On the Performance of a Two User MIMO Downlink System in Heavy Traffic

Sumit Bhardwaj, *Student Member, IEEE,* Ruth J. Williams, and Anthony S. Acampora, *Fellow, IEEE*

***Abstract*— A MIMO downlink system in which data is transmitted to two users over a common wireless channel is considered. The channel is assumed to be fixed for all transmissions over the period of interest and the ratio of anticipated average arrival rates for the two users, also known as the relative traffic rate, is the system design parameter. A packet-based traffic model is considered where data for each user is queued at the transmit end. A queueing analogue for this system leads to a coupled queueing system for which a simple policy is known to be throughput-optimal under Markovian assumptions. Since an exact expression for the performance is not available, as a measure of performance (in heavy traffic), a diffusion approximation is established. This diffusion process is a two-dimensional semimartingale reflecting Brownian motion living in the positive quadrant of two-dimensional space.**

***Index Terms*— Coupled queueing systems, diffusion approximation, heavy traffic, multi-input multi-output (MIMO), semimartingale reflecting Brownian motion (SRBM).**

I. Introduction

Current cellular systems consider each base station as a separate entity with no cooperation among base stations. Infrastructure cooperation, that is, cooperation among base stations has been proposed as a means of achieving higher throughput (see, e.g., [1], [2], [3]) where the main idea is to consider the cooperating base stations as one end of a MIMO system and then to use results from information theory for the study of cellular systems. In this correspondence, we consider a two-user MIMO downlink system where data is buffered at the transmit end and the channel is assumed to be fixed for all transmissions over the period of interest (one might view this as one period for a quasi-static channel). The two-user MIMO downlink system can be seen as a model of a cellular system with two users and two cooperating base station antennas which might be two cooperating base stations each with a single antenna or a single-cell cellular system with a multi-antenna base station. It is well known that in such a system, the sum of the rates at which data can be served for the two users is greater than the single-user capacity for any user. Thus, one can obtain improved capacity by cooperation.

This communication system 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 [3]. An exact expression for the performance of this policy is not available. In this correspondence, as a measure of performance, we prove a limit theorem justifying a diffusion approximation for a heavily loaded system operating under this policy. In particular, the diffusion is a two-dimensional semimartingale reflecting Brownian motion (SRBM) living in the positive two-dimensional quadrant (Theorem VIII.3).

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 [5], [6] (also see references therein) under restrictive assumptions. In [5], 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. [6] study a throughput-optimal scheduling rule, which they call an exponential scheduling rule, and show that 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 [5], [6]. The Maxweight policy [5] 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 V where we define the heavy traffic conditions. Moreover, a complete resource pooling (CRP) condition is assumed in [5] 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 [5] is assumed to be an ergodic Markov process while we assume that the arrival process is a renewal process. In [6], the capacity region is a convex polyhedron and a CRP condition similar to [5] is assumed; however, service is given to only one queue at a time while here we can serve both queues at the same time.

The rest of this correspondence is organized as follows. We explain the notation used in this correspondence and present some mathematical preliminaries in Section II. We describe the MIMO downlink system of interest in Section III and develop a queueing analogue for it in Section IV. We formally define the heavy traffic conditions in Section V. In Section VI, we define the scaling and present standard functional limit theorems used in proving our main results. In Section VII, we prove a fluid limit result (Lemma VII.2) for our queueing system. This plays a role in establishing the heavy traffic limit theorem through determining the fluid scale service allocations. We present the main theorem of this correspondence (Theorem VIII.3) in Section VIII which says that in the heavy traffic limit, the renormalized queuelength process converges in distribution to a SRBM living in a two-dimensional quadrant. There, we also discuss the properties of the limiting process. Finally, we summarize our conclusions in Section IX.

II. Notation and Preliminaries

We will use the following notation throughout the correspondence. 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. Vectors will be denoted by lower case bold symbols and matrices by upper case bold symbols. Let $0 = (0,0,\ldots,0) \in \mathbb{R}^d_+$.

The research of S. Bhardwaj and A. S. Acampora was supported in part by the UC Discovery Grant COM04-10174. The research of R. J. Williams was supported in part by NSF grant DMS-0604537.

S. Bhardwaj and A. S. Acampora are with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA (email: bhardwajs@ucsd.edu, acampora@ece.ucsd.edu).

R. J. Williams is with the Department of Mathematics, University of California at San Diego, La Jolla, CA 92093 USA (email: williams@math.ucsd.edu).

[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. [4, p. 26].

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 correspondence 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. functions from $[0, \infty)$ into $\mathbb{R}^d$ and we endow this space with the usual Skorokhod J_1-topology (see Ethier and Kurtz [7, 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.

A. 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 II.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

1) $z(t) = x(t) + Ry(t)$ for all $t \in \mathbb{R}_+$,
2) $z(t) \in \mathbb{R}_+^d$ for all $t \in \mathbb{R}_+$,
3) 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 [8] 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 II.2 (Harrison-Reiman (HR) Condition). A $d \times d$ matrix R satisfies the HR condition if $R = I - Q$, where I is the $d \times d$ identity matrix, Q has zeros along the diagonal, all of the entries of Q are nonnegative and Q has spectral radius strictly less than one.

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 [10, Chapter 6].

Proposition II.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 [8] and alluded to for $x \in \mathbb{D}_+^d$. A complete proof can be found in [9] 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.

III. System Model

In this section 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 correspondence, 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 (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 (the cooperating base stations) has perfect channel state information (CSI).

Weingarten et al. [11] have shown that for such a system, Dirty Paper Coding (DPC), introduced by Costa [12], 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 [13]. Figure 1 illustrates the capacity region for an example of a two-user MIMO BC with two transmit and two receive antennas. Here 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.

Let c_1^* (c_2^*) be the maximum rate at which data can be transmitted (in bits per sec (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

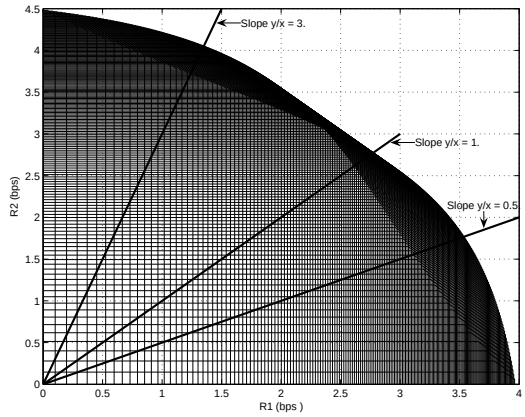

Fig. 1. An example of a capacity region of a 2-user MIMO BC for a fixed channel where R_1 and R_2 are the rates of user 1 and 2, respectively.

this third boundary the *capacity surface*. The following lemma states a key property of the capacity surface of the two-user MIMO BC.

Lemma III.1. *For any point (x,y) on the capacity surface of a two-user MIMO BC, the following holds,*

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

Proof. See Appendix I. □

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. It has been shown [3] 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 1 illustrates a few such operation points for sample values of $k_2=3,1,0.5$.

IV. Queueing Analogue

In this section we develop a queueing analogue for the system described in Section III. 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.

A. 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.

B. 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.

C. 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

$$\Lambda(q)\triangleq\begin{cases}(c_1,c_2) & \text{if } q_1>0,q_2>0,\\ (c_1^*,0) & \text{if } q_1>0,q_2=0,\\ (0,c_2^*) & \text{if } q_1=0,q_2>0,\\ (0,0) & \text{if } q_1=0,q_2=0.\end{cases} \tag{2}$$

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 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 IV.1. *For any $q\in\mathbb{R}_+^2$ and $x>0$, $\Lambda(xq)=\Lambda(q)$.*

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

D. Queuelength Process

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

$$Q_i(t)=E_i(t)-D_i(t), \tag{3}$$

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

$$D_i(t)=S_i(T_i(t)), \tag{4}$$

[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.

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

$$\begin{aligned} T_i(t) &= \int_0^t \Lambda_i(Q(s))\,ds \\ &= c_i \int_0^t 1_{\{Q_j(s)>0\ \forall j\}}\,ds + c_i^* \int_0^t 1_{\{Q_i(s)>0;\ Q_j=0\ \forall\ j\neq i\}}\,ds. \end{aligned} \tag{5}$$

V. Heavy Traffic Assumptions

A. Assumptions

We consider the operation of our queueing system in the asymptotic regime where it is heavily loaded. (Kelly and Laws [14] 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 section; 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

$$u_i^r(k) = \frac{1}{\lambda_i^r}\breve{u}_i(k) \tag{6}$$

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 $\{v_i(k)\}_{k=1}^{\infty}$, $i = 1,2$, 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 $\breve{u}_i(k)$ 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., Bell and Williams [15]). 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 V.1 (Heavy Traffic Assumption). For $i = 1,2$, there is $\theta_i \in \mathbb{R}$ such that

$$r(\lambda_i^r - \lambda_i) \to \theta_i \text{ as } r \to \infty. \tag{7}$$

Remark. This assumption does not restrict the direction in which the heavy traffic limit is approached, unlike that in Gans and Van Ryzin [16]. 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).

B. Connection to Complete Resource Pooling (CRP)

To make a connection with the work of Stolyar [5] (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 BC is shown in Fig. 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. 2, then the upper surface of the system stability region, $\tilde{\mathcal{C}}$, is shown by the dashed curve.

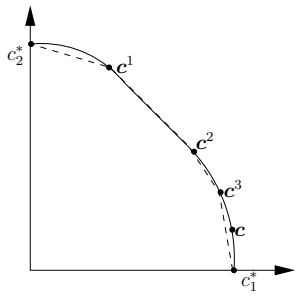

Fig. 2. The solid curve indicates the capacity surface while the surface of the system stability region is shown by the dashed line.

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 does not hold.

Lemma V.2. *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. 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. 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 [5] 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 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.

VI. Scaling and Standard Limit Theorems

A. 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.

1) 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

$$\bar{T}_i^r(t) \triangleq r^{-2} T_i^r(r^2 t), \quad \bar{Q}_i^r(t) \triangleq r^{-2} Q_i^r(r^2 t), \tag{8}$$

$$\bar{E}_i^r(t) \triangleq r^{-2} E_i^r(r^2 t), \quad \bar{S}_i^r(t) \triangleq r^{-2} S_i^r(r^2 t). \tag{9}$$

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.

2) 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

$$\hat{Q}_i^r(t) \triangleq \frac{Q_i^r(r^2 t)}{r}, \ t \geq 0, \tag{10}$$

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

$$\hat{E}_i^r(t) \triangleq \frac{E_i^r(r^2 t) - \lambda_i^r r^2 t}{r} \tag{11}$$

and

$$\hat{S}_i^r(t) \triangleq \frac{S_i(r^2 t) - \mu_i r^2 t}{r}. \tag{12}$$

B. Functional Limit Theorems for Stochastic Primitives

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

Proposition VI.1 (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.,*

$$(\hat{E}^r(\cdot), \hat{S}^r(\cdot)) \Rightarrow (B_E(\cdot), B_S(\cdot)) \text{ as } r \to \infty, \tag{13}$$

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 $\breve{u}_i(k)$ and $v_i(k)$ is required for the FCLT. Furthermore, since a Brownian motion is a continuous process, the weak-convergence of $(\hat{E}^r(\cdot), \hat{S}^r(\cdot))$ 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 [17], 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 [18]. □

As a corollary, we have the following functional law of large numbers (FLLN) for the stochastic primitives. From now for each $t \geq 0$, let $\lambda(t) = \lambda t$ and $\mu(t) = \mu t$.

Corollary VI.2 (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.,*

$$\left(\bar{E}^r(\cdot), \bar{S}^r(\cdot)\right) \Rightarrow (\lambda(\cdot), \mu(\cdot)) \text{ as } r \to \infty. \tag{14}$$

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 VI.1 implies that

$$\left(\frac{1}{r}\hat{E}^r(\cdot), \frac{1}{r}\hat{S}^r(\cdot)\right) \Rightarrow (0,0) \text{ as } r \to \infty. \tag{15}$$

The desired result follows from this and the fact that $\lambda_i^r \to \lambda_i$ as $r \to \infty$ by (7) for $i = 1,2$. □

VII. Fluid Model

Applying fluid scaling to the dynamic equation (3) satisfied by the queuelength process for the system indexed by r, we obtain for $r > 0$, $i = 1,2$, $t \geq 0$,

$$\bar{Q}_i^r(t) = \bar{E}_i^r(t) - \bar{S}_i^r(\bar{T}_i^r(t)). \tag{16}$$

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

$$\bar{T}^r(t) = \frac{1}{r^2}\int_0^{r^2 t} \Lambda(Q^r(s))\, ds, \ t \geq 0. \tag{17}$$

By the change of variables $\tilde{s} = \frac{s}{r^2}$, for $t \geq 0$, (17) becomes

$$\bar{T}^r(t) = \int_0^t \Lambda\left(\frac{r^2 Q^r(r^2\tilde{s})}{r^2}\right) d\tilde{s} = \int_0^t \Lambda\left(\bar{Q}^r(\tilde{s})\right) d\tilde{s}. \tag{18}$$

where the second equality follows from the definition of $\bar{Q}^r(\cdot)$ and the scaling property of $\Lambda(\cdot)$ (see Lemma IV.1). The following lemma follows from (18) 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 VII.1. *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 $\dot{x}$:

$$x(t) = x(0) + \int_0^t \dot{x}(s) ds, \ t \geq 0. \tag{19}$$

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

Lemma VII.2. *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 $\left(\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), Q(\cdot)\right)$ as $r \to \infty$ where*

$$\bar{E}(\cdot) = \lambda(\cdot), \quad \bar{S}(\cdot) = \mu(\cdot), \quad \bar{Q}(\cdot) = 0, \quad \bar{T}(\cdot) = c(\cdot), \tag{20}$$

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 VII.1, 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 VI.2), using (16) together with the random time change theorem of Billingsley [19, 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 $\left(\bar{E}(\cdot), \bar{S}(\cdot), \bar{T}(\cdot), Q(\cdot)\right)$ is a weak limit point of this sequence. By invoking the Skorokhod representation theorem (see, e.g., [7, 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

$$\bar{Q}_i^{r_k}(t)=\bar{E}_i^{r_k}(t)-\bar{S}_i^{r_k}(\bar{T}_i^{r_k}(t)) \text{ for } t\geq 0, i=1,2 \tag{21}$$

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

$$\left(\bar{E}^{r_k}(\cdot),\bar{S}^{r_k}(\cdot),\bar{T}^{r_k}(\cdot),\bar{Q}^{r_k}(\cdot)\right)\to\left(\lambda(\cdot),\mu(\cdot),\bar{T}(\cdot),\bar{Q}(\cdot)\right) \text{ u.o.c.} \tag{22}$$

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

$$\frac{d\bar{Q}_i(t)}{dt}=\lambda_i-\mu_i\frac{d\bar{T}_i(t)}{dt},\ i=1,2. \tag{23}$$

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,

$$0=\lambda_i-\mu_i\frac{d\bar{T}_i(t)}{dt}, \tag{24}$$

which implies that

$$\frac{d\bar{T}_i(t)}{dt}=\frac{\lambda_i}{\mu_i}=c_i. \tag{25}$$

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

$$\begin{aligned}\bar{T}_i(s)-\bar{T}_i(t)&=\lim_{r\to\infty}[\bar{T}_i^r(s)-\bar{T}_i^r(t)]=\lim_{r\to\infty}\left[\int_t^s\Lambda_i\left(\bar{Q}_i^r(z)\right)dz\right]\\&=\lim_{r\to\infty}\left[\int_t^s c_i\,dz\right]=c_i(s-t),\end{aligned} \tag{26}$$

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 (23) 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, $Q_i^r(s)>0$ for all $s\in[u,v]$, which implies by the definition of $\Lambda_i(\bar{Q}^r(\cdot))$ that

$$c_i(s-t)\leq\bar{T}_i^r(s)-\bar{T}_i^r(t)\leq c_i^*(s-t) \text{ for all } s>t \text{ in } [u,v]. \tag{27}$$

Letting $r\to\infty$ yields

$$c_i(s-t)\leq\bar{T}_i(s)-\bar{T}_i(t)\leq c_i^*(s-t), \text{ for all } s>t \text{ in } [u,v]. \tag{28}$$

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 (23), since $\lambda_i=\mu_i c_i$,

$$d\bar{Q}_i(t)/dt\leq 0. \tag{29}$$

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

$$\frac{d}{dt}\left(\bar{Q}_1^2(t)+\bar{Q}_2^2(t)\right)=2\left[\bar{Q}_1(t)\frac{d\bar{Q}_1(t)}{dt}+\bar{Q}_2(t)\frac{d\bar{Q}_2(t)}{dt}\right]\leq 0. \tag{30}$$

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 $\bar{T}_i(t)=c_i$ 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 (20). □

VIII. Diffusion Approximation

A. Pre-limit process

From (3), (4), (8), (11) and (12), the diffusion scaled queuelength process can be written for $i=1,2$, $t\geq 0$, as

$$\begin{aligned}\hat{Q}_i^r(t)&=(\hat{E}_i^r(t)+\lambda_i^r rt)-(\hat{S}_i^r(\bar{T}_i^r(t))+\mu_i r\bar{T}_i^r(t))\\&=\hat{E}_i^r(t)-\hat{S}_i^r(\bar{T}_i^r(t))+r(\lambda_i^r t-\mu_i\bar{T}_i^r(t)).\end{aligned} \tag{31}$$

Expanding the last term in (31), we have

$$\begin{aligned}r(\lambda_i^r t-\mu_i\bar{T}_i^r(t))&=\frac{r^2\lambda_i^r t-\mu_i r^2\bar{T}_i^r(t)}{r}\\&=\frac{\left(\lambda_i^r-\lambda_i\right)r^2t+\lambda_i\int_0^{r^2t}ds-\mu_i\int_0^{r^2t}\Lambda_i(Q^r(s))ds}{r}.\end{aligned} \tag{32}$$

Considering four different types of states for the queuelength vector Q^r and substituting the corresponding values for $\Lambda_i(Q^r(\cdot))$ from (2), we can rewrite (32) as

$$\begin{aligned}r(\lambda_i^r t-\mu_i\bar{T}_i^r(t))&=(\lambda_i^r-\lambda_i)rt+\frac{1}{r}\Bigg[(\lambda_i-\mu_i c_i)\int_0^{r^2t}1_{\{Q^r(s)>0\}}ds\\&+(\lambda_i-\mu_i c_i^*)\int_0^{r^2t}1_{\{Q_i^r(s)>0;Q_j^r(s)=0,j\neq i\}}ds\\&+\lambda_i\int_0^{r^2t}1_{\{Q_i^r(s)=0;Q_j^r(s)>0,j\neq i\}}ds+\lambda_i\int_0^{r^2t}1_{\{Q_j^r(s)=0\ \forall\ j\}}ds\Bigg].\end{aligned} \tag{33}$$

Define for $t\geq 0$,

$$\begin{aligned}\hat{U}_i^r(t)&\triangleq\frac{1}{r}\int_0^{r^2t}1_{\{Q_i^r(s)=0;Q_j^r(s)>0,j\neq i\}}ds\\&=r\int_0^t 1_{\{\hat{Q}_i^r(s)=0;\hat{Q}_j^r(s)>0,j\neq i\}}ds,\quad i=1,2,\end{aligned} \tag{34}$$

$$\hat{Z}^r(t)\triangleq\frac{1}{r}\int_0^{r^2t}1_{\{Q_j^r(s)=0\ \forall\ j\}}ds=r\int_0^t 1_{\{\hat{Q}_j^r(s)=0\ \forall\ j\}}ds. \tag{35}$$

Then, using the fact that $\lambda_i=\mu_i c_i$ and combining (31)–(35), we obtain for $i=1,2$, $t\geq 0$,

$$\hat{Q}_i^r(t)=\hat{X}_i^r(t)+\lambda_i\hat{U}_i^r(t)+(\lambda_i-\mu_i c_i^*)\hat{U}_j^r(t)+\lambda_i\hat{Z}^r(t), \tag{36}$$

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

$$\hat{X}_i^r(t)=\hat{E}_i^r(t)-\hat{S}_i^r(\bar{T}_i^r(t))+(\lambda_i^r-\lambda_i)rt. \tag{37}$$

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

$$\hat{Q}^r(t)=\hat{X}^r(t)+\begin{bmatrix}\lambda_1 & \lambda_1-\mu_1c_1^*\\ \lambda_2-\mu_2c_2^* & \lambda_2\end{bmatrix}\hat{U}^r(t)+\begin{bmatrix}\lambda_1\\ \lambda_2\end{bmatrix}\hat{Z}^r(t). \tag{38}$$

Define the *reflection matrix* R as

$$R\triangleq\begin{bmatrix}1 & \frac{\lambda_1-\mu_1c_1^*}{\lambda_2}\\ \frac{\lambda_2-\mu_2c_2^*}{\lambda_1} & 1\end{bmatrix} \tag{39}$$

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

$$\hat{Y}_i^r(t)\triangleq\lambda_i\left(\hat{U}_i^r(t)+\frac{c_i^*c_j}{c_1^*c_2+c_1c_2^*-c_1^*c_2^*}\hat{Z}^r(t).\right) \tag{40}$$

Then, (38) can be written as

$$\hat{Q}^r(t) = \hat{X}^r(t) + R\hat{Y}^r(t),\ t \geq 0. \tag{41}$$

Note that $c_1^* c_2 + c_1 c_2^* - c_1^* c_2^* > 0$ (from Lemma III.1) 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 queuelength processes $\{\hat{Q}^r(\cdot)\}$.

Lemma VIII.1. *The sequence $\{\hat{X}^r(\cdot)\}$ converges in distribution to a Brownian motion with diagonal covariance matrix $\Gamma \triangleq \text{diag}(\lambda_1(\alpha_1^2 + \beta_1^2), \lambda_2(\alpha_2^2 + \beta_2^2))$ and drift vector $\theta \triangleq (\theta_1, \theta_2)$, that starts from the origin.*

Proof. Let $\hat{\theta}^r(t) = r(\lambda^r - \lambda)t$, $t \geq 0$. By combining Proposition VI.1, Lemma VII.2 and Assumption V.1, we have that the sequence of processes $\left\{\left(\hat{E}^r(\cdot), \hat{S}^r(\cdot), \overline{T}^r(\cdot), \hat{\theta}^r(\cdot)\right)\right\}$ converges in distribution to $(B_E(\cdot), B_S(\cdot), c(\cdot), \theta(\cdot))$ where $B_E(\cdot)$ and $B_S(\cdot)$ are independent 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 (37), using the random time change lemma, $\{\hat{X}^r(\cdot)\}$ converges in distribution to a two-dimensional Brownian motion with diagonal covariance matrix $\text{diag}(\lambda_1\alpha_1^2 + \mu_1 c_1 \beta_1^2, \lambda_2\alpha_2^2 + \mu_2 c_2 \beta_2^2) = \text{diag}(\lambda_1(\alpha_1^2 + \beta_1^2), \lambda_2(\alpha_2^2 + \beta_2^2))$ (since $\lambda_i = \mu_i c_i$ for $i = 1,2$), drift vector (θ_1, θ_2) and starting point $(0,0)$. □

B. Limit Theorem

In this subsection, we 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 correspondence.

Define

$$Q \triangleq I - R = \begin{bmatrix} 0 & \frac{\mu_1 c_1^* - \lambda_1}{\lambda_2} \\ \frac{\mu_2 c_2^* - \lambda_2}{\lambda_1} & 0 \end{bmatrix} \tag{42}$$

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 Q are nonnegative. We next show that the matrix R satisfies the HR condition described in Section II-A.

Lemma VIII.2. *The reflection matrix R satisfies the HR condition.*

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

$$x^2 - \frac{(\mu_1 c_1^* - \lambda_1)(\mu_2 c_2^* - \lambda_2)}{\lambda_1 \lambda_2} = 0. \tag{43}$$

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

$$x = \pm\sqrt{\left(\frac{c_1^*}{c_1} - 1\right)\left(\frac{c_2^*}{c_2} - 1\right)}. \tag{44}$$

Thus the spectral radius of Q 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 Q is strictly less than one. Thus R satisfies the HR condition. □

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

Theorem VIII.3 (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 II-A. 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 (41). 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 II.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 VIII.1, 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. □

C. 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.

1) 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 VIII.4. *Almost surely, the Lebesgue measure of the time spent by $\hat{Q}$ at $(0,0)$ is zero.*

Proof. Varadhan and Williams [20] 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 [21, §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. □

2) Stationary Distribution: Harrison and Williams [22] 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 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-

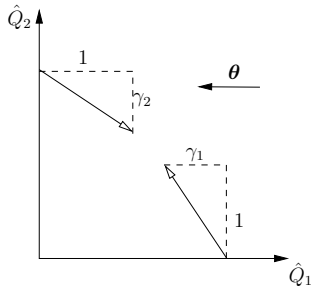

Fig. 3. Directions of reflection and drift for an example of an SRBM with $\gamma_1 = \frac{\mu_1 c_1^* - \lambda_1}{\lambda_2}$, $\gamma_2 = \frac{\mu_2 c_2^* - \lambda_2}{\lambda_1}$, and $\theta = (-1,0)$.

dimensional SRBMs, Avram et al. [23] 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 [24] 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.

IX. Concluding Remarks

In this correspondence, we studied the performance of a two-user MIMO downlink system in heavy traffic. For this coupled queueing system, we considered a simple throughput-optimal policy which, given the ratio of the bit arrival rates, depends only upon the empty/non-empty state of the queues. Since an exact expression for performance is not available, we have established a diffusion approximation as a measure of performance. The diffusion process is a two-dimensional SRBM that starts from the origin and lives in the positive orthant of two-dimensional space.

An interesting question is how does this policy compare to other policies which use more information about the system such as MaxWeight [5]? Another question that might be asked is whether or not the stationary distribution of the SRBM is the limit of the stationary distribution of the original queueing system.

Appendix I
Proof of Lemma III.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 segment and so for any point on the capacity surface we have

$$\frac{x}{c_1^*}+\frac{y}{c_2^*}\geq 1. \tag{45}$$

From (45) and the convexity of the capacity region, if there is a point on the capacity surface where (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 (1) holds.

The sum-rate capacity of the MIMO BC is defined as the maximum of the sum of a pair of rates that can be transmitted. (See Viswanath et al. [25] for details.) If the sum-rate capacity of the MIMO BC is strictly greater than the single-user capacities, c_1^* and c_2^*, then (1) holds at the point(s) achieving sum-rate capacity. This follows by noting that if only equality held in (45), 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 [25, Theorem 3], the sum-rate capacity of MIMO BC is the Sato upper bound [26] which is greater than the single-user capacities. Thus, there is a point on the capacity surface where (1) holds, and the lemma follows. □

References

[1] H. Viswanathan and K. Kumaran, "Rate scheduling in multiple antenna downlink wireless systems," *IEEE Transactions on Communications*, vol. 4, no. 53, pp. 645–655, Apr. 2005.

[2] K. Kumaran and H. Viswanathan, "Joint power and bandwidth allocation in downlink transmission," *IEEE Transactions on Wireless Communications*, vol. 4, no. 3, pp. 1008–1016, May 2005.

[3] A. S. Acampora, S. Bhardwaj, and R. M. Tamari, "On best-case throughput of cellular data networks with cooperating base stations," in *Proc. of the Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Sept. 27–29 2006.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] J. M. Harrison and M. I. Reiman, "Reflected Brownian motions on an orthant," *Ann. Probab.*, vol. 9, no. 2, pp. 302–308, 1981.

[9] A. R. K. Whitley, "Skorokhod problems and semimartingale reflecting stable processes in an orthant," Ph.D. dissertation, University of California, San Diego, 2003.

[10] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*. New York: Academic Press, 1979.

[11] H. Weingarten, Y. Steinberg, and S. Shamai (Shitz), "The capacity region of the Gaussian MIMO broadcast channel," in *Proc. of the IEEE ISIT*, July 2004.

[12] M. H. M. Costa, "Writing on dirty paper," *IEEE Transactions on Information Theory*, vol. IT-29, no. 3, pp. 439–441, 1983.

[13] N. Jindal, S. Vishwanath, and A. Goldsmith, "On the duality of Gaussian multiple-access and broadcast channels," *IEEE Transactions on Information Theory*, vol. 50(05), pp. 768–783, May 2004.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] Y. V. Prokhorov, "Convergence of random processes and limit theorems in probability theory," *Theory Probab. Appl.*, vol. 1, no. 2, pp. 157–214, 1956.

[19] P. Billingsley, *Convergence of probability measures*, 2nd ed. New York: Wiley, 1999.

[20] 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.

[21] K. L. Chung and R. J. Williams, *Introduction to Stochastic Integration*, ser. Probability and its Applications. Boston, MA: Birkhäuser, 1990.

[22] J. M. Harrison and R. J. Williams, "Brownian models of open queueing networks with homogeneous customre populations," *Stochastics*, vol. 22, pp. 77–115, 1987.

[23] 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.

[24] 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.

[25] 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.

[26] 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.