Optimal Operating Point for MIMO Multiple Access Channel With Bursty Traffic

Somsak Kittipiyakul and Tara Javidi, *Member, IEEE*

***Abstract*— Multiple antennas at the transmitters and receivers in a multiple access channel (MAC) can provide simultaneous diversity, spatial multiplexing, and space-division multiple access gains. The fundamental tradeoff in the asymptotically large SNR regime is shown by Tse et al. (2004). On the other hand, MAC scheduling can provide a statistical-multiplexing gain to improve the delay performance as shown by Bertsimas et al. (1998) and Stolyar and Ramanan (2001). In this paper, we formulate and analytically derive bounds on the optimal operating point for MIMO-MAC channel for bursty sources with delay constraints. Our system model brings together the four types of gains: diversity, spatial multiplexing, space-division multiple-access, and statistical-multiplexing gains. Our objective is to minimize the end-to-end performance as defined by the delay bound violation probability as well as the channel decoding error probability. We find the optimal diversity gain and rate region in which the system should operate. As an example, we illustrate our technique and the optimal operating point for the case of a compound Poisson source. In addition, we note an interesting interplay between the intensity of the traffic and resource pooling with regard to both multiple-access and statistical-multiplexing gains.**

***Index Terms*— cross-layer optimization, diversity-multiplexing tradeoff, MIMO multiple access channel, statistical-multiplexing.**

I. Introduction

Multiple antennas can be used to enhance the performance of wireless systems. The multiple antennas can be used to simultaneously boost the reliability (providing *diversity* gain) and the data rate (providing *spatial multiplexing* gain). In addition, in multiple access scenarios where multiple users are transmitting to a common receiver, multiple receive antennas also provide *multiple-access* gain by allowing for spatial separation of the signals of different users. Tse, Viswanath, and Zheng [2] have characterized the fundamental tradeoff between these three types of gains at high SNR.

Our goal in this paper is to answer the question first posed by Holliday and Goldsmith in [5]: “given the diversity-multiplexing region, where should one choose to operate?”. Holliday and Goldsmith answered this question in the context of the cross-layer design of a point-to-point system where a source encoder is concatenated with a MIMO link. Their goal in this context was to minimize the end-to-end distortion. In their later papers [6] and [7], Holliday et al generalized their formulation to include the distortion due to delay, where the delay is caused by random service time of the ARQ process.

In this paper, we answer the same question in a multi-user context. We consider a cross-layer queue-channel optimization problem for bursty and delay-sensitive traffic sources. We consider a system where each user has a bursty source concatenated with an infinite buffer and a MIMO multiple access channel (MIMO-MAC). The end-to-end performance metric of interest is the total bit loss probability, where loss can be due to either delay violation or decoding errors in the MIMO-MAC channel. From a user’s perspective, we face the following tradeoff: the higher the multiplexing gain the better the delay performance, but the inevitable decrease in diversity results in an increase in MIMO channel errors. At the same time, the statistical variation in the traffic patterns among users provides us with flexibility in allocating the resources.

The main contribution of this paper is the formulation of a cross-layer optimal operating point for a MIMO-MAC channel with bursty sources and delay constraints. In particular, we provide a methodology for characterizing the optimal diversity gain and rate region in which the system should operate in a MIMO-MAC channel with a given high SNR and description of the bursty traffic sources. To achieve this, we assume an optimal scheduler design which dynamically controls users’ transmission rates (or equivalently, the multiplexing gains) as a function of queue backlogs. This dynamic adaptation of multiplexing gains accounts for *statistical-multiplexing* while leveraging the known tradeoff between diversity, spatial multiplexing, and multiple-access gains given in [2]. From a scheduling perspective, statistical-multiplexing is a key mechanism by which the network resources are used to improve the delay performance for bursty users. In particular, statistical-multiplexing capitalizes on the fact that peaks in traffic of simultaneously ongoing traffic streams rarely coincide. We believe that our result can be viewed as a first step in integrating the known spatial diversity and multiplexing and multiple-access gains with that of the statistical-multiplexing. In other words, for the first time, our model brings together the four types of gains offered at a MIMO MAC.

The remainder of the paper is organized as follows. In Section II, we provide a detailed description of the system model as well as the problem formulation for general number of users. In Section III, we provide the main analytical results and bounds on the optimal channel diversity gain. We also discuss the notion of *statistical-multiplexing* and its benefits. In Section VI, we find the optimal operating diversity gain

Manuscript received June 13, 2006; revised April 4, 2007; accepted May 25, 2007. The associate editor coordinating the review of this paper and approving it for publication was Dr. Stephen Hanly. This work was supported in part by the Center for Wireless Communications, UCSD and UC Discovery Grant No. Com04-10176, ARO-MURI Grant No. W911NF-04-1-0224, NSF CAREER Award No. CNS-0533035, and AFOSR Grant No. FA9550-05-01-0430.

The authors are with the Department of Electrical and Computer Engineering, University of California/San Diego, La Jolla, CA 92093 USA (email: skittipi@ucsd.edu; tara@ece.ucsd.edu).

Digital Object Identifier 10.1109/TWC.2007.060362.

d^* or its bounds for a particular class of compound Poisson sources. Finally, in Section V, we discuss the shortcomings of the present paper, possible extensions, and future work.

II. System Model

We consider the architecture shown in Fig. 1. The system is time-slotted and consists of three main components, each shown with a different number. The first component consists of K homogeneous users, each of which has an identical but independent bursty source: each source generates information bits according to a stochastic process. When appropriate, the bits are buffered prior to transmission over the channel. The second component of interest is a MIMO multiple access channel without channel state information (CSI) at the transmitters but with perfect CSI at the receiver. The receiver consists of a joint maximum-likelihood decoder. In the absence of CSI at the transmitters, we assume that the MIMO-MAC operates at a common diversity gain, which in turn specifies the corresponding capacity region of the MIMO MAC channel as given in [2]. However, the individual rate of each user is determined dynamically by the rate scheduler which is the third component in our system. This is a centralized rate scheduler that dynamically determines the transmission rates of the individual users given queue state information (QSI) of each user.

In this paper we are interested in the following questions: given a statistical description of the sources, a large delay bound D, and a *high* SNR value of the channel, what is the optimal design of the scheduler, and what is the optimal operating diversity gain of the MIMO MAC channel. The notion of optimality needs to take into account the channel decoding error as well as the delay bound violation probabilities. We assume no restransmission of the bits in error and map our objective to the sum of the probability of delay violation and the probability of channel error. In order to mathematically define this problem, we now model each of the above components precisely.

A. Source Model

We assume that the total number of information bits generated by user i ($i = 1, \ldots, K$) is given by a sequence $S^i = \left\{S^i_t, t = 1, 2, \ldots\right\}$, where S^i_t is the total number of bits of user i generated up to timeslot t and $S^i_0 \equiv 0$. In addition, we assume that the arrival processes S^i, $i = 1, \ldots, K$, are identical and mutually independent. We also assume that each arrival process S^i has stationary increments and satisfies a *Large Deviations Principle* (LDP). In the appendix, we discuss an additional sample path LDP assumption (Assumption B) on the arrival processes. Here, to keep the flow of the paper, in this section we only provide the LDP assumption and the consequent characterization of the sources which is based on LDP.

In general, consider a source process S generating a sequence $\{S_t, t = 1, 2, \ldots\}$ of random variables, where S_t is the total number of bits generated up to timeslot t.

Definition 1: A source S is said to satisfy an LDP with *decay function*[1] $\Lambda^* : \mathbb{R} \to [0, \infty]$ if, for large enough t and for small $\epsilon > 0$,

$$\Pr\left[\frac{S_t}{t} \in (a - \epsilon, a + \epsilon)\right] \approx e^{-t\Lambda^*(a)} \tag{1}$$

where Λ^* is a lower semicontinuous function and has compact level sets[2] (see [3] and [9] for more discussions in LDP).

Fact 1: (Gärtner-Ellis theorem) Suppose a source S satisfies the following:

Assumption A:

1) The limiting log-moment generating function
$$\Lambda(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}[e^{\theta S_t}] \tag{2}$$
exists for all θ, where $\pm\infty$ are allowed both as elements of the sequence and as limit points.
2) The origin is in the interior of the domain $D_\Lambda := \{\theta | \Lambda(\theta) < \infty\}$ of $\Lambda(\theta)$.
3) $\Lambda(\theta)$ is differentiable in the interior of D_Λ and the derivative tends to infinity as θ approaches the boundary of D_Λ.
4) $\Lambda(\theta)$ is lower semicontinuous, i.e. $\liminf_{\theta_n \to \theta} \Lambda(\theta_n) \geq \Lambda(\theta)$ for all θ.

Then the source S satisfies an LDP and its decay function described by (1) is given as

$$\Lambda^*(a) = \sup_\theta \, [\theta a - \Lambda(\theta)]. \tag{3}$$

Remark 1: It can be shown that Λ^* is a convex function taking values in $[0, \infty]$ such that $\Lambda^*(E[S_1]) = 0$ where $E[S_1]$ is the average arrival rate of process S [9].

Remark 2: Many source models commonly used to model bursty traffic in communication networks satisfy Assumptions A and B. Such source models include renewal processes, Markov-modulated processes, and more generally stationary processes with mild mixing conditions [3].

Remark 3: From LDP, a source is fully characterized by either $\Lambda(\cdot)$ or $\Lambda^*(\cdot)$. An alternative to the LDP characterization of a source is the well-known *effective bandwidth* (see [8]) which was used in our previous study for a point-to-point scenario [14].

B. MIMO-MAC Channel Model and PHY Model

We use the same symmetric MIMO multiple access channel model as described in [2] which assumes symmetric transmitters seeing i.i.d. fading channels, perfect symbol synchronization and perfect CSI at the receiver but no CSI at any transmitters. Each transmitter has M transmit antennas, while the receiver has N receive antennas. Space-time coding happens over a channel coherence time which is assumed to contain T symbols[3]. We assume the duration of a timeslot is equal to the channel coherence period, i.e., a timeslot contains

[1] In large deviations literature, the Λ^* function is typically called a “rate” function. Here we use the name decay function to avoid confusion with transmission rate.

[2] The level set $\{x : \Lambda^*(x) \leq a\}$ is compact for every real a

[3] We assume either a sufficiently large symbol rate or a sufficiently small number of antennas such that $T \geq KM + N - 1$.

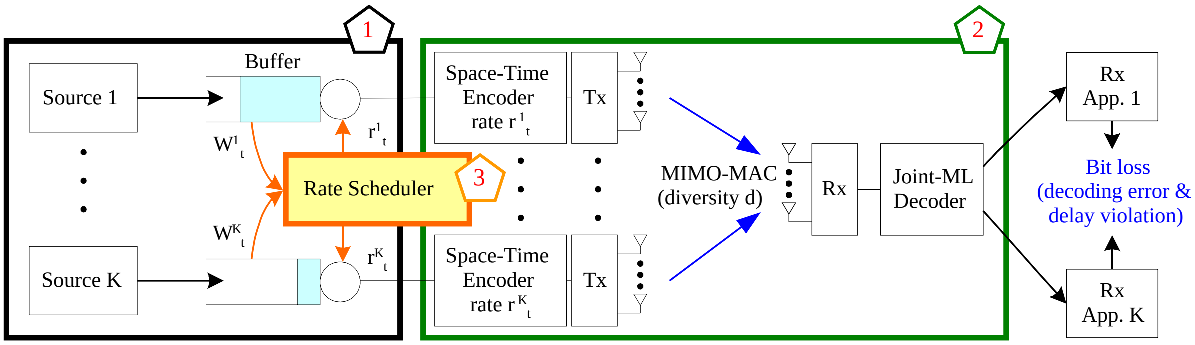


Fig. 1. System model and the two causes of bit loss: delay violation and channel decoding error.

T symbols. Since the transmitters are assumed to know only the channel statistics, including the average received SNR, they always transmit at the maximum powers which are equal for all transmitters. The channel fading processes of the transmitters are assumed to be stationary over time, mutually independent, and identical. For each transmitter, the channel fadings for different antenna paths are assumed to be slow block-fading with i.i.d. Rayleigh fading where the fadings stay constant during a timeslot and change independently and simultaneously over timeslots. We denote by ρ the average received signal-to-noise ratio (SNR) at each receive antenna.

From a system perspective, at each SNR level ρ, the PHY layer for the MIMO-MAC channel provides a tradeoff between the reliability of the transmissions and the transmission rates. Equivalently, we can say that the PHY layer provides a tradeoff between a common diversity d and the multiplexing gain region, denoted by $\mathcal{R}(d)$, where d and $\mathcal{R}(d)$ are as defined in [1] and [2]. We state these definitions below.

Definition 2: (Definition 1 in [1]) A code scheme $\{C(\rho)\}$, which is a family of codes (coding over one single coherence block) with one codebook for each SNR level ρ and provides data rate $R(\rho)$ and average error probability $P_e(\rho)$, achieves *multiplexing gain* r and *diversity gain* d if

$$\lim_{\rho\to\infty} \frac{R(\rho)}{\log\rho} = r \text{ and } \lim_{\rho\to\infty} \frac{\log P_e(\rho)}{\log\rho} = -d. \tag{4}$$

For notational simplicity we shorten (4) as $R(\rho) \doteq r\log\rho$ and $P_e(\rho) \doteq \rho^{-d}$. We also use $\dot{\leq}$ and $\dot{\geq}$ if $\leq$ and $\geq$ hold in the limit. [4]

Definition 3: (Theorem 2 in [2]) Let r^i_t be the multiplexing gain of user i, $i = 1,\dots,K$, at time t. Given a common diversity requirement d for all users, i.e.,

$$P^i_e \dot{\leq} SNR^{-d}, \quad i = 1,\dots,K, \tag{5}$$

where P^i_e is the average error probability for user i. Then the spatial multiplexing gains $(r^1_t,\dots,r^K_t)$ at any timeslot t must be within the (time-independent) multiplexing gain region

$$\mathcal{R}(d) = \left\{(r^1,\dots,r^K) : \sum_{s\in S} r^s \leq r^*_{|S|M,N}(d), \forall S \subseteq \{1,\dots,K\}\right\}. \tag{6}$$

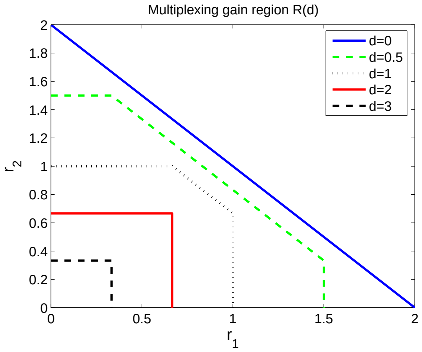


Fig. 2. Example of the multiplexing gain region $\mathcal{R}(d)$ for $M = N = 2$ case.

where $r^*_{m,n}(d)$ for any integers m and n is the largest multiplexing gain achieved for an $m \times n$ point-to-point MIMO link for a given diversity d and is defined as a piecewise linear function joining the points $((m-k)(n-k), k)$ for $k = 0,\dots,\min(m,n)$.

In this paper, we consider a system which always operates at a common diversity gain d at any time t. This d directly determines the multiplexing region $\mathcal{R}(d)$ and its shape. In particular, d determines the sum of all the rates at all time t, i.e. $\sum_{i=1}^K r^i_t \leq r^*_{KM,N}(d)$, which is independent of time. However, the individual rate at time t, $r^i_t, i = 1,\dots,K$, is determined dynamically by the rate scheduler discussed later.

In Fig. 2, we illustrate the dependence of the shape of $\mathcal{R}(d)$ and d for a simple case of $K = 2$ users and $M = N = 2$. As seen in this figure, there exists a diversity gain d_0 (in this example, $d_0 = 2$) such that, for large d ($d > d_0$), the shape of $\mathcal{R}(d)$ follows a rectangular shape (*single-user performance* regime), while, for small d ($d < d_0$), $\mathcal{R}(d)$ is a polymatroid shape (*antenna-pooling* regime). Furthermore, [2] shows that d_0 is the unique solution to

$$r^*_{KM,N}(d_0) = Kr^*_{M,N}(d_0). \tag{7}$$

Later we will see the impact of this change of shape on the working of the scheduler block.

C. Rate Scheduler

Given that each user operates at a fixed and common diversity gain d and given an average SNR of ρ in the MIMO-MAC subsystem, the function of scheduler $g_d : \mathbb{R}^K_+ \mapsto \mathcal{R}(d)$

[4] Note that we use the natural log instead of $\log_2$ and hence use nats instead of bits. This is for convenience of the presentation.

is to allocate, at the beginning of every timeslot t, the set of feasible multiplexing gains to the users. This is done, equivalently, by selecting a vector of spatial multiplexing gains $(r_t^1, \ldots, r_t^K)$ from the multiplexing gain region $\mathcal{R}(d)$. The decision is based on the delay of the head-of-the-line bit in queue i, denoted by W_t^i, $i = 1, \ldots, K$, at the beginning of timeslot t. Specifically, we assume that

$$(r_t^1, \ldots, r_t^K) = g_d(W_t^1, \ldots, W_t^K).$$

Without loss of optimality, one can assume that the rate scheduler always assigns the highest possible sum rate. At timeslot t, an amount of $r_t^i T \log \rho$ bits are taken out from head-of-the-line of the buffer of user i. We assume that if any buffers do not have enough data to transmit, the null data is used to fulfill the rates.

Remark 4: Recall that the shape of the multiplexing gain $\mathcal{R}(d)$ depends on d (e.g. see Fig. 2). As a result, the choice of diversity gain d determines the class of feasible dynamic schedulers. In the single-user performance regime ($d \geq d_0$), the users are decoupled and independent from one another, hence, reducing the scheduler to a static (and decoupled) choice of multiplexing gain $r_t^i = r^*_{M,N}(d)$ for all $i = 1, \ldots, K$ and all time $t \in \mathbb{Z}$. For the antenna-pooling regime ($d < d_0$), on the other hand, $\mathcal{R}(d)$ is a polymatroid. In other words, in this regime, the multiplexing gains of the users are dependent on one another and must be jointly allocated.

Remark 5: The model in this paper assumes that there is no CSI available at the transmitters and the central scheduler. However, the scheduler has perfect knowledge of the queue state information (QSI). This is not unrealistic given the fact that it is less bandwidth consuming and more accurate to send the QSI of each buffer (an observable scalar number) to the centralized scheduler than to *estimate* CSI for MIMO channels (K matrices, each of dimension $M \times N$) at the receiver and feed back these matrices to the transmitters.

Remark 6: Due to lack of CSI, the role of the rate scheduler in this paper is not to minimize the channel error performance; instead, the scheduler improves the delay violation probability by taking advantage of the statistical-multiplexing gain provided by the multiple bursty sources sharing the multiple access channel.

D. Arrival Rate Scaling and Stability Condition

Since the rates of transmission in the MIMO-MAC channel are scaled as $\log \rho$, we scale the arrival rates with $\log \rho$ as well. In other words, we assume that the average bit arrival rate $\tilde{\lambda}$ of each user is

$$\tilde{\lambda} = \lambda T \log \rho \tag{8}$$

bits per timeslot for a given constant positive λ.

In addition, to guarantee system stability, we require that the total average arrival rate to be no greater than the (sum) capacity of the MIMO-MAC channel [4]. In particular, we assume that

$$K\tilde{\lambda} < \min(KM, N) T \log \rho,$$

or equivalently

$$\lambda < \min(M, N/K). \tag{9}$$

Since the system is stable, it reaches a steady state. We let W^i and L^i denote the steady-state delay and queue length, respectively, for queue i, $i = 1, \ldots, K$.

For the rest of the paper, we denote $r_{av}(d)$ as the average multiplexing gain at the common diversity d, defined as

$$r_{av}(d) := \begin{cases} r^*_{M,N}(d) & \text{if } d \geq d_0, \\ \frac{1}{K} r^*_{KM,N}(d) & \text{if } d < d_0. \end{cases} \tag{10}$$

and denote

$$C_{av}(d) := r_{av}(d) T \log \rho \quad \text{(bits per timeslot)} \tag{11}$$

as the average per-queue channel capacity at diversity d.

E. Objective

Our system objective is to find the optimal operating channel diversity gain d^* and the corresponding multiplexing gain region $\mathcal{R}(d^*)$ in which the system should operate. This diversity d^* minimizes the end-to-end total bit loss probability caused by two phenomenas: 1) delay violation of the delay bound D and 2) channel decoding error. [5]

In particular, we define the following probabilities:

$$\begin{aligned} P_e^i(d) &:= \Pr[\text{decoding error for user } i] \\ P_e(d) &:= \Pr[\text{decoding error for any user}] \qquad (12) \\ P_q^i(d) &:= \Pr[\text{delay violation of user } i] \\ &= \Pr[W^i > D] \\ P_q(d) &:= \Pr[\text{delay violation for any user}] \\ &= \Pr[\max_{i=1,\ldots,K} W^i > D]. \qquad (13) \end{aligned}$$

With the above definitions, the total loss probability $P_{\text{tot}}(d)$ is expressed as

$$\begin{aligned} P_{\text{tot}}(d) &:= \max_{i=1,\ldots,K} \Pr[\text{bit loss for user } i] \\ &= \max_{i=1,\ldots,K} \left\{ P_e^i(d) + (1 - P_e^i(d)) P_q^i(d) \right\} \end{aligned}$$

where $P_e^i(d) + (1 - P_e^i(d)) P_q^i(d)$ is the total bit loss probability of user i due to channel and delay violation. We will later show that $P_e^i(d) \doteq P_e(d) \doteq \rho^{-d}$ and $P_q^i(d) \doteq P_q(d) \doteq \rho^{-g(d)}$ where g is a functional taking positive values. Hence, the asymptotic large-SNR expression of $P_{\text{tot}}(d)$ is given as

$$P_{\text{tot}}(d) \doteq P_e(d) + P_q(d) \doteq \rho^{-d} + \rho^{-g(d)}. \tag{14}$$

We note that both probabilities $P_e(d)$ and $P_q(d)$ are functions of the diversity gain d as well as the average SNR ρ. However, there is a tradeoff between the two probabilities as a function of d: Intuitively, for a fixed ρ, we expect that a high diversity gain, which translates into a smaller transmission rate region, results in faster queue build-up and larger delays. On the other hand, this higher diversity gain yields better channel performance.

In the remainder of the paper, we will derive analytically large SNR approximations for $P_q(d)$ and $P_e(d)$ and show that

[5] We assume no retransmission for the lost bits due to channel decoding errors or delay violation. Furthermore, the source processes are not effected by the lost bits.

given a fixed and high ρ, $P_q(d)$ is increasing on d while $P_e(d)$ is decreasing on d (confirming the above intuition). Furthermore, we find the best PHY layer operating point, i.e. diversity gain d^*, so as to minimize the total bit loss probability P_{tot} in the high SNR regime. In other words, we will find d^* that balances the exponents of the two probabilities.

III. Problem Analysis

In this section, we analytically derive the two loss probabilities, $P_e(d)$ and $P_q(d)$, for asymptotically large SNR. As we will see, the two probabilities decay exponentially with SNR. For the channel, the definition of diversity gain [2] gives a direct asymptotic approximation of $P_e(d)$ for large SNR. Obtaining the asymptotic $P_q(d)$, however, requires more work. Depending on the value of d, we either directly compute the asymptotic $P_q(d)$ or provide lower and upper bounds of the asymptotic $P_q(d)$.

A. Asymptotic $P_e(d)$

The asymptotic expression of $P_e(d)$ for large SNR comes directly from the definition of diversity gain given in Definitions 2 and 3. By the union bound and the symmetry among users, we have the following bounds:

$$P_e^1(d) \le P_e(d) \le K P_e^1(d)$$

where $P_e^1(d)$ is the probability of decoding error for user 1. Using $P_e^1(d) \doteq \rho^{-d}$ in Definition 2 and the fact that K is a constant independent of ρ, we have

$$P_e(d) \doteq \rho^{-d}. \tag{15}$$

B. Asymptotic $P_q(d)$

Similarly, by the union bound and the symmetry among users, we have the following bounds:

$$\Pr[W^1 > D] \le P_q(d) \le K \Pr[W^1 > D], \tag{16}$$

where $\Pr[W^1 > D]$ implicitly depends on d.

Now, let us first focus on $\Pr[W^1 > D]$. To get an analytical expression for the asymptotic $\Pr[W^1 > D]$, we consider two cases depending on the value of d.

Case 1: Single-user performance regime ($d_0 \le d < MN$)

As discussed in Section II-C, the multiplexing gain region $\mathcal{R}(d)$ in this regime is a square and the scheduler assign decoupled rates to the queues. Hence, the optimal scheduler simply assigns a fixed transmission rate of $C_{av}(d)$ given in (11) to each user (i.e. we call this the *symmetric static scheduler*). Therefore, the asymptotic approximation (when D is sufficiently large) of the delay violation probability is given as follows.

Lemma 1: For $d_0 \le d \le MN$ and sufficiently large D, the asymptotic large-SNR approximation of $\Pr[W^1 > D]$ is such that

$$\lim_{\rho \to \infty} \frac{\log \Pr\left[W^1 > D\right]}{\log \rho} = -\sigma_s(d) D T r_{av}(d) \tag{17}$$

where $\sigma_s(d)$ is defined such that

$$\Lambda(\sigma_s(d)) = \sigma_s(d) C_{av}(d). \tag{18}$$

Equivalently, we can write (17) as

$$\Pr[W^1 > D] \doteq \rho^{-\sigma_s(d) D T r_{av}(d)}. \tag{19}$$

The proof of this Lemma is given in Appendix II.

Case 2: Antenna-pooling regime ($0 < d < d_0$)

In this case, the multiplexing gain region $\mathcal{R}(d)$ is polymatroid and the transmission rates of the users must be jointly allocated by a scheduler. Since the optimal policy (with respect to the delay violation probability objective) is unknown, we provide the following lower bound and upper bound to $\Pr[W^1 > D]$.

1) Upper Bound on $\Pr[W^1 > D]$*:* An upper bound on $\Pr[W^1 > D]$ is easily found since any feasible scheduling policy can provide an upper bound. In particular, to arrive at the upper bound $P^u(d)$, we consider the same symmetric static scheduler as described in Case 1: the symmetric static scheduler always assigns the symmetric rate of $C_{av}(d)$ to each user at all time. By Lemma 1, the asymptotic approximation of $P^u(d)$ for large D is given as

$$P^u(d) \doteq \rho^{-\sigma_s(d) D T r_{av}(d)}. \tag{20}$$

We note that this upper bound becomes tighter as d increases to d_0 since $\mathcal{R}(d)$ approaches a K-dimensional hypercube.

2) Lower Bound on $\Pr[W^1 > D]$*:* The lower bound $P^l(d)$ on $\Pr[W^1 > D]$ is obtained from Fact 2, the construction of a fictitious system, and Fact 3, as follow.

Fact 2: Consider two systems whose multiplexing gain regions are given by $\mathcal{R}_1$ and $\mathcal{R}_2$, respectively, where $\mathcal{R}_1 \subseteq \mathcal{R}_2$. The delay violation probability associated with the second system is no greater than that of the first system.

For a given d, consider a fictitious system whose multiplexing gain region is given by

$$\mathcal{R}_{\text{fic}}(d) := \left\{ (r_1, \ldots, r_K) : \sum_{i=1}^{K} r_i \le r^*_{KM,N}(d) \right\}. \tag{21}$$

Since $\mathcal{R}(d) \subseteq \mathcal{R}_{\text{fic}}(d)$, Fact 2 states that the delay violation probability for this system is a lower bound for $\Pr[W^1 > D]$. Stolyar and Ramanan [4] have shown that the largest-delay-first (LDF) policy achieves the minimum asymptotic delay violation probability for this fictitious system.

Fact 3: (Theorem 2.2 in [4]) Consider a single-server queuing model with K users (illustrated in Fig. 4(iii) for $K = 2$). For the sources considered in this paper, the largest-delay-first (LDF) policy achieves the minimum delay violation probability $\Pr[\max_{i=1,\ldots,K} W^i > D]$ when D is large. [6]

Hence, we compute the minimum asymptotic delay violation probability for this fictitious system to arrive at a lower bound, $P^l(d)$, for $\Pr[W^1 > D]$, as given in the following Lemma:

Lemma 2: For $0 < d < d_0$ and sufficiently large D, the asymptotic large-SNR approximation of $P^l(d)$ is given by

$$P^l(d) \doteq \rho^{-K\sigma_s(d) D T r_{av}(d)}. \tag{22}$$

[6] See more details of this fact in Fact 4. The result in [4] is much more general than this. It works with any weighted delays, i.e. $\Pr[\max_{i=1,\ldots,K} W^i/\alpha_i > D]$, where α_i is the weight for user i.

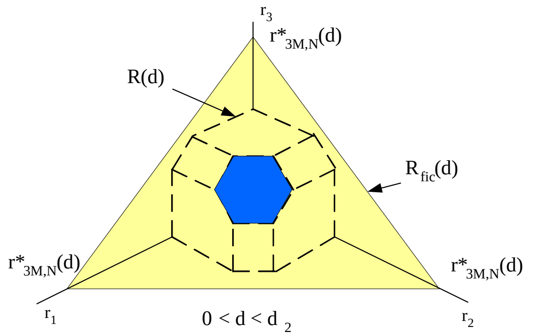


Fig. 3. The MIMO-MAC multiplexing gain region $\mathcal{R}(d)$ and the multiplexing gain region of the fictitious system $\mathcal{R}_{\text{fic}}(d)$ for $K = 3$ users.

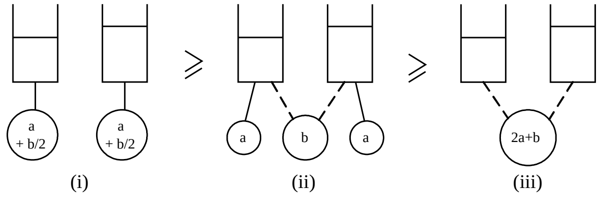


Fig. 4. Queuing models of the upper and lower bounds of $P[W^1 > D]$ for the case of $K = 2$ users and the antenna-pooling regime ($r^*_{2M,N}(d) \leq 2r^*_{M,N}(d)$). a and b are defined such that $a + b = r^*_{M,N}(d)$ and $2a + b = r^*_{2M,N}(d)$.

The proof is given in Appendix II. We also note that $P^l(d)$ becomes a tighter bound as $d \to 0$.

Remark 7: Comparing the exponents in (20) and (22), we see that the LDF scheduler improves the exponent of the delay violation probability by K times of that of the symmetric static scheduler. Talking in the language of channel diversity, the LDF scheduler improves the diversity gain by K folds by taking advantage of statistical-multiplexing of the sources. However, we want to emphasize that the lower bound in $P^l(d)$ derived from the fictitious system with the multiplexing gain region $\mathcal{R}_{\text{fic}}(d)$ becomes more loose as the number of users K grows. This is expected because the actual multiplexing gain region $\mathcal{R}(d)$ in (6) is a polymatroid (see an example of $K = 3$ users in Fig. 3) while that of the fictitious system is just the K-dimensional simplex given by the constraint $\sum_{i=1}^K r_i \leq r^*_{KM,N}(d)$. Thus, the lower-bound becomes more optimistic as the number of users increases.

Remark 8: For an example of $K = 2$ users, Fig. 4 summarizes the two bounds with the queuing models in mind. The upper bound $P^l(d)$ is the tail probability of system (i) which always serves each queue with multiplexing gain $r^*_{2M,N}(d)/2$. The lower bound $P^l(d)$ is the tail probability of system (iii) which assigns the single server of multiplexing gain $r^*_{2M,N}(d)$ based on LDF scheduling. System (ii) is the queuing model given by the multiplexing gain region $\mathcal{R}(d)$.

Now, using the above two cases and the bounds in (16), we arrive at an asymptotic characterization of $P_q(d)$ as follows

$$P_q(d) \doteq \Pr[W^1 > D], \tag{23}$$

and, in particular, when $d_0 \leq d < MN$,

$$P_q(d) \doteq \rho^{-\sigma_s(d)DTr_{av}(d)} \tag{24}$$

and, when $0 < d \leq d_0$,

$$\rho^{-K\sigma_s(d)DTr_{av}(d)} \dot{\leq} P_q(d) \dot{\leq} \rho^{-\sigma_s(d)DTr_{av}(d)}. \tag{25}$$

In summary, so far, we have seen that $P_q(d)$ and $P_e(d)$ exponentially decay with ρ. The rate of decay of $P_e(d)$ is known. When $d > d_0$, the rate of decay of $P_q(d)$ is known via (24). The rate of decay of $P_q(d)$ when $d < d_0$ is, however, unknown but is bounded as in (25).

Next, with $P_q(d)$ and $P_e(d)$ at hand, we proceed with the minimization of the total bit loss probability.

C. Minimizing Asymptotic Total Loss Probability

From the asymptotic expressions of $P_e(d)$ given in (15) and $P_q(d)$ in (24) and (25), the asymptotic characterization of the total loss probability $P_{\text{tot}}(d) \doteq P_q(d) + P_e(d)$ is immediate: For $d_0 \leq d \leq MN$,

$$P_{\text{tot}}(d) \doteq \rho^{-\sigma_s(d)DTr_{av}(d)} + \rho^{-d}. \tag{26}$$

For $0 < d < d_0$,

$$\begin{aligned} \rho^{-K\sigma_s(d)DTr_{av}(d)} + \rho^{-d} \dot{\leq}\ & P_{\text{tot}}(d) \\ & \dot{\leq} \rho^{-\sigma_s(d)DTr_{av}(d)} + \rho^{-d}. \end{aligned} \tag{27}$$

Since the term $\sigma_s(d)DTr_{av}(d)$ is decreasing in d while the term d is increasing on d, the minimum of $P_{\text{tot}}(d)$ in (26) or its bounds in (27) happen when the value of d makes the exponents of the two terms are within $o(1)$ of each other (note that if the exponents were not in the same order, one term would dominate in the sum as $\rho \to \infty$). We now introduce an algorithm which guarantees such choices of d:

Algorithm 1:

1) Solve for d which is a solution of

$$\sigma_s(d)DTr_{av}(d) = d. \tag{28}$$

If $d \geq d_0$, then $d^* = d^*_u = d^*_l = d$ and stop. Otherwise, set $d^*_l = d$. Go to Step 2.

2) Solve for d which is a solution of

$$K\sigma_s(d)DTr_{av}(d) = d \tag{29}$$

and set $d^*_u = \min(d, d_0)$.

Theorem 1: Algorithm 1 results in a closed interval $[d^*_l, d^*_u]$ in which the optimal common diversity gain d^* lies.

The proof of this theorem is given in Appendix II.

Notice that the optimal diversity d^* and its bounds depend on the statistical characteristics of the symmetric sources (Λ, μ, λ), the parameters of the MIMO-MAC channel (e.g. T, M, N), and the delay bound D.

D. Statistical-Multiplexing and Optimal Diversity Gain

From the above analysis, we obtain the following critical observation. Given a delay constraint, the statistical property of the source has a significant impact on the level of diversity a well-designed system can enjoy. In other words, the optimal scheduler which statistically multiplexes the MIMO resources allows the combined bursty sources to perceive as smaller aggregate traffic and hence a higher degree of diversity. Rigorously, this performance improvement can be attributed to a *statistical-multiplexing gain* as follows:

Definition 4: An optimal dynamic scheduler with the total bit loss probability P^*_{tot} provides *statistical-multiplexing gain* of s over the static rate scheduler with P^f_{tot}, where

$$s := -\lim_{\rho\to\infty} \frac{\log P^*_{\text{tot}} - \log P^f_{\text{tot}}}{\log \rho}. \tag{30}$$

From this definition and the fact that $P^f_{\text{tot}} \doteq \rho^{-d^*_l}$, the following proposition is immediate.

Lemma 3: Consider the system model in Section II. The optimal statistical-multiplexing gain s^* is given by

$$s^* = d^* - d^*_l. \tag{31}$$

Furthermore, it is bounded above by $d^*_u - d^*_l$.

Remark 9: The two concepts of "statistical-multiplexing gain" and "multi-user diversity gain" are related conceptually. The former takes advantage of the troughs (due to burstiness) of the traffic of different users while the latter takes advantage of the peaks (due to fadings) of the channels of different users. But their impacts on the design are sufficiently different, as multi-user diversity gain requires channel CSI at the transmitters while statistical-multiplexing requires QSI.

E. Resource Pooling and Statistical-Multiplexing

Here we discuss the effect of the arrival rate λ and the average delay bound D to the performance region of the MIMO-MAC. The relationship between (λ, D) and the system performance is summarized in Fig. 5. The system performance is divided into three main regions: the single-user performance region, the antenna-pooling with *significant* statistical-multiplexing region, and the antenna-pooling with *insignificant* statistical-multiplexing region. In the single-user performance region, the achieved optimal diversity gain d^* is equivalent to the case when only one user is in the system, i.e. the case $d^* \geq d_0$. Specifically, this case happens when λ is sufficiently small and D is sufficiently large. We denote this region as $\mathcal{A}_1$.

$$\mathcal{A}_1 := \left\{ (\lambda, D) : \lambda \leq r_0, D \geq \frac{d_0}{\sigma_s(d_0)Tr_0} \right\}$$

where $r_0 = r^*_{M,N}(d_0)$.[7] Since in this region the transmission rate of each user is independent, there is no resource sharing and hence no statistical-multiplexing gain.

On the other hand, the significance of statistical-multiplexing gain outside $\mathcal{A}_1$ is impacted by the rate of arrivals λ as well as the average delay bound D. In particular, for (λ, D) in the neighborhood of $\mathcal{A}_1$, the statistical-multiplexing gain is not significant as the queues still behave in a roughly independent manner. Similarly, as λ increases to an overload situation or the delay bound D becomes very tight, the benefits of juggling resources diminishes. In contrast, for medium values of λ and D and under the optimal dynamic scheduler, each queue perceives the whole (pooled) resource to itself, compared to $1/K$ of the resource as in case of a symmetric static scheduler.

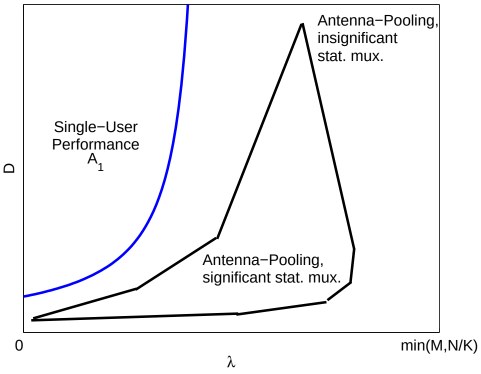


Fig. 5. The relation between (λ, D) to the system performance.

To illustrate the approach shown in this paper and the corresponding calculation, we look at a simple example of a compound Poisson source with $K = 2$ users in the next section.

IV. Example: Compound Poisson Sources and $K = 2$

In this section, we illustrate the proposed approach via an example. We consider two independent but identical source processes. For each source i, arrivals are independent across timeslots. The number of bits that arrive in a timeslot t, A^i_t, is an aggregation of a random number of packets whose sizes are also random, i.e. $A^i_t = \sum_{n=1}^{N} Y_n$. Furthermore, we assume that the number of packets at each slot, N, is an independent Poisson random variable with rate ν packets per timeslot, while the length of the packets, Y_i, $i = 1, 2, \ldots$, are i.i.d. random variables with exponential distribution of mean $1/\mu$. The average bit arrival rate $\tilde{\lambda}$ for each source is equal to ν/μ and scales with $\log \rho$ as in (8), i.e.

$$\tilde{\lambda} = \nu/\mu = \lambda T \log \rho. \tag{32}$$

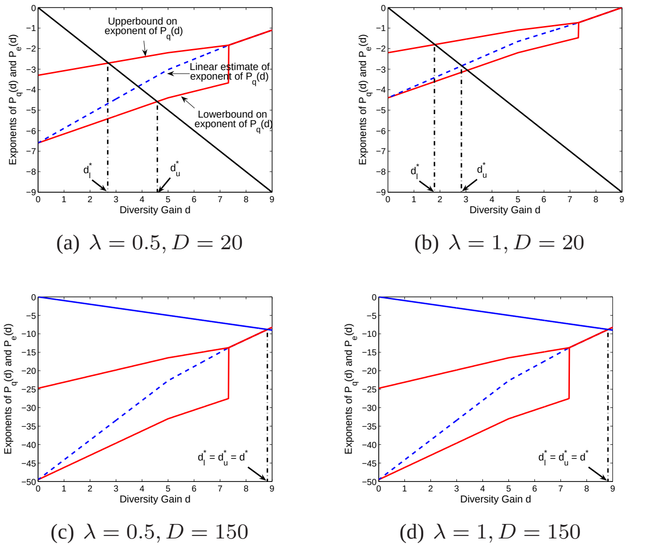
Proposition 1: For the compound Poisson source with exponential packet length, the $\sigma_s(d)$ defined in Lemmas 1 is given as

$$\sigma_s(d) = \mu(1 - \frac{\lambda}{r_{av}(d)}). \tag{33}$$

This proposition is proved in Appendix II.

Note that the ratio of the per-queue average bit arrival rate over the average service rate, $\frac{\lambda}{r_{av}(d)}$, can be called the traffic load per queue. It is important to note that the delay violation exponent in (33) is a decreasing function of the average packet size $1/\mu$, for a fixed packet arrival rate ν. A larger packet size in effectively creates more burstiness in the arrivals, hence a higher delay violation probability.

[7] Note that $\frac{d_0}{\sigma_s(d_0)Tr_0}$ is an increasing function on λ since $\sigma_s(d_0)$ which is the delay violation exponent is itself decreasing on the arrival rate λ.



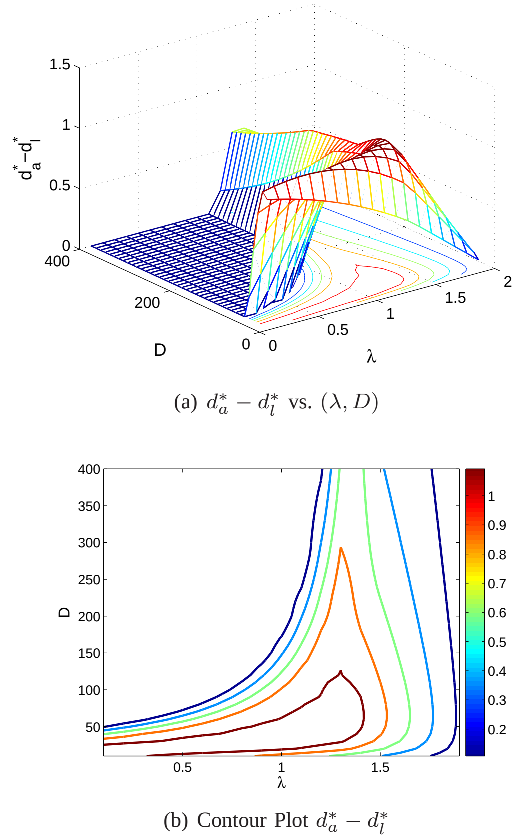
(a) $\lambda = 0.5, D = 20$

(b) $\lambda = 1, D = 20$

(c) $\lambda = 0.5, D = 150$

(d) $\lambda = 1, D = 150$

Fig. 6. Plots of the exponents of $P_q(d)$ and $P_e(d)$ for two different average arrival rates ($\lambda = 0.5$ and 1) and for two different delay bounds $D = 20$ and 150. For $d < d_0 = 7.3$, the upper and lower bounds of the exponent of $P_q(d)$ are shown. In addition, we draw a simple linear estimate (dotted line) of the exponent of $P_q(d)$ between the two bounds.



(a) $d^*_a - d^*_l$ vs. (λ, D)

(b) Contour Plot $d^*_a - d^*_l$

Fig. 7. 3D and contour plots characterizing the statistical-multiplexing gain, approximated by $d^*_a - d^*_l$, v.s. delay bound D and arrival rate λ.

With Proposition 1 in hand, we are now ready to use Algorithm 1 to obtain d^* (or its bounds). Fig. 6 shows the optimal d^* and its bounds when $M = N = 4$, the average packet size $1/\mu$ is 100 nats, and the symbol rate such that there are $T = 2M + N - 1 = 11$ symbols per timeslot. In these figures, we plot the exponents of $P_q(d)$ or its bounds, i.e. $-\sigma_s(d)DTr_{av}(d)$ and $-2\sigma_s(d)DTr_{av}(d)$, and the exponent of $P_e(d)$, i.e. $-d$. To better illustrate the procedure followed by Algorithm 1, we plot the exponents of $P_q(d)$ and $P_e(d)$ separately. Note that when $d \leq d_0$ ($d_0 = 7.3$ in this example), we only have lower and upper bounds for the exponents of $P_q(d)$. In this case, in addition to the bounds, we plot a linear approximation (dotted line) to emphasize the tightness of the lower bound around d_0 and the upper bound around 0. The optimal diversity gain d^* or its bounds (d^*_l and d^*_u) are shown in each plot as the crossing of the exponents.

As we discussed in Section III-E, the statistical-multiplexing gain s^* defined in (31) depends on the optimal choice of d^* which itself is a function of the arrival rate λ and the average delay bound D. Depending on λ and D, we may or may not have statistical-multiplexing gain. For example, Fig. 6(c) shows that, in the case of sufficiently low arrival rates and a large delay bound, the optimal P^*_{tot} happens when the users operate in the single-user performance region. Hence, in this case, there is no statistical-multiplexing gain to be achieved by dynamic scheduler. Here, the dominant form of loss occurs on the channels. On the other hand, Fig. 6(b) corresponds to the case of large arrival rate and small delay bound, where the loss probability due to delay violation dominates that of the channel. In this case, the optimal diversity d^* necessitates resource sharing in form of antenna pooling. As a result, the impact of n optimal dynamic scheduler, in a form of statistical-multiplexing gain, becomes more significant.

Fig. 7 illustrates the performance region discussed in Section III-E. In particular, the figure gives a characterization of the region shown in Fig. 5 for the compound Poisson case. We approximate the statistical-multiplexing gain $s^* = d^* - d^*_l$ achieved with an optimal dynamic scheduler with that of a simple linear approximation $d^*_a - d^*_l$, where d^*_a is the optimal diversity gain derived from the dotted line in Fig. 6.

V. Summary and Future Work

In this paper, we considered a system of bursty and delay-sensitive symmetric sources concatenated with a symmetric MIMO-MAC channel. We assumed no CSI information available to the transmitters and a block fading model with a block coding whose block lengths are matched to the coherence time of the channel. Furthermore, we assumed a fixed and equal high transmission power at each transmitter, i.e. high SNR regime. We addressed the optimal choice of the spatial diversity gain d^* such that it minimizes an end-to-end loss performance where loss can occur due to delay violation as well as channel decoding error. We showed how an optimal choice of diversity gain d^* depends on a queue-based scheduler module whose job is to statistically multiplex the resources of the MIMO-MAC. In doing so, we integrated the notion of statistical-multiplexing gain with those of spatial diversity, multiplexing, and multi-access gains provided by the MIMO-MAC.

To strengthen the result presented in this paper, we will need to analyze the performance of the optimal dynamic scheduler, rather than working with bounds. As emphasized in the introduction to this paper, we view the result of this paper as a first step to fully integrate the notions of scheduling and statistical-multiplexing with other aspects of MIMO technology. As such, strengthening the current result is not our most critical concern. Some future works that can extend the utilization of the system resources in other aspects are as follows:

- Time diversity: In the present work, we assume coding whose block length is matched to the coherence time of the channel. This is a rather limiting assumption, given the delay bound in the order of multiple coherence times we considered in this paper. We hope to extend our optimization of spatial diversity to the time-diversity, in a form of coding over multiple coherence times (see [15]) or hybrid ARQ (see [13], [6], and [7]).
- Controlling SNR: In the current formulation, the average SNR is fixed across all users and time. In the absence of CSI at the transmitters and of delay constraints ($D = \infty$), this fixed SNR assumption incurs no loss in optimality. But for small and medium D, it is natural to expect that an improvement in performance is possible when the transmit powers for users are functions of the queue states.
- Dynamic control of PHY layer operating point: In this paper, we assumed that the sole responsibility of the dynamic scheduler is to control the transmission rates or multiplexing gains while the diversity gain and the sum of the rates of all users are kept constant. It is clear that allowing for dynamic control of the diversity gain, as proposed by [7], will only improve the statistical-multiplexing gain and the performance of the system. This can naturally be extended to time-diversity, and is roughly related to a dynamic control of the tradeoff among various forms of diversity gain. In this context, our work can be viewed as providing a lower bound on the performance of an optimally designed system.
- Cooperative multiple-access channel: Cooperation among users can substantially improve the reliability of communication [16]. It is interesting to extend our study to MAC with cooperation (see [17]). However, to do that we require a similar tradeoff result as in [2]. Furthermore, there is another parameter which is the size of the cooperative cluster that needs to be taken into account (see [18] for a simple case of single source with multiple relays).

References

[1] L. Zheng and D. Tse, “Diversity-multiplexing: a fundamental tradeoff in multiple-antenna channels,” *IEEE Trans. Info. Theory*, v. 49, no. 5, May 2003.

[2] D. Tse, P. Viswanath, and L. Zheng, “Diversity-multiplexing tradeoff in multiple-access channels,” *IEEE Trans. Info. Theory*, v. 50, no. 9, Sept 2004.

[3] D. Bertsimas, I. C. Paschalidis, and J. N. Tsitsiklis, “Asymptotic buffer overflow probabilities in multiclass multiplexers: an optimal control approach,” *IEEE Trans. Automatic Control*, v. 43, no. 3, March 1998.

[4] A. L. Stolyar and K. Ramanan, “Largest weighted delay first scheduling: Large deviations and optimality.” *Ann. Applied Probability*, v. 11, pp. 1-48, 2001.

[5] T. Holliday and A. Goldsmith, “Joint source and channel coding for MIMO systems,” *Proc. Allerton Conf.*, 2004.

[6] T. Holliday and A. Goldsmith, “Optimizing End-to-End Distortion in MIMO Systems”, *IEEE ISIT'05*, 2005.

[7] T. Holliday, A. Goldsmith, and H. V. Poor, “The Impact of Delay on the Diversity Multiplexing ARQ Tradeoff”, *ICC*, 2006.

[8] F. Kelly, “Notes on effective bandwidths,” *Stochastic Networks: Theory and Applications*, vol. 4, 1996.

[9] A. Dembo and O. Zeitouni, *Large Deviation Techniques and Applications*, Boston, MA: Jones and Bartlett, 1992.

[10] A. Dembo and T. Zajic, “Large deviations: from empirical mean and measure to partial sum process,” *Stochastic Processes and their Apps.*, v. 57, 1995.

[11] A. Ganesh, N. O'Connel, and D. Wischik, *Big Queues*, Springer-Verlag, Berlin, 2004.

[12] J. Walrand and P. Varaiya, *High-Performance Communication Networks*, 2nd ed., Morgan Kaufmann Publishers, 2000.

[13] H. El Gamal, G. Caire, and M. O. Damen, “The MIMO ARQ channel: diversity-multiplexing-delay tradeoff,” *IEEE Trans. Info. Theory*, v. 52, Aug 2006.

[14] S. Kittipiyakul and T. Javidi, “Optimal operating point in MIMO channel for delay-sensitive and bursty traffic,” *IEEE ISIT'06*, July 2006.

[15] P. Elia, S. Kittipiyakul, and T. Javidi, “On the Responsiveness-Diversity-Multiplexing tradeoff,” *WiOpt*, 2007.

[16] A. Sendonaris, E. Erkip and B. Aazhang, “Increasing uplink capacity via user cooperation diversity,” *IEEE ISIT'98,* August 1998.

[17] K. Azarian and H. El Gamal, “Cooperation in outage-limited multiple-access channels,” *IEEE Inter. Zurich Seminar on Commun. (IZS06)*, 2006.

[18] P. Elia, S. Kittipiyakul, and T. Javidi, “Cooperative diversity in wireless networks with stochastic and bursty traffic,” *IEEE ISIT07*, June 2007.

Appendix I
Additional Assumption on Source Model

Here we give the additional assumption B on the source we consider in this paper. Assumption B is required in the proof of Lemma 2.

Assumption B: “Sample path LDP” (see [3], [9] and [10])

For an arrival sequence $\{S_1, S_2, \ldots\}$, for all $m \in \mathbb{N}$, for every $\epsilon_1, \epsilon_2 > 0$, and for every scalar $a_0, \ldots, a_{m-1}$, there exists $M > 0$ such that for all $n \geq M$ and all $k_0, \ldots, k_m$ with $1 = k_0 \leq k_1 \leq \cdots \leq k_m = n$,

$$\begin{aligned}
&\exp\left\{-n\epsilon_2 - \sum_{i=1}^{m-1}(k_{i+1}-k_i)\Lambda^*(a_i)\right\} \\
&\leq \Pr\left[\left|S_{k_{i+1}} - S_{k_i} - (k_{i+1}-k_i)a_i\right| \leq n\epsilon_1, i = 1, \ldots, m-1\right] \\
&\leq \exp\left\{n\epsilon_2 - \sum_{i=1}^{m-1}(k_{i+1}-k_i)\Lambda^*(a_i)\right\}
\end{aligned}$$

Appendix II
Proofs of Lemmas and Proposition

Proof of Lemma 1:

Proof: Since the symmetric static scheduler always assigns the service rate $C_{av}(d)$ to each queue, we have that

$$\Pr[W^1 > D] = \Pr[L^1 > DC_{av}(d)]. \tag{34}$$

To be more specific, the statement holds because any bits delayed more than D timeslots see at least $DC_{av}(d)$ bits before them, and any bits delayed less than D timeslots must see less than $DC_{av}(d)$ bits before them. This is valid because of the first-come-first-serve discipline assumption. Hence, the two events $\left\{W^1 > D\right\}$ and $\left\{L^1 > DC_{av}(d)\right\}$ are equivalent and have the same probability.

Now, since $\Pr[L^1 > DC_{av}(d)]$ is equal to the tail probability for a buffer which is served at fixed capacity of $C_{av}(d)$

and whose arrival process is described by $\Lambda(\cdot)$ and satisfying LDP, one can calculate the tail probability for a single queue system with a fixed service rate c as (see [8], [11] and [12])

$$\lim_{B\to\infty}\frac{1}{B}\log\Pr[L^1 > B] = -\theta^*$$

where θ^* is the largest positive root of equation $\frac{\Lambda(\theta)}{\theta} = c$. By replacing B with $DC_{av}(d)$ and c with $C_{av}(d)$ and using (11), we have

$$\lim_{\rho\to\infty}\frac{\log\Pr[L^1 > DC_{av}(d)]}{\log\rho} = -\sigma_s(d)DTr_{av}(d)$$

where $\sigma_s(d)$ is given as the solution to

$$\Lambda(\sigma_s(d)) = \sigma_s(d)C_{av}(d).$$

■

Proof of Lemma 2:

Before showing the proof of Lemma 2, we recall the following result on the asymptotic tail probability of the maximal weighted delay under the longest-weighted-delay-first (LWDF) scheduling discipline from [4] and simplify the result to our specific assumptions of symmetric users with LDF scheduling discipline.

Fact 4: (Theorem 2.2 in [4]) Consider a single server of fixed service rate 1 and K mutually independent source processes with stationary increments. The total number of information bits generated by source i ($i = 1,\dots,K$) is given by a sequence $\left\{\hat{S}_t^i, t = 1,2,\dots\right\}$ where $\hat{S}_t^i$ is the cumulative total number of work arrived until time t from source i. We assume $\left\{\hat{S}_t^i, t = 1,2,\dots\right\}$ satisfies LDP and sample path LDP (Assumptions A and B) with the convex decay function $\hat{\Lambda}_i^*$ and the convex log moment generating function $\hat{\Lambda}_i$. Assume $KE[\hat{S}_1^1] < 1$ for stability. Let α_i be the weight for user i (assuming $0 < \alpha_1 \le \alpha_2 \le \cdots \le \alpha_K$). Consider the longest-weighted-delay-first (LWDF) scheduling discipline, which always assign the server to the longest waiting (i.e. head-of-the-line)) customer of the source i which has the maximal weighted delay. Then, the LWDF scheduling discipline maximizes the exponential decay rate of the stationary distribution of the maximal delay, among all causal and work-conserving scheduling disciplines. Furthermore, the probability is given as

$$\limsup_{n\to\infty}\frac{1}{n}\log\Pr\left[\frac{1}{n}\max_{i\in 1,\dots,K}\hat{w}^i > 1\right] \le -J_* \tag{35}$$

where and $\hat{w}^i$ is the stationary delay for user i, $i = 1\dots,K$, and J_* is given as:

$$J_* = \min_{j;x_1,\dots,x_j}\frac{1}{\gamma}\sum_{i=1}^{j}(1-\alpha_i\gamma)\hat{\Lambda}_i^*(x_i) \tag{36}$$

subject to

$$j \in \{1,\dots,K\}, x_i > 0, \sum_{i=1}^{j} x_i > 1 \tag{37}$$

and

$$\frac{1}{\alpha_{j+1}} < \gamma = \frac{\sum_{i=1}^{j} x_i - 1}{\sum_{i=1}^{j}\alpha_i x_i} \le \frac{1}{\alpha_j} \tag{38}$$

with $\alpha_{K+1} \equiv \infty$.

Since in this paper we consider symmetric users and LDF scheduling which is the LWDF discipline with equal weights, the following corollary gives a specific expression of J_* which will be used to show Lemma 2.

Corollary 1: Under the assumptions of symmetric users with equal weights, i.e. $\hat{\Lambda}_i^* = \hat{\Lambda}^*$, $\hat{\Lambda}_i = \hat{\Lambda}$, and $\alpha_i = 1$ for all $i = 1,\dots,K$, J_* in Fact 4 is reduced to

$$J_* = \sup\left\{\theta : \hat{\Lambda}(\theta) \le \theta/K\right\}. \tag{39}$$

Proof: Under the assumption of equal weights (i.e. $\alpha_i = 1$ for all $i = 1,\dots,K$), there are feasible values of γ in (38) only when $j = K$. Hence, the minimization in (36) is reduced to

$$J_* = \min_{x_1,\dots,x_K}\frac{1-\gamma}{\gamma}\sum_{i=1}^{K}\hat{\Lambda}^*(x_i) \tag{40}$$

subject to

$$\sum_{i=1}^{K} x_i > 1, x_i > 0,\ \ i = 1,\dots,K \tag{41}$$

and

$$\frac{1}{\alpha_{K+1}} = 0 < \gamma = \frac{\sum_{i=1}^{K} x_i - 1}{\sum_{i=1}^{K} x_i} \le 1. \tag{42}$$

However, we notice that condition (42) is satisfied with any choices of $\{x_i\}$ satisfying condition (41). Hence, plugging the expression of γ into (40), we get

$$J_* = \min_{x_1,\dots,x_K}\frac{1}{\sum_{i=1}^{K} x_i - 1}\sum_{i=1}^{K}\hat{\Lambda}^*(x_i) \tag{43}$$

subject to (41).

We can simplify J_* further by using the convexity property of $\hat{\Lambda}^*$, i.e.

$$\frac{1}{K}\sum_{i=1}^{K}\hat{\Lambda}^*(x_i) \ge \hat{\Lambda}^*(\frac{\sum_{i=1}^{K} x_i}{K}) =: \hat{\Lambda}^*(\frac{Ka+1}{K}),$$

where we let a be such that $Ka = \sum_{i=1}^{K} x_i - 1 > 0$ by the condition in (41). The equality holds when $x_i = a + 1/K$ for all $i = 1,\dots,K$. Hence, we can rewrite (43) and its conditions concisely as

$$J_* = \min_{a>0}\frac{\hat{\Lambda}^*\left(a + \frac{1}{K}\right)}{a}. \tag{44}$$

To finish the proof, we expand $\hat{\Lambda}^*$ using its definition, as follows:

$$\begin{aligned}
J_* &= \min_{a>0}\frac{1}{a}\hat{\Lambda}^*\left(a + 1/K\right) \\
&= \min_{a>0}\ \frac{1}{a}\ \sup_{\theta\in\mathbb{R}}\theta(a + 1/K) - \hat{\Lambda}(\theta) \\
&= \sup_{\theta\in\mathbb{R}}\min_{a>0}\ \theta + \frac{\theta/K - \hat{\Lambda}(\theta)}{a} \\
&= \sup_{\theta\in\mathbb{R}}\begin{cases}-\infty, & \text{if } \theta/K < \hat{\Lambda}(\theta),\\ \theta, & \text{if } \theta/K \ge \hat{\Lambda}(\theta)\end{cases} \\
&= \sup\left\{\theta : \hat{\Lambda}(\theta) \le \theta/K\right\},
\end{aligned}$$

where the third equality holds because the function $\theta + \frac{\theta/K-\hat{\Lambda}(\theta)}{a}$ is convex on a and concave on θ (since $\hat{\Lambda}$ is convex). ■

Proof: (Lemma 2) Consider a scaled version of the system in Fact 4 where the service rate is scaled to C (which is equal to KC_{av}) and the arrivals are also scaled up by C. We can think of this scaling as a change of measurement units. We denote W_s^i as the stationary delay of arrivals of user i, $i = 1, \ldots, K$, for this scaled single-server system with LDF scheduling. Since scaling of the service and arrivals do not change the distribution of the delays, we have from Fact 4 that

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left[\frac{1}{n} \max_{i\in 1,\ldots,K} W_s^i > 1\right] \leq -J_*. \tag{45}$$

Noticing that the log moment generating function Λ of the scaled system is given as

$$\Lambda(\theta) = \lim_{t\to\infty} \frac{1}{t} \log E[e^{\theta C \hat{S}_t^1}] = \hat{\Lambda}(\theta C), \tag{46}$$

we have, by using Corollary 1,

$$\begin{aligned} J_* &= \sup\left\{\theta : \hat{\Lambda}(\theta) \leq \theta/K\right\} \\ &= \sup\left\{\theta : \Lambda(\theta/C) \leq \theta/K\right\} \\ &= C \sup\left\{\tilde{\theta} : \Lambda(\tilde{\theta}) \leq \tilde{\theta}C/K = \tilde{\theta}C_{av}\right\} \\ &= C\sigma_s \end{aligned} \tag{47}$$

where $\sigma_s > 0$ is defined as the unique solution to

$$\Lambda(\sigma_s) = \sigma_s C_{av}.$$

The second equality in (47) follows by using (46); the third equaility follows by letting $\tilde{\theta} = \theta/C$; and the last equality by using that fact that Λ is strictly convex and $\Lambda'(0) = \lambda T \log\rho < C_{av}$ (the stability condition in (9) and the fact that $\hat{\Lambda}'(0)$ is the average arrival rate per source [8]) and hence the supremum is attained with $\tilde{\theta} = \sigma_s$.

Replacing n with D and $J_* = C\sigma_s = KC_{\text{av}}\sigma_s = K\sigma_s T r_{\text{av}} \log\rho$ in (45), we have

$$P\left[\max_{i\in 1,\ldots,K} W_s^i > D\right] \doteq \rho^{-K\sigma_s D T r_{\text{av}}},$$

for large value of D.

From symmetry, on the other hand, we have

$$P[W_s^1 > D] \leq P\left[\max_{i\in 1,\ldots,K} W_s^i > D\right] \leq KP[W_s^1 > D].$$

This provides the assertion of the lemma:

$$P^l(d) := P[W_s^1 > D] \doteq \rho^{-K\sigma_s D T r_{\text{av}}}.$$

■

Proof of Theorem 1:

Proof: We first show the existence of a solution d in (28) of Algorithm 1. The LHS term is decreasing on d and equal to 0 for $d \geq \bar{d}$, for some $\bar{d}$ such that the arrival rate $\lambda T \log\rho$ is equal to the service rate $C_{av}(\bar{d})$ (in which case, $\sigma_s(\bar{d}) = 0$). On another hand, the RHS term is increasing on d and is equal to 0 when $d = 0$. Hence, (28) must hold for some $d \in (0, \bar{d})$. Next, if d solving (28) is less than d_0, this d is the lower bound d_l^* (i.e. asymptotically maximizing the RHS term in (27)) and the upper bound d_u^* is obtained from maximizing the LHS in (27). The existence of d solving (29) can be shown similarly. ■

Proof of Proposition I:

Proof: The limiting log moment generating function $\Lambda(\cdot)$ for compound Poisson source with exponential packet length is derived in [14], which is

$$\begin{aligned} \Lambda(\theta) &= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\exp(\theta S_t^1)\right] \\ &= \lim_{n\to\infty} \frac{1}{n} \log \mathbb{E}\left[\exp(\theta \sum_{t=1}^{n} A_t^1)\right] \\ &= \log \mathbb{E}\left[\exp \theta A_1^1\right] \\ &= \begin{cases} \frac{\nu\theta}{\mu-\theta} & \text{if } \theta < \mu, \\ \infty & \text{if } \theta \geq \mu. \end{cases} \end{aligned} \tag{48}$$

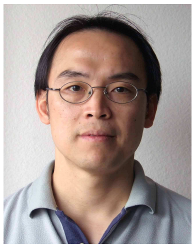
From Lemma 1, $\sigma_s(d)$ is the solution to $\Lambda(\sigma_s(d)) = \sigma_s(d)C_{av}(d)$. From (48), this reduces to finding $\sigma_s(d)$ such that

$$\begin{aligned} \frac{\nu\sigma_s(d)}{\mu - \sigma_s(d)} &= \sigma_s(d)C_{av}(d) \\ \Leftrightarrow \quad \frac{\mu\lambda T \log\rho}{\mu - \sigma_s(d)} &= r_{av}(d) T \log\rho \end{aligned}$$

where we have replaced ν and $C_{\text{av}}(d)$ from (32) and (11). Hence,

$$\sigma_s(d) = \mu(1 - \lambda/r_{av}(d)).$$

■



Somsak Kittipiyakul received his S.B. and M.Eng. degrees in electrical engineering and computer science from Massachusetts Institute of Technology in 1996. Currently, he is pursuing the Ph.D. degree at the University of California, San Diego.

His research interests include wireless communication and networking, stochastic resource allocation and scheduling, and cross-layer performance analysis.

Tara Javidi (S96-M02) studied electrical engineering at Sharif University of Technology, Tehran, Iran from 1992 to 1996. She received the MS degrees in electrical engineering (systems), and in applied mathematics (stochastics) from the University of Michigan, Ann Arbor, in 1998 and 1999, respectively. She received her Ph.D. in electrical engineering and computer science from the University of Michigan, Ann Arbor, in 2002.

From 2002 to 2004, she was an assistant professor at the Electrical Engineering Department, University of Washington, Seattle. She joined University of California, San Diego, in 2005, where she is currently an assistant professor of electrical and computer engineering. She was a Barbour Scholar during 1999-2000 academic year and received an NSF CAREER Award in 2004.

Her research interests are in communication networks, stochastic resource allocation, stochastic control theory, and wireless communications.