Delay Performance of the Multiuser MISO Downlink

Sebastian Schiessl*, James Gross* and Giuseppe Caire‡
*School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden
‡ Institute for Telecommunication Systems, Technical University Berlin, Berlin, Germany
Emails: {schiessl,jamesgr}@kth.se, and caire@tu-berlin.de

***Abstract*—We analyze a MISO downlink channel where a multi-antenna transmitter communicates with a large number of single-antenna receivers. Using linear beamforming or nonlinear precoding techniques, the transmitter can serve multiple users simultaneously during each transmission slot. However, increasing the number of users, i.e., the multiplexing gain, reduces the beamforming gain, which means that the individual data rates decrease. We use stochastic network calculus to analyze the queueing delay that occurs due to the time-varying data rates. Our results show that the optimal number of users, i.e., the optimal trade-off between multiplexing gain and beamforming gain, depends on incoming data traffic and its delay requirements.**

***Index Terms*—Multiple-input multiple-output (MIMO), multiuser diversity, zero-forcing beamforming (ZFBF), dirty paper coding (DPC), stochastic network calculus**

I. Introduction

The capacity of wireless communication systems can be significantly increased when both the transmitter and the receiver are equipped with multiple antennas. Interestingly, similar capacity gains can also be achieved when a multi-antenna transmitter communicates simultaneously with multiple receivers that have only a single antenna each. In order to achieve the capacity of such a multi-user multiple-input single-output (MU-MISO) downlink channel, the transmitter must employ nonlinear precoding techniques like dirty paper coding (DPC) [1]. However, even linear precoding techniques are sufficient to achieve a large fraction of the capacity. A commonly used linear precoding scheme is zero-forcing beamforming (ZFBF), which projects the signal intended for a user into a subspace that is orthogonal to the channels of the other users. A transmitter with M antennas can use ZFBF to send $K \leq M$ different data streams to K users at a time. Increasing K increases the multiplexing gain, but it decreases the beamforming gain due to a reduced dimensionality of the subspaces that are orthogonal to the other users. Thus, when K becomes equal to M, the linear growth in capacity is lost [2]. Previous works, e.g. [3], have studied the optimal number of scheduled users K, i.e., the optimal trade-off between the multiplexing gain and beamforming gain such that the ergodic capacity of the system is maximized.

However such an analysis of the ergodic sum capacity does not accurately reflect the performance when the system is subject to constraints on maximum delay, such as in live video or audio transmissions. This is due to two reasons. First, when the total number of users U is larger than the number of antennas M, then the transmitter can only schedule a subset of $K < U$ users in each transmission slot. Thus, the delay also depends on how often and how regularly each user is scheduled. Second, large variations in the instantaneous data rates mean that the transmitter cannot always send all the available data. When the channel conditions are bad, the data must be stored in a buffer for transmission in subsequent time slots, causing a buffering or queueing delay.

A. Related Work

Several works have studied the use of linear precoding in the multiuser MISO downlink, as nonlinear dirty paper coding techniques are difficult to implement in practice. Specifically, Yoo and Goldsmith [4] showed that when the total number of users U in the system greatly exceeds the number of antennas, then ZFBF achieves asymptotically the same performance as DPC. However, their scheme assumes that the transmitter has channel state information (CSI) of all users. The cost of collecting this CSI would be overwhelming when the number of users U is large. Sharif and Hassibi [5] reduce the overhead from collecting CSI by randomly creating a set of beamforming vectors and then transmitting only to the users which report the highest signal-to-interference-and-noise ratio (SINR) along those random beams. Although the scheduling probabilities of all users are equal, the scheduling of the users is random, which can result in unacceptably long delays for some users. Zhang et al. [6] studied the optimal number of scheduled users when the transmitter has only knowledge of the channel of the scheduled users, and also considered imperfect CSI. Ravindran and Jindal [7] also studied imperfect CSI due to quantized feedback. They found that collecting many bits of feedback (accurate CSI) from very few users is more beneficial than collecting few bits of feedback from many users, which supports the assumption in [6] that CSI should be obtained only for the scheduled users.

However, all of these works studied only the ergodic capacity of the MU-MISO downlink, and did not address the system performance under delay constraints. When the transmission rate varies due to channel fading, the transmitter cannot always transmit all data and must keep data in a buffer, causing a random queueing delay. This queueing delay can be analyzed through the frameworks of stochastic network calculus [8], [9] or effective capacity [10]. Several authors [11]–[13] studied the effective capacity of MIMO systems, considering only the single user case. Li et al. [14] investigated the effective capacity of multiuser MIMO systems. However, the authors make many assumptions that we do not consider practical, e.g., that the channel coefficients are non-fading and that there is always a backlog of data in each user's queue.

B. Contributions

In this paper, we analyze the queueing performance of the MU-MISO downlink using stochastic network calculus (SNC). We consider both linear ZFBF precoding and nonlinear dirty paper coding. We demonstrate that SNC can still be applied when the users are scheduled regularly in a round robin fashion. Based on previous results, we present closed-form expressions to analytically determine the distribution of the queueing delay. Our numerical results show that the optimal number of scheduled users changes when considering the delay performance instead of the ergodic capacity.

This paper is structured as follows: The system model is given in Sec. II. In Sec. III, we briefly summarize SNC and then derive analytical delay bounds for the considered scenarios. We present numerical evaluations in Sec. IV and our conclusions in Sec. V.

II. System Model

We consider downlink transmissions in a time-slotted system from a single base station with M antennas to U users. We consider the case $U > M$, where the transmitter cannot serve all users at once. Instead, in each time slot t, only a subset $\mathcal{K}_t \subset \{1, \ldots, U\}$ of users are scheduled for transmission, with $K_t \triangleq |\mathcal{K}_t| \leq M$. Contrary to [4], we assume that the scheduling scheme does not depend on the channel states. This is because we consider a fairly large number of users U, so that acquiring channel state information (CSI) for all U users would result in an infeasible amount of overhead. Instead, we follow [6], where the channel is estimated only for the scheduled users. We assume that the base station has perfect CSI for all K_t scheduled users.

We describe in Sec. II-A the physical layer transmission for the scheduled users $\mathcal{K}_t$. Round robin scheduling is presented in Sec. II-B. Then, we describe in Sec. II-C the queueing delay of the system on the link layer, followed by the problem statement in Sec. II-D.

A. Physical Layer Model

The received signal $\mathbf{y}_t \in \mathbb{C}^{K_t \times 1}$ at the K_t scheduled users in time slot t can be described as

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{x}_t + \mathbf{z}_t \,. \tag{1}$$

For the channel matrix $\mathbf{H}_t \in \mathbb{C}^{K_t \times M}$, we assume Rayleigh fading, i.e., all elements are independent and identically distributed (i.i.d.) with Gaussian distribution $\mathcal{CN}(0, 1)$. Furthermore, we consider the quasi-static fading model where the channel $\mathbf{H}_t$ remains constant for the duration of time slot t, consisting of n channel uses, and changes to an independent realization in the next time slot (the set $\mathcal{K}_t$ of scheduled users also changes). The input signal is denoted as $\mathbf{x}_t \in \mathbb{C}^{M \times 1}$ and must satisfy a short-term power constraint $\operatorname{tr}\left(\mathbb{E}\left[\mathbf{x}_t \mathbf{x}_t^{\mathsf{H}}\right]\right) \leq P_\Sigma$ for each realization of $\mathbf{H}_t$. The noise $\mathbf{z}_t \in \mathbb{C}^{K_t \times 1}$ has i.i.d. components $\mathcal{CN}(0, 1)$.

Given the channel matrix $\mathbf{H}_t$, the transmitter must encode the data for the K_t scheduled users into coded symbols $\mathbf{x}_t$. We now present two different encoding strategies.

1) Zero-Forcing Beamforming (ZFBF): When the transmitter applies ZFBF, the input signal vector $\mathbf{x}_t$ is given by [1]

$$\mathbf{x}_t = \mathbf{V}_t \mathbf{P}_t^{1/2} \mathbf{s}_t \tag{2}$$

where $\mathbf{V}_t$ is the precoding matrix, $\mathbf{P}_t = \operatorname{diag}(\rho_{t,1}, \ldots, \rho_{t,K})$ is the power allocation matrix, and $\mathbf{s}_t$ is the $K_t \times 1$ vector of (independently) coded Gaussian symbols for the K_t scheduled users. In this case, the input signal $\mathbf{x}_t$ is also Gaussian. The ZFBF precoder is given as [1]

$$\mathbf{V}_t = \mathbf{H}_t^{\mathsf{H}} (\mathbf{H}_t \mathbf{H}_t^{\mathsf{H}})^{-1} \mathbf{\Xi}_t^{1/2} \tag{3}$$

where $\mathbf{H}_t^{\dagger} = \mathbf{H}_t^{\mathsf{H}} (\mathbf{H}_t \mathbf{H}_t^{\mathsf{H}})^{-1}$ is the Moore-Penrose pseudo-inverse of $\mathbf{H}_t$ and $\mathbf{\Xi}_t = \operatorname{diag}(\xi_{t,1}, \ldots, \xi_{t,K})$ is the normalization matrix such that the columns of $\mathbf{V}_t$ have unit-2 norm. The variables $\xi_{t,k}$ are central chi-square distributed (scaled by a factor $1/2$) with $2m_t$ degrees of freedom, where $m_t = M - K_t + 1$. Their PDF is given by [1, Lemma 4]

$$f_m(\xi) = \frac{1}{\Gamma(m)} \xi^{m-1} e^{-\xi} \,. \tag{4}$$

We asume that the blocklength n of the channel code is sufficiently long, so that the system can achieve error-free transmission to user k at a rate [1]

$$R_k(t) = \log_2(1 + \rho_{t,k} \xi_{t,k}) \,. \tag{5}$$

2) Zero-Forcing with Dirty-Paper Coding (ZF-DPC): For comparison, we also present a scheme known originally as *ranked known interference (RKI)* [1]. Assume that the scheduled users $\mathcal{K}_t \subset \{1, \ldots, U\}$ are ordered from 1 to K_t. When a scheduled user $k \in \mathcal{K}_t$ is the κ-th ordered user, it experiences interference from the ordered users $1, \ldots, \kappa - 1$. The interference from those users is non-causally known at the transmitter. Therefore, the transmitter can employ dirty paper coding (DPC) when encoding the data for the κ-th ordered user, which allows sending data at the same rate as if no interference was present. Furthermore, if the ordered users $\kappa + 1, \ldots, K_t$ apply zero-forcing (ZF) towards the users $1, \ldots, \kappa$, then they will not interfere with the κ-th user. Thus, when user k is the κ-th ordered user, it can achieve a rate $R_k(t) = \log_2(1 + \rho_{t,k} \xi_{t,k})$, where the variables $\xi_{t,k}$ have central chi-square distribution (scaled by $1/2$) with $2m_{t,k}$ degrees of freedom with $m_{t,k} = M - \kappa + 1$. The PDF of $\xi_{t,k}$ is given by (4). Note that $m_{t,k}$ and the rates $R_k(t)$ depend on the user ordering.

B. Round Robin (RR) Scheduling

In the considered scenario, the number of users U exceeds the number of transmit antennas M. Therefore, the transmitter must schedule a subset $\mathcal{K}_t$ of users in each time slot t.

We consider round robin (RR) scheduling as in [6], where multiple users can be scheduled in each time slot. Each user k is scheduled exactly once within a superframe of T slots. The average number of scheduled users per slot is then given as $\overline{K} \triangleq U/T$, with $1 \leq \overline{K} \leq M$. As the total number of users U is fixed, $\overline{K}$ may not always be an integer, and thus the scheduler must sometimes select more than $\overline{K}$ users, sometimes less. We assume that in T_A of the subslots, $K_\mathsf{A} = \lceil \overline{K} \rceil$ users are served, in $T_\mathsf{B} = T - T_\mathsf{A}$ of the subslots, $K_\mathsf{B} = \lfloor \overline{K} \rfloor$ users are served, such that the total number of users served in the superframe is $T_\mathsf{A} K_\mathsf{A} + T_\mathsf{B} K_\mathsf{B} = U$.

In order to maintain fairness between the users, the transmitter randomly assigns the users to the slots in each superframe. Furthermore, in case of ZF-DPC, where the performance depends on the encoding order of the users, we require that the users are ordered randomly.

C. Link Layer Model

In time slot t, $A_k(t)$ data bits intended for downlink transmission to user k arrive at the transmitter. The data is stored in a transmit buffer, with individual buffers (or queues) for each user. We assume that the arrival process $A_k(t)$ is constant over time and equal for all users, with α denoting the constant number of bits that arrive in the queue of each user in each time slot. The service rate offered by the wireless system in each time slot to a scheduled user $k \in \mathcal{K}_t$ is given by $S_k(t) = nR_k(t)$, and $S_k(t) = 0$ when $k \notin \mathcal{K}_t$. The departure process $D_k(t)$ describes the amount of data that is transmitted to the receiver. Thus, $D_k(t)$ is limited both by the amount of data waiting in the buffer, as well as by the service rate $S_k(t)$. The cumulative arrival, service, and departure processes are defined as

$$\mathbf{A}_k(t_1, t_2) \triangleq \sum_{t=t_1}^{t_2-1} A_k(t) \ , \quad \mathbf{S}_k(t_1, t_2) \triangleq \sum_{t=t_1}^{t_2-1} S_k(t) \ , \tag{6}$$

$$\mathbf{D}_k(t_1, t_2) \triangleq \sum_{t=t_1}^{t_2-1} D_k(t) \ . \tag{7}$$

The queueing delay $W_k(t)$ of user k at time t is defined as the time it takes for all data that arrived prior to time t to depart from the transmit buffer and reach the receiver [9], [15]:

$$W_k(t) \triangleq \inf \{ u \geq 0 : \quad \mathbf{A}_k(0, t) \leq \mathbf{D}_k(0, t+u) \} \ . \tag{8}$$

The delay $W_k(t)$ is random. We want to find the probability $p_{\mathrm{v},k}(w)$ that the delay $W_k(t)$ of the data for user k exceeds a specified target delay w at any time t:

$$p_{\mathrm{v},k}(w) \triangleq \sup_{t \geq 0} \{ \mathbb{P} \{ W_k(t) > w \} \} \ . \tag{9}$$

D. Problem Statement

In this work, we want to find the value $\overline{K}$ that minimizes the delay violation probability $p_{\mathrm{v},k}(w)$. On the one hand, choosing a small value of $\overline{K}$ means that only few users are scheduled in each time slot, so that their signals are transmitted with high beamforming gain and transmit power. However, this also results in a small multiplexing gain and a long time to schedule all users. On the other hand, a large value $\overline{K}$ results in poor beamforming gain.

We note that the delay violation probability $p_{\mathrm{v},k}(w)$ cannot be determined directly in an analytically tractable form. However, the delay violation probability can be analytically approximated/bounded using the frameworks of effective capacity [10] or stochastic network calculus [8], [9]. Effective capacity provides an approximation for $p_{\mathrm{v},k}(w)$ that is tight for large w. In this work, we perform the optimization of $\overline{K}$ based on stochastic network calculus, as it provides a strict upper bound on $p_{\mathrm{v},k}(w)$ that holds also for small w.

III. Analysis

In Sec. III-A, we present a summary of the delay analysis through stochastic network calculus in a transform domain [9]. We demonstrate in Sec. III-B how stochastic network calculus can be used when round robin scheduling is used. In Sec. III-C, we analytically obtain the stochastic network calculus bounds for the considered scenario. Note that the transmission and scheduling strategies in Sec. II are fair, as the distribution of the service process $S_k(t)$ is the same for all users. We assume that all users are subject to the same delay requirements and thus drop the subscript k to shorten the notation.

A. Stochastic Network Calculus (SNC)

This section closely follows our previous work [15] and provides a summary of stochastic network calculus [8], [9].

The delay $W(t)$ in (8) is defined in terms of the arrival and departure processes. However, the distribution of the delay can be found directly from the statistics of the arrival and service processes. We follow [9] and describe these processes in the exponential domain, also called *SNR domain*. The arrival and service processes in the bit domain, $A(t)$ and $S(t)$, are converted to the SNR domain as

$$\mathcal{A}(t) \triangleq e^{A(t)} \ , \quad \mathcal{S}(t) \triangleq e^{S(t)} \ . \tag{10}$$

In this work, we assume constant arrivals with $A(t) = \alpha$. Consider for now a service process $S(t)$ that is independent and identically distributed (i.i.d.) between time slots. An upper bound on the delay violation probability $p_\mathrm{v}(w)$ can then be obtained in terms of the Mellin transforms of $\mathcal{A}$ and $\mathcal{S}$. The Mellin transform $\mathcal{M}_\mathcal{X}(\theta)$ of a nonnegative random variable $\mathcal{X}$ is defined as [9]

$$\mathcal{M}_\mathcal{X}(\theta) \triangleq \mathbb{E}\left[\mathcal{X}^{\theta-1}\right] \tag{11}$$

for a parameter $\theta \in \mathbb{R}$. For the analysis, we choose $\theta > 0$ and check if the stability condition $\mathcal{M}_\mathcal{A}(1+\theta)\mathcal{M}_\mathcal{S}(1-\theta) < 1$ holds. If it holds, define the kernel [9], [16]

$$\begin{aligned}\mathbb{K}(\theta, w) &\triangleq \lim_{t\to\infty} \sum_{u=0}^{t} \mathcal{M}_\mathcal{A}(1+\theta)^{t-u} \cdot \mathcal{M}_\mathcal{S}(1-\theta)^{t+w-u} \\ &= \frac{\mathcal{M}_\mathcal{S}(1-\theta)^w}{1 - \mathcal{M}_\mathcal{A}(1+\theta)\mathcal{M}_\mathcal{S}(1-\theta)} \ .\end{aligned} \tag{12}$$

For any parameter $\theta > 0$, the kernel $\mathbb{K}(\theta, w)$ provides an upper bound on the delay violation probability $p_\mathrm{v}(w)$ [9], [16].

This holds for any time slot t, including the limit $t \to \infty$ (steady-state). In order to find the tightest upper bound, one must find the parameter $\theta > 0$ that minimizes $\mathbb{K}(\theta, w)$:

$$p_{\mathrm{v}}(w) \leq \inf_{\theta>0} \left\{ \mathbb{K}(\theta, w) \right\} \, . \tag{13}$$

In the scenarios investigated in this paper, the stability condition is only satisfied for a limited range $0 < \theta < \theta_{\max}$. We can then finely quantize this range and perform an exhaustive search for the minimum/infimum in (13).

B. SNC and Round Robin Scheduling

For round robin scheduling, the delay analysis through stochastic network calculus as shown in Sec. III-A cannot be applied directly, as $S(t)$ is zero in the time slots where the user is not scheduled, i.e., $S(t)$ is not i.i.d. between time slots. However, stochastic network calculus can be applied on the superframe level. The service that a user receives in superframe i is denoted as $S^{(T)}(i)$, and is i.i.d. between superframes, because each user is scheduled exactly once per superframe of length T. The arrival process on the superframe level is given as $A^{(T)}(i) = \alpha T$ bits, and the Mellin transform of the process $\mathcal{A}$ in the SNR domain is

$$\mathcal{M}_{\mathcal{A}^{(T)}}(\theta) = e^{\alpha T(\theta - 1)} \, . \tag{14}$$

Assume first that w/T, where w is maximum delay in time slots, is an integer: Then, the queueing analysis can easily be done on the superframe level:

$$p_{\mathrm{v}}(w) \leq \mathbb{K}^{(T)}\left(\theta, \frac{w}{T}\right) \, , \tag{15}$$

with

$$\mathbb{K}^{(T)}\left(\theta, \frac{w}{T}\right) = \frac{\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta)^{\frac{w}{T}}}{1 - \mathcal{M}_{\mathcal{A}^{(T)}}(1+\theta)\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta)} \, . \tag{16}$$

In case w/T is not an integer, some users (denoted as group 1) will be served $\lceil w/T \rceil$ times before the deadline, while others (group 2) will only be served $\lfloor w/T \rfloor$ times. For the sake of fairness, we assume that the users are assigned randomly to the slots. Then, the probability of being in the second group is $p_2 = \frac{\mathrm{mod}\ (w,T)}{T}$, and $p_1 = 1 - p_2$. Thus, the overall bound on the delay violation probability is given by

$$p_{\mathrm{v}}(w) \leq p_1 \mathbb{K}^{(T)}\left(\theta, \left\lceil \frac{w}{T} \right\rceil\right) + p_2 \mathbb{K}^{(T)}\left(\theta, \left\lfloor \frac{w}{T} \right\rfloor\right) \, . \tag{17}$$

Similar to (13), this bound holds for any $\theta > 0$, such that finding the tightest possible bound requires taking the infimum over (17) with respect to θ.

C. Delay Analysis for MU-MISO Downlink

The kernel (16) depends on the Mellin transform of the service $\mathcal{S}^{(T)}$ offered to each user in each superframe. Users are scheduled exactly once in a superframe, so that $\mathcal{S}^{(T)}$ has the same distribution as the service $\mathcal{S}$ experienced by a scheduled user. The SNR-domain service process of a scheduled user is given as $\mathcal{S} = e^{S} = e^{nR}$, with $R = \log_2(1 + \rho\xi)$.

For ZFBF and ZF-DPC, ξ is a scaled central χ^2 variable with varying degrees of freedom $2m$ as outlined in Sec. II-A. For ZFBF, we have $m = M - K_{(\mathsf{A}/\mathsf{B})} + 1$, depending on the slot type (A/B). For ZF-DPC, $m \in \left\{1, \ldots, M - K_{(\mathsf{A}/\mathsf{B})} + 1\right\}$, each with probability $p_{m|(\mathsf{A}/\mathsf{B})} = 1/K_{(\mathsf{A}/\mathsf{B})}$.

The transmitter is subject to a short-term power constraint $\mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}_t \mathbf{x}_t^{\mathsf{H}}\right]\right) \leq P_{\Sigma}$. A simple power allocation strategy shares P_{Σ} equally among the K_{A} or K_{B} scheduled users:

$$\rho = \begin{cases} \rho_{\mathsf{A}} = \frac{P_{\Sigma}}{K_{\mathsf{A}}} & \text{with prob.} \quad p_{\mathsf{A}} = \frac{K_{\mathsf{A}} T_{\mathsf{A}}}{U} \\ \rho_{\mathsf{B}} = \frac{P_{\Sigma}}{K_{\mathsf{B}}} & \text{with prob.} \quad p_{\mathsf{B}} = \frac{K_{\mathsf{B}} T_{\mathsf{B}}}{U} \end{cases} \, . \tag{18}$$

The Mellin transform of the service process $\mathcal{S}^{(T)}$ can be obtained by averaging over the Mellin transforms of the service process with specific values of ρ and m:

$$\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta) = \sum_{\rho,m} p_{\rho,m} \mathcal{M}_{\mathcal{S}^{(T)}|\rho,m}(1-\theta) \, , \tag{19}$$

where $p_{\rho,m}$ denotes the joint probability of a user's channel having $2m$ degrees of freedom ($\xi \sim \frac{1}{2}\chi^2_{2m}$) and power ρ.[1]

For a specific constant power ρ and a specific m, the Mellin transform of the service process can be obtained as

$$\mathcal{M}_{\mathcal{S}^{(T)}|\rho,m}(1-\theta) = \mathbb{E}\left[\left(e^{nR}\right)^{-\theta} \middle| \rho, m \right] \tag{20}$$

$$= \mathbb{E}\left[(1+\rho\xi)^{-\frac{\theta n}{\ln 2}} \middle| m \right] \, . \tag{21}$$

We define $\tilde{\theta} \triangleq \frac{\theta n}{\ln 2}$ and follow the derivations in [17] to obtain

$$\mathbb{E}\left[(1+\rho\xi)^{-\tilde{\theta}} \middle| m \right] = \int_0^{\infty} (1+\rho\xi)^{-\tilde{\theta}} f_m(\xi) d\xi \tag{22}$$

$$= \int_0^{\infty} (1+\rho\xi)^{-\tilde{\theta}} \frac{1}{\Gamma(m)} \xi^{m-1} e^{-\xi} d\xi \tag{23}$$

$$= \sum_{\mu=0}^{m-1} \frac{\binom{m-1}{\mu}(-1)^{\mu}}{\Gamma(m)\rho^{m-1}} \int_0^{\infty} (1+\rho\xi)^{m-1-\mu-\tilde{\theta}} e^{-\xi} d\xi \tag{24}$$

$$= \sum_{\mu=0}^{m-1} \frac{\binom{m-1}{\mu}(-1)^{\mu}}{\Gamma(m)\rho^{\mu+\tilde{\theta}}} e^{\frac{1}{\rho}} \cdot \int_0^{\infty} \left(\frac{1}{\rho} + \xi\right)^{m-1-\mu-\tilde{\theta}} e^{-\left(\frac{1}{\rho}+\xi\right)} d\xi \tag{25}$$

$$= \sum_{\mu=0}^{m-1} \frac{\binom{m-1}{\mu}(-1)^{\mu}}{\Gamma(m)\rho^{\mu+\tilde{\theta}}} e^{\frac{1}{\rho}} \cdot \Gamma\left(m - \mu - \tilde{\theta}, \frac{1}{\rho}\right) \, . \tag{26}$$

In (24), we used the conversion [17]

$$x^{m-1} = \sum_{\mu=0}^{m-1} \binom{m-1}{\mu} (1+x)^{m-1-\mu} (-1)^{\mu} \tag{27}$$

and in (26), we applied the upper incomplete Gamma function

$$\Gamma(s, x) = \int_x^{\infty} t^{s-1} e^{-t} dt \, . \tag{28}$$

Thus, given the arrival rate α in bits per time slot and a specific choice of superframe length T (which determines the average number of scheduled users $\overline{K} = U/T$), the upper bound (17) on $p_{\mathrm{v}}(w)$ can be obtained analytically through (19) and (26).

[1] For ZFBF, $p_{\rho,m}$ is equal to p_{A} or p_{B} as given in (18). For ZF-DPC, the different $p_{\rho,m}$ can simply be obtained as $p_{\mathsf{A}/\mathsf{B}} \cdot p_{m|(\mathsf{A}/\mathsf{B})}$.

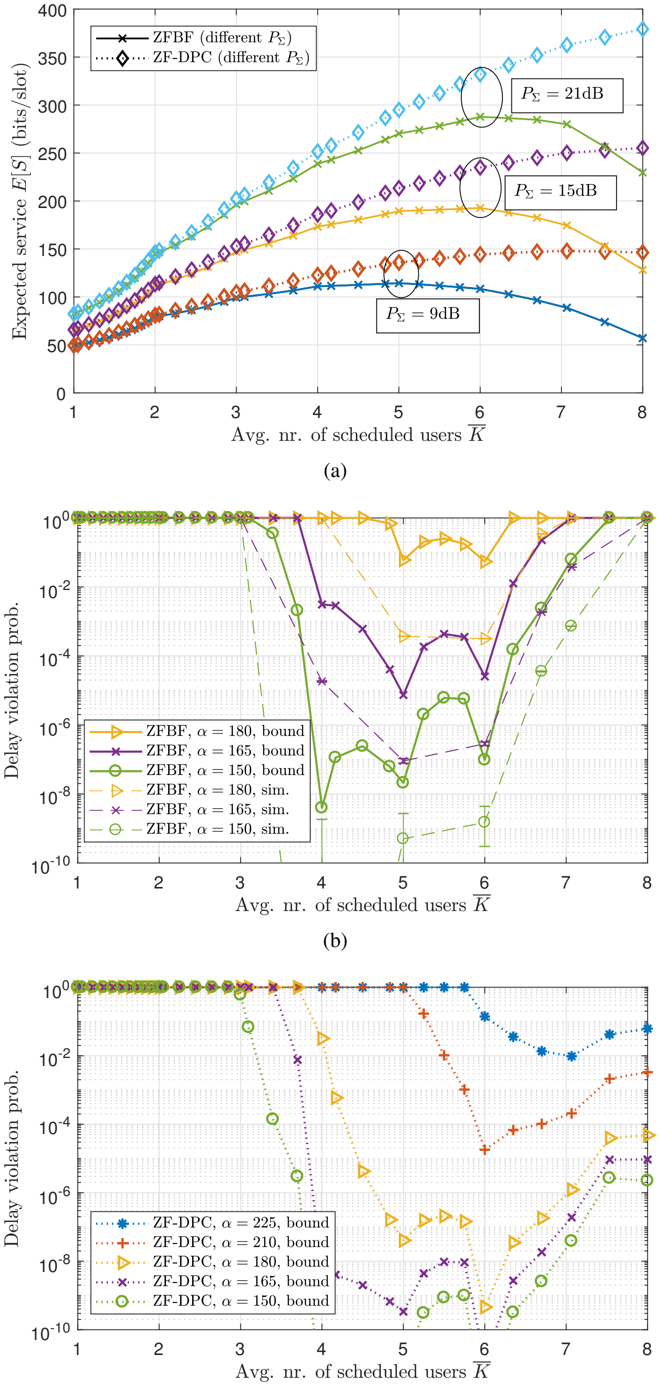
IV. Numerical Results

In Fig. 1, we show various aspects of the performance of a system with $U = 120$ users and $M = 8$ antennas. First, in Fig. 1a, we show the expected service rate per slot vs. the average number of scheduled users $\overline{K}$ for different values of the SNR $P_\Sigma \in \{9, 15, 21\}$ dB. Note that the superframe length T is an integer number, but $\overline{K} = U/T$ is not always integer. In each superframe of T time slots, each user receives nR bits. Thus, the expected service rate per user and time slot is

$$\mathbb{E}\left[S\right] = \frac{1}{T}\mathbb{E}\left[S^{(T)}\right] = \frac{1}{T}\mathbb{E}\left[nR\right] . \qquad (29)$$

For ZFBF, we observe for every SNR P_Σ that the expected service rate first increases and then decreases in $\overline{K}$. At very small $\overline{K}$, an increase in $\overline{K}$ means that more users are scheduled simultaneously, and the multiplexing gain from transmitting to multiple users outweighs the performance loss due to slightly decreased service rates of each user. However, at very large $\overline{K}$, the relative increase in the number of scheduled users is small, whereas the beamforming gain is massively reduced. The value of $\overline{K}$ that maximizes the expected service rate grows with the SNR, in line with previous results [3]. For ZF-DPC, the expected service rate is strictly increasing in $\overline{K}$ for $P_\Sigma \in \{15, 21\}$ dB. This is because the additional users do not create any interference towards the previous users. The only downside from adding more users to the ZF-DPC system is that the transmit power is shared with the new users. When $P_\Sigma = 9$ dB, this leads to a tiny reduction in $\mathbb{E}\left[S\right]$ at $\overline{K} = 8$.

In Fig. 1b and Fig. 1c, we consider the delay performance of the system for ZFBF and ZF-DPC, respectively, with different arrival rates α, a maximum delay of $w = 60$ time slots, and with $P_\Sigma = 15$ dB. For ZFBF, Fig. 1b shows that the delay violation probability, obtained from the analytical bound (17) (solid curves), remains high at $\alpha = 180$ bits/slot. However, when the arrival rate is decreased, the bound decreases significantly. Interestingly, the minimum is attained at $\overline{K} = 6$ for $\alpha = 180$, at $\overline{K} = 5$ for $\alpha = 165$, and at $\overline{K} = 4$ for $\alpha = 150$ bits/slot. Thus, the optimal value of $\overline{K}$ changes depending on the arrival rate and delay constraints imposed on the system. Many of our additional experiments also show that the optimal $\overline{K}$ under delay constraints is slightly below the value of $\overline{K}$ that maximizes the expected service rate. An explanation for this phenomenon is that even though decreasing the number of users $\overline{K}$ means that users are scheduled less often (lower multiplexing gain), the system has a higher beamforming gain, i.e., the channel gains ξ of all users have more degrees of freedom. This decreases the variance of the service S experienced by each user and thus improves the delay performance of the system. In addition to the bound (17) on the delay violation probability $p_{\mathrm{v}}(w)$, Fig. 1b shows also the actual delay violation probability (dashed curves), which was obtained empirically by simulating the queueing system over 10^{10} time slots. As this is computationally intensive, simulations were only performed for integer values of $\overline{K}$ and for $\overline{K} \in \{120/18, 120/17\}$. Although the bounds are not tight (which was observed also in other works on SNC [9], [15], [17]), the bounds correctly predict the number of scheduled users $\overline{K}$ that minimizes the actual delay violation probability $p_{\mathrm{v}}(w)$. Note that for $\alpha = 150$, one would need even longer simulations to reach sufficient confidence in $p_{\mathrm{v}}(w)$.



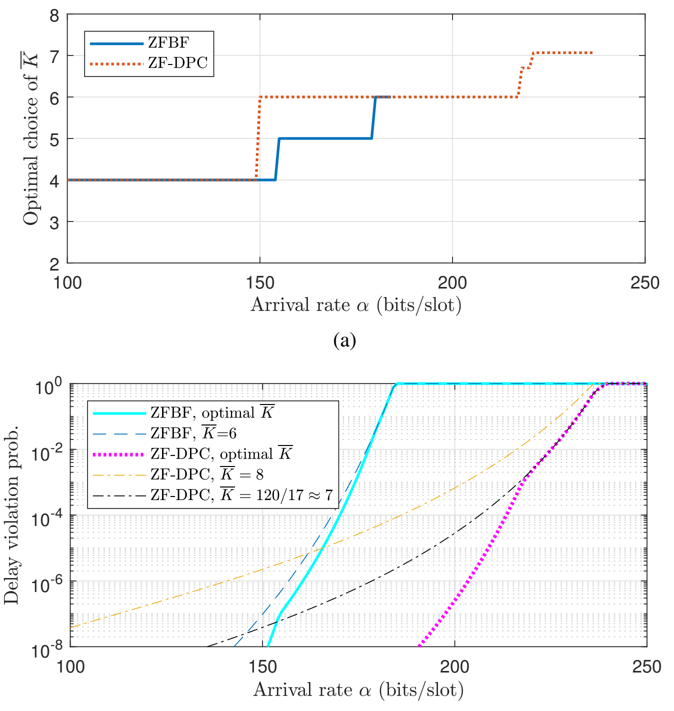
(a)

(b)

(c)

Fig. 1. $M = 8$, $U = 120$ users, $n = 1000$. (a): Expected service rate for $P_\Sigma \in \{9, 15, 21\}$ dB. (b) Delay violation probability for ZFBF, according to the SNC bound, for deadline $w = 60$ slots and different arrival rates α, $P_\Sigma = 15$ dB. Also showing $p_{\mathrm{v}}(w)$ from simulations over 10^{10} slots, with 95% confidence intervals. (c) same parameters, but for ZF-DPC.

Fig. 1c shows the delay violation probability for ZF-DPC. Here, we observe that the minimum in the delay violation



(a)

(b)

Fig. 2. $M = 8$ antennas, $U = 120$ users, $n = 1000$ symbols, $P_{\Sigma} = 15$ dB, $w = 60$. (a): Optimal choice of $\overline{K}$ such that the delay violation probability is minimized. (b) Delay violation probability, with optimal $\overline{K}$ for each point, along with suboptimal fixed values of $\overline{K}$.

probability is attained at $\overline{K} = 120/17 \approx 7$ for $\alpha = 225$ and at $\overline{K} = 6$ for $\alpha = 210$, whereas Fig. 1a showed that the expected service rate is maximized at $\overline{K} = 8$. The explanation is similar to the explanation in case of ZFBF: When scheduling $\overline{K} = 8$ users, the effective channel gains ξ for some of the users (the users which are encoded last in the ZF-DPC order) have only 2 degrees of freedom. These users may experience very low data rates, so that the delay violation probability increases.

In Fig. 2, we further investigate the optimal value of $\overline{K}$ and how the optimal choice of $\overline{K}$ influences the delay performance. Fig. 2a shows the optimal values for $\overline{K}$ for ZFBF (blue, solid lines) and ZF-DPC (red, dotted lines). We observe that the optimal value of $\overline{K}$ decreases when the arrival rate α is reduced. In Fig. 2b, we investigate how the optimal choice of $\overline{K}$ affects the delay performance of the considered systems. In case of ZFBF, we find that choosing the suboptimal value $\overline{K} = 6$ deteriorates the performance only slightly. For ZF-DPC, the selected value of $\overline{K}$ seems to have a larger impact. We observe that choosing the value $\overline{K} = 8$, i.e., the value that maximizes the expected service rate $\mathbb{E}[S]$ of the system, would lead to a massive increase in the delay violation probability.

V. Conclusions

In this work, we have presented an analytical framework to study the delay performance of the multiuser MISO downlink. We found that the optimal number of scheduled users depends on the delay requirements of the system. There are many interesting possible extensions of this work. First of all, we considered equal power allocation, whereas the transmitter could also optimize the transmission power. In another line of research, one could investigate systems with very large numbers of transmit antennas (massive MIMO). Finally, when the maximum tolerable delay becomes very short, the length of each time slot should also be chosen very small. For very short time slots, one must take into account that the channel estimates may become inaccurate, and also consider the impact of channel coding at finite blocklength in order to gain more realistic insights into the system performance.

References

[1] G. Caire and S. Shamai, "On the achievable throughput of a multiantenna gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1691–1706, 2003.

[2] C. B. Peel, B. M. Hochwald, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part I: channel inversion and regularization," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 195–202, 2005.

[3] B. Hochwald and S. Vishwanath, "Space-time multiple access: Linear growth in the sum rate," in *Proc. 40th Annual Allerton Conf. Communications, Control and Computing*, 2002.

[4] T. Yoo and A. Goldsmith, "On the optimality of multiantenna broadcast scheduling using zero-forcing beamforming," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 3, pp. 528–541, 2006.

[5] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 506–522, 2005.

[6] J. Zhang, M. Kountouris, J. G. Andrews, and R. W. Heath, "Multi-mode transmission for the MIMO broadcast channel with imperfect channel state information," *IEEE Trans. Commun.*, vol. 59, no. 3, pp. 803–814, 2011.

[7] N. Ravindran and N. Jindal, "Multi-user diversity vs. accurate channel state information in MIMO downlink channels," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3037–3046, Sept. 2012.

[8] M. Fidler, "A network calculus approach to probabilistic quality of service analysis of fading channels," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2006, pp. 1–6.

[9] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 204–217, Feb. 2016.

[10] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[11] L. Liu and J. F. Chamberland, "On the effective capacities of multiple-antenna gaussian channels," in *2008 IEEE Int. Symp. Inf. Theory*, July 2008, pp. 2583–2587.

[12] E. A. Jorswieck, R. Mochaourab, and M. Mittelbach, "Effective capacity maximization in multi-antenna channels with covariance feedback," *IEEE Trans. Wireless Commun.*, vol. 9, no. 10, pp. 2988–2993, 2010.

[13] M. C. Gursoy, "MIMO wireless communications under statistical queueing constraints," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 5897–5917, Sept. 2011.

[14] J. Li, N. Bao, W. Xia, and L. Shen, "Adaptive user scheduling and resource management for multiuser MIMO downlink systems with heterogeneous delay requirements," in *IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Apr. 2013, pp. 1351–1356.

[15] S. Schiessl, H. Al-Zubaidy, M. Skoglund, and J. Gross, "Delay performance of wireless communications with imperfect CSI and finite length coding," *IEEE Trans. Commun.*, in press. [Online]. Available: https://arxiv.org/abs/1608.08445

[16] S. Schiessl, J. Gross, and H. Al-Zubaidy, "Delay analysis for wireless fading channels with finite blocklength channel coding," in *Proc. 18th ACM Int. Conf. Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*. ACM, 2015, pp. 13–22.

[17] S. Schiessl, H. Al-Zubaidy, M. Skoglund, and J. Gross, "Finite length coding in edge computing scenarios," in *Proc. 21th Int. ITG Workshop on Smart Antennas (WSA)*, Mar. 2017, pp. 1–6.