Delay Performance of MISO Wireless Communications

Jesús Arnau and Marios Kountouris
Mathematical and Algorithmic Sciences Lab
Paris Research Center, Huawei France
Emails: jesus.arnau@ieee.org, marios.kountouris@huawei.com

***Abstract*—Ultra reliable, low latency communications (URLLC) are currently attracting significant attention due to the emergence of mission-critical applications and device-centric communication. URLLC will entail a fundamental paradigm shift from throughput-oriented system design towards holistic designs for guaranteed and reliable end-to-end latency. A deep understanding of the delay performance of wireless networks is essential for efficient URLLC systems. In this paper, we investigate the network layer performance of multiple-input, single-output (MISO) systems under statistical delay constraints. We provide a statistical characterization of MISO diversity-oriented service process through closed-form expressions of its Mellin transform and derive probabilistic delay bounds using tools from stochastic network calculus. In particular, we analyze transmit beamforming with perfect and imperfect channel knowledge and compare it with orthogonal space-time codes and antenna selection. The effect of transmit power and number of antennas on the delay distribution is also investigated. Our results provide useful guidelines for the design of communication systems that can guarantee the stringent URLLC latency requirements.**

***Index Terms*—URLLC, 5G systems, MIMO, diversity, stochastic network calculus, queueing analysis.**

I. Introduction

In order to handle the ever-increasing traffic load, existing wireless networks have typically been designed with a focus on improving spectral efficiency and increasing coverage. The latency requirements of different applications have mostly been an after-thought. Ultra reliable, low latency communications (URLLC) has not been in the mainstream of most wireless networks, due to the focus on human-centric communications, delay-tolerant content and reliability levels in the order of 95-99%. However, a plethora of socially useful applications and new uses of wireless communication are currently envisioned in areas such as industrial control, smart cities, augmented/virtual reality (AR/VR), automated transportation and algorithmic trading. In response, new releases of mobile cellular networks (mainly 5G new radio and beyond) are envisaged to support URLLC scenarios with strict requirements in terms of latency (ranging from 1 ms to few milliseconds end-to-end latency depending on the use cases) and reliability (higher than 99.9999%) [1].

URLLC and device-centric communication pose significant theoretical and practical challenges, requiring a departure from capacity-oriented system design towards a holistic view (network architecture, control, and data). Despite recent efforts, more work is needed to better understand the non-asymptotic fundamental tradeoffs between delay, reliability and throughput, including both coding delays and queueing delays.

In networking, delay is a key performance measure and queueing theory has been instrumental in providing exact solutions for backlog and delays in packet-switched networks. However, queueing network analysis is largely restricted to few interacting (coupled) queues, small network topologies and Poisson arrivals. Classical queueing models typically allow the analysis of average delay, failing to characterize delay quantiles (worst-case delay) and distributions, which are of cardinal importance in mission-critical applications. Various efforts to combine queueing with communication theory, such as stochastic network calculus [2]–[5], timely throughput [6], effective bandwidth [2], and effective capacity [7] to name a few, take on a different approach and compute performance bounds for a wide range of stochastic processes. These approaches promise significant performance gains - in terms of latency, reliability and throughput - and crisp insights for the design of low latency communication systems.

In this paper, we use stochastic network calculus, which allows non-asymptotic statistical bounds on network layer performance metrics, such as maximum delay and delay violation probability. In particular, we use the $(\min,\times)$ network calculus methodology developed in [8]. For ultra reliable communications, we focus on multiple-input multiple-output (MIMO) techniques, which have great potential to combat fading (diversity), increase spectral efficiency (multiplexing), and reduce interference. In particular, we consider three MIMO diversity techniques: (i) maximum ratio transmission (MRT), a transmit beamforming technique that maximizes the received signal and realizes diversity exploiting channel state information (CSI) at the transmitter; (ii) orthogonal space-time block coding (OSTBC), which provides diversity with no CSI; and (iii) antenna selection, which relies on low-rate CSI.

A. Related Work

Despite the extended literature on MIMO techniques at the physical layer, only few attempts have been made to characterize the upper layer performance of multi-antenna techniques taking into account the queueing effects. In [9] the service process of an adaptive MIMO system with Poisson arrivals is characterized. Bounds on the delay violation probability have been derived for MIMO multiple access with bursty traffic in [10], while [11] provides an asymptotic analysis of the diversity-multiplexing tradeoff for MIMO systems with bursty and delay-limited information. Using large deviations, [12] analyzes the queueing performance of queue-aware scheduling in multiuser MIMO systems. Bounds on the tail of delay of MIMO communication systems have been derived using the effective capacity framework [13]–[15]. Using Markov chains to reproduce the state of Gilbert-Elliott fading channels, the flow-level performance of MIMO spatial multiplexing has been analyzed using stochastic network calculus in [16], [17]. Nevertheless, to the best of our knowledge, there is no work that considered the delay performance of MIMO schemes using stochastic network calculus for wireless fading channels.

B. Contributions

We investigate the upper layer delay performance of multiple input, single output (MISO) diversity communication in the presence of statistical delay constraints. We consider MRT transmit beamforming at the physical layer and derive probabilistic delay bounds using tools from stochastic network calculus. For that, we provide a statistical characterization of the cumulative service process for MISO beamforming channels with both perfect and imperfect CSI through closed-form expressions of its Mellin transform. The impact of transmit antennas, signal-to-noise ratio (SNR), and imperfect CSI on the delay distribution of MISO MRT systems is characterized. We then show that our mathematical framework can be applied to the statistical characterization of various MIMO service processes, including antenna selection and OSTBC. This allows us to compare the delay performance of transmit beamforming with alternative diversity-achieving techniques with less stringent CSI requirements. Interestingly, MISO MRT is shown to reduce the delay violation probability as compared to single-antenna transmissions even with imperfect CSI. The derived delay bounds enable us to assess the robustness of MISO MRT delay performance with respect to channel imperfections. Our results also show under which operating parameters other diversity-techniques are preferable than MRT in terms of delay violation probability. In addition, we provide an asymptotic statistical characterization of the service process in the low/high SNR regime and for large number of antennas.

The rest of the paper is organized as follows: In Section II, we provide our system model and in Section III, a brief background on the (min,×) network calculus is presented. In Section IV, the delay performance analysis of MISO diversity systems is derived. Section V provides the delay performance in asymptotic regimes. Numerical results are presented in Section VI, followed by conclusions in Section VII.

II. System model

We consider data transmission over a point-to-point vector communication channel, where a transmitter with M antennas sends the queued data bits to a single-antenna receiver. Time is divided into time slots of duration T (discrete-time model), and at each slot i, the source generates a_i data bits and stores them in a queue. Each slot contains $L + L_m$ symbols, where L denotes the complex data symbols and L_m the metadata (headers, training, estimation, acknowledgments, etc.).

A. Signal model

The received downlink signal $y_i \in \mathbb{C}$ at slot i in a MISO wireless channel is given by

$$y_i = \sqrt{\mathsf{snr}} \cdot \mathbf{h}_i^{\mathrm{H}} \mathbf{x}_i + n_i \tag{1}$$

where $\mathbf{h}_i \in \mathbb{C}^{M\times 1}$ is the flat-fading channel between the transmitter and the receiver at the i-th slot, which is circularly-symmetric complex Gaussian distributed $\mathbf{h} \sim \mathcal{CN}(0,1)$ (Rayleigh fading). The transmitted vector is denoted by $\mathbf{x}_i \in \mathbb{C}^{M\times 1}$, and $n_i \sim \mathcal{CN}(0,1)$ is the additive background noise that may also include (Gaussian) interference from neighboring systems. Our model and analysis can be extended to take into account interference using tools from [18]. A block-fading model is assumed, where the channel remains constant during one slot and varies independently from slot to slot.

Since we focus on ultra-reliable communications, we consider one of the most prominent multi-antenna diversity techniques, namely transmit beamforming, which refers to sending linearly weighted versions of the same signal on each antenna. The transmitted signal can be written as $\mathbf{x}_i = \mathbf{w}_i s_i$, where s_i is the zero-mean data signal at slot i with power $\mathbb{E}\left[|s|^2\right] = 1$, and $\mathbf{w}_i \in \mathbb{C}^{M\times 1}$ is the unit-norm beamforming vector. Note that, since noise is assumed to have unit power, snr represents the average received SNR, whereas the instantaneous SNR in the i-th slot is given by $\gamma_i = \mathsf{snr}|\mathbf{h}_i^{\mathrm{H}}\mathbf{w}_i|^2$.

B. Transmission technique

The performance target is to maximize the instantaneous SNR of the MISO channel. This can be achieved by sending information only in the direction of the channel vector $\mathbf{h}$, as information sent in any orthogonal direction will be nulled out by the channel anyway. We thus consider maximum ratio transmission [19], which is equivalent to eigen-beamforming since beamforming along the dominant (and only) eigenmode of the $M \times 1$ vector channel is performed.

Assuming that both transmitter and receiver have perfect CSI, the MRT beamforming vector is given by $\mathbf{w}_i = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|}$. In that case, the instantaneous SNR is $\gamma_i = \mathsf{snr}\|\mathbf{h}_i\|^2$, which is gamma distributed with shape parameter M and scale parameter snr, i.e. $\gamma_i \sim \mathrm{Gamma}(M, \mathsf{snr})$. When the transmitter does not fully know the actual channel vector $\mathbf{h}$ (imperfect CSI), we can model its channel knowledge as $\hat{\mathbf{h}} = \mathbf{h} + \mathbf{e}$, where $\mathbf{e} \sim \mathcal{CN}(0, \sigma_{\mathrm{e}}^2\mathbf{I})$. MRT is then performed based on the channel estimate, so that $\mathbf{w} = \frac{\hat{\mathbf{h}}}{\|\hat{\mathbf{h}}\|}$. Particularizing [20, Eq. 7] to the MISO case, the instantaneous SNR is gamma distributed with shape parameter M and scale parameter ζ, i.e.

$$\gamma_i \sim \mathrm{Gamma}(M, \zeta) \quad \text{with} \quad \zeta = \left(\sigma_{\mathrm{e}}^2 + \frac{1+\sigma_{\mathrm{e}}^2}{\mathsf{snr}}\right)^{-1}. \tag{2}$$

This additive error model is consistent with time-division duplex (TDD) operation, where uplink and downlink transmissions take place at the same frequency, in different time instants; assuming they fall within the coherence interval of the channel, then channel reciprocity can be used to estimate the downlink channel from uplink pilot signals. This model also applies to frequency-division duplex (FDD) operation with analog feedback [21]. We only account for the effect of CSI error in MISO beamforming, which reduces the achieved SNR (SNR loss) because of not transmitting exactly in the direction of the actual channel; as we explain in the next subsection there could be another penalty in the rate selection process.

C. Data transmission

A codeword of length L symbols and rate R_i (in bits per symbol) is transmitted at each slot i. The transmitter selects a rate adapted to γ_i and we consider that no errors occur; the achievable rate is equal to the Shannon capacity of the channel, $R_i = \log_2(1+\gamma_i)$. That is valid for MISO systems with knowledge of the fading coefficients of the vector channel and the SNR realization at each slot, which makes the MISO channel behaving equivalently to an AWGN channel with SNR $\mathsf{snr}\|\mathbf{h}\|^2$. Perfect knowledge of the SNR realization implies that the transmitter can adapt the rate to it with no errors. Thus, we only account for the channel estimation error as an SNR penalty. The case of imperfect rate selection in MISO systems, which goes beyond the scope of this work, can be analyzed using for instance the techniques developed in [22].

D. Queuing model

For the analysis of queuing systems, we consider a system-theoretic stochastic model as in [8], which is widely used in the stochastic network calculus methodology. The arrival process a_i models the number of bits that arrive at the queue at a discrete time instant i. For successful transmissions, the service process s_i is equal to LR_i; in case of transmission errors, the service is considered to be zero as no data is removed from the queue. The departure process d_i describes the number of bits that arrive successfully at the destination and depends on both the service process and the number of bits waiting in the queue. Note that acknowledgments and feedback messages are assumed to be instantaneous and error-free.

We define the cumulative arrival, service and departure processes during time interval $[\tau, t)$ for any $0 \leq \tau \leq t$, as

$$A(\tau,t)=\sum_{i=\tau}^{t-1} a_i, \quad S(\tau,t)=\sum_{i=\tau}^{t-1} s_i, \quad D(\tau,t)=\sum_{i=\tau}^{t-1} d_i.$$

For lossless first-in first-out (FIFO) queuing systems, the delay $W(t)$ at time t, i.e. the number of slots it takes for an information bit arriving at time t to be received at the destination, is defined as

$$W(t)=\inf\{u>0 : A(0,t)/D(0,t+u)\leq 1\}. \tag{3}$$

The delay violation probability is given by $\Lambda(w,t) = \mathbb{P}\left[W(t)>w\right]$.

Using the dynamic server property (i.e. $D(0,t) \geq A{*}S(0,t)$ where the $(\min,+)$ convolution operator '*' is defined as $(f*g)(\tau,t)=\inf_{\tau\leq u\leq t}\{f(\tau,u)+g(u,t)\}$ [2]), the delay can be characterized through the cumulative arrival and service processes, which we have so far described in the so-called bit domain. As it is more convenient for the analysis of wireless fading channels, we follow [8] and analyze these processes in the exponential (or SNR) domain.

III. Stochastic Network Calculus in the SNR Domain

A remarkable feature of stochastic network calculus is that it allows to obtain tight bounds on the delay violation probability based on statistical characterizations of the arrival and service processes in terms of their Mellin transforms. For fading channels, it is more convenient to map and analyze these processes into a transfer domain (SNR domain [8]).

The cumulative processes in the SNR domain, denoted by calligraphic letters and converted from the bit domain through the exponential function, are

$$\mathcal{A}(\tau,t)=e^{A(\tau,t)}, \quad \mathcal{D}(\tau,t)=e^{D(\tau,t)}, \quad \mathcal{S}(\tau,t)=e^{S(\tau,t)}.$$

From these definitions, an upper bound on the delay violation probability can be computed by means of the Mellin transforms of $\mathcal{A}(\tau,t)$ and $\mathcal{S}(\tau,t)$:

$$p_{\mathrm{v}}(w)=\inf_{s>0}\left\{K(s,-w)\right\}\geq\Lambda(w) \tag{4}$$

where $K(s,-w)$ is the steady-state kernel, defined as

$$\mathcal{K}(s,-w)=\lim_{t\to\infty}\sum_{u=0}^{t}\mathcal{M}_{\mathcal{A}}(1+s,u,t)\mathcal{M}_{\mathcal{S}}(1-s,u,t+w) \tag{5}$$

where $\mathcal{M}_X(s)=\mathbb{E}\left[X^{s-1}\right]$ denotes the Mellin transform of a nonnegative random variable X for any $s \in \mathbb{C}$ for which the expectation exists.

A. Mellin transform of arrival and service processes

Assuming that $\mathcal{A}(\tau,t)$ has stationary and independent increments, the Mellin transform becomes independent of the time instance, i.e.

$$\begin{aligned}\mathcal{M}_{\mathcal{A}}(s,\tau,t) &= \mathbb{E}\left[\left(\prod_{i=\tau}^{t-1} e^{a_i}\right)^{s-1}\right]\\ &= \mathbb{E}\left[e^{a(s-1)}\right]^{t-\tau}=\mathcal{M}_{\alpha}(s)^{t-\tau}\end{aligned}$$

where we have defined $\alpha = e^{a}$, the non-cumulative arrival process in the SNR domain. We consider the traffic class of $(z(s),\rho(s))$-bounded arrivals, whose moment generating function in the bit domain is bounded by [2]

$$\frac{1}{s}\log\mathbb{E}[e^{sA(\tau,t)}]\leq\rho(s)\cdot(t-\tau)+z(s) \tag{6}$$

for some $s>0$. Restricting ourselves to the case where ρ is independent of s and $z(s)=0$, we have

$$\mathcal{M}_{\alpha}(s)=e^{\rho(s-1)}. \tag{7}$$

For the service process, we start by rewriting $s_i = B\log g(\gamma)$, where $B = L/\log 2$ and $g(\gamma)=1+\gamma$. Assuming that different s_i are independent and identically distributed (i.i.d.), we can express the Mellin transform of the cumulative service as

$$\begin{aligned}\mathcal{M}_{\mathcal{S}}(s,\tau,t) &= \mathbb{E}\left[\left(\prod_{i=\tau}^{t-1} g(\gamma)^B\right)^{s-1}\right] \\ &= \mathbb{E}\left[g(\gamma)^{B(s-1)}\right]^{t-\tau} \\ &= \mathcal{M}_{g(\gamma)}\left(1+B(s-1)\right)^{t-\tau}. \qquad (8)\end{aligned}$$

B. Delay Bound

Plugging (7) and (8) into (5), the steady-state kernel can be rewritten as [8]

$$\mathcal{K}(s,-w) = \frac{[\mathcal{M}_{g(\gamma)}(1-B\cdot s)]^w}{1-\mathcal{M}_\alpha(1+s)\mathcal{M}_{g(\gamma)}(1-B\cdot s)}, \qquad (9)$$

for any $s>0$ under the stability condition $\mathcal{M}_\alpha(1+s)\mathcal{M}_{\mathcal{S}}(1-s)<1$. The delay bound (4) thus reduces to

$$p_{\mathrm{v}}(w) = \inf_{s>0}\left\{\frac{[\mathcal{M}_{g(\gamma)}(1-B\cdot s)]^w}{1-\mathcal{M}_\alpha(1+s)\mathcal{M}_{g(\gamma)}(1-B\cdot s)}\right\}. \qquad (10)$$

IV. Delay Performance: Exact Analysis

In this section, we derive closed-form expressions for the steady-state kernel $\mathcal{K}(s,-w)$ of MISO diversity schemes based on the exact distribution of the instantaneous SNR. We start by providing a general result on the Mellin transform of the service process when the instantaneous SNR is gamma distributed. Obtaining the steady-state kernel for MRT beamforming with both perfect and imperfect CSI is a particularization of this result, which is shown to apply, as a byproduct, for obtaining the performance of other diversity techniques, including MISO OSTBC.

Consider the instantaneous SNR to be a gamma distributed random variable $\gamma \sim \mathrm{Gamma}(M,\zeta)$ with shape parameter M, scale parameter ζ and pdf

$$f_\gamma(x) = \frac{x^{M-1}e^{-\frac{x}{\zeta}}}{\Gamma(M)\zeta^M}, \quad x \geq 0 \qquad (11)$$

where $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,\mathrm{d}x$ is the (complete) gamma function; we have dropped the subindex since SNRs are independent and ergodic. First, we derive the Mellin transform of $g(\gamma)$, i.e. $\mathcal{M}_{g(\gamma)}(s) = \mathbb{E}\left[g(\gamma)^{s-1}\right]$. For notation convenience, in the remainder we assume $B = L/\log 2 = 1$, however in Sec. VI we give again relevant values to this parameter in order to obtain meaningful numerical results.

Preliminary result 1. *The Mellin transform of $g(\gamma) = 1+\gamma$, where $\gamma \sim \mathrm{Gamma}(M,\zeta)$ with $M \in \mathbb{N}^+$ and $\zeta > 0$, is given by*

$$\mathcal{M}_{g(\gamma)}(s) = \zeta^{-M}\cdot U(M, M+s, \zeta^{-1}) \qquad (12)$$

where $U(a,b,z)$ is Tricomi's confluent hypergeometric function [23, Eq. 13.2.5] (also called confluent hypergeometric function of the second kind and denoted by $\Psi(a;b;z)$).

Proof: It can be obtained from [14], where the effective capacity of a MISO system is investigated. In particular, it follows by applying a change of variables to [14, Eq. 9]. ■

A. MISO MRT

The Mellin transform derived above applies directly to the service process with MISO MRT transmission. Using this expression together with the transform of the arrival process, we obtain the kernel and consequently the bound on the delay violation probability as follows

$$p_{\mathrm{v}}(w) = \inf_{s>0}\left\{\frac{\left(\zeta^{-M}\cdot U(M,M+1-s,\zeta^{-1})\right)^w}{1-e^{\rho s}\zeta^{-M}\cdot U(M,M+1-s,\zeta^{-1})}\right\} (13)$$

Although $U(a,b,z)$ is implemented in standard software for mathematical calculations, we provide below an alternative expression for the Mellin transform in terms of the simpler upper incomplete gamma function. The following expression can be obtained by noticing that M is a positive integer and applying the binomial theorem:

$$\begin{aligned}\mathcal{M}_{g(\gamma)}(s) = \quad & e^{\frac{1}{\zeta}}\sum_{j=0}^{M-1}(-1)^{M-1-j} \\ & \times \zeta^{j+s-M}\frac{\Gamma(j+s,\zeta^{-1})}{\Gamma(M-j)\Gamma(j+1)} \qquad (14)\end{aligned}$$

where $\Gamma(s,z) = \int_z^\infty t^{s-1}e^{-t}\,\mathrm{d}t$ is the upper incomplete gamma function. Note that, for the SISO case, letting $M=1$ and $\zeta = \mathsf{snr}$ in (14) we obtain $\mathcal{M}_{g(\gamma)}(s) = e^{\frac{1}{\mathsf{snr}}}\cdot\mathsf{snr}^{s-1}\cdot\Gamma(s,\mathsf{snr}^{-1})$ which is the same expression reported in [8].

The above expressions allows us to obtain bounds on the delay violation probability for different system parameters without resorting to Monte Carlo simulations. However, due to the complexity of the kernel function, no closed-form solution for the minimum s can be found, and we must resort to numerical methods. In some asymptotic cases, we can have simpler expressions of the Mellin transform that make this process easier, as we will show later in Section V.

B. OSTBC

Orthogonal space-time block coding has been a very successful transmit diversity technique because it can achieve full diversity without CSI at the transmitter and need for joint decoding of multiple symbols. It is characterized by the number of independent symbols L_s transmitted over T time slots; the code rate is $R_c = L_s/T$. When the transmitter uses OSTBC with M transmit antennas, code parameter T, and the receiver performs MRC with N antennas, the equivalent SNR $\gamma = \frac{\mathsf{snr}}{M}\|\mathbf{H}\|_{\mathrm{F}}^2$ is gamma distributed with shape parameter MN and scale parameter $(\mathsf{snr}/M)^{-1}$ [24, Eq. 3.43]; here $\mathbf{H}$ denotes the MIMO channel matrix of $N \times M$ complex Gaussian entries. Particularizing (12) for the case of MISO OSTBC, we have the following result.

Result 1. *The Mellin transform of the service process of a MISO system employing OSTBC is given by*

$$\mathcal{M}_{g(\gamma)}^{\mathrm{OSTBC}}(s) = \left(\frac{\mathsf{snr}}{M}\right)^{-M}\cdot U\left(M, M+s, M/\mathsf{snr}\right). \qquad (15)$$

C. Antenna Selection

Antenna selection is a low-complexity, low-rate feedback diversity technique, in which the transmitter and/or the receiver select a subset of transmit/receive antennas for transmission/reception. It can be used in conjunction with other diversity techniques and can improve the performance of open-loop MIMO at the expense of very low amount of feedback. We consider here transmit antenna selection (TAS), in which the transmitter selects to transmit on the antenna (one of M) that maximizes the instantaneous SNR. The amount of CSI required to be fed back to the transmitter is $\lceil \log_2 M \rceil$ bits (index of best antenna), where $\lceil x \rceil$ denotes the smallest integer larger than x. The instantaneous SNR can be expressed as $\gamma_{\text{TAS}} = \mathsf{snr}\gamma_{\max}$, where $\gamma_{\max}$ is the largest channel gain, i.e. $\gamma_{\max} = \max\limits_{1 \le i \le M} |h_i|^2$. Since $h_i \sim \mathcal{CN}(0,1)$, we have that $|h_i|^2$ is exponentially distributed with unit mean and pdf $f_{|h_i|^2}(x) = e^{-x}$.

Result 2. *The Mellin transform for a MISO system employing TAS is given by*

$$\mathcal{M}^{\text{TAS}}_{g(\gamma)}(s) = M\zeta^{s-1} \sum_{k=0}^{M-1} \binom{M-1}{k} (-1)^k \times \frac{e^{\frac{k+1}{\zeta}}}{(k+1)^s} \Gamma\left(s, \frac{k+1}{\zeta}\right). \tag{16}$$

Proof: Suppose that $X_1, \ldots, X_n$ are n independent continuous variates, each with cdf $F(x)$ and pdf $f(x)$. The pdf of the r-th order statistic $X_{(r)}$, $r = 1, \ldots, n$ is given by [25]

$$f_{(r)}(x) = \frac{1}{B(r, n-r+1)} F^{r-1}(x)[1-F(x)]^{n-r} f(x). \tag{17}$$

Therefore, the pdf of $\gamma_{\max} = \gamma_{(M)}$ is given by $f_{\gamma_{\max}}(x) = M f_\gamma(x) F_\gamma^{M-1}(x)$. Since $\gamma \sim \text{Exp}(1)$ in the case of TAS with pdf $f_\gamma(x) = e^{-x}$, we have that

$$\begin{aligned} \mathcal{M}^{\text{TAS}}_{g(\gamma)}(s) &= M \int_0^\infty (1+\zeta x)^{s-1} f_\gamma(x) F_\gamma^{M-1}(x) \mathrm{d}x \\ &= M \int_0^\infty (1+\zeta x)^{s-1} e^{-x} (1-e^{-x})^{M-1} \mathrm{d}x. \end{aligned} \tag{18}$$

Applying binomial theorem and solving the resulting integral we finally obtain

$$\mathcal{M}^{\text{TAS}}_{g(\gamma)}(s) = M\zeta^{s-1} \sum_{k=0}^{M-1} \binom{M-1}{k} (-1)^k \times \frac{e^{\frac{k+1}{\zeta}}}{(k+1)^s} \Gamma\left(s, \frac{k+1}{\zeta}\right). \tag{19}$$

■

V. Delay Performance: Asymptotic Analysis

In the previous section, we have provided analytical expressions for the Mellin transform of the service process for various multi-antenna diversity techniques. The exact results are mainly given in terms of special functions and alternating series. To explore further the delay performance of MISO MRT, we derive in this section simplified expressions for various asymptotic regimes: low/high SNR and large M. Additionally, we obtain a general result for a Gaussian distributed service process; as we will show, the MISO MRT service process converges to this distribution as M grows large.

A. High SNR regime

We investigate here how statistical delay constraints affect the MISO performance at high SNR. We assume that this also implies large ζ, which is true as long as σ_{e}^2 does not increase[1] with the SNR.

Corollary 1. *In the high SNR regime, the Mellin transform of the service process scales as*

$$\mathcal{M}^{\text{H}}_{g(\gamma)}(s) = \begin{cases} \zeta^{s-1} \frac{\Gamma(s+M-1)}{\Gamma(M)} & s > 1-M \\ \zeta^{-M} \frac{\Gamma(1-s-M)}{\Gamma(1-s)} & s \le 1-M \\ \zeta^{-M} \frac{\log\zeta - \psi(M)}{\Gamma(M)} & s = 1-M \end{cases} \tag{20}$$

where $\psi(x)$ denotes the Digamma function [23, Sec. 6.3].

Proof: The three branches are obtained after direct application of the asymptotic properties of the $U(a,b,z)$ function listed in [23, Sec. 13.5]. The first branch can be also derived by considering the approximated service process $s_i \approx \log(\gamma_i)$, which gives that $\mathcal{M}^{\text{H}}_{g(\gamma)}(s) = \zeta^{s-1}\Gamma(s+M-1)$ for $s+M > 1$. ■

B. Low SNR regime

At low SNR, we have the following result:

Corollary 2. *In the low SNR regime, the Mellin transform of the service process is approximately given as*

$$\mathcal{M}^{\text{L}}_{g(\gamma)}(s) \approx (1-(s-1)\zeta)^{-M}, \qquad s < \zeta^{-1} - 1. \tag{21}$$

Proof: At low SNR, we use the first order Taylor series expansion $\log(1+x) \approx x$. In that case, the service process can be approximated as $s_i \approx \gamma_i \Rightarrow g(\gamma) \approx e^\gamma$, hence

$$\mathcal{M}^{\text{L}}_{g(\gamma)}(s) = \mathbb{E}\left[e^{\gamma(s-1)}\right] \approx (1-(s-1)\zeta)^{-M} \tag{22}$$

using the moment generating function (MGF) of a gamma random variable. ■

C. Large antenna regime

The distribution of the mutual information of a Rayleigh fading MIMO system is generally rather complicated, therefore approximations have been used in the literature. For example, in the large antenna regime ($M \to \infty$) and using the Central Limit Theorem (CLT), it can be shown that the distribution of the mutual information $\mathcal{I}$ converges to a Gaussian distribution; see for instance [26] and references therein. Using similar arguments here, we can obtain simpler expressions for the Mellin transform of the service process. In general, we can obtain results of the form

$$M^\alpha \frac{\mathcal{I} - \mu}{\sigma_M} \xrightarrow{d} \mathcal{N}(0,1) \tag{23}$$

[1] As a matter of fact, most frequently and in practice $\sigma_{\text{e}}^2 \propto 1/\mathsf{snr}$.

where convergence is in distribution, $\mu = \mathbb{E}(\mathcal{I})$, σ_M is a variance term, and α is a measure of the convergence speed (normally 0.5). This means that, for large M, an accurate approximation of the distribution is given by $\mathcal{I} \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma^2 = \sigma_M^2/M^\alpha$. The mean and the variance terms can be obtained in closed form from [27]. Note that, for brevity, throughout this section we will use the natural logarithm, and thus all rates are in nats.

Result 3. *The Mellin transform of a service process with rate following a Gaussian distribution with mean μ and variance σ^2 is given by*

$$\mathcal{M}^{\text{as}}_{g(\gamma)}(s) = e^{(s-1)\mu+(s-1)^2\frac{\sigma^2}{2}}. \tag{24}$$

Result 4. *For the MISO MRT case, as the number of antennas grows large, we have*

$$\lim_{M\to\infty} \mathcal{M}^{\text{as}}_{g(\gamma)}(s) \to (1+\zeta M)^{s-1}. \tag{25}$$

Proof: The mutual information can be written as $\mathcal{I} = \log(1+\gamma)$. Rewriting (24) and applying Jensen's inequality

$$\begin{aligned}
\mathcal{M}^{\text{as}}_{g(\gamma)}(s) \quad &\approx \quad e^{(s-1)\,\mathbb{E}[\log(1+\gamma)]} \cdot e^{(s-1)^2\frac{\sigma^2}{2}} && (26)\\
&\leq \quad e^{(s-1)\log(1+\mathbb{E}[\gamma])} \cdot e^{(s-1)^2\frac{\sigma^2}{2}} && (27)\\
&= \quad (1+\zeta M)^{s-1} \cdot e^{(s-1)^2\frac{\sigma^2}{2}} && (28)\\
&\overset{M\to\infty}{\approx} \quad (1+\zeta M)^{s-1} && (29)
\end{aligned}$$

where the last equality follows from the fact that $\lim\limits_{M\to\infty} \sigma^2 = 0$ [26], [27]. ∎

Note that using standard tools (e.g. continuous mapping theorem and Chebyshev inequality), it can be shown that the bound is asymptotically tight. Furthermore, the asymptotic convergence can be obtained without resorting to the Gaussian approximation by showing that convergence in distribution implies convergence in $\mathcal{M}_{g(\gamma)}(s)$. Let $y_1, y_2, \ldots$ be a sequence of positive random variables that converges in distribution to a positive random variable y. For $s > 0$, we have $\lim\limits_{M\to\infty} \mathcal{M}_{y_i}(s) = \mathcal{M}_y(s)$. By Lebesgue's dominated convergence theorem and $\frac{\|\mathbf{h}\|^2}{\mathbb{E}[\|\mathbf{h}\|^2]} \xrightarrow{\mathcal{P}} 1$, we have that $\lim\limits_{M\to\infty} \mathcal{M}_{g(\gamma)}(s) = (1+\rho M)^{s-1}$.

Interestingly, we observe that for large M, $\mathcal{M}^{\text{as}}_{g(\gamma)}(s) \sim (\zeta M)^{s-1}$, which is related to the so-called channel hardening effect, i.e. the channel behaves equivalently to an AWGN channel with SNR ζM. In the low SNR regime, the number of transmit antennas affects linearly the service process, while at high SNR, the Mellin transform of the service process grows superlinearly with M (for $s > 1$).

The approximation $s_i \sim \mathcal{N}(\mu, \sigma^2)$ allows us to simplify the delay violation probability expression (13), however its relevance and applicability goes beyond, as it allows analyzing the delay violation probability of any system whose service rate can be approximated by a Gaussian random variable. Additionally, it can provide very simple expressions for the effective capacity.

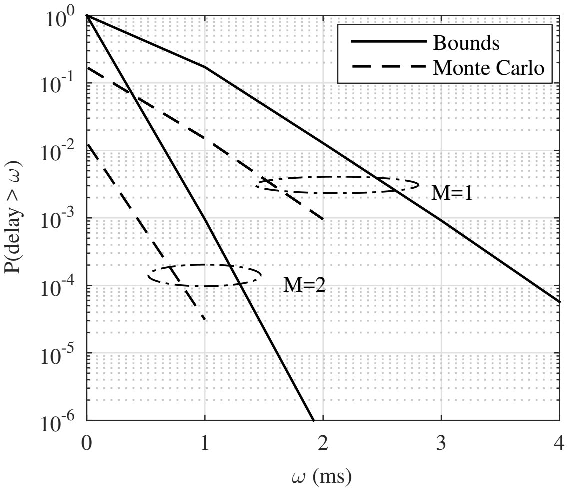

Figure 1. Delay violation probability and associated bounds as a function of the target dealy, $\rho = 24$ kbps and $\mathsf{snr} = -2$ dB.

VI. Numerical Results

In this section, we provide numerical evaluation of the performance of MISO communication systems based on the above analysis. Unless otherwise stated, the duration of a slot is set to $T = 1$ ms, the overhead is disregarded ($L_m \to 0$), and the blocklength is assumed to be $L = 168$; consequently $B = L/\log 2 \neq 1$, and we reincorporate this parameter into the equations.

We start by validating our analysis with Monte Carlo simulations. In Figure 1, we compare the delay violation probability and its bound with $\rho = 24$ kbps and $\mathsf{snr} = -2$ dB. We corroborate that the bounds follow the trend of the original curve, and we point out that the maximum difference in the x-axis seems to be of about 1 ms.

Figure 2 plots the violation bound for MISO MRT as a function of the target delay w with $\rho = 24$ kbps and $\mathsf{snr} = 5$ dB. It shows the effect of varying the number of antennas and the accuracy of the CSI. We observe the strong decrease of the delay violation probability when increasing the number of antennas: with perfect CSI, the probability of exceeding 1 ms delay roughly decreases by three orders of magnitude when adding an extra antenna.

In Figure 3, we compare the delay performance of MISO MRT with OSTBC and TAS. We can see that MRT generally performs better when the quality of the CSI is good: above a certain value of σ_e^2, TAS and OSTBC outperform MRT. The values at which this change takes place seem to be dependent on the number of antennas.

In Figure 4, we investigate further the effect of adding antennas, and compare it to that of increasing the power. For a target delay of 1 ms at 0 dB, we can see that going from three to four antennas seems to have only slightly less impact than doubling the power; at 5 dB, however, this is not the case anymore: 3 dB of extra power decrease the violation probability by one order of magnitude, but adding one antenna decreases it by two orders of magnitude.

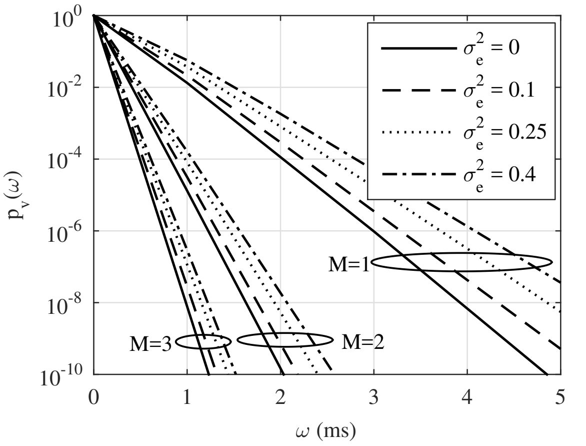

Figure 2. Delay violation probability bound as a function of the target delay, $\rho = 24$ kbps and $\mathsf{snr} = 5$ dB.

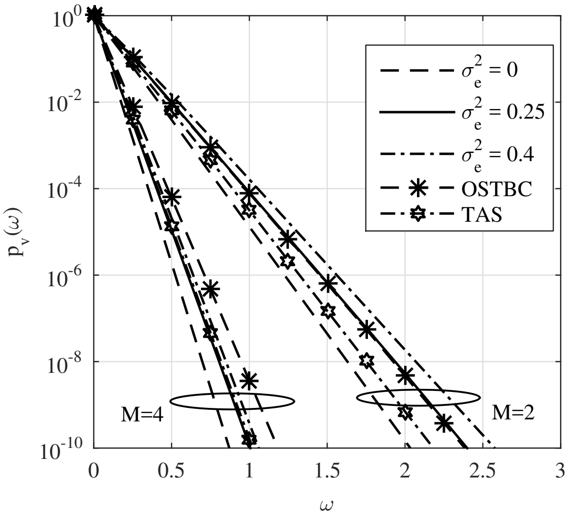

Figure 3. Delay violation probability bound as a function of the target delay for different diversity techniques, $\rho = 24$ kbps and $\mathsf{snr} = 5$ dB.

As explained in Section IV, it is important to have simple expressions for the kernel when possible. In Figures 5 and Figure 6, we illustrate the accuracy of the Gaussian approximation for $M = 3$ and $M = 10$; as expected, the error is large for the former and negligible for the latter. This justifies the use of the much simpler expression (24) whenever M is relatively large.

In Figure 7 we test the accuracy of the high and low SNR approximations derived in Section V. We can see that the high SNR approximation becomes asymptotically tight as the SNR increases, and that, remarkably, the low SNR approximation is reasonably accurate for most SNR values; this makes the low SNR approximation particularly interesting given its simple expression.

VII. Conclusions

In this work, we characterized the delay performance of MISO diversity communications under statistical delay con-

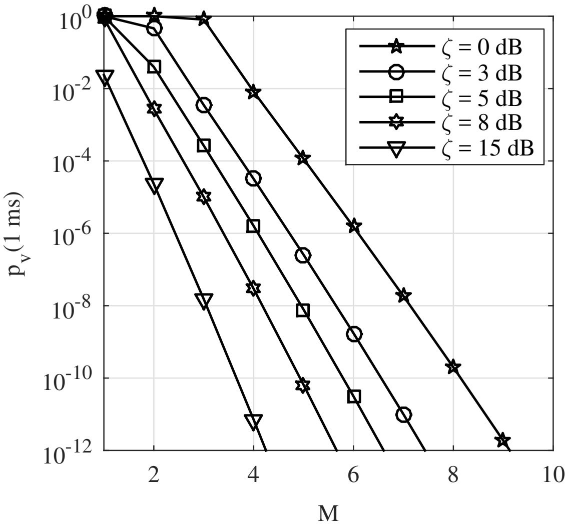

Figure 4. Bound on the probability of exceeding 1 ms delay as a function of the number of antennas,$\rho = 256$ kbps.

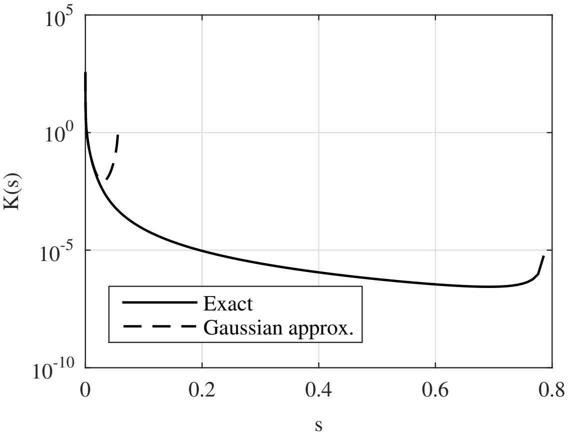

Figure 5. Mellin transform (left) and kernel (right) as a function of s, $M = 3$, $\rho = 20$ kbps, $\zeta = 0$ dB, $w = 1$.

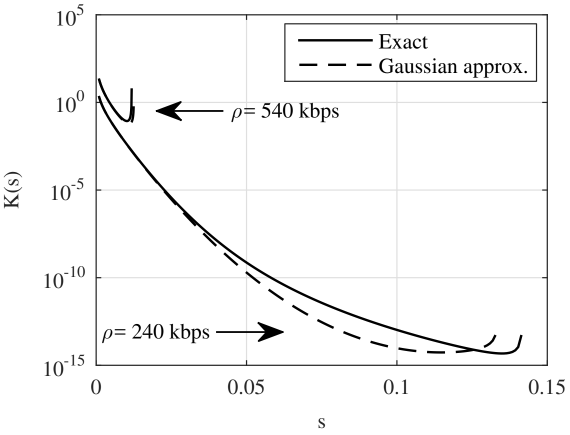

Figure 6. Mellin transform (left) and kernel (right) as a function of s, $M = 10$, $\zeta = 0$ dB, $w = 1$.

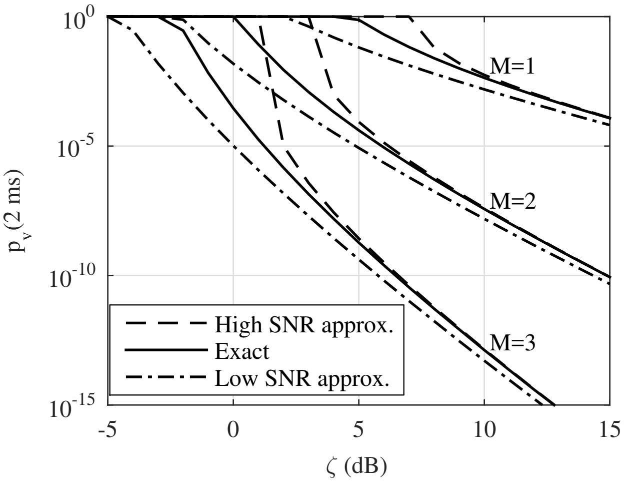

Figure 7. Bound on the probability of exceeding 2 ms delay, $\rho = 200$ kbps.

straints. Using stochastic networks calculus, we derived a statistical characterization of the service process in term of Mellin transform for multi-antenna fading channels and provided probabilistic delay bounds. We showed how the number of transmit antennas and transmit SNR may affect the delay performance. MISO MRT is shown to reduce the delay violation probability as compared to single-antenna transmissions even with imperfect CSI. Nevertheless, as channel imperfections increase, other diversity-techniques, such as OSTBC and antenna selection, perform better than MRT in terms of delay violation probability. Future work could consider the effect of imperfect CSI at the receiver and limited feedback in FDD MIMO systems. Further extensions of this framework may include the analysis of MIMO spatial multiplexing, MIMO channels with co-channel interference, and multiuser MIMO systems.

References

[1] 3GPP, "TR 38.913 (V14.2.0):study on scenarios and requirements for next generation access technologies," Tech. Rep., Mar. 2017.

[2] C.-S. Chang, *Performance Guarantees in Communication Networks*. London, UK: Springer-Verlag, 2000.

[3] Y. Jiang and Y. Liu, *Stochastic Network Calculus*. London, UK: Springer-Verlag, 2008.

[4] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. 14th IEEE Inter.Workshop on Quality of Service*, June 2006, pp. 261–270.

[5] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, First quarter 2015.

[6] I. H. Hou, V. Borkar, and P. R. Kumar, "A theory of QoS for wireless," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 486–494.

[7] 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, July 2003.

[8] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "A (min, x) network calculus for multi-hop fading channels," in *Proc. IEEE INFOCOM*, April 2013, pp. 1833–1841.

[9] S. Zhou, K. Zhang, Z. Niu, and Y. Yang, "Queuing analysis on MIMO systems with adaptive modulation and coding," in *Proc. IEEE Inter. Conf. on Commun. (ICC)*, May 2008, pp. 3400–3405.

[10] S. Kittipiyakul and T. Javidi, "Optimal operating point for MIMO multiple access channel with bursty traffic," *IEEE Trans. Wireless Commun.*, vol. 6, no. 12, pp. 4464–4474, Dec. 2007.

[11] S. Kittipiyakul, P. Elia, and T. Javidi, "High-SNR analysis of outage-limited communications with bursty and delay-limited information," *IEEE Trans. Inf. Theory*, vol. 55, no. 2, pp. 746–763, Feb. 2009.

[12] J. Chen and V. K. N. Lau, "Large deviation delay analysis of queue-aware multi-user MIMO systems with two-timescale mobile-driven feedback," *IEEE Trans. Signal Process.*, vol. 61, no. 16, pp. 4067–4076, Aug. 2013.

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

[14] M. Matthaiou, G. C. Alexandropoulos, H. Q. Ngo, and E. G. Larsson, "Analytic framework for the effective rate of MISO fading channels," *IEEE Trans. Commun.*, vol. 60, no. 6, pp. 1741–1751, Jun. 2012.

[15] M. You, H. Sun, J. Jiang, and J. Zhang, "Unified framework for the effective rate analysis of wireless communication systems over MISO fading channels," *IEEE Trans. Commun.*, vol. 65, no. 4, pp. 1775–1785, Apr. 2017.

[16] K. Mahmood, A. Rizk, and Y. Jiang, "On the flow-level delay of a spatial multiplexing MIMO wireless channel," in *Proc. IEEE Inter. Conf. on Commun. (ICC)*, June 2011, pp. 1–6.

[17] K. Mahmood, M. Vehkapera, and Y. Jiang, "Delay constrained throughput analysis of a correlated MIMO wireless channel," in *Proc. 20th Inter. Conf. on Comp. Commun. and Netw. (ICCCN)*, July 2011, pp. 1–7.

[18] S. Schiessl, F. Naghibi, H. Al-Zubaidy, M. Fidler, and J. Gross, "On the delay performance of interference channels," in *2016 IFIP Networking Conference (IFIP Networking) and Workshops*, May 2016, pp. 216–224.

[19] T. K. Y. Lo, "Maximum ratio transmission," *IEEE Trans. Commun.*, vol. 47, no. 10, pp. 1458–1461, Oct. 1999.

[20] Y. Chen and C. Tellambura, "Performance analysis of maximum ratio transmission with imperfect channel estimation," *IEEE Commun. Lett.*, vol. 9, no. 4, pp. 322–324, Apr. 2005.

[21] G. Caire, N. Jindal, M. Kobayashi, and N. Ravindran, "Multiuser MIMO achievable rates with downlink training and channel state feedback," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2845–2866, Jun. 2010.

[22] S. Schiessl, H. Al-Zubaidy, M. Skoglund, and J. Gross, "Analysis of wireless communications with finite blocklength and imperfect channel knowledge," 2016. [Online]. Available: http://arxiv.org/abs/1608.08445

[23] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover, 1964.

[24] A. Paulraj, R. Nabar, and D. Gore, *Introduction to Space-Time Wireless Communications*, 1st ed. New York, NY, USA: Cambridge University Press, 2003.

[25] H. A. David and H. N. Nagaraja, *Order Statistics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2005.

[26] B. M. Hochwald, T. L. Marzetta, and V. Tarokh, "Multiple-antenna channel hardening and its implications for rate feedback and scheduling," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 1893–1909, Sep. 2004.

[27] M. R. McKay, P. J. Smith, H. A. Suraweera, and I. B. Collings, "On the mutual information distribution of OFDM-based spatial multiplexing: Exact variance and outage approximation," *IEEE Trans. Inf. Theory*, vol. 54, no. 7, pp. 3260–3278, July 2008.