Delay Performance of the Multiuser MISO Downlink under Imperfect CSI and Finite Length Coding

Sebastian Schiessl, *Student Member, IEEE*, James Gross, *Senior Member, IEEE*, Mikael Skoglund, *Fellow, IEEE* and Giuseppe Caire, *Fellow, IEEE*

***Abstract*—We use stochastic network calculus to investigate the delay performance of a multiuser MISO system with zero-forcing beamforming. First, we consider ideal assumptions with long codewords and perfect CSI at the transmitter, where we observe a strong channel hardening effect that results in very high reliability with respect to the maximum delay of the application. We then study the system under more realistic assumptions with imperfect CSI and finite blocklength channel coding. These effects lead to interference and to transmission errors, and we derive closed-form approximations for the resulting error probability. Compared to the ideal case, imperfect CSI and finite length coding cause massive degradations in the average transmission rate. Surprisingly, the system nevertheless maintains the same qualitative behavior as in the ideal case: as long as the average transmission rate is higher than the arrival rate, the system can still achieve very high reliability with respect to the maximum delay.**

***Index Terms*—Multiple-input multiple-output (MIMO), multiuser diversity, zero-forcing beamforming (ZFBF), stochastic network calculus, imperfect CSI, finite blocklength regime**

I. Introduction

Due to the random nature of the wireless channel, it is notoriously difficult to design wireless communication systems for applications that require both very low latency and very high reliability. For example, applications in factory automation often require latencies of just a few milliseconds and reliability (with respect to this deadline) of $1 - 10^{-8}$ and above [1], [2], which is difficult to achieve in a wireless channel that is subject to fading and noise. In order to increase the reliability of the system, one can equip the transmitter with multiple antennas, which is known as multiple-input single-output (MISO). When transmitting only to a single receiver, multiple transmit antennas increase the diversity of the system and reduce the variations in the signal strength in fading channels, leading to lower error probabilities. When the transmitter has channel state information (CSI), it can use beamforming to send the signal in the direction of the user's channel, resulting not only in a diversity gain but also a power gain [3], making the system even more resilient against errors. On the other hand, a transmitter with M antennas can also serve $K \leq M$ different users at the same time, i.e., achieve a multiplexing gain. Serving multiple users at once means that each user can be scheduled more often, which can reduce the delay. An often used transmission strategy for the multiuser MISO downlink is zero-forcing beamforming (ZFBF), which ensures that the signal intended for each user does not create interference at the other (unintended) receivers. Nevertheless, increasing K reduces the beamforming gain, i.e., reduces the data rates of the individual transmissions. The trade-off between the multiplexing gain and the beamforming gain was studied in [4] with respect to the ergodic capacity.

However, the ergodic capacity does not accurately reflect the delay performance of the system. When there is a certain probability that the data rate is small, or when transmission errors occur, the transmitter must keep the data in a buffer so that it can be transmitted in subsequent time slots. This buffering causes a random delay. The queueing delay may sometimes grow until the deadline of the application is violated. For applications that require very high reliability, the communication system must be designed such that the probability of a deadline violation is minimized. For example, low beamforming gains should be avoided, as low beamforming gains increase the probability that the individual data rates are small. In [5], we studied the trade-off between the multiplexing and beamforming gains with respect to the queueing performance.

Fortunately, as the number of antennas M grows, the transmitter can schedule $K = aM$ users with $a < 1$ and thus benefit from a linear increase in both multiplexing gain and beamforming gain. As the beamforming gain increases linearly in M, the system will experience only small variations (relative to the average) in the achievable transmission rate. This effect is known as channel hardening [6]. In this case, we suspect that only the *average* of the transmission rate will determine the queueing performance: When the *average* transmission rate is higher than the incoming data rate at the transmit buffer, then long queueing delays should be very unlikely (occur with almost zero probability); otherwise, long queueing delays have probability one.

In this context, it is of critical importance that the performance of the physical layer transmissions is modeled accurately. Specifically, when the duration of each time slot

This work was funded in part by the Swedish Foundation for Strategic Research and the Swedish Research Council.

The work of Giuseppe Caire was partially funded by a Professorship Grant of the Alexander von Humboldt Foundation.

S. Schiessl, J. Gross and M. Skoglund are with the Division of Information Science and Engineering, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: schiessl@kth.se, james.gross@ee.kth.se, skoglund@kth.se).

G. Caire is with the Electrical Engineering and Computer Science Department, Technische Universität Berlin, 10587 Berlin, Germany (e-mail: caire@tu-berlin.de).

is short, channel estimation causes a significant overhead, and the transmitter can only acquire an imperfect estimate of the channel state. First of all, this means that zero-forcing beamforming cannot eliminate the interference. Second, the transmitter does not know the actual signal-to-interference-and-noise ratio (SINR) of the channel, and therefore, outages can occur when the actual channel capacity is below the selected transmission rate. The transmitter must then find a careful balance between the outage probability and the rate with respect to the queueing delay. In addition to imperfect CSI, we note that the transmitter cannot achieve error-free transmissions at the channel capacity when the blocklength of the channel code is finite. All of these effects must be taken into account when considering systems for ultra low latency communications.

A. Related Work

This paper builds on results from several research areas. On the physical layer, we consider the multiuser MISO downlink with imperfect CSI, as well as finite blocklength coding. On top of these physical-layer aspects, we investigate the queueing delay of the data at the link layer.

1) Multiuser-MISO and Imperfect CSI: Linear ZFBF precoding in the multiuser MISO downlink has been studied by several authors. Although ZFBF is not capacity-achieving, Yoo and Goldsmith [7] showed that when the total number of users K_{tot} is much larger than the number of antennas M, then ZFBF can achieve the same asymptotic performance as the capacity-achieving scheme based on dirty-paper coding (DPC) [8]. These results hold only if the transmitter has channel state information (CSI) of all users. While the authors investigated in [9] also the impact of quantized channel state feedback, it would still be impossible to receive feedback from e.g. 100 or more users when the duration of each time slot is short. The random beamforming scheme by Sharif and Hassibi [10] reduces the overhead from collecting CSI by transmitting a training sequence along a set of randomly created beams. The users then send the index and the signal-to-interference-and-noise ratio (SINR) of the beam with the highest SINR. However, in this scheme, some of the users may not be scheduled for a long period because the scheduling decision is based on the random SINR. Furthermore, even though the overhead from collecting CSI is reduced, collecting feedback from many users may still be infeasible when considering scenarios with very low latency. In fact, Ravindran and Jindal [11] found that, given a fixed budget for the total overhead, it is better to collect accurate channel estimates from only a small number of users than to collect inaccurate CSI from many users, i.e., they found that accurate CSI is more important than multiuser diversity. When the transmitter has only CSI for the users that are scheduled, Zhang et al. [12] studied whether the transmitter should send data to a single user or to $K = M$ users. The same authors studied in [13] the more general case $K = aM$ and also considered imperfect CSI. This is very close to our work, but the authors studied only the ergodic sum rate and assumed that an additional perfect feedback link provided the exact value of the channel capacity to the transmitter, such that outages did not occur. Similarly, the authors in [14] studied the ergodic capacity of a multiuser MISO system under imperfect CSI. In the ergodic case, rate adaptation is not necessary, and the performance loss is due to imperfect beamforming and due to additional noise terms at the receiver. Similarly, some other well-known works on imperfect CSI, e.g., [15]–[17] consider only the ergodic case, where the effects of channel estimation errors can be averaged over many fading instances. In contrast, we want to study the queueing performance of a system where outages occur because the transmitter must adapt the rate to an imperfect estimate of the channel, and where the receiver must decode the signal in the same time slot (as opposed to decoding over an infinite time horizon).

2) Finite Blocklength Coding: Well-known results on channel coding at finite blocklength were derived by Polyanskiy et al. [18], who showed that the loss in the achievable data rate due to finite blocklength can be approximated by a simple second-order expression. Yang et al. extended these results to fading channels [19]. These works generally assume non-Gaussian codebooks in Gaussian noise. Therefore, the results cannot be applied when the transmissions create mutual interference. Scarlett et al. [20] studied the performance of Gaussian codebooks under non-Gaussian interference. We are not aware of any results combining finite blocklength coding with multi-user MISO.

3) Delay Analysis: The queueing delay, which occurs due to transmission errors and low transmission rates, can for example be analyzed through the frameworks of stochastic network calculus [21], [22], effective capacity [23] or using the martingale technique [24]. With respect to single-user multi-antenna systems, the queueing delay was studied in several works [25]–[30]. Sending the same data to different users (multicasting) was studied in [31]. The multi-user case with line-of-sight channels was studied by Li et al. [32]. Optimal power allocation with respect to the delay performance was studied in [33]. However, all of these works assumed CSI to be perfect. In [34], we studied the delay in single-user single-antenna systems with finite length coding and imperfect CSI.

B. Contributions

In this work, we study the multiuser MISO downlink both under ideal assumptions and under more realistic assumptions with imperfect CSI and finite blocklength coding. Specifically, we make the following contributions:

- For the ideal scenario with perfect CSI and long codewords, we use a previous result [5], [29], [30] to study the effect of channel hardening. We investigate how many antennas are necessary to achieve extremely high reliability with almost zero violations of the deadline.
- For the realistic scenario with imperfect CSI, we derive two closed-form approximations, corresponding to approximate lower and upper bounds on the conditional outage probability.
- For the realistic scenario with imperfect CSI and finite blocklength of the channel code, we derive a closed-form approximation for the conditional error probability.

- We verify by extensive Monte Carlo simulations that the derived expressions are in many cases lower or upper bounds on the conditional outage or error probability.
- Our numerical analysis shows that imperfect CSI leads to substantial losses in the average transmission rate. Furthermore, we find that the additional loss due to finite blocklength effects is only relevant when the CSI becomes nearly perfect. Despite the massive performance loss compared to the ideal case, a system that shows a strong channel hardening effect in the ideal scenario maintains this behavior qualitatively in the realistic case. In other words, multiuser MISO systems can achieve very high reliability even under non-ideal assumptions.

This paper is structured as follows: In Sec. II, we present the system model for the ideal scenario with perfect CSI and long codewords. In Sec. III, we present a short summary of the delay analysis through network calculus, and perform the delay analysis for the ideal scenario. In Sec. IV, we show how imperfect CSI and finite blocklength effects can be modeled and analyzed. Numerical evaluations are presented in Sec. V, before we finally conclude the paper in Sec. VI.

II. System Model

We consider a system where data is sent from a transmitter with M antennas to K_{tot} single-antenna users, with $K_{\mathrm{tot}} \gg M$. We assume time-slotted transmissions. In each time slot, the transmitter can select only a subset $\mathcal{K} \subset \{1, \ldots, K_{\mathrm{tot}}\}$ of users, with the number of scheduled users denoted as $K \triangleq |\mathcal{K}| \leq M$. As we consider scenarios where the duration of each time slot is short, the overhead from collecting channel state information (CSI) for all K_{tot} users would be overwhelming. Thus, the transmitter cannot select the user set $\mathcal{K}$ based on the instantaneous CSI. Instead, the transmitter selects in each time slot a set of users $\mathcal{K}$ and then collects channel state information only for those users, similar to [13]. In this section, we describe a basic model with ideal assumptions based on [5], i.e., we assume that the transmitter has perfect CSI for the scheduled users and that data can be transmitted at a rate equal to the channel capacity without errors. The more realistic scenario with imperfect CSI and finite blocklength of the channel code will be modeled and analyzed in Sec. IV.

First, we describe in Sec. II-A the data transmission from a physical layer perspective. In Sec. II-B, we discuss user scheduling. In Sec. II-C, the system is described from a queueing perspective. Finally, we present the problem statement in Sec. II-D.

A. Physical Layer Model

The received signal $\mathbf{y} \in \mathbb{C}^{K\times 1}$ at the K scheduled users can be described as

$$\mathbf{y} = \mathbf{H}_{\mathrm{PL}}\mathbf{H}^{\mathsf{H}}\mathbf{x} + \mathbf{z}\,. \tag{1}$$

The path loss matrix $\mathbf{H}_{\mathrm{PL}} = \mathrm{diag}(\beta_1, \ldots, \beta_K)$ describes the path loss factors between the transmitter and the K users. As the received signal y_k at user k depends only on the path loss β_k of user k, the path loss can be taken into account by a linear scaling of the received signal power. We thus assume without loss of generality that all $\beta_k = 1$. For the channel matrix $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{C}^{M\times K}$, 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 $\mathbf{H}$ remains constant for the duration of one time slot, consisting of n_{d} channel uses, and changes to an independent realization in the next time slot (note that the set $\mathcal{K}$ of scheduled users also changes). The input signal is denoted as $\mathbf{x} \in \mathbb{C}^{M\times 1}$ and must satisfy a short-term power constraint $\mathrm{tr}\left(\mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{H}}\right]\right) \leq P_\Sigma$ for each realization of $\mathbf{H}$. The noise $\mathbf{z} \in \mathbb{C}^{K\times 1}$ has i.i.d. components $\mathcal{CN}(0,1)$.

The transmitter encodes the data for the K scheduled users into code symbols $\mathbf{x} \in \mathbb{C}^{M\times 1}$ (one symbol per antenna). In order to obtain $\mathbf{x}$, the transmitter can encode the data of the K users individually into symbols $\mathbf{s} \in \mathbb{C}^{K\times 1}$ and apply a precoding strategy to obtain $\mathbf{x}$ from $\mathbf{s}$. We focus in this work on Zero-Forcing Beamforming (ZFBF), which is a linear strategy that completely eliminates the interference of the signals at the other receivers. In this case, the input signal vector $\mathbf{x}$ is given by [5], [8]

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

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$ is the precoding matrix and $\mathbf{P} = \mathrm{diag}(\rho_1, \ldots, \rho_K)$ is the power allocation matrix. The vector $\mathbf{s}$ denotes the $K \times 1$ vector of (independently) coded Gaussian symbols for the K scheduled users. When the transmitter perfectly knows the channel matrix $\mathbf{H}$, the ZFBF precoder is given as [5], [8]

$$\mathbf{V} = \mathbf{H}^{\dagger}\mathbf{\Xi}^{1/2} \tag{3}$$

where $\mathbf{H}^{\dagger} = \mathbf{H}(\mathbf{H}^{\mathsf{H}}\mathbf{H})^{-1}$ is the Moore-Penrose pseudo-inverse of $\mathbf{H}^{\mathsf{H}}$ and $\mathbf{\Xi} = \mathrm{diag}\,(\xi_1, \ldots, \xi_K)$ is the normalization matrix such that the columns of $\mathbf{V}$ have unit-2 norm. The variables ξ_k are central chi-square distributed (scaled by a factor $1/2$) with $2m$ degrees of freedom, where $m = M - K + 1$. Their PDF is given by [8, Lemma 4]

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

We assume for now that the blocklength n_{d} of the channel code is sufficiently long, so that the system can achieve error-free transmission ($\varepsilon = 0$) to user k at a rate [5], [8]

$$R_k = \log_2(1 + \rho_k|\mathbf{h}_k\mathbf{v}_k|^2) = \log_2(1 + \rho_k\xi_k)\,, \tag{5}$$

which changes along with $\mathbf{H}$ from time slot to time slot.

B. Scheduling

In each time slot, the transmitter can schedule only a subset $\mathcal{K}$ of users. To make sure that each user is scheduled regularly, we consider superframes of length T slots, and we require that each user is scheduled exactly once within a superframe. The average number of scheduled users per slot is given as $\overline{K} = K_{\mathrm{tot}}/T$. To simplify notations and discussions, the analysis in Sec. III and IV considers only the case where $\overline{K}$ is integer and the transmitter schedules a constant number of $K = \overline{K}$ users in each time slot. The analysis of non-integer $\overline{K}$ is discussed

in Appendix A. Furthermore, in case the transmitter schedules different sets of users with different K for each set, one can simply repeat the analysis for the different values of K.

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, with α_k denoting the constant number of bits that arrive at the queue of user k in each time slot. In case the transmissions are always error-free, the service rate offered by the wireless system in each time slot to user k is given as $S_k(t) = n_{\mathrm{d}} R_k(t)$ when k is among the scheduled users, or $S_k(t) = 0$ when user k is not scheduled. In case transmission errors occur with probability $\varepsilon > 0$, the users indicate through a feedback bit[1] whether the transmission was successful. When an error occurred, the data remain in the queue for transmission in subsequent time slots. Thus, the service for a scheduled user k is $S_k(t) = n_{\mathrm{d}} R_k(t) Z$, where $Z \sim \mathrm{Bernoulli}(1-\varepsilon)$. 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 [22], [34]:

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

The delay $W_k(t)$ is random. The reliability of a communication system with respect to the deadline w of the application can be described by the probability that the random delay $W_k(t)$ of the data for user k exceeds the target delay w at any time t:

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

We note here that for the considered system, the delay violation probability $p_{\mathrm{v},k}(w)$ cannot be analyzed through closed-form expressions. We thus follow our previous works [5], [34] and use stochastic network calculus [21], [22] to obtain analytical bounds on $p_{\mathrm{v},k}(w)$.

D. Problem Statement

In the first part of this work, we consider a system with perfect CSI and long blocklength of the channel code. In this case, error-free transmissions at the channel capacity can be achieved. Based on previous results [5], we study optimal number of scheduled users $\overline{K}$, i.e., the optimal trade-off between the multiplexing gain and the beamforming gain with respect to the delay performance. Furthermore, we investigate the effect of channel hardening: when the number of antennas M grows, the data rate of the wireless channel becomes nearly constant. We therefore expect that long queueing delays occur with almost zero probability when the average transmission rate is above the arrival rate. Naturally, when the average transmission rate is below the arrival rate, the delay grows to infinity and the delay violation probability is one. We will investigate how many antennas are necessary to observe such a zero/one behavior in practice.

In the second part of this work, we consider a more realistic scenario, where the transmitter must first estimate the channel before the transmission starts, and where the blocklength of the channel code is finite. These effects may have a significant impact on the system performance in the realistic case. Most importantly, scheduling a larger number of users $\overline{K}$ will increase the interference and also result in a larger overhead for channel estimation. We therefore expect that the optimal number of scheduled users $\overline{K}$ will decrease. However, it is not clear whether these effects will just lead to a change in the optimal value of $\overline{K}$ and to a *quantitative* loss in the overall performance, or whether these effects lead to a *qualitative* change in the system performance. Specifically, we want to find out whether a realistic system maintains the approximate zero/one behavior with respect to the delay distribution, i.e., whether the system still shows extremely high reliability whenever the average transmission rate is above the arrival rate.

III. Analysis – Ideal Case

In this section, we follow [5] and outline the analytical approach to determine $p_{\mathrm{v},k}(w)$. Specifically, in Sec. III-A, we present a summary of the delay analysis through stochastic network calculus in a transform domain [22]. In Sec. III-B we show how these results can be used when the users are only scheduled once per superframe. In Sec. III-C, we present how the stochastic network calculus bounds for the ideal case can be obtained analytically using previous results [5], [29], [30]. Given a fixed scheduling scheme with K scheduled users per time slot, the delay performance of user k depends only on the channel characteristics and arrival rate of user k. We drop the subscript k to shorten the notation.

A. Stochastic Network Calculus (SNC)

This section closely follows our previous works [5], [34] and provides a summary of stochastic network calculus [21], [22]. 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 [22] and describe these processes in the exponential domain, also referred to as *SNR domain*. The arrival and service processes in the bit domain, $A(t)$ and $S(t)$, are converted to the SNR domain (denoted by calligraphic letters) as

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

[1] As in related work, this feedback bit is assumed instantaneous and error-free. A single bit can be sent at negligible error rate even with little resources.

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. Then, an upper bound on the delay violation probability $p_{\rm v}(w)$ can 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 [22]

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

for a parameter $\theta \in \mathbb{R}$. For the analysis, we choose $\theta > 0$ and define the kernel [22], [35]

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

Under the condition $\mathcal{M}_{\mathcal{A}}(1+\theta)\mathcal{M}_{\mathcal{S}}(1-\theta) < 1$, the kernel converges:

$$\mathbb{K}(\theta, w) = \frac{\mathcal{M}_{\mathcal{S}}(1-\theta)^{w}}{1 - \mathcal{M}_{\mathcal{A}}(1+\theta)\mathcal{M}_{\mathcal{S}}(1-\theta)} . \tag{13}$$

For any parameter $\theta > 0$, the kernel $\mathbb{K}(\theta, w)$ provides an upper bound on the delay violation probability $p_{\rm v}(w)$ [22], [35]. This holds also in steady-state, i.e., in the limit $t \to \infty$. The tightest upper bound can be found by iterating over the parameter $\theta > 0$:

$$p_{\rm v}(w) \leq \inf_{\theta>0} \left\{\mathbb{K}(\theta, w)\right\} . \tag{14}$$

B. SNC and Scheduling

The delay analysis through stochastic network calculus as shown in Sec. III-A cannot be applied directly because $S(t)$ 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 of block-fading and 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 $\mathcal{M}_{\mathcal{A}^{(T)}}(\theta) = e^{\alpha T(\theta-1)}$.

In case w/T is an integer, it follows directly from (14) that:

$$p_{\rm v}(w) \leq \inf_{\theta>0} \left\{\mathbb{K}^{(T)}\left(\theta, \frac{w}{T}\right)\right\} . \tag{15}$$

When the condition $\mathcal{M}_{\mathcal{A}^{(T)}}(1+\theta)\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta) < 1$ holds, the kernel $\mathbb{K}^{(T)}(\cdot)$ converges to

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

The bound depends on the Mellin transform of the service $\mathcal{S}^{(T)}$ per superframe in the SNR-domain, which is connected to the bit-domain service process as $\mathcal{S}^{(T)} = e^{S^{(T)}}$. In the bit-domain, each user experiences a service of $S^{(T)} = n_{\rm d} R Z$ bits per superframe, where $Z = 0$ when a transmission error occurs and $Z = 1$ otherwise. Therefore:

$$\mathcal{M}_{\mathcal{S}^{(T)}}(\theta) = \mathbb{E}\left[\left(e^{n_{\rm d} R Z}\right)^{\theta-1}\right] . \tag{17}$$

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 $p_2 = \frac{\mod(w,T)}{T}$, and $p_1 = 1 - p_2$. Thus, the overall bound on the delay violation probability is given by [5]

$$p_{\rm v}(w) \leq p_1 \mathbb{K}^{(T)}\left(\theta, \left\lceil \frac{w}{T} \right\rceil\right) + p_2 \mathbb{K}^{(T)}\left(\theta, \left\lfloor \frac{w}{T} \right\rfloor\right) . \tag{18}$$

C. Delay Analysis – Ideal Case

In the ideal case with perfect CSI and long codewords, the rate is given as $R = \log_2(1+\rho\xi)$, and no errors occur ($Z = 1$ with prob. 1). We assume that the signal power ρ remains constant. In this case, we can obtain the Mellin transform of $\mathcal{S}^{(T)}$ in closed form. The variable ξ is a central χ^2 variable with $2m$ degrees of freedom, where $m = M - K + 1$, as outlined in Sec. II-A. The Mellin transform of the service process in the SNR domain is then given as:

$$\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta) = \mathbb{E}\left[\left(e^{n_{\rm d} R}\right)^{-\theta}\right] = \int_0^{\infty} (1+\rho\xi)^{-\tilde{\theta}} f_m(\xi) d\xi . \tag{19}$$

As shown in [28], [30], this integral can be obtained in closed-form using Tricomi's confluent hypergeometric function. Equivalently, we can also generalize [29] and express the Mellin transform as a sum of incomplete gamma functions [5] (the same result was independently found in [30]):

$$\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta) = \sum_{\mu=0}^{m-1} \frac{\binom{m-1}{\mu}(-1)^{\mu}}{\Gamma(m)\rho^{\mu+\tilde{\theta}}} e^{\frac{1}{\rho}} \Gamma\left(m-\mu-\tilde{\theta}, \frac{1}{\rho}\right) , \tag{20}$$

where $\tilde{\theta} \triangleq \frac{\theta n_{\rm d}}{\ln 2}$, and $\Gamma(s,x)$ is the upper incomplete Gamma function

$$\Gamma(s,x) = \int_x^{\infty} t^{s-1} e^{-t} dt . \tag{21}$$

Thus, given the arrival rate α in bits per time slot and a specific choice of superframe length T, the upper bound (18) on $p_{\rm v}(w)$ can be obtained analytically through (20).

IV. Analysis – Realistic Case

In the previous section, we considered the multiuser MISO downlink with perfect CSI and infinitely long codewords. However, in a real system, the transmitter first needs to acquire an estimate $\hat{\mathbf{H}}$ of the channel matrix $\mathbf{H}$ before computing the beamforming matrix $\mathbf{V}$. Due to the channel estimation error, the ZFBF matrix $\mathbf{V}$ is not perfectly matched to the actual channel $\mathbf{H}$, and thus, the interference cannot be completely eliminated. Furthermore, the transmitter must adapt the coding rate R to the imperfect channel estimate $\hat{\mathbf{H}}$. Outages will occur whenever the rate $R = \Phi(\hat{\mathbf{H}})$ selected by the transmitter happens to be below the actual capacity. Moreover, when the blocklength of the channel code is small, then one cannot achieve error-free transmissions at a rate equal to the channel capacity. Instead, the transmitter must choose rates below the

channel capacity in order to achieve low (but still non-zero) error probabilities. All these effects have an impact on the optimal number of scheduled users K. Specifically, there are now three additional reasons to choose a small value of K:

- Channel estimation overhead: For each scheduled user, a dedicated training period is required. At large K, this overhead severely reduces the number of symbols n_{d} that remain for the data transmission. Finite blocklength effects may cause an additional performance loss when n_{d} becomes very small.
- Interference: The signal for each scheduled user creates interference at the other users. A smaller number K thus reduces the interference.
- Backoff: In order to avoid errors, the transmitter must often choose a rate $R = \Phi(\hat{\mathbf{H}})$ below the estimated capacity. Reducing the number of scheduled users K increases the individual channel capacities and thus reduces the *relative* impact of this backoff.

However, there is still a major reason to consider a large number of scheduled users K:

- Reliability: When more users K are scheduled in each time slot, the transmitter can finish serving all K_{tot} within a shorter time T, which means that more retransmission opportunities are available before the target deadline w is reached. Although scheduling more users may increase the error probability of the *individual* transmissions, it could potentially enhance the *overall* reliability of the system with respect to the deadline by allowing more retransmissions.

Taking all these effects into account, our main problem remains the same: we want to determine the optimal value of K such that the overall reliability of the system with respect to the target deadline is maximized. To solve this question, one must also solve a secondary problem: one must determine the optimal rate adaptation function $\Phi : \hat{\mathbf{H}} \to R$. When choosing a high rate $R = \Phi(\hat{\mathbf{H}})$, the corresponding error probability ε will be too high. On the other hand, choosing very low rates may also lead to violations of the deadline, because then the transmitter cannot transmit all buffered data.

In order to solve these questions, we model in the following the effects of imperfect CSI and finite blocklength channel coding. Without loss of generality, we will consider only the signal at receiver $k = 1$. The derived quantities in this section correspond to user $k = 1$, which will not be indicated by a subscript. The analysis holds also if the users are subject to different path losses, as the received signal and the channel estimation error for user 1 are affected only by the path loss of user 1, which can be taken into account by scaling all powers.

A. Imperfect CSI

We consider a time-division duplex (TDD) system, where the transmitter can estimate the channel from training sequences of length n_{t} symbols sent by the users in the uplink. The training sequences must be mutually orthogonal, thus Kn_{t} channel uses are required for the training of K users. The SNR of the uplink channel is denoted as P_{UL} and known to the transmitter. By observing the training sequence, the transmitter can obtain the MMSE estimate $\hat{\mathbf{h}}_1$ of the channel towards user 1. According to [14], the actual channel vector $\mathbf{h}_1$ is given in terms of the MMSE estimate $\hat{\mathbf{h}}_1$ as

$$\mathbf{h}_1 = \hat{\mathbf{h}}_1 + \mathbf{e}_1 \,, \tag{22}$$

with $\mathbf{e}_1 \sim \mathcal{CN}(0, \sigma_{\mathrm{e}}^2 \mathbf{I})$ independent of $\hat{\mathbf{h}}_1$, and

$$\sigma_{\mathrm{e}}^2 = \frac{1}{1 + P_{\mathrm{UL}} n_{\mathrm{t}}} \,. \tag{23}$$

The transmitter then applies zero-forcing beamforming to create the beamforming matrix $\mathbf{V}$ based on the estimated channel matrix $\hat{\mathbf{H}}$. The received signal at user 1 is given as:

$$y_1 = \sqrt{\rho_1} \mathbf{h}_1^{\mathsf{H}} \mathbf{x} + z_1 \tag{24}$$

$$= \sqrt{\rho_1} \mathbf{h}_1^{\mathsf{H}} \mathbf{v}_1 s_1 + \sum_{j=2}^{K} \sqrt{\rho_j} \mathbf{h}_1^{\mathsf{H}} \mathbf{v}_j s_j + z_1 \,. \tag{25}$$

We assume for now that the sum power P_Σ is distributed equally, i.e., all users have power $\rho = P_\Sigma / K$. We discuss different power allocations at the end of this section. The signal-to-interference-plus-noise ratio (SINR) is then:

$$\mathsf{SINR} = \frac{\rho |\mathbf{h}_1^{\mathsf{H}} \mathbf{v}_1|^2}{1 + \sum_{j=2}^{K} \rho |\mathbf{h}_1^{\mathsf{H}} \mathbf{v}_j|^2} \tag{26}$$

$$= \frac{\rho |(\hat{\mathbf{h}}_1 + \mathbf{e}_1)^{\mathsf{H}} \mathbf{v}_1|^2}{1 + \sum_{j=2}^{K} \rho |\mathbf{e}_1^{\mathsf{H}} \mathbf{v}_j|^2} = \frac{G}{1 + I_\Sigma} \,, \tag{27}$$

where we have denoted the power of the signal at the receiver as G and the sum of the interference powers as I_Σ, and where we made use of the fact that $\hat{\mathbf{h}}_1^{\mathsf{H}} \mathbf{v}_j = 0$ for $j = 2, \ldots, K$. After the uplink training, the transmitter must send a downlink training sequence of $n_{\mathrm{t,DL}}$ symbols per user, so that the receivers can also learn the channel and decode the signal [14]. In this work, we assume that when $n_{\mathrm{t,DL}} \geq n_{\mathrm{t}}$, the estimation error at the receiver is negligible compared to the estimation error at the transmitter.[2] For the moment, we ignore the effects of finite blocklength channel coding, and assume that data can be successfully transmitted when the data rate R is below the instantaneous capacity $c = \log_2(1 + \mathsf{SINR})$. In each slot of length n_{tot} symbols, only $n_{\mathrm{d}} = n_{\mathrm{tot}} - K(n_{\mathrm{t}} + n_{\mathrm{t,DL}})$ symbols remain for the data transmission.

Given the estimated channel $\hat{\mathbf{H}}$, the transmitter must also choose a certain data rate R for the transmission. However, due to the channel estimation error $\mathbf{e}_1$, the transmitter does not know the exact value of the SINR in (27). Therefore, there is a chance that the channel will be in outage, i.e., that the transmission fails. The outage probability $\varepsilon_{\mathrm{out}}$, conditioned on the channel estimate $\hat{\mathbf{H}}$, is defined as

$$\varepsilon_{\mathrm{out}} = \mathbb{P}\left\{ \log_2(1 + \mathsf{SINR}) < R \,|\, \hat{\mathbf{H}} \right\} \,. \tag{28}$$

[2]The first reason for this assumption is that the transmitter can generally transmit at higher power than the users' devices, which may be battery-powered. Second, a small estimation error at the transmitter may lead to significant interference and to outages, whereas a small estimation error at the receiver would correspond only to a small additional noise term in the decoding process. Third, the receiver can estimate the channel not only from the dedicated training sequence, but also from the codeword itself (joint estimation and decoding) [36], further improving CSI at the receiver.

We now consider the distributions of the interference power I_Σ and the signal-of interest power G. The interference I_Σ is the sum of $K-1$ random variables (which may be correlated), each being exponentially distributed with mean

$$\mathbb{E}\left[\rho|\mathbf{e}_1^{\mathsf{H}}\mathbf{v}_j|^2\right] = \rho\mathbf{v}_j^{\mathsf{H}}\mathbb{E}\left[\mathbf{e}_1\mathbf{e}_1^{\mathsf{H}}\right]\mathbf{v}_j = \rho\sigma_{\mathrm{e}}^2\,. \tag{29}$$

The signal-of-interest has power $G = \rho|(\hat{\mathbf{h}}_1+\mathbf{e}_1)^{\mathsf{H}}\mathbf{v}_1|^2$. Conditioned on the known values $\hat{\mathbf{h}}_1$ and $\mathbf{v}_1$, with $\hat{g} = \rho|\hat{\mathbf{h}}_1^{\mathsf{H}}\mathbf{v}_1|^2$ denoting the estimated receive power, G has non-central chi-square distribution with two degrees of freedom and PDF

$$f_{G|\hat{g}}(g) = \frac{1}{\rho\sigma_{\mathrm{e}}^2}e^{-\frac{g+\hat{g}}{\rho\sigma_{\mathrm{e}}^2}}I_0\left(\frac{2\sqrt{\hat{g}g}}{\rho\sigma_{\mathrm{e}}^2}\right)\,. \tag{30}$$

As this PDF is difficult to handle analytically, we extend the results in [34] to the multi-antenna scenario and derive a Gaussian approximation for G:

$$G = \rho\mathbf{v}_1^{\mathsf{H}}(\hat{\mathbf{h}}_1+\mathbf{e}_1)(\hat{\mathbf{h}}_1+\mathbf{e}_1)^{\mathsf{H}}\mathbf{v}_1 \tag{31}$$

$$= \underbrace{\rho|\hat{\mathbf{h}}_1^{\mathsf{H}}\mathbf{v}_1|^2}_{\hat{g}\ \text{(known)}} + \underbrace{2\rho|\hat{\mathbf{h}}_1^{\mathsf{H}}\mathbf{v}_1|\Re\left\{e^{-i\angle(\hat{\mathbf{h}}_1^{\mathsf{H}}\mathbf{v}_1)}\mathbf{e}_1^{\mathsf{H}}\mathbf{v}_1\right\}}_{\tilde{G}\sim\mathcal{N}(0,\sigma_G^2)} + \underbrace{\rho|\mathbf{e}_1^{\mathsf{H}}\mathbf{v}_1|^2}_{\text{negligible}}. \tag{32}$$

The term $\mathbf{e}_1^{\mathsf{H}}\mathbf{v}_1$ is Gaussian and distributed as $\mathcal{CN}(0,\sigma_{\mathrm{e}}^2)$, according to (29). The distribution of a circularly symmetric random variable is not affected by a phase shift $e^{-i\angle(\hat{\mathbf{h}}_1^{\mathsf{H}}\mathbf{v}_1)}$, thus the second term, which is denoted as $\tilde{G}$, is a real-valued Gaussian variable with variance

$$\sigma_G^2 = 4\rho^2|\hat{\mathbf{h}}_1^{\mathsf{H}}\mathbf{v}_1|^2\frac{\sigma_{\mathrm{e}}^2}{2} = 2\sigma_{\mathrm{e}}^2\rho\hat{g}\,. \tag{33}$$

The third term has variance $\rho^2\sigma_{\mathrm{e}}^4$, which becomes insignificant compared to σ_G^2 even for moderate training powers P_{UL} and training sequence lengths n_{t}. Thus, the actual receive power G can be closely approximated as Gaussian: $G \approx \hat{g}+\tilde{G} \sim \mathcal{N}(\hat{g},\sigma_G^2)$.

In order to find closed-form approximations for $\varepsilon_{\mathrm{out}}$, we assume that G and I_Σ are independently distributed. This is well motivated in the high SNR regime, and numerical results show that the resulting closed form approximations are fairly accurate even at low SNR. Assume $P_\Sigma = P_{\mathrm{UL}}$ and let $P_\Sigma \to \infty$. The mean (which is equal to the standard deviation) of each interference term in I_Σ then converges to a constant:

$$\rho\sigma_{\mathrm{e}}^2 = \frac{P_\Sigma}{K}\frac{1}{1+P_{\mathrm{UL}}n_{\mathrm{t}}} \xrightarrow[P_{\mathrm{UL}}=P_\Sigma]{P_\Sigma\to\infty} \frac{1}{Kn_{\mathrm{t}}} \tag{34}$$

The variance σ_G^2 of the signal-of-interest power G converges to $\sigma_G^2 = 2\sigma_{\mathrm{e}}^2\rho\hat{g} \approx 2\hat{g}/(Kn_{\mathrm{t}})$. We note that the mean $\hat{g}$ grows with P_Σ, but the standard deviation σ_G grows only with $\sqrt{\hat{g}}$ and thus becomes negligible at high SNR. On the other hand, the standard deviation of the interference terms in I_Σ remains constant relative to its mean. Thus, the variations in $\mathsf{SINR} = G/(1+I_\Sigma)$ will be caused mostly by the interference I_Σ, whereas the variations in $G \approx \hat{g}$ and possible correlations to I_Σ can be completely ignored. In the low SNR regime, the variations in G must be considered, and the correlation between G and I_Σ would have some impact. Nevertheless, the numerical analysis in Sec. V-B shows that our derived expressions remain good approximations even for powers as low as $P_\Sigma = P_{\mathrm{UL}} = 5$ dB.

The beamforming vectors $\mathbf{v}_j$ are not mutually orthogonal, and thus the individual interference terms are correlated, which further complicates the analysis. As I_Σ is a sum of random variables, the variance of I_Σ is minimal in case the individual interference terms are completely independent (all $\mathbf{v}_j$ are orthogonal), and the variance is maximal in case all the interference terms are completely correlated (all $\mathbf{v}_j$ point in the same direction). Minimum variance of the interference generally minimizes the chance that the interference I_Σ is very large, and should therefore minimize the outage probability $\varepsilon_{\mathrm{out}}$ compared to the correlated case. Similarly, we conjecture that maximum variance (due to completely correlated interferers) maximizes the outage probability $\varepsilon_{\mathrm{out}}$. In the following two subsections, we use these two cases to obtain two approximations for the outage probability $\varepsilon_{\mathrm{out}}$. Our numerical analysis shows that for a wide range of parameters, the approximation based on uncorrelated interference provides a lower bound on $\varepsilon_{\mathrm{out}}$, and the approximation based on correlated interference provides an upper bound on $\varepsilon_{\mathrm{out}}$ (or at least a good approximation).

1) Approximation for Uncorrelated Interference: When assuming that the vectors are mutually orthogonal, all the interference terms $\rho|\mathbf{e}_1^{\mathsf{H}}\mathbf{v}_j|^2$ are uncorrelated. In this case, the sum interference I_Σ is given as the sum of $K-1$ independent, exponentially distributed random variables, each with mean $\rho\sigma_{\mathrm{e}}^2$. Thus, the interference I_Σ has gamma distribution with shape factor $\nu = (K-1)$ and scale $\lambda = \rho\sigma_{\mathrm{e}}^2$, whose cumulative distribution function is given as:

$$F_{I_\Sigma}(x) \approx 1 - \frac{\Gamma\left(\nu,\frac{x}{\lambda}\right)}{\Gamma(\nu)}\,. \tag{35}$$

Then, we obtain the following approximation:

Result 1. *When assuming that the interferers are mutually orthogonal, the conditional outage probability for a given channel estimate $\hat{\mathbf{H}}$ and a given rate R can be approximated as*

$$\varepsilon_{\mathrm{out}}(R,\hat{\mathbf{H}}) \approx Q\left(\frac{\hat{g}-\gamma_{\mathrm{o}}}{\sigma_G}\right) + e^{-\frac{\hat{g}-\gamma_{\mathrm{o}}}{\lambda\gamma_{\mathrm{o}}}+\frac{\sigma_G^2}{2(\lambda\gamma_{\mathrm{o}})^2}} \cdot\sum_{m=0}^{\nu-1}\sum_{l=0}^{m}\frac{\binom{m}{l}(\tilde{\mu}-\gamma_{\mathrm{o}})^{m-l}}{(\lambda\gamma_{\mathrm{o}})^m m!}B_l(\gamma_{\mathrm{o}}-\tilde{\mu}) \tag{36}$$

with $\hat{g} = \rho|\mathbf{h}_1\mathbf{v}_1|^2$, $\gamma_{\mathrm{o}} = 2^R-1$, $\lambda = \rho\sigma_{\mathrm{e}}^2$, $\tilde{\mu} = \hat{g}-\frac{\sigma_G^2}{\lambda\gamma_{\mathrm{o}}}$, and

$$B_l(x) = \int_x^\infty t^l\frac{1}{\sqrt{2\pi\sigma_G^2}}e^{-\frac{t^2}{2\sigma_G^2}}dt\,. \tag{37}$$

It can be seen directly that $B_0(x) = Q(x/\sigma_G)$. Furthermore, we obtain $B_1(x) = \sqrt{\frac{\sigma_G^2}{2\pi}}e^{-\frac{x^2}{2\sigma_G^2}}$. For values $l \geq 2$, integration by parts can be applied, resulting in:

$$B_l(x) = \sqrt{\frac{\sigma_G^2}{2\pi}}x^{l-1}e^{-\frac{x^2}{2\sigma_G^2}} + (l-1)\sigma_G^2B_{l-2}(x)\,. \tag{38}$$

Thus, the values of $B_l(\gamma_{\mathrm{o}}-\tilde{\mu})$ for $l = 2,\ldots,\nu-1$ can be obtained iteratively from (38)*.*

Proof. Defining $\gamma_{\rm o} = 2^R - 1$, the conditional outage probability can be approximated as

$$\varepsilon_{\rm out} = \mathbb{P}\left\{ \mathsf{SINR} < \gamma_{\rm o} | \, \hat{\mathbf{H}} \right\} = \mathbb{P}\left\{ \frac{G}{1+I_\Sigma} < \gamma_{\rm o} \middle| \, \hat{\mathbf{H}} \right\} \tag{39}$$

$$\approx \int_0^{\gamma_{\rm o}} f_{G|\hat{g}}(g) dg + \int_{\gamma_{\rm o}}^{\infty} \mathbb{P}\left\{ \frac{g}{\gamma_{\rm o}} - 1 < I_\Sigma \middle| \, \hat{\mathbf{H}} \right\} f_{G|\hat{g}}(g) dg \tag{40}$$

$$\approx \int_0^{\gamma_{\rm o}} f_{G|\hat{g}}(g) dg + \int_{\gamma_{\rm o}}^{\infty} \frac{1}{\Gamma(\nu)} \Gamma\left(\nu, \frac{g}{\lambda \gamma_{\rm o}} - \frac{1}{\lambda}\right) f_{G|\hat{g}}(g) dg \, . \tag{41}$$

where (40) follows from assuming independence between I_Σ and G and (41) follows when the terms in I_Σ are independent. Using the Gaussian approximation for G, and the finite series [37, Eq. (8.352.7)] for the incomplete gamma function (21) at integer values of ν:

$$\Gamma(\nu, x) = \Gamma(\nu) e^{-x} \sum_{m=0}^{\nu-1} \frac{x^m}{m!} \, , \tag{42}$$

we find that the outage probability in case of uncorrelated interferers is approximated by:

$$\varepsilon_{\rm out} \approx Q\left(\frac{\hat{g} - \gamma_{\rm o}}{\sigma_G} \right) + \\ + \int_{\gamma_{\rm o}}^{\infty} e^{-\frac{g-\gamma_{\rm o}}{\lambda \gamma_{\rm o}}} \sum_{m=0}^{\nu-1} \frac{\left(\frac{g}{\lambda \gamma_{\rm o}} - \frac{1}{\lambda} \right)^m}{m! \sqrt{2\pi \sigma_G^2}} e^{-\frac{(g-\hat{g})^2}{2\sigma_G^2}} dg \, . \tag{43}$$

The result then follows after some tedious algebra, where an important step is the binomial expansion $\left(\frac{g}{\lambda \gamma_{\rm o}} - \frac{1}{\lambda} \right)^m = \frac{1}{(\lambda \gamma_{\rm o})^m} \sum_{l=0}^{m} \binom{m}{l} (g - \tilde{\mu})^l (\tilde{\mu} - \gamma_{\rm o})^{m-l}$. □

2) Approximation for Correlated Interference: In the previous subsection, we considered the case where all vectors $\mathbf{v}_j$ are orthogonal, and thus, the individual interference terms are independent. Conversely, we consider in this subsection the extreme case where the vectors $\mathbf{v}_j$ for $j = 2, \ldots, K$ are identical, resulting in completely correlated interference. This assumption results in the maximum possible variance of the interference I_Σ. If all $\mathbf{v}_j$ are equal, then the sum interference I_Σ is equal to $(K-1)$ times the first interference term $\rho |\mathbf{e}_1^{\mathsf{H}} \mathbf{v}_2|^2$, which is exponentially distributed with mean $\rho \sigma_{\rm e}^2$. Thus, I_Σ is exponentially distributed with mean $\lambda_{\rm c} = \rho \sigma_{\rm e}^2 (K-1)$, which is equivalent to a gamma-distributed variable with shape $\nu = 1$ and scale $\lambda_{\rm c}$. Thus, the outage probability can be approximated by (36), which for $\nu = 1$ simplifies to the following result:

Result 2. *When assuming completely correlated interference, the outage probability for a given channel estimate $\hat{\mathbf{H}}$ and a given rate R can be approximated as*

$$\varepsilon_{\rm out}(R, \hat{\mathbf{H}}) \approx Q\left(\frac{\hat{g} - \gamma_{\rm o}}{\sigma_G} \right) + e^{-\frac{\hat{g}-\gamma_{\rm o}}{\lambda_{\rm c} \gamma_{\rm o}} + \frac{\sigma_G^2}{2(\lambda_{\rm c} \gamma_{\rm o})^2}} Q\left(\frac{\gamma_{\rm o} - \tilde{\mu}_{\rm c}}{\sigma_G} \right) \tag{44}$$

with $\hat{g} = \rho |\mathbf{h}_1 \mathbf{v}_1|^2$, $\gamma_{\rm o} = 2^R - 1$, and $\tilde{\mu}_{\rm c} = \hat{g} - \frac{\sigma_G^2}{\lambda_{\rm c} \gamma_{\rm o}}$.

We required in (36) and (44) that all signal powers ρ_j are equal. If the ρ_j are different, one could still use the approximate lower bound (36) after replacing all interference powers with $\min_j(\rho_j)$. This remains fairly accurate when the ρ_j are similar, which is reasonable at high SNR. For the delay analysis, we will rely mostly on the approximate upper bound (44), which still holds after changing $\lambda_{\rm c}$ to $\lambda_{\rm c} = \sigma_{\rm e}^2 \sum_{j=2}^{K} \rho_j$, so (44) holds even when the ρ_j are very different.

B. Finite Blocklength Channel Coding

1) Background: When the duration of each time slot is short, the blocklength of the channel code used for the transmission is rather small. This invalidates the assumption that error-free transmissions can be achieved at a rate equal to the channel capacity $\log_2(1 + \mathsf{SINR})$. Instead, results for finite blocklength channel coding must be used. For AWGN (additive white Gaussian noise) channels, a well-known result is given by Polyanskiy et al. [18, Thm. 54], who showed that given a maximum error probability ε, the achievable coding rate with $n_{\rm d}$ complex channel uses at SNR ρ is closely approximated by

$$R_{\rm AWGN}(n_{\rm d}, \varepsilon, \rho) \approx \log_2(1+\rho) - \sqrt{\frac{\mathcal{V}_{\rm AWGN}(\rho)}{n_{\rm d}}} Q^{-1}(\varepsilon) \, , \tag{45}$$

where the channel dispersion, adapted to our notation[3], is given as [18], [19], [34]

$$\mathcal{V}_{\rm AWGN}(\rho) = \left(1 - \frac{1}{(1+\rho)^2} \right) \log_2^2(e) \, . \tag{46}$$

However, this result for AWGN channels holds only in case the interference I_Σ is zero, i.e., when $K = 1$ or when the transmitter has perfect CSI and applies ZFBF. Under imperfect CSI with $K > 1$, each receiver experiences interference from the signals intended for other users. In order to achieve the rate (45) for user $k = 1$, the transmitter would need to use a non-Gaussian codebook [18]. Thus, the other users $k = 2, \ldots, K$ would be subject to non-Gaussian interference, and then (45) would not hold for the other users.

Thus, we must employ different results to model the effects of finite blocklength coding. Specifically, Scarlett et al. [20] considered the performance of Gaussian codebooks under non-Gaussian noise and nearest-neighbor decoding. The authors also considered the case where K sender-receiver pairs transmit concurrently, using i.i.d. Gaussian codebooks, and the receiver $k = 1$ experiences i.i.d. Gaussian interference from the transmitters $k = 2, \ldots, K$. These results can be directly applied to our scenario because there is no difference between interference that originates from K independent transmitters and interference that originates from a single transmitter superimposing K independently coded signals. Therefore, when i.i.d. Gaussian codebooks are used, a second-

[3]We define R in bits instead of nats, and we have $n_{\rm d}$ complex-valued channel uses, corresponding to $2n_{\rm d}$ real channel uses.

order approximation for the achievable coding rate is given by [20, Eq. (24)]

$$R_{\mathrm{iid}}(n_{\mathrm{d}},\varepsilon,\rho)=\log_2(1+\rho)-\sqrt{\frac{\mathcal{V}_{\mathrm{iid}}(\rho)}{n_{\mathrm{d}}}}Q^{-1}(\varepsilon)+\mathcal{O}\left(\frac{\log n}{n}\right) \tag{47}$$

with dispersion [20, Eq. (27)], adapted to our notation[3]:

$$\mathcal{V}_{\mathrm{iid}}(\rho)=\frac{2\cdot\rho}{1+\rho}\log_2^2(e)\,. \tag{48}$$

When the transmitter picks an i.i.d. Gaussian codebook containing $2^{n_{\mathrm{d}}R}$ messages, the decoding error probability at the receiver at a specific SINR can thus be approximated as

$$\varepsilon(\mathsf{SINR})\approx Q\left(\frac{\log_2(1+\mathsf{SINR})-R}{\sqrt{\mathcal{V}_{\mathrm{iid}}(\mathsf{SINR})/n_{\mathrm{d}}}}\right). \tag{49}$$

We note that the choice of a Gaussian codebook depends only on its size, defined by the number of messages 2^{nR}, and not on the exact value of SINR. As a result, (49) holds even when the transmitter does not know the exact value of SINR ahead of the transmission. Thus, when the transmitter knows only the estimated channel $\hat{\mathbf{H}}$ and chooses a coding rate R, the overall error probability can be approximated as[4]

$$\varepsilon\approx\mathbb{E}\left[Q\left(\frac{\log_2(1+\mathsf{SINR})-R}{\sqrt{\mathcal{V}_{\mathrm{iid}}(\mathsf{SINR})/n_{\mathrm{d}}}}\right)\middle|\,\hat{\mathbf{H}}\right], \tag{50}$$

where the expectation is taken over the distribution of SINR, conditioned on the estimated channel matrix $\hat{\mathbf{H}}$. We note that (45) and (47) are second order approximations, i.e., as $n_{\mathrm{d}}\to\infty$, the term $\mathcal{O}\left(\log(n)/n\right)$ becomes insignificant compared to the second term, which decays as $\mathcal{O}\left(1/\sqrt{n}\right)$. For the AWGN channel, it was shown that the approximation (45) can be accurate for blocklengths as small as $n_{\mathrm{d}}\approx 200$ [18]. In our previous work [34], which considered a transmitter with only one antenna, we used a strict lower bound on the achievable coding rate to show that the approximation was very accurate for the considered parameters. However, we are not aware of any results that can be used to verify the accuracy of (47) and (50). Therefore, our results should be seen as approximations, which can help guide the transmitter in the difficult task of selecting the coding rate R and the optimal number of scheduled users K.

The error probability ε in (50) can be obtained from Monte-Carlo simulations just like in the case of infinite blocklength. However, in order to find an optimal rate adaptation function $\Phi:\hat{\mathbf{H}}\to R$, the transmitter must be able to quickly determine the error probability ε for rate R through a closed-form expression. Therefore, we apply several approximations to (50) in order to obtain a closed-form expression.

2) Closed-form Approximation: We apply the concept of *random blocklength-equivalent capacity* [34], which allows treating the effects from finite-blocklength coding in the same fashion as outages. We define the random blocklength-equivalent capacity as

$$C_{\mathrm{b}}=\log_2(1+\mathsf{SINR})-\sqrt{\frac{\mathcal{V}_{\mathrm{iid}}(\mathsf{SINR})}{n_{\mathrm{d}}}}U_{\mathrm{b}} \tag{51}$$

with $U_{\mathrm{b}}\sim\mathcal{N}(0,1)$ independent of SINR. For a fixed value of SINR, only U_{b} is random, and the blocklength-equivalent outage probability $\mathbb{P}\left\{C_{\mathrm{b}}<R\right\}$ is – by definition – equal to the error probability at finite blocklength $\varepsilon(\mathsf{SINR})$ given in (49). Furthermore, it can be easily verified that when both SINR and U_{b} are random, $\mathbb{P}\left\{C_{\mathrm{b}}<R|\hat{\mathbf{H}}\right\}$ is exactly equal to the expression for ε given in (50). Using this concept, we follow the further steps in [34] and apply the first-order Taylor approximation [34]

$$\log_2(x)-a\geq\log_2\left(x-\frac{x}{\log_2(e)}a\right) \tag{52}$$

to C_{b} around $(1+\mathsf{SINR})$ in order to bring U_{b} into the same domain as SINR:

$$C_{\mathrm{b}}\approx\log_2\left(1+\mathsf{SINR}-\frac{1+\mathsf{SINR}}{\log_2(e)}\sqrt{\frac{\mathcal{V}_{\mathrm{iid}}(\mathsf{SINR})}{n_{\mathrm{d}}}}U_{\mathrm{b}}\right) \tag{53}$$

$$=\log_2\left(1+\frac{G}{1+I_\Sigma}-\left(1+\frac{G}{1+I_\Sigma}\right)\sqrt{\frac{\mathcal{V}_{\mathrm{iid}}\left(\frac{G}{1+I_\Sigma}\right)}{n_{\mathrm{d}}\log_2^2(e)}}U_{\mathrm{b}}\right) \tag{54}$$

$$\approx\log_2\left(1+\frac{\hat{g}+\tilde{G}-\left(1+I_\Sigma+\hat{g}+\tilde{G}\right)\sqrt{\frac{\mathcal{V}_{\mathrm{iid}}\left(\frac{\hat{g}+\tilde{G}}{1+I_\Sigma}\right)}{n_{\mathrm{d}}\log_2^2(e)}}U_{\mathrm{b}}}{1+I_\Sigma}\right) \tag{55}$$

$$\approx\log_2\left(1+\frac{\hat{g}+\tilde{G}-(1+\mathbb{E}\left[I_\Sigma\right]+\hat{g})\sqrt{\frac{\mathcal{V}_{\mathrm{iid}}\left(\frac{\hat{g}}{1+\mathbb{E}[I_\Sigma]}\right)}{n_{\mathrm{d}}\log_2^2(e)}}U_{\mathrm{b}}}{1+I_\Sigma}\right) \tag{56}$$

$$=\log_2\left(1+\frac{\hat{g}+\tilde{G}_{\mathrm{IC,F}}}{1+I_\Sigma}\right). \tag{57}$$

In (53), we applied the Taylor approximation. In (55), we applied the Gaussian approximation $G\approx\hat{g}+\tilde{G}$. In (56), we replaced I_Σ and $\tilde{G}$ in the factor before U_{b} with their respective expectations. This is reasonable because this factor corresponds only to the variance of the term with U_{b}. Although even small values of $\tilde{G}$ or I_Σ can cause an outage, the same small values lead only to a small change (relative to $\hat{g}$) in the variance of the term with U_{b}, which does not significantly affect the distribution of C_{b}. Finally, in (57), we have defined $\tilde{G}_{\mathrm{IC,F}}$ as the sum of the Gaussian variable $\tilde{G}$ and the independent Gaussian variable U_{b}, multiplied by a constant factor. The sum of two independent Gaussian random variables is Gaussian, and the variance of the sum is equal to the sum of

[4]This still requires that the receivers obtain perfect CSI from the $n_{\mathrm{t,DL}}$ downlink training symbols. However, for the single-antenna case, we showed that an expression similar to (50) is accurate even when the receiver has only imperfect CSI [35].

the individual variances. Thus, $\tilde{G}_{\mathrm{IC,F}}$ is zero-mean Gaussian with variance

$$\sigma^2_{\mathrm{IC,F}} = \sigma^2_G + \left(1+\rho\sigma^2_{\mathrm{e}}(K-1)+\hat{g}\right)^2 \frac{\mathcal{V}_{\mathrm{iid}}(\frac{\hat{g}}{1+\rho\sigma^2_{\mathrm{e}}(K-1)})}{n_{\mathrm{d}}\log_2^2(e)}. \tag{58}$$

The error probability ε due to finite blocklength coding and imperfect CSI is given in (50), which is equal to the blocklength-equivalent outage probability $\mathbb{P}\left\{C_{\mathrm{b}} < R|\hat{\mathbf{H}}\right\}$. Using the approximation (57) for C_{b}, we can follow the same steps as in Sec. IV-A to approximate ε, simply replacing $\tilde{G}$ by $\tilde{G}_{\mathrm{IC,F}}$ in (44):

Result 3. *When assuming that the interference I_Σ is completely correlated, the error probability under imperfect CSI and finite blocklength coding for a given channel estimate $\hat{\mathbf{H}}$ and rate R can be approximated as:*

$$\varepsilon(R,\hat{\mathbf{H}}) \approx Q\left(\frac{\hat{g}-\gamma_{\mathrm{o}}}{\sigma_{\mathrm{IC,F}}}\right) + e^{-\frac{\hat{g}-\gamma_{\mathrm{o}}}{\lambda_{\mathrm{c}}\gamma_{\mathrm{o}}}+\frac{\sigma^2_{\mathrm{IC,F}}}{2(\lambda_{\mathrm{c}}\gamma_{\mathrm{o}})^2}} Q\left(\frac{\gamma_{\mathrm{o}}-\tilde{\mu}_{\mathrm{IC,F}}}{\sigma_{\mathrm{IC,F}}}\right) \tag{59}$$

with $\hat{g} = \rho|\mathbf{h}_1\mathbf{v}_1|^2$, $\gamma_{\mathrm{o}} = 2^R-1$, $\lambda_{\mathrm{c}} = \rho\sigma^2_{\mathrm{e}}(K-1)$, and $\tilde{\mu}_{\mathrm{IC,F}} = \hat{g} - \frac{\sigma^2_{\mathrm{IC,F}}}{\lambda_{\mathrm{c}}\gamma_{\mathrm{o}}}$.

Our numerical evaluations show that for most parameters, (59) is an upper bound to ε in (50) because we assumed correlated interference and also because the Taylor approximation (52) was designed to obtain a lower bound on C_{b}. When assuming uncorrelated interference, the error probability ε under finite blocklength coding can be approximated by replacing σ_G with $\sigma_{\mathrm{IC,F}}$ in (36) and (37). However, due to the Taylor approximation (52), the resulting expression is not a lower bound to ε. Therefore, we use only (59) to estimate the effects of finite length coding.

C. Delay Analysis and Rate Adaptation

For the scenario with perfect CSI, we presented closed-form expressions for the Mellin transform of the SNR-domain service process $\mathcal{S}^{(T)} = e^{n_{\mathrm{d}}R}$ in Sec. III-C. These closed-form expressions can be used to compute the kernel function $\mathbb{K}^{(T)}(\theta, w/T)$ in (16), which is an upper bound on the delay violation probability $p_{\mathrm{v}}(w)$.

However, in case of imperfect CSI, finding an upper bound on $p_{\mathrm{v}}(w)$ is a much more difficult task, because it is still unclear what rate $R = \Phi(\hat{\mathbf{H}})$ the transmitter should choose. The optimal rate adaptation function Φ^* must find a balance between the rate $R = \Phi(\hat{\mathbf{H}})$ and the corresponding error probability $\varepsilon(\Phi(\hat{\mathbf{H}}),\hat{\mathbf{H}})$ such that the reliability of the system with respect to the deadline w is maximized. Specifically, we want to find

$$\Phi^* = \arg\min_{\Phi} \quad p_{\mathrm{v}}(w). \tag{60}$$

Because $p_{\mathrm{v}}(w)$ cannot be determined in analytical form, we follow our previous work [34] and perform the optimization based on the analytical upper bound $\mathbb{K}^{(T)}(\theta, w/T)$ on $p_{\mathrm{v}}(w)$. First, we fix a parameter $\theta > 0$ and determine

$$\Phi^*_\theta = \arg\min_{\Phi} \quad \mathbb{K}^{(T)}(\theta, w/T, \Phi) \tag{61}$$

$$= \arg\min_{\Phi} \quad \mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta) \tag{62}$$

$$= \arg\min_{\Phi} \quad \mathbb{E}\left[(1-\varepsilon)\, e^{-\theta n_{\mathrm{d}}\Phi(\hat{\mathbf{H}})} + \varepsilon\right] \tag{63}$$

with $\varepsilon = \varepsilon(\Phi(\hat{\mathbf{H}}),\hat{\mathbf{H}})$ also depending on Φ. In (61), we specifically denoted that the kernel $\mathbb{K}^{(T)}(\cdot)$ depends implicitly on the rate adaptation function Φ. The second step (62) follows directly by inspecting (16). Then, the optimal rate adaptation function Φ^* can be found by iterating over $\theta > 0$:

$$\Phi^* \approx \arg\min_{\Phi^*_\theta} \quad \inf_{\theta>0} \quad \mathbb{K}^{(T)}(\theta, w/T, \Phi^*_\theta). \tag{64}$$

In order to solve this problem, we use the closed-form approximation for the error probability derived in the previous section, which depends on $\hat{\mathbf{H}}$ only through the estimated SNR $\hat{g} = \rho|\mathbf{h}_1\mathbf{v}_1|^2$. Thus, we need to take the expected value in (63) only with respect to the distribution of $\hat{g}$, which has χ^2 distribution with $2m$ degrees of freedom. We quantize this distribution to points $\hat{g}_i$ with $i = 1,\ldots,N_{\hat{g}}$, where the probability of each value is denoted as $p_{\hat{g}}(i)$. For each quantized value $\hat{g}_i$, we can choose a rate R_i and obtain $\varepsilon(R_i,\hat{g}_i)$ from (59). Therefore,

$$\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta) \approx \sum_i p_{\hat{g}}(i)\left((1-\varepsilon)e^{-\theta n_{\mathrm{d}}R_i}+\varepsilon\right) \tag{65}$$

with $\varepsilon = \varepsilon(R_i,\hat{g}_i)$. In order to find the optimal rate adaptation function Φ^*_θ, we need to determine the optimal rates R^*_i that minimize (65). This can be achieved by generating a vector of quantized rates R_j with $j = 1,\ldots,N_R$, and finding

$$R^*_i = \arg\min_{R_j} \quad (1-\varepsilon(R_j,\hat{g}_i))e^{-\theta n_{\mathrm{d}}R_j} + \varepsilon(R_j,\hat{g}_i)). \tag{66}$$

The optimal rate adaptation function Φ^* can then be obtained by repeating this process for different values of $\theta > 0$. This can be done efficiently, as the error probabilities $\varepsilon(R_j,\hat{g}_i)$ do not depend on θ.

V. Numerical Results

For the numerical evaluations of our results, we consider a symmetric setup where all users have the same path loss, arrival rate, and delay requirements, in order to focus on the performance difference between ideal and the realistic model. However, the analysis also applies in asymmetric scenarios, since the delay performance of each user depends only on the number K of concurrently scheduled users, but not on the path loss or arrival rate of the other users. We first investigate in Sec. V-A the channel hardening effect under the basic, ideal system model. Then, in Sec. V-B we validate the approximations of the outage probability under imperfect CSI. In Sec. V-C we investigate how imperfect CSI affects the delay performance. In Sec. V-D, we also take the effects of finite blocklength channel coding into account, and confirm by Monte Carlo simulations that the analytical approximations of the error and outage probabilities lead to upper bounds on the delay violation probability.

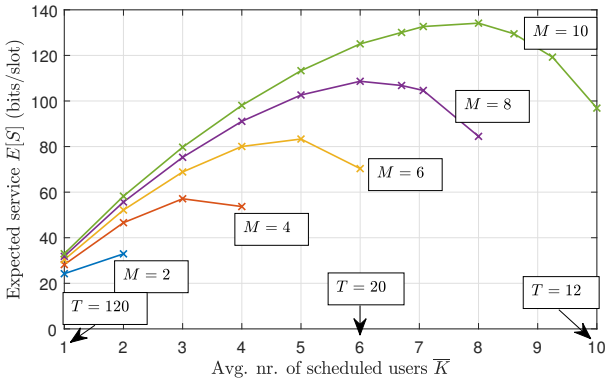
A. Ideal case

For the ideal scenario with perfect CSI, we first show in Fig. 1 the expected service rate $\mathbb{E}[S]$ of the system per time slot versus the number of scheduled users $\overline{K}$, for different numbers of transmit antennas M. The expected service per time slot is given as

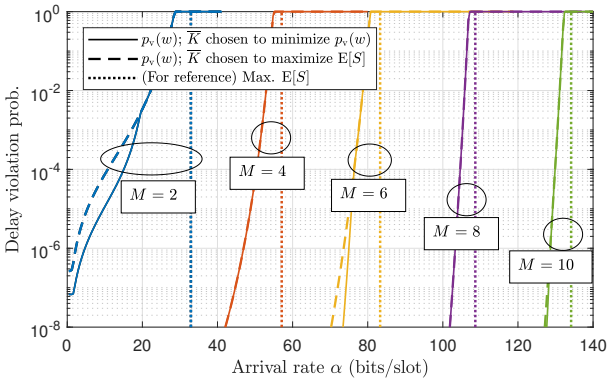
$$\mathbb{E}[S] = \frac{1}{T}\mathbb{E}\left[S^{(T)}\right] = \frac{1}{T}\mathbb{E}[n_{\mathrm{d}} R Z] \ . \tag{67}$$

In the ideal case, no errors occur, i.e., we always have $Z = 1$. The total number of users is $K_{\mathrm{tot}} = 120$. The transmitter can then choose for example a superframe length of $T = 40$ slots and schedule $\overline{K} = 3$ users in each time slot. We observe that the expected service $\mathbb{E}[S]$ first increases in $\overline{K}$ and then decreases. This is not surprising. For example, increasing $\overline{K}$ from 1 to 2 doubles the multiplexing gain, but barely affects the beamforming gain. Contrary to that, at large $\overline{K}$, an increase in $\overline{K}$ leads only to a minor increase in multiplexing gain, but a large decrease in the beamforming gain.

Fig. 1b shows the delay violation probability $p_{\mathrm{v}}(w)$ vs. the arrival rate α. For each data point, we select the number of scheduled users $\overline{K}$ such that $p_{\mathrm{v}}(w)$ is minimized. However, we note that there are only few cases where the optimization over $\overline{K}$ improves the performance compared to the value $\overline{K}$ that maximizes the expected service rate $\mathbb{E}[S]$. For $M \in \{6, 8, 10\}$, we observe that $p_{\mathrm{v}}(w) < 10^{-8}$ even when the arrival rate α is only 10% below the expected service rate $\mathbb{E}[S]$. Then, $p_{\mathrm{v}}(w)$ rises sharply to 1 as α increases. Thus, we observe significant channel hardening and an approximate zero/one queueing behavior for the considered parameters, even with a fairly small number of antennas. When the number of antennas increases further, the beamforming gain will increase even further, which will further decrease the probability of queueing delays. In the following, we will investigate whether this observation holds also when we consider a more realistic system model.



(a)



(b)

Fig. 1. $K_{\mathrm{tot}} = 120$ users, $n_{\mathrm{d}} = 400$ symbols, $P_{\Sigma} = 20$ dB. (a) Expected service $\mathbb{E}[S]$ in bits per slot for $M \in \{2, 4, 6, 8, 10\}$. (b) Bound on the delay violation probability $p_{\mathrm{v}}(w)$ for $w = 120$ slots vs. arrival rate α, compared to expected service.

B. Imperfect CSI – Validation

In Fig. 2, we validate the approximations (36) and (44) for the outage probability $\varepsilon_{\mathrm{out}}$. First, in Fig. 2a we consider $M = 8$ antennas with $K = 5$ users. We investigate the outage probability vs. the data rate when the estimated capacity $\log_2(1+\hat{g})$ of the channel is 6 bits/channel use (this is close to the mean value). For the simulations, we generate matrices $\hat{\mathbf{H}}$ and compute the corresponding beamformers $\mathbf{v}_j$ until we find a matrix $\hat{\mathbf{H}}_{\hat{R}_1 \approx 6}$ such that $\log_2(1+\hat{g}) = \log_2(1+\rho|\hat{\mathbf{h}}_1 \mathbf{v}_1|^2)$ is between 5.99 and 6.01 bits. For each such matrix, we generate at least 10^6 instances of the estimation error $\mathbf{e}_1$, in order to obtain the distribution of the actual SINR and the outage probability $\varepsilon_{\mathrm{out}}$, conditioned on the channel estimate $\hat{\mathbf{H}}_{\hat{R}_1 \approx 6}$.

Fig. 2a shows the resulting $\varepsilon_{\mathrm{out}}$ for several instances of $\hat{\mathbf{H}}_{\hat{R}_1 \approx 6}$, as well as $\varepsilon_{\mathrm{out}}$ averaged over 1000 instances of $\hat{\mathbf{H}}_{\hat{R}_1 \approx 6}$. The average over $\varepsilon_{\mathrm{out}}$ (bold dotted curve) has the following meaning: when the transmitter estimates the channel capacity to be $\log_2(1+\hat{g}) \approx 6$ bits/channel use, the transmitter must choose a rate of $R \approx 4.6$ to achieve $\varepsilon_{\mathrm{out}} < 10^{-4}$. However, repeating these Monte Carlo simulations for many different values of the estimated SNR $\hat{g}$ is computationally prohibitive. Instead, we want to determine the rate R from the two analytical approximations for $\varepsilon_{\mathrm{out}}$. The approximation (44) for correlated interference is best suited when high reliability is desired: if the transmitter wants to keep $\varepsilon_{\mathrm{out}}$ below e.g. 10^{-4}, it would choose a rate $R \approx 4.3$ according to this approximation. This is a robust choice, because (44) acts as an upper bound on $\varepsilon_{\mathrm{out}}$ (except at unreasonably large $\varepsilon_{\mathrm{out}}$), and so the actual $\varepsilon_{\mathrm{out}}$ is below the target of 10^{-4}. Conversely, the approximation (36) for uncorrelated interference provides a lower bound on $\varepsilon_{\mathrm{out}}$ for most parameters. Fig. 2a shows also that the correlation between the beamforming vectors $\mathbf{v}_j$ seems to be quite low for many instances of $\hat{\mathbf{H}}_{\hat{R}_1 \approx 6}$. Despite that, the rare instances of $\hat{\mathbf{H}}_{\hat{R}_1 \approx 6}$ where the correlation is high seem to have a strong impact on $\varepsilon_{\mathrm{out}}$. We also show results for two hypothetical models: one where the variations in G are generated independently of I_{Σ} (yellow dotted curve), and one where there are no variations in G, i.e., $G = \hat{g}$ (green dotted curve). The figure shows almost no difference between both hypothetical models and the actual system. These results confirm the rationale from Sec. IV-A that at high SNR, the variations in G are negligible, which we used to motivate that G and I_{Σ} can be treated as independent variables.

In Fig. 2b, we consider $K = 2$ users and channels $\hat{\mathbf{H}}_{\hat{R}_1 \approx 8}$

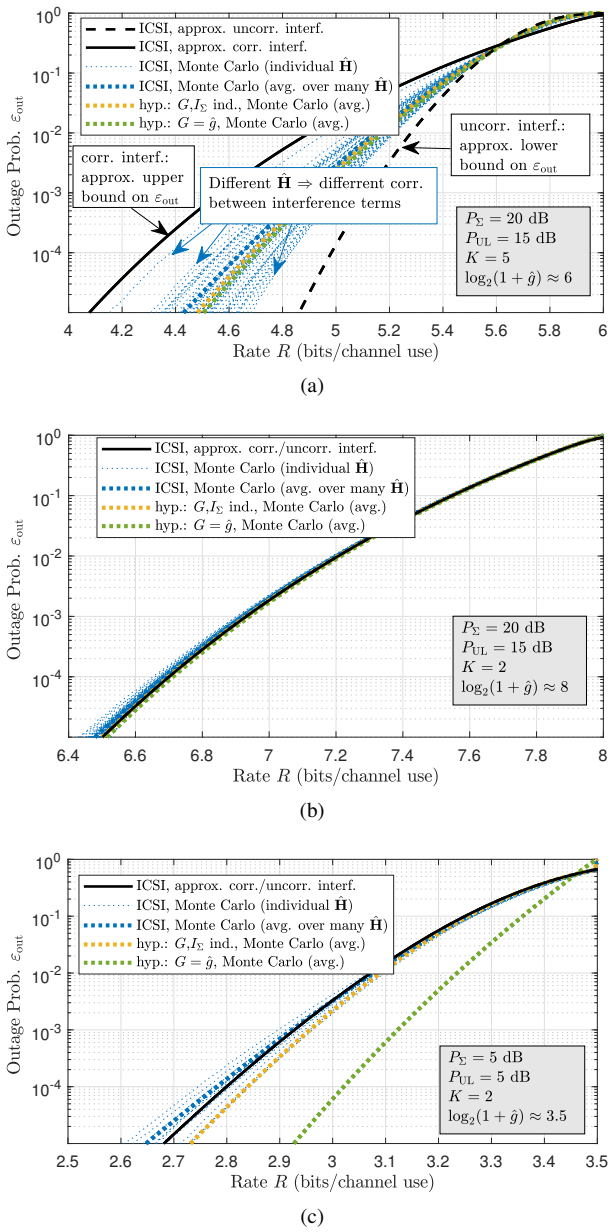


Fig. 2. Validation of the approximations for ε_{out} using Monte Carlo simulations: ε_{out} vs. R for $M = 8$, $n_{\text{t}} = 10$. (a) high SNR, $K = 5$, (b) high SNR, $K = 2$, (c) low SNR, $K = 2$.

where $\log_2(1+\hat{g})$ is between 7.99 and 8.01 bits. For $K = 2$, there is only a single interferer and therefore no difference between the two approximations (36) and (44). The Monte Carlo simulations match the analytical results almost exactly. We note that for $K = 2$, the correlation between the signal power G and the interference I_Σ can sometimes lead to a tiny increase in ε_{out}. Therefore, (44) is not an upper bound on ε_{out}, but an accurate approximation. Finally, in Fig. 2c, we investigate the approximations in the low SNR regime with $P_\Sigma = P_{\text{UL}} = 5$ dB. There is now a massive difference beween the actual system and the hypothetical model that assumes no variations in G ($G = \hat{g}$). The other hypothetical model, which considers these variations but ignores the correlation of G and I_Σ, shows similar performance as the actual system. The derived approximations are still very accurate. These results validate our analytical approach, which considers the variations in G but treats them as independent of I_Σ.

C. Imperfect CSI – Results

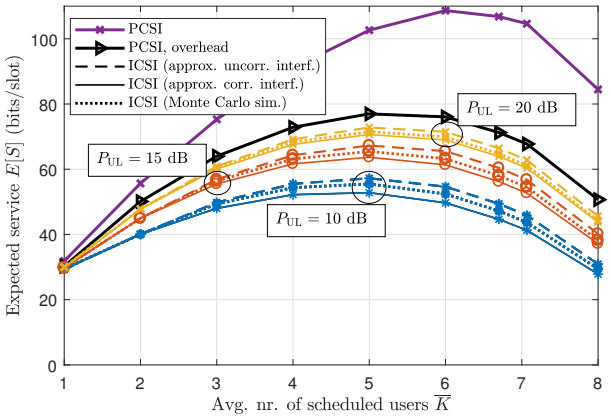
In Fig. 3, we study the how imperfect CSI affects the performance. First, in Fig. 3a we show the expected service $\mathbb{E}[S]$ per slot vs. the number of scheduled users $\overline{K}$ for $M = 8$. The curve for perfect CSI (PCSI), while assuming no overhead, was already shown in Fig. 1a in the previous section. We now find that the performance massively deteriorates when considering channel estimation and imperfect CSI. First of all, the bold black curve shows results where CSI is still assumed to be perfect, but an overhead of $n_{\text{t}} = 10$ and $n_{\text{t,DL}} = 10$ symbols for the uplink and downlink training is taken into account. The overhead already leads to a large performance loss at large $\overline{K}$, and the optimal value of $\overline{K}$ reduces to 5. Now, we consider three values of $P_{\text{UL}} \in \{10, 15, 20\}$ dB to show different levels of CSI quality. In all cases, we plot three different performance curves, corresponding to the approximate lower bound on ε_{out} in (36) (dashed curve), to the approximate upper bound on ε_{out} in (44) (thin solid curve), and to Monte Carlo simulations (dotted curve).[5] We find that the two approximations for the outage probability ε_{out} match fairly well, and that the results for the Monte Carlo simulations always lie between or very close to the approximations. Overall, we find that imperfect CSI can have quite a strong impact on the system performance. Surprisingly, this does not lead to a change in the optimal number of scheduled users $\overline{K}$, which remains at $\overline{K} = 5$ (after the training overhead was taken into account). Lastly, we note that the approximations correctly predict the optimal value $\overline{K}$ as determined from the simulations.

Finally, we investigate in Fig. 3b whether the approximate zero/one behavior with respect to the delay violation probability $p_{\text{v}}(w)$ is maintained when considering imperfect CSI. Despite the possibility of outages, and the backoff that is required to achieve low outage probabilities, we find that the approximate zero/one behavior of the system remains (even though the slopes become less steep): The delay violation probability is close to zero when the arrival rate α is around 20% below the expected service $\mathbb{E}[S]$, and is equal to one when $\alpha > \mathbb{E}[S]$. These results stand in stark contrast to the findings for the single-antenna case [34], where imperfect CSI and finite blocklength had only a moderate impact on the expected service, but a dramatic impact on the performance under delay constraints.

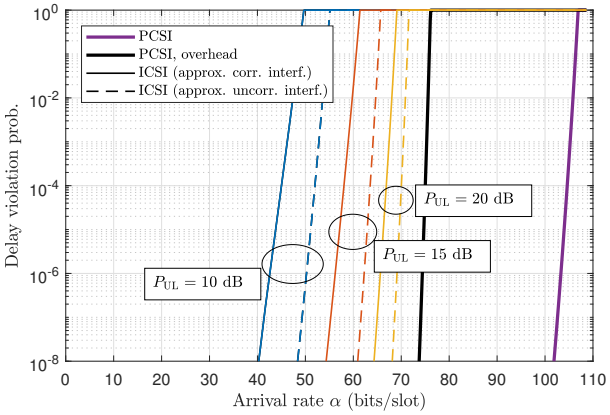
D. Finite Blocklength Coding

In Fig. 4a, we investigate the impact of finite blocklength coding. For the simulations, we use the same methods that were used for Fig. 2a, which results in random samples of SINR. The error probability ε at finite blocklength (50) can be obtained by computing $\varepsilon(\mathsf{SINR})$ in (49) for each realization of

[5]For the simulations, we use the rate adaptation function Φ that was obtained for the approximation with correlated interference (solid curve). As expected, the simulations show better performance, because the actual ε_{out} is below the approximation.

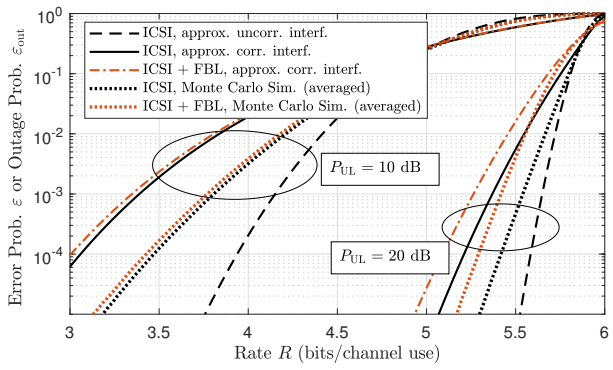


(a)

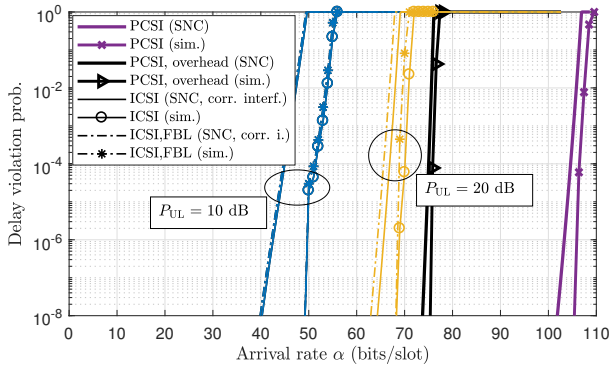


(b)

Fig. 3. $K_{\mathrm{tot}} = 120$ users, $M = 8$, $n_{\mathrm{tot}} = 400$ symbols, $P_\Sigma = 20$ dB, different values of P_{UL}. (a) Expected service $\mathbb{E}[S]$ in bits per slot vs. $\overline{K}$. (b) Delay violation probability $p_{\mathrm{v}}(w)$ (from SNC) vs. α. Deadline $w = 120$ slots.



(a)



(b)

Fig. 4. $K_{\mathrm{tot}} = 120$ users, $M = 8$, $n_{\mathrm{tot}} = 400$ symbols, $P_\Sigma = 20$ dB, different values of P_{UL}. (a) Error/outage probabilities vs. rate R when $\log_2(1+\hat{g}) \approx 6$, $K = 5$. (b) Delay violation probability (both analytical results (using SNC with the approximation for correlated interference) and Monte Carlo simulations over 10^7 time slots) vs. arrival rate α.

SINR, and then taking the average. We find that finite length coding has very little impact on the performance when the CSI quality is poor ($P_{\mathrm{UL}} = 10$ dB). However, when the quality of the channel estimates increases, finite blocklength effects are more pronounced. Nevertheless, even at $P_{\mathrm{UL}} = 20$ dB, the system loses only 0.1 bits in the rate due to finite blocklength effects. Although this performance loss is small, it cannot be ignored when high reliability is desired. We note that the approximations based on correlated interference correctly predict the performance loss of 0.1 bits that can be observed in the simulations. Therefore, we use in the following only these approximate upper bounds.

In Fig. 4b, we investigate the delay performance under finite length coding. First, we consider only the analytical results from stochastic network calculus. At $P_{\mathrm{UL}} = 10$ dB, the performance loss is dominated by imperfect CSI, and finite blocklength effects are negligible. When the accuracy of the CSI increases ($P_{\mathrm{UL}} = 20$ dB), finite blocklength effects cause a small performance penalty. However, the delay violation probability maintains its approximate zero/one behavior. We conclude that finite blocklength effects often have a much smaller performance impact than imperfect CSI. This is in line with previous results for the single-antenna case [34].

Finally, we also performed Monte Carlo simulations of the queueing system. We can confirm that the actual delay violation probability $p_{\mathrm{v}}(w)$ observed in the simulations is always below the analytical results. However, there is a gap between the analytical results and the simulation results. This is mostly because the approximate upper bounds on $\varepsilon_{\mathrm{out}}$ and ε are not perfectly tight. Nevertheless, the analytical results are useful, as they predict quite accurately that a system with $P_{\mathrm{UL}} = 10$ dB can only support an arrival rate α of 40 to 50 bits per slot, which is much less than the roughly 100 bits per slot in the ideal (perfect CSI) model. Additionally, the approximations correctly predict that finite blocklength coding causes only a small performance loss for the considered scenario.

VI. Conclusions

We considered the delay performance of the multiuser MISO downlink under ideal and under realistic assumptions. Under ideal assumptions, multiple antennas create an almost deterministic queueing behavior, i.e., the system can achieve very high reliability with respect to the deadline, as long as the average transmission rate is large enough. When considering imperfect CSI and finite blocklength coding, we observed a massive degradation of the average transmission rate. Nevertheless, we found that the system maintains the same qualitative behavior: when the average transmission rate exceeds the arrival rate, the system can still achieve very

high reliability. While it has long been known that multi-antenna technology can greatly improve the reliability, it is still surprising that the system remains extremely reliable even under non-ideal assumptions.

APPENDIX A
ANALYSIS FOR NON-INTEGER $\overline{K}$

The average number of scheduled users per slot is defined as $\overline{K} \triangleq K_{\mathrm{tot}}/T$. In case $\overline{K}$ is not an integer number, the transmitter schedules $K_{\mathsf{A}} = \lceil \overline{K} \rceil$ in T_{A} time slots, and $K_{\mathsf{B}} = \lfloor \overline{K} \rfloor$ in $T_{\mathsf{B}} = T - T_{\mathsf{A}}$ time slots, where T_{A} and T_{B} can be determined from $T_{\mathsf{A}}K_{\mathsf{A}} + T_{\mathsf{B}}K_{\mathsf{B}} = K_{\mathrm{tot}}$. Each user's signal is transmitted with power $\rho_{\mathsf{A}} = P_\Sigma/K_{\mathsf{A}}$ or $\rho_{\mathsf{B}} = P_\Sigma/K_{\mathsf{B}}$. To maintain fairness between all users, the users are randomly assigned to the A or B slots in each superframe, with probabilities $p_{\mathsf{A}} = K_{\mathsf{A}}T_{\mathsf{A}}/K_{\mathrm{tot}}$ and $p_{\mathsf{B}} = K_{\mathsf{B}}T_{\mathsf{B}}/K_{\mathrm{tot}}$, respectively [5].

The Mellin transform of the service $\mathcal{S}^{(T)}$ experienced by each user can be obtained by averaging over the Mellin transforms of the service process for the type A or B slots:

$$\mathcal{M}_{\mathcal{S}^{(T)}}(1-\theta) = p_{\mathsf{A}}\mathcal{M}_{\mathcal{S}^{(T)}|K_{\mathsf{A}}}(1-\theta) + p_{\mathsf{B}}\mathcal{M}_{\mathcal{S}^{(T)}|K_{\mathsf{B}}}(1-\theta)\,, \tag{68}$$

where the Mellin transform $\mathcal{M}_{\mathcal{S}^{(T)}|K}(\theta)$ for a specific number of users K is given by (20) for perfect CSI, or by (65) when considering imperfect CSI and finite blocklength effects.

REFERENCES

[1] A. Osseiran, F. Boccardi, V. Braun *et al.*, "Scenarios for 5G mobile and wireless communications: the vision of the METIS project," *IEEE Commun. Mag.*, vol. 52, no. 5, pp. 26–35, May 2014.

[2] 3GPP, "Study on communication for automation in vertical domains," Tech. Rep. 22.804, 2018. [Online]. Available: http://www.3gpp.org/DynaReport/22804.htm

[3] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[4] 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.

[5] S. Schiessl, J. Gross, and G. Caire, "Delay performance of the multiuser MISO downlink," *IEEE Global Commun. Conf. (GLOBECOM) (accepted). arXiv preprint arXiv:1805.01645*, 2018.

[6] 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, Sept 2004.

[7] 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.

[8] 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.

[9] T. Yoo, N. Jindal, and A. Goldsmith, "Multi-antenna downlink channels with limited feedback and user selection," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 7, 2007.

[10] 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.

[11] 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.

[12] J. Zhang, R. W. Heath, M. Kountouris, and J. G. Andrews, "Mode switching for the multi-antenna broadcast channel based on delay and channel quantization," *EURASIP J. Advances Signal Proc.*, vol. 2009, p. 1, 2009.

[13] 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.

[14] 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, 2010.

[15] E. Björnson, L. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems: Is massive MIMO the answer?" *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3059–3075, 2015.

[16] H. Q. Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, 2013.

[17] K. T. Truong and R. W. Heath, "Effects of channel aging in massive MIMO systems," *J. Commun. Networks*, vol. 15, no. 4, pp. 338–351, 2013.

[18] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, May 2010.

[19] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4243, Jul. 2014.

[20] J. Scarlett, V. Y. F. Tan, and G. Durisi, "The dispersion of nearest-neighbor decoding for additive non-gaussian channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 1, pp. 81–92, Jan 2017.

[21] 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.

[22] 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.

[23] 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.

[24] F. Ciucu, "Network calculus delay bounds in queueing networks with exact solutions," in *Int. Teletraffic Congr.* Springer, 2007, pp. 495–506.

[25] 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.

[26] 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.

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

[28] 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, June 2012.

[29] 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.

[30] J. Arnau and M. Kountouris, "Delay performance of MISO wireless communications," in *16th Int. Symp. Modeling Optim. in Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2018, pp. 1–8.

[31] M. Kountouris and A. Avranas, "Delay performance of multi-antenna multicasting in wireless networks," in *IEEE Int. Workshop Signal Process. Advances Wireless Commun. (SPAWC)*, June 2018, pp. 1–5.

[32] 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.

[33] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, 2007.

[34] 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.*, vol. 66, no. 12, pp. 6527–6541, Dec 2018.

[35] 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.

[36] M. Skoglund, J. Giese, and S. Parkvall, "Code design for combined channel estimation and error protection," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1162–1171, May 2002.

[37] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. Elsevier, 2007.

Sebastian Schiessl (S'16) received his Dipl.-Ing degree from Technical University of Munich, Germany. During his studies, he also stayed for one year at the University of Illinois at Urbana-Champaign. He joined the Royal Institute of Technology (KTH), Stockholm, Sweden in 2013, where he is working towards his PhD at the department of Information Science and Engineering. His theoretical research is focused on studying the queueing delay of low-latency wireless communication systems. In addition, he has also done some experimental implementations on the Wireless Open Access Research Platform (WARP).

James Gross (S'02–M'04–SM'15) received his Ph.D. degree from TU Berlin in 2006. From 2008-2012, he was Assistant Professor and Head of the Mobile Network Performance Group at RWTH Aachen University, as well as a member of the DFG-funded UMIC Research Centre of RWTH. Since November 2012, he has been with the Electrical Engineering School, KTH Royal Institute of Technology, Stockholm, as an Associate Professor. He also serves as Director for the ACCESS Linnaeus Centre and is a member of the board of KTHs Innovative Centre for Embedded Systems. His research interests are in the area of mobile systems and networks, with a focus on critical machine-to-machine communications, cellular networks, resource allocation, as well as performance evaluation methods. He has authored over 150 (peer-reviewed) papers in international journals and conferences. His work has been awarded multiple times, including best paper awards at ACM MSWiM 2015, the Best Demo Paper Award at IEEE WoWMoM 2015, the Best Paper Award at IEEE WoWMoM 2009, and the Best Paper Award at European Wireless 2009. In 2007, he was the recipient of the ITG/KuVS dissertation award for his Ph.D. thesis. He is also co-founder of R3 Communications GmbH, a Berlin-based start-up in the area of ultra-reliable low-latency wireless networking for industrial automation.

Mikael Skoglund (S'93–M'97–SM'04–F'19) received the Ph.D. degree in 1997 from Chalmers University of Technology, Sweden. In 1997, he joined the Royal Institute of Technology (KTH), Stockholm, Sweden, where he was appointed to the Chair in Communication Theory in 2003. At KTH, he heads the Division of Information Science and Engineering.

Dr. Skoglund has worked on problems in source-channel coding, coding and transmission for wireless communications, Shannon theory, information and control, and statistical signal processing. He has authored and co-authored some 150 journal and 350 conference papers.

Dr. Skoglund is a Fellow of the IEEE. During 2003–08 he was an associate editor for the IEEE Transactions on Communications and during 2008–12 he was on the editorial board for the IEEE Transactions on Information Theory. He has served on numerous technical program committees for IEEE sponsored conferences, he is general co-chair for IEEE ITW 2019, and he will serve as TPC co-chair for IEEE ISIT 2022.

Giuseppe Caire (S'92–M'94–SM'03–F'05) was born in Torino in 1965. He received the B.Sc. in Electrical Engineering from Politecnico di Torino in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992, and the Ph.D. from Politecnico di Torino in 1994. He has been a post-doctoral research fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands) in 1994-1995, Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, and he is currently an Alexander von Humboldt Professor with the Faculty of Electrical Engineering and Computer Science at the Technical University of Berlin, Germany.

He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society & Information Theory Society Joint Paper Award in 2004 and in 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, the Vodafone Innovation Prize in 2015, and an ERC Advanced Grant in 2018. Giuseppe Caire is a Fellow of IEEE since 2005. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications theory, information theory, channel and source coding with particular focus on wireless communications.