Delay Performance of Wireless Communications with Imperfect CSI and Finite Length Coding

Sebastian Schiessl, *Student Member, IEEE,* Hussein Al-Zubaidy, *Senior Member, IEEE,*
Mikael Skoglund, *Senior Member, IEEE,* and James Gross, *Senior Member, IEEE*

***Abstract*—With the rise of critical machine-to-machine applications, next generation wireless communication systems must meet challenging requirements with respect to latency and reliability. A key question in this context relates to channel state estimation, which allows the transmitter to adapt the code rate to the channel state. In this work, we characterize the trade-off between the training sequence length and data codeword length: shorter channel estimation leaves more time for the payload transmission but reduces the estimation accuracy and causes more decoding errors. Using lower coding rates can mitigate this effect, but may result in a higher backlog of data at the transmitter. In order to optimize the training sequence length and the rate adaptation scheme with respect to the delay performance, we employ queueing analysis on top of accurate models of the physical layer. We obtain an analytically tractable solution to the problem by deriving a closed-form approximation for the decoding error probability due to imperfect channel knowledge and finite blocklength channel coding. The optimized training sequence length and rate adaptation strategy can reduce the delay violation probability by an order of magnitude, compared to suboptimal strategies that do not consider the delay constraints.**

***Index Terms*—Finite blocklength regime, imperfect CSI, rate adaptation, quasi-static fading, queueing analysis**

I. Introduction

Traditionally, wireless networks have been optimized for the requirements of human-related applications such as voice communication as well as Internet applications. However, new applications based on machine-to-machine (M2M) communication emerged over the last decade, coming with strongly different requirements. M2M applications can be distinguished into two classes: *massive* and *critical*. Massive M2M applications relate typically to pure sensing and monitoring, for example sending utility meter readings to the utility provider. In contrast, critical M2M applications relate to closed-loop control and arise for example in the context of industrial automation. In this work, we focus exclusively on critical M2M applications. Critical M2M applications typically generate only small payload packets on a periodic basis. However, they often demand very low latency and extremely high reliability of the transmissions. For instance, applications from factory automation easily require communication latencies between a sensor and a control unit of at most a few milliseconds, as well as reliability (with respect to that deadline) of 99.9999% and above [1]. While lower reliability does not necessarily lead to costly faults, as additional safety layers, e.g., PROFIsafe [2], can be used, it leads to more frequent safety violations, which must often be resolved by human operators, resulting in significant downtimes. Therefore, we study in this paper how to design wireless communication systems for maximum reliability with respect to a given deadline.

Physical layer analysis of wireless communications has been frequently based on the assumption that error-free transmissions can be achieved through channel coding at Shannon's channel capacity, which is a fairly accurate model when the blocklength of the channel code is very large. However, with very short target latencies, systems will only be able to spend a small number of symbols per packet transmission, resulting in a significant performance loss due to finite blocklength coding [3]. The performance loss may be even larger when the signal strength on the physical layer is time-varying due to channel fading. An important question in this context relates to the most efficient transmission strategy, in particular, if and how the transmitter should estimate the channel state and adapt the coding rate according to the channel state information (CSI), which is imperfect due to the limited estimation time. On the one hand, long estimation sequences and low coding rates improve the reliability of the individual transmissions. On the other hand, when only little time is left for the actual payload transmissions and when the coding rates are low, then only small amounts of data can be transmitted. For a critical M2M system, the most efficient transmission strategy cannot be found by a pure physical layer analysis because the transmitter will be required to keep critical data in a buffer until the data is successfully received. Transmission errors and time-varying coding rates then lead to a random backlog of data at the transmitter, and thus, to a random queueing delay at the link layer. In conclusion, the transmission strategy that maximizes the overall reliability with respect to the deadline can only be determined through a combination of accurate physical layer modeling and queueing analysis.

As our work combines aspects from the physical layer and the link layer, we build on literature from both information theory and queuing theory. Concerning research in information theory, the analysis of the theoretical limits of finite blocklength channel coding by Polyanskiy et al. [3] was extended by Yang et al. to block-fading channels [4]–[6]. A summary of these works, which have inspired a number of new research papers, can be found in [7]. Wu and Jindal [8] and Makki et al. [9], [10] analyzed communications in block-fading channels when (hybrid) automatic repeat request (ARQ)

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

H. Al-Zubaidy is with the School of Information Technology, Halmstad University, Halmstad, Sweden, (email: hussein.al-zubaidy@hh.se).

protocols are used, or in spectrum sharing networks [11], [12]. Those works account for finite blocklength effects, but mostly consider the case where the transmitter does not adapt the rate to the channel. Various authors [11], [13]–[16] have studied communications in block-fading channels with rate adaptation or power/resource allocation at the transmitter, but they only took finite blocklength effects into account and considered CSI at the transmitter (CSIT) to be perfect. Lim and Lau [17] and Lau et al. [18] studied rate adaptation with imperfect or outdated CSIT, but considered neither finite blocklength effects nor the impact of transmission errors on the delay. Instead, researchers have often focused on imperfect CSI at the receiver (CSIR), which causes an error in the amplitude and phase of the signal during demodulation and decoding, which in turn can lead to decoding errors. Médard [19] studied the mutual information under imperfect CSIR, and Hassibi and Hochwald [20] analyzed the ergodic capacity of systems with channel training. Potter et al. [21] studied the achievable rate under imperfect CSIR and finite blocklength coding, assuming mismatched decoding for the achievability bound. However, none of these works on imperfect CSIR investigated rate adaptation with imperfect CSIT.

In the field of communication networks, queueing theory has been used extensively to analyze the delay performance of wireless networks. While wireless network analysis poses a significant challenge to traditional queuing theory, several techniques have been developed to address this challenge. Wu and Negi [22] presented the framework of effective capacity that provides approximations on the delay performance, which are however asymptotic, i.e. only valid for long delays. In contrast, stochastic network calculus [23] provides non-asymptotic bounds on the delay performance, and can also be extended for the analysis of multi-hop wireless links [24], [25]. Finite blocklength effects and imperfect CSI have been separately studied with respect to their impact on the queueing performance. Gursoy [26] computed the effective capacity for block-fading channels at finite blocklength and showed that there is a unique optimum for the error probability. In our own previous work [27], we extended this analysis using stochastic network calculus and provided analytical solutions for finite blocklength coding in Rayleigh block-fading channels. Ozcan and Gursoy [28] studied the effective capacity of cognitive radio systems at finite blocklength. Nevertheless, none of these works considered imperfect CSIT. The queueing performance under rate adaptation with imperfect/outdated CSIT but without finite blocklength effects was analyzed by Gross [29].

In this work, we address the delay performance of a wireless communication system with rate adaptation based on imperfect CSI. We provide three main contributions:

- Based on stochastic network calculus [23], [24], we characterize the trade-off between the rate and the error probability with respect to the delay performance as a convex optimization problem. Thus, the transmitter can efficiently optimize the transmission rate, taking the overall latency and reliability constraints of the critical M2M application into account. We also show that the maximization of the average service rate as well as the maximization of the effective capacity become efficiently solvable optimization problems.
- All these optimization results are based on our novel closed-form approximation for the error probability due to the combined effects of finite blocklength coding and imperfect CSI at the transmitter in Rayleigh fading channels. Specifically, we derive an approximation for an information-theoretic result from Yang et al. [6]. A key challenge that we overcome in this derivation is that finite blocklength effects are modeled as variations in the rate, whereas imperfect CSI corresponds to variations in the SNR. Our approximation is invertible, providing a direct mapping from the error probability to the rate.
- Our numerical results show that rate adaption, despite needing a large fraction of the resources for channel training and feedback, can significantly outperform fixed rate transmissions when low latency is required. Moreover, we find that through an optimized rate adaptation strategy and through an optimized choice of training duration versus payload transmission time, the system can improve the overall reliability by one order of magnitude. Lastly, we find that when rate adaptation is used, finite blocklength coding has a significant impact on the performance, contrasting results in [6] for fixed rate communications.

This paper is organized as follows: The system model is given in Sec. II. Our main contributions are presented in Sec. III. Numerical results are then presented in Sec. IV, followed by our conclusions in Sec. V.

Throughout the paper, we utilize the following notation: Uppercase italic letters X generally refer to random variables, whereas the corresponding lowercase letters x refer to a realization of that random variable. We write $f_{X|y}(x)$ for the probability density function (PDF) and $F_{X|y}(x)$ for the cumulative distribution function (CDF) of the variable X, conditioned on the value $Y = y$. The indicator function for the set $\{X < x\}$ is denoted as $\mathbb{1}_{X<x}$. For complex-valued x, $\Re\{x\}$ is the real part, $\angle(x)$ is the phase, and x^* is the complex conjugate. When it is necessary to consider multiple time slots, we label the variables related to a specific time slot i by the subscript i.

II. System Model

We consider the transmission of a time-critical data flow that is generated by an application process on the transmitter side, e.g., by a sensor, to an application process on the receiver side, e.g., to a control unit, over a wireless fading channel, see also Fig. 1. The delay of the data flow on the application layer, which depends on the underlying physical and link layers, should not exceed a given target deadline, which we assume in the following to be very short (e.g., a few milliseconds). In the assumed time-slotted system, the tight deadline translates into a small number of time slots (e.g. 5 or 10). As each time slot is very short, both imperfect channel state information (CSI) and finite blocklength channel coding lead to transmission errors on the physical layer, as described in Sec. II-A. Due to transmission errors and time-varying data rates on the physical layer, the transmitter needs to keep all data in a buffer

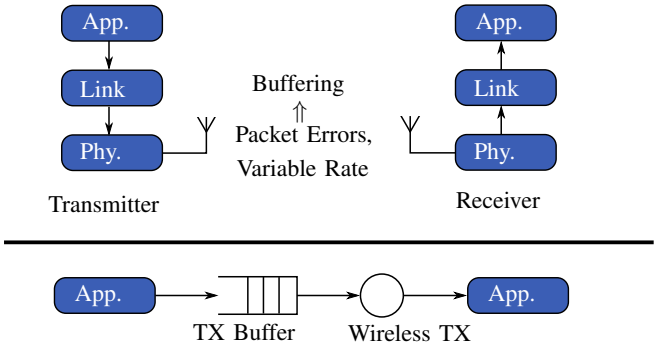


Fig. 1. Schematic illustration of the assumed system model.

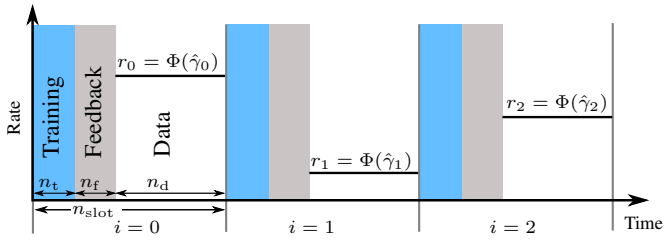


Fig. 2. Illustration of the time slot structure: after the training, the transmitter receives the SNR estimate $\hat{\gamma}$ as feedback, and then transmits a codeword at a rate determined by the rate adaptation function $\Phi : \hat{\gamma} \to r$.

until successful transmission, which causes a random queueing delay at the link layer, as outlined in Sec. II-B. Finally, in Sec. II-C, we formulate the problem of determining the system parameters that maximize the reliability with respect to the deadline of the data flow.

A. Physical Layer Model

A frequency-flat Rayleigh block-fading channel model is considered, where the channel remains constant for the duration of one time slot and changes independently from slot to slot. This model applies for example to systems that perform frequency hopping after each time slot. We consider a single-antenna system, where the fading coefficient H is scalar and has circularly symmetric Gaussian distribution $\mathcal{CN}(0,1)$. The instantaneous SNR at the receiver is given as $\Gamma = \bar{\gamma}|H|^2$ and has exponential distribution with mean $\bar{\gamma}$. The average SNR $\bar{\gamma}$ is constant and known at the transmitter and the receiver.

Each time slot contains n_{slot} symbols and is assumed to be split into three phases as shown in Fig. 2: the training/estimation phase, where the transmitter sends a known training sequence of n_{t} symbols; a feedback phase of n_{f} symbols where the receiver sends an estimate $\hat{\Gamma}$ of the channel's SNR back to the transmitter; and the data transmission phase, where the transmitter sends a codeword of length n_{d}. The rate R of the code is determined by some rate adaptation function $\Phi : \hat{\Gamma} \to R$, which remains fixed over time. We now describe the phases in detail.

1) Training Phase: The receiver estimates the fading coefficient H through a training sequence of n_{t} symbols. The minimum mean square error (MMSE) estimate for H is given as [20], [30]:

$$\hat{H} = \frac{\bar{\gamma} n_{\text{t}}}{1+\bar{\gamma} n_{\text{t}}} H + \frac{\sqrt{\bar{\gamma} n_{\text{t}}}}{1+\bar{\gamma} n_{\text{t}}} N \,, \tag{1}$$

where $N \sim \mathcal{CN}(0,1)$ is independent of H. Therefore, the channel estimate $\hat{H}$ is Gaussian distributed as $\hat{H} \sim \mathcal{CN}\left(0, \rho^2\right)$ with $\rho^2 = \bar{\gamma} n_{\text{t}}/(1+\bar{\gamma} n_{\text{t}})$, while the estimated SNR $\hat{\Gamma} \triangleq \bar{\gamma}|\hat{H}|^2$ follows an exponential distribution with mean $\rho^2\bar{\gamma}$. Due to H and $\hat{H}$ being jointly Gaussian, the distribution of H conditioned on the estimate $\hat{H}$ can be expressed as [30]

$$H = \hat{H} + Z \,, \tag{2}$$

where the estimation error $Z \sim \mathcal{CN}(0, \sigma_Z^2)$ is independent of $\hat{H}$ with

$$\sigma_Z^2 = \frac{1}{1+\bar{\gamma} n_{\text{t}}} \,. \tag{3}$$

2) Feedback Phase: After the training phase, the receiver sends the estimated coefficient $\hat{H}$ as feedback of length n_{f} symbols to the transmitter. The phase $\angle(\hat{H})$ of the channel coefficient is not used at the transmitter, so the feedback only needs to contain the magnitude of $\hat{H}$ (or equivalently, the estimated SNR $\hat{\Gamma} = \bar{\gamma}|\hat{H}|^2$). We assume for now that the feedback is sufficiently quantized (we will investigate quantization issues in Sec. IV-B) and error-free[1]. Finally, the feedback also includes an acknowledgment of the previous data transmission.

3) Data Transmission Phase: Once the transmitter knows the channel estimate $\hat{\Gamma}$, it selects a code rate R according to the given rate adaptation function $\Phi : \hat{\Gamma} \to R$, and transmits a codeword of n_{d} symbols. In the following subsection, we consider specific realizations $\hat{\gamma}$ and r of the random variables $\hat{\Gamma}$ and R, respectively.

First of all, consider the hypothetical case of perfect CSI (PCSI), i.e., the exact value of the SNR γ is known to the transmitter and receiver. Then, the channel in each time slot is an AWGN channel. Due to the short duration of each time slot, only channel codes with very short blocklength can be used. Therefore, traditional models, which often assume that channel coding provides error-free transmissions at the Shannon capacity, become inaccurate. Instead, one must use results on finite blocklength coding [3]. The achievable rate in an AWGN channel with blocklength n_{d}, SNR γ and error probability $\varepsilon_{\text{PCSI}}$ can be closely approximated by [3, Thm. 54]

$$r_{\text{PCSI,FBL}}(\gamma, n_{\text{d}}, \varepsilon_{\text{PCSI}}) \approx \log_2(1+\gamma) - \sqrt{\frac{\mathcal{V}(\gamma)}{n_{\text{d}}}} Q^{-1}(\varepsilon_{\text{PCSI}}) \,, \tag{4}$$

where $Q(\cdot)$ is the Gaussian Q-function and $\mathcal{V}(\gamma)$ is the channel dispersion, defined as

$$\mathcal{V}(\gamma) = \log_2^2(e)\left(1 - \frac{1}{(1+\gamma)^2}\right) \,. \tag{5}$$

In case of perfect CSI, the error probability can then be found by solving (4) for $\varepsilon_{\text{PCSI}}$:

$$\varepsilon_{\text{PCSI}} \approx Q\left(\frac{\log_2(1+\gamma) - r}{\sqrt{\mathcal{V}(\gamma)/n_{\text{d}}}}\right) \,. \tag{6}$$

[1]As shown in Sec. IV-B, only a small number of feedback bits is required. Thus, a code with very low coding rate can be used, such that the probability of feedback errors becomes negligible.

However, in a realistic scenario, the transmitter only has a noisy estimate $\hat{\gamma}$ of the SNR. The actual SNR Γ can be below or above the estimate $\hat{\gamma}$. In order to obtain the overall error probability in this case, it is necessary to take the expectation of the error probability over the conditional distribution of the actual SNR Γ given the estimated SNR $\hat{\gamma}$:

$$\varepsilon \approx \mathbb{E}_{\Gamma|\hat{\gamma}}\left[Q\left(\frac{\log_2(1+\Gamma)-r}{\sqrt{\mathcal{V}(\Gamma)/n_{\mathrm{d}}}}\right)\right] . \tag{7}$$

To derive this conditional distribution of Γ, note that the distribution of H conditioned on a known $\hat{h}$ is Gaussian with $H = \hat{h}+Z$, according to (2). The SNR $\Gamma = \bar{\gamma}|H|^2$ conditioned on the estimate $\hat{\gamma}$ then follows a non-central χ^2-distribution with two degrees of freedom and PDF:

$$f_{\Gamma|\hat{\gamma}}(x) = \frac{1}{\bar{\gamma}\cdot\sigma_Z^2}\cdot e^{-\frac{x+\hat{\gamma}}{\bar{\gamma}\cdot\sigma_Z^2}}\cdot I_0\left(\frac{2\sqrt{x\hat{\gamma}}}{\bar{\gamma}\cdot\sigma_Z^2}\right) , \tag{8}$$

where $I_0(\cdot)$ is the modified Bessel function of the first kind. We state two important remarks concerning (7):

(I) When conditioned on a specific estimate $\hat{h}$, the channel coefficient H has the same distribution as the fading coefficient of a Rician channel with line-of-sight (LoS) component $\hat{h}$ and non-LoS component Z. Thus, (7) corresponds to the normal approximation for fading channels by Yang et al. [6, (59-61)].[2] However, when applying this approximation, one must carefully differentiate between CSIT and CSIR. The authors in [6] assumed perfect CSIR (perfect knowledge of $\hat{h}$, Z, and H at the receiver) when proposing the approximation (7), whereas in our scenario, the receiver only knows $\hat{h}$ (imperfect CSIR), and thus (7) could be inaccurate for our scenario. In Appendix A, we verify the accuracy of (7) by computing a lower bound [5, Cor. 3] on the achievable rate under imperfect CSIR. For the considered parameters, the system model using (7) approximates the achievable performance well, i.e., imperfect CSIR has at most a small impact on the performance when using a reasonably long training sequence, e.g., $n_{\mathrm{t}} \geq 10$.

(II) The finite blocklength result (7) is an approximation and may become inaccurate for very short blocklengths n_{d} or for extremely low block error probabilities ε [3], [6], [7]. We will further discuss this issue in Sec. IV-A. For now, we assume that (7) holds with equality in order to simplify discussions.

B. Link Layer Model

In order to characterize the application-layer delay of a wireless communication system, one must account for the consequences of transmission errors and low data rates on the transmit buffer. As data is only removed from the buffer once an acknowledgment indicates the correct reception, the buffering delay is random and needs to be characterized through queueing analysis.

The buffer at the transmitter can be described by its arrival, service and departure processes. The arrival process A_i describes the number of bits that enter the transmit buffer in time slot i. We are interested in data flows as arising in industrial contexts, e.g., periodic transmissions of sensor readings, and we thus assume that the arrival process is constant with $A_i = \alpha$ bits per time slot. The service process S_i describes the number of bits that can potentially be transmitted over the wireless channel: in time slot i, the transmitter knows the random channel estimate $\hat{\Gamma}_i$, and, given the rate adaptation function Φ, the transmitter sends a codeword containing $n_{\mathrm{d}}\cdot R_i$ bits with $R_i = \Phi(\hat{\Gamma}_i)$. The transmission error probability given in (7) is now denoted by the uppercase letter $\mathcal{E}_i$, as it depends on both $\hat{\Gamma}_i$ and $\Phi(\hat{\Gamma}_i)$, and is therefore random. Thus,

$$S_i = \begin{cases} n_{\mathrm{d}}\cdot\Phi(\hat{\Gamma}_i) & \text{with prob.} \quad (1-\mathcal{E}_i) \\ 0 & \text{with prob.} \quad \mathcal{E}_i \end{cases} . \tag{9}$$

The departure process D_i is given as the number of bits leaving the queue, which is equal to the number of bits that reach the receiver successfully. The departure process is upper-bounded both by the service process and by the amount of data waiting in the queue. For the delay analysis, we also define the cumulative arrival, service, and departure processes

$$\mathbf{A}(\tau,t) \triangleq \sum_{i=\tau}^{t-1} A_i , \quad \mathbf{S}(\tau,t) \triangleq \sum_{i=\tau}^{t-1} S_i , \quad \mathbf{D}(\tau,t) \triangleq \sum_{i=\tau}^{t-1} D_i . \tag{10}$$

As the size of the arriving data packets is small, we assume that the buffer never gets full, i.e., that the buffer is of infinite size. The random delay $W(i)$ at time i is defined as the time it takes for all data that arrived prior to time i to depart from the transmit buffer, i.e., to be correctly decoded at the receiver:

$$W(i) \triangleq \inf\{u \geq 0 : \quad \mathbf{A}(0,i) \leq \mathbf{D}(0,i+u)\} . \tag{11}$$

The application data must be transmitted with high reliability within a deadline of w time slots. The delay violation probability $p_{\mathrm{v}}(w)$ describes the probability that at any time i the random delay $W(i)$ exceeds the target delay w:

$$p_{\mathrm{v}}(w) \triangleq \sup_{i\geq 0}\{\mathbb{P}\{W(i) > w\}\} . \tag{12}$$

C. Problem Statement

The main objective of this work is to determine the optimal system parameters, specifically the optimal rate adaptation function Φ and training sequence length n_{t}, such that for a given maximum delay (deadline) w, the delay violation probability $p_{\mathrm{v}}(w)$ is minimized. First, we are interested in the rate adaptation function $\Phi : \hat{\Gamma} \to R$ that provides the best performance. Choosing high rates R will lead to high error probabilities $\mathcal{E}$, and too many errors may cause a violation of the deadline. On the other hand, the deadline may also be violated if the transmitter always chooses a very low rate, as it will take longer to transmit the data. Second, there is a similar trade-off between the length of the training sequence n_{t} and the length of data transmission n_{d}: when using a long training sequence of n_{t} symbols, the channel estimates become more accurate, allowing transmissions with higher reliability but leaving fewer symbols ($n_{\mathrm{d}} = n_{\mathrm{slot}} - n_{\mathrm{f}} - n_{\mathrm{t}}$) for the data transmission. This trade-off becomes particularly interesting when short transmission slots are considered: due

[2]The authors in [6] use natural logarithms, whereas we use binary logarithms and express the rate in bits.

to finite blocklength channel coding, the shortening of the payload transmission deteriorates the communication performance even more rapidly. Does it even make sense to use training in all scenarios, or is it sometimes better to skip training and feedback ($n_\mathrm{t} = 0$, $n_\mathrm{f} = 0$) and use the entire time slot for data transmissions at a fixed rate? This leads us to an initial formulation of the reliability maximization problem given by

$$[n_\mathrm{t}^*, \Phi^*] = \underset{n_\mathrm{t}, \Phi}{\arg\min} \quad p_\mathrm{v}(w)\,. \tag{13}$$

We note that an analytical expression for the delay violation probability $p_\mathrm{v}(w)$ does not exist. However, $p_\mathrm{v}(w)$ can be captured analytically by resorting to effective capacity [22] or stochastic network calculus [23]. Although effective capacity has been successfully applied in numerous works, e.g., [26], [31], it only provides a tight approximation for the tail of the delay distribution $p_\mathrm{v}(w)$, and hence only allows for a proper analysis of relatively large delays w. Contrary to that, stochastic network calculus [23], which has been recently extended to wireless network analysis in a transform domain [24] and which has also been applied to various scenarios [25], [27], [32], [33], provides a strict upper bound on the delay violation probability $p_\mathrm{v}(w)$, even at low target delays. In the following chapter, we thus focus on solving the optimization problem (13) analytically based on stochastic network calculus.

III. Joint Analysis of Imperfect CSIT and Finite Length Coding

In this section, we present our main contributions. We briefly summarize stochastic network calculus in Sec. III-A and use it in Sec. III-B to reformulate the search for the optimal rate adaptation function Φ^*. The biggest challenge in solving this problem is the fact that no closed-form representation of the transmission error probability due to finite blocklength and imperfect CSIT exists.[3] Thus, we derive in Sec. III-C a novel closed-form approximation of the error probability due to imperfect CSIT and finite blocklength. This allows us then in Sec. III-D to address the optimal trade-off between the rate and the error probability as a problem that is convex in the error probability. In Sec. III-E, we show that the approximation for the error probability can also be used to maximize the effective capacity [22] and the expected goodput.

A. Queueing Analysis

This subsection summarizes how the delay performance, specifically the delay violation probability $p_\mathrm{v}(w)$, can be analyzed through stochastic network calculus [23], [24]. Parts of this summary are taken from our previous work in [27].

The delay $W(i)$ in (11) is defined in terms of the arrival and departure processes. However, for finding the statistical distribution of the delay, it is easier to use only the arrival and service processes. The authors in [24] characterized these processes in the exponential domain, also referred to as *SNR domain*. The cumulative arrival and service processes in the bit domain, $\mathbf{A}(\tau, t)$ and $\mathbf{S}(\tau, t)$, are converted to the SNR domain (denoted by calligraphic letters) as follows:

$$\boldsymbol{\mathcal{A}}(\tau, t) \triangleq e^{\mathbf{A}(\tau,t)}, \quad \boldsymbol{\mathcal{S}}(\tau, t) \triangleq e^{\mathbf{S}(\tau,t)}\,. \tag{14}$$

Similarly, we define $\mathcal{A}_i \triangleq e^{A_i}$ and $\mathcal{S}_i \triangleq e^{S_i}$. According to the system model, both A_i and S_i are independent and identically distributed (i.i.d.) between time slots, and we can thus drop the subscript i. An upper bound on the delay violation probability $p_\mathrm{v}(w)$ can then be computed 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 [24]

$$\mathcal{M}_\mathcal{X}(\theta) \triangleq \mathbb{E}\left[\mathcal{X}^{\theta-1}\right] \tag{15}$$

for a parameter $\theta \in \mathbb{R}$. For the analysis, choose $\theta > 0$ and check whether the stability condition

$$\mathcal{M}_\mathcal{A}(1+\theta)\mathcal{M}_\mathcal{S}(1-\theta) < 1 \tag{16}$$

holds. If it holds, define the kernel [24], [27]

$$\mathcal{K}(\theta, w) \triangleq \lim_{i\to\infty} \sum_{u=0}^{i} \mathcal{M}_\mathcal{A}(1+\theta)^{i-u}\mathcal{M}_\mathcal{S}(1-\theta)^{i+w-u} \tag{17}$$

$$= \frac{\mathcal{M}_\mathcal{S}(1-\theta)^w}{1-\mathcal{M}_\mathcal{A}(1+\theta)\mathcal{M}_\mathcal{S}(1-\theta)}\,. \tag{18}$$

For any parameter $\theta > 0$ that satisfies the stability condition, the kernel $\mathcal{K}(\theta, w)$ provides an upper bound on the probability $p_\mathrm{v}(w)$ that the delay exceeds the target delay w. This holds for any time slot i, including the limit $i \to \infty$ (steady-state). In order to find the tightest upper bound, one must find the parameter $\theta > 0$ that minimizes $\mathcal{K}(\theta, w)$ [24], [27]:

$$p_\mathrm{v}(w) \leq \inf_{\theta>0} \left\{\mathcal{K}(\theta, w)\right\}\,. \tag{19}$$

B. Parameter Optimization Problem

1) Revised Problem Statement: While the original parameter optimization problem (13) with respect to $p_\mathrm{v}(w)$ cannot be solved analytically, we propose to perform the parameter optimization based on the analytical bound (19) on $p_\mathrm{v}(w)$:

$$[n_\mathrm{t}^*, \Phi^*] \approx \underset{n_\mathrm{t}, \Phi}{\arg\min} \quad \inf_{\theta>0} \left\{\mathcal{K}(\theta, w, n_\mathrm{t}, \Phi)\right\}\,, \tag{20}$$

where we explicitly denote that the kernel function $\mathcal{K}(\cdot)$ defined in (17) depends on n_t and Φ through the service process. We will investigate the difference between (13) and (20) numerically in Sec. IV-A. The service process S in (9) can be described as $S = n_\mathrm{d} \cdot R \cdot Y$, where $R = \Phi(\hat{\Gamma})$ is the code rate adapted to the estimated SNR $\hat{\Gamma}$ and Y is a Bernoulli random variable, which is zero in case of a transmission error, i.e., with probability $\mathcal{E}$ [27]. The delay bound depends on the Mellin transform $\mathcal{M}_\mathcal{S}(\theta)$ of the service process in the SNR domain $\mathcal{S} = e^S$, which is given as

$$\mathcal{M}_\mathcal{S}(\theta) = \mathbb{E}\left[\mathcal{S}^{\theta-1}\right] = \mathbb{E}\left[e^{n_\mathrm{d}\Phi(\hat{\Gamma})\cdot Y\cdot(\theta-1)}\right] \tag{21}$$

$$= \mathbb{E}\left[(1-\mathcal{E})e^{n_\mathrm{d}\Phi(\hat{\Gamma})\cdot(\theta-1)} + \mathcal{E}\right]\,, \tag{22}$$

[3] For Rayleigh fading channels without rate adaptation, one can approximate (7) by the outage probability, which was shown to be accurate up to $\mathcal{O}(\log(n)/n)$ by Yang et al. [6]. However, this does not hold when the transmitter adapts the rate to an estimate of the instantaneous channel, as the distribution of the SNR Γ conditioned on an accurate estimate $\hat{\gamma}$ is no longer "smooth" as required in [6]. For the same reason, the approximations by Makki et al. [9], [11] do not apply here.

where the expectation is taken with respect to $\hat{\Gamma}$, with $\hat{\Gamma} \sim \exp\left(\rho^2\bar{\gamma}\right)$ as discussed in Sec. II-A1. Keep in mind that the error probability $\mathcal{E}$ given by (7) is not constant, as it depends on the estimated SNR $\hat{\Gamma}$ and on the selected rate $R = \Phi(\hat{\Gamma})$.

2) Solution Approach: Let us now consider the nature of the considered parameter optimization problem with respect to the training sequence length n_t and the rate adaptation function Φ. The optimal training sequence length n_t^* can be found quickly: Assume that $n_\mathrm{t}^\mathrm{loc}$ is the first local minimum, i.e., $n_\mathrm{t}^\mathrm{loc}$ is the smallest value such that the delay violation probability at $n_\mathrm{t}^\mathrm{loc}$ is smaller than it is at the adjacent values $n_\mathrm{t}^\mathrm{loc}-1$ or $n_\mathrm{t}^\mathrm{loc}+1$. Now, we increase n_t in two steps. First, we increase from $n_\mathrm{t}^\mathrm{loc}$ to $n_\mathrm{t}^\mathrm{loc}+1$. The delay violation probability has deteriorated. Therefore, we know that the performance gain due to the extra training symbol did not compensate for the loss due to the reduced codeword length. Second, we consider the increase by one more symbol from $n_\mathrm{t}^\mathrm{loc}+1$ to $n_\mathrm{t}^\mathrm{loc}+2$. The *relative* increase in n_t is smaller for the second increase than for the first increase, while the relative reduction in codeword length is greater. Thus, the delay violation probability must increase even further at $n_\mathrm{t}^\mathrm{loc}+2$. This argument can be extended to all values $n_\mathrm{t} > n_\mathrm{t}^\mathrm{loc}$, which means that the delay violation probability must keep strictly increasing for $n_\mathrm{t} > n_\mathrm{t}^\mathrm{loc}$ Thus, $n_\mathrm{t}^\mathrm{loc}$ must be identical to the global optimum n_t^*, and the optimal value of n_t can be found through fast optimization techniques.

Now, we assume that n_t is fixed and address the problem of finding the optimal rate adaptation function Φ^*. The optimization problem (20) requires an iteration over the parameter $\theta > 0$. We quantize the range[4] of θ and find the optimal rate adaptation function Φ_θ^* for each quantized value of θ. Afterwards, we can determine the argument θ^* that minimizes the bound $\mathcal{K}\left(\theta, w, n_\mathrm{t}, \Phi_\theta^*\right)$ over θ, and the optimal rate adaptation function is given as $\Phi^* = \Phi_{\theta^*}^*$. For a specific θ, the function Φ_θ^* that minimizes the bound (19) on $p_\mathrm{v}(w)$ is given as

$$\Phi_\theta^* = \arg\min_{\Phi} \ \mathcal{K}\left(\theta, w, n_\mathrm{t}, \Phi\right) \tag{23}$$

$$= \arg\min_{\Phi} \ \mathcal{M}_\mathcal{S}(1-\theta) \tag{24}$$

where we used the fact that the kernel $\mathcal{K}\left(\theta, w, n_\mathrm{t}, \Phi\right)$ is monotonically increasing in $\mathcal{M}_\mathcal{S}(1-\theta)$. In order to find Φ_θ^*, we rewrite (22) as

$$\mathcal{M}_\mathcal{S}(1-\theta) = \int_0^\infty \underbrace{\left((1-\varepsilon)e^{n_\mathrm{d}\cdot\Phi(\hat{\gamma})\cdot(-\theta)} + \varepsilon\right)}_{\triangleq\, g(\hat{\gamma},\Phi(\hat{\gamma}))} f_{\hat{\Gamma}}(\hat{\gamma}) d\hat{\gamma} \ . \tag{25}$$

To find the rate adaptation function Φ_θ^* that minimizes $\mathcal{M}_\mathcal{S}(1-\theta)$, we quantize the range of $\hat{\gamma}$. Then, for each value of $\hat{\gamma}$, we need to find the rate r that minimizes $g(\hat{\gamma}, r)$. However, this rate optimization problem is not tractable analytically because the error probability ε in (7) depends on both $\hat{\gamma}$ and $\Phi(\hat{\gamma})$, and is only given in form of an integral, i.e., not in closed-form. In order to find a tractable solution to this problem, we develop in the following subsection a closed-form approximation for the error probability ε in (7). Then, we show in Sec. III-D that when using the closed-form approximation, the search for the minimum in $g(\hat{\gamma}, r)$ becomes a convex problem, and thus the optimal rate adaptation function Φ^* can be found efficiently.

[4]Only a limited range of $0 < \theta < \theta^\mathrm{max}$ must be considered as the stability condition (16) does not hold for very large values of θ: $\mathcal{M}_\mathcal{A}(1+\theta)$ grows exponentially with θ, while $\mathcal{M}_\mathcal{S}(1-\theta)$ remains larger than the average error probability $\mathbb{E}\left[\mathcal{E}\right] > 0$.

C. Error Probability at Imperfect CSIT and Finite Blocklength

In order to analyze the error probability ε due to the combined impact of imperfect CSIT and finite blocklength coding, we will first derive an approximation for the Shannon outage probability, i.e., for the probability of errors which are only due to imperfect CSIT. Afterwards, we will show that the error probability ε due to imperfect CSIT and finite blocklength coding can be expressed in a form that is similar to an outage probability. We can then use the insights gained from the outage probability analysis to derive an approximation for the error probability.

1) Outage Probability Approximation for Imperfect CSIT: For now, we ignore finite blocklength effects, i.e., we assume an infinite blocklength model, and focus only on the effects of imperfect CSIT. We consider specific realizations of the estimated SNR $\hat{\gamma}$ and the rate r. When the blocklength n_d of the channel code tends to infinity, the Gaussian Q-function in (7) converges to $\mathbb{1}_{\log_2(1+\Gamma)<r}$ and the error probability (7) converges to the conditional Shannon outage probability

$$\varepsilon_\mathrm{out} \triangleq \mathbb{P}\left\{\log_2(1+\Gamma) < r \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right\} \ . \tag{26}$$

Outages occur because the transmitter must choose a rate r without knowing the SNR Γ exactly. The conditional outage probability can be determined from the cumulative distribution of Γ conditioned on $\hat{\gamma}$, which is given in terms of the Marcum Q-function $Q_1(a,b)$ [29], [34]:

$$F_{\Gamma|\hat{\gamma}}(x) = 1 - Q_1\left(\sqrt{\frac{2\hat{\gamma}}{\bar{\gamma}\sigma_Z^2}}, \sqrt{\frac{2x}{\bar{\gamma}\sigma_Z^2}}\right) \ . \tag{27}$$

However, evaluating the Marcum Q-function is fairly complex, which makes it difficult to find the optimal trade-off between the rate r and the error probability even in the case where the blocklength tends to infinity. We thus provide an upper bound for the outage probability based on the Gaussian Q-function.

Lemma 1. *Given an imperfect estimate of the channel $\hat{\gamma}$ and a rate r, the outage probability (i.e., the error probability when $n_\mathrm{d} \to \infty$) is bounded by*

$$\varepsilon_\mathrm{out} \leq Q\left(\frac{\hat{\gamma} - (2^r - 1)}{\sigma_\mathrm{ICSI}}\right) \ , \tag{28}$$

with $\sigma_\mathrm{ICSI}^2 \triangleq 2\sigma_Z^2\bar{\gamma}\hat{\gamma}$.

Proof. The random fading coefficient H is given in terms of the known measurement $\hat{h}$ and the estimation error Z according to (2). Thus:

$$\Gamma = \bar{\gamma}|H|^2 = \bar{\gamma}(\hat{h}+Z)(\hat{h}^*+Z^*) \tag{29}$$

$$= \hat{\gamma} + 2\bar{\gamma}\Re\left\{\hat{h}^* Z\right\} + \bar{\gamma}\left|Z\right|^2 \tag{30}$$

$$= \hat{\gamma} + 2\bar{\gamma}|\hat{h}|\Re\left\{\bar{Z}\right\} + \bar{\gamma}\left|\bar{Z}\right|^2 \tag{31}$$

where $\bar{Z} = e^{-j\angle(\hat{h})}Z$ is a phase-rotated version of Z. The distribution and the magnitude of a circularly symmetric random variable stay constant under phase rotation, and thus the real part $\Re\left\{\bar{Z}\right\}$ has Gaussian distribution $\mathcal{N}(0, \sigma_Z^2/2)$. It follows that the SNR $\Gamma = \bar{\gamma}\left|H\right|^2$ is given as

$$\Gamma = \hat{\gamma} + \tilde{\Gamma}_{\mathrm{G}} + \tilde{\Gamma}_{\delta} = \hat{\gamma} + \tilde{\Gamma}\,, \tag{32}$$

i.e. the estimation error $\tilde{\Gamma} = \Gamma - \hat{\gamma}$ is the sum of a Gaussian error $\tilde{\Gamma}_{\mathrm{G}} \sim \mathcal{N}(0, \sigma^2_{\mathrm{ICSI}})$ and some $\tilde{\Gamma}_{\delta} = \bar{\gamma}\left|\bar{Z}\right|^2$. The outage probability $\varepsilon_{\mathrm{out}}$ can then be bounded as:

$$\begin{aligned}
\varepsilon_{\mathrm{out}} &= \mathbb{P}\left\{\Gamma < 2^r - 1 \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right\} && (33)\\
&\leq \mathbb{P}\left\{\hat{\gamma} + \tilde{\Gamma}_{\mathrm{G}} < 2^r - 1\right\} && (34)\\
&= \mathbb{P}\left\{-\frac{\tilde{\Gamma}_{\mathrm{G}}}{\sigma_{\mathrm{ICSI}}} > \frac{\hat{\gamma} - (2^r - 1)}{\sigma_{\mathrm{ICSI}}}\right\} && (35)
\end{aligned}$$

where the inequality holds because $\tilde{\Gamma}_{\delta} \geq 0$. □

The variance of $\tilde{\Gamma}_{\delta}$ is proportional to σ_Z^4 and thus $\tilde{\Gamma}_{\delta}$ becomes small relative to $\tilde{\Gamma}_{\mathrm{G}}$ as the channel estimates become more accurate. In that case, the estimation error is approximately Gaussian. This fact will simplify the joint analysis of imperfect CSI and finite length coding.

2) Combined Analysis of Imperfect CSIT and Finite Blocklength: We now derive an approximation for the transmission error probability ε in (7) that depends on both imperfect CSIT and finite blocklength effects. In order to combine both effects, we use the following definition, which allows treating errors due to finite blocklength coding equivalently to outage events: we define the *random blocklength-equivalent capacity*

$$C_{\mathrm{b}} \triangleq \log_2(1+\Gamma) - \sqrt{\frac{\mathcal{V}(\Gamma)}{n_{\mathrm{d}}}} \cdot U_{\mathrm{FBL}} \tag{36}$$

with $U_{\mathrm{FBL}} \sim \mathcal{N}(0,1)$ independent of Γ. The *blocklength-equivalent outage probability* $\mathbb{P}\{C_b < r\}$, conditioned on the estimate $\hat{\gamma}$, is equal to the error probability ε in (7) because

$$\begin{aligned}
\mathbb{P}\left\{C_b < r \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right\} &= \mathbb{E}\left[\mathbb{1}_{C_b<r} \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right] && (37)\\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}_{C_b<r}\left|\Gamma\right.\right] \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right] && (38)\\
&= \mathbb{E}\left[\mathbb{P}\left\{C_b < r\left|\Gamma\right.\right\} \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right] && (39)\\
&= \mathbb{E}\left[Q\left(\frac{\log_2(1+\Gamma) - r}{\sqrt{\mathcal{V}(\Gamma)/n_{\mathrm{d}}}}\right) \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right], && (40)
\end{aligned}$$

where (38) follows from the law of total expectation, and (40) follows from the definition of U_{FBL} as a Gaussian variable.

The definition of the blocklength-equivalent capacity C_{b} allows treating errors due to imperfect CSI and finite blocklength as outage events that depend on two random variables, Γ and U_{FBL}. However, the blocklength-equivalent outage probability cannot be obtained easily, as C_{b} is logarithmic in the SNR Γ but linear in U_{FBL}. We overcome this problem by using the first-order Taylor approximation of $\ln(x)$ around the point x_0, which has gradient $\frac{1}{x_0}$. Due to the concavity of the $\ln$-function, this linear approximation is larger than the function itself:

$$\ln(x_0) + \frac{1}{x_0}a \geq \ln\left(x_0 + a\right)\,. \tag{41}$$

Due to $\ln(x)$ being continuous and monotonically increasing, this means that for some $\delta \geq 0$ and $b = -a\log_2(e)/x_0$:

$$\log_2(x_0) - b = \log_2\left(x_0 - x_0\frac{b}{\log_2(e)} + \delta\right)\,. \tag{42}$$

We apply this result to C_{b} given in (36) around $x_0 = 1 + \Gamma$:

$$C_{\mathrm{b}} = \log_2\left(1 + \Gamma - \sigma_{\mathrm{FBL}}(\Gamma)\cdot U_{\mathrm{FBL}} + U_{\delta}\right) \tag{43}$$

with

$$\sigma_{\mathrm{FBL}}(\Gamma) \triangleq \frac{1+\Gamma}{\log_2(e)}\sqrt{\frac{\mathcal{V}(\Gamma)}{n_{\mathrm{d}}}}\,, \tag{44}$$

and some random $U_{\delta} \geq 0$. Now, recall from (32) that the estimation error $\tilde{\Gamma}$ in the SNR Γ can be approximated as a Gaussian error: $\tilde{\Gamma} = \tilde{\Gamma}_{\mathrm{G}} + \tilde{\Gamma}_{\delta}$ with $\tilde{\Gamma}_{\delta} \geq 0$. Thus, (43) becomes

$$\begin{aligned}
C_{\mathrm{b}} &= \log_2\left(1 + \hat{\gamma} + \tilde{\Gamma} - \sigma_{\mathrm{FBL}}(\hat{\gamma} + \tilde{\Gamma})U_{\mathrm{FBL}} + U_{\delta}\right) && (45)\\
&\geq \log_2\left(1 + \hat{\gamma} + \tilde{\Gamma}_{\mathrm{G}} - \sigma_{\mathrm{FBL}}(\hat{\gamma} + \tilde{\Gamma})U_{\mathrm{FBL}}\right)\,. && (46)
\end{aligned}$$

We recall that $\varepsilon = \mathbb{P}\left\{C_b < r \,\middle|\, \hat{\Gamma} = \hat{\gamma}\right\}$. Thus, the error probability ε can be bounded as

$$\varepsilon \leq \mathbb{P}\left\{\log_2\left(1 + \hat{\gamma} + \tilde{\Gamma}_{\mathrm{G}} - \sigma_{\mathrm{FBL}}(\hat{\gamma} + \tilde{\Gamma})U_{\mathrm{FBL}}\right) < r\right\}\,. \tag{47}$$

Conjecture 1. *Inequality (47) holds when $\sigma_{\mathrm{FBL}}(\hat{\gamma} + \tilde{\Gamma})$ is replaced by $\sigma_{\mathrm{FBL}}(\hat{\gamma})$, i.e.*

$$\varepsilon \leq \mathbb{P}\left\{\log_2\left(1 + \hat{\gamma} + \tilde{\Gamma}_{\mathrm{G}} - \sigma_{\mathrm{FBL}}(\hat{\gamma})U_{\mathrm{FBL}}\right) < r\right\} \tag{48}$$

as long as the rate r is below the estimated channel capacity $\log_2(1+\hat{\gamma})$.

We have not been able to disprove the conjecture numerically, and in the following we provide intuitive reasoning why it should always hold.

Rationale. When the estimated SNR $\hat{\gamma}$ is larger than the actual SNR, then the variance is replaced by a larger term, i.e., the finite blocklength effects are overestimated. Using channel codes with short blocklength decreases the reliability of the transmission, thus, overestimating the finite blocklength effects leads in general to an overestimation of the error probability. On the other hand, when the estimated SNR $\hat{\gamma}$ is smaller than the actual SNR, then the channel is already better than predicted, and there is a high margin between the actual capacity and the rate. Errors in this regime are very rare and do not significantly contribute to the overall error probability. □

Assuming Conjecture 1 holds, we can derive the following lemma, which provides a fairly tight upper bound on the error probability ε. Even if Conjecture 1 is not considered, the following expression still provides an approximation for ε.

Lemma 2. *When the estimated SNR $\hat{\gamma}$ and the average SNR $\bar{\gamma}$ are known, then the error probability ε for a code with rate $r < \log_2(1+\hat{\gamma})$ and blocklength n_d is bounded as*

$$\varepsilon \leq Q\left(\frac{\hat{\gamma} - (2^r - 1)}{\sigma_{\mathrm{IC,F}}(\hat{\gamma})}\right) \triangleq \varepsilon' \,, \tag{49}$$

with

$$\sigma^2_{\mathrm{IC,F}}(\hat{\gamma}) = \sigma^2_{\mathrm{ICSI}} + \sigma^2_{\mathrm{FBL}}(\hat{\gamma}) = \frac{2\bar{\gamma}\hat{\gamma}}{1+\bar{\gamma}n_\mathrm{t}} + \frac{(1+\hat{\gamma})^2\mathcal{V}(\hat{\gamma})}{\log_2^2(e)n_\mathrm{d}} \,. \tag{50}$$

Proof. The variables $\tilde{\Gamma}_\mathrm{G} \sim \mathcal{N}(0, \sigma^2_{\mathrm{ICSI}})$ and $U_\mathrm{FBL} \sim \mathcal{N}(0,1)$ are independent. Thus, the difference $\tilde{\Gamma}_\mathrm{G} - \sigma_\mathrm{FBL}(\hat{\gamma})U_\mathrm{FBL}$ can be described by $U_\mathrm{IC,F} \sim \mathcal{N}(0, \sigma^2_{\mathrm{IC,F}}(\hat{\gamma}))$, where $\sigma^2_{\mathrm{IC,F}}(\hat{\gamma})$ is the sum of the two variances. Then, starting from Conjecture 1 and (48), we obtain:

$$\begin{aligned}\varepsilon &\leq \mathbb{P}\left\{\hat{\gamma} + \tilde{\Gamma}_\mathrm{G} - \sigma_\mathrm{FBL}(\hat{\gamma})U_\mathrm{FBL} < 2^r - 1\right\} \\ &= \mathbb{P}\left\{\hat{\gamma} + U_\mathrm{IC,F} < 2^r - 1\right\} \\ &= \mathbb{P}\left\{-\frac{U_\mathrm{IC,F}}{\sigma_\mathrm{IC,F}(\hat{\gamma})} > \frac{\hat{\gamma} - (2^r-1)}{\sigma_\mathrm{IC,F}(\hat{\gamma})}\right\} \,.\end{aligned}$$

□

In the context of ultra-reliable low latency systems with rate adaptation, a transmitter may want to choose the data rate r such that the error probability ε does not exceed a given target error probability. Contrary to the expression (7) for the error probability ε, which cannot be inverted to obtain the rate r given ε, the bound ε' in (49) can be inverted for the rate:

Corollary 1. *Given a channel estimate $\hat{\gamma}$, training sequence length n_t, blocklength n_d, and a target error probability ε' with $Q(\hat{\gamma}/\sigma_\mathrm{IC,F}(\hat{\gamma})) < \varepsilon' < 1/2$, the actual error probability ε is less than or equal ε' if the transmitter chooses the rate*

$$r_\mathrm{IC,F}(\hat{\gamma}, \varepsilon') = \log_2\left(1 + \hat{\gamma} - \sigma_\mathrm{IC,F}(\hat{\gamma})Q^{-1}(\varepsilon')\right) \,. \tag{51}$$

Proof. The proof follows by solving (49) for r, with $\varepsilon' > Q(\hat{\gamma}/\sigma_\mathrm{IC,F}(\hat{\gamma}))$ ensuring that $r > 0$. □

D. Optimal Rate Adaptation

As shown in Sec. III-B, the optimal rate adaptation function Φ^*_θ for a specific parameter $\theta > 0$ minimizes (25), i.e., the Mellin transform $\mathcal{M}_\mathcal{S}(1-\theta)$ of the SNR-domain service process $\mathcal{S}$. The rate adaptation problem can be solved when the error probability ε is replaced by its closed-form approximation ε' from Lemma 2. It is however not yet clear whether optimizing the rate adaptation strategy requires an exhaustive search over all possible rates r. As we will show now, such an exhaustive search is not necessary, as the rate adaptation problem is convex. For the proof, we replace the direct mapping $\Phi : \hat{\gamma} \to r$ with an indirect mapping where we first determine $\varphi : \hat{\gamma} \to \varepsilon'$ and then use Corollary 1 to obtain the rates $r = r_\mathrm{IC,F}(\hat{\gamma}, \varepsilon')$ with $\varepsilon' = \varphi(\hat{\gamma})$, yielding:

$$\mathcal{M}_\mathcal{S}(1-\theta) \leq \int_0^\infty \underbrace{\left((1-\varphi(\hat{\gamma}))e^{n_\mathrm{d}\cdot r\cdot(-\theta)} + \varphi(\hat{\gamma})\right)}_{\tilde{g}(\hat{\gamma},\varphi(\hat{\gamma}))} f_{\hat{\Gamma}}(\hat{\gamma})d\hat{\gamma} \,, \tag{52}$$

where $r = r_\mathrm{IC,F}(\hat{\gamma}, \varphi(\hat{\gamma}))$. The inequality is due to $\varepsilon \leq \varepsilon'$ as established by Lemma 2. For any $\theta > 0$, the rate adaptation function $\tilde{\Phi}^*_\theta$ based on the approximation from Lemma 2 is then given by $\tilde{\Phi}^*_\theta(\hat{\gamma}) = r_\mathrm{IC,F}(\hat{\gamma}, \varphi^*_\theta(\hat{\gamma}))$, where φ^*_θ minimizes the right-hand side of (52). The function φ^*_θ can be found by minimizing the term $\tilde{g}(\hat{\gamma}, \varepsilon')$ over ε' individually for each discretized value $\hat{\gamma}$. Building on results from [26], we find:

Lemma 3. *$\tilde{g}(\hat{\gamma}, \varepsilon')$ is convex in ε' for $Q(\hat{\gamma}/\sigma_\mathrm{IC,F}(\hat{\gamma})) < \varepsilon' < 1/2$ and $\theta > 0$.*

Proof. See Appendix B. □

The convexity property ensures that the optimal $\varepsilon' = \varphi^*_\theta(\hat{\gamma})$ for each $\hat{\gamma}$ is unique and can be found efficiently. Thus, the optimized rate adaptation function $\tilde{\Phi}^*$ can be found efficiently.

E. Further Uses of the Approximation

Lemma 2 can also be used for optimizing different performance metrics of the queuing system.

1) Maximization of Expected Goodput: The expected goodput $\overline{\eta}$ describes the expected number of bits per slot that can be successfully sent to the receiver in case there is always a sufficient backlog of data in the transmit buffer:

$$\overline{\eta} \triangleq \mathbb{E}[S] = \mathbb{E}\left[n_\mathrm{d}\Phi(\hat{\Gamma})\cdot(1-\mathcal{E})\right] \tag{53}$$

$$\leq \int_0^\infty \underbrace{n_\mathrm{d}\Phi(\hat{\gamma})\cdot\left(1 - Q\left(\frac{\hat{\gamma} - \left(2^{\Phi(\hat{\gamma})} - 1\right)}{\sigma_\mathrm{IC,F}(\hat{\gamma})}\right)\right)}_{\triangleq\, h(\hat{\gamma},\Phi(\hat{\gamma}))} f_{\hat{\Gamma}}(\hat{\gamma})d\hat{\gamma} \,, \tag{54}$$

where we applied the inequality (49) from Lemma 2. We can efficiently determine a rate adaptation function Φ that maximizes the expected goodput $\overline{\eta}$ because $h(\hat{\gamma}, r)$ defined in (54) is concave in r:

Lemma 4. *$h_1(r) \triangleq h(\hat{\gamma}, r)$ is concave in r for each $\hat{\gamma}$ and $0 < r < \log_2(1+\hat{\gamma})$.*

Proof. All terms in the second derivative of $h_1(r)$ are strictly negative in the given range. The restriction $r < \log_2(1+\hat{\gamma})$ ensures that the error probability is below $1/2$. □

2) Effective Capacity: Above, we relied on stochastic network calculus to obtain an upper bound on the delay violation probability $p_\mathrm{v}(w)$, which is valid even at small values of the target delay w. At fairly long delays, i.e., for the tail of the delay distribution, one can also use the effective capacity framework to approximate $p_\mathrm{v}(w)$. The effective capacity for the cumulative service process $\mathbf{S}(\tau, t)$ is defined as [22]

$$c_\mathrm{E}(\theta) = -\lim_{t\to\infty}\frac{1}{\theta t}\log\left(\mathbb{E}\left[e^{-\theta\mathbf{S}(0,t)}\right]\right) \,. \tag{55}$$

As we assume the service increments S to be independent, the effective capacity simplifies to

$$c_\mathrm{E}(\theta) = -\frac{1}{\theta}\log\left(\mathbb{E}\left[e^{-\theta S}\right]\right) \,. \tag{56}$$

We note that $\mathbb{E}\left[e^{-\theta S}\right] = \mathbb{E}\left[\mathcal{S}^{-\theta}\right] = \mathcal{M}_\mathcal{S}(1-\theta)$. As $\log(x)$ is monotonically increasing, the rate adaptation function Φ that minimizes $\mathcal{M}_\mathcal{S}(1-\theta)$, which can be found as shown in Sec. III-D, also maximizes the effective capacity $c_\mathrm{E}(\theta)$.

IV. Numerical Evaluation

In this section we evaluate various aspects of our derived mathematical models. Note that we do not consider scenarios where the deadline w is extremely short, e.g., only one or two time slots. For such deadlines, the system would need to achieve extremely low transmission error probabilities on the physical layer, e.g., $\varepsilon < 10^{-8}$, in order to achieve ultra high reliability of $p_{\mathrm{v}}(w) < 10^{-8}$ on the application layer. The accuracy of the system model for such small block error probabilities ε cannot be verified by the currently known information-theoretic bounds. Therefore, our numerical analysis considers only systems where the deadline is slightly longer, e.g., $w = 5$ time slots, and ultra high reliability of $p_{\mathrm{v}}(w) < 10^{-8}$ on the application layer can be achieved even when ε is higher, e.g., above 10^{-3}.

In Sec. IV-A, we address the accuracy of the error probability approximation from Lemma 2, especially with respect to its use in rate adaptation. In Sec. IV-B, we investigate the optimal length of the training sequence, as well as the impact of feedback quantization. In Sec. IV-C, we compare rate adaptation to fixed rate schemes. Finally, in Sec. IV-D, we separately investigate the performance impact of imperfect CSI and finite blocklength effects.

A. Accuracy of the Approximation

In Fig. 3, we compare the error probability ε in (7) with its approximation ε' from Lemma 2. The length of the training sequence is $n_{\mathrm{t}} \in \{10, 25\}$, the average SNR $\bar{\gamma}$ is 15 dB, and the length of the data transmission phase is fixed at $n_{\mathrm{d}} = 200$. In this example, we want to investigate the difference between ε and ε' at various rates. Therefore, we use simple, non-optimized rate adaptation functions given as $\Phi(\hat{\gamma}) = \kappa \cdot \hat{c}$, where $\hat{c} = \log_2(1 + \hat{\gamma})$ is the estimated capacity and $\kappa \in \{0.75, 0.9, 0.95\}$ is chosen arbitrarily. First of all, we confirm in all cases that ε' is indeed an upper bound on ε, as expected from Lemma 2. Second, even though we observe that this bound is not always tight, especially when $\varepsilon' < 10^{-2}$, it can be seen that the upper bound ε' predicts quite well how much the error probability ε changes when the rate is slightly reduced or when the number of training symbols n_{t} changes. Thus, Fig. 3 already provides strong indication that the approximation ε' is accurate enough to optimize the rate adaptation function.

As shown in Sec. III-B, the optimal rate adaptation function minimizes the value of $\mathcal{M}_{\mathcal{S}}(1-\theta)$. We compare in Fig. 4a the rate adaptation function $\tilde{\Phi}^*_\theta$ that was optimized based on the analytical approximation ε' in Lemma 2 (red dashed curve) to the optimal function Φ^*_θ that solves the rate adaptation problem by numerically computing the error probability ε in (7) for many different values of the rate r (black solid curve). For the selected parameters[5], Fig. 4a shows that the difference in the selected rates between the two schemes is fairly small. More importantly, the value of $\mathcal{M}_{\mathcal{S}}(1-\theta)$ for $\theta = 0.01$ increases only slightly (by 1%) from 0.0291 to 0.0294. This underpins our approach to use the approximation ε' for solving the rate adaptation problem. In fact, we validated for a much larger parameter range that the rate adaptation based on our proposed approximation is sufficient. The largest differences were observed at low average SNR $\bar{\gamma}$ and small n_{t}, but even at $\bar{\gamma} = 5$ dB and $n_{\mathrm{t}} = 5$ (where the arrival rate was set to $\alpha = 15$ bits/slot so that delay requirements of $p_{\mathrm{v}}(5) < 10^{-8}$ could be met) the value of $\mathcal{M}_{\mathcal{S}}(1-\theta)$ at $\theta^* \approx 0.23$ increased only by 4% when using the approximation ε' in Lemma 2 instead of numerical computations of ε.

[5]For these parameters, $\alpha = 350$ bits/slot, and $w = 5$, the bound $\mathcal{K}\left(\theta, w, n_{\mathrm{t}}, \Phi^*_\theta\right)$ was minimal at $\theta^* \approx 0.010$.

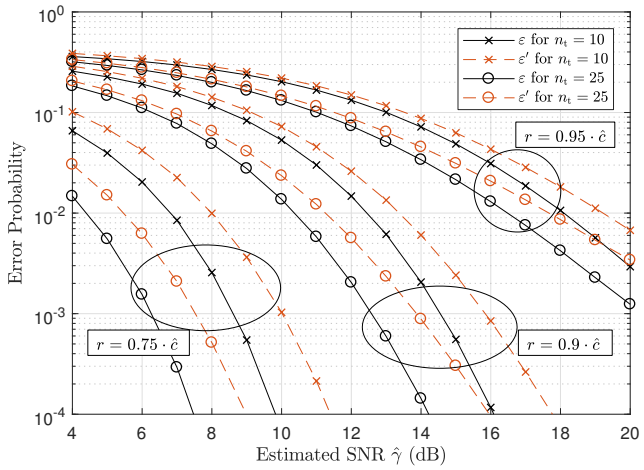
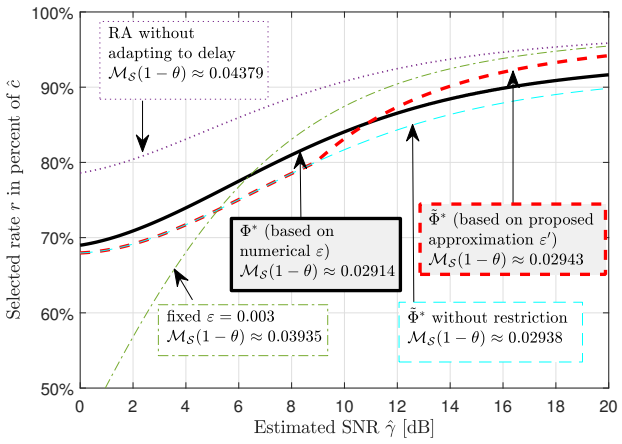


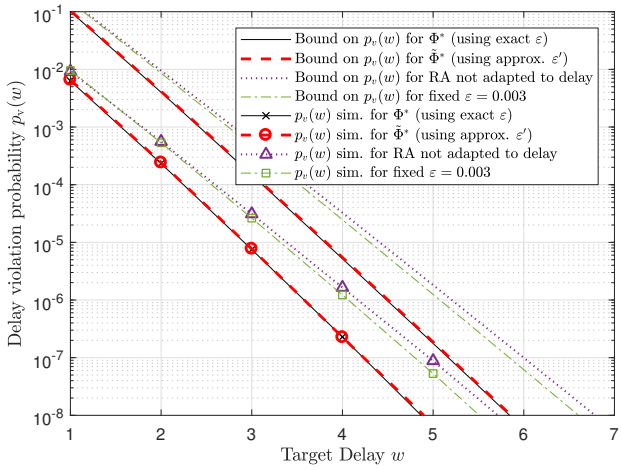
Fig. 3. Error probability ε and approximation/upper bound ε' from Lemma 2 vs. estimated SNR $\hat{\gamma}$, when the $\Phi(\hat{\gamma}) = \kappa \cdot \hat{c}$ with $\kappa \in \{0.75, 0.9, 0.95\}$. $n_{\mathrm{t}} \in \{10, 25\}$, $n_{\mathrm{d}} = 200$, $\bar{\gamma} = 15$ dB.

The system model, which is based on the normal approximation (7) for finite blocklength coding, may become inaccurate when the blocklength n_{d} is very small or when the error probability ε is extremely small. In general, the normal approximation is considered accurate when the rate is a large fraction of the channel capacity [3]. We have avoided very short blocklengths, and we observe in Fig. 4a that the selected rate always amounts to a large fraction of the estimated capacity, with $r \geq 0.7 \cdot \hat{c}$. Additionally, in order to avoid selecting parameters where ε becomes extremely small, we have restricted the search for the optimal ε' to $\varepsilon' \geq 10^{-3}$. Fig. 4a also shows the selected rates when ε' is not restricted (dashed blue curve). In this example, restricting $\tilde{\Phi}^*_\theta$ to $\varepsilon' \geq 10^{-3}$ is only relevant for $\hat{\gamma} > 9$ dB. The restriction has almost no impact on the value of $\mathcal{M}_{\mathcal{S}}(1-\theta)$, which depends mostly on the service at low values of $\hat{\gamma}$, where the system must accept error probabilities above 10^{-3} in order to avoid very low data rates. In conclusion, we find that the values of the blocklength, rate, and error probability are in a range where the normal approximation (7) is assumed to be accurate. In Appendix A, we also validate this assumption quantitatively through information-theoretic bounds.

Furthermore, Fig. 4a shows the results for two suboptimal rate adaptation schemes. The green dash-dotted curve shows the rate r that would be chosen when the transmitter always keeps the error probability at a fixed value $\varepsilon = 0.003$ for all values of $\hat{\gamma}$. The value of $\mathcal{M}_{\mathcal{S}}(1-\theta)$ increases to 0.0394 for this scheme (other fixed values of ε are even worse). The second suboptimal rate adaptation scheme (purple dotted curve) is the one that does not take the delay requirements into account, but optimizes Φ to achieve the maximum expected goodput $\overline{\eta}$



(a)



(b)

Fig. 4. (a) Choice of $r(\hat{\gamma})$ in percent of $\hat{c} = \log_2(1+\hat{\gamma})$ vs. estimated SNR $\hat{\gamma}$ for different rate adaptation schemes. $n_\mathrm{t} = 25$, $n_\mathrm{d} = 200$, $\bar{\gamma} = 15$ dB, $\theta = 0.01$. (b) Delay violation probability $p_\mathrm{v}(w)$ (obtained from simulations over 10^{11} time steps) and its respective upper bound (19) vs. target delay w. $n_\mathrm{t} = 25$, $n_\mathrm{d} = 200$, $\bar{\gamma} = 15$ dB, arrival rate $\alpha = 350$ bits/slot.

defined in (53). Such a delay-agnostic rate adaptation scheme favors high data rates over high reliability, which causes an increase in $\mathcal{M}_\mathcal{S}(1-\theta)$ to 0.0438. Due to the massive increases in $\mathcal{M}_\mathcal{S}$, we suspect that the delay performance will deteriorate with both suboptimal schemes.

This suspicion is confirmed by Fig. 4b. It shows the delay violation probability $p_\mathrm{v}(w)$, which can be obtained empirically by simulating the queueing system with random instances of the service and arrival process, and its analytical upper bound (19), versus the target delay w for those different rate adaptation schemes. We first note that while the upper bound (19) on $p_\mathrm{v}(w)$ is not tight[6], the upper bound is very useful, as it not only predicts the slope of $p_\mathrm{v}(w)$ correctly, but also predicts which parameters (here: which rate adaptation schemes) are optimal with respect to $p_\mathrm{v}(w)$. Thus, for the considered choices of rate adaptation functions Φ, the original optimization problem in (13) and its analytical approximation

[6]The difference is not due to inaccuracies in the system model or in the error probability approximation because the simulations use the same system model. Similar differences were also observed in other works on stochastic network calculus [24], [25], [27] and may be due to union bounds and moment bounds used to derive the analytical upper bound (19).

TABLE I
BUFFER VIOLATION PROBABILITY (FROM SIMULATIONS OVER $2 \cdot 10^{10}$ TIME SLOTS) FOR DIFFERENT BUFFER SIZES.

Buffer Size	$4 \cdot \alpha$	$5 \cdot \alpha$	$6 \cdot \alpha$	$7 \cdot \alpha$
Buffer Violation Prob.	$6 \cdot 10^{-6}$	$2 \cdot 10^{-7}$	$7 \cdot 10^{-9}$	0

(20) lead to the same results. We observe that the delay bounds for the rate adaptation Φ^* based on the exact error probability ε (solid black curve) and the rate adaptation $\tilde{\Phi}^*$ based on the approximation ε' (dashed red) are almost indistinguishable, which is in line with the nearly equal values of $\mathcal{M}_\mathcal{S}(1-\theta)$ we observed before. The difference between the two schemes in $p_\mathrm{v}(w)$ as obtained from simulations is also not noticeable. Contrary to that, when using the suboptimal schemes, which either use fixed $\varepsilon = 0.003$ or do not adapt the rate to the delay constraints, the delay violation probability $p_\mathrm{v}(w)$ at $w = 4$ degrades by nearly an order of magnitude, and this degradation is correctly predicted by the analytical bounds. Furthermore, instead of simply maximizing the expected goodput $\overline{\eta}$, which is the performance metric that was considered in some previous works on finite blocklength communications (e.g. [8], [9], [11]), the parameter optimization should take the delay requirements specifically into account. In conclusion, we find that optimizing the rate adaptation function leads to significant improvements in the reliability compared to the considered suboptimal approaches. The optimal rate adaptation function can be found efficiently by using the proposed approximation.

Finally, we investigate the impact of limited buffer sizes in a system with $n_\mathrm{t} = 25$, $n_\mathrm{d} = 200$, $\bar{\gamma} = 15$ dB, and arrival rate $\alpha = 350$ bits/slot. If a new packet arrives when the transmit buffer is already full, the new packet is dropped. The probability of such buffer violations is given in Table I for buffer sizes of 4α to 7α. The buffer violation probability is closely related to the delay violation probability $p_\mathrm{v}(w)$, which decays exponentially in w as observed in Fig. 4b. As a result, when the buffer is only 7 times the size of an incoming packet α (i.e., 2450 bit), we never observed a full buffer in the entire simulation of $2 \cdot 10^{10}$ time steps. Therefore, when considering systems that guarantee high reliability with respect to a short deadline w, we can ignore the impact of limited buffer sizes, i.e., we safely assume that the buffer size is infinite.

B. Optimal Training Length and Quantized Feedback

Fig. 5 shows the delay bound (19) for different values of the arrival rate α versus the training length n_t. We first consider unquantized feedback. For $\alpha = 250$ bits per slot, the smallest delay violation probability is obtained at $n_\mathrm{t}^* = 31$ training symbols, but the performance remains similar for n_t between 20 and 50. When the arrival rate is increased, fewer training symbols should be used; the delay violation probability for $\alpha = 600$ easily increases by an order of magnitude when too many training symbols are used instead of $n_\mathrm{t}^* = 17$.

Moreover, Fig. 5 shows the performance when the feedback is quantized. Here, we did not quantize the estimated SNR $\hat{\gamma}$ directly, but quantized the estimated capacity $\hat{c} = \log_2(1+\hat{\gamma})$

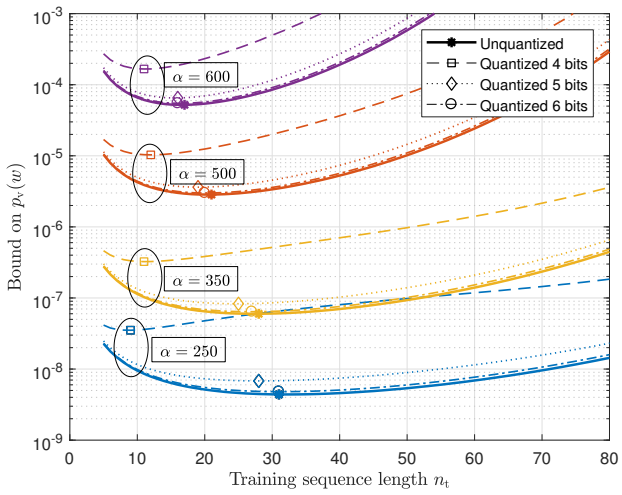


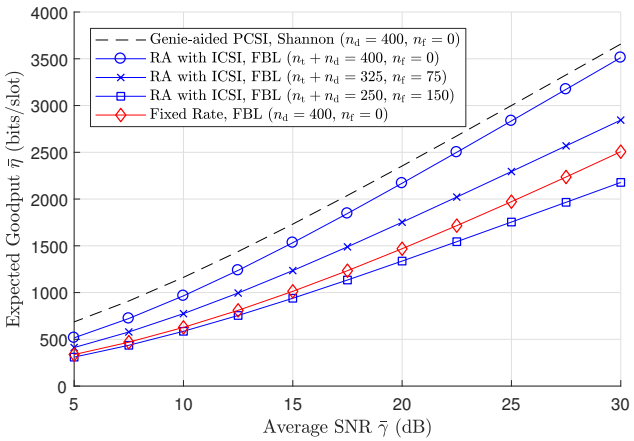
Fig. 5. Bounds on the delay violation probability vs. number of training symbols n_t for target delay $w = 5$, average SNR $\bar{\gamma} = 15$ dB, $n_{slot} = 400$, $n_f = 150$, different arrival rates α in bits/slot and different levels of feedback quantization. The markers show the minimum points (optimal n_t).

uniformly up to a chosen maximum value[7]. When only 4 bits are used, the performance easily degrades by an order of magnitude. In this case, the optimal value of n_t also decreases because it does not make sense to spend a long time on estimation and then coarsely quantize the estimate. However, we observe no significant performance differences when the feedback is quantized with at least 6 bits, even with this simple uniform quantizer, which means that rate adaptation with limited feedback is feasible. In case data is transmitted also in the opposite direction from the receiver to the transmitter, the 6 feedback bits can easily be appended to that data stream, and the cost of feedback is almost negligible.

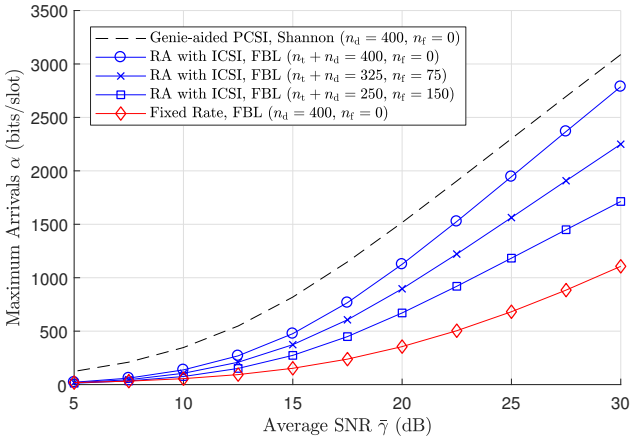
C. Rate Adaptation vs. Fixed Rate Transmission

We now investigate whether wireless systems for low latency communications should use transmissions at a fixed rate instead of rate adaptation, as the latter is more complex and requires spending a large fraction of resources for channel estimation and feedback. In Fig. 6a, we show the expected goodput $\overline{\eta}$ in (53), i.e. the performance when there are no delay constraints, versus the average SNR $\bar{\gamma}$. Fig. 6b shows the maximum supported arrival rate α such that the system still meets strict delay constraints. In both figures, we plot the performance of an unrealistic, genie-aided system which is assumed to know the channel perfectly and transmits without errors at a rate equal to the instantaneous channel capacity over the entire slot length of $n_{slot} = 400$ symbols. In addition, we show the performance of realistic systems with or without rate adaptation. For the rate adaptation systems (blue curves), we always show the results for the optimized rate adaptation scheme $\tilde{\Phi}^*$ and the optimal length n_t^* of the training sequence. We make three different assumptions for the length of the feedback phase: $n_f = 0$ symbols (piggybacking of feedback

[7]The uniform quantizer is suboptimal; we heuristically designed a non-uniform quantizer that takes the impact of quantization on $\mathcal{M}_S(\theta)$ into account, yielding much better results. Optimal design of feedback quantization is left for future work.



(a)



(b)

Fig. 6. Performance for $n_{slot} = 400$, $n_f \in \{0, 75, 150\}$; n_t and rates always chosen optimally. (a) Expected goodput $\overline{\eta}$ vs. average SNR $\bar{\gamma}$. (b) Maximum arrival rate α vs. average SNR such that for target delay $w = 5$ slots, $p_v(w) < 10^{-8}$.

– no additional cost), $n_f = 75$ symbols (medium cost), and $n_f = 150$ symbols (high cost). For comparison, we also show the performance of a system that uses the entire time slot of $n_{slot} = 400$ symbols to transmit codewords at fixed rate, with rates also optimized for maximum $\overline{\eta}$ or α. For the fixed rate transmissions, we still consider finite blocklength effects. When comparing only the expected goodput $\overline{\eta}$ in Fig. 6a, the system with fixed rate transmissions is superior to the system with rate adaptation and $n_f = 150$ symbols. However, we see drastically different results when taking the delay requirements into account. Fig. 6b shows the maximum arrivals α such that the system violates a deadline of $w = 5$ slots only with $p_v(w) < 10^{-8}$: the system with fixed rate, which uses the entire time slot for data transmissions, now performs far worse than the systems with rate adaptation, even though the rate adaptation systems use up to $n_f = 150$ symbols (more than a third of the time slot) for feedback. In fact, at $\bar{\gamma} = 20$ dB, the fixed rate system supports only half the arrival rate α compared to the rate adaptation system with $n_f = 150$. Thus, for low latency communications, systems with rate adaptation heavily outperform fixed rate systems, despite the cost of channel estimation and the large overhead from feedback.

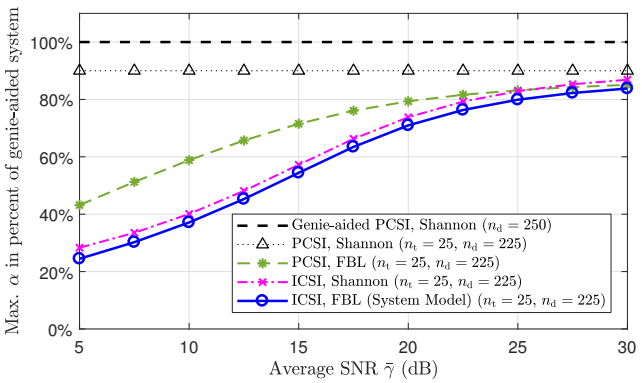


Fig. 7. Maximum arrival rate α, in percent of the max. α for the genie-aided system, vs. average SNR for $n_{\mathrm{slot}} = 250$, $n_{\mathrm{f}} = 0$ such that $p_{\mathrm{v}}(w) < 10^{-8}$ for $w = 5$. Different models assuming perfect/imperfect CSI and finite/infinite blocklength.

D. Quantitative Performance Loss due to Finite Blocklength and Imperfect CSI

The performance loss of the system with rate adaptation is due to the time required for channel estimation and feedback, but also due to imperfect CSIT and finite blocklength effects. In Fig. 7, we show how much these effects contribute to the performance loss when $n_{\mathrm{slot}} = 250$, $n_{\mathrm{f}} = 0$, and $n_{\mathrm{t}} = 25$. In this figure, we either assume CSI to be perfect (PCSI) or imperfect (ICSI). Furthermore, we either assume that the error probability depends on finite blocklength (FBL) effects and is given by (7), or we assume that the error probability is equal to the conditional outage probability, implicitly assuming an infinite blocklength (Shannon) model. For better visualization at low average SNR, we now plot the maximum arrival rate α of each of these systems in percent the maximum arrival rate α that can be achieved by a genie-aided system with $n_{\mathrm{d}} = 250$, with perfect CSI and assuming infinite blocklength. Taking only the overhead due to the $n_{\mathrm{t}} = 25$ training symbols into account, but still assuming infinite blocklength and perfect CSI, one gets 90% of the performance of the genie-aided system. In addition to the overhead, both imperfect CSI and finite blocklength effects have a substantial influence on the performance. However, considering just the finite blocklength effects (green dashed curve) only leads to an accurate estimate of the system performance when the average SNR is very high. In most cases, a better approximation of the performance is found by considering only the effects of imperfect CSI but assuming a Shannon model for channel codes (magenta dash-dotted curve). The graph shows only a small difference to the system model (blue curve). However, even though the two curves with imperfect CSI are close and show the same trend, the relative difference between them is significant at low SNR. When longer training sequences are used, this difference increases further. Thus, when rate adaptation is used, finite blocklength effects have significant influence on the performance. This stands in contrast to the results in [5], [6] for fixed rate transmissions, where it was shown that finite blocklength effects have very little impact. Nevertheless, while finite blocklength effects must not be ignored, it is in many cases more important to take the performance loss due to imperfect CSI into account.

V. Conclusions and Future Work

In this work, we studied the delay performance of rate adaptation systems with imperfect CSI and finite blocklength channel coding. We developed a closed-form approximation for the error probability due to imperfect CSI and finite blocklength, which can be used to efficiently optimize the rate adaptation function. Despite the large overhead for channel estimation and feedback that is necessary for rate adaptation, our numerical results show that rate adaptation is superior to fixed rate transmissions when low latency is required. A possible extension of this work relates to multi-antenna systems, which can offer higher reliability at the cost of even more channel estimation, while imperfect CSI at the transmitter would also affect beamforming.

Appendix A
Validation of the System Model

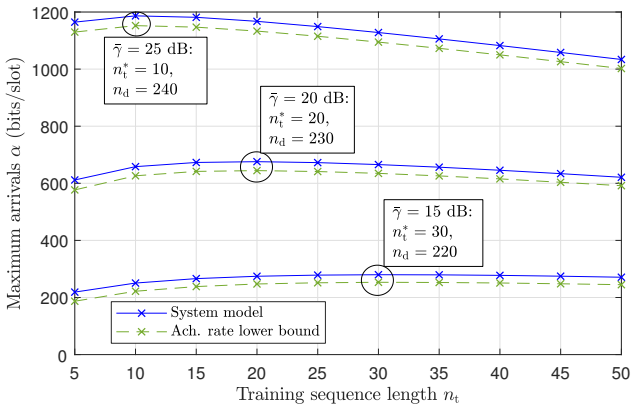
Please note that in order to understand this section, the reader is encouraged to first read the numerical evaluation in Sec. IV-A and Sec. IV-C.

We assumed throughout the paper that the decoding error probability is exactly equal to ε in (7), which is based on the assumption of perfect CSIR [6]. Using [5, Cor. 3] (implementation available in [35]), we obtain a strict lower bound on the achievable rate under imperfect CSIR. For the bound, one assumes only knowledge of the fading statistics of the corresponding Rician fading channel, i.e., of $\hat{h}$ and σ_Z^2, but no prior knowledge of the fading coefficient H. In Fig. 8, we compare the delay performance of the system model to the performance of a system where the achievable rate[8] is given by [5, Cor. 3]. In particular, we compare in Fig. 8a the maximum arrival rate α per time slot such that the system can still guarantee the quality-of-service constraints, versus the training sequence length n_{t}, while keeping $n_{\mathrm{slot}} = 400$ and $n_{\mathrm{f}} = 150$ fixed. We observe in all cases only small differences between the system model and the model based on the lower bound [5, Cor. 3]. Furthermore, both curves show the same trends, and the optimal training sequence length $n_{\mathrm{t}}{}^*$ (marked by circles) is the same. Therefore, the system model based on (7) is accurate enough to support all our findings.

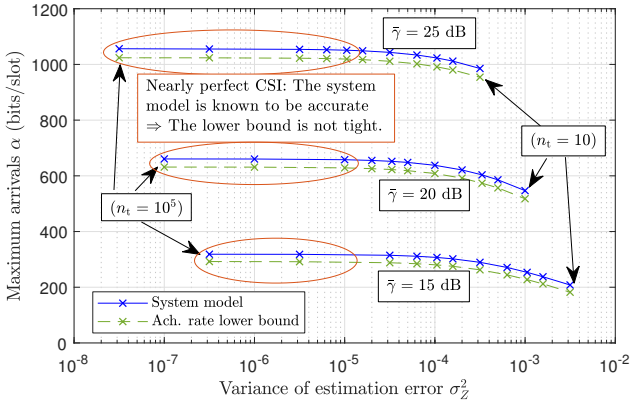
The above validation assumed that the receiver only knows the imperfect channel estimate $\hat{h}$. However, in many practical systems, the transmitter will send additional synchronization and pilot symbols during the data transmission phase, which will make the CSIR better than the CSIT. In addition, the receiver can improve the CSI through combined estimation and decoding [36]. These arguments provide additional support for the accuracy of (7) in practical systems.

In Fig. 8b, we further investigate the difference between the system model and the lower bound on the achievable rate. Specifically, we investigate the performance at nearly

[8]First, we always determine the optimal values ε' as described in Sec. III-D. For the system model, the rates are given by $r_{\mathrm{IC,F}}(\hat{\gamma}, \varepsilon')$ and the error probabilities are given by (7). For the lower bound, we set the error probability to ε' and compute the corresponding achievable rates from [5, Cor. 3]. We use the restriction $\varepsilon' > 10^{-3}$ here, because [5, Cor. 3] can only be evaluated by Monte Carlo methods. This becomes computationally prohibitive for extremely small ε'.



(a)



(b)

Fig. 8. Maximum arrivals α so that for $w = 5$, $p_{\mathrm{v}}(w) < 10^{-8}$ is satisfied. $\bar{\gamma} \in \{15, 20, 25\}$ dB. (a) Max. α vs. n_{t}, while $n_{\mathrm{slot}} = 400$ and $n_{\mathrm{f}} = 150$ remain fixed. (b) Max. α vs. σ_Z^2 (variance of the channel estimation error), with fixed $n_{\mathrm{d}} = 200$.

perfect CSI, where we keep $n_{\mathrm{d}} = 200$ fixed and decrease the variance $\sigma_Z^2 = 1/(1+\bar{\gamma}n_{\mathrm{t}})$ of the channel estimation error to extremely small values below 10^{-5} (this would correspond to an unrealistic number of e.g. $n_{\mathrm{t}} = 10^4$ training symbols). At perfect CSI, the system model is accurate because (7) converges to (6) and it was shown previously that (6) is very accurate [3]. Thus, the lower bound [5, Cor. 3] is not tight at perfect CSI; it slightly underestimates the achievable performance. We now observe that the gap between each pair of curves remains constant over the whole range of σ_Z^2. Therefore, we conclude that even at realistic values of σ_Z^2 and $n_{\mathrm{t}} \in \{10, \ldots, 50\}$, the lower bound [5, Cor. 3] underestimates the performance, while the system model based on (7) provides a more accurate estimate of the performance.

Appendix B
Proof of Lemma 3

To show convexity, we show that for fixed $\hat{\gamma}$, the second derivative of $g_1(\varepsilon') = \tilde{g}(\hat{\gamma}, \varepsilon')$ is strictly positive for $\theta > 0$:

$$g_1(\varepsilon') = (1-\varepsilon')e^{n_{\mathrm{d}} \cdot r_{\mathrm{IC,F}}(\hat{\gamma}, n_{\mathrm{d}}, \varepsilon')(-\theta)} + \varepsilon' \tag{57}$$

$$= (1-\varepsilon')(1+\hat{\gamma} - \sigma_{\mathrm{IC,F}}(\hat{\gamma})Q^{-1}(\varepsilon'))^{-\frac{n_{\mathrm{d}}}{\ln 2}\theta} + \varepsilon' \tag{58}$$

$$= (1-\varepsilon')(a - bQ^{-1}(\varepsilon'))^c + \varepsilon' \tag{59}$$

with $a, b > 0$ and $c < 0$. Due to $\varepsilon' > Q(\hat{\gamma}/\sigma_{\mathrm{IC,F}}(\hat{\gamma}))$, we have $r_{\mathrm{IC,F}}(\hat{\gamma}, n_{\mathrm{d}}, \varepsilon') > 0$ and $(a - bQ^{-1}(\varepsilon')) > 1$. The first derivative is given by

$$\dot{g}_1(\varepsilon') = (1-\varepsilon')c(a - bQ^{-1}(\varepsilon'))^{c-1}(-b\dot{Q}^{-1}(\varepsilon')) - (a - bQ^{-1}(\varepsilon'))^c + 1. \tag{60}$$

The second derivative is given by

$$\ddot{g}_1(\varepsilon') = (1-\varepsilon')c(a - bQ^{-1}(\varepsilon'))^{c-1}(-b\ddot{Q}^{-1}(\varepsilon')) + (1-\varepsilon')c(c-1)(a - bQ^{-1}(\varepsilon'))^{c-2}(-b\dot{Q}^{-1}(\varepsilon'))^2 - 2c(a - bQ^{-1}(\varepsilon'))^{c-1}(-b\dot{Q}^{-1}(\varepsilon')). \tag{61}$$

From [26], the derivatives of the inverse Q-function are:

$$\dot{Q}^{-1}(\varepsilon') = -\sqrt{2\pi}e^{\frac{Q^{-1}(\varepsilon')^2}{2}} \tag{62}$$

$$\ddot{Q}^{-1}(\varepsilon') = 2\pi Q^{-1}(\varepsilon')e^{Q^{-1}(\varepsilon')^2} \tag{63}$$

Thus, for $\varepsilon' < 1/2$, $\dot{Q}^{-1}(\varepsilon') < 0$ and $\ddot{Q}^{-1}(\varepsilon') > 0$, and therefore $\ddot{g}_1(\varepsilon') > 0$.

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] J. Åkerberg, F. Reichenbach, and M. Björkman, "Enabling safety-critical wireless communication using WirelessHART and PROFIsafe," in *IEEE Conf. Emerging Technologies and Factory Automation (ETFA)*. IEEE, 2010, pp. 1–8.

[3] 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.

[4] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Diversity versus channel knowledge at finite block-length," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Sep. 2012, pp. 572–576.

[5] ——, "Quasi-static SIMO fading channels at finite blocklength," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2013, pp. 1531–1535.

[6] ——, "Quasi-static multiple-antenna fading channels at finite block-length," *IEEE Trans. Inf. Theory*, vol. 60, no. 7, pp. 4232–4243, Jul. 2014.

[7] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, 2016.

[8] P. Wu and N. Jindal, "Coding versus ARQ in fading channels: how reliable should the PHY be?" *IEEE Trans. Commun.*, vol. 59, no. 12, pp. 3363–3374, Dec. 2011.

[9] B. Makki, T. Svensson, and M. Zorzi, "Finite block-length analysis of the incremental redundancy HARQ," *IEEE Wireless Commun. Letters*, vol. 3, no. 5, pp. 529–532, 2014.

[10] ——, "Green communication via type-I ARQ: Finite block-length analysis," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*. IEEE, 2014, pp. 2673–2677.

[11] ——, "Finite block-length analysis of spectrum sharing networks using rate adaptation," *IEEE Trans. Commun.*, vol. 63, no. 8, pp. 2823–2835, 2015.

[12] ——, "Finite block-length analysis of spectrum sharing networks: Interference-constrained scenario," *IEEE Wireless Commun. Letters*, vol. 4, no. 4, pp. 433–436, 2015.

[13] O. L. A. López, E. M. G. Fernández, R. D. Souza, and H. Alves, "Wireless powered communications with finite battery and finite blocklength," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1803–1816, April 2018.

[14] H. Shariatmadari, R. Duan, S. Iraji, Z. Li, M. A. Uusitalo, and R. Jäntti, "Resource allocations for ultra-reliable low-latency communications," *Int. J. Wireless Inf. Networks*, pp. 1–11, 2017.

[15] J. Meng and E.-H. Yang, "Constellation and rate selection in adaptive modulation and coding based on finite blocklength analysis and its application to lte," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5496–5508, 2014.

[16] E.-H. Yang and J. Meng, "New nonasymptotic channel coding theorems for structured codes," *IEEE Trans. Inf. Theory*, vol. 61, no. 9, pp. 4534–4553, 2015.

[17] A. Lim and V. K. N. Lau, "On the fundamental tradeoff of spatial diversity and spatial multiplexing of MISO/SIMO links with imperfect CSIT," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 110–117, 2008.

[18] V. K. N. Lau, M. Jiang, and Y. Liu, "Cross layer design of uplink multi-antenna wireless systems with outdated CSI," *IEEE Trans. Wireless Commun.*, vol. 5, no. 6, pp. 1250–1253, 2006.

[19] M. Medard, "The effect upon channel capacity in wireless communications of perfect and imperfect knowledge of the channel," *IEEE Trans. Inf. Theory*, vol. 46, no. 3, pp. 933–946, 2000.

[20] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inf. Theory*, vol. 49, no. 4, pp. 951–963, 2003.

[21] C. Potter, K. Kosbar, and A. Panagos, "On achievable rates for MIMO systems with imperfect channel state information in the finite length regime," *IEEE Trans. Commun.*, vol. 61, no. 7, pp. 2772–2781, 2013.

[22] 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.

[23] 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.

[24] 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.

[25] N. Petreska, H. Al-Zubaidy, R. Knorr, and J. Gross, "On the recursive nature of end-to-end delay bound for heterogenous networks," in *IEEE Int. Conf. Commun. (ICC)*, Jun. 2015.

[26] M. C. Gursoy, "Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime," *EURASIP J. Wireless Commun. and Networking*, Dec. 2013.

[27] 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.

[28] G. Ozcan and M. C. Gursoy, "Throughput of cognitive radio systems with finite blocklength codes," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2541–2554, 2013.

[29] J. Gross, "Scheduling with outdated CSI: Effective service capacities of optimistic vs. pessimistic policies," in *Proc. 20th IEEE Int. Workshop Quality of Service (IWQoS)*. IEEE, 2012, pp. 1–9.

[30] 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.

[31] M. Ozmen and M. C. Gursoy, "Throughput regions of multiple-access fading channels with Markov arrivals and QoS constraints." *IEEE Wireless Commun. Letters*, vol. 2, no. 5, pp. 499–502, 2013.

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

[33] H. Al-Zubaidy, G. Dán, and V. Fodor, "Performance of in-network processing for visual analysis in wireless sensor networks," in *IFIP Networking Conf.* IEEE, 2015, pp. 1–9.

[34] A. H. Nuttall, "Some integrals involving the Q-M function," *IEEE Trans. Inf. Theory*, vol. 21, no. 1, pp. 95–96, 1975.

[35] Y. Polyanskiy, "SPECTRE: short-packet communication toolbox," 2016. [Online]. Available: https://github.com/yp-mit/spectre

[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.

Sebastian Schiessl received his Dipl.-Ing degree from Technical University of Munich, Germany in 2012. 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 wireless communications. In addition, he has also done some experimental implementations on the Wireless Open Access Research Platform (WARP).

Hussein Al-Zubaidy (S07M'11SM'16) received the Ph.D. degree in electrical and computer engineering from Carleton University, Ottawa, ON, Canada, in 2010. He was a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, from 2011 to 2013. In the Fall of 2013, he joined the School of Electrical Engineering (EES) at the Royal Institute of Technology (KTH), Stockholm, Sweden, as a Post-Doctoral Fellow. Since Fall 2015, he has been a Senior Researcher with EES at the Royal Institute of Technology (KTH), Stockholm, Sweden. Dr. Al-Zubaidy is the recipient of many honors and awards, including the Ontario Graduate Scholarship (OGS), NSERC Visiting Fellowship, NSERC Summer Program in Taiwan, OGSST, and NSERC Post-Doctoral Fellowship.

Mikael Skoglund (S93-M97-SM04) 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 Department 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 more than 140 journal and some 330 conference papers. Dr. Skoglund has served on numerous technical program committees for IEEE sponsored conferences. During 200308 he was an associate editor with the IEEE Transactions on Communications and during 200812 he was on the editorial board for the IEEE Transactions on Information Theory.

James Gross 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 Linneaus 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 100 (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.