Delay Analysis for Wireless Fading Channels with Finite Blocklength Channel Coding

Sebastian Schiessl
schiessl@kth.se

James Gross
james.gross@ee.kth.se

Hussein Al-Zubaidy
hzubaidy@kth.se

School of Electrical Engineering
KTH Royal Institute of Technology
100 44 Stockholm, Sweden

ABSTRACT

Upcoming low-latency machine-to-machine (M2M) applications are currently attracting a significant amount of interest from the wireless networking research community. The design challenge with respect to such future applications is to allow wireless networks to operate extremely reliably at very short deadlines for rather small packets. To date, it is unclear how to design wireless networks efficiently for such novel requirements. One reason is that existing performance models for wireless networks often assume that the rate of the channel code is equal to the Shannon capacity. However, this model does not hold anymore when the packet size and thus blocklength of the channel code is small. Although it is known [1] that finite blocklength has a major impact on the physical layer performance, we lack higher-layer performance models which account in particular for the queueing effects under the finite blocklength regime.

A recently developed methodology [2] provides probabilistic higher-layer delay bounds for fading channels when assuming transmission at the Shannon capacity limit. Based on this novel approach, we develop service process characterizations for fading channels with finite blocklength channel coding, leading to novel probabilistic delay bounds that can give a fundamental insight into the capabilities and limitations of wireless networks when facing low-latency M2M applications. In particular, we show that the Shannon capacity model significantly overestimates the delay performance for such applications, which would lead to insufficient resource allocations. Finally, based on our (validated) analytical model, we study various important parameter trade-offs highlighting the sensitivity of the delay distribution under the finite blocklength regime.

Categories and Subject Descriptors

C.2.1 [**Computer-Communication Networks**]: Network Architecture and Design - Wireless communication

Keywords

Finite blocklength regime, stochastic network calculus, quality of service, queueing systems, fading channels

1. INTRODUCTION

While state-of-the-art wireless systems have mostly been designed for human users, it is expected that next generation systems will be strongly utilized by so-called machine-to-machine communications (M2M). In such applications, automated distributed processes communicate over wireless networks and thus require quite different network features than typical human-related applications. Despite these new requirements, wireless systems offer many advantages for M2M applications, such as reduced cabling cost, increased flexibility, and higher robustness [3]. Thus, M2M applications for traffic safety, the smart electricity grid, or in the context of industrial automation systems are currently of high interest in the domain of 5G cellular networks [4] .

One of the biggest distinguishing factors between M2M and human-related applications are the requirements with respect to the delay. For instance, in factory automation there are often closed-loop control systems, where sensors, controllers, and actuators must exchange information with cycle times (i.e. delays) of 5 ms and below while requiring reliability levels of $1 - 10^{-5}$ and higher (with respect to the deadline). Despite these tough requirements, packet sizes for these applications are typically rather small, i.e. only a few bytes need to be transmitted per datagram. Thus, the academic and industrial research community faces the question how wireless networks can be designed to support such novel application types, also referred to as *low-latency* applications.

This turns out to be a difficult question. Despite the huge interest, we lack a solid theoretical base for modeling the performance of such systems due to the short time spans and small packet sizes involved. Many existing performance models assume that channel coding can provide error-free transmissions in a noisy channel, and that those codes offer a data rate equal to the Shannon capacity. However, this model only holds in the limit of channel codes with infinite blocklength. In low-latency applications with small packet sizes and small blocklengths, there is always a probability that transmissions fail due to noise. Furthermore, for high reliability, data must be encoded at a rate which is significantly lower than the Shannon capacity. Regarding the pure physical layer behavior, Polyanskiy et al. [1]

MSWiM'15, November 2–6, 2015, Cancun, Mexico.
© 2015 ACM. ISBN 978-1-4503-3762-5/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2811587.2811596.

derived an information-theoretic performance model of the finite blocklength regime, which quantifies these effects.

In order to characterize the possibilities and limitations of wireless networks with respect to low-latency M2M applications, such finite-blocklength performance models need to be extended up to the application layer, where queueing effects are taken into account. One factor that causes queueing is channel fading, which means that the signal strength and thus the data rate of a wireless channel changes randomly over time. In general, it is difficult to analyze the queueing performance of fading channels due to the difficulty of finding a stochastic characterization of the random data rate. When the physical layer model also considers finite blocklength effects, the analysis at the application layer becomes even more challenging.

In this paper, we address this fundamental challenge. Al-Zubaidy et al. [2] recently provided a methodology for wireless network performance analysis in fading channels with the Shannon capacity model. By applying stochastic network calculus in a transform domain, they were able to derive probabilistic delay bounds in closed form. Based on this novel approach, we provide a performance model for wireless systems that operate at finite blocklength. In particular, the core contributions of this paper are:

- We derive probabilistic delay bounds for wireless systems that use channel coding at finite blocklength.
- We provide a fast and efficient method to compute the bounds for Rayleigh fading channels. The computation requires solving an integral, which we accomplished through several series expansions, leading to an infinite number of infinite sums. However, we demonstrate that the series converges very quickly for reasonable channel parameters.
- We validate the analytical delay bounds by simulations.
- Our results quantify the performance difference between the Shannon capacity model and the finite blocklength model in [1]. We show that finite blocklength effects can be significant and must be taken into account, in particular for low-latency M2M applications.

The rest of the paper is structured as follows: In Section 2 we discuss related work. In Section 3 we present the basic assumptions and the problem formulation of our work, while in Section 4 we present a brief review of stochastic network calculus. Our main analytical contribution follows in Section 5, while we validate this work and present further numerical results in Section 6. Finally, we conclude our work in Section 7.

2. RELATED WORK

The characterization of channel codes at finite blocklength by Polyanskiy et al. [1] has renewed the research interest in this area. Most notably, Yang et. al [5] performed studies for finite blocklength coding in fading channels. They analyzed systems where the transmitter does not adapt the rate according to the instantaneous SNR of the channel and computed the maximum achievable rate for a certain error probability. It was found that for many fading distributions, including Rayleigh, the difference between the infinite and the finite blocklength model is very small. In another work [6], they investigated the tradeoff between transmit diversity and the cost of learning the channel. However, none of these results apply to our scenario where the rate is always adapted to the current SNR of the channel. Furthermore, they do not consider queueing effects. Wu and Jindal [7] considered queueing effects in a simple ARQ system but did not address the delay.

Performance analysis of wireless networks in fading channels has often been based on discrete/finite-state channel models such as the Gilbert-Elliott channel or finite-state Markov channels (FSMC), e.g. [8, 9]. However, such discrete models cannot provide exact solutions when the fading channels show a continuous distribution of the SNR.

The (min,×) network calculus approach developed in [2] was used for transmit power minimization for process automation under delay constraints [10]. However, the service was characterized by the infinite blocklength Shannon capacity model.

Finite blocklength effects in wireless networks were studied by Gursoy [11], who applied the effective capacity framework [12] to fading channels at finite blocklength and proved that there is a unique optimal tradeoff between the rate and the error probability. This work is the closest to our work. It was extended to the scenario where a codeword is distributed across multiple coherence blocks [13]. The downside of the effective capacity framework is that it provides results only for constant arrivals and that it analyzes the tail of the delay distribution, meaning that it works only for relatively large delays [12]. Furthermore, the authors in [11] provided no analytical method to compute the effective capacity at finite blocklength, which means that numerical integration is necessary.

Apart from the work by Gursoy, only a few authors have worked on the higher-layer analysis of wireless networks in the finite blocklength regime. Zhang et al. [14] provide a network calculus analysis of an AWGN channel without fading. However, they ignored that the error rate is no longer zero and assumed a deterministic service curve.

3. SYSTEM MODEL

We consider data transmission between a data source, (e.g. a sensor in an industrial automation system) to another device (e.g. a control unit) over a wireless channel. A discrete-time model is used, i.e. time is divided into time slots with duration T. In each time slot i, the source generates a_i data bits and stores them in a queue. Then the queued data bits are transmitted over the wireless channel.

3.1 Wireless Channel Model

The wireless link is modeled as a single-antenna Rayleigh fading channel, where the signal-to-noise ratio (SNR) at the receiver varies over time. We assume a block-fading model where the SNR γ_i remains constant during each time slot and varies independently from one time slot to the other. Hence, the SNR values in different time slots are independent and identically distributed (i.i.d.) with exponential distribution:

$$f(\gamma_i) = \frac{1}{\bar{\gamma}} e^{-\frac{\gamma_i}{\bar{\gamma}}}, \tag{1}$$

where $\bar{\gamma}$ is the average SNR at the receiver, which depends on the transmit power at the source, among other parameters. In each slot, the system transmits N symbols, which consist of n symbols for data transmission and n_h symbols for headers and channel estimation, as well as for feedback and acknowledgments from the receiver. The system thus occupies a bandwidth of N/T [Hz].

In each time slot i, the transmitter uses a channel code of length n and rate R_i to encode the first nR_i bits in the queue and then transmits the codeword to the receiver. The receiver replies with an acknowledgment, which is assumed to be instantaneous and error-free. Furthermore, we assume that the transmitter has perfect estimates of the instantaneous SNR γ_i and adapts the coding rate R_i according to γ_i. A standard rate model that is often applied in wireless networking research, including in [2], assumes that the achievable rate R_i in bits per (complex-valued) symbol is equal to the Shannon capacity of the channel, and no errors occur:

$$R_{i,\text{Shannon}} = \log_2(1+\gamma_i). \tag{2}$$

We will refer to the standard rate model as *Shannon model*. The Shannon capacity is an upper bound for codes which only holds when the blocklength n tends to infinity.

At finite blocklength, there is always a probability $\epsilon > 0$ that a transmission error occurs. This error probability can be reduced by decreasing the rate of the code. It was shown by Polyanskiy et al. [1, Thm. 54] that for an AWGN channel with SNR γ_i at a blocklength n and error probability ϵ, the achievable rate in bits per symbol can be closely approximated by

$$R_i(n,\epsilon) \approx \log_2(1+\gamma_i) - \sqrt{\frac{V}{n}}Q^{-1}(\epsilon)\log_2 e, \tag{3}$$

where γ_i is the instantaneous SNR of the channel, $Q^{-1}(x)$ the inverse of the Gaussian Q-function, and the channel dispersion V is given as[1]

$$V = 1 - \frac{1}{(1+\gamma_i)^2}. \tag{4}$$

The achievable rate expression in Eq. (3) is a tight approximation for an information-theoretic bound. Even though current coding and modulation schemes cannot yet fully achieve this rate, this model provides a much better description than a simple Shannon model. A comparison between the information-theoretic bounds and current LDPC channel codes can be found in [1]. Furthermore, it was also proven that Eq. (3) closely approximates the converse bound, which means that even the best future coding schemes cannot exceed this rate.

At very low SNR γ_i, the expression for $R_i(n,\epsilon)$ can become negative. Therefore, the achievable rate must be lower-bounded by zero:

$$R_i^*(n,\epsilon) = \max\left(R_i(n,\epsilon), 0\right). \tag{5}$$

We assume in the following that there exist codes with blocklength n and error probability ϵ that achieve the rate $R_i^*(n,\epsilon)$ exactly. The throughput in time slot i is $n \cdot R_i^*(n,\epsilon)$ bits if no transmission error occurs.

3.2 Queueing Model

For the system-level analysis of a wireless communication network, we use the same stochastic system-theoretic model as in [2]. The a_i data bits that are generated at the source correspond to the arrival process of the queueing system during time slot i. The departure process d_i describes the number of bits that arrives successfully at the destination. The departures depend both on the number of bits waiting in the queue and on the service offered by the wireless link. The service process s_i is equal to $nR_i^*(n,\epsilon)$ when the transmission is successful and a positive acknowledgment is received. When there is a transmission error, we set s_i to zero. This means that the bits remain in the queue; they will be transmitted again in future time slots. Therefore, all data will eventually be transmitted to the destination and the queueing system is lossless. The wireless link transmits the data from the queue in FIFO (first-in first-out) fashion.

In order to derive delay bounds, we need to define the cumulative arrival, service and departure processes in the time interval $[\tau, t)$:

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

The delay $W(t)$ at time t describes the number of time slots it takes for an information bit arriving at time t to be received at the destination. It is defined as

$$W(t) \triangleq \inf\left\{u > 0 : \quad A(0,t) \le D(0,t+u)\right\}. \tag{6}$$

3.3 Problem Statement

We are interested in finding a probabilistic bound on the delay $W(t)$. Thus, we define a target delay $\hat{w}$. The probability that the delay is larger than $\hat{w}$, i.e. that some data bits are not received within a certain deadline, is denoted by the delay violation probability $p_{\text{v}}(\hat{w})$

$$p_{\text{v}}(\hat{w}) = \mathbb{P}\left\{W(t) > \hat{w}\right\}. \tag{7}$$

We assume that a system is reliable when only a very small percentage $p_{\text{v}}(\hat{w})$ of bits is received after the deadline $\hat{w}$. Our main goal in this work is to find an estimate for the delay violation probability $p_{\text{v}}(\hat{w})$ when the rate of the channel code is given by the finite blocklength model. Furthermore, we investigate how the proposed model can be used to aid the design of communication systems that operate at low delay.

4. STOCHASTIC NETWORK CALCULUS

In this section, we provide an overview of the results derived in [2], where stochastic network calculus in a transform domain was used to derive an upper bound for the delay. Even for coding at infinite blocklength, the major problem in deriving stochastic performance bounds for fading channels is the nonlinear mapping of SNR to achievable rate as $R = \log_2(1 + \text{SNR})$. While the probability distribution for the SNR is usually given in a simple form as in Eq. (1), the statistics of the rate cannot be stated in a simple closed form. This problem remains when the achievable rate approximation $R^*(n,\epsilon)$ for finite blocklengths is used because the mapping is essentially still logarithmic except for some penalty term.

4.1 Network Calculus in the SNR Domain

The authors in [2] solved this problem for infinite blocklength by analyzing the system in the exponential domain, also referred to as *SNR domain*. Instead of describing the

[1] We use a different notation than [1] and put $\log_2 e$ as a separate factor in Eq. (3).

cumulative service and arrival $S(\tau,t)$ and $A(\tau,t)$ in the bit domain, they are converted to the SNR domain as follows:

$$\mathcal{A}(\tau,t) = e^{A(\tau,t)}, \quad \mathcal{S}(\tau,t) = e^{S(\tau,t)}. \tag{8}$$

The arrivals can then be interpreted as a series of power or SNR demands on the system. Due to the exponential function, the cumulative arrival and service processes are now multiplicative instead of additive:

$$\mathcal{A}(\tau,t) = \prod_{i=\tau}^{t-1} e^{a_i}, \quad \mathcal{S}(\tau,t) = \prod_{i=\tau}^{t-1} e^{s_i}. \tag{9}$$

As s_i is usually a logarithmic function of the SNR, switching to the SNR domain (i.e. taking the exponential function) eliminates the logarithm. Then, closed-form statistical analysis becomes possible through stochastic network calculus.

Stochastic network calculus allows the description and analysis of queueing systems through simple linear input-output relations. In the bit domain it is based on a (min,+) dioid algebra on $(\mathbb{R} \cup \{+\infty\})$ where the standard addition is replaced by the minimum (or infimum) and the standard multiplication replaced by addition. Similar to the convolution and deconvolution in standard algebra, there are definitions for convolution and deconvolution operators in (min,+) algebra. The convolution and deconvolution operators in (min,+)-algebra are often used for performance evaluation. The reader is referred to [2] for more information.

In the SNR domain network calculus, the arrival, service and departure processes become multiplicative instead of additive. This requires using a (min,×)-algebra instead of (min,+) where × denotes the standard multiplication. The non-commutative convolution and deconvolution operators are defined as

$$\mathcal{X} \otimes \mathcal{Y}(\tau,t) \triangleq \inf_{\tau \le u \le t} \left\{ \mathcal{X}(\tau,u) \cdot \mathcal{Y}(u,t) \right\}, \tag{10}$$

$$\mathcal{X} \oslash \mathcal{Y}(\tau,t) \triangleq \sup_{u \le \tau} \left\{ \frac{\mathcal{X}(u,t)}{\mathcal{Y}(u,\tau)} \right\}. \tag{11}$$

Many of the input-output relationships of the queueing system can be expressed using these operators. The delay can be bounded as follows [2]:

$$W(t) \le \inf \left\{ u \ge 0 : \mathcal{A} \oslash \mathcal{S}(t+u,t) \le 1 \right\}, \tag{12}$$

which means that the delay violation probability $p_{\mathrm{v}}(w) = \mathbb{P}\left\{ W(t) > w \right\}$ can be bounded as [2]:

$$p_{\mathrm{v}}(w) \le \mathbb{P}\left\{ \mathcal{A} \oslash \mathcal{S}(t+w,t) > 1 \right\}. \tag{13}$$

This bound cannot be computed directly. However, it can be upper-bounded again by using the Mellin transform. The Mellin transform $\mathcal{M}_{\mathcal{X}}(s)$ of a nonnegative random variable $\mathcal{X}$ is defined as [2]

$$\mathcal{M}_{\mathcal{X}}(s) \triangleq \mathbb{E}\left[\mathcal{X}^{s-1} \right]. \tag{14}$$

We denote the Mellin transform of a bivariate process $\mathcal{X}(\tau,t)$ as $\mathcal{M}_{\mathcal{X}}(s,\tau,t)$ and choose values for $s \in \mathbb{R}$.

The Mellin transform is used to formulate the moment bound, which is given for $a > 0$ and $s > 0$ as [2]

$$\mathbb{P}(\mathcal{X} \ge a) \le a^{-s} \mathcal{M}_{\mathcal{X}}(1+s). \tag{15}$$

The moment bound follows directly from Markov's inequality as $\mathbb{P}(\mathcal{X} \ge a) = \mathbb{P}(\mathcal{X}^s \ge a^s)$ for any $s > 0$. The moment bound with $a = 1$ on Eq. (13) results in

$$p_{\mathrm{v}}(w) \le \mathcal{M}_{\mathcal{A} \oslash \mathcal{S}}(1+s, t+w, t). \tag{16}$$

The Mellin transform of the (min,×)-deconvolution of two processes can be upper-bounded for $s > 0$ [2]:

$$\mathcal{M}_{\mathcal{A} \oslash \mathcal{S}}(1+s,\tau,t) \le \sum_{u=0}^{\tau} \mathcal{M}_{\mathcal{A}}(1+s,u,t) \cdot \mathcal{M}_{\mathcal{S}}(1-s,u,\tau). \tag{17}$$

Therefore, a bound on $p_{\mathrm{v}}(w)$ can be computed from the Mellin transforms of the arrival and service processes.

4.2 Mellin Transform of the Arrival Process

Analogue to [2] we focus on $(\sigma(s),\rho(s))$-bounded arrivals where the log-moment generating function (log-MGF) of the cumulative arrivals in the bit domain is bounded by

$$\frac{1}{s} \log \mathbb{E}\left[e^{sA(\tau,t)} \right] \le \rho(s) \cdot (t-\tau) + \sigma(s). \tag{18}$$

To simplify notation, we restrict the following analysis to values (σ,ρ) that are independent of s, which is true for constant arrivals. Using Eqs. (18) and (8), the Mellin transform of the SNR-domain arrival process can be upper-bounded:

$$\mathcal{M}_{\mathcal{A}}(s,\tau,t) = \mathbb{E}\left[\mathcal{A}(\tau,t)^{s-1} \right] \le e^{(s-1)(\rho\cdot(t-\tau)+\sigma)}. \tag{19}$$

4.3 Mellin Transform of the Service Process

When the service, i.e. the achievable rate in time slot i can be written as $s_i = n \cdot \log_2 g(\gamma_i) = \frac{n}{\ln 2} \cdot \ln g(\gamma_i)$, and when the s_i of different time slots are i.i.d. (independent and identically distributed), then the Mellin transform of the cumulative service $\mathcal{M}_{\mathcal{S}}(s,\tau,t)$ can be computed from the Mellin transform of $g(\gamma_i)$, which will be derived in Sec. 5:

$$\begin{aligned}\mathcal{M}_{\mathcal{S}}(s,\tau,t) &= \mathbb{E}\left[\left(\prod_{i=\tau}^{t-1} g(\gamma_i)^{\frac{n}{\ln 2}} \right)^{s-1} \right] \\ &= \mathbb{E}\left[g(\gamma_i)^{\frac{n(s-1)}{\ln 2}} \right]^{t-\tau} = \left(\mathcal{M}_{g(\gamma_i)}\left(1 + \frac{n(s-1)}{\ln 2} \right) \right)^{t-\tau}.\end{aligned} \tag{20}$$

4.4 Delay Bound

When the Mellin transforms of the arrival and service processes are known, one can combine Eq. (17) with Eq. (16), which must hold for all $s \in \mathbb{R}^+$, to compute a bound on the delay violation probability $p_{\mathrm{v}}(w)$:

$$p_{\mathrm{v}}(w) \le \inf_{s>0} \left\{ K\left(s,t+w,t\right) \right\}. \tag{21}$$

where $K\left(s,t+w,t\right)$ is defined as

$$K\left(s,t+w,t\right) \triangleq \sum_{u=0}^{t} \mathcal{M}_{\mathcal{A}}(1+s,u,t) \cdot \mathcal{M}_{\mathcal{S}}(1-s,u,t+w). \tag{22}$$

Note: $K\left(s,t+w,t\right)$ is essentially the right side of Eq. (17), except that the upper limit of the sum was changed from $t+w$ to t. This change was proven in [2].

When using the bounded arrival model in (19), the service model in (20), and

$$Y(s) \triangleq \mathcal{M}_{g(\gamma_i)}\left(1 - \frac{n}{\ln 2} s \right),$$

then $K\left(s,t+w,t\right)$ can be computed as

$$\begin{aligned}K\left(s,t+w,t\right) &\leq \sum_{u=0}^{t} e^{\sigma s}\left(e^{\rho s}\right)^{t-u}\cdot Y(s)^{t+w-u}\\ &= e^{\sigma s}Y(s)^{w}\sum_{v=0}^{t}\left(e^{\rho s}Y(s)\right)^{v}\\ &= e^{\sigma s}Y(s)^{w}\frac{1-\left(e^{\rho s}Y(s)\right)^{t+1}}{1-e^{\rho s}Y(s)}.\end{aligned}$$

The queueing system is stable if

$$e^{\rho s}Y(s) < 1. \tag{23}$$

In a stable queueing system, we can obtain a bound on the function $K\left(s,t+w,t\right)$ by letting $t\to\infty$:

$$K\left(s,t+w,t\right) \leq e^{\sigma s}Y(s)^{w}\frac{1}{1-e^{\rho s}Y(s)}. \tag{24}$$

5. SERVICE CHARACTERIZATION IN THE FINITE BLOCKLENGTH REGIME

When using stochastic network calculus, the computation of delay bounds requires the computation of the Mellin transform of e^{s_i}, i.e. the service process in the SNR domain. We assume in the following that s_i follows the finite blocklength model.

At finite blocklength, there is always a chance that errors will occur. The error probability is denoted as ϵ. In Sec. 5.1, we will show how to compute the Mellin transform of the service process when transmission errors occur.

Apart from the chance that errors occur, coding at finite blocklength also causes a rate loss, which depends on the instantaneous SNR γ_i and makes the computation of the Mellin transform difficult. In Sec. 5.3, we show how to approximate the Mellin transform by approximating the rate loss as a constant. In Sec. 5.4, we compute the Mellin transform through a number of series expansions, which allows approximation of the Mellin transform with arbitrary accuracy.

5.1 Characterization of Transmission Errors

In Sec. 4.3, we showed that if the offered service in the bit domain is given as $s_i = n\cdot\log_2 g(\gamma_i)$, then the Mellin transform of the cumulative service $\mathcal{S}(\tau,t)$ in the SNR domain can be computed from the Mellin transform of $g(\gamma_i)$. However, when coding at finite blocklength there is always a chance that an error occurs. In case of error, the offered service s_i is zero. The service model needs to be modified to describe transmission errors. We use the Bernoulli random variable $Z_i \in \{\text{error},\text{success}\}$ to describe the error event. Then, the service in the bit domain depends on two random variables:

$$s_i = \begin{cases} n\log_2 h(\gamma_i) & \text{if } Z_i = \text{success}\\ 0 & \text{if } Z_i = \text{error}\end{cases}, \tag{25}$$

where $h(\gamma_i)$ will be specified later. Now, the service can be written as $s_i = n\cdot\log_2 g(\gamma_i, Z_i)$ with

$$g(\gamma_i, Z_i) \triangleq \begin{cases} h(\gamma_i) & \text{if } Z_i = \text{success}\\ 1 & \text{if } Z_i = \text{error}\end{cases}. \tag{26}$$

In general, the two random variables γ_i and Z_i might not be independent. However, in this work we restrict the analysis to constant values[2] of the error probability ϵ. When the transmitter chooses a code with rate $R^*(n,\epsilon)$ according to (5), then the rate of this code depends on the SNR γ_i but the error probability of this code is always ϵ. Therefore, the error event Z_i is also independent of γ_i, and the Mellin transform of $g(\gamma_i, Z_i)$ can be computed as

$$\begin{aligned}\mathcal{M}_{g(\gamma_i,Z_i)}(s) &= \mathbb{E}_{\gamma_i,Z_i}\left[g(\gamma_i,Z_i)^{s-1}\right]\\ &= (1-\epsilon)\cdot\mathbb{E}_{\gamma_i}\left[h(\gamma_i)^{s-1}\right]+\epsilon\\ &= (1-\epsilon)\cdot\mathcal{M}_{h(\gamma_i)}(s)+\epsilon. \end{aligned} \tag{27}$$

We already know from Eq. (20) that the Mellin transform of the SNR-domain service process $\mathcal{S}(\tau,t)$ can be computed from the Mellin transform of the $g()$-function, which holds also when the $g()$-function has two arguments. Now, Eq. (27) showed that the Mellin transform of the $g()$-function can in turn be easily computed from $\mathcal{M}_{h(\gamma_i)}(s)$, which we will derive in the following sections.

5.2 Service at Finite Blocklength

When the blocklength n and the error probability ϵ are fixed, the achievable rate is given by Eq. (5):

$$R_i^*(n,\epsilon) = \max\left(\log_2(1+\gamma_i)-\sqrt{\frac{V}{n}}Q^{-1}(\epsilon)\log_2 e, 0\right).$$

We define the constant

$$P = \sqrt{\frac{1}{n}}Q^{-1}(\epsilon) \tag{28}$$

and rewrite Eq. (5):

$$R_i^*(n,\epsilon) = \max\left(\log_2\left(\frac{1+\gamma_i}{e^{\sqrt{V}P}}\right),0\right).$$

Now, define

$$h(\gamma_i) \triangleq \max\left(\frac{1+\gamma_i}{e^{\sqrt{V}P}},1\right). \tag{29}$$

In case the transmission is successful, the service is given as $s_i = nR_i^*(n,\epsilon) = n\log_2 h(\gamma_i)$. In case of error, the service is zero. The Mellin transform of $h(\gamma_i)$ is however difficult to obtain because the channel dispersion V given in Eq. (4) depends on the SNR γ_i.

5.3 Approximation for High SNR Values

Our first approach to find the Mellin transform of $h(\gamma_i)$ approximates the channel dispersion as constant, which is accurate at high SNR values. Note that the second term of the channel dispersion V approaches zero when the SNR is high. Thus, at high SNR we can approximate the channel dispersion as

$$V = 1-\frac{1}{(1+\gamma_i)^2}\approx 1. \tag{30}$$

The penalty term $\sqrt{V}P$ becomes equal to the constant P.

[2] Note that varying ϵ with SNR, e.g. allowing more errors when the channel is bad, might result in better performance. Investigating this effect is left for future work.

For Rayleigh fading channels, γ_i is distributed according to Eq. (1) and the Mellin transform of $h(\gamma_i)$ is approximately

$$\begin{aligned}\mathcal{M}_{h(\gamma_i)}(s) &\approx \mathbb{E}\left[\left(\max\left(\frac{1+\gamma_i}{e^P},1\right)\right)^{s-1}\right]\\ &= \int_0^{e^P-1}\frac{1}{\bar{\gamma}}e^{-\frac{\gamma_i}{\bar{\gamma}}}d\gamma_i + \int_{e^P-1}^{\infty}\left(\frac{1+\gamma_i}{e^P}\right)^{s-1}\frac{1}{\bar{\gamma}}e^{-\frac{\gamma_i}{\bar{\gamma}}}d\gamma_i.\end{aligned}$$

The first integral is simply the cumulative distribution function of the SNR. Denote the second integral as $B_0(s)$:

$$\begin{aligned}B_0(s) &= \int_{e^P-1}^{\infty}\left(\frac{1+\gamma_i}{e^P}\right)^{s-1}\frac{1}{\bar{\gamma}}e^{-\frac{\gamma_i}{\bar{\gamma}}}d\gamma_i\\ &= e^{\frac{1}{\bar{\gamma}}}\left(\frac{\bar{\gamma}}{e^P}\right)^{s-1}\int_{e^P}^{\infty}\left(\frac{u}{\bar{\gamma}}\right)^{s-1}\frac{1}{\bar{\gamma}}e^{-\frac{u}{\bar{\gamma}}}du\\ &= e^{\frac{1}{\bar{\gamma}}}\left(\frac{\bar{\gamma}}{e^P}\right)^{s-1}\int_{\frac{e^P}{\bar{\gamma}}}^{\infty}q^{s-1}e^{-q}dq\\ &= e^{\frac{1}{\bar{\gamma}}}\left(\frac{\bar{\gamma}}{e^P}\right)^{s-1}\Gamma\left(s,\frac{e^P}{\bar{\gamma}}\right),\end{aligned} \tag{31}$$

where $\Gamma(s,x)$ denotes the upper incomplete gamma function:

$$\Gamma(s,x) = \int_x^{\infty} q^{s-1}e^{-q}dq. \tag{32}$$

The Mellin transform of $h(\gamma_i)$ is then

$$\mathcal{M}_{h(\gamma_i)}(s) \approx 1 - e^{-\frac{e^P-1}{\bar{\gamma}}} + B_0(s). \tag{33}$$

Observe that if we allow $P = 0$, we obtain the Mellin transform of the service process with the infinite blocklength model, which is given in [2].

5.4 Extension to all SNR Values

In order to extend the previous result to lower SNR values, a series expansion for the square root channel dispersion $\sqrt{V}$ is used, which is based on the following expansion of $\sqrt{1-x}$ for $-1 \le x \le 1$ (Formula 1.110 in [15]):

$$\begin{aligned}\sqrt{1-x} &= 1 + \sum_{j=1}^{\infty}\binom{1/2}{j}(-x)^j\\ &= 1 - \frac{x}{2} - \frac{x^2}{8} - \frac{x^3}{16}\dots\end{aligned}$$

With the definition[3]

$$b_j \triangleq \left|\binom{1/2}{j}\right| = \left|\frac{\left(\frac{1}{2}\right)\left(\frac{1}{2}-1\right)\dots\left(\frac{1}{2}-j+1\right)}{j!}\right|,$$

one can write $\sqrt{V}$ as

$$\sqrt{V} = \sqrt{1-\frac{1}{(1+\gamma_i)^2}} = 1 - \sum_{j=1}^{\infty}\frac{b_j}{(1+\gamma_i)^{2j}}. \tag{34}$$

The convergence of the series in Eq. (34) to the actual value is illustrated in Fig. 1. If all terms of the infinite series in (34) are ignored, then $\sqrt{V}$ is approximated to 1. This corresponds to the approximation for high SNR as discussed in the previous section. In order to get a better approximation, the terms in the infinite sum need to be included. It can be seen that three terms in the sum already lead to a very tight approximation when the instantaneous SNR γ_i is above -10 dB.

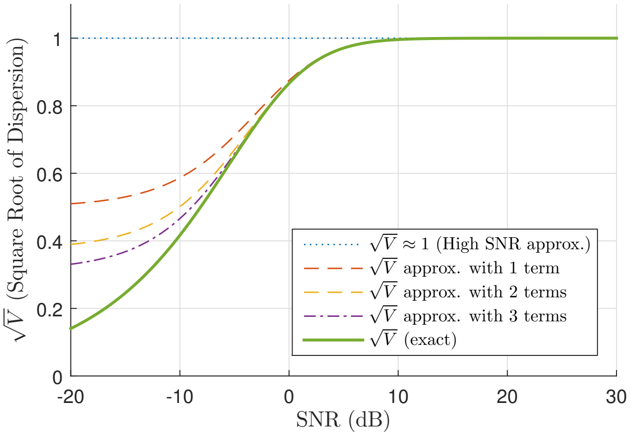

Figure 1: Approximation for $\sqrt{V}$ when the series in Eq. (34) is limited

A useful property is that when the series is approximated by a limited number of terms, the approximation is always larger than the actual value. The rate is reduced by a term that is linear in $\sqrt{V}$, so the approximation always underestimates the achievable rate. Therefore, the approximation leads to higher delays, and creates a valid upper bound on the delay.

The series expansion is used to compute the Mellin transform of $h(\gamma_i)$ given in (29). First, we must find the point where the maximum in $h(\gamma_i)$ becomes greater than 1. When the high SNR approximation was used, this point was easily found at $\gamma_i^* = e^P - 1$. Now, we suggest to do a simple line search. Near this point, the achievable rate is close to 0, and thus any inaccuracies will only have minor impact on the result. We assume that the point is found at $\gamma_i^* = e^{P'} - 1$ for some value P'.

Then, the Mellin transform of $h(\gamma_i)$ is:

$$\mathcal{M}_{h(\gamma_i)}(s) = \int_0^{e^{P'}-1}\frac{1}{\bar{\gamma}}e^{-\frac{\gamma_i}{\bar{\gamma}}}d\gamma_i + B(s)$$

with

$$\begin{aligned}B(s) &= \int_{e^{P'}-1}^{\infty}\left(\frac{1+\gamma_i}{e^{\sqrt{V}P}}\right)^{s-1}\frac{1}{\bar{\gamma}}e^{-\frac{\gamma_i}{\bar{\gamma}}}d\gamma_i\\ &= \int_{e^{P'}-1}^{\infty}\left(\frac{1+\gamma_i}{e^P}e^{\sum_{j=1}^{\infty}\frac{b_jP}{(1+\gamma_i)^{2j}}}\right)^{s-1}\frac{1}{\bar{\gamma}}e^{-\frac{\gamma_i}{\bar{\gamma}}}d\gamma_i\\ &= \int_{e^{P'}-1}^{\infty}\left(\frac{1+\gamma_i}{e^P}\right)^{s-1}\prod_{j=1}^{\infty}e^{\frac{b_jP(s-1)}{(1+\gamma_i)^{2j}}}\frac{1}{\bar{\gamma}}e^{-\frac{\gamma_i}{\bar{\gamma}}}d\gamma_i,\end{aligned} \tag{35}$$

where we used the power series of Eq. (34). Now, for each factor in the infinite product, the series expansion of the exponential function is applied, and the factors that do not depend on γ_i are denoted as $c_{j,k}$:

$$e^{\frac{b_jP\cdot(s-1)}{(1+\gamma_i)^{2j}}} = \sum_{k_j=0}^{\infty}\frac{1}{k_j!}\left(\frac{b_jP\cdot(s-1)}{(1+\gamma_i)^{2j}}\right)^{k_j} = \sum_{k_j=0}^{\infty}\frac{c_{j,k_j}}{(1+\gamma_i)^{2jk_j}}. \tag{36}$$

In addition, define the variables

[3] The signs of the binomial coefficient and $(-x)^j$ are always opposite

$$h_{j,k} \triangleq \frac{c_{j,k}}{\bar{\gamma}^{2\cdot j\cdot k}} = \frac{1}{k!}\cdot\left(\frac{b_j P\cdot(s-1)}{\bar{\gamma}^{2\cdot j}}\right)^k, \tag{37}$$

$$\mu \triangleq \left(\frac{\bar{\gamma}}{e^P}\right)^{s-1}\cdot e^{\frac{1}{\bar{\gamma}}}. \tag{38}$$

To see how the integral can be solved, first assume that the product in $B(s)$ includes only the first factor $j=1$. Call this integral $B_1(s)$:

$$B_1(s) = \int_{e^{P'}-1}^{\infty} \left(\frac{1+\gamma_i}{e^P}\right)^{s-1} e^{\frac{b_1 P(s-1)}{(1+\gamma_i)^2}} \frac{1}{\bar{\gamma}} e^{-\frac{\gamma_i}{\bar{\gamma}}} d\gamma_i$$

$$= \sum_{k_1=0}^{\infty} c_{1,k_1} \int\limits_{e^{P'}-1}^{\infty} \left(\frac{1+\gamma_i}{e^P}\right)^{s-1} \frac{1}{(1+\gamma_i)^{2\cdot 1\cdot k_1}} \cdot \frac{1}{\bar{\gamma}} e^{-\frac{\gamma_i}{\bar{\gamma}}} d\gamma_i$$

$$= \sum_{k_1=0}^{\infty} \frac{c_{1,k_1}}{\bar{\gamma}^{2\cdot 1\cdot k_1}} \cdot \left(\frac{\bar{\gamma}}{e^P}\right)^{s-1} \int\limits_{e^{P'}-1}^{\infty} \left(\frac{1+\gamma_i}{\bar{\gamma}}\right)^{s-1-2k_1} \frac{1}{\bar{\gamma}} e^{-\frac{\gamma_i}{\bar{\gamma}}} d\gamma_i$$

$$= \sum_{k_1=0}^{\infty} h_{1,k_1} \cdot \left(\frac{\bar{\gamma}}{e^P}\right)^{s-1} \cdot e^{\frac{1}{\bar{\gamma}}} \int\limits_{\frac{e^{P'}}{\bar{\gamma}}}^{\infty} q^{s-1-2k_1} e^{-q} dq$$

$$= \sum_{k_1=0}^{\infty} h_{1,k_1} \cdot \mu \cdot \Gamma\left(s - 2\cdot 1\cdot k_1, \frac{e^{P'}}{\bar{\gamma}}\right).$$

When the first J factors ($j = 1,\ldots,J$) in Eq. (35) are used, and each factor is expanded according to Eq. (36), then there are J sums, and each term includes the factor

$$\prod_{j=1}^{J} \frac{c_{j,k_j}}{(1+\gamma_i)^{2jk_j}} = \left(\frac{\bar{\gamma}}{1+\gamma_i}\right)^{\sum_{j=1}^{J} 2jk_j} \prod_{j=1}^{J} h_{j,k_j}. \tag{39}$$

Then, similar to the computation of $B_1(s)$, the integral $B_J(s)$ can be computed as

$$B_J(s) = \sum_{k_J=0}^{\infty}\cdots\sum_{k_1=0}^{\infty} \mu\cdot\Gamma\left(s - \sum_{j=1}^{J} 2jk_j, \frac{e^{P'}}{\bar{\gamma}}\right)\cdot\prod_{j=1}^{J} h_{j,k_j}. \tag{40}$$

To obtain $B(s)$, let J go to infinity. $B(s)$ contains an infinite number of infinite sums. For practical SNR values, it is however sufficient to compute only very few terms, as the value of the incomplete gamma function decreases very quickly with j and k_j. We suggest to include only terms with $\sum_{j=1}^{\infty} 2jk_j \le L$, e.g. $L = 10$, which allows fast calculations but still gives tight approximations for reasonable SNR values. Note that the computation of the incomplete gamma function $\Gamma(s,x)$ needs itself a numerical approximation. However, when computing $B_J(s)$, $\Gamma(s,x)$ needs to be computed only for the first term through this numerical approximation. The other terms can be computed much faster by using the recurrence relation, e.g. [16, Eq. 6.5.21].

6. NUMERICAL RESULTS

In this section, we evaluate the bounds on the delay violation probability numerically and compare it to simulation results. Unless stated otherwise, we use a blocklength of $N = 168$ symbols and set the length of one time slot to 1 ms. The choice of 168 symbols is inspired by the size of a resource block in an LTE system, which contains $12\cdot 7 = 84$ symbols and lasts 0.5 ms [17]. We assume that the channel stays constant for 1 ms, i.e. two LTE resource blocks, and then changes to a different value. Furthermore, for most of our results the n_h overhead symbols are ignored. Thus, the number of symbols for the channel code n is also 168.

For the arrivals, e.g. data generated at a sensor, we assume a constant and periodic process where in each time slot a packet of size a bits arrives into the queue. The Mellin transform of the arrival process is then given by (19), which is satisfied with equality, with $\rho(s) = a$ and $\sigma(s) = 0$.

The simulations use the same channel model as used by the analysis, i.e. the coding rate is assumed to be equal to information-theoretic bound in Eq. (3) for finite blocklengths.

6.1 Validity of the Bounds

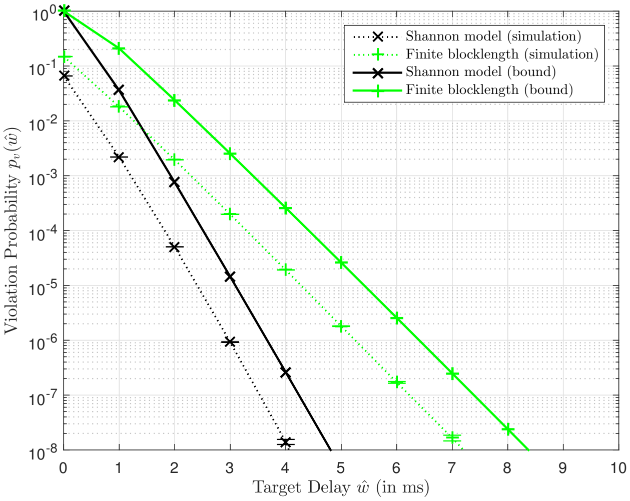

Figure 2: Simulation results and delay bounds for average SNR $\bar{\gamma} = 2$ dB, with arrivals $a = 24$ bits. $T = 1$ ms, $n = 168$. The optimal error probability was found at $\epsilon = 0.0138$. 10^{11} simulations were performed.

For an average SNR of 2 dB, Fig. 2 shows the delay violation probability $p_\text{v}(\hat{w})$ for different target delays $\hat{w}$. In each time slot, $a = 24$ bits were generated at the source. When the effects of coding at finite blocklength are taken into account, the delay increases significantly. The analytical bounds that were obtained with the Shannon capacity model would underestimate the actual delays.

Fig. 3 shows a similar effect at an average SNR of 10 dB with $a = 240$ bits arriving in each time slot. Here, the differences between the Shannon model and the finite blocklength model are smaller than in Fig. 2. This result is reasonable: at high SNR, the absolute rate penalty of finite blocklength codes is nearly constant. Thus, with higher SNR and higher rate, the relative penalty becomes smaller.

The analytical bounds are extremely useful for predicting the system performance even though there is a difference between the analytical bounds and the simulation results. The difference was also observed in [2] and [18] and seems to be unrelated to the finite blocklength model. Despite the difference, the slope of exponential decay of the delay violation

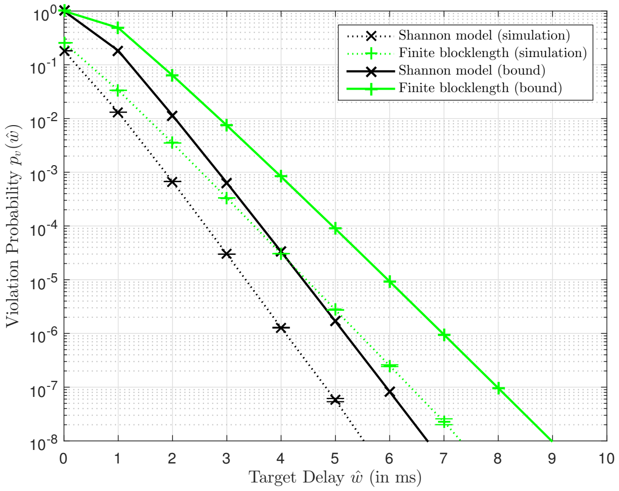

Figure 3: Simulation results and delay bounds for average SNR $\bar{\gamma} = 10$ dB, with arrivals $a = 240$ bits. $T = 1$ ms, $n = 168$. The optimal error probability was found at $\epsilon = 0.0046$. 10^{11} simulations were performed.

probability matches with the simulation results. In addition, the horizontal distance between analytical and simulation results is small. When the bounds predict a delay of e.g. 8 ms for a certain delay violation probability, the actual delay is e.g. 6 or 7 ms. Most importantly, the analytical results provide upper bounds for the delay violation probability, so a system that achieves the rate $R(n, \epsilon)$ will perform better than those bounds. When allocating resources, we would rather allocate a bit more resources than necessary and get a system that performs better than required.

To compute the delay bounds efficiently, we must use an approximation for the integral $B(s)$, which uses the infinite sum in Eq. (40) where only terms with $\sum_{j=1}^{\infty} 2jk_j \leq L$ are considered. The resulting delay bounds for different values of L are shown in Fig. 4 for different SNR values. At each SNR, the maximum possible size a of the arriving packets was chosen such that the best approximation with $L = 20$ still satisfies the delay requirements $\hat{w} = 7$ ms, $p_{\text{v}}(\hat{w}) \leq 10^{-5}$. When using fewer approximation terms, the delay bounds become more loose. Those bounds are still valid upper bounds, but they do not estimate the actual performance of the system well. It can also be seen that very few terms are sufficient. For the selected parameters, we find that the approximation with $L = 8$ is already very accurate. At high SNR, even simpler approximations are acceptable.

6.2 Resource Allocation

An analytical method can help a system in deciding how much bandwidth and resources must be allocated to a certain application. When dimensioning a system, we require that $p_{\text{v}}(\hat{w})$ must be smaller than some target delay violation probability $\hat{p}$. Thus, the delay requirements are represented by the tuple $(\hat{w}, \hat{p})$.

In Fig. 5 we show the minimum average SNR at the receiver for different requirements on the delay. We set a fixed target delay violation probability of $\hat{p} = 10^{-5}$, but we vary the target delay $\hat{w}$ at which the system must satisfy this target. In each time slot $T = 1$ ms, a packet with $a = 120$ arrives. When the system demands very small delays $\hat{w}$, the required SNR increases, so the transmitter should choose a higher transmit power. For very small target delays, the difference between the Shannon model and the finite blocklength model is more than 3dB. This is a significant difference that must be taken into account when allocating resources.

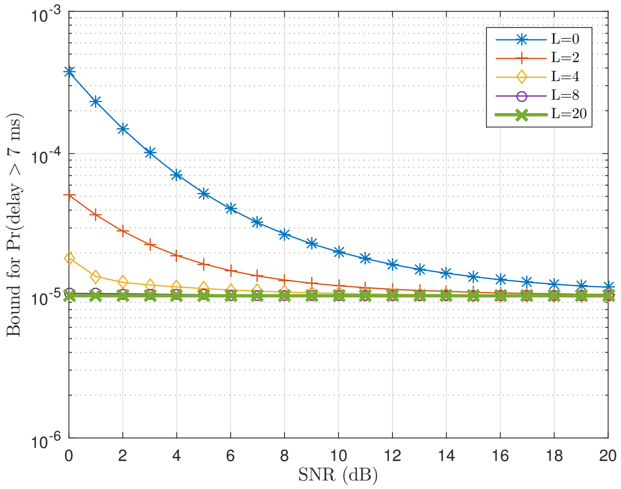

Figure 4: Delay bounds for finite blocklength $n = 168$ with different approximations for the integral $B(s)$. At each SNR, the packet size a was chosen such that the delay bound with the tightest approximation $L = 20$ is at $p_{\text{v}}(\hat{w}) = 10^{-5}$ for a target delay $\hat{w} = 7$ ms.

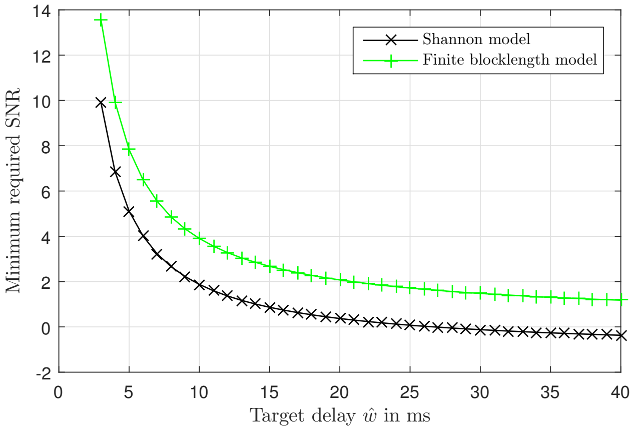

Figure 5: Minimum SNR for different requirements on the delay. For each target delay $\hat{w}$, we required $p_{\text{v}}(\hat{w}) \leq 10^{-5}$. $n = 168$, $a = 120$ bits.

Instead of allocating more resources to the system, perhaps there is a way to reduce the demand on the system. In the example of a sensor that generates data, one could perhaps reduce the accuracy of the sensor readings in order to meet the delay requirements. This is shown in Fig. 6 for different SNR values and fixed delay requirements of $\hat{w} = 7$ ms and $\hat{p} = 10^{-5}$. Here, the Shannon model would again overestimate the performance of the system.

6.3 Optimal Error Probability

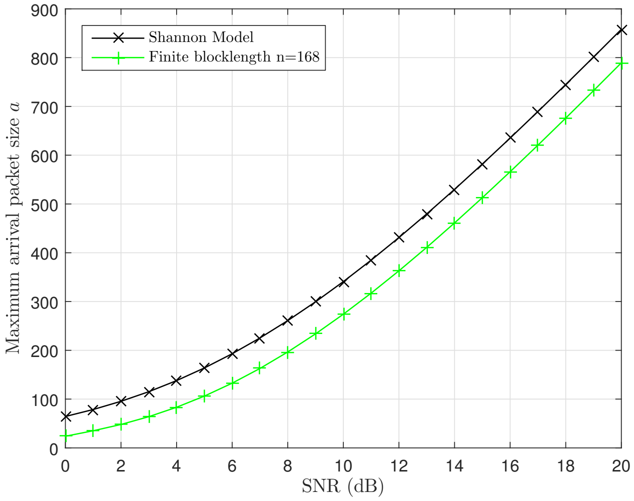

Figure 6: Maximum packet size a (in bits) of the arrival traffic for different SNR. $n = 168$, $T = 1$ ms. Delay requirements: $\hat{w} = 7$ ms and $\hat{p} = 10^{-5}$.

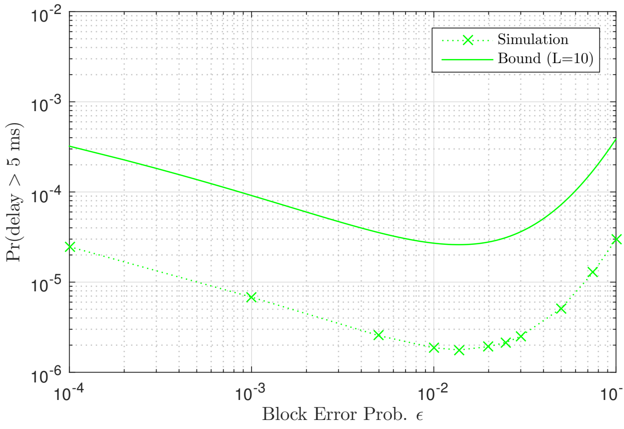

Figure 7: The delay violation probability $p_v(\hat{w})$ for $\hat{w} = 5$ ms depends on the the block error probability ϵ. Parameters: $\bar{\gamma} = 2$ dB, $T = 1$ ms, $n = 168$, $a = 24$ bits. $2 \cdot 10^{10}$ simulations were performed.

When coding at finite blocklength, there is always a probability ϵ that the data cannot be decoded at the receiver. Fig. 7 shows that the analytical delay bounds can be used to find an optimal value for ϵ. For a blocklength $n = 168$, SNR 2 dB and $a = 24$ bits arriving in every time slot, the optimum is at $\epsilon = 0.0138$. For higher values of ϵ, too many transmissions are lost, and the delay increases. For smaller values of ϵ, the system chooses very small transmission rates, such that the queue cannot be served fast enough and the delay also increases. Our simulation results confirm that the analytical bounds can be used to find the value of ϵ leads to the best performance.

In all results presented so far, we have chosen the optimum value for ϵ. Fig. 8 shows the optimal values that were used for the analysis in Fig. 5. It shows that when the delay requirements are very strict (small $\hat{w}$), the system should operate at a very small error probability ϵ and thus at very low rate. When larger delays are acceptable, then the system performs best when it uses higher rate and accepts a higher probability of error.

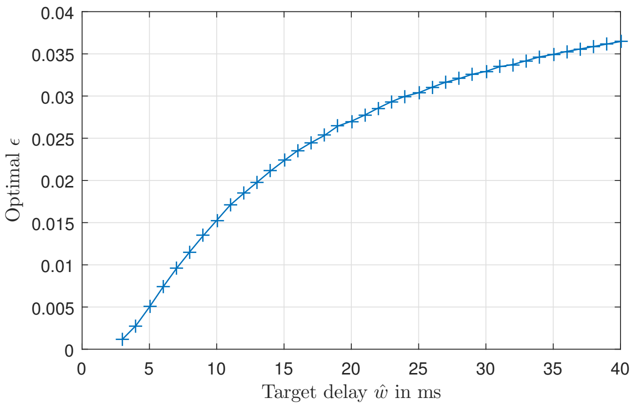

Figure 8: Optimal ϵ for the parameters in Fig. 5

6.4 Optimizing the Blocklength

In a fading channel, the instantaneous SNR varies randomly over time. If this instantaneous value remains too small for some time, then only very little data can be transmitted, and the data experiences a long delay. By making the channel variations faster, it becomes less likely to experience a long delay. Although a system cannot influence directly how fast the channel changes, it can employ frequency hopping. The SNR values of channels at sufficiently different frequencies can be assumed to be independent.

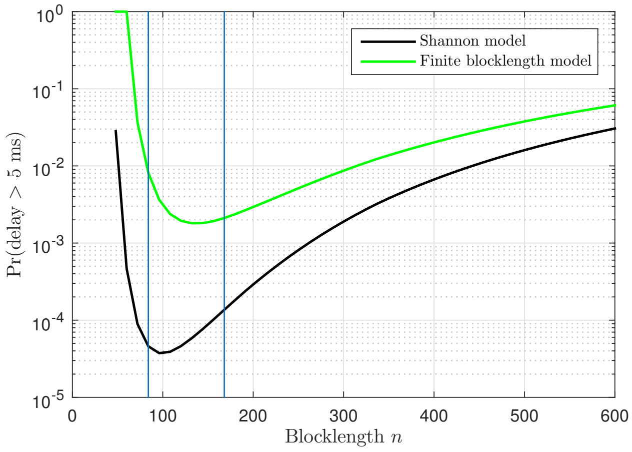

Figure 9: Delay violation probability at $\hat{w} = 5$ ms for different blocklengths. $\bar{\gamma} = 10$ dB, $a = 150$ bits, $n_h = 84$ symbols. The blue lines are located at $n = 84$ and $n = 168$.

Without the impact of finite blocklength effects and without any overhead from metadata, channel estimation and feedbacks, a queueing system should change the channel as quickly as possible, i.e. change the frequency as often as possible. This is no longer true when the overhead and the effects of finite blocklength are taken into account. In that case, varying the channel too quickly creates a lot of overhead. On the other hand, the channel still needs to change often enough to avoid long delays. How quickly should it change?

In the following example, we assume that the channel remains constant for a long time, but after each transmission of duration T, frequency hopping is employed so that the SNR changes to a different and independent value. As in the previous examples, we assume that it takes 1 ms to transmit $12 \cdot 14 = 168$ symbols. However, we change the number of symbols n in multiples of 12. We assume that the overhead is always $n_h = 84$ symbols. Thus, the duration T of one time slot is now assumed to be $\frac{n+n_h}{168}$ ms.

Fig. 9 shows the delay violation probabilities for a target delay of 5 ms for different blocklengths. When considering the Shannon model, the best performance is obtained for a blocklength of $n = 96$ symbols. Thus, it would be best to transmit $84 + 96$ symbols, so $T = 1.07$ ms, and then hop to a different frequency. In contrast to that, when finite blocklength effects are considered, the optimal blocklength is at 132 symbols, so the system should transmit a block of $84 + 132$ symbols ($T = 1.29$ ms). The difference might increase further when the system is only allowed to change the duration in multiples of 0.5 ms, as visualized by the vertical blue lines. Then the system would choose $T = 1$ ms with the Shannon model, but the finite blocklength model would perform best with $T = 1.5$ms.

7. CONCLUSION

In this work, we use a stochastic network calculus approach to compute probabilistic delay bounds for delay sensitive wireless systems in terms of the fading channel parameters. We provide a service characterization for the underlying fading channel in the finite blocklength regime. We then use a recently developed (min,×) network calculus methodology to compute the desired bounds. The finite blocklength channel model leads to analytical challenges that we overcome by using multiple series expansions which converge quite rapidly. We use simulations to validate the obtained delay bounds. Our analysis shows that the infinite blocklength assumption can significantly overestimate the performance of a system that operates at finite blocklength, especially at low SNR. Thus, low-latency M2M applications require more power or bandwidth than a simpler analysis with infinite blocklength models would predict.

The obtained results can be used for network dimensioning and parameters optimization. In future work, the obtained results can be extended to multi-hop settings with cross traffic and variable arrival traffic. Further research may also investigate less idealized channel models with effects like time-correlated fading, imperfect channel state information, or lost acknowledgment packets.

8. REFERENCES

[1] Y. Polyanskiy, H. V. Poor, and S. Verdú, “Channel coding rate in the finite blocklength regime,” *IEEE Trans. Inf. Theory*, vol. 56, pp. 2307–2359, May 2010.

[2] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, “Network-Layer Performance Analysis of Multihop Fading Channels,” *IEEE/ACM Trans. Netw.*, Oct. 2014. to appear. doi: 10.1109/TNET.2014.2360675.

[3] J. Akerberg, M. Gidlund, and M. Bjorkman, “Future research challenges in wireless sensor and actuator networks targeting industrial automation,” in *Proc. IEEE Int. Conf. on Industrial Informatics (INDIN)*, vol. 9, pp. 410–415, Jul. 2011.

[4] 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, pp. 26–35, May 2014.

[5] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, “Quasi-Static Multiple-Antenna Fading Channels at Finite Blocklength,” *IEEE Trans. Inf. Theory*, vol. 60, pp. 4232–4243, Jul. 2014.

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

[7] P. Wu and N. Jindal, “Coding versus ARQ in fading channels: how reliable should the PHY be?,” *IEEE Trans. Commun.*, vol. 59, pp. 3363–3374, Dec. 2011.

[8] J. G. Kim and M. M. Krunz, “Bandwidth allocation in wireless networks with guaranteed packet-loss performance,” *IEEE/ACM Trans. Netw.*, vol. 8, pp. 337–349, Jun. 2000.

[9] M. Hassan, M. M. Krunz, and I. Matta, “Markov-based channel characterization for tractable performance analysis in wireless packet networks,” *IEEE Trans. Wireless Commun.*, vol. 3, pp. 821–831, May 2004.

[10] Petreska, Neda and Al-Zubaidy, Hussein and Gross, James, “Power minimization for industrial wireless networks under statistical delay constraints,” in *Proc. Int. Teletraffic Congr. (ITC)*, pp. 1–9, 2014.

[11] M. C. Gursoy, “Throughput analysis of buffer-constrained wireless systems in the finite blocklength regime,” *EURASIP Journal on Wireless Communications and Networking*, Dec. 2013.

[12] D. Wu and R. Negi, “Effective capacity: a wireless link model for support of quality of service,” *IEEE Trans. Wireless Commun.*, vol. 2, pp. 630–643, Jul. 2003.

[13] D. Qiao, M. C. Gursoy, and S. Velipasalar, “Channel coding over multiple coherence blocks with queueing constraints,” in *Proc. IEEE Int. Conf. on Communications (ICC)*, pp. 1–5, Jun. 2011.

[14] Y. Zhang and Y. Jiang, “Performance of data transmission over a gaussian channel with dispersion,” in *Proc. Int. Symp. on Wireless Communication Systems (ISWCS)*, pp. 721–725, VDE, Aug. 2012.

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

[16] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions.* Courier Corporation, 1964.

[17] S. Sesia, M. Baker, and I. Toufik, *LTE-The UMTS Long Term Evolution: From Theory to Practice.* John Wiley & Sons, 2011.

[18] 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. on Communications (ICC)*, Jun. 2015. Online: http://www.researchgate.net/publication/275648083.