Effective Rate of URLLC with Short Block-Length Information Theory

Najib Odhah
System Architecture Department
IHP – Leibniz-Institut für innovative Mikroelektronik, Frankfurt Oder, Germany
odhah@ihp-microelectronics.com

Eckhard Grass
System Architecture Department
IHP – Leibniz-Institut für innovative Mikroelektronik, Frankfurt Oder, Humboldt University of Berlin, Berlin, Germany
grass@ihp-microelectronics.com

Rolf Kraemer
System Architecture Department
IHP – Leibniz-Institut für innovative Mikroelektronik, Frankfurt Oder, Germany
rolf.kraemer@me.com

***Abstract*— Shannon channel capacity estimation, based on large packet length is used in traditional Radio Resource Management (RRM) optimization. This is good for the normal transmission of data in a wired or wireless system. For industrial automation and control, rather short packages are used due to the short-latency requirements. Using Shannon's formula leads in this case to inaccurate RRM solutions, thus another formula should be used to optimize radio resources in short block-length packet transmission, which is the basic of Ultra-Reliable Low-Latency Communications (URLLCs). The stringent requirement of delay Quality of Service (QoS) for URLLCs requires a link-level channel model rather than a physical level channel model. After finding the basic and accurate formula of the achievable rate of short block-length packet transmission, the RRM optimization problem can be accurately formulated and solved under the new constraints of URLLCs. In this short paper, the current mathematical models, which are used in formulating the effective transmission rate of URLLCs, will be briefly explained. Then, using this rate in RRM for URLLC will be discussed.**

Keywords—URLLC, RRM, delay QoS exponent, decoding error rate, delay bound violation probability, short block-length, effective Bandwidth

I. INTRODUCTION

Ultra-Reliable and Low-Latency Communications (URLLC), which is one of the key communication use-cases of the 5th Generation (5G) and Beyond 5G (B5G) , and also planned for the future 6th Generation (6G) cellular networks, will be essential for the development of various emerging mission-critical and Tactile Internet applications[1]. The term Tactile Internet (TI) was firstly coined by Gerhard Fettweis et al in early 2014[2]. It outlines a new operation type of internet applications that use guaranteed Round Trip Delay (RTD) to enable hard real-time operation as required for tactile feedback. In 5G terms, it belongs to the class of URLLC interactions with guaranteed latency.

In URLLC, short packets are transmitted and thus the Shannon based channel capacity formula that was derived using large number theory for long packet transmission cannot be used. An alternative formula, which can define accurately the maximal achievable rate of short packet transmission, must be derived and validated. In the seminal work of [3], the authors deeply investigated the problem of short packets transmission over Additive White Gaussian Noise (AWGN) channel. They derived the upper bound, lower bound, and normal approximation of the maximal achievable rate of short block-length packet transmission. This seminal work has initiated an intensive research work in this topic through the last decade and contributed to the current research efforts of solving the research questions in the third use case of 5G, i.e. critical Machine Type Communications (cMTCs) or URLLCs.

The authors of [4] comprehensively reviewed all the research work, which has been accomplished on this topic since this seminal work. They pursued to study and propose more accurate formulas of the maximal achievable rate bounds and approximations for more realistic environments, i.e. quasi-static and fast fading channels [3]–[10]. They also built a free MATLAB toolbox, which can be used to reproduce their research results for helping the community to pursue the research in this hot research field [11].

In traditional Radio Resource Management (RRM) for long block-length transmission, i.e. for enhanced Mobile Broadband (eMBB), the Shannon based channel capacity and the upper bound of the achievable transmission rate can accurately define the relation between transmission resources, i.e. rate, power, and bandwidth. This relation is the building block of any RRM optimization problem, which can be solved to optimally allocate the radio resources under given system constraints. The design of the recent RRM optimization algorithms for URLLCs required a more accurate rate formula that considers the short block-length information theory. Most research work currently used the normal approximation of the achievable transmission rate as the starting point for the formulation of the optimization problem for URLLC, but this approximation is suitable only for relatively short block-length packet transmission. In very short/short block-length transmission, the normal approximation cannot be applied for an URLLC performance study, thus the authors of [9] proposed a more accurate approximation, which is called saddlepoint approximation. The saddle point approximation of the maximal achievable transmission rate is still under study, and it is not used for RRM till now and it is still a hot research challenge.

Short block-length information theory, which has been initiated by the seminal work of Polyanskiy et al. [3], played a significant role to formulate and solve many RRM optimization problems, but it could not guarantee the stringent delay Quality of Service (QoS) requirement of

real-time communications, i.e. URLLCs. Thus, the maximal achievable rate of short block-length must be derived based on a link-level channel model. The authors in [12] proposed and developed a link-layer channel model called Effective Capacity (EC)/Effective Bandwidth (EB), in which the wireless link is modeled by two functions called the probability of nonempty buffer and the QoS exponent of the wireless link. This cross-layer design will insure the End-to-End (E2E) performance of the wireless link in a real-time manner.

Both, short block-length information theory and EC/EB were individually studied, however, the joint study of both approaches is worthwhile work in progress. In this paper, the two methods will be individually reviewed, then the paper will study how they can be jointly used to derive a good estimate of the effective transmission rate as the basis of the RRM optimization problem formulation for real-time communications. The rest of the paper will be structured as follows: section 2 will review the recent work of short block-length information theory. Section 3 will review the recent work of the EC/EB. Section 4 will study the joint relation between the short block-length information theory and the EC/EB and explain how to use them to find the effective transmission rate that will be used in RRM for URLLC. The paper is concluded in section 5, and the future work is presented.

II. SHORT BLOCK-LENGTH INFORMATION THEORY

In URLLC, the well-known transmission rate performance metric for long block-length wireless communications, i.e. eMBB, will not be accurate and will lead to non-optimal radio resources allocations. Thus, a new performance metrics must be proposed and studied to cope with the advent of new traffic schemes in 5G and B5G. During the last decade, significant progress has been made within the information theory community to address the problem of transmitting short packets. Building upon Dobrushin's and Strassen's previous asymptotic results, Polyanskiy et al. [3] provided a unified approach to obtain tight upper and lower bounds on the maximal achievable rate for short block-length packet transmission. This seminal work initiated so called finite block-length information theory, which addresses the problem of quantifying a new performance metric for short packets transmission and, hence, solves the long-standing problem of accounting for latency constraints in a satisfactory way. The authors of the seminal work showed that for various channels with positive capacity $C(\rho)$, the maximal achievable rate for short block-length transmission can be approximated as [4]

$$R^*(n,\epsilon,\rho) \approx C(\rho) - \sqrt{\frac{V(\rho)}{n}}Q^{-1}(\epsilon) + \mathcal{O}\left(\frac{\log n}{n}\right) \quad (1)$$

Where $V(\rho)$ denotes the channel dispersion, and it is function of the received Signal to Noise Ratio (SNR), i.e. ρ, of the wireless link, as in the channel capacity $C(\rho)$. $\mathcal{O}(\log n/n)$ comprises remainder terms of order $(\log n/n)$, depending totally on the packet length *n*. The approximation (1) shows that to cope with the desired error probability of ϵ for a given packet size *n*, a penalty on the rate should be incurred, which is the second term of the approximation that is proportional to $\sqrt{(1/n)}$. The traditional approach of approximating $R^*(n,\epsilon,\rho) \approx C(\rho)$, for large packet sizes and small packet error rates, allows one to model a communication channel as a bit pipe that delivers reliably $C(\rho)$ bits per channel use. In short block-length regime, the communication channel can be thought of as a bit pipe of randomly varying size. Specifically, the size of the bit pipe behaves as a Gaussian random variable with mean $C(\rho)$ and variance $\sqrt{((V(\rho))/n)}$. Hence, $V(\rho)$ is a measure of the channel dispersion. The packet error probability ϵ is the probability that $R^*(n,\epsilon,\rho)$ is larger than the size of the bit pipe.

The channel capacity $C(\rho)$ and the channel dispersion $V(\rho)$ in the approximation (1) have different formulas according to the nature of the wireless channel. (i.e. if it can be AWGN channel, slow fading channel, or fast fading channel). For deeper understanding the readers are referred to the publication by Durisi et al. [4] and the references therein.

Assuming an AWGN channel, Polyanskiy et al. proved that a good approximation for $R^*(n,\epsilon,\rho)$ can be obtained by replacing the remainder terms on the right-hand side of the approximation (1) by $\log n/2n$, the new approximation is called normal approximation [3]

$$R^*(n,\epsilon,\rho) \approx C(\rho) - \sqrt{\frac{V(\rho)}{n}}Q^{-1}(\epsilon) + \frac{\log n}{2n} \quad (2)$$

The normal approximation is very beneficial as a new performance metric for finite block-length regime, and it has been used in the subsequent research in URLLCs, specifically RRM for the URLLCs[13]–[15]. It provides an accurate results for medium block-length, i.e. $n \geq 200$, but the results are inaccurate for very short block-length, i.e. $n \ll 200$. Lancho et al. [16] proposed another approximation called saddlepoint approximation, and they proved that this approximation outperforms the normal approximation, specifically in case of very short block-length. Moreover, saddlepoint approximation is an accurate performance metric for URLLs over the entire range of the system parameters.

III. LINK-LEVEL CHANNEL CAPACITY

The emergence of delay-sensitive wireless applications requires an efficient modeling of wireless channel that can take into consideration QoS metrics such as delay-violation probability, data rate, and end-to-end delay. Physical-layer channel models, for analyzing the performance of delay-sensitive wireless applications, can be complex and inaccurate. Hence, a new link-layer channel model called EC/EB has been proposed by Wu et al. [12]. They first modeled a wireless link by two EC functions, namely, the probability of nonempty buffer, and the delay QoS exponent of the wireless link. Then, they proposed a simple and efficient algorithm estimating these EC/EB functions. The physical-layer analogies of these two link-layer EC functions are the marginal distribution (e.g., Rayleigh–Ricean distribution) and the Doppler spectrum, respectively. Figure 1 depicts a link-layer channel model, where the data source generates packets at a constant rate μ (arrival rate). Generated packets are first sent to the (infinite) buffer at the transmitter, whose queue length is Q_n, where n here refers to the *nth* sample interval. The packets in the queue is transmitted over the fading channel at data rate r_n (service rate).

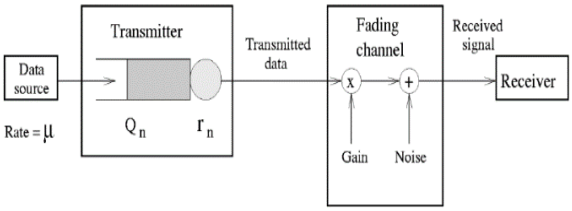

Figure 1 Link-Layer channel model

The EC is the dual concept of the EB, so these two dual mathematical tools will be briefly explained in the subsequent subsections.

A. *Effective Bandwidth (EB)*

The EB is defined as the minimum constant service, i.e., transmission rate over the channel, that is needed to satisfy a given queueing delay requirement for a given source rate (arrival rate). EB is used for obtaining optimal radio resource allocation schemes for delay-sensitive wireless communications.

Wu et. al derived expressions for both EB and EC, then finding reliable relation for the delay violation probability for delay-sensitive wireless communications. The optimal delay QoS exponent θ^* that characterizes the queue length decaying rate can be found by solving the following equation [12]:

$$\Lambda_A(\theta^*) + \Lambda_s(-\theta^*) = 0 \tag{3}$$

The first term is related to arrival process at the queue input, the second term is related to service process at the queue output, i.e. channel input. For a dynamic queuing system, the queue length process, $Q(t)$ converges in distribution to a random variable $Q(\infty)$ such that [17]

$$-\lim_{q\to 0} \frac{\ln(\Pr\{Q(\infty) > q\})}{\mathrm{q}} = \theta \tag{4}$$

Equation (4) states that the probability of the queue length exceeding a certain threshold q decays exponentially fast as q increases and the parameter θ determines the decaying rate.

Also, the delay bound violation probability can be expressed as

$$\epsilon_d = \Pr\{\mathrm{q} > \mathrm{Q}\} = e^{-\theta^* \mathrm{Q}} \tag{5}$$

B. *Effective Capacity (EC)*

The EB defines the minimum service rate that is needed to guarantee a delay requirement for a given source traffic (arrival rate). The EC model, on the other hand, can be used to find the maximum source rate (arrival rate) that the channel can handle (given service rate) with the required delay constraint.

The derivation of the EC will be as that of effective bandwidth, but for arrival process [18]

For more details on The EB/EC, the reader can be referred to the references [12] and [18]

IV. EFFECTIVE RATE FOR REAL-TIME RRM

Figure (*2*.a) shows that a real-time communications system model, which has two kinds of mobile devices, sensors for uploading real-time data, and User Equipments (UEs) for receiving (downloading) these data to act in real-time manner. Figure (*2*.b) shows the End-to-End (E2E) delay $D_{\max}$, which consists of four components, Uplink and Downlink delays, Queuing delay, and backhaul delay, where the required queuing delay can be assumed as [15]

$$D^{\mathrm{q,d}} = D_{\max} - 3T_{\mathrm{f}} \tag{6}$$

Each delay of the other delays is assumed to be one frame time, i.e. T_{f}

The main challenge here is how to find the basic building block formula of the transmission rate that will be used in the RRM for real-time communications. As discussed in the previous sections, the transmission rate here must be the reliable maximal achievable rate for short block-length packet transmission, but taking the queuing delay into the consideration, specifically in the downlink case. In this section, the so-called effective rate for real-time RRM will be discussed and its usage in RRM will be briefly explained. The achievable packet service rate (in packets/frame) of user k in the downlink can be expressed as follows

$$r_k^{\mathrm{d}} \approx \frac{\tau W_k^{\mathrm{d}}}{u \ln 2}\left[\ln(1+\rho) - \sqrt{\frac{1}{\tau W_k^{\mathrm{d}}}} Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,d}}\right)\right] \tag{7}$$

where the received SNR $\rho = \frac{\alpha_k^{\mathrm{d}} g_k^{\mathrm{d}} P_k^{\mathrm{d}}}{\phi N_0 W_k^{\mathrm{d}}}$, α_k^{d} is the large-scale channel gain, g_k^{d} is the small-scale channel gain, P_k^{d} is the transmitted power, ϕ is the SNR loss due to imperfect CSI at the transmitter, N_0 is the single-side noise spectral density, and W_k^{d} is the transmission bandwidth. Equation (7) is a special case of the normal approximation in equation (2), where the channel dispersion is assumed to be one and the packet length is replaced with τW_k^{d}, where τ is the packet time duration. Also, u is the number of bits in each packet, and it is used here to express the rate in packets per frame, instead of bits per second.

As explained in subsection 3, equation (7) alone does not reflect the delay bound and delay bound violation

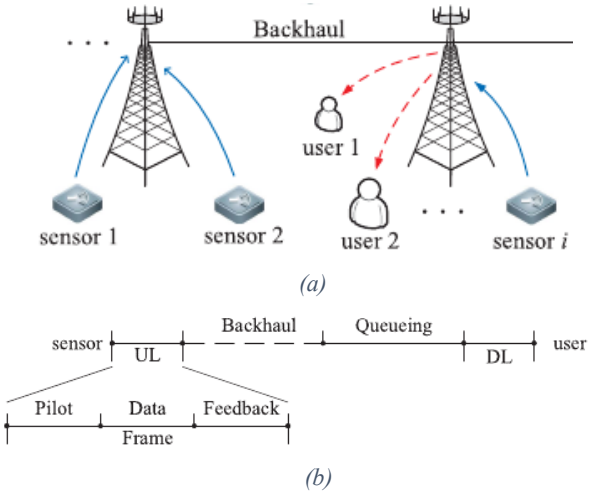

Figure 2 real-time communications. (a) System model and (b) E2E delay

probability incurred by queuing packets at base station, so EB must be used to derive so called the effective transmission rate. The EB (packets/frame) for a Poisson arrival process is given by [15]

$$E_k^{\mathrm{B}} = \frac{T_{\mathrm{f}} \ln\left(1/\varepsilon^{\mathrm{q,d}}\right)}{D^{\mathrm{q,d}} \ln\left[\frac{T_{\mathrm{f}} \ln(1/\varepsilon^{\mathrm{q,d}})}{\mu_k D^{\mathrm{q,d}}} + 1\right]} \tag{8}$$

The constraint, which is reflecting the requirements for queueing delay and DL decoding error probability, can be expressed as the SNR required to support the average packet arrival rate μ_k. The required SNR can be obtained by substituting equation (7) into $r_k^{\mathrm{d}} = E_k^{\mathrm{B}}$,

$$\gamma_k^{\mathrm{d}} = \exp\left[\frac{E_k^{\mathrm{B}} u \ln 2}{\tau W_k^{\mathrm{d}}} + \frac{Q_{\mathrm{G}}^{-1}\left(\varepsilon^{\mathrm{c,d}}\right)}{\sqrt{\tau W_k^{\mathrm{d}}}}\right] - 1 \tag{9}$$

The above required received SNR is accurate for real-time RRM optimization and solution. For more details on how to use the above equation in real-time RRM for URLLC, the reader is referred to [15] .

V. Conclusions and Future work

The two mathematical models, which are short block-length information theory and effective bandwidth/effective capacity, has been revisited and explained. The effective transmission rate, which reflects the reliability and delay requirements of link-level real-time transmission of URLLC packets, has been also briefly discussed and it has been shown that it can be the basic formula for any RRM problem formulation and optimization.

In future work, the effective transmission rate will be further studied with more accurate short block-length transmission rate approximations, such as saddlepoint approximation, which is accurate for all packet block-length packet transmission. Also, simulation results that validate the significance of the effective transmission rate for URLLCs will be presented and discussed.

Acknowledgment

This work is supported by the German Research Foundation (DFG) as part of the 5G-REMOTE project (5G-Enabled Real Time Communications for Tactile Internet). The authors are responsible for the content of the paper.

References

[1] C. She *et al.*, "A tutorial of ultra-reliable and low-latency communications in 6g: Integrating theoretical knowledge into deep learning," *arXiv preprint arXiv:2009.06010*, 2020.

[2] G. P. Fettweis, "The tactile internet: applications and challenges," *IEEE Vehicular Technology Magazine*, vol. 9, no. 1, pp. 64–70, 2014.

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

[4] 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.

[5] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static multiple-antenna fading channels at finite blocklength," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 4232–4265, 2014.

[6] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Transactions on Communications*, vol. 64, no. 2, pp. 618–629, 2015.

[7] M. C. Coskun *et al.*, "Efficient error-correcting codes in the short blocklength regime," *Physical Communication*, vol. 34, pp. 66–79, 2019.

[8] J. Östman, R. Devassy, G. Durisi, and E. G. Ström, "Short-packet transmission via variable-length codes in the presence of noisy stop feedback," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 214–227, 2020.

[9] J. Östman, A. Lancho, G. Durisi, and L. Sanguinetti, "URLLC with Massive MIMO: Analysis and Design at Finite Blocklength," *IEEE Transactions on Wireless Communications*, 2021.

[10] A. Lancho, G. Durisi, and L. Sanguinetti, "Cell-free Massive MIMO with Short Packets," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2021, pp. 416–420.

[11] A. Collins *et al.*, "SPECTRE, short packet communication toolbox," *available at github. com/yp-mit/spectre*, 2016.

[12] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Transactions on wireless communications*, vol. 2, no. 4, pp. 630–643, 2003.

[13] C. She, C. Yang, and T. Q. Quek, "Radio Resource Management for Ultra-reliable and Low-latency Communications," *IEEE Communications Magazine*, vol. 55, no. 6, pp. 72–78, 2017.

[14] C. She, C. Yang, and T. Q. Quek, "Cross-layer Optimization for Ultra-reliable and Low-latency Radio Access Networks," *arXiv preprint arXiv:1703.09575*, 2017.

[15] C. Sun, C. She, C. Yang, T. Q. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Transactions on Wireless Communications*, vol. 18, no. 1, pp. 402–415, 2018.

[16] A. Lancho, J. Östman, G. Durisi, T. Koch, and G. Vazquez-Vilar, "Saddlepoint approximations for short-packet wireless communications," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4831–4846, 2020.

[17] A. Aijaz, "Towards 5G-enabled tactile internet: radio resource allocation for haptic communications," in *Wireless Communications and Networking Conference (WCNC), 2016 IEEE*, 2016, pp. 1–6.

[18] M. Amjad, L. Musavian, and M. H. Rehmani, "Effective capacity in wireless networks: A comprehensive survey," *IEEE Communications Surveys \& Tutorials*, vol. 21, no. 4, pp. 3007–3038, 2019.