Resource Allocation and HARQ Optimization for URLLC Traffic in 5G Wireless Networks

Arjun Anand, *Student Member, IEEE*, and Gustavo de Veciana, *Fellow, IEEE*

***Abstract*— 5G wireless networks are expected to support ultra-reliable low latency communications (URLLC) traffic which requires very low packet delays (< 1 ms) and extremely high reliability (∼99.999%). In this paper, we focus on the design of a wireless system supporting downlink URLLC traffic. Using a queuing network-based model for the wireless system, we characterize the effect of various design choices on the maximum URLLC load it can support, including: 1) system parameters such as the bandwidth, link SINR, and QoS requirements; 2) resource allocation schemes in orthogonal frequency-division multiple access (OFDMA)-based systems; and 3) hybrid automatic repeat request schemes. Key contributions of this paper which are of practical interest are: 1) study of how the minimum required system bandwidth to support a given URLLC load scales with associated QoS constraints; 2) characterization of optimal OFDMA resource allocation schemes which maximize the admissible URLLC load; and 3) optimization of a repetition code-based packet re-transmission scheme.**

***Index Terms*— URLLC, resource allocation, OFDMA, HARQ, wireless networks.**

I. Introduction

5G WIRELESS networks are expected to support a new class of traffic called Ultra Reliable Low Latency Communication (URLLC) for appplications like industrial automation, mission critical traffic, virtual reality, etc., see e.g., [1]–[7]. URLLC traffic requires packet latencies less than 1 msec along with a very high reliability of 99.999 %. Wireless system design to meet such stringent Quality of Service (QoS) requirements is a particularly challenging task and is the focus of this paper. Specifically in this paper we consider downlink transmission of URLLC traffic in a Frequency Division Duplex (FDD) based system with separate frequency bands for uplink and downlink.

The QoS requirements URLLC traffic places on a wireless downlink system are specified as follows: a packet of size L bits must be successfully delivered to the receiver by the Base Station (BS) within a end-to-end delay of no more than d seconds with a probability of at least $1 - \delta$. The delay experienced by a packet includes queuing delay at the BS, transmission duration, receiver processing delay, packet decoding feedback transmission duration, and time to make further re-transmissions as needed. Typical values of QoS parameters mentioned in the literature are $L = 32$ bytes, $d = 1$ msec., and $\delta = 10^{-6}$, see [6] for more details.

This paper investigates how design choices impact the URLLC 'capacity', i.e., the maximum URLLC load the system can support and how this is affected by the stringency of the QoS requirements. In particular, the paper studies the impact of: 1) system bandwidth W, user SINR, QoS parameters d and δ; 2) resource allocation in the time-frequency plane of an Orthogonal Frequency Division Multiple Access (OFDMA) system; and 3), HARQ schemes on URLLC capacity. The three aspects are inter-related, for example, the impact of the system bandwidth W on URLLC capacity depends on the re-transmission schemes being used.

Our aim of understanding of the impact of the system parameters on URLLC capacity will help system engineering to meet URLLC's QoS requirements. Another important aspect which needs careful consideration is how resources are allocated to URLLC transmissions. 5G standards are OFDMA based and hence, users' packets are allocated different parts of a time-frequency plane for data transmission. To send a URLLC packet, one can schedule 'tall' transmissions which use a large swath of bandwidth for a short duration or 'wide' transmissions which use a small bandwidth over a longer duration. 'Tall' transmissions result in reduced transmission times for packets, however, the maximum number of concurrent transmissions is also reduced. This might result in queuing or blocking of URLLC packets due to the immediate un-availability of bandwidth. By contrast, 'wide' transmissions permit a higher number of concurrent transmissions but with longer transmission times for each packet. Hence, the average number of packets in the system is higher with 'wide' transmissions than with 'tall' transmissions, which may again lead to bandwidth scarcity. Further the transmission duration for 'wide' transmissions are constrained by d. Therefore, one would like to analyze the trade-off between 'tall' and 'wide' packet transmissions.

Finally studying the impact of HARQ schemes on URLLC capacity can help evaluate various design choices, such as the maximum number of re-transmissions allowed and the reliability (coding scheme) one should target after each transmission. This paper proposes an analytical framework to study the above mentioned design choices and trade-offs.

A. Related Work

Questions surrounding URLLC traffic have recently received a lot of attention. The 3GPP standards committee

Manuscript received April 4, 2018; revised September 9, 2018; accepted September 22, 2018. Date of publication October 12, 2018; date of current version November 30, 2018. This work was supported by FutureWei Technologies. *(Corresponding author: Arjun Anand.)*

The authors are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX 78712 USA (e-mail: arjunanand1989@gmail.com; gustavo@ece.utexas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSAC.2018.2874122

has recognized the need for a new OFDMA based frame structure to support such traffic, see e.g., [5] for a discussion of various proposals. In particular to meet the stringent latency constraints of URLLC traffic, they have specified a *mini-slot* level access to radio resources for URLLC traffic with mini-slot durations of $0.125 - 0.250$ msec. This is different from the standard *slot* level access to radio resources for eMBB traffic which has slot durations of 1 msec or higher.

System level designs for URLLC networks have been explored in [7]–[12]. Bennis *et al.* [7] have surveyed the literature on URLLC traffic and have elaborated on the various technologies and methodologies related to URLLC system design. Durisi *et al.* [12] discuss information theoretic results on sending short packets. They also discuss protocols to transmit small length packets between two nodes in a downlink broadcast setting and for random access based uplink. However, they do not focus on optimizing the resources required in an OFDMA based system supporting stochastic loads. Popovski *et al.* [8] have covered various aspects of URLLC traffic like the overhead due to packet headers, decoding failure probability of URLLC transmissions, and Channel State Information (CSI) acquisition at the transmitter. Ji *et al.* [9] discuss QoS requirements for URLLC traffic. They also specify various methods to share resources among URLLC and other traffic types. Ashraf *et al.* [10] study the effect of physical layer waveforms, OFDMA numerology, and Forward Error Correction (FEC) schemes on URLLC capacity via simulation. They further propose the use of Tail Biting Convolution Codes (TBCC) to achieve a target reliability as high as 10^{-9}.

The work in [11] is most closely related to ours. The authors have used a queue based model and simulations to study the design of wireless systems supporting URLLC traffic. In particular they introduce simple $M/M/m/k$ and $M/D/m/m$ queuing models to study trade-offs among system capacity, latency and reliability requirements for the worst case scenario where all users are at the cell edge. In particular, they have considered trade-offs among system capacity, reliability, and latency requirements. However, in the analysis of system trade-offs, they have only considered packet loss due to 'blocking' at the BS, i.e., unavailability of resources to immediately transmit a packet, and have not explicitly considered the effect of decoding failures and re-transmissions on system capacity. Also, they do not consider the optimization of re-transmission schemes. Our work is inspired by this initial work's approach.

The above mentioned work [11] also focused on multiplexing of enhanced Mobile Broadband (eMBB) and URLLC traffic. They showed that allocating dedicated frequency bands to URLLC and eMBB traffic is inefficient, and have advocated a shared wide-band resource allocation for both URLLC and eMBB traffic. In addition to [11], there are a few other works [13], [14] which address multiplexing URLLC and eMBB traffic via preemptive puncturing/superposition of eMBB traffic.

The line of work [15], [16] on HARQ design and optimization for URLLC traffic is closely related to our work. However, the key difference is that they focus only on the mean resource utilization of various HARQ schemes, whereas we focus on both the *mean* and *variance* of resource utilization of HARQ schemes in an OFDMA based system. This leads to solutions which are different from the ones obtained by minimizing just the mean resource utilization.

Another work which is related to ours is [17]. In [17], the authors have considered the design of random access strategies for uplink delay sensitive communications. In particular they have optimized the number of frequency bins and HARQ stages under various SINR regimes for chase HARQ combining. Our focus in this paper is on scheduled downlink communications which is different from their system model.

Many works focus on the industrial applications of URLLC traffic and exhibit simulation based studies for such systems, see [1]–[3]. Some authors, see e.g., [18], [19] focus exclusively on physical layer aspects like modulation and coding, fading and link budget analysis. However, the above mentioned works do not holistically address the design of wireless systems supporting URLLC traffic.

B. Our Contributions

In this paper we shall consider a simple Poisson model for URLLC packet arrivals. In line with the previous works, we shall also assume a wide-band allocation of resources to URLLC traffic by considering systems where such traffic can preemptively puncture/superpose URLLC packets upon previously scheduled eMBB traffic when necessary. We thus assume URLLC packet transmissions are scheduled immediately upon arrival. Such a model is not unreasonable due the stringent latency requirements of URLLC traffic. Based on this model the paper makes the following key contributions.

1) *Resource allocation in OFDMA systems:* We initially consider a *one-shot transmission* model (re-transmissions not permitted) we show that extending URLLC transmissions in time (while reducing the corresponding bandwidth usage) subject to deadline constraints increases the URLLC load that can be supported, i.e., 'wide' transmissions are better than 'tall' transmissions.
2) *Impact of system parameters:* Using an extension of the classical square-root staffing rule, we characterize the minimum overall system bandwidth W needed to support a given URLLC load. Further using the channel capacity results of [20] in the finite blocklength regime we study the scaling of URLLC capacity as a function of W, SINR, d and δ.
3) *Modeling re-transmissions/HARQ:* We extend the one-shot transmission model to incorporate HARQ schemes which allow re-transmissions if needed. The entire downlink system, the BS and associated users are modeled as a queuing network. In this setting we derive closed form expressions for various important parameters of the system such as average packet delay, distribution of the number of packets in the system, average bandwidth utilization, etc. Our framework allows us to explore the effect of a given HARQ scheme on the URLLC capacity.
4) *Optimization of a Repetition Coding based scheme:* Finally we consider the optimization of a repetition

coding based scheme minimize the necessary bandwidth required to support a given URLLC load. This can be viewed as the dual of URLLC capacity maximization problem. We identify two relevant regimes of operation, namely, *variance dominated* (for overall low URLLC loads) and *mean utilization dominated* (for high overall URLLC loads) regimes, and reach the following two conclusions:

a) In the variance dominated regime, a one-shot transmissions are optimal.

b) In the mean utilization dominated regime, the optimal transmission scheme leverages re-transmissions, if necessary, to meet the desired reliability requirement. Further the maximum number of re-transmissions permitted in the optimal scheme is a non-increasing function of the SINR.

C. Organization

The paper is organized as follows. In Sec. II we introduce our one-shot transmission model and develop the key associated results. In Sec. III we extend the one-shot transmission model to incorporate HARQ schemes. In Sec. IV we discuss the optimization of repitition coding based scheme to maximize URLLC capacity. This is followed by conclusions in Sec. V.

II. Performance Analysis: One-Shot Transmission

In this paper we focus on downlink transmissions in a wireless system with a single Base Station serving a dynamic population of URLLC users and their associated packets. The wireless system is OFDMA based where different parts of the time-frequency plane are allocated to URLLC users' packets based on transmission requests. A URLLC packet may suffer from queuing delays at the BS, transmission and propagation delays, and receiver processing delays. The system should be engineered such that the QoS requirements of URLLC traffic are satisfied, i.e., a URLLC packet of size L bits must be delivered successfully to the receiver within a total delay of d seconds with a success probability of at least $1-\delta$. We start by introducing our system model.

A. System Model– One Shot Transmission

We consider a system operating in a large aggregate bandwidth of say W Hz.[1] For simplicity we ignore the slotted nature of the system. To model the 'near far' effects in wireless systems, we shall consider a multi-class system with C classes of users where each class represents users with same SINR.[2] The aggregate traffic generated at the BS by class c users is modeled as a Poisson process with rate λ_c packets/sec. A Poisson has the virtue of simplicity and tractability. In practice a stream of URLLC packets corresponding to a control application would most likely be isochronous, i.e., regularly spaced packets while other applications might be more sporadic. A superposition of deterministically spaced and sporadic streams of packets, where each individual stream contributes a small fraction of the overall traffic might be relatively well modelled as a Poisson stream. Define the vector of arrival rates $\boldsymbol{\lambda} := (\lambda_1, \lambda_2, \ldots, \lambda_C)$. Let $SINR_c$ denote the SINR of a class c user's packets.

We initially assume that each URLLC packet is transmitted once. We will call this the *one-shot* transmission model. We will extend this to include re-transmissions in Sec. III. A packet destined to a class c user requires r_c channel uses in the time-frequency plane to transmit its codeword. The codeword for a transmission is chosen such that the decoding is successful with probability of at least $1-\delta$. A URLLC packet of class c is allocated a bandwidth of h_c for a period of time s_c. These values are fixed and related to r_c by $\kappa s_c h_c = r_c$, where κ is a constant which denotes the number of channel uses per unit time per unit bandwidth of the OFDMA time-frequency plane. The value of κ depends on the OFDMA frame structure and numerology. Since URLLC packets have a deadline of d seconds, we shall always choose $s_c \leq d$. For ease of analysis we shall also assume that for any class c, d is an integer multiple of s_c. Thus following vectors which characterize the system: $\mathbf{r} := (r_1, r_2, \ldots, r_C)$, $\mathbf{s} := (s_1, s_2, \ldots, s_C)$, $\mathbf{h} := (h_1, h_2, \ldots, h_C)$ and $\boldsymbol{\rho} := (\rho_1, \rho_2, \ldots, \rho_C)$, where $\rho_c := \lambda_c s_c$.

We shall make the following key assumption on the system operation.

Assumption 1: (**Immediate scheduling**) A URLLC packet transmission request is scheduled immediately upon arrival if there is spare bandwidth is available. Otherwise the packet is lost. New packets do not preempt ongoing URLLC packet transmissions.

Given the stringent latency requirements, the immediate scheduling assumption is a reasonable design choice.

B. Infinite System Bandwidth

Initially let us consider a system with infinite bandwidth, i.e., $W = \infty$. In such a system the base station can be modeled as a multi-class $M/GI/\infty$, see [21] for more details. Let $\mathbf{N} := (N_1, N_2, \ldots, N_C)$ be a random vector denoting the number of active transmissions when the system is in steady state. For any $\mathbf{n} \in \mathbb{Z}_+^C$, let $\pi(\mathbf{n}) := P(\mathbf{N} = \mathbf{n})$ be the stationary distribution. Using standard results for $M/GI/\infty$ queues (see [22]) one immediately gets the following results:

$$\pi(\mathbf{n}) = \Pi_{c=1}^{C} \left(\frac{\rho_c^{n_c}}{n_c!} \right) \exp(-\rho_c), \tag{1}$$

and the average bandwidth utilization is given by

$$\mathbb{E}\left[\mathbf{h}\mathbf{N}^T\right] = \mathbf{h}\boldsymbol{\rho}^T.$$

Observe that the number of active transmissions of any class c is Poisson distributed with mean ρ_c. Thus ρ_c as the average load of class c traffic.

C. Effect of Finite System Bandwidth

Although in practice the available system bandwidth W is not infinite but possibly large. We will consider a case

[1] This need not be a contiguous bandwidth, but result from the use of carrier aggregation across disjoint segments

[2] Ideally SINR is a continuous random variable, however, in practical systems the channel quality feedback from users are quantized to several discrete levels.

where a wide-band allocation W is available to transmit URLLC traffic. This might be made available through a puncturing/superposition scheme between URLLC and eMBB traffic. see e.g., [13]. Even large bandwidth systems can occasionally suffer from congestion due to the stochastic variations in the arrival process and occasionally there may not be enough spare bandwidth to transmit a new URLLC packet. In such cases we shall assume that packets are *blocked* and dropped from the system. Let $\mathbf{N}(t) := (N_1(t), N_2(t), \ldots, N_C(t))$ be a random vector denoting the number of packets of each class in the system at time t. A class c packet arriving at time t is blocked if the following condition holds:

$$h_c + \sum_{c'=1}^{C} h_{c'} N_{c'}(t) > W. \tag{2}$$

We address the following two questions in this section:

1) How do the choices of $\mathbf{h}$ and $\mathbf{s}$ affect the blocking of URLLC packets?
2) What is the required system bandwidth W given a desired packet reliability δ?

To study the effect of $\mathbf{h}$ and $\mathbf{s}$ on the blocking of URLLC traffic, we shall first consider the blocking probability of a typical class c packet. Observe that the blocking probability experienced by packets of a class depends on $\mathbf{h}, \mathbf{s}$ (of all classes), $\boldsymbol{\lambda}$ and W. Let $p_{b,c}(\mathbf{h}, \mathbf{s}, \boldsymbol{\lambda}, W)$ be the blocking probability experienced by a typical class c packet arrival. The fraction of class c traffic admitted is then given by $\lambda_c (1 - p_{b,c}(\mathbf{h}, \mathbf{s}, \boldsymbol{\lambda}, W))$. Hence, lowering the blocking probability increases the admitted URLLC traffic. The following result which is proved in Appendix-A gives us the key insight on optimal choices of $\mathbf{h}$ and $\mathbf{s}$ for URLLC packet transmissions.

Theorem 1: For a given $\mathbf{h}$ and $\mathbf{s}$, positive integer q, and $i \in \{1, 2, \ldots, C\}$ define $\mathbf{h}' := (h_1, h_2, \ldots, h_i/q, \ldots, h_C)$ and $\mathbf{s}' := (s_1, s_2 \ldots, qs_i, \ldots, s_C)$. Under the one-shot transmission model and Assumption 1, if $\rho_i < 1$, then for any $c \in \{1, 2, \ldots, C\}$, there exists $\tilde{W}_c$ such that for $W > \tilde{W}_c$ we have that $p_{b,c}(\mathbf{h}, \mathbf{s}, \boldsymbol{\lambda}, W) \geq p_{b,c}(\mathbf{h}', \mathbf{s}', \boldsymbol{\lambda}, W)$.

Remarks: Observe that in wide-band systems scaling h_i and s_i by an integer q as required in the above theorem increases the number of concurrent transmissions of class i and is also beneficial for all classes (including class i). To understand this let us look at the mean and variance of the bandwidth utilization of class i in a system with parameters $\mathbf{h}'$ and $\mathbf{s}'$ and infinite bandwidth. The average bandwidth utilization of class i, given by $h_i \lambda_i s_i$, does not change with scaling factor q, as the decrease in bandwidth of class i is compensated by corresponding increase in the average number of users of class i. However, the variance of the bandwidth utilization, given by $\frac{1}{q} h_i^2 \lambda_i s_i$ decreases with q. Therefore, the congestion events occur less frequently and the system admits more traffic. Note that this observation is in line with the previous work on URLLC traffic (see [7]) where the emphasis is on such events corresponding to the 'tail' of URLLC traffic demand. Further, the assumption $\rho_i < 1$ is not restrictive as one can divide a class into various 'virtual' sub-classes such that the average load in each sub-class is less than unity.

Therefore, one should scale s_i with an integer q such that $qs_i = d$. Such an integer q exists because of our assumption that d is an integer multiple of s_i. Hence, this motivates the following optimal choices of s_i and h_i:

$$s_i = d \quad \text{and } h_i = \frac{r_i}{\kappa d}. \tag{3}$$

To summarize, one might think that 'tall' transmissions are better as they take less time, however, according to the above result it is better to decrease the bandwidth per transmission and spread out the transmissions as 'wide' as possible in the time axis, i.e., increase s_i (and decrease h_i) as long as the deadline is not violated.

To meet the reliability requirements of URLLC traffic, the system bandwidth W must be chosen such that the probability of blocking of a typical URLLC packet arrival is of the order of δ. To that end we shall use a multi-class extension of the classical square-root staffing rule (see [22] for more details) to relate W, $\mathbf{r}$, $\boldsymbol{\lambda}$ and δ. Under this dimensioning rule, to support a URLLC load of $\boldsymbol{\lambda}$ with reliability δ for a given $\mathbf{r}$, the system bandwidth should satisfy the following condition:

$$W \geq \zeta^{\text{mean}}(\mathbf{r}) + c(\delta) \sqrt{\zeta^{\text{variance}}(\mathbf{r})}, \tag{4}$$

where $c(\delta) = Q^{-1}(\delta)$, $Q(\cdot)$ is the Q-function, $\zeta^{\text{mean}}(\mathbf{r}) := \sum_{c=1}^{C} \lambda_c \frac{r_c}{\kappa}$ is the mean bandwidth utilization, and $\zeta^{\text{variance}}(\mathbf{r}) := \sum_{c=1}^{C} \lambda_c \frac{r_c^2}{\kappa^2 d}$ is the variance of the bandwidth utilization.

Next we study the URLLC capacity scaling with respect to W, $SINR_c$, d, and δ. This requires a model relating r_c, $SINR_c$, and δ which is described in the next subsection.

D. Finite Block Length Model

Since the URLLC packet sizes are typically small, we shall use the capacity results for the finite blocklength regime developed in [20]. In an AWGN channel the number of information bits L that can be transmitted with a codeword decoding error probability of p in r channel uses is given by

$$L = rC(SINR_c) - Q^{-1}(p)\sqrt{rV(SINR_c)} + O(\log_2(r)), \tag{5}$$

where $C(SINR_c) = \log_2(1 + SINR_c)$ is the AWGN channel capacity under infinite blocklength assumption and $V(SINR_c) = (\log_2(e))^2 \left(1 - \frac{1}{(1+SINR_c)^2}\right)$. Using the above model one can approximate r as a function of p as follows:

$$r \approx \frac{L}{C(SINR_c)} + \frac{\left(Q^{-1}(p)\right)^2 V(SINR_c)}{2\left(C(SINR_c)\right)^2} + \frac{\left(Q^{-1}(p)\right)^2 V(SINR_c)}{2\left(C(SINR_c)\right)^2} \sqrt{1 + \frac{4LC(SINR_c)}{V(SINR_c)\left(Q^{-1}(p)\right)^2}}. \tag{6}$$

A derivation of this approximation is given in Appendix-B. We can now write r_c as a function of δ, L and $SINR_c$ for various user/packet classes.

E. Capacity Scaling

We shall define the *single class URLLC capacity* as follows.

Definition 1: For any class c, its single class URLLC capacity λ_c^* is the maximum URLLC arrival rate that can be supported by the system while satisfying the QoS requirements if only class c traffic is present in the system.

Note that λ_c^* is a function of W, d. δ, $SINR_c$, and L. We would like to study the scaling of λ_c^* with respect to various system parameters. Recall that for $f, g : \mathbb{R}_+ \to \mathbb{R}_+$, we say that $f(x) \sim \Theta(g(x))$ if there exist x_o, a, and b such that $a \leq b$ and for $x \geq x_o$ we have that

$$ag(x) \leq f(x) \leq bg(x). \tag{7}$$

The following result summarizes the scaling of λ_c^* with various system parameters. The proof of the theorem below is given in Appendix-C.

Theorem 2: Under one-shot transmission model and Assumption 1 we have that

1) $\lambda_c^* \sim \Theta\left(W - \sqrt{W}\right)$.
2) For $SINR_c \gg 1$, we have that $\lambda_c^* \sim \Theta\left(\log_2(SINR_c) - \sqrt{\log_2(SINR_c)}\right)$.
3) $\lambda_c^* \sim \Theta\left(1 - \frac{1}{\sqrt{d}}\right)$.
4) $\lambda_c^* \sim \Theta\left(\frac{1}{-\log_2(\delta)}\right)$.

Remarks: Observe that λ_c^* scales as a strictly concave function of $SINR_c$, d, and δ. Hence, while increasing $SINR_c$ and d or decreasing δ one suffers from diminishing returns. However, as expected the scaling of λ_c with respect to W does not suffer from diminishing returns. For large W, λ_c^* increases linearly with W which is the best one could hope.

III. Performance Analysis With Multiple Transmissions

Next we extend the system model to include re-transmissions and study the role of HARQ schemes. We shall first explain the extension of our system model.

A. System Model– Multiple Transmissions

Paralleling the one-shot transmission model considered in Sec. II, we shall consider a multi-class system with Poisson arrivals for URLLC traffic, where a class represents users' packets sharing the same SINR. However, by contrast to our one-shot transmission model, in this section we shall permit packet re-transmissions. Suppose a class c packet can have up to m_c transmission attempts after which it is dropped. We index transmission attempts by $m = 1, 2, \ldots, m_c$, where $m = 1$ corresponds to the *initial* transmission and any $m > 1$ corresponds to a *re-transmission*. A class c packet in the m^{th} transmission attempt is assumed to require $r_{c,m}$ channel uses in the time-frequency plane. The bandwidth used and the time to transmit in the m^{th} packet transmission are denoted by $h_{c,m}$ and $s_{c,m}$, respectively. They are related to $r_{c,m}$ by $\kappa h_{c,m} s_{c,m} = r_{c,m}$. For any $m \in \{1, 2, \ldots, m_c\}$, define $\mathbf{r}_c^{(m)} := (r_{c,1}, r_{c,2}, \ldots, r_{c,m})$. After every transmission the intended receiver sends a one bit feedback to the BS indicating success/failure of the packet decoding process.

TABLE I
Summary of Important Vectors Characterizing the System

Notation	Description of the vectors
$\mathbf{h}_c$	Bandwidths for class c packets
$\mathbf{s}_c$	Transmission times for class c packets
$\mathbf{r}_c^{(m_c)}$	Channel uses for class c packets
$\mathbf{m}$	Maximum transmission attempts for all classes
$\boldsymbol{\lambda}$	Arrival rates for all classes
$\boldsymbol{\rho}$	System load for all classes

In general, the probability of decoding failure of a class c packet after the m^{th} transmission attempt, denoted by $p_{c,m}\left(\mathbf{r}_c^{(m)}\right)$, is a function of $\mathbf{r}_c^{(m)}$. A decoding failure for a class c packet occurs if the packet has not been successfully decoded after m_c transmission attempts. This happens with probability $\Pi_{m=1}^{m_c} p_{c,m}\left(\mathbf{r}_c^{(m)}\right)$. Thus one should design the system such that $\Pi_{m=1}^{m_c} p_{c,m}\left(\mathbf{r}_c^{(m)}\right) \leq \delta$. Therefore, the values of $r_{c,m}$, $p_{c,m}\left(\mathbf{r}_c^{(m)}\right)$ and m_c jointly characterize the HARQ scheme used for class c users.

The feedback on success/failure of a transmission will incur propagation delays, receiver processing delay, and the uplink channel access and scheduling delays. We shall assume that the uplink channel is well provisioned so that there are no scheduling and channel access delays. Therefore, the total feedback delay includes only the propagation delay and the receiver processing delay which we shall denote by a deterministic value f_c for a class c user. A class dependent feedback delay is consistent with our notion that classes denote users with similar channel characteristics, for example, users at the cell edge may experience longer feedback delays.

For any class c, define the following vectors $\mathbf{s}_c := (s_{c,1}, s_{c,2} \ldots, s_{c,m_c})$, $\mathbf{h}_c := (h_{c,1}, h_{c,2} \ldots, h_{c,m_c})$, and $\boldsymbol{\rho}_c := (\rho_{c,1}, \rho_{c,2} \ldots, \rho_{c,m_c})$, where $\rho_{c,1} := \lambda_c s_{c,1}$ and for any $m > 1$ let $\rho_{c,m} := \lambda_c \left(\Pi_{k=1}^{m-1} p_{c,k}\left(\mathbf{r}_c^{(k)}\right)\right) s_{c,m}$. Using the above definitions, we further define the following vectors capturing the overall system's design and loads: $\mathbf{s} := (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_C)$, $\mathbf{h} := (\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_C)$, and $\boldsymbol{\rho} := (\boldsymbol{\rho}_1, \boldsymbol{\rho}_2, \ldots, \boldsymbol{\rho}_C)$. We further let $\mathbf{m} := (m_1, m_2, \ldots, m_C)$ denote vector of maximum transmission attempts per class. The important vectors are summarized in the Table I.

We shall also revise the *immediate scheduling* assumption for the setting with packet re-transmissions.

Assumption 2: (**Immediate scheduling**) An initial URLLC packet transmission request or a re-transmission is admitted and scheduled for transmission immediately if there is spare bandwidth available. Otherwise the packet is lost.

B. Infinite System Bandwidth

Once again consider a system with infinite system bandwidth so that there is no blocking of packets. In the multiple transmission model the BS has to wait for feedback from the intended receiver before re-transmitting a packet. We model this system with feedback using a network of two multi-class $M/GI/\infty$ queues, one modeling BS transmissions and the

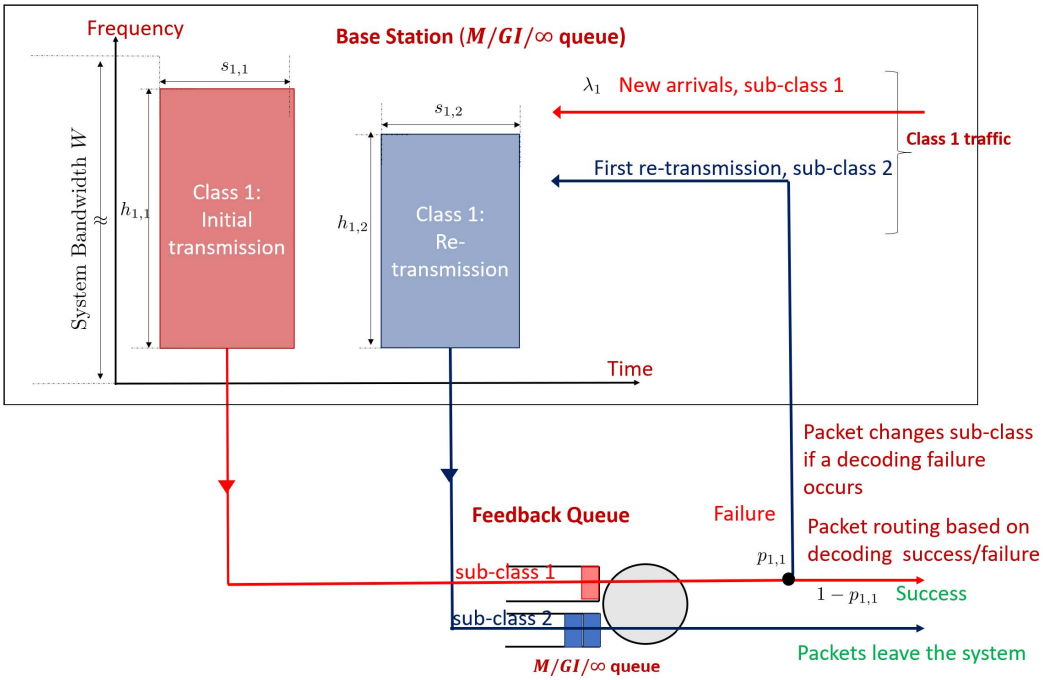

Fig. 1. A wireless system with a single class of URLLC users modeled as a network of two $M/GI/\infty$ queues. Up to two transmissions attempts are allowed for all packets, i.e., $m_1 = 2$. Packets of sub-classes one and two are shown by red and blue colors, respectively. Observe that a packet will change its sub-class after a decoding failure.

other modeling the packets awaiting feedback, which we refer to as the *feedback* queue. This is described below.

Base Station queue: Similar to Sec. II, the BS is modeled as a multi-class $M/GI/\infty$ queue where each class corresponds to a set of users with the same SINR. However, unlike the one-shot transmission model, we further divide each class into various *sub-classes* to keep track of the number of re-transmissions. In particular each class c is further divided into m_c sub-classes with the sub-classes indexed by various possible stages of packet (re)transmission. A class c packet which is being transmitted for the m^{th} time belongs to m^{th} sub-class and it will require a bandwidth $h_{c,m}$ for a period of time $s_{c,m}$ to complete transmission. Further, because of our assumption of infinite bandwidth, the BS can transmit any number of packets from any of its classes concurrently, i.e., the number of servers in the queuing model is ∞.

Feedback queue: We model the packet decoding and feedback sending processes as a multi-class $M/GI/\infty$ queue which uses the same notion of a class and sub-class in the feedback queue as in the BS queue. For a class c packet, the feedback associated with the decoding of a class c packet is received at the BS after f_c seconds. Based on the success/failure of the decoding process the BS then decides to re-transmit it or not. We abstract this process as follows. A class c packet after its m^{th} transmission is routed from the BS queue to the feedback queue where it spends f_c seconds. Note that the packet retains its class and sub-class indices in the feedback queue. After f_c seconds in the feedback queue it is then routed to the sub-class $m+1$ of class c with probability $p_{c,m}\left(\mathbf{r}_c^{(m)}\right)$ (decoding failure) or leaves the system with probability $1-p_{c,m}\left(\mathbf{r}_c^{(m)}\right)$ (successful decoding). If a class c packet in m^{th} sub-class is routed to the BS, then it changes its sub-class index to $m+1$, i.e., it is being transmitted for $(m+1)^{\text{th}}$ time. This process repeats until the packet is successfully decoded, or m_c transmission attempts have been made, whichever happens first. Thus a class c packet always leaves the system after m_c transmissions irrespective of the outcome of the decoding process of the m_c^{th} transmission. The queuing network consisting of a BS and a single class of URLLC traffic is illustrated in the Fig. 1.

Observe that it is assumed that any number of URLLC packets can be processed in parallel in the feedback queue, and hence it can also be modeled as an $M/GI/\infty$ queue. This is a reasonable assumption because the packet decoding process across users are independent of each other and done in parallel and we assume sufficient uplink bandwidth is provisioned for feedback from various users.

The queuing model described previously can be used to study various important properties of the multi-class system which are given below. Let $\mathbf{N}$ now be a random vector denoting the number of packets in different stages of re-transmissions of all classes in the steady state, i.e., $\mathbf{N} := (N_{1,1}, N_{1,2} \ldots, N_{1,m_1}, \ldots, N_{C,1}, N_{C,2} \ldots, N_{C,m_C})$. Once again, from classical queuing theory results (see [22]), it follows that the steady state probability $\pi(\mathbf{n}) = P(\mathbf{N} = \mathbf{n})$ is given by:

$$\pi(\mathbf{n}) = \Pi_{c=1}^{C} \Pi_{m=1}^{m_c} \left(\frac{\rho_{c,m}^{n_{c,m}}}{n_{c,m}!}\right) \exp(-\rho_{c,m}), \tag{8}$$

where $\rho_{c,m}$ is the average system load of class c packets in sub-class m. The average bandwidth utilized is given by $\mathbb{E}\left[\mathbf{h}\mathbf{N}^T\right] = \mathbf{h}\boldsymbol{\rho}^T$.

C. Effect of Finite System Bandwidth

Paralleling the one-shot transmission model studied earlier, a finite bandwidth system may suffer from congestion due to stochastic variations in loads and may have to block an immediate packet transmission request (a new packet or a re-transmission). Hence, we must choose W appropriately to meet the reliability requirements. A natural extension to the result in (4) for the bandwidth requirement of the multi-class system is as follows. Given a target blocking probability of δ, W is chosen such that

$$W \geq \eta^{\text{mean}}(\mathbf{r},\mathbf{m}) + c(\delta)\sqrt{\eta^{\text{variance}}(\mathbf{r},\mathbf{m},\mathbf{h},\mathbf{s})}, \tag{9}$$

where

$$\eta^{\text{mean}}(\mathbf{r},\mathbf{m}) := \frac{1}{\kappa}\sum_{c=1}^{C}\lambda_c\left(r_{c,1}+\sum_{m=2}^{m_c}\left(\Pi_{k=1}^{m-1}p_{c,k}\left(\mathbf{r}_c^{(k)}\right)\right)r_{c,m}\right) \tag{10}$$

and

$$\eta^{\text{variance}}(\mathbf{r},\mathbf{m},\mathbf{h},\mathbf{s}) := \frac{1}{\kappa^2}\sum_{c=1}^{C}\lambda_c\left(\frac{r_{c,1}^2}{s_{c,1}}+\sum_{m=2}^{m_c}\left(\Pi_{k=1}^{m-1}p_{c,k}\left(\mathbf{r}_c^{(k)}\right)\right)\frac{r_{c,m}^2}{s_{c,m}}\right). \tag{11}$$

Note that we have used the fact that $\kappa h_{c,m}s_{c,m} = r_{c,m}$ in writing the above equations.

The above characteristics follow by applying the square-root staffing rule to the multi-class system. The first term $\eta^{\text{mean}}(\mathbf{r},\mathbf{m})$ represents the mean bandwidth utilization. The term $\eta^{\text{variance}}(\mathbf{r},\mathbf{m},\mathbf{h},\mathbf{s})$ represents the variance of the bandwidth utilization. Observe that while $\eta^{\text{mean}}(\mathbf{r},\mathbf{m})$ only depends on $\mathbf{r}$ and $\mathbf{m}$, each term in η^{variance} is multiplied with $1/s_{c,m}$ and thus is affected by the choice of $\mathbf{s}$.

For $m_c = 1$, we have shown in Thm. 2 that it is advantageous in terms of blocking probability to increase $s_{c,m}$ (or decrease $h_{c,m}$) subject to the deadline constraint. This is not easily extendable to the case for $m_c > 1$. However, this result gives us a key insight on the choice of $s_{c,m}$. A natural extension of this insight to higher values of m_c is to increase the transmission times of all stages such that the cumulative transmission time of m_c stages and feedback delays add up to d, i.e.,

$$\sum_{m=1}^{m_c} s_{c,m} + m_c f_c = d. \tag{12}$$

Based on the previous discussion let us discuss the various steps one might follow to properly dimension this multi-class system appropriately to support URLLC traffic.

1) Choose $\mathbf{r}$ and $\mathbf{m}$ such that probability of decoding failure is less than or equal to δ.
2) Choose $\mathbf{s}$ such that the condition (12) is satisfied. This also determines $\mathbf{h}$ as $\mathbf{r}$ is chosen in the first step and $\kappa h_{c,m}s_{c,m} = r_{c,m}$.
3) To support an arrival rate vector $\boldsymbol{\lambda}$, determine the minimum necessary bandwidth via (9).

Although (9) and (12) provide the basic insight into the effect of re-transmissions on the URLLC capacity, however there are still many possible solutions which satisfy (9) and (12). One has to find the optimal values for $\mathbf{r}$, $\mathbf{h}$, $\mathbf{s}$, and $\mathbf{m}$ to maximize the URLLC capacity. This is discussed in the next section.

IV. URLLC Capacity Maximization/ Required Bandwidth Minimization

There are two ways to formulate the problem of optimizing HARQ schemes to maximize URLLC capacity. One can characterize the set of URLLC arrival rates which can be supported for a given system bandwidth W subject to the QoS constraints. This will define a *multi-class URLLC capacity region*. Alternatively, one can formulate the problem in terms of minimizing the bandwidth required to support a given vector $\boldsymbol{\lambda}$ of URLLC arrival rates subject to the QoS constraints. This second approach is somewhat simpler yet still allows one to study the most efficient system design for the HARQ schemes. One can then study the structural properties of the solution obtained. We shall follow this second approach in the rest of this paper. The associated optimization problem is as follows:

$$\mathcal{OP}_2: \min_{\mathbf{m},\mathbf{r},\mathbf{h},\mathbf{s}} \ \eta^{\text{mean}}(\mathbf{r},\mathbf{m}) + c(\delta)\sqrt{\eta^{\text{variance}}(\mathbf{r},\mathbf{m},\mathbf{h},\mathbf{s})} \tag{13}$$

$$\text{s.t. } \kappa h_{c,m}s_{c,m} = r_{c,m}, \quad \sum_{m=1}^{m_c} s_{c,m} + m_c f_c \leq d, \tag{14}$$

$$\text{and } h_c \leq W,\ \Pi_{m=1}^{m_c} p_{c,m}\left(\mathbf{r}_c^{(m)}\right) \leq \delta, \quad \forall c. \tag{15}$$

The above problem is a non-convex, mixed integer programming problem, and in general is analytically intractable. To get some insights on this problem we will consider a specific scheme, namely, *repetition coding with homogeneous transmissions*. The performance under this scheme provides an upper bound on the minimum bandwidth required under commonly used Chase combining, see [23].

A. Repetition Coding– Homogeneous Transmissions

In repetition coding, the same codeword is transmitted repeatedly to the receiver until the packet is successfully decoded or the maximum number of re-transmissions has been reached. We shall also further assume that the transmissions are homogeneous. This is stated formally below.

Assumption 3: (**Homogeneous transmissions**) For all c and m, we have that $r_{c,m} = r_c$, $h_{c,m} = h_c$ and $s_{c,m} = s_c$.

We also make the following assumption on the packet decoding process at the receiver.

Assumption 4: (**Independent decoding**) The receiver decodes each transmission independently of the previous transmissions, and hence, the probability of failure in any transmission attempt depends only on the codeword used in that stage.

Under the above assumptions, the decoding failure probability is independent across re-transmissions and driven by r_c,

i.e., for any c and m we have that $p_{c,m}\left(\mathbf{r}_c^{(m)}\right) = p_c(r_c)$. Assuming independence between the decoding processes simplifies the analysis further. Also, due to the stringent latency requirements, complex HARQ schemes may not be practically feasible at the receiver. Independent decoding assumption provides an upper bound on the minimum system bandwidth required under Chase combining where the re-transmissions of a packet are combined coherently at the receiver and decoded. Utilizing homogeneous transmissions reduces the overhead in control signals to indicate the allocation of bandwidth to users.

Unfortunately, under finite block length model and repetition coding, $\mathcal{OP}_2$ is still analytically intractable in a multi-class system. Therefore, we shall consider two regimes, *the variance* and *the mean utilization dominated regimes* where the solutions simplify considerably. They are formally described next.

Definition 2: 1) *Variance dominated regime:* In this regime, the objective function of $\mathcal{OP}_2$ includes only the variance of the bandwidth utilization (η^{variance}).

2) *Mean utilization dominated regime:* In this regime, the objective function $\mathcal{OP}_2$ includes only the mean of the bandwidth utilization (η^{mean}).

Note that at low loads when λ_c's are small, in (13) the term corresponding to the overall variance is dominant, therefore, at low loads we shall minimize the variance of the total bandwidth usage. At high loads, the variance of bandwidth usage (second term) is smaller than the mean (first term). Hence, we shall focus on minimizing the mean utilization at high loads. We shall also use the finite blocklength model discussed in Sec. II-D to relate $p_c(r_c)$ and r_c. Under these simplifications, one can de-couple $\mathcal{OP}_2$ for each class and optimize the HARQ schemes separately for each class. The main result in the variance dominated regime is given below.

Proposition 1: For the multiple transmissions model in Sec. III, under Assumptions 2, 3, and 4, and in the variance dominated regime, the optimization problem $\mathcal{OP}_2$ decomposes across classes. The optimization problem for class c is as follows:

$$\min_{m_c,r_c,h_c,s_c} : \sum_{m=0}^{m_c} \left(\frac{r_c^2}{s_c}\right)(p_c(r_c))^m \tag{16}$$

$$\text{s.t. } \kappa s_c h_c = r_c, \quad h_c \le W, \; m_c\,(s_c + f_c) = d, \tag{17}$$

$$(p_c(r_c))^{m_c} \le \delta. \tag{18}$$

Furthermore, under the finite block length model (6) relating $p_c(r_c)$ and r_c, for $L \le 2000$ bits, $d \le 2$ msec., $\delta \in \left[10^{-3}, 10^{-6}\right]$, $SINR_c \in [0, 20]\,dB$, $f_c \ge 0.1$ msec. the optimal solution has the following structure:

1) One shot transmission is optimal, i.e., $m_c^* = 1$.
2) The optimal values of transmission time and bandwidth, denoted by s_c^* and h_c^*, respectively, satisfy

$$s_c^* = d - f_c \quad \text{and } h_c^* = \frac{r_c^*}{d - f_c}, \tag{19}$$

where r_c^* is the smallest r such that $p_c(r) \le \delta$.

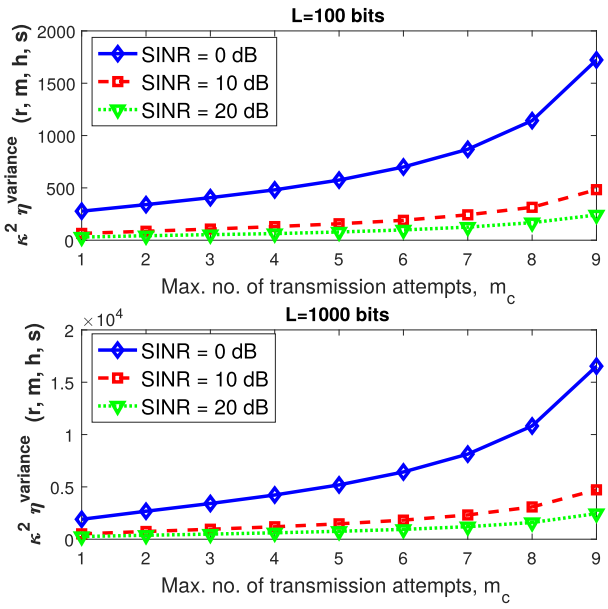

Fig. 2. Variance of bandwidth utilization (scaled) as a function of m_c for various values of $SINR_c$ and L with $\lambda_c = 1$ packet/msec., $\delta = 10^{-6}$, $d = 1$ msec., and $f_c = 0.125$ msec.

The equivalent result in the mean utilization dominated regime is given below.

Proposition 2: For the multiple transmission model in Sec. III, under Assumptions 2, 3, and 4, and in the mean utilization dominated regime, the optimization problem $\mathcal{OP}_2$ decomposes across classes. The optimization for class c is given by:

$$\min_{m_c,r_c,h_c,s_c} : \sum_{m=0}^{m_c} r_c(p_c(r_c))^m \tag{20}$$

$$\text{s.t. } \kappa s_c h_c = r_c, \quad h_c \le W, \; m_c\,(s_c + f_c) = d, \tag{21}$$

$$(p_c(r_c))^{m_c} \le \delta. \tag{22}$$

Furthermore, under the finite block length model (6) relating r_c and $p_c(r_c)$, for $L \le 2000$ bits, $d \le 2$ msec., $\delta \in \left[10^{-3}, 10^{-6}\right]$, $SINR_c \in [0, 20]\,dB$, $f_c \in [0.1, 0.25]$ msec. the optimal solution has the following structure:

1) The optimal value m_c^* is strictly more than one.
2) The optimal value m_c^* is a non-increasing function of $SINR_c$.

The decomposed optimization problems for class c in both the regimes are obtained from the definitions of $\eta^{\text{mean}}(\mathbf{r}, \mathbf{m})$ and $\eta^{\text{variance}}(\mathbf{r}, \mathbf{m}, \mathbf{h}, \mathbf{s})$. The results on m_c^* are obtained by direct substitution. Some remarks regarding the above two propositions are in order.

Comparing the objective functions for the mean and variance dominated regimes, note that each term in the variance dominated regime is multiplied with an extra term $\frac{r_c}{s_c}$. Since $\frac{r_c}{s_c} = \frac{r_c m_c}{(d - m_c f_c)}$, the objective function in the variance dominated regime increases sharply with m_c. Therefore, the optimal value of m_c is lower in the variance dominated regime than in the mean dominated regime. In the mean dominated regime, as one decreases $SINR_c$, the resources required

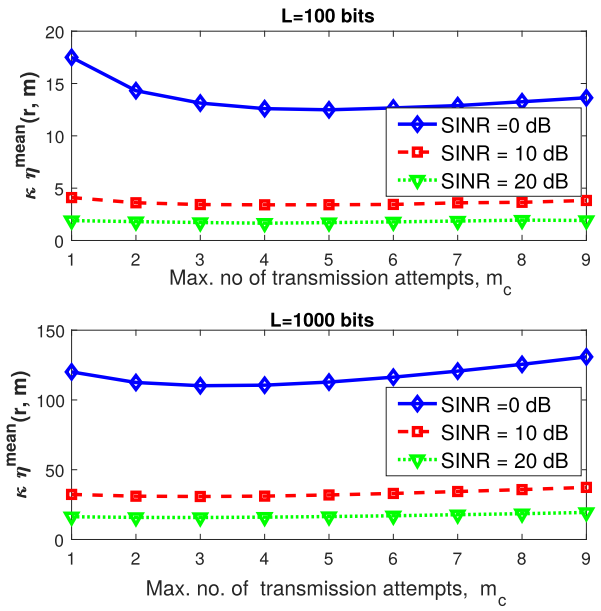

Fig. 3. Mean of bandwidth utilization (scaled) as a function of m_c for various values of $SINR_c$ and L with $\lambda_c = 100$ packet/msec., $\delta = 10^{-6}$, $d = 1$ msec., and $f_c = 0.125$ msec.

per transmission (r_c) to meet a given reliability requirement increase sharply. Hence, it is advantageous at lower SINRs to increase m_c and to choose a lower reliability target per transmission. We have plotted the variance and mean of bandwidth utilization for various packet sizes and SINR values in Figures 2 and 3, respectively.

V. Conclusions

In this paper we explored possible designs of 5G wireless systems supporting URLLC traffic. We develop a simple model for URLLC packet transmissions which captures the essential properties of such a system when pre-emptive/immediate URLLC scheduling and finite block-length transmissions are used. Based on this model we derive scaling results for the URLLC capacity (admissible load subject to QoS constraints) with respect to various system parameters such as the link SINR, system bandwidth, and the packet latency and reliability requirements. Several key findings arise which are of practical interest. First, URLLC capacity is enhanced by extending URLLC transmissions in time as much as possible (subject to latency constraints) while using the least amount of bandwidth (to meet reliability requirements). Next we consider results associated with optimizing a repetition coding based scheme. In the variance dominated regime (typically low loads with arrival rates less than 1 packet/sec.), one-shot transmissions satisfying the above mentioned requirements minimize the necessary bandwidth required to support URLLC traffic. In the mean utilization dominated regime (typically high loads with arrival rates more than 100 packets/sec.), optimal re-transmission schemes minimizing the necessary bandwidth leverage multiple re-transmissions and the maximum number of transmissions required is a non-increasing function of SINR.

Appendix

A. Proof of Theorem 1

In this section we shall explicitly show the dependence of $\pi(\mathbf{n})$ on W by denoting it as $\pi(\mathbf{n}, W)$. Using the standard results from queuing theory (see [22]), we have that:

$$\pi(\mathbf{n}, W) = G\,\Pi_{c=1}^{C}\left(\frac{\rho_c^{n_c}}{n_c!}\right), \tag{23}$$

where $G^{-1} = \sum_{\tilde{\mathbf{n}}\in\mathcal{S}}\Pi_{c=1}^{C}\left(\frac{\rho_c^{\tilde{n}_c}}{\tilde{n}_c!}\right)$ and $\mathcal{S} = \{\mathbf{n} \mid \mathbf{h}\mathbf{n}^T \le W\}$. Here $\mathcal{S}$ is the set of all user configurations such that the total bandwidth constraint is not violated. Similarly one can define $\pi'(\mathbf{n}, W)$ for the case when bandwidths and transmission times are $\mathbf{h}'$ and $\mathbf{s}'$, respectively with $q\rho_i$ replacing ρ_i in (23). Define $\mathbf{n}_{\setminus i} := (n_1, n_2, \ldots, n_{i-1}, n_{i+1}, \ldots n_C)$. With slight abuse of notation let $\pi\left(\mathbf{n}_{\setminus i}, W\right)$ and $\pi'\left(\mathbf{n}_{\setminus i}, W\right)$ bet the steady probabilities of $\mathbf{n}_{\setminus i}$ under bandwidths $\mathbf{h}$ and $\mathbf{h}'$, respectively. Based on the standard results for $M/GI/\infty$ queues, as $W \to \infty$, $\pi\left(\mathbf{n}_{\setminus i}, W\right)$ and $\pi'\left(\mathbf{n}_{\setminus i}, W\right)$ converge to a Poisson distribution, i.e., we have the following:

$$\begin{aligned}\lim_{W\to\infty}\pi\left(\mathbf{n}_{\setminus i}, W\right) &= \lim_{W\to\infty}\pi'\left(\mathbf{n}_{\setminus i}, W\right)\\ &= \exp\left(-\sum_{c:c\neq i}\rho_c\right)\Pi_{c:c\neq i}\left(\frac{\rho_c^{n_c}}{n_c!}\right).\end{aligned} \tag{24}$$

Using PASTA property (see [22]), the blocking probability experienced by a typical arrival to class 1 is given by

$$p_{b,1}\left(\mathbf{h}', \mathbf{s}', \boldsymbol{\lambda}, W\right) = \sum_{\mathbf{n}\in\mathcal{S}_1'}\pi'(\mathbf{n}, W), \tag{25}$$

where $\mathcal{S}_1' := \{\mathbf{n} \mid \mathbf{h}'\mathbf{n}^T \le W \text{ and } \mathbf{h}'(\mathbf{n}^T + \mathbf{e}_1^T) > W\}$, $\mathbf{e}_1$ is the standard unit vector with non-zero entry at position one, $(\cdot)^T$ is the transpose operator. $\mathcal{S}_1'$ is the set of blocking states for class 1. Given $\mathbf{n}_{\setminus\mathbf{i}}$, a blocking event occurs when $n_i \in \left\{\lceil\frac{q\left(W-\sum_{c:c\neq i}h_c n_c\right)}{h_i}\rceil - \lceil\frac{qh_1}{h_i}\rceil + 1, \ldots, \lceil\frac{q\left(W-\sum_{c:c\neq i}h_c n_c\right)}{h_i}\rceil\right\}$. Therefore, using the definition of $\pi'\left(\mathbf{n}, W\right)$ one can re-write (25) as follows:

$$p_{b,1}\left(\mathbf{h}', \mathbf{s}', \boldsymbol{\lambda}, W\right) = \sum_{\mathbf{n}_{\setminus i}\in\mathcal{S}_{\setminus i}}\zeta(\mathbf{n}_{\setminus i}, q, \mathbf{h}', \mathbf{s}', W)\pi'\left(\mathbf{n}_{\setminus i}, W\right), \tag{26}$$

where $\mathcal{S}_{\setminus i} := \left\{\mathbf{n}_{\setminus i} \mid \mathbf{h}'_{\setminus i}\mathbf{n}'^T_{\setminus i} \le W\right\}$ and

$$\zeta(\mathbf{n}_{\setminus i}, q, \mathbf{h}', \mathbf{s}', W) := \frac{\sum_{j=\lceil\frac{q\left(W-\sum_{c:c\neq i}h_c n_c\right)}{h_i}\rceil - \lceil\frac{qh_1}{h_i}\rceil+1}^{\lceil\frac{q\left(W-\sum_{c:c\neq i}h_c n_c\right)}{h_i}\rceil}\frac{(q\rho_i)^j}{j!}}{\sum_{j=0}^{\lceil\frac{q\left(W-\sum_{c:c\neq i}h_c n_c\right)}{h_i}\rceil}\frac{(q\rho_i)^j}{j!}}.$$

One can show that for a given q, if $\rho_i < 1$ $\zeta(\mathbf{n}_{\setminus i}, q, \mathbf{h}', \mathbf{s}', W) < \zeta(\mathbf{n}_{\setminus i}, 1, \mathbf{h}', \mathbf{s}', W)$ for large W. Using this and (24), one can conclude blocking probability for the scaled system is lower in wideband systems.

B. Approximate Expression for Blocklength

If we ignore the terms $O(\log_2(r))$ in (5), we have the following approximate expression relating blocklength r, the number of information bits L and the probability of decoding failure p.

$$L \approx rC(SINR_c) - Q^{-1}(p)\sqrt{rV(SINR_c)}. \tag{27}$$

If we substitute $\sqrt{r} = x$, then the above equation is a quadratic equation in x. Solving it we get the approximate expression for r in (6).

C. Proof of Theorem 2

From (9), we have the following relation between λ_c^* and W

$$\kappa W = \lambda_c^* r_c + c(\delta) r_c \sqrt{\frac{\lambda_c^*}{d}}, \tag{28}$$

where r_c, L, $SINR_c$, and δ are related according to (6), and the expression for r_c is re-stated below:

$$r_c = \frac{L}{C(SINR_c)} + \frac{\left(Q^{-1}(\delta)\right)^2 V(SINR_c)}{2\left(C(SINR_c)\right)^2} + \frac{\left(Q^{-1}(\delta)\right)^2 V(SINR_c)}{2\left(C(SINR_c)\right)^2}\sqrt{1+\frac{4LC(SINR_c)}{V(SINR_c)\left(Q^{-1}(\delta)\right)^2}}. \tag{29}$$

Solving for λ_c^* in (28), we have that

$$\lambda_c^* = \frac{\kappa W}{r_c} + \frac{c(\delta)^2}{d}\left(1 - \sqrt{1 + \frac{4\kappa W d}{c(\delta)^2 r_c}}\right). \tag{30}$$

Scaling with respect to W directly follows from (30).

To understand the scaling with respect to $SINR_c$, we have to first study the scaling of r_c with respect to $SINR_c$. For large $SINR_c$, we have that

$$C(SINR_c) \sim \Theta\left(\log_2\left(SINR_c\right)\right), \tag{31}$$

$$V(SINR_c) \sim \Theta(1). \tag{32}$$

Therefore, $r_c \sim \Theta\left(\frac{1}{\log_2(SINR_c)}\right)$. Using (30), we get that $\lambda_c^* \sim \Theta\left(\log_2\left(SINR_c\right) - \sqrt{\log_2\left(SINR_c\right)}\right)$. Similarly, using (30), we get the scaling with respect to d as $\lambda_c^* \sim \Theta\left(1 - \frac{1}{\sqrt{d}}\right)$. If we use the square-root staffing rule with the normal approximation (see [22]), we have that $c(\delta) = Q^{-1}(\delta) \sim \Theta\left(\sqrt{-\log(\delta)}\right)$. As we increase δ, we $c(\delta) \to 0$. Using $Q^{-1}(\delta) \sim \Theta\left(\sqrt{-\log(\delta)}\right)$ we have that $r_c \sim \Theta\left(-\log(\delta)\right)$. Therefore, from (30) we get that $\lambda_c^* \sim \Theta\left(\frac{1}{-\log(\delta)}\right)$.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous referees for their valuable comments and suggestions.

REFERENCES

[1] B. Holfeld *et al.*, "Wireless communication for factory automation: An opportunity for LTE and 5G systems," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 36–43, Jun. 2016.

[2] O. N. C. Yilmaz, Y.-P. E. Wang, N. A. Johansson, N. Brahmi, S. A. Ashraf, and J. Sachs, "Analysis of ultra-reliable and low-latency 5G communication for a factory automation use case," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 1190–1195.

[3] M. Gidlund, T. Lennvall, and J. Åkerberg, "Will 5G become yet another wireless technology for industrial automation?" in *Proc. IEEE Int. Conf. Ind. Technol. (ICIT)*, Mar. 2017, pp. 1319–1324.

[4] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska, "A flexible 5G frame structure design for frequency-division duplex cases," *IEEE Commun. Mag.*, vol. 54, no. 3, pp. 53–59, Mar. 2016.

[5] *Chairman's Notes*, document R1-1704170, 3GPP: 3GPP TSG RAN WG1 Meeting 88bis, Apr. 2017. [Online]. Available: http://www.3gpp.org/ftp/TSG_RAN/WG1_RL1/TSGR1_88b/Report/

[6] *Chairman's Notes*, document R1-1810050, 3GPP, TSG RAN WG1 Meeting 94, Aug. 2018.

[7] M. Bennis, M. Debbah, and H. V. Poor. (Jan. 2018). "Ultra-reliable and low-latency wireless communication: Tail, risk and scale." [Online]. Available: https://arxiv.org/abs/1801.01270

[8] P. Popovski *et al.* (2017). "Ultra-reliable low-latency communication (URLLC): Principles and building blocks." [Online]. Available: http://arxiv.org/abs/1708.07862

[9] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim. (2017). "Introduction to ultra reliable and low latency communications in 5G." [Online]. Available: http://arxiv.org/abs/1704.05565

[10] S. Ashraf *et al.*, "From radio design to system evaluations for ultra-reliable and low-latency communication," in *Proc. Eur. Wireless Conf.*, May 2017, pp. 1–8.

[11] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5G ultra-reliable and low-latency systems design," in *Proc. Eur. Conf. Netw. Commun. (EuCNC)*, Jun. 2017, pp. 1–5.

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

[13] A. Anand, G. de Veciana, and S. Shakkottai. (Dec. 2017). "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks." [Online]. Available: https://arxiv.org/abs/1712.05344

[14] L. You, Q. Liao, N. Pappas, and D. Yuan. (Jan. 2018). "Resource optimization with flexible numerology and frame structure for heterogeneous services." [Online]. Available: https://arxiv.org/abs/1801.02066

[15] H. Shariatmadari, S. Iraji, and R. Jäntti, "Analysis of transmission methods for ultra-reliable communications," in *Proc. PIMRC*, Aug. 2015, pp. 2303–2308.

[16] H. Shariatmadari, S. Iraji, Z. Li, M. A. Uusitalo, and R. Jäntti, "Optimized transmission and resource allocation strategies for ultra-reliable communications," in *Proc. PIMRC*, Sep. 2016, pp. 1–6.

[17] D. Malak, H. Huang, and J. G. Andrews. (Nov. 2017). "Throughput maximization for delay-sensitive random access communication." [Online]. Available: https://arxiv.org/abs/1711.02056

[18] G. Durisi, T. Koch, J. Östman, Y. Polyanskiy, and W. Yang, "Short-packet communications over multiple-antenna Rayleigh-fading channels," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 618–629, Feb. 2016.

[19] B. Singh, Z. Li, O. Tirkkonen, M. A. Uusitalo, and P. Mogensen, "Ultra-reliable communication in a factory environment for 5G wireless networks: Link level and deployment study," in *Proc. IEEE Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Sep. 2016, pp. 1–5.

[20] 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.

[21] L. Kleinrock, *Queueing System*, vol. 1. Hoboken, NJ, USA: Wiley, 1975.

[22] M. Harchol-Balter, *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[23] S. Sesia, I. Toufik, and M. Baker, *LTE—The UMTS Long Term Evolution: From Theory to Practice*. Hoboken, NJ, USA: Wiley, 2009.

Arjun Anand received the B.Tech. degree in electronics and communication engineering from the National Institute of Technology, Calicut, India, in 2011, the M.Eng. degree in telecommunications from the Indian Institute of Science, Bengaluru, India, in 2013, and the Ph.D. degree in electrical engineering from The University of Texas at Austin, Austin, TX, USA, in 2018. From 2013 to 2014, he was a Staff-I Engineer at Broadcom Communication Technologies Pvt., Ltd., Bengaluru, where he was involved in the development of products based on IEEE 802.11 ac standard. He is currently a Research Scientist at Intel Corporation, Hillsboro, OR, USA. His research interests include resource allocation in wireless networks, queuing theory, and reinforcement learning.

Gustavo de Veciana (S'88–M'94–SM'01–F'09) received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993, respectively. He joined the Department of Electrical and Computer Engineering, where he is currently a Cullen Trust Professor of engineering. He has served as the Director and an Associate Director of the Wireless Networking and Communications Group, The University of Texas at Austin, from 2003 to 2007. His research focuses on the analysis and design of communication and computing networks, data-driven decision-making in man-machine systems, and applied probability and queuing theory. In 2009, he was designated IEEE Fellow for his contributions to the analysis and design of communication networks. He was a recipient of the National Science Foundation CAREER Award 1996 and a co-recipient of five best paper awards including: the IEEE William McCalla Best ICCAD Paper Award for 2000, the Best Paper in ACM TODAES Jan 2002–2004, the Best Paper in ITC 2010, the Best Paper in ACM MSWIM 2010, and the Best Paper IEEE INFOCOM 2014. He currently serves on the Board of Trustees of IMDEA Networks Madrid. He has served as an Editor and is currently serving as the Editor-at-Large for the IEEE/ACM TRANSACTIONS ON NETWORKING.