Performance modeling and dimensioning of latency-critical traffic in 5G networks

Mohammed Abdullah, Salah Eddine Elayoubi, Tijani Chahed, Abdel Lisser

Performance modeling and dimensioning of latency-critical traffic in 5G networks

Mohammed Abdullah*†, Salah Eddine Elayoubi*, Tijani Chahed†, Abdel Lisser*
* CentraleSupélec, Université Paris-Saclay, CNRS, L2S, Gif-sur-Yvette, France
† Télécom SudParis, Institut Polytechnique de Paris, SAMOVAR, Palaiseau, France

***Abstract*—We propose a new performance model for transport of time-critical Ultra Reliable Low Latency Communications (URLLC) traffic in 5G networks and Beyond and apply it to dimensioning of such systems. The Quality of Service (QoS) requirement is formulated in terms of an outage probability which is defined as the probability that the latency exceeds a maximal allowed budget, and which should be kept very low. We develop a generic queuing model to compute this outage probability and adapt it to integrate the specificity of the 5G radio interface, taking into account the heterogeneity of users radio conditions and thus their Modulation and Coding Schemes (MCS) as well as retransmissions due to errors on the radio link. We also propose a low complexity method to calculate it using a geometric tail approach to approximate the tail distribution of the queue, for relevant arrival distributions: Poisson and Binomial. We show numerically the performance of our exact model and approximation and that they yield very accurate performance against simulations, and in comparison with other models from the state of the art. We also show the system dimensioning in terms of required resources to satisfy the outage constraint.**

I. Introduction

Ultra-Reliable Low Latency Communications (URLLC) service was introduced in 5G standardization [1] to tackle critical services such as autonomous driving, industry 4.0, smart grid, etc. A typical performance target is $1ms$ delay and $99,999\%$ reliability constraints [2]. Several features were introduced in 3GPP standardization to help reach URLLC low latency and high reliability constraints. For instance, short Transmission Time Interval (TTI) can be combined with blind repetitions for reaching reliability in case of harsh radio environments. These techniques enable that the radio latency (i.e., the time between the packet generation and its decoding by the base station) is very low (e.g. between 0.5 and 1 ms). However, the underlying assumption is that resources are always available and latency is only due to packet alignment, scheduling grant reception, over-the-air transmission and packet decoding.

When resources are scarce or traffic load is high, an additional component is added that is the queuing delay, i.e., the delay before a resource is available for the packet to be scheduled. When URLLC service is in competition with enhanced Mobile Broadband (eMBB) service, the problem of queuing is solved by preemptive scheduling, where URLLC packets are served immediately upon arrival by preempting some eMBB resources [3], [4]. However, when URLLC packets compete with other URLLC packets, preemption is not possible and over-reservation of resources may be needed. For URLLC periodic traffic scenarios, semi-persistent scheduling (SPS) is proposed and resources are pre-reserved for each user [5]. However, for sporadic traffic scenarios, SPS is inefficient and mastering the queuing delay is still an open problem. Solving this problem requires efficient performance models for resource dimensioning, and is the objective of this paper.

Many works tackled this dimensioning issue and proposed performance models for URLLC based on queuing theory. They however often make strong assumptions on arrival process, typically Poisson, and service distribution. The work in [6] proposed an M/M/1 model that is based on the assumption of Poisson arrivals of packets and an exponential model for the variation of packet sizes due to different radio conditions. [7] relaxed the exponential assumption for the service rate and adopted an M/G/1 model with vacations (to account for the presence of other users), but with two restrictive assumptions. First, the "General" service model is due to different packet sizes and not different radio conditions, and second, packets are supposed to be served by one server in continuous time, while packets in 5G are multiplexed in the spectrum dimension (several servers) and time is slotted. [8] makes use of an M/GI/∞ model in order to study resource allocation for URLLC. [9] considers a M/M/m/K queue to model the system reliability for a worst case scenario where users are assumed to be at the cell edge. The work in [10] derives generic end-to-end latency distribution of any network topology which enables to determine percentiles, and applies it to the cases of exponential and deterministic service distributions. The work in [11] considers a risk-resistant approach (risk is delay outage) to minimize the latter based on an M/G/1 queuing model. The work in [12] considers a definition of delay violation probability which includes two components: a packet delay exceeding a given constraint and that errors are undetected, at the ARQ level. The arrival process is assumed to follow a Geometric distribution with bulk arrivals (Geo$^{[X]}$/G/1 model).

Some works make use of different analytical tools, such as Extreme Value Theory [13] and stochastic network calculus ([14] and [12] in the case of perfect error detection), but we adopt queuing theory in our present work as it has the advantage of avoiding over-dimensioning that is inherent to approaches such as extreme value theory and network calculus.

We develop in this paper a queuing model for computing the outage probability (i.e., the probability that the packet delay is larger than a given target). We formulate the equations describing the evolution of the number of packets waiting in the queue and show how to derive the outage probability. We

extend our model to integrate heterogeneous radio conditions leading to different modulation and coding schemes (MCS) and retransmissions due to radio errors (fast fading). We also propose a low complexity method to calculate the outage probability using a geometric tail approach to approximate the tail distribution of the queue, for Poisson and Binomial arrival distributions. For the numerical experimentations, we develop a discrete-time simulator of the system and show that the developed queuing models, both exact and approximate, are very accurate. In the basic case (homogeneous MCS, no retransmissions), we also compare our exact model to a continuous-time M/D/c model and to the M/M/c/K model of [9] and show its superiority. We eventually illustrate the dimensioning of the system in terms of the number of required resources so as to satisfy the outage constraint.

The remainder of this paper is organized as follows. In section II, we develop a discrete time queuing model for the outage probability and apply it in section III to different 5G URLLC scenarios, in the presence of radio errors and retransmissions. Section IV illustrates the performance and accuracy of the proposed queuing model against classical models and simulations and its usage for system dimensioning. Section V eventually concludes the paper.

II. Latency outage model

In the following, we will use the notation $\mathcal{X}$ for sets, $|\mathcal{X}|$ for the cardinality of sets, $\mathbf{x}$ for line vectors and $\mathbf{x}^T$ their transpose, and $\mathbf{X}$ for matrices with elements X_{ij}.

A. System and traffic model

We consider a 5G cell where resources are organized into Resource Blocs (RB) and (mini-)slots. The slot is of size T ms and there are some reserved RBs for URLLC. Let R be the amount of resources reserved per slot for URLLC. In each slot, packets are generated following some stochastic process, and packet arrivals in different slots are independent. Packets might be of equal or variable sizes. We make no assumptions on the distributions of the arrival process and packet sizes. The number for resources required for serving new arriving packets during slot t is a discrete random variable $a(t)$, defined in some subset $\mathcal{A}$ of $\mathbb{N}$, the set of positive integers. As of the delay budget, a packet may stay for $\delta \geq 1$ slots in the system before its delay budget expires, otherwise it is in outage.

B. Outage probability formulation

As the queue follows a First Come First Serve (FCFS) discipline, a packet generated in a slot sees other packets generated within the same slot and those generated in previous slots that are still in the queue waiting for service, if any.

In slot t, knowing that there are R reserved resources, we define the queue duration $B(t)$ as the amount of resources that will be needed in the future slots to serve the backlogged traffic after using all the resources of slot t, and is given by:

$$B(t) = (a(t) + B(t-1) - R)^+ , \tag{1}$$

where $(x)^+ = \max(0, x)$, $a(t)$ is, as stated above, the amount of resources needed for serving new packets arriving in slot t, $B(t-1)$ is the queue duration in the previous slot and $a(t) + B(t-1)$ is the total amount of resources needed for serving all the backlogged packets from previous slots plus the new packets at slot t. As there are R resources available in each slot, at most R among the required resources are consumed, and the remaining packets overflow to the next slot.

We define the probability of overflow as the probability that some packets that are present in slot t will be still not served in slot $t + \delta - 1$. This is computed by:

$$O = \lim_{t\to\infty} Pr[B(t) > (\delta - 1)R] \tag{2}$$

as packets constituting $B(t)$ have been in the system at least during slot t, and some of them will be surely in outage if the resources in the next $(\delta - 1)$ slots are not sufficient to serve all of them. The limit in (2) exists as the associated Markov chain described next fulfills both the ergodicity and irreducibility conditions.

Note that the *overflow* probability of equation (2), which refers to all packets present in slot t being still not served in slot $t+\delta-1$, is slightly different from the *outage* probability, defined as the probability that a packet stays more than δ slots. Indeed, when an overflow occurs, it does not necessarily imply that all packets of slot t are lost, but that at least one of them is lost. Overflow is thus an upper bound on outage. We show next how the outage can be derived using the overflow.

C. Overflow and outage probabilities computation

In order to compute the outage probability, we need to first determine the distribution of the overflow in steady-state. Eqn. (1) involves three random variables:

- $B(t)$, a discrete integer random variable that takes its values in $[0, \infty[$. Let $q_b(t)$ be the probability that $B(t)$ takes the value $b \in [0, \infty[$.
- $B(t-1)$ has the same limiting distribution as $B(t)$
- and $a(t)$, that is the amount of resources needed for serving the new packet arrivals. $a(t)$ is independent from $B(t-1)$ and takes its values in some set $[0, a_{max}]$, where a_{max} is a positive integer (that might be infinite). Let z_j be the limiting probability that $a(t) = j$, $j \leq a_{max}$.

In steady-state, let $\mathbf{q} = (q_b, b \geq 0)$ is the vector of duration probabilities. Setting a maximal queue duration $B_{max} >> R$ and defining $\mathcal{B} = [0, B_{max}] \subset \mathbb{N}$ as the space of possible values, we write the following set of linear equations:

$$\mathbf{q} = \mathbf{q}\mathbf{Q} \tag{3}$$

$\mathbf{Q}$ is the transition matrix (Q_{jb}, j and $b \in \mathcal{B}$, is the transition probability from queue duration j at t to b at $t+1$):

$$Q_{jb} = \begin{cases} z_{b+R-j}, & \text{if } b \in]0, B_{max}[\\ \sum_{i \geq b+R-j} z_i, & \text{if } b = B_{max} \\ \sum_{i \leq R-j} z_i, & \text{if } b = 0 \\ 0, & \text{otherwise} \end{cases}$$

This set of equations can be solved by adding the normalizing equation:

$$\sum_{b \geq 0} q_b = 1 \tag{4}$$

The overflow probability (2) is obtained by:

$$O(\delta) = 1 - \sum_{b=0}^{(\delta-1)R} q_b \tag{5}$$

We now compute the outage probability defined, again, as the probability that a particular packet stays more than δ slots in the system. It can be approximated using the overflow probability by:

$$\theta(a, \delta) = \sum_{r=1}^{R-1} \frac{r}{R} q_{(\delta-1)R+r} + O(\delta + 1) \tag{6}$$

where the term $q_{\delta R+r}$ indicates that exactly $r < R$ resources are missing for serving all the traffic present in slot i during the δ subsequent slots, it is weighed by $\frac{r}{R}$ to indicate that the outage in this case occurs for a fraction of the slot. $O(\delta + 1)$ is the overflow probability, computed as in equation (5) but supposing that there is an additional slot within the delay budget, this term indicates that more than R resources are lacking, and there is at least a whole slot in outage.

III. Model adaptation to URLLC scenarios

We now show how to apply the developed discrete queuing model to URLLC scenarios. We first show the traffic and radio characterises for URLLC, and then develop a low complexity queuing model that fits well this URLLC system. We then show how to integrate retransmissions due to radio errors.

A. URLLC traffic model

1) Arrival process: We consider the most common case in industrial applications where U URLLC users (e.g., machines) are connected to an access point. If the probability of generating a packet during a slot is f, the probability that the newly generated packets require i resources is thus $Binomial(U, f)$, and the probability of having i packets arriving in a time slot is:

$$\zeta_i = \binom{U}{i} f^i (1-f)^{U-i}. \tag{7}$$

Note that, if the number of users is large and f is small, this distribution can be approximated by a Poisson of intensity $\lambda = Uf$.

2) Service process: We consider the general case when different users are subject to different radio conditions and each packet uses an MCS that is drawn from some known distribution. Our assumption here is that the statistics of the radio channel are known. Once a packet is generated, it uses MCS k with probability β_k, with $\sum_{k=1}^{K} \beta_k = 1$, K being the number of available MCS. When a packet uses MCS k, it consumes an amount of RBs equal to α_k. The amount of resources consumed by a user u in slot i is thus a random variable with distribution:

$$X_{u,i} = \begin{cases} 0, & \text{with prob. } (1-f) \\ \alpha_k & \text{with prob. } f\beta_k \end{cases} \tag{8}$$

Without loss of generality, we assume that the MCS are sorted in increasing order of spectral efficiency, meaning that $\alpha_1 > ... > \alpha_K$. The total number of resources requested by new packets generated in slot i is then given by:

$$a(i) = \sum_{u=1}^{U} X_{u,i} \tag{9}$$

Equations (3)-(4) allow the computation of the overflow probability. In these equations, z_j describes the probability that new arrivals in a slot require j resources (RBs), and ca be computed using the multinomial distribution. Let $m(u_0, u_1, ..., u_K)$ be the probability of having, in a given slot, a vector of generated packets $\vec{u} = (u_0, u_1, ..., u_K)$, where u_k, $k > 0$ is the number of packets with MCS k, and u_0 is the number of users that did not generate any packet. Let $\mathcal{U}$ be the space of all possible vectors $\vec{u}$ such that $\sum_{k=0}^{K} u_k = U$,

$$m(\vec{u}) = \frac{U!}{\prod_{k=0}^{K} u_k} (1-f)^{u_0} f^{U-u_0} \prod_{k=1}^{K} \beta_k^{u_k} \tag{10}$$

Let $\mathcal{U}_j$, $j \in [0, \alpha_1 U]$, be the subset of $\mathcal{U}$ such that $\sum_{k=0}^{K} u_k \alpha_k = j$. The probability of consuming j resources is thus computed by:

$$z_j = \sum_{\vec{u} \in \mathcal{U}_j} m(\vec{u}) \tag{11}$$

Note that in the case of homogeneous MCS, e.g. the operator uses a robust MCS for all users to ensure a very high reliability, all packets consume exactly α_1 RBs. Defining a resource unit equal to α_1 RBs, one packet consumes one resource unit and z_j becomes equal to the probability of i packets arriving ($z_i = \zeta_i$), in this case, the outage probability is calculated as in equation (6), While in the case of heterogeneous radio conditions, where different users may be subject to different radio conditions and each packet may use an MCS that is drawn from some distribution that assumed to be known, the outage probability is then given by:

$$\theta(\delta) = \sum_{r'=1}^{R'} \frac{r'}{R'} \Big[\sum_{j=Mr'}^{M(r'+1)-1} q_{(\delta-1)R+j} \Big] + O(\delta + 1) \tag{12}$$

where M is the mean number of RB's consumed by a packet and $R' = \frac{R}{M}$.

B. A low complexity model

We have seen above that Binomial and Poisson arrivals are two particularity interesting cases. The following lemma shows that, under such traffic arrivals, the complexity of the model in (3)-(4) can be drastically decreased. Indeed, both the outage and overflow probabilities depend on the probability distribution of the queue length in the steady-state, and as we are tracking a very rare event (packet loss probability smaller than 10^{-5}), the Markov chain should be truncated at a high value of the queue length B_{max}, where B_{max} depends on the arrival rate, R (resources Blocks), and δ delay budget. Instead of solving B_{max} equations, we show in the following lemma that a small number of equations is sufficient.

Lemma 1. *For binomial and Poisson arrivals, the equilibrium probabilities* $\mathbf{p}_j$*, exhibit the geometric tail behavior:*

$$q_j \sim \gamma\eta^j \;\; as \;\; j \to \infty \tag{13}$$

for some constant $\gamma > 0$ *and* $0 < \eta < 1$*. For sufficiently large* M*, we have:*

$$q_j = q_M \eta^{j-M}, \;\; j \geq M. \tag{14}$$

Proof. We here provide a sketch of the proof. In [15], Theorem $C.1$ (Appendix C) states that the geometric tail approach applies under the following conditions:

(a) The generating function $\sum_{j=0}^{\infty} q_j x^j$ for $|x| < 1$ has the form $\frac{N(x)}{D(x)}$, where $N(x)$ and $D(x)$ are two analytic functions whose domains of definition can be extended to a region $|x| > L > 1$.
(b) $D(x) = 0$ has real root x_0 on the interval $(1, L)$
(c) The zero $x = x_0$ of $D(x)$ is of multiplicity 1.

We then compute the PGF of the queue length as follows:

$$P(x) = \sum_{j=0}^{\infty} \Big(z_j \sum_{k=0}^{R} q_k + \sum_{k=R+1}^{R+j} q_k z_{j-k+R} \Big) x^j \tag{15}$$
$$= \frac{x^{-R}\Big[\sum_{j=0}^{\infty} z_j x^j\Big]\Big[\sum_{k=0}^{R-1} q_k (x^R - x^k)\Big]}{1 - x^{-R}\sum_{j=0}^{\infty} z_j x^j}$$

Poisson arrivals: In the Poisson arrival case with rate λ ($R > \lambda$) the probability generating function (15) is:

$$P(x) = \frac{x^{-R} e^{-\lambda(1-x)} \sum_{k=0}^{R-1} q_k (x^R - x^k)}{1 - x^{-R} e^{-\lambda(1-x)}} \tag{16}$$

Binomial arrivals: If packets arrive in a binomial distribution, let n be the number of users, each being active with a probability f ($R > nf$), we have:

$$P(x) = \frac{x^R (fx + 1 - f)^n \sum_{k=0}^{R-1} q_k (x^R - x^k)}{1 - x^{-R}(fx + 1 - f)^n} \tag{17}$$

It is easy to see that condition (a) is valid in both cases. For conditions (b) and (c), below is a sketch of the proof:

1) $D'(x) > 0$ for $0 < x < 1 < x_1$, and $D(1) = 0$,
2) $D'(x) < 0$ for $x > x_1$,
3) $\lim_{x\to\infty} D(x) = -\infty$,
4) $D'(x) = 0$ has only root at x_1,

Taking $x_1 = \frac{R}{\lambda}$ for the Poisson arrival case and $x_1 = \frac{R(1-f)}{fn - Rf}$ for the Binomial case, leads to the proof that the geometric tail approach is applicable, and $\eta = \frac{1}{x_0}$ in this case. □

C. Integrating radio errors and retransmissions

URLLC users usually use a robust MCS so that packets are lost with a small probability ϵ[1]. This radio loss cannot be neglected when a very high reliability is sought. In case of packet loss, it is retransmitted until it is decoded by the base station, but each retransmission generates an additional delay and outage occurs if the packet is not well received before the budget of δ slots expires. In order to model the impact of retransmissions, we introduce to the overflow model two modifications, as follows:

- The activity factor of users is increased by a factor of $1/(1-\epsilon)$ to account for the fact that a packet is repeated for a geometric number of times. The probability of generating a packet during a slot becomes: $f' = \frac{f}{1-\epsilon}$
- The overflow probability is computed accounting for the multiple retransmissions.

Knowing that the error probability is low as URLLC packets use robust MCS, and that the outage probability is low for the targeted traffic regimes, we consider only one transmission and one retransmission when computing the outage. The outage probability integrating radio errors becomes:

$$\theta'(a,\delta) = \epsilon \sum_{\delta_1=1}^{\delta-1} \Big[\Big(\sum_{b_1=(\delta_1-1)R+1}^{\delta_1 R} q_{b_1} \Big) \Big(1 - \sum_{b_2=0}^{(\delta-\delta_1)R} q_{b_2}\Big)\Big] + (1-\epsilon)\theta(a,\delta) \tag{18}$$

where $\theta(a,\delta)$ is the outage probability with no radio errors obtained by replacing f with f'. The second term accounts for the retransmission delay in case of radio loss. In this case, the first transmission consumes exactly $\delta_1 < \delta$ slots (the term $(\sum_{b_1=(\delta_1-1)R}^{\delta_1 R} q_{b_1})$), and outage occurs if the second transmission takes more than $\delta - \delta_1$ slots (the term $(1 - \sum_{b_2=0}^{(\delta-\delta_1)R} q_{b_2})$).

IV. Numerical experiments

We now consider two sets of numerical experiments. The first set (section IV-A1) aims at model validation and compares the model to a simulator that corresponds to a limited system (only the scheduler is modeled, MCS distribution is taken as input). The second set of experiments (section IV-B) aims to show how the model can be used for dimensioning of real URLLC systems, and makes use of a complete system level simulator with link adaptation and user generation.

A. Model validation and comparison to state of the art

1) Simulator description for the benchmark: In this section (figures 1 and 2), we make use of the mathematical models and a Markov chain simulator based on theoretical MCS distributions, issued from a large scale system level simulator and illustrated in Table I. In case of a fixed MCS, all users use MCS 1. When an MCS with a spectral efficiency of y bit/s/Hz is used, and knowing that the RB size is h Hz, a packet of size $s = 96$ bits occupies a number of RBs equal to $\lceil \frac{s}{Thy} \rceil$, where $\lceil x \rceil$ is the largest integer greater than or equal to x.

In the Markov chain simulator, time is divided into slots of size $T = 0.144$ ms and there are R reserved RBs for URLLC. Packets are all of equal size. In each slot, each user generates a packet following a Bernoulli law with parameter f, and if a packet is generated, it chooses at random an MCS following the input distribution. Packets are served following a FCFS discipline. When a packet is generated (following some arrival

[1]The MCS is usually selected based on a pessimistic Signal to Interference and Noise Ratio (SINR), by subtracting a margin on the estimated SINR.

process), it is put at the end of the queue. A time slot is filled with the packets at the head of the queue until all of the R RBs are occupied or the queue is empty. When a packet cannot be scheduled on one slot as the remaining resources are not sufficient, it can be scheduled on two consecutive slots. A packet is lost with probability ϵ. It is then regenerated and placed at the end of the queue for retransmission. A packet is considered in outage if it stays in the system more than δ slots. In our numerical applications, we take $\delta = 4$ mini-slots.

TABLE I: MCS distribution and resource consumption

k	1	2	3	4	5	6	7
β_k	.027	.009	.002	.003	.057	.032	.09
α_k	27	18	11	7	5	4	3
k	8	9	10	11	12	13	14
β_k	.112	.039	.0136	.010	.074	.137	.014
α_k	3	2	2	2	2	1	1

2) Comparison with the state of the art: For the comparison with the state of the art, we consider a system where all users use a common MCS, as this corresponds to an interesting practical case (no link adaptation due to stringent latency requirements), and as the state of the art does not allow to model heterogeneous MCS. The case with link adaptation will be presented in the next section. The arrival process is considered as Poisson for fitting with the queuing literature.

We first compare our model to the Monte Carlo simulation for validation. Figure 1 plots the outage probability obtained from the analytical model (equation (6)) and from simulations. We observe a very good fit between both results. We also plot the outage obtained from the geometric tail approach (section III-B), and the approximation shows to be very good, with a low coomplexity (solving only a set of 15 equations, instead of $B_{max} = 100$ equations for the original model.

We now move to the comparison with the literature. As all packets consume the same amount of resources, a natural model is a continuous time queue with $c = R$ servers. A model exists for M/D/c queues, since the work of Crommelin in 1932 [16]. We apply the algorithm of [15] (pages 378-379) and plot in figure 1 the outage probability of the M/D/c model ($c = R$ servers, Poisson arrival rate of Uf packets per slot, deterministic service time of 1 slot). Figure 1 shows also that the M/D/c model overestimates the outage probability. For further comparison with the state of the art, we also consider the M/M/c/K loss model proposed in [9] for URLLC, where the service is approximated as exponential, c is the number of servers, and K is the maximum number of packets the system can hold, computed in [9] as the number of packets upon arrival that discourages a packet from being queued as it corresponds to an outage ($K = c\delta$ in our case). Figure 1 shows that this model is not adequate, as it largely overestimates the outage for small loads (our region of interest), and underestimates it for high loads (due to blocking).

3) Model validation for more realistic scenarios: We now move to the general case where link adaptation makes different packets consume different amounts of resources. Binomial traffic model is used, for modeling a fixed number of URLLC users generating sporadic traffic. We use the MCS distribution of Table I. Figure 2 compares the outage probabilities of our model with simulations, varying the activity factor f for users. We simulate two scenarios: the first where we neglect losses due to fast fading (packets are in outage only due to queuing delay), and the other where fast fading may lead to a loss and the packet is retransmitted (the delay being considered from the packet generation to its correct decoding). We observe a perfect fit with simulations, meaning that equations (18) and (6) give tight approximations of the packet outage, with and without retransmissions due to radio errors.

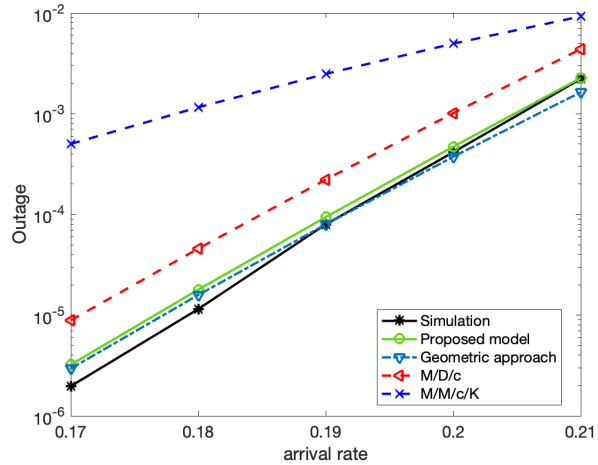

Fig. 1: Outage for the homogeneous MCS case ($R = 5$, $\delta = 4$).

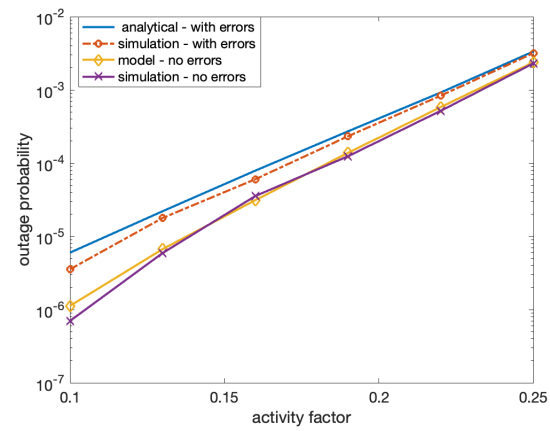

Fig. 2: Outage for common MCS ($R = 24$, $U = 20$, $\delta = 4$).

B. Resource dimensioning

We now show how to use our model for resource dimensioning. For this purpose, we make use of our performance model, and apply it on a live network modeled by a system level simulator. Note that this simulator, described in the following, is different from the Markov chain simulator used above that considered only the scheduler, without considering neither user arrivals/departures, nor link adaptation.

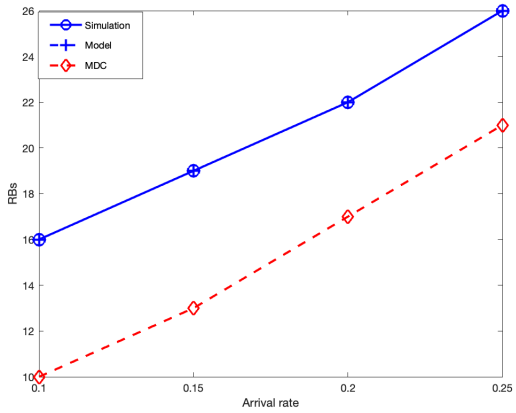

Fig. 3: Resource dimensioning for the URLLC service.

A BS serves a set of URLLC users. At the start of each run, the positions of the users are drawn randomly in the cell, and their path loss is computed. For each slot, a fast fading is generated and the resulting Channel Quality Indicator (CQI) is computed. The corresponding packet is then sent using the adequate MCS, corresponding to the computed CQI. A RAN management entity is then responsible of collecting information about the traffic and radio conditions statistics. This entity, located for instance within the NSSMF (Network Slice Subnet Management Function), computes the average arrival rate of packets and the aggregated MCS distribution (probability for a packet of using a given MCS).

These information are then sent to a URLLC slice dimensioning entity that implements the performance evaluation model. In particular, the packet arrival rate and the MCS distribution are combined with the service-related information (target delay) and the amount of URLLC reserved resources R to compute the packet outage. R is then adjusted in the model until reaching the target outage of 10^{-5}.

Figure 3 shows the resulting dimensioning in terms of resource blocks for different arrival rates. Three methods are implemented for computing the required amount of resources:

1) The proposed discrete time model,
2) the M/D/c model, making the assumption that all packets consume the same amount of resources, as if they were using a common MCS. This amount of resources is computed as the average over all MCSs.
3) a pure simulation method, based on trial and success, where different values of R are tested on the live network on a sufficiently long time, and then the minimal value of R where an acceptable outage is observed is selected. This method is not acceptable in practice and is only shown for validation purposes.

We observe that our model and simulations give the same system dimensioning. However, the M/D/c model underestimates the required resources and results in practice in an unacceptable outage.

V. Conclusion

In this paper, we developed a performance model and dimensioning for latency-critical traffic in 5G networks. We considered devices that transmit packets with stringent delay constraints and formulated the equations describing the evolution of the queue duration. We computed the outage probability, i.e., the probability that the delay in the system exceeds a maximal threshold, and proposed a low computational complexity approximation based on geometric tail approach. We accounted for cases where users use different MCS, and included the impact of retransmission due to errors. We showed numerically the accuracy of the developed models against simulations, and their better performance in comparison with models from the state of the art. We also illustrated the system dimensioning in terms of required resources for URLLC so as to meet the stringent outage probability.

Acknowledgment

This research work was partially supported by DIGICOSME under project OPITU, grant no. ANR-11-IDEX-0003-02.

References

[1] *System Architecture for the 5G System*, 3GPP, TS 23.501, Dec. 2017, version 15.0.0 Release 15.
[2] M. Bennis, M. Debbah, and H. V. Poor, "Ultra reliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.
[3] K. I. Pedersen, G. Pocovi, J. Steiner, and S. R. Khosravirad, "Punctured scheduling for critical low latency data on a shared channel with mobile broadband," in *IEEE VTC-Fall*, 2017.
[4] M. Morcos, M. Mhedhbi, A. Galindo-Serrano, and S. Elayoubi, "Optimal resource preemption for aperiodic urllc traffic in 5g networks," in *IEEE PIMRC*, 2020.
[5] Z. Li, M. A. Uusitalo, H. Shariatmadari, and B. Singh, "5g urllc: Design challenges and system concepts," in *2018 15th international symposium on wireless communication systems (ISWCS)*. IEEE, 2018, pp. 1–6.
[6] A. Chagdali, S. E. Elayoubi, A. M. Masucci, and A. Simonian, "Performance of urllc traffic scheduling policies with redundancy," in *2020 32nd International Teletraffic Congress (ITC 32)*. IEEE, 2020.
[7] H. Jang, J. Kim, W. Yoo, and J.-M. Chung, "Urllc mode optimal resource allocation to support harq in 5g wireless networks," *IEEE Access*, vol. 8, pp. 126 797–126 804, 2020.
[8] A. Anand and G. de Veciana, "Resource allocation and harq optimization for urllc traffic in 5g wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, 2018.
[9] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5g ultra-reliable and low-latency systems design," in *2017 European Conference on Networks and Communications (EuCNC)*, 2017, pp. 1–5.
[10] P. Schulz, L. Ong, B. Abdullah, M. Simsek, and G. Fettweis, "End-to-end latency distribution in future mobile communication networks," in *WSA 2020; 24th International ITG Workshop on Smart Antennas*, 2020.
[11] B. Shi, F.-C. Zheng, C. She, J. Luo, and A. G. Burr, "Risk-resistant resource allocation for embb and urllc coexistence under m/g/1 queueing model," *IEEE Trans. on Vehicular Technology*, vol. 71, no. 6, 2022.
[12] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal-Biyikoglu, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 721–734, 2019.
[13] C.-F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Communications Letters*, vol. 22, no. 6, pp. 1292–1295, 2018.
[14] J. Zeng, Y. Song, T. Wu, T. Lv, and S. Zhou, "Guaranteeing qos for noma-enabled urllc based on $\kappa - \mu$ shadowed fading model," *Sensors*, vol. 22, no. 14, 2022.
[15] H. C. Tijms, *A first course in stochastic models*. Wiley, 2003.
[16] C. Crommelin, "Delay probability formulae when the holding times are constant," *Post Office Electrical Engineer's Journal*, vol. 25, 1932.