A Large Deviations Model for Latency Outage for URLLC

Salah Eddine Elayoubi[1(✉)], Nathalie Naddeh[2,3], Tijani Chahed[2], and Sana Ben Jemaa[3]

[1] Université Paris Saclay, CentraleSupelec, L2S, CNRS, Gif-Sur-Yvette, France
salaheddine.elayoubi@centralesupelec.fr
[2] Institut Polytechnique de Paris, Telecom SudParis, Palaiseau, France
{nathalie.naddeh,tijani.chahed}@telecom-sudparis.eu
[3] Orange Labs, Chatillon, France
sana.benjemaa@orange.com

Abstract. In this paper, we develop an analytical model for radio resource dimensioning for latency-critical services in 5G networks. URLLC (Ultra-Reliable Low Latency Communications) service is introduced in the 5G networks to respond to the requirements of critical applications such as self driving cars, industry 4.0, etc. Its stringent requirements in terms of latency and reliability are challenging to meet and are usually tackled by resource over-dimensioning. In this paper, we develop large deviation bounds for the outage probability i.e., the probability that the packet delay exceeds a given target. Our numerical applications show that the bounds are sufficiently tight for mastering over-dimensioning. We then develop a resource dimensioning framework based on the developed bounds and apply it to a large scale system level simulator. Our simulation results show that the developed model, when coupled with field-based radio condition distributions, allows achieving the reliability targets with acceptable cost in terms of resource consumption.

Keywords: URLLC · large deviation bounds · latency · reliability · 5G Networks · dimensioning

1 Introduction

1.1 Context

Ultra-Reliable Low Latency Communications (URLLC) was introduced in the 3GPP 5G-NR (Third Generation Partnership Project Fifth Generation New-Radio) standardization [1,2] to tackle critical services such as autonomous driving, mission critical applications, smart grid, etc. Depending on the use case, the Quality of Service (QoS) requirements vary, such as $1ms$ and $99,999\%$ delay and reliability constraints [5]. Several features were introduced in the 3GPP standardization to help reach URLLC low latency and high reliability constraints.

E. Hyytiä and V. Kavitha (Eds.): VALUETOOLS 2022, LNICST 482, pp. 223–239, 2023.
https://doi.org/10.1007/978-3-031-31234-2_14

For instance, on the Medium Access Control (MAC) layer, short Transmission Time Interval (TTI) or a mini slot is applied which allows scheduling over 2, 4 or 7 Orthogonal Frequency Division Multiplexing (OFDM) symbols. These techniques enable latency reduction on the radio level [19], ensuring that the radio latency (i.e., the time between the packet generation and its decoding by the base station) is below 0.5 ms. However, the underlying assumption is that resources are always available and latency is only due to the packet alignment, scheduling grant reception, over-the-air transmission and packet decoding. When resources are scarce or traffic load is high, an additional component occurs is the queuing delay, i.e. the delay before a resource is available for the packet to be scheduled. This queuing delay has to be added to the other delay components and taken into consideration in the overall radio latency. When URLLC packets are in competition with large enhanced Mobile Broadband (eMBB) packets, the problem of queuing is solved by the feature of preemptive scheduling, where URLLC packets are served immediately upon arrival by preempting part of eMBB resources, provided that the resources are available and ready to serve URLLC packets, with the adequate numerology [16,17]. When URLLC packets are in competition with other URLLC packets, preemption is not possible and over-reservation of resources may be needed. When traffic is periodic, semi-persistent scheduling (SPS) is proposed and resources are pre-reserved for each of the users [11,15]. For sporadic traffic, SPS is highly inefficient and mastering the queuing delay for URLLC is still an open problem.

In this paper, we develop a mathematical model for computing the outage probability (i.e. the probability that the packet delay is larger than a target). Our model is based on the large deviations theory [21], and consists in finding an upper bound of the outage probability. We develop several bounds, suitable for the URLLC sporadic traffic model and show the tighter bounds in two cases: a very stringent delay budget where the radio procedures do not give room for further queuing delays, and a less stringent case where several slots are available for queuing within the delay budget. Our numerical results show that the derived models can be used for resource dimensioning and do not lead to an excessive over-dimensioning.

1.2 Related Works

A large number of papers in the literature deal with the mechanisms that allow reaching low latency on the radio interface, when multiplexing URLLC with eMBB. Authors in [10,18] examine the impact of changing the TTI length dynamically on serving the URLLC packets while meeting the deadline, while guarantying eMBB performance. Other works discuss the semi-persistent scheduling approach [15]. [11] computed the amount of resources to be reserved for URLLC users, knowing a deterministic traffic pattern and a target reliability. Preemptive scheduling has been analyzed in several studies. Authors in [17] present a new scheduling algorithm where URLLC traffic is dynamically multiplexed through puncturing the enhanced Mobile Broad-Band (eMBB) traffic, with added recovery mechanism for punctured eMBB packets. In [16], the

authors propose a joint optimization framework for URLLC and eMBB with preemptive scheduling in order to achieve better URLLC performance while limiting the impact on eMBB throughput. Also the authors in [3] propose a deep reinforcement learning approach for preemptive scheduling.

Priority scheduling has also been the subject of a large pan of the literature. In [8], the authors proposed a priority-based resource reservation mechanism aiming to reduce URLLC delay and packet loss, while limiting the impact on eMBB. The authors in [13] proposed a multiplexing method for eMBB and URLLC with service isolation, formulated as an Adaptive Modulation and Coding (AMC) optimization problem.

In the above-discussed body of works, the objective was to achieve flexibility for serving URLLC in the presence of lower priority eMBB services, always with the assumption of a sufficiently large amount of resources in the cell. However, in some industrial situations, URLLC traffic load may be large and the (local) network operators have to provision sufficient resources while avoiding over-dimensioning, and preemption between URLLC users is not possible. In this context, many works proposed grant-free contention-based channel access for URLLC in the uplink. Authors in [20] proposed to send these replicas in a contention-based manner on different frequency resources on consecutive time slots, while in [9] the authors considered a more flexible scheme where replicas can be sent on any of the available time-frequency resources. These schemes focused on the uplink, as the centralized orthogonal resource allocation in the downlink is supposed to avoid collisions between packets. However, in high traffic regimes, the problem of resource dimensioning is still open and it is the focus of the current paper. There have been attempts for using classical queuing theory methods for dimensioning the system, but they needed to make strong hypotheses on the traffic and system. For instance, [7] 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. In [14], the authors make use of 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. [12] relaxed the Exponential assumption for the service rate and adopted an M/G/1 model with vacations, 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 the 5G NR system can multiplex packets in the spectrum dimension (several servers) and is time-slotted.

With regards to these limitations and the difficulty to find realistic and tractable queuing models for URLLC, we adopt a large deviations approach that is suitable to analyze the tail of the system, corresponding to the URLLC outage region. We make use of two types of simulations for validating the model. First, we compare the model to numerical simulations of the discrete-time Markov process describing the system evolution. And next, we implement the dimensioning framework based on the analytical model in a large scale system level simulator, and observe the URLLC performance in a realistic setting.

1.3 Paper Organization

The remainder of the paper is organized as follows: In Sect. 2, we describe the outage model based on the large deviations theory. Section 3 compares the model with respect to numerical resolution based on a realistic radio distribution. Section 4 applies the proposed dimensioning framework to a large scale system level simulator and quantifies the resulting resource reservation gap. Section 5 concludes the paper.

2 Outage Model

2.1 System and Traffic Model

For developing the analytical model, we consider a 5G cell with U URLLC users and with a 5G-NR like frame, where time/frequency resources are organized into Resource Blocks (RB) and (mini-)slots. The slot is of size T ms and there are R reserved RBs of the total bandwidth, dedicated for URLLC traffic. We consider a sporadic traffic model, i.e. a user is active (generates a packet) during a slot with probability q.

There are I different Modulation and Coding Scheme (MCS), numbered 0 to I, and a packet belongs to a user whose MCS is i with probability p_i. We assume that the MCS distribution in the cell is known. e.g. from field measurements. If a user uses MCS i, each of its packets consumes r_i RBs[1]. Without loss of generality, we suppose that the MCSs are sorted following increasing spectral efficiency, i.e. $r_0 > ... > r_I$.

2.2 Outage Bounds for a Tight Delay Budget (No Waiting)

Let $X_u(t)$ be the number of requested RBs by user $u \in 1, 2, ..., U$ during slot t. $X_u(t)$ are i.i.d. random variable that take the following values:

$$X_u(t) = \begin{cases} 0, & \text{with prob. } 1-q \\ r_i & \text{with prob. } qp_i \end{cases} \tag{1}$$

The total number of resources requested by packets generated in a given slot is then given by:

$$\bar{R}(t) = \sum_{u=1}^{U} X_u(t) \tag{2}$$

[1] If the spectral efficiency of MCS i is equal to e_i (bit/s/Hz), a packet is of size P bit, and one RB spans over b Hz, the amount of consumed RBs is computed by:

$$r_i = \lceil \frac{P}{e_i bT} \rceil,$$

where $\lceil x \rceil$ is the smaller integer larger than x.

The outage occurs when the number of needed resources exceeds the amount of reserved resources. The objective is to ensure that the outage probability is below a small positive value ϵ:

$$Pr(\sum_{u=1}^{U} X_u > R) \leq \epsilon \tag{3}$$

Problem (3) can be solved using Large deviation techniques for which several bounds exist. We start by computing the mean and standard deviation of X_u and $\bar{R}$. For X_u, the mean value is:

$$\mu_0 = E[X_u] = q \sum_{i=1}^{I} p_i r_i \tag{4}$$

and the variance is:

$$\sigma_0^2 = E[X_u^2] - \mu_0^2 = q \sum_{i=1}^{I} p_i r_i^2 - q^2 (\sum_{i=1}^{I} p_i r_i)^2 \tag{5}$$

As for the total consumption of RBs, its mean and variance are $\mu = U\mu_0$ and $\sigma^2 = U\sigma_0^2$, respectively.

Define $x_u = X_u - \mu_0$. The outage constraint (3) can be rewritten as:

$$Pr(\sum_{u=1}^{U} x_u > R - \mu) \leq \epsilon \tag{6}$$

Define now $s = \frac{R-\mu}{\sigma}$, the constraint can be rewritten as:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \epsilon \tag{7}$$

Bienaymé-Chebychev Bound. The well-known Bienaymé-Chebychev bound [21] can be applied. Taking the bound as equal to ϵ, we have:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \frac{1}{s^2}, \tag{8}$$

leading to the required reservation:

$$R_1 = \mu + \frac{\sigma}{\sqrt{\epsilon}} \tag{9}$$

Bernstein Bound. The Bienaymé-Chebychev bound is known to be weak for a sum of random variables. x_i's have the advantage of being independent and bounded, we can apply more tight bounds. Let M be the upper bound of x_i:

$$M = r_0 - q\sum_{i=1}^{I} r_j p_j \tag{10}$$

Bernstein [6] proved that the sum of bounded independent random variables is bounded by:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \exp\left[-\frac{s^2}{2+\frac{2}{3}\frac{M}{\sigma}s}\right] \tag{11}$$

Substituting the bound by the target, this leads to the reservation:

$$R_2 = \mu - \frac{M\ln\epsilon}{3} + \frac{\sigma}{2}\sqrt{\frac{4M^2(\ln\epsilon)^2}{9\sigma^2} - 8\ln\epsilon} \tag{12}$$

Bennet Bounds. Bennet [4] proposed two enhancements on Bernstein's bound, as follows.

First, the bound can be computed as:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq \exp\left[-\frac{s^2}{1+\frac{1}{3}\frac{M}{\sigma}s+\sqrt{1+\frac{2}{3}\frac{M}{\sigma}s}}\right] \tag{13}$$

Leading to the reservation of resources:

$$R_3 = \sigma s_3 + \mu \tag{14}$$

with s_3 solution of the following equation:

$$\frac{s^2}{\ln\epsilon} + 1 + \frac{1}{3}\frac{M}{\sigma}s + \sqrt{1+\frac{2}{3}\frac{M}{\sigma}s} = 0 \tag{15}$$

Bennet [4] also proposed another bound as follows:

$$Pr(\sum_{u=1}^{U} x_u > s\sigma) \leq e^{\frac{s\sigma}{M}}\left(1+s\frac{M}{\sigma}\right)^{-(\frac{s\sigma}{M}+\frac{\sigma^2}{M^2})} \tag{16}$$

Leading to the reservation of resources:

$$R_4 = \sigma s_4 + \mu \tag{17}$$

with s_4 solution of the following equation:

$$e^{\frac{s\sigma}{M}}\left(1+s\frac{M}{\sigma}\right)^{-(\frac{s\sigma}{M}+\frac{\sigma^2}{M^2})} = \epsilon \tag{18}$$

2.3 Model with Waiting

We consider now the case with a looser constraint, i.e. where a packet can stay for $\delta > 1$ slots in the system before its delay budget expires (e.g. 7 slots for a target delay of 1 ms and a slot length of 0.144 ms). We consider the same traffic model as in the previous section.

Outage Probability Formulation. In a given slot, numbered 0, knowing that there are R reserved RBs, the "overflow" of resources, i.e. the amount of RB's that will be needed in the future to serve the backlogged traffic is equal to:

$$B_{(0)} = \left(\sum_{u=1}^{U} X_{(0),u} + B_{(-1)} - R \right)^{+} \tag{19}$$

where $X_{(0),u}$ is the amount of resources required for serving the packet of user u generated at slot 0. $B_{(-1)}$ is the amount of overflow traffic from the previous slot (denoted by slot -1), and $(x)^{+} = \max(x, 0)$. Recursively, for a previous slot $-j$, the overflow is computed by:

$$B_{(-j)} = \left(\sum_{u=1}^{U} X_{(-j),u} + B_{(-j-1)} - R \right)^{+} \tag{20}$$

with $X_{(-j),u}$ the amount of resources required for serving the packet of user u generated at slot $-j$.

The outage probability is computed by the probability that the new packet has to wait for more than δ slots:

$$Pr(B_{(0)} > \delta R) \leq \epsilon \tag{21}$$

Approximate Outage Probability. We consider a system with memory of m slots, i.e. the probability that there are packets waiting from more than m slots is negligible. In this case, we neglect the term $B_{(-m-1)}$ in the overflow. Summing up to the previous m slots, and replacing $\left(\sum_{u=1}^{U} X_{(-j),u} + B_{(-j-1)} - R \right)^{+}$ by $\sum_{u=1}^{U} X_{(-j),u} + B_{(-j-1)} - R$, the outage constraint becomes:

$$Pr(\sum_{j=1}^{m} \sum_{u=1}^{U} X_{(-j),u} > (\delta + m + 1)R) \leq \epsilon \tag{22}$$

This approximation is twofold. First, by neglecting the overflow from slots that are older than m, we suppose that the system is not in overload for a large time. This assumption is reasonable for the URLLC regime. We will see in the numerical applications that a memory of 10 slots gives a good approximation. Second, by removing the $(.)^{+}$ operator from the overflow of Eq. 20, we allow the overflow to be negative as if the whole mR resources were used to serve the

traffic arriving within the previous m slots. We shall test the validity of this approximation in numerical applications.

The delay constraint (22) can then be rewritten as:

$$Pr(\sum_{j=1}^{m}\sum_{u=1}^{U} X_{-(j),u} > (\delta + m + 1)R) \leq \epsilon \tag{23}$$

This constraint compares the sum of $U(m+1)$ independent variables with a threshold; it can be rewritten as:

$$Pr(\sum_{j=1}^{m}\sum_{u=1}^{U} x_{-(j),u} > \hat{\sigma} s) \leq \epsilon \tag{24}$$

with $\hat{\sigma} = \sqrt{U(m+1)}\sigma_0$ and

$$s = \frac{(\delta + m + 1)R - (m+1)U\mu_0}{\hat{\sigma}}, \tag{25}$$

$x_{-(j),u} = X_{-(j),u} - \mu_0$ are centered independent random variables bounded by M computed as in Eq. (10).

We can apply the same bounds of Eqs. (8), (11), (13) and (16) on the system.

3 Numerical Applications

We now compare the analytical bounds with numerical simulations. We developed a simple simulator for the cell scheduler that operates as follows:

- Inputs: the simulator takes as input the traffic profile (number of users, average number of packets per second per user) and the radio conditions. For a realistic setting, we consider a typical MCS distribution issued from a system level simulator, as discussed later. The MCS distribution is illustrated in Fig. 1.
- Traffic generation: the time is divided into slots of size $T = 0.144$ ms and there are R reserved RBs for URLLC. In each slot, each user generates a packet following a Bernoulli law with parameter q, and if a packet is generated, it chooses at random an MCS following the input distribution. Packets are all of equal size (96 bits).
- Scheduler: Packets are served following a First-Come-First-Serve (FCFS) discipline. When a packet is generated, it is put at the end of the queue. A time slot is filled with the packets at the head of queue until all of the R RBs are occupied or the queue is empty. When a packet cannot be scheduled on 1 slot as the remaining resources are not sufficient, it can be scheduled on two consecutive slots.
- Output: For each of the packets, it is counted as an outage if the delay between its generation and its service exceeds a threshold.

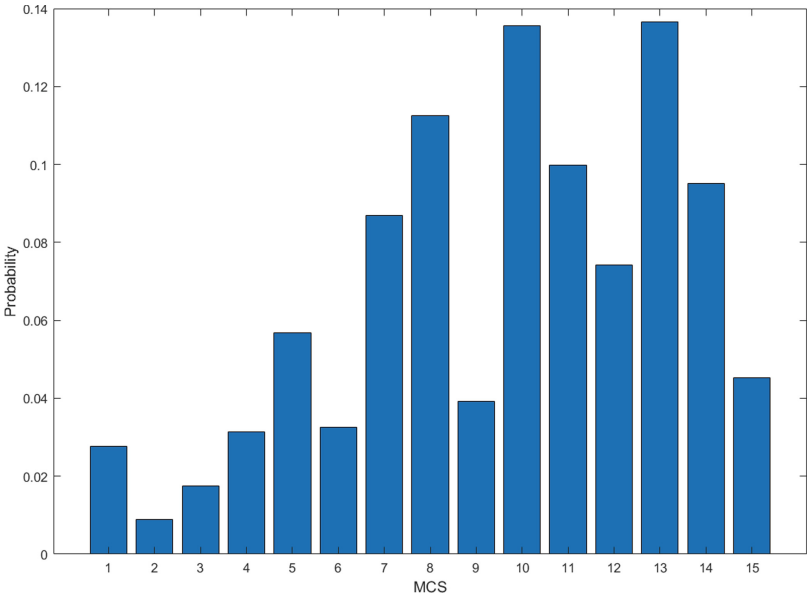

Fig. 1. MCS distribution.

3.1 Model with No Waiting

We start by the case of a very stringent delay budget, where there is no room for waiting. We illustrate in Fig. 2 the outage probability obtained by simulation, and using the bounds of Eqs. (8), (11), (13) and (16). The parameters taken for this simulation are: $U = 20$, $q = 0.072$. First, all the bounds give an outage probability that is larger than the simulation. Second, it can be observed that the second bound of Bennet (Eq. (16)) gives the closest bound to the simulation as it is adapted to a sum of independent variables. Third, the simulation stops for an outage rate that is below 10^{-7} as the outage event becomes too rare to be simulated.

Based on these results, we investigate the amount of over-dimensioning required when using the analytical bounds, compared with the simulation. For a target outage probability of 10^{-5}, the required reservation is of $R = 85$ RBs, based on simulations, while the Bennet bound (16) required 115 RBs. The Chebychev bound is so loose that the reservation requirement exceeds 500 RBs.

3.2 Model with Queuing Delay

We now move to a more common use case where there is room for multiple slots for queuing within the delay budget. Here we take the threshold on the waiting delay equal to 1 ms. Note that the threshold depends on the service requirements

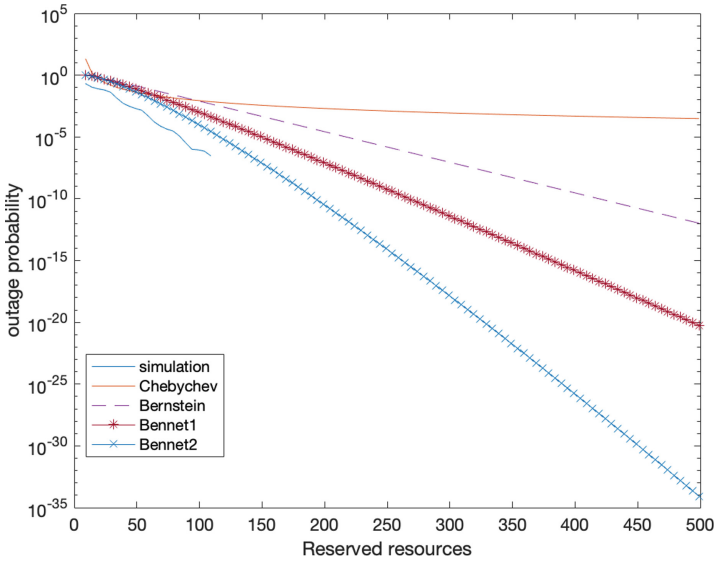

Fig. 2. Outage probability with no waiting.

and the radio settings, and the waiting delay threshold has to be computed as the difference between the service delay budget and the other non compressible delays (alignment, propagation, decoding, back-haul etc.).

We consider the same MCS distribution as previously. We first start by studying the impact of the approximation of finite memory m on the bound, considering the most tight bound of Bennet (16). We can observe in Fig. 3 that the amount of required reservation increases with m, and stabilizes starting from $m > 9$. We consider in the following $m = 10$.

We compare in Fig. 4 the analytical bounds with simulation results. We see that the difference between the Bennet2 bound achieves the closest bound to the simulation and that the gap is reduced compared to the no-waiting case.

In order to compare with queuing models used in the state of the art, we implement the M/M/c/K model proposed in [14], where arrivals are Poisson, service is approximated as exponential, c is the number of servers, and K is the maximum number of packets the system can hold. K is computed in [14] 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 for a fair comparison). However, the number of servers is not known as it is computed as the number of packets that can be served in parallel, while this number depends, for a fixed R, on the MCS. [14] considered the worst case, i.e. when all users are at the cell edge and computed c as the ratio between R and the number of resources occupied by a packet generated at cell edge (MCS 1). As this is too pessimistic, we consider

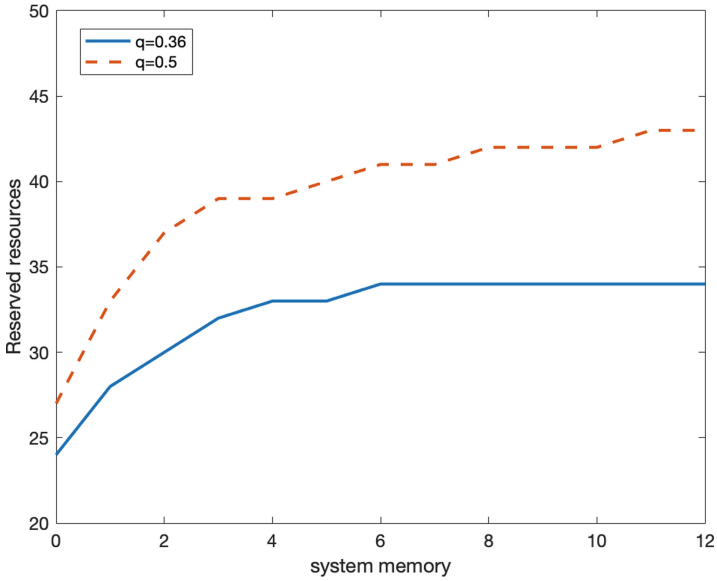

Fig. 3. Required resource reservation for a target reliability (1 ms budget).

the MCS used by the worst 10% of users (90% percentile), that corresponds to MCS 5. Figure 4 shows that the bound is too loose (very large outage). One can try to consider the average resource consumption instead of the worst case or the percentile ($c = \lceil \frac{R}{\sum_{i=1}^{I} r_i} \rceil$), but Fig. 4 shows that this method cannot be used for URLLC resource provisioning, as it sometimes largely underestimates the outage (the step-like behaviour comes from the necessity top have an integer number of servers in the M/M/c/K model).

The model can also be used for resource dimensioning, i.e. for computing the resource reservation for ensuring the target performance. Figure 5 compares the amount of reserved resources for the analytical bound (16) with the numerical simulations and shows that the bound is very tight.

4 Resource Dimensioning Framework

Having validated our analytical model based on simple numerical simulations, we now propose a resource dimensioning framework and test it on a large system level simulator.

4.1 Architecture

Figure 6 illustrates the architecture for implementing the proposed scheme. We propose that the resource allocation module for the URLLC slice be implemented within the Network Slice Subnet Management Function (NSSMF). Within the

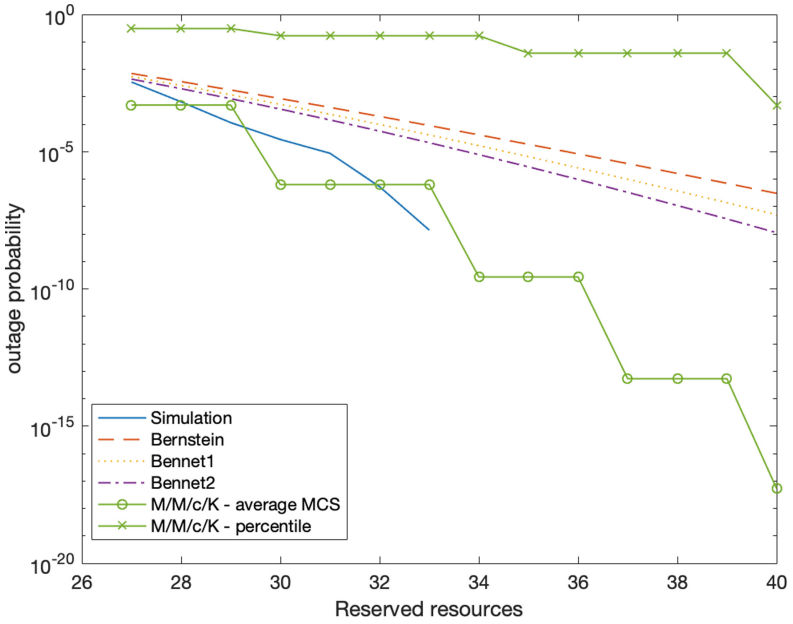

Fig. 4. Outage probability for the delayed case ($U = 20$, $q = 0.36$).

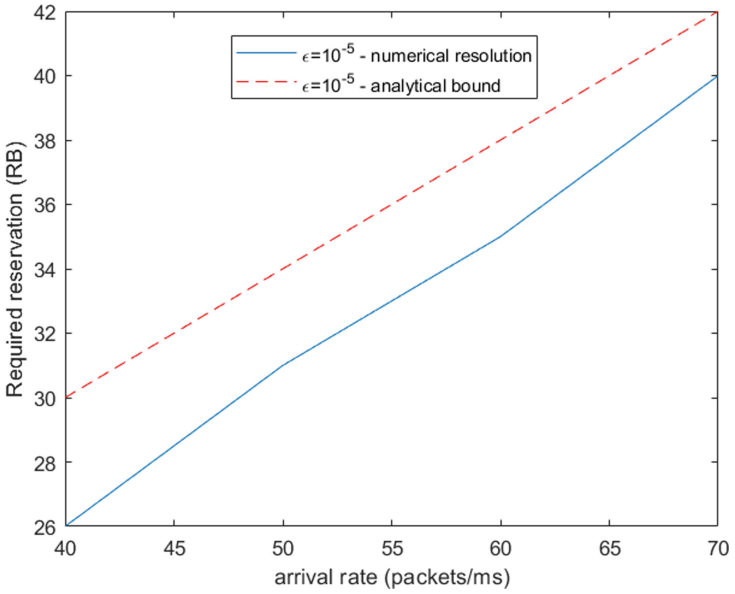

Fig. 5. Required resource reservation for a target reliability (1 ms budget).

NSSMF, two modules allow the dynamic management of the slices. First, an MCS distribution module allows building a per-gNodeB MCS distribution. Second, we use this distribution as input for the resource dimensioning module that takes as input the traffic (number of URLLC users, number of packets/user/s) and computes the needed amount of resources to be reserved for the URLLC slice in each of the gNodeBs, using the analytical model (Eq. (16)). The system applies the new configuration, dynamically changing depending on the NSSMF updates (traffic and radio conditions change).

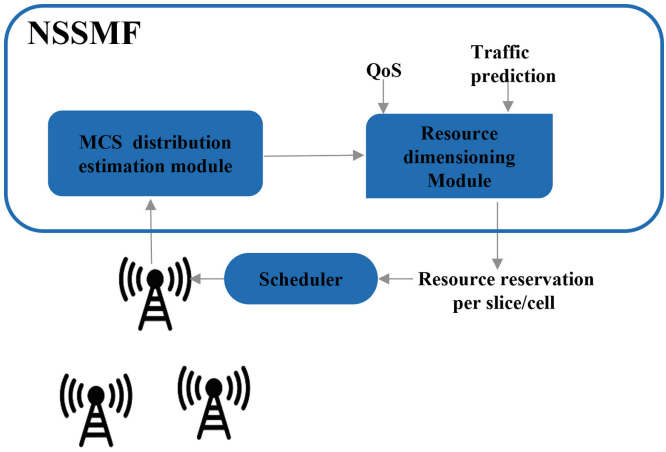

Fig. 6. Proposed architecture for the dimensioning framework.

4.2 System Level Simulation Results

Figure 7 illustrates the network created by the simulator, showing the positions of the gNodeBs and some URLLC UEs.

We perform three types of simulations. The simulation and configuration parameters are presented in Table 1.

In the first simulation, for each traffic intensity, we perform a series of simulations, changing the amount of reserved resources in each cell until reaching the target of 10^{-5} outage. This gives the system simulation resource reservation, which is not applicable in practice as it requires a large number of trials on the up and running network. Second, we apply our dimensioning framework where we extract the radio conditions distribution from the cells, and then apply the proposed analytical model to obtain the required reservation. Finally we simulate the M/M/c/K model with 90% percentile MCS. The second set of simulations is based on this analytical reservation (Eq. (16)) to verify that the outage is far below the target. Figure 8 compares the reservation obtained by extensive simulations with the analytical model and the M/M/c/K model [14] with a cell edge MCS (worst 10% of users). We first observe that the M/M/c/K model leads to

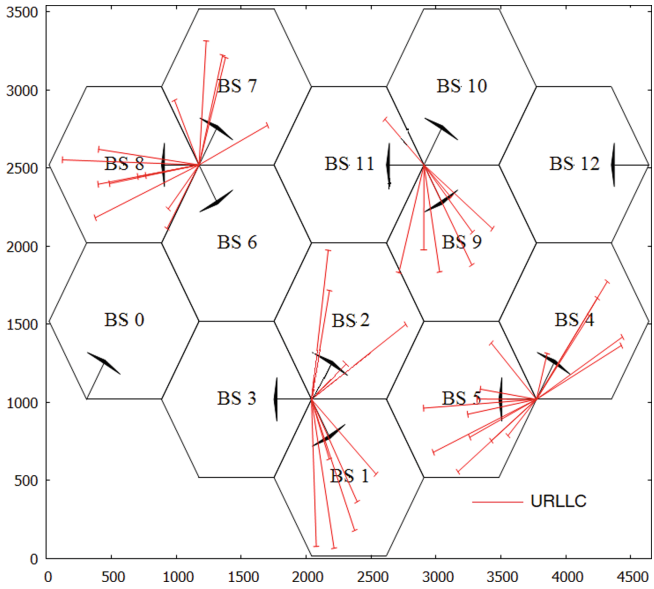

Fig. 7. Urban network with 13 gNodeBs.

Table 1. System parameters.

Parameters	URLLC
Environment	3GPP Urban Macro (UMa)
Number of gNodeBs	13
Bandwidth	20 Mhz
Sub-Carrier-Spacing (SCS)	15 Khz
Number of RBs	106
TTI size (ms)	0.143
Traffic model	Bernoulli
Packet size	96 bits
Speed	Static

a very large over-dimensioning. As of our proposed bound, we observe an average over-dimensioning ratio of 15% compared to the system simulator, which is acceptable for guaranteeing URLLC reliability, knowing that the bound is computed based only on the knowledge of the average traffic intensity and radio conditions.

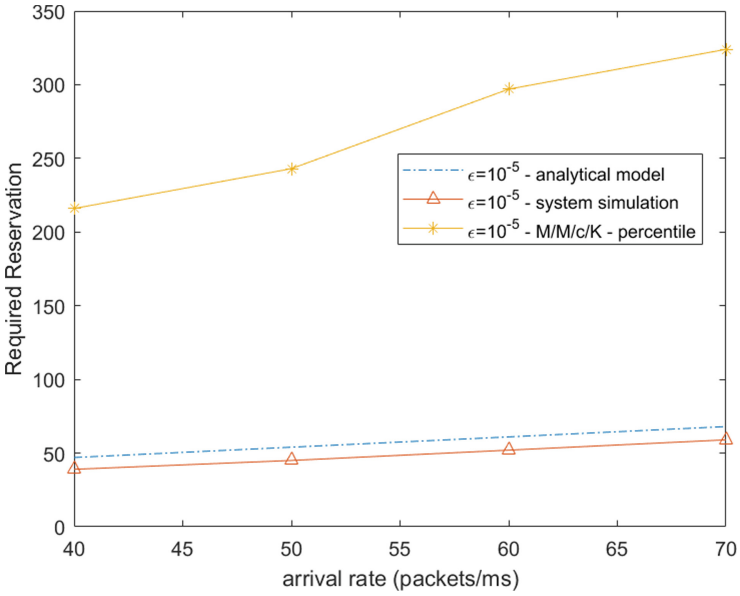

Fig. 8. System simulations versus analytical model.

5 Conclusion

In this paper, we developed a performance evaluation framework for URLLC traffic in 5G networks based on large deviation bounds. We consider the queuing delay and derive the outage probability bound, i.e. the probability that the delay exceeds a given target. We first compared the analytical model with a numerical simulation of the scheduler and showed that the proposed bound is tight. We then proposed a framework for resource dimensioning, that combines the analytical model with measurements of radio conditions issued from the network. We tested the proposed framework on a large scale system level simulator and showed that the URLLC targets are achieved with an acceptable over-dimensioning cost and a low management overhead.

References

1. 3GPP, TS 23.501: System Architecture for the 5G System (2017). Version 15.0.0 Release 15
2. 3GPP, TR 38.912: 5G; Study on New Radio (NR) access technology (2018). Version 15.0.0 Release 15
3. Alsenwi, M., Tran, N., Bennis, M., Pandey, S., Bairagi, A., Hong, C.S.: Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: a deep reinforcement learning based approach. IEEE Trans. Wirel. Commun. **PP**, 1 (2021). https://doi.org/10.1109/TWC.2021.3060514

4. Bennett, G.: Probability inequalities for the sum of independent random variables. J. Am. Stat. Assoc. **57**(297), 33–45 (1962)
5. Bennis, M., Debbah, M., Poor, H.V.: Ultra reliable and low-latency wireless communication: tail, risk, and scale. Proc. IEEE **106**(10), 1834–1853 (2018)
6. Bernštein, S.: Theory of probability. Moscow. MR0169758 (1927)
7. Chagdali, A., Elayoubi, S.E., Masucci, A.M., Simonian, A.: Performance of URLLC traffic scheduling policies with redundancy. In: 2020 32nd International Teletraffic Congress (ITC 32), pp. 55–63. IEEE (2020)
8. Chen, Y., Cheng, L., Wang, L.: Prioritized resource reservation for reducing random access delay in 5G URLLC. In: 2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), pp. 1–5 (2017). https://doi.org/10.1109/PIMRC.2017.8292695
9. Elayoubi, S.E., Brown, P., Deghel, M., Galindo-Serrano, A.: Radio resource allocation and retransmission schemes for URLLC over 5G networks. IEEE JSAC **37**(4), 896–904 (2019). https://doi.org/10.1109/JSAC.2019.2898783
10. Fountoulakis, E., Pappas, N., Liao, Q., Suryaprakash, V., Yuan, D.: An examination of the benefits of scalable TTI for heterogeneous traffic management in 5G networks. In: 2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), pp. 1–6 (2017). https://doi.org/10.23919/WIOPT.2017.7959871
11. Han, Y., Elayoubi, S.E., Galindo-Serrano, A., Varma, V.S., Messai, M.: Periodic radio resource allocation to meet latency and reliability requirements in 5G networks. In: 2018 IEEE 87th Vehicular Technology Conference (VTC Spring), pp. 1–6. IEEE (2018)
12. Jang, H., Kim, J., Yoo, W., Chung, J.M.: URLLC mode optimal resource allocation to support HARQ in 5G wireless networks. IEEE Access **8**, 126797–126804 (2020)
13. Korrai, P., Lagunas, E., Sharma, S.K., Chatzinotas, S., Bandi, A., Ottersten, B.: A RAN resource slicing mechanism for multiplexing of eMBB and URLLC services in OFDMA based 5G wireless networks. IEEE Access **8**, 45674–45688 (2020). https://doi.org/10.1109/ACCESS.2020.2977773
14. Li, C.P., Jiang, J., Chen, W., Ji, T., Smee, J.: 5G ultra-reliable and low-latency systems design. In: 2017 European Conference on Networks and Communications (EuCNC), pp. 1–5 (2017). https://doi.org/10.1109/EuCNC.2017.7980747
15. Li, Z., Uusitalo, M.A., Shariatmadari, H., Singh, B.: 5G URLLC: design challenges and system concepts. In: 2018 15th International Symposium on Wireless Communication Systems (ISWCS), pp. 1–6. IEEE (2018)
16. Morcos, M., Mhedhbi, M., Galindo-Serrano, A., Eddine Elayoubi, S.: Optimal resource preemption for aperiodic URLLC traffic in 5G networks. In: 2020 IEEE 31st Annual International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 1–6 (2020). https://doi.org/10.1109/PIMRC48278.2020.9217111
17. Pedersen, K.I., Pocovi, G., Steiner, J., Khosravirad, S.R.: Punctured scheduling for critical low latency data on a shared channel with mobile broadband. In: 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), pp. 1–6 (2017). https://doi.org/10.1109/VTCFall.2017.8287951
18. Pedersen, K.I., Berardinelli, G., Frederiksen, F., Mogensen, P., Szufarska, A.: A flexible 5G frame structure design for frequency-division duplex cases. IEEE Commun. Mag. **54**(3), 53–59 (2016). https://doi.org/10.1109/MCOM.2016.7432148
19. Sachs, J., Wikstrom, G., Dudda, T., Baldemair, R., Kittichokechai, K.: 5G radio network design for ultra-reliable low-latency communication. IEEE Netw. **32**(2), 24–31 (2018). https://doi.org/10.1109/MNET.2018.1700232

20. Singh, B., Tirkkonen, O., Li, Z., Uusitalo, M.A.: Contention-based access for ultra-reliable low latency uplink transmissions. IEEE Wirel. Commun. Lett. **7**(2), 182–185 (2018)
21. Stroock, D.W.: An Introduction to the Theory of Large Deviations. Springer, Heidelberg (2012)