Radio resource allocation for extreme URLLC under partial knowledge of arrival distributions

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

Radio resource allocation for extreme URLLC under partial knowledge of arrival distributions

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, Institut Mines Telecom, SAMOVAR, Palaiseau, France

***Abstract*—We address radio resource allocation for the transport of extreme Ultra Reliable Low Latency and Reliability (URLLC) traffic. One illustrative use case is factory automation using 6G networks. In this context, extreme URLLC has very stringent Quality of Service (QoS) requirements: 0.1ms for the delay and 10^{-7} for the reliability. Reliability can be even higher for other use cases. We model QoS in terms of outage probability, that is the likelihood of failing to serve at least one packet due to insufficient resources, and derive the minimal resource reservation that would meet such requirement. We formulate the problem as chance-constrained optimization and solve it assuming partial knowledge of arriving traffic distribution. We treat the case where traffic is described through its mean and variance, and make use of three approaches to find the optimal solution: distributionally robust using worst-case value at risk approach, distribution-based approximation and bounds from large deviation theory. We also solve the optimization problem using a data driven approach and propose a sliding window mechanism to perform it online. We compare the performance of the aforementioned approaches numerically and show the effectiveness of the data driven approach, accounting for user radio condition heterogeneity and thus different Modulation and Coding Schemes.**

I. Introduction

The emergence of new applications necessitates even stricter reliability and latency requirements than those set in 5G Ultra Reliable Low Latency and Reliability (URLLC), 1ms for delay and 10^{-5} reliability. Factory automation for instance is one use case where so-called extreme URLLC can be deployed [1], with QoS requirements as strict as 0.1ms for the delay and 10^{-7}, or even 10^{-9} reliability, depending on the application, be it autonomous vehicles, remote surgery, robot control, or other mission-critical use cases.

In 5G URLLC, short time intervals or mini-slots, repetition, and retransmission are vital for achieving stringent QoS requirements. In this context, the ability to queue packets in a buffer, awaiting service in current or upcoming mini-slots, enables some flexibility in resource reservation and allow dynamic adjustments. In serving extreme URLLC however, buffering is not an option, nor are retransmissions. Packets should be served in the mini-slot in which they arrive or are discarded.

We focus in this work on bufferless setting and aim to determine the minimal required resources in order to meet extreme URLLC QoS constraints, quantified in terms of so-called outage probability, that is the probability of serving the incoming extreme URLLC packets in the current mini-slot. We cast such problem as a Chance-Constrained Optimization (CCO) one and solve it without making assumption on the distribution of the traffic. We study the cases where only the mean and variance are known and apply two approaches to solve the problem: a first one based on worst case value at risk approach and a second one based on assumptions on the traffic distribution. For the case where the number of users is known, we make use of large deviation approach and derive several bounds for the outage probability. We also study a data driven approach and propose a sliding window mechanism in order to dynamically adjust the minimal number of required resources.

There has been only few works on resource allocation for the transport of extreme URLLC. The authors in [2] studied various aspects of extreme URLLC in a massive Multiple Input Multiple Output (MIMO) network, taking into account effective capacity, energy efficiency, extreme delay constraint and shadowing. They established an upper bound for delay using the Chernoff bound. However, their analysis assumes complete knowledge of packet arrival distribution. In [3], a novel extreme URLLC service framework has been introduced for the use case of Metaverse. To assess the users Quality of Experience (QoE), the authors introduced a metric termed "Meta-immersion" which encompasses both objective aspects of QoS and subjective ones, such as feelings of users; the latency budget however has been allowed to be greater than 0.1 ms. The authors in [4] proposed a probabilistic grant-free scheduling model, tailored for handling extreme URLLC packets. In their approach, they addressed both repetition and retransmission strategies. The latter however can be inappropriate in this setup due to the stringent latency requirement which imposes transmission of packets in the mini-slot corresponding to their arrival. Furthermore, the authors make the assumption that packets arriving in the system adhere to a binomial distribution. In [5], QoS has been modeled via latency and age of information metrics. The system incorporates however a queuing mechanism, enabling packets to wait in a buffer for the next mini-slot. The authos also assume Poisson distribution for the incoming traffic.

As of the use of bounds, in [6] for instance the authors studied the trade-off between reliability and latency in URLLC systems that employ Finite Block Length (FBL) coding, with bursty traffic. The utilized large deviation theory to derive probabilities related to length and delay violations in the FBL regime, considering URLLC QoS constraints. Additionally, a normal approximation has been used to estimate the service

rate of the wireless transmission system under consideration. In [7], the focus was on evaluating the performance of URLLC services, specifically in terms of latency, utilizing large deviation bounds. The study made use of Bernstein and Bennet bounds to establish an upper bound on the outage probability. This bound has been utilized to assess resource dimensioning. In their validation however, the authors assume packets arrive to the system following Binomial distribution.

The remainder of this paper is organized as follows. In section II, we show the system model and problem formulation and solve it using the worst case value at risk, distribution approximation and using the aforementioned bounds. Section III contains our work pertaining to the data driven approach. We show our numerical results in section IV. Section V eventually concludes the paper.

II. Problem formulation and analysis

We consider a 5G/6G cell where resources are organized into Resource Blocs (RBs) and mini-slots of size $T = 0.1$ms. We assume packets are all of the same small size, for instance 96 bits. Let R be the number of RBs reserved for extreme URLLC per mini-slot. In each mini-slot, packets are generated following some stochastic process.

We define the outage probability as follows:

$$P(\xi > R) \tag{1}$$

where ξ is the number of RBs required to serve the upcoming packets to the system in a given mini-slot. This outage probability metric contains both delay and reliability constraints: as the length of the mini-slot is 0,1ms, delay is satisfied if the packet is served in the same mini-slot where it arrives to the system. As of reliability, it is obtained by setting a bound ϵ on this probability, for instance 10^{-7}.

The aim is to minimize the number of RBs allocated for the extreme URLLC while satisfying the outage probability constraint. The optimal R is the solution to the following CCO problem:

$$\begin{aligned} R^* &= \arg\min R \\ \text{s.t.} &P[\xi > R] \leq \epsilon. \end{aligned} \tag{2}$$

Note that the variable ξ combines several parameters, including the system configuration (transmission power, numerology), the number of users and their radio conditions (their positions with respect to the base station and the interference they receive) and the service characteristics (e.g. packet size and packet arrival process). We will show in section IV-B how to combine these parameters to obtain ξ. For the ease of reading, we will focus now on the mathematical framework for deriving the optimal resource allocation, knowing some information about ξ, e.g. its moments, its distribution or samples of it.

A. Distributionally robust allocation using Value-at-Risk

Let us assume that only the first two moments of ξ are known: mean μ and variance V. We make use of the notion of worst-case Value at Risk (VaR) introduced in [8] to tackle problems with rare events, in order to solve problem (2) and obtain a closed-form solution. We denote by $\mathcal{P}$ the set of all distributions with mean μ and variance V. The VaR problem in the worst-case scenario is as follows:

$$\min R \qquad \text{subject to} \quad \sup_{P \in \mathcal{P}} P[\xi > R] \leq \epsilon \tag{3}$$

where the **sup** is taken on all probability distributions in $\mathcal{P}$. In [9], the authors provide the following equivalence relation:

$$\sup P[\xi > R] \leq \epsilon \iff k(\epsilon)\sqrt{V} - \mu \leq R, \tag{4}$$

where $k(\epsilon) = \sqrt{\frac{1-\epsilon}{\epsilon}}$ is called the risk factor. This leads for reserving a number of RBs, which we term as R_1:

$$R_1 = k(\epsilon)\sqrt{V} - \mu. \tag{5}$$

B. Approximating using known arrival distributions

We now turn to the case where we approximate ξ by known distributions.

1) Poisson Distribution: Poisson distribution is commonly used in the literature [10], [11], [12], and [13]. Building upon this and leveraging the first two moments of ξ, we aim to approximate its distribution by a Poisson distribution with mean arrival rate λ. Let $F_\xi(\cdot)$ denote the cumulative distribution function (CDF) ξ. The resource reservation in this case, denoted as R_2, is as follows:

$$R_2 = \min_x \{x, F_\xi(x) \geq 1 - \epsilon\}. \tag{6}$$

This Poisson-based approximation overlooks the variance term, potentially leading to conservative or non-feasible solutions. In order to address this limitation, we consider next the Gamma distribution, allowing for integration of both moments in solving problem (2). Under certain conditions, the Gamma distribution converges to Poisson distribution, which makes it adhere to this common assumption used in the literature.

2) Gamma Distribution: Assuming that the random variable ξ follows the Gamma distribution with shape α and scale β ($\xi \sim \text{Gam}(\alpha, \beta)$), we have that:

$$E[\xi] = \mu = \alpha\beta \quad \text{and} \quad Var(\xi) = V = \alpha\beta^2, \tag{7}$$

Consequently, the parameters of the Gamma distribution can be expressed as:

$$\alpha = \frac{\mu^2}{V} \quad \text{and} \quad \beta = \frac{V}{\mu}, \tag{8}$$

As before, we use the CDF to calculate the number of required resources using this approximation, which we term R_3, using equation (6).

3) Gaussian Distribution: The most used distribution to approximate random variables in CCO problems is the Gaussian distribution. Moreover, it is the limiting distribution of both Poisson and Gamma distributions [14]. Thus assuming that ξ follows a Gaussian distribution with a mean μ and variance V and referring to [15], the probability constraint in (2) can be expressed as follows:

$$P[\xi > R] \leq \epsilon \iff \mu + \phi^{-1}(1-\epsilon)\sqrt{V} \leq R \tag{9}$$

where $\phi^{-1}(\cdot)$ is the inverse of ϕ, the standard normal CDF. Consequently, the optimal resource reservation in this scenario, R_4, can be determined in a closed form:

$$R_4 = \mu + F^{-1}(1-\epsilon)\sqrt{V}. \tag{10}$$

C. Large deviation bound assuming known number of users

We now investigate bounds on the probability constraint, using namely employing the Bernstein and Bennet bounds [7]. We assume that the number of users u is known and define $X_i(t)$ as a random variable indicating the number of RBs required for serving user i at time t, where $X_i(t) = \{0, 1, \ldots, k\}$ ($X_i = 0$ in case user is non-active). We denote by $\xi(t) = \sum_i X_i$ the total number of RBs required to serve all packets arriving to the system at time t. It is obvious that the CCO problem (2) can be expressed as follows:

$$\begin{aligned} \min \quad & R \\ \text{s.t.} \quad & P[\sum_i X_i > R] \leq \epsilon. \end{aligned} \tag{11}$$

We assume that all X_is have identical and independent distributions (i.i.d.) with known support, we have that the expectation and variance of X_i are, respectively, given by $\mu_i = E[X_i] = \frac{\mu}{u}$ and $V_i = Var(X_i) = \frac{V}{u}$ where μ and V are the mean and variance of ξ, respectively. Let $x_i = X_i - \mu_i$. Consequently, the probability constraint can be expressed as:

$$P[\sum_i x_i > R - \mu] \leq \epsilon \iff P[\sum_i x_i > s\sigma] \leq \epsilon \tag{12}$$

where $\sigma = \sqrt{V}$ is the standard deviation of ξ and $s = \frac{R-\mu}{\sigma}$. Following the methodology presented in [7], we need to establish an upper bound for x_i, denoted as M_i, and then define M as the maximum of M_i. Observe that $M = M_i$ since all x_is are i.i.d., leading to $M = k - \frac{\mu}{u}$ where k is the maximum number of resource blocks for serving a single packet.

1) Bernstein bound: Assuming that the number of users u and the upper bound M are known, the Bernstein bound is given as follows [7]:

$$P[\sum_{i=1}^{u} x_i > s\sigma] \leq exp\Big[-\frac{s^2}{2 + \frac{2M}{3\sigma}s}\Big] \tag{13}$$

By substituting the bound by ϵ, we get the following reservation:

$$R_5 = \mu - \frac{Mln\epsilon}{3} + \frac{\sigma}{2}\sqrt{\frac{4M^2(ln\epsilon)^2}{9\sigma^2} - 8ln\epsilon}. \tag{14}$$

2) Bennet bound: For Bennet bounds, two bounds are given as enhancements to Bernstein's bound in [7]. The first bound which we call Bennet 1 can be computed as:

$$P[\sum_i x_i > s\sigma] \leq exp\Big[-\frac{s^2}{1 + \frac{Ms}{3\sigma} + \sqrt{1 + \frac{2Ms}{3\sigma}}}\Big], \tag{15}$$

which leads to the following reservation of resources:

$$R_6 = \sigma s_1 + \mu, \tag{16}$$

where s_1 is the solution of the following equation:

$$\frac{s^2}{ln\epsilon} + 1 + \frac{Ms}{3\sigma} + \sqrt{1 + \frac{2Ms}{3\sigma}} = 0. \tag{17}$$

The second bound which we call Bennet 2 is given as follows:

$$P[\sum_i x_i > s\sigma] \leq e^{\frac{s\sigma}{M}}\Big(1 + s\frac{M}{\sigma}\Big)^{-(\frac{s\sigma}{M} + \frac{\sigma^2}{M^2})}, \tag{18}$$

This leads to the following reservation:

$$R_7 = \sigma s_2 + \mu \tag{19}$$

with s_2 is the solution of the following equation:

$$e^{\frac{s\sigma}{M}}\Big(1 + s\frac{M}{\sigma}\Big)^{-(\frac{s\sigma}{M} + \frac{\sigma^2}{M^2})} = \epsilon \tag{20}$$

III. Data-Driven Method

We now assume that we are given an arrival dataset $\{\xi_i\}_{i=1}^N$ where x_i denotes an arrival instant and N is the size of the dataset. Our CCO problem (2) can be approximated by [16]

$$\begin{aligned} \min \quad & R \\ \text{s.t.} \quad & f(R, \xi_1) \leq 0, \ldots f(R, \xi_N) \leq 0, \end{aligned} \tag{21}$$

where $f(R, \xi_i) = \xi_i - R$, for $i \in \{1, ..., N\}$.

The optimal solution of (21) is feasible for all the N realizations. Solving this problem requires the convexity of the constraints $f(R, \xi)$. However, the solution of (21) is a random variable as it depends on the random samples. Thus given a solution R to the CCO problem, we should first check its feasibility, i.e., whether it satisfies the constraint, in a probabilistic way, i.e., the violation probability of the outage probability, i.e., $P(P(\xi > R) > \epsilon)$, be smaller than a given threshold which we term β.

Assumption 1. *Let Ξ denote the support of the random variable ξ, the distribution $\xi \sim \Xi$ exists and is fixed.*

Assumption 2. *Function $f(R, \xi)$ is convex in R for every instance of ξ, and the deterministic constraint defines a convex set.*

Assumption 3. *The samples ξ^i in the dataset $\{\xi^i\}_{i=1}^N$ are independent and identically distributed.*

Assumption 4. *Every scenario problem, i.e., one realization of (21) for one instance of ξ_i, is feasible, and its feasibility region has a non-empty interior. Moreover, the optimal solution exists and is unique.*

There is a trade-off between feasibility versus the degree of conservatism of the solution: as the number of realizations N gets larger, the solution will tend to be more conservative as the realizations might contain larger number of incoming packets per mini-slot. For smaller values of N however, we might not be able to capture the rare events and hence result in a solution that would violate the above mentioned probabilistic constraint. We hence need to find a value for N that would balance these two aspects.

A. A-priori feasibility guarantees

Using Assumptions 1-4, for a non-degenerate fully supported problem, it is shown in [[16], Theorem 5] that:

$$P\Big(V(x_N^*) > \epsilon\Big) = \sum_{i=0}^{n-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i}, \tag{22}$$

where x_N^* is the optimal solution of problem (21) and $V(x_N^*)$ is the violation probability of the constraints. Moreover, in [[16], Corollary 1] it is stated that for a given ϵ and a confidence parameter $\beta \in (0,1)$, N is such that:

$$\sum_{i=0}^{n-1} \binom{N}{i} \epsilon^i (1-\epsilon)^{N-i} \leq \beta, \tag{23}$$

where $n = 1$ is the dimension of the decision variable R. We thus have that:

$$P\Big(V(x_N^*) \leq \epsilon\Big) \geq 1 - \beta \tag{24}$$

One can see that our scenario problem (21) is fully supported as the support scenarios is of cardinal 1 ($|S| = 1$) which is the largest ξ_i in the dataset, moreover removing some realizations that are different from the support scenario in hand will not change the solution, thus it is a non-degenerate problem. This means that all the feasibility guarantees shown before apply to problem (2). Given ϵ and β, N is calculated using equation (23).

B. Dynamic policy update

For a dynamic resource reservation, we propose the sliding window approach, to be used in online settings, to address problem (2). To do so, we need two parameters to initiate the solution process: the violation probability bound ϵ, and the targeted confidence parameter β to calculate N (the number of realizations in the dataset). The sliding window approach involves maintaining a dataset of size N, which is continuously updated in an online fashion. For each time slot t, the approach entails removing the realizations observed at time $t-N$ and incorporating the realizations observed at time t, thus the available dataset at time t, denoted by D_t, is:

$$D_t = \{\xi_{t-N+1}, \xi_{t-N+1}, \ldots, \xi_t\}, \tag{25}$$

Following this construction, we obtain a dynamic resource reservation depending on our up-to-date knowledge (new dataset), and thus at each time t the number of resources reserved for serving the packets, denoted by $R(t)$, is:

$$R(t) = \max \ \{\xi_i \ , \ \xi_i \in D_{t-1}\}, \tag{26}$$

This approach ensures adaptability and effectiveness in managing time varying traffic. Moreover, following this setup, one can observe that with an increase in the arrival rate, the system will rapidly adjust the number of resources by increasing them as soon as the increase takes place. In the case of a decreasing arrival rate of packets, the system will wait for some time (N mini-slots) before adapting and decreasing the resource reservation, thus preventing any spurious event.

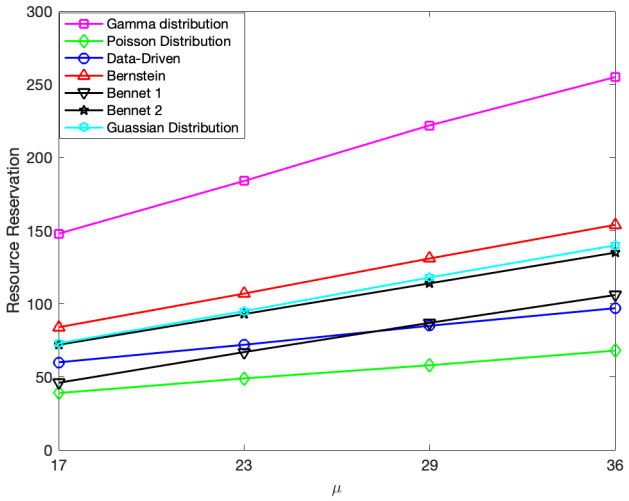

Fig. 1: Comparison of resource dimensioning methods ($\epsilon = 10^{-7}$ and $\beta = 0.03$).

IV. Numerical Experiments

In our experiments, we consider the case of factory automation deploying extreme URLLC with 0,1ms delay and 10^{-7} reliability constraints. We assume that we have $u = 200$ machines. We construct a simulator with packet arrivals consisting of a mixture of Poisson, Binomial, or Gaussian distribution at each time, so as to obtain a rather general arrival distribution with rate within the range $0 \leq \xi \leq u$. To compute the empirical mean and variance of the arrivals, and to generate the dataset for the data-driven method with a packet loss probability smaller than $\epsilon = 10^{-7}$ and a confidence interval $\beta = 0.03$, we run the simulator for $N = 3 \times 10^7$ iterations.

In a real setting, a Radio Access Network (RAN) management entity, possibly situated within the Network Slice Subnet Management Function (NSSMF) [17], is responsible for collecting traffic and radio condition statistics at each mini-slot (0.1 ms) that will be used for reserving resources online using the sliding window approach.

A. Homogenous Modulation and Coding Scheme

In the following, we assume that all packets use the same Modulation and Coding Scheme (MCS) (homogenous MCS), and thus each packet consumes the same number of resource blocks (one RB) to be served. We compare in figure 1 resource dimensioning using the Bernstein bound, Bennett bounds, Data-Driven method, as well as Poisson, Gaussian, and Gamma approximations, for increasing value of $E[\xi] =$. We did not plot the curve corresponding to worst case VaR approach as it yields a too conservative result. We plot in figure 2 the corresponding outage probability calculated using the optimal reservations shown in the previous figure for each approach, using the simulator with 10^8 iterations. In terms of resource reservations, we observe that all curves are upper

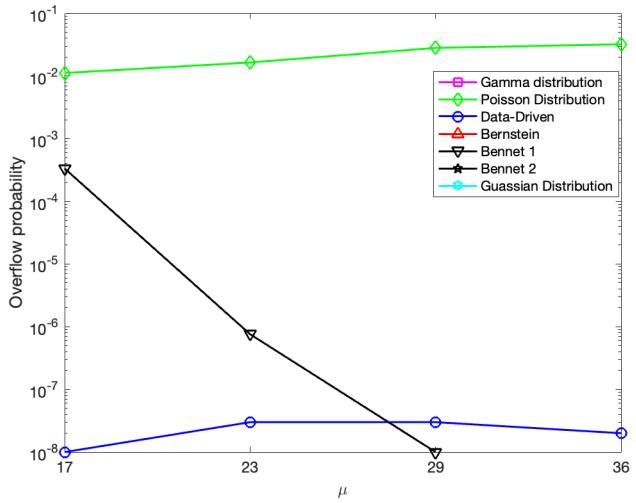

Fig. 2: Comparison of outage probability using the different methods ($\epsilon = 10^{-7}$ and $\beta = 0.03$).

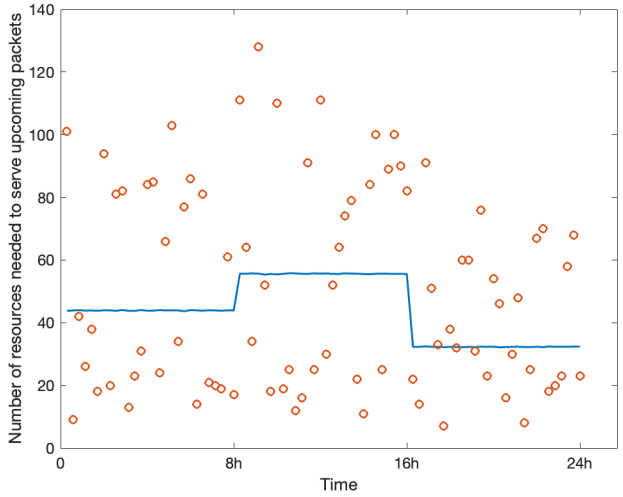

Fig. 3: Distribution of packet arrivals over 24 hours.

bounded by Gamma distribution (owing to the large variance of ξ) and lower bounded by Poisson (only approach relying on mean of ξ only). Poisson however does not reach the target outage probability (10^{-7}) whereas the latter is 0 for Gamma distribution. The optimal resource reservations for Bernstein and Bennet bounds are close to the Gaussian distribution one since $\epsilon \times M \ll Var(\xi)$. Bennet 1 is the closest to the data-driven approach, however it does not always achieve the target outage probability. The other bounds achieve zero outage probability. The data driven approach is robust against mean and variance, and yields the smallest optimal reservation of resources while meeting the target outage probability.

B. Heterogenous Modulation and Coding Scheme

We now delve into the practical scenario where users have different radio conditions, that translate towards heterogeneous MCS, and consequently distinct requirements on the number of RBs per packet. Recall that $X_i(t)$ is a random variable denoting the number of RBs required to serve a packet generated by machine i, and the sum of all X_i's gives the variable ξ. We assume that the support of $X_i(t)$ is $\{0, 1, \ldots, k\}$, i.e. when there is a generated packet ($X_i(t) > 0$, it is encoded using one of the k possible MCS.

In order to be able to apply our model, we need the distribution of the amount of requested RBs $F_\xi(\cdot)$, or a series of observed samples from the field. In order to obtain these information, we build an ad hoc simulator for a factory with machines connected to a base station. The positions of machines in the factory are fixed, with some obstacles that create shadowing, so that the average path loss for users does not change over time. However, the instantaneous path loss fluctuates due to fast fading. The path loss for user i at time t is thus expressed by the following expression:

$$p_i(t) = \frac{a.d_i^b}{G.L.S_i.F_i(t)}, \tag{27}$$

where d_i is the distance to the base station, a and b are carefully chosen coefficients to accurately represent the behavior of the communication medium as well as the antenna gain (G) and the equipment imperfections (L), S_i and $F_i(t)$ are shadowing and fast fading fading, respectively. At each time step, the simulator computes the Signal-to-Interference-plus-Noise Ratio (SINR), as follows:

$$SINR_i(t) = \frac{P/p_i(t)}{\eta + I_i} \tag{28}$$

where η is the noise, P is the transmission power and I_i is the interference from other base stations. The system then dynamically selects the MCS based on the computed SINR values, and then computes the amount of RBs necessary for serving the corresponding packets.

We consider the sliding window approach. The simulation considers three distinct periods in one day: 0 to 8 hours, 8 to 16 hours, and 16 to 24 hours. During each period, the factory adjusts the number of working machines, which influences both the number of packet arrivals and thus the required number of resource blocks. The variation in the distribution of ξ across these periods is illustrated in Figure 3, where the continuous line represents the average number of required resource blocks, ξ, changing over time, and the points depict some realizations of this distribution during these periods. The simulator is executed continuously for a total of 24 hours, about 84×10^7 mini-slot.

Figure 4 shows the corresponding dynamic dimensioning, increasing and decreasing in response to shifting periods. Figure 5 illustrates the corresponding outage probability which indeed is below the target 10^{-7}, which shows the effectiveness of our algorithm.

V. Conclusion

We present in this paper a comprehensive framework for radio resource allocation tailored to extreme URLLC

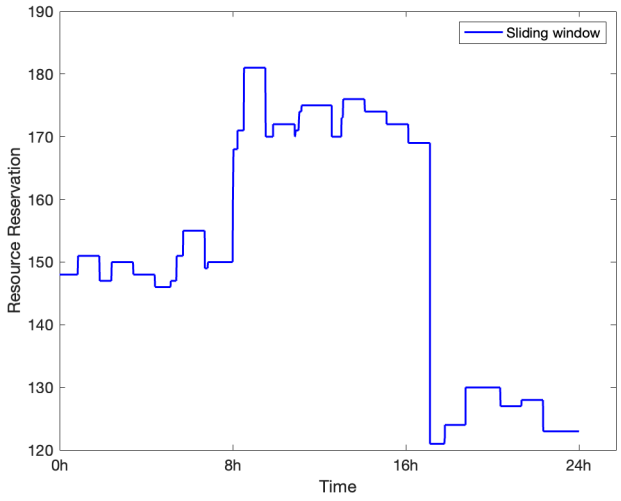

Fig. 4: The evolution of the number of resources reserved through the 24h, with $\epsilon = 10^{-7}$ and $\beta = 0.03$.

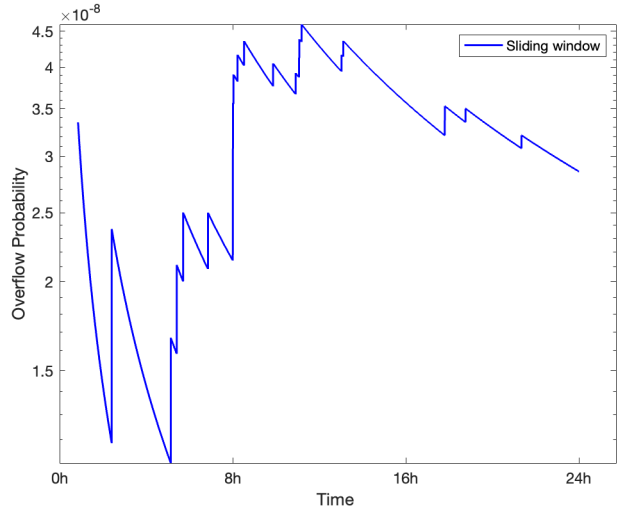

Fig. 5: The evolution of the outage through the 24h, with $\epsilon = 10^{-7}$ and $\beta = 0.03$.

traffic. Using chance-constrained optimization and various methodological approaches, including distributionally robust, distribution-based approximation and data-driven techniques, we address the challenge of finding the optimal resource reservation while meeting stringent QoS requirements. Our numerical comparisons, applied for the case of automated factory, both with homogeneous and heterogeneous MCS cases, show the effectiveness of the proposed data-driven approach, offering a practical solution adaptable to real-time conditions and different user radio conditions.

REFERENCES

[1] J. Park, S. Samarakoon, H. Shiri, M. K Abdel-Aziz, T. Nishio, A. Elgabli, and M. Bennis. "Extreme URLLC: Vision, challenges, and key enablers". In: *arXiv preprint arXiv:2001.09683* (2020).

[2] Y. Chen, H. Lu, L. Qin, C. Zhang, and C. W. Chen. "Statistical QoS provisioning analysis and performance optimization in xURLLC-enabled massive MU-MIMO networks: A stochastic network calculus perspective". In: *IEEE Transactions on Wireless Communications* (2024).

[3] H. Du, J. Liu, D. Niyato, J. Kang, Z. Xiong, J. Zhang, and D. I. Kim. "Attention-aware resource allocation and QoE analysis for metaverse xURLLC services". In: *IEEE Journal on Selected Areas in Communications* (2023).

[4] S. Eum, S. Arakawa, and M. Murata. "A probabilistic Grant Free scheduling model to allocate resources for eXtreme URLLC applications". In: *2022 IEEE Latin-American Conference on Communications (LATINCOM)*. IEEE. 2022, pp. 1–6.

[5] Y. Chen, H. Lu, L. Qin, Y. Deng, and A. Nallanathan. "When xURLLC meets NOMA: A stochastic network calculus perspective". In: *IEEE Communications Magazine* (2023).

[6] L. Li, W. Chen, and K. B Letaief. "Ultra-reliable and low latency wireless communications with burst traffics: A large deviation method". In: *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2021, pp. 01–06.

[7] S. Elayoubi, N. Naddeh, T. Chahed, and S. Ben Jemaa. "A Large Deviations Model for Latency Outage for URLLC". In: *EAI International Conference on Performance Evaluation Methodologies and Tools*. Springer. 2022, pp. 223–239.

[8] J. Liu, A. Lisser, and Z. Chen. "Distributionally robust chance constrained geometric optimization". In: *Mathematics of Operations Research* 47.4 (2022), pp. 2950–2988.

[9] L. El Ghaoui, M. Oks, and F. Oustry. "Worst-case value-at-risk and robust portfolio optimization: A conic programming approach". In: *Operations research* 51.4 (2003), pp. 543–556.

[10] 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, pp. 55–63.

[11] H. Jang, J. Kim, W. Yoo, and J. Chung. "URLLC mode optimal resource allocation to support HARQ in 5G wireless networks". In: *IEEE Access* 8 (2020), pp. 126797–126804.

[12] C. 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)*. IEEE. 2017, pp. 1–5.

[13] B. Shi, F. Zheng, C. She, J. Luo, and A. G Burr. "Risk-resistant resource allocation for eMBB and URLLC coexistence under M/G/1 queueing model". In: *IEEE Transactions on Vehicular Technology* 71.6 (2022), pp. 6279–6290.

[14] L. M Leemis and J. T McQueston. "Univariate distribution relationships". In: *The American Statistician* 62.1 (2008), pp. 45–53.

[15] S. Peng, F. Maggioni, and A. Lisser. "Bounds for probabilistic programming with application to a blend planning problem". In: *European Journal of Operational Research* 297.3 (2022), pp. 964–976.

[16] X. Geng and L. Xie. "Data-driven decision making in power systems with probabilistic guarantees: Theory and applications of chance-constrained optimization". In: *Annual reviews in control* 47 (2019), pp. 341–363.

[17] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia. "OpenRAN Gym: An open toolbox for data collection and experimentation with AI in O-RAN". In: *2022 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2022, pp. 518–523.