Providing Latency Guarantees in 6G Multi-Domain Networks

Fidan Mehmeti
Chair of Communication Networks
Technical University of Munich
fidan.mehmeti@tum.de

Wolfgang Kellerer
Chair of Communication Networks
Technical University of Munich
wolfgang.kellerer@tum.de

***Abstract*—Simultaneously with the rapid deployment of 5G networks, research and industrial players have started looking into the next generation of cellular networks, 6G. One of the main new features planned for 6G are multi-domain networks, where both public networks (owned by traditional cellular operators) and private networks (operated by different non-traditional entities) will be deployed and would need to inter-operate to provide a satisfying user experience. However, as these networks are operated by different entities, the aim of providing end-to-end guarantees to a user whose data traverse multiple networks before reaching their destination faces significant challenges. In this paper, we focus on providing latency guarantees in a multi-network setup. The approach followed here is to use limited information from other network domains to determine the resource allocation policy at the network domain of interest so that an end-to-end latency guarantee can be maintained between a communication pair at different domains. To that end, we provide analytical approaches, relying on queueing theory, for three distributions of the data rate at the receiver (leading to exponential services times, constant rates, and Round-robin), and for each one of them we propose the resource allocation policy at the transmitter side. The evaluation is performed on input data from a public dataset. Results show that our approaches outperform the benchmarks in terms of efficient resource utilization (by at least 35%), while being only $5-10\%$ worse than an (infeasible) oracle for all scenarios of interest.**

***Index Terms*—6G, Campus networks, Performance guarantees, Latency, Multi-domain networks**

I. Introduction

Although considerable performance improvements have been reported with 5G across different dimensions [1], [2], [3], there are still applications, such as holographic communications [4], whose successful operation relies on having too much resources, bringing the 5G network operation near exhaustion. The resource-intensive nature of these applications is reflected both in terms of large number of frequency bands (bandwidth-hungry applications), and in terms of the high reliability to deliver a packet within a very short time (latency-sensitive applications). Therefore, the research community backed up by industrial players [5] already started working on 6G networks, planned to be fully operational by 2030 [6].

The other aspect that 5G networks do not cover are multi-domain networks [5], comprising several and diverse single-domain networks (including traditional networks operated by "classical" network operators), ranging from molecular networks [7], body area networks [8], industrial mesh networks [9], aeronautical networks [10], up to smaller cellular networks operated by private entities [11]. The latter are known as *campus networks*. These are private networks, including Radio Access Network (RAN) and Core Network (CN), not owned by the cellular operators, providing network access within a university, hospital, etc.

From the description of campus networks, 6G will surely be heterolithic in terms of the operators of the networks through which data traverse. For instance, the data transmitter can be located within the coverage area of a network owned by a cellular network operator. The user's packets from the transmitter via the wireless interface will be sent to the Base Station (BS) that serves that user, from where the packets are forwarded to the CN. The receiver, on the other hand, may be within the coverage area of a campus network (over which the cellular operator has no control), operated by a private entity/institution. Therefore, these packets from the CN of the "transmitter network" are forwarded through the Internet to the CN of the campus network, from where they are further forwarded to the corresponding BS of the campus network. Finally, the packets are delivered to the intended receiver. The overall process is illustrated in Fig. 1.

Different operators managing the transmitter and receiver networks means that providing any end-to-end performance guarantees in a multi-domain network in terms of any of the metrics of interest poses significant challenges. For example, if there is a maximum allowed latency for a given application, it is not clear how the components of packet delay should be split among the different entities the packet traverses, i.e., how much time at most it should take in the transmitter network, backbone Internet, and receiver network. On top of that, the transmitter experiencing given channel conditions would require a given amount of RAN resources to maintain a maximum allowed latency on its side. On the other end of the communication path, the receiver will most probably experience different channel conditions. Therefore, it will require a different amount of resources to satisfy the still unknown allowed packet delay on its side. In addition, it would be cumbersome for different entities operating different campus networks to exchange all the information on the actual

This work was supported by the Federal Ministry of Education and Research of Germany (BMBF) under the project "6G-Life" with project identification number 16KISK002.

channel conditions of all their users. The other reason is that each campus network needs to maintain its privacy by disclosing only limited information to other campus networks.

The important question that arises is *how should we approach the problem analytically, whether following an integrated approach by considering the end-to-end delay and the reliability to provide it in multi-domain networks, or to split the allowed packet delay in its components over the corresponding entities?* If the decision is to use the latter approach, what values should these components attain? Answering these questions is not trivial.

In this paper, we follow a general analytical approach in which the network operation is modeled using a queueing network comprising individual queueing systems for each entity. Exploiting the nature of the queueing network for different allocation policies at the receiver campus network, we decide whether to consider the end-to-end latency in an integrated way, or to split the maximum allowed packet delays across different domains. We make this decision based on the existence or lack of analytical tractability for a given allocation policy at the receiver side (and the limited information that is sent to other domains by that network). When the policy is such that the problem is not analytically tractable, we resort to consider the latency guarantees confined to separate domains. Irrespective of how the problem at hand is approached, we provide an answer on how to allocate resources at the transmitter side without violating the end-to-end latency requirement. The approach we propose can be helpful for operators of campus networks running latency-sensitive applications across different domains. The main message of the paper is that having limited feedback from different single-domain networks suffices to support latency- and reliability-sensitive applications with users in different campus networks if the resource allocation at the transmitter is done properly. Our main contributions in this paper can be summarized as:

- We provide resource allocation policies at the transmitter-side network for three different resource allocation policies at the receiver-side network (yielding exponential service times, constant rates, and Round-robin) to guarantee a maximum latency with high reliability.
- We determine how the transmitter rate should change as a function of the receiver rate for a given scenario.
- Using simulations run on public traces, we show the advantages our approaches offer against benchmarks in terms of reliability and efficient resource utilization.

II. Problem Formulation

A. System model

We consider a multi-domain network, consisting of multiple single-domain networks, i.e., campus networks, including public networks. This is illustrated in Fig. 1. In general, campus networks are operated by different entities, and each operator entity of a campus network manages both the RAN and CN part, i.e., their operation resembles that of a traditional cellular network. In general, there can be multiple BSs associated with

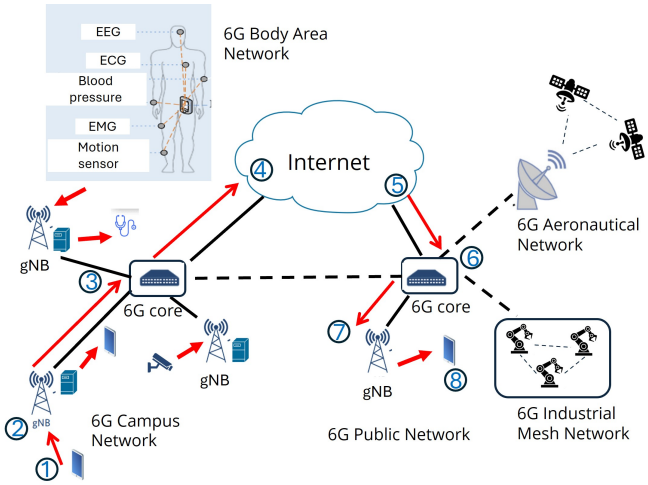

Fig. 1. Illustrating how the communication process in a 6G network, comprising multiple campus networks and a classical (public) cellular network, could look like. The segments in red depict parts of the path example of the communication process.

the unique CN in the same campus network. Such an example is a university with a large campus, where a single BS will not suffice to satisfy the traffic requirements of all its users.

While the communication is between transmitters and receivers of the same campus network, there are no changes in the operation compared to traditional cellular networks. The challenge arises when the transmitter and receiver are within coverage areas of different campus networks. In that case, the CN of the transmitter-side network forwards through the Internet the data to the CN of the receiving-side network, which further forwards them to the corresponding BS by which the receiver is being served. This closes the one-way direction of the process (illustrated in Fig. 1 via ①-⑧).

In this work, we consider the communication process between a single communication pair, a transmitter and a receiver. Therefore, to simplify the notation, we add the index tr to all the variables related to the transmitter and the index r to all the variables pertaining to the receiver. This leads to the assumption that we consider two campus networks.

Each campus network operates its own set of frequency blocks, known as Physical Resource Blocks (PRBs) that are the unit of resource allocation in contemporary cellular networks [12], distributed across all the BSs of a campus network. Each PRB consists of 12 subcarriers. The slot duration is a function of the subcarrier spacing. Specifically, if the subcarrier spacing is 15 kHz (PRB width of 180 kHz), the slot duration is 1 ms. If the subcarrier spacing is 30 kHz (PRB width of 360 kHz), the corresponding slot duration is 0.5 ms. The slot duration decreases further (by $2\times$) when switching to subcarrier spacings of 60 kHz, and another $2\times$ when switching to 120 kHz [12].[1] Different PRBs are assigned to different User Equipment (UEs), i.e., users, within a slot. The assignment varies across slots. Consequently, scheduling

[1]As still there are no indications regarding structural changes in resource allocation in 6G, for that part in this work we use the notions from 5G.

needs to be performed in two dimensions, *time* and *frequency*. In this work, the assumption is that there are K available PRBs in each BS of any campus network.

UEs experience different channel conditions across different PRBs even within the same slot. This is captured by the parameter known as Channel Quality Indicator (CQI) [12], which has values in the range $1-15$, depending on the Signal-to-Interference-Plus-Noise-Ratio (SINR), with lower values for worse channel conditions. As UEs can be mobile and given the inherent time-varying nature of the channels, per-PRB CQI changes from one slot to another, and its value depending on the Modulation and Coding Scheme used sets the per-PRB rate [13]. To keep the analysis tractable, we make a simplifying assumption in this work. Specifically, we assume that the BS splits the transmission power equally among all PRBs it transmits on, and that the UE channel characteristics remain static across all PRBs (identical CQI over all PRBs for that UE), but change randomly (according to some distribution) from one slot to another, and are mutually independent among UEs (i.e., we consider UEs with heterogeneous channel conditions). These assumptions relax the resource allocation problem to the number of allocated PRBs per UE.

The previous assumptions imply that in every slot, the transmitter will have per-PRB rate R_{tr} and the receiver R_r, which are the rates each assigned PRB brings to the respective user. The per-PRB rate can be modeled as a discrete random variable with values in $\{r_1, r_2, \ldots, r_{15}\}$ (note that there are 15 possible values of CQI), such that $r_1 < r_2 < \ldots < r_{15}$, with Probability Mass Functions (PMFs), in a slot, of $p_{R_{tr}}$ and p_{R_r}, both functions of their CQIs over time.

Traffic model: The traffic at the transmitter is generated according to a Poisson process with rate λ. The packets are of constant size Δ. If more packets are in the transmitter queue, they are served in a First-Come First-Served (FCFS) fashion.

B. Problem setup

Providing end-to-end latency guarantees in a multi-domain network is important for delay-sensitive applications. Use cases that fall into this category are known as Ultra-Reliable Low-Latency Communications (URLLC) [12], characterized by latency requirements in the order of milliseconds and reliability higher than 99%. An example would be a doctor performing a surgery to a remotely located patient [14], or controlling a factory robot remotely [5].

The next step is to define latency in the context of this work. A packet generated at the transmitter will be transmitted first to the corresponding BS. This would correspond to the link 1-2 in Fig. 1. The packet is then sent to the CN of the transmitter campus network (link 2-3). From the CN, the packet through Internet is sent to the CN of the receiver campus network, traversing the path 3-4-5-6. After being processed at the receiver CN, it is forwarded down to the BS within whose area the receiver is located. This is link 6-7. The final hop, link 7-8, is to the receiver.

Backhaul link capacities are usually much higher than rates in the RAN. The same holds for the rates between the CN and Internet, as well as the routing in Internet. Hence, the bottleneck links in this setup are 1-2 and 7-8. For other links in the path, we assume that in total they add a constant delay of T_0, which is lower than T_{tr} (the delay in link 1-2) and T_r (the delay in link 7-8). The packet delay in our setup is

$$T = T_{tr} + T_0 + T_r. \tag{1}$$

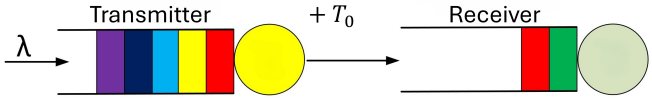

Fig. 2. The queueing network model.

When traversing link 1-2, the packet experiences *the transmission delay* at the UE and the propagation delay from UE to BS. Given the coverage areas of a BS in this setup, the propagation delays are very small (in the order of microseconds), and can therefore be neglected. This leads to T_{tr} consisting of only the queueing delay at the user and transmission time. The latter depends directly on the uplink rate of the UE.

On the other end of the path, the packets arriving to the receiver BS queue at the buffer dedicated to the receiver. FCFS is used there too, and the propagation delay between the receiving BS and the UE can be neglected. So, T_r is the sum of the queueing delay and service time at the UE, with the latter depending directly on the downlink rate at the receiver.

We denote by T_{max} the maximum allowed end-to-end latency for each packet. Having the vast majority of packets not exceed this latency can be captured by the inequality

$$\mathbb{P}\left(T \leq T_{max}\right) \geq 1 - \epsilon, \tag{2}$$

where T is given by (1), and ϵ is the *outage probability* with very low values. For example, if $\epsilon = 0.001$ only 0.1% of the packets will violate the latency requirement. The term $\mathbb{P}\left(T \leq T_{max}\right)$ is the reliability. In this example, it is $\geq 99.9\%$.

The problem in this work is to determine the data rate, and hence the resource allocation policy, at the transmitter so that (2) is satisfied. We remind the reader that in a multi-domain network the operator of the transmitter network has no control over the domain of receiver campus network; it can only obtain limited information from other campus networks.

From the previous discussion, the end-to-end process can be modeled using a queueing network, consisting of two queues and a constant delay in between. This is illustrated in Fig. 2. The first queue corresponds to the transmitter, whereas the second to the receiver. The constant delay is simply T_0.

In Section III, we consider three scenarios corresponding to different allocation policies at the receiver campus network.

III. Analyses for Different Policies

In this section, we consider three scenarios in terms of the data rates at the receiver, and based on that we determine the resource allocation policy at the transmitter in order to meet the latency requirement. These are: (i) data rates that yield exponentially-distributed service times, (b) constant data rates, and (iii) Round-robin allocation.

A. Exponential service times at the receiver

In the first scenario, the receiver data rate, U_r, is chosen such that the service time at the receiver queue is exponentially distributed[2] with mean $\mathbb{E}[t_r] = \mathbb{E}[\frac{\Delta}{U_r}]$. The probability density function (pdf) of this service time is

$$f_{t_r}(x) = \frac{1}{\mathbb{E}[t_r]} e^{-\frac{x}{\mathbb{E}[t_r]}}. \quad (3)$$

Note that we consider this case as this is the only scenario in which an analytical solution can be obtained. Also, we are working in this scenario with the average service time instead of the data rate because for the latter we would need as a parameter $\mathbb{E}\left[\frac{1}{U_r}\right]$, which is less descriptive than the transmission time and the data rate itself.

The receiver network can send the information that the receiver of interest experiences exponential service times with mean $\mathbb{E}[t_r]$. Based on this information and knowing the maximum allowed latency, T_{max}, the transmitter campus network can decide to provide a data rate, U_{tr}, that yields an exponentially distributed transmission time with pdf

$$f_{t_{tr}}(x) = \frac{1}{\mathbb{E}[t_{tr}]} e^{-\frac{x}{\mathbb{E}[t_{tr}]}}. \quad (4)$$

As the generation process at the transmitter is Poisson, and service times at both communication ends are exponential, the queueing network becomes a Jackson network [15], in which the queues can be considered in isolation, as M/M/1 queues. As is known from queueing theory [15], the system time (queueing + service time) in an M/M/1 queue is exponentially distributed with the rate equal to the difference between the service rate and the arrival rate. In our scenario, this amounts to exponentially distributed times with rates $\frac{1}{\mathbb{E}[t_{tr}]} - \lambda$ and $\frac{1}{\mathbb{E}[t_r]} - \lambda$ for the transmitter and receiver, respectively. Due to the flow preservation principle, the arrival rate to the receiver is unchanged, λ.

The constant delay T_0 simply shifts the exponentially distributed time T_{tr} by that amount, in which case the exponential property is still preserved. Therefore, the sum of latency components T_{tr} and T_0 is exponentially distributed with pdf

$$f_{T_{tr}+T_0}(x) = \left(\frac{1}{\mathbb{E}\left[t_{tr}\right]} - \lambda\right) e^{-\left(\frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\cdot(x - T_0)}. \quad (5)$$

From the previous discussion, the pdf of the latency component at the receiver is

$$f_{T_r}(x) = \left(\frac{1}{\mathbb{E}\left[t_r\right]} - \lambda\right) e^{-\left(\frac{1}{\mathbb{E}[t_r]} - \lambda\right)x}. \quad (6)$$

The pdf of the sum of two independent random variables is the convolution of their pdfs [15], which in our case would lead to the pdf of the overall latency of

$$f_T(x) = f_{T_{tr}+T_0}(x) * f_{T_r}(x). \quad (7)$$

To obtain the pdf of the total latency, we resort to using Laplace transforms. The Laplace transform of the convolution of the functions in the time domain is the product of their respective Laplace transforms. This transforms (7) into

$$\mathcal{L}_T(s) = \mathcal{L}_{T_{tr}+T_0}(s) \cdot \mathcal{L}_{T_r}(s). \quad (8)$$

The next step is to determine the Laplace transforms of the pdfs of $T_{tr} + T_0$ and T_r, which are $\mathcal{L}_{T_{tr}+T_0}(s)$ and $\mathcal{L}_{T_r}(s)$, respectively. With the Laplace transform of T_{tr} defined as

$$\mathcal{L}_{T_{tr}}(s) = \int_0^\infty f_{T_{tr}}(x) e^{-sx} dx = \frac{(\mathbb{E}[t_{tr}])^{-1} - \lambda}{s + (\mathbb{E}[t_{tr}])^{-1} - \lambda}, \quad (9)$$

and delaying by T_0 in the time domain multiplies the actual Laplace transform by e^{-sT_0} (directly from (9)), leading to[3]

$$\mathcal{L}_{T_{tr}+T_0}(s) = (s + 1/\mathbb{E}\left[t_{tr}\right] - \lambda)^{-1} e^{-sT_0} (1/\mathbb{E}[t_{tr}] - \lambda). \quad (10)$$

Substituting (10) and the Laplace transform of $f_{T_r}(x)$, which from (9) is $\frac{\frac{1}{\mathbb{E}[t_r]} - \lambda}{s + \frac{1}{\mathbb{E}[t_r]} - \lambda}$, into (8), we obtain

$$\mathcal{L}_T(s) = \frac{\left(\frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\left(\frac{1}{\mathbb{E}[t_r]} - \lambda\right) e^{-sT_0}}{\left(s + \frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\left(s + \frac{1}{\mathbb{E}[t_r]} - \lambda\right)}. \quad (11)$$

Next, we turn our attention to (2). Its left-hand side (LHS) is the cumulative distribution function (CDF) at the maximum latency, $F_T(T_{max})$. Therefore, the latency/reliability requirement (2) reduces to

$$F(T_{max}) \geq 1 - \epsilon. \quad (12)$$

The next step is to determine the Laplace transform of the CDF of the end-to-end latency, $F_T(x)$. The relation between CDF and pdf is

$$F_T(x) = \int_0^x f_T(y) dy, \quad (13)$$

or in the s-domain

$$\mathcal{L}_{F_T}(s) = \frac{\mathcal{L}_T(s)}{s}. \quad (14)$$

Substituting (11) into (14), we obtain the following:

$$\mathcal{L}_{F_T}(s) = \frac{\left(\frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\left(\frac{1}{\mathbb{E}[t_r]} - \lambda\right) e^{-sT_0}}{s\left(s + \frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\left(s + \frac{1}{\mathbb{E}[t_r]} - \lambda\right)}. \quad (15)$$

From (15), to determine the CDF of the end-to-end latency in this multi-domain network, we need to take the inverse Laplace transform of (15). As a final result, we obtain:

Lemma 1. *The CDF of the end-to-end latency at T_{max} in a multi-domain network with exponential service times at the transmitter and receiver is*

$$F(T_{max}) = 1 - \frac{\frac{1}{\mathbb{E}[t_r]} - \lambda}{\frac{1}{\mathbb{E}[t_r]} - \frac{1}{\mathbb{E}[t_{tr}]}} e^{-\left(\frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\cdot(T_{max} - T_0)} + \frac{\frac{1}{\mathbb{E}[t_{tr}]} - \lambda}{\frac{1}{\mathbb{E}[t_r]} - \frac{1}{\mathbb{E}[t_{tr}]}} e^{-\left(\frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\cdot(T_{max} - T_0)}. \quad (16)$$

Proof. The inverse Laplace transform of (15) can be written as

$$F_T(x) = \left(\frac{1}{\mathbb{E}\left[t_{tr}\right]} - \lambda\right) \cdot \left(\frac{1}{\mathbb{E}\left[t_r\right]} - \lambda\right) \cdot \mathcal{L}^{-1}\left\{\frac{e^{-sT_0}}{s\left(s + \frac{1}{\mathbb{E}[t_{tr}]} - \lambda\right)\left(s + \frac{1}{\mathbb{E}[t_r]} - \lambda\right)}\right\}. \quad (17)$$

[2]In fact, this is the transmission time at the BS in the receiver campus network directly depending on the download rate of the receiver UE. Nevertheless, we are going to refer to this as the receiver service time.

[3]Note that we denote the time argument by x as we have already used t too many times for the different time components. Also, we use the unilateral Laplace transform because we are dealing with time.

As a first step, the argument under the inverse Laplace transform can be transformed into

$$\frac{e^{-sT_0}}{\left(\frac{1}{\mathbb{E}[t_r]}-\frac{1}{\mathbb{E}[t_{tr}]}\right)s}\cdot\left(\frac{1}{s+\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}-\frac{1}{s+\frac{1}{\mathbb{E}[t_r]}-\lambda}\right). \quad (18)$$

From (18), we need to expand the products into

$$\frac{1}{s\left(s+\frac{1}{\mathbb{E}[t_{tr}]}-\lambda\right)}=\frac{1}{\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}\left(\frac{1}{s}-\frac{1}{s+\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}\right), \quad (19)$$

and

$$\frac{1}{s\left(s+\frac{1}{\mathbb{E}[t_r]}-\lambda\right)}=\frac{1}{\frac{1}{\mathbb{E}[t_r]}-\lambda}\left(\frac{1}{s}-\frac{1}{s+\frac{1}{\mathbb{E}[t_r]}-\lambda}\right). \quad (20)$$

The last two expansions transform (18) into

$$\frac{1}{\frac{1}{\mathbb{E}[t_r]}-\frac{1}{\mathbb{E}[t_{tr}]}}\left(\frac{1}{\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}\left(\frac{e^{-sT_0}}{s}-\frac{e^{-sT_0}}{s+\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}\right)-\frac{1}{\frac{1}{\mathbb{E}[t_r]}-\lambda}\left(\frac{e^{-sT_0}}{s}-\frac{e^{-sT_0}}{s+\frac{1}{\mathbb{E}[t_r]}-\lambda}\right)\right) \quad (21)$$

Shifting the function in the time domain to the right by T_0 implies multiplying the Laplace transform of the corresponding function by e^{-sT_0}. Further, the inverse Laplace transforms of $\frac{1}{s}$, $\frac{1}{s+\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}$, and $\frac{1}{s+\frac{1}{\mathbb{E}[t_r]}-\lambda}$ are $u(x)$ (i.e., the Heaviside function), $e^{-\left(\frac{1}{\mathbb{E}[t_{tr}]}-\lambda\right)x}$, and $e^{-\left(\frac{1}{\mathbb{E}[t_r]}-\lambda\right)x}$, respectively.

The arguments of the previous paragraph lead to the inverse Laplace transform of (21):

$$\frac{1}{\frac{1}{\mathbb{E}[t_r]}-\frac{1}{\mathbb{E}[t_{tr}]}}\left(\frac{1}{\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}\left(u(x-T_0)-e^{-\left(\frac{1}{\mathbb{E}[t_{tr}]}-\lambda\right)(x-T_0)}\right)-\frac{1}{\frac{1}{\mathbb{E}[t_r]}-\lambda}\left(u(x-T_0)-e^{-\left(\frac{1}{\mathbb{E}[t_r]}-\lambda\right)(x-T_0)}\right)\right), \quad (22)$$

which substituted into (17) results in the CDF of the end-to-end latency:

$$F_T(x)=u(x-T_0)-\frac{1}{\frac{1}{\mathbb{E}[t_r]}-\frac{1}{\mathbb{E}[t_{tr}]}}\cdot\left(\left(\frac{1}{\mathbb{E}\left[t_r\right]}-\lambda\right)e^{-\left(\frac{1}{\mathbb{E}[t_{tr}]}-\lambda\right)(x-T_0)}-\left(\frac{1}{\mathbb{E}\left[t_{tr}\right]}-\lambda\right)e^{-\left(\frac{1}{\mathbb{E}[t_r]}-\lambda\right)(x-T_0)}\right). \quad (23)$$

As $T_{max}>T_0$, in (23), $u(T_{max}-T_0)=1$. Therefore, for $x=T_{max}$, (23) reduces to (16). □

Finally, substituting (16) into (12), we obtain:

Theorem 2. *In a multi-domain network with exponential service times at the transmitter and receiver, where the average receiver service time is $\mathbb{E}[t_r]$, to meet the latency T_{max} with a reliability of at least $1-\epsilon$, the mean transmitter service time is the highest value $\mathbb{E}[t_{tr}]$ that satisfies the inequality*

$$\frac{\frac{1}{\mathbb{E}[t_r]}-\lambda}{\frac{1}{\mathbb{E}[t_r]}-\frac{1}{\mathbb{E}[t_{tr}]}}e^{-\left(\frac{1}{\mathbb{E}[t_{tr}]}-\lambda\right)\cdot(T_{max}-T_0)}-\frac{\frac{1}{\mathbb{E}[t_{tr}]}-\lambda}{\frac{1}{\mathbb{E}[t_r]}-\frac{1}{\mathbb{E}[t_{tr}]}}e^{-\left(\frac{1}{\mathbb{E}[t_r]}-\lambda\right)\cdot(T_{max}-T_0)}\leq\epsilon. \quad (24)$$

Inequality (24) needs to be solved numerically as it is transcendental in $\mathbb{E}\left[t_{tr}\right]$. As can be observed from this subsection, even for exponentially distributed service times with the corresponding data rates at both communication ends, it is cumbersome to maintain the tractability when analyzing the latency in a multi-domain network. Therefore, in the next two scenarios, for other distributions of the data rates at the receiver, we switch to another approach.

B. Constant data rate at the receiver

In the second scenario, at the receiver campus network, the user of interest is guaranteed a constant data rate, U_r. This is the information that is sent (only once) from the receiver campus network to the transmitter operator. With this assumption, the considered queueing network is not Jackson anymore. Therefore, we need to resort to an alternative approach in determining the resource allocation policy at the transmitter side to satisfy (2). This approach consists in splitting the maximum allowed latency into the corresponding components. To that end, there are two important issues to address. The first is what data rate to provide to the transmitter, and hence what resource allocation policy to follow, whereas the second is how to determine the allowed latency components at each queue.

We will start with addressing the first issue, that of the required data rate at the transmitter, and concretely, its distribution. The answer to this question lies in the LHS of (2), i.e., at the CDF. The problem reduces to determining the data rate distribution (among the set of all possible distributions with the same average value) that provides the highest value of $F_{T_{tr}}(T_{max})$ (i.e., the highest reliability). It turns out that this is the case with deterministic (constant) data rates. The rationale behind this is given as follows.

As is well known [15], among all the queueing systems with Poisson arrival and general service times (M/G/1), the queue whose average total packet delay is lowest is the one with deterministic service times. Further, according to [16] (Theorem B, pg.6), and in line with a variation of *stochastic dominance* principle, if a process has a lower average than another process, then it must have a higher or equal CDF at any point compared to the other process. Since, as already mentioned, the queue with deterministic (constant) service time has the lowest average total delay among all M/G/1 queues with the same average service time, it follows that it will also have the highest $F_{T_{tr}}(T_{max})$. Therefore, the best option is to provide a constant data rate to the transmitter as well. We denote it by U_{tr}.

Before determining U_{tr}, we need to look at the second issue mentioned previously, that of splitting the allowed latency. Let us denote the allowed latency at the transmitter by $T_{max}^{(tr)}$. The latency requirement at the transmitter then becomes

$$\mathbb{P}\left(T_{tr} \leq T_{max}^{(tr)}\right) \geq 1-\epsilon. \quad (25)$$

In the next step, we determine the minimum (constant) data rate that is sufficient to satisfy (25). The CDF of the queueing time in an M/D/1 queue is [15]

$$F_Q(x) = \left(1-\frac{\lambda\Delta}{U_{tr}}\right)\sum_{i=0}^{\lfloor\frac{U_{tr}x}{\Delta}\rfloor} e^{-\lambda(i\frac{\Delta}{U_{tr}}-x)}\frac{(i\frac{\Delta}{U_{tr}}-x)^i}{i!}\lambda^i, \quad (26)$$

with the floor function in the upper limit of the summation.

The CDF of the allowed latency at the transmitter is the convolution of the CDF of queueing time with the pdf of service time [15]:

$$F_{T_{tr}}(T_{max}^{(tr)}) = F_Q(T_{max}^{(tr)}) * \delta(T_{max}^{(tr)} - \frac{\Delta}{U_{tr}}) = F_Q(T_{max}^{(tr)} - \frac{\Delta}{U_{tr}}), \quad (27)$$

where $\delta(x)$ is the Dirac delta function [17]. For deterministic service times, their pdf is a shifted Dirac function. The convolution of a signal with a shifted Dirac delta is the shifted signal itself [17]. Hence, the RHS of (27).

Substituting (26) into (27), and the latter into (25), after rearranging we obtain:

Result 3. *The required transmitter data rate with traffic generation rate λ for a packet to be reliably transmitted within $T_{max}^{(tr)}$ to the BS of the transmitter network with a minimum reliability of $1-\epsilon$ is the minimum value of U_{tr} that satisfies the following inequality:*

$$\sum_{i=0}^{\lfloor\frac{U_{tr}}{\Delta}T_{max}^{(tr)}-1\rfloor} e^{-\lambda((i+1)\frac{\Delta}{U_{tr}}-T_{max}^{(tr)})}\frac{((i+1)\frac{\Delta}{U_{tr}}-T_{max}^{(tr)})^i}{i!}\lambda^i \geq \frac{1-\epsilon}{1-\frac{\lambda\Delta}{U_{tr}}}. \quad (28)$$

The value of U_{tr} as a function of $T_{max}^{(tr)}$ from (28) can be obtained numerically. The next step is to determine the latency component at the transmitter, $T_{max}^{(tr)}$, and its relation to the maximum allowed latency at the receiver, $T_{max}^{(r)}$. To that end, we depart with exploiting some knowledge from queueing theory on the variability of the arrival processes at both queues, and then propose a solution. As the first queue (the transmitter) behaves as an M/D/1 system, the coefficient of variation (the ratio of the standard deviation and the mean) of the inter-departure time from it can be approximated as [15]

$$c_{d,tr}^2 = (1-\rho_{tr}^2)c_{a,tr}^2 + \rho_{tr}^2 c_{s,tr}^2, \quad (29)$$

where $\rho_{tr} = \frac{\lambda\Delta}{U_{tr}}$ is transmitter utilization ratio, $c_{a,tr}$ is the coefficient of variation of inter-generation times at the transmitter (with a Poisson arrival process, $c_{a,tr}=1$), and $c_{s,tr}$ is the coefficient of variation of service times at the transmitter (as it is deterministic, $c_{s,tr}=0$). This reduces (29) to

$$c_{d,tr}^2 = 1-\rho_{tr}^2. \quad (30)$$

The inter-arrival time distribution at the receiver, $c_{a,r}$, is identical to the inter-departure time of packets from the transmitter as the constant delay T_0 is added to every packet on their way:

$$c_{a,r}^2 = c_{d,tr}^2 = 1-\rho_{tr}^2. \quad (31)$$

Eq.(31) reveals that $c_{a,r}$ in this scenario is not very far from 1. Namely, due to the stringent latency requirement ρ_{tr} is much lower than 1, making $\rho_{tr}^2 << 1$. This results in $c_{a,r} \approx c_{a,tr} = 1$, i.e., the arrival process at the receiver is approximately Poisson, with the receiver queue behaving as an M/D/1. Knowing U_r, the transmitter using (28), adapting the indices to the receiver side, can determine numerically $T_{max}^{(r)}$. Finally, we have

$$T_{max}^{(tr)} = T_{max} - T_{max}^{(r)} - T_0, \quad (32)$$

which substituted into (28) yields the lowest required constant rate at the transmitter, U_{tr}. In terms of the allocation policy this implies that in a slot the number of required PRBs by the transmitter UE is $K_{tr} = \frac{U_{tr}}{R_{tr}}$.

C. Round-robin allocation

In the last scenario, the receiver BS allocates its resources in a Round-robin fashion to its users. The receiver UE in a slot experiences the rate $U_r = \frac{K}{n_r}R_r$, where n_r is the number of (always active) users in the receiver BS. The receiver campus network can then send $\mathbb{E}[U_r] = \frac{K}{n_r}\mathbb{E}[R_r]$ and its coefficient of variation as information to the transmission network.

In line with the discussion from Section III-B, we choose to provide a constant data rate to the transmitter U_{tr} because with Round-robin at the transmitter side there is no control over the latency, and the required latency may not be met with a given reliability. This is more emphasized in scenarios with a large number of users being active in both domains.

From the previous discussion, the transmitter queue is M/D/1. Hence, following a similar reasoning as before, the arrival process at the receiver queue can be approximated as Poisson, making the receiver queue M/G/1. We need to take into account the distribution of the service time at the receiver too. The coefficient of variation of the receiver service time is

$$c_{s,r} = c_{t_r} = c_{\frac{\Delta}{\frac{K}{n}R_r}} = c_{\frac{n\Delta}{KR_r}} = c_{\frac{1}{R_r}}, \quad (33)$$

where the last step follows from the fact that $\frac{n\Delta}{K}$ is a constant, and multiplying a random variable by a constant does not change its coefficient of variation.

If $c_{s,r} \to 0$, the receiver queue resembles an M/D/1, in which case we can use the approach from Section III-B. However, that is rarely the case. The further away $c_{s,r}$ from 0, the lower the reliability for the same maximum latency. Therefore, the latency requirement at the receiver should be more relaxed. Empirical evidence, from running many simulations, suggest that a good way to split the latency into the components is $T_{max}^{(tr)} = \frac{c_{s,r}(T_{max}-T_0)}{1+c_{s,r}}$ and $T_{max}^{(r)} = \frac{T_{max}-T_0}{1+c_{s,r}}$. Substituting the so-obtained $T_{max}^{(tr)}$ into (28), we obtain U_{tr} and $K_{tr} = \frac{U_{tr}}{R_{tr}}$.

IV. Performance Evaluation

A. Simulation setup

In all the scenarios, there are six users both on the transmitter- and receiver-campus network. There are two pairs of users that communicate in this multi-domain network; w.l.o.g. let us assume that these are user 1 at the transmission

campus network and user 1 at the receiver campus network (user pair 1), and user 2 at the transmission campus network with user 2 at the receiver campus network (user pair 2). The PMFs of their per-PRB rates are given in Table I, and are from a publicly-available trace [18]. As for the other users, the presented results here (pertaining to the two user pairs of interest) are oblivious to their channel characteristics. Therefore, we omit their statistics due to space limitations.

The slot duration is 0.5 ms, implying a subcarrier spacing of 30 kHz. With 12 subcarriers per PRB, PRB width is 360 kHz. The number of PRBs on both campus networks is 273 [12]. MATLAB R2020b is used as the simulation environment.

B. Exponential-exponential service times

In the first scenario, we consider exponential service times at both campus networks, with mean at the receiver of $\mathbb{E}[t_r] = 0.5$ ms. Packets are generated as a Poisson process and are of (constant) size $\Delta = 12$ kbits. The reliability is ≥ 0.99. i.e., $\epsilon = 0.01$, and $T_0 = 1$ ms. Fig. 3 shows the required mean transmission time $\mathbb{E}[t_{tr}]$ as a function of T_{max}, for pair 1, for $\lambda = 10\,\text{s}^{-1}$ and $\lambda = 100\,\text{s}^{-1}$. As can be observed, $\mathbb{E}[t_{tr}]$ is an increasing function in T_{max} (with more relaxed latency requirement the transmission time can be higher). The other observation is that for higher λ the transmission time should be lower as there is more traffic and the data need to be transmitted faster to meet the latency requirement.

Next, we investigate how the required $\mathbb{E}[t_{tr}]$ varies with λ for user pair 2, for two maximum latencies, $T_{max} = 10$ ms and $T_{max} = 8$ ms. Fig. 4 depicts the results for $\mathbb{E}[t_r] = 0.4$ ms. As $T_{max,2} > T_{max,1}$, a higher data rate is required to satisfy the latency requirement for both pairs. A more relaxed latency requirement ($T_{max} = 10$ ms) requires fewer resources, and hence, allows a higher transmission time than the stricter latency ($T_{max} = 8$ ms).

C. Constant-constant rates

In this scenario, we compare the performance of our approach (deterministic rates at both sides) against three benchmarks, (1) exponential service times at the transmitter, (2) Round-robin at the transmitter, and (3) centralized approach, in which a single operator entity oversees both campus networks. In all three benchmarks, the receiver has constant rate. First, we compare the required transmitter rates with our approach from Section III-B against benchmark (3) in which the operator for a given constant receiver rate can determine the minimum U_{tr} to meet the latency requirement. We do this for two different receiver rates $U_r = 18$ Mbps and $U_r = 24$ Mbps, for user pair 1. The other parameters remain unchanged. Fig. 5 portrays the transmitter required rate as a function of T_{max}. The results with our proposed approach for this scenario are denoted by "Dist.", whereas those obtained with benchmark (3) correspond to "Cent.". Comparing the results from Fig. 5, our approach for $U_r = 24$ Mbps requires a higher transmitter rates of about 5% than "Cent." and about $5 - 13\%$ when $U_r = 18$ Mbps. The discrepancy is lower for more relaxed latency requirements (higher T_{max}). The lower values of T_{max} of $5 - 6$ ms are very restrictive and hard to be met. So, we can say that for realistic values of T_{max} our approach requires only about 5% more resources than with (3), where the latter acts as an oracle, and is infeasible in multi-domain networks.

Further, we compare the performance of our approach against benchmark (1) in terms of the required transmission time for $\lambda = 100\,\text{s}^{-1}$ and $\lambda = 200\,\text{s}^{-1}$. The rate at the receiver is $U_r = 24$ Mbps. Fig. 6 shows the results. The results with our approach are marked as "Det". As can be observed, with our approach much higher transmission times can be afforded (lower data rates needed), leading to more efficient utilization of resources at the transmitter campus network.

Benchmark (2), Round-robin at the transmitter, cannot even provide the reliability of 99% when $T_{max} < 8$ ms for $\lambda = 100\,\text{s}^{-1}$. For $\lambda = 200\,\text{s}^{-1}$, things become even worse because benchmark (2) cannot guarantee $T_{max} \leq 10$ ms with $\epsilon = 0.01$.

In relation to the previous scenario, we compare the resource utilization with our approach for both transmitter users and benchmark (2). Sample results are shown in Fig. 7. For user 1, the resource utilization is on average (out of $K = 273$) 11.93%, whereas for user 2 it is 8.5%. Because of the six users at the transmitter campus network, with Round-robin the amount of resources is always $(100/6)\%$, which is higher than with our proposed approach by at least 35%, without even guaranteeing the maximum latency, as mentioned previously.

Finally, Fig. 8 shows the dependency of the transmitter rate on the receiver rate for $\lambda = 150\,\text{s}^{-1}$, when constant rates are provided to both users in the communication pair. This is done for $T_{max} = 9$ ms and $T_{max} = 10$ ms, with the other parameters unchanged. In summary, reducing the rate on one end requires increasing that of the other and higher rates are required for stricter latencies. Similar conclusions can be drawn for our approach when at the receiver Round-robin is used. Due to space limitations, we omit those results.

V. Related Work

The goal of providing performance guarantees in cellular networks for different metrics has been targeted for a long time. One of the most exploited metrics in terms of performance guarantees is throughput. In [19] and [20], the authors focus on providing throughput guarantees to users belonging to the same and different use cases. For example, [19] proposes the same constant rate for all users for the vast majority of the time, but not 100% of the time, and shows the advantages of this approach compared to the scenario when a constant rate is guaranteed always. While in [19] the analysis determines the maximum achievable rate for all users (the same throughput to everyone), in [20] the maximum achievable data rate is determined for each user separately, depending on their channel conditions and the available resources. However, neither [19] nor [20] deal with providing latency guarantees to the users for the vast majority of the time, i.e., with high reliability.

More related in spirit to this work, [21] deals with admission control policies which guarantee the already admitted users that their packets will almost never exceed a given maximum latency. Latency guarantees for different slices in

TABLE I

PER-PRB RATES AND THE RESPECTIVE PROBABILITIES FOR USERS 1 AND 2 ON THE TRANS. NETWORK AND USERS 1 AND 2 ON THE REC. NETWORK

R (kbps)	48	73.6	121.8	192.2	282	378	474.2	612	772.2	874.8	1063.8	1249.6	1448.4	1640.6	1778.4
$p_{1,l}$ (tr.)	0	0.1	0.72	0.04	0.05	0.09	0	0	0	0	0	0	0	0	0
$p_{2,l}$ (tr.)	0	0	0.2	0.7	0.1	0	0	0	0	0	0	0	0	0	0
$p_{1,l}$ (rec.)	0	0	0	0	0.01	0.12	0.51	0.32	0.01	0.01	0.02	0	0	0	0
$p_{2,l}$ (rec.)	0.18	0.11	0.1	0.07	0.05	0.1	0.16	0.11	0.02	0.04	0	0.03	0	0.02	0.01

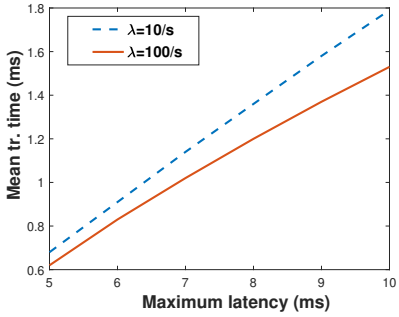

Fig. 3. The required mean transmission times with receiver mean service time of 0.5 ms and $\lambda = 10\ s^{-1}$ and $\lambda = 100\ s^{-1}$.

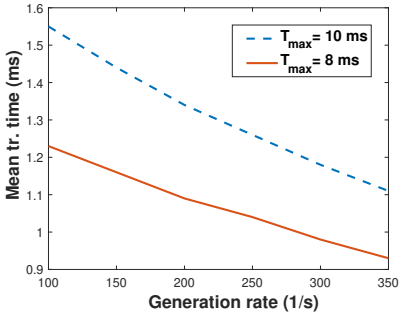

Fig. 4. The required mean transmission times with receiver mean service time of 0.4 ms and $T_{max} =$ 10 ms and $T_{max} = 8$ ms.

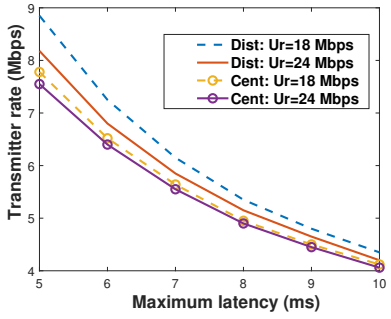

Fig. 5. The required (constant) transmitter rate with our approach ("Dist.") and a centralized approach ("Cent.") for U_r of 18 and 24 Mbps.

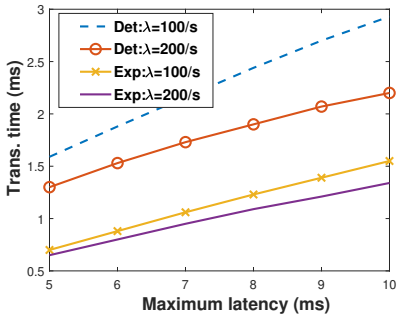

Fig. 6. The required mean transmission times at the trans. for exponential and deterministic scenarios when the receiver has a constant rate of 24 Mbps.

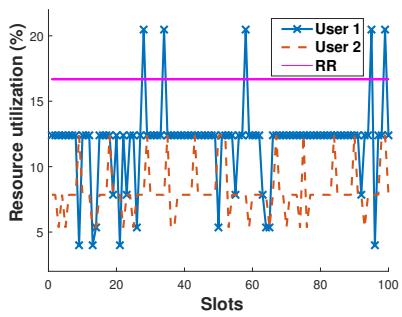

Fig. 7. Evolution of res. utilization at trans. users 1 and 2 with constant rates and Round-robin ($\lambda =$ $100\ s^{-1}$, $T_{max} = 10$ ms).

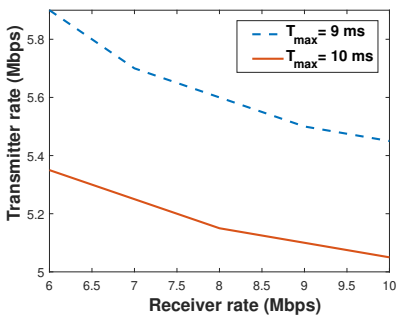

Fig. 8. The required transmission constant rate as a function of the receiver constant rate for different T_{max} and $\lambda = 150\ s^{-1}$.

a cellular network are provided in [22]. The implications of providing latency and reliability guarantees to URLLC traffic are investigated in [23], [24], [25]. Further, [26] focuses on providing latency guarantees with the added functionality of edge computing, where latency consists of the uplink and downlink transmission times and the processing delay at the edge cloud in a single cell. In [27], the authors propose approaches to minimize the latency in a mmWave cellular network. To meet the latency and reliability requirements of certain types of traffic in a cellular network, the authors in [28] propose a periodic resource allocation strategy, with constant packet sizes. Common to all these works is that they focus on the use cases of URLLC. However, despite the fact that the aforementioned works provide latency guarantees, they are all tailored only for single-domain networks, i.e., for public cellular networks, where the operator has full knowledge of the topology in the entire network over time. In the multi-domain network, this is not the case and consequently, the results from these related works cannot be applied.

In contrast, in this work we consider the problem of providing latency guarantees in a multi-domain network where the transmitting-side network obtains only partial information (such as the data rate of a user or the average number of active users) from other campus networks. We perform analyses and propose resource allocation policies on the transmitting campus network for different scenarios. To our best knowledge, there are no other works that tackle the problem of providing latency guarantees, of interest to URLLC services, in a multi-domain network, envisioned to be one of the features in 6G.

VI. CONCLUSION

In this work, we considered the problem of providing end-to-end latency guarantees with a given reliability in a multi-domain network, when the transmitter and receiver are located in different domains. We did this for three rate distributions at the receiver network (leading to exponential service times, constant rates, and with resources allocated in Round-robin fashion). For all three rate types, we derived the allocation policies at the transmitter side (the only domain the transmitter campus network has control over) so that the end-to-end latency is maintained with a given (high) reliability. Results show that our approaches outperform the benchmarks in terms of efficient resource utilization (by at least 35%), while being only $5-10\%$ worse than an oracle for all scenarios of interest.

In the future, we plan to incorporate mobility management into multi-domain networks with performance guarantees.

REFERENCES

[1] N. U. Ginige, K. B. Shashika Manosha, N. Rajatheva, and M. Latva-aho, "Admission control in 5G networks for the coexistence of eMBB-URLLC users," in *Proc. of IEEE VTC-Spring*, 2020.

[2] N. et al., "5G performance measurements in mobility for the bus transportation system in an urban environment," *IEEE Access*, 2023.

[3] J. W. et al., "Spectral efficiency improvement with 5G technologies: Results from field tests," *IEEE Journal on Sel. Areas in Communications*, vol. 35, no. 8, 2017.

[4] Z. Sun and Y. Jing, "Holographic MIMO NOMA communications: A power saving design," *IEEE Transactions on Wireless Communications*, vol. 23, no. 12, 2024.

[5] M. H. et al., "A secure and resilient 6G architecture vision of the German flagship project 6G-ANNA," *IEEE Access*, vol. 11, 2023.

[6] M. G. et al., "Toward 6G networks: Use cases and technologies," *IEEE Communications Magazine*, vol. 58, no. 3, 2020.

[7] U. U. et al., "Low-complexity timing methods for molecular communication via diffusion," *IEEE Transactions on Molecular, Biological, and Multi-Scale Communications*, vol. 9, no. 4, 2023.

[8] M. Mukhtar, M. Hanif Lashari, M. Alhussein, S. Karim, and K. Aurangzeb, "Energy-efficient framework to mitigate denial of sleep attacks in wireless body area networks," *IEEE Access*, vol. 12, 2024.

[9] G. P. et al., "A bluetooth low energy real-time protocol for industrial wireless mesh networks," in *Proc. of IECON*, 2016.

[10] A. Papa, J. von Mankowski, H. Vijayaraghavan, B. Mafakheri, L. Goratti, and W. Kellerer, "Enabling 6G applications in the sky: Aeronautical federation framework," *IEEE Network*, vol. 38, no. 1, 2024.

[11] M.-I. Corici, V. Gowtham, T. Magedanz, A. Prakash, and F. Schreiner, "NEMI: A 6G-ready AI-enabled autonomic network management system for open campus networks," in *IEEE Globecom Wkshps*, 2022.

[12] ETSI, "5G NR overall description: 3GPP TS 38.300 version 15.3.1 release 15." www.etsi.org, 2018. Technical specification.

[13] F. Mehmeti and T. La Porta, "Reducing the cost of consistency: Performance improvements in next generation cellular net. with optimal resource reallocation," *IEEE Tran. on Mob. Comp.*, vol. 21, no. 7, 2022.

[14] F. J. et al., "6G networks for the operating room of the future," *Progress in Biomedical Engineering*, vol. 6, oct 2024.

[15] J. Shortle, J. Thompson, D. Gross, and C. Harris, *Fundamentals of Queueing Theory*. Wiley, 2018.

[16] R. Szekli, *Stochastic ordering and dependence in applied probability*. Springer, 1995.

[17] A. Oppenheim and A. Willsky, *Signals and systems*. Prentice Hall, 1996.

[18] D. Raca, D. Leahy, C. J. Sreenan, and J. J. Quinlan, "Beyond throughput, the next generation: A 5G dataset with channel and context metrics," in *Proc. of ACM MMSys*, 2020.

[19] F. Mehmeti and C. Rosenberg, "How expensive is consistency? Performance analysis of consistent rate provisioning to mobile users in cellular networks," *IEEE Tran. on Mobile Computing*, vol. 18, no. 5, 2019.

[20] F. Mehmeti, A. Papa, W. Kellerer, and T. F. La Porta, "Minimizing rate variability with effective resource utilization in cellular networks," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, 2024.

[21] F. Mehmeti and T. La Porta, "Admission control for URLLC users in 5G networks," in *Proc. of ACM MSWiM 2021*.

[22] L. Dong and R. Li, "Latency guarantee service slice in 5G and beyond," in *Proc. of IEEE CCNC*, 2022.

[23] 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, 2018.

[24] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, and B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Communications*, vol. 25, no. 3, 2018.

[25] C. Sun, C. She, and C. Yang, "Energy-efficient resource allocation for ultra-reliable and low-latency communications," in *Proc. of IEEE GLOBECOM*, 2017.

[26] F. Mehmeti, V. T. Haider, and W. Kellerer, "Admission control for URLLC traffic with computation requirements in 5G and beyond," in *Proc. of IEEE/IFIP NOMS*, 2023.

[27] R. Ford, M. Zhang, M. Mezzavilla, S. Dutta, S. Rangan, and M. Zorzi, "Achieving ultra-low latency in 5G millimeter wave cellular networks," *IEEE Communications Magazine*, vol. 55, no. 3, 2017.

[28] Y. Han, S. E. Elayoubi, A. Galindo-Serrano, V. S. Varma, and M. Messai, "Periodic radio resource allocation to meet latency and reliability requirements in 5G networks," in *Proc. of IEEE VTC*, 2018.