Bandwidth Provisioning for Network Slices with Per User QoS Guarantees

Panagiotis Nikolaidis and Asim Zoulkarni and John Baras
Department of Electrical & Computer Engineering and the Institute for Systems Research
University of Maryland, College Park, MD 20742, USA
Email: {nikolaid, asimz, baras}@umd.edu

NOMS 2023-2023 IEEE/IFIP Network Operations and Management Symposium | 978-1-6654-7716-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/NOMS56928.2023.10154366

***Abstract*—A Network Slice (NS) is a set of network resources deployed to deliver premium service to a group of users as described in a Service Level Agreement (SLA). We consider that premium service entails per user Quality of Service (QoS) requirements instead of aggregate NS metrics as in the majority of the literature. Since the users in a NS may run applications with different QoS requirements, aggregate metrics are inadequate. Their inefficiency is even more pronounced in Radio Access Network Slices (RANSs). In this case, even if all users run the same application, provisioning bandwidth based on aggregate metrics may violate the QoS requirements of users experiencing poorer channel conditions. To resolve this issue, we propose a method to compute the bandwidth required to deliver per user average packet delay guarantees in enhanced Mobile Broadband (eMBB) RANSs by dimensioning a multiclass queueing system. Our method considers a varying number of NS users over time and allows the online adaptation of the NS bandwidth to its current traffic load, which significantly limits bandwidth overprovisioning. We also describe in detail how an SLA is reached between the tenant and the Network Operator (NO) based on the cdf of the required NS bandwidth. We conduct simulations to show the need for per user guarantees and the efficiency of our method.**

***Index Terms*—network slicing, radio access networks, queueing theory, LTE, 5G, eMBB, 5QI, SLA**

I. Introduction

THE emergence of applications involving information transmission with strict QoS requirements has fueled the evolution of virtual networking. In this paradigm, companies can provide their end users with Quality of Service (QoS) guarantees by requesting the deployment of a virtual network tailored for their needs from a Network Operator (NO).

To fulfill such service requests, the NO and the requesting company first need to enter into a Service Level Agreement (SLA) that specifies the expected delivered QoS and the cost of the resulting virtual network. Companies that form SLAs with NOs are called tenants in the virtual networking paradigm.

The fast deployment and configuration of virtual networks has been enabled by Software Defined Networking (SDN) and Network Function Virtualization (NFV). Although the first major virtual networks centered around data center networking, the extension of virtual networking in cellular networks has recently attracted considerable attention due to the plethora of applications that require a cellular infrastructure.

Virtual networks that are deployed on a cellular network are often referred to as Network Slices (NSs). Unlike conventional virtual networks, NSs are composed of the Radio Access Network (RAN) part, the Transport Network (TN) part and the Core Network (CN) part. The deployment of NSs that guarantee the fulfillment of their corresponding SLAs is particularly challenging in the RAN due to the randomness of the wireless access channels. For brevity, we refer to the RAN part of a NS as a Radio Access Network Slice (RANS).

Recently, many works have addressed various aspects of NSs. In [1] and [2], the competition between NOs for tenant requests is discussed and auction-based algorithms are proposed. The authors in [3] study the competition between tenants while considering the Fisher Market mechanism and inelastic traffic. In [4], the trade-off between inter-slice and intra-slice fairness is addressed for NSs requiring heterogeneous resources. The joint scheduling of the enhanced Mobile Broadband (eMBB) and Ultra Reliable Low Latency Communications (URLLC) NSs using minislot puncturing is analyzed in [5]. In [6], the authors provide insights for provisioning the URLLC NS when the blocking probability is the metric of interest.

Regarding the CN part of NSs, the authors in [7] address the problem of multiplexing end-to-end NSs with reduced SLA violations. In [8], the authors derive a containerized IP Multimedia Subsystem (cIMS) and analyze its performance using a multiclass queueing framework where each class of requests follows a different arrival distribution. The authors compare the performance of the cIMS when each class is routed to a different Home Subscriber Server (HSS) with the resulting performance when a single HSS is used.

In the previous works, the considered QoS requirements for NSs are aggregate NS level requirements. However, the primary goal of a Network Slice (NS) is to deliver premium service to all of its users. However, dimensioning the RANS based on aggregate instead of per user requirements can be detrimental for the delivered service to some users.

As a motivating example, consider a NS consisting of multiple video call users experiencing good channel conditions and a single web-browsing user experiencing poor channel conditions. Suppose the NS is dimensioned based on a threshold for the average packet delay over all users. Then, although the constraint might be met since most traffic originates from users that enjoy good channel conditions, the web-browsing user might experience large delays. In this paper, we wish to

resolve such issues and fill this gap in the literature.

Our contributions are summarized as follows. First, we propose a method to compute the bandwidth needed by an eMBB NS with per user guarantees regarding average packet delays. This bandwidth is computed fast enough to allow its online adaptation to a varying number of NS users with different channel and traffic statistics. Second, we describe the information that needs to be exchanged between the tenant and the NO to form the SLA, where the cost of the NS is computed based on the cdf of the NS bandwidth.

Overall, our paper is structured as follows. In Section II, we present a model of the eMBB NS traffic consisting of users running applications with different QoS requirements and experiencing different channel conditions. We also consider that users connect to and disconnect from the NS throughout its lifetime. We link the variety of applications within an eMBB NS with the 5G Quality Indicators (5QIs) and base our model of a NS on a multiclass queueing system.

In Section III, we solve an optimization problem to compute the required NS bandwidth for a fixed number of users. In Section IV, we numerically compute the cdf of the required NS bandwidth from the user session statistics and describe the process of forming the SLA between the tenant and the NO. In Section V, we show the necessity of per user guarantees and the resource efficiency of our method through simulations. Lastly, Section VI concludes the paper.

II. EMBB Network Slice Traffic

User Sessions: We consider a single Base Station (BS) where eMBB users connect, stay connected for a random amount of time, and then disconnect. We consider that the tenant has registered $N_{\max}$ users to their eMBB slice. To provide a value for $N_{\max}$, we first consider the average number of users that are within the coverage area of a BS.

Using the data provided by a major network operator in [9], we divide the total number of subscribers by the total number of BSs in [9] to obtain 140 users per BS. Hence, we consider the typical number of registered users to the eMBB NS to be $N_{\max} = 50$, i.e., approximately one third of the total number of users within the coverage area of the BS.

Let $N(t)$ denote the number of users that are simultaneously active at time t at the BS. Clearly $N(t) \leq N_{\max}$. We provide a model for $N(t)$ as follows. First, we consider that each of the $N_{\max}$ users connects to the BS for $p_a = 10\%$ of the time. Hence, we consider that each user's holding times and idle times have expected values $T_h = 5$ and $T_i = (1 - p_a)T_h/p_a$.

We only require that the inter-arrival times and the holding times follow iid processes independent of each other. For simulation purposes and without loss of generality, we consider that the inter-arrival times follow the exponential distribution with mean $T_i = 45$ minutes. For the holding times, we consider a heavy tailed distribution such as the lognormal distribution with mean $T_h = 5$ and standard deviation $\sigma_h = 15$ minutes.

To compute $N(t)$, we need to compute how many user sessions overlap at the same time. Assuming that each user's activity is independent of the activity of any other user, the probability that at an arbitrary time t, there are $N = k$ eMBB users active at the same time is:

$$\Pr(N = k) \coloneqq p_k = \binom{N_{\max}}{k} p_a^k (1 - p_a)^{N_{\max} - k}. \quad (1)$$

Thus, $\mathbb{E}[N] = N_{\max} p_a = 5$, where in this model we assume that the inter-arrival and holding times of a user are independent of the type of application used. Next, we describe the packet level traffic for a fixed number of active NS users.

TABLE I
Modulation order M_i (bits per symbol) vs SNR γ_i (dB) [10].

γ_i	$-\infty$	-6.7	-4.7	-2.3	0.2	2.4	4.3	5.9
M_i	0.15	0.15	0.23	0.38	0.60	0.88	1.18	1.48
γ_i	8.1	10.3	11.7	14.1	16.3	18.7	21.0	22.7
M_i	1.91	2.41	2.73	3.32	3.90	4.52	5.12	5.55

Cellular Protocol: We assume that the 5G NR protocol is deployed. In 5G NR, the time and frequency resources are divided into Physical Resource Blocks (PRBs) of 1 ms duration and 180 KHz bandwidth. Within each PRB, $N_s = 168$ symbols are sent. All bandwidths are expressed in units of PRBs. Also, we assume that all users need to maximize the transmission power to be considered for admission to the NS. The eMBB slice is allocated a total of W PRBs. Each eMBB user transmits on all of the W PRBs, thus the user channel bandwidth is W.

Modulation and Coding Schemes: Reliability is a high priority in communications. Here, reliability refers to a Block Error Rate (BLER) target. To achieve a fixed BLER target, 5G NR adapts the modulation order[1] dynamically to adapt to the varying received SNR. Let $\mathcal{M}$ denote the set of available modulation and coding schemes. Given a fixed BLER target, let γ_i denote the required SNR to achieve a modulation order of $M_i \in \mathcal{M}$ bits per symbol. In Table I, we show the modulation order M_i and the SNR γ_i when BLER is 10% [10, Table 4.7], where we consider that $M_0 = 0.15$ is robust enough for any SNR within the coverage of the BS. The previous table has been used primarily for LTE, however we expect that similar values are used for 5G NR as well. In any case, NOs are free to consider different values for Table I.

Channel Model: We assume that each PRB experiences large and small scale fading due to path losses, user mobility and multipath propagation. We also assume that the small scale fading is slow and flat within a PRB. In 5G NR, the PRB dimensions are specified such that this holds in practice.

We also assume that the SNR difference between any two PRBs of the channel bandwidth is less than 2 dB since all PRBs belong in the same frequency band. Since the modulation scheme applied to a PRB changes in practice every 2 dB as implied by Table I, then for modulation assignment purposes, the SNR is roughly the same on all the W PRBs.

Let $\mathcal{U}$ denote the set of eMBB users in the NS, where $|\mathcal{U}| = N$. Let random variable $\gamma_u(\tau)$ denote the SNR of user u at subframe τ. Due to small scale fading, we consider that $\gamma_u(\tau)$ over multiple subframes $\tau \in T$ forms a random process of iid

[1]By modulation order, we refer to the number of bits per symbol.

random variables. Let $G_u(x)$ denote the cdf of $\gamma_u(\tau)$, $\forall\tau$. Without loss of generality, we assume Rayleigh fading, thus $G_u(x)$ is the cdf of the Exponential distribution. Then, $G_u(x)$ is fully characterized by its expected value $\overline{\gamma}_u$. We note that $\overline{\gamma}_u$ can be estimated quickly by the BS since $G_u(x)$ is sampled every 1 ms. Thus, $\overline{\gamma}_u$ is considered known and so is $G_u(x)$.

Packet Arrivals: We consider that each user u generates a packet of P_u bits. Let random variable T_u denote the packet inter-arrival time of user u. Here, we consider that the T_u follows the exponential distribution with expected value $\overline{T}_u$ and that the inter-arrival process $\{T_u(\tau)\}_{\tau\in T}$ is iid:

$$\Pr(T_u \leq x) = 1 - e^{x/\overline{T}_u}. \tag{2}$$

This model approximates the traffic by a wide variety of eMBB applications. For example, for a VoIP user, the packetization period is $\overline{T}_u = 20$ ms and the typical data payload is $P_u = 2$ Kb. Typically, low P_u values model delay sensitive applications such as phone calls, video calls or online gaming. High P_u values model delay insensitive ones such as large file transfers, texting and emails.

Packet Transmission Times: The transmission time S_u of a packet of P_u bits depends on the sum of the modulation orders of all the W PRBs utilized by each user. We approximate this sum with WM_u, where M_u is the modulation order achieved by the average SNR γ_u.

Given the above, for modulation order M_u and channel bandwidth W, then WN_sM_u bits can be transmitted per ms. Thus, the transmission time S_u in ms of P_u bits is $S_u = P_u/(WN_sM_u)$. The only random variable in the previous expression is M_u. Thus, $\Pr(S_u = S_i) = \Pr(M_u = M_i)$ The modulation order M_u depends on the SNR γ_u as shown in Table I. Since $\{\gamma_u(\tau)\}_{\tau\in T}$ are iid, then $\{S_u(\tau)\}_{\tau\in T}$ are also iid. Lastly, given that γ_u follows cdf $G_u(x)$:

$$\Pr(S_u = S_i) = G_u(\gamma_{i+1}) - G_u(\gamma_i). \tag{3}$$

Packet Delay Requirements. To satisfy the user's traffic, we wish to bound the expected overall time of a user's packet in the system by $\overline{d}_u$. Let random variable Q_u denote the queueing delay and random variable S_u denote the transmission time of a packet. Then, we require that:

$$\mathbb{E}[Q_u] + \mathbb{E}[S_u] \leq \overline{d}_u. \tag{4}$$

Multiclass Queueing System: Given the above, the packet traffic of the eMBB slice can be described as an $M/G/1$ multiclass queueing system. Each user u of the eMBB slice corresponds to a class of the system. Each class u is defined by its arrival process $\{T_u(\tau)\}_{\tau\in T}$ given by (2), its service process $\{S_u(\tau)\}_{\tau\in T}$ given by (3), and its QoS requirements $\overline{d}_u$ given by (4). The 1 ms granularity is enough to consider that the system operates in continuous time.

We note that each user's arrival process, service process and QoS requirements depend only by the following vector of variables $f_u = [\overline{\gamma}_u, \overline{T}_u, P_u, \overline{d}_u]$. For this reason, we call f_u the feature vector of user u. We also note that the feature vector f_u is immediately known to the BS once the user connects, with the exception of $\overline{\gamma}_u$, where we assume fast convergence of the sample mean monitored by the BS to the expected value.

TABLE II
eMBB Feature Statistics.

5QI	Application Group	R_{5QI}	T_{5QI}	P_{5QI}	d_{5QI}	a_{5QI}
1	Interactive Voice Calls	100 Kbps	20 ms	2 Kb	80 ms	8.6%
2	Interactive Video Calls	7 Mbps	20 ms	140 Kb	130 ms	12.4%
3	Online Gaming	15 Mbps	20 ms	0.3 Mb	30 ms	1.9%
4	Buffered Video	12 Mbps	100 ms	1.2 Mb	280 ms	13%
6	Web Browsing	5 Mbps	100 ms	0.5 Mb	280 ms	64.1%

Feature Vector Statistics. Regarding the statistics of the feature vectors, we consider that the feature vectors $f_{u\in\mathcal{U}}$ are iid. The features f_u of each eMBB user relate very closely with some of the QoS flow characteristics specified by the 5G QoS Indicators (5QIs) in 5G NR [11], formerly known as QoS Class Identifiers (QCIs) in LTE. For this reason, we consider the application groups in [11, Table 5.7.4-1] for selected 5QI values that correspond to eMBB traffic. Then, we consider that the support of the delay threshold $\overline{d}_u$ are the set of the Packet Delay Budgets (PDBs) in [11, Table 5.7.4-1].

Next, for the selected applications, we consider a suggested bitrate $R_{5\text{QI}}$ and a suggested mean packetization period $T_{5\text{QI}}$. The set of all packetization periods compose the support of T_u. Then, we compute the suggested packet lengths by $P_{5\text{QI}} = R_{5\text{QI}}T_{5\text{QI}}$ which describe the support P_u. Lastly, we assign a probability $a_{5\text{QI}}$ to each group based on the traffic analysis in [12]. All these values are shown in Table II.

The last feature to be described is the mean user SNR $\overline{\gamma}_u$. We consider that it follows the Normal cdf with mean $\overline{\gamma} = 10$ dB and $\sigma = 3$ dB, assuming that the BS is placed so that most users achieve a moderate modulation order in Table I.

Given the above, the random vector $f_u = [\overline{\gamma}_u, \overline{T}_u, P_u, \overline{d}_u]$ can be further divided into two vectors; the wireless channel conditions vector f_u^c and the application features vector f_u^a. For our model, $f_u^c = [\overline{\gamma}_u]$ and $f_u^a = [\overline{T}_u, P_u, \overline{d}_u]$.

In general, we expect that these two vectors to be independent of each other since the application a user uses does not affect their channel conditions. Also, in practice, we expect that the components of f_u^c to be continuous, whereas the components of f_u^a to be discrete. Indeed this is the case for our model. Since $f_u^c = [\gamma_u]$ and f_u^a depends only on the 5QI value of user u, we compute the following probability:

$$\Pr\left(f_u^c \leq \gamma \cap f_u^a = [T_{5\text{QI}}, P_{5\text{QI}}, d_{5\text{QI}}]\right) = \Phi_{10,9}(\gamma)a_{5\text{QI}}. \tag{5}$$

In conclusion, every time a new user connects according to the arrival process with mean T_s defined earlier, the user's feature vector is generated by (5).

III. Online Bandwidth Allocation

We wish to compute the required bandwidth W to satisfy the expected delay vector $\{\overline{d}_u\}_{u\in\mathcal{U}}$ of the $M/G/1$ multiclass queueing system described previously. Its computation depends highly on the service discipline. Here, we assume that FIFO is used, thus the computation of W is feasible.

Let $\lambda_u = 1/\overline{T}_u$. The overall arrival rate to the system is $\lambda = \sum_u \lambda_u$. Also, note that the probability that a received

packet at the BS originates from user u is $p_u = \lambda_u/\lambda$. Hence, by iterated expectation, the expected value of the overall packet transmission time is $\mathbb{E}[S] = \sum_u p_u \mathbb{E}[S_u]$. Similarly, the second raw moment is $\mathbb{E}[S^2] = \sum_u p_u \mathbb{E}[S_u^2]$, where $\mathbb{E}[S_u^2] = \mathrm{Var}[\mathrm{S_u}] + \mathbb{E}[\mathrm{S_u}]^2$ and $\mathrm{Var}[\mathrm{S_u}]$ is computed using (3). Lastly, let $\rho = \lambda \mathbb{E}[S]$.

Since the Poisson Arrivals See Time Averages (PASTA) theorem holds for each class [13], then the average waiting time in the queue is the same for all classes, i.e., $\mathbb{E}[Q_u] = \mathbb{E}[Q]$, $\forall u$ [13]. We compute $\mathbb{E}[Q_u]$ for $\rho < 1$ using the Pollaczek–Khintchine formula and Little's Law [14]:

$$\mathbb{E}[Q] = \mathbb{E}[Q_u] = \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)},\ \forall u \in \mathcal{U}. \tag{6}$$

Hence, to compute the minimum required bandwidth W that satisfies the QoS requirements of each user as in (4), we need to solve the following optimization problem:

$$\begin{aligned} \underset{W}{\text{minimize}} \quad & W \\ \text{subject to:} \quad & \frac{\lambda \mathbb{E}[S^2]}{2(1-\rho)} + \mathbb{E}[S_u] \le \overline{d}_u,\ \forall u \in \mathcal{U}. \end{aligned} \tag{7}$$

Note that to satisfy the overall user delay constraint, it is necessary that $\rho < 1$. Otherwise, the queueing system is unstable and the queueing delay becomes unbounded. Let $g_u(W)$ denote the left side of the constraint in (7), i.e., the overall packet delay of user u given bandwidth W. To solve optimization problem (7), we use the following:

Proposition 1. *In a multiclass $M/G/1$ FIFO queueing system, if each user service time S_u is decreasing w.r.t. W and each user arrival process does not depend on W, then the overall packet delay of each user is decreasing w.r.t. W, i.e., the function g_u is decreasing in $\mathcal{H} = \{W : \rho(W) < 1\}$.*
Proof. See Appendix A.

In our case, $S_u = P_u/(WN_sM_u)$ is decreasing w.r.t. W and the arrival process does not depend on W. Therefore, Proposition 1 is applicable to our system, thus our $g_u(W)$ is decreasing. Given its monotonicity, it immediately follows:

Corollary 1. *The unique optimal solution of optimization problem (7) is given by $W^* = \max_{u \in \mathcal{U}} \ g_u^{-1}(\overline{d}_u)$.*

Therefore, the optimal solution can be computed easily if a closed form expression of $g_u^{-1}(\cdot)$ is available.

Proposition 2. *The quantity $g_u^{-1}(\overline{d}_u)$ is computed as follows:*

$$g_u^{-1}(\overline{d}_u) = \frac{a\overline{d}_u + c_u + \sqrt{(a\overline{d}_u - c_u)^2 + 2b\overline{d}_u}}{2\overline{d}_u},\ \forall \overline{d}_u > 0,$$

where:

$$a = \sum_{u \in \mathcal{U}} \frac{P_u \mathbb{E}[M_u^{-1}]}{N_s \overline{T}_u}, b = \sum_{u \in \mathcal{U}} \frac{P_u^2 \mathbb{E}[M_u^{-2}]}{N_s^2 \overline{T}_u}, c_u = \frac{P_u \mathbb{E}[M_u^{-1}]}{N_s}.$$

Proof. See Appendix B.

Given all the above, the optimal solution W^* can be computed directly using Corollary 1 and Proposition 2. Since constants a, b and c_u depend only on the elements of the feature vectors $\{f_u\}_{u \in \mathcal{U}}$, the knowledge of the feature vectors is sufficient and necessary to compute W^*.

Fig. 1. The BS observes the user feature vectors $\{f_u\}_{u \in \mathcal{U}(t)}$ of the connected users $\mathcal{U}(t)$. Since the feature vector of each user remains constant throughout their connection, the BS observes them only when a user connects to the NS. The bandwidth demand of the NS $W^*(t)$ is computed by Proposition 2.

Considering that each element of f_u can be measured by the BS shortly after a user connects to the network, then we conclude that optimization problem (7) can be solved online. Therefore, the BS can adapt the eMBB NS bandwidth $W^*(t)$ online as the number of connected users $N(t)$ changes, while ensuring that the desired QoS of each connected user is met. Figure 1 illustrates the online adaptation procedure.

Adapting online bandwidth $W^*(t)$ to match the current traffic load of the eMBB slice results in a considerable increase of resource efficiency, since any bandwidth that is unused by the NS can be allocated to the rest of the regular traffic.

IV. Service Level Agreement

Given the previous analysis, it follows that $W^*(t)$ is a random process that depends on the random process $N(t)$ which denotes the number of connected users to the BS at time t, and the random vectors $f_{u \in \mathcal{U}(t)}$ which denote the feature vectors of the set of connected users $\mathcal{U}(t)$ at time t.

However, the tenant wishes that the desired QoS is almost always met. Hence, the NO needs to ensure that the available bandwidth to the NS denoted by W_r is greater than the required slice bandwidth $W^*(t)$ with high probability. Let $P_H = 0.9$ denote a constant close to 1. Since the NO needs fulfill the SLA, the following inequality needs to be satisfied:

$$\Pr(W^* \le W_r) \ge P_H. \tag{8}$$

Overall, the SLA between the NO and the tenant is formed as follows. First, the NO needs to know the number of registered users to the slice $N_{\max}$, the average user connection inter-arrival time T_i, the average user holding time T_h, and the statistics of feature vector f_u as described in Sec. II. We consider that this information is either immediately known to the tenant or it is observed during a trial period, where the NO creates a temporary slice for the tenant.

Second, given this information, the NO can compute the statistics of $N(t)$ and the required bandwidth $W^*(t)$ for a given number of connected users $N(t)$ as shown in Sec. II and in Sec. III. Third, the network operator needs to compute the bandwidth W_r that needs to be available at any moment to the NS, so that the probability of $W^*(t) > W_r$ is very small, i.e., the probability that the desired QoS of the eMBB slice is not met at an arbitrary time t is very small.

Fourth, once the bandwidth W_r is computed, the NO computes the cost of the requested NS based on W_r and charge

the tenant accordingly. If the cost is accepted by the tenant, then the SLA is formed. For the remainder of this section, we describe how W_r can be computed and thus focus primarily on the third step of the SLA process.

Since the SLA requires that $\Pr(W^* \leq W_r) \geq P_H$, where P_H is a constant close to 1, we need to compute the cdf of W^* to find W_r. The derivation of a closed form expression for this cdf is hard to obtain. Therefore, we compute it numerically through Monte Carlo simulations.

To do so, we can sample the process $W^*(t)$ over time. However, we note that the random process $W^*(t)$ is not memoryless since the service process is not memoryless. Thus, periodically sampling $W^*(t)$ may produce samples that are not independent. For this reason, instead of directly sampling $W^*(t)$, we proceed as follow.

Let F denote the cdf of W^*. Due to Corollary 1 and the law of total probability when conditioning on the number of connected users N, it follows:

$$F(w) = \sum_{k=1}^{N_{\max}} \Pr\left(\max_{u \in \mathcal{U}} g_u^{-1}(\bar{d}_u) \leq w | N = k\right) p_k. \quad (9)$$

Although the expressions of $g_u^{-1}(\bar{d}_u)$ are readily available from Proposition (2), a closed-form expression of the above probabilities is still hard to obtain. Thus, we compute them numerically through Monte Carlo simulations.

Let F_k denote the cdf of W^* for $N = k$. It follows that $F(w) = \sum_{k=0}^{N_{\max}} F_k(w) p_k$. Also, let $\hat{F}_{k,s_k}$ denote the empirical cdf of W^* for $N = k$ when the number of samples is s_k:

$$\hat{F}_{k,s_k}(w) \coloneqq \frac{1}{s_k} \sum_{i=1}^{s_k} \mathbf{1}_{X_i \leq w}, \ \forall k \in \{1, ..., N_{\max}\}, \quad (10)$$

where the random variables $X_1, ..., X_{s_k}$ are iid and follow the cdf of $\max_{u \in \mathcal{U}} g_u^{-1}(\bar{d}_u)$ for $|\mathcal{U}| = k$. Given the above, the estimated cdf of W^* is $\hat{F}(w) = \sum_{k=0}^{N_{\max}} \hat{F}_{k,s_k}(w) p_k$, where $\hat{F}_{0,0}(w) = 1$. In general, we wish that the estimated cdf $\hat{F}(w)$ deviates more than ϵ from the real cdf $F(w)$ w.l.p. for any w:

$$\Pr\left(\sup_w |F(w) - \hat{F}(w)| > \epsilon\right) \leq P_L. \quad (11)$$

To achieve this, we need to determine the number of samples s_k used for the computation of each empirical cdf $\hat{F}_k$.

Proposition 3. *If* $s_k = \left\lceil \frac{N_{\max} p_k}{2\epsilon^2} \ln \frac{2N_{\max}}{P_L} \right\rceil$, *then (11) holds. Proof. See Appendix C.*

This implies that a large number of samples is required for high accuracy. For instance, the number of samples is $\sum_k s_k \approx 69078$ for $\epsilon = (1 - P_H)/2 = 0.05$ and $P_L = 1 - P_H = 0.1$ and $N_{\max} = 50$. However, this sampling is performed only whenever a new NS request is received by the NO and thus it can be performed offline.

Specifically, the sampling procedure is conducted as follows. For each $1 \leq k \leq N_{\max}$, we perform the following two steps s_k times, where s_k is given by Proposition (3). First, we draw k feature vectors that follow the distribution in (5) and second, we compute the required bandwidth W^* using the closed-form expression given by Corollary (1) and Proposition (2). The fraction of the s_k values of W^* that are less than or equal to w equals $\hat{F}_{k,s_k}(w)$ as in (10). Thus, once the sampling procedure is completed, we can compute each $\hat{F}_{k,s_k}(w)$ and $\hat{F}(w) = p_0 + \sum_{k=1}^{N_{\max}} \hat{F}_{k,s_k}(w) p_k$ for any w.

Given the estimated cdf $\hat{F}$, we can numerically solve (8). Let $\mathcal{W}$ denote the set that contains all possible values of W_r. The set $\mathcal{W}$ is discrete since W_r is a natural number. Thus, the NO wishes to find bandwidth W_r^*:

$$W_r^* \coloneqq \min\{w \in \mathcal{W} : F(w) \geq P_H\}. \quad (12)$$

However, only $\hat{F}(w)$ can be computed, thus instead we find:

$$\hat{W}_r^* \coloneqq \min\{w \in \mathcal{W} : \hat{F}(w) \geq P_H + \epsilon\}, \quad (13)$$

where it is required that $\epsilon \leq 1 - P_H$. Note that from (11) and Proposition 3, it immediately follows:

Corollary 2. *If the NO provisions* $\hat{W}_r^*$ *bandwidth for the NS, the SLA is fulfilled w.h.p., i.e.,* $\Pr\left(F(\hat{W}_r^*) \geq P_H\right) \geq 1 - P_L$.

We note that the value of ϵ affects the optimality gap, i.e., the difference between $\hat{W}_r^*$ and W_r^*. Now, it remains to describe how $\hat{W}_r^*$ is found, i.e., how (13) is solved. To solve it, we consider that $\mathcal{W}$ is upper bounded by a quantity $W_{\max}$.

Although (13) can be solved even when $\mathcal{W}$ does not have an upper bound, the knowledge of an upper bound can reduce the time and space complexity of the solution algorithm. We can either consider that the largest sample out of all the $\sum_k s_k$ samples is w.h.p an upper bound of $\mathcal{W}$ or we can leverage (2) to derive an upper bound. Here, we follow the latter approach.

Proposition 4. *An upper bound of set* $\mathcal{W}$ *is:*

$$W_{\max} = \frac{N_{\max} P_3}{M_0 N_s T_3} + \sqrt{\frac{N_{\max} P_4^2}{2 M_0^2 N_s^2 T_4 d_3}}.$$

Proof. See Appendix D.

Thus, we consider the approximation $\mathcal{W} \approx [1, ..., W_{\max}]$. Clearly $|\mathcal{W}| = W_{\max}$. Now, (13) can be solved as follows. First, we obtain $\hat{F}(w)$, $\forall w \in \mathcal{W}$, with $\mathcal{O}(\sum_k s_k + |\mathcal{W}|^2)$ computations. Second, since the estimated cdf $\hat{F}(w)$ is increasing, we perform binary search on the sorted $\hat{F}(w)$ values and set $\hat{W}_r^*$ equal to the match for target value $P_H + \epsilon$, which requires $\mathcal{O}(\log |\mathcal{W}|)$ computations. Using the expression for $\sum_k s_k$ in Proposition 3, we have:

Corollary 3. *The provisioned bandwidth* $\hat{W}_r^*$ *can be found with* $\mathcal{O}\left(\epsilon^{-2} N_{\max} \ln\left(2 N_{\max} P_L^{-1}\right) + |\mathcal{W}|^2\right)$ *time complexity.*

Once the NO provisions $\hat{W}_r^*$ bandwidth for the NS, then the SLA is fulfilled w.h.p. as implied by Corollary 2. We note that at any time t, the excess bandwidth $\hat{W}_r^* - W^*(t)$ can be utilized by other users as long as these users release it immediately whenever the NS needs it.

However, the sharing of resources between multiple NSs while satisfying all their SLAs is more complex. Here, we consider that bandwidth $\hat{W}_r^*$ is exclusively used only by

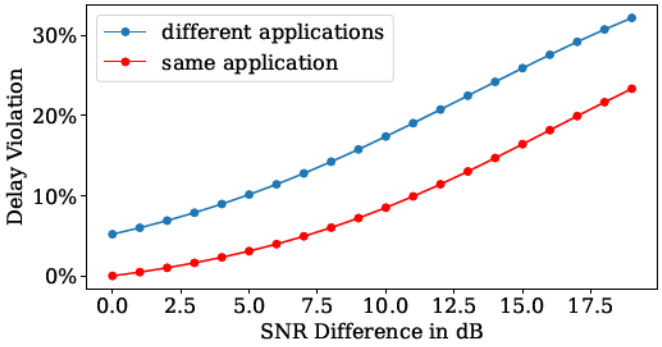

Fig. 2. First, consider a queueing system with 10 interactive video call users with $\overline{\gamma}_1 = 20$ dB and 1 online gaming user with $\overline{\gamma}_2 = \overline{\gamma}_1 - \epsilon$ dB. The video call users and the online gaming user require an average packet delay of at most 130 ms and 30 ms respectively. The different applications plot shows the delay violation for the online gaming user as the SNR difference ϵ increases when the NS is dimensioned such that the average packet delay over all users is less than 30ms. Second, we consider that all users are identical interactive video call users except from one user with $\overline{\gamma}_2 = \overline{\gamma}_1 - \epsilon$ dB. The same application plot shows the violation of the average packet delay requirement of the user with the low SNR as ϵ increases. Both plots show that although the strictest average packet delay requirement is met over all users, there is a user whose average packet delay requirement is violated.

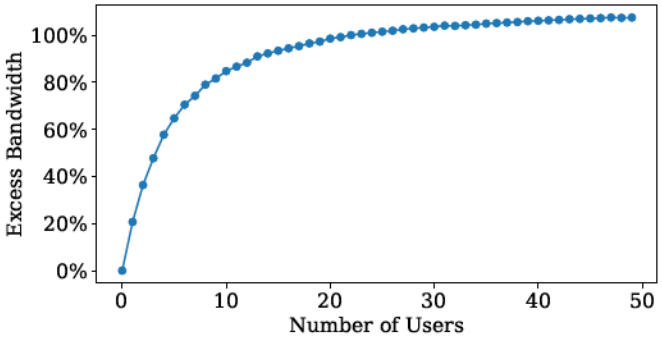

Fig. 3. Here, we consider a queueing system with N identical video calling users with $\overline{\gamma}_1 = 20$ dB and 1 video gaming user with $\overline{\gamma}_2 = 17$ dB. Suppose we approximate this multiclass queueing system by a single class system of $N + 1$ identical gaming users with $\overline{\gamma} = 17$ dB. Clearly, the bandwidth W_a required to satisfy the average packet delay of the single class queueing system is higher than the bandwidth W^* required to satisfy the per user average packet delay requirements of the original multiclass system. The plot shows the bandwidth overprovisioning $(W_a - W^*)/W^*$ as the number of video calling users N increases. The figure illustrates that the single class approximation increasingly overprovisions resources as the users increase.

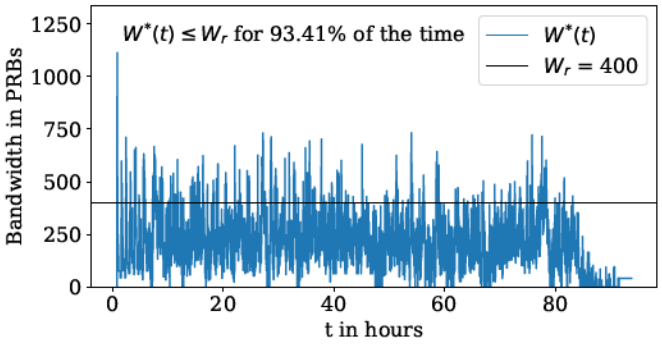

Fig. 4. The figure shows the bandwidth demand $W^*(t)$ of a NS with the default parameter values. We note that $W^*(t)$ is a piecewise constant random process that changes whenever a user connects to or disconnects from the NS. The system was run such that each of the $N_{\max} = 50$ users connects to and disconnects from the NS for 100 times resulting in a simulation time of 90 hours. This process $W^*(t)$ has high variability. The figure also shows that the provisioned bandwidth W_r indeed suffices to satisfy the SLA between the tenant and the NO. Lastly, the fast computation of $W^*(t)$ allows the utilization of the unused NS resources by the regular traffic of the cell. The unused resources form the space between the orange and the blue lines.

one NS. We note that resource exclusivity reduces resource efficiency. However it increases isolation, since traffic surges in one NS do not affect any other NS, even in the case where the surges are caused by DDoS attacks or other security breaches.

Therefore, each time a new NS request is received by the NO, the BS provisions $\hat{W}_r^*$ which is the maximum bandwidth that is available to NS at any time t. The cost of the SLA is then computed based on $\hat{W}_r^*$.

V. Simulation Results

In this section, we test various components of our proposed method and showcase their importance. All simulations were run on a home computer with an Intel i7-10700K processor using 16 GB of RAM running on Windows 10.

First, we wish to show the necessity of per user guarantees in RANSs. To do so, suppose that we instead consider the total average packet delay over all users as our aggregate metric and wish to bound it by the smallest delay bound $\overline{d} = \min_u \overline{d}_u$.

Then, the required bandwidth W_a can be easily computed by replacing c_u in Proposition 2 with $c = \sum_u c_u p_u$. For channel bandwidth W_a, we can compute the total packet delay of each user using the left side of the constraints in (7) and check for violations. In Fig. 2, we show the delay violation for a user in two different scenarios when the total average delay is used instead of per user guarantees.

Having established the importance of per user guarantees, we next illustrate the value of provisioning resources for a NS with per user guarantees by dimensioning a multiclass queueing system. To this end, suppose that instead of dimensioning a multiclass queueing system, we dimension a single class queueing system where the class used is formed by considering the worst case features over all the classes of the original multiclass queueing system. In Fig. 3, we show the bandwidth overprovisioning caused by following this approach.

Next, we begin the NS level simulations. To verify the validity of our data-driven method to obtain W_r, we simulate the random process $W^*(t)$ of a NS with the default parameter values as described in the previous sections of the paper. In Fig. 4, the random process $W^*(t)$ is depicted over a large period of time and the percentage of time that $W^*(t) \leq W_r$ is computed. In Fig. 5, we compute this percentage for multiple NSs that are parameterized randomly to verify the generality of our approach. Given that W_r requires a large amount samples to be computed, we also provide the run times of its computation in Fig. 6. Also, since we are not able to derive closed form expressions for W_r, we test its sensitivity w.r.t. the number of registered users $N_{\max}$ and the NS reliability requirement P_H in Fig. 7 and in Fig. 8 respectively.

Lastly, we conclude this section by providing motivation

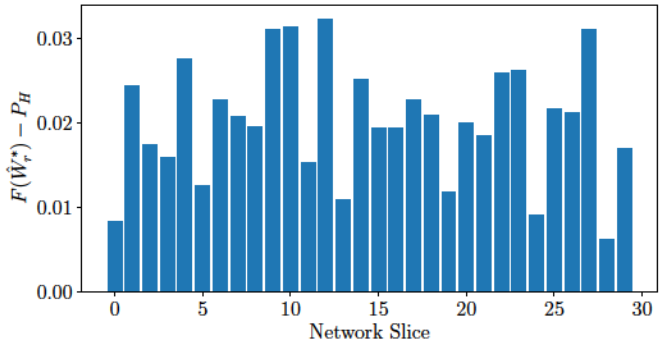

Fig. 5. The figure shows the difference $F(\hat{W}_r^*) - P_H$ for 30 NSs whose parameters T_h, σ_h, p_a, $N_{\max}$, P_H and P_L were chosen randomly. Since $F(\hat{W}_r^*) > P_H$, the SLA is fulfilled for each NS.

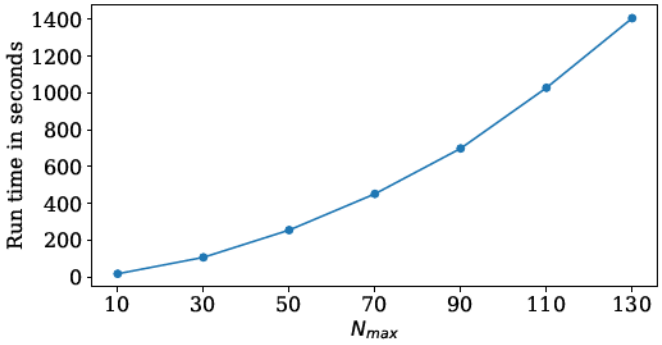

Fig. 6. The figure shows the run time of the computation of W_r as the number of registered users increases. The plot shows that although the computation of W_r is performed offline, the required computational time is less than 25 minutes even for large NSs. This implies that the NO can process new NS requests from tenants in a short period of time.

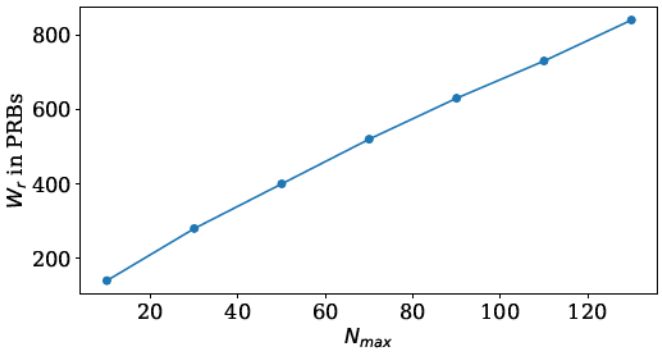

Fig. 7. For the default NS parameters, the figure shows the provisioned bandwidth W_r for a varying number of registered NS users $N_{\max}$. For the given region, the relation between W_r and $N_{\max}$ seems to be almost linear. This plot can be used by the NO to estimate the cost of registering extra users to an already deployed NS as requested by the tenant.

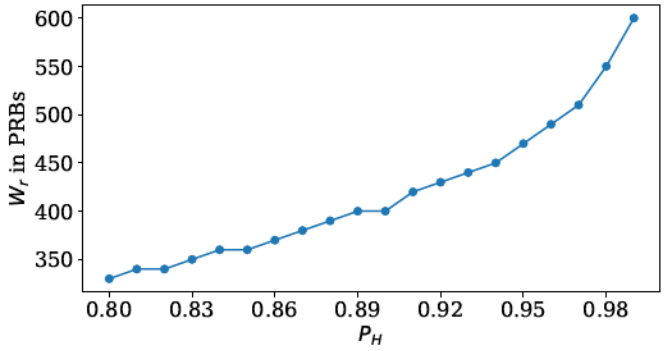

Fig. 8. The figures shows that the provisioned bandwidth W_r increases significantly for large reliability P_H. Since higher provisioned bandwidth W_r implies higher costs for the tenant, this plot is a trade-off curve for the tenant.

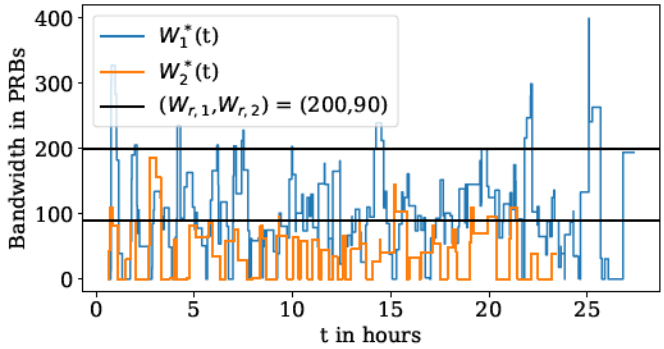

Fig. 9. The figure shows the bandwidth demands of two NSs over time. If resource sharing is not allowed, the bandwidth demand of NS 1 is not met whenever the blue line is higher than the upper black line. Similarly, the demand of NS 2 is not met whenever the orange line is higher than the lower black line. Note that there are time intervals where there is a demand violation for one NS that could be met by utilizing the idle resources of the other NS.

for enabling resource sharing between NSs. In this paper, we provided guarantees when the provisioned bandwidth W_r of each NS is not shared with other NSs. However, if multiple NSs are being served by the same BS, then the unused bandwidth of one NS can be allocated to another NS that is in need. The design of a scheduler that multiplexes NSs while satisfying each of their SLAs is left for future work. Here, we only illustrate the potential resource savings for two NSs as shown in Fig. 9 and in Fig. 10.

VI. Conclusion

In this paper, we showcased the importance of per user QoS guarantees for RANSs, where tenants and NOs form strict SLAs. To solve the bandwidth provisioning problem of a RANS, we related the NS to a multiclass queueing system, while considering channel fading under the operation of a cellular protocol. We showed that the NS traffic is a highly variable random process since users running different applications under different channel conditions connect to and disconnect from the NS multiple times throughout its lifetime. Nonetheless, we provided a data driven method to accurately estimate the cdf of the NS bandwidth demand. Based on this cdf, we derived the provisioned bandwidth that guarantees

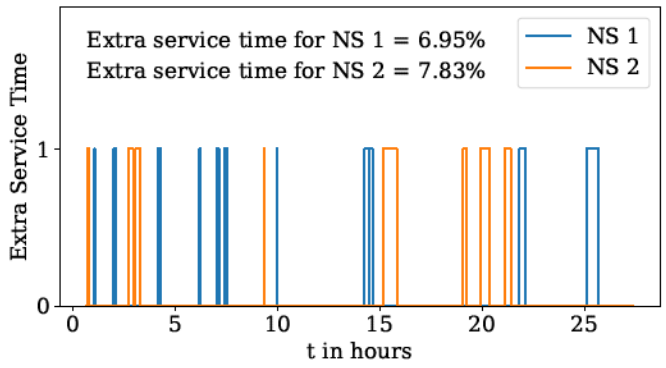

Fig. 10. This figure shows the intervals where multiplexing the NSs depicted in Fig. 9 would have resulted in extra service time. These intervals can be obtained from Fig. 9 by finding the time periods where there is a demand violation even though the overall bandwidth demand is less than the overall provisioned bandwidth. Thus, it suffices to check when $W_1^*(t)+W_2^*(t) \leq W_{r,1}+W_{r,2}$ and ($W_1^*(t) \leq W_{r,1}$ or $W_2^*(t) \leq W_{r,2}$). The figure indicates that multiplexing can considerably increase the service time of a NS.

the SLA fulfillment w.h.p. and described in detail the SLA formation process between the tenant and the NO.

For future work, we wish to loosen two key assumptions of our system model. First, we wish to consider packet arrival processes that are not exponential. Second, we wish to consider multiserver queueing systems since in practice each user's channel bandwidth is upper bounded by a threshold. Finally and most importantly, we wish to improve the resource efficiency of our method by deriving a scheduler that multiplexes NSs without violating any of their SLAs.

Appendix A
Proof of Proposition 1

Suppose that $S_u(W)$ is decreasing. Then, $\overline{S}_u(W)=\mathbb{E}[S_u]$ is decreasing as the sum of decreasing functions. Also, $S_u^2(W)$ is decreasing as the product of two positive decreasing functions. Hence, $\mathbb{E}[S_u^2]$ is decreasing w.r.t. W as the sum of decreasing functions.

It remains to show that the left term of $g_u(W)$ is decreasing. Hence, it suffices to show that $\frac{\lambda\mathbb{E}[S^2]}{2(1-\rho)}$ and $\mathbb{E}[S^2]$ are both decreasing w.r.t. W, considering that the product of two positive decreasing functions is also decreasing.

First note that $\rho(W)=\lambda\mathbb{E}[S]=\lambda\sum_u p_u\overline{S}_u(W)$ is a decreasing function of W as the sum of decreasing functions. Since $f(\rho)=1/(1-\rho)$ is increasing in $[0,1)$ and $\rho(W)$ is decreasing, then $(f\circ\rho)(W)$ is decreasing. Hence, the first factor is decreasing. Similarly as with $\mathbb{E}[S]$, the second factor $\mathbb{E}[S^2]$ is also decreasing. Thus, $g_u(W)$ is decreasing. □

Appendix B
Proof of Proposition 2

We need to obtain the expression for $g_u(W)$, solve for W and then evaluate at $d=\overline{d}_u$. This involves solving a quadratic equation where we accept only the largest root p_1 which satisfies $\lim_{d\to\infty} p_1(d)=W_{\min}$, where $\rho(W_{\min})=1$. For brevity, we omit these algebraic manipulations. By following these steps, we can obtain the expression in Proposition 2. □

Appendix C
Proof of Proposition 3

Given that the real cdf is $F(w)=\sum_{k=0}^{N_{\max}} F_k(w)p_k$ and the estimated cdf is $\hat{F}(w)=\sum_{k=0}^{N_{\max}} \hat{F}_{k,s_k}(w)p_k$, then it follows:

$$\sup_w |F(w)-\hat{F}(w)|>\epsilon \subseteq \sum_{k=0}^{N_{\max}} \sup_w |F_k(w)-\hat{F}_{k,s_k}(w)|p_k>\epsilon$$
$$\subseteq \bigcup_{k=0}^{N_{\max}} \sup_w |F_k(w)-\hat{F}_{k,s_k}(w)| > \frac{\epsilon}{N_{\max}p_k}. \tag{14}$$

In (14), the first "$\subseteq$" holds due to the triangle inequality and the fact that the supremum of the sum is less than or equal to the sum of the supremums of the summands. The second "$\subseteq$" holds since if the sum of $N_{\max}$ summands is greater than ϵ, then at least one summand is greater than $\epsilon/N_{\max}$.

From the first and third part of (14), we obtain:

$$\sum_{k=0}^{N_{\max}} \Pr\left(\sup_w |F_k(w)-\hat{F}_{k,s_k}(w)| > \frac{\epsilon}{N_{\max}p_k}\right) \leq P_L$$
$$\Rightarrow \quad \Pr\left(\sup_w |F(w)-\hat{F}(w)|>\epsilon\right) \leq P_L. \tag{15}$$

Thus, it suffices to satisfy the following inequality $\forall k$:

$$\Pr\left(\sup_w |F_k(w)-\hat{F}_{k,s_k}(w)| > \frac{\epsilon}{N_{\max}p_k}\right) \leq \frac{P_L}{N_{\max}}. \tag{16}$$

Using the Dvoretzky-Kiefer-Wolfowitz inequality [15], we obtain the number of samples s_k:

$$s_k=\left\lceil \frac{N_{\max}p_k}{2\epsilon^2}\ln\frac{2N_{\max}}{P_L}\right\rceil, \forall k\in\{1,...,N_{\max}\}. \quad \square$$

Appendix D
Proof of Proposition 4

From Proposition 2, it follows:

$$W^* \leq \max_{u\in\mathcal{U}}\left(\sqrt{\frac{b}{2\overline{d}_u}}+\max\left\{a,\frac{c_u}{\overline{d}_u}\right\}\right). \tag{17}$$

We upper bound the right side of (17) using the following:

$$a=\sum_{u\in\mathcal{U}}\frac{P_u\,\mathbb{E}[M_u^{-1}]}{N_s\overline{T}_u} \leq \frac{N_{\max}}{M_0N_s}\max_{5\text{QI}}\frac{P_{5QI}}{T_{5QI}}=\frac{N_{\max}P_3}{M_0N_sT_3},$$
$$\frac{c_u}{\overline{d}_u}=\frac{P_u\,\mathbb{E}[M_u^{-1}]}{N_s\overline{d}_u} \leq \frac{1}{M_0N_s}\max_{5\text{QI}}\frac{P_{5\text{QI}}}{d_{5\text{QI}}}=\frac{P_3}{M_0N_sd_3}. \tag{18}$$

Now, we combine (17) with (18). Since $P_3/T_3>P_3/d_3$:

$$W^* \leq \max_{u\in\mathcal{U}}\left(\sqrt{\frac{b}{2\overline{d}_u}}+\frac{N_{\max}P_3}{M_0N_sT_3}\right). \tag{19}$$

Next, we consider the following bounds:

$$b=\sum_{u\in\mathcal{U}}\frac{P_u^2\,\mathbb{E}[M_u^{-2}]}{N_s^2\overline{T}_u} \leq \frac{N_{\max}}{M_0^2N_s^2}\max_{5\text{QI}}\frac{P_{5\text{QI}}^2}{T_{5\text{QI}}}=\frac{N_{\max}P_4^2}{M_0^2N_s^2T_4}, \tag{20}$$
$$d_u \geq d_3.$$

Lastly, we combine (19) with (20) to obtain:

$$W^* \leq \frac{N_{\max}P_3}{M_0N_sT_3}+\sqrt{\frac{N_{\max}P_4^2}{2M_0^2N_s^2T_4d_3}}. \quad \square$$

REFERENCES

[1] Q. Qin and L. Tassiulas, "Auction-based network slicing architecture and experimentation on sd-rans," in *OpenWireless'20*. New York, NY, USA: ACM, 2020, p. 1–6.

[2] Q. Qin, N. Choi, M. R. Rahman, M. Thottan, and L. Tassiulas, "Network slicing in heterogeneous software-defined rans," in *IEEE INFOCOM 2020*, 2020, pp. 2371–2380.

[3] P. Caballero, A. Banchs, G. de Veciana, X. Costa-Pérez, and A. Azcorra, "Network slicing for guaranteed rate services: Admission control and resource allocation games," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6419–6432, 2018.

[4] J. Zheng and G. de Veciana, "Elastic multi-resource network slicing: Can protection lead to improved performance?" *2019 WiOpt*, pp. 1–8, 2019.

[5] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of urllc and embb traffic in 5g wireless networks," in *IEEE INFOCOM 2018*, 2018, pp. 1970–1978.

[6] C.-P. Li, J. Jiang, W. Chen, T. Ji, and J. Smee, "5g ultra-reliable and low-latency systems design," in *2017 EuCNC*, 2017, pp. 1–5.

[7] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Pérez, "Resource sharing efficiency in network slicing," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 909–923, 2019.

[8] M. Di Mauro and A. Liotta, "Statistical assessment of ip multimedia subsystem in a softwarized environment: A queueing networks approach," *IEEE Transactions on Network and Service Management*, vol. 16, no. 4, pp. 1493–1506, 2019.

[9] A. Nika, A. Ismail, B. Y. Zhao, S. Gaito, G. P. Rossi, and H. Zheng, "Understanding and predicting data hotspots in cellular networks," *Mobile Networks and Applications*, vol. 21, no. 3, pp. 402–413, 2016.

[10] A. Ghosh and R. Ratasuk, *Essentials of LTE and LTE-A*, 1st ed. USA: Cambridge Univ. Press, 2011.

[11] 3GPP, "System architecture for the 5g system (5gs)," 3rd Generation Partnership Project (3GPP), Tech. Spec. (TS) 23.501, 2021.

[12] C. Cullen, "2020 mobile internet phenomena report," Sandvine, Feb. 2020. [Online]. Available: https://www.sandvine.com/phenomena

[13] O. J. Boxma and T. Takine, "The m/g/1 fifo queue with several customer classes," *Queueing Syst. Theory Appl.*, vol. 45, no. 3, p. 185–189, nov 2003. [Online]. Available: https://doi.org/10.1023/A:1027306700614

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

[15] P. Massart, "The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality," *The Annals of Probability*, vol. 18, no. 3, pp. 1269 – 1283, 1990. [Online]. Available: https://doi.org/10.1214/aop/1176990746