Delay-Guaranteed Resource Allocation for Deterministic Communications: An Efficient Stochastic Network Calculus Method

Juan Zhu and Shaowei Wang
School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China
Email: juanzhu@smail.nju.edu.cn, wangsw@nju.edu.cn

GLOBECOM 2023 - 2023 IEEE Global Communications Conference | 979-8-3503-1090-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/GLOBECOM54140.2023.10437275

***Abstract*—Deterministic communications systems are critical for time-sensitive applications in the Internet of Things, which demand stringent delay requirements under the conditions of limited available resources. Though network slicing can provide best-effort services by focusing on average performance metrics, it generally cannot address the worse-case issues arising from the deterministic communication scenarios. In this paper, we propose an efficient inter-slice radio resource allocation scheme for mobile networks to provide delay-guaranteed services, where a stochastic network calculus model is introduced to analyze the service delay and its variation. We derive a tight and time-invariant upper bound of the delay violation probability, and develop an efficient resource allocation algorithm to meet the stringent delay requirement of different radio slices. Numerical results demonstrate our proposal achieves a promising trade-off between resource utilization and delay.**

***Index Terms*—Deterministic communications, Internet of Things, radio resource allocation, stochastic network calculus.**

I. Introduction

With the proliferation of devices and their associated data demand boosted by the Internet of Things (IoT), 5G and beyond mobile networks are envisioned to support high-quality connectivity for both human-centric and machine-type services [1]. In particular, deterministic communications systems have emerged as potential solutions to provide end-to-end performance guarantees in both the core network and the radio access network (RAN) of mobile networks, which are especially critical for real-time applications such as motion control, industrial automation and robotics where any unexpected delay or its variation would lead to serious consequences [2]. Compared to the core network equipped with powerful computing and caching components, ensuring reliable and predictable operations in the RAN is more challenging due to the limited radio resources, the highly diversified and stringent performance requirements, and the harsh radio propagation environments.

In [3], a multigroup analytical framework is developed for massive random access of machine-to-machine communications in the IoT, where the access behavior is modeled as a double-queue with tunable device backoff parameters to guarantee the mean access delay. RAN slicing has also been identified as a promising technology to efficiently handle highly diverse delay requirements of the IoT applications in mobile networks, which allows multiple self-contained logical subnets, i.e., the RAN slices run on top of the same physical infrastructure so as to provide slice-specific services. In [4] [5], online convex optimization frameworks are proposed to learn the instant resource allocation from the experience data, which achieves low packet loss ratios with given delay budgets.

However, communication networks always deal with stochastic service requests and suffer from the severe fadings of radio channels in practice [6]. As a result, the RAN slicing schemes, which generally focus on average performance metrics and serve users in a best-effort manner, cannot meet the delay requirements of the IoT devices all the time. To address this problem, network planning and scheduling for the worst-case scenarios have been studied recently. In [7], a closed-loop load frequency control scheme is developed, which considers the worst case of the transmission delay. These worst-case scenarios, however, are oversimplified and rarely happen in practice, so the derived deterministic performance bounds are too conservative to utilize the scarce network resources efficiently.

To achieve a balance between resource utilization and performance assurance, effective capacity concept has been introduced to analyze statistical delay bounds. In [8], moreover, a decentralized method is proposed to deal with the resource allocation problem in a heterogeneous wireless network, where the effective capacity is applied to convert the statistical delay constraints into equivalent average rate constraints. However, the bounds derived from effective capacity-based approaches heavily rely on the choice of model parameters, which are difficult to determine in advance. In contrast with the effective capacity theory, stochastic network calculus (SNC) is a more flexible analysis framework that can provide reliable and reproducible statistical performance bounds by transforming complex queueing systems to analytically tractable linear ones with alternate algebras [9]. In [10], a bisection search algorithm is employed to minimize the total transmit power under given delay requirements, where a closed-form upper bound for the

This work was supported in part by the National Natural Science Foundation of China under Grants 61931023 and U1936202.

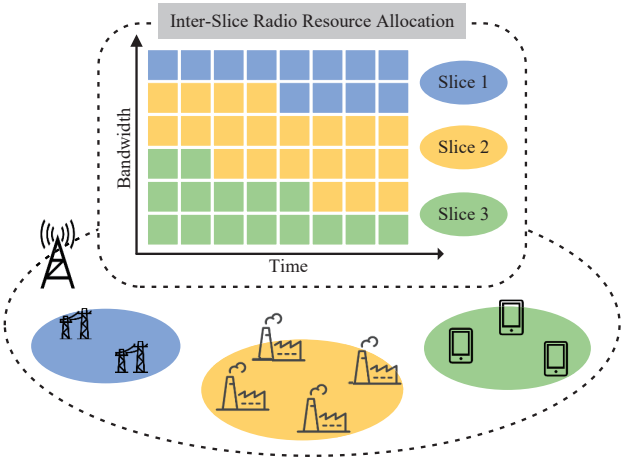

Fig. 1. Inter-slice radio resource allocation.

delay probability exceeding a given threshold is worked out by using the SNC method. In [11], another SNC-based model is proposed to compute the number of radio resources assigned to each RAN slice while keeping the delay within acceptable levels.

In this paper, we investigate the radio resource allocation problem in the IoT deterministic communications scenarios, where an efficient SNC model is employed to guarantee the diverse delay requirements of RAN slices in statistics. We first derive a tight and time-invariant upper bound of delay violation probability by using *Doob*'s martingale inequality [12] with independent and identically distributed (i.i.d.) increments. In addition, we analyze the delay variation and further improve the system robustness by restricting the delay variation bounds within a specified range. Finally, we design an efficient inter-slice radio resource allocation algorithm that achieves a trade-off between resource utilization and delay stability.

II. Delay Model and Problem Formulation

We focus on a single-cell RAN with a set of slices $\mathcal{I}$, as illustrated in Fig. 1, where each RAN slice provides a customized network service for an aggregated traffic of multiple users. A RAN slice orchestrator periodically executes inter-slice scheduling procedure to decide the dedicated radio resource quota for each slice so as to meet the delay requirements in the long term.

A. Delay Model

Consider a time-slotted, fluid-flow queuing system with an infinite buffer as shown in Fig. 2. For RAN slice $i \in \mathcal{I}$, denote by $a_i(u)$ and $s_i(u)$ the instantaneous arrival and service increments in the uth timeslot, respectively. The cumulative traffic arrival and service processes of slice i in the time interval $[\tau, t)$ are defined by bivariate functions $A_i(\tau,t) = \sum_{u=\tau}^{t-1} a_i(u)$ and $S_i(\tau,t) = \sum_{u=\tau}^{t-1} s_i(u)$, respectively. A service description $S_i(\tau,t)$ is referred to as a dynamic server if its corresponding departure process D_i satisfies the following inequality for all arrival process $A_i(\tau,t)$ [13]:

$$D_i(\tau,t) \geq \inf_{\tau \leq u \leq t} \{A_i(\tau,u) + S_i(u,t)\}. \tag{1}$$

Without loss of generality, we assume that $A_i(\tau,t)$ and $S_i(\tau,t)$ are both i.i.d. incremental processes, i.e., the instantaneous arrival and service increments $a_i(\cdot)$ and $s_i(\cdot)$ are i.i.d. across different timeslots. In the SNC analysis, a queueing node is unstable if its delay grows over time and becomes unbounded. A stability condition ensuring the queueing delay finite at all times in the i.i.d. increments scenarios can be expressed as follows:

$$\mathbb{E}[a_i(\cdot)] < \mathbb{E}[s_i(\cdot)], \tag{2}$$

where $\mathbb{E}[\cdot]$ denotes the expectation of a random variable.

In first-come-first-served order, the queueing delay $W_i(t)$ at time t is defined as the time it takes for all data that arrived prior to time t to depart from the transmit buffer and reach the receiver:

$$W_i(t) \triangleq \inf\{u \geq 0 : A_i(0,t) \leq D_i(0,t+u)\}. \tag{3}$$

We use the delay violation probability to measure the robustness of a communications system with respect to the deadline w_i^{th}, which is defined as the probability that the random delay $W_i(t)$ exceeds the delay deadline w_i^{th} at any time t:

$$p_i(w_i^{th}, t) \triangleq \mathbb{P}\{W_i(t) > w_i^{th}\}. \tag{4}$$

An SNC model introduced in [14] provides a tractable upper bound $q_i^{up}(w,t)$ of delay violation probability for RAN slice i by using the well-known union bound and *Chernoff*'s inequality:

$$q_i^{up}(w_i^{th}, t) = \inf_{\theta>0} \left\{ M_{s_i}^{w_i^{th}}(-\theta) \frac{1 - M_{a_i}^{t+1}(\theta) M_{s_i}^{t+1}(-\theta)}{1 - M_{a_i}(\theta) M_{s_i}(-\theta)} \right\}, \quad \text{s.t. } M_{a_i}(\theta) M_{s_i}(-\theta) < 1, \tag{5}$$

where $M_{a_i}(\theta) = \mathbb{E}[e^{\theta a_i}]$ and $M_{s_i}(-\theta) = \mathbb{E}[e^{-\theta s_i}]$ are the moment generating functions of random arrival increments $a_i(\cdot)$ and random service increments $s_i(\cdot)$ for RAN slice i, respectively. It is worth noting that (2) is a necessary condition for the constraint $M_{a_i}(\theta) M_{s_i}(-\theta) < 1$ to hold, which ensures that the moments of the arrival and service processes are well-behaved and that the derived bound is valid.

We derive a time-invariant delay violation probability bound $p_i^{up}(w_i^{th})$ with *Doob*'s martingale inequality [12] as stated in Theorem 1, which is tighter than $q_i^{up}(w_i^{th}, t)$ in (5). The rationale behind this improvement can be found through (2), which emphasizes that for a queueing system to remain stable, the average arrival rate should be strictly less than the service rate. When the positivity constraint imposed on the buffer

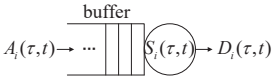

Fig. 2. A single queueing node with arrival $A_i(\tau,t)$, service $S_i(\tau,t)$ and departure $D_i(\tau,t)$ processes.

is temporarily disregarded, its expected increment becomes negative. This leads to the behavior of the buffer's content resembling that of a supermartingale. Consequently, the application of *Doob*'s martingale inequality becomes effective in deriving a stringent bound.

Theorem 1. *For RAN slice i, consider a traffic arrival process $A_i(\tau,t)$ with i.i.d. increments $a_i(u)$, and $A_i(\tau,t)$ is served by an i.i.d. incremental service process $S_i(\tau,t)=\sum_{u=\tau}^{t-1}s_i(u)$. Then the delay violation probability at any time can be upper bounded by*

$$\begin{aligned} p_i^{up}(w_i^{th}) &= \inf_{\theta>0} M_{s_i}^{w_i^{th}}(-\theta), \\ &s.t.\ M_{a_i}(\theta)M_{s_i}(-\theta)<1. \end{aligned} \tag{6}$$

Proof: Please refer to Appendix. □

It can be verified that $M_{s_i}^{w_i^{th}}(-\theta)$ is monotonically decreasing with θ and its infimum can be obtained at:

$$\theta^* = \sup\{\theta : M_{a_i}(\theta)M_{s_i}(-\theta)<1\}. \tag{7}$$

By analyzing the behavior of the function $M_{a_i}(\theta)M_{s_i}(-\theta)$ with respect to θ, we can conclude that a unique solution for $M_{a_i}(\theta)M_{s_i}(-\theta)=1$ exists within its monotonically increasing interval. As a result, a simple bisection search technique can be employed to quickly determine $p_i^{up}(w_i^{th})$, which is more computationally efficient compared to finding $q_i^{up}(w_i^{th},t)$ in (5).

To show the predictable performance, we further investigate the distribution of delay $W_i(t)$ by examining its second-order moment. Specifically, we define the delay variation $V_i(t)$ at time t as:

$$V_i(t)=\sqrt{\mathbb{E}[W_i^2(t)]}. \tag{8}$$

The metric measures the spread or dispersion of the delay times. Obviously, a lower value of $V_i(t)$ indicates a more stable delay.

To estimate $V_i(t)$, we define a random variable $W_i^{'}(t)$, such that for an arbitrary positive integer w,

$$\mathbb{P}\{W_i^{'}(t)>w\}=p_i^{up}(w), \tag{9}$$

and

$$\begin{aligned} \mathbb{P}\{W_i^{'}(t)=w\} &= \mathbb{P}\{W_i^{'}(t)>w\}-\mathbb{P}\{W_i^{'}(t)>w+1\} \\ &= p_i^{up}(w)-p_i^{up}(w+1). \end{aligned} \tag{10}$$

As $p_i^{up}(w)$ is the upper bound of the delay violation probability with respect to w, it follows that $\mathbb{P}\{W_i^{'}(t)>w\}\geq\mathbb{P}\{W_i(t)>w\}$, which indicates that variables $W'(t)$ and $W(t)$ are statistically ordered. Therefore, their second-order moments are also ordered [15],

$$\mathbb{E}[W_i^2(t)]\leq\mathbb{E}[W_i^{'2}(t)]. \tag{11}$$

Then we can derive a time-invariant upper bound V_i^{up} of the delay variation:

Theorem 2. *For RAN slice i, consider a traffic arrival process $A_i(\tau,t)$ with i.i.d. increments $a_i(u)$, and $A_i(\tau,t)$ is served by an i.i.d. incremental service process $S_i(\tau,t)=\sum_{u=\tau}^{t-1}s_i(u)$. Then the upper bound $V_i^{up}(t)$ of the delay variation defined in (8) can be derived as*

$$\begin{aligned} V_i^{up} &= \sqrt{\mathbb{E}[W_i^{'2}(t)]} \\ &= \frac{\sqrt{M_{s_i}(-\theta^*)(1+M_{s_i}(-\theta^*))}}{1-M_{s_i}(-\theta^*)}, \end{aligned} \tag{12}$$

where θ^* is given by (7).

Proof: Please refer to Appendix. □

B. Problem Formulation

We adopt a compound Poisson process to characterize the aggregated traffic of each RAN slice and the traffic process of RAN slice i is given by

$$A_i(0,t)=\sum_{n=1}^{N_i(t)}L_i(n), \tag{13}$$

where $N_i(t)$ is a counting of an independent Poisson process with rate λ_i and can be regarded as the cumulative number of arrived users up to time t; $L_i(\cdot)$'s are i.i.d. random variables that represent the data amount of users for RAN slice i. In this case, the process $A_i(0,t)$ is a Lévy process with independent stationary increments. The moment generating function of the arrival increment $a_i(\cdot)$ can be written as:

$$M_{a_i}(\theta)=e^{\lambda_i(M_{L_i}(\theta)-1)}, \tag{14}$$

where $M_{L_i}(\theta)$ is the moment generating function of $L_i(\cdot)$.

Denote by γ the signal-to-noise ratio of channel, the achievable transmission rate for RAN slice i is

$$s_i(t)=B_i\log_2(1+\gamma), \tag{15}$$

where B_i is the channel bandwidth. For block fading channels, $s_i(\cdot)$'s are i.i.d. across different timeslots. The moment generating function of the service increment $s_i(\cdot)$ for slice i can be expressed as

$$M_{s_i}(-\theta)=\mathbb{E}[(1+\gamma)^{\frac{-B_i\theta}{\ln 2}}]. \tag{16}$$

The upper bounds of the delay violation probability and delay variation for RAN slice i are obtained by substituting (14) and (16) into (6) and (12), respectively.

The bandwidth resources are organized into resource blocks (RBs) and RB allocation is denoted by $\boldsymbol{\mathcal{R}}=(\mathcal{R}_1,...,\mathcal{R}_i,...,\mathcal{R}_{|\mathcal{I}|})$, where $\mathcal{R}_i$ is the set of available RBs for the RAN slice i during each timeslot. To accommodate the diverse delay requirements of all slices, we try to minimize the maximum of the delay violation probability bounds of RAN slices while satisfying both the total resources and the delay variation constraints. When the RAN can not satisfy the delay requirements of all slices due to the limited RBs, the RAN slice orchestrator should prioritize the slices, where the slice with the higher priority would achieve a lower delay violation probability bound. To this end, we assign a priority φ_i to slice i as a tunable hyperparameter from the viewpoint of economics.

Applying Theorem 1, we define the priority-delay probability f_i for slice i as follow:

$$f_i = \begin{cases} \varphi_i \cdot \lg(p_i^{up}(w_i^{th})), & \mathbb{E}[a_i(\cdot)] < \mathbb{E}[s_i(\cdot)], \\ \varphi_i, & \text{otherwise.} \end{cases} \tag{17}$$

A balance among the delay requirements of all slices can be achieved by solving the following min-max problem:

$$\min_{|\mathcal{R}_1|,...,|\mathcal{R}_{|\mathcal{I}|}|} F = \max(f_1, ..., f_i, ..., f_{|\mathcal{I}|}), \tag{18a}$$

$$\text{s.t.} \sum_{i \in \mathcal{I}} |\mathcal{R}_i| \leq |\boldsymbol{\mathcal{R}}|, \tag{18b}$$

$$V_i^{up} \leq V_i^{th}, \quad \forall i \in \mathcal{I}, \tag{18c}$$

where V_i^{th} is the pre-specified maximum allowable delay variation of RAN slice i.

III. Delay-guaranteed Radio Resource Allocation

The objective function (18a) is non-convex and hard to deal with in general. We develop an efficient greedy algorithm that constructs promising solutions by selecting as good as possible local solutions in each step. Consider a single step and setting the delay variation constraint (18c) aside, we can redistribute RBs between two slices $i^{'}$ and $i^{''}$ while holding the stability condition:

$$\begin{aligned} \min_{|\mathcal{R}_{i'}|,|\mathcal{R}_{i''}|} & \max(\varphi_{i'} \cdot \lg p_{i'}^{up}, \varphi_{i''} \cdot \lg p_{i''}^{up}), \\ \text{s.t. } & |\mathcal{R}_{i'}| + |\mathcal{R}_{i''}| = R. \end{aligned} \tag{19}$$

The delay violation probability bound $p_i^{up}(\cdot)$ in (6) is a monotonically decreasing function of the allocated RBs $|\mathcal{R}_i|$ since more available bandwidth leads to less traffic jam and lower delay. As a result, the sufficient and necessary condition of the optimal solution to (19) is given by

$$\varphi_{i'} \cdot \lg p_{i'}^{up} = \varphi_{i''} \cdot \lg p_{i''}^{up}. \tag{20}$$

This condition guides the search for the optimal solution to (19), which offers the local best choice in a single step for (18). The details of our proposed delay-guaranteed inter-slice radio resource allocation procedure are summarized in Algorithm 1.

Initially, the available RBs are equally distributed among all RAN slices. The parameter m is used to track the number of consecutive iterations that none of the RBs is redistributed. In the subsequent steps (lines 2-22), RBs are reallocated from slice $i^{''}$ with the lowest priority-delay probability to the slice $i^{'}$ with the highest priority-delay probability. The array of priority-delay probabilities of all slices is sorted in ascending order to determine the slice $i^{'}$ and the slice $i^{''}$ (line 3). The inner *while* loop calculates the local optimal RBs allocation according to (20) while taking the delay variation constraints into account (lines 4-9). To avoid repeated reallocations between two slices, the last RB allocation in the loop is withdrawn to ensure that the priority-delay probability of slice $i^{''}$ is still higher than that of slice $i^{'}$, except for two cases: 1) slice $i^{'}$ has higher priority than slice $i^{''}$, and 2) the stability condition in slice $i^{'}$ changes from unsatisfied to satisfied while the stability condition slice $i^{''}$ remaining satisfied during the reallocation of the last RB (lines 10-12). The value of the objective function in (18a) is then updated (line 14). Finally, we check if the value of this function has decreased with respect to its value in the previous iteration (lines 15-20). If *yes*, the algorithm proceeds to the next iteration. If *no*, it implies that slice $i^{''}$ has no extra RBs for slice $i^{'}$, and we select the slice with the highest priority-delay probability among the remaining slices to donate RBs to slice $i^{'}$.

Algorithm 1: Radio Resource Allocation for Deterministic Communications

Initialization: Equal distribution of RBs among the RAN slices. Set the maximum iterations $N = |\mathcal{I}|^2$. Set $n = 0$ and $m = 0$;
1 Calculate f_i by (17) and evaluate F by (18a);
2 **while** $n < N$ *and* $m < |\mathcal{I}|$ **do**
3 Set $a = \text{sort}(f_1, ..., f_i, ..., f_{|\mathcal{I}|})$. Select $i^{'} = \arg(a(-1))$ and $i^{''} = \arg(a(m))$;
4 Calculate $f_{i'} - f_{i''}$;
5 **while** $f_{i'} - f_{i''} > 0$ *and* $V_{i'}^{up} \leq V_{i'}^{th}$ *and* $V_{i''}^{up} \leq V_{i''}^{th}$ **do**
6 Set $prev\_f_{i'} = f_{i'}$;
7 Set $|\mathcal{R}_{i'}| = |\mathcal{R}_{i'}| + 1$ and $|\mathcal{R}_{i''}| = |\mathcal{R}_{i''}| - 1$;
8 Calculate $V_{i'}^{up}$, $V_{i''}^{up}$ and $f_{i'} - f_{i''}$. See (12) and (17);
9 **end**
10 **if** *not* $(\varphi_{i'} > \varphi_{i''})$ *or not* $(prev\_f_{i'} > 0$ *and* $f_{i'} < 0$ *and* $f_{i''} < 0)$ **then**
11 Set $|\mathcal{R}_{i'}| = |\mathcal{R}_{i'}| - 1$ and $|\mathcal{R}_{i''}| = |\mathcal{R}_{i''}| + 1$;
12 **end**
13 Set $prev\_F = F$ and evaluate F by (18a);
14 **if** $F < prev\_F$ **then**
15 Set $m = 0$.
16 **end**
17 **else**
18 Set $m = m + 1$;
19 **end**
20 $n = n + 1$;
21 **end**
return: $\mathcal{R}_i \quad \forall i \in \mathcal{I}$

IV. Numerical Results

Assuming that $L_i(\cdot)$ follows a Poisson distribution with parameter ρ_i, and the values of ρ_i and λ_i in (14) are randomly generated. We compare our derived delay violation probability bound $p_i^{up}(\cdot)$ with the bound $q_i^{up}(\cdot)$ in (5) from [14] to demonstrate its effectiveness. We compare our proposed algorithm with two reference solutions related to the primary factors: traffic demand and delay deadline. Reference solution 1 involves allocating RBs in proportion to the mean traffic demand of each slice while reference solution 2 allocates RBs inversely proportional to the required delay deadline.

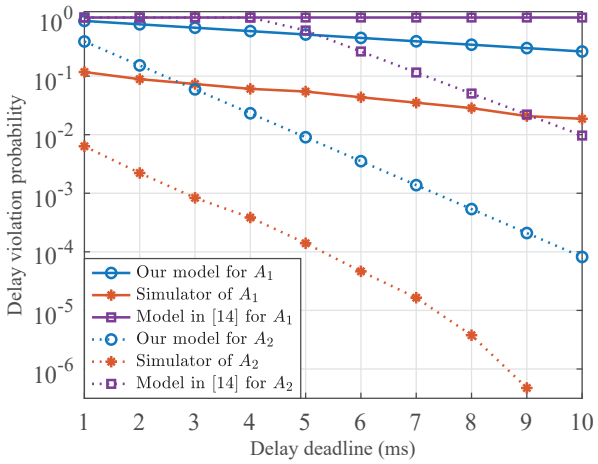

Fig. 3. Delay violation probability as a function of the required delay deadline for two different traffic processes.

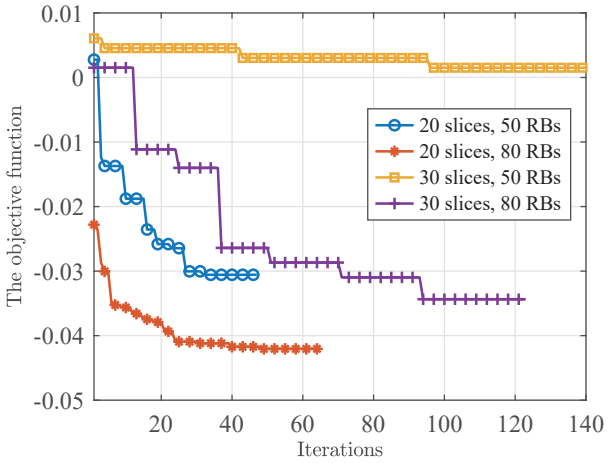

Fig. 5. Convergence of proposed algorithm.

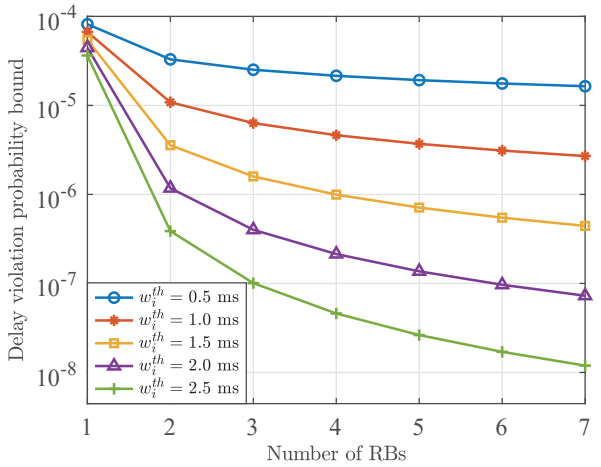

Fig. 4. Delay violation probability bound as a function of the number of allocated RBs with respect to different required deadlines.

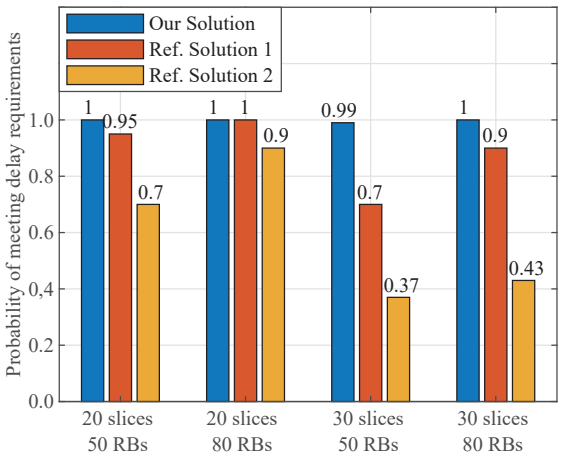

Fig. 6. Probabilities that the delay requirements of RAN slices are met.

A. Validation of the Proposed SNC-based Model

Fig. 3 shows the simulated delay violation probability and the derived upper bounds for two different traffic processes denoted by A_1 and A_2. A_1 has a larger mean than A_2. Our model consistently provides valid upper estimations of delay violation probabilities for given specific delay deadlines. This demonstrates the effectiveness of our model to calculate the required number of RBs for RAN slices to ensure their delay violation probabilities remain below a given value. In contrast, model in [14] often produces invalid upper bounds, i.e., $q_i^{up}(\cdot) = 1$.

Fig. 4 displays the delay violation probability bound as a function of the number of RBs assigned to RAN slice i for different delay deadlines w_i^{th}. As the number of allocated RBs increases, the derived delay violation probability bounds monotonically decrease with a decreasing rate, indicating that achieving a deterministic zero-violation delay bound is resource-intensive and impractical. Instead, statistical bounds are more reasonable and efficient.

B. Performance Analysis

In Fig. 5, the convergence of the proposed algorithm is shown under various scenarios with different amounts of slices and RBs. The curves head down until the algorithm converges. When the total radio resources are insufficient to meet the traffic demands of all RAN slices, the curve converges to a positive value, indicating that no enough RBs are available for the slices with the lowest priorities to satisfy the stability condition.

Fig. 6 shows the probabilities that the RAN slices satisfy their delay requirements. We can see that our proposed algorithm outperforms the reference solutions which do not take all relevant factors into account. Significant gains of our algorithm can be found when there are insufficient radio resources to support the delay-guaranteed services of all slices, indicating that our proposal can achieve a trade-off between resource utilization and delay guarantee.

V. Conclusions

In this paper, we presented an efficient SNC-based model for inter-slice radio resource allocation in deterministic communications systems. We derived an analytical relationship between the allocated radio resource quotas and the upper bounds of delay violation probability and delay variation. Additionally, a critical condition is also identified for the optimal resource allocation between two slices in scenarios with sufficient resources, which helps develop an efficient greedy algorithm for delay-guaranteed radio resource allocation. Numerical results have demonstrated that the derived bounds offer valid upper estimations of the amount of required radio resources to guarantee statistical delay bounds. Furthermore, the proposed algorithm exhibits advantages in dealing with diverse delay requirements across RAN slices.

Appendix

A. Proof of Theorem 1

With the definition of delay violation probability, we have

$$\begin{aligned}\mathbb{P}\{W_i(t) > w_i^{th}\} &\overset{(3)}{=} \mathbb{P}\{A_i(0,t) > D_i(0,t+w_i^{th})\}\\ &= \mathbb{P}\{e^{\theta A_i(0,t)-\theta D_i(0,t+w_i^{th})} > 1\}\\ &\overset{(1)}{\leq} \mathbb{P}\{\sup_{0\leq u\leq t}\{e^{\theta A_i(u,t)-\theta S_i(u,t+w_i^{th})}\} > 1\}\\ &= \mathbb{P}\{\sup_{0\leq u\leq t}\{e^{\theta A_i(t-u,t)-\theta S_i(t-u,t+w_i^{th})}\} > 1\}.\end{aligned}$$

Consider a sequence of non-negative random variables $\{U_{i,u}\}, u = 0, 1, ..., t$, formed by $U_{i,u} = e^{\theta A_i(t-u,t)-\theta S_i(t-u,t+w_i^{th})}$. Since $A_i(\tau,t)$ and $S_i(\tau,t)$ has independent stationary increments, we then have $U_{i,u+1} = U_{i,u}e^{\theta(a_i-s_i)}$ and there holds

$$\begin{aligned}\mathbb{E}[U_{i,u+1}|U_{i,0},...,U_{i,u}] &= \mathbb{E}[U_{i,u}e^{\theta(a_i-s_i)}|U_{i,0},...,U_{i,u}]\\ &= \mathbb{E}[U_{i,u}|U_{i,0},...,U_{i,u}]\mathbb{E}[e^{\theta(a_i-s_i)}]\\ &< U_{i,u}.\end{aligned}$$

The last step holds since $M_{a_i}(\theta)M_{s_i}(-\theta) = \mathbb{E}[e^{\theta(a_i-s_i)}] < 1$. Hence $U_{i,0}, U_{i,1}, ..., U_{i,t}$ form a non-negative supermartingale. Then for any real number $C > 0$, there holds [12]:

$$\mathbb{P}\{\sup_{0\leq u\leq t} U_{i,u} \geq C\} \leq \frac{\mathbb{E}[U_{i,0}]}{C}. \tag{21}$$

Thus, it holds that:

$$\begin{aligned}\mathbb{P}\{W_i(t) > w_i^{th}\} &\leq \mathbb{P}\{\sup_{0\leq u\leq t} U_{i,u} > 1\}\\ &\overset{(21)}{\leq} \mathbb{E}[U_{i,0}]\\ &= \mathbb{E}[e^{\theta A_i(t-0,t)-\theta S_i(t-0,t+w_i^{th})}]\\ &= M_{s_i}^{w_i^{th}}(-\theta).\end{aligned}$$

The last equation holds due to the i.i.d. incremental service process.

B. Proof of Theorem 2

By doubly differentiating both sides of geometric series $\sum_{w=0}^{\infty} x^w = \frac{1}{1-x}$ with respect to $\ln x$, we get,

$$\sum_{w=0}^{\infty} w^2 x^w = \frac{x(1+x)}{(1-x)^3}. \tag{22}$$

Then, we can derive an upper bound of $\mathbb{E}[W_i^2(t)]$,

$$\begin{aligned}\mathbb{E}[(W_i^2(t))] &\overset{(11)}{\leq} \mathbb{E}[W_i^{'2}(t)] = \sum_{w=0}^{\infty} w^2\mathbb{P}\{W_i^{'}(t) = w\}\\ &\overset{(10)}{=} \sum_{w=0}^{\infty} w^2(p_i^{up}(w) - p_i^{up}(w+1))\\ &\overset{(6)}{=} \sum_{w=0}^{\infty} w^2 M_{s_i}^{w}(-\theta^*)(1 - M_{s_i}(-\theta^*))\\ &\overset{(22)}{=} \frac{M_{s_i}(-\theta^*)(1+M_{s_i}(-\theta^*))}{(1-M_{s_i}(-\theta^*))^2}.\end{aligned}$$

References

[1] S. Liu, T. Wang, and S. Wang, "Hardware impairment estimation in NB-IoT: A parallel multi-task learning method," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 6859–6869, Dec. 2022.

[2] F. Song *et al.*, "Enabling heterogeneous deterministic networks with smart collaborative theory," *IEEE Netw.*, vol. 35, no. 3, pp. 64–71, sept 2021.

[3] C. Zhang *et al.*, "Throughput optimization with delay guarantee for massive random access of M2M communications in industrial IoT," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 077–10 092, Dec. 2019.

[4] T. Wang and S. Wang, "Online convex optimization for efficient and robust inter-slice radio resource management," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6050–6062, Jan. 2021.

[5] T. Wang and S. Wang, "Inter-slice radio resource allocation: An online convex optimization approach," *IEEE Wireless Commun.*, vol. 28, no. 5, pp. 171–177, Oct. 2021.

[6] L. Shen, Y. Zhang, and S. Wang, "Codebook based antenna configuration: A new network planning paradigm for mmwave mobile communication systems," *IEEE Trans. Veh. Technol.*, Mar. 2023, doi:10.1109/TVT.2023.3259435.

[7] Y. Zhang *et al.*, "Deterministic network calculus-based H_∞ load frequency control of multiarea power systems under malicious DoS attacks," *IEEE Trans. Smart Grid*, vol. 13, no. 2, pp. 1542–1554, Mar. 2022.

[8] F. Jin, R. Zhang, and L. Hanzo, "Resource allocation under delay-guarantee constraints for heterogeneous visible-light and RF femtocell," *IEEE Wireless Commun.*, vol. 14, no. 22, pp. 1020–1034, Feb. 2014.

[9] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 1, pp. 92–105, Aug. 2014.

[10] C. Xiao *et al.*, "Downlink MIMO-NOMA for ultra-reliable low-latency communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 780–794, Oct. 2019.

[11] O. Adamuz-Hinojosa *et al.*, "A stochastic network calculus (SNC)-based model for planning B5G uRLLC RAN slices," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1250–1265, Feb. 2022.

[12] V. Pena, "A general class of exponential inequalities for martingales and ratios," *Ann. Probab.*, vol. 27, no. 1, pp. 537–564, Jan. 1999.

[13] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. IEEE Int. Workshop Qual. Serv.*, New Haven, CT, USA, Jun. 2006.

[14] H. AI-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 204–217, Feb. 2016.

[15] C. Tepedelenlioglu *et al.*, "Applications of stochastic ordering to wireless communications," *IEEE Wireless Commun.*, vol. 10, no. 12, pp. 4249–4257, Dec. 2011.