QoS-Guaranteed Resource Allocation in Mobile Communications: A Stochastic Network Calculus Approach

Juan Zhu, *Student Member, IEEE*, and Shaowei Wang, *Senior Member, IEEE*

***Abstract*— Deterministic mobile networks are essential for advanced applications that demand strict quality of service (QoS) assurances under limited resource availability. Though network slicing can optimize average performance metrics to offer best-effort services, it often fails to meet the high-reliability requirements of deterministic communication scenarios. In this paper, we introduce a novel QoS-guaranteed inter-slice radio resource allocation scheme for mobile networks to deliver deterministic services over the long term. First, we develop an analytical martingale-based stochastic network calculus framework, which yields stochastic bounds for transmission delays and queue backlogs across various traffic arrival patterns. These bounds produce robust interval estimations that guide resource allocation decisions, effectively addressing channel variability and long-tail QoS effects. Then, an efficient resource allocation algorithm is proposed to approach the derived performance bounds while ensuring fairness across different radio slices with diverse QoS needs. The framework also incorporates an adaptive traffic predictor, enabling our algorithm to track and respond to network dynamics. Numerical results demonstrate that our proposed scheme achieves a promising trade-off between resource utilization and QoS guarantees.**

***Index Terms*— Deterministic communications, QoS, radio resource allocation, stochastic network calculus.**

I. INTRODUCTION

WITH the proliferation of mobile devices and the associated growing demand boosted by emerging applications, 5G and beyond mobile networks are envisioned to support high-quality connectivity for both human-centric and machine-type services [2]. These emerging applications, including industrial automation, autonomous vehicles, and telemedicine, involve multiple coexisting communication services, each with particular delay and reliability requirements. For real-time and mission-critical services, even a slight unexpected delay or packet loss would result in severe consequences [3]. Traditional mobile networks, which provide best-effort services, often struggle to cope with these highly diversified and stringent performance requirements. In response to these challenges, deterministic mobile communications has emerged as a promising approach to bound delay and minimize packet loss [4]. Recent advancements, including time-sensitive networking [5], deterministic networking [6], and deterministic IP [7], have incorporated mechanisms such as time synchronization, traffic shaping, and deterministic forwarding to establish deterministic behaviors and provide QoS guarantees across network segments.

Compared to the core network equipped with powerful computing and caching components, ensuring reliable and predictable operations for the radio access network (RAN) in mobile communication systems, however, poses greater challenges due to the limited radio resources and the harsh radio propagation environments. RAN slicing offers a promising solution by enabling multiple self-contained logical subnets within a shared network infrastructure [8], [9]. The subnets, denominated RAN slices, can be tailored to the requirements of specific communication services. While most advanced RAN slicing solutions focus on resource allocation among multiple user equipments to satisfy their traffic demands, a critical challenge remains: how to pre-determine the required resources for multiple RAN slices to meet their long-term performance requirements? In this regard, the 3rd Generation Partnership Project (3GPP) has introduced the preparation phase within the RAN slice management and orchestration [10], where mobile network operators are required to plan the slice deployment among other tasks.

Real-world communication networks always deal with stochastic service requests and suffer from the severe fadings of radio channels in mobile communication scenarios. As a result, RAN slicing schemes based on average performance metrics may encounter difficulties in consistently meeting the stringent QoS requirements. Recent research has explored planning and scheduling for worst-case scenarios. In [11], a closed-loop load frequency control scheme is devised to ensure an acceptable maximum transmission delay. The worst-case delay, however, can become unbounded if there exists a possibility of zero data transmission, which can occur when, for instance, the received signal-to-noise ratio falls below the limited sensitivity threshold of the receiver [12]. Network performance curves typically present long, low-probability tails [13], [14], suggesting that the worst-case scenarios are oversimplified and rarely happen in practice. The derived

Received 1 November 2023; revised 9 May 2024; accepted 22 August 2024; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor S. Moharir Date of publication 20 September 2024; date of current version 19 December 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61931023. Part of this work has been accepted for presentation at the IEEE Global Communications Conference 2023 [DOI: 10.1109/GLOBECOM54140.2023.10437275], Kuala Lumpur, Malaysia, December 4–8, 2023. *(Corresponding author: Shaowei Wang.)*

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China (e-mail: juanzhu@smail.nju.edu.cn; wangsw@nju.edu.cn).

Digital Object Identifier 10.1109/TNET.2024.3458922

deterministic performance bounds are thus too conservative to utilize the scarce network resources in an efficient way.

Allowing rare QoS violations theoretically would greatly improve the effective utilization of statistical multiplexing gains. In this context, a stochastic performance bound refers to a specific interval, the probability of the experienced service falling within which is not lower than a prescribed threshold determined by a bounding function associated with the interval. Such an approach is instrumental in guiding resource allocation decisions in network environments, particularly when considering the long-tail effect of general QoS metrics. The effect results in skewed and sparse data, rendering single-point estimations of averages or worst-case scenarios vulnerable to outliers. In contrast, stochastic bounds enhance resilience against these challenges by capturing a more comprehensive representation of the metrics, thus offering a more robust means of balancing resource utilization and performance assurance.

Extreme value theory and effective capacity explore stochastic performance bounds by modeling the extreme values or events that occur in the tails of distributions. In [15], the power consumption is minimized by trading off the allocated resources for local computation and task offloading, while the probabilistic constraints on task queue lengths are imposed by using extreme value theory. In [16], a reliability measure characterizing queuing delays is defined using extreme value theory to optimize network-wide power consumption for vehicular users. These extreme value theory-based methods, however, often require prior model parameters that are difficult to obtain in advance. In [17], the delay violation probabilities for vehicle-to-vehicle links, derived from effective capacity theory, are controlled while maximizing the sum ergodic capacity of vehicle-to-infrastructure links. Nevertheless, the effective capacity primarily deals with the deterministic traffic models with fixed traffic parameters. In contrast, stochastic network calculus (SNC) is a more comprehensive and flexible mathematical framework that can provide reliable and reproducible stochastic performance bounds by transforming complex queueing systems to analytically tractable linear ones with alternate algebras [18].

In the literature, the main SNC approaches can be subdivided into three main categories: tail bound-based [19], [20], moment generating functions (MGF)-based [21], and martingale-based [22] approaches. The tail bound-based SNC relies on the available upper envelopes of traffic arrivals and the lower envelopes of services to establish the bounding functions, e.g., in [23], a linear-bounding function SNC model is introduced to calculate radio resource allocation for RAN slices while maintaining acceptable delays. The MGF-based SNC employs *Chernoff* bound and *Boole*'s inequality to derive exponential bounding functions. In [24], a bisection search algorithm is proposed to minimize the transmit power for independent and identically distributed (i.i.d.) traffic arrivals, while meeting the specified constraints on the delay violation probability bound derived from the MGF-based SNC. The martingale-based SNC refines the MGF-based approach to achieve tighter stochastic performance bounds by mitigating the impact of union bound. The martingale-based SNC model in [22] is designed for constant service rate systems and is inapplicable for wireless networks with random channel fading and variable transmission rates. In [25], a delay analysis framework for i.i.d. traffic and service models is proposed to dynamically allocate transmission power. Nevertheless, all these SNC frameworks would become intractable when dealing with the well-known self-similar properties of real-world traffic scenarios [26].

In this article, we address the challenges of deterministic communications scenarios where mobile network operators have to proactively plan the deployment of multiple RAN slices, each with distinct delay and reliability requirements. We introduce an efficient resource allocation scheme aiming at achieving a balance between multi-slice QoS guarantee and resource utilization. The proposed scheme consists of two parts: derivation of essential performance metrics and bandwidth allocation. For the former, a martingale-based SNC model is proposed to establish stochastic bounds of transmission delays and queue backlogs for both i.i.d. and non-i.i.d. traffic arrivals. For the bandwidth allocation, an efficient greedy algorithm is designed to iteratively optimize the derived performance metrics while accommodating time-varying traffic arrivals and ensuring slice fairness. The main contributions of the paper can be summarized as follows.

- We develop a martingale-based SNC model that provides stochastic bounds on transmission delay, delay variation, and packet loss rate for RAN slices operating under block fading channels. The proposed model serves as a robust and reliable approach to inform QoS-guaranteed resource allocation decisions in the networks characterized by long-tail distributed performance.
- We analyze two prominent categories of arrival processes: i.i.d. incremental processes and autoregression (AR) models. While the simple i.i.d. models provide good intuition into the functioning of our martingale-based SNC model, the AR models offer a more realistic representation by capturing the self-similar characteristics of network traffic.
- We give the sufficient and necessary condition for optimal bandwidth allocation by exploiting the attributes of the derived stochastic bounds. This decisive condition serves as a guiding principle in the development of a highly efficient greedy algorithm that achieves a trade-off among slice fairness, resource utilization, and QoS assurance.
- We design an adaptive traffic predictor to monitor and respond to fluctuations in network traffic patterns, which enables our resource allocation scheme to dynamically track the evolving network demands.

The remainder of this article is organized as follows. Section II presents the problem formulation and introduces the three key metrics used to evaluate network performance. In Section III, we develop a versatile martingale-based SNC framework to derive stochastic bounds for performance metrics, with analysis focused on two illustrative traffic classes. An efficient greedy algorithm for addressing the optimal bandwidth allocation problem in deterministic mobile communications is elucidated in Section IV. Numerical results

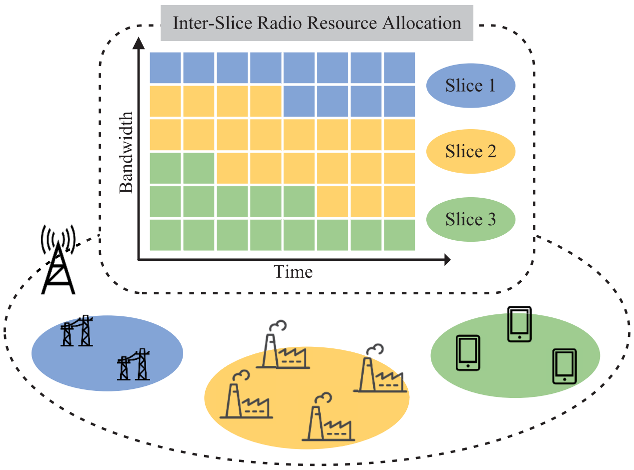


Fig. 1. Inter-slice radio resource allocation.

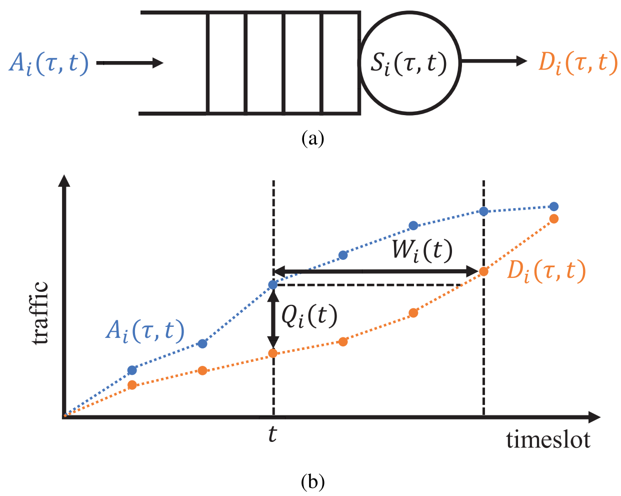


Fig. 2. A single queueing node with arrival $A_i(\tau,t)$, service $S_i(\tau,t)$ and departure $D_i(\tau,t)$ processes.

validating our proposal are provided in Section V. Section VI concludes this work.

II. System Model

Consider the RAN in a given mobile network where the radio resources between different cells are logically isolated from each other and thus each cell can be configured independently. We focus on a cell with I slices, and the total bandwidth resources of the cell are organized into R resource blocks (RBs). Denote the set of RAN slices by $\mathcal{I}$, and each $i \in \mathcal{I}$ provides a customized network service for an aggregated traffic of multiple users. The RAN slice orchestrator periodically executes an inter-slice scheduling procedure to decide the dedicated RB quota for each slice so as to meet the QoS requirements in the long term. An illustrative example with $I = 3$ slices and $R = 48$ RBs is depicted in Fig. 1.

A. Metrics for Evaluating Network Performance

Consider a time-slotted queuing system for RAN slice $i \in \mathcal{I}$, as depicted in Fig. 2(a). Let $a_i(\tau)$ and $s_i(\tau)$ denote the random instantaneous arrival and service increments in the τ-th timeslot, respectively. For a wireless channel with signal-to-noise ratio γ, its achievable transmission rate is

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

where B_i is the allocated bandwidth. For block fading channels, the channel conditions are i.i.d. across different timeslots, which carries over to $s_i(t)$'s. The cumulative traffic arrival and service processes in the time interval $[0, t)$ are defined by bivariate functions $A_i(0,t) = \sum_{\tau=0}^{t-1} a_i(\tau)$ and $S_i(0,t) = \sum_{\tau=0}^{t-1} s_i(\tau)$, respectively. A service description $S_i(0,t)$ is referred to as a dynamic server if its corresponding cumulative departure process $D_i(0,t)$ satisfies the following inequality for any arrival process $A_i(0,t)$ [21]:

$$D_i(0,t) \geq \inf_{0\leq\tau\leq t}\{A_i(0,\tau) + S_i(\tau,t)\}. \tag{2}$$

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

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

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

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

To show the predictable system 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:

$$V_i(t) \triangleq \sqrt{\mathbb{E}[W_i^2(t)]}. \tag{5}$$

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

Measuring packet loss rate $L_i(t)$ in network systems can be non-trivial due to its dependence on various factors such as buffer overflow, retransmissions, and packet corruption. To simplify the analysis, we assume that packet loss occurs only when the buffer overflows. In the SNC, the queueing backlog $Q_i(t)$ at time t is defined as

$$Q_i(t) \triangleq A_i(0,t) - D_i(0,t). \tag{6}$$

We approximate $L_i(t)$ by the probability that the random backlog $Q_i(t)$ exceeds the given buffer budget q at any time t:

$$L_i(q,t) \triangleq \mathbb{P}\{Q_i(t) > q\}. \tag{7}$$

In the SNC analysis, a queueing node is considered unstable if its backlog or delay grows over time and becomes unbounded. To ensure that the queueing backlog and delay remain finite at all times, a commonly used stability condition can be expressed as follows:

$$\lim_{t\to\infty}\frac{1}{t}\mathbb{E}[A_i(0,t)] < \lim_{t\to\infty}\frac{1}{t}\mathbb{E}[S_i(0,t)], \tag{8}$$

where $\mathbb{E}[\cdot]$ represents the expectation of a random variable. Once the stability condition is satisfied, the SNC framework can provide reliable and reproducible statistical bounds on performance metrics. Denote the upper bounds of delay violation probability, delay variation, and packet loss rate for RAN slice i by P_i^{up}, V_i^{up} and L_i^{up}, respectively. The analytical expressions are derived in the subsequent section.

B. Problem Formulation

To accommodate the diverse QoS requirements across RAN slices, we aim to minimize the maximum weighted sum of delay violation probability bound and packet loss rate bound among all slices, while subjecting to the total resource constraints and ensuring that the upper bounds of delay variation remain below predefined thresholds.

When the RAN is unable to satisfy the requirements of all slices due to the limited resources, the RAN slice orchestrator should prioritize the slices. The slice with higher priority would achieve lower performance bounds, thereby ensuring stricter QoS. To this end, we assign a priority ϕ_i to slice i as a tunable hyperparameter. Define f_i for RAN slice i as follows:

$$f_i = \phi_i(P_i^{up} + \omega_i L_i^{up}), \tag{9}$$

where ω_i is a predefined weight that balances the impact of delay violation probability and packet loss rate for slice i. The RB allocation among RAN slices is denoted by $\boldsymbol{\mathcal{R}} = (\mathcal{R}_1, \cdots, \mathcal{R}_1, \cdots \mathcal{R}_I)$, where $\mathcal{R}_i$ represents the set of available RBs for slice i. A balance among the QoS demands of all RAN slices can be achieved by solving the following min-max problem:

$$\begin{aligned} &\min_{\boldsymbol{\mathcal{R}}} \; F = \max(f_1, \dots, f_I), \\ &\text{s.t. } C_1\text{: } \sum_{i\in\mathcal{I}} |\mathcal{R}_i| \le R, \\ &\quad\;\; C_2\text{: } V_i^{up} \le V_i^{th}, \forall i \in \mathcal{I}, \end{aligned} \tag{10}$$

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

III. Martingale-Based Stochastic Network Calculus Model

The SNC serves as a unified framework to analyze the queueing behaviors across a wide range of traffic classes. Typically, such a framework relies on *Boole*'s inequality to bound stochastic processes. This inequality, however, is known to be overly conservative, especially in non-*Poisson* scenarios. To improve the closeness of upper estimations across diverse arrival patterns, we introduce a novel technique that leverages supermartingale theory [27] to derive time-invariant performance metric bounds. The underlying idea for this technique stems from the stability condition (8), which indicates that the average arrival rate must be strictly less than the average service rate to keep a queueing system in a stable regime. When the positivity constraint on the buffer is disregarded, the expected increment of the buffer content becomes negative, causing its behavior to resemble that of a supermartingale. This insight inspires the utilization of *Doob*'s martingale inequality rather than *Boole*'s inequality to establish tighter and more robust performance bounds.

We simplify notation by omitting the subscript i denoting the slice index, as the following analysis is universally applicable across all slices. Considering a p-order dependent arrival process, we introduce the following vector:

$$\vec{a}^p(\tau) = (a(t-\tau-1), \sum_{i=1}^{2} a(t-\tau-i), \dots, \sum_{i=1}^{p} a(t-\tau-i)).$$

This vector collects historical data that influences the traffic increment at time τ. For the sake of technical clarity, we establish a parallel notation for the service process:

$$\vec{s}^p(\tau) = (s(t-\tau-1), \sum_{i=1}^{2} s(t-\tau-i), \dots, \sum_{i=1}^{p} s(t-\tau-i)).$$

To compensate for potential correlations among the neighboring traffic increments, we introduce a class of functions defined as follows:

$$\begin{aligned} \mathscr{H} \triangleq \{h : \mathbb{R}^{p+1} \to \mathbb{R} \mid h(\vec{a}^p(\tau), \cdot) &\text{ is non-increasing} \\ &\text{in each element of } \vec{a}^p(\tau)\}. \end{aligned} \tag{11}$$

We proceed to present the following definition:

Definition 1 (Supermartingale Traffic): For $\theta > 0$ and a blocking fading channel characterized by an i.i.d. incremental service process $S(0,t) = \sum_{\tau=0}^{t-1} s(\tau)$, an process $A(0,t)$ is a (h, θ, S)-supermartingale-traffic arrival if there exists a function $h(\vec{a}^p(\tau), \theta) \in \mathscr{H}$ such that the process

$$\{U(\tau) = h(\vec{a}^p(\tau), \theta) e^{\theta A(t-\tau,t) - \theta S(t-\tau,t)}, 0 \le \tau < t\} \tag{12}$$

forms a non-negative supermartingale, i.e., the process $\{U(\tau), 0 \le \tau < t\}$ satisfies

$$\mathbb{E}[U(\tau+1)|U(0), \dots, U(\tau)] < U(\tau). \tag{13}$$

The rationale behind the exponential transform in (12) lies in its shape directly determines the rate of decay for queueing metrics. Since *Doob*'s inequality does not distinguish supermartingales from martingales, a smaller gap between the constructed supermartingale and a martingale would tighten the inequality. The convex nature of the exponential transform assigns more weight to larger arrival values, thereby reducing negative drift and narrowing the gap. Additionally, maximizing the decay factor θ also serves to minimize the gap. However, an excessively large θ value would violate (13) and destroy the constructed supermartingale. The feasible region of θ is given by

$$\boldsymbol{\Theta} = \{\theta > 0 | \lim_{t\to\infty} \mathbb{E}^{\frac{1}{t}}[e^{\theta A(0,t) - \theta S(0,t)}] < 1\}. \tag{14}$$

By applying *Jensen*'s inequality, it becomes evident that $\boldsymbol{\Theta} \neq \emptyset$ only when the stability condition (8) holds.

The following theorems constitute the central results of this work, delineating how the supermartingale-traffic can be leveraged to derive the upper bounds of our performance metrics.

Theorem 1 (Upper Bound for Delay Violation Probability): Consider a queueing system with an arrival process $A(0,t)$ and a dynamic server $S(0,t) = \sum_{\tau=0}^{t-1} s(\tau)$ that transmits *data over block fading channels. If $A(0,t)$ is a (h,θ,S)-supermartingale-traffic arrival, then the delay violation probability with respect to the delay deadline w at any time can be upper bounded by*

$$P^{up}(w)=\begin{cases}\inf\limits_{\theta\in\mathbf{\Theta}}\kappa(\theta)M_s^w(-\theta), & \text{if (8) holds,}\\ 1, & \text{otherwise,}\end{cases}\tag{15}$$

where

$$\kappa(\theta)=\frac{\mathbb{E}[h(\vec{a}^p(\tau),\theta)]}{\mathbb{E}[h(\vec{s}^p(\tau),\theta)]},\tag{16}$$

and $M_s(-\theta)=\mathbb{E}[e^{-\theta s}]$ represents the moment generating function of the random variable s.

Proof: Please refer to Appendix B. □

To estimate $V(t)$, we define a random variable $W'(t)$, such that for an arbitrary positive integer w,

$$P'(w)=\mathbb{P}\{W'(t)>w\}=\kappa(\theta)M_s^w(-\theta),\tag{17}$$

and

$$\begin{aligned}\mathbb{P}\{W'(t)=w\}&=\mathbb{P}\{W'(t)>w\}-\mathbb{P}\{W'(t)>w+1\}\\&=P'(w)-P'(w+1).\end{aligned}\tag{18}$$

Since the infimum of $P'(w)$ is $P^{up}(w)$ according to Theorem 1, it follows that $\mathbb{P}\{W'(t)>w\}\geq\mathbb{P}\{W(t)>w\}$, which indicates that random variables $W'(t)$ and $W(t)$ are statistically ordered. Therefore, their second-order moments are also ordered [28]:

$$\mathbb{E}[(W^2(t))]\leq\mathbb{E}[W'^2(t)].\tag{19}$$

Then we can derive a time-invariant upper bound V^{up} of the delay variation, as stated in Theorem 2.

Theorem 2 (Upper Bound for Delay Variation): Consider a queueing system with an arrival process $A(0,t)$ and a dynamic server $S(0,t)=\sum_{\tau=0}^{t-1}s(\tau)$ that transmits data over block fading channels. If $A(0,t)$ is a (h,θ,S)-supermartingale-traffic arrival, then an upper bound $V^{up}(t)$ of the delay variation defined in (5) can be derived as

$$\begin{aligned}V^{up}&=\sqrt{\mathbb{E}[W'^2(t)]}\\&=\begin{cases}\inf\limits_{\theta\in\mathbf{\Theta}}\dfrac{\sqrt{\kappa(\theta)M_s(-\theta)(1+M_s(-\theta))}}{1-M_s(-\theta)}, & \text{if (8) holds,}\\ c, & \text{otherwise,}\end{cases}\end{aligned}\tag{20}$$

where c is a sufficiently large constant. $\mathbf{\Theta}$ and $\kappa(\theta)$ are given by (14) and (16), respectively.

Proof: Please refer to Appendix C. □

For packet loss probability defined in (7), we present the following theorem:

Theorem 3 (Upper Bound for Packet Loss Rate): Consider a queueing system with an arrival process $A(0,t)$ and a dynamic server $S(0,t)=\sum_{\tau=0}^{t-1}s(\tau)$ that transmits data over block fading channels. If $A(0,t)$ is a (h,θ,S)-supermartingale-traffic arrival, then the packet loss rate with respect to the buffer budget q at any time be upper bounded by

$$L^{up}(q)=\begin{cases}\inf\limits_{\theta\in\mathbf{\Theta}}\kappa(\theta)e^{-\theta q}, & \text{if (8) holds,}\\ 1, & \text{otherwise,}\end{cases}\tag{21}$$

where $\mathbf{\Theta}$ and $\kappa(\theta)$ are given by (14) and (16), respectively.

Proof: Please refer to Appendix B. □

The performance bounds derived in Theorem 1-3 are essentially solutions to constrained optimization problems, with θ as the decision variable constrained within the feasible region $\mathbf{\Theta}$. These bounds lack closed-form expressions due to the inherent complexity of the queueing system.

The proposed martingale-based SNC model is versatile across broad classes of arrival processes. This work illustrates its applications through two representative cases: i.i.d processes and p-order autoregressive processes. While not realistic, i.i.d. models provide valuable insights into the functionality of our model. AR models capture the self-similar characteristics of network traffic, showcasing the effectiveness of the proposed SNC framework in handling realistic non-i.i.d. scenarios.

1) i.i.d. arrival process: Consider a series of non-negative i.i.d. random variables $a(\tau)$'s.

Theorem 4 (Stochastic Performance Bounds for i.i.d. Arrival Processes): For an i.i.d. incremental arrival process $A(0,t)=\sum_{\tau=0}^{t-1}a(\tau)$, let the function $h\in\mathscr{H}$ be given by,

$$h(\vec{a}^0(\tau),\theta)=1,\tag{22}$$

such that the traffic $A(0,t)$ served by an i.i.d. incremental service process $S(0,t)$ forms a $(1,\theta,S)$-supermartingale-traffic arrival. The corresponding statistical performance bounds are provided in Theorem 1-3 with

$$\kappa(\theta)=1.\tag{23}$$

Proof: Please refer to Appendix D. □

2) Autoregressive arrival process: A p-order autoregressive process, denoted as AR(p), evolves by rescaling the p previous values of the process and adding white *Gaussian* noise. Let $Z_0,Z_1,\cdots$ be i.i.d. standard *Gaussian* variables. For $p\geq 1$, $\varphi_1,\cdots,\varphi_p\in[0,1)$, $\varphi=\sum_{k=1}^{p}\varphi_k$, and $\mu,\sigma>0$, $a(\tau)$ satisfies the following relation:

$$a(\tau)=\sum_{k=1}^{p}\varphi_k a(\tau-k)+(1-\varphi)\mu+(1-\varphi)\sigma Z_\tau.\tag{24}$$

This p-order autoregressive process is a *Gaussian* process with mean μ, and its variance can be derived using the *Yule-Walker* equations [29].

Theorem 5(Stochastic Performance Bounds for Autoregressive Arrival Processes): For an autoregressive incremental arrival process $A(0,t)=\sum_{\tau=0}^{t-1}a(\tau)$ described above, let the function $h\in\mathscr{H}$ be given by

$$h(\vec{a}^p(\tau),\theta)=e^{-\frac{\theta}{1-\varphi}\sum_{k=1}^{p}\varphi_k\sum_{i=1}^{k}a(t-\tau-i)},\tag{25}$$

such that the traffic $A(0,t)$ served by an i.i.d. incremental service process $S(0,t)$ forms a (h,θ,S)-supermartingale-traffic arrival. The corresponding performance bounds are *provided in Theorem 1-3 with*

$$\kappa(\theta) = e^{-\frac{\mu\theta}{1-\varphi}+\frac{\theta^2\upsilon^2}{2(1-\varphi)^2}} M_s^{-1}\Big(-\frac{\theta}{1-\varphi}\sum_{k=1}^{p} k\varphi_k\Big). \tag{26}$$

Here, υ^2 denotes the variance of $\sum_{k=1}^{p}\varphi_k\sum_{i=1}^{k} a(t-\tau-i)$.

Proof: Please refer to Appendix E. □

The main workload of our model lies in designing an appropriate function $h \in \mathscr{H}$ to form a (h,θ,S)-supermartingale-traffic tailored to the arrival characteristics, which significantly influences the quality of the derived QoS bounds. Despite this, we emphasize at least two advantages of our method: Firstly, our model yields tighter bounds compared to the existing SNC models for the most prevalent i.i.d. traffic arrivals; also, our model is capable of providing statistical performance bounds for non-i.i.d. traffic, a challenge that other methods typically struggle to address.

IV. QoS-Guaranteed Radio Resource Allocation

To compute the required RB number in advance to meet the QoS requirements of multiple RAN slices in the long term, we develop a two-step dynamic radio resource allocation scheme, where an adaptive traffic predictor is designed to forecast the future traffic and an effective bandwidth scheduler is used to allocate RBs to RAN slices based on the predicted traffic intensity.

A. Adaptive Traffic Prediction

To accommodate the strong time variability in network traffic, the model parameters of arrivals are continuously updated using recent data. In contrast, the traffic distribution type is determined offline on the basis of a large dataset of experimental records to balance predictive capability and model complexity, which would be re-estimated if the prediction error exceeds a certain threshold during the dynamic update process of traffic parameters.

We employ the partial autocorrelation function (PACF) [30] to differentiate between i.i.d. arrivals and AR models, as well as to determine the order of AR models. The PACF values are calculated for traffic observations at lags $k = 1, 2, \ldots, K$, with K being the predefined maximum lag. An i.i.d. process exhibits negligible correlations between observations, with PACF values remaining close to zero and statistically insignificant for all lags beyond lag 0. For an AR(p) process, the lag corresponding to the last significant spike before values drop to insignificance is a reliable indicator of the order of the model, as exemplified in Fig. 3.

Once the traffic model is identified, we employ adaptive prediction methods to estimate the associated model parameters, while a sliding window is incorporated to make use of the recent information in arriving data. For i.i.d. processes, the maximum likelihood estimation [31] presents a straightforward and well-suited approach with minimal assumptions and asymptotic efficiency. When dealing with AR models, the adaptive *Kalman* filter proves to be a powerful solution due to its ability to provide uncertainties for parameter estimates and handle missing data and irregular sampling. Specifically,

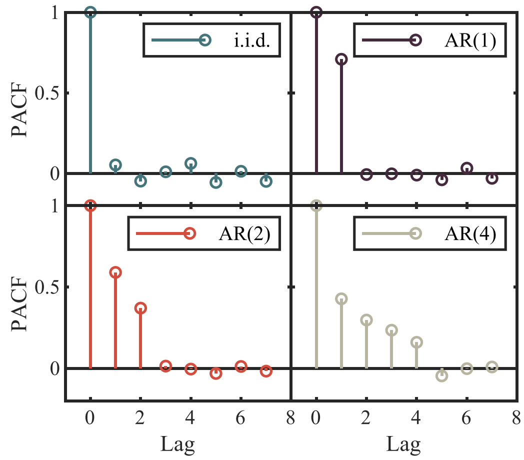


Fig. 3. PACF of i.i.d. processes and AR models of different orders.

we define the state vector as $x(t) = [1-\varphi(t), \varphi_1(t), \ldots, \varphi_p(t)]$ and consider $H(t) = [1, a(t-1), \ldots, a(t-p)]$. The corresponding state transition equation and observation equation are formulated as follows:

$$\begin{aligned} &x(t) = x(t-1),\\ &a(t) = H(t)x(t) + v(t), \end{aligned} \tag{27}$$

where $v(t) = (1-\varphi)\sigma Z_\tau$ represents the measurement noise.

B. QoS-Guaranteed Radio Resource Allocation

Due to the absence of closed-form expressions for the derived performance bounds, the objective and constraints in (10) become nontrivial to handle. To address this challenge, we develop an efficient greedy algorithm that constructs promising solutions by making a series of as good as possible local choices in each step. This approach allows us to achieve practical solutions despite the inherent complexity of the problem.

To effectively address the constraints imposed by variations in delay, we introduce a penalty function for RAN slice i defined as follows:

$$f_i' = \phi_i(P_i^{up} + \omega_i L_i^{up}) + \phi_i(1+\omega_i)\mathbb{1}(V_i^{up} > V_i^{th}), \tag{28}$$

where $\mathbb{1}(V_i^{up} > V_i^{th})$ is an indicator function equaling 1 if $V_i^{up} > V_i^{th}$ and 0 otherwise. The penalty function f_i' monotonically decreases as the number of allocated RBs increases, since more available bandwidth leads to fewer traffic jams and lower delay. From (28), if neither the stability condition nor the delay variation constraint is met, $f_i' = 2\phi_i(1+\omega_i)$; if one holds and the other does not, $\phi_i(1+\omega_i) \le f_i' < 2\phi_i(1+\omega_i)$; and if both conditions are fulfilled, $f_i' < \phi_i(1+\omega_i)$. These inherent characteristics greatly facilitate the design of our algorithm.

Focusing on a single step of the greedy algorithm, we redistribute RBs between two specific slices i' and i'':

$$\begin{aligned} &\min_{|\mathcal{R}_{i'}|,|\mathcal{R}_{i''}|} \max(f_{i'}', f_{i''}'),\\ &\quad \text{s.t. } |\mathcal{R}_{i'}| + |\mathcal{R}_{i''}| = R', \end{aligned} \tag{29}$$

where R' represents the total number of RBs allocated to i' and i'' in the last cycle. Since the monotonically decreasing behavior of the penalty f_i', the optimal solution to (29) is governed by the following sufficient and necessary condition:

$$f_{i'}' = f_{i''}'. \tag{30}$$

This condition offers the local best choice in a single step for (10). Due to the discretization in RB allocation, however, (30) may not be strictly met. Instead, our goal is to maximize the proximity between $f_{i'}'$ and $f_{i''}'$ in each cycle.

The key steps of the proposed algorithm are outlined as follows: Initially, the available RBs are evenly distributed among all RAN slices. In the subsequent iterations, RBs are reallocated from slice i'' with the lowest penalty to slice i' with the highest penalty, until their penalties become nearest. Subsequently, the objective function, the maximum penalty across all slices, is updated. Finally, we check if this value has decreased compared to the previous iteration. 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 then select the slice with the highest penalty among the remaining slices to donate RBs to slice i'. The algorithm ceases when all other slices have no longer redundant RBs that could be allocated to the slice with the highest penalty.

To avoid repetitive allocation between these slices, the penalty of i'' should not exceed that of i' at the end of each redistribution step, except in two cases: 1) slice i' has a higher priority than slice i'', and during the reallocation of the last RB, the stability condition and delay variation constraint of slice i' transition from being unsatisfied to satisfied, and 2) during the reallocation of the last RB, the stability condition and delay variation constraint of slice i' transition from unsatisfied to satisfied, while the stability condition of slice i'' remains satisfied. Typically, the iteration count N is much less than I^2 due to such mechanisms involved.

The details of our proposed QoS-guaranteed inter-slice radio resource allocation procedure are summarized in Algorithm 1. Since the decision variables, the number of RBs, are discrete and finite, we precompute the QoS bounds corresponding to different RB numbers offline and store them in a lookup table. This lookup table allows our adaptive algorithm to directly retrieve the required information during runtime, thereby mitigating the complexity associated with real-time statistical performance bounds calculations and significantly enhancing overall efficiency. In each iteration, selecting two specific slices and reallocating RBs has a complexity of $\mathcal{O}(I)$ and $\mathcal{O}(R)$, respectively. The total complexity of algorithm 1 is $\mathcal{O}(N(I+R))$.

Algorithm 1 Radio Resource Allocation for Deterministic Communications

Initialization: Equal distribution of RBs among the RAN slices. Set $m = 0$;

Calculate f_i' by (28) and evaluate $F' = \max(f_1', \dots, f_I')$;
while $m < I$ **do**
  Set $prev\_F = F'$;
  Set $a = \text{sort}(f_1', \dots, f_I')$. Select $i' = \arg(a(-1))$ and $i'' = \arg(a(m))$;
  Calculate $f_{i'}' - f_{i''}'$;
  while $f_{i'}' - f_{i''}' > 0$ **do**
    Set $prev\_f_{i'}' = f_{i'}'$;
    Set $|\mathcal{R}_{i'}| = |\mathcal{R}_{i'}| + 1$ and $|\mathcal{R}_{i''}| = |\mathcal{R}_{i''}| - 1$;
    Calculate $f_{i'}'$, $f_{i''}'$ by (28);
  end
  if *not* $(\phi_{i'} > \phi_{i''}$ *and* $prev\_f_{i'}' > \phi_i(1+\omega_i)$ *and* $f_{i'}' < \phi_i(1+\omega_i))$ *or not* $(prev\_f_{i'}' > \phi_i(1+\omega_i)$ *and* $f_{i'}' < \phi_i(1+\omega_i)$ *and* $f_{i''}' < \phi_i(1+\omega_i))$ **then**
    Set $|\mathcal{R}_{i'}| = |\mathcal{R}_{i'}| - 1$ and $|\mathcal{R}_{i''}| = |\mathcal{R}_{i''}| + 1$;
  end
  Evaluate $F' = \max(f_1', \dots, f_I')$;
  if $F' < prev\_F$ **then**
    Set $m = 0$.
  end
  else
    Set $m = m + 1$;
  end
end
ruturn: $\mathcal{R}_i \quad \forall i \in \mathcal{I}$

V. Numerical Results

A. Experimental Settings

Wireless channels are assumed to follow *Nakagami* fading [32] with the fading parameter of 5 and the average power of 3.16 dBm. Each RB has a bandwidth of 180 kHz, and time is discretized into slots with the duration of 0.5ms. The traffic parameters and QoS requirements of RAN slices are detailed in Table I.

To demonstrate the effectiveness of our proposed SNC model, we compare the derived performance bounds with the results obtained from simulations. We consider two traffic arrival types, the composite Poisson process and the AR models. In scenarios with i.i.d. incremental traffic arrivals, we also contrast our results with the bounds established in [33], which are derived using *Boole*'s inequality.

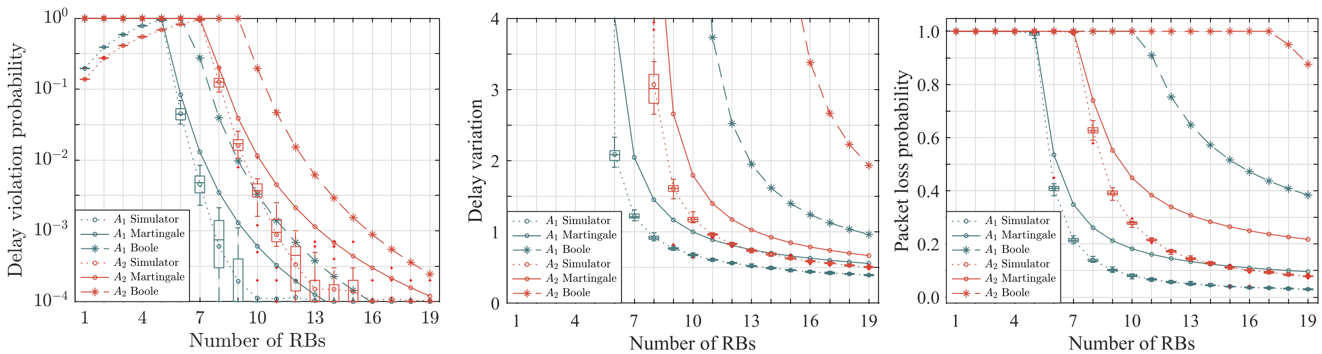
To evaluate our resource allocation algorithm, we first compare the prediction accuracy of our adaptive traffic predictor against the recursive least square (RLS) algorithm under different traffic fluctuation levels. We then assess the convergence of the proposed greedy algorithm. Finally, we perform comparative analyses with baseline approaches and state-of-the-art methods by simulating the RB allocation schemes provided by these methods. The distinct QoS requirements of various slices and the traffic arrivals at different time slots are randomly generated from the distribution specified in Table I. The baselines considered are:

- Equal RB allocation across slices.
- Allocation proportional to mean traffic demand.
- Allocation based on the average performance metrics derived using the M/G/1 queue theory model [34], which determines RB allocation by minimizing the maximum value of the weighted sum of average delay and average packet loss rate, paralleling with the objective in (10).

To further substantiate the merits of our proposal, we compare it with the following algorithms:

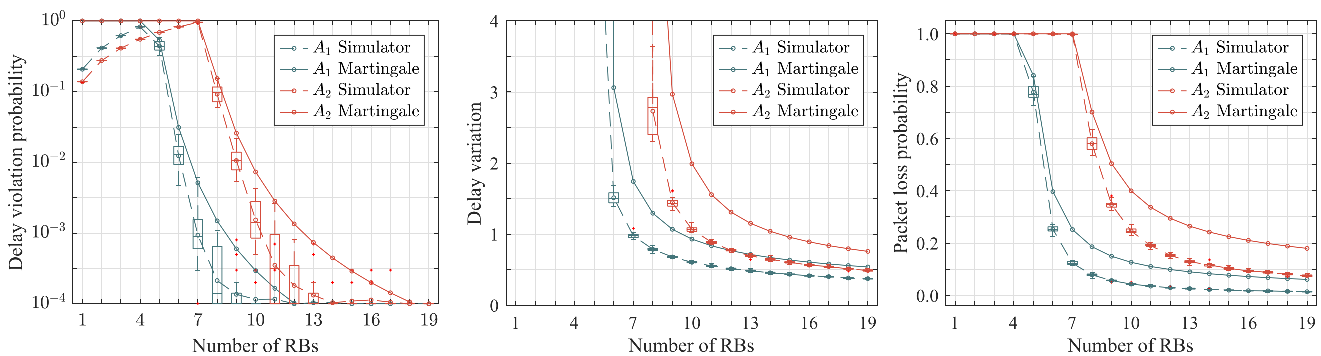
TABLE I
SETTINGS OF TRAFFIC PARAMETERS AND QOS REQUIREMENTS FOR RAN SLICES

	Fig. 4	Fig. 5	Fig. 7 and Fig. 8	Fig. 9	Fig. 10
i.i.d. incremental arrivals $a(\tau)=\sum_{n=1}^{N} X_n$ $N \sim Poisson(\lambda_1)$ $X_n \sim Poisson(\lambda_2)$(kbps)	A_1: $\lambda_1 = 7$ $\lambda_2 = 70$; A_2: $\lambda_1 = 10$ $\lambda_2 = 90$	/	$\lambda_1 \sim Uniform(5, 15)$ $\lambda_2 \sim Uniform(50, 150)$	/	$\lambda_1 \sim Uniform(10, 20)$ $\lambda_2 \sim Uniform(50, 200)$
AR incremental arrivals Eq. (24) μ(Mbps)	/	A_1: $p = 1$ $\phi = 0.7$ $\mu = 1.2$ $\sigma = 30$; A_2: $p = 1$ $\phi = 0.7$ $\mu = 2.0$ $\sigma = 30$	/	$p = 1$ $\phi \sim Uniform(0.4, 0.9)$ $\mu \sim Uniform(0.5, 2.0)$ $\sigma = 30$	/
QoS requirements of slices	w_i=4 $q_i = 70$kbit		$w_i \sim Uniform(1, 10)$ $V_i^{th} \sim Uniform(10, 40)$ $q_i \sim Uniform(2, 8) * 20$kbit		



(a) Delay violation probability. (b) Delay variation. (c) Packet loss probability.

Fig. 4. QoS metrics of i.i.d. incremental arrival processes as functions of the number of allocated RBs for two different traffic processes.



(a) Delay violation probability. (b) Delay variation. (c) Packet loss probability.

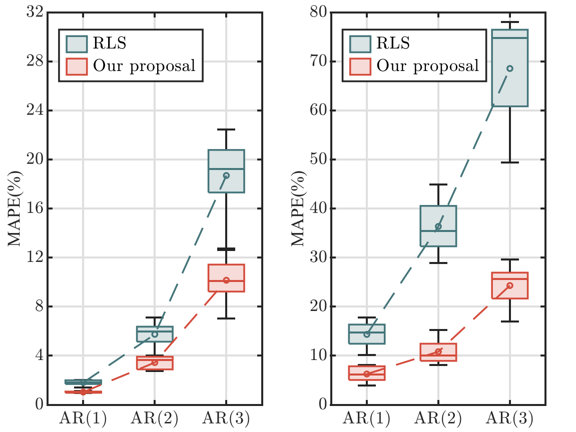
Fig. 5. QoS metrics of 1-order AR arrival models as functions of the number of allocated RBs for two different traffic processes.

- FairLog [35]: The logarithmic assignment is employed to ensure equal fair among slices in RB allocation, while the number of each slice is limited by its average requirement. A Lagrangian function is utilized for optimization.
- ShareHeu [36]: A low-complexity heuristic algorithm is proposed for network-level resource preallocation problem, where overflowed packet transmission is scheduled by borrowing RBs from other services. The algorithm aims to preallocate the minimal RB number, and we fairly compare it with our proposal by modifying its termination condition to allocate a fixed number of RBs.
- UtilAlloc [37]: Utility functions are designed to quantify vehicle service satisfaction considering transmission rate, age of information, and packet loss rate. A heuristic resource allocation is proposed to maximize weighted utility functions for efficient resource distribution among slices.

Since most of these comparison methods rely on i.i.d. traffic arrivals, fair comparisons are limited to i.i.d. traffic conditions and exclude AR models. This limitation of the existing methods already highlights one of the key advantages of our proposed approach, which is its ability to handle non-i.i.d. traffic patterns.

B. Validation of the Proposed SNC-Based Model

Fig. 4 presents the simulated performance metrics and the corresponding derived upper bounds for two distinct traffic



(a) Slight traffic fluctuations. (b) Substantial traffic fluctuations.

Fig. 6. MAPE of adaptive traffic predictor for different traffic models.

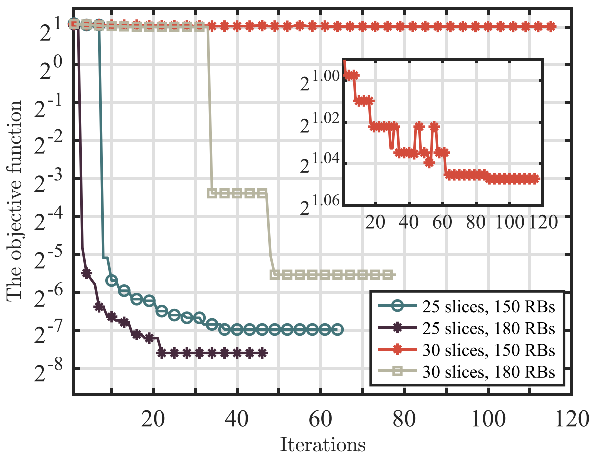


Fig. 7. Convergence of proposed greedy algorithm.

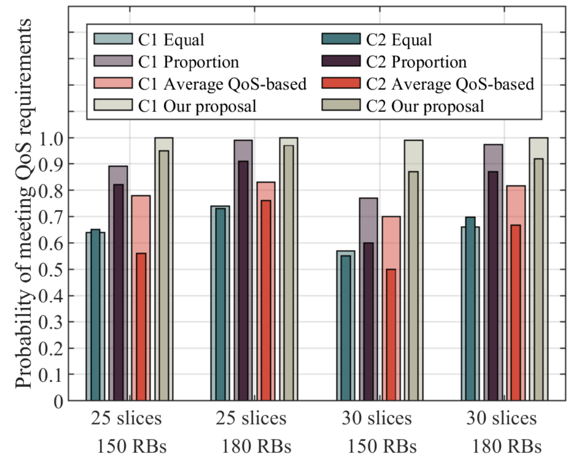


Fig. 8. Probabilities of meeting QoS requirements for RAN slices under Case 1 (C1) with a known time-invariant i.i.d. traffic pattern and Case 2 (C2) with a predicted time-varying i.i.d. traffic pattern.

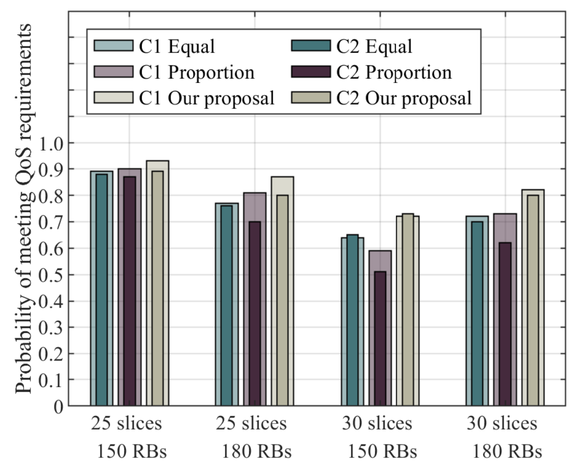


Fig. 9. Probabilities of meeting QoS requirements for RAN slices under Case 1 (C1) with a known time-invariant AR traffic pattern and Case 2 (C2) with a predicted time-varying AR traffic pattern.

processes denoted as A_1 and A_2, with A_2 exhibiting higher intensity than A_1. Our proposed model consistently offers valid upper estimations of delay violation probabilities, delay variations, and packet loss rate. This demonstrates the effectiveness of our model in determining the required number of RBs for RAN slices to ensure their QoS within the specified intervals. In contrast, the SNC model based on *Boole*'s inequality often yields much looser upper bounds. While the SNC model based on *Boole*'s inequality cannot handle non-i.i.d. traffic, our proposed SNC model continues to perform effectively in capturing their boundary characteristics, as shown in Fig. 5.

In these scenarios, the derived performance bounds monotonically decrease with a diminishing rate as the allocated RB number increases, which implies that achieving a deterministic zero-violation delay is resource-intensive and impractical. Instead, statistical bounds serve as more reasonable and efficient metrics to guide radio resource allocation decisions under the conditions of limited available resources. Furthermore, we point out that these bounds match simulations at the starting point of the true distribution. As the martingale-based SNC is based on exponential transforms, it can only render bounds in the form of the generalized exponential distribution. This implies that the longer the true distribution aligns with its initial trend, the smaller the gap is between the distribution and the obtained bounds.

C. Performance Analysis

Fig. 6 shows the mean absolute percentage errors (MAPEs) of traffic predictors under different traffic models and fluctuation levels. When the traffic arrivals are modeled as AR models, our adaptive traffic predictor exhibits superior accuracy and stability in comparison to RLS prediction. The advantage arises from the capability of the adaptive *Kalman* filter to incorporate uncertainty into parameter estimation, making it more robust in handling non-stationary traffic data.

In Fig. 7, the convergence behaviors of the proposed algorithm are shown under various scenarios involving varying numbers of slices and RBs. The curves head down until the algorithm converges. When the total radio resources are insufficient to accommodate the traffic demands of all RAN slices, the curve converges to a positive value. The outcome indicates that the available RBs are not enough for the slices with the lowest priorities to meet their stability conditions or delay

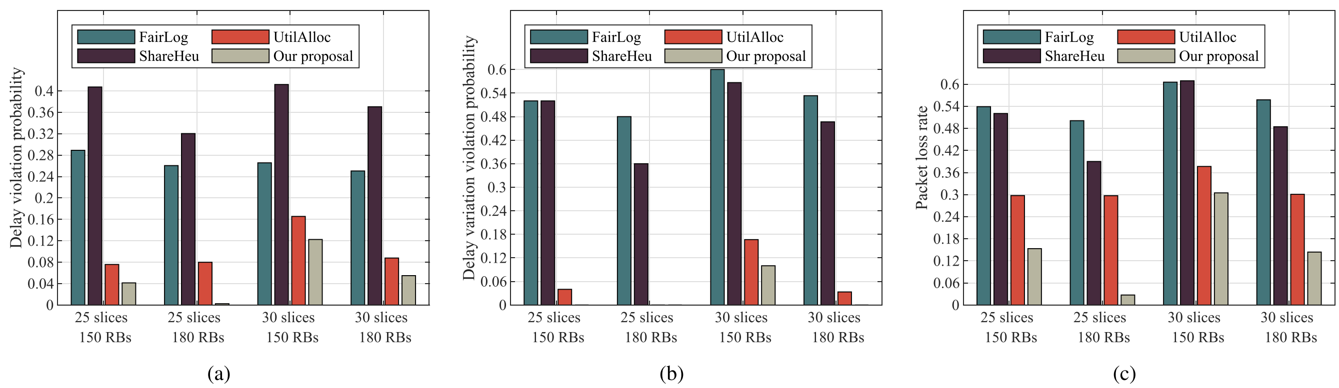


Fig. 10. QoS comparison of the existing methods and the proposed approach for known i.i.d. traffic arrivals.

variation constraints. The observed iteration counts remain far below I^2, which can be further reduced by employing a more sophisticated initialization than equal allocation. This allows the algorithm to converge before the commencement of the subsequent preparation phase, showcasing its effectiveness in managing dynamic radio resource allocation.

Fig. 8 and Fig. 9 show the probabilities of RAN slices satisfying their QoS requirements under cases of known time-invariant traffic arrivals and predicted time-varying traffic arrivals. A slice is deemed to meet its requirements in a slot only if its delay, delay variation, and packet loss rate are below specified thresholds. Proportional allocation tends to perform better than equal allocation by taking the differentiated traffic demands among RAN slices into account. Allocation based on average QoS exhibits the poorest outcomes, which can be explained as follows: the random channel fadings cause low probability but very large transmission delay and packet loss rate, skewing their average values and yielding unreasonable RB allocation. Our proposed algorithm consistently outperforms the baselines in all scenarios. Significant advantages of our algorithm can be found when there are insufficient radio resources to support the QoS-guaranteed services of all slices, indicating that our proposal can achieve a trade-off between resource utilization and stringent performance requirements. In addition, comparing results between the two cases, the performance of proportional allocation and average performance-based allocation deteriorates significantly, whereas our proposal exhibits more moderate degradation. The resilience to traffic prediction errors can be attributed to our performance metrics providing interval estimations rather than point estimations.

The QoS comparison between the proposed algorithm and state-of-the-art algorithms is depicted in Fig. 10. With the same resource budget, our algorithm achieves lower delay violation probability, delay variance violation probability, and packet loss rate compared to the existing methods. This highlights the effectiveness and robustness of our proposed QoS metric and the formulated resource allocation paradigm of minimizing the maximum violation probabilities among slices in ensuring excellent long-term performance in random environments. Furthermore, our method excels in handling non-i.i.d. traffic arrivals, a challenge that the existing methods struggle to address.

VI. Conclusion

In this paper, we presented a martingale-based SNC model for inter-slice radio resource allocation in deterministic communications systems. We derive an analytical relationship between the allocated radio resource quotas and the stringent upper bounds of delay violation probability, delay variation, and packet loss rate for both i.i.d. and non-i.i.d traffic arrivals. Additionally, a critical condition is identified for the optimal bandwidth allocation between two slices, which helps develop an efficient greedy algorithm for QoS-guaranteed radio resource allocation. Numerical results demonstrate that the derived bounds offer valid upper estimations of the required amount of radio resources to guarantee QoS. Furthermore, the proposed algorithm exhibits advantages in dealing with diverse QoS requirements across RAN slices.

Appendix A
Lemma 3

For any $\theta > 0$ and $k \geq 1$, the following inequality holds:

$$\mathbb{E}[h(\vec{a}^k(\tau), \theta)] > \mathbb{E}[h(\vec{s}^k(\tau), \theta)]. \tag{31}$$

Proof: Consider $\sigma > 0$ and define the stopping time T as

$$T \triangleq \sup\{\tau \geq 0 | A(t-\tau, t) - S(t-\tau, t) \geq \sigma\}. \tag{32}$$

In this context, $t - T$ represents the first time the queueing backlog exceeds σ. Assume that $\sum_{\tau=1}^{k} a(t-T-\tau) \geq \sum_{\tau=1}^{k} s(t-T-\tau)$ for some $k \geq 1$, then We have:

$$\begin{aligned} &A(t-T-k, t) - S(t-T-k, t) \\ &\quad = \left(\sum_{\tau=1}^{k} a(t-T-\tau) - \sum_{\tau=1}^{k} s(t-T-\tau)\right) + A(t-T, t) \\ &\qquad - S(t-T, t) \\ &\quad \geq \sigma, \end{aligned}$$

which contradicts the maximality of T. Hence, it follows that:

$$\sum_{\tau=1}^{k} a(t-T-\tau) < \sum_{\tau=1}^{k} s(t-T-\tau),$$

for any sample path. Since $h \in \mathscr{H}$ is non-increasing in each element of $\vec{a}^k(\tau)$, (31) holds for any $k \geq 1$. ■

APPENDIX B
PROOF OF THEOREM 1 AND THEOREM 3

By applying *Doob*'s optional sampling theorem to the supermartingale $\{U(\tau), 0 \leq \tau < t\}$, we establish that for every $k \in \mathbb{N}$, the following inequality holds:

$$\mathbb{E}[U(0)] \geq \mathbb{E}[U(T \wedge k)], \tag{33}$$

where $\wedge$ denotes the minimum operator and T is a stopping time as defined in (32). Utilizing this result, we can proceed as follows:

$$\begin{aligned}
&\mathbb{E}[h(\vec{a}^p(\tau),\theta)] \\
&\overset{(12)}{=} \mathbb{E}[U(0)] \overset{(33)}{\geq} \mathbb{E}[U(T \wedge k)] \\
&\geq \mathbb{E}[U(T \wedge k) 1_{T<k}] \\
&\overset{(12)}{=} \mathbb{E}[h(\vec{a}^p(T),\theta) e^{\theta A(t-T,t) - \theta S(t-T,t)} 1_{T<k}] \\
&= \mathbb{E}[h(\vec{a}^p(T),\theta)] \mathbb{E}[e^{\theta A(t-T,t) - \theta S(t-T,t)} | h(\vec{a}^p(T),\theta)] \\
&\quad \mathbb{P}(T<k) \\
&\overset{(31)}{\geq} \mathbb{E}[h(\vec{s}^p(T),\theta)] \mathbb{E}[e^{\theta A(t-T,t) - \theta S(t-T,t)}] \mathbb{P}(T<k) \\
&\overset{(32)}{\geq} \mathbb{E}[h(\vec{s}^p(T),\theta)] e^{\theta\sigma} \mathbb{P}(T<k) \\
&= \mathbb{E}[h(\vec{s}^p(\tau),\theta)] e^{\theta\sigma} \mathbb{P}(T<k).
\end{aligned}$$

The last equation holds due to the independent and identical distribution of s. Let $k \to \infty$, we arrive at the result:

$$\mathbb{P}\{Q(t) \geq q\} = \mathbb{P}\{T < \infty\} \leq \frac{\mathbb{E}[h(\vec{a}^p(\tau),\theta)]}{\mathbb{E}[h(\vec{s}^p(\tau),\theta)]} e^{-\theta q} = \kappa(\theta) e^{-\theta q}.$$

Since this inequality holds for each θ positive, we can optimize over θ to get Theorem 3.

Using the definition of delay violation probability, we can establish the following relationships:

$$\begin{aligned}
&\mathbb{P}\{W(t) > w\} \\
&\overset{(3)}{=} \mathbb{P}\{A(0,t) > D(0,t+w)\} \\
&\overset{(2)}{\leq} \mathbb{P}\{\sup_{0 \leq \tau \leq t} \{A(\tau,t) - S(\tau,t+w)\} > 0\} \\
&= \mathbb{P}\{\sup_{0 \leq \tau \leq t} \{A(\tau,t) - S(\tau,t)\} - S(t+1,t+w) > 0\} \\
&= \mathbb{P}\{Q(t) > S(t+1,t+w)\}.
\end{aligned}$$

Since $s(\tau)$'s are i.i.d., $Q(t)$ and $S(t+1, t+w)$ are independent. Consequently, we have:

$$\begin{aligned}
&\mathbb{P}\{Q(t) > S(t+1,t+w)\} \\
&\quad = \int \mathbb{P}(Q(t) > x) f_{S(t+1,t+w)}(x) \mathrm{d}x \\
&\quad \overset{(21)}{=} \kappa(\theta) \int e^{-\theta x} f_{S(t+1,t+w)}(x) \mathrm{d}x \\
&\quad = \kappa(\theta) \mathbb{E}[e^{-\theta S(t+1,t+w)}] \\
&\quad = \kappa(\theta) M_s^w(-\theta).
\end{aligned}$$

By combining the results obtained from the two formulas above, we can get Theorem 1.

APPENDIX C
PROOF OF THEOREM 2

By taking a double derivative of both sides of the geometric series $\sum_{w=0}^{\infty} x^w = \frac{1}{1-x}$ with respect to $\ln x$, we obtain the following result:

$$\sum_{w=0}^{\infty} w^2 x^w = \frac{x(1+x)}{(1-x)^3}. \tag{34}$$

Using this result, we can derive the upper bound of $\mathbb{E}[W^2(t)]$ as follows:

$$\begin{aligned}
\mathbb{E}[W^2(t)] &\overset{(19)}{\leq} \mathbb{E}[W'^2(t)] \\
&= \sum_{w=0}^{\infty} w^2 \mathbb{P}\{W'(t) = w\} \\
&\overset{(18)}{=} \sum_{w=0}^{\infty} w^2 (p'(w) - p'(w+1)) \\
&\overset{(15)}{=} \sum_{w=0}^{\infty} w^2 \kappa(\theta) (M_s^w(-\theta) - M_s^{w+1}(-\theta)) \\
&= \kappa(\theta)(1 - M_s(-\theta)) \sum_{w=0}^{\infty} w^2 M_s^w(-\theta) \\
&\overset{(34)}{=} \kappa(\theta) \frac{M_s(-\theta)(1 + M_s(-\theta))}{(1 - M_s(-\theta))^2}.
\end{aligned}$$

Since this inequality holds for each θ positive, we can optimize over θ to get Theorem 2.

APPENDIX D
PROOF OF THEOREM 4

Given that both $A(0,t)$ and $S(0,t)$ are i.i.d. incremental processes, Eq. (14) indicates that any feasible θ satisfies

$$\mathbb{E}[e^{\theta a - \theta s}] < 1, \tag{35}$$

When $h(\vec{a}^0(\tau),\theta) = 1$, the quantity $U(\tau) = e^{\theta(A(t-\tau,t) - S(t-\tau,t))}$. This leads to the following observation:

$$\begin{aligned}
U(\tau+1) &= e^{\theta(A(t-\tau-1,t) - S(t-\tau-1,t))} \\
&= U(\tau) e^{\theta a - \theta s}.
\end{aligned}$$

The conditional expected value of $U(\tau+1)$ is given by:

$$\begin{aligned}
&\mathbb{E}[U(\tau+1) | U(0), \cdots, U(\tau)] \\
&= \mathbb{E}[U(\tau) e^{\theta a - \theta s} | U(0), \cdots, U(\tau)] \\
&\quad = \mathbb{E}[U(\tau) | U(0), \cdots, U(\tau)] \mathbb{E}[e^{\theta a - \theta s}] \\
&\quad \overset{(35)}{<} U(\tau),
\end{aligned}$$

which suggests that $\{U(\tau), 0 \leq \tau < t\}$ forms a supermartingale. This further indicates that the i.i.d. incremental arrival process $A(0,t)$ can be characterized as a $(1,\theta,S)$-supermartingale-traffic arrival. By substituting $h(\vec{a}^0(\tau),\theta) = 1$ into Theorem 1 to 3, we arrive at the upper bounds for the delay violation probability, delay variation, and packet loss probability, as presented in Theorem 4.

APPENDIX E
PROOF OF THEOREM 5

Given the stationary AR model, Eq. (14) indicates that any feasible θ satisfies

$$\lim_{t\to\infty} \mathbb{E}^{\frac{1}{t}}[e^{\theta A(0,t)-\theta S(0,t)}] = \mathbb{E}[e^{\theta\mu-\theta s}] < 1. \quad (36)$$

For simplicity of notation, write $\tau' = t-\tau-1$. As established,

$$U(\tau) \overset{(12)}{=} h(\vec{a}^p(\tau),\theta)e^{\theta(A(\tau'+1,t)-S(\tau'+1,t))},$$

where

$$h(\vec{a}^p(\tau),\theta) \overset{(25)}{=} e^{-\frac{\theta}{1-\varphi}\sum_{k=1}^{p}\varphi_k\sum_{i=1}^{k}a(\tau'+1-i)}.$$

The following relationships can be derived:

$$\begin{aligned}
U(\tau+1) &= h(\vec{a}^p(\tau+1),\theta)e^{\theta(A(\tau',t)-S(\tau',t))}\\
&= U(\tau)\frac{h(\vec{a}^p(\tau+1),\theta)}{h(\vec{a}^p(\tau),\theta)}e^{\theta(a(\tau',t)-s)}\\
&= U(\tau)e^{\frac{\theta}{1-\varphi}\sum_{k=1}^{p}\varphi_k(a(\tau')-a(\tau'-k))}e^{\theta(a(\tau')-s)}\\
&= U(\tau)e^{\frac{\theta}{1-\varphi}(\varphi a(\tau')-\sum_{k=1}^{p}\varphi_k a(\tau'-k))}e^{\theta(a(\tau')-s)}\\
&\overset{(24)}{=} U(\tau)e^{\frac{\varphi\theta}{1-\varphi}a(\tau')-\frac{\theta}{1-\varphi}(a(\tau')-(1-\varphi)\mu-(1-\varphi)\sigma Z_{\tau'})}\\
&\quad e^{\theta(a(\tau')-s)}\\
&= U(\tau)e^{\theta(\mu+\sigma Z_{\tau'}-s)}.
\end{aligned}$$

Now, consider the conditional expectation of $U(\tau+1)$:

$$\begin{aligned}
&\mathbb{E}[U(\tau+1)|U(0),\cdots,U(\tau)]\\
&= \mathbb{E}[U(\tau)e^{\theta(\mu+\sigma Z_{\tau'}-s)}|U(0),\cdots,U(\tau)]\\
&= \mathbb{E}[U(\tau)|U(0),\cdots,U(\tau)]\mathbb{E}[e^{\theta(\mu+\sigma Z_{\tau'}-s)}]\\
&= U(\tau)\mathbb{E}[e^{\theta(\mu+0-s)}]\\
&\overset{(36)}{<} U(\tau),
\end{aligned}$$

which shows that $\{U(\tau), 0\le\tau<t\}$ is a supermartingale. Consequently, the AR incremental arrival process $A(0,t)$ can be characterized as a (h,θ,S)-supermartingale-traffic arrival.

To derive analytical expressions for the upper bounds of performance metrics, we denote the variable Y as follows:

$$Y = \sum_{k=1}^{p}\varphi_k\sum_{i=1}^{k}a(t-\tau-i).$$

Note that Y follows a normal distribution with a mean of $\mathbb{E}[Y] = \mu\sum_{k=1}^{p}k\varphi_k$ and a variance of $\mathbb{V}[Y] = v_Y^2$. The value of v_Y^2 can be calculated using the *Yule-Walker* equations. Therefore, according to Theorems 1-3, the upper bounds of the performance metrics are shown in (11), where the term $\kappa(\theta)$ is given by:

$$\begin{aligned}
\kappa(\theta) &= \frac{\mathbb{E}[h^p_{\vec{a}(\tau)}(\theta)]}{\mathbb{E}[h^p_{\vec{s}(\tau)}(\theta)]}\\
&= \frac{\mathbb{E}[e^{-\frac{\theta}{1-\varphi}Y}]}{\mathbb{E}[e^{-\frac{\theta}{1-\varphi}(\sum_{k=1}^{p}k\varphi_k s)}]}\\
&= \frac{M_Y(-\frac{\theta}{1-\varphi})}{M_s(-\frac{\theta}{1-\varphi}\sum_{k=1}^{p}k\varphi_k)}\\
&= e^{-\frac{\mu\theta\sum_{k=1}^{p}k\varphi_k}{1-\varphi}+\frac{\theta^2 v_Y^2}{2(1-\varphi)^2}}M_s^{-1}(-\frac{\theta}{1-\varphi}\sum_{k=1}^{p}k\varphi_k).
\end{aligned}$$

To provide examples, we present the closed-form analytical expressions for v_Y^2 considering the cases: $p=1$ and $p=2$.

1) $p=1$: The scenario with $p=1$ offers a straightforward calculation for $\kappa(\theta)$. In the context of a stationary 1-order AR model, we have the following expressions:

$$\begin{aligned}
a(\tau) &= \varphi_1 a(\tau-1)+(1-\varphi_1)\mu+(1-\varphi_1)\sigma Z,\\
Y &= \varphi_1 a(t-\tau-1).
\end{aligned}$$

This allows us to compute the variance v_a^2:

$$\begin{aligned}
v_a^2 &= \mathbb{V}(a(\tau))\\
&= \mathbb{V}(\varphi_1 a(\tau-1)+(1-\varphi_1)\mu+(1-\varphi_1)\sigma Z)\\
&= \varphi_1^2 v_a^2 + 0 + (1-\varphi_1)^2\sigma^2,
\end{aligned}$$

and thus $v_a^2 = \frac{(1-\varphi_1)^2\sigma^2}{1-\varphi_1^2}$. Consequently, the variance of Y can be expressed as:

$$v_Y^2 = \mathbb{V}[Y] = \varphi_1^2 v^2 = \frac{(1-\varphi_1)\varphi_1^2}{1+\varphi_1}\sigma^2.$$

2) $p=2$: For the more intricate situation of $p=2$, let γ_m denote the autocovariance function of $a(\tau)$, and we consider:

$$a(\tau) = \varphi_1 a(\tau-1)+\varphi_2 a(\tau-2)+(1-\varphi)\mu+(1-\varphi)\sigma Z$$

and

$$Y = (\varphi_1+\varphi_2)a(t-\tau-1)+\varphi_2 a(t-\tau-2).$$

The *Yule-Walker* equations for a 2-order AR process are

$$\gamma_1 = \varphi_1\gamma_0+\varphi_2\gamma_1, \gamma_2 = \varphi_1\gamma_1+\varphi_2\gamma_0.$$

This leads to $\gamma = \frac{\varphi_1}{1-\varphi_2}\gamma_0$. Hence, the variance of Y can be derived as follows:

$$\begin{aligned}
v_Y^2 &= \mathbb{V}[Y]\\
&= (\varphi_1+\varphi_2)^2\gamma_0+\varphi_2^2\gamma_0+2(\varphi_1+\varphi_2)\varphi_2\gamma_1\\
&= ((\varphi_1+\varphi_2)^2+\varphi_2^2+\frac{2(\varphi_1+\varphi_2)\varphi_1\varphi_2}{1-\varphi_2})\gamma_0,
\end{aligned}$$

where γ_0 satisfies the equation:

$$\begin{aligned}
\gamma_0 &= \mathbb{V}(a(\tau))\\
&= \mathbb{V}(\varphi_1 a(\tau-1)+\varphi_2 a(\tau-2)+(1-\varphi)\mu+(1-\varphi)\sigma Z)\\
&= \varphi_1^2\gamma_0+\varphi_2^2\gamma_0+2\varphi_1\varphi_2\gamma_1+0+(1-\varphi)^2\sigma^2\\
&= \varphi_1^2\gamma_0+\varphi_2^2\gamma_0+2\varphi_1\varphi_2\frac{\varphi_1}{1-\varphi_2}\gamma_0+(1-\varphi)^2\sigma^2,
\end{aligned}$$

leading to the expression $\gamma_0 = \frac{(1-\varphi_2)(1-\varphi_1-\varphi_2)^2}{1-\varphi_2-\varphi_1^2-\varphi_2^2+\varphi_2^3-\varphi_1^2\varphi_2}\sigma^2$.

References

[1] J. Zhu and S. Wang, "Delay-guaranteed resource allocation for deterministic communications: An efficient stochastic network calculus method," in *Proc. GLOBECOM*, Kuala Lumpur, Malaysia, Dec. 2023, pp. 7061–7066, doi: 10.1109/GLOBECOM54140.2023.10437275.

[2] S. Liu, T. Wang, and S. Wang, "Hardware impairment estimation in NB-IoT: A parallel multitask learning method," *IEEE Internet Things J.*, vol. 10, no. 8, pp. 6859–6869, Apr. 2023.

[3] F. Song, L. Li, I. You, and H. Zhang, "Enabling heterogeneous deterministic networks with smart collaborative theory," *IEEE Netw.*, vol. 35, no. 3, pp. 64–71, Jun. 2021.

[4] B. Hu and H. Gharavi, "A hybrid wired/wireless deterministic network for smart grid," *IEEE Wireless Commun.*, vol. 28, no. 3, pp. 138–143, Jun. 2021.

[5] L. Deng, G. Xie, H. Liu, Y. Han, R. Li, and K. Li, "A survey of real-time Ethernet modeling and design methodologies: From AVB to TSN," *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–36, Jan. 2022.

[6] A. Nasrallah et al., "Ultra-low latency (ULL) networks: The IEEE TSN and IETF DetNet standards and related 5G ULL research," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 88–145, 1st Quart., 2019.

[7] J. Krolikowski et al., "Joint routing and scheduling for large-scale deterministic IP networks," *Comput. Commun.*, vol. 165, pp. 33–42, Jan. 2021.

[8] 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, Sep. 2021.

[9] 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.

[10] *Management and Orchestration; Concepts, Use Cases and Requirements (Release 16)*, document 3GPP TS 28.530, Version 16.1.0, Dec. 2019.

[11] Y. Zhang, C. Peng, S. Xie, and X. Du, "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.

[12] L. Li, W. Chen, and K. B. Letaief, "Simple bounds on delay-constrained capacity and delay-violation probability of joint queue and channel-aware wireless transmissions," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2744–2759, Apr. 2023.

[13] E. Baccarelli, N. Cordeschi, and M. Biagi, "Conditionally optimal minimum-delay scheduling for bursty traffic over fading channels," *IEEE Trans. Veh. Technol.*, vol. 59, no. 7, pp. 3294–3310, Sep. 2010.

[14] J. Xie, D. Guo, X. Li, Y. Shen, and X. Jiang, "Cutting long-tail latency of routing response in software defined networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 3, pp. 384–396, Mar. 2018.

[15] C. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4132–4150, Jun. 2019.

[16] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Distributed federated learning for ultra-reliable low-latency vehicular communications," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1146–1159, Feb. 2020.

[17] C. T. Guo, L. Liang, and G. Y. Li, "Resource allocation for low-latency vehicular communications: An effective capacity perspective," *IEEE J. Sel. Area Commun.*, vol. 37, no. 4, pp. 905–917, Apr. 2019.

[18] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 92–105, 1st Quart., 2015.

[19] F. Ciucu, A. Burchard, and J. Liebeherr, "A network service curve approach for the stochastic analysis of networks," in *Proc. ACM SIGMETRICS*, Banff, AB, Canada, Jun. 2005, pp. 279–290.

[20] C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," *IEEE/ACM Trans. Netw.*, vol. 15, no. 6, pp. 1442–1453, Dec. 2007.

[21] M. Fidler, "An end-to-end probabilistic network calculus with moment generating functions," in *Proc. IEEE IWQoS*, New Haven, CT, USA, Jun. 2006, pp. 261–270.

[22] F. Ciucu, F. Poloczek, and J. Schmitt, "Sharp per-flow delay bounds for bursty arrivals: The case of FIFO, SP, and EDF scheduling," in *Proc. IEEE INFOCOM*, Toronto, ON, Canada, Apr. 2014, pp. 1896–1904.

[23] O. Adamuz-Hinojosa, V. Sciancalepore, P. Ameigeiras, J. M. Lopez-Soler, and X. Costa-Pérez, "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. 2023.

[24] 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, Apr. 2019.

[25] B. Yu, X. Chi, and X. Liu, "Martingale-based bandwidth abstraction and slice instantiation under the end-to-end latency-bounded reliability constraint," *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 217–221, Jan. 2022.

[26] K. Wang, X. Li, H. Ji, and X. Du, "Modeling and optimizing the LTE discontinuous reception mechanism under self-similar traffic," *IEEE Trans. Veh. Technol.*, vol. 65, no. 7, pp. 5595–5610, Jul. 2016.

[27] V. H. de la Peña, "A general class of exponential inequalities for martingales and ratios," *Ann. Probab.*, vol. 27, n o. 1, pp. 537–564, Jan. 1999.

[28] C. Tepedelenlioglu, A. Rajan, and Y. Zhang, "Applications of stochastic ordering to wireless communications," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4249–4257, Dec. 2011.

[29] M. Kallas, P. Honeine, C. Francis, and H. Amoud, "Kernel autoregressive models using Yule–Walker equations," *Signal Process.*, vol. 93, no. 11, pp. 3053–3061, Nov. 2013.

[30] S. Saha, A. Haque, and G. Sidebottom, "Deep sequence modeling for anomalous ISP traffic prediction," in *Proc. IEEE Int. Conf. Commun.*, Seoul, South Korea, May 2022, pp. 5439–5444.

[31] Y. Zhou, Z. Mo, Q. Xiao, S. Chen, and Y. Yin, "Privacy-preserving transportation traffic measurement in intelligent cyber-physical road systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3749–3759, May 2016.

[32] N. C. Beaulieu and C. Cheng, "Efficient Nakagami-m fading channel simulation," *IEEE Trans. Veh. Technol.*, vol. 54, no. 2, pp. 413–424, Mar. 2005.

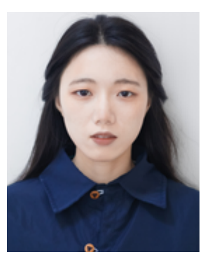
[33] H. Al-Zubaidy, J. Liebeherr, and A. Burchard, "Network-layer performance analysis of multihop fading channels," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 204–217, Feb. 2016.

[34] T. Jiang, L. Liu, and J. Li, "Analysis of the M/G/1 queue in multi-phase random environment with disasters," *J. Math. Anal. Appl.*, vol. 430, no. 2, pp. 857–873, Oct. 2015.

[35] M. K. Motalleb, V. Shah-Mansouri, S. Parsaeefard, and O. L. A. López, "Resource allocation in an open RAN system using network slicing," *IEEE Trans. Netw. Service Manag.*, vol. 20, no. 1, pp. 471–485, Mar. 2023.

[36] W. Shi et al., "Two-level soft RAN slicing for customized services in 5G-and-beyond wireless communications," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4169–4179, Jun. 2022.

[37] Y. Cui, X. Yang, P. He, D. Wu, and R. Wang, "O-RAN slicing for multi-service resource allocation in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 9272–9283, Jul. 2024.



Juan Zhu (Student Member, IEEE) received the B.S. degree from Nanjing University, Nanjing, China, in 2022, where she is currently pursuing the M.S. degree with the School of Electronic Science and Engineering. Her research interests include machine learning and deterministic networking.

Shaowei Wang (Senior Member, IEEE) received the Ph.D. degree from Wuhan University, Wuhan, China, in 2006. In 2006, he joined the School of Electronic Science and Engineering, Nanjing University, Nanjing, China, as a Faculty Member, where he is currently a Full Professor. From 2012 to 2013, he was a Visiting Scholar/a Professor with Stanford University, Stanford, CA, USA, and The University of British Columbia, Vancouver, BC, Canada. His research interests include communications and networking, operations research, and machine learning.