A Generalized Delay and Backlog Analysis for Multiplexing URLLC and eMBB: Reconfigurable Intelligent Surfaces or Decode-and-Forward?

Haoran Peng, *Member, IEEE*, Ching-Chieh Hsia,
Zhu Han, *Fellow, IEEE*, and Li-Chun Wang, *Fellow, IEEE*

***Abstract*—By creating multipath backscatter links and amplify signal strength, reconfigurable intelligent surfaces (RIS) and decode-and-forward (DF) relaying are shown to degrade the latency of the ultrareliable low-latency communications (URLLCs) and enhanced mobile broadband (eMBB) multiplexing system. This study investigates the delay and backlog violation behavior of URLLCs and eMBB multiplexing systems supported by different technologies, e.g. RIS and DF relay, for different scheduling policies of static priority, nonpreemption, and earliest deadline first. A tight analysis approach based on the Martingale theory was proposed to evaluate the serviceability of URLLC and eMBB multiplexing systems. On this basis, the Martingale theory analyzes the delay and backlog bounds by transforming the arrival and service processes into exponential forms of the moment generating function. Furthermore, this study derives the closed-form expression of delay and backlog bound for the URLLCs and eMBB multiplexing in two-hop heterogeneous communication networks. Numerical results demonstrate that the proposed Martingale-based tightly analytical method outperforms the state-of-the-art classic stochastic network calculus for evaluating delay and backlog violation in URLLC and eMBB multiplexing systems.**

***Index Terms*—Martingale, reconfigurable intelligent surfaces, relay, URLLC, eMBB, stochastic network calculus**

I. Introduction

The Third-Generation Partnership Project (3GPP) defines that a main service for the fifth-generation cellular networks (5G) new radio (NR) is to efficiently support multiplexing ultrareliable low-latency communications (URLLC) and enhanced mobile broadband (eMBB) [1]–[4]. URLLC traffic aims to achieve an extremely low latency (0.25-0.3 msec/packet) while guaranteeing high reliability of the 99.999% packet success probability [5]. The eMBB transmission aims to provide high-throughput (gigabit per second) data rates with millisecond-level latency [6], [7]. To improve spectrum efficiency, URLLC traffic is scheduled to puncture an ongoing eMBB traffic [5], [8]–[10]. However, eMBB services are likely to be interrupted several times during transmission to guarantee high reliability and low latency for URLLC traffic [6]. Therefore, URLLC and eMBB multiplexing suffer from a violation behavior of delay and backlog.

A. Motivations

Reconfigurable intelligent surfaces (RIS) and decode-and-forward (DF) relay were shown to efficiently improve the system capacity by strengthening the received signal [11], [12]. RIS creates multipath backscatter links by intelligently controlling an array of passive reflecting elements [13], [14]. DF relays improve the propagation path by decoding, remodulating, and retransmitting the implied radio frequency (RF) signals [15], [16]. However, the latency and backlog violation behavior persists even when the URLLC and eMBB coexistence system is supported by RIS or relay [17], [18]. Furthermore, the service capabilities of the RIS and DF relays depend on their characteristics, including hardware complexity, noise generation, spectral efficiency, and power budget [19]. The number of passive reflection elements also affects the performance of RIS-assisted URLLC and eMBB coexistence systems [15]. Besides, latency and backlog violations in multi-hop heterogeneous communication networks are difficult to be estimated due to the different serviceability of each node. Therefore, it is critical to model and analyze the end-to-end transmission latency and backlog to evaluate the performance of the RIS and DF relay in the URLLC and eMBB multiplexing system, as well as the multi-hop heterogeneous communication networks.

Various solutions have been explored in the literature to analyze latency and backlog violation behavior in wireless systems [20]–[28]. Queuing theory has been widely used to analyze the latency performance of wireless systems due to its ability to profile the system serving customers [23]. However,

This work has been partially funded by the National Science and Technology Council under the Grants MOST 110-2221-E-A49-039-MY3, and MOST 111-2221-E-A49-071-MY3, and NSTC 111-2634-F-A49-010, and NSTC 112-2218-E-A49-023-, Taiwan. This work was also financially supported by the Center for Open Intelligent Connectivity from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

This work was supported by the Higher Education Sprout Project of the National Yang Ming Chiao Tung University and Ministry of Education (MOE), Taiwan.

This work is partially supported by NSF CNS-2107216, CNS-2128368, CMMI-2222810, ECCS-2302469, US Department of Transportation, Toyota and Amazon.

The evaluation code and data are included in the repository (https://github.com/Haoran-Peng/MartingaleMultiHop).

Haoran Peng is with the College of Electrical and Electronic Engineering, Wenzhou University, Wenzhou City, Zhejiang Province, China 325035. (hrmpeng@gmail.com).

Ching-Chieh Hsia, and Li-Chun Wang are with the Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu 300, Taiwan. (cchsia.ee10@nycu.edu.tw, wang@nycu.edu.tw).

Zhu Han is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701. (hanzhu22@gmail.com).

Corresponding author: Li-Chun Wang, wang@nycu.edu.tw.

complex queuing problems result to difficulty in achieving a steady state in wireless systems, thus challenging queuing theory. Furthermore, the second-order statistical analysis of the probability density function and variance in queuing theory is complicated and loses effectiveness in nonlinear hybrid service systems. Stochastic network calculus (SNC) can transform a complex nonlinear network system into a linear system by taking advantage of the min-plus algebra [29]–[31]. However, SNC renders loose bounds because the curves for the service lower bound and the arrival upper bound are calculated using the Boole's inequality, which regards each time instance of the stochastic process separately and brings a looseness in the tail probability [32]. Furthermore, SNC was shown to lost effectiveness in delay and backlog analysis for multi-hop heterogeneous communication networks [33].

The Martingale theory effectively overcomes the looseness problem in SNC by applying the optimal stopping theorem of supermartingales, which is a variant of the Doob's inequality and sharper than the Boole's inequality [31], [32], [34], [35]. The conditional expectation of the future state values in a supermartingale process is bounded by the current state value [32]. Martingale envelopes can provide tight bounds of delay and backlog by exceeding a given value over a time interval, while SNC transforms the moment generating function (MGF) to the Chernoff bound [28]. Therefore, this study leverages Martingale to accurately evaluate the delay and backlog violation behavior of RIS and DF relays in the URLLC and eMBB coexistence system as well as the service capability of RIS with different numbers of reflective elements.

B. State-of-the-art methods

1) URLLC and eMBB multiplexing systems: Various resource management schemes were developed based on the superposition/puncturing scheme to improve the spectral efficiency for URLLC and eMBB traffic coexistence [5], [8]. In [5], a joint URLLC and eMBB traffic scheduler was developed to guarantee URLLC priority and eMBB utility maximization for different models, such as linear, convex, and threshold models. In [8], the URLLC/eMBB scheduling problem was formulated as a mixed integer nonlinear programming to minimize the loss of eMBB data rate while guaranteeing the quality of service (QoS) constraints of URLLC and eMBB traffics. In [3], deep reinforcement learning (DRL) was explored to maximize the eMBB data rate while satisfying the URLLC reliability constraint. In [36], a block coordinated descent algorithm was proposed to minimize URLLC power consumption under various QoS constraints in the downlink radio access network of URLLC and eMBB traffics. In [37], the resource allocation of URLLC and eMBB network slices was formulated as a multitimescale problem, and a DRL-based algorithm was proposed to efficiently solve the problem and achieve high throughput.

2) RIS/relay-assisted multiplexing URLLC and eMBB: RIS has been shown to significantly reduce URLLC latency and improve channel gains in URLLC and eMBB multiplexing systems [12], [18], [38]. In [39], RIS was used to maximize URLLC reliability and minimize eMBB rate loss by jointly optimizing RIS phase shift, frequency, and base station transmission power. RIS-aided radio access network was shown to effectively increase the uplink URLLC reliability and eMBB throughput simultaneously under both the heterogeneous orthogonal multiple access (OMA) and heterogeneous nonorthogonal multiple access (NOMA) frameworks [11]. In [18], a two-phase relay-assisted protocol was developed to support URLLC uplink and minimize transmission power consumption by jointly scheduling relay, transmission power, frequency, and decoding error probability. In [12], a multiunmanned aerial vehicle (UAV) relay network was developed to improve system throughput and reduce power consumption for the URLLC and eMBB multiplexing system by jointly optimizing transmit power, user scheduling, and bandwidth.

3) Delay and backlog violation analysis: In [29], SNC was adopted to assess the probability of delay violation of mobile edge computing networks. An SNC model was applied to estimate the upper bounds of the violation probability of both the peak age of information (AoI) and the delay for URLLC services supported by AoI and finite blocklength coding [40]. An SNC-based propagation delay embedded min-plus convolution approach was presented to analyze the leftover services received by the per-flow traffic in satellite data relay networks [20]. In [41], Martingale was used to derive the delay bounds of cloud centers, edge nodes, and vehicular fog nodes in heterogeneous vehicular networks. A Martingale-based approximation theory was adopted to analyze the end-to-end delay in the multiqueuing edge computing node system [35]. In [42], the Martingale theory was applied to analyze the stochastic end-to-end delay bound with the ALOHA-NOMA scheme in an edge computing scenario.

4) Limitations: The above studies [3], [5], [8], [11], [12], [18], [36]–[39] achieved outstanding contributions in the improvement of URLLC and eMBB multiplexing system performance. However, these studies lack a comprehensive analysis of delays and backlog violation behaviors. The latency and backlog analysis in [20], [29], and [40] suffers from the looseness boundaries derived from the SNC model. The Martingale theory provided a tight delay probability bound in [35], [41], and [42], while the analysis of the service ability of multiplexing URLLC and eMBB was ignored. To the best of our knowledge, this study is the first to compare the RIS and DF relay from the perspective of tight latency and backlog analysis in multiplexing URLLC and eMBB.

C. Contributions and organizations

This study aims to analyze the delay and backlog violation behaviors of the URLLC and eMBB multiplexing system supported by RIS or DF relays. As multiplexing methods for URLLC and eMBB have already been investigated in [4], the focus of this study is on the performance analysis of URLLC and eMBB multiplexing systems, rather than proposing any multiplexing approaches. Various scheduling schemes for URLLC and eMBB multiplexing were investigated. These schemes include static priority (SP), nonpreemption, and earliest deadline first (EDF). A tight-bound analysis based on the Martingale theory was then proposed to compare the

performance of RIS and relay by depicting the arrival and service processes. The proposed Martingale envelope model derives an exponential transformation with the multiplication of arrival and service flows, whereas the SNC model treats each time instance of the arrival and service processes separately and cannot capture their main properties [32]. Therefore, the Martingale envelope is tighter than the classic SNC envelope model. The numerical results demonstrated that the proposed Martingale theory-based model outperforms the classic SNC in the queuing system calculus. This study also provides a comprehensive comparison between RIS and DF relay with respect to delays and backlog violations in the URLLC and eMBB multiplexing system, as well as the two-hop heterogeneous communication network. The main contributions of this study are summarized as follows.

- This study investigated the behaviors of delay and backlog violation behaviors in the URLLC and eMBB multiplexing system that is supported by RIS and DF relays. Furthermore, to explore the effectiveness of RIS and DF relays, various multiplexing scheduling schemes, such as SP, nonpreemption, and EDF, were studied. This study derived the closed-form expressions of the delay and backlog bounds of the URLLC and eMBB multiplexing queuing system for single-hop and two-hop communication networks based on the Martingale theory. The Martingale envelope provides a tight bound on delay and backlog probability by taking the multiplication of arrival and service processes into an exponential transformation.
- The simulation results show that the proposed Martingale theory-based approach outperforms the state-of-the-art SNC in tightly analyzing the latency and backlog violation probability for the URLLC and eMBB coexistence system. Furthermore, this study empirically compared the performance of the DF relay and RIS with respect to improving the service capability of the URLLC and eMBB multiplexing system.

The accurate analysis and estimation of delay and backlog can benefit several practical applications. For example, autonomous vehicles require URLLC and eMBB to ensure safe driving and entertainment, respectively [43]. Virtual and augmented reality applications also require high data rates and low latency to provide a seamless user experience [44]. Additionally, industrial automation and control systems can benefit from 5G's ability to support eMBB communication and URLLC [45]. The remainder of this paper is organized as follows. Section II presents the system model. The Section III presents the SNC-based queuing network model. Section IV derives the Martingale envelope of arrival and service processes. Section V illustrates the Martingale theory-based analysis model proposed for different scheduling policies. Section VI evaluates the effectiveness of the proposed Martingale theory-based analysis model, and discusses the performance of DF relay and RIS in the URLLC and eMBB multiplexing system. Finally, Section VII gives the concluding remarks.

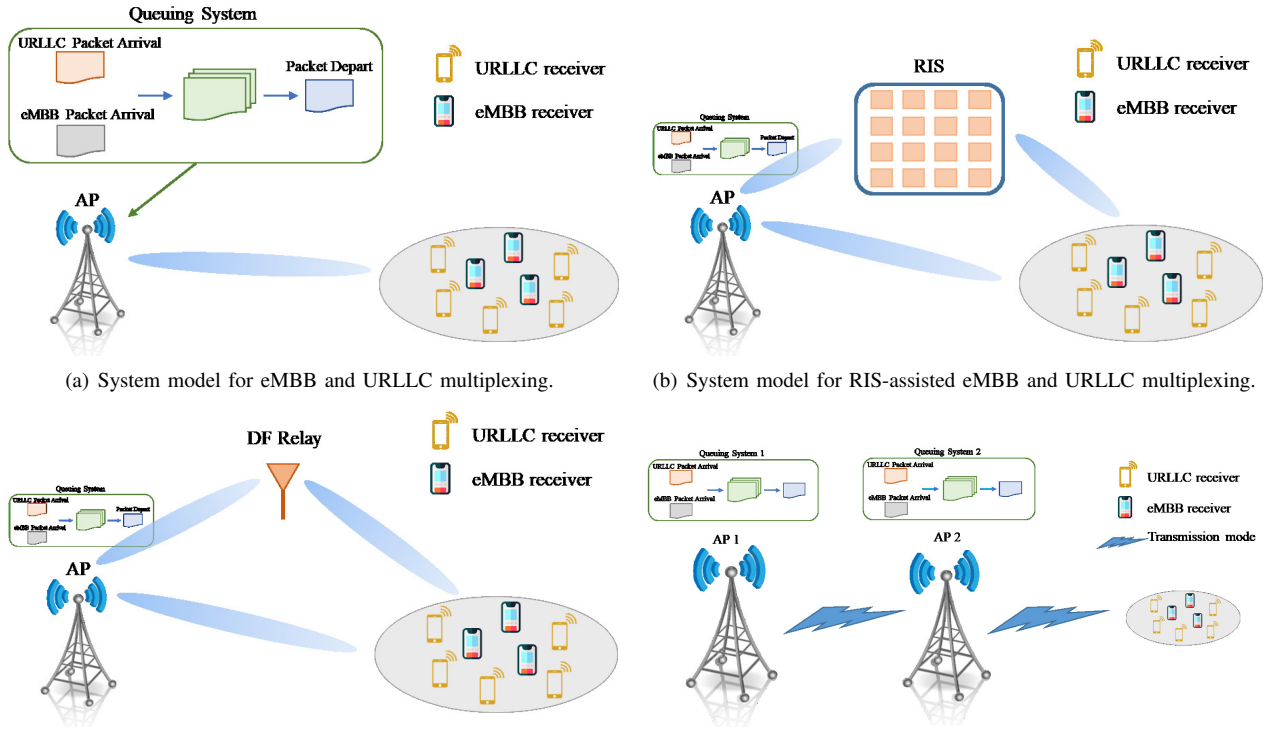
Notations: Table I summarizes a partial of the important notations in this paper.

TABLE I
GLOSSARY OF NOTATIONS

Symbol	Description
I_v	The inter-beam interference
$\boldsymbol{w}_b$	The active precoding/beamforming vectors for receiver v
$\boldsymbol{s}_v$	The signals for receiver v
p_b^u	The transmit power for URLLC receiver
p_b^e	The transmit power for eMBB receiver
$\boldsymbol{h}_{b,v}$	The channel between the AP and the receiver v
F_B	The bandwidth
M	The number of antennas at AP
N	The number of antennas at RIS
$\mathcal{L}$	The number of antennas at DF relay
$\boldsymbol{G}$	The channel from AP to RIS
$\boldsymbol{\Phi}$	The diagnonal phase-shifting matrix of the RIS
$\boldsymbol{h}_{r,v}$	The channel from the RIS to receiver v
$\boldsymbol{h}_{b,d}$	The channel from the AP to DF relay
$\boldsymbol{h}_{d,v}$	The channel from the DF relay to receiver v
A^u	The arrival processes for the URLLC traffic
A^e	The arrival processes for the eMBB traffic
p_{ij}	The probability of Markov chain from the state i transit to state j
ζ	The number of states in the Markov chain
$B^i(t)$	The backlog of the i hop system at time slot t
$A^i(t)$	The arrival process for the ith hop at the tth time slot
$S^i(t)$	The service process for the ith hop at the tth time slot
$D^i(t)$	The departure process for the ith hop at the tth time slot
$\bar{s}_\eta^i$	The instant service rate
$T_{a^u}^\theta$	The exponentially transformed transition probability matrix for A^u
$T_{a^e}^\theta$	The exponentially transformed transition probability matrix for A^e
Γ	The optional stopping time
$G_{A^u}(\eta)$	The supermartingale envelopes of URLLC arrival process
$G_{A^e}(\eta)$	The supermartingale envelopes of eMBB arrival process
$G_S(\eta)$	The supermartingale envelopes of service processes

II. SYSTEM MODEL

In this study, a multiantenna access point (AP) and multiple URLLC and eMBB receivers of a single antenna are considered. The arrival of the data packet for URLLC and eMBB traffic follows the Poisson process, which is a type of Markov process. In a Markov process, the future state of the system depends only on its current state and not on any previous state. For a Poisson process, arrival packets occur randomly over time with a constant rate parameter. By the memoryless property of the Poisson process, this conditional



(a) System model for eMBB and URLLC multiplexing.

(b) System model for RIS-assisted eMBB and URLLC multiplexing.

(c) System model for DF relay-assisted eMBB and URLLC multiplexing.

(d) System model for two-hop heterogeneous networks.

Fig. 1. (a) The AP serves the URLLC and eMBB receivers via multiple input single output (MISO) communications. (b) An RIS is deployed to improve the serviceability of the eMBB and URLLC multiplexing system by constructing multiple line-of-sight (LoS) links between the AP and users. (c) DF relay supports the eMBB and URLLC multiplexing system. (d) The transmission between two nodes without the help of RIS and DF relay.

probability depends only on the current state of the system and not on any previous states. Poisson process has been widely used as a mathematical model for random events that occur over time, such as the URLLC and eMBB traffic arrival process. Furthermore, using the Poisson process model can capture the randomness and unpredictability of traffic arrivals and help analyze network performance [10], [17], [37], [46]. Figure 1(a) shows the URLLC and eMBB multiplexing system model. The AP directly serves receivers through multiinput, single-output (MISO) communications. RIS and relay are explored to enhance the serviceability of the URLLC and eMBB multiplexing system. Therefore, the delay and backlog of the queue of packets that arrive at the AP can be reduced by improving the system capacity. Figures 1(b) and 1(c) illustrates the URLLC and eMBB multiplexing system supported by the RIS and DF relay, respectively. The transmission between two nodes can be established through MISO communications, RIS, and DF relay. Figure 1(d) depicts the two-hop heterogeneous communication network for a URLLC and eMBB multiplexing system without the help of RIS and DF relay. Two APs are connected in tandem and have different service capabilities. Figure 2 illustrates the two-hop system model for the RIS-assisted URLLC and eMBB multiplexing. In the first hop, the first AP serves as the transmitter of the signal, while the second AP serves as the receiver, obtaining the signal from both the direct link from the first AP and the reflected link from the first RIS. In the second hop, the second AP serves as the transmitter of the signal, while the URLLC/eMBB user serves as the receiver, receiving the signal from both the direct link from the second AP and the reflected link from the second RIS. This study assumes a different serviceability of the two APs, which results in different channel bandwidths in the first and second hops. To achieve a two-hop heterogeneous network for DF relay-aided URLLC and eMBB multiplexing, the RIS can be replaced with a multiantenna relay.

A. MISO transmission models

The AP is equipped with $\mathcal{M} = \{1, 2, \cdots, M\}$ antennas. The sets of URLLC and eMBB receivers with a single antenna are denoted as $\mathcal{U} = \{1, 2, \cdots, u, \cdots, U\}$ and $\mathcal{E} = \{1, 2, \cdots, e, \cdots, E\}$, respectively. U and E represent the number of URLLC and eMBB receivers, respectively. This study assumes that the AP uses the same transmit power level to serve each URLLC or eMBB receiver. Therefore, the set of all receivers can be expressed as $\mathcal{V} = \{\mathcal{U}, \mathcal{E}\} = \{1, 2, \cdots, v, \cdots, V\}$. The number of receivers in this system is $V = U + E$. The estimation of the channel state information (CSI) brings latency to the multiplexing system. However, this study focuses on the analysis of delay and backlog for URLLC and eMBB multiplexing. Following [35], [47], this study assumes that the AP perfectly obtains the CSI of the channel. The baseband signal transmitted from the AP to each URLLC/eMBB receiver v can be defined as

$$\boldsymbol{x} = \boldsymbol{w}_b \boldsymbol{s}_v, \tag{1}$$

where $\boldsymbol{s}_v \sim \mathcal{CN}(0, 1)$ are signals for receiver v, which are assumed to be independent and identically distributed

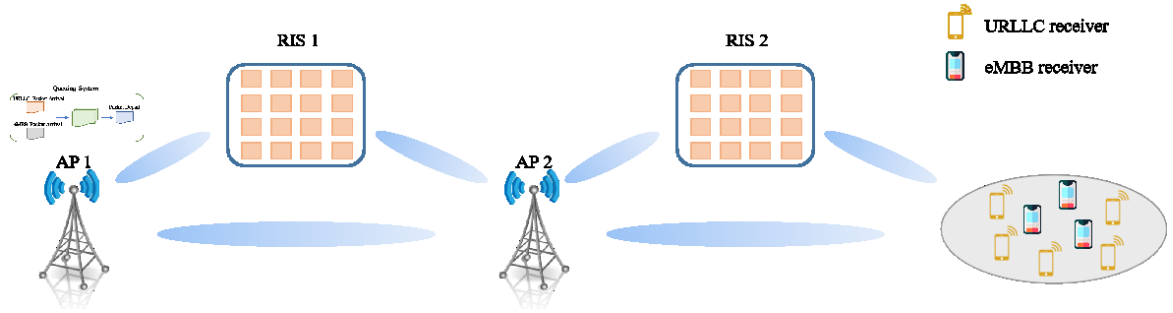


Fig. 2. The considered two-hop system model for RIS-assisted URLLC and eMBB multiplexing.

circularly symmetric complex Gaussian random variables with zero mean and unit variance [48]. $\boldsymbol{w}_b \in \mathbb{C}^{M\times 1}$ denotes the active transmit precoding/beamforming vectors for receiver v. $p_b = \|\boldsymbol{w}_b\|^2$ is the transmission power of the AP for each receiver v. To ensure QoS for different receivers, the transmit power budget for URLLC and eMBB receivers is p_b^u and p_b^e, respectively. Therefore, the signal received at each URLLC/eMBB user v can be denoted as

$$y_m = \boldsymbol{w}_b \boldsymbol{h}_{b,v}^H \boldsymbol{s}_v + n, \tag{2}$$

where $\boldsymbol{h}_{b,v} = \left[h_{b,v}^1, h_{b,v}^2, \cdots, h_{b,v}^M\right] \in \mathbb{C}^{M\times 1}$ is the channel vector between the AP and the receiver v with flat Rayleigh fading and path loss. Following [49], this study uses maximum ratio transmission (MRT) precoding to optimize the overall performance of the multiantenna system by reducing interference, increasing signal quality, and improving reliability and stability. MRT precoding is easy to implement by calculating the complex weights for each antenna and combining the signals of each antenna using the calculated weights, such as $\mathbf{w}_b = \mathbf{h}_{b,v}^*/\|\mathbf{h}_{b,v}\|$. $n \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) at each receiver's device. Therefore, the system capacity of the MISO transmission model can be obtained as

$$C_{MISO} = F_B \log_2\left(1 + \frac{p_b \left|\boldsymbol{h}_{b,v}^H\right|^2}{\sigma^2 + I_v}\right), \tag{3}$$

where F_B and I_v is the bandwidth and inter-beam interference, respectively. Following [50], this study considers the worst case of inter-beam interference for standard sidelobe level of uniform linear arrays (ULA), which is around 12.3 dB.

B. RIS-assisted transmission model

The RIS is equipped with $\mathcal{N} = \{1, 2, \cdots, N\}$ passive metasurface elements to reflect the impinging RF signal to the receivers. The RIS can provide multipath virtual line-of-sight (LoS) links to improve transmission throughput between the AP and receivers. The desired signal received at each URLLC/eMBB user can be denoted as

$$y_r = \boldsymbol{w}_b^r \left(\boldsymbol{h}_{b,v}^H + \boldsymbol{G}^H \boldsymbol{\Phi}^H \boldsymbol{h}_{r,v}^H\right) \boldsymbol{s}_v + n, \tag{4}$$

where $\boldsymbol{h}_{r,v} = \left[h_{r,v}^1, h_{r,v}^2, \cdots, h_{r,v}^N\right] \in \mathbb{C}^{N\times 1}$ is the channel vector from the RIS to receiver v with flat Rayleigh fading and path loss. $\boldsymbol{G} \in \mathbb{C}^{M\times N}$ is the channel from the AP to the RIS with flat Rayleigh fading and path loss. The precoding vector $\boldsymbol{w}_b^r$ is optimized by MRT for the receiver v. $\boldsymbol{\Phi}$ is the diagonal phase-shifting matrix of the RIS and is given by

$$\boldsymbol{\Phi} = \mathrm{diag}(\alpha_1 e^{j\theta_1}, \cdots, \alpha_N e^{j\theta_N}) \in \mathbb{C}^{N\times N}, \tag{5}$$

where $\theta_n \in [0, 2\pi]$ is the proper phase shift producing a correct passive beamforming for each receiver v. Specifically, phase shift in RIS is discrete because the RIS operates by adjusting the phase of the reflected signal to control the direction of the beam. The use of discrete phase shifters simplifies the design and control of the RIS, making it easier to implement and operate [51]. $j = \sqrt{-1}$ is the imaginary unit, and $a_n \in [0, 1]$ is the fixed-amplitude reflection coefficient of the metasurface element n in the RIS. In this study, $a_n = 1$ is set for each element n in the RIS to the maximum signal reflection efficiency [52]. Therefore, the capacity of the RIS-assisted URLLC and eMBB multiplexing system can be expressed as

$$C_{RIS} = F_B \log_2\left(1 + \frac{p_b \left|\boldsymbol{h}_{b,v}^H + \boldsymbol{G}^H \boldsymbol{\Phi}^H \boldsymbol{h}_{r,v}^H\right|^2}{\sigma^2 + I_v}\right). \tag{6}$$

This study aims to investigate the maximum serviceability of the RIS-assisted URLLC and eMBB multiplexing system. Various state-of-the-art methods, such as generalized Benders decomposition, were used to explore the phase shift design of RIS-assisted wireless systems [52]. Therefore, in this study, the design of phase shifts will not be elucidated.

C. DF relay-assisted transmission model

Following [15], this study considers a half-duplex and repetition-encoded DF relay consisting of two equal-sized transmission phases of decoding and forwarding. The DF relay is equipped with $\mathcal{L}$ antennas.

1) Decoding phase: The signal transmitted directly from the AP to each receiver v can be expressed as

$$y_v^1 = \boldsymbol{h}_{b,v}^H \boldsymbol{w'}_b \boldsymbol{s}_v + n_v^1, \tag{7}$$

where $\boldsymbol{w'}_b \in \mathbb{C}^{M\times 1}$ denotes the active transmit precoding/beamforming vectors in the AP and $p_1 = \|\boldsymbol{w'}_b\|^2$ and $n_v^1 \in \mathcal{CN}(0, \sigma^2)$ are the transmit power and the v-th receiver AWGN, respectively. The signal received on the DF relay is expressed as

$$y_d^1 = \boldsymbol{h}_{b,d}^H \boldsymbol{w'}_b \boldsymbol{s}_v + n_d^1, \tag{8}$$

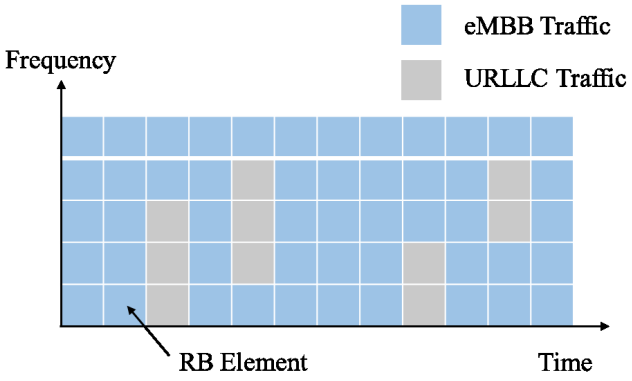


Fig. 3. The static priority model for high-priority URLLC and low-priority eMBB multiplexing.

where $\boldsymbol{h}_{b,d} \in \mathbb{C}^{M\times\mathcal{L}}$ is the channel from the AP to the DF relay with flat Rayleigh fading and path loss. $n_d^1 \in \mathcal{CN}(0,\sigma^2)$ is the AWGN on the relay. The $\boldsymbol{h}_{b,d}$ is given by

$$\boldsymbol{h}_{b,d} = \begin{bmatrix} h_{b,d}^{11} & h_{b,d}^{12} & \cdots & h_{b,d}^{1\mathcal{L}} \\ h_{b,d}^{21} & h_{b,d}^{22} & \cdots & h_{b,d}^{2\mathcal{L}} \\ \vdots & \vdots & \ddots & \vdots \\ h_{b,d}^{M1} & h_{b,d}^{M2} & \cdots & h_{b,d}^{M\mathcal{L}} \end{bmatrix}. \tag{9}$$

The received signal y_d^1 is decoded in the current phase and then encoded for transmission to the receiver in the next phase.

2) Forwarding phase: The received signal at each receiver v of the DF relay is denoted as

$$y_v^2 = \boldsymbol{h}_{d,v}^H \boldsymbol{w}_d \boldsymbol{s}_v + n_v^2, \tag{10}$$

where $\boldsymbol{w}_d \in \mathbb{C}$ denotes the active precoding vector for the receiver v and $p_2 = \|\boldsymbol{w}_d\|^2$ and $n_v^2 \in \mathcal{CN}(0,\sigma^2)$ are the transmit power of the DF relay and the noise at receiver v in the forwarding phase, respectively. $\boldsymbol{h}_{d,v} = \{h_{d,v}^1, h_{d,v}^2, \cdots, h_{d,v}^{\mathcal{L}}\} \in \mathbb{C}^{\mathcal{L}\times 1}$ represents the channel of the DF relay to the receiver v with flat Rayleigh fading and path loss. The achievable rate of the URLLC and eMBB multiplexing system supported by the DF relay can be denoted as

$$C_{DF} = \frac{1}{2} F_B \log_2 \left[1 + \min\left(\frac{p_1 \left|\boldsymbol{h}_{b,d}\right|^2}{\sigma^2 + I_v}, \frac{p_1 \left|\boldsymbol{h}_{b,v}\right|^2 + p_2 \left|\boldsymbol{h}_{d,v}\right|^2}{\sigma^2 + I_v} \right) \right]. \tag{11}$$

Following [15], this study sets $p1 + p2 = 2 \times p_b$.

D. URLLC and eMBB multiplexing scheduling model

This study investigates three scheduling policies, namely, SP, nonpreemption, and EDF, for URLLC and eMBB multiplexing to discuss the serviceability of MISO, RIS, and DF relay-assisted systems.

1) SP scheduling: As shown in Figure 3, the URLLC packet preempts a part of the ongoing eMBB transmission and spans multiple frequency bands. In the multiplexing system, high-priority URLLC transmission may interrupt low-priority eMBB transmission several times. URLLC traffic has the highest priority and occupies the channel as long as it is present, while eMBB traffic can only use the channel if there are no URLLC packets. Without generality loss, this study assumes that the new arriving URLLC packet will be blocked when other URLLC packets are transmitting via the current spectrum resource.

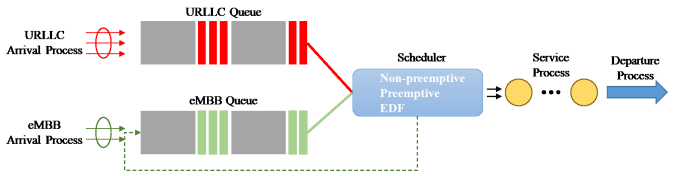


Fig. 4. The discrete-time network model for coexisting URLLC and eMBB.

2) Nonpreemptive scheduling: The arriving URLLC and eMBB packets are scheduled according to their arrival order in the nonpreemptive scheduling policy. No transmission interruption occurs in this scheduling because all services have the same priority.

3) Earliest deadline first scheduling: EDF schedules each service by its arrival time, the required execution time, and the deadline to ensure that all tasks are completed by the expected deadline. The priority of each packet depends on the current execution progress and its deadline.

III. Queuing Network Model for URLLC and eMBB Multiplexing

Figure 4 illustrates the discrete-time network model for the URLLC and eMBB multiplexing system. The discrete-time network model consists of three parts, namely, the arrival, service, and departure processes. The URLLC and eMBB traffic arrival processes are defined as

$$A^u(x,y) = \sum_{\eta=x}^{y} a_\eta^u \tag{12}$$

and

$$A^e(x,y) = \sum_{\eta=x}^{y} a_\eta^e, \tag{13}$$

respectively. Here x and y are the time interval of the arrival process, and a_η and b_η represent the instantaneous Markov arrival process in the time interval η for URLLC and eMBB, respectively. $A_u(0,\eta) = A_u(\eta)$ and $A_e(0,\eta) = A_e(\eta)$ represent the cumulative arrival curves of URLLC and eMBB from the initial time interval to η, respectively. The transition matrix of the arrival Markov chain can be represented as

$$T_a = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1\zeta} \\ p_{21} & p_{22} & \cdots & p_{2\zeta} \\ \vdots & \vdots & \ddots & \vdots \\ p_{\zeta 1} & p_{\zeta 2} & \cdots & p_{\zeta\zeta} \end{bmatrix}, \tag{14}$$

where p_{ij} is the probability of Markov chain from the state i transit to state j for all $i,j \in \{1,\zeta\}$, while ζ is the number of states in the Markov chain. The exponentially transformed transition probability matrix of the Markov arrival process A^u and A^e can be obtained by

$$T_{a^u}^{\theta} = T_a e^{\theta a_\eta^u} \tag{15}$$

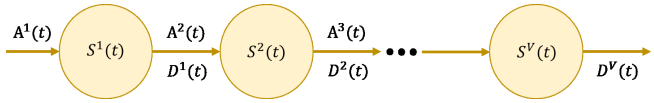


Fig. 5. The V nodes of a multi-hop system.

and

$$T_{a^e}^{\theta} = T_a e^{\theta a_\eta^e}, \tag{16}$$

respectively. The service process of i-th hop the multiplexing system is defined as

$$S^i(x, y) = \sum_{\eta=x}^{y} \bar{s}_\eta^i \tag{17}$$

where $\bar{s}_\eta^i$ is the instant service rate, which can be C_{MISO}, C_{RIS}, and C_{DF} for the MISO, RIS, and DF relay-assisted transmission model, respectively. For the i-th hop service process S^i, the Markov chain transition probability matrix T_{s^i} and the exponentially transformed transition probability matrix $T_{s^i}^{\theta}$ are similarly defined with that of (14), (15) and (16). Furthermore, this study assumes that the arrival and service processes follow the Markov arrival process and have an independent, stationary, and reversible property.

A. Min-plus algebra convolution for queuing network model

A min-plus convolution algebra is used in this study to describe the relationship between the arrival and service processes. The convolution of $f(\cdot)$ and $g(\cdot)$ in the theory of linear time-invariant systems is written as

$$(f \star g)(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau)\, d\tau, \tag{18}$$

where $0 < \tau < t$, $\star$ represents the convolution operation and $f(\cdot)$ and $g(\cdot)$ are measurable functions on $\mathbb{R}^n$. In the queuing network model, the min-plus convolution $\otimes$ of $f(\cdot)$ and $g(\cdot)$ is defined as [53]

$$(f \otimes g)(t) = \inf_{0 \le \tau \le t} \{f(\tau) + g(t-\tau)\}. \tag{19}$$

Unlike (18), the min-plus convolution substitutes the operator *infimum* and *sum* for *sum* and *product*, respectively. Therefore, the arrival, service, and departure processes can be modeled as nonincreasing and nonnegative cumulative functions. Besides, the arrival and service curves in the SNC model are used to describe the system delay and the length of the service queue. $A(\eta)$, $S(\eta)$, and $D(\eta)$ are used to represent the cumulative arrival, service, and departure curves for brevity, respectively. The mathematical expressions of the departure, backlog, and delay processes are derived as follows.

- *Departure process*: The leaving process of the URLLC and eMBB multiplexing system is formed by arrival and service processes as

$$\begin{aligned} D(\eta) \ge (A \otimes S)(\eta) &= \inf_{0 \le \tau \le \eta} \{A(\tau) + S(\eta - \tau)\} \\ &= \inf_{0 \le \tau \le \eta} \{A(\tau) + S(\tau, \eta)\}. \end{aligned} \tag{20}$$

Referring to (20), the number of packets that leave during slot η is equal to or greater than the sum of arriving and served packets during time slot τ and $\eta - \tau$, respectively.

- *Backlog process*: The queue length waiting for service is called the system backlog and can be written as

$$\begin{aligned} B(\eta) &= A(\eta) - D(\eta) \\ &\le \sup_{0 \le \tau \le \eta} \{A(\tau, \eta) - S(\tau, \eta)\}, \end{aligned} \tag{21}$$

where $A(\tau, \eta)$ is the shorthand for $A(\eta) - A(\tau)$. Referring to (20), the upper bound of the backlog can be obtained by replacing $D(\eta)$ with the min-plus convolution form.

- *Delay process*: The delay process $W(\eta)$ is the total time it takes for a unit packet to stay in the system, which is the horizontal distance between $A(\eta)$ and $D(\eta)$ and can be expressed as

$$W(\eta) = \inf \{\tau \ge 0 : A(\eta - \tau) \le D(\eta)\}. \tag{22}$$

B. Multi-hop heterogeneous network

Figure 5 shows the concept of the V nodes of a multi-hop system. $A^i(t)$, $S^i(t)$, and $D^i(t)$ are used to represent the cumulative arrival, service, and departure curves for the ith hop at the tth time slot, respectively [31]. The first and second hop of the service curve for the departure process are defined as

$$\exists t \ge 0, \quad 0 \le \tau \le t, \quad D^1(t) \ge \inf_{0 \le \tau \le t} \left\{A^1(\tau) + S^1(\tau, t)\right\} \tag{23}$$

and

$$\exists l \ge 0, \quad 0 \le t \le l, \quad D^2(l) \ge \inf_{0 \le t \le l} \left\{A^2(t) + S^2(t, l)\right\}, \tag{24}$$

respectively. From Figure 5, the departure process in the first hop is seamlessly connected to the arrival process in the second hop $D^1(t) = A^2(t)$. Referring to (23) and (24), the expression for the two-hop heterogeneous network can be written as

$$\begin{aligned} D^2(l) &\ge \inf_{0 \le t \le l} \left\{A^2(t) + S^2(t, l)\right\} \\ &\ge \inf_{0 \le \tau \le t \le l} \left\{A^1(\tau) + S^1(\tau, t) + S^2(t, l)\right\} \\ &= \left(A^1 \otimes S^1 \otimes S^2\right)(l). \end{aligned} \tag{25}$$

Therefore, the min-plus convolution form of the multi-hop system with V nodes holds for

$$\hat{S}(\eta) = S^1 \otimes S^2 \otimes \cdots \otimes S^V(\eta) \tag{26}$$

$$\le \inf_{\beta_1 + \beta_2 + \ldots + \beta_V = \eta} \left\{S^1(\beta_1) + S^2(\beta_2) + \cdots + S^V(\beta_V)\right\}. \tag{27}$$

Therefore, the backlog of a two-hop system can be obtained by

$$\begin{aligned} B^2(l) &= A^1(l) - D^2(l) \\ &= A^1(l) - \inf_{0 \le \tau \le t \le l} \left\{A^1(\tau) + S^1(\tau, t) + S^2(t, l)\right\} \\ &\le \sup_{0 \le \tau \le t \le l} \left\{A^1(\tau, l) - S^1(\tau, t) - S^2(t, l)\right\}, \end{aligned} \tag{28}$$

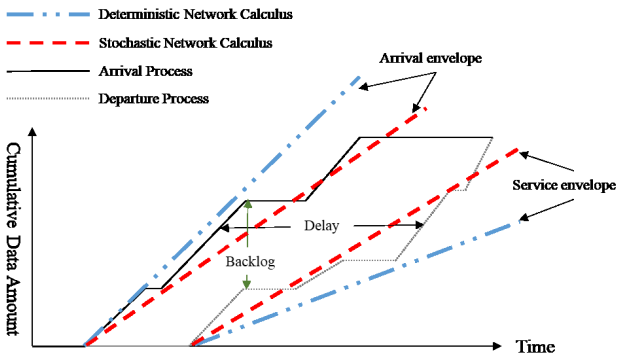


Fig. 6. The concept of the arrival and departure curves in the system.

IV. The Basic of SNC and Martingale Envelopes

Figure 6 illustrates the concept of arrival and departure curves in the system. The horizontal and vertical distances between the cumulative arrival and departure curves are the delay and the backlog, respectively. From the figure, the deterministic network calculus (DNC)-based envelope provides the worst-case curve of the system serviceability, and ensures that all arrival and departure curves do not exceed the boundaries of the DNC. The SNC extends the DNC to the probabilistic domain and leverages the statistical multiplexing gain [54]. DNC is appropriate to describe URLLC traffic, as it has little tolerance for longer delays. However, the DNC does not take into account the efficiency of statistical multiplexing when calculating the amount of resources requirements for a service to run on a network node and result in an overestimation. SNC takes into account the statistical nature of traffic, which is often more realistic in practice [55]. As shown in Fig. 6, the probabilistic bounds are tighter and more reasonable than the DNC-based bounds for describing the arrival and service processes.

A. Stochastic network calculus fundamental

The classical SNC model uses min-plus algebra convolution to obtain the performance boundaries of the system backlog and the delay. The bounded values of the SP can be calculated using the Boole's inequality as [56]

$$P\left(\sup_{\eta} X_\eta \geq \sigma\right) \leq \sum_{\eta} P\left(X_\eta \geq \sigma\right), \tag{29}$$

where X_η represents a stochastic process. The *supremum* of a stochastic process is estimated by the extended tail probability $P\left(\sup_{\eta} X_\eta \geq \sigma\right)$ of the single random variable. If the dependency between each stochastic process is not considered, the SNC model cannot capture the correlation properties of X_η and bring a significant deviation in the tail.

B. Martingale fundamental

The Martingale envelope theory was demonstrated to reduce the derivation in the tail and improve the practicality of the standard SNC model by transforming the MGF into Martingale [32], [34], [42]. The key definitions of the Martingale envelope theory are given below.

Definition 1 (Martingale Process)**.** *Let F_η represent a filtration in the given probability space, where $F_\eta \subset F_{\eta+1}$, $\eta \in \mathbb{N}$. The discrete-time random process $\boldsymbol{X} = \{X_\eta,\ \eta \geq 0\}$ is a discrete-time Martingale when the following conditions are satisfied.*

(i) Integrability condition: $\forall \eta,\ X_\eta$ is F_η-measurable,
(ii) Measurable condition: $E\left[\left|X_\eta\right|\right] \leq \infty$,
(iii) Martingale property: $E\left[X_{\eta+1}|F_\eta\right] = X_\eta$, and $X_\eta = X_0$,

where $E[\cdot]$ represents the expectation operator. The property (iii) can be proved by the tower property of conditional expectation.

Definition 2 (Supermartingale Process)**.** *The discrete-time random process $\boldsymbol{X} = \{X_\eta,\ \eta \geq 0\}$ is a discrete-time supermartingale process when the following conditions are satisfied.*

$$\begin{aligned} &E\left[\left|X_\eta\right|\right] \leq \infty, \\ &E\left[X_{\eta+1}|F_\eta\right] \leq X_\eta, \\ &X_\eta \leq X_0. \end{aligned} \tag{30}$$

The upper bound for the tail probability of delay process or backlog processes can be obtained using the supermartingale.

Definition 3 (Submartingale Process)**.** *The discrete-time random process $\boldsymbol{X} = \{X_\eta,\ \eta \geq 0\}$ is a discrete-time submartingale process when the following conditions are satisfied.*

$$\begin{aligned} &E\left[\left|X_\eta\right|\right] \leq \infty, \\ &E\left[X_{\eta+1}|F_\eta\right] \geq X_\eta, \\ &X_\eta \geq X_0. \end{aligned} \tag{31}$$

C. Envelope of arrival and service processes

The arrival and service processes of a queue system can be bounded by the supermartingale process $G_A(\eta)$ and $G_S(\eta)$, respectively, and are described as follows.

Definition 4 (Arrival Martingale Envelope)**.** *For a monotonically increasing function $h_A(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and every exponential decay factor $\theta > 0$, the arrival process $A(\eta)$ admits a (h_A, θ, K_A)-martingale envelope if*

$$G_A(\eta) := h_A(a_\eta) e^{\theta(A(\eta) - \eta K_A)}, \forall \eta \geq 0 \tag{32}$$

is a supermartingale process. $h_A(\cdot)$ represents the correlation in the stochastic arrival process, and $K_A \geq 0$ is the allocated capacity for the traffic flow $A(\eta)$, respectively.

Definition 5 (Service Martingale Envelope)**.** *For a monotonically increasing function $h_S(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ and every exponential decay factor $\theta > 0$, the service process $S(\eta)$ admits a (h_S, θ, K_S)-martingale envelope if*

$$G_S(\eta) := h_S(s_\eta) e^{\theta(\eta K_S - S(\eta))}, \forall \eta \geq 0 \tag{33}$$

is a supermartingale process. $h_S(\cdot)$ represents the correlation in the stochastic service process, and $K_S > 0$ is the effective system capacity for the service curve $S(\eta)$.

D. Optional stopping theorem

The Martingale analysis model transforms the arrival and service processes into exponential forms of MGF. The complementary cumulative distribution function is estimated using

the Doob's inequality to analyze the violation of the delay and backlog processes.

$$P\left(\sup_{\eta} X_\eta \geq \sigma\right) \leq \frac{E[X_o]}{\sigma}. \tag{34}$$

For the stochastic process, $\mathbf{X} = \{X_\eta,\ \eta \geq 0\}$ is similar to the Markov inequality for a single random variable. The Doob's optional stopping theorem is employed to the supermartingale to properly bound the stochastic process. Applying (21), (32), and (33), the distribution of the backlog of a supermartingale process is obtained as

$$\begin{aligned} B(\eta) &= A(\eta) - D(\eta) \\ &\leq \sup\left\{A(\eta) - \eta K_A + \eta K_S - S(\eta)\right\}. \end{aligned} \tag{35}$$

Therefore, the optional stopping theorem is derived in this study as follows.

Theorem 1 (Optional Stopping Theorem)**.** *For any threshold $\sigma \geq 0$, the stopping time Γ for a supermartingale process is*

$$\Gamma := \inf\left\{\eta \geq 0 : A(\eta) - \eta K_A + \eta K_S - S(\eta) \geq \sigma\right\}. \tag{36}$$

Since it is possible that $\Gamma = \infty$, a new bounded stopping time $\Gamma \wedge \eta := \min\{\Gamma, \eta\}, \forall \eta \geq 0$ is presented.

Proof. Let X_n be a supermartingale with respect to a filtration F_n in the given probability space, and let Γ be the stopping time defined in equation (36). We want to show that Γ is indeed a stopping time, that is, $\{\Gamma \leq n\} \in F_n$ for all n. X_n is a supermartingale means $E(X_{n+1}|F_n) \leq X_n$ for all n.
Let n be any nonnegative integer and let η be any nonnegative real number. We define the event $A(\eta) := \{A(\eta) - \eta K_A + \eta K_S - S(\eta) \geq \sigma\}$, where K_A and K_S are constants, and $S(\eta)$ is a function of η. Then, by the definition of Γ, we have $\Gamma \geq \eta$ if and only if $A(\eta)$ holds. Therefore, we can write

$$\{\Gamma \leq n\} = \bigcup_{k=0}^{n} \{A(k) \cap \{\Gamma = k\}\}. \tag{37}$$

We want to show that each set in this union is in F_n. Let k be any integer between 0 and n, and consider the event $A(k) \cap \{\Gamma = k\}$. Since $\Gamma = k$, we have $A(k-1) \cap \{\Gamma = k\} = \emptyset$, which means that $A(k-1)$ does not hold at time k. Therefore, we can write

$$E(X_{k+1} 1_{A(k)}|F_k) = X_k 1_{A(k)} + E(X_{k+1} 1_{A(k)^c}|F_k). \tag{38}$$

Since X_n is a supermartingale, we have $E(X_{k+1}|F_k) \leq X_k$, which implies $E(X_{k+1} 1_{A(k)^c}|F_k) \leq X_k 1_{A(k)^c}$. Combining this with the previous equation, we get

$$E(X_{k+1} 1_{A(k)}|F_k) \leq X_k. \tag{39}$$

Therefore, X_n is a supermartingale with respect to increased filtration G_n, where $G_n := F_n \vee \sigma(A(k) : k \leq n)$ is the smallest sigma-algebra that contains both F_n and all events $A(k)$ up to time n. Since stopping times are defined with respect to filtrations, we conclude that $\{\Gamma = k\}$ is in $G_k \subseteq F_n$, which implies that $\{\Gamma \leq n\}$ is in F_n for all n. Hence, Γ is indeed a stopping time. Hence, we have shown that the stopping time defined in Theorem 1 is well-defined and satisfies the necessary properties. □

V. Martingale-Based End-to-End Backlog and Delay Analysis

The Martingale theory is applied to analyze the backlog and delay bounds in different scheduling policies, such as SP, nonpreemption, and EDF, for the URLLC and eMBB multiplexing system. The supermartingale envelopes of the URLLC arrival, the eMBB arrival, and the service processes are defined as $G_{A^u}(\eta)$, $G_{A^e}(\eta)$, and $G_S(\eta)$, respectively. K_S^1 and K_S^2 represents the service capability of the first and second AP in the two-hop heterogeneous network, respectively. Following [57], a proportion ξ is introduced to constrain K_S^1 and K_S^2 for reflecting the different service capability of each AP. Therefore, the K_{A^u}, K_{A^2}, K_S^1 and K_S^2 must satisfy $K_{A^u} \geq \frac{\ln sp(T_{a^u}^\theta)}{\theta}$, $K_{A^e} \geq \frac{\ln sp(T_{a^e}^\theta)}{\theta}$, $0 \leq K_S^1 \leq \frac{\ln sp(T_{s^1}^\theta)}{-\theta\xi}$, and $0 \leq K_S^2 \leq \frac{\ln sp(T_{s^2}^\theta)}{-\theta(1-\xi)}$. $sp(T^\theta)$ is the spectral radius of T^θ and reflects the maximum eigenvalue of the transition matrix.

A. Martingale-based backlog analysis

Based on the independent assumption of arrival and service processes, a discrete-time supermartingale $\mathcal{H}(\eta)$ with related to $A^u(\eta)$, $A^e(\eta)$ and $S(\eta)$ can be formed as

$$\mathcal{H}(\eta) = h_{A^u}(a_\eta^u) h_{A^e}(a_\eta^e) h_s(s_\eta) \psi, \tag{40}$$

where $\psi = e^{\theta^*(A^e(\eta) - \eta K_{A^e} + A^u(\eta) - \eta K_{A^u} + \eta K_S - S(\eta))}$. Therefore, the backlog can be formulated as

$$\begin{aligned} &P\Big(B\left(\eta \geq \sigma\right)\Big) \\ &= P\Big(A(\eta) - S(\eta) \geq \sigma\Big) \\ &= P\Big(A^u(\eta) + A^e(\eta) - S(\eta) \geq \sigma\Big) \\ &\leq P\Big(A^u(\eta) + A^e(\eta) - S(\eta) - \eta(K_{A^u} + K_{A^e} - K_S) \geq \sigma\Big). \end{aligned} \tag{41}$$

Applying the optional stopping Theorem 1 to the supermartingale $\mathcal{H}(\eta)$, $\forall \eta \in \mathbb{N}$, we have

$$\begin{aligned} &E[h_{A^u}(a_0^u) h_{A^e}(a_0^e) h_S(s_0)] \\ &= E[\mathcal{H}(0)] \\ &= E[\mathcal{H}(\Gamma \wedge \eta)] \geq E[\mathcal{H}(\Gamma \wedge \eta) 1_{\{\Gamma < \eta\}}] \\ &= E\Big[h_{A^u}(a_0^u) h_{A^e}(a_0^e) h_S(s_0) \\ &\qquad e^{\theta^*(A^e(\Gamma) - \Gamma K_{A^e} + A^u(\Gamma) - \Gamma K_{A^u} + \Gamma K_S - S(\Gamma))} 1_{(\Gamma < \eta)}\Big] \\ &\geq H e^{\theta^* \sigma} P(\Gamma < \eta). \end{aligned} \tag{42}$$

To put it briefly, we can define ρ as the product of the expected values of $h_{A^u}(a_0^u)$, $h_{A^e}(a_0^e)$, and $h_S(s_0)$, respectively, denoted by $E\left[h_{A^u}(a_0^u)\right]$, $E\left[h_{A^e}(a_0^e)\right]$, and $E\left[h_S(s_0)\right]$. Therefore, we have $\rho = E\left[h_{A^u}(a_0^u)\right] E\left[h_{A^e}(a_0^e)\right] E\left[h_S(s_0)\right]$. Applying Theorem 1 and (35) to (42), for $\eta \to \infty$, the distribution of the backlog violation of a single-hop system can be obtained as

$$P\Big(B(\eta) \geq \sigma\Big) = P(\Gamma < \infty) \leq \frac{\rho}{H e^{\theta^* \sigma}}, \tag{43}$$

where H and θ^* in (43) are given by

$$H = \min\{h_{A^u}(x) h_{A^e}(y) h_S(z) : x + y - z > 0\} \tag{44}$$

and

$$\theta^* = \sup\{\theta > 0 : K_{A^u} + K_{A^e} \leq K_S\}, \tag{45}$$

respectively. Furthermore, the distribution of the backlog violation of a multi-hop system with V service nodes can be obtained as

$$P\Big(B^2(l) \geq \sigma\Big) \leq \frac{E\left[h_{A^u}(a_0^u)\right] E\left[h_{A^e}(a_0^e)\right] \prod_{i=1}^{V} E\left[h_{S^i}(s_0)\right]}{H' e^{\hat{\theta}^* \sigma}}, \tag{46}$$

where H' and $\hat{\theta}^*$ in (46) are given by

$$H' = \min\{h_{A^u}(x) h_{A^e}(y) \prod_{i=1}^{V} h_{S^i}(z^i) : x + y > z^1 > \cdots > z^V\} \tag{47}$$

and

$$\hat{\theta}^* = \sup\left\{\theta > 0 : K_{A^u} + K_{A^e} \leq \min\{K_S^1, K_S^2, \ldots, K_S^V\}\right\}, \tag{48}$$

respectively. $K_{A^u} + K_{A^e}$ should be less than or equal to any value of K_S^1 and K_S^2 to guarantee the stability condition in a multi-hop heterogeneous network [57]. H' holds the smallest value of $h_{A^u}(x) h_{A^e}(y) \prod_{i=1}^{V} h_{S^i}(z^i)$ because the instantaneous arrival must larger than any value of the stochastic process $\{z^i\}_{i=1}^{V}$ to drive the service process of the first hop. Furthermore, each instantaneous value z^i must larger than any value of the following stochastic process $\{z^i\}_{i+1}^{V}$ (i.e., $z^i > z^{i+1}$) to drive the next service process.

B. Martingale-based delay analysis for SP

In the SP scheduling policy, the eMBB packet is interrupted by the arriving URLLC packet and will wait for the spectrum resources until all URLLC traffics are fully served. Therefore, the remaining service processes for URLLC and eMBB traffics are defined as

$$S^u(x, y) = [S(x, y) - A^e(x, y)]^+ \tag{49}$$

and

$$S^e(x, y) = [S(x, y) - A^u(x, y)]^+, \tag{50}$$

respectively. $[x]^+$ denotes the operation to obtain the positive part of x. According to the definition of the delay process (22), the URLLC and eMBB service delay for the SP scheduling policy are written as

$$\begin{aligned} &P\Big(W^u(\eta) \geq \kappa\Big) \\ &= P\Big(A^u(\kappa, \eta) - S^u(\eta) \geq 0\Big) \\ &= P\Big(A^u(\kappa, \eta) + A^e(\eta) - S(\eta) \geq 0\Big) \\ &\leq P\Big(\sup_{0 \leq \kappa \leq \eta} \Big\{A^u(\kappa, \eta) + A^e(\eta) - S(\eta) \\ &\quad - (\eta - \kappa) K_{A^u} - \eta K_{A^e} + \eta K_S\Big\} \geq \kappa(K_S - K_{A^e})\Big) \end{aligned} \tag{51}$$

and

$$\begin{aligned} &P\Big(W^e(\eta) \geq \kappa\Big) \\ &= P\Big(A^e(\kappa, \eta) - S^e(\eta) \geq 0\Big) \\ &\leq P\Big(\sup_{0 \leq \kappa \leq \eta} \Big\{A^e(\kappa, \eta) + A^u(\eta) - S(\eta) \\ &\quad - (\eta - \kappa) K_{A^e} - \eta K_{A^u} + \eta K_S\Big\} \geq \kappa(K_S - K_{A^u})\Big), \end{aligned} \tag{52}$$

respectively. According to the exponential transforms of MGF and the optional stopping Theorem 1, the delay distributions of URLLC and eMBB for the SP scheduling policy are defined as

$$P\left(W^u(\eta) \geq \kappa\right) \leq \frac{\rho}{H} e^{-\theta^*(\kappa K_S - \kappa K_{A^e})}, \tag{53}$$

and

$$P\left(W^e(\eta) \geq \kappa\right) \leq \frac{\rho}{H} e^{-\theta^*(\kappa K_S - \kappa K_{A^u})}, \tag{54}$$

respectively. Furthermore, the delay distributions of URLLC and eMBB traffics for the two-hop system can be derived by

$$P\left(W^u(l) \geq \kappa\right) \leq \frac{\varpi}{H'} e^{-\hat{\theta}^*(\kappa \xi K_S^1 + \kappa(1-\xi) K_S^2 - \kappa K_{A^e})}, \tag{55}$$

and

$$P\left(W^e(l) \geq \kappa\right) \leq \frac{\varpi}{H'} e^{-\hat{\theta}^*(\kappa \xi K_S^1 + \kappa(1-\xi) K_S^2 - \kappa K_{A^u})}, \tag{56}$$

respectively, where ϖ is given by

$$\varpi = E\left[h_{A^u}(a_0^u)\right] E\left[h_{A^e}(a_0^e)\right] E\left[h_{S^1}(s_0)\right] E\left[h_{S^2}(s_0)\right]. \tag{57}$$

C. Martingale-based delay analysis for nonpreemption

The nonpreemptive scheduling follows the first-in-first-out policy, which states that all packets in the queue system have the same priority. Therefore, the service processes of URLLC and eMBB in the nonpreemptive scheduling policy can be expressed as

$$S^u(x, y) = [S(x, y) - A^e(x, y - z)]^+ I_{\{y - x > z\}} \tag{58}$$

and

$$S^e(x, y) = [S(x, y) - A^u(x, y - z)]^+ I_{\{y - x > z\}}, \tag{59}$$

respectively. z indicates that a traffic flow stays in the queue system from y to $y + z$. Without generalization loss, z is set as κ for convenience. $[x]^+$ represents the operation to obtain the positive part of x. I_E is the indicator function of condition E. Therefore, the delay of the URLLC and eMBB processes can be defined as

$$\begin{aligned} P\Big(W^u(\eta) \geq \kappa\Big) &= P\Big(A^u(\kappa, \eta) - S^u(\eta) \geq 0\Big) \\ &= P\Big(A^u(\kappa, \eta) + A^e(\kappa, \eta) - S(\eta) \geq 0\Big) \\ &\leq P\Big(\sup_{0 \leq \kappa \leq \eta} \Big\{A^u(\kappa, \eta) + A^e(\kappa, \eta) - S(\eta) \\ &\quad - (\eta - \kappa)(K_{A^u} + K_{A^e}) + \eta K_S\Big\} \geq \kappa K_S\Big) \end{aligned} \tag{60}$$

and

$$P\Big(W^e(\eta) \geq \kappa\Big) = P\Big(A^e(\kappa,\eta) - S^e(\eta) \geq 0\Big) \leq P\Big(\sup_{0\leq\kappa\leq\eta}\Big\{A^u(\kappa,\eta) + A^e(\kappa,\eta) - S(\eta) - (\eta-\kappa)(K_{A^u}+K_{A^e}) + \eta K_S\Big\} \geq \kappa K_S\Big), \tag{61}$$

respectively. The delay distribution for URLLC and eMBB traffic in the single-hop system can be obtained by applying (42) and is given by

$$P\left(W^u(\eta) \geq \kappa\right) \leq \frac{\rho}{H} e^{-\theta^*\kappa K_S}, \tag{62}$$

and

$$P\left(W^e(\eta) \geq \kappa\right) \leq \frac{\rho}{H} e^{-\theta^*\kappa K_S}, \tag{63}$$

respectively. Furthermore, the delay distribution for URLLC and eMBB traffic in the two-hop system can be obtained by

$$P\left(W^u(l) \geq \kappa\right) \leq \frac{\varpi}{H'} e^{-\hat{\theta}^*(\kappa\xi K_S^1 + \kappa(1-\xi)K_S^2)}, \tag{64}$$

and

$$P\left(W^e(l) \geq \kappa\right) \leq \frac{\varpi}{H'} e^{-\hat{\theta}^*(\kappa\xi K_S^1 + \kappa(1-\xi)K_S^2)}, \tag{65}$$

respectively.

D. Martingale-based delay analysis for EDF

The waiting time for the URLLC packet a_η^u and the eMBB packet a_η^e in the queuing system are defined as d_η^u and d_η^e, respectively. The priority of a packet in the EDF scheduling policy depends on its remaining deadline. To guarantee the critical low-latency requirement of URLLC transmission, a relative deadline threshold μ for each URLLC packet is designed in this study. Specifically, a URLLC packet a_η^u has a higher priority than an eMBB packet a_η^e when $d_\eta^u - d_\eta^e > \mu$. The eMBB packet a_η^e has a high priority when $d_\eta^u - d_\eta^e < \mu$. Otherwise, URLLC and eMBB packets have the same priority. Therefore, the bivariate random service processes of URLLC and eMBB in the EDF scheduling are given by

$$S^u(x,y) = [S(x,y) - A^e(x, y-z+\min(z,\delta)]^+ I_{\{y-x>z\}} \tag{66}$$

and

$$S^e(x,y) = [S(x,y) - A^u(x, y-z+\min(z,\delta)]^+ I_{\{y-x>z\}}, \tag{67}$$

respectively. $\delta = d_\eta^u - d_\eta^e$ is the difference between the URLLC and eMBB packet waiting time. Because the URLLC and eMBB packets have the same priority when $\delta - \mu = 0$, the delay distribution can be obtained using (62) and (63) in the nonpreemtive scheduling policy. $[x]^+$ is the operation to obtain the positive part of x. I_E represents the indicator function of condition E. By transforming the MGF into exponential form, the delay of the eMBB service process is discussed for two cases, i.e., $\delta - \mu > 0$ and $\delta - \mu < 0$, respectively.

For the case $\delta - \mu > 0$, the delay distribution of the single-hop system can be obtained by

$$P(W^e(\eta) \geq \kappa) \leq \frac{\rho}{H} e^{-\theta^*(\kappa K_S - \min(\kappa,\delta)K_{A^u})}. \tag{68}$$

TABLE II
SIMULATION PARAMETERS

Environment Parameters	Default Value
Bandwidth at the first hop	900 KHz
Bandwidth at the second hop	720 KHz
Carrier frequency	3 GHz
Packet size of URLLC	32 bytes
Packet size of eMBB	1600 bytes
The number of URLLC processes	12
The number of eMBB processes	6
Ratio of URLLC to eMBB arrival rate	10
The number of RIS reflection elements	[25, 100]
AWGN	-94 dBm
Noise figure N_f	10 dB
Transmit power at AP p_b	10 dBm
AP antenna gain	5 dBi
RIS antenna gain	5 dBi
DF relay antenna gain	5 dBi
Receiver antenna gain	0 dBi
Distance between AP and RIS	50 ~ 150 m
Distance between AP and each receiver	50 ~ 150 m
Distance between RIS and each receiver	50 ~ 150 m
Deadline threshold μ in EDF	25 ms
Transmit power for URLLC	20 dBm
Transmit power for eMBB	23 dBm

The delay distribution of the two-hop system can be derived by

$$P(W^e(l) \geq \kappa) \leq \frac{\varpi}{H'} e^{-\hat{\theta}^*(\kappa\xi K_S^1 + \kappa(1-\xi)K_S^2 - \min(\kappa,\delta)K_{A^u})}. \tag{69}$$

For the case $\delta - \mu < 0$, the delay distribution can be obtained by

$$P(W^u(\eta) \geq \kappa) \leq \frac{\rho'}{H_1} e^{-\theta_1^*\kappa K_S} + \frac{\rho}{H_2} e^{-\theta_2^*(\kappa K_S - \delta K_{A^e})}, \tag{70}$$

where H_1, θ_1^*, H_2, θ_2^*, and ρ' are given by

$$H_1 = \min\left\{h_{A^u}(x)h_S(z) : x - z > 0\right\}, \tag{71}$$

$$\theta_1^* \leq \sup\{\theta > 0 : K_{A^u} \leq K_S\}, \tag{72}$$

$$H_2 = \min\left\{h_{A^u}(x)h_{A^e}(y)h_S(z) : x + y - z > 0\right\}, \tag{73}$$

$$\theta_2^* \leq \sup\{\theta > 0 : K_{A^u} + K_{A^e} \leq K_S\}, \tag{74}$$

and

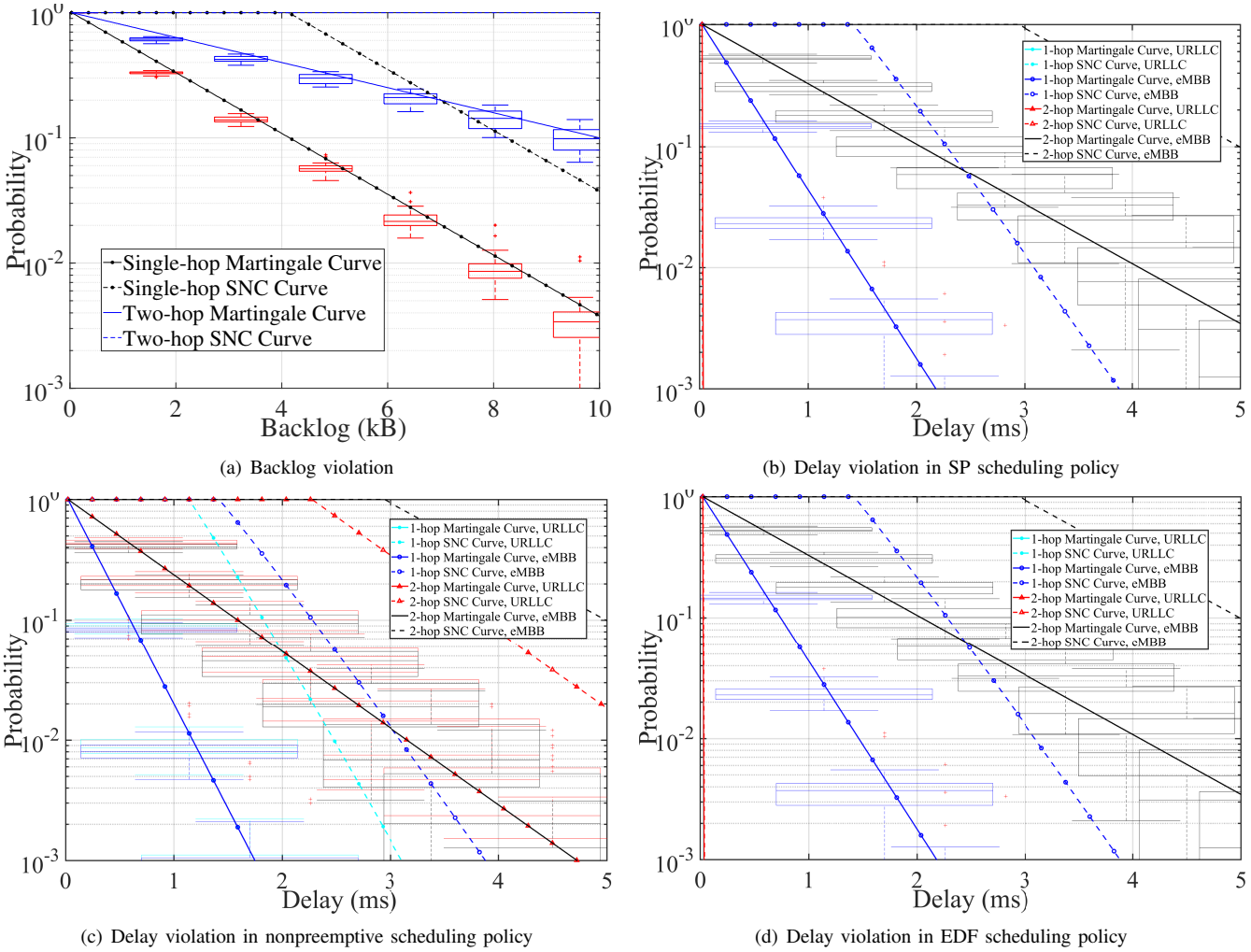
$$\rho' = E[h_{A^u}(a_0^u)]E[h_S(s_0)], \tag{75}$$

respectively. The delay distribution of the two-hop system for the case $\delta - \mu < 0$ can be obtained by

$$P(W^u(l) \geq \kappa) \leq \frac{\varpi'}{H_1'} e^{-\hat{\theta}_1^*(\kappa\xi K_S^1 + \kappa(1-\xi)K_S^2)} + \frac{\varpi}{H_2'} e^{-\hat{\theta}_2^*(\kappa\xi K_S^1 + \kappa(1-\xi)K_S^2 - \delta K_{A^e})}, \tag{76}$$

where H_1', $\hat{\theta}_1^*$, H_2', $\hat{\theta}_2^*$, and ϖ' are given by

$$H_1' = \min\{h_{A^u}(x)h_{S^1}(z^1)h_{S^2}(z^2) : x > z^1 > z^2\}, \tag{77}$$



(a) Backlog violation

(b) Delay violation in SP scheduling policy

(c) Delay violation in nonpreemptive scheduling policy

(d) Delay violation in EDF scheduling policy

Fig. 7. The backlog and delay analysis of the MISO system for multiplexing URLLC and eMBB.

$$\hat{\theta}_1^* \leq \sup\left\{\theta > 0 : K_{A^u} \leq \min\{K_S^1, K_S^2\}\right\}, \tag{78}$$

$$H_2' = \min\{h_{A^u}(x)h_{A^e}(y)h_{S^1}(z^1)h_{S^2}(z^2) : x + y > z^1 > z^2\}, \tag{79}$$

$$\hat{\theta}_2^* \leq \sup\left\{\theta > 0 : K_{A^u} + K_{A^e} \leq \min\{K_S^1, K_S^2\}\right\}, \tag{80}$$

and

$$\varpi' = E[h_{A^u}(a_0^u)]E[h_{S^1}(s_0)]E[h_{S^2}(s_0)] \tag{81}$$

respectively.

VI. Numerical Results

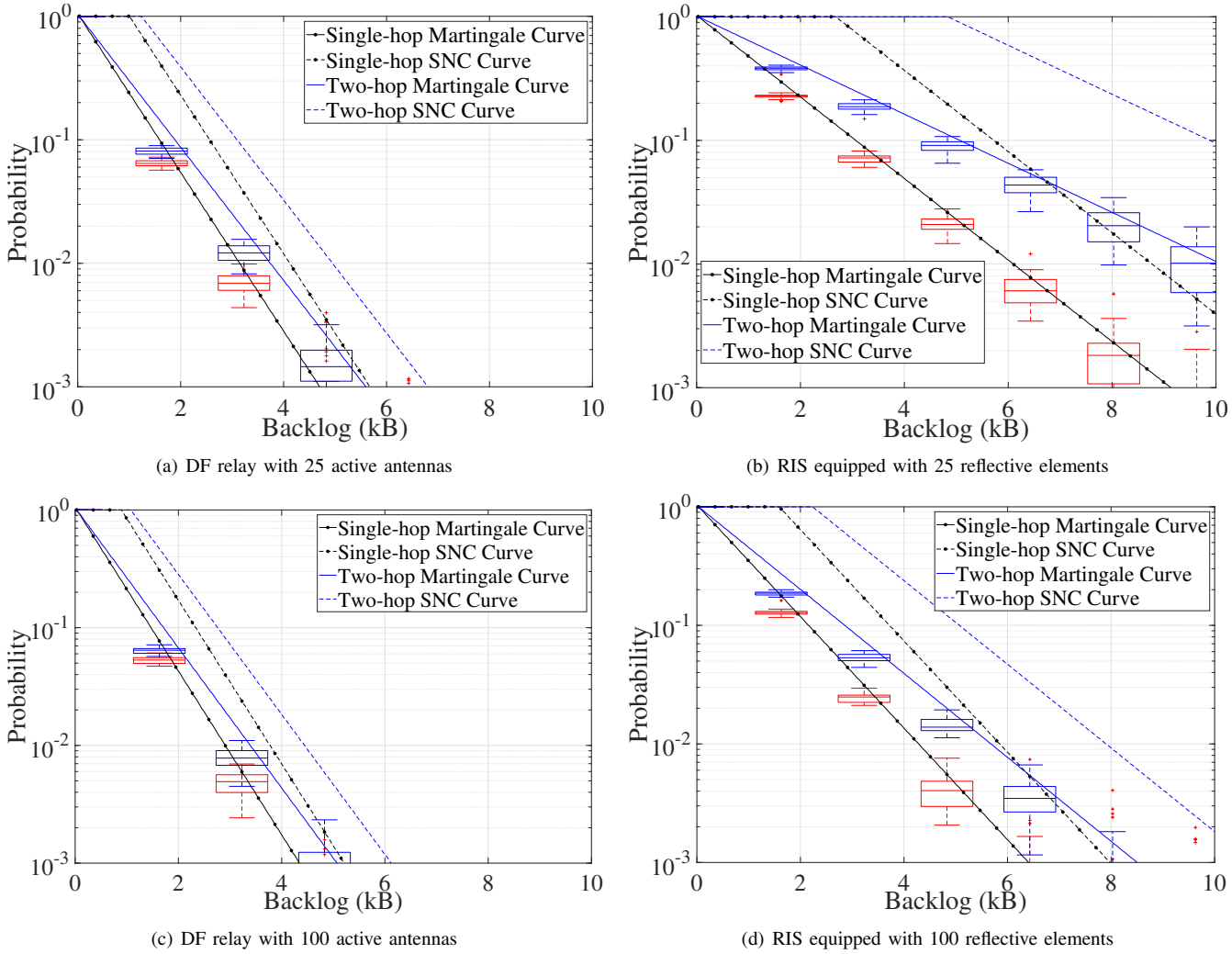
Table II presents the simulation setting of the partial parameters [15], [58]–[60]. Following [59], [60], the transmit power budget for URLLC and eMBB receiver is 20 dBm and 23 dBm, respectively. Additionally, this study analyzed the performance of the URLLC and eMBB multiplexing system with respect to different transmit power budgets ranging from 20 dBm to 33 dBm. Following [15], the AWGN (in dBm) at the receiver is given by

$$N_r = -174 + 10 * log_{10}(B) + N_f, \tag{82}$$

where -174dBm/Hz is the noise density, B is the bandwidth, and $N_f = 10$ is the noise figure. According to 3GPP, a resource block (RB) is defined as 12 consecutive subcarriers in the frequency domain and each subcarrier spacing is 15 KHz [58]. Therefore, the channel bandwidth of each RB is 180 KHz [61]. Assuming 12 URLLC and 6 eMBB users, we set channel bandwidths of 900 KHz and 720 KHz for the first and second hops, respectively [62]. Without generality loss, the DF relay was deployed in the same position as the RIS. In this study, the serviceability of the RIS with respect to different metasurface sizes is explored by setting the number of reflection elements at 25 and 100. According to [15], [63], this study assumes that both BS and RIS use ULA for their antenna distribution. The reason for considering an ULA for the RIS is that it provides a simple and efficient way to control the phase shift of the reflected signal. By adjusting the phase shift of each antenna element in the ULA, the reflected signal can be directed in a specific direction and the signal strength can be optimized at the receiver. Additionally, the use of an ULA for the RIS simplifies the analysis and modeling of the system, allowing this study to focus on estimating the delay and backlog of this multiplexing system. Following [15], the channel gain $\mathcal{G}$ related to the distance is defined as

$$\mathcal{G} = \gamma_t + \gamma_r + \begin{cases} -37.5 - 22\log_{10}(\mathcal{D}/1m) & \text{if LoS,} \\ -35.1 - 36.7\log_{10}(\mathcal{D}/1m) & \text{if NLoS.} \end{cases} \tag{83}$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3313964



(a) DF relay with 25 active antennas

(b) RIS equipped with 25 reflective elements

(c) DF relay with 100 active antennas

(d) RIS equipped with 100 reflective elements

Fig. 8. The comparison of backlog violation analysis between DF relay and RIS for multiplexing URLLC and eMBB.

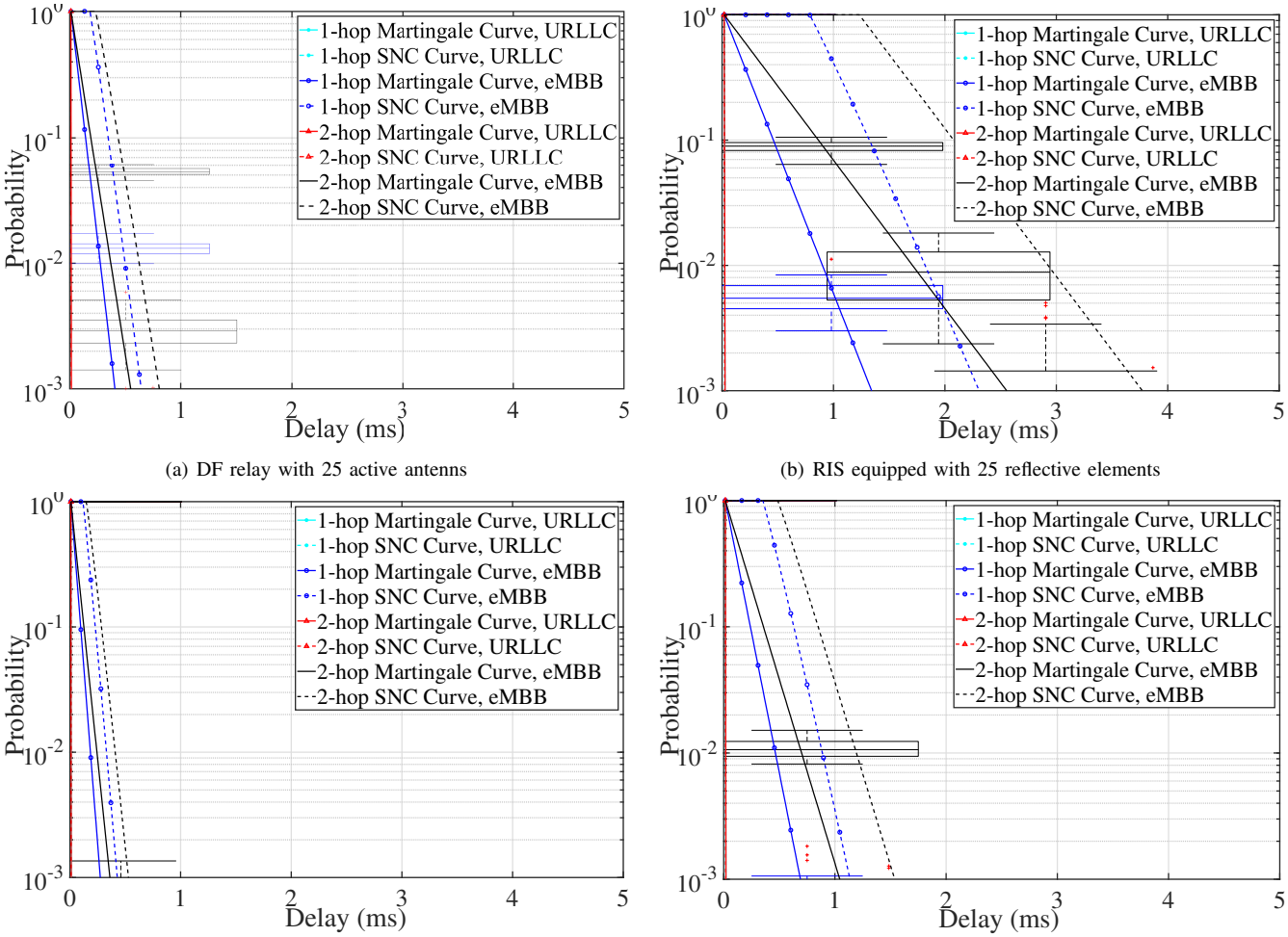
Here γ_t and γ_r are the antenna gains of the transmitter and the receiver, respectively. $\mathcal{D}$ is the Euclidean distance between the transmitter and the receiver. The Poisson arrival rates for URLLC and eMBB are $\lambda_u = 10$ and $\lambda_e = 1$, respectively. The packet sizes of URLLC and eMBB are 32 bytes and 1600 bytes, respectively. The deadline threshold in the EDF scheduling policy is set at 25 ms [64]. This means that a URLLC packet in the queue has a higher priority than an eMBB packet when its arrival time minus the eMBB arrival time is less than 25 ms. The proposed Martingale-based SNC model was used to evaluate the serviceability of MISO, RIS, and DF relay-assisted URLLC and eMBB multiplexing systems. This section discusses the backlog and delay violation probabilities of the URLLC and eMBB multiplexing system in terms of various scheduling policies, such as SP, nonpreemption, and EDF. Furthermore, in each case a single-hop system and a two-hop heterogeneous system were experimentally studied.

A. Performance evaluation between Martingale and SNC

Figures 7 show the backlog and delay violations of MISO-assisted URLLC and eMBB multiplexing. The distance between the AP and the receivers is set to 150 meters. The transmit power for the URLLC and eMBB receivers is 20 and 23 dBm, respectively [59], [60]. It is observed that the SNC curves have significant gaps with the simulation results, whereas the Martingale curves regress the simulation results accurately and tightly. Due to their high priority and low latency requirements, URLLC traffic can be immediately served in SP and EDF scheduling policies. However, in a two-hop network, delays and backlog violations are greater due to the presence of a low serviceability node, which acts as a bottleneck in the heterogeneous transmission system.

Figures 8 illustrate the backlog violation behavior of the URLLC and eMBB multiplexing system in terms of various communication system models. It is assumed that the DF relay has the same power budget as the AP. The red and blue box plots represent the simulation results for the single-hop network and the heterogeneous two-hop network, respectively. Solid and dot lines are the boundaries derived from the proposed Martingale and SNC models, respectively. From these figures, it can be observed that the Martingale curve is tightly closed to the curve of the simulation result, whereas the SNC curve is loose and has a gap from the simulation result. Furthermore, the probability of backlog violation reached the

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3313964



(a) DF relay with 25 active antenns

(b) RIS equipped with 25 reflective elements

(c) DF relay with 100 active antennas

(d) RIS equipped with 100 reflective elements

Fig. 9. The delay violation analysis with the SP scheduling policy.

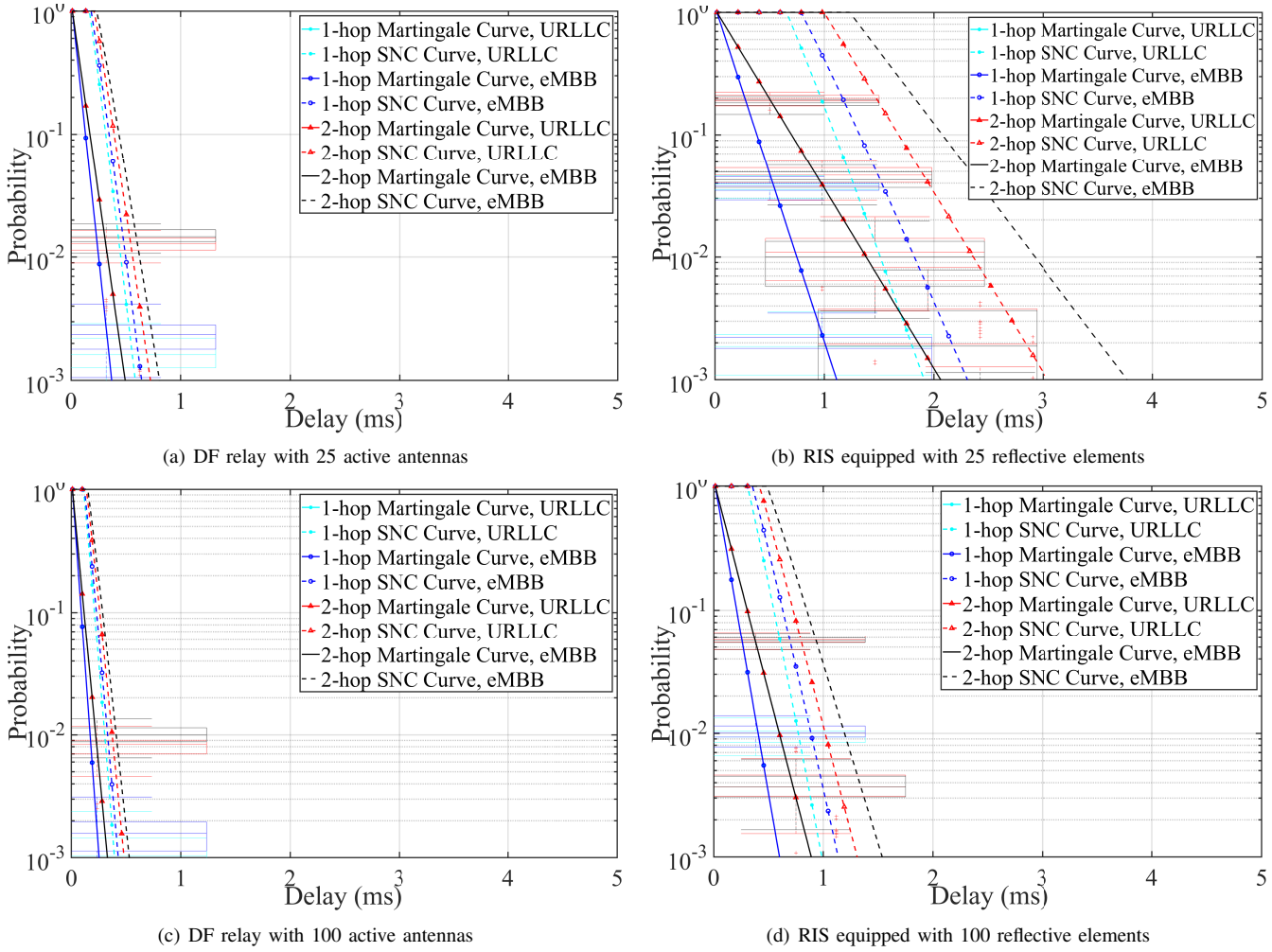
lowest value when the DF relay was equipped with 100 reflective elements, while the RIS with 25 reflective elements suffered from the highest probability of backlog violation. It should be noted that the DF relay outperformed RIS despite having the same number of passive antennas. This is because the DF relay is equipped with active antennas, which can provide additional power to enhance signal strength, while the passive reflective elements in the RIS only change the transmit direction of the signals. However, the DF relay consumes energy while the RIS can trade off energy consumption and transmission performance. As shown in Figures 7(a) and 8, the MISO system suffers the highest probability of a backlog violation. Consequently, the backlog violation behavior of the URLLC and eMBB multiplexing system can be reduced by both the relay and the RIS.

Figures 9 depict and evaluate the delay violation behavior with the SP scheduling policy. The box plot with the colors cyan, red, blue, and black represents the simulation results for the URLLC traffic of a single-hop system, the URLLC traffic of a two-hop system, the eMBB traffic of a single-hop system, and the eMBB traffic of a two-hop system, respectively. From these figures, the simulation results and probability bounds' curves with respect to URLLC traffics in all cases are close to zero. The reason is that URLLC's packet size is much smaller than that of eMBB and has high priority. As a result, each arriving URLLC packet is immediately served by the multiplexing system. Furthermore, the delay violation behavior of the eMBB traffic can be seen in all transmission models and is particularly high in the RIS system with 25 reflective elements. This is because eMBB packets have a low priority and massive data that cannot be fully served by URLLC and eMBB multiplexing systems.

Figures 10 shows the analysis of the delay violation behavior for the nonpreemption scheduling, where URLLC packets have the same priority as eMBB packets. Therefore, the delay violation probability for URLLC traffic is the same as that of eMBB traffic. The probability of delay is extremely high for all transmissions due to the limitation of serviceability when the RIS is only equipped with 25 reflective elements. Furthermore, the RIS equipped with 100 reflective elements significantly reduces the probability of delay violations in the nonpreemptive scheduling policy. It can be observed that DF relay can immediately serve all URLLC and eMBB packets in this scheduling policy.

Figures 11 show the probability of delays in the EDF scheduling policy. It can be seen that URLLC traffic can

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3313964



(a) DF relay with 25 active antennas

(b) RIS equipped with 25 reflective elements

(c) DF relay with 100 active antennas

(d) RIS equipped with 100 reflective elements

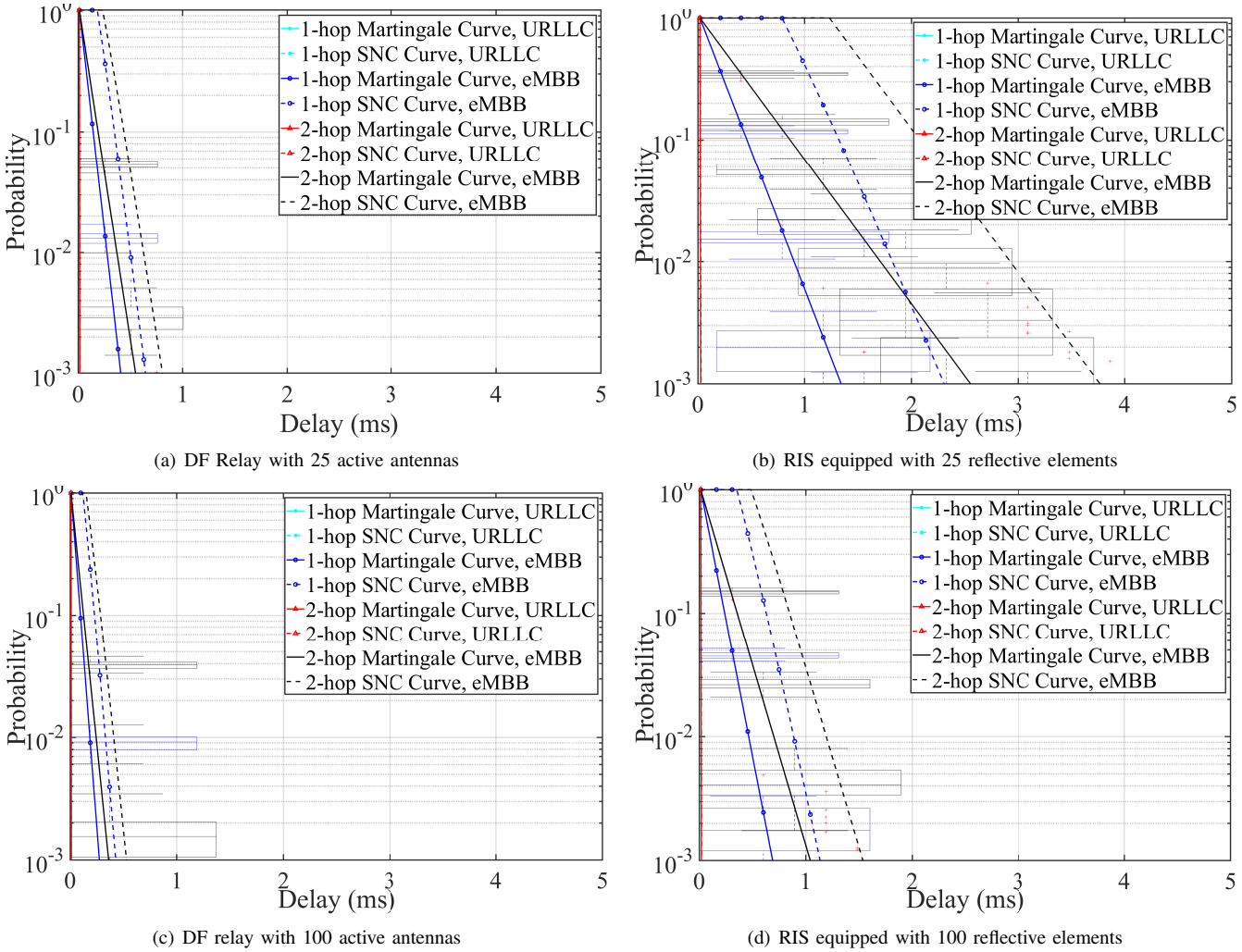
Fig. 10. The delay violation analysis with the non-preemption scheduling policy.

be fully served in both the DF relay and RIS systems. Furthermore, eMBB delay violation behavior occurs in two-hop systems. This is because the system capacity on the second hop is assumed to be less than that of the first hop and cannot immediately serve all arriving eMBB packets. The proposed Martingale-based analysis model can provide extremely tight bounds in terms of the eMBB delay probability curves, while the SNC model suffers from loose bounds.

The numerical results show that the serviceability of the RIS system increased with increasing number of reflective elements. The DF relay equipped with 100 reflective elements achieved the best performance in reducing backlog and delay violations compared to the RIS equipped with 100 reflective elements. Furthermore, the DF relay equipped with 25 reflective elements outperformed the RIS equipped with 100 reflective elements in terms of backlog and delay reductions. Due to the transmit power required on the relay in the forwarding phase, the DF relay must be energy-consuming. Furthermore, the proposed Martingale-based analysis model outperformed the SNC model in providing accurate bounds for backlog and delay violations. The results show that the classical SNC model is extremely loose, such as the SNC curve in Figure 9(b), which overestimated the delay of the eMBB packet by 99% at 1 ms. This is because the SNC model derived the upper/lower bounds by transforming the MGF to the Chernoff bound, regarding each stochastic process as a separate point, whereas Martingale adopted Doob's optional sampling theorem and considered the correlation between each stochastic process.

In summary, the low-latency requirement of the URLLC traffic can be met by the SP scheduling policy in all three communication system models, whereas the eMBB traffic suffers a high probability of delay violations. The eMBB transmission reached the lowest probability of delay violation in the non-preemption scheduling policy, while it is difficult to satisfy the low-latency requirement of URLLC traffic. Although URLLC packets suffer from delay violations in the MISO system, the EDF scheduling policy trades off the latency behavior between URLLC and eMBB transmission. Furthermore, the DF relay and the RIS can reduce delay violations by improving the serviceability of URLLC and eMBB multiplexing systems. Lastly, the two-hop heterogeneous network suffers from more delay violations than the single-hop network. This is because the serviceability of the second hop is less than that of the first hop and cannot immediately serve all arrival packets.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3313964



(a) DF Relay with 25 active antennas

(b) RIS equipped with 25 reflective elements

(c) DF relay with 100 active antennas

(d) RIS equipped with 100 reflective elements

Fig. 11. The delay violation analysis with the EDF scheduling policy.

B. Comparison of different scheduling policies, transmit power, and receiver distributions

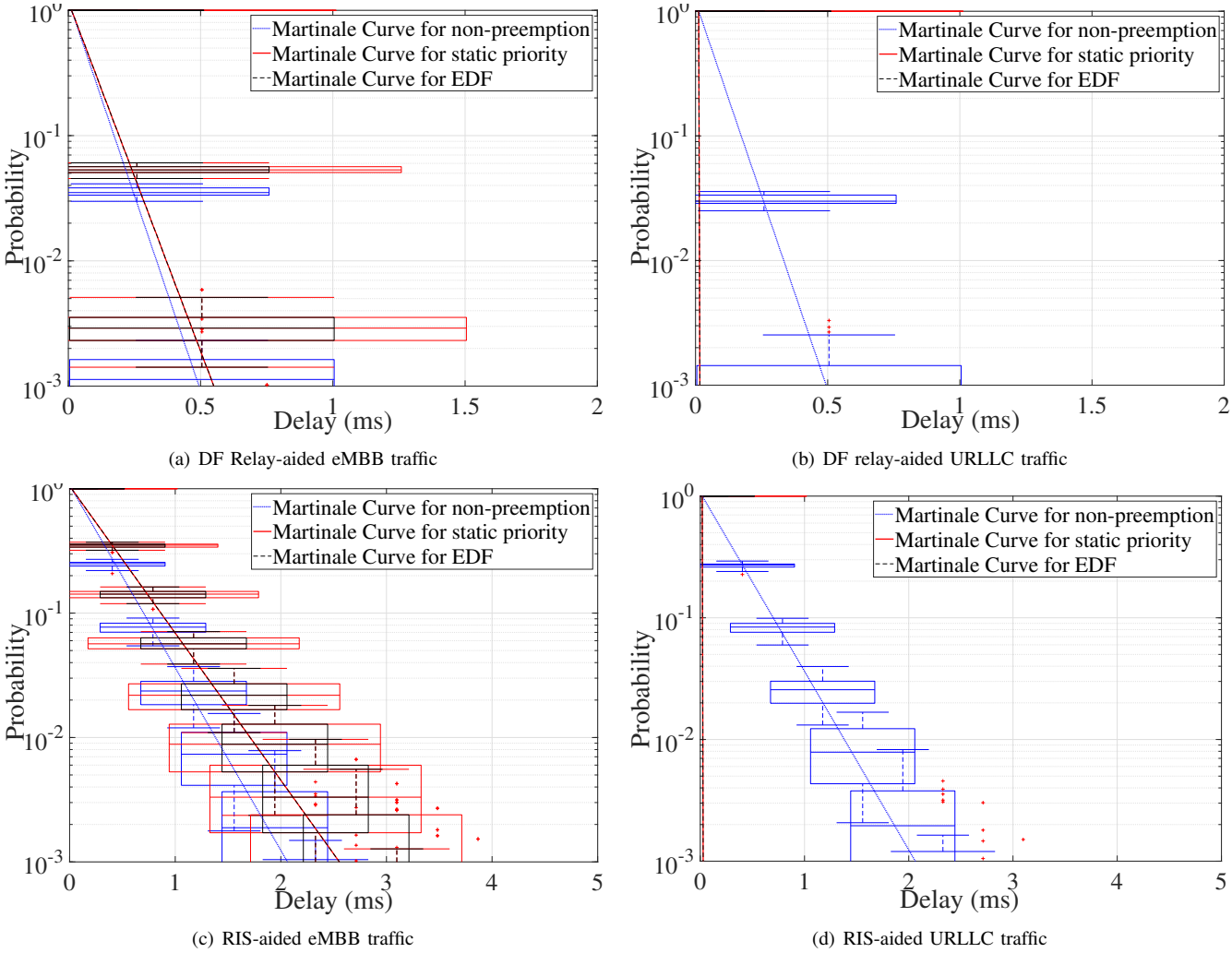
Figures 12 illustrate the delay violations for two-hop URLLC and eMBB multiplexing with different scheduling policies. The DF relay and RIS are equipped with 25 active antennas and 25 passive reflective elements, respectively. URLLC packets can be fully served by the relay and RIS systems in the EDF and SP scheduling policies, while URLLC suffers from delays in the nonpreemptive scheduling policy. In the nonpreemptive scheduling policy, eMBB services experience the lowest delay, while in the SP scheduling policy, eMBB traffic experiences the highest delay. This is because eMBB traffic has the same priority as URLLC services in the nonpreemptive scheduling policy, but has the lowest priority in the SP scheduling policy.

Figures 13 depict the backlog violation in different receiver distributions. It can be observed that the volume of backlog increases as distance increases in RIS-assisted systems, while there is only a slight increase in DF relay-assisted systems. This is because the DF relay strengthens the signals by using additional power, which can improve service coverage. Furthermore, increasing the number of passive elements can extend service coverage and reduce backlog violations. Figures 14 show the backlog violation with respect to different transmit power budgets. It can be seen that the backlog violation decreases as the transmit power and the number of antennas increase. The performance of RIS is more sensitive than that of the DF relay because the RIS passively reflects the incoming signals, while the DF relay can actively increase the signal strength.

VII. Conclusion

This study comprehensively investigated the serviceability of URLLC and eMBB multiplexing systems supported by MISO, RIS, and DF relay, as well as the single-hop homogeneous and the two-hop heterogeneous communication networks. The backlog and delay violation behaviors for URLLC and eMBB multiplexing were accurately analyzed by applying Martingale theory to the SNC model. Furthermore, this study discussed the backlog and delay violations distribution in terms of different scheduling policies, such as SP, nonpreemption, and EDF. The numerical results demonstrated that the RIS and DF relay significantly improve the serviceability of the URLLC and eMBB multiplexing system. The DF

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2023.3313964



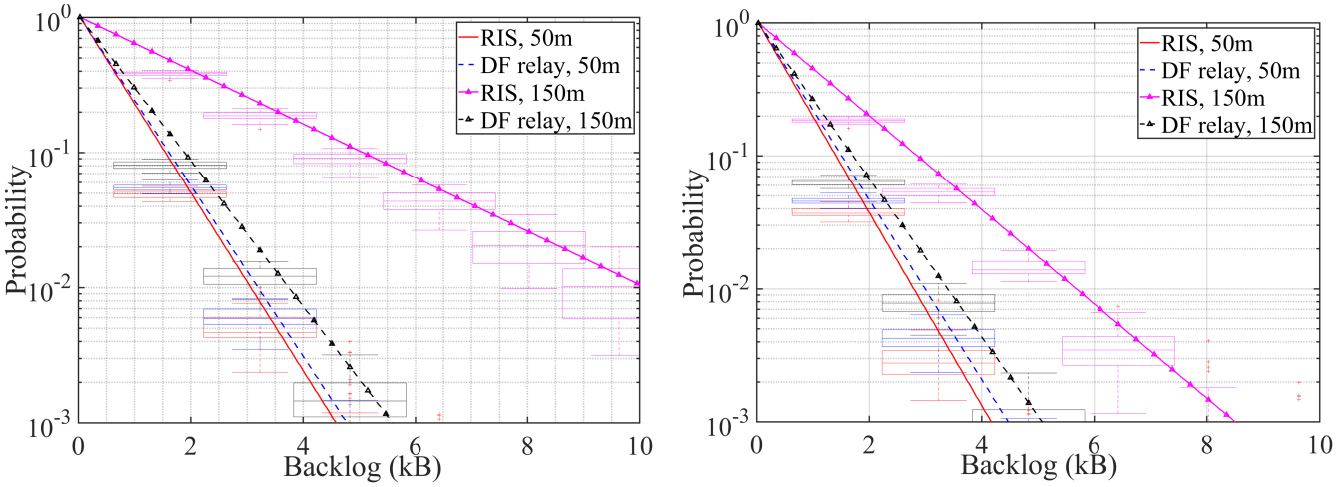
(a) DF Relay-aided eMBB traffic

(b) DF relay-aided URLLC traffic

(c) RIS-aided eMBB traffic

(d) RIS-aided URLLC traffic

Fig. 12. The delay violation analysis with different scheduling policies in the two-hop system.

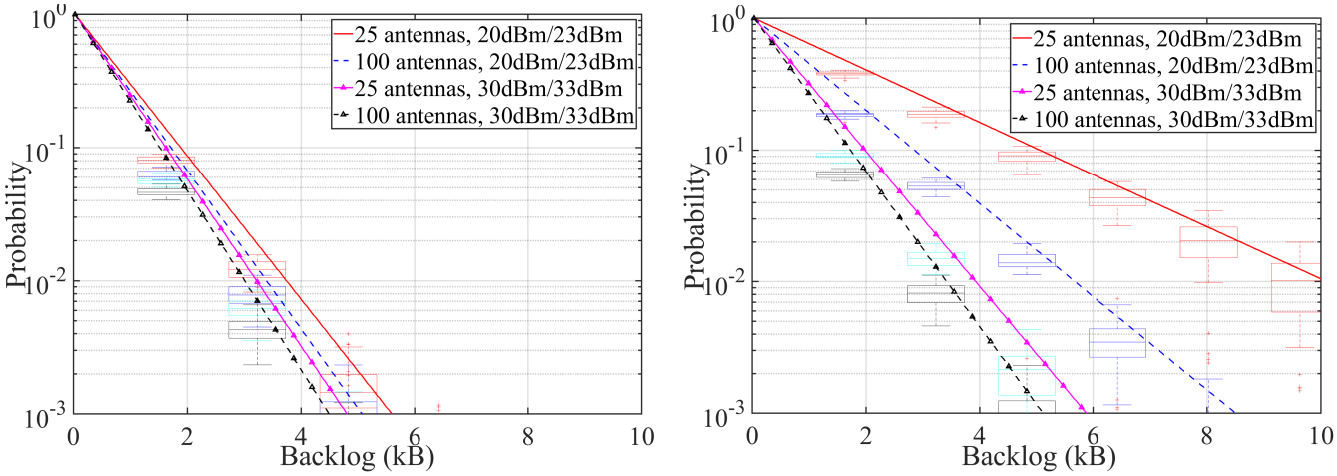


(a) DF Relay and RIS with 25 antennas

(b) DF Relay and RIS with 100 antennas

Fig. 13. The backlog violation analysis with different distribution of receivers.

relay equipped with 100 reflective elements achieved the best performance in reducing backlog and delay violations, whereas the RIS equipped with 100 reflective elements can trade off energy consumption and transmission QoS. Furthermore, the EDF scheduling policy trades off the delay performance between URLLC and eMBB traffics by introducing a proper



(a) DF relay-aided system

(b) RIS-aided system

Fig. 14. The backlog violation analysis with different power levels.

deadline threshold. Lastly, the proposed Martingale model provided extremely tight bounds for the distribution of backlog and delay violations, whereas the classical SNC model suffers from loose bounds. As this study aims to analyze the delay and backlog in URLLC and eMBB multiplexing systems, perfect CSI is not always achievable in practical implementations. Estimating CSI and accounting for channel changes can introduce additional latency in wireless systems. Therefore, the impact of imperfect CSI and its estimation on system performance will be investigated in future work.

REFERENCES

[1] "3GPP Release 17," https://www.3gpp.org/release-17, retrieved: 2022-08-12.

[2] Y. Zhao, X. Chi, L. Qian, Y. Zhu, and F. Hou, "Resource allocation and slicing puncture in cellular networks with eMBB and URLLC terminals coexistence," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18 431–18 444, Oct. 2022.

[3] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4585–4600, Jul. 2021.

[4] H. Peng, L.-C. Wang, and Z. Jian, "Data-driven spectrum partition for multiplexing URLLC and eMBB," *IEEE Trans. Cogn. Commun. Netw.*, early access, Dec. 2022, doi:10.1109/TCCN.2022.3231690.

[5] A. Anand, G. de Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.

[6] Y. Huang, Y. T. Hou, and W. Lou, "A deep-learning-based link adaptation design for eMBB/URLLC multiplexing in 5G NR," in *Proc. IEEE Conf. Computer Commun. (INFOCOM)*, Vancouver, Canada, May 2021.

[7] T. N. Weerasinghe, I. A. Balapuwaduge, and F. Y. Li, "Priority-based initial access for URLLC traffic in massive IoT networks: Schemes and performance analysis," *Comput. Netw.*, vol. 178, Sep. 2020, Art. no. 107360.

[8] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Joint resource and power allocation for URLLC-eMBB traffics multiplexing in 6G wireless networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, Canada, Jun. 2021.

[9] J. Li and X. Zhang, "Deep reinforcement learning-based joint scheduling of eMBB and URLLC in 5G networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 9, pp. 1543–1546, Sep. 2020.

[10] K. Li, P. Zhu, Y. Wang, F.-C. Zheng, and X. You, "Joint uplink and downlink resource allocation toward energy-efficient transmission for URLLC," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 7, pp. 2176–2192, Jul. 2023.

[11] V. D. P. Souto, S. Montejo-SÃąnchez, J. L. Rebelatto, R. D. Souza, and B. F. UchÃť a-Filho, "IRS-aided physical layer network slicing for URLLC and eMBB," *IEEE Access*, vol. 9, pp. 163 086–163 098, Dec. 2021.

[12] X. Xi, X. Cao, P. Yang, J. Chen, T. Q. S. Quek, and D. Wu, "Network resource allocation for eMBB payload and URLLC control information communication multiplexing in a multi-UAV relay network," *IEEE Trans. Commun.*, vol. 69, no. 3, pp. 1802–1817, Mar. 2021.

[13] H. Peng, L.-C. Wang, G. Ye Li, and A.-H. Tsai, "Long-lasting UAV-aided RIS communications based on SWIPT," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Austin, TX, Apr. 2022, pp. 1844–1849.

[14] H. Peng and L.-C. Wang, "Energy harvesting reconfigurable intelligent surface for UAV based on robust deep reinforcement learning," *IEEE Trans. Wireless Commun.*, early access, Feb. 2023, doi:10.1109/TWC.2023.3245820.

[15] E. BjÃűrnson, Ã. ÃŰzdogan, and E. G. Larsson, "Intelligent reflecting surface versus decode-and-forward: How large surfaces are needed to beat relaying?" *IEEE Wireless Commun. Lett.*, vol. 9, no. 2, pp. 244–248, Feb. 2020.

[16] G. Levin and S. Loyka, "Amplify-and-forward versus decode-and-forward relaying: Which is better?" in *Proc. Int. Zurich Seminar commun. (IZS)*, Zurich, Switzerland, Mar. 2012.

[17] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1304–1319, Feb. 2022.

[18] J. Cheng and C. Shen, "Relay-assisted uplink transmission design of URLLC packets," *IEEE Internet Things J.*, vol. 9, no. 19, pp. 18 839–18 853, Oct. 2022.

[19] M. Di Renzo, K. Ntontin, J. Song, F. H. Danufane, X. Qian, F. Lazarakis, J. De Rosny, D.-T. Phan-Huy, O. Simeone, R. Zhang, M. Debbah, G. Lerosey, M. Fink, S. Tretyakov, and S. Shamai, "Reconfigurable intelligent surfaces vs. relaying: Differences, similarities, and performance comparison," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 798–807, Jun. 2020.

[20] Y. Zhu, D. Zhou, M. Sheng, J. Li, and Z. Han, "Stochastic delay analysis for satellite data relay networks with heterogeneous traffic and transmission links," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 156–170, Jan. 2021.

[21] M. Khabbaz, J. Antoun, S. Sharafeddine, and C. Assi, "Modeling and delay analysis of intermittent V2U communication in secluded areas," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3228–3240, Feb. 2020.

[22] C. She, Y. Duan, G. Zhao, T. Q. S. Quek, Y. Li, and B. Vucetic, "Cross-layer design for mission-critical IoT in mobile edge computing systems," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9360–9374, Dec. 2019.

[23] R. Devassy, G. Durisi, G. C. Ferrante, O. Simeone, and E. Uysal, "Reliable transmission of short packets through queues and noisy channels under latency and peak-age violation guarantees," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 721–734, Apr. 2019.

[24] M. Alaslani, F. Nawab, and B. Shihada, "Blockchain in IoT systems: End-to-end delay evaluation," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8332–8344, Oct. 2019.

[25] NGMN Alliance, "5G E2E technology to support verticals URLLC requirements," *NGMN Alliance*, pp. 1–50, Oct. 2019.

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

[27] M. Mei, Q. Yang, M. Qin, K. S. Kwak, and R. R. Rao, "QoS-driven stochastic analysis for heterogeneous cognitive radio networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Seoul, South Korea, May 2020.

[28] S. Schiessl, J. Gross, M. Skoglund, and G. Caire, "Delay performance of the multiuser MISO downlink under imperfect CSI and finite-length coding," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 4, pp. 765–779, Apr. 2019.

[29] M. Mei, M. Yao, Q. Yang, M. Qin, K. S. Kwak, and R. R. Rao, "Delay analysis of mobile edge computing using poisson cluster process modeling: A stochastic network calculus perspective," *IEEE Trans. Commun.*, vol. 70, no. 4, pp. 2532–2546, Apr. 2022.

[30] A. Iqbal, U. Javed, S. Saleh, J. Kim, J. S. Alowibdi, and M. U. Ilyas, "Analytical modeling of end-to-end delay in openflow based networks," *IEEE Access*, vol. 5, pp. 6859–6871, Dec. 2017.

[31] Y. Hu, H. Li, Z. Chang, and Z. Han, "End-to-end backlog and delay bound analysis for multi-hop vehicular ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6808–6821, Oct. 2017.

[32] F. Poloczek and F. Ciucu, "A Martingale-envelope and applications," *ACM SIGMETRICS Perform. Eval. Rev.*, vol. 41, no. 3, pp. 43–45, Jan. 2014.

[33] S. Zoppi, J. P. Champati, J. Gross, and W. Kellerer, "Dynamic scheduling for delay-critical packets in a networked control system using WirelessHART," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Dublin, Ireland, Jun. 2020, pp. 1–7.

[34] 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, Oct. 2022.

[35] B. Picano, R. Fantacci, and Z. Han, "Aging and delay analysis based on Lyapunov optimization and Martingale theory," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8216–8226, Aug. 2021.

[36] F. Saggese, M. Moretti, and P. Popovski, "Power minimization of downlink spectrum slicing for eMBB and URLLC users," *IEEE Trans. Wireless Commun.*, early access, Jul. 2022, doi:10.1109/TWC.2022.3189396.

[37] M. Setayesh, S. Bahrami, and V. W. Wong, "Resource slicing for eMBB and URLLC services in radio access network using hierarchical deep learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 11, pp. 8950–8966, Nov. 2022.

[38] M. Almekhlafi, M. A. Arfaoui, C. Assi, and A. Ghrayeb, "Enabling URLLC applications through reconfigurable intelligent surfaces: Challenges and potential," *IEEE Internet Things Mag.*, vol. 5, no. 1, pp. 130–135, Mar. 2022.

[39] M. Almekhlafi, M. A. Arfaoui, M. Elhattab, C. Assi, and A. Ghrayeb, "Joint scheduling of eMBB and URLLC services in RIS-aided downlink cellular networks," in *Proc. Int. Conf. Comput. Commun. Netw. (ICCCN)*, Athens, Greece, Aug. 2021.

[40] X. Zhang, J. Wang, and H. V. Poor, "Joint optimization and tradeoff modeling for peak AoI and delay-bound violation probabilities over URLLC-enabled wireless networks using FBC," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, Canada, Jun. 2021.

[41] T. Liu, L. Sun, R. Chen, F. Shu, X. Zhou, and Z. Han, "Martingale theory-based optimal task allocation in heterogeneous vehicular networks," *IEEE Access*, vol. 7, pp. 122 354–122 366, May 2019.

[42] R. Fantacci, T. Pecorella, B. Picano, and L. Pierucci, "Martingale theory application to the delay analysis of a multi-hop aloha NOMA scheme in edge computing systems," *IEEE/ACM Trans. Netw.*, vol. 29, no. 6, pp. 2834–2842, Dec. 2021.

[43] H. Peng, H. Rahbari, S. J. Yang, and L.-C. Wang, "Non-cooperative learning for robust spectrum sharing in connected vehicles with malicious agents," in *Proc. IEEE Glob. Commun. Conf.*, Rio de Janeiro, Brazil, Dec. 2022, pp. 1769–1775.

[44] B. Shi, F.-C. Zheng, C. She, J. Luo, and A. G. Burr, "Risk-resistant resource allocation for eMBB and URLLC coexistence under M/G/1 queueing model," *IEEE Trans. Veh. Techn.*, vol. 71, no. 6, pp. 6279–6290, Mar. 2022.

[45] B. S. Khan, S. Jangsher, A. Ahmed, and A. Al-Dweik, "URLLC and eMBB in 5G industrial IoT: A survey," *IEEE Open J. Commun. Soc.*, vol. 3, no. 1, pp. 1134–1163, Jul. 2022.

[46] R. Abreu, T. Jacobsen, G. Berardinelli, K. Pedersen, I. Z. KovÃącs, and P. Mogensen, "Power control optimization for uplink grant-free URLLC," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Barcelona, Spain, Apr. 2018, pp. 1–6.

[47] B. Liu, P. Zhu, J. Li, D. Wang, and Y. Wang, "Energy-efficient optimization via joint power and subcarrier allocation for eMBB and URLLC services," *IEEE Wireless Commun. Lett.*, vol. 11, no. 11, pp. 2340–2344, Nov. 2022.

[48] H. Xie, J. Xu, and Y.-F. Liu, "Max-min fairness in IRS-aided multi-cell MISO systems with joint transmit and reflective beamforming," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1379–1393, Feb. 2021.

[49] E. BjÃűrnson and L. Sanguinetti, "Power scaling laws and near-field behaviors of massive MIMO and intelligent reflecting surfaces," *IEEE Open J. Commun. Soc.*, vol. 1, no. 1, pp. 1306–1324, Sep. 2020.

[50] M. Y. Javed, N. Tervo, and A. PÃd'rssinen, "Inter-beam interference reduction in hybrid mmw beamforming transceivers," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Bologna, Italy, Sep. 2018, pp. 220–224.

[51] C. You, B. Zheng, and R. Zhang, "Channel estimation and passive beamforming for intelligent reflecting surface: Discrete phase shift and progressive refinement," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2604–2620, Nov. 2020.

[52] H. Peng, C.-Y. Ho, Y.-T. Lin, and L.-C. Wang, "Energy-efficient symbiotic radio using generalized benders decomposition," in *Proc. IEEE 96th Veh. Technol. Conf. (VTC2022-Fall)*, London/Beijing, Sep. 2022.

[53] Y. Jiang, "Stochastic network calculus for performance analysis of internet networks–An overview and outlook," in *Proc. IEEE Int. Conf. Comput., Netw. Commun. (ICNC)*, Maui, HI, Jan. 2012, pp. 638–644.

[54] 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.

[55] O. Adamuz-Hinojosa, V. Sciancalepore, P. Ameigeiras, J. M. Lopez-Soler, and X. Costa-PÃl'rez, "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.

[56] Y. Jiang, "A basic stochastic network calculus," *SIGCOMM Comput. Commun. Rev.*, vol. 36, no. 4, pp. 123–134, Aug. 2006.

[57] Q. Zhang, Y. Zhu, D. Zhou, Y. Dong, F. Xie, and Z. Han, "Martingale theory-based delay bound analysis for two-hop heterogeneous networks," *has submitted to IEEE Wireless Commun. Lett.*

[58] "3GPP TS 38.211 Release 15," http://www.etsi.org/standards-search, retrieved: 2018-07.

[59] A. A. Nasir, H. D. Tuan, H. H. Nguyen, M. Debbah, and H. V. Poor, "Resource allocation and beamforming design in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2021.

[60] "Simulation evaluation of 802.11ax for IMT-2020 eMBB dense urban scenario," https://mentor.ieee.org/802.11/dcn/19/11-19-0871-00-AANI-802-11ax-for-imt-2020-embb-dense-urban.pptx, retrieved: 2019-11-07.

[61] W. Zhang, M. Derakhshani, and S. Lambotharan, "Stochastic optimization of URLLC-eMBB joint scheduling with queuing mechanism," *IEEE Wireless Commun. Lett.*, vol. 10, no. 4, pp. 844–848, Apr. 2021.

[62] W. Sui, X. Chen, S. Zhang, Z. Jiang, and S. Xu, "Energy-efficient resource allocation with flexible frame structure for hybrid embb and urllc services," *IEEE Trans. Green Commun. Netw.*, vol. 5, no. 1, pp. 72–83, Mar 2021.

[63] W. Wang and W. Zhang, "Intelligent reflecting surface configurations for smart radio using deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 8, pp. 2335–2346, Aug. 2022.

[64] Y. L. Lee, D. Qin, L.-C. Wang, and G. H. Sim, "6G massive radio access networks: Key applications, requirements and challenges," *IEEE Open J. Veh. Techn.*, vol. 2, no. 1, pp. 54–66, Dec. 2021.

Haoran Peng (Member, IEEE) received the B.Eng. degree in software engineering from the University of Electronic Science and Technology of China, in 2015, and the Ph.D. degree (with honored) in electrical and computer engineering from the National Yang Ming Chiao Tung University, in 2022.

From 2015 to 2018, he was a full-time software engineer. From June 2021 to August 2021, he was a Visiting Student Research Collaborator with the Global Cybersecurity Institute in the Golisano College of Computing and Information Sciences, Rochester Institute of Technology. His current research interests mainly include optimization and machine learning for wireless communications. He has served or is currently serving as a reviewer for several top-tier peer-reviewed journals, and has also been a TPC (Technical Program Committee) member for several international conferences. He was awarded the IEEE VTS Student Travel Grant in VTC2022-Fall. He was honored with the scholarship award for excellence in research at the 2021 CTCI Foundation Science and Technology Scholarship, in recognition of his outstanding performance. He was also honored the Outstanding Graduate Student at National Yang Ming Chiao Tung University in 2022.

Ching-Chieh, Hsia received the B.S. degree in computer science and information engineering from Tamkang University in 2021, and the M.S. degree in electronics and electrical engineering from National Yang Ming Chiao Tung University in 2023. His research interests include stochastic network calculus, and martingale theory analysis.

Zhu Han (S'01–M'04–SM'09–F'14) received a B.S. degree in electronic engineering from Tsinghua University, in 1997, and M.S. and Ph.D. degrees in electrical and computer engineering from the University of Maryland, College Park, in 1999 and 2003, respectively.

From 2000 to 2002, he was an R&D Engineer of JDSU, Germantown, Maryland. From 2003 to 2006, he was a Research Associate at the University of Maryland. From 2006 to 2008, he was an assistant professor at Boise State University, Idaho. Currently, he is a John and Rebecca Moores Professor in the Electrical and Computer Engineering Department as well as in the Computer Science Department at the University of Houston, Texas. Dr. Han's main research targets on the novel game-theory related concepts critical to enabling efficient and distributive use of wireless networks with limited resources. His other research interests include wireless resource allocation and management, wireless communications and networking, quantum computing, data science, smart grid, security and privacy. Dr. Han received an NSF Career Award in 2010, the Fred W. Ellersick Prize of the IEEE Communication Society in 2011, the EURASIP Best Paper Award for the Journal on Advances in Signal Processing in 2015, IEEE Leonard G. Abraham Prize in the field of Communications Systems (best paper award in IEEE JSAC) in 2016, and several best paper awards in IEEE conferences. Dr. Han was an IEEE Communications Society Distinguished Lecturer from 2015-2018, AAAS fellow since 2019, and ACM distinguished Member since 2019. Dr. Han is a 1% highly cited researcher since 2017 according to Web of Science. Dr. Han is also the winner of the 2021 IEEE Kiyo Tomiyasu Award (an IEEE Technical Field Award), for outstanding early to mid-career contributions to technologies holding the promise of innovative applications, with the following citation: "for contributions to game theory and distributed management of autonomous communication networks."

Li-Chun Wang (M'96 – SM'06 – F'11) received Ph. D. degree from the Georgia Institute of Technology, Atlanta, in 1996. From 1996 to 2000, he was a Senior Technical Staff Member at AT&T Laboratories. Since August 2000, he has joined National Yang Ming Chiao Tung University (NYCU) in Taiwan. He is now the Chair Professor and serves the Dean of College of Electrical and Computer Engineering of NYCU.

Dr. Wang was elected to the IEEE Fellow in 2011 for his contributions to cellular architecture and radio resource management in wireless networks. He won two Distinguished Research Awards from National Science and Technology Council (2012, 2017), IEEE Communications Society Asia-Pacific Board Best Award (2015), Y. Z. Hsu Scientific Paper Award (2013), and IEEE Jack Neubauer Best Paper Award (1997). He was recognized as Top 2% Scientists Worldwide in a study from Stanford University. His recent research interests are in the areas of cross-layer optimization for wireless systems, AI-enabled radio resource management for heterogeneous mobile networks, and big data analysis for industrial Internet of things. He holds 26 US patents, and has published over 300 journal and conference papers, and co-edited the book, âĂIJKey Technologies for 5G Wireless Systems,âĂİ (Cambridge University Press 2017).