E2E Performance Modeling for Slice-Based Video Streaming With Layered Encoding

Yannan Wei, *Graduate Student Member, IEEE*, Qiang Ye, *Senior Member, IEEE*, Kaige Qu, *Member, IEEE*, Weihua Zhuang, *Fellow, IEEE*, and Xuemin Shen, *Fellow, IEEE*

***Abstract*—In this paper, we present a performance analytical model for end-to-end (E2E) service provisioning (i.e., processing or transmission) of layer-encoded video packets over a network slice in the core network. The disparate service reliability requirements of base layer (BL) and enhancement layer (EL) packets are considered in the proposed analytical model for the E2E packet delays, deadline violation probabilities, and throughputs of BL and EL packets. Specifically, a network function virtualization (NFV) node along the routing path of the video streaming slice is split into two consecutive logical nodes, one for packet processing and the other for transmission, based on which a segment-based analysis framework is proposed for E2E service performance modeling. A two-stage queuing model is established to obtain the approximate steady-state probability distribution of queue length at the first node in the first segment, upon which the BL/EL packet delay, deadline violation probability, and throughput at the segment are derived. In addition, the inter-departure time of successive packets departing from the first segment is analyzed based on an approximate M/D/1 system, and the packet departure process at the first segment is approximated as a Poisson process under the assumption of a large packet service rate of the first node. The independence between two consecutive segments is then achieved for analysis tractability, based on which the E2E performance measures are derived. Extensive simulation results demonstrate the accuracy of our proposed performance analytical model and its effectiveness such as in transport parameter determination.**

***Index Terms*—Performance analytical modeling, video streaming with layered encoding, network slice, deadline violation probability.**

I. Introduction

FUTURE mobile communication networks are expected to support diversified emerging services with stringent quality-of-service (QoS) requirements, such as integrated sensing and communication (ISAC) [1], connected autonomous vehicles (CAVs) [2], and mixed reality (XR) [3]. According to the Ericsson's report, over half of all traffic crossing mobile networks is video related, and the proportion is projected to be around 80% by the end of 2029 [4]. The pervasiveness of video streaming services manifests various media formats in different aspects, from traditional on-demand/live streaming to short-form videos (e.g., TikTok and e-commerce) and to XR-related videos (e.g., volumetric streaming) which provide viewers with full immersiveness and six degrees of freedom (6DoF) [5], [6], [7], [8].

A video streaming service is delay-sensitive and requires a large service (i.e., processing or transmission) rate for smooth high-resolution (e.g., 4K) video content playback, which, however, suffers from network condition variations (e.g., throughput and congestion level). To cope with network dynamics, layered video encoding is widely adopted to support adaptive video streaming. Specifically, for traditional on-demand video streaming, a video clip is temporally divided into a sequence of video segments, each of which is encoded into one base layer (BL) and several interdependent enhancement layers (ELs) with scalable video coding (SVC) [9], [10]. For omnidirectional 360° video streaming, a video segment is further spatially partitioned into multiple non-overlapping tiles. Each video tile is similarly encoded into one BL tile and several EL tiles with scalable high-efficiency video coding (SHVC) [11], [12]. The BL corresponds to the lowest bitrate for ensuring video robustness which is an essential layer of decoding and can be decoded independently. The EL, as a supplement to BL, increases the coding bitrate and thus provides improved video quality, which can be decoded only if the lower-level BL is decoded. Layered video encoding enables additional flexibility in response to network traffic dynamics, where some in-transit EL packets can be intelligently discarded for congestion alleviation [13], [14]. Therefore, BL and EL packets are not only delay-sensitive with strict delivery deadlines but also have differentiated service reliability requirements. BL packets need to be reliably delivered to avoid long video rebuffering time, while deadline-violated EL packets that no longer enhance the video quality should be discarded to increase the service efficiency.

In order to support the coexistence of different services over a common physical substrate network (e.g., a core network) with satisfactory service performance, network slicing has been proposed and seen as a promising enabler, where the software-defined networking (SDN) and network function virtualization (NFV) are two key technologies to facilitate agile and scalable service deployment and network management

Received 5 July 2024; revised 8 March 2025 and 8 May 2025; accepted 28 June 2025; approved by IEEE TRANSACTIONS ON NETWORKING Editor J. Wu. Date of publication 31 July 2025; date of current version 18 December 2025. This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada. *(Corresponding author: Kaige Qu.)*

Yannan Wei, Weihua Zhuang, and Xuemin Shen are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: y272wei@uwaterloo.ca; wzhuang@uwaterloo.ca; sshen@uwaterloo.ca).

Qiang Ye is with the Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: qiang.ye@ucalgary.ca).

Kaige Qu was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: k2qu@uwaterloo.ca).

Digital Object Identifier 10.1109/TON.2025.3591401

[15], [16]. Multiple virtual networks, also known as *network slices*, can be created over a common physical network to support different services with diverse performance demands [17], [18]. Each service is represented by a service function chain (SFC) which is a specific sequence of virtual network functions (VNFs) interconnected in a predefined order. An SFC is embedded in the physical network by the SDN/NFV controller as a virtual network for supporting a particular service, where optimal routing paths can be established with dedicated processing and transmission resources reserved on NFV nodes and physical links, respectively [19], [20].

A network slice with flexible resource orchestration can be deployed to support the layered-encoding video streaming with fine-grained QoS provisioning. To evaluate the service provisioning performance and optimize the virtual network deployment, an end-to-end (E2E) performance analytical model is essential to evaluate the E2E delays and E2E deadline violation probabilities of BL and EL packets. Deadline violation probability is a critical performance metric to evaluate, since video packets delivered within an average delay bound may not always improve video smoothness/quality due to service deadline violations. Specifically, deadline-violated BL packets result in the decoding failure of other timely-delivered EL packets, while deadline-violated EL packets (if not discarded) do not enhance video quality. In addition, customized protocol functionalities and operations can be deployed and implemented on certain NFV nodes and SDN switches within the network slice for enhanced video packet delivery [14], [21], [22], [23]. A performance analytical model can also be helpful in determining key transport parameters in those protocol functionalities.

The impact of disparate service reliability requirements of BL and EL packets in multi-hop layer-encoded video packet transmissions is not considered in existing analytical approaches [24], which makes the E2E performance modeling, especially the analysis of deadline violation probability, challenging to perform. The packet service performance at one service node is challenging to analyze due to deadline-violated EL packet dropping. BL and EL packets may experience different average packet delays. The steady-state probability distribution of queue length at the node is difficult to obtain, thus making it challenging to determine the packet deadline violation probability. In addition, queuing dynamics at one node are closely correlated to those at its preceding node, which makes the E2E performance modeling even more challenging. The packet arrival process at one node is equivalent to the packet departure process at its preceding node, which is difficult to analyze especially when deadline-violated EL packets are directly discarded without being further forwarded, leading to reduced throughput. Therefore, the correlations between packet services at two consecutive nodes need to be investigated in the E2E performance analytical modeling.

In this paper, we present an E2E service performance modeling for layer-encoded video traffic traversing a video streaming slice in the core network, where the discrepant service reliability requirements of BL and EL packets are considered. A customized caching-based packet retransmission functionality is proposed and deployed in the considered video streaming slice to achieve efficient and prompt packet retransmissions for enhancing video packet delivery. The key transport parameters including the cycle of the periodic timer and the minimum required caching buffer size for caching buffer release are determined based on the proposed performance analytical model. Specifically, each NFV node along the routing path of the video streaming slice is split into two consecutive logical nodes, one for packet processing and the other for packet transmission, with respective packet service rate, and the E2E packet processing and transmission over the video streaming slice is abstracted as a multi-hop packet service process with various service reliability requirements. Based on the analysis of the correlations between two consecutive nodes, a segment-based analysis framework is proposed. A segment consists of several consecutive nodes along the routing path of the video streaming slice where the packet service rate of the first node is smaller than those of the rest of nodes in the segment. Within each segment, a two-stage queuing model is established to obtain the approximate steady-state probability distribution of queue length at the first node, based on which the BL/EL packet delay, deadline violation probability, and throughput at the segment are derived. In addition, the departure process from a segment is discussed based on an approximated M/D/1 queuing system, upon which the E2E performance is derived. The main contributions of this paper are summarized as follows.

- A segment-based analysis framework is proposed for E2E service performance modeling over a video streaming slice in the core network. A two-stage queuing model is established to derive the approximate steady-state probability distribution of queue length at the first node in each segment with the consideration of disparate service reliability requirements of BL and EL packets;
- The departure process of successive packets departing from a segment is analyzed as a mixed process alternating between Poisson and deterministic processes. The independence between two consecutive segments is achieved under the assumption of a large packet service rate of the first node in the preceding segment for E2E analysis tractability;
- The proposed performance analytical model determines the E2E packet delays, the E2E deadline violation probabilities, and the E2E throughputs of BL and EL packets;
- The accuracy of the proposed E2E performance analytical model and its effectiveness in (slice) resource reservation and calculation of key transport-layer parameters in the proposed caching-based packet retransmission scheme are demonstrated via extensive simulations.

The rest of this paper is organized as follows. Section II reviews the related work. The system model is described in Section II. Section IV presents our proposed E2E performance analytical model for video packet services over a video streaming slice with layered encoding. Simulation results are discussed in Section V, and conclusions are drawn in Section VI. A list of important notations is given in Table I.

TABLE I
LIST OF IMPORTANT NOTATIONS

Symbol	Definition
W	The number of transmission links in a video streaming slice
k_c	Cycle of the periodic timer
$C^b_{\min}/C^e_{\min}$	Minimum required caching buffer size (in packet) for BL/EL packets
μ^g	Packet service rate (in packet/s) of the first node in segment g
λ^g	Total packet arrival rate at segment g. $\lambda^g = \lambda^g_b + \lambda^g_e$ where λ^g_b and λ^g_e are the arrival rates of BL and EL packets arriving at segment g, respectively
$\bar{d}^{b,g}/\bar{d}^{e,g}$	Average remaining service deadline of BL/EL packets arriving at segment g
N^1_e	Queue length bound for EL packets that can successfully pass through the first node in the first segment without being discarded due to service deadline violations
$\hat{q}_i, i \geq 0$	Approximated steady-state probability distribution of queue length at the first node in a segment
$P_j, j = 0, 1, \cdots, N^1_e$	Steady-state probability distribution of queue length for stage one queuing
$\pi_n, n \geq 0$	Steady-state probability distribution of queue length for stage two queuing
ρ^1_1/ρ^1_2	Traffic intensity for stage one/two queuing at the first node in the first segment
$\tilde{T}^1$	Total service time for a packet passing through the rest of nodes except the first one in the first segment
T^1_i	Time consumption for the ith packet in the queue of the first node traversing the first segment
$T^1_\theta, \theta \in \{b, e\}$	Average BL/EL packet delay for traversing the first segment
$j^\theta - 1, \theta \in \{b, e\}$	Queue length bound for BL/EL packets at the first node that can successfully traverse a segment without violating their service deadlines
$V^1_\theta, \theta \in \{b, e\}$	Deadline violation probability of BL/EL packets when traversing the first segment
λ^1_t	Total throughput at the first segment. $\lambda^1_t = \lambda^1_{t,b} + \lambda^1_{t,e}$ where $\lambda^1_{t,b}$ and $\lambda^1_{t,e}$ are the throughputs of BL and EL traffic after traversing the first segment, respectively.
Y^1	Inter-departure time of successive packets departing from the first segment
$T_\theta/V_\theta/\lambda_{t,\theta}$, $\theta \in \{b, e\}$	E2E packet delay, deadline violation probability, and throughput of BL/EL packets traversing the video streaming slice

II. RELATED WORK

Many works have investigated network slice deployment to support diversified services with various service requirements such as video streaming [14], [25]. Before the deployment of a network slice for a particular service, in the planning stage, a network operator needs to ensure that the service requirements in terms of delay, throughput, and/or reliability are met in the long run. Thus, it is necessary to conduct E2E performance analytical modeling with given amount of slicing resources and service traffic characteristics. Some works focus on the radio access network (RAN) while others focus on the core network [26], [27], [28], [29]. For RAN slicing, it is usually modeled as a single-hop packet service process. For core network slicing scenario, a network slice is described as an SFC embedded over a physical network, where a service traffic flow passes through a sequence of VNFs of the SFC in a multi-hop manner. For either a single-hop or a multi-hop case, existing analytical approaches can be categorized into stochastic network calculus (SNC)-based modeling and queuing theory-based modeling [24].

With the assumption of Poisson traffic arrivals, queuing theory-based approaches (e.g., M/M/1/K, M/D/1/K) are used to derive closed-form steady-state probability distribution of queue length (or packet delay) of a service node in many cases, and certain performance metrics are analyzed, e.g., average packet delay [30], [31], [32]. For a multi-hop setting, existing works model a packet service as a tandem of independent queues, and the service process at each hop is modeled as an M/M/1 or M/D/1 queuing system [33], [34], [35], [36]. Then, the E2E packet delay is calculated as a summation of per-hop packet delays. Different from queuing theory-based approaches, SNC-based approaches aim at a non-asymptotic performance bound on packet delay in the form of *Probability[delay > budget] ≤ violation probability* by characterizing the (affine) arrival and service envelopes [26], [37], [38]. It is able to deal with various types of packet arrivals and services, and provides a conservative estimation of such delay bound, which may deviate from the real measurements, at the expense of obtaining the exact distribution of queue dynamics and average delay performance [26], [29].

The disparity in service reliability requirements between BL and EL packets in layer-encoded video streaming poses significant challenges on E2E performance analysis, especially the deadline violation probability, which is not considered and addressed in existing analytical approaches. In addition, as discussed in Section I, the correlations between packet services at two consecutive nodes require further investigation in the E2E performance modeling. In this paper, we present a practical E2E performance analytical model for slice-based layer-encoded multi-hop video packet transmissions, where the impact of distinct service reliability requirements of BL and EL packets is studied and captured.

III. SYSTEM MODEL

A. Network Model

We consider an SDN/NFV-based core network, as shown in Fig. 1. With the SDN/NFV controller, multiple SFCs are embedded in the core network to create different network slices for supporting diverse services (e.g., video streaming service), each of which consists of a sequence of NFV nodes and SDN switches interconnected in a predefined order. The virtual network topology of each core network slice (assumed

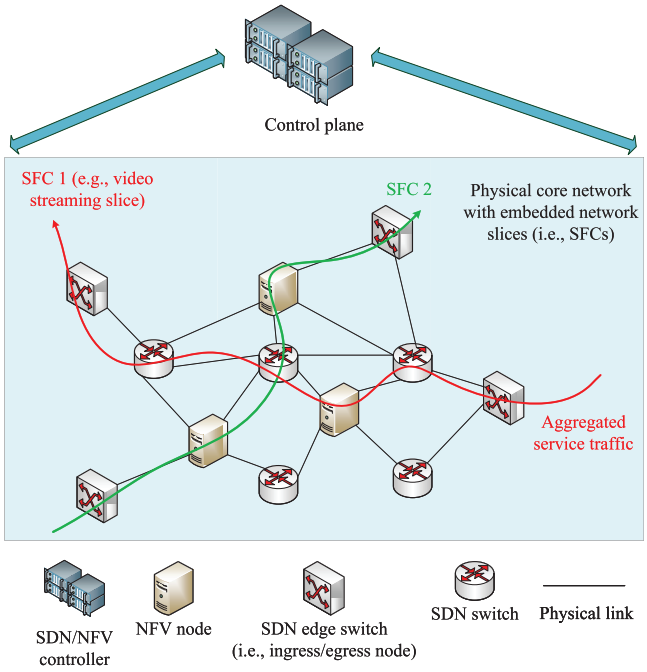


Fig. 1. An SDN/NFV-based core network with embedded network slices.

linear for simplicity) is configured where dedicated virtual network resources including processing and link transmission resources are reserved on NFV nodes and physical links along the route for data processing and transmission, respectively. Customized protocol functionalities (e.g., retransmission [21] and selective caching [14]) can be deployed on certain network nodes (e.g., an NFV node or an SDN edge switch) within a core network slice for supporting fine-grained transmission control and enhanced service provisioning. Here, we consider that a network slice for supporting the video streaming service is deployed between each pair of ingress and egress edge switches in the core network.[1]

B. Video Traffic Model

We consider a video streaming service with layered encoding. Specifically, each video stored in a remote server on the Internet is temporally divided into a sequence of video segments, each with a playback time of 1-10 seconds [8]. Each video segment is encoded into multiple interdependent layers, including one BL and several ELs, to support adaptive video streaming. BL ensures video robustness with basic video quality, while EL enhances video quality. An EL can be decoded only if the corresponding BL and all lower-layer ELs are received and decoded. Therefore, EL packets are optional in terms of video decoding for smooth video playback. BL packets have a higher service reliability requirement than EL packets, indicating that BL packets need to be reliably delivered to avoid long video rebuffering time, while deadline-violated EL packets that do not improve video quality should be discarded to save processing/transmission resources.

[1]For simplicity, each network switch is also referred to as a node.

The virtual routing path of a video streaming network slice is shown in Fig. 2. The BL and EL traffic flows aggregated from multiple video clients served by the same video streaming slice are considered as two independent Poisson arrival processes at the ingress node [26], [30], [32], [39], with average packet arrival rates denoted by λ_b and λ_e in packet/s, respectively. The virtual routing path between the ingress (s_0) and egress (s_W) nodes consists of $(W+1)$ nodes including NFV nodes for data packet processing and transmission nodes for packet forwarding. Denote the packet processing/transmission rate of NFV/transmission node i ($i = 0, 1, \cdots, W$) by μ_i in packet/s. At each node, packets are transmitted/processed based on the first-in-first-out (FIFO) principle, and the remaining service deadline of each EL packet is checked. Deadline-violated EL packets are directly discarded without being further forwarded.

C. Caching-Based Packet Retransmission Functionality

To support enhanced layer-encoded video packet services, we propose a customized caching-based packet retransmission functionality to achieve efficient packet retransmission. Specifically, we consider that the ingress node has an additional caching buffer for aggregated video traffic of each video streaming slice which temporarily stores video packets sent but not acknowledged for possible retransmissions. Besides, the egress node is considered to be equipped with the capability to detect any packet loss due to random link failures (referred to as random packet loss) and differentiate from the EL packet losses due to service deadline violations.[2] The workflow of the proposed caching-based packet retransmission functionality is given in Fig. 3. When the egress node detects a random BL/EL packet loss, it sends a retransmission request to the ingress node, and the ingress node retransmits the lost packet by retrieving the corresponding packet copy in the caching buffer. If the requested packet copy is unavailable at the ingress node, the lost packet is retransmitted by a remote server, with a longer packet retransmission delay.

Due to limited caching resources, the caching buffer at the ingress node should be regularly released to avoid buffer overflow. We consider a time-based caching buffer release mechanism. Specifically, a timeout timer is set at the ingress node to periodically release the cached packet copies in the caching buffer. The workflow of time-based caching buffer release is shown in Fig. 4. The timer initialization phase is the time interval where the ingress node transmits BL/EL packets received from remote servers for the first time and stores their packet copies in its caching buffer. Then, the ingress node starts the periodic timer and continues putting the copies of subsequently received packets into the caching buffer. Every time the periodic timer timeouts, the ingress node releases the cached packet copies that are stored before the timer starts from the caching buffer. The periodic timer restarts when it timeouts (or after each caching buffer release).

In order to balance between the performance of the proposed caching-based packet retransmission scheme and the caching

[2]Due to the low service reliability requirement, deadline-violated EL packets do not need to be retransmitted.

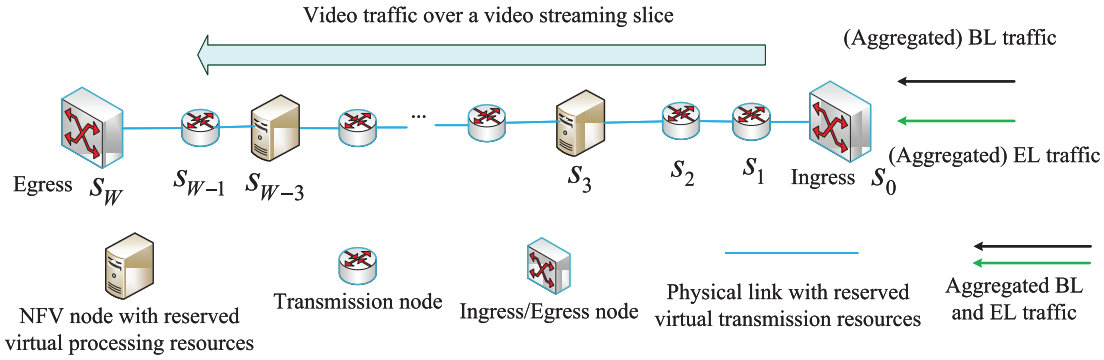


Fig. 2. The virtual routing path of a video streaming slice with aggregated BL and EL traffic.

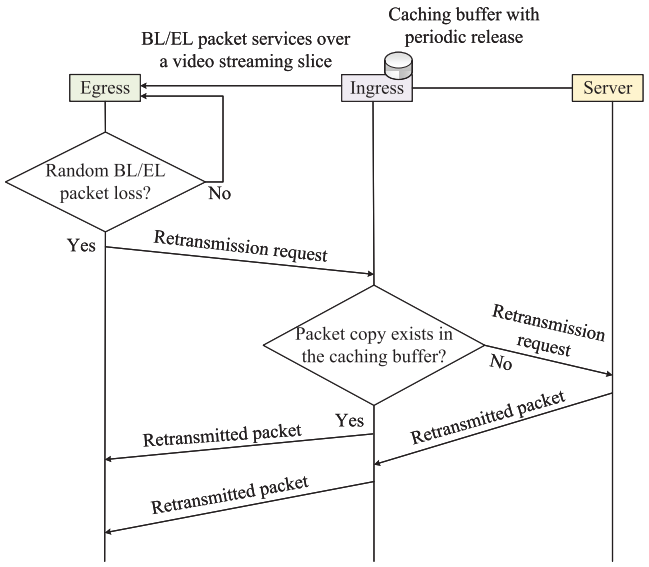


Fig. 3. The workflow of caching-based packet retransmission.

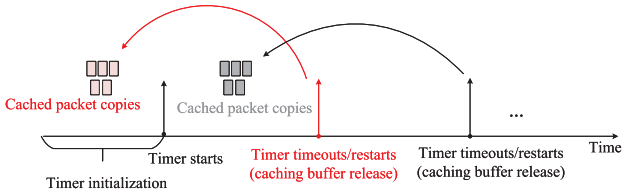


Fig. 4. The workflow of time-based caching buffer release.

resource utilization, two key transport parameters need to be determined. The first key transport parameter is the cycle of the periodic timer, which needs to be set based on the estimated average E2E BL/EL packet delay,[3] such that, on average, the cached packet copies in the caching buffer are released after the original packets are successfully received by the egress node. If the cycle of the periodic timer is smaller than the estimated average E2E BL/EL delay, video packets transmitted by the ingress node will arrive at the egress node after the corresponding cached packet copies are released from the caching buffer (referred to as '*release-before-arrival*' packets), and any lost packets will be retransmitted from remote servers on the Internet, compromising the performance of the proposed caching-based packet retransmission scheme. Correspondingly, a minimum required caching buffer size (in packets) for BL/EL packets, as the other key transport parameter, needs to be decided, with which copies of arriving packets at the ingress node can be put into the caching buffer instead of being dropped due to buffer overflow.[4] If the caching buffer size is set to be larger than the minimum required value, the caching resources are not efficiently utilized.

Suppose we obtain the average E2E BL and EL packet delays for traversing the video streaming slice between a pair of ingress and egress nodes, denoted by T_b and T_e respectively. Let k_I and k_c be the duration of the timer initialization phase and the cycle of the periodic timer, respectively. Based on the above discussions, we have $k_I = k_c$, $k_c = T_b$ for BL packets, and $k_c = T_e$ for EL packets. Thus, the minimum required caching buffer size for BL packets, $C^b_{\min}$, is given by

$$C^b_{\min} = \lceil \lambda_b \cdot 2T_b \rceil \tag{1}$$

where $\lceil \cdot \rceil$ is the ceiling function. Besides, Let $\bar{d}^e$ be the average service deadline of arriving EL packets at the ingress node. Since deadline-violated EL packets are directly released from the caching buffer due to a low service reliability requirement, the minimum required caching buffer size for EL packets, $C^e_{\min}$, is given by

$$C^e_{\min} = \left\lceil \min\left(\lambda_e \cdot 2T_e, \lambda_e \bar{d}^e\right) \right\rceil. \tag{2}$$

Based on Eqs. (1) - (2), we obtain the minimum required total caching buffer size for both BL and EL packets, denoted by $C_{\min} = C^b_{\min} + C^e_{\min}$.

IV. E2E Service Performance Modeling

In this section, we discuss our E2E performance modeling for (aggregated) video traffic with layered encoding delivered

[3]The time for the egress node to detect a random packet loss until the ingress node receives the retransmission request is neglected.

[4]Additional caching resources need to be assigned against the burstiness of packet arrivals, which is out of the scope of this work.

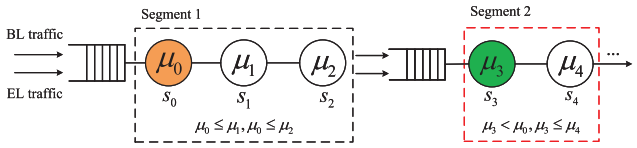


Fig. 5. A segment-based analysis framework.

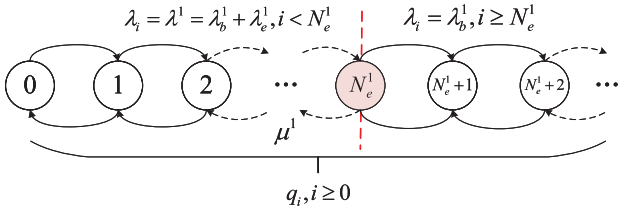


Fig. 6. The state transition diagram of queue length at the first node in the first segment.

over a video streaming slice. The analytical model determines the average E2E packet delay, the average E2E deadline violation probability, and the average E2E throughput, while considering the discrepant service reliability requirements of BL and EL packets. For any NFV node along the routing path of the video streaming slice, arriving packets are first queued for processing, and the processed packets are queued for transmission to the subsequent node. Accordingly, each NFV node is split into two consecutive logical nodes, one for packet processing and the other for transmission. In the following, any processing or transmission node is referred to as service node s_i ($i = 0, 1, \cdots, W'$), where $(W'+1)$ is the total number of nodes composing the video streaming slice with the consideration of the NFV node logic splitting. The E2E service over the video streaming slice is abstracted as a multi-hop packet service system with various service reliability requirements, where each hop/node is either a processing or transmission node. Here, we suppose each (abstracted) node s_i along the considered video streaming slice has a deterministic service rate μ_i (i.e., constant service time), allocated and reserved by the SDN/NFV controller at the time of slice deployment.

A. Segment-Based Analysis Framework

During multi-hop packet service, the queuing dynamics of each intermediate node are closely correlated to the packet service rate or packet departure process of its preceding node. Specifically, for any intermediate node between the ingress and egress nodes, s_i ($i > 0$), the packet arrival process is a slightly delayed version of the packet departure process at its preceding node s_{i-1} (due to propagation delay). In addition, given the packet service rates of nodes, if the packet service rate μ_i of s_i is greater than or equal to that of s_{i-1}, i.e., $\mu_i \ge \mu_{i-1}$, a packet arriving at s_i is immediately served before the next packet arrival. Note that the case of $\mu_i > \mu_{i-1}$ can happen when extra resources are sliced and reserved on node i to accommodate short-term traffic load fluctuations for enhancing the E2E service performance. There is no packet queuing at s_i but only deterministic packet service delay $\frac{1}{\mu_i}$, and the packet departure process at s_i statistically remains the same as that at s_{i-1}. Whether a packet exceeds its service deadline when passing through s_i directly depends on its remaining deadline and the queue length when it arrives at s_{i-1}. If $\mu_i < \mu_{i-1}$, newly arrived packets at s_i may wait in the service queue of s_i when a packet is being served. In this case, both packet queuing and service delays exist.

We propose a segment-based analysis framework, as shown in Fig. 5. Given the allocated packet service rates of nodes along the routing path of the video streaming slice, the E2E multi-hop packet service is grouped into several sequential segments. We denote by μ^g the service rate of the first node in segment g. Segment g consists of a consecutive sequence of nodes within which the packet service rates of the rest of nodes except the first one are no less than the first node's packet service rate, i.e., μ^g. Besides, we have $\mu^{g+1} < \mu^g$ for two consecutive segments g and $g+1$. Let λ_b^g and λ_e^g be the average packet arrival rates of BL and EL traffic arriving at segment g, respectively, and the total packet arrival rate at segment g is thus given by $\lambda^g = \lambda_b^g + \lambda_e^g$. Note that λ^g may not always be equal to λ^{g+1} for two consecutive segments g and $g+1$ due to possible deadline-violated EL packet dropping. In addition, let $\bar{d}^{b,g}$ and $\bar{d}^{e,g}$ be the average (remaining) service deadlines of BL and EL packets arriving at segment g, respectively. Within each segment, packet service performance at the first node is critical to analyze since there are only deterministic packet service delays at the other nodes, which is discussed in the following.

B. Service Performance Modeling at The First Segment

Given the average service deadline of EL packets arriving at the first segment, $\bar{d}^{e,1}$, the corresponding queue length bound for EL packets that can successfully pass through the first node is given by $N_e^1 = \lceil \bar{d}^{e,1}\mu^1 \rceil$. When the queue length of the first node increases to the bound, N_e^1, arriving EL packets are dropped before passing through the first node due to service deadline violations. In that case, only BL packets with a higher service reliability requirement can pass through. The state transition diagram of queue length at the first node is shown in Fig. 6. It can be seen that the transition rate from state i to state $i+1$ is $\lambda_i = \lambda^1 = \lambda_b^1 + \lambda_e^1$ for $i < N_e^1$, as both arriving BL and EL packets can be served through the first node. The transition rate from state i to state $i+1$ is $\lambda_i = \lambda_b^1$ for $i \ge N_e^1$, as only arriving BL packets can pass through the first node while arriving EL packets will be discarded due to service deadline violations. Therefore, packet service at the first node is a special state-dependent M/D/1 queuing system where the packet arrival rate is state-dependent only at state $i = N_e^1$, and the packet service rate is not state-dependent and keeps the same for all states (i.e., μ^1). Let the steady-state probability distribution of queue length at the first node (at arbitrary times) be denoted by $\{q_i, i \ge 0\}$. Note that $\{q_i, i \ge 0\}$ is difficult to obtain, which involves complex Laplace Transforms (LTs) and recursive computations, and the related steady-state averages such as the mean sojourn time are challenging to derive [40]. Therefore, according to the state transition diagram shown in Fig. 6, we propose an approximate approach to obtain the

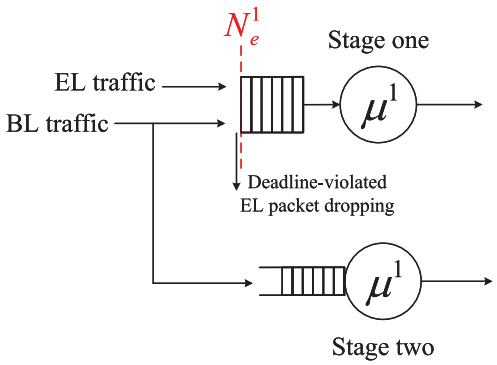


Fig. 7. Two-stage queuing at the first node of the first segment.

steady-state probability distribution of queue length at the first node, denoted by $\{\hat{q}_i, i \geq 0\}$.

Specifically, we consider that packet services at the first node go through two sequential stages, as shown in Fig. 7. At stage one when the queue length of the first node is smaller than N_e^1, both BL and EL packets can pass through the first node. At stage two when the queue length of the first node is larger than or equal to N_e^1, only arriving BL packets are admitted to the queue while an arriving EL packet is discarded before being served due to service deadline violation. The stage one queuing is modeled as an M/D/1/N_e^1 system with packet arrival rate of λ^1 and service rate of μ^1. The steady-state probability distribution of queue length, denoted by $P_j, j = 0, 1, \cdots, N_e^1$, is given by [41]

$$\begin{aligned} P_0 &= \frac{1}{1+\rho_1^1 b_{N_e^1-1}} \\ P_{N_e^1} &= 1 - \frac{b_{N_e^1-1}}{1+\rho_1^1 b_{N_e^1-1}} \\ P_j &= \frac{b_j - b_{j-1}}{1+\rho_1^1 b_{N_e^1-1}}, j = 1, \cdots, N_e^1 - 1 \end{aligned} \tag{3}$$

where $\rho_1^1 = \frac{\lambda^1}{\mu^1}$ is the traffic intensity for stage one queuing. The coefficients $\{b_n, n \geq 0\}$ are given by $b_n = \sum_{k=0}^{n} P_k\left((k-n)\frac{1}{\mu^1}\right)$ with $P_k(t) = \frac{(\lambda^1 t)^k}{k!}e^{-\lambda^1 t}$. In addition, the stage two queuing is modeled as an M/D/1 system with packet arrival rate of λ_b^1 and service rate of μ^1. The steady-state probability distribution of queue length, denoted by $\pi_n, n \geq 0$, is given by [42]

$$\begin{aligned} \pi_0 &= 1 - \rho_2^1 \\ \pi_1 &= \left(e^{\rho_2^1} - 1\right)\left(1 - \rho_2^1\right) \\ \pi_n &= \left(1 - \rho_2^1\right)\left[e^{n\rho_2^1} + \sum_{k=1}^{n-1}(-1)^{n-k}e^{k\rho_2^1}\right. \\ &\quad \left.\left(\frac{(k\rho_2^1)^{n-k}}{(n-k)!} + \frac{(k\rho_2^1)^{n-k-1}}{(n-k-1)!}\right)\right], n \geq 2 \end{aligned} \tag{4}$$

where $\rho_2^1 = \frac{\lambda_b^1}{\mu^1}$ is the traffic intensity for stage two queuing.

With Eqs. (3) - (4), the approximate steady-state probability distribution of queue length at the first node, i.e., $\{\hat{q}_i, i \geq 0\}$, is obtained as

$$\hat{q}_i = \begin{cases} P_i, i < N_e^1 \\ P_{N_e^1}\pi_{i-N_e^1}, i \geq N_e^1. \end{cases} \tag{5}$$

Remark 1: In our approximate approach, the state transitions from N_e^1+i to $N_e^1+(i+1)$ for $i \geq 0$ (see Fig. 6) are neglected, leading to approximation errors. When the packet service rate of the first node, μ^1, increases, such that the queue length is seldomly larger than N_e^1, more accurate approximations of $\{q_i, i \geq 0\}$ can be obtained. In addition, since the packet arrival process (or rate) is not state-dependent in each queuing stage, the Poisson Arrivals See Time Averages (PASTA) property holds for both the modeled M/D/1/N_e^1 system and the modeled M/D/1 system. The approximated steady-state probability distribution at arbitrary times, $\{\hat{q}_i, i \geq 0\}$, is also the approximated steady-state probability distribution at arrival epochs.

Next, we derive the average BL/EL packet delay, deadline violation probability, and throughput for traversing the first segment. Let $\tilde{T}^1$ denote the total service time for a packet passing through the rest of nodes except the first one in the first segment. If there are i packets (including the one that is in service) at the first node, then the time for the ith packet to traverse the first segment, denoted by T_i^1, is given by

$$T_i^1 = \begin{cases} 0, i = 0 \\ (i-1)\frac{1}{\mu^1} + \tilde{T}^1, i > 0 \end{cases} \tag{6}$$

where the expected residual service time of the packet in service when there are i packets in the system, denoted by $E[R_i]$, is neglected in the case $i > 0$. With the approximated steady-state probability distribution of queue length at the first node, $\{\hat{q}_i, i \geq 0\}$, there is a one-to-one mapping between $\hat{q}_i$ and T_i^1 for each $i \geq 0$.

The average BL/EL packet delay for traversing the first segment, denoted by $T_\theta^1, \theta \in \{b, e\}$, is given by

$$T_e^1 = \tilde{T}^1 + T^{M/D/1/N_e^1} + P_{N_e^1}\left(\left(N_e^1 - 1\right)\frac{1}{\mu^1}\right)(\pi_0 - 1) \tag{7}$$

and

$$T_b^1 = P_{N_e^1}T^{M/D/1} + T^{M/D/1/N_e^1} + \tilde{T}^1 \tag{8}$$

where $T^{M/D/1/N_e^1}$ and $T^{M/D/1}$ are the average packet delays of the considered $M/D/1/N_e^1$ system for stage one queuing and the considered $M/D/1$ system for stage two queuing, respectively, given by [43]

$$T^{M/D/1/N_e^1} = \frac{1}{\lambda^1}\frac{1+\rho_1^1 b_{N_e^1-1}}{b_{N_e^1-1}}\frac{N_e^1 + N_e^1\rho_1^1 b_{N_e^1-1} - \sum_{k=0}^{N_e^1-1} b_k}{1+\rho_1^1 b_{N_e^1-1}} \tag{9}$$

and

$$T^{M/D/1} = \left(1 + \frac{1}{2}\frac{\rho_2^1}{1-\rho_2^1}\right)\frac{1}{\mu^1}. \tag{10}$$

In terms of deadline violation probability, take EL packets as an example, we first find $j^e = \min_j\left\{T_j^1 > \bar{d}^{e,1}\right\}$. The queue length bound for EL packets at the first node that can successfully traverse the first segment is thus $j^e - 1$. Then, the

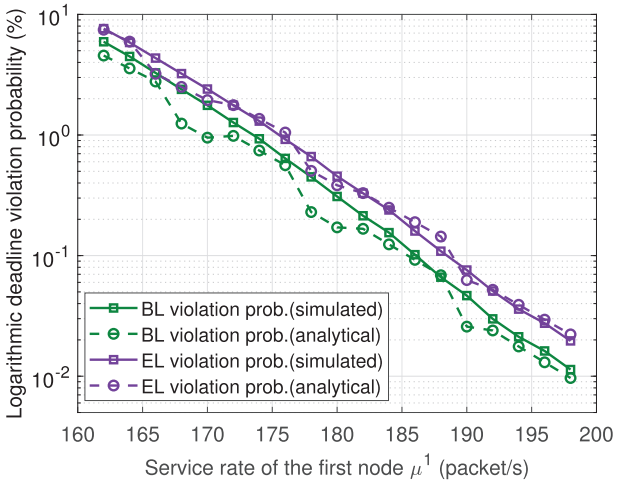


Fig. 8. Deadline violation probability (in log-scale) of BL and EL packets at the example segment with varying packet service rates of the first node (before approximations).

TABLE II
THE VALUES OF N_e^g AND j^e FOR DIFFERENT PACKET SERVICE RATES OF THE FIRST NODE μ^1 IN THE EXAMPLE SEGMENT

μ^1	162	164	166	168	170	172	174	176	178
N_e^1	17	17	17	17	17	18	18	18	18
j^e	15	15	16	16	16	16	16	16	17

deadline violation probability of EL packets when traversing the first segment, denoted by V_e^1, is calculated by

$$V_e^1 = V_{e,A}^1 + V_{e,B}^1 = \left(1 - \sum_{j=0}^{j^e-2} \hat{q}_j\right) + \hat{q}_{j^e-2} a^b \qquad (11)$$

where $a^b = \frac{\lambda_b^1}{\lambda_b^1+\lambda_e^1}$ is the probability of a BL packet arriving at the first segment before an EL packet. The EL packet deadline violation probability at the first segment, V_e^1, consists of two parts, i.e., $V_{e,A}^1$ and $V_{e,B}^1$, which represent the probability of queue length (of the first node) no less than $j^e - 1$ and the probability of observing $j^e - 2$ packets in the queue by an arriving BL packet, respectively. The deadline violation probability of BL packets when traversing the first segment, denoted by V_b^1, can be calculated similarly according to Eq. (11) with j^e and a^b replaced by $j^b = \min_j \left\{T_j^1 > \bar{d}^{b,1}\right\}$ and $a^e = \frac{\lambda_e^1}{\lambda_b^1+\lambda_e^1}$, respectively.

We take a segment consisting of four nodes with packet service rates $\left[\mu^1, 200, 200, 200\right]$ as an example. We set $\lambda_b^1 = 100$, $\lambda_e^1 = 50$, $\bar{d}^{b,1} = 105$ ms, and $\bar{d}^{e,1} = 100$ ms. The values of N_e^1 and j^e for different μ^1 are listed in Table II. We notice that, due to the $\lceil\cdot\rceil$ operator, N_e^1 does not change or increase as μ^1 increases, and the changing points of N_e^1 are different from those of j^e, which leads to the relative estimation error between the estimated V_e^1 by Eq. (11) and the real value oscillating (the same for V_b^1), as shown in Fig. 8. To address this and obtain a more accurate deadline violation probability estimate, we make some approximations based on the above observations. Specifically, let $\left(N_e^1, j^e\right)$ and $\left(N_e^{1'}, j^{e'}\right)$ be the tuples corresponding to two different packet service rates of the first node in the first segment, denoted by μ^1 and $\mu^{1'}$ where $\mu^{1'} = \mu^1 + \Delta_\mu$ (Δ_μ set as 2 packet/s in the simulations). Due to term $\tilde{T}^1$ in Eq. (6), intuitively, j^e should be smaller than N_e^1. Thus, if $j^e = N_e^1$, we let $j^e = j^e - 1$. Besides, if $j^{e'} - j^e > 0$ while $N_e^{1'} = N_e^1$, we let $j^{e'} = j^{e'} - 1$. The deadline violation probability estimations after the approximations are presented in the next section.

Finally, due to the disparate service reliability requirements, deadline-violated EL packets are directly discarded without being further transmitted. Therefore, the throughputs of BL and EL traffic after traversing the first segment, denoted by $\lambda_{t,b}^1$ and $\lambda_{t,e}^1$ respectively, are given by

$$\begin{aligned} \lambda_{t,b}^1 &= \lambda_b^1 \\ \lambda_{t,e}^1 &= \lambda_e^1 \left(1 - V_e^1\right). \end{aligned} \qquad (12)$$

Thus, the total throughput, λ_t^1, at the first segment is $\lambda_t^1 = \lambda_{t,b}^1 + \lambda_{t,e}^1$. In addition, the updated average service deadline of BL/EL packets arriving at the second segment is $\bar{d}^{\theta,2} = \left(\bar{d}^{\theta,1} - T_\theta^1\right)^+, \theta \in \{b, e\}$ where $(x)^+ = \max(0, x)$.

C. Service Performance Modeling Between Segments

Next, we analyze the packet departure process at the first segment (also the packet arrival process at the second segment), which is non-trivial due to two main reasons. First, the steady-state probability distribution of queue length seen by departing packets from stage one queuing of the first node, denoted by $q_m, m = 0, 1, \cdots, N_e^1 - 1$, differs from the steady-state probability distribution of queue length at arbitrary times, i.e., $P_j, j = 0, 1, \cdots, N_e^1$. Second, certain EL packets may exceed their service deadlines and be discarded when passing through the other nodes in the first segment rather than the first node. To address this, we adopt an approximate approach to analyze the packet departure process at the first segment. Specifically, from the viewpoint of departing packets from the first segment, stage one queuing of the first node can be approximated by an M/D/1 system with the packet arrival rate being λ_t^1 and service rate being μ^1, such that the probability of queue length of the first node no less than j^e is small and can be neglected. Then, according to the PASTA property of Poisson arrivals and the level crossing property [43], the steady-state probability distribution of queue length seen by departing packets in the approximated M/D/1 system is the same as that seen by arrival epochs or at arbitrary times.

Based on the approximate M/D/1 system, let random variable Y^1 be the inter-departure time of successive packets departing from the first segment. If a departing packet, k, sees a nonempty queue, then $Y^1 = \frac{1}{\mu^1}$. Otherwise, $Y^1 = X^1 + \frac{1}{\mu^1}$, where X^1 represents the time interval from the departure of packet k to the arrival of packet $k+1$. Due to the memoryless property of Poisson arrivals, X^1 follows the same exponential distribution as the packet inter-arrival time with parameter λ_t^1. Therefore, it is seen that packet departure from the first segment is a mixed process alternating between deterministic

and Poisson processes. The mean and variance of the packet inter-departure time, Y^1, are given by

$$E[Y^1] = \rho_t^1 \frac{1}{\mu^1} + \left(1-\rho_t^1\right) E\left[X^1 + \frac{1}{\mu^1}\right] = \frac{1}{\lambda_t^1}$$
$$D[Y^1] = E\left[\left(Y^1 - E[Y^1]\right)^2\right] = \frac{1}{(\lambda_t^1)^2} - \frac{1}{(\mu^1)^2} \quad (13)$$

where $\rho_t^1 = \frac{\lambda_t^1}{\mu^1}$. From Eq. (13), when the packet service rate of the first node, i.e., μ^1, is large, packet departures from the first segment approach to a Poisson process. Besides, when λ_t^1 increases to approach to μ^1, i.e., when the queuing system is heavily loaded, packet departures from the first segment approach to a deterministic process, which is aligned with the preceding analysis.

In order to achieve the independence between two consecutive segments for analysis tractability, under the assumption of a large packet service rate of the first node in the first segment, i.e., μ^1, we approximate the packet departures from the first segment as a Poisson process with rate parameter λ_t^1.

D. Average E2E Performance

Based on the proposed segment-based analysis framework, we derive the (average) BL/EL packet delay, deadline violation probability, and throughput at the first segment. The independence between two consecutive segments is achieved under the assumption of a large packet service rate of the first node in the preceding segment. For a video streaming slice in the core network which consists of N_s segments, the average E2E packet delay, T_θ, average E2E deadline violation probability, V_θ, and average E2E throughput, $\lambda_{t,\theta}$, of BL/EL packets are given by

$$T_\theta = \sum_{g=1}^{N_s} T_\theta^g, \theta \in \{b, e\}$$
$$V_\theta = 1 - \prod_{g=1}^{N_s} (1 - V_\theta^g), \theta \in \{b, e\}$$
$$\lambda_{t,\theta} = \begin{cases} \lambda_b^1, \theta = b \\ \lambda_e^1 (1 - V_e), \theta = e \end{cases} \quad (14)$$

where T_θ^g and V_θ^g denote the average BL/EL packet delay and deadline violation probability at the gth segment, respectively.[5]

The E2E performance metrics given in Eq. (14) depend only on the number of segments composing a video streaming slice, i.e., N_s, which incurs a time complexity of $O(N_s)$.

E. Discussion

In the E2E performance analytical modeling, the key feature of layer-encoded video streaming, i.e., the distinct service reliability requirements between BL and EL packets, is captured. However, the inter-dependence between different ELs, that is, an EL can only be decoded if the corresponding BL and all lower-layer ELs are successfully decoded, is not explicitly considered. In that case, EL packets of multiple ELs may also have different levels of service reliability requirements. If any EL packet of an EL is discarded due to deadline violation, the EL packets of higher ELs should also be discarded for transmission efficiency.

[5]Note that when the number of segments composing a video streaming slice N_s increases, the cumulative effect of successive approximations of multiple segments in the E2E performance analytical modeling can lead to larger accuracy gaps.

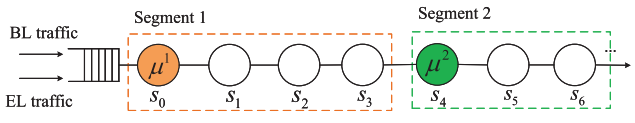


Fig. 9. The considered virtual network topology of a video streaming slice in the simulation.

The proposed analytical model can be used for performance evaluation of a video streaming slice. Given the characteristics of aggregated BL and EL traffic flows (i.e., packet arrival rates λ_b and λ_e) and the slice resource configuration (i.e., the reserved packet service rates of nodes within the video streaming slice), the following key steps should be followed:

- *Step 1*: Determine the segmentation of the video streaming slice according to the slice resource configuration, as discussed in Subsection IV-A;
- *Step 2*: Calculate the segment-level performance metrics for each segment based on Eqs. (7), (8), (11) and (12);
- *Step 3*: Calculate the E2E performance metrics based on Eq. (14).

The proposed analytical model can also be helpful for slice resource reservation. Given the characteristics of aggregated BL and EL traffic flows and the specific service requirements in terms of average packet delay and/or deadline violation probability, a network operator (or slice orchestrator) can configure slice resource starting from the case where there is only one segment composing the video streaming slice, and utilize the proposed analytical model to evaluate the E2E performances until the service requirements are met.

In the next section, two examples are presented to show the effectiveness of the proposed E2E analytical model for practical applications.

V. Simulation Results

In this section, simulation results are presented to verify the accuracy of our proposed performance analytical model for E2E layer-encoded video packet services over a video streaming slice. The considered virtual network topology is shown in Fig. 9 which consists of two segments. Within each segment, each node is either a processing node or a transmission node with a deterministic packet service rate. In addition, aggregated BL and EL traffic flows at the ingress node (s_0) are considered as two independent Poisson processes with packet arrival rates of $\lambda_b^1 = 100$ and $\lambda_e^1 = 50$ packet/s [16]. The average service deadlines of BL and EL packets at the first segment are set as $\bar{d}^{b,1} = 105$ ms and $\bar{d}^{e,1} = 100$ ms [44], [45], [46], [47]. All simulations are carried out with SimPy 4.0.1 and Python 3.10.10.[6] In the evaluation, we validate the

[6]Please refer to https://simpy.readthedocs.io/en/latest/index.html for more details.

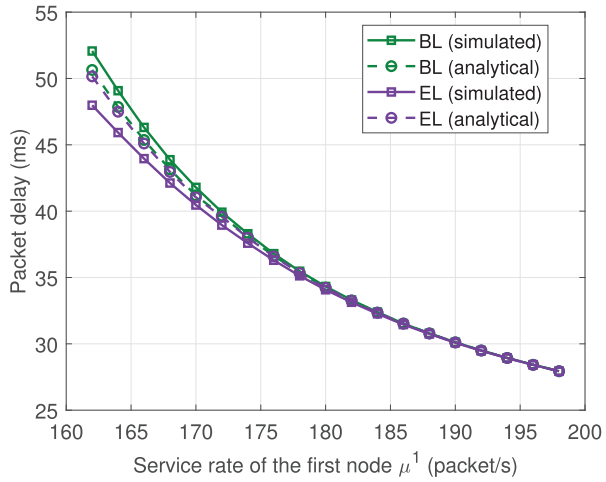


Fig. 10. Packet delay at the first segment with varying packet service rates of the first node μ^1.

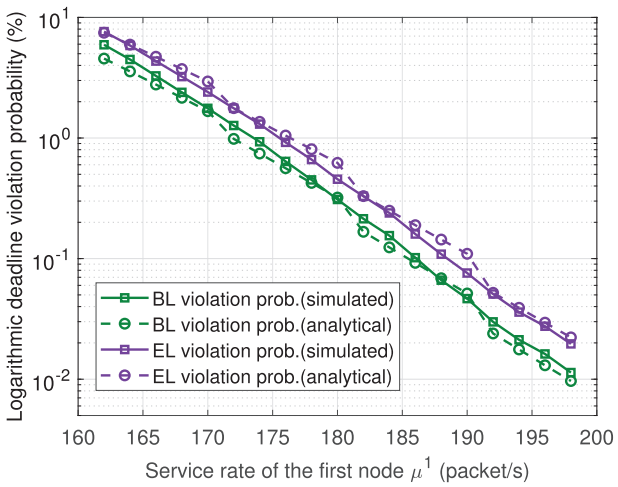


Fig. 11. Deadline violation probability at the first segment with varying packet service rates of the first node μ^1.

proposed performance analytical model on a segment basis. For each segment, we vary the packet service rate of the first node (e.g., μ^1 or μ^2) to compare the performances under different traffic load conditions. Then, the accuracy of E2E performance analysis is verified. In addition, the effectiveness of the proposed performance analytical model is shown with two application examples, where M/M/1 and M/D/1, two widely used mathematical models in existing works [34], [35], are adopted as benchmarks.

A. Validation of The Proposed Performance Analytical Model

Fig. 10 shows the (average) BL and EL packet delays at the first segment with varying μ^1 where the packet service rates of the nodes in the first segment are $[\mu^1, 200, 200, 200]$ in packet/s. We can see that close estimations of BL and EL packet delays are achieved, especially when μ^1 is large. For small values of μ^1, the queuing system of the first node is heavily loaded, and there are more BL packets on stage two queuing, leading to larger BL packet delay than EL packet delay. For a large μ^1, packet services at the first node are more closely seen as an M/D/1/N_e^1 system with a large value of N_e^1 and a very small value of $P_{N_e^1}$. Thus, BL and EL packets have a close packet delay, and more accurate packet delay estimations are achieved (see Eq. (7)). The deadline violation probabilities (in log-scale) of BL and EL packets at the first segment with varying μ^1 are shown in Fig. 11. It can be seen that the proposed performance analytical model achieves close BL/EL packet deadline violation probability estimation, even for large values of μ^1. Besides, compared to Fig. 8, the relative estimation errors between the analytical results and the simulation results are reduced, which proves the effectiveness of the approximations made in Subsection IV-B.[7] The throughputs of BL and EL packets at the first segment with varying μ^1 are shown in Fig. 12. We see the analytical

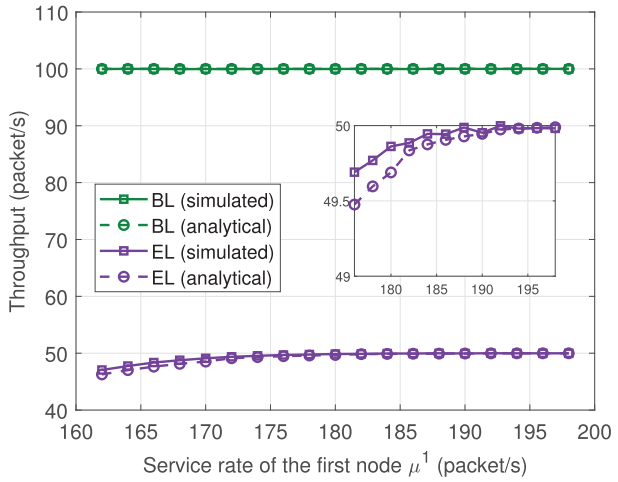


Fig. 12. Throughput at the first segment with varying packet service rates of the first node μ^1.

results match with the simulation results well, and the estimation error reduces as μ^1 increases due to a smaller V_e^1. In addition, due to the high service reliability requirement, BL packets are reliably transmitted, and thus the BL packet throughput equals to its packet arrival rate (i.e., $\lambda_{t,b}^1 = \lambda_b^1$). On the other hand, due to the low service reliability requirement, deadline-violated EL packets that do not improve video quality are discarded. Thus, the EL packet throughput is no greater than its arrival rate (i.e., $\lambda_{t,e}^1 \leq \lambda_e^1$). The results presented in Fig. 10 - Fig. 12 verify the effectiveness of the established two-stage queuing in Subsection IV-B as a service modeling approximation of the first node, and also the accuracy of our proposed performance analytical model.

Furthermore, the BL/EL packet throughput at the first segment with varying BL packet arrival rate (λ_b^1) is shown in Fig. 13. The packet service rates of the nodes in the first segment are set as $[180, 200, 200, 200]$ packet/s. We can see that due to the high service reliability requirement, the BL packet throughput increases linearly with its arrival rate. On the other hand, as λ_b^1 increases, the inter-arrival time of

[7]Please note that the non-smoothness of analytical results (in Figs. 11, 15, and 19) is caused by the $\lceil \cdot \rceil$ operator in determining the queue length bound N_e^1 for stage one queuing at the first node and the approximations we made for obtaining a more accurate deadline violation probability estimate (see the explanations for Fig. 8).

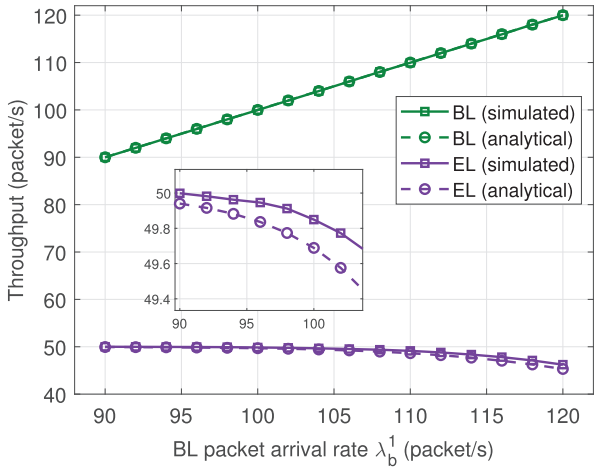


Fig. 13. Throughput at the first segment with varying BL packet arrival rates λ_b^1.

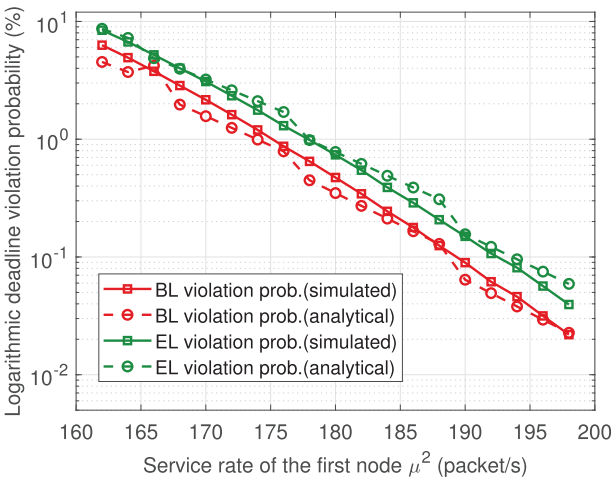


Fig. 15. Deadline violation probability at the second segment with varying packet service rates of the first node μ^2.

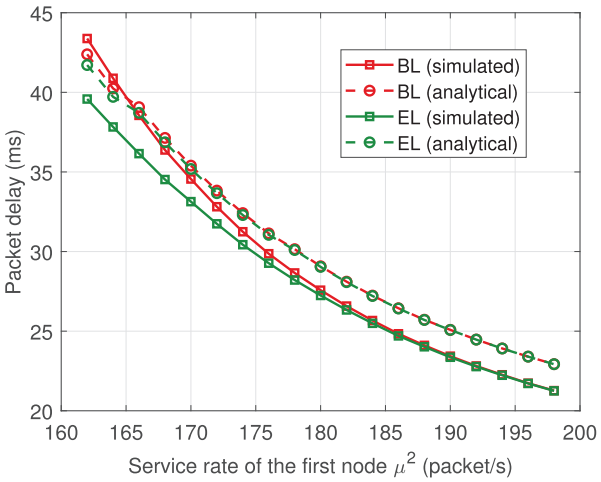


Fig. 14. Packet delay at the second segment with varying packet service rates of the first node μ^2.

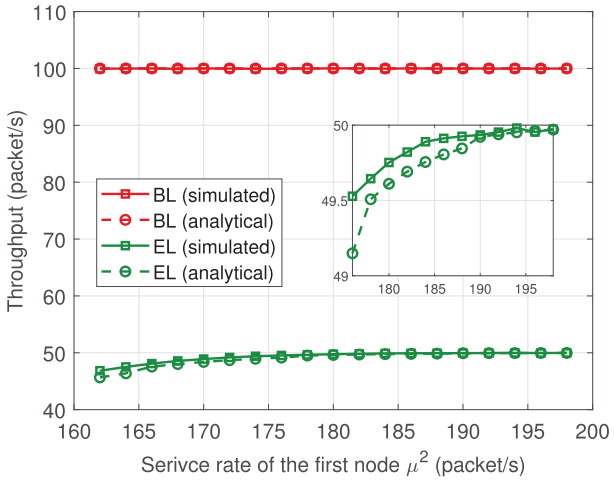


Fig. 16. Throughput at the second segment with varying packet service rates of the first node μ^2.

successive BL packets is smaller. More BL packets enter stage one queuing while the queue length bound, N_e^1, for EL packets is unchanged. Less EL packets can be served through the first node without exceeding their service deadlines. Hence, the EL packet throughput decreases with λ_b^1.

The packet delays, deadline violation probabilities (in log-scale), and throughputs of BL and EL packets at the second segment with varying μ^2 are demonstrated in Fig. 14 - Fig. 16, respectively. The packet service rates of the nodes in the first segment are set as $[300, 300, 300, 300]$ packet/s while the service rates of the nodes in the second segment are $[\mu^2, 200, 200]$ packet/s. We can see that due to smaller packet queuing and service delays, the BL/EL packet delay increases with μ^2, and the BL/EL packet deadline violation probability decreases with μ^2. Besides, the BL packet throughput maintains the same as its arrival rate (i.e., $\lambda_{t,b}^2 = \lambda_b^2 = \lambda_b^1$) due to the high service reliability requirement, while the EL packet throughput gradually approaches to its arrival rate (λ_e^2) as μ^2 increases due to a smaller V_e^2. The close estimations of BL/EL packet delay, deadline violation probability, and throughput using our proposed performance analytical model compared to the simulation results validate the independence approximation between two consecutive segments on the premise of a large μ^1. As discussed in Subsection IV-C, when μ^1 is large enough, the packet departure process at the first segment (or the packet arrival process at the second segment) is approximated as a Poisson process. Thus, the simulation results obtained for the second segment are similar to those for the first segment. In addition, the BL/EL packet throughput at the second segment with varying BL packet arrival rate (λ_b^1) is shown in Fig. 17, where the packet service rates of the nodes in the second segment are set as $[170, 200, 200]$ packet/s. Similar results as in Fig. 13 are observed where the BL packet throughput increases linearly with its arrival rate, while the EL packet throughput decreases with λ_b^1, as more BL packets join stage one queuing which has a finite and fixed queue length bound, i.e., N_e^2.

Finally, the performances of E2E packet delays, deadline violation probabilities (in log-scale), and throughputs of BL and EL packets with varying μ^2 are evaluated in Fig. 18 - Fig. 20, respectively. The packet service rates of the nodes composing the considered virtual network topology (see Fig. 9) of the video streaming slice are set as

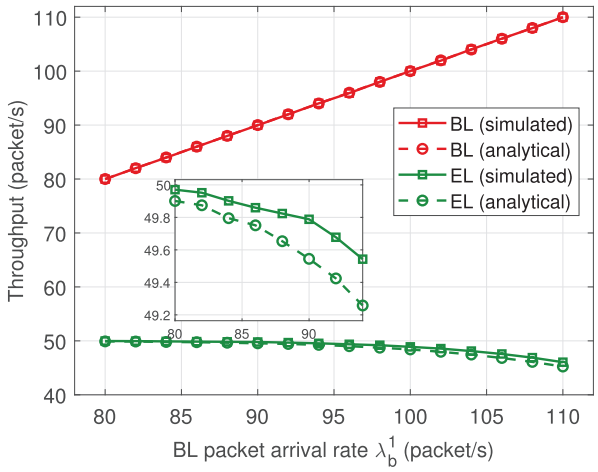


Fig. 17. Throughput at the second segment with varying BL packet arrival rates λ_b^1.

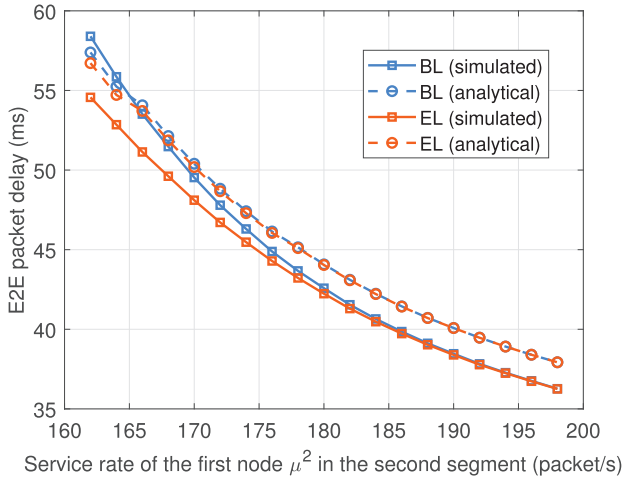


Fig. 18. E2E packet delay with varying packet service rates of the first node in the second segment μ^2.

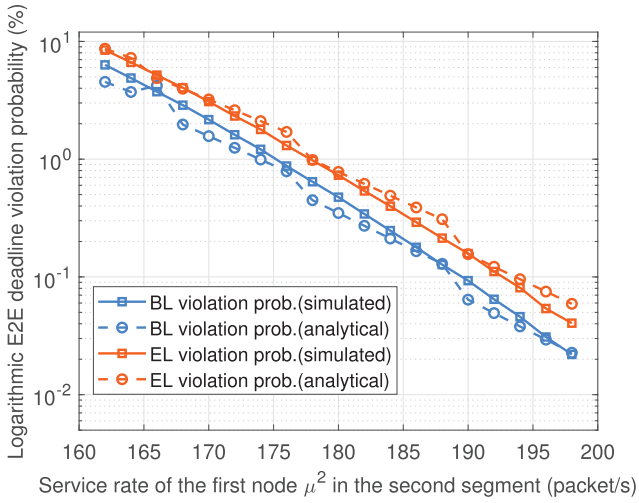


Fig. 19. E2E deadline violation probability with varying packet service rates of the first node in the second segment μ^2.

$[300, 300, 300, 300, \mu^2, 200, 200]$ in packet/s. The close analytical results compared to the simulation results indicate the accuracy of our proposed performance analytical model which bases on the segment-based analysis framework in

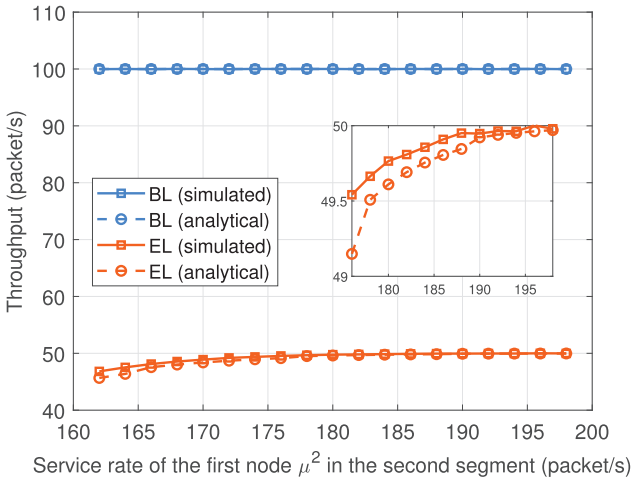


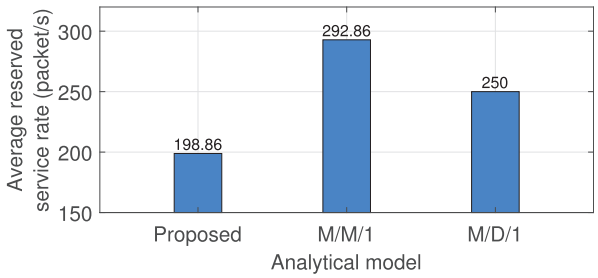
Fig. 20. E2E throughput with varying packet service rates of the first node in the second segment μ^2.

Subsection IV-A for slice-based E2E video packet service with layered encoding.

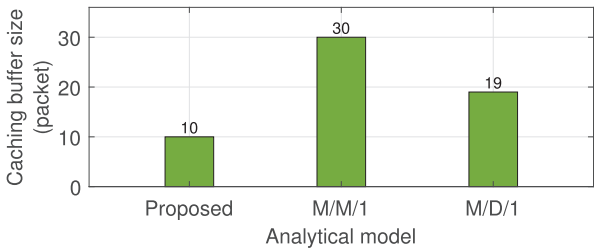
B. Effectiveness of The Proposed Performance Analytical Model

In the following, the effectiveness of the proposed performance analytical model is shown with two application examples, where M/M/1 and M/D/1 are selected as the benchmark models. First, the proposed performance analytical model can be applied for slice resource reservation. We consider the virtual network topology in Fig. 9 for layer-encoded video streaming service and suppose a target E2E delay of 49 ms for both BL and EL packets. For the proposed performance analytical model, we assume that the virtual network topology consists of only one segment for simplicity. A feasible node packet service rate configuration to achieve the target BL/EL delay is $[180, 212, 200, 200, 200, 200, 200]$ packet/s. In the benchmark models, packet service at each node is considered independent, and the differentiated service reliability requirements between BL and EL packets are neglected. The target E2E BL/EL delay is equally divided to each node, i.e., a per-node delay target of 7 ms, based on which the packet service rate reserved on each node is determined accordingly. The average reserved packet service rate on each node in the considered virtual network topology is shown in Fig. 21(a). We can see that due to accurate estimations of BL and EL packet delays using our proposed performance analytical model, fewer resources (i.e., packet service rate) are required for each node to achieve the target E2E packet delay, compared to the benchmark models.

In addition, the proposed performance analytical model can be utilized to determine the key transport parameters in the customized caching-based packet retransmission functionality for supporting enhanced video packet delivery, including the cycle of the periodic timer and the minimum required caching buffer size for time-based caching buffer release. Specifically, we set the packet service rates of the nodes in the considered virtual network topology as $[180, 212, 200, 200, 200, 200, 200]$ packet/s and consider only BL packets for an example. The



(a) Average reserved packet service rate on each node.



(b) Minimum required caching buffer size for BL packets.

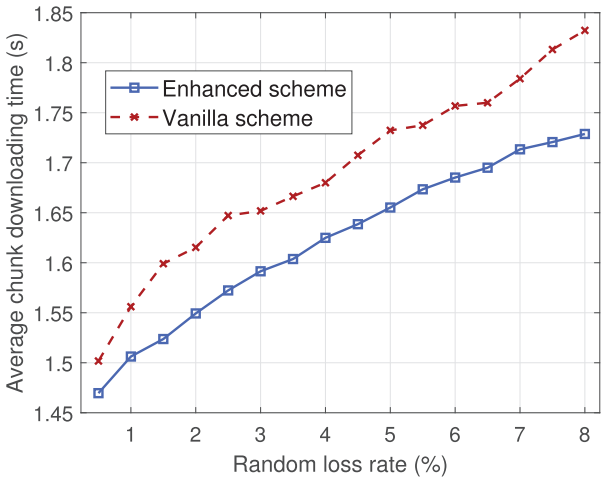
Fig. 21. Average reserved per-node packet service rate and the minimum required caching buffer size determined using different analytical models.

TABLE III
THE PERFORMANCE OF THE CACHING-BASED PACKET RETRANSMISSION SCHEME

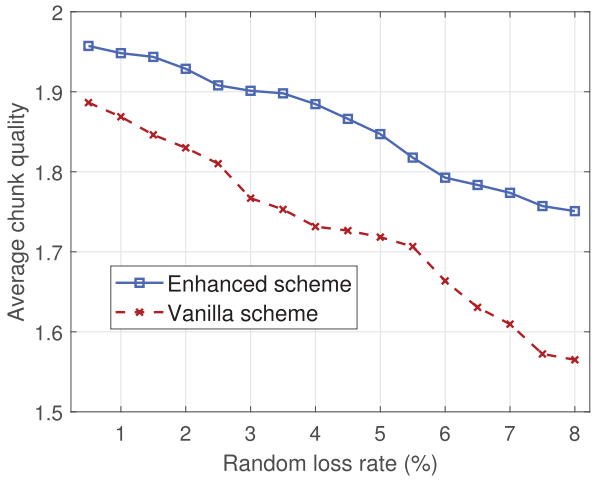
Caching miss ratio (%)	1.4597
Caching buffer utilization (%)	82.0005
Packet retransmission delay (ms)	49.2115
Optimal packet retransmission delay (ms)	49.0655

packet delay between a remote video server on the Internet and the ingress node is set to 10 ms. The cycle of the periodic timer is set by the estimated E2E BL packet delay, and the minimum required caching buffer size for BL packets is determined according to Eq. (1). Fig. 21(b) shows the minimum required caching buffer sizes (for BL packets) calculated based on different analytical models. The performance of the caching-based packet retransmission scheme with the caching buffer size set based on our proposed performance analytical model is given in Table III. It is seen that compared to the benchmark models, fewer caching resources are required with our proposed performance analytical model due to more accurate E2E BL packet delay estimation. Meanwhile, the proposed caching-based packet retransmission scheme performs well and achieves low caching miss ratio caused by '*release-before-arrival*' BL packets, high caching buffer utilization, and near-optimal packet retransmission delay, compared to the optimal case when the caching miss ratio is 0.

Moreover, we consider a use case of tile-based 360° video streaming to examine the outcomes of the proposed caching-based packet retransmission scheme and deadline-violated EL packet dropping without retransmissions (*Enhanced scheme*) on video streaming performance in terms of average chunk downloading time and quality. Chunk downloading time refers to the time to receive all the BL tiles of a chunk, indicating the time when a chunk is ready to play with basic quality.



(a) Chunk downloading time



(b) Chunk quality

Fig. 22. Average chunk downloading time and quality vs. random packet loss rate.

Chunk quality is measured by the ratio of the number of (expected) EL tiles timely delivered to the total number of EL tiles requested for a chunk, given that the corresponding BL tiles have already been received. For the benchmark scheme (*Vanilla scheme*), deadline-violated EL packets are not discarded during packet transmissions, which does not improve the chunk quality. Besides, random BL/EL packet losses are retransmitted from a remote server on the Internet to a user. As shown in Fig. 22, as the random (link) loss rate increases, the average chunk downloading time increases and the average chunk quality decreases, while our proposed enhanced scheme outperforms the benchmark scheme. As the random loss rate increases, more BL/EL packets are lost and re-injected into the video streaming slice for retransmissions. Retransmitted BL packets result in increasing chunk downloading time, while retransmitted EL packets degrade chunk quality due to deadline violations. With the proposed caching-based packet retransmission scheme, randomly lost BL/EL packets are retransmitted by the ingress node using the cached packet copies in the caching buffer, instead of by the remote server. Thus, a smaller packet retransmission

delay is achieved. In addition, deadline-violated EL packets are directly discarded without being further transmitted, including the retransmitted ones. In contrast, in the Vanilla scheme, all lost BL/EL packets including the deadline-violated ones are retransmitted from the remote server. Hence, the proposed enhanced scheme brings better streaming performance than the Vanilla scheme.

VI. Conclusion

In this paper, we have developed an E2E performance analytical model for layer-encoded video service over a core network slice. A segment-based E2E analysis framework is proposed, where a two-stage queuing model is established to determine per-segment service performance with the consideration of discrepant service reliability requirements of BL and EL packets. The independence between two consecutive segments is achieved, based on which the E2E service performance is determined, including the E2E BL/EL packet delay, deadline violation probability, and throughput. Simulation results demonstrate the proposed analytical model outperforms the benchmarks as well as its effectiveness in determining the sliced node service rates and transport-layer protocol parameters. For future work, we plan to extend the proposed analytical model to incorporate the impact of decoding dependency among multiple ELs. The impact of traffic load dynamics caused by other service traffic traversing a common physical path will also be considered. Moreover, the proposed analytical model will be extended to include a more general case where preemptive services of BL/EL packets may occur.

Acknowledgment

The work of Kaige Qu was completed when she was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Canada.

References

[1] S. Lu et al., "Integrated sensing and communications: Recent advances and ten open challenges," *IEEE Internet Things J.*, vol. 11, no. 11, pp. 19094–19120, Jun. 2024.

[2] K. Qu, W. Zhuang, Q. Ye, W. Wu, and X. Shen, "Model-assisted learning for adaptive cooperative perception of connected autonomous vehicles," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 8820–8835, Aug. 2024.

[3] T. Taleb et al., "Toward supporting XR services: Architecture and enablers," *IEEE Internet Things J.*, vol. 10, no. 4, pp. 3567–3586, Feb. 2023.

[4] Ericsson. (Nov. 2023). *Ericsson Mobility Report*. Accessed: May 20, 2024. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports

[5] R. Bhattacharyya et al., "QFlow: A learning approach to high QoE video streaming at the wireless edge," *IEEE/ACM Trans. Netw.*, vol. 30, no. 1, pp. 32–46, Feb. 2022.

[6] L. De Cicco and S. Mascolo, "An adaptive video streaming control system: Modeling, validation, and performance evaluation," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 526–539, Apr. 2014.

[7] L. Rossi, J. Chakareski, P. Frossard, and S. Colonnese, "A Poisson hidden Markov model for multiview video traffic," *IEEE/ACM Trans. Netw.*, vol. 23, no. 2, pp. 547–558, Apr. 2015.

[8] J. Van Der Hooft et al., "A tutorial on immersive video delivery: From omnidirectional video to holography," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 2, pp. 1336–1375, 2nd Quart., 2023.

[9] L. Yu, "SVC-based dynamic caching for smart media streaming over the Internet of Things," *Future Gener. Comput. Syst.*, vol. 114, pp. 219–228, Jan. 2021.

[10] A. Elgabli, V. Aggarwal, S. Hao, F. Qian, and S. Sen, "LBP: Robust rate adaptation algorithm for SVC video streaming," *IEEE/ACM Trans. Netw.*, vol. 26, no. 4, pp. 1633–1645, Aug. 2018.

[11] X. Zhang, X. Hu, L. Zhong, S. Shirmohammadi, and L. Zhang, "Cooperative tile-based 360° panoramic streaming in heterogeneous networks using scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 217–231, Jan. 2020.

[12] J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable extensions of the high efficiency video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 20–34, Jan. 2016.

[13] Y. Xu et al., "RAPID: Avoiding TCP incast throughput collapse in public clouds with intelligent packet discarding," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 8, pp. 1911–1923, Aug. 2019.

[14] S. Yan, Q. Ye, and W. Zhuang, "Learning-based transmission protocol customization for VoD streaming in cybertwin-enabled next-generation core networks," *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16326–16336, Nov. 2021.

[15] M. A. Habibi et al., "Toward an open, intelligent, and end-to-end architectural framework for network slicing in 6G communication systems," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1615–1658, 2023.

[16] Q. Ye, W. Shi, K. Qu, H. He, W. Zhuang, and X. Shen, "Joint RAN slicing and computation offloading for autonomous vehicular networks: A learning-assisted hierarchical approach," *IEEE Open J. Veh. Technol.*, vol. 2, pp. 272–288, 2021.

[17] Q. Ye, J. Li, K. Qu, W. Zhuang, X. S. Shen, and X. Li, "End-to-end quality of service in 5G networks: Examining the effectiveness of a network slicing framework," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 65–74, Jun. 2018.

[18] A. A. Abdellatif, A. Abo-Eleneen, A. Mohamed, A. Erbad, N. V. Navkar, and M. Guizani, "Intelligent-slicing: An AI-assisted network slicing framework for 5G-and-beyond networks," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 2, pp. 1024–1039, Jun. 2023.

[19] H. Li et al., "Slice-based service function chain embedding for end-to-end network slice deployment," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 3, pp. 3652–3672, Sep. 2023.

[20] M. Dai, G. Sun, H. Yu, and D. Niyato, "Maximize the long-term average revenue of network slice provider via admission control among heterogeneous slices," *IEEE/ACM Trans. Netw.*, vol. 32, no. 1, pp. 745–760, Feb. 2024.

[21] J. Chen et al., "SDATP: An SDN-based traffic-adaptive and service-oriented transmission protocol," *IEEE Trans. Cognit. Commun. Netw.*, vol. 6, no. 2, pp. 756–770, Jun. 2020.

[22] Y. Wei, Q. J. Ye, K. Qu, W. Zhuang, and X. S. Shen, "Transmission protocol customization for on-demand tile-based 360° VR video streaming," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2024, pp. 1069–1074.

[23] M.-H. Wang, L.-W. Chen, P.-W. Chi, and C.-L. Lei, "SDUDP: A reliable UDP-based transmission protocol over SDN," *IEEE Access*, vol. 5, pp. 5904–5916, 2017.

[24] H. Wang, Y. Wu, G. Min, and W. Miao, "A graph neural network-based digital twin for network slicing management," *IEEE Trans. Ind. Informat.*, vol. 18, no. 2, pp. 1367–1376, Feb. 2022.

[25] Y. Wei, Q. Ye, K. Qu, W. Zhuang, and X. Shen, "Customized transmission protocol for tile-based 360° VR video streaming over core network slices," *IEEE Trans. Netw.*, vol. 33, no. 1, pp. 340–354, Feb. 2025.

[26] O. Adamuz-Hinojosa, V. Sciancalepore, P. Ameigeiras, J. M. Lopez-Soler, and X. Costa-Pérez, "A stochastic network calculus (SNC)-based model for planning B5G uRLLC RAN slices," *IEEE Trans. Wireless Commun.*, vol. 22, no. 2, pp. 1250–1265, Feb. 2023.

[27] S. A. Hashemian and F. Ashtiani, "Analytical modeling and improvement of interference-coupled RAN slicing," *IEEE Trans. Mobile Comput.*, vol. 23, no. 12, pp. 13472–13486, Dec. 2024.

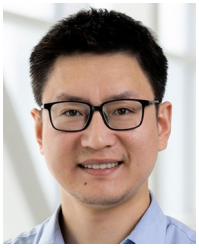
[28] K. Qu, W. Zhuang, X. Shen, X. Li, and J. Rao, "Dynamic resource scaling for VNF over nonstationary traffic: A learning approach," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 2, pp. 648–662, Jun. 2021.

[29] Q. Xu, J. Wang, and K. Wu, "Learning-based dynamic resource provisioning for network slicing with ensured end-to-end performance bound," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 28–41, Jan. 2020.

[30] D. Ayepah-Mensah, G. Sun, G. O. Boateng, S. Anokye, and G. Liu, "Blockchain-enabled federated learning-based resource allocation and trading for network slicing in 5G," *IEEE/ACM Trans. Netw.*, vol. 32, no. 1, pp. 654–669, Feb. 2024.

[31] I. Afolabi, J. Prados-Garzon, M. Bagaa, T. Taleb, and P. Ameigeiras, "Dynamic resource provisioning of a scalable E2E network slicing orchestration system," *IEEE Trans. Mobile Comput.*, vol. 19, no. 11, pp. 2594–2608, Nov. 2020.
[32] H. H. Esmat and B. Lorenzo, "Self-learning multi-mode slicing mechanism for dynamic network architectures," *IEEE/ACM Trans. Netw.*, vol. 32, no. 2, pp. 1048–1063, Apr. 2024.
[33] J. Li, W. Shi, Q. Ye, N. Zhang, W. Zhuang, and X. Shen, "Multiservice function chain embedding with delay guarantee: A game-theoretical approach," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11219–11232, Jul. 2021.
[34] R. Gouareb, V. Friderikos, and A.-H. Aghvami, "Virtual network functions routing and placement for edge cloud latency minimization," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2346–2357, Oct. 2018.
[35] Q. Ye, W. Zhuang, X. Li, and J. Rao, "End-to-end delay modeling for embedded VNF chains in 5G core networks," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 692–704, Feb. 2019.
[36] S. Kamath, S. Singh, and M. S. Kumar, "Multiclass queueing network modeling and traffic flow analysis for SDN-enabled mobile core networks with network slicing," *IEEE Access*, vol. 8, pp. 417–430, 2020.
[37] Z. Wang, J. Zhang, and T. Huang, "Determining delay bounds for a chain of virtual network functions using network calculus," *IEEE Commun. Lett.*, vol. 25, no. 8, pp. 2550–2553, Aug. 2021.
[38] S. Ma, X. Chen, Z. Li, and Y. Chen, "Performance evaluation of URLLC in 5G based on stochastic network calculus," *Mobile Netw. Appl.*, vol. 26, no. 3, pp. 1182–1194, Jun. 2021.
[39] M. Lecci, M. Drago, A. Zanella, and M. Zorzi, "An open framework for analyzing and modeling XR network traffic," *IEEE Access*, vol. 9, pp. 129782–129795, 2021.
[40] H. Abouee-Mehrizi and O. Baron, "State-dependent M/G/1 queueing systems," *Queueing Syst.*, vol. 82, nos. 1–2, pp. 121–148, Feb. 2016.
[41] O. Brun and J.-M. Garcia, "Analytical solution of finite capacity M/D/1 queues," *J. Appl. Probab.*, vol. 37, no. 4, pp. 1092–1098, Dec. 2000.
[42] K. Nakagawa, "On the series expansion for the stationary probabilities of an M/D/1 queue," *J. Oper. Res. Soc. Jpn.*, vol. 48, no. 2, pp. 111–122, 2005.
[43] D. Bertsekas and R. Gallager, *Data Networks*. Belmont, MA, USA: Athena Scientific, 2021.
[44] L. Sun et al., "A two-tier system for on-demand streaming of 360 degree video over dynamic networks," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 1, pp. 43–57, Mar. 2019.
[45] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman, "BOLA: Near-optimal bitrate adaptation for online videos," *IEEE/ACM Trans. Netw.*, vol. 28, no. 4, pp. 1698–1711, Aug. 2020.
[46] B. Cheng, M. Wang, X. Lin, and J. Chen, "Context-aware cognitive QoS management for networking video transmission," *IEEE/ACM Trans. Netw.*, vol. 29, no. 3, pp. 1422–1434, Jun. 2021.
[47] O. S. Peñaherrera-Pulla, C. Baena, S. Fortes, E. Baena, and R. Barco, "KQI assessment of VR services: A case study on 360-video over 4G and 5G," *IEEE Trans. Netw. Service Manage.*, vol. 19, no. 4, pp. 5366–5382, Dec. 2022.

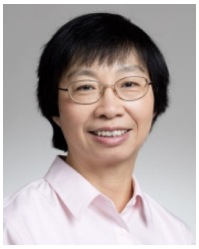
Yannan Wei (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. His research interests include resource management and service provisioning in future telecommunication systems.



Qiang (John) Ye (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, ON, Canada, in 2016.

Since 2023, he has been an Assistant Professor with the Department of Electrical and Software Engineering, University of Calgary, AB, Canada. He received the Best Paper Award in the IEEE ICCC in 2024 and the IEEE TCCN Exemplary Editor Award in 2023. He received the Early Career Research Excellence Award, Schulich School of Engineering, University of Calgary, in 2024. He is/was the general, publication, publicity, TPC, or symposium co-chair for different reputable international conferences and workshops. He has been serving/served as the IEEE VTS Region 7 Chapter Coordinator since 2024, the IEEE ComSoc Southern Alberta Chapter Vice Chair since 2024, and the VTS Regions 1-7 Chapters Coordinator (2022–2023). He serves as an Associate Editor for prestigious IEEE journals, such as IEEE INTERNET OF THINGS JOURNAL, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, and IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY. He has been selected as an IEEE ComSoc Distinguished Lecturer for the class of 2025–2026.

Kaige Qu (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2021. From February 2021 to December 2023, she was a Post-Doctoral Fellow and then a Research Associate with the Department of Electrical and Computer Engineering, University of Waterloo. Her research interests include connected and autonomous vehicles, and network intelligence.



Weihua Zhuang (Fellow, IEEE) received the B.Sc. and M.Sc. degrees from Dalian Marine University, Dalian, China, and the Ph.D. degree from the University of New Brunswick, Canada, all in electrical engineering. She is the University Professor and the University Research Chair of Wireless Communication Networks at the University of Waterloo, Canada. Her research interests include network architecture, algorithms, and protocols, and service provisioning in future communication systems. She is an Elected Member of the Board of Governors and the Past President of the IEEE Vehicular Technology Society. She is a Fellow of the Royal Society of Canada, Canadian Academy of Engineering, and the Engineering Institute of Canada.

Xuemin (Sherman) Shen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Rutgers University, New Brunswick, NJ, USA, in 1990.

He is currently the University Professor of the Department of Electrical and Computer Engineering, University of Waterloo, Canada. His research interests include network resource management, wireless network security, the Internet of Things, AI for networks, and vehicular networks. He is a registered Professional Engineer of Ontario, Canada, an Engineering Institute of Canada Fellow, a Canadian Academy of Engineering Fellow, a Royal Society of Canada Fellow, a Chinese Academy of Engineering Foreign Member, an International Fellow of the Engineering Academy of Japan, and a Distinguished Lecturer of the IEEE Vehicular Technology Society and Communications Society. He is the Past President of the IEEE Communications Society. He was the Vice President for Technical and Educational Activities, the Vice President for Publications, a Member-at-Large on the Board of Governors, the Chair of the Distinguished Lecturer Selection Committee, and a member of the IEEE Fellow Selection Committee of the ComSoc.