Statistical QoS Provisionings for Wireless Unicast/Multicast of Multi-Layer Video Streams

Qinghe Du, *Student Member, IEEE*, and Xi Zhang, *Senior Member, IEEE*

***Abstract*—Due to the time-varying wireless channels, deterministic quality of service (QoS) is usually difficult to guarantee for real-time *multi-layer* video transmissions in wireless networks. Consequently, *statistical* QoS guarantees have become an important alternative in supporting real-time video transmissions. In this paper, we propose an efficient framework to model the statistical delay QoS guarantees, in terms of QoS exponent, effective bandwidth/capacity, and delay-bound violation probability, for multi-layer video transmissions over wireless fading channels. In particular, a separate queue is maintained for each video layer, and the same delay bound and corresponding violation probability threshold are set up for all layers. Applying the effective bandwidth/capacity analyses on the incoming video stream, we obtain a set of *QoS exponents* for all video layers to effectively characterize this delay QoS requirement. We then develop a set of optimal adaptive transmission schemes to minimize the resource consumption while satisfying the diverse QoS requirements under various scenarios, including video unicast/multicast with and/or without loss tolerance. Simulation results are also presented to demonstrate the impact of statistical QoS provisionings on resource allocations of our proposed adaptive transmission schemes.**

***Index Terms*—Mobile multicast, wireless networks, rate control, layered video streaming, statistical QoS guarantees.**

I. Introduction

RECENTLY, supporting real-time video services with diverse QoS constraints has become one of the essential requirements for wireless communications networks. Consequently, how to efficiently guarantee QoS for video transmission attracts more and more research attention [1]-[14]. However, the unstable wireless environments and the popular layer-structured video signals [2]-[4] impose a great deal of challenges in delay QoS provisionings. Due to the highly-varying wireless channels, the *deterministic* delay QoS requirements are usually hard to guarantee. As a result, *statistical* delay QoS guarantees [9]-[14], in terms of effective bandwidth/capacity and queue-length-bound/delay-bound violation probabilities, have been proposed and demonstrated as the powerful way to characterize delay QoS provisionings for wireless traffics. While many related existing research works mainly focused on the scenarios with single-layer streams [11]-[14], the modern video coding techniques usually generate layer-structured traffics [2]-[4]. Unfortunately, how to design efficient schemes to support statistical delay QoS for layered video traffics over wireless networks has been neither well understood nor thoroughly studied.

In video transmissions, video source is usually encoded into a number of data layers [2]-[4] in the application protocol layer. By applying layered video coding, the receivers under poorer channel conditions can get only lower video quality, while those under better channel conditions can achieve higher video quality. Although the layered coding techniques are efficient in handling diverse channel conditions, they also raise new challenges for statistical delay QoS guarantees, which are not encountered in single-layer video transmissions. First, we need to keep the synchronous transmissions across different video layers, implying the same delay-bound violation probability for all layers. Second, for multi-layer video stream, it is a natural requirement that different video layers can tolerate different loss levels. Therefore, the scheduling and resource allocation need to be aware of the diverse loss constraints. Third, how to minimize the consumption of scarce wireless-resources while satisfying the specified delay QoS requirements is a widely cited open problem.

Besides the general challenges in statistical delay QoS guarantees for the unicast transmission of layered video, multicasting layered video over wireless networks further complicates the problem significantly due to the heterogeneous channel qualities across multicast receivers at each time instant. Unlike in the wireless multicast, there are relatively more research results for the multicast over wireline networks. A number of multicast protocols were proposed over wireline networks. The authors of [3] developed the efficient receiver-driven layered multicast over the Internet, where the video source is encoded to a hierarchical signal with different layers. Each layer corresponds to a multicast group and multicast receivers can join/leave the group based on their bandwidths. In [5], we proposed a novel flow control scheme for multicast services over the asynchronous transfer mode (ATM) networks. The kernel parts of this scheme include the optimal second-order rate control algorithm and the feedback soft-synchronization protocol [6], [7], which can achieve *scalable* and *adaptive* multicast flow control over bandwidth and buffer occupancies and utilizations. The above designs are shown to be efficient in the wireline networks. However, the multicast strategies in wireline networks cannot be directly applied into wireless networks. This is because highly and rapidly time-varying wireless-channels qualities result in unstable bandwidths and thus unsatisfied loss and delay QoS. For wireless video multicast, at the multicast sender we need to design the transmission

Manuscript received 1 March 2009; revised 1 November 2009. The research reported in this paper was supported in part by the U.S. National Science Foundation CAREER Award under Grant ECS-0348694. As this paper was co-authored by a guest editor of this issue, the review of this manuscript was handled by senior editor Prof. Larry Milstein.

The authors are with the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mails: duqinghe@tamu.edu; xizhang@ece.tamu.edu).

Digital Object Identifier 10.1109/JSAC.2010.100413.

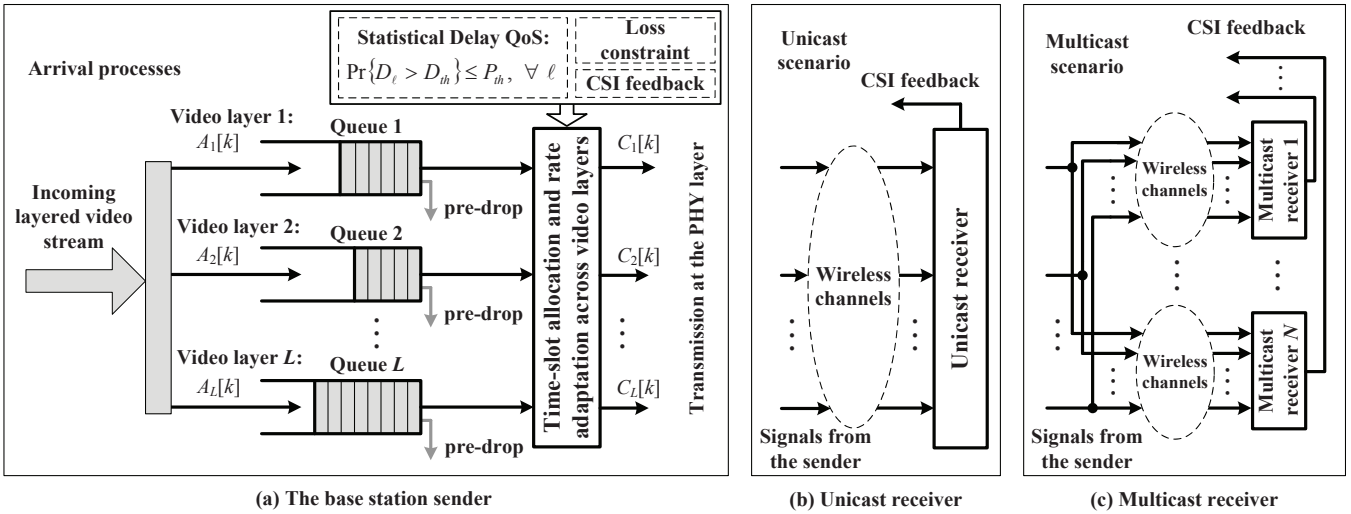

Fig. 1. The system modeling framework for layered-video transmission over wireless networks; (a) The layered-video arrival stream and the sender's processing. (b) Unicast scenario. (c) Multicast scenario.

scheme to control the loss and/or delay performance for all multicast receivers at each video layer based on their instantaneous channel qualities.

In [8] and [14], we applied the effective capacity theory to propose and evaluate rate-adaptation schemes for statistical delay QoS guarantees in mobile multicast. However, the analyses only focused on single-layer stream. It remains one of the major challenges to extend the statistical QoS theory into multi-layer video multicast in developing QoS-driven transmission strategies. In [4], the authors proposed a cross-layer architecture for adaptive video multicast over multirate wireless LANs. In particular, two-layer video signals are considered, which include the base layer (more important) and enhancement layer (less important). The authors derived the transmission rate for the base layer according to the worst-case signal-to-noise ratio (SNR) among all receivers, while dynamically regulating the transmission rate for the enhancement layer based on the best-case SNR to benefit receivers with better channel qualities. However, under this strategy, the loss rate of the enhancement layer will vary significantly with the statistical characteristics of wireless channels, and thus is hard to control.

To overcome the above problems, in this paper we propose an efficient framework to model the statistical delay QoS guarantees, in terms of QoS exponent, effective bandwidth/capacity, and delay-bound violation probability, for multi-layer video transmission over wireless networks. In particular, a separate queue is maintained for each video layer, and the same delay bound and the corresponding violation probability are set up for all video layers. We then develop a set of optimal adaptive transmission schemes to minimize the resource consumption while satisfying the diverse QoS requirements under various scenarios, including video unicast/multicast with and/or without loss tolerance.

The rest of this paper is organized as follows. Section II describes the system model. Section III proposes the framework of statistical delay QoS guarantees for multi-layer video unicast and multicast. Section IV presents the design procedures for multi-layer video by applying effective bandwidth/capacity theory. Sections V and VI derive the optimal adaptive transmission schemes for video unicast and multicast, respectively. Section VII presents the simulation evaluations. The paper concludes with Section VIII.

II. The System Model

We consider the unicast/multicast system model for multi-layer video distribution in wireless networks, as shown in Fig. 1. Specifically, the base station sender is responsible for transmitting a multi-layer video stream to a single receiver (unicast) or multiple receivers (multicast) over broadcast fading channels. The video stream generated by upper protocol layers (e.g., application layer) consists of L video layers, each having the specific QoS requirements. The L-layer video stream will be injected to the physical (PHY) layer. Then, as depicted in Fig. 1(a), we aim at developing strategies to efficiently allocate limited wireless resources for multi-layer video transmission while satisfying the specified QoS requirements for each video layer.

At the PHY layer, the sender uses a constant transmit power with the signal bandwidth equal to B Hz. The wireless broadcast channels are assumed to be flat fading. Then, we can use an SNR vector $\boldsymbol{\gamma} \triangleq (\gamma_1, \gamma_2, \ldots, \gamma_N)$ to characterize the channel state information (CSI) of receivers, where N denotes the number of unicast/multicast receivers, γ_n is the received SNR of the nth receiver for $n = 1, 2, \ldots, N$, and $\{\gamma_n\}_{n=1}^{N}$ are independent and identically distributed (i.i.d.) for the cases of $N > 1$. When N is equal to 1, the scenario reduces to video unicast,[1] as illustrated in Fig. 1(b); while N is larger than 1, we get the multicast scenario depicted in Fig. 1(c). The CSI $\boldsymbol{\gamma}$ is modeled as an ergodic and stationary block-fading process, where $\boldsymbol{\gamma}$ does not change within a time-frame with the fixed length T, but varies independently from frame to frame. Moreover, $\boldsymbol{\gamma}$ follows Rayleigh fading model, which is one the most generally used models to characterize wireless fading channels. In addition, we assume that $\boldsymbol{\gamma}$ can

[1]When $N = 1$, we write SNR as γ instead of γ_1 to simplify notation.

be perfectly estimated by the receivers and reliably fed back to the sender without delay through the dedicated feedback control channels.

III. Modeling Framework for Wireless Unicast/Muticast of Multi-Layer-Video

We propose the following framework for transmitting multi-layer video over fading channels by integrating the adaptive resource allocations, statistical QoS guarantees, and loss constraints.

A. Multi-Queue Model for Multi-Layer Video Arrival Processes

The modern video coding techniques [2] usually encode the video source into a number of video layers with different relevance and importance. The most important layer is called based layer and the other layers are called enhancement layers. Because of the diverse importance, different strategies need to be proposed for the corresponding video layers in the PHY-layer transmission, depending on the specified QoS requirements (to be detailed in Sections III-B and III-D). Then, to achieve the efficient video transmission, the sender manages a separate queue for each video layer. As shown in Fig. 1(a), the data arrival rate of the ℓth video layer is characterized by a discrete-time process, denoted by $A_\ell[k]$ (nats/frame), where $\ell = 1, 2, \ldots, L$, and $[k]$, $k = 1, 2, \ldots$, is the index of time frames; the service rate process (departure process) of the ℓth layer is denoted by $C_\ell[k]$ (nats/frame). Moreover, we determine $C_\ell[k]$ based on CSI, total available wireless resources, and QoS constraints.

B. Statistical Delay QoS Guarantees for Video Transmissions

For video transmissions, delay is one of the most important QoS metrics. However, due to the highly varying wireless channels, usually the hard delay bound cannot be guaranteed. Therefore, the statistical metric, namely *delay-bound violation probability* [9]-[12], has been widely applied in QoS evaluations for real-time services. In our framework, we also use the delay-bound violation probability to statistically characterize the delay QoS provisionings for each video layer. In particular, a queueing delay bound, denoted by $D_{\rm th}$, is specified. Accordingly, over all video layers, the delay-bound violation probability cannot exceed the threshold denoted by $P_{\rm th}$:

$$\Pr\{D_\ell > D_{\rm th}\} \le P_{\rm th}, \quad P_{\rm th} \in (0, 1), \tag{1}$$

for all $\ell \in \{1, 2, \ldots, L\}$, where D_ℓ denotes the queueing delay at the ℓth video layer. The delay-bound violation probability in Eq. (1) is evaluated over the entire transmission process, which is assumed to be long enough. Note that $D_{\rm th}$ and $P_{\rm th}$ are application-dependent parameters. Moreover, the pair of $D_{\rm th}$ and $P_{\rm th}$ is set to be the same for all video layers, because the synchronous transmission across different video layers is usually required. Also note that the delay in wireless video transmissions may result from multiple factors such as transmissions, queueing, and decoding. In this paper, we mainly focus on queueing delay, which reflects the capability of the wireless channel (transmission bottleneck) in supporting video distributions.

C. Adaptive Resource Allocation and Transmissions

To efficiently use the limited wireless resources for video unicast/multicast, we employ the adaptive transmission strategy (based on the CSI), consisting of three folds: transmission rate adaptation, dynamic time-slot allocation, and adaptive pre-drop queue management strategy, as detailed below.

1) Time Slot Allocation for Video Layers: Each time frame is divided into L time slots, the lengths of which are denoted by $\{T_\ell[k]\}_{\ell=1}^L$, where $0 \le T_\ell[k] \le T$ and $\sum_{\ell=1}^L T_\ell[k] \le T$. The time slot with length $T_\ell[k]$ is used for transmitting data of the ℓth video layer. For convenience of presentation, we further define time proportion $t_\ell[k] \triangleq T_\ell[k]/T$, and thus we have $\sum_{\ell=1}^L t_\ell[k] \le 1$. Notice that our target is to minimize the wireless-resource consumption while satisfying the QoS requirements imposed by video qualities. Thus, $\sum_{\ell=1}^L t_\ell[k]$ may be smaller than 1 for some $\boldsymbol{\gamma}$.

2) Rate Adaptation of Unicast/Multicast: We denote the total amount of transmitted data at the ℓth video layer in the kth time frame by $\mathcal{R}_\ell[k]$ (with the unit nats/frame). Moreover, we use the normalized transmission rate, denoted by $R_\ell[k]$ (nats/s/Hz), to characterize the transmission rate adaptation, where $R_\ell[k] \triangleq \mathcal{R}_\ell[k]/(BT)$. We assume that capacity-achieving codes are used for transmission at the PHY layer. Accordingly, for unicast, the normalized transmission rate of the ℓth video layer is set equal to the Shannon capacity under the current SNR γ:

$$R_\ell[k] = \log(1 + \gamma) \qquad (\text{nats/s/Hz}). \tag{2}$$

Clearly, $R_\ell[k]$ does not vary with ℓ, and thus we only focus on time-slot allocation for unicast.

For the multicast case, the rate adaptation becomes more complicated. In particular, the time slot for video layer ℓ is further partitioned into N sub-slots. The length of the nth sub-slot, denoted by $T_{\ell,n}[k]$, is equal to $T_\ell[k]t_{\ell,n}[k]$, where $0 \le t_{\ell,n}[k] \le 1$ and $\sum_{n=1}^N t_{\ell,n}[k] = 1$. Within the nth sub-slot, the transmission rate is set equal to the Shannon capacity under SNR γ_n, and thus the data transmitted in this sub-slot can be correctly decoded only by receivers with SNR higher than or equal to γ_n. Then, the normalized transmission rate $R_\ell[k]$ for the ℓth video layer becomes

$$\begin{aligned} R_\ell[k] &= \sum_{n=1}^N t_{\ell,n}[k] R_{\ell,n}[k] \\ &= \sum_{n=1}^N t_{\ell,n}[k] \log(1 + \gamma_n) \quad (\text{nats/s/Hz}), \end{aligned} \tag{3}$$

where $R_{\ell,n}[k] \triangleq \log(1 + \gamma_n)$ is the normalized transmission rate for the nth sub-slot of the ℓth video layer. As a result, we need to not only adjust $t_\ell[k]$'s for each layer, but also regulate $t_{\ell,n}[k]$'s within every time slot.

Unlike the wireline multicast networks, in this paper we focus on the layered video transmissions over wireless networks, which has a single-hop cellular network structure. Due to the broadcast nature of wireless channels, the sender only needs to transmit a *single copy* of data and *all* multicast receivers can hear the transmitted signal for each video layer. Under this model, our scheme employs the *sender-oriented* multicast approach because the sender needs to dynamically adjust the

transmissions rate in controlling the loss rate (to be detailed in Section III-D) and guaranteeing the delay-QoS requirements (see Section III-B) *across* different multicast receivers.

3) Pre-drop Strategy: In [14], for multicasting single-layer-data we developed the pre-drop strategy to gain a more robust queueing behavior. In this paper, we further extend the pre-drop strategy to multi-layer video transmission. Specifically, based on the CSI, in each time frame the sender can drop some data (see Fig. 1(a)) from the head of each queue, but treat them *as if* they were transmitted. We denote the amount of dropped data in ℓth video layer by $Z_\ell[k]$ (nats/frame) and define the normalized drop rate, denoted by $z_\ell[k]$, as $z_\ell[k] \triangleq Z_\ell[k]/(BT)$ (nats/s/Hz). Then, the service process $C_\ell[k]$ of the ℓth video layer is given by

$$C_\ell[k] = BT(t_\ell[k]R_\ell[k] + z_\ell[k]) \qquad (\text{nats/frame}). \tag{4}$$

Clearly, the pre-drop strategy suppresses the growing speed of the queue for a more robust queueing behavior, but this strategy also causes data loss to all multicast receivers. As a result, $z_\ell[k]$ needs to be determined by not only the CSI, but also the loss constraints (see Section III-D).

D. Loss Rate Constraint

Although a certain loss is usually tolerable for delay-sensitive services, the loss level cannot be arbitrarily high. Consequently, we require the loss rate of the ℓth video layer for each receiver to be limited lower than or equal to an application-dependent threshold, denoted by $q_{\text{th}}^{(\ell)}$. The loss rate of the ℓth video layer for the nth receiver, denoted by $q_{\ell,n}$, is defined as the ratio of the amount of data correctly received by this receiver to that of the data transmitted at this video layer. Data loss for unicast will be caused only by the pre-drop strategy, while data loss for multicast will be introduced by both pre-drop operation and heterogeneous channel fading across multicast receivers.

Since various efficient forward-error control (FEC) codes [16]-[18] at upper protocol layers were proposed and widely applied to multicast communications in wired/wireless networks, in our framework we suppose that FEC mechanisms are already employed at the upper protocol layers. The error-control redundancies added in the FEC codes at different video layers are inherently related among video layers and are jointly determined by the targeted video-delivery qualities at different video layers. Correspondingly, the tolerable loss-rate levels $q_{\text{th}}^{(\ell)}$'s for different video layers (indexed by ℓ) are jointly specified based on the video delivery quality requirements and the error control redundancy degrees across different video layers. Under this framework, we then mainly focus on how to use the minimum wireless resources with QoS guarantees to unicast/multicast multi-layer video over wireless channels.

IV. Statistical Delay QoS Guarantees Through Effective Bandiwdth/Capacity

A. Preliminary for Statistical Delay QoS Guarantees

The theory on statistical delay QoS guarantees [9]-[11] provides a powerful approach in analyzing the queueing behavior for time-varying arrival and/or service processes. Specifically, consider a stable queueing system with the stationary and ergodic arrival and service processes. Asymptotic analyses [9] show that with sufficient conditions, the queue length process $Q[k]$ converges to a random variable $Q[\infty]$ in distribution such that

$$-\lim_{Q_{\text{th}}\to\infty} \frac{\log\left(\Pr\{Q[\infty] > Q_{\text{th}}\}\right)}{Q_{\text{th}}} = \theta \tag{5}$$

for a certain $\theta > 0$, where Q_{th} is the queue-length bound. Moreover, the queue-length bound violation probability can be approximated by

$$\Pr\{Q > Q_{\text{th}}\} \approx e^{-\theta Q_{\text{th}}}, \tag{6}$$

where we remove the index $[k]$ for $Q[k]$ to simplify notations. In the above two equations, θ is called *QoS exponent* and can be used to characterize delay QoS. The larger θ corresponds to the more stringent QoS requirement, while the smaller θ imposes the looser delay constraint. When the QoS metric of interest becomes delay-bound, the similar expressions can be obtained.

B. Preliminary for Effective Bandwidth/Capacity

Effective bandwidth [10] and effective capacity [11] are a pair of dual concepts. Given an arrival process $A[k]$, effective bandwidth of $A[k]$, denoted by $\mathcal{A}(\theta)$ (nats/frame), is defined as the *minimum constant service rate* required to guarantee a specified QoS exponent θ, i.e., Eqs. (5)-(6) are satisfied for the given θ. In contrast, given a service process $C[k]$, effective capacity of $C[k]$, denoted by $\mathcal{C}(\theta)$ (nats/frame), is defined as the maximum *constant arrival rate* which can be supported by $C[k]$ subject to the specified QoS exponent θ. Moreover, effective bandwidth [10] and effective capacity [11] can be expressed as $\mathcal{A}(\theta) = \lim_{k\to\infty}(1/(k\theta))\log\left(\mathbb{E}\left\{e^{\theta S_A[k]}\right\}\right)$ and $\mathcal{C}(\theta) = -\lim_{k\to\infty}(1/(\theta k))\log\left(\mathbb{E}\left\{e^{-\theta S_C[k]}\right\}\right)$, respectively, where $\mathbb{E}\{\cdot\}$ denotes the expectation, $S_A[k] \triangleq \sum_{i=1}^{k} A[i]$, and $S_C[k] \triangleq \sum_{i=1}^{k} C[i]$. In addition, for effective capacity and effective bandwidth scenarios, the delay-bound violation probability can be approximated [10], [11] as:

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta\mathcal{C}(\theta)D_{\text{th}}}, \text{ for effective capacity;} \tag{7}$$

and

$$\Pr\{D > D_{\text{th}}\} \approx e^{-\theta\mathcal{A}(\theta)D_{\text{th}}}, \text{ for effective bandwidth.} \tag{8}$$

Equations (6), (7), and (8) are good approximations for relatively large Q_{th} and D_{th} as shown in [11], [22]. When Q_{th} and D_{th} are relatively small, the more accurate approximations expressions than Eqs. (6), (7), and (8) are given in [11], [22] as follows:

$$\begin{cases} \Pr\{Q > Q_{\text{th}}\} & \approx & \varrho e^{-\theta Q_{\text{th}}}; \\ \Pr\{D > D_{\text{th}}\} & \approx & \varrho e^{-\theta\mathcal{C}(\theta)D_{\text{th}}}; \\ \Pr\{D > D_{\text{th}}\} & \approx & \varrho e^{-\theta\mathcal{A}(\theta)D_{\text{th}}}, \end{cases}$$

where ϱ denotes the probability that the queue is nonempty. These approximations are upper-bounded by the corresponding approximations given in Eqs. (6)-(8). Thus, directly using Eqs. (6)-(8) for the system design often guarantees more stringent QoS than the specified requirements. For wireless video transmissions, the delay bound D_{th} is typically hundreds of milliseconds (ms), which are thus much larger than the

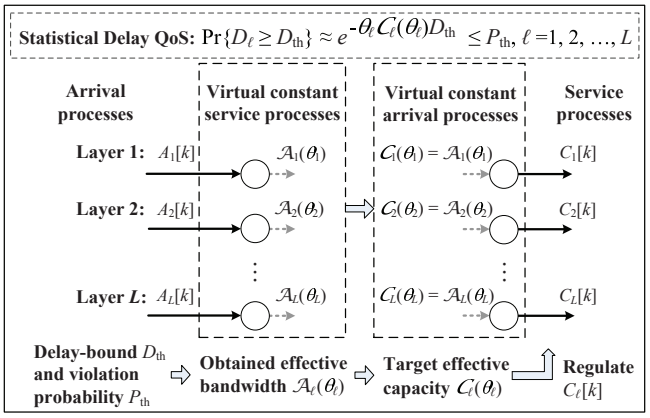

Fig. 2. Illustration of design procedures to guarantee statistical delay QoS by using effective capacity and effective bandwidth theories.

adaptive-transmission period scale (e.g., the physical-layer time-frame length) of the wireless transmission system, where the adaptive-transmission period typically varies from a few milliseconds (ms) to tens of milliseconds (ms). Therefore, Eqs. (6)-(8) are good approximating expressions in designing efficient wireless video-transmission schemes with statistical QoS guarantees.

C. Design Procedures for Transmitting Layered Video with Statistical QoS Guarantees

For a dynamic queueing system, in order to guarantee the QoS exponent θ given in Eqs. (5)-(6), the following equation need to be satisfied [10], [12]:

$$\mathcal{C}(\theta) = \mathcal{A}(\theta). \tag{9}$$

Inspired by this property, the statistical delay QoS guarantees can be characterized through the arrival process and service process separately. As shown in Fig. 2, the queueing system for the ℓth video layer can be decomposed to two virtual queueing systems. The one on the left-hand side of Fig. 2 is composed by the true arrival process $A_\ell[k]$ and one virtual constant-rate service process, the rate of which is equal to the effective bandwidth $\mathcal{A}_\ell(\theta_\ell)$ of $A_\ell[k]$; the right one consists of the true service process $C_\ell[k]$ and one constant-rate virtual arrival process, the rate of which is equal to the effective capacity $\mathcal{C}_\ell(\theta_\ell)$ of $C_\ell[k]$. Using the above concept, we develop the design procedures to provide statistical delay QoS guarantees for transmitting multi-layer video stream as shown in Fig. 3.

Among the procedures in Fig. 3, Steps 1 and 2 first identify the effective bandwidth $\mathcal{A}_\ell(\theta_\ell)$ and QoS exponent θ_ℓ required to satisfy the delay-bound $D_{\rm th}$ and its violation probability $P_{\rm th}$. Then, to satisfy the delay QoS in Eq. (1), we need to either satisfy Eq. (9) or guarantee that the effective capacity is larger than the effective bandwidth, which results in Steps 3 and 4.

The analytical expressions of effective bandwidth for many typical arrival processes, such as constant-rate process, autoregressive (AR) process, and Markovian process, can be found in [10]. Note that if $A_\ell[k]$ is time varying, in Step 2 θ_ℓ can be determined through Eq. (8). However, if $A_\ell[k]$ is a

Step 1:	Determine effective bandwidth functions $\mathcal{A}_\ell(\theta)$ for the arrival processes $A_\ell[k]$, $\ell = 1, 2, \ldots, L$.
Step 2:	Apply Eq. (7) or Eq. (8) to find the solution θ_ℓ to the equation $\Pr\{D_\ell > D_{\rm th}\} = P_{\rm th}$ and get the corresponding effective bandwidth $\mathcal{A}_\ell(\theta_\ell)$.
Step 3:	Set the target effective capacity $\overline{\mathcal{C}}_\ell = \mathcal{A}_\ell(\theta_\ell)$ for the service processes of each video layer.
Step 4:	Jointly design adaptive service process $\mathcal{C}_\ell[k]$ for each video layer, such that $\mathcal{C}_\ell(\theta_\ell) \geq \overline{\mathcal{C}}_\ell$ is satisfied while minimizing the total consumed wireless resources.

Fig. 3. The design procedures to provide statistical delay QoS guarantees for transmitting multi-layer video stream.

constant-rate process equal to $\overline{A}_\ell$, we have $\mathcal{A}_\ell(\theta) = \overline{A}_\ell$ for all θ, implying that the delay-bound violation probability of the virtual queueing system on the left-hand side of Fig. 2 is always equal to 0. Therefore, we cannot derive θ_ℓ directly through Eq. (8). In contrast, the QoS exponent θ_ℓ to guide the adaptive transmission needs to be determined by using Eq. (7) under the condition of $\mathcal{C}_\ell(\theta_\ell) = \mathcal{A}_\ell(\theta_\ell) = \overline{A}_\ell$.

V. Unicasting Multi-Layer Video Stream

Assuming that the target effective capacity $\{\overline{\mathcal{C}}_\ell\}_{\ell=1}^L$ and QoS exponent θ_ℓ have been determined, we next focus on developing the optimal adaptive time-slot allocation and pre-drop strategy to satisfy the QoS requirements while minimizing the wireless-resource consumption. Unless otherwise mentioned, we drop the time-frame index $[k]$ for the corresponding variables in the rest of this paper to simplify notations. Based on our previous work [13], if a stationary and ergodic service rate process C_ℓ is uncorrelated across different time frames, we can write its effective capacity as follows:

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell} \log\left(\mathbb{E}\left\{e^{-\theta_\ell C_\ell}\right\}\right) \quad \text{nats/frame.} \tag{10}$$

Since the block-fading channel model described in Section II satisfies the time-uncorrelated condition, we can apply Eq. (10) for our framework to derive the adaptive unicast/mutlicast schemes with the statistical QoS guarantees.

A. Unicasting Layered Video Stream Without Loss Tolerance

We first consider the cases without loss tolerance for multi-layer video transmissions, i.e., $q_{\rm th}^{(\ell)} = 0$ for all ℓ and the pre-drop strategy will not be applied. Thus, we only need to focus on regulating the time-slot proportion $\{t_\ell\}_{\ell=1}^L$ for each video layer. Following the design target and QoS constraints characterized in Sections III and IV, we derive the adaptive transmission strategy by solving the following optimization problem.

Problem 1: $\boldsymbol{P1}$

$$\min_{\boldsymbol{t}} \left\{\sum_{\ell=1}^{L} \mathbb{E}_\gamma\{t_\ell\}\right\} \tag{11}$$

$$\text{s.t.:} \quad \mathcal{C}_\ell(\theta_\ell) \geq \overline{\mathcal{C}}_\ell, \quad \forall\, \ell, \tag{12}$$

$$\sum_{\ell=1}^{L} t_\ell - 1 \leq 0,\ 0 \leq t_\ell \leq 1,\ \forall\, \gamma, \tag{13}$$

where $\boldsymbol{t} \triangleq (t_1, t_2, \ldots, t_L)$ and $\mathbb{E}_\gamma\{\cdot\}$ denotes the expectation over the random variable γ.

Using Eq. (10), the constraint in (12) can be equivalently rewritten as

$$\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell \le 0, \tag{14}$$

where $\beta_\ell \triangleq \theta_\ell TB$ is termed normalized QoS exponent and $V_\ell \triangleq e^{-\theta_\ell \overline{\mathcal{C}}_\ell}$. It is not difficult to see: 1) the objective function in $\boldsymbol{P1}$ is convex over $\boldsymbol{t}$; 2) the functions on the left-hand side of all inequality constraints (Eqs. (13) and (14)) are convex over $\boldsymbol{t}$. Therefore, $\boldsymbol{P1}$ is a convex problem [20] and the optimal solution can be obtained by using the Lagrangian method and the Karush-Kuhn-Tucker (KKT) conditions [20], which is summarized in Theorem 1.

Theorem 1: The optimal solution $\boldsymbol{t}^*$ to problem $\boldsymbol{P1}$, if existing, is determined by

$$t_\ell^* = t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma^*) \triangleq \left[-\frac{\log\left(\frac{1+\psi_\gamma^*}{\beta_\ell \lambda_\ell^* \log(1+\gamma)}\right)}{\beta_\ell \log(1+\gamma)}\right]^+, \tag{15}$$

where $[\cdot]^+ \triangleq \max\{\cdot, 0\}$. The parameters ψ_γ^* and $\{\lambda_\ell^*\}_{\ell=1}^L$ are the optimal Lagrangian multipliers associated with Eqs. (13) and (14), respectively. Given SNR γ in a fading state and $\{\lambda_\ell^*\}_{\ell=1}^L$, if

$$\sum_{\ell=1}^{L} t_\ell(\gamma, \lambda_\ell^*, 0) \ge 1 \tag{16}$$

holds, ψ_γ^* is the unique solution to

$$\sum_{\ell=1}^{L} t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma^*) = 1, \quad \psi_\gamma^* \ge 0; \tag{17}$$

otherwise, we get

$$\psi_\gamma^* = 0. \tag{18}$$

Under the above strategy to determine $\boldsymbol{t}^*$ and ψ_γ^*, the optimal $\{\lambda_\ell^*\}_{\ell=1}^L$ are selected to satisfy

$$\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell^* \log(1+\gamma)}\right\} - V_\ell = 0, \ \forall \ell. \tag{19}$$

Proof: We construct the Lagrangian function for $\boldsymbol{P1}$, denoted by J, as follows:

$$\begin{aligned} J &= \mathbb{E}_\gamma\left\{\sum_{\ell=1}^{L} t_\ell\right\} + \mathbb{E}_\gamma\left\{\psi_\gamma\left(\sum_{\ell=1}^{L} t_\ell - 1\right)\right\} \\ &\quad + \sum_{\ell=1}^{L} \lambda_\ell\Big(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\Big), \end{aligned} \tag{20}$$

where $\psi_\gamma \ge 0$ and $\lambda_\ell \ge 0$, $\ell = 1, 2, \ldots, L$, are Lagrangian multipliers associated with Eqs. (13) and (14), respectively. Then, the optimal $\boldsymbol{t}^*$ and Lagrangian multipliers of optimization problem $\boldsymbol{P1}$ satisfy the following KKT conditions [20]:

$$\begin{cases} \frac{\partial J}{\partial t_\ell}\big|_{t_\ell = t_\ell^*} = 0, \forall\, \ell, \forall\, \gamma \\ \psi_\gamma^* \ge 0 \text{ and } \lambda_\ell^* \ge 0; \\ \psi_\gamma^*\left(\sum_{\ell=1}^{L} t_\ell^* - 1\right) = 0, \ \forall\, \ell, \ \gamma; \\ \lambda_\ell^*\left(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell^* \log(1+\gamma)}\right\} - V_\ell\right) = 0, \ \forall\, \ell. \end{cases} \tag{21}$$

Taking the derivative of J with respect to (w.r.t.) t_ℓ, we get

$$\frac{\partial J}{\partial t_\ell} = \Big(1 + \psi_\gamma - \lambda_\ell \beta_\ell (1+\gamma)^{-\beta_\ell t_\ell} \log(1+\gamma)\Big) f_\Gamma(\gamma) d\gamma \tag{22}$$

where $f_\Gamma(\gamma)$ is the probability density function (pdf) of γ. Plugging Eq. (22) into the first line of Eq. (21) and solving for t_ℓ^* under the boundary condition $t_\ell \ge 0$, we get Eq. (15).

According to Eq. (15), $t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma)$ is a strictly decreasing function of ψ_γ for $t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma) > 0$, and $\psi_\gamma^* \to \infty$ leads to $t_\ell^* \to 0$. Therefore, if Eq. (16) holds, we can always find the unique ψ_γ^* to satisfy Eq. (18); otherwise, the inequality $\sum_{\ell=1}^{L} t_\ell(\gamma, \lambda^*, \psi_\gamma) - 1 < 0$ follows for any $\psi_\gamma \ge 0$, implying $\psi_\gamma^* = 0$ by applying the third line of Eq. (21). Through Eq. (15), $\lambda_\ell^* = 0$ results in $t_\ell = 0$ for all γ, and thus the constraint in Eq. (12) will be violated, implying an infeasible solution. Therefore, λ_ℓ^* has to be positive. Then, to satisfy the fourth line of Eq. (21), Eq. (19) must hold and thus Theorem 1 follows. ■

Note that given $\{\lambda_\ell^*\}_{\ell=1}^L$, ψ_γ^* is easy to solve because $t_\ell(\gamma, \lambda_\ell^*, \psi_\gamma)$ is a decreasing function of ψ_γ. However, how to find $\{\lambda_\ell^*\}_{\ell=1}^L$ is still unknown. Moreover, Theorem 1 does not state whether the optimal solution exists. Next, we discuss how to get $\{\lambda_\ell^*\}_{\ell=1}^L$ and examine the existence of the optimal solution, which can be performed either off-line or on-line. Based on the optimization theory [19], [20], the Lagrangian dual problem to $\boldsymbol{P1}$ is given by

$$\max_{(\boldsymbol{\lambda}, \psi_\gamma)} \left\{\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)\right\}, \tag{23}$$

where $\boldsymbol{\lambda} \triangleq (\lambda_1, \lambda_2, \ldots, \lambda_L)$ and $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$ is the Lagrangian dual function defined by $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma) \triangleq \min_{\boldsymbol{t}}\{J\} = J|_{t_\ell = t_\ell(\gamma, \lambda_\ell, \psi_\gamma)}$. We can further convert Eq. (23) into $\max_{(\boldsymbol{\lambda}, \psi_\gamma)}\left\{\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)\right\} = \max_{\boldsymbol{\lambda}}\left\{\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda}))\right\}$, where $\psi_\gamma(\boldsymbol{\lambda})$ denotes the maximizer of $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$ given $\boldsymbol{\lambda}$. Moreover, we can obtain $\psi_\gamma(\boldsymbol{\lambda})$ by using the same procedures as those used in determining ψ_γ^*, which are given by Eqs. (16)-(18) in Theorem 1.

Since problem $\boldsymbol{P1}$ is convex, there is no duality gap between $\boldsymbol{P1}$ and its dual problem given by Eq. (23) if the optimal solution exists. Thus, the optimal Lagrangian multipliers $\{\lambda_\ell^*\}_{\ell=1}^L$ and ψ_γ^* to problem $\boldsymbol{P1}$ also maximize the objective function $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$ in Eq. (23). Consequently, we can obtain $\{\lambda_\ell^*\}_{\ell=1}^L$ through maximizing $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma)$. Following convex optimization theory [20], $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda}))$ is a concave function over $\boldsymbol{\lambda}$, and thus we can track the optimal $\boldsymbol{\lambda}^*$ by using the subgradient method [21]:

$$\lambda_\ell := \lambda_\ell + \epsilon\left(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\right)\Big|_{t_\ell = t_\ell(\gamma, \lambda_\ell, \psi_\gamma(\boldsymbol{\lambda}))} \tag{24}$$

where ϵ is a positive real number close to 0. and $\left(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\right)$ in the above equation is a subgradient of $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda}))$ w.r.t. λ_ℓ [19] (see the definition of subgradient in Section VI-B). If the optimal solution to $\boldsymbol{P1}$ exists, the above iteration will converge to the optimal $\boldsymbol{\lambda}^*$ with properly selected ϵ because of the concavity of $\widetilde{J}(\boldsymbol{\lambda}, \psi_\gamma(\boldsymbol{\lambda}))$. Correspondingly, $\left(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell\right)$ will converge to 0. If the optimal solution to $\boldsymbol{P1}$ does not exist, we cannot

support such a statistical QoS requirement even we use up all time slots. Then, $(\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\} - V_\ell)$ is always larger than 0 for some ℓ. As a result, λ_ℓ will approach infinity. So, if any λ_ℓ does not converge and keeps increasing, we can conclude that the optimal solution does not exist and current wireless resources are not enough to support the specified statistical delay QoS for the incoming multi-layer video stream.

To find the optimal $\boldsymbol{\lambda}^*$, we need the pdf of $\boldsymbol{\gamma}$. In realistic systems, although the pdf of $\boldsymbol{\gamma}$ is usually unknown, we can still apply Eq. (24) to implement online tracking. In particular, the iterative update of Eq. (24) will be performed in each time frame. However, the expectation of $e^{-\beta_\ell t_\ell \log(1+\gamma)}$ in Eq. (24) needs to be substituted by its estimation obtained based on the statistics from previous time frames. Denoting the estimation of $\mathbb{E}_\gamma\left\{e^{-\beta_\ell t_\ell \log(1+\gamma)}\right\}$ in the kth time frame by $\mathcal{S}_\ell[k]$, we obtain $\mathcal{S}_\ell[k+1]$ through a first-order autoregressive filter (low-pass filter) as follows:

$$\mathcal{S}_\ell[k+1] := (1-\alpha)\mathcal{S}_\ell[k] + \alpha e^{-\beta_\ell t_\ell [k+1]\log(1+\gamma[k+1])}, \quad (25)$$

where $\ell = 1, 2, \ldots, L$ and $\alpha \in (0,1)$ is a small positive number close to 0. If the optimal solution exists, the online tracking method converges with properly selected α and ϵ. Section VII will present some examples of tracking the optimal Lagrangian multipliers through simulations.

B. Unicasting Layered Video Stream With Loss Tolerance

When $q_{\text{th}}^{(\ell)} > 0$, the transmission strategy becomes more complicated, but will use less wireless resources. After integrating the pre-drop strategy, the loss rate q_ℓ of the ℓth layer is derived as

$$q_\ell = 1 - \frac{BT\mathbb{E}_\gamma\{t_\ell R_\ell\}}{\mathbb{E}_\gamma\{C_\ell\}} = 1 - \frac{\mathbb{E}_\gamma\{t_\ell R_\ell\}}{\mathbb{E}_\gamma\{t_\ell R_\ell + z_\ell\}}. \quad (26)$$

Next, we identify the adaptive transmission policy by solving optimization problem $\boldsymbol{P2}$:

Problem 2: $\boldsymbol{P2}$

$$\min_{(\boldsymbol{t},\boldsymbol{z})} \left\{\sum_{\ell=1}^{L} \mathbb{E}_\gamma\{t_\ell\}\right\} \quad (27)$$

$$\text{s.t.:} \quad \mathbb{E}_\gamma\left\{e^{-\beta_\ell(z_\ell + t_\ell \log(1+\gamma))}\right\} - V_\ell \le 0,\ z_\ell \ge 0,\ \forall\,\ell, \quad (28)$$

$$q_\ell \le q_{\text{th}}^{(\ell)},\ \forall\,\ell \quad (29)$$

$$\sum_{\ell=1}^{L} t_\ell \le 1,\ 0 \le t_\ell \le 1, \quad \forall\,\gamma, \quad (30)$$

where $\boldsymbol{z} \triangleq (z_1, z_2, \ldots, z_\ell)$.

Applying Eq. (26), we can rewrite Eq. (29) as follows:

$$\left(1 - q_{\text{th}}^{(\ell)}\right)\mathbb{E}_\gamma\{z_\ell\} - q_{\text{th}}^{(\ell)}\mathbb{E}_\gamma\{t_\ell \log(1+\gamma)\} \le 0,\ \forall\,\ell. \quad (31)$$

It is also not hard to prove that problem $\boldsymbol{P2}$ is still a convex problem and the Lagrangian method is still effective in finding the optimal solutions, which is summarized in Theorem 2.

Theorem 2: The optimal solution $(\boldsymbol{t}^*, \boldsymbol{z}^*)$, if existing, is expressed by a set of functions of γ, $\{\lambda_\ell\}_{\ell=1}^L$, $\{\phi_\ell\}_{\ell=1}^L$, and ψ_γ as follows::

$$t_\ell^* = t_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, \psi_\gamma^*), \quad z_\ell^* = z_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, \psi_\gamma^*), \quad \forall\,\ell, \quad (32)$$

where

$$t_\ell(\gamma, \lambda_\ell, \phi_\ell, \psi_\gamma) \triangleq \begin{cases} \infty, & \text{if } -\infty < \frac{1+\psi_\gamma}{\log(1+\gamma)} \le \phi_\ell q_{\text{th}}^{(\ell)}; \\ \left[-\frac{1}{\beta_\ell \log(1+\gamma)} \log\left(\frac{1+\psi_\gamma - q_{\text{th}}^{(\ell)}\phi_\ell \log(1+\gamma)}{\beta_\ell \lambda_\ell \log(1+\gamma)}\right)\right]^+, & \\ & \text{if } \phi_\ell q_{\text{th}}^{(\ell)} < \frac{1+\psi_\gamma}{\log(1+\gamma)} < \phi_\ell; \\ 0, & \text{if } \phi_\ell \le \frac{1+\psi_\gamma}{\log(1+\gamma)} < \infty, \end{cases} \quad (33)$$

and

$$z_\ell(\gamma, \lambda_\ell, \phi_\ell, \psi_\gamma) \triangleq \begin{cases} 0, & \text{if } -\infty < \frac{1+\psi_\gamma}{\log(1+\gamma)} < \phi_\ell\,; \\ \left[-\frac{1}{\beta_\ell}\log\left(\frac{\phi\left(1-q_{\text{th}}^{(\ell)}\right)}{\beta_\ell \lambda_\ell}\right)\right]^+, & \text{if } \phi_\ell \le \frac{1+\psi_\gamma}{\log(1+\gamma)} < \infty. \end{cases} \quad (34)$$

Given γ, $\{\lambda_\ell^*\}_{\ell=1}^L$, and $\{\phi_\ell^*\}_{\ell=1}^L$, if $\sum_{\ell=1}^L t_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, 0) \ge 1$, ψ_γ^* is selected such that the equation $\sum_{\ell=1}^L t_\ell(\gamma, \lambda_\ell^*, \phi_\ell^*, \psi_\gamma^*) = 1$ holds; otherwise, we have $\psi_\gamma^* = 0$. The optimal $\{\lambda_\ell^*\}_{\ell=1}^L$ and $\{\phi_\ell^*\}_{\ell=1}^L$ need to be jointly selected such that "=" holds in both Eqs. (28) and (31).

Proof: The proof of Theorem 2 can be readily obtained by using the standard Lagrangian-multiplier based method and KKT conditions, and thus is omitted due to lack of space. The detailed proof of Theorem 2 is provided online in [24]. ■

In order to search for the optimal Lagrangian multipliers and check the existence of the optimal solution, we can also design the adaptive tracking method similar to problem $\boldsymbol{P1}$.

VI. QoS Guarantees for Multicasting Layered-Video Stream

We consider the multicast scenario in this section. If no loss is tolerated, the transmission rate in each time frame is limited by the worst-case SNR among all multicast receivers. Thus, the system throughput will be degraded very quickly as the multicast group size increases. Therefore, we mainly focus on the multicast scenario with loss tolerance.

A. Problem Formulation for Multicast Scenario

Under the multicast rate-adaptation strategy given in Section III, the loss rate of the nth receiver at the ℓth video layer becomes

$$q_{\ell,n} = 1 - \frac{\mathbb{E}_{\boldsymbol{\gamma}}\left\{t_\ell \sum_{i=1}^N t_{\ell,i}\log(1+\gamma_i)\delta_{\gamma_n \ge \gamma_i}\right\}}{\mathbb{E}_{\boldsymbol{\gamma}}\{z_\ell + t_\ell R_\ell\}}, \quad (35)$$

where $\mathbb{E}_{\boldsymbol{\gamma}}\{\cdot\}$ denotes the expectation over all fading states of the random vector variable $\boldsymbol{\gamma}$, $\delta_{\gamma_n \ge \gamma_i}$ is the indication function (for a given statement ϖ, $\delta(\varpi) = 1$ if ϖ is true, and $\delta(\varpi) = 0$ otherwise), and $R_\ell = \sum_{i=1}^N t_{\ell,i}\log(1+\gamma_i)$ is the total normalized transmission rate in a time frame (see Eq. (3)). Accordingly, the following loss-rate constraint needs to be satisfied for each multicast receiver at every video layer, which is specified by the inequality as follows:

$$q_{\ell,n} \le q_{\text{th}}^{(\ell)}, \quad \forall\, n,\ \forall\,\ell. \quad (36)$$

To simplify the derivations, we first use a *relaxed* constraint to replace Eq. (36) by:

$$q_{\ell,0} \triangleq \frac{1}{N}\sum_{n=1}^{N} q_{\ell,n} \le q_{\text{th}}^{(\ell)}, \quad \forall\, \ell \tag{37}$$

where $q_{\ell,0}$ is called *group loss rate* (average loss rate over receivers) at the ℓth video layer. We will show later that the optimal adaptation policy derived under the group-loss-rate constraint given in Eq. (37) does not violate Eq. (36), and thus is also optimal under the original loss-rate constraint given by Eq. (36). Plugging Eq. (35) into Eq. (37), we have

$$q_{\ell,0} = 1 - \frac{\mathbb{E}_{\boldsymbol{\gamma}}\left\{t_\ell \sum_{i=1}^{N} t_{\ell,i} m_i \log(1+\gamma_i)\right\}}{N\mathbb{E}_{\boldsymbol{\gamma}}\left\{z_\ell + t_\ell R_\ell\right\}}, \tag{38}$$

where m_i is the number of receivers with SNR higher than or equal to γ_i. In addition, it is clear that R_ℓ falls in the following range:

$$R_\ell \in \left[R_{\pi(N)}, R_{\pi(1)}\right], \tag{39}$$

where $R_{\pi(N)} \triangleq \min_{1\le n\le N}\{\log(1+\gamma_n)\}$ and $R_{\pi(1)} \triangleq \max_{1\le n\le N}\{\log(1+\gamma_n)\}$. Note that when we attempt to use a normalized transmission rate equal to R_ℓ in a time frame, there are many different choices for $\{t_{\ell,n}\}_{n=1}^{N}$ to get the same R_ℓ. In order to minimize the loss for the entire multicast group, among all these choices we need to select the one which maximizes the numerator of the second term on the right-hand side of Eq. (38), which represents the sum rate of data correctly received by each multicast receiver. Accordingly, we define

$$\begin{aligned}\widetilde{g}_s(R_\ell) \quad \triangleq \quad & \max_{\boldsymbol{t}_\ell:\ \sum_{i=1}^{N} t_{\ell,i}=1} \left\{\sum_{i=1}^{N} t_{\ell,i} m_i \log(1+\gamma_i)\right\} \\ & \text{s.t.:} \sum_{i=1}^{N} t_{\ell,i}\log(1+\gamma_i) = R_\ell \end{aligned} \tag{40}$$

where $\boldsymbol{t}_\ell \triangleq (t_{\ell,1}, t_{\ell,2}, \ldots, t_{\ell,N})$. Therefore, $\widetilde{g}_s(R_\ell)$ denotes the maximum sum of achieved rates over all multicast receivers under the given normalized transmission rate R_ℓ. In our previous work [15], where we studied the tradeoff between the average throughput and loss rate for single-layer multicast, we showed that $\widetilde{g}_s(R_\ell)$ can be derived through the concept of convex hull [20]. Using the properties of convex hull, in [15] we proved that $\widetilde{g}_s(R_\ell)$ is a *continuous*, *piecewise linear*, and *concave function* over R_ℓ. Thus, we can obtain $\widetilde{g}_s(R_\ell)$ as follows:

$$\widetilde{g}_s(R_\ell) = \begin{cases} \widetilde{g}_s(r_i) + \eta_i(R_\ell - r_i), & \text{if } R_\ell \in [r_i, r_{i-1}), \\ & 2 \le i \le \mathcal{N}; \\ \widetilde{g}_s(r_2) + \eta_2(r_1 - r_2), & \text{if } R_\ell = r_1, \end{cases} \tag{41}$$

where $R_{\pi(1)} = r_1 > r_2 > \cdots > r_{\mathcal{N}} = R_{\pi(N)}$. Fig. 4 depicts an example for the function $\widetilde{g}_s(R_\ell)$. As shown in Fig. 4, within each interval $[r_i, r_{i-1})$, $\widetilde{g}_s(R_\ell)$ is a linear function of R_ℓ with the slope equal to η_i, and $(\mathcal{N}-1)$ is equal to the number of such intervals. Note that $\{(r_i, \widetilde{g}_s(r_i))\}_{i=1}^{\mathcal{N}}$ are actually the vertices on the upper boundary of the convex hull of the 2-dimensional point set $\left\{(\log(1+\gamma_i), m_i\log(1+\gamma_i))\right\}_{i=1}^{N}$ (see [15]). For the complete procedures to identify $\widetilde{g}_s(R_\ell)$ and the corresponding time slot allocation policy, please refer to our previous work [15]. The above discussions imply that we need to consider only the transmission policies yielding $\widetilde{g}_s(R_\ell)$, because only these policies can minimize the total loss for the entire multicast group. Moreover, Eqs. (40)-(41) suggest that we can focus on regulating the scalar R_ℓ instead of the N-dimension time-proportion vector $\boldsymbol{t}_\ell$. After R_ℓ is determined, we can use Eq. (40) to obtain $\boldsymbol{t}_\ell$.

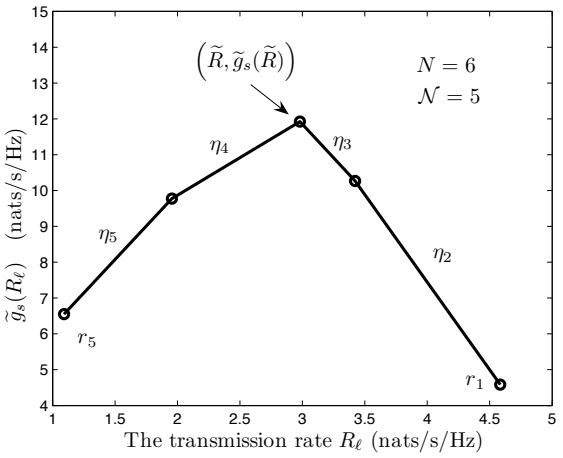

Fig. 4. An example for the function $\widetilde{g}_s(R_\ell)$ against R_ℓ, where $N = 6$ and $\boldsymbol{\gamma} = (1.98, 6.07, 18.71, 29.63, 45.43, 96.93)$.

Following previous analyses in this section, we formulate problem $\boldsymbol{P3}$ to derive the adaptation policy for multi-layer video multicast as follows:

Problem 3: $\boldsymbol{P3}$

$$\min_{(\boldsymbol{t},\boldsymbol{R},\boldsymbol{z})} \left\{\sum_{\ell=1}^{L} \mathbb{E}_{\boldsymbol{\gamma}}\left\{t_\ell\right\}\right\} \tag{42}$$

$$\text{s.t.:}\ \ \mathbb{E}_{\boldsymbol{\gamma}}\left\{e^{-\beta_\ell(z_\ell + t_\ell R_\ell)}\right\} - V_\ell \le 0,\ z_\ell \ge 0,\ \forall\, \ell, \tag{43}$$

$$N\left(1 - q_{\text{th}}^{(\ell)}\right)\mathbb{E}_{\boldsymbol{\gamma}}\left\{t_\ell R_\ell + z_\ell\right\} - \mathbb{E}_{\boldsymbol{\gamma}}\left\{t_\ell \widetilde{g}_{\text{s}}(R_\ell)\right\} \le 0, \forall\, \ell, \tag{44}$$

$$\sum_{\ell=1}^{L} t_\ell - 1 \le 0,\ 0 \le t_\ell \le 1, \quad \forall\, \boldsymbol{\gamma}, \tag{45}$$

where $\boldsymbol{R} \triangleq (R_1, R_2, \ldots, R_L)$ and Eq. (44) is the group-loss-rate constraint (equivalent to Eq. (37)) for the policies corresponding to $\widetilde{g}_s(R_\ell)$.

B. Derivation of the Optimal Solution for Multicast Video

Notice that Problem $\boldsymbol{P3}$ is not convex, because the functions on the left-hand side of Eqs. (43) and (44) are not convex over $(\boldsymbol{t}, \boldsymbol{R}, \boldsymbol{z})$. However, we show in the following that the optimal solution can still be obtained through Lagrangian dual problem.

B.1 Lagrangian Characterization of Problem $\boldsymbol{P3}$

The Lagrangian function of $\boldsymbol{P3}$, denoted by W, is constructed as

$$W = \mathbb{E}_{\boldsymbol{\gamma}}\{w\} \tag{46}$$

where

$$w \triangleq \sum_{\ell=1}^{L} t_\ell + \psi_{\boldsymbol{\gamma}} \left(\sum_{\ell=1}^{L} t_\ell - 1 \right) + \sum_{\ell=1}^{L} \lambda_\ell \Big(e^{-\beta_\ell (z_\ell + t_\ell R_\ell)} - V_\ell \Big) + \sum_{\ell=1}^{L} \phi_\ell \Big(N \left(1 - q_{\text{th}}^{(\ell)} \right) (t_\ell R_\ell + z_\ell) - t_\ell \widetilde{g}_{\text{s}}(R_\ell) \Big). \quad (47)$$

In Eq. (47), $\psi_{\boldsymbol{\gamma}} \geq 0$, $\phi_\ell \geq 0$, and $\lambda_\ell \geq 0$ are the Lagrangian multipliers associated with the constraints given by Eqs. (45), (44), and (43), respectively. The Lagrangian dual function, denoted by U, is then determined by

$$U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \triangleq \min_{(\boldsymbol{t}, \boldsymbol{z}, \boldsymbol{R})} \Big\{ W \Big\} = \mathbb{E}_{\boldsymbol{\gamma}} \Big\{ u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \Big\}, \quad (48)$$

where $\boldsymbol{\phi} \triangleq (\phi_1, \phi_2, \ldots, \phi_L)$ and

$$u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \triangleq \min_{(\boldsymbol{t}, \boldsymbol{z}, \boldsymbol{R})} \Big\{ w \Big\}. \quad (49)$$

It is clear that $u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})$ is a concave function over $(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})$, and so is $U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})$ Moreover, the Lagrange dual problem is defined as:

$$\boldsymbol{P3}\textbf{-Dual}: \; U^\star \triangleq U(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star, \psi_{\boldsymbol{\gamma}}^\star) = \max_{(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}})} \Big\{ U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \Big\}, \quad (50)$$

where $\left(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star, \psi_{\boldsymbol{\gamma}}^\star\right)$ is the maximizer. We then solve for the optimal adaptation strategy to $\boldsymbol{P3}$ through the dual problem. If the optimal solution to $\boldsymbol{P3}$ exists, we will show later that there is no duality gap between the primal problem $\boldsymbol{P3}$ and the dual problem $\boldsymbol{P3}$-**Dual**. As a result, the optimal solution to $\boldsymbol{P1}$ must minimize the Lagrangian function W under the optimal Lagrangian multipliers $\left(\boldsymbol{\lambda}^\star, \boldsymbol{\phi}^\star, \psi_{\boldsymbol{\gamma}}^\star\right)$.

B.2 Derivation of the Lagrangian Dual Function $U(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) = \mathbb{E}_{\boldsymbol{\gamma}} \{ u(\boldsymbol{\lambda}, \boldsymbol{\phi}, \psi_{\boldsymbol{\gamma}}) \}$

Since $\widetilde{g}_{\text{s}}(R_\ell)$ is nondifferentiable at some R_ℓ (as shown in Fig. 4), w is also nondifferentiable at some R_ℓ. Alternatively, we need to use the subgradient and subdifferential [19] instead of gradient to derive the minimizer to Eq. (49), which is denoted by $(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)$.

Definition 1: Consider a convex function $h : \mathcal{D} \rightarrow \mathbb{R}$, where $\mathbb{R}$ denotes the set of real numbers and $\mathcal{D} \subset \mathbb{R}^n$ is a convex set. Then, an $n \times 1$ vector $\boldsymbol{\xi}$, for $\boldsymbol{\xi} \in \mathbb{R}^n$, is called a *subgradient* [19] at $\mathbf{d} \in \mathcal{D}$ if $h(\mathbf{d}') \geq h(\mathbf{d}) + \boldsymbol{\xi}^{\text{T}}(\mathbf{d}' - \mathbf{d})$ for all $\mathbf{d}' \in \mathcal{D}$, where $(\cdot)^{\text{T}}$ represents the transpose. The collection of subgradients at $\mathbf{d}$ form a set called the *subdifferential* [19] of $h(\cdot)$ at $\mathbf{d}$, denoted by $\partial h(\mathbf{d})$. If $h(\cdot)$ is differentiable at $\mathbf{d}$, the subgradient at $\mathbf{d}$ is unique and becomes the gradient. Moreover, the sufficient and necessary condition that $\mathbf{d}^*$ minimizes $h(\mathbf{d})$ is that $\mathbf{0} \in \partial h(\mathbf{d}^*)$. When $h(\cdot)$ is a *concave* function, the subgradient and subdifferential at $h(\mathbf{d})$ is defined in the similar way as the convex case, except that the required inequality becomes $h(\mathbf{d}') \leq h(\mathbf{d}) + \boldsymbol{\xi}^{\tau}(\mathbf{d}' - \mathbf{d})$.

Then, using Eq. (40) and Definition 1, we obtain

$$\partial \widetilde{g}_{\text{s}}(r) = \begin{cases} \big[\eta_{\mathcal{N}}, \infty\big], & \text{if } r = r_{\mathcal{N}}; \\ \big\{\eta_i\big\}, & \text{if } r_i < r < r_{i-1}, \; 2 \leq i \leq \mathcal{N}; \\ \big[\eta_i, \eta_{i+1}\big], & \text{if } r = r_i, \; 2 \leq i \leq \mathcal{N} - 1; \\ \big[-\infty, \eta_1\big], & \text{if } r = r_1. \end{cases} \quad (51)$$

Applying Eq. (47) into Definition 1, we get the subdifferential of w w.r.t. R_ℓ, denoted by ∂w_{R_ℓ}, as:

$$\partial w_{R_\ell} = \Big\{ y \Big| y = t_\ell \Big(\phi_\ell N \left(1 - q_{\text{th}}^{(\ell)} \right) - \lambda_\ell \beta_\ell e^{-\beta_\ell (z_\ell + t_\ell R_\ell)} - \phi_\ell x \Big), \; \forall\, x \in \partial \widetilde{g}_{\text{s}}(R_\ell) \Big\}, \; \forall\, \ell, \boldsymbol{\gamma}. \quad (52)$$

It is clear that w is differentiable w.r.t. t_ℓ and R_ℓ. Taking the derivative of w w.r.t. t_ℓ and z_ℓ, respectively, we get

$$\frac{\partial w}{\partial t_\ell} = 1 + \psi_{\boldsymbol{\gamma}} - \lambda_\ell \beta_\ell R_\ell e^{-\beta_\ell (z_\ell + t_\ell R_\ell)} - \phi_\ell \widetilde{g}_{\text{s}}(R_\ell) + \phi_\ell N \left(1 - q_{\text{th}}^{(\ell)} \right) R_\ell, \; \forall\, \ell, \boldsymbol{\gamma}; \quad (53)$$

$$\frac{\partial w}{\partial z_\ell} = \phi_\ell N \left(1 - q_{\text{th}}^{(\ell)} \right) - \lambda_\ell \beta_\ell e^{-\beta_\ell (z_\ell + t_\ell R_\ell)}, \; \forall\, \ell, \boldsymbol{\gamma}. \quad (54)$$

Clearly, the minimization of w can be performed separately for each video layer. Now consider the ℓth video layer. Since the function w is not convex over the 3-tuple (t_ℓ, z_ℓ, R_ℓ), the equations $\partial w / t_\ell = 0$, $\partial w / z_\ell = 0$, and $0 \in \partial w_{R_\ell}$ are only the necessary conditions for $(t_\ell^*, z_\ell^*, R_\ell^*)$. However, if t_ℓ^* is given, w becomes a convex function over the 2-tuple (z_ℓ, R_ℓ). Using this property, we can decompose the minimization of w into several easier sub-problems. Applying the above principle, we discuss the cases with the fixed $t_\ell^* = 0$ and $t_\ell^* > 0$, respectively, as follows.

1) $t_\ell^* = 0$: The variable R_ℓ vanishes in Eq. (47) when $t_\ell = 0$. Then, we only need to find the minimizer z_ℓ^*. By solving $\partial w / \partial z_\ell = 0$ under the condition $z_\ell \geq 0$, we get

$$z_\ell^* = \left[-\tfrac{1}{\beta_\ell} \log \left(\frac{\phi_\ell N \left(1 - q_{\text{th}}^{(\ell)} \right)}{\lambda_\ell \beta_\ell} \right) \right]^+. \quad (55)$$

2) $t_\ell^* > 0$: We jointly solve $\partial w / \partial z_\ell = 0$ and $0 \in \partial w_{R_\ell}$ under the condition $z_\ell \geq 0$ and $R_\ell \geq 0$, and then get the minimizer, which is summarized in Eqs. (56)-(57) as follows:

if $\widehat{R}_\ell > \widetilde{R}$, then

$$\begin{cases} R_\ell^* = \widetilde{R}; \\ z_\ell^* = \left[-\tfrac{1}{\beta_\ell} \log \left(\frac{\phi_\ell N \left(1 - q_{\text{th}}^{(\ell)} \right)}{\lambda_\ell \beta_\ell} \right) - t_\ell^* R_\ell^* \right]^+; \end{cases} \quad (56)$$

if $\widehat{R}_\ell \leq \widetilde{R}$, then

$$\begin{cases} R_\ell^* = \widehat{R}_\ell; \\ z_\ell^* = 0, \end{cases} \quad (57)$$

where $\widehat{R}_\ell$ is the unique solution to

$$0 \in \left(\partial w_{R_\ell} \right) |_{z_\ell = 0} \quad (58)$$

under the given t_ℓ, and

$$\widetilde{R} \triangleq \arg\max_r \big\{ \widetilde{g}_{\text{s}}(r) \big\}. \quad (59)$$

The detailed derivations for Eqs. (56) and (57) are omitted due to lack of space, but are provided online in [24]. Note that $\widetilde{R}$ depends only on $\widetilde{g}_{\text{s}}(r)$, but not on t_ℓ^*. Then, through Eqs. (56)-(57), we can see that with $t_\ell^* > 0$, the minimizer must satisfy either $z_\ell^* = 0$ or $R_\ell^* = \widetilde{R}$. Further note that the above results provide not only the mathematical convenience, but also the insightful observations for the adaptive multicast transmission. For multicast, zero loss can be achieved only

when setting transmission rate $R_\ell = R_{\pi(N)}$, which is determined by the worst-case SNR over all multicast receivers. The higher the transmission rate or the higher the drop rate we use, the higher the loss rate we get (observed from Eq. (38)). Therefore, when not violating the statistical delay QoS guarantees, we need to choose the transmission rate R_ℓ and the drop rate z_ℓ as small as possible. Moreover, R_ℓ and z_ℓ need to be jointly derived for performance optimization. Following the above strategies, we first derive $\widehat{R}_\ell$ which optimizes the system performance (equivalently, minimizes the Lagrangian function) given the zero drop rate. However, if $\widehat{R}_\ell > \widetilde{R}$, we can see that the achieved sum rate $\widetilde{g}_s(R_\ell)$ over all multicast receivers under $R_\ell = \widehat{R}_\ell$ decreases when $\widehat{R}_\ell$ increases, as depicted in Fig. 4. When this happens, we need to set $R_\ell = \widetilde{R}$ and apply the nonzero drop rate to avoid the degradation of $\widetilde{g}_s(R_\ell)$ while supporting the satisfied service rate.

Based on the above results for $t_\ell^* = 0$ and $t_\ell^* > 0$, the minimizer $(t_\ell^*, R_\ell^*, z_\ell^*)$ at the ℓth video layer must fall into one of the following three Sub-domains:

$$\begin{cases} \text{Sub-domain 1:} & t_\ell = 0,\ R_\ell \geq 0,\ z_\ell \geq 0; \\ \text{Sub-domain 2:} & t_\ell \geq 0,\ R_\ell = \widetilde{R},\ z_\ell \geq 0; \\ \text{Sub-domain 3:} & t_\ell \geq 0,\ R_\ell \geq 0,\ z_\ell = 0. \end{cases} \tag{60}$$

In Eq. (60), Sub-domain 1 is associated with the case of $t_\ell^* = 0$. For the case with $t_\ell^* > 0$, Sub-domains 2 and 3 correspond to the conditions $\widehat{R}_\ell \geq \widetilde{R}$ and $\widehat{R}_\ell < \widetilde{R}$, respectively. In order to get the minimizer $(t_\ell^*, R_\ell^*, z_\ell^*)$ of w, we can first find the minimizer within each Sub-domain, which is denoted by $\left(t_\ell^{(j)}, R_\ell^{(j)}, z_\ell^{(j)}\right)$, $j = 1, 2, 3$. After identifying the minimizers of each Sub-domain, $(t_\ell^*, R_\ell^*, z_\ell^*)$ can then be obtained through

$$(t_\ell^*, R_\ell^*, z_\ell^*) = \left(t_\ell^{(j^*)}, R_\ell^{(j^*)}, z_\ell^{(j^*)}\right) \quad \forall\, \ell, \tag{61}$$

where

$$j^* = \arg \min_{j=1,2,3} \left\{ w\big|_{(t_\ell, R_\ell, z_\ell) = \left(t_\ell^{(j)}, R_\ell^{(j)}, z_\ell^{(j)}\right)} \right\}.$$

In Sub-domains 1, 2, and 3, the variables t_ℓ, R_ℓ, and z_ℓ are fixed, respectively, implying that there are only two optimization variables in each Sub-domain. Therefore, the minimization problem within each Sub-domain becomes tractable. For Sub-domain 1, the minimizer $\left(0, R_\ell^{(1)}, z_\ell^{(1)}\right)$ is given in Eq. (55). For Sub-domain 2, since R_ℓ is fixed, deriving the minimizer $\left(t_\ell^{(2)}, R_\ell^{(2)}, z_\ell^{(2)}\right)$ is equivalent to solving a convex problem. For Sub-domain 3, the optimization problem can be readily solved by applying the piecewise linear property of $\widetilde{g}(R_\ell)$. The detailed derivations for $\left(t_\ell^{(2)}, R_\ell^{(2)}, z_\ell^{(2)}\right)$ and $\left(t_\ell^{(3)}, R_\ell^{(3)}, z_\ell^{(3)}\right)$ are omitted due to lack of space, but are provided online in [24].

B.3. The Optimal Solution to $\boldsymbol{P3}$.

In Section VI-B.2, we have obtained the minimizer $(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)$ for w. Then, based on the optimization theory [19], the necessary and sufficient conditions for zero duality gap are as follows: there exists the feasible policy $(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)|_{\{\phi_\gamma = \phi_\gamma^\star, \lambda_\ell = \lambda_\ell^\star, \phi_\ell = \phi_\ell^\star,\ \forall\, \ell, \gamma\}}$ such that

$$\begin{cases} \psi_{\boldsymbol{\gamma}}^\star \left(\sum_{\ell=1}^L t_\ell^* - 1 \right) = 0,\ \forall\, \ell,\ \gamma; \\ \lambda_\ell^\star \left(\mathbb{E}_{\boldsymbol{\gamma}} \left\{ e^{-\beta_\ell (z_\ell^* + t_\ell^* R_\ell^*)} \right\} - V_\ell \right) = 0,\ \forall\, \ell; \\ \phi_\ell^\star \mathbb{E}_{\boldsymbol{\gamma}} \left\{ N \left(1 - q_{\text{th}}^{(\ell)} \right) (t_\ell^* R_\ell^* + z_\ell^*) - t_\ell^* \widetilde{g}_{\text{s}}(R_\ell^*) \right\} = 0,\ \forall\, \ell; \\ \psi_{\boldsymbol{\gamma}}^\star \geq 0,\ \lambda_\ell^\star \geq 0,\ \phi_\ell^\star \geq 0. \end{cases} \tag{62}$$

Then, the optimal policy to $\boldsymbol{P3}$ is given by

$$(\boldsymbol{t}^*, \boldsymbol{z}^*, \boldsymbol{R}^*)|_{\{\phi_\gamma = \phi_\gamma^\star, \lambda_\ell = \lambda_\ell^\star, \phi_\ell = \phi_\ell^\star,\ \forall\, \ell, \gamma\}}. \tag{63}$$

We can solve for the optimal Lagrangian multipliers by using the similar arguments in the proof of Theorem 1. Specifically, for each channel realization if $\sum_{\ell=1}^L t_\ell^*|_{\psi_{\boldsymbol{\gamma}}=0} \leq 1$, we have $\psi_{\boldsymbol{\gamma}}^\star = 0$; otherwise, $\psi_{\boldsymbol{\gamma}}^\star$ is the unique solution to $\sum_{\ell=1}^L t_\ell^* = 1$. Moreover, $\{\phi_\ell^\star\}_{\ell=1}^L$ and $\{\lambda_\ell^\star\}_{\ell=1}^L$ will be selected such that "=" holds for constraints given in Eqs. (43)-(44). Furthermore, we can design the adaptive tracking method similar to problem $\boldsymbol{P1}$ to examine the existence of the optimal solution and find the optimal Lagrangian multipliers.

Note that under the above optimal solution, different $\{\gamma_n\}_{n=1}^N$, which have the same ordered permutation, will generate the same function $\widetilde{g}_s(R_\ell)$ defined by Eq. (40) and thus the same adaptation policy. Then, since γ_n's are i.i.d. (as assumed in Section II), this policy will benefit all receivers evenly, implying $q_{\ell,0} = q_{\ell,1} = q_{\ell,2} = \cdots = q_{\ell,N} = q_{\text{th}}^{(\ell)}$. Therefore, the original loss-rate constraint is not violated for all multicast receivers. Moreover, since the group-loss-rate constraint given by Eq. (44) in problem $\boldsymbol{P3}$ is a relaxed version of the original loss-rate constraint for each multicast receiver (given by Eq. (36)), the optimal solution to problem $\boldsymbol{P3}$ is also optimal even if we replace Eq. (44) by using the original loss-rate constraint given in Eq. (36).

VII. Simulation Evaluations

We use simulation experiments to evaluate the performances of our proposed optimal adaptive transmission schemes and to investigate the impact of QoS requirements on resource allocations. Note that the metric "delay" investigated/simulated in simulations represents the queueing delay, as addressed previously in Section III-B for the framework of this paper. In simulations, we set the signal bandwidth B and time-frame length T equal to 2×10^5 Hz and 10 ms, respectively. The arrival video stream includes two layers, both of which have constant arrival rates, where $A_1[k] = 250$ Kbps and $A_2[k] = 150$ Kbps. Then, the effective bandwidths of $A_1[k]$ and $A_2[k]$ are determined by $\mathcal{A}_1(\theta_1) = 1.733 \times 10^3$ nats/frame $\mathcal{A}_2(\theta_2) = 1.040 \times 10^3$ nats/frame, respectively. The values of θ_1 and θ_2 can be derived from solving Eq. (7), depending on the QoS requirements specified by D_{th} and P_{th}, $\ell = 1, 2, \ldots, L$. The wireless channel follows the Rayleigh fading model and we denote the average SNR by $\overline{\gamma}$. Fig. 5 plots the iterative on-line tracking of the optimal Lagrangian multipliers λ_ℓ^*'s based on the method used in Section V-A with $\epsilon = 0.01$ and $\alpha = 0.02$. As shown in Fig. 5, the Lagrangian multipliers quickly converge to the optimal value and oscillate slightly within the small dynamic ranges, which demonstrates the effectiveness of our tracking method.

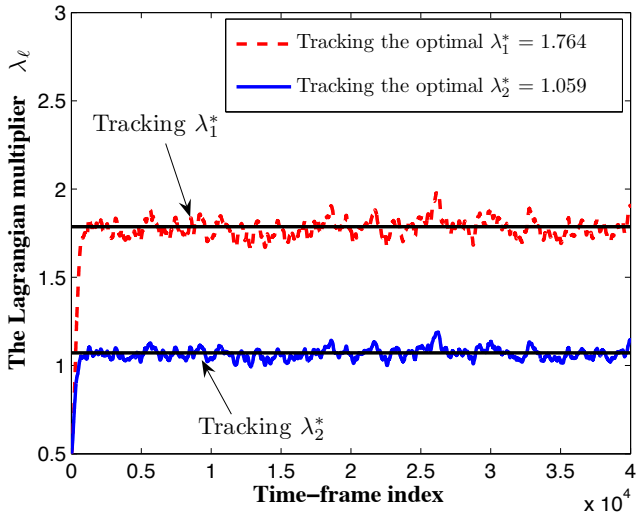

Fig. 5. Illustration of tracking the optimal Lagrangian multiplier λ_ℓ^* for unicast with zero loss, where the average SNR is $\overline{\gamma} = 10$ dB, the required delay bound is $P_{\rm th} = 10^{-4}$, and the required threshold for the delay-bound violation probability is $D_{\rm th} = 250$ ms.

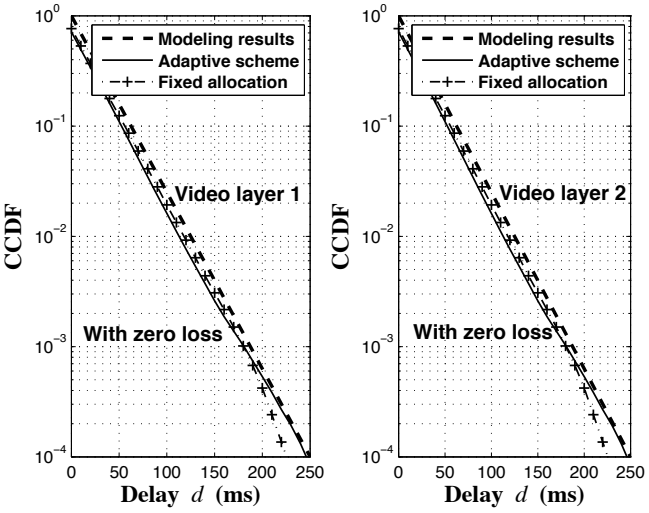

Fig. 6. The complementary cumulative distribution function (CCDF) $\Pr\{D_\ell > d\}$ of the queueing delay for the unicast scenario with zero loss for video layer 1 and video layer 2, respectively, where $\overline{\gamma} = 15$ dB, $P_{\rm th} = 10^{-4}$, and the required delay-bound is $D_{\rm th} = 250$ ms.

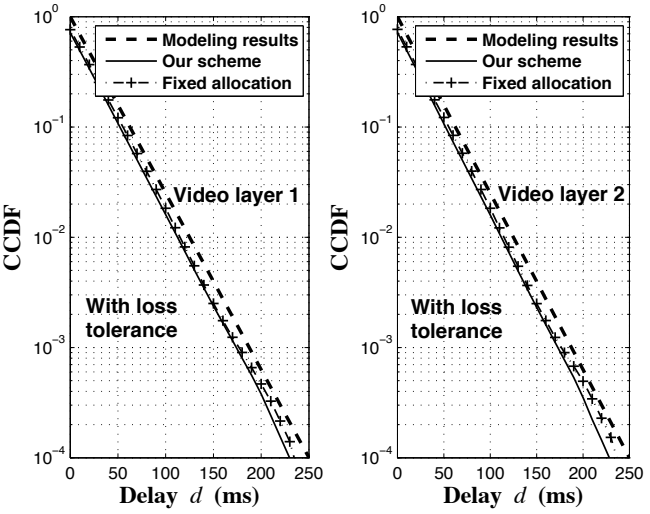

Fig. 7. The complementary cumulative distribution function $\Pr\{D_\ell > d\}$ of the queueing delay for the unicast scenario with loss tolerance for video layer 1 and video layer 2, respectively, where $\overline{\gamma} = 15$ dB, $P_{\rm th} = 10^{-4}$, $D_{\rm th} = 250$ ms, and $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.02)$.

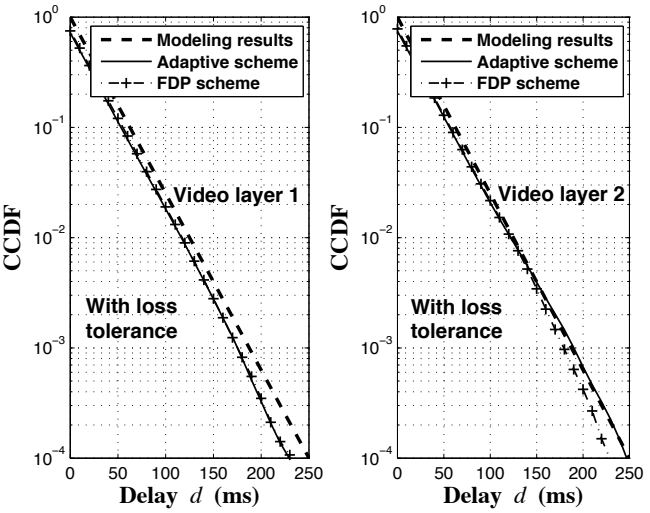

Fig. 8. The complementary cumulative distribution function $\Pr\{D_\ell > d\}$ of the queueing delay for multicast scenario with loss tolerance for video layer 1 and video layer 2, respectively, where $N = 20$ receivers, $\overline{\gamma} = 20$ dB, $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.05)$, $P_{\rm th} = 10^{-4}$, and $D_{\rm th} = 250$ ms.

We also investigate some straightforward time-slot allocation schemes as the baseline schemes for comparative analyses. We will compare the average resource consumption between our derived optimal schemes and these baseline schemes under the same QoS satisfactions.

1) Fixed time-slot allocation for unicast without loss: This scheme uses constant time-slot length $t_\ell[k] = \overline{t}_\ell$, $k = 1, 2, \ldots$, in all fading states. The normalized transmission rate is set to $R_\ell = \log(1+\gamma)$ nats/s/Hz for the ℓth video layer. The parameters $\overline{t}_\ell$, $\ell = 1, 2$, are selected such that the effective capacity $\mathcal{C}_\ell(\theta_\ell)$ of the ℓth video layer's service process is just equal to $\overline{C}_\ell$:

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell}\log\left(\mathbb{E}\left\{e^{-\theta_\ell \overline{t}_\ell BT\log(1+\gamma)}\right\}\right) = \overline{C}_\ell, \qquad (64)$$

where $\overline{C}_\ell = \mathcal{A}_\ell(\theta_\ell)$ and $\ell = 1, 2, \ldots, L$. We can obtain the unique $\overline{t}_\ell$'s by numerically solving the above equation. If we get $\sum_{\ell=1}^{L} \overline{t}_\ell > 1$, this scheme cannot guarantee the QoS requirements under current channel conditions, even using up all time-slot resources.

2) Fixed time-slot allocation for unicast with loss tolerance: This scheme uses both the constant time-slot length $t_\ell[k] = \overline{t}_\ell$ and the constant per-drop rate $z_\ell[k] = \overline{z}_\ell$ in all fading states. The normalized transmission rate is also set to $R_\ell = \log(1+\gamma)$ nats/s/Hz for the ℓth video layer. The parameters $\overline{t}_\ell$ and $\overline{z}_\ell$ can be obtained by solving

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell}\log\left(\mathbb{E}\left\{e^{-\theta_\ell\left(\overline{t}_\ell BT\log(1+\gamma) + BT\overline{z}_\ell\right)}\right\}\right) = \overline{C}_\ell \qquad (65)$$

and $q_\ell = q_{\rm th}^{(\ell)}$ for all video layers, where $\ell = 1, 2, \ldots, L$.

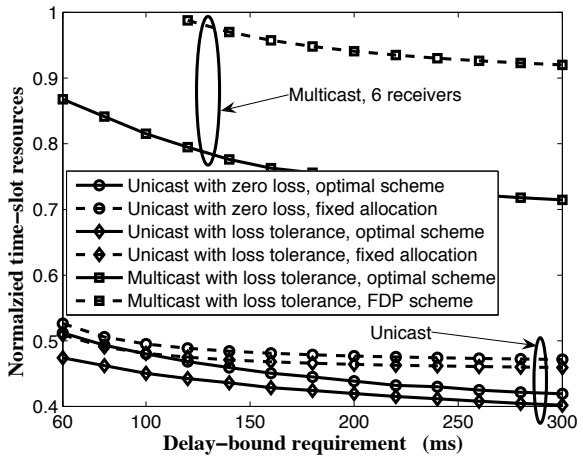

Fig. 9. The normalized time-slot resource consumption $\mathbb{E}_{\gamma}\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus delay-bound requirement $D_{\rm th}$, where $\overline{\gamma} = 15$ dB, $P_{\rm th} = 10^{-4}$, and $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.05)$.

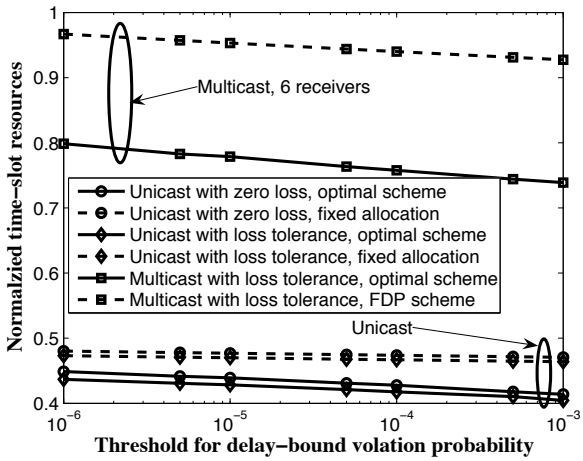

Fig. 10. The normalized time-slot resource consumption $\mathbb{E}_{\gamma}\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus the threshold $P_{\rm th}$ for the delay-bound violation probability, where $\overline{\gamma} = 15$ dB, $D_{\rm th} = 250$ ms, and $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right) = (0.01, 0.02)$.

3) Fixed dominating position scheme for multicast with loss tolerance: The fixed dominating position (FDP) scheme always sets $R_\ell = \log(1 + \gamma_{\pi(i_\ell)})$ nats/s/Hz, where $\gamma_{\pi(i)}$ denotes the ith largest instantaneous SNR among all multicast receivers. The index i_ℓ is fixed at $i_\ell = \left\lceil N\left(1 - q_{\rm th}^{(\ell)}\right)\right\rceil$ such that the loss-rate QoS is not violated. Moreover, the FDP scheme also adopts the constant time-slot length $t_\ell[k] = \overline{t}_\ell$ and the constant per-drop rate $z_\ell[k] = \overline{z}_\ell$, which can be obtained by solving

$$\mathcal{C}_\ell(\theta_\ell) = -\frac{1}{\theta_\ell}\log\left(\mathbb{E}\left\{e^{-\theta_\ell\left(\overline{t}_\ell BT\log(1+\gamma_{\pi(i_\ell)})+BT\overline{z}_\ell\right)}\right\}\right) = \overline{C}_\ell \tag{66}$$

and $q_{\ell,0} = q_{\rm th}^{(\ell)}$ for all video layers, where $\ell = 1, 2, \ldots, L$.

Figures 6, 7, and 8 depict the complementary cumulative distribution function (CCDF) of the queueing delay, i.e., the probability $\Pr\{D_\ell > d\}$ given a threshold d, for unicast with zero loss, unicast with loss tolerance, and multicast with loss tolerance, respectively. We can observe from Figs. 6-8 that the CCDF's of all schemes agree well with the modeling results (see Eqs. (7)-(8)) at each video layer, where the delay-bound violation probability decreases exponentially against the delay bound. Moreover, for the required delay bound $D_{\rm th} = 250$ ms, the violation probability of all schemes can be upper-bounded by the targeted $P_{\rm th} = 10^{-4}$ for each video layer, which demonstrates the validity of all schemes in terms of statistical delay-QoS guarantees. Having shown that all schemes can meet the same QoS requirements, we then focus on the performance of the average time-slot consumption.

Figures 9 and 10 illustrate the impact of delay-bound $D_{\rm th}$ and its violation probability $P_{\rm th}$ on the resource consumption, respectively. As shown in Figs. 9 and 10, either smaller $D_{\rm th}$ or $P_{\rm th}$ will cause more resource consumption, which is expected because the smaller $D_{\rm th}$ or $P_{\rm th}$ implies the more stringent delay QoS requirement. Moreover, under various QoS conditions, our proposed optimal schemes always use much less wireless resources than the baseline schemes. For multicast services, our derived optimal scheme consumes at least 15% of total resources less than the FDP scheme. For unicast services, more resources can be saved by using the optimal scheme when delay QoS becomes looser, while less resources are saved under the more stringent QoS constraints.

After demonstrating the superiority of our proposed optimal schemes over the baseline schemes, Fig. 11(a) plots the average time-slot consumption versus the average SNR for multi-layer video unicast and multicast. We observe from Fig. 11(a) that under the same channel conditions, video unicast uses much less time-slot resources than multicast. When the average SNR is relatively low around 7-8 dB, video unicast does not need to consume all available time-slot resources. In contrast, video multicast almost uses up *all* resources even with $\overline{\gamma} = 13.5$ dB for the 6-receiver case. For the 10-receiver case, the average SNR needs to be larger than 15 dB to provide the QoS-guaranteed multicast services. The above observations reflect the key challenges on wireless multicast. In wireless broadcast channels, since all receivers can hear the sender, it is ideal that only one copy of data is transmitted such that sizable resources can be saved. However, due to the heterogeneous fading channels across multicast receivers, the transmission rate has to be limited within the relatively low range to avoid too much data loss for receivers with poorer instantaneous channel qualities. As a result, more time-slot resources are consumed to meet the QoS requirements. In addition, more multicast receivers result in more resource consumption, as depicted in Fig. 11(a). But note that although the wireless multicast faces many challenges, it still uses much less wireless resources than the strategy which uses multiple unicast links to implement wireless multicast. For example, if using multiple unicast links to implement multicast, we need the time-slot resources at least N times as much as the resource consumption for a unicast link. Clearly, for environments simulated in Fig. 11(a), even with $\overline{\gamma} = 18$ dB, we still do not have enough resources for such a unicast-based multicast scheme with just 6 receivers. Fig. 11(b) shows the resource consumption for each video layer. We can see that video layer 1 requires more resources

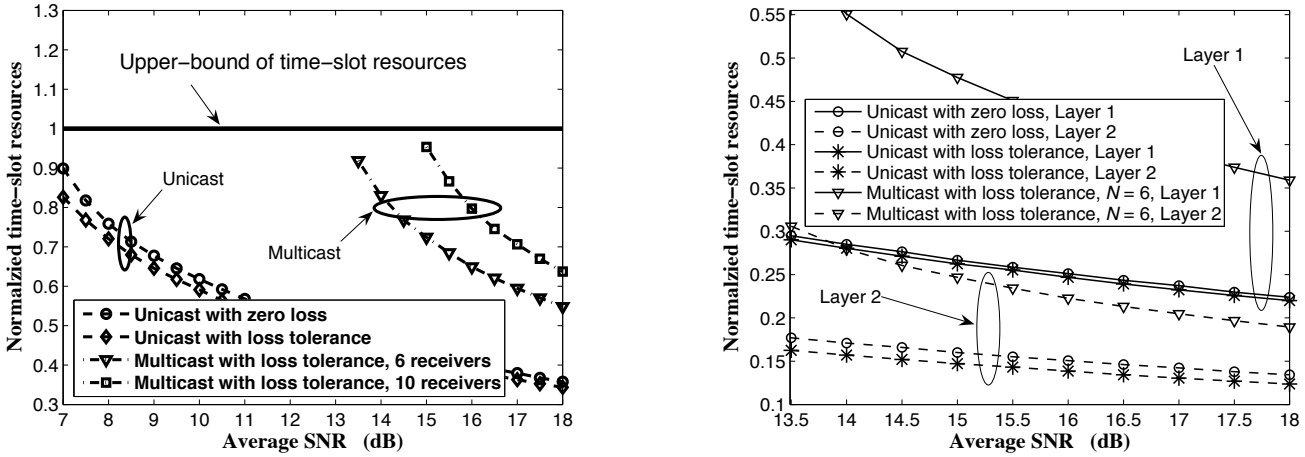

Fig. 11. (a) The normalized time-slot resource consumption $\mathbb{E}_\gamma\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus $\overline{\gamma}$, where $P_{\rm th}=10^{-4}$, $D_{\rm th}=250$ ms, $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right)=(0.01, 0.05)$. (b) $\mathbb{E}_\gamma\{t_\ell\}$ for each video layer versus $\overline{\gamma}$, where $P_{\rm th}=10^{-4}$, $D_{\rm th}=250$ ms, $\left(q_{\rm th}^{(1)}, q_{\rm th}^{(2)}\right)=(0.01, 0.05)$.

than video layer 2, which is because in our settings the traffic load of layer 1 is higher and the loss-rate QoS of layer 1 is more stringent. Furthermore, in multicast the difference of resource consumption between the two video layers is larger than the difference in unicast.

Figure 12 shows the impact from loss-rate constraints on video multicast. As shown in Fig. 12, even slightly increasing $q_{\rm th}^{(\ell)}$ can significantly reduce the total consumed wireless resources. This is because the higher loss-tolerance level will enable larger multicast transmission rate and thus consume less time-slot resources. The above observations suggest that there exists a tradeoff between loss-rate control at the physical layer and error recovery at the upper protocol layers. As mentioned previously, the loss-rate $q_{\rm th}^{(\ell)}$ depends largely on the capability of erasure-correction codes used at upper protocol layers, especially for multicast services. Thus, Fig. 12 suggests that using more redundancy for forward error-control at upper protocol layers can effectively decrease the total wireless-resource consumption with QoS guarantees, while enabling the repair of more data losses at the physical layer.

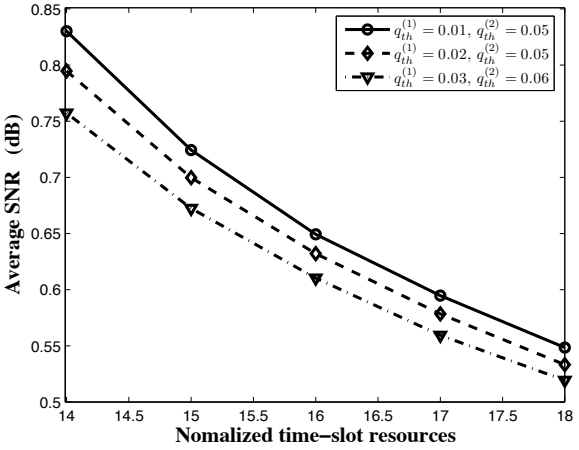

Fig. 12. The normalized time-slot resource consumption $\mathbb{E}_\gamma\left\{\sum_{\ell=1}^{L} t_\ell\right\}$ versus $\overline{\gamma}$ for multicast with $N=6$, $P_{\rm th}=10^{-4}$, and $D_{\rm th}=250$ ms.

VIII. Conclusions

We proposed a framework to model the wireless transmission of multi-layer video stream with statistical delay QoS guarantees. A separate queue is maintained for each video layer and the same statistical delay QoS-requirement needs to be satisfied by all video layers, where the statistical delay QoS is characterized by the delay-bound and its corresponding violation probability through the effective capacity bandwidth/capacity theory. Under the proposed framework, we derived a set of optimal rate adaptation and time-slot allocation schemes for video unicast/multicast with and/or without loss tolerance, which minimizes the time-slot resource consumption. We also conducted extensive simulation experiments to demonstrate the impact of statistical QoS provisionings on wireless resource allocations by using our derived optimal adaptive transmission schemes.

References

[1] J. Shin, J.-W. Kim, C.-C.J. Kuo, "Quality-of-service mapping mechanism for packet video indifferentiated services network", *IEEE Trans. Multimedia*, vol. 3, no. 2, pp. 291-231, Jun. 2001.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103-1120, Sep. 2007.

[3] S. McCanne, M. Vetterli, and V. Jacobson, "Low-complexity video coding for receiver-driven layered multicast," *IEEE J. Sel. Areas Commun.*, vol. 15, no. 6, pp. 983-1001, Aug. 1997.

[4] J. Villalon, P. Cuenca, L. L. Orozco-Barbosa, Y. Seok, and T. Turletti, "Cross-layer architecture for adaptive video multicast streaming over mutlirate wireless LANs,", *IEEE J. Sel. Area. Commun.*, vol. 25, no. 4, pp. 699-711, May 2007.

[5] X. Zhang and K. G. Shin, D. Saha, and D. Kandlur, "Scalable flow control for multicast ABR services in ATM networks," *IEEE/ACM Trans. on Netw.*, vol. 10, no. 1, pp. 67-85, Feb. 2002.

[6] X. Zhang and K. G. Shin, "Delay analysis of feedback-synchronization signaling for multicast flow control," *IEEE/ACM Trans. on Netw.*, vol. 11, no.3, pp. 436-460, Jun. 2003.

[7] X. Zhang and K. G. Shin, "Markov-chain modeling for multicast signaling delay analysis," *IEEE/ACM Trans. on Netw.*, vol. 12, no. 4, pp. 667-680, Aug. 2004.

[8] X. Zhang and Q. Du, "Cross-layer modeling for QoS-driven multimedia multicast/broadcast over fading channels," *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 62-70, Aug. 2007.
[9] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Auto. Control*, vol 39, no. 5, pp. 913-931, May 1994.
[10] C.-S. Chang, *Performance Guarantees in Communication Networks*, Springer-Verlag London, 2000.
[11] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of Quality of Service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630-643, Jul. 2003.
[12] X. Zhang, J. Tang, H.-H. Chen, S. Ci, and M. Guizni, "Cross-layer-based modeling for quality of service guarantees in mobile wireless networks," *IEEE Commun. Mag.*, pp. 100-106, Jan. 2006.
[13] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. Wireless Commun.*, vol. 6, no. 8, pp. 3058-3068, Aug. 2007.
[14] Q. Du and X. Zhang, "Effective capacity of superposition coding based mobile multicast in wireless networks," in *Proc. IEEE International Conference on Commun., (ICC'09)*, Dresden, Germany, Jun. 2009.
[15] Q. Du and X. Zhang, "Cross-layer design based rate control for mobile multicast in cellular networks," in *Proc. IEEE GLOBECOM 2007*, Washington DC, USA, Nov. 26-30, 2007, pp. 5180-5184.
[16] J. Nonenmacher, E. Biersack, and D. Towsley, "Parity-based loss recovery for reliable multicast transmission," *IEEE/ACM Trans. Networking*, vol. 6, no. 4, pp. 349-361, Aug. 1998.
[17] J. W. Byers, M. Luby, and M. Mitzenmacher, "A digital Fountain approach to asynchronous reliable multicast." *IEEE J. Select. Areas Commun.*, vol. 20, no. 8, pp. 1528-1540, Oct. 2002.
[18] X. Zhang and Q. Du, "Adaptive Low-Complexity Erasure-Correcting Code Based Protocols for QoS-Driven Mobile Multicast Services Over Wireless Networks," *IEEE Trans. Veh. Technol.*, vol. 55, no. 5, pp. 1633-1647, Sep. 2006.
[19] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*, 3rd ed., John Wiley & Sons, Inc., 2006.
[20] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
[21] Stephen Boyd, Lin Xiao, and Almir Mutapcic, *Subgradient Methods*, http://www.stanford.edu/class/ee392o/subgrad_method.pdf.
[22] G. L. Choudhury, D. M. Lucantoni, and W. Whitt, "Squeezing the most out of ATM," *IEEE Trans. Commun.*, vol. 44, pp. 203-217, Feb. 1996.
[23] T.H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. Cambridge, Mass. and New York: MIT Press and McGraw-Hill, 2001.
[24] Q. Du and X. Zhang, "Statistical QoS provisionings for wireless unicast/multicast of multi-layer video streams," *Networking and Information Systems Labs., Dept. Electr. and Comput. Eng., Texas A&M Univ., College Station, Tech. Rep.* [Online.] Available: http://www.ece.tamu.edu/∼xizhang/papers/multi_layer_video.pdf.

Qinghe Du [S'09] received B.S. and M.S. from Xian Jiaotong University, and is currently working towards to the Ph.D. degree under supervising of Prof. Xi Zhang in Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering at Texas A&M University.

His research interests include mobile wireless communications and networks with emphasis on mobile multicast, statistical QoS provisioning, QoS-driven resource allocations, cognitive radio techniques, and cross-layer design over wireless networks. His work co-authored with his Ph.D. advisor Prof. Xi Zhang received the Best Paper Award in the IEEE Globecom 2007 for the paper "Cross-Layer Design Based Rate Control for Mobile Multicast in Cellular Networks". He has published multiple papers in IEEE Transactions, IEEE J-SAC, IEEE Comm Magazine, IEEE INFOCOM, IEEE Globecom, IEEE ICC, etc.

Xi Zhang [S'89-SM'98] received the B.S. and M.S. degrees from Xidian University, Xi'an, China, the M.S. degree from Lehigh University, Bethlehem, PA, all in electrical engineering and computer science, and the Ph.D. degree in electrical engineering and computer science (Electrical Engineering-Systems) from The University of Michigan, Ann Arbor.

He is currently an Associate Professor and the Founding Director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station. He was an Assistant Professor and the Founding Director of the Division of Computer Systems Engineering, Department of Electrical Engineering and Computer Science, Beijing Information Technology Engineering Institute, China, from 1984 to 1989. He was a Research Fellow with the School of Electrical Engineering, University of Technology, Sydney, Australia, and the Department of Electrical and Computer Engineering, James Cook University, Australia, under a Fellowship from the Chinese National Commission of Education. He worked as a Summer Intern with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hills, NJ, and with AT&T Laboratories Research, Florham Park, NJ, in 1997. He has published more than 170 research papers in the areas of wireless networks and communications systems, mobile computing, network protocol design and modeling, statistical communications, random signal processing, information theory, and control theory and systems.

Prof. Zhang received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He received the Best Paper Awards in the IEEE Globecom 2009 and the IEEE Globecom 2007, respectively. He also received the TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering at Texas A&M University, College Station, in 2006. He is currently serving as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, an Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS for the special issue on "wireless video transmissions", an Associate Editor for the IEEE COMMUNICATIONS LETTERS, a Guest Editor for the IEEE WIRELESS COMMUNICATIONS MAGAZINE for the special issue on "next generation of CDMA versus OFDMA for 4G wireless applications", an Editor for the JOHN WILEY'S JOURNAL ON WIRELESS COMMUNICATIONS AND MOBILE COMPUTING, an Editor for the JOURNAL OF COMPUTER SYSTEMS, NETWORKING, AND COMMUNICATIONS, an Associate Editor for the JOHN WILEY'S JOURNAL ON SECURITY AND COMMUNICATIONS NETWORKS, an Area Editor for the ELSEVIER JOURNAL ON COMPUTER COMMUNICATIONS, and a Guest Editor for JOHN WILEY'S JOURNAL ON WIRELESS COMMUNICATIONS AND MOBILE COMPUTING for the special issue on "next generation wireless communications and mobile computing". He has frequently served as the Panelist on the U.S. National Science Foundation Research-Proposal Review Panels. He is serving or has served as the Technical Program Committee (TPC) Chair for IEEE Globecom 2011, TPC Vice-Chair for IEEE INFOCOM 2010, TPC Co-Chair for IEEE INFOCOM 2009 Mini-Conference, TPC Co-Chair for IEEE Globecom 2008 - Wireless Communications Symposium, TPC Co-Chair for the IEEE ICC 2008 - Information and Network Security Symposium, Symposium Chair for IEEE/ACM International Cross-Layer Optimized Wireless Networks Symposium 2006, 2007, and 2008, respectively, the TPC Chair for IEEE/ACM IWCMC 2006, 2007, and 2008, respectively, the Demo/Poster Chair for IEEE INFOCOM 2008, the Student Travel Grants Co-Chair for IEEE INFOCOM 2007, the General Chair for ACM QShine 2010, the Panel Co-Chair for IEEE ICCCN 2007, the Poster Chair for IEEE/ACM MSWiM 2007 and IEEE QShine 2006, Executive Committee Co-Chair for QShine, the Publicity Chair for IEEE/ACM QShine 2007 and IEEE WirelessCom 2005, and the Panelist on the Cross-Layer Optimized Wireless Networks and Multimedia Communications at IEEE ICCCN 2007 and WiFi-Hotspots/WLAN and QoS Panel at IEEE QShine 2004. He has served as the TPC members for more than 70 IEEE/ACM conferences, including IEEE INFOCOM, IEEE Globecom, IEEE ICC, IEEE WCNC, IEEE VTC, IEEE/ACM QShine, IEEE WoWMoM, IEEE ICCCN, etc.

Prof. Zhang is a Senior Member of the IEEE and a Member of the Association for Computing Machinery (ACM).