A Cross-Layer Quality-of-Service Mapping Architecture for Video Delivery in Wireless Networks

Wuttipong Kumwilaisak, *Student Member, IEEE*, Y. Thomas Hou, *Member, IEEE*, Qian Zhang, *Member, IEEE*, Wenwu Zhu, *Senior Member, IEEE*, C.-C. Jay Kuo, *Fellow, IEEE*, and Ya-Qin Zhang, *Fellow, IEEE*

***Abstract*—Providing quality-of-service (QoS) to video delivery in wireless networks has attracted intensive research over the years. A fundamental problem in this area is how to map QoS criterion at different layers and optimize QoS across the layers. In this paper, we investigate this problem and present a cross-layer mapping architecture for video transmission in wireless networks. There are several important building blocks in this architecture, among others, QoS interaction between video coding and transmission modules, QoS mapping mechanism, video quality adaptation, and source rate constraint derivation. We describe the design and algorithms for each building block, which either builds upon or extend the state-of-the-art algorithms that were developed without much considerations of other layers. Finally, we use simulation results to demonstrate the performance of the proposed architecture for progressive fine granularity scalability video transmission over time-varying and nonstationary wireless channel.**

***Index Terms*—Channel capacity, effective capacity, quality-of-service (QoS) mapping, QoS, scalable video, video adaptation, wireless networks.**

I. Introduction

With the development of third-generation (3G) [4], [10], [13] and fourth-generation (4G) [2] wireless standards, new broadband video applications can be offered to mobile users. In addition to delivering high bit rate video applications, 3G and 4G systems are also expected to provide multiple quality-of-service (QoS) guarantees to different types of user applications. For example, the packet-switched connection in the Universal Mobile Telecommunications System (UMTS) provides four different services differentiated by delay sensitivity: conversational, streaming, interactive, and background classes [10]. An important issue in providing multiple QoS guarantees to video applications in wireless systems is dynamic QoS management for services with mobility support [2]. A dynamic QoS management system allows video applications and the underlying prioritized transmission system to interact with each other in order to cope with service degradation and resource constraint in a time-varying wireless environment [2], [9].

Being different from wired networks, wireless networks typically have time-varying and nonstationary links due to the following factors: 1) fading effects coming from path loss, large-scale fading, and small scale fading [15]; 2) roaming between heterogeneous mobile networks [e.g., from wireless local-area network (LAN) to wireless wide-area network (WAN)]; and 3) the variation in mobile speed, average received power, and surrounding environments [12], [15]. Consequently, the quality of wireless link varies, which can be measured by the variation of the signal-to-noise ratio (SNR) or the bit-error rate (BER). These variations result in time-varying available transmission bandwidth at the link layer (also called the channel service rate [3], [31]), which also leads to time-varying delay of arrival video packets at the application layer, especially when retransmission is employed at the link layer. Since the buffer size at the link layer is typically finite, the time-varying channel service rate can induce buffer overflow (and therefore, video packet loss) due to the bit rate mismatch between the transmitting video packet and the channel service rate. At the application layer, due to variation in arrival time of video packets, some packets may become useless during playback if its arrival time exceeds certain threshold.

With time-varying wireless link quality, providing QoS for video applications in the form of *absolute* guarantee [25], [28] may not be feasible. Thus, it is more reasonable to provide QoS in the form of *soft* (or "elastic") guarantee, which allows QoS parameters in the priority transmission system to be adjusted along with changing channel conditions. The relative QoS differentiation discussed in [2], [8], and [28] is one of the possible solutions for next generation adaptive QoS system. Similarly, on the application layer, it is desirable to have a video bitstream be adaptive to changing channel conditions. Among several possible approaches for video quality adaptation [21], [30], we will employ scalable video in this paper due to its low complexity and high flexibility in rate adaptation.

To coordinate effective adaptation of QoS parameters at video application layer and priority transmission system, cross-layer interaction and QoS mapping mechanism are required. Unfortunately, a good cross-layer QoS mapping and adaptation mechanism that offers a good compromise between the video quality requirement and the available transmission resource is a challenging task. This is because at the priority transmission layer, QoS is expressed in terms of probability of buffer overflow

Manuscript received October 1, 2002; revised June 30, 2003. The work of W. Kumwilaisak and Y. T. Hou was done while they were visiting Microsoft Research Asia, Beijing, China, during the Summer of 2002.

W. Kumwilaisak was with University of Southern California, Los Angeles, CA 90089–2564 USA. He is now with Samsung Electronics, Suwon, Korea (email: wuttipong.k@samsung.com).

Y. T. Hou is with The Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24061 USA (email: thou@vt.edu).

Q. Zhang, W. Zhu, and Y.-Q. Zhang are with Microsoft Research Asia, Beijing 100080, China (e-mail: wwzhu@microsoft.com).

C.-C. J. Kuo is with Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA 90089-2564 USA (email: cckuo@sipi.usc.edu).

Digital Object Identifier 10.1109/JSAC.2003.816445

and/or the probability of delay violation at the link layer. On the other hand, at the video application layer, QoS is measured objectively by the mean squared error (MSE) and/or the peak-signal-to-noise ratio (PSNR). Despite of recent research efforts in mapping QoS parameters across these two domains (see related work in Section VII), there are still some important issues that remain unanswered, which we summarize as follows.

1) A QoS-based adaptation model, which shows how QoS parameters of both priority transmission systems and video applications should be adjusted based on time-varying wireless channel.
2) A coordination mechanism between priority transmission system and video applications, which provides interaction between the two layers.
3) A resource allocation within the priority transmission system, which provides soft QoS guarantee based on time-varying wireless channel.

To address these issues, we present a QoS mapping architecture that address cross-layer QoS issues for video delivery over wireless networks. We present details for each important building blocks under this architecture, which include: 1) the derivation of the rate constraint of a priority transmission system; 2) the development of a QoS mapping mechanism that optimally maps video classes to statistical QoS guarantees of a priority transmission system; and 3) the QoS interaction procedure between video applications and the priority transmission system to provide the best tradeoff between the video application quality and the transmission capability under time-varying wireless channel.

The rest of the paper is organized as follows. In Section II, we describe the cross-layer QoS mapping architecture proposed in this paper. In the subsequent three sections, we present the details for the important building blocks in this architecture. In particular, Section III derives the rate constraint of a priority transmission system, Section IV presents the QoS mapping between video applications and the priority transmission system under time-varying wireless channel, and Section V shows the interaction procedure between video applications and the priority transmission system. Simulation results are given in Section VI. Section VII reviews related work and Section VIII concludes this paper.

II. Architectural Description

Fig. 1 shows the proposed cross-layer QoS mapping architecture for video delivery over a single-hop wireless networks. This architecture considers an end-to-end delivery system for a video source from the sender to the receiver, which includes source video encoding module, cross-layer QoS mapping and adaptation module, link layer packet transmission module, wireless channel (time varying and nonstationary), adaptive wireless channel modeling module, and video decoder/output at the receiver. In this section, we will give an overview of key modules in this cross-layer QoS mapping architecture. Since the main challenge here is the time-varying and nonstationary behavior of the wireless link, we will describe its modeling first. Then, we will discuss the link layer packet transmission module, and cross-layer QoS mapping and adaptation module.

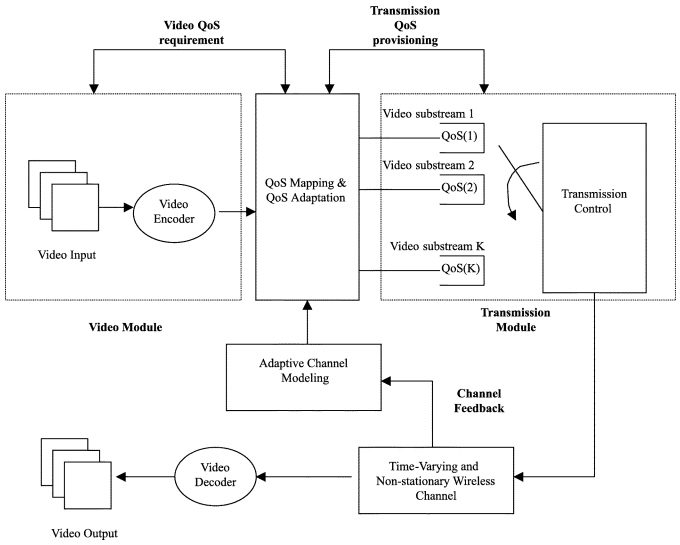


Fig. 1. A schematic of a cross-layer QoS management architecture for video delivery over wireless channel.

In this investigation, we consider to model the wireless channel at the link layer (instead of physical layer) since the link layer modeling is more amenable for analysis and simulations of the QoS provisioning system (e.g., delay bound or packet loss rate) [31]. Here, the wireless link is expected to be fading, time-varying, and nonstationary, which will provide a time-varying available transmission bandwidth for video service. We assume that the fading, time-varying, and nonstationary characteristics of the wireless channel can be modeled by a discrete-time Markov model (see Section III-A), where each state represents the available transmission rate under current channel conditions. This channel modeling process is performed by the adaptive channel modeling module in Fig. 1. Since the Markov model for the channel can be fully characterized by its transition probability matrix, the adaptive channel modeling module will periodically measure and update the transition probability matrix to keep track of the current channel characteristics based on the algorithm proposed in [12].

We now describe the link-layer transmission control module in the architecture. In this module, we employ a class-based buffering and scheduling mechanism to achieve differentiated services. In particular, we maintain K QoS priority classes with each class of traffic being maintained in separate buffers. A strict (nonpreemptive) priority scheduling policy is employed to serve packets among the classes. That is, packets in a higher priority queue will always be sent first; packets in the lower priority queue will be sent only if there is no packet in the higher priority queues. Also, packets within the same class queue are served in a first-in–first-out (FIFO) manner. For a packet that experiences excess queueing delay (i.e., will miss its scheduled playback time) will be flushed out of the buffer (discarded) without being sent over the wireless channel. Based on this class-based buffering and strict priority scheduling mechanism, we expect that each QoS priority class will have some sort of statistical QoS guarantees in terms of probability of packet loss and packet delay. As we shall see in Section III-B, statistical QoS guarantees of multiple priority classes can be translated into rate constraints based on the effective capacity theory [5], [31]. The

calculated rate constraints will in turn specify the maximum data rate that can be transmitted reliably with statistical QoS guarantee over the time-varying wireless channel. Consequently, this will enable us to classify video substreams into classes and allocate transmission bandwidth for each class.

It is worth pointing out that the adaptive wireless channel modeling module and link-layer transmission control module are generically designed and application independent. They are installed at wireless end system as a common platform to support a wide range of applications (not limited to video delivery). There are many advantages for such design, such as universal applicability, modularity, and economy of scale (i.e., can be massively produced).

We now consider the QoS-mapping and adaptation module, which is the key component to achieve cross-layer QoS mapping in this video delivery architecture. Unlike the adaptive channel modeling module and link-layer transmission module, the QoS-mapping and adaptation module is application-specific. In this case, it is designed to optimally match video application layer QoS and the underlying link-layer QoS. Since the QoS measure at the video application layer (e.g., distortion and uninterrupted video service perceived by end users) is not directly related to QoS measure in the link layer (e.g., packet loss/delay probability), a mapping and adaptation mechanism must be in place to maximize application layer QoS with the time-varying available link layer transmission bandwidth. To be more specific, at the video application layer, each video packet is characterized based on its loss and delay properties, which contribute to the end-to-end video quality and service. Then, these video packets are classified and optimally mapped to the classes of link transmission module under the rate constraint. The video application layer QoS and link-layer QoS are allowed to interact with each other and adapt along with the wireless channel condition. The objective of these interaction and adaptation is to find a satisfactory QoS tradeoff so that each end user's video service can be supported with available transmission resources. In Section IV, we will show in details how the video application layer QoS can be optimally mapped into link-layer QoS for video packet transmission. Then, in Section V, the adaptive QoS module via cross-layer QoS interaction will be described.

III. Video Bitstream Rate Constraint Under Priority Transmission

In this section, we derive the video substream rate constraint in the strict priority transmission module. The rate constraint specifies the maximum input data rate to a particular buffer class that can be transmitted with certain statistical QoS guarantee. It will be used as the basis to allocate the channel bandwidth for data transmission. Since the wireless channel is expected to be fading, time-varying, and nonstationary, in Section III-A, we first characterize the time-varying available transmission rate using a Markov chain [12]. Then, in Section III-B, we outline key results from effective capacity theory that will be used for our derivation for the class rate constraint. Finally, in Section III-C, we derive the rate constraint for each class traffic under a strict priority scheduler. Table I lists the notations that we will use in this paper.

TABLE I
Summary of Notations I

$X_c(t)$	:	random channel state at time t.
r_i	:	achievable channel transmission rate of channel state i.
$r_{channel}$	:	expected link-layer transmission rate.
$P\{\cdot\}$	:	probability of the event $\{\cdot\}$.
p_{ij}	:	transition probability from channel state i to channel state j.
p_i	:	state probability of channel state i.
$P_{transition}$	:	transition probability matrix.
$B_i(t)$	:	queue length of priority class i at time t.
B_i^{max}	:	buffer size of priority class i.
$\tilde{T}_i^{max}$	:	delay bound provided by priority class i.
$\tilde{T}_i(t)$	:	delay bound provided by priority class i at time t.
θ_i	:	QoS exponent corresponding to guaranteed packet loss probability of priority class i.
ϕ_i	:	QoS exponent corresponding to guaranteed packet delay probability of priority class i.
κ_i	:	QoS exponent corresponding to required packet loss probability of substream i.
ξ_i	:	probability that a buffer of priority class i is not empty.
$S_i(t)$	:	channel service of class i in bits over the time interval [0,t).
$\alpha_i(t)$	:	generated source rate of class i in bits/s at time t.
$\alpha_i^{(c)}(u)$	:	random channel service rate of class i in bits/s.
$\mu_i(u)$	:	effective capacity of a channel of class i with QoS exponent u.
$\Lambda_i(u)$	:	asymptotic log-moment generating function of a stochastic process of class i.

A. Time-Varying Nonstationary Wireless Channel Rate

Although the wireless channel is expected to be time-varying and nonstationary, we assume that within each small time interval, say g, the channel rate is stationary and time-varying. Furthermore, within each small time interval g, we assume that service rate for the time-varying wireless channel can be modeled by a first-order L-state Markov model as suggested in [29].

Within the small time interval g, denote $X_c(u)$ as the state of the channel at time u and $X_c(u) \in \{1, \ldots, L\}$. Each state $X_c(u) = i$ corresponds to a channel link condition, which can be characterized by an achievable channel transmission rate of r_i. The achievable channel transmission rate at state i (in unit of bits per second) can be computed as

$$r_i = R \cdot \log_2(1 + \gamma_i) \tag{1}$$

where R is the transmission bandwidth in Hz and γ_i is the SNR value of the wireless channel condition at state i (physical layer parameter) [12].

For the L-state discrete-time Markov chain, denote p_{ij} as the state transition probability from state i (at time $u-1$) to state j (at time u) with a transition time interval of 1 time unit and $1 < g$. That is, $p_{ij} = P\{X_c(u) = j | X_c(u-1) = i\}$. Then, the L-sate Markov chain can be completely characterized by the $L \times L$ state transition matrix

$$P_{\text{transition}} = \begin{pmatrix} p_{11}, & \cdots & p_{1L} \\ \vdots & \vdots & \vdots \\ p_{L1} & \cdots & p_{LL} \end{pmatrix}. \tag{2}$$

Using the state transition matrix, we can calculate the state probability for Markov model within the time interval g [27], which we denote as $[p_1, p_2, \ldots, p_L]$. Therefore, the expected

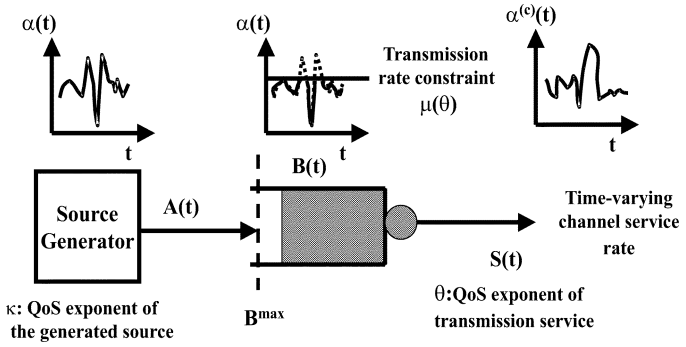


Fig. 2. A queueing transmission model.

link-layer transmission rate r_{channel} during this time interval g is

$$r_{\text{channel}} = \sum_{i=1}^{L} r_i \cdot p_i \tag{3}$$

where r_i is the achievable link layer transmission rate in (1).

At the end of each time interval g, the state transition matrix in (2) will be updated by the adaptive channel modeling module to reflect the nonstationary nature of the wireless environment [12].

B. Effective Service Capacity: Some Background

In this section, we provide some background on effective service capacity, which will be used to derive the rate constraint for multiple classes under strict priority scheduling in the next section.

Fig. 2 shows a queueing system for time-varying source rate and channel service rate. The accumulated amount of data generated by the source from time 0 to t is a random variable of the form

$$A(t) = \int_0^t \alpha(u)du \tag{4}$$

where $\alpha(u)$ is the source data generation rate. The amount of data $A(t)$ will be stored in the buffer of size $B^{\max}$ awaiting for transmission. On the other hand, the accumulated channel service from time 0 to t is of the form

$$S(t) = \int_0^t \alpha^{(c)}(u)du \tag{5}$$

where $\alpha^{(c)}(u)$ is the channel service rate at time u. Recall that the time-varying channel service rate has been modeled by a L-state discrete-time Markov chain in Section III-A, where $\alpha^{(c)}(u) \in \{r_1, r_2, \ldots, r_L\}$.

Based on the results in [5] and [31], the stochastic behavior of the accumulated channel service $S(t)$ can be described by the concept of *effective capacity*, which can be written in the form of

$$\mu(\theta) = -\frac{\Lambda^{(c)}(\theta)}{\theta} \tag{6}$$

where $\Lambda^{(c)}(\theta)$ is the asymptotic log-moment generating function of $S(t)$, defined as

$$\Lambda^{(c)}(\theta) = \lim_{t\to\infty} \frac{\log E\left[e^{-\theta S(t)}\right]}{t}$$

and θ is called the QoS exponent corresponding to the effective capacity $\mu(\theta)$. The parameter θ is related to the statistical QoS guarantee (e.g., packet loss probability) of the time-varying channel. By using large deviation theory [5], [31], the statistical QoS guarantee in terms of packet loss probability can be derived as a function of θ as follows:

$$P\{B(t) > B^{\max}|\theta\} \approx \xi \cdot e^{-\theta \cdot B^{\max}} \tag{7}$$

where $B(t)$ is the buffer occupancy at time t, $B^{\max}$ is the maximum buffer size, ξ is the probability that the buffer is not empty, and $\xi \cdot e^{-\theta B^{\max}}$ is the approximate packet loss probability guarantee.

As indicated by (6) and (7), the effective channel capacity is related with the statistical QoS guarantee through the QoS exponent θ. Intuitively, it says that the effective capacity in (6) imposes a limit for maximum amount of data that can be transmitted over time-varying channel with statistical QoS guarantee in (7).

In general (see Fig. 2), the statistical QoS guarantee required by the source (e.g., characterized by a QoS exponent κ) may mismatch with the statistical QoS guarantee provided by the channel (i.e., characterized by the QoS exponent θ). In particular, if the source generating rate corresponding to the effective capacity of QoS exponent κ is greater than the effective channel capacity (i.e., $\mu(\kappa) > \mu(\theta)$), part of the source rate would be expected to be cut-off (or shaped). The following result from [5] and [6] shows the maximum source rate that can be transmitted when there is a mismatch between the QoS exponents corresponding to the source and channel

$$\mu(\kappa) = \begin{cases} \mu(\theta), & 0 \le \kappa \le \theta \\ \mu(\theta)\frac{\theta}{\kappa} + \frac{\kappa-\theta}{\kappa} e_{S(t)}(\kappa-\theta), & \kappa > \theta \end{cases} \tag{8}$$

where $\kappa > 0$ is the QoS exponent corresponding to the packet loss probability required by source generation rate and $\mu(\kappa)$ is the source generation rate with QoS exponent κ. Note that

$$e_{S(t)}(\kappa-\theta) = \frac{\Lambda^{(c)}(\theta-\kappa)}{\kappa-\theta}$$

can be viewed as the effective bandwidth [5], [11], [31] of $S(t)$ with the QoS exponent $\kappa - \theta$.

Note that when the time-varying service rate is modeled as a Markov chain as in Section III-A, the closed form of effective service capacity and effective bandwidth can be obtained via [5] and [11] as

$$\mu(\theta) = \frac{\ln\left[\Omega(e^{-\theta\Lambda}P_{\text{transition}})\right]}{-\theta} \tag{9}$$

and

$$e_{S(t)}(\theta) = \frac{\ln\left[\Omega(e^{\theta\Lambda}P_{\text{transition}})\right]}{\theta} \tag{10}$$

where $\mu(\theta)$ and $e_{S(t)}(\theta)$ are the effective capacity and the effective bandwidth of $S(t)$ corresponding to QoS exponent θ, respectively, $P_{\text{transition}}$ is the transition probability matrix of the

discrete Markov model and $\Omega(U)$ is the spectral radius of matrix U. Λ is defined as a diagonal matrix of achievable channel transmission rate of each Markov state obtained from (1). In case of L Markov states, the diagonal matrix can be shown as

$$\Lambda = \begin{pmatrix} r_1 & 0 & \dots & 0 \\ 0 & r_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots \\ 0 & 0 & \dots & r_L \end{pmatrix}. \tag{11}$$

C. Rate Constraint Derivation of Multiple Priority Classes

The rate constraint of multiple priority classes under a time-varying service rate channel $\alpha^{(c)}(u)$ is derived in this section. We assume that the channel characteristic is stationary in a period of derivation but time-varying. We start the derivation by first assuming that there are only two priority classes with QoS exponents θ_1 and θ_2 corresponding to their guaranteed packet loss probabilities. The randomly generated rate of data substream at time t for transmitting over the first and second priority classes are $\alpha_1(t)$ and $\alpha_2(t)$ and stored in different buffers of sizes $B_1^{\max}$ and $B_2^{\max}$, respectively. The statistical QoS guarantee of each priority class is provided in form of the packet loss probability as shown in (7), which is computed based on its corresponding QoS exponent and the buffer size. With the strict priority scheduling, the second priority class has a lower priority than the first priority class and will be served only after all data in the buffer of the first priority class is served.

For the first substream in the high priority buffer, it is easy to see that the rate constraint of substream 1 with QoS exponent requirement κ_1 transmitted over the priority class 1 with QoS exponent θ_1 and buffer size $B_1^{\max}$ can be shown based on (8) as

$$\min\{\mu_1(\kappa_1), \alpha_1(t)\} < r_{\text{channel}} \tag{12}$$

or

$$\min\{\mu_1(\theta_1), \alpha_1(t)\} < r_{\text{channel}},\ 0 \le \kappa_1 \le \theta_1 \tag{13}$$

$$\min\left\{\mu_1(\theta_1)\frac{\theta_1}{\kappa_1} + \frac{\kappa_1-\theta_1}{\kappa_1} \cdot e_{S_{1(t)}}(\kappa_1-\theta_1), \alpha_1(t)\right\} < r_{\text{channel}},\ \kappa_1 > \theta_1 \tag{14}$$

where $\mu_1(\kappa_1)$ is the rate constraint of substream 1 and r_{channel} is the expected channel service rate computed from (3). $S_1(t)$ is the random variable of information that can be transmitted over priority class 1 under the time-varying service rate $\alpha_1^{(c)}(u) = \alpha^{(c)}(u)$ from time 0 to t.

For the low priority substream, the existence of substream 1 affects the rate constraint of substream 2 due to the strict priority scheduling algorithm. The derivation of the rate constraint of substream 2 can be simply viewed as trying to transmit substream 2 alone with time-varying channel service rate

$$\alpha_2^{(c)}(u) = \alpha_1^{(c)}(u) - \min\{\mu_1(\kappa_1), \alpha_1(t)\} \tag{15}$$

where $\alpha_2^{(c)}(u)$ is the time-varying channel service rate, which is seen by substream 2 with the existence of substream 1. Suppose that substream 2 has its own QoS exponent requirement equaling κ_2. Hence, from (8), the rate constraint of substream 2 can be computed based on $\alpha_2^{(c)}(u)$ as

$$\mu_2(\kappa_2) = \begin{cases} \mu_2(\theta_2), & 0 \le \kappa_2 \le \theta_2 \\ \mu_2(\theta_2)\frac{\theta_2}{\kappa_2} + \frac{\kappa_2-\theta_2}{\kappa_2} \cdot e_{S(t)}(\kappa_2-\theta_2), & \kappa_2 > \theta_2 \end{cases} \tag{16}$$

where $S_2(t)$ is the random variable of information that can be transmitted over priority class 2 under the time-varying service rate $\alpha_2^{(c)}(u)$ from time 0 to t, $\mu_2(\cdot)$ is the effective capacity computed from $S_2(t)$ with (6), and $e_{S_2(t)}(\kappa_2-\theta_2)$ is the effective bandwidth of $S_2(t)$ with QoS exponent provisioning $\kappa_2-\theta_2$.

Together with (12), the rate constraint on both substreams 1 and 2 can be expressed as

$$\min\{\mu_1(\kappa_1), \alpha_1(t)\} < r_{\text{channel}} \tag{17}$$

and

$$\min\{\mu_1(\kappa_1), \alpha_1(t)\} + \min\{\mu_2(\kappa_2), \alpha_2(t)\} < r_{\text{channel}}. \tag{18}$$

(17) and (18) show that the transmission rates of substreams 1 and 2 are limited by $\mu_1(\kappa_1)$ and $\mu_2(\kappa_2)$, respectively. Moreover, the summation of the constraint on the rate of both substreams 1 and 2 should not exceed the expected channel service rate r_{channel}. Therefore, when the substream demands to send more data than the rate constraint, in which the priority class can allow with the statistical QoS guarantee, the rate shaper algorithm has to be applied to shape the information rate to meet with the rate constraints.

The procedure for deriving the rate constraint for two data substreams can be easily extended to K substreams via

$$\sum_{i=1}^{k} \min\{\mu_i(\kappa_i), \alpha_i(t)\} < r_{\text{channel}}, \qquad k = 1, 2, \dots, K \tag{19}$$

where $\mu_i(\kappa_i)$ is the rate constraint of substream i computed by assuming that the channel service rate seen by substream i can be written as

$$\alpha_i^{(c)}(u) = \alpha^{(c)}(u) - \sum_{j=1}^{i-1} \min\{\mu_j, (\kappa_j), \alpha_j(t)\} \tag{20}$$

where κ_i is the QoS exponent corresponding to the guaranteed packet loss probability required by source substream i and $\alpha_i(t)$ is the random data rate generated by the source of class i.

As shown in our analysis, the rate constraints for multiple priority classes are dependent on each other. Channel occupation by higher priority classes (i.e., rate constraints) affects the rate constraints of lower priority classes. The higher channel occupation from higher priority classes, the lower opportunity of channel resource usage from lower priority ones. It is worth pointing out that although we only consider rate constraint under strict priority scheduling in this paper, it is possible to extend this result under other scheduling disciplines.

IV. Mapping Video Layers for QoS Classes

In this section, we study how to optimally map each video layer to one of the priority classes. Although our focus is on scalable video coding, the underlying technique is applicable

TABLE II
SUMMARY OF NOTATIONS II

F	:	video playback frame rate in frames/s.
ΔD_j	:	the distortion reduction if video layer j is correctly received.
$\nu_i(\cdot)$	:	the probability of packet delay violation under priority class i.
$\varepsilon_i(\cdot)$	:	the probability of packet loss under priority class i.
N_{GOP}	:	the number of video frames in one group of pictures (GOP).
π_j	:	the priority class that video layer j is transmitted.
$t_d(j)$	:	the playback deadline of video layer j.
D_0	:	the expected distortion if no video data is received.
$D_{GOP}(\cdot)$	:	the total expected distortion from mapping N_{GOP} scalable video frames to priority classes.
$\beta_j(\cdot)$	:	the probability that video layer j is lost due to either buffer overflow or playback deadline violation.
b_j^i	:	the size of video layer j, which will be conveyed by priority class i.
$\Upsilon(t)$	:	the QoS bound of the video quality requirement at time t.
Ω_i	:	the range of statistical QoS guarantee in terms of buffer overflow probability of priority class i.
Θ_i	:	the range of rate constraint corresponding to the range of statistical QoS guarantee of priority class i.
Ψ	:	the set of the guaranteed packet loss probability of priority networks.
N_s	:	the number of possible QoS parameters of multiple priority classes used for QoS adaptation.

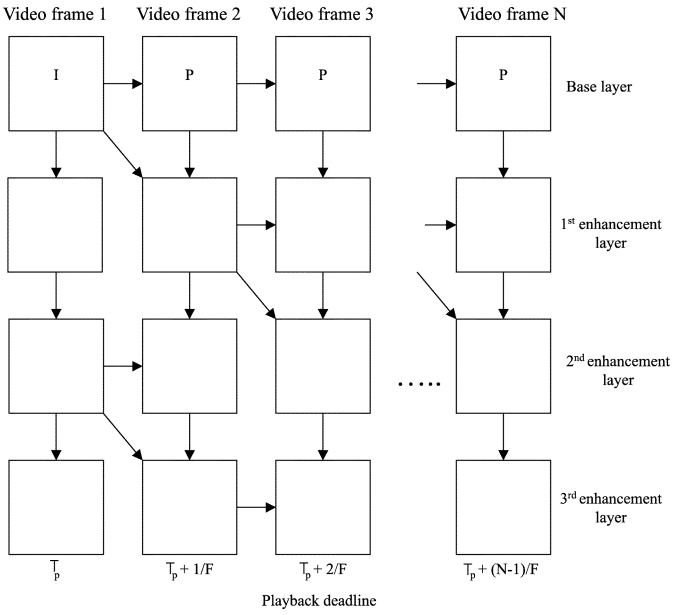


Fig. 3. GOP structure of MPEG-4 PFGS scalable video.

to other prioritized video coding schemes (e.g., relative priority index [24]). Some additional notations are listed in Table II.

A. Preliminaries

Fig. 3 shows a group-of-picture (GOP) structure of MPEG-4 PFGS [30]. Suppose that there are $N_{\rm GOP}$ video frames and there are M video layers in one GOP. Therefore, the loss of any video portion will affect the end-to-end video quality due to the interdependency within the encoding structure. Each video layer is packetized into several fixed-size packets before transmission. Each video packet cannot contain video data across video layers or video frames. Packets from the same video layer are put onto the same priority class.[1] Furthermore, suppose that the video playback frame rate at the end user is fixed at F frames/s. If the mobile terminal starts to play back the first video frame of a GOP at time T_p, video frame n in the same GOP should be received and be ready to be displayed before time $T_d(n) = T_p + ((n-1)/F)$ for uninterrupted playback.

Let $\vec{\pi} = [\pi_1, \ldots, \pi_M]$ be the mapping policy from M video layers to K priority classes, where $\pi_j \in \{0, 1, \ldots, K\}$ is the priority class that video layer j is transmitted. $\pi_j = 0$ represents the fact that video layer j is abstained from transmission. The overall expected distortion from the mapping scheme $\vec{\pi}$ can be derived using the dependent structure of scalable video [7] as shown in Fig. 3, which can be expressed as follows

$$D_{\rm GOP}(\vec{\pi}) = D_0 - \sum_{j=1}^{M} \Delta D_{j' \preceq j}^{\pi_j} \tag{21}$$

where

$$\Delta D_{j' \preceq j}^{\pi_j} = \Delta D_j \Pi_{j' \preceq j} \left(1 - \beta_{j'}\left(\theta_{\pi_{j'}}, \phi_{\pi_{j'}}\right)\right) \tag{22}$$

$D_{\rm GOP}(\vec{\pi})$ is the total expected distortion from mapping $N_{\rm GOP}$ scalable frames to K different priority classes with an allocation policy $\vec{\pi}$, D_0 is the expected distortion if no video data are received, ΔD_j is the distortion reduction if video layer j is correctly received. Note that the term $\Pi_{j' \preceq j}(1 - \beta_{j'}(\theta_{\pi_{j'}}, \phi_{\pi_{j'}}))$ in (22) is the probability that video layer j and all video layers j', on which video layer j depends, are correctly received while video layer j' is transmitted over the priority class $\pi_{j'}$. The priority class $\pi_{j'}$ has QoS exponent $\theta_{\pi_{j'}}$ and $\phi_{\pi_{j'}}$, which correspond to its guaranteed buffer overflow and delay bound probability, respectively. On the other hand, $\beta_{j'}(\theta_{\pi_{j'}}, \phi_{\pi_{j'}})$ is the probability that video layer j' is lost due to either buffer overflow or playback deadline violation when transmitted over priority class $\pi_{j'}$. Since we map all video packets from the same video layer to the same priority class, the probability that video layer j' will be lost can be computed as

$$\beta_{j'}\left(\theta_{\pi_{j'}}, \phi_{\pi_{j'}}\right) = \varepsilon_{\pi_{j'}}\left(\theta_{\pi_{j'}}\right) + \left(1 - \varepsilon_{\pi_{j'}}\left(\theta_{\pi_{j'}}\right)\right) \cdot \nu_{\pi_{j'}}\left(\phi_{\pi_{j'}}, t_d(j')\right) \tag{23}$$

where $\varepsilon_{\pi_{j'}}(\theta_{\pi_{j'}})$ and $\nu_{\pi_{j'}}(\phi_{\pi_{j'}}, t_d(j'))$ denote the probabilities that video packets corresponding to video layer j' are lost due to buffer overflow and playback deadline violation (i.e., the deadline is $t_d(j')$)) when transmitted over priority class $\pi_{j'}$, respectively.

To derive $\varepsilon_{\pi_{j'}}(\theta_{\pi_{j'}})$ and $\nu_{\pi_{j'}}(\phi_{\pi_{j'}}, t_d(j'))$, we use our results in Section III-B, which was based on the theory of large deviation [5], [31]. First, $\varepsilon_{\pi_{j'}}(\theta_{\pi_{j'}})$ can be directly obtained by using (7) as

$$\varepsilon_{\pi_{j'}}\left(\theta_{\pi_{j'}}\right) = P\left\{B_{\pi_{j'}}(t) > B_{\pi_{j'}}^{\max} | \theta_{\pi_{j'}}\right\} \approx \xi_{\pi_{j'}} \cdot e^{-\theta_{\pi_{j'}} B_{\pi_{j'}}^{\max}} \tag{24}$$

where $\xi_{\pi_{j'}}$ is the probability that the buffer of priority class $\pi_{j'}$ is not empty, $\theta_{\pi_{j'}}$ is the QoS exponent corresponding to buffer overflow probability of priority class $\pi_{j'}$, $B_{\pi_{j'}}(t)$ is the buffer

[1]Although the wireless channel is nonstationary in nature, it is reasonable to assume that on the time scale of one GOP (e.g., 1 s), the channel characteristic is stationary.

occupancy under priority class $\pi_{j'}$ at time t, and $B^{\max}_{\pi_{j'}}$ is the maximum buffer size of priority class $\pi_{j'}$.

Then, $\nu_{\pi_{j'}}(\phi_{\pi_{j'}}, t_d(j'))$ can be computed by using the relationship between the experienced packet delay and the buffer occupancy of priority class $\pi_{j'}$ as [33]

$$\tilde{T}_{\pi_{j'}}(t) \leq \frac{B_{\pi_{j'}}(t)}{\mu_{\pi_{j'}}\left(\kappa_{\pi_{j'}}\right)} \tag{25}$$

where $\mu_{\pi_{j'}}(\kappa_{\pi_{j'}})$ is the rate constraint of priority class $\pi_{j'}$ as derived in Section III-C and $\tilde{T}_{\pi_{j'}}(t)$ is the experienced packet delay at time t under priority class $\pi_{j'}$. The upper bound of (25) can be obtained when $B_{\pi_{j'}}(t) = B^{\max}_{\pi_{j'}}$, (i.e., $\tilde{T}^{\max}_{\pi_{j'}} = B^{\max}_{\pi_{j'}}/(\mu_{\pi_{j'}}(\kappa_{\pi_{j'}}))$). $\tilde{T}^{\max}_{\pi_{j'}}$ is the maximum delay a video packet may experience under priority class $\pi_{j'}$. By substituting the parameters from (25) and its upper bound to (24), the probability of packet delay violation under priority class $\pi_{j'}$ can be computed as

$$\begin{aligned}\nu_{\pi_{j'}}\left(\phi_{\pi_{j'}}, \tilde{T}^{\max}_{\pi_{j'}}\right) &= P\left\{\tilde{T}_{\pi_{j'}}(t) > \tilde{T}^{\max}_{\pi_{j'}} | \phi_{\pi_{j'}}\right\} \\ &\approx \xi_{\pi_{j'}} \cdot e^{-\phi_{\pi_{j'}} \tilde{T}^{\max}_{\pi_{j'}}}\end{aligned} \tag{26}$$

where $\phi_{\pi_{j'}}$ is the QoS exponent of the guaranteed delay bound of priority class $\pi_{j'}$ and can be expressed as

$$\phi_{\pi_{j'}} = \theta_{\pi_{j'}} \cdot \mu_{\pi_{j'}}\left(\kappa_{\pi_{j'}}\right). \tag{27}$$

With the QoS exponent of the guaranteed bound in (27), when video packets corresponding to video layer j' are transmitted over priority class $\pi_{j'}$, the probability of its playback delay violation can be computed as follows [31], [33]:

$$\begin{aligned}\nu_{\pi_{j'}}\left(\phi_{\pi_{j'}}, t_d(j')\right) &= P\left\{\tilde{T}_{\pi_{j'}}(t) > t_d(j') | \phi_{\pi_{j'}}\right\} \\ &\approx \xi_{\pi_{j'}} \cdot e^{-\phi_{\pi_{j'}} t_d(j')}\end{aligned} \tag{28}$$

where $t_d(j')$ is the playback deadline of video layer j' (e.g., when a video layer corresponding to video frame n, $t_d(j') = T_d(n)$).

B. Problem Formulation

Based on the parameters described above, the optimal mapping problem can be formally stated as follows. Given the set of rate constraints under the priority transmission system in Section III-C and the expected channel service rate r_{channel}, which can be considered stationary in a time period g corresponding to one GOP, what is the optimal mapping policy $\vec{\pi}^*$ from one GOP with N_{GOP} scalable frames (coded in M video layers) to K priority classes such that $D_{\text{GOP}}(\vec{\pi})$ is minimized? That is

$$\text{Min} \quad D_{\text{GOP}}(\vec{\pi}) \tag{29}$$

$$\text{s.t.} \quad \sum_{\forall j, \pi_j = i} b_j^{\pi_j} \leq \mu_i(\kappa_i) \cdot g, \; i = 1, \ldots, K \tag{30}$$

$$\sum_{i=1}^{K} \mu_i(\kappa_i) < r_{\text{channel}} \tag{31}$$

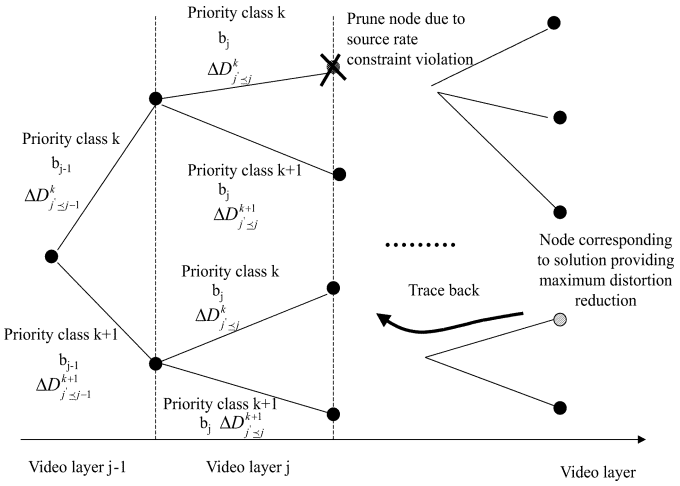


Fig. 4. Illustration of the tree search approach to derive the optimal mapping algorithm.

where $\mu_i(\kappa_i)$ is the rate constraint of priority class i, and $b_j^{\pi_j}$ is the size of video layer j, which will be conveyed by priority class π_j.

There are two sets of constraints in the above problem formulation. The first set of constraints say that the source rate of video bitstreams under each priority class must not exceed the rate constraint of the corresponding priority class. The second constraint says that the summation of rate constraints of all priority classes has to be bounded by the expected channel service rate (see Section III-C). Recall that, under optimal allocation of the wireless channel resources to video applications, the maximum bit rate of video streams transmitted over priority class i occurs when $\kappa_i = \theta_i$. Thus, we will set $\kappa_i = \theta_i$ in our QoS mapping algorithm.

C. Solution to the Optimization Problem

Our solution to the optimization problem follows a constrained-based search that exploits the dependency among the layers. Referring to Fig. 4 [18], the tree represents all possible QoS mapping solutions. Each stage of the tree corresponds to one of the video layers. Each node of the tree at a given stage represents a possible cumulative buffer occupancy in each priority class. For example, in Fig. 4, for each node at stage $j-1$, we create branches in order to account for all possible accumulated buffer occupancies due to QoS mapping from video layer 1 to j, where the number of the branches is equal to $K+1$.[2] Then, we compute the buffer occupancy corresponding to each node at stage j by summing the size of video layer j and accumulated buffer occupancy corresponding to nodes at stage $j-1$. This is equivalent to adding the size of video layer j to the buffer of all possible priority classes at stage $j-1$.

Each branch at stage j has a cost to account for the expected distortion reduction when video layer j is mapped to a particular priority class. The reduction in distortion is zero if video layer j is abstained from transmission. Therefore, as we traverse the tree from the root to leaves, we can compute the accumulated expected distortion reduction for each possible mappings. For

[2]Recall that $K+1$ classes include the case where video layer is abstained from transmission.

example, in Fig. 4, branch k from stage $j-1$ to j has the associated distortion reduction by transmission of video layer j via priority class k. The level of occupied resource is the size of video layer j, which is equal to b_j. The expected distortion reduction of video layer j in association with branch k can be found by (22) and (23) and is

$$\Delta D^k_{j' \preceq j} = \Delta D_j [1-\beta_j(\theta_k, \phi_k)] \prod_{j' \prec j} \left[1-\beta_{j'}\left(\theta_{\pi_{j'}}, \phi_{\pi_{j'}}\right)\right] \tag{32}$$

where $\Delta D^k_{j' \preceq j}$ is the expected distortion reduction when transmission video layer j over priority class k and video layers j' is correctly received.

It is worth pointing out that an exhaustive search of each node for a complete tree is not necessary, due to the rate constraint for each priority class given (30) and (31). That is, it is sufficient to prune the branch when the accumulated rate exceeds its corresponding rate constraint—a branch cannot be created if it violates the rate constraint of corresponding priority class. Once we find the maximum accumulated distortion reduction, the optimal mapping solution can be found by traversing back from the leaf node to the root of the tree (see Fig. 4).

V. Video Adaptation

A. QoS Bounds

We use a set of QoS bounds to characterize the range of video quality requirements and transmission service capabilities. Within this set of bounds, QoS parameters of video and transmission service can be adjusted to cope with the time-varying and nonstationary wireless link quality. Due to the time-varying characteristics of video content and time-varying wireless channel, the set of bounds are also time-varying. Specifically, the QoS bound for video application at time t can be defined as the video distortion of GOP as

$$\Upsilon(t) = \left[D^L_{\mathrm{GOP}}(t), D^U_{\mathrm{GOP}}(t)\right] \tag{33}$$

where $D^L_{\mathrm{GOP}}(t)$ and $D^U_{\mathrm{GOP}}(t)$ are the respective lower and upper bounds at time t. For transmission service with K priority classes, priority class i with buffer size $B_i^{\max}$ can provide statistical QoS guarantee bounds in terms of buffer overflow probability

$$\Omega_i = [\varepsilon_{i,L}(\theta_{i,L}), \varepsilon_{i,U}(\theta_{i,U})] \tag{34}$$

where $\varepsilon_{i,L}(\theta_{i,L})$ and $\varepsilon_{i,U}(\theta_{i,U})$ are the respective lower and upper bounds of the guaranteed buffer overflow probability by priority corresponding to QoS exponent $\theta_{i,L}$ and $\theta_{i,U}$. Similarly, the rate constraint corresponding to the statistical QoS guarantee (see Section III-C) can be expressed as

$$\Theta_i = [\mu_i(\theta_{i,L}), \mu_i(\theta_{i,U})] \tag{35}$$

where $\mu_i(\theta_{i,L})$ and $\mu_i(\theta_{i,U})$ are the respective rate constraints corresponding to $\varepsilon_{i,L}(\theta_{i,L})$ and $\varepsilon_{i,U}(\theta_{i,U})$. Note that the range for guaranteed packet delay can also be obtained from the guaranteed buffer overflow probability (see Section IV).

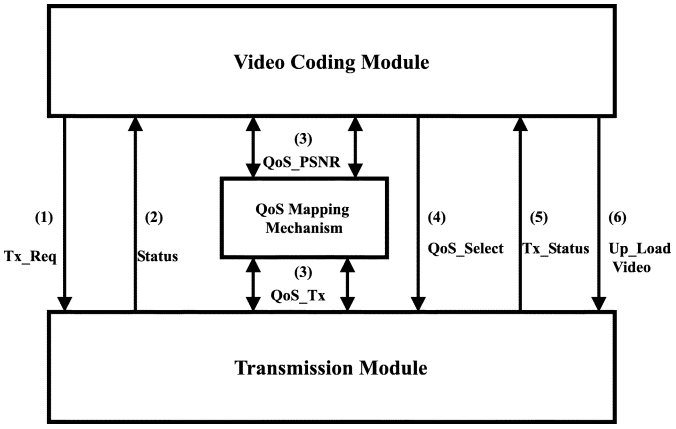


Fig. 5. Interaction between video applications and priority networks for QoS adaptation.

B. Video QoS Adaptation Through Cross-Layer Interaction

Using the set of QoS bounds, we propose an adaptation algorithm for video source through interaction with the underlying transmission network. The proposed algorithm optimally adjusts its video encoding behavior based on the QoS bounds.

Suppose that there are K priority classes in priority network. From the defined QoS bound of packet loss probability described in Section V-A, it can be set up as $\Psi = \{\vec{\varepsilon}^{(1)}, \ldots, \vec{\varepsilon}^{(Ns)}\}$, where $\vec{\varepsilon}^{(i)} = [\varepsilon_1^{(i)}(\theta_1^{(i)}), \ldots, \varepsilon_K^{(i)}(\theta_K^{(i)})]$, $\varepsilon_l^{(i)}(\theta_l^{(i)}) \in \Omega_l$ and is the candidate of guaranteed packet loss probability of priority class l, and N_s is the number of elements in Ψ. In theory, Ψ can be an infinite set (i.e., $N_s = \infty$) due to the continuous value of guaranteed packet loss probability in the QoS bound. However, in real-world implementation, N_s must be limited to a finite set of QoS options to reduce complexity. In this paper, we assume Ψ is a finite set.

The optimal video adaptation algorithm can be formulated as follows. Given that the current expected channel service rate at time t equal to $r_{\mathrm{channel}}(t)$,[3] find a set of QoS parameters for the priority network $\vec{\varepsilon}^*$ from Ψ such that the expected video distortion is minimized while satisfying the QoS bound and current available wireless channel rate. That is

$$\text{Min} \quad D_{\mathrm{GOP}}(\vec{\pi}_{\vec{\varepsilon}}) \tag{36}$$

$$\text{s.t.} \quad D_{\mathrm{GOP}}(\vec{\pi}_{\vec{\varepsilon}}) \in \Upsilon(t) \tag{37}$$

$$\sum_{i=1}^{K} \mu_i(\theta_i) < r_{\mathrm{channel}}(t) \tag{38}$$

where $\mu_i(\theta_i)$ is the rate constraint of priority class i (corresponding to its statistical QoS guarantee in $\vec{\varepsilon}$) and $D_{\mathrm{GOP}}(\vec{\pi}_{\vec{\varepsilon}})$ is the optimal expected video distortion from (29), which is obtained by using the set of QoS parameters of guaranteed packet loss $\vec{\varepsilon}$ and those of its counterpart guaranteed packet delay at time t. The computation of $D_{\mathrm{GOP}}(\vec{\pi}_{\vec{\varepsilon}})$ is done based on video layer mapping scheme in Section IV.

In the following, we describe an adaptation algorithm for video encoding based on interaction with the underlying transmission network to achieve the above problem formulation (also, see Fig. 5).

[3]Now, we consider the longer time scale in video transmission (i.e., more than one GOP). Therefore, the wireless channel condition is no longer be stationary and the expected channel service rate is time-varying.

Algorithm I: (**Video Adaptation**)

- Step 1: The video coding module sets up the QoS bound $\Upsilon(t)$ in terms of the expected video distortion (or the expected PSNR). Then, it sends the request for transmission (Tx_{Req}) to the transmission module to set up the transmission process.
- Step 2: After transmission module receives (Tx_{Req}), transmission module offers a set of statistical QoS guarantees for each priority class based on Ψ to the video coding module. Then, the video layer mapping interface translates the QoS provisioning of each priority class to the expected distortion value as described in Section IV. Note that all of the possible solutions in Ψ will be considered. The QoS parameters of the priority network that provide the lowest distortion and satisfy the range of video quality requirement $\Upsilon(t)$ will be chosen as the QoS parameters for video transmission. If there are no QoS parameters satisfying all constraints simultaneously. It implies that the available transmission capabilities cannot meet the video quality requirement of the video coding module under the current channel condition. Then, we go to Step 3. If all constraints are met, we go to Step 4 directly.
- Step 3: The transmission module requests the video coding module to adjust the video quality requirement $\Upsilon(t)$. The video coding module complies with this request and adjusts the QoS bound (i.e., the expected distortion range) and repeat the process in Step 2.
- Step 4: The video coding module sends selected QoS parameters (QoS_{Select}) to the transmission module to set up QoS parameters of each priority class in the priority network.
- Step 5: The transmission module sends the acknowledgment to the video coding module after its QoS parameters are set up.
- Step 6: The prioritized video bitstream is uploaded to the priority network based on agreed QoS parameters and the video layer mapping policy.
- Step 7: Upon the change of the transmission channel service rate is detected during transmission, adaptation of QoS parameters for both the video application and the priority network will be needed. That is, we go back to Step 2.

VI. Experimental Results

In this section, we present the simulation results of the proposed cross-layer QoS mapping for prioritized video transmission over time-varying and nonstationary wireless channels. The 100 video frames of CIF foreman sequence are used for simulation. Video sequence is encoded by PFGS video codec with frame rate 10 frames/s and there are ten frames in each GOP. The nonstationary behavior of wireless channels is simulated by randomly changing the normalized Doppler frequency and average power. The normalized Doppler frequency is chosen from the set of $\{10^{-3}, 5{\cdot}10^{-3}, 10^{-2}\}$ reflecting the time-varying mobile speed while the average SNR of the received signal varies from 10 to 20 dB.

A. Rate Constraint of Multiple Priority Classes

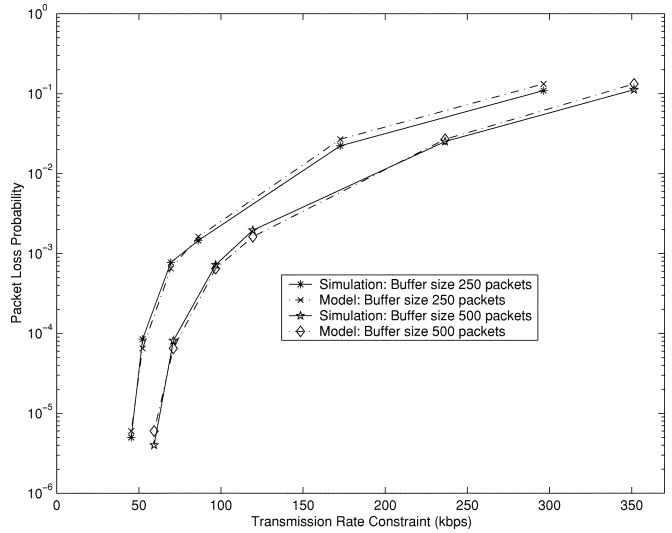
In this section, we conduct experiments to study the derived rate constraint of the multiple classes under time-varying wireless channel as described in Section III. In particular, we adopt a time-varying service-rate channel modeled by the Markov process from the work in [12].

First, let us consider two priority classes with strict priority scheduling for packet transmission under the link layer transmission. The first class has a higher priority than the second class. The packet size is 200 bytes. The expected service rate of the wireless channel is set to $r_{\text{channel}} = 380$ kb/s at normalized Doppler frequency 10^{-2}, in this simulation. As seen in Fig. 6(a), the rate constraint of the first priority class (i.e., the high priority class) computed from the closed form of effective bandwidth and effective capacity [(8), (9), and (10)] and those obtained from the simulation are close to each other over a wide range of packet loss probabilities.[4] Under time-varying channel characteristic, the lower the packet loss probability requirement, the less reliably allowable the transmitted data rate. Simulation results given in Fig. 6(a) also study the buffer size effect on the rate constraint. The larger the buffer size, the more data rate we can transmit under the same packet loss probability guarantee.

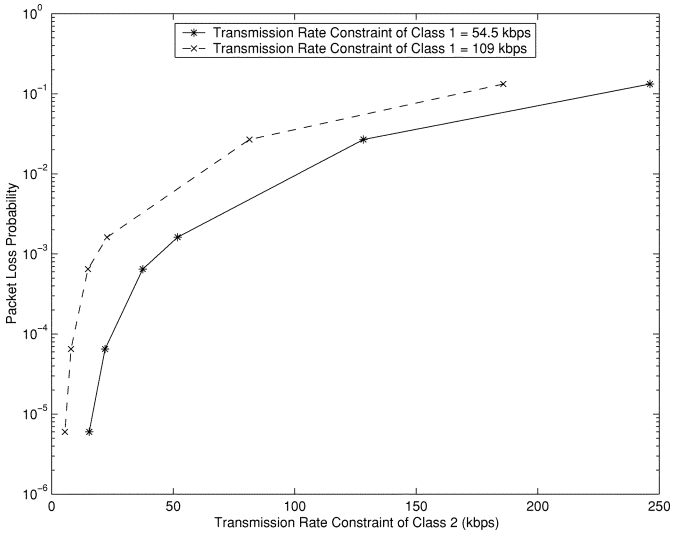
Based on the rate constraint shown in Section III-C, Fig. 6(b) shows the rate constraint of priority class 2 (i.e., the low priority class) over a wide range of guaranteed packet loss probability requirement. As shown in these simulation results, the rate constraint of priority class 2 with a buffer size equal to 250 packets is dependent on how much priority class 1 occupies the wireless link. The rate constraint of priority class 2 with a lower QoS guarantee of priority class 1 (with guaranteed packet $\text{loss probability} = 10^{-2}$ and $\text{rate constraint} = 109$ kb/s) can provide a higher transmission rate than that with a higher QoS guarantee of priority class 1 (with guaranteed packet $\text{loss probability} = 10^{-4}$ and $\text{rate constraint} = 54.5$ kb/s).

The rate constraint of priority class 1 in different wireless channel environments is investigated below. The maximum buffer size is set at 500 packets for our studies. First, the effect of different normalized Doppler frequencies (i.e., indicating the changing speed of wireless channel condition) to the rate constraint is shown in Fig. 7. If the channel changes slowly (with a low normalized Doppler frequency), the rate constraint is lower than that with a higher normalized Doppler frequency given the same guaranteed buffer overflow probability. This results from the longer period that the wireless channel stays in the bad channel condition, which leads to less overall reliable

[4]Recall that probability of packet loss and QoS exponent is related through the large deviation theory as in Section III-B.



(a)



(b)

Fig. 6. (a) Rate constraint of a high priority class from two priority classes, which is computed from the discrete Markov wireless channel model corresponding to the normalized $\mathrm{Doppler\ frequency} = 10^{-2}$ and $\mathrm{average\ power} = 16\ \mathrm{dB}$ and the buffer sizes are chosen to be 250 and 500 packets. (b) Rate constraint of a lower priority class from two priority classes and a buffer size of 250 packets based on absolute priority scheduling when the packet loss rate guarantee of class 1 is equal to 10^{-2} and 10^{-4}.

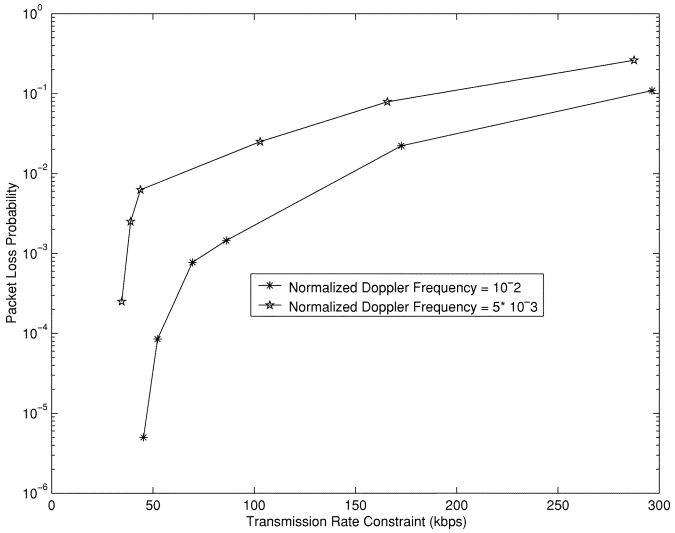


Fig. 7. Rate constraint of a wireless channel of priority class 1 computed from the discrete Markov channel model with the normalized Doppler frequencies 10^{-2} and $5 \cdot 10^{-3}$, where the average power is equal to 16 dB.

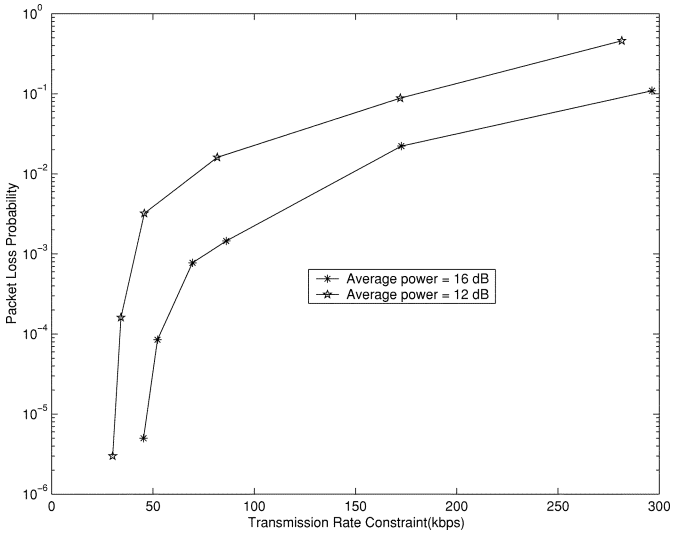


Fig. 8. Rate constraint of the wireless channel for priority class 1 computed from the discrete Markov channel model with the average power equal to 16 dB and 12 dB, where the normalized Doppler frequency is 10^{-2}.

transmission data rate. The effect of the average power on the rate constraint is shown in Fig. 8. As shown in Fig. 8, given the probability of packet loss guarantee, the rate constraint can be increased by enhancing the average power transmission. Note that the curves of rate constraint corresponding to probability of packet delay has the same tendency of those corresponding to probability of packet loss considered in this section.

B. Mapping Video Layers to QoS Classes

To study the mapping between video layers and QoS classes, three priority classes with strict priority scheduling are considered. The guaranteed buffer overflow probabilities are 10^{-4}, 10^{-3}, and 0.1 from the highest to the lowest priority class, respectively. The guaranteed delay bound probabilities are derived based on its counterpart buffer overflow probabilities and video structure. The rate constraint of each priority class can be obtained from the corresponding set of QoS guarantees and used in the QoS mapping mechanism (see Section IV).

The video packet size is set at 200 bytes, whereas the buffer size is equal to 1000 packets. The expected service rate of the wireless channel is set to $r_{\mathrm{channel}} = 380$ kb/s at normalized Doppler frequency 10^{-2} as in the previous simulation section. The target bit rate of base layer of video is equal to the rate constraint of the highest priority classes.

In Fig. 9, we compare the expected PSNR obtained from the optimal mapping algorithm proposed in Section IV and the unprioritized mapping, where all of the video layers are treated equally and randomly mapped to QoS classes. The study is conducted in a wide range of channel condition through the average channel SNR. The experimental results of each specific channel SNR are obtained from averaging over 100 video frames and 30 channel realizations. We can see from simulation results that

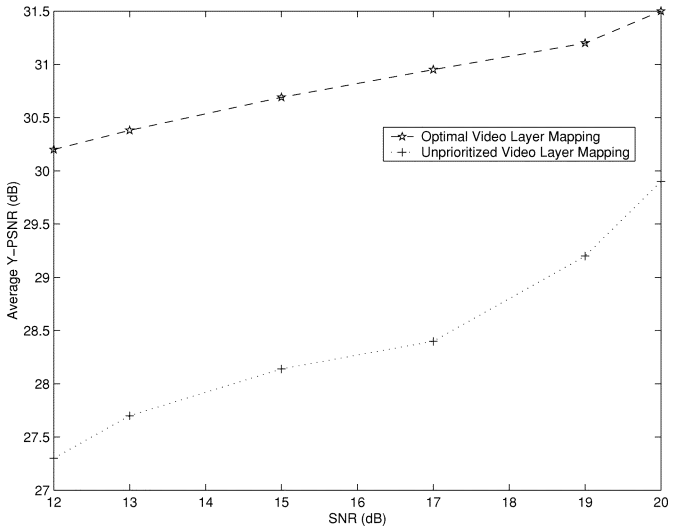


Fig. 9. Average PSNR comparison of video layer mapping to QoS classes between optimal video layer mapping and unprioritized video layer mapping.

there are significant performance difference between optimal video layer mapping to QoS classes and the mapping without considering the difference among video layers.

C. Video Adaptation

To evaluate the performance of dynamic QoS adjustment for adaptive scalable video transmission, three priority classes with strict priority scheduling are considered. The range of guaranteed buffer overflow probabilities are $[10^{-5}, 10^{-3}]$, $[10^{-4}, 10^{-2}]$, and $[10^{-2}, 2 \cdot 10^{-1}]$ for priority classes 1, 2, and 3, respectively. The number of possible QoS parameters of multiple priority classes used for QoS adaptation obtained from QoS ranges is 20 (i.e., $N_s = 20$). The rate constraints of priority classes and probability of guaranteed delay bound are derived from the set of buffer overflow probabilities. With described parameters, the QoS mapping mechanism proposed in Section IV is used as a QoS interface between the video and the transmission modules during video transmission.

The video sequence is pre-encoded with the target bit rate of the base-layer equal to 100 kb/s. The time-varying and nonstationary wireless channel is simulated by varying the average power and the normalized Doppler frequency and modeled by the discrete Markov chain as described before. The channel is assumed to be stationary during one GOP interval (i.e., 1 s). The buffer size is set to be 1000 packets with the packet size equal to 200 bytes.

We compare three video transmission systems with different characteristics:

- System 1: no QoS interaction and adaptation with the expected PSNR requirement equal to 29 dB;
- System 2: no QoS interaction but with QoS adaptation system with the expected PSNR requirement equal to 29 dB;
- System 3: with both QoS interaction and QoS adaptation with the PSNR requirement within the range of [27, 37] dB.

The guaranteed multiple QoS provisioning of the first transmission system is fixed at 10^{-4}, 10^{-3}, and 0.1 for the

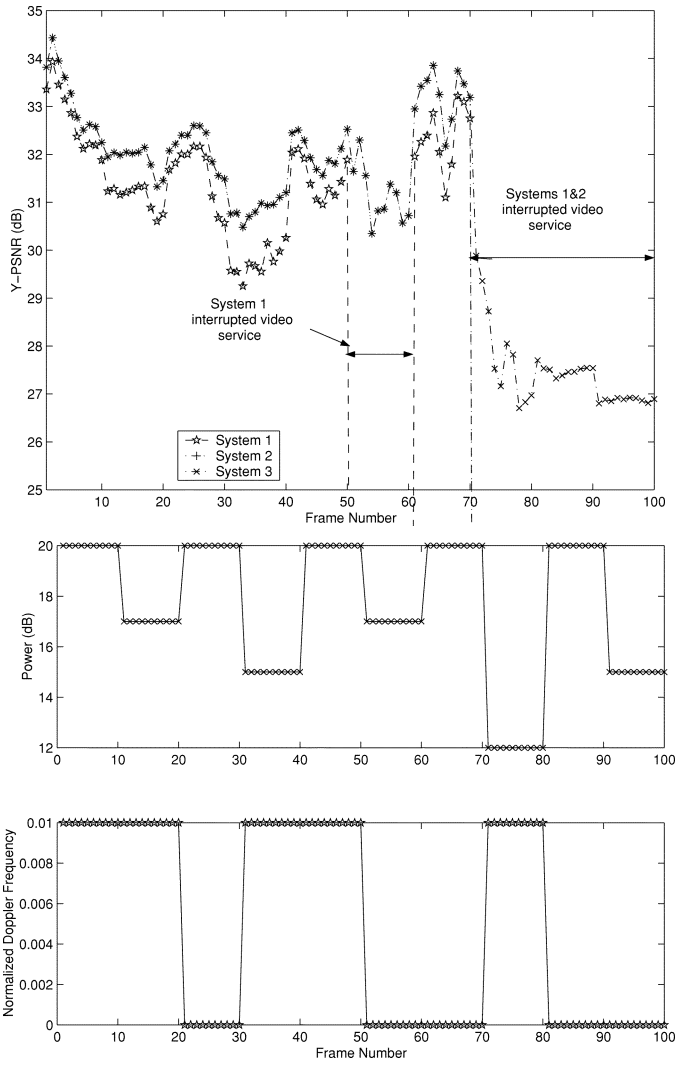


Fig. 10. Y-PSNR comparison of three video transmission systems under a nonstationary wireless environment with the time-varying power and speed.

three priority classes, respectively, while the QoS provisioning of the second and third systems are adaptively changed based on channel conditions. However, QoS interaction for adjusting the video requirement in System 2 is not applied. The adaptation scheme is performed to achieve the expected PSNR requirement.

The simulation results are given in Fig. 10. First, let us compare Systems 1 and 2. Due to adaptation capability of System 2, System 2 provides better PSNR and more consistent video service than System 1 under changing wireless network environments. However, Systems 1 and 2 are based on fixed expected video QoS requirement. Therefore, when the wireless network environment is not in a good state (i.e., average SNR is low) together with the video contents are changed, the expected video requirement can not be maintained (i.e., interrupted video service during frame 50–60 and 70–100 in System 1 and interrupted video service during frame 70–100 in System 2). With both QoS adaptation and interaction, the adaptive video transmission system (i.e., System 3) provides more consistent video service and enhanced video quality than the other two systems. Even though the original range of

PSNR requirement of System 3 may not be satisfied in some periods of video service, with the interaction between modules to adjust the PSNR range, the video service can be sustained during communication.

VII. Related Work

There have been many studies on the cross-layer design for efficient multimedia delivery with QoS assurance over wired and wireless networks in recent years [1], [14], [16], [17], [19], [20], [22]–[24], [26], [32]. In [14], [16], [19], [22]–[24], and [26], the efforts have been focused on the utilization of the differentiated service architecture to convey multimedia data. The common approach in these previous works is the partitioning of multimedia data into smaller units, and then maps these units to different classes for prioritized transmission. The partitioned multimedia units are prioritized based on its contribution to the expected quality at the end user, while the priority transmission system provides different QoS guarantees depending on its corresponding service priority. Servetto *et al.* [23] proposed an optimization framework to segment a variable bit rate source to several substreams. Then, the resulting substreams are transmitted in multiple priority classes with ATM connections. The objective of this scheme was to minimize the expected distortion of the variable bit rate source due to transmission. Shin *et al.* [24], Tan *et al.* [26], Sehgal *et al.* [22], Padmannabhan *et al.* [19], Martin [16], and Masala *et al.* [14] used the different priority classes of DiffServ architecture [8] to deliver multimedia data. Shin *et al.* [24] prioritized each video packet based on its error propagation effect if it is lost. Video packets were mapped differently to transmission priority classes with the objective to maximize end-to-end video quality under the cost and/or price constraint. Tan *et al.* [26] and Masala *et al.* [14] examined the same problem as that formulated in [24] with different approaches for video prioritization. The other types of multimedia delivery over DiffServ network such as prioritized speech and audio packets were considered by Martin [16] and Sehgal *et al.* [22], respectively. However, the authors did not take into an account of the stochastic behavior of wireless network in their cross-layer design. In other words, adaptive resource allocation, adaptive QoS scheme in both application and link layer transmission, and the interaction between layers under time-varying wireless environment were not addressed. Consequently, when applying these proposed algorithms to time-varying and nonstationary wireless environment, the systems may fail to sustain QoS assurance of multimedia applications.

In [1], [17], [20], and [32], the authors introduce cross-layer design with adaptive QoS assurance for multimedia transmission. In [32], Xiao *et al.* studied the rate-delay tradeoff curve offered from the lower-layer protocol to the applications. Then, the application layer chose the operating point from this curve as a guaranteed QoS parameter for transmission. These curves can change as the wireless network environment changes.

In [1], [17], and [20], the authors investigated the dynamic QoS framework to adaptively adjust QoS parameters of the wireless network to match with time-varying wireless channel condition. In their studies, the application was given the flexibility to adapt to the level of QoS provided by the network. Even though these efforts considered the cross-layer design based on QoS adaptation framework, their QoS parameters are based on absolute QoS. Furthermore, the mutual awareness of QoS parameters between application and link transmission layer were not established. In other words, there is no interaction between layers to obtain the operating QoS tradeoff points. Therefore, their proposed system may not be able to maintain QoS when wireless channel is highly dynamic. On the contrary, by utilizing the statistical QoS in cross-layer design and interaction between layers as we have done in this paper, the multimedia transmission system tends to be more robust in maintaining QoS parameter (i.e., quality of multimedia and uninterrupted service) under highly dynamic wireless environment.

VIII. Conclusion

In this paper, we proposed a cross-layer QoS mapping architecture for video delivery over wireless environment. There are several components under this architecture, including a proposal of an adaptive QoS service model that allows QoS parameters to be adaptively adjusted according to the time-varying wireless channel condition, an interaction mechanism between the priority network and video applications to provide proper QoS selection, and a resource management scheme to assign resources based on the QoS guarantee for each priority class under the time-varying wireless channel. This architecture enables to perform QoS mapping between statistical QoS guarantees at the network level to a corresponding priority class with different video quality requirements. Simulation results demonstrated that the proposed dynamic QoS management system can provide consistent video service and enhanced end-to-end video quality over time-varying and nonstationary wireless channels.

References

[1] A. Alwin, R. Bagrodia, N. Bambos, M. Gerla, L. Kleinrock, J. Short, and J. Villasenor, "Adaptive mobile multimedia networks," *IEEE Pers. Commun.*, vol. 3, no. 2, pp. 34–51, Apr. 1996.

[2] B. Arroyo-Fernandez, J. Dasilva, J. Fernandes, and R. Prasad, "Life after third-generation mobile communications," *IEEE Commun. Mag.*, vol. 39, no. 8, Aug. 2001.

[3] Y. Le Boudec and P. Thiran, "A short tutorial on network calculus I: Fundamental bounds in communication networks," in *Proc. IEEE ISCAS 2000*, Geneva, Switzerland, May 2000, pp. IV-93–IV-96.

[4] P. Chaudhury, W. Mohr, and S. Onoe, "The 3GPP proposal for IMT-2000," *IEEE Commun. Mag.*, vol. 37, pp. 72–81, Dec. 1999.

[5] C.-S. Chang and J. Thomas, "Effective bandwidth in high speed digital networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1091–1100, Aug. 1995.

[6] C.-S. Chang and T. Zajic, "Effective bandwidths of departure processes from queues with time varying capacities," in *Proc. IEEE INFOCOM'95*, Boston, MA, Apr. 1995, pp. 1001–1009.

[7] P. A. Chou and Z. Miao, Rate-distortion optimized streaming of packetized media, submitted for publication.

[8] C. Dovrolis, D. Stiliadis, and P. Ramanathan, "Proportional differentiated services: Delay differentiation and packet scheduling," *IEEE/ACM Trans. Networking*, vol. 10, pp. 12–26, Feb. 2002.

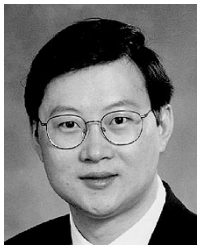
[9] A. Goldsmith and S. Wicker, "Design challenges for energy-constrained ad hoc wireless networks," *IEEE Wireless Commun.*, vol. 9, pp. 8–27, Aug. 2002.

[10] H. Holma and A. Toskala, Eds., *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*, 2nd ed. New York: Wiley, Sept. 2000.

[11] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, pp. 424–428, Aug. 1993.
[12] W. Kumwilaisak and C.-C. Jay Kuo, Adaptive variable length Markov chain for non-stationary fading channel modeling, submitted for publication.
[13] S. Manistis, E. Nikolouzou, and I. Venieris, "QoS issues in the converged 3G wireless and wired networks," *IEEE Commun. Mag.*, vol. 40, pp. 44–53, Aug. 2002.
[14] E. Masala, D. Quaglia, and J. C. de Martin, "Adaptive picture slicing for distortion-based classification of video packets," in *Proc. IEEE Workshop Multimedia Signal Processing*, Cannes, France, Oct. 2001, pp. 111–116.
[15] M. J. Mason, G. C. Hess, and S. S. Gilbert, "Shadowing variability in an urban land mobile environment at 900 MHz," *Electron. Lett.*, vol. 26, pp. 646–648, May 1990.
[16] J. C. de Martin, "Source-driven packet marking for speech transmission over differentiated-service networks," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Salt Lake City, UT, May 2001, pp. 753–756.
[17] M. Mirhakkak, N. Schult, and D. Thomson, "Dynamic bandwidth management and adaptive applications for a variable bandwidth wireless environment," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 1984–1997, Oct. 2001.
[18] A. Ortega and K. Ramchandran, "Rate distortions for image and video compression," *IEEE Signal Processing Mag.*, vol. 15, pp. 23–50, Oct. 2001.
[19] V. Padmanabhan, "Using differentiated services mechanisms to improve network protocol and application performance," presented at the *IEEE RTAS Workshop QoS Support for Real-Time Internet Applications*, Vancouver, Canada, June 1999.
[20] R. Ramanathan and R. Hain, "An ad hoc wireless testbed for scalable, adaptive QoS support," in *Proc. IEEE WCNC*, Chicago, IL, Nov. 2000, pp. 998–1002.
[21] D. L. Reyes, A. R. Reibman, S.-F. Chang, and J. I.-I. Chuang, "Error-resilient transcoding for video over wireless channels," *IEEE J. Select. Areas Commun.*, vol. 18, pp. 1063–1074, June 2000.
[22] A. Sehgal and P. A. Chou, "Cost-distortion optimized streaming media over DiffServ networks," in *Proc. IEEE Int. Conf. Multimedia and Expo.*, Lausanne, Aug. 2002, pp. 857–860.
[23] S. D. Servetto, K. Ramchandran, K. Nahrstedt, and A. Ortega, "Optimal segmentation of a VBR source for its parallel transmission over multiple ATM connections," in *Proc. IEEE Int. Conf. Image Processing*, Santa Barbara, CA, Oct. 1997, pp. 5–8.
[24] J. Shin, J. Kim, and C.-C. Jay Kuo, "Quality-of-service mapping mechanism for packet video in differentiated services network," *IEEE Trans. Multimedia*, vol. 3, pp. 219–231, June 2001.
[25] I. Stoica, S. Shenkar, and H. Zhang, "Core-stateless fair queueing: Achieving approximately fair bandwidth allocations in high speed networks," *IEEE/ACM Trans. Networking*, vol. 11, pp. 33–46, Feb. 2003.
[26] W. Tan and A. Zhakor, "Packet classification schemes for streaming MPEG video over delay and loss differentiated networks," presented at the *IEEE Packet Video Workshop 2001*, Kyongju, Korea, Apr. 2001.
[27] W. Turin, *Digital Transmission Systems*. New York: McGraw-Hill, 1998.
[28] B. Vandalore, R. Jain, S. Fahmy, and S. Dixit, "AQuaFWiN: Adaptive QoS framework for multimedia in wireless networks and its comparison with other QoS frameworks," in *Proc. IEEE Local Computer Networks*, Boston, MA, Oct. 1999, pp. 88–97.
[29] H. Wang and N. Moayeri, "Finite state Markov channel—A useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, pp. 163–171, Feb. 1995.
[30] F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 332–344, Mar. 2001.
[31] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, pp. 630–643, July 2003.
[32] L. Xiao, M. Johansson, H. Hindi, S. Boyd, and A. Goldsmith, Joint optimization of communication rates and linear systems, submitted for publication.
[33] Z.-L. Zhang, "End-to-end support for statistical quality-of-service guarantees in multimedia networks," Ph.D. Dissertation, Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, Feb. 1997.

Wuttipong Kumwilaisak (S'02–M'03) received the B.E. degree from Chulalongkorn University, Bangkok, Thailand, in 1995, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, in 2003, all in electrical engineering.

He was a Research Intern at Ericssion Euro Laboratory, Aachen, Germany, and at Microsoft Research Asia, Beijing, China, in the summers of 2001 and 2002, respectively. He is currently with the mobile platform solution team at Samsung Electronics, Suwon, Korea. His current research interests include multimedia communication, multimedia processing, and wireless communication.



Y. Thomas Hou (S'91–M'98) received the B.E. degree (*summa cum laude*) from the City College of New York in 1991, the M.S. degree from Columbia University, NY, in 1993, and the Ph.D. degree from Polytechnic University, Brooklyn, NY, in 1998, all in electrical engineering.

From 1997 to 2002, he was a Research Scientist and Project Leader at Fujitsu Laboratories of America, IP Networking Research Department, Sunnyvale, CA (Silicon Valley). He is currently an Assistant Professor at Virginia Tech, The Bradley Department of Electrical and Computer Engineering, Blacksburg, VA. His research interests include wireless video sensor networks, multimedia delivery over wireless networks, scalable architectures, protocols, and implementations for differentiated services Internet, and service overlay networking.

Dr. Hou has published extensively in the above areas and is a corecipient of the 2002 IEEE International Conference on Network Protocols Best Paper Award and the 2001 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award. He is a Member of the Association for Computing Machinery (ACM).

Qian Zhang (M'00) received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 1994, 1996, and 1999, respectively, all in computer science. She joined Microsoft Research, Asia, Beijing, China, in July 1999, as an Associate Researcher in the Internet Media Group and now is a Researcher of the Wireless and Networking Group.

She has published over 40 refereed papers and is the inventor of several pending patents. Her current research interest includes multimedia delivery over wireless, Internet, next-generation wireless networks, P2P network/ad hoc network. Currently, she is participating in TCP/IP header compression in ROHC WG in IETF. She is the principal contributor of the IETF ROHC-TCP WG draft.

Wenwu Zhu (S'91–M'96–SM'01) received the B.E. and M.E. degrees from National University of Science and Technology, China, in 1985 and 1988, respectively, the M.S. degree from Illinois Institute of Technology, Chicago, IL, and the Ph.D. degree from Polytechnic University, Brooklyn, NY, in 1993 and 1996, respectively, all in electrical engineering.

He joined Microsoft Research Asia, Beijing, China, in October 1999, as a Researcher in the Internet Media Group and now is a Research Manager of the Wireless and Networking Group. Prior to his current post, he worked at Bell Labs, Lucent Technologies, Holmdel, NJ, as a Member of Technical Staff from July 1996 to October 1999. While he was at Bell Labs, he performed research and development in the areas of Internet video, video conferencing, and video streaming over IP networks. He has published over 100 refereed papers in the international leading journals and conferences in the areas of wireless/Internet video transport, wireless/Internet multimedia communications and networking, and multimedia signal processing. He is the inventor of more than a dozen of pending patents. His current research interests are in the areas of wireless/Internet multimedia communications and networking.

C.-C. Jay Kuo (S'83–M'86–SM'92–F'99) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, R.O.C., in 1980, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, MA, in 1985 and 1987, respectively, all in electrical engineering.

He was a Computational and Applied Mathematics (CAM) Research Assistant Professor in the Department of Mathematics, University of California, Los Angeles, from October 1987 to December 1988. Since January 1989, he has been with the Department of Electrical Engineering, Systems and the Signal and Image Processing Institute, University of Southern California, Los Angeles, where he currently has a joint appointment as Professor of Electrical Engineering and Mathematics. His research interests are in the areas of digital signal and image processing, audio and video coding, media communication technologies and delivery protocols, and network computing. He is a coauthor of more than 500 technical publications in international conferences and journals, as well as the following books *Content-Based Audio Classification and Retrieval for Audio-visual Data Parsing* (with T. Zhang) (Norwell, MA: Kluwer, 2001), *Semantic Video Object Segmentation for Content-Based Multimedia Applications* (with J. Guo) (Norwell, MA: Kluwer, 2001), *Intelligent Systems for Video Analysis and Access over the Internet* (with W. Zhou) (Englewood Cliffs, NJ: Prentice Hall, 2002), *Quality of Service Provisioning for Multimedia Applications in Service Differentiation Networks* (in preparation with J. Shin and D. Lee) (Englewood Cliffs, NJ: Prentice Hall, 2003).

Dr. Kuo received the National Science Foundation Young Investigator Award (NYI) and the Presidential Faculty Fellow (PFF) Award in 1992 and 1993, respectively. He is a Member of the Society for Industrial and Applied Mathematics (SIAM), Association for Computing Machinery (ACM), and a Fellow of The Internationl Society for Optical Engineers (SPIE). He is Editor-in-Chief for the *Journal of Visual Communication and Image Representation*, Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, Editor for the *Journal of Information Science and Engineering*, and the RURASIP *Journal of Applied Signal Processing*. He served as Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1995 to 1998, and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 1995 to 1997.

Ya-Qin Zhang (S'87–M'90–SM'93–F'97) received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China (USTC) in 1983 and 1985, and the Ph.D. degree in electrical engineering from George Washington University, Washington DC, in 1989. He received executive business training from Harvard University, Cambridge, MA.

He is the Managing Director of Microsoft Research Asia, Beijing, China, which he joined in January 1999. Before that, he was Director of Multimedia Technology Laboratory, Sarnoff Corporation, Princeton, NJ (formerly David Sarnoff Research Center, and RCA Laboratories). He has been engaged in research and commercialization of MPEG2/DTV, MPEG4/VLBR, and multimedia information technologies. He was with GTE Laboratories, Inc., Waltham, MA, and Contel Technology Center, VA, from 1989 to 1994. He has authored and coauthored over 200 refereed papers in leading international conferences and journals. He has been granted over 40 U.S. patents in digital video, Internet, multimedia, wireless and satellite communications. Many of the technologies he and his team developed have become the basis for start-up ventures, commercial products, and international standards.

Dr. Zhang served as the Editor-in-Chief for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from July 1997 to July 1999. He was the Chairman of Visual Signal Processing and Communications Technical Committee of IEEE Circuits and Systems. He serves on the Editorial Boards of seven other professional journals and over a dozen conference committees. He has been a key contributor to the ISO/MPEG and ITU standardization efforts in digital video and multimedia. He received numerous awards, including several industry technical achievement awards and IEEE awards such as Jubilee Golden Medal. He was awarded as the "Research Engineer of the Year" in 1998 by the New Jersey Engineering Council for his "leadership and invention in communications technology, which has enabled dramatic advances in digital video compression and manipulation for broadcast and interactive television and networking applications." He received the Prestigious National Award as "The Outstanding Young Electrical Engineering of 1998," given annually to one Electrical Engineer in the U.S.