Cross-Layer Modeling for QoS-Driven Multimedia Multicast/Broadcast over Fading Channels in Mobile Wireless Networks

Xi Zhang and Qinghe Du, Texas A&M University

ABSTRACT

In this article we propose a cross-layer design model for multimedia multicast/broadcast services to efficiently support the diverse quality of service requirements over mobile wireless networks. Specifically, we aim at achieving high system throughput for multimedia multicast/broadcast while satisfying QoS requirements from different protocol layers. First, at the physical layer, we propose a dynamic rate adaptation scheme to optimize the average throughput subject to the loss rate QoS constraint specified from the upper-layer protocol users. We investigate scenarios with either independent and identically distributed (i.i.d.) or non-i.i.d. fading channels connecting to different multicast receivers. Then, applying the *effective capacity theory* at the data link layer, we study the impact of the delay QoS requirement (i.e., QoS exponent) on the multimedia data rate of mobile multicast/broadcast that our proposed scheme can support. Also presented are simulation results which show the trade-off among different QoS metrics and the performance superiority of our proposed scheme as compared to the other existing schemes.

INTRODUCTION

As the demands of group-oriented and wireless-connection-based multimedia services with diverse quality of service (QoS) requirements increase explosively, mobile multimedia multicast/broadcast (one-to-many) communications such as teleconferencing, live video broadcasting, air/highway traffic control, and online gaming have become a critically important component of the next-generation wireless networks. One of the most important characteristics of mobile multicast[1] is highly efficient group communications with significantly reduced wireless resource consumption. In particular, by taking advantage of wireless channels' broadcast nature, radio signals sent by the sender can be captured by all multicast receivers within reception range. Clearly, this feature is very attractive for downlink transmissions in widely used cellular architectures (e.g., third-/fourth-generation cellular networks and 802.16x networks). Consequently, mobile multicast in these types of network architectures has received more and more research attention [1–3].

Besides broadcast, time-varying fading is the other essential feature of wireless channels. Usually, adaptive modulation and coding (AMC) [4, 5] at the physical layer provides efficient and promising ways to *adapt transmission rates* to variations in wireless channel quality such that the specified QoS requirements can be supported, including high throughput, low delay, and limited loss rate. However, the heterogeneous channel qualities, both instantaneous and statistical, among different receivers present great challenges in supporting the diverse QoS requirements (multiple QoS metrics) of multimedia multicast services. For instance, consider a multicast session in cellular networks (downlink transmission) where the sender transmits a single stream to all multicast receivers. On one hand, to guarantee reliability at the physical layer, the transmission rate has to be confined to the capability of the worst case instantaneous channel quality over all multicast receivers, which results in severely low multicast throughput and long delay. In particular, from an information theory perspective, the authors of [1] investigated physical-layer multicast capacity with multiple transmit antennas at the sender, where the worst case instantaneous achievable throughput over all multicast receivers is optimized. They showed that with a finite number of antennas, the multicast capacity tends to be zero as the multicast group size becomes large. While they proved that a non-zero limiting capacity with an infinite multicast group size for multicast can be achieved by using an infinite number of transmit antennas, only a small number of antennas are typically used in most practical systems.

On the other hand, if applying higher transmission rates for multicast, data losses often occur at receivers with poorer instantaneous

The research reported in this article was supported in part by the U.S. National Science Foundation CAREER Award under Grant ECS-0348694.

[1] *Broadcast (to all receivers) is a special or extreme case of multicast (to a group of specified receivers); thus, we use the term* multicast *to refer to broadcast and all types of one-to-many communications services in the rest of this article.*

0163-6804/07/$20.00 © 2007 IEEE

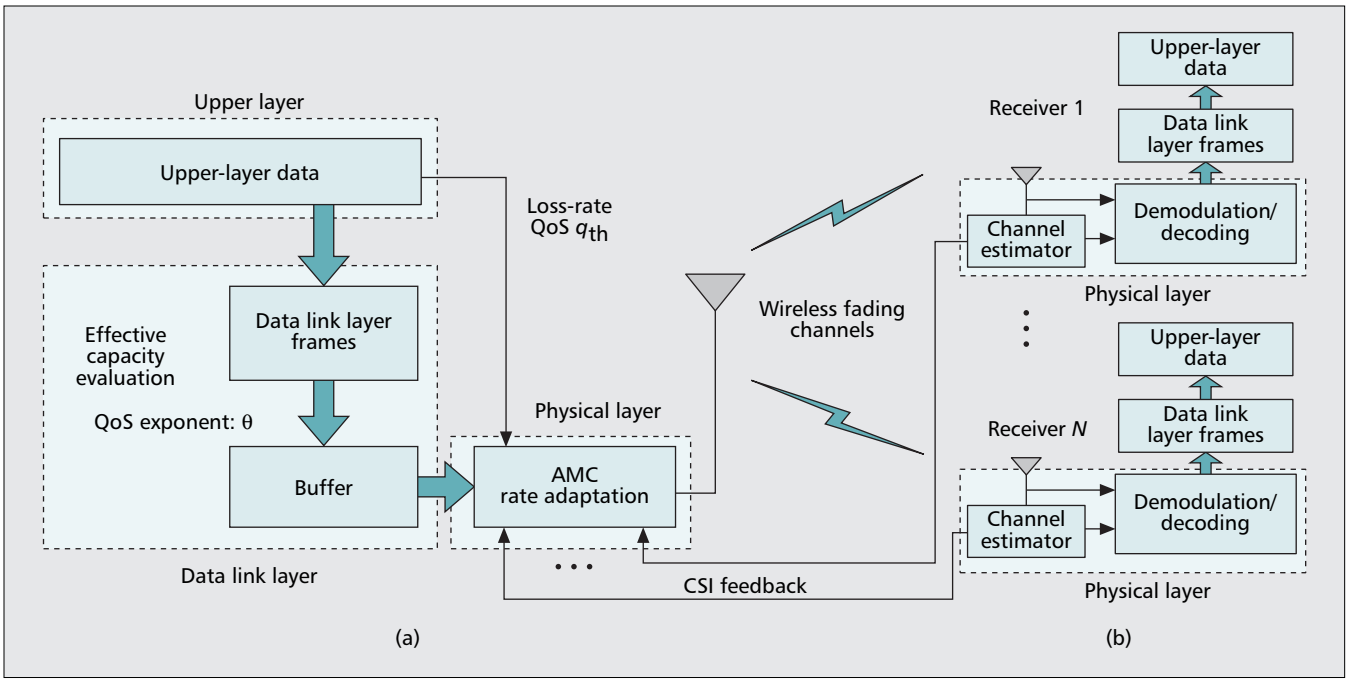

Figure 1. *The cross-layer design model for multimedia multicast in cellular wireless networks: a) the base station sender; b)* N *mobile multicast receivers.*

channel qualities. Although with real-time requirements multimedia services can usually tolerate a certain level of data loss to trade off for high throughput (or low delay), the data loss rate cannot be arbitrarily high, depending on the requirements of the upper-layer protocols. Consequently, the efficient design and implementation of adaptive transmissions for mobile multimedia multicast cannot be achieved only at the physical layer, but needs to be integrated with other QoS constraints from the upper protocol layers such that diverse QoS requirements can be supported. In [3] the authors studied the throughput-delay trade-off for cellular multicast. For a single multicast session, a static rate control algorithm was studied in [3], where the transmission rate is always dominated by the signal-to-noise ratio (SNR) which is in a fixed quantile of the ordered SNRs over all multicast receivers. However, this static rate control algorithm is not as efficient as the dynamic rate control/adaptation strategy.

To overcome the aforementioned problems, we propose a cross-layer design model for multimedia multicast services to efficiently support diverse QoS requirements over mobile wireless networks. Specifically, we aim at maximizing multimedia multicast throughput through QoS information exchanges across different protocol layers. First, at the physical layer, we propose a *dynamic* rate adaptation scheme to optimize the average throughput subject to the loss rate QoS constraint specified by upper-layer protocol users. Then, applying the effective capacity theory [6, 7] at the data link layer, we study the impact of a delay QoS constraint (i.e., QoS exponent) on the data rate of mobile multimedia multicast traffic that can be supported. Simulation results are also provided to show the trade-off among different QoS metrics and the performance superiority of our proposed scheme over existing schemes.

The rest of this article is organized as follows. We start with presenting the cross-layer system model. Then we propose a dynamic rate adaptation scheme at the physical layer subject to the loss rate QoS constraint specified at upper layers. Finally, employing the QoS exponent as the delay QoS constraint at the data link layer, we evaluate the effective capacity performance of our proposed scheme through simulations, which is then followed by the article's conclusion.

The Cross-Layer Design System Model

System Model

We concentrate on a discrete time mobile multicast system in cellular networks, where the sender (base station) attempts to directly transmit a *single* multimedia data stream to N (the multicast group size) mobile multicast receivers through broadcast wireless channels, which is illustrated in Fig. 1. The sender and all multicast receivers use single-transmit/-receive antennas for multicast transmission. As shown in Fig. 1, the upper layer data with *constant* arrival rate is divided into frames at the data link layer and stored in the buffer. We assume that the time used to transmit any frame is the same and equal to T, and the amount of data contained in a frame varies with the transmission rate used at the physical layer. The frames are then converted into a bitstream and finally fed into the physical layer. We denote the total physical layer signal bandwidth for multimedia multicast by B. The sender uses AMC techniques to adapt the transmission rate according to the channel state

It should be noted that the delay-constrained multimedia services usually can tolerate a certain level of data loss. However, how much data loss can be tolerable depends on the requirement from upper layer protocols.

information (CSI) fed back from all multicast receivers and the loss rate QoS constraint, denoted q_{th}, specified by the upper-layer protocols. Each multicast receiver performs the reverse operation to recover the data for the upper-layer user when it receives the multicast signal transmitted through wireless fading channels. The constant data rate of multimedia source data that can be supported is eventually evaluated at the data-link layer by using the effective capacity theory subject to a QoS delay constraint, named the QoS exponent and denoted θ. As discussed above, there are two QoS constraints of interest, the loss rate constraint q_{th} and the QoS exponent θ, which are both statistical QoS metrics. The value of q_{th} characterizes the loss rate constraints from upper protocol layers, while the QoS exponent θ can be used to estimate and evaluate the violation probability of the specified delay bound or queue (first in first out) length threshold. These two metrics will be elaborated on in more detail later.

The Wireless Channel Model

The fading channel between the sender and a receiver for a communication pair is modeled as a discrete time, ergodic, flat block fading stochastic process. In particular, the instantaneous received SNR (defined as the ratio of the received signal power to additive white Gaussian noise power within the signal bandwidth B at the receiver end) of the nth receiver, denoted γ_n, is invariant within a frame's time duration, but varies independently among different frames. Without loss of generality, we form all γ_ns as an N-dimension SNR vector $\boldsymbol{\gamma}$ to characterize a particular fading state of the corresponding CSI. Moreover, we use the subscript $\pi(\cdot)$ to denote the permutation of γ_ns in a fading state such that $\gamma_{\pi(i)}$ decreases as i increases from 1 to N. Throughout this article, we assume that $\boldsymbol{\gamma}$ follows continuous distribution (the most widely used fading models, e.g., Rayleigh and Nakagami models, fall into this category); that is, the cumulative distribution function (CDF) is continuous. Also, we assume that CSI $\boldsymbol{\gamma}$ is perfectly estimated at multicast receivers and is reliably fed back to the sender without delay.

Cross-Layer Design Problem Formulation of Rate Adaptation for Mobile Multimedia Multicast with Statistical QoS Provisioning

Physical-Layer Rate Adaptation Subject to Loss Rate QoS Requirements from Upper Layers

The sender uses a constant transmit power, but dynamically regulates the transmission rate through AMC as depicted in Fig. 1. In this article we focus on rate adaptation for Shannon-capacity-based multicast systems. Specifically, for the transmission in a frame, the sender dynamically picks an SNR from γ_ns as the *dominating* SNR, denoted γ_{dom}, and then set the transmission rate equal to the Shannon capacity with γ_{dom}, that is, $B \log_2(1 + \gamma_{\text{dom}})$ b/s. Thus, given a $\boldsymbol{\gamma}$, the transmission rate can take N possible values within a frame. To get a more general framework, we allow the *time-sharing* strategy [10] to be applied among these N possible values. In particular, given any fading state we allocate the time proportion, denoted λ_n ($0 \le \lambda_n \le 1$, $n = 1, 2, \ldots, N$), for the transmission rate corresponding to $\gamma_{\text{dom}} = \gamma_n$, where λ_ns are the functions of $\boldsymbol{\gamma}$, and the sum of λ_ns is equal to 1. We further use all λ_ns to construct an N-dimension function vector $\boldsymbol{\lambda}$ to completely characterize the *rate adaptation policy*. Thus, given a transmission rate with γ_{dom}, the nth receiver can correctly receive the signal with this rate only when γ_n is greater than or equal to γ_{dom}. On the other hand, if γ_n is smaller than γ_{dom}, the nth receiver has to drop the received signal, implying zero instantaneous data rate. Now, we need to introduce some variables as follows. We call the transmission rate averaged within a frame the *instantaneous throughput* and denote it R. The instantaneous data rate achieved by the nth receiver, denoted g_n, is then called *instantaneous goodput*. Correspondingly, we use $\mathbb{E}\{R\}$ and $\mathbb{E}\{g_n\}$ to represent the average throughput and average goodput of the nth receiver, respectively, where $\mathbb{E}\{\cdot\}$ denotes the expectation over all fading states.

If setting $\lambda_\pi(N) = 1$ for all $\boldsymbol{\gamma}$s, where $\lambda_\pi(i)$ represents the time proportion for the transmission rate with $\gamma_{\text{dom}} = \gamma_{\pi(i)}$, the transmission rate is always determined by the *worst case* instantaneous SNR among all receivers; thus, there are no losses for any receivers. We call it the worst case SNR dominating (WSD) scheme. Otherwise, data loss is usually unavoidable for receivers with low instantaneous SNRs. It should be noted that the delay-constrained multimedia services can usually tolerate a certain level of data loss. However, how much data loss can be tolerated depends on the requirements of upper-layer protocols. In this article we use a loss rate threshold, termed the *loss rate QoS* and denoted q_{th}, to characterize this requirement. Specifically, the loss rate of the nth receiver, denoted q_n, is defined as the ratio of the amount of data (averaged over all fading states) not correctly received by the nth receiver to that of the transmitted multicast data (averaged over all fading states), and q_ns of all receivers are required not to exceed q_{th}. Under the above system settings, we can derive efficient rate adaptation schemes by solving the following average throughput optimization problem:

$$\max_{\boldsymbol{\lambda}:\lambda_n \ge 0, \sum_{n=1}^{N} \lambda_n = 1} \{\mathbb{E}\{R\}\}, \text{ subject to:}$$
$$q_n \triangleq 1 - \frac{\mathbb{E}\{g_n\}}{\mathbb{E}\{R\}} \le q_{th}, n = 1, 2, \ldots N. \quad (1)$$

Effective Capacity with Statistical Delay QoS Guarantees

By optimizing average throughput, we can achieve low average delay for multimedia multicast transmission. However, the metric of average throughput itself is not enough to characterize the delay QoS provisioning. Because

of random fading over wireless channels, the hard delay bound required for delay-sensitive multimedia services is usually difficult to guarantee. Thus, we consider statistical delay QoS guarantees instead. A statistical QoS requirement measure, denoted θ and called the QoS exponent, was developed to characterize delay QoS provisioning in [6, 8]. Specifically, θ is a positive real-valued number that connects a specified queue length threshold, denoted Q_{th}, with the probability that the queue length, denoted Q, exceeds Q_{th} during transmission. Note that for our mobile multimedia multicast model, the queue is measured at the buffer in the data link layer as shown in Fig. 1. It was shown that with stationary and ergodic arrival and service processes under certain conditions, the probability that Q exceeds Q_{th} can be approximated as $e^{-\theta Q_{th}}$ for a large Q_{th} [8] and $\beta e^{-\theta Q_{th}}$ for a small Q_{th} [6], where β is the probability that the buffer is not empty. Note that if we map the queue length threshold to the delay bound, we can obtain similar approximations to those above in terms of delay bound violation probability. Clearly, the parameter θ measures the exponential decaying rate of the threshold violation probability. A larger θ corresponds to a more stringent delay QoS requirement, which implies a higher decaying rate of the QoS violation probability as a function of the specified Q_{th}, and vice versa.

Using the QoS exponent θ, the authors of [6] proposed a useful concept, *effective capacity*, defined as the maximum constant arrival rate which can be supported by a certain random service process (e.g., an adaptive transmission rate over a wireless fading channel) subject to a specified θ. For the flat block fading channel with independent variation across different frame durations, the effective capacity (bits per frame), denoted $E_C(\theta)$, of a rate adaptation scheme with the specified QoS exponent θ is determined [7] by

$$E_C(\theta) = \frac{1}{\theta}\log\left(\mathbb{E}\left\{e^{-\theta TR}\right\}\right). \quad (2)$$

Correspondingly, $E_C(\theta)/(TB)$ denotes the normalized effective capacity (bits per second per hertz). In this article, in order to observe the impact of delay QoS constraint on the achievable data rate for mobile multimedia multicast, we use effective capacity as the main performance metric to evaluate our derived rate adaptation schemes.

To better understand the effective capacity, we summarize a number of main properties of effective capacity [5] as follows. The effective capacity $E_C(\theta)$ is a monotonically decreasing function of θ, which implies that more stringent QoS requirements result in lower supportable service rates. As θ approaches 0, the service does not impose any constraint on the queue length and delay bound, and thus the effective capacity converges to the average throughput $\mathbb{E}\{R\}$. In contrast, when θ approaches ∞, implying the zero delay requirement, the effective capacity degrades to the minimum service rate over all fading states. Therefore, in [5] we proposed using the effective capacity functions as a bridge for cross-layer design modeling between the physical and data link layers with the delay statistical QoS guarantees over unicast wireless networks.

Rate Adaptation Subject to Loss Rate QoS Constraints

We focus on ideas about the trade-off between average throughput and loss rate such that efficient rate control can be achieved. We begin with presenting the solution to the optimization problem given by Eq. 1 for the scenario where the channel fading for different sender-receiver pairs are i.i.d. Also, we discuss the existing rate adaptation schemes. Moreover, we investigate how to efficiently apply the results obtained for i.i.d. fading environments in more practical, but more complex, non-i.i.d. fading scenarios.

Independent and Identically Distributed Fading Channels across Multicast Receivers

Adaptive Dominating Position Scheme — Using the time-sharing strategy, the highest average throughput without loss rate constraint is obtained by setting $\lambda_{\pi(1)} = 1$ for all γs. We name this policy the best case SNR dominating (BSD) scheme. If we employ the BSD in i.i.d. fading environments (i.e., γ_ns are i.i.d. over all sender-receiver pairs), the loss rates of all receivers are equal to $(N-1)/N$. Therefore, even if $q_{th} \geq (N-1)/N$, the optimal solution to Eq. 1 still yields the BSD scheme. In contrast, when $q_{th} = 0$, no data loss is tolerable for any receiver, and then we obtain $\lambda_\pi(N) = 1$ for all γs (i.e., the WSD scheme). Thus, we only need to focus on the cases with $0 \leq q_{th} \leq (N-1)/N$.

Under i.i.d. fading environments, it is clear that the optimal solution to Eq. 1 should benefit all receivers evenly. That is, the average goodput $\mathbb{E}\{g_n\}$s are equal for all receivers, and so are the loss rates q_ns. Consequently, the loss-rate constraints in Eq. 1 can be equivalently rewritten as:

- $\mathbb{E}\{g_n\}$'s are equal for all N receivers.
- $q_0 \leq q_{th}$, where $q_0 \triangleq 1 - \mathbb{E}\{g_{sum}\}/\mathbb{E}\{R\}$ called the *group loss rate* and $g_{sum}(R)$ is the sum of all g_ns normalized by N.

Correspondingly, we call $g_{sum}(R)$ and $\mathbb{E}\{g_{sum}\}$ the *instantaneous sum goodput* (normalized by N) and *average sum goodput* (normalized by N), respectively. In the following, we first concentrate on maximizing the average throughput only based on the second constraint above (i.e, $q_0 \leq q_{th}$). We will point out later that with i.i.d. γ_n's, the derived policy also satisfies the first constraint.

The value of q_0 is determined by the instantaneous throughput R and sum goodput g_{sum} in all fading states. In each fading state, we define N operating points as $(R_{\pi(i)}, g_{sum}(R_{\pi(i)}))$, $i = 1, 2, \ldots, N$, in the "instantaneous throughput — sum goodput" plane, as shown in an example for $N = 5$ in Fig. 2, where $R_{\pi(i)}$ is equal to the transmission rate with $\gamma_{dom} = \gamma_{\pi(i)}$, and $g_{sum}(R_{\pi(i)})$ represents g_{sum} obtained with $\gamma_{\pi(i)} = 1$. Note that in Fig. 2, the horizontal and vertical axes represent the instantaneous throughput and instantaneous sum goodput, respectively. Since we employ the time sharing strategy, points resulting from all possible rate adaptation policies form the convex hull [9] of the set of the N operating points, which is marked as the shaded region in Fig. 2.

Because of random fading over wireless channels, a hard delay bound required for delay-sensitive multimedia services is usually difficult to guarantee. Thus, we consider the statistical delay QoS guarantees instead.

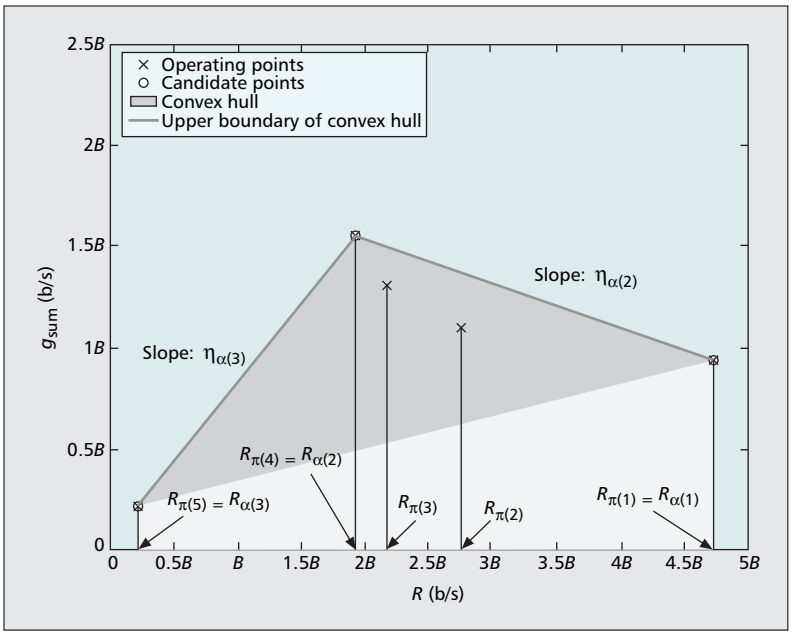

Figure 2. *An example of the upper boundary of convex hull formed by all operating points in the "instantaneous throughput – sum goodput" plane for a fading state, where* B *is the signal bandwidth,* N = *5,* $\lambda_{\pi(1)}$ = *14.08 dB,* $\lambda_{\pi(2)}$ = *7.56 dB,* $\lambda_{\pi(3)}$ = *5.45 dB,* $\lambda_{\pi(4)}$ = *4.48 dB,* $\lambda_{\pi(5)}$ = *–8.03 dB, and* N = *3.*

Because operating points are discrete, the convex hull must be in the shape of a polygon and its vertexes must be some operating points.

However, rather than the entire convex hull, we are more interested in its upper boundary, which is marked with bold lines in Fig. 2. Specifically, the upper boundary of the convex hull is defined as the set of all points each of which has the maximum instantaneous sum goodput over all points in the convex hull with the same instantaneous throughput. To completely characterize the upper boundary of the convex hull, we only need the operating points that are vertices of the convex hull and also on the upper boundary. We call these operating points *candidate points* (Fig. 2) and use N to represent the number of them. Note that N may be different for different $\boldsymbol{\gamma}$s. Moreover, we denote the transmission rates corresponding to the candidate points by $R_{\alpha(j)}$, j = 1, 2, …, N, where $\alpha(\cdot)$ represents the permutation of the transmission rates associated with candidate points such that $R_{\alpha(j)}$ decreases as j increases. It is not difficult to show that only under the constraint $q_0 \le q_{th}$, the point generated by the rate adaptation policy that maximizes the average throughput must fall onto the upper boundary in any fading state. Thus, we only need to perform time-sharing strategies among candidate points. Accordingly, we let $\lambda_{\alpha(j)}$ denote the time proportion allocated to transmission rate $R_{\alpha(j)}$, j = 1, 2, …, N, which can equivalently characterize the rate adaptation policy $\boldsymbol{\lambda}$.

We further introduce the concept of instantaneous sum goodput-throughput gain (IGTG), denoted $\eta_{\alpha(j)}$, j = 2, …, N, which represents the slope of the straight line connecting two neighboring candidate points indexed by $\alpha(j)$ and $\alpha(j-1)$. It is worth noting that the upper boundary characterizes a concave function of the instantaneous throughput; hence, $\eta_{\alpha(j)}$ is a decreasing and an increasing function of R and j, respectively. In addition, we derive that $1 > \eta_{\alpha(j)} > -\infty$ always holds and correspondingly define $\eta_{\alpha(1)} \triangleq -\infty$ and $\eta_{\alpha(N+1)} \triangleq 1$.

Using the concept of IGTG, the policy maximizing the average throughput only under the constraint $q_0 \le q_{th}$ can be characterized by a specified global IGTG threshold, denoted η_{th}. In particular, in each fading state we set the rate adaptation policy as $\lambda_{\alpha(j^*)} := 1$ and $\lambda_{\alpha(j)} := 0$ for $j \ne j^*$, where j^* is obtained by solving the following inequality:

$$\eta_{\alpha(j^*+1)} \ge \eta_{th} > \eta_{\alpha(j^*)}, \tag{3}$$

where η_{th} is uniquely determined through the condition $q_0 = q_{th}$. The procedures to perform the above policy are summarized in Algorithm 1, which is called the adaptive dominating position (DP) scheme because the dominating SNR γ_{dom} is not located in a fixed position in the ordered SNR sequence of $\gamma_{\pi(i)}$s.

We explain the optimality of the above policy as follows. Clearly, $\eta_{th} \to -\infty$ and $\eta_{th} = 1$ result in the BSD and WSD schemes, respectively. Starting from the WSD policy, gradually decreasing the IGTG threshold η_{th} increases both instantaneous throughput R in all fading states and the group loss rate q_0. Also, since γ_ns are assumed to have continuous CDF, q_0 is a continuous function of η_{th}, and there must exist such a η_{th} resulting in $q_0 = q_{th}$. Considering the fact that the IGTG $\eta_{\alpha(j)}$ monotonically decreases with the increase of R, we observe that Algorithm 1 guarantees obtaining the maximum $\mathbb{E}\{g_{sum}\}$ over all policies achieving the same $\mathbb{E}\{R\}$. In other words, Algorithm 1 always takes advantage of larger IGTG such that with the same average throughput, q_0 is minimized. Therefore, among all policies achieving the same q_0, Algorithm 1 obtains the maximum average throughput. Moreover, because Algorithm 1 operates only on the ordered SNR sequence $\{\gamma_{\pi(i)}\}_{i=1}^{N}$, in i.i.d. fading environments all multicast receivers will evenly benefit. Thus, all receivers eventually achieve equal data rates; that is, the first constraint discussed earlier in this section is satisfied, which implies that Algorithm 1 is also the solution to Eq. 1. Based on the above discussions, Algorithm 1 maximizes the average throughput with $q_0 = q_{th}$, and the i.i.d. fading channels guarantee all receivers to achieve the same loss rate under Algorithm 1. For non-i.i.d. fading environments, Algorithm 1 usually does not generate the same loss rates for multicast receivers. However, if statistical channel qualities of different receivers are similar, their loss rates are often still close to the loss rate threshold q_{th}. In contrast, if the differences among their statistical channel qualities are significant, the loss rates of receivers with statistically poorer channels will usuallybe much higher than q_{th} under Algorithm 1.

Adaptively Tracking the IGTG Threshold — The general analytical expression for the global IGTG threshold η_{th} does not exist. In contrast, the value of η_{th} can be determined through numerical or simulation-based methods. These methods need the channels' statistical information, such as the probability density function (pdf) of $\boldsymbol{\gamma}$, which is usually not available in prac-

tical systems. To solve this problem, we employ a simple gradient-descent-based method [9] to dynamically track the IGTG threshold even in the absence of channel statistical information, which is described as follows. We use t, $t = 1, 2, \ldots$, as the time index for frames. For the tth frame, we use the symbol $[t]$ to index the estimate of corresponding variables. We initialize $\eta_{th}[1]$ as 1 (i.e., starting with WSD scheme). For every frame, we apply the estimate $\eta_{th}[t]$ as the IGTG threshold to perform rate adaptation based on Algorithm 1, and then calculate $\eta_{th}[t + 1]$ for the next frame as $\eta_{th}[t] + \varepsilon(q_0[t] - q_{th})$. The above iterative update attempts to minimize the cost function $(q_0 - q_{th})^2$ such that the true IGTG threshold can be approached. In particular, $(q_0[t] - q_{th})$ estimates the negative direction of the cost function's gradient with respect to η_{th}, and ε is a small positive step size.

After transmitting the tth frame, we first estimate the average goodput $\mathbb{E}\{g_n\}$ for each receiver and the average throughput by using first-order low-pass autoregressive (AR) filters. With the obtained results, we further estimate the average sum goodput and group loss rate $q_0[t]$ through their definitions, which will then be used for IGTG tracking. Under the above tracking algorithm, $\eta_{th}[t]$ is expected to converge and oscillate within a small range around the true value of the IGTG threshold. It is also worth noting that the above tracking method can adapt to the variation of multicast group size or channels' statistical features. In our simulations, the above method converges quickly for $q_{th} < 0.4$ under various settings, which has covered a wide range of loss rate constraints tolerable in practical systems. If q_{th} is too high, especially when it is close to $(N - 1)/N$, η_{th} becomes a very large negative real number; then the convergence speed is slow, which causes a certain throughput degradation. However, most practical multimedia services will not usually support such a high loss rate threshold.

Fixed Dominating Position Scheme — Compared to the adaptive DP scheme, a simpler suboptimal scheme is to always let $\lambda_{\pi(k)} = 1$ with a fixed k (a similar strategy was discussed in [3]) over all fading states. Thus, we call this strategy the fixed dominating position scheme. The loss rate under this scheme equals $(1 - k/N)$ for all receivers in i.i.d. fading environments, and k needs to be selected such that the achieved loss rate is just below q_{th}, that is, $k = \lceil \Upsilon(1 - q_{th})N \rceil$.

Non-Independent and Identically Distributed Fading Channels across Multicast Receivers

For non-i.i.d. fading environments, Algorithm 1 is not optimal to the problem given in Eq. 1 because Algorithm 1 is not aware of differences among the statistical channel information of multicast receivers. Once the differences among receivers' statistical channel qualities are significant, Algorithm 1 often causes severe intrasession unfairness. In particular, multicast receivers with statistically poorer channels suffer from high-level data losses, and the loss rate QoS will be violated.

In fading state γ:
Step 1: Determine N operating points, N candidate points, and IGTG $\eta_{\alpha(j)}$, $j = 1; 2, \ldots, N + 1$.
Step 2: Find j^* satisfying Eq. 3. Set $\lambda_{\alpha(j^*)} := 1$ and $\lambda_{\alpha(j)} := 0$ for $j \neq j^*$.

Algorithm 1. *Adaptive DP scheme.*

However, note that the derivation of the optimal solution to Eq. 1 with non-i.i.d. channels is much more challenging. Hence, we turn to designing suboptimal schemes. Specifically, we employ a suboptimal strategy, called *adaptive subgrouping*. Integrating this strategy, we can efficiently apply Algorithm 1 to non-i.i.d. fading environments. The idea of this scheme is explained as follows.

In non-i.i.d. fading environments, an efficient method of rate adaptation is to determine the transmission rates only based on the CSI from receivers with statistically poorer channel qualities. On one hand, this can efficiently support loss rate QoS for these receivers. On the other hand, the loss rate constraint is not usually violated for other receivers, because they have statistically better channels. To use the above strategy, we need to estimate receivers' statistical channel qualities. After transmitting the tth frame, we calculate $\bar{g}_n[t]$ of the nth receiver, which equals the time average of its instantaneous goodput in previous frames. If we apply Algorithm 1 to the entire multicast group, after transmitting a certain number of frames lower $\bar{g}_n[t]$ usually implies statistically poorer channel qualities.

Motivated by the above analyses, we develop the following method. Let us start with applying Algorithm 1 on the entire multicast group. In every $(t + 1)th$ frame duration, we denote the minimum among all $\bar{g}_n[t]$s ($\bar{g}_n[0]$s are set to 0) $\bar{g}_{min}[t]$ and pick all receivers with $\bar{g}_n[t]$ less than or equal to $(1 + \xi)\bar{g}_{min}[t]$ to form a subgroup Ω, where ξ is a small positive number and we denote the number of receivers in Ω $\tilde{N}$. Then we apply Algorithm 1 on Ω (use $\tilde{N}$ instead of N) for rate adaptation. Since Ω usually includes receivers with similarly poor channel qualities, using Algorithm 1 on Ω can efficiently support loss rate QoS for these receivers and increase throughput as well. Moreover, because of the dynamic updating of Ω, we need to apply the method of adaptively tracking the IGTG threshold developed earlier.

As discussed before, Algorithm 1 attempts to maintain the group loss rate as q_{th} while optimizing the average throughput. It should also be noted that receivers in Ω may achieve different, although close, loss rates. As a result, even when the group loss rate of Ω approaches q_{th}, some receivers may still get a loss rate a little higher than q_{th}. To overcome this problem, we can use a more conservative (lower) loss rate threshold $\tilde{q}_{th}$ equal to the maximum between 0 and $1 - (1 + \xi)(1 - q_{th})$ instead of the true q_{th} in Algorithm 1. Finally, we summarize the adaptive subgrouping scheme in Algorithm 2.

In addition, this subgrouping strategy is applicable for the fixed DP scheme in non-i.i.d. fading channels to avoid severe intrasession

unfairness. To attain this we only need to modify step 2 of Algorithm 2 as: using $\tilde{N}$ and $\tilde{q}_{th}$ instead of N and q_{th}, respectively, applying the fixed DP scheme on Ω to determine the transmission rate.

For the transmission of the $(t+1)$-th frame:

Step 1: Determine Ω and the number $\tilde{N}$ of receivers in Ω.

Step 2: Using $\tilde{N}$ and $\tilde{q}_{th}$ instead of N and q_{th}, respectively, apply Algorithm 1 (integrated with the IGTG threshold tracking method developed earlier) on Ω for rate adaptation.

Step 3: Calculate $\bar{g}_n[t+1]$ for all multicast receivers.

Algorithm 2. *Adaptive subgrouping algorithm.*

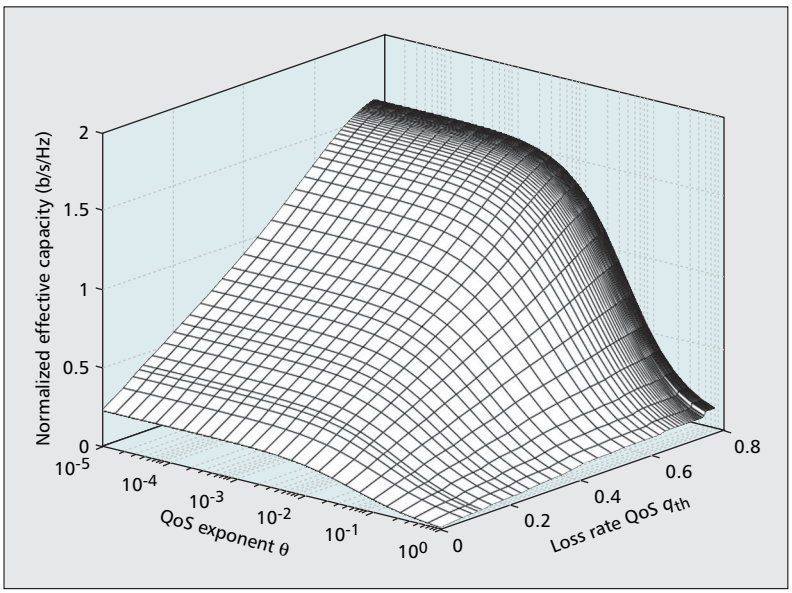

Figure 3. *Normalized effective capacity* $E_C(\theta)/(TB)$ *vs. loss rate QoS* q_{th} *and QoS exponent* θ*, where* $N = 5$ *and* $\{\gamma_n\}_{n=1}^{N}$ *follow i.i.d. Rayleigh distribution with* $\bar{\gamma} = 0$ *dB for each multicast receiver.*

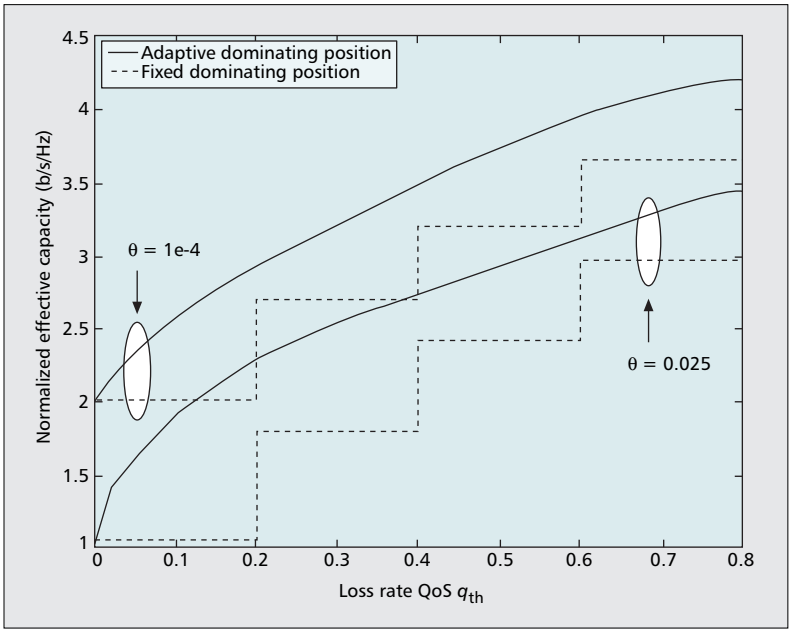

Figure 4. *Normalized effective capacity* $E_C(\theta)/(TB)$ *vs. loss rate QoS* q_{th}*, where* $N = 5$*,* θ *=1e-4, 0.025, and* $\{\gamma_n\}_{n=1}^{N}$ *follow i.i.d. Nakagami-*m *fading with* $m = 2$ *and* $\bar{\gamma} = 10$ *dB for each multicast receiver.*

Note that since Ω is only a subgroup of the entire multicast group and varies dynamically, the dominating SNR may not be always located at a fixed position in the ordered SNR sequence. However, we still use *fixed DP* to refer to this scheme for the convenience of presentation.

Simulation Evaluations

We evaluate the effective capacity performance of our proposed rate adaptation schemes through simulations. In the simulations we set the signal bandwidth $B = 100$ kHz and the frame duration $T = 2$ ms, respectively. Figure 3 plots the dynamics of normalized effective capacity of our derived adaptive DP scheme against the QoS exponent θ and the loss rate QoS requirement q_{th}, where the multicast group size N is 5, fading channels follow i.i.d. Rayleigh distribution across different receivers, and the average SNR, denoted $\bar{\gamma}$, is equal to 0 dB. From Fig. 3, we observe that QoS requirements from different protocol layers have significant impact on the multimedia throughput that can be supported. As shown in Fig. 3, when the QoS requirements becomes stringent (i.e., either θ increases or q_{th} decreases), the effective capacity decreases. When θ gets larger, the delay bound or queue length threshold violation probability needs to be maintained at a lower level, and the system can only support a lower constant multimedia data rate. If q_{th} gets smaller, the transmission rate must be confined to a lower range to avoid severe data loss, which also leads to a degradation of effective capacity. When θ approaches ∞, the effective capacity will approach 0, as shown in Fig. 3, which implies zero delay cannot be guaranteed for this case.

Figure 4 compares the normalized effective capacity between the fixed DP scheme and our proposed adaptive DP scheme with variation of loss rate QoS requirement q_{th}, where $N = 5$ and γ_ns follow i.i.d. Nakagami-m fading with $m = 2$ and $\bar{\gamma} = 10$ dB for all receivers. We see from Fig. 4 that only at $q_{\text{th}} = 0$ and $q_{\text{th}} = (N-1)/N = 0.8$ do both schemes achieve the same effective capacity. This is because for $q_{\text{th}} = 0$ both schemes reduce to the WSD scheme, and for $q_{\text{th}} = (N-1)/N$ both of them converge to the BSD scheme. However, in most other cases where $0 < q_{\text{th}} < (N-1)/N$, our proposed adaptive DP scheme significantly outperforms the fixed DP scheme, which also confirms the advantages of using the adaptive dominating position strategy. Moreover, the adaptive DP scheme can flexibly adapt to the loss rate QoS variation, which makes the achieved effective capacity a continuous function of q_{th}. However, for the fixed DP scheme, the achieved effective capacity is only a step function of q_{th} due to the inflexible strategy. Figure 5 compares the normalized effective capacity between these two schemes against the variation of QoS exponent θ in i.i.d. Nakagami-m fading channels, where $m = 2$, $N = 5$, and $\bar{\gamma} = 10$ dB. As illustrated in Fig. 5, our proposed scheme achieves higher effective capacity than the fixed DP scheme for various θ. Also, as θ increases, the degrading speed of our scheme is clearly much slower than that of the fixed DP scheme. Figure 6 plots the normalized effective

capacity vs. QoS exponent θ in a non-i.i.d. fading environment, where $N = 18$ and γ_ns follow independent Nakagami-m distribution with $m = 2$. All receivers are divided into three subsets, within each of which there are $N/3$ receivers with identical distributions. The average SNRs of receivers for the three subsets are 7 dB, 12 dB, and 15 dB, respectively. We then apply our proposed subgrouping strategy to both the adaptive (with $\xi = 0.01$ and $\varepsilon = 0.01$) and fixed DP schemes (with $\xi = 0.01$). As shown in Fig. 6, the adaptive DP scheme also significantly outperforms the fixed DP scheme, which further validates the superiority of our proposed adaptive DP scheme.

Conclusions

We propose and analyze a cross-layer design scheme to comprehensively characterize QoS-driven mobile multimedia multicast services over fading channels in wireless networks. Specifically, we integrate rate adaptation in the physical layer with the loss rate QoS constraints from upper protocol layers. Moreover, we evaluate rate adaptation schemes by using the concept of effective capacity, which connects the throughput measure of multimedia multicast services with the delay-related QoS exponent requirement from data link layers. Based on the proposed cross-layer design model, we develop efficient rate adaptation schemes for both i.i.d. and non-i.i.d. fading environments. The obtained simulation results show that our proposed scheme significantly outperforms existing strategies.

References

[1] N. Jindal and Z. Q. Luo, "Capacity Limits of Multiple Antenna Multicast," *IEEE Int'l. Symp. Info. Theory*, Seatlle, WA, July 2006, pp. 1841–45.

[2] X, Zhang and Q. Du, "Adaptive Low-Complexity Erasure-Correcting Code Based Protocols for QoS-Driven Mobile Multicast Services Over Wireless Networks," *IEEE Trans. Vehic. Tech.*, vol. 55, no. 5, Sept. 2006, pp. 1633–47.

[3] P. K. Gopala and H. El Gamal, "On the Throughput-Delay Trade-off in Cellular Multicast," *Proc. Int'l. Conf. Wireless Networks, Commun. and Mobile Comp.*, vol. 2, HI, June 2005, pp. 1401–06.

[4] A. J. Goldsmith and S.-G. Chua, "Variable-Rate Variable-Power MQAM for Fading Channels," *IEEE Trans. Commun.*, vol. 45, Oct. 1997, pp. 1218–30.

[5] X. Zhang *et al.*, "Cross-Layer-Based Modeling for Quality of Service Guarantees in Mobile Wireless Networks," *IEEE Commun. Mag.*, Jan. 2006, pp. 100–06.

[6] 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, July 2003, pp. 630–43.

[7] J. Tang and X. Zhang, "Quality-of-Service Driven Power and Rate Adaptation Over Wireless Links," *IEEE Trans. Wireless Commun.*, July 2007.

[8] C.-S. Chang, "Stability, Queue Length, and Delay of Deterministic and Stochastic Queuing Networks," *IEEE Trans. Auto. Control*, vol. 39, no. 5, May 1994, pp. 913–31.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge Univ. Press, 2004.

[10] T. M. Cover, "Broadcast Channels," *IEEE Trans. Info. Theory*, vol. IT-18, Jan. 1972, pp. 2–14.

Biographies

Xi Zhang [S'89, SM'98] (xizhang@ece.tamu.edu) received his B.S. and M.S. degrees from Xidian University, Xi'an, China, his M.S. degree from Lehigh University, Bethlehem, Pennsylvania, all in electrical engineering and computer science, and his Ph.D. degree in electrical engineering and computer science (electrical engineering — systems) from the University of Michigan, Ann Arbor. He is currently an assistant professor and the founding director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University. 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, Queensland,

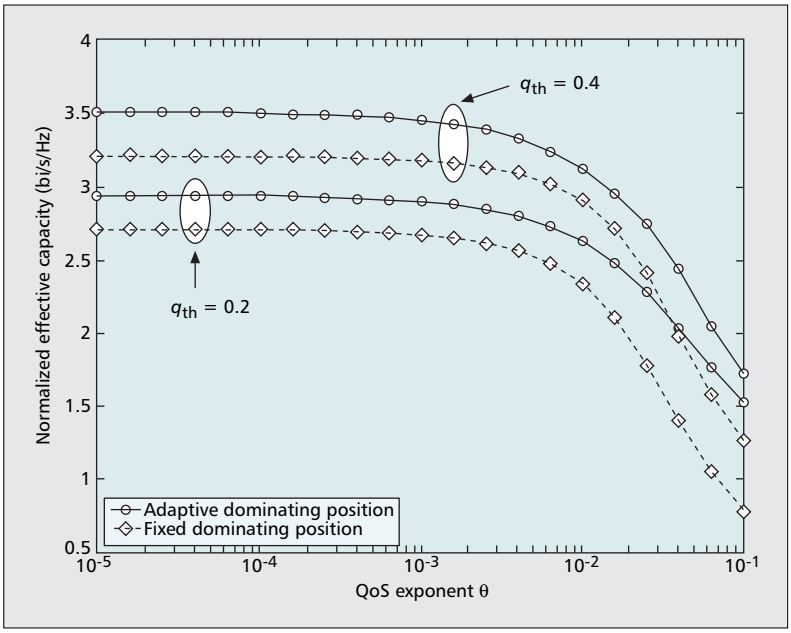

Figure 5. *Normalized effective capacity* $E_C(\theta)/(TB)$ *vs. QoS exponent* θ*, where* $N = 5$*,* $q_{th} = 0.2, 0.4$*, and* $\{\gamma_n\}_{n=1}^{N}$ *follow i.i.d. Nakagami-*m *fading with* $m = 2$ *and* $\bar{\gamma} = 10$ *dB for each multicast receiver.*

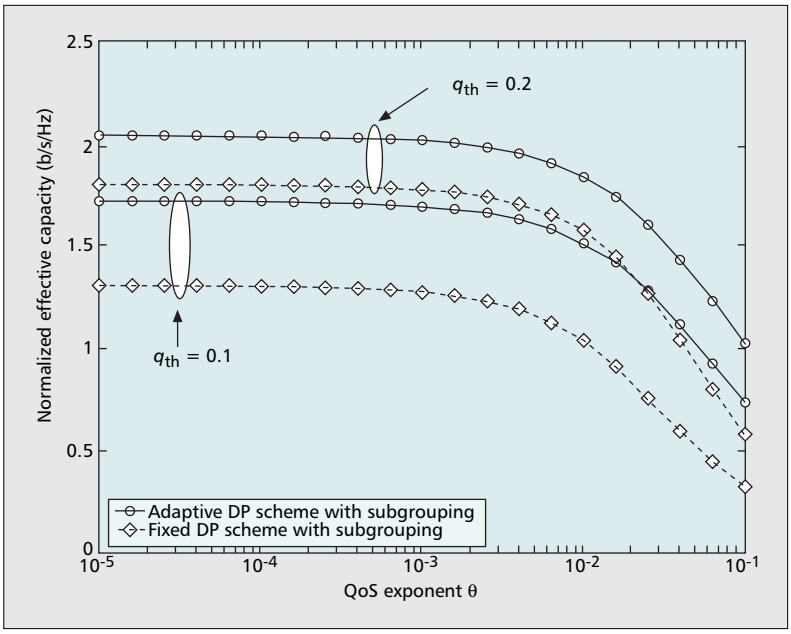

Figure 6. *Normalized effective capacity vs. QoS exponent* θ *in a non-i.i.d. fading environment, where* $N = 18$ *and* $\{\gamma_n\}_{n=1}^{N}$ *follow independent Nakagami-*m *distribution with* $m = 2$*. All receivers are divided into three subsets. Within each subset, there are* $N/3$ *receivers with identical distribution for each multicast receiver. The average SNRs of receivers in the three subsets are 7 dB, 12 dB, and 15 dB, respectively.*

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, New Jersey, and AT&T Laboratories Research, Florham Park, New Jersey in 1997. He has published more than 100 research papers in the areas of wireless networks and communications, mobile computing, cross-layer optimizations for QoS guarantees over mobile wireless networks, effective capacity and effective bandwidth theories for wireless networks, DS-CDMA, MIMO-OFDM and space-time coding, adaptive modulation and coding, wireless diversity techniques and resource allocations, wireless sensor and ad hoc networks, cognitive radio and cooperative communications/relay networks, vehicular ad hoc networks, multichannel MAC protocols, wireless and wired network security, wireless and wired multicast networks, channel coding for mobile wireless multimedia multicast, network protocols design and modeling, statistical communications theory, information theory, random signal processing, and control theory and systems. He 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 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 in 2006. He is currently serving as an Editor for *IEEE Transactions on Wireless Communications*, an Associate Editor for *IEEE Transactions on Vehicular Technology*, an Associate Editor for *IEEE Communications Letters*, and an Editor for Wiley's *Journal on Wireless Communications and Mobile Computing*, an Editor for *Journal of Computer Systems, Networking, and Communications*, and is also serving as the Guest Editor for the *IEEE Wireless Communications* Special Issue on Next Generation of CDMA vs. OFDMA for 4G Wireless Applications. He has frequently served as a panelist on U.S. National Science Foundation (NSF) Research-Proposal Review Panels. He also served as the WiFi-Hotspots/WLAN and QoS panelist at IEEE QShine '04. He is serving or has served as Co-Chair for the IEEE GLOBECOM 2008 Wireless Communications Symposium 2008 and IEEE ICC 2008 Information and Network Security Symposium 2008, respectively; as Chair of the IEEE International Cross-Layer Optimized Wireless Networks Symposium 2006 and 2007; TPC Chair for IEEE International Wireless Communications and Mobile Computing Conference '06 and '07, respectively; Poster Chair for IEEE INFOCOM 2008; Student Travel Grants Committee Co-Chair for IEEE INFOCOM 2007; Panel Co-Chair for the IEEE 16th International Conference on Computer Communications and Networks, Poster Chair for IEEE QShine 2006 and the IEEE/ACM 10th International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems '07, and Publicity Chair for IEEE WirelessCom 2005 and QShine 2007. He has served as a TPC member for more than 40 IEEE/ACM conferences, including IEEE INFOCOM, IEEE GLOBECOM, IEEE ICC, IEEE WCNC, IEEE VTC, IEEE/ACM QShine, IEEE WoWMoM, IEEE ICCCN, and others He is a member of the Association for Computing Machinery.

QINGHE DU (duqinghe@ece.tamu.edu) received B.S. and M.S. degrees in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 2001 and 2004, respectively. He is currently a research assistant working toward a Ph.D. degree in the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University. His research interests include mobile wireless communications and networks, with emphasis on forward error control coding, cross-layer design, wireless transmit diversity techniques, and wireless resource allocation for mobile multicast communications over wireless networks.