Cross-Layer Modeling for Quality of Service Guarantees over Wireless Links

Jia Tang, *Student Member, IEEE,* and Xi Zhang, *Senior Member, IEEE*

***Abstract*— We propose a cross-layer approach to investigate the impact of physical-layer infrastructure on data-link-layer quality-of-service (QoS) performance over wireless links in mobile networks. At the physical layer, we take multiple-input-multiple-output (MIMO) diversity schemes as well as adaptive-modulation-and-coding (AMC) techniques into account. At the data-link layer, our focus is on how this physical-layer infrastructure influences the real-time multimedia *delay-bound* QoS performance. To achieve this goal, we first model the physical-layer service process as a finite-state Markov chain (FSMC). Based on this FSMC model, we then characterize the QoS performance at data-link-layer using the *effective capacity* approach, which turns out to be critically important for the statistical QoS guarantees over wireless links in mobile networks. We also investigate the impact of physical-layer power control and channel-state information (CSI) feedback delay on the QoS performance. The numerical results obtained demonstrate that our proposed cross-layer model can efficiently characterize the interactions between the physical-layer infrastructure and data-link-layer QoS performance.**

***Index Terms*— Cross-layer design and optimization, mobile wireless networks, quality-of-service (QoS), effective capacity, adaptive-modulation-and-coding (AMC), multiple-input-multiple-output (MIMO), real-time multimedia delay-bound.**

I. Introduction

TO SUPPORT the diverse quality-of-service (QoS) requirements for heterogeneous mobile users, a large number of advanced schemes are developed at physical layer to overcome the impact of wireless fading channels. Among them, the multiple-input-multiple-output (MIMO) infrastructure [1], [2] and adaptive-modulation-and-coding (AMC) scheme [3], [4] are promising techniques that have received significant research attention. While the main focus is how to utilize those techniques to improve the spectral efficiency, the problem of how to efficiently employ the unique nature of such techniques for enhancing upper-layer protocol design, and to determine what is the impact of these physical-layer techniques on supporting the diverse QoS requirements, have been neither well understood, nor thoroughly studied. Consequently, it becomes increasingly important to develop the *cross-layer* system model to integrate the QoS provisioning algorithms/protocols at higher network-protocol layers with MIMO and AMC implemented at physical layer. In this paper, our focus is on designing the cross-layer model that can characterize the interactions across physical-layer and data-link-layer, and on mapping the physical-layer parameters to the data-link-layer's real-time multimedia *delay-bound* QoS requirements.

There have been a variety of research works focusing on wireless system modeling in both physical-layer and data-link-layer [5]–[8]. In [9], [10], Wu and Negi proposed a very interesting concept termed as "effective capacity". This concept turns out to be the *dual problem* of the so-called "effective bandwidth", which has been extensively studied in the early 90's with the emphasis on wired asynchronous transfer mode (ATM) networks [11]–[15]. The effective capacity and effective bandwidth enable us to analyze the statistical delay-bound violation probability, which is critically important for real-time multimedia wireless networks. In this paper we propose a cross-layer approach to investigate the impact of physical-layer infrastructure on data-link-layer QoS performance in mobile wireless networks. At the physical layer, we integrate the MIMO transmit/receive-diversity with AMC and develop a unified finite-state Markov chain (FSMC) model that characterizes the channel state variation. At the data-link layer, our focus is on how the physical-layer infrastructure influences the real-time multimedia delay-bound QoS provisioning performance. Based on our FSMC model developed at physical-layer, we characterize the QoS provisioning performance at data-link-layer by applying and extending the effective capacity method [9]. We show how the effective capacity can function as a bridge that connects the physical-layer across to the data-link-layer. The numerical and simulation results obtained demonstrate that our approach can efficiently capture the interactions across different network-protocol-layers and accurately characterizes the QoS provisioning performance. Based on our proposed cross-layer model, the advanced mechanisms such as adaptive resource-allocation, admission control, and packet scheduling schemes can be developed to guarantee the diverse QoS requirements for future wireless networks.

The rest of this paper is organized as follows. Section II describes the physical-layer system model. Section III investigates the effective capacity and its relationship with the cross-layer design. Section IV-A presents the numerical results on effective capacity and statistical QoS guarantees. Section V discusses the impact of power-control and feedback-delay on the effective capacity. The paper concludes with Section VI.

II. The Physical-Layer System Model

The system model is shown in Fig. 1. In this paper, we concentrate on a point-to-point wireless downlink with N_t

Manuscript received Feb. 2, 2006; revised Dec. 12, 2006; accepted December 18, 2006. The associate editor coordinating the review of this paper and approving it for publication was S. Hanly. The research reported in this paper was supported in part by the U. S. National Science Foundation CAREER Award under Grant ECS-0348694.

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

Digital Object Identifier 10.1109/TWC.2007.06087.

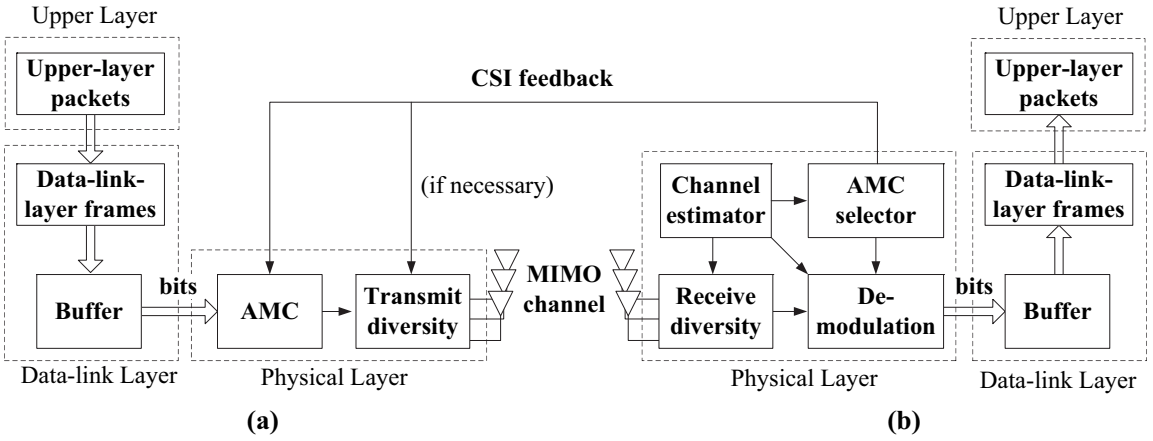


Fig. 1. The system model. (a) Basestation transmitter. (b) Mobile wireless receiver.

antennas (Tx) at the basestation transmitter and N_r antennas (Rx) at the mobile receiver. As shown by Fig. 1, the upper-protocol-layer packets are first divided into a number of frames at the data-link layer. The frames are stored at the transmitter-buffer and then split into bit-streams at the physical layer, where the AMC and MIMO-diversity are employed, respectively. The reverse operations are executed at the receiver side. Also, the channel-state information (CSI) is estimated at the receiver and fed back to the transmitter for AMC and MIMO-diversity (if necessary, depending on the specific MIMO-diversity scheme used). The upper-protocol-layer packets have the same packet-size, which consists of N_p bits. Also, the frame at data-link layer has the same time-duration, which is denoted by T_f. Due to the employment of AMC, the number of bits per frame varies depending on the modulation-and-code modes selected. Therefore, each frame comprises various portions of the packet. Furthermore, at the data-link layer, the system integrates forward-error control (FEC) with automatic retransmission request (ARQ) strategies, which will be detailed in Section II-B.

We assume that the wireless channel is flat-fading with Nakagami-m distribution, which is independent identically distributed (i.i.d.) between each transmit/receive antenna-pair. Also, the channel is invariant within a frame's time-duration T_f, but varies from one frame to another. We use the Nakagami-m channel model because this model is very general and often best fits the land-mobile and indoor-mobile multipath propagations [18], [19]. We assume that the CSI is perfectly estimated at the receiver and reliably fed back to the transmitter. However, the CSI feedback can be delayed, which is particularly addressed in Section V.

A. MIMO Diversity

We only focus on diversity-based MIMO systems. There exists a number of promising transmit/receive diversity schemes. For example, when the CSI is available at both sides of the wireless link, maximal-ratio transmission (MRT, also known as beamforming) and maximal-ratio combining (MRC) are known as the optimal transmit- and receive-diversity schemes [19], respectively. When the CSI is not available at the transmitter side, space-time block coding (STBC) is a powerful approach to achieve transmit diversity [2]. Moreover, the selection combining (SC) at either the transmitter or receiver side emerges as a good tradeoff between performance and complexity [18]. For a variety of MIMO-diversity schemes, we can show that the probability density function (pdf) of the combined signal-to-noise ratio (SNR), denoted by $p_\Gamma(\gamma)$, can be derived as a *unified* expression as follows:

$$p_\Gamma(\gamma) = \frac{M}{\Gamma(mL)} \sum_{i=0}^{M-1} (-1)^i \binom{M-1}{i} \exp\left(-(i+1)\frac{\beta m}{\overline{\gamma}}\gamma\right) \cdot \sum_{j=0}^{i(mL-1)} \xi_{ji} \left(\frac{\beta m}{\overline{\gamma}}\right)^{j+mL} \gamma^{j+mL-1} \quad (1)$$

where $\Gamma(\cdot)$ represents the Gamma function, $\overline{\gamma}$ denotes the average SNR of the combined signal, m denotes the fading parameter, ξ_{ji} is the multinomial expansion coefficients determined by $\xi_{ji} = \sum_{p=a}^{b} \xi_{p\,(i-1)}/[(j-p)!]$ with $a = \max\{0, j-(M-1)\}$, $b = \min\{j, (i-1)(M-1)\}$, $\xi_{j0} = \xi_{0i} = 1$, $\xi_{j1} = 1/(j!)$, and $\xi_{1i} = i$, and finally, the parameters M, L, and β are MIMO-diversity-scheme dependent, which are specified in TABLE I. Note that in Eq. (1), M and L denote the *selection diversity* order and *combining diversity* order, respectively, and β only affects STBC scheme that reduces the variation of the channel. The total diversity order is determined by $M \times L$ (also equal to $N_t \times N_r$).

TABLE I
PARAMETER IDENTIFICATIONS FOR UNIFIED MIMO DIVERSITY.

MIMO Diversity Schemes	M	L	β
Tx-MRT/Rx-1	1	N_t	1
Tx-STBC/Rx-MRC	1	N_tN_r	N_t
Tx-SC/Rx-MRC	N_t	N_r	1
Tx-MRT/Rx-SC	N_r	N_t	1
Tx-STBC/Rx-SC	N_r	N_t	N_t
Tx-SC/Rx-SC	N_tN_r	1	1
Performance Upper-Bound	1	N_tN_r	1

B. Adaptive Modulation and Coding

The AMC technique has emerged as one of the key solutions to increase the spectral-efficiency of wireless networks.

In [4], the authors studied adaptive modulation integrated with convolutional-code-based FEC strategy. Following the work of [4], the specific modulation-and-coding modes for the AMC scheme used in this paper are constructed as follows.

The entire SNR range is partitioned by, e.g., $K = 7$ non-overlapping consecutive intervals, resulting in $K+1$ boundary points denoted by $\{\Gamma_k\}_{k=1}^{K+1}$, where $\Gamma_1 < \Gamma_2 < \cdots < \Gamma_{K+1}$ with $\Gamma_1 = 0$ and $\Gamma_{K+1} = \infty$. Correspondingly, the AMC is selected to be in mode k if the SNR, denoted by γ, falls into the range: $\Gamma_k \le \gamma < \Gamma_{k+1}$, where $k = 1, 2, ..., K$. More specifically, the code rates of the available modes are 0, 0.5, 0.5, 0.75, 0.5625, 0.75, and 0.75, respectively, and their corresponding constellations are "outage", BPSK, QPSK, QPSK, 16-QAM, 16-QAM, and 64-QAM, respectively. As the SNR increases, the system selects the AMC mode with higher spectral-efficiency to transmit data. On the other hand, as the SNR gets worse, the system decreases the transmission rate to adapt to the degraded channel conditions. In the worst case, the transmitter can stop transmitting data, which corresponds to the "outage" mode of the system. The packet-error rate (PER) when using the kth AMC mode ($k = 2, 3, ..., K$), denoted by $\mathrm{PER}_k(\gamma)$, can be approximated as follows [4, eq. (3)]:

$$\mathrm{PER}_k(\gamma) = \begin{cases} 1, & \text{if } 0 < \gamma < \gamma_k \\ a_k \exp(-g_k \gamma), & \text{if } \gamma \ge \gamma_k \end{cases} \tag{2}$$

where a_k, g_k, and γ_k are mode-dependent parameters [4, TABLE II]. Accordingly, the AMC is in mode k if the SNR γ falls into the range of $\Gamma_k \le \gamma < \Gamma_{k+1}$. Based on the pdf of the SNR in Eq. (1), the probability π_k, that the SNR falls into mode k is determined by

$$\begin{aligned} \pi_k &= \int_{\Gamma_k}^{\Gamma_{k+1}} p_\Gamma(\gamma) d\gamma \\ &= \left[\frac{\gamma\left(mL, \frac{\beta m}{\overline{\gamma}} \Gamma_{k+1}\right)}{\Gamma(mL)} \right]^M - \left[\frac{\gamma\left(mL, \frac{\beta m}{\overline{\gamma}} \Gamma_k\right)}{\Gamma(mL)} \right]^M \end{aligned} \tag{3}$$

where $k = 1, 2, ..., K$ and $\gamma(\cdot, \cdot)$ denotes the incomplete Gamma function. We select the boundaries such that $\Gamma_k \ge \gamma_k$ for all $k = 2, 3, ..., K$. Then, using Eq. (2), we obtain the average PER of mode k, denoted by $\overline{\mathrm{PER}}_k$, as follows:

$$\begin{aligned} \overline{\mathrm{PER}}_k &= \frac{1}{\pi_k} \int_{\Gamma_k}^{\Gamma_{k+1}} a_k \exp(-g_k \gamma) p_\Gamma(\gamma) d\gamma \\ &= \frac{a_k M}{\pi_k \Gamma(mL)} \sum_{i=0}^{M-1} (-1)^i \binom{M-1}{i} \sum_{j=0}^{i(mL-1)} \xi_{ji} \\ &\quad \cdot \left(\frac{\beta m}{b_k}\right)^{j+mL} \left[\gamma\left(j + mL, \frac{b_k \Gamma_{k+1}}{\overline{\gamma}}\right) \right. \\ &\quad \left. - \gamma\left(j + mL, \frac{b_k \Gamma_k}{\overline{\gamma}}\right) \right] \end{aligned} \tag{4}$$

where $b_k = g_k \overline{\gamma} + (i+1)\beta m$. Thus, the average PER can be expressed as follows:

$$\mathrm{PER} = \frac{\sum_{k=2}^{K} R_k \pi_k \overline{\mathrm{PER}}_k}{\sum_{k=2}^{K} R_k \pi_k} \tag{5}$$

where R_k denotes the spectral-efficiency of the kth mode. We can numerically obtain the boundaries $\{\Gamma_k\}_{k=2}^{K}$ such that the average PER satisfies the reliability QoS requirement. When taking the ARQ into account, the achieved spectral-efficiency of the kth mode, denoted by $\widetilde{R}_k$, can be expressed as

$$\widetilde{R}_k = R_k \left(1 - \overline{\mathrm{PER}}_k\right). \tag{6}$$

C. Service Process Modeling Using FSMC

In this paper, we employ the FSMC model to characterize the variations of the MIMO-diversity and AMC-based wireless-channel service process. The state of FSMC corresponds to the mode of AMC, where the effective transmission rate of the kth mode is $\widetilde{R}_k$. Let $p_{i,j}$ denote the transition probability of the FSMC from state i to state j. We assume a slow-fading channel model such that transition only happens between adjacent states [5][6]. Under such an assumption, we have $p_{ij} = 0$ for all $|i - j| > 1$. The adjacent transition probability can be approximated as [5]

$$\begin{cases} p_{k,k+1} \approx \frac{N_\Gamma(\Gamma_{k+1}) T_f}{\pi_k}, & \text{where } k = 1, 2, ..., K-1, \\ p_{k,k-1} \approx \frac{N_\Gamma(\Gamma_k) T_f}{\pi_k}, & \text{where } k = 2, 3, ..., K \end{cases} \tag{7}$$

where $N_\Gamma(\gamma)$ is the level-crossing rate (LCR) calculated at SNR value of γ [6]. Then, the remaining transition probability can be derived as

$$\begin{cases} p_{1,1} = 1 - p_{1,2} \\ p_{K,K} = 1 - p_{K,K-1} \\ p_{k,k} = 1 - p_{k,k-1} - p_{k,k+1}, \ k = 2, ..., K-1. \end{cases} \tag{8}$$

Thus, applying Eqs. (7) and (8), we obtain the transition probability matrix of the FSMC, which is denoted by $\mathbf{P} = [p_{ij}]_{K \times K}$. Correspondingly, the stationary distribution of the FSMC, denoted by $\boldsymbol{\pi}$, is determined by $\boldsymbol{\pi} = [\pi_1, \pi_2, ..., \pi_K]$, where π_k is given by Eq. (3) for $k = 1, 2, ..., K$.

In order to obtain the transition probability matrix $\mathbf{P}$, it is necessary to find the LCR $N_\Gamma(\gamma)$ in Eq. (7). We obtain the *unified* closed-form expression for the LCR, $N_\Gamma(\gamma)$, which is given as follows:

$$\begin{aligned} N_\Gamma(\gamma) &= \frac{\sqrt{2\pi} f_d M}{\Gamma(mL)} \sum_{i=0}^{M-1} (-1)^i \binom{M-1}{i} \\ &\quad \cdot \exp\left(-(i+1)\frac{\beta m \gamma}{\overline{\gamma}}\right) \sum_{j=0}^{i(mL-1)} \xi_{ji} \left(\frac{\beta m \gamma}{\overline{\gamma}}\right)^{j+mL-\frac{1}{2}} \end{aligned} \tag{9}$$

where f_d denotes the maximum Doppler frequency of the channel. Substituting Eq. (9) into Eq. (7), the transition matrix $\mathbf{P}$ is determined for different MIMO diversity schemes.

III. The Effective Capacity and Cross-Layer Designs

A. Statistical QoS Guarantees

The real-time multimedia services such as video and audio require bounded delay, or equivalently, guaranteed service bandwidth. Once a received real-time packet violates its delay-bound, it is considered as useless and will be discarded. However, over the mobile wireless networks, a hard delay-bound guarantee is not practically achieveable due to the impact of the time-varying fading channels. For example, over the Rayleigh or Rician fading channels, the only lower-bound of

the system bandwidth that can be *deterministically* guaranteed is a bandwidth of zero. Therefore, we consider an alternative solution by providing the *statistical* QoS guarantees, where we guarantee the delay-bound with a small violation probability.

During the early 90's, the statistical QoS guarantees theories have been extensively studied in the contexts of so-called *effective bandwidth theory*, with the emphasis on wired ATM networks [11]–[15]. The asymptotic results in [11] showed that, for stationary arrival and service processes under sufficient conditions, the probability that the queue size Q exceeds a certain threshold B decays exponentially fast as the threshold B increases, i.e.,

$$\Pr\{Q > B\} \approx e^{-\theta B}, \text{ for a large } B \tag{10}$$

where θ is a certain positive constant called "QoS exponent" [9]. For a small B, the following approximation is shown to be more accurate [9, eq. (9)]:

$$\Pr\{Q > B\} \approx \varepsilon e^{-\theta B} \tag{11}$$

where ε denotes the probability that the buffer is not empty, which can be approximated by the ratio of the average arrival-rate to the average service-rate [11, eq. (9.184)]. Furthermore, when delay-bound is the main QoS metric of interest (i.e., when the focus is on delay-bound violation probability), an expression similar to Eq. (11) can be obtained as

$$\Pr\{\text{Delay} > \tau_{\max}\} \approx \varepsilon e^{-\theta\delta\tau_{\max}} \tag{12}$$

where $\tau_{\max}$ denotes the delay-bound, and δ is jointly determined by both arrival and service processes, which will be detailed below.

From Eqs. (10)–(12) we can see that the parameter θ plays an important role for the statistical QoS guarantees, which indicates the decaying-rate of the QoS violation probability. A smaller θ corresponds to a slower delaying-rate, which implies that the system can only provide a *looser* QoS requirement, while a larger θ leads to a faster delaying-rate, which means a more *stringent* QoS requirement can be guaranteed. Consequently, θ is called *QoS exponent* [9].

B. Effective Capacity and Cross-Layer Designs

Inspired by the effective bandwidth theory, Wu and Negi in [9] developed the concept of *effective capacity*, which is a *dual problem* of the original effective bandwidth. The effective capacity function, denoted by $\mathrm{E_C}(\theta)$, characterizes the attainable service-rate as a function of the QoS exponent θ. Specifically, in [9] the effective capacity $\mathrm{E_C}(\theta)$ is defined as the *constant* arrival-rate that the channel can support in order to guarantee a QoS requirement specified by θ.

Although the original concept of effective capacity is proposed based on *constant*-arrival assumption, it actually can be generalized to investigate the QoS performance of any *stationary* arrival process. Under such a condition, the arrival process should be represented by its effective bandwidth while the service process should be characterized by its effective capacity, respectively. Note that for a constant arrival-process, the corresponding effective-bandwidth is equal to its constant arrival-rate. Thus, the problem discussed in [9] can be considered as the special case of our more general scenario addressed in this paper, where both arrival and service processes are time-varying. To help demonstrate the principles and identify the relationships between effective bandwidth and effective capacity, let us consider the case as illustrated in Fig. 2.

For any given arrival process and service process, we depict their effective-bandwidth function, denoted by $\mathrm{E_B}(\theta)$, and effective-capacity function, denoted by $\mathrm{E_C}(\theta)$, in Fig. 2, respectively. Let us define two limiting values as follows:

$$\begin{cases} \mu_A \triangleq \lim_{\theta\to 0} \mathrm{E_B}(\theta) \\ \mu_C \triangleq \lim_{\theta\to 0} \mathrm{E_C}(\theta). \end{cases} \tag{13}$$

The effective bandwidth theory demonstrates that μ_A is equal to the average arrival-rate of the traffic process [12], [15]. Also, we will show in the next section that μ_C is equal to the average service-rate of the service process. Therefore, using the approximation in [11, eq. (9.184)], the buffer non-empty probability ε in Eqs. (11) and (12) can be expressed as

$$\varepsilon \approx \frac{\mu_A}{\mu_C}. \tag{14}$$

The effective-bandwidth function $\mathrm{E_B}(\theta)$ intersects with the effective-capacity function $\mathrm{E_C}(\theta)$ at the point where the QoS-exponent is θ^* and the rate is δ.

In general, the delay-bound violation probability can be calculated in the following algorithm:

Algorithm 1: Calculation of the delay-bound violation probability:

Step 1: According to the statistical characteristics of the arrival and service processes, find the effective-bandwidth function $\mathrm{E_B}(\theta)$ and effective-capacity function $\mathrm{E_C}(\theta)$. Determine the solution of the rate and QoS-exponent pair (δ, θ^*) such that $\mathrm{E_B}(\theta^*) = \mathrm{E_C}(\theta^*) = \delta$.

Step 2: Approximate the buffer non-empty probability ε by using Eq. (14).

Step 3: For any pre-determined delay-bound $\tau_{\max}$ and (δ, θ^*) obtained in **Step 2**, the delay-bound violation probability can be derived using Eqs. (12) and (14) as follows:

$$\Pr\{\text{Delay} > \tau_{\max}\} \approx \varepsilon e^{-\theta^*\delta\tau_{\max}}. \tag{15}$$

From Fig. 2 we can gain insights about how the statistical QoS performance changes according to the service and arrival processes. As shown by Fig. 2, increasing the service-process bandwidth (as shown by the arrow at the lower position) results in higher effective capacity, which will lead to a larger QoS-exponent solution θ^*. This implies that the higher bandwidth service-process can support a more *stringent* QoS for a given arrival process. On the other hand, increasing the arrival-process bandwidth (as shown by the arrow at the upper position) makes the effective bandwidth increase, which generates a smaller QoS-exponent solution θ^* for a given service process. This implies that only a *looser* QoS can be guaranteed. When the bandwidth of the arrival process further increases such that $\mu_A > \mu_C$, there is no solution for $\theta^* > 0$ existing. Thus, the service process cannot support any QoS for the given arrival process, which is consistent with the queueing

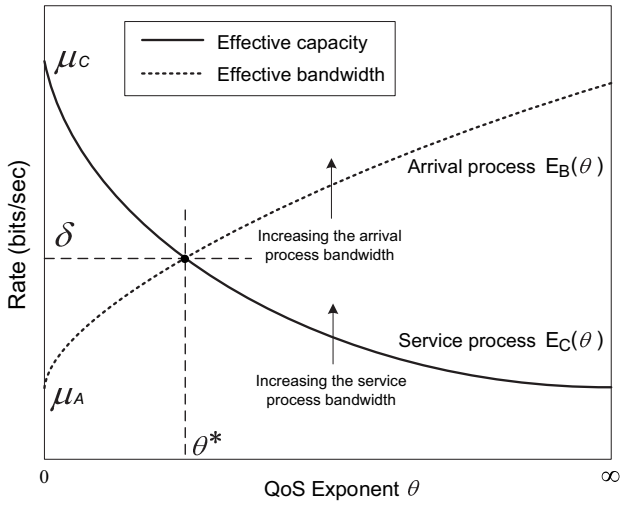


Fig. 2. The relationships between effective bandwidth and effective capacity as a function of the QoS exponent θ.

theory that if $\mu_A > \mu_C$, both queue length and the queueing delay will approach to infinity.

Inspired by the above analyses and observations, we propose to use the effective capacity as a bridge for the cross-layer modeling. The characterizations of the QoS performance guarantees are equivalent to investigating the dynamics of the effective capacity function, which turns out to be a simple and efficient cross-layer modeling approach. In [9], the authors employed an experimental-based method to measure the effective capacity. In fact, it is feasible to formulate the effective capacity problem in a more systematic manner. In the next section, we analytically investigate the effective capacity function for our FSMC-based wireless-channel service process.

IV. Effective Capacity for FSMC-Based Wireless-Channel Service Process

Analytically, the effective capacity can be formally defined as follows. Let the sequence $\{R[i],\ i = 1, 2, ...\}$ denote a discrete-time stationary and ergodic stochastic service process and $S[t] \triangleq \sum_{i=1}^{t} R[i]$ be the partial sum of the service process. Assume that the Gärtner-Ellis limit of $S[t]$, expressed as $\Lambda_{\rm C}(\theta) = \lim_{t\to\infty}(1/t)\log\left(\mathbb{E}\left\{e^{\theta S[t]}\right\}\right)$ exists and is a convex function differentiable for all real θ [12, pp. 921]. Then, the effective capacity of the service process, denoted by $\mathrm{E_C}(\theta)$, where $\theta > 0$, is defined as [9, eq. (12)]

$$\mathrm{E_C}(\theta) \triangleq -\frac{\Lambda_{\rm C}(-\theta)}{\theta} = -\lim_{t\to\infty}\frac{1}{\theta t}\log\Big(\mathbb{E}\left[e^{-\theta S[t]}\right]\Big). \quad (16)$$

Based on our physical-layer FSMC model developed in Section II-C, we have the following proposition:

Proposition 1: Let $\{\mu_k, k = 1, 2, ..., K\}$ be the number of bits per frame transmitted at the state k of the FSMC-based service process and define $\mathbf{\Phi}(\theta) \triangleq \mathrm{diag}\left\{e^{-\mu_1\theta}, e^{-\mu_2\theta}, ..., e^{-\mu_K\theta}\right\}$. Then, the effective capacity of the FSMC-based service process is determined by

$$\mathrm{E_C}(\theta) = -\frac{1}{\theta}\log\Big(\rho\{\mathbf{P}\,\mathbf{\Phi}(\theta)\}\Big) \quad (17)$$

where $\mathbf{P}$ is the transition probability matrix determined by Eqs. (7) and (8), and $\rho\{\cdot\}$ denotes the spectral radius of the matrix.

Proof: This result follows along the lines of [12, Example 3.3]; see [16, Appendix II] for the details. ■

Based on our system model in Section II, the transmission rate μ_k can be expressed as

$$\mu_k = \widetilde{R}_k T_f W, \quad \text{for all } k = 1, 2, ..., K \quad (18)$$

where W denotes the system spectral-bandwidth, and $\widetilde{R}_k$ is derived in Eq. (6) which takes the ARQ into consideration.

We can characterize the monotonic and asymptotic properties of the effective-capacity function by Proposition 2 that follows below.

Proposition 2: Let $\overline{\mu}$ and $\mu_{\min}$ be the average and minimum number of bits per frame transmitted by the FSMC-based service process, respectively. Then, the following claims hold for the effective-capacity function $\mathrm{E_C}(\theta)$ of the FSMC-based service process:

$$\underline{\text{Claim 1.}} \quad \frac{d\mathrm{E_C}(\theta)}{d\theta} \le 0, \quad \text{for all } \theta > 0. \quad (19)$$

$$\underline{\text{Claim 2.}} \quad \sup_{\theta>0}\mathrm{E_C}(\theta) = \lim_{\theta\to 0}\mathrm{E_C}(\theta) = \overline{\mu}. \quad (20)$$

$$\underline{\text{Claim 3.}} \quad \inf_{\theta>0}\mathrm{E_C}(\theta) = \lim_{\theta\to\infty}\mathrm{E_C}(\theta) = \mu_{\min}. \quad (21)$$

Proof: This result follows along the lines of [11], [15]; see [16, Appendix III] for the details. ■

A. Numerical Evaluations

We evaluate the effective capacity by numerical solutions under different physical-layer diversity schemes and parameters, where we set the total system spectral-bandwidth $W = 100$ KHz, the upper-layer packet-size $N_p = 1080$ bits, and the data-link-layer frame time-duration $T_f = 2$ ms. Unless explicitly stated on the legend of the figures, the other system parameters are set as follows: the fading parameter $m = 1$, indicating the Rayleigh fading channel, the average SNR $\overline{\gamma} = 10$ dB, the Doppler frequency $f_d = 5$ Hz, and the average packet-error rate PER = 10^{-3}. To ease comparison with the spectral-efficiency, in the following discussions we plot the normalized effective capacity (*which is defined as the effective capacity divided by the spectral-bandwidth W and the frame duration T_f, and thus has the unit of "bits/sec/Hz"*).

Fig. 3 plots the effective capacity against the QoS exponent θ under different spatial diversity schemes. As shown in Fig. 3, the physical-layer antenna infrastructures have significant impact on the effective capacity. The effective capacities of MIMO (i.e., Tx-2/Rx-2) or multiple-input-single-output (MISO, i.e., Tx-2/Rx-1) systems are significantly larger than those of the SISO systems. Also, different diversity schemes can achieve different effective capacities, depending on how much of the CSI information is utilized.

An interesting observation is that when the QoS exponent θ is small, the effective capacity of some MIMO systems (e.g., STBC/SC) is lower than that of the MISO systems (e.g., MRT/MRC), which is because STBC/SC does not efficiently

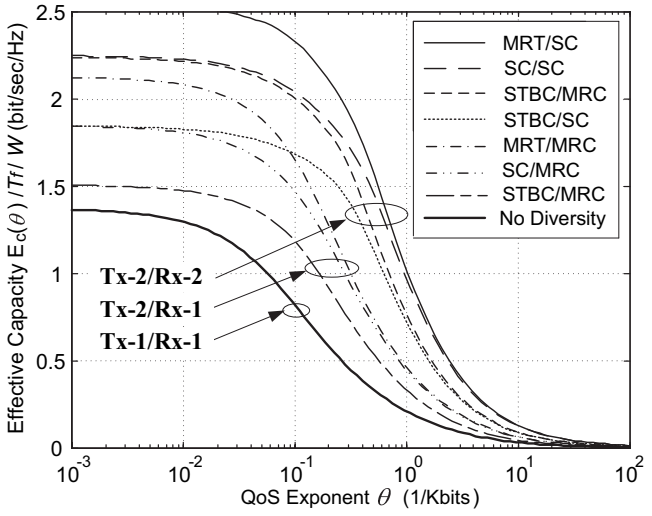


Fig. 3. The normalized effective capacity as a function of the QoS exponent θ under different spatial diversity schemes.

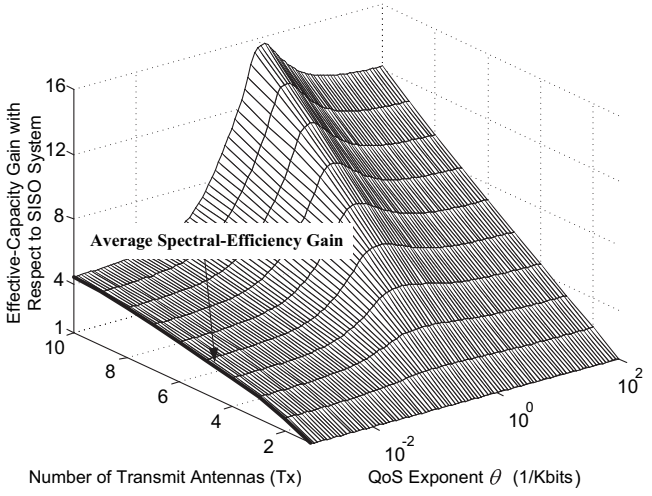


Fig. 4. The effective-capacity gain with respect to the SISO system. MRT/MRC (Rx-1) is employed to fully utilize the CSI. The average SNR $\overline{\gamma} = 5$ dB.

utilize the CSI while MRT/MRC fully utilizes the CSI. However, this situation changes as the QoS exponent θ increases. As shown in Fig. 3, for a large θ, the effective capacities of all MIMO systems are larger than those of the MISO systems. This implies that even under the condition that the MIMO system has lower spectral-efficiency than the MISO system, it offers more significant advantages in supporting the stringent QoS requirement.

To further investigate the impact of antenna diversity on the effective capacity, Fig. 4 plots the effective-capacity gain (defined as the ratio of the effective-capacity with antenna-diversity-based systems to that with SISO-based systems) against the number of transmit antenna N_t and the QoS exponent θ, where we employ MRT/MRC (Rx-1) to fully utilize the CSI. Notice from Proposition 2 that as the QoS-exponent θ approaches to 0, the effective-capacity converges to the *average service-rate*. Furthermore, the *average spectral-efficiency* is the average service-rate normalized by the spectral-bandwidth W. Therefore, the boldface line highlighted in Fig. 4 is actually the average spectral-efficiency gain achieved by antenna diversity. We observe from Fig. 4 that the effective-capacity gain with large θ is significantly higher than the average spectral-efficiency gain (indicated by the boldface line in Fig. 4 as $\theta \to 0$). Thus, Fig. 4 implies that *the superiority/gain of employing antenna diversity in terms of QoS-guarantees is even more significant than that in terms of the spectral-efficiency*.

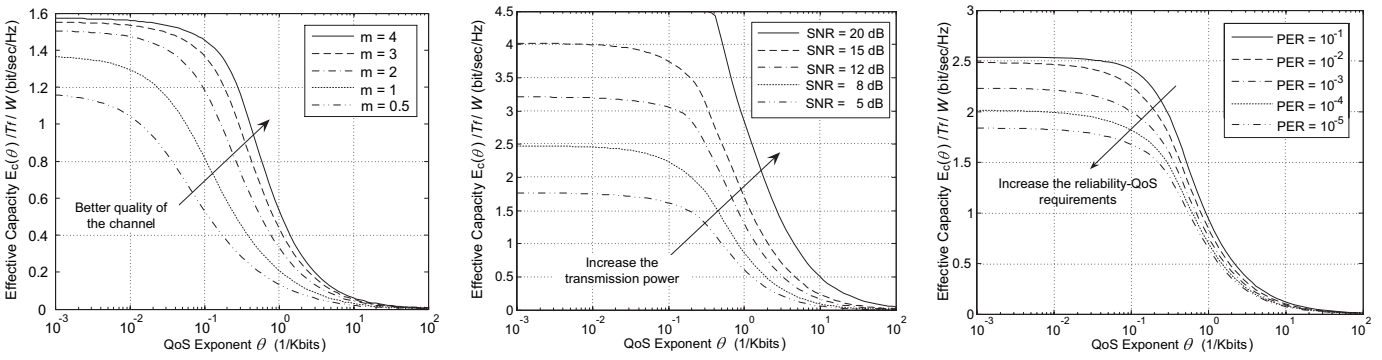
Fig. 5(a) plots the effective capacity of SISO system against the QoS exponent θ with different channel distributions. From Fig. 5(a), we can observe that as the fading parameter m increases (the channel quality gets better), the effective capacity increases correspondingly, which is expected since the stabler channel can support a more stringent QoS. Fig. 5(b) plots the effective capacity against the QoS exponent θ when the SNR varies. As shown in Fig. 5(b), increasing the SNR of the wireless channel, or equivalently, increasing the transmission power, can improve the effective capacity. When the SNR $\overline{\gamma} = 20$ dB, we can see from Fig. 5(b) that the effective capacity gets saturated at spectral-efficiency of 4.5 bits/sec/Hz, which is because 4.5 bits/sec/Hz is the highest spectral-efficiency that can be obtained by the underlying AMC scheme. Fig. 5(c) depicts the effective capacity versus the QoS exponent θ with different reliability-QoS requirements of PER's, where the more stringent reliability-QoS results in the lower effective capacity. In summary, from Fig. 5 we can observe that the physical-layer variations have significant impact on the effective capacity, and thus on the QoS provisioning performance of wireless networks at higher-protocol-layers.

It is well known that there exists a diversity-multiplexing tradeoff over fading channels [17], where high diversity is achieved at the expense of losing throughput/rate, and vice versa. Being consistent with the above fact, from Figs. 3 and 5 we also observe that there is a tradeoff between throughput (effective capacity) and QoS (θ), where the channel can support lower rate as QoS becomes more stringent. This observation can be considered as a cross-layer assessment for the impact of the diversity-multiplexing tradeoff on QoS at the network layer in terms of queuing delay. From the information theoretic point-of-view, for the fast fading channel, we can adapt the rate and apply ergodic capacity, which essentially assumes an infinite buffer model. On the other hand, for slow fading channel, we cannot adapt the rate and need to apply the outage capacity, which assumes a zero buffer model. In fact, increasing the diversity makes the output process/rate more deterministic, which in turn requires less buffering, implying the more stringent QoS. This also agrees, from a different angle, with our observation that the channel can only support lower throughput when the delay constraint QoS is more stringent, which is shown in Figs. 3 and 5.

V. Impact of Power Control and Feedback Delay on the Effective Capacity

A. The Impact of Power Control on the Effective Capacity

In previous sections, we employ the AMC scheme which uses *constant* power. However, it is well known that the optimal power-control, i.e., water-filling-based scheme, can



(a) SISO (b) MRT/MRC. Tx-4/Rx-1. (c) STBC/MRC. Tx-2/Rx-2.

Fig. 5. The normalized effective capacity $\mathrm{E_c}(\theta)$ as a function of QoS exponent θ with different physical-layer diversity schemes and parameters.

achieve higher spectral-efficiency than the constant-power schemes. A natural question to ask is whether the water-filling-based power-control is still optimal when we include QoS-guarantees? We will show via our effective capacity analysis that this is not true.

In [3, eq. (5)], the authors provided the time-domain water-filling power-control strategy for un-coded adaptive QAM modulation. We integrate the approach in [3] with our adaptive-modulation-based FSMC and analytically derive the corresponding effective capacity, which is numerically plotted against θ as shown in Fig. 6. We employ un-coded adaptive modulation instead of coded scheme used in previous sections because the un-coded scheme is analytically convenient for water-filling-based power-control. We can see from Fig. 6 that the effective-capacity of the scheme with optimal power-control is larger than that of the constant-power scheme when the QoS-exponent θ is small, which is due to the fact that the water-filling scheme always has better average spectral-efficiency than the constant-power scheme. However, as θ increases, the effective-capacity of the scheme with optimal power-control is lower than the constant-power scheme. This implies that *the optimal power-control that maximizes the spectral-efficiency is not necessarily optimal for QoS guarantees*. The reason behind this counter-intuitive observation is because the water-filling scheme increases the variation (instability) of the service-rate, which is undesired in terms of QoS-guarantees.

B. The Impact of Feedback Delay on the Effective Capacity

In previous sections, we assume that the CSI is *reliably* fed back to the transmitter without error and delay. However, in practical wireless networks, this assumption does not hold. In particular, the CSI *feedback delay* is un-avoidable in most of the situations.

Denote the feedback delay by τ. After the delay τ, the original SNR γ changes to a new value denoted by $\widehat{\gamma}$. Given γ, the pdf of $\widehat{\gamma}$ can be expressed as

$$p_{\widehat{\Gamma}|\Gamma}(\widehat{\gamma}|\gamma) = \frac{1}{(1-\rho)}\left(\frac{m}{\overline{\gamma}}\right)\left(\frac{\widehat{\gamma}}{\rho\gamma}\right)^{\frac{m-1}{2}} \cdot \exp\left(-\frac{m(\rho\gamma+\widehat{\gamma})}{(1-\rho)\overline{\gamma}}\right) I_{m-1}\left(\frac{2m\sqrt{\rho\gamma\widehat{\gamma}}}{(1-\rho)\overline{\gamma}}\right) \quad (22)$$

where $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind with order ν and ρ represents the correlation coefficient between $\widehat{\gamma}$ and γ, which is given by $\rho = J_0^2(2\pi f_d \tau)$ [19] with $J_0(\cdot)$ denoting the zero-th-order Bessel function of the first kind.

In order to guarantee the reliability QoS requirement, the system needs to maintain the same PER or bit-error-rate (BER) as the case without feedback delay. As a result, the boundary points for the AMC should be re-calculated. In [3], the authors analyzed the impact of CSI feedback delay on BER performance for the adaptive modulation. In [18], we also investigated the feedback delay issue for SC/MRC scheme from BER perspective. Using the similar approach to [3] and [18], we derive analytical expressions for the effective capacity when considering CSI feedback delays, where the normalized effective capacity of the AMC-based SISO system is numerically plotted as a function of θ and $f_d\tau$ in Fig. 7. We can observe from Fig. 7 that as long as the normalized feedback delay, measured by $f_d\tau$, is within certain threshold (e.g., $f_d\tau \leq 10^{-2}$), the effective capacity is virtually unchanged with $f_d\tau$. When the normalized feedback delay further increases, the effective capacity decreases accordingly. Note that in our system model, we have $T_f \times f_d = 10^{-2}$. Thus, over the Rayleigh fading channel with Doppler frequency of $f_d = 5$ Hz, our system can tolerant CSI feedback-delay with approximately one frame's time-duration while still maintaining virtually the same statistical QoS performance.

VI. Conclusions

We proposed the cross-layer design approach to study the interactions between physical-layer AMC and MIMO-diversity and higher-protocol-layer on the statistical QoS performance of the mobile wireless networks. We identified the critical relationships between effective bandwidth and effective capacity. Our numerical results showed that the AMC and MIMO-diversity employed at physical-layer have significant impact on the statistical QoS performance at upper-protocol-layers. The proposed cross-layer modeling accurately characterize the influence of physical-layer infrastructure on statistical QoS performance at higher-protocol layers.

While in this paper, we only investigate the single user QoS provisioning, our developed cross-layer modeling technique can be readily extended to the scenarios with multiple users

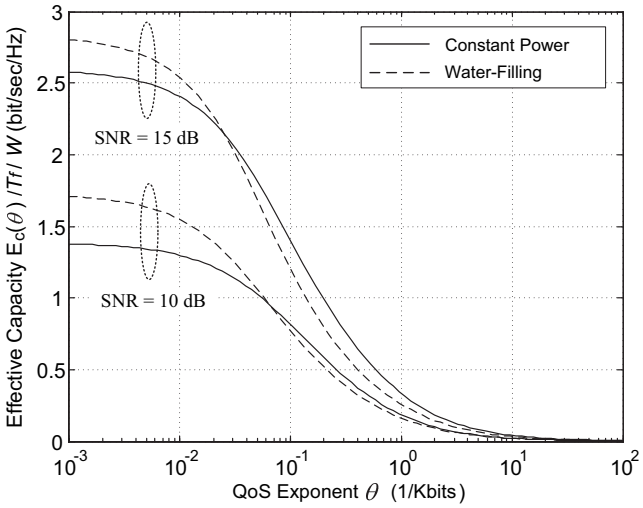


Fig. 6. The normalized effective capacity of the SISO system as a function of the QoS exponent θ with un-coded adaptive QAM modulation. The BER requirement is set to BER=10^{-3}.

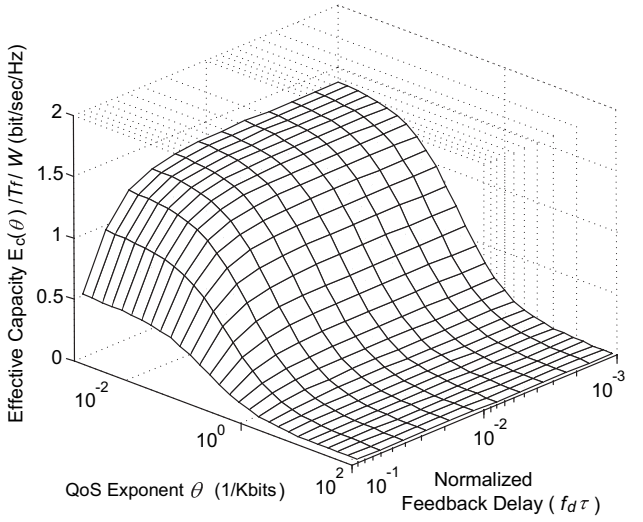


Fig. 7. The normalized effective capacity of the AMC-based SISO system when considering the CSI feedback delay.

sharing the wireless media in, e.g., dynamic TDMA-based wireless networks. More importantly, our developed cross-layer modeling technique also offers a practical and effective approach to develop highly-efficient admission-control, packet scheduling, and adaptive resource-allocation schemes to guarantee the QoS for real-time multimedia traffics over mobile wireless networks.

REFERENCES

[1] I. E. Telatar, "Capacity of multi-antenna Gaussian channels," AT&T Bell Labs Tech. Rep. BL0112170-950615-07TM, 1995.

[2] V. Tarokh, H. Jafarkhani, and A. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Inform. Theory*, vol. 45, no. 5, pp. 1456–1467, July 1999.

[3] A. J. Goldsmith and S. Chua, "Vairable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[4] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer combining of adaptive modulation and coding with truncated ARQ over wireless links," *IEEE Trans. Wireless Commun.*, vol. 3, no. 5, pp. 1746-1755, Sept. 2004.

[5] H. S. Wang and N. Moayeri, "Finite-state Markov channel—a useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[6] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[7] J. G. Kim and M. Krunz, "Bandwidth allocation in wireless networks with gauranteed packet-loss performance," *IEEE/ACM Trans. Networking*, vol. 8, pp. 337–349, June 2000.

[8] Q. Liu, S. Zhou, and G. B. Giannakis, "Queuing with adaptive modulation and coding over wireless link: cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, May 2005.

[9] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, Jul. 2003.

[10] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality-of-service assurance," *IEEE Trans. Veh. Technol.*, vol. 53, no. 5, pp. 1547–1557, Sep. 2004.

[11] C.-S. Chang, *Performance Guarantees in Communication Networks*. Berlin, Germany: Springer-Verlag, 2000.

[12] C.-S. Chang, "Stability, queue length, and delay of deterministic and stochastic queueing networks," *IEEE Trans. Automatic Control*, vol. 39, no. 5, pp. 913–931, May 1994.

[13] F. Kelly, S. Zachary, and I. Ziedins, "Stochastic Networks: Theory and Applications", *Royal Statistical Society Lecture Notes Series*, vol. 4, Oxford University Press, pp. 141–168, 1996.

[14] G. Kesidis, J. Walrand, and C.-S. Chang, "Effective bandwidths for multiclass Markov fluids and other ATM sources," *IEEE/ACM Trans. Networking*, vol. 1, no. 4, pp. 424–428, Aug. 1993.

[15] A. I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high speed networks," *IEEE/ACM Trans. Networking*, vol. 1, no. 3, pp. 329–343, June 1993.

[16] J. Tang and X. Zhang, "Cross-Layer Modeling for Quality of Service Guarantees Over Wireless Links," *Technical Report*, Texas A&M University. [online]: http://dropzone.tamu.edu/∼xizhang/papers/model.pdf

[17] L. Zheng and D. Tse, "Diversity and multiplexing: A fundamental tradeoff in multiple antenna channels," *IEEE Trans. Inform. Theory*, vol. 49, no. 5, pp.1073–1096, May 2003.

[18] J. Tang and X. Zhang, "Transmit selection diversity with maximal-ratio combining for multicarrier DS-CDMA wireless networks over Nakagami-m Fading Channels," *IEEE J. Select. Areas Commun.*, vol. 24, no. 1, pp. 104–112, Jan. 2006.

[19] M. K. Simon and M. S. Alouini, *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*. New York: Wiley, 2nd Ed., 2005.

Jia Tang (S'03) received the B.S. degree in Electrical Engineering from Xi'an Jiaotong University, Xi'an, China, in 2001. He is currently a research assistant working toward the Ph.D. degree in Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station, Texas, USA.

His research interests include mobile wireless communications and networks, with emphasis on cross-layer design and optimizations, wireless quality-of-service (QoS) provisioning for mobile multimedia networks and wireless resource allocation.

Mr. Tang received Fouraker Graduate Research Fellowship Award from Department of Electrical and Computer Engineering, Texas A&M University in 2005.

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

He is currently an Assistant Professor and the Founding Director of the Networking and Information Systems Laboratory, Department of Electrical and Computer Engineering, Texas A&M University, College Station. He was an Assistant Professor and the Founding Director of the Division of Computer Systems Engineering, Department of Electrical Engineering and Computer Science, Beijing Information Technology Engineering Institute, Beijing, 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, Australia, under a Fellowship from the Chinese National Commission of Education. He worked as a Summer Intern with the Networks and Distributed Systems Research Department, AT&T Bell Laboratories, Murray Hills, NJ, and with AT&T Laboratories Research, Florham Park, NJ, in 1997. He has published more than 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 modulations and coding (AMC), wireless diversity techniques and resource allocations, wireless sensor and Ad Hoc networks, cognitive radio and cooperative communications/relay networks, vehicular Ad Hoc networks, multi-channel 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.

Prof. Zhang received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He received the TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering at Texas A&M University, College Station, in 2006. He also received the Best Paper Award from the IEEE Globecom 2007. He is currently serving as an Editor for the *IEEE Transactions on Wireless Communications*, an Associate Editor for the *IEEE Transactions on Vehicular Technology*, an Associate Editor for the *IEEE Communications Letters*, an Editor for the *Wiley's Journal on Wireless Communications and Mobile Computing*, an Editor for the *Journal of Computer Systems, Networking, and Communications*, and an Associate Editor for the *John Wiley's Journal on Security and Communications Networks*, and is also serving as the Guest Editor for the *IEEE Wireless Communications Magazine* for the special issue on "next generation of CDMA versus OFDMA for 4G wireless applications". He has frequently served as the Panelist on the U.S. National Science Foundation Research-Proposal Review Panels. He is serving or has served as the Co-Chair for the IEEE Globecom 2008 – Wireless Communications Symposium and the Co-Chair for the IEEE ICC 2008 – Information and Network Security Symposium, respectively, the Symposium Chair for the IEEE/ACM International Cross-Layer Optimized Wireless Networks Symposium 2006, 2007, and 2008, respectively, the TPC Chair for the IEEE/ACM IWCMC 2006, 2007, and 2008, respectively, the Poster Chair for the IEEE INFOCOM 2008, the Student Travel Grants Co-Chair for the IEEE INFOCOM 2007, the Panel Co-Chair for the IEEE ICCCN 2007, the Poster Chair for the IEEE/ACM MSWiM 2007 and the IEEE QShine 2006, the Publicity Chair for the IEEE/ACM QShine 2007 and the IEEE WirelessCom 2005, and the Panelist on the Cross-Layer Optimized Wireless Networks and Multimedia Communications at IEEE ICCCN 2007 and WiFi-Hotspots/WLAN and QoS Panel at the IEEE QShine 2004. He has served as the TPC members for more than 50 IEEE/ACM conferences, including the IEEE INFOCOM, IEEE Globecom, IEEE ICC, IEEE WCNC, IEEE VTC, IEEE/ACM QShine, IEEE WoWMoM, IEEE ICCCN, etc.

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