Cross-Layer-Model Based Adaptive Resource Allocation for Statistical QoS Guarantees in Mobile Wireless Networks

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

***Abstract*— We propose a cross-layer-model based adaptive resource-allocation scheme for the diverse quality-of-service (QoS) guarantees over downlink mobile wireless networks. Our proposed scheme dynamically assigns power-levels and time-slots for heterogeneous real-time mobile users to satisfy the variation of statistical delay-bound QoS requirements. To achieve this goal, we apply Wu and Negi's *effective capacity* approach to derive the admission-control and power/time-slot allocation algorithms, guaranteeing the statistical delay-bound for heterogeneous mobile users. When designing such an algorithm, we study the impact of physical-layer issues such as adaptive power-control and channel-state information (CSI) feedback delay on the QoS provisioning performance. Through numerical and simulation results, we observe that the adaptive power adaptation has a significant impact on statistical QoS-guarantees. In addition, the analyses indicate that our proposed resource-allocation algorithms are shown to be able to efficiently support the diverse QoS requirements for various real-time mobile users over different wireless channels. Also, in an in-door mobile environment, e.g., the widely used wireless local-area networks (WLAN), our proposed algorithm is shown to be robust to the CSI feedback delay.**

***Index Terms*— Wireless networks, resource allocation, quality-of-service (QoS), power control, cross-layer design and optimization, effective capacity, real-time multimedia delay-bound.**

I. Introduction

THE DIVERSE quality-of-service (QoS) guarantees for the real-time multimedia transmissions play a critically important role in the next-generation mobile wireless networks. Unlike its wired counterpart networks, supporting the QoS requirement in wireless environment is much more challenging since the time-varying fading channel has the significant impact on the network performance. For wireless QoS guarantees, link adaptation (LA) techniques have been widely considered as the key solution to overcome the impact of the wireless channel. At the physical layer, the most scarce resources are power and spectral-bandwidth. As a result, the LA techniques such as adaptive modulation and power control are developed to enhance the spectral efficiency while maintaining a certain target error performance [1]. However, for real-time wireless multimedia services, the main QoS metric is bounded-delay, instead of high spectral efficiency [2]. Therefore, to support the real-time wireless multimedia QoS, we need to consider the LA techniques not only at the physical-layer, but also at the upper-protocol-layers such as data-link layer when designing the wireless networks. To achieve this goal, in this paper we develop the cross-layer-model based adaptive resource-allocation scheme to support the real-time multimedia QoS in the downlink heterogeneous mobile wireless networks.

Manuscript received May 26, 2006; accepted Aug. 14, 2006. The associate editor coordinating the review of this paper and approving it for publication was D. Wu. 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.2008.060293.

A. The Related Work

QoS provisioning in wireless networks has been widely studied from different perspectives, such as packet scheduling, admission control, traffic specifications, resource reservations, etc. [2]–[11]. In [2] and [3], the authors investigated the real-time and non-real-time QoS provisioning for code-division-multiple-access (CDMA)-based wireless networks. In [4]–[6], several architectures/algorithms were discussed for either implicit or explicit QoS provisioning. In [7]–[8], the authors integrated the finite-state Markov chain (FSMC) with adaptive modulation and coding (AMC), and then jointly considered the physical-layer channel and data-link-layer queuing characteristics. The idea of resource allocation in [7] and [8] is to calculate the reserved bandwidth for each user by appropriate admission control and scheduling. This scheme is developed across the physical-layer and data-link-layer and is thus capable of characterizing the impact of physical-layer variation on the data-link-layer QoS performance. However, the main QoS requirement addressed in [7] and [8] is the *average delay* of the wireless transmission, which does not effectively support the real-time multimedia services, where the key QoS metric is the *bounded delay*, as addressed in this paper.

In [9], [10], the authors proposed a powerful concept termed "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 in the contexts of wired asynchronous transfer mode (ATM) networks [13]–[14]. The effective capacity and effective bandwidth enable us to analyze the *statistical* delay-bound violation and buffer-overflow probabilities, which are critically important for multimedia wireless networks. Based on [10], the authors in [11], [12] proposed a set of resource-allocation schemes for statistical QoS guarantees in wireless networks. The key techniques

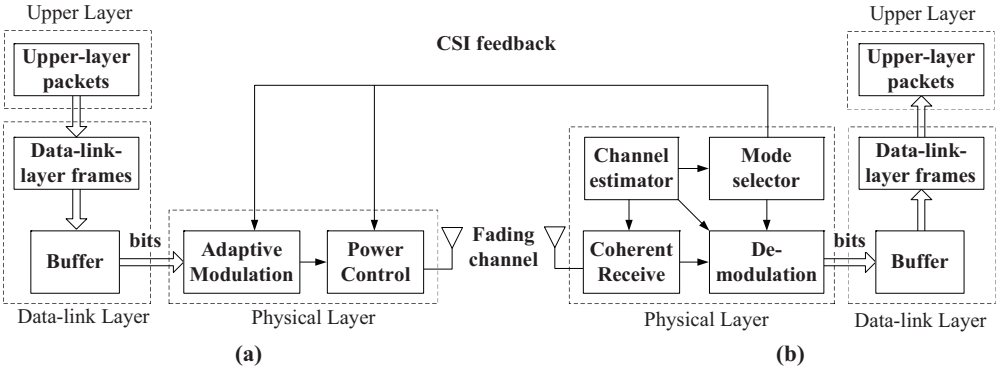


Fig. 1. The system model. (a) Basestation transmitter. (b) The kth mobile receiver.

used in [11], [12] are the integration of effective capacity with multiuser diversity [15], such that the scheme not only provides the statistical QoS for different mobile users, but also increases the total wireless-network's throughput. However, the effective capacity approach has not been explored in cross-layer modeling and design for adaptive resource allocation and QoS guarantees in mobile wireless networks.

B. The Contribution of This Paper

To overcome the aforementioned problems, in this paper we propose a cross-layer-model based adaptive resource-allocation scheme for downlink heterogeneous mobile wireless networks. Based on our application of the effective capacity method [16], [17], the system resources are allocated according to the heterogeneous fading channel statistics, the diverse QoS requirements, and different traffic characteristics. Specifically, our scheme adaptively assigns power-level and time-slots for real-time mobile users in a dynamic time-division multiple access (TDMA) mode to guarantee the bounded delays. We analytically derive the admission-control and power/time-slot allocation conditions to guarantee the *statistical* delay-bound for real-time mobile users. In this paper, we do not employ multiuser diversity because of the following reasons. In a *centralized heterogenous* multiuser network, the multiuser diversity will cause the serious fairness problem — the users with good channels may occupy most of the resources, while the users with poor channels may hardly have opportunity for information transmission, which will result in large queueing delay and thus the user's delay-bound QoS cannot be guaranteed. On the other hand, the advantages of multiuser diversity only contribute to a small portion of mobile users whose channel quality is good, which may not lead to a significant QoS performance improvement from the entire network perspectives. Note that in [18], the authors proposed to use multiuser diversity under the "proportional fairness" constraint. However, this scheme can only support a *loose* delay-bound QoS requirements, which is also not suitable for real-time multimedia services where the delay-bound QoS requirement is *stringent*.

When designing the adaptive resource-allocation algorithm, we address the problems of the physical-layer impact on the statistical QoS provisioning performance. Specifically, we study how adaptive power-control and channel-state information (CSI) feedback delay influence our proposed scheme. Based on our previous work [17], we apply our proposed *QoS-driven power adaptation* for heterogeneous mobile users and compare its performance with conventional water-filling and constant power schemes. Our numerical and simulation results show that our proposed QoS-driven power control has significant advantages over the conventional power controls in terms of QoS-guarantees. On the other hand, our effective-capacity-based adaptive resource-allocation algorithm can efficiently support the QoS requirements for diverse real-time mobile users. In an in-door mobile environment, e.g., the widely used wireless local-area networks (WLAN), the proposed algorithm also provides sufficient robustness to the CSI feedback delay.

The rest of the paper is organized as follows. Section II describes our system model. Sections III briefly introduces the concept of effective capacity. Section IV develops the admission control and time-slot allocation algorithm with fixed average transmission-power. Section V proposes the joint power-level and time-slot allocation scheme. Section VI analyzes the impact of feedback delay on the proposed scheme. The paper concludes with Section VII.

II. System Model

The system model is shown in Fig. 1. In this paper, we concentrate on single-input-single-output (SISO) antenna system with the downlink transmission from the basestation to the mobile users. We denote the total number of mobile users by K, the total spectral-bandwidth of the system by B, and the average transmission-power of the basestation by $\overline{P}$, respectively. We first assume that the average transmission power $\overline{P}$ is fixed. In Section V, we will remove this constraint and let $\overline{P}$ vary within a discrete set. The K users are assumed to be heterogenous, i.e., they may experience different fading conditions and demand different QoS requirements.

As shown by Fig. 1, the upper-protocol-layer packets are first divided into a number of frames at data-link layer. The frames are stored at the transmitter infinite-buffer and then split into bit-streams at physical layer, where the adaptive-modulation and power-control are employed, respectively, to enhance the system performance. The reverse operations are executed at the receiver side. Also, the CSI is estimated at the receiver and fed back to the transmitter for adaptive modulation and adaptive power-control, respectively.

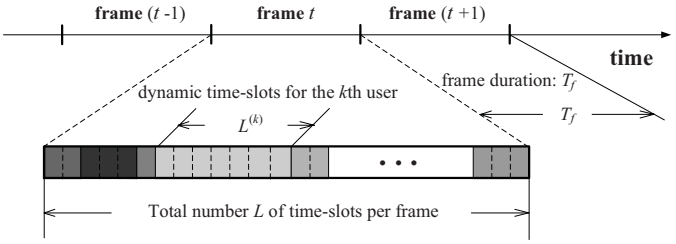


Fig. 2. The frame structure of the proposed system.

A. Data-Link Layer Frame Structure

The frame structure of our proposed system is shown by Fig. 2. In our system, each frame at data-link layer consists of L number of time-slots. The time-duration of each frame is denoted by T_f. Due to the employment of adaptive modulation, the number of bits per frame varies depending on each user's modulation modes selected. As shown in Fig. 2, within the frame duration T_f, the system runs in a dynamic TDMA mode. The kth mobile user is assigned with a number $L^{(k)}$ of time-slots. The number $L^{(k)}$ is determined by the kth mobile user's QoS requirement, which will be detailed in Section IV. Clearly, we have $\sum_{k=1}^{K} L^{(k)} \leq L$.

B. Channel Model

We assume that the wireless fading channel is flat-fading with Nakagami-m distribution. The fading statistics of different mobile users are independent of each other. In this section, we omit the user index k for simplicity. The probability density function (pdf) of the signal-to-noise ratio (SNR), denoted by $p_\Gamma(\gamma)$, can be expressed as [19]

$$p_\Gamma(\gamma) = \frac{\gamma^{m-1}}{\Gamma(m)} \left(\frac{m}{\overline{\gamma}}\right)^m \exp\left(-\frac{m}{\overline{\gamma}}\gamma\right), \ \gamma \geq 0 \tag{1}$$

where $\Gamma(\cdot)$ represents the complete Gamma function, m denotes the fading parameter of Nakagami-m distribution, and $\overline{\gamma}$ denotes the average SNR of the combined signal, which can be expressed as $\overline{\gamma} = \overline{P}E\{\alpha^2\}/(N_0 B)$, where $E\{\alpha^2\}$ is the average path-gain of the Nakagami fading channel and N_0 is the single-sided power spectral density (PSD) of the complex additive white Gaussian noise (AWGN). Note that when the average power-level $\overline{P}$ varies, the corresponding average SNR $\overline{\gamma}$ will change accordingly.

We use Nakagami-m channel model because this model is very general and often best fits the land-mobile and indoor-mobile multipath propagations [19]. As the fading parameter m varies, where $m \in [1/2, +\infty)$, the model spans a wide range of fading environments, including one-sided Gaussian fading channel ($m = 1/2$, the worst fading case), the Rayleigh fading channel ($m = 1$), the precise approximations of Rician and lognormal fading channels ($m > 1$), and the conventional Gaussian channel ($m = \infty$, no fading). The channel is assumed to be invariant within a frame's time-duration T_f, but varies from one frame to another. Furthermore, we assume that the CSI is perfectly estimated at the receiver and reliably fed back to the transmitter with a time-delay denoted by τ. First, we assume $\tau = 0$, implying the perfect CSI feedback. We will address the scenario with delayed CSI feedback in Section VI.

C. Adaptive Modulation

Adaptive modulation is an efficient LA technique to improve the spectral-efficiency at physical layer. In this paper, we employ the adaptive QAM modulation proposed in [1]. The specific modulation modes for the adaptive-modulation scheme are constructed as follows. We partition the entire SNR range by N non-overlapping consecutive intervals, resulting in $N+1$ boundary points denoted by $\{\Gamma_n\}_{n=0}^{N}$, where $\Gamma_0 < \Gamma_1 < \cdots < \Gamma_N$ with $\Gamma_0 = 0$ and $\Gamma_N = \infty$. Correspondingly, the adaptive modulation is selected to be in mode n if the SNR, denoted by γ, falls into the range of $\Gamma_n \leq \gamma < \Gamma_{n+1}$. The zero-th mode corresponds to the "outage" mode of the system, i.e., the transmitter stops transmitting data in Mode 0. The constellation used for the nth mode is M_n-QAM, where $M_n = 2^n$ with $n \in \{1, 2, ..., N-1\}$. Let us further define $M_0 = 0$ and $M_N = \infty$. Thus, the spectral-efficiency of the adaptive modulation ranges from 0 to $N-1$ bits/sec/Hz. As the SNR increases, the system selects the 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 stops transmitting data as in the "outage" mode.

The bit-error rate (BER) when using the nth mode for $n \in \{1, 2, ..., N-1\}$, denoted by BER_n, can be approximated as follows [1]

$$\mathrm{BER}_n \approx 0.2 \exp\left(-g_n \gamma\right) \tag{2}$$

where $g_n = 3/[2(M_n - 1)]$. Based on the pdf given in Eq. (1), the probability π_n, that the SNR falls into mode n is determined by

$$\pi_n = \int_{\Gamma_n}^{\Gamma_{n+1}} p_\Gamma(\gamma) d\gamma = \frac{\Gamma\left(m, \frac{m}{\overline{\gamma}}\Gamma_n\right)}{\Gamma(m)} - \frac{\Gamma\left(m, \frac{m}{\overline{\gamma}}\Gamma_{n+1}\right)}{\Gamma(m)} \tag{3}$$

where $\Gamma(\cdot, \cdot)$ represents the incomplete Gamma function and $n \in \{0, 1, ..., N-1\}$.

In general, the forward-error control (FEC) and automatic retransmission request (ARQ) are also employed at the physical/data-link layer. However, in this paper we only focus on uncoded system due to the following reasons. First, there exist the simple *analytical* power-control policies [1][17] for uncoded transmissions, while for coded transmission, it is difficult to find such a policy. Thus, we assume uncoded transmission for analytical convenience. Second, based on our study in [16], we observe that the performance trends of FEC/ARQ-based transmission is similar to that of uncoded systems, as long as the link BER is not too high. Therefore, the investigation of the uncoded system also provides a guideline on designing the coded system.

D. Power Control

We mainly investigate three different power-control strategies, namely, our proposed QoS-driven power control [17], the water-filling power control, and the constant-power approach. For different power-control strategies, the power-control law as well as the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ are different. We study how to adjust the power and decide the boundary points

for the above three power-control strategies, respectively, as follows.

Strategy I: QoS-Driven Optimal Power Control. In [17], we develop the QoS-driven optimal power-control strategy for the adaptive QAM modulation. Let the BER QoS requirement of the system be denoted by $\mathrm{P_{tgt}}$. In order to achieve the target BER, i.e., $\mathrm{P_{tgt}}$, the power-control law, denoted by $\mu_n(\gamma)$, for the nth mode can be derived as [17, eq. (22)]

$$\mu_n(\gamma)=\begin{cases}(M_n-1)\dfrac{1}{\nu_n\gamma}, & M_n\le\frac{\gamma}{\gamma_0}<M_{n+1},\ (n\neq 0)\\ 0, & \frac{\gamma}{\gamma_0}<M_1,\ (n=0)\end{cases} \tag{4}$$

where $\nu_n=-1.5/\log(5\mathrm{P_{tgt}})$ and γ_0 is the cut-off threshold, which can be numerically obtained by meeting the following mean power constraint:

$$\sum_{n=1}^{N-1}\int_{\Gamma_n}^{\Gamma_{n+1}}\mu_n(\gamma)p_\Gamma(\gamma)d\gamma=1 \tag{5}$$

where we have [17, eq. (33)]

$$\Gamma_n=\gamma_0 M_n^{\frac{\kappa T_f B\theta}{\log 2}+1} \tag{6}$$

where θ is called QoS-exponent [9], [10], which will be detailed in Section III, and $\kappa\ge 1$ is a parameter to deal with the impact of channel correlation. Specifically, when the channel process is uncorrelated (i.e., block fading channel), then we have $\kappa=1$. Otherwise, when the channel process is correlated, κ should be adjusted according to the channel Doppler frequency f_d, see [17] for details. Once the cut-off threshold γ_0 is determined, the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ can be obtained by using Eq. (6). The QoS-driven power control makes the BER of each mode equal to $\mathrm{P_{tgt}}$. Then, the resulting system BER is also equal to $\mathrm{P_{tgt}}$.

Strategy II: Water-Filling Power Control. In [1], the authors proposed the optimal power-control strategy for adaptive MQAM that can maximize the spectral-efficiency, which is actually based on the time-domain water-filling algorithm. However, based on our study in [17], we find that the water-filling power control can be considered as a special case of our proposed QoS-driven power control by letting the QoS exponent $\theta\to 0$. Thus, the power-control law and mean power constraint of the water-filling scheme are the same as those given by Eqs. (4) and (5), respectively. The boundary points are determined by

$$\Gamma_n=\lim_{\theta\to 0}\gamma_0 M_n^{\frac{\kappa T_f B\theta}{\log 2}+1}=\gamma_0 M_n. \tag{7}$$

Strategy III: Constant-Power Approach. Constant power-control approach is to keep the transmission power at the basestation as a constant. Using Eqs. (1) and (2), the average BER of the mode n, denoted by $\overline{\mathrm{BER}}_n$, can be derived as

$$\begin{aligned}\overline{\mathrm{BER}}_n &= \frac{1}{\pi_n}\int_{\Gamma_n}^{\Gamma_{n+1}}0.2\exp(-g_n\gamma)p_\Gamma(\gamma)d\gamma\\ &= \frac{0.2\left(\frac{m}{b_n}\right)^m}{\pi_n\Gamma(m)}\left[\Gamma\left(m,\frac{b_n\Gamma_n}{\overline{\gamma}}\right)-\Gamma\left(m,\frac{b_n\Gamma_{n+1}}{\overline{\gamma}}\right)\right]\end{aligned} \tag{8}$$

where $b_n=g_n\overline{\gamma}+m$ for $n\in\{1,2,...,N-1\}$ and the boundary points are determined by

$$\Gamma_n=\frac{\eta}{g_n} \tag{9}$$

where the parameter η ($\eta>0$) in Eq. (9) is numerically obtained by meeting the following constraint on the average BER requirement $\mathrm{P_{tgt}}$:

$$\mathrm{P_{tgt}}=\frac{\sum_{n=1}^{N-1}n\pi_n\overline{\mathrm{BER}}_n}{\sum_{n=1}^{N-1}n\pi_n}. \tag{10}$$

where $\overline{\mathrm{BER}}_n$ is the function of η through Eqs. (8) and (9). Once the parameter η is determined, the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ can be obtained by using Eq. (9).

E. Service Process Modeling by Using FSMC

In this paper, we employ the FSMC model to characterize the variation of the wireless service process. Each state of FSMC corresponds to a mode of the adaptive-modulation scheme. Let $p_{i,j}$ denote the transition probability from state i to state j. We assume a slow-fading channel model such that the transition only happens between adjacent states [20]. Under such an assumption, we have $p_{ij}=0$ for all $|i-j|>1$. The adjacent transition probability can be approximated as [20]

$$\begin{cases}p_{n,n+1}\approx\frac{N_\Gamma(\Gamma_{n+1})T_f}{\pi_n}, & \text{where } n=0,1,...,N-2,\\ p_{n,n-1}\approx\frac{N_\Gamma(\Gamma_n)T_f}{\pi_n}, & \text{where } n=1,2,...,N-1\end{cases} \tag{11}$$

where $N_\Gamma(\gamma)$ is the level-crossing rate (LCR) determined by SNR of γ, which is given by [19]

$$N_\Gamma(\gamma)=\frac{\sqrt{2\pi}f_d}{\Gamma m}\left(\frac{m\gamma}{\overline{\gamma}}\right)^{m-\frac{1}{2}}\exp\left(-\frac{m\gamma}{\overline{\gamma}}\right) \tag{12}$$

where f_d is the maximum Doppler frequency of the mobile user. Then, the remaining transition probabilities can be derived by using Eq. (11) as follows:

$$\begin{cases}p_{0,0}=1-p_{0,1}\\ p_{N-1,N-1}=1-p_{N-1,N-2}\\ p_{n,n}=1-p_{n,n-1}-p_{n,n+1},\ n=1,...,N-2.\end{cases} \tag{13}$$

Applying Eqs. (11) and (13), we obtain the probability transition matrix of the FSMC, denoted by $\mathbf{P}=[p_{ij}]_{N\times N}$. Correspondingly, we obtain the stationary distribution of the FSMC, denoted by $\boldsymbol{\pi}$, as follows:

$$\boldsymbol{\pi}=[\,\pi_0,\pi_1,...\pi_{N-1}\,] \tag{14}$$

where π_n is given by Eq. (3) for $n\in\{0,1,...,N-1\}$.

III. A Preliminary of the Effective Capacity

A. Statistical QoS Guarantees

The real-time multimedia services such as video and audio require the bounded delay, or equivalently, the guaranteed 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 practically infeasible to be achieved due to the impact of the time-varying fading channels. For example,

over the Rayleigh fading channel, the only lower-bound of the system bandwidth that can be *deterministically* guaranteed is a bandwidth of zero [10]. Thus, 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 [13]–[14]. The asymptotic results in [13] showed that, for *stationary* arrival and service processes with the average arrival-rate less than the average service-rate, the probability that the queue size Q exceeds a certain threshold C decays exponentially fast as the threshold C increases, i.e.,

$$\Pr\{Q > C\} \approx e^{-\theta C} \tag{15}$$

where θ is a certain positive constant called *QoS exponent* [9], [10] to be detailed below. 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. (15) can be obtained as

$$\Pr\{\text{Delay} > D_{\max}\} \approx e^{-\theta\delta D_{\max}} \tag{16}$$

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

From Eqs. (15)–(16), 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. The smaller θ corresponds to the 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], [10].

B. Effective Bandwidth and Effective Capacity

In [9], [10], the authors proposed a powerful concept termed as *effective capacity*, which turns out to be the *dual problem* of the effective bandwidth. The effective capacity characterizes the attainable wireless-channel service-rate as a function of the QoS exponent θ, and thus we can use it as a bridge in cross-layer design modeling between physical-layer system infrastructure and data-link-layer's statistical QoS performance.

To help demonstrate the principles and identify the relationships between effective bandwidth and effective capacity, let us consider the case as illustrated in Fig. 3. For any given arrival process and service process, we sketch their effective-bandwidth function, denoted by $\mathrm{E_B}(\theta)$, and effective-capacity function, denoted by $\mathrm{E_C}(\theta)$, in Fig. 3, respectively. 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 as shown in Fig. 3. In general, the delay-bound violation probability can be calculated as following steps:

Step1: 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 (δ, θ^*) such that $\mathrm{E_B}(\theta^*) = \mathrm{E_C}(\theta^*) = \delta$.

Step2: For any pre-determined delay-bound $D_{\max}$, the delay-bound violation probability can be derived using Eq. (16) as

$$\Pr\{\text{Delay} > D_{\max}\} \approx e^{-\theta^*\delta D_{\max}} \tag{17}$$

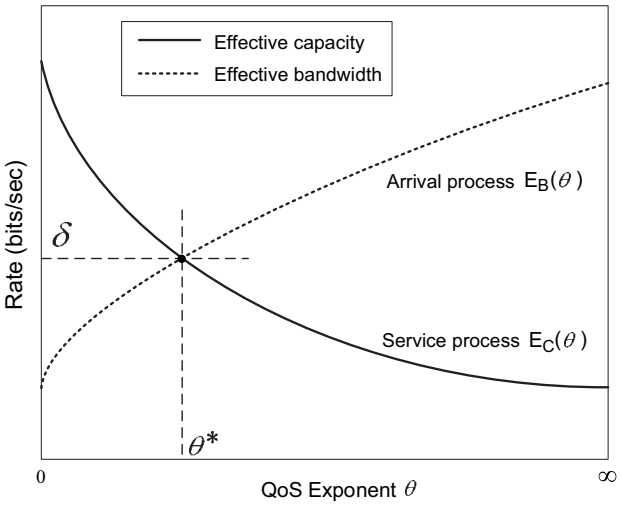


Fig. 3. Relationships between effective bandwidth and effective capacity as a function of the QoS exponent θ.

From the above observations and analyses, we propose to use the effective bandwidth and 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 very simple and efficient cross-layer approach.

C. Effective Capacity of Our Proposed Scheme

As described above, the effective capacity is the dual problem of the effective bandwidth. Thus, utilizing the well-established effective bandwidth theory, it is feasible to formulate the effective capacity problem analytically. We showed in [16] that based on our physical-layer FSMC model, the effective capacity function $\mathrm{E_C}(\theta)$ can be expressed as follows:

$$\mathrm{E_C}(\theta) = -\frac{1}{\theta}\log\Big(\rho\{\mathbf{P}\,\mathbf{\Phi}(\theta)\}\Big),\ \theta > 0 \tag{18}$$

where $\rho\{\cdot\}$ denotes the spectral radius of the matrix, $\mathbf{P}$ is the transition matrix of our developed FSMC, and

$$\mathbf{\Phi}(\theta) \triangleq \mathrm{diag}\left\{e^{-\lambda_0\theta}, e^{-\lambda_1\theta}, ..., e^{-\lambda_{N-1}\theta}\right\} \tag{19}$$

where λ_n with $n \in \{0, 1, ..., N-1\}$ is the number of bits per frame transmitted by the nth mode of the adaptive-modulation scheme.

IV. Adaptive Resource Allocation with Fixed Average Power

The cross-layer modeling introduced in Section III establishes the analytical framework to investigate the impact of physical-layer infrastructure variations on the statistical QoS provisioning performance at the data-link-layer through the effective capacity function. In this section, we develop the

adaptive resource-allocation algorithms based on our developed cross-layer model to guarantee the desired QoS requirements. Since our focus is mainly on resource allocation in this paper, we only adopt the simple round-robin (RR) scheduling for the real-time mobile users.

A. The Effective Capacity of the Service Process

As described in Section II-A, our proposed system operates in a dynamic TDMA mode. As shown in Fig. 2, the kth user is assigned with $L^{(k)}$ of time-slots per frame for information transmission. In order to determine the number $L^{(k)}$ of time-slots allocated to the kth user to support its statistical QoS, we first need to derive the effective capacity of the service-process. Consider only allocating $L^{(k)} = 1$ time-slot as a basic-unit to the kth user, the effective capacity of the kth user, denoted by $\mathrm{E}_{\mathrm{C}}^{(k,1)}(\theta)$, can be expressed using Eq. (18) as

$$\mathrm{E}_{\mathrm{C}}^{(k,1)}(\theta) = -\frac{1}{\theta}\log\left(\rho\left\{\mathbf{P}^{(k)}\,\mathbf{\Phi}^{(1)}(\theta)\right\}\right),\ \theta > 0 \tag{20}$$

where $\mathbf{P}^{(k)}$ is the transition probability matrix of the kth user, which is determined by the kth user's channel statistics and is independent of $L^{(k)}$, and $\mathbf{\Phi}^{(1)}(\theta)$ is given by $\mathbf{\Phi}^{(1)}(\theta) = \mathrm{diag}\left\{e^{-\lambda_0^{(1)}\theta}, e^{-\lambda_1^{(1)}\theta}, ..., e^{-\lambda_{N-1}^{(1)}\theta}\right\}$, where $\lambda_n^{(1)} = nT_fB/L,\ n \in \{0, 1, ..., N-1\}$, which is independent of the channel statistics.

When allocating $L^{(k)} = l$ time-slots for the user, applying the results developed in [12], the effective capacity, denoted by $\mathrm{E}_{\mathrm{C}}^{(k,l)}(\theta)$, can be expressed as

$$\mathrm{E}_{\mathrm{C}}^{(k,l)}(\theta) = l\mathrm{E}_{\mathrm{C}}^{(k,1)}(l\theta). \tag{21}$$

B. Admission-Control and Time-Slot Allocation

Let the kth user's statistical QoS requirement be denoted by $\{D_{\max}^{(k)}, \varepsilon^{(k)}\}$, where $D_{\max}^{(k)}$ is the delay-bound and $\varepsilon^{(k)}$ is the violation probability. Similar to the procedure described in Section III-B, the time-slot allocation algorithms can be designed in the following steps:

S1: Denote the effective bandwidth of the kth user's arrival-process by $\mathrm{E}_{\mathrm{B}}^{(k)}(\theta)$. Find the solution of the rate and QoS-exponent (δ_l, θ_l) such that $\mathrm{E}_{\mathrm{B}}^{(k)}(\theta_l) = \mathrm{E}_{\mathrm{C}}^{(k,l)}(\theta_l) = \delta_l$.

S2: Using $L^{(k)} = l$ number of time-slots, the delay-bound violation probability can be derived as

$$\Pr\{\mathrm{Delay} > D_{\max}^{(k)}\} \approx \exp\left(-\theta_l\delta_l D_{\max}^{(k)}\right) \tag{22}$$

S3: The number $L^{(k)}$ is determined by

$$L^{(k)} = \min_{1\le l\le L}\left\{\, l \,\right\},\ \text{s.t. } \exp\left(-\theta_l\delta_l D_{\max}^{(k)}\right) \le \varepsilon^{(k)}. \tag{23}$$

For each real-time user, $L^{(k)}$ can be calculated using Eq. (23). Clearly, the total number of time-slots that are allocated to the real-time users needs to satisfy the following equation:

$$\sum_{k=1}^{K} L^{(k)} \le L. \tag{24}$$

TABLE I
QOS REQUIREMENTS FOR AUDIO AND VIDEO SERVICES.

	BER $\mathrm{P_{tgt}}$	Delay-bound $D_{\max}$	Violation Prob. ε
Audio	10^{-3}	50 ms	10^{-2}
Video	10^{-4}	150 ms	10^{-3}

When a new mobile real-time user applies to join the system, the admission-control algorithm examines if the number of available time-slot resources is sufficient to support the new real-time mobile user's statistical QoS. If yes, the new real-time mobile user is admitted to join the system; otherwise, this new real-time mobile user is rejected to join the system.

C. Numerical and Simulation Results

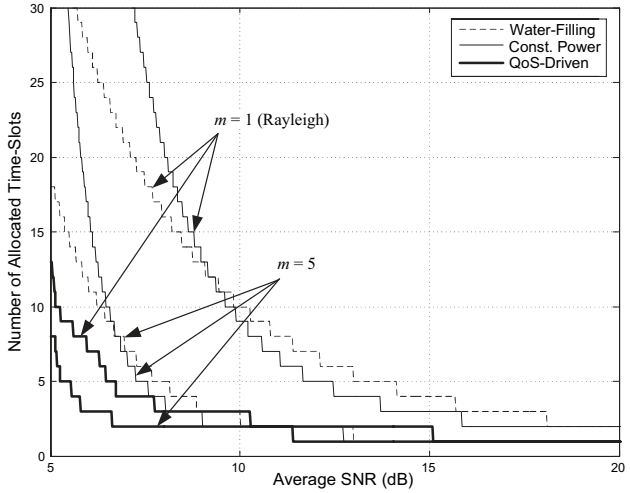
We evaluate the proposed time-slot allocation algorithms through numerical solutions and simulations. In the following, we set the number of adaptive-modulation modes $N = 8$, the total system spectral-bandwidth $B = 1000$ KHz, the data-link-layer frame time-duration $T_f = 2$ ms, the number of time-slots per frame $L = 100$, and the maximum Doppler frequency $f_d = 15$ Hz. Moreover, we generate two types of real-time services. The first type simulates the low speed audio service, where we model the arrival traffic by the well-known ON-OFF fluid model. The holding times in "ON" and "OFF" states are exponentially distributed with the mean equal to 8.9 ms and 8.4 ms, respectively. The "ON" state traffic is modeled as a constant-rate of 32 Kbps. The second one simulates a high-speed video traffic flow. We employ a first-order auto-regressive (AR) process to simulate video traffic characteristics [21], the bit-rate of which can be expressed as

$$\nu(t) = a\nu(t-1) + bw \tag{25}$$

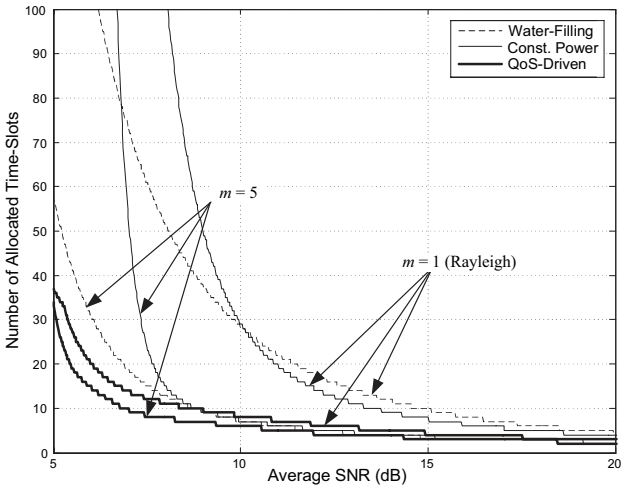
where $a = 0.8781$, $b = 0.1108$ [21] and w is a Gaussian random variable with the mean 80 Kbps and the standard deviation of 30 Kbps. The effective bandwidth of the audio and video traffic is derived according to [13] and [14], respectively. The QoS requirements of these two types of services are shown in TABLE I.

Using the time-slot allocation algorithm proposed in Section IV-B, Fig. 4 shows the numerical results of allocated time-slots for audio and video services as a function of the average SNR. As shown by Fig. 4, for both audio and video services, the required time-slots for supporting the QoS decreases as the average SNR increases. The better quality channel (fading parameter $m = 5$) needs the fewer number of time-slots than the Rayleigh fading channel (fading parameter $m = 1$). When the SNR is low, the time-slot allocation algorithms may not find the feasible solution of the $L^{(k)}$ to support the required QoS, since $L^{(k)}$ must satisfy $1 \le L^{(k)} \le L$. From Fig. 4 we can also observe that our proposed QoS-driven power control has significant superiorities over both the conventional water-filling scheme and constant power approach.

To evaluate whether the allocated time-slots can support the required statistical QoS, Fig. 5 plots the simulated delay-bound violation probabilities for video and audio services using our proposed QoS-driven power control. We can obverse from Fig. 5 that for both audio and video services the delay-bound violation probabilities are below the required upper-bounds



(a) Audio time-slot allocation.



(b) Video time-slot allocation.

Fig. 4. The numerical time-slot allocation results for audio and video services.

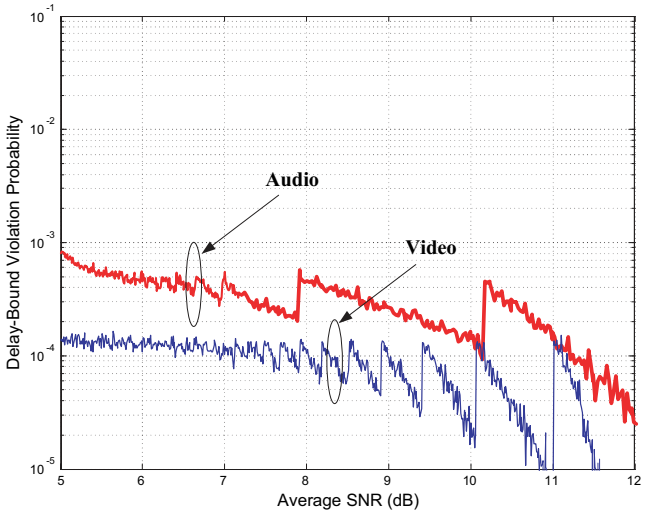


Fig. 5. Simulation results of the delay-bound violation probability for QoS-driven power control. The fading parameter $m = 1$ (Rayleigh fading channel).

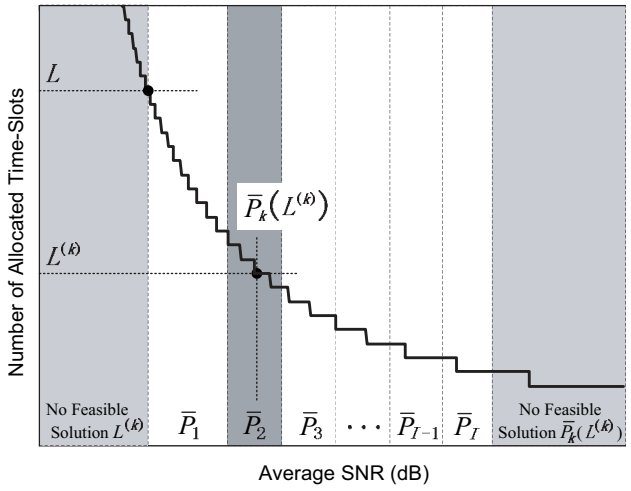


Fig. 6. The time-slot and power-level mapping relations.

ε's. The simulated delay-bound violation probability is lower than the designated delay-bound violation probability ε, which is due to the fact that the approximations in Eqs. (15) and (16) are conservative [10], [16]. Interestingly, Fig. 5 shows that the QoS-violation probability *fluctuates* according to the time-slot allocation outcomes, which is because our time-slot allocation results vary within a discrete set. For the conventional water-filling scheme and constant power approach, we can observe the similar delay-bound violation probability performance, which is omitted for lack of space. Note that the conventional power control schemes achieve the similar QoS violation performance by using much more resources (i.e., time-slots, see Fig. 4) than our proposed QoS-driven power control scheme.

V. Joint Power-Level and Time-Slot Allocation

A. Power-Level and Time-Slot Allocation Using Dynamic Programming

In previous sections, we assume that the average transmission power $\overline{P}$ at the basestation transmitter is fixed. In this section, we remove this constraint and let the average transmission power vary within a discrete set. In fact, setting the initial power-level has already been adopted in, e.g., UMTS 3GPP standard [22] for cellular networks. However, in [22] it does not mention how to adjust the power-level to guarantee the QoS requirement. In this paper, the idea of joint power-level and time-slot allocation can be described as follows. To guarantee the kth user's QoS requirement, the basestation may assign a larger number of time-slots while using a lower power-level; it is also possible to allocate a fewer number of time-slots while using a higher power-level. The goal of our proposed joint power-level and time-slot allocation algorithm is to assign each user with time-slots and power-levels such that the user's QoS requirement is guaranteed while minimizing the total transmission energy. Thus, when the number of users is large or the channel quality is poor, the basestation can increase its transmission power-level to admit more mobile users. On the other hand, when the number of mobile users is small or the channel quality is good enough, the basestation can decrease its transmission power-level while still guaranteeing the desired QoS requirements. In a multi-

cell wireless networks, e.g., the cellular networks, this will not only save the power resources at the basestation, but also generate less interference to the other cells.

It is clear that under current problem formulation, we can also use different power-control policies *for each given power-level*. However, in this section, we will only focus on our proposed QoS-driven power control, since this scheme offers the optimal performance. Let the set of the discrete average power-levels be denoted by $\mathcal{P} = \{\overline{P}_1, \overline{P}_2, ..., \overline{P}_I\}$, where $0 < \overline{P}_1 < \overline{P}_2 < ... < \overline{P}_I$. Moreover, let $\overline{P}_k(L^{(k)})$ denote the minimum power-level that is required to guarantee the kth user's QoS requirement when allocating $L^{(k)}$ time-slots to the mobile user. Then, the problem of our dynamic resource-allocation can be formulated as follows:

$$\textbf{Objective:} \qquad \min \left\{ \sum_{k=1}^{K} L^{(k)} P_k(L^{(k)}) \right\} \tag{26}$$

subject to:

$$\begin{cases} 1 \le L^{(k)} \le L, \ \forall k \in \{1, 2, ..., K\} \\ \sum_{k=1}^{K} L^{(k)} \le L \end{cases} \tag{27}$$

where

$$\overline{P}_k(L^{(k)}) = \min \left\{ \overline{P} \in \mathcal{P} \,\middle|\, \exp\left(-\theta_{L^{(k)}} \delta_{L^{(k)}} D_{\max}^{(k)}\right) \le \varepsilon^{(k)} \right\}. \tag{28}$$

To obtain the feasible solutions of the time-slots $L^{(k)}$ and the power-level $\overline{P}_k(L^{(k)})$, let us consider the procedure illustrated in Fig. 6. Given a time-slot-allocation table obtained from Section IV (e.g., Fig. 4), we can partition the average-SNR range by a number of consecutive intervals, with each interval corresponding to a power-level. At the range where the average SNR is too low (as shown by the shaded-area in the left-hand-side of Fig. 6), there is no feasible solution of $L^{(k)}$ due to the constraint of Eq. (27) that $L^{(k)}$ must satisfy $L^{(k)} \le L$. On the other hand, at the range where the average SNR is too large (as shown by the shaded-area on the right-hand-side of Fig. 6), there is no feasible solution of $\overline{P}_k(L^{(k)})$ due to the condition of Eq. (28) that $\overline{P}_k(L^{(k)})$ must satisfy $\overline{P}_k(L^{(k)}) \le \overline{P}_I$. At the range in between, each average SNR-interval is achieved by using certain power-level $\overline{P}_i$, where $i \in \{1, 2, ..., I\}$. Then, for a given $L^{(k)}$, the required power-level $\overline{P}_k(L^{(k)})$ can be obtained by mapping $L^{(k)}$ into the corresponding SNR-interval. For example, for the case shown in Fig. 6, the power-level $\overline{P}_k(L^{(k)})$ falls into the SNR-interval belonging to $\overline{P}_2$ (as shown by the shaded-area in the middle of Fig. 6). Therefore, the required minimum power-level is $\overline{P}_k(L^{(k)}) = \overline{P}_2$.

Once $\overline{P}_k(L^{(k)})$ is attained, this minimization problem can be solved by the dynamic programming (DP) approach [23]. Let us define $u_k(l) \triangleq l\overline{P}_k(l)$, where $l = 1, 2, ..., L$. The cost function of the first mobile user, denoted by $\mathcal{J}_1(l)$, can be expressed as

$$\mathcal{J}_1(l) = u_1(l). \tag{29}$$

Then, the cost function for the kth mobile user can be derived iteratively as:

$$\mathcal{J}_k(l) = \min_{1 \le t \le l-1} \left\{ u_k(t) + \mathcal{J}_{k-1}(l-t) \right\}, \ \text{for } k \le l \le L \tag{30}$$

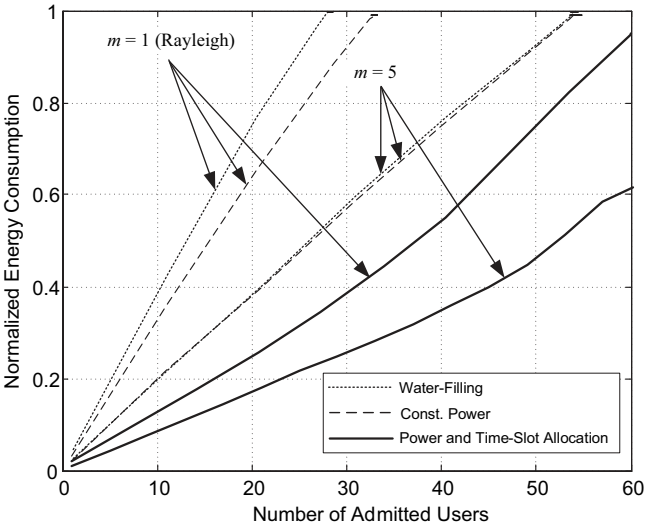


Fig. 7. The average energy consumption comparisons.

where $k = 2, 3, ..., K$. The resource-allocation algorithm is executed every time when the new mobile user arrives or the old mobile user leaves. In the case when the new user tries to join the network, it is possible that there is no feasible solution for the above problem. Thus, the basestation cannot support the QoS requirement for the admission-testing mobile user and therefore this mobile user is rejected to join the wireless networks. Otherwise, the new mobile user is assigned with certain time-slots and power-level for transmissions.

B. Complexity Discussions

In general, the complexity of finding the optimal power-level and time-slots for multiple mobile users *exponentially* increases with the dimension of the searching space. For example, for K users each being assigned with L time-slots and I power-levels, the complexity is approximately proportional to $(LI)^K$. In contrast, by using our proposed dynamic-programming based allocation scheme, the complexity is *linearly* increased with LK. The key reasons of this complexity decreasing include the followings. First, the employment of dynamic programming reduces the exponential complexity to linear complexity. Second, by using the power-level mapping procedure introduced in Section V-A, the burden of finding the minimum power-level (with complexity proportional to I) is transferred to look up the "time-slot allocation table" as shown by Fig. 6. Therefore, the complexity of dynamic-programming is independent of I. In practical systems, this time-slot allocation table can be calculated off-line and stored at the basestation in advance, without costing run-time CPU resources.

C. Simulation Results

We also conduct simulations to evaluate our proposed joint power-level and time-slot allocation algorithms. In the simulations, the traffic types are randomly selected between audio and video services with probability of 50% for each type. We set the discrete average-power varying within a dynamic range of ± 3 dB, with 7 discrete levels

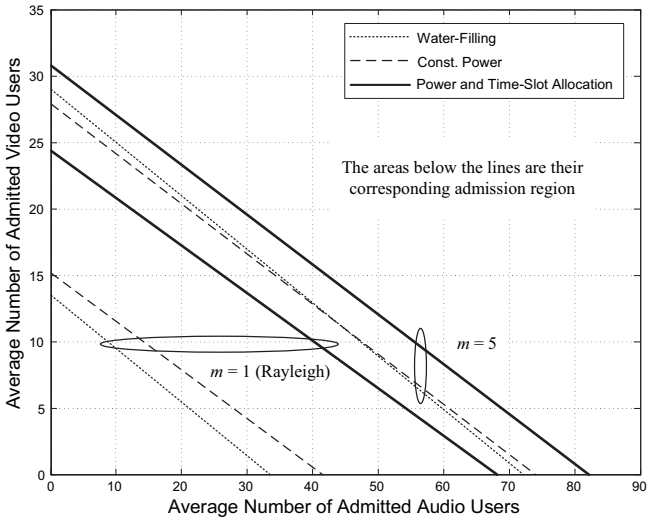


Fig. 8. The average admission region of the system.

$\{-3\,\mathrm{dB}, -2\,\mathrm{dB}, ..., 2\,\mathrm{dB}, 3\,\mathrm{dB}\}$ relative to the central power-level (0 dB). Also, for a fair comparison with the results in previous sections, we let the SNR of each user be uniformly distributed between 5 dB and 25 dB when using the central power-level (0 dB). Note that in UMTS 3GPP standard [22], the power-level dynamic range is ± 9 dB (normal condition) and ± 12 dB (extreme condition), which is much larger than that used in our simulation. Therefore, our simulation results are still conservative in terms of performance improvements.

Fig. 7 plots the average energy consumption comparisons between the above three schemes, where the power is normalized by the central power-level (0 dB). We can also observe from Fig. 7 that the joint power-level and time-slot allocation has significant advantage over the water-filling and constant-power approaches. Fig. 8 depicts the simulation results of the average admission-regions for the video and audio users. As shown by Fig. 8, the averaged admission region can be enlarged by the dynamic-programming-based resource allocation. When the fading parameter $m = 5$, the improvement is not as significant as that in Rayleigh fading channel, which is due to the system capacity limit ($L = 100$). However, our simulations show that this admission region is achieved by using only 65% of the power as compared to that in Rayleigh channel.

VI. The Impact of Feedback Delay

In previous sections, we assume that the CSI is reliably fed back to the transmitter without error and delay. However, in practice, this assumption hardly holds. In particular, the CSI feedback delay is un-avoidable in most situations. Without loss of generality, we discuss the impact of feedback delay on a single user and omit the user-index for simplicity.

In order to guarantee the reliability QoS, the system needs to maintain the same BER as that for the case without feedback delay. As a result, the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ for the adaptive modulation should be re-calculated. In [1], the authors analyzed the impact of CSI feedback delay on BER performance for the adaptive modulation. In [24], we also investigated the feedback delay issue for transmit-selection-combining (SC)/receive-maximal-ratio combining (MRC)-based multiple-input-multiple-output (MIMO) scheme from BER perspective. Using the similar approach to [1], [24], we study the impact of feedback delay on the system's delay-bound QoS performance for different power-control policies as follows.

A. QoS-Driven and Water-Filling Power Controls

We first investigate our proposed QoS-driven power control. When considering the feedback delay, the transmission procedure can be described as follows. The constellation M_n is determined based on the SNR γ at time t, but the constellation is transmitted at time $t+\tau$ with actual SNR denoted by $\widehat{\gamma}$. In order to achieve the actual BER of $\mathrm{P_{tgt}}$ as in the case without delay, the system needs to be designed to operate at a lower target BER, which is denoted by $\mathrm{P'_{tgt}}$. According to Eq. (2), the instantaneous BER at time $t+\tau$, denoted by $\mathrm{BER}_n(\widehat{\gamma}|\gamma)$, is given by

$$\begin{aligned}\mathrm{BER}_n(\widehat{\gamma}|\gamma) &= 0.2\exp\Big(-g_n\mu_n(\gamma)\widehat{\gamma}\Big)\\ &= 0.2\exp\left(\frac{\log(5\mathrm{P'_{tgt}})\widehat{\gamma}}{\gamma}\right) \qquad (31)\end{aligned}$$

where $\mu_n(\gamma)$ is the QoS-driven power-control law given by Eq. (4), except that $\mathrm{P_{tgt}}$ in the parameter ν_n should be replaced by the new target BER $\mathrm{P'_{tgt}}$. Then, we obtain the average BER with a given γ, denoted by $\mathrm{BER}_n(\gamma)$, as follows:

$$\mathrm{BER}_n(\gamma) = \int_0^\infty \mathrm{BER}_n(\widehat{\gamma}|\gamma)\; p_{\widehat{\Gamma}|\Gamma}(\widehat{\gamma}|\gamma)\, d\widehat{\gamma} \qquad (32)$$

where $p_{\widehat{\Gamma}|\Gamma}(\widehat{\gamma}|\gamma)$ is the pdf of $\widehat{\gamma}$ conditioned on γ, which is given by [24]

$$\begin{aligned}p_{\widehat{\Gamma}|\Gamma}(\widehat{\gamma}|\gamma) &= \frac{1}{(1-\rho)}\left(\frac{m}{\overline{\gamma}}\right)\left(\frac{\widehat{\gamma}}{\rho\gamma}\right)^{\frac{m-1}{2}}\\ &\quad\cdot\exp\left(-\frac{m(\rho\gamma+\widehat{\gamma})}{(1-\rho)\overline{\gamma}}\right)I_{m-1}\left(\frac{2m\sqrt{\rho\gamma\widehat{\gamma}}}{(1-\rho)\overline{\gamma}}\right) \qquad (33)\end{aligned}$$

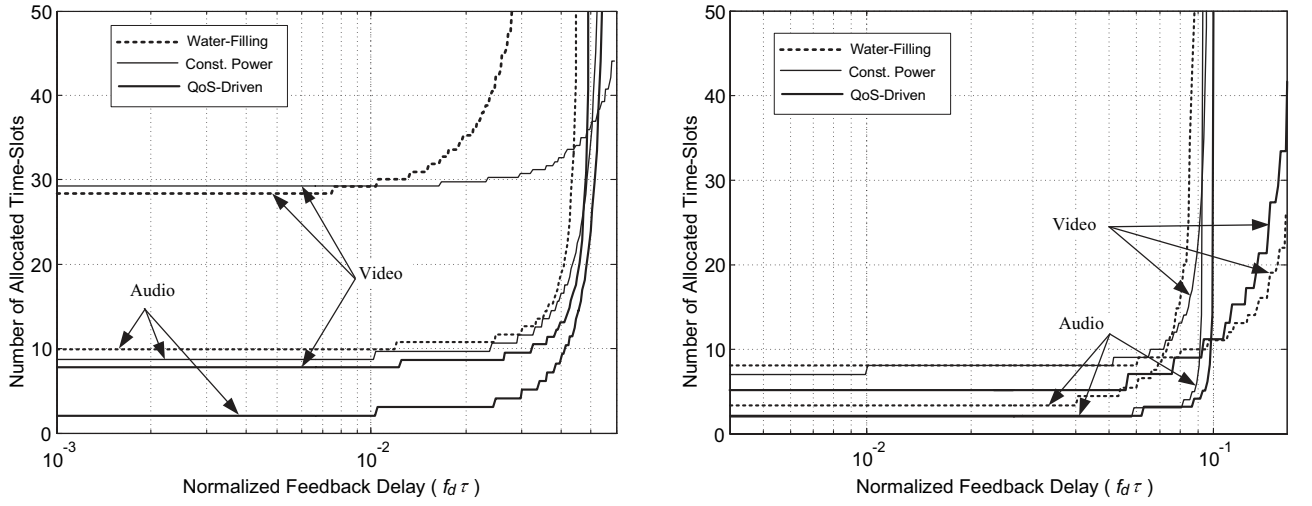
where $I_\nu(\cdot)$ denotes the modified Bessel function of the first kind with order ν and ρ represents the correlation coefficient between $\widehat{\gamma}$ and γ, which is given by $\rho = J_0^2(2\pi f_d\tau)$ [19] with $J_0(\cdot)$ denoting the zero-th-order Bessel function of the first kind. Omitting the tedious derivations for lack of space, we obtain $\mathrm{BER}_n(\gamma)$ in Eq. (32) as a closed-form as follows:

$$\begin{aligned}\mathrm{BER}_n(\gamma) &= 0.2\left(\frac{m\gamma}{m\gamma-(1-\rho)\overline{\gamma}\log(5\mathrm{P'_{tgt}})}\right)^m\\ &\quad\cdot\exp\left(\frac{m\rho\log(5\mathrm{P'_{tgt}})\gamma}{m\gamma-(1-\rho)\overline{\gamma}\log(5\mathrm{P'_{tgt}})}\right). \qquad (34)\end{aligned}$$

Averaging Eq. (34) with respect to the pdf $p_\Gamma(\gamma)$ of γ given by Eq. (1), we can express the average BER, denoted by $\overline{\mathrm{BER}}_n$, when γ falls into the nth mode, as follows:

$$\overline{\mathrm{BER}}_n = \frac{1}{\pi_n}\int_{\Gamma_n}^{\Gamma_{n+1}} \mathrm{BER}_n(\gamma)p_\Gamma(\gamma)d\gamma \qquad (35)$$

where π_n and Γ_n are given by Eqs. (3) and (6), respectively. It is hard to find the closed-form expression for Eq. (35).



(a) $m = 1$. (b) $m = 5$.

Fig. 9. The impact of CSI feedback delay on the time-slot allocation. The average SNR is set to $\overline{\gamma} = 10$ dB.

However, it can be solved by a single finite-integral as

$$\begin{aligned}\overline{\mathrm{BER}}_n &= \frac{0.2(1-\rho)^m[\log(5\mathrm{P}'_{\mathrm{tgt}})]^m}{\pi_n\Gamma(m)} \\ &\quad\cdot\int_{x_n}^{x_{n+1}}\exp\left(\frac{\log(5\mathrm{P}'_{\mathrm{tgt}})x(1-\rho x)}{1-x}\right) \\ &\quad\cdot\frac{x^{2m-1}}{(1-x)^{m+1}}dx \end{aligned} \tag{36}$$

where $x_n = m\Gamma_n/\left[m\Gamma_n - (1-\rho)\overline{\gamma}\log(5\mathrm{P}'_{\mathrm{tgt}})\right]$ and $x_N = 1$. The numerical searching procedure is used to search for the new target BER $\mathrm{P}'_{\mathrm{tgt}}$ such that the actual BER after delay satisfies

$$\mathrm{P}_{\mathrm{tgt}} = \frac{\sum_{n=1}^{N-1} n\pi_n\overline{\mathrm{BER}}_n}{\sum_{n=1}^{N-1} n\pi_n}. \tag{37}$$

Once the new target BER $\mathrm{P}'_{\mathrm{tgt}}$ is obtained, we can find the new boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ and thus reconstruct the FSMC of the service-process. Then, the resource-allocation algorithms can be re-executed based on the new FSMC. For water-filling power control, the procedure is the similar, but omitted for lack of space.

B. Constant Power-Control

Based on the similar approach to Section VI-A, we can show that the average BER for the nth mode can be derived as

$$\begin{aligned}\overline{\mathrm{BER}}_n &= \frac{1}{\pi_n}\int_{\Gamma_n}^{\Gamma_{n+1}}\mathrm{BER}_n(\gamma)p_\Gamma(\gamma)d\gamma \\ &= \frac{0.2}{\pi_n\Gamma(m)}\left(\frac{m}{b'_n}\right)^m \\ &\quad\cdot\left[\Gamma\left(m,\frac{b'_n\Gamma_n}{\overline{\gamma}}\right) - \Gamma\left(m,\frac{b'_n\Gamma_{n+1}}{\overline{\gamma}}\right)\right] \end{aligned} \tag{38}$$

where Γ_n is given by Eq. (9), $b'_n = m(\overline{\gamma}g_n + m)/\zeta_n$, and $\zeta_n = m + (1-\rho)\overline{\gamma}g_n$. The searching procedure is also to find the new boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ such that Eq. (37) is satisfied. Also, after the boundary points $\{\Gamma_n\}_{n=1}^{N-1}$ are determined, we can reconstruct the FSMC of the service-process and then we can re-execute the resource-allocation algorithms based on the new FSMC.

C. Numerical and Simulation Results

The above analyses are verified by the numerical and simulation results. In Fig. 9, we investigate the impact of feedback delay on time-slot allocations. We can see from Fig. 9 that the time-slot allocation results remain unchanged when the normalized feedback delay is below certain threshold. When $f_d\tau$ further increases, the number $L^{(k)}$ starts increasing in order to maintain the same statistical QoS requirements. From Fig. 9, we know that for all power-control policies, the better quality channel ($m = 5$) can tolerant larger feedback delay than the Rayleigh fading channel ($m = 1$). Specifically, the Rayleigh channel can only tolerant feedback delay with $f_d\tau \leq 0.01$, while the channel with $m = 5$ can tolerant the delay $f_d\tau \geq 0.04$. Note that in our system, we have $T_f \times f_d = 0.03$, implying that the channel with $m = 5$ can tolerant the feedback delay which is even larger than one frame's time duration. Thus, the proposed scheme provides sufficient robustness to the system in an in-door mobile environment, e.g., the widely used WLAN.

VII. Conclusions

We proposed and analyzed a cross-layer-model based adaptive resource-allocation scheme for diverse QoS guarantees over downlink mobile wireless networks. Our scheme jointly allocates power-levels and time-slots for real-time users to guarantee the diverse statistical delay-bound QoS requirements. We developed the admission-control and power/time-slot allocation algorithms by extending the effective capacity method. We also studied the impact of adaptive power control and CSI feedback delay at physical-layer on the QoS provisioning performance. Compared to the conventional water-filling and constant power approach, our proposed QoS-driven power adaptation shows significant advantages. The joint power/time-slot allocation scheme can significantly reduce

the transmit power, or equivalently, increase the admission region. Also, in an in-door mobile environment, our proposed algorithm is shown to be robust to the CSI feedback delay.

REFERENCES

[1] 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.

[2] S. Choi and K. G. Shin, "An uplink CDMA system architecture with diverse QoS guarantees for heterogeneous traffic," *IEEE/ACM Trans. Networking*, vol. 7, no. 5, pp. 616–628, Oct. 1999.

[3] X. Zhang and J. Tang, "QoS-driven asynchronous uplink subchannel allocation algorithms for space-time OFDM-CDMA systems in wireless networks," *ACM/Kluwer J. Wireless Networks*, vol. 12, no. 5, pp. 412–425, May 2006.

[4] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," *IEEE Commun. Mag.*, vol. 41, no. 10, pp. 74–80, Oct. 2003.

[5] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Commun.*, vol. 9, no. 5, pp. 76–83, Oct. 2002.

[6] J. Razavilar, K. J. R. Liu, and S. I. Marcus, "Jointly optimized bit-rate/delay control policy for wireless packet networks with fading channels," *IEEE Trans. Commun*., vol. 50, no. 3, pp. 484–494, Mar. 2002.

[7] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks," *IEEE J. Sel. Areas Commun*., to appear.

[8] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer modeling of adaptive wireless link for QoS support in multimedia networks," *ACM/Kluwer J. Wireless Networks*, to appear.

[9] D. Wu, "Providing Quality-of-Service Guarantees in Wireless Networks," Ph.D. Dissertation, Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, Aug., 2003.

[10] 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, July 2003.

[11] 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, Sept. 2004.

[12] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Trans. Veh. Technol.*, vol. 54, no. 3, pp. 1198–1206, May 2005.

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

[14] C. Courcoubetis and R. Weber, "Effective bandwidth for stationary sources," *Probability in Engineering and Information Sciences*, vol. 9, no. 2, pp. 285 – 294, 1995.

[15] R. Knopp and P. A. Humblet,"Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC 1995*, pp. 331–335, June 1995.

[16] J. Tang and X. Zhang, "Cross-layer modeling for quality of service guarantees over wireless links," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 12, pp. 4504–4512, Dec. 2007.

[17] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Trans. on Wireless Commun.*, vol. 6, no. 8, pp. 3058–3068, Aug. 2007.

[18] P. Viswanath, D. N. C. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, no. 6, 2002, pp 1277-1294.

[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.

[20] 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.

[21] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun*., vol. 36, pp. 834–843, 1988.

[22] 3GPP TS 25.101, V. 5.2.0. UE Radio Transmission and Reception (FDD, release 5), 2002.

[23] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd Ed., Athena Scientific, 2000.

[24] 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. Sel. Areas Commun.*, vol. 24, no. 1, pp. 104–112, Jan. 2006.

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.

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 was 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. His current research interests are in the areas of wireless networks, Internet protocols, communications theory, signal processing, information theory, multimedia, and stochastic control systems.

Prof. Zhang received the U.S. National Science Foundation CAREER Award in 2004 for his research in the areas of mobile wireless and multicast networking and systems. He received the Best Paper Award in the IEEE Globecom 2007. He also received the TEES Select Young Faculty Award for Excellence in Research Performance from the Dwight Look College of Engineering at Texas A&M University, College Station, in 2006. He is currently serving as an Editor for the *IEEE Transactions on Wireless Communications*, an Associate Editor for the *IEEE Transactions on Vehicular Technology*, an Associate Editor for the *IEEE Communications Letters*, an Editor for the *John Wiley's Wireless Communications and Mobile Computing Journal*, an Editor for the *John Wiley's Journal of Computer Systems, Networking, and Communications*, and an Associate Editor for the *John Wiley's Journal on Security and Communications Networks*. He is also serving as a Guest Editor for the *IEEE Journal on Selected Areas in Communications* for the Special Issue on "Resource Allocation for Wireless Video Traffic", a Guest Editor for the *IEEE Wireless Communications Magazine* for the Special Issues on "Next Generation of CDMA versus OFDMA for 4G Wireless Applications", a Guest Editor for the *Wiley's Journal on Wireless Communications and Mobile Computing* for the Special Issue on "Next Generation Wireless Communications and Mobile Computing Networks". He has frequently served as the Panelist on the U.S. National Science Foundation Research-Proposal Review Panels. He is currently serving as a Program Co-Chair for the IEEE INFOCOM 2009 Mini-Conference. He is also serving as the Symposium Co-Chair for the IEEE ICC 2008 – Information and Network Security Symposium and the Symposium Co-Chair for the IEEE Globecom 2008 – Wireless Communications Symposium. He is serving or has served as the Symposium Chair for the IEEE International Cross-Layer Optimized Wireless Networks Symposium 2008, 2007, and 2006, the TPC Chair for the IEEE IWCMC 2008, 2007, and 2006. He is serving or has served as a Demo/Poster Co-Chair for INFOCOM 2008 and as a 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 QShine 2006 and the IEEE/ACM MSWiM 2007. 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, etc.

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