Cross-Layer Scheduling With Prescribed QoS Guarantees in Adaptive Wireless Networks

Qingwen Liu, *Student Member, IEEE*, Shengli Zhou, *Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Providing *guaranteed* quality-of-service (QoS) for multimedia applications over wireless *fading* channels is challenging. To this end, we develop a cross-layer design for multiuser scheduling at the data link layer, with each user employing adaptive modulation and coding (AMC) at the physical layer. By classifying users into: QoS-guaranteed and best-effort users, the proposed scheduler enables *prescribed* QoS guarantees and *efficient* bandwidth utilization simultaneously. Furthermore, our cross-layer scheduler enjoys low-complexity implementation and analysis, provides service isolation and scalability, decouples delay from dynamically-scheduled bandwidth, and is backward compatible with existing separate-layer designs. Accuracy of the performance analysis is verified by simulations and pertinent robustness issues are briefly discussed. Numerical examples illustrate the steady-state statistical performance for a single and multiple users, as well as the asymptotic behavior for a large number of users.**

***Index Terms*—Adaptive modulation and coding (AMC), admission control, cross-layer design, quality-of-service (QoS), scheduling, wireless networks.**

I. Introduction

QUALITY-OF-SERVICE (QoS) metrics of a connection (flow or session) include data throughput, packet error/loss rate, and delay performance. Usually, multimedia applications can be classified in two categories: QoS-guaranteed and best-effort[1] ones [20]. The first category includes voice (e.g., VoIP), video/audio streaming, video/audio telephony, and conferencing; while applications such as web-browsing, e-mail, and file transfer protocol (FTP) belong to the second category.

For QoS guarantees in high-rate multimedia applications, the scarcity of transmission capacity, multipath fading and Doppler effects are common challenges to most communication networks, military or civilian, when mobile devices communicate a wide range of information over wireless links. The "bottleneck" common to both networks is the wireless link, not only because wireless resources (bandwidth and power) are more scarce and expensive than their wired counterparts, but also because the overall system performance degrades markedly due to time- and frequency-dispersive fading effects introduced by the wireless air interface. Unlike wired networks, even if large bandwidth is allocated to a certain connection, the loss and delay requirements may not be satisfied when the wireless channel experiences deep fades. Powerful forward error correction (FEC) coding or automatic-repeat-request (ARQ) protocols can reduce the loss rate, at the expense of increased bandwidth and delay [13]. On the other hand, allocating a fixed amount of bandwidth to each user may not be as efficient, because the queues may be empty from time to time due to the dynamic nature of the traffic [14]. These considerations testify to the difficulty in simultaneously *guaranteeing* QoS and utilizing resources *efficiently*.

Scheduling plays an important role in QoS provision. Although many traffic scheduling algorithms are available for wireline networks [29], they cannot be directly applied to wireless networks because of the fundamental differences between the two [9]. An overview of scheduling techniques for wireless networking can be found in [9], where a number of desirable properties have been summarized and many classes of schedulers have been compared on the basis of these properties. A challenge to scheduler designs is *predicting* all three aspects of QoS, namely, throughput, loss, and delay. For example, various scheduler designs can optimally utilize resources under certain metrics; however, the induced QoS can not be prescribed prior to scheduling, which usually results in underdesigning with respect to certain QoS aspects and overdesigning with respect to others. This motivates the design and performance evaluation of schedulers guaranteeing *prescribed* QoS with *efficient* resource utilization over wireless *fading* links [9].

Efficient bandwidth utilization for a prescribed error performance at the physical layer can be accomplished with adaptive modulation and coding (AMC) schemes, that match transmission parameters to the wireless channel conditions; see, e.g., [1], [3]–[5], [7], [11], [12], [18], and references therein. However, most existing AMC designs are tailored for the physical layer. Their impact on, and interaction with, higher protocol layers remain largely unexplored, e.g., the impact of AMC on delay performance of real-time applications.

In this paper, we consider multiuser scheduling at the medium access control (MAC) sublayer of the data link layer, where each user employs AMC at the physical layer (Section II). We develop a cross-layer scheduler design which accounts for both

Manuscript received April 1, 2004; revised December 28, 2004. The work of Q. Liu and G. B. Giannakis was supported in part by the Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The work of S. Zhou was supported in part by University of Connecticut Research Foundation internal Grant 445157. This paper was presented in part at the GLOBECOM Conference, Dallas, TX, November 29–December 3, 2004.

Q. Liu and G. B. Giannakis are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: qliu@ece.umn.edu; georgios@ece.umn.edu).

S. Zhou is with the Department of Electrical and Computer Engineering, University of Connecticut, Storrs, CT 06269 USA (e-mail: shengli@engr.uconn.edu).

Digital Object Identifier 10.1109/JSAC.2005.845430

[1]The service for a connection is considered to be "best-effort" when no guarantees on delay and throughput are provided; nearly error-free reception of information is typically required for such services [9], [20].

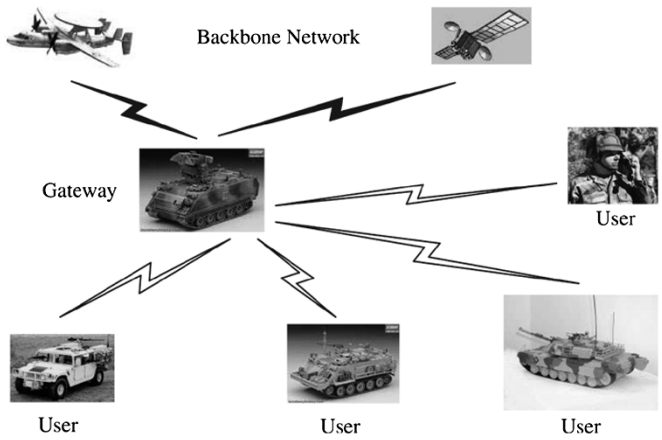


Fig. 1. System diagram.

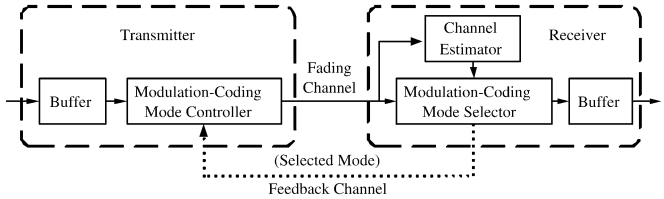


Fig. 2. Wireless link from the gateway to each user.

TABLE I
TRANSMISSION MODES WITH CONVOLUTIONALLY CODED MODULATION

Mode n	1	2	3	4	5
Modulation	BPSK	QPSK	QPSK	16QAM	64QAM
Coding Rate R_c	1/2	1/2	3/4	3/4	3/4
R_n	0.5	1.0	1.5	3.0	4.5

(The generator polynomial of the mother code is $g = [133, 171]$.)

channel variations and the status of users' queues, and includes admission control and scheduling policies for multiple users with different QoS requirements (Section III). We also analyze QoS and optimize the bandwidth allocation (Section IV). We summarize the desirable features of our scheduler design and derive its steady-state statistical characteristics (Section V). Finally, we verify its performance via simulations, discuss a pertinent robustness issue and illustrate steady-state statistical performance for single and multiple users, as well as the asymptotic behavior with a large number of users (Section VI).

Our scheduler design is *distinct* from most existing scheduling algorithms, because: 1) it guarantees the prescribed QoS and 2) combines scheduling with AMC, while at the same time it accounts for the wireless fading channel and queueing effects across layers.

II. MODELING PRELIMINARIES

A. System Description

Let us consider a military network which must provide high rate and high QoS communications of command and control information as detailed in [25]. The branch units (users), e.g., frontline force, artillery bastion, field hospital, field airport, etc., communicate to the headquarters through a gateway located at the frontline command. All the information collected from the frontline command is sent to the headquarters through a high-speed backbone. Fig. 1 depicts the system under consideration, where multiple users (sessions) are connected to the gateway over wireless channels, using time-division multiplexing/time-division multiple-access (TDM/TDMA). We focus on the downlink here, although our results can be extended to the uplink as well.

The wireless link from the gateway to each user is depicted in Fig. 2. A finite-length buffer (queue) is implemented at the gateway for each user, and operates in a first-in–first-out (FIFO) mode. The AMC controller follows the queue at the gateway (transmitter), and the AMC selector is implemented at each user (receiver).

At the *data link layer*, the processing unit is a packet comprising multiple information bits. We assume that the queue has finite-length (capacity) of K packets per user. The customers of the queue are packets, served by the AMC module at the physical layer.

At the *physical layer*, the processing unit is a frame consisting of multiple transmitted symbols. We assume that multiple transmission modes are available to each user, with each mode representing a pair of a specific modulation format and a FEC code, as in the HIPERLAN/2 and the IEEE 802.11a standards. Based on channel estimates obtained at the receiver, the AMC selector determines the modulation-coding pair (mode), which is sent back to the transmitter through a feedback channel, for the AMC controller to update the transmission mode. Coherent demodulation and maximum-likelihood (ML) decoding are employed at the receiver. The decoded bit streams are mapped to packets, which are pushed upward to the data link layer.

We consider the following group of transmission modes.

TM Convolutionally coded M_n-ary rectangular/square QAM, adopted from the HIPERLAN/2, or, the IEEE 802.11a standards [7], listed in Table I, where the transmission rate R_n (bits/symbol) is in ascending order with the index n [13]. Although we focus on TM in this paper, other transmission modes can be similarly constructed [1], [3], [4], [12].

We detail the packet and frame structures, as in Fig. 3.

1) At the *data link layer*, each packet contains a fixed number of bits (N_b), which include packet header, payload, and cyclic redundancy check (CRC) bits. After modulation and coding with mode n of rate R_n at the gateway, each packet is mapped to a symbol-block containing N_b/R_n symbols.
2) At the *physical layer*, the data are transmitted frame by frame through the wireless link, where each frame contains a fixed number of symbols (N_s). Given a fixed symbol rate, the frame duration (T_f seconds) is constant, and represents the *time-unit* throughout this paper. With TDM, each frame is divided into $N_c + N_d$ *time slots*, where for convenience we let each time-slot contain a fixed number of N_b/R_1 symbols. As a result, each time slot can transmit exactly R_n/R_1 packets with transmission mode n. For the TM in particular, one time-slot can accommodate $R_1/R_1 = 1$ packet with mode $n = 1$, $R_2/R_1 = 2$ packets with mode $n = 2$ and so on. The N_c

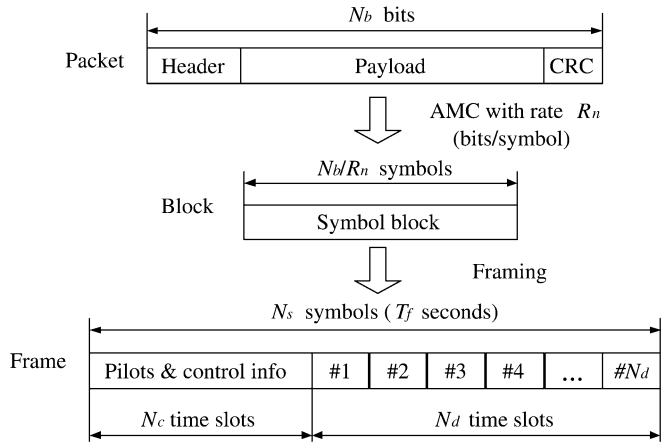


Fig. 3. Processing units at each layer.

time slots contain control information and pilots. The N_d time slots convey data, which are scheduled to different users with TDMA dynamically. Each user is allocated a certain number of time slots during each frame. The scheduler design is the main focus of this paper and will be addressed in Section III.

We next list our operating assumptions.

A1) For each user, the channel is frequency flat and remains invariant per frame, but is allowed to vary from frame to frame. This corresponds to a block fading model, which is suitable for slowly varying wireless channels [16]. Thus, AMC is adjusted on a frame-by-frame basis.

A2) Perfect channel state information (CSI) is available at the receiver relying on training-based channel estimation. The corresponding mode selection is fed back to the transmitter without error and latency, as in [5].

The assumption that the feedback channel is error free could be (at least approximately) satisfied by using heavily coded feedback. On the other hand, the feedback latency could be compensated by long range channel prediction; see, e.g., [8], [10], and references therein.

A3) If the queue is full, arriving packets will be dropped, so that the overflow content is lost.

This can be afforded by the user datagram protocol (UDP) for instance, where retransmission is usually not provided [30].

A4) Error detection based on CRC is perfect, provided that sufficiently reliable error detection CRC codes are used [17].

A5) If a packet is received incorrectly at the receiver after error detection, we drop it and declare packet loss.

Assumption A5 is reasonable and can be afforded by UDP-based video transmissions, because the underlying bit streams represent highly correlated contents.

B. Channel Model Induced by AMC

For flat fading channels adhering to A1, the channel quality can be captured by a single parameter, namely the received signal-to-noise ratio (SNR) γ. Since the channel varies from frame to frame, we adopt the general Nakagami-m model, which encompasses a large class of fading channels [21]. The received SNR γ per frame is, thus, a random variable with a Gamma probability density function

$$p_\gamma(\gamma) = \frac{m^m \gamma^{m-1}}{\bar{\gamma}^m \Gamma(m)} \exp\left(-\frac{m\gamma}{\bar{\gamma}}\right) \tag{1}$$

where $\bar{\gamma} := E\{\gamma\}$ is the average received SNR, $\Gamma(m) := \int_0^\infty t^{m-1}\exp(-t)\,dt$ is the Gamma function, and m is the Nakagami fading parameter ($m \geq 1/2$).

Each user relies on AMC at the physical layer. The objective of AMC is to maximize the data rate by adjusting transmission parameters to channel variations, while maintaining a prescribed packet error rate P_0. Let N denote the total number of transmission modes available ($N = 5$ for TM). As in [5], we assume constant power transmission, and partition the entire SNR range into $N+1$ nonoverlapping consecutive intervals, with boundary points denoted as $\{\gamma_n\}_{n=0}^{N+1}$. In this case

$$\text{mode } n \text{ is chosen}, \quad \text{when } \gamma \in [\gamma_n, \gamma_{n+1}). \tag{2}$$

To avoid deep channel fades, no data are sent when $\gamma_0 \leq \gamma < \gamma_1$, which corresponds to the mode $n = 0$ with rate $R_0 = 0$ (bits/symbol). The design objective of AMC is to determine the boundary points $\{\gamma_n\}_{n=0}^{N+1}$. Given P_0, $\bar{\gamma}$, and m, we have developed a threshold searching algorithm in [14], which determines $\{\gamma_n\}_{n=0}^{N+1}$ so that the average packet error rate for each mode is exactly P_0.

The SNR region $[\gamma_n, \gamma_{n+1})$ corresponding to transmission mode n constitutes the channel state indexed by n. To describe the transition of these channel states, we rely on a finite-state Markov chain (FSMC) model, which is developed in [14]. The state transition matrix of such FSMC is

$$\mathbf{P}_c = [P_{l,n}]_{(N+1)\times(N+1)} \tag{3}$$

which depends on the statistical channel parameters: average received SNR $\bar{\gamma}$, Nakagami fading parameter m and mobility-induced Doppler spread f_d [14].

Although we adopt the channel transition matrix in (3) for the Nakagami-fading channel, the ensuing results apply also to more general channel transition matrices including those described in [27].

III. Scheduler Design

In this section, we develop a multiuser scheduler at the medium access control (MAC) sublayer of the data link layer, with each user adopting AMC at the physical layer and the channel modeled as an FSMC.

In general, it is difficult for a scheduler to achieve the following two objectives simultaneously: 1) guaranteed QoS per user and 2) efficient bandwidth utilization. Targeting a desirable tradeoff, we classify the users admitted by the gateway into two categories of service: QoS-guaranteed and best-effort users [9], [20]. The QoS-guaranteed users reserve a certain amount of bandwidth, while the best-effort users do not, as will become clear later on.

In order to allocate the right amount of bandwidth to each QoS-guaranteed user, the key is to determine the relationship

between QoS and the reserved bandwidth. This task will be pursued in Section IV-A, where we will express QoS as a function of the throughput, the packet loss rate and the average packet delay, given a certain number of time slots. We will then allocate the minimal required b_i^* time slots per frame, which will guarantee the prescribed QoS of user $i \in \mathcal{I}$, where $\mathcal{I}$ is the set of indexes for all QoS-guaranteed users admitted by the gateway. In order to guarantee QoS for all QoS-guaranteed users, we propose the following two-step admission control policy.

Admission Policy:

Step 1) When the QoS-guaranteed user k requests a certain QoS, the admission control module determines the minimal required b_k^* time slots per frame via the procedure that will be described in Section IV-B. If the inequality

$$b_k^* + \sum_{i \in \mathcal{I}} b_i^* \leq N_d \tag{4}$$

holds, then user k is admitted into the set $\mathcal{I}$; otherwise, user k is rejected.

Step 2) No bandwidth is reserved for the best-effort users and the corresponding admission control scheme is up to the designer's choice.

This admission control policy guarantees the *prescribed* QoS for all QoS-guaranteed users in the system, because b_i^* time slots are reserved for each user $i \in \mathcal{I}$, at any time. Thus, determining b_i^* is the key to ensuring *prescribed* QoS guarantees. However, due to the dynamic behavior of the queue and the channel, the b_i^* time slots may not be completely occupied by each user. For efficient bandwidth utilization, we will dynamically allocate the unused time slots to best-effort users, that are admitted to the system with no preassigned time slots, and expect only the best-effort service. We summarize our proposed scheduling policy as follows.

Scheduling Policy:

1) For each QoS-guaranteed user $i \in \mathcal{I}$, b_i^* time slots are allocated all the time.
2) The number of *actually scheduled* time slots $\tilde{b}_{t,i}$ for QoS-guaranteed user i at time t, depends on both the data availability in the queue and the channel transport capability according to

$$\begin{aligned} \tilde{b}_{t,i} &= f(U_{t-1,i}, C_{t,i}) \\ &:= \begin{cases} 0, & \text{if } C_{t,i} = 0 \\ b_i^*, & \text{if } U_{t-1,i} \geq C_{t,i}, \quad C_{t,i} \neq 0 \\ \lceil b_i^* U_{t-1,i}/C_{t,i} \rceil, & \text{if } U_{t-1,i} < C_{t,i}, \quad C_{t,i} \neq 0 \end{cases} \end{aligned} \tag{5}$$

where $\lceil x \rceil$ denotes the smallest integer not less than x; $U_{t-1,i}$ is the number of packets in the queue of user i at the end of time-unit $t-1$; and $C_{t,i} = b_i^* R_n / R_1$ is the maximal number of packets, which can be transported to user i, with R_n being determined by AMC based on the channel quality at time-unit t. It is clear that $\tilde{b}_{t,i} \leq b_i^*$.

3) The total number of unused time slots $N_d - \sum_{i \in \mathcal{I}} \tilde{b}_{t,i}$ is shared by best-effort users at time t. We notice that $N_d - \sum_{i \in \mathcal{I}} \tilde{b}_{t,i} \geq N_d - \sum_{i \in \mathcal{I}} b_i^* \geq 0$.

This proposed scheduler provides efficient bandwidth utilization: when $U_{t-1,i} \geq C_{t,i}$, the maximum possible bandwidth b_i^* is scheduled; however, when $U_{t-1,i} < C_{t,i}$, only the *needed* bandwidth is scheduled for each user $i \in \mathcal{I}$, with the remaining slots shared by best-effort users.

Remark 1: Scheduling of time slots shared by best-effort users is up to the designer's choice, and many existing scheduling algorithms are available to this effect [9]. For example, the weighted-fair-queueing (WFQ) scheduling policy may be adopted. The weight may depend on both channel quality and the number of packets in the queue of individual users. WFQ provides a balanced resource utilization between fairness and efficiency.

Remark 2: Low probability of detection (LPD) is desirable in tactical networks. It is possible to combine bandwidth scheduling for QoS support with power control to account for LPD. For example, the transmit power of an individual user can be negotiated in advance. With predetermined power per user, our current bandwidth scheduling algorithm can be used because the bandwidth for QoS support is assigned based on the received SNR per user, regardless of transmit power. Thus, both LPD and QoS concerns can be accommodated by such an extended design.

Notice that our scheduler depends on both the queue state at the data link layer, and the channel state at the physical layer; hence, it offers a *cross-layer scheduler design*. The next step in our design is to figure out how to choose b_i^* with a prescribed QoS requirement.

IV. Bandwidth Minimization With QoS Guarantees

We first derive the QoS of a certain user in terms of its throughput, packet loss rate and average delay, for a given bandwidth allocation. We then propose an algorithm for determining the minimal required bandwidth to guarantee the prescribed QoS. Without loss of generality, we will confine ourselves to a single QoS-guaranteed user as in [14], so that we omit the user subscript $i \in \mathcal{I}$, for notational simplicity; e.g., $C_{t,i}$ and $U_{t,i}$ will be replaced by C_t and U_t, respectively.

A. QoS Over Wireless Links

Our goal here is to model the queueing arrival process and service process, in order to derive the steady-state behavior of the queueing system, as in [14] and, thus, obtain the QoS metrics.

Let t index time units and A_t denote the number of packets arriving at time t. We assume that the process A_t is stationary with $E\{A_t\} = \lambda$, and is independent of the queue state, as well as the channel state. If, for example, A_t is Poisson distributed with parameter λ (packets/frame), then ([6], p. 164)

$$P(A_t = a) = \lambda^a e^{-\lambda}/a!, \quad a \in \mathcal{A} \tag{6}$$

where $E\{A_t\} = \lambda$ and $\mathcal{A} := \{0, 1, \ldots, \infty\}$.

Different from nonadaptive modulations, AMC dictates a dynamic, rather than deterministic, service process for the queue, capable of transmitting a variable number of packets per time unit (frame). Let C_t (packets/frame) denote this transmission capability, i.e., the number of packets that can be transmitted at time t. Corresponding to each transmission mode n, let c_n

(packets/frame) denote the number of packets transmitted with AMC mode n per time unit. We then have

$$C_t \in \mathcal{C}, \quad \mathcal{C} := \{c_n : c_n = bR_n/R_1, n = 0, \ldots, N\} \tag{7}$$

where b is the number of time slots reserved for this user. We term b as *bandwidth coefficient*. As specified by (7), the AMC module yields a queue server with a total of $N+1$ states $\{c_n\}_{n=0}^{N}$, with the service process C_t representing the evolution of server states. Since the AMC mode n is chosen when the channel enters the state n, we model the service process C_t as an FSMC with transition matrix given by (3).

Let $U_t, U_t \in \mathcal{U} := \{0, 1, \ldots, K\}$, denote the queue state and (U_{t-1}, C_t) denote the pair of queue and server states, whose variation is modeled as an augmented FSMC [14]. We have proved that the steady-state distribution of (U_{t-1}, C_t) exists and is unique; see [14, eq. (24)] for the calculation of the steady-state distribution denoted as

$$P(U = u, C = c) := \lim_{t\to\infty} P(U_{t-1} = u, C_t = c). \tag{8}$$

Based on the steady-state distribution $P(U = u, C = c)$, it becomes possible to evaluate the QoS for each user.

Let P_d denote the packet dropping (overflow or blocking) probability upon the queue. Based on $P(A_t = a)$ in (6) and $P(U = u, C = c)$ in (8), we can readily compute P_d, as illustrated in [14, eq. (31)]. A packet is correctly received by a user, only if it is not dropped from the queue (with probability $1 - P_d$) and is correctly received through the wireless channel (with probability $1 - P_0$). Hence, we can obtain the packet loss rate as [14, eq. (13)]

$$\xi = 1 - (1 - P_d)(1 - P_0) \tag{9}$$

and the throughput as [14, eq. (14)]

$$\eta = \lambda(1 - \xi). \tag{10}$$

Let us now derive the average delay τ. With the steady-state distribution $P(U = u, C = c)$ in (8), we can compute the average number of packets in the queue and in transmission as [15, eq. (21)]

$$N_{\text{wl}} = \sum_{u\in\mathcal{U}, c\in\mathcal{C}} uP(U = u, C = c) + \sum_{u\in\mathcal{U}, c\in\mathcal{C}} \min\{u, c\}P(U = u, C = c) \tag{11}$$

where the subscript "wl" stands for wireless link. Based on Little's Theorem [6], the average delay per packet through the wireless link can be calculated as ([15, eq. (22)]

$$\tau = \frac{N_{\text{wl}}}{\lambda(1 - P_d)}. \tag{12}$$

In summary, given the bandwidth coefficient b, target packet error rate P_0, Doppler spread f_d, average SNR $\bar{\gamma}$, Nakagami parameter m, buffer length K, and data arrival rate λ, we can ascertain QoS (e.g., ξ, η, and τ) over the wireless link analytically.

Remark 3: For simplicity, we used the Poisson arrival process here. However, the same steps for QoS evaluation are directly applicable to other memoryless arrival processes. The analytical framework can be even extended to a Markov arrival process, which includes typical time-bursty traffic models; e.g., the on–off model. In this case, the state should be augmented to a triplet (U_{t-1}, C_t, A_t) in order to take into account the state transition of the arrival process A_t. Using the steady-state distribution of this state triplet, the corresponding QoS derivation can be carried out similarly.

Having established the relationship between the bandwidth coefficient b and the achieved QoS, we will next develop a procedure to find the minimal b guaranteeing the prescribed QoS.

B. Bandwidth Minimization With QoS Guarantees

The parameters $b, P_0, f_d, \bar{\gamma}, m, K, \lambda$, can be grouped in two categories: 1) the channel condition parameters, which include $f_d, \bar{\gamma}$ and m and the QoS parameter λ and 2) the resource management parameters b, P_0 and K. The parameters in the first category depend on the application, while the parameters in the second category are controllable via radio resource management (RRM).

If the QoS requirements of a certain user over the wireless link are requested by the signaling through the resource reservation protocol (RSVP) for instance, how can we determine, and even minimize, the required radio resources, namely, b and K, in order to guarantee the prescribed QoS? For simplicity, we fix the buffer-length K and focus on the effect of the bandwidth coefficient b on QoS. Let us illustrate the minimization with respect to b by the following numerical example.

Example 1: We consider $N_b = 1\,080$ (bits/packet), $T_f = 2$ (ms), packet arrival rate $\lambda = 2.5$ (packets/frame), average SNR $\bar{\gamma} = 12$ (dB), normalized Doppler frequency $f_d T_f = 0.01$, Nakagami fading parameter $m = 1.0$, and buffer length $K = 100$ (packets). The target packet error rate P_0 varies in the typical region $\mathcal{P} \subseteq [10^{-4}, 10^{-1}]$. The bandwidth coefficient b takes values in $\mathcal{D} = \{1, 2, 3, 4, 5, 6\}$. Suppose that the prescribed QoS requirements are: 1) $\eta \geq \eta_0 = 2.5$ (packets/frame) $= 1.35$ (Mbps); 2) $\xi \leq \xi_0 = 0.01$; and 3) $\tau \leq \tau_0 = 4T_f = 8$ (ms) (delay through the wireless link, not the end-to-end delay), which are typical QoS requirements for real-time video transmissions [2], [9].

We plot the packet loss rate ξ versus P_0 in Fig. 4; and the average delay τ versus P_0 in Fig. 5. From Fig. 5, we notice that τ decreases as P_0 increases, because a high value of P_0 leads to a high probability of selecting high rate modes in TM, i.e., high service rate for the queue. However, ξ depends on both P_0 and P_d [cf. (9)]; their joint effects on ξ are depicted by Fig. 4. From the shapes of these plots, we infer that P_0 dominates ξ when P_0 has large values, while P_d dominates ξ when P_0 has small values. This observation agrees with the intuition behind (9). Increasing the value of b, i.e., increasing bandwidth, results in increased service rate upon the queue, so that P_d and, thus, ξ decreases, as in Fig. 4, [14]; and τ decreases, as in Fig. 5.

From Fig. 4, $\xi \leq \xi_0$ can be guaranteed only for $\{b = 5, 3 \cdot 10^{-4} \leq P_0 \leq 4 \cdot 10^{-3}\}$ and $\{b = 6, 1 \cdot 10^{-4} \leq P_0 \leq 5 \cdot 10^{-3}\}$. The union of these sets is denoted as *the loss-guaranteed region*, and is depicted by the dashed lines in Fig. 6. From Fig. 5, $\tau \leq \tau_0$ can be guaranteed, only for $\{b = 3, P_0 \geq 1 \cdot 10^{-2}\}$, $\{b = 4, P_0 \geq 2 \cdot 10^{-3}\}$, $\{b = 5, P_0 \geq 5 \cdot 10^{-4}\}$, and $\{b = 6, P_0 \geq 2 \cdot 10^{-4}\}$. The union of these sets is denoted as *the delay-guaranteed region*, and is illustrated by the dotted lines in Fig. 6.

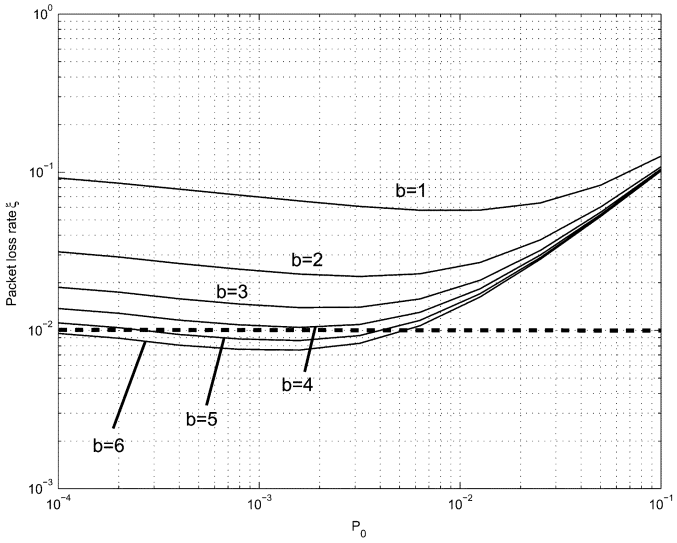


Fig. 4. Packet loss rate ξ versus target PER P_0.

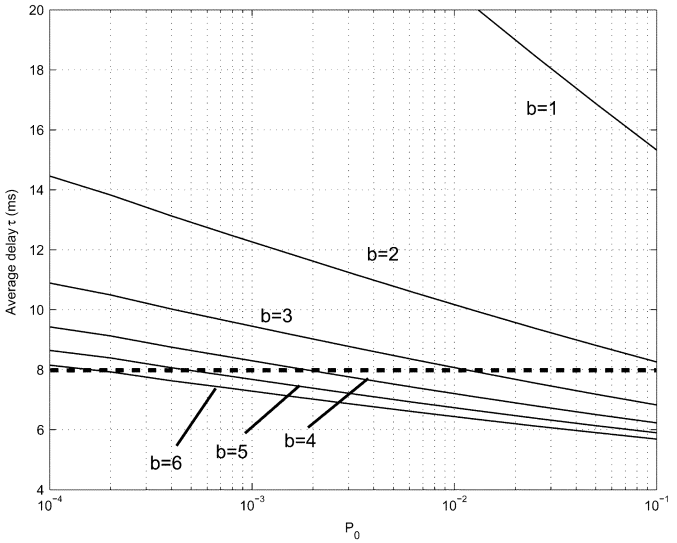


Fig. 5. Average delay τ versus target PER P_0.

In the loss-guaranteed region, $\xi \leq \xi_0 = 10^{-2}$ is guaranteed, i.e., more than 99% packets go through the wireless link, so that the throughput requirement is practically guaranteed in most applications. Thus, the prescribed QoS can be guaranteed, only if the RRM selects b and P_0 at the intersection of the loss-guaranteed and delay-guaranteed regions. This intersection is denoted as *the QoS-guaranteed region*: $\{b = 5, 5 \cdot 10^{-4} \leq P_0 \leq 4 \cdot 10^{-3}\}$ and $\{b = 6, 2 \cdot 10^{-4} \leq P_0 \leq 5 \cdot 10^{-3}\}$, and is indicated by the solid lines in Fig. 6. It is clear that selecting the minimal value of b, denoted as b^*, in the QoS-guaranteed region leads to the best bandwidth utilization. In this case, $b^* = 5$ is the minimal required bandwidth coefficient guaranteeing the prescribed QoS.

We notice that any P_0 associated with b^* in the QoS-guaranteed design, i.e., $P_0 \in \mathcal{P}^* := [5 \cdot 10^{-4}, 4 \cdot 10^{-3}]$ associated with $b^* = 5$, is a candidate to guarantee the prescribed QoS. However, the left end of $\mathcal{P}^*$, $P_0 = 5 \cdot 10^{-4}$, is the critical point

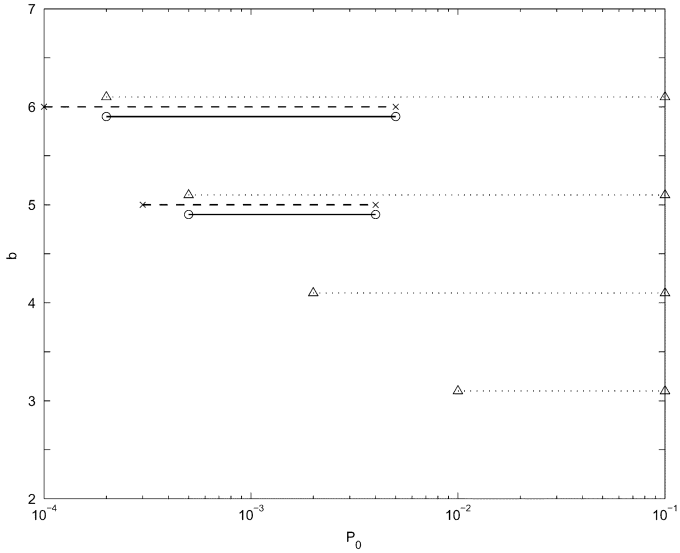


Fig. 6. Loss-guaranteed (dashed), delay-guaranteed (dotted), and QoS-guaranteed (solid) regions.

for $\tau \leq \tau_0$ or $\xi \leq \xi_0$; and the right end of $\mathcal{P}^*$, $P_0 = 4 \cdot 10^{-3}$, is the critical point for $\xi \leq \xi_0$. For robustness, it is desirable to let P_0 be far from both ends. We simply select the middle of $\mathcal{P}^*$ in the $\log_{10}$ domain, i.e., $P_0^* := 1.4 \cdot 10^{-3}$, which provides the least sensitivity to QoS requirements of both $\tau \leq \tau_0$ and $\xi \leq \xi_0$. We, thus, determine the optimal operating point as $(b^*, P_0^*) = (5, 1.4 \cdot 10^{-3})$. ■

In general, given the prescribed QoS requirements (η_0, ξ_0, τ_0), and the system parameters $(f_d, \bar{\gamma}, m, K$ and $\lambda = \eta_0/(1 - \xi_0) \approx \eta_0)$, we can obtain the minimal required bandwidth coefficient b^* and the corresponding target PER P_0^* analytically through the following steps.

Step 1) Initialization: Let $\mathcal{P}$ be the set of possible target PER values; and let $\mathcal{D}$ be the set of possible values of the bandwidth coefficient, in an increasing order, i.e., $b_k < b_{k+1}$, $\forall k$, $b_k \in \mathcal{D}$. Set the initial index $k = 1$.

Step 2) Based on (9) and (12), calculate

$$J = \sum_{P_0 \in \mathcal{P}} 1\{\xi(P_0, b_k) \leq \xi_0\} \times 1\{\tau(P_0, b_k) \leq \tau_0\} \qquad (13)$$

where $1\{\cdot\}$ stands for the indicator function.

Step 3) If $J > 0$, let $b^* = b_k$ and go to Step 4). If $J = 0$ and $k < |\mathcal{D}|$, set $k = k + 1$ and go to Step 2). Otherwise, no bandwidth coefficient in $\mathcal{D}$ can afford the prescribed QoS and stop searching.

Step 4) Let P_0^* be the middle of $\mathcal{P}^* := \{P_0 : \{\xi(P_0, b^*) \leq \xi_0\} \cap \{\tau(P_0, b^*) \leq \tau_0\}\}$ in the $\log_{10}$ domain.

The computational complexity mainly comes from computing the steady-state distribution π to obtain $\xi(P_0, b_k)$ and $\tau(P_0, b_k)$, that amounts to solving linear equations in queueing analysis [14]. The parameters (b^*, P_0^*) only need to be updated based on the slow-varying system parameters, namely, f_d, $\bar{\gamma}$, m, and K. On the other hand, look-up tables can be used for (b^*, P_0^*) in practice.

V. Features and Steady-State Performance

In this section, we will summarize the features of the proposed scheduler and analyze its steady-state statistical performance.

A. Desirable Features

The desirable attributes of a "good" scheduler have been summarized in [9] and [20]. Following the order in [9], the attributes of the proposed scheduler are as follows.

1) The QoS expressed in terms of throughput, packet loss rate and average delay is guaranteed for each QoS-guaranteed user $i \in \mathcal{I}$, via the optimal bandwidth reservation of b_i^* time slots. Therefore, service degradation is resolved by QoS-guarantees.
2) The bandwidth is utilized efficiently due to the dynamically scheduled $\tilde{b}_{t,i}$ time slots based on (5) for each QoS-guaranteed user $i \in \mathcal{I}$. For example, if the queue is empty $U_{t-1,i} = 0$ or the channel is in deep fading $C_{t,i} = 0$, then user i will not use any time slot since $\tilde{b}_{t,i} = 0$. On the other hand, the residual time slots $N_d - \sum_{i\in\mathcal{I}} \tilde{b}_{t,i}$ can be shared by best-effort users.
3) The implementation is simple, because the scheduled bandwidth for each user $i \in \mathcal{I}$, can be easily determined by $(U_{t-1,i}, C_{t,i})$ from (5); while the minimal required bandwidth coefficient b_i^* for the prescribed QoS can be analytically derived, as shown in Section IV, and can be easily realized by look-up tables.
4) The service for each QoS-guaranteed user is isolated from those of other users, through the scheduling policy of Section III; so that "misbehaving" users only affect their own QoS.
5) The energy (primarily transmission energy) can be saved, either when the channel is in deep fades, i.e., $\gamma < \gamma_1$, which corresponds to $C_{t,i} = 0$; or when the buffer is empty $U_{t-1,i} = 0$.
6) The delay and dynamically-scheduled bandwidth are decoupled for each QoS-guaranteed user $i \in \mathcal{I}$. Although the bandwidth is assigned to each user dynamically as in (5), each user's delay requirement is guaranteed because of the upper-bounded bandwidth reservation of b_i^* time slots as shown in Section IV-B.
7) Scalability is provided for the QoS-guaranteed users, when a new QoS-guaranteed user requests service from the system through the admission policy of Section III.
8) Our cross-layer scheduler design is compatible with separate-layer designs, because the scheduler can be implemented in existing systems by simply adding the corresponding functions for cross-layer information exchange, e.g., $C_{t,i}$ from the physical layer to the MAC sublayer.
9) The performance analysis of our scheduler design will turn out to be relatively simple. This is because the steady-state statistical characteristics of the bandwidth occupied by both QoS-guaranteed and best-effort users can be obtained analytically, as we show in the ensuing subsection.

B. Steady-State Statistical Performance

Given $(b_i^*, P_{0,i}^*)$ obtained as we explained in Section IV-B for each QoS-guaranteed user $i \in \mathcal{I}$, the actually scheduled time slots $\tilde{b}_{t,i}$ are determined by $(U_{t-1,i}, C_{t,i})$ dynamically as in (5). Therefore, the variation of $\tilde{b}_{t,i}$ is also an FSMC and its steady-state distribution, $P(\tilde{b}_i = j) := \lim_{t\to\infty} P(\tilde{b}_{t,i} = j)$, can be derived using (5) and (8) as

$$P(\tilde{b}_i = j) = \sum_{u\in\mathcal{U}, c\in\mathcal{C}} 1\{f(u,c) = j\} \times P(U = u, C = c) \tag{14}$$

where $0 \le j \le b_i^*$.

Supposing that channel fading, traffic and the QoS requirements of different users are independent, the corresponding queueing systems are independent, and so are the steady-state distributions $\tilde{b}_i := \lim_{t\to\infty} \tilde{b}_{i,t}$. Define the $\mathcal{Z}$-transform of $\tilde{b}_i$ as

$$\tilde{D}_i(z) := \sum_{j\ge 0} P(\tilde{b}_i = j) z^{-j}, \quad \forall\, i \in \mathcal{I} \tag{15}$$

and let the number of time slots occupied by all QoS-guaranteed users be denoted as $\tilde{b}_{\mathcal{I}} := \sum_{i\in\mathcal{I}} \tilde{b}_i$. Then, the $\mathcal{Z}$-transform of $\tilde{b}_{\mathcal{I}}$ can be derived from (15) as

$$\tilde{D}_{\mathcal{I}}(z) := \sum_{j\ge 0} P(\tilde{b}_{\mathcal{I}} = j) z^{-j} = \prod_{i\in\mathcal{I}} \tilde{D}_i(z). \tag{16}$$

Equation (16) allows us to express the corresponding steady-state distribution of $\tilde{b}_{\mathcal{I}}$ via the inverse $\mathcal{Z}$-transform of $\tilde{D}_{\mathcal{I}}(z)$ as

$$\left\{ P(\tilde{b}_{\mathcal{I}} = j), 0 \le j \le \sum_{i\in\mathcal{I}} b_i^* \right\} = \mathcal{Z}^{-1}\{\tilde{D}_{\mathcal{I}}(z)\} \tag{17}$$

which implies that the probability of having j time slots occupied by all QoS-guaranteed users is the coefficient of z^{-j} in $\tilde{D}_{\mathcal{I}}(z)$. Correspondingly, the k th moment of $\tilde{b}_{\mathcal{I}}$ can be evaluated as

$$E\left\{\tilde{b}_{\mathcal{I}}^k\right\} = \sum_{j\ge 0} j^k P(\tilde{b}_{\mathcal{I}} = j). \tag{18}$$

The best-effort users share a total of $N_d - \tilde{b}_{\mathcal{I}}$ time slots. Based on the distribution of $b_{\mathcal{I}}$, the distribution and moments of $N_d - \tilde{b}_{\mathcal{I}}$ can be easily derived from (14)–(18).

The steady-state analysis characterizes statistically the bandwidth utilization of the QoS-guaranteed users. On the other hand, this information can be used to assist admission control and scheduling of best-effort users.

VI. Numerical Results

A. Simulation-Based Verification

To offer QoS guarantees, our scheduler design in Section III relies on analytical QoS computations. To validate our QoS derivation, suppose that the following three users have been admitted by the system. We assume the following common parameters for these users as: bandwidth coefficient $b = 1$, frame

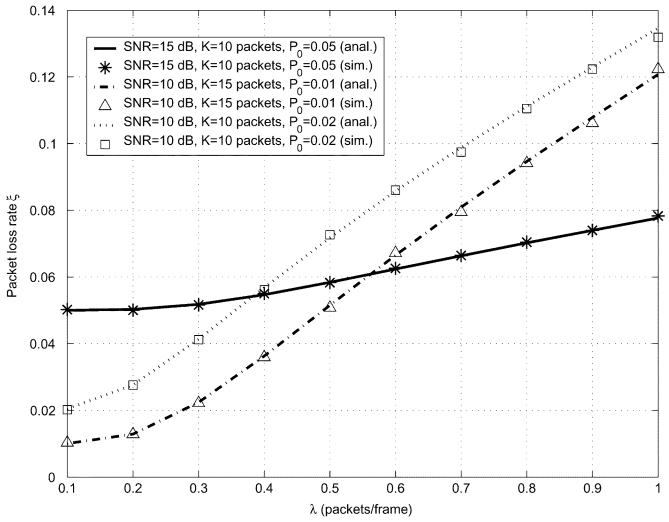


Fig. 7. Packet loss rate ξ versus packet arriving rate λ.

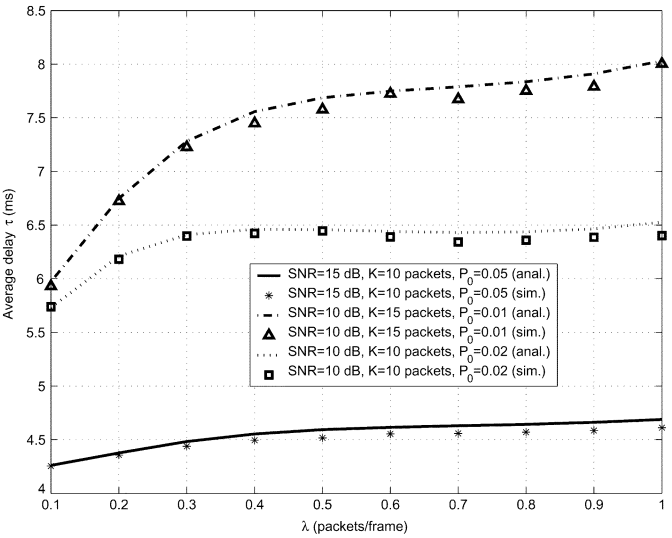


Fig. 8. Average delay τ versus packet arriving rate λ.

length $T_f = 2$ (ms), Nakagami parameter $m = 1$ and Doppler frequency $f_d = 10$ (Hz), i.e., $f_d T_f = 0.02$. The simulation setting for the first user is average received SNR $\bar{\gamma} = 15$ (dB), buffer size $K = 10$ (packets) and the prescribed PER for AMC is $P_0 = 0.05$; the setting for the second user is $\bar{\gamma} = 10$ (dB), $K = 15$ (packets) and $P_0 = 0.01$; and the setting for the third user is $\bar{\gamma} = 10$ (dB), $K = 10$ (packets) and $P_0 = 0.02$. The packet arrival rate λ (packets/frame) varies from 0.1 to 1.0 with increasing step 0.1. In each simulation, the system was run for a time period equivalent to 100 000 ms. Figs. 7 and 8, compare analytical versus simulation based results for the packet loss rate and average delay, respectively; the throughput can be derived from the packet loss rate directly based on (10). In these figures, the analytical "lines" match well the simulated "points," which verify that the prescribed QoS is guaranteed for each admitted user.

TABLE II
OPTIMAL BANDWIDTH COEFFICIENT b^* VERSUS $(\bar{\gamma}, m, f_d)$

$\bar{\gamma}$ (dB)	15	14	16	15	15	15	15
m	2.0	2.0	2.0	1.8	2.2	2.0	2.0
f_d (Hz)	10	10	10	10	10	5	15
b^* for S_1	2	2	2	2	2	2	2
b^* for S_2	2	3	2	3	2	2	2

B. Robustness to Parameter Mismatch

Since our scheduler design depends on channel statistics which must be estimated in practice, it is important to investigate the robustness of our design when there is mismatch between estimated and actual channel statistics. To this end, let us illustrate how the optimal bandwidth coefficient b^* is affected when mismatch is present between the estimated channel parameters $(\bar{\gamma}, m, f_d)$ and their actual values (notice that K and λ are of less concern, since those are readily available or can be negotiated as part of the admission contract).

With exact knowledge of $T_f = 2$ (ms), $\lambda = 3$ (packets/frame) and $K = 10$ (packets), suppose that a set of reference channel parameters for one user is $\bar{\gamma} = 15$ (dB), $m = 2$ and $f_d = 10$ (Hz), which can be treated as actual values. We consider two sets of prescribed QoS requirements $S_1 := \{\xi \le \xi_0 = 0.01 \text{ and } \tau \le \tau_0 = 4.2 \text{ (ms)}\}$ and $S_2 := \{\xi \le \xi_0 = 0.005 \text{ and } \tau \le \tau_0 = 4.2 \text{ (ms)}\}$, where the QoS requirement under S_2 is more stringent than that under S_1. For comparison, we vary only one entry of the triplet $(\bar{\gamma}, m, f_d)$ each time from the reference parameters, where the parameters we perturb can be viewed as the estimated values. We determine the corresponding optimal bandwidth coefficient b^* for S_1 and S_2, respectively, as shown in Table II.

From Table II, we notice that b^* remains the same under S_1 for different settings of $(\bar{\gamma}, m, f_d)$. However, b^* changes under S_2 for some of these settings. The fact that b^* is more robust for S_1 than S_2 means that less stringent QoS requirements lead to better robustness. In general, we observe that the robustness of the proposed scheduler depends on the accuracy in estimating statistical channel parameters *as well as* the prescribed QoS. Here, we only studied a simple example; the robustness issue of the proposed scheduler warrants further thorough investigation.

C. Steady-State Statistical Performance

We next depict the steady-state statistical performance that has been discussed in Section V-B. For convenience, we consider each QoS-guaranteed user $i \in \mathcal{I}$ having the same parameters as those in Example 1. We also assume that $\sum_{i\in\mathcal{I}} b_i^* \le N_d$, so that each user $i \in \mathcal{I}$ will operate with $(b_i^*, P_{0,i}^*) = (b^*, P_0^*)$ determined in Example 1.

Let $N_{\mathcal{I}} := |\mathcal{I}|$ denote the number of QoS-guaranteed users in the system. For $N_{\mathcal{I}} = 1, 20$ and based on (14)–(17), we plot the steady-state distribution of the number of time slots occupied by all QoS-guaranteed users $\{P(\tilde{b}_{\mathcal{I}} = j),\ 0 \le b_{\mathcal{I}} \le \sum_{i\in\mathcal{I}} b_i^*\}$ in Figs. 9 and 10, respectively. From (18), the corresponding means $\mu = E\{\tilde{b}_{\mathcal{I}}\}$ are 1.2386 and 24.7715, respectively; and the normalized standard

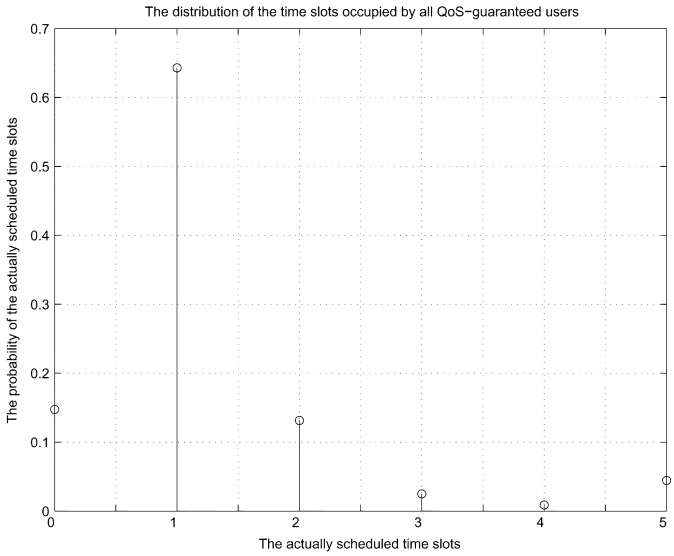


Fig. 9. Probability $P(\bar{b}_{\mathcal{I}} = j)$ versus j, $0 \le j \le \sum_{i\in\mathcal{I}} b^*$, for $N_{\mathcal{I}} = 1$.

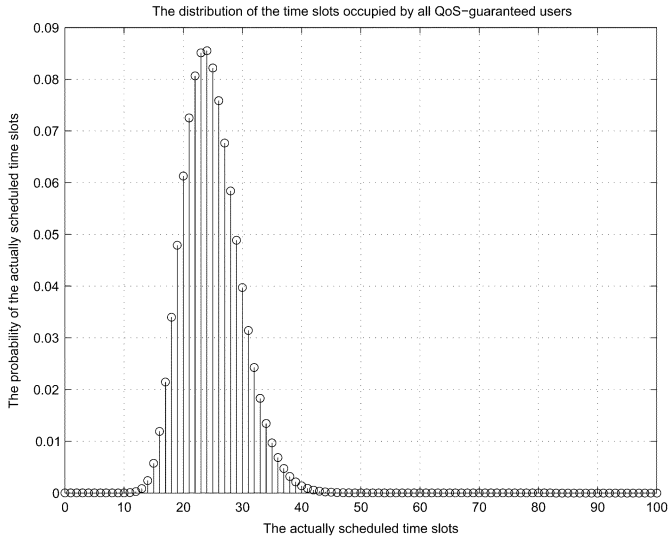


Fig. 10. Probability $P(\bar{b}_{\mathcal{I}} = j)$ versus j, $0 \le j \le \sum_{i\in\mathcal{I}} b^*$, for $N_{\mathcal{I}} = 20$.

deviations $\bar{\sigma} = \sigma/E\{\tilde{b}_{\mathcal{I}}\} = \sqrt{E\{\tilde{b}_{\mathcal{I}}^2\} - E^2\{\tilde{b}_{\mathcal{I}}\}}/E\{\tilde{b}_{\mathcal{I}}\}$ are 0.8519 and 0.1905, respectively.

For $N_{\mathcal{I}} = 1$, in order to guarantee the prescribed QoS of any given user, $b_i^* = 5$ time slots should be reserved at any time; however, the actually used time slots are only $E\{\tilde{b}_{\mathcal{I}}\} = 1.2386$ on average. Based on Fig. 9, the transmission power and bandwidth will be saved in $P(\tilde{b}_{\mathcal{I}} = 0) = 0.1473 \approx 15\%$ time units, when the buffer is empty ($U = 0$) or when deep fades occur ($C = 0$). We notice that $P(\tilde{b}_{\mathcal{I}} = j)$ is decreasing from $j = 1$ to $j = 4$; however, $P(\tilde{b}_{\mathcal{I}} = 5) = 0.0445 \approx 4.5\%$ takes a larger value than when $j = 3, 4$, which shows that the maximum possible bandwidth coefficient b^* indeed plays a role in guaranteeing the QoS of a single user.

From Fig. 10 with $N_{\mathcal{I}} = 20$, one can observe that $P(\tilde{b}_{\mathcal{I}} = j)$ decreases very fast and converges to 0 for large j values. Intuitively, it is very likely that some users with large bandwidth will be scheduled at the same time with other users having small bandwidth, which will result in a modest sum. The probability of having all users occupying a large amount of reserved b_i^* slots at the same time is very small. By the central limit theorem (CLT), the sum of $N_{\mathcal{I}}$ independent identically distributed (i.i.d.) random variables with mean μ and normalized standard deviation $\bar{\sigma}$, tends to be Gaussian distributed with mean $\mu N_{\mathcal{I}}$ and normalized standard deviation $\bar{\sigma}/\sqrt{N_{\mathcal{I}}}$ for large $N_{\mathcal{I}}$. Because of the decreasing normalized standard deviation $\bar{\sigma}/\sqrt{N_{\mathcal{I}}}$, the Gaussian "bell" becomes concentrated around the mean $\mu N_{\mathcal{I}}$, for large $N_{\mathcal{I}}$.

Recall that our scheduler meets the QoS requirement for each QoS-guaranteed user. The total number of QoS-guaranteed users is controlled by the admission control module. The plots in Fig. 10 bring out additional insights. Let P_{out} be the bandwidth-request outage probability of all QoS-guaranteed users, i.e., the probability that the sum of bandwidth requests of all users in $\mathcal{I}$ can not be provided, which may degrade the QoS for some of these QoS-guaranteed users. Notice that a practically acceptable P_{out} may allow more QoS-guaranteed users to be served with a reduced amount of total bandwidth. Specifically, if we define a constant j_0 satisfying

$$\sum_{j=j_0+1}^{\sum_{i\in\mathcal{I}} b_i^*} P(\tilde{b}_{\mathcal{I}} = j) < P_{\text{out}}$$

and

$$\sum_{j=j_0}^{\sum_{i\in\mathcal{I}} b_i^*} P(\tilde{b}_{\mathcal{I}} = j) \ge P_{\text{out}}$$

then reserving j_0 time slots provides QoS guarantee with outage probability P_{out}. For example, when $N_{\mathcal{I}} = 20$ as in Fig. 10, $P_{\text{out}} = 10^{-4}$ results in $j_0 = 47$, although $\sum_{i\in\mathcal{I}} b_i^* = 100$. This tolerance provides flexibility to admit more QoS-guaranteed users in the system, under the given resource N_d. How to analyze and utilize this behavior is an interesting future topic deserving further investigation.

VII. Concluding Remarks and Future Directions

In this paper, we developed a cross-layer scheduler design for multimedia applications in adaptive wireless networks, where the QoS-guaranteed multiuser scheduler at the MAC sublayer is coupled with adaptive modulation and coding (AMC) at the physical layer. We introduced novel admission and scheduling policies, and analyzed QoS of a wireless connection in terms of throughput, packet loss rate and average delay, given a certain reserved bandwidth. This analysis enabled us to determine the minimal required bandwidth per user. Our cross-layer scheduler guarantees prescribed QoS for admitted QoS-guaranteed users, achieves efficient bandwidth utilization, enjoys low-complexity implementation, isolates QoS-guaranteed service, decouples delay from dynamically-scheduled bandwidth, provides scalability, is compatible with existing separate-layer designs, and can afford simple performance analysis. Our performance analysis was verified by simulations and an example was tested to confirm robustness to channel parameter mismatch. Numerical examples illustrated the steady-state statistical performance, including the distribution, the mean and the normalized standard deviation of actually scheduled bandwidth, in

single- and multiuser scenarios. The importance of bandwidth reservation for QoS guarantees of a single user was pointed out. The asymptotic behavior of bandwidth allocation for multiple QoS-guaranteed users was also outlined.

Opportunistic scheduling (see [22], [26] and references therein) are suitable for best-effort services without QoS guarantees as pointed out in [20], while our proposed scheduling provides QoS-guaranteed service. These two kinds of scheduling algorithms have *complementary* value in integrated networks which include both QoS-guaranteed and best-effort services; and their interaction is a subject worthy of future study.

In this paper, we explored single-hop connections, which offer the building blocks and insights toward QoS support for end-to-end connections in multihop networks. Since the usable bandwidth may be shared by multiple end-to-end connections from a certain node to multiple nodes in multihop networks, it is necessary to adopt a scheduler that efficiently assigns bandwidth to support the QoS corresponding to different connections. From this viewpoint our scheduling design forms the building block for end-to-end QoS support in multihop networks, but further study is due.

Our proposed scheme relies on perfect channel state information (CSI) and traffic estimation. Although we briefly touched upon the possible mismatch between estimated and true channel statistics, it is more important to study thoroughly robustness with respect to the traffic model considered and imperfect instantaneous CSI in the AMC module; our preliminary work in this direction can be found in [31].

While average delay was included in our QoS metrics, bounded delay for a given outage probability is a direction deserving further investigation. Although fixed buffer-lengths were considered, joint resource allocation of bandwidth and buffer-lengths may lead to scheduling algorithms with further enhanced efficiency.

ACKNOWLEDGMENT

The authors would like to thank Dr. X. Wang, Postdoctoral Researcher with the SPINCOM Group, University of Minnesota, for insightful discussion and for providing a useful simulation code. They are also grateful to the Guest Editor and anonymous reviewers for their helpful suggestions which helped improve the presentation of this paper.

REFERENCES

[1] *Physical Layer Aspects of UTRA High Speed Downlink Packet Access (Release 4)*, 2001. 3GPP TR 25.848 V4.0.0.

[2] *Error Resilience in Real-Time Packet Multimedia Payloads*, 1999. 3GPP TSG-S4 Codec Working Group.

[3] *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, July 1999. 3GPP2 C.S0002-0 Version 1.0.

[4] *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broad-band Wireless Access Systems*, 2002. IEEE Standard 802.16 Working Group.

[5] M. S. Alouini and A. J. Goldsmith, "Adaptive modulation over Nakagami fading channels," *Kluwer J. Wireless Commun.*, vol. 13, no. 1–2, pp. 119–143, May 2000.

[6] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, 1992.

[7] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the HIPERLAN/2 and IEEE 802.11a wireless LAN standards," *IEEE Commun. Mag.*, vol. 40, no. 5, pp. 172–180, May 2002.

[8] A. Duel-Hallen, S. Hu, and H. Hallen, "Long-range prediction of fading signals," *IEEE Signal Process. Mag.*, vol. 17, no. 3, pp. 62–75, May 2000.

[9] 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.

[10] S. Falahati, A. Svensson, T. Ekman, and M. Sternad, "Adaptive modulation systems for predicted wireless channels," *IEEE Trans. Commun.*, vol. 52, no. 2, pp. 307–316, Feb. 2004.

[11] K. J. Hole, H. Holm, and G. E. Oien, "Adaptive multidimensional coded modulation over flat fading channels," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 7, pp. 1153–1158, Jul. 2000.

[12] J. Karaoğuz, "High-rate wireless personal area networks," *IEEE Commun. Mag.*, vol. 39, no. 12, pp. 96–102, Dec. 2001.

[13] 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, Sep. 2004.

[14] ——, "Queueing with adaptive modulation and coding over wireless links: Cross-layer analysis and design," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, May 2005.

[15] ——, "TCP performance in wireless access with adaptive modulation and coding," in *Proc. Int. Conf. Commun.*, vol. 7, Paris, France, Jun. 20–24, 2004, pp. 3989–3993.

[16] E. Malkamaki and H. Leib, "Performance of truncated type-II hybrid ARQ schemes with noisy feedback over block fading channels," *IEEE Trans. Commun.*, vol. 48, no. 9, pp. 1477–1487, Sep. 2000.

[17] H. Minn, M. Zeng, and V. K. Bhargava, "On ARQ scheme with adaptive error control," *IEEE Trans. Veh. Technol.*, vol. 50, no. 6, pp. 1426–1436, Nov. 2001.

[18] M. B. Pursley and J. M. Shea, "Adaptive nonuniform phase-shift-key modulation for multimedia traffic in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 8, pp. 1394–1407, Aug. 2000.

[19] 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.

[20] 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.

[21] G. L. Stüber, *Principles of Mobile Communication*, 2nd ed. Norwell, MA: Kluwer, 2001.

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

[23] 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.

[24] X. Wang and J. K. Tugnait, "Joint design of bandwidth distribution, truncated ARQ protocol and adaptive modulation and coding scheme for multiple delay sensitive traffics," presented at the 38th Conf. Inf. Sci. Syst., Princeton, NJ, Mar. 17–19, 2004.

[25] K. Wongthavarawat and A. Ganz, "IEEE 802.16 based last mile broadband wireless military networks with quality of service support," in *Proc. Military Commun. Conf.*, Oct. 13–16, 2003, pp. 779–784.

[26] M. Xiao, N. B. Shroff, and E. K. P. Chong, "A utility-based power-control scheme in wireless cellular systems," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 210–221, Apr. 2003.

[27] F. Xie, J. L. Hammond, and D. L. Noneaker, "Steady-state analysis of a split-connection scheme for Internet access through a wireless terminal," *IEEE/ACM Trans. Netw.*, vol. 12, no. 3, pp. 515–525, June 2004.

[28] M. D. Yacoub, J. E. V. Bautista, and L. G. de R. Guedes, "On higher order statistics of the Nakagami-m distribution," *IEEE Trans. Veh. Technol.*, vol. 48, no. 3, pp. 790–794, May 1999.

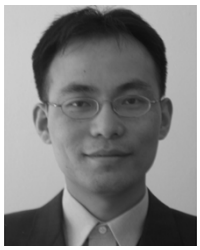
[29] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proc. IEEE*, vol. 83, no. 10, pp. 1374–1396, Oct. 1995.

[30] H. Zheng and J. Boyce, "An improved UDP protocol for video transmission over Internet-to-wireless networks," *IEEE Trans. Multimedia*, vol. 3, no. 3, pp. 356–365, Sep. 2001.

[31] S. Zhou and G. B. Giannakis, "How accurate channel prediction needs to be for transmit-beamforming with adaptive modulation in Rayleigh MIMO channels?," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1285–1294, Jul. 2004.

Qingwen Liu (S'04) received the B.S. degree in electrical engineering and information science from the University of Science and Technology of China (USTC), Hefei, in 2001 and the M.S. degree in electrical engineering from the University of Minnesota (UMN), Minneapolis, in 2003. He is currently working towards the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Minnesota (UMN), Minneapolis.

His research interests lie in the areas of communications, signal processing, and networking, with emphasis on cross-layer analysis and design, quality-of-service support for multimedia applications over wired-wireless networks, and resource allocation.



Shengli Zhou (M'03) received the B.S. and M.Sc. degrees from the University of Science and Technology of China (USTC), Hefei, in 1995 and 1998, respectively, both in electrical engineering and information science, and the Ph.D. degree in electrical engineering from the University of Minnesota (UMN), Minneapolis, in 2002.

He joined the Department of Electrical and Computer Engineering at the University of Connecticut, Storrs, in 2003. His research interests lie in the areas of communications and signal processing, including channel estimation and equalization, multiuser and multicarrier communications, space–time coding, adaptive modulation, and cross-layer designs.

Dr. Zhou has served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since February 2005.

Georgios B. Giannakis (S'84–M'86–SM'91–F'97) received the Diploma degree in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1981, the M.Sc. degree in electrical engineering, the M.Sc. degree in mathematics, and the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 1983, 1986, and 1986, respectively, where he was from September 1982 to July 1986.

After lecturing for one year at USC, he joined the University of Virginia in 1987, where he became a Professor of Electrical Engineering in 1997. Since 1999, he has been a Professor with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, where he now holds an ADC Chair in Wireless Telecommunications. His general interests span the areas of communications and signal processing, estimation and detection theory, time-series analysis, and system identification—subjects on which he has published more than 200 journal papers, 350 conference papers, and two edited books. Current research focuses on transmitter and receiver diversity techniques for single- and multiuser fading communication channels, complex-field and space–time coding, multicarrier, ultrawideband wireless communication systems, cross-layer designs, and distributed sensor networks.

Dr. Giannakis is the (co-) recipient of six paper awards from the IEEE Signal Processing Society (1992, 1998, 2000, 2001, 2003). He also received the Signal Processing Society's Technical Achievement Award in 2000. He served as Editor-in-Chief for the IEEE SIGNAL PROCESSING LETTERS, as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE SIGNAL PROCESSING LETTERS, as Secretary of the Signal Processing Conference Board, as Member of the Signal Processing Publications Board, as Member and Vice-Chair of the Statistical Signal and Array Processing Technical Committee, as Chair of the Signal Processing for Communications Technical Committee, and as a Member of the IEEE Fellows Election Committee. He has also served as a member of the IEEE Signal Processing Society's Board of Governors, the Editorial Board for the PROCEEDINGS OF THE IEEE, and the Steering Committee of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.