A Cross-Layer Scheduling Algorithm With QoS Support in Wireless Networks

Qingwen Liu, *Student Member, IEEE*, Xin Wang, *Member, IEEE*, and Georgios B. Giannakis, *Fellow, IEEE*

***Abstract*—Scheduling plays an important role in providing quality of service (QoS) support to multimedia communications in various kinds of wireless networks, including cellular networks, mobile *ad hoc* networks, and wireless sensor networks. The authors propose a scheduling algorithm at the medium access control (MAC) layer for multiple connections with diverse QoS requirements, where each connection employs adaptive modulation and coding (AMC) scheme at the physical (PHY) layer over wireless fading channels. Each connection is assigned a priority, which is updated dynamically based on its channel and service status; the connection with the highest priority is scheduled each time. The authors' scheduler provides diverse QoS guarantees, uses the wireless bandwidth efficiently, and enjoys flexibility, scalability, and low implementation complexity. Its performance is evaluated via simulations.**

***Index Terms*—Adaptive modulation and coding (AMC), cross-layer design, IEEE 802.16, quality of service (QoS), scheduling algorithm, wireless networks, Worldwide Interoperability for Microwave Access (WiMAX).**

I. Introduction

THE RAPID growth in demand for high-speed and high-quality multimedia communications is creating opportunities and challenges for next-generation wired–wireless network designs. Multimedia communications entail diverse quality of service (QoS) requirements for different applications including voice, data and real time, or streaming video/audio. Providing QoS-guaranteed services is necessary for future wireless networks, including cellular networks, mobile *ad hoc* networks, and wireless sensor networks, e.g., IEEE 802.16, IEEE 802.11, and IEEE 802.15 standard wireless networks. Such networks are envisioned to support multimedia services with different QoS requirements. However, the aforementioned standards define only QoS architecture and signaling, but do not specify the scheduling algorithm that will ultimately provide QoS support.

Scheduling plays an important role in QoS provision. Although many traffic scheduling algorithms are available for wireline networks [25], they cannot be directly applied to wireless networks because of the fundamental differences between the two [9], [11]. For example, traditional schedulers for wireline networks only consider traffic and queuing status; however, channel capacity in wireless networks is time varying due to multipath fading and Doppler effects. Even if large bandwidth is allocated to a certain connection, the prescribed delay or throughput performance may not be satisfied, and the allocated bandwidth is wasted when the wireless channel experiences deep fades. An overview of scheduling techniques for wireless networking can be found in [9], where a number of desirable features have been summarized, and many classes of schedulers have been compared on the basis of these features. To schedule wireless resources (such as bandwidth and power) efficiently for diverse QoS guarantees, the interactive queuing behavior induced by heterogenous traffic as well as the dynamic variation of wireless channel should be considered in scheduler design.

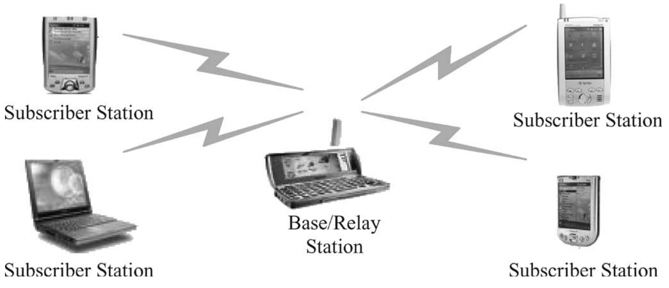

Fig. 1. Network topology.

In this paper, we introduce a priority-based scheduler at the medium access control (MAC) layer for multiple connections with diverse QoS requirements, where each connection employs adaptive modulation and coding (AMC) scheme at the physical (PHY) layer. We define a priority function (PRF) for each connection admitted in the system and update it dynamically depending on the wireless channel quality, QoS satisfaction, and service priority across layers. Thus, the connection with the highest priority is scheduled each time. Our scheduler provides prescribed QoS guarantees and utilizes the wireless bandwidth efficiently while enjoying low implementation complexity, flexibility, and scalability.

II. System Architecture

A. Network Configuration

Fig. 1 illustrates the wireless network topology under consideration. Multiple subscriber stations (SS) are connected to the base station (BS) or relay station over wireless channels, where multiple connections (sessions, flows) can be supported by each

Manuscript received September 12, 2005; revised December 28, 2005. This work was prepared through collaborative participation in the Communications and Networks Consortium sponsored by the U.S. Army Research Laboratory under the Collaborative Technology Alliance Program, Cooperative Agreement DAAD19-01-2-0011. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation thereon. The material in this paper was presented in part at the 2nd International Conference on QoS in Heterogeneous Wired/Wireless Networks (QShine), Orlando, FL, August 22–24, 2005. The review of this paper was coordinated by Prof. X. Shen.

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: qliu@ece.umn.edu; xinwest@ece.umn.edu; georgios@ece.umn.edu).

Digital Object Identifier 10.1109/TVT.2006.873832

0018-9545/$20.00 © 2006 IEEE

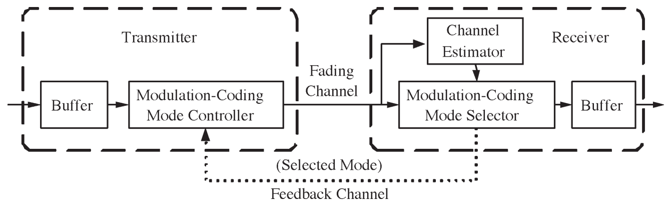

Fig. 2. Wireless link from BS to SS.

SS. This kind of star topology is not only applicable to cellular networks but is also used to describe the connections between each relay station and multiple SS in mobile *ad hoc* networks and wireless sensor networks.

All connections communicate with the BS using time-division multiplexing/time-division multiple access (TDM/TDMA). We will focus on the downlink here, although our results can be extended to the uplink as well. The wireless link of each connection from the BS to each SS is depicted in Fig. 2. A buffer is implemented at the BS for each connection and operates in a first-input-first-output (FIFO) mode. The AMC controller follows the buffer at the BS (transmitter), and the AMC selector is implemented at the SS (receiver).

At the PHY, multiple transmission modes are available to each user, with each mode representing a pair of a specific modulation format and a forward error control (FEC) code, as in IEEE 802.11/15/16, 3GPP, and 3GPP2 standards. Based on channel estimates obtained at the receiver, the AMC selector determines the modulation-coding pair (mode or burst profile), whose index is sent back to the transmitter through a feedback channel, for the AMC controller to update the transmission mode. Coherent demodulation and soft-decision Viterbi decoding are employed at the receiver. The decoded bit streams are mapped to packets, which are pushed upward to the MAC.

We consider the following group of transmission modes as in the IEEE 802.16 standard [3].

Transmission modes (TM): The modulations are M_n-ary rectangular/square quadrature amplitude modulators (QAMs), and the FEC codes are Reed–Solomon (RS) concatenated with convolutional codes (CC) (see Table I). Although we focus on this TM, other transmission modes can be similarly constructed [1]–[3], [13].

At the PHY, the processing unit is a frame consisting of multiple transmitted symbols. At the MAC, the processing unit is a packet comprising multiple information bits. Fig. 3 details the packet and frame structures.

1) At the MAC, 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 as in Table I, each packet is mapped to a symbol block containing N_b/R_n symbols.
2) At the PHY, the data are transmitted frame by frame through the wireless channel, with each frame containing a fixed number of symbols N_s. Given a fixed symbol rate, the frame duration T_f (in 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 $2N_b/R_1$ symbols. As a result, each time slot can transmit exactly $2R_n/R_1$ packets with transmission mode n. For the TM in particular, one time slot can accommodate $2R_1/R_1 = 2$ packets with mode $n = 1$, $2R_2/R_1 = 3$ packets with mode $n = 2$, and so on. The N_c time slots contain control information and pilots. The N_d time slots convey data, which are scheduled to different connections dynamically. Each connection 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.

B. QoS Architecture at the MAC

At the MAC, each connection belongs to a single service class and is associated with a set of QoS parameters that quantify its characteristics. Four QoS classes are provided by the MAC in the IEEE 802.16 standard.

1) Unsolicited grant service (UGS) supports constant bit rate (CBR) or fixed throughput connections such as E1/T1 lines and voice over IP (VoIP). This service provides guarantees on throughput, latency, and jitter to the necessary levels as TDM services. The QoS metrics here are the packet error rate (PER) and the service rate.
2) Real-time polling service (rtPS) provides guarantees on throughput and latency, but with greater tolerance on latency relative to UGS, e.g., MPEG video conferencing and video streaming. The delayed packets are useless and will be dropped. The QoS metrics are the PER and the maximum delay (or the maximum delay for a given outage probability).
3) Nonreal-time polling service (nrtPS) provides guarantees in terms of throughput only and is therefore suitable for mission critical data applications, such as File Transfer Protocol (FTP). These applications are time-insensitive and require minimum throughput. For example, an FTP file can be downloaded within a bounded waiting time if the minimum reserved rate is guaranteed. The QoS metrics are the PER and the minimum reserved rate.
4) Best effort (BE) service provides no guarantees on delay or throughput and is used for Hypertext Transport Protocol (HTTP) and electronic mail (e-mail), for example. BE applications receive the residual bandwidth after the bandwidth is allocated to the connections of the previous three service classes. Although no delay and rate is specified for BE connections, a prescribed PER should be maintained over wireless channels.

The signaling and procedure for the service setup and maintenance of each connection are defined as in the IEEE 802.16 standard [3]. However, the standard does not define the scheduling mechanism or the admission control and traffic policing processes. The signaling overhead is not included in our design and analysis.

C. AMC Design at the PHY

Efficient bandwidth utilization for a prescribed PER performance at the PHY can be accomplished with AMC schemes, which match transmission parameters to the time-varying

TABLE I
TRANSMISSION MODES IN THE IEEE 802.16 STANDARD

Mode n	1	2	3	4	5	6
Modulation	QPSK	QPSK	16QAM	16QAM	64QAM	64QAM
RS Code	(32,24,4)	(40,36,2)	(64,48,8)	(80,72,4)	(108,96,6)	(102,108,6)
CC Code Rate	2/3	5/6	2/3	5/6	3/4	5/6
Coding Rate R_c	1/2	3/4	1/2	3/4	2/3	3/4
R_n (bits/symbol)	1.0	1.5	2.0	3.0	4.0	4.5
a_n (dB)	232.9242	140.7922	264.0330	208.5741	216.8218	220.7515
g_n	22.7925	8.2425	6.5750	2.7885	1.0675	0.8125
γ_{pn}(dB)	3.7164	5.9474	9.6598	12.3610	16.6996	17.9629

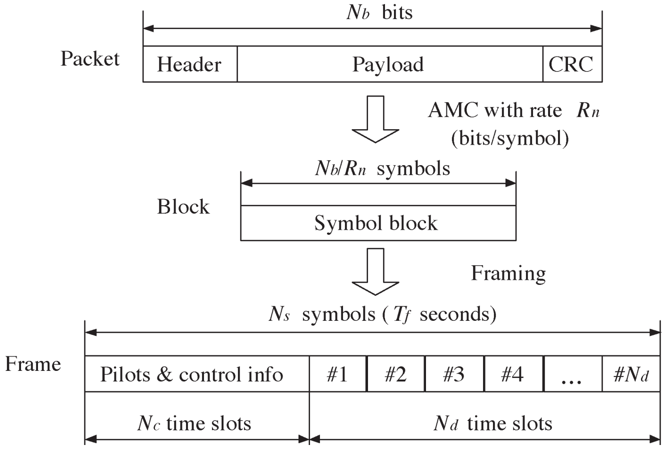

Fig. 3. Processing units at MAC and PHY.

wireless channel conditions adaptively and have been adopted by many standard wireless networks, such as IEEE 802.11/15/16 and 3GPP/3GPP2 [1]–[3], [7], [13].

Each connection with rtPS, nrtPS, and BE services relies on AMC at the PHY. The objective of AMC is to maximize the data rate by adjusting transmission modes to channel variations while maintaining a prescribed PER P_0, and the design procedure is similar to that proposed in [5] and [16].

Let N denote the total number of transmission modes available ($N = 6$ for TM). As in [5], we assume constant power transmission and partition the entire signal-to-noise ratio (SNR) range in $N + 1$ nonoverlapping consecutive intervals, with boundary points denoted as $\{\gamma_n\}_{n=0}^{N+1}$. In this case

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

To avoid deep-channel fades, no data are sent when $\gamma_0 \le \gamma < \gamma_1$, which corresponds to the mode $n = 0$ with rate $R_0 = 0$ bit/symbol. The design objective of AMC is to determine the boundary points $\{\gamma_n\}_{n=0}^{N+1}$.

To simplify the AMC design, we approximate the PER expression in AWGN channels as

$$\text{PER}_n(\gamma) \approx \begin{cases} 1, & \text{if } 0 < \gamma < \gamma_{pn} \\ a_n \exp(-g_n \gamma), & \text{if } \gamma \ge \gamma_{pn} \end{cases} \tag{2}$$

where n is the mode index and γ is the received SNR. Parameters a_n, g_n, and γ_{pn} in (2) are mode-dependent and are obtained by fitting (2) to the exact PER via simulations presented in [4]. With packet length $N_b = 128$ bytes/packet, the fitting parameters for transmission modes in TM are provided in Table I. Using the approximate yet simple expression (2) facilitates the mode selection. This approach has also been adopted by Hole *et al.* [10], where bit error rate (BER) is used as a figure of merit.

We set the region boundary (switching threshold) γ_n for the transmission mode n to be the minimum SNR required to guarantee P_0. Inverting the PER expression in (2), we obtain

$$\begin{aligned} \gamma_0 &= 0 \\ \gamma_n &= \frac{1}{g_n} \ln\left(\frac{a_n}{P_0}\right), \qquad n = 1, 2, \ldots, N \\ \gamma_{N+1} &= +\infty. \end{aligned} \tag{3}$$

With the boundaries $\{\gamma_n\}_{n=0}^{N}$ specified by (3), one can verify that the AMC in (1) guarantees that the PER is less than or equal to P_0. Maintaining the target PER performance, the proposed AMC transmissions with (1) and (3) are designed to maximize the spectral efficiency, with the given finite transmission modes.

In summary, our AMC design guarantees that the PER is less than or equal to P_0 by determining $\{\gamma_n\}_{n=0}^{N}$ as in (3) and updating the transmission mode as in (1).

III. SCHEDULER DESIGN

In this section, we describe our scheduler for multiple connections with diverse QoS requirements.

A. *Scheduling UGS Connections*

In UGS service, the transmission mode at the PHY is fixed to meet the prescribed "average" PER requirement as in traditional TDMA cellular networks, e.g., Global System for Mobile Communications (GSM). For example, the transmission mode could be selected in the initial service access phase via training to meet the average PER requirement. Then, the transmission mode is fixed during the whole service time. The AMC design in Section II-C is not adopted for UGS connections because voice services may tolerate some "instantaneous" packet loss and AMC feedback overhead should be reduced for low-rate voice traffic, e.g., 8-kbps voice stream. For these reasons, the time slots allocated for UGS connections are fixed, based on their constant-rate requirements at the MAC.

We denote the total time slots allocated to UGS connections as N_{UGS} per frame. The residual time slots $N_r = N_d - N_{\text{UGS}}$ are scheduled for the other three QoS classes, with UGS connections given higher priority than the other three QoS classes (rtPS, nrtPS, and BE).

B. Scheduling rtPS, nrtPS, and BE Connections

Each connection i, where i denotes the connection identification (CID) of rtPS, nrtPS, and BE services, adopts AMC at the PHY. Given a prescribed PER ξ_i, the SNR thresholds $\{\gamma_n\}_{n=0}^{N+1}$ for connection i are determined as in Section II-C by setting $P_0 = \xi_i$. Thus, the possible transmission rate (capacity), i.e., the number of packets that could be carried by N_r time slots for connection i at time t (frame index), can be expressed as

$$C_i(t) = N_r R_i(t) \tag{4}$$

where $R_i(t) \in \{2R_n\}_{n=0}^{N}$ is the number of packets that can be carried by one time slot and is determined by the channel quality of connection i via AMC as in (1). Notice that either $R_i(t)$ or $C_i(t)$ indicates the channel quality or capacity, which will be accounted for by the scheduler as we will see next.

At the MAC, the scheduler simply allocates all N_r time slots per frame to the connection

$$i^* = \arg\max_i \phi_i(t) \tag{5}$$

where $\phi_i(t)$ is the PRF for connection i at time t, which will be specified soon. If multiple connections have the same value $\max_i\{\phi_i(t)\}$, the scheduler will randomly select one of them with even opportunity.

For each rtPS connection, the scheduler timestamps each arriving packet according to its arrival time and defines its timeout if the waiting time of such a packet in queue is over the maximum latency (deadline) T_i. The PRF for a rtPS connection i at time t is defined as

$$\phi_i(t) = \begin{cases} \beta_{rt} \frac{R_i(t)}{R_N} \frac{1}{F_i(t)}, & \text{if } F_i(t) \ge 1,\ R_i(t) \ne 0 \\ \beta_{rt}, & \text{if } F_i(t) < 1,\ R_i(t) \ne 0 \\ 0, & \text{if } R_i(t) = 0 \end{cases} \tag{6}$$

where $\beta_{rt} \in [0,1]$ is the rtPS-class coefficient and $F_i(t)$ is the delay satisfaction indicator, which is defined as

$$F_i(t) = T_i - \Delta T_i - W_i(t) + 1 \tag{7}$$

with $\Delta T_i \in [0, T_i]$ denoting the guard time region ahead of the deadline T_i, and $W_i(t) \in [0, T_i]$ denoting the longest packet waiting time, i.e., the head of the line (HOL) delay. If $F_i(t) \ge 1$, i.e., $W_i(t) \in [0, T_i - \Delta T_i]$, the delay requirement is satisfied, and the effect on priority is quantified as $1/F_i(i) \in [0,1]$: Large values of $F_i(t)$ indicate high degree of satisfaction, which leads to low priority. If $F_i(t) < 1$, i.e., $W_i(t) \in (T_i - \Delta T_i, T_i]$, the packets of connection i should be sent immediately to avoid packet drop due to delay outage, so that the highest value of PRF β_{rt} is set. Parameter $R_N := \max_n\{2R_n\}_{n=0}^{N}$, and the factor $R_i(t)/R_N \in [0,1]$ quantifies the normalized channel quality because high received SNR induces high capacity, which results in high priority. When $R_i(t) = 0$, the channel is in deep fade and the capacity is zero, so that connection i should not be served regardless of delay performance. Notice that the value of $\phi_i(t)$ for rtPS connection i lies in $[0, \beta_{rt}]$.

For each nrtPS connection, guaranteeing the minimum reserved rate η_i means that the average transmission rate should be greater than η_i. In practice, if data of connection i are always available in queue, the average transmission rate at time t is usually estimated over a window size t_c based on (4) and (5) as

$$\hat{\eta}_i(t+1) = \begin{cases} \hat{\eta}_i(t)(1 - 1/t_c), & \text{if } i \ne i^* \\ \hat{\eta}_i(t)(1 - 1/t_c) + C_i(t)/t_c, & \text{if } i = i^*. \end{cases} \tag{8}$$

We would like to guarantee $\hat{\eta}_i(t) \ge \eta_i$ during the entire service period. The PRF for an nrtPS connection i at time t is defined as

$$\phi_i(t) = \begin{cases} \beta_{\mathrm{nrt}} \frac{R_i(t)}{R_N} \frac{1}{F_i(t)}, & \text{if } F_i(t) \ge 1,\ R_i(t) \ne 0 \\ \beta_{\mathrm{nrt}}, & \text{if } F_i(t) < 1,\ R_i(t) \ne 0 \\ 0, & \text{if } R_i(t) = 0. \end{cases} \tag{9}$$

where $\beta_{\mathrm{nrt}} \in [0,1]$ is the nrtPS-class coefficient, and $F_i(t)$ is the ratio of the average transmission rate over the minimum reserved rate

$$F_i(t) = \hat{\eta}_i(t)/\eta_i. \tag{10}$$

Quantity $F_i(t)$ here is the rate satisfaction indicator. If $F_i(t) \ge 1$, the rate requirement is satisfied, and its effect on priority is quantified as $1/F_i(t) \in [0,1]$. If $F_i(t) < 1$, the packets of connection i should be sent as soon as possible to meet the rate requirement; in this case, the upper-bound value β_{nrt} is set for $\phi_i(t)$. Once again, the value of $\phi_i(t)$ lies in $[0, \beta_{\mathrm{nrt}}]$.

For BE connections, there are no QoS guarantees. The PRF for a BE connection i at time t is

$$\phi_i(t) = \beta_{\mathrm{BE}} \frac{R_i(t)}{R_N} \tag{11}$$

where $\beta_{\mathrm{BE}} \in [0,1]$ is the BE-class coefficient. Notice that $\phi_i(t)$ varies in $[0, \beta_{\mathrm{BE}}]$, which only depends on the normalized channel quality regardless of delay or rate performance.

The role of β_{rtPS}, β_{nrtPS}, and β_{BE} in (6), (9), and (11), respectively, is to provide different priorities for different QoS classes. For example, if the priority order for different QoS classes is rtPS > nrtPS > BE, the coefficients can be set under the constraint $\beta_{\mathrm{rtPS}} > \beta_{\mathrm{nrtPS}} > \beta_{\mathrm{BE}}$; e.g., $\beta_{\mathrm{rtPS}} = 1.0 > \beta_{\mathrm{nrtPS}} = 0.8 > \beta_{\mathrm{BE}} = 0.6$. Thus, the QoS of connections in a high-priority QoS class can be satisfied prior to those of a low-priority QoS class because the value of $\phi_i(t)$ for QoS unsatisfied connections will equal the upper-bound β_{rtPS}, β_{nrtPS}, and β_{BE} for rtPS, nrtPS, and BE connections, respectively. The purpose of normalizing $\phi_i(t)$ in $[0, \beta_{\mathrm{rtPS}}]$, $[0, \beta_{\mathrm{nrtPS}}]$, and $[0, \beta_{\mathrm{BE}}]$, respectively, is to provide comparable priorities among connections with different kinds of services, which enable exploiting multiuser diversity among all connections with rtPS, nrtPS, and BE services.

In the same service class of rtPS or nrtPS, $\phi_i(t)$ for different connections can only depend on the normalized channel quality $R_i(t)/R_N$ and QoS satisfaction indicator $F_i(t)$, where the principle is similar to that of the scheme in [6]. However, the major difference with our PRF design is that the value of $\phi_i(t)$ for the QoS unsatisfied connection is set to its upper-bound β_{rtPS}, β_{nrtPS}, and β_{BE} for rtPS, nrtPS, and BE connections, respectively. This design can result in better QoS guarantees than [6]. For example, when QoS is not satisfied for a connection, bad channel quality may lead to low priority in [6], so that QoS will even decrease. However, our design assigns the highest priority to such a connection, which increases its QoS as soon as possible, thus providing better QoS guarantees. A similar observation has been made also in [14] for delay-sensitive traffic.

IV. Desirable Features

In this section, we will summarize the features of our proposed scheduler with reference to the scheduler design criteria suggested in [9]. Here are the attributes of the proposed scheduler.

1) Efficient bandwidth utilization is achieved through the normalized channel quality factor $R_i(t)/R_N$ in $\phi_i(t)$ for each connection, so that the scheduler will not assign a time slot to the connection with bad channel quality and multiuser diversity can be exploited.
2) Delay bound T_i is provided for rtPS connections. When the HOL delay $W_i(t)$ approaches T_i $(W_i(t) \in (T_i - \Delta T_i, T_i])$, the highest value β_{rtPS} is set for connection i, which will be served as soon as possible. Because wireless channels can experience deep fades, the delay outage event, e.g., $W_i(t) = T_i$ and $C_i(t) = 0$, cannot be avoided. Thus, the hard delay bound may not be guaranteed. However, controlling the delay outage probability below the practically acceptable values could be realized by adjusting ΔT_i.
3) Throughput is guaranteed for nrtPS connections if sufficient bandwidth is provided. When $\hat{\eta}_i(t) < \eta_i$, the highest priority β_{nrt} will be set for connection i, which will be served as soon as possible until its throughput requirement is satisfied.
4) Implementation complexity is low because our priority-based scheduler simply updates the priority of each connection per frame and allocates N_r time slots to the connection with the highest priority as in (5).
5) Flexibility is provided because the scheduling does not depend on any traffic or channel model.
6) Scalability is achieved. When the available bandwidth decreases by adding new connections to the system for instance, the performance of connections with low-priority service classes will be degraded prior to those with high-priority service classes, as we will verify by simulations in the ensuring section.

V. Simulations

Because the design and performance of fixed TDMA bandwidth allocation for UGS connections are well understood, here, we only focus on the scheduling for rtPS, nrtPS, and BE connections.

We list the assumptions we employed in simulations.

A1: The wireless channel quality of each connection remains constant per frame, but is allowed to vary from frame to frame. This corresponds to a block-fading channel model, which is suitable for slowly varying wireless channels. Thus, AMC is implemented on a frame-by-frame basis [16].

A2: Perfect channel state information (CSI) is available at the receiver via training-based channel estimation. The corresponding transmission mode selection is fed back to the transmitter without error and latency [5]. The assumption that the feedback channel is error free could be (at least approximately) satisfied by using heavily coded feedback streams [3]. On the other hand, the feedback latency can be compensated by channel prediction; see, e.g., [8] and references therein.

A3: Error detection based on CRC is perfect, provided that sufficiently reliable error detection CRC codes are used per packet [17].

A4: If a packet is received incorrectly after error detection, we declare packet loss.

A. Channel Model

For fading channels adhering to A1, the channel quality is captured by a single parameter, namely, the "instantaneous" SNR γ, which remains invariant during a frame. We adopt the general Nakagami-m model to describe γ statistically [5]. The received SNR γ per frame is thus a random variable with a Gamma probability density function, i.e.,

$$p_\gamma(\gamma) = \frac{m^m \gamma^{m-1}}{\bar{\gamma}^m \Gamma(m)} \exp\left(-\frac{m\gamma}{\bar{\gamma}}\right) \tag{12}$$

where $\bar{\gamma} := E\{\gamma\}$ is the "average" received SNR, $\Gamma(m) := \int_0^\infty t^{m-1} e^{-t} dt$ is the Gamma function, and m is the Nakagami fading parameter $(m \geq 1/2)$. This model includes the Rayleigh channel when $m = 1$. A one-to-one mapping between the Ricean factor and the Nakagami fading parameter m allows Ricean channels to be well approximated by the Nakagami-m channels [22]. This channel model is suitable for flat-fading channels as well as frequency-selective fading channels encountered with orthogonal frequency-division multiplexing (OFDM) systems [12].

With our AMC design in Section II-C, 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 considering mobility-induced Doppler effects, we rely on a finite-state Markov chain (FSMC) model, which is developed in [15]. The state transition matrix of such FSMC is

$$\mathbf{P}_c = [P_{l,n}]_{(N+1)\times(N+1)} \tag{13}$$

which depends on the statistical channel parameters: average received SNR $\bar{\gamma}$, Nakagami fading parameter m, and mobility-induced Doppler spread f_d [15].

Although we adopt the channel transition matrix in (13) for the Nakagami fading channel, the ensuing results apply also to other kinds of channel transition matrices.

B. Parameter Setting

The wireless channel of connection i is modeled using an FSMC, as in Section V-A, with parameters: the average

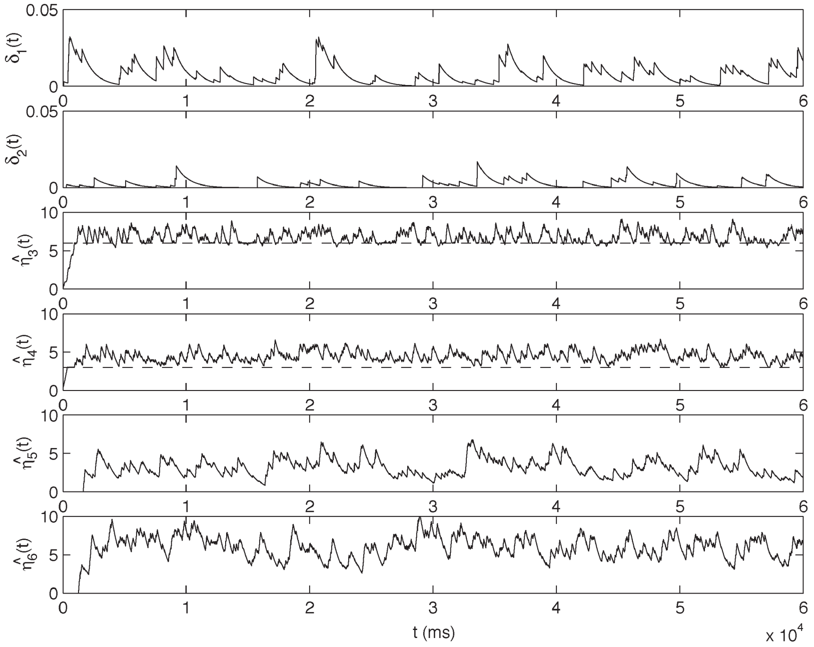

Fig. 4. $\delta_1(t)$, $\delta_2(t)$, $\hat{\eta}_3(t)$, $\hat{\eta}_4(t)$, $\hat{\eta}_5(t)$, and $\hat{\eta}_6(t)$ versus t for $N_r = 3$ (dashed lines indicate $\eta_3 = 6$ and $\eta_4 = 3$).

received SNR $\bar{\gamma}_i$, the Nakagami fading parameter m_i, and the Doppler frequency $f_{d,i}$. The frame length is $T_f = 1$ ms. The packet length at the MAC is fixed to $N_b = 128$ bytes.

For each rtPS connection i, as in [21], we assume that the arrival process to the queue is Bernoulli distributed with a given average rate η_i and parameter $p_i \in (0, 1)$. As a result, the instantaneous arriving rate at time t can be expressed as

$$A_i(t) = \begin{cases} 0, & \text{with probability } p_i \\ \eta_i/(1-p_i), & \text{with probability } 1-p_i. \end{cases} \tag{14}$$

The QoS parameters are the PER ξ_i and the maximum delay T_i. We consider that two rtPS connections are admitted in the system with $i = 1$ and 2, respectively. Their channel, QoS, and traffic parameters are as follows:

$$\begin{aligned} &\bar{\gamma}_1 = 15 \text{ dB}, \quad m_1 = 1.2, \qquad f_{d,1} = 10 \text{ Hz}, \quad \xi_1 = 10^{-2} \\ &T_1 = 30 \text{ ms}, \quad \eta_1 = 2 \text{ Mbps}, \quad p_1 = 0.4 \\ &\bar{\gamma}_2 = 20 \text{ dB}, \quad m_2 = 1, \qquad f_{d,2} = 10 \text{ Hz}, \quad \xi_2 = 10^{-2} \\ &T_2 = 50 \text{ ms}, \quad \eta_2 = 1 \text{ Mbps}, \quad p_2 = 0.5. \end{aligned}$$

The guard time is set to $\Delta T_i = 1$ ms. The delay performance of rtPS connections is evaluated by the delay outage probability $\delta_i(t)$ over a window size $t_c = 1000$ ms as

$$\delta_i(t+1) = \begin{cases} \delta_i(t)(1-1/t_c), & \text{if } W_i(t) < T_i \\ \delta_i(t)(1-1/t_c)+1/t_c, & \text{if } W_i(t) = T_i \text{ and } i \neq i^*. \end{cases} \tag{15}$$

The delay outage event happens at time $t+1$ when $W_i(t) = T_i$ and $i \neq i^*$ because $C_i(t) = 0$, or multiple rtPS connections compete with the same PRF value β_{rtPS}.

For each nrtPS connection i, we assume that the data are always available, which is reasonable for FTP applications, for example. The QoS parameters are the PER ξ_i and the minimum reserved rate η_i. Two nrtPS connections are admitted in the system with $i = 3$ and 4, respectively. Their channel and QoS parameters are as follows:

$$\begin{aligned} &\bar{\gamma}_3 = 15 \text{ dB}, \ m_3 = 1, \ f_{d,3} = 10 \text{ Hz}, \ \xi_3 = 10^{-3}, \ \eta_3 = 6 \text{ Mbps} \\ &\bar{\gamma}_4 = 20 \text{ dB}, \ m_4 = 1, \ f_{d,4} = 10 \text{ Hz}, \ \xi_4 = 10^{-3}, \ \eta_4 = 3 \text{ Mbps}. \end{aligned}$$

The rate performance of nrtPS connections is evaluated by the average service rate $\hat{\eta}_i(t)$ over a window size of $t_c = 1000$ ms based on (8).

For each BE connection i, we assume that the data are always available for HTTP or e-mail applications. The pertinent QoS parameter here is just the PER ξ_i. We consider two BE connections in the system with $i = 5$ and 6, respectively. Their channel and QoS parameters are as follows:

$$\begin{aligned} &\bar{\gamma}_5 = 16 \text{ dB}, \quad m_5 = 1, \quad f_{d,5} = 10 \text{ Hz}, \quad \xi_5 = 10^{-3} \\ &\bar{\gamma}_6 = 18 \text{ dB}, \quad m_6 = 1, \quad f_{d,6} = 10 \text{ Hz}, \quad \xi_6 = 10^{-3}. \end{aligned}$$

The rate performance of BE connections is also evaluated by the average service rate $\hat{\eta}_i(t)$ over a window size $t_c = 1000$ ms based on (8).

The system is simulated over 60 000 ms with bounds $\beta_{\text{rtPS}} = 1.0$, $\beta_{\text{nrtPS}} = 0.8$, and $\beta_{\text{BE}} = 0.6$, respectively.

C. Results and Discussion

Because AMC in Section II-C is employed by each connection of rtPS, nrtPS, and BE services at the PHY, the prescribed PER is guaranteed by setting $P_0 = \xi_i$; for this reason, we only focus on delay and rate performance here.

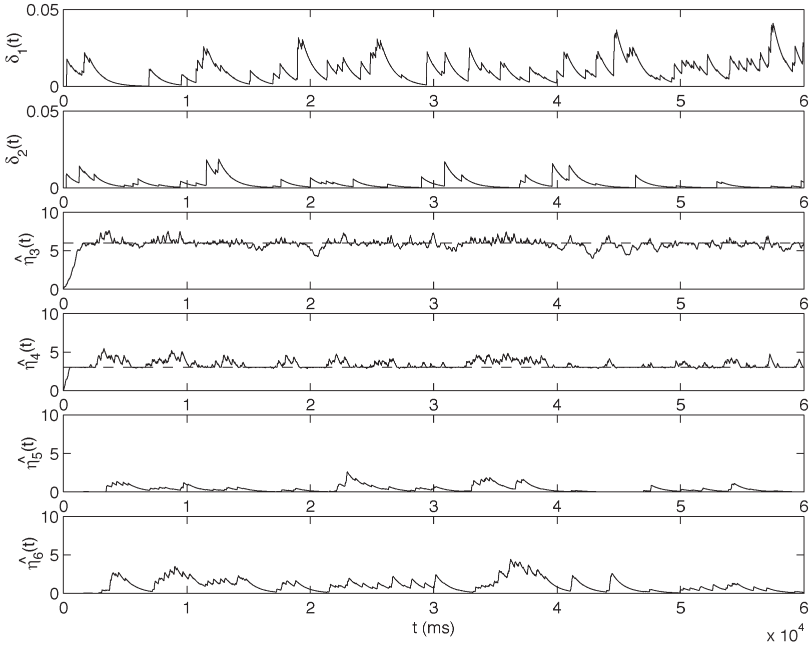

Fig. 5. $\delta_1(t)$, $\delta_2(t)$, $\hat{\eta}_3(t)$, $\hat{\eta}_4(t)$, $\hat{\eta}_5(t)$, and $\hat{\eta}_6(t)$ versus t for $N_r = 2$ (dashed lines indicate $\eta_3 = 6$ and $\eta_4 = 3$).

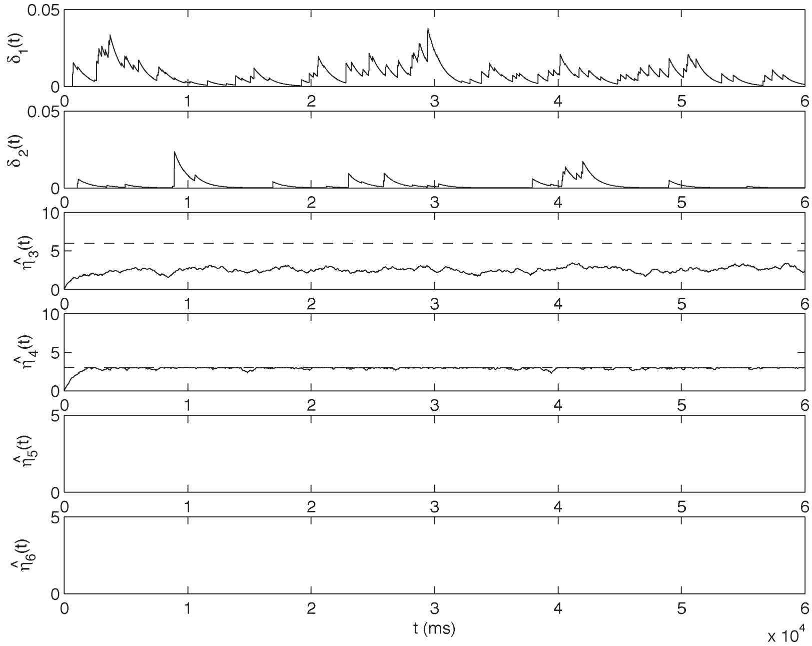

Fig. 6. $\delta_1(t)$, $\delta_2(t)$, $\hat{\eta}_3(t)$, $\hat{\eta}_4(t)$, $\hat{\eta}_5(t)$, and $\hat{\eta}_6(t)$ versus t for $N_r = 1$ (dashed lines indicate $\eta_3 = 6$ and $\eta_4 = 3$).

The delay outage probability $\delta_i(t)$ of rtPS connections $i = 1$ and 2 and the average transmission rate $\hat{\eta}_i(t)$ of nrtPS and BE connections $i = 3, 4, 5,$ and 6 are plotted in Figs. 4–6 for $N_r = 3, 2,$ and 1, respectively.

Fig. 4 depicts the performance for $N_r = 3$ and shows that the delay outage probabilities $\delta_1(t)$, $\delta_2(t)$ are always below 5%, which is satisfied for normal rtPS applications. The average values of $\delta_1(t)$ and $\delta_2(t)$ are 0.77% and 0.24% for connection $i = 1$ and 2, respectively, which indicate good delay performance for rtPS connections. Notice that $\hat{\eta}_3(t)$ and $\hat{\eta}_4(t)$ vary around the minimum reserved rates $\eta_3 = 6$ and $\eta_4 = 3$, but the variations are small. The average values of $\hat{\eta}_3(t)$ and $\hat{\eta}_4(t)$ are 6.7 and 4.5, which illustrate good rate guarantees for nrtPS connections. Furthermore, $\hat{\eta}_5(t)$ and $\hat{\eta}_6(t)$ show that the average transmission rates for $i = 5$ and 6 have large variations and even approach zero sometimes. This confirms that the rate performance is not guaranteed for BE connections, in the presence of channel fading and connections with higher priority service classes. The average values of $\hat{\eta}_5(t)$ and $\hat{\eta}_6(t)$ are 3.1 and 5.9, respectively.

Fig. 5 illustrates the performance for $N_r = 2$. Here, one time slot is reduced from $N_r = 3$, which may be interpreted as new UGS connections admitted, and the system has to increase N_{UGS}, which in turn reduces $N_r = N_d - N_{\mathrm{UGS}}$. We find that the performance of $\delta_1(t)$ and $\delta_2(t)$ for rtPS connections is similar to that for $N_r = 3$, and their average values for $N_r = 2$ are 1.06% and 0.29%, respectively. Notice that $\hat{\eta}_3(t)$ and $\hat{\eta}_4(t)$ for nrtPS connections still vary around the minimum reserved rates, and their average values are 5.8 and 3.4, respectively; whereas $\hat{\eta}_5(t)$ and $\hat{\eta}_6(t)$ for BE connections have average values 0.4 and 1.0, respectively, which are much smaller than those for $N_r = 3$.

Comparing the results for $N_r = 2$ and $N_r = 3$, it is clear that the scheduler guarantees the prescribed QoS for rtPS and nrtPS connections. However, the rate performance of BE connections for $N_r = 2$ is worse than that for $N_r = 3$ due to the bandwidth reduction.

Fig. 6 depicts the performance for $N_r = 1$. The average values of $\delta_1(t)$ and $\delta_2(t)$ are 0.78% and 0.18%, respectively, so that the delay performance is still good for rtPS connections. However, the average values of $\hat{\eta}_3(t)$ and $\hat{\eta}_4(t)$ for nrtPS connections become 2.5 and 2.9, respectively, which implies performance degradation for the prescribed rate requirements due to insufficient available bandwidth. Notice that BE connections are not served at all because $\hat{\eta}_5(t) = 0$ and $\hat{\eta}_6(t) = 0$.

From the results for $N_r = 3, 2,$ and 1, we find that scalability is also achieved. When the available bandwidth decreases by adding new connections for instance, the connections with low-priority service classes experience degraded performance prior to those in high-priority classes.

VI. Conclusion and Future Directions

In this paper, we developed a cross-layer scheduling algorithm at the MAC layer for multiple connections with diverse QoS requirements, which can be used in cellular networks, mobile *ad hoc* networks, and wireless sensor networks. Each connection admitted in the system is assigned a priority, which is updated dynamically depending on its channel quality, QoS satisfaction, and service priority; thus, the connection with the highest priority is scheduled first each time. Our proposed scheduler offers prescribed delay, and rate guarantees for real-time and nonreal-time traffic; at the same time, it uses the wireless bandwidth efficiently by exploiting multiuser diversity among connections with different kinds of services. Furthermore, our scheduler enjoys flexibility, scalability, and low implementation complexity. Performance of our scheduler was evaluated via simulations in the IEEE 802.16 standard setting, where the upper-bound β_{rtPS}, β_{nrtPS}, β_{BE}, and the delay guard time ΔT_i were set heuristically.

Their effects on performance are worthy of further research. Furthermore, our scheduler allocates all N_r time slots to one connection each time for simplicity; however, scheduling multiple connections each time may lead to better performance, which is under current investigation. The fairness issue for the users in the same service class is another topic in our research agenda. The effects of imperfect channel state information due to estimation error and feedback latency are also worth further study.

References

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

[2] 3GPP2 C.S0002-0 Version 1.0, *Physical Layer Standard for cdma2000 Spread Spectrum Systems*, Jul. 1999.

[3] IEEE Standard 802.16 Working Group, *IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems (Revision of IEEE Standard 802.16-2001)*, 2004.

[4] ——, *FEC Performance With ARQ and Adaptive Burst Profile Selection*, 2001.

[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] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 150–154, Feb. 2001.

[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] 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.

[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] 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.

[11] E. Hossain and V. K. Bhargava, "Link-level traffic scheduling for providing predictive QoS in wireless multimedia networks," *IEEE Trans. Multimedia*, vol. 6, no. 1, pp. 199–217, Feb. 2004.

[12] Z. Kang, K. Yao, and F. Lorenzelli, "Nakagami-m fading modeling in the frequency domain for OFDM system analysis," *IEEE Commun. Lett.*, vol. 7, no. 10, pp. 484–486, Oct. 2003.

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

[14] A. Khattab and K. Elsayed, "Channel-aware scheduling schemes with statistical delay-bound guarantees in wireless multimedia networks," in *Proc. ACM/IEEE MSWiM*, Venice, Italy, Oct. 4–6, 2004, pp. 31–38.

[15] Q. Liu, S. Zhou, and G. B. Giannakis, "Cross-layer modeling of adaptive wireless links for QoS support in multimedia networks," in *Proc. 1st Int. Conf. QShine*, Dallas, TX, Oct. 18–20, 2004, pp. 65–75.

[16] ——, "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.

[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] J. Padhye, V. Firoiu, D. F. Towsley, and J. F. Kurose, "Modeling TCP Reno performance: A simple model and its empirical validation," *IEEE/ACM Trans. Netw.*, vol. 8, no. 2, pp. 133–145, Apr. 2000.

[19] T. Rappaport, *Wireless Communications: Principles and Practice*. Upper Saddle River, NJ: Prentice-Hall, 1996.

[20] 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.

[21] S. Shakkottai and A. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," in *Proc. 17th ITC*, Salvador da Bahia, Brazil, Sep. 24–28, 2001, pp. 793–804.

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

[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] 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.

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

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

His research interests are in the areas of communications, signal processing, and networking, with emphasis on cross-layer design and optimization, quality of service provisioning, resource allocation, and link adaptation.

Xin Wang (S'03–M'04) received the B.Sc. and M.Sc. degrees from Fudan University, Shanghai, China, in 1997 and 2000, respectively, and the Ph.D. degree from Auburn University, Auburn, AL, in 2004, all in electrical engineering.

He is currently a Research Associate at the University of Minnesota, Minneapolis. His research interests include medium access control, cross-layer design, energy-efficient resource allocation, and signal processing for communication networks.

Georgios B. Giannakis (S'84–M'86–SM'91–F'97) received the Diploma in electrical engineering from the National Technical University of Athens, Athens, Greece, in 1981 and 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.

After lecturing for one year at USC, he joined the University of Virginia, Charlottesville, 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. He has published more than 250 journal papers, 400 conference papers, and two edited books. His current research focuses on diversity techniques for fading channels, complex-field and space–time coding, multicarrier, ultrawide band wireless communication systems, cross-layer designs, and sensor networks.

Dr. Giannakis is the (co)recipient of six paper awards from the IEEE Signal Processing (SP) and Communications Societies (1992, 1998, 2000, 2001, 2003, and 2004). He also received Technical Achievement Awards from the SP Society in 2000 and from EURASIP in 2005. 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 SP Conference Board, as member of the SP Publications Board, as Member and Vice-Chair of the Statistical Signal and Array Processing Technical Committee, as Chair of the SP for Communications Technical Committee, and as a member of the IEEE Fellows Election Committee. He has also served as a member of the IEEE-SP 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.