EURASIP Journal on Applied Signal Processing 2005:2, 167–182

Joint Cross-Layer Design for Wireless QoS Content Delivery

Jie Chen
Division of Engineering, Brown University, Providence, RI 02912, USA
Email: jie_chen@brown.edu

Tiejun Lv
School of Information Engineering, Beijing University of Posts and Communications, Beijing 100876, China
Email: lvtiejun@mail.tsinghua.edu.cn

Haitao Zheng
Microsoft Research Asia, Sigma Center Beijing, 49 Zhichun Road, Beijing 100080, China
Email: htzheng@microsoft.com

Received 14 August 2003; Revised 27 June 2004

We propose a joint cross-layer design for wireless quality-of-service (QoS) content delivery. Central to our proposed cross-layer design is the concept of *adaptation*. Adaptation represents the ability to adjust protocol stacks and applications to respond to channel variations. We focus our cross-layer design especially on the application, media access control (MAC), and physical layers. The network is designed based on our proposed fast frequency-hopping orthogonal frequency division multiplex (OFDM) technique. We also propose a QoS-awareness scheduler and a power adaptation transmission scheme operating at both the base station and mobile sides. The proposed MAC scheduler coordinates the transmissions of an IP base station and mobile nodes. The scheduler also selects appropriate transmission formats and packet priorities for individual users based on current channel conditions and the users' QoS requirements. The test results show that our cross-layer design provides an excellent framework for wireless QoS content delivery.

Keywords and phrases: cross-layer design, wireless QoS, MAC scheduler, power adaptation, frequency-hopping OFDM.

1. INTRODUCTION

As mobile cellular networks are evolving to carry voice, video, and data services, an all-Internet protocol- (IP-)based system, including a radio access network and a core network, akin to the Internet, is likely to become the most favorable solution for future wireless Internet-centric systems [1, 2]. The advantages of all-IP networks include cost efficiency, improved reliability, ease of new service implementation, and ease of integration with heterogenous networks. In this paper, we adopt an IP-based network as our cross-layer design platform. With its TCP/IP-based design, the IP network can easily adapt to changing wireless transmission conditions and can offer users a "LAN-like" experience for broadband data and voice services. However, significant difficulties in developing IP-based wireless data networks to meet quality-of-service (QoS) requirements remain and must be addressed. A summary of these difficulties follows.

Dynamic link characteristics

Because a mobile device transmits and receives radio signals over the air, wireless transmission is vulnerable to noise and other types of interference. The loss-of-sight effect, multipath fading, and interference from other devices make channel conditions vary unpredictably over time. Changing the transmission rate as a channel varies does improve transmission efficiency but could also result in data rate oscillation. The mobility of handsets further increases difficulties in channel estimation and prediction, thus increasing the bit error rate (BER). Two approaches, channel coding and link-layer automatic retransmission request (ARQ), have commonly been used to address this problem. The first approach employs sophisticated channel coding and interleaving techniques. For example, turbo coding, despite its complexity, is now the standard channel coding technique in the 3G Universal Mobile Telecommunications System (UMTS) [3]. This approach, however, heavily relies on the accuracy of channel

estimation. The second approach, a link-layer ARQ mechanism, performs error control by retransmitting lost frames [4]. Although it is insensitive to inaccuracies in channel estimation, this approach introduces latency and delay jitter to IP packet flows. The tradeoff between latency and reliability depends on ARQ persistence, which is defined as a protocol's willingness to retransmit lost frames to ensure reliable transmission [5]. The persistence can also be expressed in terms of time or the maximum number of retransmissions.

Resource contention

As in wireline networks, users share channel resources in wireless networks. When multiple users run different applications, the most salient issue is QoS variability, including error rate, latency, and bandwidth. The resource contention problem has been quite challenging in the design of wireline networks. Taking mobility, unpredictable link variation, and dynamic network topology into consideration makes wireless networks even more difficult to design. The media access control (MAC) layer, for instance, uses a scheduler to determine the next user to be served based on the channel condition and the individual user's QoS requirements [6, 7, 8]. Conventional schedulers have usually been developed only for downlink (DL) transmission because the base station alone gathers all user information. Uplink (UL) transmission has typically been made through contentions, yielding high delay jitters. Most scheduler algorithms have been designed to maximize system throughput, but do not take any QoS requirements into design consideration.

One important cause of the difficulties previously mentioned is the *dynamic behavior* of networks. Dynamic resource allocation is a general way to overcome these difficulties. The basic idea behind dynamic resource allocation is to optimize resource usage, such as transmit power and symbol rate as well as modulation and coding schemes, and so forth. Many proposed approaches have been based on techniques like power control, rate adaptation, dynamic channel allocation [9], beamforming [10] and multiuser detection [11]. There are also many joint design methods including adaptive rate and power control [12] and power control and beamforming [13]. Network performance can be further improved by adopting the concept of *cross-layer design*. Assuming that the data received follows independent and identically distributed (i.i.d.) Bernoulli processes, an optimal scheduler was proposed in [14]. A proportionally fair QoS scheduler has been described in [8]. However, individual users have specific QoS requirements and the fair scheduling mechanism is too simple to handle these. Other more complicated schemes have also been proposed in [15]. The high data rate (HDR) versions of CDMA2000 and the enhanced general packet radio service (EGPRS) have provided scheduling mechanisms to accommodate cross-layer design. A recent survey of cross-layer design is provided in [16].

The previously mentioned schemes are not true "cross-layer" design schemes. A cross-layer design, in general, should involve at least two out of the five key layers, such as the application layer, transport layer, network layer, MAC layer, and physical layer. Cross-layer design can be implemented within different protocol layers in the overall protocol stacks. We briefly explain the different approaches.

(1) *Application layer.* Application transmission adaptation refers to an application's capability to adjust its behavior to changing network and channel characteristics. Wireless networks often have to deal with adverse conditions in which handoffs, deep fading, and bad carrier signals can lead to high packet losses. Only adaptive applications that are network/channel aware can cope with these challenging circumstances. For multimedia delivery, a media server can track packet losses and adjust the media source rate accordingly [17, 18, 19, 20, 21]. To reduce information loss, the media server can employ packet forward error correction (FEC) coding and unequal error protection.

(2) *Transport layer.* The underlying transport layer could also be adapted, making it transparent to the application layer. As a result, applications originally developed for wireline networks could remain intact. One drawback of adaptation at this level is that it is impossible to implement a complete adaptation if the part of transport layer is application specific. The protocol should differentiate between various packet loss patterns, that is, packet losses due to network congestion or due to channel errors [22, 23, 24]. The protocol should also invoke congestion control and rate adaptation accordingly. Several cross-layer approaches, such as EBSN [25], snoop [26], and freeze TCP [27], have been proposed as TCP alternatives; they are designed to distinguish congestion loss from noncongestion loss, and to invoke different flow control mechanisms in each case. Note that the transport layer can only adapt effectively if it can (i) observe network-layer and link-layer conditions, (ii) propagate information to the application layer, and (iii) identify and accommodate the application layer's needs in the meantime.

(3) *Link and network layers.* The application characteristics, such as its QoS requirements and packet priorities, could be used in coordinating the behavior of the link and network layers. In particular, the persistence level of the MAC-layer ARQ adapts to an individual application's latency and reliability requirements. The link layer scheduler also allocates radio resources to various packet flows based on their QoS priorities. This adaptation, however, requires the MAC and network layers to be able to distinguish different packet flows. This differentiation, in general, can be achieved by an explicit indication of the QoS requirements associated with each packet flow [5]. Note that, in some systems, the transport and MAC layers both can conduct error recovery using FEC coding and packet retransmissions. The balance between both schemes is important for achieving optimal usage of overall communication resources. The network could also operate efficiently by using link- and physical-layer information, such as rate fluctuation and error condition, to distribute channel resources.

(4) *Joint MAC and physical layer.* A cross-layer design for a physical-layer transceiver and a MAC protocol for multiple-input multiple-output (MIMO-)OFDM ad hoc networks has been proposed [28]. The joint design results in better BER, and high MAC throughput and fairness. A cross-layer design for efficient resource allocation has also been discussed in

[16] where the authors propose to share MAC- and physical-layer information with higher layers. The impact of signal processing on physical layer and MAC protocol design was studied in [29].

(5) *Predict future network and channel conditions.* An essential factor in adaptation is each layer's ability to estimate current network and channel conditions and even to predict ones, as well as to exchange information across different layers. A receiving entity then evaluates current conditions to invoke a reception mechanism accordingly, while a transmitting entity uses current and future conditions to adjust its transmission flow. A condition report based on a receiver's feedback is normally more accurate than estimations performed by a transmitter alone.

In order to build an efficient wireless network, we strive to create a series of protocol layers that can communicate and interact with each other so that we can continuously achieve service improvements. Such adaptation, however, can be much easier to achieve for wireline networks than for wireless networks. In this paper, we propose a cross-layer design, especially focusing on interactions among the application, MAC, and physical layers, for wireless QoS content delivery. We propose a QoS-awareness scheduler and a power adaptation scheme for both UL and DL MAC layers. The proposed scheme will achieve resource efficiency by coordinating transmissions at the physical layer. For instance, the MAC scheduler may select an appropriate transmission format and packet priority for each individual user depending on current channel condition and the user's QoS requirements. The rest of this paper is organized as follows. In Section 2, we describe our physical and MAC layer designs in detail. In Section 3, we show how the cross-layer design can meet users' QoS requirements under different wireless transmission conditions. To avoid unnecessary QoS degradations, we propose a QoS-aware scheduler and a link transmission design that adapt to rapidly changing channel conditions. Test results in Section 4 show that the proposed cross-layer design provides a good transmission scheme for wireless QoS content delivery. Finally, we conclude our paper in Section 5.

2. SYSTEM ARCHITECTURE

In this section, we will first describe our design platform, which is based on the residue number system (RNS) frequency-hopping (FH) OFDM technique. Then, we will present the system architecture, especially focusing on the physical and MAC layer designs.

2.1. Physical layer design

In conventional OFDM design, an OFDM symbol starting at $t = t_s$ can be written as

$$s(t) = \sum_{-N_s/2}^{N_s/2-1} d_{i+N_s/2} e^{j2\pi i(t-t_s)/T}, \quad t_s \le t \le t_s + T, \qquad s(t) = 0, \quad t < t_s,\ t > t_s + T. \tag{1}$$

Here d_i is a complex quadrature amplitude modulation (QAM) symbol, N_s is the subcarrier number, and T is the symbol duration. An OFDM system usually has higher data throughput than a single-carrier system like the CDMA and time division multiple access (TDMA) designs [6, 30, 31, 32]. To further improve data throughput and enable rapid ARQ, we propose an RNS-OFDM design. The RNS-OFDM is referred to as a fast-hopping OFDM design; it can also be viewed as a wideband spread-spectrum technique designed specifically for broadband mobile data communications. In a wireless cellular network, the cell spectrum W is divided into a set of *subcarriers*. The time is divided into *symbols*. A number of these subcarriers or *tones* are assigned to a user when there is data to send or to receive. Subcarriers comprising a user's channel hop over the entire band W as time progresses. The pattern of subcarriers in the time/frequency space is referred to as the *hopping pattern* based on the RNS design. For OFDM systems, there is no intracell interference as users within the same cell are orthogonal to each other. However, intercell interference still exists, especially for users located near cell boundaries. FH has been proposed to alleviate this problem. The FH pattern's design must [33] (i) avoid ambiguity when identifying users; (ii) reduce the chances of a collision between two FH patterns; (iii) distribute subcarriers evenly; and (iv) keep adjacent FH patterns far from each other. After surveying various FH pattern construction techniques, such as the *m-sequence* [33], Reed-Solomon codes [34], and the number theory [35], we have selected the RNS algorithm [36] for FH pattern design, as it has been shown that the RNS can meet all four FH pattern criteria listed above [37].

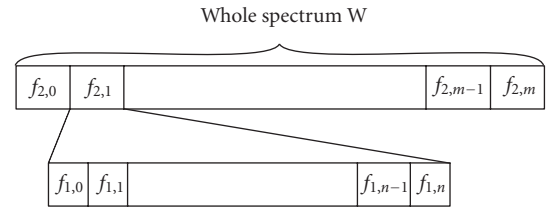


Figure 1: Spectrum allocation for RNS-OFDM.

The RNS is designed based on the Chinese remainder theorem (CRT) [38]. It has recently been used in various wireless communication systems [37, 39]. In proposed FH-OFDM system, RNS defines a rule in choosing positive integers referred to as *moduli* so that all the moduli are pairwise relative primes. In an FH design, the moduli correspond to tones or subcarriers. An individual user's FH address is uniquely and unambiguously represented by a residue sequence determined by the RNS algorithm. We take a numerical example to illustrate the procedure. The total frequency band W can be divided into $F_1 = W/m$ and F_1/n. The carrier frequency of user i at a certain time can be defined by 2-stage FH as $f_i = f_{2,x} * F_1 + f_{1,y}$, $x \in m$, $y \in n$, as shown in Figure 1. For instance, the frequency bandwidth is 5 MHz ($W = 5$ MHz). m and n are selected to be 5 and 10, respectively. Then, we can get $F_1 = 1$ MHz and $F_1/n = 0.1$ MHz. If we take $f_{2,x} = 3$ and $f_{1,y} = 1$ as an example (these numbers are assigned by the RNS), a user's current hopping tone

Table 1: RNS-OFDM hopping pattern assignment (cell 1). In this example, we show that user 2 starts some time later than user 1. Here $F_1 = W/m_1$.

User 1 of cell 1 (FH address 15 $\Leftrightarrow$ (3,2)) $m_1 = 12, m_2 = 13$	User 2 of cell 1 (FH address 21 $\Leftrightarrow$ (9,8)) $m_1 = 12, m_2 = 13$	Accumulator $m_1 = 12, m_2 = 13$
$f_{2,1} * F_1 + f_{1,12}$	—	10, 10
$f_{2,2} * F_1 + f_{1,0}$	—	11, 11
$f_{2,3} * F_1 + f_{1,1}$	$f_{2,9} * F_1 + f_{1,7}$	0, 12
$f_{2,4} * F_1 + f_{1,2}$	$f_{2,10} * F_1 + f_{1,8}$	1, 0
$f_{2,5} * F_1 + f_{1,3}$	$f_{2,11} * F_1 + f_{1,9}$	2, 1

Table 2: RNS-OFDM hopping pattern assignment (cell 2). Here, user 2 starts some time later and $F_1 = W/m_1$.

User 1 of cell 2 (FH address 17 $\Leftrightarrow$ (8,7)) $m_1 = 9, m_2 = 10$	User 2 of cell 2 (FH address 34 $\Leftrightarrow$ (7,4)) $m_1 = 9, m_2 = 10$	Accumulator $m_1 = 9, m_2 = 10$
$f_{2,0} * F_1 + f_{1,8}$	—	1, 1
$f_{2,1} * F_1 + f_{1,9}$	—	2, 2
$f_{2,2} * F_1 + f_{1,0}$	—	3, 3
$f_{2,3} * F_1 + f_{1,1}$	$f_{2,2} * F_1 + f_{1,8}$	4, 4
$f_{2,4} * F_1 + f_{1,2}$	$f_{2,3} * F_1 + f_{1,9}$	5, 5

would be $f_i = f_{2,x} * F_1 + f_{1,y} = 3.1$ MHz. $f_{2,x}$ determines the integer portion and $f_{1,y}$ decides the decimal portion of the subcarrier on a MHz scale or one is for a coarse, and the other is for a fine frequency selection.

In an OFDM system, FH provides frequency diversity that helps reduce fading effects [40, 41, 42]. The salient feature of this type of proposed design is an absence of *in-cell interference* because of OFDM design's orthogonal nature. The orthogonality is preserved even in the presence of multipaths. However, as we have mentioned above, this principle cannot be applied to intercell users because interference can be caused by the subcarriers that are reused from cell to cell. This interference can be significantly reduced across cells if each different user has their own unique hopping pattern. Fast FH leads to improved physical-layer spectral efficiency. A numerical example of FH pattern assignment is shown in Table 1 for cell 1, and in Table 2 for cell 2. A user-specific FH address that identifies each individual mobile node, such as 15 and 21 in Table 1, is assigned by an IP base station once a data traffic channel is allocated. Although the FH pattern of all cell sites follows the RNS, the different cell sites in this example have (i) different moduli, such as $m_1 = 12, m_2 = 13$ for cell 1 and $m_1 = 9, m_2 = 10$ for cell 2, and (ii) initial accumulator settings, such as $(10, 10)$ for cell 1 and $(1, 1)$ for cell 2. As long as the number of active users is less than $m_1 * m_2$, there are no collisions among neighboring cells [37]. We envision an active user who has data to send or receive. The RNS operation proceeds on a residue-by-residue basis; for instance FH address 15 has residues $(3, 2)$, that is $15 \Leftrightarrow (3, 2)$, assuming moduli $m_1 = 12$ and $m_2 = 13$. The subcarrier frequency is determined by both the residue of the FH address and the current accumulator assignment. For instance, $f_{2,1} * F_1 + f_{1,12}$ was chosen because the residue of user 1's FH address is $(3, 2)$ and the current accumulator of cell 1 is $(10, 10)$. By taking the summation of both terms $(3 + 10, 2 + 10)$, we obtain a residue of $(1, 12)$ (equivalent to $f_{2,1} * F_1 + f_{1,12}$) with respect to the moduli of $m_1 = 12$ and $m_2 = 13$ for cell 1. The system is fast hopping since the subcarrier hops every symbol. Notice that the accumulator increases in a circular fashion with modulo (m_1, m_2): for example, we note that $(0, 12)$ follows $(11, 11)$ in Table 1. In the past, FH and OFDM techniques have been used separately in communication systems, such as the 3G and IEEE 802.11a systems. The proposed RNS-OFDM design *effectively* combines the merits of orthogonal and spectrum-spreading designs. In particular, since FH is performed after QAM, we can distinguish different users based on their unique user-signature FH patterns on the receiver side. Those signature patterns can be precomputed and retrieved later by a look-up-table method. Let the pattern of the kth user be expressed as $\vec{a}_k = [a_k(0), a_k(1), \ldots, a_k(L-1)]$. The transmitted symbol x_k^m is then decoded using the kth user's signature code, which can be expressed as

$$\mathbf{Y}_k = \left[y_k(0), y_k(1), \ldots, y_k(L-1)\right] = x_k^m \cdot \mathbf{1} \oplus \vec{a}_k, \qquad (2)$$

where $\mathbf{1}$ is a unit vector of length L, $y_k(l)$ for $1 \le l \le L$, and $\oplus$ denotes modulo-2 addition.

Downlink design

In addition to the RNS-OFDM design, we also propose to use *shared* pilots in the physical layer. Pilots are reserved subcarriers sent by an *IP base station*. They enable a mobile unit to

(i) perform channel estimation for coherent modulation;
(ii) achieve symbol/frame/carrier synchronization;
(iii) perform IP base station identification because adjacent base stations have unique pilot characteristics including RNS moduli (m_1, m_2) and accumulator settings.

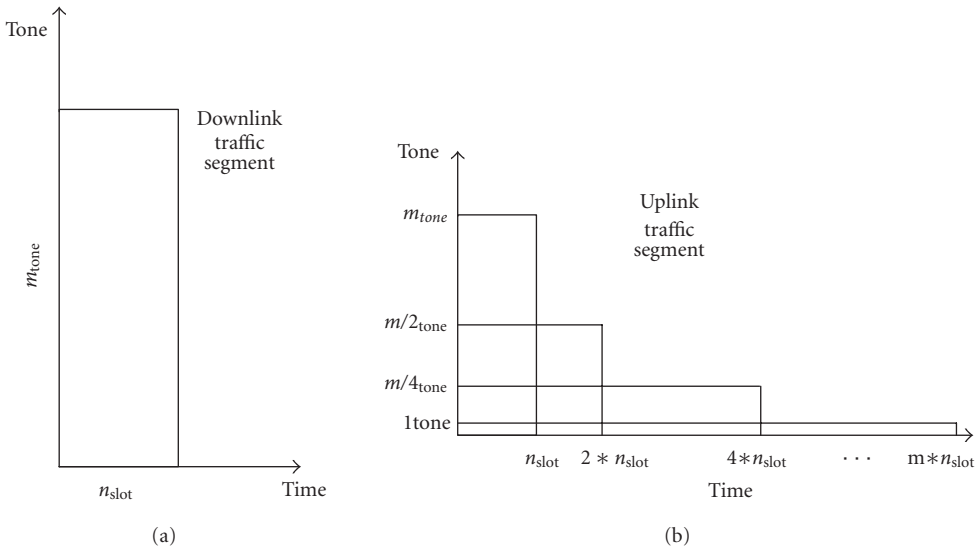


Figure 2: Downlink and uplink transmission options. In order to adapt to time-varying channel conditions and QoS requirements, different uplink transmission options are supported. For downlink transmission, an IP base station can usually transmit at much higher power. Many other design choices also compensate for signal loss, including the antenna array and so forth. However, this is not the case for uplink transmission. We thus choose to support more transmission options for uplink than downlink traffic to better use the physical resources.

DL pilots have specific hopping patterns associated with each IP base station. The pilots are shared by all users within a base station's coverage area. Such a design eliminates pilot overhead for data and control frames. During each symbol period, both pilots and user signals hop from one subcarrier frequency to another, covering the entire transmission bandwidth. Pilots are also "continuous" in the time domain so that mobiles can always listen to pilots, allowing delay-free reception even in a high-Doppler environment [43, 44]. In the proposed system, DL transmission operates in a pure scheduled mode. Only one transmission option is possible for DL traffic, as shown in Figure 2a. During transmission, users have a dedicated traffic channel that is both power controlled and timing controlled for synchronous transmission. The control channel performs in a TDMA fashion, and an active user has designated time slot when transmitting control information, such as ACK/NACK.

Uplink design

In both the frequency and time domains, an UL OFDM symbol is formed in the same way as a DL one. The difference between the domains is that the *UL* physical layer does not use pilots because it is too difficult to establish phase reference for coherent modulation. As a result, a receiver needs to use energy detection during a symbol time, which is very inefficient. Alternatively, in our design, a mobile node stays in one tone for several symbol periods, a so called *quiet time*. Data is differentially modulated in each quiet time using the middle symbol as a reference. For instance, if a quiet time lasts for 5 symbols, then the third symbol is a reference symbol. Since UL traffic does not hop as quickly as DL traffic, an IP base station's receiver can establish a phase reference easily during the quiet period. The tradeoff is that it takes somewhat more time to achieve frequency diversity and interference averaging compared to symbol-by-symbol hopping in DL transmissions. But we have carefully selected the shortest symbol durations and quiet time so that the most effective averaging can still be achieved without long delays. By taking users' QoS requirements and channel conditions into design consideration, we have designed a mobile node to support variable-rate channel coding schemes for transmitting UL signals. The different power-adaptive transmission options shown in Figure 2b are supported for UL data transmission. Signal transmission is *power* and *timing* controlled. Timing control or synchronization ensures that symbols from different mobile nodes arrive at an IP base station at the same time. Mobile terminals periodically transmit wideband multi-subcarrier signals to allow closed-loop timing control. It should be pointed out that, since mobile nodes move much more slowly than electromagnetic waves, round-trip delays change very slowly [9]. Timing control hence can be performed at very slow rates, say every few seconds, constituting a very low resource overhead.

2.2. MAC layer design

In the proposed system, a *segment* is the minimum size of a data unit for transmission over the traffic channels. A DL traffic channel segment consists of m_{tone} subcarriers by n_{slot} time slots, as shown in Figure 2a. Within each time slot, we can have either a data segment or a control segment.

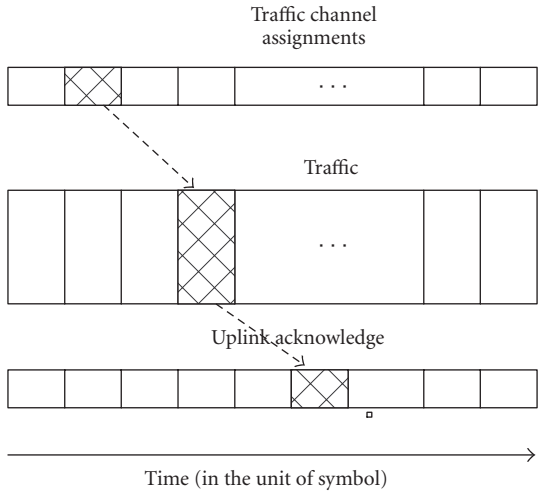


Figure 3: The tight structure for efficient system resource management. Once a traffic channel is assigned, data will be put on the channel; a mobile node then sends an ACK/NACK via the uplink channel. This fast FH allows for quick ACK/NACK message turnaround time. Such a master-slave structure is tightly coupled, providing improved link reliability. Fast ARQ also provides better support for interactive applications.

Out of the total available subcarriers over each symbol period, a certain number of subcarriers are reserved for signaling purposes, such as acknowledging UL data (ACK/NACK) and assigning both UL and DL segments. Unlike the DL traffic channels, each UL traffic channel supports multiple-transmission options, such as option 1 (m_{tone} by n_{slot}) and option 2 ($m/2_{\text{tone}}$ by $2 * n_{\text{slot}}$). Such settings provide efficient and flexible QoS support (refer to our discussion in Section 3.3). Although there are many transmission options, an UL traffic channel uses the same number of physical resources, $m_{\text{tone}} * n_{\text{slot}}$, as the other options.

The MAC layer performs data segmentation (from IP packets to MAC frames) and assembly (from MAC frames to IP packets). It uses a structure that tightly links contention-free requests, assignments, traffic channels, and acknowledgments. This provides the potential for efficient resource management (Figure 3). Traffic segments are instantaneously assigned to *active* users. This scheme not only provides flexibility in traffic multiplexing, which results in high channel utilization, but also facilitates DL multicast. The proposed MAC layer also employs ARQ to provide rapid recovery of corrupted MAC frames and thus increased link reliability. In addition, the proposed FH technique embeds ARQ (ACK/NACK) information in at least one subcarrier (see Figure 3), so that ARQ signaling round-trip time is within the millisecond range. Consequently, the MAC schedule can quickly adjust resource allocation to adapt to time-varying channel conditions. Such features can only be obtained by utilizing fast-hopping multicarrier and cross-layer design at the MAC and physical layers. Overall, fast ARQ minimizes channel latency, maximizes spectrum efficiency, and provides better support of interactive applications.

3. CROSS-LAYER QoS DESIGN

In wireless communications, it is much easier to schedule DL QoS content delivery than to schedule UL traffic because of UL's many-to-one nature. Extensive research has been done on DL scheduler design based on associated QoS requirements [45, 46, 47]. In this paper, we focus on the problem of scheduling multiple classes of DL and UL traffic. In particular, we consider the problem of joint power and bandwidth allocation according to each user's QoS requirements and the wireless channel conditions. This is achieved by cross-layer efforts involving the application, MAC, and physical layers. We introduce a set of QoS-aware MAC states that provides a basis upon which the MAC scheduler selects appropriate transmission formats and packet priorities based on channel conditions and QoS requirements. Next, we will explain our design in detail.

3.1. QoS-aware and power-adaptive MAC states

As we have mentioned in the introduction, the cross-layer design needs built-in supporting mechanisms at different layers. At the MAC layer, we propose a set of QoS-aware MAC states, that is, *high QoS*, *medium QoS*, and *low QoS*, for both UL and DL transmissions. The IP base station dynamically schedules users in different MAC states based on resource availability and a system QoS measure ($\text{QoS} = \text{QoS}_{\text{class}} * \text{QoS}_{\text{stream}}$). The system QoS measure includes general criteria, $\text{QoS}_{\text{class}}$, and application-specific criteria, $\text{QoS}_{\text{stream}}$. In other words, it consists of an *economy* factor (i.e., $\text{QoS}_{\text{class}}$) and a technical factor (i.e., $\text{QoS}_{\text{stream}}$). $\text{QoS}_{\text{class}}$ is determined by a service priority/pricing, such as how much a user pays in monthly service charges, which is independent of applications. A *secondary* consideration is $\text{QoS}_{\text{stream}}$, a traffic class that is determined by applications' characteristics. $\text{QoS}_{\text{stream}}$ is secondary to $\text{QoS}_{\text{class}}$ because a MAC scheduler will grant higher priority to users who pay more (or higher $\text{QoS}_{\text{class}}$). We will discuss $\text{QoS}_{\text{stream}}$ assignment in Table 4 later. QoS information is provided by mobile nodes during an initial registration stage. Although both mobile nodes and an IP base station work together in making decisions about QoS state assignment, the IP base station acts as a "master" in the decision-making process, while mobile nodes provide information only. For instance, a mobile node can request its desired MAC state but the final decision is made by a MAC scheduler at an IP base station. On the mobile node side, there are also three different QoS-aware MAC states: high QoS, medium QoS, and low QoS, corresponding to the states at the IP base station. We can also say that there is a one-to-one correspondence between the state at the IP base station and that at the mobile node. An IP base station uses *QoS identification numbers* to distinguish different users.

We have proposed various MAC states and the transitions between them as illustrated in Figure 4. The MAC states are defined as follows.

(i) *High-QoS-state* users, or the so-called *active users*, can actively send and receive data in this state. This is a high-QoS state in which users have dedicated traffic and control

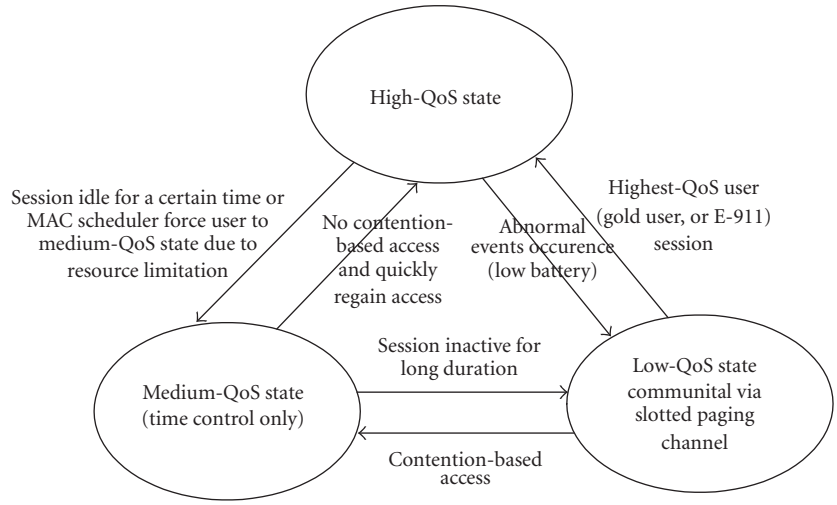


Figure 4: Transition among different QoS states.

channels that are power and timing controlled. A base station assigns traffic segments to any high-QoS user with data to send or receive. It is the MAC scheduler's task to determine how many active users can be in this high-QoS state, and to allocate resources among those users. The MAC scheduler selects an appropriate transmission format and packet priorities for each user depending on channel conditions and users' QoS requirements (refer to the MAC scheduler design below).

A *high-QoS identifier* is a number, from 0 to n_{high}, assigned by an IP base station to uniquely identify an active user. Here, n_{high} is the maximum number of active users a system can accommodate. If there are too many high-QoS users, those users with low $\text{QoS}_{\text{class}}$ and $\text{QoS}_{\text{stream}}$ will be moved to a medium-QoS state or even a low-QoS state. The high-QoS identifier is dynamically allocated to a user when it migrates to the high-QoS state from other states, and is revoked when it migrates out of the high-QoS state.

(ii) *Low-QoS-state* users maintain only basic connectivity, such as a page channel, with an IP base station. Users in the low-QoS state have shared DL *paging* slots so that they can wake up periodically and listen to incoming pages from an IP base station. The low-QoS state is proposed primarily for power saving. A mobile stays in this mode until there is an incoming session via the page channel, or it has data to send.

A *low-QoS identifier* is a number, from 0 to n_{low}, uniquely identifying a low-QoS mobile. Here n_{low} is the maximum number of inactive users that a system can accommodate. In theory, n_{low} can be an infinite number. When a mobile migrates among high-QoS, medium-QoS, and low-QoS states, for instance, an IP base station automatically downgrades the user's MAC state after his mobile has been inactive for a period of time, while the low-QoS identifier remains unchanged during transitions.

(iii) *Medium-QoS state*. This lies between the high-QoS state and the low-QoS state. While in the medium-QoS state users have *contention-free* UL request slots to indicate immediately to an IP base station that they have data to send, users also have shared DL message slots that are timing controlled but are not power controlled in order to save battery power. Our traffic channel is designed to be dedicated to a single user while a message channel can be shared among different users. Unlike active users in the high-QoS state, users in the medium-QoS state have no dedicated traffic channel and must wait for the MAC scheduler to upgrade them into the high-QoS state. One of the benefits of having such a state is power saving—a mobile node will be moved from the high-QoS state to the medium-QoS state if it is idle for a while.

A "session-on" user is the same as an active user, one who is actively transmitting/receiving data. A "session-hold" user was an active user but is now idle either because the user has no on-going traffic or because the MAC scheduler cannot grant the user permission to access physical resources since doing so would prevent it from serving high-QoS users. The *medium-QoS identifier* is a number, from 0 to n_{medium}, which uniquely identifies a medium-QoS mobile. n_{medium} is usually much larger than n_{high} because systems often can accommodate many more session-hold users than session-on users. In our design, we make $n_{\text{medium}} = 10 * n_{\text{high}}$. Here we choose 10 users arbitrarily to indicate that our system can support more "medium-QoS" users than "high-QoS" users. Unlike the high-QoS identifier, the medium-QoS identifier does not change when a mobile migrates between high-QoS and medium-QoS states.

The proposed MAC scheduler takes into account QoS awareness and power adaptation. The physical resource assignments, such as traffic and control channel assignments, that are associated with different MAC states are shown in Table 3. Although both low-QoS and medium-QoS states are designed for power saving, they differ in the amount of time they require to return to the high-QoS state, or their latency to start data transmission. In general, transitions between low-QoS and high-QoS states take more time since a mobile

Table 3: Channels are assigned under different MAC states.

Channel	High QoS	Medium QoS	Low QoS
DL and UL traffic	Yes	No	No
DL SNR report	Yes	No	No
DL and UL assignments and power control	Yes	No	No
Acknowledgment for DL and UL traffic channels	Yes	No	No
Syncronization channel	Yes	Yes	No
Timing control	Yes	Yes	No
Traffic channel request	No	Yes	No
Traffic channel grant	No	Yes	No
Fast paging	No	No	Yes
Slow paging	No	No	Yes
Access request	No	No	Yes
Access confirmation	No	No	Yes

has fewer physical resources in the low-QoS state (refer to Table 3). For example, an active user will enter the medium-QoS state before reaching the low-QoS state if the user's session is idle for a certain period of time. The reason not to put an active user directly into the low-QoS state is to enable the user to quickly regain physical resources if the idle time is relatively short. All MAC state transitions are managed jointly by a MAC scheduler at an IP base station and one at a mobile node. The MAC state is maintained at both IP base station and mobile nodes. During mobile-IP handoff (from one base station to another), a user's MAC state is transferred from its old base station to the new one (detailed mobile IP design information is beyond the scope of this paper). The new IP base station will decide whether to keep the user's old MAC state or to reinitiate a new MAC state. For instance, a high-QoS user can keep original the MAC state assignment if the new base station has enough physical resources to keep the user in the high-QoS state. Otherwise, the user will be downgraded to the medium-QoS state.

3.2. Assignment of MAC states

In this subsection, we will address the problem of how to assign MAC states for each mobile. QoS provision and support remain technical challenges in wireless communications. Previous work has generally adopted simple QoS parameters, such as the fixed target signal-to-interference-plus-noise ratio (SINR) [13]. In previous research transmissions were also designed based on current channel conditions [9, 48]. In addition, existing schemes cannot adaptively adjust according to users' QoS satisfaction levels. In the proposed system, we designate control channels, as shown in Table 3, for users to give feedback on their QoS satisfaction. The system thus can continuously adjust users' MAC states and allocate physical resources to best fit users' QoS needs. Here we propose a general framework for resource allocation based on users' MAC states. The proposed framework employs a diverse set of objective functions and QoS requirements. Specifically, we will consider the following two factors: service priority/pricing and traffic class.

Service priority/pricing (QoS_{class})

Price or service charge plays an essential role in resource allocation [49, 50]. A user who pays a higher monthly fee, in general, is entitled to better QoS support. We therefore rank users in terms of "gold," "silver," and "bronze" service priorities. The service priority allows the MAC scheduler to differentiate between users, particularly when a resource conflict occurs. For instance, a system initially contains one gold user and one silver user, running the same type of application. Both users are assigned the high-QoS MAC state initially. When a third user initiates an emergency 911 call that qualifies for Gold service priory, all three users begin competing for limited physical resources. Assuming that the system can only support two or fewer high-QoS users, the MAC scheduler will temporarily downgrade a low-QoS_{class} user, the silver user in this case, from the high-QoS state to the medium-QoS state until the third user finishes using the E-911 service.

Traffic class (QoS_{stream})

The characteristics of data traffic or applications also play an important role in MAC state assignments. In our design, the minimum resolution of traffic is called a *MAC stream*. Corresponding transmission requirements are characterized by factors such as the BER, segment loss (SL), segment order (SD), and delay (DEL). An example of MAC stream assignments is shown in Table 4. Here, the "stream QoS-ID" is associated with a certain service with specific QoS factors. A higher QoS-ID indicates that a stream is more important. For instance, an E-911 session is more important than voice and data services. All entries in Table 4 are stored in a look-up table, and service providers or operators can modify the parameters to reflect the importance of each application, or even expand the table to include additional network parameters. This information will ultimately be mapped onto a set of QoS_{stream}. Different MAC streams, or *marked packets*, can have different packet lengths, with stream QoS-IDs embedded in the packet headers.

Table 4: Example of MAC stream assignments. QoS_{stream}, for instance, is in the range of [15, 1]: 15 is the highest QoS-MAC stream while 5 is the lowest one (4 to 1 are reserved for future use). Network performance, in terms of BER, SL, SO, and DEL, is scored with a value from 4 to 1: 4 stands for the "most important," while 1 stands for the "least important." The setting can be adjusted based on channel conditions. Again, we show just one design example in this table.

Stream QoS-ID (QoS_{stream})	Description	BER	SL	SO	DEL
15	E-911 session	4	4	4	4
14	Layer 2 control	4	4	4	4
13	Layer 3 control	4	4	4	4
12	Circuit voice (G729)	3	3	4	4
11	VoIP (G711-coded)	3	4	3	4
10	Stream video (H263)	3	3	3	4
9	Interactive data	3	4	4	3
8	Multicast RTP	3	3	3	3
7	Internet control	4	4	4	4
6	Internet data	3	4	4	2
5	Network management (OA&M) traffic	2	2	3	3

Having defined the QoS_{class} and QoS_{stream}, we can determine a user's MAC state using the following QoS criterion:

$$QoS = QoS_{class} * QoS_{stream}. \quad (3)$$

For example, on a revenue basis, we define $QoS_{class} = 3$ (gold user), $QoS_{class} = 2$ (silver user), $QoS_{class} = 1$ (bronze user). QoS_{stream} is listed in Table 4. Note that our proposed system supports either a native IP mode or a point-to-point protocol- (PPP-)like mode for non-IP traffic.

3.3. MAC scheduler

Users can actively send and receive traffic only in the high-QoS state. However, there is resource contention when multiple high-QoS users compete for channel resources. A conventional scheduler is developed solely for "centralized allocation," where an IP base station gathers all user information, that is, channel response and backlog size, and decides which user is to be served [51, 52]. The disadvantage of the centralized approach lies in the difficulty of getting channel response for all users and IP base stations across multiple cells. Computational complexity also grows significantly when the number of users increases. A "distributed scheduler," on the other hand, allows each user to optimize resource usage individually. In [53, 54], a pricing system was developed, in which users would have to pay higher prices when demanding more physical resources. The disadvantage of this system is that there are many local optimization points but global system optimization is difficult to achieve. Considering the tradeoff between performance and complexity, a distributed scheduler is suitable for systems with large numbers of users. The previously mentioned centralized and distributed schedulers were designed to maximize system throughput but do not take any QoS requirements into design consideration. We propose a "scheduled" approach to achieve QoS-aware and power-adaptive transmission.

In the proposed QoS framework, network level QoS parameters include BER, SL, SO, and DEL. Different applications have different QoS network requirements (see Table 4). The goal of the MAC scheduler is to select appropriate users' transmission power/format and packet priorities based on present channel conditions and users' QoS requirements. For instance, real-time applications are more sensitive to service DEL than SL, while non-real-time applications are more sensitive to SO and SL. Our cross-layer design adjusts MAC states and transmission formats to observe and respond to channel variations based on the QoS vector (BER, SL, SO, DEL). Network QoS is computed as follows:

$$QoS_{network} = w_1 * BER + w_2 * SL + w_3 * SO + w_4 * DEL \quad \forall\ w_1 + w_2 + w_3 + w_4 = 1. \quad (4)$$

Here the normalized weighting vector is $\vec{w} = (w_1, w_2, w_3, w_4)$. For the example in Table 4, we select weights equally among BER, SL, SO, and DEL. We can further adjust the $\vec{w}$ to set priorities differently among BER, SL, SO, and DEL when computing the $QoS_{network}$.

Network parameters, combined with QoS_{class} and QoS_{stream}, constitute a multicriteria decision for a given user. Cost can be measured by $Cost = QoS_{class} * QoS_{stream} * QoS_{network}$. We can thus obtain a unique cross-layer cost measurement, which reflects each user's QoS as well as ongoing application and current network conditions. The MAC scheduler determines the power and coding rate required to transmit a given frame with a specific amount of reliability based on this multicriteria cost measurement, Cost. The architecture of the MAC, scheduler is shown in Figure 5. The MAC scheduler computes the real cost value, $Cost_{real}$, based on feedback from QoS monitors at the physical, MAC, and network layers. Then, it compares the current $Cost_{real}$ with the desirable cost measurement Cost. By adjusting the transmission power of both the UL and DL traffic channels, the scheduler is able to minimize the difference between the desired cost and real cost measurements for each user,

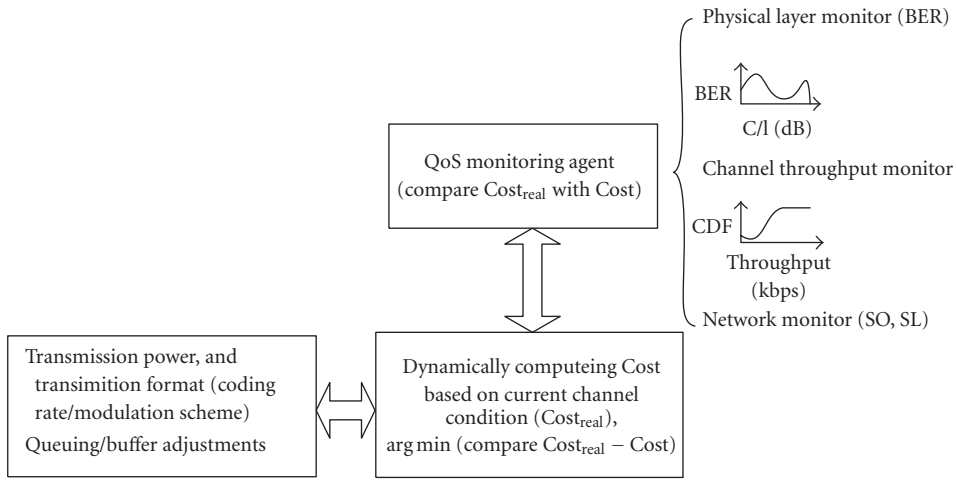


Figure 5: Architecture of a MAC scheduler (applicable to both uplink and downlink transmissions).

$\arg\min \|\text{Cost}_{\text{real}} - \text{Cost}\|$. Although the power control is applied to both DL and UL traffic, the MAC scheduler operates under the constraint that the total transmit power in each slot cannot exceed a fixed peak power value. If power adjustment cannot improve BER performance effectively, the MAC scheduler will select a different transmission format for a given data frame, such as a coding rate and modulation scheme—from aggressive approaches, like Turbo code and 64-QAM, to conservative approaches, like 5/6 convolutional code and QPSK.

In addition to conventional power and transmit rate control, our novel design supports unique UL transmission options to achieve scalable QoS performance. An IP base station can usually transmit at much higher power levels and can employ other sophisticated means for data transmission. Therefore, we have designed DL transmission options to be much simpler than those for UL. As a result, we only support one transmission option as shown in Figure 2, and it operates in a pure scheduled mode with fixed transmission time. For UL transmission, however, we support different transmission options besides rate and the power controls, in order to take spectrum efficiency and QoS awareness into design consideration. For instance, when a user experiences bad channel conditions, the MAC scheduler will use option 4 instead of option 1, as shown in Figure 2b, for UL data transmission because the user probably cannot receive all tones correctly after adjustments have been made to power and transmission rates. Under such conditions, physical resources can be allocated to other users that need them. To accommodate different channel conditions, we propose to use multiple UL transmission options with variable frequencies and time allocation for UL traffic. Each transmission format can employ a different number of OFDM tones and symbols. For example, the MAC scheduler can use the following four transmission formats/options in our QoS-aware design: (1) m_{tone} tones by n_{slot} time slots, (2) $m/2_{\text{tone}}$ tones by $2 * n_{\text{slot}}$ time slots, (3) $m/4_{\text{tone}}$ tones by $4 * n_{\text{slot}}$ time slots, and (4) 1 tone by $m * n_{\text{slot}}$ time slots, as shown in Figure 2b. Unlike single-carrier CDMA design, our design can group subcarriers (tones) in a different manner: according to performance. Options 1 and 2 are suitable for users with good instantaneous channel conditions while options 3 and 4 provide stable transmission over a longer period of time. It is not wise, for instance, to assign a large number of OFDM tones (option 1) to a given user under relatively bad channel conditions. This simply wastes physical resources and leads to overall system performance degradation because other users could be deprived of needed resources. The MAC scheduler at a mobile node will select option 1 or option 2 by assigning more tones over a short time period for data transmission, assuming that good channel conditions usually do not last long. The QoS scheduler jointly considers MAC states and physical layer conditions, and performs cross-layer interaction. Otherwise, we would have to design each layer separately for the worst possible conditions, which leads to resource inefficiency.

If all means mentioned above fail, the MAC scheduler will adjust queuing/buffer either (i) by throwing away some symbols for real-time applications to avoid further delay, or (ii) by buffering information in reserved memory for non real-time applications so that buffered information can be transmitted later when the network conditions get better.

Table 5: Basic system parameters.

Carrier frequency	Up to 5 GHz
Bandwidth	1.25 MHz UL, 1.25 MHz DL
Number of subcarriers	113
FFT window length	128 samples
Cyclic prefix	16 samples
DL peak rate	2.7 Mbps
UL peak rate	817 kbps

4. SIMULATION RESULTS AND DISCUSSIONS

Basic system parameters are listed in Table 5. Out of the 113 subcarriers available over each symbol period, 36 subcarriers

Table 6: Performance at different locations within a cell.

Location	Distance from an IP base station (miles)	SNR (dB)
A	0.5	About 20
B	1.4	About 20
C	2.0	About 15
D	1.6	About 10

Table 7: User throughput by location and applications (kbps).

Throughput	FTP UL	FTP DL	128 kbps media	Web page
Location A	354	2 470	142	32
Location B	225	1 910	138	27
Location C	142	1 280	141	23
Location D	47	675	136	29

will be reserved for signaling, such as for acknowledging DL data and dedicated control signals. The test system consists of a cluster of several wireless routers supported by a managed IP network. Testing was conducted in a 700 MHz guard band over a 1.25 MHz channel. Next, we will discuss the baseline tests followed by related QoS measurements.

Baseline tests with flat QoS requirements

This test set serves as a baseline for all other tests in order to calibrate the potential throughput of representative applications. The tests are performed under single-user transmission conditions without competing physical resources: QoS_{class} does not matter under such test conditions. We have measured the system's performance under a wide variety of channel conditions. To execute these tests, we located four different SNR testing spots, from location A to location D within one cell's coverage, as shown in Table 6. In addition, we selected four different types of services or sessions to represent typical wireless content delivery.

(1) FTP UL: an FTP service uploaded a 10 MB file.
(2) FTP DL: an FTP service downloaded 50 distinct 2 MB files.
(3) 128 kbps media: a 128 kbps real-time media stream was played from a server within a core network.
(4) Web page: a web page of 205 KB was periodically refreshed throughout the baseline test. An effort was made to have this event happen roughly every 120 seconds.

Throughput results are summarized in Table 7. Under good SNR conditions, such as at location A listed in Table 6, FTP DL works at an average throughput rate of 2.4 Mbps with instantaneous bursts of throughput up to 2.6 Mbps, as shown in Figure 6. The 128 kbps media stream can play smoothly regardless of location because it does not attempt to compete for more bandwidth than its assigned encoding rate, 128 kbps. As shown in Figure 7, the application's

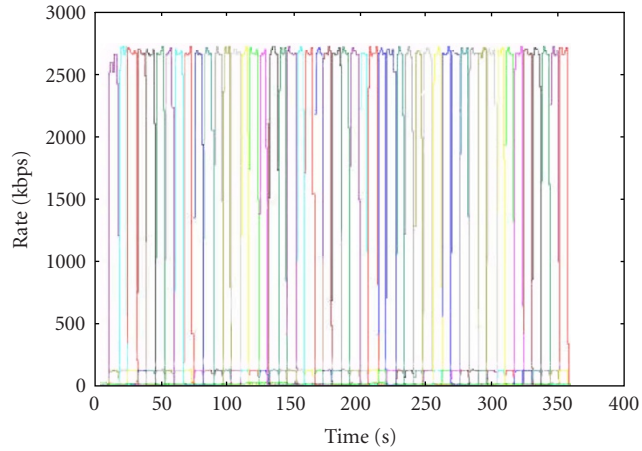


Figure 6: Throughput of the FTP download service (FTP DL) at location A.

throughput was approximately 140 kbps, representing a real-time multimedia service with total bandwidth requirements including application, headers, and other control information. For streaming media applications based on the user datagram protocol (UDP), minimal packet loss was experienced. Media streams thus can be played smoothly at all times, without relying heavily on the internal caching mechanisms of host software. Web traffic demonstrated its characteristic "bursty" nature. Test instructions required periodic (at intervals of about 120 seconds) reloads using a new 205 KB image for each new trail. This ensured that the image was never pulled from the local cache of a host computer. As shown in Figure 8, users downloading 205 KB images at random times will produce very brief spikes of throughput, which approaches the channel capacity. When the download is completed, system resource demand immediately returns to zero. For applications like web surfing and file transfers, the service rate is subject to the limitations of TCP, for example, a slow start. We observed a very limited number of packet losses and TCP retransmissions.

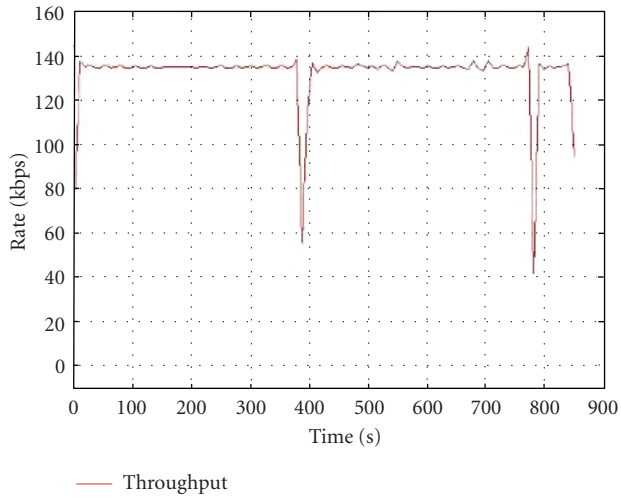


Figure 7: Media stream of 128 kbps. Dips around $t = 400$ seconds and $t = 800$ seconds indicate pauses between two audio pieces or between two songs.

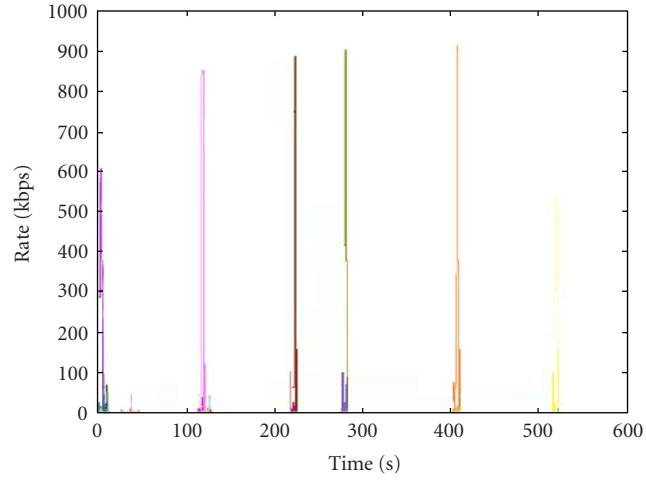


Figure 8: Multiple downloads of 205 KB web image.

Cross-layer example with gold and bronze QoS services

This test set was designed to demonstrate that cross-layer design provides better QoS support. We compared the performance of a gold user, that is, $\text{QoS}_{\text{class}} = 3$, with a bronze user, that is, $\text{QoS}_{\text{class}} = 1$, along with other lower QoS background users competing for the remaining bandwidth. To execute this test case, a single load mobile was placed in an excellent SNR environment (SNR = 23 dB). This load mobile, at location A, listed in Table 6, was used to create a controlled load on the cell site. The amount of load was controlled by a MAC QoS scheduler at an IP base station. In this QoS test case, each user traveled from one location to another while performing an assigned task, as summarized in Table 8.

These QoS tests illustrate how cross-layer QoS assignments can impact an end user's experience as well as the experience of other users in the system. In the first test, all users registered in a cell site have equal QoS priority assignments, and they simultaneously request physical resources. Figure 9 depicts the application level throughput of a gold user (namely, mobile 1) running the FTP download service. In this figure, the oscillations of throughput after $t = 100$ seconds show that the MAC scheduler dynamically adjusted the physical resources while all mobile nodes moved around. Mobile 1 had an initial throughput around 2.5 Mbps. Then, another user (mobile 2) joined the network and started to download a 128 kbps audio stream. Mobile 2 operates under poor SNR conditions and was able to maintain sustainable throughput at 160 kbps (see Figure 10). The different UL transmission options listed in Figure 2, along with power control and channel coding schemes, were utilized to offset channel variations. This provides a good example of cross-layer design because without assistance from different layers, the system will work under the worst possible conditions, leading to the inefficient use of physical resources. If there were only two active users in the system, there would be no resource competition since the resources would be perfectly distributed to these two users. Mobile 1 could thus maintain throughput at 2.6 Mbps (refer to Figure 9). After 105 seconds ($t = 105$ seconds), mobile 3, performing a 100 MB FTP as shown in Figure 11, entered the system. The MAC scheduler effectively assigned resources so that all three users were supported according to their QoS assignments. Although, in this case all users had the same QoS class, they had different $\text{QoS}_{\text{stream}}$, so the FTP user got the most bandwidth.

The above demonstration illustrates the system's ability to balance diverse users' needs. User 2 continues to receive a minimal 128 kbps audio stream, while load 1 and the FTP user 3 share the remaining bandwidth. Both user 1 and user 3 thus achieve throughput at 1.2 to 1.3 Mbps because the throughput of mobile 1 dropped from 2.7 Mbps to 1.4 Mbps after 3 joined at $t = 105$ seconds; for further details, refer to Figure 9. In this example, the aggregate system throughput of three mobile nodes fluctuated between 2.7 and 2.8 Mbps, very close to the theoretical maximum capacity. When mobile 3 finished its task at $t = 905$ seconds, the MAC scheduler instantly reallocated the excess bandwidth back to mobile user 1, and its throughput bounced back to 2.6 Mbps, as shown in Figure 9.

Next, we assign different QoS classes, $\text{QoS}_{\text{class}}$, to the mobiles. To accomplish this, mobile 1 in this case was assigned bronze QoS status, and we assumed that the user would never receive more than 150 kbps of resources regardless of cell capacity. Figure 12 depicts how the throughput dropped, as a step function, once the QoS class was changed from gold to bronze at $t = 80$ seconds. The throughput available to that user was reduced from 2.6 Mbps to 150 kbps, reflecting the new QoS assignment. By examining Figure 13, we observe that the throughput of the streaming media session was always maintained at around 128 kbps. However, Figure 14 shows that the gold user 3 experienced

Table 8: Task assignments for QoS tests.

Mobile	Activity
File transfer (moving from location B to location C)	An FTP service downloaded 50 distinct 2 MB files.
Web surfing (moving from location C to location D)	A web page of 205 KB was periodically refreshed throughout the test.
Streaming media (moving from location D to location A)	A 128 kbps media stream was played.

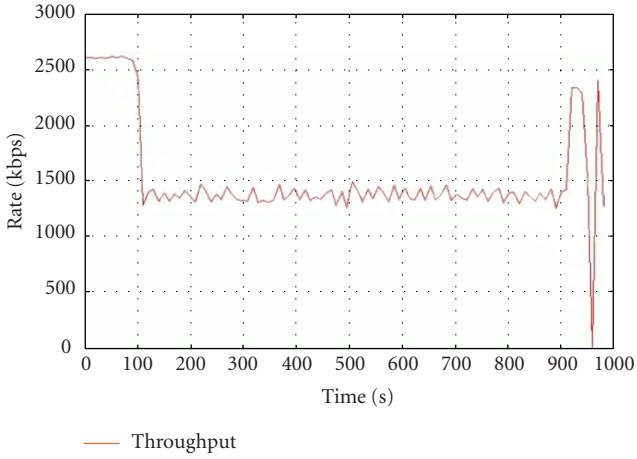


Figure 9: QoS gold user (mobile 1) running an FTP DL.

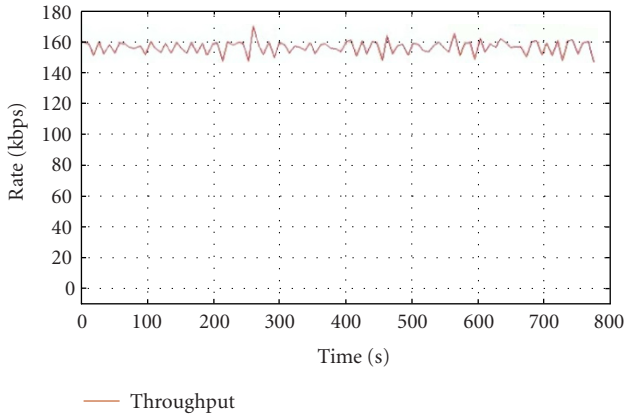


Figure 10: Streaming audio user (mobile 2) moving from location D to location A.

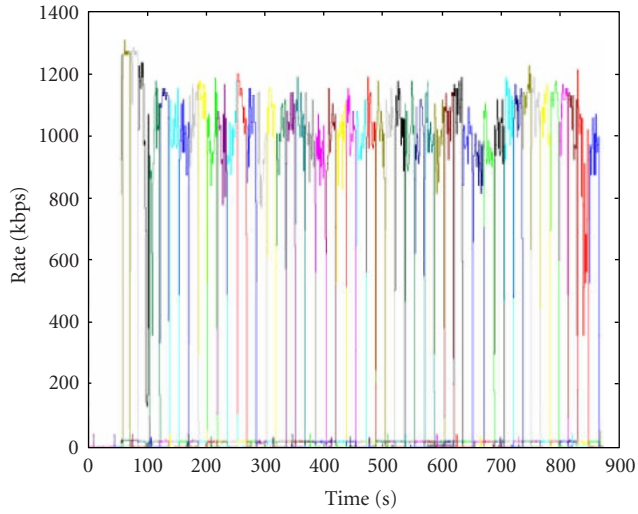


Figure 11: FTP transform (mobile 3) moving from location B to location C.

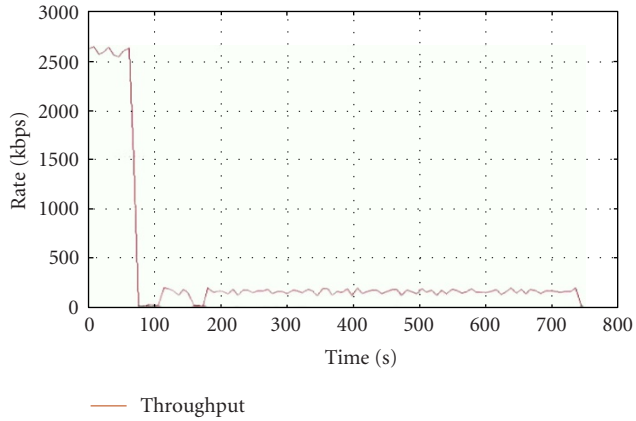


Figure 12: QoS change from gold to bronze for mobile user 1 running an FTP DL.

much higher throughput at 2.2–2.5 Mbps, which is dramatically different from the previous example.

5. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

In this paper, we propose a cross-layer design for QoS-aware content delivery. Central to our proposed cross-layer design is the concept of *adaptation*. We propose a cross-layer platform, utilizing the application, MAC, and physical layers, for wireless QoS content delivery. The proposed QoS-awareness scheduler and power adaptation scheme organize the behavior of the lower physical layer for resource efficiency.

We also provide extensive test results in this paper to illustrate that the cross-layer design improves QoS satisfaction. In essence, most of the simulations focus on presenting results that show that a user's QoS assignments can be more efficiently managed with a cross-layer design. The claim that the proposed design leads to a more efficient use of resources has not yet been substantiated quantitatively. In our future work, we plan to quantitatively measure efficiency and to compare our results with other approaches, such as a design

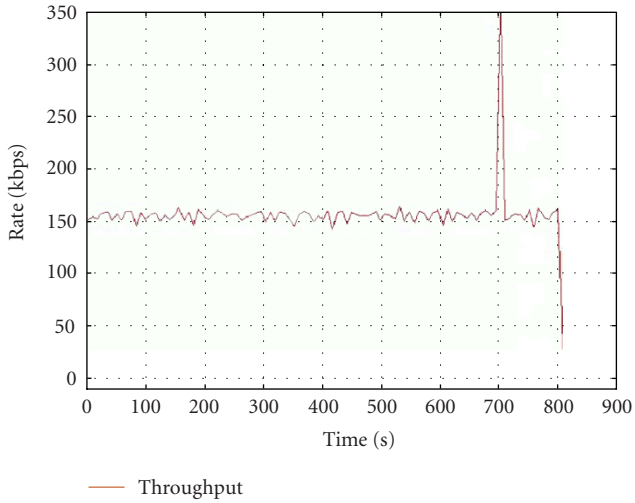


Figure 13: Streaming audio user (mobile 2) moving from location D to location A.

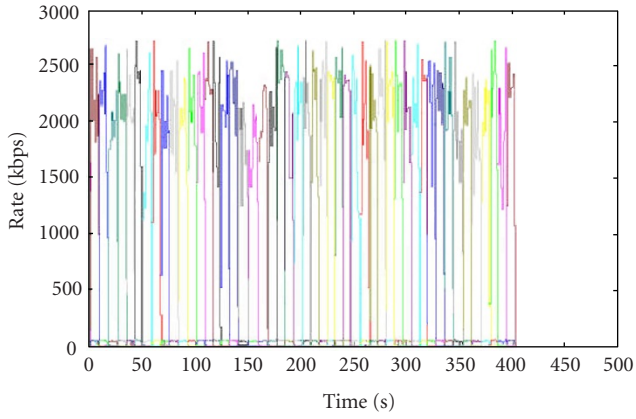


Figure 14: FTP transform (mobile 3) moving from location B to location C.

without cross-layer assignment, and a design that applies changes in the system parameters (e.g., changing the design in Table 4).

REFERENCES

[1] S. Uskela, "Key concepts for evolution toward beyond 3G networks," *IEEE Wireless Communications*, vol. 10, no. 1, pp. 43–48, 2003.

[2] M. Kangas, H. Krummel, M. Mahkonen, et al., "Functional partitioning of all-IP mobile devices," in *Proc. IASTED International Conference Communications, Internet and Information Technology (CIIT '02)*, pp. 376–129, St. Thomas, Virgin Islands, USA, November 2002.

[3] 3GPP, "Physical layer aspects of UTRA high speed downlink packet access," TR 25.848, v0.6.0, 2001.

[4] W. Yu-Ming and L. Shu, "A modified selective-repeat type-II hybrid ARQ system and its performance analysis," *IEEE Trans. Communications*, vol. 31, no. 5, pp. 593–608, 1983.

[5] G. Fairhurst and L. Wood, "Advice to link designers on link automatic repeat reQuest (ARQ)" , RFC 3366, August 2002, ftp://ftp.rfc-editor.org/in-notes/rfc3366.txt.

[6] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE 51st Vehicular Technology Conference (VTC '00)*, vol. 3, pp. 1854–1858, Tokyo, Japan, 2000.

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

[8] D. Tse, "Forward link multiuser diversity through proportional fair scheduling," *Bell Laboratories Journal*, August 1999.

[9] R. D. Yates, "A framework for uplink power control in cellular radio systems," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 7, pp. 1341–1347, 1995.

[10] B. Suard, A. Naguib, G. Xu, and A. Paulraj, "Performance of CDMA mobile communication systems using antenna arrays," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '93)*, vol. 4, pp. 153–156, Minneapolis, Minn, USA, April 1993.

[11] W. Yu and J. Cioffi, "FDMA capacity of Gaussian multiple-access channels with ISI," *IEEE Trans. Communications*, vol. 50, no. 1, pp. 102–111, 2002.

[12] S.-L. Kim, Z. Rosberg, and J. Zander, "Combined power control and transmission rate selection in cellular networks," in *Proc. IEEE 50th Vehicular Technology Conference (VTC '99)*, vol. 3, pp. 1653–1657, Amsterdam, The Netherlands, September 1999.

[13] F. Rashid-Farrokhi, L. Tassiulas, and K. J. R. Liu, "Joint optimal power control and beamforming in wireless networks using antenna arrays," *IEEE Trans. Communications*, vol. 46, no. 10, pp. 1313–1324, 1998.

[14] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Transactions on Information Theory*, vol. 39, no. 2, pp. 466–478, 1993.

[15] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1277–1294, 2002.

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

[17] B. Girod and N. Färber, "Wireless video," in *Compressed Video over Networks*, pp. 465–512, Marcel Dekker, New York, NY, USA, 2000.

[18] B. Girod, M. Kalman, Y. J. Liang, and R. Zhang, "Advances in channel-adaptive video streaming," in *Proc. IEEE International Conference on Image Processing (ICIP '02)*, vol. 1, pp. I9–I12, Rochester, NY, USA, September 2002.

[19] H. Liu and M. E. Zarki, "Adaptive source rate control for real-time wireless video transmission," *Mobile Networks and Applications*, vol. 3, no. 1, pp. 49–60, 1998.

[20] S. Aramvith, I.-M. Pao, and M.-T. Sun, "A rate-control scheme for video transport over wireless channels," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 11, no. 5, pp. 569–580, 2001.

[21] C.-Y. Hsu, A. Ortega, and M. Khansari, "Rate control for robust video transmission over burst-error wireless channels," *IEEE Journal on Selected Areas in Communications*, vol. 17, no. 5, pp. 756–773, 1999.

[22] S. Cen, P. C. Cosman, and G. M. Voelker, "End-to-end differentiation of congestion and wireless losses," in *Proc. SPIE Multimedia Computing and Networking (MMCN '02)*, vol. 4673 of *Proceedings of SPIE*, pp. 1–15, San Jose, Calif, USA, January 2002.

[23] N. K. G. Samaraweera, "Non-congestion packet loss detection for TCP error recovery using wireless links," *IEE Proceedings on Communications*, vol. 146, no. 4, pp. 222–230, 1999.

[24] S. Biaz and N. Vaidya, "Discriminating congestion losses from wireless losses using inter-arrival times at the receiver," Tech. Rep. 98-014, Computer Science Department, Texas A&M University, College Station, Tex, USA, June 1998.

[25] B. S. Bakshi, P. Krishna, N. H. Vaidya, and D. K. Pradhan, "Improving performance of TCP over wireless networks," Tech. Rep. 96-014, Computer Science Department, Texas A&M University, College Station, Tex, USA, May 1996.

[26] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, "A comparison of mechanisms for improving TCP performance over wireless links," *IEEE/ACM Transactions on Networking*, vol. 5, no. 6, pp. 756–759, 1997.

[27] T. Goff, J. Moronski, D. S. Phatak, and V. Gupta, "Freeze-TCP: A true end-to-end TCP enhancement mechanism for mobile environments," in *Proc. 19th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM '99)*, vol. 3, pp. 1537–1545, Tel Aviv, Israel, March 2000.

[28] S. Shakkottai, T. S. Rappaport, and P. C. Karlsson, "Cross-layer design for wireless networks," Tech. Rep. TR-2003-04-00001, Wireless Networking and Communications Group, Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, Tex, USA, 2004, http://www.wncg.org.

[29] T. Lang, Q. Zhao, and G. Mergen, "Multipacket reception in random access wireless networks: from signal processing to optimal medium access control," *IEEE Communication Magazine*, vol. 39, no. 11, pp. 108–112, 2001.

[30] QualComm, *1xEV-DO*, 2001, http://www.qualcomm.com/technology/1xev-do/.

[31] H. Chekal and Y. Han, "An efficient power allocation scheme in HDR systems," in *Proc. IEEE 54th Vehicular Technology Conference (VTC '01)*, vol. 4, pp. 2197–2201, Atlantic City, NJ, USA, October 2001.

[32] W. Chung, H. W. Lee, and J. Moon, "Downlink capacity of CDMA/HDR," in *Proc. IEEE 53th Vehicular Technology Conference (VTC '01)*, vol. 3, pp. 1937–1941, Rhodes, Greece, May 2001.

[33] M. K. Simon, J. K. Omura, R. A. Scholtz, and B. K. Levitt, *Spread Spectrum Communications*, Computer Science Press, Rockville, Md, USA, 1985.

[34] S. B. Wicker and V. K. Bhargava, *Reed-Solomon Codes and Their Applications*, John Wiley & Sons, New York, NY, USA, 1999.

[35] S. V. Maric and E. L. Titlebaum, "A class of frequency hop codes with nearly ideal characteristics for use in multiple-access spread-spectrum communications and radar and sonar systems," *IEEE Trans. Communications*, vol. 40, no. 9, pp. 1442–1447, 1992.

[36] M. A. Soderstrand, W. K. Jenkins, G. A. Jullien, and F. J. Taylor, *Residue Number System Arithmetic: Modern Applications in Digital Signal Processing*, IEEE Press, Piscataway, NJ, USA, 1986.

[37] L.-L. Yang and L. Hanzo, "Residue number system assisted fast frequency-hopped synchronous ultra-wideband spread-spectrum multiple-access: a design alternative to impulse radio," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 9, pp. 1652–1663, 2002.

[38] R. W. Watson and C. W. Hastings, "Self-checked computation using residue arithmetic," *Proceedings of the IEEE*, vol. 54, pp. 1920–1931, 1996.

[39] L.-L. Yang and L. Hanzo, "Residue number system arithmetic assisted M-ary modulation," *IEEE Communications Letters*, vol. 3, no. 2, pp. 28–30, 1999.

[40] A. J. Viterbi, *CDMA: Principle of Spread-Spectrum Communications*, Addison-Wesley Publishing Company, Reading, Mass, USA, 1995.

[41] M. D. Yacoub, *Foundations of Mobile Radio Engineering*, CRC Press, Boca Raton, Fla, USA, 1993.

[42] T. S. Rappaport, *Wireless Communications: Principle and Practice*, Prentice-Hall, Upper Saddle River, NJ, USA, 1996.

[43] R. Steele, *Mobile Radio Communications*, Pentech Press, London, UK, 1992.

[44] W. C. Jakes, *Microwave Mobile Communications*, IEEE Press, New York, NY, USA, 1994.

[45] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: the exponential rule," in *Analytic Methods in Applied Probability: In Memory of Fridrikh Karpelevich* Yu. M. Suhov, Ed., vol. 207 of *Amer. Math. Soc. Transl. Ser. 2*, pp. 185–201, American Mathematical Society, Providence, RI, USA, 2002.

[46] M. Assaad, B. Jouaber, and D. Zeghlache, "Effect of TCP on UMTS-HSDPA system: Services performance and system capacity," in *Proc. IEEE Global Telecommunications Conference (GLOBECOM '04)*, Dallas, Tex, USA, November–December 2004.

[47] H. Viswanathan and K. Kumaran, "Rate scheduling in multiple antenna downlink wireless systems," in *Proc. 39th Annual Allerton Conference on Communications and Control and Computing (Allerton '01)*, Monticello, Ill, USA, October 2001.

[48] A. Yener, R. Yates, and S. Ulukus, "Interference management for CDMA systems through power control, multiuser detection, and beamforming," *IEEE Trans. Communications*, vol. 49, no. 7, pp. 1227–1239, 2001.

[49] D. Clark, *Internet Cost Allocation and Pricing*, MIT Press, Cambridge, Mass, USA, 1997.

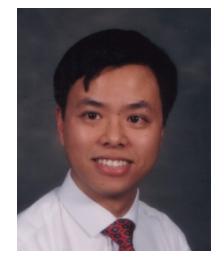
[50] K. Nichols, V. Jacobson, and L. Zhang, "A two-bit differentiated services architecture for the Internet," Internet Draft, November 1997, ftp://ftp.ee.lbl.gov/papers/.

[51] Z. Han and K. J. R. Liu, "Joint adaptive power and modulation management in wireless networks with antenna diversity," in *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM '02)*, pp. 278–282, Rosslyn, Va, USA, August 2002.

[52] S. Kandukuri and S. Boyd, "Optimal power control in interference-limited fading wireless channels with outage-probability specifications," *IEEE Transactions on Wireless Communications*, vol. 1, no. 1, pp. 46–55, 2001.

[53] C. Saraydar, N. Mandayam, and D. Goodman, "Efficient power control via pricing in wireless data networks," *IEEE Trans. Communications*, vol. 50, no. 2, pp. 291–303, 2002.

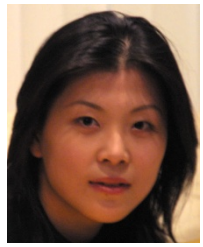
[54] D. Goodman and N. Mandayam, "Power control for wireless data," *IEEE Personal Communications*, vol. 7, no. 2, pp. 48–54, 2000.



Jie Chen received his M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park. He is currently an Assistant Professor in the Division of Engineering, Brown University. Dr. Chen's research interests include wireless communications and networking, genomic signal processing, nanoscale device and circuit design, and nanotechnology for interdisciplinary bio-medical applications. He has received the IEEE Distinguished Lecturer Award for circuit and system society, and division award from Bell Labs. He has been invited as the speaker in different conferences and workshops. Dr. Chen has published about 40 scientific papers in refereed journals

and conference proceedings, and has coauthored two books: *Design of Digital Video Coding Systems: A Complete Compressed Domain Approach* and *Genomic Signal Processing and Statistics*, and two book chapters: "A probabilistic-based design for chapters nanoscale computation" and "Cancer genomics, proteomics, and clinic applications." Currently, he is the Associate Editor for the IEEE Signal Processing Magazine. He has served as an Associate Editor of the IEEE Transactions on Multimedia and EURASIP Journal on Applied Signal Processing. He is an IEEE Senior Member of the Signal Processing Society. He also serves as the Technical Program Cochair of the IEEE Genomic Signal Processing and Statistics Workshop 2005, and Chair-Elect of Life-Science Systems and Applications Technical Committee, IEEE Circuits and Systems Society.

Tiejun Lv received the B.S. degree from Southwest Jiaotong University, Sichuan Province, China, in 1991, and the M.S. and Ph.D. degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC) in 1997 and 2000, respectively. From January 2001 to December 2002, he was a Postdoctor in the Department of Automation, Tsinghua University, Beijing, China. Since April 2003, he has been an Associate Professor in the School of Information Engineering, Beijing University of Posts and Telecommunications (BUPT), Beijing, China. His research interests focus mainly on statistical array and signal processing for digital wireless communication systems.



Haitao Zheng received her M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1998 and 1999, respectively. From August 1999 to March 2004, she was with the Wireless Research Laboratory, Bell Labs, Lucent Technologies, Holmdel, NJ. Since March 2004, she has joined the Wireless and Networking Research Group, Microsoft Research Asia. Her research interests include wireless communications and networking, multimedia communications, and signal processing. Dr. Zheng was admitted to a highly gifted class of Xi'an Jiaotong University at the age of 15. She received the 1998–1999 George Harhalakis Outstanding Systems Engineering Graduate Student Award in recognition of her outstanding contributions in cross-disciplinary research from the University of Maryland, College Park. Recently, she received Bell Laboratories' 2002 President's Gold Award in recognition of her outstanding level of innovation, technical excellence, and business impact. She currently serves as the TPC Member of the IEEE Multimedia Signal Processing Technical Committee, Guest Editor of the IEEE JSAC Special Issue on Advanced Mobility Management and QoS Protocols for Wireless Internet, and Guest Editor of the EURASIP Journal on Applied Signal Processing Special Issue on Cross-Layer Design for Communications and Signal Processing. She has served as a TPC Member of ICME, ICC, and ICASSP.