Opportunistic Scheduling with QoS Constraints for Multiclass Services HSUPA System

Dan Liao and Lemin Li

This paper focuses on the scheduling problem with the objective of maximizing system throughput, while guaranteeing long-term quality of service (QoS) constraints for non-realtime data users and short-term QoS constraints for realtime multimedia users in multiclass service high-speed uplink packet access (HSUPA) systems. After studying the feasible rate region for multiclass service HSUPA systems, we formulate this scheduling problem and propose a multi-constraints HSUPA opportunistic scheduling (MHOS) algorithm to solve this problem. The MHOS algorithm selects the optimal subset of users for transmission at each time slot to maximize system throughput, while guaranteeing the different constraints. The selection is made according to channel condition, feasible rate region, and user weights, which are adjusted by stochastic approximation algorithms to guarantee the different QoS constraints at different time scales. Simulation results show that the proposed MHOS algorithm guarantees QoS constraints, and achieves high system throughput.

Keywords: Opportunistic scheduling, multiclass services, long-term QoS, short-term QoS, HSUPA.

I. Introduction

One of the most important features of 3G and beyond 3G systems is to support high-speed packet data services. A series of specifications have been released by the 3rd Generation Partnership Project (3GPP) during the evolution of wideband code-division multiple-access (WCDMA) to address the increasing demand for improvement in system coverage and performance and reduction in packet delay. Due to the asymmetric capability of packet data transmission in downlink and uplink of WCDMA Release 5 and further requirements for peer-to-peer applications, 3GPP is working on improving the uplink packet data transmission capability in high-speed uplink packet access (HSUPA) systems [1], [2], which was introduced in WCDMA Release 6.

Uplink scheduling in wireless systems is gaining importance due to rising demand for uplink intensive data services (ftp, image uploads, and so on). There have been some previous studies on the subject of uplink scheduling [3]-[8]. The authors of [3] propose and study algorithms for efficient uplink packet-data scheduling in a CDMA cell. The algorithms attempt to maximize system throughput under the transmit power limitations of mobiles assuming instantaneous knowledge of user queues and channels. In [4], a low-complexity distributed scheduling control is proposed to maximize the efficiency of medium access control (MAC) on the uplink in a future WCDMA-based third-generation radio access network, and hence maximize spectral efficiency. In [5], the authors study the problem of uplink system performance maximization, while maintaining fairness among the various users despite their time-varying channel conditions. In [6], the authors determine the optimal adaptive rate and power control strategies to maximize the total throughput in a multirate

Manuscript received July 16, 2006: revised Dec. 15, 2006.

This work was supported by National Natural Science Foundation of China (NSFC), Research Grants Council (RGC) of Hong Kong Joint Research (NO. 60218002), and National Key Basic Research Program of China (2007CB307104 of 2007CB307100).

Dan Liao (phone: + 86 288 320 3008, email: liaodan@uestc.edu.cn) and Lemin Li (email: lml@uestc.edu.cn) are with School of Communication and Information Engineering, University of Electronics Science and Technology of China, SiChuan, China.

CDMA system. However, in [3]-[6], realtime multimedia users are not taken into consideration.

Wireless communication networks are expected to support multimedia traffic, such as video and voice, which have a variety of quality of service (QoS) requirements, and to make efficient use of the radio resource. Various studies have been carried out on short-term QoS support for multimedia users with a scheduling scheme. Some of these studies have considered uplink transmission [7], [8]. In [7], a dynamic fair resource allocation scheme to efficiently support realtime and non-realtime traffic with guaranteed statistical QoS in uplinks of wideband CDMA cellular networks was proposed. The scheme uses the generalized processor sharing (GPS) fair service discipline to allocate uplink channel resources, taking into account the characteristics of channel fading and inter-cell interference. In [8], the authors formulated and studied the power allocation and scheduling problem in real-time CDMA wireless networks. The authors reduced the optimization problem to a linear programming problem and resolved it. However, in [7] and [8], the coexistence of non-realtime data users and realtime multimedia users and the differing needs for long-term and short-term QoS guarantees were not considered. None of these previously proposed schemes [3]-[8] are designed for multiclass services in HSUPA systems.

Recently, a few researchers have focused on the scheduling problem for HSUPA systems [9], [10]. In [9], three different time and code division schedulers are evaluated for data transmission on an enhanced dedicated channel (E-DCH) in WCDMA uplink. In [10], the authors study the performance of WCDMA enhanced uplink systems deploying Node B based scheduling and fast hybrid ARQ (HARQ) retransmission protocols. However, none of these studies consider multiclass services which include realtime multimedia users and non-realtime data users, nor do they support the different QoS guarantee constraints.

In this paper, we study the scheduling problem with the objective of maximizing system throughput, while guaranteeing QoS constraints in multiclass service HSUPA systems, and propose an opportunistic scheduling (OS) algorithm for multiclass service HSUPA systems. Recently, various OS schemes have been developed for wireless networks (see [11-18] and references therein). Opportunistic scheduling is intended to balance two conflicting goals: high system throughput, and QoS constraints (such as fairness and minimal throughput guarantee). On the one hand, the wireless resource normally should be allocated to users with good channel conditions to achieve high system throughput. On the other hand, the throughput-based strategy will inevitably cause constraint problems. A tradeoff between throughput maximization and users' constraints is needed.

In [11]-[18], the authors study the OS problem in different systems. We can classify these scheduling schemes as being either single-server scheduling or multi-server scheduling based on the underlying multiple access scheme used. In single-server scheduling, only one user is served at a time, whereas in multi-server scheduling, multiple users can be served simultaneously.

In [11]-[14], single-server OS problems are studied. In a single-server scheduling scheme, there is no interference between users, so only user selection problems are considered without considering the power resource that should be allocated to each user. In contrast, we must decide which and how many users should be scheduled and what power resource should be allocated to each of the scheduled user in a given time-slot. In [15]-[18] the multi-server OS problem is studied. In [15], the authors first develop a general methodology to design opportunistic fair wireless schedulers using an adaptive control framework. They next formulate and solve the multi-channel scheduling problem. However, interference between users is not considered and a linear relationship of the scheduled rate and power consumption is assumed. In [16] an opportunistic power scheduling scheme is presented in which multiple transmissions are allowed, namely, a joint time-slot and power allocation scheme for downlink communication in wireless systems. In [17], the authors study, analyze, and demonstrate the tradeoffs between the achievable throughput and fairness that occur in the realization of multi-user uplink transmission scheduling in CDMA systems. However, in [16] and [17] the short-term QoS for realtime multimedia users is not considered. In [18], the authors generalize the OS to include multiple constraints, multiple interfaces, and short term fairness. However, in [18] realtime multimedia users are not considered. The short-term fairness constraint is not suitable for realtime multimedia users. None of the existing OS schemes are suitable for multiclass service HSUPA systems because none of them consider the coexistence of non-realtime data users and realtime multimedia users or the feasible rate region for HSUPA systems.

The main contributions of this paper are as follows. We first study the feasible rate region in multiclass service HSUPA systems. We then formulate the scheduling problem with the objective of maximizing the system throughput, while guaranteeing long-term QoS constraints for non-realtime data users and short-term QoS constraints for realtime multimedia users in HSUPA. Finally, we propose a new OS algorithm, multi-constraint HSUPA opportunistic scheduling (MHOS), with QoS constraints for multiclass service HSUPA systems and analyze the performance of MHOS via simulation.

The rest of this paper is organized as follows. In section II, we study the feasible rate region and formulate the scheduling

problem. In section III, we propose the new MHOS algorithm for multiclass service HSUPA systems. In section IV, we analyze the performance of MHOS via simulation. Finally, we give our conclusions in section V.

II. Formulation of Feasible Rate Region and Scheduling Problem

In this section, we first study the feasible rate region for multiclass service HSUPA systems. Then, considering the feasible rate region constraint, we formulate the scheduling problem with the objective of maximizing system throughput, while guaranteeing long-term QoS constraints for non-realtime data users and short-term QoS constraints for realtime multimedia users.

1. Feasible Rate Region for HSUPA

In HSUPA, multi-server transmission is adopted. In multi-server scheduling, multiple users can be served simultaneously. In transmission, simultaneously scheduled users could interfere with each other. Because there is a maximum transmission power limit, the signal-to-interference ratio (SIR) each user can achieve is limited. Since the transmission rate is decided by SIR, in addition to the QoS constraints there is the feasible rate region constraint for the scheduling problem in HSUPA.

In a single cell in an HSUPA system which employs adaptive modulation and coding (AMC) and multicode, we assume that there are K types of services supported in the cell. Of service type k, it is assumed that there are N_k users. Let q_j denote the j-th user of service type q. The uplink SIR, $\gamma_m^{q_j}$, for multicode m at user q_j, is given by

$$\gamma_m^{q_j} = \frac{P_m^{q_j} h^{q_j}}{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{m=1}^{M^{k_i}} h^{k_i} P_m^{k_i} - \sum_{m=1}^{M^{q_j}} P_m^{q_j} h^{q_j} + N_0}, \tag{1}$$

where $P_m^{q_j}$ is the transmit power for multicode m at user q_j, h^{q_j} is the path gain for user q_j, M^{q_j} is the number of multicodes allocated to user q_j, and N_0 is the background noise power. In (1), the effect of "self-noise" is not included.

Let γ^{q_j} denote the SIR requirement for user q_j to achieve the target bit error rate (BER). In order to meet the SIR requirement of all users, we must then have for any multicode of each user

$$\frac{P_m^{q_j} h^{q_j}}{\sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{m=1}^{M^{k_i}} h^{k_i} P_m^{k_i} - \sum_{m=1}^{M^{q_j}} P_m^{q_j} h^{q_j} + N_0} \geq \gamma^{q_j}. \tag{2}$$

The feasible rate region for frequency division duplex (FDD) wideband CDMA has been derived in [20] and that for uplink of a single CDMA cell has been derived in [3]. However, those studies do not consider the specific characteristics of HSUPA, such as AMC. We first formulate the feasible SIR vectors γ specified by (2) for HSUPA. We then give the feasible rate region and point out the specific aspects utilized later in our scheduling algorithm.

To minimize the power levels of each user, the equality in (2) must hold. This yields

$$\frac{P_m^{q_j} h^{q_j}}{\gamma^{q_j}} = \sum_{k=1}^{K}\sum_{i=1}^{N_k}\sum_{m=1}^{M^{k_i}} h^{k_i} P_m^{k_i} - \sum_{m=1}^{M^{q_j}} P_m^{q_j} h^{q_j} + N_0. \tag{3}$$

Note that (3) is satisfied for all multicodes of user q_j. Thus, for multicode m of user q_j, the right side of (3) is the same as that for multicode n. Thus,

$$P_m^{q_j} = P_n^{q_j} = P_1^{q_j}. \tag{4}$$

From (4), (3) becomes

$$\begin{aligned} \frac{P_1^{q_j} h^{q_j}}{\gamma^{q_j}} &= \sum_{k=1}^{K}\sum_{i=1}^{N_k} M^{k_i} P_1^{k_i} h^{k_i} - M^{q_j} P_1^{q_j} h^{q_j} + N_0, \\ (\frac{1}{\gamma^{q_j}} + M^{q_j}) P_1^{q_j} h^{q_j} &= \sum_{k=1}^{K}\sum_{i=1}^{N_k} M^{k_i} P_1^{k_i} h^{k_i} + N_0. \end{aligned} \tag{5}$$

It should be noted that (5) is also satisfied for any user k_i, that is, the left side of (5) can be $[1/\gamma^{k_i} + M^{k_i}]P_1^{k_i} h^{k_i}$. Thus,

$$(\frac{1}{\gamma^{q_j}} + M^{q_j}) P_1^{q_j} h^{q_j} = (\frac{1}{\gamma^{k_i}} + M^{k_i}) P_1^{k_i} h^{k_i}. \tag{6}$$

Defining Θ^{q_j} as $[1/\gamma^{q_j} + M^{q_j}]h^{q_j}$, (6) becomes

$$P_1^{k_i} = \frac{\Theta^{q_j}}{\Theta^{k_i}} P_1^{q_j}. \tag{7}$$

From (7), (5) becomes

$$\begin{aligned} \Theta^{q_j} P_1^{q_j} &= \sum_{k=1}^{K}\sum_{i=1}^{N_k} M^{k_i} h^{k_i} \frac{\Theta^{q_j}}{\Theta^{k_i}} P_1^{q_j} + N_0, \\ (\Theta^{q_j} - \sum_{k=1}^{K}\sum_{i=1}^{N_k} M^{k_i} h^{k_i} \frac{\Theta^{q_j}}{\Theta^{k_i}}) P_1^{q_j} &= N_0, \\ P_1^{q_j} &= \frac{N_0}{\Theta^{q_j}(1 - \sum_{k=1}^{K}\sum_{i=1}^{N_k} M^{k_i} h^{k_i} \frac{1}{\Theta^{k_i}})}. \end{aligned} \tag{8}$$

Let P^{q_j} denote the transmit power for user q_j. We assume

that the maximum transmission power for user q_j is $p_{\max}^{q_j}$. $P^{q_j} = \sum_{m=1}^{M^{q_j}} P_m^{q_j} \le P_{\max}^{q_j}$. From (4), $P_m^{q_j} \le P_{\max}^{q_j} / M^{q_j}$. From (8),

$$P_1^{q_j} = \frac{N_0}{\Theta^{q_j}(1-\sum_{k=1}^{K}\sum_{i=1}^{N_k}\frac{M^{k_i}h^{k_i}}{\Theta^{k_i}})} \le \frac{P_{\max}^{q_j}}{M^{q_j}},$$
$$\sum_{k=1}^{K}\sum_{i=1}^{N_k}\frac{M^{k_i}h^{k_i}}{\Theta^{k_i}} \le 1-\frac{N_0 M^{q_j}}{P_{\max}^{q_j}\Theta^{q_j}}, \quad (9)$$

which must be satisfied for all users. Thus, defining $\phi^{k_i} = \frac{M^{k_i}h^{k_i}}{\Theta^{k_i}} = \frac{M^{k_i}}{1/\gamma^{k_i} + M^{k_i}}$, (9) becomes

$$\sum_{k=1}^{K}\sum_{i=1}^{N_k}\phi^{k_i} \le 1-\max_{q_j}\frac{N_0\phi^{q_j}}{P_{\max}^{q_j}h^{q_j}}, \quad (10)$$

which must be satisfied in order to have an existing power allocation.

Considering the AMC, according to [21], there is a relationship between rate and SIR as (11) to achieve the target BER:

$$R^{q_j} = M^{q_j} W \log_2(1-\frac{3\gamma^{q_j}}{2\ln 5\varepsilon^{q_j}}), \quad (11)$$

where W is bandwidth of the HSUPA system, R^{q_j} is the transmit rate for user q_j, and ε^{q_j} is the target BER for user q_j.

From $\phi^{k_i} = \frac{M^{k_i}}{1/\gamma^{k_i} + M^{k_i}}$, (11) becomes

$$R^{q_j} = M^{q_j} W \log_2(1-\frac{3\phi^{q_j}}{2(M^{q_j} - M^{q_j}\phi^{q_j})\ln 5\varepsilon^{q_j}}). \quad (12)$$

Let $M_{\max}$ be the number of multicodes provided in the HSUPA system. The multicode constraint is

$$\sum_{k=1}^{K}\sum_{i=1}^{N_k} M^{k_i} \le M_{\max}. \quad (13)$$

Since our scheduling scheme is used at each slot, we should get the feasible rate region at a certain slot. Let $\Re_t$ denote the feasible rate region at slot t. To get $\Re_t$, we must get the rate constraints at slot t. Let $\phi_t^{q_j}$ be ϕ^{q_j} at slot t, $M_t^{q_j}$ be M^{q_j} at slot t, and $h_t^{q_j}$ be h^{q_j} at slot t. At slot t, constraints (10), (12), and (13) can be changed to (14)-(16), respectively:

$$\sum_{k=1}^{K}\sum_{i=1}^{N_k}\phi_t^{q_j} \le 1-\max_{q_j}\frac{N_0\phi_t^{q_j}}{P_{\max}^{q_j}h_t^{q_j}}, \quad (14)$$

$$R_t^{q_j} = M_t^{q_j} W \log_2[1-\frac{3\phi_t^{q_j}}{2(M_t^{q_j} - M_t^{q_j}\phi_t^{q_j})\ln 5\varepsilon^{q_j}}], \quad (15)$$

$$\sum_{k=1}^{K}\sum_{i=1}^{N_k} M_t^{q_j} \le M_{\max}. \quad (16)$$

Let $\vec{R}_t = \{R_t^{q_j} | q_j \in U\}$ be the transmit rate vector, where $R_t^{q_j}$ is a random variable representing the transmit rate value for user q_j at time slot t and U is the set containing all users in the HSUPA system.

Combining (14)-(16), the feasible rate region at slot t, $\Re_t$ is given by

$$\Re_t = \{\vec{R}_t | \vec{R}_t \text{ satify (14)} - \text{(16)}\}. \quad (17)$$

2. Formulation of the Opportunistic Scheduling Problem with QoS Requirements

Given $\vec{R}_t$, the scheduling algorithm Q determines which user subset $U_Q(\vec{R}_t)$ should be scheduled: if $q_j \in U_Q(\vec{R}_t)$, then user q_j should be served at the time slot t.

We consider an HSUPA system that supports multimedia and data users. For data users, long-term QoS constraints are required. In this paper, long-term fairness constraints for data users are considered. Given the target weight ω^{q_j} of data user q_j in the system, deterministic fairness requires

$$E[R_t^{q_j} 1(q_j \in U_Q(\vec{R}_t))] / E[R_t^{k_i} 1(k_i \in U_Q(\vec{R}_t))] = \omega^{q_j} / \omega^{k_i}, \quad (18)$$

where $1(\Lambda)$ is the indicator function of the event if Λ occurs, $1(\Lambda) = 1$; otherwise, $1(\Lambda) = 0$.

In this paper, we introduce a short-term QoS constraint and a probabilistic throughput guarantee constraint for realtime multimedia users. For realtime multimedia users q_j, assume the data is discharged with the constant rate D^{q_j}, thus, during E time slots duration, multimedia user q_j has throughput requirement $B^{q_j} = D^{q_j} E\Delta$, where Δ is the scheduling interval. The data transmitted during E time slots duration is a frame. If the size of a frame of user q_j is less than B^{q_j}, we consider that frame an error. Let $\Pr^{q_j}$ denote the frame error rate for multimedia user q_j. The probabilistic throughput guarantee constraint for multimedia user q_j requires that the probability $\Pr^{q_j}$ is less than the target probability π^{q_j}. The constraint is formulated as

$$P\{\sum_{t=T}^{E+T-1} R_t^{q_j} \Delta 1(q_j \in U_Q(\vec{R}_t)) < B^{q_j}\} \le \pi^{q_j} \text{ (for any slot T)} \quad (19)$$

Assume that set A contains the services type of realtime multimedia users, and set B contains the services type of non-realtime data users. Considering the feasible rate region for

HSUPA, the scheduling problem with the objective of maximizing system throughput, while guaranteeing long-term QoS constraints for data users and short-term QoS constraints for multimedia users, can be formulated as an optimization problem as

$$\max_{Q} \sum_{k=1}^{K} \sum_{i=1}^{N_k} E[R_t^{k_i} 1(k_i \in U_Q(\vec{R}_t))] \tag{20}$$

$$\text{s.t. } P\{ \sum_{t=T}^{E+T-1} R_t^{q_j} \Delta 1(q_j \in U_Q(\vec{R}_t)) < B^{q_j} \} \le \pi^{q_j},\ q \in A, \tag{21}$$

$$E[R_t^{q_j} 1(q_j \in U_Q(\vec{R}_t))] / E[R_t^{k_i} 1(k_i \in U_Q(\vec{R}_t))] = \omega^{q_j} / \omega^{k_j},\ q \in B, k \in B \tag{22}$$

$$\vec{R}_t \in \Re_t, \tag{23}$$

where (21) is the short-term probabilistic throughput guarantee constraints for realtime multimedia users, (22) is the long-term fairness constraints for non-realtime data users, and (23) is the feasible rate region constraint for the HSUPA system.

III. Proposed MHOS Algorithm

Our objective in designing the scheduler is to ensure QoS under feasible rate region constraints, while employing opportunistic scheduling strategies to increase the total system throughput by selecting users with high-quality channels when possible. To solve the problem, we take an approach similar to that taken in [11]-[18]. We decouple the MHOS scheduler and treat it as two separate entities: the user weight updating block and the throughput optimization block. The throughput optimization block selects the users with high-quality channels under a feasible rate region constraint to achieve the highest weighted system throughput. The user weight updating block using a stochastic approximation algorithm updates the user weights to guarantee the different QoS constraints for multimedia and data users.

1. Design of Throughput Optimization Block

According to the framework, as in [11]-[18], the functionality of the throughput optimization block is to ensure a throughput-optimal selection of users and rates. Let $W_t^{q_j}$ denote the user weight for q_j at slot t. The objective of the throughput optimization block is set to

$$\max_{Q} \sum_{k=1}^{K} \sum_{i=1}^{N_k} W_t^{k_i} R_t^{k_i} 1(k_i \in U_Q(\vec{R}_t)), \tag{24}$$

such that in each time slot, the weighted system throughput is maximized. The constraints of this optimization are given by (23). Notice that there is no QoS constraint as QoS is treated in the updating block.

From objective (24) and constraint (23), our optimal policy is defined as

$$Q^*(\vec{R}_t) = \arg\max_{u_t \in \Omega, \vec{R}_t \in \Re_t} \sum_{k_i \in u_t} W_t^{k_i} R_t^{k_i}, \tag{25}$$

where Ω is all possible user subsets, and u_t is the user subset selected at slot t.

The policy defined in (25) selects users and rates to achieve the highest weighted system throughput. We observe that this formulation is an NP-hard knapsack problem [22]. Consequently, since a complete search of the solution space is infeasible in practice, we develop an approximation to the optimal solution as follows. We decouple the policy as two separate entities: rate selection and user selection. For a certain user subset, we give a rate selection policy to increase the weighted system throughput. Based on the rate selection policy, we select the optimal user subset to serve.

A. Rate Selection

For user subset $\tilde{u}$, the policy (25) becomes

$$Q^*(\vec{R}_t) = \arg\max_{\vec{R}_t \in \Re_t} \sum_{k_i \in \tilde{u}_t} W_t^{k_i} R_t^{k_i}. \tag{26}$$

Comparing policy (26) with (25), we can see that policy (26) does not select the user subset; it only selects the optimal rate for users in a certain user subset.

From (15), the transmit rate $R_t^{q_j}$ is the function of $\phi_t^{q_j}$ and $M_t^{q_j}$, so the rate selection is equal to $\phi_t^{q_j}$ calculation and $M_t^{q_j}$ allocation. We first calculate $\phi_t^{q_j}$, and then allocate the multicodes for all users to achieve the highest weighted system throughput. In this paper, $\phi_t^{q_j}$ is the main power resource that is allocated to users like the power index defined in [23].

We develop an approximation to the optimal $\phi_t^{q_j}$ calculation as follows. Define $\beta_t^{q_j}$ as the temporary weight used for resource allocation for user q_j at slot t. Let $\beta_t^{q_j} = W_t^{q_j} h_t^{q_j}$ and $\Phi_t = \sum_{q_j \in \tilde{u}} \phi_t^{q_j}$. Considering (15), because $R_t^{q_j}$ is the increasing function of $\phi_t^{q_j}$, the larger $h_t^{q_j}$ is, the larger Φ_t is. Observe that the larger $\beta_t^{q_j}$ is, the more likely the user will be selected, as it adds increasingly to the objective function. That means we allocate more resources to users with good channel condition and user weight to achieve high weighted system throughput. To simplify the allocation, we allocate the resource to users fairly according to temporary weight $\beta_t^{q_j}$. We calculate $\phi_t^{q_j}$ as $\phi_t^{q_j} = \theta_t^{q_j} \Phi_t$, where $\theta_t^{q_j} = \beta_t^{q_j} / \sum_{k_i \in \tilde{u}} \beta_t^{k_i}$. To maximize Φ_t, the equality in (14)

must hold

$$\begin{aligned}
&\Phi_t = 1 - \max_{q_j \in \tilde{u}} \frac{N_0 \theta_t^{q_j} \Phi_t}{P_{\max}^{q_j} h_t^{q_j}}, \\
&\Phi_t \left(1 + \max_{q_j \in \tilde{u}} \frac{N_0 \theta_t^{q_j}}{P_{\max}^{q_j} h_t^{q_j}}\right) = 1, \\
&\Phi_t = \min_{q_j \in \tilde{u}} \frac{1}{1 + \dfrac{N_0 \theta_t^{q_j}}{P_{\max}^{q_j} h_t^{q_j}}}, \\
&\phi_t^{q_j} = \theta_t^{q_j} \Phi_t = \theta_t^{q_j} \min_{q_j \in \tilde{u}} \frac{1}{1 + \dfrac{N_0 \theta_t^{q_j}}{P_{\max}^{q_j} h_t^{q_j}}}.
\end{aligned} \tag{27}$$

After calculating $\phi_t^{q_j}$ according to (27), we allocate the multicodes with the objective of maximizing the weighted system throughput. The allocation policy is as follows:

Step 1. Sort all users in the system in decreasing order of $\beta_t^{q_j}$.

Step 2. Initialization: Allocate one multicode to each user (The user selection policy guarantees $|\tilde{u}| \le M_{\max}$).

Step 3. Select $M_t^{q_j}$ for users in order starting from the top of the list. The selection goal is to maximize $R_t^{q_j}$ according to (15), when $\phi_t^{q_j}$ is calculated according to (27).

Step 4. Stop if all the multicodes are allocated, that is, $\sum_{q_j \in \tilde{u}} M_t^{q_j} = M_{\max}$, or we have allocated multicodes for all users.

B. User Selection

To solve (25), we present the user selection policy in this section. We first present the stochastic approximation algorithm for user subset selection based on [24] and [25], which generates a random walk over the different user subsets where each new subset is obtained from the old one by moving toward the global optimizer. We then present a low-complexity approximation of user subset selection.

1) Adaptive User Subset Selection

Because there is a multicode limit, $M_{\max}$, the number of selected users must be less than $M_{\max}$. Let $U' = \{u | u \in \Omega; |u| \le M_{\max}\}$. We use the $|U'|$ unit vectors to denote all $|U'|$ possible user subsets. That is, $e_1, e_2, \cdots, e_{|U'|}$, where e_i denotes a vector, with a one at the i-th position and zeros elsewhere. At each iteration, the algorithm updates the vector $\mathrm{P}(n) = [\rho(n,u_1), \rho(n,u_2), \cdots, \rho(n,u_{|U'|})]$, which represents the state occupation probabilities with the elements satisfying $\rho(n,u_i) \in [0,1]$ and $\sum_{i=1}^{|U'|} \rho(n,u_i) = 1$. The user subset visited is denoted as $u(n)$ and the optimal user subset at the n-th iteration is denoted as $\hat{u}(n)$. For notational simplicity, we map a sequence of user subsets $\{u(n), n = 1,2,\cdots\}$ to the sequence $\{D(n), n = 1,2,\cdots\}$, the element of which is a unit vector satisfying $D(n)=e_i$, if $u(n) = u_i, i = 1,2,\cdots,|U'|$. The discrete stochastic approximation algorithm for solving the optimum user subset selection problem within each time slot is summarized as follows:

Step 1. Initialization: n=1. Randomly select the initial user subsets $u(n) \in U'$ and $\hat{u}(n) = u(n)$. Set the probability vector P(n) by $\rho(n,u(n)) = 1$ and $\rho(n,u) = 0$ for all $u \ne u(n)$.

Step 2. Sampling and evaluation: Given $u(n)$, calculate the weighted system throughput $WSR = \sum_{q_j \in u(n)} W_t^{q_j} R_t^{q_j}$ based on the rate selection. Uniformly choose another candidate $u'(n) \in U' \setminus u(n)$ and compute the corresponding weighted system throughput WSR'.

Step 3. Acceptance: If $WSR' > WSR$, then set $u(n+1) = u'(n)$; otherwise, set $u(n+1)=u(n)$.

Step 4. Adaptive filter for updating state occupation probabilities: Update the state occupation probabilities by $\rho(n+1) = \rho(n) + a(n)[D(n+1) - \rho(n)]$, where $a(n)=1/n$ is a decreasing step size.

Step 5. Updating the estimate of the optimizer at the current iteration: If $\rho(n+1,u(n+1)) > \rho(n+1,\hat{u}(n+1))$, then set $\hat{u}(n+1) = u(n+1)$; otherwise, set $\hat{u}(n+1) = \hat{u}(n)$.

Step 6. $n=n+1$, and go to step 2.

Let $N = \sum_{k=1}^{K} N_k$, thus, $|U'| = \sum_{i=1}^{M_{\max}} C_N^i$. If the number of users in the HSUPA system is large, the number of user subsets is very large. There are many parameters which need to be updated. That leads to hard implementation of adaptive user subset selection; therefore we present a low-complexity approximation of user subset selection. In the simulation, we implement the low-complexity approximation of user subset selection in the MHOS scheme rather than the adaptive user subset selection.

2) Low-complexity approximation

We propose the following simple user subset selection policy, similar to that in [3], which may be more suitable for practical implementation:

Step 1. Sort all users in system in decreasing order of $\beta_t^{q_j}$.

Step 2. Add user i in order starting from the top of the list, while maintaining and updating the weighted system

throughput $WSR = \sum_{j \le i} W_j^t R_j^t$ based on the rate selection.

Step 3. Stop if adding the next user reduces *WSR* or all users in the system are added, and allow transmission of all added users at their rates as computed.

In our simulations section, we use this simple user subset selection policy in the MHOS scheme.

2. Design of User Weight Updating Block

The function of the updating block is to update the control parameters in order for the output of the throughput optimization block to satisfy the QoS requirement. Recall that the requirement of the updating block is to guarantee that the user weights vector converges to the optimal value while satisfying the QoS constraint. To ensure this, we employ the stochastic approximation algorithm in the updating block. We note that such a technique is employed in other studies [11]-[18] for long-term QoS guarantee. At the first slot, $W_t^{q_j}$ is equal to the original user weight $W_0^{q_j}$ assigned by the system.

For data users, there are long-term QoS constraints:

$$E[R_t^{q_j} 1(q_j \in U_Q(\vec{R}_t))] / E[R_t^{k_i} 1(k_i \in U_Q(\vec{R}_t))] = \omega^{q_j} / \omega^{k_i}.$$

We use the stochastic approximation algorithm to adaptively find optimal user weights for data users as

$$W_{t+1}^{q_j} = W_t^{q_j} - a_t \left(\frac{\omega^{q_j}}{\sum_{k \in B} \omega^{k_i}} - \frac{R_t^{q_j} 1(q_j \in U_Q(\vec{R}_t))}{\sum_{k \in B} R_t^{k_i} 1(k_i \in U_Q(\vec{R}_t))} \right), \quad (28)$$

where a_t is chosen to converge to zero, that is, a^t=1/t.

For multimedia users, there are short-term QoS constraints:

$$P\left\{ \sum_{t=T}^{E+T-1} R_t^{q_j} \Delta 1(q_j \in U_Q(\vec{R}_t)) < B^{q_j} \right\} \le \pi^{q_j}.$$

We update the user weights for multimedia users at slot t as

$$W_{t+1}^{q_j} = W_t^{q_j} \quad \text{if } t \bmod E \neq 0, \quad (29)$$

$$W_{t+1}^{q_j} = W_t^{q_j} - a_t(\pi^{q_j} - 1(\sum_{k=t-M+1}^{t} R_k^{q_j} \Delta 1(q_j \in U_Q(\vec{R}_k)) < B^{q_j})) \quad \text{if } t \in \{t | t \bmod E = 0\} \cap Z^+. \quad (30)$$

The basic purpose of the user weight update is to increase the user weight value for users who do not achieve the QoS requirement and to decrease the user weight value for users who do achieve it.

In this study, we verify by simulation that the proposed MHOS scheme for multiclass service HSUPA systems can achieve high system throughput while guaranteeing the different QoS constraints.

IV. Simulation Results

In this section, three baseline algorithms are proposed: the HSUPA proportional fair (HPF) algorithm; the HSUPA multimedia user first (HMUF) algorithm modified from proportional fair (PF) [19], [26]; and the time-span minimization and best effort (TSMABE) [27] for HSUPA algorithm. In the simulation, comparisons are made between HPF, HMUF, and MHOS in terms of system throughput and QoS guarantee.

1. Two Baseline Algorithms

In this subsection, we modify two scheduling algorithms for HSUPA as baseline algorithms and compare the proposed MHOS algorithm with them by simulation in the next subsection.

A. HSUPA Proportional Fair (HPF) Algorithm

The proportional fair (PF) scheduling rule [19], [26] is implemented in commercial mobile networks including CDMA 2000 1xEV-DO (or HDR). The PF scheduling rule is widely accepted because it is simple and efficient and it has few parameters that need to be optimized. However, since the PF algorithm schedules users one at a time, it is modified for uplink [3]. In this subsection, we modify the PF algorithm for HSUPA.

This HPF algorithm has the following steps:

Step 1. Sort users in descending order of the measure $\vartheta_t^i = R_t^i(0) / \overline{R}^i$, assuming no interference from other users while computing $R_t^i(0)$. The average rate of user i is denoted by $\overline{R}^i$. It is updated through a low pass filter in each scheduling interval [19].

Step 2. Add i users in order starting from the top of the list, while maintaining and updating weighted system throughput $WSR = \sum_{j \le i} W_j^t R_j^t$ based on the rate selection, where W_t^i is equal to original user weight W_0^i at any slot and temporary user weight $\beta_t^i = \vartheta_t^i = R_t^i(0) / \overline{R}^i$.

Step 3. Stop if adding the next user reduces *WSR* or all users in the system are added, and allow transmission of all added users at their rates as computed.

B. The HSUPA Multimedia User First (HMUF) Algorithm

In [2], the authors propose the TSMABE algorithm for a scheduling problem with minimum throughput requirement

constraint. The TSMABE serves the other users only if all users' constraints are guaranteed during the whole time duration. In this subsection, we propose the HMUF algorithm according to the basic idea of TSMABE. In this algorithm, we serve the data users only if all multimedia users' constraints are satisfied.

Let MU_t be the set of all the multimedia users who need more time slots to satisfy their QoS requirement at time slot t. At the first time slot, MU_t is equal to MU, which is the set of all the multimedia users in the system. If the QoS requirement for a multimedia user is satisfied, then the user is removed from MU_t.

This HMUF algorithm has the following steps:

Step 1. If $MU^t \neq \phi$, go to step 2; otherwise, go to step 3.

Step 2. Only serve users in set MU^t. Sort users belonging to set MU^t in descending order of the measure h_t^i. Add i users in order starting from the top of the list, while maintaining and updating weighted system throughput $WSR = \sum_{j \le i} W_j^t R_j^t$ based on the rate selection, where W_t^i is equal to original user weight W_0^i at any slot, and temporary user weight $\beta_t^i = h_t^i$. Stop if adding the next user reduces WSR or all users in set MU^t are added, and allow transmission of all added users at their rates as computed.

Step 3. Only serve data users. Compute R_t^i for all data users based on the rate selection, where both W_t^i and β_t^i are equal to original user weight W_0^i at any slot, and allow transmission of all data users at their rates as computed.

2. Numerical Results and Discussion

This subsection compares the proposed MHOS algorithm with two baseline algorithms, HPF and HMUF. The simulation demonstrates that the MHOS algorithm shows good performance.

The system parameters used in the simulation are as follows. We assume that bandwidth W=3.84 MHz, scheduling interval Δ is 2 ms, the number of multicodes $M_{\max}$=50, and the maximum transmission power $P_i^{\max} = 2W$ for all users. Assume that the one-sided power spectral density of noise is 10^{-6}, so the background noise power $N_0 = W^* 10^{-6} = 3.84$.

Assume that there is only one service type of multimedia user and one service type of data user in the system. The number of multimedia users is N_u, and the number of data users is N_d. For the multimedia user i, we assume the target FER $\pi^i = 0.01$, the target BER $\varepsilon^i = 10^{-3}$, time slot duration E=100, original user weight $W_0^i = 2$, constant rate D^i=50 KBps, and throughput requirement B^i=10 Kb. For the data user i, we assume the target BER $\varepsilon^i = 10^{-6}$, target weight $\omega^i = 1/N_d$, and original user weight $W_0^i = 1$. Assume that the path gains for all users follow Rayleigh distribution with Rayleigh scale parameter b = 0.2.

In our simulation, we assume the number of data users is a constant value, N_d= 30, and the number of multimedia users is changed from 25 to 40.

Figure 1 shows the system throughput for MHOS, HPF, and HMUF. From Fig. 1, we can see that MHOS achieves the highest system throughput, employing opportunistic scheduling strategies to increase the total system throughput by selecting users with high-quality channels when possible. The system throughput for MHOS increases when the number of multimedia users increases. This is because when the number of users in the system increases, the probability of there being high-quality channels increases. Because HMUF selects multimedia users with high-quality channels when possible, the system throughput increases with the increasing of the number of multimedia users. When the number of multimedia users increases, the system throughput for HPF increases at first.

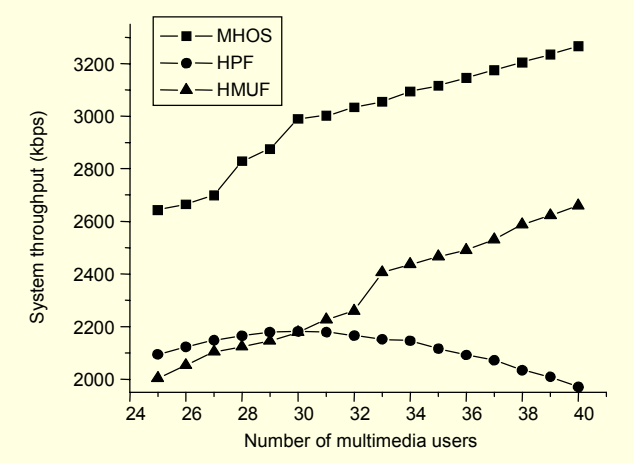


Fig. 1. System throughput for MHOS, HPF, and HMUF.

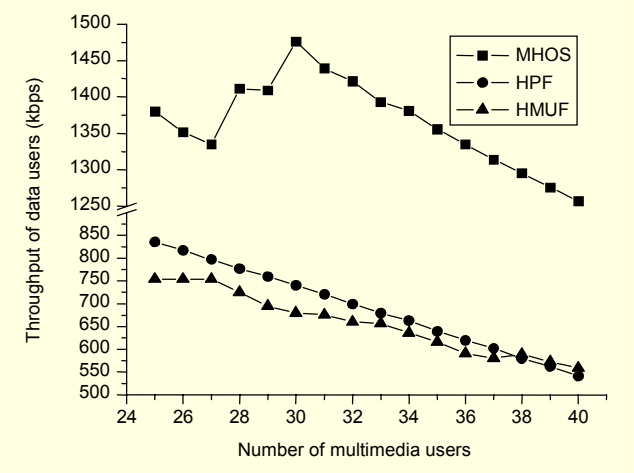


Fig. 2. Throughput of data users for MHOS, HPF, and HMUF.

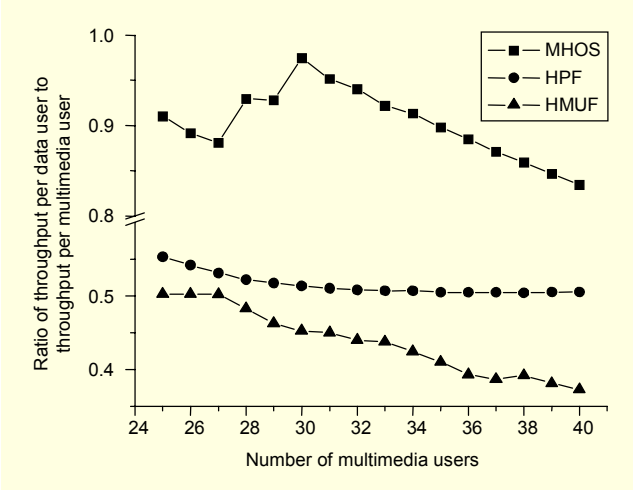


Fig. 3. Ratio of throughput per data user to throughput per multimedia user for MHOS, HPF, and HMUF.

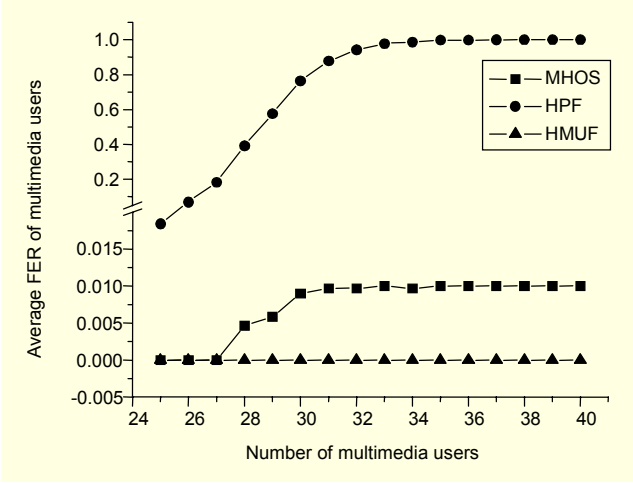


Fig. 4. Average FER of multimedia users for MHOS, HPF, and HMUF.

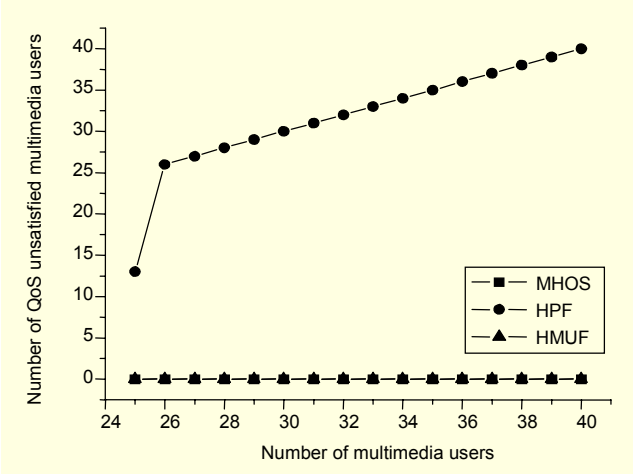


Fig. 5. Number of QoS unsatisfied multimedia users for MHOS, HPF, and HMUF.

Then, the system throughput decrease, because the HPF sometimes selects some users with poor channel conditions to guarantee fairness for all users when the N_u is too big.

From Fig. 2, we can see that the reason MHOS has higher system throughput than HPF and HMUF is that the MHOS has higher throughput of data users. Unlike HMUF, MHOS exploits the channel conditions of all users rather than multimedia users only, so MHOS achieves higher throughput for data users. The HMUF serves data users only if the QoS requirement of all multimedia users is satisfied, so the data users get lower throughput than HPF. With lower throughput for data users, the HMUF can achieve higher system throughput than HPF because the HMUF has higher throughput for multimedia users. The HMUF selects multimedia users with high-quality channels to guarantee the service of multimedia users, but the HPF does not guarantee that.

Figure 3 shows the ratio of throughput per data user to throughput per multimedia user. For MHOS, the throughput per data user is close to the throughput per multimedia user. For HPF, the ratio of throughput per data user to throughput per multimedia user is close to a constant value. That is because the HPF does not distinguish between multimedia user and data user. All the users receive the power resource $\phi_t^{q_j}$ fairly. Because data users have a lower BER than multimedia users, data users get lower throughput than multimedia users with the same power resource $\phi_t^{q_j}$, according to (15). So the ratio is close to a constant value less than 1. Because the HMUF serves data users only if the QoS requirement of all multimedia users is satisfied, the ratio is the smallest and decreases with the increase in the number of multimedia users.

From Figs. 4 and 5, we can see that the MHOS can guarantee the short-term QoS constraint for multimedia users. The average FER is less than the target 0.01 and all users achieve the QoS requirement. Because the HMUF serves the

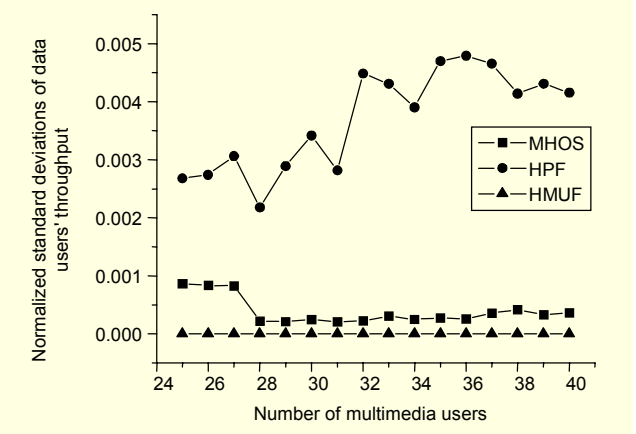


Fig. 6. Normalized standard deviations of data users' throughput for MHOS, HPF, and HMUF.

multimedia user first, the FER is 0. The HPF does not consider the QoS for multimedia users, so the QoS requirement is hard to satisfy for them.

In Fig. 6, we can see that the normalized standard deviations of data users' throughput for MHOS, HPF, and HMUF are all close to 0, that is, every data user's throughput is close to the average throughput per data user. This demonstrates that fairness for data users is guaranteed using our proposed scheme.

V. Conclusion

In this paper, we presented our study of the opportunistic scheduling problem in multiclass service HSUPA systems. We first studied the feasible rate region for HSUPA. Then, considering the feasible rate region constraint, we formulated the opportunistic scheduling problem with QoS constraints and offered a solution.

Our scheduler was designed to ensure QoS under feasible rate region constraints, while employing opportunistic scheduling strategies to increase the total system throughput by selecting users with high-quality channels when possible.

Via simulation, we compare the proposed MHOS algorithm with two baseline algorithms in terms of QoS guarantee and system throughput. Simulations show that the proposed MHOS algorithm guarantees the different QoS constraints, and achieves higher performance than the baseline algorithms.

References

[1] 3GPP, *TS25.309 v6.2.0, FDD Enhanced Uplink Overall Description Stage 2*.

[2] 3GPP, *TR25.896 v6.0.0, Feasibility Study for Enhanced Uplink for UTRA FDD*.

[3] K. Kumaran and L. Qian, "Uplink Scheduling in CDMA Packet-Data Systems," *INFOCOM 2003*, vol. 1, 2003, pp. 292-300.

[4] T. O'Farrell and P. Omiyi, "Low-Complexity Medium Access Control Protocols for QoS Support in Third-Generation Radio Access Networks," *IEEE Trans. on Wireless Communications*, vol. 4, no. 2, Mar. 2005, pp. 743 - 756.

[5] C. Li and S. Papavassiliou, "Joint Throughput Maximization and Fair Scheduling in Uplink DS-CDMA Systems," *IEEE/Sarnoff Symp. on Advances in Wired and Wireless Communication*, 26-27 Apr. 2004, pp. 193-196.

[6] S.A. Jafar and A. Goldsmith, "Adaptive Multirate CDMA for Uplink Throughput Maximization," *IEEE Trans. on Wireless Communications*, vol. 2, no. 2, Mar. 2003, pp. 218-228.

[7] L. Xu, X. Shen, and J.W. Mark," Dynamic Fair Scheduling with QoS Constraints in Multimedia Wideband CDMA Cellular Networks," *IEEE Trans. on Wireless Communications*, vol. 3, no. 1, Jan. 2004, pp. 60-73.

[8] H.C. Akin and K.M. Wasserman, "Resource Allocation and Scheduling in Uplink for Multimedia CDMA Wireless Systems," *2004 IEE Symp. on Advances in Wired and Wireless Communication*, Apr. 2004, pp. 185-188.

[9] C. Rosa, J. Outes, T.B. Sorensen, J. Wigard, and P.E. Mogensen, "Combined Time and Code Division Scheduling for Enhanced Uplink Packet Access in WCDMA," *VTC 2004* , vol. 2, Sep. 2004, pp. 851-855.

[10] C. Rosa, J. Outes, K. Dimou, T.B. Sorensen, J, Wigard, F, Frederiksen, and P.E. Mogensen, "Performance of Fast Node B Scheduling and L1 HARQ Schemes in WCDMA Uplink Packet Access," *VTC 2004*, vol. 3, May 2004, pp. 1635-1639.

[11] X. Liu, E.K.P. Chong, and N.B. Shroff, "Opportunistic Transmission Scheduling with Resource Sharing Constraints in Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, Oct. 2001, pp. 2053-2064.

[12] X. Liu, "Opportunistic Scheduling in Wireless Communication Networks," Ph.D. dissertation, Purdue University, 2002.

[13] X. Liu, E.K.P. Chong, and N.B. Shroff, "A Framework for Opportunistic Scheduling in Wireless Networks," *Computer Networks*, vol. 41, no. 4, Mar. 2003, pp. 451-474.

[14] A. Farrokh and V. Krishnamurthy, "Opportunistic Scheduling for Streaming Users in High-Speed Downlink Packet Access (HSDPA)," *GLOBECOM '04. IEEE*, vol. 6, 2004, pp. 4043-4047.

[15] Y. Liu and E. Knightly, "Opportunistic Fair Scheduling over Multiple Wireless Channels," *INFOCOM 2003*, vol. 2, 2003, pp. 1106-1115.

[16] J.W. Lee, R.R. Mazumdar, and N.B. Shroff, "Opportunistic Power Scheduling for Dynamic Multi-Server Wireless Systems," *IEEE Trans. on Wireless Communications*, vol. 5, no. 6, June 2006, pp. 1506-1515.

[17] C.Z. Li and S. Papavassiliou, "On the Fairness and Throughput Tradeoff of Multi-User Uplink Scheduling in WCDMA Systems," *VTC 2005*, vol. 1, Sep. 2005, pp. 206-210.

[18] S. Kulkarni and C. Rosenberg, "Opportunistic Scheduling: Generalizations to Include Multiple Constraints, Multiple Interfaces, and Short Term Fairness," *Springer Wireless Networks*, vol. 11, no. 5, Sep. 2005, pp. 557-569.

[19] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR a High Efficiency-High Data Rate Personal Communication Wireless System," *Proc. of IEEE VTC 2000*, vol. 3, May 2000.

[20] X. Wang, "An FDD Wideband CDMA MAC Protocol with Minimum-Power Allocation and GPS-Scheduling for Wireless Wide Area Multimedia Networks," *IEEE Trans. Mobile Computing*, vol. 4, no. 1, Jan. 2005, pp. 16-28.

[21] M. Alouini and A. Goldsmith, "Adaptive Modulation over Nakagami Fading Channels," *Wireless Personal Communications*, vol. 13, May 2000, pp. 119-143.

[22] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*, John Wiley and Sons, Ltd., New York, 1990.

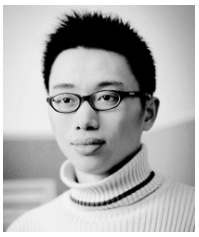
[23] C.Z. Li and S. Papavassiliou, "Fair Channel-Adaptive Rate Scheduling in Wireless Networks with Multirate Multimedia Services," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 10, Dec. 2003, pp. 1604-1614.

[24] C.X. Li and X.D. Wang, "Adaptive Opportunistic Fair Scheduling over Multiuser Spatial Channels," *IEEE Trans. on Communications*, vol. 53, no. 10, Oct. 2005, pp. 1708-1717.

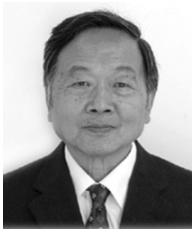
[25] S. Andradóttir, "A Global Search Method for Discrete Stochastic Optimization," *SIAM J. Control Optim*., vol. 6, no. 6, May 1996, pp. 513-530.

[26] F.P. Kelly, A.K. Maulloo, and D.K.H. Tan, "Rate Control in Communication Networks: Shadow Prices, Proportional Fairness and Stability," *J. of the Operational Research Society*, vol. 49, Apr. 1998, pp. 237-252.

[27] E. Lim and S.H. Kim, "Transmission Rate Scheduling with Fairness Constraints in Downlink of CDMA Data Networks," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 1, Jan. 2005, pp. 328-337.



Dan Liao received the BS and MS degrees in electrical engineering from University of Electronic Science and Technology of China, in 2001 and 2004, respectively. He is currently working toward the PhD degree in the School of Communications and Information Engineering, UESTC. His research interests include wireless networks, resource allocation and scheduling, and QoS provisioning.



Lemin Li received the BS degree in electrical engineering from Shanghai Jiaotong University, China, in 1952. From 1952 to 1956, he was with the Department of Electrical Communications at Shanghai Jiaotong University. Since 1956 he has been with the University of Electronic Science and Technology of China. He has been a Visiting Scholar in the Department of Electrical Engineering and Computer Science at the University of California at San Diego, USA, doing research on digital and spread spectrum communications. He is currently a Professor of the School of Communications and Information Engineering, UESTC. He is a member of the Chinese Academy of Engineering. His present research interests are in the areas of communication networks including broadband networks and wireless networks.