Wireless Netw (2011) 17:611–620
DOI 10.1007/s11276-010-0300-0

A low-complexity beamforming-based scheduling to downlink OFDMA/SDMA systems with multimedia traffic

Wen-Ching Chung · Li-Chun Wang · Chung-Ju Chang

Published online: 16 November 2010

Abstract In this paper, we propose a low-complexity beamforming-based scheduling scheme utilizing a semi-orthogonal user selection (SUS) algorithm in downlink orthogonal frequency division multiple access (OFDMA)/space division multiple access (SDMA) systems to support multimedia traffic. One of the challenges in the multi-dimensional (space, time, and frequency) radio resource allocation problem for OFDMA/SDMA systems is its high complexity, especially to simultaneously satisfy the quality of services (QoS) requirements for various traffic classes. In the literature, the SUS algorithm is usually applied to the single-class traffic environment, but extending the SUS algorithm to the multimedia environment is not straightforward because of the need to prioritize the real-time (RT) users and the non-real-time (NRT) users. To solve this problem, we propose the concept of urgency value to guarantee the fairness of the NRT as well as the best effort (BE) users while satisfying the delay requirement for the RT users. Simulation results show that, when traffic load is greater than 0.5, the proposed scheduling algorithm can improve the fairness performance by more than 100% over the most recently proposed algorithms.

Keywords Resource allocation · OFDMA/SDMA · Semi-orthogonal user selection · QoS · Fairness

The original version of this paper was presented at IEEE GLOBECOM 2009, Honolulu, Hawaii, USA, Nov. 30–Dec. 4, 2009.

W.-C. Chung (✉) · L.-C. Wang · C.-J. Chang
Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan, ROC
e-mail: wenching.ece88g@nctu.edu.tw

1 Introduction

The study of orthogonal frequency division multiple access (OFDMA) combined with space division multiple access (SDMA) has recently attracted much attention. OFDMA can mitigate multi-path fading, while SDMA can improve spectral efficiency and thus increase system throughput. Intuitively, combining OFDMA and SDMA can use the advantages of both OFDMA and SDMA to provide high rate transmission for wireless services. However, with multiple degrees of freedom, the problem of radio resource allocation (RRA) in space, frequency, and time for the OFDMA/SDMA system is very challenging, especially when multimedia traffic is considered. How to develop a low-complexity scheduling scheme for supporting multimedia traffic in the OFDMA/SDMA systems thus becomes a crucial issue [1].

In the literature, quality-of-service (QoS) [2–5] and fairness [6–9] are two important performance measures for the RRA algorithms in the OFDMA/SDMA system. In a modern wireless system with multimedia traffic, the QoS requirements could be considered as bit error rate (BER), the minimum required transmission rate, the maximum packet delay tolerance, and the maximum packet dropping ratio. To resolve this, heuristic algorithms to increase the total system data rate and guarantee the minimum data rates for each user have been proposed [3]. In addition, a joint resource allocation algorithm to guarantee the maximum allowable delay for delay sensitive services and the minimum data rate requirement for rate sensitive services has been suggested [4]. Other research took account of the BER requirement in the design of RRA scheme [5]. On the other hand, fairness is another performance measure to fairly allocate radio resources [6–12]. The fairness could be considered as proportional fairness, max-min fairness, and weighted fairness. However, providing fairness among users will sacrifice the system throughput. Thus, in the

context of a multimedia environment, designing a scheduling scheme to achieve an overall optimal trade-off among system throughput, QoS guarantee, and fairness is a challenging issue.

Design of the grouping algorithm is an important issue for SDMA systems [1–6, 13]. An optimal set of cochannel users has been obtained by using the singular value decomposition (SVD) of the channel vectors [14]. The authors of [15, 16] adopted a semi-orthogonal user selection (SUS) algorithm to determine the users of which channel vectors are nearly orthogonal. Computational time for the SUS algorithm is greatly reduced since it directly calculates the orthogonality of channel vectors without SVD. An SDMA grouping algorithm has been proposed according to the correlation and gains of the uses' spatial channels [6]. However, these proposed algorithms only considered a single service class. Extending the grouping algorithm to the multimedia environment must deal with the issue of prioritizing real-time (RT) users and non-real-time (NRT) users.

In this paper, we utilize an SUS algorithm and propose a low-complexity beamforming-based scheduling (LCBS) scheme to support multimedia traffic in OFDMA/SDMA systems. The goal of the LCBS scheme is to maximize system throughput, while guaranteing the QoS requirements for all services and ensuring fairness for the non-real-time (NRT) services and best effort (BE) services as well. Here, the multimedia traffic is grouped into two service sets: the delay sensitive set for real-time (RT) services, and the delay-insensitive services. In the LCBS scheme, to save computational time, we simplify the prioritizing scheme by introducing the urgency value for the users. The LCBS scheme first allocates subchannels to the delay-sensitive users according to the delay requirement. Next, it assigns the remaining resources to the delay-insensitive users according to the ratio of data rates. Therefore, the LCBS scheme not only can guarantee the delay requirements for RT users, but can also provide the proportional fairness to the NRT users and BE users. Moreover, the LCBS scheme utilizes the SUS scheme to select the cochannel users, which reduces the complexity and increases the system throughput for the LCBS scheme.

The remainder of the paper is organized as follows. In Sect. 2, we introduce the system model of OFDMA/SDMA system with multimedia traffic. In Sect. 3, the LCBS algorithm is presented to maximize the system throughput. Performance of the LCBS algorithm is analyzed in Sect. 4 Finally, concluding remarks are given in Sect. 5.

2 System model

2.1 OFDMA/SDMA system

Figure 1 shows the downlink OFDMA/SDMA system with LCBS algorithm, where the base station (BS) is equipped with Q transmitting antennas and N subchannels. Assume that zero-forcing (ZF) beamforming technology [15] is adopted to transmit data streams to K single-antenna mobile stations. The BS allocates subchannel and power, determines modulation order, and controls the beamforming weights. Let the basic unit of resource allocation be one subchannel. In this paper, each subchannel is composed of a adjacent OFDM subcarriers. It has been found that grouping adjacent subcarriers has the highest multiuser diversity and maximum system throughput [17].

In an OFDMA/SDMA system, the time axis is divided into fixed-length frames, each of which has L symbols for OFDMA downlink transmission. Let Ψ_n be the set of subcarriers in subchannel n and $\Omega_n^{(l)}$ be the set of users that are selected to multiplex subchannel n for the lth OFDMA symbol. Note that $|\Omega_n^{(l)}| \leq Q$. The LCBS algorithm is executed at the beginning of each frame. When the coherent time of wireless channel is longer than the frame duration, the channel can then be regarded as fixed within a frame. Let $\mathbf{h}_{k,i}$ be a $1 \times Q$ channel gain vector for user k at subcarrier i. Note that $\mathbf{h}_{k,i}$ is not a function of l since the channel is fixed within a frame. With the ZF transmitted beamforming, the received signal of user k at subcarrier i for the lth OFDMA symbol, denoted by $Y_{k,i}^{(l)}$, is given by

$$Y_{k,i}^{(l)} = \mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(l)}\sqrt{p_{k,i}^{(l)}}X_{k,i}^{(l)} + Z_{k,i}^{(l)} \tag{1}$$

where $\mathbf{w}_{k,i}^{(l)}$ is a $Q \times 1$ ZF beamforming vector, $p_{k,i}^{(l)}$ is the allocated power, $X_{k,i}^{(l)}$ is the data symbol, and $Z_{k,i}^{(l)}$ is the additive white Gaussian noise (AWGN) with zero mean and variance σ^2. The received SNR, denoted by $\mathrm{SNR}_{k,i}^{(l)}$, can be calculated as

$$\mathrm{SNR}_{k,i}^{(l)} = \frac{p_{k,i}^{(l)}|\mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(l)}|^2}{\sigma^2} \tag{2}$$

From (2), it can be seen that the receiver SNR is affected by the beamforming weight. If user k has a high spatial correlation, its received SNR will be poor since the term $\mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(l)}$ is small.

2.2 Power allocation

The allocated power to user k must satisfy the minimum SNR requirement. Let BER_k^* be the BER requirement of user k. According to [2], the minimum required SNR for user k with M-QAM modulation is then obtained by

$$\mathrm{SNR}_k^* = -\frac{\ln(5\mathrm{BER}_k^*)}{1.5}(M-1) \tag{3}$$

Note that (3) represents the approximated required SNR, which is suitable for the design of power allocation in this study. Assume that the received SNR in (2) is equal to the

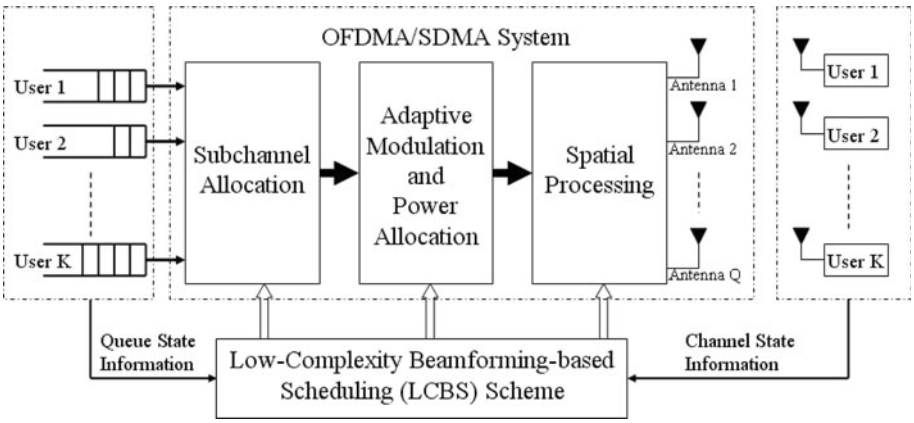


Fig. 1 The downlink OFDMA/SDMA system

minimum required SNR in (3). Then, the allocated power $p_{k,i}^{(l)}$ can be immediately obtained as follows:

$$p_{k,i}^{(l)} = -\frac{\ln(5\mathrm{BER}_{\mathrm{k}}^{*})(\mathrm{M}-1)}{1.5} \cdot \frac{\sigma^2}{|\mathbf{h}_{k,i}\mathbf{w}_{k,i}^{(l)}|^2} \tag{4}$$

Note that the BER requirement of user k is automatically guaranteed by using the power allocation scheme in (4).

2.3 Service classes

In this study, we consider that the OFDMA/SDMA system can support three classes of service: RT, NRT and BE [2]. Each traffic class has different QoS requirements. The QoS requirements for RT services are BER, the maximum packet delay tolerance, and the maximum packet dropping ratio. For NRT services, the QoS requirements include BER and the minimum required transmission rate. The QoS requirement for BE services considers only BER. Denote BER_k^*, R_k^*, D_k^*, and PD_k^* as the BER requirements for user k, the minimum required transmission rate, the maximum packet delay tolerance, and the maximum packet dropping ratio, respectively.

There are four types of traffic in the OFDMA/SDMA system, voice traffic, video traffic, HTTP traffic, and FTP traffic. Each user has one individual queue at the BS. Arriving packets for each user are stored in their own queue in a first-in first-out manner. When the packet delay of RT services exceeds the maximum packet delay, this packet will be dropped by the system. If the buffer of NRT services or BE services has not overflowed, the packets of these services are stored in the queue in the BS without being dropped.

3 Service class awareness scheduling

In this section, we first formulate the RRA problem of the OFDMA/SDMA system as an optimization problem, and then propose the LCBS algorithm to solve this optimization problem.

Since the basic unit of resource allocation is one subchannel, the modulation order assigned to user k at subcarrier i, $i \in \Psi_n$, should be the same. Define $c_{k,n}^{(l)}$ as the assignment variable of modulation order for user k on subchannel n for the lth OFDMA symbol, where $c_{k,n}^{(l)} \in \{0, 1, 2, 3\}$. When $c_{k,n}^{(l)} = 0$, no data is transmitted. For $c_{k,n}^{(l)} = 1$, 2, and 3, the modulation scheme of QPSK, 16-QAM, and 64-QAM are adopted, respectively. Define the assignment vector $\mathbf{c}^{(l)}$ as the solution of the LCBS algorithm for the lth symbol, where $\mathbf{c}^{(l)} \equiv [c_{k,n}^{(l)}]$ is a $K \times N$ vector. Then the throughput of user k can be defined as

$$C_k = \sum_{l=1}^{L}\sum_{n=1}^{N} b \cdot c_{k,n}^{(l)} \tag{5}$$

where $b = 2a$ is the number of transmission bits with the basic QPSK modulation over a subcarries in one subchannel.

Assume that user's traffic type can be detected by the BS. In order to reduce the time to search for the candidate user, the LCBS algorithm groups users into two service sets according to their traffic type. The first set is the delay-sensitive RT users, denoted by Ω_{ds}. The second set is the delay-insensitive NRT and BE users, denoted by Ω_{dis}. Let $P_{k,n}^{(l)}$ denote the allocated power for user k at the lth symbol of subchannel n, where

$$P_{k,n}^{(l)} = \sum_{i \in \Psi_n} p_{k,i}^{(l)} \tag{6}$$

Before the modulation order $c_{k,n}^{(l)}$ is allocated to user k on subchannel n for the lth OFDMA symbol, the system will check $P_{k,n}^{(l)}$ first. Only when the system can support the required transmission power $P_{k,n}^{(l)}$ on subchannel n, we have $c_{k,n}^{(l)} > 0$. If some subcarriers of the nth subchannel are in deep fading, the system requires to ensure the required transmission power at the subcarriers at the subchannel is

large enough to support the selected modulation order subject to the total transmission power constraint. Once the system cannot support $P_{k,n}^{(l)}$ with $c_{k,n}^{(l)} > 0$, no date is transmitted on subchannel n and we set $c_{k,n}^{(l)} = 0$. We then formulate the RRA problem as an optimization problem as follows:

$$(\mathbf{c}^{*(1)}, \ldots, \mathbf{c}^{*(L)}) = \arg \max_{\mathbf{c}^{(1)},\ldots,\mathbf{c}^{(L)}} \sum_{k=1}^{K} C_k \tag{7}$$

subject to the following constraints:

1.

$$|\Omega_n^{(l)}| \leq Q, \quad \forall n, l$$

2.

$$\sum_{n=1}^{N} \sum_{k=1}^{K} P_{k,n}^{(l)} \leq P_T, \quad \forall l$$

3.

$$D_k^* - D_k > 0, \quad \forall k \in \Omega_{ds}$$

4.

$$R_1 : \cdots : R_k = \eta_1 : \cdots : \eta_k, \quad \forall k \in \Omega_{dis}$$

5.

$$\frac{|\mathbf{h}_{k,i}^H \mathbf{h}_{k',i}|}{||\mathbf{h}_{k,i}|| \cdot ||\mathbf{h}_{k',i}||} \leq \alpha, \quad \forall \{k, k'\} \in \Omega_n^{(l)}, i \in \Psi_n, n, l$$

where P_T is the total transmission power for every OFDMA symbol, α is the threshold for semi-orthogonality, $(\cdot)^H$ is the conjugate transpose, D_k is the delay time from the arrival of head-of-line (HOL) packet to the current frame for user k, R_k is the data transmission rate, and η_k is the proportional data rate. Constraint (1) is the *subchannel allocation constraint* since a subchannel at each OFDMA symbol can be allocated to at most Q users. Constraint (2) is the *total system power constraint* so that the total power for each OFDMA symbol has a limitation for downlink data transmission. Constraint (3) refers to the *service fulfillment constraint* for the RT services, and constraint (4) denotes that NRT services and BE services have the requirement for proportional fair rate. We set $\eta_k = 1$, for $k \in \Omega_{dis}$. Constraint (5) is the *semi-orthogonal user constraint* [15] for choosing a group of users such that their channel directions are matching zero-forcing beam directions. Notice that, if α is too large, effective channel gains are reduced due to the loss of the zero-forcing channel inversion. Moreover, if α is too small, the multiuser diversity gain decreases [15]. We set $\alpha = 0.3$.

To get an exact solution for optimization problem (7) is complicated and somewhat impracticable. Hence, we propose the LCBS algorithm to find the suboptimal set of assignment vector $(\mathbf{c}^{*(1)}, \ldots, \mathbf{c}^{*(L)})$. The LCBS algorithm first allocates the radio resources to RT users to satisfy the delay requirement. After that, the LCBS algorithm assigns the remaining resources to the NRT users and BE users to satisfy the requirement for a proportional fair rate. The proposed LCBS algorithm has two parts: (1) an urgency value assignment (UVA) scheme, and (2) an urgency-based resource allocation (URA) scheme.

3.1 Urgency value assignment scheme

In the UVA scheme, the urgency value for each user is calculated according to the packet delay or data transmission rate. The urgency value for delay sensitive set is defined as follows. Let $V_k = D_k^* - D_k$ be a time-to-expiration parameter, which indicates the deadline of HOL packet. If the packet is delivered by the deadline, the QoS requirement in terms of packet delay is satisfied. Define the urgency value for RT traffic as

$$u_k = \begin{cases} 1, & \text{if } V_k \leq V_{th} \\ \frac{D_k}{D_k^*}, & \text{if } V_k > V_{th} \end{cases} \tag{8}$$

where V_{th} is the threshold for V_k. When the value of D_k increases, the urgency degree of user k also increases. The threshold value V_{th} can be set to one if there are sufficient resources to deliver the packet within one frame. However, since the channel condition may be poor or the system may be busy, V_{th} is set to a larger value to satisfy the packet delay requirement.

Now, we define the urgency value for delay insensitive set. Let $\beta_k = R_k/\eta_k$ and $\beta_{max} = \max \beta_k$ for $k \in \Omega_{dis}$. A user with larger β_k is assigned a lower urgency value, since its data rate is higher than other users. Therefore, the urgency value for users in the insensitive set is defined as

$$u_k = 1 - \frac{\beta_k}{\beta_{max}} \tag{9}$$

However, NRT users also have a QoS requirement in terms of the minimum transmission rate. When an NRT user does not achieve the minimum transmission rate, the urgency value for this user should be set to a largest value to get the best service chance. The value of u_k for NRT users is then modified as

$$u_k = \begin{cases} 2, & \text{if } R_k < R_k^* \\ 1 - \frac{\beta_k}{\beta_{max}}, & \text{if } R_k \geq R_k^* \end{cases} \tag{10}$$

3.2 Urgency-based resource allocation scheme

In the URA scheme, radio resources are allocated to selected users based on their traffic type, urgency value,

and spatial diversity gain. In the LCBS algorithm, RT users are serviced first to satisfy the delay requirement. After that, NRT users and BE users are serviced to achieve fair transmission rate. The procedures of the URA scheme for delay-insensitive set are the same as those for the delay-sensitive set. Hence, in this subsection, we describe only the procedures of the URA scheme for the delay-sensitive set.

For user k, denote q_k as the user's buffer occupancy at the beginning of a frame and denote B_k as the number of residual bits of HOL packet. The user with the maximum urgency value should be serviced first. We then have $\Omega = \{k|q_k > 0 \text{ and } u_k = \max u_k, k \in \Omega_{ds}\}$. Since a user with a large HOL packet needs more service time, the candidate set for service is selected based on $\Omega_c = \{k|B_k = \max B_k, k \in \Omega\}$. Define $N_f^{(l)}$ as the set of the free subchannel available for the lth OFDMA symbol. The optimal pair of user and subchannel is selected according to the maximum channel condition,

$$(k^*, n^*) = \arg \max_{k \in \Omega_c, n \in N_f^{(l)}} \sum_{i \in \Psi_n} ||\mathbf{h}_{k,i}|| \tag{11}$$

Set $\Omega_{n^*}^{(l)} = \{k^*\}$ and $N_f^{(l)} = N_f^{(l)} - n^*$. After the optimal pair (k^*, n^*) is obtained from (11), a cochannel user k_{cu}^* is selected for subchannel n^* according to the semi-orthogonal user constraint (5). Note that with larger values for α, the orthogonality is worse. As α increases, the transmitted power increases since the orthogonality condition is not maintained. In this paper, the orthogonality of the selected users in the subchannel is obtained from the average of the orthogonality of each subcarrier in this subchannel. The semi-orthogonal user k_{cu}^* is then selected according to

$$k_{cu}^* = \arg \min_{k \in \Omega_c} \max_{k' \in \Omega_{n^*}^{(l)}} \frac{1}{a} \sum_{i \in \Psi_{n^*}} \frac{|\mathbf{h}_{k,i}^H \mathbf{h}_{k',i}|}{||\mathbf{h}_{k,i}|| \cdot ||\mathbf{h}_{k',i}||} \leq \alpha \tag{12}$$

where a is the number of subcarriers in one subchannel. If k_{cu}^* exists, set $\Omega_{n^*}^{(l)} = \Omega_{n^*}^{(l)} + \{k_{cu}^*\}$; otherwise, find a new candidate set Ω_c and perform (12) again. The procedure of finding the cochannel user is repeated until Q users are selected for subchannel n^*, or no candidate user exists.

After subchannel n^* is allocated to selected users, the modulation order for each selected user is roughly given so that the allocated power of subchannel n^* is less than P_T/N. Since the value of α may be set inappropriately, the selected user for subchannel n^* can be removed if it needs too much power. The system state will be updated after modulation order is assigned. After all the subchannels are allocated to users, the modulation for the selected user will be modified if the total transmission power for this symbol is lower than P_T. An index for increment power is defined by increasing one modulation order for user k at the l-th symbol of subchannel n at the lth symbol as

$$\gamma_{k,n}^{(l)} = \begin{cases} P_{k,n}^{(l)}(\text{BER}_k^*, c_{k,n}^{(l)} + 1) - P_{k,n}^{(l)}(\text{BER}_k^*, c_{k,n}^{(l)}), & \text{if } c_{k,n}^{(l)} = 1 \text{ or } 2 \\ \infty, & \text{otherwise} \end{cases} \tag{13}$$

Here, $P_{k,n}^{(l)}$ is denoted by $P_{k,n}^{(l)}(\text{BER}_k^*, c_{k,n}^{(l)})$ since $P_{k,n}^{(l)}$ is a function of $\text{BER}^*{}_k$ and $c_{k,n}^{(l)}$. The pair of candidate user and subchannel to increase one modulation order is selected by

$$(k^*, n^*) = \arg \min_{k \in \Omega_n^{(l)}, n} \gamma_{k,n}^{(l)} \tag{14}$$

Let the current allocated power be $\hat{P}^{(l)}$. If $\hat{P}^{(l)} + \gamma_{k^*,n^*}^{(l)} < P_T$, set $c_{k^*,n^*}^{(l)} = c_{k^*,n^*}^{(l)} + 1$. Note that the allocated modulation order to a user should not exceed its required transmission rate. After the modulation order for the selected user is modified, the system states for this user are updated immediately. In order to reduce the complexity of the LCBS algorithm, the URA scheme pre-assigns the same resource to the user on the next symbol if this user's buffer is still occupied.

4 Simulation results

4.1 Simulation environment

In this study, the downlink OFDMA/SDMA system is set to be compatible to the IEEE 802.16 standard [18]. Table 1 gives the simulated parameters in the physical layer according to [19]. The path loss is modeled as $128.1 + 37.6 \log R$ dB, where R (in unit of kilometers) is the distance between the BS and the user [20]. The log-normal shadowing has zero mean and standard deviation of 8 dB. In the Stanford University Interim (SUI) channel models, the multipath channel for SUI3 is modeled by 3 taps model [21]. In our simulations, the multipath channel for each antenna has 6 taps of Rayleigh-faded paths with an exponential power delay profile.

Assume that the OFDMA/SDMA system can support four traffic types. The system requirements for each traffic type are given in Table 2 [22]. The first is the voice traffic of RT service. Each voice traffic is modeled as an ON-OFF model. Lengths of ON period and OFF period follow an exponential distribution with means 1.026 and 1.171 s, respectively [22]. The second type is the streaming video traffic [20] of RT service. Each frame of video data arrives at a regular interval of 100 ms and is composed of eight slices (packets). The slice size and the inter-arrival time between slices in a frame are distributed in a truncated Pareto distribution. The third type is the HTTP traffic [20]

Table 1 Parameters for OFDMA/SDMA System

Parameters	Values
Cell radius	1.6 km
Number of antenna to BS (Q)	2
Frame duration	2 ms
System bandwidth	5 MHz
FFT size	512
Subcarrier frequency spacing	11.16 kHz
OFDMA symbol duration	100.8 μ sec
Number of data subcarriers	384
Number of subchannel (N)	8
Number of data subcarriers per subchannel (a)	48
Number of OFDMA symbol for downlink transmission per frame (L)	8
Power allocation to data transmission (P_T)	43.1 dBm
Thermal noise density	−174 dBm/Hz

of NRT service. The HTTP data is modeled as a sequence of page download, and each page download can be considered as a sequence of page arrivals. Both the main object size and the embedded object size act according to a truncated lognormal distribution. Moreover, the reading time and the parsing time in web browsing are distributed in an exponential distribution. The last traffic type is the FTP traffic [20] of BE service. The FTP data is modeled as a sequence of file download, where the file size follows a truncated lognormal distribution. Similarly, the reading time in FTP traffic is distributed in an exponential distribution. The distribution parameters for video, HTTP, and FTP traffics can be found in [20], and so details are omitted here.

In this paper, the LCBS algorithm is compared to the exhaustive search (ES) algorithm and the ARRA algorithm [2]. The ES algorithm is an optimal method which obtains the optimal solution by exhaustively solving the optimal problem (7). However, it is not easy to find a solution such that $R_k : R_{k'} = \eta_k : \eta_{k'}$, for $k \neq k'$, and $k, k' \in \Omega_{dis}$. Hence, we relax the constrain (4) for the ES algorithm so as to maximize system throughput. On the other hand, the performance of the ARRA algorithm is better than conventional algorithms in terms of system throughput and satisfaction extent of QoS requirements. Based on a time-to-expiration (TTE) parameter and the radio resource required by each user, the ARRA algorithm dynamically adjusts the user priority. Moreover, the ARRA algorithm uses the greedy principle to find the best allocation. This can be considered as a joint design of power, subchannel and bit allocation in the physical layer.

4.2 Performance evaluation

In this simulation environment, the maximum system transmission rate is equal to 18.432 Mbps, which is achieved when Q users are multiplexed for each subchannel and the system adopts the highest modulation order to transmit data to each user. The average data rate of each voice, video, HTTP, or FTP traffic is equal to 5.7 kbps, 64 kbps, 14.5 kbps, or 88.9 kbps, respectively. Define the traffic load of the system as the ratio of the total average data rate of users over the maximum system transmission rate. Suppose that the number of users in each traffic type is the same, and let the number of users be increased from 40 to 400. Therefore, the traffic load is varied from 0.094 to 0.939. The following performance measures are investigated in the simulations: (1) system throughput, (2) packet dropping ratio of RT users, (3) mean packet delay of RT users, (4) average transmission rate of NRT users, (5) average transmission rate of BE users, and (6) Jain fairness index (JFI) of NRT and BE users, which is defined as [6]

$$JFI = \frac{\left(\sum_{k=1}^{K} R_k\right)^2}{K \sum_{k=1}^{K} R_k^2} \tag{15}$$

Figure 2 depicts the system throughput versus traffic load for the LCBS algorithm, the ARRA algorithm [2], and the ES algorithm. It is found that the LCBS algorithm decreases the system throughput by up to only 2.7% compared to the ARRA algorithm, and by up to 5.3% compared to the ES scheme when the traffic load is over 0.75. Both of the LCBS algorithm and the ARRA algorithm take throughput maximization as one of the design objectives. However, the system throughput for the LCBS algorithm is sacrificed when the transmission fairness for delay insensitive is considered. The LCBS algorithm utilizes the property of beamforming in the physical layer to enhance the system throughput. On the

Table 2 The system requirement for each traffic type

	Voice (RT)	Video (RT)	HTTP (NRT)	FTP (BE)
Required BER	10^{-3}	10^{-4}	10^{-6}	10^{-6}
Maximum packet delay tolerance	40 ms	10 ms	N/A	N/A
Maximum packet dropping ratio	1%	1%	N/A	N/A
Minimum required transmission rate	N/A	N/A	100 kbps	N/A

N/A Not Applicable

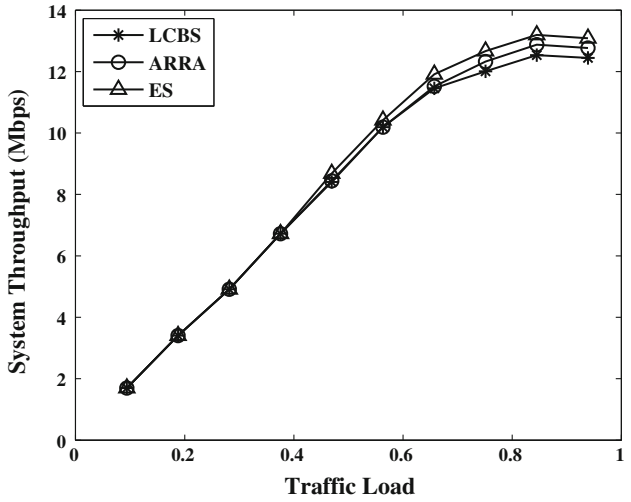


Fig. 2 Comparison of system throughput for LCBS, ARRA, and ES algorithms

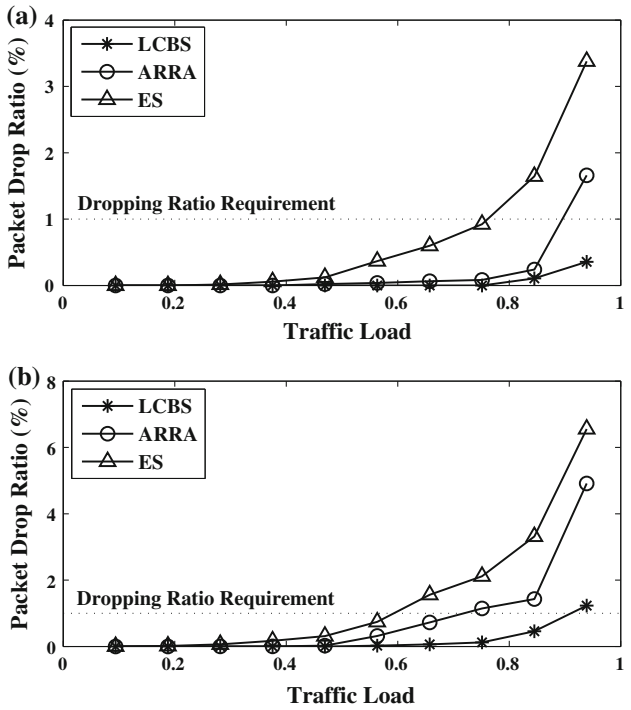


Fig. 3 Comparison of LCBS, ARRA, and ES algorithms in terms of **a** packet drop ratio of voice traffic, and **b** packet drop ratio of video traffic

other hand, the LCBS algorithm has the best system performance in term of packet drop ratio of RT users, as given in Fig. 3. This also can enhance the system throughput for the LCBS algorithm. Hence, the system throughput of the LCBS algorithm is close to that of the ARRA algorithm and the ES algorithm.

The packet drop rate versus traffic load is given in Fig. 3. It can be seen that, when the traffic load is great than 0.72 (0.59), the ARRA (ES) algorithm violates the QoS requirement of maximum packet dropping ratio for RT users. However, the LCBS algorithm can guarantee the dropping ratio requirement for voice traffic when traffic load is heavy. Even for video traffic, this requirement is still satisfied by the LCBS algorithm until the traffic load is over 0.91. The reason for this phenomenon is that the LCBS algorithm first allocates the radio resource to RT users. In the LCBS algorithm, according to the design of urgency value for RT users, the service time for voice traffic is longer than that for video traffic since the maximum delay tolerance for voice traffic is longer than that for video traffic. This means that the voice packet can be delivered before the maximum delay requirement. On the other hand, the ARRA algorithm promotes the RT users with larger packet delays so as to deliver the NRT packet first. This causes that the priority of NRT users to be higher than that of RT users when the NRT users are urgent. As the number of NRT users increases, the RT packet can not be delivered within the maximum delay tolerance, and so the packet drop ratio of RT users for the ARRA algorithm is higher than that for the LCBS algorithm. However, the ES algorithm is mainly to maximize the system throughput. Therefore, the amount of bandwidth granted to the RT user might not be sufficient enough. This makes that the ES algorithm has the worst packet drop rate. The packet delay versus traffic load is shown in Fig. 4. It is found that the LCBS algorithm improves the delay performance of voice traffic by up to 85%. The reasons for this are the same as those given in Fig. 3 for that the increase of packet dropping ratio.

The average transmission rate of NRT and BE users is given in Fig. 5. All three compared algorithms can satisfy the QoS requirement of minimum transmission rate for HTTP traffic. The transmission rate of FTP traffic (resp. HTTP traffic) for the ARRA algorithm is greater (resp. less) than that for the LCBS algorithm, due to the reasons given below. In order to enhance the system throughput, the ARRA algorithm allocates the radio resource to users according to their priority. After the radio resource is allocated to a user, the user priority is adjusted immediately. When priorities of RT users and NRT users are the same as that of BE users, the ARRA algorithm allocates more radio resources to users with larger effective linkgain to maximize the system throughput. Since the FTP traffic has a higher arrival rate, it has the higher transmission rate when the priorities of all users are the same. This situation is usually observed when the traffic load is light. On the other hand, the LCBS algorithm dynamically adjusts the urgency value of each FTP user according to its transmission rate. Therefore, the transmission rate of FTP user for the LCBS algorithm is decreased.

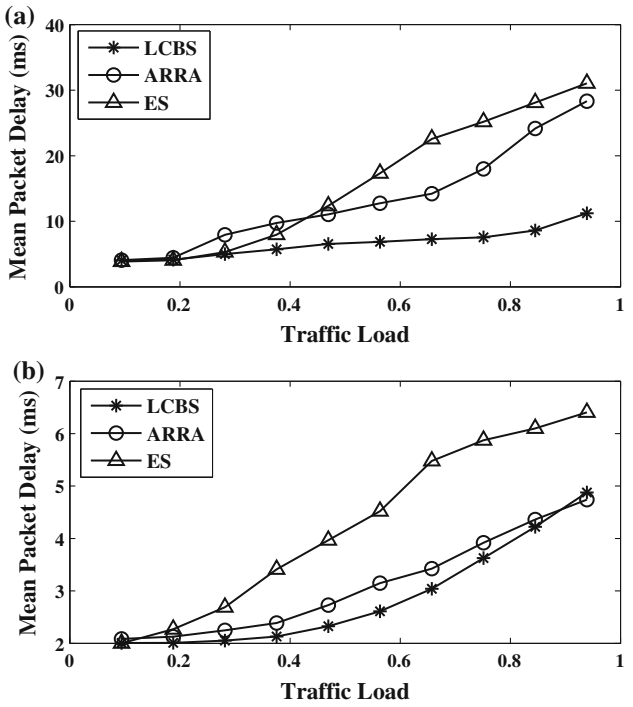


Fig. 4 Comparison of LCBS, ARRA, and ES algorithms in terms of **a** mean packet delay of voice traffic, and **b** mean packet delay of video traffic

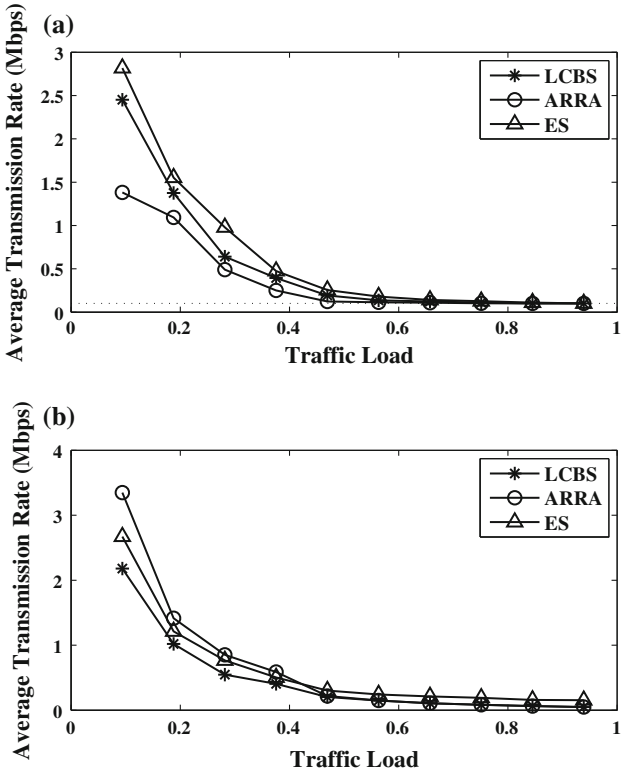


Fig. 5 Comparison of LCBS, ARRA, and ES algorithms in terms of **a** average transmission rate of HTTP traffic, and **b** average transmission rate of FTP traffic

Figure 6 shows the Jain fairness index (JFI) versus traffic load. The LCBS algorithm has the best performance for JFI. It can be seen that the JFI for the ARRA algorithm is heavily decreased as the traffic load is increased. When the traffic load is greater than 0.5, the LCBS algorithm improves the fairness performance by more than 100% for the reasons given below. The LCBS algorithm considers the transmission fairness of NRT users and BE users, and assigns a higher priority to the delay-insensitive user with the minimum transmission rate. Therefore, the LCBS algorithm can substantially improve the fairness performance for HTTP traffic and FTP traffic. Moreover, since the urgency value for each user is good design and the property of beamforming is used, the system throughput for the LCBS algorithm decreases only 2.7% compared to the ARRA algorithm, and 5.3% compared to the ES algorithm. Although the ES algorithm tries to find an optimal solution for the optimal problem (7), it relaxes the fair rate constraint so as to maximize the system throughput. Therefore, the JFI for the ES algorithm is in the middle. On the other hand, the ARRA algorithm gives a higher transmission rate to users with larger effective linkgain to enhance throughput when the QoS requirement of minimum required transmission rate for NRT users is satisfied. This, however, leads to an unfair transmission rate between the NRT users and BE users for the ARRA algorithm.

4.3 Complexity

The computational complexity of the LCBS algorithm is discussed here. The UVA scheme sequentially sets the urgency value for all users. Hence, the complexity of the UVA scheme is $O(K)$. The URA scheme first looks for a best pair of user k^* and subchannel n^* from K users and N subchannels. This complexity is $O(KN)$. Based on this best pair, the URA scheme uses (12) to search for other semi-orthogonal users, but this complexity can be ignored compared with $O(KN)$. The complexity of power allocation in the URA scheme can be ignored compared with that of subchannel allocation, since the power is allocated to selected users according to their effective linkgain. The number of these iterations is bounded by LN since each frame has L symbols and each symbol has N subchannels. In the worse case, the computational complexity for the URA scheme is $O(LKN^2)$, so the overall complexity of LCBS algorithm is $O(LKN^2)$. In practice, the URA scheme allocates the same radio resource to select users for the next several symbols. Hence, the complexity of LCBS algorithm would be greatly reduced by almost L times. On the other hand, the ES algorithm has a computational complexity in the order of $O((LKN)^Q)$, while the ARRA algorithm has a computational complexity in the order of

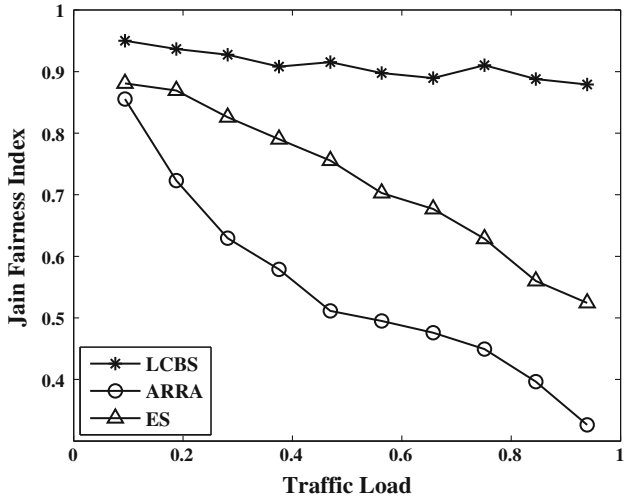


Fig. 6 Comparison of LCBS, ARRA, and ES algorithms in terms of Jain fairness index for delay insensitive set

$O(LKQN^2)$. As described above, the LCBS algorithm can efficiently solve the multi-dimensional (space, time, and frequency) RRA problem and has lower complexity than the ES and ARRA algorithms.

5 Conclusions

In this paper, we propose a low-complexity beamforming-based scheduling (LCBS) scheme to reduce the complexity of scheduling scheme for an OFDMA/SDMA system with multimedia traffic. By utilizing a semi-orthogonal user selection (SUS) algorithm, the multi-dimensional (space, time, and frequency) RRA problem can be efficiently solved by the LCBS algorithm. The purpose of the LCBS algorithm is to maximize the system throughput while guaranteeing the QoS requirement for all users and providing a fair transmission rate for both NRT users and BE users. Therefore, the LCBS scheme is designed with a good trade-off between the implementation complexity and system performance in throughput, QoS guarantee, and fairness guarantee.

The UVA scheme sets the simple urgency value for all users according to their QoS requirements or their transmission rate. The URA scheme first looks for a best pair of user and subchannel. Based on this best pair, the URA scheme searches other cochannel users by adopting semi-orthogonal users selection scheme. After that, the transmitted power is allocated to selected users based on their effective linkgain. Simulation results show that, compared to the ARRA algorithm, the LCBS algorithm can not only achieve the similar system throughput, but also improve the system performance, especially in the terms of packet dropping ratio of voice and video traffics, packet delay of voice traffic, and fair transmission rate of HTTP and FTP traffic.

Acknowledgments This work was supported by the National Science Council of Taiwan, ROC, under contract number NSC 96-2628-E-009-004-MY3, NSC 97-2221-E-009-098-MY3, NSC 97-2221-E-009-099-MY3, and the Ministry of Education under the ATU plan.

References

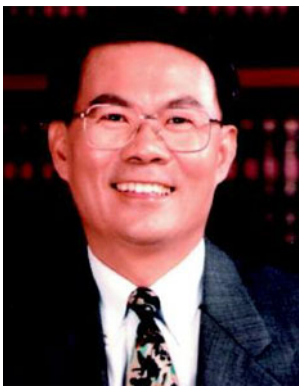
1. Lau, V. K. N. (2005). Optimal downlink space-time scheduling design with convex utility functions-multiple-antenna systems with orthogonal spatial multiplexing. *IEEE Transations on Vehicular Technology, 54*(4), 1322V1333.
2. Tsai, C.-F., Chang, C.-J., Ren, F.-C., & Yen, C.-M. (2008). Adaptive radio resource allocation for downlink OFDMA/SDMA systems with multimedia traffic. *IEEE Transactions on Wireless Communications, 7*(5), 1734–1743.
3. Koutsopoulos I., & Tassiulas, L. (2008). The impact of space division multiplexing on resource allocation: A unified treatment of TDMA, OFDMA and CDMA. *IEEE Transactions on Communications, 56*(2), 260–269.
4. Xu, H., Tian, H., Feng, Y., Gao, Y., & Zhang, P. (2008). An efficient resource management scheme with Guaranteed QoS of heterogeneous services in MIMO-OFDM system. In *Proceedings of IEEE WCNC'08*, NV. pp. 1838–1843.
5. Tsang, Y. M., & Cheng, R. S. (2004). Optimal resource allocation in SDMA/muilti-input-single-output/OFDM systems under QoS and power constraints. In *Proceedings of IEEE WCNC'08*, GA. pp. 1595–1600.
6. Maciel, T. F., & Klein, A. (2007). A resource allocation strategy for SDMA/OFDMA systems. In *Proceedings of IEEE ISTMWC'07*, Budapest, Hungary.
7. Kim, H., & Han, Y. (2005). A proportional fair scheduling for multicarrier transmission systems. *IEEE Communications Letters, 9*(3), 210–212.
8. Nguyen, T.-D., & Han, Y. (2006). A proportional fairness algorithm with QoS provision in downlink OFDMA systems. *IEEE Communications Letters, 10*(11), 760–762.
9. Thoen, S., Perre, L. V., Engels, M., & Man, H. D. (2002). Adaptive loading for OFDM/SDMA-based wireless networks. *IEEE Transactions on Communications, 50*(11), 1798–1810.
10. Radunovic, B., & Le Boudec, J.-Y. (2007). A unified framework for max-min and min-max fairness with applications. *IEEE/ACM Transactions on Networking, 15*(5), 1073–1083.
11. Cheng, R.-G., Chang, C.-J., Shih, C.-Y., & Chen, Y.-S. (2006). A new scheme to achieve weighted fairness for WLAN supporting multimedia services. *IEEE Transactions on Wireless Communications, 5*(5), 1095–1102.
12. Banchs, A. (2002). User fair queuing: Fair allocation of bandwidth for users. In *Proceedings of IEEE INFCOM'02*, NY. pp. 1668–1677.
13. Song, X., Hu, N., He, Z., Niu, K., Wang, X., & Wu, W. (2007). A cross-layer design for downlink scheduling in SDMA packet access networks. In *Proceedings of IEEE 65th VTC*, Dublin, Ireland. pp. 1016–1020.
14. Kim, J., Part, S., Lee, J. H., Lee, J., & Jumg, H. (2005). A scheduling algorithm combined with zero-forcing beamforming for a multiuser MIMO wireless system. In *Proceedings of IEEE 62nd VTC*, TX. pp. 211–215. .
15. Yoo, T., & Goldsmith, A., (2006). On the optimality of multi-antenna broadcast scheduling using zero-forcing beamforming.

IEEE Journal on Selected Areas in Communications, 24(3), 528–541.
16. Swannack, C., Uysal-Biyikoglu, E., & Wornell, G.W. (2004). Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel. In *Proceedings of Allerton Conference on Communications, Control and Computer*, 756–765.
17. Shen, M., Li, G., & Liu, H. (2005). Effective of traffic channel configuration on the orthogonal frequency division multiple access downlink performance. *IEEE Transaction on Wireless Communications, 4*(4), 1901–1913.
18. IEEE 802.16-2004. (2004). "IEEE Standard for local and metropolitan area networks-part 16: Air interface for fixed broadband wireless access systems."
19. Yaghoobi, H. (2004). Scalable OFDMA physical layer in IEEE 802.16 WirelessMAN. *Intel Techonology Journal, 8*(3), 201–212.
20. 3GPP TR 25.892. (Jun. 2004). "3GPP Technical Report for feasibility study for OFDM for UTRAN enhancement."
21. IEEE 802.16.3c-01/29r3. (2001). "Channel models for fixed wireless application."
22. WiMAX forum V.2.0. (Dec. 2007). "WiMAX Technical Report for Wimax system evaluation methodology."

Author Biographies

Wen-Ching Chung was born in Taiwan, ROC, in June 1977. He received B.E. and Ph.D. degrees in electrical engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1999 and 2006, respectively. Since January 2008, he has joined the Department of Electrical Engineering of National Chiao Tung University in Taiwan as a Post Doctor. His research interests are in the areas of radio resources management for wireless communication networks and nonlinear control.

Li-Chun Wang (S'92-M'96-SM'06) received the B.S. degree in electrical engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 1986, the M.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1988, and the M.Sc. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1995 and 1996, respectively. In 1995, he was affiliated with Northern Telecom in Richardson, Texas. From 1996 to 2000, he was with AT&T Laboratories, where he was a Senior Technical Staff Member in the Wireless Communications Research Department. Since August 2000, he has joined the Department of Communication Engineering of National Chiao Tung University in Taiwan as an Associate Professor and has been promoted to a full professor since August 2005. Dr. Wang was a corecipient of the Jack Neubauer Best Paper Award from the IEEE Vehicular Technology Society in 1997. His current research interests are in the areas of cellular architectures, radio network resource management, cross-layer optimization for cooperative and cognitive wireless networks. He is the holder of three U.S. patents with three more pending.



Chung-Ju Chang was born in Taiwan, ROC, in August 1950. He received the B.E. and M.E. degrees in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1972 and 1976, respectively, and the Ph.D. degree in electrical engineering from National Taiwan University, Taiwan, in 1985. From 1976 to 1988, he was with Telecommunication Laboratories, Directorate General of Telecommunications, Ministry of Communications, Taiwan, as a Design Engineer, Supervisor, Project Manager, and then Division Director. He also acted as a Science and Technical Advisor for the Minister of the Ministry of Communications from 1987 to 1989. In 1988, he joined the Faculty of the Department of Communication Engineering, College of Electrical Engineering and Computer Science, National Chiao Tung University, as an Associate Professor. He has been a Professor since 1993 and a Chair Professor since 2009. He was Director of the Institute of Communication Engineering from August 1993 to July 1995, Chairman of Department of Communication Engineering from August 1999 to July 2001, and Dean of the Research and Development Office from August 2002 to July 2004. Also, he was an Advisor for the Ministry of Education to promote the education of communication science and technologies for colleges and universities in Taiwan during 1995–1999. He is acting as a Committee Member of the Telecommunication Deliberate Body, Taiwan. Moreover, he once served as Editor for IEEE Communications Magazine and Associate Editor for IEEE Transactions Vehicular Technology. His research interests include performance evaluation, radio resources management for wireless communication networks, and traffic control for broadband networks. Dr. Chang is members of the Chinese Institute of Engineers (CIE) and the Chinese Institute of Electrical Engineers (CIEE).