GGRA: A Feasible Resource-Allocation Scheme by Optimization Technique for IEEE 802.16 Uplink Systems

Yin Chiu, Chung-Ju Chang, *Fellow, IEEE*, Kai-Ten Feng, *Member, IEEE*, and Fang-Ching Ren, *Member, IEEE*

***Abstract*—Generally, optimization techniques for resource allocation of orthogonal frequency-division multiple-access (OFDMA) systems are infeasible for real-time applications. In this paper, with consideration of grouping for subscriber stations (SSs), a resource-allocation scheme by an *optimization technique* of a genetic algorithm is proposed for the uplinks of IEEE 802.16 OFDMA systems. The genetic algorithm with SS grouping resource-allocation (GGRA) scheme first designs a rate assignment strategy that is applied with a predefined residual lifetime to dynamically allocate resource to each service. It then aggregates high-correlation SSs into the same group, where the SSs will be allocated to different slots to avoid mutual user interference. Finally, the GGRA scheme finds an optimal assignment matrix for the system by the genetic algorithm based on the SS groups to greatly lessen the computation complexity. The GGRA scheme can also maximize the system throughput and fulfill the quality-of-service (QoS) requirements. Simulation results show that the proposed GGRA scheme performs better than the efficient and fair scheduling (EFS) algorithm and the maximum largest weighted delay first (MLWDF) algorithm in system throughputs, voice/video packet drop rates, unsatisfied ratios of hypertext transfer protocol (HTTP) users/packets, and file transfer protocol (FTP) throughputs. The computation complexity of the GGRA scheme is also tractable and, thus, feasible for real-time applications.**

***Index Terms*—Genetic algorithm (GA), orthogonal frequency-division multiple access (OFDMA), quality of service (QoS), resource allocation, subscriber station (SS) grouping, uplink (UL).**

I. Introduction

THE IEEE standard 802.16 has become a popular broadband wireless-access technology. The IEEE 802.16-2004 [1] specification defines the frame structure for uplink (UL) and downlink (DL) and contains data-transmission techniques such as quality of service (QoS), power control, adaptive modulation and coding, and orthogonal frequency-division multiple access (OFDMA). IEEE 802.16e [2] is specified to support high-mobility communications for wireless metropolitan area networks. On the other hand, multiple-input multiple-output (MIMO) has also been a popular technique in recent years. It provides space diversity to recover the data from a time-variant channel. The technique provides us degrees of freedom to make resource allocations.

In the UL of IEEE 802.16, subscriber stations (SSs) send bandwidth requests to the base station (BS), and the BS takes the responsibility of allocating the service traffic of UL SSs to the UL subframe. There are four types of service traffic with differentiated QoS requirements in IEEE 802.16. The BS should allocate the resource and guarantee the QoS requirements. Since IEEE 802.16 defines a grant-per-SS (GPSS) structure, the BS allocates the resource to each SS, and each SS should have a scheduler to map the service flow to the dedicated resource. Therefore, a layered structure is considered: a BS bandwidth allocator and an SS scheduler. It is shown in [3] that this structure can achieve better performance.

Usually, such a resource allocation is mathematically formulated into an optimization problem. If the solution of the optimization problem were found by the exhaustive search, it would take a lot of computation time and become infeasible. Therefore, a lot of research by heuristic algorithms for suboptimal solution was proposed. Sun *et al.* [4] considered a single-carrier modulation scheme (WirelessMAN-SC) and proposed a queue-aware scheduling algorithm using a threshold set to each service flow to define their urgent levels. The method of a threshold set is useful in controlling the queue length such that the service flows may not wait too long and violate the QoS requirements. However, the algorithm considered a single-carrier modulation scheme, which is not applicable for the 2-D allocation for the OFDMA scheme. Ben-Shimol *et al.* [5] considered a 2-D mapping scheme using a rectangular-shaped allocation, which is consistent with the allocation format of the IEEE 802.16 UL OFDMA mode. The scheme separates the allocation into two parts, i.e., a scheduler and a mapper, and it simply considers the frame length and tries to maximize the allocation efficiency with or without service priority. However, the channel response due to the frequency-selective fading channel is not considered, which is also impractical. Yen *et al.* [6] proposed a dynamic priority resource-allocation (DPRA) scheme for IEEE 802.16 UL systems. The DPRA scheme dynamically gives priority values to four types of service traffic based on their urgency degrees and allocates system radio resources according to their priority values. It can maximize the system throughput and satisfy differentiated QoS requirements. Tsai *et al.* [7] also proposed an adaptive radio resource allocation (ARRA) for the DL OFDMA/space-division multiple-access system. It defines a priority value to

Manuscript received July 1, 2009; revised October 10, 2009. First published January 12, 2010; current version published March 19, 2010. The review of this paper was coordinated by Prof. P. Langendoerfer.

Y. Chiu, C.-J. Chang and, K.-T. Feng are with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu 300, Taiwan (e-mail: cjchang@mail.nctu.edu.tw).

F.-C. Ren is with the Industrial Technology Research Institute, Hsinchu 31040, Taiwan.

Digital Object Identifier 10.1109/TVT.2009.2040005

dynamically allocate the resource to urgent SSs. Simulation results show that ARRA can guarantee QoS and achieve good system throughput. However, it considers the DL case, and the allocation can be further improved by global allocation methods.

Furthermore, some research made effort to reduce the computational complexity for practical applications. For example, Wang *et al.* [8] proposed a tree-structure-based method. It has great reduction since its complexity is in a linear increment, rather than exponential, with the increase in the number of users and antennas. However, it is designed for antenna selection only and lacks QoS requirement consideration. As a result, it is not suitable for systems that support multimedia traffic with QoS constraints. Singh and Sharma [9] proposed an efficient and fair scheduling (EFS) algorithm. This algorithm defines a fixed-priority allocation policy such that high-priority service with high-channel gain will be allocated first. This heuristic approach can achieve high system throughput and satisfy the QoS requirement in most cases. However, the fixed-priority allocation policy would lack flexibility. Park *et al.* [10] proposed an algorithm based on the maximum largest weighted delay first (MLWDF) algorithm, which can dynamically adjust the priority of unsolicited grant service (UGS), real-time polling service (rtPS), and nonreal-time polling service (nrtPS) traffic. This algorithm can improve the QoS fulfillment and maintain the good performance. Both the EFS and MLWDF algorithms use the heuristic approach, which choose one SS at a time. The performance can be further improved if optimization techniques are used, which completely plan the allocation for the whole frame.

An optimization technique that takes advantage of the nature behavior is called the genetic algorithm (GA). The GA is based on the nature of survival of the fittest. It randomly generates the population, using crossover and mutation methods, and evaluates chromosomes by a well-defined fitness function. It performs well for finding an optimal solution if a good convergence strategy is defined. Moreover, the computation time can be limited by setting an iteration upper bound. Therefore, it is helpful in many dimensions to solve the optimization problem. It is widely used in manufacturing problems such as job-shop scheduling or a flow-shop problem. There is also some GA applied in communications. Reddy *et al.* [11] proposed an efficient orthogonal frequency-division multiplexing (OFDM) scheduling algorithm based on the GA to do the subchannel selection and power loading; Guo *et al.* [12] proposed an antenna-selection method based on the GA for MIMO systems. However, it is found that this previous research faces several problems, such as QoS requirement guarantee and computational tractability, while conforming to the IEEE 802.16 UL frame format.

In this paper, we propose the GA with SS grouping resource-allocation (GGRA) scheme for ULs of IEEE 802.16 wireless communication systems. The GGRA scheme first designs a rate-assignment strategy (RAS) and formulates the resource-allocation problem as an optimization equation that is subject to system constraints. The RAS is to allocate a resource for QoS guarantee and UL frame structure conformity. Also, to make the computation tractable, the GGRA scheme considers the interference due to user channel *correlation* and *aggregates* those SSs that have high correlation in the *same group*. Then, it adopts the GA to solve the optimization problem and obtain an optimal assignment matrix for the UL of the IEEE 802.16 communication system. Notice that these groups turn to the genes of chromosomes in the GA, where the number of groups (genes) would be much smaller than the number of original SSs. Then, the individuals in the GA are represented by arbitrary permutation of group orders and used to decide the frame allocation. If there is still remaining resource after the GA, a residual resource reallocation (RRR) method, which is contained in the GGRA, is designed to enhance the system throughput. Consequently, the GGRA scheme can efficiently find the optimal solution and maximize the system throughput. Simulation results show that the proposed GGRA scheme achieves higher system throughput than the EFS algorithm [9] and the MLWDF algorithm [10] by 18% and 9%, respectively. The unsatisfied ratio of hypertext transfer protocol (HTTP) users of the GGRA scheme is also the lowest among the three schemes. Furthermore, the GGRA scheme attains higher file transfer protocol (FTP) throughput than the EFS and MLWDF algorithms by 30% and 14%, respectively. The computation time of the GGRA is also reasonable and feasible for real-time applications.

The remaining part of this paper is organized as follows. Section II describes the system model, including the virtual MIMO system, service traffic types, and user correlation. Section III introduces the proposed RAS and formulates the UL resource allocation into an optimization problem. Section IV presents the proposed GGRA scheme to find a kind of optimal solution for the problem. Section V illustrates simulation results and discussions. Finally, concluding remarks are given in Section VI.

II. System Model

A. OFDMA System With Virtual MIMO

Assume that the UL of the IEEE 802.16 wireless communication system is in the OFDMA mode with adjacent subcarrier grouping. There are N subchannels in the system, and one subchannel is composed of q subcarriers. A single cell consisting of one BS and K SSs (users) is considered. Each SS is equipped with one transmit antenna only, whereas a BS has M receiving antennas. These single-antenna SSs and the M-antenna BS form a virtual-MIMO system with an M-by-K channel matrix if the distributed antennas among K SSs can be seen as a virtual multiple input. Due to the space diversity, the MIMO system can have many SSs transmit data through the same subchannel as long as the rank constraint is fulfilled.

The OFDMA/time-division-duplex frame structure defined in IEEE 802.16 contains two subframes: one for DL and the other for UL. The duration of DL and UL subframe is decided by the BS and broadcast through the DL map (DL-MAP) and UL map (UL-MAP) messages to each associated SS. Assume that there are L symbols in the UL map and that the basic allocation unit is a slot, which is rectangular in shape and composed of one subchannel and one OFDMA symbol.

IEEE 802.16 supports four types of services, including real time and nonreal time. Each service type is stated as follows.

1) UGS, which supports constant-bit-rate real-time traffic, such as voice transmission. UGS flows should be granted a fixed amount of resource in each frame, and there is no bandwidth request that is needed for UGS.
2) rtPS, which is designed for variable-size real-time data transmission, such as video streaming. It is a delay-sensitive traffic so that the delay requirement is an important QoS issue for the rtPS. The resource allocation for this type of service should be dynamically arranged according to the packet-delay requirement of the SS, and the packet dropping rate should be under its requirement. The rtPS can use a polling mechanism to request more bandwidths if there is a large amount of service flow in the buffer or if the delay requirement is going to be violated.
3) nrtPS, which is designed to support delay-tolerant data streams, and a requirement of a minimum transmission rate for the SS is needed. There are nrtPS flows such as the HTTP traffic. The nrtPS flow can use a polling service to request bandwidth according to the buffer condition or the minimum rate requirement.
4) Best effort (BE), which is non-QoS guaranteed. The BE service is always the lowest priority and will be transmitted when the system resource is still available after scheduling UGS, rtPS, and nrtPS. Therefore, the BE can only use the remaining resource in the system.

In this paper, the proposed GGRA scheme will use an RAS with a predefined *residual lifetime*, which represents the number of available frame delays according to the head-of-line packet size and the delay bound (minimum rate requirement) for UGS and rtPS (nrtPS) traffic, to dynamically adjust the priority of each service traffic to guarantee the QoS requirement. The details of the RAS will be described in Section III.

B. Power Allocation

To achieve the target bit error rate (BER), the transmission power for each SS is needed to be determined to combat noise and interference while maximizing the system throughput. A normalized quadratic-amplitude modulation (QAM) is used so that the data symbol has unitary mean energy. The QAM includes quadrature phase shift keying (QPSK), 16-QAM, and 64-QAM. Also, the channel response of the system is assumed to be almost constant during the interval between the SS report and the transmission. Therefore, the channel gain of user k on subchannel n, which is denoted by $h_{k,n}$, is constant throughout the time interval. The additive white Gaussian noise (AWGN) with zero mean and variance σ^2 is assumed to be applied to the transmitted signal in the channel. Finally, the received signal-to-interference-plus-noise ratio (SINR) of user k on subchannel n at the ℓth OFDMA symbol, which is denoted by $\mathrm{SINR}_{k,n}^{\ell}$, can be obtained by

$$\mathrm{SINR}_{k,n}^{\ell} = \frac{\rho_{k,n}^{\ell} |h_{k,n}|^2}{\sum\limits_{k' \neq k} \rho_{k',n}^{\ell} |h_{k',n}|^2 + \sigma^2} \tag{1}$$

where $\rho_{k,n}^{\ell}$ is the allocated power of user k of each subcarrier on subchannel n at the ℓth OFDMA symbol, $1 \leq k$, $k' \leq K$, $1 \leq n \leq N$, and $1 \leq \ell \leq L$. There is a tight bound of the BER when using Q-QAM at a given SINR value, which is derived in [13] and is given by

$$\mathrm{BER} \leq 0.2 e^{-1.5 \times \frac{SINR}{Q-1}}. \tag{2}$$

Therefore, we can get the minimum SINR of user k, which is denoted by SINR_k^*, to achieve the target BER for user k, which is denoted by BER_k^*, by

$$\mathrm{SINR}_k^* = \frac{-(Q-1)\ln\left(5\mathrm{BER}_k^*\right)}{1.5}. \tag{3}$$

Since SINR_k^* is determined by the power of user k and other users that are scheduled in the same subchannel, it is quite complicated to determine the transmission power by a total-user consideration. Zhang and Letaief [14] have proposed a method to deal with the spatial multiple access of MIMO-OFDM systems. The mutual channel correlation for each pair of users k_1 and k_2, which is denoted by Cor_{k_1,k_2}, is assumed to be independent of the subchannel and is defined by (4), shown at the bottom of the page, where j is the imaginary part of the complex number, d is the distance between the two transmit antennas, $\theta_{k_i}^{Rx}$ is the direction of arrival of k_i's signal, $i = 1$ or 2, and λ is the wavelength of the subcarrier. It also shows that the user interference can be almost perfectly cancelled by the multiuser detector if we put low-correlation users into the same subchannel [11], [14]. Thus, if low (high) correlation SSs are aggregated into the different (same) group and allocated to the same (different) slots, the interference term in (1) can be neglected, and $\rho_{k,n}^{\ell}$ to achieve BER^* can be obtained by

$$\rho_{k,n}^{\ell} = \frac{-(Q-1)\ln(5\mathrm{BER}^*)\sigma^2}{1.5|h_{k,n}|^2}. \tag{5}$$

Finally, the power of user k in subchannel n, which is denoted by $p_{k,n}^{\ell}$, can be obtained by

$$p_{k,n}^{\ell} = q \cdot \rho_{k,n}^{\ell}. \tag{6}$$

$$|Cor_{k_1,k_2}| = \begin{cases} \frac{1}{M} \left| \frac{1-\exp\left(j2\pi\frac{d}{\lambda}M\left(\sin\left(\theta_{k_1}^{Rx}\right)-\sin\left(\theta_{k_2}^{Rx}\right)\right)\right)}{1-\exp\left(j2\pi\frac{d}{\lambda}\left(\sin\left(\theta_{k_1}^{Rx}\right)-\sin\left(\theta_{k_2}^{Rx}\right)\right)\right)} \right|, & \text{if } \theta_{k_1}^{Rx} \neq \theta_{k_2}^{Rx} \\ 1, & \text{if } \theta_{k_1}^{Rx} = \theta_{k_2}^{Rx} \end{cases} \tag{4}$$

III. Rate-Assignment Strategy and Problem Formulation

An RAS is designed for QoS guarantee and efficient resource allocation. The RAS gives a priority to the traffic of each service of the SS, based on a predefined *residual lifetime*, and dynamically adjusts the priority frame by frame. The residual lifetime represents the number of frames remaining to finish the transmission for the head-of-line (HOL) packet of the service traffic; otherwise, the HOL packet will be discarded. The SS with a smaller residual lifetime should have a higher priority, and the RAS will choose the SS with the highest priority to serve first.

Denote $F_{k,s}$ as the residual lifetime for service type s of SS k, $s \in \{\text{UGS, rtPS, nrtPS, BE}\}$. $F_{k,\text{UGS}}$ is initially set to its packet-delay requirement, which is denoted by $D^*_{k,\text{UGS}}$, and is decreased one unit frame by frame. $F_{k,\text{rtPS}}$ can be obtained by

$$F_{k,\text{rtPS}} = D^*_{\text{rtPS}} - \tilde{F}_{k,\text{rtPS}} \tag{7}$$

where D^*_{rtPS} is the packet-delay requirement of the rtPS traffic of SS k, and $\tilde{F}_{k,\text{rtPS}}$ is the number of frames experienced by the rtPS packet of SS k at the current scheduled frame.

For $F_{k,\text{nrtPS}}$, it can be derived from the nrtPS's requirement of a minimum transmission rate of SS k, which is denoted by $R^*_{k,\min}$, which would be a long-term average. Denote the number of bits remaining unserved for the nrtPS HOL packet for service class s of user k by $\text{HOL}_{k,s}$, $s \in \{\text{UGS, rtPS, nrtPS, BE}\}$, the number of bits previously transmitted for the nrtPS packet of user k by $T_{k,\text{nrtPS}}$, and the number of previously active frames for the nrtPS transmission of user k by $\tau_{k,\text{nrtPS}}$. Since the system should fulfill the minimum rate requirement for the nrtPS, the maximum number of frames that are allowed for the HOL packet to transmit, which is denoted by D^*_{nrtPS}, should make the long-term average transmission rate larger than $R^*_{k,\min}$. Therefore, D^*_{nrtPS} can be calculated by

$$D^*_{\text{nrtPS}} = \left\lfloor \frac{T_{k,\text{nrtPS}} + \text{HOL}_{k,\text{nrtPS}}}{R^*_{k,\min}} \right\rfloor - \tau_{k,\text{nrtPS}}. \tag{8}$$

Then, $F_{k,\text{nrtPS}}$ can be obtained by

$$F_{k,\text{nrtPS}} = D^*_{\text{nrtPS}} - \tilde{F}_{k,\text{nrtPS}} \tag{9}$$

where $\tilde{F}_{k,\text{nrtPS}}$ is the number of frames that are experienced by the nrtPS HOL packet of user k. For F_{BE}, we do not ensure its QoS. However, a very low minimum-transmission rate can be set for the BE traffic to avoid starvation of its service packet transmission. Therefore, $D^*_{k,\text{BE}}$ and $F_{k,\text{BE}}$ can be similarly calculated by (8) and (9) as for $D^*_{k,\text{nrtPS}}$ and $F_{k,\text{nrtPS}}$, respectively.

The RAS determines the number of bits for service class s of user k, which is denoted by $R_{k,s}$, $s \in \{\text{UGS, rtPS, nrtPS, BE}\}, 1 \le k \le K$. $R_{k,s}$ is given by (10), shown at the bottom of the page, where $D^*_{k,s}$ denotes the packet delay requirement of SS_k with service type s. When an SS is selected, this RAS can help the scheduler give more resource to more urgent service to ensure the QoS. It divides $D^*_{k,s}$ into four levels, gives the average minimum rate to level 1 (L1) of each service, and doubles the rate when the service enters the next level. At level 4 (L4), the HOL packet for service type s of SS_k will be forced to be totally sent out to avoid QoS violation. This RAS saves the resource when the traffic is not urgent and gives it to urgent traffic by assigning more transmission bits. Notice that, to reduce the transmission overhead, the data of different services for the same SS should be aggregated and be consistently allocated.

Subsequently, in this paper, the proposed GGRA scheme intends to maximize the system throughput and fulfill QoS requirements for all SSs while being subject to system constraints such as the power, rank, and buffer constraints. Therefore, the resource-allocation problem is to decide an optimal user–subchannel pair and the corresponding modulation order. Denote $x^{\ell}_{k,n}$ as the number of bits that are carried by the modulation order for each subcarrier in subchannel n of SS_k in the ℓth OFDMA symbols $1 \le k \le K$, $1 \le n \le N$, and $1 \le \ell \le L$, which is given by

$$x^{\ell}_{k,n} = \begin{cases} 0, & \text{if not assigned} \\ 2, & \text{if QPSK modulation is assigned} \\ 4, & \text{if 16-QAM is assigned} \\ 6, & \text{if 64-QAM is assigned.} \end{cases} \tag{11}$$

The assignment vector in the ℓth OFDMA symbol for all users, which is denoted by $\mathbf{x}^{\ell}$, can be expressed by (12), shown at the bottom of the page. Then, the assignment matrix for the OFDMA UL frame, which is denoted by $\mathbf{x}$, is given by

$$\mathbf{x} = [\mathbf{x}^1, \mathbf{x}^{\ell}, \ldots, \mathbf{x}^L]. \tag{13}$$

$$R_{k,s} = \begin{cases} \text{HOL}_{k,s}/F_{k,s}\text{bits}, & \text{if } 0.75D^*_{k,s} \le F_{k,s} \le D^*_{k,s} \quad \text{(L1)} \\ \min\{2 \cdot \text{HOL}_{k,s}/F_{k,s}, \text{HOL}\}\text{bits}, & \text{if } 0.5D^*_{k,s} \le F_{k,s} \le 0.75D^*_{k,s} \quad \text{(L2)} \\ \min\{4 \cdot \text{HOL}_{k,s}/F_{k,s}, \text{HOL}\}\text{bits}, & \text{if } 0.25D^*_{k,s} \le F_{k,s} \le 0.5D^*_{k,s} \quad \text{(L3)} \\ \text{HOL}_{k,s}\text{bits}, & \text{if } D_{k,s} \le 0.25D^*_{k,s} \quad \text{(L4)} \end{cases} \tag{10}$$

$$\mathbf{x}^{\ell} \equiv \left[x^{\ell}_{1,1}, x^{\ell}_{1,2}, \ldots, x^{\ell}_{1,N}, \ldots, x^{\ell}_{k,1}, x^{\ell}_{k,2}, \ldots, x^{\ell}_{k,N}, \ldots, x^{\ell}_{K,1}, x^{\ell}_{K,2}, \ldots, x^{\ell}_{K,N}\right]^T \tag{12}$$

Since one subchannel is composed of q subcarriers, given $\mathbf{x}$, the total allocated bits to user k in the UL subframe, which is denoted by $R_k(\mathbf{x})$ or R_k, can be calculated by

$$R_k = \sum_{l=1}^{L} \sum_{n=1}^{N} q \cdot x_{k,n}^{l}. \tag{14}$$

Finally, the allocation problem can be mathematically formulated as optimization equations given by the following.

Objective

$$\mathbf{x}^* = \arg\max_{\mathrm{x}} \sum_{k=1}^{K} R_k(\mathbf{x}). \tag{15}$$

Subject to the system constraints

$$\begin{aligned} &1)\ \text{power constraint}: \sum_{n=1}^{N} p_{k,n}^{\ell} \leq p_{k,\max}, 1 \\ &\qquad \leq k \leq K, 1 \leq \ell \leq L \\ &2)\ \text{rank constraint}: \sum_{k=1}^{K} \operatorname{sgn}\left(x_{k,n}^{\ell}\right) \leq M, 1 \\ &\qquad \leq n \leq N, 1 \leq \ell \leq L \\ &3)\ \text{buffer constraint}: R_k \leq B_k, 1 \leq k \leq K \end{aligned} \tag{16}$$

and the RAS

$$\begin{aligned} &1)\ \text{SS with the lowest } F_{k,s} \text{ is selected first} \\ &2)\ \text{when } \mathrm{SS_k} \text{ is selected, determine } R_{k,s} \text{ by } (11) \\ &3)\ \text{each SS should obey consistent allocation} \end{aligned} \tag{17}$$

where $\mathbf{x}^*$ is the optimal assignment matrix, $p_{k,\max}$ is the maximum allowable power for SS k, B_k is the total number of bits in the buffers for all service types of SS k, and $R_k = \sum_s R_{k,s}$.

IV. Genetic Algorithm With Subscriber Station Grouping for Resource Allocation

The GGRA scheme is here devised to solve the optimization problem given in (15)–(17). The GGRA scheme first performs SS grouping by aggregating high-correlation SSs together as a group since the user correlation in a virtual MIMO system may lead to a large interference and severe performance degradation. The SS grouping is executed by choosing an SS_k and including all other SSs that have high correlation with the SS_k. Usually, a correlation threshold δ is given, and those SSs with a correlation value that is higher than δ are aggregated into the same group [14]. If there is SS that has a correlation value with other SSs higher than δ, then the SS will become one group by itself. The GGRA scheme will assign these SSs in the same group to different slots to prevent high mutual interference. Notice that since the system performance of the BER is a function of the correlation amplitude, this correlation threshold δ can be determined such that the corresponding BER is just below the target BER, i.e., BER^*.

Assume that there are a total of G groups in the system. Then, the GGRA scheme adopts the optimization technique of the GA to find the solution, where the G groups will be the G genes of a chromosome (an individual) for the GA, and the initial chromosome is denoted by C_0. Therefore, the GGRA scheme not only avoids interference among high-correlation users but also reduces the number of genes in GA chromosomes.

Based on C_0, I chromosomes are firstly formed to be the population of the first generation by random permutation of the genes (groups). Then, the GA uses a two-point crossover method to generate the population of the next generation. It randomly chooses two chromosomes and performs gene exchanges between the two chromosomes. Afterward, those duplicated genes in one chromosome are randomly permutated and put back to replace the duplicated genes of the other chromosome. There are $2I$ chromosomes in total with a different order of genes generated for each generation. The ith chromosome, which is denoted by C_i, $1 \leq i \leq 2I$ consists of G number of SS groups and is expressed as

$$C_i = [\Phi_{i,1}, \ldots, \Phi_{i,g}, \ldots, \Phi_{i,G}] \tag{18}$$

where $\Phi_{i,g}$ is the gth chromosome in the ith individual, which is the basic element in each individual. Assume that $\Phi_{i,g}$ contains κ SSs and is expressed as

$$\Phi_{i,g} = \{SS_{g,1}, SS_{g,2}, \ldots, SS_{g,\kappa}]. \tag{19}$$

Notice that different individuals represent various scheduling orders of groups. The GA performs the allocation based on the RAS for each C_i, $1 \leq i \leq 2I$. It chooses the first group and allocates resource to the SS with the smallest residual lifetime from the first slot of the UL subframe. After choosing the SS, for example, SS k, its available modulation order can be obtained by the power allocation given in (5) and (6) under the power constraint. Then, the RAS is applied to determine the allocated data bits to SS k. Finally, the required number of slots for SS k can be calculated by $\lceil R_k/q \cdot x_{k,n}^{l} \rceil$. If there are remaining bits in the slots, they will be given to each service class of SS k proportionally to their service queue length. When finishing the allocation for this SS, the GA will continue finding the next SS in the same group by the residual lifetime. If every SS in this group is scheduled, then the GA will allocate the resource to SSs in the next group until all SSs are scheduled or no remaining resource is left.

If there are still free slots available when all SSs are allocated, the GGRA performs RRR for utilization maximization. In each available free slot, the GGRA will check the power and buffer constraints of each SS and the rank constraint for the system. Then, it gives the slot to the SS that supports the highest modulation order among all SSs under the system constraints. Each selected SS can transmit its four kinds of service in the slot. The RRR will iteratively choose the SS until there are no remaining slots or packets in the queue. Consequently, a UL allocation map for each individual is determined. Notice that one individual gets one allocation result. After finishing the allocation, the GA should evaluate the fitness of each individual related with one allocation result. Since we want to maximize the throughput with QoS guarantee, a predefined

TABLE I
SYSTEM-LEVEL PARAMETERS

Parameters	Value
Cell size	1.6 km
Number of Antenna in BS	2
Frame duration	5 ms
System bandwidth	5MHz
FFT size	512
Subcarrier frequency spacing	10.9375 KHz
Number of data subcarriers	384
Number of subchannels	8
Number of data subcarriers per subchannel	48
OFDMA symbol duration	102.86 μs
Number of slots for uplink transmission per frame	16
Maximum transmission power for each SS	23 dBm
Thermal noise density	-174 dBm/Hz
Shadowing model distribution	Log-normal
Fast fading model distribution	Rayleigh

TABLE II
QoS REQUIREMENTS OF EACH SERVICE TYPE

Traffic Type	Requirement	Value
Voice (UGS)	Required BER	10^{-3}
	Max. Delay Tolerance	20 ms
	Max. Allowable packet dropping rate	1%
Video (rtPS)	Required BER	10^{-4}
	Max. Delay Tolerance	100 ms
	Max. Allowable packet dropping rate	1%
HTTP (nrtPS)	Required BER	10^{-6}
	Min. Required Transmission Rate	128 Kbps
FTP (BE)	Required BER	10^{-6}

fitness function, which is denoted by Π, is defined to evaluate the individuals, which is given by

$$\Pi\left(\sum_{k} R_k, \nu\right) = \frac{\sum\limits_{k} R_k}{q \cdot W} \cdot e^{-\alpha \cdot \nu} \qquad (20)$$

where α is a constant to control the exponential decay, ν is the unsatisfied ratio, which is the number of bits for UGS, rtPS, and nrtPS unsatisfied service traffic divided by the total number of bits of the original HOL packet, and W represents the total number of used slots in the current scheduled frame. Π contains two terms: The first term gives a value about *spectrum efficiency*, and the second term gives a value that is related with the *unsatisfied ratio*. As a result, I individuals with the highest fitness value will be chosen to become the offspring. If the maximum iteration number is reached, or the optimal solution is found, which leads to 64-QAM for each slot and zero unsatisfied ratio, then the GGRA will terminate and choose the best individual with the highest fitness value. The corresponding allocation result will be adopted for UL resource allocation. Otherwise, the GGRA will go back to the crossover and mutation and produce the new individuals for allocation in the next iteration.

V. SIMULATION RESULTS AND DISCUSSION

We perform the simulations by using C++ language in the workstation. In the simulations, the system-level parameters of the UL IEEE 802.16 system are set to be compatible with the IEEE 802.16m Evaluation Methodology Document [15]. Both the large- and small-scale fadings are considered, and the path loss is modeled as $128.1 + 37.6 \log X$ (dB), where X is the distance between the BS and the SS in units of kilometers. The power delay profile follows the exponential decay rule. Finally, the channel state is assumed to be fixed within one frame and varied frame by frame with time according to the fading model. Table I shows these system-level parameters.

Four types of traffic corresponding to the UGS, the rtPS, the nrtPS, and the BE are considered in the simulations [15]. The first one is the voice traffic for the UGS, which is modeled as the ON–OFF model, with ON (OFF) period exponentially distributed with a mean value of 1 s (1.35 s). During the ON period, one packet is generated every 20 ms with a fixed size of 28 B, with both a payload and a header included. The second is the video-streaming traffic for the rtPS. It consists of a sequence of video frames regularly generated every 100 ms. Each video frame is composed of eight slices, and each slice is related to a single packet. The size of each slice packet is truncated Pareto distributed, and the interarrival time of two consecutive packets is also truncated Pareto distribution. The third service type is the Web-browsing HTTP traffic for the nrtPS. It is modeled as a sequence of Web-page downloads, and each page is composed of several packet arrivals. It can be separated as a main object and some embedded objects that can be divided into several packets, and the maximum transmission unit is 1500 B, with the header included. The interarrival time between two pages is exponentially distributed with mean time of 30 s. The main and embedded objects are both truncated lognormal-distributed with mean of 10 710 and 7758 B, respectively. The last type is the FTP traffic for the BE. It is modeled as a sequence of file downloads. The size of each file is truncated lognormal-distributed with mean of 2 MB, standard deviation of 0.722 MB, and a maximum value of 5 MB.

The interarrival time of two consecutive files is exponentially distributed with mean of 180 s. We can calculate the arrival rates of the UGS, the rtPS, the nrtPS, and the BE by the traffic models mentioned above, which are 4.8, 64, 14.5, and 88.9 kb/s, respectively. The QoS requirements of each service type are defined in Table II. Furthermore, the minimum transmission rate of the BE traffic is set as 1 B per frame for calculating F_{BE} in the GGRA scheme.

The number of SSs varies from 10 to 80, which are uniformly distributed over the cell. Each of the SS contains all the four types of services and has a speed of 60 km/h. In this paper, the maximum system transmission rate in a frame, which is equal to 14.7456 Mb/s, is achieved when the highest modulation order is assigned to the M SSs with low correlation in each slot according to the rank constraint. The traffic intensity is defined as the ratio of the total average arrival rate of four service types of all users over the maximum system transmission rate. We compare the proposed GGRA scheme with the EFS [9] and MLWDF algorithms [10], which use a heuristic approach and define the service priority statically and dynamically, respectively.

Fig. 1 shows the system throughput of the three resource-allocation schemes. It can be seen that the three schemes have similar throughput in low-medium traffic-intensity cases because the system can sufficiently support the four types of services in these cases and easily fulfill the QoS requirements of the UGS, the rtPS, and the nrtPS. However, in the

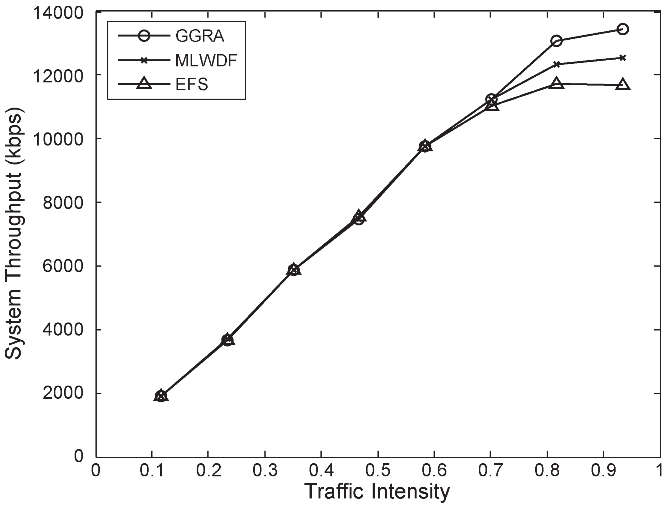

Fig. 1. System throughput.

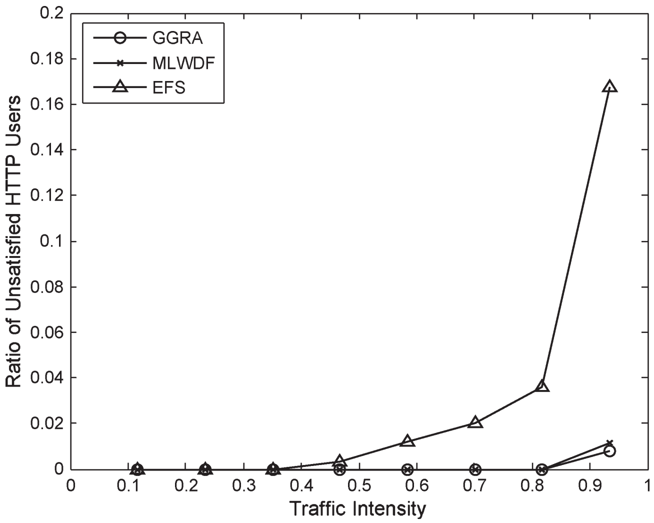

Fig. 2. Ratio of unsatisfied HTTP users.

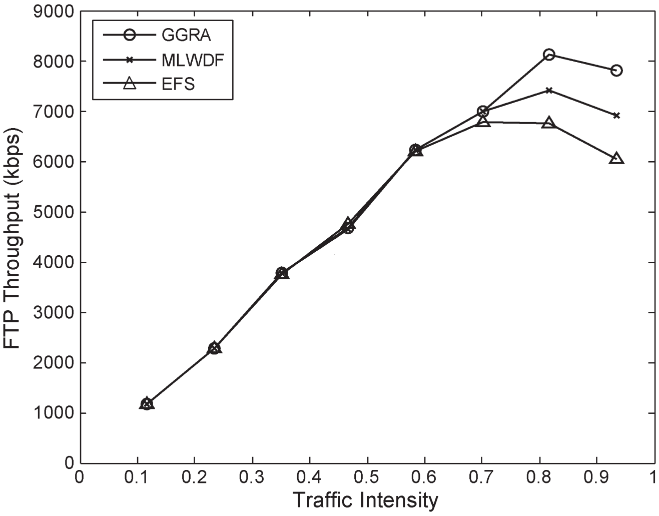

Fig. 3. FTP throughput.

high-traffic-intensity case, the GGRA scheme has the highest throughput among the three allocation schemes. It is because the GGRA adopts the residual lifetime of the service traffic to dynamically allocate the resource, and thus, their QoS requirements can be satisfied. Also, the GGRA uses the GA, instead of heuristic algorithms, to find the optimal allocation map, and thus, the spectrum efficiency can be maximized. The GGRA scheme can allocate the service traffic more precisely and efficiently. On the other hand, the MLWDF can dynamically adjust the traffic priority as the GGRA by a predefined function. However, it considers one allocation slot a time rather than the whole frame, as the GGRA does. Thus, such a local optimal solution done by the MLWDF cannot achieve higher throughput than the global optimal solution done by the GGRA. On the other hand, the EFS serves the highest priority traffic first. Therefore, it may allocate slots to the SS that have a higher priority service but cannot support a high modulation order. Moreover, the EFS considers one slot a time, as the MLWDF does. These significant weak points cause the EFS scheme the throughput degradation in the high-traffic-intensity case.

Notice that all the three allocation schemes can achieve almost zero voice and voice packet dropping rates for all cases. It is because the proposed GGRA scheme grants resource according to the RAS based on the predefined residual lifetime. The RAS can efficiently grant resource to the voice and the video traffic because the traffic of these services has small delay tolerances and small packet sizes. Also, the GA is applied to make efficient allocation. Thus, the urgent voice and video packets can be sent out before reaching the delay bound. On the other hand, the MLWDF scheme dynamically considers the channel gain and the weighted delay. Thus, it can take care of the UGS, the rtPS, and the nrtPS traffic in a fair way. Therefore, the MLWDF can achieve very low voice and video dropping rates as the GGRA. Also, the EFS scheme always serves the voice packet prior to other service traffic. Thus, it can achieve a zero voice packet dropping rate for all traffic cases. Furthermore, the video packet size is small enough and still has higher priority than the HTTP and FTP packets; therefore, it can also be sent out with almost zero dropping rate.

Fig. 2 shows the ratio of unsatisfied HTTP users. It can be seen that the GGRA can maintain zero unsatisfied users under the traffic intensity up to 0.8. This is due to the fact that the GGRA considers a long-term average minimum rate requirement for each HTTP user, and the RAS can give more resources to HTTP users when the long-term rate is below the minimum rate requirement. Therefore, if one packet is considered unsatisfied, the GGRA scheme will give a lower residual lifetime to the next HTTP packet of this user and force the RAS to send it out more quickly. Finally, the overall HTTP rate will be compensated and kept above the minimum rate requirement. The MLWDF scheme fairly considers HTTP as the GGRA scheme; therefore, it can keep a low unsatisfied ratio as well. However, the EFS scheme serves HTTP packets at a lower priority than voice and video packets. Thus, the average rate of HTTP users will severely decrease in the high-traffic case and cause a high ratio of unsatisfied HTTP users.

Fig. 3 shows the FTP throughputs of the three allocation schemes. The arrival rate of the FTP traffic is much higher than

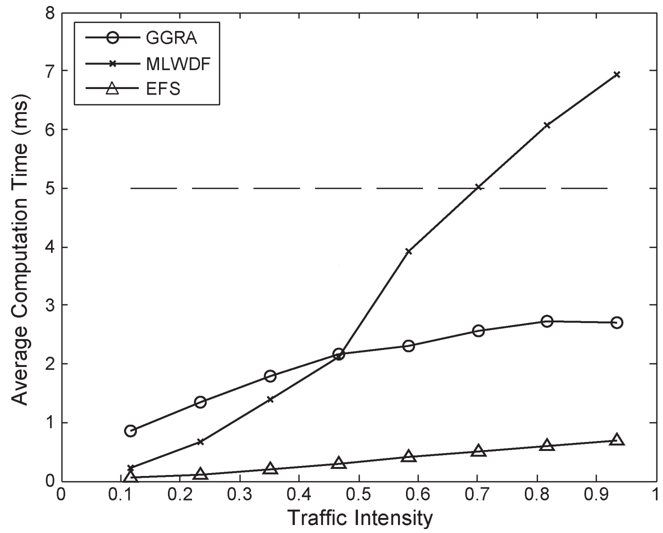

Fig. 4. Average computation time.

the other service traffic, and the FTP throughput degradation will severely affect the system throughput. It can be seen that the FTP throughput of the GGRA increases until traffic intensity larger than 0.8. This phenomenon is due to the cooperation of the GA and the RAS, which save more resources for the FTP traffic. The MLWDF considers a local optimal allocation and dynamically assigns resources to each service. Furthermore, if the high-priority service is forbidden to be sent out due to power or interference issues, the FTP packets can still be allocated in the current frame. This relaxation raises the FTP throughput of the MLWDF scheme. In the EFS scheme, it sends out a high-priority packet as soon as possible rather than just fulfilling the QoS, which may cause some high-priority traffic to use a low-modulation order. It not only reduces the spectrum efficiency but also suppresses the transmission opportunity of the FTP traffic, which further degrades the performance of the FTP service. As a result, the EFS scheme has the worst FTP throughput among the three schemes.

Finally, Fig. 4 shows the computation time of the three resource-allocation schemes, where the dotted line denotes the 5-ms frame duration. The GGRA needs the largest computation time in the low-traffic case since it needs to do crossover and mutation to generate new individuals, and each individual has a new fitness value that is calculated by the new scheduling permutation. However, each individual in the GA is linked to one allocation order, and these orders can be computed in parallel with no interaction. Therefore, as the traffic intensity increases, the average computation time by the GGRA can still be maintained without varying too much and is kept less than 3 ms, which is lower than the frame duration. The GGRA is feasible for real-time applications. The EFS uses a heuristic approach to assign resources to the user with the highest priority first. Therefore, the allocation can be done in a short computation time for the EFS under all traffic-intensity cases. On the other hand, the MLWDF also adopts the heuristic approach to allocate resources to the user with the largest MLWDF function value. However, the computation of the MLWDF increases tremendously because of the relaxation of the FTP packet transmission opportunity, where the relaxation needs to check the FTP queues of each SS after sorting the QoS guarantee traffic, which takes more computation time.

VI. Conclusion

In this paper, a GGRA scheme has been proposed for the IEEE 802.16 UL system. With the design of the SS grouping, the GGRA scheme makes the adoption of the GA optimization technique computationally tractable and, thus, feasible in real-time applications. The GGRA scheme can maximize the system throughput under the system constraints and QoS requirements, where an RAS equipped with a predefined residual lifetime is originally designed. Simulation results show that the GGRA scheme performs better than the MLWDF and EFS schemes. The GGRA scheme can achieve high system throughput and fulfill QoS requirements for all types of traffic.

References

[1] IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems, IEEE Std. 802.16-2004, Oct. 2004.

[2] IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands, IEEE Std. 802.16e, Oct. 2005.

[3] D. Niyato and E. Hossain, "Queue-aware uplink bandwidth allocation for polling services in 802.16 broadband wireless networks," in *Proc. IEEE GLOBECOM*, 2005, pp. 3072–3076.

[4] J. Sun, Y. Yao, and H. Zhu, "Quality of service scheduling for 802.16 broadband wireless access systems," in *Proc. IEEE VTC-Spring*, 2006, vol. 3, pp. 1221–1225.

[5] Y. Ben-Shimol, I. Kitroser, and Y. Dinitz, "Two dimensional mapping for wireless OFDMA systems," *IEEE Trans. Broadcast.*, vol. 52, no. 3, pp. 388–396, Sep. 2006.

[6] C. M. Yen, C. J. Chang, F. C. Ren, and J. A. Lai, "Dynamic priority resource allocation for uplinks in IEEE 802.16 wireless communication systems," *IEEE Trans. Veh. Technol.*, vol. 58, no. 8, pp. 4587–4597, Oct. 2009.

[7] C. F. Tsai, C. J. Chang, F. C. Ren, and C. M. Yen, "Adaptive radio resource allocation for downlink OFDMA/SDMA systems with multimedia traffic," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1734–1743, May 2008.

[8] J. Wang, K. Araki, Z. Zhang, Y. Chang, H. Zhu, and T. Kashima, "A low complexity tree-structure based user scheduling algorithm for up-link multi-user MIMO systems," *IEICE Trans. Commun.*, vol. E90-B, no. 6, pp. 1415–1423, Jun. 2007.

[9] V. Singh and V. Sharma, "Efficient and fair scheduling of uplink and downlink in IEEE 802.16 OFDMA networks," in *Proc. IEEE WCNC*, 2006, vol. 2, pp. 984–990.

[10] W. Park, S. Cho, and S. Bahk, "Scheduler design for multiple traffic classes in OFDMA networks," in *Proc. IEEE ICC*, 2006, vol. 2, pp. 790–795.

[11] Y. B. Reddy, N. Gajendar, P. Taylor, and D. Madden, "Computationally efficient resource allocation in OFDM systems: Genetic algorithm approach," in *Proc. ITNG*, 2007, pp. 36–41.

[12] Q. Guo, S. C. Kim, and D. C. Park, "Antenna selection using genetic algorithms," *IEICE Trans. Fundam.*, vol. E89-A, no. 6, pp. 1773–1775, Jun. 2006.

[13] A. J. Goldsmith and S. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218–1230, Oct. 1997.

[14] Y. J. Zhang and K. B. Letaief, "An efficient resource-allocation scheme for spatial multiuser access in MIMO/OFDM systems," *IEEE Trans. Commun.*, vol. 53, no. 1, pp. 107–116, Jan. 2005.

[15] IEEE 802.16m Evaluation Methodology Document (EMD), IEEE Std. 802.16m-08/004r1, Mar. 17, 2008.

Yin Chiu was born in Taichung, Taiwan. He received the B.E. and M.E. degrees from the National Chiao Tung University, Hsinchu, Taiwan, in 2006 and 2008, respectively.

His research interests include radio resource management, scheduling, and wireless communication systems.

Chung-Ju Chang (F'06) was born in Taiwan in August 1950. He received the B.E. and M.E. degrees in electronics engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 1972 and 1976, respectively, and the Ph.D. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, in 1985.

From 1976 to 1988, he was with the Telecommunication Laboratories, Directorate General of Telecommunications, Ministry of Communications, Taiwan, as a Design Engineer, Supervisor, Project Manager, and then Division Director. From 1987 to 1989, he was a Science and Technical Advisor to the Minister of the Ministry of Communications. In 1988, he joined the Faculty of the Department of Electrical Engineering, College of Electrical and Computer Engineering, National Chiao Tung University, as an Associate Professor, where, since 1993, he has been a Professor. From August 1993 to July 1995, he was the Director of the Institute of Communication Engineering, from August 1999 to July 2001 the Chairman of the Department of Communication Engineering, and, from August 2002 to July 2004, the Dean of the Research and Development Office. During 1995–1999, he was also an Advisor to the Ministry of Education to promote education in communication science and technologies for colleges and universities in Taiwan. His research interests include performance evaluation, radio resource management for wireless communication networks, and traffic control for broadband networks.

Dr. Chang is a member of the Chinese Institute of Engineers. He is currently a Committee Member of the Telecommunication Deliberate Body, Taiwan. He also serves as an Editor for the *IEEE Communications Magazine* and as an Associate Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY.

Kai-Ten Feng (M'03) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, in 1992, the M.S. degree from the University of Michigan, Ann Arbor, in 1996, and the Ph.D. degree from the University of California, Berkeley, in 2000.

Between 2000 and 2003, he was with OnStar Corporation, a subsidiary of General Motors Corporation, as an In-Vehicle Development Manager/Senior Technologist, working on the design of future telematics platforms and in-vehicle networks. Between February 2003 and July 2007, he was an Assistant Professor with the Department of Electrical Engineering, National Chiao Tung University, Hsinchu, Taiwan, where, since August 2007, he has been an Associate Professor. His current research interests include cooperative and cognitive networks, mobile *ad hoc* and sensor networks, embedded system design, wireless location technologies, and intelligent transportation systems.

Dr. Feng was a recipient of the Best Paper Award from the Spring 2006 IEEE Vehicular Technology Conference, which ranked his paper first among the 615 accepted papers. He was also the recipient of the Outstanding Young Electrical Engineer Award in 2007 from the Chinese Institute of Electrical Engineering. He has served on the technical program committees of the IEEE Vehicular Technology Conference, the IEEE International Conference on Communications, and the Asia-Pacific Conference on Wearable Computing Systems.

Fang-Chin Ren (M'03) was born in Hsinchu, Taiwan. He received the B.E., M.E., and Ph.D. degrees in communication engineering from the National Chiao Tung University, Hsinchu, in 1992, 1994, and 2001, respectively.

Since 2001, he has been a Protocol Design Engineer with the Industrial Technology Research Institute, Hsinchu, where he is involved in the design and development of wireless code-division multiple access chipsets, WiMAX mobile multihop relay technology, and fourth-generation access technology. His current research interests include system performance analysis, protocol design, and mobile radio networks.