Opportunistic Scheduling with Economized CSI Feedback for OFDMA/TDD Downlink Systems

Yao-Hsing Chung and Chung-Ju Chang
Department of Communication Engineering
National Chiao Tung University
Hsinchu 300, Taiwan
E-mail: cjchang@mail.nctu.edu.tw

***Abstract*— In this paper we propose an *economized-CSI (channel state information) opportunistic scheduling* (ECOS) scheme for OFDMA/TDD downlink systems. The ECOS scheme consists of a *quality-of-service (QoS) guarantee scheduling* (QGS) algorithm and a *CSI overhead reduction* (COR) algorithm. The QGS algorithm guarantees QoS requirements by adjusting *priority value* and *anticipated rate* of users dynamically. In addition, it achieves balance between QoS guarantee and throughput maximization by employing a priority threshold. The COR algorithm performs *admissible user selection* to economize the amount of CSI overhead. Since all users do not need to feed back CSI every frame, the transmission power used for CSI reporting is economized. Besides, the COR algorithm can reduce the computation complexity of the QGS algorithm because the number of users to be scheduled is reduced. Simulation results show that the proposed ECOS scheme greatly reduces the uplink bandwidth occupancy of CSI feedback. Also, it achieves high system throughput and maintains QoS guarantee at high traffic load.**

I. Introduction

Orthogonal frequency division multiplexing (OFDM) system exploiting multiuser diversity (named multiuser-OFDM or OFDMA system) requires *channel state information* (CSI) of all subchannels of each user (full CSI) to maximize the system throughput or minimize the transmission power. In some situations, it may be not practical that all users feed unquantized CSI of all subchannels back to the base station (BS), especially when the bandwidth for CSI feedback is limited. Therefore, several researches focus on reducing CSI overhead in wireless communication systems.

Ideas of best-n feedback and opportunistic feedback were proposed to mitigate the CSI overhead problem [1]-[5]. In the best-n feedback, only the information of subchannel with top n channel conditions were reported [1]. On the other hand, the opportunistic feedback costs only one bit representing the state of a subchannel [2]-[4]. If the channel gain of a subchannel of a user exceeds the prespecified threshold, the user will feed back with '1' [2]-[4]. An opportunistic feedback scheme with multiple SNR threshold was proposed in [5] to improve the success probability of feedback when the feedback channels are contention-based. The derivations of the SNR threshold of the opportunistic feedback provide excellent CSI overhead reduction with high system throughput, but they should have involved *quality-of-service* (QoS) consideration to support classes of multimedia service. Also, only users who successfully fed backcan be allocated resource, which means it is hard to guarantee QoS requirements of users. In addition, since opportunistic feedback is a contention-based scheme, it needs extra bits for identifying feedback users.

We propose an *economized CSI opportunistic scheduling* (ECOS) scheme in this paper. The ECOS scheme consists of a *quality-of-service guarantee scheduling* (QGS) algorithm and a *CSI overhead reduction* (COR) algorithm. The QGS algorithm adjusts priorities of users dynamically and achieves balance between QoS guarantee and throughput maximization. The COR algorithm reduces the CSI overhead by proper selection of *admissible users* according to the degree of QoS fulfillment of each user. Besides, the COR algorithm reduces the computation complexity of the QGS algorithm because the number of users to be scheduled is reduced. By the design of *admissible bit map*, the BS can recognize which reported CSI is belonged to which *admissible user* without additional identifying bits. Simulation results show that the proposed ECOS scheme can greatly reduce the uplink bandwidth occupancy of CSI feedback. Also, it can still achieve high system throughput and keep QoS guaranteed.

II. System Model

An OFDMA/TDD system is considered, such as the one defined in IEEE 802.16 standard [6]. There are total K active mobile users and N subchannels, each with n_s adjacent subcarriers in a cell. The time axis is divided into frames, and each frame consists of a downlink subframe followed by a uplink subframe. A downlink (uplink) subframe has L_D (L_U) OFDM symbols. One subchannel in frequency domain and one OFDM symbol in time domain form a *tile*, which is a basic unit for resource allocation. The system supports adaptive modulation orders of M-QAM, where $M \in \{4, 16, 64\}$. The CSI of users is modulated by QPSK to ensure its reliability and is transmitted to BS through feedback channels, which are dedicated from a portion of uplink bandwidth [6]. In addition, we assume that the BS has the knowledge of the distance between it and each user.

The system is assumed to support classes of service with a variety of traffic: (i) real-time (RT) service, (ii) non-real-time (NRT) service, and (iii) best-effort (BE) service. There are two kinds of traffic for RT service, voice and video. The HTTP (hypertext-transport-protocol) and FTP (file-transfer-protocol) traffics are belonged to NRT and BE services, respectively. QoS requirement for the user with RT (NRT) services is the maximum delay tolerance (minimum required transmission rate). Besides, the maximum required bit-error-rate (BER) of user k, denoted by BER_k^*, is given according to the traffic type belonged to the user as well.

In a real wireless communication system, the CSI is represented by a quantized value rather than a real value [6]. Herein, we

978-1-4244-5213-4/09/ $26.00 ©2009 IEEE

assume that the CSI is simply quantized to indicate the supportable maximum modulation order of a subchannel. Assume that the equal power allocation is adopted for all subchannels. Since a minimum required SNR of applying M-QAM modulation order while satisfying BER_k^* has been given in [7], the CSI of subchannel n reported by user k at frame t, denoted by $\gamma_{k,n}(t)$, is represented by quantizing real SNR value into 0, 1, 2, or 3 to indicate no-transmission, QPSK, 16-QAM, and 64-QAM, respectively. Therefore, the feedback-rate of a user is $2N$ bits/frame (bpf) for N subchannels.

III. Economized-CSI Opportunistic Scheduling Scheme

The proposed *economized-CSI opportunistic scheduling* (ECOS) scheme consists of a *QoS guaranteed scheduling* (QGS) algorithm and a *CSI overhead reduction* (COR) algorithm. As shown in Fig. 1, at the beginning of downlink subframe t, *admissible users* for CSI report in $\Omega(t)$ will be informed, where $\Omega(t)$ is the set of *admissible users* who were selected to report CSI at uplink subframe t. After CSI feedbacks of $\Omega(t)$ are received at uplink subframe t, the QGS algorithm is performed to allocate resource to users in $\Omega(t)$ for downlink transmission at frame $(t+1)$. The COR algorithm determines $\Omega(t+1)$ afterward. The QGS algorithm and the COR algorithm are introduced as follows.

A. QoS Guaranteed Scheduling Algorithm

With a goal to maintain QoS, the QGS algorithm performed at frame t allocates resource for frame $(t+1)$ according to $Q_k(t)$ and $\widehat{R}_k(t)$, where $Q_k(t)$ and $\widehat{R}_k(t)$ are the *priority value* and the *anticipated rate* of the user k at frame t), respectively.

The *priority value* of user k at frame t, denoted by $Q_k(t)$, is defined as how much the user k starves for the resource in order to fulfill the QoS requirements. If the number of residual bits of the HoL packet of user k, denoted by $B_k(t)$, is zero, then $Q_k(t)$ is set to be zero; otherwise, $Q_k(t)$ is defined as

$$Q_k(t) = \begin{cases} q_{RT} \cdot e^{(\frac{D_k(t)+1}{D_k^*})}, & \text{if } k \text{ is a RT user,} \\ q_{NRT} \cdot e^{(\frac{\alpha[R_k^* - \bar{R}_k(t)]}{R_k^*})}, & \text{if } k \text{ is a NRT user,} \\ q_{BE} \cdot e^{-d_k(t)}, & \text{elsewise,} \end{cases} \quad (1)$$

where q_{RT}, q_{NRT}, and q_{BE} are the basic priorities for RT, NRT, and BE services, respectively; $D_k(t)$ is the delay that the head-of-line (HoL) packet of user k has experienced till frame t; $d_k(t)$ is the normalized (by the radius of the cell) distance between user k and BS at frame t, $0 \le d_k(t) < 1$; D_k^* and R_k^* are maximum delay tolerance and minimum required transmission rate of user k, respectively; α is a designed parameter which adjusts the decadent rate of $Q_k(t)$ of NRT users; and $\bar{R}_k(t)$ is the average transmission rate of user k till frame t. It is clear that $q_{RT} > q_{NRT} > q_{BE}$. The $\bar{R}_k(t)$ is given by

$$\bar{R}_k(t) = \frac{\sum_{w=1}^{W} R_k(t-w)}{\sum_{w=1}^{W} F_k(t-w)}, \quad (2)$$

where W is the observing window size, $R_k(t)$ is number of allocated bits of user k, and $F_k(t)$ is buffer-state indicator which is set to 1 if user k has data in buffer in frame t, otherwise is set to 0. The $Q_k(t)$ of users with BE service is proportional to $d_k(t)$ by the fact that the user who is closer to BS should have higher probability to get service because of his/her better channel condition. Besides, the $Q_k(t)$ of BE users is varied frame by frame because it is assumed that each user is moving in the cell, which implies that each BE user has chance to be scheduled.

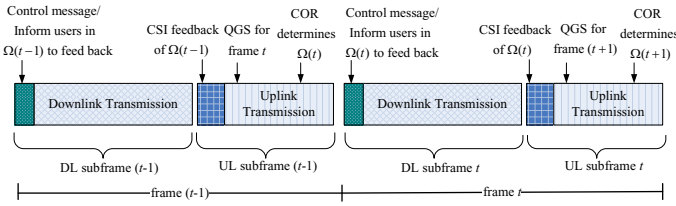

Fig. 1. The frame structure of the OFDMA/TDD system

The *anticipated rate* of user k at frame t, denoted by $\widehat{R}_k(t)$, is the number of bits that is expected to allocate to user k at frame t in order to avoid QoS violation if user k is scheduled at frame t. The $\widehat{R}_k(t)$ is defined as

$$\widehat{R}_k(t) = \begin{cases} \left\lceil \{\frac{D_k(t)}{D_k^*} U(x_1) + [1 - U(x_1)]\} \frac{B_k(t)}{b} \right\rceil \cdot b, & \text{if } k \text{ is RT user,} \\ \left\lceil \{F_k^T(t)[R_k^* - \bar{R}_k(t)] U(x_2) + \beta R_k^* [1 - U(x_2)]\} \frac{U(B_k(t))}{b} \right\rceil \cdot b, & \text{if } k \text{ is NRT user,} \\ \left\lceil e^{-d_k(t)} \cdot \frac{B_k(t)}{b} \right\rceil \cdot b, & \text{if } k \text{ is BE user,} \end{cases} \quad (3)$$

where b is the minimum number of bits of a tile for resource allocation, $b = 2n_s$; $F_k^T(t)$ denotes $\sum_{w=0}^{W} F_k(t-w)$; β is a designed parameter which adjusts the amount of $\widehat{R}_k(t)$ of NRT users; the $\lceil m \rceil$ is the smallest integer larger than m; the $U(x) = 1$ if $x > 0$, otherwise the $U(x) = 0$; and x_1 and x_2 are $(\lceil \frac{D_k^*}{2} \rceil - D_k(t))$ and $(R_k^* - \bar{R}_k(t))$, respectively. The amount of the $\widehat{R}_k(t)$ for RT users is adjusted by $U(x_1)$ according to the HoL packet delay. When the HoL packet delay of a RT user is less than x_1, $\widehat{R}_k(t)$ is set to be partial $B_k(t)$; otherwise, $\widehat{R}_k(t)$ is set to be entire $B_k(t)$. Likewise, the amount of the $\widehat{R}_k(t)$ for NRT users is regulated by $U(x_2)$ according to the minimum transmission rate satisfaction.

However, in order to enhance the system throughput, whenever the highest *priority* of the users is lower than a priority threshold, denoted by Q_{th}, the QGS algorithm just sets aside these high priority users but allocates resources according to the channel state of the users. The flow chart of QGS algorithm is shown in Fig. 2, where $\mathcal{K}$ is the set of users to be scheduled, $\mathcal{R}$ is the set of users whose *anticipated rate* is not satisfied for $\mathcal{R} \subseteq \mathcal{K}$, $Q_{\max}$ denotes the highest *priority value* of the users in $\mathcal{R}$, and $\mathcal{N}$ is the set of subchannels which have free symbols to allocate. The QGS algorithm mainly contains four functions: *initialization*, *priority-based user-subchannel selection*, *SNR-based user-subchannel selection*, and *adjacent symbol allocation*, which are described below.

Initialization: The QGS algorithm is initialized such that $\mathcal{K} = \mathcal{R} = \Omega(t-1)$, $\mathcal{N} = \{1, \cdots, N\}$, $\mathcal{S}_n = \{1, \cdots, L_D\}$, and $r_k = 0$, $1 \le k \le K$. Note that the QGS algorithm allocates resource of frame t to $\Omega(t-1)$ rather than all users. $\mathcal{S}_n$ is the set of free symbols in subchannel n, and r_k is the accumulated number of allocated bits to user k during the scheduling. In addition, $Q_k(t)$ and $\widehat{R}_k(t)$ of all users have been calculated and are known by

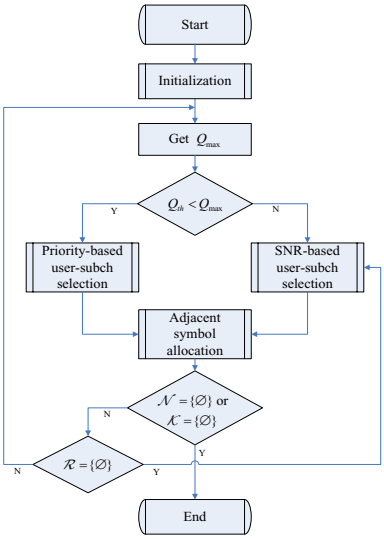

Fig. 2. Flow chart of the QGS algorithm

the QGS algorithm.

Priority-based user-subchannel selection: This function is to select users with highest priority and allocate resource. Assume user k is selected. If user k has the best CSI $\gamma_{k,n}$ in subchannel n, then the user k is allocated with the subchannel n.

SNR-based user-subchannel selection: This function is to select users with best CSI and allocate resource. If there are several users whose CSI of the subchannel n is the best, the subchannel n is allocated to the one who has the highest *priority value*.

Adjacent symbol allocation: Assume that the subchannel n is assigned to user k and $\min(\mathcal{S}_n) = \ell$. To fulfill QoS requirements of the user k, the i^{th} OFDM symbol of subchannel n, $\ell \leq i \leq L_D$, is assigned to user k until $\widehat{R}_k(t) \leq r_k$ is satisfied. Whenever the i^{th} symbol is assigned to user k, $\mathcal{S}_n = \mathcal{S}_n - \{i\}$, $r_k = r_k + 2b\gamma_{k,n}$, and $B_k = \max(B_k - 2b\gamma_{k,n}, 0)$. If the *anticipated rate* of user k is satisfied, says $\widehat{R}_k(t) \leq r_k$, the $\mathcal{R} = \mathcal{R} - \{k\}$. If user k's buffer becomes empty ($B_k = 0$), then $\mathcal{K} = \mathcal{K} - \{k\}$ and $\mathcal{R} = \mathcal{R} - \{k\}$. Also, if $\mathcal{S}^{(n)} = \{\emptyset\}$, then $\mathcal{N} = \mathcal{N} - \{n\}$.

After the QGS algorithm is done, it can be known that $R_k(t) = r_k$, $\forall k$, where $R_k(t)$ is the total number of allocated bits of user k at frame t.

B. CSI Overhead Reduced Algorithm

The COR algorithm, performed after the QGS algorithm, determines which users are admitted to feed back CSI at frame t. The *admissible users* allowed to feed back at frame t are the users who have higher priority to be scheduled by QGS algorithm performed at the downlink subframe $(t+1)$. From the perspective of QoS guarantee, the QGS algorithm performed at frame t allocates resource according to $Q_k(t)$ and $\widehat{R}_k(t)$, and the high priority user is assigned sufficient bits fulfilling the *anticipated rate* prior to low priority users. Therefore, we have (i) the opportunity of the high priority user k being scheduled by the QGS algorithm at frame t is proportional to $Q_k(t)$, and (ii) $\widehat{R}_k(t) \leq R_k(t)$ should be satisfied if user k is scheduled. Here $R_k(t) = \widehat{R}_k(t)$ is assumed. Let C_D denote the upper bound of downlink capacity while all the NL_D *tiles* are assigned with the highest modulation order 64-QAM; in other words, $C_D = 3bNL_D$ bits per frame (bpf). Before the *admissible user selection*, the COR algorithm calculates $Q_k(t)$ and $\widehat{R}_k(t)$ of all users. The *admissible user selection* is done by solving

$$\Omega(t) = \arg\max_{\Theta} \sum_{k\in\Theta} Q_k(t), \tag{4}$$

subject to

$$\sum_{k\in\Theta} \widehat{R}_k(t) \leq C_D, \quad \forall \widehat{R}_k(t) > 0, \tag{5}$$

where $\Omega(t)$ is the set of *admissible users* who need to feed back at frame t, Θ denotes a possible subset of $\{1, 2, \cdots, K\}$. Therefore, there are 2^K combinations of Θ, which grows exponentially when K increases.

In order to determine $\Omega(t)$ within extremely short time, the *admissible user selection* can be implemented by a heuristic greedy method. This method first sorts all users according to $Q_k(t)$. Denote the sorted user set by $\Lambda(t)$, which is given by

$$\Lambda(t) = \{\lambda_i,\ 1 \leq i \leq K |\ \ Q_{\lambda_i}(t) \leq Q_{\lambda_{i+1}}(t)\}, \tag{6}$$

where λ_i is an identifier of the i^{th} descending order user. After that, the number of *admissible users*, denoted by $K_a(t)$, can be obtained by

$$K_a(t) = \arg\max_{k}(0 \leq C_D - \sum_{i=1}^{k} \widehat{R}_{\lambda_i}(t)), \tag{7}$$

subject to

$$\widehat{R}_{\lambda_k}(t) > 0. \tag{8}$$

Finally, the $\Omega(t)$ is constructed by the front $K_a(t)$ ordered users in $\Lambda(t)$, where $\Omega(t) = \{\lambda_1, \lambda_2, \cdots, \lambda_{K_a(t)}\}$, and the *admissible user selection* is done. Also, COR algorithm transfers $Q_k(t)$, $\widehat{R_k}(t)$, and $\Omega(t)$ to the QGS algorithm for resource allocation.

In uplink subframe t, the BS will arrange $K_a(t)$ feedback channels whose position is pre-defined and is known by users. Define *admission bit map* $\vec{a}(t) = [a_1(t), \ldots, a_K(t)], a_k(t) \in \{0, 1\}$, and user k is allowed to feed back if the corresponding admission bit $a_k(t) = 1$, otherwise $a_k(t) = 0$. After *admission bit map* is broadcasted at the beginning of frame t, the user k receiving $a_k(t) = 1$ will report CSI through $s_k(t)^{th}$ feedback channel, where $s_k(t) = \sum_{m=1}^{k} a_m(t)$, and the BS can recognize that reported CSI is belonged to which *admissible user* by the order of the feedback channels without additional identifying bits.

IV. Simulation Results

A. Simulation Environment

In the simulations, the system level parameters of downlink OFDMA environment are set to be compatible with the IEEE 802.16 standard [6], and the scalable parameters, referred to [8], are given as follows: the cell size is 1.6 km, the frame duration is 5 ms, the system bandwidth is 5MHz, the FFT size is 512, the subcarrier frequency spacing is 10.9375 KHz, the number of data subcarriers is 384, the number of subchannels (N) is 8, the number of data subcarriers per subchannel (n_s) is 48, the number of slots for downlink (L_D) and uplink (L_U) transmission are 20 and 10, respectively, the maximum transmission power (P_T) is 23 dBm, and the thermal noise density is -174 dBm/Hz. Therefore, the maximum downlink capacity C_D is 9.216 Mbps. Furthermore, since the CSI is modulated by QPSK, the maximum CSI feedback

capacity is achieved while the whole uplink subframe is used for CSI feedback, and it is equal to bNL_U bpf or 1.536 Mbps in this case.

The traffic models of voice, video, HTTP, and FTP are given below. The constant rate voice traffic is modeled as an ON-OFF model. The streaming video traffic is assumed that its arrive is a regular interval, each frame is decomposed into eight slices (packets), and the size of a packet is distributed in a truncated Pareto distribution. The HTTP traffic is modeled as a sequence of page downloads, and each page download is modeled as a sequence of packet arrivals. The FTP traffic of BE service is modeled as a sequence of file downloads, and the size of a file is distributed in a truncated lognormal distribution. In the simulations, the number of users in each traffic type is assumed to be the same. The traffic intensity of the system is defined as the ratio of the total average data rate of users over C_D. The average date rate of each voice, video, HTTP, or FTP arrival user is equal to 5.2 kbps, 64 kbps, 14.5 kbps, or 88.9 kbps, respectively. Thus, the traffic intensity varies from 0.18 to 0.93 as the number of users varies from 40 to 200.

For the QoS requirements, the maximum delay tolerance of voice (video) is 40 ms (20 ms), both the maximum allowable packet dropping rate of voice and video are set to be 1%, the minimum required transmission rate of HTTP is 100 Kbps (500 bpf), and the required BER for voice, video, HTTP, and FTP are 10^{-3}, 10^{-4}, 10^{-6}, and 10^{-6}, respectively. The parameters of the QGS algorithm are set as follows: $q_{RT} = 5$, $q_{NRT} = 3$, $q_{BE} = 1$, $\alpha = 1.5$, $\beta = 0.3$.

B. Performance Evaluation

We will compare the performance of the following scheduling schemes. (i) ECOS-Q_{th} is the proposed ECOS scheme with $Q_{th} = 1$, 10, or 20, (ii) QGS+BS1-Q_{th} is the QGS algorithm with best-n feedback ($n = 1$) [1] and $Q_{th} = 1$, (iii) QGS+FL-Q_{th} is the QGS algorithm with full CSI feedback and $Q_{th} = 1$ or 20, and (iv) ARRA+FL is the ARRA algorithm proposed in [8] with full CSI feedback. The ARRA algorithm is a priority based scheduling scheme, where the priority is determined by time-to-expired value and the remaining bit of the HoL packet. However, the design of priority of the ARRA algorithm may cause that the small size RT packet is incurred with a lower priority than NRT packets even through the RT packet almost violates the delay requirement. For the CSI overhead, the size of CSI of QGS+BS1-Q_{th} is $2+\log N$ bits/user/frame, where $\log N$ is the overhead for subchannel identification, and the size of CSI of other scheduling schemes is $2N$ bits/user/frame.

Fig. 3 shows the uplink occupancy ratio for CSI feedback, defined as the ratio of the average number of CSI feedback users multiplied the size of CSI, over the maximum CSI feedback capacity. It shows that the ECOS scheme achieves the best performance which consumes less than 5% uplink bandwidth for CSI feedback at high traffic load, while QGS+BS1-Q_{th} (QGS+FL-Q_{th} or ARRA+FL) consumes about 10%-15% (35%-42%) uplink bandwidth for CSI feedback. As a result, the COR algorithm in the ECOS scheme economizes about 37% uplink bandwidth in CSI feedback with respect to full CSI feedback, which equals to a gain of about 1.7 Mbps uplink bandwidth for data transmission (under the assumption of that the highest modulation order of uplink is 64-QAM; therefore, the maximum

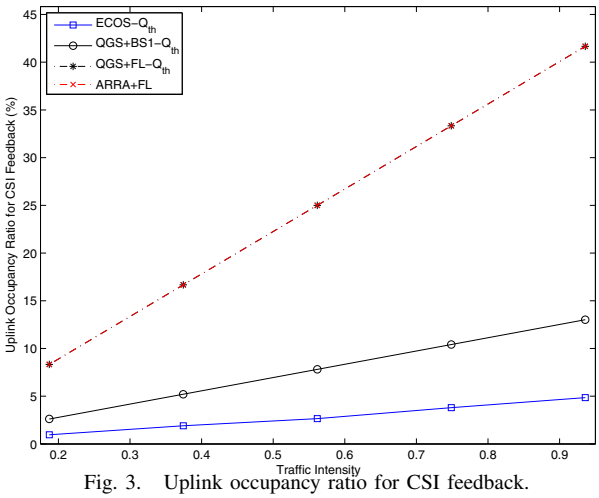

Fig. 3. Uplink occupancy ratio for CSI feedback.

uplink capacity is $3bNL_U = 4.608$ Mbps). Table I shows the average number of CSI feedback users (*admissible users*) of ECOS-Q_{th}. It can be seen that the average *admissible users* number is about 12% of total number of users. This can be regarded as that the ECOS scheme economizes 88% transmission power of a user which is used for feedback on average. Also, the computation complexity of the QGS algorithm can be reduced by 98% when the computation complexity grows with K^2.

TABLE I. AVERAGE NUMBER OF ADMISSIBLE USERS

Total user no.	40	80	120	160	200
Average admissible user no.	4.7	8.9	13.2	18.5	24.5

Fig. 4 (a) and 4 (b) show the guarantee ratio of voice and video users, respectively, where the ratio is defined as the ratio of the number of voice (video) users whose packet dropping rate is below the maximum allowable packet dropping rate over the total number of voice (video) users. It can be seen that most of RT users of QGS+FL-20 violate the QoS requirement because QGS+FL-20 prefers to maximize throughput rather than to guarantee QoS, in contrast to QGS+FL-1. Similarly, the ECOS-20 has poor performance on the QoS guarantee of RT users, while ECOS-1 and ECOS-10 are good. The poor performance on QoS guarantee of $Q_{th} = 20$ is caused by $Q_{th} > 13.6$, where 13.6 is maximum achievable priority value and can be obtained by Eq.(1); in this case, the goal of the QGS algorithm is mainly to maximize throughput without QoS guarantee. Although the Q_{th} of the QGS+BS1-1 is set small to maintain QoS, some video users of QGS+BS1-1 are unsatisfied with maximum allowable packet dropping rate when the traffic intensity is above medium high. The reason is that there is only 1 of 8 subchannels can be allocated to a user by QGS+BS1-1, which may be insufficient for a RT user to transmit a HoL packet within the maximum delay tolerance. Some RT (both voice and video) users of ARRA+FL are also unsatisfied at high traffic intensity. It is caused by inadequate design of priority of the ARRA algorithm, which we have mentioned before. For NRT users, the minimum transmission rate of all scheduling schemes is able to be guaranteed.

Fig. 5 shows the system throughput of downlink. Note that the throughput degradation caused by dropped RT packets has been considered in the simulations. The QGS+FL-20 has the best performance at high traffic load, but unfortunately QGS+FL-20

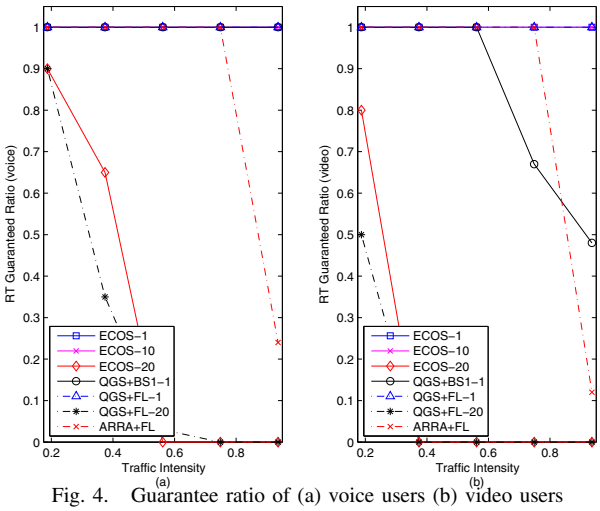

Fig. 4. Guarantee ratio of (a) voice users (b) video users

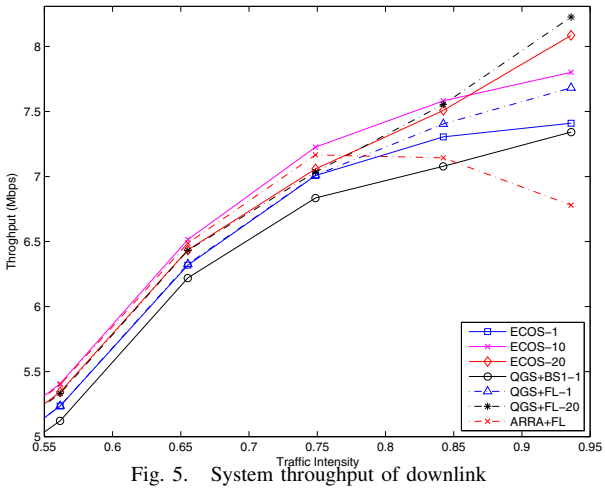

Fig. 5. System throughput of downlink

does not guarantee QoS for RT users. To observe the effectiveness of the COR algorithm by comparing ECOS-1 and QGS+FL-1, it can be found that ECOS-1 has only about 0.2 Mbps throughput degradation, but it gains about 1.7 Mbps uplink bandwidth. Moreover, the Q_{th} provides a trade-off between QoS guarantee and throughput maximization. It can be seen that, by comparing the performance between ECOS-1, ECOS-10, and ECOS-20, the throughput of ECOS-20 (ECOS-1) is the highest (lowest), but ECOS-20 has the worst QoS guarantee. In this case, ECOS-10 would have the best balance performance, where the throughput of ECOS-10 is about 0.4 Mbps ahead of ECOS-1, and ECOS-10 also guarantees QoS requirements.

V. Conclusions

In this paper we propose an *economized-CSI opportunistic scheduling* (ECOS) scheme for OFDMA/TDD downlink systems. The ECOS scheme consists of a *QoS guarantee scheduling* (QGS) algorithm and a *CSI overhead reduction* (COR) algorithm. From the perspective of base station, the COR algorithm economizes the uplink bandwidth used for CSI reporting. From the perspective of mobile users, the COR algorithm saves the transmission power used for CSI reporting since they do not need to feed back every frame. Simulation results show that the ECOS scheme can achieve high system throughput. Besides, it only consumes less than 5% uplink bandwidth for CSI feedback. It gains about 1.7 Mbps uplink bandwidth for data transmission with only about 0.2 Mbps downlink throughput penalty at high traffic intensity. On the other hand, the average number of *admissible users* of the ECOS scheme is about 12% of total number of users, which means (i) 88% transmission power of a user which is used for feedback is saved on average, and (ii) 98% of the computation complexity of the QGS algorithm is reduced when the computation complexity grows with K^2. Also, the Q_{th} provides a balance between QoS guarantee and throughput maximization. How to get an optimal value of Q_{th} is a future work. To summarize, the ECOS scheme can provide high system throughput and guarantee satisfactory QoS requirements with minimum CSI feedback overhead for OFDMA/TDD downlink systems.

References

[1] Z. H. Han and Y. H. Lee, "Opportunistic scheduling with partial channel information in OFDMA/FDD systems," *IEEE VTC 2004 Fall*, vol. 1, pp.511-514.

[2] Y. Rong, S. A. Vorobyov, and A. B. Gershman, "Adaptive OFDM techniques with one-bit-per-subcarrier channel-state feedback," *IEEE Trans. Commun.*, vol. 54, no. 11, pp. 1993-2003, Nov. 2006.

[3] Y. Xue and T. Kaiser, "Exploiting multiuser diversity with imperfect one-bit channel state feedback," *IEEE Trans. Veh. Tech.*, vol. 56, no. 1, pp.183-193, Jan. 2007.

[4] S. Sanayei and A. Nosratinia, "Opportunistic downlink transmission with limited feedback," *IEEE Trans. Info. Theory*, vol. 53, no. 11, pp. 4363-4372, Nov. 2007.

[5] R. Agarwal, V. R. Majjigi, Z. Han, R. Vannithamby, and J. M. Cioffi, "Low complexity resource allocation with opportunistic feedback over downlink OFDMA networks," *IEEE J. Select. Area Commun.*, vol. 26, no. 8, pp. 1462-1472, Oct. 2008.

[6] IEEE Standard Std. 802.16e, "Local and metropolitan area networks-part 16: air interface for fixed and mobile broadband wireless access systems," 2005.

[7] A. J. Goldsmith and S. G. Chua, "Variable-rate variable-power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, pp. 1218-1230, Oct. 1997.

[8] 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, May 2008.