Feedback Reduction for Multiuser OFDM Systems

Jeongho Jeon, *Student Member, IEEE*, Kyuho Son, *Student Member, IEEE*,
Hyang-Won Lee, *Member, IEEE*, and Song Chong, *Member, IEEE*

***Abstract*—Feedback reduction in multiuser orthogonal frequency-division multiplexing (OFDM) systems has become an important issue due to the excessive amount of feedback required to use opportunistic scheduling, particularly when the number of users and carriers is large. In this paper, we propose a novel feedback-reduction scheme for efficient downlink scheduling. In the proposed scheme, each user determines the amount of feedback based on the so-called *feedback efficiency* in a distributed manner. The key idea is to give more of an opportunity for feedback to users who are more often scheduled. Simulation results demonstrate that the proposed scheme can substantially decrease the feedback load while achieving almost the same scheduling performance as in the case of full feedback. In addition, the proposed scheme offers unique advantages over existing ones. First, it is not tailored to a specific scheduling policy; thus, it has adaptability to the change of the underlying scheduling policy. Second, the total feedback load can be maintained below a target level, regardless of the number of users in the system.**

***Index Terms*—Channel quality indicator (CQI) feedback, opportunistic scheduling, orthogonal frequency-division multiplexing (OFDM) system.**

I. Introduction

OPPORTUNISTIC scheduling exploits independently fading wireless channels to enhance the throughput performance of wireless systems [1]. For this, each user should send the channel quality indicator (CQI) to the base station (BS) in each time slot. The amount of CQI feedback increases as the number of users in the system increases, and obviously, it becomes a more serious problem in multicarrier (interchangeably, multichannel) systems where the CQI information of every subcarrier (interchangeably, subchannel) must be known to the BS. Clearly, the goal of feedback reduction is to reduce the feedback load, but at the same time, the performance of a scheduler can be deteriorated due to the insufficient feedback information. Hence, the key challenge is to minimize the adverse impact on the scheduler (i.e., to enable a scheduler to achieve comparable performance to the case of full feedback), while keeping the feedback information minimal.

There have been several works in this context. In the *best-M* feedback scheme [2], each user selects M best subchannels having the highest signal strength and sends the CQIs of these subchannels to the BS. Under this scheme, it can be conjectured that the fairness-oriented scheduling policies such as the proportional fairness scheduler (PFS)[1] [3] will behave as designed since all users are given an equal amount of feedback opportunity M. However, the performance of the maximum-rate scheduler (MRS) [4], [5] (which serves the user with the highest achievable rate in each time slot) can be degraded, because some of the subchannels can be scheduled to a user in deep fading, thereby yielding throughput loss, unless M is close to the total number of subchannels. In [6], the *absolute SNR thresholding* scheme was proposed, wherein user k sends feedback if $\gamma_k^t \geq \gamma_{\mathrm{th}}$, where γ_k^t denotes the instantaneous SNR of user k at time slot t, and γ_{th} is the threshold. This can easily be extended to multichannel systems by applying the same threshold to each subchannel. Since, however, users in deep fading will have no chance of being scheduled, the fairness-oriented scheduling policies will exhibit severe performance degradation. To solve this problem, the absolute SNR thresholding scheme was modified such that user k sends feedback when the normalized SNR $\gamma_k^t/\bar{\gamma}_k$ exceeds a certain threshold A, where $\bar{\gamma}_k$ is the average of γ_k^t from time slot 0 to t. However, this *normalized SNR thresholding* scheme has the same problem as that of the best-M scheme in the long term, because each user sends approximately the same amount of feedback on the average. Consequently, it may work well with the PFS but not with the MRS.

Note that all of these schemes have a common drawback that the total amount of feedback tends to increase in proportion with the number of users in the system, which significantly undermines their practicality for a large system. To limit the total feedback load, a random-access-based feedback scheme, known as *opportunistic feedback*, was proposed in [8], where the feedback opportunities are granted for the users who succeed in the underlying random access competition, and the users satisfying $\gamma_k^t \geq \gamma_{\mathrm{th}}$ are allowed to send the message

Manuscript received April 7, 2008; revised January 30, 2009, May 2, 2009, and July 13, 2009. First published August 11, 2009; current version published January 20, 2010. This work was supported in part by the Ministry of Knowledge Economy, Korea, through the Information Technology Research Center support program supervised by the Institute of Information Technology Advancement (IITA) under Grant IITA-2009-C1090-0902-0037 and in part by the Defense Acquisition Program Administration and Agency for Defense Development under the contract. The material in this paper was presented in part at the IEEE International Conference on Communications Glasgow, U.K., June 2007. The review of this paper was coordinated by Prof. T. J. Lim.

J. Jeon is with the Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742 USA, and also with the Institute for Systems Research, University of Maryland, College Park, MD 20742 USA (e-mail: jeongho@umd.edu).

K. Son and S. Chong are with the School of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: skio@netsys.kaist.ac.kr; song@ee.kaist.ac.kr).

H.-W. Lee is with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: hwlee@mit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2009.2029860

[1] Under the PFS, users compete for resources based on the achievable rate normalized by their respective average rate.

0018-9545/$26.00 © 2009 IEEE

containing its identity during the competition. After that, the BS randomly allocates the feedback opportunities among the successful users. This scheme can explicitly control the feedback load, regardless of the number of users in the system, but inherently possesses the same drawback as that of the absolute SNR thresholding scheme and suffers from long feedback delays due to the underlying competition process.

More importantly, it should be mentioned that all the previous schemes use the SNR as a feedback decision metric, and this SNR-based approach seems to have a fundamental limitation in guaranteeing the performance of various types of scheduling policies. We take a different approach to this problem by observing that, ideally, only the users that will be served in each time slot need to report their CQIs. To mimic this ideal case, our scheme gives more feedback opportunities to the users who are more likely to be scheduled. For this, we first introduce *feedback efficiency*, which is defined as the ratio of the average number of allocated subchannels to the amount of feedback. Under our proposed scheme, all the active users are expected to maintain the same *target feedback efficiency* by adjusting their feedback amount in each time slot. It is shown that our proposed scheme achieves almost the same scheduling performance as in the full feedback while substantially reducing the total feedback load. Moreover, there are two additional unique advantages over the previous schemes. First, it has adaptability to the change of the underlying scheduling policy and thus works well with a broad class of scheduling policies, ranging from the MRS aiming at maximum efficiency to the maximum-fairness scheduler (MFS)[2] [4] aiming at maximum fairness. Second, it explicitly controls the total feedback load below a target level, regardless of the number of users in the system.

This paper is organized as follows. In Section II, we describe the system model and define some notations. In Section III, as a preliminary step to develop a feedback algorithm, we investigate how users share resources under scheduling policies of different fairness criteria. Section IV proposes and analyzes the *efficiency-based feedback* algorithm in detail. In Section V, we present numerical results that evaluate the performance of our proposed algorithm in terms of the total throughput, the per-user throughput, fairness, and the total feedback load. We conclude this paper in Section VI.

II. System Model

We consider an orthogonal-frequency-division-multiplexing-based single-cell downlink with K users and N subchannels. Denote by $\mathcal{K}$ and $\mathcal{N}$ the set of all users and subchannels, respectively. A time-slotted system is considered, and an example frame structure is shown in Fig. 1. Let $\mathbf{h}_t = [\mathbf{h}_k^t, \forall k \in \mathcal{K}]$ be the channel state vector at time slot t, where $\mathbf{h}_k^t$ is an N-by-1 vector. We assume that the channel state is fixed during a slot. Let $\Gamma_{\mathbf{h}_t}$ be the feasible region of the transmission rate vector $\mathbf{r}_t = [r_k^t, \forall k \in \mathcal{K}]$ for given channel state $\mathbf{h}_t$ at time slot t, i.e., $\mathbf{r}_t \in \Gamma_{\mathbf{h}_t}$. The transmission rate of user k at time slot t is determined by a scheduler as $r_k^t = \sum_{n=1}^{N} x_{k,n}^t r_{k,n}^t$, where $r_{k,n}^t$ is the achievable rate of user k over subchannel n during time slot t, and $x_{k,n}^t$ is an indicator function such that $x_{k,n}^t = 1$ if user k is chosen to be served over subchannel n at that time slot and $x_{k,n}^t = 0$ otherwise. We restrict that any subchannel cannot be shared by multiple users during a slot, i.e., $\sum_{k=1}^{K} x_{k,n}^t \leq 1, \forall n, \forall t$. Let R_k^t be the average throughput of user k up to time slot t, i.e., $R_k^t = (1/t)\sum_{\tau=1}^{t} r_k^\tau$, and let $\mathbf{R}_t = [R_k^t, \forall k \in \mathcal{K}]$. It is assumed that each user is associated with a utility function $U_k(R_k^t)$ of their average throughput, and let $U(\mathbf{R}_t) = \sum_{k=1}^{K} U_k(R_k^t)$, where the utility function $U_k(\cdot)$ is an increasing, strictly concave, and continuously differentiable function on $\mathfrak{R}_+$. Define the long-term average throughput vector as $\mathbf{R} = \lim_{t\to\infty} \mathbf{R}_t$ and let Γ be the long-term feasible region of the long-term average throughput vector $\mathbf{R}$ by considering all the possible scheduling policies. It can easily be shown that the long-term feasible region Γ is a convex and compact set [9].

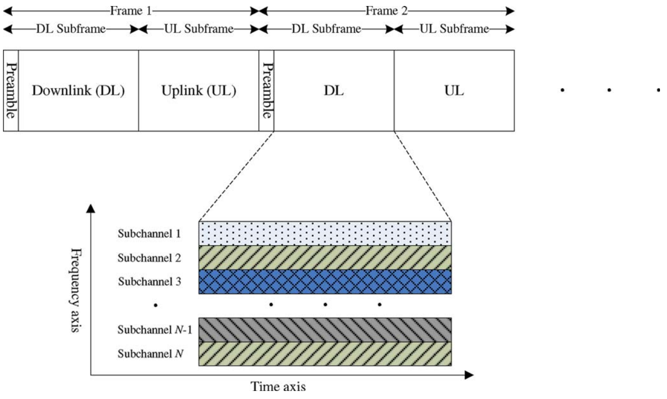


Fig. 1. Example frame structure of a multichannel system.

We assume that users know which subchannels are allocated to them in each time slot. Note that this is not merely an assumption because, for example, in the IEEE 802.16e system, the DL-MAP (downlink map) message contains such resource-allocation information. It is also assumed that the feedback transmission from a user to the BS is collision-free and that the number of subchannels that is fed back to the BS need not be fixed. This can be done by sending the feedback information through the uplink data channel. Indeed, in the IEEE 802.16e system, users periodically send the REP-RSP (channel measurement report response) message containing the CQIs of the five best subchannels and indicating bitmaps through the data channel.[3]

III. Understanding Schedulers

In designing a wireless packet scheduler, the total throughput and fairness among users are two competing interests; the MRS can be used to maximize the throughput, whereas the MFS can be used to maximize fairness. To trade off between them, the PFS was proposed [3]. The α-PFS is a generalized and

[2] As an extreme pursuit of fairness, the MFS allocates resources in each time slot to maximize the *minimum* user's average data rate.

[3] In the IEEE 802.16e system, users report the differential of channel states for the five selected subchannels on its dedicated fast-feedback channel (CQICH) with a step of 1 dB in between the REP-RSP messages [19].

unified form of those schedulers [14]. It is often referred to as the generic proportional fair scheduler in the IEEE 802.16 Task Group m [17]. This section investigates how users share resources under the α-PFS of different fairness criteria.

Many of the scheduling policies—including the PFS—can be viewed as a gradient-based algorithm [10], [11]. It aims to maximize the sum utility by maximizing $U(\mathbf{R}_{t+1}) - U(\mathbf{R}_t)$ in each time slot, which is equivalent to choosing the transmission rate vector having the maximum projection onto the gradient of the sum utility as

$$\arg \max_{\mathbf{r}_t \in \Gamma_{\mathbf{h}_t}} \nabla U(\mathbf{R}_t)^T \cdot \mathbf{r}_t \tag{1}$$

where T denotes the vector transpose. With a strictly concave utility function, it has been proved that such a gradient-based algorithm maximizes sum utility $U(\mathbf{R})$ over the long-term feasible region Γ [10], [12].

By changing utility functions, the gradient-based scheduler can achieve various objectives. For example, the following function enables achieving any compromise between fairness and efficiency [13], [14]:

$$U(\mathbf{R}_t) = \begin{cases} \sum\limits_{k\in\mathcal{K}} (1-\alpha)^{-1} \left(R_k^t\right)^{1-\alpha}, & \alpha > 0,\ \alpha \neq 1 \\ \sum\limits_{k\in\mathcal{K}} \log\left(R_k^t\right), & \alpha = 1. \end{cases} \tag{2}$$

With this utility function, the scheduler in (1) becomes

$$\arg \max_{k\in\mathcal{K}} \left(R_k^t\right)^{-\alpha} r_k^t \tag{3}$$

where we have assumed the single-channel system $(N = 1)$ for simplicity of illustration. As $\alpha \to 0$, the total throughput is maximized (MRS), and when $\alpha = 1$, the *proportionally fair* throughput allocation is achieved (PFS). As α increases, the fairness is improved at the cost of reduced total throughput, and, particularly as $\alpha \to \infty$, the throughput allocation becomes *max-min fair* (MFS). Under the scheduling policy in (3), each user shares the channel as in Observation 1 below.

Observation 1: For the two-user case, each user receives time slots in a long-term sense as in the following ratio:

$$\theta_1' = \frac{1}{1 + \left(\frac{R_1^{\max}}{R_2^{\max}}\right)^{(\alpha-1)/\alpha}}, \quad \theta_2' = 1 - \theta_1' \tag{4}$$

where $R_k^{\max}$ is defined as the maximum achievable long-term throughput of user k, i.e., $R_k^{\max} = \lim_{t\to\infty}(1/t)\sum_{\tau=1}^{t} \sum_{n=1}^{N} r_{k,n}^{\tau}$. This is simply the long-term average throughput of user k when user k is scheduled for all subchannels and for all time slots.

Proof: See the Appendix for the proof. ■

Fig. 2 plots (4) when the long-term maximum achievable rate of user 1 is twice than that of user 2, i.e., $R_1^{\max}/R_2^{\max} = 2$. When $\alpha = 0$ (MRS), the resource-sharing ratio becomes $\theta_1' = 1$ and $\theta_2' = 0$, i.e., the *strong user* monopolizes all the resources. Therefore, the best-M and normalized SNR thresholding schemes will reveal redundant feedback, because the *weak user* does not need to send feedback at all. Note that the absolute SNR thresholding scheme will be well matched with the MRS because the *strong user* is given more feedback opportunities. On the other hand, when $\alpha = 1$ (PFS), each user receives an equal fraction of time slots. From this, it can be conjectured that if there are totally K users in the system, each user will equally receive fraction $1/K$ of the time slots, which is referred to as the *equal time sharing* property of the PFS [16]. Hence, the absolute SNR thresholding scheme can cause a serious fairness problem because the *weak user* has no chance of being scheduled at all due to the lack of feedback information. Nevertheless, the best-M and normalized SNR thresholding schemes will be harmonious with PFS because these schemes give an equal or approximately equal number of feedback opportunities to users, irrespective of their relative channel strengths. The case of $\alpha \to \infty$ (MFS) can also analogously be explained. This observation indicates that the previous schemes have their own single ideal operating point (i.e., $\alpha = 0$ for the absolute SNR thresholding scheme and $\alpha = 1$ for the best-M and normalized SNR thresholding schemes). More importantly, the SNR (which was the feedback decision metric in the previous schemes) is not the only criterion in determining the users to be served, but the criterion can be differed according to the underlying scheduling policy. This shows the necessity of designing a feedback algorithm that can work well with the schedulers of different fairness criteria.

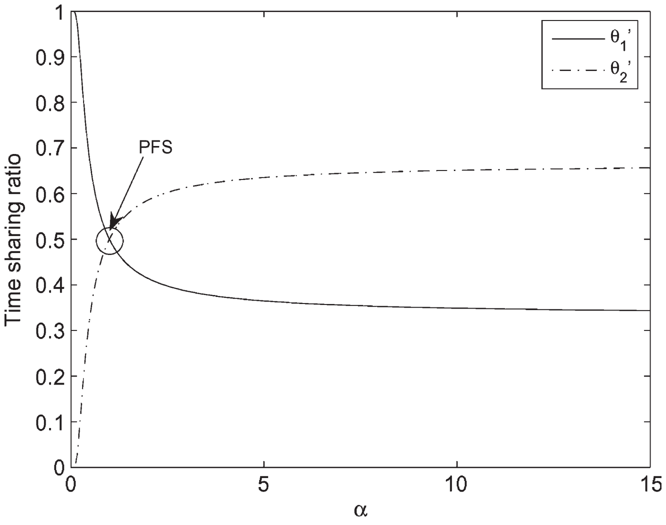


Fig. 2. Time sharing ratio in the two-user case where $R_1^{\max}/R_2^{\max} = 2$.

IV. Efficiency-Based Feedback-Reduction Algorithm

A. Description of the Algorithm

The key idea behind our proposed algorithm is to give more feedback opportunities to users who are more often scheduled. To realize such a concept in determining each user's feedback amount, we introduce *feedback efficiency*, which is defined as the ratio of the average number of allocated subchannels $\bar{s}_k^t$ to the amount of feedback f_k^t for user k at time slot t. Specifically, f_k^t denotes the number of subchannels whose CQI information was sent back to the BS at time slot t, and the average number of allocated subchannels is given by $\bar{s}_k^t = (1/t)\sum_{\tau=1}^{t} s_k^{\tau}$.

A proper *target efficiency* value will be predetermined based on the system objective, and all users are expected to maintain the same *efficiency* by adjusting their feedback amount in each time slot.

The efficiency-based feedback reduction (EFR) scheme is presented in Algorithm 1. For the given target efficiency e, each user k computes the probability p_k^t at time slot t as

$$p_k^t = \left[\frac{1}{N} \cdot \frac{\bar{s}_k^t}{e}\right]_1^+ \tag{5}$$

where $[\cdot]_1^+$ denotes the orthogonal projection onto the interval [0, 1].[4] After that, each user k generates the binomial random variable X_k with parameters N and p_k^t and sends the CQIs of the best X_k subchannels to the BS. Therefore, the scheduling of each of the subchannels is done for the reduced set of users who sent feedback on that subchannel. Once a scheduling decision is made, each user updates its average number of allocated subchannels as $\bar{s}_k^{t+1} = (1/t+1)\sum_{\tau=1}^{t+1} s_k^\tau$. Note that if target efficiency e increases, then each user sends less feedback than before, and by doing so, the feedback efficiency is maintained at its target level.

```
Algorithm 1 EFR Algorithm
  1: At time slot $t$,
     – Compute the probability $p_k^t$ using (5).
     – Generate the binomial random variable $X_k$ with
       parameters $N$ and $p_k^t$.
  2: At time slot $t+1$,
     – Select $f_k^{t+1} = X_k$ best subchannels and send back
       the CQIs of these subchannels to the BS.
     – Update the average number of allocated subchannels
       $\bar{s}_k^{t+1}$.
     – Set $t = t+1$ and go to 1.
```

Recall Section III, where we derived the long-term time sharing ratio $[\theta'_k, \forall k \in \mathcal{K}]$ between users, which is equal to the long-term average number of allocated subchannels $\bar{s}_k^t$ for the single channel system. In this case, the probability p_k^t of user k is given by $[(\theta'_k/e)]_1^+$ in the long-term sense. From this, one can easily notice that the expected amount of feedback is proportional to the resource-sharing ratio under our proposed scheme. This argument can readily be extended to a multichannel system without loss of generality. Moreover, unlike previous works, our scheme takes into account the underlying scheduling policy through the average scheduling frequency $\bar{s}_k^t$, which is used to decide the amount of feedback.

B. Feedback Load Estimation

A common drawback of the previous schemes is that the total feedback load increases in proportion to the number of users. However, under our proposed scheme, the total feedback load can be maintained below a target level, even when the number of users changes. Note that, as the number of users increases, the average number of allocated subchannels $\bar{s}_k^t$ of each user will be decreased. Consequently, the probability p_k^t will be lowered, which results in a reduced amount of feedback f_k^t. Owing to this self-regulating mechanism, the total feedback load, which is the sum of all the individual users' feedback, can be maintained below a target level (which relies on the target efficiency), regardless of the number of users in the system.

Note that, under our proposed scheme, each user sends X_k amount of feedback, which follows the binomial distribution with parameters N and p_k, where $p_k = [(\bar{s}_k/Ne)]_1^+$, and time index t is dropped for brevity. Thus, the kth user's expected amount of feedback is given by

$$E[X_k] = Np_k \leq \frac{\bar{s}_k}{e}$$

and the total feedback load is

$$\sum_{k\in\mathcal{K}} E[X_k] \leq \frac{1}{e}\sum_{k\in\mathcal{K}} \bar{s}_k = \frac{N}{e} \tag{6}$$

where the equality, i.e., $\sum_{k\in\mathcal{K}} \bar{s}_k = N$, in (6) follows from the fact that the available number of subchannels is always N, regardless of the number of users. Finally, the total feedback load per channel of our efficiency-based feedback scheme is bounded by

$$F_{\text{EFR}} \leq \frac{1}{e}. \tag{7}$$

Notice that the total feedback load does not depend on the number of users in the system, but it is simply the inverse of the target efficiency that is under our control.

We now look at the feedback load of previous schemes by assuming that users experience independent identically distributed fading. For a Rayleigh fading channel, the instantaneous SNR γ_k is exponentially distributed with the following probability density function [18]:

$$f(\gamma_k) = \frac{1}{\bar{\gamma}_k}\exp\left(-\frac{\gamma_k}{\bar{\gamma}_k}\right), \quad \gamma_k \geq 0$$

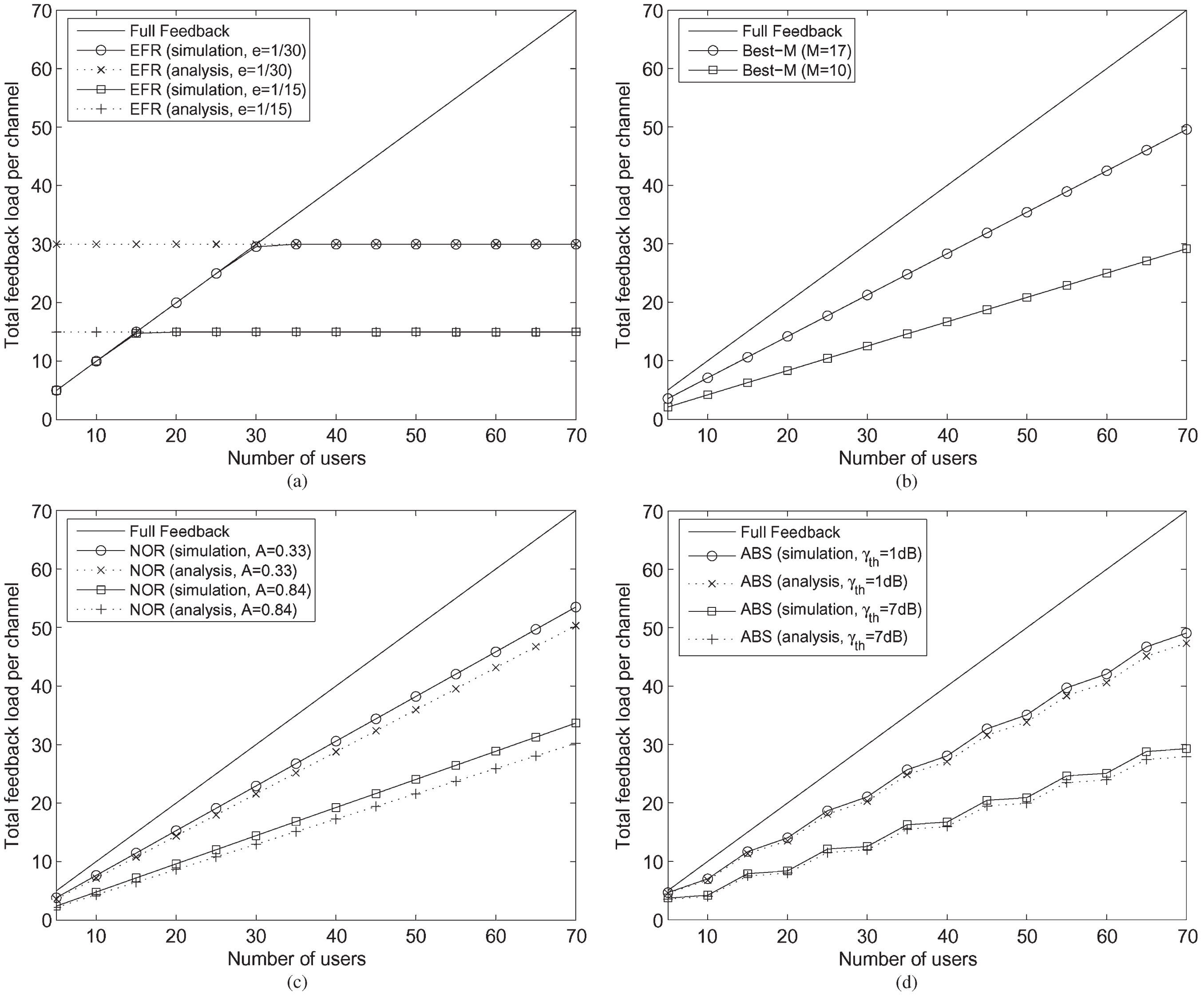
where $\bar{\gamma}_k$ is the average SNR. Under the absolute SNR thresholding scheme, each user sends the CQI if the instantaneous SNR exceeds a certain threshold γ_{th}. Thus, the total feedback load per channel of the absolute SNR thresholding scheme is given by

$$F_{\text{ABS}} = \sum_{k=1}^{K}\left(1 - \int_0^{\gamma_{\text{th}}} f(\gamma_k)d\gamma_k\right) = \sum_{k=1}^{K}\exp\left(-\frac{\gamma_{\text{th}}}{\bar{\gamma}_k}\right). \tag{8}$$

On the other hand, under the normalized SNR thresholding scheme, each user sends the CQI when the normalized SNR $\gamma_k/\bar{\gamma}_k$ exceeds a certain threshold A. By simply replacing γ_{th} in (8) with $\bar{\gamma}_k A$ for each user, the total feedback load per channel of the normalized SNR thresholding scheme is given by

$$F_{\text{NOR}} = \sum_{k=1}^{K}\exp(-A) = K\exp(-A). \tag{9}$$

[4] For example, if $N = 4$, $e = (1/5)$, and $\bar{s}_k^t = 1$, then $p_k^t = [(5/4)]_1^+ = 1$.



(a)

(b)

(c)

(d)

Fig. 3. Comparison of the total feedback load. (a) Efficiency-based feedback scheme. (b) Best-M scheme. (c) Normalized SNR thresholding scheme. (d) Absolute SNR thresholding scheme.

The total feedback load per channel of the best-M feedback scheme is deterministically given by

$$F_{\text{Best-}M} = \frac{KM}{N}. \tag{10}$$

Note that both the best-M and normalized SNR thresholding schemes let the users send equal amounts of feedback, irrespective of their relative channel strengths. Specifically, by simply setting $A = -\ln(M/N)$, each user's feedback amount and the total feedback load of the normalized SNR thresholding scheme becomes identical to those of the best-M scheme in the average sense.

To verify the analysis given above, we compare our scheme with the previous ones for a 24-subchannel system. In the case of the absolute SNR thresholding or normalized SNR thresholding scheme, the resulting total feedback load was averaged over multiple subchannels to obtain the per-channel feedback load. Fig. 3 plots the total feedback load per channel of our proposed and previous schemes. As expected, our scheme controls the feedback load below a target level, whereas under other schemes, the feedback load increases with the number of users. Note that, when there are a large number of users, all the previous schemes suffer from the excessive feedback load. On the other hand, when there are a small number of users, their feedback-reduction mechanism can cause the underutilization of downlink resources due to the insufficient feedback. Note that the shortcomings of previous schemes can be solved by adaptively selecting the thresholds or other parameters. To do this, however, it is necessary to continuously monitor the number of users in the system. Unlike previous schemes, our scheme does not require this dynamic parameter selection, because each user automatically adapts its feedback load in response to the scheduling frequency that changes according to the number of users.

C. Selection of the Feedback-Efficiency Factor

Clearly, our scheme can easily trade off between the feedback load and the scheduling performance by controlling the target

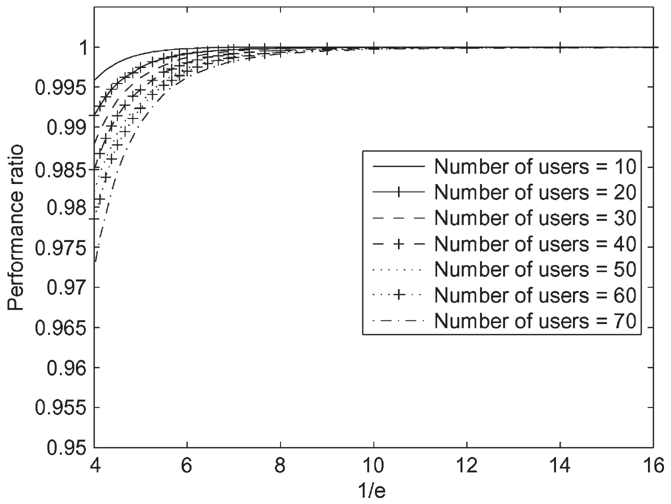


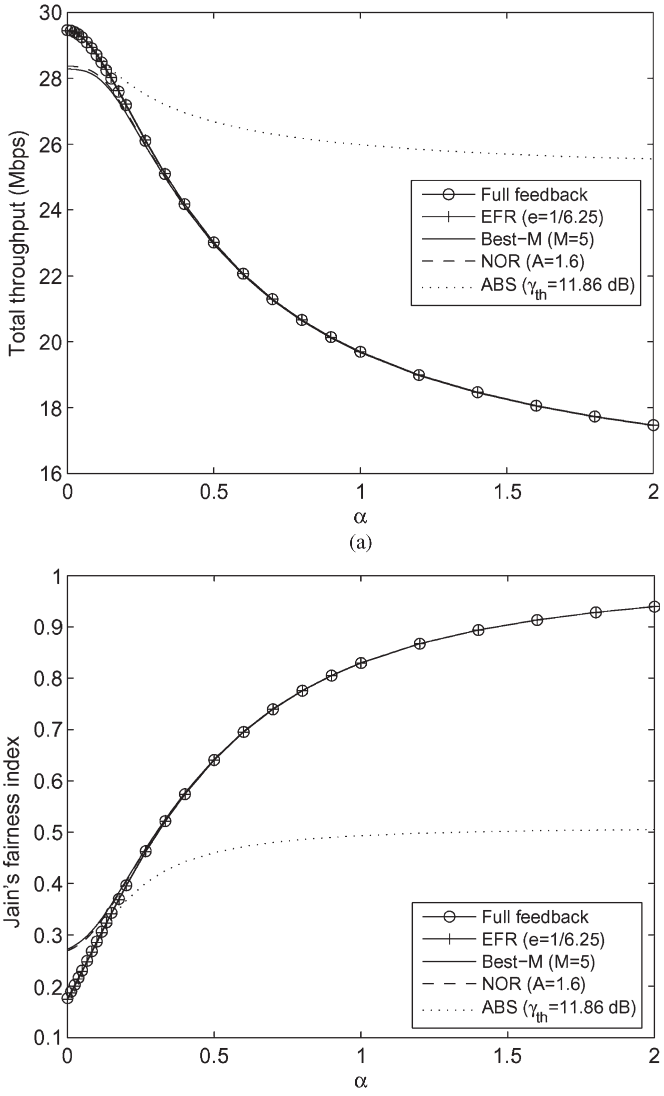
Fig. 4. Performance ratio of the efficiency-based feedback scheme with PFS.

efficiency e. For example, if e is high, the feedback load will decrease, while the scheduling performance can possibly be degraded due to the insufficient feedback. For a small e, the opposite will happen. Hence, the target efficiency must carefully be selected by taking into account the system objectives. If we can find a closed-form expression for the system performance (e.g., the sum utility under the gradient-based scheduler), it will relatively be easy to decide e. However, finding such an expression seems to be intractable. Due to this difficulty, we select the target efficiency by comparing the empirical throughput performance. We used the PFS for this comparison, and thus, the performance metric is the sum utility. Fig. 4 plots the performance ratio, which is defined as the utility under our scheme normalized by that under the full feedback, where the same simulation scenario as in Section V was applied. As $1/e$ increases (low feedback efficiency), our scheme better approximates the full feedback. In particular, at $1/e = 6$, the performance ratio is higher than 99%, regardless of the number of users, and the improvement becomes marginal above $1/e = 10$. Hence, we will use a $1/e$ falling between 6 and 10 later in the simulation. Note that the feedback efficiency can also be chosen using the analysis in Section IV-B to meet the maximum feedback load allowed by the system.

D. Considerations

Our scheme takes a significant step toward the design of scheduler-compatible feedback reduction, but there still remain several challenges for its implementation.

1) In Algorithm 1, each user generates a binomial random variable X_k with parameters N and p_k^t to decide the amount of feedback f_k^t in each time slot t. Since the variance of X_k is given by $Np_k^t(1-p_k^t)$, the algorithm may induce substantial variations in the feedback load when N is large. To deal with this problem, one can consider a more deterministic approach by letting $f_k^t = \text{nint}(Np_k^t)$, where $\text{nint}(\cdot)$ is the nearest integer function. This way, the total amount of feedback in each time slot will fluctuate less, staying close to the target level.

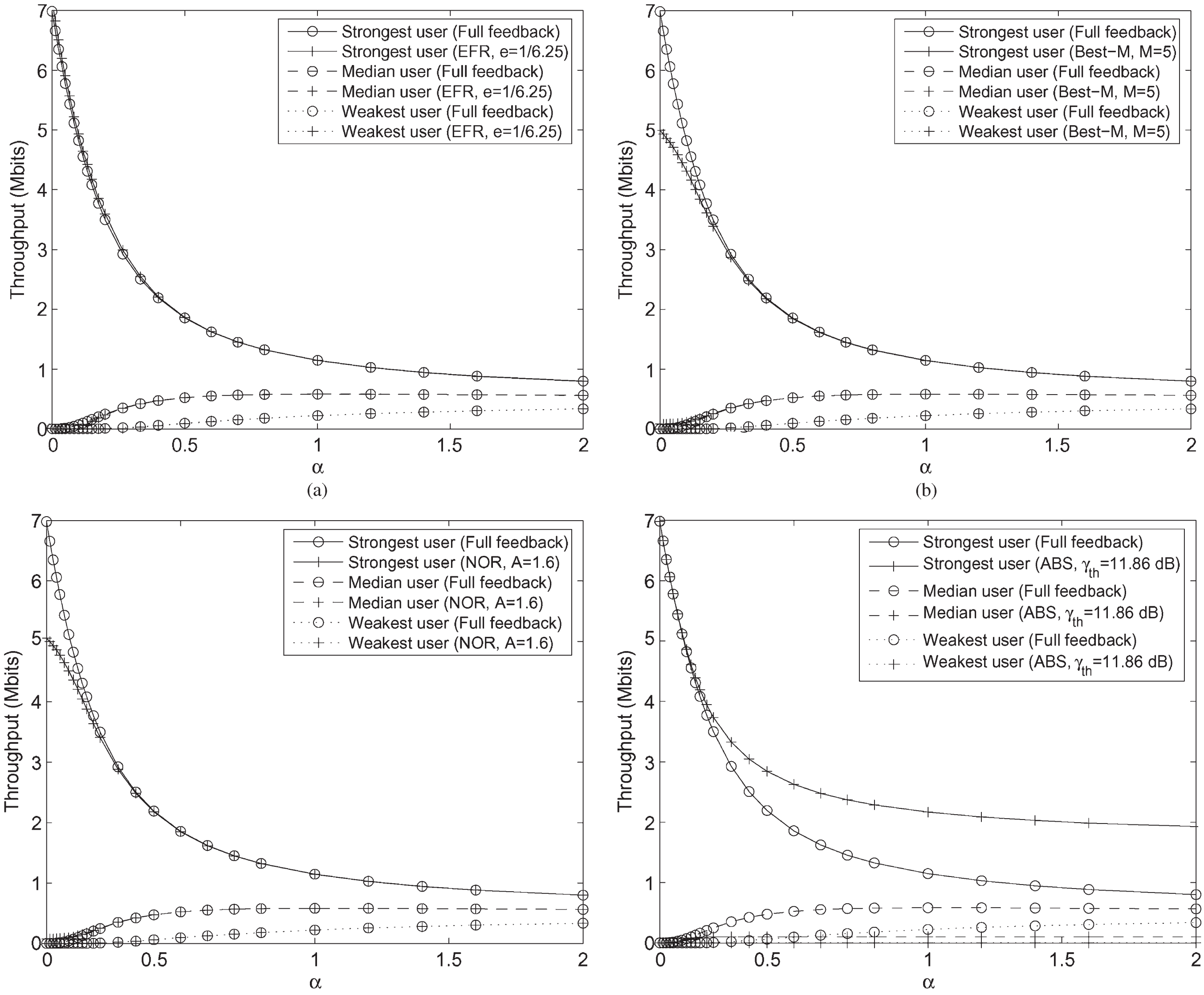


(a)

(b)

Fig. 5. Comparison of total throughput and Jain's fairness index with the scheduling policies of different fairness criteria. (a) Total throughput (number of users = 30). (b) Jain's fairness index (number of users = 30).

2) This paper implicitly assumes that the channel state $\mathbf{h}_t$ is a stationary and ergodic process. However, in reality, it is not likely to be stationary and ergodic, for example, due to user mobility. In this case, the moving average filter can be used to track the dynamics of the average number of allocated subchannels $\bar{s}_k^t$. For example, one can use the exponentially weighted moving average (EWMA) filter as $\bar{s}_k^t = (1-(1/t_e))\bar{s}_k^{t-1} + (1/t_e)s_k^t$, where t_e is the length of the window for the update.
3) Consider an extreme case where a user, e.g., k, has not been scheduled for a long period of time, for example, due to deep fading. Then, $\bar{s}_k^t$ can decrease close to zero, and consequently, the user will not be scheduled hereafter because of no received feedback. To cope with such an extreme case, a *detection mode* can be considered in which such users are allowed to send at least one feedback to check whether the channel statistics were changed for the better.



(a)

(b)

(c)

(d)

Fig. 6. Comparison of per-user throughput with the scheduling policies of different fairness criterions. (a) Efficiency-based feedback scheme (number of users $= 30$). (b) Best-M scheme (number of users $= 30$). (c) Normalized SNR thresholding scheme (number of users $= 30$). (d) Absolute SNR thresholding scheme (number of users $= 30$).

4) We implicitly assume that all users are active in the sense that every user has an infinite amount of data at the BS. However, in reality, some users may have no data for a while, and the feedback from such users will not be needed. In this case, the BS can stop them from sending the feedback. Note that such a variation on the number of active users will be reflected to the remaining users' s_k^t, and our scheme will enable each user to adapt to the new environment.

The modified EFR algorithm is presented in Algorithm 2 by considering the practical issues discussed in this section, and we compare the performance of the modified algorithm with the original one in the following section.

Algorithm 2 Modified EFR Algorithm

```
1: At time slot $t$,
     – Compute the probability $p_k^t$ using (5).
     – Set $f_k^{t+1} = \mathrm{nint}(Np_k^t)$. If $f_k^{t+1} = 0$, set $f_k^{t+1} = 1$.
2: At time slot $t+1$,
     – Select $f_k^{t+1}$ best subchannels and send back the CQIs of these subchannels to the BS.
     – Update the average number of allocated subchannels $\bar{s}_k^{t+1}$ using the EWMA filter.
     – Set $t = t + 1$ and go to 1.
```

V. Simulation Results

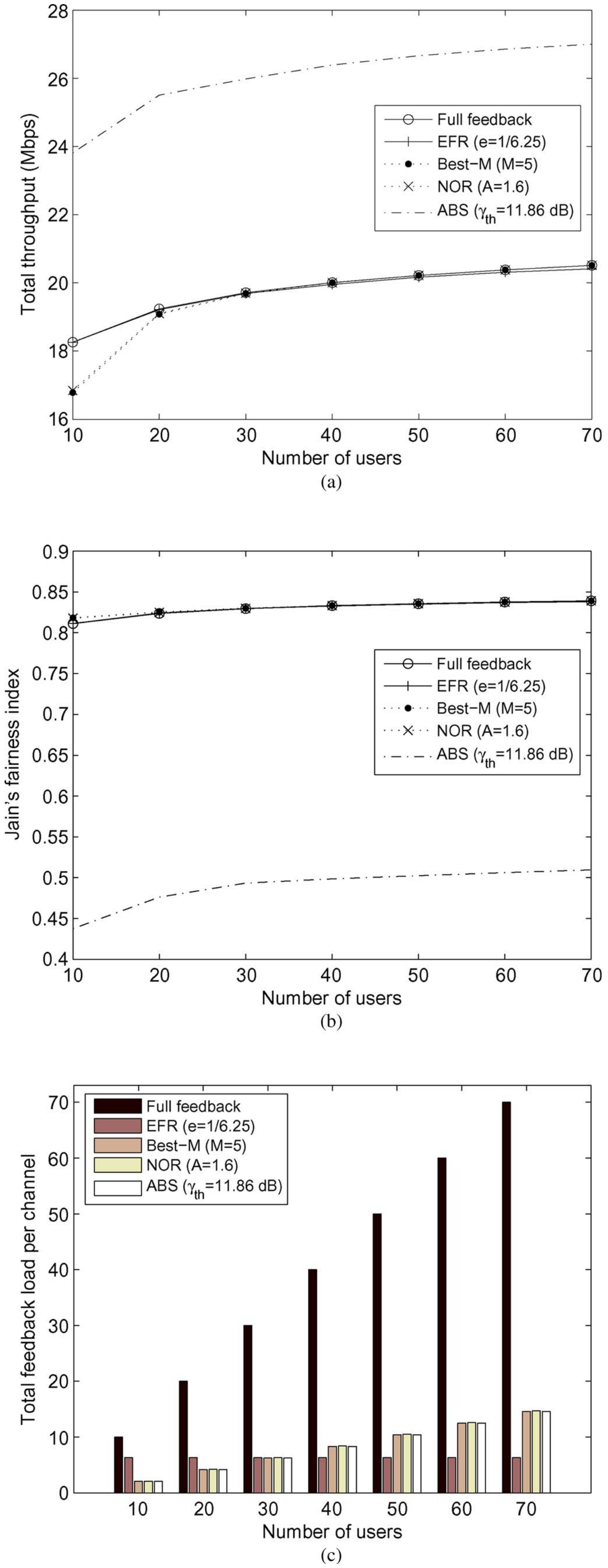
For the simulation, the IEEE 802.16e 1024-FFT orthogonal frequency-division multiple-access adaptive modulation and coding (AMC) mode [19] is used, where the data subcarriers are grouped into 24 subchannels. The AMC mode is possible only if the channel coherence time is much longer than the lag between the time the channel is measured and the time when the packet is actually transmitted [3], [19]. Therefore, the International Telecommunication Union (ITU) pedestrian B model is chosen for the user mobility [20]. Note that, for high-mobility users (usually served in the *diversity mode* in the

IEEE 802.16e system), it is sufficient to predict the average CQI of the overall subchannels, because the channel follows the stationary distribution [3].

The radius of the cell is 1 km, and the distance d_k between user k and the BS is a 2-D uniformly distributed random variable. We use $PL(d_k) = 16.62 + 37.6log_{10}(d_k)$[dB] for the path loss model, and the shadowing has a log-normal distribution with standard deviation $\sigma_s = 8$ dB. The ITU pedestrian B parameters are applied to the standard delay-spread model to generate the frequency-selective fast fading [21]. For a fair comparison, the threshold of each scheme is chosen such that they have the same feedback load at 30 number of users. Using the analysis in Section IV-B, such parameters can be obtained as follows: $M = 5$ (best-M scheme), $\gamma_{\text{th}} = 11.86$ dB (absolute SNR thresholding scheme), $A = 1.6$ (normalized SNR thresholding scheme), and $e = 1/6.25$ (efficiency-based feedback scheme). The length of a time slot is set to 5 ms. We used Jain's fairness index [22] for the fairness comparison, which is given by $((\sum_{k\in\mathcal{K}} R_k)^2/K\sum_{k\in\mathcal{K}}(R_k)^2)$. Note that the index ranges from $(1/K)$ (worst case) to 1 (best case).

In Fig. 5, we compare the throughput and fairness performance under the scheduling policies of different fairness criteria. First, observe that our scheme closely approximates the full feedback for every α. The best-M and normalized SNR thresholding schemes show significant throughput loss near $\alpha = 0$, while the throughput and fairness approach that of the full feedback as α increases. On the other hand, the absolute SNR thresholding scheme shows the opposite behavior, i.e., its performance is more degraded as α increases. In Fig. 6, we plot the per-user throughput of our proposed and the previous schemes. The data are taken from the same simulation as in Fig. 5. Three representative users with the highest, the median, and the lowest SNR are selected. Fig. 6(a) shows that the individual throughput under our scheme follows that under the full feedback [as we observed in Fig. 5(a)]. However, under the best-M and normalized SNR thresholding schemes in Fig. 6(b) and (c), the strongest user's throughput at $\alpha = 0$ (MRS) is much lower than that under the full feedback. This is because these schemes let the users send equal or approximately equal amounts of feedback, irrespective of channel qualities. Under the absolute SNR thresholding scheme in Fig. 6(d), the throughput of the median and weakest users is not raised, even though the fairness exponent α increases. This is because such users are not allowed to send an enough amount of feedback under the absolute SNR thresholding scheme, regardless of the fairness criterion of the scheduler.

Fig. 7 depicts the total throughput, the fairness index, and the total feedback load per channel for various numbers of users, where the PFS is used. In Fig. 7(a)–(b), our scheme shows almost identical performance to that of the full feedback. However, the best-M and normalized SNR thresholding schemes experience considerable throughput degradation when there are a small number of users. On the other hand, the total throughput of the absolute SNR thresholding scheme outperforms the full feedback, whereas the fairness performance is worse than that of the full feedback. This is because the weak users have no chance of being scheduled at all due to the lack of the CQI information, and consequently, the resulting throughput

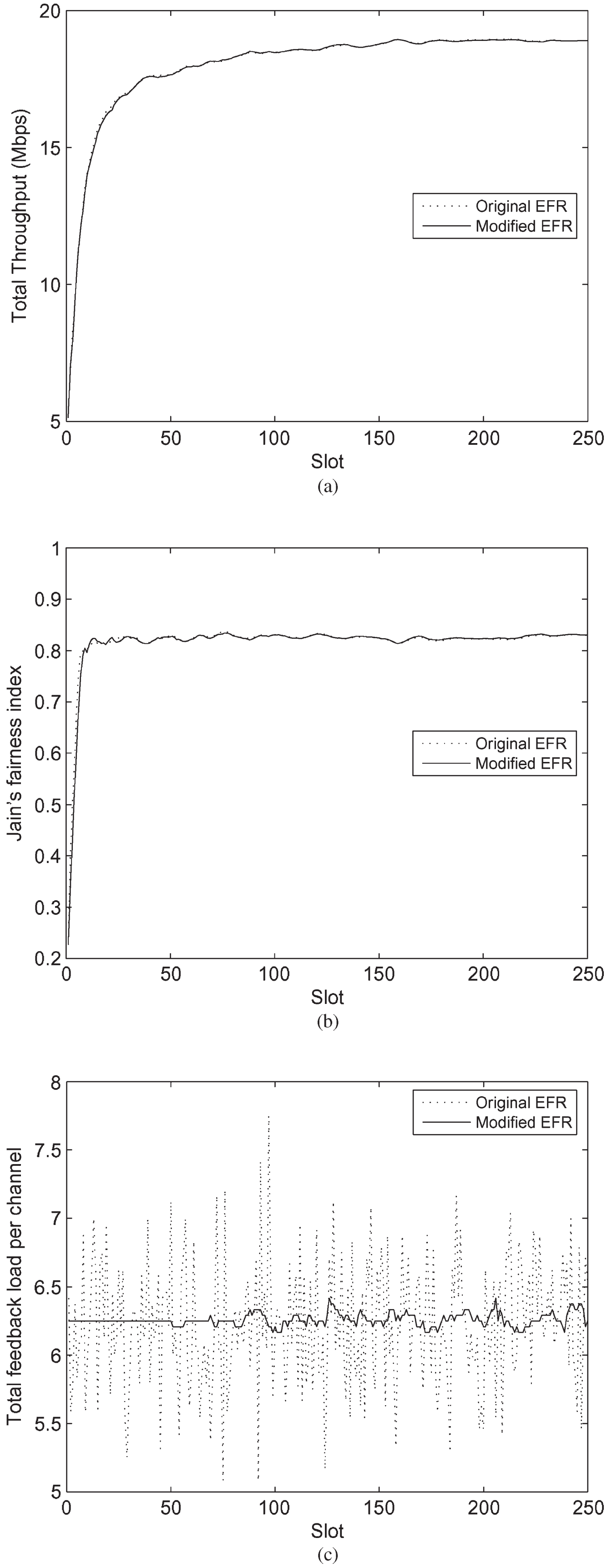


(a)

(b)

(c)

Fig. 7. Comparison of total throughput, Jain's fairness index, and total feedback load per channel for various number of users. (a) Total throughput with PFS (i.e., $\alpha = 1$). (b) Jain's fairness index with PFS (i.e., $\alpha = 1$). (c) Total feedback load per channel with PFS (i.e., $\alpha = 1$).



(a)

(b)

(c)

Fig. 8. Comparison of Algorithm 1 (original EFR) with Algorithm 2 (modified EFR). (a) Total throughput with PFS (number of users $= 30$). (b) Jain's fairness index with PFS (number of users $= 30$). (c) Total feedback load per channel with PFS (number of users $= 30$).

vector is quite distant from the proportionally fair throughput allocation [see Figs. 5 and 6(d) at $\alpha = 1$]. From Fig. 7(c), it can be seen that our scheme controls the per-channel feedback load below $1/e$, whereas other schemes exhibit increasing feedback load as the number of users increases.

In Fig. 8, we compare the performance of the modified efficiency-based feedback scheme in Algorithm 2 with the original one in Algorithm 1. In Fig. 8(a) and (b), the modified scheme exhibits almost identical total throughput and fairness performance to those of the original scheme. However, from Fig. 8(c), it can be seen that the original scheme induces substantial variation in feedback load, whereas the feedback load fluctuates much less under the modified scheme, which follows from the construction of the modified scheme in Section IV-D.

VI. Concluding Remarks

In this paper, we have proposed an innovative feedback-reduction algorithm that uses feedback efficiency as a feedback-decision metric instead of the received SNR used in the previous schemes. The key idea behind our algorithm is to give more feedback opportunity to the users who are more often scheduled. As a preliminary step, we have first studied the impact of feedback reduction on the performance of the scheduler by adopting the viewpoint of the gradient-based scheduling theory. We have then proposed and analyzed the efficiency-based feedback algorithm that exhibits almost the same scheduling performance as in the case of full feedback while substantially reducing the amount of feedback. The simulation results have demonstrated that our scheme works well with a broad class of scheduling policies, ranging from the MRS aiming at maximum efficiency to the MFS aiming at maximum fairness. Moreover, it can control the total feedback load below a target level, regardless of the number of users. All these advantages are achieved without additional control overhead.

Appendix
Proof of Observation 1

Let $[\theta'_k, \forall k \in \mathcal{K}]$ be the long-term average ratio of time slots allocated to users by scheduling policy (3), where $0 \leq \theta'_k \leq 1, \forall k \in \mathcal{K}$, and $\sum_{k\in\mathcal{K}} \theta'_k = 1$. Consequently, we can express the resulting long-term throughput vector as $\mathbf{R}' = [\theta'_k R_k^{\max}, \forall k \in \mathcal{K}]$. Similarly, an arbitrary long-term throughput vector by an arbitrary allocation ratio $[\theta_k, \forall k \in \mathcal{K}]$ can be expressed as $\mathbf{R} = [\theta_k R_k^{\max}, \forall k \in \mathcal{K}]$, where $0 \leq \theta_k \leq 1, \forall k \in \mathcal{K}$, and $\sum_{k\in\mathcal{K}} \theta_k = 1$. Because the utility functions in (2) are strictly concave, it is easy to show that $\mathbf{R}^*$ is the optimal solution of the sum utility maximization if and only if (see [10] and the references therein)

$$\nabla U(\mathbf{R}^*)^T \cdot (\mathbf{R} - \mathbf{R}^*) \leq 0 \qquad \forall \mathbf{R} \in \Gamma. \tag{11}$$

With the strictly convex long-term feasible region, it follows that $\mathbf{R}^*$ is the unique maximizer and $\mathbf{R}_t \rightarrow \mathbf{R}^*$ as $t \rightarrow \infty$, starting with any initial state $\mathbf{R}_0 \in \Gamma$. Therefore, the long-term throughput vector $\mathbf{R}'$ should satisfy (11) for all possible throughput vectors $\mathbf{R} \in \Gamma$, because $\mathbf{R}'$ is the unique optimal long-term throughput vector obtained by the gradient-based scheduler in (3). For the two-user case, (11) can be rewritten as follows:

$$\frac{\theta_1 R_1^{\max} - \theta_1' R_1^{\max}}{(\theta_1' R_1^{\max})^{\alpha}} + \frac{(1-\theta_1) R_2^{\max} - (1-\theta_1')\, R_2^{\max}}{((1-\theta_1')\, R_2^{\max})^{\alpha}} \le 0 \tag{12}$$

for all $\theta_k \in [0,1]$. After some manipulation, (12) reduces to

$$(\theta_1 - \theta_1')\, A \le 0 \tag{13}$$

where $A = (1-\theta_1')^{\alpha} (R_2^{\max})^{\alpha-1} - \theta_1'^{\alpha} (R_1^{\max})^{\alpha-1}$. The inequality in (13) is always satisfied, for all $\theta_k \in [0,1]$, if and only if $A = 0$, and this leads us to the result of Observation 1.

References

[1] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE ICC*, Seattle, WA, Jun. 1995, pp. 331–335.

[2] 3GPP Tdoc R1-051045 CQI Report and Scheduling Procedure, 2005.

[3] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[4] J. G. Andrews, A. Ghosh, and R. Muhamed, *Fundamentals of WiMAX: Understanding Broadband Wireless Networking*. Englewood Cliffs, NJ: Prentice-Hall, 2007.

[5] S. Shakkottai and A. L. Stolyar, "A study of scheduling algorithms for a mixture of real and non-real time data in HDR," Bell Labs., Lucent Technol., Murray Hill, NJ, Oct. 2000.

[6] D. Gesbert and M.-S. Alouini, "How much feedback is multi-user diversity really worth?" in *Proc. IEEE ICC*, Paris, France, Jun. 2004, pp. 234–238.

[7] L. Yang, M.-S. Alouini, and D. Gesbert, "Further results on selective multiuser diversity," in *Proc. ACM MSWiM*, Venice, Italy, Oct. 2004, pp. 25–30.

[8] T. Tang and R. W. Heath, Jr., "Opportunistic feedback for downlink multiuser diversity," *IEEE Commun. Lett.*, vol. 9, no. 10, pp. 948–950, Oct. 2005.

[9] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Comput. Netw.*, vol. 41, no. 4, pp. 451–474, Mar. 2003.

[10] R. Agrawal and V. Subramanian, "Optimality of certain channel aware scheduling policies," in *Proc. Allerton Conf. Commun.*, Oct. 2002, pp. 1532–1541.

[11] J. Huang, V. Subramanian, R. Agrawal, and R. Berry, "Downlink scheduling and resource allocation for OFDM systems," in *Proc. Conf. Inf. Sci. Syst.*, Princeton, NJ, Mar. 2006, pp. 1272–1279.

[12] H. J. Kushner and P. A. Whiting, "Convergence of proportional-fair sharing algorithms under general conditions," *IEEE Trans. Wireless Commun.*, vol. 3, no. 4, pp. 1250–1259, Jul. 2004.

[13] R. Agrawal, A. Bedekar, R. La, and V. Subramanian, "A class and channel-condition based weighted proportional fair scheduler," in *Proc. ITC*, Salvador, Brazil, Sep. 2001.

[14] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Trans. Netw.*, vol. 8, no. 5, pp. 556–567, Oct. 2000.

[15] F. Kelly, "Charging and rate control for elastic traffic," *Eur. Trans. Telecommun.*, vol. 8, no. 1, pp. 33–37, Jan. 1997.

[16] S. Borst, "User-level performance of channel-aware scheduling algorithm in wireless data networks," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar. 2003, pp. 636–647.

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

[18] Q. Zhang and S. A. Kassam, "Finite-state Markov model for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[19] Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems, IEEE Std. 802.16e-2005 and IEEE Std. 802.16-2004/Cor 1-2005, Dec. 2005.

[20] Guidelines for the Evaluation of Radio Transmission Technologies for IMT-2000, Recommendation ITU-R M.1225, 1997.

[21] M. Pätzold, *Mobile Fading Channels*. Baffins Lane, U.K.: Wiley, 2002.

[22] R. Jain, D. Chiu, and W. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared system," DEC, Maynard, MA, TR-301, Sep. 1984.

Jeongho Jeon (S'06) received the B.S. degree (*summa cum laude*) in electronic engineering from Sogang University, Seoul, Korea, in 2006 and the M.S. degree in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2008. He is currently working toward the Ph.D. degree with the Department of Electrical and Computer Engineering, University of Maryland, College Park.

Since 2008, he has been a graduate research assistant with the Institute for Systems Research, University of Maryland. His current research interests are in the areas of communication theory and its applications.

Kyuho Son (S'03) received the B.S. and M.S. degrees in electrical engineering and computer science in 2002 and 2004, respectively, from Korea Advanced Institute of Science and Technology, Daejeon, Korea, where he is currently working toward the Ph.D. degree with the School of Electrical Engineering and Computer Science.

His current research interests include interference management and load balancing in multicell networks, as well as spectrum sharing in cognitive radio networks.

Mr. Son is the Web Chair of the 2009 Seventh International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks.

Hyang-Won Lee (M'07) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Korea, in 2001, 2003 and 2007, respectively.

He is currently a Postdoctoral Research Associate with the Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge. His research interests are in the areas of wireless networks and optical networks.

Song Chong (M'93) received the B.S. and M.S. degrees in control and instrumentation engineering from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively, and the Ph.D. degree in electrical and computer engineering from the University of Texas at Austin in 1995.

Since March 2000, he has been with the School of Electrical Engineering and Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, where he is a Professor and the Director of the Communications and Computing Group. Prior to joining KAIST, he was with the Performance Analysis Department, AT&T Bell Laboratories, Holmdel, NJ, as a Member of Technical Staff. He is the author of more than 70 papers published in international journals and conference proceedings. He is the holder three U.S. patents. He is an Editor of the *Journal of Communications and Networks*. His current research interests include wireless networks, future Internet, human mobility, and performance evaluation.

Dr. Chong is currently the Chair of Wireless Working Group of the Future Internet Forum and the Vice President of Information and Communication Society of Korea. He is the General Chair of the 2009 Seventh International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks. He has served on the Technical Program Committees of a number of key international conferences, including the IEEE Conference on Computer Communications, the ACM International Conference on Emerging Networking Experiments and Technologies, the International Test Conference, etc.