Proceedings of APCC2008 copyright © 2008 IEICE 08 SB 0083

Opportunistic Packet Scheduling Algorithm for Load Balancing in a Multi-hop Relay-enhanced Cellular OFDMA-TDD System

Tae W. Kim, Tae-Young Min, and Chung G. Kang
Department of Electrical Engineering, Korea University
E-mail: ccgkang@ korea.ac.kr

***Abstract* – In this paper, we consider a multi-hop relay-enhanced cellular OFDMA-TDD system with the full frequency-reuse capability, in which a TDD frame can be asymmetrically divided into two different intervals, one for access link to mobile stations and the other for relay link to base station (BS)-relay station (RS) communication, while the same radio resource is fully reused by every RS in the cell. Since a single common boundary between access and relay links is employed for the system, some access link associated with an individual RS can be either overloaded or under-loaded when traffic load is non-uniformly distributed, which causes an inefficient resource allocation. This paper proposes a load-balancing opportunistic (LoBO) scheduling algorithm that improves the overall system throughput in a weighted proportional fairness manner while balancing the traffic load over the access link to be shared by all RS's. The proposed algorithm dynamically determines the common access-to-relay interval boundary as a part of packet scheduling, which has been shown to outperform the conventional system in which the boundary selection and packet scheduling are implemented as the separate processes.**

I. INTRODUCTION

Multi-hop relay systems are considered as a useful means of enhancing coverage, throughput and capacity of the mobile wireless broadband, e.g., IEEE 802.16e mobile wireless MAN [1]. Advantages of coverage expansion and throughput enhancement can be leveraged to reduce total deployment cost for a given system performance requirement and thereby improve the economic viability of those systems. IEEE 802.16j multi-hop relay (MR) task group is one particular example of standardization activities towards relay-enhance cellular system (RECS), which aims to enable exploitation of such advantages by specifying OFDMA physical layer and medium access control layer enhancements to IEEE Std 802.16 for licensed bands to enable the operation of relay stations [2][3].

As compared with the conventional optical repeaters in the cellular system, benefits of introducing the MR into the field include an easy network deployment and a significant reduction in an infrastructure cost by replacing the wireline relay link with the wireless hops using the same frequency assignment (FA) as the access link [3]. Furthermore, throughput and capacity enhancement can be achieved through the frequency-reuse capabilities of relay stations (RS's). In this paper, we consider a multi-hop relay-enhanced cellular system with the full frequency-reuse capability, in which a TDD frame can be asymmetrically divided into two different intervals, one for access link to mobile stations and the other for relay link to BS-RS communication. The relay link is orthogonally shared among all RS's while the access link is fully reused independently by every RS. Since a single common boundary between access and relay links is employed for the system, radio resource in some access link might be wasted, incurring an inefficiency of overall resource allocation, unless the traffic load of mobile stations is uniformly distributed among the RS's. In order to deal with the underlying inefficiency, there must be some means of load-balancing among all RS's while dynamically configuring the optimal boundary between access and relay links.

[1] This work was partly supported by the IT R&D program of MKE/IITA [2008-F-015-01, Research on Ubiquitous Mobility Management Methods for Higher Service Availability].

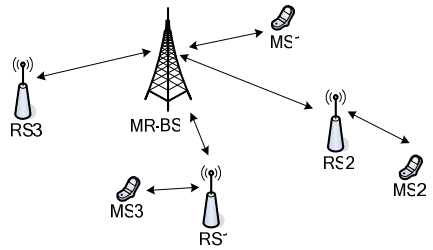

Figure 1. Multi-hop Relay-enhanced System: Illustration

In this paper, we propose a load-balancing opportunistic (LoBO) scheduling algorithm, which allows for dynamically selecting a set of users to be served in each RS while maintaining the best common boundary of access and relay link at the same time. Its design objective is to improve the system throughput in a weighted proportional fairness manner while balancing the load over the access link to be shared by all RS's. The proposed algorithm dynamically determines the access-to-relay interval ratio as a part of packet scheduling. The simulation result for IEEE 802.16j-based cellular OFDMA-TDD system shows that LoBO scheduling improves the overall system throughput approximately by 30% over a conventional approach with a proportional fairness scheduling algorithm without any load-balancing capability.

II. MR SYSTEM ARCHITECTURE AND MODELING

A. Multi-hop Relay (MR) System: Overview

Fig. 1 illustrates a typical cellular system that employs the layer-2 (L2) relay station (RS). Mobile station (MS) can communicate with base station (MR-BS) either directly or over 2-hop via RS. Even though more than two hops can be exploited via multiple RS's, we only consider a simple 2-hop scenario via a single RS in the current analysis. The direct link between MS and MR-BS is referred to as an access link while the link between RS and MR-BS is referred to as a relay link. The L2 RS works as a half-duplexing relay, in which a signal from MR-BS is decoded first and then forwarded toward MS, vice versa. As opposed to the amplify-and-forward (AF) type of relay, a.k.a. L1 relay, adaptive modulation and coding (AMC) can be applicable to L2 RS by taking the link condition between RS and MS, which allows for fully exploiting the bandwidth by opportunistic scheduling among

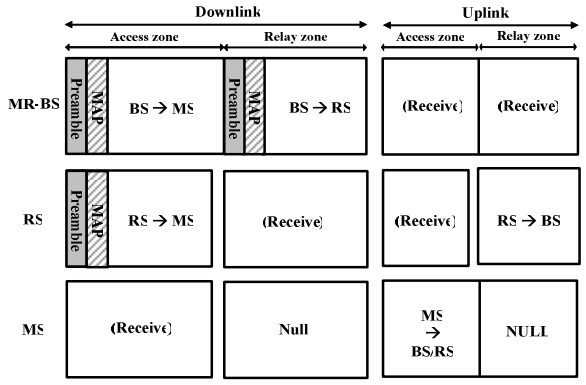

Figure 2. MR Frame Structure: Non-transparent/Overlapped Allocation

multiple users. Furthermore, bandwidth assignment can be controlled centrally by a packet scheduler in the MR-BS. Bandwidth assignment information on OFDMA data region is known to all RS's and MS's via a MAC management message, e.g., MAP message in IEEE 802.16 standard and thus, each RS can selectively serve the MS's located only within its own coverage, which allows for reducing co-channel interference from neighbor cells. In the current analysis, we assume a non-transparent transmission mode, in which all MS's served by RS can receive the MAP message only via RS, not directly from MR-BS.

We consider a frame structure harmonized in IEEE 802.16j [4], one especially designed for a two-hop scenario with a single FA. As illustrated in Fig. 2, each TDD frame is divided into downlink and uplink intervals, each of which is further divided into access link interval for MS (i.e., used for BS-to-MS and RS-to-MS links) and relay link interval for RS (i.e., used for BS-to-RS link). One main characteristics of this particular frame structure is that access link interval is shared with all MS's that are communicating directly with BS or indirectly via RS's. For a non-transparent case in which some MS's might not be able to directly receive the MAP message, bandwidth allocation results must be also relayed. Therefore, two different MAPs for MS must be transmitted in the downlink, one for MS that is directly served by BS, and the other for MS that is served by RS. Furthermore, there must be another MAP for the relay link, so as to indicate which OFDMA data regions are dedicated to individual RS.

B. Resource Allocation Schemes

In this paper, we consider a downlink for the Wireless Broadband (WiBro) system, which is a mobile version of WiMAX derived from the IEEE 802.16e standard [1]. It is the OFDMA/TDD system with 768 useful subcarriers over a nominal bandwidth of 8.75MHz at 2.3 GHz band, designated for mobile broadband Internet services in Korea. Each frame is composed of 42 OFDM symbols, corresponding to 5ms. For the asymmetric characteristics of typical internet traffic, we assume a downlink/uplink ratio of 2:1, i.e., 24 symbols for downlink subframe and 12 symbols for uplink subframe, with the rest of symbols used for preamble and control information. A basic resource allocation unit is given by a subchannel, defined as a set of 48 subcarriers selected either by a diversity mode or a band AMC mode. Note that there are 384 subchannels available for each frame (e.g., 768 subcarriers/symbol * 1 subchannel/48 subcarriers * 24 symbols/frame = 384 subchannels/frame) in a downlink. The common boundary between access and relay link intervals is fixed, but it must be determined so as to best utilize the overall resources at least in the average sense for the varying traffic load.

Depending on the frequency reusability over the access link intervals of downlink and uplink, we consider two different allocation schemes: the overlapped and orthogonal allocation schemes. The orthogonal allocation scheme corresponds to the case of not reusing the subchannels for the access link interval, i.e., no subchannels can be shared among the MS's that are directly served by BS and those that are served by RS's. Meanwhile, the overlapped allocation scheme corresponds to the case that all subchannels in the access link interval are shared among all MS's throughout the cell coverage, regardless of whether MS is directly served by BS or not. In spite of attempting to maximize the bandwidth efficiency, it tends to suffer from co-channel interference, which reduces the overall system throughput and furthermore, induces the outage events around sub-cell edges as well as a boundary of each cell. Meanwhile, all subchannels of relay link are orthogonally divided for RS's and thus, they are not subject to any co-channel interference. Given the access and relay link intervals, a packet scheduling algorithm is applied to determine which users are served in each interval.

C. Resource Management Issues for MR System

- Access link selection

In the multi-hop relay system, the overall coverage can be divided into two different types, one covered by a BS and the others covered by RS's. If the overall traffic is uniformly loaded throughout a cell, then all available resources can be fairly shared among BS and RS's by centralized scheduling. Otherwise, however, resource demands for BS and individual RS might be dynamically varying on the traffic distribution. For the MR system subject to overlapped allocation, in which BS and individual RS reuse the same resource for access link, a non-uniform traffic distribution may incur inefficient resource utilization. This particular defect can be resolved by load balancing among BS and RS's with some centralized control schemes. In fact, it is equivalent to the access link selection problem that deals with which MS is served by BS or RS. It is obvious that the access link selection and scheduling problems must be jointly considered, since both affect utilization of the same resource. In general, however, the joint optimization of link selection and scheduling is prohibitively complicated. Therefore, one problem is separated from the other in this paper, simply because we want to focus just on the short-term resource management in the sense that the link selection problem deals with a rather long-term behavior of traffic load variation. In other words, we assume that the access link for each MS is already set up by some long-term load balancing scheme and thus, we only focus on packet scheduling on a rather short-term basis.

- Boundary selection

Given the traffic load and channel conditions, a boundary between relay and access intervals is supposed to be determined by resource allocation subject packet scheduling. Depending on the traffic distribution, it might suffer from resource scarcity or waste for the given boundary. In the course of packet scheduling, therefore, there must be some means of selecting the boundary in association with load balancing, which deals with resource allocation for both access and relay links.

Fig. 3 illustrates one particular example of determining the dynamic boundary for the downlink. Herein, we define two different makers, left marker (LM) and right marker (RM). The LM represents a marker to indicate a size of access interval for BS and RS, denoting its size by $S_k(t)$ for node k in a frame t (k = 0 reserved for BS). In a similar manner, RM is used to indicate a size of relay interval, denoting its size by $S_r(t)$. Since all RS's are sharing the same relay link in an orthogonal manner, only a single relay interval with a length of $S_r(t)$ is given, i.e., one RM for the system. For the access link, meanwhile, BS or each RS is associated with the different LM's, i.e., each of them with a length

of $S_k(t)$ for RS k. Assuming that there are K RS's, let us denote the k-th RS by RS_k, k = 1, 2, …., K. For the given ($S_r(t)$, $S_k(t)$), however, only a single common boundary between access and relay intervals is set system-wide. If the length of downlink frame is given by T_{frame} , LM and RM move to the right and to the left, respectively, until the following relation is satisfied while performing resource allocation:

$$S_r(t) + \max_{k \in \{0,1,2,\cdots,K\}} \{S_k(t)\} \le T_{frame} \tag{1}$$

and the boundary is set at the point where one of LM meets with RM. In other words, the boundary is dynamically configured by depending on whether the packet scheduler allocates the current subchannel to BS or one of RS's. Therefore, an amount of resources available to each link is governed by the boundary, which subsequently limits amount of the resource available to each MS.

III. LOAD-BALANCING OPPORTUNISTIC (LOBO) SCHEDULING FOR MR SYSTEM

In order to maximize the bandwidth efficiency in the multi-hop relay system, load balancing and packet scheduling must be jointly performed as discussed in the previous section. We first consider the corresponding optimization framework. Under the proportional fairness (PF) packet scheduling algorithm, for example, the boundary and link selection problem to maximize the average system throughput can be formulated as follows:

$$\max_{\left(\beta, x_n^{(k)}\right)} \left\{ \sum_{i=1}^{N_c} \sum_{k=0}^{K} \sum_{n=1}^{N} \bar{R}_{i,n}^{(k)}\left(\beta, I_{oc}\right) \cdot x_n^{(k)} \right\} \tag{2}$$

where

$\bar{R}_{i,n}^{(k)}$: Average throughput of MS n served by node k in the cell i

$\beta = T_{access} / T_{relay}$: Ratio of access to relay intervals

$$x_n^{(k)} = \begin{cases} 1, & n \in U^{(k)} \\ 0, & n \notin U^{(k)} \end{cases}$$

$U^{(k)}$: A set of users to be by BS (k = 0) and RS k (k = 1, 2,.., K)

N : The number of MS's in the cell

N_c : The number of cells

I_{oc} : Other cell interference

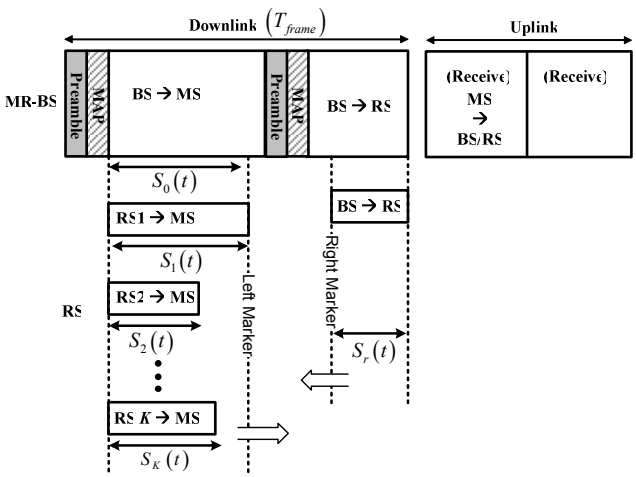

Figure 3. Boundary for Access and Relay Link of OFDMA-TDD Multi-hop Relay System

Note that $x_n^{(k)}$ is a binary decision variable for access link selection and β is the variable to represent the access-to-relay interval ratio in (2). It is obvious that the average system throughput is a function of $x_n^{(k)}$ and β , which is subsequently governed by which user is served by the packet scheduling algorithm. In fact, a complexity of solution approach to (2) is prohibitively too high, since many different decision variables are related to each other in a compound manner.

The access link selection problem is to determine whether an MS is served by either BS or RS. In general, a link with the best CINR is selected. Since it may be unrealistic to select the best link in each frame, the access link selection problem can be considered as a part of long-term resource scheduling. In this paper, we focus only on a short-term scheduling problem which deals with the dynamic resource allocation in each frame. In other words, we assume that the access link is already set up by some other means of scheduling algorithm, which is beyond the current discussion. In the problem formulation of (2), therefore, { $x_n^{(k)}$ } are not the variables anymore and we are interested in determining β only. In fact, β can be varied in each frame, which will depend on the resource allocation by the packet scheduling algorithm. It implies that β must be determined by the packet scheduling algorithm that maximizes the network utility. For the overlapped allocation as in Fig. 2, some relay links can waste the resource, as each of relay coverage is scheduled individually while fixing the common boundary.

Therefore, the boundary variable β must be determined so as not to waste the overall resource throughout the cell. Unless the traffic is uniformly loaded for every RS, the common boundary will be bounded by the most overloaded RS, while underutilizing the resource of access link for the relatively under-loaded RS. The over-specified access link due to the imbalanced load among the RSs in the current frame reduces the resource available for relay link, which will influence on the boundary in the following frame. In other words, a waste of the access interval subsequently limits available resource for the relay interval, which implies that some means of scheduling must be designed so as to balance the load over both intervals.

In the practical system, meanwhile, we note that the access link data packets to be served by BS and those by RS are allocated in the different frames. Referring to Fig. 4, once a BS-MS access intervals and relay intervals for the frame t are determined by a packet scheduling algorithm, resource allocation for the corresponding relay interval subsequently determines RS-MS access interval in the frame $(t+\tau)$ where τ represents a system delay that is required to process the channel allocation information notified by BS. In the frame t, therefore, an access link resource waste for MS n associated with BS is measured by the access interval given in this particular frame while resource allocation for individual RS-MS interval in this frame already has been set in some earlier frame.

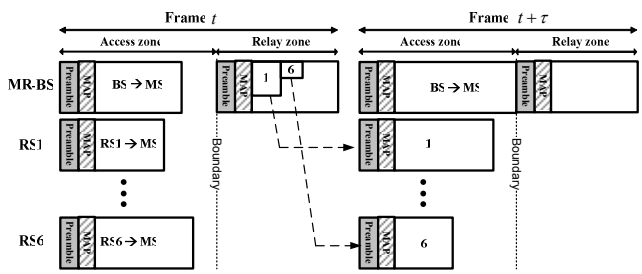

Figure 4. Illustration for Resource Allocation

Meanwhile, in case that MS n is served by RS in the frame t, we have to consider a corresponding resource waste in the frame $(t+\tau)$. When an MS is selected by a scheduler, therefore, two different cases must be separately considered, depending on whether it is served by BS or RS. Let $G_{BS}(t)$ and $G_{RS_k}(t)$ denote a group of MS's that are served by BS and RS k in the frame t, respectively. Furthermore, let $S_{RS_k}(t)$ and $S_{BS}(t)$ denote the numbers of subchannels assigned for RS k and BS, respectively, as an access link. If the number of subchannels used by some RS is less than the total number of subchannels used by RS or BS that serves MS n, then a sum of their difference represents overall resource waste, which will be referred to as the *resource allocation gap* (RAG) in this paper. Denoting the RAG for each MS n in the frame t by $W_n(t)$, it is given as follows:

If $n \in G_{BS}(t)$,

$$W_n(t) = \begin{cases} \sum_{k=1}^{K} \max\left\{0, S_{BS}(t) - S_{RS_k}(t)\right\}, & S_{BS}(t) \geq \max_{k\in\{1,2,\cdots,K\}} S_{RS_k}(t) \\ 0, \text{ otherwise} \end{cases} \tag{3}$$

If $n \in G_{RS_k}(t)$,

$$W_n(t) = \sum_{k=1, k\neq m}^{K} \max\left\{0, S_{RS_m}(t+\tau) - S_{RS_k}(t+\tau)\right\} \tag{4}$$

In order to apply an opportunistic packet scheduling algorithm to the two-hop transmission, meanwhile, the end-to-end instantaneous channel condition must be taken into account by concatenating the different channel condition of each hop at the same time. Toward this end, let us define the end-to-end effective transmission rate for MS n which is served by RS k, denoting it by $\overline{r}_{k,n}$, as follows:

$$\overline{r}_{k,n} = \left(\frac{1}{r_{k,n}} + \frac{1}{r_k}\right)^{-1} \tag{5}$$

where $r_{k,n}$ and r_k denote the instantaneous transmission rate for the access link between RS k and MS n, and that for the relay link for RS k, respectively. In order to maximize the network utilization, we consider a problem that maximizes a sum of utilization for an individual user. For a long-term average transmission rate of MS n, denote by R_n, let us denote the corresponding utility function by $U(R_n)$. As utilization of individual user may be reduced by resource waste in the course of packet scheduling for each MS, we define the overall network utility function as a weighted sum of individual utility, where each user n is associated with a weighting factor α_n which summarizes an inefficiency of individual user, as follows:

$$\sum_{n=1}^{N} \alpha_n U(R_n) \tag{6}$$

The network utility function (6) takes the inefficiency into account by weighting the utility of individual user in terms of the corresponding RAG, i.e., $\alpha_n = 1/W_n$ In other words, resource allocation to one with the less RAG contributes more to the overall utility. Invoking a notion of proportional fairness among all users, we set the utility function to $U(R_n) = \log R_n$. It leads to an optimal scheduling scheme with respect to (6), selecting an MS with the following criterion in each schedule interval:

$$i^* = \arg\max_n \frac{\overline{r}_{k,n}}{W_n R_n} \tag{7}$$

The packet scheduler given by (7) implies that each MS will be given the different access priority depending on both the instantaneous channel condition and the RAG for each MS in each instance of scheduling. Intuitively, higher priority will be given to the MS with the least RAG. As this particular approach allows for determining the boundary that reduces the resource waste caused by a non-uniform distribution of traffic load, it is referred to as load-balancing opportunistic (LoBO) scheduling.

IV. Simulation Results and Performance Analysis

Each site is modeled by an ideal hexagonal cell and only two tiers of cells are considered, i.e., a total of 19 cells, with respect to a reference cell in the center. Omni-directional antenna is used in each cell. Due to the finite number of cells, accurate level of interference from all other cells cannot be captured in the model. In order to remove such a boundary effect, we consider a so-called *wrap-around* structure, which allows for capturing a more accurate level of inter-cell interference. MS's are uniformly distributed throughout each cell and move at a velocity of 3km/h along a direction randomly selected in each frame. Furthermore, a path loss is given by the WINNER model while a log-normal shadow fading model with a standard deviation of 8dB is considered for large-scale fading. Meanwhile, a multi-path fading model follows ITU-R M.1225 recommendation for pedestrian A (PED-A). An individual multi-path is subject to the independent Rayleigh fading, whose time-domain correlation is implemented by Jake's model.

We assume that 100 mobile users are uniformly distributed throughout each cell. We deploy 6 relays in each cell, located around a BS at a 2/3 position between BS and cell boundary. The transmit power of BS and RS is limited to 20W and 10W, respectively. Meanwhile, we consider two different types of traffic models: full buffer and Ethernet traffic models. For the full buffer traffic model, we assume that each MS always has packets to transmit in the buffer. Ethernet traffic model deals with a simple Internet traffic that is characterized by empirical data of Ethernet traffic. The details of Ethernet traffic model can be referred to [5].

Table 1 is the MCS table used in simulation, which shows CINR levels required for the given modulation and coding set (MCS) subject to the given channel model. In the current simulation, a group of 10 subchannels is considered as a basic unit of resource allocation for packet scheduling. The system processing delay is set to $\tau = 1$ frame.

In the current simulation, we investigate the system throughput, resource efficiency, and fairness for the traffic models of full buffer and Ethernet, respectively, as varying the number of users. The performance of our proposed LoBO scheduling scheme is compared with that for PF scheduling with the fixed boundary, which is determined so as to maximize the average system throughput. Meanwhile, fairness among the users is measured in terms of the following performance measure [6]:

$$F \triangleq 10 \cdot \log\left(\frac{\mu}{\sigma}\right) = 10 \cdot \log\left(\frac{\frac{1}{N}\sum_{n=1}^{N} R_n}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(R_n - \frac{1}{N}\sum_{n=1}^{N} R_n\right)^2}}\right) \tag{8}$$

Table 1 MCS Table for Adaptive Modulation & Coding

Downlink MCS		Required CINR (dB)
QPSK	1/12	-3.46
QPSK	1/6	-1.0
QPSK	1/3	1.73
QPSK	1/2	5.40
16-QAM	1/2	10.5
64-QAM	1/2	15.0
64-QAM	2/3	20.0
64-QAM	5/6	28.5

In Figure 5, it is shown that the system throughput performance is improved with LoBO scheduling approximately by up to 31% and 21% for the full buffer and Ethernet traffic models, respectively. It is obvious that the performance gain is obtained by ensuring the efficiency by reducing the resource waste with the dynamic boundary for loading balancing, which deals with any resource waste for the access link of each RS. The corresponding gain becomes clearer from Fig. 6, which shows the resource efficiency of each access link associated with individual node for the different schemes with the different traffic models. It is observed that most of the access links for relay stations are underutilized for PF scheduling with the fixed boundary.

In Fig. 7, overall system fairness is compared in terms of the performance measure defined by (8). For both traffic models, it is obvious that the LoBO scheduling scheme does not much compromise its system fairness performance even with the throughput gain observed in Fig. 5.

V. CONCLUSION

In a multi-hop relay-enhanced cellular OFDMA-TDD system with the full frequency-reuse capability, i.e., sharing the same access interval among BS and all RS's, a group of users to be served by each RS is scheduled while simultaneously reconfiguring the boundary between access and relay links, so as to balance the non-uniform traffic load distribution. The proposed load-balancing opportunistic (LoBO) scheduling scheme is based on a measure of resource allocation gap as a weight factor that controls the opportunistic scheduling priority for individual user. It intends to provide a weighted proportional fairness among all users while reducing the resource allocation inefficiency caused by non-uniform traffic load distribution. The simulation result for IEEE 802.16j-based cellular OFDMA system shows that LoBO scheduling improves the overall system throughput approximately by 30% over a conventional approach with a proportional fairness scheduling algorithm with the fixed common boundary. In the current analysis, a link selection problem to determine whether an MS is served by BS or RS has not been considered together with scheduling and boundary selection problems. In the future, therefore, a more thorough optimization is required by considering a joint control problem.

REFERENCES

[1] "Definition of terminology used in Mobile Multihop Relay," IEEE 802.16j MMR Contribution, May, 2006.

[2] J.W. Cho and Z.J. Haas, "On the Throughput Enhancement of the Downstream Channel in Cellular Radio Networks Through Multihop Relaying," IEEE Journal on Selected Areas in Communications, vol, 22, no. 7, pp. 1206-1219, September 2004.

[3] K. Park, C.G. Kang, D. Chang, S. Song, J. Ahn, and J. Ihm, "Relay-enhanced Cellular Performance of OFDMA TDD System for Mobile Wireless Broadband Services," ICCCN 2007, August 2007.

[4] IEEE 802.16j Baseline document "IEEE 802.16j-06/026r3". IEEE 802.16 Relay TG, May, 2007.

[5] Chung G. Kang, Jin M. Ku, Pil K. Kim, Se J. Lee, and Simon Shin, "On the Performance of Broadband Mobile Internet Access System," Wireless Personal Communications, February 2008.

[6] S.K. Lee, I.-S. Cho, J.W. Cho, Y.W. So, and D.Y. Hong, "CDMA Bunched Systems for Improving Fairness Performance of the Packet Data Services," LNCS, pp.94-102, Oct. 2006

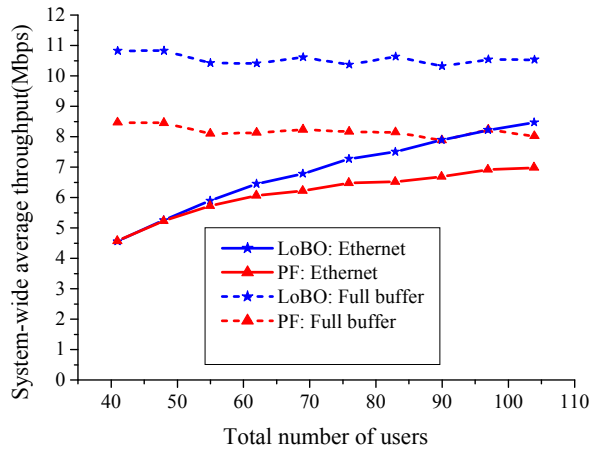

Figure 5. System Throughput for Varying the Number of Users

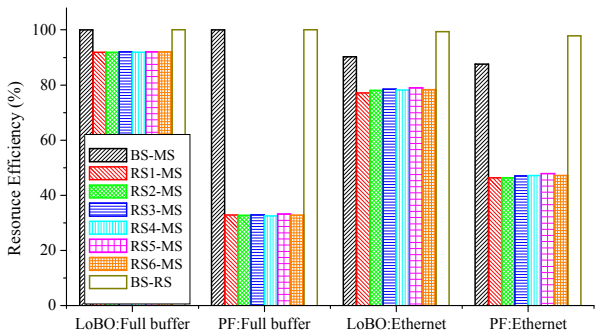

Figure 6. Resource Efficiency for Individual Access Link

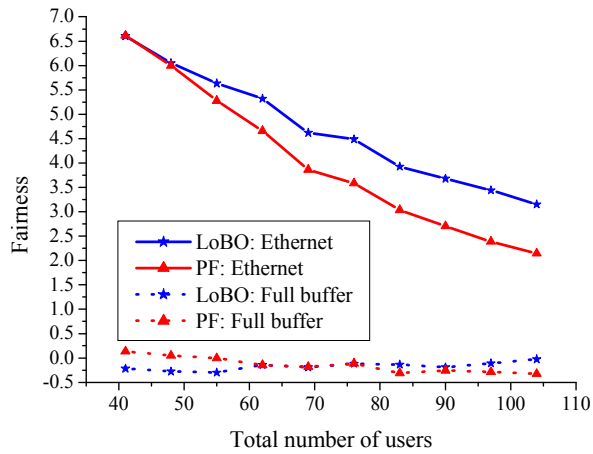

Figure 7. Overall System Fairness for the Varying Number of Users