A slot allocation mechanism for diverse QoS types in OFDMA based IEEE 802.16e systems

Harsha Gowda *Ramya Lakshmaiah* *Manjot Kaur* *Chandrashekar Mohanram*
Manjeet Singh *Shashidhara Dongre*
Communications Strategic Business Unit
Larsen and Toubro Infotech Limited
Bangalore, India.
Phone: 91-80-66242424, Fax: 91-80-28413555
E-mail: {harsha.gowda, ramya.lakshmaiah, manjot.kaur, chandrashekar.mohanram, manjeet.singh, shashidhara.dongre}@lntinfotech.com

***Abstract*— In this paper, we investigate slot allocation for diverse Quality of Service (QoS) types in Orthogonal Frequency Division Multiple Access (OFDMA) based IEEE 802.16e WirelessMAN systems. For a downlink scenario, we propose flow metrics that dynamically capture the extent to which flows of diverse QoS types merit bandwidth allocation. Using these flow metrics, we propose a scheduling rule for a mixture of real-time, non real-time and best effort traffic. Our system model uses queues implemented as finite length buffers for each of the downlink flows. Simulation results are presented to compare the performance of the proposed rule with that of existing rules.**

***Index Terms*— Slot allocation, MLWDF, OFDMA, IEEE 802.16e WirelessMAN, WiMAX, QoS, cross-layer optimization.**

ba

I. Introduction

Orthogonal Frequency Division Multiple Access (OFDMA) is one of the physical layer transmission techniques in the IEEE 802.16e WirelessMAN standard [1], [2]. The IEEE 802.16e WirelessMAN standard specifies the mechanism for broadband wireless access (BWA) for fixed and mobile subscribers at vehicular speeds of up to 100 km/hr.

The standard aims to provide voice and packet data services to users while meeting the latency and throughput requirements of flows. Traffic can be broadly classified into real-time (RT), non real-time (NRT) and best effort (BE) flows at the time of service flow connection setup. Associated with each of the flows is a set of Quality of Service (QoS) parameters: for RT flows, the parameters are a latency requirement and a throughput requirement while for NRT flows, the parameter is a throughput requirement. BE flows have a throughput requirement but are given least priority and are handled on a space availability basis.

Cross-layer optimization techniques have been investigated recently for providing QoS over wireless links [3], [4]. Utility functions based on the mean packet waiting times of RT flows have been proposed in [4] to optimize Medium Access Control (MAC) layer - Physical (PHY) layer cross-layer performance. In [3], the design and implementation of a simulator based on a cross-layer protocol between MAC and PHY layers in a OFDMA based IEEE 802.16e WirelessMAN system is considered. In this paper, for a downlink scenario, we extend the work in [3], [5] to propose, a) A set of metrics specific to RT, NRT and BE flows to measure the extent to which a downlink flow merits OFDMA slot allocation[1]. b) A scheduling rule based on flow metrics to allocate OFDMA slots to RT, NRT and BE flows in accordance with the flow QoS requirements.

Throughput optimal[2] schemes for a mixture of RT, NRT and BE traffic have been proposed for time slot allocation in Code Division Multiple Access - High Data Rate (CDMA-HDR) systems [5] [6]. The Modified Largest Weighted Delay First (MLWDF) rule [5], originally proposed for CDMA-HDR systems, can be used in OFDMA based IEEE 802.16e WirelessMAN systems to perform slot allocation. Token based queues are assumed for NRT and BE flows where tokens arrive at a constant rate as dictated by the throughput requirements for NRT and BE flows. The MLWDF rule works well for a mixture of RT, NRT and BE flows when infinitely backlogged queues are assumed for NRT and BE flows. Under such an assumption, all bandwidth assigned to an NRT or BE flow will result in throughput for the flow. However, the infinite data assumption will not hold for practical wireless systems where the traffic arrival is bursty and finite buffers are employed. This is because, when token queues are employed, it is possible that an NRT or BE flow is assigned bandwidth even when the flow has no data in its actual queues. In this paper, we overcome this problem by allowing an NRT or BE flow to contend for bandwidth i.e, OFDMA slots, only when its queue is non-empty. This helps minimize the wastage of bandwidth and increases throughput when compared to the MLWDF rule. In addition, the short term traffic arrival and throughput statistics are maintained [7] for each of the flows to overcome the burstiness in traffic.

In this paper, we define a flow metric that dynamically measures the extent to which a flow merits bandwidth allocation.

[1]A slot as defined in the IEEE 802.16e WirelessMAN OFDMA PHY is the least possible unit of bandwidth allocation and is a two-dimensional allocation spanning subchannels in the frequency domain and OFDMA symbol durations in the time domain. The definition of a slot is dependent on the subcarrier permutation and varies depending on whether the mode of operation is the distributed subcarrier permutation or the adjacent subcarrier permutation.

[2]A scheme is said to be throughput optimal in the sense that it makes queues stable if at all it is feasible to do so with any other rule.

Furthermore, we propose a scheduling rule for a mixture of RT, NRT and BE traffic. When all of the bandwidth needs of RT and NRT traffic are met i.e, RT and NRT flows have empty queues and OFDMA slots remain, BE flows are serviced to make use of the remaining bandwidth.

The organization of the remainder of this paper is as follows. In Section II, we introduce the system model. In Section III, we briefly discuss the MLWDF scheduling rule. The flow metric definition for RT, NRT and BE traffic and the scheduling rule for a mixture of RT, NRT and BE flows are discussed in Section IV. Simulation results are presented in Section V. Conclusions are drawn in Section VI.

II. System model

We assume the Time Division Duplex (TDD) mode of operation at the base station (BS) with the BS alternating between transmission on the downlink and reception on the uplink . We assume that the BS has knowledge of the channels of all users in the system. Since channel conditions vary over a period of time, the channel state information from each of the users is updated once per frame by means of channel-to-interference-plus-noise ratio (CINR) reports on the uplink feedback channels. Using these CINR reports, the modulation-coding scheme (MCS) levels are chosen based on CINR threshold values for a specific target bit error rate (BER). In this paper, we assume that the wireless channel for a user is constant over an entire downlink OFDMA sub-frame.

Let M be the number of users in the system, K be the number of slots available for allocation during the downlink OFDMA sub-frame and $\mathcal{M} = \{1, 2, 3, ..., M\}$ be the user index set. $f_{1,m}, f_{2,m}$ and $f_{3,m}$ are RT, NRT and BE flows where indices 1, 2, and 3 correspond to RT, NRT and BE flows destined to user m respectively. For the sake of simplicity, we impose a condition that a user can have no more than one active flow of each type i.e, RT, NRT and BE. $\mathcal{F}_m \subset \{f_{1,m}, f_{2,m}, f_{3,m}\}$ is the set of all active flows destined to user m. $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2 \cup \cdots \cup \mathcal{F}_M$ is the set of all downlink flows.

The model and form of delay QoS for RT flows used in this paper is as given below:

$$P(W_{1,m} \geq T_{1,m}) = 0 \tag{1}$$

where $W_{1,m}$ is the delay of a typical packet for user m's RT flow, $T_{1,m}$ is the corresponding latency requirement for the flow. If the latency requirement for the RT flow is violated by a packet, the packet is dropped. For RT, NRT and BE flows, throughput requirements are of the form,

$$D_{i,m} \approx A_{i,m} \forall i \in \{1, 2, 3\} \tag{2}$$

where $D_{i,m}$ is the long-term throughput for user m's flow i and $A_{i,m}$ is the corresponding arrival rate for the flow.

Queues implemented as finite length buffers are maintained in the BS for each of the downlink flows and are updated each time a flow is assigned a slot. The queues operate on a first-in first-out (FIFO) basis i.e, if a flow is assigned bandwidth, packets are pulled from the head of the queue for transmission. Also, if a queue becomes full, the packet at the head of the queue is dropped. Packets are admitted into a queue at the beginning of every frame. We assume that a packet is already delayed by one frame duration when it is admitted into a flow's queue. Let $a_{i,m}(t) \forall i \in \{1, 2, 3\}$ be the number of bits in a packet that arrived during frame $t-1$ for user m's flow i. The number of bits in the queue corresponding to user m's flow i during frame t is given by,

$$Q_{i,m}(t) = Q_{i,m}(t-1) + a_{i,m}(t) \tag{3}$$

Before slot allocation begins during a frame, the throughput for user m's flow i during frame t is initialized as,

$$d_{i,m}(t) = 0 \tag{4}$$

When OFDMA slot allocation is in progress during a frame, if user m's flow i is allocated an OFDMA slot, $d_{i,m}(t)$ and $Q_{i,m}(t)$ are updated sequentially and in order as given below.

$$d_{i,m}(t) = d_{i,m}(t) + \min(Q_{i,m}(t), \mu_m(t)) \tag{5}$$

$$Q_{i,m}(t) = \begin{cases} Q_{i,m}(t) - \mu_m(t) \text{ if} & Q_{i,m}(t) \geq \mu_m(t) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where $\mu_m(t)$ (bits/OFDMA slot) is the maximum number of bits that can be transmitted on an OFDMA slot during frame t as determined by user m's MCS level chosen based on the CINR feedback report.

For RT flows, delay QoS is implemented as given in (1). We define the head-of-line (HOL) packet delay for user m's RT flow as,

$$W_{1,m}(t) = \begin{cases} 1 + \arg \max\limits_{0 \leq d < T_{1,m}-1} r_{1,m}(t-d) & Q_{1,m}(t) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $r_{1,m}(t-d)$ is the number of bits remaining in user m's RT flow from a packet arrival during frame $t-d-1$ such that $r_{1,m}(t-d) \leq a_{1,m}(t-d) \forall m \in \mathcal{M}, 0 \leq d < T_{1,m} - 1$ and, $T_{1,m}$ (in frame durations) is the latency requirement for user m's RT flow.

Short-term traffic arrival and throughput statistics are maintained by measuring traffic arrival and departure over a sliding window for each of the downlink flows. The traffic arrival rate (bits/frame) and throughput (bits/frame) for user m's flow i during frame t is defined as,

$$A_{i,m}(t) = \sum_{d=0}^{D-1} a_{i,m}(t-d) \tag{8}$$

$$D_{i,m}(t) = \sum_{d=1}^{D} d_{i,m}(t-d) \tag{9}$$

where D is the sliding window length (in frame durations).

III. The MLWDF rule

The MLWDF rule [5] for OFDMA slot allocation during frame t for RT flows is,

$$m^* = \arg \max_{m \in \mathcal{M}} \frac{\mu_m(t) W_{1,m}(t)}{\bar{\mu}_m T_{1,m}} \tag{10}$$

where $\bar{\mu}_m$ is the mean number of bits that can be transmitted on an OFDMA slot during a frame for user m. The rule

assigns an OFDMA slot to the flow with the largest MLWDF parameter taking into account the user MCS level and flow HOL delays. When the ratio of the HOL packet delay to the latency constraint is nearly the same for all flows, the rule favours the flow whose user has the best channel condition relative to the mean channel condition. However, when a flow's HOL packet delay approaches the flow's latency constraint, the term $\frac{W_{1,m}(t)}{T_{1,m}}$ approaches 1, resulting in higher priority to the flow thereby over-riding channel conditions.

The MLWDF rule can be modified to meet the QoS requirements of RT flows and throughput requirements of NRT and BE flows. This is done by assuming token queues for NRT and BE flows where tokens are assumed to arrive at a constant rate $A_{i,m} \forall i \in \{2,3\}$, the arrival rate for user m's NRT or BE flow. During each frame, the token queues are incremented by $A_{i,m}\Delta t \forall i \in \{2,3\}$ where Δt is the frame duration. Let $Q'_{i,m}(t) \forall i \in \{2,3\}$ be the number of tokens in user m's NRT or BE flow. The HOL token waiting time for user m's NRT or BE flow is,

$$W_{i,m}(t) = \frac{Q'_{i,m}(t)}{A_{i,m}} \forall m \in \mathcal{M}, i \in \{2,3\} \tag{11}$$

The MLWDF rule to meet the throughput requirements of NRT flows and QoS requirements of RT flows is given by,

$$i^*, m^* = \arg \max_{i \in \{1,2\}, m \in \mathcal{M}} \frac{\mu_m(t) W_{i,m}(t)}{\bar{\mu}_m T_{i,m}} \tag{12}$$

If user m's NRT flow is assigned a slot, the number of tokens in the token queue is reduced by an amount $\mu_m(t)$. If the flow's actual queue $Q_{2,m}(t)$ is empty, any bandwidth that is assigned to user m's NRT flow will go waste. After all RT and NRT flows are serviced to the point where RT flows have empty queues and NRT flows have empty token queues, any remaining bandwidth during a frame can be used to allocate slots to BE flows. The MLWDF rule to meet the throughput requirements of BE flows is,

$$m^* = \arg \max_{m \in \mathcal{M}} \frac{\mu_m(t) W_{3,m}(t)}{\bar{\mu}_m T_{3,m}} \tag{13}$$

IV. Proposed Solution

In this section, we propose a scheduling rule that will meet the QoS requirements of the form (1) and (2) for RT flows and throughput requirements of the form (2) for NRT and BE flows.

For this purpose, we propose a set of flow metrics for RT, NRT and BE flows. These metrics capture the extent to which a flow merits allocation of OFDMA slots and is updated dynamically after the allocation of an OFDMA slot to the flow.

The flow metric definition for RT flows is,

$$FM_{1,m} = \frac{\mu_m(t)}{\bar{\mu}_m} \frac{A_{1,m}(t)}{D_{1,m}(t)} \frac{T_{1,m}}{(T_{1,m} - W_{1,m}(t))} \tag{14}$$

The flow metric definition for NRT and BE flows is,

$$FM_{i,m} = \frac{\mu_m(t)}{\bar{\mu}_m} \frac{A_{i,m}(t)}{D_{i,m}(t)} \forall i \in \{2,3\} \tag{15}$$

With this flow metric definition for individual flows, it is possible to define a scheduling rule that will allocate OFDMA slots to RT and NRT flows during a frame. The scheduling rule for OFDMA slot allocation to RT and NRT flows during frame t is,

$$i^*, m^* = \arg \max_{i \in \{1,2\}, m \in \mathcal{M}} FM_{i,m} \tag{16}$$

If after satisfying the bandwidth needs of RT and NRT flows i.e, all RT and NRT flows have empty queues, OFDMA slots remain, the remaining slots can be used to satisfy the throughput requirements of BE flows as given by the following scheduling rule.

$$m^* = \arg \max_{m \in \mathcal{M}} FM_{3,m} \tag{17}$$

As mentioned earlier, the flow metric definition for flows takes into consideration the short-term arrival and throughput statistics in addition to individual users' channel conditions. Given the bursty nature of arrival of packets and the finite buffer constraint, any short-term increase or decrease in the arrival of packets is taken into consideration by the proposed scheduling rule in making scheduling decisions. If the packet arrival for a flow increases in the short-term, the term $\frac{A_{i,m}(t)}{D_{i,m}(t)}$ increases, giving the flow a better chance of bandwidth allocation. This ensures that flows whose queues are close to getting full are given greater priority when compared to flows with relatively empty queues. Furthermore, to handle QoS requirements of the form (1), an additional term, $\frac{T_{1,m}}{(T_{1,m} - W_{1,m}(t))}$, is introduced for RT flows. When the HOL packet of an RT flow approaches its deadline as given by the latency constraint, the term $\frac{T_{1,m}}{(T_{1,m} - W_{1,m}(t))}$ blows out to give higher priority to such RT flows.

V. Numerical Results

Simulation results are shown for 512 subcarrier 5 MHz bandwidth OFDMA based IEEE 802.16e WirelessMAN system. We assume a 5 millisecond frame where the downlink sub-frame is assumed to be of duration 3.2 millisecond and each OFDMA symbol is of duration 0.1 millisecond. We assume the distributed subcarrier permutation for subchannelization of subcarriers. A total of 240 OFDMA slots are available for allocation out of which 40 slots are set aside to account for MAC overheads such as MAP messages, protocol data unit (PDU) MAC headers and losses due to gaps in tiling downlink bursts etc i.e, during each downlink sub-frame, only 200 slots are allocated to flows. We assume a system operating at 2.5 GHz with $M = 30$ users traveling at vehicular speeds in the range $2 - 100$ km/hr. Users are grouped into 5 groups of 6 users each. All users in a group are assumed to have the same path-loss. Relative to users in group 1, users in groups $2, 3, 4$, and 5 have an additional path-loss of $2, 4, 6$, and 8 dB respectively. FIFO buffers for each flow are designed to hold up to 100 packets. The following table shows the CINR feedback threshold and the corresponding MCS level used in this paper.

CINR feedback threshold (dB)	MCS level
36.0	5/6 64-QAM
32.0	3/4 64-QAM
28.0	2/3 64-QAM
24.0	1/2 64-QAM
20.0	1/2 16-QAM
16.0	3/4 QPSK
12.0	1/2 QPSK

For RT flows, periodic arrival of variable size packets is assumed in the traffic model. For NRT and BE flows, we assume a Bernoulli arrival model in which uniform size packets arrive during a frame with probability $\lambda_{i,m}\ \forall i \in \{2,3\}, m \in \mathcal{M}$. In the subsections to follow, we compare the performance of the proposed scheduling rule with that of the MLWDF rule.

A. *Case 1: $M = 30$ users with one RT flow each*

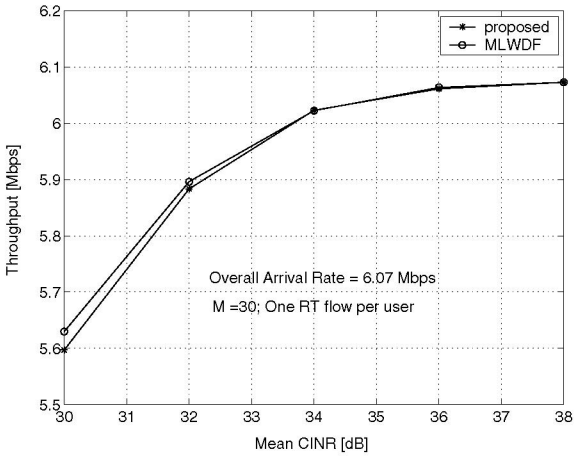

Fig. 1. $M = 30$ users with one RT flow each

In this section, an RT flow with arrival rate $A_{1,m} = 200\,\text{kbps}\ \forall m \in \mathcal{M}$ is assumed for each one of the $M = 30$ users. The packet inter-arrival time is assumed to be 7 frames for each of the flows. The latency requirement for each of the flows is assumed to be $T_{1,m} = 5\,\text{frames}\ \forall\ m \in \mathcal{M}$. Under these assumptions, any packet drop for these flows will occur because of a failure to meet the latency requirement in (1) and not because of the finite buffer constraint. The arrival model assumed here is consistent with the traffic model for high bit rate streaming video applications.

Figure 1 shows the sum throughput of all the flows for increasing CINR values. It is evident that the performance of the proposed rule is comparable with that of the MLWDF rule in that requirements (1) and (2) for RT flows are met at the same CINR$=$ 38 dB.

B. *Case 2: $M = 30$ users with one NRT flow each*

An NRT flow with arrival rate $A_{2,m} = 200\,\text{kbps}\ \forall m \in \mathcal{M}$ is assumed for each one of the $M = 30$ users. A packet arrival probability of $\lambda_{2,m} = 0.1\ \forall m \in \mathcal{M}$ is assumed for each of the flows during a frame. The data rate requirement assumed here is consistent with the requirements for a file transfer of size $\approx 1 - 1.5$ MB in one minute.

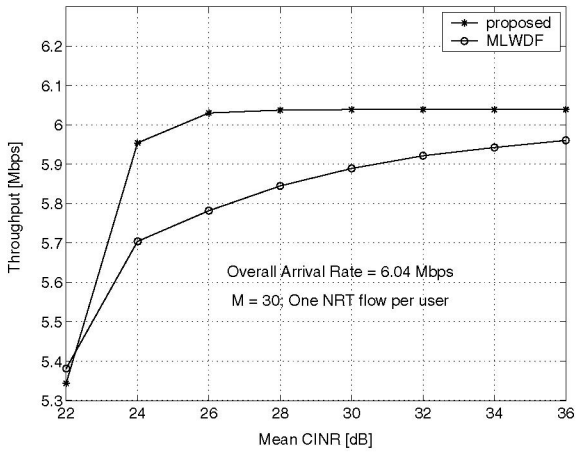

Fig. 2. $M = 30$ users with one NRT flow each

Figure 2, shows the sum throughput of all the flows for increasing CINR values. The proposed scheduling rule is able to meet the throughput requirements of NRT flows at a CINR value of 26 dB while the MLWDF rule is unable to do so even at 36 dB. It is obvious that a gain of over 10 dB over the MLWDF rule can be achieved using the proposed rule for NRT flows.

C. *Case 3: $M = 30$ users with one RT, one NRT, and one BE flow each*

In this section, we consider a system with $M = 30$ users with one RT, one NRT and one BE flow for each user. The arrival model for RT flows assumes an arrival rate of $A_{1,m} = 30\,\text{kbps}\ \forall m \in \mathcal{M}$ with a packet arriving once in every 6 frames and a latency constraint of $T_{1,m} = 5\,\text{frames}\ \forall m \in \mathcal{M}$. For NRT flows, an arrival rate $A_{2,m} = 130\,\text{kbps}\ \forall m \in \mathcal{M}$ is assumed with a packet arrival probability of $\lambda_{2,m} = 0.1\ \forall m \in \mathcal{M}$. For BE flows, an arrival rate $A_{3,m} = 70\,\text{kbps}\ \forall m \in \mathcal{M}$ is assumed with a packet arrival probability of $\lambda_{3,m} = 0.1\ \forall m \in \mathcal{M}$. The arrival model for RT flows is consistent with the traffic model for low bit rate voice call applications. The traffic model for NRT flows is consistent with the requirements for a file download application of size $\approx 0.5 - 1$ MB in one minute while the model for BE flows is consistent with user requirements for World Wide Web surfing.

Figure 3(a) shows the sum throughput of RT and NRT flows, Figure 3(b) shows the sum throughput of BE flows and Figure 3(c) shows the overall throughput for increasing CINR values. It can be seen that the proposed scheduling rule works well for a mixture of RT and NRT traffic in that is able to meet the requirement (1) and (2) for RT flows and the requirement (2) for NRT flows at a CINR of 20 dB while the MLWDF rule is able to do so only at a CINR value of 26 dB. Furthermore, the proposed scheduling rule allocates bandwidth to BE flows only after the requirements of RT and NRT flows are met.

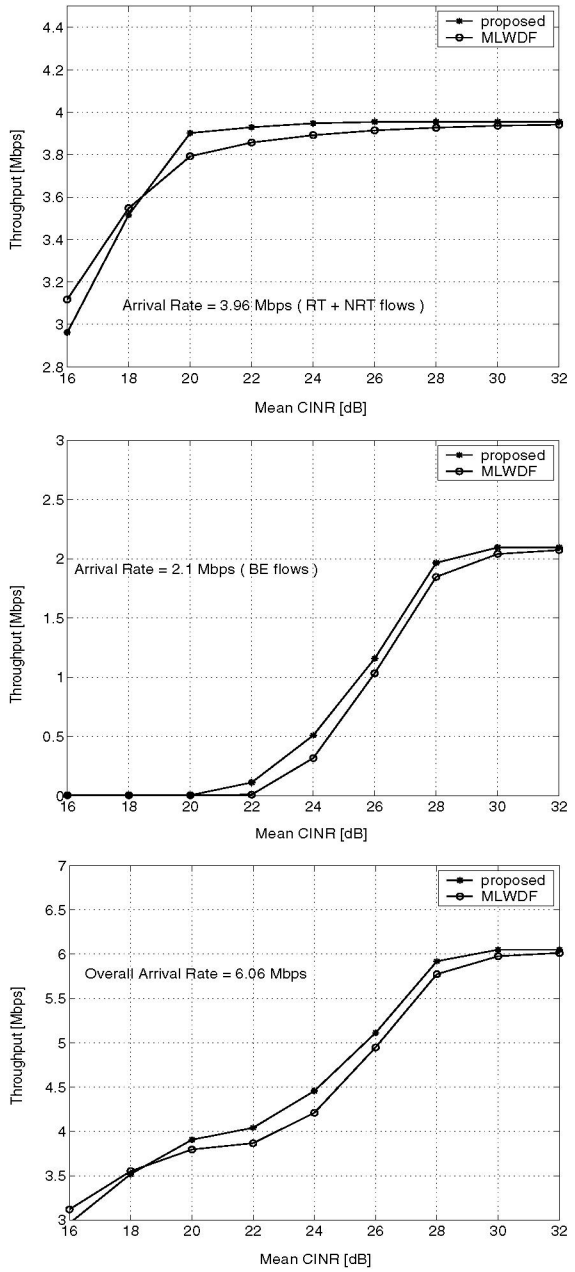

Fig. 3. $M = 30$ users with one RT, one NRT, and one BE flow each. (a) Throughput vs. CINR - RT and NRT flows (b) Throughput vs. CINR - BE flows (c) Overall throughput vs. CINR

VI. Conclusions

In this paper, we proposed a set of flow metrics to capture the extent to which an RT, NRT or BE flow merits OFDMA slot allocation. Using these metrics, we proposed a scheduling rule for OFDMA slot allocation to these flows. Through simulations, we have been able to show that the performance of the proposed scheduling rule is comparable to the performance of the MLWDF rule for purely RT traffic. For a mixture of RT, NRT and BE traffic, the proposed rule can achieve a gain of atleast 5 dB over the MLWDF rule. Furthermore, we assumed a simple Bernoulli arrival model for NRT and BE flows. However, in practice, traffic arrival for these flows could be highly bursty in nature. Since the proposed metric computation takes into account the burstiness in traffic by maintaining the short-term arrival and throughput statistics, good performance and larger gains over the MLWDF rule can be expected.

References

[1] IEEE Std 802.16-2004, IEEE Standard for Local and metropolitan area networks: Part 16: "Air Interface for Fixed Broadband Wireless Access Systems."

[2] IEEE Std 802.16e-2005 and IEEE Std 802.16-2004/Cor1-2005 (Amendment and Corrigendum to IEEE Std 802.16-2004): Part 16: "Air Interface for Fixed and Mobile Broadband Wireless Access Systems."

[3] Taesoo Kwon, Howon Lee, Sik Choi, Juyeop Kim, Dong-Ho Cho, Sunghyun Cho, Sangboh Yun, Won-Hyoung Park, and Kiho Kim, "Design and Implementation of a Simulator Based on a Cross-Layer Protocol between MAC and PHY Layers in a WiBro Compatible IEEE 802.16e OFDMA System," *IEEE Communications Magazine,* vol. 43, no. 12, pp. 136-146, December 2005.

[4] Guocong Song, Ye Li, L. J. Cimini, H. Zheng, "Joint Channel-Aware and Queue-Aware Data Scheduling in Multiple Shared Wireless Channels," *Proceedings of the IEEE Wireless Communications and Networking Conference,* vol. 3, pp. 1939-1944, March 2004.

[5] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, R. Vijayakumar, "Providing Quality of Service over a Shared Wireless Link," *IEEE Communications Magazine,* vol. 39, no. 2, pp. 150154, February 2001.

[6] S. Shakkottai, A. Stolyar, "Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR," *Proceedings of the 17th International Teletraffic Congress,* Salvador da Bahia, Brazil, December 2001.

[7] P. Parag, S. Bhashyam, and R. Aravind, "A Subcarrier Allocation Algorithm for OFDMA Using Buffer and Channel State Information," *Proceedings of the* 62^{nd} *IEEE Vehicular Technology Conference,* Dallas, Texas, USA, vol. 1, pp. 622-625, September 2005.