Buffer Sharing on an OFDMA Downlink

Tolga Girici
Department of Electrical and Electronics Engineering
TOBB University of Economics and Technology
Ankara Turkey
tgirici@etu.edu.tr

Omur Ozel, Elif Uysal-Biyikoglu
Department of Electrical and Electronics Engineering
Middle East Technical University
Ankara Turkey
omur,elif@eee.metu.edu.tr

***Abstract*—In this work we consider the allocation of buffer space to data streams sharing a common high-speed wireless transmitter. As an example, we focus on an OFDMA-based downlink system scenario. Scheduling for maximum throughput has been extensively studied in the literature. However, the practically interesting case of a finite buffer has not been sufficiently addressed before. Especially in the case of overloaded packet queues, the choice of buffer management policy substantially affects the throughput performance. We consider a physical-layer scheduling scheme that allocates users to subcarriers based on channel state, in order to make the most use of multiuser diversity. We then consider optimal buffer partitioning to accommodate the resulting rates. We study the system throughput by simulations. As a benchmark, we also simulate MaxWeight, a well-known cross-layer channel and queue-aware scheduling policy that is throughput-optimal in the absence of a finite buffer constraint. We observe that a suitable buffer management policy with a simple channel-aware queuing policy achieves cross-layer scheduling performance, and can exceed it.**

***Index Terms*—Buffer Partitioning, Optimization, Downlink, Broadband Wireless Access, OFDM , Queueing Analysis.**

I. Introduction

A typical broadband wireless access base station (such as in WiMax [1] and LTE [2]) serves a metropolitan area, where hundreds of users are demanding high speed multimedia applications. As advanced physical layer techniques such as array processing and multicarrier transmission have facilitated the delivery of high speed data over wireless links, the higher layer issues of scheduling and buffer management for multiple users to maximize efficiency and service quality are largely open. Among the main problems which have been successfully addressed[3], whilst under somewhat restrictive assumptions, is achieving network capacity (such as the stability region of a fading broadcast channel, for example) while packet arrivals are discrete stochastic processes, and queues are required to be finite with probability one. However, the related problem of throughput maximization under constraints on queue length ('finite buffer constraint') remains as a challenge.

Meanwhile, finite buffers are a practical reality. One example is relay stations that are used to improve coverage in IEEE 802.16j-based mobile multihop relay networks [4]. These relay stations have to be low in cost, therefore they come with a number of capacity limitations and memory can be one of them. In wireless broadband access networks instantaneous capacity may be far from expected, because of the unpredictable channel conditions. Because of the limitations in transport control protocols, base stations often work in the oversubscribed regime, where some packet buffers will be working with loads greater than one. Therefore, finding implementable scheduling and buffer management strategies in the oversubscribed regime is important.

Resource allocation in broadband wireless access networks is a challenging problem. Next generation broadband technologies are mostly based on Orthogonal Frequency Division Multiplexing (OFDM) in the physical layer. OFDM provides immunity to intersymbol interference and multipath fading; it also can be used as a multiple access scheme (OFDMA), where the subchannels are allocated to individual users, based on channel condition, buffer occupancy and service requirements. As for optimal throughput, for single channels Largest weighted delay first (LWDF) in [5] is shown to be throughput optimal. This scheme was extended for OFDMA-based (multichannel) systems in [6] and [7]. These schemes use head-of-line delay or average delay along with channel condition in scheduling metrics and use those metrics to allocate subchannels to users one by one. In wireless metropolitan area networks fairness is also an important criteria because channel-aware scheduling schemes normally favor users close to the base station. Proportional fairness [8] is a good tradeoff between throughput and fairness. Such schemes usually give equal chance to access the channel and hence users receive rates proportional to their average rate or SNR. Proportional fairness in OFDMA based systems was formulated in [9].

It is conceivable that the optimal solution of the scheduling-throughput maximization problem requires a cross-layer algorithm: one which uses information on channel and queue/delay states at the same time. It is even more so in the finite buffer case. Even if the arrival rates are the same, in a metropolitan area network, the service rates of users in the cellular area vary considerably with distance. If a purely channel state-based scheduler is applied, then the unserved packets of the distant users may fill up the buffer, resulting in, (1) unfairness: Some users do not get any service, and (2) loss of network capacity: as only a subset of the users are using the total system bandwidth, the throughput is limited by these users' total achieved rate. In addition, there is a loss in total achievable rate in the wireless channel due to the reduced multiuser diversity. Specifically, in our example, considering equal service demand by all users, eventually the only scheduled users will tend to be among the ones whose receivers are most distant from the transmitter, as they typically have the lowest data rates hence the highest loading factors. On the other hand, when rate

978-1-4244-8016-6/10/$26.00 ©2010 IEEE

demands are unequal, even in the case of symmetric channels, users with high demand will tend to fill up, or **hog**, the buffer.

One way of solving the described problem is judicious buffer management. It is worth defining the two opposite extremes of using a buffer: Complete Sharing (CS) and Complete Partitioning(CP) [10] [11]. In CS, the complete memory pool can be used by the BS for all sessions. This improves the degree of resource utilization but a session with high traffic intensity (e.g. in a BWA network, a distant user) can completely *hog* the buffer degrading the performance as mentioned in the previous paragraph. On the other hand, in CP, each session has its own reserved buffer partition. This is a definite solution to the hogging problem, for example, in the simplest case of Equal Partitioning (EP). However, EP is not necessarily the throughput-optimal one among all CP solutions.

Optimal buffer partitioning has been studied in [13][1]. In this paper, building on the results of [13], we consider the joint problem of buffer management and scheduling in an OFDMA based finite-buffer downlink system. Jointly optimal scheduling and buffer management is at present an open problem. Here our goal is not to fully solve this problem, but suggest that its solution could lead to technologically attractive single-layer schemes that achive cross-layer performance in the high-speed wireless downlink. We do this by presenting a proof-of-concept algorithm and compare its performance with the MaxWeight benchmark. Specifically, we want to answer the following questions: 1) How effective is the buffer sharing policy on throughput and fairness performances in the over-subscribed regime? 2) Can a simple scheduling policy with suitable buffer management perform better than current best known cross-layer scheduling policies?

We first consider a simple normalized SNR-based scheduling as in [15] and estimate its throughput performance under finite buffer assumptions. Based on this estimation, we consider allocating the buffer resource optimally. In addition to this scheme, we consider Max-Weight schemes to which both queue and channel state information is made available, and which make use of these jointly. The latter type of schemes are *cross-layer* and they possibly have better performance. However, a method that keeps layers *separate* is certainly preferable- if not necessary, as system designers often argue- from an implementability perspective. We shall target such a layered approach and see if a purely channel aware scheduling can perform comparably to a cross layer scheduling, if suitable buffer management is applied. In the next section we will present the formulation of buffer management as an optimization problem.

II. Buffer Partitioning

In [12] it was observed that throughput is an increasing concave function of traffic intensity ρ and buffer capacity m for $M/M/1/m$ and $M/G/1/m$ systems. Let $B(\rho, m)$ be the overflow probability and let $T(\rho, m) = \lambda(1 - B(\rho, m))$ be the throughput. $T(\rho, m+1) - T(\rho, m)$ is the increase in throughput by adding one more unit to the buffer capacity. In [12] it is proven that $T(\rho, m+1) - T(\rho, m)$ is a decreasing function, therefore increasing buffer capacity brings diminishing returns. Another interesting finding is that the traffic intensity ρ^* that maximizes this throughput improvement in $M/M/1/m$ is between 1 and 1.8. This implies that in the oversubscribed case that we consider, optimal buffer partitioning is very important.

A. Optimal Buffer Allocation Algorithm

In [13], these two properties are used and a buffer allocation algorithm with optimal throughput is proposed. Suppose that the BS has memory capacity of M packets of fixed length. The algorithm initially allocates space for one packet to each of the N nodes. Then it calculates $\Delta T_i(m_i) = T(\rho_i, m_i + 1) - T(\rho_i, m_i)$ for each node, where m_i is the buffer space allocated to node i at the current step. The BS allocates one more buffer space to the node maximizing $\Delta T_i(m_i)$. This continues until all the buffer spaces are allocated. If the throughput function $T(\rho, m)$ is concave in m, then the algorithm is optimal [13].

III. OFDMA-Based Downlink Resource Allocation with Finite Buffers

In this case we consider a base station transmitting to N users. A wideband channel of bandwidth W is divided into K subchannels of W_{sub}Hz. We assume that the channel gain from Base Station to each user consists of a slow and a fast-varying component. Slow component consists of pathloss and fading, while the fast part consists of Rayleigh fading. We assume in this work that the slow part is actually time-invariant for each user and the fast part is independent and identically distributed for each user, time slot and subchannel. Assuming a fixed transmission power per subchannel (as we assume in this work) the SNR of user i at subchannel k is $\gamma_{i,k} = \gamma_i^0 h_{i,k}$, where γ_i^0 is the average SNR of user i (combination of pathloss and shadowing) and $h_{i,k}$ represents fast Rayleigh fading at subchannel k for user i. Therefore $h_{i,k}$ is an exponential random variable with mean one. The achievable rate by user i at subchannel k is assumed to be $\log_2(1 + \gamma_{i,k})$ bits/sec/Hz. We assume a normalized SNR-based scheduling , where at each subchannel, the user $\arg\max_i\{\frac{\gamma_{i,k}}{\gamma_i^0} = h_{i,k}\}$ is scheduled. Because of the i.i.d. nature of normalized SNR, each user gets any subchannel with probability $1/N$. We prefer this scheduling method because it maintains a balance between fairness and throughput, by allocating equal resource to each user on the average. Besides it is an analytically tractable method. Other types of scheduling methods and joint scheduling and buffer management is a subject of future research. We assume packets of constant length L bits. Let R be the random variable representing the achievable spectral efficiency given that a node with average SNR γ^0 wins subchannel k. The expression for R is $R = \log_2(1+\gamma^0 \max_i h_{i,k})$. Using extreme value theory [14], [15] the probability distribution function of

[1] There are also hybrid schemes like buffer sharing with minimum buffer guarantees and push-out type of policies , where some packets in the buffer can be removed for some higher priority packets that arrive. These schemes are beyond the scope of this work.

number of bits that can be transmitted by the maximizing user from a subchannel converges to the following expression as number of users go to infinity,

$$\lim_{N\to\infty} F_R^N(r) = \exp\left(-\exp\left(-\frac{r-a^N}{b^N}\right)\right) \tag{1}$$

where $\exp(-\exp(-x))$ is the normalized Gumbel distribution. Here a^N and b^N (in bits/sec/Hz) is [14],

$$\begin{aligned} a^N &= \log_2(1+\gamma^0 \ln N) \\ b^N &= \log_2\left(\frac{1+\gamma^0(1+\ln N)}{1+\gamma^0 \ln N}\right) \end{aligned} \tag{2}$$

The probability density function of R is as follows:

$$\begin{aligned} &\lim_{N\to\infty} f_R^N(r) \\ &= \frac{1}{b^N}\exp\left(-\exp\left(-\frac{r-a^N}{b^N}\right)\right)\exp\left(-\frac{r-a^N}{b^N}\right) \end{aligned} \tag{3}$$

The average spectral efficiency converges to $a^N+E_0 b^N$, where $E_0 = 0.5772...$ is the Euler number [14]. The standard deviation becomes $b^N\frac{\pi}{\sqrt{6}}$. As seen from these expressions, the mean increases and the standard deviation decreases with increasing N. Therefore achievable rate per subchannel converges to a deterministic quantity as N increases.

A. M/G/1/m Model

We assume Poisson arrivals of rate λ bits/sec for each user. Packets are of constant length L bits. The services for each time slot and subchannel can be in fractions of a packet. Let M be the total buffer capacity and for the case of complete partitioning, let m_i be the buffer capacity allocated to user i by the BS. If a packet does not completely fit into the residual buffer memory, it is dropped.

The system described can be modeled as an $M/G/1/m_i$ system, where m_i is the buffer allocated to user i. Gelenbe's formula is known as an accurate approximation to the packet drop probability for $M/G/1/m_i$ [19]:

$$P_d(\lambda,\mu,m_i) = \frac{\lambda(\mu-\lambda)\exp(-2\frac{(\mu-\lambda)(m_i-1)}{\lambda A^2+\mu S^2})}{\mu^2-\lambda^2\exp(-2\frac{(\mu-\lambda)(m_i-1)}{\lambda A^2+\mu S^2})}, \forall i \tag{4}$$

In the above, $S = \frac{Var[T]}{E[T]^2}$ where T is service time, and $A = \frac{Var[T_a]}{E[T_a]^2}$, where T_a is inter-arrival time. As inter-arrivals are exponential, $A = 1$. To apply the above approximation for drop probability, we shall thus need the first and second moments of the service time.

The service time (*i.e.*, packet transmission duration) is the number of timeslots from the start of the transmission of the first bit of a packet until the end of the time slot in which all L bits are sent. Considering packets that are long compared to the maximum number of bits that can be sent in a subchannel, it will take a number of subchannel uses to transmit a packet. The probability that user i gets a given subchannel in a given time slot is $\frac{1}{N}$ and K is the number of subchannels. We know that each packet is of length L bits, and user n will on average transmit $\frac{WT_s}{K}(a_n^N + E_0 b_n^N)$ bits of data when it is assigned a subchannel in any given timeslot [2]. These statistical values are obtained using Extreme Value Theory , and they are exact as number of users go to infinity. In the rest, we will assume that these values are exact.

Now consider T time slots. The number of bits collected by a given user over this duration is the random variable $S_T = \sum_{j=1}^{T} R_j$, where R_j is the number of bits that the user receives in the j^{th} timeslot. Note now that T is a stopping rule associated with the process S_T [18] In order to be able to use Gelenbe's formula, we need to find the mean and variance of time until a L bits are received (i.e. process is stopped) by a user. The following theorem was proven in [18] and references therein.

Theorem 1: Let $x_1, x_2, \ldots$ be independent random variables with mean μ and variance σ^2. Let $s_n = \sum_{k=1}^{n} x_k$ and for any $c > 0$ define $T = T(c)$ as the first $n \geq 1$ such that $s_n > c$. Then, as $c \to \infty$, $E\{T\} \sim \frac{c}{\mu}$ and $Var\{T\} \sim \frac{c\sigma^2}{\mu^3}$

In our problem, stopping time is packet completion time, and packets are of constant length L, so $c = L$. Note that the result in Theorem 1 is asymptotic in c, so we are considering the regime where packet length is much larger than the mean number of bits transmitted, μ, per subcarrier allocated. Note that this mean number of bits is approximately $\frac{WT_s}{N}(a + E_0 b)$, where a, and b are defined in (2). As for the variance of the number of bits per time slot, σ^2, again let us again use the large-n stochastic limit for this quantity as an approximation. Variance of number of transmitted bits over an allocated subchannel is approximately $(\frac{WT_s}{K})^2 b^2 \frac{\pi^2}{6}$ as mentioned above, which follows from extreme value theory. The second moment is found as $(\frac{WT_s}{K})^2(b^2\frac{\pi^2}{6} + (a+E_0 b)^2)$, by adding the square of its mean. A subchannel is allocated with probability $1/N$. Thus, the second moment of transmitted bits over any subchannel is found by multiplying the above quantity by $1/N$. We find the parameter σ^2 by subtracting the square of mean and then multiplying the result by K.

$$\begin{aligned} \sigma^2 &\sim K\left[\frac{1}{N}\left(\frac{WT_s}{K}\right)^2\left(b^2\frac{\pi^2}{6} + (a+E_0 b)^2\right)\right. \\ &\qquad \left. -\frac{1}{N^2}\left(\frac{WT_s}{K}\right)^2 (a+E_0 b)^2\right] \\ &= \frac{K}{N}\left(\frac{WT_s}{K}\right)^2\left(\frac{b^2\pi^2}{6} + (a+E_0 b)^2\left(1-\frac{1}{N}\right)\right) \end{aligned}$$

As a result, the mean and variance of the service time are equal to,

$$\begin{aligned} E[T] &\sim \frac{LN}{WT_s}\frac{1}{a+E_0 b} \\ Var[T] &\sim \frac{L\frac{K}{N}\left(\frac{WT_s}{K}\right)^2\left(\frac{b^2\pi^2}{6} + (a+E_0 b)^2\left(1-\frac{1}{N}\right)\right)}{\left((\frac{WT_s}{K})^2 b^2\frac{\pi^2}{6}\right)^3} \end{aligned} \tag{5}$$

Using these, the parameter S in (4) can be calculated.

[2] From now on, we will omit the subscripts and superscripts in a_n^N and b_n^N for the sake of simplicity. In fact, the parameters a and b are possibly different for each user and depend on the user distance.

In the performance evaluations, we will compute the optimal partitions by computing the blocking probability using Gelenbe's formula given above, with the service time mean and variance we have just derived. The computation of optimal partitions uses the observation that the throughput $T(\lambda, \rho_i, m_i) = \lambda(1 - P_d(\lambda, \mu_i, m_i))$ is a concave function of m_i in the $M/G/1/m_i$ system [13], and the algorithm provided in chapter II of [13].

IV. Performance Studies

We consider a system of 40 users and a cellular area of radius 2000m. In order to better observe the performance, the users are located at discrete distances from the BS. 20% of the users are located at 400, 800, 1200, 1600 and 2000 meters from the BS. We arranged the BS power, channel noise and Non Line of Sight path loss model so that the users at each distance level have approximately 34.4, 23.9, 17.7, 13.4 and 9.9dB SNR. A 1MHz channel is divided into 100 subcarriers.

We will evaluate the performance of the proposed optimal buffer partitioning scheme by comparing some joint-scheduling and buffer management schemes. Scheduling schemes include channel-aware and joint channel and queue-aware policies. Buffer management schemes include complete sharing, and equal and optimal partitioning schemes. Performance comparison will be performed by gradually increasing the system arrival rate for a fixed number of users, and memory resource. But first, we need to observe the accuracy of the M/G/1/m analysis that we presented in the previous section.

A. Accuracy of The M/G/1/m Analysis

While analyzing the drop rate and throughput, we used approximation techniques such as extreme value theory and Gelenbe's formula. In this part, we will test the accuracy of these approximations. We will consider the normalized SNR based scheduling mentioned in Section 3. Let's call this scheme as MC. As the buffer partitioning scheme, we consider equal partitioning, which will be denoted as EP. In Figure 1 we see the per user throughput for different users as a function of buffer space for the MC-EP scheme. For a system of 40 users and 10Mbps total arrival rate (i.e. 250kbps per user) buffer space is changed from 4 to 12 packets per user, which means a change from 40KB to 120KB total memory resource. The dotted lines are the simulation results. It can be seen that analytical results closely follow the simulation results. M/G/1 model is especially more accurate for distant users, which have higher load and therefore higher drop rate.

Figure 2 shows the throughput as a function of arrival rate. For a system of 40 users and a memory space of 8 packets/user (i.e. 80KB of total resource) , the total arrival rate is changed from 4 to 12 Mbps. We observe even a closer match between the results of the simplified analysis of the previous section and the simulation results, where the deviation is always below 1%. These results show that the M/G/1 analysis can be used in optimum buffer partitioning.

B. Benchmark Algorithms

Let MC denote the scheduling scheme mentioned above. As an alternative to MC we will consider Max-Weight (MW) scheduling, where, each subchannel, at each time slot is allocated to the user

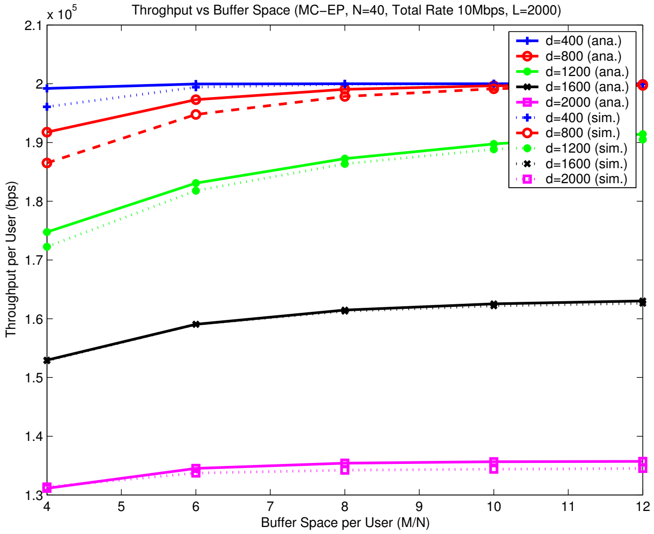

Fig. 1. Accuracy of the analytical approximation is tested by increasing the total buffer space. Accuracy increases as the buffer space and user distance (hence user load ρ) increases. For most of the cases the analytic result deviates less than 1% from the simulation results.

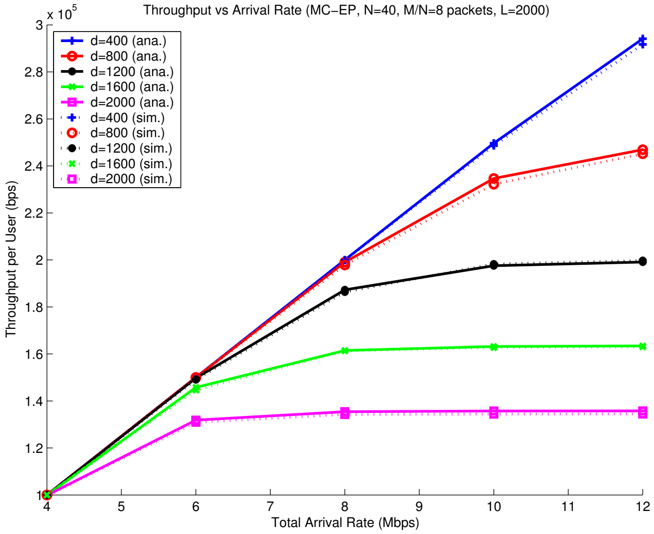

Fig. 2. Accuracy of the analytical approximation is tested by increasing the bit arrival rate. Analytic result always deviates less than 1% from the simulation results.

$$\arg\max_i \{q_i(t) \log_2(1 + \gamma_{i,k})\} \tag{6}$$

, where $q_i(t)$ is the buffer occupancy for user i at time slot t. We also consider a proportional fair alternative called MW-PF, where the metric is

$$\arg\max_i \left\{ \frac{q_i(t) \log_2(1 + \gamma_{i,k})}{R_i(t)} \right\} \tag{7}$$

$R_i(t)$ is the average received rate for user i at time slot t, which is updated at each time slot as $R_i(t + 1) = \alpha R_i(t) +$

$(1-\alpha)r_i(t)$, where $r_i(t)$ is the rate allocated to user i at time slot t and α is a constant typically close to 1. In both of these queue-aware scheme, after the allocation of each subchannel (e.g. subchannel k to user i^*) the buffer occupancy of the user is updated as $q_{i^*}(t) = \max(0, q_{i^*}(t) - \frac{WT_s}{K}\log_2(1+\gamma_{i^*,k}))$, so that users are not overallocated.

As for the buffer management schemes, CS, EP, OP mean Complete Sharing, Equal Partitioning and Optimal Partitioning, respectively.

C. Simulation results and performance comparisons

The arrival rate for each user varies from 175 to 350kbps. Packets are of fixed length of 2000 bits. Considering the average SNR and the multiuser diversity gain calculated in (2), the total traffic intensity of the system for the MC scheme goes approximately from 0.5 to 1.8. Of course, individual traffic intensities for the near (far) users are lower (higher) than these average numbers for each arrival rate.

Figure 3 shows the throughput performance MC-CS, MW-CS, MW-PF-CS and MC-OP, in order to show the effect of using Optimal Partitioning with a channel-aware scheduler. Total buffer capacity is $M = 200$ packets (50KB). For the optimal partitioning, we used the M/G/1/m based model. We made the following observations,

1) Complete sharing policies (MC-CS, MW-CS, MW-PF-CS) all saturate early. This is due to the hogging effect of distant user traffic. Especially the performance of MC-CS is poor.
2) Applying optimal partitioning to MC (MC-OP) we obtain significant performance improvement with respect to MC-CS (60% improvement). Moreover, in the high load case MC-OP performs better than max weight/complete sharing policies. Considering that the max-weight scheduler uses full knowledge of queue states and the channel, and is optimal (under infinite buffers), outperforming it by simple channel-state based scheduling is notable.

This time, in Figure 4, we apply equal partitioning to the Max-Weight policies and compare their performance to MC-EP and MC-OP. We make the following observations.

1) The algorithm that we achieve by allowing MaxWeight to also use a partitioned buffer (MW-EP) is the best out of all policies considered. This algorithm is really a cross between queue-and-channel aware scheduling, and buffer management. This shows that even an asymptotically throughput-optimal cross-layer scheme can benefit from buffer partitioning, when buffers are finite. This observation is encouraging, both because it suggests that partitioning has significant impact, and because it suggests a direction in which to look for the solution of the globally optimal joint scheduling and buffer management problem. The simple, MC-EP and MC-OP policies are seen to achieve comparable performance to MW-EP, which answers the question posed in the Introduction. At this point, one may ask why no MW-OP policy is shown. The answer is twofold: first, the service time distribution for MaxWeight with partitioned buffers is quite intractable thus making it intractable to compute optimal buffer partitioning using our methods, and second, as MaxWeight adapts to queue size, and service time depends on queue size itself, the circulant approach of setting limits to queue sizes based on service time does not make much sense.
2) MC-EP and OP policies are better than MW-PF-EP at high load. Since MC is a normalized SNR based scheduling, it inherently respects proportional fairness. Therefore comparing it with MW-PF-EP is a somewhat fairer comparison.

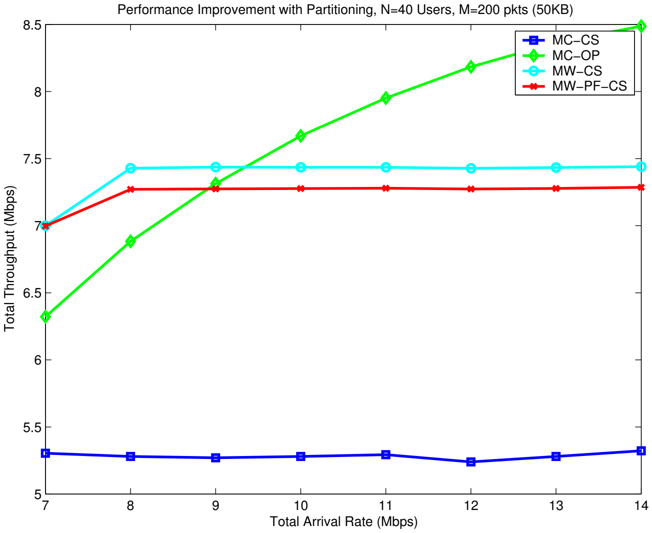

Fig. 3. Total Throughput vs. Arrival Rate: Max Channel-Optimal Partitioning (MC-OP), our channel-based scheduling and optimal buffer partitioning policy, significantly outperforms channel-aware scheduling under no buffer partitions (Max Channel- complete sharing, MC-CS) as load increases. This is due to the *hogging* of the buffers by high-load sessions.

Figure 5 plots the sum of logarithms of user throughputs (a measure of proportional fairness [9]) for the same system parameters as Figures 3 and 4. The most interesting result is that the MC-OP scheme performs even better than MW-EP in the high load case. While complete sharing schemes still perform rather poorly, contrary to Figure 4, MW-EP scheme also becomes poor in terms of proportional fairness as the system load increases. This is quite intuitive: as load increases, eventually each user has a buffer that is almost surely full, and at that point, queue size is no longer a distinguishing factor between users in the MaxWeight scheme. Therefore MW-EP turns into a maximum-instantaneous-throughput scheme. It does attain maximum throughput (as shown in Figure 4) but distant users are almost never scheduled, which degrades the fairness.

V. Conclusions

In this work we considered the problem of subchannel and buffer allocation in an OFDMA-based downlink system. We developed an accurate finite buffer queueing model for

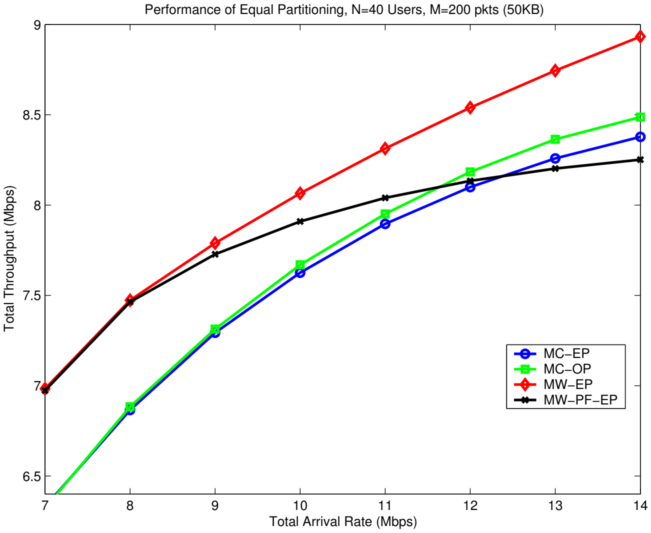

Fig. 4. Total Throughput vs. Arrival Rate: MaxWeight, the cross-layer scheduling benchmark, can also make use of partitioned buffers. This way, the best performance out of all algorithms considered is achieved. MC-OP still outperforms MW-PF-EP in the high load case.

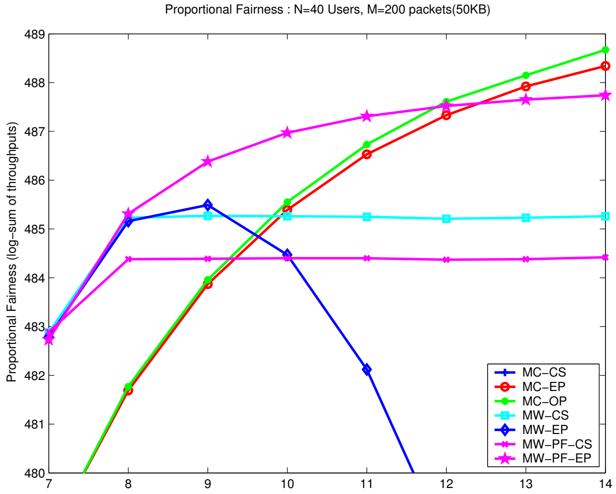

Fig. 5. Proportional Fairness (log-sum throughput) vs arrival rate. MC-OP is the best for higher loads. MW-EP has a surprisingly poor proportional fairness performance.

a channel-based scheduling scheme and obtained the optimal buffer partitioning based on that model. By simulations, we compared the throughput performance of this scheme with other scheduling/buffer management schemes. The simulation results reveal that using simple channel aware scheduling jointly with equal partitioning, provided better proportional fairness performance than joint channel *and queue* aware scheme with complete buffer sharing, as load increases. It is also seen that with optimal buffer partitioning scheme, comparable, and even better, throughput can be achieved with respect to MaxWeight, which is optimal for infinite buffers.

There are plenty of possible interesting directions for future work: for one thing, the question of optimal joint scheduling and buffer control is open. Secondly, while the results here seem promising, as the optimal cross-layer scheme for finite buffers is unknown, the question of whether single-layer schemes are sufficient is still unanswered. Of course, there are the more practical aspects of all these questions, which involve transport-layer mechanisms for controlling the load that enters the buffer.

Acknowledgment

The work in this paper was supported in part by TUBITAK under grants no.108E208 and no.106E119, and in part by NSF.

References

[1] C. Eklund, R. B. Marks, K.L. Stanwood, and S. Wang.," IEEE Standard 802.16: A Technical Overview of the WirelessMAN Air Interface for Broadband Wireless Access', IEEE Comm. Magazine, June 2002.

[2] H. Ekstrom, A. Furuskar, J. Karlsson, M. Meyer, S. Parkvall, J. Torsner, and M. Wahlqvist', Technical solutions for the 3g long-term evolution', Communications Magazine, IEEE, 44(3):3845, March 2006.

[3] E. Yeh and A. Cohen, 'Throughput and delay optimal resource allocation in multiaccess fading channels', Information Theory, 2003. Proceedings. IEEE International Symposium on, pp. 245245, June-4 July 2003.

[4] Peters, S.W.; Heath, R.W.; , 'The future of WiMAX: Multihop relaying with IEEE 802.16j,' Communications Magazine, IEEE , vol.47, no.1, pp.104-111, January 2009

[5] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, 'Providing Quality of Service over a Shared Wireless Link', IEEE Communcations Magazine, pages 150154, Feb. 2001.

[6] G. Song, Y. Y. Li, L. J. Cimini, and H. Zheng,' Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels', Proc. IEEE WCNC, 3:19391944, Mar. 2004.

[7] P. Parag, S. Bhashyam, and R. Aravind, 'A subcarrier allocation algorithm for OFDMA using buffer and channel state informatio', Proc. 62nd IEEE Veh. Technol. Conf., 1:622625, Sep. 2005.

[8] A. Jalali, R. Padovani, and R. Pankaj,' Data througput of CDMA-HDR: A high efficiency-high data rate personal communication wireless system', In Proceedings of the IEEE Semiannual Vehicular Technology Conference, VTC2000-Spring, Tokyo, Japan, May 2000.

[9] H. Kim and Y. Han, ' A Proportional Fair Scheduling for Multicarrier Transmission Systems', IEEE Comm. Letters, pages 210212, Mar. 2005.

[10] F. Kamoun and L. Kleinrock, 'Analysis of shared finite storage in a computer network node environment under general traffic conditions', IEEE Trans. Commun, July 1979

[11] G. Foschini and B. Gopinath, 'Sharing memory optimally', IEEE Transactions on Communications, vol. 31, pp. 352360, March 1983

[12] Ziya, S., 'On the relationships among traffic load, capacity, and throughput for the M/M/1/m, M/G/1/m-PS, and M/G/c/c queues', IEEE Transactions on Automatic Control 53 (2008),2696-2701.

[13] Ozel,O., 'Optimal Resource Allocation Algorithms for Efficient Operation of Wireless Networks ', M.Sc. Thesis, Middle East Technical University, 2009

[14] Galambos, J.,'The Asymptotic Theory of Extreme Order Statistics', John Wiley & Sons, 1978

[15] G. Song and Y. (G.) Li, 'Asymptotic throughput analysis for channel-aware scheduling,' IEEE Transactions on Communications, vol.54, no.10, pp.1827-1834, Oct 2006.

[16] R. G. Gallager, 'Discrete Stochastic Processes', Boston/ Dordrecht/ London: Kluwer Academic Publishers, 1996.

[17] NIST/SEMATECH, '6.3.3.1. Counts Control Charts, e-Handbook of Statistical Methods', http:// www.itl.nist.gov / div898 / handbook/pmc/section3/pmc331.htm

[18] D. Siegmund, 'The Variance of One-Sided Stopping Rules', in *The Annals of Math. Statistics*, Vol. 40, No. 3 (Jun., 1969), pp. 1074-1077.

[19] Guoqing Li; Hui Liu; , 'Dynamic resource allocation with finite buffer constraint in broadband OFDMA networks,' Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE , vol.2, no., pp.1037-1042 vol.2, 20-20 March 2003