Downlink Scheduling with Guarantees on the Probability of Short-Term Throughput

Na Chen and Scott Jordan, *Member, IEEE*

***Abstract*—We consider the problem of scheduling multiple transmissions on the downlink of a wireless network with performance guarantees in the form of the probabilities that short term throughputs exceed user specified thresholds. Many interactive data applications have some degree of a latency requirement, and measure performance by throughput over a relatively short time interval. We refer to the fraction of time such user throughput reaches a predefined rate threshold or higher as *tail probability*. The problem is formulated as maximizing the minimum ratio of tail probability to the user specified probability threshold. We present necessary and sufficient optimality conditions for the case in which the time interval of interest is consistent with the time scale of channel variation. An online algorithm is proposed which can achieve the optimality. For the case in which the time interval of interest is large compared to the time scale of channel variation, we develop an online algorithm which attempts to maximize the minimum normalized tail probability by taking the advantage of channel variation over users and over time. Simulation results demonstrate that the proposed algorithm can achieve better performance than other algorithms such as the proportional fair algorithm and the Max C/I algorithm.**

***Index Terms*—Opportunistic scheduling, statistical performance guarantees.**

I. Introduction

We consider a wireless scheduling problem with performance guarantees in the form of the probabilities that short term throughputs exceed user specified thresholds. It has often been noted that for elastic data applications channel variation over users and over time can be exploited to improve the long-term system performance. Opportunistic algorithms take advantage of channel variation and increase long-term average throughput at the cost of unfairness among users. Many opportunistic algorithms (see e.g. [1]–[5]) limit the extent of this unfairness by providing lower-bound guarantees on individual long-term performances.

While much research has been done for scheduling problems concerning performance guarantees on individual *long-term* average throughput or related utility (see e.g. [1], [2], [4], [5]), little work has addressed how to guarantee the *short-term* average throughput seen by users. One algorithm which focusses on short-term throughput is a variant of the proportional fair algorithm [6], in which the average served data rate is updated by an exponentially weighted low-pass filter; the algorithm provides no guarantees, but achieves good average throughput over a time scale that is relatively large compared to the time scale of channel fluctuation. Another approach, wireless credit-based fair queuing [7] achieves short-term fairness, defined as a probabilistic bound on the difference between the weighted time in service of two users. Other algorithms provide probabilistic guarantees on the delay experienced by each packet, e.g. Largest Weighted Delay First [8].

In contrast, in this paper we consider a performance guarantee on the probability that short-term throughput exceeds a rate threshold specified by each user. Each user can request that this probability exceed a specified level. The performance guarantee thus concerns a tail probability of short-term throughput, as opposed to the short-term guarantees on delay considered in the literature. In addition, the guarantees can be tailored to each user.

The rest of this paper is organized as follows. In Section II, we introduce the system model, define the performance requirements, and formulate the problem. In Section III, we consider the case in which the channels fluctuate very slowly compared to the time interval of interest to users; the necessary and sufficient optimality conditions are presented, followed by an optimal online algorithm. In Section IV, we develop an algorithm for a more general case in which the channels fluctuate quickly compared to the time interval of interest to users.

II. Problem Description

We consider the downlink in a single-cell CDMA system consisting of a base station and a fixed number, M, of users. The base station schedules transmissions in a time slots of fixed duration on the order of the channel's coherence time. Users are assumed to have an infinite backlog of data. A user is *active* when the base station is transmitting data to it. For each time slot, the scheduler makes decisions of which users are active, their transmit power levels and their transmission rates. We assume that the channel gains of users are discrete-time random processes, independent of each other, and that they are estimated by the base station. Let $x_i(t)$ and $s_i(t)$ denote the transmission rate and transmit power, respectively, of user i in time slot t, with:

$$x_i(t) = \frac{W}{\gamma} \frac{s_i(t) h_i(t)}{I_0 + N_0} \tag{1}$$

where W is the spreading bandwidth, γ is the required bit energy-to-interference density ratio, I_0 is the interference

Manuscript received October 26, 2007; revised August 12, 2008; accepted August 25, 2008. The associate editor coordinating the review of this letter and approving it for publication was Y. Fang.

N. Chen is with the Department of Electrical Engineering and Computer Science, University of California, Irvine, CA 92697, USA (e-mail: nac@uci.edu).

S. Jordan is with the Department of Computer Science, University of California, Irvine, CA 92697, USA (e-mail: sjordan@uci.edu).

Portions of this paper appeared in the 2008 Wireless Communications and Networking Conference. This material is based upon work supported by the National Science Foundation.

Digital Object Identifier 10.1109/TWC.2009.071117

power, N_0 is the background noise power, and $h_i(t)$ is the channel gain of user i in time slot t. (Consideration of variable interference power, non-orthogonal codes, and estimation errors is beyond the scope of this paper.) The throughput measured over t_c time slots seen by user i in time slot t is then given by $m_i(t;t_c) = \sum_{\tau=t-t_c+1}^{t} x_i(\tau)/t_c$.

The fraction of time that $m_i(t;t_c)$ exceeds a rate threshold r_i is given by:

$$P_i(r_i;t_c) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} I(m_i(t;t_c) \geq r_i)$$

where $I(\cdot)$ is the indicator function. The performance requirements of users are in the form of $P_i(r_i;t_c) \geq q_i,\ \forall i$, where r_i and q_i are specified by user i. We consider two resources, the total transmit power S_{tot} and the total channels (or codes) N_{tot}:

$$\sum_{i=1}^{M} s_i(t) \leq S_{tot},\ \forall t \tag{2}$$

$$\sum_{i=1}^{M} I(x_i(t) > 0) \leq N_{tot},\ \forall t \tag{3}$$

The objective is to minimize the cost of the resources consumed to satisfy the users' performance requirements on tail probabilities. Let Q denote a scheduling policy, which determines resource allocation among users over time. Denote $z(S_{tot}, N_{tot}) = \min_i P_i(r_i;t_c)/q_i$. The problem is:

$$(\mathbf{P1}) \max_{Q} z(S_{tot}, N_{tot}) \text{ s.t. } (2) \ \&\ (3)$$

We consider the special case $t_c = 1$ in section III, and return to the case $t_c > 1$ in section IV.

III. The case of $t_c = 1$

In this case, $P_i(r_i;t_c) = P(x_i(\tau) \geq r_i)$. The optimal transmission rate of a user is thus either r_i or 0, and the scheduler only need determine which users are active in each time slot. Let $D = (d_1, d_2, ..., d_M)$, where $d_i = I(\text{user } i \text{ active})$. Channel processes are assumed stationary and ergodic. As a result, the scheduler makes decisions based on the current channel state. In channel state $H = (h_1, h_2, ..., h_M)$, a decision vector D is feasible *iff* resource constraints are satisfied:

$$\sum\nolimits_{i=1}^{M} s_i(H;r_i)d_i \leq S_{tot},\ \sum\nolimits_{i=1}^{M} d_i \leq N_{tot} \tag{4}$$

where $s_i(H;r_i)$ is the transmit power required to achieve r_i in channel state H, which can be obtained from (1).

Denote by $\mathcal{D}_H$ the set of feasible decision vectors D such that (4) holds, and denote by $p(D|H)$ the probability that the scheduler's decision in channel state H is D. A policy is feasible *iff* $p(D|H) > 0$ only for $D \in \mathcal{D}_H$. The tail probability of user i is:

$$P_i(r_i;1) = \sum_{H} \sum_{D\in\mathcal{D}_H} d_i p(D|H)p(H) \tag{5}$$

Problem ($\mathbf{P1}$) is equivalent to:

$$(\mathbf{P2}) \quad \max_{\{p(D|H)\}} \ z(S_{tot}, N_{tot})$$
$$\text{s.t.} \quad P_i(r_i;1)/q_i \geq z(S_{tot}, N_{tot}),\ \forall i \tag{6}$$
$$\sum\nolimits_{D\in\mathcal{D}_H} p(D|H) = 1,\ \forall H \tag{7}$$

To solve ($\mathbf{P2}$), we first consider its dual problem, given by:

$$(\mathbf{P3}) \quad \min_{U,V} \ \sum\nolimits_{H} v_H$$
$$\text{s.t.} \quad 1 - \sum\nolimits_{i=1}^{M} u_i \leq 0,\ u_i \geq 0,\ \forall i$$
$$p(H) \sum\nolimits_{i=1}^{M} \frac{u_i d_i}{q_i} - v_H \leq 0,\ \forall D \in \mathcal{D}_H, H \tag{8}$$

where $U = \{u_i\}$ and $V = \{v_H\}$ are the vectors of Lagrangian multipliers associated with constraints (6) and (7), respectively.

Theorem 1 (Necessary Condition): If a policy Q^* is optimal, then in each channel state, Q^* only chooses the feasible decision vector D which maximizes $\sum_{i=1}^{M} u_i^* d_i/q_i$, where U^* is the set of optimal Lagrangian multipliers.

Proof: For a given U, (8) implies:

$$v_H(U) = p(H) \max_{D\in\mathcal{D}_H} \sum_{i=1}^{M} \frac{u_i d_i}{q_i} \tag{9}$$

Thus, ($\mathbf{P3}$) is equivalent to a minimization problem over U. Since the feasible region of U is non-empty and lower-bounded, there must exist at least one optimal solution, U^*. Moreover, optimality forces $\sum_{i=1}^{M} u_i^* = 1$. Let $\{p^*(D|H)\}$ denote an optimal solution to the primal problem ($\mathbf{P2}$). Using the complementary slackness theorem and substituting $v_H(U^*)$ by (9) yields:

$$p^*(D|H) \left(\sum_{i=1}^{M} \frac{u_i^* d_i}{q_i} - \max_{D\in\mathcal{D}_H} \sum_{i=1}^{M} \frac{u_i^* d_i}{q_i} \right) = 0$$

which implies that $p^*(D|H) > 0$ only if D is feasible and maximizes $\sum_{i=1}^{M} u_i^* d_i/q_i$. The theorem follows. ■

Theorem 2 (Sufficient Condition): If a policy Q^* i) only chooses the feasible decision vector D which maximizes $\sum_{i=1}^{M} u_i^* d_i/q_i$ in each channel state, where U^* is the optimal Lagrangian multiplier vector associated with (6), and ii) can balance the achieved normalized tail probability $P_i^*(r_i;1)/q_i = P_j^*(r_j;1)/q_j, \forall i,j$, then Q^* is an optimal solution.

Proof: Suppose that $P_i(r_i;1)$ is achieved by any given feasible policy. The optimal Lagrangian multiplier vector U^* satisfies $u_i^* \geq 0\ \forall i$ and $\sum_{i=1}^{M} u_i^* = 1$. Thus:

$$\begin{aligned} z(S_{tot}, N_{tot}) &= \min_i \left\{ \frac{P_i(r_i;1)}{q_i} \right\} \leq \sum_{i=1}^{M} u_i^* \frac{P_i(r_i;1)}{q_i} \\ &= \sum_{i=1}^{M} u_i^* \left(\sum_{H} \sum_{D\in\mathcal{D}_H} \frac{d_i}{q_i} p(D|H)p(H) \right) \\ &= \sum_{H} p(H) \left(\sum_{D\in\mathcal{D}_H} p(D|H) \left(\sum_{i=1}^{M} \frac{u_i^* d_i}{q_i} \right) \right) \end{aligned}$$

where the penultimate equality comes from (5). On the other hand, with equal normalized tail probabilities, the objective function value achieved by Q^* is given by:

$$\begin{aligned} z^*(S_{tot}, N_{tot}) &= \sum_{i=1}^{M} u_i^* \frac{P_i^*(r_i;1)}{q_i} \\ &= \sum_{H} p(H) \max_{D \in \mathcal{D}_H} \left\{ \sum_{i=1}^{M} \frac{u_i^* d_i}{q_i} \right\} \end{aligned}$$

Obviously $z(S_{tot}, N_{tot}) \leq z^*(S_{tot}, N_{tot})$, which means that any feasible policy cannot achieve a better objective function value than Policy Q^* does. The theorem follows. ■

We next proceed to an online algorithm which can find the optimal Lagrangian multipliers U^* as well as achieve the optimal value $z^*(S_{tot}, N_{tot})$. Theorem 2 provides a sufficient condition for a policy to be optimal. However, in most cases, the optimal Lagrangian multipliers are unknown. Define a set of functions on U as $f_i(U) = P_i^{Q(U)}(r_i;1)/q_i - \sum_{j=1}^{M} u_j P_j^{Q(U)}(r_j;1)/q_j, \forall i$, where $P_i^{Q(U)}(r_i;1)$ is the tail probability of user i achieved by a policy, $Q(U)$, which chooses a feasible decision vector that maximizes $\sum_{i=1}^{M} u_i d_i/q_i$ in each time slot. Break ties at random. It is easily shown that the optimal Lagrangian multipliers vector U^* is the zero of the functions $\{f_i(U)\}$. Moreover, the feasible region of U^* is $\{U : \sum_{i=1}^{M} u_i = 1, u_i \geq 0, \forall i\}$. This leads us to use the truncated Robbins-Monro (RM) algorithm [9], which can find the zero of an unknown function with the root region known. The basic idea of this method is to replace $f_i(U)$ by a *noise-corrupted* observation in the root-finding process. As U approaches U^*, $f_i(U)$ approaches zero, which implies that normalized tail probabilities are balanced. Hence, the corresponding policy $Q(U)$ is optimal by Theorem 2.

Our algorithm includes two stages in each time slot. The first stage is to determine the optimal decision vector $D(t)$ under Policy $Q(U(t))$, i.e., to solve the following problem:

$$(\mathbf{P4}) \max_{D \in \mathcal{D}_{H(t)}} \left\{ \sum_{i=1}^{M} \frac{u_i(t) d_i}{q_i} \right\}$$

If we consider $u_i(t)/q_i$ as the price that user i would be willing to pay to transmit in time slot t, then the system's objective can be viewed as maximizing the total revenue. With constraints on the total transmit power and on the total channels, (**P4**) is a two-constraint 0-1 knapsack problem, which is NP-complete. When the number of users is large, a greedy algorithm is an appropriate solution method, since it can achieve a reasonable approximation with a complexity of $O(M)$:

1: *Sort users in ascending order according to* $(u_i(t)/q_i)/s_i(H(t);r_i)$*;*
2: *Assign the transmit power to users in order until the total power or total channels run out.*

Note that multiuser diversity is realized here, i.e., the users with better channel conditions are more likely to be chosen due to less transmit power required.

The second stage is to update $U(t)$. We define the noise-corrupted observation of $f_i(U(t))$ as:

$$\hat{f}_i(U(t)) = \frac{d_i(t)}{q_i} - \sum_{j=1}^{M} \frac{u_j(t) d_j(t)}{q_j} \tag{10}$$

It can be shown that $E_H[f_i(U(t)) - \hat{f}_i(U(t))|U(t)] = 0$ with probability one. The second stage is:

3: $u_i(t+1) = u_i(t) - \alpha(t)\hat{f}_i(U(t)),\ \forall i$;
4: *If* $u_i(t+1) > 1$, *then set* $u_i(t+1) = 1$; *if* $u_i(t+1) < 0$, *then set* $u_i(t+1) = 0,\ \forall i$;
5: *Normalize* $\sum_{i=1}^{M} u_i(t+1)$ *to 1 by setting* $u_i(t+1) = u_i(t+1)/\sum_{i=1}^{M} u_i(t+1),\ \forall i$.

Choosing a step size $\alpha(t) = 1/(t+1)$ results in $U(t)$ converging to U^* with probability one [9]. The update of $U(t)$ can be interpreted as the process of users adjusting their prices to achieve fairness in normalized tail probabilities.

Finally, consider the minimum normalized tail probability achieved by this algorithm. Define the minimum normalized tail probability in time slot t as $z(t; S_{tot}, N_{tot}) = \min_i\{[\sum_{\tau=1}^{t} d_i(\tau)/q_i]/t\}$. As discussed above, as $t \to \infty$, the corresponding policy approaches the optimal policy $Q(U^*)$. As a result, $z(t; S_{tot}, N_{tot})$ approaches the optimal value $z^*(S_{tot}, N_{tot})$. Numerical results illustrating the convergence properties of our algorithm as well as the feasible region of (S_{tot}, N_{tot}) can be found in [10].

IV. The Case of $t_c > 1$

We now consider the case $t_c > 1$. In this case, the throughput depends not only on the current transmission rate, but also on the transmission rates in the last $t_c - 1$ time slots. In order to achieve good performance, the scheduler can no longer be *memoryless*, but should consider the transmission rates in past time slots when making the current scheduling decision. Moreover, if the current channel condition of a user is good relative to its own average channel quality, then the scheduler could transmit to it at a high rate to take advantage of channel fluctuation over time. Based on these observations, we develop an online algorithm. Our approach is to set a goal that all users achieve their target throughput r_i over a sliding window consisting of the last $0 \leq \Delta t < t_c$ time slots and the current slot. The target rate for user i at time t is thus:

$$\begin{aligned} y_i(t) &= \max[t_c r_i - \sum_{\tau=t-\Delta t}^{t-1} x_i(\tau), 0] \\ &= \max[t_c r_i - \Delta t m_i(t-1;\Delta t), 0] \end{aligned} \tag{11}$$

where $m_i(t-1;\Delta t)$ is the average throughput over the last Δt time slots seen at $t-1$. If $\Delta t = 0$, the target rate is $t_c r_i$, in which case users ignore the achieved rates in the preceding time slots; the resulting algorithm is identical to the optimal policy in the case of $t_c = 1$ except that r_i is replaced by $t_c r_i$. As Δt increases, the influence of transmission rates in the preceding time slots increases; in particular, when $\Delta t = t_c - 1$, the target rate is exactly the amount required to enable the average throughput seen in the current time slot to meet the rate threshold. We call this algorithm *Maxmin normalized tail probability* (MMNTP):

1: *Calculate the target rate $y_i(t)$ using (11) and the transmit power $s_i(t)$ using (1);*
2: *Set the residual power equal to the total transmit power and the residual channels equal to the total number of channels;*
3: *Sort users in ascending order according to the value of $(u_i(t)/q_i)/(h_i(t)/r_i)$, and assign the first user a transmit power of $\min\{s_i(t), \text{the residual power}\}$.*
4: *Continue to assign the residual transmit power to the following users until the total transmit power or the total channels run out; if the residual power > 0 after all users achieve their targets, assign the residual power to the user with the best channel condition.*
5: *Update $U(t)$ using the method in the case of $t_c = 1$; update $m_i(t-1;\Delta t)$.*

In addition to the use of (11), there are another two distinctions between this algorithm and the optimal algorithm for $t_c = 1$. One is that in Step 4 the residual transmit power is assigned to the user even though it is not enough to support the required transmission rate. The other difference lies in Step 5. When the truncated Robbins Monro algorithm is used, the noise-corrupted observation in (10) should change to:

$$\hat{f}_i(U(t)) = \frac{I(m_i(t;t_c) \geq r_i)}{q_i} - \sum_{j=1}^{M} \frac{u_j(t) I(m_j(t;t_c) \geq r_j)}{q_j}$$

so that the normalized tail probabilities can be equalized.

We first investigate the effect of Δt on the performance. We simulate a system with 5 homogeneous users ($M = 5$). Users are equidistant from the base station and thus have the same distance-based attenuation, $\bar{h}_i = \bar{h}\ \forall i$. The channel processes are modeled by Rayleigh fading and assumed independent over users and over time. User rates are normalized by $W/(\gamma(I_0 + N_0))$, and the normalized rate thresholds $\tilde{r}_i = 1.5 S_{tot}\bar{h}_i/M\ \forall i$ are identical for all users. User also specify the tail probability thresholds, $q_i = 0.7\ \forall i$. Figure 1 shows that the minimum *normalized* tail probability achieved by MMNTP for varying Δt under different values of t_c. For a given t_c, the minimum normalized tail probability achieved by MMNTP is unimodal in $\Delta t/t_c$; the peak is marked by a big cross. When $\Delta t = 0$, users require a transmission rate of $t_c r_i$ with the ambition of making the average throughputs seen in the next t_c time slot meet the rate thresholds. This case can take the most advantage of channel variation among users, but may not be optimal without considering the transmission rates in the preceding time slots. In contrast, when $\Delta t = t_c - 1$, users request a rate of $t_c r_c - m_i(t-1; t_c - 1)$ to make the average throughput seen in the current slot meet the threshold. This case takes the least advantage of channel variation among users and ignores the effect of the current rate on average throughputs seen in the following time slots. Both of these extreme cases usually cannot achieve the optimal value; and the optimal Δt lies between 0 and $t_c - 1$ in most cases.

We now consider optimality in terms of minimum normalized tail probability. To do so, we vary the number of users M, set $\tilde{r}_i = S_{tot}\bar{h}/M\ \forall i$ and $t_c = M$. It is easily shown that in this symmetric case the optimal algorithm is Round-Robin, which schedules each user to be transmitted once every

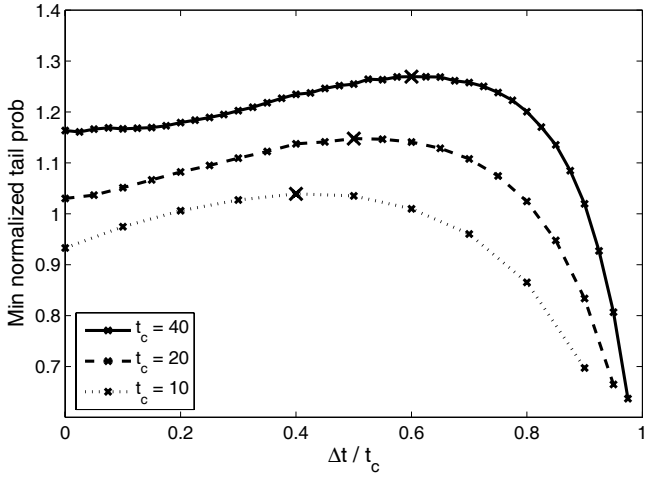

Fig. 1. Minimum normalized tail probability vs. Δt

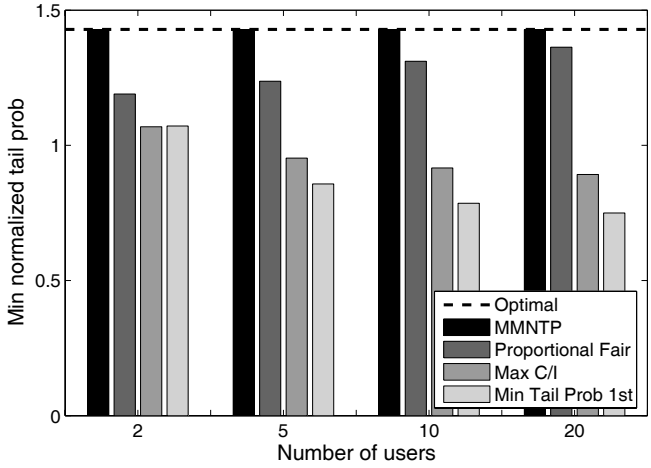

Fig. 2. Minimum normalized tail probability achieved by different algorithms

t_c time slots, and the resulting tail probability is 1 for all users. It follows that the optimal normalized tail probability is $1/0.7 \approx 1.43$. We will compare the performance of MMNTP with two well-known algorithms which do not consider short-term performance: the proportional fair algorithm and the max C/I algorithm which always assigns a time slot to the user with the best channel condition. We also compare to a simple forcing algorithm, called *minimum tail probability first*, which assigns a time slot to the user with the minimum current normalized tail probability. Figure 2 shows the minimum normalized tail probabilities achieved by different algorithms for different numbers of users. The red dashed line represents the optimal value. For all values of M, MMNTP can achieve optimality by setting $\Delta t = t_c - 1$. This is because it takes into account the transmission rates in the preceding time slots, which results in users taking turns to require a transmission rate of $t_c r_i$. Hence, the algorithm achieves the same minimum normalized tail probability as the Round-Robin algorithm. The proportional fair algorithm can achieve good but suboptimal performance in terms of minimal normalized tail probability because it also considers the transmission rates in the past in power allocation, which positively impacts the performance.

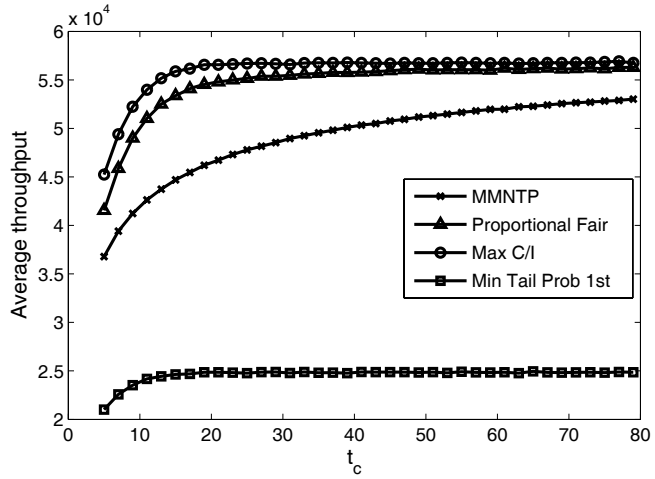

Fig. 3. Scenario 1: long-term average throughput vs. t_c

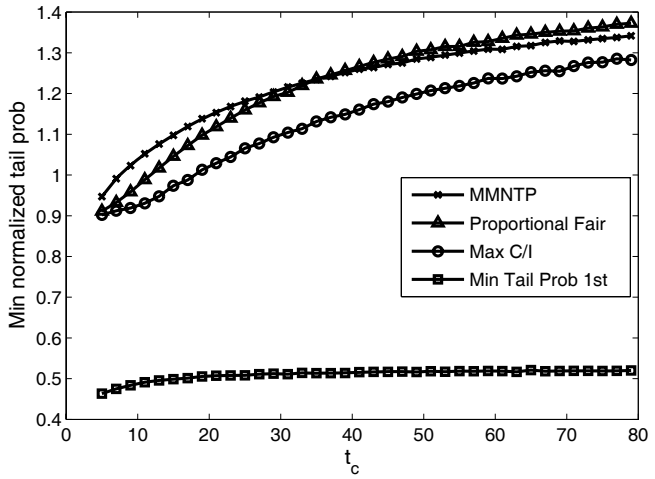

Fig. 4. Scenario 1: minimum normalized tail probability vs. t_c

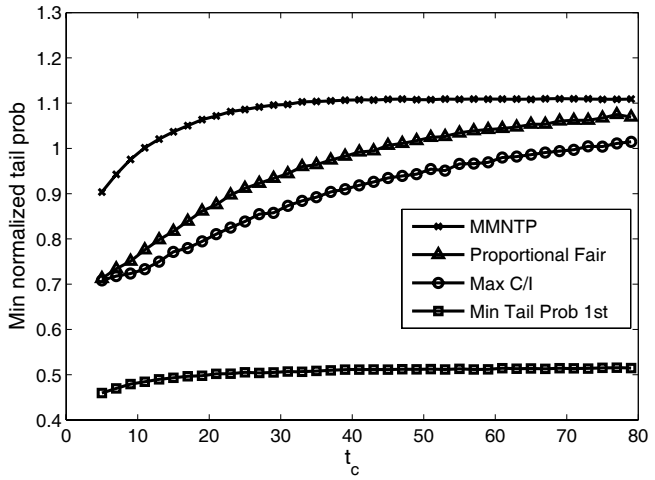

Fig. 5. Scenario 2: minimum normalized tail probability vs. t_c

Tail probability, of course, is only one performance measure. The proportional fair and max C/I algorithms are focussed on long-term average throughput, and so should not necessarily be expected to achieve high minimum normalized tail probabilities. Figure 3 shows the long-term average throughputs achieved by different algorithms. Max C/I achieves the highest long-term average throughput as it takes most advantage of multiuser diversity, with the proportional fair algorithm not far behind. As expected, MMNTP achieves a significantly lower average long-term throughput. This illustrates the tradeoff between the minimum normalized tail probability and the long-term average throughput.

Finally, we compare MMNTP to other algorithms under varying t_c. A 5-user system ($M = 5$) is simulated, and Δt is set to $\lfloor t_c/2 \rfloor$. Moreover, we set the maximum transmission rate to $t_c r_i$ for user i to improve the achieved performance by all algorithms. We next consider four scenarios. Scenario 1 is a completely symmetric case, which has been introduced above. All users are set 400m away from the base station, requiring the same rate thresholds, $\tilde{r}_i = 1.5 S_{tot} \bar{h}/M \ \forall i$ with $\bar{h}$ being the distance-based attenuation at 400m, and the same tail probability thresholds, $q_i = 0.7 \ \forall i$. Scenario 2 changes equal tail probability thresholds in Scenario 1 to uniformly distributed tail probability thresholds, i.e., $q_i \sim U(0.5, 0.9)$. Scenario 3 changes the equal distance in Scenario 1 to uniformly distributed distance $U(400, 800)$, and sets $\tilde{r}_i = 0.8 S_{tot} \sum_{i=1}^{M} \bar{h}_i / M^2 \ \forall i$ with $\bar{h}_i$ being the distance-based attenuation experienced by user i to keep rate thresholds same for users. Scenario 4 changes equal rate thresholds in Scenario 1 are changed to uniformly distributed rate thresholds, i.e., $\tilde{r}_i \sim U(1, 1.8) \times S_{tot} \bar{h}/M$.

Figure 4 shows the minimum normalized tail probabilities achieved by different algorithms with varying t_c in Scenario 1. For all algorithms, the minimum tail probabilities monotonically increase with t_c. This is reasonable because a longer observation time provides the scheduler with a better chance to exploit multiuser diversity and channel fluctuation. In particular for MMNTP, as t_c increases, the transmission rates required by users in (11) increase accordingly. As a result, users with better channel conditions can obtain a higher data rate if selected to transmit. Thus, multiuser diversity is utilized more fully and the performance is improved. The figure also shows that MMNTP outperforms other algorithms. In this symmetric case, the Max C/I algorithm is in fact MMNTP with $\Delta t = 0$; therefore, it is not surprising that MMNTP outperforms Max C/I.

Figure 5 shows the minimum normalized tail probabilities varying with t_c in Scenario 2. In this case, MMNTP rapidly approaches the upper bound of the minimum normalized tail probability, which comes from the inverse of the maximum tail probability threshold, i.e., $1/0.9 \approx 1.11$. In this asymmetric environment, MMNTP significantly outperforms other algorithms for all t_c displayed in this figure; this is because it can effectively balance the normalized tail probabilities among users. The minimum tail probability first algorithm can also balance the achieved normalized tail probabilities, but in an aggressive manner which results in a poor performance.

The advantage of MMNTP is more notable when users become more heterogeneous, as in Scenario 3 where the differences in distance away from the base station lead to significant differences between users. Figure 6 shows that MMNTP achieves much better minimum normalized tail probability than other algorithms do under this setting. The

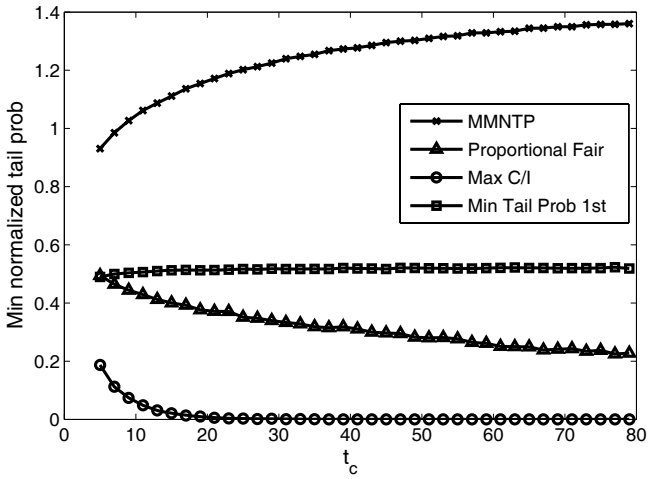

Fig. 6. Scenario 3: minimum normalized tail probability vs. t_c

Max C/I algorithm and the proportional fair algorithm display poor performance, even worse than the Min tail probability first algorithm. This is because these two algorithms do not balance the tail probabilities of users. As a result, the user farthest away from the base station receives a very low tail probability, which drags down the achieved performance. Finally, as expected, MMNTP can achieve better minimum normalized tail probabilities than other algorithms in Scenario 4 with heterogeneous rate thresholds (results not shown).

V. Conclusions

In this paper, we have studied a scheduling problem with performance guarantees in the form of the probabilities that short term throughputs exceed user specified thresholds. For the case in which the time interval of interest is one time slot, we presented necessary and sufficient optimality conditions. The optimal policy can be considered to maximize total system revenue while balancing the achieved normalized tail probabilities of users. An online algorithm was proposed to achieve the optimal value by dynamically adjusting weights, which can be interpreted as the prices that users are willing to pay for service satisfying their rate requirements. For the case in which the time interval is more than one time slot, we developed an algorithm which effectively combines efficiency with fairness. Simulation results show that it achieves higher minimum normalized tail probability at the expense of long-term average throughput, with the differences becoming larger in the case of heterogeneous users.

References

[1] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, pp. 451-474, 2003.

[2] Y. Liu and E. Knightly, "Opportunistic fair scheduling over multiple wireless channels," in *Proc. INFOCOM 2003*, vol. 2, pp. 1106-1115, 2003.

[3] V. Tsibonis and L. Georgiadis, "Optimal downlink scheduling policies for slotted wireless time-varying channels," *IEEE Trans. Wireless Commun.*, vol. 4, pp. 1808-1817, 2005.

[4] S. Borst and P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization," in *Proc. INFOCOM 2001*, vol. 2, pp. 22-26, 2001.

[5] P. Zhang and S. Jordan, "Throughput guarantee targeted hybrid scheduling for downlink WCDMA data networks," in *Proc. WCNC 2006*, vol. 3, pp. 1699-1704, 2006.

[6] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[7] Y. Liu, S. Gruhl, and E. W. Knightly, "WCFQ: an opportunistic wireless scheduler with statistical fairness bounds," *IEEE Trans. Networking*, vol. 2, pp. 1017-1028, 2003.

[8] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Scheduling in a queuing system with asynchronously varying service rates," *Probab. Eng. Inf. Sci.*, vol. 18, no. 2, pp. 191-217, 2004.

[9] H.-F. Chen, *Stochastic Approximation and Its Applications*. Springer, 2002.

[10] N. Chen and S. Jordan, "Downlink scheduling with probabilistic guarantees on short-term average throughputs," in *Proc. 2008 IEEE Wireless Commun. Networking Conf. (WCNC)*, 2008.