Scheduling with QoS Constraints over Rayleigh Fading Channels

A. Eryilmaz and R. Srikant

Abstract— **We consider the problem of scheduling packets from multiple flows over a Rayleigh fading wireless channel. Recently, there has been much interest in opportunistic scheduling, i.e., scheduling packets from a user who has the highest SNR (signal-to-noise ratio), to maximize the network's throughput. In this paper, we compare the throughput achievable under fair opportunistic scheduling (i.e., a modification of opportunistic scheduling to ensure fair resource allocation) with the throughput under time-division multiplexing (TDM) scheduling. Using large deviations to characterize the probability that the QoS constraint (an upper bound on delay) is violated, we numerically compare the performance of the two scheduling algorithms under various channel conditions. We show that the opportunistic scheduler outperforms the TDM scheduler when the number of users is small but the TDM scheduler performs better when the number of users exceeds a threshold which depends on the channel parameters.**

I. Introduction

In the recent years, the idea of *opportunistic scheduling* [13], where the channel gain fluctuations are exploited by serving the flow that has the best channel state, has attracted much attention. When there are several users in the network and the channel state for each of them fades independently of each other, it is likely that at least one of them has a good channel state, and hence this type of gain is called the *multi-user diversity gain*. Systems such as HDR over CDMA [1], and EDGE over GSM have been developed that takes advantage of the channel fluctuations in the multi-user context. Even if the the wireless channel does not fluctuate fast enough, channel fluctuations can be artificially induced in order to improve the overall performance [13], [6]. However, large fluctuations result in large mean delay expressions, hence the Quality of Service (QoS) constraints becomes more and more difficult to satisfy with increasing number of users.

There are two issues with opportunistic scheduling: *fairness* and *Quality-of-Service (QoS) degradation* caused by flows having to wait till their channel becomes sufficiently good for them to receive service. The fairness problem has been studied extensively in [12], [4], [13], where instead of serving the user with the best channel state, the instantaneous bit-rate is weighted by some factor to achieve to achieve *proportional fairness* or some other fairness criterion. However, the latency experienced by the flows might still be bad under a fair allocation, for most of the service may be given to a single flow with a good channel. Hence, there is a trade-off between the multi-user diversity gain exploited in opportunistic scheduling and the delay experienced by the flows. We study this tradeoff between throughput gain and QoS degradation in this paper. Specifically, we compare the performance of fair opportunistic scheduling with the performance of TDM scheduling, where by TDM scheduling we mean a resource allocation scheme where each user is guaranteed a fixed fraction of the time slots periodically.

Suppose there is a delay constraint in the form of a small tail probability of the steady state delay being larger than, say 100 msecs. The arrival rate that can be supported under this constraint will be called the *supportable* rate. The above delay condition is reasonable both for real-time traffic and for the efficient operation of TCP. The latter is due to the fact that TCP re-transmits a packet for which an ACK has not been received within a few round trip times (RTT), which are typically of the order of 100 msecs. If the queueing delay due to opportunistic scheduling is large, then repeated timeouts will occur and cause degradation in end-to-end TCP throughput [5].

With a delay constraint of this sort on each of the n flows, it should be obvious that when n becomes too large, the opportunistic scheduler should perform much worse than the TDM scheduler, because of the randomness in the service process caused by opportunistic scheduling. However, it is not clear whether the advantage of multi-user diversity provides better performance for small enough n compared to the TDM type of schedulers. And if so, can a threshold on n be specified such that above the threshold the TDM scheduler should be preferred and below which vice versa?

In this paper, we evaluate the probability of QoS violation using large deviations theory [3] and perform numerical computations as well as simulations using typical values for the system parameters. We observe that the opportunistic scheduler takes advantage of the multi-user diversity for small enough n and supports much higher total throughput levels compared to the TDM scheduler. The optimum (in the sense of maximizing the total supportable throughput) number of users that can be supported as well as the threshold up to which opportunistic scheduler supersedes the TDM scheduler can be clearly identified in our results. Simulation results, which turn out to be in exact agreement with the theoretical results, are presented afterwards.

In [4], opportunistic scheduling under a different type of QoS constraint is studied. A number of other works suggest the use of queue length information in scheduling the transmissions [11], [10], [9], [2] for throughput optimality.

This work was supported by a Vodafone Fellowship, NSF Grant ITR 00-85929, DARPA Grant F30602-00-2-0542 and AFOSR URI F49620-01-1-0365

A. Eryilmaz and R. Srikant are with the Department of Electrical and Computer Engineering and the Coordinated Science Lab, University of Illinois at Urbana,Champaign, IL, 61801. Email: {eryilmaz,rsrikant}@uiuc.edu

The basic idea is to give priority to the flows that have a longer backlog. However, these schedulers do not provide isolation among flows and may be unfair to those flows that have low arrival rates. In [7], [8], QoS constrained scheduling is considered for the case where the channel conditions and arrival rates are i.i.d. among the users. The author concentrates on the MaxQueue rule for ON-OFF channels in detail. This rule gives service to the flow with the maximum backlog over all flows whose channel state is ON. The *effective capacity* of this rule is derived and compared with the greedy rule, which serves the user with the best channel in each slot. It is shown that there is a non-negligible difference between the effective capacity of the two rules as the QoS constraint (total queue length tail distribution) gets more strict. However, in our paper we consider the more realistic Rayleigh fading channel and delay constraints and study the characteristics of the supportable throughput with increasing number of users.

The paper is organized as follows: Section II introduces the system model and the large deviations result that will be used to evaluate the QoS provided to the users of the wireless network. In Section III, opportunistic and TDM schedulers are described and analyzed for the case where the channel state processes for all the users are i.i.d. The effect of channel spectrum splitting is considered in Section IV. Section V generalizes the analysis of the opportunistic scheduler to the case where the channel state processes of the different users are independent, but not identical. Simulation results are presented in Section VI. And finally, conclusions are drawn in Section VII.

II. SYSTEM MODEL

Consider a system that consists of n flows sharing a common time-varying medium. A typical example of such a system is the *downlink* in cellular wireless communication. Let the time be slotted, with each slot being of size T_c msec. The bandwidth available for communication is W MHz. The number of arrivals to each flow is assumed to be a constant and equal to λ packets per time slot. It should also be noted that the assumption of symmetric arrivals to each queue is made for ease of presentation. All of the analysis holds for the case when each flow has a different arrival rate as well as a different delay constraint. Here, we assume each *packet* is of size $W \times T_c$ nats for convenience. We have introduced the notion of packet for notational convenience; however, the network is allowed to serve a fraction of a packet in a time slot. In fact, in HDR and other real wireless systems, the amount served in each time slot is quantized in a manner that is independent of the size of the arriving packets. As will become clear later, here we assume that each user who is chosen for service in a time slot is given an amount of service (in bits) equal to the information theoretic capacity of the channel under the channel condition at that time instant. We simply measure this amount of service in units of packets which has been defined above.

Next, we consider a single flow and describe the queueing-theoretic model that we will use to evaluate the QoS violation for that flow. Since we are considering the packets as being infinitely divisible, i.e., as a fluid, we define the delay experienced in a time slot t to be the maximum delay experienced by any bit served in the time slot. Let D_t denote this delay in slot t. We assume that even if a bit is served within the same slot that it entered the queue, it has experienced a delay of 1. Define $A(t)$ and $S(t)$ as the cumulative arrivals and departures within $[1, t]$, respectively. Obviously, due to the deterministic nature of arrivals, we have $A(t) = \lambda t$. Thus, the event $D_t > d$, is equivalent to the event $S(t) - A(t-d) \leq 0$. Hence we can write the following

$$\begin{aligned} P(D_t > d) &= P(S(t) - A(t-d) \leq 0) \\ &= P(A(t) - S(t) \geq \lambda d). \end{aligned}$$

But notice that $A(t) - S(t)$ is nothing but the length of a queue at time t, given that the queue was empty at the beginning of time slot 1. Let us denote this queue length at time t as $Q[t]$. If we define QoS violation to be the event that the delay exceeds some threshold d, then the steady-state probability of QoS violation can be described as

$$\lim_{t\to\infty} P(D_t > d) \quad = \quad \lim_{t\to\infty} P(Q[t] \geq \lambda d)$$

We denote the (random) amount of offered service at time slot k by $\mu[k]$. Therefore, the total offered service up to time t is given by $M(t) = \sum_{k=1}^{t} \mu[k]$. For ease of representation, let us define $X(t) = A(t) - M(t) = \sum_{k=1}^{t} x[k]$, where $x[k] = \lambda - \mu[k]$. Then we can utilize a well-known upper bound of large deviations which is also known to be asymptotically tight [3]:

$$P(q[0] \geq \lambda d) \quad \leq \quad e^{-\lambda d \inf\limits_{y\geq 0} \left\{ \frac{\mathcal{I}_x(y)}{y} \right\}}, \qquad (1)$$

where $\mathcal{I}_x(\theta) := \theta x - \Lambda_x(\theta)$ where $\Lambda_x(\theta)$ is defined as

$$\Lambda_x(\theta) = \lim_{t\to\infty} \frac{1}{t} \sum_{k=1}^{t} \ln E\left(e^{\theta x[k]}\right).$$

Precise conditions on the process $\{x[k]\}$ for the above result to hold are given in [3].

Suppose that the QoS constraint is of the form

$$\lim_{t\to\infty} P(D_t > d) \quad \leq \quad 10^{-\varepsilon},$$

for some fixed $\varepsilon > 0$. Then, from (1), the QoS constraint can also be expressed as

$$\lambda \inf_{y\geq 0} \left\{ \frac{\mathcal{I}_x(y)}{y} \right\} \quad \geq \quad \delta, \qquad (2)$$

where $\delta := \frac{\varepsilon}{d} \ln(10)$. The following fact is also established in [3].

Fact 1: The condition (2) is equivalent to $\Lambda_x\left(\frac{\delta}{\lambda}\right) \leq 0$.

Since the arrival process is deterministic, the condition in Fact 1 can be equivalently written as

$$\Lambda_\mu\left(-\frac{\delta}{\lambda}\right) \leq -\delta. \tag{3}$$

Given the scheduler and the delay constraint parameters d and ζ, this inequality specifies the throughput λ that is supportable. Next, we use this model to analyze the opportunistic and TDM scheduling policies.

III. Analysis of Scheduling Policies

The aim of this section will be to analyze the maximum allowable total throughput as a function of the number of flows, n, for different scheduling policies.

A. Analysis of Opportunistic Scheduling

The transmission medium is assumed to be time-varying independently among different time slots, but is constant within a given time slot. Let h denote the random variable representing the channel gain in a given slot. Then the rate that can be achieved over this channel is given by $WT_c \ln(1 + SNR \cdot h)\, nats/sec = \ln(1 + SNR \cdot h)\, packets/slot$. Notice that achieving this rate implies the availability of channel side information at the receiver and the transmitter. Here, SNR is the received signal to noise ratio assuming a constant channel gain of 1. For a Rayleigh fading channel, h is exponentially distributed with unit mean. Furthermore, we assume that the channel gains for different flows are independently and identically distributed. Since the channel gains are i.i.d., fairness considerations such as proportional fairness do not enter into the picture and thus, we can use the following rule to describe the opportunistic scheduler we will consider.

Opportunistic Scheduler: At each time slot, choose the flow that has the best channel gain for transmission. In other words, if $h^{(i)}[k]$ denotes the channel gain for flow $i \in \{1, \cdots, n\}$ in slot k, then only flow $i^*[k]$ is allowed transmit in slot k, where $i^* := \arg\max_i h^{(i)}[k]$. ◇

Proposition 1: The service rate distribution of any given flow, say i, operating under the Opportunistic Scheduling Policy is given by

$$P(\mu^{(i)} \leq r) \quad = \quad \frac{1}{n}(P(Z^{(i)} \leq r))^n + 1 - \frac{1}{n} \qquad \forall r \geq 0,$$

where $Z^{(i)} = \ln\left(1 + SNR \cdot h^{(i)}\right)$.

Proof: The result trivially follows from the i.i.d assumption on the channel processes. ■

The probability distribution of $Z^{(1)}$ can easily be computed to be $P(Z^{(1)} \leq r) = 1 - e^{-\frac{1}{SNR}(e^r - 1)} \quad \forall r \geq 0$.

Now that we get an expression for the service rate distribution, the computation of $\Lambda_\mu(\cdot)$ is straight forward. Due to the difficulties in getting a closed form expression, we will numerically solve (3) and obtain the total supportable throughput for the network as a function of the number of users under the opportunistic scheduling policy. These results will be demonstrated in a later section.

B. Analysis of TDM Scheduling

TDM Scheduler: Each flow is periodically scheduled to transmit once every n slots, regardless of the channel gain realizations. ◇

One way of approach is to assume that the system started at time 1 and the flow assigned to that slot is uniformly and randomly chosen among the n flows. Once this choice is known the rest of the assignments are periodically done according to the TDM scheduler described above. Equivalently, given that we are interested in the moment generating function of flow j, it is equally possible that the first time slot to which flow j is assigned can be one of $1, 2, \cdots, n$.

We are interested in $\Lambda_{\mu^{(j)}}(\theta)$, which is defined as

$$\lim_{t\to\infty} \frac{1}{t} \ln\left(\frac{1}{n}\sum_{i=1}^{n} E\left[e^{\theta \sum_{k=1}^{t} \mu^{(j)}[k]} \,|\text{Serve flow j at time i}\right]\right)$$

where $\mu^{(j)}[k]$ denotes the amount of service offered to the flow i at time slot k. The log-moment generating function of $\{\mu^{(j)}[k]\}$ under the TDM scheduling policy is given by the following proposition.

Proposition 2: For the TDM scheduler

$$\Lambda_{\mu^{(j)}}(\theta) \quad = \quad \frac{1}{n}\Lambda_Z(\theta),$$

where $\Lambda_Z(\theta)$ is the log moment generating function of $Z^{(1)} := \ln\left(1 + SNR \cdot h^{(1)}\right)$.

Proof: Due to the nature of the scheduler, given that flow j is first served at time i implies that only $\mu^{(j)}[i], \mu^{(j)}[i{+}N], \mu^{(j)}[i{+}2N], \cdots$ can take positive values, while the service at the rest of the time slots must be zero.

This observation enables us to write $\Lambda_{\mu^{(j)}}(\theta)$ as

$$\lim_{t\to\infty} \frac{1}{t} \ln\left(\frac{1}{n}\sum_{i=1}^{n}\left(E\left[e^{\theta Z^{(1)}}\right]\right)^{\lfloor \frac{t-i}{n} \rfloor + 1}\right), \tag{4}$$

where $Z^{(1)} = \ln(1 + SNR \cdot h)$. It is not difficult to see that the following upper and lower bounds can be applied to the floor function: for $i \in \{1, 2, \cdots, n\}, \quad t \geq 1:$

$$\frac{t}{n} - 2 \leq \lfloor \frac{t}{n} \rfloor - 1 \leq \lfloor \frac{t-i}{n} \rfloor \leq \lfloor \frac{t}{n} \rfloor \leq \frac{t}{n}.$$

If we use these bounds on (4), we get for each $j \in \{1, \cdots, n\}:$

$$\lim_{t\to\infty}\left(\frac{1}{n} - \frac{1}{t}\right)\Lambda_Z(\theta) \leq \quad \Lambda_{\mu^{(j)}}(\theta) \quad \leq \lim_{t\to\infty}\left(\frac{1}{n} + \frac{1}{t}\right)\Lambda_Z(\theta)$$

which gives the desired result. ■

C. Numerical computation and Comparison

In this section, we numerically compare the performance of the two scheduling policies using reasonable values for the channel parameters (such as the SNR value, coherence time, delay constraint, etc.). We let $T_c = 1$ msec and $W = 1.25$ MHz, which are reasonable values for the coherence time and bandwidth. For many applications, typically the delay experienced by each flow should not be not more than 100 msec. Examples applications include TCP where excessive delays can cause frequent time-outs, transmission of streaming video where excessive jitter should be avoided while a small amount of jitter can be tolerated by an appropriately designed play-out buffer. Hence, in all the simulations in this paper, we take $d = 100$ and the QoS constraint is always assumed to be as follows: we require that the probability that the delay exceeds d is constrained to be less than or equal to 10^{-5}. Modifying this threshold to a less stringent level (e.g. 10^{-3}) does not change the performance curves significantly. Figure 1 depicts the total

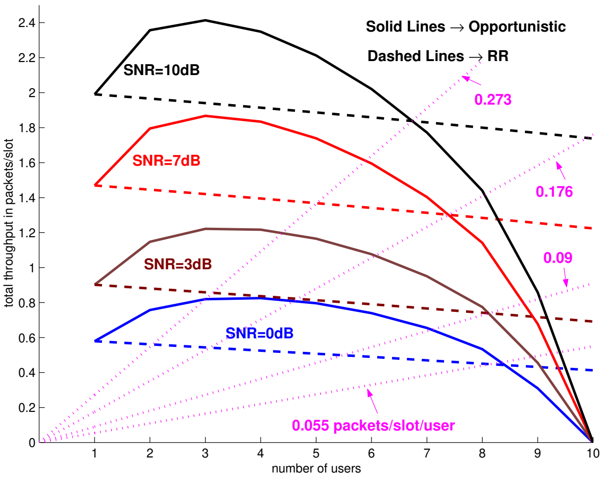

Fig. 1. Total throughput as a function of the number of users n, for the two schedulers with various SNR levels.

supportable throughput (i.e., the sum of the supportable rates for all users in the network) versus number of users curve for different levels of SNR. In the figure, results with different SNR levels are color coded and the solid lines represent the opportunistic scheduler case, whereas the dashed lines are for the TDM scheduler. For each SNR level, we have drawn a dotted line whose slope indicates a threshold throughput per user such that for any rate above this threshold, the opportunistic scheduler can admit more users with that guaranteed rate per user compared to the TDM scheduler. For any rate below the threshold, the TDM scheduler performs better. We make the following observations:

- The total throughput is larger under opportunistic scheduling when the number of users is small. On the other hand, the TDM scheduler performs better when the number of users is large. This is due to the fact that, under opportunistic scheduling, the initial increase in the number of users improves the multi-user diversity gain more than it hurts the delay constraint. However, as n get larger, the fluctuations in the offered service under opportunistic scheduling affect the performance much more severely and supportable throughput drops sharply. Such a phenomenon does not occur in the TDM scheduler because it does not exploit the multi-user diversity gain.
- Whether one should use opportunistic scheduling or not depends upon the throughput that the wireless service provider expects to provide to the users of the network. When the SNR is 0 dB, from Figure 1, when the required throughput per user is 0.055 packets-per-slot (which is approximately 100 kbps), the network is indifferent between using opportunistic scheduling or TDM scheduling. Thus, if the required throughput is smaller than this, then the network should use opportunistic scheduling if the number of users is below eight and it should use TDM when the number of users is larger. On the other hand, if the required throughput is larger than 0.055 packets per slot, then the network should use opportunistic scheduling and use admission control to restrict the number of users to be at a supportable level. Similar conclusions can be drawn for other SNR levels from the figure.

IV. Scheduling in Time and Frequency

In actual systems, a single carrier may possess more than one frequency band that fades independently, each of which is referred to as a *channel*. Then the question is how to distribute the services among the users over these channels. In this section, we will study modified versions of the opportunistic and the TDM scheduler of Section III.

A. Analysis of the Modified Opportunistic Scheduler

Suppose there are M channels available for transmission. Furthermore, we assume that the channel gains for each user over different channels are identically and independently distributed as exponential random variables with unit mean. Let $h_m^{(i)}[k]$ denote the channel gain random variable for user i over channel m at time slot k. Then the user assignments to channel for a given slot is given by

Modified Opportunistic Scheduler: For a given time slot, and over a given channel, say m, serve the user that satisfies $i^* \in \arg\max_j h_m^{(j)}$. ◇

Then it is easy to see that the service offered to queue i is given by $\mu^{(i)} = \sum_{m=1}^{M} \mu_m^{(i)}$, where $\mu_m^{(i)}$ denotes the service offered to queue i over channel m. Therefore, we have

$$\Delta_{\mu^{(i)}}(\theta) \quad := \quad \ln E\left[e^{\theta\mu^{(i)}}\right] = M\Delta_{\mu_m^{(i)}}(\theta).$$

The distribution of $\mu_m^{(i)}$ is exactly in the same form as the distribution of μ^i of Section III-A. Here, it should be noted that we assume that the SNR is the same over the different channels, which is reasonable.

Hence, we can utilize the derivation of Section III-A to compute $\Delta_{\mu^{(i)}}(\theta)$. The numerical computation of the corresponding supportable rates will be provided in Section IV-C, along with comparisons with earlier schedulers as well as new schedulers.

B. Analysis of Modified TDM Scheduling

A natural modification of the TDM Scheduler studied in Section III-B is given below.

MODIFIED TDM SCHEDULER: Suppose the users are numbered from 1 to n, and the channels are numbered from 1 to M. Then, in time slot 1, the scheduler starts from user 1 and assigns the first available channel to it (then that channel is no longer available to other users in that time slot), and continues to assign users to channels sequentially until either all the n users are scheduled, or the available channels in the current slot is exhausted. In the former case, the scheduler wraps around and continues the same assignment procedure starting from user 1. In the latter case, no other assignment is made and in the next time slot the same assignment procedure continues starting from the user that was blocked due to unavailability of channels in the current slot. $\diamond$

A closer look into the operation of the above scheduler reveals the fact that, each user will be (roughly) periodically assigned to a channel once every $\frac{n}{M}$ slots. This is similar to the situation in Section III-B and similar calculations yield

$$\Delta_{\mu^{(i)}}(\theta) = \frac{M}{n} \ln E(e^{\theta Z^{(1)}}),$$

where $Z^{(1)}$ was defined earlier.

C. Numerical Computation and Comparison of the Modified Schedulers

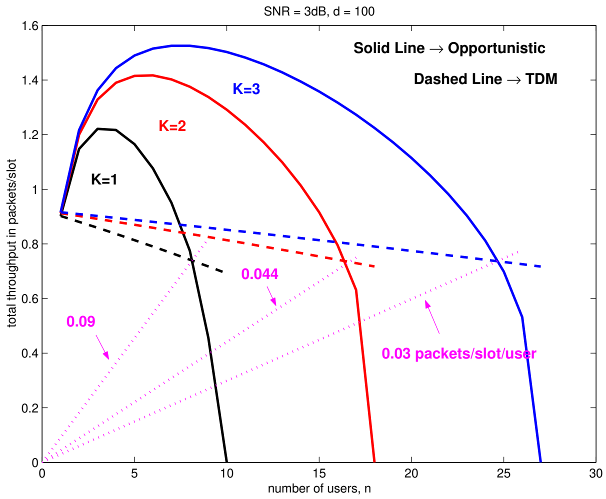

Fig. 2. Performance of the modified schedulers with varying number of sub-bands

In order to exhibit the advantage of implementing the opportunistic scheduler over multiple independently varying channels instead of a single frequency band of the same total size, we consider a frequency band of size W Hz and partition it into K sub-bands of equal size with $K \in \{1, 2, 3\}$. Then, we implement the modified scheduler for different values of K. Figure 2 depicts the results as well as the threshold on the expected throughput level per user, above which the opportunistic scheduler is more advantageous compared to TDM. Our first observation is that the threshold level decreases with increasing K. Also, we can clearly see that as K increases the opportunistic scheduler performs significantly better even for smaller levels of throughput per user, whereas the TDM scheduler changes only slightly. Hence, it is more advantageous in terms of total throughput to split the available frequency band into independently varying sub-bands and then use opportunistic scheduling.

V. FAIR SCHEDULING

In the previous sections, we have assumed that mean signal-to-noise ratio (denoted by SNR) is the same for all users. However, this is not very realistic and in this section, we study the situation where the SNR could be different for different users. Here, we will only study the single channel case, while the extension to M channels can be carried in a straightforward manner along the same lines.

In a wireless medium, the channel conditions of different users need not be symmetric, because some users may experience lower channel gains due to their distances from the base station. This situation can be captured in our model by assuming that user i's channel gain, denoted by $h^{(i)}$, is exponentially distributed with mean $1/m_i$ for some $m_i > 0$. Hence, we have $P\left(h^{(i)} \leq x\right) = 1 - e^{-m_i x}$. In order to compensate for the asymmetry of the channel processes, we add weighting factors to the scheduling decisions as in [12], [4]. Further, each user may demand a different delay constraint since some may require more stringent delay characteristics than others. We may take this into account by defining $d^{(i)}$ to be the delay constraint for user i. This in turn results in different $\delta^{(i)}$ values for different i.

FAIR OPPORTUNISTIC SCHEDULER: At each time slot, the service is given to the user that satisfies

$$\arg \max_{1 \leq j \leq n} \beta^{(j)} \mu^{(j)} + \nu^{(j)},$$

where $\{\beta^{(j)}, \nu^{(j)}\}$ is a set of non-negative constants associated with different users. Here, it should be noted that proportional fairness [12] or other forms of fairness [4] can be achieved in this model if these parameters are chosen appropriately. $\diamond$

We want to answer the following question: given a number of users and their associated channel processes, what is the maximum set of rates that can be supported? To answer this question, we start by modifying the multi-user model to a single server case from the perspective of a single user. For simplicity, we only consider the case $\nu^{(j)} = 0$ for all j. The general case can also be treated easily. The following proposition provides an expression for the probability that user i is chosen among the n users in a given slot. This expression will be used in the next proposition to compute the service rate distribution for user i.

Proposition 3: For the opportunistic scheduler, in a given slot, the probability that user i is served is given by

$$\alpha^{(i)} := \int_0^\infty m_i e^{-m_i x} \prod_{j=1, j \neq i}^{n} \left(1 - e^{-m_j g^{(j)}(x)}\right) dx,$$

where $g^{(j)}(x) := \frac{(1+xSNR)^{(\beta^{(i)}/\beta^{(j)})}-1}{SNR}$.

For the special case when $\beta^{(j)}$ are equal for all j, the expression simplifies to

$$\alpha^{(i)} = 1 + \sum_{k=1}^{n-1} \left(\sum_{(\gamma^1,\cdots,\gamma^k)\in\Gamma(k)} \frac{(-1)^k m_i}{m_i + m_{\gamma^1} + \cdots + m_{\gamma^k}} \right),$$

where $\Gamma(k)$ contains all $k-$tuples of the set $\{1,\cdots,n\}\backslash\{i\}$.

Proof: The proof follows from conditioning on the realization on $h^{(i)}$. ■

Proposition 4: For the opportunistic scheduling, the probability density function of the service rate process for user i, denoted as $f_{\mu^{(i)}}(r)$, is given by

$$\frac{m_i e^{r-\frac{m_i(e^r-1)}{SNR}}}{SNR} \prod_{j\neq i} \left(1 - e^{-m_j g^{(j)}\left(\frac{e^r-1}{SNR}\right)}\right) + (1-\alpha^{(i)})\delta(r),$$

for any $r \geq 0$ and zero elsewhere, $\alpha^{(i)}$, $g^{(j)}(x)$ are defined in Proposition 3 and $\delta(r)$ is the Kronecker delta function.

Proof: The proof follows from considering cumulative distribution function of $\mu^{(i)}$ and using Proposition 3 to condition on the i^{th} user being chosen. Details are omitted due to space constraints. ■

VI. Simulations

In this section, we compare the numerical results obtained in the previous sections using large deviations with simulation results. In Figure 3, we compare the numerical results of Figure 2 with our simulations. The numerical results are provided in solid in the figure for comparison and the simulation results are plotted in dashed lines. It can

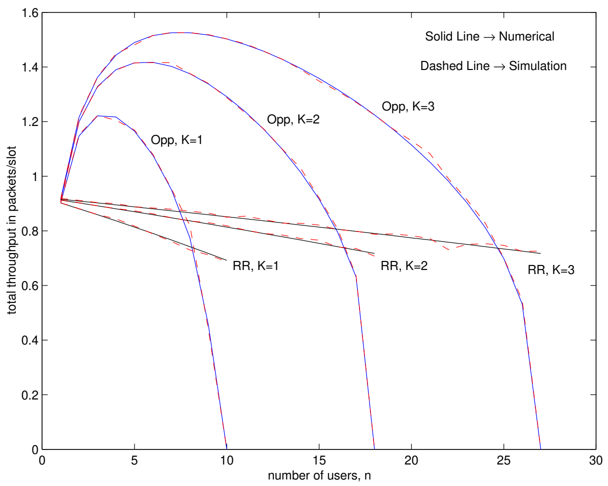

Fig. 3. Simulations versus numerical computations for all cases. The simulation interval was taken to be 300 million slots

be seen from the figure that the simulation results are almost identical with the numerical results. Hence, our formulation captures the actual system dynamics quite accurately. Due to space limitations, we cannot present the simulation results for all the different cases, but in all cases, the simulation results match the large-deviations estimates almost exactly.

VII. Conclusions

In this paper, we studied guaranteed throughput levels for delay-constrained traffic, where the transmission is over a Rayleigh fading channel. The delay constraint is in the form of a small tail probability for the delay distribution. We compared the supportable throughput for two schedulers: Opportunistic and TDM schedulers. The numerical computations revealed that the opportunistic scheduler exhibits an initial increase in the total supportable throughput level as the number of users increase and then it drops sharply, whereas TDM scheduler's performance decreases much more slowly. We observed that the opportunistic scheduler is better than the TDM scheduler for a small number of users.

Acknowledgments: We thank Prof. Pramod Viswanath for useful discussions during the course of this work.

References

[1] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi. Cdma/hdr: A bandwidth efficient high speed wireless data service for nomadic users. *IEEE Communications Magazine*, pages 70–77, July 2000.

[2] A. Eryilmaz, R. Srikant, and J. Perkins. Stable scheduling policies for fading wireless channels. To be published at ACM/IEEE Transactions on Networking, available at http://www.comm.csl.uiuc.edu/˜srikant.

[3] P. W. Glynn and W. Whitt. Logarithmic asymptotics for steady-state tail probabilities in single-server queues. *Journal of Applied Probability*, 31A:131–156, 1994.

[4] X. Liu, E. Chong, and N. Shroff. Opportunistic transmission scheduling with resource-sharing constraints in wireless networks. *IEEE Journal on Selected Areas in Communications*, 19(10):2053–2064, October 2001.

[5] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose. Modeling tcp throughput: A simple model and its empirical validation. In *Proceedings of ACM SIGCOMM*, 1998.

[6] S. Sanghavi and B. Hajek. Adaptive induced fluctuations for multi-user diversity. In *Proceedings of ISIT*, Lausanne, Switzerland, July 2002.

[7] S. Shakkottai. Effective capacity of the max-queue rule. In *Proceedings of the Allerton Conference on Control, Communications and Computing*, pages 665–674, 2003.

[8] S. Shakkottai. Modes of overflow, effective capacity and qos for wireless scheduling. In *ISIT*, page 334, 2003.

[9] S. Shakkottai, R. Srikant, and A. Stolyar. Pathwise optimality of the exponential scheduling rule for wireless channels. In *Proceedings of ISIT*, Lausanne, Switzerland, July 2002. Longer version available at http://www.comm.csl.uiuc.edu/˜srikant.

[10] S. Shakkottai and A. Stolyar. Scheduling for multiple flows sharing a time-varying channel: The exponential rule. *Translations of the American Mathematical Society*, 2001. To appear.

[11] L. Tassiulas and A. Ephremides. Dynamic server allocation to parallel queues with randomly varying connectivity. *IEEE Transactions on Information Theory*, 39:466–478, March 1993.

[12] D. N. Tse. Multi-user diversity and proportional fairness. Unpublished work. Also US Patent 6449490.

[13] P. Viswanath, D. Tse, and R. Laroia. Opportunistic beamforming using dumb antennas. *IEEE Transactions on Information Theory*, 48(6):1277–1294, June 2002.