Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Joint Scheduling of Rate-guaranteed and Best-effort Services over a Wireless Channel

Murtaza Zafer and Eytan Modiano[1]

Abstract— **We address multi-user scheduling over the downlink channel in wireless data systems. Specifically, we consider a time-slotted system with a single transmitter serving multiple users. With fixed power transmission, the channel condition of a user determines the reliable rate of communication to that user in a particular time-slot. The user set consists of, (i) throughput guaranteed (QoS) users, and, (ii) best effort (BE) users. For such a system we obtain the optimal policy that serves the QoS users with minimum time-slot utilization, thereby, maximizing the remaining fraction of time-slots allocated to the BE users. We present a simple geometric visualization of the optimal policy. In the special scenario of symmetric Rayleigh fading, we obtain explicit formulas that relate the achievable throughput rate guarantee to the number of QoS users supportable and the fraction of time-slots allocated to the BE users. Finally, we compare the throughput results for the optimal policy with the random-scheduling policy and show that gains on the order of $\ln(N)$ can be achieved by exploiting multi-user diversity, where N is the number of QoS users.**

Index Terms— **Downlink, Opportunistic scheduling, Multi-user diversity, Quality of Service, Wireless fading channel.**

I. Introduction

Rapid growth of the internet and multi-media applications has created an ever increasing demand for wireless data systems. Development of data systems, such as the 1xEV-DO system in [3], introduces new challenges in providing Quality of Service (QoS) over a wireless channel. In contrast to conventional voice traffic, data streams are inherently bursty and can tolerate much higher delays. Hence, reserving resources to provide QoS is inefficient which means that to share a common resource one needs efficient scheduling algorithms. Also, as the wireless channel is time-varying, one can exploit the varying channel conditions among various users to increase the system throughput. In the literature, such an approach is referred to as *Opportunistic scheduling* [1], [2], [4] or exploiting *Multi-user diversity* [6].

In this work, we consider the downlink scenario with a single server that represents the base station and multiple users that represent the mobile handsets. The set of users are divided into two classes: (i) throughput rate guaranteed QoS users and (ii) "best effort" (BE) users. The QoS users have high priority service and are guaranteed average throughput rates if these rates are feasible; while, the BE users have a low priority service and are served when the resources are available. The goal of this work is to design a scheduling policy that serves the QoS users with the least time-slot utilization so as to maximize the remaining fraction of time-slots available for the BE users.

Down-link scheduling is an active area of research with recent work that includes [1], [2], [4], [5]. The work in [1] presented various formulations based on utility maximization. The work in [2] considered the objective of maximizing the minimum throughput rate, [4] maximized the throughput with fairness constraints while [5] presented algorithms with delay considerations. Our work differs in presenting a simple formulation that combines the QoS and the BE users by abstracting the service of BE users as the fraction of allocated time-slots. We give the optimality conditions and show that a policy is optimal if and only if it satisfies a certain simple geometric structure. Under symmetric Rayleigh fading, we obtain explicit formulas that relate the achievable throughput rate guarantee to the number of QoS users supportable and the fraction of time-slots assigned to the BE users. Finally, we analytically compare the optimal and the random-scheduling policy and quantify the gains achieved by exploiting multi-user diversity.

II. System and Problem Description

A. System Model

We consider the wireless downlink scenario, i.e. communication from the base station to the mobile handsets in a time slotted system. There are multiple users in the system, each user experiencing time varying channel conditions. The channel state of a user is assumed constant in a single time slot but varies over multiple time slots. We assume that the underlying stochastic process driving the channels' states is stationary. This, however, does not preclude the possibility of channel correlations over time and among users. At the beginning of a time-slot, the transmitter knows the channel state of each user for that particular slot[1]. In a time-slot, it serves at most one user with full power P. Since the users have different channel conditions the reliable rate of communication per time slot to the users is variable. Clearly, the transmitter can exploit this variability and select the "best user" for transmission in a time-slot based on some performance measure. The above system models a TDMA system and the recently proposed 1xEV-DO data system [3] and is a commonly used model in the literature to study *opportunistic scheduling* in wireless networks [1], [2], [4].

[1] Murtaza Zafer {murtaza@mit.edu} and Eytan Modiano {modiano@mit.edu} are with the Laboratory for Information and Decision Systems, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, USA. This work was supported by NSF ITR grant CCR-0325401, by DARPA/AFOSR through the University of Illinois grant no. F49620-02-1-0325 and by NASA Space Communication Project grant number NAG3-2835.

[1] This is a simplifying assumption that models one step channel prediction

0-7803-9568-9/05/$20.00 ©2005 IEEE

Let $\bar{\mathbf{r}} = \{r_i\}$ denote the vector of reliable rate of communication to the users in a generic time-slot, say for example the k^{th} time-slot. This means that if user i is chosen to be served in time-slot k, the throughput for that user is simply r_i. The transmitter has knowledge of $\bar{\mathbf{r}}$ at the beginning of slot k but does not know this vector for future slots. Let Ω be the set comprising of all possible rate vectors. In the k^{th} time-slot, $\bar{\mathbf{r}}$ is a particular realization from the set Ω which has a probability distribution induced by the underlying stochastic model of the channels' states. A scheduling policy, denoted as $\Gamma^k(\bar{\mathbf{r}})$, is a rule that specifies which user the transmitter serves in time-slot k. A *stationary scheduling policy*, denoted $\Gamma(\bar{\mathbf{r}})$, is one that does not depend on the time index and can be represented as a map from the set Ω to the user index; i.e. each $\bar{\mathbf{r}} \in \Omega$ is mapped to a unique user index. As the underlying processes are stationary, it is well-known that a stationary optimal policy exists, hence, it suffices to focus on stationary policies. In the rest of the paper, a scheduling policy refers to the above map.

Let X_i denote the throughput per time-slot of user i, then,

$$X_i(\bar{\mathbf{r}}) = \begin{cases} r_i, & \text{if } \Gamma(\bar{\mathbf{r}}) = i \text{ (i.e. user } i \text{ selected)} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The expected throughput per time slot is $E[X_i]$. Under ergodicity of the channel process and stationarity of the scheduling rule, it's well known that $E[X_i]$ equals the long term throughput per slot (called throughput rate) of user i.

B. Problem Description

As mentioned earlier, the set of users are divided into two priority classes: (i) the throughput rate guaranteed (QoS) users and (ii) the "best effort" (BE) users. The QoS users are guaranteed average throughput rates while the BE users have no such guarantees. Let there be N QoS users that are guaranteed throughput rates $\bar{\mathbf{R}} = (R_1, .., R_N)$, if such a vector is *feasible*. By feasibility we mean that there exists a scheduling policy such that $E[X_i] \geq R_i, \forall i = 1, .., N$, where X_i is defined as in (1). The objective, now, is to serve the QoS users with the least time-slot utilization and share the remaining time-slots among the BE users. This objective provides a simple and tractable way of integrating the two classes of service. Also, typically in most practical systems, the population of BE users is large and a natural objective while serving such users is simply maximizing the sum-throughput. Clearly, under a large population of BE users maximizing the time-slot allocation is equivalent to maximizing the total throughput of such users[2].

Let I_i be the indicator function for selection of user i,

$$I_i(\bar{\mathbf{r}}) = \begin{cases} 1, \text{ if } \Gamma(\bar{\mathbf{r}}) = i \\ 0, \text{ otherwise} \end{cases} \quad (2)$$

With this notation we can re-write X_i as $X_i = r_i I_i$. The optimization problem can now be stated as follows,

$$\begin{aligned} \min \quad & \sum_{i=1}^{N} E[I_i] \\ \text{subject to} \quad & E[r_i I_i] \geq R_i, \; i = 1, .., N \end{aligned} \quad (3)$$

where the expectation is taken over the joint distribution of $\bar{\mathbf{r}}$ for the N QoS users. Note that minimizing $\sum_{i=1}^{N} E[I_i]$ is equivalent to maximizing $1 - \sum_{i=1}^{N} E[I_i]$ which equals the fraction of time-slots available for the BE users. We assume that $\bar{\mathbf{R}} > 0$, i.e. $(R_1 > 0, .., R_N > 0)$. If some $R_k = 0$, we can neglect that user and the problem reduces to $N - 1$ dimensions. We assume that $\bar{\mathbf{R}}$ is feasible and away from the boundary of the set comprising all achievable throughput rate vectors. This assumption is solely to simplify the mathematical exposition by avoiding the limiting conditions at the boundary and does not affect the results presented throughout this paper.

III. Optimal Policy

The QoS users experience different time-varying channel conditions, hence, intuitively the optimal policy must exploit the variable communication rates to the users by selecting the best user to have a high throughput per time slot. The choice of which user to serve must also account for the different throughput rate guarantees among the users and their varying channel statistics. Clearly, for optimality the inequality in (3) must also be met with equality.

Let $\bar{\mathbf{r}} = (r_1, \ldots, r_N)$ be the rate vector in a generic time-slot for the N QoS users[3]; this vector lies in the set $\Omega \subseteq \mathbb{R}^{+N}$. Let the joint probability density function be $f(\bar{\mathbf{r}})$ such that the probability of some region $Z \subset \Omega$ is given as $\int_Z f(\bar{\mathbf{r}}) d\bar{\mathbf{r}}$. The restriction on $f(\bar{\mathbf{r}})$ is that subsets with zero volume in Ω (or individual points) have zero probability. Since a scheduling policy maps $\bar{\mathbf{r}} \in \Omega$ to a unique user index, we can represent it as a partition of the set Ω into $N + 1$ regions denoted as $Z_1, .., Z_N, Z_f$. In a particular time-slot, if the transmission rate vector $\bar{\mathbf{r}} \in Z_i$, user i is selected for service whereas if $\bar{\mathbf{r}} \in Z_f$, no QoS user is selected and the slot is used to serve the BE users. The problem thus reduces to choosing these regions optimally to minimize the objective function and satisfy the throughput rate constraint, $\int_{Z_i} r_i f(\bar{\mathbf{r}}) d\bar{\mathbf{r}} \geq R_i, \; i = 1, \ldots, N$.

As individual points in Ω have zero probability, we will refer to regions within Ω[4]. The notation $\bar{\mathbf{r}} \rightarrow Z$ ($\bar{\mathbf{r}} \not\rightarrow Z$) means that there is a neighborhood around $\bar{\mathbf{r}}$ that lies (does not lie) in Z. Formally, $\bar{\mathbf{r}} \rightarrow Z$ implies that there exists $\epsilon > 0$ such that $\hat{\mathbf{r}} \in \Omega, \; ||\hat{\mathbf{r}} - \bar{\mathbf{r}}|| < \epsilon \Rightarrow \hat{\mathbf{r}} \in Z$. The following lemma gives the necessary condition for the optimality of region Z_f. It states that if $\bar{\mathbf{r}}$ is mapped to Z_i, all rate vectors with i^{th} component larger than r_i cannot be mapped to Z_f.

***Lemma* 1:** Under the optimal policy, suppose $\bar{\mathbf{r}} = (r_1, .., r_N) \rightarrow Z_i$ then $\hat{\mathbf{r}} = (\hat{r}_1, .., (\hat{r}_i > r_i), .., \hat{r}_N) \not\rightarrow Z_f$.

[2] Time slots allocated for BE users can be shared in a greedy fashion, thus, maximizing the sum throughput of these users.

[3] To make the notations simple, $\bar{\mathbf{r}}$, depending on the context denotes a random vector and also a particular realization for a generic time-slot.

[4] Regions with zero probability density can be removed from Ω as their mapping does not affect optimality.

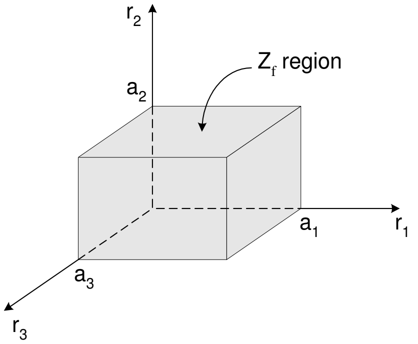

Fig. 1. The Z_f region for $N = 3$, threshold vector $\bar{\mathbf{a}} = (a_1, a_2, a_3)$ and $\Omega = \mathbb{R}^{+N}$. Note $Z_f = \{\bar{\mathbf{r}} : 0 \le r_i \le a_i,\ \forall i = 1, \ldots, N\}$.

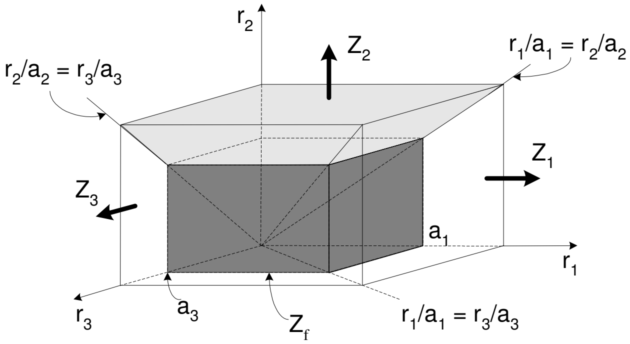

Fig. 2. Optimal policy structure for $N = 3$, threshold vector $\bar{\mathbf{a}} = (a_1, a_2, a_3)$ and $\Omega = \mathbb{R}^{+N}$. The Z_i regions are top truncated pyramids.

Proof: We omit a rigorous proof for brevity but the main idea is that if there is a $\hat{\mathbf{r}} \to Z_f$ with $\hat{r_i} > r_i$ then we can re-map the regions such that the objective function in (3) decreases. This is achieved by mapping a small neighborhood of $\hat{\mathbf{r}}$ to Z_i and mapping a neighborhood of $\bar{\mathbf{r}}$ to Z_f while ensuring that the throughput constraints still hold. As $\hat{r_i} > r_i$ one can show that the objective function under the new map is strictly lower than the earlier map. ■

Interestingly, Lemma 1 implies a special structure on Z_f as follows. Let a_1 be the infimum value of the first component among all vectors $\bar{\mathbf{r}} \to Z_1$; i.e. $a_1 = \inf_{(\bar{\mathbf{r}} \to Z_1)} r_1$. Now, any $\hat{\mathbf{r}} \to Z_f$ must be such that $\hat{r}_1 \le a_1$; otherwise the above lemma will be violated. As this holds for all Z_i, the optimal policy is such that there exists constants $\{a_i\}$ and if $r_i \le a_i, \forall i$ then $\bar{\mathbf{r}} \in Z_f$. The region Z_f is shown in Figure 1. This implication is quite intuitive as it suggests that when the rate vector of the QoS users is below some threshold vector (bad channel conditions), the QoS users must not be scheduled and the slot must be used to serve the BE users.

The vector $\bar{\mathbf{a}}$ depends on the throughput vector $\bar{\mathbf{R}}$ and the density function $f(\bar{\mathbf{r}})$. Given that $\bar{\mathbf{R}}$ does not lie on the boundary of feasible throughput rates, it follows that $\bar{\mathbf{a}}$ is at least a positive vector $(a_1 > 0, \ldots, a_N > 0)$ and the region $Z_f = \{\bar{\mathbf{r}} : \bar{\mathbf{r}} \in \Omega, r_i \le a_i \forall i\}$ is not null (non-zero probability). We now proceed to obtain the structure of the regions $Z_i,\ i = 1, \ldots, N$.

***Lemma* 2:** Consider regions $Z_i,\ Z_j,\ j \ne i$ and the corresponding thresholds a_i, a_j. Suppose $\bar{\mathbf{r}} \notin Z_f$ and satisfies,

$$\frac{r_i}{a_i} > \frac{r_j}{a_j} \tag{4}$$

then under the optimal policy $\bar{\mathbf{r}} \not\to Z_j$

Proof: Appendix I ■

The above lemma states that if the weighted comparison of the i^{th} and the j^{th} component of $\bar{\mathbf{r}}$ is in favour of user i, it is not optimal to serve user j. The weights are the inverse values of the corresponding components of the threshold vector $\bar{\mathbf{a}}$. The above implication is intuitive as condition (4) means that in some sense user i has a better channel condition than user j and hence serving user j is not optimal. Combining the above two lemmas, we obtain the following necessary conditions for the optimal policy.

Theorem *I: (Necessary Conditions)* Consider $\bar{\mathbf{r}} = (r_1, \ldots, r_N)$ then the optimal policy is such that there exists a threshold vector $\bar{\mathbf{a}}$ with the following structure,

1) $\bar{\mathbf{r}} \to Z_f$ if it satisfies,

$$r_i < a_i,\ \forall i = 1, \ldots, N \tag{5}$$

2) $\bar{\mathbf{r}} \to Z_i$, $(i = 1, \ldots, N)$ if it satisfies,

$$\frac{r_i}{a_i} > \frac{r_j}{a_j},\ \forall j = 1, \ldots, N, j \ne i \tag{6}$$

$$r_i > a_i \tag{7}$$

3)

$$\int_{Z_i} r_i f(\bar{\mathbf{r}}) d\bar{\mathbf{r}} = R_i,\ \forall i = 1, \ldots, N \tag{8}$$

Proof: Conditions 1 and 2 follow from Lemmas 1 and 2. Clearly, as stated in Condition 3, for optimality the throughput constraint must be met with equality. ■

The set of $\bar{\mathbf{r}}$ that lie on the boundaries for which there is equality in (5) and (6) can be mapped to any Z_i without affecting optimality. It can also be observed that the set of conditions in Theorem I are exhaustive and map every $\bar{\mathbf{r}} \in \Omega$ to a unique user index. Thus, given $\bar{\mathbf{a}}$, we have a unique partition of Ω into regions $Z_1, \ldots, Z_N, Z_f$. In Figure 2, we present a geometric picture of these regions for $N = 3$. As seen from the figure the Z_i regions are top truncated pyramids and it can be verified (say, for example Z_2 region) that (6) is satisfied.

Next, we present the sufficiency argument by proving that a scheduling policy of the form as in Theorem I minimizes the objective in (3) and hence is optimal. First, observe that a scheduling policy outlined in Theorem I can be re-written in a simplified way as a maximum weighted rule as follows,

$$\Gamma(\bar{\mathbf{r}}) = \begin{cases} Z_f \text{ (no QoS user)}\ , & \text{if } r_i \le a_i, \forall i = 1, .., N \\ \text{argmax}_i\ \frac{r_i}{a_i}\ , & \text{otherwise} \end{cases} \tag{9}$$

where $\{a_i\}$ are such that $E[r_i I_i] = R_i, \forall i$.

Theorem *II: (Sufficiency)* Consider the optimization problem in (3) and let $\bar{\mathbf{R}}$ be feasible, then policy Γ defined in (9) is optimal.

Proof: Appendix II. ■

Thus, Theorem I states that the optimal policy must satisfy certain conditions which impose a weighted comparison structure on the policy and conversely, Theorem II completes the argument by stating that any policy with that structure is optimal. Now, vector $\bar{\mathbf{a}}$ is chosen such that $\int_{Z_i} r_i f(\bar{\mathbf{r}}) d\bar{\mathbf{r}} = R_i,\ i = 1,..,N$. This can be solved using techniques of finding the positive root of a non-linear vector equation. For general density functions, it is difficult to obtain analytical expressions for $\bar{\mathbf{a}}$. In practice, however, vector $\bar{\mathbf{a}}$ can be adjusted in real time using stochastic approximation algorithms similar to those outlined in [1], [2], [8], [9]. Interestingly, as discussed next in Section IV, one can solve for $\bar{\mathbf{a}}$ in closed form under a symmetric Rayleigh fading model. From a system perspective, this analytical study helps us obtain explicit results for various important performance measures such as the achievable throughput rate guarantee, the number of QoS users supportable and the fraction of time-slots allocated to the BE users.

IV. Dimensioning

We have shown that an optimal policy has a weighted structure as represented in (9) for some threshold vector $\bar{\mathbf{a}}$. Here, we consider a symmetric Rayleigh fading scenario under which closed form expressions can be obtained for various performance measures. To proceed, we make the following specializations to the earlier model. The rate per time slot of a user is assumed proportional to the fade state (square magnitude); i.e. $r = k(|h|^2 P)$, where k is a constant, $|h|$ is the magnitude of the fade state and P is the transmission power. This linear relationship is a good approximation of the Shannon capacity formula in the low SNR regime and in ultra-wideband transmission and has been studied earlier in the literature [7]. The users experience independent identically distributed (i.i.d) flat Rayleigh fading, hence, $|h|^2$ is Exponentially distributed. As r is proportional to $|h|^2$, the distribution of r is also Exponential and is given as $f(r) = e^{-r/\mu}/\mu,\ r \geq 0$ where $\mu = E[r]$ is the average throughput rate of a user if it is served in all the time-slots. Finally, the guaranteed throughput rate is the same for all N QoS users, i.e. $\bar{\mathbf{R}} = (R, \ldots, R)$.

A. Throughput Characterization

Intuitively, the fraction of time-slots remaining for the BE users, denoted as γ, will depend on the parameters R, N, μ. As R, N increases, γ should decrease whereas if μ increases (higher communication rates to the QoS users), the throughput guarantee can be achieved in fewer slots and γ should increase. Equivalently, given γ, N, μ, one can also ask for the maximum throughput-rate guarantee achievable for the QoS users. Our goal in the subsequent analysis is to obtain expressions for all these performance measures.

It's clear that due to symmetry in $f(\bar{\mathbf{r}})$ and $\bar{\mathbf{R}}$, the regions $Z_i,\ i = 1,..,N$ are identical ($\Omega = \mathbb{R}^{+N}$). Hence, the $\{a_i\}$'s are equal and the threshold vector is given as $\bar{\mathbf{a}} = (a,..,a)$. The following lemma relates the threshold value a with γ.

***Lemma* 3:** Let γ be the fraction of time-slots allocated to the BE users, the threshold value a for the optimal policy is given by,

$$a = \mu \ln \left(\frac{1}{1 - \gamma^{1/N}} \right) \tag{10}$$

Proof: From Theorem I, the region Z_f is given as $Z_f = \{\bar{\mathbf{r}} : 0 \leq r_i \leq a,\ \forall i = 1, \ldots, N\}$. By ergodicity, the probability of this region equals γ and by the i.i.d channel assumption, $f(\bar{\mathbf{r}}) = \prod_i f_i(r_i) = \prod_i f(r_i)$. Thus we get,

$$\int_0^a \ldots \int_0^a \prod_i f(r_i) dr_i = \gamma \tag{11}$$

Evaluating the integrals for the exponential distribution gives,

$$\gamma = \left(1 - e^{-a/\mu}\right)^N \tag{12}$$

Re-writing the above expression gives the result in (10). ■

Observe from (10) that $\gamma = 0 \Rightarrow a = 0$ and $\gamma = 1 \Rightarrow a \to \infty$ which corroborates the intuition that $\gamma = 0$ implies Z_f is null and $\gamma = 1$ (all slots for BE users) implies $Z_f = \mathbb{R}^{+N}$.

***Lemma* 4:** Under the optimal policy, the throughput rate guarantee R for a given threshold value a is given by,

$$R = \sum_{k=0}^{N-1} \binom{N-1}{k} (-1)^k \left(a + \frac{\mu}{k+1}\right) \frac{e^{-(k+1)a/\mu}}{k+1} \tag{13}$$

Proof: Given a threshold vector $\bar{\mathbf{a}} = (a, \ldots, a)$, the region Z_i is given as, $Z_i = \{\bar{\mathbf{r}} : a \leq r_i < \infty,\ 0 \leq r_j \leq r_i,\ j \neq i\}$. As $R = E[r_i I_i]$ we get,

$$R = \int_a^\infty \int_0^{r_i} \ldots \int_0^{r_i} r_i f(r_i) dr_i \prod_{j \neq i} f(r_j) dr_j \tag{14}$$

where $f(\bar{\mathbf{r}}) = \prod_i f_i(r_i) = \prod_i f(r_i)$ by the i.i.d assumption. For the exponential distribution, (14) simplifies to,

$$R = \int_a^\infty \frac{r_i e^{-r_i/\mu}}{\mu} \left(1 - e^{-r_i/\mu}\right)^{N-1} dr_i \tag{15}$$

Using the binomial expansion, $(1 - e^{-r_i/\mu})^{N-1} = \sum_{k=0}^{N-1} \binom{N-1}{k} (-1)^k e^{-kr_i/\mu}$, (15) can be solved to give (13). ■

Conversely, one can also solve (13) to obtain the value of a that would achieve rate R. As R is monotonically decreasing in a, the value of $a \geq 0$ that achieves R in (13) is unique.

Eliminating a from (10) and (13) we obtain a unified relationship among the system quantities: (i) Throughput rate R, (ii) Fraction of time-slots, γ, allocated to the BE users and (iii) Number of QoS users, N, in the system.

Theorem *III:* Under the model assumptions stated earlier with N QoS users in the system and $\gamma \in [0, 1]$ fraction of time-slots allocated to the BE users, the maximum throughput rate R for each QoS user is given as,

$$\begin{aligned}\frac{R}{\mu} =& \sum_{k=0}^{N-1} \binom{N-1}{k} (-1)^k \times \\ & \left(\frac{-\ln(1 - \gamma^{1/N})}{k+1} + \frac{1}{(k+1)^2} \right) (1 - \gamma^{\frac{1}{N}})^{(k+1)}\end{aligned} \tag{16}$$

Proof: The result follows from Lemmas 3 and 4. ■

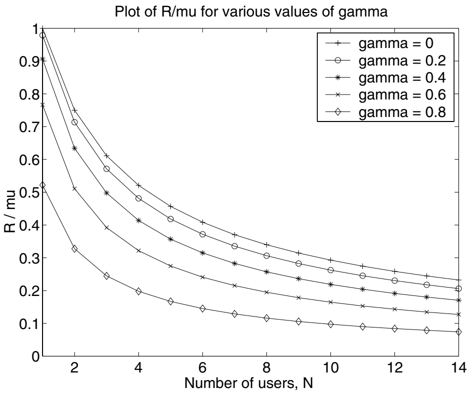

Fig. 3. Plot of R/μ versus N for the optimal policy for various γ values.

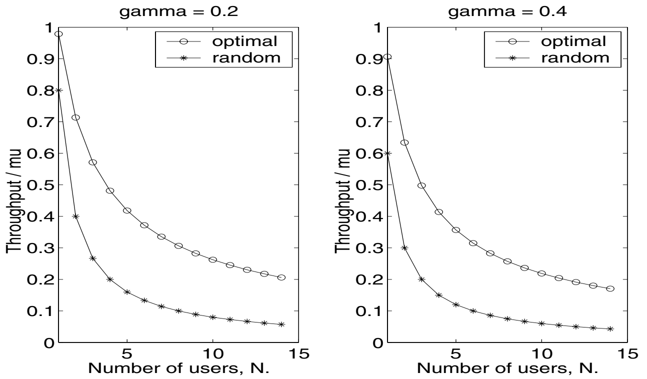

Fig. 4. Plot comparing R/μ for the optimal and the random policy.

An interesting observation is that R varies linearly with μ where μ is the average channel condition of the QoS users. Re-phrasing (16) we see that given R_0 and γ, $N_{max} = \max_{N\geq 1}\ (R \geq R_0)$ is the maximum number of supportable QoS users with rate guarantee R_0, if a solution exists. Finally, given R and N, the value of γ in (16) is the maximum fraction of slots that can be allocated to the BE users. Figure 3 is a plot of R/μ versus N for different γ values.

B. Comparison with Random-scheduling

We, now, compare the performance of the optimal policy with the random scheduling policy that is very simple to implement and does not exploit the varying channel conditions among the users. Specifically, the random policy assigns a time-slot to the BE users with probability γ and to the QoS users with probability $1-\gamma$. Among the QoS users the slot is then randomly assigned to one of the users with equal probability $1/N$. Due to the random nature of the assignment each QoS user gets $(1-\gamma)/N$ fraction of time-slots and the users have statistically identical channel conditions. Thus the throughput rate of each QoS user, denoted R_r, is given as,

$$R_r = \mu\frac{(1-\gamma)}{N} \tag{17}$$

Figure 4 plots R^{opt}/μ and R_r/μ versus N for $\gamma = 0.2, 0.4$, where R^{opt} is the throughput for the optimal policy as given in (16). We, next, quantify the gain, defined as R^{opt}/R_r, for large N and show that it is on the order of $\ln(N)$.

***Proposition* 1:** The throughput gain, defined as R^{opt}/R_r, of the optimal policy as compared to the random policy is,

$$\frac{R^{opt}}{R_r} = \Theta(\ln(N)) \tag{18}$$

Proof: Starting with (16), the summation over the first terms can be evaluated as follows. Let $\alpha = (1-\gamma^{\frac{1}{N}})$, then, taking $\gamma \in (0,1)$ we have $\alpha \in (0,1)$.

$$\begin{aligned}\sum_{k=0}^{N-1}\binom{N-1}{k}(-1)^k\frac{\alpha^{(k+1)}}{k+1} &= \sum_{k=0}^{N-1}\binom{N-1}{k}\int_0^{\alpha}(-x)^k dx\\ = \int_0^{\alpha}(1-x)^{N-1}dx &= \frac{1-(1-\alpha)^N}{N} = \frac{1-\gamma}{N} \quad (19)\end{aligned}$$

We can now re-write (16) as,

$$\frac{R}{\mu} = \frac{1-\gamma}{N}\left(\ln\left(\frac{1}{\alpha}\right) + \frac{N}{1-\gamma}\sum_{k=0}^{N-1}\binom{N-1}{k}\frac{-1^k\alpha^{k+1}}{(k+1)^2}\right) \tag{20}$$

Since (19) holds for all α, we get the identity, $\sum_{k=0}^{N-1}\binom{N-1}{k}(-1)^k\frac{x^{(k+1)}}{k+1} = \frac{1-(1-x)^N}{N}$. Dividing both sides of this equation by x and integrating from 0 to α, we get,

$$\begin{aligned}\sum_{k=0}^{N-1}\binom{N-1}{k}\frac{-1^k\alpha^{k+1}}{(k+1)^2} &= \int_0^{\alpha}\left(\frac{1-(1-x)^N}{Nx}\right)dx\\ &\leq \int_0^{\alpha}dx = \alpha = (1-\gamma^{\frac{1}{N}}) \quad (21)\end{aligned}$$

The inequality above follows by noting that $\frac{1-(1-x)^N}{Nx}$ is positive, monotonically decreasing for $x \in [0,1], N \geq 1$ and has a maximum value equal to 1 at $x = 0$. Using (21) we can bound the summation term in (20) as, $\frac{N}{1-\gamma}\sum_{k=0}^{N-1}\binom{N-1}{k}\frac{-1^k\alpha^{k+1}}{(k+1)^2} \leq \frac{N}{1-\gamma}(1-\gamma^{\frac{1}{N}}) \xrightarrow{N\to\infty} \frac{-\ln(\gamma)}{1-\gamma}$ (which is finite for $\gamma > 0$). Considering the log term in (20) we see that, $\ln(\frac{1}{\alpha}) = -\ln(1-\gamma^{\frac{1}{N}}) = \gamma^{1/N} + \frac{\gamma^{2/N}}{2} + \frac{\gamma^{3/N}}{3} + \ldots = \Theta(\ln(N))$. Thus, for any $0 < \gamma < 1$ and large N, the log term in (20) dominates and we can express R^{opt} as,

$$\frac{R^{opt}}{\mu} = \frac{1-\gamma}{N}\Theta(\ln(N)) \tag{22}$$

From (17) and (22) we get the result in (18), ■

Observe that as $N \to \infty$ the throughput for both the optimal and the random policy tends to zero. Equation (22) simply states that R^{opt} decreases as $\ln(N)/N$ while (17) states that R_r decreases as $1/N$. Hence, we get a gain on the order of $\ln(N)$. The above logarithmic behavior arises due to the infinite support and the exponential distribution of the rate under Rayleigh fading. While such channel statistics are simplified models, in practice one could expect gains along these orders for moderate QoS user population.

V. Conclusion

We addressed the issue of downlink scheduling over a wireless channel incorporating the QoS and best effort services. We considered a set of N rate guaranteed users and obtained an optimal policy that serves these users with the least time-slot utilization, thereby, maximizing the time-slot allocation to the BE users. This work opens up interesting questions about QoS guarantees over wireless channels. While we considered long-term rate guarantee as a QoS measure, future work seeks to address scheduling over a wireless channel with more general QoS requirements, for example, strict delay constraints on the data such as those that arise in video streaming and multimedia applications.

REFERENCES

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

[2] S. Borst, P. Whiting, "Dynamic rate control algorithms for HDR throughput optimization", *IEEE INFOCOM*, Alaska, April 2001.

[3] A. Jalali, R. Padovani, R. Pankaj, "Data throughput of CDMA-HDR a high efficiency high data rate personal communication wireless system", *IEEE Vehicular Technology Conf.*, vol. 3, 2000.

[4] Y. Liu, E. Knightly, "Opportunistic fair scheduling over multiple wireless channels", *IEEE INFOCOM 2003*, San Francisco, 2003.

[5] S. Shakkottai, A. Stolyar, "Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR", *Proc. International Teletraffic Congress (ITC-17)*, Brazil, Sept. 2001.

[6] P. Viswanath, D. Tse and R. Laroia, "Opportunistic Beamforming using Dumb Antennas", *IEEE Tran. on Information Theory*, 48(6), June, 2002.

[7] P. Liu, R. Berry, M. Honig, "Delay-Sensitive Packet Scheduling in Wireless Networks", *IEEE WCNC* 2003, New Orleans, LA.

[8] H. Kushner, G. Yin, "Stochastic approximation algorithms and applications", Springer, New York, 1997.

[9] I. Wang, E. Chong, S. Kulkarni, "Weighted averaging and stochastic approximation", Math. of Control, Signals and Systems 1 (10), 1997.

APPENDIX I
PROOF OF LEMMA 2

For brevity, we simply outline the steps involved in the proof and omit the technical steps. The proof is based on a contradiction argument. To begin, consider $\bar{\mathbf{r}} \notin Z_f$ and suppose that for the optimal policy, $\bar{\mathbf{r}} \rightarrow Z_j$ such that $\frac{r_i}{a_i} > \frac{r_j}{a_j}$. We now give a re-mapping of the regions such that the objective function decreases or equivalently the probability of Z_f region increases, thereby, showing that the earlier mapping cannot be optimal. As the lemma involves only the i^{th} and j^{th} component, we will focus only on these components. Let the neighborhood around $\bar{\mathbf{r}}$ that is mapped to Z_j be denoted as S_1. We can represent S_1 as $S_1 = \{\bar{\mathbf{x}} : \bar{\mathbf{x}} \in \Omega, ||\bar{\mathbf{x}} - \bar{\mathbf{r}}|| < \delta_1\}$ for some $0 < \delta_1 \le \delta_1^m$ where δ_1^m is the largest δ_1 such that $S_1 \in Z_j$. By the assumption $\bar{\mathbf{r}} \rightarrow Z_j$, there exists $\delta_1^m > 0$. Now, since the optimal policy satisfies Lemma 1 we know that a_i is the infimum value of the i^{th} component among $\bar{\mathbf{x}} \rightarrow Z_i$. Thus, there exists a point $\bar{\mathbf{m}}$ with $m_i = a_i$ and a region around $\bar{\mathbf{m}}$, denoted S_2, that maps to Z_i. The region S_2 can be represented as $S_2 = \{\bar{\mathbf{x}} : \bar{\mathbf{x}} \in Z_i, 0 < (x_i - m_i) < \delta_2\}$ for $\delta_2 > 0$. Finally, since $\bar{\mathbf{R}}$ does not lie on the boundary of feasible throughput vectors there exists $\bar{\mathbf{n}}$ with $n_j = a_j > 0$ and a region around $\bar{\mathbf{n}}$, denoted S_3, that maps to Z_f. The region S_3 is $S_3 = \{\bar{\mathbf{x}} : \bar{\mathbf{x}} \in Z_f, 0 < (n_j - x_j) < \delta_3\}$ for $\delta_3 > 0$. Thus, we have regions S_1, S_2, S_3 that are not null and as defined above. Now re-map these regions as follows. Map $S_1 \Rightarrow Z_i$, $S_2 \Rightarrow Z_f$ and $S_3 \Rightarrow Z_j$ as shown in Figure 5(b). By appropriately choosing the $\delta_i's$, one can ensure that the throughput constraints are satisfied and also show that the objective function is smaller under the new mapping.

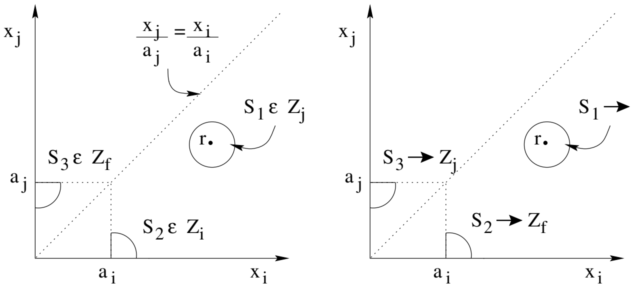

Fig. (a): Original mapping   Fig. (b): New mapping

Fig. 5. Figure showing the mappings for the proof of Lemma 2.

APPENDIX II
PROOF OF THEOREM II

We will prove optimality of policy Γ, defined in (9), by showing that for any other feasible policy $\tilde{\Gamma}$ we have $\sum_{i=1}^N E[I_i] \le \sum_{i=1}^N E[\tilde{I}_i]$ where $I_i(\bar{\mathbf{r}})$ and $\tilde{I}_i(\bar{\mathbf{r}})$ are the indicator functions for the respective policies. We know that policy Γ satisfies the throughput-rate constraints with equality, i.e. $E[r_i I_i] = R_i$. If $\tilde{\Gamma}$ does not, its trivial to prove that $\tilde{\Gamma}$ cannot be optimal. Now, suppose $\tilde{\Gamma}$ also satisfies the rate constraints with equality, i.e. $E[r_i \tilde{I}_i] = R_i$, then, the objective function for policy $\tilde{\Gamma}$ can be re-written as,

$$\sum_{i=1}^N E[\tilde{I}_i] = \sum_{i=1}^N E[\tilde{I}_i] - \sum_{i=1}^N \frac{1}{a_i}(E[r_i \tilde{I}_i] - R_i) \tag{23}$$

where $\{a_i\}$ is the threshold vector for policy Γ. Note that the second term in (23) is zero. Re-arranging (23) we get,

$$\sum_{i=1}^N E[\tilde{I}_i] = E\left[\sum_{i=1}^N \left(1 - \frac{r_i}{a_i}\right) \tilde{I}_i\right] + \sum_{i=1}^N \frac{R_i}{a_i} \tag{24}$$

For any vector $\bar{\mathbf{r}}$ we have the following two cases.

Case 1: Suppose $r_i \le a_i, \forall i$, then, policy Γ does not choose any QoS user (Equation (9)) and $I_i = 0, \forall i = 1, \ldots, N$. Now, since $r_i \le a_i$, we have $(1 - \frac{r_i}{a_i}) \ge 0, \forall i$. This implies that whether $\tilde{\Gamma}$ chooses or does not choose a QoS user we have the following inequality,

$$\sum_{i=1}^N \left(1 - \frac{r_i}{a_i}\right) \tilde{I}_i \ge 0 = \sum_{i=1}^N \left(1 - \frac{r_i}{a_i}\right) I_i \tag{25}$$

Case 2: Suppose $r_i > a_i$ for some index i. Let j be the chosen index for policy Γ, then, from (9) we see that r_j/a_j has the maximum value. Thus, $(1 - \frac{r_j}{a_j}) \le (1 - \frac{r_i}{a_i}), \forall i$ and also $(1 - \frac{r_j}{a_j}) < 0$. Again irrespective of what $\tilde{\Gamma}$ chooses,

$$\sum_{i=1}^N \left(1 - \frac{r_i}{a_i}\right) \tilde{I}_i \ge \left(1 - \frac{r_j}{a_j}\right) = \sum_{i=1}^N \left(1 - \frac{r_i}{a_i}\right) I_i \tag{26}$$

From (24), (25) and (26) we get,

$$\sum_{i=1}^N E[\tilde{I}_i] \ge E\left[\sum_{i=1}^N \left(1 - \frac{r_i}{a_i}\right) I_i\right] + \sum_{i=1}^N \frac{R_i}{a_i} = \sum_{i=1}^N E[I_i]$$

where the last equality follows from (23) replacing $\tilde{I}_i$ with I_i. This completes the proof.