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

Murtaza Zafer and Eytan Modiano[1]
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
Email:{murtaza@mit.edu , modiano@mit.edu}

***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, where the channel condition of each user is time varying. Based on the throughput requirements, the user set is divided into two classes (i) throughput guaranteed (QoS) users, and, (ii) best effort (BE) users. For this system we obtain the optimal policy that serves the QoS users with the minimum time-slot utilization and maximizes the total fraction of time-slots allocated to the BE users. We show that the optimal policy has a simple geometric structure that can be easily visualized graphically. In the special case of Rayleigh fading, we obtain closed-form formulas that relate the achievable throughput-rate guarantee of the QoS users as a function of other system parameters, thus, providing closed-from relationships to understand the various system tradeoffs. Analytical comparison between the optimal and the random-scheduling policy shows that gains on the order of $\ln(N)$ can be achieved, where N is the number of QoS users. Finally, we present simulation results comparing the optimal policy under Rayleigh and Nakagami fading with other heuristic policies including a well known opportunistic-scheduling policy.**

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

I. Introduction

Rapid growth of the Internet and multi-media applications has created a fast increasing demand for data services over wireless systems. Development of wireless data systems, such as the 1xEV-DO system in [1], WiMAX etc., introduces new challenges in providing Quality of Service (QoS) over a wireless channel [2]. 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. Therefore, in order to share a common resource, one needs efficient scheduling algorithms. Furthermore, in a wireless system the scheduling problem has an additional complexity associated with time-varying communication rates since the channel conditions are time-varying. With multiple users in the system, the transmitter can look at the communication rates of the various users and opportunistically choose the "best user" to transmit to based on a required set of objectives. In the literature, such an approach is referred to as *Opportunistic scheduling* [4], [5], [7] or exploiting *Multi-user diversity* [10].

In this work, we utilize opportunistic scheduling to address the following downlink scenario: there is a single server that represents the base station transmitting to multiple users that represent the mobile handsets. The system operates in a time-slotted manner and in each time-slot the base station can serve only one user. The set of users are divided into two classes: (i) throughput rate guaranteed QoS users and (ii) "best effort" (BE) users. The QoS users in the system represent session applications such as FTP, high data-rate web-browsing, throughput-constrained data transfers etc., which require the base station to provide a certain data rate on the downlink. In contrast, the BE users represent on-the-fly applications such as email transfers, low priority and latency tolerant data transfers etc. which do not have rate requirements and are short-lived. The goal of this work is to design a scheduling policy that provides the required throughput rates to the QoS users with the least time-slot utilization and maximizes the remaining fraction of time-slots assigned for the BE class.

Down-link scheduling and power/rate adaptation is an active area of research in wireless systems with recent work that includes [4]–[9], [11]–[14]. The work in [4] studied opportunistic scheduling under a utility maximization framework and presented various formulations therein. In [5], the authors considered the objective of maximizing the minimum throughput rate among a set of users while [6] extended the framework to include a dynamic user population. In [7], multiple simultaneous transmissions employing spread spectrum with fairness constraints was considered and [8] presented algorithms for scheduling users with average delay considerations. The works in [9], [11]–[14] studied transmission power/rate adaptation. In [9], [11] the goal of the scheduling policy was to ensure queue stability, in [12] the aim was to minimize transmission power subject to average delay constraints whereas [13], [14] considered explicit hard deadline constraints over point to point communication. Our work in this paper differs from the above by presenting a different formulation that combines the QoS and the BE classes of service. We adopt a geometric approach to the problem and show that the optimal policy satisfies a special structure. The geometric analysis is valid for a general fading model and hence is applicable for a wide set of scenarios. In the special case of Rayleigh fading we further obtain closed-form formulas for the various performance metrics. Part of the work in this paper has been presented earlier in [3].

The rest of the paper is organized as follows. In Section II, we present the system model and the problem description. In Section III, we present the geometric approach to the problem and obtain the optimal policy. The throughput results for Rayleigh fading are presented in Section IV; simulation results

[1] 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.

comparing the optimal and the random scheduling policy for Rayleigh and Nakagami fading are presented in Section V; and Section VI concludes the work.

II. System and Problem Description

A. System Model

We consider the wireless downlink scenario, namely, communication from the base station (the transmitter) to the mobile handsets (the receivers, also referred as users) in a time slotted system. There are multiple users in the system, each user experiencing time-varying channel condition. The channel state of a user remains constant for a single time slot but changes over multiple time slots. We assume that the channel stochastic process is stationary and ergodic. This assumption does not preclude channel correlations over time and among the users, thus allowing the possibility of channel states over multiple time-slots to be dependent. 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 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 [1] and is a commonly used model in the literature for the wireless downlink [4], [5], [7], [8].

Let $\bar{\mathbf{r}} = \{r_i\}$ denote the vector of communication rates 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 in that slot is simply r_i. We will refer to r_i as the "*channel rate*" for user i and $\bar{\mathbf{r}}$ as the "*channel rate vector*". The transmitter has knowledge of $\bar{\mathbf{r}}$ at the beginning of slot k but does not know this vector for future slots. In the k^{th} time-slot, $\bar{\mathbf{r}}$ is a particular realization from the set comprising all possible channel rate vectors whose probability distribution depends on the 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 given that the channel rate vector in that slot is $\bar{\mathbf{r}}$. A *stationary scheduling policy*, denoted $\Gamma(\bar{\mathbf{r}})$, is one that depends solely on $\bar{\mathbf{r}}$ but does not depend on the time index. Clearly, such a policy can be represented as a map from the set of channel rate vectors to the user index; namely, each $\bar{\mathbf{r}}$ is mapped to a unique user index. As the underlying processes are stationary, we restrict attention in this paper to stationary scheduling policies and such a restriction suffices.

B. Problem Description

The set of users in the system are divided into two service classes: (i) throughput rate guaranteed (QoS) users and (ii) "best effort" (BE) users. As mentioned earlier, QoS users represent session applications that require the base station to provide a certain data rate on the downlink, whereas, the BE users represent low priority data transfer applications which do not have a rate requirement and are short-lived. The number of BE users is assumed large and being short-lived it changes rapidly over time. In such a setup, the objective at the base station is to provide the throughput rates to the QoS users with the least time-slot utilization so that the remaining fraction of time-slots allocated for serving the BE class is maximized[2]. The scheduling problem now is to obtain a rule that assigns time-slots dynamically over time to meet the above objective.

Let there be N QoS users in the system and denote the channel rate vector for these users as $\bar{\mathbf{r}} = (r_1, \ldots, r_N)$. Let $X_i(\bar{\mathbf{r}})$ denote the throughput per time-slot of user i. We have[3],

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

The expected throughput per time slot is $E[X_i(\bar{\mathbf{r}})]$. Under stationarity of the scheduling rule, it is easy to see that $X_i(\bar{\mathbf{r}})$ is stationary and ergodic and that $E[X_i(\bar{\mathbf{r}})]$ equals the long term time-average throughput per slot (called throughput rate) of user i. Let $\bar{\mathbf{R}} = (R_1, .., R_N)$ be the guaranteed throughput rates to the QoS users. We will assume that $\bar{\mathbf{R}}$ is *feasible* and by feasibility we mean that there exists at least one scheduling policy that achieves the throughput rates, i.e. $E[X_i(\bar{\mathbf{r}})] \geq R_i, \forall i = 1, .., N$ for some policy.

Let $I_i(\bar{\mathbf{r}})$ be the indicator function for selection of user i,

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

With this notation we can re-write $X_i(\bar{\mathbf{r}})$ as $X_i(\bar{\mathbf{r}}) = r_i I_i(\bar{\mathbf{r}})$. The optimization problem can now be formally stated as follows,

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

The expectation above is taken over the joint distribution of the channel rate vector, $\bar{\mathbf{r}}$, for the N QoS users. Note that minimizing $\sum_{i=1}^{N} E[I_i(\bar{\mathbf{r}})]$ is equivalent to maximizing $1 - \sum_{i=1}^{N} E[I_i(\bar{\mathbf{r}})]$ 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 also assume that $\bar{\mathbf{R}}$ is away from the boundary of the set, which is characterized later, comprising all feasible 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 the paper.

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

[2] We assume that among the BE users a greedy algorithm is used to share the slots that are allocated for the BE class. With a large population of BE users there is a high probability of at least one user experiencing good channel condition. Thus, maximizing the time-slot allocation is then equivalent to maximizing the sum total throughput of BE users.

[3] For notational simplicity, explicit dependence of $X_i(\cdot)$ on Γ is not indicated. Also, since the service of BE users is simply the fraction of allocated time-slots to that class, their channel rate vector is not required for the optimization.

III. Optimal Policy

The QoS users experience different time-varying channel conditions, hence, intuitively the optimal policy must exploit this variability giving preference to users with better channel conditions. This would ensure a high throughput per slot and would lead to a fewer fraction of time-slots being utilized to provide the throughput guarantee. However, simply choosing the best user is not sufficient since the throughput requirements of the QoS users and their channel statistics might be very different which necessitates that these parameters must also be taken into account.

Let Ω be the set comprising all possible channel rate vectors, $\bar{\mathbf{r}}$; we have $\Omega \subseteq \mathbb{R}^{+N}$. Let the joint probability density function be $f(\bar{\mathbf{r}})$ [4] so that the probability of a subset $Z \subset \Omega$ is given as $\int_Z f(\bar{\mathbf{r}}) d\bar{\mathbf{r}}$. We assume that $f(\bar{\mathbf{r}})$ is such that subsets with zero volume in Ω (or individual points) have zero probability, thus, excluding point mass distributions. Since a scheduling policy maps $\bar{\mathbf{r}} \in \Omega$ to a unique user index, we will represent a scheduling policy 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 channel rate vector $\bar{\mathbf{r}}$ lies within region 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[5]. The problem thus reduces to choosing these regions optimally to minimize the objective function and satisfy the throughput rate constraint, $\int_{Z_i} r_i f(\bar{\mathbf{r}}) d\bar{\mathbf{r}} \geq R_i,\ i = 1, \ldots, N$.

In the rest of the paper, the notation $\bar{\mathbf{r}} \rightarrow Z$ ($\bar{\mathbf{r}} \not\rightarrow Z$) means that there is a neighborhood around $\bar{\mathbf{r}}$ that is mapped (is not mapped) to region Z and the probability of this neighborhood is non-zero. Formally, $\bar{\mathbf{r}} \rightarrow Z$ implies that there exists $\epsilon > 0$ such that all $\hat{\mathbf{r}} \in \Omega,\ ||\hat{\mathbf{r}} - \bar{\mathbf{r}}|| < \epsilon \Rightarrow \hat{\mathbf{r}} \in Z$ and $\int_{||\hat{\mathbf{r}} - \bar{\mathbf{r}}|| < \epsilon} f(\hat{\mathbf{r}}) d\hat{\mathbf{r}} > 0$; where the norm $||\cdot||$ is the Euclidean distance norm in $\mathbb{R}^{\mathbb{N}}$. The following two lemmas give the properties of the optimal $Z_1, \ldots, Z_N, Z_f$ regions. The first lemma deals with the region Z_f and it states that if $\bar{\mathbf{r}}$ is mapped to Z_i, all rate vectors with the 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* $\hat{\mathbf{r}} = (\hat{r}_1, .., (\hat{r}_i > r_i), .., \hat{r}_N) \not\rightarrow Z_f$.

Proof: Appendix I ■

A careful observation of Lemma 1 yields a special structure on Z_f as follows. Let a_1 be the infimum value of the first component among all vectors $\bar{\mathbf{r}} \rightarrow Z_1$; i.e. $a_1 = \inf_{(\bar{\mathbf{r}} \rightarrow Z_1)} r_1$. Now, any $\hat{\mathbf{r}} \rightarrow Z_f$ must be such that $\hat{r}_1 \leq a_1$; otherwise Lemma 1 will be violated. As this holds for all Z_i, an optimal policy has constants $\{a_i\}$ where $a_i = \inf_{(\bar{\mathbf{r}} \rightarrow Z_i)} r_i$ such that if $r_i \leq a_i, \forall i$, then $\bar{\mathbf{r}} \in Z_f$. The region Z_f is shown in Figure 1.

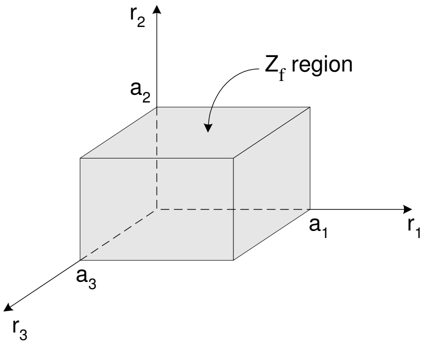

Fig. 1. The Z_f region for $N = 3$, threshold vector $\bar{\mathbf{a}} = (a_1, a_2, a_3)$ and $\Omega = \mathbb{R}^{+N}$. Note that $Z_f = \{\bar{\mathbf{r}} : 0 \leq r_i \leq a_i,\ \forall i = 1, \ldots, N\}$.

This implication is quite intuitive as it suggests that when the channel 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 required throughput vector $\bar{\mathbf{R}}$ for the QoS users 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 $Z_f = \{\bar{\mathbf{r}} \mid \bar{\mathbf{r}} \in \Omega, r_i \leq a_i \forall i\}$ 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* $Z_i,\ Z_j,\ j \neq i$ *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 $\bar{\mathbf{r}} \not\rightarrow Z_j$

Proof: Appendix II ■

The above lemma states that if the weighted comparison of i^{th} and j^{th} component of $\bar{\mathbf{r}}$ is in favour of the i^{th} component (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 geometric structure for the optimal policy.

Theorem *I: **(Optimal Structure)** Consider a channel rate vector* $\bar{\mathbf{r}} = (r_1, \ldots, r_N)$, *then, under the optimal policy there exists a threshold vector* $\bar{\mathbf{a}}$ *with the following structure.*

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

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

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

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

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

3)

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

[4] To avoid excessive notations, $\bar{\mathbf{r}}$, depending on the context denotes both a random vector and a particular realization for a generic time-slot.

[5] To eliminate uninteresting partitions the following technical assumptions are made. The set Ω can be partitioned into a finite set of components, where, each component is a connected set with non-zero volume and every point of this set is arbitrarily close to an interior point. Such an assumption removes the trivial point/zero volume sets. A scheduling policy is a partition as above and each region Z_i is a finite union of the component sets of the partition. Further, we assume that for set Ω non-zero volume sub-sets that have zero probability have already been removed as their mapping plays no role in the optimization.

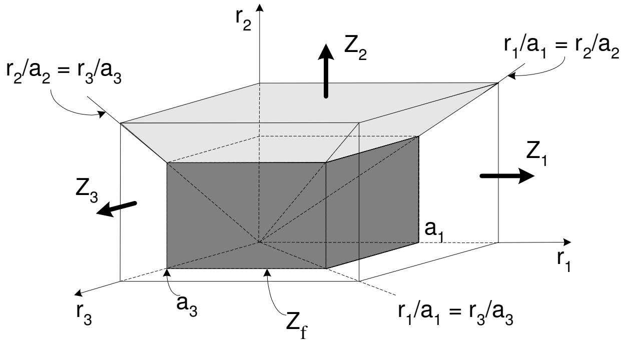

Fig. 2. Optimal policy structure for $N = 3$, threshold vector $\bar{\mathbf{a}} = (a_1, a_2, a_3)$ and $\Omega = \mathbb{R}^{+N}$. The Z_i regions are top truncated pyramids.

Proof: Conditions 1 and 2 follow from Lemmas 1 and 2. Condition 3 states the obvious requirement that for optimality the throughput constraint must be met with equality; since, otherwise the excess fraction of slots that lead to a throughput above R_i can be assigned to the BE users. ■

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 (see, for example the light shaded Z_2 region) and it can be verified that in this region, (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 function 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 (with ties broken arbitrarily) as follows,

$$\Gamma(\bar{\mathbf{r}}) = \begin{cases} Z_f \text{ (serve BE class)}\ , & \text{if } r_i \leq a_i, \forall i = 1, .., N \\ \text{argmax}_i\ \frac{r_i}{a_i}\ , & \text{otherwise} \end{cases} \tag{9}$$

where $\{a_i\}$ are such that $E[r_i I_i] = R_i, \forall i = 1, \ldots, N$.

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 III. ■

Thus, Theorem I states that the optimal policy must satisfy certain conditions which impose a weighted comparison structure on it and conversely, Theorem II completes the argument by stating that a policy with that structure is optimal.

The results presented so far for the optimal policy assumed that $\bar{\mathbf{R}}$ was feasible, that is, it assumed that the optimization problem in (3) had a solution and the throughput rate $\bar{\mathbf{R}}$ could be guaranteed by some scheduling policy. We now go back and characterize the set of all such feasible throughput rate vectors. Let Π denote this set; we claim that the interior of Π is generated by considering each threshold vector $\bar{\mathbf{a}} > 0$ and obtaining the corresponding $\bar{\mathbf{R}}$ that can be achieved for the policy in (9) for that particular $\bar{\mathbf{a}}$. To see why this is true consider the following. Given any $\bar{\mathbf{a}} > 0$, we first construct a policy as given in (9). Since this is a valid scheduling policy the corresponding $\bar{\mathbf{R}}$ with $R_i = E[r_i I_i]$ is feasible; hence, Π must at least include all such $\bar{\mathbf{R}}$. Now, conversely, pick a feasible $\bar{\mathbf{R}}$ in the interior of Π, then, from Theorem I we see that a scheduling policy can be re-mapped to have the optimal geometric structure or equivalently there exists $\bar{\mathbf{a}} > 0$ for which the policy in (9) is optimal.

For a given $\bar{\mathbf{R}}$, we know from (8) that the threshold vector $\bar{\mathbf{a}}$ for the optimal policy 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 numerous techniques of finding the positive root of a non-linear vector equation. In practice, however, the density function $f(\bar{\mathbf{r}})$ may not be known apriori in which case the vector $\bar{\mathbf{a}}$ can be adjusted in real time using stochastic approximation algorithms similar to those outlined in [4], [5]. For a comprehensive and thorough treatment of stochastic approximation algorithms see [17]. We now consider the special case of Rayleigh fading in the next section and obtain explicit expressions for various system metrics.

IV. Dimensioning

In this section, we apply the general results obtained in the last section to a Rayleigh fading scenario. From a practical perspective while such a fading model might be restrictive, nevertheless, from a systems viewpoint the closed form formulas obtained provide important tradeoff limits between the allocation of resources to the QoS and the BE users and can be used as a first cut calculation in system design. For other fading distributions a similar analysis can be carried out, albeit, closed form expressions may not always be possible and certain quantities would need to be evaluated numerically, as done in Section V for an illustrative Nakagami fading scenario.

To proceed, we consider the following specializations to the earlier model. The users experience independent identically distributed (i.i.d) flat Rayleigh fading, hence, $|h|^2$ is Exponentially distributed, where $|h|$ is the magnitude of the channel gain/fade state. 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 and P is the transmission power. A linear power-rate relationship is a good model in various scenarios such as the low SNR regime in which most CDMA systems operate, ultra-wideband transmission and fixed modulation schemes and has been studied earlier in the literature [15]. 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. Lastly, we take the guaranteed throughput rate the same for all N QoS users, namely, $\bar{\mathbf{R}} = (R, \ldots, R)$.

A. Throughput Characterization

Let γ denote the fraction of time-slots allocated to the BE users. We first obtain the threshold vector in terms of γ as follows. Due to symmetry in $f(\bar{\mathbf{r}})$ and $\bar{\mathbf{R}}$, clearly, the regions

$Z_i,\ i = 1,..,N$ are identical, hence, the $\{a_i\}$'s for the optimal policy are equal and the threshold vector is given as $\bar{\mathbf{a}} = (a,..,a)$. Now, the threshold value a in terms of γ is as follows.

***Lemma* 3:** *Let γ be the fraction of time-slots allocated to the BE users, then 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 $Z_f = \{\bar{\mathbf{r}} : 0 \le r_i \le a,\ \forall i = 1,\dots,N\}$. 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,

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

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 $\gamma = 0 \Rightarrow a = 0$ which agrees with the fact that $\gamma = 0$ (no slot for the BE users) implies Z_f is null and similarly, $\gamma = 1 \Rightarrow a \to \infty$ which agrees with the fact that $\gamma = 1$ (all slots for the BE users) implies $Z_f = \mathbb{R}^{+N}$.

Now, using the optimal structure of region Z_i we can obtain an expression for the required throughput rate R in terms of the threshold value a.

***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 $\bar{\mathbf{a}} = (a,\dots,a)$, the region Z_i is given as, $Z_i = \{\bar{\mathbf{r}} : a \le r_i < \infty,\ 0 \le r_j \le r_i,\ j \ne i\}$. As $R = E[r_i I_i]$ we get,

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

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 get (13). ■

Note from (13) that R is monotonically decreasing in a, hence there is a one to one relationship between R and a. Stated equivalently, given a certain R value, there is a unique threshold $a \ge 0$ that achieves it. 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 (iii) Number of QoS users, N, and (iv) The average channel condition, μ, of the users.

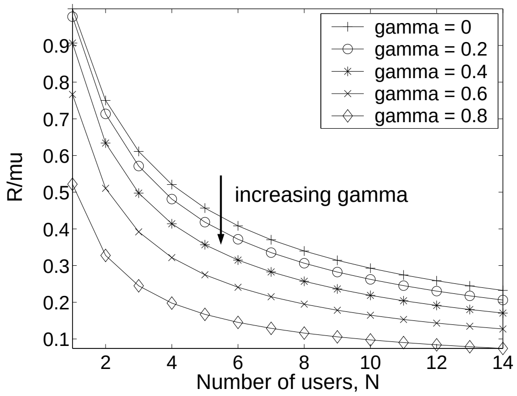

Fig. 3. Plot of R/μ versus N for the optimal policy for various γ values.

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 by,*

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

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

From (16), we see that R depends linearly on μ, thus as expected, for a given N, γ, the throughput guarantee is higher if μ is increased. Now, re-phrasing (16), theoretical limits for various performance measures can be deduced as follows.

Maximum Throughput Rate: By setting $\gamma = 0$, we can obtain the maximum throughput rate $R_{max}(N)$ for each QoS user when no slots are allocated for the BE users. This is given as,

$$R_{max}(N) = \mu \left(\sum_{k=0}^{N-1} \binom{N-1}{k} (-1)^k \frac{1}{(k+1)^2}\right) \tag{17}$$

Figure 3 is a plot of R/μ versus N for different γ values. The function $R_{max}(N)/\mu$ is the topmost curve corresponding to $\gamma = 0$. As $R_{max}(N)$ is monotonically decreasing in N, its maximum value is at $N = 1$ and equals $R_{max}/\mu = 1$. This is expected as the maximum rate achievable when all the slots are assigned to just one QoS user equals $E[r]$ $(= \mu)$.

Maximum Number of QoS Users: Fix a value of R_0 and γ, the maximum number of QoS users such that throughput of each is at least R_0 is given by,

$$N_{max}(R_0, \gamma) = \max_N \ (R \ge R_0) \tag{18}$$

Obviously if the values of R_0, γ are such that there is no integer $N \ge 1$ that achieves it, the system values in this case are infeasible. Figure 4 is a plot of R/μ versus γ for various values of N. Infeasibility arises when $(\gamma, R_0/\mu)$ point lies above the $N = 1$ curve (in Fig. 4).

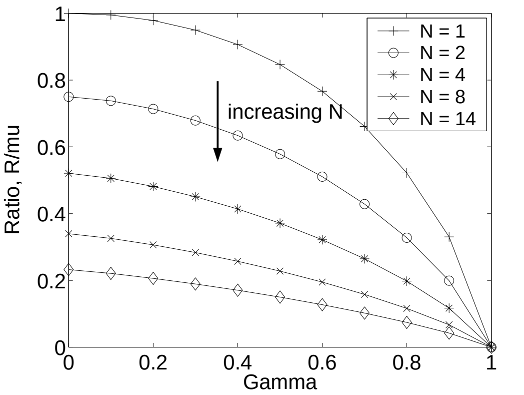

Fig. 4. Plot of R/μ versus γ for values of $N = 1, 2, 4, 8, 14$.

Maximum Value of γ: Given R and N, the value of γ that solves the equation in (16) gives the maximum fraction of slots that can be allocated to the BE users. Figure 4 with its axes inverted gives a plot of γ versus R/μ for different N.

B. Comparison with Random-scheduling Policy

To understand how much gain can be achieved, we present an analytical comparison of the optimal policy with the random scheduling policy. 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 $1/N$. Clearly, this policy does not exploit the varying channel conditions for scheduling the users. Due to the random nature of the assignment each QoS user gets $(1 - \gamma)/N$ fraction of time-slots and since the users have statistically identical channel conditions, the throughput rate of each QoS user, denoted R_r, is given as,

$$R_r = \mu \frac{(1-\gamma)}{N} \tag{19}$$

Let us now fix a value of γ for both the optimal and the random policies, i.e. under both policies, γ fraction of slots are assigned to the BE class. Let R^{opt}, R_r denote the corresponding throughput rate provided to each QoS user. Then, as shown below, the gain defined as R^{opt}/R_r is on the order of $\ln(N)$. To show this result, we need the following lemma.

***Lemma* 5:** *For any $\gamma \in (0, 1)$, we have the following relationship*[6]

$$\ln\left(\frac{1}{1-\gamma^{\frac{1}{N}}}\right) = \Theta(\ln(N)) \tag{20}$$

Proof: Appendix IV ■

Theorem *IV: The throughput gain of the optimal policy as compared to the random policy, defined as R^{opt}/R_r, for $\gamma \in (0, 1)$ satisfies the relationship,*

$$\frac{R^{opt}}{R_r} = \Theta(\ln(N)) \tag{21}$$

Proof: Appendix V ■

Observe that as $N \to \infty$ the throughput rate per QoS user for both the optimal and the random policy tends to zero. Equation (19) states that R_r decreases as $1/N$ whereas (45) in Appendix V states that by using the optimal policy R^{opt} decreases more slowly as $\ln(N)/N$. Hence, we get a gain on the order of $\ln(N)$. The above logarithmic behavior can be attributed to the exponential distribution of the rate under Rayleigh fading and while such channel statistics are simplified models, in practice one could expect gains along these orders for moderate QoS user population.

V. Simulation Results

To validate the theoretical results derived in the earlier sections, we present simulation results obtained for two fading distributions, Rayleigh and Nakagami. The setup for the simulations is as follows: we consider a time duration of 10 seconds and divide it into 10,000 slots, thus, each time-slot is of length 1 millisecond. For the sake of simplicity, the QoS users all experience i.i.d channel fading. We assume a linear relationship between the channel rate and the fade state (squared magnitude); i.e. $r \propto |h|^2$. Thus, for Rayleigh fading the rate, r, at which data can be transmitted in a slot is Exponentially distributed with density $f(r) = \frac{e^{-r/\mu}}{\mu},\ r \geq 0$; while for Nakagami fading, r has a Gamma distribution given as $f(r) = \left(\frac{m}{\mu}\right)^m \frac{r^{m-1}}{\Gamma(m)} e^{-mr/\mu},\ r \geq 0$, where m is the fading parameter [16]. The mean channel rate, μ, for each user is taken as, $\mu = 800$ Kbits/sec for both the distributions. At each time-slot, a random vector of channel rates for the QoS users is drawn from the respective distribution. Given this channel rate vector, the particular scheduling policy decides which QoS user to serve or to allocate the slot to the BE class. In the former case, the chosen QoS user, say user i, receives a throughput rate of r_i while for the others the throughput rate is 0 in that slot. In the latter case, all QoS users get a 0 throughput in that slot.

We simulate the optimal, the random, the greedy Time Division Multi-Access (TDMA) and an opportunistic scheduling policy studied in [10] which we refer to as "*Opportunistic Proportional Fair*" (OPF) policy. In case of the optimal policy, the scheduling decision is taken as given in (9) where the threshold vector $\bar{\mathbf{a}}$ is computed using the formulas in Section IV. The random policy makes a scheduling decision as described in Section IV-B. For the greedy TDMA and the OPF policy the scheduling decision is taken as follows. Let T_k denote the running time-average of the throughput rate for the k^{th} QoS user. At the beginning of each time-slot, consider all QoS users for which $T_k < R$ where R is the required throughput guarantee. In the greedy TDMA policy the user with the maximum channel rate is selected whereas for the OPF policy the user that maximizes the metric r_k/T_k is selected. If for all QoS users $T_k \geq R$, the slot is allocated to the BE class.

We first numerically validate the theoretical results obtained in Section IV. We consider Rayleigh fading with 3 QoS users each having a throughput rate guarantee of $R = 200$ Kbits/sec.

[6] The following notation is followed: (i) $f(N) = O(g(N))$ means that there exists a constant c and integer N_0 such that $f(N) \leq cg(N)$ for $N > N_0$, (ii) $f(N) = \Theta(g(N))$ means that $f(N) = O(g(N))$ and $g(N) = O(f(N))$

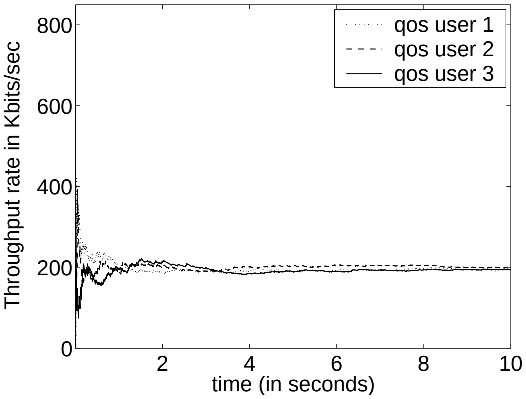

Fig. 5. Running time-average of throughput rate for Rayleigh fading with 3 QoS users, $R = 200$ Kbits/sec.

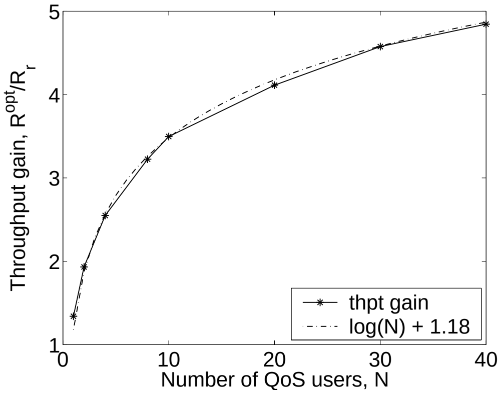

Fig. 6. Throughput gain, R^{opt}/R_r, for Rayleigh fading with $\gamma = 0.3$.

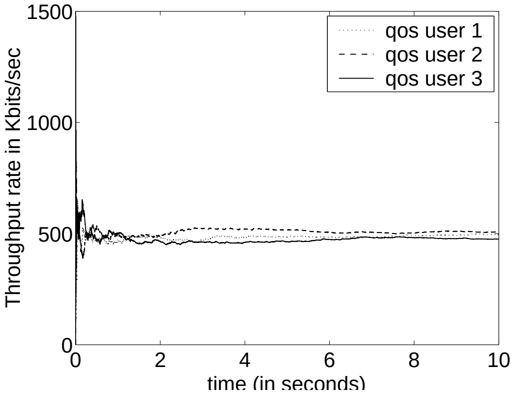

Fig. 7. Running time-average of throughput rate for Nakagami fading with fade parameter $m = 0.6$, $\gamma = 0.3$ and 3 QoS users.

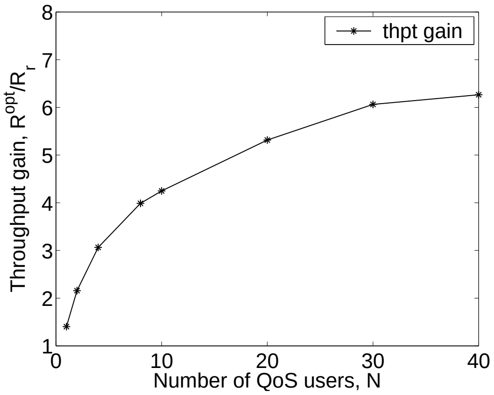

Fig. 8. Throughput gain, R^{opt}/R_r, for Nakagami fading with fade parameter $m = 0.6$ and $\gamma = 0.3$.

Figure 5 gives a plot of the running time-average of throughput rate under the optimal policy. As can be seen from the plot, the long-term required rate is achieved very quickly in time within almost a second and is maintained thereafter within a close range. Thus, within a very short time interval the required throughput rate can be provided to the QoS users. A similar trend is observed when the parameter values are varied. In Figure 6, we fix $\gamma = 0.3$, i.e. the BE class is assigned 30% of the slots. The figure gives a plot of the simulated throughput gain R^{opt}/R_r as a function of N; where R^{opt}, R_r is the throughput rate of each QoS user under the optimal and the random policy respectively. In conformation with the result in (21), we see from the plot that $\frac{R^{opt}}{R_r}$ grows logarithmic in N. We next consider Nakagami fading with the fading parameter $m = 0.6$. In Figure 7, we fix $\gamma = 0.3$ and plot the running time-average of the throughput rate for the optimal policy with 3 QoS users. For the case of Nakagami fading, (11) becomes, $\int_0^{\frac{ma}{\mu}} t^{m-1}e^{-t}dt = \gamma^{\frac{1}{N}}\Gamma(m)$ from which the optimal threshold a is evaluated numerically by finding the root of the above non-linear equation. The long-term rate provided to each QoS user in this case is $R = 494$ Kbits/sec. Again as before, the throughput rate is achieved very quickly in time and is maintained thereafter within a close range. In Figure 8 we compare the throughput gain of the optimal policy versus the random policy. As seen from the plot the optimal policy achieves a substantial gain in throughput even with Nakagami distribution. In fact, the gain is higher now because the Gamma distribution with $m = 0.6$ has a larger variance than the Exponential with the same mean. As a result, the optimal policy which opportunistically exploits rate variations gives a higher gain in comparison to random assignment.

We now present simulation results that compare the performance of the optimal, random, TDMA and OPF policies. We consider 3 QoS users with Rayleigh fading and the mean channel rate of each QoS user, $\mu = 800$ Kbits/sec. Figure 9 plots the total fraction of slots utilized by the QoS users under each policy versus the throughput rate requirement of each QoS user. The quantity, (1− total fraction of slots used by QoS users), is the time-slot allocation to the BE class. First, as expected the random policy has the worst performance and utilizes the maximum time-slots to provide the throughput rate guarantees. Since the OPF, TDMA and optimal policy exploit the channel variations and opportunistically schedule the users, the time-slot utilization is lower as compared to the random policy. The OPF policy performs worse than the TDMA policy which is expected since the TDMA policy by being greedy has a high throughput per slot and hence utilizes fewer time-slots. Finally, as expected the optimal policy uses a substantially lower fraction of time-slots than all the policies.

VI. Conclusion

We addressed the issue of downlink scheduling over a wire-

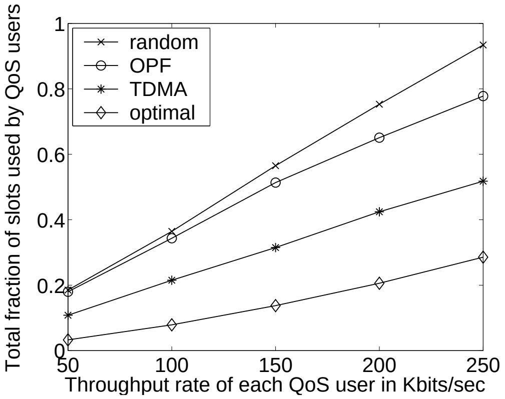

Fig. 9. Comparison of the fraction of slots utilized by the random, OPF, TDMA and optimal policies.

less channel incorporating the QoS and best effort services. We considered a set of N rate guaranteed users and obtained the optimal policy that serves these users with the least time-slot utilization, thereby, maximizing the time-slot allocation to the BE users. Equivalently, the optimal policy also solves the problem of maximizing the rate guarantee for the QoS users given that a certain fraction of time-slots must be allocated to the BE users. We presented a geometric visualization of the optimal policy and under Rayleigh fading we derived analytical expressions quantifying the various system metrics. Analytical comparison with the random-scheduling policy showed that throughput gains on the order of $\ln(N)$ can be achieved by exploiting multi-user diversity. Finally simulation results show substantial gains achieved by the optimal policy as compared to other well-known policies in the literature.

References

[1] 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.

[2] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijaykumar, "Providing quality of service over a shared wireless link", *IEEE Communications Magazine*, pp. 150-154, Feb. 2001.

[3] M. Zafer and E. Modiano, "Joint Scheduling of Rate-guaranteed and Best-effort Services over a Wireless Channel", *IEEE CDC-ECC 2005*, Seville, Spain, Dec. 2005.

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

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

[6] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks", *IEEE/ACM Transactions on Networking*, vol. 13, no. 3, pp. 636-647, 2005.

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

[8] 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.

[9] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity", *IEEE Transactions on Information Theory*, vol. 39, pp. 466478, Mar. 1993.

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

[11] M. Neely, E. Modiano, and C. Rohrs, "Power Allocation and Routing in Multi-Beam Satellites with Time Varying Channels," *IEEE Transactions on Networking*, vol. 11, no. 1, pp. 138-152, Feb. 2003.

[12] D. Rajan, A. Sabharwal, B. Aazhang, "Delay bounded packet scheduling of bursty traffic over wireless channels", *IEEE Transactions on Information Theory*, , vol. 50, 1, pp. 125-144, Jan. 2004.

[13] M. Zafer, E. Modiano, "A Calculus Approach to Minimum Energy Transmission Policies with Quality of Service Guarantees", *Proceedings of the IEEE INFOCOM 2005*, vol. 1, pp. 548-559, March 2005.

[14] M. Zafer, E. Modiano, "Continuous-time Optimal Rate Control for Delay Constrained Data Transmission", *43rd Annual Allerton Conference on Communication, Control and Computing, Monticello*, Sept. 2005.

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

[16] M. Nakagami, "The m-distribution - A general formula of intensity distribution of fading," *Statistical Methods in Radio Wave Propagation*, W. C. Hoffman, Ed. London, England: Pergamon, 1960

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

[18] Gradshteyn I.S., Ryzhik I.M., "*Table of Integrals, Series and Products*", Academic Press,

Appendix I
Proof of Lemma 1

The proof is based on a contradiction argument where we begin by supposing that for the optimal policy there is a $\hat{\mathbf{r}} \rightarrow Z_f$ with $\hat{r}_i > r_i$. By re-mapping the regions we will show that the objective function in (3) decreases, thus, contradicting the optimality claim and proving $\hat{\mathbf{r}} \not\rightarrow Z_f$.

We are given that $\bar{\mathbf{r}} \rightarrow Z_i$, hence, there is a neighborhood of $\bar{\mathbf{r}}$, which we denote as S_1, that is mapped to Z_i, i.e. $S_1 \in Z_i$ and $S_1 = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in \Omega, ||\bar{\mathbf{x}} - \bar{\mathbf{r}}|| < \delta_1\}$ for some $\delta_1 > 0$. Further, by assumption $\hat{\mathbf{r}} \rightarrow Z_f$, there is a neighborhood of $\hat{\mathbf{r}}$ given as, $S_2 = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in \Omega, ||\bar{\mathbf{x}} - \hat{\mathbf{r}}|| < \delta_2\}$ for some $\delta_2 > 0$, such that $S_2 \in Z_f$.

Now re-map the regions as follows. Map $S_1 \Rightarrow Z_f$ and $S_2 \Rightarrow Z_i$. To ensure the new mapping is feasible we must satisfy the QoS rate constraint for user i which entails the following equality.

$$\int_{S_2} x_i f(\bar{\mathbf{x}}) d\bar{\mathbf{x}} = \int_{S_1} x_i f(\bar{\mathbf{x}}) d\bar{\mathbf{x}} \tag{22}$$

The left side above is the throughput achieved over region S_2 under the new map and the right side is the throughput lost by re-mapping S_1 to Z_f. A set of $\delta_1, \delta_2 > 0$ exist that satisfy (22); to see this note that the integral over any region $\{S_k\}_{k=1}^2$ is a positive, continuous function with respect to δ_k, non-increasing as δ_k decreases and tends to zero as $\delta_k \downarrow 0$. Hence, starting with the largest δ_1, δ_2 values (that satisfy the S_1, S_2 definition) and then decreasing these values one can obtain $\{\delta_1, \delta_2 > 0\}$ such that each integral above is positive and the two are equal. Now, viewing δ_2 as a function of δ_1, it's clear that if a solution exists for some δ_1^0 then for all $\delta_1 \leq \delta_1^0$ a solution exists by the continuity and decreasing property of the integrals. We now proceed by choosing $\delta_1 \leq \delta_1^0$.

Using the First Mean Value theorem, [18], we can take the x_i outside the integrals as follows, $\int_{S_1} x_i f(\bar{\mathbf{x}}) d\bar{\mathbf{x}} = (r_i + \epsilon_1) \int_{S_1} f(\bar{\mathbf{x}}) d\bar{\mathbf{x}}$ and $\int_{S_2} x_i f(\bar{\mathbf{x}}) d\bar{\mathbf{x}} = (\hat{r}_i + \epsilon_2) \int_{S_2} f(\bar{\mathbf{x}}) d\bar{\mathbf{x}}$, where the $\{\epsilon_k\}_{k=1}^2$ depend on $\{\delta_k\}_{k=1}^2$ or equivalently on δ_1 (as δ_2 depends on δ_1 through (22)). With this, we can re-write (22) as,

$$(\hat{r}_i + \epsilon_2) \int_{S_2} f(\bar{\mathbf{x}}) d\bar{\mathbf{x}} = (r_i + \epsilon_1) \int_{S_1} f(\bar{\mathbf{x}}) d\bar{\mathbf{x}} \tag{23}$$

Now, looking at the objective function in (3), the change in its value due to the re-map equals the probability of region

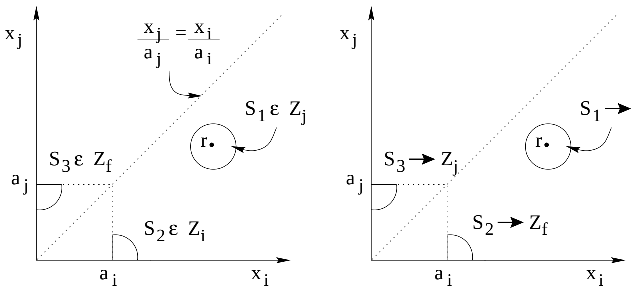

Fig. (a): Original mapping

Fig. (b): New mapping

Fig. 10. Figure showing the mappings for the proof of Lemma 2.

S_2 (added from Z_f to Z_i) minus the probability of region S_1 (removed from Z_i). Thus,

$$\begin{aligned}\Delta J &= -\int_{S_1} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} + \int_{S_2} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \\ &= -\left(\frac{\hat{r}_i + \epsilon_2}{r_i + \epsilon_1} - 1\right)\int_{S_2} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \qquad (24)\end{aligned}$$

Let $c = \hat{r}_i - r_i$, then, $c > 0$ (since by assumption $\hat{r}_i > r_i$). Using the First Mean Value theorem, we also have $\epsilon_k \to 0$ as $\delta_k \to 0$. Thus, for any c we can scale δ_1 to be small enough such that $\left(\frac{\hat{r}_i+\epsilon_2}{r_i+\epsilon_1} - 1\right) > 0$. Further, since the integral in (24) is the probability of S_2 which is strictly positive (regions with zero probability are uninteresting and have been removed from Ω), we finally get, $\Delta J < 0$. This completes the contradiction argument.

Appendix II
Proof of Lemma 2

The proof is based on a contradiction argument. To begin, consider $\bar{\mathbf{r}} \notin Z_f$ and suppose that for the optimal policy, $\bar{\mathbf{r}} \to Z_j$ such that,

$$\frac{r_i}{a_i} > \frac{r_j}{a_j} \qquad (25)$$

We now give a re-mapping of the regions such that the objective function in (3) decreases or equivalently the probability of Z_f region increases, thus, proving 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 $\bar{\mathbf{x}} \in \Omega$ denote a generic rate vector. Since by assumption $\bar{\mathbf{r}} \to Z_j$, there is a neighborhood around $\bar{\mathbf{r}}$ given as $S_1 = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in \Omega, ||\bar{\mathbf{x}} - \bar{\mathbf{r}}|| < \delta_1\}$ for some $\delta_1 > 0$, such that $S_1 \in Z_j$. Next, since the optimal policy satisfies Lemma 1 (its violation would make the policy non-optimal to start with) we know that a_i is the infimum value of the i^{th} component among $\bar{\mathbf{x}} \to Z_i$. 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; i.e. $S_2 \in Z_i$ and $S_2 = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in \Omega, 0 < (x_i - m_i) < \delta_2\}$ for some $\delta_2 > 0$. Finally, since $\bar{\mathbf{R}}$ does not lie on the boundary of feasible throughput vectors the region Z_f is not null. Hence, 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; namely, $S_3 \in Z_f$ and $S_3 = \{\bar{\mathbf{x}} \mid \bar{\mathbf{x}} \in \Omega, 0 < (n_j - x_j) < \delta_3\}$ for some $\delta_3 > 0$. The regions S_1, S_2, S_3 are depicted in Figure 10(a).

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 10(b). To ensure the new mapping is feasible we must satisfy the QoS rate constraints for user i and user j, which entails the following equalities.

$$\begin{aligned}\int_{S_2} x_i f(\bar{\mathbf{x}})d\bar{\mathbf{x}} &= \int_{S_1} x_i f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \qquad (26)\\ \int_{S_3} x_j f(\bar{\mathbf{x}})d\bar{\mathbf{x}} &= \int_{S_1} x_j f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \qquad (27)\end{aligned}$$

Equation (26) matches the throughput lost for user i due to the re-map of $S_2 \Rightarrow Z_f$ and the throughput gained by $S_1 \Rightarrow Z_i$, while (27) gives a similar equality for user j. To see why a set of $\{\delta_k\}_{k=1}^{3}$ exist that solve the above equations, note that the integral over any region S_k is a continuous, positive function of δ_k, decreasing (or non-increasing) as δ_k decreases and tends to zero as $\delta_k \downarrow 0$. Hence, starting with the largest δ_1 (that satisfies the S_1 definition), decrease it until a δ_2 is obtained that solves (26). By the non-nullity of S_1, S_2 and the above property of the integrals such a solution $\delta_1, \delta_2 > 0$ exists. Similarly obtain a δ_1, δ_3 that solves (27). Finally, taking δ_1 as the minimum of the two solutions, re-obtain δ_2, δ_3 such that both (26) and (27) are satisfied. Now, viewing δ_2, δ_3 as functions of δ_1, it's clear that if a solution exists for some δ_1^0, then, for all $\delta_1 \leq \delta_1^0$ a solution exists by the continuity and decreasing property of the integrals. We now proceed by choosing $\delta_1 \leq \delta_1^0$.

Using the First Mean Value theorem, [18], we can re-write the above integrals as,

$$\begin{aligned}(a_i + \epsilon_2)\int_{S_2} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} &= (r_i + \epsilon_1)\int_{S_1} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \qquad (28)\\ (a_j + \epsilon_3)\int_{S_3} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} &= (r_j + \epsilon_4)\int_{S_1} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \qquad (29)\end{aligned}$$

where the $\{\epsilon_k\}$ above depend on the $\{\delta_k\}$ or equivalently on δ_1 (as δ_2, δ_3 depend on δ_1 through (26) and (27)). Next, looking at the objective function in (3), the change in its value due to the re-map equals the probability of region S_3 (added from Z_f to Z_j) minus the probability of region S_2 (removed from Z_i). Thus,

$$\begin{aligned}\Delta J &= -\int_{S_2} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} + \int_{S_3} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \\ &= -\left(\frac{r_i + \epsilon_1}{a_i + \epsilon_2} - \frac{r_j + \epsilon_4}{a_j + \epsilon_3}\right)\int_{S_1} f(\bar{\mathbf{x}})d\bar{\mathbf{x}} \qquad (30)\end{aligned}$$

Let $c = \frac{r_i}{a_i} - \frac{r_j}{a_j}$, then, from (25) we have $c > 0$. From the First Mean Value theorem we also have $\epsilon_k \to 0$ as $\delta_k \to 0$. Thus, for any given c we can scale δ_1 to be small enough such that $\left(\frac{r_i+\epsilon_1}{a_i+\epsilon_2} - \frac{r_j+\epsilon_4}{a_j+\epsilon_3}\right) > 0$. Further, since the integral in (30) is the probability of S_1 which is strictly positive, we finally get $\Delta J < 0$. This completes the proof.

Appendix III
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 $\sum_{i=1}^{N} E[I_i] \leq \sum_{i=1}^{N} E[\tilde{I}_i]$ 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, it is 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,

$$\sum_{i=1}^{N} E[\tilde{I}_i] = \sum_{i=1}^{N} E[\tilde{I}_i] - \sum_{i=1}^{N} \frac{1}{a_i}(E[r_i\tilde{I}_i] - R_i) \tag{31}$$

where $\{a_i\}$ is the threshold vector for policy Γ. Note that the second term in (31) is zero. Re-arranging (31) we get,

$$\sum_{i=1}^{N} E[\tilde{I}_i] = E\left[\sum_{i=1}^{N}\left(1 - \frac{r_i}{a_i}\right)\tilde{I}_i\right] + \sum_{i=1}^{N}\frac{R_i}{a_i} \tag{32}$$

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,

$$\sum_{i=1}^{N}\left(1 - \frac{r_i}{a_i}\right)\tilde{I}_i \ge 0 = \sum_{i=1}^{N}\left(1 - \frac{r_i}{a_i}\right) I_i \tag{33}$$

Case 2: Suppose $r_i > a_i$ for some index i. Let j be the chosen user 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,

$$\sum_{i=1}^{N}\left(1 - \frac{r_i}{a_i}\right)\tilde{I}_i \ge \left(1 - \frac{r_j}{a_j}\right) = \sum_{i=1}^{N}\left(1 - \frac{r_i}{a_i}\right) I_i \tag{34}$$

From (32), (33) and (34) 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 (31) replacing $\tilde{I}_i$ with I_i. This completes the proof.

Appendix IV
Proof of Lemma 5

To prove the lemma we need to show the following two relationships, $\ln(1/(1-\gamma^{\frac{1}{N}})) = O(\ln(N))$ and $\ln(N) = O(\ln(1/(1-\gamma^{\frac{1}{N}})))$. We begin by proving the first relationship. Since $\gamma \in (0,1)$ and $N \ge 1$ is a positive integer, we have $0 < \gamma^{\frac{1}{N}} < 1$. Taking a power series expansion of $\left(\frac{1}{1-\gamma^{\frac{1}{N}}}\right)$ we get,

$$\ln\left(\frac{1}{1-\gamma^{\frac{1}{N}}}\right) = \ln\Big(1+\gamma^{1/N} + \ldots + \gamma^{(N-1)/N} + \gamma(1 + \gamma^{1/N} + \ldots) + \gamma^2(\ldots)\Big) \tag{35}$$

$$= \ln\left(\frac{1 + \gamma^{1/N} + \ldots + \gamma^{(N-1)/N}}{1-\gamma}\right) \tag{36}$$

$$\le \ln\left(\frac{N}{1-\gamma}\right) = \ln(N) - \ln(1-\gamma) \tag{37}$$

The inequality above follows, since $\gamma < 1 \Rightarrow \left(1 + \gamma^{1/N} + \ldots + \gamma^{(N-1)/N}\right) \le N$; thus we get $\ln(1/(1-\gamma^{\frac{1}{N}})) = O(\ln(N))$. To prove the reverse relationship, i.e. $\ln(N) = O(\ln(1/(1-\gamma^{\frac{1}{N}})))$, proceed as follows. Using the standard inequality, $\ln(N) \le 1 + \frac{1}{2} + \ldots + \frac{1}{N-1}$, we get,

$$\begin{aligned}\gamma\ln(N) &\le \gamma + \frac{\gamma}{2} + \ldots + \frac{\gamma}{N-1} \\ &\le \gamma^{1/N} + \frac{\gamma^{2/N}}{2} + \ldots + \frac{\gamma^{N/N}}{N-1} \quad \text{(since } 0 < \gamma < 1\text{)} \\ &\le \ln\left(\frac{1}{1-\gamma^{\frac{1}{N}}}\right) \end{aligned} \tag{38}$$

where the last inequality above follows by truncating the power series expansion of $-\ln(1-\gamma^{\frac{1}{N}})$. Thus, $\ln(N) \le \frac{1}{\gamma}\ln(1/(1-\gamma^{\frac{1}{N}}))$ which gives $\ln(N) = O(\ln(1/(1-\gamma^{\frac{1}{N}})))$.

Appendix V
Proof of Theorem IV

Starting with (16) we can write it as,

$$\begin{aligned}\frac{R}{\mu} &= \ln\left(\frac{1}{1-\gamma^{\frac{1}{N}}}\right)\sum_{k=0}^{N-1}\binom{N-1}{k}\frac{(-1)^k(1-\gamma^{\frac{1}{N}})^{(k+1)}}{k+1} \\ &+ \sum_{k=0}^{N-1}\binom{N-1}{k}\frac{(1-\gamma^{\frac{1}{N}})^{(k+1)}}{(k+1)^2}\end{aligned} \tag{39}$$

Consider the first term in (39) above; it can be evaluated as follows. Let $\alpha = (1-\gamma^{\frac{1}{N}})$, then, since $\gamma \in (0,1)$ we have $\alpha \in (0,1)$.

$$\begin{aligned}\sum_{k=0}^{N-1}\binom{N-1}{k}(-1)^k\frac{\alpha^{(k+1)}}{k+1} &= \sum_{k=0}^{N-1}\binom{N-1}{k}\int_0^\alpha(-x)^k dx \\ &\overset{(a)}{=} \int_0^\alpha (1-x)^{N-1}dx \qquad (40) \\ &= \frac{1-(1-\alpha)^N}{N} = \frac{1-\gamma}{N} \qquad (41)\end{aligned}$$

Equality (a) above follows by interchanging the summation and the integral and using the Binomial expansion. Thus, we get, $\ln\left(\frac{1}{\alpha}\right)\sum_{k=0}^{N-1}\binom{N-1}{k}(-1)^k\frac{\alpha^{(k+1)}}{k+1} = \ln\left(\frac{1}{\alpha}\right)\frac{1-\gamma}{N}$. Now, consider the second term in (39) and proceed as follows. First, since (41) 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 by x and integrating from 0 to α, gives,

$$\begin{aligned}&\int_0^\alpha\sum_{k=0}^{N-1}\binom{N-1}{k}(-1)^k\frac{x^k}{k+1} = \int_0^\alpha\left(\frac{1-(1-x)^N}{Nx}\right)dx \\ &\Rightarrow \sum_{k=0}^{N-1}\binom{N-1}{k}\frac{-1^k\alpha^{k+1}}{(k+1)^2} = \int_0^\alpha\left(\frac{1-(1-x)^N}{Nx}\right)dx \\ &\le \int_0^\alpha dx = \alpha = (1-\gamma^{\frac{1}{N}})\end{aligned} \tag{42}$$

The inequality above follows by noting that $\frac{1-(1-x)^N}{Nx}$ is positive, monotonically non-increasing over $x \in [0,1]$, for fixed $N \ge 1$, and has a maximum value equal to 1 at $x = 0$. Equation (42) further gives, $\frac{N}{1-\gamma}\left(\sum_{k=0}^{N-1}\binom{N-1}{k}\frac{-1^k\alpha^{k+1}}{(k+1)^2}\right) \le$

$\frac{N}{1-\gamma}(1-\gamma^{\frac{1}{N}}) \xrightarrow{N\to\infty} \frac{-\ln(\gamma)}{1-\gamma}$ (which is finite for $0 < \gamma < 1$) and since $\frac{N}{1-\gamma}(1-\gamma^{\frac{1}{N}})$ is monotonically increasing in N with a finite limiting value, it is bounded for all N. Thus, we get,

$$\frac{N}{1-\gamma}\left(\sum_{k=0}^{N-1}\binom{N-1}{k}\frac{-1^k\alpha^{k+1}}{(k+1)^2}\right) \leq \frac{\ln(1/\gamma)}{1-\gamma} \tag{43}$$

Now, using the above simplifications we can re-write (39) as,

$$\frac{R}{\mu} = \frac{1-\gamma}{N}\left(\ln\left(\frac{1}{\alpha}\right) + \frac{N}{1-\gamma}\sum_{k=0}^{N-1}\binom{N-1}{k}\frac{-1^k\alpha^{k+1}}{(k+1)^2}\right) \tag{44}$$

For $\gamma \in (0,1)$, the first term within brackets above, grows as $\ln(\frac{1}{\alpha}) = \Theta(\ln(N))$ (using Lemma 5) whereas the second term is bounded (from (43)). Hence, for large N, R^{opt} can be expressed as,

$$\frac{R^{opt}}{\mu} = \frac{1-\gamma}{N}\Theta(\ln(N)) \tag{45}$$

From (19) and (45) we get the result in (21),