Utilizing Multiuser Diversity for Efficient Support of Quality of Service over a Fading Channel

Dapeng Wu* Rohit Negi†

Abstract

We consider the problem of quality of service (QoS) provisioning for K users sharing a downlink time-slotted fading channel. We develop simple and efficient schemes for admission control, resource allocation, and scheduling, which can yield substantial capacity gain. The efficiency is achieved by virtue of recently identified *multiuser diversity*. A unique feature of our work is *explicit* provisioning of statistical QoS, which is characterized by a data rate, delay bound, and delay-bound violation probability triplet. The results show that compared with a fixed-slot assignment scheme, our approach can substantially increase the statistical delay-constrained capacity of a fading channel (*i.e.*, the maximum data rate achievable with the delay-bound violation probability satisfied), when delay requirements are not very tight, while yet guaranteeing QoS at any delay requirement. For example, in the case of low signal-to-noise-ratio (SNR) and ergodic Rayleigh fading, our scheme can achieve approximately $\sum_{k=1}^{K} \frac{1}{k}$ gain for K users with loose-delay requirements, as expected from the classic paper [10] on multiuser diversity. But more importantly, when the delay bound is not loose, so that simple-minded multiuser-diversity scheduling does not directly apply, our scheme can achieve a capacity gain, and yet meet the QoS requirements.

Key Words: Multiuser diversity, QoS, effective capacity, fading, scheduling, resource allocation.

*Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Tel. (412) 268-7107, Fax (412) 268-1679, Email: dpwu@cs.cmu.edu. URL: http://www.cs.cmu.edu/~dpwu.

†Please direct all correspondence to Prof. Rohit Negi, Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Tel. (412) 268-6264, Fax (412) 268-2860, Email: negi@ece.cmu.edu. URL: http://www.ece.cmu.edu/~negi.

1 Introduction

Providing quality of service (QoS), such as delay and rate guarantees, is an important objective in the design of future packet cellular networks [8]. However, this requirement poses a challenge in wireless network design, because wireless channels have low reliability, and time varying signal strength, which may cause severe QoS violations. Further, the capacity of a wireless channel is severely limited, making efficient bandwidth utilization a priority.

An effective way to increase the capacity of a time-varying channel is the use of diversity. The idea of diversity is to create multiple *independent* signal paths between the transmitter and the receiver so that higher channel capacity can be obtained. Diversity can be achieved over time, space, and frequency. These traditional diversity methods are essentially applicable to a single-user link. Recently, however, Knopp and Humblet [10] introduced another kind of diversity, which is inherent in a wireless network with multiple users sharing a time-varying channel. This diversity, termed *multiuser diversity* [6], comes from the fact that different users usually have *independent* channel gains for the same shared medium. With multiuser diversity, the strategy of maximizing the total Shannon (ergodic) capacity is to allow at any time slot only the user with the best channel to transmit. This strategy is called Knopp and Humblet's (K&H) scheduling. Results [10] have shown that K&H scheduling can increase the total (ergodic) capacity dramatically, in the absence of delay constraints, as compared to the traditionally used (weighted) round robin (RR) scheduling where each user is *a priori* allocated fixed time slots.

The K&H scheduling intends to maximize ergodic capacity, which pertains to situations of infinite tolerable delay. However, under this scheme, a user in a fade of an arbitrarily long period will not be allowed to transmit during this period, resulting in an arbitrarily long delay; therefore, this scheme provides no delay guarantees and thus is not suitable for delay-sensitive applications, such as voice or video. To mitigate this problem, Bettesh and Shamai [2] proposed an algorithm, which strikes a balance between throughput and delay constraints. This algorithm combines K&H scheduling with an RR scheduling, and it can achieve lower delay than K&H scheduling while obtaining a capacity gain over a pure RR scheduling. However, it is very complex to theoretically relate the QoS obtained by this algorithm to the control parameters of the algorithm, and thus cannot be used to guarantee a specified QoS. Furthermore, a direct (Monte Carlo) measurement of QoS obtained, using the queueing behavior resulting from the algorithm, requires an excessively large number of samples, so that it becomes practically infeasible.

Another typical approach is to use dynamic programming [3] to design a scheduler that can increase capacity, while also maintaining QoS guarantees. But this approach suffers from the curse of dimensionality, since the size of the dynamic program state space grows exponentially with the number of users and with the delay requirement.

To address these problems, this paper proposes an approach, which simplifies the task of explicit provisioning of QoS guarantees while achieving efficiency in utilizing wireless channel resources. Specifically, we design our scheduler based on K&H scheduling, but shift the burden of QoS provisioning to the resource allocation mechanism, thus simplifying the design of the scheduler. Such

a partitioning would be meaningless if the resource allocation problem now becomes complicated. However, we are able to solve the resource allocation problem efficiently using the recently developed method of *effective capacity* [22]. Effective capacity captures the effect of channel fading on the queueing behavior of the link, using a computationally simple yet accurate model, and thus, is the critical device we need to design an efficient resource allocation mechanism.

Our results show that compared to RR scheduling, our approach can substantially increase the statistical delay-constrained capacity (defined later) of a fading channel, when delay requirements are not very tight. For example, in the case of low signal-to-noise-ratio (SNR) and ergodic Rayleigh fading, our scheme can achieve approximately $\sum_{k=1}^{K} \frac{1}{k}$ gain for K users with loose-delay requirements, as expected from [10]. But more importantly, when the delay bound is not loose, so that simple-minded K&H scheduling does not directly apply, our scheme can achieve a capacity gain, and yet meet the QoS requirements.

The remainder of this paper is organized as follows. In Section 2, we discuss multiuser diversity and the recently introduced concept of effective capacity. Multiuser diversity, using K&H scheduling, is our key technique to increase capacity, while effective capacity is our critical device for QoS provisioning over a K&H scheduled wireless channel. Section 3 presents efficient QoS provisioning mechanisms and shows how to use multiuser diversity to achieve a performance gain while yet satisfying QoS constraints. In Section 4, we present the simulation results that demonstrate the performance gain of our scheme. Section 5 discusses the related work. In Section 6, we conclude the paper and point out future research directions.

2 Multiuser Diversity with QoS Constraints

In this section, we quantitatively discuss the performance gain obtained by multiuser diversity and describe the technique of effective capacity.

2.1 Multiuser Diversity

We first describe the model. Fig. 1 shows the architecture for scheduling multiuser traffic over a fading (time-varying) time-slotted wireless channel. A cellular wireless network is assumed, and the downlink is considered, where a base station transmits data to K mobile user terminals, each of which requires certain QoS guarantees. The channel fading processes of the users are assumed to be stationary, ergodic and independent of each other. A single cell is considered, and interference from other cells is modelled as background noise. In the base station, packets destined to different users are put into separate queues. We assume a block fading channel model [4], which assumes that user channel gains are constant over a time duration of length T_s (T_s is assumed to be small enough that the channel gains are constant, yet large enough that ideal channel codes can achieve capacity over that duration). Therefore, we partition time into 'frames' (indexed as $t = 0, 1, 2, \ldots$),

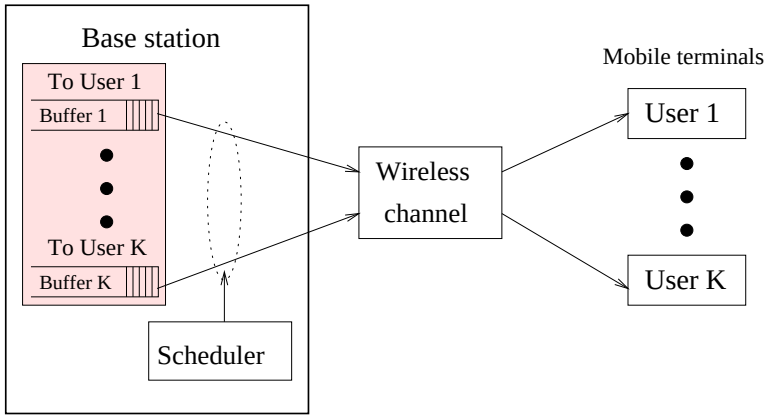


Figure 1: Downlink scheduling architecture for multiple users sharing a wireless channel.

each of length T_s. Thus, each user k has a time-varying channel power gain[1] $g_k(t), k = 1, \ldots, K$, which varies with the frame index t. The base station is assumed to know the current and past values of $g_k(t)$. The capacity of the channel for the k^{th} user, $c_k(t)$, is

$$c_k(t) \quad = \quad \log_2(1 + g_k(t)) \qquad \text{bits/symbol} \tag{1}$$

We divide each frame of length T_s into infinitesimal time slots, and assume that the channel can be shared by several users, in the same frame. Further, we assume a *fluid model* for packet transmission, where the base station can allot *variable fractions* of a channel frame to a user, over time. The system described above could be, for example, an idealized time-division multiple access (TDMA) system, where the frame of each channel consists of TDMA time slots which are infinitesimal. Note that in a practical TDMA system, there would be a finite number of finite-length time slots in each frame.

To provide QoS guarantees, we propose an architecture, which consists of scheduling, admission control, and resource allocation (presented in Section 3). Since the channel fading processes of the users are assumed to be independent of each other, we can potentially utilize multiuser diversity to increase capacity, as mentioned in Section 1. Thus, *to maximize the ergodic capacity* (*i.e.*, in the absence of delay constraints), the (optimal) K&H schedule at any time instant t, is to transmit the data of the user with the largest gain $g_k(t)$ [10]. The ergodic channel capacity achieved by such a K&H scheduler is $c_{max} = \mathbf{E}[\max[c_1(t), c_2(t), \cdots, c_K(t)]]$. The ergodic channel capacity gain of the K&H scheduler over a RR scheduler is $c_{max}/\mathbf{E}[c_1(t)]$. The following proposition specifies the ergodic channel capacity gain achieved by the K&H scheduler.

Proposition 1 *Assume that the K users in the system have i.i.d. channel gains, which are stationary processes in time t. For Rayleigh fading channels (*i.e.*, having exponentially-distributed channel power gains), at low SNR, we have the approximation, $c_{max}/\mathbf{E}[c_1(t)] \approx \sum_{k=1}^{K} \frac{1}{k} \approx \log(K+1)$ for large K.*

[1] $g_k(t) = |h_k(t)|^2 P_0/\sigma^2$, where the maximum transmission power P_0 and noise variance σ^2 are assumed to be constant and equal for all users. $h_k(t)$ is the voltage gain of the channel for the k^{th} user.

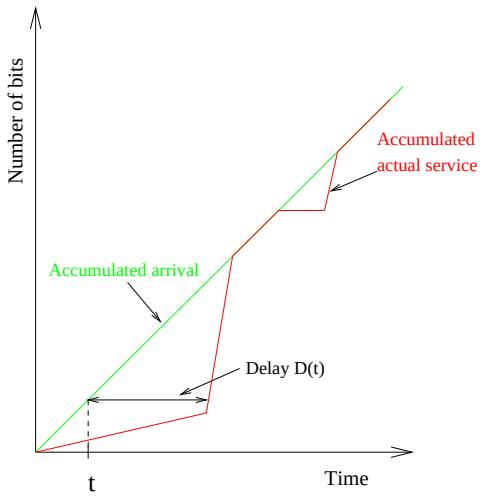


Figure 2: Delay $D(t)$ due to mismatch between the arrival and departure.

For a proof of Proposition 1, see Appendix. At high SNR, the ergodic channel capacity gain is smaller.

Notice that K&H scheduling can result in a user experiencing an arbitrarily long duration of outage, because of its failure to obtain the channel. Thus, it becomes important to efficiently compute the QoS obtained by the user, in a K&H scheduled system. A direct approach may be to model each $g_k(t)$ as a Markov process, and analyze the Markov process resulting from the K&H scheduler. It is apparent that this direct approach is computationally intractable, since the large state space of the joint Markov process of all the users would need to be analyzed and the complexity of this queueing analysis is exponential in the number of users. In essence, the main contribution of this paper is to show that we can compute the QoS obtained by the user, in a K&H scheduled system, efficiently and accurately, using the concept of effective capacity.

2.2 Effective Capacity

We first formally define statistical QoS, which characterizes the user requirement. First, consider a single-user system, where the user is allotted a single time varying channel (thus, there is no scheduling involved). Assume that the user source has a fixed rate r_s and a specified delay bound D_{max}, and requires that the delay-bound violation probability is not greater than a certain value ε, that is,

$$\sup_t Pr\{D(t) \geq D_{max}\} \leq \varepsilon, \tag{2}$$

where $D(t)$ is the delay experienced by a source packet arriving at time t (see Fig. 2), and $Pr\{D(t) \geq D_{max}\}$ is the probability of $D(t)$ exceeding a delay bound D_{max}. Then, we say that the user is specified by the (statistical) QoS triplet $\{r_s, D_{max}, \varepsilon\}$. Even for this simple case, it is not immediately obvious as to which QoS triplets are feasible, for the given channel, since a rather complex queueing system (with an arbitrary channel capacity process) will need to be analyzed.

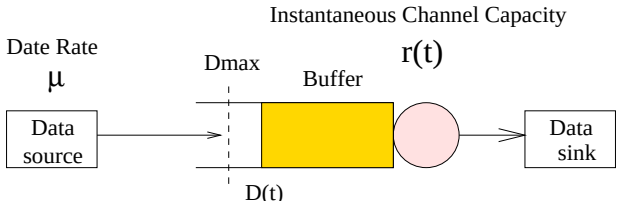


Figure 3: A queueing system model.

The key contribution of [22] was to introduce a concept of statistical delay-constrained capacity termed *effective capacity*, which allows us to obtain a simple and efficient test, to check the feasibility of QoS triplets for a single time-varying channel. That paper did not deal with scheduling and the channel processes resulting from it.

In this paper, we show that the effective capacity concept can be applied to the K&H scheduled channel, and is precisely the critical device that we need to solve the QoS constrained multiuser diversity problem. Therefore, we briefly explain the concept of effective capacity, and refer the reader to [22] for details.

Let $r(t)$ be the instantaneous channel capacity at time t. The *effective capacity function* of $r(t)$ is defined as [22]

$$\alpha(u) = -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}], \ \forall\ u \geq 0. \tag{3}$$

In this paper, since t is a discrete frame index, the integral above should be thought of as a summation.

Consider a queue of infinite buffer size supplied by a data source of *constant* data rate μ (see Fig. 3). It can be shown [22] that if $\alpha(u)$ indeed exists (*e.g.*, for ergodic, stationary, Markovian $r(t)$), then the probability of $D(t)$ exceeding a delay bound D_{max} satisfies

$$\sup_t Pr\{D(t) \geq D_{max}\} \approx e^{-\theta(\mu)D_{max}}, \tag{4}$$

where the function $\theta(\mu)$ of source rate μ depends only on the channel capacity process $r(t)$. $\theta(\mu)$ can be considered as a "channel model" that models the channel at the link layer (in contrast to "radio layer" models specified by Markov processes, or Doppler spectra). The approximation (4) is accurate for large D_{max}.

In terms of the effective capacity function (3) defined earlier, the *QoS exponent function* $\theta(\mu)$ can be written as [22]

$$\theta(\mu) = \mu\alpha^{-1}(\mu) \tag{5}$$

where $\alpha^{-1}(\cdot)$ is the inverse function of $\alpha(u)$. Once $\theta(\mu)$ has been measured for a given channel, it can be used to check the feasibility of QoS triplets. Specifically, a QoS triplet $\{r_s, D_{max}, \varepsilon\}$ is feasible if $\theta(r_s) \geq \rho$, where $\rho \doteq -\log\varepsilon/D_{max}$. Thus, we can use the effective capacity model $\alpha(u)$ (or equivalently, the function $\theta(\mu)$ via (5)) to relate the channel capacity process $r(t)$ to statistical QoS.

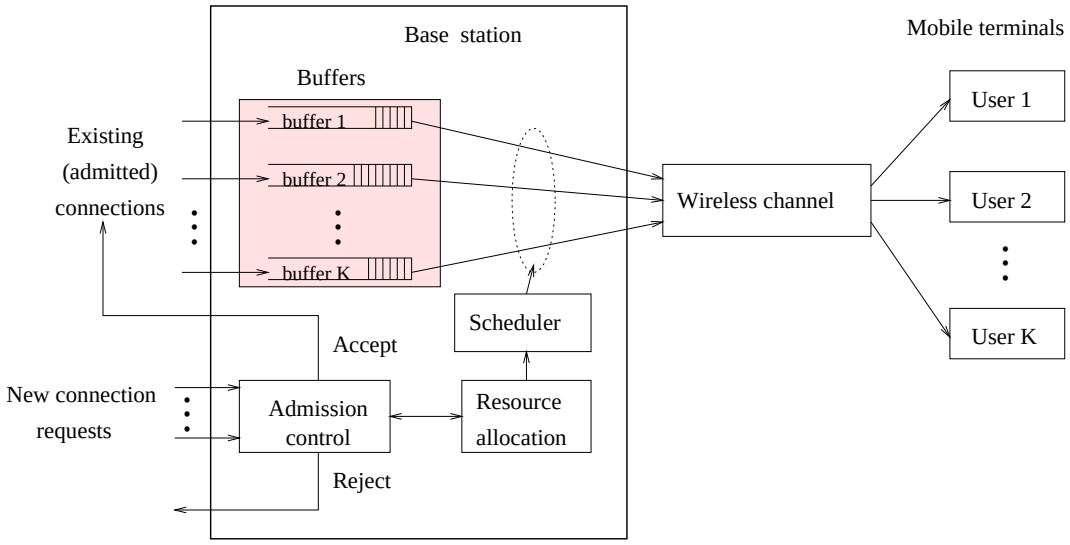


Figure 4: QoS provisioning architecture in a base station.

Since our effective capacity method predicts an exponential dependence (4) between $\{D_{max}, \varepsilon\}$, we can henceforth consider the QoS *pair* $\{r_s, \rho\}$ to be equivalent to the QoS triplet $\{r_s, D_{max}, \varepsilon\}$, with the understanding that $\rho = -\log \varepsilon / D_{max}$.

In Appendix, we present a simple and efficient algorithm to estimate $\theta(\mu)$ by direct measurement on the queueing behavior resulting from $r(t)$. In Section 4.2.1, we show that the estimation algorithm converges quickly, as compared with directly measuring the QoS.

Now, having described our basic techniques, *i.e.*, multiuser diversity using K&H scheduling, and effective capacity, in the next section, we present a QoS architecture consisting of admission control, resource allocation and scheduling, which utilizes these techniques for efficient support of QoS.

3 QoS Provisioning with Multiuser Diversity

The key problem is, how to utilize multiuser diversity while yet satisfying the individual QoS constraints of the K users. To cope with this problem, we design a QoS provisioning architecture, which utilizes multiuser diversity and effective capacity.

We assume the same setting as in Section 2.1. Fig. 4 shows our QoS provisioning architecture in a base station, consisting of three components, namely, admission control, resource allocation, and scheduling. When a new connection request comes, we first use a resource allocation algorithm to compute how much resource is needed to support the requested QoS. Then the admission control module checks whether the required resource can be satisfied. If yes, the connection request is accepted; otherwise, the connection request is rejected. For admitted connections, packets that

belong to different connections[2] are put into separate queues. The scheduler decides, in each frame t, how to schedule packets for transmission, based on the *current* channel gains $g_k(t)$ and the amount of resource allocated.

In the following sections, we describe our schemes for scheduling, admission control and resource allocation in detail. In Section 3.1, we consider the homogeneous case, in which all users have the same QoS requirements $\{r_s, D_{max}, \varepsilon\}$ or equivalently the same QoS pair $\{r_s, \rho = -\log \varepsilon / D_{max}\}$ and also the same channel statistics (*e.g.*, similar Doppler rates), so that all users need to be assigned equal channel resources. Section 3.2 addresses the heterogeneous case, in which users have different QoS pairs $\{r_s, \rho\}$ and/or different channel statistics.

3.1 Homogeneous Case

3.1.1 Scheduling

As explained in Section 1, we simplify the scheduler, by shifting the burden of guaranteeing user QoS to the resource allocation module. Therefore, our scheduler is a simple combination of K&H and RR scheduling.

Section 2 explained that in any frame t, the K&H scheduler transmits the data of the user with the largest gain $g_k(t)$. However, the QoS of a user may be satisfied by using only a fraction of the frame $\beta \leq 1$. Therefore, it is the function of the resource allocation algorithm to allot the minimum required β to the user. This will be described in Section 3.1.2. It is clear that K&H scheduling attempts to utilize multiuser diversity to maximize the throughput.

On the other hand, the RR scheduler allots to every user k, a fraction $\zeta \leq 1/K$ of *each* frame, where ζ again needs to be determined by the resource allocation algorithm. Thus RR scheduling attempts to provide tight QoS guarantees, at the expense of decreased throughput, in contrast to K&H scheduling.

Our scheduler is a joint K&H/RR scheme, which attempts to maximize the throughput, while yet providing QoS guarantees. In each frame t, its operation is the following. First, find the user $k^*(t)$ such that it has the largest channel gain among all users. Then, schedule user $k^*(t)$ with $\beta + \zeta$ fraction of the frame; schedule each of the other users $k \neq k^*(t)$ with ζ fraction of the frame. Thus, a fraction β of the frame is used by K&H scheduling, while simultaneously, a total fraction $K\zeta$ of the frame is used by RR scheduling. The total usage of the frame is $\beta + K\zeta \leq 1$.

3.1.2 Admission Control and Resource Allocation

The scheduler described in Section 3.1.1 is simple, but it needs the frame fractions $\{\zeta, \beta\}$ to be computed and reserved. This function is performed at the admission control and resource allocation phase.

Since Section 3.1 addresses the homogeneous case with K users, without loss of generality,

[2]We assume that each mobile user is associated with only one connection.

denote $\alpha_{K,\zeta,\beta}(u)$ the effective capacity function of user $k = 1$ under the joint K&H/RR scheduling (henceforth called 'joint scheduling'), with frame shares ζ and β respectively, *i.e.*, denote the capacity process allotted to user 1 by the joint scheduler as the process $r(t)$ and then compute $\alpha_{K,\zeta,\beta}(u)$ using (3). The corresponding QoS exponent function $\theta_{K,\zeta,\beta}(\mu)$ can be found via (5). Note that $\theta_{K,\zeta,\beta}(\mu)$ is a function of number of users K. Then, the admission control and resource allocation scheme for users requiring the QoS pair $\{r_s, \rho\}$ is as below,

$$\underset{\{\zeta,\beta\}}{\text{minimize}} \quad K\zeta + \beta \tag{6}$$

$$\text{subject to} \quad \theta_{K,\zeta,\beta}(r_s) \geq \rho, \tag{7}$$

$$K\zeta + \beta \leq 1, \tag{8}$$

$$\zeta \geq 0, \qquad \beta \geq 0 \tag{9}$$

The minimization in (6) is to minimize the total frame fraction used. (7) ensures that the QoS pair $\{r_s, \rho\}$ of each user is feasible. Furthermore, Eqs. (7)–(9) also serve as an admission control test, to check availability of resources to serve this set of users. Since we have the following relation for $\lambda > 0$ (see Appendix for a proof)

$$\theta_{K,\zeta,\beta}(\mu) = \theta_{K,\lambda\zeta,\lambda\beta}(\lambda\mu), \tag{10}$$

we only need to measure the $\theta_{K,\zeta,\beta}(\cdot)$ functions for different ratios of ζ/β.

To summarize, given the fading channel and QoS of K homogeneous users, we use the following procedure to achieve multiuser diversity gain with QoS provisioning:

1. Estimate $\theta_{K,\zeta,\beta}(\mu)$, directly from the queueing behavior, for various values of $\{\zeta, \beta\}$.
2. Determine the optimal $\{\zeta, \beta\}$ pair that satisfies users' QoS, while minimizing frame usage.
3. If admission control is passed, provide the joint scheduler with the optimal ζ and β, for simultaneous RR and K&H scheduling, respectively.

This summary indicates that our approach needs to address the following issues. Our paper [22] showed the usefulness of the effective capacity concept, only for a single-user system. But, it is not obvious that the $\theta_{K,\zeta,\beta}(\mu)$ estimate will converge quickly in the multiuser scenario, or even that $\theta_{K,\zeta,\beta}(\mu)$ can accurately predict QoS via (4) (although, theoretically, the prediction is accurate asymptotically for large D_{max}). Further, it needs to be seen whether the QoS can be controlled by $\{\zeta, \beta\}$. Last, we also need to show that our scheme can provide a substantial capacity gain, over RR scheduling. These issues will be addressed via simulations in Section 4.

3.1.3 Improvement

The aforementioned admission control and resource allocation, *i.e.*, Eqs. (6)–(9), may not be efficient in terms of resource usage, when K is large and ρ takes a medium value. The reason is the

following. On one hand, as K increases, the maximum data rate $\mu_{K,\zeta,\beta}(\theta = \rho)$ achievable with a specified QoS exponent ρ, also increases due to $K\&H$ scheduling; but this increase of $\mu_{K,\zeta,\beta}(\theta = \rho)$ is only achievable when ρ is small, *i.e.*, the delay requirement is loose. Here, the function $\mu_{K,\zeta,\beta}(.)$ is an inverse function of $\theta_{K,\zeta,\beta}(\mu)$. On the other hand, as K increases, the probability that a user will be allowed to transmit is $1/K$ since there are K identical users share the channel; hence, the probability that a user has a large queue length, increases as K increases; so, to achieve the same QoS exponent ρ, the data rate $\mu_{K,\zeta,\beta}(\theta = \rho)$ has to be decreased as K increases. For a medium-valued ρ, if K is small, the effect of capacity gain is dominant since the probability that a user is allowed to transmit, *i.e.*, $1/K$, is large; if K is large, then the effect of less probability to transmit is dominant, resulting in the reduction of data rate $\mu_{K,\zeta,\beta}(\theta = \rho)$. Hence, it may be better to partition the users into groups and schedule the groups independently.

Before describing the partitioning scheme, we need to introduce some notations. Suppose that the K users are partitioned into M groups (obviously, $1 \leq M \leq K$). Each group m ($m = 1, 2, \cdots, M$) has K_m users with channel characterization (*i.e.*, QoS exponent function) $\theta_{K_m,\zeta_m,\beta_m}(\mu)$, where $\{\zeta_m, \beta_m\}$ are the frame shares assigned to group m, for the joint K&H/RR scheduling. Obviously, $\sum_{m=1}^{M} K_m = K$.

Next, we describe scheduling and resource allocation/admission control, respectively.

Scheduling

For *each group* m, we use the joint K&H/RR scheduler with frame shares $\{\zeta_m, \beta_m\}$. In each frame t, it works as follows. First, find the user $k_m^*(t)$ such that it has the largest channel gain among K_m users of group m (not among K users). Then, schedule user $k_m^*(t)$ with $\beta_m + \zeta_m$ fraction of the frame; schedule each of other group-m users $k \neq k_m^*(t)$ with ζ_m fraction of the frame. Thus, the total usage of the frame by all M group is $\sum_{m=1}^{M}(K_m\zeta_m + \beta_m) \leq 1$.

Admission Control and Resource Allocation

The scheduler described above requires the frame fractions $\{\zeta_m, \beta_m\}$ to be computed and reserved. This function is performed at the admission control and resource allocation phase. With dividing users into M groups, the admission control and resource allocation scheme for K users requiring the QoS pair $\{r_s, \rho\}$ is as below,

$$\underset{\{M,K_m,\zeta_m,\beta_m\}}{\text{minimize}} \quad \sum_{m=1}^{M}(K_m\zeta_m + \beta_m) \tag{11}$$

$$\text{subject to} \quad \theta_{K_m,\zeta_m,\beta_m}(r_s) \geq \rho, \qquad \forall m \tag{12}$$

$$\sum_{m=1}^{M}(K_m\zeta_m + \beta_m) \leq 1, \tag{13}$$

$$\zeta_m \geq 0, \qquad \beta_m \geq 0, \qquad \forall m, \tag{14}$$

$$M \in \{1, 2, \cdots, K\}, \qquad \sum_{m=1}^{M} K_m = K \tag{15}$$

Eq. (11) is to minimize the total frame fraction used. (12) ensures that the QoS pair $\{r_s, \rho\}$ of each user is feasible. Furthermore, Eqs. (12)–(15) also serve as an admission control test, to check availability of resources to serve the K users. If $M = 1$, Eqs. (11)–(15) reduced to Eqs. (6)–(9). Therefore, the optimal solution in Eqs. (6)–(9) is a feasible solution of Eqs. (11)–(15). As a result, the resource allocation/admission control in Eqs. (11)–(15) is at least as efficient as that in Eqs. (6)–(9), if not more efficient.

Note that the improvement on efficiency achieved in Eqs. (11)–(15) is at the cost of complexity.

3.2 Heterogeneous Case

For the heterogeneous case, in which users have different QoS pairs $\{r_s, \rho\}$ and/or different channel statistics, the admission control/resource allocation problem can also be formulated, similar to Eqs. (11)–(15), as minimizing the resource usage over M, partitioning of the K users,ζ_m and β_m. But solving this minimization problem has an exponential complexity, *i.e.*, $O(K^K)$, since we have to try all the possible combinations. To reduce the complexity, we design a sub-optimal algorithm, which has a complexity of $O(K \log K)$.

We first consider the case where the K users have different channel statistics but the same $\{r_s, \rho\}$. Figure 5 shows the flow chart of our algorithm for the resource allocation. The basic operations of this algorithm are sorting and partitioning. Sorting the users is to facilitate partitioning the users. Partitioning is achieved through tests, which recursively check whether adding a user to a K&H scheduled group can reduce the channel usage.

Next we describe the algorithm. According to Figure 5, we first measure the function $\mu_k(\theta = \rho)$ for each user k, where $\mu_k(\theta)$ is the inverse function of $\theta(\mu)$ defined in (5); note that there is no scheduling involved in (5). Then we sort the users in descending order of $\mu_k(\theta = \rho)$, which results in an ordered list denoted by L_{user}. We set a variable m to count the number of groups, each of which uses the K&H scheduling with a fraction β_m. Denote $\mathcal{S}(m)$ the set that contains the users in group m.

We partition the K users recursively, starting from group $m = 1$. Each time when we form a new group m, we first remove the head element of list L_{user} and put it into an empty set $\mathcal{S}(m)$; if L_{user} is not empty, we again remove the head ν of list L_{user} and put ν into $\mathcal{S}(m)$; now we have two users and can apply the K&H scheduling to the two users; if the resulting channel usage is greater than or equal to that due to applying the RR scheduling to the two users, we move the user ν from $\mathcal{S}(m)$ back to the head of L_{user}, *i.e.*, the new user ν should not be added to the group m and group m will only have one user; otherwise, we recursively continue this procedure, that is, adding a new user and testing whether the resulting channel usage is reduced, *i.e.*, $\beta_m(\mathcal{S}(m)) < \beta_m(\mathcal{S}'(m)) + r_s/\mu_\nu(\theta = \rho)$, where $\beta_m(\mathcal{S}(m))$ is the channel usage of set $\mathcal{S}(m)$ and set $\mathcal{S}'(m)$ is the difference $\mathcal{S}(m) - \{\nu\}$, and $r_s/\mu_\nu(\theta = \rho)$ is the channel usage for user ν if the RR scheduling is used. We continue the partitioning until L_{user} is empty.

In the process of partitioning, we determine ζ_m and β_m for scheduling as below. If group m has only one user, say user k, we only use the RR scheduling with $\zeta_m = r_s/\mu_k(\theta = \rho)$, and set

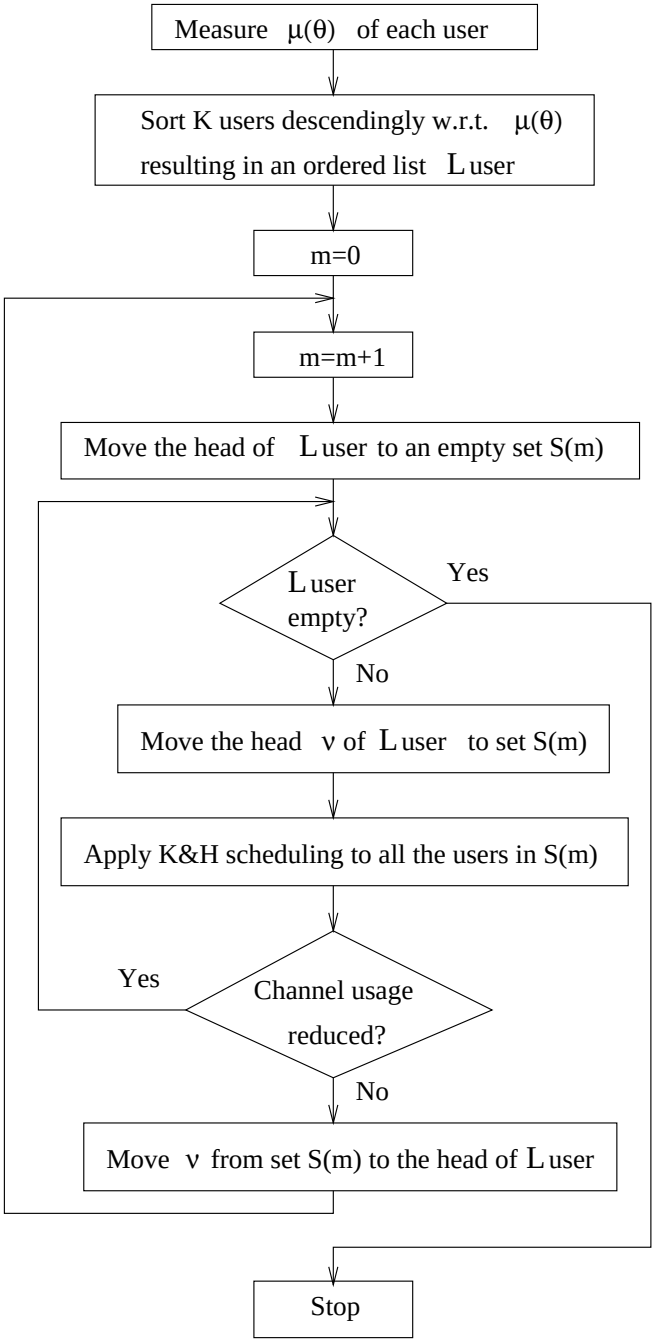


Figure 5: Flow chart of resource allocation for the heterogeneous case.

$\beta_m = 0$; if group m has more than one user, we only use the K&H scheduling with the β_m obtained in the test for $\mathcal{S}(m)$, and set $\zeta_m = 0$. The joint K&H/RR scheduling is not used, in order to reduce complexity.

The outputs of the algorithm are 1) a partition of the K users, say, M groups, and 2) $\{\zeta_m, \beta_m\}$ $(m = 1, \cdots, M)$. After running the resource allocation algorithm, we do admission control by testing whether the total channel usage of the M groups is not greater than unity.

If the admission control is passed, we schedule each of the M groups as follows. If group m has only one user, we use the RR scheduling with ζ_m; if group m has more than one user, we apply the K&H scheduling with β_m to all the users in group m.

Our algorithm reduces complexity at the cost of optimality. Specifically, the resource allocation algorithm achieves $O(K \log K)$ complexity due to sorting. Note that we only have to try at most K tests in partitioning the K users. The performance of the algorithm may not be optimal but our simulations show that for practical range of Doppler rates, our algorithm improves the performance by a factor of two, compared to the RR scheduling.

For the case where the K users have different $\{r_s, \rho\}$ and different channel statistics, we could classify the users into N classes so that the users in the same class have the same QoS pair $\{r_s, \rho\}$. Then apply the resource allocation algorithm to each class. The admission control is simply to check whether the total channel usage of the N classes is not greater than unity.

To summarize, given the fading channel and QoS of K users, we use the following procedure to achieve multiuser diversity gain with QoS provisioning:

1. Use the resource allocation algorithm to partition the K users and determine $\{\zeta_m, \beta_m\}$ that satisfies users' QoS.

2. If admission control is passed, provide the scheduler with ζ_m and β_m, for the RR or K&H scheduling.

4 Simulation Results

4.1 Simulation Setting

We simulate the system depicted in Fig. 4, in which each connection[3] is simulated as plotted in Fig. 6. In Fig. 6, the data source of user k generates packets at a *constant* rate $r_s^{(k)}$. Generated packets are first sent to the (infinite) buffer at the transmitter, whose queue length in frame t is $Q_k(t)$. The head-of-line packet in the queue is transmitted over the fading channel at data rate $r_k(t)$. The fading channel has a random power gain $g_k(t)$ (the noise variance is absorbed into $g_k(t)$). We use a fluid model, that is, the size of a packet is infinitesimal. In practical systems, the results presented here will have to be modified to account for finite packet sizes.

[3]Assume that K connections are set up and each mobile user is associated with a connection.

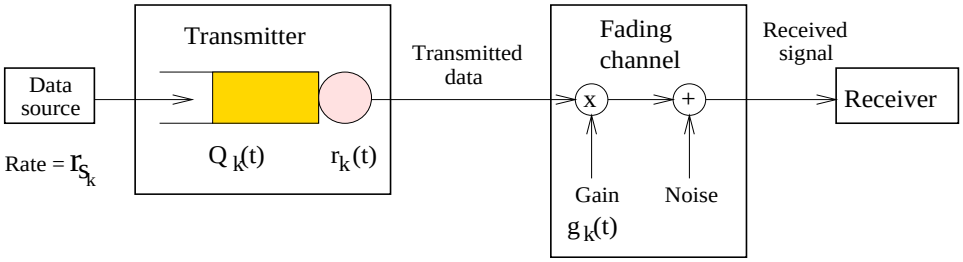


Figure 6: The queueing model used for simulations.

We assume that the transmitter has perfect knowledge of the current channel gains $g_k(t)$ in frame t. Therefore, it can use rate-adaptive transmissions, and ideal channel codes, to transmit packets without decoding errors. For the homogeneous case, under joint scheduling, the transmission rate $r_k(t)$ of user k is equal to a fraction of its instantaneous capacity, as below,

$$r_k(t) = \begin{cases} (\zeta + \beta)c_k(t) & \text{if } k = \arg\max_{i \in \{1,\cdots,K\}} g_i(t); \\ \\ \zeta c_k(t) & \text{otherwise.} \end{cases} \tag{16}$$

where the instantaneous channel capacity $c_k(t)$ is

$$c_k(t) = B_c \log_2(1 + g_k(t)) \tag{17}$$

where B_c is the channel bandwidth. For the heterogeneous case, the rate $r_k(t)$ is computed within each class as described in Section 3.2.

The average SNR is fixed in each simulation run. We define r_{awgn} as the capacity of an equivalent AWGN channel, which has the same average SNR. *i.e.*,

$$r_{awgn} = B_c \log_2(1 + SNR_{avg}) \tag{18}$$

where $SNR_{avg} = E[g_k(t)]$. Then, we can eliminate B_c using Eqs. (17) and (18) as,

$$c_k(t) = \frac{r_{awgn} \log_2(1 + g_k(t))}{\log_2(1 + SNR_{avg})}. \tag{19}$$

The sample interval (frame length) T_s is set to 1 milli-second and each simulation run is 100-second long in all scenarios except in Section 4.2.1. Rayleigh flat-fading voltage-gains $h_k(t)$ are generated by a first-order auto-regressive (AR(1)) model as below:

$$h_k(t) = \kappa \times h_k(t-1) + v_k(t), \tag{20}$$

where $v_k(t)$ are zero-mean i.i.d. complex Gaussian variables. The coefficient κ determines the Doppler rate, *i.e.*, the larger the κ, the smaller the Doppler rate. Specifically, the coefficient κ can be determined by the following procedure: 1) compute the coherence time T_c by [19, page 165]

$$T_c \approx \frac{9}{16\pi f_m}, \tag{21}$$

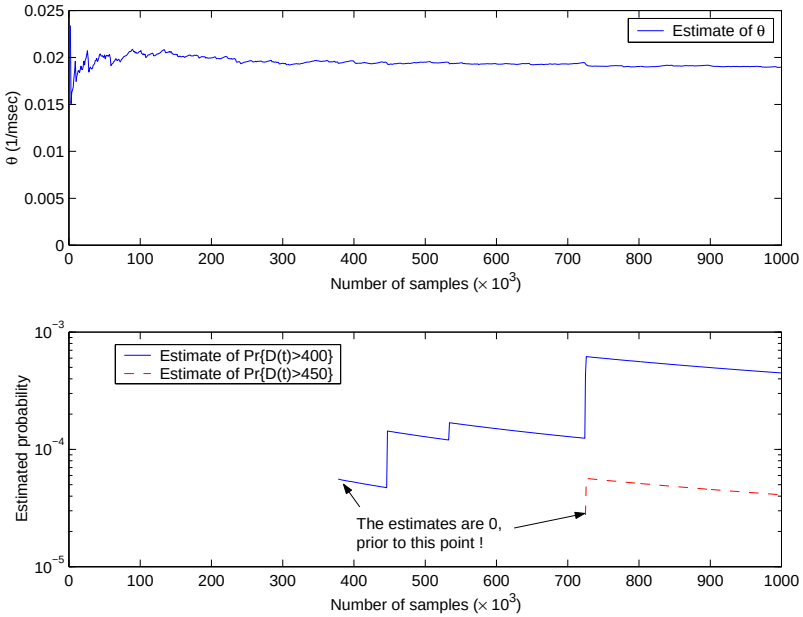


Figure 7: Convergence of estimates.

where the coherence time is defined as the time, over which the time auto-correlation function of the fading process is above 0.5; 2) compute the coefficient κ by[4]

$$\kappa = 0.5^{T_s/T_c}. \tag{22}$$

For Ricean fading, the voltage-gains $h_k(t)$ are generated by adding a deterministic signal component to Rayleigh-fading voltage-gains (see [17] for detail).

4.2 Performance Evaluation

We organize this section as follows. Section 4.2.1 shows the convergence of our estimation algorithm. In Section 4.2.2, we assess the accuracy of our QoS estimation (4). Section 4.2.3 investigates the effectiveness of the resource allocation scheme in QoS provisioning. In these sections (4.2.1 to 4.2.3), we only consider the homogeneous case, *i.e.*, all users have the same QoS requirements $\{r_s, \rho\}$ and also the same channel statistics. In Sections 4.2.4 through 4.2.6, we evaluate the performance of our scheduler in the homogeneous and the heterogeneous case, respectively.

4.2.1 Convergence of Estimates

This experiment is to show the convergence behavior of estimates. We do simulations with the following parameters fixed: $r_{awgn} = 1000$ kb/s, $K = 10$, $\kappa = 0.8$, and $SNR_{avg} = -40$ dB.

[4]The auto-correlation function of the AR(1) process is κ^n, where n is the number of sample intervals. Solving $\kappa^{T_c/T_s} = 0.5$ for κ, we obtain (22).

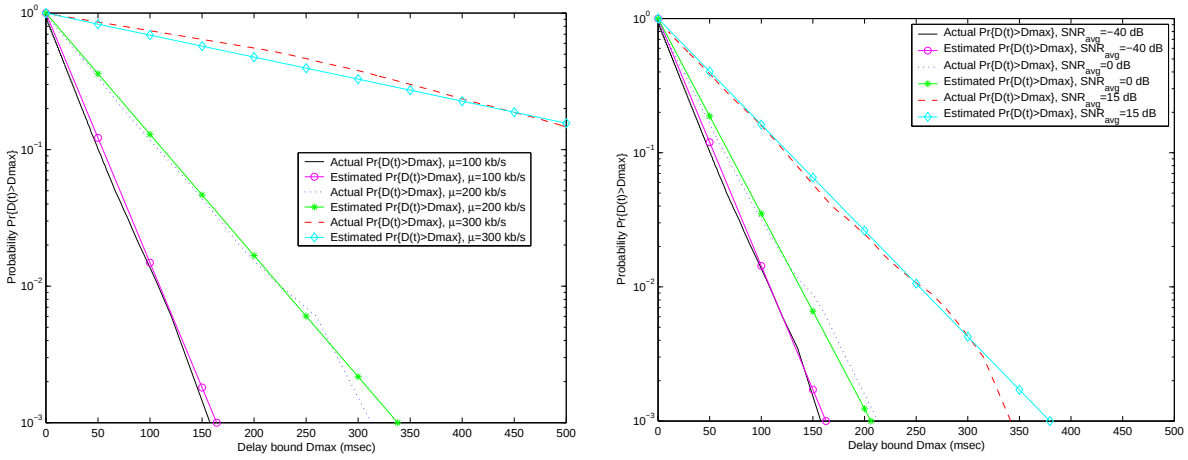


Figure 8: Actual and estimated delay-bound violation probability for (a) different source rates, and (b) different SNR_{avg}.

Fig. 7 shows the convergence of the estimate of θ ($\theta(\mu)$ for $\mu = 200$ kb/s) for the queue. It can be seen that the estimate of θ converges within 2×10^4 samples/frames (20 sec). The same figure shows the (lack of) convergence of direct (Monte Carlo) estimates of delay-bound-violation probabilities, measured for the same queue (the two probability estimates eventually converge to 10^{-3} and 10^{-4}, respectively). This precludes using the direct probability estimate to predict the user QoS, as alluded to in Section 1. The reason for the slow convergence of the direct probability estimate is that the K&H scheduling results in a user being allotted the channel in a bursty manner, and thus increases the correlation time of $D(t)$ substantially. Therefore, even 10^6 samples are not enough to obtain an accurate estimate of a probability as high as 10^{-3}.

4.2.2 Accuracy of Channel Estimation

The experiments in this section are to show that the estimated effective capacity can indeed be used to accurately predict QoS.

We do experiments under five different settings: 1) AR(1) Rayleigh fading channel with changing source rate and fixed SNR_{avg}, 2) AR(1) Rayleigh fading channel with changing SNR_{avg} and fixed source rate, 3) AR(2) Rayleigh fading channel, 4) Ricean fading channel, and 5) Nakagami-m fading channel (chi-distribution) [21, page 22].

Under the first setting, we do simulations with the following parameters fixed: $r_{awgn} = 1000$ kb/s, $K = 10$, $\kappa = 0.8$, and $SNR_{avg} = -40$ dB. By changing the source rate μ, we simulate three cases, *i.e.*, μ =100, 200, and 300 kb/s. Fig. 8(a) shows the actual delay-bound violation probability $\sup_t Pr\{D(t) > D_{max}\}$ vs. the delay bound D_{max}. From the figure, it can be observed that the actual delay-bound violation probability decreases exponentially with D_{max}, for all the cases. This

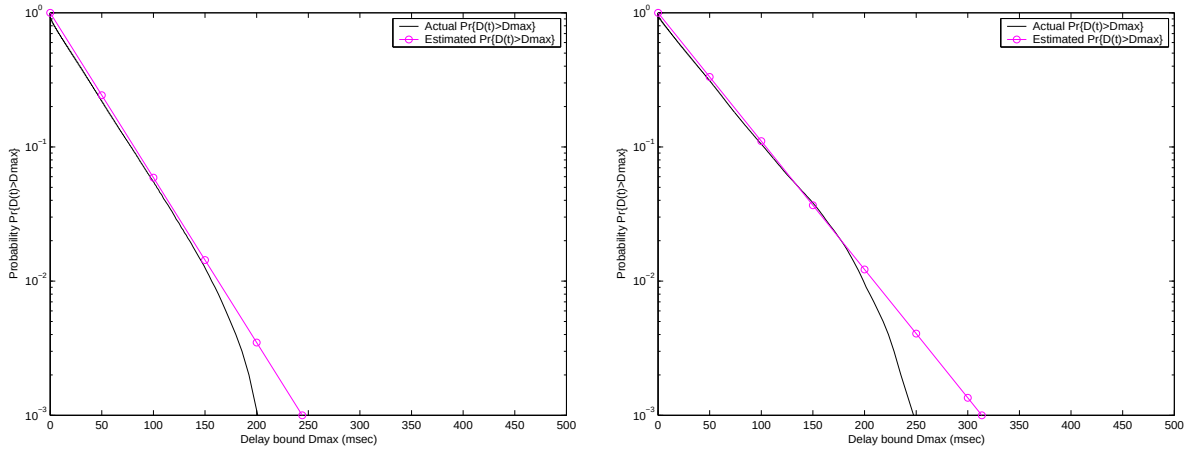


Figure 9: Actual and estimated delay-bound violation probability for (a) AR(2) channel, and (b) Ricean channel.

confirms the exponential dependence shown in (4).

In addition, we use the estimation scheme, *i.e.*, Eqs. (36) through (40), to obtain an estimated θ; with the resulting θ, we predict the probability $\sup_t Pr\{D(t) > D_{max}\}$ (using (4)). As shown in Fig. 8(a), the estimated $\sup_t Pr\{D(t) > D_{max}\}$ is quite close to the actual $\sup_t Pr\{D(t) > D_{max}\}$. This demonstrates that our estimation is accurate, which justifies the use of (7) by the resource allocation algorithm to guarantee QoS.

Notice that the (negative) slope of the $\sup_t Pr\{D(t) > D_{max}\}$ plot increases with the decrease of the source rate μ. This is because the smaller the source rate, the smaller the probability of delay-bound violation, resulting in a sharper slope (*i.e.*, a larger decaying rate θ).

Under the second setting, we do simulations with the following parameters fixed: $r_{awgn} = 1000$ kb/s, $K = 10$, $\kappa = 0.8$, and $\mu = 100$ kb/s. By changing SNR_{avg}, we simulate three cases, *i.e.*, SNR_{avg} = -40, 0, and 15 dB. Fig. 8(b) shows that the conclusions drawn from the first set of experiments still hold. Thus, our estimation scheme gives consistent performance over different SNRs also.

In the third setting, AR(2) Rayleigh fading voltage-gains $h_k(t)$ are generated as below:

$$h_k(t) = \kappa_1 \times h_k(t-1) + \kappa_2 \times h_k(t-2) + v_k(t), \tag{23}$$

where $v_k(t)$ are zero-mean i.i.d. complex Gaussian variables. The parameters of the simulation are $r_{awgn} = 1000$ kb/s, $\kappa_1 = 0.7$, $\kappa_2 = 0.2$, $K = 10$, μ =100 kb/s and $SNR_{avg} = -40$ dB. Fig. 9(a) shows that the conclusions drawn from the first set of experiments still hold. Thus, our estimation scheme consistently predicts the QoS metric under different autoregressive channel fading models.

Under the fourth setting, the parameters of the simulation are Ricean factor[5]= 7 dB, $r_{awgn} =$

[5]Ricean factor is defined as the ratio between the deterministic signal power A^2 and the variance of the multipath

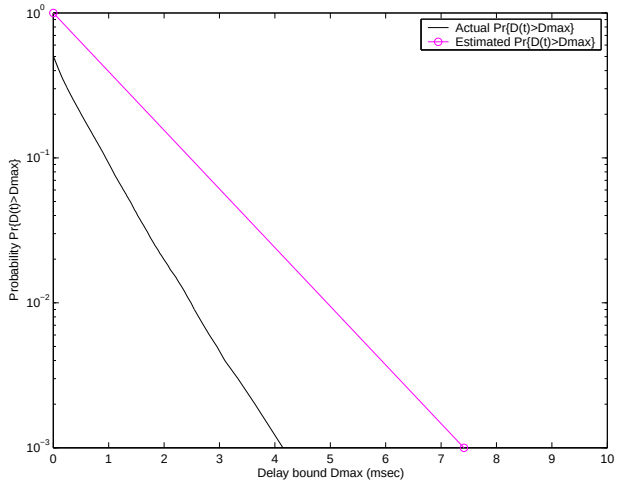


Figure 10: Actual and estimated delay-bound violation probability for Nakagami-m channel ($m = 32$).

1000 kb/s, $K = 10$, $\kappa = 0.8$, and $\mu = 100$ kb/s. Fig. 9(b) shows that the conclusions drawn from the first set of experiments also hold for Ricean fading channels.

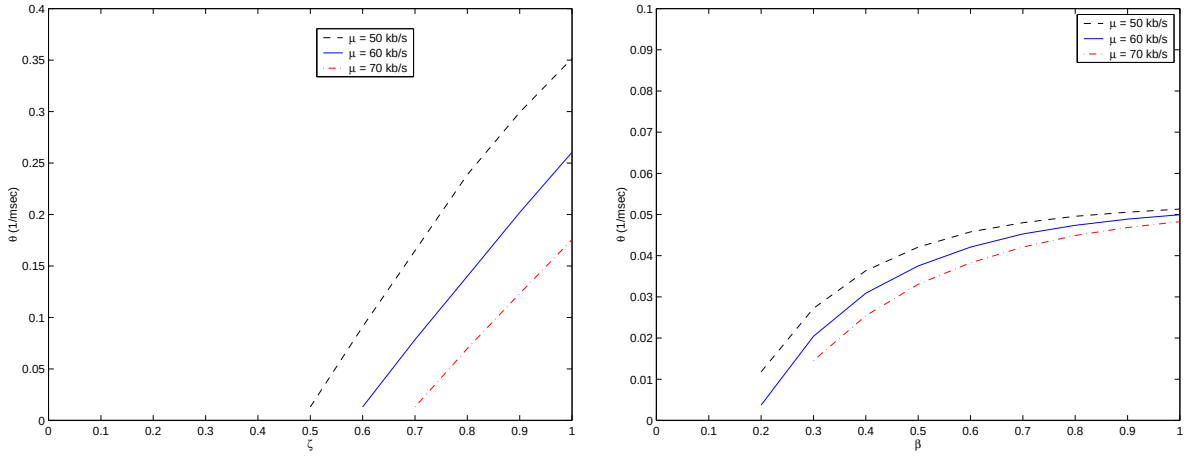
In the fifth setting, Nakagami-m fading power gains $g_k(t)$ are generated as below:

$$g_k(t) = \sum_{i=1}^{m} \hat{g}_i(t), \tag{24}$$

where m is the parameter of Nakagami-m distribution and takes values of positive integers, $\hat{g}_i(t)$ are AR(1) Rayleigh fading power gains. The parameters of the simulation are $m = 32$, $r_{awgn} = 1000$ kb/s, $K = 10$, $\kappa = 0.8$, $SNR_{avg} = -40$ dB, and μ =90 kb/s. Fig. 10 shows that the estimate does not give a good agreement with the actual $\sup_t Pr\{D(t) > D_{max}\}$. The reason is that the high diversity in high-order Nakagami fading models averages out the randomness in the fading process. The higher diversity a fading channel has, the more like an AWGN channel the fading channel is. It is known that for an AWGN channel, the actual $\sup_t Pr\{D(t) > D_{max}\}$ does not decay exponentially with D_{max} but takes values of 0 or 1. Therefore, the higher diversity a fading channel possesses, the less accurate the exponential approximation (4) is, hence the less accurate the estimate is.

In summary, the results for Rayleigh/Ricean flat-fading channels have shown the exponential behavior of the actual $\sup_t Pr\{D(t) > D_{max}\}$ and the accurateness of our estimation. We caution however that such a strong agreement between the estimate and the actual QoS may not occur in all situations with practical values of D_{max} (although the theory predicts the agreement asymptotically for large D_{max}). We have shown that in the case of high-diversity channel fading models (*e.g.*, high-order Nakagami fading models), the estimation is not accurate.

$2\sigma_m^2$, *i.e.*, Ricean factor $= A^2/(2\sigma_m^2)$.



(a) (b)

Figure 11: (a) θ vs. ζ and (b) θ vs. β.

4.2.3 Effectiveness of Resource Allocation in QoS Provisioning

The experiments here are to show that a QoS pair $\{r_s, \rho\}$ can be achieved (within limits) by choosing ζ or β appropriately. In the experiments, we fix the following parameters: $K = 10$, $\kappa = 0.8$, and $SNR_{avg} = -40$ dB. We simulate three data rates, *i.e.*, μ =50, 60, and 70 kb/s, respectively. We do two sets of experiments: one for the RR scheduling and the other for the K&H scheduling.

In the first set of experiments, only the RR scheduling is used; we change ζ from 0.1 to 1 and estimate the resulting θ for a given μ, using Eqs. (36) through (40). Fig. 11(a) shows that θ increases with ζ. Thus, Fig. 11 can be used to allot ζ to a user to satisfy its QoS requirements when using RR scheduling.

In the second set of experiments, only the K&H scheduling is used; we change β from 0.1 to 1 and estimate the resulting θ, for a given μ. Fig. 11(b) shows that θ increases with the increase of β, and thus the figure can be used to allot β to a user to satisfy its QoS requirements when using K&H scheduling.

4.2.4 Performance Gains of Scheduling: Homogeneous Case

Under the setting of identical QoS requirements $\{r_s, \rho\}$ and i.i.d. channel gain processes, the experiments here demonstrate the performance gain of joint scheduling over RR scheduling, using the optimum $\{\zeta, \beta\}$ values specified by the resource allocation algorithm, *i.e.*, Eqs. (6)–(9). In particular, the experiments show that for loose delay constraints, the large capacity gains promised by the K&H scheme can indeed be approached.

To evaluate the performance of the scheduling schemes under different SNRs and different Doppler rates (*i.e.*, different κ), we simulate three cases: 1) $\kappa = 0.8$, and $SNR_{avg} = -40$ dB, 2)

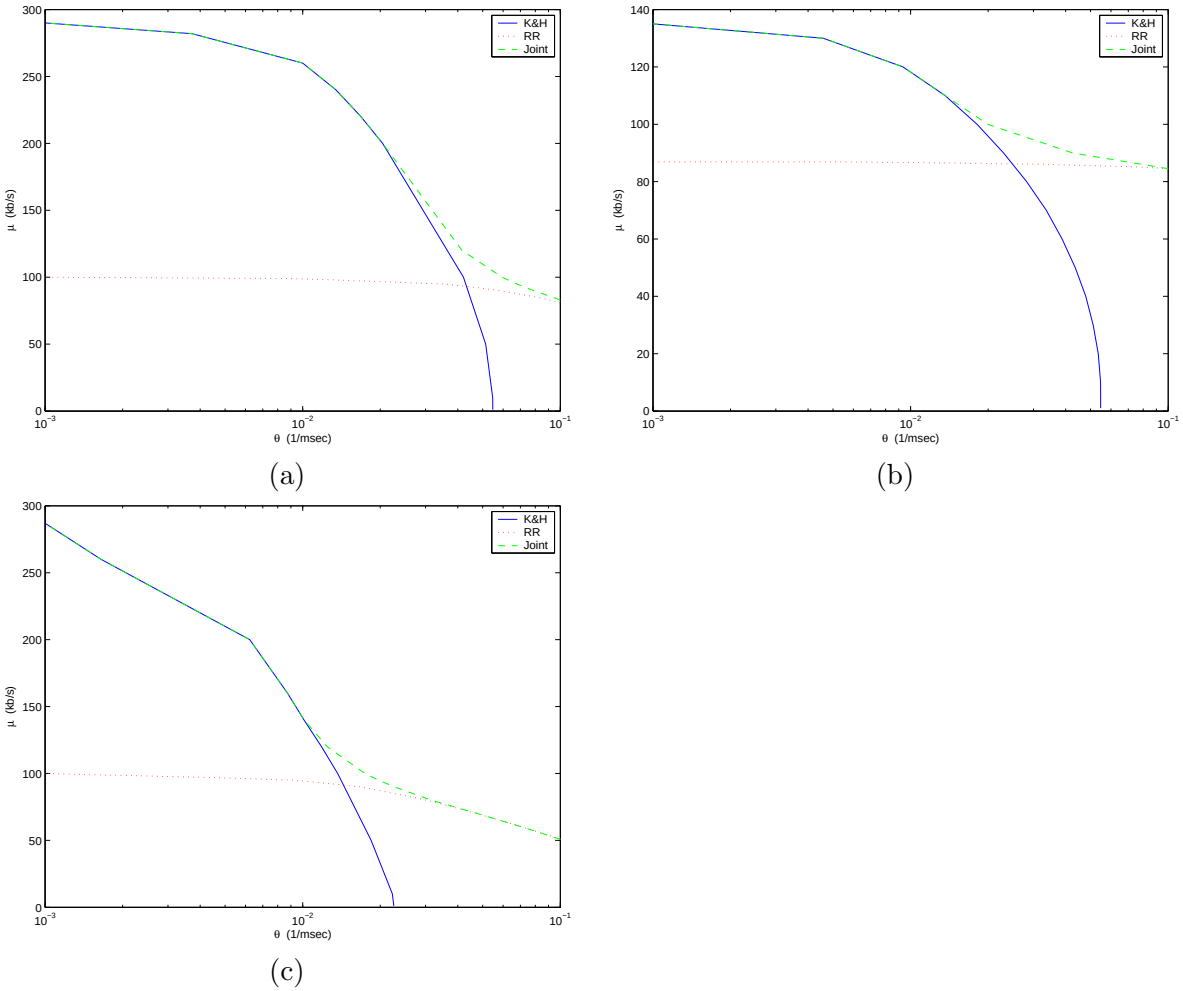


Figure 12: $\theta(\mu)$ vs. μ for (a) $\kappa = 0.8$, $SNR_{avg} = -40$ dB, (b) $\kappa = 0.8$, $SNR_{avg} = 15$ dB, and (c) $\kappa = 0.95$, $SNR_{avg} = -40$ dB.

$\kappa = 0.8$, and $SNR_{avg} = 15$ dB, and 3) $\kappa = 0.95$, and $SNR_{avg} = -40$ dB. In all the experiments, we set $r_{awgn} = 1000$ kb/s and $K = 10$.

In Fig. 12, we plot the function $\theta(\mu)$ achieved by the joint, K&H, and RR schedulers, for a range of source rate μ, when the entire frame is used (*i.e.*, $K\zeta + \beta = 1$). The function $\theta(\mu)$ in the figure is obtained by the estimation scheme, *i.e.*, Eqs. (36) through (40). In the case of the joint scheduling, each point in the figure corresponds to a specific optimum $\{\zeta, \beta\}$, while for the RR and the K&H scheduling, we set $K\zeta = 1$ and $\beta = 1$ respectively. The curve of $\theta(\mu)$ can be directly used to check for feasibility of a QoS pair $\{r_s, \rho\}$, by checking whether $\theta(r_s) > \rho$ is satisfied. Furthermore, for a given θ, the ratio of $\mu(\theta)$ of the joint scheduler to the $\mu(\theta)$ of the RR scheduler (both obtained from the figure), represents the delay-constrained **capacity gain** that can be achieved by using joint scheduling.

Four important observations can be made from the figure. First, the range of θ can be divided into three segments: small, medium, and large θ, which correspond to three categories of the QoS constraints: loose-delay, medium-delay, and tight-delay requirements. For small θ, our joint scheduler achieves substantial gain, *e.g.*, approximately $\sum_{k=1}^{K} \frac{1}{k}$ capacity gain for Rayleigh fading channels at low SNR. For example, in Fig. 12(a), when $\theta = 0.001$, the capacity gain for the joint scheduler is 2.9, which is close to $\sum_{k=1}^{10} \frac{1}{k} = 2.929$. For medium θ, our joint scheduler also achieves gain. For example, in Fig. 12(a), when $\theta = 0.01$, the capacity gain for the joint scheduler is 2.6. For large θ, such as $\theta = 0.1$, our joint scheduler does not give any gain. Thus, the curve of $\theta(\mu)$ shows the range of θ (delay constraints) for which a K&H type scheme can provide a performance gain. When the scheduler is provided with the optimum $\{\zeta, \beta\}$ values, the QoS pair $\{\mu, \theta(\mu)\}$ guaranteed to a user is indeed satisfied; the simulation result that shows this fact, is similar to Fig. 8, and therefore, is not shown.

Second, we observe that the joint scheduler has a larger effective capacity than both the K&H and the RR for a rather small range of θ. Therefore, in practice, it may be sufficient to use either K&H or RR scheduling, depending on whether θ is small or large respectively, and dispense with the more complicated joint scheduling. However, we have designed more sophisticated joint schedulers, such as splitting the channel between the best two users in every slot, which perform substantially better than either K&H and RR scheduling, for medium values of θ. This is shown in Fig. 13. This indicates that more sophisticated joint schedulers may squeeze out more channel capacity gain. We leave this for future study.

Third, the figure of $\theta(\mu)$ can be used to satisfy the QoS constraint (7), even though it only represent the $K\zeta + \beta = 1$ case, as follows. For the QoS pair $\{r_s, \rho\}$, we compute the ratio $\lambda \doteq \frac{r_s}{\mu(\theta=\rho)}$ using the $\mu(\theta)$ function in the figure. Suppose the $\mu(\theta = \rho)$ point in the figure corresponds to the optimum pair $\{\bar{\zeta}, \bar{\beta}\}$. Since we have the relation $\theta_{\bar{\zeta},\bar{\beta}}(\mu) = \theta_{\lambda\bar{\zeta},\lambda\bar{\beta}}(\lambda\mu)$, *i.e.*, Eq. (10), we assert that instead of using the entire frame (as in the figure), if we use a total fraction λ of the frame, then we can achieve the desired QoS $\{r_s, \rho\}$. The joint scheduler then needs to use the $\{\lambda\bar{\zeta}, \lambda\bar{\beta}\}$ pair to do RR and K&H scheduling respectively. This indicates a compelling advantage of our QoS provisioning scheme over direct-measurement based schemes, which require experiments for different λ, even if the ratio ζ/β is fixed.

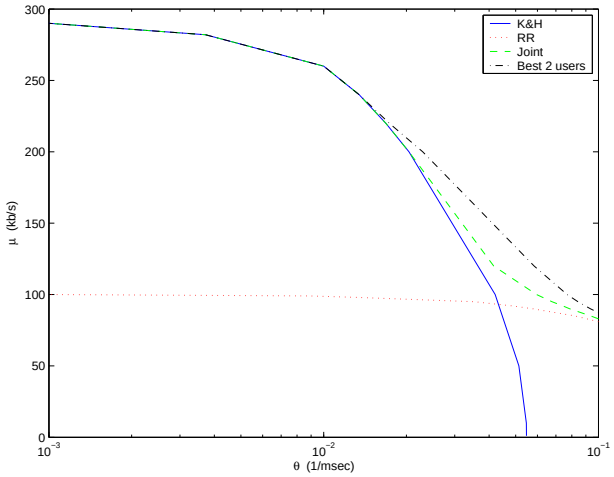


Figure 13: $\theta(\mu)$ vs. μ for splitting the channel between the best two users.

Fourth, we observe that if θ is larger than a certain value, the corresponding data rate $\mu(\theta)$ achieved by the K&H, approaches zero. This is because the probability that a user will be allowed to transmit is $1/K$ since there are K identical users share the channel; hence, on average, a user has to wait $K-1$ frames before its next packet is scheduled for transmission, making a guarantee of tight-delay requirements (large θ) impossible.

4.2.5 Performance Improvement due to Partitioning

This experiment demonstrates the performance improvement due to partitioning of users as mentioned in Section 3.1.3. In particular, the experiment indicates the trade-off between performance improvement and complexity.

Again, the setting of this experiment is i.i.d. channel gain processes and identical QoS requirements $\{r_s, \rho\}$. We divide K users into groups and see how much gain can be achieved, compared with non-partitioning. The parameters of the experiment is the following: $r_{awgn} = 1000$ kb/s, $\kappa = 0.8$, $SNR_{avg} = -40$ dB, $r_s = 100$ kb/s, $\rho = 0.03$.

The simulation result shows that the channel can support at most 13 users if the users is not partitioned. If users are allowed to be partitioned, solving Eqs. (11)–(15), we find the maximum number of users that the channel can support is 16, where the 16 users are partitioned into two groups, each of which consists of 8 users. So partitioning increases the maximum admissible number of users by 3. Note that this performance improvement is at the cost of complexity of solving Eqs. (11)–(15).

Table 1: Parameters for AR(2) Rayleigh fading channels.

	User 1	User 2	User 3	User 4	User 5	User 6
κ_1	0.8	0.75	0.7	0.6	0.5	0.45
κ_2	0.1	0.05	0.2	0.2	0.3	0.4

4.2.6 Performance Gains of Scheduling: Heterogeneous Case

The experiments here show the performance gain of the K&H/RR scheduling over the RR scheduling, using the $\{\zeta_m, \beta_m\}$ values specified by the resource allocation algorithm in Figure 5, under the setting of identical QoS requirements $\{r_s, \rho\}$ and non-identical, independent channel gain processes.

We do two sets of experiments. In the experiments, we fix the following parameters: $r_{awgn} =$ 1000 kb/s, $SNR_{avg} = -40$ dB, $r_s = 30$ kb/s, $\rho = 0.01$.

The first set of experiments is done under three different settings: 1) $K = 10$, 2) $K = 19$, 3) $K = 37$. We use AR(1) Rayleigh fading channel and each user has different κ, *i.e.*, different Doppler rate. We let κ change from 0.6 to 0.99 with equal spacing for the three settings, that is, user 1 has $\kappa = 0.6$, the last user (10th, 19th, or 37th user) has $\kappa = 0.99$, and other users' κ are determined by equal spacing between 0.6 and 0.99.

In the first setting ($K = 10$), the resource allocation algorithm in Figure 5 results in a partition with two groups, where the number of users in each group are $K_1 = 8$ and $K_2 = 2$. The total channel usage under the K&H scheduling is 0.195, while the total channel usage under the RR scheduling is 0.319.

In the second setting ($K = 19$), the resource allocation algorithm leads to a partition with three groups, where $K_1 = 14$, $K_2 = 4$, and $K_3 = 1$. The total channel usage under the K&H/RR scheduling is 0.344, while the total channel usage under the RR scheduling is 0.597.

In the third setting ($K = 37$), the resource allocation algorithm obtains a partition with five groups, where $K_1 = 27$, $K_2 = 6$, $K_3 = 2$, $K_4 = 1$, and $K_5 = 1$. The total channel usage under the K&H/RR scheduling is 0.69, while the total channel usage under the RR scheduling is 1.15, which is rejected by admission control.

The second set of experiments is done under the following setting. We have $K = 10$ users in the system. Among the ten users, four users have AR(1) Rayleigh fading channels and each of them has different κ. We let κ change from 0.6 to 0.99 with equal spacing, that is, user 1 has $\kappa = 0.6$, the fourth user has $\kappa = 0.99$, and other users' κ are determined by equal spacing between 0.6 and 0.99. The other six users have AR(2) Rayleigh fading channels, specified by (23). Table 1 lists the parameters for the AR(2) Rayleigh fading channels of the six users. From the simulation, the resource allocation algorithm obtains a partition with two groups, where $K_1 = 9$ and $K_2 = 1$. The total channel usage under the K&H/RR scheduling is 0.1842, while the total channel usage under the RR scheduling is 0.3215.

Hence, our resource allocation algorithm in Figure 5 achieves smaller channel usage than that using the RR scheduling; as a result, the system can admit more users.

In summary, the K&H/RR scheduler achieves performance gain when delay requirements are not very tight, while yet guaranteeing QoS at any delay requirement.

5 Related Work

There have been many proposals on QoS provisioning in wireless networks. Since our work is centered on scheduling, we will focus on the literature on scheduling with QoS constraints in wireless environments. Besides K&H scheduling and Bettesh & Shamai's scheduler that we discussed in Section 1, previous works on this topic also include wireless fair queueing [12, 14, 18], modified largest weighted delay first (M-LWDF) [1], opportunistic transmission scheduling [11] and lazy packet scheduling [16].

Wireless fair queueing schemes [12, 14, 18] are aimed at applying Fair Queueing [15] to wireless networks. The objective of these schemes is to provide fairness, while providing loose QoS guarantees. However, the problem formulation there does not allow explicit QoS guarantees (*e.g.*, explicit delay bound or rate guarantee), unlike our approach. Further, their problem formulation stresses fairness, rather than efficiency, and hence, does not utilize multiuser diversity to improve capacity.

The M-LWDF algorithm [1] and the opportunistic transmission scheduling [11] implicitly utilize multiuser diversity, so that higher efficiency can be achieved. However, the schemes do not provide explicit QoS, but rather optimize a certain QoS parameter.

The lazy packet scheduling [16] is targeted at minimizing energy, subject to a delay constraint. The scheme only considers AWGN channels and thus allows for a deterministic delay bound, unlike fading channels and the general statistical QoS considered in our work.

Static fixed channel assignments, primarily in the wireline context, have been considered [9], in a multiuser, multichannel environment. However, these do not consider channel fading, or general QoS guarantees.

Time-division scheduling has been proposed for 3-G WCDMA [8, page 226]. The proposed time-division scheduling is similar to the RR scheduling in this paper. However, their proposal did not provide methods on how to use time-division scheduling to support statistical QoS guarantees explicitly. With the notion of effective capacity, we are able to make explicit QoS provisioning with our joint scheduling.

6 Concluding Remarks

In this paper, we examined the problem of QoS provisioning for K users sharing a single time-slotted fading downlink channel. We developed simple and efficient schemes for admission control, resource allocation, and scheduling, to obtain a gain in delay-constrained capacity. Multiuser diversity

obtained by the well-known K&H scheduling is the key that gives rise to this performance gain. However, the unique feature of this paper is explicit support of the statistical QoS requirement $\{r_s, D_{max}, \varepsilon\}$, for channels utilizing K&H scheduling. The concept of effective capacity is the key that explicitly guarantees the QoS. Thus, the paper combines crucial ideas from the areas of communication theory and queueing theory to provide the tools to increase capacity and yet satisfy QoS constraints. The statistical QoS requirement is satisfied by the channel assignments $\{\zeta, \beta\}$, which are determined by the resource allocation module at the admission phase. Then, the joint scheduler uses the channel assignments $\{\zeta, \beta\}$ in scheduling data at the transmission phase, with guaranteed QoS. Simulation results have shown that our approach can substantially increase the delay-constrained capacity of a fading channel, compared with the RR scheduling, when delay requirements are not very tight.

Our future work will focus on the design of scheduling, admission control and resource allocation, for multiple users sharing multiple channels. In addition, work is underway in applying our approach to practical settings, such as non-negligible packet size.

Appendix

Proof of Proposition 1

In this proof, the time index t for channel gains and capacities is dropped, due to the assumption of stationarity of the channel gains. Without loss of generality, the expectation of channel power gain g_k can be normalized to one, *i.e.*, $\mathbf{E}[g_k] = 1, \forall k$. Then, for Rayleigh fading channels, the CDF of each channel power gain g_k is

$$F_G(g) \quad = \quad 1 - e^{-g}. \tag{25}$$

For low SNR, *i.e.*, very small g_k, the channel capacity of the k^{th} user can be approximated by

$$c_k \quad = \quad \log_2(1 + g_k) \approx g_k. \tag{26}$$

Then we have

$$\begin{aligned}
c_{max}/\mathbf{E}[c_1] &= \mathbf{E}[\max[c_1, c_2, \cdots, c_K]]/\mathbf{E}[c_1] && (27)\\
&\approx \mathbf{E}[\max[g_1, g_2, \cdots, g_K]]/\mathbf{E}[g_1] && (28)\\
&\stackrel{(a)}{=} \int_0^\infty g\mathrm{d}(F_G^K(g)) && (29)\\
&= \int_0^\infty (1 - F_G^K(g))\mathrm{d}g && (30)\\
&= \int_0^\infty (1 - F_G(g))\frac{(1 - F_G^K(g))}{(1 - F_G(g))}\mathrm{d}g && (31)\\
&= \int_0^\infty e^{-g}(F_G^{K-1}(g) + F_G^{K-2}(g) + \cdots + 1)\mathrm{d}g && (32)\\
&= \int_0^\infty (F_G^{K-1}(g) + F_G^{K-2}(g) + \cdots + 1)\mathrm{d}F_G(g) && (33)\\
&= 1 + \frac{1}{2} + \cdots + \frac{1}{K} && (34)
\end{aligned}$$

where (a) follows from the fact that the CDF of the random variable $\max[g_1, g_2, \cdots, g_K]$ is $F_G^K(g)$ [20, page 248], and $\mathbf{E}[g_1] = 1$. It can be easily proved that

$$\log(K+1) \leq \sum_{k=1}^{K} \frac{1}{k} \leq 1 + \log K. \tag{35}$$

So for large K, we can approximate $\sum_{k=1}^{K} \frac{1}{k}$ by $\log(K+1)$. This completes the proof. ■

Estimation of QoS Exponent $\theta(\mu)$

We briefly describe a simple algorithm to estimate the function $\theta(\mu)$ (see [22] for details of derivation). Assume that the time-varying channel capacity process $r(t)$ is stationary and ergodic. For a given (unknown) fading channel and a given source rate μ, we take measurements from the queue (see Fig. 3). First, take a number of samples, say N, over an interval of length T, and record the following quantities at the nth sampling epoch: S_n the indicator of whether a packets is in service[6] ($S_n \in \{0, 1\}$), Q_n the number of bits in the queue (excluding the packet in service), and T_n the remaining service time of the packet in service (if there is one in service). Then, compute the following sample means,

$$\hat{\gamma} = \frac{1}{N}\sum_{n=1}^{N} S_n, \tag{36}$$

$$\hat{q} = \frac{1}{N}\sum_{n=1}^{N} Q_n, \tag{37}$$

[6] A packet in service refers to a packet in the process of being transmitted.

and

$$\hat{\tau_s} = \frac{1}{N}\sum_{n=1}^{N} T_n. \tag{38}$$

Finally, we obtain the estimate of $\theta(\mu)$ by

$$\hat{\theta} = \frac{\hat{\gamma} \times \mu}{\mu \times \hat{\tau_s} + \hat{q}} \tag{39}$$

Eqs. (36) through (39) constitute our algorithm for estimating the QoS exponent function $\theta(\mu)$. Note that, to get the function $\theta(\mu)$, we need to estimate θ for different source rate μ.

The estimated $\hat{\theta}$ can be used to predict the QoS by approximating Eq. (4) with

$$\sup_t Pr\{D(t) \geq D_{max}\} \approx e^{-\hat{\theta} D_{max}}. \tag{40}$$

Proof for Eq. (10)

We use the notation in Section 3.1.2. From (3), we have

$$\alpha_{K,\zeta,\beta}(u) = \frac{-\lim_{t\to\infty} \frac{1}{t} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}]}{u}, \ \forall\ u \geq 0. \tag{41}$$

Since we only consider the homogeneous case, without loss of generality, denote $\alpha_{K,\lambda\zeta,\lambda\beta}(u)$ the effective capacity function of user $k = 1$ under K&H/RR scheduling, with frame shares $\lambda\zeta$ and $\lambda\beta$ respectively ($\lambda > 0$). Denote the resulting capacity process allotted to user 1 by the joint scheduler as the process $\hat{r}(t)$. Then for $u \geq 0$, we have

$$\begin{aligned}
\alpha_{K,\lambda\zeta,\lambda\beta}(u) &\overset{(a)}{=} \frac{-\lim_{t\to\infty} \frac{1}{t} \log E[e^{-u\int_0^t \hat{r}(\tau)\mathrm{d}\tau}]}{u} \\
&\overset{(b)}{=} \frac{-\lim_{t\to\infty} \frac{1}{t} \log E[e^{-u\int_0^t \lambda r(\tau)\mathrm{d}\tau}]}{u} \\
&= \frac{-\lim_{t\to\infty} \frac{\lambda}{t} \log E[e^{-(\lambda u)\int_0^t r(\tau)\mathrm{d}\tau}]}{\lambda u} \\
&\overset{(c)}{=} \lambda \times \alpha_{K,\zeta,\beta}(\lambda u)
\end{aligned} \tag{42}$$

where (a) from (3), (b) using $\hat{r}(t) = \lambda r(t)$, which is obtained via (16), and (c) from (41). Then by $\alpha_{K,\lambda\zeta,\lambda\beta}(u) = \lambda \times \alpha_{K,\zeta,\beta}(\lambda u) \doteq \mu$, we have

$$u = \alpha^{-1}_{K,\lambda\zeta,\lambda\beta}(\lambda\mu) \tag{43}$$

and

$$\lambda u = \alpha^{-1}_{K,\zeta,\beta}(\mu). \tag{44}$$

Removing u in (43) and (44) results in

$$\alpha^{-1}_{K,\zeta,\beta}(\mu) = \lambda \times \alpha^{-1}_{K,\lambda\zeta,\lambda\beta}(\lambda\mu) \tag{45}$$

Thus, we have

$$\begin{aligned} \theta_{K,\zeta,\beta}(\mu) &\overset{(a)}{=} \mu \times \alpha^{-1}_{K,\zeta,\beta}(\mu) \\ &\overset{(b)}{=} \mu \times \lambda \times \alpha^{-1}_{K,\lambda\zeta,\lambda\beta}(\lambda\mu) \\ &\overset{(c)}{=} \theta_{K,\lambda\zeta,\lambda\beta}(\lambda\mu) \end{aligned} \tag{46}$$

where (a) from (5), (b) from (45), and (c) from (5). This completes the proof. ■

References

[1] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine,* vol. 39, no. 2, pp. 150–154, Feb. 2001.

[2] I. Bettesh and S. Shamai, "A low delay algorithm for the multiple access channel with Rayleigh fading," in *Proc. IEEE Personal, Indoor and Mobile Radio Communications (PIMRC'98),* 1998.

[3] I. Bettesh and S. Shamai, "Optimal power and rate control for fading channels," in *Proc. IEEE Vehicular Technology Conference,* Spring 2001.

[4] E. Biglieri, J. Proakis, and S. Shamai, "Fading channel: information theoretic and communication aspects," *IEEE Trans. Information Theory,* vol. 44, pp. 2619–2692, Oct. 1998.

[5] L. Georgiadis, R. Guerin, V. Peris, and R. Rajan, "Efficient support of delay and rate guarantees in an Internet," in *Proc. ACM SIGCOMM'96,* Aug. 1996.

[6] M. Grossglauser and D. Tse, "Mobility increases the capacity of wireless adhoc networks," in *Proc. IEEE INFOCOM'01,* April 2001.

[7] S. Hanly and D. Tse, "Multi-access fading channels: part II: delay-limited capacities," *IEEE Trans. on Information Theory,* vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[8] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications,* Wiley, 2000.

[9] L. M. C. Hoo, "Multiuser transmit optimization for multicarrier modulation system," *Ph. D. Dissertation,* Department of Electrical Engineering, Stanford University, CA, USA, Dec. 2000.

[10] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE International Conference on Communications (ICC'95),* Seattle, USA, June 1995.

[11] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications,* vol. 19, no. 10, pp. 2053–2064, Oct. 2001.

[12] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. on Networking,* vol. 7, no. 4, pp. 473–489, Aug. 1999.

[13] B. L. Mark and G. Ramamurthy, "Real-time estimation and dynamic renegotiation of UPC parameters for arbitrary traffic sources in ATM networks," *IEEE/ACM Trans. on Networking,* vol. 6, no. 6, pp. 811–827, Dec. 1998.

[14] T. S. E. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," in *Proc. IEEE INFOCOM'98,* pp. 1103–1111, San Francisco, CA, USA, March 1998.

[15] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single node case," *IEEE/ACM Trans. on Networking,* vol. 1, no. 3, pp. 344–357, June 1993.

[16] B. Prabhakar, E. Uysal-Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. IEEE INFOCOM'01,* April 2001.

[17] R. J. Punnoose, P. V. Nikitin, and D. D. Stancil, "Efficient simulation of Ricean fading within a packet simulator," in *Proc. IEEE Vehicular Technology Conference (VTC'2000),* Boston, MA, USA, Sept. 2000.

[18] P. Ramanathan and P. Agrawal, "Adapting packet fair queueing algorithms to wireless networks," in *Proc. ACM MOBICOM'98,* Oct. 1998.

[19] T. S. Rappaport, *Wireless Communications: Principles & Practice,* Prentice Hall, 1996.

[20] G. G. Roussas, "A course in mathematical statistics," 2nd ed., Academic Press, 1997.

[21] M. K. Simon and M.-S. Alouini, "Digital communication over fading channels," Wiley, 2000.

[22] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," to appear in *IEEE Trans. on Wireless Communications,* Available at http://www.cs.cmu.edu/~dpwu/publications.html.