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

Dapeng Wu and 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.**

***Index Terms*— Multiuser diversity, QoS, effective capacity, fading, scheduling, resource allocation.**

I. INTRODUCTION

Providing quality of service (QoS), such as delay and rate guarantees, is an important objective in the design of future packet cellular networks [5]. However, this requirement poses a challenge in wireless network design, because wireless channels have low reliability, and time varying signal attenuation (fading), 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 [6] 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* [4], 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 [6] have shown that the 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 [1] proposed an algorithm, which strikes a balance between throughput and delay constraints. This algorithm combines the K&H scheduling with an RR scheduling, and it can achieve lower delay than the 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, achieved by the queueing behavior resulting from the algorithm, requires an excessively large number of samples, so that

Dapeng Wu is with University of Florida, Dept. of Electrical & Computer Engineering, P.O.Box 116130, Gainesville, FL 32611, USA. Email: wu@ece.ufl.edu.

Rohit Negi is with Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Email: negi@ece.cmu.edu.

it becomes practically infeasible.

Another typical approach is to use dynamic programming [2] 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 the 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* [7]. 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 the 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 [6]. 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 II, we discuss multiuser diversity and the recently introduced concept of effective capacity. Multiuser diversity, using the 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 III presents efficient QoS provisioning mechanisms and shows how to use multiuser diversity to achieve a performance gain while yet satisfying QoS constraints. In Section IV, we present the simulation results that demonstrate the performance gain of our scheme. Section V concludes the paper.

II. Multiuser Diversity with QoS Constraints

In this section, we describe multiuser diversity and the technique of effective capacity.

A. 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 with constant variance. In the base station, packets destined to different users are put into separate queues. We assume a block fading channel model [3], 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 $t = 0, 1, 2, \ldots$), each of length T_s. Thus, each user k has a time-varying channel power gain $g_k(t), k = 1, \ldots, K$, which varies with the frame index t; and $g_k(t) = |h_k(t)|^2$, where $h_k(t)$ is the voltage gain of the channel for the k^{th} user. 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

$$c_k(t) = \log_2(1 + g_k(t) \times P_0/\sigma_n^2) \quad \text{bits/symbol} \tag{1}$$

where the maximum transmission power P_0 and noise variance σ_n^2 are assumed to be constant and equal for all users. 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 III). 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 I. 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 $g_k(t)$ [6]. 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 $c_{max}/\mathbf{E}[c_1(t)]$. In particular, for Rayleigh fading channels, at low SNR, we have the approximation, $c_{max}/\mathbf{E}[c_1(t)] \approx \sum_{k=1}^{K} \frac{1}{k}$ for large K [8, page 121]. 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. 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.

B. 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,

$$Pr\{D(\infty) > D_{max}\} \leq \varepsilon, \tag{2}$$

where $D(\infty)$ is the steady-state delay experienced by a flow, and $Pr\{D(\infty) > D_{max}\}$ is the probability of $D(\infty)$ 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. The key contribution of [7] 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 [7] for details.

Let $r(t)$ be the instantaneous channel capacity at time t. The *effective capacity function* of $r(t)$ is defined as [7]

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

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. It can be shown [7] that if $\alpha(u)$ indeed exists (*e.g.*, for ergodic, stationary, Markovian $r(t)$), then the probability of $D(\infty)$ exceeding a delay bound D_{max} satisfies

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta(\mu)D_{max}}, \tag{4}$$

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 "physical 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 [7]

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

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. Since our effective capacity method

predicts an exponential dependence (4) between ε and D_{max}, 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 [8, page 81], we presented a simple and efficient algorithm to estimate $\theta(\mu)$ by direct measurement of the queueing behavior resulting from $r(t)$. In Section IV-B.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.

III. 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 II-A. Fig. 1 shows our QoS provisioning architecture in the 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 so, the connection request is accepted; otherwise, the connection request is rejected. For admitted connections, packets that belong to different connections[1] 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. We only 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. For the heterogeneous case, see [8, page 130].

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

A. Scheduling

As explained in Section I, 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 II 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 III-B. It is clear that the 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 the RR scheduling attempts to provide tight QoS guarantees, at the expense of decreased throughput, in contrast to the 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 the K&H scheduling, while simultaneously, a total fraction $K\zeta$ of the frame is used by the RR scheduling. The total usage of the frame is $\beta + K\zeta \leq 1$.

B. Admission Control and Resource Allocation

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

Since we only consider the homogeneous case, without loss of generality, denote $\alpha_{\zeta,\beta}(u)$ 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 $\alpha_{\zeta,\beta}(u)$ using (3). The corresponding QoS exponent function $\theta_{\zeta,\beta}(\mu)$ can be found via (5). Then, the admission control and resource allocation scheme for users requiring the QoS pair $\{r_s, \rho\}$ is as below,

$$\begin{aligned} \underset{\{\zeta,\beta\}}{\text{minimize}} \quad & K\zeta + \beta && (6)\\ \text{subject to} \quad & \theta_{\zeta,\beta}(r_s) \geq \rho, && (7)\\ & K\zeta + \beta \leq 1, && (8)\\ & \zeta \geq 0, \qquad \beta \geq 0 && (9) \end{aligned}$$

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 [8, pp. 270–271] for a proof)

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

we only need to measure the $\theta_{\zeta,\beta}(\cdot)$ 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 $\theta_{\zeta,\beta}(\mu)$, 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, by solving (6) to (9).
3. 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 [7] showed the usefulness of the effective capacity concept, only for a single-user system. But, it is not obvious that the $\alpha_{\zeta,\beta}(u)$ estimate will converge quickly in the multiuser scenario, or even that effective capacity 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 the RR scheduling. These issues are addressed via simulations.

IV. Simulation Results

A. Simulation Setting

We simulate the system depicted in Fig. 1. Under the joint scheduling, the transmission rate $r_k(t)$ of user k is equal to a fraction of its instantaneous capacity, as below,

$$r_k(t) = \begin{cases} (\zeta + \beta)c_k(t) & \text{if } k = \arg\max\limits_{i\in\{1,\cdots,K\}} g_i(t); \\ \zeta c_k(t) & \text{otherwise.} \end{cases} \tag{11}$$

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

$$c_k(t) = B_c \log_2(1 + g_k(t) \times P_0/\sigma_n^2) \tag{12}$$

where B_c denotes the channel bandwidth, and the transmission power P_0 and noise variance σ_n^2 are assumed to be constant. 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.*,

$$r_{awgn} = B_c \log_2(1 + SNR_{avg}) \tag{13}$$

where $SNR_{avg} = E[g_k(t) \times P_0/\sigma^2] = P_0/\sigma^2$. We set $E[g_k(t)] = 1$.

The sample interval (frame length) T_s is set to 1 millisecond. Most simulation runs are 1000-second long; some simulation runs are 10000-second long in order to obtain good estimate of the actual delay-violation probability $Pr\{D(\infty) \geq D_{max}\}$ by the Monte Carlo method. Rayleigh fading $h_k(t)$ are generated by the following auto-regressive model

$$h_k(t) = \kappa \times h_k(t-1) + v_k(t), \tag{14}$$

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. In this paper, we do simulations with the following parameters fixed: $r_{awgn} = 1000$ kb/s, $K = 10$, $\kappa = 0.8$, and $SNR_{avg} = -40$ dB.

B. Performance Evaluation

We organize this section as follows. Section IV-B.1 shows the convergence of our estimation algorithm. In Section IV-B.2, we assess the accuracy of our QoS estimation (4). Section IV-B.3 investigates the effectiveness of the resource allocation scheme in QoS provisioning. In Section IV-B.4, we evaluate the performance of our scheduler.

1) Convergence of Estimates: This experiment is to show the convergence behavior of estimates. Fig. 2 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 (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 I. 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}.

2) Accuracy of Channel Estimation: This experiment is to show that the estimated effective capacity can indeed be used to accurately predict QoS.

By changing the source rate μ, we simulate three cases, *i.e.*, μ =100, 200, and 300 kb/s. Fig. 3 shows the actual delay-bound violation probability $Pr\{D(\infty) > D_{max}\}$ 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 confirms the exponential dependence shown in (4).

In addition, we use the estimation scheme in [8, page 81] to obtain an estimated θ; with the resulting θ, we predict the probability $Pr\{D(\infty) > D_{max}\}$ (using (4)). As shown in Fig. 3, the estimated $Pr\{D(\infty) > D_{max}\}$ is quite close to the actual $Pr\{D(\infty) > D_{max}\}$. 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 $Pr\{D(\infty) > D_{max}\}$ 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).

We also did simulations under different SNR, different Doppler rates, and different autoregressive channel fading models (a range of AR(1), AR(2) models). Refer to [8, page 137] for details. All the results have shown the exponential behavior of the actual $Pr\{D(\infty) > D_{max}\}$ and the accurateness of our estimation. Due to the space limit, we only show the results for different source rates in Fig. 3. We caution however that such a strong agreement between the bound 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}), such as in the case of high diversity channel fading models (*e.g.*, higher order Nakagami fading models). See [8, page 137] for details.

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. We simulate three data rates, *i.e.*, μ =50, 60, and 70 kb/s, respectively. We do two sets of experiments as below.

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. Fig. 4(a) shows that θ increases with ζ. Thus, Fig. 4(a) 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. 4(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) Performance Gains of Scheduling: This experiment demonstrates the performance gain of the joint scheduling over the RR scheduling, using the optimum $\{\zeta, \beta\}$ values specified by the resource allocation algorithm. In particular, it shows that for loose delay constraints, the large capacity gains promised by the K&H scheme can indeed be approached.

In Fig. 5, 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.*, $K\zeta + \beta = 1$). 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 figure can be directly used to check for feasibility of a QoS pair $\{r_s, \rho\}$, by checking that it satisfies $\theta(r_s) > \rho$. In particular, 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 the joint scheduling.

Three 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 $\sum_{k=1}^{K} \frac{1}{k}$ channel capacity gain for Rayleigh fading channels at low SNR. For example, when $\theta = 0.001$, the channel 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, when $\theta = 0.01$, the channel 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 figure 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 guaranteed to the user are indeed satisfied. Not surprisingly, the simulation result that shows this fact is similar to Fig. 3, 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 the K&H and the RR scheduling, for medium values of θ [8, page 143].

Third, the figure can be used to satisfy the QoS constraint (7), even though it only represents 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.

In summary, the joint scheduler achieves performance gain when delay requirements are not very tight, while yet guaranteeing QoS at any delay requirement.

V. Concluding Remarks

In this paper, we examined the problem of QoS provisioning for K users sharing a single time-slotted Rayleigh 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 the 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.

Acknowledgment

This work was supported by the National Science Foundation under the grant ANI-0111818.

References

[1] 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.

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

[3] 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.

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

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

[6] 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.

[7] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. on Wireless Communications*, vol. 2, no. 4, pp. 630–643, July 2003.

[8] D. Wu, "Providing quality of service guarantees in wireless networks," *Ph.D. Dissertation*, Dept. of Electrical & Computer Engineering, Carnegie Mellon University, Aug. 2003. Available at `http://www.wu.ece.ufl.edu/mypapers/Thesis.pdf`.

PLACE PHOTO HERE

Dapeng Wu (S'98–M'04) received B.E. in Electrical Engineering from Huazhong University of Science and Technology, Wuhan, China, in 1990, M.E. in Electrical Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1997, and Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, PA, in 2003. Since August 2003, he has been with Electrical and Computer Engineering Department at University of Florida, Gainesville, FL, as an Assistant Professor. His research interests are in the areas of networking, communications, multimedia, signal processing, and information and network security. Dr. Wu received the IEEE Circuits and Systems for Video Technology (CSVT) Transactions Best Paper Award for Year 2001.

PLACE PHOTO HERE

Rohit Negi received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Bombay, India in 1995. He received the M.S. and Ph.D. degrees from Stanford University, CA, USA, in 1996 and 2000 respectively, both in Electrical Engineering. He has received the President of India Gold medal in 1995.

Since 2000, he has been with the Electrical and Computer Engineering department at Carnegie Mellon University, Pittsburgh, PA, USA, where he is an Assistant Professor. His research interests include signal processing, coding for communications systems, information theory, networking, cross-layer optimization and sensor networks.

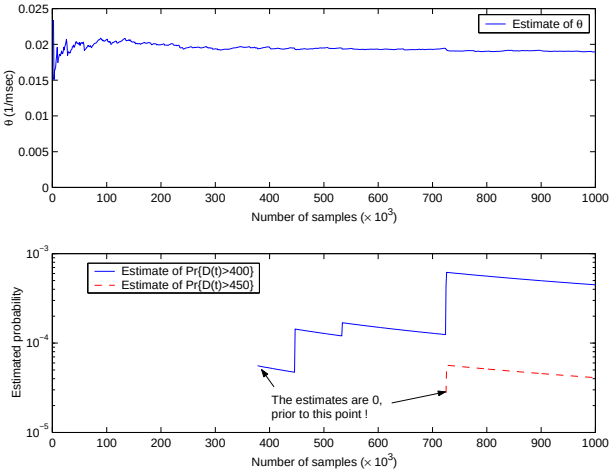

Fig. 2. Convergence of estimates.

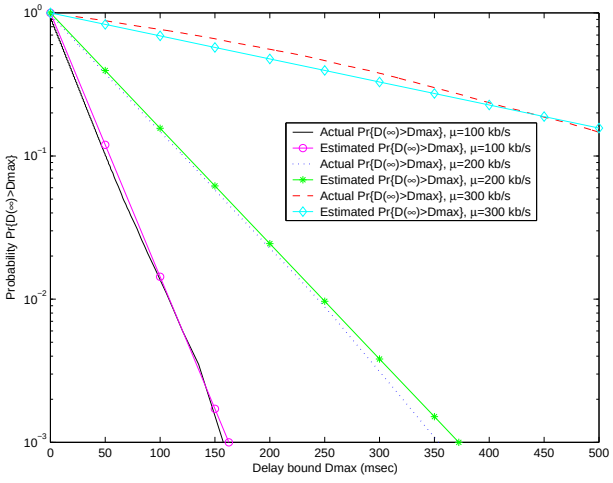

Fig. 3. Actual and estimated delay-bound violation probability.

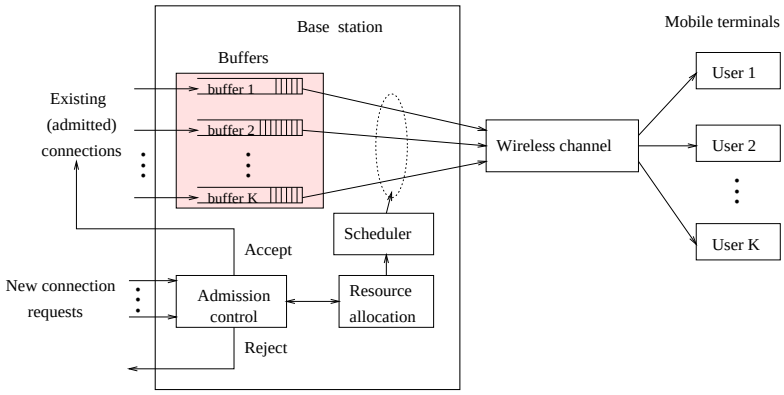

Fig. 1. QoS provisioning architecture in a base station.

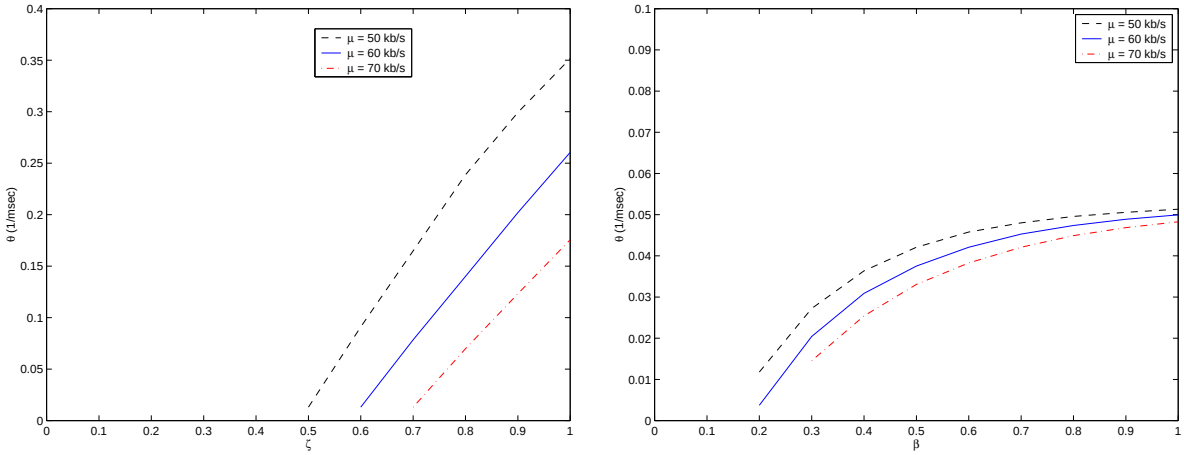

Fig. 4. (a) θ vs. ζ and (b) θ vs. β.

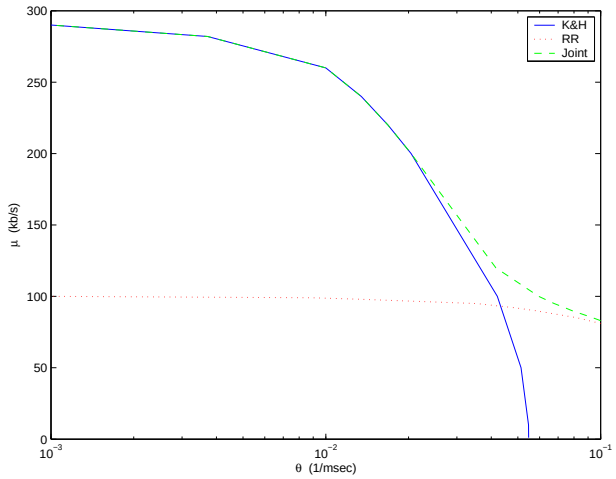

Fig. 5. $\theta(\mu)$ vs. μ.