Downlink Scheduling in a Cellular Network for Quality of Service Assurance

Dapeng Wu* Rohit Negi†

Abstract

We consider the problem of scheduling data in the downlink of a cellular network, over parallel time-varying channels, while providing quality of service (QoS) guarantees, to multiple users in the network. We design simple and efficient admission control, resource allocation, and scheduling algorithms for guaranteeing requested QoS. In our design, a joint K&H/RR scheduler, composed of K&H scheduling and Round Robin (RR) scheduling, utilizes both multiuser diversity and frequency diversity to achieve capacity gain when delay constraints are loose or moderate. However, for tight delay constraints, an additional Reference Channel (RC) scheduler is required to obtain additional frequency diversity gain. The key advantage of our formulation is that the desired QoS constraints can be *explicitly* enforced, by utilizing the concept of effective capacity.

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

*Please direct all correspondence to Prof. Dapeng Wu, University of Florida, Dept. of Electrical & Computer Engineering, P.O.Box 116130, Gainesville, FL 32611, USA. Tel. (352) 392-4954, Fax (352) 392-0044, Email: wu@ece.ufl.edu. URL: http://www.wu.ece.ufl.edu.

†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

Next-generation cellular wireless networks are expected to support multimedia traffic with diverse QoS requirements. Due to wireless channel fading, wherein the channel gains very with time, achieving this goal requires different approaches to QoS provisioning in wireless networks, compared to the wireline counterpart. One of such approaches is to use multiuser diversity [14], which is inherent in a wireless network with multiple users sharing a time-varying channel. 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 [14]. Results [5] 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.

However, channel fading also makes it challenging to guarantee required QoS. In [14], we proposed a joint K&H/RR scheduler to provide explicit QoS guarantees for multiple users sharing *one* channel; essentially, we simplified the scheduler design by shifting the burden to the resource allocation mechanism, and were able to solve the resource allocation problem efficiently, thanks to the recently developed method of *effective capacity* [13]. 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. This paper is intended to extend our work in [14] to the setting of multiple users sharing *multiple* parallel channels, by utilizing both multiuser diversity and frequency diversity. We first begin by applying the joint K&H/RR scheduler in [14] to the multiple channel case. Due to the frequency diversity inherent in multiple wireless channels, the joint K&H/RR scheduler in the new setting can achieve higher capacity gain than that in [14], when delay requirements are loose or moderate. However, we then note that when users' delay requirements are stringent, the joint K&H/RR reduces to the RR

scheduling, and so the high capacity gain due to multiuser diversity associated with the K&H scheduling, vanishes.

To extract more capacity in this case with tight delay requirements, it is desirable to have a scheduler, which at each instant, dynamically selects the best channel among multiple channels for each user to transmit, so as to obtain frequency diversity. In other words, this scheduler must find a channel-assignment schedule, at each time-slot, which minimizes the channel usage while yet satisfying users' QoS requirements. We therefore formulate this scheduling problem as a linear program, in order to avoid the 'curse of dimensionality' associated with optimal dynamic programming solutions. The key idea that allows us to do this, is what we call the Reference Channel (RC) approach, wherein the QoS requirements of the users, are captured by resource allocation (channel assignments). The additional RC approach allows us to obtain capacity gain under tight QoS constraints, by utilizing frequency diversity.

The remainder of this paper is organized as follows. In Section 2, we present efficient QoS provisioning mechanisms and show how to use both multiuser diversity and frequency diversity to achieve a capacity gain while yet satisfying QoS constraints, when transmitting over multiple parallel channels. Section 3 describes our reference-channel-based scheduler that is added to the joint K&H/RR scheduler, to provide a performance gain when delay requirements are tight. In Section 4, we present the simulation results that illustrate the performance improvement of our scheme over the single channel case in [14]. Section 5 discusses the related work. Section 6 concludes the paper.

2 QoS Provisioning with Multiuser Diversity and Frequency Diversity

This section is organized as below. Section 2.1 describes the assumptions and our QoS provisioning architecture. In Section 2.2, we briefly describe the effective capacity technique, which is used in the design of our QoS provisioning schemes. Section 2.3 presents efficient schemes for guaranteeing QoS over multiple parallel channels.

2.1 Architecture

Fig. 1 shows the architecture for transporting multiuser traffic over time-slotted fading channels. A cellular wireless network is assumed, and the downlink is considered, where a base station transmits data over N parallel, independent channels 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. For example, the N channels could be frequency bands which are separated by more than the coherence bandwidth [12, page 202]. A single cell is considered, and interference from other cells is modelled as background noise. We assume a block fading channel model, which assumes that user channel gains are constant over a time duration of length T_s. Therefore, we partition time into 'frames' (indexed as $t = 0, 1, 2, \ldots$), each of length T_s. Thus, each user k has time-varying channel power gains $g_{k,n}(t)$, for each of the N independent channels, which vary with the frame index t. Here $n \in \{1, 2, \ldots, N\}$ refers to the n^{th} channel. The base station is assumed to know the current and past values of $g_{k,n}(t)$. The capacity of the n^{th} channel for the k^{th} user, $c_{k,n}(t)$, is

$$c_{k,n}(t) \;=\; \log_2(1 + g_{k,n}(t) \times P_0/\sigma^2) \;\text{ bits/symbol} \tag{1}$$

where the transmission power P_0 and noise variance σ^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 same channel n can be shared by several users, in the same frame. This is illustrated in Fig. 1, where data from buffers 1 to K can be simultaneously transmitted over channel 1. 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 FDMA-TDMA system, where the N parallel, independent channels represent N frequencies, which are spaced apart (FDMA), and where the frame of each channel consists of TDMA time slots which are infinitesimal.

As shown in Fig. 1, our QoS provisioning architecture consists 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 destined to different mobile users are put into separate queues. The scheduler decides, in each frame t, how to schedule packets for transmission, based on the *current* channel gains $g_{k,n}(t)$ and the amount of resource allocated.

2.2 Effective Capacity

Scheduling requires a computationally efficient procedure to guarantee QoS. In our scheme, we use the recently developed method of effective capacity [13] to guarantee QoS.

We first formally define statistical QoS, which characterizes the user requirement. Consider a single-user system, where the user is allotted a single time varying channel. 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 $\{r_s, D_{max}, \varepsilon\}$.

To test whether this QoS triplet can be satisfied by a given channel and a given scheduler, we use the effective capacity technique, developed in [13]. For convenience, we briefly describe the effective capacity technique as below.

Let $r(t)$ be the instantaneous channel capacity at time t. The *effective capacity function* of $r(t)$ is defined as [13]

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

Consider a queue of infinite buffer size supplied by a data source of *constant* data rate μ (see Fig. 2). It can be shown [13] that if $\alpha(u)$ indeed exists, 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 $r(t)$.

In terms of the effective capacity function (3) defined earlier, the *QoS exponent function* $\theta(\mu)$ can be written as [13]

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

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, the channel can support a QoS triplet $\{r_s, D_{max}, \varepsilon\}$ 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 [15, page 81], we presented a simple and efficient algorithm to estimate $\theta(\mu)$ by direct measurement of the queueing behavior resulting from $r(t)$. Thus, effective capacity provides a computationally efficient procedure to guarantee QoS.

Next, we show our schemes for efficient support of QoS, with the aid of the effective capacity technique.

2.3 QoS Provisioning Schemes

2.3.1 Scheduling

We first explain K&H and RR scheduling separately. In any frame t, the K&H scheduler transmits the data of the user with the largest gain $g_{k,n}(t)$ $(k = 1, 2, \cdots, K)$, for *each* channel n. 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 allocation will be described in Section 2.3.2. It is clear that K&H scheduling attempts to utilize multiuser diversity to maximize the throughput of each channel. Compared to the K&H scheduling over single channel as described in [14], the K&H scheduling here achieves higher throughput when delay requirements are loose. This is because, for fixed ratio[1] N/K, as the number of channel N increases, the number of users K increases, resulting in a larger multiuser diversity gain, which is approximately $\sum_{k=1}^{K} 1/k$ at low SNR.

On the other hand, for *each* channel n, 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. Compared to the RR scheduling over

[1]We fix the ratio N/K so that each user is allotted the same amount of channel resource, for fair comparison.

single channel as described in [14], the RR scheduling here utilizes frequency diversity (each user's data simultaneously transmitted over multiple channels), thereby increasing effective capacity when delay requirements are tight.

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

As a result of the joint K&H/RR scheduling, in frame t, user k has an instantaneous channel capacity, denote by $r^{(k)}(t)$, as below,

$$r^{(k)}(t) = \sum_{n=1}^{N} (\zeta + \beta \times \mathbf{1}(k = k^*(n,t)))c_{k,n}(t), \tag{6}$$

where $\mathbf{1}(\cdot)$ is an indicator function, *i.e.*, $\mathbf{1}(k = a) = 1$ if $k = a$, and $\mathbf{1}(k = a) = 0$ if $k \neq a$. Note that $r^{(k)}(t)$ is the total capacity allocated to user k using the N scheduled channels.

2.3.2 Admission Control and Resource Allocation

The scheduler described in Section 2.3.1 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. Computing $\{\beta, \zeta\}$ could potentially require exponentially complex operations to analyze the joint capacity process of all users and channels. However, the method of effective capacity simplifies this computation dramatically.

Since we only consider the homogeneous case, without loss of generality, denote $\theta_{\zeta,\beta}(\mu)$

the QoS exponent function of user $k = 1$ under the joint K&H/RR scheduling (henceforth called 'joint scheduling'), with frame shares ζ and β respectively. Note that $\theta_{\zeta,\beta}(\mu)$ is the QoS exponent function of the channel, resulting from the joint K&H/RR scheduling with frame shares ζ and β respectively. Assume that each user has homogeneous QoS requirements, characterized by data rate r_s, delay bound D_{max}, and delay-bound violation probability ε. Let $\rho = -\log_e \varepsilon / D_{max}$. The admission control and resource allocation scheme for users requiring the QoS pair $\{r_s, \rho\}$ is given as below,

$$\underset{\{\zeta,\beta\}}{\text{minimize}} \quad K\zeta + \beta \tag{7}$$

$$\text{subject to} \quad \theta_{\zeta,\beta}(r_s) \geq \rho, \tag{8}$$

$$K\zeta + \beta \leq 1, \tag{9}$$

$$\zeta \geq 0, \qquad \beta \geq 0. \tag{10}$$

The minimization in (7) is to minimize the total frame fraction used. (8) ensures that the QoS pair $\{r_s, \rho\}$ of each user is feasible. See [13] for details on the validity of this test. Furthermore, Eqs. (8)–(10) also serve as an admission control test, to check availability of resources to serve this set of users. Since we have the relation $\theta_{\zeta,\beta}(\mu) = \theta_{\lambda\zeta,\lambda\beta}(\lambda\mu)$ (see [15, pp. 270–271] for a proof), we only need to measure the $\theta_{\zeta,\beta}(\cdot)$ functions for different ratios of ζ/β. Notice that solving (7)–(10) is easy, since the function $\theta_{\zeta,\beta}(\mu)$ has been characterized. Thus, the effective capacity method allows considerable simplification in resource allocation, compared to analyzing the joint capacity process of all users and channels.

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

1. Estimate $\theta_{\zeta,\beta}(\mu)$, directly from the queueing behavior, for various values of $\{\zeta, \beta\}$ [13].
2. Determine the optimal $\{\zeta, \beta\}$ pair that satisfies users' QoS, while minimizing frame usage.

3. Provide the joint scheduler with the optimal ζ and β, for simultaneous RR and K&H scheduling, respectively.

It can be seen that the above joint K&H/RR scheduling, admission control and resource allocation schemes utilize both multiuser diversity and frequency diversity. We will show, in Section 4, that such a QoS provisioning achieves higher effective capacity than the one in [14], which utilizes multiuser diversity only.

On the other hand, we observe that when users' delay requirements are stringent (*i.e.*, large ρ), the joint K&H/RR reduces to the RR scheduling (fixed slot assignment) (see Fig. 3). This is because the K&R scheduler is only effective when the delay is large enough to allow each user to achieve the largest capacity among all the users some time during the delay window. Therefore, for tight delay, the high capacity gain associated with the K&H scheduling cannot be achieved. Can the scheduling be modified, so that even with stringent delay requirements, gains over simple RR scheduling can be achieved? To answer this question, we provide an analogy to diversity techniques used in physical layer designs. The careful reader may notice that the RR scheduler proposed in Section 2.3.1 has a similar flavor to equal power distribution used in multichannel transmission, since the RR scheduler equally distributes the traffic of a user over multiple channels in each frame. Since transmitting over the best channel often achieves better performance than equal power distribution, one could ask whether choosing the best channel for each user to transmit would bring about performance gain in the case of tight delay requirements. This is the motivation of designing a reference-channel-based scheduler for tight delay requirements, which we present next.

3 Reference-channel-based Scheduling

Section 2.3 basically extends the K&H/RR scheduling technique of [14], to the case with multiple parallel channels. The drawback of this straight-forward extension was that, al-

though the capacity gain is high for loose or moderate delay requirements (see Section 4.2), the gain vanishes when delay requirements become stringent. This section therefore proposes a scheduler, which squeezes more out of frequency diversity, to provide capacity gains under stringent delay requirements. The rest of the section is organized as follows. We first formulate the reference channel scheduler in Section 3.1. Then in Section 3.2, we analyze the performance of the scheduler.

3.1 'Reference Channel' Approach to Scheduling

Section 2.3 presented our admission control and resource allocation scheme to determine the optimal channel allocation $\{\zeta, \beta\}$ that satisfies users' QoS pair $\{r_s, \rho\}$; the optimal ζ and β are used for simultaneous RR and K&H scheduling, and the resulting scheduled channel can guarantee the QoS pair $\{r_s, \rho\}$ of each user. Hence, with the optimal $\{\zeta, \beta\}$ found by resource allocation, the instantaneous capacity of the scheduled channel $r^{(k)}(t)$ in Eq. (6) is enough to guarantee the QoS pair $\{r_s, \rho\}$ of each user k.

Then, the question is whether the same instantaneous channel capacity specified by Eq. (6) can be provided by a different *instantaneous* schedule, by scheduling the N channels appropriately in each frame t. Assume each frame consists of slots of variable length. Denote $w_{k,n}(t)$ the length of a slot in channel n, allocated to user k in frame t. Then, given that Eq. (6) needs to be satisfied for each user k, the optimal scheduling problem is to find, for each frame t, the set of slot lengths $\{w_{k,n}(t)\}$ that minimizes the channel usage $\sum_{k=1}^{K} \sum_{n=1}^{N} w_{k,n}(t)$,

while satisfying the QoS constraints, that is,

$$\underset{\{w_{k,n}(t)\}}{\text{minimize}} \quad \sum_{k=1}^{K}\sum_{n=1}^{N} w_{k,n}(t) \tag{11}$$

$$\text{subject to} \quad \sum_{n=1}^{N} w_{k,n}(t)c_{k,n}(t) \geq \sum_{n=1}^{N}(\zeta + \beta \times \mathbf{1}(k = k^*(n,t)))c_{k,n}(t), \forall k \tag{12}$$

$$\sum_{k=1}^{K} w_{k,n}(t) \leq 1, \forall\ n \tag{13}$$

$$0 \leq w_{k,n}(t) \leq 1, \forall\ k,\ \forall\ n \tag{14}$$

The constraint (12) represents the QoS constraints since the instantaneous channel capacity specified by Eq. (6) [right hand side in (12)] is enough to satisfy the QoS pair $\{r_s, \rho\}$ of user k. The constraint (13) arises because the total usage of any channel n cannot exceed unity. The intuition of the formulation (11) through (14) is that, the less is the channel usage in supporting QoS for the K users, the more is the bandwidth available for use by other data, such as best-effort traffic. It is obvious that our optimal scheduling problem (*i.e.*, the minimization problem (11)) is simply a linear program.

The key idea in the above optimal scheduler design is to map the QoS requirements $\{r_s, \rho\}$ into a new form, based on the actual time-varying channel capacity specified by Eq. (6); that is, the channel resulting from the K&H/RR scheduling is regarded as a reference channel. Thus, we call the optimal scheduler specified by (11) through (14) as the Reference Channel (RC) scheduler. Thus, when the delay constraints are tight and the admission control allots $\{\beta, \zeta\}$ to the joint K&H/RR scheduler, the RC scheduler minimizes the channel usage at each frame, while yet providing as much capacity to each user as specified by the joint K&H/RR scheduler.

Note that if $\zeta = 0$, *i.e.*, the admission control algorithm allocates channel resources to K&H scheduling only, then the RC scheduler is equivalent to the K&H scheduling since we

have

$$w_{k,n}(t) = \beta \times \mathbf{1}(k = k^*(n,t))), \qquad \forall\ k, \forall n, \tag{15}$$

which means for each channel, the RC scheduler chooses the best user to transmit, and this is exactly the same as the K&H scheduling. So the relation between the joint K&H/RR scheduling and the RC scheduling is that 1) if the admission control allocates channel resources to the RR scheduling due to tight delay requirements, then the RC scheduler can be used to minimize channel usage; 2) if the admission control allocates channel resources to the K&H scheduling only, due to loose delay requirements, then there is no need to use the RC scheduler. The second statement is formally presented in the following proposition.

Proposition 1 *Assume K users share N parallel channels and the K users are scheduled by the K&H scheduling specified in Section 2.3.1. If $\beta = 1$, then there does not exist a channel assignment $\{w_{k,n}(t) : k = 1, \cdots, K; n = 1, \cdots, N\}$ such that $\sum_{k=1}^{K} \sum_{n=1}^{N} w_{k,n}(t) < N$. In this case, there is no need to use the RC scheduler, in addition to the K&H scheduling.*

For a proof of Proposition 1, see the Appendix. Proposition 1 states that if the K users are scheduled by the K&H scheduler with $\beta = 1$, then no channel assignment $\{w_{k,n}(t) : k = 1, \cdots, K; n = 1, \cdots, N\}$ can reduce the channel resource usage.

Next we show a toy example of the capacity gain achieved by the RC scheduler over the RR scheduler. Suppose $K = N = 2$ and the channel capacities at frame t are listed in Table 1. Also assume that channel allocation $\zeta = 1/2$ so that the two channels are completely allocated. Then, using the RC scheduler, at frame t, user 1 will be assigned with 2/3 of channel 1 and user 2 will be assigned with 3/5 of channel 2. Hence, the resulting channel usage is $2/3+3/5 = 19/15 < 2$. So the channel usage of the RC scheduler is reduced, as compared to the RR scheduler.

Table 1: Channel capacities $c_{k,n}(t)$.

	Channel 1	Channel 2
User 1	9	3
User 2	1	5

3.2 Performance Analysis

To evaluate the performance of the RC scheduling algorithm, we introduce two metrics, expected channel usage $\eta(K,N)$ and expected gain $L(K,N)$ defined as below,

$$\eta(K,N) \;=\; \frac{\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbf{E}[\sum_{k=1}^{K}\sum_{n=1}^{N}w_{k,n}(t)]}{N}, \tag{16}$$

where τ is the connection life time and the expectation is over the channel power gains $g_{k,n}(t)$, and

$$L(K,N) \;=\; \frac{1}{\eta(K,N)} \tag{17}$$

The quantity $1-\eta(K,N)$ represents average free channel resource (per channel), which can be used for supporting the users, other than the QoS-assured K users. For example, the frame fractions $\{1-\sum_k w_{k,n}(t)\}$ of each channel n, which are unused after the K users have been supported, can be used for either Best Effort (BE) or Guaranteed Rate (GR) traffic [2]. It is clear that the smaller the channel usage $\eta(K,N)$ (the larger the gain $L(K,N)$), the more free channel resource is available to support BE or GR traffic. The following proposition shows that minimizing $\eta(K,N)$ or maximizing $L(K,N)$ is equivalent to maximizing the capacity available to support BE/GR traffic.

Proposition 2 *Assume that the unused frame fractions* $\{1-\sum_{k=1}^{K}w_{k,n}(t)\}$ *are used solely by* K_B *BE/GR users (indexed by* $K+1, K+2, \cdots, K+K_B$*), whose channel gain processes*

are i.i.d. (in user k and channel n), strict-sense stationary (in time t) and independent of the K QoS-assured users. If the BE/GR scheduler allots each channel to the contending user with the highest channel gain among the K_B users, then the 'available expected capacity',

$$C_{exp} \;=\; \mathbf{E}\left[\sum_{n=1}^{N}(1-\sum_{k=1}^{K} w_{k,n}(t))c_{k^*(n,t),n}(t)\right], \tag{18}$$

is maximized by any scheduler that minimizes $\eta(K,N)$ or maximizes $L(K,N)$. Here, $k^*(n,t)$ denotes the index of the BE/GR user with the highest channel gain among the K_B BE/GR users, for the n^{th} channel in frame t.

For a proof of Proposition 2, see the Appendix.

Next, we present bounds on $\eta(K,N)$ and $L(K,N)$ of the RC scheduler. We consider the case where K users have i.i.d. channel gains which are stationary processes in frame t. The following proposition specifies a lower bound on $\eta(K,N)$ of the RC scheduler.

Proposition 3 *Assume that K users have N i.i.d. channel gains which are strict-sense stationary processes in frame t. Also assume $\zeta = 1/K$; that is, only the RR scheduler is used and the N channels are fully assigned to the K users. Then a lower bound on $\eta(K,N)$ of the RC scheduler specified by (11) through (14), is*

$$\eta(K,N) \geq \mathbf{E}[c_{mean}/c_{max}], \tag{19}$$

where $c_{mean} = \sum_{n=1}^{N} c_{1,n}/N$ and $c_{max} = \max\{c_{1,1}, c_{1,2}, \cdots, c_{1,N}\}$. The time index has been dropped here, due to the assumption of stationarity of the channel gains. Hence, an upper bound on $L(K,N)$ of the RC scheduler specified by (11) through (14), is

$$L(K,N) \leq \frac{1}{\mathbf{E}[c_{mean}/c_{max}]}. \tag{20}$$

For a proof of Proposition 3, see the Appendix.

Furthermore, the following proposition states that the upper bound on gain $L(K,N)$ in (20) monotonically decreases as the average SNR increases.

Proposition 4 *The lower bound on $\eta(K,N)$ in (19),* i.e., $\mathbf{E}[c_{mean}/c_{max}]$, *monotonically increases to 1 as SNR_{avg} increases from 0 to ∞, where $SNR_{avg} = P_0/\sigma^2$ [see Eq. (1)]. Hence, the upper bound on $L(K,N)$ in (20),* i.e., $1/\mathbf{E}[c_{mean}/c_{max}]$, *monotonically decreases to 1 as SNR_{avg} increases from 0 to ∞.*

For a proof of Proposition 4, see the Appendix. Proposition 4 shows that for large SNR_{avg}, there is not much gain to be expected by using the RC scheduler.

So far, we have considered the effect of $\eta(K,N)$ and $L(K,N)$ on the available expected capacity, and derived bounds on $\eta(K,N)$ and $L(K,N)$. In the next section, we evaluate the performance of the RC scheduler and the joint K&H/RR scheduler through simulations.

4 Simulation Results

4.1 Simulation Setting

We simulate the system depicted in Fig. 1. We set the average SNR of each fading channel, fixed at –40 dB. We define r_{awgn} as the capacity of an equivalent AWGN channel, which has the same average SNR, *i.e.*, –40 dB. We set $r_{awgn} = 1000$ kb/s in all the simulations.

The sample interval (frame length) T_s is set to 1 milli-second and each simulation run is 100-second long in all scenarios. Denote $h_{k,n}(t)$ the voltage gain of the n^{th} channel for the k^{th} user. We generate Rayleigh flat-fading voltage-gains $h_{k,n}(t)$ by a first-order auto-regressive (AR(1)) model as below:

$$h_{k,n}(t) = \kappa \times h_{k,n}(t-1) + u_{k,n}(t), \tag{21}$$

where $u_{k,n}(t)$ are i.i.d. complex Gaussian variables with zero mean per dimension. In all the simulations, we set $\kappa = 0.8$, which roughly corresponds to a Doppler rate of 58 Hz [15, page 90].

We only consider the homogeneous case, *i.e.*, each user k has the same QoS requirements $\{r_s, \rho\}$, and the channel gain processes $\{g_{k,n}(t)\}$ are i.i.d in channel n and user k (note that $g_{k,n}(t)$ is not i.i.d. in t).

4.2 Performance Evaluation

4.2.1 Performance Gain of Joint K&H/RR Scheduling

The experiments here are intended to show the performance gain of the joint K&H/RR scheduler in Section 2.3.1 due to utilization of multiple channels. This can be compared with the scheme in [14] where only a single channel was assumed.

The experiments use the optimum $\{\zeta, \beta\}$ values specified by the resource allocation algorithm, *i.e.*, Eqs. (7)–(10). For a fair comparison, we fix the ratio N/K so that each user is allotted the same amount of channel resource for different $\{K, N\}$ pairs. We simulate three cases: 1) $K = 10$, $N = 1$, 2) $K = 20$, $N = 2$, 3) $K = 40$, $N = 4$. For Case 1, the joint K&H/RR scheduler in Section 2.3.1 reduces to the joint scheduler presented in [14].

In Fig. 3(a), we plot the function $\theta(\mu)$ achieved by the joint, K&H, and RR schedulers under Case 3, for a range of source rate μ, when the entire frame of each channel is used (*i.e.*, $K\zeta + \beta = 1$). The function $\theta(\mu)$ in the figure is obtained by the estimation scheme described in [13]. In the case of joint scheduling, each point in the curve of $\theta(\mu)$ corresponds to a specific optimum $\{\zeta, \beta\}$, while $K\zeta = 1$ and $\beta = 1$ are set for RR and K&H scheduling 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. From the figure, 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. Cases 1 and 2 have similar behavior to that plotted in Fig. 3(a).

Fig. 3(b) plots the function $\theta(\mu)$ achieved by the joint K&H/RR scheduler in three cases, for a range of source rate μ, when the entire frame is used (*i.e.*, $K\zeta + \beta = 1$). This figure shows that the larger N is, the higher capacity the joint K&H/RR scheduler in Section 2.3.1 achieves, given each user allotted the same amount of channel resource. This is because the larger N is, the higher diversity the scheduler can achieve. For small θ, the capacity gain is due to multiuser diversity, *i.e.*, there are more users as N increases for fixed N/K; for large θ, the capacity gain is achieved by frequency diversity, *i.e.*, there are more channels to be simultaneously utilized as N increases.

The simulation results in this section demonstrate that the joint K&H/RR scheduler can significantly increase the effective capacity of fading channels, compared with the RR scheduling, for any delay requirement; and the joint K&H/RR scheduler for the multiple channel case achieves higher capacity gain than that for the single channel case.

4.2.2 Performance Gain of RC Scheduling

The experiments in this section are aimed to show the performance gain achieved by the RC scheduler.

We simulate three scenarios for the experiments. In the first scenario, we change the QoS requirement θ while fixing other source/channel parameters. We fix the data rate $r_s =$ 30 kb/s to compare the difference in channel usage achieved by different schedulers. In this scenario, the N channels are not fully allocated by the admission control. Figure 4 shows the expected channel usage $\eta(K, N)$ vs. θ for the RR scheduler, joint K&H/RR scheduler (denoted by "joint" in the figure), and the combination of joint K&H/RR and the

RC scheduler (denoted by "joint+RC" in the figure). It is noted that for $N \geq 2$, the joint K&H/RR scheduler uses less channel resources than the RR scheduler for any θ, and the combination of the joint K&H/RR and the RC scheduler further reduces the channel usage, for large θ. We also observe that 1) for small θ, the K&H scheduler suffices to minimize the channel usage (the RC scheduling does not help since the RC scheduling only improves over the RR scheduling); 2) for large θ, the RC scheduler with fixed channel assignment achieves the minimum channel usage (the K&H scheduler does not help since the K&H scheduler is not applicable for large θ).

In the second and third scenarios, we only simulate the RC scheduler with channel assignments $\zeta = 1/K$ and $\beta = 0$. Hence, the N channels are fully allocated to the K users. We set $K = N$ since the performance gain $L(K, N)$ will remain the same for the same N and any $K \geq N$, if the channels are fully allocated to the K users by the admission control. In the experiments, we choose $\{r_s, \rho\}$ so that $\theta_{\zeta,\beta}(r_s) = \rho$.

In the second scenario, we change the average SNR of the channels while fixing other source/channel parameters. Figure 5(a) shows performance gain $L(K, N)$ vs. average SNR. Just as Proposition 4 indicates, the gain $L(K, N)$ monotonically decreases as the average SNR increases from –40 dB to 15 dB. Intuitively, this is caused by the concavity of the capacity function $c = \log_2(1 + g)$. For high average SNR, a higher channel gain does not result in a substantially higher capacity. Thus, for a high average SNR, scheduling by choosing the best channels (with or without QoS constraints) does not result in a large $L(K, N)$, unlike the case of low average SNR. In addition, Figure 5(a) shows that the gain $L(K, N)$ falls more rapidly for larger N. This is because a larger N results in a larger $L(K, N)$ at low SNR while at high SNR, $L(K, N)$ goes to 1 no matter what N is (see Proposition 4). Figure 5(b) shows the corresponding expected channel usage vs. average SNR.

In the third scenario, we change the number of channels N while fixing other source/channel parameters. Figure 6 shows the performance gain $L(K, N)$ versus number of channels N,

for different average SNRs. It also shows the upper bound (20). From the figure, we observe that as the number of channels increases from 2 to 16, the gain $L(K, N)$ increases. This is because a larger number of channels in the system, increases the likelihood of using channels with large gains, which translates into higher performance gain. Another interesting observation is that the performance gain $L(K, N)$ increases almost linearly with the increase of $\log_e N$ (note that the X-axis in the figure is in a log scale). We also plot the corresponding expected channel usage $\eta(K, N)$ vs. number of channels in Figure 7. The lower bound in Figure 7 is computed by (19). One may notice that the gap between the bound and the actual metric in Figs. 6 and 7 reduces as the number of channels increases. This is because the more channels there is, the less the channel usage is, and hence the more likely each user chooses its best channel to transmit, so that the actual performance gets closer to the bound.[2]

In summary, the joint K&H/RR scheduler for the multiple channel case achieves higher capacity gain than that for the single channel case; the RC scheduler further squeezes out the capacity from multiple channels, when the delay requirements are tight.

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 that we discussed in Section 1, previous works on this topic also include wireless fair queueing [7, 8, 11], modified largest weighted delay first (M-LWDF) [1], opportunistic transmission scheduling [6] and lazy packet scheduling [10].

Wireless fair queueing schemes [7, 8, 11] are aimed at applying Fair Queueing [9] to

[2]In the proof of Proposition 3, we show that the bound corresponds to the case where each user chooses its best channel to transmit.

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 [6] 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 [10] 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 [4], 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 [3, 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.

The RC scheduling approach has similarities to the various scheduling algorithms, which use a 'Virtual time reference', such as Virtual Clock, Fair Queueing (and its packetized versions), Earliest Deadline Due, etc. These scheduling algorithms handle source randomness, by prioritizing the user transmissions, using an easily-computed sequence of transmission times. A scheduler that follows the transmission times, is guaranteed to satisfy the QoS requirements of the users. Similarly, in our work, channel randomness is handled by allotting users an easily-computed 'Virtual channel reference', *i.e.*, the channel assignment $\{\zeta, \beta\}$. A

scheduler (of which the RC scheduler is the optimal version) that allots the time-varying capacities specified by $\{\zeta, \beta\}$, at each time instant, is guaranteed to satisfy the QoS requirements of the users (assuming an appropriate admission control algorithm was used in the calculation of $\{\zeta, \beta\}$).

6 Concluding Remarks

In this paper, we examined the problem of providing QoS guarantees to K users over N parallel time-varying channels. We designed simple and efficient admission control, resource allocation, and scheduling algorithms for guaranteeing requested QoS. We developed two sets of scheduling algorithms, namely, joint K&H/RR scheduling and RC scheduling. The joint K&H/RR scheduling utilizes both multiuser diversity and frequency diversity to achieve capacity gain, and is an extension of our previous work [14]. The RC scheduling is formulated as a linear program, which minimizes the channel usage while satisfying users' QoS constraints. The relation between the joint K&H/RR scheduling and the RC scheduling is that 1) if the admission control allocates channel resources to the RR scheduling due to tight delay requirements, then the RC scheduler can be used to minimize channel usage; 2) if the admission control allocates channel resources to the K&H scheduling only, due to loose delay requirements, then there is no need to use the RC scheduler. Simulation results have demonstrated that substantial gain can be achieved by the joint K&H/RR scheduler and the RC scheduler.

Acknowledgment

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

Appendix

Proof of Proposition 1

We prove it by contradiction. Suppose there exists a channel assignment $\{w_{k,n}(t) : k = 1, \cdots, K; n = 1, \cdots, N\}$ such that $\sum_{k=1}^{K}\sum_{n=1}^{N} w_{k,n}(t) < N$ and $\sum_{n=1}^{N} w_{k,n}(t)c_{k,n}(t) \geq \sum_{n=1}^{N} \mathbf{1}(k = k^*(n,t))c_{k,n}(t), \forall k$ where $k^*(n,t)$ is the index of the user whose capacity $c_{k,n}(t)$ is the largest among K users, for channel n. Since $\sum_{k=1}^{K}\sum_{n=1}^{N} w_{k,n}(t) < N$, there must exist at least one channel n_0 such that $\sum_{k=1}^{K} w_{k,n_0}(t) < 1$. For that channel n_0, $\sum_{k=1}^{K} w_{k,n_0}(t)c_{k,n_0}(t) < c_{k^*(n_0,t),n_0}(t)$, where $k^*(n_0,t)$ is the index of the user with the largest capacity (among K users) in channel n_0, at frame t. For $n \neq n_0$, we have $\sum_{k=1}^{K} w_{k,n}(t)c_{k,n}(t) < c_{k^*(n,t),n}(t)$. Therefore, we obtain

$$\sum_{k=1}^{K}\sum_{n=1}^{N} w_{k,n}(t)c_{k,n}(t) \quad < \quad \sum_{n=1}^{N} c_{k^*(n,t),n}(t) \tag{22}$$

Note the strict inequality in (22). But, since we have the K&H scheduling with $\beta = 1$, we must have

$$\sum_{k=1}^{K}\sum_{n=1}^{N} w_{k,n}(t)c_{k,n}(t) \quad \geq \quad \sum_{n=1}^{N} c_{k^*(n,t),n}(t) \tag{23}$$

(22) and (23) are contradictory. ■

Proof of Proposition 2

By definition of $k^*(n,t)$, the capacities $c_{k^*(n,t),n}(t)$ are independent of $\{c_{k,n}(t), k \leq K\}$, and hence is independent of $\{w_{k,n}(t), k \leq K\}$. Thus, (18) becomes

$$\begin{aligned} C_{exp} &= \sum_{n=1}^{N} \left[\left(1 - \mathbf{E}\left[\sum_{k=1}^{K} w_{k,n}(t) \right] \right) \mathbf{E} c_{k^*(n,t),n}(t) \right] \\ &\overset{(a)}{=} \mathbf{E}[c_{k^*(n,t),n}(t)] \times \left(N - \mathbf{E} \sum_{n=1}^{N} \sum_{k=1}^{K} w_{k,n}(t) \right) \\ &= \mathbf{E}[c_{k^*(n,t),n}(t)] \times (N - N \times \eta(K,N)) \end{aligned}$$

where (a) is due to the fact that $c_{k,n}(t)$ $(k = K+1, \cdots, K+K_B)$ are i.i.d. and strict-sense stationary, and hence $c_{k^*(n,t),n}(t)$ are i.i.d and strict-sense stationary. Therefore, minimizing the expected channel usage $\eta(K,N)$ is equivalent to maximizing the available expected capacity C_{exp}.■

Proof of Proposition 3

It is clear that the minimum value of the objective (11) under the constraint of (12) and (14) [*i.e.*, omitting (13)] is a lower bound on that of (11) under the constraints of (12) through (14). The solution for (11), (12) and (14), is simply that each user only chooses its best channel to transmit (even though the total usage of a channel by all users could be more than 1), *i.e.*,

$$w_{k,n}(t) = \frac{\sum_{m=1}^{N} \zeta c_{k,m}(t)}{c_{k,n}(t)} \times \mathbf{1}(n = \bar{n}(k,t)), \qquad \forall k, \forall n \tag{24}$$

where $\bar{n}(k,t)$ is the index of the channel whose capacity $c_{k,n}(t)$ is the largest among N channels for user k. So we get $\eta(K,N)$ for the scheduler specified by (11) through (14) as

below,

$$\begin{aligned}
\eta(K,N) &\overset{(a)}{=} \frac{\mathbf{E}[\sum_{k=1}^{K}\sum_{n=1}^{N} w_{k,n}(t)]}{N} \\
&\overset{(b)}{\geq} \frac{\mathbf{E}[\sum_{k=1}^{K}(\frac{\sum_{n=1}^{N}\zeta c_{k,n}(t)}{c_{k,\bar{n}(k,t)}(t)})]}{N} \\
&\overset{(c)}{=} \frac{(\sum_{k=1}^{K} N\zeta)\mathbf{E}[\frac{\sum_{n=1}^{N} c_{k,n}/N}{c_{max}}]}{N} \\
&\overset{(d)}{=} \mathbf{E}[\frac{c_{mean}}{c_{max}}]
\end{aligned}$$

where (a) due to the fact that $c_{k,n}(t)$ are stationary, thereby $w_{k,n}(t)$ being stationary, (b) since the assignment in (24) gives a lower bound, (c) since $c_{k,n}(t)$ are i.i.d. and stationary, and (d) due to $\zeta = 1/K$. This completes the proof. ■

Proof of Proposition 4

We first present a lemma and then prove Proposition 4.

Let $\gamma = P_0/\sigma^2$. Denote g_1 and g_2 channel power gains of two fading channels, respectively.

Lemma 1 *If $g_1 > g_2 > 0$, then $\log(1+\gamma\times g_2)/\log(1+\gamma\times g_1)$ monotonically increases from g_2/g_1 to 1, as γ increases from 0 to ∞.*

Proof:

Define $f(\gamma) = \log(1+\gamma g_2)/\log(1+\gamma g_1)$. By L'Hospital's rule, it is obvious that $f(\gamma) \to g2/g1$ as $\gamma \downarrow 0$ and $f(\gamma) \to 1$ as $\gamma \to \infty$. Now, we only need to show $f'(\gamma) > 0$ for $\gamma > 0$. Taking the derivative results in

$$f'(\gamma) = \frac{\frac{g_2}{1+\gamma g_2}\log(1+\gamma g_1) - \frac{g_1}{1+\gamma g_1}\log(1+\gamma g_2)}{\log^2(1+\gamma g_1)} \tag{25}$$

Then, we only need to show

$$\frac{g_2}{1+\gamma g_2}\log(1+\gamma g_1) > \frac{g_1}{1+\gamma g_1}\log(1+\gamma g_2) \tag{26}$$

or equivalently, that,

$$\frac{\frac{g_2}{(1+\gamma g_2)\log(1+\gamma g_2)}}{\frac{g_1}{(1+\gamma g_1)\log(1+\gamma g_1)}} > 1 \tag{27}$$

Define $h(x) = \frac{x}{(1+\gamma x)\log(1+\gamma x)}$. If $h'(x) < 0$ for $x > 0$, then $g_1 > g_2 > 0$ implies $h(g_2)/h(g_1) > 1$, which is the inequality in (27). So we only need to show $h'(x) < 0$ for $x > 0$. Taking the derivative, we have

$$h'(x) = \frac{\frac{1+\gamma x-\gamma x}{(1+\gamma x)^2}\log(1+\gamma x) - \frac{\gamma}{1+\gamma x}\frac{x}{1+\gamma x}}{\log^2(1+\gamma x)} \tag{28}$$

$$= \frac{\frac{1}{(1+\gamma x)^2}(\log(1+\gamma x) - \gamma x)}{\log^2(1+\gamma x)} \tag{29}$$

For $\gamma > 0$ and $x > 0$, we have $\log(1+\gamma x) - \gamma x < 0$, which implies $h'(x) < 0$. This completes the proof.■

Next, we prove Proposition 4.

Since $g_{1,n}$ $(n = 1, \cdots, N)$ are identically distributed processes, we have

$$\mathbf{E}\left[\frac{c_{mean}}{c_{max}}\right] = \mathbf{E}\left[\frac{\log(1+\gamma \times g_{1,1})}{\log(1+\gamma \times g_{max})}\right] \tag{30}$$

where $g_{max} = \max_{n\in\{1,2,\cdots,N\}} g_{1,n}$. Since $g_{max} \geq g_{1,1} > 0$, then by Lemma 1, $\log(1+\gamma \times g_{1,1})/\log(1+\gamma \times g_{max})$ monotonically increases from $g_{1,1}/g_{max}$ to 1, as γ increases from 0 to ∞. Hence, $\mathbf{E}[c_{mean}/c_{max}]$ monotonically increases from $\mathbf{E}[g_{1,1}/g_{max}]$ to 1, as γ increases from 0 to ∞. 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] 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.

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

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

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

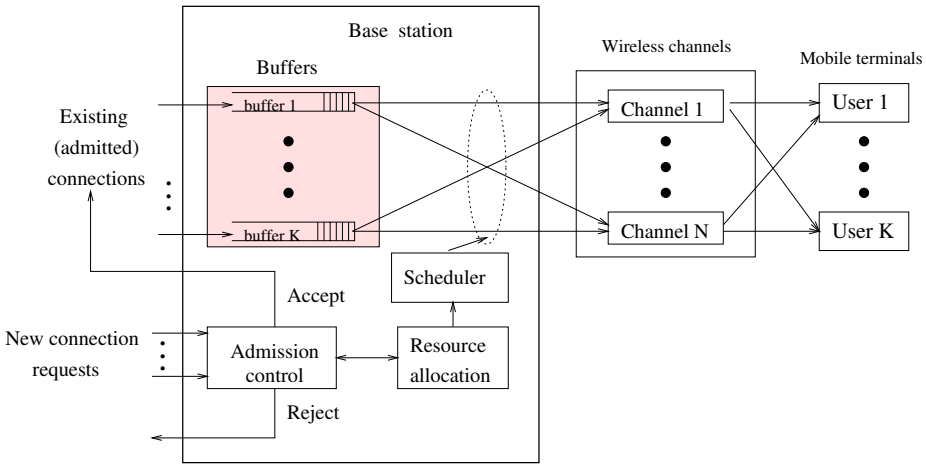


Figure 1: QoS provisioning architecture in a base station.

[10] 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.

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

[12] T. S. Rappaport, *Wireless Communications: Principles & Practice,* 2nd Ed., Prentice Hall, 2002.

[13] 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.

[14] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE ICC'03,* Anchorage, Alaska, USA, May 2003.

[15] 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`.

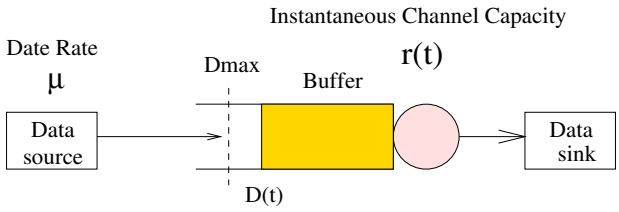
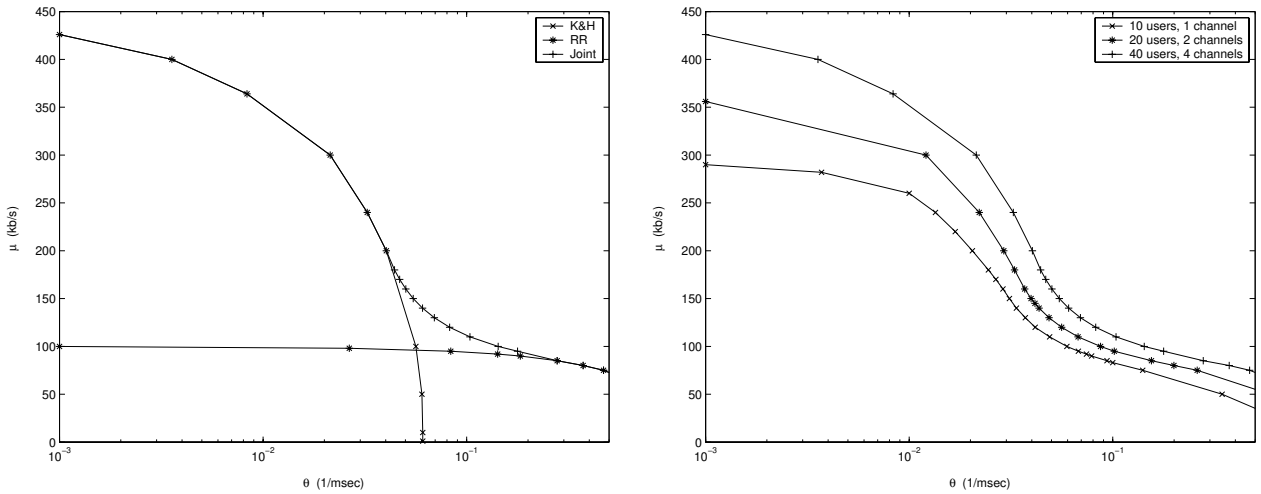


Figure 2: A queueing system model.



(a) (b)

Figure 3: (a) $\theta(\mu)$ vs. μ for K&H, RR, and joint scheduling ($K = 40, N = 4$), and (b) $\theta(\mu)$ vs. μ for joint K&H/RR scheduling.

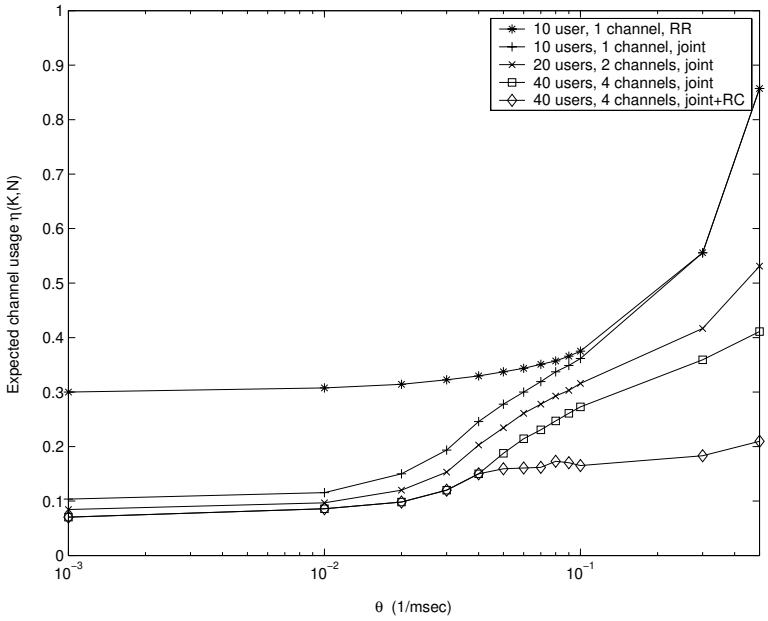
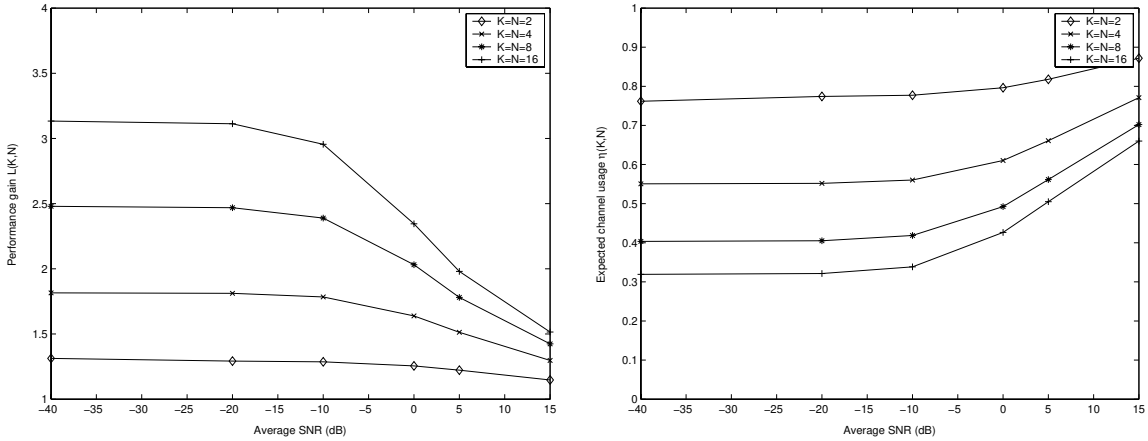
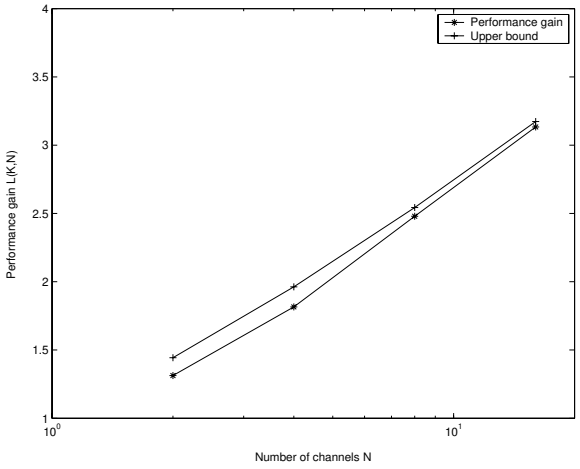


Figure 4: Expected channel usage $\eta(K, N)$ vs. θ.

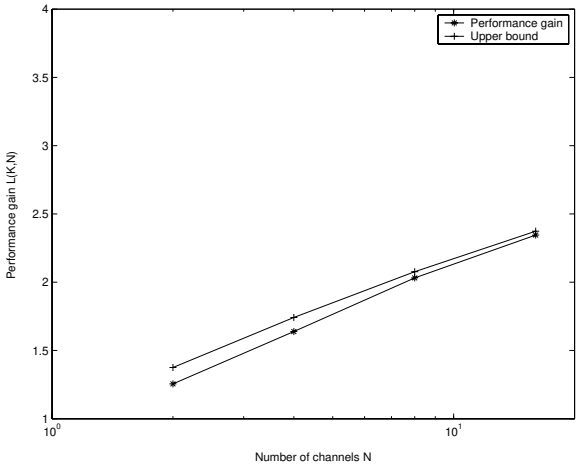


(a) (b)

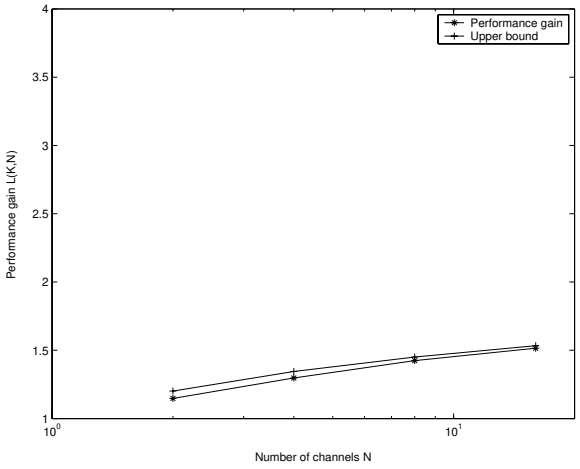
Figure 5: (a) Performance gain $L(K, N)$ vs. average SNR, and (b) $\eta(K, N)$ vs. average SNR.



(a)

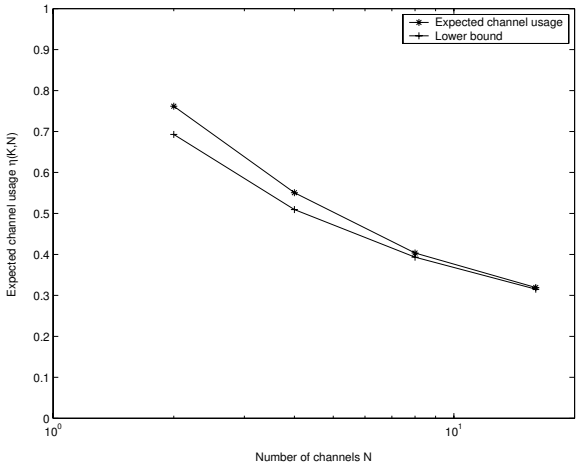


(b)

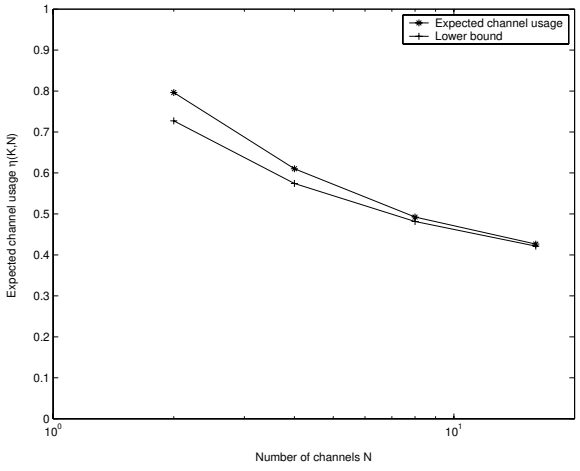


(c)

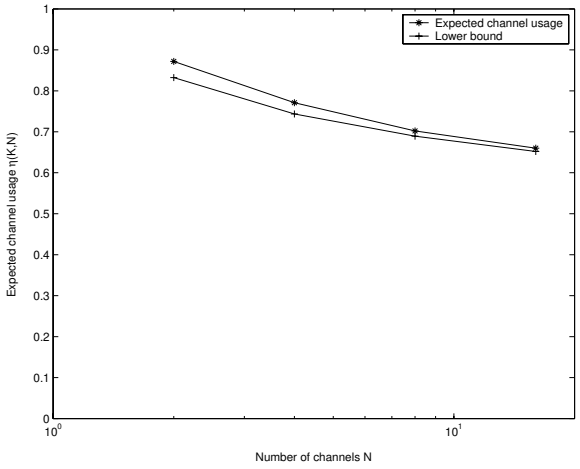
Figure 6: $L(K, N)$ vs. number of channels N for average SNR = (a) –40 dB, (b) 0 dB, and (c) 15 dB.



(a)



(b)



(c)

Figure 7: $\eta(K, N)$ vs. number of channels N for average SNR = (a) –40 dB, (b) 0 dB, and (c) 15 dB.