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. Our scheduling algorithms consists of two sets, namely, (what we call) joint K&H/RR scheduling and Reference Channel (RC) scheduling. The joint K&H/RR scheduling, composed of K&H scheduling and Round Robin (RR) scheduling, utilizes both multiuser diversity and frequency diversity to achieve capacity gain, and the RC scheduling 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.

In designing the RC scheduler, we propose a *reference channel* approach and formulate the scheduler as a linear program, dispensing with complex dynamic programming approaches, by the use of a resource allocation scheme. An advantage of this formulation is that the desired QoS constraints can be *explicitly* enforced, by allotting sufficient channel resources to users, during call admission.

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

*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

Next-generation cellular wireless networks are expected to support multimedia traffic with diverse quality-of-service (QoS) requirements. Due to time variations of wireless channel condition, achieving this goal requires different approaches to QoS provisioning in wireless networks, compared to the wireline counterpart. One of such approaches is to use diversity.

Diversity techniques using the time, space, or frequency dimensions can be used to increase the outage capacity [4] of a fading channel, by minimizing the probability of deep fades. These traditional diversity methods are essentially applicable to a single-user link. In a wireless network with multiple users sharing a time-varying channel, another diversity, termed *multiuser diversity* [8], was proposed by Knopp and Humblet [12] to increase the channel capacity. 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 [23]. Results [12] 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.

It is known [23] that the K&H scheduling maximizes ergodic capacity but it provides no delay guarantees. To combat this problem, a natural solution is to combine the K&H scheduling with the RR scheduling, since it can leverage the best features of K&H scheduling (maximizing capacity) and RR scheduling (achieving low delay) [3]. However, designing such a scheduler with explicit QoS guarantees to each user, is not a trivial task. To explicitly enforce QoS guarantees, a typical procedure of QoS provisioning design involves four steps:

1. Channel measurement: e.g., measure the channel capacity process [11].
2. Channel modeling: e.g., use a Markov-modulated Poisson process to model the channel capacity process [11].
3. Deriving QoS measures: e.g., analyze the queue and derive the delay distribution, given the Markov-modulated Poisson process as the service model [11].
4. Relating the control parameters of QoS provisioning mechanisms to the derived QoS measures: e.g., relate the control parameters of the joint scheduler to the QoS measures.

Steps 1 to 3 are intended to analyze the QoS provisioning mechanisms, whereas step 4 is aimed at designing the QoS provisioning mechanisms. However, the main obstacle of applying the four steps in QoS provisioning, is high complexity in characterizing the relation between the control parameters and the calculated QoS measures. For example, one could use queueing analysis (having a complexity that is exponential in the number of users [23]) to determine what percentage of the channel resource should be allocated to the K&H and RR scheduling respectively, so that a specified QoS can be satisfied. But the queueing analysis does not result in a close-form relation between the control parameters and the QoS measures [24].

Recognizing that the key difficulty in explicit QoS provisioning, is the lack of a method that can easily relate the control parameters of a QoS provisioning system to the QoS measures, we proposed an approach in [23], which simplifies the task of explicit provisioning of QoS guarantees. Specifically, we simplify the design of joint K&H/RR scheduler by shifting the burden to the resource allocation mechanism. Furthermore, we are able to solve the resource allocation problem efficiently, thanks to 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.

Different from [23], which addressed QoS provisioning for multiple users sharing *one* channel, this paper extends the joint K&H/RR scheduling method to the setting of multiple users sharing *multiple* channels, by utilizing both multiuser diversity and frequency diversity. As a result, the joint scheduler in the new setting achieves higher capacity gain than that in [23]. Moreover, when users' delay requirements are stringent, wherein channel resources have to be allocated for the RR scheduling (fixed slot assignment) [23], the high capacity gain associated with K&H scheduling vanishes. To squeeze out more capacity in this case, a possible solution is to design a scheduler, which dynamically selects the best channel among multiple channels for a user to transmit. In other words, this scheduler is intended to find a channel-assignment schedule, at each time-slot, which minimizes the channel usage under users' QoS constraints.

We 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' approach, wherein the QoS requirements of the users, are captured by resource allocation (channel assignments). The scheduler obtained, as a result of the Reference Channel approach, is sub-optimal. Therefore, we analyze the performance of this scheduler, by comparing its performance gain with a bound we derived. We show by simulations, that the performance of our sub-optimal scheduler is quite close to the bound. This demonstrates the effectiveness of our scheduler. The performance gain is obtained, as a result of dynamically choosing the best channel to transmit.

The remainder of this paper is organized as follows. In Section 2, we present efficient QoS provisioning mechanisms and show how to use multiuser diversity and frequency diversity to achieve a capacity gain while yet satisfying QoS constraints. Section 3 describes our reference-channel-based scheduler that provides 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 that in [23]. Section 5 discusses the related work. In Section 6, we conclude the paper.

2 QoS Provisioning with Multiuser Diversity and Frequency Diversity

This section is organized as below. Section 2.1 describes the assumptions and the QoS provisioning architecture we use. In Section 2.2, we overview the technique of effective capacity. Section 2.3 presents efficient schemes for guaranteeing QoS.

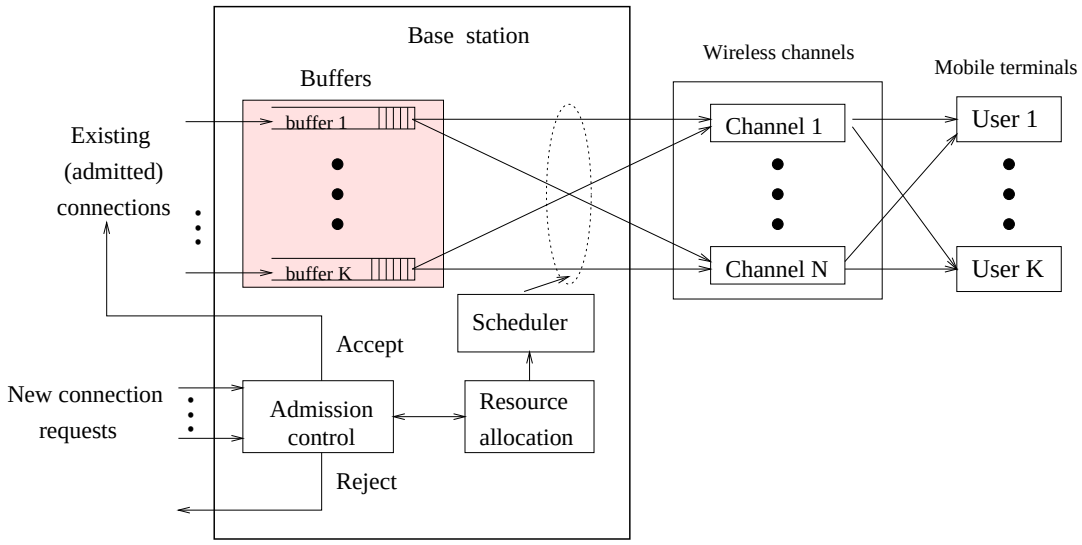


Figure 1: QoS provisioning architecture in a base station.

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. A single cell is considered, and interference from other cells is modelled as background noise. 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 $t = 0, 1, 2, \ldots$), 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

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

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[1] 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. Note that in a practical FDMA-TDMA system, there would be a finite number of finite-length time slots in each frame, rather than the infinite number of infinitesimal time slots, assumed here.

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[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 $g_{k,n}(t)$ and the amount of resource allocated.

Next, we describe the technique of effective capacity, which is a crucial tool in designing our QoS provisioning mechanisms.

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, 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. The key contribution of Ref. [22] was to introduce a concept of statistical delay-constrained capacity termed *effective capacity*, which allowed us to obtain a simple and efficient test, to check the feasibility of QoS triplets for a single time-varying channel. Furthermore, in [23], we showed how to apply the effective capacity concept to the K&H scheduled channel. Therefore, we briefly explain the concept of effective capacity, and refer the reader to [22, 23] for details.

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

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

[1]FDMA is frequency-division multiple access and TDMA is time-division multiple access.

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

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 [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. 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 [22], we present a simple and efficient algorithm to estimate $\theta(\mu)$ by direct measurement on the queueing behavior resulting from $r(t)$.

Now, having described our basic technique, *i.e.*, effective capacity, in the next section, we present schemes for scheduling, admission control and resource allocation, which utilize this technique for efficient support of QoS. 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.

2.3 QoS Provisioning Schemes

2.3.1 Scheduling

As explained in Section 1, we simplify the scheduler, by shifting the burden of guaranteeing users' QoS to resource allocation. Therefore, our scheduler is a simple combination of K&H and RR 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 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 [23], the K&H scheduling here achieves higher throughput when delay requirements are loose. This is because, for fixed ratio[3] N/K, as the number of channel N increases, the number of users K increases, resulting in a larger capacity gain, which is approximately $\sum_{k=1}^{K} 1/k$.

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 single channel as described in [23], 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, its operation is the following. 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 $\beta + K\zeta \leq 1$.

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.

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). Note that since the capacity process $r(t)$ depends on the number of users K and the number of channels N, $\theta_{\zeta,\beta}(\mu)$ is actually a function of K and N. However, since we assume K and N are fixed, there is no need to put the extra arguments K and N in the function $\theta_{\zeta,\beta}(\mu)$. With this simplified notation, the admission

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

control and resource allocation scheme for users requiring the QoS pair $\{r_s, \rho\}$ is given as below,

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

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

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

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

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 relation $\theta_{\zeta,\beta}(\mu) = \theta_{\lambda\zeta,\lambda\beta}(\lambda\mu)$ (its proof is similar to that in [23]), we only need to measure the $\theta_{\zeta,\beta}(\cdot)$ functions for different ratios of ζ/β.

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 $\{\zeta, \beta\}$.
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 described in [23], which utilizes multiuser diversity only.

On the other hand, we observe that when users' delay requirements are stringent, the RR scheduling (fixed slot assignment) has to be used (see Fig. 4). Then the high capacity gain associated with K&H scheduling cannot be achieved (see Fig. 4). A careful reader may notice that the RR scheduler proposed in Section 2.3.1 has a similar flavor to equal gain combining used in multichannel receivers [21, page 262], since the RR scheduler equally distributes the traffic of a user over multiple channels in each frame. Since selection combining (choosing the channel with the highest SNR) [21, page 262] achieves better performance than equal gain combining, one could ask whether choosing the best channel for a 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, which we present next.

3 Reference-channel-based Scheduling

This section is organized as follows. We first formulate the downlink scheduling problem in Section 3.1. Then in Section 3.2, we propose a Reference channel approach to the problem and with

this approach we design the scheduler by a linear program. In Section 3.3, we investigate the performance of the scheduler.

3.1 The Problem of Optimal Scheduling

Let $w_{k,n}(t)$ ($w_{k,n}(t)$ are real numbers in the interval $[0,1]$) be the fraction of channel n, allotted by the base station to user k, in frame t.

The scheduling problem is to find, for each frame t, the set of $\{w_{k,n}(t)\}$ that minimizes the time-averaged expected channel usage $\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbf{E}[\sum_{k=1}^{K}\sum_{n=1}^{N}w_{k,n}(t)]$ (where τ is the connection life time), given the QoS constraints, as below,

$$\underset{\{w_{k,n}(t)\}}{\text{minimize}} \quad \frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbf{E}[\ \sum_{k=1}^{K}\sum_{n=1}^{N}w_{k,n}(t)\ ] \tag{10}$$

$$\text{subject to} \quad \sup_{t} Pr\{D_k(t) \geq D_{max}^{(k)}\} \leq \varepsilon_k, \qquad \text{for a fixed rate } r_s^{(k)}, \qquad \forall\ k \tag{11}$$

$$\sum_{k=1}^{K} w_{k,n}(t) \leq 1, \qquad \forall\ n,\ \forall\ t \tag{12}$$

$$w_{k,n}(t) \geq 0, \qquad \forall\ k,\ \forall\ n,\ \forall\ t \tag{13}$$

The constraint (11) represents statistical QoS constraints, that is, each user k specifies its QoS by a triplet $\{r_s^{(k)}, D_{max}^{(k)}, \varepsilon_k\}$, which means that each user k, transmitting at a fixed data rate $r_s^{(k)}$, requires that the probability of its packet delay $D_k(t)$ exceeding the delay bound $D_{max}^{(k)}$, is not greater than ε_k. The constraint (12) arises because the total usage of any channel n cannot exceed unity. The intuition of the formulation (10) through (13) 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 or Guaranteed Rate traffic [7].

We call any scheduler, which achieves the minimum in (10), as the *optimal scheduler*. To meet the statistical QoS requirements of the K users, an optimal scheduler needs to keep track of the queue length, for each user, using a state variable. It would make scheduling decisions (*i.e.,* allocation of $\{w_{k,n}(t)\}$), based on the current state. Dynamic programming often turns out to be a natural way to solve such an optimization problem [5, 6]. However, the dimensionality of the state variable is typically proportional to the number of users (at least), which results in very high (exponential in number of users) complexity for the associated dynamic programming solution [2]. Simpler approaches, such as [1], which use the state variable sub-optimally, do not enforce a given QoS, but rather seek to optimize some form of a QoS parameter.

This motivates us to seek a simple (sub-optimal) approach, which can enforce the specified QoS constraints explicitly, and yet achieve an efficient channel usage. This idea is elaborated in the next section.

3.2 'Reference Channel' Approach to Scheduling

The key idea in the scheduler design is to specify the QoS constraints, using (what we call) the 'Reference Channel' approach. In the original optimal scheduling problem (10), the statistical QoS constraints (11) are specified by triplets $\{r_s^{(k)}, D_{max}^{(k)}, \varepsilon_k\}$. However, we map these constraints into a new form, based on the actual time-varying channel capacities of the K users. To elaborate, we assume that the base station can measure the statistics of the time-varying channel capacities (for example, the QoS exponent function $\theta(\mu)$ described in Section 2.2). Further, it is assumed that an appropriate admission control and resource allocation algorithm (such as that in Section 2.3.2), allots a fraction $\xi_{k,n}$ ($\xi_{k,n}$ are real numbers in the interval $[0, 1]$) of channel n, to user k, for the duration of the connection time. In other words, the key idea of the admission control and resource allocation algorithm is that, if a given user k were allotted the *fixed* channel assignment $\{\xi_{k,n}\}$ during the entire connection period, then the time-varying capacity $\sum_{n=1}^{N} \xi_{k,n} c_{k,n}(t)$, which it would obtain, would be sufficient to fulfill its QoS requirements specified by $\{r_s^{(k)}, D_{max}^{(k)}, \varepsilon_k\}$. A necessary condition on $\xi_{k,n}$ is that,

$$\sum_{k=1}^{K} \xi_{k,n} \leq 1, \qquad \forall\ n \tag{14}$$

Thus, our approach shifts the complexity of satisfying the QoS requirements (11), from the scheduler to the admission control algorithm, which needs to ensure that its choice of channel assignment $\{\xi_{k,n}\}$, meets the QoS requirements of all the users. Since the QoS constraint (11) is embedded in the channel assignment $\{\xi_{k,n}\}$, hence we call our approach to scheduling as a 'Reference Channel' approach. A careful reader may note a similarity of this approach, to other virtual reference approaches [25, 26], which are used to handle *source randomness* in wireline scheduling. Our motivation, on the other hand, is to handle *channel randomness* in wireless scheduling. This point is discussed in more detail in Section 5.

Thus, with the QoS constraints embedded in the $\{\xi_{k,n}\}$, the QoS constraint (11) can be replaced by the specific set of constraints,

$$\sum_{n=1}^{N} w_{k,n}(t) c_{k,n}(t) \geq \sum_{n=1}^{N} \xi_{k,n} c_{k,n}(t), \quad \forall\ k \tag{15}$$

Note that the channel fractions $w_{k,n}(t)$ and $\xi_{k,n}$ perform different functions. The fractions $w_{k,n}(t)$ are assigned by a *scheduler*, depending on the channel gains it observes, and they specify the actual fractions of the N channel frames used by different users at time t. Thus, they will (in general) vary with time. On the other hand, the fractions $\xi_{k,n}$ are assigned by an *admission control and resource allocation algorithm*, and they represent the channel resources *reserved* for different users, rather than the actual fractions of the N channel frames *used* by the users. Thus, $\xi_{k,n}$ are fixed during the life time of a connection.

It is clear that (15) ensures that in every frame t, the scheduler will allot each user k a capacity, which is not less than the capacity specified by the $\xi_{k,n}$. Thus, a scheduler that satisfies (15) is

guaranteed to satisfy the QoS requirements of all the K users. However, in the process of replacing the QoS constraint (11), by the constraint (15), we have conceivably tightened the constraints on the scheduler (since the latter constraint needs to be at least as tight as the former), which means that the scheduler we will derive will be sub-optimal, with respect to the optimal scheduler (10) through (13). However, as will be shown, this modification results in a simpler scheduler, which achieves a performance close to a bound we derived.

To summarize, we derive a sub-optimal scheduler, which we call Reference Channel (RC) scheduler, based on the optimization problem below: for each frame t,

$$\underset{\{w_{k,n}(t)\}}{\text{minimize}} \quad \sum_{k=1}^{K}\sum_{n=1}^{N} w_{k,n}(t) \tag{16}$$

$$\text{subject to} \quad \sum_{n=1}^{N} w_{k,n}(t)c_{k,n}(t) \geq \sum_{n=1}^{N} \xi_{k,n}c_{k,n}(t), \qquad \forall\ k \tag{17}$$

$$\sum_{k=1}^{K} w_{k,n}(t) \leq 1, \qquad \forall\ n \tag{18}$$

$$w_{k,n}(t) \geq 0, \quad \forall\ k,\ \forall\ n \tag{19}$$

Notice that the cost function in (16) is different from the one in (10), since we have dispensed with the expectation and time-averaging in (16). This can be done, because the fractions $w_{k,n}(t)$ at time t, can be optimally chosen *independent of future channel gains*, thanks to the Reference Channel formulation. Thus, interestingly, whereas the optimal scheduler state would need to incorporate the channel states of the $N \times K$ fading channels (if they are correlated between different frames t), our sub-optimal scheduler does not need to do so, since the correlations in the channel fading process have been already accounted for by the admission control algorithm!

It is obvious that our sub-optimal scheduling problem (*i.e.*, the minimization problem (16)) is simply a linear program. The solution (scheduler) can be found with low complexity, by either the simplex method or interior-point methods [16, pp. 362–417].

The constraint (17) is for the case of fixed channel assignment (associated with RR scheduling). If the admission control and resource allocation algorithm in Section 2.3.2 is used, the constraint (17) becomes

$$\sum_{n=1}^{N} w_{k,n}(t)c_{k,n}(t) \geq \sum_{n=1}^{N} (\zeta + \beta \times \mathbf{1}(k = k^*(n,t)))c_{k,n}(t), \qquad \forall\ k \tag{20}$$

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, and $\mathbf{1}(\cdot)$ is an indicator function such that $\mathbf{1}(k=a)=1$ if $k=a$, and $\mathbf{1}(k=a)=0$ if $k \neq a$. 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

$$w_{k,n}(t) = \beta \times \mathbf{1}(k = k^*(n,t))), \qquad \forall\ k, \forall n, \tag{21}$$

which means for each channel, choosing 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.

In the next section, we investigate the performance of our RC scheduler. In particular, since the optimal scheduler (based on dynamic programming) is very complex, we present a simple bound for evaluating the performance of the RC scheduler. Then, in Section 4 we show that the performance of the RC scheduler is close to the bound.

3.3 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,

$$\eta(K,N) \;=\; \frac{\frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbf{E}[\sum_{k=1}^{K}\sum_{n=1}^{N}w_{k,n}(t)]}{N}, \tag{22}$$

where the expectation is over $g_{k,n}(t)$, and

$$L(K,N) \;=\; \frac{1}{\eta(K,N)} \tag{23}$$

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 [7]. It is clear that the smaller channel usage $\eta(K,N)$ (the larger gain $L(K,N)$), the more free channel resource 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 1 *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',*

$$C_{exp} \;=\; \mathbf{E}\left[\sum_{n=1}^{N}(1-\sum_{k=1}^{K}w_{k,n}(t))c_{k^*(n,t),n}(t)\right], \tag{24}$$

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 1, see the Appendix.

Next, we present bounds on $\eta(K,N)$ and $L(K,N)$, which will be used to evaluate the performance of the RC scheduler.

Computing (22) for the optimal scheduler (10) through (13) is complex, because the optimal scheduler itself has high complexity. For this reason, we seek to derive a lower bound on $\eta(K,N)$ 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 2 *Assume that K users have N i.i.d. channel gains which are strict-sense stationary processes in frame t. Each user k has channel assignments $\{\xi_{k,n}\}$, where $\xi_{k,n}$ are equal for fixed k and all n ($n = 1, 2, \cdots, N$). Assume that the N channels are fully assigned to the K users, i.e.,*

$$\sum_{k=1}^{K}\sum_{n=1}^{N}\xi_{k,n} = N \tag{25}$$

Then a lower bound on $\eta(K,N)$ of the RC scheduler specified by (16) through (19), is

$$\eta(K,N) \geq \mathbf{E}[c_{mean}/c_{max}], \tag{26}$$

where $c_{mean} = \sum_{n=1}^{N} c_{k,n}/N$ and $c_{max} = \max\{c_{k,1}, c_{k,2}, \cdots, c_{k,N}\}$. 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 (16) through (19), is

$$L(K,N) \leq \frac{1}{\mathbf{E}[c_{mean}/c_{max}]}. \tag{27}$$

For a proof of Proposition 2, see the Appendix.

Furthermore, the following proposition states that the upper bound on $L(K,N)$ in (27) monotonically decreases as average SNR increases.

Proposition 3 *The lower bound on $\eta(K,N)$ in (26),* i.e., *$\mathbf{E}[c_{mean}/c_{max}]$, monotonically increases to 1 as SNR_{avg} increases from 0 to ∞, where $SNR_{avg} = P_0/\sigma^2$. Hence, the upper bound on $L(K,N)$ in (27),* 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 3, see the Appendix.

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.

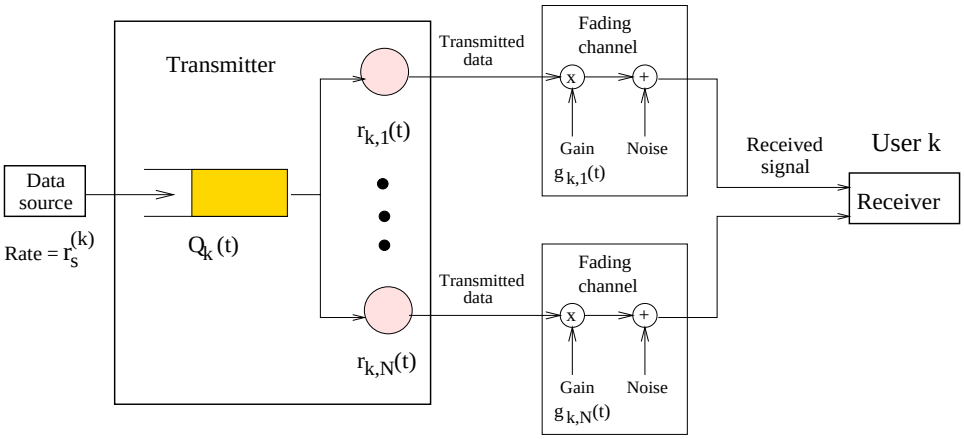


Figure 2: The queueing model used for simulations.

4 Simulation Results

4.1 Simulation Setting

We simulate the system depicted in Fig. 1, in which each connection[4] is simulated as plotted in Fig. 2. In Fig. 2, 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 N fading channels at data rate $\sum_{n=1}^{N} r_{k,n}(t)$. Each fading channel n has a random power gain $g_{k,n}(t)$ (the noise variance is absorbed into $g_{k,n}(t)$). 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.

We assume that the transmitter has perfect knowledge of the current channel gains $g_{k,n}(t)$ in frame t. Therefore, it can use rate-adaptive transmissions, and ideal channel codes, to transmit packets without decoding errors. Under the joint K&H/RR scheduling, the transmission rate $r_{k,n}(t)$ of user k over channel n, is given as below,

$$r_{k,n}(t) = (\zeta + \beta \times \mathbf{1}(k = k^*(n,t)))c_{k,n}(t), \tag{28}$$

where the instantaneous channel capacity $c_{k,n}(t)$ is

$$c_{k,n}(t) = B_c \log_2(1 + g_{k,n}(t) \times P_0/\sigma^2) \tag{29}$$

where B_c is the channel bandwidth. On the other hand, for the combination of joint K&H/RR and RC scheduling, the transmission rate $r_{k,n}(t)$ of user k over channel n, is set as,

$$r_{k,n}(t) = w_{k,n}(t)c_{k,n}(t). \tag{30}$$

where $\{w_{k,n}(t)\}$ is a solution to the linear program specified by (16), (18), (19) and (20).

[4]Assume that K connections are set up and each mobile user is associated with only one connection.

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.*,

$$r_{awgn} = B_c \log_2(1 + SNR_{avg}) \tag{31}$$

where $SNR_{avg} = E[g_{k,n}(t) \times P_0/\sigma^2] = P_0/\sigma^2$, assuming that the transmission power P_0 and noise variance σ^2 are constant and equal for all users, in a simulation run. We set $E[g_{k,n}(t)] = 1$. Then, we can eliminate B_c using Eqs. (29) and (31) as,

$$c_{k,n}(t) = \frac{r_{awgn} \log_2(1 + g_{k,n}(t) \times SNR_{avg})}{\log_2(1 + SNR_{avg})}. \tag{32}$$

In all the simulations, we set $r_{awgn} = 1000$ kb/s.

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:

$$h_{k,n}(t) = (\kappa \times h_{k,n}(t-1) + v_{k,n}(t)) \times \sqrt{\frac{1-\kappa^2}{2}}, \tag{33}$$

where $v_{k,n}(t)$ are i.i.d. complex Gaussian variables with zero mean and unity variance per dimension. It is clear that (33) results in $E[g_{k,n}(t)] = E[|h_{k,n}(t)|^2] = 1$. 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 [20, page 165]

$$T_c \approx \frac{9}{16\pi f_m}, \tag{34}$$

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[5]

$$\kappa = 0.5^{T_s/T_c}. \tag{35}$$

In all the simulations, we set $\kappa = 0.8$, which roughly corresponds to a Doppler rate of 58 Hz.

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

4.2 Performance Evaluation

We organize this section as follows. In Section 4.2.1, we assess the accuracy of our QoS estimation (4). In Section 4.2.2, we evaluate the performance of our joint K&H/RR scheduler. In Section 4.2.3, we evaluate the performance of our RC scheduler.

[5]The auto-correlation function of the AR(1) process is κ^t, where t is the number of sample intervals. Solving $\kappa^{T_c/T_s} = 0.5$ for κ, we obtain (35).

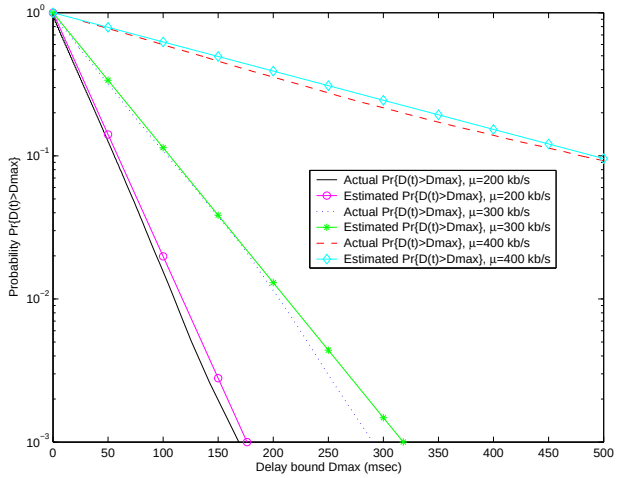


Figure 3: Actual and estimated delay-bound violation probability.

4.2.1 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.

In the experiments, the following parameters are fixed: $K = 40$, $N = 4$, and $SNR_{avg} = -40$ dB. By changing the source rate μ, we simulate three cases, *i.e.*, μ =200, 300, and 400 kb/s. Fig. 3 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 confirms the exponential dependence shown in (4). In addition, the estimated $\sup_t Pr\{D(t) > D_{max}\}$ is quite close to the actual $\sup_t Pr\{D(t) > D_{max}\}$, which demonstrates the effectiveness of our channel estimation algorithm.

4.2.2 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.

We set $SNR_{avg} = -40$ dB. The experiments use the optimum $\{\zeta, \beta\}$ values specified by the resource allocation algorithm, *i.e.*, Eqs. (6)–(9). 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 [23].

In Fig. 4, 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 [23]. In the case of joint scheduling, each point in the curve of $\theta(\mu)$ corresponds to a specific optimum $\{\zeta, \beta\}$, while

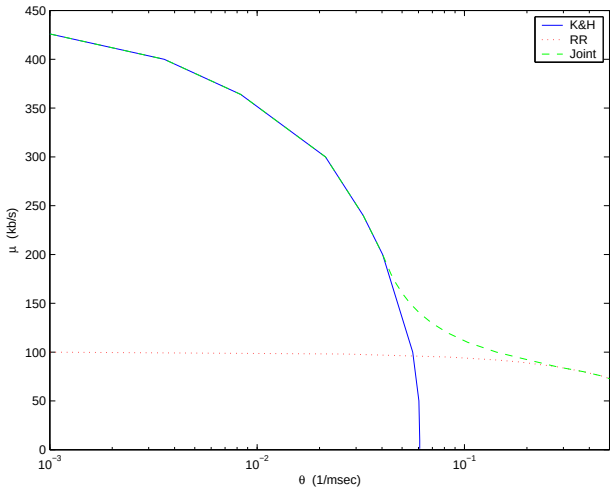


Figure 4: $\theta(\mu)$ vs. μ for K&H, RR, and joint scheduling ($K = 40, N = 4$).

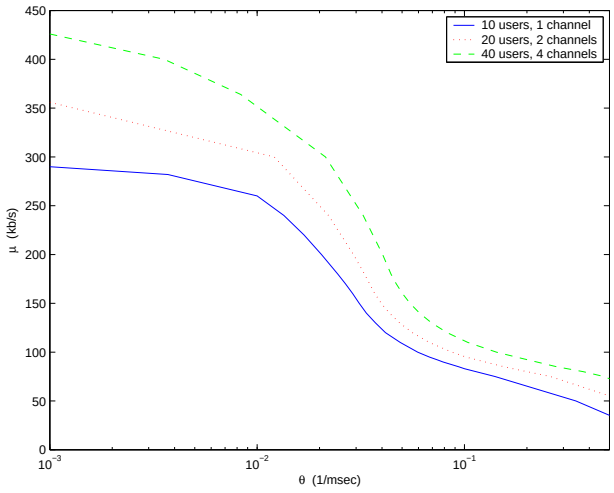


Figure 5: $\theta(\mu)$ vs. μ for joint K&H/RR scheduling.

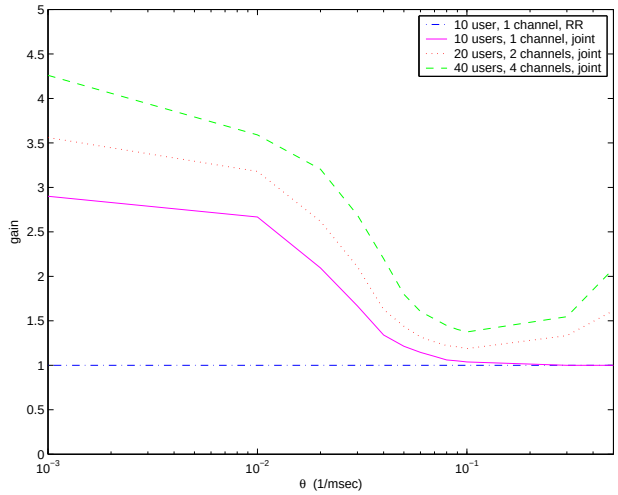


Figure 6: Gain for joint K&H/RR scheduling over RR scheduling.

$K\zeta = 1$ 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. 4.

Fig. 5 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.

On the other hand, using the RR scheduler for single channel as a benchmark, we plot the capacity gain achieved by the joint K&H/RR scheduler in Fig. 6. The capacity gain of the joint scheduler is the ratio of $\mu(\theta)$ of the joint scheduler to the $\mu(\theta)$ of the RR scheduler. For $N \geq 2$, the figure shows that 1) in the range of small θ, the capacity gain decreases with the increase of θ, which is due to the fact that multiuser diversity is less effective as θ increases, 2) in the range of large θ, the capacity gain increases with the increase of θ, which is due to the fact that the effect of frequency diversity kicks in as θ increases, 3) in the middle range of θ, the capacity gain is the least since both multiuser diversity and frequency diversity are less effective.

The simulation results in this section demonstrate that the joint K&H/RR scheduler can significantly increase the delay-constrained 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

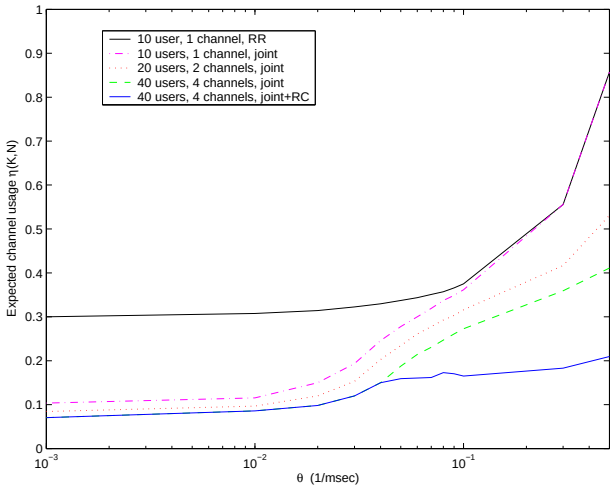


Figure 7: Expected channel usage $\eta(K, N)$ vs. θ.

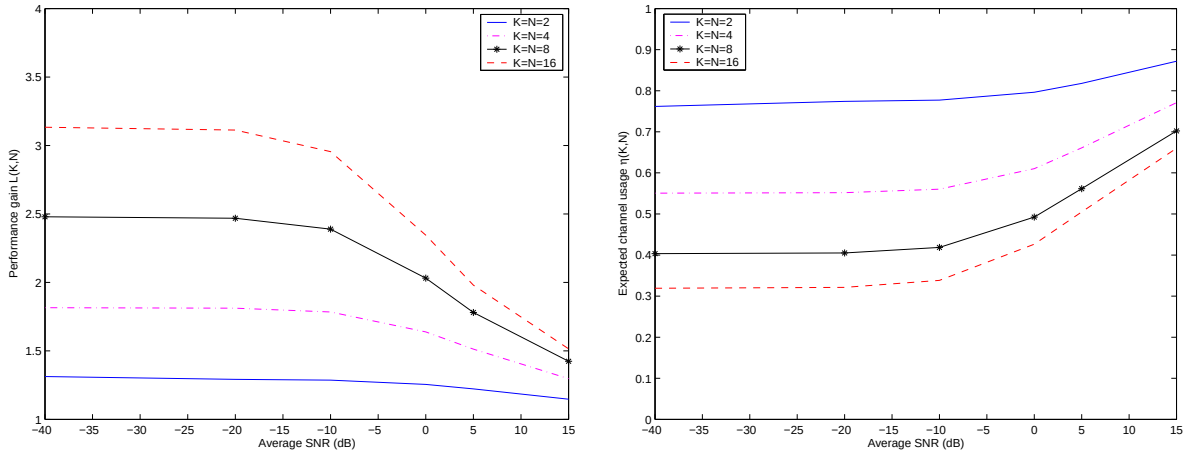
higher capacity gain than that for the single channel case.

4.2.3 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 $r_s^{(k)} = 30$ 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. Fig. 7 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 the RR scheduling or fixed channel assignment); 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 fixed channel assignment. In the experiments, we choose $\{r_s^{(k)}, \rho_k\}$ so that $\theta_{\zeta,\beta}(r_s^{(k)}) = \rho_k$, where $\zeta = 1/K$ and $\beta = 0$. Hence, the N channels are fully allocated to K users by the admission control, and we have fixed channel assignment $\xi_{k,n} = \zeta$, $\forall k, \forall n$. 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.



(a) (b)

Figure 8: (a) Performance gain $L(K,N)$ vs. average SNR, and (b) $\eta(K,N)$ vs. average SNR.

In the second scenario, we change the average SNR of the channels while fixing other source/channel parameters. Fig. 8(a) shows performance gain $L(K,N)$ vs. average SNR. Just as Proposition 3 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, Fig. 8(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 3). Fig. 8(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 9 shows the performance gain $L(K,N)$ versus number of channels N, for different average SNRs. It also shows the upper bound (27). 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 Fig. 10. The lower bound in Fig. 10 is computed by (26). One may notice that the gap between the bound and the actual metric in Figs. 9 and 10 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[6].

For all the simulations, we verify that the actual QoS achieved by the RC scheduler meets the users' requirements. The actual delay-bound violation probability curve is similar to that in Fig. 3 and is upper-bounded by the requested delay-bound violation probability.

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 [14, 15, 19], modified largest weighted delay first (M-LWDF) [1], opportunistic transmission scheduling [13] and lazy packet scheduling [18].

Wireless fair queueing schemes [14, 15, 19] are aimed at applying Fair Queueing [17] 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 [13] 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.

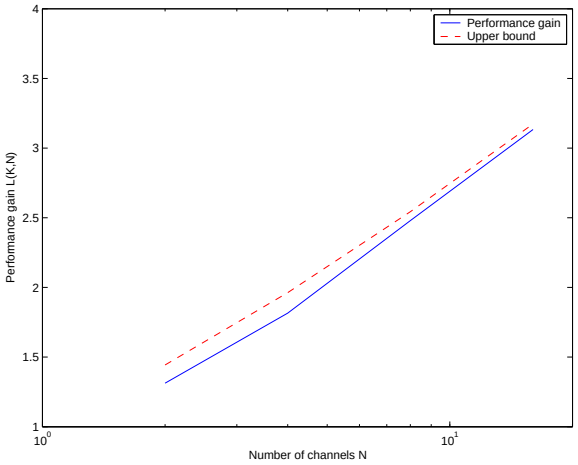
The lazy packet scheduling [18] is targeted at minimizing energy, subject to a delay constraint. The scheme only considers AWGN channels and thus allows for a deterministic delay bound, unlike fading channels and the general statistical QoS considered in our work.

Static fixed channel assignments, primarily in the wireline context, have been considered [10], in a multiuser, multichannel environment. However, these do not consider channel fading, or general QoS guarantees.

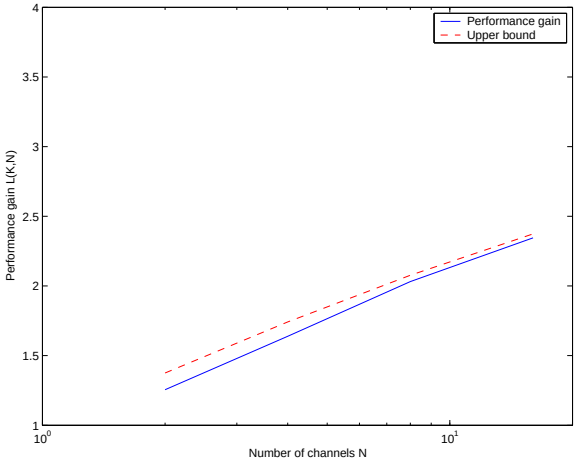
Time-division scheduling has been proposed for 3-G WCDMA [9, page 226]. The proposed time-division scheduling is similar to the RR scheduling in this paper. However, their proposal did not provide methods on how to use time-division scheduling to support statistical QoS guarantees explicitly. With the notion of effective capacity, we are able to make explicit QoS provisioning with our joint scheduling.

As mentioned in Section 3.2, 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

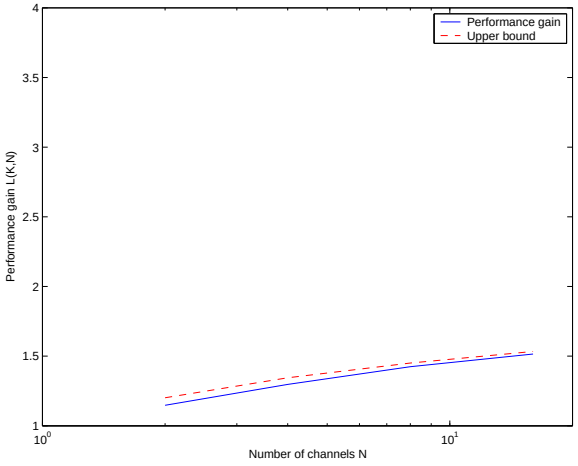
[6]In the proof of Proposition 2, we show that the bound corresponds to the case where each user chooses its best channel to transmit.



(a)

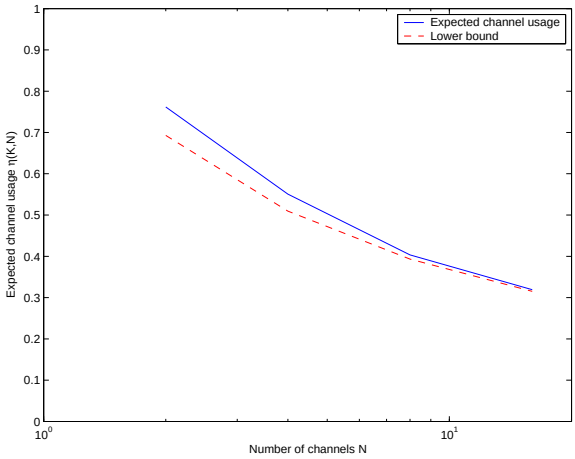


(b)

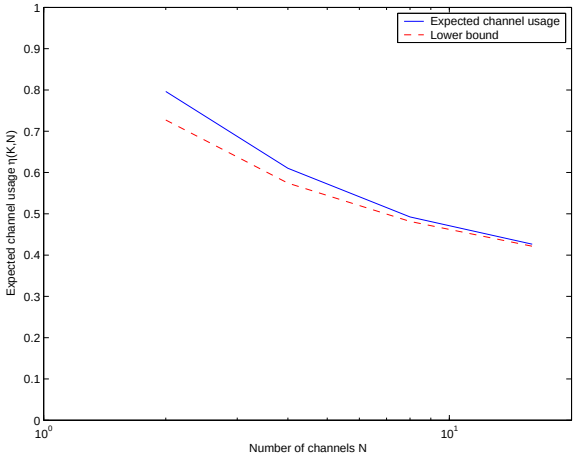


(c)

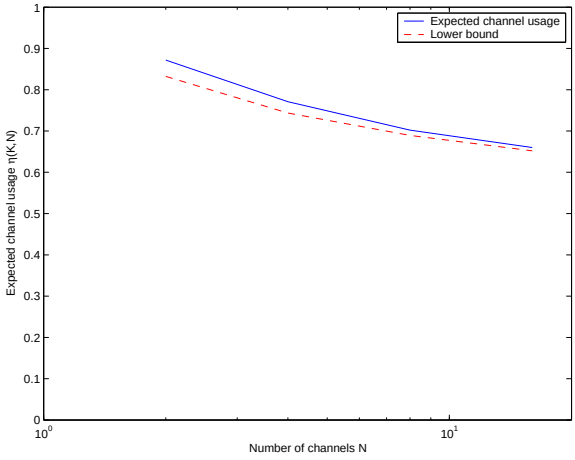
Figure 9: $L(K, N)$ vs. number of channels N for average SNR = (a) –40 dB, (b) 0 dB, and (c) 15 dB.



(a)



(b)



(c)

Figure 10: $\eta(K, N)$ vs. number of channels N for average SNR = (a) –40 dB, (b) 0 dB, and (c) 15 dB.

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 $\{\xi_{k,n}\}$. A scheduler (of which the RC scheduler is the optimum version) that allots the time-varying capacities specified by $\{\xi_{k,n}\}$, 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 $\{\xi_{k,n}\}$).

6 Concluding Remarks

With the increasing popularity of wireless networks, the issue of efficiently supporting QoS over scarce and shared wireless channels has come to the fore. 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 [23]. 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. The key features of the RC scheduler are,

- High efficiency. This is achieved by dynamically selecting the best channel to transmit.
- Simplicity. Dynamic programming is often required to provide an optimal solution to the scheduling problem. However, the high complexity of dynamic programming (exponential in the number of users) prevents it from being used in practical implementations. On the other hand, the RC scheduler has a low complexity (polynomial in the number of users), and yet performs very close to the bound we derived. This indicates that the RC scheduler is simple and efficient.
- Statistical QoS support. The RC scheduler is targeted at statistical QoS support. The statistical QoS requirements are represented by the channel assignments $\{\xi_{k,n}\}$, which appear in the constraints of the linear program of the scheduler.

Simulation results have demonstrated that substantial gain can be achieved by the joint K&H/RR scheduler and the RC scheduler, and have validated our analysis of the RC scheduler performance.

Our future work will focus on the design of admission control, resource allocation and K&H/RR scheduler, for the heterogeneous case, *i.e.*, different users have different QoS requirements and different channel statistics.

Appendix

Proof of Proposition 1: 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, (24) becomes

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

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 2: It is clear that the minimum value of the objective (16) under the constraint of (17) and (19) is a lower bound on that of (16) under the constraints of (17) through (19). The solution for (16), (17) and (19), 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.*,

$$w_{k,n}(t) = \frac{\sum_{m=1}^{N} \xi_{k,m} c_{k,m}(t)}{c_{k,n}(t)} \times \mathbf{1}(n = \bar{n}(k,t)), \qquad \forall k, \forall n \tag{36}$$

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 (16) through (19) as below,

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

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 (36) gives a lower bound, (c) since $c_{k,n}(t)$ are i.i.d. and stationary, and (d) due to (25). This completes the proof. ■

Expected channel usage $\eta(K,N)$ decreases as average SNR increases:
We first present a lemma and then prove Proposition 3.

Let $\gamma = P_0/\sigma^2$. Denote g_1 and g_2 channel power gains of two fading channels, respectively. Lemma 1 tells that for fixed channel gain ratio g_1/g_2, the corresponding capacity ratio $\log(1+\gamma \times g_2)/\log(1+\gamma \times g_1)$ monotonically increases from g_2/g_1 to 1, as average SNR γ increases from 0 to ∞.

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: We prove it by considering three cases: 1) $0 < \gamma < \infty$, 2) $\gamma = 0$, and 3) γ goes to ∞.

Case 1: $0 < \gamma < \infty$

Define $f(\gamma) = \log(1+\gamma g_2)/\log(1+\gamma g_1)$. To prove the lemma for Case 1, we only need to show $f'(\gamma) > 0$ for $\gamma > 0$. Taking the derivative results in

$$f'(\gamma) = \frac{\frac{g_2}{1+\gamma g_2}\log(1+\gamma g_1) - \frac{g_1}{1+\gamma g_1}\log(1+\gamma g_2)}{\log^2(1+\gamma g_1)} \tag{37}$$

Since $\log^2(1+\gamma g_1) > 0$ for $\gamma > 0$, we only need to show

$$\frac{g_2}{1+\gamma g_2}\log(1+\gamma g_1) > \frac{g_1}{1+\gamma g_1}\log(1+\gamma g_2) \tag{38}$$

or equivalently,

$$\frac{\frac{g_2}{1+\gamma g_2}\log(1+\gamma g_1)}{\frac{g_1}{1+\gamma g_1}\log(1+\gamma g_2)} = \frac{\frac{g_2}{(1+\gamma g_2)\log(1+\gamma g_2)}}{\frac{g_1}{(1+\gamma g_1)\log(1+\gamma g_1)}} > 1 \tag{39}$$

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 $0 < h(g_1) < h(g_2)$, *i.e.*, $h(g_2)/h(g_1) > 1$, which is the inequality in (39). So we only need to show $h'(x) < 0$ for $x > 0$. Taking the derivative, we have

$$h'(x) = \frac{\frac{1+\gamma x-\gamma x}{(1+\gamma x)^2}\log(1+\gamma x) - \frac{\gamma}{1+\gamma x}\frac{x}{1+\gamma x}}{\log^2(1+\gamma x)} \tag{40}$$

$$= \frac{\frac{1}{(1+\gamma x)^2}(\log(1+\gamma x) - \gamma x)}{\log^2(1+\gamma x)} \tag{41}$$

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

Case 2: $\gamma = 0$

$$\lim_{\gamma \to 0} f(\gamma) \overset{(a)}{=} \lim_{\gamma \to 0} \frac{\frac{g_2}{1+\gamma g_2}}{\frac{g_1}{1+\gamma g_1}} = \frac{g_2}{g_1} \tag{42}$$

where (a) is from L'Hospital's rule.

Case 3: γ goes to ∞

$$\lim_{\gamma\to\infty} f(\gamma) \stackrel{(a)}{=} \lim_{\gamma\to\infty} \frac{\frac{g_2}{1+\gamma g_2}}{\frac{g_1}{1+\gamma g_1}} \tag{43}$$

$$= \lim_{\gamma\to\infty} \frac{g_2(1+\gamma g_1)}{g_1(1+\gamma g_2)} \tag{44}$$

$$\stackrel{(b)}{=} \frac{g_2}{g_1} \times \frac{g_1}{g_2} \tag{45}$$

$$= 1 \tag{46}$$

where (a) and (b) are from L'Hospital's rule.

Combining Cases 1 to 3, we complete the proof.■

Next, we prove Proposition 3.

Proof of Proposition 3: From the definition of c_{max}, we have

$$\begin{aligned} c_{max} &= \max_{n\in\{1,2,\cdots,N\}} c_{k,n} \\ &= \max_{n\in\{1,2,\cdots,N\}} \log(1+\gamma g_{k,n}) \\ &= \log(1+\gamma g_{max}) \end{aligned} \tag{47}$$

where $g_{max} = \max_{n\in\{1,2,\cdots,N\}} g_{k,n}$. Also, from the definition of c_{mean}, we get

$$\begin{aligned} c_{mean} &= \frac{1}{N}\sum_{n=1}^{N} c_{k,n} \\ &= \frac{1}{N}\sum_{n=1}^{N} \log(1+\gamma g_{k,n}) \\ &= \log(1+\gamma g_{mean}) \end{aligned} \tag{48}$$

where

$$g_{mean} = \frac{1}{\gamma}\left(\prod_{n=1}^{N}(1+\gamma g_{k,n})^{1/N} - 1\right) \tag{49}$$

It is obvious that $g_{max} > g_{mean} > 0$. So from Lemma 1, we have $\log(1+\gamma\times g_{mean})/\log(1+\gamma\times g_{max})$, *i.e.*, c_{mean}/c_{max}, monotonically increases from g_{mean}/g_{max} to 1, as γ increases from 0 to ∞. Hence, $\mathbf{E}[c_{mean}/c_{max}]$ monotonically increases from $\mathbf{E}[g_{mean}/g_{max}]$ to 1, as γ increases from 0 to ∞.■

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] D. Bertsekas, *Dynamic Programming and Optimal Control, Vol. 1, 2,* Athena Scientific, 1995.

[3] 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.

[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] B. E. Collins and R. L. Cruz, "Transmission policies for time-varying channels with average delay constraints," in *Proc. 1999 Allerton Conference on Communication, Control, and Computing,* Monticello, IL., USA, Sept. 1999.

[6] M. Elaoud and P. Ramanathan, "Adaptive allocation of CDMA resources for network-level QoS assurances," in *Proc. ACM Mobicom'00,* Aug. 2000.

[7] 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.

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

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

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

[11] Y. Y. Kim and S.-Q. Li, "Capturing important statistics of a fading/shadowing channel for network performance analysis," *IEEE Journal on Selected Areas in Communications,* vol. 17, no. 5, pp. 888–901, May 1999.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] J. Nocedal and S. J. Wright, *Numerical optimization,* Springer, 1999.

[17] 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.

[18] 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.

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

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

[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.

[23] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *Technical Report,* Carnegie Mellon U., Aug. 2002. Available at `http://www.cs.cmu.edu/~dpwu/publications.html`.

[24] D. Wu, "Effective capacity approach to providing statistical quality-of-service guarantees in wireless networks," *Ph.D. thesis proposal,* Carnegie Mellon U., Sept. 2002.

[25] H. Zhang, "Service disciplines for guaranteed performance service in packet-switching networks," *Proceedings of the IEEE,* vol. 83, no. 10, Oct. 1995.

[26] Z.-L. Zhang, Z. Duan, and Y. T. Hou, "Virtual time reference system: a unifying scheduling framework for scalable support of guaranteed services," *IEEE Journal on Selected Areas in Communications,* vol. 18, no. 12, pp. 2684–2695, Dec. 2000.