Measurement-Based Opportunistic Scheduling for Heterogenous Wireless Systems

Shailesh Patil and Gustavo de Veciana, *Senior Member, IEEE*

***Abstract*—We study the performance of an opportunistic scheduling scheme *maximum quantile scheduling*, i.e., scheduling a user whose current rate is in the highest quantile relative to its current rate distribution, in a wireless system. In a practical scenario it is unlikely that users' rate distributions are known at the scheduler, and have to be estimated via measurement. Under the assumption of fast fading, we prove a bound on the relative penalty associated with such estimates, showing that number of independent samples need only grow linearly with the *number of active users*. This is a fairly limited cost, suggesting one could track distributional changes in users' channels. By contrast other opportunistic scheduling schemes require estimating or setting weights/thresholds that implicitly depend not only on the number of users, but also *their rate distributions, and possibly their traffic characteristics*. In other words the penalty associated with tuning weights for other schemes can be higher than that associated with estimating users' rate distributions for maximum quantile scheduling. This statement is supported by our simulation results. Furthermore we prove that if rates are bounded and number of users is high, maximum quantile scheduling is sum average throughput maximizing subject to temporal fairness.**

***Index Terms*—Scheduling, resource management, stochastic majorization, cellular systems, wireless access.**

I. Introduction

MOTIVATION. The scheduling of users' data transmissions at a wireless access point has recently attracted a substantial amount of attention, e.g., [1]. A key feature of wireless systems relative to the traditional wireline systems is that, the channel capacity, or service rate, may exhibit temporal variations. This allows one to consider scheduling policies that choose to send to, or receive from, a user which at a given point in time has the 'best', e.g., highest, capacity. Such 'opportunistic scheduling' can lead to good increases in the aggregate capacity of a wireless systems, such as CDMA-HDR, HSDPA [5].

In practice users' channel capacity variations are unknown and heterogenous, e.g., users close to an access point see significantly different channel capacity than those further off. Thus it is important to devise opportunistic schedulers that do not starve some users, e.g., those with poor channels, to achieve some degree of fairness among users sharing an access point. To this end many opportunistic scheduling schemes have been devised that make decisions by selecting the user that currently has the highest weighted channel capacity. In practice the weights may be hard to determine, because they depend in a complex way on the users' channel capacity distributions, the number of users, and the characteristics of their traffic. Thus they either need to be estimated or tuned based on the service users have received.

Unfortunately, the complex dependence of weights may make them very sensitive to changes in the system, i.e., if a user's traffic characteristics changes, or a user leaves or enters the system (e.g., a mobile user comes out of the shadow of a building), or the channel characteristics of a user change, then the weights associated with *all* users may need to change. Therefore, it is likely that a significant fraction of time will be spent in estimating/tuning weights to their 'ideal' values. In fact, if the system is dynamic enough and/or the tuning algorithm is not sensitive enough, one may never converge, leading to poor throughput performance. Consider a simple example. Due to the stochastic or time varying nature of channel capacity and user's traffic a measurement-based opportunistic scheduler may be biased in favor of a user who has not received service in the recent past or one that currently has a high queue. While, this myopic approach is good for short term fairness, the scheduler may end up serving a user even though it is not currently experiencing a high channel rate. This in turn decreases the achieved opportunism and long term throughput the system can sustain. In heavily loaded systems, at a given moment of time, it is very likely that there exists a group of users which are starved. If those users are served, others may become starved, leading to a cycle, in which the level of opportunism and throughput are low.

Recently, distribution based opportunistic schedulers have been proposed by several researchers under different guises [8][6][9]. In this paper, we shall refer to this scheme as *maximum quantile scheduler*. The idea is to schedule a user whose current rate is highest relative to his *own* distribution, i.e., in the highest quantile. This allows the scheme to exploit opportunism and achieve fairness without weights, however, instead one needs to estimate each user's channel capacity distribution. In this paper we will show that the throughput penalty incurred from estimating users' distributions is limited.

Contributions. Following are the contributions of this paper:

- We first show that if the achievable instantaneous rate of users' is bounded, then among the class of scheduling policies that serve each user an equal fraction of time, maximum quantile scheduling maximizes the long term

Paper approved by T.-S. P. Yum, the Editor for Packet Access and Switching of the IEEE Communications Society. Manuscript received February 21, 2008; revised July 21, 2008 and October 27, 2008.

S. Patil is with Qualcomm Flarion Technologies, 500 Somerset Corporate Blvd., Bridgewater NJ 08807, USA, (e-mail: patil@qualcomm.com.).

G. de Veciana is with Wireless Networking & Communications Group, Dept. of Electrical & Computer Engineering, University of Texas at Austin, 1 University Station C0803 Austin, TX 78712, USA, (e-mail: gustavo@ece.utexas.edu).

This research was supported in part by National Science Foundation Award CNS-0721532.

Digital Object Identifier 10.1109/TCOMM.2009.080900

system throughput when there is a large number of users. Furthermore, we show that the marginal distribution for the rate when users are selected for service under maximum quantile scheduling can not be stochastically dominated by any other non-idling scheduler.

- Under the assumption of fast fading, we prove a bound on each user's relative throughput penalty when maximum quantile scheduling is based on empirical estimates of users' channel capacity distribution. The bound shows that such penalties can be controlled if the number of independent samples used to estimate the empirical distribution is roughly proportional to the *number of users in the system*.
- Using simulations, we compare the performance of maximum quantile scheduling and other opportunistic scheduling schemes when weights/distributions are estimated via measurement. We find that maximum quantile scheduling can have significantly better performance in terms of throughput penalty and file transfer delay, e.g., up to 30% improvement. In other words, the penalty associated with estimating distributions can be lower compared to that associated with estimating weights.

Paper Organization. In Section II we discuss some prior work in the area of opportunistic scheduling and introduce some known features of maximum quantile scheduling. Throughput performance and optimality of the scheme is studied in Section III. We prove our bound on the relative throughput penalty associated with measuring distributions in Section IV. Simulation results comparing the performance of maximum quantile scheduling to other schemes are presented in Section V. Section VI concludes the paper.

II. Revisiting Opportunistic Scheduling

A. System Model and Notation

We begin by introducing our system model and some notations. For simplicity, we focus on downlink scheduling from an access point to multiple users. Suppose time is divided into equal sized slots, e.g., CDMA-HDR systems have a slot duration of 1.67 ms [5]. During each slot, all users feedback the data rate they can support and the access point decides to serve at most one user in the slot. In the sequel we use the terms 'channel capacity' and 'rate' interchangeably. For analysis purposes, we make the following assumptions on users' channel capacity distribution(s) across.

Assumption 2.1: We assume the channel capacity (rate) for each user is a stationary ergodic process and these processes are independent across users. Further we assume that the marginal rate distribution function for each user is continuous, increasing, and is known a priori at the access point.

Discussion on the assumption. First the rate distributions seen by users might indeed be roughly stationary and independent across user. The assumption that the access point knows the marginal distributions of the channel capacity processes may seem unreasonable, but simple book keeping of the users' current rate feedback can be used to estimate distributions. This will be discussed in more detail in Section IV. Note that channel capacities are not restricted to any specific distribution, i.e., users can undergo any fading process. This makes the analysis presented applicable to real world scenarios. Further note that we require the marginal distribution function of rates to be continuous and increasing only for simplicity sake, the results presented here can be extended to the discrete, non increasing case also (see [7]).

System Scenario. Unless specified otherwise, we will mostly focus on the 'fixed saturated' case, where the number of users in the system does not change with time and each user is infinitely backlogged. Such a scenario is an approximation where the number of users in the system changes slowly and packet queues for each user are always non empty at the access point. This idealization is often studied in literature.

Notation. In the sequel we will let $x^i(t)$ denote the realization of the rate of user i at time slot t, and let X^i be a random variable whose distribution is that of the rate of user i on a *typical* slot. Recall that we will be assuming X^i to be independent across users but need not be identically distributed. We denote the distribution function of X^i by $F_{X^i}(\cdot)$. Note by assumption $F_{X^i}(\cdot)$ is an increasing continuous function, this allows us to have an inverse $F^{-1}_{X^i}(\cdot)$ defined. The number of users in the system is given by n.

B. Previous Work

The first opportunistic scheduling *maximum rate scheduling* was first proposed in [1]. Here the user with maximum current rate is served, i.e., user $k(t)$ is selected for service on time slot t if

$$k(t) \in \arg \max_{i=1,\ldots,n} x^i(t).$$

This maximizes system throughput in a fixed saturated system, but in a system where users have heterogeneous rate distributions, may neglect those with poor channels.

A myriad of approaches have been proposed to address both unfairness/performance issues. A widely cited scheme is *proportional fair scheduling* [2][3] which serves the user whose current rate normalized by a moving average of his allocated rate is the highest, i.e., user $k(t)$ is served during time slot t if

$$k(t) \in \arg \max_{i \in i=1,\ldots,n} \frac{x^i(t)}{\mu^i(t)}, \tag{1}$$

where

$$\mu^i(t) = (1 - \frac{1}{t_c})\mu^i(t-1) + \frac{1}{t_c} x^i(t) \mathbf{1}_{S^i_{pf}(t)}$$

and t_c is the moving average parameter, $S^i_{pf}(t)$ is the event that user i gets served on slot t by the scheme, and $\mathbf{1}_{S^i_{pf}(t)}$ is the indicator function of $S^i_{pf}(t)$.

More recently, [4] proposed strategies that maximize sum throughput under fairness constraints. For example, they show that a scheduling policy of the form

$$k(t) \in \arg \max_{i=1,\ldots,n} [x^i(t) + \nu^i], \tag{2}$$

maximizes the sum throughput subject to constraints on the fraction of time each user i is served in a fixed saturated regime. Here ν^i is a weight associated with user i that ensures that users get served the desired fraction of time. Similar optimal schemes were proposed for rate and utility based fairness.

While the optimality characteristics of these schemes are desirable, in practice they would require estimating thresholds ν^i which are complicated functions of users' requirement and rate distributions. In Section V, we show that such estimates may converge slowly, leading to loss in performance.

C. Maximum Quantile Scheduling

Maximum quantile scheduling has been independently proposed by [8] as a 'CDF based scheme', while [6] proposed a 'score based scheduler', and [9] proposed a 'distribution fairness' based scheduler.

Let us briefly introduce this scheme in the fixed saturated regime. The main idea is to schedule a user whose rate is highest compared to his *own* distribution, i.e., serve user $k(t)$ during slot t if

$$k(t) \in \arg \max_{i=1,\ldots,n} F_{X^i}(x^i(t)). \tag{3}$$

It is well known that $F_{X^i}(X^i)$ is uniformly distributed on $[0,1]$. Let $U^i := F_{X^i}(X^i)$, then U^i is also uniformly distributed on $[0,1]$. Under Assumption 2.1, maximum quantile can be thought of as picking the maximum among independent realizations of users' (i.i.d.) U^i's on every slot. Thus, maximum quantile is equally likely to serve any user on a typical slot, and all users get served an equal fraction, i.e., $\frac{1}{n}^{th}$ of time. Furthermore, let $U^{(n)} := \max[U^1, \ldots, U^n]$, then again from Assumption 2.1

$$\Pr(U^{(n)} \leq u) = u^n, \ \forall u \in [0,1], \tag{4}$$

and the rate distribution seen by user i on a slot that it gets served is the same as $F^{-1}_{X^i}(U^{(n)})$. Therefore, the average throughput seen by user i is given by $G^i_{mq}(n)$[8],

$$G^i_{mq}(n) = \frac{E[F^{-1}_{X^i}(U^{(n)})]}{n} = \frac{E[X^{i,(n)}]}{n},$$

where $X^{i,(n)}$ is maximum of n i.i.d. copies of X^i, i.e., $X^{i,(n)} := \max[X^i_1, \ldots, X^i_n]$, where $X^i_j \sim X^i$, $\forall j = 1, \ldots, n$. Note that by contrast, even if users' rate distributions were known, it is not easy to evaluate the individual and system throughput for other schemes discussed in the previous subsection.

Maximum quantile scheduling has several desirable properties and simulation results show that it has good throughput performance (see [8][7] for details). However, as discussed earlier, it is important to understand performance penalty associated with estimating users' rate distributions.

III. Performance of Maximum Quantile Scheduling in Fixed Saturated System

In this section we study the performance of maximum quantile scheduling in terms of the amount of opportunism exploited, and the throughput achieved by the scheme. Due to lack of space, we omit the proofs for theorems presented in this section, the reader can refer to Chapter 2 of [7] for details.

'Opportunistically' Optimal. Suppose we consider as measure of opportunism achieved by user i as the quantile of the rate achieved by the user, i.e., $F_{X^i}(x^i(t))$ whenever it is served. A high quantile means a high degree of opportunism and $E[\sum_{i=1}^n F_{X^i}(X^i)\mathbf{1}_{S^i_\beta}]$ denotes the overall expected opportunism realized by a scheduling scheme β. (Here S^i_β is the event that user i is selected for service on typical slot by β.) It should be clear that maximum quantile scheduling maximizes the system opportunism.

Not Stochastically Dominated. Maximum quantile scheduling has an optimality in terms of the rates seen by users in the typical slots in which they are served. Let us first introduce the concept of stochastic dominance, we say that a random variable Y stochastically dominates random variable V, if $\forall v$, $\Pr(Y > v) \geq \Pr(V > v)$, this is denoted as $Y \geq^{st} V$. Let R^i_{mq} represent the rate distribution seen by user i when selected for service on a typical slot by maximum quantile scheduling, and let $\overrightarrow{R}_{mq} = (R^1_{mq}, \ldots, R^n_{mq})$, i.e., the vector of random variables representing the rate distributions. Let $\overrightarrow{R}_\beta = (R^1_\beta, \ldots, R^n_\beta)$ be the same quantity for another distinct non idling scheduling scheme β that may *not* serve all users an equal fraction of time. By distinct we mean that the scheme does not always pick the same user as maximum quantile, and by non idling, we mean that the scheme will never choose to *not serve* a user in the slot. Then our claim (formally stated below) is that $\overrightarrow{R}_\beta \not\geq^{st} \overrightarrow{R}_{mq}$, i.e., $\exists j = 1, \ldots, n$, such that $R^j_\beta \not\geq^{st} R^j_{mq}$.

Theorem 3.1: Consider a fixed saturated system with n users, whose channel capacity variations satisfy Assumption 2.1. Then for any distinct non idling scheduling scheme β, $\overrightarrow{R}_\beta \not\geq^{st} \overrightarrow{R}_{mq}$.

Note that a scheduling scheme γ is known to be Pareto optimal if there exists no other scheduling scheme that is able to give an equal or higher average throughput to *all* the users than that received by users under γ. Theorem 3.1 can be thought to be a weak form of Pareto optimality in terms of rate seen in a typical slot. We now show that maximum quantile is not Pareto optimal in terms of average throughput.

Not Pareto Optimal. We illustrate this with a simple two user system with ON-OFF channels. (The example can be extended to the continuous case.) The ON and OFF channel states correspond to rates 1 and 0 respectively. User 1 and 2 have an ON probability of 0.6 and 0.4 respectively. Here maximum quantile will serve User 1 a rate of 0.42, and User 2 a rate of 0.32. However, it can be shown that maximum quantile may sometimes serve User 2 in OFF state, even though User 1's channel is ON. Therefore, it is possible to improve performance while still serving each user an equal fraction of time. Consider a scheme that always serves the user with the highest instantaneous rate and breaks ties $\frac{7}{24}^{th}$ of times in favor of User 1. Such a scheme will give User 1 a rate of 0.43, and User 2 will get a rate of 0.33. Hence one can give better performance to both the users, while maintaining temporal fairness.

Throughput Optimal for Large Number of Users. Even though maximum quantile is not Pareto optimal, it does achieve good system throughput performance. If the rates achievable by users in a system are bounded, then maximum quantile scheduling is sum throughput optimal among policies that serve all users an equal fraction of time as the number of users increases. Our claim is formally stated below.

Theorem 3.2: Consider a n user fixed saturated system using maximum quantile scheduling. Suppose each user has a finite maximum instantaneous rate. Then under Assumption 2.1 as $n \to \infty$, each user is likely to be served at its maximum rate, so maximum quantile scheduling is sum throughput optimal among scheduling policies that serve all users equally.

Summarizing, even though maximum quantile is not Pareto optimal, it gives good throughput.

IV. Penalty due to Estimating Distributions

Let us study the throughput penalty incurred by maximum quantile scheduling due to estimation of rate distributions of users. Recall that Assumption 2.1 required the rate distribution functions, i.e., $F_{X^i}(\cdot)$ of each user be known at the access point. This is unlikely, however suppose the quantile of the current rate of a user is estimated using the previous m samples of the user's rate. The empirical distribution of user i during slot t based on m previous samples is denoted by $\tilde{F}^{m,t}_{X^i}(\cdot)$ and is given by

$$\tilde{F}^{m,t}_{X^i}(x) = \frac{1}{m}\sum_{j=1}^{m} 1\{X^i(t-j) \le x\}. \quad (5)$$

Note that the above way of estimating is similar to the score function described in [6], however there, no attempt was made to theoretically evaluate the throughput penalty due to incorrect estimation.

Thus maximum quantile scheduling of users based on estimated distributions, would choose user $k(t)$ for service during slot t if

$$k(t) \in \arg\max_{i=1,\ldots,n} \tilde{F}^{m,t}_{X^i}(x^i(t)),$$

with ties being broken arbitrarily. It can be shown that for any user on any slot t, $\tilde{F}^{m,t}_{X^i}(X^i(t))$ is uniformly distributed on $\{0, \frac{1}{m} \ldots, 1\}$. Therefore, it is easy to see that even with estimated distributions, maximum quantile scheduling will still serve each user an equal fraction of time.

Calculating the penalty due to estimation seems to be intractable under slow fading, *therefore we add an additional assumption of fast fading, i.e., rate realizations of a user in a slot is independent across slots.* Even though this is usually not true, independence of samples can be roughly ensured by taking samples that are sufficiently apart in time or using some scheme, e.g. 'opportunistic beamforming' [3].

We now calculate the average throughput achieved by users under maximum quantile scheduling based on estimated distributions. Since we are interested in the stationary behavior, we simplify notation from $\tilde{F}^{m,t}_{X^i}(\cdot)$ to $\tilde{F}^{m}_{X^i}(\cdot)$. Following theorem gives the performance of this scheme, see Theorem 2.4.1 in [7] for proof.

Theorem 4.1: Consider a n user fixed saturated system using maximum quantile scheduling, where the rate distributions in such a system are estimated via (5) based on m *independent* samples of a user's channel. Then under Assumption 2.1, the average throughput achieved by user k is given by

$$\tilde{G}^k_{mq}(n,m) = \frac{E[F^{-1}_{X^k}(\tilde{U}_{n,m})]}{n},$$

where $\tilde{U}_{n,m}$ is a continuous r.v. on $[0,1]$ having a probability density function

$$f_{\tilde{U}_{n,m}}(u) = \sum_{j=0}^{m} \binom{m}{j} u^j (1-u)^{m-j} \frac{((j+1)^n - j^n)}{(m+1)^{n-1}}. \quad (6)$$

Recall that R^i_{mq} represents the rate distribution seen by user i when selected for service on a typical slot by maximum quantile scheduling (with perfect distribution knowledge). Let $\tilde{R}^{i,m}_{mq}$ denote the same quantity for maximum quantile scheduling when distributions are estimated using m independent samples.

We show that $\tilde{R}^{i,m}_{mq}$ and R^i_{mq} are 'closely related' random variables, i.e., the rate seen by a user when served under empirical distributions is similar to that seen when distributions are perfectly known. This is used to show that the average throughput of a user when empirical distributions are used is less than or equal to that achieved when distributions are perfectly known, i.e., $\tilde{G}^k_{mq}(n,m) \le G^k_{mq}(n)$ and bound the relative throughput penalty due to estimation. Our result is stated below, the proof is given in Appendix A.

Theorem 4.2: Consider a fixed saturated system with n users using maximum quantile scheduling. Then under Assumption 2.1 and fast fading $\forall n, m$,

$$(\frac{m+1}{n}(1-(\frac{m}{m+1})^n)) \le \frac{\Pr(\tilde{R}^{i,m}_{mq} \le r)}{\Pr(R^i_{mq} \le r)} \le 1, \ \forall r,$$

and

$$G^k_{mq}(n) \ge \tilde{G}^k_{mq}(n,m), \ \forall m,$$

and the relative throughput penalty is bounded by

$$\frac{|G^k_{mq}(n) - \tilde{G}^k_{mq}(n,m)|}{G^k_{mq}(n)} \le 1 - \frac{m+1}{n}(1-(\frac{m}{m+1})^n).$$

To understand the scaling of the number of independent samples m required to limit the throughput penalty, note that for a reasonably large n, if m scales linearly with n, then $(\frac{m}{m+1})^n = (1+\frac{1}{m})^{-n} \approx e^{-\frac{n}{m}}$. Expanding $e^{-\frac{n}{m}}$ and simplifying, we get that the relative throughput penalty is equal to

$$1 - \frac{m+1}{m} + \frac{m+1}{n}(\frac{1}{2}(\frac{n}{m})^2 - \ldots),$$

which is upper bounded by $\frac{n}{2m}$. Then to achieve a relative error less than ϵ, at most $\frac{n}{2\epsilon}$ samples are needed, e.g., to achieve an error less than 5%, at most $10n$ samples are needed. In other words for a given error bound, the number of samples required will at worst grow linearly with the number of users.

Simulations. To validate these results, we ran some simulations. Our setup consists of two classes of users having a mean signal to noise ratio (SNR) of 2 and 0.1, with both classes experiencing Rayleigh fading and containing an equal number of users. The channel capacity for all users is fast fading, i.e., rate supported by users is independent across slots, and the slot size is set to 1.67 msec. The bandwidth associated with each user is 500 KHz and we assume that coding achieves the Shannon rate. Unless specified otherwise, this setup will be used throughout the paper for simulations.

We first observed the throughput penalty for different values of n and m. The value of n is varied from 8 to 16 to 32,

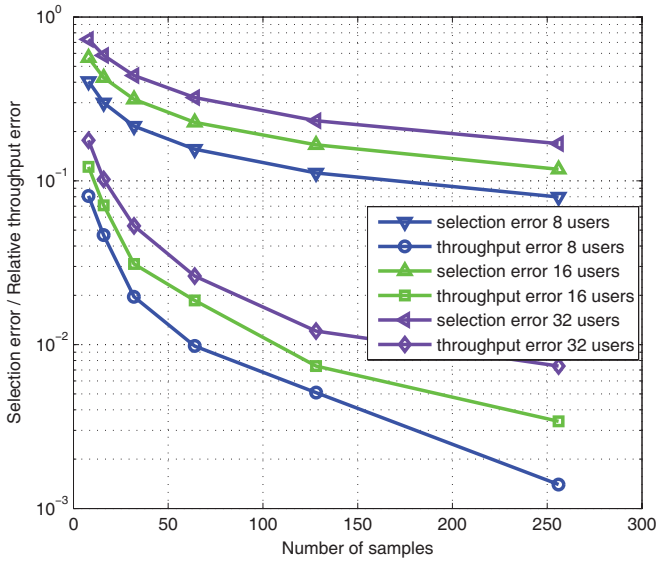

Fig. 1. The top three curves plot the selection error probability for maximum quantile scheduling, due to estimated distributions with increasing number of users. The bottom three curves plot the relative throughput penalty for the same.

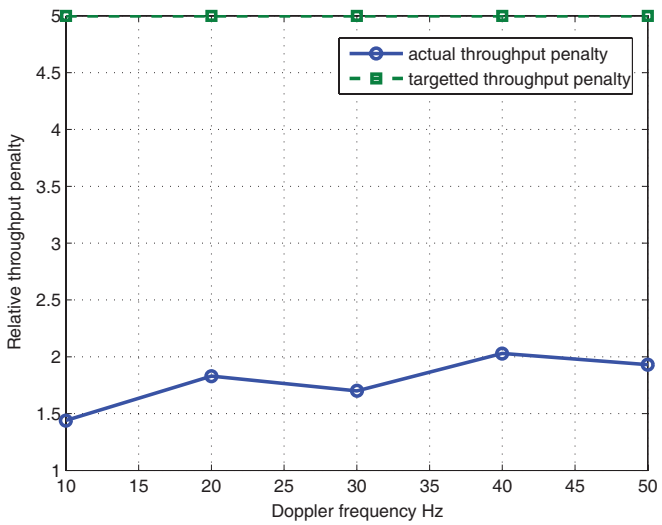

Fig. 2. Relative throughput penalty for 10 users with slow Rayleigh fading channel capacities.

while m is varied by a factor of 2 from 8 to 256 for a given value of n. As shown in Figure 1, the bound is clearly met, in fact the results indicate that our bound is quite conservative (which is not surprising, since the bound is distribution free). For example, a relative throughput penalty of around 1% is achieved with only 64 samples for 8 users, whereas the bound suggests 5%.

We also plot the selection error probability in the figure, i.e., the fraction of slots where the user selected with maximum quantile is *not* chosen due to error in estimation of distribution. As the plot indicates, this can be quite high. Our analysis (not included in this paper) shows that the number of samples required to achieve a given error probability grows roughly as $O(n^2)$. Therefore, even though mistakes may be made in selecting the user with the highest quantile, the throughput penalty in making an error is not large.

Let us consider the bound under slow fading now. The need for m independent samples immediately suggests the need for sampling m coherence time intervals to achieve the required penalty. We ran simulations to confirm this conjecture. The simulation consisted of two (earlier described) classes of slow Rayleigh fading users with 5 users each, we aimed for a penalty of 5%. The Doppler spread for the channels was varied from 10 Hz to 50 Hz in steps of 10 Hz. Let f_D denote the Doppler spread, then the coherence time can be estimated using the formula $\frac{9}{16\pi f_D}$ [10]. Given the coherence time, the number of slots needed to estimate the rate distributions can be ascertained. The simulation results are plotted in Figure 2, which shows that the required penalty is met in all cases. Note that in our simulations we found that for Doppler spread of 10 Hz, 932 slots were needed. This corresponds to 1.55 seconds (slot size is 1.67 msec), it may be reasonable to expect the system to be stationary for such a period because the Doppler spread is quite low, i.e., users/objects are moving quite slowly. In other words, even though very slowly fading systems may require a large number of samples to achieve the desired penalty, it may also be reasonable to expect such channels to be stationary over large periods of time.

Discussion of the bound. Theorem 4.2 has several interesting implications, which we discuss below.

- The bounds shows that the relative throughput penalty due to estimation of users' distribution can be bounded for i.i.d. samples of *any* distribution.
- The theorem is strong in the sense that it shows a relationship between distributions (and not just the average) of rates seen by the user in both the empirical and perfectly known distribution cases.
- The number of samples needed to achieve small penalty is *only linear in the number of users*. This is limited (at least for the fast fading case) because slot sizes are usually milliseconds long.
- The dependence of penalty on the number of users is significant. Even if the number of users are changing with time, to achieve a certain penalty, a system designer only needs to estimate the 'average' number of users that will be competing for service. We reiterate here that this is unlike other weight based schemes which are dependent on users' distribution or traffic characteristics.
- The dependence on only the number of users also allows us to conjecture that if users' channel are stationary for roughly $O(n^2)$ slots (under fast fading), then the desired penalty will be met.

Summarizing, maximum quantile scheduling under estimated distribution case is not only fair, suffers from fairly limited penalty, but is quite easy to design for and to implement.

V. Simulation Based Comparison with Weight Based Schemes

Let us now compare the performance of maximum quantile scheduling to other schemes via simulations. First we compare the maximum sum throughput scheme described by (2), with maximum quantile scheduling when weights and distributions have to be estimated. Next we modify the setup to consider the case where the number of users dynamically vary with time, and observe the time of file transfer.

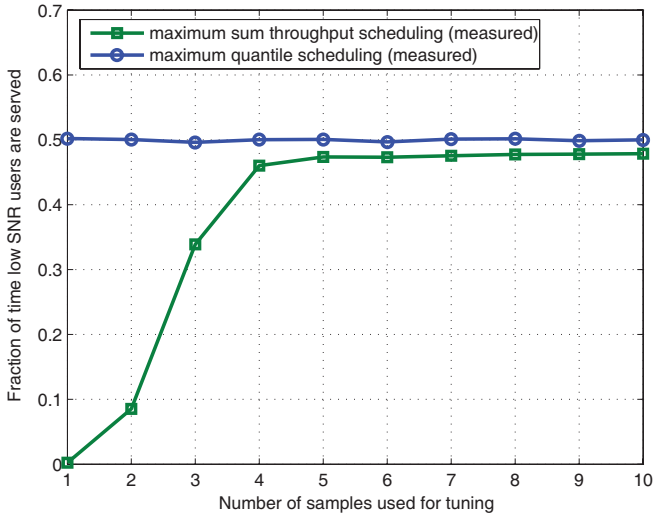

Fig. 3. Fraction of time low SNR users are served by measurement based maximum sum throughput and maximum quantile scheduling schemes, with increasing number of tuning samples.

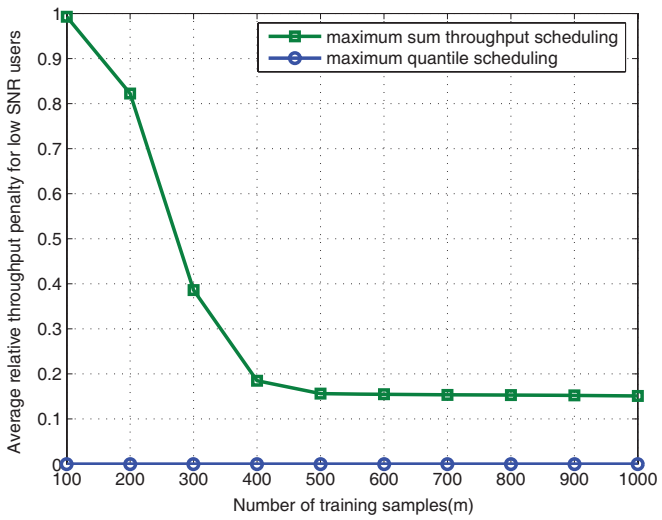

Fig. 4. Average relative throughput penalty incurred by the class of low SNR users for increasing number of tuning samples.

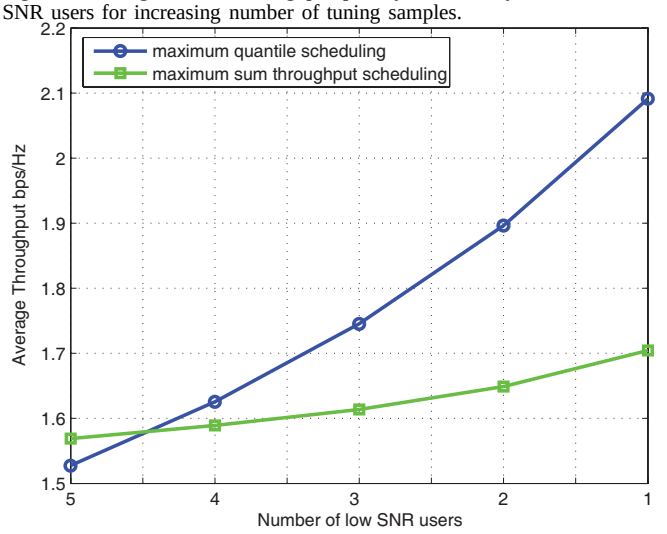

Fig. 5. Throughput achieved by maximum sum throughput under temporal fairness and maximum quantile scheduling with decreasing number of low SNR users.

A. Throughput Penalty Comparisons

Recall that if the users' weight ν^i are properly set in (2), then the scheme maximizes sum throughput under temporal fairness. However in practice the weights for each user needs to be estimated, we investigate the sensitivity of system throughput to errors in these weights by performing two controlled experiments.

In the first experiment, there are 5 users in each class (with the previously described setup), and the weights ν^i for all users are initialized to 0. We train the weights for m slots according to the stochastic approximation algorithm suggested in [4], and observe the average penalty in performance due to errors in weights on the $(m+1)^{st}$ slot. We refer the reader to [4] for details on the training algorithm. We evaluate two performance parameters, the fraction of time low SNR users are served, and the relative penalty in throughput achieved by those users as compared to that achieved when weights are perfectly known.

The algorithm for estimating the ν^i's has parameters (w, δ, δ_i) that need to be set, we first set these parameters equal to those suggested in [4]. However, the scheduling scheme served the users with low average SNR less than 0.1% of time even with $m = 2000$ (which demonstrates the difficulty in setting measurement based weights). Hence we changed the parameters to $w = 0.005$, $\delta = 0.2$ and $\delta_i = 0.1$.

Figure 3 shows the fraction of time low average SNR users are served as an increasing number of training samples m is used. We also plot the corresponding results for maximum quantile scheduling, which always serves low average SNR users close to 0.5 fraction of time. By contrast, maximum sum throughput takes around 400 samples to converge to approximately 0.47 and then shows negligible improvement. This is because the granularity of training is not sufficiently small, however as suggested in the previous paragraph, if one reduces these updates, then the convergence time may be much larger. Figure 4 plots the throughput penalty for the low average SNR users with training samples m. While the throughput penalty is virtually 0 under maximum quantile scheduling, there is penalty of 15% even for $m = 1000$ training samples. Note that a 3% loss in temporal fairness can lead to a 15% loss in throughput.

Second experiment is a sensitivity assessment of maximum sum throughput's performance. Suppose there are 5 users in each class, and estimates for ν^i's have converged. If a user leaves, the ideal values of weights would change. However if the scheme does not immediately tune ν^i a throughput penalty is incurred. To asses these penalties we simulated a scenario where weights ν^i had converged for the case where there were 5 users in each class, and then we dropped the number of users having low average SNR to 4, 3, 2, 1. Figure 5 shows the throughput achieved by maximum sum throughput and maximum quantile scheduling for the scenarios. Note that maximum quantile scheduling starts doing better as soon as the number of low average SNR users goes from 5 to 4, i.e., maximum sum throughput no longer remains optimal. We observed a similar trend when the high average SNR users were reduced. In summary, the sensitivity of throughput to the optimal weights is relatively high, and changes in the numbers of active users can lead to reduced throughput. By contrast maximum quantile scheduling does not require

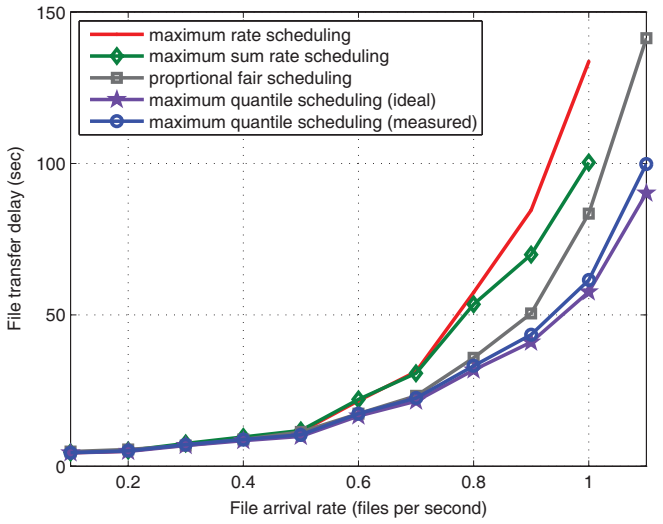

Fig. 6. File transfer delay performance of maximum quantile, maximum rate and maximum sum throughput scheduling.

estimating weights, and estimation of the users' distribution has low impact on the overall performance.

B. Performance Comparison with Varying Number of Users

Let us compare the performance of maximum quantile to maximum rate, proportional fair and maximum sum throughput under temporal constraints in a system where the number of users vary with time. One can think of this as users with files coming to an access point and leaving once their files have been transferred. Here instead of throughput, a good metric for performance is the average file transfer delay.

Again our setup is the same as before, except that the number of users will change with time. Users arrive to the system according to a Poison process, and are equally likely to belong to one of the two classes. Each user has a file associated with it. The file sizes are exponentially distributed with a mean size of 60KB. We keep track of the time taken from a user's arrival to departure. For maximum quantile, estimate for users rate distributions are generated by keeping track of previous samples. For an arrival rate of 0.1, 50 samples are used. This is increased linearly by 50 samples for every arrival rate increase of 0.1. While the weights for maximum sum throughput are trained using the stochastic approximation algorithm referred to in Subsection V-A, and the values of the parameters same as before.

The average file transfer delay experienced by users is plotted with increasing average arrival rate in Figure 6. Maximum quantile outperforms all other schemes, in fact the reduction in delay is almost 30% (as compared to proportional fair) at an arrival rate of 1.1. Also note that due to non convergence of weights, maximum rate and maximum sum throughput up to a load of 0.8 have quite similar performance. Therefore maximum sum throughput can easily degrade to maximum rate in a dynamic scenario. We also plot the delay experienced by users under maximum quantile with perfect rate distribution knowledge. Observe that measurement based performance is close to ideal performance.

VI. Conclusion

In this paper we evaluated measurement based maximum quantile scheduling. The key take away is that, perhaps surprisingly, maximum quantile scheduling which would require estimation of each users channel rate distribution, is not only robust to estimation errors, and can give an excellent performance relative to other weight based schemes. However it remains to be seen whether the gains over simple schemes such as proportionally fair are justified by additional complexity needed to estimate distributions.

Appendix A
Proof of Theorem 4.2

We present a few useful lemmas before proving Theorem 4.2.

Lemma A.1: Let H be a binomial random variable with parameters (m,u). Consider the moment generating function of H, $M(s) := (1-u+ue^s)^m$. Its l^{th} derivative is given by

$$\frac{d^l M(s)}{ds^l} = \sum_{j=1}^{l} b_{j,l} \frac{m!}{(m-j)!}(1-u+ue^s)^{m-j}(ue^s)^j. \quad (7)$$

Here $b_{j,l}$'s are constants with the following properties:

- $b_{1,1} = 1$
- $b_{j,l} = jb_{j,l-1} + b_{j-1,l-1}$, $\forall j = 1,\ldots,l$, $\forall l$
- $b_{0,l} = b_{l+1,l} = 0$, $\forall l$.

Note that since $b_{1,1} = 1$ and $b_{l+1,l} = 0$, $\forall l$, from the second property we get that $b_{l,l} = b_{l-1,l-1} = 1$, $\forall l$.

Proof: We give a proof by induction on l. The lemma holds for $l = 1$. Assume the lemma holds for l. Then, to prove the lemma for $l+1$, we differentiate (7) and after some rearrangement get

$$\frac{d^{l+1} M(s)}{ds^{l+1}} = \sum_{j=1}^{l+1} [(jb_{j,l} + b_{j-1,l}) \frac{m!}{(m-j)!}(1-u+ue^s)^{m-j}(ue^s)^j].$$

This completes the proof. ∎

From Lemma A.1 it follows that the l^{th} order moment of H is given by

$$E[H^l] = \sum_{j=1}^{l} b_{j,l} \frac{m!}{(m-j)!} u^j. \quad (8)$$

The following lemma exhibits an inequality between the moments of H.

Lemma A.2: Let H be a binomial r.v. with parameters (m,u). Then for all l such that $l \leq m$,

$$E[H^{l+1}] \leq (mu + l(1-u))E[H^l]. \quad (9)$$

Proof: The right side of (9) can be expressed as $((m-l)u+l)E[H^l]$. Using (8), we get

$$\frac{m!}{(m-l-1)!}u^{l+1} + \sum_{j=1}^{l} [lb_{j,l}\frac{m!}{(m-j)!} + (m-l)b_{j-1,l}\frac{m!}{(m-j+1)!}]u^j. \quad (10)$$

If one splits $lb_{j,l}\frac{m!}{(m-j)!} = jb_{j,l}\frac{m!}{(m-j)!} + (l-j)b_{j,l}\frac{m!}{(m-j)!}$, then (10) in turn can be expressed as

$$\frac{m!}{(m-l-1)!}u^{l+1} + \sum_{j=1}^{l}[(jb_{j,l}\frac{m!}{(m-j)!} + (m-l)b_{j-1,l}$$
$$\frac{m!}{(m-j+1)!})u^j + (l-j+1)b_{j-1,l}\frac{m!}{(m-j+1)!}u^{j-1}].$$

Now since $0 \le u \le 1$, then $\forall j$, $u^{j-1} \ge u^j$. So, from the above equation we get

$$(mu + l(1-u))E[H^l] \ge \frac{m!}{(m-l-1)!}u^{l+1} +$$
$$\sum_{j=1}^{l}[(jb_{j,l}\frac{m!}{(m-j)!} + (m-l)b_{j-1,l}\frac{m!}{(m-j+1)!})$$
$$+(l-j+1)b_{j-1,l}\frac{m!}{(m-j+1)!}]u^j.$$

Combining the last two terms in the summation of the above inequality, we get

$$(mu + l(1-u))E[H^l] \ge$$
$$\frac{m!}{(m-l-1)!}u^{l+1} + \sum_{j=1}^{l}(jb_{j,l} + b_{j-1,l})\frac{m!}{(m-j)!}u^j$$

This proves (9). ∎

Next we show that $U^{(n)}$ dominates $\tilde{U}_{n,m}$ in a likelihood ratio ordering sense, i.e., $U^{(n)} \ge^{lr} \tilde{U}_{n,m}$ [11][12]. This is a strong form of dominance which means that $f_{U^{(n)}}(u)/f_{\tilde{U}_{n,m}}(u)$ is non decreasing in u, or $f_{\tilde{U}_{n,m}}(u)/f_{U^{(n)}}(u)$ is non increasing in u (here $f_{U^{(n)}}(u)$ is the probability density function of $U^{(n)}$). If $U^{(n)} \ge^{lr} \tilde{U}_{n,m}$, it follows that $U^{(n)} \ge^{st} \tilde{U}_{n,m}$.

Lemma A.3: For the random variables $U^{(n)}$ and $\tilde{U}_{n,m}$ given by (4) and (6) respectively, then $\forall n, m$ $U^{(n)} \ge^{lr} \tilde{U}_{n,m}$.

Proof: To prove the lemma, we need to show

$$\frac{d}{du}\left[\frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)}\right] \le 0,$$

$\forall u \in (0,1]$. To prove this, it is sufficient to show

$$f_{U^{(n)}}(u)\left[\frac{df_{\tilde{U}_{n,m}}(u)}{du}\right] - f_{\tilde{U}_{n,m}}(u)\left[\frac{df_{U^{(n)}}(u)}{du}\right] \le 0.$$

Note that $f_{U^{(n)}}(u) = nu^{n-1}$. Then expanding, we get

$$\frac{1}{(m+1)^{n-1}}[nu^{n-1}(-m(1-u)^{m-1} +$$
$$\sum_{j=1}^{m-1}\binom{m}{j}u^{j-1}(1-u)^{m-j-1}(j-mu)((j+1)^n - j^n) +$$
$$mu^{m-1}((m+1)^n - m^n)) - n(n-1)u^{n-2}$$
$$(\sum_{j=0}^{m}\binom{m}{j}u^j(1-u)^{m-j}((j+1)^n - j^n))] \le 0.$$

Simplifying and multiplying both sides by $(1-u)$, we get

$$(-mu(1-u)^m +$$
$$\sum_{j=1}^{m-1}\binom{m}{j}(j-mu)u^j(1-u)^{m-j}((j+1)^n - j^n) +$$
$$(m-mu)u^m((m+1)^n - m^n)) - (n-1)(1-u)$$
$$(\sum_{j=0}^{m}\binom{m}{j}u^j(1-u)^{m-j}((j+1)^n - j^n)) \le 0.$$

The above inequality can be rewritten as

$$\sum_{j=0}^{m}\binom{m}{j}(j-mu-(n-1)(1-u))u^j(1-u)^{m-j}$$
$$((j+1)^n - j^n) \le 0.$$

Then the inequality clearly holds for $m < n$. However the more interesting case is when $m \ge n$, and this requires a few more steps. Note that $\binom{m}{j}u^j(1-u)^{m-j}$ is the probability that a binomial r.v. with parameter (m,u) has a value j, i.e., the same as that of H. Then the inequality can be rewritten in terms of expectations as

$$E[(H-mu)((H+1)^n - H^n)] -$$
$$(n-1)(1-u)E[(H+1)^n - H^n] \le 0.$$

This can be further rewritten as

$$E[H((H+1)^n - H^n)] \le$$
$$(mu + (n-1)(1-u))E[(H+1)^n - H^n]. \quad (11)$$

Expanding $(H+1)^n$ and simplifying, one can show that (11) will hold if

$$E[H^{l+1}] \le (mu + l(1-u))E[H^l],$$

$\forall l < n \le m$. This follows from Lemma A.2. This completes the proof. ∎

We now prove Theorem 4.2.

Proof: To prove the first claim, define $u := F_{X^i}(r)$ and consider

$$F_{U^{(n)}}(u) - F_{\tilde{U}_{n,m}}(u), \ \forall u \in (0,1].$$

This is equivalent to

$$\int_0^u (f_{U^{(n)}}(u) - f_{\tilde{U}_{n,m}}(u))du.$$

Which in turn is equivalent to

$$\int_0^u f_{U^{(n)}}(u)(1 - \frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)})du.$$

Then

$$F_{U^{(n)}}(u) - F_{\tilde{U}_{n,m}}(u) \le \int_0^u f_{U^{(n)}}(u)\max_u(1 - \frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)})du.$$

Note from Lemma A.3,

$$\min_u \frac{f_{\tilde{U}_{n,m}}(u)}{f_{U^{(n)}}(u)} = \frac{f_{\tilde{U}_{n,m}}(1)}{f_{U^{(n)}}(1)} = \frac{m+1}{n}(1 - (\frac{m}{m+1})^n).$$

Then

$$F_{U^{(n)}}(u) - F_{\tilde{U}_{n,m}}(u) \leq F_{U^{(n)}}(u)(1 - \frac{m+1}{n}(1 - (\frac{m}{m+1})^n)).$$

Simplifying, one gets

$$F_{U^{(n)}}(u)(\frac{m+1}{n}(1 - (\frac{m}{m+1})^n)) \leq F_{\tilde{U}_{n,m}}(u).$$

Now from Lemma A.3, it follows that $U^{(n)} \geq^{st} \tilde{U}_{n,m}$, combining this with the above equation we get

$$\frac{m+1}{n}(1 - (\frac{m}{m+1})^n) \leq \frac{F_{\tilde{U}_{n,m}}(u)}{F_{U^{(n)}}(u)} \leq 1.$$

Using the definition of u, and the fact that $F_{X^i}(\cdot)$ is an increasing function, the above equation can be rewritten as

$$\frac{m+1}{n}(1 - (\frac{m}{m+1})^n) \leq \frac{\Pr(F^{-1}_{X^i}(\tilde{U}_{n,m}) \leq r)}{\Pr(F^{-1}_{X^i}(U^{(n)}) \leq r)} \leq 1.$$

Note that $R^i_{mq} = F^{-1}_{X^i}(U^{(n)})$ and $\tilde{R}^{i,m}_{mq} = F^{-1}_{X^i}((\tilde{U}_{n,m})$, then the above equation can be written as

$$\frac{m+1}{n}(1 - (\frac{m}{m+1})^n) \leq \frac{\Pr(\tilde{R}^{i,m}_{mq} \leq r)}{\Pr(R^i_{mq} \leq r)} \leq 1.$$

To prove the second claim, recall that $G^k_{mq}(n) = \frac{E[F^{-1}_{X^k}(U^{(n)})]}{n}$. Note that $F^{-1}_{X^k}(\cdot)$ is an increasing function. Therefore it is sufficient to prove that $U^{(n)} \geq^{st} \tilde{U}_{n,m}$ to prove the theorem, which is shown to be true from Lemma A.3.

We now prove the third part of the theorem. Note from the second part of the theorem, it is suffices to study

$$\frac{G^k_{mq}(n) - \tilde{G}^k_{mq}(n,m)}{G^k_{mq}(n)}.$$

Consider the difference between the two throughput, i.e., $E[F^{-1}_{X^k}(U^{(n)})] - E[F^{-1}_{X^k}(\tilde{U}_{n,m})]$. The difference can be expressed as

$$\int_0^1 F^{-1}_{X^k}(u) f_{U^{(n)}}(u) du - \int_0^1 F^{-1}_{X^k}(u) f_{\tilde{U}_{n,m}}(u) du.$$

Then following the methodology used in the first part of the proof one can show

$$E[F^{-1}_{X^k}(U^{(n)})] - E[F^{-1}_{X^k}(\tilde{U}_{n,m})] \leq$$
$$\int_0^1 F^{-1}_{X^k}(u) f_{U^{(n)}}(u)(1 - \frac{m+1}{n}(1 - (\frac{m}{m+1})^n)) du,$$

or

$$E[F^{-1}_{X^k}(U^{(n)})] - E[F^{-1}_{X^k}(\tilde{U}_{n,m})] \leq$$
$$E[F^{-1}_{X^k}(U^{(n)})](1 - \frac{m+1}{n}(1 - (\frac{m}{m+1})^n)).$$

This completes the proof. ■

REFERENCES

[1] R. Knopp and P. Humblet, "Information capacity and power control in single cell multi-user communications," *Proc. IEEE International Computer Conf.*, vol. 1, pp. 331-335, June 1995.

[2] A. Jalali, R. Padovani and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," in *Proc. Veh. Technol. Conf., 2000. VTC 2000-Spring Tokyo*, vol. 3, pp. 1854-1858, May 2000.

[3] P. Viswanath, D. Tse and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, pp. 1277-1294, June 2002.

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

[5] P. Bender, P. Black, M. Grob and R. Padovani and N. Sindhushayana and A. Viterbi, "CDMA-HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, pp. 70-77, July 2000.

[6] Thomas Bonald, "A score-based opportunistic scheduler for fading radio channels," *Proc. European Wireless*, 2004.

[7] S. Patil, "Opportunistic scheduling and resource allocation among heterogeneous users in wireless networks, Ph.D. thesis, Univeristy of Texas at Austin," [Online]. Available: http://users.ece.utexas.edu/ğustavo/papers/phd/Shailesh.Patil2006.pdf May 2006.

[8] D. Park, H. Seo and H. Kwon and B. G. Lee, "A new wireless packet scheduling algorithm based on the CDF of user transmission rates," in *Proc. IEEE Globecom*, pp. 528-532, Nov. 2003.

[9] X. Qin and R. Berry, "Opportunistic splitting algorithms for wireless networks with heterogeneous users," in *Proc. Conf. Inform. Sciences Systems (CISS)*, Mar. 2004.

[10] T. S. Rappaport, *Wireless Communications, Principles and Practice*, Pearson Education, 2002.

[11] A. Marshall and I. Olkin, *Inequalities: Theory of Marjorization and Its Applications*, Academic Press, 1979.

[12] S. M. Ross, *Stochastic Processes*, John Wiley, 1983.

Shailesh Patil received his Bachelor in Electronics & Communications Engineering from Delhi University, India in 2001, and M.S. and Ph.D. both in Electrical & Computer Engineering from University of Texas at Austin in 2004 and 2006 respectively. His research interests include developing MAC level algorithms in wireless networks. He is a recipient of Texas Telecommunications Engineering Consortium (TxTEC) Fellowship in 2002. He is currently with Qualcomm Flarion Technologies.

Gustavo de Veciana (S'88-M'94-SM'2001) is the Temple Foundation Professor in Electrical and Computer Engineering at the University of Texas at Austin. He received his B.S., M.S, and Ph.D. in electrical engineering from the University of California at Berkeley in 1987, 1990, and 1993 respectively. His research focuses on the design, analysis and control of telecommunication networks. Current interests include: measurement, modeling and performance evaluation; wireless and sensor networks. Dr. de Veciana has been an editor for the IEEE/ACM Transactions on Networking. He is the recipient of a 1996 NSF CAREER Award, co-recipient of the IEEE William McCalla Best ICCAD Paper Award for 2000, and co-recipient of the Best Paper in ACM Transactions on Design Automation of Electronic Systems, Jan 2002-2004.