Delay-Sensitive Packet Scheduling in Wireless Networks

Peijuan Liu, Randall Berry, Michael L. Honig
ECE Department, Northwestern University
2145 Sheridan Road, Evanston, IL 60208 USA
{peijuan,rberry,mh}@ece.nwu.edu

***Abstract*— We consider "opportunistic" downlink scheduling of data traffic in a wireless network. In particular, we focus on the delay performance of such schedulers. First a channel-dependent scheduling algorithm is considered which maximizes throughput by always transmitting to the user with the best channel conditions. The delay distribution of this scheduling rule is analyzed and asymptotic results are given when the number of competing users becomes large. Simulations show these asymptotic results are a good approximation for even a small number of users. This scheduling rule may result in unfair treatment of users that have relative bad channels for a long period of time; to remedy this we propose a simple utility-based scheduling algorithm. The motivation is to maximize the time-averaged utility, where utility is a decreasing function of the delay incurred when serving a request. The scheduling algorithm takes into account both the utility function and the channel state. We give simulation results which characterize the performance of the scheduling algorithm. The effect of the temporal correlation of the channel on the performance is also studied.**

I. INTRODUCTION

In this paper we consider downlink wireless data scheduling. For delay tolerant applications, scheduling algorithms can exploit channel variations across the user population and attempt to transmit to users when they have "good" channel conditions. Such "opportunistic" scheduling algorithms have received much attention recently (e.g., [14],[15]) and are part of most third generation wireless standards.

Much of the prior work on scheduling has focused on throughput or fairness for users with "elastic" traffic. In this paper, we consider the delay performance of opportunistic schedulers. First, we consider a simple scheduling rule (the Max R Rule) that transmits to the user with the best channel at any time. It is easily shown that the Max R Rule maximizes throughput given a constraint on the total downlink transmission power. We derive the resulting delay distribution for this scheduling rule; in particular, we consider a system where the number of competing users becomes large and characterize the asymptotic delay distribution. Although the Max R Rule is very efficient in utilizing the limited radio resources, it is biased against users that remain in an unfavorable channel condition; such users may be blocked for a long time and experience indefinite delay. This unfairness presents a significant problem especially in a low-tier mobility environment. To overcome this, we introduce utility function that is expressed in terms of delay. The longer the delay, the lower the derived utility. Using this utility, we propose a simple utility-based scheduler (the U'R Rule), which attempts to maximize the time-average utility. The scheduler takes into account both the channel conditions and the utility received by each user. The utility function indicates the sense of "urgency" of a request, while the channel state influences the efficiency of resource utilization given that the request is served.

A sampling of other work that has addressed downlink scheduling for wireless networks includes [3], [4], [5], [11]. In all of this work, scheduling algorithms attempt to exploit the "multiuser diversity" that is present in a fading environment. The authors in [3], [4], [5] consider scheduling schemes that give preferences to users with favorable channel conditions where the channel is modeled as a simple two-state process with a "good" and "bad" channel state. This simplification, although it provides insight, often becomes inadequate in realistic systems. More refined scheduling rules, which make use of combined knowledge of the queue length, waiting time and channel conditions are presented in [7], [8], [9], [10]. Fairness is an important issue for scheduling in general. For example, it is often desirable to provide worst-case guarantees on throughput and delay and to achieve some degree of separation between flows. The authors of [11] propose a framework for achieving this end by emulating the generalized processor sharing (GPS) model ([12], [13]) proposed for wireline allocations. In our work, we also consider scheduling based on waiting time and channel conditions. However, our work differs from the previous work in that we adopt a utility-based approach. Maximizing the utility rate automatically trades-off fairness for throughput, where the nature of the tradeoff is embedded in the definition of the utility function.

The rest of the paper is organized as follows. In Sect. II, we describe a downlink single-cell model. In Sect. III, we consider the Max R Rule and analyze the delay performance. In Sect. IV, we propose the U'R rule which makes use of both the channel condition and the utility function. We also present simulation results that show how the temporal correlation of the channel affects the performance of different scheduling rules. Conclusions are given in the last section.

This work was supported by the Motorola-Northwestern Center for Telecommunications, and by NSF under grant CCR-9903055.

II. SYSTEM MODEL

We consider the downlink of a single-cell using Time Division Multiplexing (TDM). We assume that there are K users with packets to transmit. A scheduler, which resides in the base station (or base station controller), decides at the start of each scheduling interval which of the Head of Line (HOL) packets to serve. We assume that the base station transmits to one user at a time with the full available power.

The packets are assumed to have fixed length L, where L is small enough so that the channel stays relatively constant during a packet transmission. Each packet transmission time is therefore $T = L/R$, where R is the rate at which data can be transmitted to the scheduled user with very small probability of error. The rate R depends on the channel gain h and the specific coding and modulation scheme chosen for a user. We assume that R increases with h, and their relation is given by $R = \mathcal{C}(hP)$, where P is the total available downlink transmission power, and $\mathcal{C}(\cdot)$ is an increasing function. We assume that R (or equivalently h) is available to the scheduler.

III. THE MAX R RULE

A. Throughput Optimization

In this section, to simplify our analysis we assume that once the base station starts transmitting to user i, it must complete the transmission of one packet at rate R_i. Upon completion, a new decision is made as to which is the next scheduled packet. A scheduling interval ("time slot") is therefore the packet service time, which varies in duration according to the available rates. The amount of data served in each time slot, L, stays fixed. This differs from systems such as HDR ([1], [2]), where the time slots are of fixed duration (1.67ms) and the packets served in each slot may differ in size.*

Let $R_{i,j}$ be the transmission rate at which user i could receive data if scheduled in time slot j. We define *the Max R Rule*, to be the scheduling policy which serves user i^* given by:

$$i^* = \arg \max_{i \in \{1,\dots,K\}} R_{i,j}, \quad \text{for all} \quad j. \tag{1}$$

where any ties can be broken arbitrarily. It is easily seen that the Max R Rule maximizes the total system throughput.

Next, we study the delay distribution resulting from this rule and show that the mean delay grows at rate $\Theta(K/\ln(K))$ † as $K \to \infty$.

B. Delay Distribution Analysis

To simplify our analysis we consider a system with K backlogged queues. In such a system, delay is calculated from the time a packet arrives at the head of the queue. We derive the delay distribution across packets at the output of the scheduler.

Let $\tilde{R}_j = \max_{i\in\{1,\dots,K\}} R_{i,j}$. The duration of the j^{th} slot when using the Max R rule is then $\tilde{T}_j = \min_{i\in\{1,\dots,K\}} T_{i,j} = L/\tilde{R}_j$, where $T_{i,j} = L/R_{i,j}$. We assume that $\{R_{i,j}\}$ is a set of independent and identically distributed (*i.i.d.*) random variables with *p.d.f.* $f_R(r)$. Therefore, $\{T_{i,j}\}$ are also *i.i.d.* and have *p.d.f.* $f_T(t) = f_R(\frac{L}{t})\frac{L}{t^2}$. Here we assume that the independence holds not only across users, but over time slots as well. Conditioned on a packet getting scheduled in the N^{th} slot after it becomes the HOL packet, the total delay experienced by this packet, D, is given by

$$D = \sum_{j=1}^{N} \tilde{T}_j, \tag{2}$$

where N is a random variable denoting the number of time slots required until a packet gets scheduled. Since the $R_{i,j}$'s are *i.i.d.*, the *p.m.f.* of N is given by:

$$p_n = \Pr[N = n] = \frac{1}{K}\left(\frac{K-1}{K}\right)^{n-1}, \text{ for } n = 1, \dots, \infty. \tag{3}$$

This is a geometric distribution with parameter $1/K$.

The packet delay seen at the output of the scheduler, D, is therefore the sum of a random number (N) of *i.i.d.* random variables ($\tilde{T}_j$). We point out that N is independent from $\tilde{T}_j$. We want to derive $f_D(x)$, the *p.d.f.* of the total delay D.

Let $F_{\tilde{T}}(x)$ and $f_{\tilde{T}}(x)$ denote the *c.d.f.* and *p.d.f.* of $\tilde{T}_j$, respectively. Let $D_n = \sum_{j=1}^{n} \tilde{T}_j$ be the sum of n independent $\tilde{T}_j$'s. The *p.d.f.* of D_n, $f_{D_n}(x)$, is given by the i-fold self convolution of $f_{\tilde{T}}(x)$:

$$f_{D_n}(x) = \underbrace{f_{\tilde{T}}(x) \star f_{\tilde{T}}(x) \star \cdots \star f_{\tilde{T}}(x)}_{n}. \tag{4}$$

Using (2) and (3), the *p.d.f.* of D can be written in terms of $f_{D_n}(x)$:

$$f_D(x) = \sum_{n=1}^{\infty} f_{D_n}(x) p_n \tag{5}$$

Hence, the moments of D are given by:

$$\begin{aligned} \mathbb{E}(D^k) &= \sum_{n=1}^{\infty} \mathbb{E}(D^k \mid N = n) p_n \\ &= \sum_{n=1}^{\infty} \mathbb{E}(D_n^k) p_n. \end{aligned} \tag{6}$$

In particular,

$$\mathbb{E}(D) = \sum_{n=1}^{\infty} n\mathbb{E}(\tilde{T}_j) p_n = K\mathbb{E}(\tilde{T}_j), \tag{7}$$

and

$$\begin{aligned} \mathbb{E}(D^2) &= \sum_{n=1}^{\infty} \left(n\mathbb{E}(\tilde{T}_j^2) + n(n-1)\mathbb{E}^2(\tilde{T}_j) \right) p_n \\ &= K\mathbb{E}(\tilde{T}_j^2) + (2K^2 - 2K)\mathbb{E}^2(\tilde{T}_j). \end{aligned} \tag{8}$$

*This analysis can be generalized to the case of fixed duration time-slots at the expense of more complicated notation.

†We use the notation $f(n) = \Theta(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} < \infty$ and $\lim_{n\to\infty} \frac{g(n)}{f(n)} < \infty$, for $f(n)$ and $g(n) > 0$, $\forall n$.

Higher moments of D can be expressed similarly in terms of moments of $\tilde{T}_j$.

Now we only need to specify $f_{\tilde{T}_j}(x)$ to complete the analysis of the delay distribution. Recall that $\tilde{T}_j$ is the minimum of K *i.i.d.* random variables; hence its *c.d.f.* is given by ([18], [19]):

$$F_{\tilde{T}_j,K}(x) = \Pr\left[\tilde{T}_j \le x\right] = 1 - [1 - F_T(x)]^K, \qquad (9)$$

where $F_T(x) = \int_0^x f_T(t)\,dt$ is the *c.d.f.* for $T_{i,j}$.

So far, we have completely characterized the delay distribution given a distribution on the channel rates, $R_{i,j}$. However, this computation becomes complex as K becomes large. The distribution for the duration of one time slot ($F_{\tilde{T}_j,K}(x)$) contains K as an exponent, therefore it becomes increasingly difficult to evaluate moments of $\tilde{T}_j$ analytically. However, using extreme value theory from [18], [19], we can analyze the distribution of $\tilde{T}_j$ for large K. Then using (7) and (8), the asymptotic behavior of D can also be obtained.

C. Extreme Value Analysis

As $K \to \infty$, the distribution specified by (9) degenerates to:

$$\lim_{K\to\infty} F_{\tilde{T}_j,K}(x) = \begin{cases} 0 & \text{if } F_T(x) = 0 \\ 1 & \text{if } F_T(x) \le 1. \end{cases}$$

Extreme value theory focuses on the linear transformation $y = c_K + d_K x$ where c_K and d_K are constants that depend on K. Appropriate choice of these constants may result in a non-degenerate limiting distribution $L(x) = \lim_{K\to\infty} F_{\tilde{T}_j,K}(c_K + d_K x)$. A given *c.d.f.*, $F_T(x)$, is said to belong to, or lie in, the *domain of attraction for minima* of a given *c.d.f.* $L(x)$ if $L(x) = \lim_{K\to\infty} F_{\tilde{T}_j,K}(c_K + d_K x) = \lim_{K\to\infty} 1 - [1 - F_T(c_K + d_K x)]^K$ for given sequences $\{c_K\}$ and $\{d_K > 0\}$ ([18], [19]).

Assume flat Raleigh fading for each user, and that $R_{i,j}$ is proportional to the received SINR; this results in $f_R(r)$ having an exponential distribution, i.e.,

$$f_R(r) = \frac{1}{R_0}\exp\left(-\frac{r}{R_0}\right), \qquad (10)$$

where $R_0 = \mathbb{E}(R)$.

The *c.d.f.* for T is then given by the Frechet distribution:

$$F_T(t) = \exp\left(-\frac{T_0}{t}\right), \qquad (11)$$

which has *p.d.f.* $f_T(t) = \frac{T_0}{t^2}\exp\left(-\frac{T_0}{t}\right)$, where $T_0 = L/R_0$. Note that $\mathbb{E}(T) = \infty$.

We use the following Theorem from [18], to show that $F_T(t)$ given by (11) lies in the domain of attraction for minima of the Gumbel distribution.

Theorem 1: [18] A necessary and sufficient condition for a continuous *c.d.f.*, $F(x)$, to lie in the domain of attraction for minima of a Gumbel distribution is that

$$\lim_{\varepsilon\to 0} \frac{F^{-1}(\varepsilon) - F^{-1}(2\varepsilon)}{F^{-1}(2\varepsilon) - F^{-1}(4\varepsilon)} = 1. \qquad (12)$$

Given $F_T^{-1}(u) = -T_0/\ln(u)$, it can easily be shown that (12) holds. Therefore, we have the following Corollary:

Corollary: For $F_T(x)$ given by (11), the random variable $T_{i,j}$ lies in the domain of attraction for minima of a Gumbel distribution with *c.d.f.*

$$\begin{aligned} L(x) &= \lim_{K\to\infty} \Pr\left[\min_{i\in\{1,\dots,K\}} T_{i,j} \le c_K + d_K x\right] \qquad (13) \\ &= 1 - \exp(-\exp(x)), \end{aligned}$$

for $-\infty < x < \infty$.

It can be also shown using results from [18] that the appropriate constants c_K and d_K are given by:

$$c_K = \frac{T_0}{\ln(K)} \quad \text{and} \quad d_K = \frac{T_0}{\ln(K)(1+\ln(K))}. \qquad (14)$$

As $K \to \infty$, $\tilde{T}_j \to 0$ in mean square. From this analysis it follows that as $K \to \infty$, the mean slot duration $\mathbb{E}(\tilde{T}_j)$ decreases at rate $\Theta\left(\frac{1}{\ln(K)}\right)$. Therefore, using (7), we have shown that the average total delay $\mathbb{E}(D) = \Theta\left(\frac{K}{\ln(K)}\right)$ as $K \to \infty$. Note that if round robin scheduling was used, the mean delay $\mathbb{E}(D)$ would increase at rate $\Theta(K)$. Similarly, using (8), the variance of D grows at rate $\Theta\left(\frac{K^2}{\ln^2(K)}\right)$.

D. Numerical Results for Finite Users

In the previous subsection, we showed that asymptotically the expected delay grows at rate $\Theta(\frac{K}{\ln(K)})$ for a Max R scheduler. We now present numerical results that show this asymptotic growth rate holds even for a small number of users.

We simulate a backlogged system with packet length $L = 1024$ (bits). The Max R Rule is used. In each time slot, rates are chosen independently across users according to an exponential distribution with mean $R_0 = 50,000$ (bps).

Fig. 1 shows $\mathbb{E}(\tilde{T}_j)$ vs. the number of users, K. The solid line represents the simulation results. The dash-dot line is the analytical results that we get using the *c.d.f.* given by (9)‡. The dotted curve is $L/R_0 \ln(K)$ vs. K, where $L/R_0 = 0.02048$ is a normalizing constant. This corresponds to the asymptotic growth rate.

Notice the simulation and analytical results agree perfectly. The approximation given by the asymptotic analysis also performs very well even for small values of K.

IV. The U'R Rule

The Max R Rule makes the most efficient use of the channel since in each time slot the user with the highest data rate gets transmitted. However, it does not account for fairness and may perform poorly in terms of delay and stability (see [6]). To balance fairness with efficiency, we introduce a utility-based scheduling algorithm that takes into account both the channel condition and the user's utility function.

‡The analytical results are calculated using Maple. We do not show the results for $K \ge 20$, since the computational complexity becomes too high.

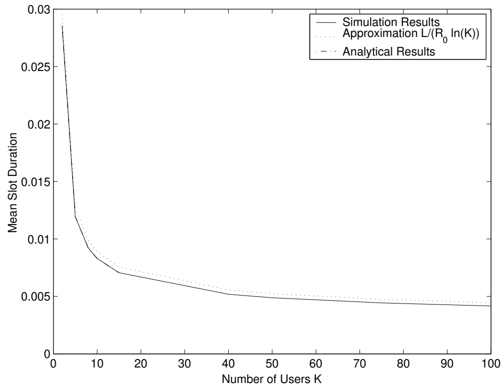

Fig. 1. $\mathbb{E}(\tilde{T}_j)$ vs. K.

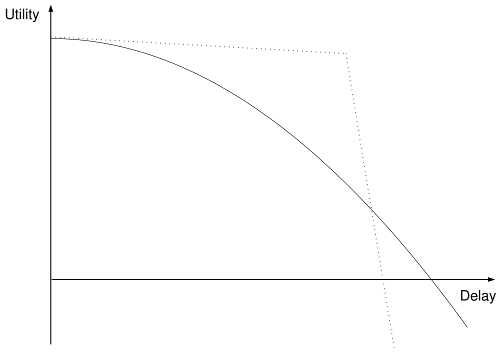

Fig. 2. Two example utility functions.

A. Utility Function

Assume that associated with each packet k is a utility function denoted by $U_k(D)$, where D is the total time between the packet arrival and service completion. This includes both the queueing time and the transmission time. The utility is decreasing in D, i.e. the longer the delay, the lower the utility. Two examples of $U_k(D)$ are depicted in Fig. 2. Unlike conventional definitions of utility functions in microeconomics, we assume that $U_k(D)$ can take on negative values. In that case, the utility is interpreted as the level of dissatisfaction that a user experiences due to packet transmission delay. By varying the shape of the utility function, different delay requirements can be reflected. For example if a packet has a deadline, then the utility function could be relatively flat before the deadline and drop sharply beyond that point, as shown by the dotted line in Fig. 2.

B. The U'R Rule – Motivation

Our objective is to schedule transmissions at each time to maximize the total utility rate:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{k=1}^{N(T)} U_k(D_k), \tag{15}$$

where $N(T)$ denotes the total number of packets served up to time T, and D_k represents the delay experienced by packet k. With this objective in mind, we first look at a simple case where only two users are present, and they both have packets waiting. Suppose that at the current scheduling instant their HOL packets have waited for W_1 and W_2 time units in the queue, respectively. Let $U_1(D)$ and $U_2(D)$ be the associated utility functions. Looking two time slots ahead, and assuming the channel for each user stays fixed across the two time slots, the utility derived by transmitting in the order $1 \to 2$ is

$$U_1(W_1 + \frac{L}{R_{1,1}}) + U_2(W_2 + \frac{L}{R_{1,1}} + \frac{L}{R_{2,1}}). \tag{16}$$

Similarly, transmitting in the reverse order gives the total utility

$$U_2(W_2 + \frac{L}{R_{2,1}}) + U_1(W_1 + \frac{L}{R_{2,1}} + \frac{L}{R_{1,1}}). \tag{17}$$

The total time for transmitting the two packets is the same in both cases.

To maximize utility rate, user 1 should be scheduled first if (16) $\geq$ (17). Assuming $L/R_{1,1}$ and $L/R_{2,1}$ are small, we can approximate (16) and (17) in terms of their first-order Taylor expansions to get that user 1 should be scheduled if

$$\begin{aligned} & U_1(W_1) + U_1^{'}(W_1)\tfrac{L}{R_{1,1}} + U_2(W_2) + U_2^{'}(W_2)(\tfrac{L}{R_{1,1}} + \tfrac{L}{R_{2,1}}) \\ \geq\ & U_2(W_2) + U_2^{'}(W_2)\tfrac{L}{R_{2,1}} + U_1(W_1) + U_1^{'}(W_1)(\tfrac{L}{R_{2,1}} + \tfrac{L}{R_{1,1}}). \end{aligned}$$

Notice that $U'(D) < 0$ for all $D > 0$. Cancelling terms and reorganizing, we get

$$| U_1^{'}(W_1) | R_{1,1} \geq | U_2^{'}(W_2) | R_{2,1} \tag{18}$$

as an approximate condition for when user 1 should be scheduled first. The preceding argument easily extends to a system with K users, assuming all packets belonging to user i have the utility function $U_i(\cdot)$. In that case, this approximation results in the base station scheduling user i^* given by

$$i^* = \arg \max_{i\in\{1,\dots,K\}} | U_i^{'}(W_{i,j}) | R_{i,j}, \quad \text{for all} \quad j, \tag{19}$$

where $W_{i,j}$ is the delay for the HOL packet of user i at the j^{th} slot. Ties can be broken arbitrarily. We call the scheduling rule in (19) the *U'R rule*. We can interpret $| U^{'}(W_{i,j}) |$ as a bid price per information bit submitted by the corresponding user. This rule can then be interpreted as scheduling the request which gives the most revenue per second in the next time slot. If we further assume that $U_i(\cdot)$ is concave, then the bid price increases with the waiting time $W_{i,j}$. In other words, as a packet waits in the queue, it becomes more urgent, and the potential utility loss by postponing its transmission becomes more substantial. The rate at which the urgency increases is decided by the second derivative of $U_i(\cdot)$.

We review the assumptions leading to the U'R rule to see when this rule should perform well. First, we have assumed that every packet eventually gets sent. If packets can be ignored forever or dropped from the queue without penalty, then we can serve the most users and achieve a higher utility rate by always sending to the user with the best channel. Second, we have used a Taylor expansion with only the first-order terms, so that L/R must be relatively small. Otherwise, second-order terms should be taken into consideration as well. In that case, we may not get a scheduling rule with a comparably simple structure. Last, we have assumed that the channel is constant over 2 scheduling periods. This is a reasonable assumption for low-tier mobility. In the next subsection we explore the effects on the scheduler's performance of relaxing this assumption.

The virtue of the U'R rule lies in its simplicity and flexibility. First, the information needed to make the decision is readily available at the base station. The channel gain, and hence the rate $R_{i,j}$, can be estimated via a pilot signal. The base station could learn a user's utility function during call set-up. Alternatively, users could provide the bid $\mid U_i^{'}(W_{i,j}) \mid$ when needed. In the latter case, the base station need not have complete knowledge of the utility function. The U'R rule is also flexible in that by choosing different utility functions, we get different scheduling rules. For example, the Max R Rule is a special case of the U'R Rule when $U_i(D) = -aD + b$, where $a > 0$ and b are constants for all i. If $U_i(D) = -a_i D + b_i$, where $a_i > 0$ and b_i are constants which depend on the user, then the U'R rule becomes scheduling the user with the largest $a_i R_i$. Clearly, a_i can be used to indicate a user's priority level. Specifically, if a_i's are chosen such that $1/a_i$ is the mean rate user i gets over a certain (relatively long) time window, then we get the *Proportional Fair Rule* proposed in [16], [17]. If $U_i(D) = -\beta_i D^2$, for all i, where $\beta_i > 0$, then we get the *Modified Largest Weighted Delay First (M-LWDF) rule* which was proposed in [9]. It is shown that the M-LWDF rule is *throughput optimal*, i.e., it makes the queues stable if it is feasible to do so with any other scheduling rule.

C. Simulation Results

We simulate a backlogged system with 2 users; both the Max R rule and the U'R rule will be considered. Instead of completely transmitting one packet at once, we use a fixed time-slot scheme similar to HDR. In this setting, the Max R Rule stays the same at each decision instant. The U'R Rule is now to transmit to the user i^* such that

$$i^* = \arg \max_{i \in \{1,\dots,K\}} \mid U_i^{'}(W_{i,j}) \mid R_{i,j}/L_{i,j}, \quad \text{for all} \quad j,$$

where $L_{i,j}$ denotes the remaining packet length of user i at the j^{th} decision epoch.

The packet length is set to be $L = 20$ (unit lengths) and slot duration is normalized to 1 (slot). The supportable rates $R_{i,j}$ varies with time and has a marginal exponential distribution with mean $\mathbb{E}(R_{i,j}) = 5$ (unit length/slot) for all i and j. The rates are correlated over time. Specifically, $R_{i,j} = |X_{i,j}|^2$, where for each i, $\{X_{i,j}\}_{j=1}^{\infty}$ is a complex Gauss-Markov random process. For $j = 1$, the real and imaginary parts of $X_{i,1}$ are *i.i.d.* Gaussian random variables with distribution $N(0, \mathbb{E}(R_{i,j})/2)$. For $j > 1$, $X_{i,j} = \varrho X_{i,j-1} + \xi_{i,j}$, where $\xi_{i,j}$'s are independent complex Gaussian random variables.

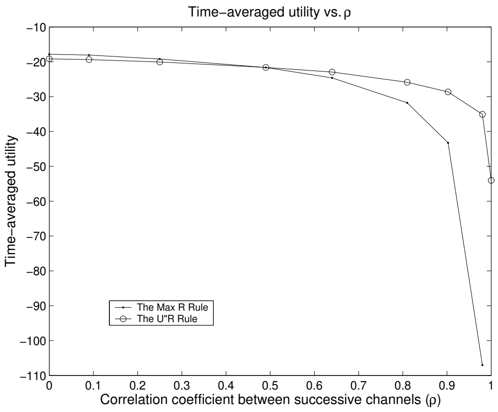

Fig. 3. $\lim_{T\to\infty} \frac{1}{T} \sum_{i=k}^{N(T)} U(D_k)$ vs. ρ.

The real and imaginary parts of $\xi_{i,j}$ are also *i.i.d.* with distribution $N(0, (1-\varrho^2)\mathbb{E}(R_{i,j})/2)$. This results in $R_{i,j}$ having the desired marginal distribution. It can also be shown that the covariance coefficient between $R_{i,j}$ and $R_{i,j-1}$ is given by $\rho = \varrho^2$ for all $j > 1$. We assume that the utility function for each user is the same and is given by $U(D) = -D^2$.

Fig. 3 shows the time-averaged utility vs. ρ for the different scheduling rules. Each data point is averaged over ten different random seeds. First observe that the average utility decreases as the channel correlation increases. That is as expected, since the more independent the channels are over time, the more advantage an "opportunistic" scheduler gains via multiuser diversity. The U'R Rule performs better when the channels are highly correlated (> 0.5). For example, when $\rho = 0.9025$, the utility gain of the U'R Rule over the Max R Rule can be as large as 34%. Furthermore, when the channels are constant ($\rho = 1$), the Max R Rule causes infinite delay for the user with the worse channel and therefore the derived utility goes to negative infinity. The U'R Rule schedules users in a periodic fashion in this case resulting in a finite utility-rate. The Max R Rule achieves higher utility-rate when the channel correlations are low, although the difference is not as significant.

Fig. 4 shows two typical delay complementary *c.d.f.*'s with $\rho = 0.81$ under the U'R Rule and the Max R Rule, respectively. Notice the Max R curve has a sharper drop at the start and a heavier tail. In other words, the delay for the U'R Rule is distributed more evenly than with the Max R Rule. Considering fairness, a more equitable delay distribution is a desirable feature; the U'R Rule clearly achieves better performance in this aspect.

V. Conclusions

We have considered downlink scheduling for wireless data. We analyzed the delay distribution for the Max R Rule which maximizes throughput. By applying extreme value theory, we obtained the asymptotic rate at which the mean total delay increases with the number of users. To enforce fairness, we considered a maximum utility-rate formulation which led to the

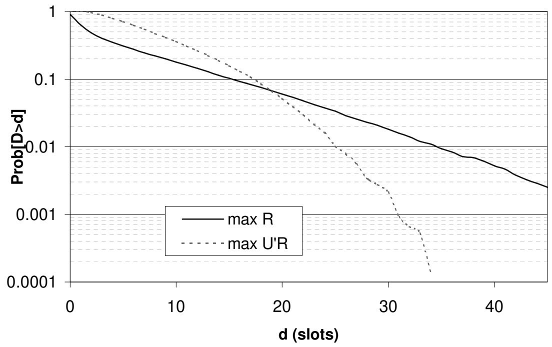

Fig. 4. Complimentary c.d.f. of delay under Max R and U'R scheduling rules.

U'R scheduling rule. The U'R Rule takes into account both the channel conditions and the utility functions. Note that the Max R rule is a special case of the U'R Rule. Simulation results have shown that the U'R Rule tends to outperform the Max R Rule in terms of maximizing time-averaged utility when the channel correlation is high. When the channel correlation is low, the Max R Rule performs better.

REFERENCES

[1] P. Bender, *et al*. "CDMA HDR: a bandwidth-efficient high-speed wireless data service for nomadic users", in *IEEE Commun. Mag.*, Pages 70-77, July 2000.

[2] TIA/EIA IS-856 CDMA 2000: High rate packet data air interface specification, Nov. 2000.

[3] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queue with randomly varying connectivity", in *IEEE Transactions on Information Theory*, Vol. 39, pp. 466-478, March 1993.

[4] P. Bhagwat, P. Bhattacharya, A. Krishna and S. K. Tripathi, "Enhancing throughput over wireless LANs using channel state dependent packet scheduling", in *Proceedings of Infocom*, San Franciso, CA, March 1996, pp. 1133-1140.

[5] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel", in *Proceedings of WowMoM*, Seattle, WA, 1999, pp. 35-42

[6] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR", Bell Laboratories Technical Report, Oct. 2000.

[7] S. Shakkottai and A. L. Stolyar, "Scheduling for multiple flows sharing a time-varying channel: The exponential rule", submitted.

[8] A. L. Stolyar, "MaxWeight scheduling in a generalized switch: state space collapse and equivalent workload minimization under complete resource pooling", submitted.

[9] M. Andrews, K. Kumaran, K. Ramanan, A. L. Stolyar, R. Vijayakumar and P. Whiting, "CDMA data QoS scheduling on the forward link with variable channel conditions", Bell Laboratories Technical Report, April, 2000.

[10] K. Lee and M. El Zarki, "Packet Scheduling Schemes for Real Time Services in Cellular IP Networks", preprint, 2001.

[11] V. Bharghavan, S. Lu and T. Nandagopal, "Fair queuing in wireless networks: Issues and approaches", in *IEEE Personal Communications*, Vol. 6, pp. 44-53, Feb 1999.

[12] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single-node case", in *IEEE/ACM Transactions on Networking* , Vol. 1, pp. 344-357, June 1993.

[13] A. Parekh and R. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the multiple -node case", in *IEEE/ACM Transactions on Networking* , Vol. 2, pp. 137-150, April, 1994.

[14] P. Humblet and R. Knopp, "Information capacity and power control in single-cell multiuser communications", in *Proc. ICC '95*, pp. 331-335, June, 1995.

[15] D. Tse, "Multiuser diversity in wireless networks: smart scheduling, dumb antennas and epidemic communication.", Presented at *IMA workshop on wireless networks*, Aug. 10, 2001.

[16] D. Tse, "Forward link multiuser diversity through proportional fair scheduling.", August 1999. Presentation at Bell Labs.

[17] A. Jalali, R. Padovani, R. Pankaj, "Data throughput of CDMA-HDR a high efficiency - high data rate personal communication wireless system.", in *Proc. VTC '2000*, Spring, 2000.

[18] Enrique Castillo, *Extreme Value Theory in Engineering*. New York: Academic, 1988.

[19] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*. New York: John Wiley and Sons, 1978.