Class and Channel Condition Based Weighted Proportional Fair Scheduler

Rajeev Agrawal[a] , Anand Bedekar[a], Richard J. La[a], and Vijay Subramanian[a]

[a] Mathematics of Communication Networks
Motorola Inc., Arlington Heights, IL, USA 60004.

In this paper we outline a scheme to perform packet-level scheduling and resource allocation at the wireless node that takes into account the notions of both efficiency and fairness and presents a means to explore the trade-off between these two notions. As a part of this scheme we see the scheduling problem as deciding not just the packet transmission schedule but also the power allocation, the modulation and coding scheme allocation and the spreading code determination since the latter three directly influence the radio resources consumed. Using a utility maximization formulation based on the data-rates that the mobiles can transmit at, we decide on the weights for a weighted proportionally fair allocation based scheduling algorithm. We conclude with a simulation based performance analysis for infinitely-backlogged sources on a UMTS system.

1. INTRODUCTION

The explosion of multimedia services on the Internet is leading to a demand of the same services in the non-tethered wireless space. There are, however, many peculiarities that a wireless channel possesses which makes supporting such services much tougher than on wireline networks. One of the key elements in this is the scheduler used at various nodes. Scheduling in traditional wireline networks consists mainly of deciding the order in which users access the channel. This is because it is quite easy to use these channels very close to their (information theoretic) capacity at any given power (used on the channel). Thus, it is best to operate at the maximum capacity by using the maximum power all the time. In addition, the channels, and thus the data rates, are not time-varying either. On wireless channels, however, there are many considerations which do not allow for such a mode of operation. The bandwidth available for transmission on a wireless channel and the power levels allowed (both regulated) put a hard limit on the capacity[1]. Another important element is the mobility of the end-user equipment which results in time-varying multipath and fading. Further, the size, battery power, and processing power of the end devices place additional constraints on system performance. Limited battery capacity also makes it necessary to use transmission schemes that would prolong battery life as much as possible. Finally, the multiuser nature of a wireless channel makes it interference-limited. Thus, one user transmitting at maximum power could severely impair the transmissions

[1]The propagation characteristics of the atmosphere, and other media, are also deciding factors for the band of operation and hence, the bandwidth.

of other users. Thus, using a traditional wireline scheduler is not a good approach on a wireless channel.

Some of recent developments in wireless link scheduling include the work by Holtzman [1], Jalali *et al.* [2], Tse [3], Shakkottai and Stolyar [4], Chawla *et al.* [5], Leelahakriengkrai [6], and Berry and Gallager [7]. All of these works show that substantial benefits are achieved when the higher layers are aware of the radio conditions and can adapt the powers, the modulation schemes, the coding schemes, and spreading gains (and hence the data rates) based upon this knowledge. The upshot of this is that scheduling policies should be devised using the knowledge of channel conditions. In a cellular context there is an additional benefit to the network layer control of transmission strategies. In such a situation it is possible to trade capacity among cells (by changing the power levels, for instance) to alleviate periods of congestion or high demand. From the discussion above it is clear that a scheduler that jointly performs packet-level scheduling and radio-resource allocation is the solution for the wireless-link scheduling problem. To be able to implement such scheduling policies it is necessary to have a system that has controls in place to allow for changing the transmission parameters easily. The third-generation (3G) technologies are a first step in this direction. Nanda *et al.* [8] provide a fairly comprehensive overview of the 3G technologies and how they have been designed with multimedia-type services in mind. In all proposals it is possible for connections to not only choose from a variety of data rates but also change the data rate in a flexible and quick fashion. There is also added feedback, in terms of more frequent error and measurement reports, which in conjunction with flexible data rate allocation, can in turn be used for a better monitoring of QoS guarantees and provisioning of resources.

Since the radio resource is quite expensive the efficient management of this resource is critical. This, however, cannot be the only concern when QoS parameters have been agreed to for various users and when the operator is obliged to uphold them. Thus, it is imperative to have some fairness in the arbitration of resources amongst the various users. In this paper we outline a scheme to perform packet-level scheduling and resource allocation at the wireless (access) node that takes into account the notions of both efficiency and fairness and presents a means to explore the trade-off between these two notions. This trade-off between efficiency and fairness was not a concern of earlier cellular systems because voice was the major application and thus, only coverage (which is in reality just another terminology for fairness) was critical. It is only with the emergence of the 3G technologies and packetized data services that such a trade-off makes any sense. In wireline networking there are two broad philosophies when QoS provisioning is considered. One follows the IntServ approach and attempts to provide strict QoS guarantees [9]. Another approach uses the ideas in DiffServ to provide a class-based differentiation of services [10]. Differentiated services are supported through various per-hop-behaviors (PHBs) in DiffServ capable networks. For instance, expedited forwarding (EF) is aimed at supporting real-time applications such as video conferencing. Other PHBs, such as assured forwarding (AF) [11] and best effort (BE), support non-real-time applications that do not require strict delay guarantees. AF can further have different services, for instance, Gold, Silver, and Bronze services.

The work in [4,6] is more in the IntServ context and they schedule users based upon their current backlog to satisfy statistical delay guarantees. With such scheduling mechanisms

one needs a good admission control policy and policing mechanism in place. The work in this paper is based on the DiffServ philosophy where we provide a differentiation based upon both class and channel-state. Further, we restrict our attention to non-real-time (rate-adaptive) services. The reason we choose to provide a service differentiation based upon channel-state is because this determines how much network resource is utilized by the application. This is very much in keeping with similar differentiation in wireline networks with protocols like TCP using quantities like round-trip time estimates and hop-count as surrogates for measuring how much of the network resource is utilized by various applications. The work in [1–3,5] also adheres to the DiffServ philosophy. In the proposed algorithm in this paper we suggest a flexible way of trading off efficiency for fairness as well as a flexible means of exploiting temporary fluctuations in channel conditions.

In Section 2 we give a detailed introduction to the wireless link scheduling problem. Thereafter, we introduce a related resource allocation problem in Section 3 and present our scheduling algorithm in Section 4. A performance analysis of the algorithm with infinitely-backlogged sources is discussed in Section 5. Finally, we conclude in Section 6.

2. THE WIRELESS LINK SCHEDULING PROBLEM

Consider a cellular system. In a given cell b let J_b be the set of users on the downlink. Time is slotted into radio blocks (in GPRS and EDGE) or frames in (cdma2000 and UMTS) of fixed duration (20, 10, or 3.33 ms depending on the technology). We shall refer to these time slots as frames hereafter. The wireless link scheduling problem is one of deciding which of these users transmit in each frame. In a TDMA system like GPRS and EDGE only one user is allowed to transmit in a frame, whereas in a CDMA system like cdma2000 or UMTS multiple users may transmit in a frame. When a user transmits, we also need to decide what power level, modulation and coding scheme, time slot (in case of TDMA) and spreading factor (in case of CDMA) it will use.

Due to different base site to user distances, shadow fading and multipath, the channel conditions of different users vary with time. This fluctuation in channel conditions results in a variation of the *effective data rate (per channel per unit power)* $\hat{R}_j(t)$ available to the different users $j \in J_b$ in different frames $t = 0, 1, \ldots$. This effective data rate per unit resource may be calculated in a variety of ways. We outline two simple ones below. Consider the signal to interference plus noise ration (SINR) of user j

$$SINR_j(t) = \frac{P_j(t)G_{jj}(t)}{\sum_{i \neq j} P_i(t)G_{ij}(t) + \sigma^2} \ , \tag{1}$$

where $P_j(t)$ is the transmit power of user j, σ^2 is the receiver noise variance and $G_{ij}(t)$ is the energy gain from base station of mobile station i to mobile station j. Once $SINR_j(t)$'s are available for the users (either using the above formula or direct measurements), one can compute the data rates and frame error rates (FER) corresponding to different choices of modulation and coding schemes (MCSs) and/or spreading factors (SFs). Hence, one can find the choice of MCS and/or SF that maximizes $R_j(1 - FER)$, where R_j and FER are the data rate and frame error rate of user j corresponding to the MCS and/or SF, respectively.

In case of a TDMA system where we do not share the transmit power of the base station across multiple users in the same frame, we may consider the transmit power fixed. In this case the resulting optimum above may be considered to be the effective data rate per unit resource $\hat{R}_j(t)$.

In a CDMA system we may consider the energy per bit to noise power-spectral-density ratio (E_b/N_0) given by:

$$\left(\frac{E_b}{N_0}\right)_j(t) = \frac{W}{R_j(t)} SINR_j(t) = \frac{W}{R_j(t)} \frac{P_j(t)G_{jj}(t)}{\sum_{i\neq j} P_i(t)G_{ij}(t) + \sigma^2} \tag{2}$$

where W is the channel bandwidth and W/R_j is user j's spreading factor. As a surrogate to controlling the FER we may attempt to satisfy an E_b/N_0 target Γ_j. Let the smallest bit rate allowed (corresponding to the largest SF) be $R_{\min}$. We may think of this as a CDMA channel. Then we can define the effective data rate per channel per unit power as

$$\hat{R}_j(t) = \frac{R_j(t)/R_{\min}}{P_j(t)} = \frac{W}{R_{\min}\Gamma_j} \frac{G_{jj}(t)}{\sum_{i\neq j} P_i(t)G_{ij}(t) + \sigma^2} \tag{3}$$

What follows does not depend on precisely how $\hat{R}_j(t)$ is defined and measured. The key idea is that it will be monotonically increasing in the users' own channel gain and decreasing in interference plus noise.

Let M_b channels be available and let P_b be the power available per channel at base site b. In each frame t we have to decide what fraction $\rho_j(t) > 0$ of the resources (channels and powers) will be allocated to the different users $j \in J_b$; $\sum_{j\in J_b} \rho_j(t) \leq 1$. In which case it gets a throughput $r_j(t) = \hat{R}_j(t)\rho_j(t)P_bM_b$ (if all of the channel and power resources are given to user j, then it would get a throughput $r_j = \hat{R}_j(t)P_bM_b$). Typically there will be additional constraints on $\rho_j(t)$ depending on the technology as described below.

In a TDMA system, like GPRS or EDGE, M_b denotes the number of *time slot* channels available ($M_b = 1, \ldots, 8$). One and only one user can transmit in a time slot in a frame, so $\rho_j(t)M_b$ must be an integer. Additional restrictions on allocation of time slots to users may further constrain $\rho_j(t)$.

In a CDMA system like cdma2000 or UMTS, M_b may be the number of spreading codes (a code representing the smallest data rate allocation R_{min}) available and P_b be the power per code available, in which case M_bP_b is the total power budget at base site b. Let $f_j(t)$ be the fraction of spreading codes given to user j and let $p_j(t)$ be the fraction of per code power to be used by user j. Then $f_j(t)M_b$ will be the number of codes given to user j and must be an integer (it may even have to be a power of 2). $p_j(t)$ may be an arbitrary real number; $p_j(t)P_b$ will be the power per code and $f_j(t)M_bp_j(t)P_b$ the total power given to user j. We may require that the power per code given to user j be such that it satisfies a certain E_b/N_0 target (2). Of course we will need that $\sum_{j\in J_b} f_j(t)M_bp_j(t)P_b \leq M_bP_b$ or that $\sum_{j\in J_b} \rho_j(t) \leq 1$, where $\rho_j(t) = f_j(t)p_j(t)$.

In summary, the wireless link scheduling problem requires that in each frame t we decide what fraction $\rho_j(t) \geq 0$ of the resources (channels and powers) will be allocated to the different users $j \in J_b$; $\sum_{j\in J_b} \rho_j(t) \leq 1$, in which case it gets a throughput $r_j(t) = \hat{R}_j(t)\rho_j(t)P_bM_b$. Since $\hat{R}_j(t)$ varies with both the user and time we would like to do the scheduling in such a way that we capitalize on these variations to get high system throughput while providing some level of QoS differentiation.

3. BASIC ALGORITHM

In the previous section we have considered the resource allocation problem on a frame-by-frame basis. This requires that the number of resources allocated to the users need be an integer. However, in the following sections as a first step of designing the scheduling algorithm we relax this constraint and consider the framework where we are interested in finding the fraction of resources to be allocated to the users over a sufficiently large period.

Given the effective data rate per unit resource $\hat{R}_j$ of the users as described in Section 2 we compute the fraction of the resources that will be allocated to each user j for transmission by solving the following optimization problem:

$$\begin{aligned} \max_{\rho_j} \quad & \sum_{j \in J_b} U_j(\rho_j \hat{R}_j P_b M_b) \\ \text{subject to} \quad & \sum_{j \in J_b} \rho_j \leq 1 \\ & \rho_j \geq 0, \end{aligned} \tag{4}$$

where $U_j(\cdot)$ is the utility function of user j as a function of the throughput it receives. The optimization problem in (4) computes the solution that maximizes the aggregate utility of the users given the resource and non-negativity constraints.

We first characterize the solution of (4) with the most commonly used utility functions of

$$U_i(r_i) = f_\alpha(r_i) = \begin{cases} \text{sgn}(\alpha) \cdot r_i^\alpha, & \text{if } \alpha \neq 0, \alpha < 1 \\ \log(r_i), & \text{if } \alpha = 0. \end{cases} \tag{5}$$

With the utility functions of $f_\alpha(\cdot)$, one can show that the solution of the optimization problem in (4) is given by

$$\rho_j = \frac{(\hat{R}_j)^{\beta-1}}{\sum_{k \in J_b} (\hat{R}_k)^{\beta-1}} \propto (\hat{R}_j)^{\beta-1}, \tag{6}$$

where $\beta = \frac{1}{1-\alpha}$. Note that if α is greater than zero, the allocation favors users with higher $\hat{R}_j$, and if α is less than zero, the allocation favors users with lower $\hat{R}_j$. The value of α equal to one leads to efficiency only solution in that all slots are allocated to the users with the highest $\hat{R}_j$, while a value of α close to $-\infty$ yields a fairness only solution in that every user receives approximately the same rate. In this sense the parameter α controls the extent to which this bias is enforced and hence how efficiency, i.e., throughput, is traded off in favor of fairness.

After computing the solution ρ^* to (4) we compute the credits C_j for the users, where the credit of user j is $C_j = \rho_j^* \cdot \hat{R}_j$. Note that the credit, C_j, of user j would be the throughput of the user normalized by $P_b M_b$ if it indeed received ρ_j^* of the resources. However, due to various system constraints a user' throughput may differ from its credit. For instance, in EDGE users are placed on one or more time slots, depending on whether they are single or multiple slot capable. A user's actual rate depends both on its ρ_j^* and time slot configuration. In a CDMA system users may have maximum data-rate

constraints or maximum power constraints. Incorporating these system constraints into the optimization problem leads to a weighted proportionally fair[2] (WPF) allocation [12, Eqn. (2)] with weights $\frac{C_j}{\hat{R}_j}$ as proved in the following proposition.

Proposition 3.1 *The weighted proportionally fair rates with the weights* $\frac{C_j}{\hat{R}_j}s$ *are also the optimal solution to the problem in (4) with the addition of the system constraints mentioned above.*

Proof: See [13].

4. THE CLASS AND CHANNEL CONDITION BASED WEIGHTED PROPORTIONAL FAIR (C^3WPF) SCHEDULER

In our algorithm that is described in this section we use users' credits to allocate the available bandwidth. The idea behind the algorithm is to mimic the behavior of weighted fair queueing (WFQ) without explicitly computing the virtual times for the arriving protocol data units (PDUs). The credits C_j are similar to the weights ϕ_j in WFQ. We show that our algorithm leads to a weighted proportionally fair (WPF) rate allocation in the sense that the users that have the same set of bottlenecks or system constraints receive rates that are proportional to their credits.

We first consider the simple case where the channel conditions and thus the effective data rates per unit resource of the users are time-invariant so as to explain the key idea behind our algorithm. Then, we describe the actual algorithm that uses the values of current and average effective data rates per unit resource of the users.

4.1. A Simple Algorithm for Time-invariant Channels

Each user has a traffic class associated with its connection. For instance, in EDGE there are six traffic classes: conversational, streaming, interactive best-effort (I1, I2, and I3), and background best-effort. Associated with each traffic class is a weight w. This weight may reflect the price charged to the traffic class per unit time of usage [14,15] and is used in the computation of credits in order to provide differentiated services among the traffic classes.

The scheduling algorithm described below attempts to deliver throughtputs proportional to credits $C_j = w_j\hat{R}_j^{\beta}$, which reflects both users' traffic classes and channel conditions. However due to additional constraints on slots, powers, codes, etc., this precise proportion may not be obtainable. We would therefore like to obtain the weighted proportional fair (WPF) solution with weights equal to the credits. In order to achieve this we use the following algorithm described below:

Let $W_j(t)$ be the total throughput of user j up to time t. Let $\bar{W}_j(t) := W_j(t)/C_j$ be the throughput normalized by credits. At time $t+1$ we sort users in increasing order of their $\bar{W}_j(t)$. The scheduler then picks the user at the front of this list and schedules it for transmission in frame $t+1$. At the same time it determines the channel and power resources needed for this user. Should resources remain, it goes down the

[2]A vector of rates r^* is said to be weighted proportionally fair with weights p if and only if it is feasible and for any other feasible rate vector r it satisfies $\sum_j p_j \frac{r_j - r_j^*}{r_j^*} \leq 0$.

sorted list, in order, to select additional users for transmission in that frame. Users selected for transmission should obviously have data to send in that frame. Note that by favoring users with low $\bar{W}(t)$ for transmission, this algorithm tries to equalize the normalized throughputs $\bar{W}_j(t)$ over all users $j \in J_b$ as time $t \to \infty$ so as to get throughputs proportional to their credits C_j. However as mentioned earlier this may not be feasible due to additional constraints. The best achievable in that case would be the WPF throughput allocation in the sense that the users with the same system constraints would receive rates proportional to their credits. We show below that this algorithm does deliver the WPF throughputs asymptotically.

Proposition 4.1 *The average throughputs of the users, i.e., $\frac{W_j(t)}{t}$, converge asymptotically to the weighted proportionally fair rates with the weights $\frac{C_j}{\hat{R}_j}$s as $t \to \infty$.*

Proof: See [13].
From this we have the following corollary.

Corollary 4.1 *The average throughputs of the users, i.e., $\frac{W_j(t)}{t}$, converge asymptotically to the optimal rates as $t \to \infty$.*

Note from the definition of the credits that two users with the same channel conditions ($\hat{R}_j$s) can ask for different service based upon the weight of the class they subscribe to.

4.2. The Actual Algorithm for Time-varying Channels

In the algorithm described above we have assumed that the channel conditions, as captured in the effective data rate per unit resource $\hat{R}_j$, do not vary with time. We now describe the actual algorithm that can take advantage of the time-varying channel conditions of the users to improve the system throughput. The key change is in the values of $\hat{R}_j$ used above and in the update equation for $\bar{W}_j$. Let $\hat{R}_j(t)$ be the current effective data rate per unit resource based on current channel conditions. Let $\hat{R}_j^{\text{av}}(t)$ be corresponding average obtained using geometric IIR filtering, i.e.,

$$\hat{R}_j^{\text{av}}(t+1) = \psi \cdot \hat{R}_j^{\text{av}}(t) + (1-\psi)\hat{R}_j(t) \quad . \tag{7}$$

The credits calculated use both $\hat{R}_j(t)$ and $\hat{R}_j^{\text{av}}(t)$ as

$$C_j(t) = w_j(\hat{R}_j^{\text{av}}(t))^\beta \left(\frac{\hat{R}_j(t)}{\hat{R}_j^{\text{av}}(t)}\right)^\gamma = w_j(\hat{R}_j^{\text{av}}(t))^{\beta-\gamma}(\hat{R}_j(t))^\gamma = C_j^1(t)C_j^2(t) \quad , \tag{8}$$

where $0 \leq \gamma \leq \beta$. The value of γ should depend on how accurate or reliable $\hat{R}_j(t)$'s are. We use different factorizations of $C_j(t)$ into $C_j^1(t)$ and $C_j^2(t)$ to construct different scheduling algorithms.

Let $D_j(t)$ be the amount of data transmitted in frame t for user j. $\bar{W}$ is updated as per the following algorithm.

$$\bar{W}_j(t+1) = \phi \cdot \bar{W}_j(t) + (1-\phi)\frac{D_j(t)}{C_j^1(t)} \tag{9}$$

In keeping with the algorithm described earlier, at the beginning of time frame $t+1$, we sort users in increasing order of $\bar{W}_j(t+1)/C_j^2(t)$ and select users for transmission based on available resources.

Currently we have three different scheduling algorithms:

- **Variant 1**: For this we use $C_j^1(t) = C_j(t)$ and $C_j^2(t) = 1$.
- **Variant 2**: Here we use $C_j^1(t) = w_j(\hat{R}_j^{\text{av}}(t))^\beta$ and $C_j^2(t) = (\frac{\hat{R}_j(t)}{\hat{R}_j^{\text{av}}(t)})^\gamma$.
- **Variant 3**: Here we use $C_j^2(t) = C_j(t)$ and $C_j^1(t) = 1$. Note that this resembles the algorithm analyzed earlier; the time-invariant credit value is substituted by a time-varying definition of credit.

Note that choosing the credits as described above does two things. Since the effective data rate per unit resource $\hat{R}_j(t)$ is time varying, it uses an estimate $\hat{R}_j^{\text{av}}(t)$ for its *average*. The estimate is based on an IIR filter; other estimates may easily be substituted. If we set $\gamma = 0$ above, then we will simply be biasing the throughput in proportion to this average (raised to the power β). However, by dividing by non-unity values of the factor $C^2(j(t)$ (in variants 2 and 3), we tend to favor those users that have better channel conditions relative to their own average conditions. This second idea is also exploited in the scheduling algorithms proposed by Holtzman [1], Tse [3], and Jalali *et al.* [2]. In fact the algorithm proposed is similar to Variant 3 proposed above with $\beta = \gamma = 1$. One major difference is that they use a time-average computation of $\bar{W}$ instead of the IIR filter estimator in (9). The time constant in the IIR filter estimator for the average must be chosen depending on how frequently channel condition measurements are available and the time constants involved in the channel fluctuations (distance based path loss, shadow fading, and fast fading).

5. PERFORMANCE ANALYSIS

In this section we describe our experiments with the proposed scheduling algorithm. In addition to evaluating the performance of the algorithm in a simplistic setting, we are also interested in comparing the performance of a scheduling algorithm that is aware of the channel conditions with a scheduling approach that does packet-level scheduling separate from radio-resource allocation. Henceforth in this document we refer to the class of schedulers that are not aware of channel conditions as split-schedulers. The specific split-scheduler that we consider is one that tries to equalize the data rates that all the users get. Note that this can be easily achieved in the framework of our algorithm by setting $\beta = 0$. This is equivalent to a WFQ scheduler with equal weights. Throughout the performance analysis section we concentrate on a UMTS system.

5.1. Simulation Set-up

In this sub-section we describe the set-up for our performance analysis. First we describe the physical constraints, and thereafter the network-level characteristics like traffic sources, fragmentation and ARQ-mechanisms.

We consider a system with 25 cells, which use the same carrier and the same sector, and 15 mobiles in each cell. The positions of the mobiles are chosen at random, spatially

Table 1
Parameters used for UMTS simulation experiments.

Number of data mobiles per cell	15
Number of voice users per cell	32
BTS max. transmit power	40 dBm = 10 W
Non-orthogonality factor α	0.4
Allowed bits / frame	2400, 1800, 1200, 600, 300
Max. total bits / frame	14400 (3/4 of code tree)
Power consumed by overhead channels	30% of BTS power
Inter Base-site spacing	2800 m
Shadow Fading	off
Simulation time	25 s
Chip rate	3.84 Mcps
System bandwidth	5 Mhz
Noise Power	1.214×10^{-10} mW = -99 dBm
Mobile speed	3 km/h

uniform in an annulus that extends half-way into the cell starting from the edge of the cell and kept fixed for the entire simulation. The UMTS frames are 10 ms in duration. Since the system is a CDMA system we use a reuse factor of one and for simplicity also assume that we have no sectorization. We model fast fading using the Jakes simulator [16] but do not incorporate log-normal shadow fading terms in the propagation loss coefficients. Each cell has a power budget that cannot be exceeded. A part of this power budget is consumed by control channels. After allocating power to the control channels we allocate non-zero power values, such that their sum does not exceed the remaining power budget, to the voice users and then add the data users in one at a time using the $\bar{W}$ ordering. In addition to a power-budget we also have a code-allocation check that checks to see if a valid code is available. For all our simulation experiments we assume that there are 32 voice users in the system. Open-loop power control on a fast time-scale (1500 Hz) is also implemented in the simulations. The parameters used for the experiments are summarized in Table 1.

We assume that all the sources for our experiments are infinitely back-logged sources. We also assume that we get instantaneous feedback for the transmissions on the wireless channel and the erroneous frame is immediately put back at the head-of-the-line for transmission. Thus, we are assuming that our ARQ mechanism transmits packets till they are correctly received. The ARQ mechanism also reacts instantaneously to bearer-type (data rate) changes recommended by the power-budget calculation and code-check algorithm fragmenting in a way such that it transmits the right number of bits in each frame.

5.2. Simulation Experiments

In this set-up we try $\beta = 0$, 1, and 2, which would allow us to compare the split round-robin scheduler ($\beta = 0$) with a scheduler that uses $U(x) = \log(x)$ ($\beta = 1$) as the utility function and another that uses $U(x) = \sqrt{x}$ as the utility function ($\beta = 2$). We also seek

Table 2
Comparison of throughput per cell - mean and standard deviation - for different values of β and variations of the scheduling algorithm. In the case of "Variant 1 - shuffling", "Variant 2" and "Variant 3", $\gamma = 1$, for $\beta = 1, 2$; and for $\beta = 0, \gamma = 0$. The maximum throughput possible is 1.44 Mbps.

β	Variant 1, $\gamma = 0$	Variant 1 - shuffling	Variant 2	Variant 3
0	294.39, 5.83 Kbps	294.39, 5.83 Kbps	294.39, 5.83 Kbps	294.39, 5.83 Kbps
1	353.14, 7.97 Kbps	385.84, 8.85 Kbps	421.82, 9.27 Kbps	442.36, 9.77 Kbps
2	423.43, 10.00 Kbps	457.74, 10.83 Kbps	480.86, 10.95 Kbps	522.15, 11.73 Kbps

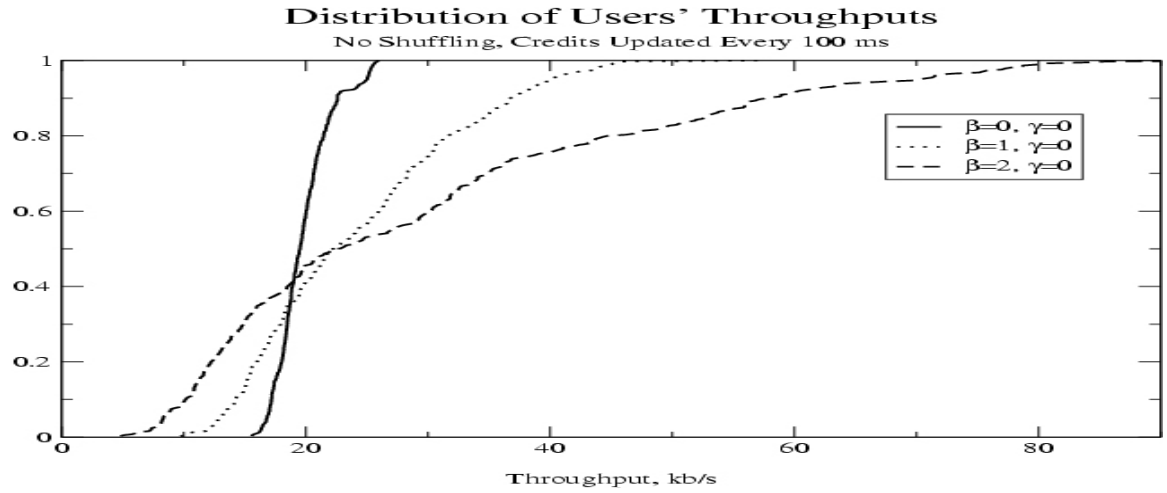

Figure 1. Distribution of individual user throughputs for Variant 1 with $\gamma = 0$.

to document the improvements obtained using knowledge of current channel conditions while making the scheduling decisions. The answers we seek are the rate-distance profiles in each cell, the user throughput distribution as well as sum throughput average and standard deviation (across multiple runs) per cell.

For the first set of experiments we assume that we do not have the knowledge of the current channel but only the knowledge of the average channel conditions obtained by the IIR filter (7). We also assume that the average rate per unit resource $\hat{R}^{av}(t)$ is available to the scheduler once in 10 frames, i.e., once in 100 ms. In Table 2 this case is listed under "Variant 1, $\gamma = 0$." In Figure 1 we plot the distribution of individual user throughputs. As expected the distribution gets more skewed as we increase β. An important point to notice is that sacrificing a little bit for a user with a bad channel we can substantially increase the throughput of the good user. In Figure 2 we plot the rate-distance profile that is achieved in one cell.

For our next set of experiments we assume that we know the current channel conditions and we use Variant 1 of the scheduling algorithm with $\gamma = 1$ for both $\beta = 1$ and $\beta = 2$. The case of $\beta = 0, \gamma = 0$ is included for comparison. Thus, in the case of $\beta = 2$ scheduling

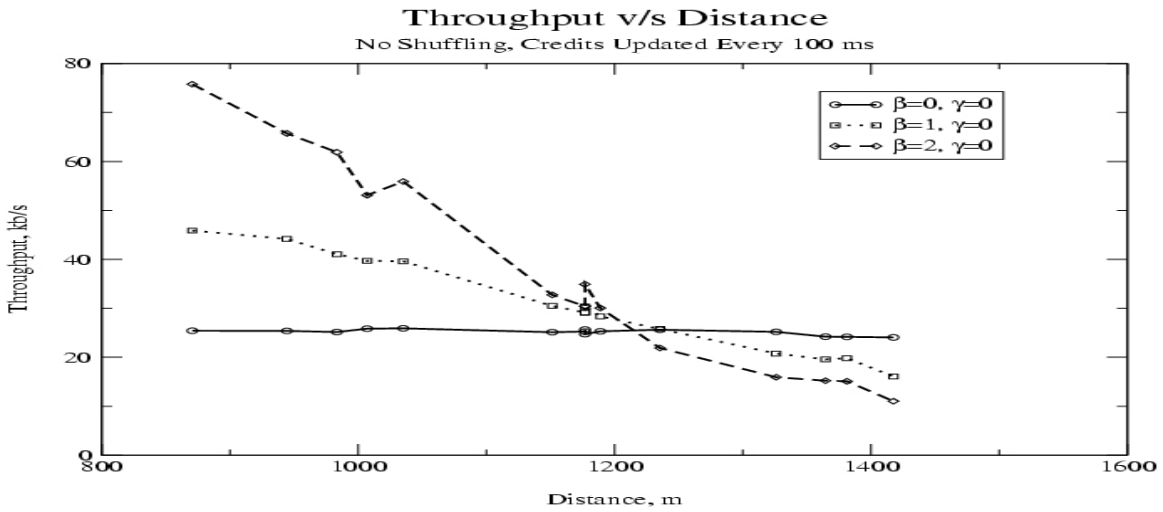

Figure 2. Typical rate-distance profile for Variant 1 with $\gamma = 0$.

is based on both the current and average channel conditions whereas for $\beta = 1$ it is based only on the current conditions. In Table 2 this case is listed under "Variant 1- shuffling." Figures 3 and 4 display the same quantities as in the earlier two figures respectively.

For our final set of experiments we consider Variant 3 of the scheduling algorithm with $\gamma = 1$ for $\beta = 1, 2$. Variant 3 with $\beta = \gamma = 1$ is similar to the algorithm proposed in [1,3,2]. An important difference is that the algorithm in [1,3,2] allows for only one user per frame. With maximum power constraints or maximum rate constraints this would be wasteful of power. They also concentrated on a CDMA2000 system with 3.33ms frames. In our simulation of their algorithm we allow for many users to be scheduled during each frame and, in addition, use an IIR filter estimate of $\bar{W}$. This falls under the case of $\beta = 1$ and $\gamma = 1$. In Table 2 this set of experiments is listed under "Variant 3." Figures 5 and 6 once again plot the same quantities as in Figures 1 and 2, respectively.

We see from Table 2 that the system throughputs increase dramatically with the knowledge of channel conditions. It is interesting to note that with the knowledge of current channel conditions the throughput of the worst user (in terms of distance) also improves. Thus, a scheduler that does not use the knowledge of channel conditions has very poor performance. We expect Variants 2 and 3 to outperform Variant 1. A close inspection of Variant 1 reveals that the current channel conditions are only reflected for users who get scheduled. On the other hand for the other two variants the relative gains of all the users are accounted for in each scheduling step. It is interesting to note that Variant 3 performs the better than Variant 2. Note that all three variants of the algorithm rely on an IIR filter for $\hat{R}_{\text{av}}$ and $\bar{W}$ calculation. The time constants in those updates need to be chosen in a way that best captures the other time-scales in the system. This needs to be quantified by further analysis and simulation.

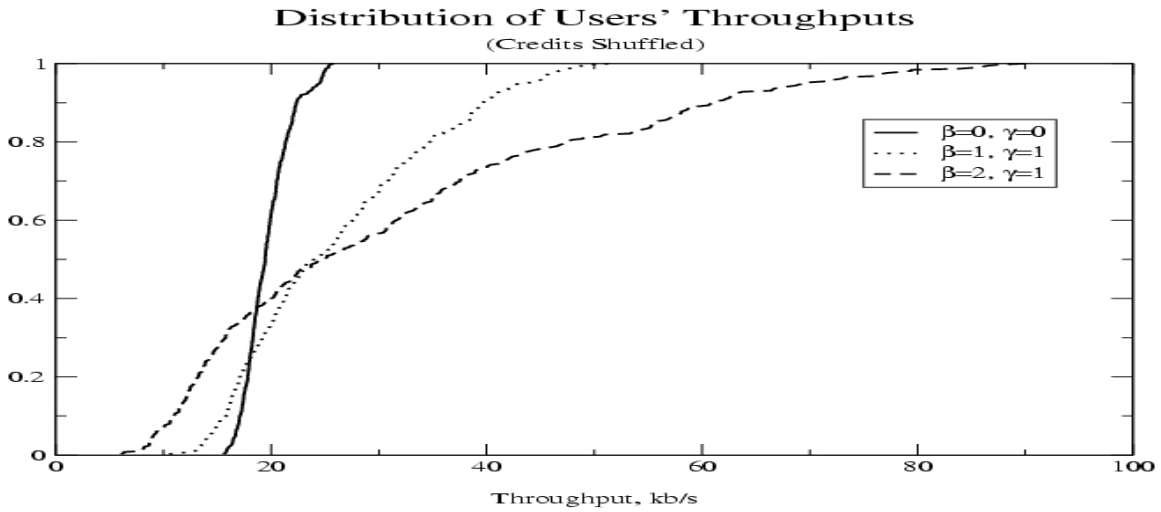

Figure 3. Distribution of individual user throughputs for Variant 1 with $\gamma = 1$ for $\beta = 1, 2$ and $\gamma = 0$ for $\beta = 0$.

6. CONCLUSION

We have proposed a scheduling algorithm that provides a flexible means of trading off efficiency for fairness as well as a flexible way of exploiting temporary fluctuations in channel conditions. The efficiency-fairness trade-off is based on a utility optimization with appropriate choices of utility functions and thus β parameters. The exploitation of the variation in channel conditions is based on biasing the algorithm with the γ parameter, in favor of users with better relative current channel condition. The analysis shows that with an appropriate choice of utility function a substantial gain in system throughput can be achieved while maintaining reasonable fairness amongst the users. The performance can be further improved by using the current channel conditions in the scheduling decisions.

Acknowledgement

We thank Eric Villier, Peter Legg, and Stephen Barrett for their help in developing the software used to obtain the simulation results.

REFERENCES

1. J. M. Holtzman, CDMA Forward Link Waterfilling Power Control, in: Proceedings of IEEE VTC 2000, 2000.
2. A. Jalali, R. Padovani, P. Rankaj, Data Throughput of CDMA-HDR a High Efficiency - High data Rate Personal Communication Wireless System, in: Proceedings of IEEE VTC 2000, 2000.
3. D. Tse, Forward-Link Multiuser Diversity Scheduling, *submitted for publication to IEEE JSAC*.

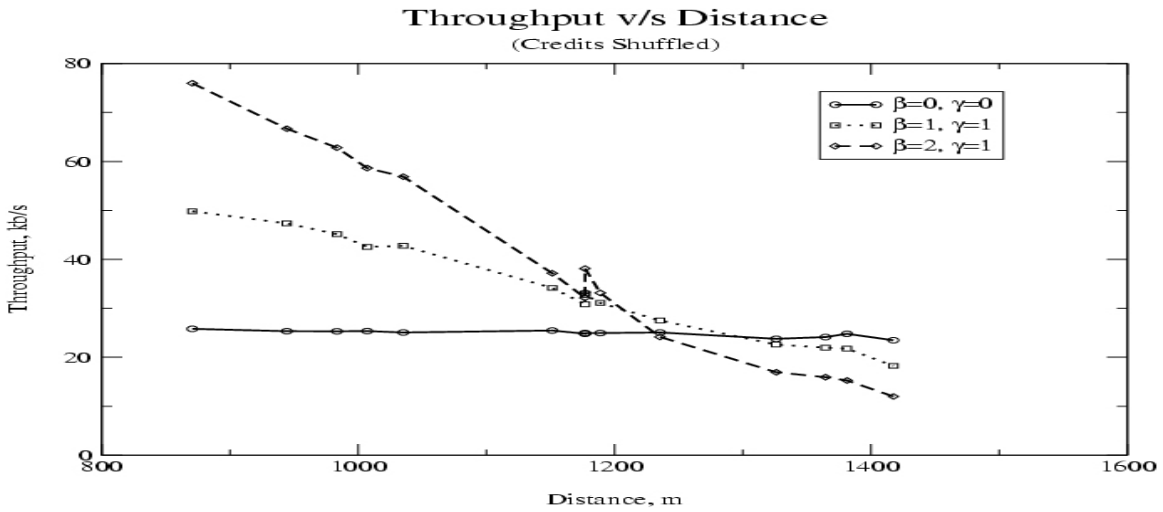

Figure 4. Typical rate-distance profile for Variant 1 with $\gamma = 1$ for $\beta = 1, 2$ and $\gamma = 0$ for $\beta = 0$.

4. S. Shakkottai, A. Stolyar, Scheduling for Multiple Flows Sharing a Time-Varying Channel: the Exponential Rule, *Preprint* (January 2001).
5. X. Qui, K. Chawla, L. Chuang, N. Sollenberger, J. Whitehead, RLC/MAC Design Alternatives for Supporting Integrated Services over EGPRS, IEEE Personal Communications Magazine 7 (2) (2000) 20–33.
6. R. Leelahakriengkrai, Scheduling in Multimedia CDMA Wireless Networks, Ph.D. thesis, University of Wisconsin, Madison,WI (2001).
7. R. Berry, Power and Delay Trade-offs in Fading Channels, Ph.D. thesis, MIT, Cambridge,MA (2000).
8. S. Nanda, K. Balachandran, S. Kumar, Adaptation Techniques in Wireless Packet Data Services, IEEE Communications Magazine .
9. R. 1633, Integrated Services in the Internet Architecture: an Overview (June 1994).
10. R. 2475, An Architecture for Differentiated Service (June 1999).
11. R. 2597, Assured Forwarding PHB Group (June 1999).
12. F. Kelly, A. Maulloo, D. Tan, Rate control in communication networks: shadow prices, proportional fairness and stability, Journal of the Operations Research Society 49 (1998) 237–252.
13. R. Agrawal, A. Bedekar, R. La, V. Subramanian, Analysis of a class and channel condition based weighted proportional fair scheduler, *Document in preparation*.
14. F. Kelly, Charging and rate control for elastic traffic, European Transactions on Telecommunications 8 (1997) 33–37.
15. R. J. La, V. Anantharam, Charge-sensitive TCP and rate allocation in the Internet, in: Proceedings of IEEE INFOCOM 2000.
16. W. C. Jakes, Microwave Mobile Communications, Wiley-Interscience, 1974.

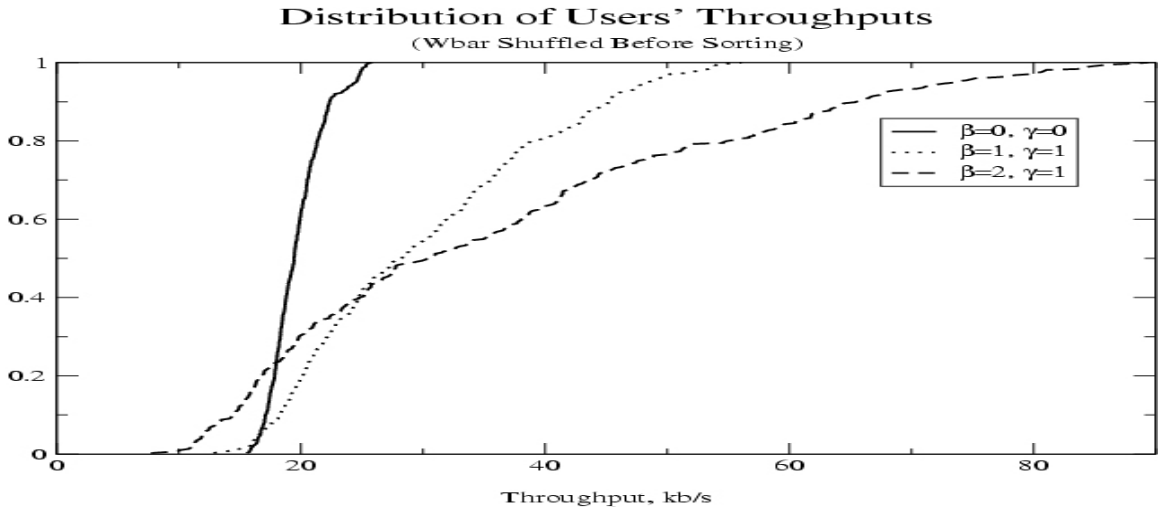

Figure 5. Distribution of individual user throughputs with $\gamma = 1$ for $\beta = 1, 2$ and $\gamma = 0$ for $\beta = 0$.

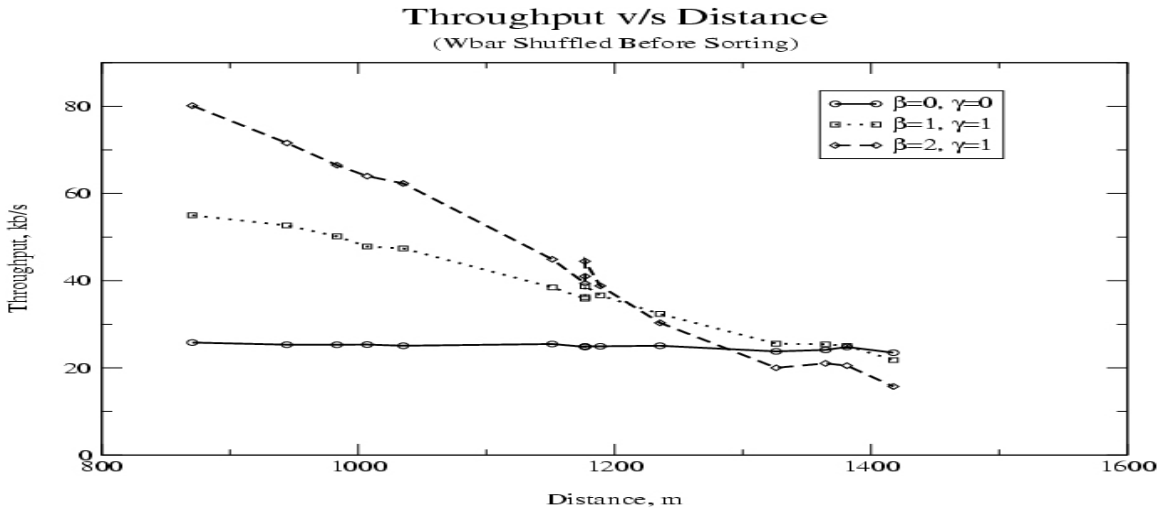

Figure 6. Typical rate-distance profile with $\gamma = 1$ for $\beta = 1, 2$ and $\gamma = 0$ for $\beta = 0$.