Channel-Quality Dependent Earliest Deadline Due Fair Scheduling Schemes for Wireless Multimedia Networks

Ahmed K. F. Khattab
ahmedkhattab@eng.cu.edu.eg

Khaled M. F. Elsayed, Senior Member, IEEE
khaled@ieee.org

Department of Electronics and Communications Engineering, Cairo University, Giza, Egypt 12613

Abstract— Providing delay guarantees to time-sensitive traffic in future wireless multimedia networks is a challenging issue. This is due to the time-varying link capacities and the variety of real-time applications expected to be handled by such networks. We propose and evaluate the performance of two channel-aware scheduling schemes that are capable of providing such delay guarantees in wireless networks. In the first proposed scheme, the Channel-Dependent Earliest-Due-Date (CD-EDD) discipline, the expiration time of the head of line packets of users' queues is taken into consideration in conjunction with the current channel states of users in the scheduling decision. This policy attempts to guarantee the targeted delay bounds in addition to exploiting multiuser diversity to make best utilization of the variable capacity of the channel. In the second scheme we attempt to ensure that the number of packets dropped due to deadline violation are fairly distributed among the users. The proposed schemes can provide statistical guarantees on delays, achieves high throughput, and exhibits good fairness performance with respect to throughput and deadline violations. We provide extensive simulation results to study the performance the proposed schemes and compare it with two of the best known scheduling schemes [16,17] in the literature.

Index Terms— Fairness, multiuser diversity, QoS provisioning, scheduling, wireless networks.

I. INTRODUCTION

The rapid growth of wireless technology, when coupled with the explosive growth of the Internet, has increased the demand for wireless data services. Traffic on beyond 3G wireless networks is expected to be a mix of real-time traffic such as voice, multimedia teleconferencing, and games, and data-traffic such as WWW browsing, messaging and file transfers. All of these applications will require widely varying and very diverse quality of service (QoS) guarantees for the different types of offered traffic, and we are now in the early days of this eventual amalgamation. Various scheduling disciplines have been developed in order to guarantee certain required QoS over wireline networks. However, these service disciplines, such as Weighted Fair Queuing (WFQ), virtual clock, and Earliest-

Due-Date First (EDD) [1], are not directly applicable in wireless networks because they do not consider the characteristics of wireless channel. These characteristics include high error rate, bursty errors, location-dependent and time-varying wireless link capacity, scarce bandwidth, user mobility, and power limitation of the mobile hosts.

All of the above characteristics make developing efficient and effective scheduling algorithms for wireless networks very challenging. Recently there has been increased interest in protocols for wireless networks which rely on significant interactions between various layers of the network stack. Generically termed cross layer design, many of these proposals are aimed at achieving performance improvements. For example, if wireless scheduling performed based on the physical layer information (users' channel states), the efficiency of the wireless system in utilizing the system resources increases. This idea was firstly exploited by Knopp and Humblet [2] when they have introduced a new diversity scheme. This diversity, termed multiuser diversity, is inherent in a wireless network with multiple users sharing a time-varying channel. Multiuser diversity comes from the fact that different users usually have independent channel gains for the same shared medium (e.g. downlink). With multiuser diversity, the strategy of maximizing the total Shannon (ergodic) capacity turns out to be a greedy scheduling rule where the scheduler allows at any time slot only the user with the best channel to transmit [3]. Results in [4] have shown that such scheduling technique can increase the total (ergodic) capacity dramatically, in the absence of delay constraints, as compared to the traditionally used scheduling techniques.

One problem with such greedy scheduling is the unfairness in resource sharing between users in the network. This is due the fact that the user with the best channel conditions will always receive the biggest share of network resources; while the user suffering from bad channel conditions will not be able to be served. The research reported in [5-10] was concerned with the problem of achieving throughput and/or temporal fairness among users. These works report on scheduling algorithms that attempt to guarantee that the difference in the services obtained by users with different channel conditions are as close as possible either on short-term basis and/or on long-term basis. However, these schemes provide no delay guarantees and thus are not suitable for delay-sensitive applications, such as voice and video.

In this paper we propose a scheduling scheme based on Earliest-Due-Date First (EDD) that exploits multi-user diversity and can provide statistical guarantees on delays, achieves high throughput, and exhibits good fairness performance with respect to throughput and deadline violations.

The rest of this paper is organized as follows: is section II we describe the model of the system under consideration. Then in section III, a survey of scheduling delay-sensitive traffic in wireless network is presented. In section IV, we propose two new scheduling policies and describe their operation principles. In section V, a number of simulation experiments are carried out to investigate the performance of these disciplines. Section VI summarizes the main findings of this paper.

II. SYSTEM MODEL

We first describe the cellular wireless network model used, and more specifically the downlink of such a network. A base station transmits data to N mobile terminal users, each of which requires certain QoS guarantees. In cell-structured wireless networks, the service area is divided into cells, and each is served via a base station. A single cell is considered in which a centralized scheduler at the base station controls the downlink scheduling, whereas uplink scheduling uses an additional mechanism such as polling to collect transmission requests from mobile terminals [5], [6]. We assume that downlink and uplink transmission don't interfere with each other.

We consider a time slotted system, where time is the resource to be shared among users. A time-slotted cellular system can have more than one channel (frequency band), but at any given time, only one user can occupy a given channel within a cell. Here, we focus on the scheduling problem for a single channel over which a number of users could be time-division multiplexed. Time division multiple access (TDMA) systems divides the time into time slots of length T_s, during which data transmission of a single user, the scheduled user, is carried out using all resources available to the base station at that time instant. For downlink scheduling, packets destined to different users are put in separate queues, one corresponding to a user's data flow.

The time varying channel conditions of wireless links are related to three basic phenomena: fast fading on the order of milliseconds, shadow fading on the order of tens to hundreds of milliseconds, and finally, long time-scale variations due to user mobility. The channel fading processes of the users are assumed to be stationary, ergodic and independent of each other, and we also assume that the channel gains are constants over one time slot's duration. Since our algorithm will exploit the users' channel conditions in making the scheduling decision, we consider wireless systems with mechanisms to make predicted channel conditions available to the base station as is commonly the case with technologies such as HDR [11], UMTS-HS-DPA [12], (E)GPRS [13], etc. The particular mechanism employed by a system depends on the communication standard. For example, in HDR and UMTS-HS-DPA, the underlying physical channel uses explicit channel notification so that the scheduler has the best possible knowledge about the channel conditions. In UMTS-DCH, there is a logical control channel assigned with every user that allows a coarse estimation of the channel condition. The packet extensions (E)GPRS to GSM-TDMA systems offer various coding schemes to support data transmission over a wide range of channel conditions. These are typically switched on a slower timescale, e.g., based on experienced frame error rates. Regardless, the recently selected coding scheme that determines the 'throughput per RLC-block' can serve as a coarse indicator of the channel condition for the scheduler. In general, the faster and more precisely the channel quality can be predicted, the better the scheduler can incorporate this information into its decision as to which user to schedule next [5]. Thus, we will assume that base station has the current (or delayed) channel state information of each user.

Figure 1 shows the architecture for channel-state aware scheduling of multiuser traffic over a fading time slotted wireless channel. The scheduler makes a decision to serve a particular queue at the beginning of every time slot. This decision could depend on HoL packet delay information, such as its waiting time and its time to expire, as well as channel states. Once a decision has been made, the chosen queue is serviced in that slot at the maximum possible rate corresponding to the state of its channel.

A very important and challenging problem in the design of high speed communications networks is that of providing Quality of Service (QoS) guarantees, usually specified in terms of rate guarantees, loss probabilities or delays of packets in the network. The control of delays is often of crucial importance, especially for real-time applications such as audio and video streaming. Real-time traffic classes are modeled as a stream of packets, with each packet having an expiry time beyond which the packet is of no use to the end user. The objective of the scheduler is to transmit each packet before its expiry, and if this is not possible, to minimize the number of lost packets due to deadline expiry. Expiry occurs when a packet have been waiting in the base station queue for a time greater than its deadline without being served. Such a packet is dropped by the system.

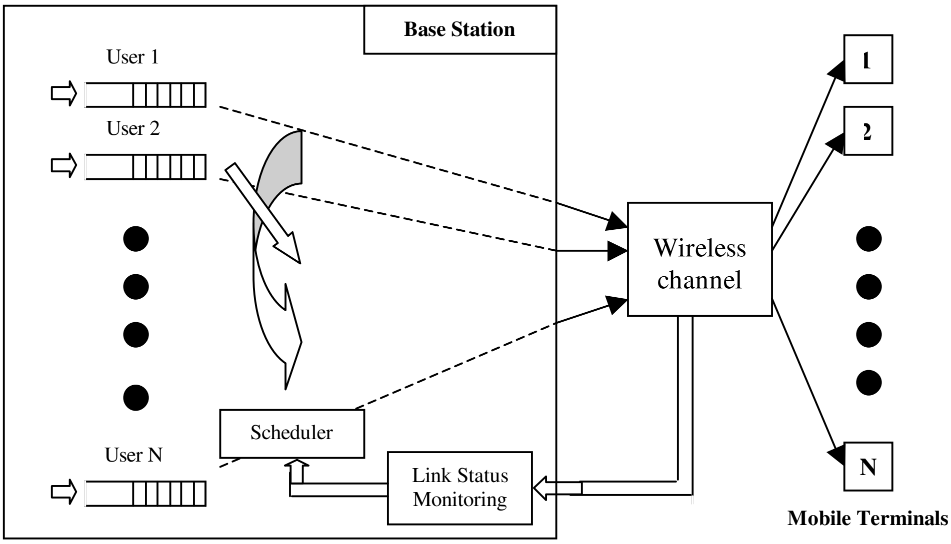


Figure 1. Channel state dependent downlink scheduling architecture for multiple users sharing a wireless TDMA channel.

Such QoS requirements can be specified in terms of deadline T_i (deterministic QoS requirements) or accompanied with allowed violation probability δ_i (statistical QoS) for each user traffic flow. In this paper, we will use the following model in defining QoS requirement of user i

$$\mathrm{P}(W_i > T_i) \leq \delta_i \tag{1}$$

for $i = 1, 2, \ldots, N$, where W_i to be the delay encountered by user i packets. Under this constraint, the problem is to minimize the violation occurrences.

III. PREVIOUS WORK

In this section, we present a survey of existing scheduling disciplines applicable in wireless networks with delay-sensitive users' traffic. In [14], the authors proposed to apply the wireline EDD discipline scheduling in wireless networks. The scheme is called Feasible Earliest Due Date (FEDD) scheduling. They assumed a simple channel state model in which the channel can be either "good" or "bad". Thus, FEDD scheduling chooses to schedule the packet which has the earliest time to expiry from the set of queues whose channels are marked good only. Such a scheduler is similar to the wireline one. Wireline EDD was proved to be optimal for i.i.d. Bernoulli arrival processes. This algorithm showed unfairness in throughput sharing, since the user suffering from long periods of bad state will not be compensated by any mechanism, thus the authors suggested to use a rate-proportional scheduler which provides guaranteed minimum bandwidth to each connection and distributes the residual bandwidth using the FEDD criterion.

The authors in [16] proposed a modification to the Largest Weighted Delay First (LWDF) scheduling discipline that take the time varying characteristics of wireless channels into account. The LWDF [15] discipline is a parameterized version of the first-input first-output (FIFO) that works as follows: at the beginning of the time slot starting at time t, serve at the maximal possible rate the queue of user j, where

$$j = \arg\max_i \{a_i W_i(t)\} \tag{2}$$

where $W_i(t)$ is the waiting time of head of line (HoL) packet of the i^{th} user at the time slot starting at time t, and $a_i > 0$, $i = 1,2, \ldots, N$, are a fixed set of constants. If the delay QoS requirements for all users is as expressed by (1), it was proved in [15] that the choice of weights a_i that makes LWDF discipline nearly throughput optimal (note however that this choice of weights is valid only for large values of the delay bound T_i and very small values of δ_i) is:

$$a_i = -\log \frac{\delta_i}{T_i} \tag{3}$$

The proposed modification of [16] was to use multiuser diversity in order to increase the efficiency of channel utilization (and hence the system throughput) and also compensate delayed users. The proposed Modified Largest Weighted Delay First (M-LWDF) discipline schedules the j^{th} user, where

$$j = \arg\max_i \{\gamma_i \mu_i(t) W_i(t)\} \tag{4}$$

where $\mu_i(t)$ is the state of the channel of user i at time t, i.e. the actual rate supported by the channel. This rate is assumed to be constant over one slot. It had been proven in [16] that setting $\gamma_i = a_i / \overline{\mu}_i$, where a_i is given by (3) and $\overline{\mu}_i$ is the mean rate supported over the i^{th} channel, makes the M-LWDF discipline throughput optimal. In practice, the mean rate can be measured over a certain, but relatively long, time window [4] by averaging the rate actually given to that user in that window. The

M-LWDF scheduling rule could be rewritten as schedule the j^{th} user, where

$$j = \arg\max_i \{a_i \frac{\mu_i(t)}{\overline{\mu}_i} W_i(t)\} \tag{5}$$

The scheme provides good QoS for delay sensitive users only with properly chosen parameters and can be easily implemented. It was shown in [16] that M-LWDF rule is optimal in the sense that it can handle all the offered traffic and renders the stability of all queues if this is feasible to any other rule, i.e. has the largest stable admission region, such a rule is called a throughput optimal one. The authors in [16] also showed how M-LWDF can be used to achieve alternative QoS defined in terms of a predefined minimum long-term throughput for each user. Unfortunately, M-LWDF scheduling was found to be highly dependent on the value of the parameters a_i, and its performance change significantly with the QoS requirements of users' flows.

In order to reduce the dependency of the M-LWDF rule on the settings of the parameters a_i, the authors in [16] also proposed a new scheduling discipline, which was further investigated and implemented in [17] for CDMA/HDR system and modified in [18]. The proposed scheme was called the exponential rule scheduling discipline. This scheme schedules the j^{th} user at the time slot starting at time t for transmission, where

$$j = \arg\max_i \{a_i \frac{\mu_i(t)}{\overline{\mu}_i} e^{\frac{a_i W_i(t) - \overline{aW}}{1+\sqrt{\overline{aW}}}}\} \tag{6}$$

and

$$\overline{aW} = \frac{1}{N}\sum_{i=1}^{N} a_i W_i(t) \tag{7}$$

This policy attempts to equalize the weighted delays $a_iW_i(t)$ of all the queues when their differences are large. If one of the queues would have larger (weighted) delay than the others by more than order $\sqrt{\overline{aW}}$, then the exponent term becomes very large and overrides the channel considerations (as long as its channel can support a non-zero rate), hence giving priority to that queue. (It can be easily noticed that the $\overline{aW}$ term in the exponent can be dropped without changing the rule since it is common for all queues. This term is present only to emphasis the motivation of the rule.)On the other hand, for small weighted delay differences (i.e. less than order $\sqrt{\overline{aW}}$), the exponential term is close to unity, and the policy behaves as the proportionally fair rule. Hence, the exponential rule policy gracefully adapts from a proportionally fair one to one which balances delays [16], [17]. It was proven in [19] that the exponential rule is a throughput optimal policy for general assumptions on the channel and arrival processes. Moreover, simulation results in [17] showed that the exponential rule scheduling exhibits better delay tails compared to any other scheduling policy in the sense that the delays of all users are about the same and are all reasonably small. This occurs, however, for large values of T_i and very small values of δ_i, which is not desired practically. Moreover, as in the M-LWDF discipline, the exponential rule scheduling was found to be highly dependent on parameter settings. Therefore, It was advised that identifying good scheduling rules which are less dependent on the "proper" parameter setting would be desirable [16].

IV. THE PROPOSED SCHEDULING SCHEMES

The goal of our work is to design scheduling schemes for delay sensitive traffic that exhibits “good performance” and that exploits multi-user diversity inherent with wireless communications. From the above study of such a problem, we mean by the word "good performance" that such a discipline should attempt to achieve the following objectives:

1) Maximize the overall system throughput: This could be easily achieved if the scheduling discipline utilizes the multiuser diversity, inherently existing in systems under consideration. Multiuser diversity efficiently utilize the channel capacity by giving higher priority to the user with the best channel conditions at a certain time instant, which means that this user can transmit with the highest possible rate, and thus increase the system throughput.

2) Graceful compensation of large delays: For real-time traffic packet, it is necessary that the delivery of such a packet be done within its deadline, otherwise, the information contained in this packet will be irrelevant for the receiver. Thus, a good scheduling discipline should have a mechanism to compensate queues whose packets are experiencing long delays in the system in order to prevent their packets from being lost. This case may be encountered by users far from the base station, which yields that their channels are suffering from long periods of bad conditions. Such a mechanism will guarantee the QoS requirements if defined as delay bound or packet loss ratio. Thus it minimizes the number of packets dropped due to deadline violation, which in turn increases the system throughput.

3) Fairness in resource sharing: Fairness is an intuitively desirable property of scheduling disciplines. A fair scheduling discipline should distribute the resources available to the system, such as capacity and time, fairly among different users. Fairness may be accomplished in delay distributions, service rates, number of packets lost, etc.

4) Weak dependency on the parameters setting: As advised in [16], it is desirable to identify good scheduling disciplines which are less dependent on the proper parameter setting. In other words their performance does not change significantly for wide range of QoS requirements, and could thus be employed in systems serving wide range of applications.

A. Channel Dependent Earliest Due Date (CD-EDD) Scheduling Discipline

In classical wireline earliest due date scheduling, each packet is assigned a deadline, and the scheduler serves packets in order of their deadlines. The queue with the smallest deadline is served first by the maximum available rate. If the scheduler is overcommitted, then some packets miss their deadlines. This policy was shown to be optimal in the wireline case (for independent identically distributed Bernoulli arrival and channel processes). EDD cannot be efficiently employed in wireless networks since it does not consider the time varying characteristics of wireless links. It was not reported in the literature the existence of a scheduling discipline that combines the EDD scheduling concept with a mechanism to adapt with the characteristics of wireless networks (with the exception of the attempt in [14] which does not actually adapt with the time varying nature of wireless channel, since it assumed the simplified channel model of good or bad).

We propose a new scheduling discipline, which we call the channel dependent earliest due date first (CD-EDD) policy. This is basically a channel-state dependent EDD policy where the scheduler chooses to schedule the queue whose HoL packet has the earliest time to expire and the best channel conditions, and consequently the highest transmission rate, among all queues. The proposed CD-EDD scheduling policy is as follows:

At the time slot starting at time *t*, schedule with the maximum possible rate the queue of the j^{th} user, where

$$j = \arg\max_i \{ a_i \frac{\mu_i(t)}{\overline{\mu}_i} \frac{W_i(t)}{d_i(t)} \} \tag{8}$$

where

a_i is the weighting parameter reflecting the statistical QoS requirements of the i^{th} user.

$\mu_i(t)$ is the actual rate that could be used for transmission by the i^{th} user at time *t*, which reflects the current channel state of the user's channel.

$\overline{\mu}_i$ is the mean rate supported or previously offered to the i^{th} user.

$W_i(t)$ is the delay experienced by the HoL packet since its entrance to the i^{th} user queue in the base station.

$d_i(t)$ is the time to expire of the i^{th} user HoL packet, which is the difference between the deadline, T_i, and the delay experienced till time *t*, $W_i(t)$, i.e.

$$d_i(t) = T_i - W_i(t) \tag{9}$$

The behavior of the CD-EDD policy can be explained as follows: when a certain queue has its HoL packet waiting in the system for a relatively long period (but have not expired yet), its time to expire will decrease significantly. in such a situation, the term $W_i(t)/d_i(t)$ will grow significantly due to the contribution of $1/d_i(t)$ until it overcomes other terms in (8). This has an effect akin to reducing the number of dropped packets due to deadline violation. On the other hand, if the delay characteristics of all users are about the same, i.e. their time to expire and waiting times are close, the term $W_i(t)/d_i(t)$ will be common to all users, and the policy then reduces to a proportionally fair one that exploit multiuser diversity to efficiently utilize the channel bandwidth of multiuser systems in a fair manner. It is worth mentioning that weights a_i doesn't contribute significantly in the decision. A rule of thumb for choosing a_i which works in practice is the one given in (3) since this choice is suggested by large deviations of optimality results.

In other words, the CD-EDD is a scheduling discipline that can be used to provide QoS guarantees, defined in terms of delay bounds, for real-time traffic in wireless networks. This is achieved by increasing the priority of delayed users to get access to the medium over time. An important feature of the CD-EDD policy is its weak dependency on the value of QoS required, and thus can be used for a wide variety of QoS requirements.

B. A Set of Violation-Fair Rules

Another new idea than can be applied in conjunction with any scheduling discipline in order to enhance the fairness characteristics of these policies is proposed here, which is based on the number of deadline violations occurring to packets of different queues. This requires that each queue in the base station to be accompanied with a counter that counts the number of packets lost in this queue's flow. This may be implemented practically be means of sliding windows basis. Let us define

$NV_i(t)$ to be the number of deadline due date violations encountered in the flow of the i^{th} user up to time t.

$\overline{NV(t)}$ to be the average of the number of violations in all N queues, i.e.

$$\overline{NV(t)} = \frac{1}{N}\sum_{i=1}^{N} NV_i(t) \tag{10}$$

The scheduler may use the number of deadline violations to find a way to compensate users suffering from unfairness in the number of dropped packets. For example the scheduler could give more credit or increase the priority level so that such a user could access the system resources. This could be achieved be a scheduling discipline that uses a term like $NV_i(t)/\overline{NV(t)}$ in making the scheduling decision. We call such a scheduling discipline a violations-fair (VF) discipline.

Initially, we applied this idea directly to the proportionally fair [4] scheduling discipline, yielding the violations-fair proportionally fair policy. So, in each time slot the scheduler chooses the j^{th} user, where

$$j = \arg\max_i \{\frac{\mu_i(t)}{\overline{\mu_i}} \frac{NV_i(t)}{\overline{NV(t)}}\} \tag{11}$$

This reduces the number of packets lost for users with bad channel conditions, therefore, enhancing the fairness characteristics of the proportionally fair policy. On the other hand, it still lacks a mechanism for provisioning of QoS guarantees for delay sensitive traffic.

We further apply the proposed modification to both the M-LWDF and the exponential rule policies. In the violations-fair modified largest weighted delay first (VF-M-LWDF) discipline, the scheduling decision is such that at the time slot starting at time t, schedule with the maximum possible rate the queue of the j^{th} user, where

$$j = \arg\max_{i}\{a_i \frac{\mu_i(t)}{\overline{\mu_i}} \frac{NV_i(t)}{\overline{NV(t)}} W_i(t)\} \tag{12}$$

In the violations-fair exponential (VF-EXP) discipline, the scheduling decision is such that at the time slot starting at time *t*, schedule with the maximum possible rate the queue of the j^{th} user, where

$$j = \arg\max_{i}\{a_i \frac{\mu_i(t)}{\overline{\mu_i}} \frac{NV_i(t)}{\overline{NV(t)}} e^{\frac{a_i W_i(t) - \overline{aW}}{1+\sqrt{\overline{aW}}}}\} \tag{13}$$

The proposed modification enhance their performance since the addition of the violations-fair term will ensure fairness in both the delay times and throughput. This could be explained since both the M-LWDF and the exponential disciplines minimize the packet delay, and when the number of dropped packets is fairly distributed among the users, the long term service rate will be equal for all users. Another very important gain of the violations-fair version of these discipline, is that their performance is not much dependent on the parameter setting as was the case in the original rules (this will be seen in the simulation results).

Finally, if the proposed violations-fair technique is applied on the proposed CD-EDD, we can get a scheduling discipline applicable in wireless network that explicitly provide QoS to delay sensitive traffic, with excellent fairness characteristic with respect to data rate, delay bound, and delay bound violation. The violations-fair –channel-dependent earliest deadline due date (VF-CD-EDD) scheduler chooses, at the time slot starting at time *t*, the j^{th} user for transmission, where

$$j = \arg\max_{i}\{a_i \frac{\mu_i(t)}{\overline{\mu_i}} \frac{W_i(t)}{d_i(t)} \frac{NV_i(t)}{\overline{NV(t)}}\} \tag{14}$$

In the next section, we provide an extensive set of simulation that explore the performance of the proposed CD-EDD discipline compared with the M-LWDF and the exponential rule disciplines as a reference. We also show the advantages achieved by their violations- fair versions, namely the VF-CD-EDD, the VF-M-LWDF, and the VF-exponential scheduling disciplines.

V. PERFORMACE EVALUATION

A. Simulation Setup

First, we describe the system model used in simulating the wireless cell-structured channel-aware scheduler described in section II. We chose the High Data Rate (HDR) CDMA system model. HDR technology has recently been proposed as a TDM-based overlay to CDMA with the goal of providing packet data services to mobile users. HDR is a downlink packet data service that occupies a single data carrier of a CDMA system, where users share the channel in a time division multiple access manner, i.e. at each time slot T_s only one user can transmit its data at the full power available to the base station. A very attractive feature of HDR is enabling the use of efficient scheduling algorithms since it provides a mechanism for link status monitoring as described in section II.

The cell serves N mobile users each receiving a data flow. The base station contains N queues, one corresponding to a different data flow and an associated scheduler. The scheduler makes a decision every 1.667 millisecond based on the current information available at the start of the time slot. As we are mainly interested in scheduling users with time sensitive traffic, we model the packet arrival processes to each of the N user's queues as a Bernoulli processes with a mean rate of 28.8 Kbps. This rate corresponds to the typical rate required for streaming audio over the Internet. Real time users, like streaming audio, will indeed generate a smooth traffic, and hence, a Bernoulli model seems reasonable for such traffic. Like the original EDD, the CD-EDD is expected to be throughput optimal for such traffic model. The HDR packet size is 128 bytes. The QoS requirements of each user are expressed in the form of the probability that the waiting time encountered by a typical packet of the i^{th} user stream exceeds the deadline T_i is less than or equal to δ_i . We assume for simplicity that all users require the same service quality, i.e. they all have the same T_i and δ_i.

Even thought all users share a common channel, the channel capacity, of that channel seen by different users is different. This is due to the wireless link characteristics described above. The instantaneous capacity of a wireless channel is given by

$$C(t) = B\log_2(1+|h(t)|^2\, SNR) \tag{15}$$

where $C(t)$ is the channel capacity or the data rate (in bits per second) that can be transmitted on a channel of bandwidth B (Hertz). The bandwidth of HDR/CDMA channel is 1.25 MHz. The term $|h(t)|$ is the normalized gain (or fading level) of the wireless channel at time t, and SNR is the required signal to noise ratio at the receiver antenna (13 dB for HDR/CDMA system). For simulation purposes, we use the typical HDR and cell parameters given in [11]. The average fade level distribution of a typical mobile in HDR cell can be easily found. The fading process of each user's channel can be represented by a Rayleigh process. So, in order to simulate N channels, we pick N fading levels according to the above distribution, and generate N Rayleigh processes with means equal to these fading levels after being normalized. Accordingly, the mean data rate $\overline{\mu}_i$ that can be supported on the a channel of mean fading level $|\overline{h}_i|$ is

$$\overline{\mu}_i = B\log_2(1+|\overline{h}_i|^2\, SNR) \tag{16}$$

It is worth mentioning that HDR doesn't not support arbitrary transmission rates, i.e. the scheduled user cannot transmit with the rate computed above but with the maximum possible rate from a set of discrete rates. An HDR user can transmit data at a rate of $9.6 * 2^i$ Kbps, $i = 0, 1,\ldots,$ with a maximum rate of 2 Mbps. Thus the state of channel $\mu_i(t)$ at the start of the time slot at time t will be the actual rate that the channel can support, rather than the channel capacity at the that time instant. We have assumed that the channel conditions do not change significantly within a time slot duration. Finally, all simulations will be carried out for a duration of 10 minutes.

B. *Performance Metrics*

Here, we discuss the performance metrics used to evaluate the performance of various scheduling algorithms. These are the delay, throughput, and packet loss criteria. We briefly outline them in turn.

For real-time traffic, a good measure of performance is the delays packets incur at the base station. A good scheduling algorithm should keep all delays below the delay bound T_i with high probability. The delay distribution curves can be used to illustrate the delay behavior of the scheduling disciplines under consideration. As remarked earlier, scheduling algorithms which keep the delays of all the users about the same and keep them all reasonably small are superior to those which may have better delay tails for one of the users but have very bad delays for other users. (Note that we have assumed for simplicity that all users require the same delay.)

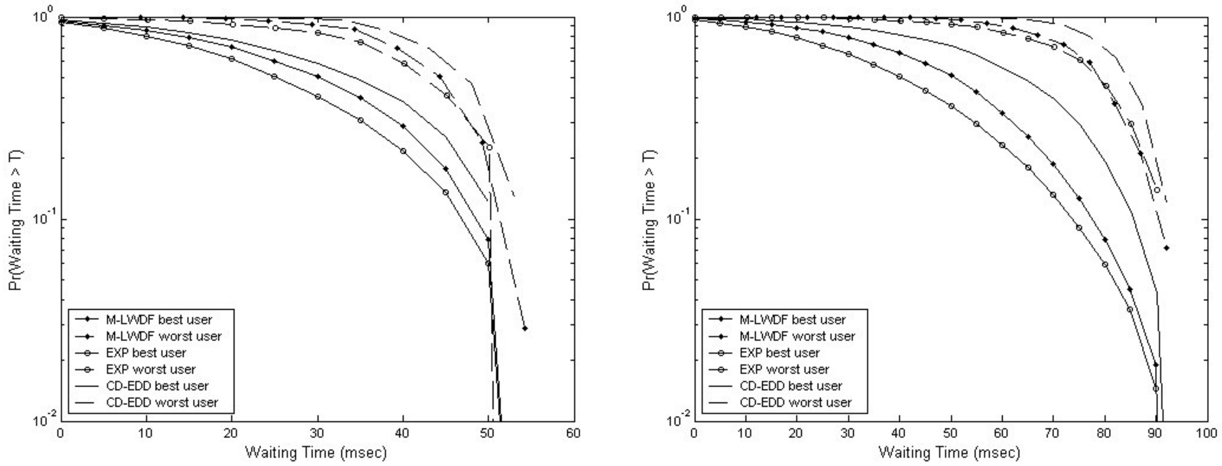
Some parameters of the delay distribution, such as the worst-case delay, the mean delay, or the 95-percentile delay, could be used to evaluate the QoS received by a user. We will consider the 95-percetile delay as our measure of the delay guarantees offered by a scheduling discipline. The 95-percentile delay is defined as the value of the delay which ninety-five percent of the users' packets experienced delays smaller than it.

Unlike non-real-time users, which may have their QoS requirements in the form of a guaranteed minimum rate, real-time users do not need the scheduler to preserve certain bandwidth for their packets' transmission. So that, we will only take the total throughput achieved be the system as a measure of the throughput performance of the scheduling disciplines at hand. The fairness of bandwidth sharing among users is also a good indication of the efficiency of any scheduling discipline.

The fraction of packets dropped, due to deadline violation, for a user can be used to evaluate the loss performance of a scheduling discipline. This fraction is required to be as small as possible in order to say that the scheduler is suitable for scheduling real-time traffic. From fairness point of view, it is better to equalize the fraction of packets lost in different queues.

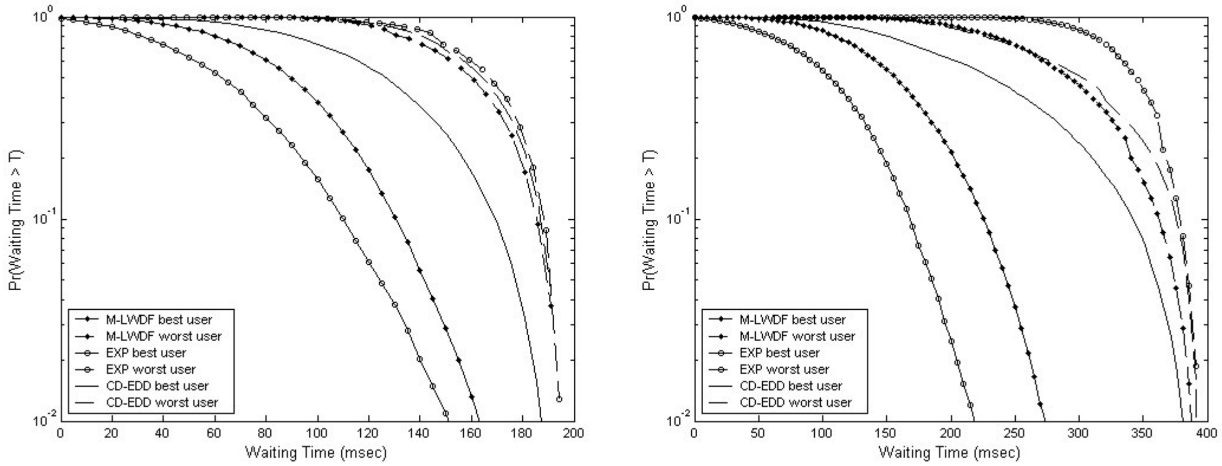
C. *Results and Discussions*

First, we estimate the number of users, with traffic like the one described above, that can be supported by a single HDR cell under each of the previously mentioned scheduling policies. In this experiment, we simulate N users, uniformly distributed throughout the cell, and monitor the average service rate received by a single user for different values of N. As N increases, and as long as the channel capacity can support such a number of users, it is expected that the service rate received by each user will be close to its arrival rate. Any further increase of the number of user, while keeping the channel capacity unchanged, will make the scheduler unable to serve such users in the appropriate time so more packets will be dropped and thus the average throughput share of each user will decrease. So, we will take the number of users beyond which the average service rate received by any



(a) Delay bound = 60 m sec.

(b) Delay bound = 100 m sec.



(c) Delay bound = 200 m sec.

(d) Delay bound = 400 m sec.

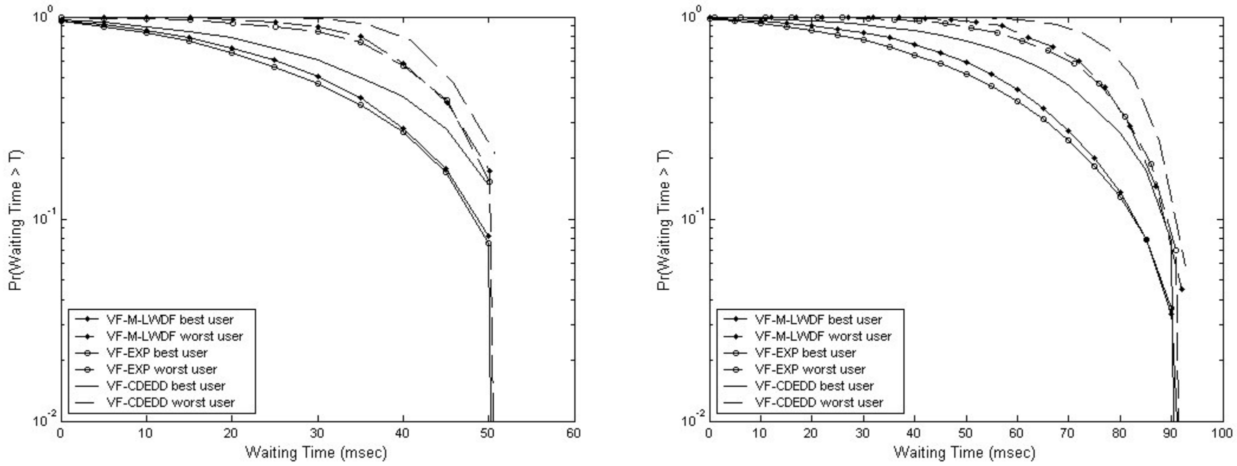
Fig. 2. Delay distributions of the best user and worst user for different delay bounds.

user in the system start to decrease as the system capacity.

Table I summarizes the simulation results of system capacity achieved by different disciplines for different values of the delay bound and for a violation probability (δ_i) of 95% (which will be used for all experiments).

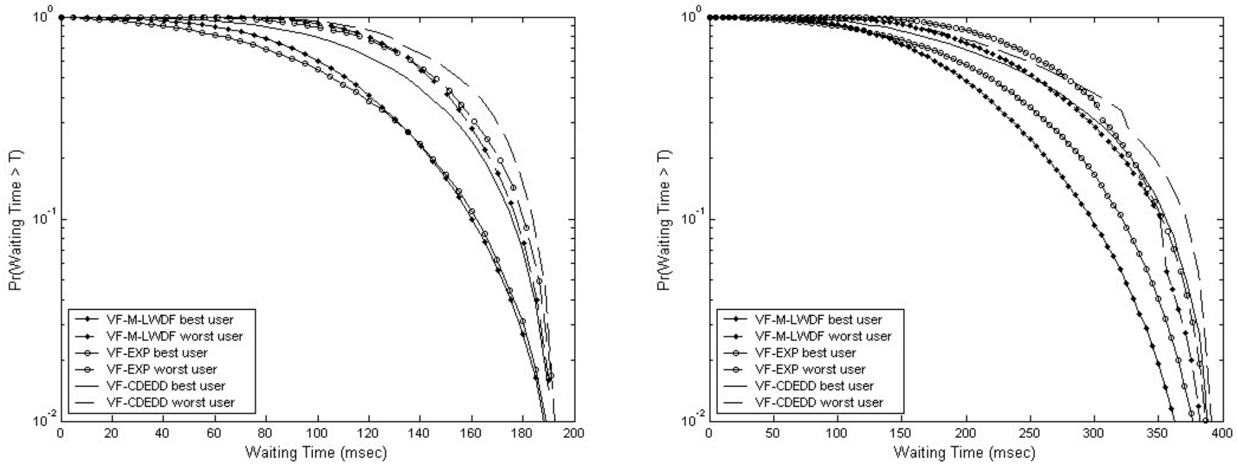
TABLE I SYSTEM CAPACITY FOR DIFFERENT DELAY BOUNDS

Delay bound (m sec)	20	60	100	200	300	400
CD-EDD	4	12	16	16	16	16
M-LWDF	4	10	12	16	16	16
EXP	4	8	12	16	16	16



(a) Delay bound = 60 m sec.

(b) Delay bound = 100 m sec.



(c) Delay bound = 200 m sec.

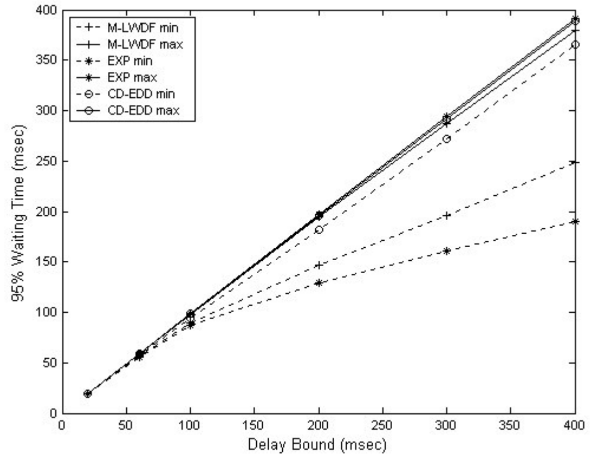
(d) Delay bound = 400 m sec.

Fig. 3. Delay distributions of the best user and worst user for different delay bounds for the violations-fair policies.

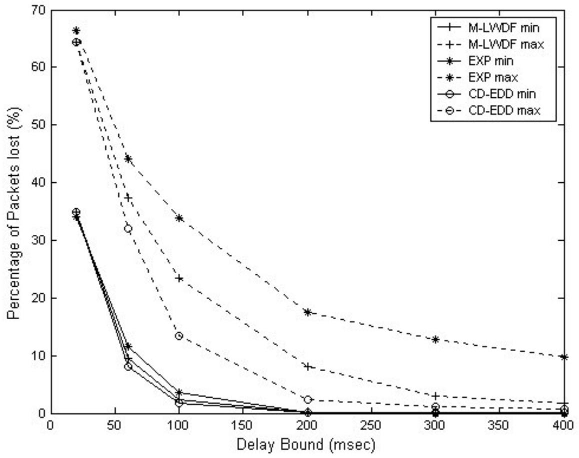
In our second experiment, we investigate the delay performance of various scheduling disciplines. We begin with a comparison of the proposed CD-EDD scheduling disciplines versus to both the M-LWDF and the exponential rule scheduling disciplines reported as the most suitable policies for scheduling delay sensitive traffic in the literature. In order to be able to evaluate the performance of the scheduling techniques, the system should be loaded with its maximum capacity. Based on the results of the first experiment in Table I, we assume that the cell is serving 14 mobile terminals, i.e. N =14, and generate 14 i.i.d. Bernoulli processes each with a men rate of 28.8 *Kbps*. Using the procedure described before, we also generate 14 Rayleigh-faded channels for the users uniformly distributed in the cell.

In Figure 2, we plot the delay distribution tails for both user 1 (with the best channel conditions) and user 14 (with the worst channel conditions) for the M-LWDF, exponential rule, and the proposed CD-EDD scheduling disciplines for delay bounds of 60, 100, 200, and 400 milliseconds.

These bounds are encountered practically in multimedia streams. It is obvious that all policies will have the same performance for very small bounds, e.g. 20 milliseconds, with such a small

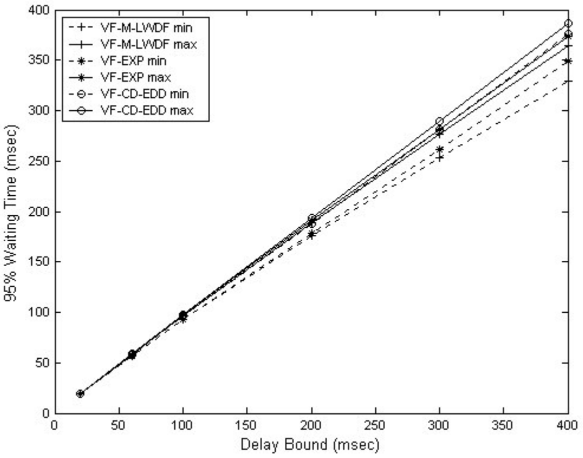


(a) The maximum and the minimum 95-percentile delays.

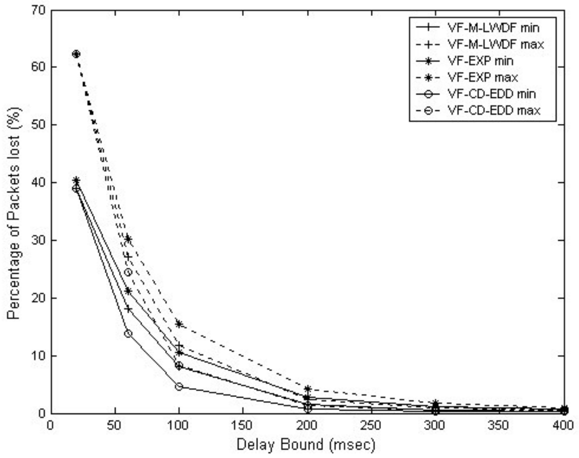


(b) The maximum and the minimum percentages of lost packets .

Fig. 4. Fairness behavior of the original disciplines



(a) The maximum and the minimum 95-percentile delays.



ɔ) The maximum and the minimum percentages of lost packets .

Fig. 5. Fairness behavior of the violation fair disciplines

deadline, the networks cannot serve such a number of users with this high QoS. We observe that for moderate delay bounds, e.g. 60 and 100 milliseconds, the performance of the exponential rule scheduling slightly outperforms both the M-LWDF and CD-EDD disciplines. In this case, all delays are kept at small value and about the same for all users. For higher bounds, e.g. 200, and 400 milliseconds, we find that the delay performance of the CD-EDD scheduler does not change significantly. On the other hand, the performance of both the M-LWDF and the exponential rule schedulers degrades severely as the gab between the tails of the best user and the worst user becomes wider. This severe degrading in performance is caused by the dependency of both rules on the quality of service required, which affects the value of the weights $\{a_i\}$ that controls the performance of these rules. (Keep in mind that the goal of the M-LWDF is to minimize the weighted delays while the exponential rule scheduling tries to keep all the delays around the average of these weighted delays ($\overline{aW}$), and they both do not target a certain dealy bound to serve as much packet as possible before this bound is violated.) On the other hand, the CD-EDD policy does not suffer from such a dependency on the QoS, and consequently the weight values, since the philosophy of this rule is mainly to serve more packets before their deadlines expire. This may cause the packets of the best channel users to experience relatively higher delays than in the case of other rules, but still lower than the worst user. This occurs on the expense of preventing more packets from being dropped due to deadline violation.

This can be further demonstrated when we plot the maximum and the minimum (corresponding to the best and worst users) 95-percentile delay and percentage of packets dropped due to deadline violations versus the delay bounds as shown in Figure 4-a and 4-b, respectively. It is clear that it is not desirable to keep one or some users' delays below a value much smaller than the required bound while leaving one or some users suffering from dropping a large percentage of their packets as the case with both the M-LWDF and exponential rule schedulers. While in CD-EDD discipline, the maximum and the minimum 95 percentile delay are about the same and so close to the delay bound, besides the packet loss ratios are very small and very close to each other. **So, the base stations can guarantee strong delay bounds for all delay sensitive users in a fair manner by using the CD-EDD scheduling discipline, regardless of the value of these bounds.**

When the same experiments were carried out for the violations-fair versions of the above disciplines, it was found that the performance of all the violations-fair policies is not much dependent on the QoS required (including the disciplines which was originally suffering from that dependency). Figure 3 shows their delay distribution tails of the best user and the worst user for different delay bounds. It is observed that the tails became much more closer for the VF-CD-EDD, but with a little bit higher delays than those of either the VF-M-LWDF or the VF-EXP policies. As shown in Figures 5-a and 5-b, where the maximum and the minimum 95-percentile delay and percentage of packets dropped due to deadline violations are plotted versus the delay bounds, the service quality offered to different users, in terms of 95% delay, is about the same. Furthermore, the amount of packets dropped in the system due to deadline violations becomes very small. Moreover, this amount is distributed among all users in a fair manner. Like the CD-EDD, the VF-CD-EDD have the superiority since it achieves the smallest number of packets lost due to deadline violations.

Finally, we study the throughput characteristics of the aforementioned scheduling disciplines. Here we are interested in studying the overall throughput of the system as well as how the throughput is divided among users with different channel conditions, i.e. fairness in throughput sharing. The results of this experiment are illustrated in Figures 6 for 100 milliseconds bounds by plotting the overall throughput of the system versus the number of users using the simulation setup of the first experiment. When the system is operating with number of users less than the system capacity, the total throughput of the system equal the sum of the arrival rate. When the system is serving more users than the system capacity, we observe that, regardless of the delay bound, the total throughput achieved using the CD-EDD is higher than any other rule because this policy causes the smallest number of packet to be lost due to deadline expiry. Moreover, the total throughput achieved using any violations-fair discipline is less than the throughput achieved with the non-violations fair counterpart. This is

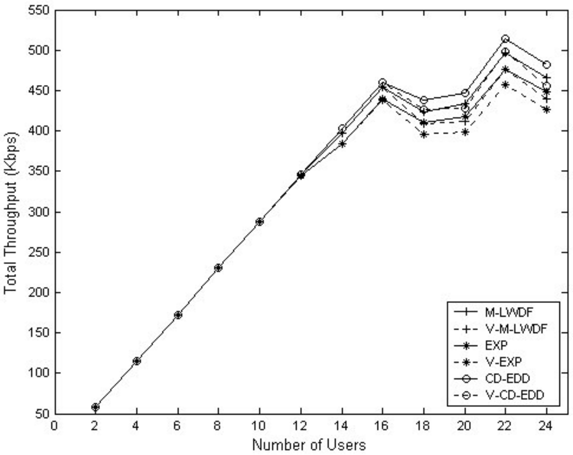


Fig. 6. The total throughput achieved by different disciplines at 100 m sec delay bound..

because in order to achieve fairness in the ratio of packets dropped among different user, the violations-fair polices may prevent users with good channel conditions from transmitting their data for the sake of users with bad channel conditions (such channels support low transmission rates only). So the overall throughput achieved by the system will be lower than the case where the scheduling policy do not intend to make users with good channel condition drop some packet for the purpose of fairness.

TABLE II THROUGHPUT FAIRNESS INDEX FOR A 100 M SEC DELAY BOUND.

	CD-EDD	**EXP**	**M-LWDF**
Original Discipline	0.8749	0.6781	0.7797
Violations-Fair Discipline	0.9656	0.9386	0.9520

In order to evaluate the throughput fairness performance of the proposed scheduling schemes, let us define the throughput fairness index as the ratio between the lowest achieved rate and the highest achieved rate. The more the fairness index approaches unity, the better the scheduling discipline in the sense that all users almost have been served with the similar rate as long as they have the same QoS. Table II list the computed values of the throughput fairness index for both the original scheduling disciplines as well as their violations-fair counterparts. We use the same simulation model used in the previous experiment where the base station serves 14 users.

Concerning the non-violations fair disciplines, it was found that the CD-EDD discipline provides the highest fairness index compared to both the exponential rule and the M-LWDF disciplines regardless the delay bound (we only include the results at 100 m sec for illustration purposes). On the other hand, even though the violations-fair disciplines leads to a slightly lower throughput per user as previously discussed, they ensure fairness in throughput sharing among all users regardless of the delay bound for the three considered violations-fair disciplines. This is because in such disciplines, the number of packets dropped due to deadline expiry is almost equal for all users. However, we note that the VF-CD-EDD has slightly better throughput fairness performance. We summarize the main results of our simulation experiments in Table III.

VI. CONCLUSIONS

This paper addresses the problem of scheduling real-time users over TDM-based wireless multimedia networks. We introduced the Channel Dependent Earliest-Due-Date first (CD-EDD) scheduling discipline, a discipline that attempts to provide statistical delay bound guarantees for time-sensitive traffic in networks with time-varying channels. Gains in throughput and realized delay are achieved by exploiting multi-user diversity techniques in which the scheduling decision takes into account the current channel state for each user in the system. By considering the packets dropping due to deadline violation, we also presented a set of scheduling policies that has satisfactory fairness characteristics in delays, throughput, and packet loss ratios among different users regardless of the value of the delay bound.

Simulation results of the proposed schemes showed that the services received by different real-time users, namely, delays, rates, and loss ratios, can be fairly achieved for a wide rang of applications. The proposed disciplines outperforms other existing polices. The computational complexity of the proposed policies are low and are suitable for application in future broadband fixed or mobile wireless systems such as 802.16a and 802.20.

TABLE III SIMULATION RESULTS SUMMARY

	CD-EDD	M-LWDF	EXP	VF-CD-EDD	VF-M-LWDF	VF-EXP
Delay bound guarantee	√	X	X	√	√	√
Independency of QoS requirement	√	X	X	√	√	√
Total system throughput	Highest	Moderate	Lowest	Less than original counterparts		
Delay fairness	√	Depend on QoS		√	√	√
Throughput fairness	Best	Good	Worst	√	√	√
Delay bound violation fairness	√	Bad	Worst	√	√	√

REFERNCES

[1] S. Kashave, *An Engineering Approach to Computer Networking: ATM Networks, the Internet, and the telephone Network.* Addison Wesley, 1997, ch. 9.

[2] R. Knopp and P. A. Humblet, “Information capacity and power control in single-cell multiuser communications,” *in Proceedings of IEEE International Conference on Communications (ICC’95),* Seattle, USA, June 1995.

[3] D. Wu and R. Negi "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *in Proceedings of IEEE ICC'03,* Anchorage, Alaska, USA, May 11-15, 2003.

[4] A. Jalali, R. Padovani, R. Pankaj "Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system," *in Proceedings of Vehicular Technology Conference 2000-spring,* vol. 3.

[5] Y. Liu, S. Gruhl, and E. Knightly. "WCFQ: An opportunistic wireless scheduler with statistical fairness bounds," *IEEE Transactions on Wireless Communication,* September 2003.

[6] Y. Liu and E. Knightly. "Opportunistic fair scheduling over multiple wireless channels," *in Proceedings of IEEE INFOCOM,* 2003.

[7] X. Liu, E. K. P. Chong, and N. B. Shroff "Transmission scheduling for efficient wireless utilization," *in Proceedings of IEEE INFOCOM,* 2000.

[8] X. Liu, E. K. P. Chong, and N. B. Shroff "Optimal opportunistic scheduling in wireless networks," *in Proceedings of IEEE INFOCOM,* 2000.

[9] S. S. Kulkarni and C. Rosenberg "Opportunistic scheduling policies for wireless systems with short term fairness constraints," *in Proceedings of IEEE GLOBECOM,* December 2003.

[10] Z. Jiang and N. K. Shankaranarayana, "Channel quality dependent scheduling for flexible wireless resource control," Globecom 2001, Texas.

[11] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, “CDMA/HDR: a bandwidth efficient high speed wireless data service for nomadic users,” *IEEE Communications Magazine,* vol. 38, no. 7, pp. 70–77, July 2000.

[12] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network, “Physical layer aspects of utra high speed downlink packet access (release 4) tr25.848 v4.0.0,” Available: http://www.3gpp.org, Mar. 2001.

[13] 3rd Generation Partnership Project; Technical Specification Group Radio Access Network, “Services provided by the physical layer (release 4) ts 25.302 v4.1.0,” Available: http://www.3gpp.org, June 2001.

[14] S. Shakkottai and R. Srikant, "Scheduling real-time traffic with deadlines over a wireless channel," *in Proceedings of ACM Workshop on Wireless and Mobile Multimedia,* Seatle, WA, August 1999.

[15] A.L. Stolyar and K. Ramanan, "Largest Weighted Delay First Scheduling: Large Deviations and Optimality," *Annals of Applied Probability,* 2001, Vol.11, No.1, pp. 1-48.

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

[17] S. Shakkottai and A.L. Stolyar, "Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR," *in Proceedings of the 17th International Teletraffic Congress - ITC-17,* Salvador da Bahia, Brazil, 24-28 September, 2001, pp. 793-804.

[18] K. Chang and Y. Han, "QoS-based adaptive scheduling for a mixed service in HDR system," PIMRC 2002.

[19] S. Shakkottai and A. L. Stolyar, "Scheduling for Multiple Flows Sharing a Time-Varying Channel: The Exponential Rule," *Analytic Methods in Applied Probability, vol.* 207, pp. 185-202., 2002.