Downlink Scheduling With Economic Considerations for Future Wireless Networks

Bader Al-Manthari, *Member, IEEE*, Nidal Nasser, *Member, IEEE*, and Hossam Hassanein, *Senior Member, IEEE*

***Abstract*—Future wireless cellular systems, i.e., high-speed downlink packet access (HSDPA) and 1x EVolution Data Optimized Revision A (1xEV-DO Rev A), promise to revolutionize the mobile users' wireless experience by offering high data rates that are beyond the capabilities of third-generation (3G) systems. These systems, however, require proficient radio resource management schemes to provide the high data rates that they promise. A key component of radio-resource management is packet scheduling, which is responsible for distributing the shared radio resources among the mobile users. In this paper, we propose a novel centralized downlink packet scheduler (CDPS) scheme to be implemented at the base stations of future wireless cellular systems. CDPS is designed to balance between the requirements of connections (e.g., throughput, fairness, etc.) and the requirements of service providers (e.g., revenues). CDPS aims at maximizing the system capacity while ensuring a fair distribution of the wireless resources. To achieve its objective, CDPS employs a flexible utility function that competently incorporates the channel quality conditions of the mobile users, as well as a unique fairness measure. In addition, CDPS utilizes an opportunity cost function that allows service providers to control its degree of fairness and hence control the system capacity. We analytically show that CDPS can be configured to reduce to the maximum carrier-to-interference ratio (Max CIR) and proportional fairness (PF) schemes, hence providing service providers the flexibility to choose between different scheduling schemes. Simulation results and comparisons with existing schemes show the effectiveness and strengths of the CDPS scheme.**

***Index Terms*—Fairness, opportunity cost, packet scheduling, utility.**

I. Introduction

THE INCREASING demand for high-speed mobile data applications has led to the development of new wireless cellular systems that can support high data rates beyond the capabilities of traditional 2.5G and third generation (3G) wireless networks. For example, the Third-Generation Partnership Project (3GPP) has standardized a 3.5G system called high-speed downlink packet access (HSDPA) [1] as an extension to the existing 3G Universal Mobile Telecommunication System (UMTS). HSDPA can theoretically support up to 14.4 Mb/s, which is seven times higher than the data rate offered by UMTS. Another example is the 1x EVolution Data Optimized Revision A (1xEV-DO Rev A), which is an HSDPA-like system standardized by 3GPP 2 (3GPP2) [2]. 1xEV-DO Rev A can achieve peak downlink data rates of up to 3.1 Mb/s. The high data rates offered by these systems allow them to deliver a competitive advantage for mobile data service providers by boosting the network performance to improve the user experience of new converged services such as streaming video, mobile Internet browsing, and Voice over Internet Protocol (VoIP).

However, to maximize the capacity and accommodate a higher number of connections while maintaining the quality of ongoing connections, future wireless cellular systems will require efficient radio-recourse management schemes. A key component of any radio-recourse management scheme is packet scheduling. Packet scheduling will play an important role in future wireless cellular systems since these systems are characterized by high-speed downlink-shared channels to support the increasing number of mobile data users. A centralized downlink packet scheduling scheme is implemented at the base stations of these systems to control the allocation of the downlink-shared channels to the connections by deciding which of them should transmit during a given time interval, and thus, to a large extent, the scheduler determines the overall behavior of these systems. Packet scheduling, therefore, should carefully be designed to maximize the efficiency of the wireless cellular systems and, hence, maximize the obtained revenues.

One important factor that should be considered in the design of a packet scheduling scheme is the connections' channel quality conditions. Mobile users experience varying channel conditions that temporally affect their supportable downlink data rates due to their mobility, interference caused by other connections in the system, obstacles that block or divert their transmitted or received signals, etc. The packet scheduling scheme should track the instantaneous channel conditions of the connections and select for transmission those that are experiencing good channel conditions to maximize system capacity (see Fig. 1). However, favoring connections with good channel conditions may prevent those with bad channel conditions from being served and may, therefore, result in starvation. A good design of a packet-scheduling scheme should take into account not only maximization of the system capacity but also fairness to connections, which use the same service and pay the same amount of money. That is, the packet-scheduling scheme should balance the tradeoff between maximizing capacity and fairness.

Manuscript received May 21, 2007; revised November 4, 2007, March 5, 2008, and April 28, 2008. First published June 6, 2008; current version published February 17, 2009. This work was supported in part by the government of the Sultanate of Oman, by the Natural Science and Engineering Research Council (NSERC) of Canada, and by Bell Canada under the Bell University Lab program. The review of this paper was coordinated by Prof. V. Wong.

B. Al-Manthari and H. Hassanein are with the Telecommunications Research Laboratory, School of Computing, Queen's University, Kingston, ON K7L 3N6, Canada (e-mail: manthari@cs.queensu.ca; hossam@cs.queensu.ca).

N. Nasser is with the Department of Computing and Information Science, University of Guelph, Guelph, ON N1G 2W1, Canada (e-mail: nasser@cis.uoguelph.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2008.927039

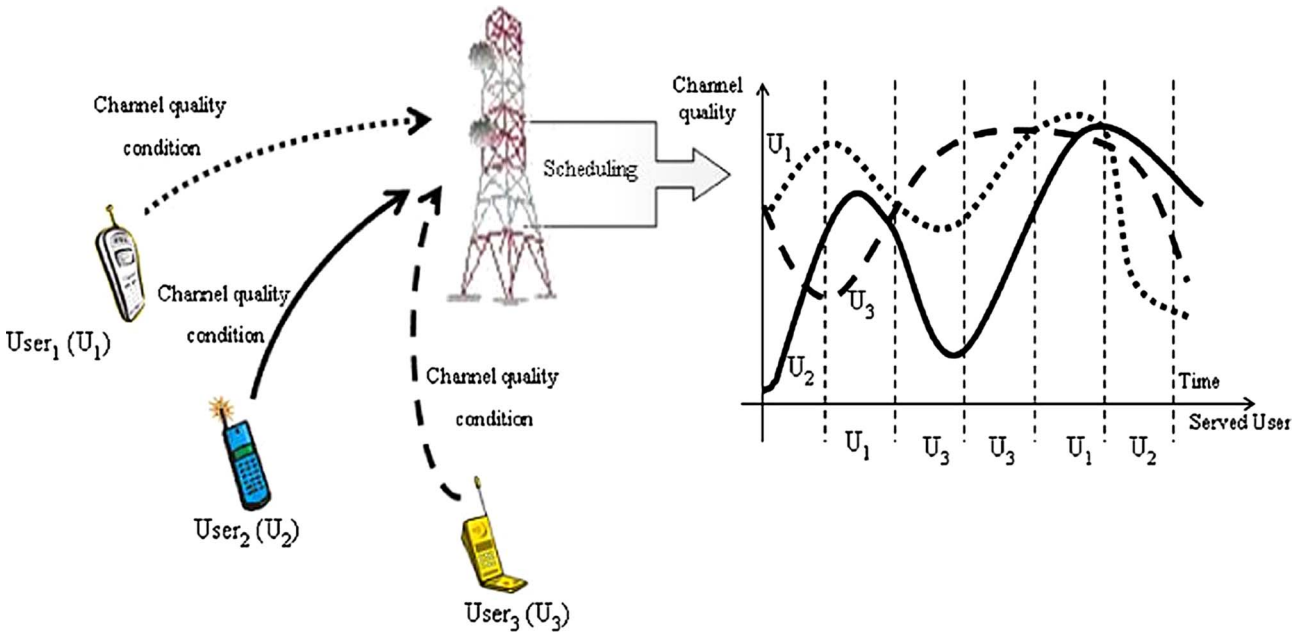

Fig. 1. Exploiting the connection's channel quality condition for scheduling decisions.

A. Related Work

Several packet scheduling schemes have been proposed in the literature to control the capacity–fairness tradeoff in future wireless cellular systems. An overview of code-division multiple-access (CDMA)-related quality-of-service (QoS) provisioning techniques, including state-of-the art packet scheduling, is presented in [3]. In our recent study [4], we surveyed HSDPA scheduling schemes, which we classified into two groups, that is, real-time and non–real-time scheduling schemes. In this section, however, we emphasize recent scheduling schemes for data (non–real-time) services in future wireless cellular networks, which are directly relevant to the contributions in this paper.

The two most well-known packet-scheduling schemes for future wireless cellular networks are the maximum carrier-to-interference ratio (Max CIR) [5] and the proportional fairness (PF) [6] schemes. Max CIR tends to maximize the system's capacity by serving the connections with the best channel quality condition at the expense of fairness since those connections with bad channel quality conditions may not get served. PF tries to increase the degree of fairness among connections by selecting those with the largest relative channel quality where the relative channel quality is the ratio between the connection's current supportable data rate (which depends on its channel quality conditions) and its average throughput. However, a recent study shows that the PF scheme gives more priority to connections with high variance in their channel conditions [7]. To solve the problem of unfairness in PF, a Data Rate Control Exponent scheme [7] is proposed, which adds a fixed exponent term to the current supportable data rate of the connection in PF to control its weight on the scheduling decisions. However, the authors in [8] showed that this scheme has two main drawbacks. First, having a fixed value for the exponent term makes the scheme less adaptable to the instantaneous channel condition of each connection. Second, it is not possible to ensure fairness among all the connections using the same value for the fixed exponent.

In [9], a score-based (SB) packet scheduling scheme is proposed. The SB scheme computes the rank of the current channel quality condition of each connection among the past channel conditions observed over a window of size W. Then, it selects for transmission the connection with the highest rank. This way, the selected rates for each connection are the best possible rates. SB is studied in [10], where an analytical model and simulation results are provided to show that it can achieve higher system throughput than PF while maintaining a similar fairness performance. In [11], the authors propose a scheduling scheme that consists of two procedures, that is, connection selection and resource allocation. The connection selection procedure orders the connections by assigning them time-varying weights. A linear combination of three terms is used to determine the weights. These terms are the channel quality conditions of the connections to increase the capacity of the system, the queue statues of the connections to prevent the selection of connections with a few packets to send, and the time of pending retransmission to reduce the packet loss. Once the connections are ordered based on their assigned weights, the resource-allocation procedure determines the amount of power and the number of codes to maximize the transmission rate at which the ordered connections can send.

An auction-based resource allocation mechanism is proposed in [12], where the assignment of the shared channels' resources is performed by means of auctions. In this scheme, a sequence of "mini-auctions" is conducted, where each mini-auction is pertinent to the reservation of wireless resources within the 1-s interval. Since it is not feasible for the connections to participate in all the mini-auctions, the authors define different utility functions for different service types. These functions are provided

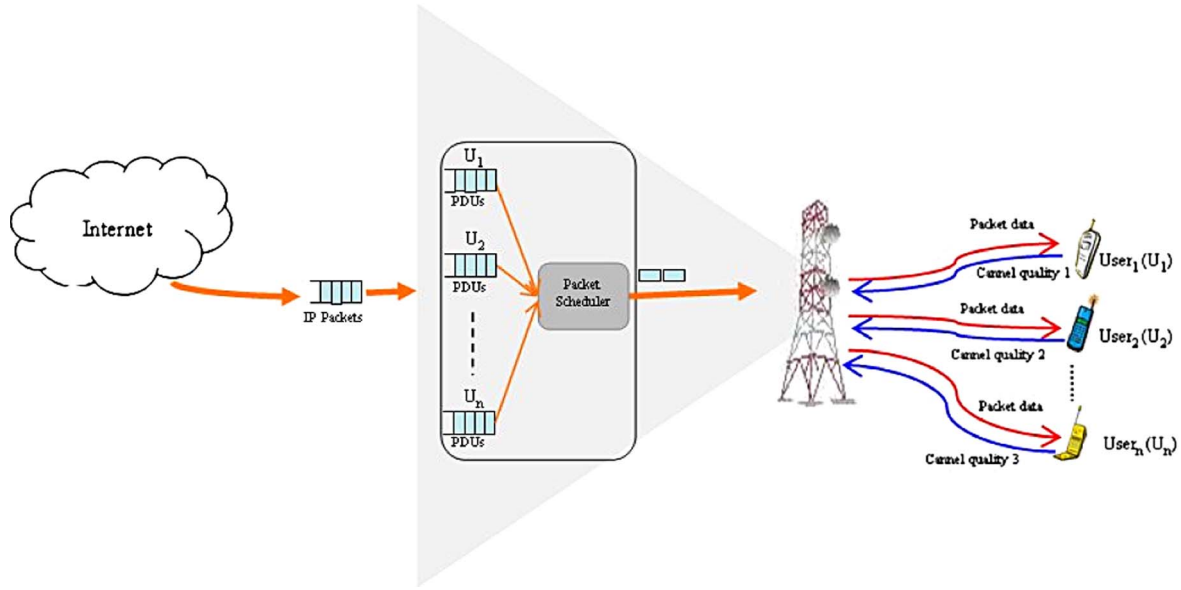

Fig. 2. Packet scheduler model.

by the service providers for connections to choose from and are used as bidding functions on behalf of the connections. The functions are scaled by the connections' total willingness to pay (determined by their chosen utility function) for a certain service.

B. Contributions

In this paper, we propose a novel centralized downlink packet scheduler (CDPS) scheme to be implemented at the base stations of future wireless cellular systems. CDPS lies on a basic economic fact that other existing schemes do not take into consideration. The fact is that mobile users, which are running the same service, may have different preferences depending on their locations, budgets, etc. These preferences are expressed in the CDPS scheme by a realistic utility function, which consists of two major components. The first component is the user connection's preferences that the service provider wants to optimize (e.g., average throughput, average delay, packet loss, etc.). In this paper, our objective is to maximize the connection's average throughputs and the system capacity. The second component is a fairness measure that we defined to ensure a fair distribution of the radio resources among connections. CDPS tries to balance between the user connection's preferences (as perceived by the service provider) and the fairness by formulating an optimization problem that can be solved in real time. In addition, CDPS gives the service provider a complete freedom to determine the degree by which the user connection's preferences and/or fairness affect the connection selection for transmission through the use of an opportunity cost function. That is, CDPS allows the service provider to choose the maximum opportunity cost that it can tolerate such that the highest level of user satisfaction is achieved, while at the same time, the maximum possible amount of revenue is obtained. This gives CDPS a unique feature, which is flexibility. This feature allows the service provider to choose the degree of fairness of the CDPS and, hence, control the capacity–fairness tradeoff. Therefore, the flexibility of CDPS results in a higher user satisfaction and higher revenues. Finally, we analytically show that CDPS can be configured to reduce to the Max CIR and PF schemes, which is another attractive feature of our proposed scheme.

C. Organization of the Paper

The rest of this paper is organized as follows. Section II discusses the system and the packet scheduler models. In Section III, we introduce our proposed packet-scheduling scheme and show its unique properties and effectiveness compared to other schemes. The simulation model, results, and comparisons with existing schemes are given in Section IV. Finally, the conclusions drawn from this paper and future work are discussed in Section V.

II. System and Packet-Scheduler Models

We assume that the base station simultaneously serves n connections $n \geq 1$ and selects one or more connections for transmission in a frame of some fixed time duration. For simplicity, we assume that only one connection is scheduled for transmission at each frame. However, our proposed scheduling scheme will equally work if more than one connection is scheduled. The base station maintains one queue for every connection, as shown in Fig. 2. Upon call arrival, the wireless system receives traffic in the form of IP packets from higher layers, which are segmented into fixed-size Protocol Data Units (PDUs). These PDUs are stored in the transmission queue of the corresponding connection in a first-in–first-out fashion. Subsequently, the PDUs are transmitted to the appropriate connection according to the adopted scheduling scheme.

Packet scheduling in future wireless cellular networks works as follows. Each connection regularly informs the base station of its channel quality condition by sending a report in the uplink to the base station. The report contains information about the instantaneous channel quality of the connection. This information includes the size of the transport block that the base station

should send to the connection, the number of simultaneous channel codes, and the type of modulation and coding schemes that the connection can support. The base station would then select the appropriate connection according to the adopted scheduling discipline and send data to the selected connection at the specified rates. For example, in HSDPA, the connection is able to measure its current channel condition by measuring the power of the received signal from the base station and then using a set of models described in [13] to determine its current supportable data rate (i.e., the rate that it can receive from the base station given its current channel condition).

III. Centralized Downlink Packet Scheduler

In this section, we propose a novel packet-scheduling scheme for future wireless cellular systems, which we call CDPS.[1] The proposed scheme employs practical economic models through the use of utility and opportunity cost functions. The use of such models is intended to enhance the satisfaction of the mobile users while at the same time increasing the obtained revenues through maximizing the system capacity. We first begin by outlining the general formulation of CDPS. Next, we provide a definition for a possible utility function, an opportunity cost function, and a fairness measure that fit into the general formulation of the CDPS. Finally, we mathematically show that our defined utility function for CDPS reduces to the Max CIR and PF scheduling schemes as special cases.

The user's i $(1 \leq i \leq n)$ preferences at time t as perceived by the service provider can be expressed by a utility function $U_i(X_{i1}(t), X_{i2}(t), \ldots, X_{im}(t))$, where n is the total number of users' connections in the system, $X_{i1}(t), \ldots, X_{im-1}(t)$ are the chosen quantitative measures of the user connection's preferences in this system such as the average throughput, current data rate, average delay, etc., $X_{im}(t)$ is a fairness measure that represents how fair the scheduling scheme is to the user connection, and m is the maximum number of chosen quantitative measures. We assume that the utility function is additive. Thus, we can express the aggregate utility of the system, for n connections, as $\sum_{i=1}^{n} U_i(X_{i1}(t), X_{i2}(t), \ldots, X_{im}(t))$ [14]. The scheduling scheme can then be formulated as an optimization problem

$$\begin{aligned} &\text{Maximize} \sum_{i=1}^{n} U_i\left(X_{i1}(t), X_{i2}(t), \ldots, X_{im}(t)\right) \\ &\text{Subject to} \quad OC_i(t) \leq K \\ &\qquad \forall i,\ 1 \leq i \leq n \end{aligned} \tag{1}$$

where $OC_i(t)$ is the opportunity cost of serving connection i at time t, and K is a predefined constant value. The concept of opportunity cost[2] arises because of the tradeoff between capacity and fairness. If the scheduling scheme tries to be fair, then the system capacity will decrease, and hence, the service provider's revenues might decrease (if the connections are charged per bit of data). Therefore, there is an opportunity cost to fairness, which, in this case, equals the loss of system capacity because a certain connection has been served rather than the one with the best channel condition. It should be noted that the opportunity cost limits the cost of serving connections to K, and hence, it increases the obtained revenues of the service provider since the connections with opportunity cost less than K are not served.

A. Cobb–Douglas Utility Function for Downlink Scheduling

In this section, we propose a utility function, a fairness measure, and an opportunity cost function. The proposed functions are designed to simultaneously achieve four objectives, namely, efficiency, fairness, user satisfaction, and flexibility. Efficiency refers to exploiting the variations of the channel quality conditions of the connections so that they can be sent at higher rates. Fairness is defined in terms of the distribution of the connections' average throughputs. User satisfaction is defined in terms of the user's connection average throughput exceeding a predefined value. Finally, flexibility refers to the freedom to control the degree of fairness of the scheduling scheme and, hence, the system throughput and the service provider's revenues. The CDPS formulation may accommodate other utility functions, fairness measures, and opportunity cost functions and is not limited to those described below.

We adopt the Cobb–Douglas utility function [14] in our proposed scheme. The reason for using the Cobb–Douglas utility function is that it fits the design goals of our proposed scheduling scheme since it consists of different parameters that each contribute to the total utility, depending on its weight, as follows. Assuming $m = 2$ in our formulation of CDPS, the Cobb–Douglas utility function is expressed as $U_i(X_1, X_2) = X_1^c \cdot X_2^d$, where $c, d \geq 0$. Let X_1 be any performance metric that the service provider wants to optimize, such as the average connection throughput or average delay. Let X_2 be a fairness measure that increases as the connection's or system's perception of fairness increases, which results in an increase in U. Then, we can express the preferences of user connection i at time t, $1 \leq i \leq n$ by

$$U_i\left(X_{i1}(t), X_{i2}(t)\right) = X_{i1}(t)^c \cdot X_{i2}(t)^d. \tag{2}$$

To maximize the system's overall utility, we need to achieve the highest possible values of $X_{i1}(t)$ and $X_{i2}(t)$ for all connections. However, it is not possible to achieve high values of both $X_{i1}(t)$ and $X_{i2}(t)$ for all connections because of the tradeoff between capacity and fairness, as mentioned earlier. In this case, the scheduling scheme needs to find the connection, which if served the system's utility will be maximized. We provide our definition of $X_{i1}(t)$ and $X_{i2}(t)$ in the next section.

To define the fairness measure for connection i, we proceed as follows:

1) $S_i(t)$ is the average throughput for connection i up to time t.

[1] A simplified version of the proposed scheduling scheme appeared at the Globecom 2006 conference.

[2] The opportunity cost for a good is defined as the value of any other goods or services that a person must give up to produce or get that good [14]. For example, when going to a movie theater, the opportunity cost is how much money could have been made if one works for the same amount of time spent watching the movie.

2) $\max_j S_j(t)$ is the maximum average throughput achieved among all connections up to time t.

Given these two definitions, the fairness measure for connection i at time t, $\alpha_i(t)$ can be defined as

$$\alpha_i(t) = S_i(t)/\left(\max_j S_j(t)\right). \tag{3}$$

That is, the fairness measure for connection i is the ratio of its average throughput to the maximum throughput achieved among all the connections in the system. We call this measure the "relative fairness." The higher the relative fairness is, the "happier" the user will be. As shown later in this section, the objective of the utility function is to achieve high values of the relative fairness to all the connections to increase the fairness of the scheduling scheme. The opportunity cost of serving connection i at time t (i.e., the opportunity cost of fairness) is defined as

$$OC_i(t) = \left(\max_j R_j(t)\right) - R_i(t) \tag{4}$$

where $R_i(t)$ is the current data rate for connection i at time t, which depends on its channel condition, and $\max_j R_j(t)$ is the maximum current data rate of all connections at time t. That is, the opportunity cost is how much data rate the system would compromise if connection i is selected for transmission given that there is a connection j with a higher current data rate. The service provider can determine the appropriate level of opportunity cost of fairness by choosing K and, hence, the appropriate level of fairness capacity required to maximize its obtained revenues.

B. Definitions of $X_{i1}(t)$ and $X_{i2}(t)$

In addition to the preceding parameters, we define the following.

1) c: the Cobb–Douglas utility function's constant, where $c \geq 0$. The value of this constant determines the weight on $X_{i1}(t)$ in the Cobb–Douglas utility function;
2) d: the Cobb–Douglas utility function's constant, where $d \geq 1$. The value of this constant determines the weight on $X_{i2}(t)$ in the Cobb–Douglas utility function. We restrict the value of this constant to an odd integer because our defined $X_{i2}(t)$ in the adopted Cobb–Douglas utility function is a negative function, as shown later, and therefore, d must be odd to preserve this;
3) n: total number of connections in the system;
4) $X_{i1}(t) = R_i(t)$: the current data rate of connection i at time t. The utility of connection i being served increases as $R_i(t)$ increases. It should be noted that other performance metrics could be used. However, we use the current data rate as the first component in the Cobb–Douglas utility function to increase the system capacity and, hence, achieve the efficiency objective (see Section III-C for details);
5) $X_{i2}(t) = f(\alpha_i(t), \gamma_i(t)) = 1 - \gamma_i^{-\ln(\alpha_i(t))}$, $\gamma_i > 1$: the fairness measure, which is a function of the relative fairness that we defined to increase fairness in the system. The fairness measure is designed such that it increases as the connection's relative fairness increases and thus increases its utility function. This measure is used to ensure fairness among connections. The parameter γ_i is used to control the shape of $X_{i2}(t)$ and, hence, the level of fairness in the system. γ_i can be set to different values for different connections to allow the service provider to maintain different levels of fairness for different connections depending on the type of traffic they have, the amount of money they are expected to pay, their loyalty, etc.

Therefore, we can express the utility of connection i at time t as

$$\begin{aligned} U_i\left(X_{i1}(t), X_{i2}(t)\right) &= X_{i1}^c(t) \cdot X_{i2}^d(t) \\ &= \left(R_i(t)\right)^c \cdot \left(1 - \gamma_i^{-\ln(\alpha_i(t))}\right)^d. \end{aligned} \tag{5}$$

For an additive utility function, the aggregate utility of the system is

$$\sum_{i=1}^{n} \left(R_i(t)\right)^c \cdot \left(1 - \gamma_i^{-\ln(\alpha_i(t))}\right)^d. \tag{6}$$

Given the opportunity cost constraint, then at each scheduling decision CDPS will find the connection that would maximize the following objective function:

$$\begin{aligned} &\sum_{i=1}^{n} \left(R_i(t)\right)^c \cdot \left(1 - \gamma_i^{-\ln(\alpha_i(t))}\right)^d \\ &\text{Subject to} \quad OC(i,t) \leq K \\ &\quad \forall i,\ 1 \leq i \leq n. \end{aligned} \tag{7}$$

Note that the scheduling decision occurs every time frame according to our assumed packet-scheduling model. Thus, it is important to note that, at each time frame, the current supportable data rate $[R_i(t)]$ of every connection i is known, and its average throughput $S_i(t)$ [and, hence, its relative fairness $\alpha_i(t)$] can be calculated by using any throughput averaging method, such as that in [15]. Therefore, a solution to (7) can found by computing the aggregate utility of the system if connection i is scheduled [see (7)] $\forall i$ and then finding the connection with the highest aggregate utility. Hence, a solution to (7) can be found by choosing connection i for transmission such that

$$\begin{aligned} i = \arg\max_i \Bigg[&\left(R_i(t)\right)^c \cdot \left(1 - \gamma_i^{-\ln(\alpha_i(t))}\right)^d \\ &+ \sum_{j=1, j\neq i}^{n} \left(R_j(t)\right)^c \cdot \left(1 - \gamma_j^{-\ln(a_j(t))}\right)^d \Bigg] \end{aligned} \tag{8}$$

where connection i is selected for transmission, and all the other connections j, $j \neq i$ are not selected. If connection i is selected for transmission, then $\alpha_i(t)$ will increase, and $\alpha_i(t) \forall j \neq i$ will decrease as the other connections are not served. Therefore, the run-time complexity of our scheme is $O(n^2)$ since (8) has to be computed for every connection.

Two important factors affect the scheduling decision (i.e., the choice of connection, e.g., i). The first factor is the connection's current supportable data rate $R_i(t)$, which depends on its channel condition. The connection with a higher current supportable data rate has a higher chance of maximizing the aggregate

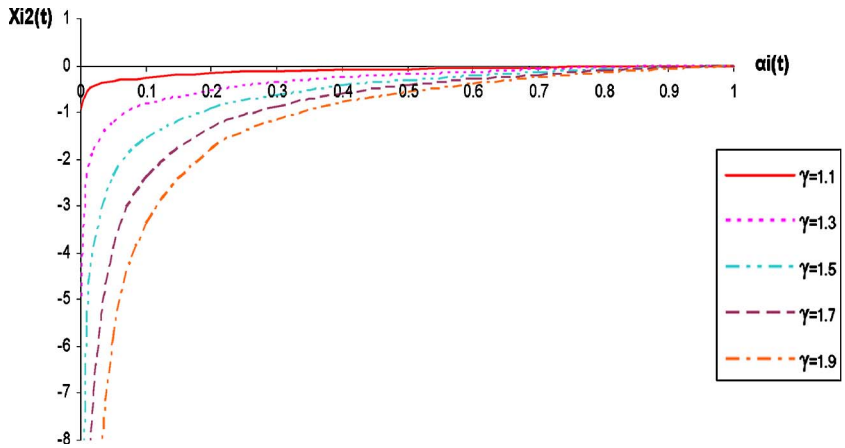

Fig. 3. Plot of $X_{i2}(t)$ for $0 < \alpha \leq 1$ and different γ_i values.

utility of the system. The second factor is the connection's relative fairness $[X_{i2}(t)]$. In our utility function, $X_{i2}(t)$ is a function of $\alpha_i(t)$ and γ_i. $X_{i2}(t)$ is designed such that it increases at a much faster rate if a connection with a relatively low average throughput (e.g., $\alpha_i(t)$ close to 0) is served than a connection with a relatively high average throughput (e.g., $\alpha_i(t)$ close to 1). γ_i determines the rate of decrease in $X_{i2}(t)$. Larger values in γ_i result in higher rates of decrease in $X_{i2}(t)$ (particularly as $\alpha_i(t)$ gets close to 0), which results in a higher fairness. Fig. 3 plots $X_{i2}(t)$ for $0 < \alpha \leq 1$ for γ_i fixed at 1.1, 1.3, 1.5, 1.7, and 1.9. As we can see, $X_{i2}(t)$ decreases at a much faster rate as the connection's relative fairness decreases from 1 to 0, and it approaches $-\infty$ as the connection's relative fairness approaches 0. This will ensure fairness among connections since if a connection with a high average throughput is served, although its utility will increase, the overall utility will not be maximized due to the rapid decrease of the utilities of those with low average throughputs. Therefore, the scheduling scheme will be forced to serve those with low average throughputs, since, if served, their utility function will sharply increase, resulting in maximizing the system aggregate utility, although those connections may not have the best channel conditions.

As Fig. 3 shows, the larger the value of γ_i, the higher the rate of decrease in $X_{i2}(t)$ (the rate of decrease in $X_{i2}(t)$ increases as $\alpha_i(t)$ gets close to 0), which allows the scheduling scheme to be fairer to the connections with low α values (i.e., low average throughputs compared with connections with high average throughputs). The service provider can choose different fixed values of γ_i for different connections, depending on their traffic class, history, etc. The values of γ_i can also be dynamically changed as needed by the service provider and according to the network statues.

C. Properties of CDPS

Earlier in the paper we identified the main objectives of CDPS. In this section, we investigate the extent to which CDPS satisfies the aforementioned objectives.

Efficiency: CDPS takes into account the instantaneous channel conditions of connections (through their current supportable data rates). CDPS makes efficient use of the bandwidth by relatively favoring connections with good channel conditions.

TABLE I
CDPS PARAMETER SETTINGS

c, d, γ_i	$1,1,6^3$

Fairness: Fairness in CDPS results from the fact that CDPS considers not only the instantaneous channel condition of the connections but also their average throughputs compared with the maximum average throughput. Both values are used in the scheduling decision.

User Satisfaction: User satisfaction, as perceived by the service provider, is taken into account by using both the instantaneous channel condition and the user's connection relative fairness. Exploiting the information about the channel conditions results in achieving high-connection average throughputs. In addition, taking into account the connections' relative fairness results in distributing the wireless resources fairly among them. This prevents starvation and, hence, improves user satisfaction.

Flexibility: Introducing the concept of opportunity cost to our proposed scheme gives it a high degree of flexibility. This gives the service provider the flexibility to choose the degree of fairness and thereby control the capacity–fairness tradeoff and effect the obtained revenues. In our scheme, the opportunity cost function is defined as $OC_i(t) = (\max_j R_j(t)) - R_i(t)$, which is the loss of throughput if the connection with the maximum data rate is not served. A smaller value of K results is a higher opportunity cost and in turn a higher system capacity and a lower degree of fairness. This is because only those whose current supportable data rates are close (depending on K) to the maximum are chosen for transmission. The service provider may choose the appropriate values for K (for each base station) to correspond to a certain degree of fairness to maximize its revenues.

D. Flexibility of CDPS

In this section, we provide two lemmas showing that the Max CIR and PF schemes are special cases of our proposed CDPS scheme by setting the CDPS parameters to specific values (Table I). The proofs of both lemmas are provided in the Appendix.

Lemma 1: If K is set to 0, then the CDPS reduces to the Max CIR scheme.

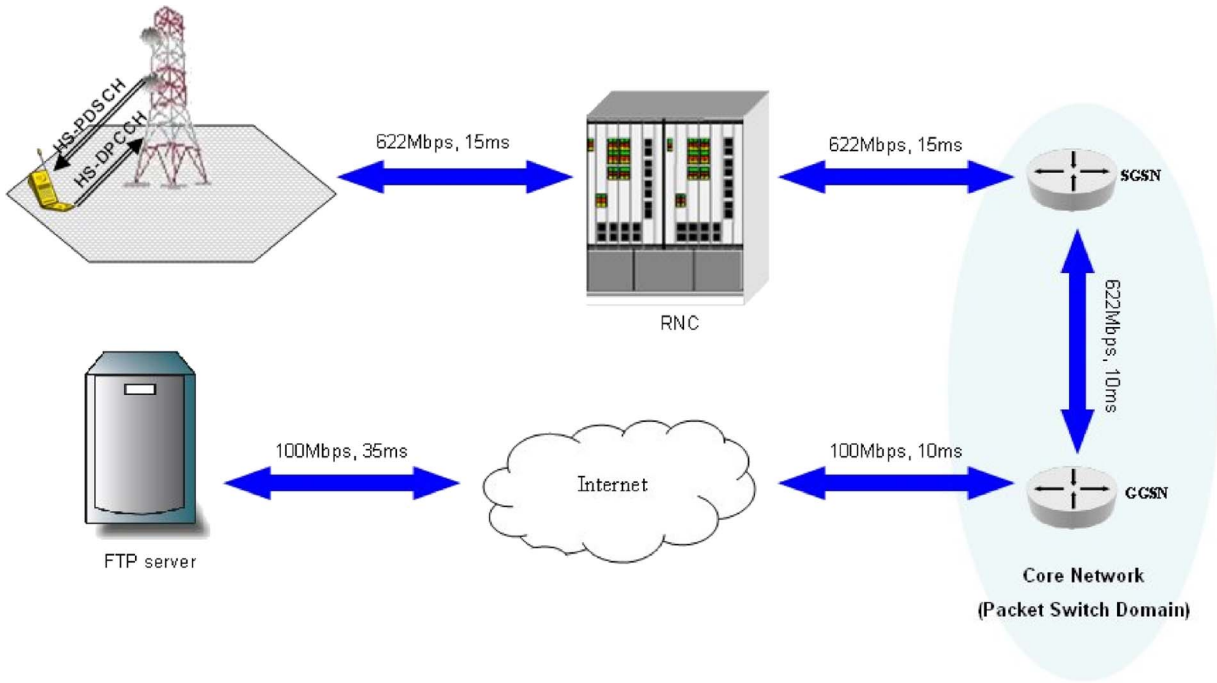

Fig. 4. Simulation model.

Lemma 2: If c is set to 0, d is set to 1, and γ_i is set to $e^{(-\ln(1-\ln\alpha_i(t))/\ln\alpha_i(t))}$, then the CDPS reduces to the PF scheme.

IV. Performance Evaluation

In this section, we evaluate the performance of our proposed scheduling scheme by means of a dynamic discrete-event simulation with the help of Network Simulator 2 (NS-2) [17] and its Enhanced UMTS Radio Access Network Extensions (EURANE) [18]. We tested our scheme on the HSDPA system. As aforementioned, HSDPA is a 3.5G wireless system that has been introduced by the 3GPP as an extension to UMTS [1]. More information about HSDPA can be found in [4].

A. Simulation Model

Fig. 4 shows the simulation model. We simulated a one-cell case, and for simplicity, we did not distinguish new and handoff calls. The cell radius is 1 km. The base station (known as Node-B in HSDPA networks) is located at the center of the cell. Therefore, only one Node-B is involved in allocating the radio resources. The connections are connected to Node-B on the downlink by high-speed physical downlink shared channel (HS-PDSCH), which is the actual physical channel for HSDPA, and on the uplink by high-speed physical dedicated control channel (HS-PDCCH), which is used to send the connections' current estimates of their channel conditions to Node-B. The Node-B is connected to the radio network controller (RNC) by a duplex link of 622-Mb/s bandwidth and 15-ms delay. The RNC is connected to the serving GPRS support node (SGSN) by a duplex link with 622-Mb/s bandwidth and 15-ms delay. The SGSN is connected to the gateway GPRS support node (GGSN) by a duplex link of 622-Mb/s bandwidth and 10-ms delay (SGSN and GGSN are part of the core network (CN) and are used to support packet-switched services). The CN is connected to the Internet via a duplex link of 100-Mb/s bandwidth and 10-ms delay. At the Internet end, an FTP server is connected to it by a duplex link of 100-Mb/s bandwidth and 35-ms delay. All of these values can be found in [19]. Each connection sends a request for one FTP file, and then, the connection terminates after the download is complete. The size of each FTP file is 50 MB.

TABLE II
Simulation Parameters

Simulation time	300s
Node-B transmission power	38 dBm
Antenna gain	17 dBi
Node-B buffer size	30 MB
Shadowing	Lognormal distribution
Intra-cell interference	30 dBm
Inter-cell interference	-70 dBm
Packet discard time	6s
Node-B buffer size	30 MB
HS-DSCH codes	10
Arrival rate	Poisson with mean 1 s

At the initialization, n connections are uniformly distributed over users in the cell. Every mobile user moves inside the cell with a constant speed of 3 km/h, which is the recommended value for a Pedestrian A (Ped A) environment by the 3GPP [19]. The simulation time step is one time frame, which is 2 ms in HSDPA, and the simulation time is 300 s (Table II).

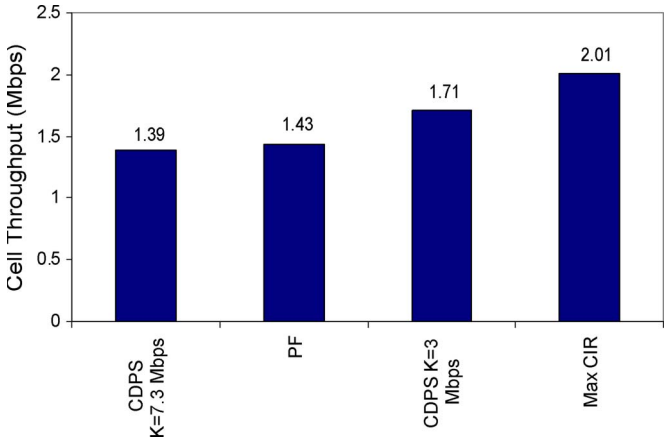

Fig. 5. Cell throughput.

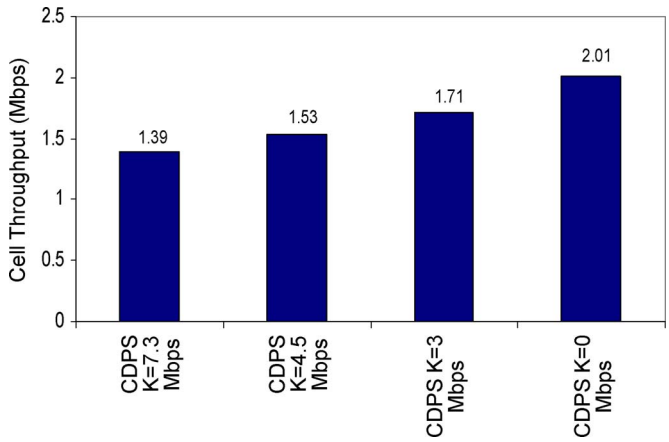

Fig. 6. Cell throughput with different values of K.

B. Channel Model

The channel model describes the attenuation of the radio signal on its way from Node-B to the user, and therefore, it describes how the channel condition of the user's connection changes with time, depending on its environment and the speed of the user. In our simulation, we adopt the propagation model in [4], which consists of five parts, that is, distance loss, shadowing, multipath fading, intracell interference, and intercell interference.

C. Simulation Results

In this section, we compare the performance of CDPS with that of Max CIR and PF. Two tested environments are used, that is, Ped A [19] and fixed differentiated channel conditions. The mobile users in the Ped A environment move at a fixed speed of 3 km/h, which is the recommended value by the 3GPP. The fixed channel environment is created to evaluate the performance of the CDPS under different fixed channel conditions (as opposed to Ped A in which the channel conditions of users vary with time, according to the models specified by the 3GPP). We compare the schemes in terms of cell throughput, distribution of user connections' average throughputs, users' satisfactions in terms of providing minimum average throughput guarantees, and percentage of packet loss due to buffer overflow and packet discarding.

Case 1—Ped A: Fig. 5 compares the cell throughput of the evaluated schemes for the Ped A environment with 25 user connections. The figure shows that Max CIR achieves the highest cell throughput (2.01 Mb/s). This is expected since Max CIR only serves connections at their best channel conditions at the expense of ignoring those with bad channel conditions. The cell throughput achieved by CDPS with $K = 7.3$ Mb/s is slightly lower than PF (1.39 Mb/s compared to 1.43 Mb/s). This is because CDPS serves connections with low average throughputs even more than PF by giving them more time slots to increase their relative fairness and maximize the overall utility of the system. However, for lower values of K, for example, 3 Mb/s, the cell throughput increases from 1.39 to 1.71 Mb/s. This is because when $K = 3$ Mb/s, only connections with good channel conditions are served (i.e., their instantaneous channel conditions are good enough such that the opportunity cost of serving them does not exceed 3 Mb/s). The effect of different values of K on cell throughput is shown in Fig. 6. The figure confirms that the service provider can control the cell throughput by controlling K, which is a unique feature of CDPS.

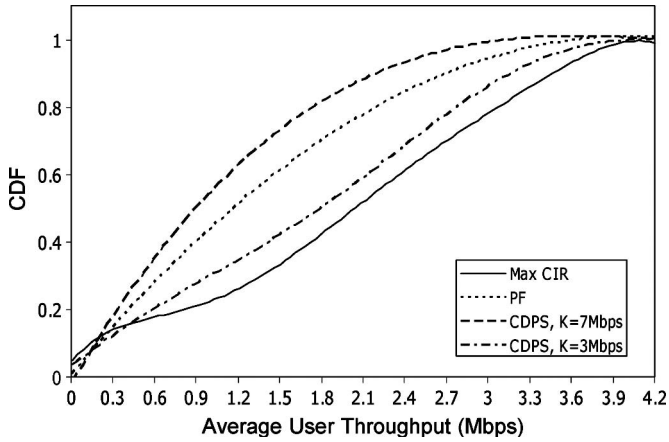

Fig. 7. Distribution of connection average throughputs.

Fig. 7 depicts the cumulative distribution function (cdf) of the connections' average throughputs for the Ped A with 25 connections. The steeper the cdf curve is, the fairer the scheduling scheme, because that means the connections' average throughputs are distributed over a small interval (i.e., connections get relatively equal average throughputs to each other). CDPS has a steeper slope than Max CIR and PF because of the effect of relative fairness, which gives more time slots to connections with low average throughputs to compensate them for their bad channel conditions. Fig. 8 shows the cdf curves of CDPS at different values of K. Clearly, we can control the degree of fairness that CDPS provides and, hence, the system throughput by changing the values of K.

The user satisfaction with a minimum average expected throughput of 128 kb/s (i.e., a user is satisfied if its average connection throughput is greater than or equal to 128 kb/s) is shown in Fig. 9. CDPS outperforms Max CIR and PF because it increases the chance of those connections with low average throughputs of getting served because of the effect of relative

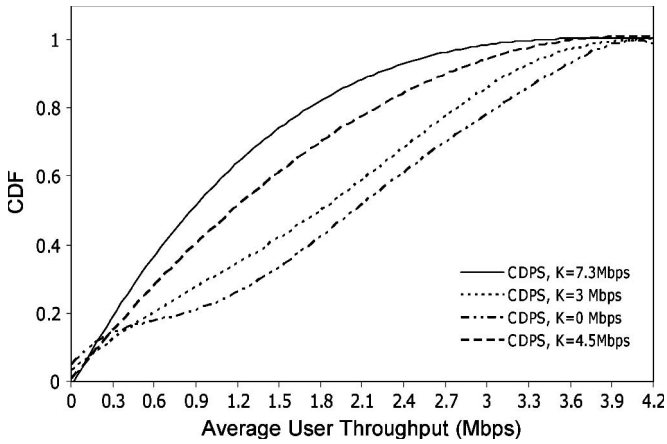

Fig. 8. Distribution of connection average throughputs with different values of K.

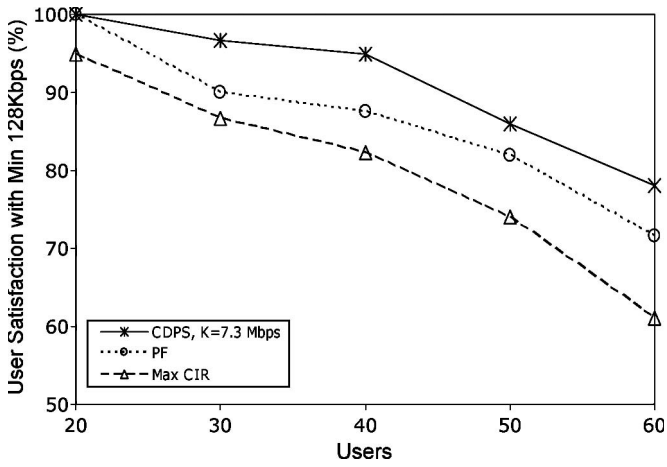

Fig. 9. User satisfaction with minimum throughput of 128 kb/s.

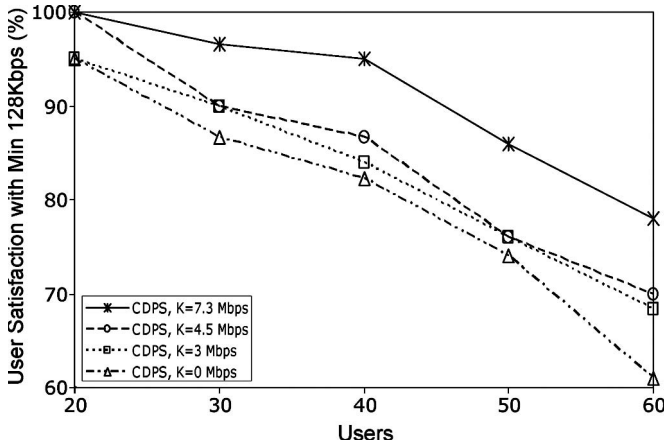

Fig. 10. User satisfaction with minimum throughput of 128 kb/s with different values of K.

fairness. However, as shown in Fig. 10, as we decrease K, fewer connections are satisfied because for low K values only those with good channel conditions are selected for transmission at the expense of ignoring the rest of the connections. Figs. 11 and 12 depict the user satisfaction with a minimum average throughput of 356 kb/s. These two figures show a similar behavior to Figs. 9 and 10, except that the percentages of satisfied

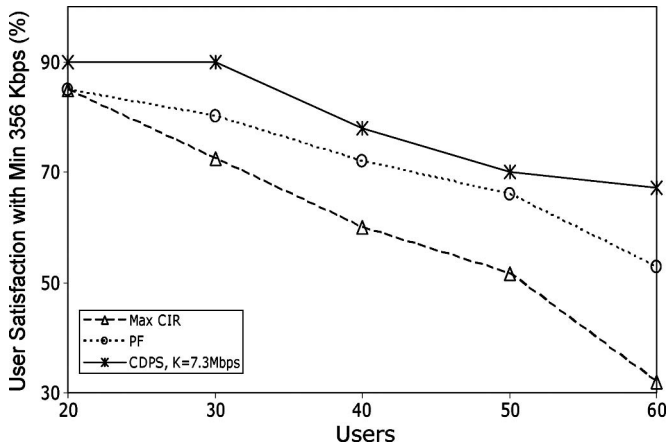

Fig. 11. User satisfaction with minimum throughput of 356 kb/s.

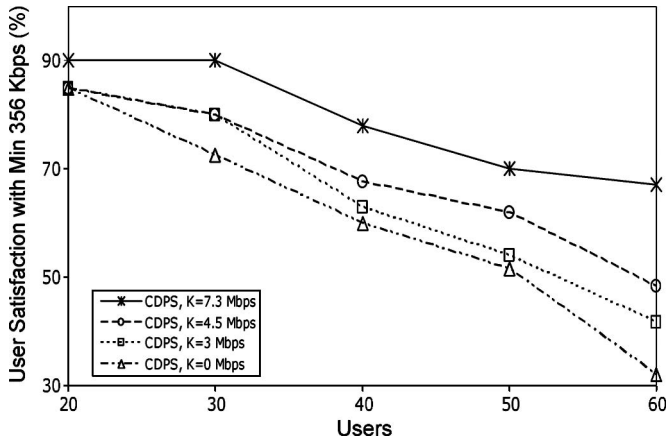

Fig. 12. User satisfaction with minimum throughput of 356 kb/s with different values of K.

users are lower in Figs. 11 and 12 because it is harder to achieve a minimum throughput guarantee of 356 kb/s than 128 kb/s.

Case 2—Differentiated Channel Conditions: The scheduling schemes are evaluated in this environment based on the average connection throughput and the percentage of packet loss. Seven values are used for the SNR: -7, -4, -1, 2, 5, 8, and 11 dB (i.e., the channel conditions of the connections are differentiated and fixed at these values). This experiment demonstrates how the scheduling schemes serve connections with different channel conditions. For each SNR value, there are ten connections (a total of 70 connections in the cell). The results for each group of ten connections based on their SNR are separately collected.

Figs. 13 and 14, respectively, depict the average throughputs and percentage of packet loss for connections with different SNR values. Clearly, CDPS achieves better performance in terms of average throughput and percentage of packet loss for connections with low SNR values (e.g., -7, -4, and -1). This is because of the effect of the fairness measure, which ensures that the connections that are having low average throughputs get more time slots to increase their relative fairness. The performance of Max CIR and PF is worse than CDPS for connections with low SNR values since more time slots are given to connections with good SNR values.

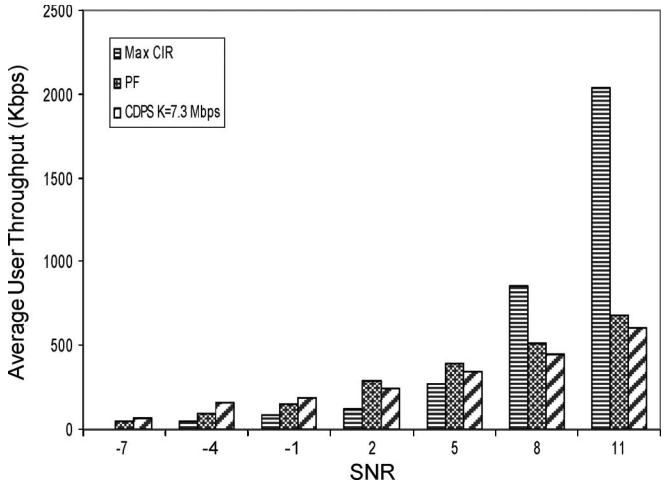

Fig. 13. Average connection throughput for connections with different SNR values.

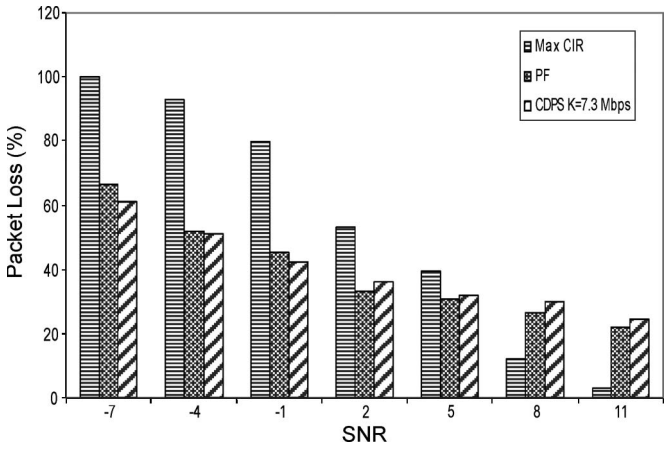

Fig. 14. Percentage of packet loss for connections with different SNR values.

V. Conclusion and Future Work

Packet scheduling is a key component in any future radio resource management module. Future wireless cellular systems exploit the bursty nature of the data traffic by utilizing high-speed downlink shared channels that are shared among the connections according to the packet scheduling scheme being used. Therefore, the design of an efficient packet scheduling scheme is crucial to the functionality of these systems. In this paper, we have proposed a CDPS scheme for future wireless cellular systems that is based on a utility function to represent the satisfactions of the mobile users as perceived by the service provider. Our scheme also utilizes an opportunity cost function to represent the satisfactions of the service provider. Therefore, unlike existing schemes, we try to simultaneously satisfy the conflicting requirements of the mobile users (fairness, throughput, etc.) and service providers (revenues). CDPS can simultaneously meet four design objectives, that is, efficiency, fairness, users' satisfactions, and flexibility. We have shown that CDPS reduces to two well-known scheduling schemes as special cases, that is, Max CIR and PF. This gives the service provider more flexibility in using different scheduling schemes.

We are currently conducting research on the effectiveness of the proposed scheme in supporting different types of traffics that have different QoS requirements and determine the appropriate parameter values in the scheme that are required to meet the QoS for such types of traffic. CDPS should be also tested on different charging schemes, and short- and long-run revenues should be examined for different parameter values of the scheme. We remark that although CDPS is designed as a packet scheduling scheme, the use of economic models and concepts can be adapted to other types of radio-resource management functions as well as to other types of systems. Therefore, we would like to use the utility and opportunity cost functions in CDPS in call admission control and handoff decisions to satisfy users and service providers as well. We would also like to investigate the potential of using CDPS as a polling mechanism to determine which connections will be polled in systems like 802.11e. In this case, other components may be used in the utility function than the connections' channel quality conditions, such as the probability that the connection will have data to send if polled, network statues, queue lengths, fairness, etc.

Appendix

Proof of Lemma 1: Set $K=0$. Then, CDPS will find connection i for which $OC_i(t)=(\max_j R_j(t))-R_i(t)\leq 0$. The only connection that satisfies this constraint is the connection with the maximum current supportable data rate [i.e., $\max_j R_j(t)$]. Therefore, CDPS is equivalent to Max CIR in choosing the connection with the highest supportable data rate. ■

Proof of Lemma 2: As we will show later, the PF scheme requires that $U_i(t)=\ln\alpha_i(t)$. Therefore, we first need to find the value of γ_i such that $1-\gamma_i^{-\ln(\alpha_j(t))}=\ln\alpha_i(t)$. We get this by solving $\gamma_i^{-\ln(\alpha_j(t))}=1-\ln\alpha_i(t)$. $\therefore -\ln(\alpha_i(t))\ln(\gamma_i)=\ln(1-\ln\alpha_i(t))$ $\therefore \gamma_i=e^{(-\ln(1-\ln\alpha_i(t))/\ln\alpha_i(t))}$.

Therefore, if we set $c=0$, $d=1$, and $\gamma_i=e^{(-\ln(1-\ln\alpha_i(t))/\ln\alpha_i(t))}$, and we ignore the opportunity cost by setting K to the highest data rate that the system can support so that all the connections are considered for transmission, then the utility function in the CDPS becomes $\ln(\alpha_i(t))$, since $\max_j S_j(t)$ is common to every connection in $\alpha_i(t)$, and we can then take it off. Therefore, the CDPS will find connection i such that

$$\text{Maximize} \sum_{i=1}^{n} \ln\left(S_i(t)\right)$$

$$\forall i,\ 1\leq i\leq n. \tag{9}$$

Maximizing the aggregate utility of the system is equivalent to maximizing the objective function F, where F is a function of $\vec{S}(t)$, and $\vec{S}(t)$ is a vector of the connections' average throughputs at time t. That is, if we find a vector $\vec{S}(t)$ that maximizes F, then the aggregate utility function will also be maximized. In other words, we want to choose some values for the connections' average throughputs such that they maximize F and, hence, the aggregate utility of the system. In our case, choosing such values is done by choosing a connection for

transmission that results in the vector $\vec{S}(t)$ maximizing the objective function F. Therefore, the problem can be formulated as

$$\text{Maximize} F\left(\vec{S}(t)\right) \equiv \sum_{i=1}^{n} \ln\left(S_i(t)\right) \quad \forall i,\ 1 \le i \le n \tag{10}$$

where $S_i(t+1)$ is the average throughput for connection i at slot $t+1$ and can be calculated by using an exponentially smoothed filter as follows [15]:

$$S_i(t+1) = \begin{cases} (1-1/t_c)S_i(t) \\ \quad +1/t_c \times R_i(t), & \text{if user } i \text{ is served} \\ (1-1/t_c)S_i(t), & \text{otherwise} \end{cases} \tag{11}$$

where t_c is the time constant of the filter, and $R_i(t)$ is the current supportable data rate of connection i.

Since $\ln(S_i(t))$ is strictly concave and is differentiable, then so is the objective function F. In addition, since the feasible region is bounded, then an optimal solution exits. Furthermore, the solution is unique, and we can use a gradient ascent method to find it as explained in [16]. However, a global optimal solution cannot be found since the number of connections and the channel capacity are varying with time. Nevertheless, we can look for a locally optimal solution. That is, at each time slot, schedule the connection that would result in a movement toward the optimal solution (i.e., a movement along the maximum objective function F gradient direction).

Let $F_i'(\vec{\mathrm{S}}(t))$ be the gradient of the objective function in the direction of serving connection i. We would like to find the value of i with the largest gradient and moving to the maximal point along that direction. Since we know what the connection's average throughput would be if served or not, then the optimization problem can be reduced to find the maximum gradient in the direction of serving connection i [i.e., maximize $F_i'(\vec{\mathrm{S}}(t))$]. We first find the gradient in the direction of serving connection i. We can do this by parameterizing the movement along the ray in the direction of serving connection i by μ, and then, F_i can be written as a function of μ as

$$F_i(\mu) = \sum_{i=1}^{n} \ln\left(S_i(t) + \mu\left(S_i(t+1) - S_i(t)\right)\right). \tag{12}$$

Taking the derivative with respect to μ and evaluating it at $\mu = 0$ (to find the critical point, in this case maxima), we get

$$\begin{aligned} F_i'(\mu) &= \sum_{i=1}^{n} \frac{1}{\left(\left(S_i(t) + \mu(S_i(t+1) - S_i(t))\right)\right)} \\ &\quad \cdot \left(\left(S_i(t+1) - S_i(t)\right)\right) \\ &= \frac{1}{\left(\left(S_i(t) + \mu(S_i(t+1) - S_i(t))\right)\right)} \\ &\quad \cdot \left(\left(S_i(t+1) - S_i(t)\right)\right) \\ &\quad + \sum_{j=1, j\neq i}^{n} \frac{1}{\left(\left(S_j(t) + \mu(S_j(t+1) - S_j(t))\right)\right)} \\ &\quad \cdot \left(\left(S_j(t+1) - S_j(t)\right)\right) \end{aligned}$$

$$\begin{aligned} \therefore\ F_i'(0) &= \frac{1}{S_i(t)} \cdot \left(\left(S_i(t+1) - S_i(t)\right)\right) \\ &\quad + \sum_{j=1, j\neq i}^{n} \frac{1}{S_j(t)} \cdot \left(\left(S_j(t+1) - S_j(t)\right)\right) \\ &= \frac{1}{S_i(t)} \cdot \left(\frac{R_i(t)}{t_c} - \frac{S_i(t)}{t_c}\right) \\ &\qquad (\text{connection is } i \text{ served}[(11)]) \\ &\quad + \sum_{j=1, j\neq i}^{n} \frac{1}{S_j(t)} \cdot \left(-\frac{S_j(t)}{t_c}\right) \\ &\qquad (\text{connection } j \text{ is not served}[(11)]) \\ &= \frac{1}{S_i(t)} \cdot \left(\frac{R_i(t)}{t_c}\right) - \frac{1}{S_i(t)} \cdot \left(\frac{S_i(t)}{t_c}\right) \\ &\quad - \sum_{j=1, j\neq i}^{n} \frac{1}{S_j(t)} \cdot \left(\frac{S_j(t)}{t_c}\right). \end{aligned}$$

Therefore, the gradient in the direction of serving connection i can be written as

$$\frac{1}{S_i(t)} \cdot \left(\frac{R_i(t)}{t_c}\right) - \sum_{j=1}^{n} \frac{1}{S_j(t)} \cdot \left(\frac{S_j(t)}{t_c}\right). \tag{13}$$

The summation term and the constant scalar t_c are common terms for all connections and can be ignored. Consequently, the maximum gradient direction (i.e., the connection that would result in a movement along the maximum gradient direction) is reduced to $\arg\max_i F_i'(\vec{\mathrm{S}}(t)) = \arg\max_i(R_i(t))/(S_i(t))$, which is the same as the PF scheme. ■

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers, whose feedback helped improve the quality of this paper.

REFERENCES

[1] *High Speed Downlink Packet Access (HSDPA); Overall Description*, Mar. 2003. 3GPP TS 25.308, Rel. 5.
[2] *CDMA2000 High Rate Packet Data Air Interface Specification*, Apr. 2004. 3GPP2 CS0024, ver. 1.0.
[3] H. Jiang, W. Zhuang, X. Shen, and Q. Bi, "Quality-of-service provisioning and efficient resource utilization in CDMA cellular communications," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 1, pp. 4–15, Jan. 2006.
[4] B. Al-Manthari, N. Nasser, and H. Hassanein, "Packet scheduling in 3.5G high speed downlink packet access networks: Breadth and depth," *IEEE Netw. Mag.*, vol. 21, no. 1, pp. 41–46, Feb. 2007.
[5] S. Borst, "User-level performance of channel-aware scheduling schemes in wireless data networks," in *Proc. IEEE Conf. Comput. Commun. INFOCOM*, Mar. 2003, vol. 1, pp. 321–331.
[6] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of CDMA-HDR a high efficiency-high date rate personal communication wireless system," in *Proc. IEEE VTC*, May 2000, pp. 1854–1858.
[7] M. Kazmi and N. Wiberg, "Scheduling schemes for HSDSCH in a WCDMA mixed traffic scenario," in *Proc. IEEE Int. Symp. PIMRC*, Beijing, China, Sep. 2003, pp. 1485–1489.

[8] G. Aniba and S. Aissa, "Fast packet scheduling assuring fairness and quality of service in HSDPA," in *Proc. CCECE*, May 2004, vol. 4, pp. 2243–2246.
[9] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," in *Proc. EW Conf.*, Sep. 2002, pp. 2244–2248.
[10] M. Assaad and D. Zeghlache, "Scheduling study in HSDPA," in *Proc. IEEE Int. Symp. PIMRC*, Sep. 2005, vol. 3, pp. 1890–1894.
[11] G. Manfredi, P. Annese, and U. Spagnolini, "A channel aware scheduling scheme for HSDPA system," in *Proc. IEEE Int. Symp. PIMRC*, Sep. 2005, vol. 4, pp. 2136–2140.
[12] M. Dramitinos, G. Stamoulis, and C. Courcoubetis, "Auction-based resource allocation in UMTS high speed downlink packet access (HSDPA)," in *Proc. NGI Netw. Conf.*, Apr. 2005, pp. 434–441.
[13] *Physical Layer Procedures*, Jun. 2003. 3GPP TS25.214, Rel. 5, ver. 5.5.0.
[14] H. Varian, *Intermediate Microeconomics: A Modern Approach*, 6th ed. New York: Norton, 2003.
[15] P. Jose, "Packet scheduling and quality of service in HSDPA," Ph.D. dissertation, Aalborg Univ., Aalborg, Denmark, Oct. 2003.
[16] P. A. Hosein, "QoS control for WCDMA high speed packet data," in *Proc. IEEE Int. Workshop MWCN*, Stockholm, Sweden, Sep. 2002, pp. 169–173.
[17] *Network Simulator 2*. [Online]. Available: http://www.isi.edu/msnam/ns
[18] *Enhanced UMTS Radio Access Network Extensions for NS2*. [Online]. Available: http://www.ti-wmc.nl/eurane/
[19] *End-to-end Network Model for Enhanced UMTS*. Deliverable D3. 2v2. [Online]. Available: http://www.ti-wmc.nl/eutrane/

Bader Al-Manthari (M'06) received the B.Sc. and M.Sc. degrees (with Honors) in 2004 and 2005, respectively, from Queen's University, Kingston, ON, Canada, where he is currently working toward the Ph.D. degree.

His research interests include economic-based radio resource management in next-generation wireless cellular networks, wireless ad hoc and sensor networks, and performance evaluation of communication protocols and schemes.

Nidal Nasser (M'00) received the B.Sc. and M.Sc. degrees (with Honors) in computer engineering from Kuwait University, Kuwait City, Kuwait, in 1996 and 1999, respectively, and the Ph.D. degree from Queen's University, Kingston, ON, Canada, in 2004.

He is currently an Associate Professor with the Department of Computing and Information Science, University of Guelph, Guelph, ON, Canada. He is a Technical Editor of Wiley's *International Journal of Wireless Communications and Mobile Computing* and Wiley's *Security and Communication Networks Journal*. He has authored several journal publications and refereed conference publications, as well as seven book chapters. He has also given tutorials at major international conferences.

Dr. Nasser is a member of several IEEE technical committees. He has been a member of the technical program and organizing committees of several international IEEE conferences and workshops. He received the Fund for Scholarly and Professional Development Award in 2004 from Queen's University. He received the Best Research Paper Award at the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA'08).

Hossam Hassanein (M'87–SM'06) received the Ph.D. degree in computing science from the University of Alberta, Edmonton, AB, Canada, in 1990.

From 1991 to 1993, he was with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada. From 1993 to 1999, he was with the Department of Mathematics and Computer Science, Kuwait University, Kuwait City, Kuwait. Since 1999, he has been with the School of Computing, Queen's University, Kingston, ON, Canada, where he is currently a Leading Researcher in the areas of broadband, wireless, and variable topology network architecture, protocols, control, and performance evaluation. He is the Founder and Director of the Telecommunications Research (TR) Laboratory. He has more than 300 publications in reputable journals, conference proceedings, and workshops in the areas of computer networks and performance evaluation. He has organized and served on the program committees of a number international conferences and workshops. He also serves on the editorial board of a number of international journals.

Dr. Hassanein is currently the Vice-Chair of the IEEE Communications Society Technical Committee on Ad hoc and Sensor Networks (TC AHSN). He was the recipient of Communications and Information Technology Ontario (CITO) Champions of Innovation Research Award in 2003. In 2007, he received the Best Paper Awards at the IEEE Wireless Communications and Networks and the IEEE Global Communication Conferences (both flagship IEEE communications society conferences).