Dynamic Fair Scheduling With QoS Constraints in Multimedia Wideband CDMA Cellular Networks

Liang Xu, *Member, IEEE*, Xuemin (Sherman) Shen, *Senior Member, IEEE*, and Jon W. Mark, *Fellow, IEEE*

***Abstract*—A class of dynamic fair scheduling schemes based on the generalized processor sharing (GPS) fair service discipline, under the generic name code-division GPS (CDGPS), is proposed for a wideband direct-sequence code-division multiple-access (CDMA) cellular network to support multimedia traffic. The CDGPS scheduler makes use of both the traffic characteristics in the link layer and the adaptivity of the wideband CDMA physical layer to perform fair scheduling on a time-slot by time-slot basis, by using a dynamic rate-scheduling approach rather than the conventional time-scheduling approach. Soft uplink capacity is characterized for designing efficient CDGPS resource allocation procedure. A credit-based CDGPS (C-CDGPS) scheme is proposed to further improve the utilization of the soft capacity by trading off the short-term fairness. Theoretical analysis shows that, with the C-CDGPS scheme, tight delay bounds can be provided to delay-sensitive traffic, and short-term unfairness can be bounded so that long-term weighted fairness for all users can still be satisfied. Simulation results show that bounded delays, increased throughput, and long-term fairness can be achieved for both homogeneous and heterogeneous traffic.**

***Index Terms*—Code-division multiple-access (CDMA), credit-based scheduling, dynamic fair scheduling, generalized processor sharing (GPS), quality-of-service (QoS), soft capacity.**

I. Introduction

FUTURE wireless networks are expected to support multimedia traffic, such as video, voice, and data, and to make efficient use of the radio resource. One of the promising approaches to efficiently support quality-of-service (QoS) for multimedia traffic in wireless networks is to employ a central scheduler, which can dynamically allocate bandwidth to mobile users in accordance with the variation of traffic load and channel conditions. The scheduler should be efficient in utilizing the radio resources and be fair in scheduling services. An ideal fair scheduling discipline is the well-known generalized processor sharing (GPS) [1], also known as weighted fair queuing (WFQ) [2]. The basic principle of GPS is to assign each user a fixed weight, instead of a fixed bandwidth, and to dynamically allocate bandwidth (or service rate) to all the users according to their weights and traffic load. Several GPS-based fair scheduling schemes have been proposed for wireline packet networks [1]–[3], and been adapted to wireless networks [4]–[7]. These conventional GPS schemes are implemented based on a time-scheduling approach, which features in a high complexity due to the extensive computation for the virtual time of each packet [3]. The time-scheduling approach is suitable for time-division multiple access (TDMA)-based wireless networks [4]–[6] and can be adapted to a hybrid TD/CDMA system [7]. For general CDMA networks, however, the radio resources are mainly related to the spreading bandwidth and transmission power, and time-division multiplexing is rarely involved [8]. Therefore, the time-scheduling approach is not fully suited to the CDMA systems, and new fair scheduling schemes which should take into account the QoS requirements of multimedia traffic, as well as changing conditions in CDMA wireless links are needed. Recently, dynamic bandwidth allocation (DBA) by varying the channel rate in DS-CDMA systems has been proposed for supporting multiple QoS [9]–[12]; however, the fair scheduling issue is not well addressed. In [13], a credit-based scheduling scheme is proposed to guarantee throughput and fairness of bursty data traffic in the downlinks [from base station (BS) to mobile stations (MSs)] of a wideband CDMA system, where the downlink capacity is considered as a constant. In that scheme, each traffic flow is associated with an amount of credits which represents the amount of service that the flow is owed, and service for different traffic flows is arranged in decreasing order of their credits. The credit of each flow may be increased at each scheduling instant if the allocated bandwidth for the flow is less than the agreed upon bandwidth and be decreased otherwise. A similar credit-based scheduling scheme is proposed in [14] for each uplink channel in a wideband CDMA system. The credit-based scheduling schemes proposed in [13] and [14] are simple to implement, since they do not require extensive computation as that in GPS-based time scheduling. However, it may be difficult for these schemes to provide tight delay bounds which are crucial for delay-sensitive traffic. In [15], delay and throughput constraints are taken into account by a token-based scheduling scheme proposed for a shared CDMA downlink, which is based on a time-scheduling approach and fairness constraint is not considered.

An important feature of CDMA cellular networks is the so-called soft capacity; the capacity of each cell, especially in uplink, is subject to the variation of signal-to-interference ratios (SIRs) and bandwidth demands of users in the cell. For heterogeneous multimedia traffic with time-varying bandwidth demands and diverse QoS requirements, a useful interpretation of the uplink capacity in a cell is the maximum throughput that can be achieved. The uplink capacity in terms of the maximum throughput will vary with the bandwidth demands from multimedia users, even though the number of multimedia users

Manuscript received April 24, 2002; revised August 6, 2002; accepted October 9, 2002. The editor coordinating the review of this paper and approving it for publication is E. Hossain. This work was supported in part by a grant from the Communications and Information Technology Ontario (CITO), and in part by an Ontario Graduate Scholarship (OGS).

The authors are with the Centre for Wireless Communications, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada (e-mail: lxu@bbcr.uwaterloo.ca; xshen@bbcr.uwaterloo.ca; jwmark@bbcr.uwaterloo.ca).

Digital Object Identifier 10.1109/TWC.2003.819028

admitted to the network can be fixed. For instance, if one user is allocated a large bandwidth while the other users are allocated relatively small bandwidths, the total achievable throughput could be larger than that for uniform bandwidth allocation (i.e., all the users are allocated the same bandwidth). This feature has been exploited for improving the uplink data throughput [16] and designing high-data-rate CDMA (HDR-CDMA) systems [17]. The softness of the uplink capacity due to variable bandwidth demands and interference makes fair scheduling in the multimedia DS-CDMA cellular networks a nontrivial problem. The main reasons are as follows.

1) In the conventional fair scheduling problem, the service rate (i.e., the link capacity) is assumed to be known by the scheduler. In the multimedia CDMA cellular networks, however, it is difficult for the scheduler to know the exact total service rate in the uplink, because the uplink capacity may be different for different bandwidth allocations. In order to avoid overbooking the capacity, a conventional scheduler may have to make a conservative estimate of the soft uplink capacity, which may result in an inefficient utilization of the radio resources. How to design a scheduler more efficient in utilizing the soft uplink capacity is an important issue.
2) Although unbalanced bandwidth allocation (e.g., in the HDR-CDMA system,) can improve the total throughput as compared with balanced bandwidth allocation in the CDMA uplink, it may contradict the fairness requirement. In other words, the total uplink throughput may have to be sacrificed for fairness. On the other hand, a short-term unfair bandwidth allocation may improve the total throughput. How to tradeoff the fairness with throughput is, thus, an important issue for fair scheduling in the multimedia CDMA uplink.

In this paper, a code-division GPS (CDGPS) fair scheduling scheme for the uplink is investigated, which employs both DBA and GPS to provide fair services in wideband CDMA networks [18]. The CDGPS scheduler makes use of both the traffic characteristics in the link layer and the adaptivity of the wideband CDMA physical layer to achieve an efficient utilization of radio resources. It adjusts the channel rate (service rate) of each traffic flow on a time-slot by time-slot basis by varying the spreading factor and/or using a multiple of orthogonal code channels, rather than allocates service time for each packet. This results in a lower computational complexity of the CDGPS scheme than that of the conventional GPS-based time-scheduling scheme, while the performance is comparable in terms of tightly bounded delay and guaranteed bandwidth provision [18]. It should be mentioned that a low-complexity GPS-based bandwidth scheduling scheme similar to the CDGPS has also been proposed in [19] for a multicarrier CDMA system. The CDGPS scheme considered here is for the uplink of DS-CDMA cellular systems, and addresses the soft-capacity issue. The main contributions of this work are as follows.

1) A formulation of the soft uplink capacity is given, where a "nominal capacity" is defined to interpret the uplink resource in each cell. Based on the nominal capacity, a resource allocation procedure is proposed for adapting the CDGPS scheme to exploit the soft-capacity in a more efficient way. The importance of introducing the concept of nominal capacity is twofold. First, the nominal capacity is measurable based on the measurement of intercell interference. Second, the nominal capacity is more accurate in describing the uplink resource of each cell than the actual soft-capacity which is uncertain due to the various bandwidth demands of the users in the cell.
2) A modified CDGPS scheme with the generic name credit-based CDGPS (C-CDGPS) is proposed. With the C-CDGPS scheme, the soft uplink capacity is allocated by using a combination of credit-based scheduling and CDGPS fair scheduling. The C-CDGPS guarantees a minimum bandwidth to each flow according to the GPS discipline, and distributes the extra bandwidth to traffic flows in decreasing order of the credits assigned to the users. The C-CDGPS scheme can improve the delay and throughput, as compared with the original CDGPS scheme. Long-term fairness among all users can also be guaranteed due to the bounded user credits, although short-term unfairness may be introduced by the credit-based scheduling. The novelty of the C-CDGPS scheme lies in that fairness and throughput requirements are well balanced in the CDMA uplink system. The proposed scheme differs from the previous credit-based schemes presented in [13]–[15] in that the softness feature of uplink capacity is taken into account by the C-CDGPS scheme.

The rest of this paper is organized as follows. The system model is briefly described in Section II. In Section III, the formulation of soft capacity is given. In Section IV, the CDGPS and C-CDGPS schemes are described in details. Simulation results are given in Section V to demonstrate the performance of CDGPS and C-CDGPS schemes, followed by concluding remarks in Section VI.

II. System Model

We consider a frequency-division-duplex (FDD) wideband DS-CDMA cellular network supporting a large number of multimedia users. The system model is different from the wideband code-division multiple-access (W-CDMA) standardized for the third-generation (3G) mobile communication systems, mainly in that a centralized bandwidth scheduling mechanism is used to allocate the CDMA system resources. In particular, a pair of bandwidth schedulers are assumed to reside in each BS, allocating the resource of the uplink and downlink, respectively, to all MSs in the cell covered by the BS. Since this paper focuses on the uplink scheduling, we shall describe the details of the considered uplink system and the media access control (MAC) scheme. The physical layer of the wideband DS-CDMA system is similar to the standard W-CDMA system with matched-filter receivers [20]. The physical data channels in the uplink, distinguished by pseudonoise (PN) codes, comprise a small number of random access channels and a large number of dedicated channels. Each mobile user is assigned to at least one dedicated data channel. As in the standard W-CDMA system, there are also some control channels which are used for transmitting system control signals, such as pilot signals and power control commands. It is assumed that the random access channels and control channels consume

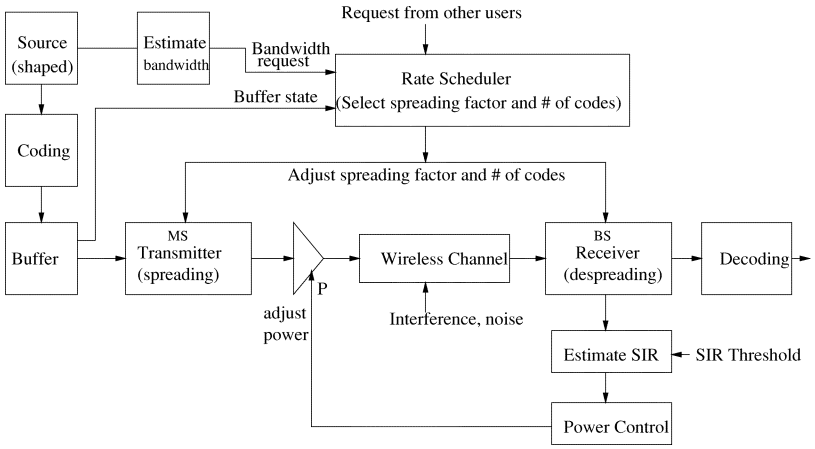

Fig. 1. Rate-scheduled wideband DS-CDMA system.

only a small portion of the uplink capacity, which is negligible compared with the capacity reserved for the dedicated channels bearing multimedia user traffic. In the remainder of this paper, maximum achievable throughout on all dedicated channels is referred as the uplink capacity. Due to the multiple-access interferences (MAI) among uplink channels, the uplink capacity is considered to interference-limited, i.e., the capacity is not limited by the number of available PN codes.

A rate-scheduled wideband DS-CDMA system model is considered. The transmission rates of the dedicated uplink data channels are dynamically allocated by the uplink scheduler, as shown in Fig. 1. The source of each MS generates a sequence of packets which enter a buffer after error control coding. The buffered information bits are then converted to a DS-CDMA signal with the symbol rate R_c/N_s, where R_c is the chip rate and N_s is the spreading factor. Since the chip rate (spread bandwidth) is fixed, the spreading factor N_s (number of chips per symbol) determines the service rate (or channel rate) at which the buffer is emptied. If multicode CDMA technique [20] is employed, the service rate can be mR_c/N_s, where m is the number of code channels. The receiver for each user is a single-user matched filter with despreading and decoding. Signal-to-interference ratio (SIR)-based close-loop fast power control mechanism is applied to achieve a target bit-error rate (BER) at the receiver side of each dedicated uplink channel. An SIR threshold corresponding to the target BER is set at the receiver side.[1] When the actual SIR is lower (higher) than the threshold, a feedback control signal is sent back to the transmitter to increase (decrease) the transmission power. The SIR threshold also corresponds to the user channel rate. When the channel rate is changed, the SIR threshold is adjusted accordingly. It is well known [21] that the power control can be nearly perfect for low-speed mobiles to maintain the target BER. For high-speed mobiles, however, the variance of the actual SIR will increase due to fast channel fading, and the SIR threshold should be increased accordingly to maintain the target BER. Without loss of generality, we assume slow-speed MSs in the remainder of this paper.

The channel rate of each MS is scheduled on a time-slot by time-slot basis. A demand-assignment slotted-MAC scheme similar to those proposed in [10] and [22] is applied to support the scheduling. The general procedure of the MAC is described as follows: By the end of each time slot, the MS sends a short message to the BS through the random access channel or its own dedicated channel. The message reports the buffer state and the amount of traffic that arrived in the previous slot. Upon receiving bandwidth request from all the active users, the scheduler allocates the channel rate for the next slot, taking into account user QoS requirements and the available uplink capacity. The resource allocation message, which may be a few bits in length, will be sent back to the MS before the start of the next slot. The resource allocation message may be sent via a downlink control channel or be piggy-backed on the downlink data channel of the MS [22]. Upon receiving a resource allocation message indicating that the rate needs to be changed, the MS immediately adjusts its spreading factor and/or number of codes at the start of the next slot. The rates of different channels are changed simultaneously, and all the uplink channels are synchronized at the slot level. This, however, does not require strict bit synchronization among different channels, and can be supported by asynchronous DS-CDMA systems. It should be mentioned that, although the demand-assignment MAC scheme may increase signalling cost to both uplink and downlink systems, the signalling cost can be minor when multimedia traffic are large in amount.

Packetized multimedia traffic, such as packet voice, video, and data, is considered. QoS requirements of the packet traffic generally consist of packet delay and loss rate. The packet loss is mainly due to the transmission errors and delay is due to the transmission and queuing time. The packet loss rate can be constrained to a satisfactory level by using the SIR-based power control to maintain a target BER and using forward error control (FEC) to correct transmission errors. For the packetized multimedia traffic, all the users can be considered to have the same BER requirement.[2] However, the packet delay requirements may be different for different types of traffic. For delay-

[1]The SIR threshold for a soft-handoff MS can be set lower than that of a nonhandoff MS [8].

[2]Although the physical layer BER requirements may be the same, multimedia users can use different error control protocols to achieve different packet loss rates.

sensitive traffic, packet delay needs to be bounded by the scheduler. In order to achieve a bounded delay for a user, it is required that each traffic source is shaped by a leaky-bucket regulator [23] with parameter (σ, ρ), where σ and ρ are token buffer size and token generation rate, respectively.

III. Soft Capacity

As the uplink capacity of a CDMA system is soft in nature, an important issue is how to formulate and deploy the capacity. In this section, we introduce the concept of "nominal capacity" which describes the softness feature of the uplink capacity. Consider a cell with N active users in the uplink. Let the spread bandwidth be $W = R_c/\alpha$, where R_c is the chip rate and α is a constant depending on the shape of the chip. Denote R_i the channel rate and P_i the received signal power of the ith user in the cell, respectively. The SIR of the ith user can be written as

$$\mathrm{SIR}_i = \frac{P_i}{\sum\limits_{j \neq i} P_j + P_n + I_{\mathrm{inter}}} \tag{1}$$

where P_n is the background noise power at the BS and I_{inter} is the intercell interference power from other cells. The bit energy of user i is $E_b = P_i/R_i$. Let I_e be the equivalent spectral density of the interference plus background noise for user i. The equivalent power of interference plus noise is WI_e. Then, the QoS at the BS receiver can be expressed in terms of E_b/I_e and (1) can be manipulated to yield

$$\left(\frac{E_b}{I_e}\right)_i = \frac{W}{R_i}\mathrm{SIR}_i. \tag{2}$$

To achieve the target BER, a minimum E_b/I_e requirement, $(E_b/I_e)_0$, needs to be guaranteed for all users. Let γ_i be the desired SIR of user i, i.e., the SIR threshold. For QoS satisfaction, γ_i must be greater than or equal to the specification $\gamma_i^0 = (E_b/I_e)_0 R_i/W$, so that the actual SIR_i can be achieved as

$$\mathrm{SIR}_i \geq \gamma_i^0 = \left(\frac{E_b}{I_e}\right)_0 \frac{R_i}{W}. \tag{3}$$

The uplink capacity can be defined as

$$C = \sum_{i=1}^{N} R_i f\left(\left(\frac{E_b}{I_e}\right)_i\right) \tag{4}$$

where $R_i = \gamma_i W/(E_b/I_e)_0$, and $f((E_b/I_e)_i)$ is an increasing function of the achieved $(E_b/I_e)_i$ which accounts for the effect of forward error control. In general, $f((E_b/I_e)_i) < 1$, since some redundant bits need to be transmitted for error control. With $(E_b/I_e)_i$ converging to $(E_b/I_e)_0$ under the SIR-based power control, $f((E_b/I_e)_i$ converges to $f((E_b/I_e)_0)$.

Due to the fluctuation of I_{inter} and the variation of bandwidth demands, the uplink capacity is soft, i.e., C of a cell is uncertain. The following simple example illustrates that different bandwidth demands may result in different values of C. Consider a simplified system model, where only two users sharing the DS-CDMA uplink and there are no other interference and background noise except for the MAI between the two users. Let R_1 and R_2 be channel rates allocated to the two users, respectively. For simplicity, assume that the two users achieve the same E_b/I_e at $(E_b/I_e)_0$. Let SIR_i equal to γ_i^0 in (3), it can be obtained that $R_1 = (W/(E_b/I_e)_0)^2/R_2$. Fig. 2 shows the R_1 versus R_2 curve, which defines the capacity region of the system. It can be seen that the R_1 versus R_2 curve is convex. This is because the uplink capacity of the DS-CDMA system with matched-filter receivers is MAI-limited. Therefore, the achievable $(R_1 + R_2)$ is smaller at point A (i.e., $R_1 = R_2$) than that at point B (i.e., $R_1 \neq R_2$). In other words, the uplink capacity C may be different for different bandwidth demands of the two users, which results in the softness of uplink capacity.

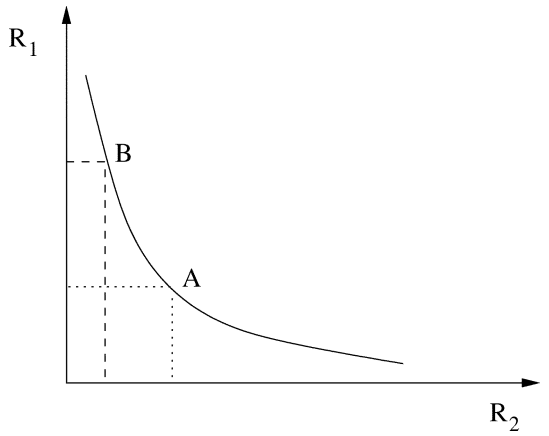

Fig. 2. Illustration of soft capacity.

In the scheduled wideband DS-CDMA system, a fair rate-allocation is based on the knowledge of the available resource. Due to the uncertainty of C, we introduce a *nominal capacity* to represent the available uplink resource. The *nominal capacity* of a cell is independent of the user bandwidth demands in the cell, and can be determined by measuring the intercell interference.

A. Nominal Capacity

From (1), if SIR_i can be maintained at the desired level γ_i, for any $i = 1, \ldots, N$, the optimal (minimal) received powers are [24]

$$P_i^* = \frac{P_n + I_{\mathrm{inter}}}{1 - \sum\limits_{j=1}^{N} \frac{\gamma_j}{1+\gamma_j}} \frac{\gamma_i}{1+\gamma_i}, \quad i = 1, \ldots, N. \tag{5}$$

Since the power value should be positive and limited, the following necessary condition must be satisfied.

$$\sum_{i=1}^{N} \frac{\gamma_i}{1+\gamma_i} < 1. \tag{6}$$

However, when the left-hand side of (6) is close to 1, the optimal power levels may be too high to be sustainable. Moreover, the increase of the total power of users in one cell may adversely affect the surrounding cells and stimulate an increase of I_{inter}. Therefore, it is necessary to impose the inequality

$$\sum_{i=1}^{N} \frac{\gamma_i}{1+\gamma_i} \leq 1 - \delta \tag{7}$$

where δ is a small positive number. We define $(1 - \delta)$ as the *nominal capacity*. It should be noted that in practice, SIR_i is a random variable and cannot be fixed at the designated γ_i,

because of the imperfection of power control. However, if γ_i, $i = 1, \ldots, N$, is set to be the SIR threshold for all i, and (7) is satisfied, SIR_i can be kept close to γ_i under the SIR-based power control [21]. Since γ_i is proportional to channel rate R_i, it can also represent the bandwidth allocated to the ith user. Therefore, the bandwidth allocation can be performed by assigning γ_i, for $i = 1, \ldots, N$, under the nominal capacity constraint (7).

In general, the nominal capacity $(1-\delta)$ of a target cell can be determined using the following approach. Let $P_u = \sum_{i=1}^{N} P_i^*$, and let (7) hold with equality. Then, from (5), (7) and some manipulation, we have

$$1 - \delta = \frac{P_u}{P_u + I_{\text{inter}} + P_n}. \tag{8}$$

In general, P_u can be set by the service provider. As P_u represents a target limit on the sum of the received signal powers from all the mobiles within the cell, the benefit of setting this system parameter is limiting the interference from this cell to other cells. In practice, P_u can also be dynamically adjusted according to the location of mobiles. For instance, if most of the mobiles move closer to the BS and, hence, their interference to other cells decrease, P_u can be increased accordingly. On the other hand, $(I_{\text{inter}} + P_n)$ can be estimated as follows [17]. First, the total received power which is the sum of user signal powers and $(I_{\text{inter}} + P_n)$ can be measured at the BS. Second, each user's signal strength is measured. Then, $(I_{\text{inter}} + P_n)$ can be obtained by subtracting the sum of the intracell user signal powers from the total received power. It should be mentioned that the actual $(I_{\text{inter}} + P_n)$ may vary from time to time. For the celluar network supporting a large number of users, the accumulated interference from MSs in the surrounding cells can be well approximated by a Gaussian process whose statistics are changing slowly [8]. I_{inter} may also experience fast fluctuation due to fast channel fading and the rapid variation of traffic load of the interfering users. When the number of interferers is large, the fast fluctuation of I_{inter} can be averaged out if the slot duration and the temporal window for measuring the interference are large enough. In practice, the slot resolution should be carefully designed for the scheduling scheme, so that the long-term variation of $(I_{\text{inter}} + P_n)$ can be tracked and the fast fluctuation can be mitigated by averaging. Since I_{inter} may vary dramatically from slot to slot when a small number of interferers dominates, the $(I_{\text{inter}} + P_n)$ should be estimated in a more conservative way for this situation.

Based on the estimated $(I_{\text{inter}} + P_n)$, the nominal capacity $(1-\delta)$ can be determined by (8) for each time slot. From (8), it can be seen that the nominal capacity $(1-\delta)$ does not depend on bandwidth demands, although the actual uplink throughput C does. Given that (7) holds with equality, it can be shown that C is a convex function of $(\gamma_1, \ldots, \gamma_N)$. Therefore, C can be different for different allocations of $(\gamma_1, \ldots, \gamma_N)$, and a lower bound of C may be found at the point where all the γ_i, $i = 1, \ldots, N$, are the same (e.g., the point A in Fig. 2).

IV. CDGPS Schemes

In this section, the CDGPS scheme for fixed capacity is introduced first. A resource allocation procedure is proposed to adapt the CDGPS scheme to the uplink system with soft-capacity. The credit-based CDGPS scheme is then developed to improve the soft capacity utilization.

A. CDGPS Fair Scheduling for Fixed Capacity

The basic principle of GPS is to assign a fixed positive real number (namely weight), instead of a fixed bandwidth, to each flow, and to dynamically allocate bandwidth for all the flows according to their weights and traffic load. Consider that N packet flows are sharing a network link with a total capacity C. Let the weight for flow i be ϕ_i. Let $S_i(\tau, t)$ denote the amount of flow i traffic that is served during an interval $(\tau, t]$. If all the flows are continuously served by the GPS server during the interval $(\tau, t]$, then $S_i(\tau, t)$ is exactly proportional to ϕ_i. Due to the burstiness of packet traffic, a user may not have packets to transmit during an interval and the unused bandwidth can be distributed among all of the backlogged sessions in proportion to their individual weight ϕ_i. This makes the GPS server efficient and fair in the bandwidth allocation. Furthermore, it has been proved that a tight delay bound can be guaranteed by the GPS server if a traffic flow is shaped by a leaky-bucket regulator [1]. Given that the shaped flow i is constrained by (σ_i, ρ_i) and $\phi_i C / \sum_{j=1}^{N} \phi_j \geq \rho_i$, the delay bound is σ_i / ρ_i. With the bounded delay guarantee, the GPS server can effectively support the QoS requirements of real-time traffic. It should be mentioned that the ideal GPS is not realizable in practical systems.

The CDGPS scheme is based on the GPS fair scheduling discipline [1], and developed for the rate-scheduled wideband DS-CDMA system. Fig. 3 shows the queuing model of the CDGPS scheme, where the total uplink capacity C is shared by N flows. In the CDGPS scheme, the uplink capacity C is in terms of total affordable throughput, and is assumed fixed. To insure that the uplink system will not be overloaded during the dynamic scheduling, C has to be chosen as the lower bound of the soft capacity, i.e., given the nominal capacity $(1-\delta)$, C can be solved from

$$N \times \frac{\frac{\left(\frac{E_b}{I_e}\right)_0}{NW} C}{1 + \frac{\left(\frac{E_b}{I_e}\right)_0}{NW} C} = 1 - \delta \tag{9}$$

Note that (9) corresponds to the case that all the N users are persistently and simultaneously transmitting at the same channel rate (e.g., the point A in Fig. 2), which gives a safe but conservative estimate of the uplink capacity. Each flow maintains a connection with link rate $R_i(k)$ during the kth time slot. The sum of $R_i(k)$ over all the users should not exceed C. For any flow i, it is entering a single-server queuing system with service rate $R_i(k)$. Different from the conventional single-server queue, the service rate $R_i(k)$ may be adjusted periodically. For each slot, the scheduler allocates adequate service rates to the N flows. The following is a detailed description of CDGPS.

1) CDGPS Rate Scheduling Procedure: Let $S_i(k)$ be the amount of flow i traffic that would be served during the time slot k. According to the GPS scheduling discipline, the following inequality should hold for any flow i that is continuously backlogged in the time slot k.

$$\frac{S_i(k)}{S_j(k)} \geq \frac{\phi_i}{\phi_j}, \quad j = 1, 2, \ldots, N. \tag{10}$$

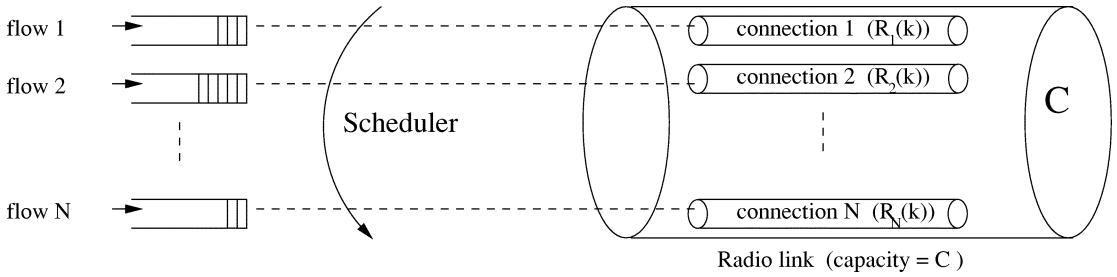

Fig. 3. Queueing model of CDGPS scheme.

Let one scheduling period, i.e., the slot length, be T. The CDGPS scheduler allocates each $R_i(k)$ using the following steps:

Step 1) Let $B_i(k)$ be the total amount of service request from user i for the time slot k. $B_i(k)$, $i = 1, 2, \ldots, N$, can be determined as follows:

$$B_i(k) = Q_i(\tau_k) + r_i(k)T \tag{11}$$

where τ_k is the end time of the slot $k-1$, and $Q_i(\tau_k)$ is the backlogged traffic at time τ_k; $r_i(k)$ is the estimated traffic arrival rate of flow i during time slot k and can be estimated from the past traffic measurement [18].

Step 2) Based on $B_i(k)$, $i = 1, 2, \ldots, N$, determine each $S_i(k)$ which is the expected amount of service received by the ith user as follows.

a) If $B_i(k) = 0$, then $S_i(k) = 0$.

b) If $B_i(k) > 0$, then $S_i(k) = g_i T$, where $g_i = \phi_i C / \sum_{j=1}^{N} \phi_j$ is the minimum rate guaranteed to user i.

c) If $\sum_i S_i(k) < CT$, then the remaining network resource is distributed to the users who expect more than their guaranteed service $g_i T$. The distribution of the remaining network resource should be proportional to each user's weight ϕ_i, according to the GPS service discipline.

Step 3) The allocated channel rate to user i can be determined by $R_i(k) = S_i(k)/T$.

2) Delay Bound: Let traffic flow i be regulated by a Leaky-Bucket with token buffer size σ_i and token generation rate ρ_i. The following delay bound for the traffic flow i can be derived.

Theorem 1: If $g_i \geq \rho_i$, then the packet delay of flow i is bounded by

$$\text{Delay}_{\max} \leq \frac{\sigma_i}{g_i} + T. \tag{12}$$

Proof: See Section A of the Appendix.

3) Fairness: The CDGPS scheduler assures weighted fairness to heterogeneous traffic. It can guarantee each backlogged flow with at least the minimum rate g_i which is proportional to its weight ϕ_i. In addition, the excess bandwidth available from flows not using their guaranteed rates is distributed among all of the backlogged flows in each time slot in proportion to their individual weights. This results in short-term fairness during a slot period T. The long-term fairness property of the CDGPS scheme can be explained by comparing the delay bound (12) with the ideal GPS delay bound σ_i/g_i [1]. It can be seen that the difference between the two bounds is within T, which implies that in the long term the service provided to user i by the CDGPS scheduler is equivalent to that provided by the ideal GPS server. In other words, the CDGPS scheduler also assures weighted fairness in the long term. The weighted fairness feature of the proposed CDGPS scheme is important for providing differentiated services in the wideband DS-CDMA cellular networks, since different traffic flows can be isolated by assigning to them different weights when they are sharing the same radio resource.

B. Resource Allocation for Soft Capacity

The CDGPS scheme is based on a fixed uplink capacity C. The advantage of setting C to be a fixed value is that: 1) the QoS is easy to be controlled and guaranteed by limiting the total throughput and hence the total interference level of uplink users to a conservative amount and 2) the rate allocation procedure is simple. However, since C has to be set conservatively as the lower bound of the *soft* capacity in order to guarantee user QoS in the worst case, the uplink capacity may be used inefficiently. To overcome this problem, the following optimal resource allocation approach is developed for fair sharing of the soft capacity.

Let $B_i(k)$ be the amount (in bits) of flow i traffic requested for transmission in slot k, $R_i(k)$ and $\gamma_i(k)$ be the allocated channel rate and SIR of user i, respectively. Let $H_i(k)$ be the throughput of user i, i.e., the number of bits successfully transmitted during the kth slot. The effective transmission rate for user i after error control is $R_i(k) f((E_b/I_e)_i)$. Then, the throughput $H_i(k)$ of user i during slot k can be written as

$$H_i(k) = \min\left(R_i(k) f\left(\left(\frac{E_b}{I_e} \right)_i \right) T, B_i(k) \right). \tag{13}$$

The optimization problem is to choose the vectors $\underline{S} = \{\gamma_1(k), \ldots, \gamma_N(k)\}$ and $\underline{R} = \{R_1(k), \ldots, R_N(k)\}$, such that

$$\text{MAX}_{\underline{S},\underline{R}} \sum_{i=1}^{N} H_i(k) \tag{14}$$

subject to (E_b/I_e) constraints

$$\left(\frac{E_b}{I_e}\right)_i \geq \left(\frac{E_b}{I_e}\right)_0, \quad i = 1, \ldots, N \tag{15}$$

capacity constraint (7) and GPS fairness constraints which state that, if $R_i(k)f((E_b/I_e)_i)T < B_i(k)$, then

$$\frac{\gamma_i(k)}{\gamma_j(k)} \geq \frac{\phi_i}{\phi_j}, \quad j = 1, \ldots, N. \tag{16}$$

Theorem 2: Let SIR vector $\underline{S}^* = \{\gamma_1(k)^*, \ldots, \gamma_N^*(k)\}$, rate vector $\underline{R}^* = \{R_1^*(k), \ldots, R_N^*(k)\}$ and throughput $H_i^*(k)$ be the optimal solutions of (14). Then, at least one of the following statements is true.

1) $H_i^*(k) = B_i(k)$ for all $i = 1, \ldots, N$.
2) The capacity constraint (7) is met with equality, i.e., $\sum_{i=1}^N \gamma_i^*(k)/[1+\gamma_i^*(k)] = 1-\delta$, for all $i = 1, \ldots, N$.

Theorem 2 indicates that the capacity constraint (7) is critical for the resource allocation. When the traffic load is high, the nominal capacity $(1-\delta)$ should be shared by users in such a way that (7) holds with equality and the GPS constraints (16) are satisfied. Based on Theorem 2, the following resource allocation procedure is proposed to find the optimal SIR vector $\underline{S}^*$ and rate vector $\underline{R}^*$. First, $B_i(k)$ is converted to the bandwidth request in terms of SIR for each user. Second, the scheduler checks if the total bandwidth requests exceed the nominal capacity $(1-\delta)$. If the total bandwidth requests do not exceed $(1-\delta)$, then each user can be granted the bandwidth requested. Otherwise, the SIRs for all the users are computed iteratively. In each iteration, a fair share of the available capacity will be computed for each user whose SIR has not been determined, according to its GPS weight and the capacity constraint (7), and compared with its requested bandwidth. For the user whose requested bandwidth is less than its fair share of the available capacity computed in any iteration, its request will be fully granted; otherwise, the user's SIR will be determined in a later iteration. In the final iteration, each remaining user will be given a fair share of the remaining capacity, which is normally no more than its request bandwidth.

1) Soft-Capacity Allocation Procedure: Let $S_i(k) = \gamma_i(k)$ be the allocated uplink resource for user i in slot k. Let $S_{\text{req},i}(k)$ be the requested SIR of user i, G_i be a guaranteed bandwidth (also in terms of SIR) for user i, C_n be the available capacity in the nth iteration, $\mathcal{B}$ be a set of users whose bandwidth request is larger than its fair share of C_n in each iteration.

Step 1) Compute $S_{\text{req},i}(k)$ in slot k, $i = 1, \ldots, N$.

$$S_{\text{req},i}(k) = \left(\frac{E_b}{I_e}\right)_0 \frac{B_i(k)}{f\left(\left(\frac{E_b}{I_e}\right)_0\right)TW}. \tag{17}$$

Step 2) Determine $S_i(k)$, $i = 1, \ldots, N$, as follows.

1) If $\sum_{i=1}^N S_{\text{req},i}(k)/[1+S_{req,i}(k)] < 1-\delta$, then $S_i(k) = S_{\text{req},i}(k)$, $i = 1, \ldots, N$, and go to step 3. Otherwise,
2) If any $S_{\text{req},i}(k) \leq G_i$, where

$$G_i = \frac{\phi_i C_0}{\sum_{j=1}^N \phi_j} \tag{18}$$

and C_0 is defined as the positive solution of the following equation:

$$\sum_{i=1}^N \frac{\phi_i C_0}{1+\phi_i C_0} = 1-\delta \tag{19}$$

then $S_i(k) = S_{\text{req},i}(k)$ and define a set $\mathcal{B} = \{j : S_{\text{req},j}(k) > G_j\}$.

3) Perform the following iteration until all $S_j(k)$, $j \in \mathcal{B}$, are assigned. In the nth iteration, find C_n to be the positive solution of the following equation:

$$\sum_{i\in\mathcal{B}} \frac{\frac{\phi_i C_n}{\sum_{j\in\mathcal{B}}\phi_j}}{1+\frac{\phi_i C_n}{\sum_{j\in\mathcal{B}}\phi_j}} = 1-\delta-\sum_{i\notin\mathcal{B}} \frac{S_i(k)}{1+S_i(k)}. \tag{20}$$

If for any $i \in \mathcal{B}$, $S_{\text{req},i}(k) \leq \phi_i C_n / \sum_{j\in\mathcal{B}}\phi_j$, then $S_i(k) = S_{\text{req},i}(k)$, remove i from $\mathcal{B}$ and continue to the $(n+1)$th iteration; else, let $S_i(k) = \phi_i C_n / \sum_{j\in\mathcal{B}}\phi_j$ for all $i \in \mathcal{B}$ and stop the iteration.

Step 3) The allocated channel rate to user i can be determined by

$$R_i(k) = \frac{WS_i(k)}{\left(\frac{E_b}{I_e}\right)_0}. \tag{21}$$

2) GPS Fairness: The above resource allocation procedure is aimed at finding a solution for (14), i.e., to maximize data throughput while satisfying the GPS fairness constraint (16). Based on Theorem 2, the capacity constraint (7) is addressed explicitly by (19) and (20) in the step 2 of the resource allocation procedure. The following theorem shows that the GPS fairness constraint (16) is also guaranteed.

Theorem 3: (GPS Fairness) Flow i is called a *greedy* flow, if the requested traffic amount $B_i(k)$ of flow i is always larger than the throughput $H_i(k)$ that the uplink system can support for any time slot. For any greedy flow i in slot k, it can be guaranteed that

$$\frac{S_i(k)}{S_j(k)} \geq \frac{\phi_i}{\phi_j}, \qquad j \neq i.$$

C. Credit-Based CDGPS (C-CDGPS)

Although the weighted fairness can be guaranteed by the soft-capacity resource allocation procedure in each time slot, in some situation the short-term fairness may not be necessary. For instance, for delay-insensitive data users, a short-term unfair allocation of resource in a slot may not degrade their QoS in terms of throughput. On the other hand, the total data throughput in the uplink may be increased if only some of the data users are allowed to transmit at their maximal rate, and the other users to transmit at a minimum rate [16]. In other words, a short-term unfairness may benefit user QoS and increase resource utilization, as long as the long-term fairness can still be guaranteed. In the following, a C-CDGPS scheduling scheme is introduced to improve the throughput of the delay-insensitive data users

and, at the same time, to guarantee the long-term fairness. In this scheme, a credit counter is associated with each traffic flow, and service for traffic flows is arranged by taking into account both GPS discipline and the order of their credits. The credit of a user is increased if it does not receive as much as its deserved fair share of bandwidth in a time slot and is decreased if it receives more than its fair share. With the GPS discipline, the minimum bandwidth requirement of each user is guaranteed by assigning a weight to it. For a delay-sensitive user, a large weight is assigned so that the guaranteed bandwidth is large enough to assure packet delay to be bounded. In the case that a user does not use up its guaranteed bandwidth, the extra bandwidth is allocated to the other greedy users in decreasing order of their credits. Note that there has to be a limit on the credit accumulation, otherwise, a user with very large credit can capture bandwidth for a long period of time. With the C-CDGPS, it can be shown that a credit bound exists so that the long-term fairness can be guaranteed.

Let $K_i(k)$ be the amount of credits that user i has by the end of slot k, $K_i(0) = 0$. Define a variable $e(k)$ to be the total amount of extra bandwidth which can be shared by greedy flows besides their guaranteed bandwidth. Note that $e(k)$ has the same unit as that of SIR. Define a set $\mathcal{B}_0 = \{j : S_{req,j}(k) > G_j\}$ for the time slot k, where $S_{\text{req},j}(k)$ and G_j are given by (17) and (18), respectively. The flows in set $\mathcal{B}_0$ are called greedy flows. Then, the extra bandwidth is

$$e(k) = \sum_{i \in \mathcal{B}_0} (S_i(k) - G_i) \tag{22}$$

and the credits of the flows in set $\mathcal{B}_0$ are updated by

$$K_i(k) = K_i(k-1) + e(k)\frac{\phi_i}{\sum_{j \in \mathcal{B}_0} \phi_j} - (S_i(k) - G_i) \quad i \in \mathcal{B}_0. \tag{23}$$

Note that $e(k)$ is nonnegative since the greedy flow i in set $\mathcal{B}_0$ can always obtain $S_i(k)$, which is no less than G_i. According to the GPS discipline, $e(k)$ should be distributed proportionally to ϕ_i. If a greedy flow does not receive its fair share of $e(k)$, it will receive more credits.

The detailed C-CDGPS scheduling procedure is presented as follows. The strategy is to give the greedy flow that has the largest amount of credits the first priority to obtain extra bandwidth, i.e., in the first iteration, all the other greedy flows are only given the guaranteed bandwidth. If the user with the highest credit does not use up $e(k)$, the user with the second highest credit will be given a share of $e(k)$ in the next iteration. The iterations continue until the extra bandwidth is used up and $K_i(k)$ will be updated by (23).

1) C-CDGPS Rate Scheduling Procedure:

Step 1) Convert traffic request to SIR target, i.e., compute $S_{\text{req},i}(k)$ using (17).

Step 2) Determine $S_i(k)$, $i = 1, 2, \ldots, N$, as follows:

a) If $\sum_{i=1}^{N} S_{\text{req},i}(k)/[1 + S_{\text{req},i}(k)] < 1 - \delta$, then $S_i(k) = S_{\text{req},i}(k)$, $i = 1, 2, \ldots, N$ and go to step 3. Otherwise,

b) If any $S_{\text{req},i}(k) \le G_i$, then $S_i(k) = S_{\text{req},i}(k)$, and define greedy flow set $\mathcal{B}_0 = \{j : S_{\text{req},j}(k) > G_j\}$.

c) Let $\mathcal{B} = \mathcal{B}_0$. Sort the elements in set $\mathcal{B}$ in decreasing order of user credits $K_i(k-1)$, $i \in \mathcal{B}$. Perform the following iterations until all $S_j(k)$, $j \in \mathcal{B}$, are assigned.

In the nth iteration, let t be the first item in $\mathcal{B}$, i.e., flow t has the highest credit so far. Let $S_t(k)$ be the solution of the following equation:

$$\frac{S_t(k)}{1 + S_t(k)} = 1 - \delta - \sum_{i \notin \mathcal{B}} \frac{S_i(k)}{1 + S_i(k)} - \sum_{i \in \mathcal{B}, i \neq t} \frac{G_i}{1 + G_i}. \tag{24}$$

If $S_{\text{req},t}(k) \le S_t(k)$, then $S_t(k) = S_{\text{req},t}(k)$, remove t from $\mathcal{B}$ and continue to $(n+1)$th iteration; otherwise, let $S_i(k) = G_i$ for all $i \neq t$ and $i \in \mathcal{B}$, and stop the iteration.

d) Compute $e(k)$ by (22) and update the credit for flow i, $i \in \mathcal{B}_0$, by (23). For $j \notin \mathcal{B}_0$, let $K_j(k) = K_j(k-1)$.

Step 3) The allocated channel rate to user i can be determined by

$$R_i(k) = \frac{W S_i(k)}{\left(\frac{E_b}{I_e}\right)_0}.$$

2) Fairness (C-CDGPS): In the C-CDGPS, the credit associated with each flow reflects the difference between the actual resource allocated to it and its desired fair share of the resource. The amount of credits can be positive, zero, or negative. Positive credit means that the corresponding flow has been given less resource than its fair share, and negative credit means that the flow has used more resource than its fair share. Therefore, fairness is not guaranteed within each slot by the C-CDGPS rate-allocation procedure. However, the short-term unfairness can be limited if the credits of greedy flows are bounded. The following theorem gives the credit bound for the fairness.

Theorem 4: Let $\mathcal{Y}$ be a set of greedy traffic flows, i.e., any flow in $\mathcal{Y}$ always has enough backlogged traffic to satisfy whatever transmission rate allocated to it. Assume that any flow $j \notin \mathcal{Y}$ requests bandwidth less than its guaranteed bandwidth, i.e., $S_{\text{req},j}(k) \le G_j$ for any slot k. Then, there exist positive constants $K_{\max}$ and K_{diff} such that, for any slot k:

1) $K_i(k) < K_{\max}$ for any $i \in \mathcal{Y}$;
2) The difference of the maximum and the minimum credits of any flow in $\mathcal{Y}$ is K_{diff} or less.

Proof: See Section B of the Appendix.

Since $K_{\max}$ and K_{diff} are bounded, Theorem 4 implies that the weighted fairness constraint (16) can be approximately satisfied in the long term. Although the GPS fairness is not strictly guaranteed by C-CDGPS, the QoS requirements in terms of delay and throughput can still be satisfied. The following theorem gives the delay bound guaranteed by C-CDGPS. The throughput performance is shown by simulation in Section V.

3) Delay Bound (C-CDGPS): Let traffic flow i be regulated by a leaky-bucket with token buffer size σ_i and token generation rate ρ_i. The following delay bound can be guaranteed to flow i.

Theorem 5: If $G_i W/(E_b/I_e)_0 f((E_b/I_e)_0) \ge \rho_i$, then the packet delay of flow i is bounded by

$$\text{Delay}_{\max} \le \frac{\sigma_i}{\rho_i} + 2T. \tag{25}$$

Proof: See Section C of the Appendix.

D. Implementation Issues

Practical implementation of the CDGPS schemes may involve a more sophisticated system which requires more detailed design of MAC protocols, power control mechanism, signalling formats, and transport formats (e.g., spreading factors and FEC code rates), among others. Although elaborating system designs are not within the scope of this paper, we shall discuss a few implementation issues as follows.

In general, a CDGPS or C-CDGPS scheduler should adjust channel rates for the uplink users on a slot-by-slot basis. Both the corresponding SIR thresholds and spreading factors of uplink channels should be changed at the same time, so that the target E_b/I_0 is maintained. Theoretically, if the target E_b/I_0 does not change, the FEC code rate will not change. In practice, a more flexible approach is to adjust both the spreading factor and the FEC code rate to achieve different effective channel rates [18] Although this approach is not considered in the proposed analysis, the CDGPS schemes can be implemented in this way. On the other hand, the proposed CDGPS schemes are based on a centralized MAC scheme instead of the decentralized MAC scheme adopted in the current W-CDMA standard. Although the proposed centralized scheduling schemes may introduce additional signalling cost compared with that in the W-CDMA standard, it may have potentials for increasing the overall utilization of radio resources. Furthermore, the decentralized MAC protocol for in the current W-CDMA standard may be unified into the centralized MAC scheme to reduce the signalling cost. For example, the mobile user can autonomously decide the spreading factor of its uplink channel based on the feedback resource allocation information in regards to the SIR and the feedback signal may be simplified to only 1 bit, which indicates the increase or decrease of the SIR threshold. It should be mentioned that the signalling overhead for the centralized scheduling scheme is an important issue and should be carefully considered in practice. The signalling overhead should be minimized without compromising the performances of the centralized scheduling schemes. Design of efficient signalling protocols for the centralized scheduling schemes is left for future consideration.

Another practical issue is related to channel fading. The channel fading is not explicitly considered in the proposed scheduling schemes, because it is assumed to be mitigated by the power-control and FEC mechanisms. In practice, mobile users may be experiencing deep fading which cannot be fully compensated for by power control. In this case, the user entering a bad channel state may need to give up some capacity and require compensation when the channel reverts to a good state. The C-CDGPS scheme can be extended to address this issue by introducing a compensation mechanism similar to that proposed in [6].

V. Numerical Results

In this section, simulation results are presented to demonstrate the performance of the proposed CDGPS and C-CDGPS schemes in terms of delay, fairness, and system utilization. A one-step traffic rate estimation approach [18] is used for the CDGPS and C-CDGPS schemes in the simulations. The scheduling period T is 10 ms for both schemes. Three simulations are performed. The first simulation is presented to demonstrate the performance of the CDGPS scheme for a fixed capacity. In the second and third simulations, the C-CDGPS scheme is compared with the CDGPS scheme for the CDMA uplink with soft-capacity, in homogeneous and heterogeneous traffic environments, respectively.

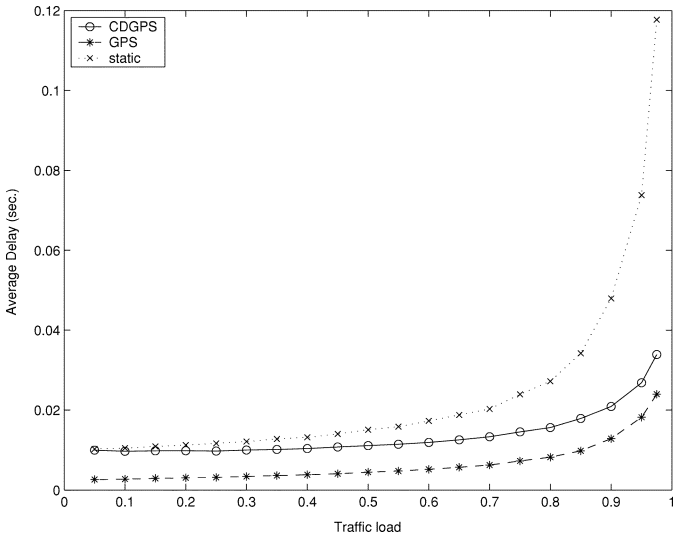

Fig. 4. Average delay comparison.

In the first simulation, the uplink capacity is assumed to be a constant $C = 2$ Mb/s. Four homogeneous best-effort packet data flows are considered, and assigned the same weight. Each of these flows is modeled by a Poisson process with average arrival rate λ and packet length L, shaped by a leaky-bucket regulator with parameter (σ, ρ). In the simulation, $L = 5120$ bits, $\sigma = 20L$, and $\rho = C/4$. λ can be varied in order to change the system load. A hypothetical GPS scheduler with service rate $C = 2$ Mb/s is also simulated, providing a performance benchmark. In addition, a static rate-allocation scheduler is simulated, which assigns a fixed channel rate that equals to the guaranteed rate g_i in the CDGPS scheme, to each packet flow. Figs. 4 and 5 show the average delay and maximum delay, respectively, with different system loads. In these figures, the system load is normalized to be $\mathrm{Load} = 4*\lambda/C$. It can be seen that the delay performance of CDGPS is similar to that of GPS, and better than that of the static scheme. As expected, the idealized GPS can achieve lower average and maximum delays. This is because the GPS performs hypothetically bit-by-bit scheduling which can instantly respond to the traffic variation. The difference between the delay performance of CDGPS and GPS is mainly due to the scheduling period T and does not significantly change with load. The simulated maximum delay of CDGPS is less than the theoretical delay bound from (12), which is 0.21 s. The average and maximum delays with the static scheme are close to that with CDGPS when the system load is light, but increase much faster than that with CDGPS when the system load increases. This is because CDGPS is more flexible in allocating bandwidth and can make use of idle network resource to improve delay performance. Fig. 6 gives the throughput comparison of CDGPS and GPS. It can be seen that both CDGPS and GPS

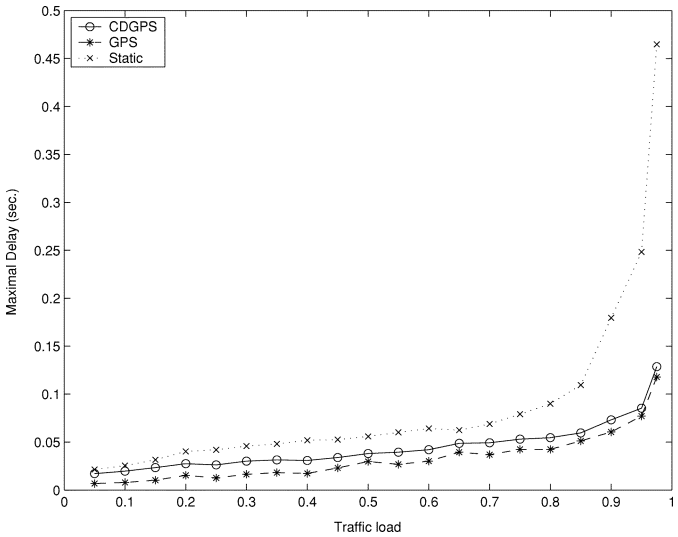

Fig. 5. Maximum delay comparison.

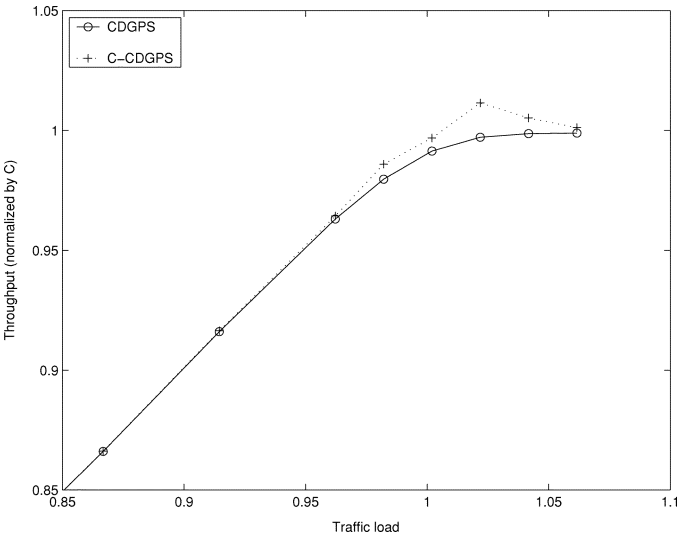

Fig. 7. Throughput comparison: CDGPS and C-CDGPS.

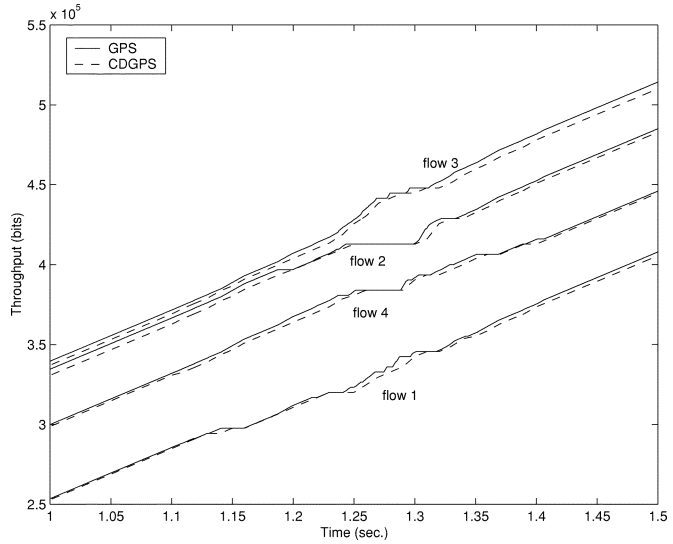

Fig. 6. Throughput comparison: CDGPS and GPS.

schedulers can fairly allocate service rates to the different flows, according to their assigned weights. When the traffic arrival rates fluctuate (e.g., in the period from 1.25 to 1.35 s in Fig. 6), the CDGPS scheduler responds instantly to the traffic variation as the ideal GPS scheduler does. Therefore, the throughput by using the CDGPS scheduler is very close to that of the ideal GPS scheduler.

In the second and third simulations, the simulated wideband DS-CDMA system supports three classes of packetized traffic, voice, video, and best-effort data. The system bandwidth W is 5 Mb/s. The $(E_b/I_e)_0$ is 10 dB. Without loss of generality, perfect SIR-based power control and $f((E_b/I_e)_0) = 1$ are assumed. For the CDGPS, a constant capacity C is estimated by solving (9). A typical network scenario is considered, where the users are uniformly distributed in the area covered by the cellular network, and all the cells have the same traffic load and user mobility status. For simplicity, let the channel characteristics such as path loss exponent, shadowing, and fading parameters be the same for all the cells, and the nominal capacity $(1-\delta)$ be a constant, i.e., $I_{\text{inter}} + P_n$ be invariant. $(1-\delta)$ can be determined by the following analytical approach for the considered network scenario. Let $P_u = I_{\text{inter}}/\eta$, where η is a constant that can be determined from the path loss exponent, shadowing parameters and geometric size of the uniform cellular networks. Some typical values of η have been given by analysis and simulation in [25]. Substituting I_{inter} by ηP_u into (8), we have

$$P_u = (P_n + \eta P_u)\frac{1-\delta}{\delta}. \tag{26}$$

Neglecting the thermal noise power P_n, we have

$$\delta = \frac{\eta}{1+\eta}. \tag{27}$$

For a macrocell networks with path loss exponent of four, the typical η is about 0.5 and $\delta = 0.33$. Therefore, the nominal capacity is $(1-\delta) = 0.67$ in this case.

In the second simulation, four homogeneous Poisson data traffic flows are simulated, each of which is guaranteed 1/4 of the capacity C. Fig. 7 shows the throughput comparison of the C-CDGPS and CDGPS. The traffic load (normalized by C) is the sum of average arrival rates of the four data flows. It is shown that C-CDGPS can improve the uplink throughput (normalized by C), especially when traffic load is around one. Figs. 8 and 9 show average delay and maximum delay, respectively. It can be seen that, when the traffic load is high (around one), the average delay and maximum delay can be reduced up to 70% and 50%, respectively, by C-CDGPS as compared with the CDGPS. This implies that the short-term unfairness introduced by the credit mechanism can actually benefit the delay performance in the uplink. Fig. 10 shows the maximum and minimum credits of the four data flows. With the increase of traffic load, the maximum and minimum credits are nearly constant and bounded, which agrees with Theorem 4. The irregularity of the curves shown in Fig. 10 is due to the randomness of the arrival process of traffic flows. With bounded credits, the short-term unfairness is also bounded and the long-term fairness can be guaranteed.

In the third simulation, ten voice flows, one VBR video flows and four best-effort data flows are simulated. Each voice flow is generated by using an ON–OFF model, where the activity factor

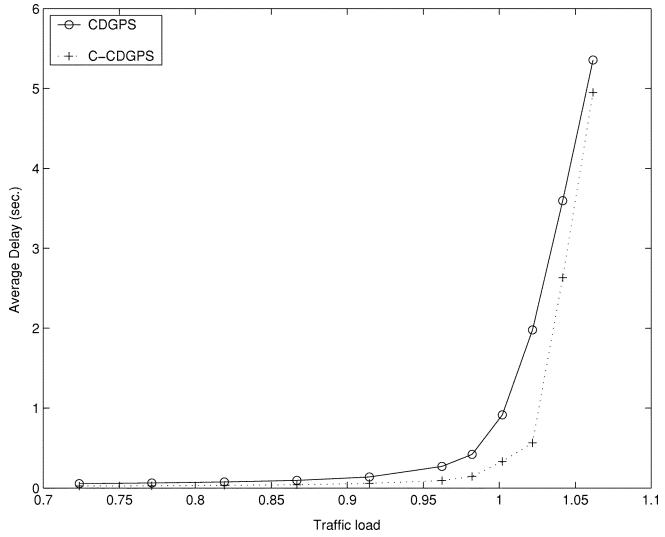

Fig. 8. Average delay comparison: CDGPS and C-CDGPS.

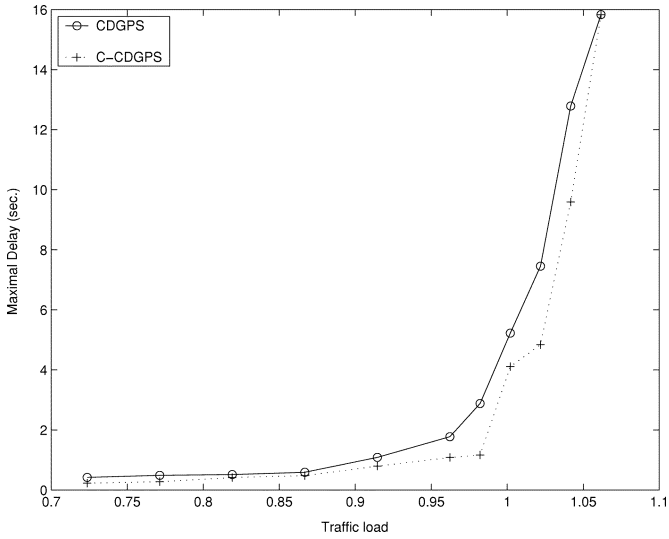

Fig. 9. Maximum delay comparison: CDGPS and C-CDGPS.

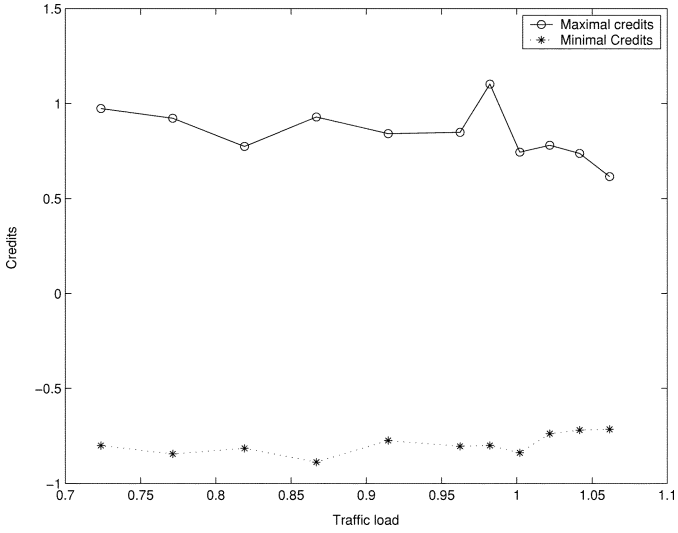

Fig. 10. Maximum and minimum credits in C-CDGPS.

TABLE I
DELAY PERFORMANCE OF HETEROGENEOUS TRAFFIC SCHEDULED BY CDGPS

Flow ID	Traffic type	Average delay (sec.)		Maximum delay (sec.)		Theoretical delay bounds
		CDGPS	C-CDGPS	CDGPS	C-CDGPS	
1	Voice	0.01459	0.01459	0.01962	0.01962	0.030
2		0.01541	0.01541	0.01973	0.01973	0.030
3		0.01481	0.01481	0.01974	0.01974	0.030
4		0.01472	0.01472	0.01999	0.01999	0.030
5		0.01492	0.01492	0.01979	0.01979	0.030
6		0.01509	0.01509	0.01994	0.01994	0.030
7		0.01468	0.01468	0.01994	0.01994	0.030
8		0.01526	0.01526	0.01971	0.01971	0.030
9		0.01541	0.01541	0.01899	0.01899	0.030
10		0.01498	0.01498	0.01992	0.01992	0.030
11	Video	0.01598	0.01068	0.1143	0.02041	0.120
12	Data	9.610	0.04616	25.852	0.4303	N/A
13		10.483	0.04470	27.343	0.3823	N/A
14		10.377	0.04466	27.743	0.3730	N/A
15		10.525	0.04457	28.831	0.3203	N/A
Utilization		0.9995	1.1551			

is 0.4, and packets are generated in the ON-state at a constant rate $R_{vo} = 8$ kb/s $= 100$ Packets/s with the packet length $L_{vo} = 80$ bits. The weight assigned to each voice flow is $\phi_{vo} = 8$. The VBR video traffic is generated by using an eight-state MMPP model and a leaky-bucket regulator. The average duration in each state is 40 ms, which is equivalent to the length of one frame of the video sequence with frame rate of 25 frame/s. The maximum rate of the VBR video is 384 kb/s and the average rate is $\rho_{vi} = 180$ kb/s. The parameter of the leaky-bucket regulator for user i is (σ_{vi}, ρ_{vi}), where $\sigma_{vi} = 18$ kbits. The weight assigned to the video flow is $\phi_{vi} = 260$. Each best-effort data flow is generated by using a Poisson process model, with packet size 1600 bits and average arrival rate 50 kb/s. The weight of each best-effort data flow is $\phi_{be} = 2$.

Simulation results for the heterogeneous traffic are summarized in Table I, where theoretical delay bounds obtained from (25) for real-time traffic are also given for comparison. It can be seen that, with the weighted fair scheduling by CDGPS and C-CDGPS, the real-time traffic flows, voice and video, can achieve lower packet delays than the delay-insensitive Poisson data traffic. With both CDGPS and C-CDGPS, the maximum packet delays of the real-time traffic are less than the theoretical bounds. For the voice traffic, the delay performances are the same for both CDGPS and C-CDGPS schemes. This is because the channel rate of voice traffic is guaranteed by assigning a relative large weight to each voice flow. For the video flow, the guaranteed rate is less than the maximum arrival rate. Therefore, the maximum delay of video traffic with the CDGPS scheme is quite large. However, the maximum delay, as well as average delay of video traffic can be significantly reduced by C-CDGPS. The average delay of the Poisson data traffic can be also significantly reduced by C-CDGPS, which implies that the throughput of data traffic with C-CDGPS may further increase if the arrival rate increases. Network resource (bandwidth) utilization, in terms of total uplink throughput normalized by the constant capacity C, is also increased by C-CDGPS as compared with the CDGPS.

VI. CONCLUSION

Efficient dynamic fair scheduling schemes have been proposed to support QoS of multimedia traffic in the uplink of

wideband DS-CDMA cellular networks. The analysis and simulation results show that bounded delay can be provisioned for real-time application by using the CDGPS service discipline with a high resource utilization, and weighted fairness can be assured among different users. The resource utilization and delay can be further improved by C-CDGPS which exploits the *soft* uplink capacity. The CDGPS schemes are easier to implement than the conventional GPS-based time-scheduling schemes, and are more suitable for the wideband DS-CDMA systems.

APPENDIX

A. Proof of Theorem 1

We first prove the following lemma.

Lemma 1: Let $Q_i(t)$ denote the flow i backlog, i.e., the queue length of flow i, Q_i^* the maximum of $Q_i(t)$. If $g_i \geq \rho_i$, where g_i is given by (18), then the maximum backlog $Q_i^* \leq \sigma_i + \rho_i T$.

Proof of Lemma 1: Without loss of generality, let flow i start to backlog at time t_0, $\tau_{k-1} < t_0 \leq \tau_k$, where τ_{k-1} and τ_k are the beginning of slot $(k-1)$ and slot k, respectively. Assume t^* is the time when the flow i backlog reaches the maximum Q_i^*. Define the arrival function $A_i(t_0, t)$ to be the total amount of arrived traffic of flow i during interval $[t_0, t]$. Since flow i is regulated by the leaky-bucket, then

$$A_i(t_0, t^*) \leq \sigma_i + \rho_i(t^* - t_0). \tag{28}$$

Let service function $S_i(t_0, t)$ represent the total amount of serviced traffic during interval $[t_0, t)$. According to CDGPS service discipline, a minimum service rate g_i is assigned to flow i at time τ_k. Therefore

$$S_i(t_0, t^*) \geq \max(0, (t^* - \tau_k)g_i) \tag{29}$$

where $\max(x, y)$ is the larger one of x and y.

Since $Q_i^* = A_i(t_0, t^*) - S_i(t_0, t^*)$ and $g_i \geq \rho_i$, we have

$$\begin{aligned} Q_i^* \leq & \sigma_i + \rho_i(t^* - t_0) - \max(0, (t^* - \tau_k)g_i) \\ \leq & \sigma_i + \rho_i(\tau_k - t_0). \end{aligned} \tag{30}$$

Further more, since $\tau_{k-1} < t_0 \leq \tau_k$ and $\tau_k - \tau_{k-1} = T$, we have $Q_i^* \leq \sigma_i + \rho_i T$. □

With Lemma 1, we proceed to prove Theorem 1.

Proof of Theorem 1: Let $D_i(t)$ be the delay experienced by flow i traffic arrived at time t, $t \geq t_0$. Since flow i starts to backlog at time t_0, $\tau_{k-1} < t_0 \leq \tau_k$, we have

$$A_i(t_0, t) = S_i(t_0, t + D_i(t)) \tag{31}$$

where $A_i(t_0, t)$ and $S_i(t_0, t)$ are the arrival function and service function of flow i, respectively. Since the backlog $Q_i(t)$ is

$$Q_i(t) = A_i(t_0, t) - S_i(t_0, t) \tag{32}$$

and

$$S_i(t_0, t + D_i(t)) = S_i(t_0, t) + S_i(t, t + D_i(t)) \tag{33}$$

we have

$$Q_i(t) = S_i(t, t + D_i(t)). \tag{34}$$

If $t \geq \tau_k$, then $S_i(t, t + D_i(t)) \geq g_i D_i(t)$. This is because the CDGPS scheduler can guarantee a minimum rate g_i to flow i during interval $[t, t+D_i(t)]$. From Lemma 1, $Q_i(t) \leq \sigma_i + \rho_i T$. Thus, we have

$$\begin{aligned} D_i(t) \leq & \frac{\sigma_i + \rho_i T}{g_i} \\ \leq & \frac{\sigma_i}{g_i} + T. \end{aligned} \tag{35}$$

If $t < \tau_k$, we have

$$Q_i(t) \leq \sigma_i + (t - t_0)\rho_i \tag{36}$$

and

$$S_i(t, t + D_i(t)) \geq [D_i(t) - (\tau_k - t)]g_i. \tag{37}$$

Then

$$\begin{aligned} D_i(t) \leq & \frac{\sigma_i + (t - t_0)\rho_i}{g_i} + (\tau_k - t) \\ \leq & \frac{\sigma_i}{g_i} + (\tau_k - t_0) \\ \leq & \frac{\sigma_i}{g_i} + T. \end{aligned} \tag{38}$$

From (35) and (38), we conclude that the maximum delay is bounded by $\sigma_i/g_i + T$.

B. Proof of Theorem 4

In the rate-scheduling procedure, $\mathcal{B}_0 = \mathcal{Y}$ for any slot k, and all the flows that are not in $\mathcal{Y}$ are allocated bandwidth in Step 2b of C-CDGPS in Section IV-C1 without changing their credits. Then, from (22) and (23) we have

$$\sum_{i \in \mathcal{Y}} K_i(k) = \sum_{i \in \mathcal{Y}} K_i(k-1) = 0 \tag{39}$$

since $K_i(0) = 0$.

From (39), if $K_{\max}$ exists, there must exist $K_{diff} \leq 2K_{\max}$. In what follows, we will show that $K_{\max} \leq (|\mathcal{Y}| + 1)S_{\max}$, where $|\mathcal{Y}|$ denotes the number of flows in $\mathcal{Y}$, and $S_{\max}$ is the maximum total bandwidth (in terms of SIR) that the system can allocate to all the uplink users.

To prove that $K_{\max}$ exists, we use an approach which is similar to that in [26, Ch. 6]. Consider the summation of all the nonnegative credits

$$Z(k) = \sum_{i \in \mathcal{Y}} \max(0, K_i(k)). \tag{40}$$

We will prove that $Z(k) \leq (|\mathcal{Y}| + 1)S_{\max}$, for any k. Since $Z(k)$ is the sum of nonnegative terms, each term can not exceed $Z(k)$ and, therefore, $K_{\max} \leq (|\mathcal{Y}| + 1)S_{\max}$.

To prove that $Z(k)$ is bounded, rewrite (23) as

$$K_i(k) = K_i(k-1) + F_i(k) - S_i(k) \tag{41}$$

where

$$F_i(k) = e(k)\frac{\phi_i}{\sum\limits_{j \in \mathcal{Y}} \phi_j} + G_i \tag{42}$$

and consider that

$$\begin{aligned}\Delta Z =& Z(k) - Z(k-1)\\ =& \sum_{i\in\mathcal{Y}}[\max(0, K_i(k)) - \max(0, K_i(k-1))]\\ \le& \sum_{i\in\mathcal{Y}|K_i(k)>0}[K_i(k) - K_i(k-1)]\\ =& \sum_{i\in\mathcal{Y}|K_i(k)>0}[F_i(k) - S_i(k)]. \end{aligned} \tag{43}$$

$Z(k) \le (|\mathcal{Y}|+1)S_{\max}$, $k \ge 1$, can be proved by induction. The initial step of the induction is $Z(0) = 0$, since $K_i(0) = 0$ for all $i \in \mathcal{Y}$. For each induction step, there are two cases.

Case 1) $Z(k-1) \le |\mathcal{Y}|S_{\max}$. Then, from (43) and the fact that $\sum_{i\in\mathcal{Y}} F_i(k) = \sum_{i\in\mathcal{Y}} S_i(k)$, we have $\Delta Z \le \sum_{i\in\mathcal{Y}|K_i(k)>0} F_i(k) \le \sum_{i\in\mathcal{Y}} F_i(k) = \sum_{i\in\mathcal{Y}} S_i(k) \le S_{\max}$. Therefore, $Z(k) \le Z(k-1) + \Delta Z \le (|\mathcal{Y}|+1)S_{\max}$ as required.

Case 2) $|\mathcal{Y}|S_{\max} < Z(k-1) \le (|\mathcal{Y}|+1)S_{\max}$. Let $K_m(k-1)$, $m \in \mathcal{Y}$, be the maximum credit in slot $k-1$. Since flow m is greedy, according to the rate-scheduling procedure of C-CDGPS, $S_m(k) \ge G_m$ and $S_j(k) \le G_j < F_j(k)$ for any $j \in \mathcal{Y}|j \ne m$.

Since $Z(k-1)$ is a summation of $|\mathcal{Y}|$ nonnegative terms, then $K_m(k-1) \ge Z(k-1)/\mathcal{Y} > S_{\max}$ and, thus, $K_m(k) \ge K_m(k-1) - S_m(k) > S_{\max} - S_m(k) \ge 0$. Then, we have

$$\begin{aligned}\Delta Z \le& \sum_{i\in\mathcal{Y}|K_i(k)>0}[F_i(k) - S_i(k)]\\ =& [F_m(k) - S_m(k)] + \sum_{j\in\mathcal{Y}|K_j(k)>0, j\ne m}[F_i(k) - S_i(k)]\\ <& [F_m(k) - S_m(k)] + \sum_{j\in\mathcal{Y} j\ne m}[F_i(k) - S_i(k)]\\ =& \sum_{i\in\mathcal{Y}}[F_i(k) - S_i(k)] = 0 \end{aligned} \tag{44}$$

since, for any $j \ne m$, $F_j(k) - S_j(k) > G_i - S_j(k) \ge 0$, and $\sum_{i\in\mathcal{Y}} F_i(k) = \sum_{i\in\mathcal{Y}} S_i(k)$. Therefore, $Z(k) < Z(k-1) < (|\mathcal{Y}|+1)S_{\max}$ as required.

In both cases, $Z(k)$ is bounded and, thus, both $K_{\max}$ and K_{diff} are bounded.

C. Proof of Theorem 5

Let $g_i = G_i W/(E_b/I_e)_0 f((E_b/I_e)_0)$ be the guaranteed service rate for greedy traffic, and consider a period during which the traffic flow i is greedy, i.e., the traffic request $B_i(k)$ for each slot in this period is greater than its guaranteed service $g_i T$. Then, according to the rate-allocation procedure of C-CDGPS, g_i can be guaranteed for any slot in this period. Following the same approach used for proving Theorem 1, the maximum delay for the Leaky-Bucket shaped traffic can be shown to be less than $\sigma_i/g_i + T$. If $B_i(k) < g_i T$ for some slots, according to the rate-allocation procedure of C-CDGPS, a service rate of $B_i(k)/T$ is guaranteed, i.e., the backlogged traffic at the start of this slot can be completely serviced in the same slot. In other words, if there are any extra delay due to the situation that the allocated service rate in some slot is less than g_i, then the extra delay will be less than T. Thus, the delay bound is $\sigma_i/g_i + 2T$. Since $g_i > \rho_i$, the delay bound can be also written as $\sigma_i/\rho_i + 2T$.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their helpful comments in improving the presentation of the paper.

REFERENCES

[1] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated sevices networks: The single-node case," *IEEE/ACM Trans. Networking*, vol. 1, pp. 344–357, June 1993.

[2] A. Demers and S. Shenker, "Analysis and simulation of a fair queuing algorithm," *Internetworking: Res. Exper.*, vol. 1, no. 1, pp. 3–26, 1990.

[3] D. Stiliadis and A. Varma, "Efficient fair queueing algorithms for packet-switched networks," *IEEE/ACM Trans. Networking*, vol. 6, pp. 175–185, Apr. 1998.

[4] P. Ramanathan and P. Agrawal, "Adapting packet fair queueing algorithms to wireless networks," in *Proc. ACM/IEEE MOBICOM'98*, Dallas, TX, Oct 1998, pp. 1–9.

[5] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. Networking*, vol. 7, pp. 473–89, Aug. 1999.

[6] Nandagopal, S. Lu, and V. Bharghavan, "A unified architecture for the design and evaluation of wireless fair scheduling algorithms," *Wireless Networks*, vol. 7, pp. 231–247, Aug. 2002.

[7] M. A. Arad and A. Leon-Garcia, "A generalized processor sharing approach to time scheduling in hybrid CDMA/TDMA," in *Proc. IEEE INFOCOM'98*, San Francisco, CA, Mar. 1998, pp. 1164–1171.

[8] A. J. Viterbi, *Principles of Spread Sprectrum Communication*. Reading, MA: Addison-Wesley, 1995.

[9] S.-J. Oh and K. M. Wasserman, "Dynamic spreading gain control in multiservice CDMA networks," *IEEE J. Select. Areas Commnun.*, vol. 17, pp. 918–927, May 1999.

[10] I. F. Akyildiz, D. A. Levine, and I. Joe, "A slotted CDMA protocol with BER scheduling for wireless mulimedia networks," *IEEE/ACM Trans. Networking*, vol. 7, pp. 146–58, Apr. 1999.

[11] O. Gurbuz and H. Owen, "Dynamic resource scheduling schemes for W-CDMA systems," *IEEE Commun. Mag.*, vol. 38, pp. 80–84, Oct. 2000.

[12] J. B. Kim and M. L. Honig, "Resource allocation for multiple classes of DS-CDMA traffic," *IEEE Trans. Veh. Technol.*, vol. 49, pp. 506–519, Mar. 2000.

[13] A. C. Kam, T. Minn, and K.-Y. S. Siu, "Supporting rate guarantee and fair access for bursty data traffic in W-CDMA," *IEEE J. Select. Areas Commun.*, vol. 19, pp. 2121–2130, Nov. 2001.

[14] O. Sallent, J. Perez-Romero, R. Agusti, and J. Sanchez, "Uplink RRM for conversational and interactive services in UTRA-FDD," in *Multiaccess, Mobility and Teletraffic for Wireless Communications*, X. Lagrange and B. Jabbari, Eds. Norwell, MA: Kluwer, 2002, vol. 6.

[15] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, pp. 150–154, Feb. 2001.

[16] S. Ramakrishna and J. M. Holtzman, "A scheme for throughput maximization in a dual-class CDMA system," *IEEE J. Select. Areas Commnun.*, vol. 16, pp. 830–844, Aug. 1998.

[17] S. Kumar and S. Nanda, "High data-rate packet communications for cellular networks using CDMA: Algorithms and performance," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 472–492, Mar. 1999.

[18] L. Xu, X. Shen, and J. W. Mark, "Dynamic bandwidth allocation with fair scheduling for WCDMA systems," *IEEE Wireless Commun.*, vol. 9, pp. 26–32, Apr. 2002.

[19] A. Stamoulis and G. Giannakis, "Packet fair queueing scheduling based on multirate multipath-transparent CDMA for wireless networks," *Proc. IEEE INFOCOM 2000*, vol. 3, pp. 1067–1076, 2000.

[20] H. Holma and A. Toskala, *WCDMA For UMTS: Radio Access For Third Generation Mobile Communications*. New York: Wiley, 2000.

[21] S. Ariyavisitakul, "Signal and interference statistics of a CDMA system with feedback power control—Part II," *IEEE Trans. Commun.*, vol. 42, pp. 597–605, Feb./Mar./Apr. 1994.
[22] Z. Liu, M. J. Karol, M. E. Zarki, and K. Y. Eng, "Channel access and interference issues in multi-code DS-CDMA wireless packet(ATM) networks," *Wireless Networks*, vol. 2, pp. 173–193, Aug. 1996.
[23] M. Schwartz, *Broadband Integrated Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1996.
[24] J. Zander, "Performance of optimum transmitter power control in cellular radio systems," *IEEE Trans. Veh. Technol.*, vol. 41, pp. 57–62, Feb. 1992.
[25] P. Newson and M. R. Heath, "The capacity of a spread spectrum CDMA system for cellular mobile radio with consideration of system imperfections," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 673–684, May 1994.
[26] A. C.-K. Kam, "Efficient Scheduling Algorithms for Quality-of-Service Guarantees in the Internet," Ph.D. dissertation , Mass. Inst. Technol., Cambridge, 2000.

Liang Xu (S'99–M'03) received the B.Eng. degree from Tsinghua University, Beijing, China, in 1994 and the M.Eng. degree from Southeast University, China, in 1997, both in electrical engineering. He is currently working toward the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada.

His research interests include radio resource management, access control, and QoS provision for wireless networks.

Xuemin (Sherman) Shen (M'97–SM'02) received the B.Sc. degree from Dalian Marine University, Dalian, China, in 1982, and the M.Sc. and Ph.D. degrees from Rutgers University, Piscataway, NJ, in 1987 and 1990, respectively, all in electrical engineering.

From September 1990 to September 1993, he was first with Howard University, Washington DC, and then the University of Alberta, Edmonton, AB, Canada. Since October 1993, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, where he is a Full Professor. His research focuses on mobility and resource management in interconnected wireless/wireline networks. In specific, his interests are traffic flow control, connection admission and access control, handoff, user location estimation, end-to-end performance modeling and evaluation, voice over mobile IP, stochastic process, and H_∞ filtering. He is a coauthor of two books, and has many publications in communications networks, control and filtering.

Dr. Shen is a registered Professional Engineer of the Province of Ontario, Canada.

Jon W. Mark (S'60–SM'80–F'88) received the B.A.Sc. degree from the University of Toronto, Toronto, ON, Canada, in 1962, and the M.Eng. and Ph.D. degrees from McMaster University, Hamilton, ON, in 1968 and 1970, respectively, both in electrical engineering.

From 1962 to 1970, he was with Canadian Westinghouse Company, Ltd., Hamilton, where he was an Engineer and then a Senior Engineer. Since 1970, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, where he is currently a Professor. He was the Department Chairman from July 1984 to June 1990. In 1996, he established the Centre for Wireless Communications, University of Waterloo, where he is currently serving as its Founding Director. He had previously worked in the areas of adaptive equalization, image coding, and spread-spectrum communications. His current research interests are in broadband communications, wireless communications, and wireless/wireline interworking.

Dr. Mark is a former Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS. He is currently a Member of the Inter-Society Steering Committee of the *IEEE/ACM Transactions on Networking*, an Editor of *Wireless Networks*, and an Associate Editor of *Telecommunication Systems*.