QoS-Oriented Packet Scheduling for Wireless Multimedia CDMA Communications

Vincent Huang and Weihua Zhuang, *Senior Member*, *IEEE*

Abstract—In the third-generation (and beyond) wireless communication systems, there will be a mixture of different traffic classes, each having its own transmission rate characteristics and quality-of-service (QoS) requirements. In this paper, a QoS-oriented medium access control (MAC) protocol with fair packet loss sharing (FPLS) scheduling is proposed for wireless code-division multiple access (CDMA) communications. The QoS parameters under consideration are the transmission bit error rate (BER), packet loss, and delay requirements. The MAC protocol exploits both time-division and code-division statistical multiplexing. The BER requirements are guaranteed by properly arranging simultaneous packet transmissions and controlling their transmit power levels, whereas the packet loss and delay requirements are guaranteed by proper packet scheduling. The basic idea of FPLS is to schedule the transmission of multimedia packets in such a way that all the users have a fair share of packet loss according to their QoS requirements, which maximizes the number of the served users under the QoS constraints. Simulation results demonstrate effectiveness of the FPLS scheduler, in comparison with other previously proposed scheduling algorithms.

Index Terms—Code-division multiple access (CDMA), medium access control (MAC), multimedia traffic, packet scheduling, quality-of-service (QoS) provisioning.

1 INTRODUCTION

The third-generation (and beyond) wireless communication systems are anticipated to provide a broad range of multimedia services including voice, data, and video to mobile users. Unlike wireline communication systems, wireless systems have very scarce bandwidth of available frequency spectrum. The limited network resources have to be used efficiently to provide satisfactory services to the users. Wideband code division multiple access (CDMA) has been selected as the major multiple access technique for the third-generation wireless systems. For a CDMA system, power control has to be used to combat the near-far problem. In supporting voice only services, the received signal power from each and every mobile user in the cell is maintained at the same constant level in the uplink at the base station [1]. However, in a multimedia CDMA system, different traffic classes require different transmission accuracies specified by bit error rate (BER). If the received power is kept at the same level for all the traffic classes, the capacity is limited by the most stringent BER requirement and cannot be used efficiently. Recently, several approaches have been proposed for optimal power control for multimedia traffic to maximize the capacity or to minimize the total transmit power [2], [3], [4].

Packetized transmission over wireless links makes it possible to achieve a high statistical multiplexing gain. Packet flows generated by mobile users can be classified to several traffic classes. Each of these classes has its unique traffic characteristics and quality-of-service (QoS) requirements. Due to the heterogeneous and bursty nature of multimedia traffic flows, the traditional voice-based medium access control (MAC) protocols do not perform well in a multimedia environment. A flexible MAC protocol which can efficiently accommodate multimedia traffic is required. One important MAC issue is the packet scheduling. The order of packet transmissions for multimedia traffic has a great impact on the efficiency and performance of the wireless system. However, the design of a packet scheduler involves balancing a number of conflicting requirements. Common criterion for packet scheduler design include maximization of throughput, QoS provisioning, scheduling according to a predefined priority structure, and low implementation complexity as packet scheduling is implemented in real-time. Most packet scheduling strategies, such as first-in-first-out (FIFO), round robin, and generalized processor sharing (GPS) [5], [6] have been originally proposed for wireline networks. Random access protocols have been widely used in the past for wireless networks. Packet reservation multiple access (PRMA) is a well-known time-division multiple access (TDMA) based protocol proposed for voice and data traffic. For CDMA-based systems, some methods have been proposed without considering time-division component [7]. Hybrid time-division/code-division multiple-access schemes have been proposed in [8], [9]. In [8], only traffic rate and delay constraints are considered in resource allocation, while packets with different BER requirements are not differentiated. In [9], a wireless multimedia access control protocol with BER scheduling (called WISPER) is proposed for a single-cell system, where packets with the same or similar BER requirements are transmitted in the same time-slot with the same received power level for all the packets.

In this paper, a MAC protocol with fair packet loss sharing (FPLS) is proposed for CDMA multimedia communications. The MAC protocol exploits both time-division and code-division multiplexing for efficient resource utilization. FPLS is a QoS requirement-based packet

- *The authors are with the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. E-mail: {huang, wzhuang}@bbcr.uwaterloo.ca.*

Manuscript received 21 Apr. 2003; revised 1 Aug. 2003; accepted 22 Sept. 2003.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number 0055-0403.

TABLE 1
Summary of Important Mathematical Notations

symbol	definition
C	total number of packets that can be scheduled in a frame
$\bar{C}$	average amount of total required resources in each frame
$\bar{c}_i$	average amount of required resources for user i in each frame
E_b/I_0	received signal bit energy to interference-plus-noise density ratio
e_i	effective bandwidth of user i
I	total number of active users in the system
L_d	total number of downlink time-slots
L_u	total number of uplink time-slots
$N_{\max}$	maximum number of code-slots in a time-slot
$P_L^{(i)}$	packet loss probability (PLP) upper bound required by user i
$\hat{P}_L^{(i)}$	expected PLP for user i according to the FPLS algorithm
$\hat{P}_{M\|R}^{(i)}(m)$	expected PLP of the most urgent packets (MUPs) for user i given the total MUP load $R = m$
$P_{\min}$	received power for packets with least stringent BER requirement
p_i	code-slots required by packet i
p^l	total code-slot size of scheduled packets in time-slot l
p_{out}	required threshold for outage probability
R	the number of total MUPs in a frame
$R_{\max}$	the maximum value of R
$\bar{r}_i$	average packet generation rate of user i
$\bar{r}_{M_i\|R}(m)$	conditional average MUP rate of user i given $R = m$
$(r_{M_i\|R})_{\max}$	the maximum value of $\bar{r}_{M_i\|R}(m)$
$r_{i\max}$	maximum packet generation rate of user i
γ_i	required E_b/I_0 for user i
$\bar{\gamma}_i$	actual received E_b/I_0 for user i
κ_i	difference between the calculated number of packet that should be scheduled based on FPLS and the actual scheduled packet number for user i
$\lambda_{M_i\|R}(m)$	weighting factor for dropped MUPs of user i
μ_i	average transmission rate of user i
$\xi_{M_i\|R}(m)$	ratio of $\bar{r}_{M_i\|R}(m)$ to $\bar{r}_i$
σ_i	standard deviation of the transmission rate of user i
σ_p	standard deviation of the error in received E_b/I_0 caused by imperfect power control

scheduling strategy. The objectives of the scheduling are to provide QoS guarantee in terms of BER, packet delay, and packet loss and to maximize the system resource utilization. The fact that the capacity of CDMA systems is interference limited and, therefore, QoS dependent poses a significant challenge in the QoS-based packet scheduling. In a wireless environment, a packet is expected to be delivered to the destination with a certain accuracy within a required time frame. Any violation of these two requirements will cause the packet to be useless and will be discarded. As QoS satisfaction and high resource utilization are, in general, conflicting goals, high utilization of the limited wireless bandwidth often means that the system resources cannot accommodate the resource demands of the admitted users from time to time and some packets have to be dropped occasionally. To support as many satisfied users as possible, a fair sharing of the dropped packets among all the users is essential. The main features of the FPLS principle are that: 1) the packet losses are shared fairly among all the users according to each and every user's QoS requirements and 2) the number of users supported by the system with guaranteed QoS is maximized which, in turn, maximizes the resource utilization. With FPLS, the bandwidth is shared among all users in such a way that, when the QoS requirements are guaranteed for one user, they will be guaranteed for all other users at the same time. No user will be allocated more resources than needed if the amount of resources is not enough for other users. A heuristic bin-packing algorithm is used in the packet scheduling so that the usage of time-slots is optimized, thus maximizing the resource utilization.

Our main contribution in this paper is to introduce the FPLS principle. We focus on a single-cell system to illustrate how to implement the FPLS principle in packet scheduling. The remainder of this paper is organized as follows: In Section 2, we describe the system model, including the MAC protocol and the QoS provisioning. The proposed FPLS principle and its implementation are presented in Section 3. In Section 4, the performance of the FPLS scheduler is studied by computer simulation and is compared with that of the previously proposed GPS and WISPER scheduling schemes. Section 5 gives the conclusions of this research. As there are many variables used in this paper, Table 1 gives a summary of the important symbols.

2 The System Model

The wideband CDMA and hybrid time-division/code-division multiple access (TD/CDMA) have been selected as the radio techniques for the Universal Mobile Telecommunication Systems (UMTS) [10], [11]. The hybrid TD/CDMA will operate in the time division duplex (TDD) mode. In this research, we consider a hybrid TD/CDMA wireless system with packetized transmission [12]. The system operates in the TDD mode to best accommodate asymmetric traffic [13], as in a multimedia system, the

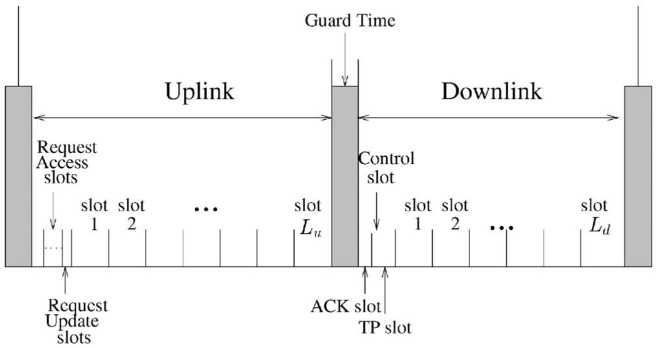


Fig. 1. Time-slots for the uplink and downlink in each frame.

traffic loads for uplink and downlink are highly asymmetric. Time is partitioned into frames of a constant duration. Each frame is divided into time slots. Multiple access within each time-slot is accomplished by assigning unique pseudorandom noise (PN) code sequence(s) to each user. The source information from and to mobile users is segmented into packets of equal length. Although, in the current proposed TDD mode, the uplink spreading factor is ranged from 1 to 16, we use a constant spreading factor for all the packets. In this way, all packets with the same BER requirement will have the same received power. To support different data rate, multicode and multislot assignments are used. The packets are transmitted at a constant bit rate, each packet requiring a time-slot for transmission. Packet transmission from and to mobile users is synchronized in time. The decision on packet transmission in each time-slot for both uplink and downlink is made at the base station and is broadcast to the mobile users as described in the following section.

2.1 MAC Protocol

Fig. 1 shows the TDD multicode TD/CDMA for multirate packet transmission. In the uplink, there are several request access minislots of a constant duration in the beginning of the frame. All packet transmission requests are sent in these slots. When there are active video traffic sources, some of the request access slots are reserved as request update slots. Since the video traffic has a variable packet generation rate, each request update slot is used exclusively by video sources to constantly inform the base station of the number of packets arrived at each of the user terminals during the previous time frame. The request slots are followed by a number of packet transmission slots of a constant duration. The downlink transmission in each frame starts with a control slot. The control slot is a broadcast time-slot which consists of a request acknowledgement (ACK) subslot and a transmission permission (TP) subslot. The ACK subslot is used to acknowledge the received requests from the mobile terminals and the TP subslot is used to broadcast the packet scheduling for the uplink transmission in the next frame.

For the request access slots, a multicode direct sequence (DS)-CDMA with a slotted ALOHA random access protocol is used. Dedicated codes are used for the requests. These codes belong to a common code pool and are not associated with any terminal. When a user is ready to send a request, it randomly chooses a code from the code pool and a request access slot for transmission. More than one user can send their requests in the same request access slot if different codes are used. The user ID is included in the request. If the request is received successfully, the base station will broadcast the user ID in the ACK slot in the current frame. If a terminal does not receive its ID in the ACK slot, it will retransmit the request in the next frame. When the terminal has received its ID in the ACK slot, it will listen to the TP slot for transmission permission. The base station uses the TP slot to inform each user of the allocated time-slot(s) and power level, and how many packets to transmit in each allocated slot. Requests for transmission are sent at the beginning of each frame for packets that have arrived at the terminal buffer in the previous frame. Orthogonal multicodes are used for simultaneous transmission of multiple packets from and to one user [12]. During the call admission control phase, the base station obtains necessary information of each mobile traffic source, including the traffic class, the initial number of packets at the terminal, the terminal buffer size, and the QoS requirements.

2.2 QoS Provisioning

The QoS parameters under consideration are transmission accuracy and delay requirements. The overall transmission accuracy requirement is represented by packet loss and transmission BER (over the wireless link) requirements. In real-time communications, a user requires its packets to be delivered to the destination within a certain time period; otherwise, the packets will be worthless and the service quality will be poor. Here, we consider the delay over the wireless link only. The delay requirement can be represented by the *life span* of each packet. The life span is a set of frames from the moment that the packet is generated to the moment that the required delay bound is reached. The *residual delay bound* is the difference between the required delay bound and the total accumulated queueing delay up to the time of packet scheduling and is also referred to as *time-out value* of the packet. Packet loss rate due to buffer overflow and exceeding the delay bound is referred to as packet loss probability (PLP).

In a wireless environment, the transmission error is caused by transmission through the fading dispersive medium. The required BER can be mapped one-to-one to the required received signal bit energy to interference-plus-noise density ratio, E_b/I_0. The mapping is a function of the modulation and channel coding scheme used, diversity, Rake receiver structure, etc. [14]. In a wideband CDMA system with a relatively large number of active mobile users in each cell, the multiple access interference plus background noise is much larger than the received signal power of any packet. In this case, the received signal power level required for BER satisfaction of each packet is approximately proportional to the required E_b/I_0 value. Consider a single-cell system where there is no intercell interference. Given the background noise power spectral density, the maximum number of packets to be transmitted in a time-slot from sources of the same traffic type can be determined in order to achieve the required BER value. The maximum number decreases with a more stringent BER requirement. Let $N_{\max}$ denote the maximum number of packets requiring the least stringent BER, and $P_{\min}$ denote the corresponding required received signal power level for each packet. All other required power levels for a more stringent BER can be represented in terms of the minimum power level. In general, the required received power level for the ith packet can be represented as $p_i P_{\min}$, where p_i (≥ 1) is a constant. If the propagation path gain between the mobile and the base station is known, the received signal power level can be translated into the transmit power level at the mobile. Mobile users of different service classes have different BER requirements and therefore require different p_i values. To transmit packets with different BER requirements in the same time slot, the number of simultaneously transmitted packets and the power level of each packet should be determined properly so that the BER requirements of all the packets are met. For example, if a packet can tolerate $(N_{\max} - 1)$ other simultaneously transmitted packets with power $P_{\min}$, it can tolerate $(N_{\max} - 1)/p_i$ other simultaneously transmitted packets with power $p_i P_{\min}$. Transmission of a packet with $P_{\min}$ is referred to as one code slot and transmission of a packet with $p_i P_{\min}$ requires p_i code slots. The summation of the code slots from all the packets transmitted in each time-slot cannot be larger than $N_{\max}$ for satisfactory transmission accuracy of all the packets in the time slot. Note that the concept of code slot is an approach to achieve BER satisfaction under the assumption that the interference level in a time slot is constant for all packets transmitted in the slot. If a packet requires a large number of code slots, then the BER for the packet is lower than the required value as the packet does not introduce interference to itself. Furthermore, with the orthogonal multicodes for parallel transmission from and to each user [12], a properly designed Rake receiver can keep the interference among the parallel transmitted signals (due to the propagation delay dispersion introduced by the wireless channel) at a low level. Hence, if a user sends more than one packet in the same time-slot, the actual BER for the user is lower than the required value. The BER for other simultaneously transmitting users remains unchanged. Therefore, it is possible to increase capacity by exploiting the orthogonality property in the multicode CDMA signaling, at the cost of the scheduling flexibility and complexity, which needs further investigation.

In summary, the BER requirements are to be guaranteed by properly arranging simultaneous packet transmissions and controlling their transmit power levels; the delay and PLP requirements are to be guaranteed by proper packet scheduling. In the following, we focus on the uplink transmission. The downlink packet scheduling can be done in a similar way at the base station.

3 The FPLS Scheduler

The design of a packet scheduler depends on many factors such as available resources, instantaneous traffic load, traffic characteristics, and QoS requirements. All these factors have to be weighted and balanced to achieve the fair sharing of the available resources by all the users in service. The main objectives of the packet scheduler proposed here are to guarantee the QoS requirements of all users and to maximize the resource utilization so that the system can support as many satisfied users as possible. There is a trade off between the two objectives. High QoS requirements will cause low resource utilization. The scheduler will provide each user with just enough resources to satisfy its QoS requirements without over allocating resources.

To achieve the objectives, the scheduler first decides the order of the transmissions, i.e., the priorities for the users to transmit their packets. In a high traffic load condition when the number of packets waiting for transmission is in the neighborhood of the system capacity, the priority for transmission should be a function of the delay requirements, packet loss requirements, and traffic load characteristics. Once the priorities are determined, a packet from the user with the highest priority will be scheduled for transmission. The scheduler then determines in which time-slot the packet will be transmitted so that the total number of the code slots for the packets in each time-slot is maximized. In the case of a heavy traffic load, packets should be scheduled to fully utilize each time slot via multiplexing in the code domain.

3.1 Packet Loss Calculation

When the instantaneous total traffic load exceeds the amount that the system can accommodate, some packets have to be dropped. In this section, we calculate the number of the packets to be dropped in the current frame for each user, given a fixed total capacity. To focus on the PLP requirement, we first assume that all the packets have the same BER requirements and, therefore, the system capacity (total resources), denoted by C, represents the maximum number of packets that can be transmitted in a frame. With heterogeneous traffic, the users have different BER requirements. The effect of different BER requirements (represented in terms of the different number of code slots) will then be considered in the next section. As described in Section 2, by the end of the current frame, the packet scheduler schedules the packet transmission for the next frame. The packets with the time-out value equal to one are referred to as *most urgent packets* (MUPs). The MUPs must be scheduled for transmission in the next frame; otherwise, they will be useless and will be dropped. Under the assumption that each user has a large enough terminal buffer size (e.g., equal to the product of the maximum rate and the delay requirement), the packet loss happens only due to scheduling when the total number of MUPs exceeds the system capacity. To guarantee the PLP requirements, we need to control the dropped MUPs for each and every user. Even though the overload of MUPs is caused by some bursty traffic sources during their bursty periods, it is fair to distribute the packet loss among all the users who can tolerate some degree of packet loss, according to their PLP requirements. The packets from a user with a shorter delay requirement become MUPs sooner and, therefore, should be scheduled for transmission sooner. The MUPs have the highest priority and are scheduled first. Only if not all the time-slots in the frame are fully utilized after all MUPs are scheduled, will the scheduler consider non-MUPs in the order of sequentially increased time-out values, starting with those of the time-out value that equals two.

Let I denote the number of active users in the system. The number of MUPs from each user depends on its traffic rate and delay requirement. In order to determine the number of the MUPs to be dropped for each user based on the PLP requirements of all the users, we first need to establish a relation between the overall PLP requirements (with respect to all the packets including both MUPs and non-MUPs) and the packet loss probabilities with respect only to the MUPs. Let the integer random variable R denote the rate of MUPs in packets/frame from all the users. Let $P_L^{(i)}$ (> 0) denote the PLP upper bound required by user i, $1 \leq i \leq I$. Given the MUP traffic load in a frame, $R = m$, the conditional MUP packet loss probability for user i is denoted by $\hat{P}_{M|R}^{(i)}(m)$ whose value depends on the scheduling mechanism. Thus, the actual PLP for user i, $\hat{P}_L^{(i)}$, is the average number of lost MUPs divided by the average number of generated packets in each frame and is given by

$$\hat{P}_L^{(i)} = \frac{\sum_{m=C+1}^{R_{\max}} \bar{r}_{M_i|R}(m)\hat{P}_{M|R}^{(i)}(m)P(R=m)}{\bar{r}_i}, \qquad (1)$$

where the capacity C in packets/frame is assumed to be an integer, $R_{\max}$ is the maximum value of the rate R, $\bar{r}_i$ is the average traffic generation rate in packets/frame of user i, $\bar{r}_{M_i|R}(m)$ is the conditional average transmission rate of the MUPs from user i given $R = m$, and $P(R = m)$ is the probability distribution function of the traffic load R. Letting $\xi_{M_i|R}(m) = \frac{\bar{r}_{M_i|R}(m)}{\bar{r}_i}$, $\hat{P}_{M|R}^{(i)}(m)$ is chosen in such a way that the following relation is satisfied

$$\frac{\xi_{M_i|R}(m)}{\xi_{M_j|R}(m)} \frac{\hat{P}_{M|R}^{(i)}(m)}{\hat{P}_{M|R}^{(j)}(m)} = \frac{P_L^{(i)}}{P_L^{(j)}}, \qquad \forall i, \forall j \in \{1, \cdots, I\}. \qquad (2)$$

From (1) and (2), it can be observed that, when the PLP requirement is satisfied for user i, i.e., $\hat{P}_L^{(i)} \leq P_L^{(i)}$, it will be satisfied for all other users, i.e., $\hat{P}_L^{(j)} \leq P_L^{(j)}$, $1 \leq j \leq I$. This means that, if the system can guarantee the PLP requirement for one user, it can guarantee the PLP requirements for all the users. The FPLS guarantees the fair packet losses among all the users. The term *fair* is used here in the sense that the packet losses are arranged according to the PLP requirements of all the users.

In the single-cell system, the system capacity (amount of the total system resources), C, is described in terms of the number of packets being transmitted in each frame, subject to the BER constraint. That is, the system capacity is the product of the number of time slots in each frame and the number of code slots in each time slot. The resource utilization efficiency is defined as the expected ratio of the resources used for transmitting the scheduled packets in each frame to the capacity C, under the QoS constraints. To demonstrate that the scheduling based on (1) and (2) maximizes the resource utilization efficiency, we argue alternatively that a minimum amount of resources is used to guarantee the PLP requirements of the scheduled packets. Let $\bar{c}_i$ denote the average amount of resources required by user i in each time frame to guarantee its PLP requirement ($\hat{P}_L^{(i)} \leq P_L^{(i)}$), while using the minimum received power level p_i for the BER satisfaction. Since we consider that all the packets have the same BER requirement, the average amount of resources can be equivalently expressed by the average number of transmitted packets in each frame. $\bar{c}_i$ achieves the minimum value when $\hat{P}_L^{(i)} = P_L^{(i)}$. The average amount of resources required for all the scheduled packets in each frame is

$$\bar{C} = \sum_{i=1}^{I} \bar{c}_i. \qquad (3)$$

$\bar{C}$ is minimized when all the $\bar{c}_i$ values simultaneously reach their minimums, i.e., when $\hat{P}_L^{(i)} = P_L^{(i)}, \forall i \in \{1, 2, \ldots, I\}$. For a given $\bar{c}_i$ value, if $\hat{P}_L^{(i)}$ is less than $P_L^{(i)}$, more packets than what are required for user i have been transmitted. That is, $\bar{c}_i$ is larger than the minimum value. For users with different BER requirements, the number of transmitted packets can be weighted by its required code slots. For example, one packet requiring two code slots to transmit is equivalent to two packets requiring one code-slot. Scheduling based on (1) and (2) satisfies the condition for the maximal resource utilization.

Letting m_i denote the MUP load from user i given $R = m$, we have $\sum_{i=1}^{I} m_i = m$. With $\hat{P}_{M|R}^{(i)}(m)$, the number of dropped packets from user i is $m_i\hat{P}_{M|R}^{(i)}(m)$. The total number of the lost packets is equal to the sum of the lost packets from all the users and is equal to $m - C$, where $m > C$, i.e.,

$$\sum_{i=1}^{I} m_i \hat{P}_{M|R}^{(i)}(m) = m - C, \qquad m > C. \qquad (4)$$

From (2)-(4), we can obtain $\hat{P}_{M|R}^{(i)}(m)$ by

$$\hat{P}_{M|R}^{(i)}(m) = \frac{\frac{1}{\xi_{M_i|R}(m)} P_L^{(i)}}{\sum_{j=1}^{I} \frac{1}{\xi_{M_j|R}(m)} m_j P_L^{(j)}} (m - C), \qquad m > C. \qquad (5)$$

From (5), the number of dropped packets for user i is proportional to a weighting factor $\lambda_{M_i|R}(m)$, which is given by

$$\lambda_{M_i|R}(m) = \frac{1}{\xi_{M_i|R}(m)} m_i P_L^{(i)} = \frac{m_i}{\bar{r}_{M_i|R}(m)} \bar{r}_i P_L^{(i)}. \qquad (6)$$

The term $\bar{r}_i P_L^{(i)}$ represents the average number of dropped packets per frame for user i based on the PLP requirement. Thus, $\lambda_{M_i|R}(m)$ is a rate *(m)*-dependent weight factor of packet dropping for user i.

Given $R = m$, from the above analysis, we can decide the number of dropped packets for each and every user in the frame for a given C value, without considering the impact of different BER requirements on the resource allocation. Taking into account that packets from different sources may require different numbers of code slots, the actual value of the capacity C changes with the BER requirements of the scheduled packets due to the fact that the capacity of CDMA systems is interference limited. As a result, the capacity C in terms of the maximum number of packets that can be transmitted in the time frame is unknown before the scheduling. To address the issue, in the following section, we present a procedure to schedule packets for guaranteeing both BER and PLP requirements.

3.2 Packet Scheduling Algorithm

To consider both BER and PLP requirements, we propose a bin-packing scheduling algorithm. The original bin packing problem is a well-known combinatorial problem which

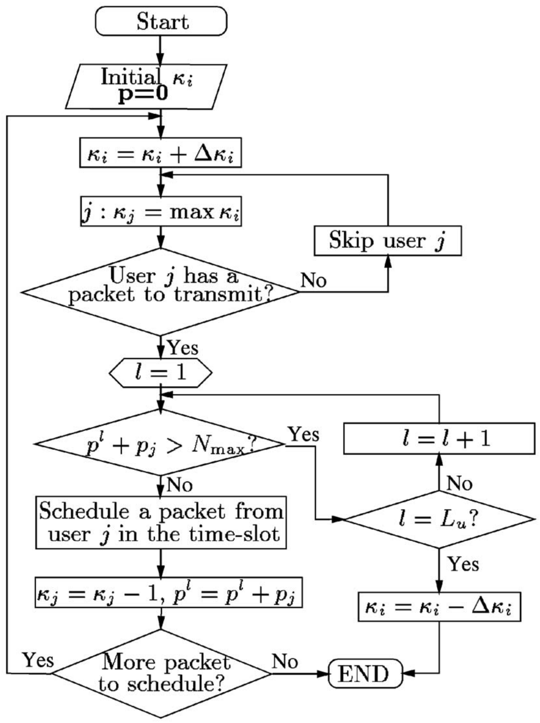


Fig. 2. The proposed FPLS bin-packing packet scheduling algorithm for each frame.

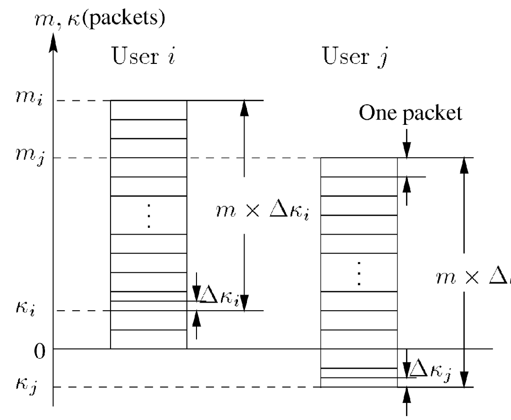


Fig. 3. Illustration of κ and $\Delta\kappa$ for user i with $\kappa_i > 0$ and user j with $\kappa_j < 0$.

deals with the way of packing a set of indivisible blocks into the minimum number of bins. It is known to be NP-complete [15]. In the packet scheduling for each time frame, we consider the time-slots as bins and the packets as blocks. The size of each bin is $N_{\max}$ code slots and the size of each block is the number of code slots required for the packet. The number of bins is fixed. We want to pack as many blocks as possible in the bins without splitting and without exceeding the size of each bin. Fig. 2 illustrates a heuristic algorithm proposed for this problem, where i is the user index, $1 \le i \le I$; l is the time-slot index in each frame and L_u the total number of time-slots in each frame for the uplink transmission; p_j is the number of code slots for each packet from user j; p^l is the total size of the scheduled packets in time-slot l, and $\mathbf{p} = (p^1, p^2, \ldots, p^{L_u})$ which has an initial value $\mathbf{0} = (0, 0, \ldots, 0)$. The packets are scheduled according to their urgency. If the buffer sizes are large enough for all terminals, the urgency depends only on the packet time-out values. The MUPs have the highest priority, and we will schedule them according to (5) given the total number of MUPs, $R = m$. First, we define a priority index κ_i for user i, which determines the order of transmission for the user among all the users. The value of κ_i is calculated according to the PLP requirements. From (5), when the system capacity C is decreased by one packet, the share of packet loss for user i is given by

$$\Delta\kappa_i = \frac{m_i \frac{1}{\xi_{M_i|R}(m)} P_L^{(i)}}{\sum_{j=1}^{I} \frac{1}{\xi_{M_j|R}(m)} m_j P_L^{(j)}}. \tag{7}$$

Correspondingly, if we can increase the system capacity by one packet, the transmitted packet number for user i will also be increased by $\Delta\kappa_i$. As the total system capacity C is unknown before packet scheduling, we schedule the packets by first assuming $C = 0$ and then increasing the value of C one by one until we cannot schedule any more packets. Under the assumption that all the m MUPs are to be dropped (i.e., with the initial value for C being 0), the initial value of κ_i for the first time frame is given by

$$\kappa_i = m_i \left(1 - \frac{m \frac{1}{\xi_{M_i|R}(m)} P_L^{(i)}}{\sum_{j=1}^{I} \frac{1}{\xi_{M_j|R}(m)} m_j P_L^{(j)}} \right), \tag{8}$$

where $\sum_{i=1}^{I} \kappa_i = 0$. When the value of C is increased by 1, the value of κ_i is increased by $\Delta\kappa_i$, for $i = 1, 2, \ldots, I$. If the ith user has the maximum κ value, i.e.,

$$i = \{ i | \kappa_i = \max_{1 \le j \le I} \kappa_j \}, \tag{9}$$

then the capacity increase of one packet is used to schedule an MUP from the user. After that, κ_i is decreased by 1. As a result, the value of κ_i indicates the difference between the calculated number of packets for transmission according to FPLS and the actual number of scheduled packets for user i. User i has been overscheduled if $\kappa_i < 0$ and underscheduled if $\kappa_i > 0$, according to the FPLS principle. The relation among m_i, κ_i, and m is illustrated in Fig. 3 for two users. The total number of MUPs in the current frame is m. When m packets are to be dropped, the share of user i is $m \times \Delta\kappa_i$, which may not be exactly the same as its number of MUPs (m_i). The difference between m_i and $m \times \Delta\kappa_i$ is κ_i. Here, we have user i with $\kappa_i > 0$ and user j with $\kappa_j < 0$. Note that $\sum_{i=1}^{I} \Delta\kappa_i = 1$, where $\Delta\kappa_i \in [0, 1]$, and $\sum_{i=1}^{I} \kappa_i = 1$ before scheduling the packet and $\sum_{i=1}^{I} \kappa_i = 0$ after that. The priority index κ_i is not used to determine the priority for each packet from user i, but rather the priority for user i to transmit its MUPs. When the user with highest κ value does not have any more MUP to transmit, MUPs from the user with next highest κ value will be scheduled. If there are resources available after all the MUPs are scheduled, the same scheduling algorithm will be used to schedule the

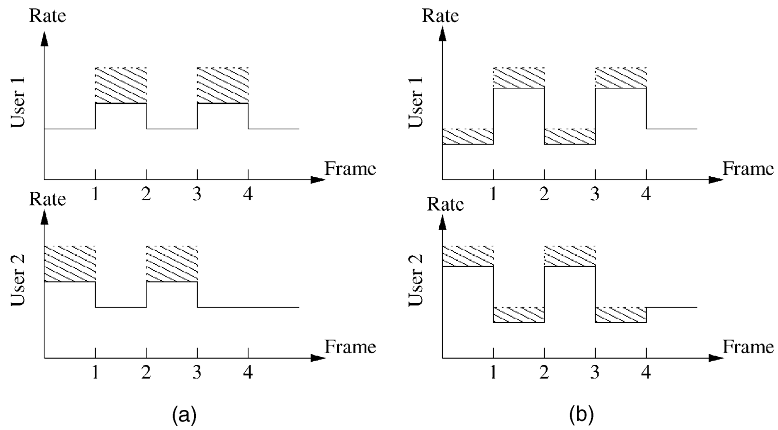


Fig. 4. Comparison of resource allocation using (a) GPS and (b) FPLS. Solid lines represent the allocated rates and dotted lines represent the input traffic rates.

non-MUPs of the time-out value equal to two, etc. In order to avoid overscheduling or underscheduling for each user in the long run, the information whether each user is overscheduled or underscheduled in the current frame should be carried over for scheduling in the next frame. As a result, for the subsequent time frames, the initial κ_i value is the summation of two components, one given by (8) and the other being the κ_i value from the previous frame.

4 Performance Evaluation

4.1 The WISPER and GPS

To evaluate the performance of the proposed packet scheduler, it is desired to compare it with other packet scheduling schemes. Unfortunately, there are very few previously proposed MAC protocols operating in a hybrid TD/CDMA scenario. Here, we consider the comparison with 1) the WISPER [9] which is a MAC protocol with packet scheduling for multimedia traffic in a single-cell system similar to the one considered here and 2) the discretized GPS [5], [6] which is a well-known work conserving protocol for wire-line bandwidth allocation. The three packet scheduling schemes are compared via computer simulation using the same single-cell system model and guaranteeing the same QoS requirements for the same traffic flows.

The three schemes (i.e., FPLS, WISPER, and GPS) schedule packets based on different principles. In the WISPER scheduling, a priority is calculated for each user in each frame. The priority is a function of timeout value and the number of packets to transmit. Higher priority is assigned to a user with small timeout value and a large number of packets in its buffer. WISPER does not consider the PLP requirement and, therefore, the scheduling remains the same for different PLP requirements. Transmission order is determined according to the packet time-out values and the number of packets ready for transmission at each mobile terminal. Packets with the same or similar BER requirements are transmitted in the same slots with the same power level.

The original GPS assumes that the server can serve multiple sessions simultaneously and that the traffic is infinitely divisible. A weight is assigned to each user, the total available bandwidth is then distributed to each backlogged user according to the assigned weight. However, in the hybrid TD/CDMA system, the slots are defined in both time and code domains. Hence, the discretized GPS is considered for the large common pool of available resources. Although many methods have been reported recently for choosing the GPS bandwidth allocation weight factor, a fair and efficient algorithm has yet to be developed. Here, the GPS weights are chosen to be proportional to the effective bandwidths of the traffic flows [16]. The effective bandwidth is the bandwidth required for a given performance objective. For real-time traffic without buffering, the effective bandwidth for user i is given by [17]

$$e_i = \bar{r}_i + \sigma_i \sqrt{-\ln 2\pi - 2\ln P_L^{(i)}}, \tag{10}$$

where σ_i is the standard deviation of the transmission rate of user i. However, for non-real-time traffic with delay and PLP requirements, the calculation of effective bandwidth is very complex. For example, the effective bandwidth for lossless multiplexing is given in [16]. In the following comparison for non-real-time traffic, the bandwidth required for homogeneous traffic to guarantee the given QoS requirements is obtained by computer simulation and then this bandwidth is used as the weighting factor in the simulation with heterogeneous traffic. Using the effective bandwidth as the weighting factor takes into account both delay and PLP requirements. However, GPS does not take the current traffic load into consideration.

The main feature of the FPLS scheduler is to even out the packet loss over a large time period for each user and to let all the users share the packet loss depending on their PLP requirements. When a user is in a bursty period and the system capacity is not sufficient enough to accommodate the total traffic requirement, all other users will have a share of packet loss with this user; on the other hand, this user will share the loss when any other users are in their bursty periods. Thus, instead of a large number of lost packets during a bursty period, the packet loss for each user traffic is smoothed over a longer time period. Based on the real-time traffic load information, the FPLS scheduler allocates the resources to the users according to their actual needs and QoS requirements.

The difference between the GPS and FPLS is illustrated in Fig. 4 for two users as an example, where packet loss happens whenever the allocated rate is below the input traffic rate. Using GPS, packet loss happens mainly during bursty periods. There will be little (or no) packet loss for a nonbursty traffic flow even though the user may tolerate packet loss to some degree. With FPLS, a nonbursty traffic flow will experience packet loss in order to give more resources to other traffic flows at their bursty periods, as long as the user's PLP requirement can be guaranteed. In GPS, when the resources allocated to a user are not used, it will be shared among all other users, and the user will not be compensated. In FPLS, when a user has dropped too many packets (i.e., is underscheduled), it will be compensated with more allocated resources in the future time frames in order to give a fair share of the system resources to every user. In this way, the FPLS scheduling is expected to achieve a high resource utilization than GPS. The performace of both GPS and FPLS depends on the number of users admitted to the system. With a fixed number of users and a fixed weight for each user, GPS guarantees a minimum bandwidth to each user. However, in order to acheive a high multiplexing gain among bursty multimedia traffic flows, the minimum bandwidth allocation is not so meaningful. If all the traffic flows have a constant rate, the performance of FPLS is the same as that of GPS. When there exist bursty traffic flows in the system, using FPLS the bandwidth will be allocated fairly to all the users in terms of the QoS provisioning, while using GPS the bandwidth may be underallocated to bursty traffic from time to time and overallocated to nonbursty traffic. Three situations can occur:

1. In a light traffic load condition, if the QoS of all the users can be satisfied in GPS, it can also be satisfied in FPLS.
2. As the traffic load increases, when GPS can guarantee the QoS only for nonbursty traffic but not for bursty traffic, it is possible that FPLS can guarantee the QoS for all the users.
3. As the traffic load further increases, when GPS can still guarantee the QoS for only nonbursty traffic, FPLS cannot guarantee QoS for any user.

If it is required that the QoS requirements of all the users be satisfied, efficient call admission control should be in place to limit the number of users in service before using the FPLS scheduling. In fact, the FPLS scheduling can be considered as an improved GPS scheme. It allocates the resources frame by frame dynamically (based on the traffic load condition) among all the users to satisfy their PLP requirements. For a long message period, it should be possible to find a particular weighting factor assignment (related to the PLP) for GPS to achieve the same performance as FPLS and the FPLS scheduling is one way to find the weighting factors in the resource allocation.

4.2 The Simulation Environment

We consider three traffic types: voice, video, and data. Each time frame is 10 ms in length and is partitioned into eight time-slots, each slot having a duration of 1.25 ms. The voice traffic is simulated by the on-off model. During the on-state, one packet is generated in each frame which is equivalent to a rate of 100 packets per second. On average, an on-state lasts 10 frames and an off-state lasts 15 frames. The video traffic can be modeled by a Markov-modulated Poisson process (MMPP) [18]. Here, to be able to calculate the conditional rate distribution, we use a simplified MMPP model with four states. The video traffic rate varies among four rates $(0, 4, 8, 12)$, in packets per frame, with a probability of $\left(\frac{1}{6}, \frac{1}{3}, \frac{1}{3}, \frac{1}{6}\right)$, respectively. The short message transfer protocol (SMTP) is used to simulate the data traffic. The data burst arrival from each mobile user is a Poisson process, and each data burst size (for SMTP traffic) is modeled by $\log_2$-normal distribution [19]. Both real-time and non-real-time traffic types are considered. For the case of real-time traffic, the time-out value is chosen to be one frame for both voice and video traffics. All the packets are delay intolerable (i.e., MUPs) and must be transmitted immediately or will be dropped. For the case of non-real-time traffic, voice traffic and data traffic are considered. The time-out value is two frames for voice traffic and 20 frames for data traffic. The simulation parameters are summarized in Table 2. Using the FPLS scheduling requires the MUP rate distribution $\bar{r}_{M_i|R}(m)$ given $R = m$. This rate distribution is a function of scheduling, which can be very complex to calculate. For the real-time traffic, Fig. 5 shows the rate distributions of each voice user and video user, respectively, with a total of 200 voice users and 15 video users, based on computer simulation. It is observed that the rate distribution for user i can be approximated by

$$\bar{r}_{M_i|R}(m) \approx \left[\frac{(r_{M_i|R})_{\max}}{R_{\max}}\right] m \tag{11}$$

to reduce the implementation complexity of the scheduling algorithm, where $(r_{M_i|R})_{\max}$ is the maximum value of $\bar{r}_{M_i|R}(m)$ and is equal to one and 12 packets/frame for voice and video users, respectively, and $R_{\max} = 380$ packets/frame. The approximation is verified to be reasonably accurate through simulations for other mixtures of user traffic flows and is used in the following simulations. To

TABLE 2
The Simulation Parameters

Parameter		Value
Number of time-slots per frame (L_u)		8
Maximum number of code-slots per time-slot ($N_{\max}$)		22 or 30
Simulation time in frames		10000
Voice traffic:	Time-out value in frames (D_i)	1 or 2
	Required code-slots (p_i)	1
	Required packet loss probability upper bound ($P_L^{(i)}$)	10^{-2}
	Average talk spurt length in frames	10
	Average silent period in frames	15
Video traffic:	Time-out value in frames (D_i)	1 or 20
	Required code-slots (p_i)	1.9
	Required packet loss probability upper bound ($P_L^{(i)}$)	10^{-2}
	Average rate in packets/frame	6
	Peak rate in packets/frame	12
Data traffic:	Time-out value in frames (D_i)	20
	Required code-slots (p_i)	3.2
	Required packet loss probability upper bound ($P_L^{(i)}$)	10^{-5}
	Geometric mean rate in packets/frame	1
	Geometric standard deviation	$\times 3$
	Peak rate in packets/frame	20

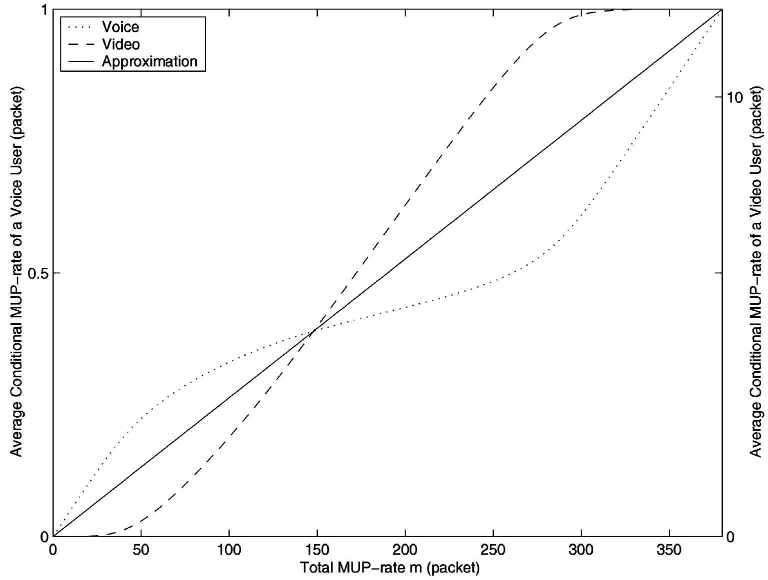


Fig. 5. The conditional MUP rate distribution $\bar{r}_{M_i|R}(m)$ (in packets/frame) for each voice/video user given the total MUP traffic load $R = m$.

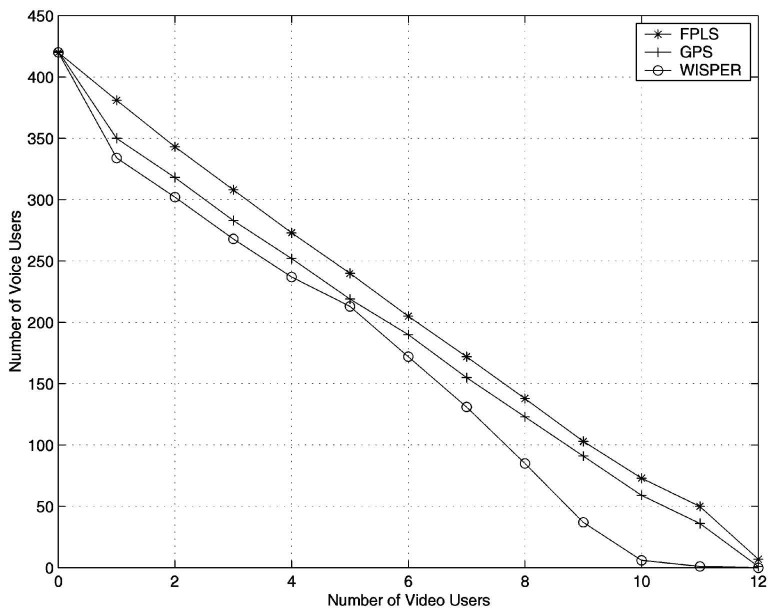


Fig. 6. The numbers of real-time traffic users supported using FPLS versus WISPER and GPS.

intuitively explain the approximation, consider a non-real-time traffic case where all the packets have a required delay bound of two frames. Initially, all the arrival packets are non-MUPs, and the non-MUPs are scheduled according to the FPLS principle. The non-MUPs, which cannot be scheduled to transmit become MUPs in the next frame. Therefore, the arrival rate of MUPs is equal to the loss rate in the current non-MUPs scheduling. In scheduling the non-MUPs, $\hat{P}^{(i)}_{M|R}(m)$ in (2) represents the percentages of arrival packets that become MUPs and $\bar{r}_{M_i|R}(m)$ represents the average arrival rate. The percentages of total arrival packets which become MUPs are approximately proportional to each other for all the traffic sources over a long time period (the proportional coefficients depend on the PLP requirements) since the FPLS can guarantee the PLP requirements of all the users to be satisfied at the same time. When the average conditional MUP rates are proportional to the PLP requirements, the linear approximation given in (11) can be used. Similar approximation can be made for packets with a required delay bound larger than two frames.

Using each scheduling method, the maximum numbers of users that can be supported by the system under the QoS constraints are found by searching through all possible combinations of the user numbers with different traffic types. For simplicity, the resource overhead necessary for signaling and control in MAC is not considered.

4.3 Real-Time Traffic

Fig. 6 shows the maximum numbers of the voice and video users that can be supported by the system with QoS satisfaction. FPLS outperforms GPS and WISPER in most

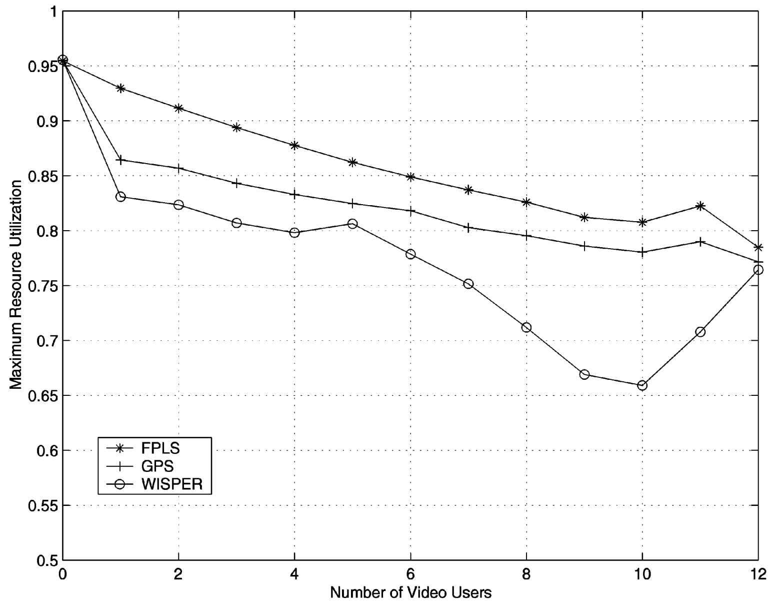


Fig. 7. The percentage of system resources utilized for real-time traffic using FPLS versus WISPER and GPS.

situations. When there is only one traffic type (voice or video) in the system, the total number of lost packets is the same for all three algorithms and, therefore, there is no performance difference among the algorithms. However, if there exists a mixture of the two traffic types in the system with different QoS requirements, FPLS performs better because of the more effective statistical multiplexing. With the real-time traffic flows, all the arrived packets are MUPs to be either transmitted in the next frame or dropped. The average rate of MUPs is the same as the average rate of the arrived packets. The system capacity depends mainly on the proper choice of packet dropping. FPLS balances the dropped packets between the two traffic types at all times according to their PLP requirements and therefore achieves the best performance. For GPS, the weighting factor given by (10) is a constant and, therefore, does not adapt to traffic dynamics. The corresponding resource allocation is not optimal when there is a mixture of different traffic types. WISPER does not take the PLP requirements into consideration. As the order of packet transmission remains the same for different packet loss requirements, the system capacity is limited by the most stringent packet loss requirement. Fig. 7 shows the corresponding resource utilization efficiency with guaranteed QoS requirements for all the users. The FPLS provides a higher resource utilization to accommodate more users than both GPS and WISPER. It is expected that, with an increase of the system capacity (e.g., a larger L_u value), the performance improvement of FPLS over GPS and WISPER will increase because a higher statistical multiplexing gain can be achieved with more voice and video users in service. Note that the discrete nature of the packetized traffic flows and code slots causes some unexpected fluctuations of the curves in Fig. 7.

4.4 Non-Real-Time Traffic

Using the data from Table X in [19], we simulate the SMTP traffic with an average arrival interval of 10 frames and size of each data burst following a $\log_2$-normal distribution with a geometric mean of 10 packets and a geometric standard deviation $\times 3$. The geometric mean and geometric standard deviation are respectively the mean and the standard deviation of the logarithmic transformed quantity of the original random variable [19]. Since the $\log_2$-normal distribution has a large variance, each simulation result is the average value of those from five independent 10000-frame runs. Fig. 8 shows the maximum numbers of the voice and data users which can be supported by the system using FPLS, GPS, and WISPER, respectively, for the same QoS requirements. FPLS can provide service for the largest number of the users. Fig. 9 shows the corresponding resource utilization efficiency. The resource utilization decreases significantly with an increase in the data user number, due to the highly bursty nature of the data traffic. Here, we choose a relative short delay requirement for data traffic to reduce the simulation time, yet we can study the performance of the three algorithms. The FPLS scheduling clearly outperforms both GPS and WISPER. In particular, when the difference between the voice and data traffic loads is large, WISPER gives a higher priority to the traffic with high load, thus causing high packet loss for users with the lower traffic load. As the number of the traffic types and/or the differences in the PLP requirements increase, further performance improvement achieved by FPLS over WISPER is expected. On the other hand, in the GPS scheduling, the allocated resources to each traffic type is proportional to its effective bandwidth and depends on the total traffic load in the frame. Since data users can tolerate a larger delay, effective statistical multiplexing among data users can be achieved in both time and code domains; on the other hand, the small number of data users in service and the highly bursty nature

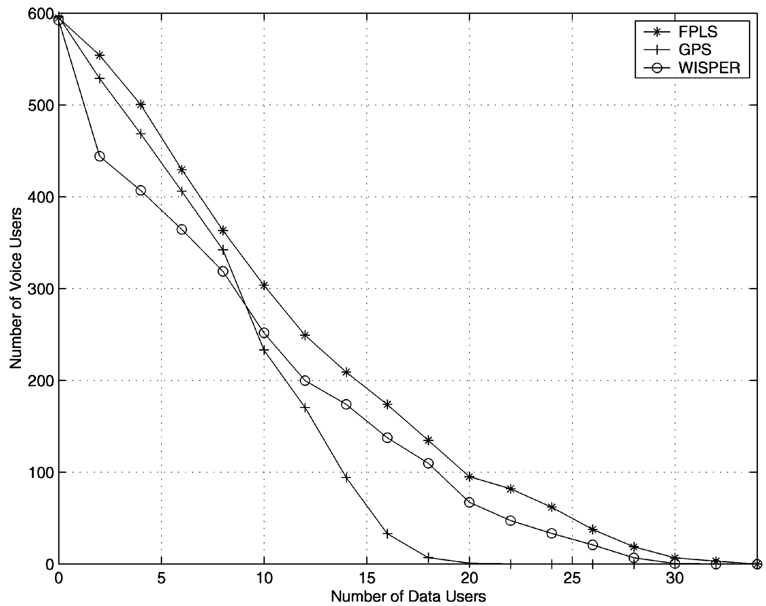


Fig. 8. The numbers of non-real-time traffic users supported using FPLS versus WISPER and GPS.

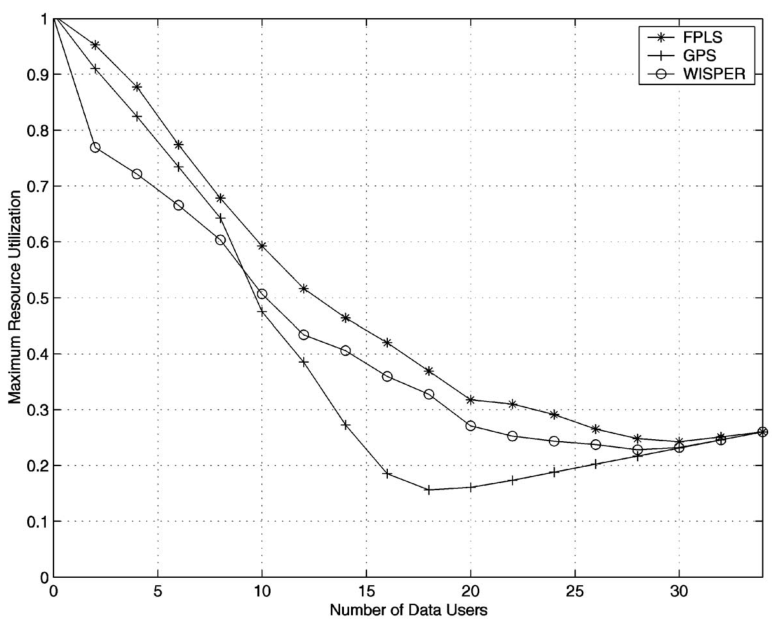


Fig. 9. The percentage of system resources utilized for non-real-time traffic using FPLS versus WISPER and GPS.

of the traffic limit the multiplexing gain. As the voice users have the more stringent delay requirement, the statistical multiplexing is achieved mainly in the code domain. However, when there is a much larger number of voice users in service (due to the relatively low rate from each voice user) than that of data users, a larger multiplexing gain can be achieved. As a result, in Fig. 9, the resource utilization efficiency decreases when the number of data users increases (corresponding to a significant decrease in the number of voice users, as shown in Fig. 8). Using GPS, the resource utilization efficiency decreases when the number of data users deviates from the maximum value. This is because the large number of data users dominate the resource usage and the voice traffic may not get its fair share in the resource allocation from frame to frame due to the work-conserving discipline and discrete nature of the GPS protocol used in the simulation. As a result, the QoS for the data users can be higher than required, which translates to a reduced resource utilization efficiency.

4.5 Imperfect Power Control

We have so far assumed perfect power control in the system, where the desired E_b/I_0 value (i.e., minimum of the required E_b/I_0) is achieved at the base station receiver. In a practical system, the power control is often not accurate. The power control error in decibel is defined as the difference between the actually received power level in decibel and the desired received power level in decibel. It is caused by the error in propagation path loss estimation and

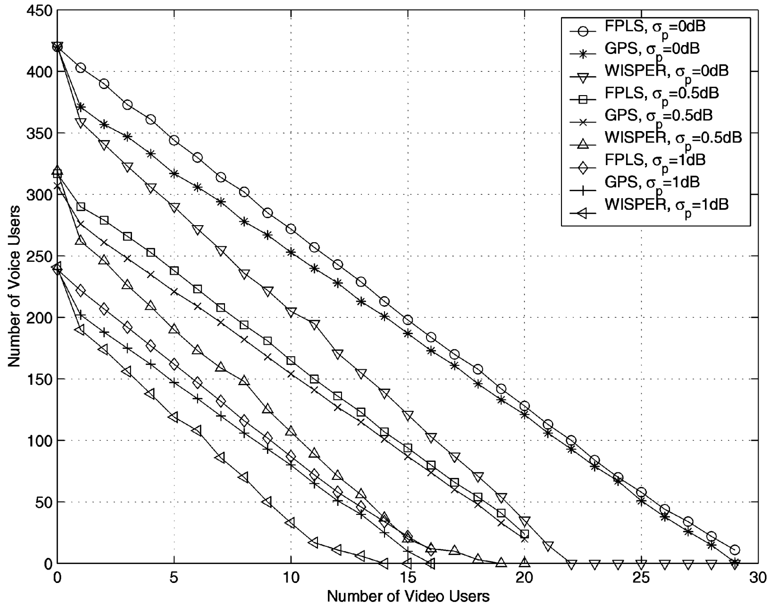


Fig. 10. Comparison of the numbers of users supported by the system with perfect and imperfect power control.

by the delay in power control process, and is usually assumed to be a zero-mean Gaussian random variable [20]. Since the E_b/I_0 in decibel is approximately proportional to the received power in decibel in the system, the E_b/I_0 also approximately follows a Gaussian distribution.

Let $\bar{\gamma}_i$ be the actual received E_b/I_0 for user i, γ_i be the desired E_b/I_0 for user i, and p_{out} be the required upper bound for the outage probability. The outage probability is defined as the probability of $\bar{\gamma}_i$ being less than γ_i. Let σ_p denote the standard deviation of the power control error in decibel. To guarantee the outage probability, we have

$$\mathrm{P}\{\bar{\gamma}_i < \gamma_i\} \leq p_{\text{out}}. \tag{12}$$

Given the values of required E_b/I_0, p_{out}, and σ_p, the target E_b/I_0 can be determined to satisfy (12). Consider real-time voice and video traffic, with the required E_b/I_0 of 7 dB. With the outage probability being one percent, Fig. 10 shows the numbers of supported users with σ_p equals to 0 dB, 0.5 dB, and 1 dB, respectively, where 0 dB means perfect power control. The simulation is carried out in the same way as with perfect power control, but using the target E_b/I_0 value which is larger than the required E_b/I_0 in order to compensate for the power control error. We studied the performance for FPLS, GPS, and WISPER. The system capacity is decreased by approximately 43 percent when σ_p is 1 dB. This result is similar to those observed in [21]. The FPLS scheduling outperforms GPS and WISPER even with imperfect power control.

5 Conclusion

In this paper, we propose the packet scheduler based on the FPLS principle for the MAC protocol in the hybrid TD/CDMA wireless multimedia communications. Statistical multiplexing in both time domain and code domain is exploited. Transmission accuracy requirement over the wireless link is guaranteed by proper transmit power allocation, while packet loss probability and delay requirement are guaranteed by proper packet scheduling. Based on the transmission rate statistics and real-time traffic load information, the FPLS scheduler allocates a minimum amount of resources to each user for QoS provisioning and high resource utilization by letting each user have a fair share in packet loss, assigning a minimum required transmit power level, and achieving maximum multiplexing in the code domain for each time-slot. Simulation results demonstrate that the FPLS scheduler outperforms both WISPER and discrete GPS scheduling schemes in terms of resource utilization. By assigning the priority to each user instead of each individual packet in the scheduling, the computational complexity is greatly reduced. The FPLS scheduler requires the rate statistics to calculate the packet loss rate over a long time period. For the traffic models considered in this paper, a linear approximation for the conditional rate distribution can be used. Since the conditional rate distribution depends only on the number of users and their traffic models using the proposed scheduling algorithm, accurate estimation of the rate distribution can be obtained offline, so as to keep the online scheduling procedure relatively simple.

Using FPLS, all the admitted users are treated equally. In the case of heavy traffic load, the performance of all the users will degrade. Therefore, it is important that the FPLS scheduling is used together with an efficient admission control.

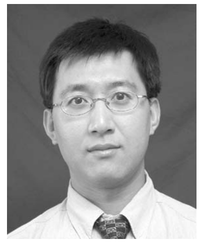
We have so far introduced the FPLS principle and its implementation in the single-cell system, where the concept of code slots can be applied. In a multicell system, the packet scheduling based on the FPLS principle should take into account intercell interference, time-varying propagation path loss, and power allocation, which needs further investigation.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their thorough reviews and helpful suggestions. This work was supported by a research grant (RGPIN155131) from the Natural Science and Engineering Research Council (NSERC) of Canada.

REFERENCES

[1] K.S. Gilhousen, I.M. Jacobs, R. Padovani, A.J. Viterbi, L.A. Weaver Jr., and C.E. Wheatley III, "On the Capacity of a Cellular CDMA System," *IEEE Trans. Vehicular Technology,* vol. 40, pp. 303-312, May 1991.
[2] M. Soleimanipour, W. Zhuang, and G.H. Freeman, "Optimal Resource Management in Wireless Multimedia Wideband CDMA Systems," *IEEE Trans. Mobile Computing,* vol. 1, no. 2, pp. 143-160, Apr.-June 2002.
[3] L.C. Yun and D.G. Messerschmitt, "Variable Quality of Service in CDMA Systems by Statistical Power Control," *Proc. IEEE Int'l Conf. Comm.,* vol. 2, pp. 713-719, 1995.
[4] S. Yao and E. Geraniotis, "Optimal Power Control Law for Multi-Media Multi-Rate CDMA Systems," *Proc. IEEE Vehivular Technology Conf.,* vol. 1, pp. 392-396, 1996.
[5] A.K. Parekh and R.G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case," *IEEE/ACM Trans. Networking,* vol. 1, pp. 344-357, June 1993.
[6] A.K. Parekh and R.G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Multiple Node Case," *IEEE/ACM Trans. Networking,* vol. 2, pp. 137-150, Apr. 1994.
[7] M.J. Karol, Z. Liu, and K.Y. Eng, "Distributed-Queueing Request Update Multiple Access (DQRUMA) for Wireless Packet (ATM) Networks," *Proc. Int'l Conf. Comm.,* vol. 1, pp. 1224-1231, June 1995.
[8] A.E. Brand and A.H. Aghvami, "Multidimensional PRMA with Prioritized Bayesian Broadcast—A MAC Strategy for Multiservice Traffic over UMTS," *IEEE Trans. Vehicular Technology,* vol. 47, pp. 1148-1161, Nov. 1998.
[9] I.F. Akyildiz, D.A. Levine, and I. Joe, "A Slotted CDMA Protocol with BER Scheduling for Wireless Multimedia Networks," *IEEE/ACM Trans. Networking,* vol. 7, pp. 146-158, Apr. 1999.
[10] *Network Architecture.* 3GPP TS 23.002 V5.9.0, Dec. 2002.
[11] J.D. Vriendt, P. Laine, C. Lerouge, and X. Xu, "Mobile Network Evolution: A Revolution on the Move," *IEEE Comm. Magazine,* pp. 104-111, Apr. 2002.
[12] V. Huang and W. Zhuang, "Optimal Resource Mangement in Packet-Switching TDD CDMA Systems," *IEEE Personal Comm.,* pp. 26-31, Dec. 2000.
[13] G.J.R. Povey and M. Nakagawa, "A Review of Time Division Duplex—C DMA Techniques," *Proc. 1998 IEEE Fifth Int'l Symp. Spread Spectrum Techniques and Application,* vol. 2, pp. 630-633, 1998.
[14] S. Manji and W. Zhuang, "Power Control and Capacity Analysis for a Packetized Indoor Multimedia DS-CDMA Network," *IEEE Trans. Vehicular Technology,* vol. 49, pp. 911-935, May 2000.
[15] M.R. Garey and D.S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness.* New York: W.H. Freeman and Company, 1979.
[16] A. Elwalid and D. Mitra, "Design of Generalized Processor Sharing Schedulers which Statistically Multiplex Heterogeneous QoS Classes," *Proc. 18th Ann. Joint Conf. IEEE Computer and Comm. Soc. (INFOCOM '99),* vol. 3, pp. 1220-1230, 1999.
[17] M. Schwartz, *Broadband Integrated Networks.* Prentice-Hall Int'l, Inc., 1996.
[18] P. Skelly, M. Schwartz, and S. Dixit, "A Histogram-Based Model for Video Traffic Behavior in an ATM Multiplexer," *IEEE/ACM Trans. Networking,* vol. 1, pp. 446-459, Aug. 1993.
[19] V. Paxson, "Empirically Derived Analytic Models of Wide-Area TCP Connections," *IEEE/ACM Trans. Networking,* vol. 2, pp. 316-336, Aug. 1994.
[20] A.J. Viterbi, A.M. Viterbi, and E. Zehavi, "Performance of Power-Controlled Wideband Terrestrial Digital Communication," *IEEE Trans. Comm.,* vol. 41, pp. 559-569, Apr. 1993.
[21] W.-M. Tam and F.C.M. Lau, "Analysis of Power Control and Its Imperfections in CDMA Cellular Systems," *IEEE Trans. Vehicular Technology,* vol. 48, pp. 1706-1717, Sept. 1999.



Vincent Huang received the MSc degree in electrical engineering in 1994 from Chalmers University of Technology, Gothenburg, Sweden, and the PhD degree in electrical engineering in 2003 from the Univeristy of Waterloo, Waterloo, Ontario, Canada. During 1994-1999, he was a research assistant in the Department of Signals and Systems at Chalmers University of Technology. His current research interests are MAC protocols and resource management for multimedia wireless communications.

Weihua Zhuang (M'93-SM'01) received the BSc and MSc degrees from Dalian Maritime University, Liaoning, China, in 1982 and 1985, respectively, and the PhD degree from the University of New Brunswick, Fredericton, NB, Canada, in 1993, all in electrical engineering. Since October 1993, she has been with the Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada, where she is a professor. She is a coauthor of the textbook *Wireless Communications and Networking* (Prentice Hall, 2003). Her current research interests include multimedia wireless communications, wireless networks, and radio positioning. Dr. Zhuang is a licensed professional engineer in the Province of Ontario, Canada. She received the Premier's Research Excellence Award (PREA) in 2001 from the Ontario Government for demonstrated excellence of scientific and academic contributions. She is a senior member of the IEEE.