In this article we propose a medium access control protocol for time-division duplex wideband code-division multiple access multimedia wireless communications. The MAC protocol exploits both time-division and code-division statistical multiplexing to efficiently accommodate real-time voice and video traffic and non-real-time bursty data traffic. An optimal packet scheduling algorithm is proposed to allocate system resources to each user for QoS provisioning and high resource utilization. For a practical solution, we propose suboptimal resource allocation and develop a heuristic bin-packing algorithm to solve the packet scheduling problem. Numerical results are presented to demonstrate the performance of the suboptimal packet scheduling algorithm for mixed voice and data traffic.

## Optimal Resource Management in Packet-Switching TDD CDMA Systems

**Vincent Huang and Weihua Zhuang, University of Waterloo**

Future wireless personal communications systems are expected to provide a broad range of multimedia services including voice, data, and video to mobile users with guaranteed quality of service (QoS). Wideband code-division multiple access (CDMA) has been proposed as the multiple access technique for third-generation wireless communications systems, as specified in the International Mobile Telecommunications in 2000 (IMT-2000) proposals. As tetherless communication and computing become more and more ubiquitous, future wideband CDMA systems are required to efficiently utilize the limited wireless spectrum due to the rapidly growing demands for services. User mobility, the hostile wireless propagation environment, and heterogeneous characteristics of multimedia traffic pose significant challenges in resource allocation. Taking into account the bursty nature of data traffic and time-varying demands for resources from video traffic, packetized transmission over wireless links makes it possible to achieve a high statistical multiplexing gain. To simultaneously maximize wireless resource utilization and guarantee QoS satisfaction, packet transmission needs to be scheduled properly through medium access control (MAC). Even though the current radio transmission technology (RTT) proposals for third-generation wireless mobile systems specify the roles of the MAC layer, the development of an efficient MAC protocol for future wideband CDMA systems remains an open research area.

In this article we propose a MAC protocol for a time-division duplex (TDD) wideband CDMA (WCDMA) system. The protocol exploits both time-division and code-division statistical multiplexing to efficiently accommodate real-time voice and video traffic and non-real-time bursty data traffic. A packet scheduling algorithm for the uplink transmission is proposed to guarantee the QoS requirements of each transmitted packet while achieving high resource utilization.

**The CDMA System Model**

We consider a TDD direct-sequence (DS) WCDMA cellular system with packetized transmission. Multiple access is accomplished by assigning unique pseudo noise (PN) code sequence(s) to each user. Time is partitioned into frames of a constant duration. The source information from and to mobile users is segmented into packets of equal length. The packets are transmitted at a constant bit rate; that is, the processing gain of the spread spectrum modulation in the wideband CDMA is a constant. As a result, each information packet requires a constant time duration (referred to as 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 mobile users by a MAC protocol.

The reasons for TDD transmission are to achieve asymmetric resource allocation between the uplink and downlink, and to make use of the reciprocal nature of the channel. For multimedia services, the resource demands in the uplink and downlink are generally not the same and are time-varying. By adjusting the boundary between the uplink and downlink, the uplink and downlink capacities can be traded against each other. Furthermore, the reciprocal nature of the channel makes it possible:

- To achieve mobile terminal simplicity by predistorting the transmitted signal in the downlink so that the channel will behave as a matched filter to the transmitted signal
- To exploit space diversity on the downlink without requiring multiple antennas at mobile terminals
- To implement accurate power control in the packetized transmission

The multimedia services supported in the system include voice, data, and video, each having its unique traffic characteristics and QoS requirements. A voice call has talk spurts and silent periods. During talk spurts, packets are generated at a constant rate; during silent periods, no packet is generated. The durations of talk spurts and silent periods can be modeled as independent random variables with exponential distributions. Different from voice traffic, data traffic is normally bursty. The arrivals of packet bursts from a single data source can be modeled by a Poisson process. The number of packets in each burst usually has an exponential distribution. For video traffic, the packet generation of each source is unique and changes with time. Normally, a video source has a much higher packet generation rate than a voice source. Voice and video are real-time traffic
and thus have strict transmission delay requirements. However, they can tolerate a certain degree of transmission errors. On the other hand, data traffic is non-real-time in nature but requires high transmission accuracy. The transmission delay requirement depends on each particular data application.

The Hybrid Time-Division/Code-Division MAC Protocol

Over the last decade, several wireless MAC protocols have been proposed. Packet reservation multiple access (PRMA) [1] is a time-division multiple access (TDMA)-based protocol proposed for voice and data traffic. Multicode-division (MD) PRMA is proposed for hybrid TDMA/CDMA systems supporting low-bit-rate traffic for applications such as voice and data [2]. Slots are defined in both time and code domains. Efficient statistical multiplexing can be achieved from the large common pool of available resources. PRMA suffers from various channel access delays and low resource utilization in the contention mode. In distributed queuing request update multiple access (DORUMA) proposed for wireless asynchronous transfer mode (ATM) networks [3], the uplink and downlink periods are configured on a slot-by-slot basis. By using a piggyback field, a video user can update the traffic slot demand without extra contention. All packets are assumed to have the same bit error rate (BER) requirement. A packet scheduling algorithm based on BER requirements for CDMA systems is proposed in [4]. The algorithm is limited to only one received power level for each slot.

Here, we consider a hybrid time-/code-division MAC protocol for the TDD wideband CDMA system. The basic idea behind the hybrid technique is to control interference in the CDMA system using TDMA-type multiplexing. Packet flows generated by and transmitted to mobile users are time-division multiplexed on spread spectrum codes. Packet transmissions in code space are scheduled to control interference among the users in each time slot in order to achieve satisfactory transmission accuracy.

Time-Division Statistical Multiplexing

Figure 1 illustrates the time slots for the uplink and downlink in each time frame. The frame duration is chosen such that a voice source generates 1 packet/frame during its talk spurt period. For the uplink transmission, there are several request access (RA) slots in the beginning of the frame for all packet transmission requests. For each active video traffic source, there is one RA slot reserved for it as a request update (RU) slot. RU slots are used exclusively by video traffic. Since each video source has a variable packet generation rate, it has to constantly inform the base station of the number of packets that have arrived at the terminal over the time duration of the previous frame. The request slots are followed by $L_u$ packet transmission (PT) slots. For the downlink transmission, there are two control slots in the beginning of the transmission for request acknowledgment (ACK) and transmission permission (TP), followed by $L_d$ PT slots. An RA slot is much shorter in time than a PT slot. Dedicated PN codes form a common code pool for the uplink requests. The base station broadcasts the available codes of the pool (excluding those used by video sources for the RU slots) to the terminals. When a mobile user is ready to send a request, the terminal randomly chooses a code from the code pool and an RA slot for transmission. More than one mobile user can send their requests in the same RA slot if different codes are used. The request includes information such as user ID and traffic type. QoS requirements associated with each traffic type are stored in the base station. If the request has been received successfully, the base station will broadcast the user ID in the ACK slot in the next downlink frame. If the user receives its ID in the ACK slot, it will listen to the TP slot for transmit permission; otherwise, it will retransmit the request in the next uplink frame. The base station broadcasts the transmission permission in the TP slot. It informs each acknowledged user of the allocated time slot(s) and how many packets to transmit in each allocated time slot.

For voice traffic, each user uses the slotted ALOHA random access protocol for an RA slot only at the beginning of each talk spurt period. Since each user generates one voice packet in each frame during the talk spurt, the base station automatically allocates the resource for the user to transmit one packet in each frame. The allocated time slot may vary from frame to frame depending on packet scheduling for that frame. The user listens to the broadcast in the TP slot in each frame to know the time slot for transmission. When the talk spurt period is over, the resources reserved for the user will be wasted, which informs the base station not to reserve the resource for the user in the following frames. The process repeats for each talk spurt. The protocol is similar to PRMA in terms of the packet reservation, but different in two aspects:

- The contention is limited only on the short RA slots for higher efficiency.
- The reserved time slot for transmission is not fixed, allowing for flexible resource allocation.

For data traffic, each user sends the request in an RA slot when a data burst is generated. The mobile user informs the base station of the number of packets to be transmitted in the request. For video traffic, the request will be sent to the base station in a way integrating those of voice and data users. A video user uses an RA slot to send its very first request to the base station in the same way a data user does. It informs the

![Figure 1. Time slots for the uplink and downlink in each frame.](image)
base station of how many packets have arrived in the buffer. After that, the base station assigns a fixed RU slot and a PN code to the user. Using the RU slot, the terminal sends only the change in the number of packets arrived in the previous frame from that in the frame before. Based on the information in the first RA slot and the RU slots, the base station will allocate resources to the user and broadcast the allocation information in the TP slot. Upon completion of the call, the user sends a termination request in the RU slot to indicate the end of transmission. Thus, unnecessary contention for request slots can be avoided and collisions in the RA slots greatly reduced.

**Code-Division Statistical Multiplexing**

In the wideband CDMA system, over each time slot, statistical multiplexing can be achieved by transmission of multiple packets, with each packet modulated by a unique PN code. For a high-rate traffic source, the user may be assigned more than one time slot in each frame, and over each assigned time slot parallel transmission of \( m > 1 \) packets is possible using the multicode (MC) CDMA technique. Let \( C^p_n \) denote the primary PN code of mobile user \( n \), the spreading codes for \( m \) parallel-transmitted packets, \( C^i_n \) (\( i = 1, \ldots, m \)), are derived from \( C^p_n \) by \( C^i_n = C^p_n \times W_i \), where \( W_i \) are from a set of orthogonal codes (e.g., Walsh codes). Using this scheme, \( C^i_n \) is orthogonal to \( C^j_n \) if \( i \neq j \). This orthogonality is maintained at the receiver since the propagation path is the same for the parallel-transmitted packets. MC CDMA transmission can support different traffic rates with the same chip rate and can avoid self-interference among packets from the same mobile.

To guarantee QoS to mobile users and achieve high wireless spectrum utilization efficiency, the MAC protocol requires a packet scheduling algorithm which determines how to allocate the time slots and the number of packets for each slot to each mobile user based on the resource demands and QoS requirements of all the users, as discussed in the next section.

**Optimal Resource Allocation**

Optimal resource management for wireless multimedia CDMA systems has been an active research area in recent years. For a single-cell system, minimizing the total transmitted power and maximizing the total transmission rate in the cell are treated as two separate optimization problems for the reverse link transmission [5]. In [6], resource management is combined with base station assignment for reverse link transmission in a multicell system. The previous work assumes a continuous-time transmission with the processing gain depending on the allocated time-varying transmission rate. The system model cannot efficiently accommodate bursty data traffic with short bursts. Furthermore, it requires high implementation complexity for the variable processing gain. Here, we investigate packet transmission with MAC to efficiently support bursty data traffic and with a constant processing gain for low implementation cost. For simplicity, we consider the uplink of the single-cell environment in the resource allocation discussed below. The approaches can also be applied to the downlink.

![Figure 2. The dynamic programming problem for packet scheduling.](image-url)
faction, packet transmission should be scheduled to meet the delay requirement and avoid the buffer overflow for each user in the cell.

**Optimization Problem Formulation**

Given the total system resources, the objective of the optimal scheduler is to maximize the system throughput while guaranteeing QoS requirements. Since the system resources used to transmit each packet are proportional to the received signal power level, the throughput is defined as the average total transmitted packets weighted by the associated received power levels in each frame, where the minimum received power level for each packet is used to satisfy the required BER. From the viewpoint of a service provider, if the revenue generated by transmission of one packet is proportional to the minimum resources required subject to QoS satisfaction, then maximizing the throughput corresponds to the maximum profit. Given the constant total available resources, we want to maximize the total code slots of all transmitted packets over all the frames under consideration. Let $K$ denote the total number of time frames to be optimized, and $M_K$ the total number of packets to be transmitted during the $K$ frames. The objective function is

$$\text{Maximize} \sum_{k=1}^{K} \sum_{i=1}^{M_K} \sum_{j=1}^{L_s} p_i S_{i,j}^{k},$$

where $S_{i,j}^{k} = 1$ if packet $i$ assigned to time slot $j$ in frame $k$, and = 0 otherwise. The QoS requirements constitute the constraints. For the delay requirement, packet $i$ should be scheduled to transmit within its life span, denoted by $E_i$, which is determined based on the transmission delay requirement and terminal buffer size, so that the delay requirement can be guaranteed and packet loss due to buffer overflow can be avoided. A dynamic programming problem can be formulated for the resource allocation problem. As shown in Fig. 2, we want to place $M_K$ packets into $KL_u$ slots. Each packet has its own life span and required $p_i$ value represented by the size of the packet. Each slot has a fixed available capacity (i.e., $N_{\text{max}}$). The objective in scheduling the transmission of the packets is to maximize the total resource utilization in all the time slots subject to the QoS constraints.

The dynamic programming problem can be solved by the deterministic or probabilistic approach [8]. However, the deterministic approach suffers from complexity which increases exponentially with the total number of time slots and from a scheduling delay up to $K$ frames. On the other hand, the probabilistic approach requires the statistical information on the packet arrivals for each traffic type, which can be difficult to obtain as it depends on characteristics of each service class and user mobility pattern in a practical system. The difficulties in optimal resource allocation result from the fact that it is necessary to consider resource allocation over a large number of frames together, because of various delay requirements of the packets and the dynamics in new packet arrivals. For a practical solution, we consider the following suboptimal approach.

**Suboptimal Solutions and Numerical Results**

**The Frame-by-Frame Problem**

To overcome the difficulties of the optimal solution, we consider a suboptimal frame-by-frame optimization. Even though the packets are to be scheduled for transmission frame by frame, the life span of each packet has to be incorporated in the problem formulation. The decision on transmitting the $i$th packet is based on three aspects:

- The resources required are proportional to $p_i$.
- The delay requirement can be represented by its timeout value in frames, denoted by $D_i$, which is the time after which the packet will not be useful and will be discarded. $D_i$ is the number of frames larger than the current frame in $E_i$.
- To avoid packet loss due to buffer overflow at mobile terminals, it is desirable to give higher priority to the packets in an almost full buffer than to those in an empty buffer, if all other conditions are the same.

Let $P_{l,i}$ denote the packet loss probability caused by the terminal buffer overflow for packet $i$. $P_{l,i}$ is a function of the buffer size and traffic characteristics.

A new weighting factor should be introduced in the throughput calculation in order to incorporate the three aspects of resource allocation. The weighting factor for transmitting the $i$th packet, denoted $w_i$, is a function of $p_i$, $D_i$, and $P_{l,i}$. $w_i$ has a larger value if it takes more system resources (or costs more) to transmit the packet. For example, if a packet has a more stringent delay requirement, it gives the system less flexibility in resource allocation, which may result in lower resource utilization efficiency. As a result, it requires more resources to transmit the packet. Let $S_{i,j}$ be 1 if packet $i$ is assigned to time slot $j$ and 0 otherwise.
and let $M$ denote the number of packets waiting for transmission in the current frame. The objective in the resource allocation is to

\[
\text{Maximize } \sum_{i=1}^{M} \sum_{j=1}^{L} w_i S_{i,j},
\]

subject to the constraints

\[
\begin{align*}
\sum_{i} S_{i,j} & \leq 1, \quad \forall i = 1 \ldots M, \\
\sum_{i} p_i S_{i,j} & \leq N_{\text{max}}, \quad \forall j = 1 \ldots L.
\end{align*}
\]

This is an integer linear programming problem. The problem is NP-complete and generally difficult to solve [9]. Below, we view the resource allocation as a bin-packing problem, and propose a heuristic approach for an efficient solution.

The Bin Packing Problem

The original bin packing problem is a well-known combinatorial problem which seeks the way to pack a set of indivisible blocks into the minimum number of bins. It is known to be NP-complete [9]. In our problem, we can consider the time slots as bins and the packets as blocks with size represented by their code slots ($p_i$ for the $i$th packet). The number of bins is fixed. We want to pack as many blocks as possible in the bins without splitting or exceeding the number of code slots in each bin ($N_{\text{max}}$). In resource allocation to maximize the total weights of the transmitted packets, the weights can be considered as the priorities of the packets in packet scheduling. If the buffer sizes are large for all terminals so that packet loss due to buffer overflow is

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of time slots per frame</td>
<td>4</td>
</tr>
<tr>
<td>Maximum number of code slots per time slot ($N_{\text{max}}$)</td>
<td>6</td>
</tr>
<tr>
<td>Simulation time (frame)</td>
<td>500</td>
</tr>
</tbody>
</table>

| Weighting factor [10]: $w_i = p_i(e^{-A D_i} + B P_{L,i})$ | $e^{-A} = 0.4, B = 0$ |
| Voice traffic: Received power ($p_i$) | 1      |
| Timeout value ($D_i$)                  | 1      |
| Average talk spurt length (frame)      | 2      |
| Average silent period (frame)          | 3      |
| Data traffic: Received power ($p_i$)   | 2      |
| Timeout value ($D_i$)                  | 10     |
| Average arrival interval (frame)       | 10     |
| Average file sizes (packet)            | 5      |

---

**Table 1.** System parameters used in the simulations.

![Figure 4](image_url). The percentage of lost packets per frame: a) voice packets; b) data packets.

![Figure 5](image_url). a) The average waiting time for transmitted data packets; b) the percentage of utilized resources.
very unlikely to happen, the urgency to transmit packet \( i \) depends mainly on its timeout value \( D_i \). In this case, packet scheduling is based on the BER requirement (represented by the \( p_i \) value) and delay requirement (represented by the \( D_i \) value):

- Packet \( i \) should be transmitted before packet \( j \) if \( D_i < D_j \).
- For packets with the same timeout value, we will schedule them by the best-fit-decreasing-height (BFDH) algorithm.

The packets with the largest weight will be scheduled first since the objective is to utilize the system resources as much as possible. Each packet is scheduled into the bin so that the bin will be as full as possible without exceeding the bin capacity; that is, after packing, the unused capacity in the bin is minimum.

**Simulation Results**

In the simulations, we consider two types of traffic: voice and data. The simulation parameters are summarized in Table 1. The simulations are carried out using GAMS with a mixed integer programming (MIP) solver. Figure 3 shows the average transmitted packets per frame for voice and data users. The number of transmitted voice packets increases with the number of voice users until it reaches the maximum, but decreases with the number of data users. Similarly, the number of transmitted data packets increases with the number of data users and decreases with the number of voice users. The system guarantees certain numbers of voice packets and data packets being transmitted even when the traffic load is very high. Figure 4 shows the packet loss rate for voice and data users. The packet loss rate increases with the traffic load. The packet loss rate for voice users is higher than that for data users, because data packets can tolerate a longer transmission delay. Figure 5a shows the average waiting time for transmitted data packets. The waiting time increases with the traffic load. When the traffic load is very high, most data packets are either transmitted with the maximum tolerable delay or lost (the packets cannot be scheduled within their timeout limit). Figure 5b shows the percentage of utilized system resources without considering the overheads due to the request slots. Almost all system resources can be utilized when the traffic load becomes high. However, after the total throughput reaches its maximum, further increasing the traffic load will result in an increased number of lost packets. Given the system resources, a call admission control policy should be designed to control the number of admitted users, in order to achieve high resource utilization efficiency and low packet loss probability.

**Conclusions**

In this article we propose a MAC protocol with a packet scheduling algorithm to support a wide variety of traffic types in the TDD wideband CDMA wireless system. The MAC protocol can efficiently accommodate two-rate (on-off) voice, variable rate video, and bursty data traffic. Packets are scheduled for transmission according to their required transmission accuracy, transmission delay, and probability of buffer overflow. Due to the difficulties in obtaining the optimal resource allocation solution, a suboptimal problem formulation and a heuristic bin-packing algorithm are developed for the packet scheduling problem. For mixed voice and data traffic, the simulation results show that the network resources can be utilized efficiently; voice packets are transmitted before data packets because voice packets have a higher delay requirement; and the transmission delay for data packets increases with traffic load.

**Acknowledgments**

This research was supported by Communications and Information Technology Ontario (CITO) and by Natural Sciences and Engineering Research Council (NSERC) of Canada. This article is based on our previously published material from 3Gwireless 2000 by Delson Group.

**References**


**Biographies**

VINCENT HUANG [S’00] (huang@bbcr.uwaterloo.ca) received an M.Sc. degree in electrical engineering in 1994 from Chalmers University of Technology, Gothenburg, Sweden. He is currently working toward a Ph.D. degree at the University of Waterloo, Canada. From 1994 to 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] (weihua@ieee.org) received a Ph.D. degree (1993) from the University of New Brunswick, Canada, in electrical engineering. Since 1993 she has been a faculty member at the University of Waterloo where she is currently an associate professor in the Department of Electrical and Computer Engineering. Her current research interests include wireless networking for multimedia personal communications and digital transmission over fading dispersive channels.