To appear in *IEEE/ACM Trans. on Networking*, October, 1999.

An Uplink CDMA System Architecture with Diverse QoS Guarantees for Heterogeneous Traffic

Sunghyun Choi and Kang G. Shin

***Abstract*— Addressed are the uplink access control problems for cellular CDMA systems that service heterogeneous traffic with various types of Quality-of-Service (QoS) and use Multi-Code CDMA to support variable bit rates. Considering its distinct QoS requirements, class-I real-time traffic (e.g., voice and video) is differentiated from class-II non-real-time traffic (e.g., data). Connection-oriented transmission is achieved by assigning mobile-oriented code channels for class-I traffic where each corresponding mobile needs to pass an admission test. Class-II traffic is transmitted in a best-effort manner through a transmission-rate request access scheme which utilizes the bandwidth left unused by class-I traffic: whenever a mobile has class-II messages to transmit, the mobile requests code channels via a base station (BS)-oriented transmission-request code channel, then, according to the base station scheduling, the transmission is scheduled and permitted. Addressed are the admission test for class-I connections, transmission power allocation, and how to maximize the aggregate throughput for class-II traffic. The admission region of voice and video connections and the optimum target signal-to-interference ratio (SIR) of class-II traffic are derived numerically. The performance of class-II traffic transmissions in terms of average delay is also evaluated and discussed.**

***Keywords*— CDMA systems, wireless/mobile communication, Multi-Code CDMA, admission control, power control, QoS guarantees, transmission-rate request access protocol, Reed-Solomon/convolutional concatenated code, Automatic Retransmission reQuest (ARQ).**

1 Introduction

Code-Division Multiple-Access (CDMA) is emerging as a promising technique for future cellular mobile/wireless communication systems. As claimed by many researchers [8, 24], CDMA in cellular environments offers several advantages over other wireless access techniques such as high spectral efficiency, soft capacity, soft handoff, and increased system capacity. Here, we consider a direct-sequence (DS) CDMA uplink (mobile-to-base) system to provide three types of QoS for heterogeneous traffic according to the application requirements: (1) bounded packet-delivery delay; (2) guaranteed transmission rate; (3) packet error probability (or bit error probability, BER). Table 1 summarizes the classification of heterogeneous traffic. Basically, class I and II are differentiated according to the required delay and rate performance. Class-I traffic requires bounded delay and guaranteed rate, but "looser" error performance. Voice and video fall into this class, although video requires better error performance and higher rate than voice. Currently, low-rate voice coding schemes like QCELP and VSELP are being used to deliver speech over a wireless link at a rate of 8 Kbps or less [18]. There has also been significant research on low-rate video coding techniques to transmit video at a rate lower than 64 Kbps including the new ITU-T recommendation H.263 [22]. Class-II traffic, like the conventional data service, does require loss-free transmission, but does not require guaranteed rate nor bounded delay. Class II is divided further into two subclasses: (1) class II-A which is delay-sensitive like FTP and remote log-in; and (2) class II-B which is delay-tolerant like paging and e-mail. Class II-A receives priority over class II-B.

To support multi-rate transmissions with DS-CDMA systems, many techniques have been proposed such as multi-modulation CDMA,[1] multi-processing gain CDMA, and multi-code CDMA [10, 11, 17]. We adopt the multi-code (MC) CDMA because it has some advantages over other methods as claimed in the literature; the multi-modulation CDMA degrades the performance for the users with high data rates [17], and the multi-processing gain CDMA is expected to cause problems to users with very high source rate to have too small a processing gain to maintain good cross correlation among different user codes [11]. MC-CDMA is expected to work well with multimedia traffic; when it integrates multimedia traffic, traffic streams with significantly different transmission rates can be easily integrated into a unified architecture, with all the transmission channels having the same bandwidth and spread spectrum processing gain. One weakness of general CDMA techniques is that the data rate of an individual user can hardly be commensurate with the system bandwidth because of the spreading factor even with above-mentioned multi-rate CDMA techniques. In this regard, a carefully-developed Time-Division Multiple Access/Time-Division Duplexing (TDMA/TDD) system like those found in [5, 21] may yield less delays for a small population of users generating very bursty traffic (e.g., a typical wireless LAN environment).

For class-I real-time traffic, connection-oriented transmission is achieved by assigning a set of mobile-oriented

A subset of materials in this paper was presented at *ACM/IEEE MobiCom'97*, Budapest, Hungary, September 26-30, 1997. The work reported in this paper was supported in part by the US Department of Transportation under Grant No. DTFH61-93-X-00017. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the funding agency.

Sunghyun Choi and Kang G. Shin are with the Real-Time Computing Laboratory, Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109 USA (e-mail: shchoi@eecs.umich.edu; kgshin@eecs.umich.edu).

[1] This technique seems to be less known compared to the other two. Basically, it varies the modulation depending on the required maximum rate. For example, the modulation is selected among BPSK, QPSK, and 16-QAM to support a basic rate, up to two times the basic rate, and up to four times the basic rate, respectively [17].

	Class I		Class II-A	Class II-B
Examples	voice	video	remote login	e-mail
Delay	bounded		sensitive	tolerable
Rate	guaranteed		not guaranteed	
	0-8 Kbps	0-64 Kbps	8-128Kbps	
BER	$\leq 10^{-3}$	$\leq 10^{-5}$	≈ 0	

Table 1: Traffic classification.

code channels and a corresponding receiver, like assigning circuits in a circuit-switched network, to each corresponding mobile after it passes the admission test. (One can imagine a pseudo-noise (PN) code as a channel through which information can be transmitted, hence called a *code channel*. Assigning a code channel to a mobile means that the base station (BS) gives the mobile permission to use a particular PN code, and gets ready to receive packets by tuning a receiver to the PN code.) On the other hand, class-II non-real-time traffic does not require bounded delays nor guaranteed rates. Moreover, it is likely to be bursty and randomly-arriving, so it is inefficient to assign a mobile-oriented code channel and a receiver to each mobile for class-II transmissions. We, therefore, adopt a transmission-rate request access scheme, similar, in principle, to the Distributed-Queueing Request Update Multiple Access (DQRUMA) [14, 15], for class-II traffic, in which a mobile, upon generation of a class-II message, requests its transmission in a transmission-request mini-slot, according to a link access scheme similar to the slotted ALOHA, and the BS permits the transmission according to some scheduling principle. The transmission-request scheme is also used for setting up new class-I connections. The third QoS related to error performance is supported by the combination of power control, error control codes, and Automatic Retransmission reQuest (ARQ) scheme (only for class II). A Forward Error Correction (FEC) scheme is used for class I due to the uncertain delivery delays of ARQ schemes.

Our previous work [5] focused on how to make various QoS guarantees for heterogeneous traffic in wireless LANs by allocating mobiles time slots to control their access to the LANs. In an earlier version of this paper [4], we did not consider multipath fading, which is the most prominent characteristic of a wireless channel. In case of fading, we use different strategies for the same system and observe performance degradation. Other researchers have also considered how to support different types of traffic in CDMA systems. The authors of [23] considered how to support voice and data in a CDMA system using the slotted ALOHA for data transmission. A demand-assignment access control scheme was proposed in [15] to support efficient multi-rate transmissions according to the mobile's demand. The authors of [9] proposed a receiver-oriented code channel and receiver scheme for data traffic using the ALOHA access. But, all of these have some form of deficiencies on their own: multi-rate is not supported [9, 23], or class-I traffic is not supported [15].

The paper is organized as follows. Section 2 describes the CDMA system specification, the assumed fading model, as well as the traffic models under consideration. The communication protocols for both class I and II traffic are presented in Section 3. The error-control schemes and the associated equations are given in Section 4. Section 5 describes run-time control schemes of the BS including the allocation of a power level to uplink accesses, admission control of class-I connection requests, and scheduling of class-II packet-transmission requests. In Section 6, we present both the analytical and simulation results. Finally, the paper concludes with Section 7.

2 Modeling and Specifications

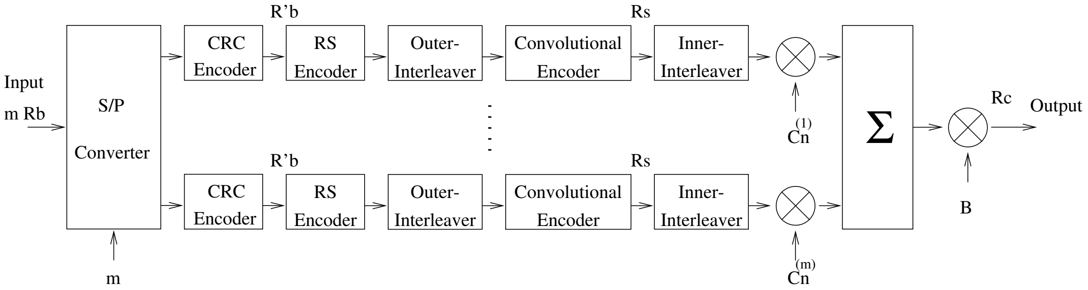
We now describe the specification of the MC-CDMA transmitter which is equipped in each mobile and the Rayleigh fading channel assumed, and also present the class-I traffic models, which are essential to establish the admission test of class-I connections.

2.1 MC-CDMA Transmitter

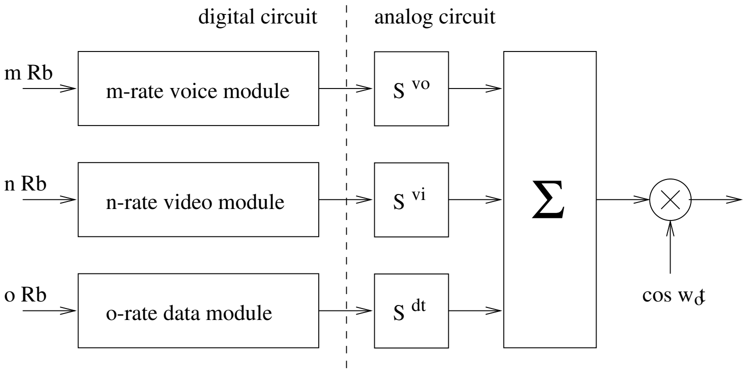
Only the uplink (mobile-to-base) is considered in this paper; the downlink (base-to-mobile) can be supported similarly but more easily, thanks to the broadcast nature of the downlink. Figure 1 shows the block diagram of a transmitter equipped in each mobile. The system is based on the MC-CDMA [11, 12] and the Reed-Solomon/convolutional concatenated coding scheme [1, 6]. Each of these components is described below in detail.

A mobile can transmit its packets at integer multiples of the *basic rate* R_b up to $M \cdot R_b$. We assume $R_b = 8$ Kbps, which is equivalent to the rate for a voice transmission, and $M = 16$ to make $M \cdot R_b = 128$ Kbps, the maximum data transmission rate. We start with an m-rate[2] transmitter module, shown in Figure 1 (a), which accepts an incoming message stream at rate mR_b. First, the stream is converted into m basic-rate streams. Each basic-rate stream is encoded by cyclic redundancy check (CRC) encoders, and then by (N, K, q) Reed-Solomon (RS) encoders. The RS encoder maps K symbols into N symbols each of which consists of a b-binary sequence, where $b = \log_2 q$. An RS codeword corresponds to a packet, and accommodates $K \cdot b - H$ information bits assuming that an H-bit header (including CRC) is added in each packet. A header is appended to all m basic-rate streams

[2] m-rate means 'rate mR_b' throughout this paper.



(a) An m-rate transmitter module



(b) The integrated transmitter

Figure 1: Block diagram for a Multi-Code CDMA transmitter: (a) an m-rate transmitter module with RS/convolutional concatenated encoders and interleavers, where $R_b' = R_b(Kb)/(Kb-H)$; (b) the integrated transmitter, where $m+n+o \leq M$.

once every $K \cdot b - H$ bits to form packets in the serial-to-parallel converter. Note that the RS encoder's input rate is $R_b' = R_b(Kb)/(Kb-H)$. The output of the RS encoder is interleaved by an outer-interleaver on a symbol-by-symbol basis, then encoded by a rate $1/n$ convolutional code with memory size M_{mem}. The output of the convolutional encoder is interleaved by an inner-interleaver on a bit-by-bit basis. The inner-interleaver renders the fading channel memoryless while the outer-interleaver randomizes burst errors at the output of the convolutional (Viterbi) decoder.

When a mobile enters the cell under consideration,[3] it is assigned a primary pseudo-noise (PN) code of chip rate R_c, which is used exclusively by the mobile as long as it remains within the cell. Each of m streams is multiplied by one of orthogonal codes. The set of orthogonal codes is generated by a sub-code concatenation scheme: when C_n^{PN} is the primary PN code of a mobile n, the spreading codes $\{C_n^{(i)}, i = 1, 2, \cdots, M\}$ are obtained by

$$C_n^{(i)} = C_n^{PN} \times D_i, \quad D_i \perp D_j, \quad i \neq j, \tag{1}$$

where D_i's are from a set of Walsh (i.e., orthogonal) codes. Note that $C_n^{(i)} \perp C_n^{(j)}, i \neq j$, is also guaranteed. Using this sub-code concatenation, it was claimed that there was zero interference across the received orthogonal codes so that multi-rate transmission can be achieved with a smaller power than the basic rate transmission [4,11,12,15]. However, this is not the case when the orthogonal codes are transmitted over a multi-path fading channel which is typical in wireless cellular networks.

In fact, multi-rate transmission can still benefit from the sub-code concatenation since, among the various delayed versions of codes, those arriving simultaneously at the BS will not work as interferences to each other. Especially, using the RAKE receiver [19] which can resolve multiple delayed versions of a signal transmitted with a resolution of chip duration T_c $(= 1/R_c)$, this can really happen. Depending on the underlying wireless channel characteristics and the receiver techniques, this benefit could be significant or very slim; for example, if (1) there is a line-of-sight transmission path between a mobile and the BS, and (2) multi-path versions of the transmitted codes are relatively weak, there will be virtually no self-interference. In this paper our CDMA system is assumed to not benefit from self-code concatenation due to multi-path fading (in order to err on the safe side). In the system designed under this assumption, higher-rate transmission can, in practice, have better error performance than lower-rate transmission due to (possibly) lower interferences. After being spread by orthogonal codes, all parallel streams are summed together, then multiplied by a base station code, which is used to indicate which BS the mobile is communicating with.

Each mobile can transmit voice, video, and data traffic at the same time as long as the total transmission rate

[3]Physically, when the user turns on the switch of a handset or the mobile is handed off from an adjacent cell.

does not exceed M. As shown in Figure 1 (b), for each traffic type, a separate transmitter module is used because each traffic type requires a different transmission power level due to its different BER requirement (to be described later). Note that when transmitter modules are in use, orthogonal codes $C_n^{(i)}$ from the same set of M codes will be used. The blocks S^{vo}, S^{vi}, and S^{dt} represent the analog amplifiers with those values of power level for voice, video, and data traffic, respectively. Note that the blocks before the amplifiers are implemented with a digital circuit, while the others are implemented with an analog circuit. All the amplified outputs of modules are summed together, multiplied by the carrier frequency, then transmitted.

2.2 Fading Channel and Power Control

A frequency non-selective and slowly-fading channel with Rayleigh-distributed envelope statistics is assumed in this paper. The frequency non-selective channel results in multiplicative distortion of the transmitted signal at the BS. Furthermore, the condition that the channel fades slowly implies that the multiplicative process may be regarded as a constant during at least one signaling interval [19]. In fact, this is one of the simplest among the known fading channel models. Depending on the actual channel model and the receiver techniques used (e.g., with or without a RAKE receiver), actual performance may vary, but the general performance trends, and the system architecture considered in this paper are still valid irrespective of the assumed channel model and the receiver techniques.

A wireless channel, in reality, is subject to (1) the long-term propagation loss due to the distance from mobiles to the BS and (2) the shadowing effects of buildings and other objects in addition to the above-mentioned short-term fading. We assume that both can be adjusted through power control, which is the case in real CDMA systems [24]. Both open- and closed-loop power-control mechanisms will be used to control the received power at the BS. Time-averaged versions of the received power are used for closed-loop power control, and hence the short-term fading is not mitigated by power control. However, we assume that the *average* received signal-to-interference ratio (and hence, the average received signal power) at the BS can be maintained according to a target value assigned to the traffic type at a given traffic load. How to control the transmission power itself is out of the scope of this paper. We only design the mechanism to determine and assign the target power level for each traffic type depending on a given traffic load.

2.3 Class-I Traffic Modeling

In our proposed scheme, class-I traffic is handled by assigning specific code channels to mobiles similarly to assigning frequency slots in FDMA, time slots in TDMA, and more generally, circuits in a circuit-switched network. However, the underlying concept is totally different because in CDMA systems, all mobiles share the entire frequency and time slots. Depending on their traffic characteristics, we can admit many more mobiles simultaneously into the system than the physical link capacity without violating the required BER because the CDMA system is interference-limited. In a sense, the CDMA system achieves statistical multiplexing. However, to achieve the required error performance, we need to control the admission of class-I traffic which, in turn, requires appropriate traffic modeling. Class-II traffic does not affect the admission control of class-I traffic because the transmission of each class-II packet is permitted by the BS, depending on the available bandwidth left unused by class-I traffic in our protocol.

It is well-known that voice traffic is characterized by an "on-off" model since a speech signal is in either *talk-spurt* or *silent* mode during the conversation of a mobile [2]. Each voice mobile is modeled by a two-state Markov chain where one state represents ON (transmitting at rate R_b) and the other represents OFF (idle). The utilization ρ_{vo} (i.e., fraction of time during which the voice mobile is active) is assumed to be 0.5. Given the number K_{vo} of voice mobiles, the number K_{vo}^a of active voice mobiles at current time can be computed by a binomial distribution:

$$P_{K_{vo}^a}(k) = \binom{K_{vo}}{k}(0.5)^{K_{vo}}. \tag{2}$$

Since there does not exist any proper model for low-rate video coding, we adopt a very simple model to represent the variable bit rate (VBR) characteristics of video traffic: each video mobile is modeled by a three-state Markov chain where in the first state, the mobile transmits packets at the rate of $4R_b$; in the second, at the rate $6R_b$; and in the third, at the rate $8R_b$ (= 64 Kbps). The mobile will reside in each of the three states with probability 0.25, 0.5, and 0.25, respectively. Given the number K_{vi} of video mobiles, the number $K_{vi}^a(j)$ of j-rate video mobiles is given by a 3rd-order multinomial distribution:

$$P_{K_{vi}^a(4),K_{vi}^a(6),K_{vi}^a(8)}(k_4,k_6,k_8) = \binom{K_{vi}}{k_4,k_6,k_8}(0.5)^{k_6}(0.25)^{k_4+k_8}, \tag{3}$$

where $k_4 + k_6 + k_8 = K_{vi}$. For other video traffic models, this distribution can still be obtained if the rate distribution is given for a model used.

3 Protocol Description

Connection-oriented transmission is achieved by assigning a set of mobile-oriented code channels for each class-I mobile. On the other hand, whenever a mobile has class-II messages to transmit, it requests code channels via a BS-oriented transmission-request code channel (or a piggyback rate-request code channel), then, according to the BS's scheduling policy, the transmission is scheduled and granted. Note that for a mobile to send a packet through a code channel, the BS should be ready to receive it by tuning a receiver to the corresponding PN code. As shown in

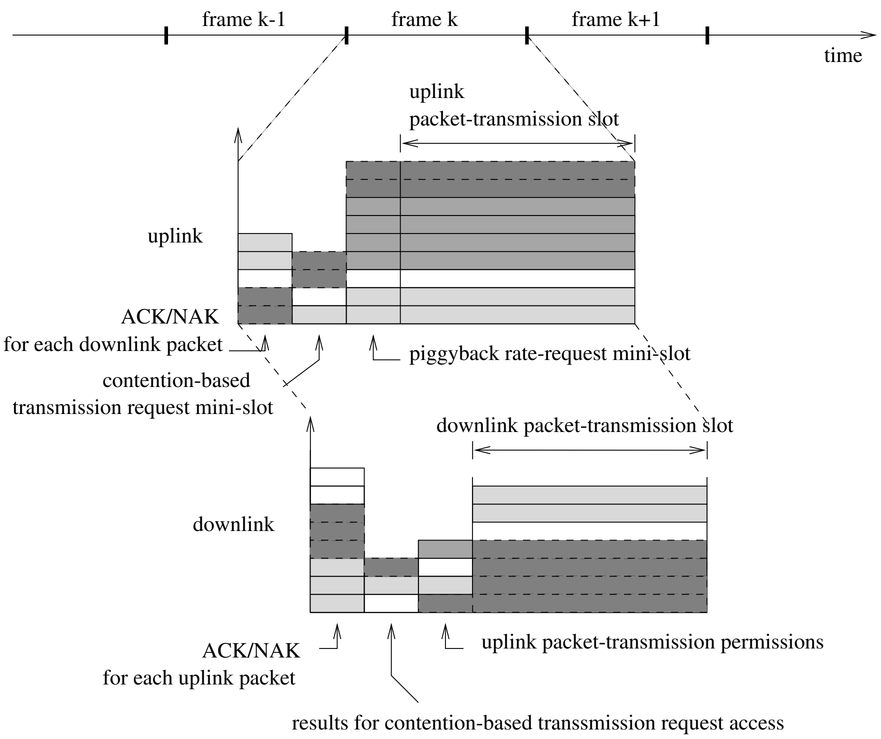


Figure 2: The uplink and downlink frames along the time-axis.

Figure 2, our scheme is based on the uplink and downlink frame formats, which are drawn basically from [15]. The uplink time-axis is divided into frames which consist of a mini-slot for ACK/NAK, a contention-based transmission-request mini-slot, a piggyback rate-request mini-slot, and a packet-transmission slot. On the other hand, the downlink time-axis is divided into frames which consist of a mini-slot for ACK/NAK transmission, a mini-slot for result announcements of contention-based transmission-requests, a mini-slot for uplink packet-transmission permissions, and a slot for downlink packet transmission. In the figure, the rectangles arranged vertically in a slot or mini-slot correspond to different packet or transmission request or ACK/NAK transmissions (with different PN codes in most cases except for the contention-based transmission request mini-slot.) Basically, rectangles with the same shade represent transmissions from (to) the same mobile if they are for uplink (downlink) except for the contention-based transmission request mini-slot. The role of each slot and mini-slot will be explained below in detail.

3.1 Contention-Based Transmission-Request Accesses

A set of BS-oriented code channels in each contention-based transmission-request mini-slot are used for class-II transmission-request access: a finite number, say, N_{rcvr}^{rqst}, of receivers (and hence code channels) are dedicated for mobiles' transmission-request accesses. An ALOHA-like protocol is used for the request accesses. This contention-based request is basically used by a mobile which is not permitted to transmit any packet during the next frame and falls into one of the following three cases: (1) when the mobile wants to establish a new class-I connection; (2) when it generated new class-II messages after transmitting all messages for which it previously requested; and (3) when it wishes to transmit class-II messages at a higher rate than the rate it previously requested due to the generation of new messages. Such a mobile waits for the next contention-based transmission request mini-slot then decides whether to send the request or not with probability p_r (i.e., in a p_r-persistent manner). If it decided to send, the mobile chooses one of N_{rcvr}^{rqst} code channels, and transmits the request via the code channel in the mini-slot. Each request includes the relevant information such as mobile ID, class type (voice, video, class II-A, or class II-B), and the required transmission rate, up to MR_b (or equivalently, the number of packets, up to M packets, to be transmitted in the next frame) if it is for class-II traffic. The packets to be retransmitted by the ARQ scheme will not be included in the calculation of the required transmission rate.

A request can be received erroneously by the BS for the following three reasons. First, the total interference level exceeds the link capacity[4] and many of the requests in the mini-slot may be received with error. Second, the request collided with another request transmitted in the same request code channel, and only the collided requests are received with error.[5] Third, a request may be received

[4]Due to the CDMA's soft capacity characteristic, the request might be received correctly even under an over-threshold situation.

[5]Transmission by different users on the same code could not collide if there is a relative delay of a few chips between the signals arriving at the receiver if PN codes with good auto-correlation property are used.

erroneously due to random noise and interference. In Figure 2, we see there are four transmission requests in frame k. Two upper requests with the same shade are found to have used the same code channel, so will collide with each other. Only (lower) two requests out of four can be successful. Only successful requests are broadcast to the mobiles through the downlink channel before the next contention-based transmission request mini-slot. The value of p_r is determined by each mobile in a distributed fashion. For the first attempt of a message transmission request, p_r is set to 1. Upon failure of a request, however, the value of p_r is decreased to

$$p_r := \frac{p_r}{p_r + 1}, \tag{4}$$

following $1, 1/2, 1/3, \cdots$ (i.e., the "harmonic" backoff) [14]. Each failed request is retransmitted in a p_r-persistent manner until it is successfully received, and a mobile with a failed request is said to be *backlogged*.

Note that the choice of the slotted ALOHA with the harmonic backoff for the contention-based request access is not mandatory. We could use other access techniques such as the Binary Stack Algorithm [3, 14] or the slotted ALOHA with other backoffs such as the binary exponential backoff as used for the IEEE 802.3 Ethernet. Depending on the request access method, actual performance will vary. However, in our system, this contention-based access is used for request access only. Moreover, the BS can control the number N^{rqst}_{rcvr} of receivers for request access depending on the observed request access delay so that the delay would not be very large. So, the actual effect of choosing a request access method could be kept minimal. We adopted the slotted ALOHA with the harmonic backoff for simplicity.

3.2 Collision-Free Piggyback Class-II Rate Requests

When a mobile is scheduled and permitted to transmit class-II packets in the packet-transmission slot of the subsequent frame, it is supposed to update the required transmission rate (for the next scheduled transmission-slot of that mobile) through the piggyback rate-request mini-slot. These piggyback requests are collision-free because, during the piggyback mini-slots, mobiles use code channels dedicated to the mobiles in the subsequent transmission slot. Through the piggyback mini-slot, a currently-transmitting mobile can request a transmission rate including the rate of newly-generated messages without involving the contention-based request accesses, thus reducing the average request access delay. A class-I mobile can also request the transmission of newly-generated class-II messages through the piggyback rate-request mini-slot, so class-I mobiles do not need the contention-based transmission-request accesses. A mobile which has voice and/or video connections can request the transmission of class-II data up to rate $15R_b$ (with a voice connection), $8R_b$ (with a video connection), and $7R_b$ (with both voice and video connections).

Each piggyback request includes such information as the class type (voice, video, class II-A or II-B) and the requested transmission rate. As shown in Figure 2, as many piggyback requests as the number of packet transmissions permitted for a mobile during a frame can be transmitted within a piggyback mini-slot. However, up to four piggyback requests will be transmitted: one for each class type.

3.3 Class-I Connection Establishment

The required QoS for class-I traffic is the bounded delivery delay of packet transmission. To this end, the BS will set up a class-I connection for each class-I transmission request by assigning a set of mobile-oriented code channels to the corresponding mobile. An important matter here is that the class-I connection request should be delivered to the BS as fast as possible. In our system, there are two ways to transmit requests to the BS, and both class I and II share these. To reduce the access delay of a class-I connection request, (1) a class-I connection is requested using either of two ways whichever becomes available to the mobile earlier. Note that a mobile already in the BS queue cannot make a new class-II transmission request using the contention-based request access. (2) Priority is given to class-I connection requests over class-II transmission requests. For example, a mobile which has both requests and not yet in the BS queue will transmit the class-I connection request only during a transmission-request mini-slot.

Upon receiving a request to establish a class-I connection in a mini-slot, the BS performs an admission test on the request according to the link capacity. The admission test checks if the total required bandwidth for class-I traffic (or *equivalent bandwidth* BW_{eq} in Section 5) is less than, or equal to, the link capacity C. If the test result is positive, the BS admits the mobile via a transmission-permission mini-slot, and assigns a receiver to the corresponding mobile while the receiver tunes to the mobile's code channels using the mobile's primary code and the corresponding orthogonal codes, then the mobile continues to transmit packets whenever it wants. If the test result is negative, the BS declares *connection blocking*, and notifies it to the mobile.

An admitted mobile will transmit its packets in each packet-transmission slot with the power level controlled by the BS, depending on the traffic condition. The transmission rate (or the number of packets transmitted in a slot) can vary according to the amount of traffic generated. Because each class-I mobile can transmit packets at a rate up to the contracted value (i.e., R_b for voice and $8R_b$ for video) after its admission into the system, there is no queueing delay, so the delivery delay is bounded, i.e., the first QoS is guaranteed.

3.4 Class-II Message Transmission

Each of successful transmission/rate requests is queued in one of two base station service queues, i.e., one for class II-A and the other for class II-B. The BS serves (i.e., gives the transmission permission to) the requests according to a transmission scheduling policy with the bandwidth available for class-II, BW_{dt} $(= C - BW_{eq})$, which depends on the admitted pair (K_{vo}, K_{vi}) of voice and video mobiles. The relation between BW_{dt} and the transmission scheduling of class-II traffic is detailed in Section 5.3. Transmission permissions for the subsequent frame are announced to mobiles in the middle of each slot via the downlink as shown in Figure 2. After sending the transmission permissions, the BS assigns a number of receivers, so that they tune to the scheduled mobiles' code channels. Each permission includes such information as mobile ID and transmission rate.

Since class-II traffic uses a selective-repeat ARQ scheme, the feedback from the BS should be announced. For each transmission of packets, ACK (acknowledge) or NAK (non-acknowledge) is fed back. NAKed packets are retransmitted with priority over other packets in the next frame during which the mobile is permitted to send packets; if a mobile has NAKed packets, only after retransmitting all of them, it will transmit other newly-generated packets. An erroneous packet transmitted in the i-th frame can be retransmitted as early as in the $(i+2)$-th frame because the corresponding NAK is fed back to the mobile during the $(i+1)$-th frame.

4 Error Performance

As described in Section 2, the RS/convolutional concatenated coding is adopted to protect data from the errors caused by the Rayleigh fading and additive white Gaussian noise (AWGN). A dual-mode RS decoder is used; for class-I traffic, an error-correction-only decoder is used while for class-II, an error-correction-and-detection decoder is used for retransmitting corrupted packets with selective-repeat ARQ.

4.1 Demodulator and Convolutional Decoder

The bit error probability at the output of the convolutional decoder can be upper-bounded [19] by:

$$P_b \le \sum_{d=d_{free}}^{\infty} b_d P_d, \tag{5}$$

where d_{free} is the free distance of the convolutional code, b_d is the total number of nonzero information bits on all weight-d paths in the trellis diagram of the convolutional code, and P_d is the pairwise error probability, which is, the probability that an incorrect path at distance d from the correct path being chosen by the Viterbi decoder. When the binary phase shift keying (BPSK) demodulator and soft decision decoding is applied, $P_d(\gamma_{o,i}, i = 1, \cdots, d)$ for a given set of the signal-to-noise ratio (SIR) $\gamma_{o,i}$ of the i-th error symbol in the incorrect path is represented by

$$\begin{aligned} P_d(\gamma_{o,i}, i = 1, \cdots, d) &= Q\left(\sqrt{2\sum_{i=1}^{d}\gamma_{o,i}}\right), \\ &\le \frac{1}{2}\prod_{i=1}^{d} e^{-\gamma_{o,i}}. \end{aligned} \tag{6}$$

By averaging Eq. (6) over $\gamma_{o,i}$'s which are exponentially distributed in a Rayleigh fading environment, we obtain

$$\begin{aligned} P_d &\le \frac{1}{2}E\left[\prod_{i=1}^{d} e^{-\gamma_{o,i}}\right], \\ &= \frac{1}{2}\left(E\left[e^{-\gamma_o}\right]\right)^d, \\ &= \frac{1}{2}\left(\frac{1}{1+\overline{\gamma}_o}\right)^d, \end{aligned} \tag{7}$$

where Eq. (7) is obtained since $\gamma_{o,i}$'s are independent due to the ideal inner-(de)interleaver, and Eq. (7) is obtained by using the pdf of exponentially-distributed γ_o with the average SIR $\overline{\gamma}_o$. Now, when the BPSK demodulator, an ideal inner-(de)interleaver, and soft decision convolutional decoding are used, the upper bound of the average bit error probability at the output of the convolutional decoder is given by

$$P_b \le \frac{1}{2}\sum_{d=d_{free}}^{\infty} b_d \left(\frac{1}{1+\overline{\gamma}_o}\right)^d. \tag{8}$$

4.2 Reed-Solomon and CRC Decoders

We adopt a bounded distance RS decoder of the (N, K, q) RS code over GF(q) with the error correction capability t $(= \lfloor (N-K)/2 \rfloor)$. That is, the decoder looks for a codeword within distance t of the received word; if there is such a codeword, the decoder finds it, and if not, the decoder declares "decoding failure." A q-ary symbol is mapped to b bits, so $q = 2^b$. If P_s denotes the symbol error rate at the input of the RS decoder, then the probability P_T of the total error of the RS decoder which includes both decoding error (i.e., decoding into a wrong codeword) and decoding failure (i.e., inability to decode) is given by

$$P_T = \sum_{i=t+1}^{N} \binom{N}{i} P_s^i (1-P_s)^{N-i}, \tag{9}$$

where $t = \lfloor (n-k)/2 \rfloor$ is the RS error-correction capability, and P_s can be upper-bounded by

$$P_s \le bP_b, \tag{10}$$

with the bit error probability P_b of the Viterbi decoder given in Eq. (8). Now, the probability P_E of RS decoding error and the probability P_F of RS decoding failure are given [16] by:

$$P_E \le P_T \sum_{i=0}^{t} \binom{N}{i} (2^b - 1)^{i-(N-K)}, \tag{11}$$

$$P_F = P_T - P_E. \tag{12}$$

When a decoding error happens with probability P_E, the errors in the received word are not detected by the RS decoder. In most cases, these undetected errors are detected by the CRC decoder, which is located after the RS decoder. Through two-level error detection by a combination of the RS decoder in the error-correction-and-detection mode and the CRC decoder, virtually all uncorrectable errors can be detected.

A detected error in a class-II packet will be informed to the mobile via the downlink ACK/NAK-announcing mini-slot, and will trigger the retransmission of the packet by the selective-repeat ARQ. Note that the packet retransmission probability P_R can be approximated by the total error probability, i.e.,

$$P_R \approx P_T, \tag{13}$$

and, the bit error probability at the output of the CRC decoder for class-II traffic will be virtually zero, i.e.,

$$P_{eb,II} \approx 0. \tag{14}$$

Now, when the RS decoder is used for class-I traffic, the decoder will be switched to the error-correction-only mode, in which, instead of declaring a decoding failure, the decoder will just bypass the received word to the output of the decoder. (The CRC decoder will do the same.) Without loss of the generality, we derive the probability $P_{eb,I}^1$ that the first bit of the received word is in error at the output of the RS decoder in the error-correction-only mode in order to obtain the bit error probability $P_{eb,I}$ at the output of the CRC decoder for class-I traffic as in Eq. (15).

5 Run-Time Control Schemes

At run-time, the BS exercises the following three types of control: (1) allocation of a power level for each packet transmission; (2) admission control of each newly-requested class-I connection; and (3) class-II traffic transmission permission control. We use a *unit of bandwidth* or a BU which is the amount of bandwidth to support an active voice mobile.

5.1 Power-Level Allocation

Let's consider how to set the target SIR for each traffic class while meeting different error-performance requirements for different traffic classes. Given modulation and coding schemes, the BER of class-I traffic is determined directly from the received average SIR as shown in Eq. (15). So, for each class-I traffic, the *target* SIR of a convolutionally coded symbol received is assigned according to its required BER. On the other hand, a virtually error-free class-II communication is achieved via error detection and retransmission basically irrespective of the received average SIR. However, we can determine the optimum target SIR of class-II traffic which maximizes the aggregate throughput of class-II traffic. How to determine the target SIRs will be presented in Section 6.1.

Now, we assume that there are K_{vo} voice mobiles with target SIR $\hat{\gamma}_o^{vo}$ and K_{vi} video mobiles with target SIR $\hat{\gamma}_o^{vi}$. In the next frame, $K_{dt,k}^a$ k-rate data mobiles, $k = 1, 2, \cdots, M$ $(= 16)$, with target SIR $\hat{\gamma}_o^{dt}$ are scheduled to transmit packets. Note that a single mobile can simultaneously have all of voice, video, and data traffic. Interference from the other mobiles can be approximated by a Gaussian noise with a zero mean. This approximation was used first in [20], and widely accepted. When random PN codes are used, the required average energy of convolutionally-coded symbol at the receiver for a voice packet can be obtained as:

$$\overline{E}_s^{vo} = \frac{1.5PGN_o}{(\hat{\gamma}_o^{vo} + 1.5PG)/\hat{\gamma}_o^{vo} - I_{est}}, \tag{16}$$

where I_{est} is the total estimated interference level during the subsequent slot as a function of K_{vo}, K_{vi}, and $K_{dt,k}^a$ (to be defined later), N_o is the single-side power density of white Gaussian noise, and PG is the processing gain of the CDMA system. The processing gain is defined as $PG = R_c/R_s$, where R_c is the chip rate of the PN sequence and R_s is the convolutionally-coded symbol rate (as shown in Figure 1). The analysis of DS/CDMA system similar to that of an MC-CDMA system can be found in [15] which did not consider traffic heterogeneity or classification.

For a mobile which will transmit voice at i-rate, video at j-rate, and data at k-rate, where $i + j + k \leq M$, during the subsequent slot, the corresponding average symbol energies (i.e., $\overline{E}_s^{vi}$ for video, and $\overline{E}_s^{dt}$ for data) relative to $\overline{E}_s^{vo}$ can be given by

$$\begin{aligned} \overline{E}_s^{vi} &= \beta_{vi}\overline{E}_s^{vo}, \\ \overline{E}_s^{dt} &= \beta_{dt}\overline{E}_s^{vo}, \end{aligned} \tag{17}$$

where the relative energy ratios are

$$\begin{aligned} \beta_{vi} &= \frac{\hat{\gamma}_o^{vi}}{\hat{\gamma}_o^{vo}}, \\ \beta_{dt} &= \frac{\hat{\gamma}_o^{dt}}{\hat{\gamma}_o^{vo}}. \end{aligned} \tag{18}$$

Now, the value of the total estimated interference level I_{est} during the subsequent slot is given by

$$I_{est} = BW_{eq} + \beta_{dt}\sum_{k=1}^{M} kK_{dt,k}^a, \tag{19}$$

where BW_{eq} is the equivalent bandwidth which was reserved for class-I traffic, and $K_{dt,k}^a$ is the number of k-rate data mobiles scheduled. BW_{eq} is determined for a given pair (K_{vo}, K_{vi}) of admitted voice and video mobiles during the admission control phase, as described in the next subsection. Note that for simplicity the expression of I_{est} does not account for the inter-cell interference. By considering the inter-cell interference, some additive terms will be included in the equation, depending on the traffic load in adjacent cells. The power level to be calculated is for the next frame, so the BS does not know the actual rates of class-I connections in that frame. Hence, the BS just assumes that BW_{eq} is always consumed by class-I connections. Note that Eqs. (16) and (17) represent the average symbol energies, not the instantaneous symbol energies.

$$
\begin{aligned}
P_{eb,I} &= P^1_{eb,I}, \\
&= Pr(\text{1st bit error } \& \ i \geq t+1 \text{ errors out of } N \text{ symbols}), \\
&= Pr(\text{1st bit error}) \cdot Pr(i \geq t+1 \text{ errors out of } N \text{ symbols}|\text{1st bit error}), \\
&= P_b \cdot Pr(i \geq t \text{ errors out of the last } N-1 \text{ symbols}), \\
&= P_b \cdot \sum_{i \geq t}^{N-1} \binom{N-1}{i} P_s^i (1-P_s)^{N-1-i}, \\
&= P_b \cdot \sum_{i \geq t+1}^{N} \frac{i}{N} \binom{N}{i} P_s^{i-1} (1-P_s)^{N-i}.
\end{aligned} \tag{15}
$$

Depending on the distance between the BS and a mobile, and the channel condition, the instantaneous symbol energies for a given slot will vary. The BS actually controls the transmission power of each mobile's each traffic class according to these average symbol energies using combined open-loop and closed-loop power control techniques [24]. We do not consider how to control the transmission power in detail according to the allocated power level because it is beyond the scope of this paper.

In Eq. (16), the following condition should hold:

$$
I_{est} \leq \frac{\hat{\gamma}_o^{vo} + 1.5PG}{\hat{\gamma}_o^{vo}} - \Delta_1 = C, \tag{20}
$$

where Δ_1 is called the *link capacity reservation factor* as it determines the bandwidth that should remain unallocated, and the value C represents the link capacity. Note that (i) if $\Delta_1 = 0$ then $\frac{E_s^{vo}}{N_o}$ could be infinite when Eq. (20) is satisfied with the equality, and (ii) using Δ_1, we can bound the maximum transmission energy. The required power levels of a voice mobile for the subsequent frame calculated using Eq. (16) is announced to the mobiles through the downlink, and the mobiles will calculate their corresponding powers using Eq. (17).

5.2 Admission Control of Class-I Connections

Before admitting each class-I connection, the corresponding mobile needs to pass the admission test, determining if the new connection can be established without affecting the QoS of the existing connections. The admission test can be derived easily by modifying Eqs. (19) and (20) as:

$$
BW_{eq} \leq \frac{\hat{\gamma}_o^{vo} + 1.5PG}{\hat{\gamma}_o^{vo}} - \Delta_1 - \Delta_2 = C - \Delta_2, \tag{21}
$$

where Δ_2 is the link capacity reservation factor for class-II traffic which is a design or dynamic parameter that depends on the underlying traffic condition. BW_{eq} is a function of the pair (K_{vo}, K_{vi}) including the newly-requested connection.

Suppose there are K^a_{vo} active voice mobiles and $K^a_{vi,j}$ j-rate video mobiles, such that $\sum_{j=0}^{8} K^a_{vi,j} = K_{vi}$ in a given slot. Assuming that all the available bandwidth BW_{dt} ($= C - BW_{eq}$) for class-II traffic is used by class-II traffic for a specific value of BW_{eq}, average SIRs for voice and video mobiles, respectively, are given by:

$$
\begin{aligned}
\overline{\gamma}_o^{vo} &= \left[\frac{N_o}{\overline{E}_s^{vo}} + \frac{K^a_{vo} - 1 + \sum_{j=1}^{8} \beta_{vi} j K^a_{vi,j} + BW_{dt}}{1.5PG} \right]^{-1}, \\
\overline{\gamma}_o^{vi} &= \frac{1.5PG}{\left(\frac{1.5PG}{\overline{\gamma}_o^{vo}} + 1 \right) / \beta_{vi} - 1}.
\end{aligned} \tag{22}
$$

The BERs for each state $(K^a_{vo}, K^a_{vi,j}, j = 0, \cdots, 8)$ can be obtained by using the equations in Section 4. Because we can calculate the distributions of active voice mobiles and actual video rates from Eqs. (2) and (3), the average BERs can be obtained for a given set of $(K_{vo}, K_{vi}, BW_{eq})$. Using an iterative search, the minimum BW_{eq} satisfying the admission condition in Eq. (21) and all of the required BERs can be obtained for a given pair (K_{vo}, K_{vi}). If there does not exist such a BW_{eq}, the corresponding numbers of voice and video mobiles cannot be admitted into the system. Because it is computationally expensive to calculate BW_{eq} for each pair of voice and video mobiles, one can calculate it off-line and store the results in the BS to be used at run-time.

5.3 Class-II Transmission Permission Control

Class-II traffic transmission-permission is controlled under a round-robin scheduling policy with a set of four prioritized queues. When there are M_o mobiles in the cell, the BS is equipped with the following four queues (in that order of priority): (1) a first-in-first-out (FIFO) queue for class II-A retransmissions; (2) a class II-A round-robin queue (with M_o entities, the number of mobiles in the cell); (3) a FIFO queue for class II-B retransmissions; and (4) a class II-B round-robin queue (with M_o entities). The i-th entity of a round-robin queue, corresponding to mobile i, contains its requested transmission rate $R_{req}[i]$, and is served in a round-robin fashion. If a number of transmitted packets by a mobile results in an error, the mobile ID and the number of the packets (R_{ret}) will be placed at one of two FIFO queues, depending on the packets' priority, for re-

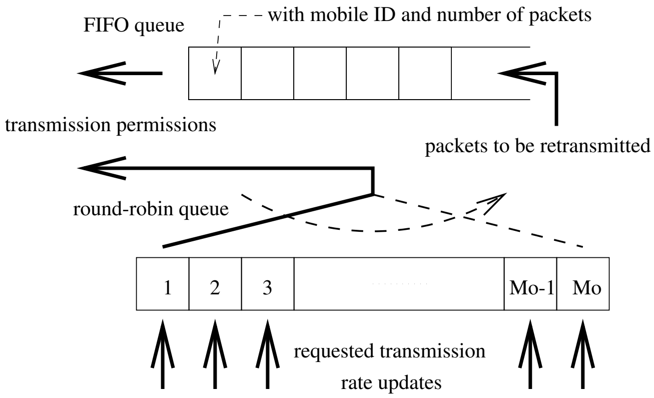


Figure 3: A FIFO queue for retransmission and a round-robin queue.

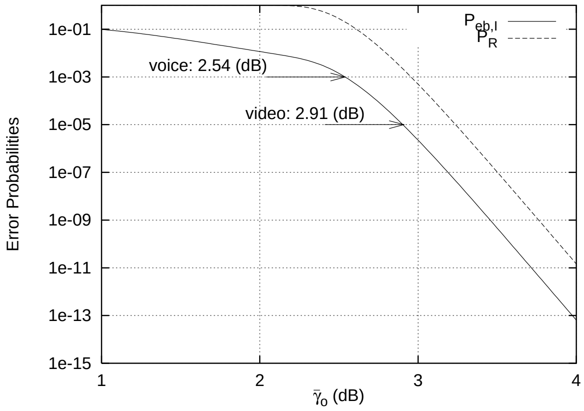


Figure 4: Error probabilities vs. average SIR $\overline{\gamma}_o$ for the RS/convolutional concatenated coding system.

transmission. Figure 3 shows a pair of FIFO and round-robin queues for M_o mobiles in the cell.

First, the permitted transmission-rate index $R_{per}[i]$ for mobile i, the maximum transmission-rate index $R_{max}[i]$ for mobile i, and the residual bandwidth BW_{res} are defined. For the subsequent frame, transmission-permissions are scheduled recursively as follows:

1. With the available bandwidth BW_{dt} for class-II traffic, the residual bandwidth $BW_{res} := BW_{dt}$, and for every i, the permitted rate index $R_{per}[i] := 0$. For every mobile i, the maximum transmission-rate index is assigned as in Eq. (23).
2. A mobile is selected according to the order of the class II-A FIFO queue, the class II-A round-robin queue, the class II-B FIFO queue, and the class II-B round-robin queue, where the requested transmission rate is determined by R_{ret} for FIFO queues and $R_{req}[i]$ for round-robin queues, respectively. When mobile i is selected, $R_{per}[i]$ should be less than $R_{max}[i]$. Otherwise, mobile i is skipped, and the next mobile will be selected.
3. The available rate R_{ava} using the residual bandwidth BW_{res} is calculated as:
$$R_{ava} = BW_{res}/\beta_{dt}. \tag{24}$$
4. For an m-rate transmission-request of mobile i, $R_{tmp} := R_{per}[i]$. Check if $m \leq R_{ava}$, then the transmission-permission rate is set accordingly (i.e., if yes, then $R_{per}[i] := m + R_{tmp}$, else $R_{per}[i] := \lfloor R_{ava} \rfloor + R_{tmp}$). Next, if $R_{per}[i] > R_{max}[i]$, then $R_{per}[i] := R_{max}[i]$. Then, the residual bandwidth is adjusted as:
$$BW_{res} := BW_{res} - (R_{per}[i] - R_{tmp}) \cdot \beta_{dt}. \tag{25}$$
5. If $BW_{res} \geq \beta_{dt}$, then go to Step 3 after selecting the next mobile using Step 2, else go to the next step.
6. The value of $R_{per}[i]$ is the transmission-rate to be permitted for mobile i, so the values of $K^a_{dt,k}$, $k = 1, 2, \cdots, M$, can be calculated. Then, calculate the voice-only mobile's power level using Eqs. (16) and (19). The transmission permission and power level are announced to mobiles through the downlink.

6 Performance Evaluation and Discussions

To derive numerical results, we used the best rate-1/2 convolutional code with $d_{free} = 10$ and $(256, 240, 256)$ extended RS code with the error-correction capability $t = 8$. The values of $\{b_d\}$ in Eq. (8) are obtained from [7]. Notice that the total code rate is $1/2 \cdot 240/256 = 15/32 \approx 0.47$. The PN codes with the processing gain $PG = 128$ and the link capacity reservation factor $\Delta_1 = 1$ are used.

6.1 Error Performance and Target SIRs

In Figure 4, we plotted the following two error probabilities as the average SIR $\overline{\gamma}_o$ increases:

Bit error probability $P_{eb,I}$ at the output of the CRC decoder when the error-correction-only mode RS decoder for class-I traffic is used;

Packet retransmission probability P_R when the error-correction-and-detection mode RS decoder and CRC decoder for class-II traffic is used.

According to each BER requirement in Table 1, we can determine the target SIR for each class-I traffic type. For class-I voice and video, we obtain $\hat{\gamma}_o^{vo} = 2.54$(dB) and $\hat{\gamma}_o^{vi} = 2.91$(dB), respectively, from the $P_{eb,I}$ curve because the error-correction-only scheme is used in the RS decoder. In [4], we reported results on the same CDMA system (with a different power allocation) without considering fading. The target SIRs for voice and video traffic were found to be -0.3 and -0.12, respectively, under that condition. About 3 (dB) additional power was found to be necessary to attain the same error performance for each type of traffic in the fading environment. Note the link capacity $C \approx 107$ (BUs) is determined from Eq. (20).

Now, we study the relationship between the power level and the throughput for class-II traffic to determine the target SIR for class-II traffic. Assume that there are only the mobiles with data to be transmitted at the basic rate, and the ideal selective-repeat ARQ is used. We obtain the max-

$$R_{max}[i] \quad := \quad \begin{cases} M, & \text{if mobile } i \text{ has no class-I connection,} \\ M-1, & \text{if mobile } i \text{ has a voice connection,} \\ M-8, & \text{if mobile } i \text{ has a video connection,} \\ M-9, & \text{if mobile } i \text{ has both voice and video connections.} \end{cases} \tag{23}$$

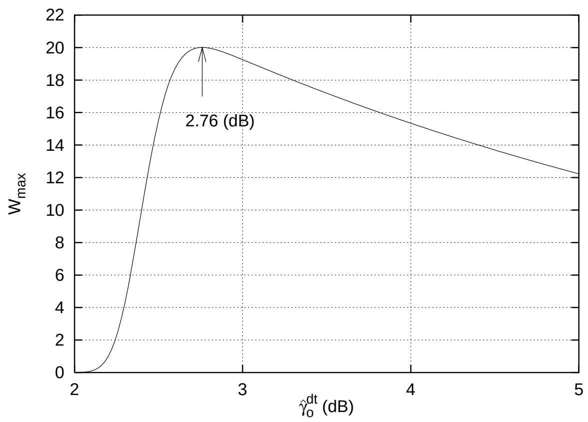


Figure 5: Maximum total throughput at peak traffic W_{max} vs. class-II traffic target SIR $\hat{\gamma}_o^{dt}$ with $BW_{dt} = 0.2C$.

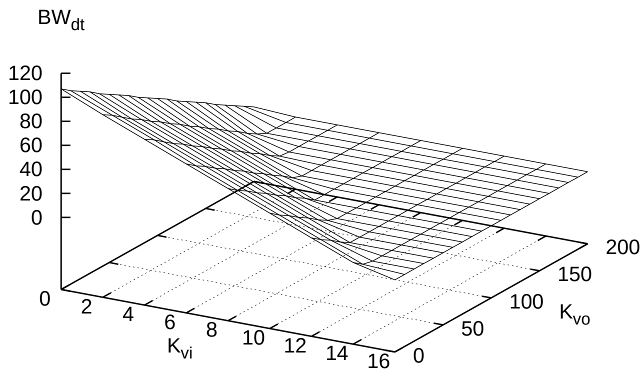


Figure 6: Admission regions for the pairs (K_{vo}, K_{vi}).

imum total throughput at peak traffic for class II, which is defined as the expected total number of packets successfully received at the receiver during one slot in the presence of more class-II traffic demands than the capacity available for class II:

$$W_{max} = (1 - P_R) K^*_{dt,1}, \tag{26}$$

where $K^*_{dt,1}$ is the maximum number of mobiles which can simultaneously transmit data. For a specific value of BW_{dt}, $K^*_{dt,1}$ can be expressed, by modifying Eqs. (19) and (20), as follows:

$$K^*_{dt,1} = \lfloor BW_{dt}/\beta_{dt} \rfloor. \tag{27}$$

There is a tradeoff between $\hat{\gamma}_o^{dt}$ and $K^*_{dt,1}$, i.e., for a low $\hat{\gamma}_o^{dt}$, $K^*_{dt,1}$ would be high, and vice versa. But, because P_R would be high for low $\hat{\gamma}_o^{dt}$, there might exist an optimum $\hat{\gamma}_o^{dt}$ that maximizes W_{max}. In Figure 5, we find the optimum $\hat{\gamma}_o^{dt}$ to be 2.76 (dB). We can easily see that the result will be the same for different transmission rates and BW_{dt} values.

6.2 Admission Region for Class-I Connections

Using three target SIRs and the procedure in Section 5.2, we obtained the admission region shown in Figure 6, where the link capacity $C \approx 107$. The surface represents the bandwidth, BW_{dt}, which can be used for class II, given the admitted voice and video mobiles (K_{vo}, K_{vi}). For a given capacity reservation factor Δ_2 for class II, the pairs (K_{vo}, K_{vi}) with $BW_{dt} \geq \Delta_2$ can be admitted to the system. From the numerical results, we found that a voice mobile consumes roughly 0.53 BU while a video mobile consumes 7.07 BUs, and we can use the following equation for the admission test:

$$\lceil 0.53 K_{vo} + 7.07 K_{vi} \rceil \leq C - \Delta_2. \tag{28}$$

It can reduce the BS's storage cost at the expense of some possible connection blocks due to the inaccuracy caused by using the ceiling function in the above equation.

6.3 Performance of Class-II Traffic Transmission

Finally, let's consider class-II performance in terms of the average message transmission delay. A message transmission delay is defined by the time duration from the generation of a message to the transmission of its last packet. We used the following assumptions and parameters for the simulations:

1. There are 50 mobiles which generate only class-II traffic throughout the simulation.
2. We neglect the noise and fading effects in the request transmissions for simplicity; each transmission request is received correctly unless it collides with another request in the same code channel.
3. Throughout the simulation, the value of BW_{dt} is kept constant; the class-I connections are not changed.
4. The number of packets in a class II-A (II-B) message is geometrically distributed with average 2 (18).
5. For a total message generation rate λ (messages/frame), class II-A (II-B) messages are generated independently in each of 50 mobiles according to a Poisson process with rate $0.9\lambda/50$ $(0.1\lambda/50)$.

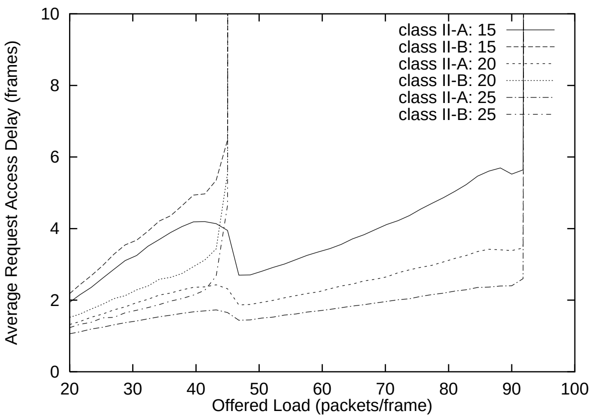


Figure 7: Average request access delay vs. offered load with $BW_{dt} = 50$.

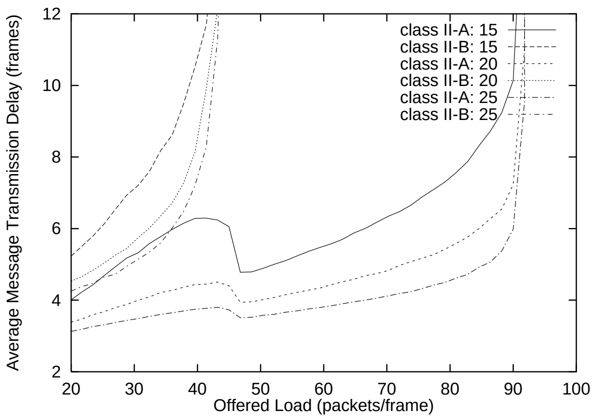


Figure 8: Average message transmission delay vs. offered load with $BW_{dt} = 50$.

Note that the offered load, defined by the overall expected number of the packets to be transmitted within a frame, for the overall message-generation rate λ (messages/frame) is $(0.9 \cdot 2 + 0.1 \cdot 18)\lambda = 3.6\lambda$ (packets/frame), in which a half of the messages is for class II-A and the other half is for class II-B.

A message-transmission delay consists of three components: (1) a request access delay, i.e., from the generation of a message to the time at which the mobile notifies the required transmission rate to the BS through a contention-based transmission-request mini-slot or a contention-free piggyback rate-request mini-slot. Physically, a mobile requests the required transmission rate, not the transmission of each message, but from an analytical point of view, we can imagine that a successful transmission of a rate request means the successful transmission requests for all those messages that have already been generated; (2) a queueing delay of the request in the BS queues, or equivalently, the delay from the successful rate request after the generation to the transmission of the first packet of the message; and (3) a transmission delay, i.e., the time from the transmission of the first packet to the completion of the last packet transmission.

We first consider the request access delay of the protocol. Figure 7 plots the average request access delay as the offered load increases for both class II-A and II-B with the transmission-request access channel number $N_{rcvr}^{rqst} = 15$, 20, and 25. The average request access delay is always greater than, or equal to, 0.5 because message-generation times are uniformly distributed within a frame time. Because a request for class II-A is given priority over that for class II-B, the request access delays reflect this fact. We observe that for the marginal offered load, i.e., near 47 for class II-B, and around 95 for class II-A, the delays approach infinity, because at these loads, the existing messages are rarely served, so the mobiles can rarely request newly-generated messages. The reason why the marginal load for class II-A is twice that for class II-B is somewhat obvious because each half of the offered load is from each subclass. We observe that the larger N_{rcvr}^{rqst}, the lower the average request access delay, since the larger N_{rcvr}^{rqst}, the less likely a request results in a collision. One can see that the delay for class II-A increases until around the offered load of 45 is reached, then decreases somewhat abruptly, and starts to increase slowly. The reason why it decreases in the middle is that the role of the piggyback rate-request becomes effective as the number of packet transmissions within a frame increases. However, as the offered load increases further, the number of transmitted packets in a frame also increases, then the total number of mobiles which transmit in a frame will decrease. So, the chance for a mobile to request through a piggyback rate-request mini-slot decreases, so the request access delay increases. The delay eventually goes to infinity at the marginal load. A similar phenomenon was also reported in [15]. It is observed that this behavior of the request access delay showing a peak is almost negligible for the case with a large N_{rcvr}^{rqst}, e.g., $N_{rcvr}^{rqst} = 25$ in the figure, since there will rarely be request collisions in this case.

From the above figure, we observed that the larger N_{rcvr}^{rqst}, the better in terms of the request access delay. However, the physically important measure is the message transmission delay, not the request access delay. We only considered the latter to understand the system better. Figure 8 shows the average message transmission delay vs. the offered load from the same simulation as in the previous figure. We observe a similar tendency and dependency on the value of N_{rcvr}^{rqst} to those from the previous figure. Note that the average delay should be at least 2.5 because the average request access delay is at least 0.5, the queueing delay is at least 1, and the transmission delay is also at least 1. We observe that especially for class II-A in the moderately-loaded region (i.e., offered load < 100), the message-transmission delay appears as the request access delay shifted by 2 frames. This is because there will be no more than one queueing delay for class II-A in this case. Note that for class II-B in this region, the message-transmission delay is more than a shifted version of the request access delay because class II-B gets less chance to be transmitted.

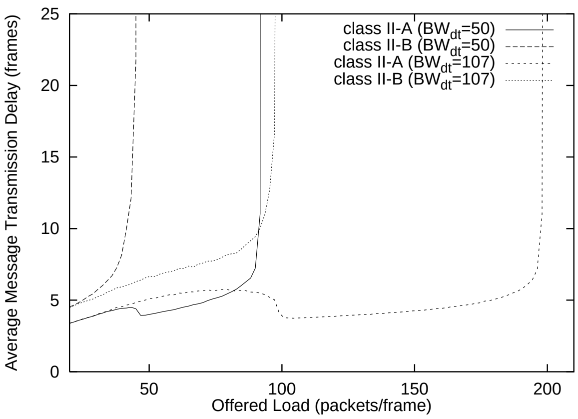


Figure 9: Average message transmission delay vs. offered load.

Next, we examine the effect of different BW_{dt} values. Figure 9 plots the average delay as the offered load increases for $BW_{dt} = 50$ and 107 with $N_{rcvr}^{rqst} = 20$. $BW_{dt} = 107$ corresponds to no class-I connections, and $BW_{dt} = 50$ corresponds to about a half of the total bandwidth is used by class-I connections, e.g., 94 voice mobiles. We see that for $BW_{dt} = 50$, the delay of class II-B becomes infinite at around 50 of offered load and that of class II-A becomes infinite around 100 of the offered load, similar to the request access delay cases. For $BW_{dt} = 107$, the offered loads at which the delays become infinite are observed to be almost doubled, because the bandwidth for class-II traffic is doubled.

7 Conclusion

In this paper, we studied a CDMA uplink system to provide diverse QoS guarantees for heterogeneous traffic, where the MC-CDMA is used to support the variable bit rates. To satisfy the pre-defined BER QoS metrics for each traffic class, we used a concatenated RS/convolutional code and developed a new scheme for allocating mobiles power levels. Connection-oriented transmission is achieved by assigning mobile-oriented code channels for class-I traffic where each corresponding mobile needs to pass an admission test. We calculated an admission region for pairs of voice and video mobiles. For class-II traffic, best-effort transmissions are supported: whenever each class-II mobile has messages to transmit, the mobile requests code channels via a transmission-request code channel, then the transmission is granted according to the underlying BS scheduling policy. We derived the target SIR by maximizing the aggregate throughput of class-II traffic and evaluated class-II transmission performance.

We are planning to extend our system to include RAKE receivers in order to combat multipath fadings more efficiently. The performance of this extension under various multipath fading channel environments will be studied. The effect of inter-cell interferences will also be studied. Considering these, the performance values reported in this paper will change, but the general performance trends are expected to remain the same. We observed these trends by comparing the performances of our system without (in [4]) and with (in this paper) considering the fading environment. We are also planning to extend our design to support a measurement-based admission control (e.g., [13]) of class-I connections, because it is not generally possible to have *a priori* information on class-I traffic. This admission control will have such a form as: (1) a connection is admitted based on its peak rate, (2) its time-varying transmission rate is observed during runtime, then (3) its equivalent bandwidth is adjusted accordingly.

Acknowledgment

The authors would like to thank Professor Wayne Stark at the University of Michigan for proofreading and finding an error in Section 4 in an earlier version of the paper.

References

[1] S. Aridhi and C. L. Despins, "Performance analysis of type-I and type-II hybrid ARQ protocols using concatenated codes in a DS-CDMA Rayleigh fading channel," in *Proc. IEEE ICUPC'95*, pp. 748–752, 1995.

[2] P. T. Brady, "A statistical analysis of on-off patterns in 16 conversations," *Bell Syst. Tech. J.*, vol. 47, pp. 73–91, January 1968.

[3] J. I. Capetanakis, "Tree algorithms for packet broadcast channels," *IEEE Trans. on Information Theory*, vol. IT-25, pp. 505–515, September 1979.

[4] S. Choi and K. G. Shin, "Uplink CDMA systems with diverse QoS guarantees for heterogeneous traffic," in *Proc. ACM/IEEE MobiCom'97*, pp. 120–130, Budapest, Hungary, September 1997.

[5] S. Choi and K. G. Shin, "A cellular wireless local area network with QoS guarantees for heterogeneous traffic," *ACM Mobile Networks and Applications (MONET)*, vol. 3, no. 1, pp. 89–100, 1998. special issue on Mobile Multimedia Communications.

[6] R. D. Cideciyan and E. Eleftheriou, "Concatenated Reed-Solomon/convolutional coding scheme for data transmission in CDMA cellular systems," in *Proc. IEEE VTC'94*, pp. 1369–1373, 1994.

[7] J. Conan, "The weight spectra of some short low-rate convolutional codes," *IEEE Trans. on Communications*, vol. 32, pp. 1050–1053, September 1984.

[8] 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 Journal on Selected Areas in Communications*, vol. 40, no. 2, pp. 303–312, May 1991.

[9] N. Guo, S. D. Morgera, and P. Mermelstein, "Common packet data channel (CPDC) for integrated wireless DS-CDMA networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 4, pp. 735–749, May 1996.

[10] C.-L. I and K. K. Sabnani, "Variable spreading gain CDMA with adaptive control for true packet switching wireless network," in *Proc. IEEE ICC'95*, pp. 725–730, June 1995.

[11] C.-L. I and R. D. Gitlin, "Multi-code CDMA wireless personal communications networks," in *Proc. IEEE ICC'95*, pp. 1060–1064, June 1995.

[12] C.-L. I, P. Pollini, L. Ozarow, and R. D. Gitlin, "Performance of multi-code CDMA wireless personal communications networks," in *Proc. IEEE VTC'95*, pp. 907–911, July 1995.

[13] S. Jamin, P. Danzig, S. Shenker, and L. Zhang, "Measurement-based admission control algorithm for integrated service packet networks," *IEEE/ACM Trans. on Networking*, vol. 5, no. 1, pp. 56–70, February 1997.

[14] M. J. Karol, Z. Liu, and K. Y. Eng, "Distributed-queueing request update multiple access (DQRUMA) for wireless packet (ATM) networks," in *Proc. IEEE ICC'95*, pp. 1224–1231, 1995.

[15] Z. Liu, M. J. Karol, M. El 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, 1996.

[16] R. J. McEliece and L. Swanson, "On the decoder error probability for Reed-Solomon codes," *IEEE Trans. on Information Theory*, vol. IT-32, no. 5, pp. 701–703, September 1986.

[17] T. Ottosson and A. Svensson, "Multi-rate schemes in DS/CDMA systems," in *Proc. IEEE VTC'95*, pp. 1006–1010, 1995.

[18] K. Pahlavan and A. H. Levesque, *Wireless Information Networks*, Wiley-Interscience, New York, NY, 1995.

[19] J. G. Proakis, *Digital Communications, 2nd ed.*, McGraw Hill, New York, NY, 1989.

[20] M. B. Pursley, "Performance evaluation for phase-coded spread-spectrum multiple-access communication – part I: system analysis," *IEEE Trans. on Communications*, vol. 25, pp. 795–799, August 1977.

[21] D. Raychaudhuri *et al.*, "WATMnet: A prototype wireless ATM system for multimedia personal communication," *IEEE Journal on Selected Areas in Communications*, pp. 83–95, January 1997.

[22] K. Rijkse, "H.263: Video coding for low-bit-rate communication," *IEEE Communications Magazine*, pp. 42–45, December 1996.

[23] M. Soroushnejad and E. Geraniotis, "Multi-access strategies for an integrated voice/data CDMA packet radio network," *IEEE Trans. on Communications*, vol. 43, no. 2/3/4, pp. 934–945, February/March/April 1995.

[24] A. J. Viterbi, *CDMA: Principles of Spread Spectrum Communication*, Addison-Wesley, Reading, MA, 1995.

Sunghyun Choi (S'96/ACM S'98) is a Ph.D. candidate at the Department of Electrical Engineering and Computer Science in the University of Michigan, Ann Arbor. He received his B.S. (summa cum laude) and M.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST) in 1992 and 1994, respectively. His research interests are in the area of wireless/mobile networks with emphasis on the QoS guarantee and adaptation, connection and mobility management, adaptive error control, multi-media CDMA, and wireless MAC protocols.

He is a member of IEEE Communication Society, ACM SIGMOBILE, and ACM SIGCOMM. Since 1997, he has been a recipient of the Korea Foundation for Advanced Studies Scholarship. During 1994-1997, he received the Korean Government Overseas Scholarship. He is also a winner of the Humantech Thesis Prize from Samsung Electronics in 1997. His email address is: shchoi@eecs.umich.edu.

Kang G. Shin (S'75–M'78–SM'83–F'92) received the B.S. degree in Electronics Engineering from Seoul National University, Seoul, Korea in 1970, and both the M.S. and Ph.D degrees in Electrical Engineering from Cornell University, Ithaca, New York in 1976 and 1978, respectively.

He is currently a Professor and the Director of the Real-Time Computing Laboratory, Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor. He has authored and coauthored approximately 600 technical papers and numerous book chapters in the areas of distributed real-time computing and control, computer networking, fault-tolerant computing, and intelligent manufacturing. He has also co-authored (jointly with C. M. Krishna) a textbook "Real-Time Systems," McGraw Hill, 1997.

His current research focuses on Quality of Service (QoS) sensitive computing and networking with emphases on timeliness and dependability. He has also been applying the basic research results to telecommunication and multimedia systems, intelligent transportation systems, embedded systems, and manufacturing applications. Dr. Shin was a Distinguished Visitor of the Computer Society of the IEEE, an Editor of *IEEE Trans. on Parallel and Distributed Computing*, and an Area Editor of *International Journal of Time-Critical Computing Systems*.