Resource Allocation for Multiple Classes of DS-CDMA Traffic

Joon Bae Kim, *Student Member, IEEE,* and Michael L. Honig, *Fellow, IEEE*

***Abstract*—We consider a packet data direct-sequence code-division multiple-access (DS-CDMA) system which supports integrated services. The services are partitioned into different traffic classes according to information rate (bandwidth) and quality of service (QoS) requirements. Given sufficient bandwidth, QoS requirements can be satisfied by an appropriate assignment of transmitted power and processing gain to users in each class. The effect of this assignment is analyzed for both a single class of data users and two classes of voice and data users. For a single class of data users, we examine the relationship between average delay and processing gain, assuming that ARQ with forward error correction is used to guarantee reliability. The only channel impairment considered is interference, which is modeled as Gaussian noise. A fixed user population is assumed and two models for generation of data packets are considered: 1) each user generates a new packet as soon as the preceding packet is successfully delivered and 2) each user generates packets according to a Poisson process. In each case, the packets enter a buffer which is emptied at the symbol rate. For the second traffic model, lowering the processing gain below a threshold can produce multiple operating points, one of which corresponds to infinite delay. The choice of processing gain which minimizes average delay in that case is the smallest processing gain at which multiple operating points are avoided. Two classes of users (voice/data and two data classes) are then considered. Numerical examples are presented which illustrate the increase in the two-dimensional (2-D) capacity region achievable by optimizing the assignment of powers and processing gains to each class.**

***Index Terms*—DS-CDMA, integrated voice and data, power control, resource allocation.**

I. Introduction

FUTURE wireless services are likely to integrate different types of traffic, such as voice, data, image, and compressed video. Because these traffic streams have different requirements on information rate and performance, allocation of scarce resources, such as power and bandwidth, among the wireless users can have a significant effect on system capacity. In this paper, we consider a particular class of resource allocation problems associated with direct-sequence code-division multiple access (DS-CDMA).

There are many ways in which the information rate and performance associated with different traffic streams can be controlled in DS-CDMA. For example, a particular information rate can be achieved through an appropriate choice of chip rate [1], processing gain [2], number of codes [3], and modulation format [4]. Here, we assume that the chip rate, number of spreading codes, and modulation format are fixed for all users. The information rate is then determined by the selection of processing gain. By fixing the chip rate, all sources are spread across the entire available bandwidth, which enhances the advantages of spread-spectrum signaling/CDMA (i.e., resistance to fading and interference, and increased trunking efficiency).

The quality of service (QoS) in DS-CDMA can be controlled by an appropriate selection of transmitted power and processing gain. The idea of assigning different transmitted powers to achieve different QoS's in DS-CDMA was proposed in [5]. Here, we also consider how the choice of processing gain affects QoS. Specifically, the processing gain is varied by changing the symbol duration. Doubling the processing gain doubles the symbol duration (i.e., halves the symbol rate). For data, decreasing the symbol rate can actually increase throughput. This is because increasing the processing gain increases the received signal-to-interference plus noise-ratio (SINR) and therefore decreases the probability of retransmission. Other ways to vary QoS, which we do not consider here, are to use different coding/decoding and detection schemes that vary in complexity.

We assume that each user generates a sequence of fixed-length packets, and reliability of data communications is guaranteed through error detection and retransmission (ARQ). The model considered along with some of the results reported here were presented in [6] and [7]. User traffic is partitioned into different classes (e.g., voice, low-priority data, and high-priority data), each of which requires a particular information rate and QoS. Traffic characteristics (i.e., packet generation rates and packet lengths) and system parameters, such as desired QoS, processing gain, and modulation format, are the same for all users in each class. However, these may differ from class to class.

The general problem we wish to address is how to assign powers and processing gains to different classes of traffic so as to maximize capacity (number of users that can be supported), given rate and QoS constraints. An equivalent problem is to maximize information rate (or the information rate region) given a fixed number of users in each class. The performance (QoS) measures we consider for data traffic are average delay and the probability that the delay exceeds a given threshold. For voice traffic, the bit error rate is constrained to be below a target threshold.

We start by considering a single class of data users and studying the preceding performance measures as a function of processing gain, assuming that the received powers associated

Manuscript received August 6, 1997; revised January 25, 1999. This work was supported by the U.S. Army Research Office under Grant DAAH04-96-1-0378 and the National Science Foundation under Grant NCR-9628365.

The authors are with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208-3118 USA (e-mail: jbkim@ece.nwu.edu; mh@ece.nwu.edu).

Publisher Item Identifier S 0018-9545(00)02564-0.

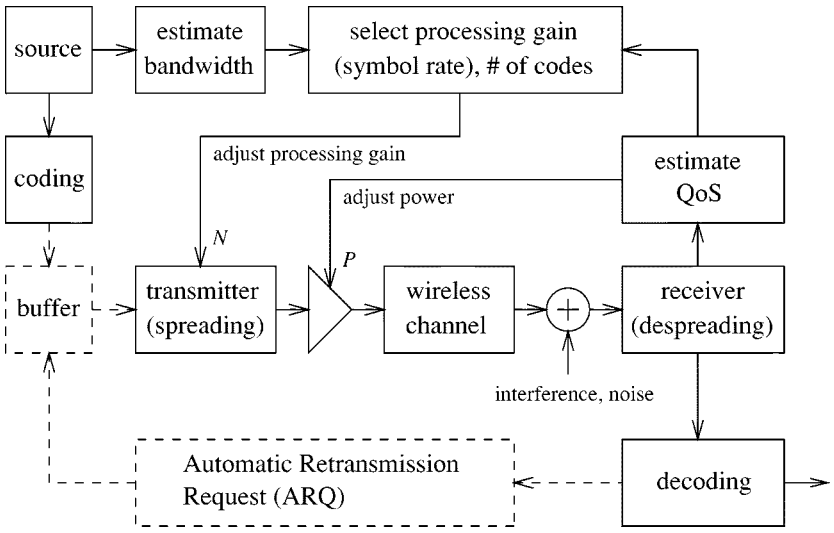

Fig. 1. Parameter selections for multirate/multi-QoS DS-CDMA.

with all users is constant for all users (i.e., perfect power control). The user population is fixed, and we consider two traffic models: 1) each user generates a new packet as soon as the previous packet has been successfully delivered and 2) each user generates packets according to a Poisson process. In the latter case, the packets enter a buffer, which is emptied at the symbol rate. The choice of processing gain therefore affects the activity factor (expected busy period) for each user, which in turn affects the SINR for all users. It is shown that lowering the processing gain below a threshold can produce multiple operating points, one of which corresponds to infinite delay. The choice of processing gain which minimizes average delay in this case is the smallest processing gain at which multiple operating points are avoided.

We then consider two classes of users and solve for the "capacity region," subject to constraints on the QoS for each class.[1] A "QoS region" for a fixed number of users is also characterized, which shows the maximum achievable QoS for users in one class given a QoS for users in the other class.

In the next section, we describe our approach to providing multirate/multi-QoS services with DS-CDMA. In Section III, we present the traffic and system model, and in Section IV, we analyze a single class of data users. Sections V and VI then presents results for two classes of voice and data users, and two different classes of data users.

II. Multirate/Multi-QoS DS-CDMA

In this section, we present our approach to providing multirate services with different QoS requirements in the context of DS-CDMA. We focus on the reverse link of a single cell, in which multiple asynchronous mobile subscribers transmit to a single base station. The chip rate (or equivalently, the bandwidth) is assumed to be the same for all sources, and the information rate is determined by selecting the processing gain (number of chips per symbol). In this way, all sources are multiplexed onto the same wideband channel, which maximizes trunking efficiency and frequency diversity.

A block diagram which illustrates the selection of parameters used to control the information rate and QoS is shown in Fig. 1. Each source generates a sequence of fixed-length packets of length L symbols, where L depends on the source. The packets generated by each source enter a buffer after error control coding. The buffer contents are then converted to a DS-CDMA signal at the symbol rate R_c/N, where R_c is the chip rate and N is the processing gain, which is being varied. Since the chip rate (spread bandwidth) is fixed, the processing gain (number of chips per symbol) determines the *symbol duration*. If the packet arrival rate is λ, then the processing gain should be no larger than $R_c/(\lambda L)$. Otherwise, the rate at which packets arrive to the buffer exceeds the rate at which the buffer is emptied (even without retransmissions).

After despreading and decoding the received packet, the receiver (base station) may request the transmitter (user) to retransmit the packet if it contains errors (dashed part of Fig. 1). This is necessary for error-sensitive (data) applications, but may be undesirable for voice due to the additional associated delay. The QoS, which may be specified in terms of error rate, throughput, and/or delay can be controlled through the selection of processing gain N, transmitted power P, and the number of spreading codes (as a form of diversity). Here, we will assume that each user is assigned only one spreading code. For example, the choice of N significantly affects packet delay as well as the information rate. That is, reducing N (which decreases the symbol duration) reduces the packet transmission time, but also decreases the SINR, which increases the probability of retransmission. Therefore, selection of N controls the tradeoff between transmission time and retransmission rate. Of course, the SINR can be improved by increasing the transmitted power P, but this increases interference to other users. Here, we examine performance (e.g., error rate and average delay) as a function of both N and P.

Depending on source characteristics, the choice of processing gain for a particular user (or traffic class) can also affect the

[1]The capacity region shows the maximum number of users in one class as a function of the number of users in the other class. This capacity region is analogous to, but not the same as, the capacity region in multiuser Shannon theory.

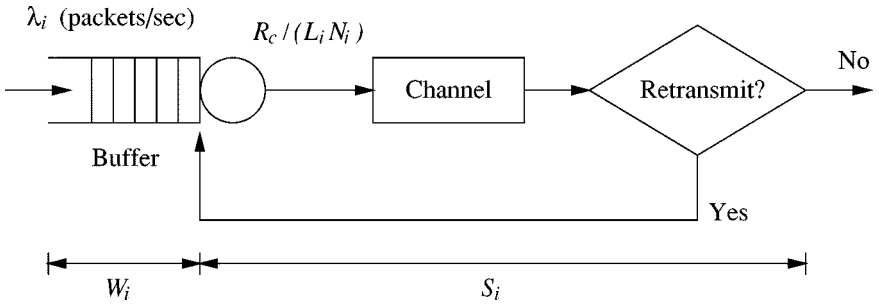

Fig. 2. Data traffic model: W_i is the average waiting time in the queue and S_i is the average packet "transfer" time, including retransmissions.

QoS of *other* types of traffic. For example, packet voice traffic with a high chip rate and low processing gain appears as bursty interference to other sources since packets are transmitted with large gaps between them. As the processing gain increases, the duration of the voice packets increases, and the intervening gaps decrease. Benefits associated with this type of traffic "smoothing" are: 1) it reduces the transmitted peak-to-average power; 2) it reduces peak-to-average interference power (associated with other voice users); and 3) it facilitates adaptive interference suppression, since the interference does not vary as rapidly, and therefore is easier to track [8]. The primary disadvantage of this smoothing is that the fraction of time the source is transmitting (activity factor) increases.

Here, we examine how the choice of processing gain (and hence symbol duration) for voice users affects the performance of data users. Assuming only additive Gaussian noise and multiple-access interference, our results indicate that the delay for data users decreases as the processing gain assigned to voice users increases. This is in addition to the preceding benefits associated with traffic smoothing.

A. Related Work

Providing for multirate/multi-QoS services within the context of DS-CDMA has started to attract considerable attention. Here, we give a very brief overview of recent work related to that presented here. In addition to [5], mentioned earlier, [9]–[11] also consider the use of power control to satisfy QoS (SINR) requirements. In [9] and [10], the power allocation which maximizes system capacity and minimizes power consumption (summed over all users) is derived, and a similar problem is considered in [11], where total system throughput is maximized. In [12], two different power control strategies, equalizing received signal powers and equalizing packet error rates, are compared for voice/data traffic with fading channels. Dynamic assignment of processing gain in a time-slotted CDMA model is studied in [13].

Additional work on access protocols for integrating voice and data traffic in DS-CDMA is presented in [14], [15], and [16]. This latter work, in addition to [11] and [17], models all data traffic as a single stochastic arrival process.

The model we consider in this paper differs from the models considered in these other references in one or more of the following ways: 1) the data traffic model accounts for both burstiness and retransmissions; 2) throughput and delay is studied as a function of both power and processing gain; 3) delay includes the additional time caused by retransmissions and queueing; and 4) performance and rate regions for two classes of traffic (analogous to voice and data) are computed. We also remark that due to the different modeling assumptions, our results appear to be quite different from results presented in the preceding references.

III. System Model and Performance Measures

We now specify a system model for the DS-CDMA reverse link based on the multirate/multi-QoS approach described in the last section. We assume that there are C classes of asynchronous users within a single cell and that there is a fixed number of users K_i in the ith class where $1 \leq i \leq C$. The different classes may represent different traffic types, or the same traffic type, but with different QoS requirements.

A. Traffic Models

Fig. 2 shows a block diagram which illustrates the data traffic models considered. Each user generates a sequence of fixed-length packets which enter an infinite-length buffer. The buffer is emptied at the (packet) rate $R_c/(L_iN_i)$ packets per second, where L_i is the packet length for traffic in class i. Two models for packet generation are considered. In the first, a new packet is generated as soon as the preceding packet is successfully delivered. The number of *active* users in the system is therefore constant. This model will be referred to as the "continuously active" model, since the users continuously transmit packets.

In the second model for packet generation, each user in class i generates a sequence of fixed-length packets according to a Poisson process with rate λ_i. The packets may therefore experience a queueing delay in addition to retransmission delays. As shown in Fig. 2, for both models each packet is retransmitted until it is received without errors.

The continuously active model is easier to analyze than the Poisson model, and corresponds to the situation in which all users are transmitting long files. (An alternative interpretation is that a flow control protocol is used which ensures that the buffers always contain packets to transmit.) The Poisson model is more appropriate for traffic which consists of short bursts of data.

For each model of data traffic, voice traffic is assumed to be generated in an analogous fashion. However, voice packets with errors are not retransmitted. The steady-state analysis of the second traffic model which follows depends only on the first-order statistics of the number of active voice users. Consequently, the arrival process for voice packets need not

be Poisson. (It may be more convenient to assume periodic arrivals of voice packets, as generated by a vocoder. In that case, no additional buffer for voice packets is required.)

B. SINR and Delay

Assuming matched-filter detection, the QoS for a user in class i is a function of the SINR, which is defined in terms of the bit energy-to-noise density ratio [18]

$$\text{SINR}_i = \frac{P_i N_i}{\alpha\left(\sum_{k=1}^{K_i-1}\chi_{k_i}g_{k_i}P_i + \sum_{j\neq i}^{C}\sum_{k=1}^{K_j}\chi_{k_j}g_{k_j}P_j\right)+\sigma^2} \tag{1}$$

where P_i is the transmitted power, N_i is the processing gain, α is a constant which depends on the shape of the DS-CDMA chips, χ_{k_i} is an on/off indicator (i.e., χ_{k_i} is one if user k_i is active and zero if inactive), g_{k_i} is the attenuation from user k_i to the base station relative to the desired signal, and σ^2 is the background noise power which includes other cell interference. The subscript k_i denotes the kth user in class i. In this paper, we assume that $\alpha = 2/3$, which corresponds to rectangular chips [19], [20]. We assume perfect power control, which is used to compensate for channel attenuation. This implies $g_{k_i} = 1$ for all users in the cell, so that P_j in the expression for SINR_i becomes the *received* power at the base station associated with all users in class j.

The on/off indicators are assumed to be independent from user to user, and identically distributed within a class

$$\Pr\{\chi_{k_i}=1\} = p_{\text{on}_i} \quad \Pr\{\chi_{k_i}=0\} = 1-p_{\text{on}_i} \tag{2}$$

where p_{on_i}, a user's on/off probability (activity factor), denotes the probability that a user in class i is active at any time instant. The distribution for χ_{k_i} is assumed to be the same for each transmitted symbol. In other words, we implicitly assume a large interleaving depth which eliminates symbol-to-symbol correlations in steady-state activity probabilities. In practice, finite interleaving depths (e.g., constrained to single packet) will compromise the accuracy of this assumption. However, as the number of users increases, this assumption should become more accurate since the number of active users changes more rapidly within the interleaving window. Under this independence assumption, SINR_i is a random variable with distribution which depends on the on/off probabilities $p_{\text{on}_i}, i = 1, \cdots, C$.

Let λ_i denote the rate at which a source in class i generates packets, and let S_i denote the average time it takes to successfully transmit a class i packet including retransmissions (average "transfer" time shown in Fig. 2). Little's rule [21] implies that

$$p_{\text{on}_i} = \begin{cases} \lambda_i S_i, & \text{if } \lambda_i S_i < 1 \\ 1, & \text{otherwise.} \end{cases} \tag{3}$$

Note that $\lambda_i S_i \geq 1$ implies instability in the sense that the average waiting time in the queue is infinite.

For voice traffic, the packet transfer time S_i is equal to the packet transmission time T_i (assuming negligible propagation and processing delays). Since the packets are transmitted at the symbol rate R_c/N_i, we have that

$$T_i = \frac{L_i N_i}{R_c}. \tag{4}$$

The stability condition $\lambda_i S_i < 1$ implies that the processing gain $N_i \leq N_{\max_i} = \lfloor R_c/(\lambda_i L_i)\rfloor$ where $\lfloor x\rfloor$ denotes the largest integer less than or equal to x.

For data traffic, S_i can be greater than T_i because of possible retransmissions. To simplify the analysis, we assume that (negative and positive) acknowledgments are immediate and are perfectly reliable. This assumption is reasonable provided that the channel propagation time is very small relative to the packet length, and acknowledgments are coded [22]. Let Δ_i denote the transfer time of a single packet for a user in class i. Then Δ_i has the geometric distribution

$$\Pr\{\Delta_i = jT_i\} = p_{r_i}^{j-1}(1-p_{r_i}) \tag{5}$$

where $j = 1, 2, \cdots$ is the number of transmissions necessary for successful reception and p_{r_i} is the probability of retransmission associated with class i. The average packet transfer time is

$$S_i = \text{E}\{\Delta_i\} = \frac{T_i}{1-p_{r_i}} = \frac{L_i N_i}{R_c(1-p_{r_i})}. \tag{6}$$

In addition to S_i, we also consider the "overlimit" probability Q_i, which is the probability that the packet transfer time Δ_i exceeds a given threshold δ_i, the maximum acceptable packet transfer time. From (5) we have that

$$Q_i = \Pr\{\Delta_i > \delta_i\} = p_{r_i}^m, \qquad \text{where } m = \lfloor \delta_i/T_i\rfloor. \tag{7}$$

For data traffic, there are two sources of delay, namely, retransmissions and queueing delay. (Note that the delay for voice traffic without retransmissions is simply T_i.) The average waiting time in the queue W_i can be calculated from the standard M/G/1 queueing model with an infinite-length buffer [21]. From (5)

$$W_i = \frac{\lambda_i \text{E}\{\Delta_i^2\}}{2(1-\lambda_i S_i)} = \frac{\lambda_i T_i^2(1+p_{r_i})}{2(1-\lambda_i S_i)(1-p_{r_i})^2}. \tag{8}$$

The average packet delay D_i for a user in class i is therefore

$$D_i = W_i + S_i = \frac{T_i(2-\lambda_i T_i)}{2(1-p_{r_i}-\lambda_i T_i)} \tag{9}$$

assuming stability $(\lambda_i S_i < 1)$. Note that we can accommodate a higher rate source (higher throughput) by decreasing D_i.

C. Error Probability

For both voice and data traffic, channel coding includes forward error correction (FEC). For data traffic, error detection coding is needed for ARQ. Instead of defining a specific modulation and FEC scheme, we will assume that the bit error probability (BEP) is an exponentially decaying function of the SINR. This corresponds to the common assumption that the interference plus noise is Gaussian [18] combined with the upper bound (used here as an approximation)

$1/\sqrt{2\pi}\int_x^\infty e^{-y^2/2}\,dy < 1/\sqrt{4\pi}e^{-x^2/2}$. Specifically, we assume that for a user in class i

$$p_{b_i} = \mathrm{E}_{\mathrm{SINR}_i}\{\mathcal{F}(\mathrm{SINR}_i)\}, \quad \text{where } \mathcal{F}(\mathrm{SINR}_i) = \kappa\exp(-\beta\mathrm{SINR}_i) \tag{10}$$

and κ and β are parameters which can be adjusted to match a particular coding scheme. The asymptotic coding gain of the error correcting code is determined by β.

We remark that the particular choice of function $\mathcal{F}(\cdot)$ does not influence our analytical approach. This choice simply enables us to obtain numerical results. By changing $\mathcal{F}(\cdot)$, we can, in principle, account for different types of channel impairments, such as fading.

Assuming that each user's on/off indicator is independent from symbol to symbol (perfect interleaving), the probability of retransmission p_{r_i} is

$$p_{r_i} = 1 - [1 - \mathrm{E}_{\mathrm{SINR}_i}\{\mathcal{F}(\mathrm{SINR}_i)\}]^{L_i r_i} \tag{11}$$

where r_i is the FEC code rate for class i traffic, and the expectation is with respect to the distribution for SINR_i. We assume that all errors are detected. For a given SINR_i, Fig. 3 illustrates the relationship between the approximation for p_{b_i} and SINR_i [i.e., $\mathcal{F}(\cdot)$]. (We are only interested in the curve at low error rates.) The dashed line corresponds to uncoded DPSK modulation, and the solid line shows coded BEP. The dashed–dotted line indicates the probability of retransmission when $L_i = 768$ and $r_i = 1/2$. The parameters $\kappa = 1/2$ and $\beta = 2$, which correspond to an asymptotic coding gain of 3 dB.

To satisfy the BEP requirement for data traffic, the probability of an undetected error must be no greater than the target BEP. Since we assume that all errors are detected, the BEP requirement for data traffic is automatically satisfied in our model, and we focus on satisfying a delay constraint (i.e., on average delay or overlimit probability). In contrast, for voice traffic we assume that the maximum delay, $D_i = L_i N_{\max_i}/R_c$, is acceptable and focus on satisfying a constraint on the average BEP p_{b_i}.

IV. Single Class of Data Users

We start by considering a single class of data users. Namely, there are K users, and each transmits with the same power and processing gain. The continuously active packet generation model is first analyzed, and is followed by an analysis of the Poisson model.

A. Continuously Active Users

Given K continuously active users ($p_{\mathrm{on}} = 1$), (1) implies

$$\mathrm{SINR} = \frac{PN}{\alpha(K-1)P + \sigma^2} = \xi N \tag{12}$$

where ξ denotes the slope of SINR versus processing gain N. The SINR is an insensitive function of P when the background noise level σ^2 is small.

For the continuously active model, the average packet delay is $D = S$, given by (6). The retransmission probability p_r is given by (11), and from (10), $p_b = \kappa\exp(-\beta\mathrm{SINR}(K))$ since $\mathrm{SINR}(K)$ is a deterministic function of K. We can therefore write

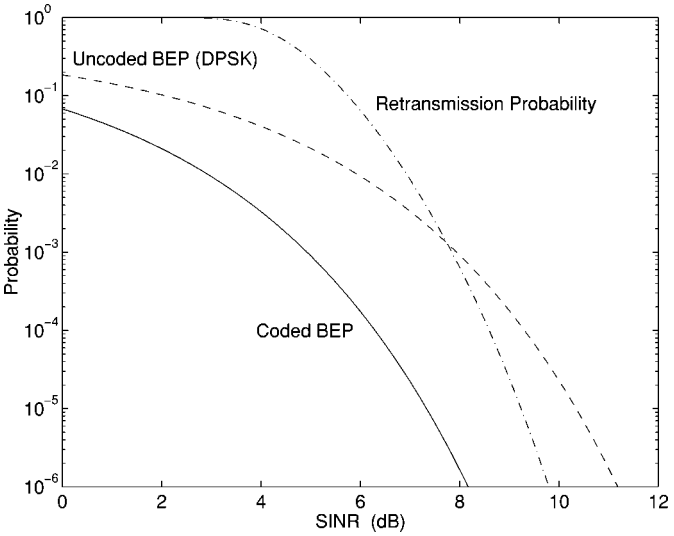

Fig. 3. Bit error probability versus SINR.

$$D = \frac{LN}{R_c(1-\kappa\exp(-\beta\xi N))^{Lr}}. \tag{13}$$

Average throughput η, defined as the average number of information bits (including overhead) successfully transferred per second, is given by

$$\eta = \frac{Lr}{D} = \frac{rR_c}{N}(1-\kappa\exp(-\beta\xi N))^{Lr}. \tag{14}$$

As N increases, both the BEP p_b, given by (10), and the retransmission probability p_r, given by (11), decrease. However, increasing N also increases the symbol duration, so that there is a tradeoff between the decrease in D caused by the decrease in p_r and the increase in D caused by the increase in T. The former dominates for small N, and the latter dominates for large N. The processing gain that minimizes average packet delay (or maximizes throughput) is easily computed by setting the derivative of (13) with respect to N equal to zero. The optimal processing gain N^* must satisfy

$$\kappa(1 + Lr\beta\xi N^*) = \exp(\beta\xi N^*). \tag{15}$$

There can be at most two solutions to (15), however, for system parameters of interest, one of these solutions is less than one and is therefore not relevant.

Fig. 4 shows a plot of average throughput η versus N for a system with 20 users. Fig. 5 shows a plot of the overlimit probability Q versus N where δ in (7) is arbitrarily taken to be 15 ms. For these plots, the packet length $L = 768$ symbols, and the code rate $r = 1/2$, which corresponds to 48 information bytes per packet (including error detection overhead such as CRC). A microcellular personal communication system is assumed [11], [22], where the chip rate R_c is 5 Mchips/s and $P/\sigma^2 = 10$ dB. Plots are shown for two different coding schemes corresponding to asymptotic coding gains of 3 dB ($\beta = 2$) and 6 dB ($\beta = 4$).

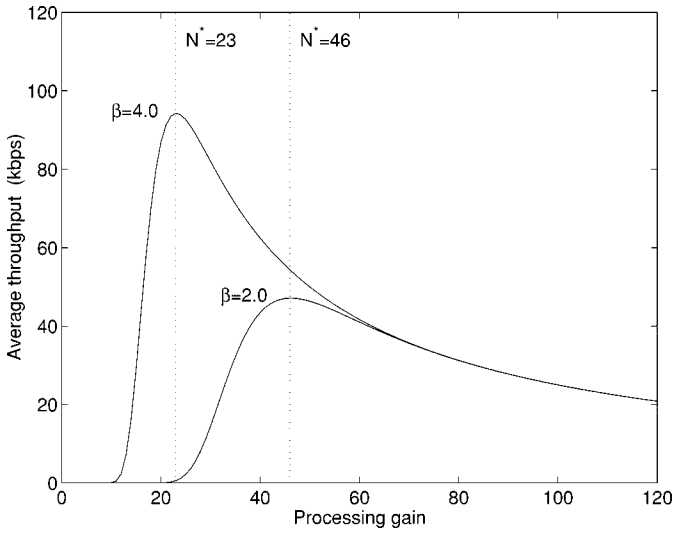

Fig. 4. Average throughput versus processing gain.

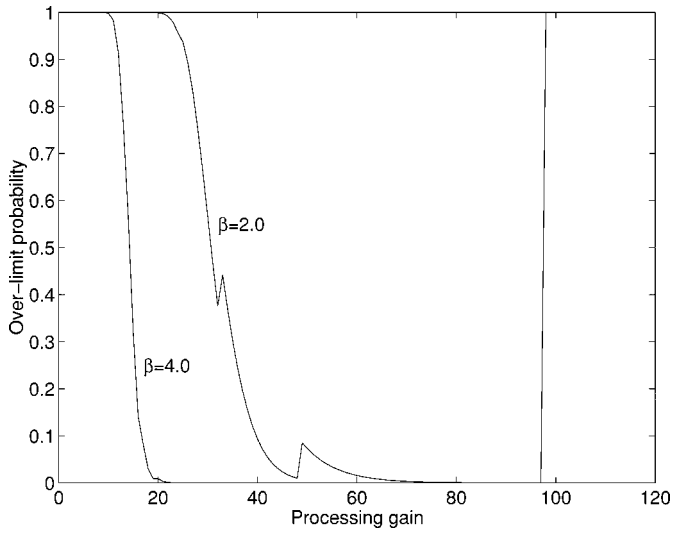

Fig. 5. Overlimit probability versus processing gain ($\delta = 15$ ms).

Fig. 4 shows that η decreases rapidly when N decreases below the optimal point N^*. This happens because p_r approaches one. For large N, η decreases slowly with N, since p_r is close to zero and $D \approx T$, which increases linearly with the symbol duration. The reduction in η as N recedes below N^* is more severe for the powerful coding scheme since a small change in SINR causes a more dramatic change in p_r. These results indicate that the choice of processing gain can significantly affect throughput (delay).

The discontinuity in Q shown in Fig. 5 is due to the fact that the packet transfer time Δ assumes discrete values, which are integer multiples of the packet transmission time T [see (5)]. Depending on the application, which determines δ, it is possible that N which minimizes D gives an unacceptably large Q. Fig. 5 suggests that N must then be increased so that the QoS constraint is satisfied, but with a corresponding reduction in throughput.

The optimality condition (15) implies that given the parameters L, r, κ, and β, the SINR ξN^* is uniquely determined. That is, when N is chosen to minimize average delay, the associated

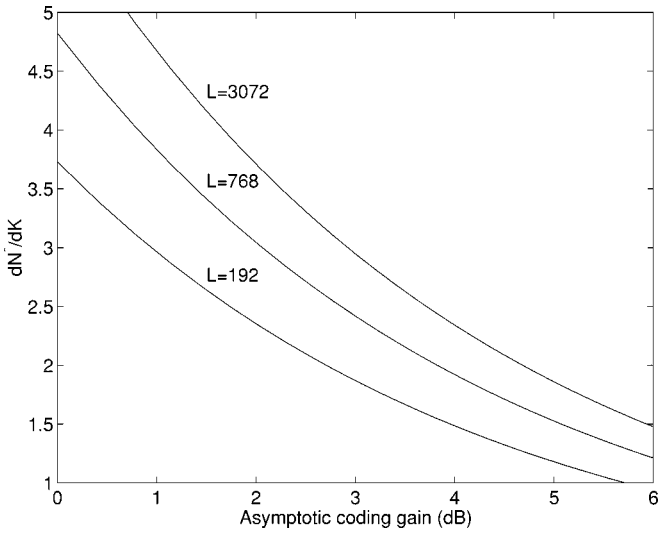

Fig. 6. dN^*/dK versus asymptotic coding gain $(10 \log \beta)$.

SINR is independent of the number of users K. Combining (12) and (15) gives

$$N^* = (\gamma^* \alpha)K + \gamma^* \left(\frac{\sigma^2}{P} - \alpha \right) \tag{16}$$

where $\gamma^* = \xi N^*$ is the solution to (15) for given L, r, κ, and β. We therefore conclude that the optimal processing gain increases linearly with the number of active users. Fig. 6 shows a plot of the slope $\gamma^* \alpha$ versus error exponent β for different packet lengths L. As the error exponent β increases, which implies an increase in coding gain, the slope decreases. That is, a more powerful code allows the use of a smaller processing gain.

B. Randomly Active Users

We now assume that each user in the cell generates packets according to a Poisson process with the same average rate λ. In steady state, each user is active (on) with probability p_{on}, and inactive (off) with probability $1 - p_{\text{on}}$. The number of active users in the system is therefore a random variable with distribution that depends on p_{on}. From (1), we have that

$$\text{SINR} = \frac{PN}{\alpha I P + \sigma^2}, \quad I = \sum_{k=1}^{K-1} \chi_k \tag{17}$$

where I denotes the number of active interferers. Since the χ_k's are assumed to be independent from user to user and have the same distribution within a packet, I has a binomial distribution

$$\begin{aligned} \Pr\{I = j\} &= \mathcal{B}(K-1, j, p_{\text{on}}) \\ &= \binom{K-1}{j} p_{\text{on}}^j (1 - p_{\text{on}})^{K-1-j} \end{aligned} \tag{18}$$

where $j = 0, 1, \cdots, K-1$ and $0 \le p_{\text{on}} \le 1$. From (10) and (11), the retransmission probability p_r can be written as

$$p_r = 1 - \left[1 - \sum_{j=0}^{K-1} \mathcal{F}\left(\frac{PN}{\alpha j P + \sigma^2} \right) \mathcal{B}(K-1, j, p_{\text{on}}) \right]^{Lr}. \tag{19}$$

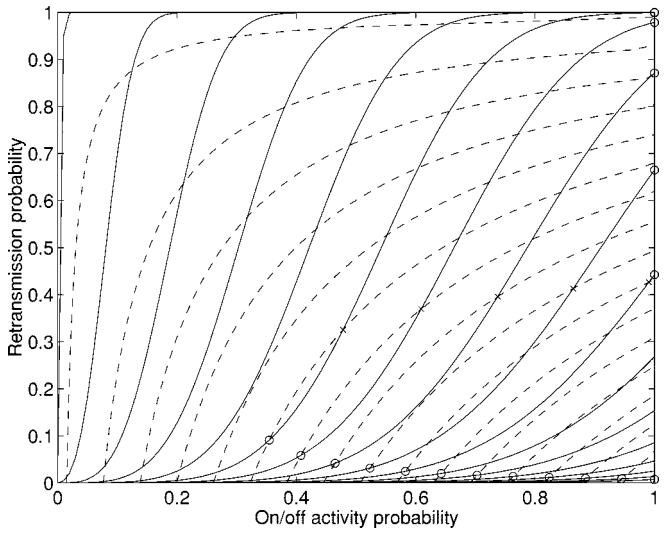

Fig. 7. Plots of p_r versus $p_{\rm on}$, according to (19) and (20), for different values of N (intersection points are marked.).

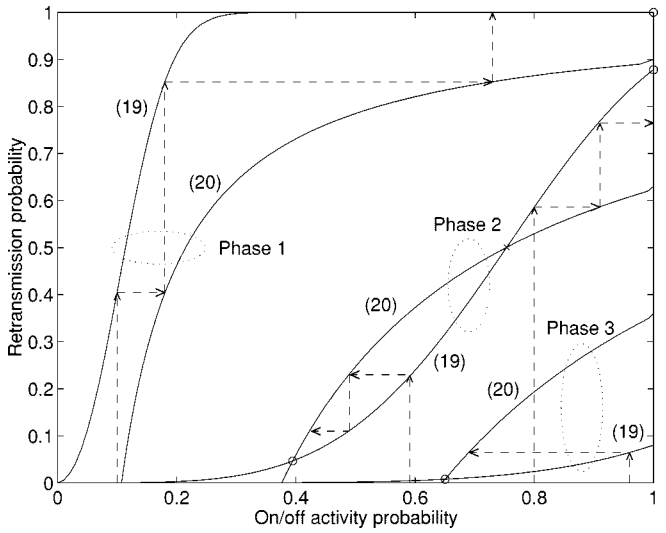

Fig. 8. Illustration of solutions to (19) and (20).

Note that p_r is expressed as a function of $p_{\rm on}$. Combining (3) and (6), $p_{\rm on}$ can also be expressed as a function of p_r

$$p_{\rm on} = \begin{cases} \dfrac{\lambda LN}{R_c(1-p_r)}, & \text{if } p_r < 1 - \dfrac{\lambda LN}{R_c} \\ 1, & \text{otherwise.} \end{cases} \tag{20}$$

We now have two equations [(19) and (20)] with two unknowns, p_r and $p_{\rm on}$. We would like to determine all values of p_r and $p_{\rm on}$ that satisfy these two equations. Fig. 7 shows plots of curves corresponding to (19) and (20) with the same parameters used to generate the plots in the preceding section, and $\lambda = 100$ packets/s. The different curves correspond to different values of N. The solid lines correspond to (19) and the dashed lines correspond to (20).

We now make some observations concerning (19) and (20). For both relations p_r is an increasing function of $p_{\rm on}$. Furthermore, Fig. 7 shows that there is at most one break-point in (19) for this particular set of parameters. That is, its second derivative has at most one zero-crossing point. It is easily seen from (20) that $p_{\rm on}$ is a convex function of p_r for $0 \le p_r < 1 - \lambda LN/R_c$. Note also that λ only affects (20).

The preceding observations imply that the curves corresponding to (19) and (20) can intersect in at most three points. For the examples considered, we have observed only cases with one and three intersections, which are illustrated in Fig. 8. (Two intersection points can occur, but only for a specific value of N, which generally is not an integer.) The cases observed, which are illustrated in Fig. 8, will be referred to as "phases" associated with a particular system state.

Referring to Fig. 8, when the system is in Phase 1 there is a single intersection point at $p_{\rm on} = 1$. The average packet delay in this case is infinite since the packet generation rate exceeds the rate at which packets are successfully transmitted. In Phase 2, there are two intersection points for which $0 < p_{\rm on} < 1$ and one for which $p_{\rm on} = 1$. The middle intersection point marked by $\times$ is "unstable" in the sense that any short-term perturbation in λ will push the system to the left or right equilibrium points. The left and right intersection points, marked by $\circ$, are "stable"

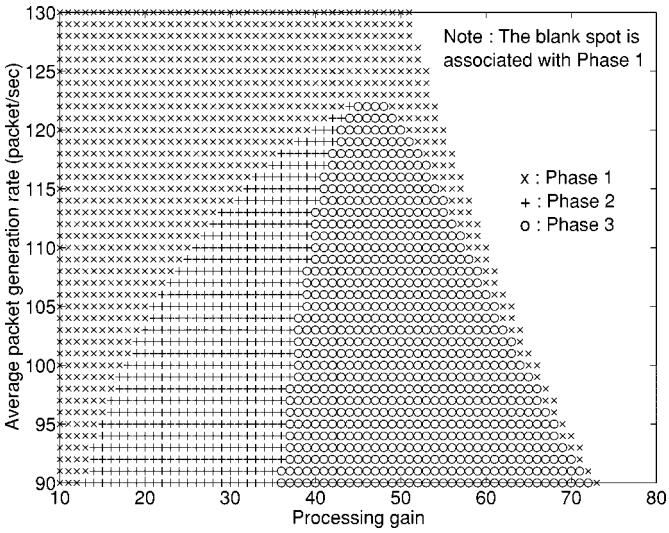

Fig. 9. Phase space diagram.

in the sense that a slight perturbation will not move the system to another equilibrium state. The right point, corresponding to $p_{\rm on} = 1$, is associated with infinite delay.

In Phase 3, there is a single stable operating point for which $0 < p_{\rm on} < 1$. Only Phase 3 corresponds to a desirable operating region since Phase 1 and Phase 2 have a stable equilibrium associated with infinite delay.

For a fixed processing gain N, as λ decreases, the curve corresponding to (20) rises so that there are transitions from Phase 1 to Phase 2 to Phase 3. (A direct transition from Phase 1 to Phase 3 can also occur since the processing gain is discrete.) If λ is fixed, then both curves corresponding to (19) and (20) move to the right as the processing gain increases. The corresponding phase transitions are difficult to predict. However, it is generally observed that transitions are made from Phase 1 to Phase 2 to Phase 3 and back to Phase 1 as N increases. Note also that Phase 1 corresponds to both very small and very large N. Specifically, Phase 1 is associated with $N \ge N_{\max} = \lfloor R_c/(\lambda L) \rfloor$.

The preceding phase transitions are illustrated in Fig. 9 which labels the "phase space" in terms of λ and N. The numerical

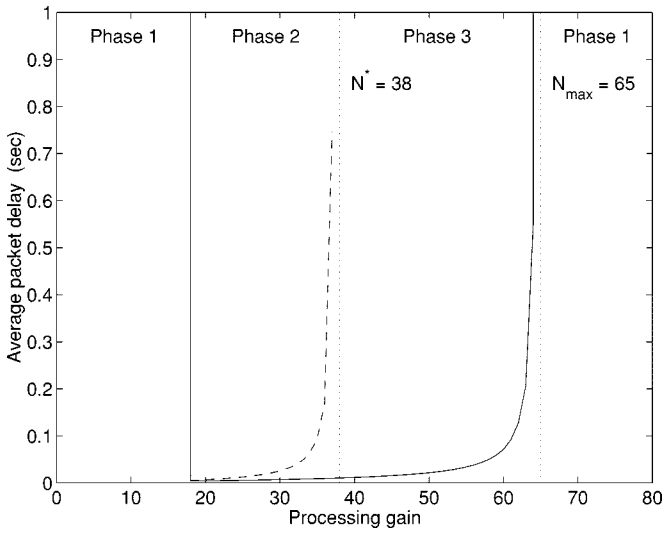

Fig. 10. Average packet delay versus processing gain ($\lambda = 100$ packets/s).

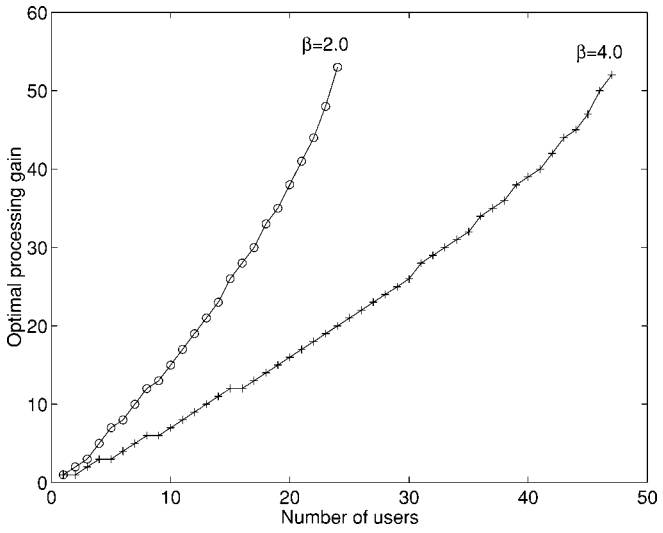

Fig. 11. Optimal processing gain versus number of users.

values of system parameters are the same as in the preceding section. The results are insensitive to the received power P, provided that P/σ^2 is large. As λ increases, the range of values of N in Phase 3 becomes smaller.

Given the system parameters, the following algorithm is used to compute solutions to (19) and (20) which correspond to stable operating points.

1) Choose an initial $p_{\rm on}$ such that $0 \le p_{\rm on} \le 1$.
2) Compute p_r from (19).
3) Compute a new value for $p_{\rm on}$ from (20).
4) Iterate steps 2) and 3) until convergence.

It makes physical sense to choose $p_{\rm on}$ initially, ignoring retransmissions. In this way, the preceding algorithm corresponds to the logical order in which $p_{\rm on}$ and p_r converge to steady-state values. The arrows in Fig. 8 illustrate the convergence of the iterative algorithm. Note in particular that the middle intersection point for Phase 2 is not stable. Any initial value of $p_{\rm on}$ not equal to the value associated with this operating point will lead to another operating point.

Fig. 10 shows average packet delay D versus processing gain N for a fixed λ. The dashed line corresponds to the unstable operating points in Phase 2 whereas the solid line corresponds to the stable operating points in Phase 2 or in Phase 3. A small processing gain corresponding to Phase 2 can potentially give a very low delay; however, the system can be perturbed to an equilibrium state in which the delay is infinite.

Fig. 10 indicates that delay is significantly affected by the choice of processing gain. In order to avoid multiple operating points, the processing gain should be selected larger than a threshold N^*. This threshold can be computed from (19) and (20) by observing that N^* corresponds to the transition from Phase 2 to Phase 3. Graphically, this means that the right end-points ($p_{\rm on} = 1$) corresponding to (19) and (20) in Fig. 8 are the same. This implies that

$$\left[1 - \kappa \exp\left(\frac{-\beta P N^*}{\alpha(K-1)P + \sigma^2}\right)\right]^{Lr} = \frac{\lambda L N^*}{R_c}. \qquad (21)$$

In general, there can be at most three solutions to (21). However, since the delay increases with N for $N > N^*$, the optimal choice of N is $\lceil N^* \rceil$ where N^* is the smallest solution to (21) that satisfies $1 < N^* \le N_{\max}$, where $N_{\max} = \lfloor R_c/(\lambda L) \rfloor$. ($N > N_{\max}$ implies that the source packet rate exceeds the rate at which the queue is being emptied.) When the system is in Phase 1 (e.g., $\lambda > 123$ in Fig. 9), the delay is infinite independent of N.

Fig. 11 shows a plot of optimal processing gain N^* versus number of users K for two different values of the coding coefficient β. The slope in this case depends on the packet generation rate λ. For $\lambda = 100$ (assumed in the figure), the maximum number of users that can be supported with finite delay is $K_{\max} = 24\,(\beta = 2)$, and $K_{\max} = 47\,(\beta = 4)$. For $K > K_{\max}$, the system is in Phase 1 for all values of N, so that the corresponding delay is infinite.

V. Two Classes: Continuously Active Users

We now consider two traffic classes in a cell, and assume the continuously active user model. The model for voice users differs from the model for data users in that voice users do not retransmit packets with errors. The same packet length and coding scheme (i.e., the same L, r, κ, and β) are assumed for both classes.

A. Two Voice/Data Classes

Given K_v active voice users and K_d active data users in a cell (i.e., $p_{{\rm on}_v} = p_{{\rm on}_d} = 1$), from (1), the SINR for users in each class is given by

$$\begin{aligned} \text{SINR}_v &= \frac{P_v N_v}{\alpha(K_v - 1)P_v + \alpha K_d P_d + \sigma^2} = \xi_v N_v \\ \text{SINR}_d &= \frac{P_d N_d}{\alpha K_v P_v + \alpha(K_d - 1)P_d + \sigma^2} = \xi_d N_d \end{aligned} \qquad (22)$$

where the subscripts v and d indicate voice and data classes, respectively. Unlike the single-class case, the power assignment

(P_v and P_d) can change SINR$_v$ and SINR$_d$ significantly even if there is no background noise ($\sigma^2 = 0$).

The resource allocation problem in this case is to determine an assignment of powers and processing gains P_v, P_d, N_v, and N_d so that QoS and rate requirements are satisfied. Of course, this assignment in general depends on the number of users, and we would like to determine a "capacity region" or the set of K_d and K_v such that QoS requirements can be satisfied through an appropriate assignment of powers and processing gains. We do this by maximizing K_d given a fixed K_v, subject to constraints on QoS (i.e., data throughput and voice error probability). An equivalent problem is to maximize the throughput of data users for fixed K_v and K_d, subject to a QoS constraint for voice users.

We first consider maximizing the throughput of data users for fixed K_v and K_d. Specifically, the problem is

$$\max_{P_v,P_d,N_v,N_d} \eta_d \quad \text{subject to} \quad p_{b_v} \le \varepsilon_v, \quad P_v \le \mathcal{P}_v, \quad \text{and} \quad P_d \le \mathcal{P}_d \tag{23}$$

where ε_v is the maximum acceptable BEP for voice, and $\mathcal{P}_v$ and $\mathcal{P}_d$ are the maximum transmitted power levels for voice and data users, respectively. Observe that for the constantly active case considered, N_v is determined by the desired information rate for voice. If voice packets are generated at rate λ_v, then

$$N_v^* = \left\lfloor \frac{R_c}{\lambda_v L} \right\rfloor. \tag{24}$$

From (10), the BEP requirement for voice users is equivalent to the SINR constraint

$$\text{SINR}_v \ge \gamma_v \quad \text{where } \gamma_v = \frac{1}{\beta} \ln\left(\frac{\kappa}{\varepsilon_v}\right). \tag{25}$$

Problem (23) is therefore equivalent to

$$\max_{P_v,P_d,N_d} \eta_d \quad \text{subject to} \quad \text{SINR}_v \ge \gamma_v, \quad P_v \le \mathcal{P}_v, \quad \text{and} \quad P_d \le \mathcal{P}_d. \tag{26}$$

The constraint (25) can be combined with (22) to obtain the following:

$$P_d \le AP_v - B \quad \text{where } A = \frac{N_v - \alpha(K_v - 1)\gamma_v}{\alpha K_d \gamma_v} \quad B = \frac{\sigma^2}{\alpha K_d}. \tag{27}$$

The inequalities $B > 0$, $P_d > 0$, and $P_v \le \mathcal{P}_v$ imply that $A \ge B/\mathcal{P}_v$, which gives the following upper bound on the number of voice users:

$$K_v \le \left\lfloor 1 + \frac{N_v^*}{\alpha\gamma_v} - \frac{\sigma^2}{\alpha \mathcal{P}_v} \right\rfloor. \tag{28}$$

Note that both relations (27) and (28) are independent of N_d.

For fixed P_v and N_d, in order to maximize SINR$_d$, P_d should be selected so that equality holds in (27). Furthermore, by substituting $AP_v - B$ for P_d in (22), it is easily shown that SINR$_d$ is an increasing function of P_v. Since η_d is an increasing function of SINR$_d$, it follows that the P_v and P_d that maximize η_d subject to the constraint (26) are given by

$$\begin{aligned} P_v^* &= \mathcal{P}_v \quad P_d^* = A\mathcal{P}_v - B, && \text{if } A < \frac{\mathcal{P}_d + B}{\mathcal{P}_v} \\ P_d^* &= \mathcal{P}_d \quad P_v^* = \frac{\mathcal{P}_d + B}{A}, && \text{otherwise.} \end{aligned} \tag{29}$$

That is, either P_v or P_d is taken to be the maximum allowable value, and the other variable is selected according to (27).

With N_v^*, P_v^*, and P_d^* determined according to (24) and (29), it only remains to maximize η_d with respect to N_d. According to the optimality condition (15) discussed in Section IV-A, the optimal N_d depends only on ξ_d. From (22), ξ_d is a constant given N_v^*, P_v^*, and P_d^*, so that this last optimization is given by (15) where ξ and N are replaced by ξ_d and N_d.

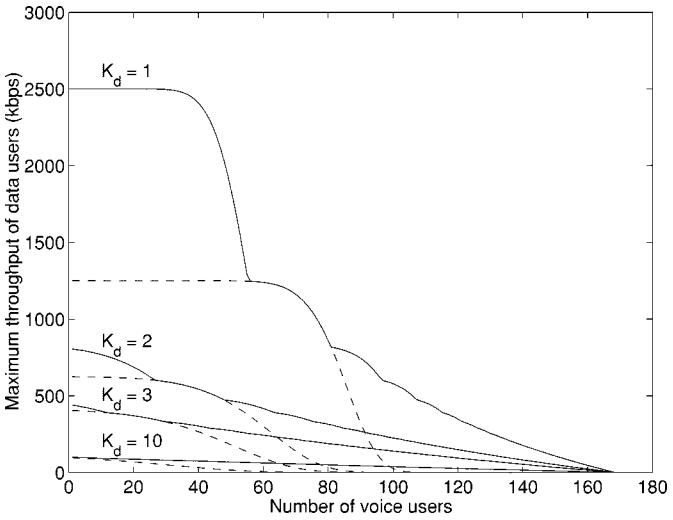

Fig. 12. Maximum throughput of data users versus number of voice users ($\lambda_v = 14.4$ kbps, $R_c = 5$ Mchips/s, $L_v = L_d = 768$ b, $r_v = r_d = 1/2$, and $\mathcal{P}_v/\sigma^2 = \mathcal{P}_d/\sigma^2 = 17$ dB).

Fig. 12 shows plots of the maximum throughput of data users versus K_v with K_d as a parameter. In this example, $\mathcal{P}_v/\sigma^2 = \mathcal{P}_d/\sigma^2 = 17$ dB, the packet generation rate for voice traffic λ_v is 18.75 packets/s, which corresponds to 14.4 kbps, and the upper bound for voice BEP, $\varepsilon_v = 10^{-3}$. Where unspecified, the remaining parameters are the same as those used in the preceding section. As expected, the maximum η_d decreases as K_v or K_d increases. In this particular case, η_d decreases more when K_d is increased (for fixed K_v) than when K_v is increased (for fixed K_d). The "kinks" in the curves correspond to values of K_d and K_v at which the optimal processing gain for data users changes. The dashed line corresponds to the power allocation with a fixed data processing gain ($N_d = 2K_d$). These results show that the optimal assignment of processing gain N_d can significantly increase data throughput.

Given the preceding results, it is straightforward to compute the capacity region, namely, the maximum number of data users that can be accommodated for a fixed number of voice users. In addition to the constraints in (23), we add the throughput constraint $\eta_d \ge \mu_d$. Fig. 13 shows the capacity region for the same parameters used in Fig. 12, and for $\mu_d = 57.6$ kbps (four

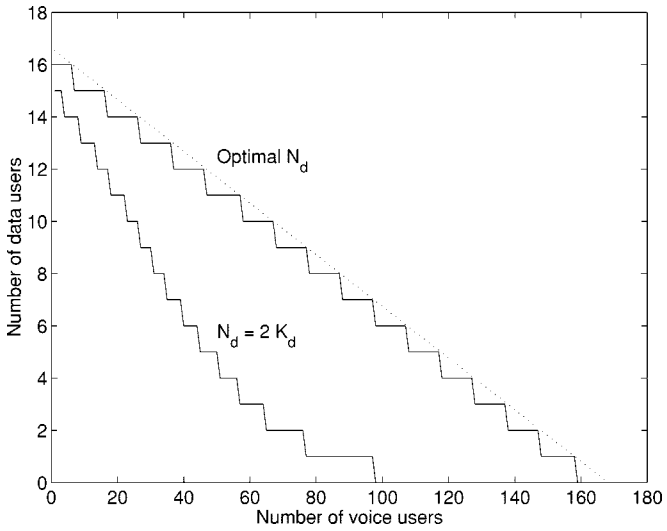

Fig. 13. Capacity region ($\max K_d$ versus K_v) for continuously active users.

times the voice rate). The results indicate that optimizing the processing gain N_d significantly expands the capacity region.

If the QoS constraints are satisfied with equality, then it is possible to derive an explicit expression for the capacity region. From (22) and (25), the QoS requirement of voice users can be rewritten in terms of the power to interference ratio for voice users ξ_v defined in (22). Namely

$$\xi_v = \frac{P_v}{\alpha(K_v-1)P_v + \alpha K_d P_d + \sigma^2} \geq \zeta_v \tag{30}$$

where

$$\zeta_v = \frac{1}{\beta N_v^*} \ln\left(\frac{\kappa}{\varepsilon_v}\right) \quad N_v^* = \left\lfloor \frac{R_c}{\lambda_v L} \right\rfloor. \tag{31}$$

Similarly, the data throughput constraint $\eta_d \geq \mu_d$ combined with (14) and (22), gives

$$\xi_d = \frac{P_d}{\alpha K_v P_v + \alpha(K_d-1)P_d + \sigma^2} \geq \zeta_d \tag{32}$$

where

$$\zeta_d = \frac{1}{\beta N_d^*} \ln\left[\frac{\kappa}{1-(\mu_d N_d^*/(rR_c))^{\frac{1}{Lr}}}\right] \tag{33}$$

and N_d^* is chosen to minimize ξ_d, since it is easily shown that the maximum number of data users K_d is a decreasing function of ξ_d. An expression for N_d^* can be obtained by setting the derivative of the right side of (33) to zero.

The preceding inequalities (30) and (32) can be rewritten as

$$\begin{aligned} P_d &\leq \frac{1-\alpha(K_v-1)\zeta_v}{\alpha K_d \zeta_v} P_v - \frac{\sigma^2}{\alpha K_d} \\ P_d &\geq \frac{\alpha K_v \zeta_d}{1-\alpha(K_d-1)\zeta_v} P_v + \frac{\sigma^2 \zeta_d}{1-\alpha(K_d-1)\zeta_d}. \end{aligned} \tag{34}$$

The capacity region can now be found analytically by adding the power constraints $P_v \leq \mathcal{P}_v$ and $P_d \leq \mathcal{P}_d$. This is illustrated in Fig. 14, which shows plots of P_d versus P_v from (34). All points which satisfy the QoS constraints lie in the shaded region. The

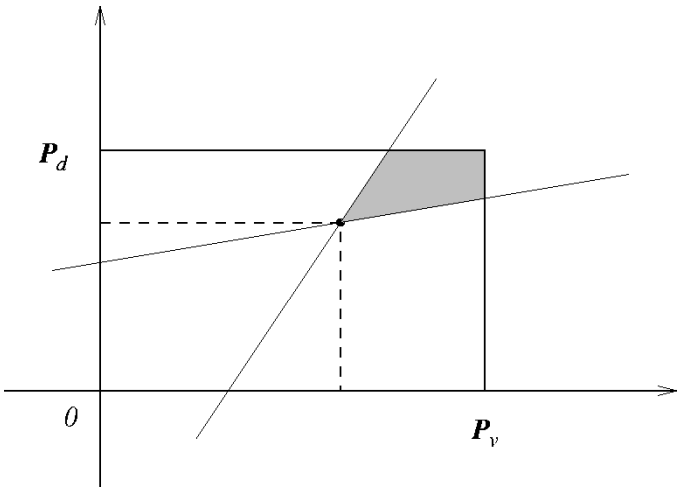

Fig. 14. Illustration of the set of points (P_d, P_v) that satisfy QoS constraints.

intersection of the two lines must lie inside the box defined by $\mathcal{P}_v$ and $\mathcal{P}_d$ to guarantee the existence of a solution that satisfies the QoS constraints.

For a fixed K_d, note that as K_v increases, the intercept point in Fig. 14 moves toward the boundary. Consequently, K_v can be increased to the point where the intersection of the two lines lies on the boundary region defined by the power constraints. This implies that when K_v is maximized, either $P_v = \mathcal{P}_v$ or $P_d = \mathcal{P}_d$. It is then straightforward to show that the capacity region is given by

$$\begin{aligned} K_d &\leq -\frac{\zeta_v(1+\alpha\zeta_d)}{\zeta_d(1+\alpha\zeta_v)} K_v + \frac{(1+\alpha\zeta_d)[\mathcal{P}_v(1+\alpha\zeta_v) - \sigma^2\zeta_v]}{\mathcal{P}_v \alpha \zeta_d (1+\alpha\zeta_v)}, \\ &\qquad \text{if } \frac{\zeta_v(1+\alpha\zeta_d)}{\zeta_d(1+\alpha\zeta_v)} > \frac{\mathcal{P}_v}{\mathcal{P}_d} \\ K_d &\leq -\frac{\zeta_v(1+\alpha\zeta_d)}{\zeta_d(1+\alpha\zeta_v)} K_v + \frac{(1+\alpha\zeta_v)[\mathcal{P}_d(1+\alpha\zeta_d) - \sigma^2\zeta_d]}{\mathcal{P}_d \alpha \zeta_d (1+\alpha\zeta_v)} \\ &\qquad \text{otherwise.} \end{aligned} \tag{35}$$

The dotted line in Fig. 13 corresponds to (35). The line is continuous, as opposed to the staircase plot from the previous analysis, since we have assumed that the inequalities in (34) are satisfied with equality (corresponding to noninteger values of K_v and K_d).

B. Two Data Classes

In this section, we consider two classes of data users in a cell. The same packet length and coding scheme (i.e., the same L, r, κ, and β) are assumed for both classes. Given K_1 users in class 1 and K_2 users in class 2, from (1) we have that

$$\begin{aligned} \text{SINR}_1 &= \frac{P_1 N_1}{\alpha(K_1-1)P_1 + \alpha K_2 P_2 + \sigma^2} = \xi_1 N_1 \\ \text{SINR}_2 &= \frac{P_2 N_2}{\alpha K_1 P_1 + \alpha(K_2-1)P_2 + \sigma^2} = \xi_2 N_2 \end{aligned} \tag{36}$$

where ξ_i denotes the desired user power to interference power plus noise ratio for class i users. From (14), the average throughput η_i for users in class i can be rewritten as

$$\eta_i = \frac{rR_c}{N_i}(1-\kappa\exp(-\beta\xi_i N_i))^{Lr}, \qquad \text{for } i = 1, 2. \tag{37}$$

We wish to select powers and processing gains to maximize the data throughput for users in class 2, given a fixed number of users in each class, and given a constraint on throughput for users in class 1. That is, the optimization problem is

$$\max_{P_1,P_2,N_1,N_2} \eta_2 \quad \text{subject to} \quad \eta_1 = \mu_1, \quad P_1 \le \mathcal{P}_1, \quad \text{and} \quad P_2 \le \mathcal{P}_2 \tag{38}$$

where $\mathcal{P}_1$ and $\mathcal{P}_2$ are the upper power limits for class 1 and 2 users, respectively. This optimization problem is similar to (23) discussed in Section V-A. It is different in that the processing gain for users in class 1 (N_1) is not fixed. (In the previous problem, the processing gain for voice users is chosen to satisfy the rate constraint.)

From (37), it is easily seen that η_2 is an increasing function of ξ_2. Since ξ_2 does not depend on N_2, we first maximize ξ_2 with respect to P_1 and P_2. We subsequently maximize the resulting η_2 with respect to N_2. A crucial observation is that ξ_i, $i = 1, 2$, is an increasing function of P_i and a decreasing function of P_j, $j \neq i$. Consequently, for all values of ξ_1 that satisfies the throughput constraint $\eta_1 = \mu_1$, the minimum value corresponds to a set of $P_i, i = 1, 2$, that maximizes ξ_2.

From the constraint $\eta_1 = \mu_1$ and (37), we have

$$\xi_1 = \frac{1}{\beta N_1} \ln\left[\frac{\kappa}{1 - (\mu_1 N_1/(rR_c))^{\frac{1}{Lr}}}\right]. \tag{39}$$

According to the preceding argument, N_1 should be chosen to minimize ξ_1. Let N_1^* be the processing gain which minimizes ξ_1, and $\zeta_1 = \min_{N_1} \xi_1$. Next, we solve for the P_1, P_2 combination which maximizes ξ_2 subject to $P_1 \le \mathcal{P}_1$, $P_2 \le \mathcal{P}_2$. From (36) and (39), we have that P_2 is a linear function of P_1

$$P_2 = \frac{1 - \alpha(K_1 - 1)\zeta_1}{\alpha K_2 \zeta_1} P_1 - \frac{\sigma^2}{\alpha K_2}. \tag{40}$$

The optimal power assignment is then given by (29), as discussed in Section V-A, where A and B are defined from (40). Finally, maximization of η_2 over N_2 is again determined by (15).

Fig. 15 illustrates the throughput region of class 1 and class 2 users. $K_1 = K_2 = 10$, $\mathcal{P}_1/\sigma^2 = \mathcal{P}_2/\sigma^2 = 17$ dB, and other parameters are the same as in the preceding section. The dashed line corresponds to the optimal power allocation with the suboptimal processing gain assignment $N_1 = N_2 = 2(K_1 + K_2)$. The optimal processing gain results in a substantial increase in maximum throughput.

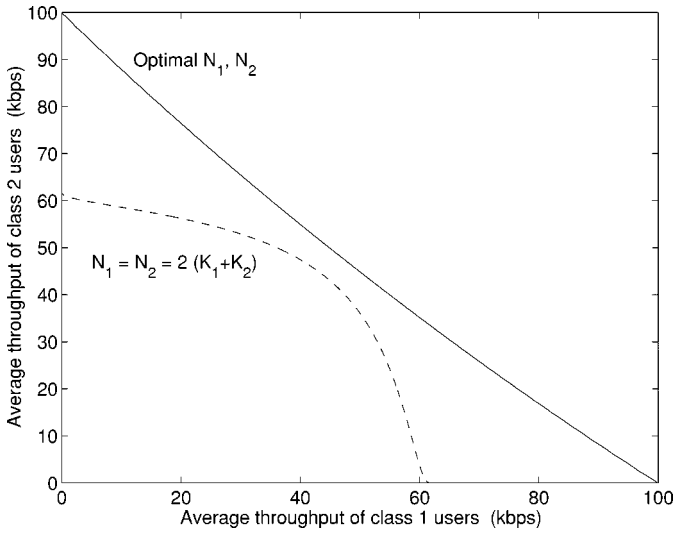

Fig. 15. Throughput region ($\max \eta_2$ versus η_1).

VI. Two Classes: Randomly Active Users

In this section, we consider two classes of voice and data traffic in a cell assuming randomly active users. The number of active voice (data) users within a cell is a random variable which depends on the on/off probability p_{on_v} (p_{on_d}) (see (3)). From (1), the SINR for users in each class is given by

$$\text{SINR}_v = \frac{P_v N_v}{\alpha I_v P_v + \alpha I_d P_d + \sigma^2}$$
$$\text{SINR}_d = \frac{P_d N_d}{\alpha I'_v P_v + \alpha I'_d P_d + \sigma^2} \tag{41}$$

where $I_v = \sum_{k=1}^{K_v - 1} \chi_{k_v}$, $I_d = \sum_{k=1}^{K_d} \chi_{k_d}$, $I'_v = \sum_{k=1}^{K_v} \chi_{k_v}$, and $I'_d = \sum_{k=1}^{K_d - 1} \chi_{k_d}$. The subscripts k_v and k_d denote the kth voice and data user, respectively. In what follows, we assume that within a packet, the on/off indicators χ_{k_v} and χ_{k_d} are independent with distribution $\Pr\{\chi_{k_v} = 1\} = p_{\text{on}_v}$, and $\Pr\{\chi_{k_d} = 1\} = p_{\text{on}_d}$. Strictly speaking, this assumption is not true since the number of voice users affects the retransmission probabilities of data users, which in turn affects the on/off probabilities. However, this assumption is accurate for a small number of users and simplifies the analysis considerably.

Unlike the continuously active case considered in the preceding section, N_v affects *both* SINR_v and SINR_d for randomly active users. Namely, increasing N_v not only increases SINR_v, but also increases p_{on_v}, which increases the duration of the interference and in turn decreases SINR_d. Note that the average transmitted power of voice users $(P_v p_{\text{on}_v})$ also increases as N_v increases. Similarly, SINR_v depends on N_d. Consequently, given a target BEP for voice users ε_v, increasing N_v allows a decrease in P_v or a increase in P_d.

The particular resource allocation problem considered here is

$$\min_{P_v,P_d,N_v,N_d} D_d \quad \text{subject to} \quad p_{b_v} \le \varepsilon_v, \quad P_v \le \mathcal{P}_v, \quad \text{and} \quad P_d \le \mathcal{P}_d \tag{42}$$

where $\mathcal{P}_v$ and $\mathcal{P}_d$ are the maximum transmitted power levels for voice and data users, respectively. This problem appears to be difficult in general, so instead we minimize D_d with respect to N_d, and plot the result as a function of the voice processing gain N_v. The following alternatives for power assignments are adopted: 1) set $P_d = \mathcal{P}_d$ and select P_v to satisfy the voice BEP requirement and 2) set $P_v = \mathcal{P}_v$ and adjust P_d to satisfy the voice BEP requirement. The processing gain for data users is then optimized. In this way, the effect of increasing the *symbol rate* for voice calls on data delay can be examined.

From (9) and (11), the average packet delay for data users can be rewritten as

$$D_d = \frac{L_d N_d (2 - \lambda_d L_d N_d / R_c)}{2R_c(1 - p_{r_d} - \lambda_d L_d N_d / R_c)} \tag{43}$$

where the retransmission probability for data users

$$p_{r_d} = 1 - \left[1 - \sum_{i=0}^{K_v} \sum_{j=0}^{K_d-1} \mathcal{F}\left(\frac{P_d N_d}{\alpha i P_v + \alpha j P_d + \sigma^2} \right) \times \mathcal{B}(K_v, i, p_{\mathrm{on}_v}) \mathcal{B}(K_d - 1, j, p_{\mathrm{on}_d}) \right]^{L_d r_d} \quad (44)$$

and from (10), the average BEP for voice users can be expressed as

$$p_{b_v} = \sum_{i=0}^{K_v-1} \sum_{j=0}^{K_d} \mathcal{F}\left(\frac{P_v N_v}{\alpha i P_v + \alpha j P_d + \sigma^2} \right) \times \mathcal{B}(K_v - 1, i, p_{\mathrm{on}_v}) \mathcal{B}(K_d, j, p_{\mathrm{on}_d}) \quad (45)$$

where the activity probability of voice and data users

$$p_{\mathrm{on}_v} = \frac{\lambda_v L_v N_v}{R_c} \quad p_{\mathrm{on}_d} = \frac{\lambda_d L_d N_d}{R_c(1 - p_{r_d})}. \quad (46)$$

In what follows, we assume that $L_v = L_d = 768$, $K_v = K_d = 10$, and $\mathcal{P}_v/\sigma^2 = \mathcal{P}_d/\sigma^2 = 17$ dB. The remaining parameters are the same as those used in the preceding section. The following iterative procedure was used to compute the minimum D_d as a function of N_v.

1) Choose an initial p_{on_d} such that $0 \le p_{\mathrm{on}_d} \le 1$.
2) For a fixed $P_d = \mathcal{P}_d$ ($P_v = \mathcal{P}_v$), compute P_v (P_d) to achieve $p_{b_v} = \varepsilon_v$.
3) Compute p_{r_d} from (44).
4) Compute a new value for p_{on_d} from (46).
5) Iterate steps 2)–4) until convergence.
6) Compute D_d with the converged values.

To check for multiple operating points, this algorithm was initialized at different values of p_{on_d}. For small values of N_d, the algorithm produced multiple operating points, which correspond to the three different phases shown in Section IV. The optimal processing gain for data users N_d^* is the smallest processing gain associated with Phase 3.

Fig. 16 shows the minimum average data packet delay as a function of N_v with different values for the data arrival rate. The solid line corresponds to the power assignment $P_d = \mathcal{P}_d$, and the dashed line corresponds to the power assignment $P_v = \mathcal{P}_v$. The optimal processing gain is assigned to data users. The discontinuities in Fig. 16 are caused by discrete changes in the optimal processing gain assigned to the data users. Fig. 16 indicates that delay is a decreasing function of N_v. The decrease in delay is most significant for small N_v, which corresponds to $P_v = \mathcal{P}_v$ (dashed line). For small λ_d, the delay decreases slightly as N_v increases. These results indicate that the decrease in transmitted power P_v associated with an increase in N_v offsets the increase in delay caused by the increase in voice activity factor (p_{on_v}).

Fig. 17 shows a plot of average transmitted power, defined as $P_i p_{\mathrm{on}_i}/\sigma^2$, versus N_v. When $P_d = \mathcal{P}_d$ (solid line), there is a slight decrease in average power as N_v increases. As N_v increases, the P_v needed to satisfy the voice BEP requirement decreases. However, the average power does not change much due to the increase in p_{on_v}. The data on/off probability decreases slightly as N_v increases, causing the average transmitted power for data users to decrease slightly. When $P_v = \mathcal{P}_v$ (dashed line), there is an increase in average power. The average voice power increases linearly as N_v increases [see (46)]. The average data power also increases in this region because a larger N_v allows a larger P_d to satisfy the voice BEP requirement.

Fig. 18 shows the capacity region, defined by the maximum number of data users that the system can accommodate as a function of the number of voice users K_v. The outer curve in Fig. 18 corresponds to choosing the optimal N_d, whereas the inner curve corresponds to the assignment $N_d = 2K_d$. The maximum processing gain N_v is assigned to voice users, so that the voice packets are fully *spread* in time, and $p_{\mathrm{on}_v} \approx 1$. The

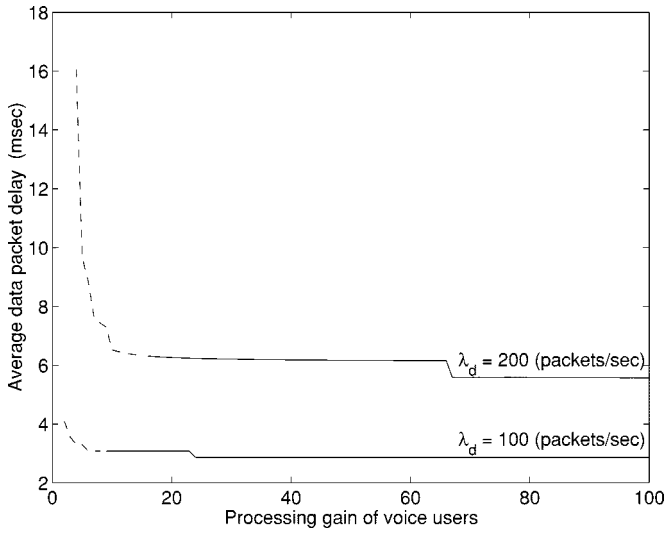

Fig. 16. Minimum data packet delay versus processing gain of voice users with different data packet generation rate ($\lambda_d = 100, 200$ packets/s).

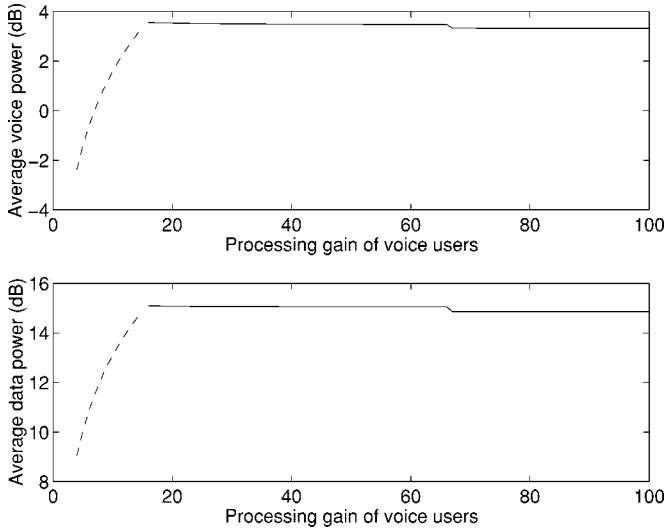

Fig. 17. Average power of voice and data users versus processing gain of voice users ($\lambda_d = 200$ packets/s).

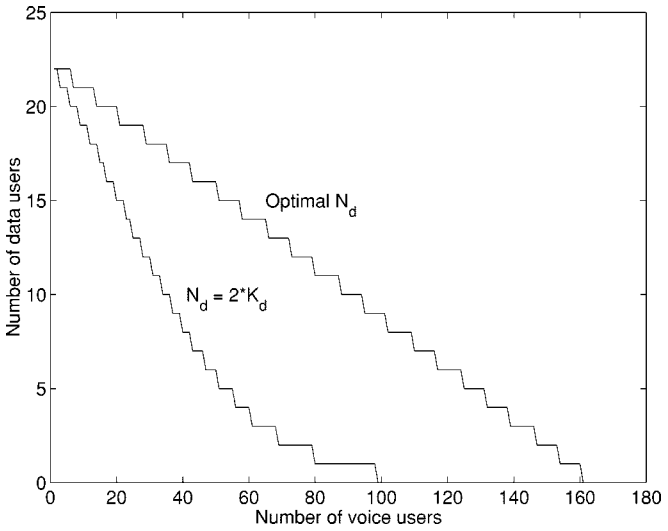

Fig. 18. Capacity region ($\max K_d$ versus K_v) for randomly active users.

data packet generation rate λ_d is 100 packets/s, and the average packet delay for data users must be less than 20 ms ($D_d \leq 20$ ms). The other parameters are the same as for the preceding plots. Fig. 18 indicates that optimizing the data processing gain expands the capacity region significantly.

VII. Conclusion

We have analyzed a model for a multirate/multi-QoS packet data DS-CDMA system. The symbol rate is determined by the selection of processing gain, and the QoS is determined by both the processing gain and transmitted power (energy per bit). Our results indicate that selection of the processing gain as well as transmitted power can significantly affect performance. For both continuously active and randomly active traffic models, we have characterized average delay and throughput for a single class of data users as a function of processing gain and error probability parameters. For continuously active users, it has been shown that the data throughput becomes a more sensitive function of processing gain as the coding gain associated with error control scheme increases. For randomly active users, it has been shown that small processing gains can give rise to multiple equilibria.

Two classes of voice/data and data/data users has also been considered. The "capacity region" in this case requires a joint optimization over the powers and processing gains assigned to the two classes, subject to QoS constraints. It has been shown numerically that this optimization can lead to a significant expansion in capacity region relative to a fixed assignment. For randomly active users, the numerical results indicate that maximizing the processing gain for voice users, or equivalently, *spreading* in time, minimizes the delay of data users and average transmitted power.

In this work, we have not considered assigning multiple codes to different classes to achieve multiple information rates. Other possibilities for future work include consideration of more than two classes of users, channel impairments such as fading and imperfect power control, and more sophisticated (multiuser) detection schemes.

References

[1] T.-H. Wu and E. Geraniotis, "CDMA with multiple chip rates for multi-media communications," in *Proc. Information Science and Systems*, Princeton, NJ, 1994, pp. 992–997.

[2] C.-L. I and K. K. Sabnani, "Variable spreading gain CDMA with adaptive control for integrated traffic in wireless networks," in *Proc. IEEE VTC*, vol. 2, Chicago, IL, July 1995, pp. 794–798.

[3] C.-L. I, G. P. Pollini, L. Ozarow, and R. D. Gitlin, "Performance of multi-code CDMA wireless personal communications networks," in *Proc. IEEE VTC*, vol. 2, Chicago, IL, July 1995, pp. 907–911.

[4] T. Ottosson and A. Svensson, "Multi-rate schemes in DS/CDMA systems," in *Proc. IEEE VTC*, vol. 2, Chicago, IL, July 1995, pp. 1006–1010.

[5] L. C. Yun and D. G. Messerschmitt, "Power control for variable QOS on a CDMA channel," in *Proc. IEEE MILCOM*, vol. 1, Fort Monmouth, NJ, Oct. 1994, pp. 178–182.

[6] M. L. Honig and J. B. Kim, "Resource allocation for packet data transmission in DS-CDMA," in *Proc. Allerton Conf.*, Urbana-Champaign, IL, Oct. 1995, pp. 925–934.

[7] ——, "Allocation of DS-CDMA parameters to achieve multiple rates and qualities of service," in *Proc. IEEE GLOBECOM*, vol. 3, London, U.K., Nov. 1996, pp. 1974–1978.

[8] M. L. Honig, "Adaptive linear interference suppression for packet DS-CDMA," *European Trans. Telecommun.*, vol. 9, no. 2, pp. 173–181, Mar.–Apr. 1998.

[9] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. IEEE PIMRC*, vol. 1, Toronto, Canada, Sept. 1995, pp. 21–25.

[10] S. Yao and E. Geraniotis, "Optimal power control law for multi-media multi-rate CDMA systems," in *Proc. IEEE VTC*, vol. 1, Atlanta, GA, Apr. 1996, pp. 392–396.

[11] Q. Shen and W. A. Krzymien, "Power assignment in CDMA personal communication systems with integrated voice/data traffic," in *Proc. IEEE GLOBECOM*, London, U.K., Nov. 1996, pp. 168–172.

[12] J. T.-H. Wu and E. Geraniotis, "Power control in multi-media CDMA networks," in *Proc. IEEE VTC*, vol. 2, Chicago, IL, July 1995, pp. 789–793.

[13] S.-J. Oh and K. M. Wasserman, "Dynamic spreading gain control in multi-service CDMA networks," , to be published.

[14] N. B. Mandayam and J. M. Holtzman, "Analysis of a simple protocol for short message data service in an integrated voice/data CDMA system," in *Proc. IEEE MILCOM*, vol. 3, San Diego, CA, Nov. 1995, pp. 1160–1164.

[15] A. Sampath, N. B. Mandayam, and J. M. Holtzman, "Analysis of an access control mechanism for data traffic in an integrated voice/data wireless CDMA system," in *Proc. IEEE VTC*, vol. 3, Atlanta, GA, Apr. 1996, pp. 1448–1452.

[16] N. B. Mandayam, J. M. Holtzman, and S. Barberis, "Erlang capacity for an integrated voice/data DS-CDMA wireless system with variable bit rate sources," in *Proc. IEEE PIMRC*, vol. 3, Toronto, Canada, Sept. 1995, pp. 1078–1082.

[17] W.-B. Yang and E. Geraniotis, "Admission policies for integrated voice and data traffic in CDMA packet radio networks," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 654–664, May 1994.

[18] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley III, "On the capacity of a cellular CDMA system," *IEEE Trans. Veh. Technol.*, vol. 40, pp. 303–312, May 1991.

[19] M. B. Pursley, "Performance evaluation for phase-coded spread-spectrum multiple access communication—Part I: System analysis," *IEEE Trans. Commun.*, vol. COM-25, pp. 795–799, Aug. 1977.

[20] R. D. Cideciyan, E. Eleftheriou, and M. Rupf, "Concatenated Reed-Solomon/convolutional coding for data transmission in CDMA-based cellular systems," *IEEE Trans. Commun.*, vol. 45, pp. 1291–1303, Oct. 1997.

[21] R. W. Wolff, *Stochastic Modeling and the Theory of Queues*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[22] C. A. F. J. Wijffels, H. S. Misser, and R. Prasad, "A micro-cellular CDMA system over slow and fast rician fading radio channels with forward error correcting coding and diversity," *IEEE Trans. Veh. Technol.*, vol. 42, pp. 570–580, Nov. 1993.

Joon Bae Kim (S'95) received the B.S.E. degree in electrical engineering from the University of Michigan, Ann Arbor, in 1988 and the M.S. degree in electrical engineering from the University of Wisconsin, Madison, in 1990. He is currently working towards the Ph.D. degree in electrical engineering at Northwestern University, Evanston, IL.

From 1990 to 1991, he served in the Korean Army. From 1991 to 1994, he was with Samsung Advanced Institute of Technology and Samsung Electronics. His research interests include resource allocation in wireless communication networks.

Michael L. Honig (F'97) received the B.S. degree in electrical engineering from Stanford University, Stanford, CA, in 1977 and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 1978 and 1981, respectively.

He joined Bell Laboratories, Holmdel, NJ, where he worked on local area networks and voiceband data transmission. In 1983, he joined the Systems Principles Research Division at Bellcore, where he worked on digital subscriber lines and wireless communications. He was a Visiting Lecturer at Princeton University, Princeton, NJ, during the Fall of 1993. Since the Fall of 1994, he has been with Northwestern University, Evanston, IL, where he is a Professor in the Electrical and Computer Engineering Department. He is an Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY and has served as an Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS.

Dr. Honig was a Guest Editor for the *European Transactions on Telecommunications and Wireless Personal Communications*. He has also served on the Digital Signal Processing Technical Committee for the IEEE Signal Processing Society. He is currently serving as a member of the Board of Governors for the Information Theory Society.