Efficient Design of Voice Carrying Fixed-Network Links in CDMA Mobile Communication Systems *

Khaled M. F. Elsayed [a] Notker Gerlich [b,**] Phuoc Tran-Gia [c]

[a] *Department of Electronics and Communications Engineering, Faculty of Engineering, Cairo University, Giza, Egypt 12613*
E-mail: khaled@ieee.org

[b] *Siemens AG, Information and Communication Networks, Hofmannstrasse 51, D-81359 München, FRG*
E-mail: Notker.Gerlich@icn.siemens.de

[c] *Institute of Computer Science, University of Würzburg, Am Hubland, D-97074 Würzburg, FRG*
E-mail: trangia@informatik.uni-wuerzburg.de

Code Division Multiple Access (CDMA) technology is gaining momentum as the preferred wireless system for the next generation Personal Communication Systems (PCS). In CDMA systems, voice is encoded and packaged in variable length packets that are transported between the mobile station and the switching center. Although the packetization provides a great flexibility in resource allocation, it poses a Quality of Service (QoS) problem on voice. In this paper, we discuss link dimensioning for a typical CDMA system. We consider a T1/E1 link between a CDMA Basestation Transceiver System (BTS) and the Base Station Controller (BSC). Traffic from various voice sources is subject to a framing scheme, which presents a semi-periodic batch input at the T1/E1 interface cards. We analyze the resulting queuing system using discrete-time analysis and large deviations theory and verify our results by simulation. We provide results for the minimum link capacity needed to support a given number of CDMA voice sources. Our results show the potential statistical gain that can be achieved by voice packetization for all practical values of link capacities.

Keywords: CDMA, Cellular Network Capacity Planning, Large Deviations, Discrete-time Analysis

1. INTRODUCTION

The Code Division Multiple Access (CDMA) cellular system promises many advantages over its AMPS and TDMA counterparts. The advantages include enhanced privacy, resistance to jamming, improved voice quality, improved handoff performance, and soft (and increased) capacity [16,5]. One important aspect of CDMA is the usage of a variable bit rate (VBR) voice encoder which reduces the required bandwidth (on the fixed network) and interference (on the air link). The vocoder detects speech and silence in the voice process and adjusts its rate accordingly. It also tunes out background noise and dynamically varies its data transmission rate to operate at one of four different levels.

The VBR vocoder has impact on the air link interface capacity as shown in [17].

* Parts of this paper are based on research supported by the Deutsche Forschungsgemeinschaft (DFG) under grant TR-257/3.

** Work done while the author was with the Institute of Computer Science of the University of Würzburg

Though the air link capacity is the scarce resource in a cellular system, it is nonetheless important to optimize the usage and design of the fixed interconnecting network. A Basestation Transceiver System (BTS) is connected to a Base Station Controller (BSC) via leased lines. These leased lines are quite expensive and add to the cost of operating the cellular system.

In this paper we study the issue of BTS–BSC interconnection. We provide a methodology for performance analysis and dimensioning of the involved communication links to satisfy the quality of service requirements. The quality of service is expressed in terms of a delay budget (maximum allowable delay) and a packet loss probability. It should be pointed out that in this paper, we will use the terms packet delay and waiting time interchangeably and they should refer to exactly the same thing. In this context both terms refer to the total time a packet would need from its arrival at the multiplexer till reaching its destination, i.e. the queueing plus transmission time.

To this end we develop and analyze two stochastic models: a fluid flow model which is analyzed by applying large deviation techniques [18]; and a discrete-time queueing model which we study using the Discrete Time Analysis [1,11,12] approach.

The rest of this paper is organized as follows. In Section 2, we describe the problem under consideration and state the objectives of studying it. In Section 3, the fluid flow model and its analysis is presented. Section 4 presents the discrete-time model and its analysis. In Section 4, we provide a numerical study and validation of the models introduced. Section 5 concludes the paper.

2. PROBLEM DESCRIPTION

Consider a basestation transceiver system (BTS) in a CDMA system. The major functionality of the BTS is to perform the IS-95 [10] air interface specifics and provide connection between the mobile users and the central office switch. Let us focus our attention on the BTS–BSC link in the reverse direction (i.e., from BTS to MSC). Let the link capacity be C bps. The capacity typically comes in multiples of the DS0 channel rate which is 64 *kbps* or 56 *kbps* (depending on line coding used). The link is statistically shared among both the packetized voice traffic of the connected sources and the signaling packets. A limited buffer of size B bits is provided to store the incoming packets until the link becomes available.

The link interface card implements a data link layer which is very similar to the HDLC protocol. Higher priority is given for signaling packets with regard to buffer sharing. Signaling packets can push out already existing voice packets if they find a full buffer.

Let us now describe the sequence of operations performed on voice packets until they reach the link buffer. At the mobile station the vocoder detects talk spurts and silence in the voice process and dynamically adapts its data rate according to speech activity and noise. In steady state, an 8K vocoder (see [10], Appendix A) transmits at one of four rates. Depending on the rate the vocoder generates variable length packets out of 160 voice samples accumulated during a 20 *msec* interval. Table 1 lists the packet lengths and rate probabilities of the 8K vocoder. The packet lengths shown include 10 octets header

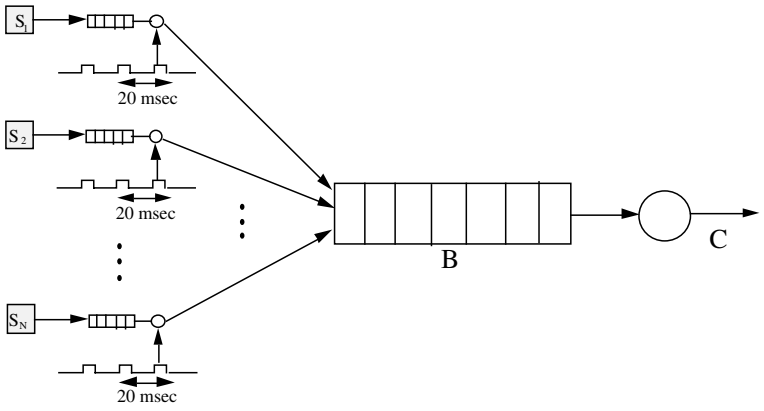

Figure 1. Voice Path from Mobiles to the BTS–BSC link

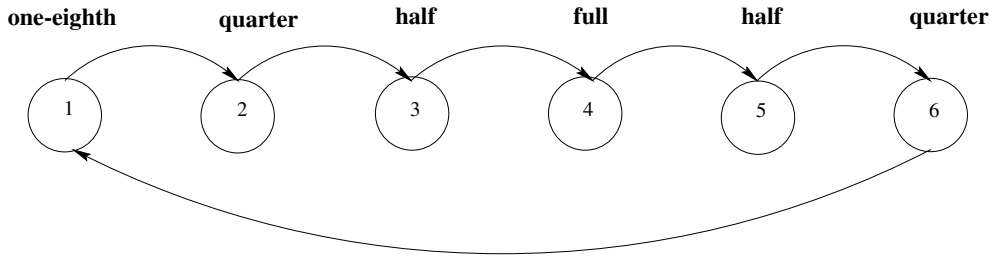

Figure 2. A Markov chain representation of the voice source.

information added by the HDLC protocol. This causes the effective rate seen at the fixed transmission links to be higher than those on the radio link. The effective rate is obtained by dividing the packet length by the fixed frame interval length of 20 *msec*. The effective rate is shown in Table 1.

Table 1
Rate distribution and corresponding packet lengths

Vocoder Rate [*bps*]	Packet Length [*bit*]	Effective Rate [*bps*]	Probability
9600 (Full)	256	12800	0.291
4800 (Half)	160	8000	0.039
2400 (Quarter)	120	6000	0.072
1200 (One-eighth)	96	4800	0.598

Another version of the vocoder that provides better voice quality operates at a maximum rate of 14.4 *kbps* (13K vocoder). In this paper, we focus on the 8K vocoder without loss of generality.

Due to voice activity detection the packet lengths of contiguous voice packets are correlated. During a talk spurt the frequencies of the larger packets are higher than during a silence period; and vice versa.

The above source can be modelled by a discrete-time Markov chain as follows. The sources pass the vocoder states in a cyclic manner: the vocoder ramps from one-eighth rate up to full rate and down back to one-eighth rate passing the intermediate rates. Thus, the Markov chain has six states as shown in Figure 2. The corresponding rates are: one-eighth,

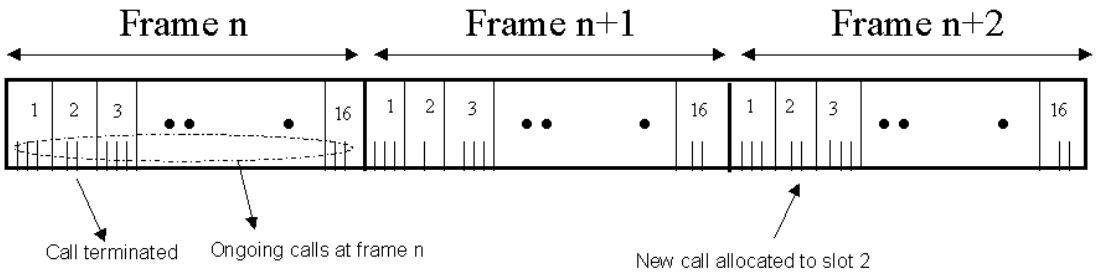

Figure 3. Illustration of the frame structure and the 16 slots. In frame n, slot 2 has 2 calls allocated. In frame n+1, one of the calls in slot 2 is terminated. A new call is started in frame n+2 and is allocated to slot 2.

quarter, half, full, half, quarter. We construct the transition probability matrix from the steady state probabilities given by Table 1. The steady state probabilities of states 1 and 4 of the Markov chain are the state probabilities for one-eighth and full rate respectively. The state probabilities of states 2 and 6 (3 and 5) are given by the state probability of the quarter (half) rate state divided by 2. Assuming that the mean sojourn time of state 1 is 620 *msec* the sojourn time means of states 2 to 6 are chosen proportional to the state probabilities. Since the sojourn time in a state of a discrete time Markov chain has a geometric distribution, the transition probability matrix of the chain can be calculated to read

$$\mathcal{P} = \begin{pmatrix} 0.968 & 0.032 & 0 & 0 & 0 & 0 \\ 0 & 0.464 & 0.536 & 0 & 0 & 0 \\ 0 & 0 & 0.011 & 0.989 & 0 & 0 \\ 0 & 0 & 0 & 0.934 & 0.066 & 0 \\ 0 & 0 & 0 & 0 & 0.011 & 0.989 \\ 0.536 & 0 & 0 & 0 & 0 & 0.464 \end{pmatrix}$$

At the base station, for each mobile station a digital signal processor unit (ASIC) is allocated that performs IS-95 processing and retrieves the packetized voice data from the raw IS-95 stream. A processor attached to the ASICs will schedule packet transmission from the different sources according to a framing mechanism described as follows. A system wide frame of $T = 20$ *msec* is used to multiplex the packets. See Figure 3 for illustration of the framing concept. The frame is divided into M slots with length T/M *msec*. For each voice source one slot is assigned during call setup and the source is only allowed to transmit packets in the assigned slot. This slot may change only when a call goes through a hard handoff. It is possible that more than one source is assigned to the same slot.

The slot assignment is done such that the load is distributed evenly among the slots. However, the free assignment of slots to connections is restricted by the fact that calls going through soft handoff require the same slot in all BTSs to which they are connected. This particular assignment is required to ensure that packets originating from the same voice source arrive at the same time at the Selector Bank Subsystem (SBS) of the BSC.

At the BSC, digital signal processors decompress the voice packets to retrieve the original voice samples. If there is more than one packet due to soft handoff the selector chooses the packet promising the best voice quality; the other packets are dropped. In the opposite direction of the connection the packet is copied for each BTS. Scheduling of packets is introduced to shape and smooth the traffic before submitting it to the link.

The BTS-BSC link buffer receives the packetized voice and signaling (related to call processing) in its common buffer and transmits the packets in a FIFO manner.

The quality of service for the BTS–BSC link is defined as follows:

- The maximum delay (or delay budget) for an arbitrary packet should be less than d *msec*. Typically, d is set to 4 *msec*. Since the delays involved are random, we express the maximum delay as the 99.99% quantile of the delay distribution of an arbitrary packet (delay budget).
- The packet loss probability due to the finite buffer should be kept below ϵ, where ϵ is typically in the range $[10^{-3} - 10^{-6}]$.

When designing a system the following questions need to be addressed. What is the minimum link speed C required for a given number N of voice channels which the BTS is serving? The other way round, what is the number N of voice channels that can be supported by a given link capacity C? Finally, what is the appropriate buffer size B to sustain the required quality of service? Large buffers would increase the maximum possible delay while enhancing the loss performance.

To answer these questions appropriate stochastic models are required. In the following sections we develop and analyze two models: The fluid flow model covers the correlations inherent in the stream of vocoder packets issued by an individual vocoder but cannot account for the framing structure imposed by the scheduling prior to transmission on the link. The discrete-time queuing model on the other hand accounts for the framing but does not cover the vocoder stream correlations. The results are validated by simulation which covers both correlations in the vocoder streams and the framing of the scheduler.

Signaling traffic is assumed to generate α% of the voice traffic. Typically α is about 1% to 10%. We concentrate on the voice traffic only and we can scale the obtained results to reflect the effect of signaling. (As an approximation the obtained number of channels is multiplied by $\frac{1}{1+\alpha}$ to account for the signalling traffic.)

3. FLUID FLOW QUEUING MODEL

Large deviations theory provides techniques for estimating properties of rare events such as their frequency and the manner in which they occur. Recently, large deviations theory gained popularity as a valid methodology for analysis of ATM networks [4,7,19]. Elwalid et al. [3] proposed the Chernoff-Dominant Eigenvalue approximation to model the buffering behavior of some queuing systems in the asymptotic case. The queue length distribution of a queuing system fed by a large number of Markov-modulated sources is approximated by:

$$G(b) \approx Ae^{zb} \tag{1}$$

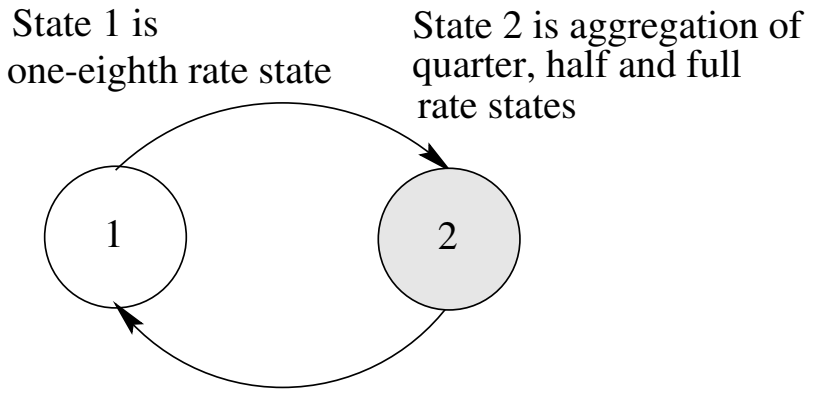

Figure 4. Aggregated 2-state Markov chain representation of the voice source.

where $G(b)$ is the cumulative distribution function that the queue length is larger than b, A is the loss in a bufferless multiplexing system as estimated from Chernoff's theorem, and z is the dominant eigenvalue in a buffered multiplexing system. The value z determines the large buffer behavior of the multiplexing system. A summary of the methodology used to determine the values of A and z is given in Appendix A.

3.1. Approximation of the CDMA Link Multiplexing using Chernoff-Dominant Eigenvalue Approach

First we need to justify the use of large deviations theory in this context. The Erlang capacity of a 3-sector base station will be in the range of 36-72 [17]. Thus on the average the system sees a large number of connected voice users at any time (including those in handoff). The first approximation we take is that we assume that the framing structure imposed on the voice sources (see section 2) can be ignored in the long term. The second approximation performed is that we observe from Table 1, that the source is mainly in full rate of 9.6 (effectively 12.8) *kbps* or one-eighth rate 1.2 (effectively 4.8) *kbps*. We assume that the half and one fourth rate are aggregated into the full rate state and their probabilities are added to the probability of being in full rate. The sources we get are therefore a worst case source w.r.t the actual source. This allows simplification and reduction of the calculations needed to evaluate A and z. The resulting source (see Figure 4) is a two-state source with state 1 having a steady state probability of $p_1 = 0.598$ and (effective) rate $R_1 = 4800$ *bps* and state 2 with a steady state probability of $p_2 = 0.402$ and (effective) rate $R_2 = 12800$ *bps*.

Discretization of the Problem

At the full rate, let the maximum packet size in bits generated over a 20-msec period be S_{max} and let S_{min} be the minimum packet size in bits generated over the same period in the one-eighth rate. Define $S_{data\_unit} = gcd(S_{max}, S_{min}, B)$, where gcd is the greatest common divisor function. We artificially introduce small packets of size S_{data_unit} which we call data units. An incoming packet is (artificially) divided into a number of data units. In the worst case, S_{data_unit} is equal to one bit. Let Δ be the transmission time of a data unit on the link speed with rate C. The buffer size is expressed in terms of the data units

as follows, $\hat{B} = B/S_{data\_unit}$. The link speed C is normalized such that the multiplexer is capable of transferring one data unit each Δ time units. The rates at full and one-eighth states are normalized with respect to the new link speed. Let us denote the one-eighth and full rate state as states 1 and 2. Let the normalized rates at these states be denoted by $\hat{r_1}$ and $\hat{r_2}$ respectively, then $\hat{r_i} = r_i/C$, $i = 1, 2$. Furthermore, let the mean length of the sojourn time of the source in state i be $\bar{t_i}$ data units. We are now in a position to show the results for our system.

Using the above characterization we have a homogeneous set of N sources each characterized by a 2-state Markov chain. The probability transition matrix of the source is given by $P = \begin{bmatrix} \alpha_1 & 1-\alpha_1 \\ 1-\alpha_2 & \alpha_2 \end{bmatrix}$, where $\alpha_i = 1 - 1/\bar{t_i}$, $i = 1, 2$ and its rate vector R is given by $R = (\hat{r_1}, \hat{r_2})$. We apply the Chernoff-dominant eigenvalue approach as follows.

Finding the value of A

For the purpose of evaluating the constant A, the 2-state CDMA voice source can be mapped to an on-off source as follows. Since $\hat{r_1} < \hat{r_2}$, for a given number of sources N, there is always at least $N \times \hat{r_1}$ data units out of the link capacity being used by the sources. We can use the following equivalent system for the purpose of calculating the value of A. We set the link capacity to $\tilde{C} = 1 - N \times \hat{r_1}$ with N on-off sources where the rate of the on state is equal to $\tilde{r_2} = \hat{r_2} - \hat{r_1}$ and the rate of the off state is of course $\tilde{r_1} = 0$.

Weiss [18] reports that for such a system of N on-off sources with link speed $\tilde{C}$, the value of A is given by:

$$A = \exp\left(-N\left[\tilde{c}\ln(\frac{\tilde{c}}{\pi_1}) + (1-\tilde{c})\ln(\frac{1-\tilde{c}}{1-\pi_1})\right]\right) \tag{2}$$

where $\tilde{c} = \frac{\tilde{C}}{N\hat{r_2}}$ and π_1 is the probability of being in state 1.

Finding the value of z

To find the value of the dominant eigenvalue of the system we solve for z which is a solution of $N\left(-\frac{\log \mu(z)}{z}\right) = C$ where $\mu(z)$ is the PF-eigenvalue of the matrix $X(z) = e^{-zR_d}P$. We have $X(z) = \begin{bmatrix} e^{-r_1 z}\alpha_1 & e^{-r_1 z}(1-\alpha_1) \\ e^{-r_2 z}(1-\alpha_2) & e^{-r_2 z}\alpha_2 \end{bmatrix}$. To find the eigenvalues of $X(z)$, we solve the equation $|\mu I - X(z)| = 0$. This gives the PF-eigenvalue equal to

$$\mu(z) = \frac{1}{2}\left[\alpha_1 e^{-r_1 z} + \alpha_2 e^{-r_2 z} + \sqrt{(\alpha_1 e^{-r_1 z} + \alpha_2 e^{-r_2 z})^2 - 4(1-\alpha_1-\alpha_2)e^{-(r_1+r_2)z}}\right] \tag{3}$$

from which we obtain the value of z that verifies the equation $N\left(-\frac{\log \mu(z)}{z}\right) = C$ via a numerical iteration.

As a first approximation to z, we can use the value of z for the continuous time fluid flow model which can be expressed in a closed form (see Appendix B).

Once the values of A and z are found, then the packet loss probability PLP is upper-bounded by the data unit overflow probability which is approximately equal to $G(\hat{B}) \approx Ae^{z\hat{B}}$, i.e. $PLP \leq Ae^{z\hat{B}}$. To find the delay bound we note that the waiting time can be expressed in terms of the data unit transmission time Δ, so the probability that the waiting time is larger than $x\Delta$ is given by $P(delay \geq x\Delta) \approx Ae^{zx}$. So if we would

like to find the value of the delay D such that $u\%$ of delays is less than D, we solve for $x\Delta \approx \ln(1-u)/A$ and we set $D = \max(0, x\Delta)$.

4. DISCRETE-TIME QUEUING MODEL

Due to the framing structure the voice packet arrivals are a deterministic arrival process with a given number of arrivals at each slot. This is only true, however, for the limited period of time where we have a particular call mix. Since a very large number of slot allocations is possible, considering all possible realizations results in a binomial distribution of the number of packets in a given slot. The service time is given by the packet length divided by link speed C.

We consider the buffer as a finite capacity queuing model operating in discrete time. Time is discretized into intervals of unit length Δ, which is the transmission time of a single data-unit. The size of a data-unit is given by the greatest common divisor of the packet lengths as given by a discrete r.v. (random variable) V (in this work the distribution of V is given by Table 1) and the buffer size B. The buffer accommodates $\hat{B}$ data-units. During a single duty cycle of constant length a, which is a multiple of Δ, packets transmitted by the active portion of N sources are collected and submitted to the buffer. Thus, we have arrivals of batches of packets, each of which is a batch of data-units. The number of active sources, and hence the number of packets in a batch, is governed by a binomially distributed r.v. $X \sim \mathsf{BIN}(N, 1/16)$, where $\mathsf{BIN}(n, p)$ denotes the binomial distribution with parameters n and p.

Two different admission policies are considered, if an arriving packet does not fit into the buffer:

1. *Partial packet loss*: free positions of the buffer are filled; the remaining data-units are lost.
2. *Full packet loss*: the packet is lost as a whole.

It should be noted that partial packet loss policy does not make sense from the implementation's point of view. Nevertheless, it makes sense in terms of the state analysis which is simplified for partial packet loss. Some related queuing problems have already been considered in the literature before. These problems usually cope with a buffer whose waiting spaces are capable of holding one packet each. In contrast we have to deal with variable length packets to be queued in a buffer, where the elementary storing unit is 1 *data-unit* (in the worst case, the data unit size is equal to 1 *bit*). In [15] a single server infinite queue with Poisson batch arrivals and general service times ($M^{[X]}/G/1$) is studied. [20] derives the generating function of the state probabilities of the $GI^{[X]}/M/c$ system. Statistical multiplexers for packetized voice connections are investigated in [6], [9], and [14]. The former analysis applies only to a constant packet length while we have variable length packets; the latter two have infinite buffers in common whereas we have to deal with a finite buffer.

In [8] and [13] finite queuing systems with batch arrivals are investigated. The blocking probabilities for full and partial packet loss of the $M^{[X]}/M/1-S$ and $G^{[X]}/D/1-S$ queuing systems are derived, respectively. The difference to the queuing system investi-

gated here is that in our case the batching process has two stages: batches of packets, each of which consists of a number of data-units.

4.1. State Analysis

4.1.1. Partial packet loss

When analyzing the buffer occupancy distribution using discrete time analysis technique [1,11,12] one keeps track of the time-dependent unfinished work process. This process is described by the r.v.'s

U_n^- r.v. for the number of data-units present in the buffer immediately *prior* to the arrival instant of the nth batch;

U_n^+ r.v. for the number of data-units in buffer immediately *after* the arrival instant of the nth batch;

Y_n r.v. for the number of data-units in batch n.

The discrete distributions of these r.v.'s are denoted by $u_n^-(k)$, $u_n^+(k)$, and $y_n(k)$, respectively. The relation between these r.v.'s is given by

$$U_n^+ = \min\{U_n^- + Y_n,\ \hat{B}\}, \tag{4a}$$

$$U_{n+1}^- = \max\{U_n^+ - a,\ 0\}. \tag{4b}$$

In terms of the discrete distributions the last equation reads

$$u_n^+(k) = \pi^{\hat{B}}[\, u_n^-(k) \circledast y_n(k)\,] \tag{5a}$$

$$u_{n+1}^-(k) = \pi_0[\, u_n^+(k) \circledast \delta(k+a)\,], \tag{5b}$$

where $\pi^s[\cdot]$ and $\pi_0[\cdot]$ are linear sweep operators on probability distributions defined by

$$\pi^m[z(k)] = \begin{cases} z(k) & \text{for } k < m, \\ \displaystyle\sum_{i=m}^{\infty} z(i) & \text{for } k = m, \\ 0 & \text{for } k > m; \end{cases} \tag{6a}$$

$$\pi_m[z(k)] = \begin{cases} 0 & \text{for } k < m, \\ \displaystyle\sum_{i=-\infty}^{m} z(i) & \text{for } k = m, \\ z(k) & \text{for } k > m; \end{cases} \tag{6b}$$

and '$\circledast$' denotes the discrete convolution

$$z(k) = z_1(k) \circledast z_2(k) \ = \ \sum_{i=-\infty}^{+\infty} z_1(k-i) \cdot z_2(i). \tag{7}$$

Note, that the convolution of a distribution $z(k)$ and the distribution defined by the Kronecker-function

$$\delta(k) = \begin{cases} 1 & \text{for } k = 0, \\ 0 & \text{for } k \neq 0 \end{cases} \tag{8}$$

denotes a shift of $z(k)$ by a indices.

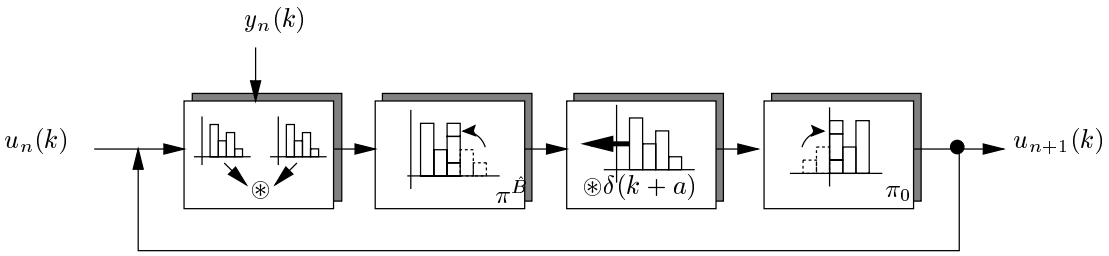

Figure 5. computational diagram for buffer occupancy distribution

Since Eqn. (5) represents a recursive relation between the system states seen upon arrival by consecutive batches,

$$u^-_{n+1}(k) = \pi_0 \left[\pi^{\hat{B}} [u^-_n(k) \circledast y_n(k)] \circledast \delta(k+a) \right], \tag{9}$$

it gives rise to the algorithm depicted in the computational diagram in Figure 5. It should be noted that the convolutions involved can be computed efficiently by using fast Fourier transforms (FFTs). In our case of identically and independently distributed batch sizes Y_n (in data-units) the computational diagram describes an iterative algorithm to determine the equilibrium buffer occupancy distribution

$$u^-(k) = \lim_{n\to\infty} u^-_n(k). \tag{10}$$

To complete the derivation we have to give the distribution $y_n(k)$ of r.v. Y_n. Since Y_n is the sum of X r.v.'s V, $y_n(k)$ is the compound distribution of the corresponding distributions $x(k)$ and $v(k)$:

$$y_n(k) = \sum_{i=0}^{N} v^{\circledast i}(k) \cdot x(i), \tag{11}$$

where, $v^{\circledast i}(k)$ denotes the i-fold convolution of $v(k)$ with itself and, naturally, $v^{\circledast 0}(k) = \delta(k)$.

4.1.2. Full packet loss

If we employ this policy the whole packet is lost if it does not fit into the buffer upon arrival. We obtain the following equations for the system-state r.v.'s as defined above

$$U^+_n = \begin{cases} U^-_n + Y_n & \text{if} \quad U^-_n + Y_n \le \hat{B} \\ U^-_n + \tilde{Y}_n & \text{if} \quad U^-_n + Y_n > \hat{B} \end{cases} \tag{12a}$$

$$U^-_{n+1} = \max\{U^+_n - a,\ 0\}, \tag{12b}$$

where r.v. $\tilde{Y_n}$ denotes the fraction of Y_n that is accepted. Distributions of these r.v.'s are given as

$$u^+_n(k) = u^-_n(k) \circledast y_n(k) \quad + \left[\sum_{i=\hat{B}-k+1}^{\infty} v(i) \right] \cdot \sum_{j=1}^{N} x(j) \sum_{i=0}^{j-1} [u^-_n(k) \circledast v^{\circledast i}(k)] \qquad k = 0, 1 \ldots \hat{B}, \tag{13a}$$

$$u^-_{n+1}(k) = \pi_0 [\, u^+_n(k) \circledast \delta(k+a) \,], \tag{13b}$$

where Eqn. (13a) is derived as follows. Consider the arrival of a batch of j packets which occurs with probability $x(j)$. If the batch is small enough to fit into the buffer completely $U_n^+ = k$ if $U_n^- + Y_n = k$. The batch does not fit into the buffer and is truncated to leave $U_n^+ = k$ if for any $0 \leq i \leq j-1$ i packets together with U_n^- sum up to k and the length of the $(i+1)$th packet exceeds $\hat{B} - k$. Unconditioning with respect to variables i and j gives rise to Eqn. (13a).

As with partial packet loss, Eqn. (13) represents a recursive relation between system states seen upon arrival of consequent batches and, hence, gives rise to an iterative algorithm to calculate the state probabilities in equilibrium.

4.2. Packet Loss Probability and Waiting Time Distribution

The derivation of the packet loss probability and the waiting time distribution applies for both partial and full packet loss policy.

Observing a tagged packet, we define the r.v. Y^* to be the end, i.e., the last data-unit of this packet within its batch of packets; $y^*(k)$ denotes the corresponding distribution. Clearly, given a buffer occupancy of U upon arrival of the batch the tagged packet is lost if $U + Y^* > \hat{B}$. This leads to the loss probability as given by

$$p_{\text{loss}} = \sum_{i=\hat{B}+1}^{\infty} [u(i) \circledast y^*(i)], \tag{14}$$

where distribution $y^*(k)$ remains to be derived.

To that end we define the conditional distribution $y^*_{|X=j}(k)$, which denotes the distribution of the tagged packet's end arriving within a batch of j packets. Since the position of the tagged packet within the batch is distributed uniformly by applying complete probability formula we obtain

$$y^*_{|X=j}(k) = \sum_{i=1}^{j} v^{\circledast i}(k) \cdot \frac{1}{j}. \tag{15}$$

The probability for the tagged packet to arrive within a batch of j packets is $j \cdot x(j)/E[\ X\ ]$, where the operator $E[\ Z\ ]$ denotes the expectation of r.v. Z. Thus, unconditioning with respect to j gives

$$\begin{aligned} y^*(k) &= \frac{1}{E[\ X\ ]} \sum_{j=1}^{N} y^*_{|X=j}(k) \cdot j \cdot x(j) \\ &= \frac{1}{E[\ X\ ]} \sum_{j=1}^{N} x(j) \sum_{i=1}^{j} v^{\circledast i}(k). \end{aligned} \tag{16}$$

The waiting time distribution $w(k)$ for packets transmitted is given by the buffer occupancy U that a tagged packet which is granted admission sees plus the work brought into the system by the packets ahead of it (including the tagged packet itself), Y^*. Defining thus

$$\tilde{U}^- = U^- + Y^* \tag{17a}$$

$$\tilde{u}^- = u^-(k) \circledast y^*(k) \tag{17b}$$

the waiting time distribution reads

$$w(k) = \begin{cases} \dfrac{\tilde{u}^-(k)}{\sum_{i=0}^{\hat{B}} \tilde{u}^-(i)} & 0 \le k \le \hat{B}, \\ 0 & k > \hat{B}. \end{cases} \tag{18}$$

5. RESULTS

The quality of service for the BTS–BSC link is defined in terms of both the delay of an arbitrary packet and the packet loss probability: The delay should be less than $d = 4$ *msec* for 99.99% of the packets, i.e. a 10^{-4} delay budget of 4 *msec* must be kept, and the packet loss should be below some $\epsilon \in [10^{-6}, 10^{-3}]$. In the results. the link capacity C is given as a multiple of DS0 channels (64 Kbps) and is denoted by $n \times 64$, where n is an integer ≥ 1.

Using the discrete-time model, Figures 6 and 7 show the probability for a packet to experience a delay of more than $d = 4$ *msec* and the packet loss probability versus the number of voice channels, respectively. The packet length distribution is the distribution of Table 1 and the buffer size is $B = 2$ *KBytes*. The diagrams clearly indicate that the limitation of the packet delay constitutes the stronger constraint. Take for instance the 16×64 *kbps* curve: the loss probability constraint suggests a support of 130 voice channels whereas the delay limitation allows only a number of 110 voice channels to be supported. Since both the loss and the delay curves are steep the quality of service improves significantly if one allows a slightly smaller number of voice channels to be supported. Consider again the 16×64 *kbps* curve: the loss probability drops 3 orders of magnitude when supporting 130 channels instead of 135.

Similarly, using the large-deviations model, Figures 8 and 9 show the probability for a packet to experience a delay of more than $d = 4$ *msec* and the packet loss probability versus the number of voice channels, respectively. The buffer size is taken to be $B = 2$ *KBytes*. Here, we note that the packet loss places a more stringent requirement on the system dimensioning. The delay is governed more or less by an equation that is a function of the buffer contents.

In the following, we will review the questions issued in Section 2. Table 2 shows the maximum number of voice channels that can be supported given a certain link capacity (in multiples of 64 *kbps*) as computed by the two models and a simulation. Table 2 also shows the number of channels that can be supported using circuit switching/peak rate allocation. This is obtained by dividing link capacity by the maximum effective voice channel bit rate of 12800 *bps* (see Table 1).

The simulation covers both correlations in the vocoder packet streams and the framing imposed by the scheduler. To this end each voice source is modeled by a first order discrete time Markov chain as described in Section 2.

The statistics are collected from simulation runs with a simulated time of over three hours each and confidence intervals of 95%. The two values in the simulation column are the smallest and largest number of sources obtained by the simulation. We note that for link speeds less than 8 DS0 channels, the large deviation model allows more voice channels

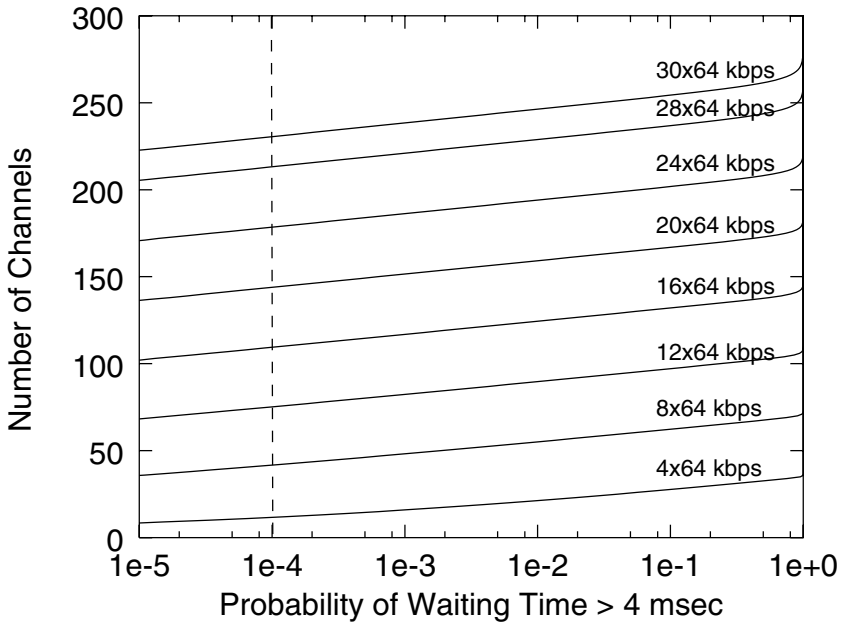

Figure 6. Complementary Packet Waiting Time Distribution [Discrete-time]

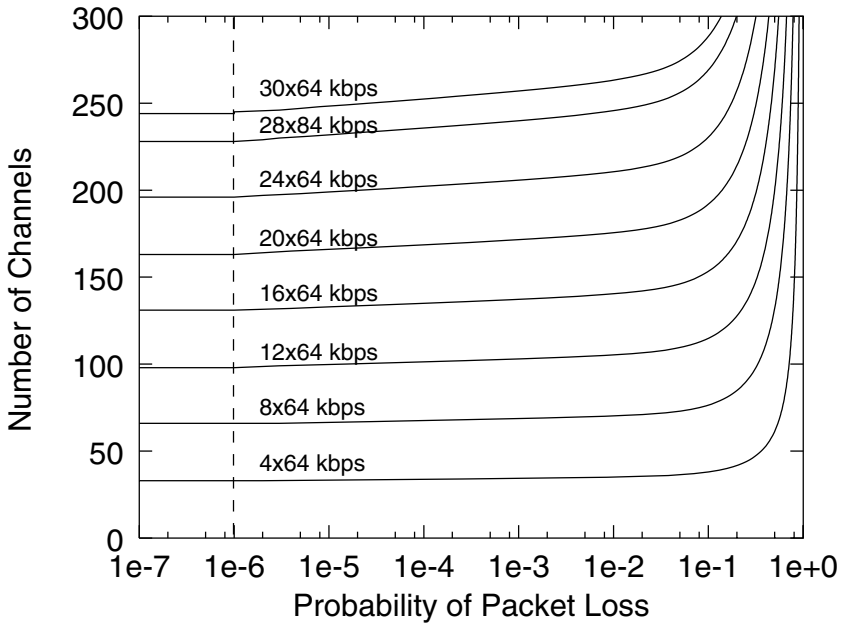

Figure 7. Packet Loss Probability [Discrete-time]

to be multiplexed and for link speeds greater than 16 DS0 channels, it allows less voice channels to be multiplexed than the discrete-time model. In between both models are very close. The simulation results are generally higher than the result of the models. This is in accordance with what we expected: the scheduling of packets to be transmitted at particular slots of the 20 *msec* tends to reduce the effect of long term correlations inherent in the superposition of the the multiple state Markov process modeling the voice source. Also, we knew in advance that our assumption of batch arrivals in the discrete-time model is an over-estimation of the slotted arrival process introduced by the packet scheduler.

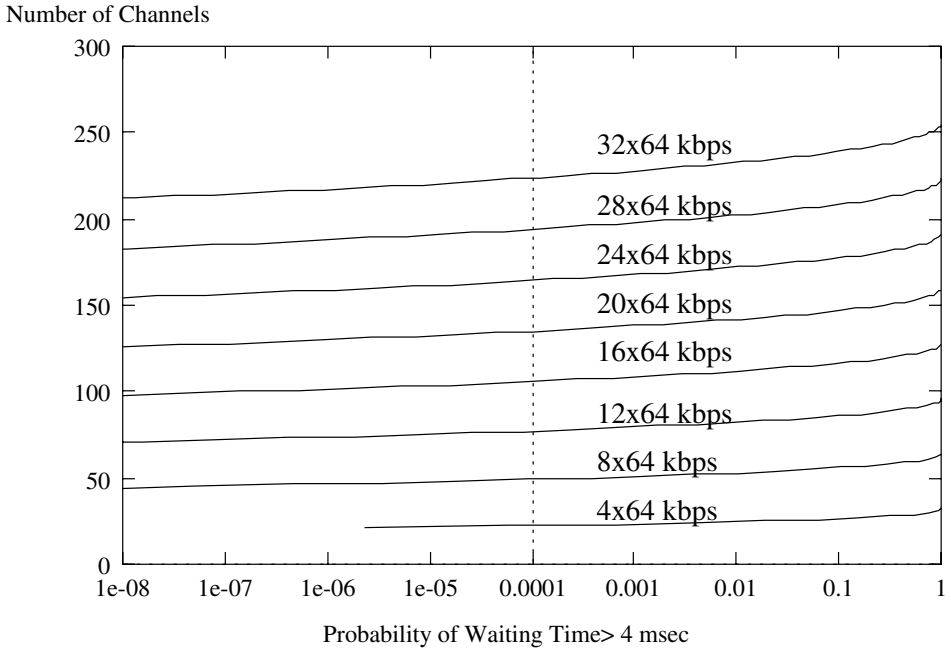

Figure 8. Complementary Packet Waiting Time Distribution [Large deviations]

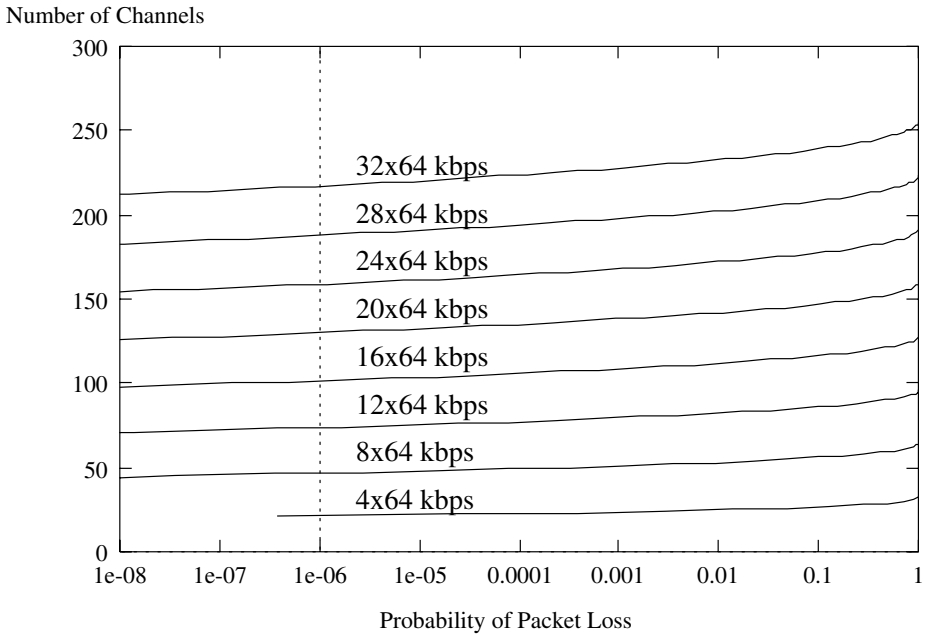

Figure 9. Packet Loss Probability [Large deviations]

This shows that both analytic models provide a lower bound on the allowable number of voice channels for a given link capacity (except for small link capacities). This is a desired property since wireless operators would have stringent requirements on the performance of the links. We take the minimum number of channels obtained by the two models as our estimate of the allowable number of voice channels for a given link capacity. If N_{LD} and N_{DT} are the number of voice channels obtained by the large deviations model and

Table 2
Maximum Number of Multiplexable Voice Channels

Link Speed (multiple of 64*kbps*)	Voice Channels			
	Circuit Switching	Large deviations	Discrete-time	Simulation
4	20	20	11	[9, 14]
8	40	45	41	[47, 49]
12	60	75	75	[83, 87]
16	80	105	109	[119, 122]
20	100	130	143	[154, 154]
24	120	160	178	[188, 188]
28	140	190	213	[220, 222]
30	150	205	230	[237, 237]

the discrete-time model respectively, then we set the number of voice channels N as:

$$N = \min(N_{LD}, N_{DT})$$

For link speeds less than 8 DS0 channels the short term packet correlations predicted by the discrete-time model is an over-estimation of the actual correlation predicted by the simulation. This is due to the assumption of batch arrivals which is quite restrictive for small link speeds. For large link speeds, the large deviations model over-estimates the correlation in the vocoder packet streams since the framing will smooth the traffic input to the multiplexer. If the link speed is larger than 10 DS0 channels the whole (2 *KB*) buffer can be emptied in less than the frame period of 20 *msec*. Hence, the buffer 'forgets' about the last packet of a source when the next packet of the same source is arriving. Consequently, the correlation has only a small, if at all, effect.

The statistical multiplexing gain resulting from dimensioning the links using the two models is depicted in Figure 10. The multiplexing gain is defined as the ratio of the maximum number of voice channels that can be supported and the minimum number of voice channels supported when applying peak bit rate allocation (circuit switching). The latter value is obtained by dividing the link capacity by the maximum voice channel bit rate. Statistical multiplexing provides a reasonable gain over peak rate allocation for rates larger than 8×64 *kbps* and the multiplexing gain reaches 1.5 at 30×64 *kbps*. The multiplexing gain smaller than 1 for low channel rates cannot be interpreted as advantage of peak rate allocation over statistical multiplexing. With peak rate allocation the calculation of the maximum number of supported voice channels only considers packet loss probability but not packet delay. To avoid exceeding the delay limit, all packets of a batch must be transmitted within 4 *msec*, which is not possible if the batch is large (i.e. the number of supported sources is large) and the channel rate is low.

Figure 11 illustrates the influence of the buffer size on the quality of service as expressed by the packet loss probability for the discrete-time model. The link speed is 30×64 *kbps* which is equivalent to 300 data-units per slot (the data-unit is 8 bits and the time-slot is 1.25 *msec*, thus the link speed is equivalent to $30 \times 64 \times 1.25/8 = 300$ data-unit per slot). Consequently, a buffer smaller than 300 data-units is emptied in each slot; each batch of packets finds an empty buffer. This explains the bends at $B = 300$ data-units for 200, 225, 250, and 300 channels. For buffers larger than 300 data-units a queue builds

up in the buffer and the typical exponential tail can be observed. The 100 channels curve exhibiting no bend is due to the fact that the length of the maximum possible batch is 320 data-units. Similarly, Figure 12 illustrates the influence of the buffer size on the packet loss probability as predicted by the large deviations model.

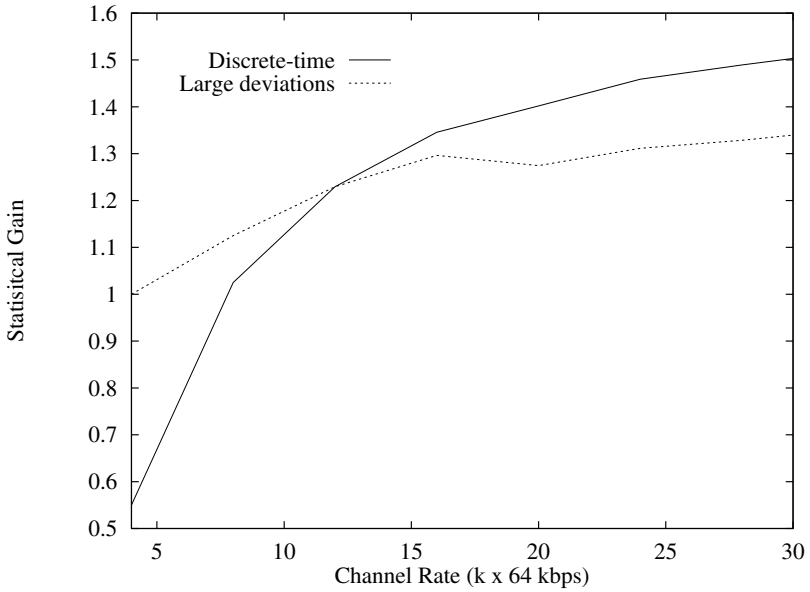

Figure 10. Statistical Multiplex Gain

6. CONCLUSIONS AND OUTLOOK

CDMA voice encoding and statistical multiplexing provide a means for making the most out of the communication links interconnecting the BTS and BSC. We introduced a methodology based on large deviations theory and discrete-time analysis that can be used to study the performance of these links. In the large deviations model, we have introduced some approximations of the original problem and ignored the framing imposed on the voice packets in the base station. The resulting system is easy to analyze and provides insight on the system behavior due to the correlation in the voice process. Also, for the discret-time system, we assumed a batch arrival process that is an upper-bound on the actual arrival stream.

Comparing the results of the large deviations and discrete-time analytical models with the simulation shows the accuracy of the proposed models. We have seen that for almost all practical link speeds, statistical multiplexing provides a reasonable gain over peak rate allocation (or circuit switching).

The obvious extension of this work would be to study the integration of wireless data and voice in CDMA systems. Data services and Internet access will drive the evolution of mobile services to third-generation mobile within the prospects of IMT-2000. Another important area is the ATM transport for wireless voice and data. The performance impacts

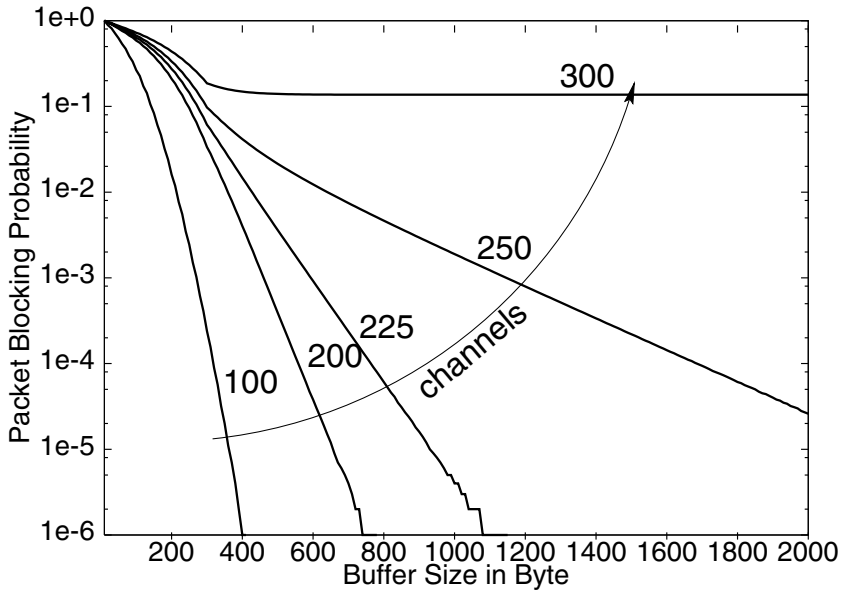

Figure 11. Influence of Buffer Size [Discrete-time]

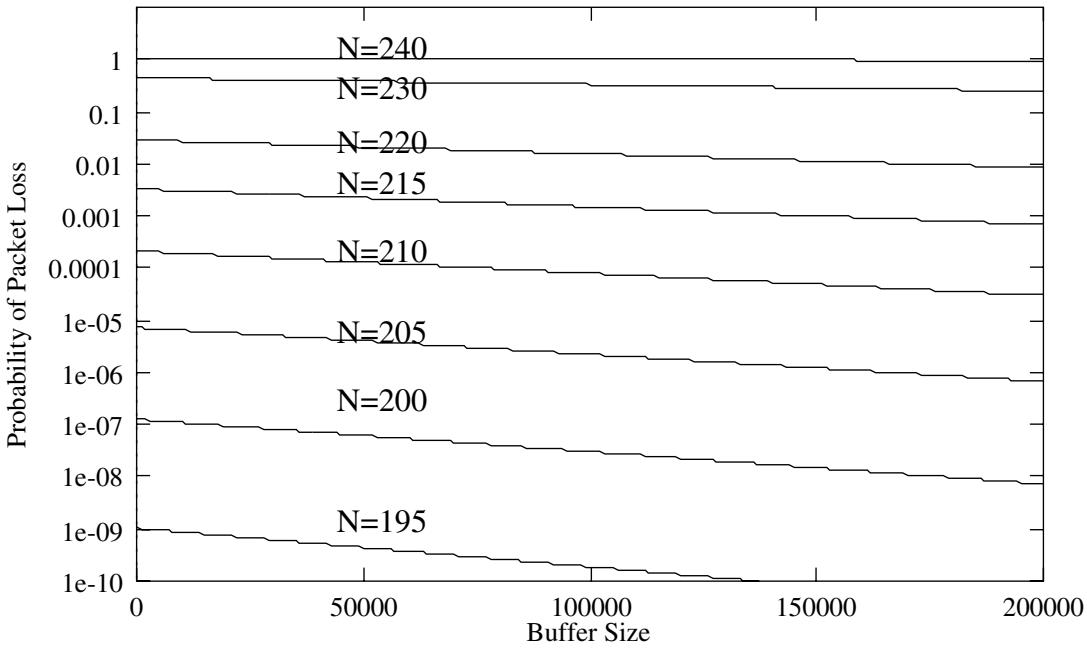

Figure 12. Influence of Buffer Size [Large deviations]

of packing the voice packets into ATM mini-cells as defined by ATM Adaptation Layer 2 (AAL2) [2] needs to be explored.

Appendix A

Consider K classes of sources where each class k is comprised of N_k sources characterized by $(Q^{(k)}, R^{(k)})$ where $Q^{(k)}(i,j), i \neq j$ is the rate at which a source in state i jumps to state j, $i,j = 1,2,\cdots,n_k$ and $Q^{(k)}(i,i) = -\sum_{j\neq i} Q^{(k)}(i,j)$. The vector $R^{(k)} = (R_1^{(k)}, R_2^{(k)}, \cdots, R_{n_k}^{(k)})$, where $R_i^{(k)}$ is the rate at which a source in state i gener-

ates traffic. Let $W_{k,i}(t)$ be the rate of traffic generated by source i of class k at time t. Let $\{W_{k,i}\}$ be the stationary distribution of $W_{k,i}(t)$. Let $W = \sum_k \sum_i W_{k,i}(t)$, then in a buffer-less system, loss occurs when $W > C$. Therefore, we estimate $P(W > C)$. Let $\pi^{(k)}$ be the stationary probability vector of class k sources associated with the matrix $Q^{(k)}$, then $W_{k,i}$ has the moment generating function

$$M_k(s) = E(e^{sW_{k,i}}) = \sum_i \pi_i^{(k)} e^{sR_i^{(k)}} \tag{19}$$

Chernoff's theorem states that

$$\log P(W > C) \leq -F(s^*) \tag{20}$$

where $F(s) = sC - \sum_k N_k \log M_k(s)$ and $F(s^*) = Sup_{s \geq 0} F(s)$, from which we get

$$A = exp(-F(s^*)) \tag{21}$$

The dominant eigenvalue of the system is obtained as follows. Consider a generic source (Q, R) and let $R_d = diag(R_1, R_2, \cdots, R_n)$, for z real and negative, the matrix $(R_d - \frac{1}{z}Q)$ is an irreducible matrix with non-negative off-diagonal elements. This matrix has a real eigenvalue, called the maximum real eigenvalue (MRE) that is greater than the real part of all the other eigenvalues. Let $g(z) = MRE(R_d - \frac{1}{z}Q)$. For K classes of heterogeneous sources, the dominant eigenvalue z is obtained by solving the equation:

$$\sum_{k=1}^{K} N_k g^{(k)}(z) = C \tag{22}$$

where $g(z) = MRE(R_d - \frac{1}{z}Q)$ (see [7] for details).

For discrete-time Markov sources, the calculation for A does not change, however z is calculated differently. Let a source be characterized by (P, R) where P is the probability transition matrix governing the underlying Markov chain of the source. Let $\mu(z)$ be the Perron-Frobenius eigenvalue of the matrix $e^{-zR_d}P$. For real z, the matrix $e^{-zR_d}P$ is non-negative and irreducible, hence its PF-eigenvalue is real, positive, and simple. We thus have z as the unique solution of

$$\sum_{k=1}^{K} N_k \{-\frac{\log \mu^{(k)}(z)}{z}\} = C \tag{23}$$

Appendix B

For a 2-state Markov fluid source described by (Q, R) where $Q = \begin{bmatrix} -\alpha_1 & \alpha_1 \\ \alpha_2 & -\alpha_2 \end{bmatrix}$ and $R = (\hat{r}_1, \hat{r}_2)$, the value of the dominant eigenvalue of the system is given by:

$$z = \frac{\hat{c}(\alpha_1 + \alpha_2) - (\alpha_1 \hat{r}_2 + \alpha_2 \hat{r}_1)}{\hat{c}^2 - \hat{c}(\hat{r}_1 + \hat{r}_2) + \hat{r}_1 \hat{r}_2}$$

where $\alpha_i = \frac{1}{t_i}$, $\hat{c} = \frac{C}{N}$, and $\hat{r}_i = \frac{r_i}{C}$. This can be verified by finding $g(z) = MRE(R_d - \frac{1}{z}Q)$ and solving for $g(z) = \hat{c}$.

References

[1] M. H. Ackroyd. Computing the waiting time distribution for the G/G/1 queue by signal processing methods. *IEEE Transactions on Communications*, COM-28(1):52–58, January 1980.

[2] ATM Forum. *AF-VTOA-0113.000, ATM Trunking Using AAL2 for Narrowband Services*, February 1999.

[3] A. Elwalid, D. Heyman, T. V. Lakshman, D. Mitra, and A. Weiss. Fundamental Bounds and Approximations for ATM Multiplexers with Applications to Video Teleconferencing. *IEEE Journal on Selected Areas in Communications*, 13:1004–1016, 1995.

[4] A. I. Elwalid and D. Mitra. Effective Bandwidth of General Markovian Traffic Sources and Admission Control of High Speed Networks. *IEEE Transactions on Networking*, 1:329–343, 1993.

[5] J. D. Gibson, editor. *The Mobile Communications Handbook*. IEEE Press / CRC Press, College Station, TX, 1996.

[6] F. Hübner and P. Tran-Gia. Quasi–stationary analysis of a finite capacity asynchronous multiplexer with modulated deterministic input. In *Proc. ITC 13*, pages 723–729, 1991.

[7] G. Kesidis, J. Walrand, and C.-S. Chang. Effective Bandwidths for Multiclass Markov Fluids and Other ATM Sources. *IEEE Transactions on Networking*, 1:424–428, 1993.

[8] D. R. Manfield and P. Tran-Gia. Analysis of a finite storage system with batch input arising out of message packetization. *IEEE Transactions on Communications*, COM-30(3):456–463, 1982.

[9] G. Ramamurthy and B. Sengupta. Delay analysis of a packet voice multiplexer by the $\sum D_i/D/1$ queue. *IEEE Transactions on Communications*, COM-39(7):1107–1114, 1991.

[10] Telecommunications Industry Association. *TIA/EIA/IS-95. Mobile Station – Base Station Compatibility Standard for Dual Mode Wideband Spread Spectrum Cellular Systems*, July 1993.

[11] P. Tran-Gia. Discrete-time analysis for the interdeparture distribution of GI/G/1 queues. In O. J. Boxma, J. W. Cohen, and H. C. Tijms, editors, *Teletraffic Analysis and Computer Performance Evaluation*, pages 341–357. North–Holland, Amsterdam, 1986.

[12] P. Tran-Gia. Discrete-time analysis technique and application to usage parameter control modelling in ATM systems. In *Proc. 8th Australian Teletraffic Seminar*, Melbourne, Australia, December 1993.

[13] P. Tran-Gia and H. Ahmadi. Analysis of a discrete-time $G^{[X]}/D/1-S$ queueing system with applications in packet–switching systems. In *Proc. INFOCOM '88*, pages 861–870, 1988.

[14] S. Tsakiridou and I. Stavrakakis. Mean delay analysis of a statistical multiplexer with batch arrivals processes — a generalization to Viterbi's formula. *Performance Evaluation*, 25:1–15, 1996.

[15] J. Van Ommeren. Simple approximations for the batch-arrival $M^X/G/1$ queue. *Operations Research*, 38(4):678–685, 1990.

[16] A. M. Viterbi. *CDMA Principles of Spread Spectrum Communication*. Addison–Wesley, Reading, MA, 1995.

[17] A. M. Viterbi and A. J. Viterbi. Erlang capacity of a power controlled CDMA system. *IEEE Journal on Selected Areas in Communications*, 11(6):892–899, August 1993.

[18] A. Weiss. An introduction to large deviations for communication networks. *IEEE Journal on Selected Areas in Communications*, 13(6):938–952, 1995.

[19] W. Whitt. Tail Probabilities with Statistical Multiplexing and Effective Bandwidth for Multi-Class Queues. *Telecommunications Systems*, 2:71–107, 1993.

[20] Y. Zhao. Analysis of the $GI^X/M/c$ model. *Queueing Systems*, 15:347–364, 1994.