Impacts of Data Call Characteristics on Multi-Service CDMA System Capacity

Yujing Wu, Carey Williamson

Department of Computer Science, University of Calgary,
2500 University Drive NW, Calgary, AB, Canada T2N1N4

Abstract

The capacity of multi-service Code-Division Multiple Access (CDMA) systems has been extensively studied in the literature. However, few studies address the fundamental issue of how the stochastic properties of non-Poisson data traffic affect the system capacity. This paper studies a CDMA system supporting voice and data traffic. Results show that increased variability in the data call arrival process decreases the system capacity, while increased variability in data call holding times increases the system capacity. The extent of these effects depends on other system parameters, such as transmission rates and communication quality requirements. These observations motivate a simple buffer-based resource management scheme that enhances the system capacity in the presence of high-variability data traffic, providing controllable performance tradeoffs between voice and data calls. This study uses both simulation and theoretical analysis, which is based on a Markov Regenerative Process (MRGP) model.

Key words: CDMA, Capacity Planning, Loss System, Markov Regenerative Process, Variability of Stochastic Processes

1 Introduction

Mobile service providers have recently made substantial investments to deploy 3G CDMA networks in North America. For example, CDMA2000 1xRTT networks have been available for several years, while data-optimized CDMA2000 1xEV-DO networks are being rolled out now.

Email address: {ywu,carey}@cpsc.ucalgary.ca
(Yujing Wu, Carey Williamson).

Despite the large-scale and rapid deployment of these networks, capacity planning for CDMA data networks is not well understood. Current planning only supports the availability of high-speed data services. However, given the growth trends in data traffic volume and the inherent scarcity of radio spectrum resources, better capacity planning tools are required.

In voice-only 2G cellular networks, the Erlang B formula has been used for capacity planning for many years. The success of this simple tool lies in the fact that voice traffic is well-modeled by a Poisson process. Internet-like data traffic, however, exhibits high variability over many timescales, which is not conducive to Poisson-based modeling [12]. This makes capacity planning in 3G networks a challenging task.

Current research on multi-service CDMA network capacity falls into two main categories. The first category is the analytical approach [1,2,5,6,8]. In most of these models, data traffic differs from voice traffic only in the transmission rate and quality of service (QoS) requirements, not in the stochastic traffic behavior. The Poisson assumption is still made for the data call arrival process. The second category is the simulation approach. Traffic generators are integrated into the simulation environment, which models the CDMA protocol stack and the radio channel. Traffic models are used to emulate data applications, such as WWW, email, and WAP. Simulation results show the relationship between offered traffic load and required radio resources. These two types of studies, however, have not addressed the fundamental issue of how the stochastic properties of non-Poisson data traffic affect the overall system capacity.

In this paper, we study a multi-service CDMA system supporting voice and non-Poisson data traffic over dedicated channels. Our results show that the variability of the data call arrival process adversely affects the system capacity, while the variability of data call holding times increases the system capacity. The extent of these effects depends on system configurations, such as traffic mix, transmission rates, and QoS requirements. Based on these observations, we propose and evaluate a simple buffer-based resource management scheme that effectively increases the system capacity in the presence of high-variability data traffic. Our study is carried out by simulation and theoretical analysis based on a Markov Regenerative Process (MRGP) model.

The rest of the paper is structured as follows. Section 2 introduces the key concepts of CDMA network capacity and the system model. Section 3 presents the analytic study for a simple system. Section 4 is devoted to the simulation study for a more general model. Section 5 presents and evaluates the data call buffering scheme. Finally, Section 6 concludes the paper.

2 Capacity and System Model

2.1 CDMA Network Capacity

A CDMA network consists of base stations (BS) each providing service to mobile stations (MS). Transmissions from the home BS to a MS traverse the *forward link*, while transmissions from a MS to the home BS traverse the *reverse link*.

One of the principal characteristics of a CDMA network is that the capacity in each direction depends on the total interference experienced by an MS or the BS. The interference level depends on the cell layout, the MS spatial distribution, and the radio propagation characteristics. To facilitate capacity analysis, an approximate model for the interference is often employed. The approximation is based on the following assumptions: the network cells are homogeneous; the MSs are uniformly placed within the cell; Rayleigh fading is ignored; and shadow fading is modeled by the lognormal distribution.

This paper uses two types of capacity measures. The first one, referred to as the *capacity bound*, is the maximum number of concurrent users that the system can support at a specified transmission rate and communication quality E_b/N_o (the ratio of bit energy to noise plus interference density). The bound is calculated assuming a fixed number of active MSs in each traffic class. The second type of measure, referred to as *Erlang capacity*, is the average traffic load that can be supported at a given communication quality and service availability probability. The Erlang capacity is evaluated in a dynamic user scenario, with calls initiated and terminated according to stochastic processes. Next we summarize the capacity bounds for the reverse and forward links, and then discuss the relationship between the Erlang capacity and a loss system.

We introduce common notation for the analysis of both link directions. Assume that the system supports $N > 1$ traffic classes. For the i-th class, let n_i be the number of active users, r_i be the transmission bit rate, α_i be the traffic activity factor, and η_i be the required E_b/N_o. Let W be the spreading bandwidth, and let $Q^{-1}(x)$ denote the inverse Q-function defined by $Q(x) = \int_x^\infty (1/\sqrt{2\pi}e^{-y^2/2})dy$.

The reverse-link capacity bound [2,8] is given by:

$$\sum_{i=0}^{N-1} n_i r_i \alpha_i \eta_i \leq \left\langle \frac{1}{1+f} \right\rangle W \times 10^{Q^{-1}(\beta)\sigma_x/10.0 - a\sigma_x^2} \tag{1}$$

where $\langle \frac{1}{1+f} \rangle$ is the average frequency reuse factor, σ_x is the standard deviation of the received signal to noise power ratio (SIR) in dB, a is a constant

with typical value 0.012, and β is the required system reliability such that $Pr\left((E_b/N_o)_i \geq \eta_i\right) = \beta$. The term $10^{Q^{-1}(\beta)\sigma_x/10.0 - a\sigma_x^2}$ reflects the impact of the power control error. The standard deviation σ_x has a typical value between 0.3 dB and 2 dB.

Next, we express the forward-link capacity bound based on the work by Lee *et al.* [10]. Let P_i represent the BS transmission power for a class i MS at the cell edge, and P_{tot} be the total transmission power of the home BS. The E_b/N_o for a class i MS at the edge of the cell must satisfy:

$$\left(\frac{E_b}{N_o}\right)_i = \frac{W}{r_i}\frac{P_i}{K_f P_{tot}} \geq \eta_i \times 10^{-\frac{Q^{-1}(\beta)\sigma_y}{10}} \ , \tag{2}$$

where σ_y is the standard deviation of the lognormally distributed propagation loss in dB, β is the required system reliability and K_f is the ratio of the received power at the edge MS from other cells and from the home cell. Typically, $K_f = 2.778$. The transmission power at the home BS satisfies: $P_{tot} = K_t \sum_{i=0}^{N-1} n_i \alpha_i P_i$, where the average forward-link power factor $K_t < 1$ accounts for the fact that not all MSs are located at the cell boundary (i.e., closer MSs require less BS transmission power). We assume that P_{tot} is not larger than the BS power limit [10]. The maximum number of users are supported when constraint (2) achieves equality. The forward-link capacity bound is given by:

$$\sum_{i=0}^{N-1} n_i r_i \alpha_i \eta_i \leq \frac{W}{K_f K_t} 10^{Q^{-1}(\beta)\sigma_y/10.0} \ . \tag{3}$$

The capacity bounds in each direction (1) and (3) have the common form:

$$\sum_{i=0}^{N-1} n_i w_i \leq \xi \ , \tag{4}$$

where $\xi > 0$, $w_i > 0$, and $n_i \geq 0$, $\forall i = 0, \ldots, N-1$. The right hand side of (4) can be viewed as the total system effective bandwidth and w_i as the effective bandwidth required by a call of the i-th class. The effective bandwidth depends on the assigned transmission rate, traffic activity factor, and required E_b/N_o. An admissible state $(n_0, \ldots, n_{N-1})$ satisfies the bound (4). The admission region Ω is the set of all admissible states. A call arrival that would move the system state out of the admissible region is blocked and cleared from the system. Otherwise, the call is accepted, and it consumes its effective bandwidth for the holding time of the call. The maximum tolerable blocking probability of the system determines the average traffic load that can be accommodated. This probability can be a maximum aggregate blocking probability, or a vector where the ith element is the maximum blocking probability for class i. Erlang capacity can be expressed as an arrival rate vector producing the maximum tolerable blocking probability.

By using the capacity bounds, we convert the capacity analysis in either direction of a CDMA system to the study of a loss system, where multiple traffic classes completely share the resources subject to the admission requirement (4). Modeling a CDMA cell as a loss system is not new. Several authors have used this approach especially for the reverse-link capacity analysis [1,2,6,7]. Many of these studies assume that the call arrival process for each class is Poisson, and then model the system as a multi-dimensional continuous time Markov chain.

A Poisson arrival process may adequately model data traffic at the session level [12]. However, CDMA data calls do not necessarily correspond to such sessions. For example, consider Web browsing in a CDMA2000 1xRTT system. A down-link data call may transmit a Web page or several successive Web objects. The arrival process of data calls is not that of browsing sessions. Since data traffic exhibits high variability over many timescales, it is questionable to use the Poisson model for the traffic at levels other than the session level. In this paper, we assess the impact of non-Poisson data traffic on the CDMA system capacity.

2.2 System Model

There is a correspondence between the Erlang capacity of a multi-service CDMA system and a loss system. As a consequence, we study a loss system whose parameters are configured based on the capacity bounds of the CDMA system. Specifically, the system is characterized as follows:

- The system supports two types of calls: data (class 0) and voice (class 1). They have different transmission rates r_i and E_b/N_o requirements η_i.
- Data calls are generated as a renewal process with rate λ_0. The inter-arrival time X_0 has a general cumulative distribution function $G(x)$. The number of bits Y_0 transmitted by a data call, referred to as the workload size, is an i.i.d. random variable with a general distribution.
- Voice calls are generated according to a Poisson process with rate λ_1, and have exponentially distributed workload sizes Y_1.
- Once a call is accepted into the system, it remains in the system for duration $Y_i/(r_i\alpha_i)$, where α_i is the traffic activity factor. The mean service rate for class i calls is μ_i, where $\mu_i = r_i\alpha_i/E[Y_i]$.

Table 1 lists the model parameters for the forward link. Substituting the corresponding values into (3), the capacity bound is obtained. All simulations and numerical studies use the values listed in Table 1 unless otherwise specified. The mean workload size of voice calls is based on a mean call holding time of 120 seconds, a typical value in cellular networks. The mean workload size of

Table 1
Model Parameters

System parameters	Value	
Spreading bandwidth W	1.2288 MHz	
K_f	2.778	
K_t	0.35	
σ_y	0	
Traffic parameters	Data calls (Class 0)	Voice calls (Class 1)
Transmission rate r_i (kbps)	100	9.6
Traffic activity factor α_i	1.0	0.5
E_b/N_o requirement η_i (dB)	3	4
Arrival rate ratio $\frac{\lambda_i}{\lambda_0+\lambda_1}$	20%	80%
Mean workload size $E[Y_i]$ (bits)	440,000	576,000
Target blocking (case I)	aggregate blocking 2%	
Target blocking (case II)	5%	2%

data calls is based on a mean Web page size of about 50 KB. We assume that 80% of the calls offered to the system are voice calls, and 20% are data calls. Since this ratio is fixed, the maximum aggregate arrival rate denoted by Λ determines the Erlang capacity as $[0.2\Lambda, 0.8\Lambda]$. We use this maximum aggregate rate to indicate the system capacity, and all the following studies are based on this performance measure. We consider two different blocking requirements as indicated in Table 1. The first case concerns the overall blocking rate while the second case considers class-specific blocking rates.

Our study focuses on the impact of the variability of inter-arrival times X_0 and workload sizes Y_0 of data calls. *Coefficient of Variation* (CV) is a measure of variability for a random variable. CV is defined as the ratio of the standard deviation to the mean. Let c_a (arrival) and c_s (size) denote the CV of X_0 and Y_0, respectively. Our studies illustrate the relationship between the system capacity Λ and these two parameters. We use second-order hyperexponential distributions to model interarrival times and workload sizes with CV > 1.

3 Theoretical Analysis via MRGP

We use a Markov Regenerative Process (MRGP) to study the system described in Section 2.2. However, there is a restriction on the data call model: the

holding time must be exponentially distributed. We leave the study of a more general system to Section 4, where simulation is used.

Trivedi *et al.* [3] developed solution methods for MRGPs, and applied them to the performance and reliability analysis of various computer systems. Our study is another effort along this line. For the details of MRGP theory and solution techniques, the reader may refer to [4,9]. We briefly introduce the MRGP technical background here, and then describe the MRGP model for a simple CDMA system.

3.1 Introduction to MRGP

In a MRGP, there exist time points where the process satisfies the Markov property [3]. These time points are referred to as regeneration points. The stochastic evolution between two successive regeneration points depends only on the state at regeneration, not on the evolution before regeneration. Furthermore, due to the time homogeneity of the embedded Markov renewal process, the evolution of the MRGP becomes a probabilistic replica after each regeneration. The key concepts of MRGP are given in the following two definitions [3].

Definition 3.1 *A sequence of bivariate random variables* $\{(u_n, t_n), n \geq 0\}$ *is called a Markov renewal sequence if: (I)* $t_0 = 0, t_{n+1} \geq t_n$*;* $u_n \in \Psi \subset \Omega$*, where* Ω *is a countable set represented by* $\{0, 1, 2, \ldots\}$*; and (II)* $\forall n \geq 0$*,*

$$\begin{aligned} &P\{u_{n+1} = j, t_{n+1} - t_n \leq t | u_n = i, t_n, \ldots, u_0, t_0\} \\ &= P\{u_{n+1} = j, t_{n+1} - t_n \leq t | u_n = i\} \qquad \textit{(Markov property)} \\ &= P\{u_1 = j, t_1 \leq t | u_0 = i\} \qquad \textit{(time homogeneity).} \end{aligned} \tag{5}$$

Definition 3.2 *A stochastic process* $\{z(t), t \geq 0\}$ *on* Ω *is called a Markov regenerative process if there exists a Markov renewal sequence* $\{(u_n; t_n), n \geq 0\}$ *of random variables such that all conditional finite-dimensional distributions of* $\{z(t_n + t); t \geq 0\}$ *given* $\{z(v); 0 \leq v \leq t_n; u_n = i\}$ *are the same as those of* $\{z(t), t \geq 0\}$ *given* $u_0 = i$*,* $i \in \Psi \subset \Omega$*.*

The above definition implies that $\{z(t_n^+), n \geq 0\}$ or $\{z(t_n^-), n \geq 0\}$ is an embedded Markov chain (EMC), and that t_n is a regeneration point of $z(t)$.

The global kernel $K(t)$ and the local kernel $E(t)$ determine the evolution of a MRGP. Kernel $K(t)$ describes the behavior of $z(t)$ at the regeneration instants while kernel $E(t)$ describes it between two consecutive regeneration instants. Entries of matrix $K(t) = [K_{i;j}(t)]$, $i, j \in \Psi$, are given by (5). Matrix $K(\infty)$ is the one-step transition probability matrix of the EMC. Entries of matrix $E(t) = [E_{i;j}(t)]$, $i \in \Psi$, $j \in \Omega$, are given by $E_{i;j}(t) = P\{z(t) = j, t_1 > t | u_0 =$

$i\}$. If local state transitions (between two consecutive regeneration points) are governed by a homogenous continuous time Markov chain (CTMC), the MRGP has a subordinate CTMC.

Knowledge of the kernels allows us to obtain three new variables, which lead to the solution of the steady state probabilities of the MRGP. The first variable $\alpha_{i;j}$ is given by

$$\alpha_{i;j} = \int_0^\infty E_{i;j}(\tau)d\tau \ , \quad i \in \Psi, j \in \Omega \ . \tag{6}$$

This variable is the mean time that $z(t)$ spends in state j between two successive regeneration instants, given that it started in state i after the last regeneration. The second one is defined as $\beta_i = E[t_1|u_0 = i]$, $i \in \Psi$. It is the mean duration of the next state of the renewal sequence given that the current state is i. The third variable is the steady state probability vector $\vec{\nu} = (\nu_k)$ of the EMC, which satisfies:

$$\vec{\nu} = \vec{\nu}K(\infty), \qquad \sum_{k\in\Psi} \nu_k = 1. \tag{7}$$

Theorem 1 in the book by Kulkarni [9] gives the steady state probabilities of the MRGP based on $\alpha_{i;j}$, β_i and $\vec{\nu}$.

3.2 MRGP Analysis of a CDMA System

3.2.1 MRGP Model

Denote by (i, j) a state with i data calls and j voice calls in the system. According to (4), the admission region is given by $i \times w + j \leq l$, where w and l are the data call bandwidth and total system bandwidth normalized to the voice call bandwidth, respectively. Figure 1 depicts the state transition diagram. A dotted arc represents a transition triggered by a data call arrival.

Let $\lfloor\cdot\rfloor$ and $|\cdot|$ denote the floor of a number and the cardinality of a set, respectively. Define $\omega_d = \{0, 1, \ldots, \lfloor l/w \rfloor\}$. We list important state sets:

- Ω denotes the set of all feasible states. $\Omega = \{(i, j)|i \in \omega_d, j = 0, 1, \ldots, \lfloor l - iw\rfloor\}$.
- S_i denotes the set of states with exactly i data calls. $S_i = \{(i, j)|j = 0, 1, \ldots, \lfloor l - iw\rfloor\}$, $|S_i| = \lfloor l - iw\rfloor + 1$, $\forall i \in \omega_d$.
- Ω_i denotes the set of states with at most i data calls. $\Omega_i = \bigcup_{j=0}^{i} S_j$, $|\Omega_i| = \sum_{m=0}^{i} \lfloor l - mw\rfloor + i + 1$, $\forall i \in \omega_d$.
- Ω_{B0} and Ω_{B1} denote the sets of states that block data and voice calls, respectively. For example, $\Omega_{B1} = \{(i, j)|i \in \omega_d, j = \lfloor l - iw\rfloor\}$.

Let $\{z(t), t \geq 0\}$ denote the two-dimensional state process on Ω. Let $t_0 = 0$ and

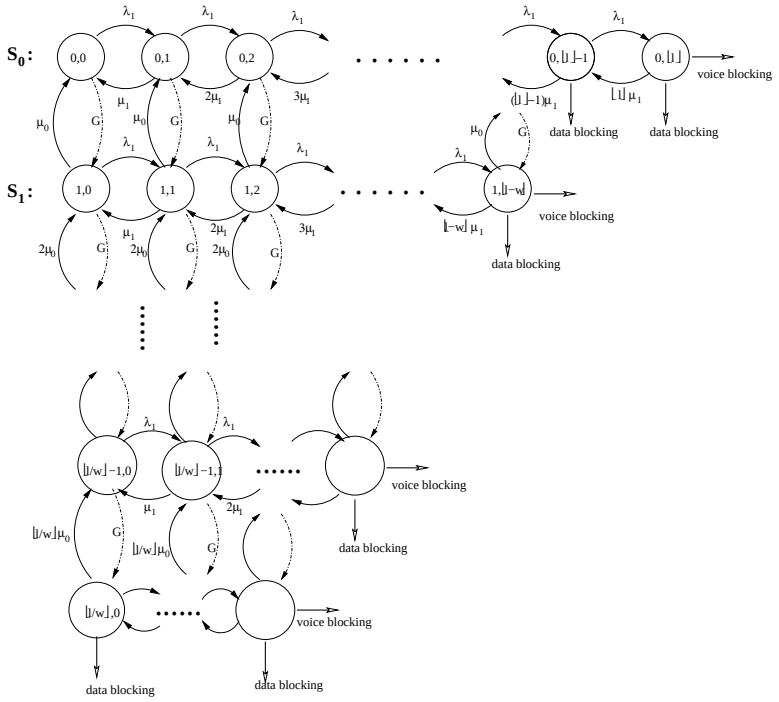


Fig. 1. MRGP state transition diagram

define t_n $(n > 0)$ as the arriving instant of the n-th data call. By Definition 3.1, $\{(z(t_n^+), t_n), n \geq 0\}$ is a Markov renewal sequence since it satisfies the time homogeneous Markovian properties. Furthermore, $\{z(t), t \geq 0\}$ is a MRGP associated with this sequence by Definition 3.2. The state space of the Markov renewal sequence is $\Psi = \Omega - \{(0, j)|j = 0, 1, \ldots, \lfloor l - w \rfloor\}$. The MRGP has a subordinate CTMC since local state transitions are caused by exponentially distributed events (voice call arrivals/departures, data call departures).

3.2.2 Solution Method

We first obtain the transient probabilities of the subordinate CTMC. We then derive the MRGP kernels, and finally calculate the MRGP stationary probabilities.

3.2.2.1 Subordinate CTMC: Assume that the state is $(i, j) \in \Psi$ just after a data call arrives. Before the arrival of the next data call, the state

evolves as a CTMC on Ω_i with the infinitesimal matrix Q_i given by

$$Q_i = \begin{bmatrix} B_0 & & & & \\ A_1 & B_1 & & & \\ & A_2 & B_2 & & \\ & & \ddots & \ddots & \\ & & & A_i & B_i \end{bmatrix} , \quad \forall i \in \omega_d , \tag{8}$$

where A_k ($k \in \omega_d - \{0\}$), a $|S_k| \times |S_{k-1}|$ block matrix, refers to the departure of a data call when there are k data calls in the system; B_k ($k \in \omega_d$), a $|S_k| \times |S_k|$ matrix, refers to no change in the number of data calls when there are k data calls in the system.

Entry $A_k(m,n)$ of A_k is the transition rate from state (k,m) to state $(k-1,n)$ before the arrival of the next data call. The transition is due to exponentially distributed events. Therefore, $A_k(m,n) = k\mu_0$ if $m = n$; otherwise $A_k(m,n) = 0$. Then $\forall k \in \omega_d - \{0\}$,

$$A_k = [k\mu_0 \mathbf{I}_k \quad \mathbf{0}] ,$$

where $\mathbf{I}_k$ is a $|S_k| \times |S_k|$ identity matrix and $\mathbf{0}$ is a $|S_k| \times (|S_{k-1}| - |S_k|)$ block matrix with all zeros. Entry $B_k(m,n)$ of matrix B_k is the transition rate from state (k,m) to (k,n) before the arrival of the next data call. When $m \neq n$, the transition is due to the arrival and departure of voice calls. The rate can be easily determined. When $m = n$, $B_k(m,n)$ is the rate of staying in state (k,m), which needs to be calculated. Note that matrix Q_i has the property $Q_i \mathbf{e} = \mathbf{0}$, where $\mathbf{e}$ is a column vector with all ones. Utilizing this property and the expression for A_k, we get B_k as follows: $\forall k \in \omega_d$,

$$B_k = \begin{bmatrix} -\lambda_1 & \lambda_1 & & & & \\ \mu_1 & -\lambda_1 - \mu_1 & \lambda_1 & & & \\ & 2\mu_1 & -\lambda_1 - 2\mu_1 & \lambda_1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \lfloor l - kw \rfloor \mu_1 & -\lfloor l - kw \rfloor \mu_1 \end{bmatrix} - k\mu_0 \mathbf{I}_k . \tag{9}$$

With Q_i known, we can obtain the transient probabilities of the subordinate CTMC. Let $P_{(i,j);(i',j')}(t)$, $(i',j') \in \Omega_i$, be the probability that the CTMC will be in state (i',j') at time t given that it was in state (i,j) initially. Define

$$\vec{P}_{(i,j)}(t) = \left[\vec{P}_{(i,j);S_0}(t) \quad \vec{P}_{(i,j);S_1}(t) \quad \dots \quad \vec{P}_{(i,j);S_i}(t) \right] , \tag{10}$$

where $\vec{P}_{(i,j);S_k}(t) = \left[P_{(i,j);(k,0)}(t) \quad P_{(i,j);(k,1)}(t) \quad \dots \quad P_{(i,j);(k,\lfloor l-kw \rfloor)}(t) \right]$. We

have, $\forall (i,j) \in \Psi$:

$$\frac{d}{dt}\vec{P}_{(i,j)}(t) = \vec{P}_{(i,j)}(t)Q_i \ , \tag{11}$$

with the initial condition: $P_{(i,j);(i',j')}(0) = 1$ if $(i',j') = (i,j)$; $P_{(i,j);(i',j')}(0) = 0$ if $(i',j') \neq (i,j)$. The transient solution is $\vec{P}_{(i,j)}(t) = \vec{P}_{(i,j)}(0) \times e^{Q_i t}$.

3.2.2.2 Kernels and Performance Measures: Entries of global kernel $K(t)$ are defined by (5). In this specific system, the entry $K_{(i,j);(i',j')}(t)$ is the probability that the system will be in state (i',j') immediately after the next data call arrives at time t, given that the system was in state (i,j) just after the previous data call arrived at time 0. Depending on the new state, the probability has different expressions as follows: $\forall (i,j) \in \Psi$, $\forall (i',j') \in \Psi$,

$$K_{(i,j);(i',j')}(t) = \begin{cases} \int_0^t P_{(i,j);(i',j')}(\tau) dG(\tau) \ , & i' = 0 \ , \\ \int_0^t P_{(i,j);(i'-1,j')}(\tau) dG(\tau) \ , & 1 \leq i' \leq i, (i',j') \notin \Omega_{B0} \ , \\ \int_0^t \left[P_{(i,j);(i'-1,j')}(\tau) + P_{(i,j);(i',j')}(\tau) \right] dG(\tau) \ , & 1 \leq i' \leq i, (i',j') \in \Omega_{B0} \ , \\ \int_0^t P_{(i,j);(i'-1,j')}(\tau) dG(\tau) \ , & i' = i+1 \ , \\ 0 \ , & \text{otherwise} \ . \end{cases} \tag{12}$$

The entry $E_{(i,j);(i',j')}(t)$ of local kernel $E(t)$ is the probability that the system will be in state (i',j') at time t and the next data call will arrive after t, given that the system was in state (i,j) just after the previous data call arrived at time 0. Thus, $\forall (i,j) \in \Psi$ and $\forall (i',j') \in \Omega$,

$$E_{(i,j);(i',j')}(t) = \begin{cases} P_{(i,j);(i',j')}(t)(1 - G(t)) \ , & (i',j') \in \Omega_i \ , \\ 0, & \text{otherwise} \ . \end{cases} \tag{13}$$

We are ready to obtain the performance measures. Let $s = (i,j) \in \Psi$ and $s' = (i',j') \in \Omega$. Variable $\alpha_{s;s'}$ and the EMC steady state probability vector $\vec{\nu} = (\nu_s)$ are calculated according to (6) and (7), respectively. By Theorem 1 in [9], the steady state probabilities (at an arbitrary time point) follow: $\forall s' \in \Omega$,

$$P\{z(t) = s'\} = \frac{\sum_{s \in \Psi} \nu_s \alpha_{s;s'}}{\sum_{s \in \Psi} \nu_s \beta_s} = \frac{\lambda_0 \sum_{s \in \Psi} \nu_s \alpha_{s;s'}}{\sum_{s \in \Psi} \nu_s} \ . \tag{14}$$

The latter step is due to $\beta_s = 1/\lambda_0$.

The arrival process for voice calls is Poisson. By PASTA (Poisson Arrivals See Time Average), the blocking probability for voice calls is $PB_v = \sum_{s \in \Omega_{B1}} P\{z(t) = s\}$. The blocking probability for data calls PB_d is the probability that the

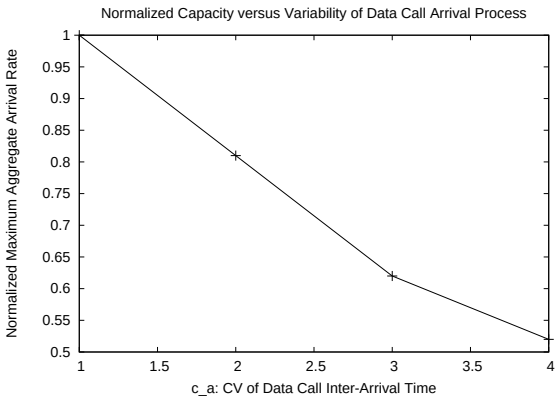


Fig. 2. Normalized capacity $\frac{\Lambda}{\Lambda_p}$ versus CV of data call inter-arrival times (2% aggregate blocking)

system is in the data call blocking set Ω_{B0} just before a data call arrives. $PB_d = \sum_{s' \in \Omega_{B0}} \sum_{s \in \Psi} \nu_s \times \int_0^\infty P_{s;s'}(\tau) dG(\tau)$.

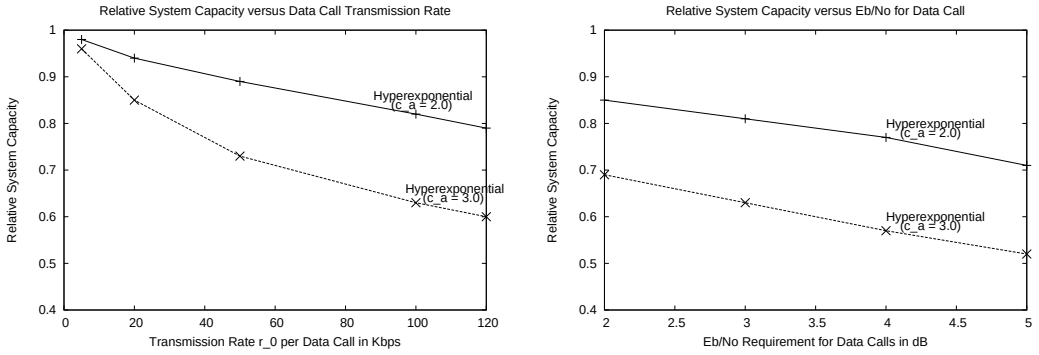
We summarize the procedure to calculate the blocking probabilities.
1. Obtain $P_{s;s'}$, $K(\infty)$ and $E(t)$ according to (11), (12) and (13), respectively;
2. Obtain $\alpha_{s;s'}$ based on (6) and solve (7) for $\vec{\nu}$;
3. Obtain the steady state probability based on (14);
4. Obtain the blocking probabilities of voice and data calls, respectively.

Once the blocking probabilities are calculated, the Erlang capacity or the maximum aggregate arrival rate Λ (given the traffic mix ratio) can be obtained numerically.

3.3 Impact of Data Call Arrival Variability

We numerically study the impact of data call arrival variability on the system capacity Λ. The workload size of data calls is exponentially distributed. A hyperexponential distribution is used for the call inter-arrival times with CV$>$ 1 . The aggregate blocking probability is 2%. The other parameters are listed in Table 1. Let Λ_p represent the maximum aggregate arrival rate when data calls arrive according to a Poisson process. We compare the difference between Λ_p and Λ. Figure 2 plots the normalized capacity $\frac{\Lambda}{\Lambda_p}$ versus CV of data call inter-arrival times. As the variability of inter-arrival times increases, the system capacity decreases. For example, the capacity for CV $c_a = 3.0$ is about two-thirds of that for CV $c_a = 1.0$.

The extent of the capacity reduction caused by the variability of data call arrivals depends on other system parameters, such as the transmission rate and E_b/N_o. Figure 3(a) shows that the normalized capacity $\frac{\Lambda}{\Lambda_p}$ decreases as the transmission rate r_0 for data calls increases. Similarly, Figure 3(b) shows that



(a) (b)

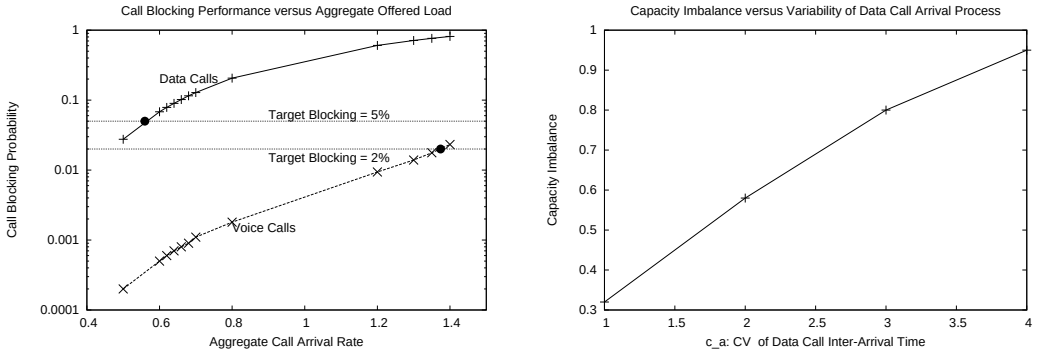
Fig. 3. Factors affecting the capacity reduction: (a) transmission rate; (b) E_b/N_o .

the normalized capacity decreases when the E_b/N_o requirement η_0 increases. These results demonstrate that modeling non-Poisson data traffic as Poisson generates large errors in capacity estimation, particularly for high transmission rates and high E_b/N_o. It is easy to understand these phenomena by checking the previous capacity bounds. As shown in (3) and (1), the effective bandwidth of data traffic is proportional to r_0 and η_0. Deviation from the Poisson process affects the capacity to a larger extent when the weight given to data calls is larger.

Network service providers may specify different blocking rates for different traffic classes. For instance, suppose that the maximum blocking probability for voice calls is 2% while that for data calls is 5%. Let $\Lambda^{(i)}$ denote the maximum aggregate arrival rate with respect to the requirement of class i calls. To meet the requirements of all traffic classes, clearly $\Lambda = \min_{i\in\{0,1\}}\Lambda^{(i)}$. Figure 4(a) plots the per-class blocking probability versus the aggregate arrival rate. The maximum aggregate arrival rate meeting the voice traffic requirement is about 1.4 calls/sec, while it is around 0.55 calls/sec for the data traffic. It is important to balance $\Lambda^{(i)}$ in order to enhance the system capacity Λ.

We measure capacity imbalance between data and voice traffic using $\triangle\Lambda = \max_{i\in\{0,1\}}\Lambda^{(i)} - \min_{i\in\{0,1\}}\Lambda^{(i)}$. The larger $\triangle\Lambda$ indicates that the capacities with respect to individual traffic classes differ more widely; that is, the total system capacity is more severely limited by the restricting class. In our scenarios, the data traffic constricts the total capacity due to its larger effective bandwidth per call. The variability of the data call arrival process exacerbates this issue. Figure 4(b) plots $\triangle\Lambda = \Lambda_1 - \Lambda_0$ versus c_a. Increasing the variability of the data call arrival process causes greater imbalance between voice and data capacities.

The foregoing numerical results clearly show that the variability of the data call arrival process degrades the system performance, decreasing the maximum traffic load that can be accommodated, and making capacities unbalanced



(a) (b)

Fig. 4. (a) Capacity imbalance illustration ($c_a = 3.0$, $c_s = 1.0$); (b) effect of data call arrival variability on $\triangle\Lambda$.

among different traffic classes. These results are consistent with our intuition.

The main limitation of the above MRGP-based analysis is that the holding time of data calls must be exponentially distributed. It is a natural extension to study a system without this restriction. The state diagram of the generalized system has at most two generally distributed timed transitions enabled at any state. According to [13], the state process is still a Markov regenerative process. The difficulty, however, arises from the calculation of the local and global kernels. Studying the generalized system analytically requires further investigation. In the rest of the paper, we use simulation to understand the performance of the generalized system.

4 Simulation Study

We simulate a system where the distribution of data call workload sizes is not necessarily exponential. Three distributions are considered: Constant ($c_s = 0$), Exponential ($c_s = 1$), and Hyperexponential ($c_s > 1$). Similar to Fig. 2, Fig. 5 plots the normalized capacity $\frac{\Lambda}{\Lambda_p}$ versus CV of data call inter-arrival times. Variable Λ_p represents the maximum aggregate arrival rate when data calls arrive according to a Poisson process. It is the same for the different workload size distribution according to the insensitivity property of the loss system. The simulation results match the analytical results for the case of exponentially distributed workload sizes. As the variability of inter-arrival times increases, the system capacity decreases in all three cases. This is consistent with the observations made from the numerical study. The further simulations also verified the other results based on the previous analysis: the variability of the data call arrival process makes capacity more unbalanced between data and voice traffic, and the transmission rate and E_b/N_o requirement influence the extent of capacity reduction.

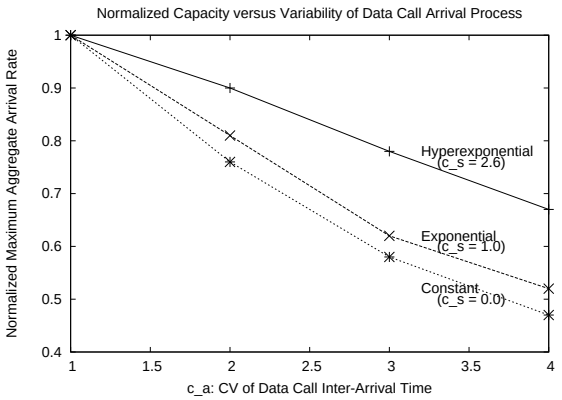


Fig. 5. Normalized capacity $\frac{\Lambda}{\Lambda_p}$ versus CV of data call inter-arrival times (2% overall blocking)

The simulations also illustrate a counter-intuitive phenomena, which could not be observed in the numerical analysis. As shown in Fig. 5, the capacity decrease is more pronounced for Constant workload sizes, and less pronounced for Hyperexponential workload sizes. The variability of the workload size (service time) of data calls decreases the aggregate blocking probability. This is contrary to intuition. It is generally believed that the more variable the arrival or service time process is, the worse any of the usual queueing performance measures will be. Our counter-intuitive result motivates further exploration of the relationship between the service time variability and the blocking probability in a G/GI/s/s loss system. We seek a systematic view that can corroborate our simulation findings.

To the best of our knowledge, Wolff [17] was the first to study counter-intuitive behaviors in queueing systems. Other simulation and numeric results [11,14–16] also show similar phenomena. As far as we know, no one has rigorously proved the relationship between the service time variability and the blocking probability for the G/GI/s/s system. Thus we pursue this study based on an approximation formula provided by Srikant and Whitt [14].

For the G/GI/s/s queue, let λ denote the mean arrival rate, μ the mean service rate, $G(x)$ the service time cumulative distribution function, and ρ the system utilization, where $\rho = \lambda/(\mu s)$. Next we characterize the variability of the arrival and service-time processes. For the arrival process, we use *normalized arrival asymptotic variance*[1] denoted by c_a^2 to partially characterize the variability. For a deterministically evenly spaced process, $c_a^2 = 0$; for a Poisson process, $c_a^2 = 1$. For a renewal process, c_a^2 coincides with the *Squared Coefficient of Variation* (SCV) of the interarrival times. For a non-renewal process, c_a^2 captures correlations between different inter-arrival times. The larger c_a^2 is,

[1] $c_a^2 = \lim_{t\to\infty} \frac{Var(A(t))}{\lambda t}$, where $A(t)$ is the cumulative arrival count up to time t, and λ is the mean arrival rate.

the more variable the arrival process is. The service times are independent, and thus we use SCV (c_s^2) to measure service time variability.

Since it is difficult to study the G/GI/s/s system directly, the loss system is often associated with the $G/GI/\infty$ model with the same arrival and service time processes. The system variability is partially characterized by the peakedness parameter z, which is defined as the ratio of the variance to the mean number of busy servers in the associated $G/GI/\infty$. The heavy-traffic approximation [14] for the peakedness is:

$$z = 1 + \mu(c_a^2 - 1) \int_0^\infty [1 - G(x)]^2 \, dx \ . \tag{15}$$

The value of $\int_0^\infty [1 - G(x)]^2 \, dx$ decreases as the service time distribution gets more variable.

According to [14], the blocking probability can be approximated by:

$$B \approx \sqrt{\frac{z}{\rho s}} \frac{\phi(-\gamma/\sqrt{z})}{\Phi(\gamma/\sqrt{z})}, \tag{16}$$

where $\gamma = \sqrt{s}(1-\rho)/\sqrt{\rho}$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the density and cumulative distribution functions of the standard normal distribution. Formula (16) is asymptotically correct under the constraint $\sqrt{s}(1-\rho)/\sqrt{\rho} \to \gamma$ as $s \to \infty$. When this constraint is satisfied, the system must be in the heavy traffic region. Peakedness z expressed by (15) can be used as an approximation. This approximation is reasonable if z is not very large.

We combine (15) and (16) to study the qualitative behavior of the blocking probability as a function of the service-time variability. Use $\uparrow$ for 'increases', $\downarrow$ for 'decreases' and $\Rightarrow$ for 'results in'. From (15), we have: $c_s^2 \uparrow \Rightarrow z \downarrow$, if $c_a^2 > 1$; $c_s^2 \uparrow \Rightarrow z = 1$, if $c_a^2 = 1$; $c_s^2 \uparrow \Rightarrow z \uparrow$, if $c_a^2 < 1$. Also from (16), we have $z \uparrow \Rightarrow B \uparrow$ since $\phi(-\gamma/\sqrt{z})$ and $\Phi(\gamma/\sqrt{z})$ are increasing and decreasing functions of z, respectively. Combining the relationships among c_s^2, z, and B, we have:

$$c_s^2 \uparrow \Rightarrow B \downarrow, \qquad \text{if } c_a^2 > 1, \tag{17}$$

$$c_s^2 \uparrow \Rightarrow B \text{ no change}, \quad \text{if } c_a^2 = 1, \tag{18}$$

$$c_s^2 \uparrow \Rightarrow B \uparrow, \qquad \text{if } c_a^2 < 1. \tag{19}$$

Expression (17) is consistent with our simulations and (18) is consistent with the insensitivity property of the $M/GI/s/s$ system. As shown by (19), the counter-intuitive phenomenon does not occur when the arrival process is less variable than a Poisson process.

This analysis corroborates our simulation results, and enhances our understanding of the fundamental issues. It is generally believed that data traffic

is much variable than voice traffic. We conclude that increased variability in data call holding times decreases the blocking probability and increases the effective system capacity.

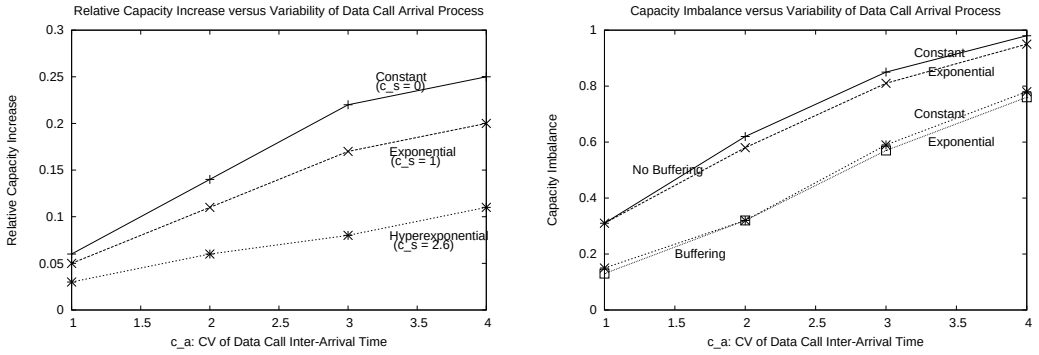
5 Data Call Buffering

Our previous results show that increased variability in the data call arrival process reduces the system capacity and exacerbates capacity imbalance among traffic classes. These observations motivate a simple buffer-based resource management scheme for data calls. The rationale for the data call buffering scheme is three-fold. First, data traffic is generally more tolerant to delay than voice traffic. Second, buffering can effectively mitigate the variability in the data call arrival process. Third, buffering data calls temporarily rather than immediately blocking them provides these calls a better opportunity to enter the system later. In other words, buffering can provide a controllable performance tradeoff between voice and data calls.

The buffering scheme works as follows. Arriving data calls that encounter a full system must enter a FIFO queue. The buffer size is infinite, so no data call is blocked due to insufficient buffer space. Each of the buffered calls has an associated timer set to a maximum delay t_B. The timer starts when the call enters the queue. When a call releases a channel from the system, the system checks whether it has room for buffered data calls. If there are M_2 data calls in the buffer, and there is room for M_1 calls, then $\min(M_1, M_2)$ data calls are removed from the buffer to access channels. Otherwise, no data call is accepted at that moment. A call is cleared from the buffer once its timer expires. Thus, data call blocking can occur t_B seconds after arriving. Furthermore, only unexpired data calls can access channels. For voice calls, the system works as an ordinary loss system. If there are available resources at the time of arrival, the call is accepted. Otherwise, the call is blocked right away. Voice calls have priority over buffered data calls to access channels.

Buffering reduces the aggregate blocking rate by slightly increasing the data call delay. We measure its efficiency by the relative capacity increase $\frac{\Lambda^{(b)}-\Lambda}{\Lambda}$, where $\Lambda^{(b)}$ is the maximum aggregate arrival rate when data call buffering is applied. Simulation is employed to study the impact of the stochastic properties of data calls on the efficiency of the buffering strategy.

Figure 6(a) shows the relationship between the relative capacity increase and the variability of the data call arrival process, for three different workload size distributions. The maximum buffer delay is 2 seconds and the overall blocking requirement is 2%. For Poisson arrivals, the capacity improvement is marginal (about 5%). When the data call arrival process is more variable, buffering



(a) (b)

Fig. 6. Effect of data call buffering (2s maximum delay) on (a) relative capacity increase $\frac{\Lambda^{(b)}-\Lambda}{\Lambda}$; (b) capacity imbalance $\Delta\Lambda = \Lambda_1 - \Lambda_0$.

offers capacity improvements of 10-25%. The greatest capacity improvement is observed for the Constant workload size distribution.

The buffering scheme also provides control over the tradeoff between blocking rates for different traffic classes. In our earlier results, capacity imbalance between voice and data traffic is inherent due to the higher effective bandwidth of data calls. This imbalance is further aggravated by data call arrival variability. However, the buffering scheme gives data calls more chances to enter the system. The blocking probability of data traffic is thus reduced, while that of voice traffic increases. As a result, this scheme balances the capacity restrictions from different traffic class requirements. Figure 6(b) compares the capacity imbalance, $\Delta\Lambda$, for cases with and without buffering. In all cases, buffering mitigates the capacity imbalance between data and voice traffic, and thus enhances the overall capacity of the CDMA system.

In Section 4, we observed a counter-intuitive phenomenon with respect to the variability of data call workload sizes. The phenomenon still exists when data call buffering is applied. However, this effect is less pronounced when the maximum buffer delay increases.

Though simple, the data call buffering scheme can effectively enhance the system capacity when the data traffic arrival process is highly variable. Adjusting the maximum buffer delay provides controllable performance tradeoffs between blocking probability and delay, and between voice and data traffic.

6 Conclusions

This paper studies a multi-service CDMA system supporting voice and data traffic. The main emphasis in our work is on understanding the impacts of

non-Poisson data traffic on overall CDMA system capacity.

We first study the system based on a Markov Regenerative Process (MRGP) model, and then explore a more general system using simulation. Our results show that increased variability in the data call arrival process decreases the system capacity, while increased variability in data call holding times increases the capacity. The extent of the phenomena observed depends on other system parameters, such as transmission rates and E_b/N_o requirements. We also study a buffer-based resource management scheme that enhances the system capacity in the presence of high-variability data traffic.

Ongoing work is exploring the impact of correlated data call interarrival times and workload sizes on the capacity. We also plan to investigate the capacity of the CDMA2000-1xEVDO system using more realistic traffic models.

References

[1] E. Altman, "Capacity of Multi-service Cellular Networks with Transmission Rate Control: A Queueing Analysis", *Proceedings of ACM MOBICOM*, Atlanta, GA, pp. 205-214, September 2002.

[2] X. Bo and Z. Chen, "On Call Admission and Performance Evaluation for Multiservice CDMA Networks", *ACM SIGMOBILE Mobile Computing and Communications Review*, Vol. 8, No. 1, pp. 98-108, 2004.

[3] S. Dharmaraja, D. Logothetis, and K. Trivedi, "Performance Modelling of Wireless Networks with Generally Distributed Handoff Interarrival Times", *Computer Communications*, Vol. 26, pp. 1747-1755, 2003.

[4] R. German, *Performance Analysis of Communication Systems: Modeling with Non-Markovian Stochastic Petri Nets*, Wiley, 2000.

[5] N. Hegde and E. Altman, "Capacity of Multiservice WCDMA Networks with Variable GoS", *IEEE Wireless Communications and Networking Conference*, Vol. 4, No. 1, pp. 1402-1407, March 2003.

[6] Y. Ishikawa and N. Umeda, "Capacity Design and Performance of Call Admission Control in Cellular CDMA Systems", *IEEE Journal on Selected Areas in Communications*, Vol. 15, No. 8, pp. 1627-1635, 1997.

[7] W. Jeon and D. Jeong, "Call Admission Control for CDMA Mobile Communications Systems Supporting Multimedia Services", *IEEE Transactions on Wireless Communications*, Vol. 1, No. 4, pp. 649-659, October 2002.

[8] I. Koo, J. Ahn, J. Lee, and K. Kim, "Analysis of Erlang Capacity for the Multimedia DS-CDMA Systems", *IEICE Transactions on Fundamentals*, Vol. E82-A, No. 5, pp. 849-855, 1999.

[9] V. Kulkarni, *Modeling and Analysis of Stochastic Systems*, Chapman & Hall, London, 1995.

[10] J. Lee and L. Miller, "Solutions for Minimum Required Forward Link Channel Power in CDMA Cellular and PCS Systems", *Journal of Communications and Networks*, Vol. 1, No. 1, pp. 42-51, March 1999.

[11] H. Masuyama and T. Takine, "Analysis of an Infinite-Server Queue with Batch Markovian Arrival Streams", *Queueing Systems*, Vol. 42, No. 3, pp. 269-296, 2002.

[12] V. Paxson and S. Floyd, "Wide Area Traffic: The Failure of Poisson Modeling", *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp. 226-244, 1994.

[13] A. Puliafito, M. Scarpa, and K. Trivedi, "Petri Nets with k Simultaneously Enabled Generally Distributed Timed Transitions", *Performance Evaluation*, Vol. 32, No. 1, pp. 1-34, 1998.

[14] R. Srikant and W. Whitt, "Simulation Run Lengths to Estimate Blocking Probabilities", *ACM Transactions on Modeling and Computer Simulation*, Vol. 6, pp. 7-52, January 1996.

[15] W. Whitt, "Heavy Traffic Approximations for Service Systems with Blocking", *Bell Labs Technical Journal*, Vol. 63, pp. 689–708, 1984.

[16] W. Whitt, "A Diffusion Approximation for the G/GI/n/m Queue", *Operation Research*, Vol. 52, No. 6, pp. 922-941, November/December 2004.

[17] R. Wolff, "The Effect of Service Time Regularity on System Performance", *Computer Performance* (edited by K. Chandy and M. Reiser), pp. 297-304, North-Holland, 1977.

Yujing Wu received the Ph.D. degree in Electrical and Computer Engineering from the University of Massachusetts at Amherst. She is a postdoctoral fellow in the Department of Computer Science at the University of Calgary, where she does research on capacity planning of 3G cellular networks.

Carey Williamson is an iCORE Professor in the Department of Computer Science at the University of Calgary, specializing in *Wireless Internet Traffic Modeling*. He holds a B.Sc.(Honours) in Computer Science from the University of Saskatchewan, and a Ph.D. in Computer Science from Stanford University. His research interests include Internet protocols, wireless networks, network traffic measurement, network simulation, and Web server performance.