Performance Implications of Fluctuating Server Capacity

Jingxiang Luo Carey Williamson
Department of Computer Science
University of Calgary
$\{$jxluo,carey$\}$@cpsc.ucalgary.ca

Abstract

In this paper, we consider a variant of an M/M/c/c loss system with fluctuating server capacity. Given a set of primary inputs, such as arrival rate, service rate, and capacity fluctuation rate, we develop a detailed model using MRGP, such that blocking and dropping metrics can be calculated explicitly. To gain insight into performance implications of a stochastic capacity queue, we conduct an analysis on a simple, approximate model. We investigate the functional behaviour of the system, using both the approximate and the detailed model. The significance of results in the context of network performance evaluation and capacity planning is highlighted.

Keywords: network capacity planning, stochastic capacity, loss system, MRGP

1. INTRODUCTION

Server capacity is the primary determinant of system performance. Network provisioning involves determining a suitable configuration of system capacity (service dimensioning) in order to achieve desired performance targets for metrics such as blocking or delay.

In most systems, the server capacity is a constant value. An illustrative example is used in telephony: the Erlang B formula expresses the call blocking probability in terms of traffic intensity and the trunk capacity. The number of lines required for a target blocking probability can be decided by using such a formula.

For the design and QoS provisioning of next-generation networks, capacity planning is an important issue. However, there is no corresponding formula as robust as the Erlang formula. This is mainly due to the complicated nature of computer networks: heterogeneous traffic, service differentiation policies, and the noisy environment for wireless networks.

Server capacity is the number of parallel service channels plus the number of buffers. Consider a server with capacity C, where C is a random variable. This system experiences two types of losses: *blocking* when an arrival encounters a full system, and *dropping* when the system capacity decreases while it is full. Many complicated factors in networks, such as server downtime, link failure, or competition among prioritized applications, can be represented using stochastic capacity models. For example, in a priority service discipline, higher-priority applications may have pre-emptive priority over low-priority applications. We can use a stochastic capacity model to represent the resource availability for low-priority applications.

There have been several prior studies for buffered and unbuffered systems that involve the change of server capacity, either implicitly or explicitly. Among them are the unreliable server system, where servers have downtime and later recover [12] [13] [14]. A recent development is performability modelling [6] [11] [18], which accounts for both system performance and system availability. For mobile users using cell phones and mobile devices, mobility leads to many handoff calls. To provide uninterrupted service, certain channels are reserved for handoff only [1]. Hence the number of channels available for new calls may vary. Mobile users access systems by wireless/cellular connections as in CDMA, where interference from neighbouring cells can severely limit the number of calls accepted by a base station. Hence whether the environment is quiet or noisy contributes to capacity fluctuations (see [9] and the references therein). The impact of data calls on the capacity of a CDMA multi-service system was evaluated in [19]. The concept of stochastic capacity has been recently proposed for studying the impacts of capacity fluctuation [17]. Due to the complicated nature of the issue, performance was evaluated mainly through simulation.

In the infinite buffer case, there is no loss, hence we call it a *pure delay system*. For a pure delay system, there have been a couple of studies [2] [10] dedicated to the delay performance in a so-called *perturbed system*, i.e., a small fraction of channels can be taken away to accommodate prioritized traffic. If the system is unbuffered, we have a *pure loss system*. We assume *no buffer*, hence a loss system.

Loss systems behave quite differently from delay systems. In the stochastic capacity queue, dropping also has an impact on the system performance. As far as we know, this has not been studied in the literature. Some system input parameters, such as the ratio of the load to the mean capacity, and the fluctuation timescale of server capacity (i.e., how frequently the capacity changes relative to the service rate), have major influences on performance. The effects of these factors (e.g., higher blocking when the load increases), are mostly intuitive. Some other factors, such as the coefficient of variation (CoV) of the service time distribution, and other details, also have implications. The effects of the latter factors (also referred to as "fine characteristics") are less obvious.

We are interested in studying the system's functional behaviour in response to the primary inputs, e.g. traffic load, capacity fluctuation, service time distribution, and service differentiation policies. In particular, we will focus on the effects of the *"fine characteristics"* on the system performance.

Two models, an approximate model and a detailed model, are developed in this paper. The detailed model, which uses a Markov regenerative process (MRGP), applies widely to different phase-type service time distributions and capacity variations, to yield precise performance metrics. The detailed model serves essentially as a computation tool for deriving the performance metrics. On the other hand, the approximate model yields insight into the behaviour of a stochastic capacity loss system; the insight is difficult, if not impossible, to obtain from the detailed model.

The rest of this paper is organized as follows. In Section 2, we describe the loss model of a stochastic capacity queue. In Section 3, we develop an approximate model to gain insight about the behaviour of this system. In Section 4, we develop a detailed formulation for the model using an MRGP. In Section 5, we express the loss metrics precisely in terms of the equilibrium state distribution. In Section 6, we present numerical examples. Section 7 concludes the paper.

2. DESCRIPTION OF A LOSS MODEL OF A STOCHASTIC CAPACITY QUEUE

In this section, we describe our model of a loss queueing system with stochastic capacity. The system is denoted generically as A/B/∼C/∼C, where A represents the inter-arrival time distribution, B represents the service time distribution, and C is the number of parallel service channels (i.e., server capacity). The notation ∼C indicates that the server capacity fluctuates with time. We assume an unbuffered system. We use the general terms "servers" and "customers" in this paper. A "server" can mean a service station, a compute server, or a processing node in an application scenario. In the context of communication, a customer refers to a call or a connection request to an access point or a network server.

2.1 Stochastic capacity variation

To describe the status of a stochastic-capacity queue, we use state (n,c) where n is the number of customers in service, and c is the current server capacity. Two processes can be identified, namely, the capacity fluctuation process $C(t)$ and the queuing process $N(t)$. It should be clear that the evolution of the underlying model is subject to $N(t) \leq C(t)$ for any time t, since $C(t)$ is the capacity limit. It is assumed that T_c, the time elapsed between capacity changes, follows a general distribution as:

$$\Pr\{T_c < t\} = F_c(t). \tag{1}$$

We assume that $C(t)$ takes integer values in $[0, c_M]$. Denote $Y(t) = c_M - C(t)$, which can be regarded as an interference process competing for the server capacity c_M with the main queue process $N(t)$; the interference process has pre-emptive priority over the main queue process. In a heterogeneous network, the fluctuation of $C(t)$ can be at different time scales, depending on the application.

In the rest of this paper, we assume that $C(t)$ is skip-free, taking integer values in $[0, c_M]$. T_c follows a known distribution $F_c(t)$ that does not depend on system state. Provided that the capacity is c currently, at the next capacity change, $C(t)$ increases to $c+1$ with probability $f_\uparrow(c)$, and decreases to $c-1$ with probability $f_\downarrow(c)$. We always have $f_\uparrow(c) + f_\downarrow(c) = 1$. Also, $f_\uparrow(0) = 1$ and $f_\downarrow(c_M) = 1$. $C(t)$ is a simple quasi-birth-death process that is independent of $N(t)$.

2.2 Phase type services and dropping policies

Service times follow a distribution $\Pr\{T_b < t\} = F_b(t)$. For non-exponential distributions, the standard techniques date back to Cox. One technique is to supplement system size n with an ordered set of time variables $t_g, 1 \leq g \leq n$, denoting the time of service already expended on each customer. Alternatively, we may consider a k-phase Cox distribution (see Figure 1, for $k = 2$), where the service time is approximated with (finite) k exponential phases. Service always starts at phase 1; after finishing phase $j-1$, a customer goes to the next phase with probability β_j, and skips the remaining phases with probability $1 - \beta_j$. The service rate at each phase is μ_j. The Cox distribution is known to form a dense subset within the set of all distributions with real, non-negative support. Further, algorithms have been provided [16] to approximate an arbitrary distribution with a Cox distribution, and examples were given for both light-tailed and heavy-tailed distributions.

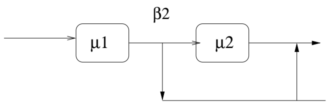

Figure 1. Example of two-phase Cox distribution model

Dropping is possible in a stochastic capacity system. Now we discuss the dropping rules (i.e., which victim to choose when the system is forced to drop). Let $\chi_i, 1 \leq i \leq k$ denote the probability that a customer is dropped from phase i. A dropping policy is specified by setting χ_i:

1. *Random dropping:* Let $n_j, 1 \leq j \leq k$ be the number of customers in phase j of the Cox service model immediately prior to the dropping instant. In random dropping,

$\chi_i = \frac{n_i}{n}$ for any i. That is, each client in the system is victimized with equal probability.

2. *Fixed-ratio dropping:* The victim to be dropped is chosen from different phases according to a pre-determined ratio. A fixed ratio υ_i is given for each i where $\sum_i \upsilon_i = 1$. We set $\chi_i = \upsilon_i$ if $n_i > 0$ for any i.
   Sometimes there exists some phase i such that $n_i = 0$. When this is the case, we re-normalize as below:
$$\chi_i = \left[\begin{array}{l} 0, \text{ if } \ n_i = 0 \\ \frac{\upsilon_i}{\sum_{\{j:n_j>0\}} \upsilon_j}, \text{ if } \ n_i > 0 \end{array} \right.$$

For fixed-ratio dropping to be imposed, it is necessary for the system to know the current service phase of each customer. However, knowing the current phase is not necessary for the policy of random dropping.

2.3 Cox service: coefficient of variation

We consider a two-phase Cox distribution, for which we calculate the coefficient of variation, which is a key characteristic of the service time. (The result derived in this Section is used in our approximate analysis in Section 3.). A random variable T drawn from this distribution (shown in Figure 1) can be expressed as $T_1 + X \ T_2$, where $T_1 \sim \text{Exp}(\mu_1), T_2 \sim \text{Exp}(\mu_2)$, and X takes a value from a 0-1 bivariate distribution: $X = 1$ with probability β_2, and $X = 0$ otherwise.

Let $E(.)$ and Var(.) denote the expectation and variance respectively. Let $\bar{\mu} = 1/E[T]$ be the average rate of service completion. The following holds for the Cox distribution:

$$1/\bar{\mu} = 1/\mu_1 + \beta_2/\mu_2.$$

Multiplying both sides by $\bar{\mu}$, we have $1 = \frac{\bar{\mu}}{\mu_1} + \frac{\bar{\mu}\beta_2}{\mu_2}$. We introduce a variable v:

$$v = \frac{\bar{\mu}\beta_2}{\mu_2}, 1 - v = \frac{\bar{\mu}}{\mu_1}. \tag{2}$$

Here v is the expected proportion of time spent in phase 2 for an average customer.

We calculate the variance:

$$\begin{aligned} \text{Var}(T) &= \text{Var}(T_1) + \text{Var}(X \ T_2) \\ &= \text{Var}(T_1) + E(X^2 \ T_2^2) - E^2(X \ T_2) \\ &= \text{Var}(T_1) + \beta_2 \text{Var}(T_2) + (\beta_2 - \beta_2^2) E^2(T_2). \end{aligned}$$

Let σ_b be the coefficient of variation. We have,

$$\sigma_b^2 = \text{Var}(T)/E^2(T) = \bar{\mu}^2 \text{Var}(T).$$

Since both T_1 and T_2 are exponentially distributed, $\text{Var}(T_1) = E^2(T_1)$ and $\text{Var}(T_2) = E^2(T_2)$. Substitution of these relations into the earlier formula for $\text{Var}(T)$ yields

$$\begin{aligned} \sigma_b^2 &= (\bar{\mu}/\mu_1)^2 + \beta_2(\bar{\mu}/\mu_2)^2 + (\beta_2 - \beta_2^2)(\bar{\mu}/\mu_2)^2 \\ &= (1 - v)^2 + ((2/\beta_2) - 1)v^2. \end{aligned} \tag{3}$$

We refer to (σ_b, v) as an alternative parameterization for any two-phase Cox distribution.

3. APPROXIMATE MODEL

We conduct in this section an analysis based on some assumed approximations. The emphasis is not on the preciseness of the model. Rather, we want to examine some key factors that have relatively simple characteristics, and by using them, we gain insight for predicting performance trends as those factors change.

We assume that arrivals are Poisson. Service times, and times between successive capacity changes, follow distributions F_b and F_c respectively.

3.1 Effect of F_b on blocking

It is well known that *Poisson arrivals see time averages (PASTA)*. Hence, the probability that an arrival is blocked is equal to the stationary probability that the system is full. For a fixed capacity M/G/c/c system, we know that the blocking probability is insensitive to the service time distribution (i.e., only the mean matters) [3]. However, this is no longer valid in a stochastic capacity system.

Drops in a stochastic capacity system reduce the mean time that a client spends in the system. In our model, besides the arrival rate and the designated stochastic variation of capacity, the mean time spent in the system is the most important factor affecting the performance. If each client stays in the system longer on average, then the system becomes more congested, leading to higher blocking.

Let T_b be the service time, i.e., the time needed for a client to *complete* service. Let T_b' be the amount of time that an accepted client spends in the stochastic capacity system. We have $T_b' \leq T_b$, as a client might be dropped before service completion. The reduction from $E[T_b]$ to $E[T_b']$ is called the *dropping-induced speedup effect*.

We claim that in a stochastic capacity system, the mean time spent in the system may depend on the distribution of F_b and not just the mean service rate. Now we estimate $\mu' = E[T_b']$. Let r_{drop} be the dropping rate, i.e., the number of drops per unit time. We assume that: (1) the dropping rate is fixed, and the time between drops is exponentially distributed (as an approximation); (2) T_b follows a two-phase Cox distribution; and (3) a victim for dropping is chosen from phase i with (long-run) probability ϕ_i, $i = 1, 2$. (Statistically, $\phi_i = E[\chi_i]$, where χ_i is the decision variable used in discussion of dropping policies.) Under these assumptions, a client (after being accepted) departs phase 1 with rate $\mu_1 + r_{\text{drop}}\phi_1$, where $r_{\text{drop}}\phi_1$ is the rate that a customer in phase 1 is dropped as a victim. For a client entering phase 2, the rate of departure from this phase is $\mu_2 + r_{\text{drop}}\phi_2$. Among all clients in phase 1, only a proportion $\frac{\mu_1\beta_2}{\mu_1 + r_{\text{drop}}\phi_1}$ enter phase 2. We assume that

$r_{\text{drop}}/\mu << 1$. For the expected time that an accepted client spends in the system, we have:

$$\begin{aligned}
\frac{1}{\mu'} &= \frac{1}{\mu_1 + r_{\text{drop}}\phi_1} + \frac{\mu_1\beta_2}{\mu_1 + r_{\text{drop}}\phi_1} \times \frac{1}{\mu_2 + r_{\text{drop}}\phi_2} \\
&= \frac{1}{\mu_1} \times \frac{1}{1 + \frac{r_{\text{drop}}}{\mu_1}\phi_1} \\
&\quad + \frac{1}{\mu_1}\frac{\mu_1\beta_2}{\mu_2} \times \frac{1}{(1 + \frac{r_{\text{drop}}}{\mu_1}\phi_1)(1 + \frac{r_{\text{drop}}}{\mu_2}\phi_2)} \\
&= (\frac{1}{\mu_1} + \frac{\beta_2}{\mu_2}) - \frac{r_{\text{drop}}}{\mu_1}\phi_1(\frac{1}{\mu_1} + \frac{\beta_2}{\mu_2}) - \frac{\beta_2}{\mu_2}\frac{r_{\text{drop}}}{\mu_2}\phi_2 \\
&\quad + \frac{1}{\bar{\mu}}\, o(\frac{r_{\text{drop}}}{\bar{\mu}}). \qquad (4)
\end{aligned}$$

In the above derivation, we have applied the expansion $1/(1+x) = 1 - x + o(x)$ as $x << 1$. To simplify, we substitute Equation (2) into Equation (4), and obtain after some manipulations:

$$\frac{1}{\mu'} = \frac{1}{\bar{\mu}} - \frac{r_{\text{drop}}}{\bar{\mu}^2}\Gamma + o(\frac{r_{\text{drop}}}{\bar{\mu}^2}), \qquad (5)$$

where

$$\Gamma = (1-v)\phi_1 + \frac{v^2}{\beta_2}\phi_2.$$

From Equation (3), we see that $v^2/\beta_2 = (\sigma_b^2 - (1-2v))/2$. Substituting this into the expression for Γ yields

$$\Gamma = (1-v)\phi_1 + v\phi_2 + \phi_2(\sigma_b^2 - 1)/2. \qquad (6)$$

The amount

$$E[T_b] - E[T_b'] = \frac{1}{\bar{\mu}} - \frac{1}{\mu'} = \frac{r_{\text{drop}}}{\bar{\mu}^2}\Gamma + o(\frac{r_{\text{drop}}}{\bar{\mu}^2}) \qquad (7)$$

measures the dropping-induced speedup effect. Note this effect is negligible when $\sigma_b \leq 1$ and $r_{\text{drop}} << \mu$.

For the fixed ratio dropping policy, ϕ_i is roughly proportional to υ_i. For the random dropping policy, ϕ_i is proportional to the expected number in phase i at equilibrium, which (according to queueing network theory), is proportional to the traffic intensity going through phase i. (This does not strictly apply to a loss system, however it holds approximately when the loss ratio is low.) We have,

$$\frac{\phi_1}{\phi_2} \approx \frac{\lambda/\mu_1}{\lambda\beta_2/\mu_2} = \frac{\mu_2}{\mu_1\beta_2} = \frac{1-v}{v}.$$

Note that $\phi_1 + \phi_2 = 1$, hence $\phi_1 \approx (1-v), \phi_2 \approx v$. With this, we find that

$$\Gamma \approx (1-v)^2 + (v/2)(\sigma_b^2 - 1) + v^2. \qquad (8)$$

For a large σ_b and a fixed non-negligible r_{drop}, when factor Γ increases (i.e., as σ_b^2 increases or v increases), the dropping-induced speedup effect becomes more obvious. Consequently the blocking probability decreases.

3.2 Effect of F_c on dropping

We relate dropping metrics to the probability of observing that the system is full. Let us assume an omniscient observer, who is sent immediately prior to each capacity decrease. We refer to this as the *dropping observance scenario*, in which the number of times seeing a full system is equal to the number of drops in the system.

Denote τ_* the instant of the most recent capacity decrease. In this section, we denote t the time elapsed since τ_* (i.e., imagine a clock that is reset to 0 every time that a capacity decrease occurs: t is the reading of this clock). We introduce a conditional probability as follows: let

$$\begin{aligned}
p^*(t|c) &= \Pr\{\text{ system is full at } \tau_* + t, \qquad (9) \\
&\qquad \text{no capacity decrease in } (\tau_*, \tau_* + t) | C(\tau_*^+) = c\}
\end{aligned}$$

A typical curve of $p^*(t|c)$ is illustrated in Figure 2: the probability of the system being full is at its highest immediately following a capacity decrease, but will decrease until another capacity decrease occurs. Let T^D be the time elapsed between two successive capacity decreases. The system is seen full with probability $p^*(T^D|c)$ (in the dropping observance scenario). Also note that (intuitively) the distribution of T^D has a heavy-tail if F_c is a heavy-tailed distribution.

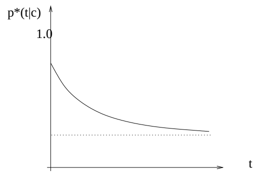

Figure 2. The curve of $p^*(t|c)$ as t elapses

The blocking probability in the previous discussion relates only to the time averaged probability that the system is full, due to the Poisson arrival assumption. However, the dropping relates to both $p^*(t|c)$ as just defined, and the timing of capacity decreases (which is presumed to be the arrival instant of the omniscient observer). But if capacity fluctuates slowly, then T^D typically takes large values. Intuitively, we know that $p^*(T^D|c)$ decreases (asymptotically) to $p^*(\infty|c)$ as $T^D \to \infty$. Hence, provided that at the instant immediately prior to a capacity decrease, the capacity is c, the probability of observing that the system is full is close to $p^*(\infty|c)$, irrespectively of F_c. This applies when capacity fluctuates slowly.

As capacity fluctuates faster, the distribution of F_c has a greater impact on the dropping rate. In the model, when distribution F_c varies both $p^*(t|c)$ and the distribution of T^D may change. However, provided that the mean of F_c does not change, and if r_{drop} is low compared to $\bar{\mu}$, and $\sigma_b < 1$ (thus the dropping-induced speedup effect is minimal), then we may consider that the change of dropping rate while F_c varies is mainly due to the timing of arrival instants of the "observer" (i.e., at what time the capacity decreases occur). Probability $p^*(t|c)$ decreases as t increases (refer to Figure 2). The observer at larger t in the *dropping observance scenario* sees that the system is full with relatively lower probability. If the observer comes soon after the most recent capacity decrease, there is a high probability of seeing a full system.

Strictly speaking, it is the function $F_c(t)$ over the whole region $0 \le t < \infty$ that has an impact on the dropping rate, and not just the tail. For instance, for a distribution F_c with mean $1/\lambda_f$, we consider three domains: $[0, 1/\lambda_f)$ (the "head"), $[1/\lambda_f, 3/\lambda_f]$ (the "body"), and $(3/\lambda_f, \infty)$ (the "tail") respectively. Hence, dropping is typically higher when F_c follows a Gamma distribution, as a Gamma distribution has a larger density function at the "head" than an exponential distribution. On the other hand, a typical heavy-tailed distribution is Pareto distribution with shape parameter $1 < \kappa < 2$, which has a larger density function at the "tail". (To refer to the definition of Pareto distribution, please see Formula (20) later in this paper.) When F_c follows such a Pareto distribution, the dropping rate is typically lower.

4. DETAILED MODEL USING MRGP

In this section, we develop a detailed model of stochastic capacity loss system. The state space of a stochastic capacity system is denoted by:

$$\mathcal{S} = \{(n,c) : 0 \le n \le c \le c_M\}, \qquad (10)$$

Let us consider a row vector $\vec{p}(t)$ indexed by each $s \in \mathcal{S}$, with $\vec{p}_s(t)$ recording the probability of being in state s at instant t. We describe the system dynamics:

$$\frac{d}{dt}\,\vec{p}(t) = \vec{p}(t)\tilde{\mathbf{Q}}(t), \qquad (11)$$

where matrix $\tilde{\mathbf{Q}}(t)$ is the infinitesimal generator matrix.

In later discussion, we will extend $\mathcal{S}$ where appropriate, for the purpose of capturing dropping (Section 4.2) and for incorporating phase variables in service times (Section 4.3).

4.1 Process of stochastic capacity queue

In Equation (11), if the process has an equilibrium and $\tilde{\mathbf{Q}}$ does not depend on t, then the equilibrium state distribution $\vec{\pi}$ can be solved from $\vec{\pi}\tilde{\mathbf{Q}} = 0$. However, only in the M/M/∼C/∼C case, and where T_c follows an exponential distribution, would $\tilde{\mathbf{Q}}(t)$ be independent of t (for this system, all events are Markovian). In the following, we identify those special instants that satisfy the Markov property, and construct an MRGP in order to solve for the equilibrium.

We briefly introduce the Markov regenerative process, and refer readers to rich literature in this area, both in theory development and in applications [4] [5] [8] [11] [15]. In an MRGP, there exist time points where the process satisfies the Markov property. We call these time instants regeneration points. Consider a bivariate sequence $\{(u_h, t_h), h = 0, 1, 2, ..\}$, where t_h is a time point and u_h describes the system status at time t_h. The sequence $\{(u_h, t_h)\}$ defines a Markov chain, satisfying both Markov property and time homogeneity. An MRGP $Z(t)$ associated with the sequence $\{(u_h, t_h)\}$ is characterized by the following property: all conditional finite dimensional distributions of $\{z(t_h + t), t \ge 0\}$ given $\{z(\cdot), 0 \le \cdot < t_h; u_h = i\}$ are the same as those $\{z(t), t \ge 0\}$ given $u_0 = i$, i.e. the evolution of $Z(t)$ after t_h depends on the state at t_h, the most recent regeneration point, but not on evolution before that t_h. The time period from t_h to t_{h+1} is called a *regenerative cycle*. Evolution of $Z(t)$ is determined by the global kernel $\mathbf{K}(t)$ and the local kernel $\mathbf{E}(t)$. Kernel $\mathbf{K}(t)$ describes the behaviour of $Z(t)$ at the regeneration instants through an embedded (discrete-time) Markov chain (EMC), while the local kernel $\mathbf{E}(t)$ describes it through the state distribution at time t between two consecutive regeneration instants.

Let $Z(t) = (N(t), C(t))$, in which $N(t)$ and $C(t)$ are the queueing process and capacity variation process in our discussion. In capacity variation $C(t)$, T_c: the time between capacity changes, is allowed to follow an arbitrary distribution. Consider a sequence $\{Z(t_h), t_h\}, h = 1, 2, ..,$ where $\{t_h\}$ denotes the sequence of capacity change instants. This sequence is a Markov renewal sequence under certain assumptions e.g., for an M/M/∼C/∼C queue; for an M/PH/∼C/∼C queue (service is phase-type distribution), a Markov renewal sequence can be formed if phase variables are incorporated (refer to Section 4.3, where we will address the case that service is Cox distribution). The stochastic capacity model $Z(t) = (N(t), C(t))$ can be viewed as the MRGP associated with the Markov renewal sequence $\{Z(t_h), t_h\}$.

Now we specify the kernels $\mathbf{K}(t)$ and $\mathbf{E}(t)$ for this system. Assume $0 = t_0 < t < t_1$. (Consideration of this interval is sufficient, due to the Markov property and time homogeneity). The entry of matrix $\mathbf{E}(t)$ is given by $E_{(n,c,n',c')}(t)$, which is the probability that, given that the system was in state (n, c) just after the previous capacity change occurred at time 0, the system will be in state (n', c') at time t, and from instant 0 to t there is no capacity change. From this definition it is obvious that for any $n \ne n'$ we have $E_{(n,c,n',c')}(t) = 0$. For time interval $(0, t)$, the dynamics governing the system are identical to those in a fixed capacity system, which have been studied extensively in literature. The entry of matrix $\mathbf{K}(t)$ is given by $K_{(n,c,n',c')}(t)$, which is the probability that the system will be in state (n', c') *immediately after* the next capacity change at time t, given that the system was in state (n, c) just after the previous capacity change occurred at time 0. Now consider the Markov chain embedded (EMC) at each instant *immediately after* each capacity change. The one-step transition probability for this EMC is given by $\mathbf{K}(\infty)$.

Related to the global kernel, we define the following. Let

t_1^- be the instant immediately prior to the capacity change. Now denote $\psi_{n,n'}(c) = \Pr\{N(t_1^-) = n' | N(t_0) = n, C(t_0^+) = c\}$ and for $0 \le c \le c_M$, define the matrix $\psi(c) = (\psi_{n,n'}(c))$. There is no capacity change for transitions described in ψ's. These ψ matrices facilitate the construction of the global kernel. Concerning the derivation of ψ matrices, we discuss as follows. Assume that $C = c$ during time interval $[t_0, t_1)$. We denote $p(t, n'|n, c)$ the transient probability of observing queue size n' at instant $t_0 + t$ prior to t_1. These transient probabilities are needed in the construction of $\psi(c)$, as shown below:

$$\psi_{n,n'}(c) = \int_0^\infty p(t, n'|n, c) dF_c(t). \qquad (12)$$

To calculate the transient probabilities, the approach of discretization (also known as uniformization) can be applied. Please see e.g. [7] for details.

However, we still have not explicitly captured the dropping, which is important. For elaboration on this point and how to construct the global matrix, please see Section 4.2.

The equilibrium solution

We address the calculation of $\vec{\pi}$, which is the equilibrium state distribution for the MRGP specified by $(\mathbf{K}, \mathbf{E})$. We follow the standard solution process, e.g. [8], with some adaptation of notation. Define $\vec{\gamma}$ as a row vector of proper size, with entry γ_s denoting the probability of being in state s, when the EMC described by $\mathbf{K}$ is in equilibrium. The following equation gives the solution of $\vec{\gamma}$,

$$\vec{\gamma}\mathbf{K} = \vec{\gamma};$$

From the local kernel $\mathbf{E}$, we can calculate:

$$\alpha_{ij} = \int_0^\infty E_{ij}(t) dt,$$

which is the expected accumulated time spent in j in a regenerative cycle that starts from i. Finally, the equilibrium state distribution $\vec{\pi}$ for the MRGP under consideration is given by:

$$\pi_j = \frac{\sum_i \gamma_i \alpha_{ij}}{\sum_i \gamma_i \sum_k \alpha_{ik}}, \qquad (13)$$

For this model, if we require that $\sum_{j \in S} \gamma_j = 1$, then the denominator in formula (13) reduces to $E[T_c]$.

4.2 Capturing dropping explicitly

Now we develop a new EMC that captures droppings explicitly. Let us denote Δ as a special value, where (c is Δ) is used to indicate that a dropping is to occur immediately. Denote $\mathcal{S}_\Delta = \{(c, \Delta) : 0 < c \le c_M\}$, in which c is the capacity at t_1^-. (See Figure 3, where we show an example for $c_M = 3$). Let $\mathcal{S}^* = \mathcal{S} \cup \mathcal{S}_\Delta$. Suppose $N(t_0) = n, C(t_0^+) = c$ and $N(t_1) = n', C(t_1^-) = c'$. If dropping occurs, it must hold that $n' = n - 1 = c'$. (Note that at the dropping instant, the system is full and at t_1 the capacity decreases by 1.) This dropping will be recognized in the new process by forcing a two-step transition, as below:

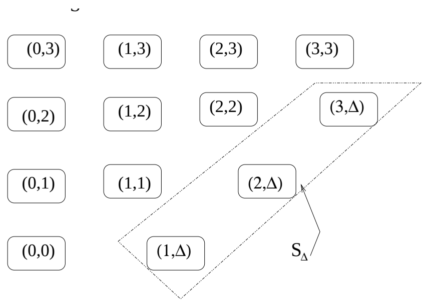

Figure 3. Extended state space $\mathcal{S}^*$ in model

$$\begin{array}{rcl} (n,c) & \rightarrow & (c,\Delta), \\ (c,\Delta) & \rightarrow & (n' = c-1, c' = c-1). \end{array} \qquad (14)$$

In this way, we can find the number of droppings by counting the number of visits to $\mathcal{S}_\Delta$ (but the time spent in $\mathcal{S}_\Delta$ is presumed 0).

We define on $\mathcal{S}^*$ a new matrix $\mathbf{K}^* = \mathbf{K}^*(\infty)$, based on the matrices $(\psi(c)), 0 \le c \le c_M$. $\mathbf{K}^*$ is constructed as:

$$K^*_{(n,c,n',c')} = \left[\begin{array}{l} 0; \quad \text{if } |c - c'| \ne 1, c \text{ or } c' \text{ is not } \Delta \\ f_\uparrow(c)\psi_{nn'}(c); \;\; \text{if } c' = c+1, n' \le c \\ \quad \text{(case 1)} \\ 0; \quad \text{if } c' = c+1 = n' \text{ (case 1a)} \\ f_\downarrow(c)\psi_{nn'}(c); \;\; \text{if } c' = c-1, c > 0 \\ \quad \text{(case 2)} \\ f_\downarrow(c)\psi_{nc}(c); \;\; \text{if } c' \text{ is } \Delta, n' = c, c > 0 \\ \quad \text{(case 2a)} \\ 1; \quad \text{if } c \text{ is } \Delta, n' = n-1 = c' \text{ (case 2b)} \\ 0; \quad \text{otherwise} \end{array} \right. \qquad (15)$$

We claim that $\mathbf{K}^*$ is the transition probability matrix of the new EMC $Z^*(t)$, which extends $Z(t)$ for $\mathcal{S}$ to the new state space $\mathcal{S}^*$. This is readily verified by each case. We point out that the two-step transition as stated in (14), corresponds to case (2a) followed by case (2b) in Equation (15).

Why is this construction necessary? In order to tell whether a dropping occurs at an instant, we must know the system status *immediately prior* to that dropping instant. However, such information is not available in the original EMC defined on $\mathcal{S}$ (recall that the EMC in our model is embedded *immediately after* instants of each capacity change). Droppings are captured in $\mathcal{S}^*$ when any transition involving dropping is forced to visit $\mathcal{S}_\Delta$.

4.3 Incorporation of phase variables

We extend the model above, which deals with only exponential service, to a model incorporating phase-type service. The purpose is to accommodate the multi-phase Cox service described in Section 2.2. For a $k-$phase Cox distribution, we can use $(n, n_2, .., n_k)$ for state description. Here n_1 is left out to remove the redundancy from $n_1 = n - \sum_{j \neq 1} n_j$.

We make adaptations to the step of constructing the kernel. This is analogous to the global kernel given in Equation (15), except for two major differences. First, in the place of matrix $\psi(c)$, we need a new matrix that incorporates phase-type services. Second, case 2a) in Equation (15) must be split into several cases corresponding to different dropping rules. These details are tedious, but straightforward.

A state is described as (n, θ), where $\theta = (n_2, n_3, .., n_k)$. Now consider the process $\bar{Z}(t) = (N(t), \Theta(t), C(t))$ defined on the expanded state space $\bar{\mathcal{S}} = \{(n, \theta, c) : 0 \le n \le c \le c_M, \theta = (n_i, i = 2, 3, ..k) : \sum_{i=2}^{k} n_i = n - n_1\}$. The description of phases is included here in order to be general. A state j in this space expands to $j = (n, \theta, c)$. We can define $\vec{\bar{\gamma}}$, $\bar{\alpha}_{ij}$ and $\vec{\bar{\pi}}$, which have meanings similar to $\vec{\gamma}$, α_{ij} and $\vec{\pi}$, but are for $\bar{Z}(t)$. Through extending the solution process for $Z(t)$ on $\mathcal{S}$ to the extended space $\bar{\mathcal{S}}$, these metrics can be worked out.

Finally, we define

$$\bar{\mathcal{S}}_\Delta = \{(n, \theta, \Delta) : 0 < n \le c_M, \theta = (n_i, i = 2, 3, ..k) : \sum_{i=2}^{k} n_i = n - n_1\} \tag{16}$$

and $\bar{\mathcal{S}}^* = \bar{\mathcal{S}} \cup \bar{\mathcal{S}}_\Delta$. State space $\bar{\mathcal{S}}_\Delta$ are similar to $\mathcal{S}_\Delta$ when we address the methodology of capturing dropping in Section 4.2, but corresponds to the situation incorporating phase variables.

5. LOSS METRICS

We calculate loss metrics as follows.

Blocking

Define:

$$\bar{\mathcal{S}}_{\text{block}} = \{(n, \theta, c) : n = c\} \tag{17}$$

Then the blocking probability p_{block} is:

$$p_{\text{block}} = \sum_{j \in \bar{\mathcal{S}}_{\text{block}}} \bar{\pi}_j = \sum_{j \in \bar{\mathcal{S}}_{\text{block}}} \sum_{i \in \bar{\mathcal{S}}} \bar{\gamma}_i \frac{\bar{\alpha}_{ij}}{E[T_c]}. \tag{18}$$

Dropping

The following formula calculates the dropping rate r_{drop}:

$$r_{\text{drop}} = \frac{\sum_{j \in \bar{\mathcal{S}}_\Delta} \bar{\gamma}_j}{\sum_{i \in \bar{\mathcal{S}}} \sum_{k \in \bar{\mathcal{S}}} \bar{\alpha}_{ik}} = \frac{\sum_{j \in \bar{\mathcal{S}}_\Delta} \bar{\gamma}_j}{E[T_c]}. \tag{19}$$

Here, state $j \in \bar{\mathcal{S}}_\Delta$ expands to $j = (n, \theta, \Delta)$, respectively state $k \in \bar{\mathcal{S}}$ expands to $k = (n, \theta, c), 0 \le c \le c_M$.

The dropping probability is the ratio of customers dropped among all customers admitted. From this, we have

$$p_{\text{drop}} = r_{\text{drop}} / (\lambda - r_{\text{block}}).$$

6. NUMERICAL EXAMPLES

We investigate by numerical examples the loss metrics in response to the input parameters in a variant of M/M/∼C/∼C system. The service time follows a Cox type distribution.

For calculation of loss metrics, we apply the detailed model developed in Section 4. Let $\mu = \bar{\mu}, \lambda, \lambda_f$ be respectively the mean service rate, the arrival rate and the mean capacity fluctuation rate. Let T_b, T_c denote respectively the random variable of service times, and times elapsed between successive capacity changes. T_b and T_c follow distributions F_b and F_c respectively. We choose a default set of parameter values, and then selectively vary some of them. The default settings are as follows. We set $c_M = 10$, $f_\uparrow(c) = 0.65$ for $0 < c < c_M$, $f_\uparrow(0) = 1.0$ and $f_\downarrow(c_M) = 1.0$. Distribution F_b is Gamma with mean 1.0 and shape parameter $k = 2$. Note that $\sigma_b < 1$ for this particular distribution. F_c is exponential with mean 1.0. The load level is $\lambda/\mu = 4.25$, and the dropping policy is random dropping, unless otherwise mentioned.

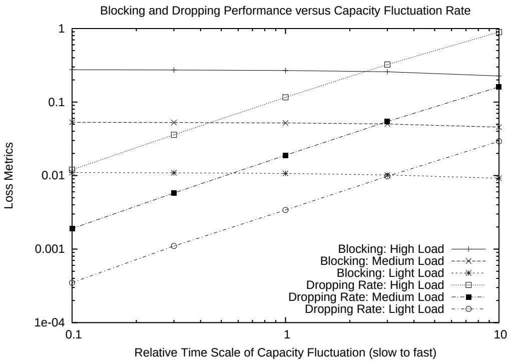

Figure 4. Loss metrics vs. rate of capacity fluctuation

We conducted four experiments.

The first experiment is designed to study the effects of the traffic load (λ/μ) and the capacity fluctuation timescale (λ_f/μ) on loss metrics. We consider three different load levels (High: $\lambda/\mu = 9.0$; Medium: $\lambda/\mu = 4.25$; Low: $\lambda/\mu = 2.10$), and the timescale λ_f/μ varies from 0.1 to 10. All other parameters are the same as defaults. The results are reported in Figure 4. From the figure, we see that as the capacity fluctuates more frequently, the dropping rate increases significantly, however the blocking probability decreases slightly. Note $\sigma_b < 1$ in this experiment.

The second experiment focuses on the effect of distribution F_c on the dropping rate. The service time distribution F_b is Gamma ($k = 2$). We consider a set of different distributions for F_c, where the parameter λ_f varies from 0.1 to 10. The first two choices are: (1) an exponential distribution with rate λ_f;

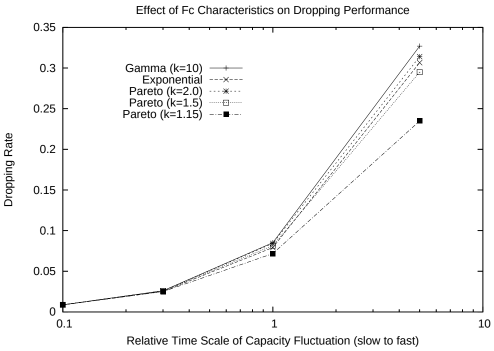

Figure 5. Effect of F_c on the dropping rate

and (2) a Gamma distribution with density function

$$x^{k-1}\frac{e^{-x/\xi}}{\xi^k\Gamma(k)} \text{ for } x > 0,$$

where the shape parameter $k = 10$ and $\xi = 0.1/\lambda_f$. The mean for this Gamma distribution is $1/\lambda_f$. For these two choices, we will write $F_c \sim$ Exp and $F_c \sim$ Gamma respectively. Other choices are taken from the family of Pareto distributions, with density function

$$(\kappa/x_m)(x_m/x)^{\kappa+1} \text{ for } x > x_m, \tag{20}$$

and 0 otherwise. We refer to κ as the shape parameter: κ takes the values of 1.15, 1.5 or 2.0; x_m is chosen such that the mean fluctuation rate λ_f varies from 0.1 to 10. For these choices, we will write $F_c \sim \text{Pareto}(\kappa = 1.15)$, $F_c \sim \text{Pareto}(\kappa = 1.5)$, $F_c \sim \text{Pareto}(\kappa = 2.0)$. Note that the Gamma distribution and the Pareto distribution are two families that have quite different characteristics. A Gamma distribution is light-tailed. A Pareto(κ) distribution is heavy-tailed, and when κ decreases, the tail becomes heavier.

The results are reported in Figure 5. When the timescale $\lambda_f/\mu < 1$, dropping rates for different F_c are almost the same. When $\lambda_f/\mu > 1$, it is in the case $F_c \sim \text{Pareto}(\kappa = 1.15)$ (the heaviest tail) that we see the lowest dropping rate, while in the case $F_c \sim \text{Gamma}(k = 10)$ we see higher dropping rate than in the case when F_c is an exponential distribution. These are consistent with earlier analysis using the approximate model.

In the third experiment, we study the effect of details of F_b (beyond mean) on the blocking. We choose $\lambda_f/\mu = 1$. Particularly, we examine the system behaviour when the coefficient of variation grows larger (specifically $\sigma_b > 1$, unlike in first two experiments, where $\sigma_b < 1$). For simplicity, we examine only the situation that F_b is taken from a two-phase Cox distribution; for this situation, we may use the alternative parameterization (σ_b, v) for F_b. We choose four different configurations. For each configuration, a curve of blocking probability is generated when σ_b^2 varies from 1 to 100, with the configuration annotated. The results are reported in Figure 6. For this experiment, in the region $\sigma_b^2 \leq 1$, different configurations of F_c and the v parameter (with the same λ_f) produce little difference in the blocking probability. When σ_b^2 is larger, the difference becomes more obvious. The blocking probability decreases while v parameter increases.

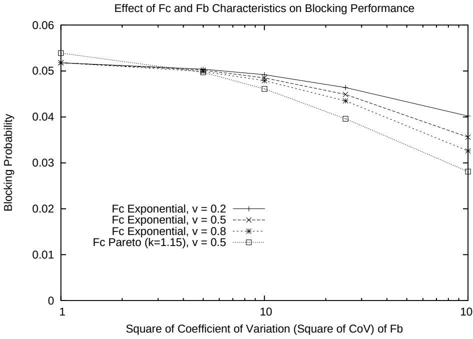

Figure 6. Effect of F_b and F_c characteristics on the blocking

In the fourth experiment, we look at the impact of different dropping policies. We use the default setting, except that distribution F_b is taken from a two-phase Cox distribution. Using the alternative parameterization, we set the parameter $v = 0.2$ and σ_b^2 varying from 1 to 100. The following dropping policies are considered: (1) the random dropping; (2) dropping with a fixed fraction: $\phi_1 = 0.5, \phi_2 = 0.5$, (3) similar to the second policy except that $\phi_1 = 0.8, \phi_2 = 0.2$; and (4) similar to the second policy except that $\phi_1 = 0.9, \phi_2 = 0.1$. One curve is generated for each policy. The results are reported in Figure 7. We note that for large σ_b, the blocking probability decreases more when victims are chosen more often from phase 2, which is the slow phase.

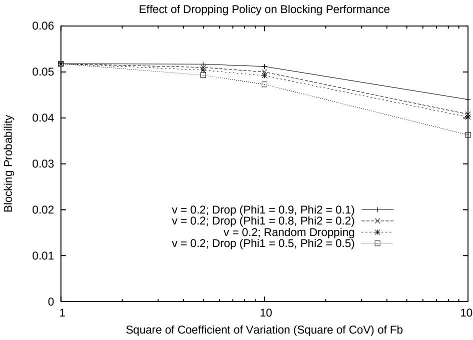

Figure 7. Effect of dropping policy on blocking

Lastly, we check on the consistency of the two models in this paper. The trend (from the detailed model) in the change

Table 1. Effect of F_b distribution: comparison of models

Parameters in F_b	% reduction in $E[T_b']$	% reduction in p_{block}
$\sigma_b^2 = 1$	0	almost none
$\phi_2 = 0.2, \sigma_b^2 = 5$	2.8	7.4 - 9.1
$\phi_2 = 0.2, \sigma_b^2 = 10$	6.3	10 - 17
$\phi_2 = 0.5, \sigma_b^2 = 5$	7.0	11 - 18

of blocking and dropping when F_b and F_c varies, is exactly as we have analyzed using the approximate model. A quantitative comparison follows regarding different choices of F_b, the dropping-induced speedup effect and reduction of p_{block}. We can calculate the reduction of time spent in the system (i.e., $E[T_b']$) using Formula (7) in Section 3.1, and we find out the reduction in the blocking probability (p_{block}) from the detailed model. We choose the following parameters: $c_M = 10$, $\lambda/\mu = 4.25, \lambda_f/\mu = 1$; for this case $r_{\text{drop}} \approx 0.07$. Parameters σ_b^2 and v relating to F_b, together with the results are listed in Table 1. We see that the decrease in $E[T_b']$ is reflected in the reduction in p_{block}.

7. SUMMARY AND OUTLOOK

The model studied in this paper provides a useful tool for performance study in areas such as rapidly developing cellular networks. In such networks, the system capacity may be affected through many factors of complex nature such as communication link downtime, guard channels to protect high-priority users or loss sensitive applications, thermal noise and fading effect in wireless channels. Thus modeling the capacity fluctuation by a general distribution is necessary. For some of the individual capacity-affecting factors mentioned above, there has been some research. However, we are more interested to present a framework, from which we can study the performance implications for any combination of these individual factors. The stochastic queue model is an abstraction from these scenarios as well as others.

To address the issues above, we formulate the model of a loss queueing system with stochastic capacity. The MRGP model is developed to calculate the exact loss metrics. We also developed an approximate model to gain insight and infer the performance implications of system inputs, especially for the "fine characteristics". For service times, we limit the discussion to two-phase Cox distributions in the approximate analysis and in numerical examples, but the method developed applies to general phase-type distributions.

An obvious observation is that the dropping rate increases fairly sharply as λ_f (rate of capacity fluctuation) increases. This holds consistently irrespective of distributions F_b and F_c. In contrast, the blocking probability responds differently to the increase of λ_f for different situations, ranging from being negligible (Figure 4, where $\sigma_b < 1$) to being significant (Figure 6 where σ_b is large).

We focus on the system's functional behaviour w.r.t. the fine characteristics of F_b and F_c. On this aspect, there are several main findings:

(a) Regarding the dropping rate, the difference in F_c (provided the mean is the same) has a modest to medium effect. The dropping rate is typically lower when the distribution F_c has a heavier-tail (Figure 5).

(b) Provided that coefficient of variation (CoV) in F_b is small ($\sigma_b \leq 1$), there is relatively little change in p_{block} when different distributions F_b and F_c are assumed in the model. A particular case is when F_b is exponentially distributed. For this situation, p_{block} is insensitive to details of F_c as well as insensitive to the dropping policy.

(c) For sufficiently large CoV in F_b ($\sigma_b >> 1$), we observe significantly lower blocking probability than otherwise. The behaviour in this situation is complicated. The change in p_{block} depends on σ_b as well as on the details of F_b beyond the first two moments (Figure 6) and on the dropping policy (Figure 7).

This work reveals some interesting and complicated behaviour for a stochastic capacity loss system. Qualitatively, we show that in a stochastic capacity system, large σ_b can be beneficial in the reduction of blocking. However, in the random dropping policy, customers in the slow phase (interpreted as customers having large unfinished workload) are more likely to be dropped: a price paid to reduce blocking.

These findings have practical significance, e.g., regarding capacity planning. We can identify certain regimes in which some loss metrics show the approximate insensitivity. One example is that dropping rate is approximately insensitive to details of F_c beyond mean, if capacity fluctuation is slow ($\lambda_f/\mu < 1$). Another example is that p_{block} is approximately insensitive to the distribution of F_b provided $\sigma_b \leq 1$. For these regimes, we expect that the simpler model developed in Section 3, after some adaptation, may be applied to evaluate loss metrics. This deserves further attention.

For future work, we also plan to further illustrate the applicability of the stochastic capacity queue, by deriving the performance implications for specific applications that have capacity variations.

ACKNOWLEDGEMENTS

Financial support for this work was provided by iCORE (Informatics Circle of Research Excellence) in the Province of Alberta, and by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

REFERENCES

[1] A. Alfa and B. Liu. "Performance analysis of a mobile communication network: the tandem case." *Computer Communications*, 27:208 – 221, 2004.

[2] N. Antunes, C. Fricker, F. Guillemin, and Ph. Robert. "Integration of streaming services and TCP data transmission in the internet." *Performance Evaluation*, 62:263–277, 2005.

[3] A. Erlang. "Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges." In E. Brockmeyer, H.L. Halstrom, and A. Jenson, editors, *The life and works of A. K. Erlang*. 1948. First published in Danish, 1917.

[4] R. Fricks, M. Telek, A. Puliafito, and K. Trivedi. "Markov renewal theory applied to performability evaluation." In K. Bagchi and G. Zobrist, editors, *State-of-the-art in Performance Modeling and Simulation, Modeling and Simulation of Advanced Computer Systems: Applications and Systems*, pages 13 – 226. Gordon & Breach, NJ, 1998.

[5] R. German, editor. *Performance Analysis of Communication Systems: Modeling with non-Markovian Stochastic Petri Nets*. Wiley, Chichester, 2000.

[6] R. German, J. Luethi, and M. Telek, editors. *Proceedings of the Fifth International Workshop on Performability Modelling and Communication Systems*. 2001.

[7] W. Grassmann. "Finding transient solutions in Markovian event systems through randomization." In W. Stewart, editor, *Numerical solution of Markov chains*. Marcel Dekker, New York, 1991.

[8] V. Kulkarni. *Modelling and analysis of stochastic systems*. Chapman & Hall, 1995.

[9] J. Luo and A. Alfa. "Performance analysis of a base station in a mobile network with realistic handoff calls." In *Proceedings of Symposium of Performance Evaluation for Computer and Communication Systems (SPECTS) 2005*, pages 235–244.

[10] W. Massey and R. Srinivasan. "A packet delay analysis for cellular digital packet data." *IEEE J. on Selected Areas in Communications*, 15 (7):1364 – 1372, 1997.

[11] J. F. Meyer. "Performability modelling of distributed real-time systems." In G. Iazeolla, P. J. Courtois, and A. Hordijk, editors, *Mathematical Computer Performance and Reliability*. North Holland, 1984.

[12] N. Mikou. "A two-node Jackson's network subject to breakdowns." *Stochastic Models*, 4:523 – 552, 1988.

[13] I. Mitrani and B. Avi-Itzhak. "A many-server queue with service interruptions." *Operation Research*, 16:628 – 638, 1968.

[14] I. Mitrani and P. Wright. "Routing in the presence of server breakdowns." *Performance Evaluation*, 20:151–164, 1994.

[15] R. Pyke. "Markov renewal processes." *Annals Math. Statist.*, 32:1231 – 1242, 1961.

[16] Y. Sasaki, H. Imai, M. Tsunoyama, and I. Ishii. "Approximation method for probability distribution functions using Cox distribution to evaluate multimedia systems." In *Proc. of 2001 Pacific Rim Sympos. on Dependable Computing*, pages 333–340, 2001.

[17] H. Sun and C. Williamson. "Simulation evaluation of call dropping policies for stochastic capacity networks". In *Proceedings of Symposium of Performance Evaluation for Computer and Telecommunication Systems (SPECTS) 2005*, pages 327–336, 2005.

[18] K. Trivedi, X. Ma, and S. Dharmaraja. "Performability modelling of wireless communication systems." *Internat. J. of Communication Systems*, 16:561–577, 2003.

[19] Y. Wu and C. Williamson. "Impacts of data call characteristics on multi-service CDMA system capacity". *Performance Evaluation*, 62:83 – 99, 2005.

AUTHOR BIOGRAPHIES

Jingxiang Luo works as a Research Associate at the Department of Computer Science, University of Calgary, Canada. He received his Ph.D. (2004) and M.Sc. (2000) in Computer Science, both from the University of Saskatchewan. His interests are in the areas of performance evaluation, queueing and network theories, modelling and simulation methodology in stochastic process. Currently, he is investigating the traffic analysis and performance modelling in mobile and wireless networks.

Carey Williamson is an iCORE Chair 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 performance.