Computers and Mathematics with Applications 64 (2012) 1352–1363

Contents lists available at SciVerse ScienceDirect

Computers and Mathematics with Applications

journal homepage: www.elsevier.com/locate/camwa

Estimating the loss probability under heavy traffic conditions

Chia-Hung Wang [a,*], Hsing Paul Luh [b]

[a] College of Management, National Chiao Tung University, No. 1001, Ta Hsueh Road, Hsinchu 30010, Taiwan, ROC

[b] Department of Mathematical Sciences, National Chengchi University, No. 64, Sec. 2, ZhiNan Road, Wen-Shan District, Taipei 11605, Taiwan, ROC

ARTICLE INFO

Keywords:
Asymptotic analysis
Multiple-server queue
Loss probability
Heavy traffic

ABSTRACT

This paper studies a multiple-server queueing model under the assumptions of renewal arrival processes and limited buffer size. An approximation for the loss probability and the asymptotic behavior are studied under the heavy traffic conditions. We present an asymptotic analysis of the loss probability when both the arrival rate and number of servers approach infinity. In illustrative examples, the loss probabilities are estimated with heavy traffic under three common distributions of inter-arrival times: exponential, deterministic and Erlang-r distributions, respectively.

1. Introduction

Motivated by the growing development of modern telecommunication systems, the studies of queueing systems with many servers and especially analysis of the loss probability have been conducted significantly under investigation [1–4]. Connections over Internet are typically generated in mounting up population of users independently communicating with an equivalently large population of servers and correspondents for a variety of applications [5]. According to traffic demand and network management settings, it requires suitable bandwidth allocation of individual connection to achieve guaranteed Quality of Service (QoS) level (see [6] for example). Due to the budget constraint, it is too costly for the network service providers to assert a 100% guaranteed availability for all connections at any time although it is the network managers' mission to provide available servers with suitable bandwidth. This is also not necessary because traffic flow fluctuates with time, and connections do not last forever but occur at random times and vanish in the network once the corresponding digital document has been transferred completely [4].

Hence, it is desirable to bring out an analytic stochastic model to determine the loss probability as an important performance measure of network systems. For example, Maglaras and Zeevi [7] studied the equivalent behavior of communication systems in a single-class Markovian model under revenue and social optimization objectives. Faragó [2] gave an estimated loss probability and link utilization for general multi-rate and heterogeneous traffic, where the individual bandwidth demands may aggregate in complex ways. Bruni et al. [8] designed a connection admission control procedure for resource management on a telecommunication network. Taking the loss probability into account, Wang and Luh [4] presented a solution analysis of bandwidth allocation on communication networks, where the authors obtained monotone and concave properties of the loss probability in $M/G/s/s$ under the Erlang loss model.

In real-world communication networks, it becomes difficult to compute numerically the loss probability for large number of servers even though by computers [9,10]. As mentioned in [11], the main drawback with exact methods of analyzing the $GI/G/s/s$ queues is the often-excessive computation times required. Indeed, many problems become intractable with small to medium-sized values of number of servers s [2].

Choi et al. [1] and Kim and Choi [12] obtained some results related to the $GI/M/s/n$ and $GI^X/M/s/n$ queues with batch size X, where s is the fixed (and small) number of servers and n is a variable denoting the capacity of waiting space. As the

* Corresponding author. Tel.: +886 953230277; fax: +886 3 5715544.
E-mail addresses: chwang728@nctu.edu.tw, jhwang728@hotmail.com, 93751502@nccu.edu.tw (C.-H. Wang), slu@nccu.edu.tw (H.P. Luh).

0898-1221/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.camwa.2012.03.081

waiting capacity n increases to infinity, Choi et al. [1] obtained the estimation for the convergence rate of the stationary $GI/M/s/n$ queue-length distribution to the stationary queue-length distribution of the $GI/M/s$ queueing system. In [12], Kim and Choi gave an analysis of the loss probability in the $GI^X/M/s/n$ queueing systems. Recently, Abramov [13] provided an asymptotic analysis of the loss probability of the $GI/M/s/n$ queue as the waiting capacity n approaches infinity. However, in those papers, the number of servers s is fixed and hence the traffic intensity is also fixed.

The main contribution of this paper is the asymptotic analysis of the loss probability as both the arrival rate and number of servers approach infinity. We consider the $GI/M/s/s$ queueing systems as the number of servers s increases to infinity, where the traffic intensity depends on s. The aim of this paper is to provide an approximation for the loss probability as the number of servers is huge. We present an approximation for the loss probability with the stationary probability of $GI/M/\infty$ queues. Computational effort with guaranteed precision level of this approximation is much less than the one for determining the exact value of the loss probability in $GI/M/s/s$ queueing systems as s is large. Needless to say, it has a significant advantage when solving the huge matrix is impossible.

The remainder of the paper is organized as follows. Section 2 presents the assumptions and definitions of the proposed queueing model under the heavy traffic conditions. An approximation of the loss probability with heavy-traffic limits is introduced in Section 3. Three examples are given in Section 4 to demonstrate the derivation of the approximated loss probabilities under assumptions of exponential, deterministic and Erlang-r distributions of the inter-arrival times, respectively. Sensitivity analysis with numerical illustrations are conducted in Section 5. Concluding remarks are drawn in Section 6. We give proofs for each proposition and theorem while providing most of them in Appendices in order not to interrupt the flow of presentation.

2. A queueing model under heavy traffic conditions

The assumptions of renewal arrival process, exponential service times, finite servers and limited buffer size are commonly used in queueing systems, e.g., [1,11,13,14], etc. In this paper, we assume that the inter-arrival times of customers are independent and identically distributed (i.i.d.) random variables with cumulative distribution function (c.d.f.) $A(t)$, probability density function $a(t)$ for $t > 0$, and mean $1/\lambda$. We also assume that the sojourn times are i.i.d. random variables following exponential distribution with mean $1/\mu$, which corresponds to the packet transmission time. Suppose that the inter-arrival time and sojourn time are mutually independent. Customers occupy those s servers in the order they occur, that is, the service discipline is First Come First Served.

Network managers are interested in knowing the behavior of the loss probability in heavy loaded systems, and it is natural to look for insight into system performance by considering the asymptotic behaviors as the number of servers is allowed to increase. The most commonly used limit theorem for large-scale queueing systems under heavy traffic is that in Halfin and Whitt [14], who considered the $GI/M/s$ queue as $s \to \infty$ and $\rho_s \to 1$ such that

$$(1 - \rho_s)\sqrt{s} \to \gamma \tag{1}$$

with $-\infty < \gamma < \infty$. For the $M/M/s$ queue with $\rho_s < 1$, they showed that the steady-state probability that a customer must wait in the queue approaches a limit κ with $0 < \kappa < 1$ as $s \to \infty$ if and only if $0 < \gamma < \infty$. For the $GI/M/s$ queue, they showed that a properly centered and normalized version of the queue length process converges to a one-dimensional diffusion. Several applications under this heavy-traffic assumption can also be found in [3,15], and reference therein.

Here, we consider a sequence of queueing models indexed by the number of servers, s. Assume that we have the mean arrival rate $\lambda_s = s\mu - \gamma\mu\sqrt{s}$, where $0 < \gamma < \sqrt{s}$, the traffic intensity of the queueing system indexed by s servers is defined as follows.

Definition 1. The *traffic intensity* of the system is defined as the fraction of the time in which servers are occupied. Namely, the traffic intensity of the system is

$$\rho_s \triangleq \frac{\lambda_s}{s\mu} = 1 - \gamma/\sqrt{s}, \tag{2}$$

which is the average occupancy of s servers in the system.

In such a case, there exists an interesting nondegenerate limit in Halfin–Whitt heavy traffic regimes [14,15], namely, $\rho_s \to 1$ and $(1 - \rho_s)\sqrt{s} \to \gamma$ as $s \to \infty$.

Assumption 1. As the number of servers, s, increases to infinity, we assume that the traffic intensities ρ_s approach 1 from below, i.e.,

$$\lim_{s\to\infty} \rho_s = 1. \tag{3}$$

Assumption 1 is the so-called heavy traffic condition, which is taken from the Halfin–Whitt heavy traffic regimes [14]. Throughout the paper, we will determine the loss probability and derive its asymptotic analysis under the stability condition $\rho_s < 1$ but close to 1 when s approaches infinity. Assumption 1 explicitly is applied to the main results, e.g., Proposition 2, Theorem 4, Proposition 5, and Theorem 5.

Consider the characteristic equation

$$z = \int_0^\infty e^{-(s\mu - s\mu z)t} dA(t), \tag{4}$$

where the variable z is assumed to be real, and Φ_s is denoted the least in absolute value root of the characteristic equation (4). It is well-known that the root Φ_s belongs to the open interval $(0, 1)$ if the traffic intensity $\rho_s < 1$, and it is equal to 1 otherwise [13,16]. Once we obtain the root Φ_s and its limit (to be discussed later), the loss probability $\mathcal{P}(\rho_s, s)$ can be estimated immediately for any value of s. An approximation is provided in the following section for the loss probability of $GI/M/s/s$ queueing systems, where the traffic intensity ρ_s depends on s.

3. Approximation of the loss probability with heavy-traffic limits

The objective of this section is to estimate the loss probability $\mathcal{P}(\rho_s, s)$ that all s servers are occupied. Let $\beta_n(s)$, $n = 0, 1, 2, \ldots$, be the probability that n customers have completed service when there are s servers in the system. From [16], we have the following lemma.

Lemma 1. *If the c.d.f. of inter-arrival times, $A(t)$, is non-lattice, then the integral*

$$\beta_n(s) = \int_0^\infty e^{-s\mu t} \frac{(s\mu t)^n}{n!} dA(t), \tag{5}$$

exists for each $n = 0, 1, 2, \ldots$.

Let $A^*(z)$ be the Laplace–Stiltjes' transform of the c.d.f. $A(t)$. The generating function of $\beta_i(s)$, $i = 0, 1, 2, \ldots$, is given by

$$\sum_{n=0}^{\infty} \beta_n(s) z^n = A^*(s\mu - s\mu z), \quad |z| \le 1. \tag{6}$$

Next, we consider the characteristic equation

$$z - A^*(s\mu - s\mu z) = 0, \quad |z| \le 1, \tag{7}$$

which is equivalent to (4).

Lemma 2. *If $A(t)$ is non-lattice, then there exists a positive real number $0 < \Phi_s < 1$ such that*

$$\Phi_s - A^*(s\mu - s\mu\Phi_s) = 0, \tag{8}$$

equivalently, which may be written as

$$\Phi_s - \sum_{n=0}^{\infty} \beta_n(s) \Phi_s^n = 0. \tag{9}$$

Next, we are going to estimate the loss probability by investigating the asymptotic behavior of the stationary probabilities of s customers in service, as $s \to \infty$. Let $\mathbf{P}^{(s)} = (P_0^{(s)}, \ldots, P_s^{(s)})$ be the stationary probability vector of customers in service. It is easy to see that $\mathcal{P}(\rho_s, s) = P_s^{(s)}$. Let $\mathbf{T}^{(s)} = [P_{m,n}]_{(s+1)\times(s+1)}$ be the one-step transition probability matrix of the embedded Markov chain. The one-step transition probability matrix $\mathbf{T}^{(s)}$ can be represented as follows

$$\begin{bmatrix} P_{0,0}^{(\infty)} & P_{0,1}^{(\infty)} & 0 & \cdots & 0 \\ P_{1,0}^{(\infty)} & P_{1,1}^{(\infty)} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ P_{s-2,0}^{(\infty)} & P_{s-2,1}^{(\infty)} & \cdots & P_{s-2,s-1}^{(\infty)} & 0 \\ P_{s-1,0}^{(\infty)} & P_{s-1,1}^{(\infty)} & \cdots & P_{s-1,s-1}^{(\infty)} & \beta_0(s-1) \\ P_{s,0} & P_{s,1} & \cdots & P_{s,s-1} & \beta_0(s) + \beta_1(s) \end{bmatrix}, \tag{10}$$

where those elements $P_{m,n}^{(\infty)}$, $0 \le m \le s-1$, $0 \le n \le s-1$, and $P_{s,n}$, $0 \le n \le s-1$, can be explicitly determined in terms of model parameters. It can be found that the stationary probability vector $\mathbf{P}^{(s)} = (P_0^{(s)}, \ldots, P_s^{(s)})$ is the unique solution of $\mathbf{P}^{(s)}\mathbf{T}^{(s)} = \mathbf{P}^{(s)}$ and $\sum_{n=0}^{s} P_n^{(s)} = 1$. By applying the coupling method in [16] on the limiting distribution of $\mathbf{P}^{(s)}$ as $s \to \infty$, it gives the following lemma, which will be borrowed for further development in this paper.

Lemma 3. *The stationary distribution of $\mathbf{P}^{(s)} = (P_0^{(s)}, \ldots, P_s^{(s)})$ goes weakly to that of the $GI/M/\infty$ queue as $s \to \infty$, i.e.,*

$$\lim_{s\to\infty} P_n^{(s)} = P_n^{(\infty)}, \quad n = 0, 1, \ldots, s. \tag{11}$$

Then, similar to the derivation in [1], it gives

$$\mathcal{P}(\rho_s, s) = P_0^{(s)} \frac{\beta_0(s)^s}{(1-\beta_0(s)-\beta_1(s))(1-\beta_1(s))^{s-1}}. \tag{12}$$

Proposition 1. *If $A(t)$ follows one of the exponential, deterministic, or Erlang-r distributions, then it gives the limit*

$$\lim_{s\to\infty} \frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)} = C_1, \tag{13}$$

where C_1 is a constant number.

Proposition 2. *If $A(t)$ is non-lattice and $(1-\rho_s)\sqrt{s} \to \gamma$ as $s \to \infty$, then there exists a positive number $\varepsilon > 0$ such that, for all $0 < p < \varepsilon$, we have*

$$\left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s \approx O(\rho_s^p) = C_2, \tag{14}$$

as $s \gg 1$, where $O(\cdot)$ is the big O notation used to describe the limiting behavior of a function.

Theorem 4. *Consider $GI/M/s/s$ queueing systems with non-lattice c.d.f. of inter-arrival times. Assume that $(1-\rho_s)\sqrt{s} \to \gamma$ as $s \to \infty$. Then, as $s \gg 1$, we have the approximation of the loss probability*

$$\mathcal{P}(\rho_s, s) \approx P_0^{\infty} C_1 C_2, \tag{15}$$

where P_0^{∞} is the stationary probability that the system is empty in $GI/M/\infty$ queues.

Proof. From (12), it can be derived that

$$\mathcal{P}(\rho_s, s) = P_0^{(s)} \frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)} \left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s. \tag{16}$$

First, by Lemma 3, it gives that $P_0^s \to P_0^{\infty}$ as $s \to \infty$. Next, by Proposition 1, there exists a constant number C_1 such that

$$\frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)} \to C_1$$

as $s \to \infty$. In addition, by Proposition 2, we have

$$\left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s \to C_2$$

as $s \gg 1$. Hence, the approximation for the loss probability can be determined as $s \gg 1$. □

In this section, we have introduced an analytic approach for determining the loss probability as the number of servers, s, is large. An approximation is provided in Theorem 4 for the loss probability $\mathcal{P}(\rho_s, s)$ under the heavy traffic conditions that $(1-\rho_s)\sqrt{s} \to \gamma$ as $s \to \infty$. By Propositions 1 and 2, the limits C_1 and C_2 can be computed with probabilities $\beta_0(s)$ and $\beta_1(s)$. Then, by Theorem 4, we can estimate the loss probability of the $GI/M/s/s$ queueing system for large s with C_1, C_2, and the stationary probability $P_0^{(\infty)}$. For $M/M/\infty$ queueing systems, it gives $P_0^{(\infty)} \approx e^{-\frac{\lambda s}{\mu}}$. For practical purposes, three examples are given in the following section to demonstrate the derivation of the approximated loss probabilities.

4. Three illustrative examples

In this section, we determine the probability $\beta_n(s)$, limit constants C_1 and C_2, and the loss probability $\mathcal{P}(\rho_s, s)$ under three common distributions for inter-arrival times: exponential, deterministic and Erlang-r distributions, respectively.

Example 1 (*Exponential Distribution*)**.** Suppose that the inter-arrival time is exponentially distributed with parameter λ. Then

$$\begin{aligned}\beta_n(s) &= \int_0^{\infty} e^{-s\mu t} \frac{(s\mu t)^n}{n!} \lambda e^{-\lambda t} dt \\ &= \frac{\lambda}{n!(1+\rho_s)^n} \int_0^{\infty} e^{-\lambda\left(1+\frac{1}{\rho_s}\right)t} \left(\lambda\left(1+\frac{1}{\rho_s}\right)t\right)^n dt \\ &= \frac{\rho_s}{(1+\rho_s)^{n+1}},\end{aligned} \tag{17}$$

for $n = 0, 1, 2, \ldots$, where $\Gamma(n+1) = n!$ is the Gamma function. From (17), we derive that

$$\beta_0(s) = \frac{\rho_s}{1+\rho_s} \tag{18}$$

and

$$\beta_n(s) = \frac{1}{1+\rho_s}\beta_{n-1}(s) = \frac{\rho_s}{(1+\rho_s)^{n+1}}, \tag{19}$$

for $n = 1, 2, \ldots$.

From (19), the limit of the probability $\beta_n(s)$ can be derived in the following result.

Corollary 1. *If $A(t)$ is the c.d.f. of exponential inter-arrival times and $\rho_s < 1$ for all s, then the limit of the probability $\beta_n(s)$ exists as $s \to \infty$, and*

$$\lim_{s\to\infty} \beta_n(s) = \frac{1}{2^{n+1}}, \tag{20}$$

for $n = 0, 1, 2, \ldots$.

Example 2 (*Deterministic Case*)**.** Assume that the inter-arrival time is deterministic with constant $1/\lambda = d$,

$$a(t) = \delta(t-d) = \begin{cases} \infty, & t = d, \\ 0, & t \neq d. \end{cases}$$

In this case, the number of served customers during an inter-arrival time d follows a Poisson distribution with mean $1/\rho_s = s\mu d$. So, the probability $\beta_n(s)$ is determined as the distribution with parameter $s\mu d$, i.e.,

$$\beta_n(s) = \frac{(s\mu d)^n e^{-s\mu d}}{n!} \tag{21}$$

for $n = 0, 1, 2, \ldots$. From (21), we derive that

$$\beta_0(s) = e^{-1/\rho_s}, \tag{22}$$

and

$$\beta_n(s) = \frac{1}{n\rho_s}\beta_{n-1}(s) = \frac{e^{-1/\rho_s}}{n!\rho_s^n} \tag{23}$$

for $n = 1, 2, \ldots$.

In the following result, we determine the limit of the probability $\beta_n(s)$ from (23).

Corollary 2. *If the inter-arrival time is constant and $\rho_s < 1$ for all s, then the limit of the probability $\beta_n(s)$ exists as $s \to \infty$, and*

$$\lim_{s\to\infty} \beta_n(s) = \frac{e^{-1}}{n!}, \tag{24}$$

for $n = 0, 1, 2, \ldots$.

Example 3 (*Erlang-r Distribution*)**.** Assume that the inter-arrival time is Erlang-r with mean $1/\lambda$. Then

$$\begin{aligned} \beta_n(s) &= \int_0^\infty e^{-s\mu t}\frac{(s\mu t)^n}{n!}\frac{\lambda^r t^{r-1}e^{-\lambda t}}{(r-1)!}dt \\ &= \frac{\lambda^{n+r}}{n!(r-1)!\rho_s^n}\int_0^\infty e^{-\lambda(1+1/\rho_s)t}t^{n+r-1}dt \\ &= \frac{\rho_s^r}{nB(n,r)(1+\rho_s)^{n+r}}, \end{aligned} \tag{25}$$

for $n = 0, 1, 2, \ldots$, where $B(n, r) = \Gamma(n)\Gamma(r)/\Gamma(n+r)$ is the Beta function. From (25), it is derived that

$$\beta_0(s) = \left(\frac{\rho_s}{1+\rho_s}\right)^r \tag{26}$$

and

$$\beta_n(s) = \frac{n+r-1}{n(1+\rho_s)}\beta_{n-1}(s) = \frac{\rho_s^r}{n!(r-1)!(1+\rho_s)^{n+r}}\Gamma(n+r), \tag{27}$$

for $n = 1, 2, \ldots$.

From (27), we have the limit of the probability $\beta_n(s)$ as follows.

Corollary 3. *If $A(t)$ is the c.d.f. of Erlang-r inter-arrival times for $r < \infty$ and $\rho_s < 1$ for all s, then the limit of the probability $\beta_n(s)$ exists as $s \to \infty$, and*

$$\lim_{s\to\infty} \beta_n(s) = \frac{\Gamma(n+r)}{n!(r-1)!2^{n+r}}, \tag{28}$$

for $n = 0, 1, 2, \ldots$.

For each non-lattice $A(t)$ and fixed positive integer s, there exists a sufficiently large integer N_s such that $\beta_n(s)$ is non-increasing for all $n \geq N_s$.

Proposition 3. *Under the condition that the c.d.f. of inter-arrival times, $A(t)$, is non-lattice, we find that* (i) *if $A(t)$ is the c.d.f. of exponential inter-arrival times, the probability $\beta_n(s)$ is non-increasing for all $n \geq 0$;* (ii) *if the inter-arrival time is constant, there exists an integer $N_d = \lceil 1/\rho_s \rceil$ such that the probability $\beta_n(s)$ is non-increasing for all $n \geq N_d$, where the ceiling function $\lceil x \rceil$ outputs the smallest integer greater than or equal to x;* (iii) *if $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, there exists an integer*

$$N_r = \left\lceil \frac{r-1}{\rho_s} \right\rceil$$

such that the probability $\beta_n(s)$ is non-increasing for all $n \geq N_r$.

Proof. (i) Because $\rho_s > 0$ for each positive integer s, it is clear that

$$\beta_n(s) = \frac{\rho_s}{(1+\rho_s)^{n+1}} \geq \frac{\rho_s}{(1+\rho_s)^{n+2}} = \beta_{n+1}(s)$$

for all $n = 0, 1, 2, \ldots$. So, $\beta_n(s)$ is non-increasing for all $n \geq 0$.

(ii) If $n \geq N_d = \lceil 1/\rho_s \rceil$, it implies $n\rho_s \geq 1$ and it can be derived that

$$\beta_{n-1}(s) = \frac{e^{-1/\rho_s}}{(n-1)!\rho_s^{n-1}} \geq \frac{e^{-1/\rho_s}}{n!\rho_s^n} = \beta_n(s).$$

Hence, $\beta_n(s)$ is non-increasing for all $n \geq N_d = \lceil 1/\rho_s \rceil$.

(iii) It can be derived that

$$\beta_n(s) = \frac{n+r-1}{n(1+\rho_s)}\beta_{n-1}(s)$$

for $n = 1, 2, \ldots$. If $n \geq N_r = \lceil \frac{r-1}{\rho_s} \rceil$, it implies $\beta_n(s) \leq \beta_{n-1}(s)$. Therefore, $\beta_n(s)$ is non-increasing for all $n \geq N_r = \lceil \frac{r-1}{\rho_s} \rceil$. □

Proposition 4. *If $A(t)$ follows one of the exponential, deterministic, or Erlang-r distributions, we have the limit in (13) as follows.* (i) *If $A(t)$ is the c.d.f. of exponential inter-arrival times, it gives the limit*

$$C_1^{\text{Exp}} = 3. \tag{29}$$

(ii) *If the inter-arrival time is constant, it gives the limit*

$$C_1^{\text{Det}} = \frac{e-1}{e-2}, \tag{30}$$

where the number e is the base of the natural logarithms. (iii) *If $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, it gives the limit*

$$C_1^{\text{Er}} = \frac{2^{r+1}-r}{2^{r+1}-2-r}, \tag{31}$$

for positive integer $r \geq 2$.

Proof. (i) Because $\rho_s \to 1$ as $s \to \infty$, it gives the following limit

$$C_1^{\text{Exp}} = \lim_{s\to\infty} \frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)} = \lim_{s\to\infty} (\rho_s^2 + \rho_s + 1) = 3.$$

(ii) As $s \to \infty$, it gives the following limit

$$C_1^{\text{Det}} = \lim_{s\to\infty} \frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)} = \lim_{s\to\infty} \frac{\rho_s e^{1/\rho_s} - 1}{\rho_s e^{1/\rho_s} - \rho_s - 1} = \frac{e-1}{e-2}.$$

(iii) For positive integer $r \geq 2$, it gives the following limit

$$\begin{aligned} C_1^{\text{Er}} &= \lim_{s\to\infty} \frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)} \\ &= \frac{(1+\rho_s)^{r+1} - r\rho_s^r}{(1+\rho_s)^{r+1} - \rho_s^r(1+\rho_s) - r\rho_s^r} \\ &= \frac{2^{r+1} - r}{2^{r+1} - 2 - r}. \quad \square \end{aligned}$$

Proposition 5. *If $A(t)$ is non-lattice and $(1-\rho_s)\sqrt{s} \to \gamma$ as $s \to \infty$, we have the constant number in (14) as follows.* (i) *If $A(t)$ is the c.d.f. of exponential inter-arrival times, it gives*

$$C_2^{\text{Exp}} \approx \left(\frac{\rho_s + \rho_s^2}{1 + \rho_s + \rho_s^2}\right)^s, \quad s \gg 1. \tag{32}$$

(ii) *If the inter-arrival time is constant, it gives*

$$C_2^{\text{Det}} \approx \left(\frac{\rho_s}{\rho_s e^{1/\rho_s} - 1}\right)^s, \quad s \gg 1. \tag{33}$$

(iii) *If $A(t)$ is the c.d.f. of Erlang-r inter-arrival times, it gives*

$$C_2^{\text{Er}} \approx \left(\frac{\rho_s^r(1+\rho_s)}{(1+\rho_s)^{r+1} - r\rho_s^r}\right)^s, \quad s \gg 1. \tag{34}$$

Proof. (i) Because $\rho_s \to 1$ as $s \gg 1$, it gives the following limit

$$C_2^{\text{Exp}} \approx \left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s = \left(\frac{\rho_s + \rho_s^2}{1 + \rho_s + \rho_s^2}\right)^s.$$

(ii) If the inter-arrival time is constant, it gives the limit

$$C_2^{\text{Det}} \approx \left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s = \left(\frac{\rho_s}{\rho_s e^{1/\rho_s} - 1}\right)^s$$

as $s \gg 1$.

(iii) For positive integer $r > 1$, it gives the following limit

$$C_2^{\text{Er}} \approx \left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s = \left(\frac{\rho_s^r(1+\rho_s)}{(1+\rho_s)^{r+1} - r\rho_s^r}\right)^s,$$

as $s \gg 1$. $\square$

Theorem 5. *Consider three queueing systems with inter-arrival times of exponential, deterministic and Erlang-r, $r \geq 2$, distributions respectively. Given the traffic intensity $\rho_s = 1 - \frac{\gamma}{\sqrt{s}} \to 1$ from below, for $0 < \gamma < \sqrt{s}$, as the number of servers $s \gg 1$, we have*

$$\mathcal{P}^{\text{Exp}}(\rho_s, s) \geq \mathcal{P}^{\text{Det}}(\rho_s, s) \geq \mathcal{P}^{\text{Erlang}}(\rho_s, s). \tag{35}$$

Proof. By Theorem 4, the loss probability can be determined as $\mathcal{P}(\rho_s, s) \approx P_0^\infty C_1 C_2$, where the probability P_0^∞ is constant for given fixed s and ρ_s in $GI/M/\infty$ queues. From Proposition 4, we have shown that the sequence of

$$\frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)}$$

converges to $C_1^{\text{Exp}} = 3$ as $s \gg 1$ for the exponential inter-arrival times. In addition, as $s \gg 1$, the sequence converges to $C_1^{\text{Det}} = (e-1)/(e-2)$ for the deterministic inter-arrival times, and it converges to $C_1^{\text{Er}} = (2^{r+1} - r)/(2^{r+1} - 2 - r)$ for the

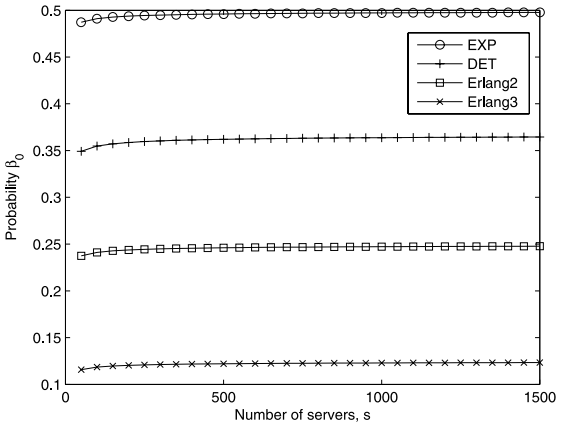

Fig. 1. Probability $\beta_0(s)$ versus a number of servers s.

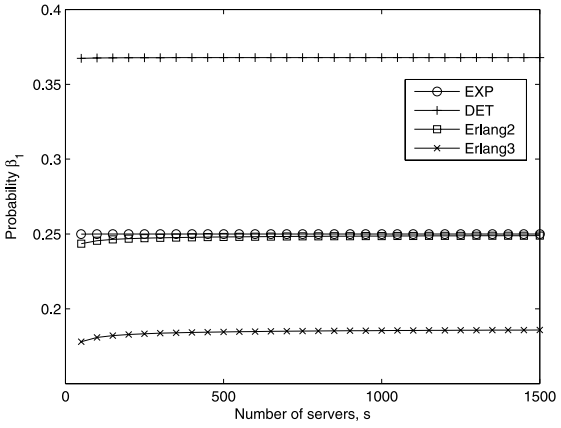

Fig. 2. Probability $\beta_1(s)$ versus a number of servers s.

Erlang-r inter-arrival times. It can be easily checked that

$$3 \geq \frac{e-1}{e-2} \geq \frac{2^{r+1}-r}{2^{r+1}-2-r} \geq \frac{2^{r+2}-(r+1)}{2^{r+2}-2-(r+1)} \tag{36}$$

for all positive integers $r \geq 2$. Hence, we have the inequality $C_1^{\text{Exp}} \geq C_1^{\text{Det}} \geq C_1^{\text{Er}}$. Next, for comparison of limit C_2^{Exp}, C_2^{Det} and C_2^{Er}, we consider the term

$$\frac{\beta_0(s)}{1-\beta_1(s)}$$

in (14) for exponential, deterministic and Erlang-r, inter-arrival times, respectively. It can be derived that the term $\beta_0(s)/(1-\beta_1(s))$ for exponential distributions is the largest, and that is the least for Erlang-r distributions. Then, we have $C_2^{\text{Exp}} \geq C_2^{\text{Det}} \geq C_2^{\text{Er}}$. Therefore, inequality (35) holds by applying approximation (15) in Theorem 4. □

5. Numerical illustrations of heavy-traffic limits

We compare the numerical results of the probability $\beta_n(s)$ and loss probability $\mathcal{P}(\rho_s, s)$ in cases of Exponential, Deterministic, Erlang-2 and Erlang-3 inter-arrival times. Here, the traffic intensity $\rho_s = 1 - \frac{\gamma}{\sqrt{s}}$ is applied from (1), where $\gamma = 0.35$ is given. The number of servers, s, varies from 50 to 1500.

In Fig. 1, it can be observed that the probability $\beta_0(s)$ of Exponential inter-arrival times is always the largest among four examples, and $\beta_0(s)$ of Erlang-3 inter-arrival times is the least. Moreover, the probability $\beta_0(s)$ is increasing as the number of servers s increases. Regarding the probability $\beta_1(s)$, we find that in the Deterministic case it is the largest and it is the least in the Erlang-3 case, which is shown in Fig. 2. It is also observed that the probability $\beta_1(s)$ is increasing as the number of servers s increases.

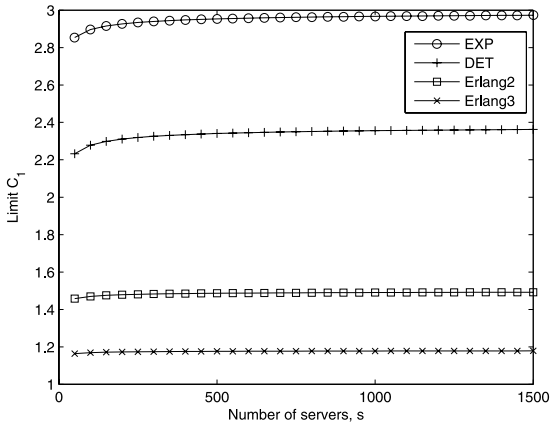

Fig. 3. Limit C_1 versus a number of servers s.

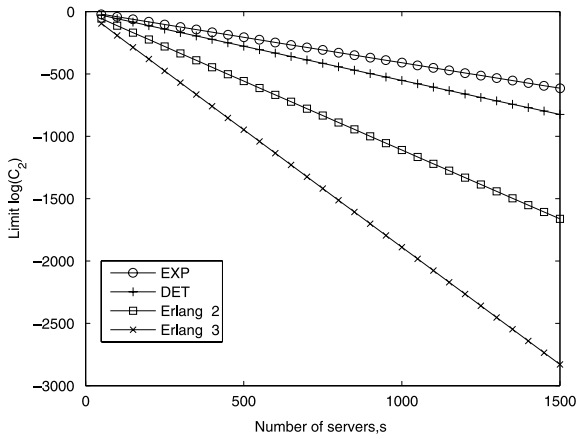

Fig. 4. The logarithm of C_2 versus a number of servers s.

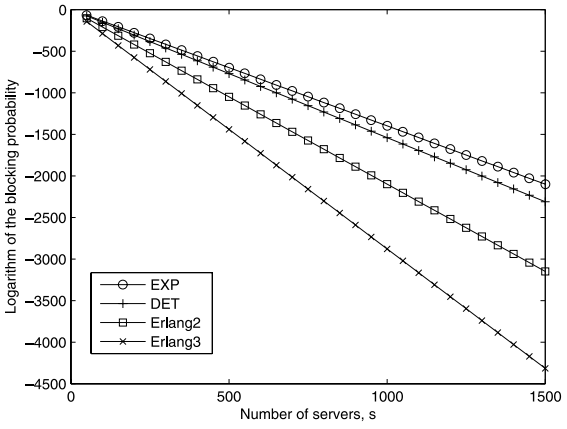

Fig. 5. The logarithm of loss probability $\log(\mathcal{P}(\rho_s, s))$ versus a number of servers s.

From Fig. 3, it can be seen that the limit C_1 is increasing as the number of servers s becomes larger. We find that the limit C_1 determined from Exponential inter-arrival times is always larger than that of Deterministic case, and the limit C_1 of Erlang-3 case is the least. It can be seen in Fig. 4 that the number C_2 is very small for all four examples. Numerical results are observed with the logarithm of C_2, i.e., $\log(C_2)$. It implies that the loss probability becomes very small, which is shown in Fig. 5. Hence, the loss probability is also depicted with the logarithm $\log(\mathcal{P}(\rho_s, s))$. We find that C_2 determined from Exponential inter-arrival times is always larger than others, and C_2 of Erlang-3 case is the least. Similarly, the loss probability under Exponential

inter-arrival time distribution is the largest, and the loss probability determined from Erlang-3 inter-arrival times is the least. The order of those loss probabilities has been shown theoretically in Theorem 5.

6. Conclusion

In this paper, we introduce a heavy-traffic queueing model as the number of servers is huge. An approximation and its asymptotic analysis are derived for the loss probability of the queueing system, where the traffic intensity increases to one from below. In illustrative examples, the loss probabilities are estimated numerically under the assumptions of exponential, deterministic and Erlang-r distributions for the inter-arrival times, respectively. For the class of problems studied with different parameters, it is concluded that the approximation is adequate for practical purposes. The asymptotic analysis of the loss probability could be applied to investigate the optimal buffer size in capacitated communication systems so that the loss probability is kept below a specific threshold. Future works would be conducted in the direction of the design of reservation protocols, scheduling policies, or feedback algorithms to guarantee the convergence of approximated solutions.

Acknowledgments

Special thanks go to Professor Zhe George Zhang at Simon Fraser University, Canada for helpful discussions, which led to the significant improvement of this paper in numerous ways. The first author is grateful for the financial support received from Aiming for the Top University Program of the National Chiao Tung University and Ministry of Education, Taiwan. The second author is thankful for a partial support from National Science Council, Taiwan under grant number NSC 100-2221-E-004-003.

Appendix. Proofs of propositions

Proof of Lemma 2. Let $Y(z) = A^*(s\mu - s\mu z)$. Since $Y(0) = \beta_0 > 0$, $Y(1) = 1$, and $Y'(1) = 1/\rho_s > 1$, there exists at least one real root between 0 and 1 for the characteristic equation (7). For real z, $0 < z \leq 1$, by changing variables via $z = e^s$, $-\infty < s \leq 0$, (7) becomes

$$e^s = Y(e^s), \quad -\infty < s \leq 0, \tag{37}$$

and so

$$s = \ln Y(e^s), \quad -\infty < s \leq 0. \tag{38}$$

Note that, Hölder's inequality gives

$$Y(e^{(ps_1+(1-p)s_2)}) \leq \left(Y(e^{s_1})\right)^p \left(Y(e^{s_2})\right)^{1-p} \tag{39}$$

for all $s_1, s_2 \leq 0$ and $0 \leq p \leq 1$. Thus

$$\ln Y(e^{ps_1+(1-p)s_2}) \leq p \ln Y(e^{s_1}) + (1-p) \ln Y(e^{s_2}) \tag{40}$$

for all $s_1, s_2 \leq 0$ and $0 \leq p \leq 1$. Therefore, the right hand side of (38) is convex. Then (38) has exactly one negative root and the root is simple. Thus, (7) has exactly one real root between 0 and 1, and the root is also simple. Denote by Φ_s the real root of (7) between 0 and 1. By Rouché's theorem, the number of zeros of $z - A^*(s\mu - s\mu z)$ and z, counted by multiplicities, on $\{z \in \mathbb{C} : |z| < \eta\}$ are the same for any η, $\Phi_s < \eta < 1$. It is easy to see that $z - A^*(s\mu - s\mu z)$ has no other zeros on $\{z \in \mathbb{C} : |z| = \Phi_s\}$ except the simple zero Φ_s. Thus, we conclude that the characteristic equation (7) has exactly one root Φ_s on $\{z \in \mathbb{C} : |z| < 1\}$. □

Proof of Proposition 1. From (5), we can determine $\beta_0(s)$ and $\beta_1(s)$ for different examples of inter-arrival time distributions. By Corollaries 1–3, there exist those limit $\lim_{s\to\infty} \beta_n(s)$ for $n = 0, 1$. Because the limits of $\beta_0(s)$ and $\beta_1(s)$ exist as $s \to \infty$, and $1 - \beta_0(s) - \beta_1(s) > 0$ for all s, the limit of

$$\frac{1-\beta_1(s)}{1-\beta_0(s)-\beta_1(s)}$$

exists as $s \to \infty$ and equals to a constant number. □

Proof of Proposition 2. It is obvious that the following inequalities hold

$$0 \leq \frac{\beta_0(s)}{1-\beta_1(s)} = \frac{\beta_0(s)}{\beta_0(s) + \sum_{n=2}^{\infty} \beta_n(s)} \leq 1.$$

Then, we have

$$0 \leq \left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s < 1.$$

Hence, the sequence

$$\left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s$$

is bounded. Moreover, it can be derived that

$$1 \geq \left(\frac{\beta_0(2)}{1-\beta_1(2)}\right)^1 \geq \left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s$$

for all $s \gg 1$, and we have

$$\left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s$$

decreases when increasing $s \gg 1$. Therefore, it have been shown that $[\beta_0(s)/(1-\beta_1(s))]^s$ is bounded and monotone as $s \gg 1$. So, the limit of

$$\left(\frac{\beta_0(s)}{1-\beta_1(s)}\right)^s$$

exists as $s \gg 1$ and equals to a constant number. □

Proof of Corollary 1. From (19), we derive

$$\lim_{s\to\infty} \beta_n(s) = \lim_{s\to\infty} \frac{\rho_s}{(1+\rho_s)^{n+1}} = \frac{1}{2^{n+1}}$$

with the help of assumption that traffic intensities ρ_s approach 1 from below as $s \to \infty$. □

Proof of Corollary 2. From (23), we derive

$$\lim_{s\to\infty} \beta_n(s) = \lim_{s\to\infty} \frac{e^{-1/\rho_s}}{n!\rho^n(s)} = \frac{e^{-1}}{n!}$$

with the help of assumption that traffic intensities ρ_s approach 1 from below as $s \to \infty$. □

Proof of Corollary 3. From (27), we have

$$\begin{aligned}\lim_{s\to\infty} \beta_n(s) &= \lim_{s\to\infty} \frac{\rho_s^r}{n!(r-1)!(1+\rho_s)^{n+r}}\Gamma(n+r) \\ &= \frac{\Gamma(n+r)}{n!(r-1)!2^{n+r}}\end{aligned}$$

with the help of assumption that traffic intensities ρ_s approach 1 from below as $s \to \infty$. □

References

[1] B.D. Choi, B. Kim, J. Kim, I.S. Wee, Exact convergence rate for the distributions of $GI/M/c/K$ queue as K tends to infinity, Queueing Systems 44 (2003) 125–136.
[2] A. Faragó, Efficient blocking probability computation of complex traffic flows for network dimensioning, Computers and Operations Research 35 (2008) 3834–3847.
[3] C. Maglaras, A. Zeevi, Diffusion approximations for a multiclass Markovian service system with "guaranteed" and "best-effort" service levels, Mathematics of Operations Research 29 (2004) 786–813.
[4] C.H. Wang, H.P. Luh, Analysis of bandwidth allocation on end-to-end QoS networks under budget control, Computers and Mathematics with Applications 62 (2011) 419–439.
[5] S.H. Hung, C.S. Shih, J.P. Shieh, C.P. Lee, Y.H. Huang, Executing mobile applications on the cloud: framework and issues, Computers and Mathematics with Applications 63 (2012) 573–587.
[6] B. Al-Manthari, N.A. Ali, N. Nasser, H. Hassanein, Dynamic multiple-frame bandwidth provisioning with fairness and revenue considerations for broadband wireless access systems, Performance Evaluation 68 (2011) 768–781.
[7] C. Maglaras, A. Zeevi, Pricing and capacity sizing for systems with shared resources: approximate solutions and scaling relations, Management Science 49 (2003) 1018–1038.
[8] C. Bruni, F.D. Priscoli, G. Koch, I. Marchetti, Resource management in network dynamics: an optimal approach to the admission control problem, Computers and Mathematics with Applications 59 (2010) 305–318.
[9] J.G. Dai, T. Tezcan, State space collapse in many-server diffusion limits of parallel server systems, Mathematics of Operations Research 36 (2011) 271–320.
[10] W. Whitt, A diffusion approximation for the $G/GI/n/m$ queue, Operations Research 52 (2004) 922–941.
[11] J.B. Atkinson, Two new heuristics for the $GI/G/n/0$ queueing loss system with examples based on the two-phase Coxian distribution, Journal of the Operational Research Society 60 (2009) 818–830.
[12] B. Kim, B.D. Choi, Asymptotic analysis and simple approximation of the loss probability of the $GI^X/M/c/K$ queue, Performance Evaluation 54 (2003) 331–356.

[13] V.M. Abramov, Asymptotic analysis of loss probabilities in $GI/M/m/n$ queueing systems as n increases to infinity, Quality Technology and Quantitative Management 4 (2007) 379–393.
[14] S. Halfin, W. Whitt, Heavy-traffic limits for queues with many exponential servers, Operations Research 29 (1981) 567–588.
[15] W. Whitt, Stochastic-Process Limits: An Introduction to Stochastic-Process Limits and their Application to Queues, Springer-Verlag New York, Inc., 2002.
[16] F. Simonot, A comparison of the stationary distributions of $GI/M/c/n$ and $GI/M/c$, Journal of Applied Probability 35 (1998) 510–515.