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.