Computation of Blocking Probability in a Loss System [1]

SEONGRYONG NA [2]

Abstract

A loss system where two types of customers arrive in accordance with two independent Poisson processes is considered. An efficient recursive formula is developed for calculating the loss probability when the number of servers is large. Some practical examples regarding the performance evaluation of telecommunications networks are discussed.

Keywords : blocking probability; loss system; recursive formula

1. Introduction

Practical examples concerning loss systems arise in many areas, such as computer and telecommunications networks, resource managements, machine repairing problems, and so on. Among others, Kelly (1991) considered a loss network where customers of distinct classes occupy different number of servers during the service period. Kelly's loss systems can be useful analytical models in the performance analysis of diverse telecommunication networks supporting multi-class services, such as voice, low-speed data and high-speed data services, since actual systems are equipped with limited resources and distinct service may request different amount of resources. Koo, Yang and Kim (2000) analyzed the capacity of some multi-media system by calculating loss probabilities for the system considered by Kelly. Other studies of telecommunications networks formulated as such loss systems can be found in Lee, Na, Choi and Lim (2002) and Koo, Yang, Ahmad and Kim (2002).

The steady-state distribution for this multi-class loss system is well established under the usual Markovian assumption of Poisson arrivals and exponential service times. See, for example, Kelly (1991) and Ross (1995). However, actual computation of blocking probability needs additional computational algorithm especially when the number of total servers in the system is large. Note that Erlang's loss formula for $M/M/c/c$ system and its recursive version serve as a good example (cf. Medhi, 2003). In this paper we focus on the blocking probability that an arriving customer cannot receive the service and the corresponding efficient computation algorithm.

In section 2, the loss system is specified and an efficient computation algorithm for the blocking probability is presented. Section 3 gives practical applications of the algorithm. Some quantities to approximate the blocking probability are also provided. The numerical method developed in section 2 is shown to be a powerful tool for the performance evaluation and the resource dimensioning of network systems.

2. Blocking probability and recursive formula

Consider a queueing system with kc servers where distinct customers of two types arrive. The customers of Type I and Type II arrive at the system in accordance with two independent Poisson processes with rate λ_1 and λ_2, respectively. Each customer of Type I occupies only one server during the service time that is exponentially distributed with mean μ_1^{-1}. On the contrary, the customer of Type II uses k servers simultaneously during the period exponentially distributed with mean μ_2^{-1}. All service times are assumed to be mutually independent. Let $\rho_i = \lambda_i/\mu_i, i = 1, 2$ denote the traffic intensity or offered load to the system of Type i. The system considered is a loss system in that any arriving customer finding fewer idle servers in the system than it requires is blocked and leaves the system without being served. Thus, a customer of Type I is lost to the system if on arrival it finds

[1]Supported by University ITRC project and in part by Yonsei University research fund of 2003

[2]Department of Information and Statistics, Yonsei University, Wonju, 220-710, Korea (nasr@dragon.yonsei.ac.kr)

all the servers busy, whereas an arriving customer of Type II is blocked without service when there are idle servers fewer than k.

We now derive the steady-state distribution for the numbers of customers being served by applying a simple truncation method to a two-dimensional time-reversible Markov chain. Let $N_i(t), i = 1,2$ be the number of the customers of Type i being served in the system at time t. Then, $(N_1(t), N_2(t))$ is a two-dimensional continuous-time Markov chain with state space

$$S = \{(n_1, n_2) \in \mathbf{Z}_0^2; n_1 + kn_2 \leq kc\},$$

where $\mathbf{Z}_0$ is the set of nonnegative integers. It is easy to see that this chain is the truncated version of a two-dimensional Markov chain $(\bar{N}_1(t), \bar{N}_2(t))$, where $\bar{N}_1(t)$ and $\bar{N}_2(t)$ denote the numbers of busy servers of two independent M/M/kc/kc and M/M/c/c loss systems with traffic intensities ρ_1 and ρ_2 , respectively. Note that since $\bar{N}_1(t)$ and $\bar{N}_2(t)$ are birth-and-death processes and time-reversible in steady-state, the two-dimensional $(\bar{N}_1(t), \bar{N}_2(t))$ is also time-reversible. Then by Kelly (1979), the truncated $(N_1(t), N_2(t))$ is time-reversible and its steady-state distribution is just the stationary distribution of $(\bar{N}_1(t), \bar{N}_2(t))$ truncated to S.

Let (N_1, N_2) and $(\bar{N}_1, \bar{N}_2)$ denote $(N_1(t), N_2(t))$ and $(\bar{N}_1(t), \bar{N}_2(t))$ in steady state, respectively. Then, we have

$$\begin{aligned} p(n_1, n_2) &= P(N_1 = n_1, N_2 = n_2) \\ &= P(\bar{N}_1 = n_1, \bar{N}_2 = n_2 | (\bar{N}_1, \bar{N}_2) \in S), \qquad (n_1, n_2) \in S. \end{aligned} \tag{1}$$

We define

$$p(n; r, \rho) = \frac{\rho^n}{n!} [\sum_{j=0}^{r} \frac{\rho^j}{j!}]^{-1}, \qquad n = 0, 1, \ldots, r. \tag{2}$$

Note that $\{p(n; r, \rho), n = 0, 1, \ldots, r\}$ is the stationary probability for the number of busy servers of M/M/r/r loss system with traffic intensity ρ. From (1) and (2),

$$p(n_1, n_2) = p(n_1; kc, \rho_1) p(n_2; c, \rho_2) / p_S, \tag{3}$$

where

$$p_S = P((\bar{N}_1, \bar{N}_2) \in S) = \sum_{(n_1, n_2) \in S} p(n_1; kc, \rho_1) p(n_2; c, \rho_2)$$

is the normalizing constant.

Since the arrival streams are Poisson processes, it is an easy task to find explicit forms of the blocking probabilities. First, define two sets of states in S

$$B_1 = \{(n_1, n_2); n_1 + kn_2 = kc\} \subset S$$

and

$$B_2 = \{(n_1, n_2); (k-1)c + 1 \leq n_1 + kn_2 \leq kc\} \subset S.$$

Then, the blocking probability that an arriving customer of Type I cannot receive the service is

$$L_1 = \sum_{(n_1, n_2) \in B_1} p(n_1, n_2) \tag{4}$$

and the blocking probability corresponding to Type II is given by

$$L_2 = \sum_{(n_1, n_2) \in B_2} p(n_1, n_2), \tag{5}$$

which come from the PASTA property(cf. Wolff, 1989).

Even though the explicit equations (3), (4) and (5) may be thought to be simple, the actual computation of (4) and (5) is not so easy when the number of servers, kc, becomes large. This difficulty arises mainly from the calculation of (2) for large n and r. Although a useful recursive algorithm to compute blocking probability is presented in Ross (1995), it seems that the algorithm there would confront the same numerical problem as the direct calculation of Erlang's loss formula if the number of servers is very large. Hence, we should devise another reliable method for the computation.

If $n < r$, then from (2)

$$\begin{aligned} p(n;r,\rho)^{-1} &= (\sum_{j=0}^{r-1}\frac{\rho^j}{j!}+\frac{\rho^r}{r!})(\frac{\rho^n}{n!})^{-1} \\ &= p(n;r-1,\rho)^{-1}+p(r;r,\rho)p(n;r,\rho)^{-1}. \end{aligned}$$

Thus, we have

$$p(n;r,\rho)=p(n;r-1,\rho)[1-p(r;r,\rho)]. \tag{6}$$

For $n = r$, we use the well-known recursive formula

$$p(r;r,\rho)=\rho p(r-1;r-1,\rho)[r+\rho p(r-1;r-1,\rho)]^{-1} \tag{7}$$

with the convention $p(0;0,\rho)=1$. Simple iteration based on (6) and (7) gives exact numerical values of $p(0;0,\rho)$; $p(1;1,\rho)$, $p(0;1,\rho)$; $p(2;2,\rho)$, $p(1;2,\rho)$, $p(0;2,\rho)$; ... ; $p(r;r,\rho)$, ... , $p(0;r,\rho)$ for any large r. Using these values for $r=kc, \rho=\rho_1$ and for $r=c, \rho=\rho_2$, one can compute (4) and (5) by

$$L_1=\sum_{n_2=0}^{c}p(k(c-n_2),n_2)=p_S^{-1}\sum_{n_2=0}^{c}p(k(c-n_2);kc,\rho_1)p(n_2;c,\rho_2) \tag{8}$$

and

$$\begin{aligned} L_2 &= \sum_{n_2=0}^{c-1}\sum_{n_1=k(c-1-n_2)+1}^{k(c-n_2)} p(n_1,n_2) + p(0,c) \\ &= p_S^{-1}\{\sum_{n_2=0}^{c-1}\sum_{n_1=k(c-1-n_2)+1}^{k(c-n_2)} p(n_1;kc,\rho_1)p(n_2;c,\rho_2) + p(0;kc,\rho_1)p(c;c,\rho_2)\}, \end{aligned} \tag{9}$$

where

$$p_S=\sum_{n_2=0}^{c}\sum_{n_1=0}^{k(c-n_2)}p(n_1;kc,\rho_1)p(n_2;c,\rho_2).$$

Note that the numerical computation of (8) and (9) spends memory of $O(kc)$ and time of $O(k^2c^2)$, where for example, the symbol $O(kc)$ means that the quantity is bounded by a constant multiplication of kc as kc increases to infinity.

3. Application and the numerical study

We now introduce approximating quantities for the blocking probabilities (4) and (5). First, the system considered here can be approximately regarded as M/M/kc/kc (or M/M/c/c) loss system from the viewpoint of the customer of Type I (Type II). Thus, the values in (4) and (5) are approximated respectively by

$$L_{1,\text{app}}=p(kc;kc,\rho_1+k\rho_2) \tag{10}$$

and

$$L_{2,\text{app}}=p(c;c,\rho_2+\rho_1/k), \tag{11}$$

where some modifications of the offered loads are made. Note that this approximation is not so elaborate but can serve as a simple substitute. Furthermore, by comparing the values from (10) and (11) with ones from (8) and (9), we may draw some meaningful conclusions on the performance characteristics of the loss system with multiple types of customers relatively to the system with single type.

The computation procedure discussed so far makes possible the accurate and rapid analysis of large capacity loss systems with multiple types of inputs. Practical examples can arise in telecommunications networks, where the performance evaluation and the resource dimensioning are of main concern.

1. Performance evaluation The evaluation of performance measures, such as blocking probability and mean number of busy servers, is the most elementary step in performance analysis. Above all, the evaluation of blocking probability is of primary concern in loss systems. In our case, two types of blocking probabilities are determined by (8) and (9), or approximately by (10) and (11) for any given system parameters. For example, if system parameters are given by $k = 2$, $c = 200$, $\rho_1 = 266$ and $\rho_2 = 57$, then we obtain $L_1 = 0.0142538$, $L_2 = 0.0290861$, $L_{1,\text{app}} = 0.0139316$ and $L_{2,\text{app}} = 0.0279682$. If $k = 5$, $c = 200$, $\rho_1 = 485$ and $\rho_2 = 97$, then $L_1 = 0.0086591$, $L_2 = 0.0442019$, $L_{1,\text{app}} = 0.00952919$ and $L_{2,\text{app}} = 0.0376642$. Besides, the mean number of busy servers, that is $E(N_1 + kN_2)$, is easily calculated using (3) and the recursive formulas (6) and (7), which numerically determines the realistic utilization factor or server utilization of the loss system.

2. Number of servers for QoS requirement Another important application can be found in the design of telecommunications system, where the necessary capacity of the system should be decided to satisfy the QoS(Quality of Service) criterion. Consider a telecommunications system supporting two types of real-time services such as voice and data. This system is assumed to be modelled as the loss system discussed in this paper. Suppose that some system parameters are prescribed beforehand. For example, the resources allocated to a data service $k = 3$, the offered server utilization $u = (\rho_1 + k\rho_2)/(kc) = 95\%$, the proportion of voice traffic $\beta = \rho_1/(\rho_1 + k\rho_2) = 70\%$. The QoS criterion is given as $p_{1,\text{req}} = 1\%$ and $p_{2,\text{req}} = 2\%$, which means that the system should have plenty of resources (or servers) so that $L_1 \leq p_{1,\text{req}}$ and $L_2 \leq p_{2,\text{req}}$. The problem is to find the number c satisfying the requirements under the prescribed conditions. This can be done by tabulating the blocking probabilities for varying values of c. The blocking probabilities under the above conditions computed by (8), (9), (10) and (11) are contained in Table 1. It is easily seen that c should be lie between 270 and 280 for the given level of QoS. In fact, the exact value is $c = 273$ and therefore, the system should be equipped with at least 819 servers to guarantee the desired QoS with the prescribed server utilization.

Table 1. Blocking probabilities under $k = 3, u = 95\%, \beta = 70\%$

c	L_1	L_2	$L_{1,\text{app}}$	$L_{2,\text{app}}$
⋮				
150	0.0126531	0.0392119	0.0121237	0.0357016
160	0.0118548	0.0367373	0.0111983	0.0338596
170	0.0111345	0.0345044	0.0103702	0.0321867
180	0.0104811	0.0324790	0.0096252	0.0306594
190	**0.0098856**	0.0306331	0.0089521	0.0292585
⋮				
270	0.0065259	0.0202206	0.0052903	0.0211414
280	0.0062266	**0.0192930**	0.0049787	0.0203947
290	0.0059460	0.0184236	0.0046895	0.0196900
⋮				

3. Calculation of admissible utilization factor We now consider another inverse problem in a slightly different direction. We assume that the parameters $k = 5$, $c = 300$, $\beta = 60\%$ are given beforehand. The problem is to search for the admissible utilization factor u or equivalently, the overall offered load $\rho_1 + k\rho_2$ under the QoS requirement $p_{1,\text{req}}$ and $p_{2,\text{req}}$. Say, for example, $p_{1,\text{req}} = 1\%$ and $p_{2,\text{req}} = 3\%$. Blocking probabilities are tabulated in Table 2 as the utilization factor varies from 99% to 96%. From Table 2, we conclude that the utilization factor must be controlled to be less than approximately 96.5% in our setting. Thus, the system should not admit the overall offered load more than 1447.5.

Table 2. Blocking probabilities under $k = 5, c = 300, \beta = 60\%$

u	L_1	L_2	$L_{1,\text{app}}$	$L_{2,\text{app}}$
⋮				
0.99	0.0102690	0.0517575	0.0145247	0.0388326
0.985	**0.0091898**	0.0464983	0.0119619	0.0360223
0.98	0.0081672	0.0414859	0.0096526	0.0332961
0.975	0.0072045	0.0367395	0.0076141	0.0306598
0.97	0.0063046	0.0322773	0.0058566	0.0281195
0.965	0.0054700	**0.0281153**	0.0043815	0.0256805
0.96	0.0047025	0.0242667	0.0031803	0.0233484
⋮				

References

KELLY, F.P. (1979). *Reversibility and stochastic networks.* Wiley, Chichester.

KELLY, F.P. (1991). Loss networks. *Ann. Appl. Probab.* **1**, 319-378.

KOO, I., YANG, J., AHMAD, A. and KIM, K. (2002). Erlang capacity analysis of hybrid FDMA/CDMA systems supporting multi-class services according to channel assignment methods. *Int. J. Commun. Syst.* **15**, 867-880.

KOO, I., YANG, J. and KIM, K. (2000). Analysis of Erlang capacity for DS-CDMA systems supporting multi-class services with the limited number of channel elements. *IEEE Proc. of WCN, 2000, USA*, 355-359.

LEE, J., NA, S., CHOI, Y. and LIM, S. (2002). Performance analysis on channel distribution algorithm of selector in CDMA2000 1x system. *KICS Proc., summer, 2002.*

MEDHI, J. (2003). *Stochastic models in queueing theory, 2nd.* Academic press, San Diego.

ROSS, K.W. (1995). *Multiservice loss models for broadband telecommunication networks.* Springer, London.

WOLFF, R.W. (1989). *Stochastic modeling and the theory of queues.* Prentice hall, Englewood Cliffs.