Iranian Journal of Science & Technology, Transaction A, Vol. 31, No. A2
Printed in The Islamic Republic of Iran, 2007

MINIMIZING LOSS PROBABILITY IN QUEUING SYSTEMS WITH HETEROGENEOUS SERVERS[*]

V. SAGLAM[**] AND A. SHAHBAZOV

Department of Statistic, Faculty of Arts and Sciences, Ondokuz
Mayis University, Kurupelit, 55139-Samsun, Turkey,
Email: vsaglam@omu.edu.tr

Abstract – The probability of losing a customer in *M/G/n/*0 and *GI/M/n/*0 loss queuing systems with heterogeneous servers is minimized. The first system uses a queue discipline in which a customer who arrives when there are free servers chooses any one of them with equal probability, but is lost otherwise. Provided that the sum of the servers rates are fixed, loss probability in this system attains minimum value when all the service rates are equal. The second system uses queue discipline, in which a customer who enters into the system is assigned to the server with the lowest number. Loss probability in this system takes the minimum value in the case when the *fastest server rule* is used in which an incoming customer is served by the free server with the shortest mean service time. If the mean of the arrival distribution is fixed, then loss probability is minimized by deterministic arrival distribution.

Keywords – Service rate, Erlang's loss formula, heterogeneous servers, loss probability, recurrent input, exponential server, overflow distribution

1. INTRODUCTION

In analyzing many queuing models, it is usually assumed that all the servers (channels) in the queuing system are identical (homogenous) in the sense that they have the same service time (s.t.) distribution. However, the servers of many real-life systems are different (heterogeneous). Such a situation appears when servers of the same mark were made at different factories, or before exploitation in the system they were used in at different systems, and therefore have a non-identical degree of wearing out. The Queuing models with heterogeneous servers also arise in a number of important applications such as computer systems, communications systems and production lines. Fundamental loss queuing systems, $M/G/n/0$ and $GI/M/n/0$, with identical servers, have been studied almost completely. A very important measure of effectiveness for these systems is the loss probability meant for the stationary probability of losing a customer or the probability that all servers of the system are busy. Stationary probability in which k servers of the system $M/G/n/0$ (Erlang's loss model) are busy is given by well-known Erlang's formula

$$p_k = \frac{\rho^k}{k!} \bigg/ \sum_{k=0}^{n} \frac{\rho^k}{k!} \quad (0 \le k \le n), \tag{1}$$

where $\rho = \lambda/\mu$ is the offered load, $1/\lambda$ is the mean interarrival time, and $1/\mu$ is mean ST. This formula, first derived in [1] for the case when all servers have the same exponential s.t., plays an important role in analyzing communication systems and its properties have been studied extensively and exhaustively. One

[*] Received by the editor March 27, 2005 and in final revised form April 8, 2007
[**] Corresponding author

of the surprising properties of formula (1) is that the limiting distribution of the number of busy servers is invariant to the s.t. distribution *G*, i.e., limiting probabilities $p_0, \cdots, p_n$ are independent of the form G depending on G only through its mean. This remarkable result and many connected questions have been studied by several authors. For example, the validity of (1) for absolute continuous s.t. distribution has been proved in [2]. An exact mathematical proof of this formula for arbitrary s.t. distribution with a finite mean has been given in [3]. Erlang's formula for system $GI/M/n/0$ with recurrent input and with the identical exponential servers has been obtained in [4]. In [5] Erlang's formula has been extended to the case of dependent service times. In [6-8], Erlang's loss model has been studied using discrete-time process at arrival and departure epochs.

From formula (1) we can find the loss probability p_n in the system $M/G/n/0$

$$\frac{1}{p_n} = n! \sum_{k=0}^{n} \frac{(1/\rho)^k}{(n-k)!} \ . \tag{2}$$

This formula is called Erlang's loss formula and is expressed in terms of the mean s.t. and the mean interarrival time.

In this paper we consider the problem of minimizing loss probability in $M/G/n/0$ and $GI/M/n/0$ queuing models with different servers. We denote the loss system with recurrent input by $GI/\vec{G}/n/0$, and with s.t. distribution G_k at *k*th server, where $\vec{G} = (G_1, \cdots, G_n)$ symbolizes the heterogeneity of the servers. In the case when G_k = G for $k = 1, \cdots, n$, we have an $M/G/n/0$ system with identical servers which we shall call homogeneous. Non-homogeneous queuing systems have been studied mainly for exponential servers. In [9, 10] a limitied distribution of the number of customers in the system $M/\vec{M}/n$ have been found with heterogeneous exponential servers and an unbounded waiting room. In [11] the problem of minimization of the loss probability in the $M/\vec{M}/n/0$ system has been solved, provided the sum of the service rates (total service rate) is fixed and the arriving customer is assigned to the free server with the shortest mean s.t.

Most of the non-homogeneous queues have been analyzed for two-server cases. In [12], explicit expressions for the steady-state probabilities for an $M/G/2/0$ queue with two classes of Poisson arrivals has been derived. The result of these authors is for the case of an arbitrary number of arrival classes in [13]. The queuing models with two non-identical exponential servers have been analyzed in [14-17], where a new queue discipline has been introduced in which a customer who arrives when both servers are free, chooses his server with some probability. The systems $M/\vec{G}/2/0$ and $M/\vec{G}/2$ with this discipline have been investigated in [18, 19]. The system $GI/M/2$ with recurrent input and a service rate depending on the number of busy servers was studied in [20].

2. LOSS PROBABILITY IN THE MODEL $M/\vec{G}/n/0$

Consider the loss queuing system $M/\vec{G}/n/0$ consisting of n heterogeneous servers labeled by numbers $1, 2, \cdots, n$. The arrival process is Poisson with rate λ, and the s.t. of any customer at the *k*th server has a distribution function G_k with finite mean $1/\mu_k$ for $k = 1, \cdots, n$. For this system the following discipline is used: An arriving customer chooses any one of the free servers with equal probability and is lost if all *n* servers are busy. Let $X_k(t) = 1$ if *k*th server is busy at time *t*, and $X_k(t) = 0$ otherwise. Then limiting probability that k servers with numbers $i_1, \cdots, i_k$ are busy can be written as

$$p_k(i_1, \cdots, i_k) = \lim_{t \to \infty} P\{X_{i_1}(t) = 1, \cdots, X_{i_k}(t) = 1, X_{i_k+1}(t) = 0, \cdots, X_{i_n}(t) = 0\}, \tag{3}$$

where $(i_1, \cdots, i_n)$ is a permutation of $(1, \cdots, n)$. Note that $X(t) = X_1(t) + \cdots + X_n(t)$ is the number of busy servers at time t . Let p_k denote the limiting probability that k servers are busy, i.e.,

$$p_k = \lim_{t\to\infty} P\{X(t)=k\}, \quad 0 \le k \le n .$$

In particular, p_n is the loss probability in the $M/\vec{G}/n/0$ queue with different servers.

In this section we consider the problem of minimizing the loss probability p_n subject to $\mu_1 + \cdots + \mu_n = n\mu$, where μ is constant. The solution of this problem is based on the explicit expression for p_n .

In [21] it has been shown that the limiting probabilities (3) exist and are given by

$$p_k(i_1,\cdots,i_k) = \frac{(n-k)!}{n!}\rho_{i_1}\cdots\rho_{i_k} p_0 \quad (0 \le k \le n), \tag{4}$$

where $\rho_k = \lambda / \mu_k$. Then limiting probabilities $p_0,\cdots,p_n$ are given by

$$p_k = (n-k)!E_k / \sum_{k=0}^{n}(n-k)!E_k \qquad (0 \le k \le n), \tag{5}$$

where $E_k = E_k(\rho_1,\cdots,\rho_n)$ is the kth elementary symmetric function of the $\rho_1,\cdots,\rho_n$, which is defined as

$$E_0 = 1, \quad E_k = \sum_{1\le i_1<\ldots<i_k\le n} \rho_{i_1}\cdots\rho_{i_k} \ , \ 0 \le k \le n ,$$

where the summation extends (over the $C(n,k)$) all combinations of k distinct elements $\{i_1,\cdots,i_k\}$ from $\{1,\cdots,n\}$. Letting $S_k = E_k / C(n,k)$ in (5), we obtain p_k in the form

$$p_k = \frac{S_k}{k!} \Bigg/ \sum_{k=0}^{n}\frac{S_k}{k!} \ , \quad 0 \le k \le n. \tag{6}$$

It is a generalization of Erlang's formula (1) to the heterogeneous servers' case. In particular, from (6) we conclude that the limiting distribution of the number of busy servers which are independent of the form $G_1,\cdots,G_n$, depends only their mean values $1/\mu_1,\cdots,1/\mu_n$. In particular, if the service times have the same mean $1/\mu$, then $S_k = \rho^k$ and probabilities $p_0,\cdots,p_n$ are calculated by Erlang's formula (1). From formula (6) we can find the loss probability in the queue $M/\vec{G}/n/0$ with different servers

$$\frac{1}{p_n} = \frac{n!}{\rho_1\cdots\rho_n}\sum_{k=0}^{n}\frac{S_k}{k!}. \tag{7}$$

It is a generalization of Erlang's loss formula (2) to the different servers case.

Our main result about the problem of minimizing the loss probability can be expressed by the following theorem.

Theorem 1. If sum service rates $\mu_1 + \cdots + \mu_n = n\mu$ is fixed, then loss probability p_n in the system $M/\vec{G}/n/0$ attains its minimum value for $\mu_1 = \cdots = \mu_n = \mu$.

Proof: Rewrite (7) in a way that is more convenient for analysis. Using the relation

$$S_k(a_1,\cdots,a_n) = a_1\cdots.a_n S_{n-k}(1/a_1,\cdots,1/a_n)$$

for positive number $a_1,\cdots,a_n$ and setting $\overline{S}_k = S_k(1/\rho_1,\cdots,1/\rho_n)$, we can write (7) in form

$$\frac{1}{p_n} = n! \Big/ \sum_{k=0}^{n} \frac{\bar{S}_k}{(n-k)!}\,. \tag{8}$$

Taking account of the inequality [22]

$$S_n^{1/n} \le S_{n-1}^{1/(n-1)} \le \cdots \le S_2^{1/2} \le S_1,$$

we have

$$\left(\bar{S}_k\right)^{1/k} \le \bar{S}_1 = \frac{1}{n}\left(\mu_1/\lambda + \cdots + \mu_n/\lambda\right) = \frac{\mu}{\lambda}.$$

From this and formula (8) we obtain

$$p_n \ge \left[n! \sum_{k=0}^{n} \frac{(\mu/\lambda)^k}{(n-k)!} \right]^{-1}. \tag{9}$$

For the case when $\mu_1 = \cdots = \mu_n = \mu$ the loss probability p_n takes the value that is equal to the expression in the right side of (9), so that p_n takes minimum value when $\mu_1 = \cdots = \mu_n = \mu.$ Note that the expression on the right side of (9) is the loss probability with (2) in the $M/G/n/0$ system with homogeneous servers. We conclude that an homogeneous system is better than the corresponding heterogeneous system, provided that the total service rate is fixed.

3. LOSS PROBABILITY IN THE MODEL $GI/\vec{M}/n/0$ WITH ORDERED ENTRY

We consider the loss queuing system consisting of n heterogeneous exponential servers labeled by $1, \cdots, n$ and arranged in series in that order. The following queue discipline is used: Each arriving customer is served by the lowest numbered server that is free. The customer initially arrives at the 1th server. If this server is free, he is served and departs. If this server is busy, he overflows and arrives at the 2nd server and so forth. The output stream from the *k*th server is the input stream to the (*k*+1)th server, for $k = 1, \cdots, n-1$. Finally, the customer who finds all servers busy is lost from the system. The overflow process from the nth server is the same as that from the loss system $GI/M/n/0.$ Interarrival times to the system are independent random variables and have distribution function F with mean $1/\lambda$. The s.t. of any customer at the kth server is exponential with μ_k parameter for $k = 1, \cdots, n$. Suppose that F and $\mu_1, \cdots, \mu_n$ are fixed. Then loss probability p_n has a varying value depending on the order of the servers. How does the order of the servers minimize p_n? The solution of this problem is based on the explicit expression for p_n. Such an explicit formula in the case that servers are identical is given in [4, 23]. In [24, 25], the generating function of the first passage time from any state to a full state in the loss system with identical exponential channels has been found and have shown that this time is independent of the placement of call policy used. Laplace-Stieltjes (LS) transform of the interoverflow times from a $GI/M/1/K$ queuing system was derived in [26].

It was shown in [4] that loss probability in the homogeneous system $GI/M/n/0$ is given by

$$\frac{1}{p_n} = \sum_{k=0}^{n} \binom{n}{k} c_k\,, \tag{10}$$

where f is the LS transform of the interarrival time and

$$c_0 = 1, \quad c_k = \frac{1 - f(\mu)}{f(\mu)} \cdots \frac{1 - f(k\mu)}{f(k\mu)}, \ \ 1 \le k \le n \cdot$$

$f_k(s)$ denotes the LS transform of the interoverflow times distribution from the first k servers for $k = 1, \cdots, n$. Then $f_k(s)$ satisfy Palm's difference equations

$$f_k(s) = \frac{f_{k-1}(s + \mu_k)}{1 - f_{k-1}(s) + f_{k-1}(s + \mu_k)} \qquad (1 \le k \le n), \tag{11}$$

where $f_0(s) = f(s)$ is the LS transform of *F*. This is the extension of Palm's equation to the heterogeneous servers case. From this equation we can find the mean of overflow time from the first k servers

$$a_k = -f_k'(0) = \lim_{s \to 0} \frac{1 - f_k(s)}{s} = a_{k-1} / f_{k-1}(\mu_k).$$

From this recurrent equation we obtain an important formula for the mean of the overflow times from queue $GI / \vec{M} / n / 0$.

$$1/a_n = \lambda f(\mu_1) f_1(\mu_2) \ldots f_{n-1}(\mu_n).$$

Using this simple relationship $1/a_n = \lambda p_n$, we obtain the loss probability in the heterogeneous system $GI / \vec{M} / n / 0$

$$p_n = f(\mu_1) f_1(\mu_2) \cdots f_{n-1}(\mu_n). \tag{12}$$

In particular, from (11) and (12), we can find the loss probability in queues $GI / \vec{M} / 1 / 0$ and $GI / \vec{M} / 2 / 0$

$$p_1 = f(\mu_1), \tag{13}$$

$$p_2 = \frac{f(\mu_1) f(\mu_1 + \mu_2)}{1 - f(\mu_2) + f(\mu_1 + \mu_2)}. \tag{14}$$

In the case that $\mu_1 = \mu_2 = \mu$ the last formula yields Palm's loss formula (10) with $n = 2$.

Theorem 2. Loss probability p_n in the queue $GI / \vec{M} / n / 0$ system achieves its minimum value when the servers are operated in the order $(i_1, \cdots, i_n)$, an arbitrary permutation of $(1, \cdots, n)$ for which $\mu_{i_1} \ge \mu_{i_2} \ge \cdots \ge \mu_{i_n}$.

We shall prove the theorem by an interchange argument. Let $p_n(i)$ denote the value of the p_n when the servers are operated in order $i = (i_1, \cdots, i_n)$. Without loss of generality, assume that the optimal order of the servers is given by a sequence $\alpha = (1, \cdots, n)$. Interchanging $k - 1$ and k in this sequence gives a new sequence $\beta = (1, \cdots, k-2, k, k-1, k+1, \cdots, n)$ which will not be optimal, i.e.,

$$p_n(\alpha) \le p_n(\beta). \tag{15}$$

Using (12) we have

$$p_n(\beta) = f(\mu_1) \cdots f_{k-3}(\mu_{k-2}) f_{k-2}(\mu_k) f_{k-1}^*(\mu_{k-1}) f_{k+1}(\mu_{k+2}) \cdots f_{n-1}(\mu_n), \tag{16}$$

where $f_{k-1}^*(.)$ is the LS transform of the overflow distribution from the first $k - 1$ servers with service rates $\mu_1, \ldots, \mu_{k-2}, \mu_k$ respectively. This transform is obtained from Palm's equation (11):

$$f_{k-1}^{*}(s) = \frac{f_{k-2}(s+\mu_k)}{1 - f_{k-2}(s) + f_{k-2}(s+\mu_k)}. \tag{17}$$

Substituting (12) and (16) into (15) leads to the inequality

$$f_{k-2}(\mu_{k-1})\, f_{k-1}(\mu_k) \le f_{k-2}(\mu_k)\, f_{k-1}^{*}(\mu_{k-1}).$$

Using the expression of the $f_{k-1}(\mu_k)$ and $f_{k-1}^{*}(\mu_{k-1})$ from (11) and (17) respectively, and denoting $f_{k-2}(.)$ by $\varphi(.)$, we have

$$\frac{\varphi(\mu_{k-1})}{1-\varphi(\mu_k)+\varphi(\mu_{k-1}+\mu_k)} \le \frac{\varphi(\mu_k)}{1-\varphi(\mu_{k-1})+\varphi(\mu_{k-1}+\mu_k)}.$$

Subtracting both sides of this inequality from 1 gives inequality

$$\frac{b}{1-\varphi(\mu_k)+\varphi(\mu_{k-1}+\mu_k)} \ge \frac{b}{1-\varphi(\mu_{k-1})+\varphi(\mu_{k-1}+\mu_k)}, \tag{18}$$

where

$$b = 1-\varphi(\mu_{k-1})-\varphi(\mu_k)+\varphi(\mu_{k-1}+\mu_k)$$

$$= \int_0^{\infty}(1-e^{-\mu_{k-1}t})(1-e^{-\mu_k t})dF_{k-2}(t) > 0$$

and $F_{k-2}(t)$ denote overflow distribution from the first $k-2$ servers. Dividing both sides of inequality (18) by factor b and performing simple algebra we obtain

$$f_{k-2}(\mu_{k-1}) \le f_{k-2}(\mu_k),\ 2 \le k \le n. \tag{19}$$

Since $f_{k-2}(s)$ is non-increasing, in $s \ge 0$ we get $\mu_{k-1} \ge \mu_k$, $2 \le k \le n$, i.e., $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_n$ which concludes the proof.

According to this theorem the loss probability in the queue $G/\vec{M}/n/0$ attains its minimum value if servers are operated in the order of shortest mean service time. In other words, in order to minimize the loss probability we must use the fastest-server rule in which each arriving customer is served by the free server with the shortest mean of service time.

We now consider the above queuing system with a given mean interarrival time and given service rates $\mu_1, \cdots, \mu_n$. Which interarrival time distribution minimizes the loss probability in this system? This question has been solved in [27] for the case of identical servers. Let H_a denote the class of all interarrival time distributions, F having a fixed mean a, and $p_n(F)$ be loss probability in the above system with interarrival time distribution $F \in H_a$. Let

$$A(t) = \begin{cases} 0, & t \le a \\ 1, & t > a. \end{cases}$$

Clearly, $A \in H_a$ and e^{-as} is the LS transform of $A(t)$.

Theorem 3. Loss probability $p_2(F)$, $F \in H_a$ in the queue $GI/\vec{M}/2/0$ is minimized by F=A.

Proof: Using Jensen's inequality we have $f(s) \geq e^{-as}$ for any $s \geq 0$, where $a = -f'(0)$ is the mean of interarrival time. Formula (14) for loss probability $p_2(F)$ in the queue $GI/\vec{M}/2/0$ can be written as

$$p_2(F) = \frac{f(\mu_1)}{1 + \frac{1 - f(\mu_2)}{f(\mu_1 + \mu_2)}}. \tag{20}$$

From Jensen's inequality we can find inequality

$$\frac{1 - f(\mu_2)}{f(\mu_1 + \mu_2)} \leq \frac{1 - e^{-a\mu_2}}{e^{-a(\mu_1 + \mu_2)}}.$$

Using this inequality in (20) we have

$$p_2(F) \geq \frac{e^{-a\mu_1}}{1 + \frac{1 - e^{-a\mu_2}}{e^{-a(\mu_1 + \mu_2)}}} = \frac{e^{-a\mu_1} e^{-a(\mu_1 + \mu_2)}}{1 - e^{-a\mu_2} + e^{-a(\mu_1 + \mu_2)}}$$

Since the LS transform of $A(t)$ is e^{-as}, we see that the expression on the right side of the last inequality has the value of $p_2(F)$ for $F = A$. Consequently, loss probability $p_2(F)$ achieves its minimum value for $F = A$:

$$\min_{F \in H_a} p_2(F) = p_2(A).$$

REFERENCES

1. Erlang, A. K. (1917). Solution of some probability problems of significance for automatic telephone exchanges. *Elektroteknikeren*, *13*, 5-13.
2. Fortet, R. M. (1956). Random distribution with an application to telephone engineering, *Proc. Berkeley Sympos., Math. Stat. and Prob*. (81-88). Los Angeles, Berkeley.
3. Sevastyanov, B. A. (1957). An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Theor. Prob. Appl. (in Russion),* 2, 109-112.
4. Palm, C. (1943). Intensitätschwankungen Fernsprechverkehr. *Ericsson Technics*, *44*, 1-189.
5. Konig, D. & Matthes, K. (1963). Werallgemeiherungen der erlangschen formelu. *Math. Nachr*., *26*, 45-56.
6. Takacs, L. (1969). On Erlang's formula. *Ann. Math. Stat*., *40*, 71-78.
7. Shanbhag, D. N. & Tambouratzis. D. G. (1973). Erlang's formula and some results on the departure process for a loss system. *J. Appl. Prob*., *10*, 223-240.
8. Brumelle, S. L. (1978). A Generalization of Erlang's loss system to state dependent arrival and service rates. *Math. Operat. Res*., *3*, 10-16.
9. Gumbel, M. (1960). Waiting lines with heterogeneous servers. *Opns. Res*., *8*, 504-511.
10. Blanc, J. P. C. (1987). A Note on waiting times in systems with queues in parallel. *J. App. Prob*., *24*, 540 -546.
11. Nath, G. B. & Enns, E. B. (1982). Optimal service rates in the multiserver loss system with heterogeneous servers. *J. App. Prob*., *18*, 776-781.
12. Chaiken, J. M. & Ignall, E. (1972). An Extension of Erlang's formulas which distinguishes individual servers. *J. Appl. Prob*., *9*, 192-197.
13. Wolff, R. W. and Wringhtsonon, C. W. (1976). An extension of Erlang's loss formula. *J. Appl. Prob*., *13*, 628-

632

14. Mors, P. M. (1958). *Queues, Inventories and Maintenance*. New York, Wiley.
15. Saaty, T. L. (1960). Time depended solution of many-server Poisson queue. *Opns. Res*., *8*, 768-771.
16. Singh, V. P. & Prasad, J. (1976). A heterogeneous system with finite waiting space. *J. Engin. Math*., 10, 125-134.
17. Singh, V. P. (1970). Two-server Markovian queues with balking: Heterogeneous and homogenous servers. *Opns. Res*., *18*, 145-159.
18. Shahbazov, A. A. (1983). Minimizing of loss probability in queuing system with two heterogeneous servers*. Izv. Acad. Sci. Azerb. SSR, Ser. Phiz. Tech and Mat .Sci. (in Russion), 2*, 130-135.
19. Knessl, C., Matkowsky, J., Schuss, Z. & Tier, C. (1990). On integral equation approach to the *M*/*G*/2 queue. *Opns. Res*., *38*, 506-518.
20. Bhat, U. N**.** (1966). The queue G/M/2 with service rate depending on the number of busy Servers*. Ann. Ins. Stat. Math*., *18*, 211-221.
21. Shahbazov, A. A. (1984). Loss queue with different servers. *Zasthosaw. Mat*., *18,* 177-186
22. Hardy, G. H., Littlewood, J. E. & Polya, G. (1954). *Inequalities*. London, Cambridge University Press.
23. Takacs, L. (1957). On a probability concerning telephone traffic, *Acta Mat. Acad. Sci. Hung, 7*, 419-433.
24. Smith, D. R. (1968). Optimal repair of a series system. *Opns. Res., 26*, 653-662.
25. Anantharam, V., Goppinath, B. & Hajela, D. J. (1988). A generalization of the Erlang formula of traffic engineering queuing system. *Queueing Systems, 3*, 277-288.
26. Cinlar, E. & Isney, R. L. (1967). Streams of overflows from a finite queue, *Opns. Res., 15*, 131-134.
27. Benes, V. E. (1959). On trunks with negative exponential holding times serving a renewal process. *Bell system Tech. J*., *38*, 211-258.