THE ROBUSTNESS OF BLOCKING PROBABILITY IN A LOSS SYSTEM WITH REPEATED CUSTOMERS

Akira Takahashi, Yoshitaka Takahashi
Graduate School of Commerce , Waseda University
Shinjuku, Tokyo 169-8050, Japan

Shigeru Kaneda, Yoshikazu Akinaga and Noriteru Shinagawa
Network Laboratories , NTT DoCoMo,Inc.
3-5, Hikarinooka, Yokosuka, Kanagawa, 239-8536, Japan

Keywords: Teletraffic analysis, loss system, repeated customers, Little's formula.

Abstract: In this paper, we analyze and synthesize a multi-server loss system with repeated customers, arising out of NTT DoCoMo-developed telecommunication networks. We first provide the numerical solution for a Markovian model with exponential retrial intervals. Applying Little's formula, we derive the main system performance measures (blocking probability and mean waiting time) for general non-Markovian models. We compare the numerical and simulated results for the Markovian model, in order to check the accuracy of the simulations. Via performing extensive simulations for non-Markovian (non-exponential retrial intervals) models, we find $robustness$ in the blocking probability and the mean waiting time, that is, the performance measures are shown to be insensitive to the retrial intervals distribution except for the mean.

1 INTRODUCTION

When the service system becomes extremely congested, a lot of customers cannot receive immediate service. Some of them may give up the service to leave the system, while others may stay in the system and retry their requests. This behavior of repeated customers leads to an additional load on the system and worsens its congestion. The importance of repeated requests on the performance of the service system was pointed out in the late 1940, and many researches have been performed since then. Pioneering studies on the multi-server loss system with repeated calls brought some kind of positive expressions of performance measures (See (Falin and Templeton, 1997), (Artalejo and Pozo, 2002), and (Udagawa and Miwa, 1965)). However, they are not necessarily convenient to calculate performance measures. Retrial queuing models including one discussed here are usually very complicated for queuing analysis and its results are not always suitable to numerical calculation. Many authors reported numerical approaches of approximation and truncation methods. For details on the numerical approaches, readers are referred to (Artalejo and Pozo, 2002) and (Stepanov, 1999).

Most of them assume that the time intervals between repeated attempts are mutually independent and exponentially distributed. However, affected by many factors and circumstances, customers' behavior in repeating is so complex that these assumptions may lead to a risky assessment. There is necessity for generalization of the retrial interval distribution.

This assumption of exponential retrial intervals is a kind of simplification for queuing analysis. There is no guarantee that repeating customers behave in such a manner. Under this assumption, the number of repeated requests emerging in a unit time changes by the state of the retrial queue (See (Artalejo et al., 2001)). There is another type of retrial queuing model in which retrial rate is constant. In this type of model, blocked customers who want to repeat must wait in line and only the customer at the head of the line can retry to hunt, if any, an idle server. It has a wide range of applications like communication protocols. However, still there are systems more appropriately modeled by the classical type. On the constant retrial policy, there are fruitful investigations of models with single-server non-exponential retrial intervals like (Gómez-Correl and Ramalhoto, 1999). However, it remains open problem to investigate the effect of the retrial times distribution on the performance of the system. Customers' behavior in repeating is expected to be highly complex and it may be risky or inefficient to build and operate the system upon the results of exponential assumption. Thus, one finds it necessary to study sensitivity and robustness of the retrial

Takahashi A., Takahashi Y., Keneda S. and Akinaga Y. (2005).
THE ROBUSTNESS OF BLOCKING PROBABILITY IN A LOSS SYSTEM WITH REPEATED CUSTOMERS.
In *Proceedings of the Second International Conference on e-Business and Telecommunication Networks*, pages 61-66
DOI: 10.5220/0001416300610066

time distribution

The main goal of this paper is to investigate the robustness (insensitivity) property between the performance measure and the retrial time distribution in a loss system with repeated customers seen in an NTT DoCoMo developed telecommunication network.

The rest of the paper is organized as follows. Section 2 gives teletraffic analysis of the retrial model. The main performance measures of practical interest are then derived. In Section 3, simulation results of non-exponential (deterministic/ two-stage Erlang/ two-stage hyper-exponential) retrial models are compared to find robustness of the system.

2 NUMERICAL RESULTS

2.1 Model Description

Consider the following loss system with repeated calls : (1) There are c servers in parallel; (2) Customers' service times, which are identical and independent from one another, are exponentially distributed with rate μ; (3) Customer arrivals follow a Poisson process of rate λ; (4) Customers who find all severs busy at their arrival epoch choose either to repeat their requests with probability p or to give up the service with probability $(1-p)$; (5) When they decide to repeat, customers wait in the retrial queue for a random time called "retrial interval", which is exponentially (generally) distributed with parameter γ in Section 2.2(Section 3) before making repeated requests; (6) Retrial customers who find again all servers busy choose either to waint in the retrial queue and repeat their requests with probability p or to stop repeating and leave the system with probability $(1-p)$; (7) Give-up customers leave the system immediately.

We introduce following notations. Suppose the existence of the stationary state, the state of the system is characterized by (1) the number of the busy servers and (2) the number of the customers waiting to make a repeated attempt. The system will be said to be in state (i,j) , if i servers busy and j customers waiting to repeat. If there are c servers in the system then the system is somewhere in the state space $\{0,1,\cdots,c\}\times\{0,1,\cdots\}$. Let $\pi_{i,j}$ denote the probability that the system is in state (i,j) from now on.

2.2 Calculation of the Stationary Distribution

By focusing on the possible state transition in a minute time Δt, we get the state-transition probabilities as shown in Table 1. Figure 1 illustrates the state-transition diagram of this model.

Table 1: State-transition probabilities

state-transition	probability
$(i+1,j)$ $\uparrow \quad (0 \leqq i \leqq c-1, 0 \leqq j)$ (i,j)	$\lambda\Delta t + o(\Delta t)$
$(i+1,j-1)$ $\nwarrow \quad (0 \leqq i \leqq c-1, 1 \leqq j)$ (i,j)	$j\gamma\Delta t + o(\Delta t)$
(i,j) $\downarrow \quad (1 \leqq i \leqq c, 0 \leqq j)$ $(i-1,j)$	$i\mu\Delta t + o(\Delta t)$
$(c,j) \quad \rightarrow \quad (c,j+1)$ $(0 \leqq j)$	$\lambda\alpha\Delta t + o(\Delta t)$
$(c,j-1) \quad \leftarrow \quad (c,j)$ $(1 \leqq j)$	$j\gamma(1-\alpha)\Delta t$ $+o(\Delta t)$
$(c,j-1) \quad \leftarrow \quad (c,j)$ $(1 \leqq j)$	$j\gamma(1-\alpha)\Delta t$ $+o(\Delta t)$

By Table 1, the state-equilibrium equations are expressed as below.

$$
\begin{aligned}
(\lambda + j\gamma)\pi_{0,j} &= \mu\pi_{1,j}. \\
(\lambda + i\mu + j\gamma)\pi_{i,j} &= \lambda\pi_{i-1,j} + (j+1)\gamma\pi_{i-1,j} \\
&\quad + (i+1)\mu\pi_{i+1,j} \\
&\quad (1 \leqq i \leqq c-1). \\
(\lambda\alpha + c\mu + j\gamma(1-\alpha))\pi_{c,j}. & \\
= \lambda\pi_{c-1,j} + \gamma\pi_{c-1,j} + \lambda\alpha\pi_{c,j-1} & \\
+ (j+1)\gamma(1-\alpha)\pi_{c,j+1}. &
\end{aligned}
$$

Our research purpose here is to study the effect of the retrial interval distribution and the existence

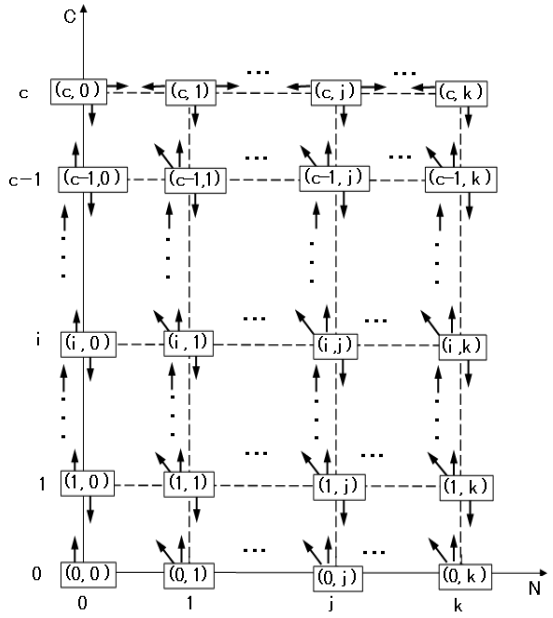

Figure 1: State-transition diagram.

of robustness. To this end, we adopt an simple approximation method of replacing the infinite space for customers waiting to repeat requests by a finite number k. It is an extension of the way to calculate the steady-state probabilities introduced in (Hashida and Kawashima, 1979) and closely explained in (Falin and Templeton, 1997), so we only show the outline of the algorithm. Here, k is assumed to be a sufficiently large positive integer so that the overflow probability is small enough to be ignored. From a finite capacity argument,

$$\text{for } 1 \leqq j \leqq k-1,$$
$$(\lambda + j\gamma)\pi_{0,j} = \mu\pi_{1,j}, \tag{1}$$

$$(\lambda + i\mu + j\gamma)\pi_{i,j} = \lambda\pi_{i-1,j} + (j+1)\gamma\pi_{i-1,j} + (i+1)\mu\pi_{i+1,j} \quad (1 \leqq i \leqq c-1), \tag{2}$$

$$(\lambda\alpha + c\mu + j\gamma(1-\alpha))\pi_{c,j} = \lambda\pi_{c-1,j} + \gamma\pi_{c-1,j} + \lambda\alpha\pi_{c,j-1} + (j+1)\gamma(1-\alpha)\pi_{c,j+1}. \tag{3}$$

For $j = k$,

$$(\lambda + k\gamma)\pi_{0,k} = \mu\pi_{1,k}, \tag{4}$$

$$(\lambda + i\mu + k\gamma)\pi_{i,k} = \lambda\pi_{i-1,k} + (i+1)\mu\pi_{i+1,k} \quad (1 \leqq i \leqq c-1), \tag{5}$$

$$(c\mu + k\gamma(1-\alpha))\pi_{c,k} = \lambda\pi_{c-1,k} + \lambda\alpha\pi_{c,k-1}. \tag{6}$$

These recurrence equations enable us to compute the stationary distribution via the following steps.

(I) Take the appropriate k and introduce auxiliary variables $\phi_{i,j} \triangleq pi_{i,j}/\pi_{0,k}$.
(II) By definition , $\phi_{0,k} = 1$.
From (4), $\phi_{1,k}$ can be determined.
From (5), one can get $\phi_{i,k}(i = 2, 3, \cdots, c)$ sequentially.
(III) Equations (1) and (2) for $i = 1, \cdots, c-1$ constitute a set of c equations with $c+1$ unknown variables $\phi_{0,k}, \phi_{1,k}, \cdots, \phi_{c-1,k}$, $\phi_{c,k}$. Thus, with $\phi_{c,k}$ obtained by (6), one finds the set of equations become solvable .
Hence, (3) gives $\phi_{c,k-2}$D
(IV) Operating steps (I), (II), and (III) repeatedlyC we get all of the $\phi_{i,j}$.
The normalization condition;
$\sum_{i=0}^{c}\sum_{j=0}^{k}\phi_{i,j} = 1/\phi_{0,k}$
settles $\pi_{0,k}$ and $\phi_{i,j} \times \pi_{0,k}$ gives $\pi_{i,j}$.
(V) By repeating from (I) to (IV) with k plus 1 until the value of $\pi_{c,k}$ becomes less than 10^{-10}, $\pi_{i,j}$ can be calculated with an accuracy enough for our purpose.

2.3 Performance Measures

We are now in a position to derive the performance measures of the system.

Time congestion (B_T)

Letting B_T be the time congestion, so called, the probabilities that all the servers are busy, we have

$$B_T = \sum_{j=0}^{k} \pi_{c,j}.$$

Blocking probability (B)

When they blocked due to all servers busy, customers can wait for some random time and retry. After several retrials, some of them may give up the service demand, and leave the system. Here, we define the

blocking probability B as the probability that customers finally leave without getting served due to successive blockings.

To the best of authors' knowledge, there are few investigations for a general retrial model. Here, applying Little's formula (Little, 1961) enables us to prove the following proposition.

Proposition 1 Consider a general retrial queuing system which has input with rate λ, service with rate μ and retrial with rate γ. Customers who try to receive a service and get blocked due to all servers busy, choose either to repeat their requests with probability p or to stop repeating and leave the system with probability $(1-p)$. Blocking probability B, that is, the probability that arriving customers finally leave the system with not receiving the service, is expressed by

$$B = 1 - \frac{1}{\rho}\overline{C}.$$

Here, ρ denotes traffic intensity defined by λ/μ and $\overline{C}$ stands for the mean number of busy servers on the stationary condition. See Appendix 1 for the proof.

It should be noted that Proposition 1 has a different expression on the blocking probability from that in (Hashida and Kawashima, 1979), where the $PASTA$ ($Poisson\ Arrivals\ See\ Time\ Averages$) property is heuristically used to provide an approximation. Our expression on the loss probability shown in Proposition 1 is exact (not approximate).

Mean waiting time (Wq)

Denote by Wq the mean waiting time, namely, the mean elapsed time from a customer's arrival epoch until the epoch where the customer gets served or stops repeating without receiving its service to leave the system.

Like B above, Wq is also derived from Little's formula and its relation to other parameters is preserved under more general situation. So we find the following proposition.

Proposition 2 Consider a general retrial queuing system which has input with rate λ, service with rate μ and retrial with rate γ. Customers who try to receive a service and get blocked due to all servers busy, choose either to repeat their requests with probability p or to stop repeating and leave the system with probability $(1-p)$. The mean waiting time Wq, that is, the time that customers have to spend on average until they finally get served or decide to stop repeating and leave, is expressed by

$$Wq = \frac{\overline{K}}{\lambda}.$$

$\overline{K}$ is the mean number of customers in the retrial area in the steady state. See Appendix 2 for the proof.

3 SIMULATION RESULTS

In the previous section, we get the numerical solution of the loss system with exponential retrial intervals. Next, we change the assumption about retrial. In this section we compare performance measures between the exponential retrial interval model and the models with non-exponential retrial intervals. Even under the exponential retrial interval assumption, multi-server property involves great complicity and analytical solutions are obtained only a few special cases like (Falin and Templeton, 1997) and (Choi and Kim, 1998). So we employ computer simulation to estimate the performance measure of non-exponential retrial interval models. The assumptions for simulation are all the same with those for numerical calculation introduced in Section 2 except for the distribution of retrial intervals. It assumes a Poisson arrival of customers with rate λ and an identically independently distributed exponential service time with rate μ.

On the distribution of retrial intervals, in this paper we take four different models; the exponential retrial interval model (Exp model) the constant retrial interval model (D model), the 2-stage Erlang distribution model ($E2$ model), and the 2-stage hyper exponential distribution model ($H2$ model). Among $H2$ models, we also have three different types whose coefficient of variation (C_X) of the retrial interval distribution is lager than 1, equal to 1, or smaller than 1. In other words, the variance of the retrial interval distribution is large, equal or small in comparison to its mean. $H2(C_X = \sqrt{2})$, $H2(C_X = \sqrt{20})$ and $H2(C_X = \sqrt{200})$ denote the model with hyper-exponential retrial intervals whose C_X equals to $\sqrt{2}, \sqrt{20}$ and $\sqrt{200}$, respectively .

Through this section, $\tau \triangleq \mu/\gamma$ is used for the indicator of the mean retrial interval and $\rho \triangleq \lambda/\mu$ for the traffic offered to the whole system.

In simulation, c(= the number of servers) is set to 10, μ 0.01and p 5/6, which means the service time average is 100 and under the condition of successive blocking customers continue to repeat 5 times on average. An individual simulation results (expressed as points in each figure) is based on 50 runs(approx. 5 hours on IBM Thnkpad PC).

First, the accuracy of the simulation should be investigated. Figure 2 shows the blocking probability B by numerical calculation and simulation with the mean retrial time $\tau = 1$. As seen in Figure 2, we cannot see significant difference between our numerical and simulation results. Therefore, our simulation results are very accurate. The accuracy of simulation is confirmed on other performance measures.

Now that we see the accuracy of the simulation, comparisons are performed when the mean retrial interval τ is 0.01, 1, and 100.0, which corresponds to

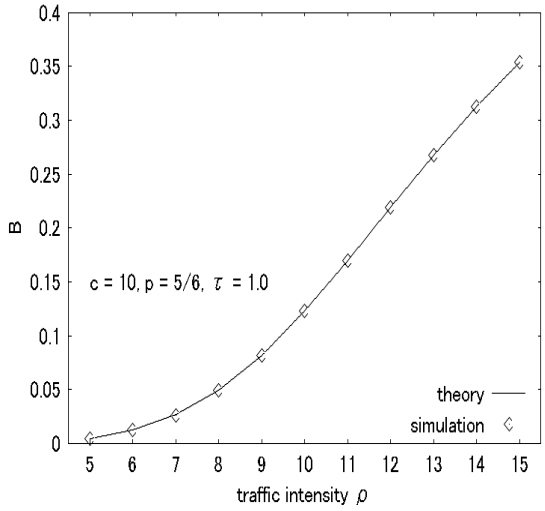

Figure 2: The blocking probability B by numeric al calculation and simulation.

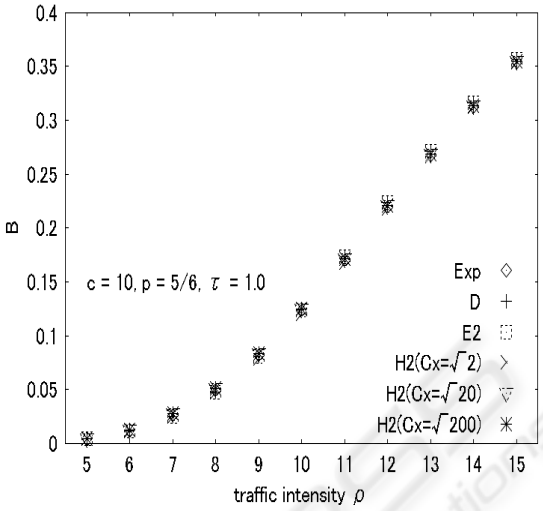

Figure 3: The blocking probability B versus the traffic intensity ρ.

the situations that repeated requests arises 100 times sooner than the service time on average, that they arise with the interval as long as the service time on average, and that they arise after a time 100 times longer than the service time on average.

Figures 3 and 4 show the relationship of the blocking probability B and the mean waiting time Wq to the traffic intensity ρ. One finds the retrial interval distribution makes little difference.

4 CONCLUSION

We have analyzed and synthesized a multi-server loss system with repeated customers, arising out of NTT DoCoMo-developed telecommunication networks. We have first provided the numerical solution for a Markovian model with exponential retrial intervals. Applying Little's formula, we have derived the main system performance measures (blocking probability and mean waiting time) for general non-Markovian models. We have compared the numerical and simulated results for the Markovian model, in order to check the accuracy of the simulations. Via performing extensive simulations for non-Markovian (non-exponential retrial intervals) models, we have found *robustness* in the blocking probability and the mean waiting time, that is, the performance measures have been shown to be insensitive to the retrial intervals distribution except for the mean.

It is left for future work to investigate the robustness for a more general (e.g., a general service time distribution) model.

ACKNOWLEDGEMENT

The present research was partially supported by a Grant-in-Aid for Scientific Research (C) from Japan Society for the Promotion of Science under Grant No. 1458049.

REFERENCES

Artalejo, J., Gómez-Correl, A., and Neuts, M. (2001). Analysis of multiserver queues with constant retrial rate. *European Journal of Operational Research*, 135:569–581.

Artalejo, J. and Pozo, M. (2002). Numerical calculation of the stationary distribution of the main multiserver retrial queue. *Annals of Operations Research*, 116:41–56.

Choi, B. and Kim, Y. (1998). The M/M/c retrial queue with geometric loss and feedback. *Computers and Mathematics with Applications*, 36:41–52.

Falin, G. and Templeton, J. (1997). *Retrial Queues*. Chapman and Hall, London, 1st edition.

Gómez-Correl, A. and Ramalhoto, M. (1999). The stationary distribution of a markovian process arising in the theory of multiserver retrial queueing systems. *Mathematical and Computer Modelling*, 30:141–158.

Hashida, O. and Kawashima, K. (1979). Buffer behavior with repeated calls. *The IECE Transactions*, J62-B:222–228.

Little, J. D. C. (1961). A proof for the queuing formula: L = λ W. *The Journal of the Operations Research Society of America*, 9:383–387.

Stepanov, S. (1999). Markov model with retrials:the calculation of stationary performance measures based on the concept of truncation. *Mathematical and Computer Modelling*, 30:207–228.

Udagawa, K. and Miwa, E. (1965). A complete group of trunks and poisson-type repeated calls which influence it. *The IECE Transactions*, 48:1666–1675.

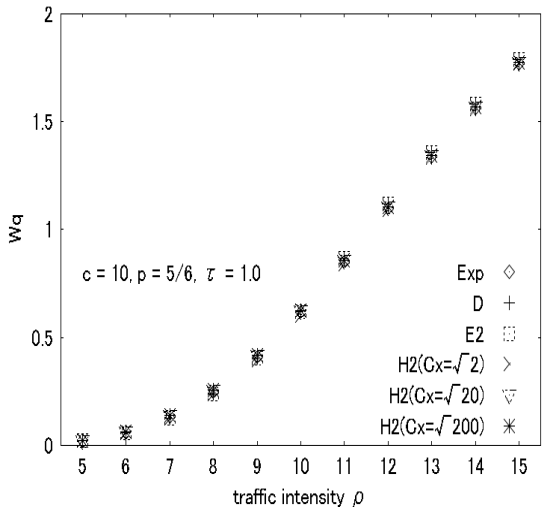

Figure 4: The mean waiting time Wq versus the traffic intensity ρ.

APPENDICES

Appendix 1

Proof of Proposition 1

We restrict ourselves to the sub-system only composed of c servers. We apply Little's formula [$L = \lambda W$] (Little, 1961) to this sub-system.

Let λ' and $\overline{C}$ respectively denote the sub-system throughput and the mean number of busy servers. By Little's formula, we have

$$\overline{C} = \lambda' \frac{1}{\mu}, \tag{A.1}$$

now that $\overline{C}$ [the mean number of customers in the subsytem] corresponds to L, λ' [the effective arrival rate of the sub-system] corresponds to λ, and $1/\mu$ [the mean sojourn time in the sub-system] corresponds to W.

The blocking probability B is defined as the probability that an arriving customer cannot finally receive its service, however often it may repeat the retrial process [being blocked, waiting, and retrying].

Since on average λ customers arrive at the system in unit time, the mean number of customers who leave the system without being served is given by λB. The mean number of customers who receive their services is obtained as

$$\lambda(1 - B) = \lambda'. \tag{A.2}$$

Substituting (A.2) into (A.1), and solving for B we finally get

$$B = 1 - \frac{1}{\rho}\overline{C}. \qquad \square$$

Appendix 2

Proof of Proposition 2

We restrict ourselves to the sub-system only composed of the retrail queue (with a finite capacity of k customers). We apply Little's formula [$L = \lambda W$] to this sub-system.

Let S and $\overline{K}$ denote the mean number of retrials and the mean number of customers in the retrial queue, respectively. Since on average λ customers arrive at the system in time unit, then each one of them go through the retrial queue S times on average. That is, the mean number of customers who go to the retrial queue in time unit equals to $S\lambda$. By Little's formula, we have

$$\overline{K} = S\lambda \frac{1}{\mu} \tag{A.3}$$

now that $\overline{K}$ [the mean number of customers in the subsytem] corresponds to L, $S\lambda$ [the effective arrival rate of the sub-system] corresponds to λ, and $1/\mu$ [the mean sojourn time in the sub-system] corresponds to W. From (A.3), we have

$$S = \frac{\overline{K}\gamma}{\lambda} \tag{A.4}$$

Since the mean retrial interval is $1/\mu$ and customers repeat their requests S times on average, then the mean waiting time Wq is

$$Wq = S\frac{1}{\gamma} \tag{A.5}$$

Substituting (A.4) to (A.5), we get

$$Wq = \frac{\overline{K}}{\lambda}. \qquad \square$$