International Journal of Applied Mathematics

Volume 30 No. 1 2017, 1-9
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
doi: http://dx.doi.org/10.12732/ijam.v30i1.1

REDUCING SERVER REDUNDANCIES IN ATM LINES VIA QUEUE DISCIPLINES

Sulaiman Sani[1 §], Hamisu Musa[2], Hali Ibrahim Abba[3]

[1]Department of Physical Sciences
Kampala International University
Kampala, UGANDA

[1,2,3]Department of Mathematics & Computer Science
Umaru Musa Yar'Adua University
Katsina, NIGERIA

Abstract: We solve a machine redundancy problem in ATM centers via service policy design. Generally, redundancy has the tendency to increase the cost of service. Using the supplementary variable technique, we derive models for the customer process. A comparative analysis with certain existing models shows that machines operating under our design are better utilized. Thus, the design is a good alternative for reasons do to with managing investment resources.

AMS Subject Classification: 60K25
Key Words: server redundancy, flexible service policy, cost of service, buffer

1. Introduction

We consider an ATM center with a fixed and a flexible machine. Customers arrive according to a Poisson process at a rate λ for service on any available machine present in the system upon arrival. We suppose that the fixed machine has a faster service rate compared with the flexible one. Thus, the fixed machine

Received: September 18, 2016

[§]Correspondence author

is always busy when there is at least a customer in the system. The service times of customers are assumed to be a sequence of mutually independent and identically distributed random variables with finite moments. In addition, services are without preemption. If a customer is served by the fixed machine, his service time follows the exponential distribution with a mean service rate μ_1. Similarly, for customers served by the flexible machine, the mean service time is β. We supposed that the service rate for the flexible machine is $\mu_2 = \frac{1}{\beta}$, the utilization parameter for the fixed machine is $\rho_1 = \frac{\lambda}{\mu_1}$ and the combined utilization rate is $\rho_s = \frac{\lambda}{\mu_1 + a\mu_2} < 1$; for some $a \in (0, 1)$.

Generally, the need to manage cost of providing services cannot be over emphasized especially at present when scarcity of investment resources round the globe is a reality. Consequently, new management models such as the one proposed in this work are essentially not out of place at least for reasons to do with effective management. In ATM centers for instance, one area where high wastage tendency is evident and effective management is required is the schedule for service. For instance, a bank fixes two ATM machines to serve an expected customer size say $n \in N$. Suddenly, another bank fixes her own machines in the neighborhood of the two already fixed machines independently. Obviously, realizing n by the former station becomes difficult or even impossible. In addition, the realization tendency for such $n \in N$ becomes less and less probable with increasing independent installations in an open neighborhood. Consequently, machine redundancy is imminent in the system a.s.

To show this, suppose that the two fixed machines are jointly expected to serve specifically 50 customers on average. Suddenly, 2 additional machines are fixed independently in an open neighborhood containing the two already fixed machines. This reduces the expected 50 customers to a value below 50 leading to lost of work. The scenario is similar to a football player that spends most time on the bench devoid of playing time. In continuous time, the player losses gear due to low level of work. As expected, additional resources must be spent to rebuild his capacity to fit the team again because of his redundancy situation. This and similar scenarios make the realization of business expectations hand in hand with management really difficult under fixed server installation policies.[1] Hence, the need to design alternative service policies that can reduce policy-based wastage tendencies for reasons to do with prudent management of investment resources.

Similarly, in a two-machine ATM center, one may arrive when both ma-

[1] Because of the high probability of wasting an already constrained resources through embedded server redundancies.

chines are busy or at least only the most reliable machine[2] is available for service. If there is only one customer for service, then he is serviced by the first machine provided that no any other customer arrives during his service period. But if at least one additional customer arrives, then the second customer may choose to wait for the remaining service time of the first customer or takes service if another machine is present. As a schedule, if both machines are available anytime, then both are put to service. Otherwise, the first machine is the only available machine in the system.[3] We refer to this service policy as *flexible* as implied in *the flexible queue discipline* below. Note that the class of flexible queue policies in service systems is not extensively studied in the literature; Emrah et al. [2] and Mohammad and Ali [4].

Our motivation stemmed from the quest to minimize waste of investment resources in ATM centers as a result of service policies. An in-depth analysis therefore is an excellent tool for decision making relative to resource management, better server utilization, etc. Basically, service providers hate to waste resources due to preventable in-effectiveness.

2. The Flexible Queue Discipline

An arbitrary customer arrives to find:

1. **The fixed machine idle:** The customer is allocated the fixed machine.

2. **The fixed machine busy:** With probability $P_{1,1}$, the customer is allocated the flexible machine.

3. **Both machines busy:** The customer waits for his service turn according to the condition whether both machines are available with probability $P_{1,1}$ or only the first machine is available with probability $P_{1,0}$.

4. The buffer is either finite or otherwise in capacity.

[2] The first machine in our case.

[3] The installation/de-installation of the second machine may be on the basis of the customer parameter n as seen by the service provider.

3. The Stationary Customer Analysis

Suppose an ordered paired process $\{X(t), \zeta(t)\}_{t\geq 0}$ is given. Where $X(t)$ denotes the number of customers in the system at time t and $\zeta(t)$ is the past service time of an arbitrary customer on the flexible machine. Looking at the system at departure instants on the flexible machine[4], the bi-variate process $\{X(t), \zeta(t)\}_{t\geq 0}$ is a Markov process. If the service time of customers is continuous and the system is empty at time zero. Then one can apply the supplementary variable technique to analyze the process $\{X(t), \zeta(t)\}_{t\geq 0}$.

Let

$$P_{0,0}(t) = P(X(t) = 0 : \text{ both machines idle}),$$

$$P_{1,0}(t) = P(X(t) = 1, \text{ only fixed machine busy}),$$

$$P_{1,1}(t) = P(X(t) \geq 2; \text{ both machines busy}).$$

If $\lambda < \mu_1 + \frac{1}{\beta}$, then as $t \to \infty$, $P_{0,0}(t)$, $P_{1,0}(t)$ and $P_{1,1}(t)$ converge to $P_{0,0}$, $P_{1,0}$ and $P_{1,1}$ respectively. Let $P_j = P[X = j : j = 0, 1, 2, ...]$ denotes the stationary probability that there are j customers in the system. Suppose that the rate equality principle holds. Then by the ergodic theorem, the following set of difference equations are satisfied

$$\lambda P_0 = \mu_1 P_{1,0}; \qquad j = 0. \tag{1}$$

$$(\lambda + \mu_1) P_{1,0} = \lambda P_0 + \mu_1 P_2 + \frac{1}{\beta} P_{1,1} P_2; \qquad j = 1. \tag{2}$$

For $j \geq 2$, we have

$$\left(\lambda + \mu_1 + \frac{1}{\beta}\right) P_j = \lambda P_{j-1} + \left(\mu_1 + \frac{1}{\beta} P_{1,1}\right) P_{j+1}. \tag{3}$$

Combining (1) through (3) together with the Markov property of the customer states, we have

$$P_j = \left(\frac{\lambda}{\mu_1 + \frac{1}{\beta} P_{1,1}}\right)^{j-1} \left(\frac{\lambda}{\mu_1}\right) P_0; \qquad j \geq 1. \tag{4}$$

[4] That is; immediately a service is completed in a service period when both machines are busy.

Put $\frac{\lambda}{\mu_1+\frac{P_{1,1}}{\beta}} = \rho_s$ and $\frac{\lambda}{\mu_1} = \rho_1$ in (4) and apply the normalization condition. One obtains that

$$P_j = \frac{\rho_s^{j-1}\rho_1(1-\rho_s)}{(1-\rho_s)+\rho_1}. \tag{5}$$

Additionally, if $j \leq K$ for some $K \in \Re^+$, then from (4), it can be shown that

$$P_K = \frac{\rho_s^{K-1}\rho_1(1-\rho_s)}{(1-\rho_s)+\rho_1(1-\rho_s^K)}; \qquad K \geq 1. \tag{6}$$

Suppose we wish to compute the stationary number of customer $E[X]$ in the system. Let $W(z)$ denotes the PGF for X. From (5), we have by definition of $W(z)$ that

$$W(z) = P_0 + \rho_1 z P_0 + \rho_1 \rho_s z^2 P_0 + ..., \tag{7}$$

so that upon simplification, we have

$$W(z) = \frac{P_0[(1-\rho_s z)+\rho_1 z]}{1-\rho_s z}. \tag{8}$$

And upon differentiating (8) at $z = 1$, one obtains that

$$E[X] = \frac{\rho_1(1-\rho_s)}{(1-\rho_s)^3+\rho_1(1-\rho_s)^2}. \tag{9}$$

Suppose that the buffer for the ATM center is finite. Let $j \leq K < \infty$ and $U(z)$ denote the size of the buffer and the PGF for the number of customers in the system for such K respectively. We have from (4) that

$$U(z) = P_0 + \rho_1 z P_0 + ... + \rho_1 \rho_s^{K-1} Z^K P_0, \tag{10}$$

so that

$$U(z) = \frac{P_0(1-\rho_s z)+\rho_1 z P_0(1-\rho_s z)^{K-1}}{1-\rho_s z}. \tag{11}$$

Differentiating (11) w.r.t. z at 1, we have

$$E[X] = \frac{\rho_1 P_0[1-\rho_s^{K-1}(K-\rho_s(K+1))]}{(1-\rho_s)^2}. \tag{12}$$

Additionally, if $K_{\min}$ denotes the minimal buffer for the queuing system in question. Then $K_{\min}$ exists. To show this, we have from (6) that

$$K = 1 + \frac{\ln\left[\frac{P_K(1-\rho_s+\rho_1)}{\rho_1(P_K\rho_s+1-\rho_s)}\right]}{\ln \rho_s}. \tag{13}$$

Suppose that $P_K \to \frac{\rho_1}{1+\rho_1}$ in an arbitrary service period. Then $\rho_s \to \rho_1$ a.s. Consequently, $K \to K_{\min}$.

4. Results

Lemma 1. *Suppose that the flexible machine breaks down in a service period with $K > 0$ customers in the system. Then*

$$P_K = \frac{(1-\rho_1)\rho_1^K}{1-\rho_1^{K+1}}. \tag{14}$$

Proof. If the flexible server breaks in a service epoch, then it is trivial that $\rho_s \to \rho_1$. Under this condition[5], the flexible machine is not necessary for the stability of the system. The lemma holds if ρ_1 substitutes ρ_s in (6) upon simplification. □

The blocking probability in (14) corresponds to the well known blocking probability for the $M/M/1/K$ model in MacGregor [3] for finite values of K and $\rho_1 < 1$. Interestingly, the flexible server policy introduced in this work has minimal buffer parameter compared with the fixed server policy of Daman and Sulaiman [1]. No doubt then, it is a good policy for effective server management.

Lemma 2. *Suppose $\lambda = \mu_1$. Then the realization $\{X = j\}$ of the Markov Chain $\{X_j\}$ is ergodic if and only if $P_{1,1} > 0$.*

Proof. By the stability condition $\rho_s = \frac{\lambda}{\mu_1 + a\mu_2} < 1$, we have for $\lambda = \mu_1$ that $a\mu_2 > 0$. Consequently, $a > 0$ is necessary. Conversely, suppose that $a > 0$. That means for any non zero number μ_2, we have $a\mu_2 > 0$. Consequently, if $\lambda < a\mu_2$ is a real number, then $\frac{\lambda}{\mu_1 + a\mu_2} < 1$. Then the system is stable. □

Lemma 3. *According to Lemma 2, the flexible server installation policy here is equivalent to any fixed server installation policy.*

Proof. Intuitively, the stationary customer process (X, ζ) in both models is saddled on the fixed exponential machine. Given that $\lambda < \mu_1 + \frac{1}{\beta}$ holds, it is trivial that $\lambda < \mu_1$ for the $M/M/1$ is implied. Consequently, $\rho_1 = \frac{\lambda}{\mu_1}$ for both models. Denote by ρ the occupation rate of the classical $M/M/1$ model. By definition $\rho = \frac{\lambda}{\mu}$. This means that $\rho_1 = \rho$. Hence, the proof follows. □

[5] The condition when $P_{1,1} = 0$.

λ	$fixed$	$flex.$
3.7	0.23871	0.30753
4.9	0.31613	0.44928
5.0	0.32258	0.46243
5.5	0.35484	0.53172
7.5	0.48387	0.88561
7.9	0.50968	0.97606

Table 1: Utilization ρ_s

λ	$flex.$	$fixed$
3.7	0.57829	0.67419
4.9	0.95613	0.98369
5.0	1.00005	1.01163
5.5	1.27026	1.15672
7.5	7.79127	1.84382
7.9	40.78550	2.00757

Table 2: Expectations $E[X]$

5. Simulations & Remarks

We compare the performance of the flexible policy in this work with fixed policies of Sivasamy et al. [6] and Sulaiman and Daman [5]. For $\mu_1 = 8$ and $\mu_2 = 7.5$, the ρ_s, $E[X]$ and K are approximated. The results are shown in Tables 1 and 2.

Remark 4. The flexible server policy is operationally better under light traffic.

This can be seen from Table 2 that when the arrival rate λ is far from the combined service rate[6] $(\mu_1 + P_{1,1}\mu_2)$, the stationary values for $E[X](flex)$ is smaller than those of $E[X](fixed)$. Consequently, one can conclude that the flexible server policy is operationally better than the fixed server policy for such interval of arrival of customers in the system.

Remark 5. For any arrival rate $\lambda \in \Re^+$, the flexible policy is better

[6] In Sivasamy et al. [6], $P_{1,1} = 1$.

utilized.

From Table 1, one can see that the server utilization rates ρ_s under the flexible server policy is significantly higher than those of the fixed server policy. For instance, when the arrival rate is 7.9 as in table 1, it can be seen that the flexible server policy is about 98% utilized as against that of the fixed server policy of about 51% utilization.

Remark 6. The flexible server policy requires a smaller buffer during light traffic.

This is in view of (13) in comparison with (43) of Sulaiman and Daman [5]. Consequently, one can conclude that the flexible policy is operationally better in terms of buffer sizes for such interval of arrivals.

Corollary 7. *There exists a unique $\lambda \in \Re^+$ such that the two models are identical.*

There is a scope in extending the results here to a case with a single fixed server and several multiple flexible servers for comparison with multiple server fixed models for management reasons.

Acknowledgments

The authors are grateful to the journal editor for the valuable contributions, suggestions and editing this article. Also is to the anonymous referees.

References

[1] O.A. Daman and S. Sani, An M/G/2 queue with a finite buffer under dual control: Stationary lost and cost analysis, *International Journal of Applied Mathematics*, **28**, No 5 (2015), 527-540, **doi:** 10.12732/ijam.v28i5.6.

[2] B.E. Emrah, O. Ceyda and O. Irem, Parallel machine scheduling with additional resources: Notation, classification, models and solution methods, *European Journal of Operational Research*, **230** (2003), 449-463.

[3] J.S. MacGregor, M/G/c/K blocking probability models and system performance, *Performance Evaluation*, **52** (2003), 237-267.

[4] K. Mohammad and M. Ali, A hybrid method for solving optimal control problems, *IAENG International Journal of Applied Mathematics*, **42**, No 2 (2012), 80-86.

[5] S. Sani and O.A. Daman, The M/G/2 queue with heterogeneous servers under a controlled service discipline: Stationary performance analysis, *IAENG International Journal of Applied Mathematics*, **45**, No 1 (2015), 31-40.

[6] R. Sivasamy, O.A. Daman and S. Sani, An M/G/2 queue where customers are served subject to a minimum violation of the FCFS queue discipline, *European Journal of Operational Research*, **240** (2015), 140-146.