Should I add an Intermittent Server

Raymond A. Marie
IRISA/University Rennes 1
Campus de Beaulieu
35042 Rennes Cedex, France
Raymond.Marie@irisa.fr

ABSTRACT

This paper concerns the situation of a queue with one regular single server and, in order to decrease the mean response time, with a second server i) leaving the back office to join the first server when the number of customers reaches the threshold K, ii) leaving the front office when he has no more customers to serve. This study produces a closed form solution for the steady state probability distribution and for different metrics such as expected response times for customers or expectation of busy periods. Then, for a given value of K, the influence of the intermittent server on the response time is exhibited. The consequences on the primary task of the intermittent server are investigated through metrics such as mean working and pseudo-idle periods. Finally, a determination of an optimal value of the threshold K is proposed.

CCS Concepts

•**Mathematics of computing** → **Queueing theory;** •**Computing methodologies** → *Simulation evaluation;*

Keywords

Performance Evaluation; Response Time; Markovian model; Intermittent Server.

1. INTRODUCTION

For a course on discrete event simulation, a good example is the one of the supermarket check-out counters where a counter can be activated/deactivated based on the states of the different queues. This argument stays because the queuing model is easy to elaborate and has no (known) analytical solution in its general configuration. This help students to realize all the advantages of a simulation approach. In addition, such a model is easily adaptable to other fields such as those of telecommunication or of data centers. Of course we have to remember that, when possible, an analytical solution must be looked for since its cost is generally lower than the one of the simulation approach.

Although most of the research work in the domain of the M/M/r queue with intermittent servers has been done through the use of simulation, we noted some developments connected to the subject. In 1971, J. Blackburn published a report [1] relative to a M/G/1 queue the server of which is an intermittent server who starts working when the number of customers crosses some threshold. This threshold is the value realizing the optimum of an objective function. A more recent analytical study investigated the case of an airline check-in counters set in an airport [4]. In this study, Parlar et. al elaborated a Markovian model and its transient solution. A major difference with the supermarket check-out system is that the number of customers to be served is known in advance (number of customers who have a reserved seat for a given flight). The problem is to control the number of open check-in counters such that all the customers that will show up before a deadline T will be served on time (such that the plane can take off on time). But most of the literature involving intermittent servers concerns studies where the activations of the servers depend on reliability/availability of the set of servers rather than on the states of the systems.

Another related class of models is the "coupled processor model" where each processor can help the other when it is idle. The two queues have their own arrival processes and service time distributions. Such a class has been the object of intensive analytical works in the past. Close to that is the case where the behaviors of the servers are no more symmetrical and only one processor can, when it becomes idle, give time to the other processor until its own queue reaches a given threshold (see the intensive study of Osogami et al. [3]). Note also the different model known as "the slow server problem" (see [5]) where, depending on the values of the parameters, the use of the slow server may increase the response time.

The present study is different in the sens that the server who gives some part of his time is not idle but works on tasks which are not directly impacting customers (the notion of response time is in some sens meaningless). This study is less general than the one cited above ([3]) but produces a closed form solution for the steady state probability distribution and for different metrics such as expected response times for customers or expectation of busy periods. One objective is to promote a better understanding of the benefits of such a strategy. A second objective is to produce a way to check the simulation software with respect to the case where the number of servers is reduced to two.

The presented application is supposed to come from the banking sector but could come from a more industrial area. The basic situation corresponds to a service counter with one regular single server and, in order to decrease the mean response time, with a second server i) leaving the back office to join the first server when the number of customers reaches the threshold K, ii) leaving the front office when he has no more customers to serve. The aim of this paper is to bring answers to questions such as : for a given value of K, what is the influence of the intermittent server on the response

VALUETOOLS '16 October 26–28, 2016, Taormina, Italy

DOI: 00.000/000_0

DOI 10.4108/eai.25-10-2016.2266975

time? What is the frequency of the interventions of the intermittent server ? What is his mean working time during one stay in the back office ? Respectively, what is his mean working time during one passage in the front office? And finally, what is the cost of such an organization ?

The paper is organized as follows : in Section 2 we present a Markovian model of the investigated system while in the following section we exhibit the steady state probability distribution of the stochastic process and the expression of the mean number of customers (or mean response time) in terms of the different parameters. In Section 4, we conduct the determination of the expectation of the time spent by the second server in one passage in the back office and those of the expectation of one sojourn time at the front office. In the following section we propose a cost function providing an optimal solution as a tool to help a manager in charge of the economical decision. Some conclusions close the paper (Section 6).

2. HYPOTHESES AND MODEL

We consider that the two servers are equivalent and that the service times are independent and identically distributed random variables following an exponential distribution with rate μ. The first server affected to the front office stays available for serving the arriving customers.

When there are $(K-1)$ customers, if the server affected to the back office is not already serving in the front office, then this server leaves the back office at the instant of arrival of a new customer and starts serving it in the front office. Once it is in the front office, the second server stays there until it has no more customers to serve and re-integrates the back office.

We assume the customer arrival process is Poisson with rate λ.

Under these hypotheses, the stochastic process modeling the number of customers in the office is a continuous time Markov chain (CTMC) $\{X(t), t \geq 0\}$ ([2]). Its transition graph is given in Figure 1.

State $(i,0)$ (resp. $(i,1)$) denotes a state with i present customers and where the second server is in the back office (resp. present). State (0) refers to the empty system and, for $i \geq K$, state i designs the system when i customers and the second server are present. Note that the first server is idle in state $(1,1)$. In addition, E_0 (resp. E_1) will denote the subset of states where the second server is in the back office (resp. present) : $E_0 = \{(0),(1,0),...(K-1,0)\}$, $E_1 = \{(1,1),...(K-1,1),(K),(K+1),...\}$.

The steady state probability distribution of this CTMC is determined in the following section.

Note that the case $K=2$ corresponds to a $M/M/2$ queue with a little specificity : once the queue is empty, the first server deals with the new arrival, the second server arriving only when a new arrival finds the first server busy, and going back as soon as there is no more customer to serve in the front office. But from the customer point of view, this specificity does not affect the performance of the queue.

3. STEADY STATE PROBABILITY DISTRIBUTION, MEAN NUMBER OF CUSTOMERS

For any state e, π_e will denote the steady state probability of state e. Defining $\rho = \lambda/2\mu$, note that the steady state probability will exist if $\rho < 1$. Using the Chapmann-Kolmogorov equations (CK eqns) of states $(i,0)$, $i = 2,...,K-1$, it is not difficult to prove by induction the relation :

$$\pi_{K-i,0} = (\sum_{j=0}^{i-1} \phi^j)\pi_{K-1,0}, \qquad i = 2,...,K-1 \ . \tag{1}$$

Use of the cut theorem on the partition $\{E_0,\ E_1\}$ and of the steady state CK eqn of the state $(1,1)$ gives us

$$\pi_{1,1} = \frac{1}{(1+2\phi)}\pi_{K-1,0} \ . \tag{2}$$

Then, using using equations (2) and (1) and the CK eqn. for state (0), we can express probability in term of $\pi_{K-1,0}$, for the case $\phi \neq 1$:

$$\pi_{K-1,0} = \pi_0 \qquad \frac{(1+2\phi)(1-\phi)}{D_0} \ , \tag{3}$$

where $D_0 = \phi[(1-\phi) + (1+2\phi)(1-\phi^{K-1})]$. Consider now the CK eqns of states $(i,1)$, $i = 2,...,K-1$, we can prove by induction that :

$$\pi_{i,1} = \pi_{1,1}\frac{(1+\rho)-2\rho^i}{(1-\rho)} \qquad i = 2,...,K \ , \tag{4}$$

Then, using equations (2) and (3), we express the probability $\pi_K = \pi_{K,1}$ as a function of probability π_0 (for the case where $\phi \neq 1$) :

$$\pi_K = \pi_0 \ \frac{(1+\rho)-2\rho^K}{(1-\rho)} \ \frac{(1-\phi)}{D_0} \ . \tag{5}$$

Using the cut theorem, it is easy get the equations :

$$\pi_i \ = \ \rho^{i-K}\pi_K, \qquad i > K \ . \tag{6}$$

Thanks to the normalizing equation, it is finally possible to show that the probability π_0 can be written as :

$$\pi_0 = \qquad \frac{(1-\rho)(1-\phi)D_0}{D_1} \ , \tag{7}$$

where

$$\begin{aligned} D_1 \ = \ & (1-\rho)\{\phi(1-\phi)^2 + \\ & +(1+2\phi)[(K-1)(1-\phi) - \phi^2(1-\phi^{K-1})]\} + \\ & +(1-\phi)^2[1+(K-1)(1+\rho)] \ . \end{aligned}$$

The particular case $\phi = 1$ is easier to deal with and we get at the end :

$$\pi_0 = \frac{2(3K-2)}{3(K(K+3)-2)} \ .$$

When $K=2$, some of the equations obtained for the general case are no longer valid but it is not difficult to find again the well known result of the $M/M/2$ queue : $\pi_0 = (1-\rho)/(1+\rho)$.

The determination of the mean number of customers is purely technique. For $\phi \neq 1$, it satisfies the following relation :

$$\begin{aligned} E[N] \ = \ & \frac{(1-\phi)}{D_1}\Big\{(1-\rho)(1+2\phi)\Big(\frac{K(K+1)}{2} - \frac{K}{(1-\phi)} \\ & +\frac{\phi(1-\phi^K)}{(1-\phi)^2}\Big) + \\ & +(1-\phi)\left((1+\rho)\frac{K(K-1)}{2} + \frac{K+\rho(K-1)}{(1-\rho)}\right)\Big\} \ . \end{aligned}$$

When $K=2$, it is not difficult to find again the well known result of the $M/M/2$ queue : $\mathbb{E}[N] = 2\rho/(1-\rho^2)$.

For a given value of ρ we expect that the expected number of customers is greater than the value given by the $M/M/2$ queue. While, as long as ρ is lower than 0.5, the expected number of customers is lower than the ratio $\frac{2\rho}{1-2\rho}$, which corresponds to the value given by the $M/M/1$ queue with 2ρ as the utilization factor. The plotting of the expectation of the number of customers as a function of ρ, for different values of the integer K, would show that this expectation is increasing with ρ and with K. Note that without the second server, the mean number of customers would tend to infinity when ρ tends to 0.5.

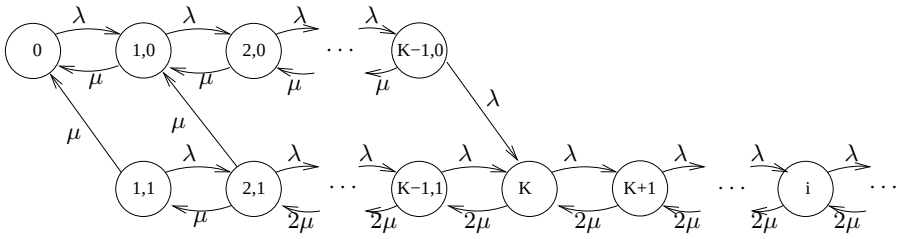

Figure 1: Transition graph of the CTMC**.**

4. PSEUDO-IDLE AND BUSY PERIODS FOR THE SECOND SERVER

By pseudo-idle period, for the second server, we mean a period of time during which this server is working in the back-office. We are interested by the expectation of such a period because we understand that too short of a period would have a negative effect on the productivity of the server. Such a period corresponds to a sojourn time of the CTMC in the subset E_0 and therefore we need to obtain the expectation of this sojourn time.

4.1 Mean Time of a Passage in the Back Office

First let us determine the probability that a pseudo-idle period starts in state (0) (respectively in state $(1,0)$). Given that the CTMC is in state $(2,1)$, if a service completes before a new arrival, the CTMC joins either state $(1,0)$ if the second server finishes his service first or state $(1,1)$ in the other case. These two events have equal probabilities (0.5 each). If the CTMC joins state $(1,1)$ from state $(2,1)$, this means that the permanent server becomes idle. Then either the second server becomes idle (with probability $\frac{\mu}{\lambda+\mu}$) or the regular server becomes busy again, the CTMC revisiting state $(2,1)$ (with probability $\frac{\lambda}{\lambda+\mu}$).

So, given a service completes when the CTMC is in state $(2,1)$, the CTMC goes to state $(1,0)$ with probability 0.5, goes to state (0) without coming back to state $(2,1)$ with probability $0.5 \times \frac{\mu}{\lambda+\mu}$ or comes back to state $(2,1)$ with probability $0.5 \times \frac{\lambda}{\lambda+\mu}$. Considering these three eventualities, we see that when the CTMC enters subset E_0, it enters it through state (0) with probability $\frac{0.5(\mu/(\lambda+\mu))}{0.5(1+\mu/(\lambda+\mu))}$ or enters it through state $(1,0)$ with probability $\frac{0.5}{0.5(1+\mu/(\lambda+\mu))}$. These two expressions reducing respectively to $\frac{\phi}{1+2\phi}$ and $\frac{1+\phi}{1+2\phi}$.

Let assume that $X(0)=0$. Let T_A be the sojourn time in the subset E_0 : $T_A = inf\{t|X(t)=K\}$. In order to express the expectation of T_A, we first consider the random variable T_i defined as the time it takes to the CTMC to reach state $(i+1,0)$ given $X(0)=(i,0)$. We also denote the expectation of T_i by α_i. Introducing the discrete random variable I_i such that, for $i \geq 0$, $I_i = 1$ (resp. $I_i = 0$) if the first transition of the CTMC from state $(i,0)$ is a jump to state $(i+1,0)$ (resp. $(i-1,0)$), we get when conditioning w.r.t. I_i : $\mathbb{E}[T_i|I_i=1] = \frac{1}{\lambda+\mu}$, and $\mathbb{E}[T_i|I_i=0] = \frac{1}{\lambda+\mu} + \alpha_{i-1} + \alpha_i$.

For $i=0$, we have immediately $\mathbb{E}[T_0] = \frac{1}{\lambda}$. Since the departure rate from state $(i,0)$ equals $(\lambda+\mu)$ while the transition rate from state $(i,0)$ to state $(i+1,0)$ equals λ, the probability that the first transition of the CTMC from state $(i,0)$ is a jump to state $(i+1,0)$ is $\mathbb{P}(I_i=1) = \frac{\lambda}{\lambda+\mu}$. Therefore, deconditioning the expectation $\alpha_i = \mathbb{E}[T_i]$ gives us, for $i>0$,

$$\alpha_i = \frac{1}{\lambda+\mu}\frac{\lambda}{\lambda+\mu} + \left(\frac{1}{\lambda+\mu} + \alpha_{i-1} + \alpha_i\right)\frac{\mu}{\lambda+\mu} ,$$

that reduces to $\alpha_i = \frac{1}{\lambda}(1+\mu\,\alpha_{i-1})$.

Since $\alpha_0 = \mathbb{E}[T_0] = \frac{1}{\lambda}$, we can compute successfully $\alpha_0, \alpha_1, \alpha_2, ...$

It is not difficult to prove that $\alpha_i = \frac{1}{\lambda}\sum_{j=0}^{i}\phi^j$.

In addition, $\mathbb{E}[T_A]$ depends on the way the CMTC enters the subset E_0 since $\mathbb{E}[T_A|X(0)=0] = \sum_{j=0}^{K-1}\alpha_j$, while $\mathbb{E}[T_A|X(0)=(1,0)] = \sum_{j=1}^{K-1}\alpha_j$. Therefore, after deconditioning we obtain $\mathbb{E}[T_A]$ that we scale by expressing this time expectation in term of a mean number of service times :

$$\mu\mathbb{E}[T_A] = \frac{\phi}{2(1+\rho)} + \phi\left((K-1) + \sum_{i=1}^{K-1}(K-i)\phi^i\right) . \quad (8)$$

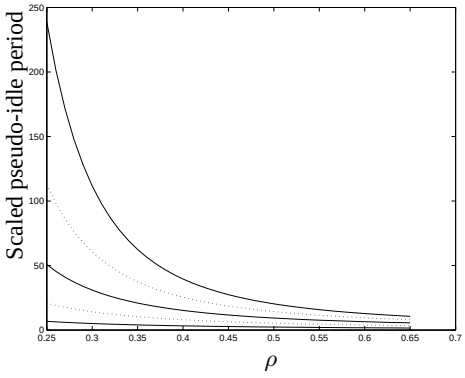

Figure 2: Scaled expectation of the pseudo-idle period of the second server as a function of ρ. For (bottom-up) $K = 2, 3, 4, 5$ and 6.

In Figure 2, we have plotted the scaled expectation of the pseudo-idle period of the second server as a function of ρ, for different values of the integer K. We can say that the expectation of the pseudo-idle period of the second server is important when ρ is between 0 and around 0.4, Remember that when $\rho = 0.4$, the utilization factor of the single server of the $M/M/1$ queue equals 0.8. As we would expect, this expectation is decreasing with ρ and increasing with K.

Note that if the manager decides to change the rule by switching from K to $(K+1)$, then the scaled expectation will be increased of the quantity $\Delta(\mu\mathbb{E}[T_A]) = \mu\mathbb{E}[T_A(K+1)] - \mu\mathbb{E}[T_A(K)]$ corresponding to $\phi\left(\sum_{i=0}^{K}\phi^i\right)$. Even in the case where $\phi = 1$ (*i.e.,* $\rho = 0.5$), this increase can be shown to correspond to $(K+1)$ mean service times!

4.2 Mean Time of a Passage in the Front Office

Now let $\mathbb{E}[T_P]$ be the expectation of a period spent in the front office by the second server. This server starts such a period with the frequency $\lambda\pi_{K-1,0}$. Using the fact that this frequency must be equal to $(\mathbb{E}[T_A] + \mathbb{E}[T_P])^{-1}$, we can obtain first the expression $\lambda\mathbb{E}[T_P]$ and

then the expression of the expectation scaled in term of the mean service time : $\mu\mathbb{E}[T_P] = \frac{1}{2(1-\rho)}\left((K-1)+\frac{1}{(1+\rho)}\right)$. Note that $\mu\mathbb{E}[T_P]$ represents also the expected number of customers served by the intermittent server during a passage in the front office.

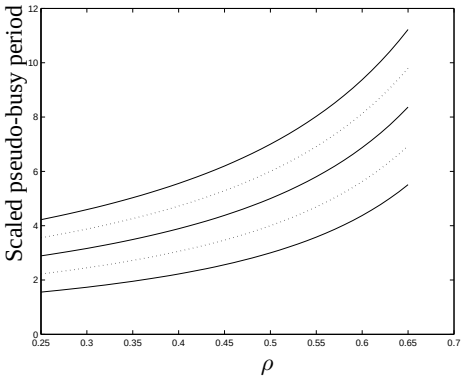

Figure 3: Scaled expectation of the pseudo-busy period of the second server as a function of ρ. For (bottom-up) $K = 2, 3, 4, 5$ and 6.

In Figure 3, we have plotted the scaled expectation of the pseudo-busy period of the second server as a function of ρ, for different values of the integer K. As we would expect, this expectation is increasing with ρ and with K. Moreover, we can say that the expectation of the pseudo-busy period of the second server is relatively small when ρ is between 0 and around 0.4, when we compare it with the one of the pseudo-idle period. From that we understand the interest of introducing the policy of the intermittent server.

5. COST FUNCTION

We have to consider two somewhat different situations. The first one is when the second server is not necessary for the system to be stable (*i.e.,* when ρ is lower than 0.5). The second situation is when the second server is necessary to the system ($\rho \geq 0.5$).

In the first situation, the second server just helps to decrease the mean response time seen by the customers. We have to compare this help with the perturbation of the work done in the back office.

We assume here that there is a fixed penalty c_0 to pay each time the second server has to leave the back office and that the cost per unit of time of this second server is c_1. We also assume that c_2 is the cost per unit of waiting time. Then the function to minimize corresponds to the total variable cost per time unit and is given by :

$$C(K) = c_0(T_A + T_P)^{-1} + c_1 S_1 + c_2 \mathbb{E}[N] \ ,$$

where $S_1 = \sum_{i=1}^{K-1} \pi_{i,1} + \sum_{i=K}^{\infty} \pi_i$. Note that the sum of probabilities S_1 is also the mean time per time unit spent by the second server in the front office and $\mathbb{E}[N]$ is also the total waiting time per time unit. When the variable K is increased, the first two terms are decreasing (asymptotically to zero) while the term $c_2\mathbb{E}[N]$ is increasing (from $c_2 2\rho/(1-\rho^2)$ when $K = 2$ to the asymptotic value $c_2 2\rho/(1-2\rho)$ when K tends to infinity). In this situation The optimal K may not be finite if the penalty coefficient c_2 is not large enough.

The second situation is different in the sense that K has to be finite in order to have a stable solution. In this case, the intermittent server has to work in the front office a percentage of time S_1 greater than $(\lambda/\mu - 1)$ such that the system has a steady state solution. The maximal feasible value $K_{\max}$ of K is given by $K_{\max} = \max\{K | S_1(K) > \lambda/\mu - 1\}$. Practically, if $K_{\max}$ is large enough (*i.e.,* when $(\lambda/\mu - 1)$ is not close to unity), the cost $c_2\mathbb{E}[N]$ should be large when $K = K_{\max}$ and we may expect the cost function to be convex. However, the convexity of $C(K)$ has not been investigated theoretically. Also, from a practical point of view, the parameter c_2 has again to be not too small with respect to c_0 and c_1 in order to avoid the limit behavior where the second server would come once a year to empty the waiting room. In Figure 4, we have plotted the values of

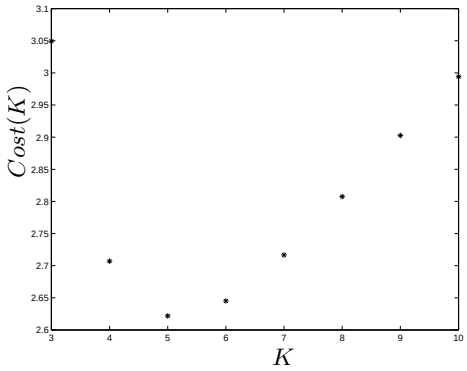

Figure 4: Variable cost function, with $\rho = 0.35$.

$C(K)$ when $\rho = 0.35$ (this means that the single server queue would have a utilization factor of 0.7), for $c_0 = 10$, $c_1 = 5$ and $c_2 = 1.5$. It can be seen that the minimum is achieved with $K = 5$. From Figures (2) and (3), we can check that for this optimal solution the mean pseudo-idle period of the second server is around 70 times the mean service time while the mean pseudo-busy period is close to 4 times the mean service time.

6. CONCLUSIONS

We have shown in this paper the importance of intermittent servers in order to reduce the response times without increasing significantly the idle times of servers. For such situations where a single server would satisfy the stability condition ($\lambda < \mu$), a non trivial result is that the pseudo-idle period of the second server is significantly longer than what would be generally expected by the management and also that the pseudo-busy period stays small ; and so the second server can keep his main activity in the back office.

We can think of applications in architectures for large telecommunication switches where we have "guard" processors to help the congested input queues on demand. Not only these results are interesting by themselves if such a situation occurs in a real situation but also, this study can be used to check simulation models used for a more complex situation.

7. REFERENCES

[1] J. D. Blackburn. Optimal control of queueing systems with intermittent service. Technical report, DTIC Document, 1971.

[2] E. Cinlar. *Introduction to stochastic Processes*. Prentice Hall, New-Jersey, 1975.

[3] T. Osogami, M. Harchol-Balter, and A. Scheller-Wolf. Analysis of cycle stealing with switching times and thresholds. *Performance Evaluation*, 61(4):347–369, 2005.

[4] M. Parlar and M. Sharafali. Dynamic allocation of airline check-in counters: a queueing optimization approach. *Management Science*, 54(8):1410–1424, 2008.

[5] M. Rubinovitch. The slow server problem: a queue with stalling. *Journal of Applied Probability*, 22(4):879–892, 1985.