Performance of a Single Server Queue Supported by an Intermittent Server

Raymond A. Marie

Performance of a Single Server Queue Supported by an Intermittent Server

Raymond A. MARIE

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

Abstract

This chapter concerns the situation of a queue with one regular single server supported by an additional intermittent server who, in order to decrease the mean response time, i) leaves the back office to join the first server when the number of customers reaches the threshold K, ii) leaves 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 cost function is proposed from which an optimal value of the threshold K is obtained.

Keywords: Performance Evaluation - Response Time - Markovian model - Intermittent Server - Single Server - Optimal Threshold - Case Study.

1 Introduction

Let us consider a single server queue where the server can be supported by a second one who **i**) leaves his current work to join the first server when the number of customers reaches a threshold K, **ii**) leaves the queuing system when he has no more customers to serve. A typical example of such a situation comes from the banking sector where the unique server from the front office is supported by a second server regularly assigned to the back office who joins the front office as soon as the number of customers reaches a given threshold (denoted here by the integer K). But such a situation could come from a more industrial area. The introduction of an intermittent server allows to decrease the expected waiting times of customers at a lower cost than affecting an extra permanent server. And the aim of this study is to determine the efficiency of such a policy.

Note that a closed situation 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 larger model is a good example to be used in a course on discrete event simulation as a practical exercise 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. Nevertheless, 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 in which the server is an intermittent one 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 [5]. 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. [4]). Note also the different model known as "the slow server problem" (see [6]) 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 sense 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 sense meaningless). This study is less general than the one cited above ([4]) but produces a closed form solution for the steady state probability distribution and for different metrics such as expected waiting times for customers or expectation of busy periods for the intermittent server. Our objective is to promote a better understanding of the benefits of such a strategy. In particular, we have to consider the trade-off between the help to the customer and the perturbation of the work in the back office. This is achieved thanks to a cost function providing an optimal value of the threshold K as a tool to help a manager in charge of the economical decision.

The paper is organized as follows: in Sect. 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 Sect. 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 introduce a cost function allowing us to provide an optimal threshold K^*. Finally, we conclude by summarizing the advantages of using an intermittent server (Sect. 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 assigned 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 him in the front office. Once he is in the front office, the second server stays there until he 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], [3], [7]). Its transition graph is given in Figure 1.

A couple $(i,0)$ (respectively $(i,1)$) denotes a state where i customers are present and where the second server is in the back office (respectively present). State (0) refers to the empty system and, for $i \geq K$, state i denotes 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 (respectively E_1) will denote the subset of states where the second server is in the back office (respectively present):

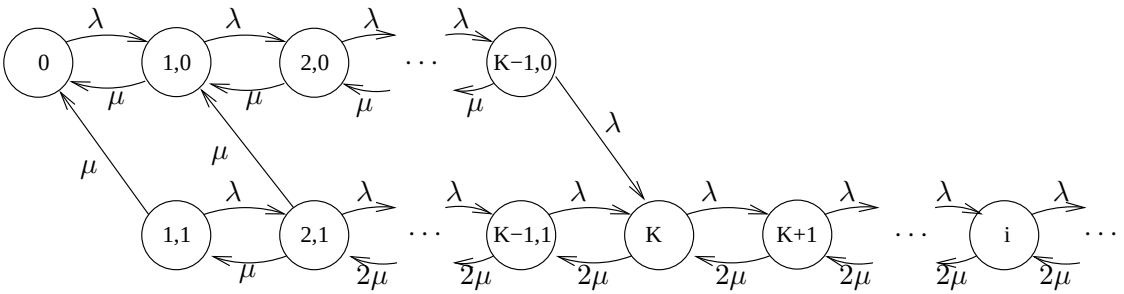

Figure 1: Transition graph of the CTMC.

$$E_0 = \{(0), (1,0), ... (K-1,0)\}, \quad E_1 = \{(1,1), ...(K-1,1), (K), (K+1), ...\} .$$

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

3.1 Steady State Probability Distribution

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

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

where $\phi = \mu/\lambda$. Use of the cut theorem on the partition $\{E_0,\ E_1\}$ and of the steady state C-K equation of state $(1, 1)$ gives us

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

Then, using equations (2) and (1) and the C-K equation for state (0), we can express probability $\pi_{K-1,0}$ in term of π_0 as :

$$\pi_{K-1,0} = \pi_0 \quad \left[\frac{\phi}{(1+2\phi)} + \phi \left(\sum_{j=0}^{K-2} \phi^j \right) \right]^{-1} , \tag{3}$$

or, for the case $\phi \neq 1$, as :

$$\pi_{K-1,0} = \pi_0 \quad \frac{(1+2\phi)(1-\phi)}{D_0} , \tag{4}$$

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

$$\pi_{i,1} = \pi_{1,1} \frac{(1+\rho) - 2\rho^i}{(1-\rho)} \quad i = 2, ..., K , \tag{5}$$

Since $\rho = 1$ is a root of the numerator, let us note that this probability can also expressed as :

$$\pi_{i,1} = \pi_{1,1} \left(1 + 2 \sum_{j=1}^{i-1} \rho^j \right) \qquad i = 2, ..., K \ , \tag{6}$$

When $i = K$, we get in particular the probability $\pi_{K,1}$ that we can rename π_K without any ambiguity :

$$\pi_K = \pi_{1,1} \frac{(1+\rho) - 2\rho^K}{(1-\rho)} = \pi_{1,1} \left(1 + 2 \sum_{j=1}^{K-1} \rho^j \right) \ . \tag{7}$$

Then, using equations (2) and (4), we express the probability π_K as a function of probability π_0 (again for the case $\phi \neq 1$) :

$$\pi_K = \pi_0 \ \frac{(1+\rho) - 2\rho^K}{(1-\rho)} \ \frac{(1-\phi)}{D_0} \ . \tag{8}$$

Considering the probabilities π_i, $i > K$, their expressions are easily obtained thanks to the use of the cut theorem :

$$\pi_i \quad = \quad \rho^{i-K} \pi_K, \qquad i > K \ . \tag{9}$$

Let us now consider the normalizing equation that we can write as :

$$S_0 + S_1 = 1 \ , \tag{10}$$

where $S_0 = \pi_0 + \sum_{i=1}^{K-1} \pi_{i,0}$ and $S_1 = \sum_{i=1}^{K-1} \pi_{i,1} + \sum_{i=K}^{\infty} \pi_i$.

Note that S_0 is the steady state probability that the intermittent server is working in the back office and that S_1 is the steady state probability that the intermittent server is working in the front office. This last sum S_1 will be also used latter when looking for the optimal threshold.

Using equations (1), (2), (4), (5), (8) and (9), we show in Sect. 7 that the probability π_0 can be written as :

$$\pi_0 = \qquad \frac{(1-\rho)(1-\phi)D_0}{D_1} \ , \tag{11}$$

where

$$\begin{aligned} D_1 \quad = \quad & (1-\rho)\{\phi(1-\phi)^2 + (1+2\phi)[(K-1)(1-\phi) - \phi^2(1-\phi^{K-1})]\} + \\ & + (1-\phi)^2[K + \rho(K-1)] \ . \end{aligned} \tag{12}$$

For the special case where $\phi = 1$, equations (1), (2), (4) and (5) reduce to:

$$\pi_{K-1,0} = \frac{3}{3K-2} \pi_0 \ , \qquad \text{and} \ \ \pi_{K-i,0} = i \pi_{K-1,0} \ , \qquad i = 2, ..., K-1 \ , \tag{13}$$

$$\pi_{1,1} = \frac{1}{3} \pi_{K-1,0} \ , \qquad \text{and} \ \ \pi_{i,1} = (3 - 2^{-(i-2)}) \pi_{1,1} \ , \qquad i = 2, ..., K \ , \tag{14}$$

while it is shown in Sect. 7 that probability π_0 satisfies :

$$\pi_0 = \frac{2(3K-2)}{3(K(K+3)-2)} \ . \tag{15}$$

For the case where $K = 2$, the transition graph of the CTMC is given on Figure 2. Some of the equations given for the general case become simpler (in particular because the expression D_0 equals $2\phi(1-\phi^2)$ when $K = 2$) and it is not difficult to find again the well known result of the $M/M/2$ queue :

$$\pi_0 = \frac{(1-\rho)}{(1+\rho)} \ . \tag{16}$$

Let us remark that for $\rho = 1/2$, we obtain $\pi_0 = 1/3$. In that case $\phi = 1$, and this result agrees with the one obtained thanks to relation (15) when $K = 2$.

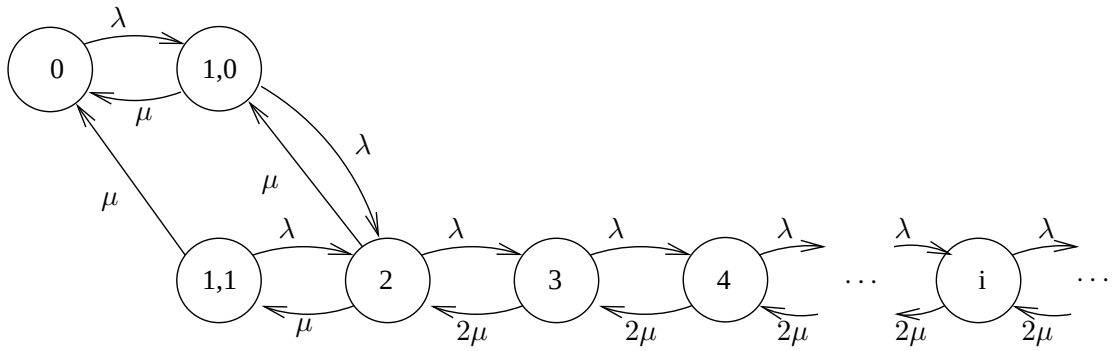

Figure 2: Transition graph of the CTMC when $K = 2$.

3.2 Mean Number of Customers, Mean Waiting Time

The determination of the mean number of customers $\mathbb{E}[N]$ is purely technique. For $\phi \neq 1$, it is shown in Sect. 8 that this expectation satisfies the following relation :

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

When $K = 2$, it is not difficult to find again the well known result of the $M/M/2$ queue :

$$\mathbb{E}[N] = 2\rho/(1-\rho^2) \ . \tag{17}$$

For the special case where $\phi = 1$, it is also shown in Sect. 8 that

$$\mathbb{E}[N] = \frac{K(K(K+3)+8)-4}{3(K(K+3)-2)} \ . \tag{18}$$

Note that for $K = 2$, $\mathbb{E}[N] = 4/3$. This result agrees with the one obtained thanks to relation (17) when $\rho = 1/2$, *i.e.*, ($\phi = 1$).

Because the aim of using an intermittent server is to decrease the waiting time of the customer in the front office, it is also interesting to consider the expected waiting time $\mathbb{E}[W]$. For that we first obtain the expected response time by use of the Little's formula and then subtract the mean service time :

$$\mathbb{E}[W] = \frac{1}{\lambda}\mathbb{E}[N] - \frac{1}{\mu} \ . \tag{19}$$

We may prefer to consider what we will call a "normalized" expected waiting time $\mathbb{E}[W_N]$ by taking the mean service time (*i.e.*, $1/\mu$) as the time unit. This gives us :

$$\mathbb{E}[W_N] = \mu\mathbb{E}[W] = \frac{\mu}{\lambda}\mathbb{E}[N] - 1 = \phi\mathbb{E}[N] - 1 \ .$$

Note that the "normalized" expected waiting time has no dimension and is therefore independent of the initial time unit.

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 1/2, the expected number of customers is lower than the ratio $\dfrac{2\rho}{1-2\rho}$, which corresponds to the value given by the $M/M/1$ queue with 2ρ as the utilization factor.

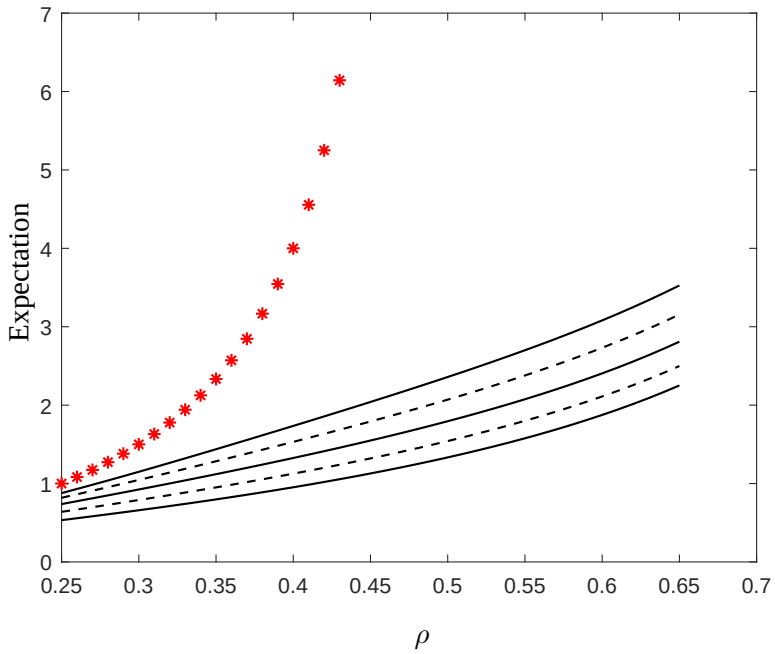

Figure 3: Mean number of customers as a function of ρ. For (bottom-up) $K = 2, 3, 4, 5$ and 6. Curve with stars corresponds to infinite K, the second server being never called.

In Figure 3, we have plotted the expectation of the number of customers 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. Note that without the second server, the mean number of customers would tend to infinity when ρ tends to $1/2$.

4 Pseudo-idle and Busy Periods of the Intermittent Server

The pseudo-idle period of the second server is defined as the 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 a too short 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 $\dfrac{\mu}{\lambda+\mu}$) or the regular server becomes busy again, the CTMC revisiting state $(2,1)$ (with probability $\dfrac{\lambda}{\lambda+\mu}$).

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 $0.5 \times \dfrac{\mu}{\lambda+\mu}$ or comes back to state $(2,1)$ with probability $0.5 \times \dfrac{\lambda}{\lambda+\mu}$. Considering these three eventualities, we see that when the CTMC enters subset E_0, it enters it through state (0) with probability $\dfrac{0.5(\mu/(\lambda+\mu))}{0.5(1+\mu/(\lambda+\mu))}$ or enters it through state $(1,0)$ with probability $\dfrac{0.5}{0.5(1+\mu/(\lambda+\mu))}$. These two expressions reducing respectively to $\dfrac{\phi}{1+2\phi}$ and $\dfrac{1+\phi}{1+2\phi}$.

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 $i \geq 0$:

$$I_i = \begin{cases} 1 & \text{if the first transition of the CTMC from state } (i,0) \\ & \text{is a jump to state } (i+1,0); \\ 0 & \text{if the first transition of the CTMC from state } (i,0) \\ & \text{is a jump to state } (i-1,0); \end{cases}$$

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 $\mathbb{E}[T_0] = \dfrac{1}{\lambda}$. 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 $\mathbb{P}(I_i=1) = \dfrac{\lambda}{\lambda+\mu}$. 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 $\alpha_0 = \mathbb{E}[T_0] = \dfrac{1}{\lambda}$, we can compute successfully α_0, α_1 ,α_2, ... It is not difficult to prove that $\alpha_i = \dfrac{1}{\lambda}\displaystyle\sum_{j=0}^{i} \phi^j$.

In addition, $\mathbb{E}[T_A]$ depends on the way the CMTC enters the subset E_0 since $\mathbb{E}[T_A|X(0)=0] = \displaystyle\sum_{j=0}^{K-1} \alpha_j$, while $\mathbb{E}[T_A|X(0)=(1,0)] = \displaystyle\sum_{j=1}^{K-1} \alpha_j$.

Therefore, after deconditioning we obtain :

$$\mathbb{E}[T_A] = \frac{1}{\lambda}\frac{1}{2(1+\rho)} + \frac{1}{\lambda}\left((K-1) + \sum_{i=1}^{K-1}(K-i)\phi^i\right) . \tag{20}$$

We can scale this result by expressing this time expectation in term of a number of mean service times :

$$\mu\mathbb{E}[T_A] = \frac{\phi}{2(1+\rho)} + \phi\left((K-1) + \sum_{i=1}^{K-1}(K-i)\phi^i\right) . \tag{21}$$

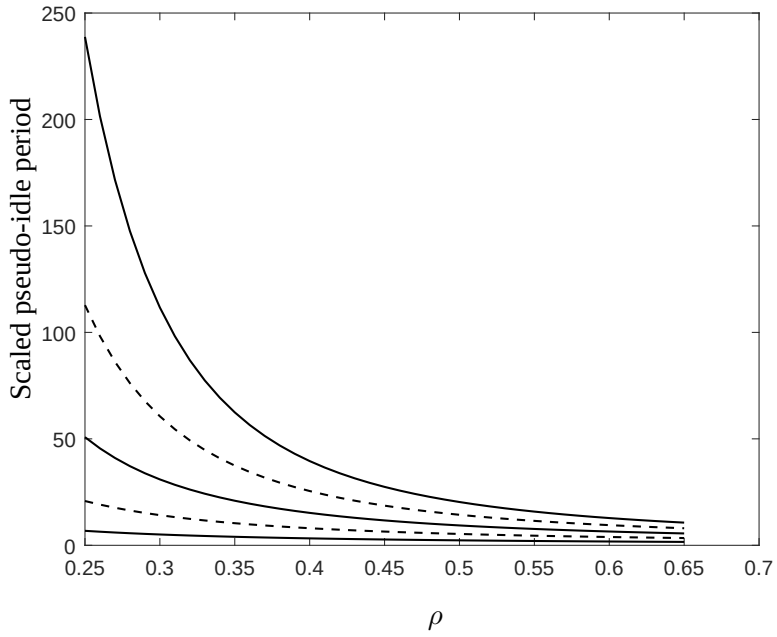

Figure 4: Scaled expectation of the pseudo-idle period of the second server as a function of ρ. For (bottom-up) $K = 2, 3, 4, 5$ and 6.

In Figure 4, 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 also 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:

$$\Delta_K(\mu\mathbb{E}[T_A]) = \mu\mathbb{E}[T_A(K+1)] - \mu\mathbb{E}[T_A(K)] = \phi\left(\sum_{i=0}^{K}\phi^i\right) .$$

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 intermittent 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 obtain a first expression for $\lambda\mathbb{E}[T_P]$:

$$\lambda\mathbb{E}[T_P] = [\pi_{K-1,0}]^{-1} - \lambda\mathbb{E}[T_A] .$$

Then, starting from equations (4) and (11) we express the inverse of probability $\pi_{K-1,0}$ as :

$$\begin{aligned}
[\pi_{K-1,0}]^{-1} &= \frac{D_1}{(1-\rho)(1+2\phi)(1-\phi)^2} , \\
&= \frac{\phi}{(1+2\phi)} + \frac{(K-1)(1-\phi) - \phi^2(1-\phi^{K-1})}{(1-\phi)^2} + \frac{[1+(K-1)(1+\rho)]}{(1-\rho)(1+2\phi)} .
\end{aligned}$$

Using equation (20) we develop the expression of $\lambda\mathbb{E}[T_A]$ as :

$$\begin{aligned}
\lambda\mathbb{E}[T_A] &= \frac{1}{2(1+\rho)} + \left((K-1) + \sum_{i=1}^{K-1}(K-i)\phi^i\right) \\
&= \frac{\phi}{(1+2\phi)} + \left((K-1) + K\sum_{i=1}^{K-1}\phi^i - \sum_{i=1}^{K-1} i\phi^i\right) \\
&= \frac{\phi}{(1+2\phi)} + \left((K-1) + K\left(\frac{1-\phi^K}{(1-\phi)} - 1\right) - \frac{(K-1)\phi^{K+1} - K\phi^K + \phi}{(1-\phi)^2}\right) \\
&= \frac{\phi}{(1+2\phi)} + \left((K-1) + \frac{K\phi(1-\phi^{K-1})}{(1-\phi)} - \frac{(K-1)\phi^{K+1} - K\phi^K + \phi}{(1-\phi)^2}\right) \\
&= \frac{\phi}{(1+2\phi)} + \frac{(K-1)(1-\phi)^2 + K\phi(1-\phi^{K-1})(1-\phi)}{(1-\phi)^2} - \frac{(K-1)\phi^{K+1} - K\phi^K + \phi}{(1-\phi)^2} \\
&= \frac{\phi}{(1+2\phi)} + \frac{(K-1) - (K-1)\phi - \phi^2 + \phi^{K+1}}{(1-\phi)^2} \\
&= \frac{\phi}{(1+2\phi)} + \frac{(K-1)(1-\phi) - \phi^2(1-\phi^{K-1})}{(1-\phi)^2} \,. \qquad (22)
\end{aligned}$$

Subtracting this last expression to the one obtained for $[\pi_{K-1,0}]^{-1}$ we get the expression of $\lambda\mathbb{E}[T_P]$:

$$\lambda\mathbb{E}[T_P] = \frac{[1+(K-1)(1+\rho)]}{(1-\rho)(1+2\phi)} = \frac{\rho}{(1-\rho)}\left((K-1) + \frac{1}{(1+\rho)}\right) \,, \qquad (23)$$

and then the expression of the expectation scaled in term of a number of mean service time :

$$\mu\mathbb{E}[T_P] = \frac{1}{2(1-\rho)}\left((K-1) + \frac{1}{(1+\rho)}\right) \,. \qquad (24)$$

Note that $\mu\mathbb{E}[T_P]$ represents also the expected number of custumers served by the intermittent server during a passage in the front office.

In Figure 5, 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 (*cf.* Figure 4). This shows the benefit of the intermittent server since the use of a low percentage of his time significantly decreases the mean waiting time.

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 $\rho < 0.5$). 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 waiting time $\mathbb{E}[W]$ seen by the customers. We have to compare this help to the customers with respect to 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. During a unit time, the expectation of the cumulative value of the waiting times equals $\lambda\mathbb{E}[W]$; this expectation being nothing else than the expectation of the number of waiting

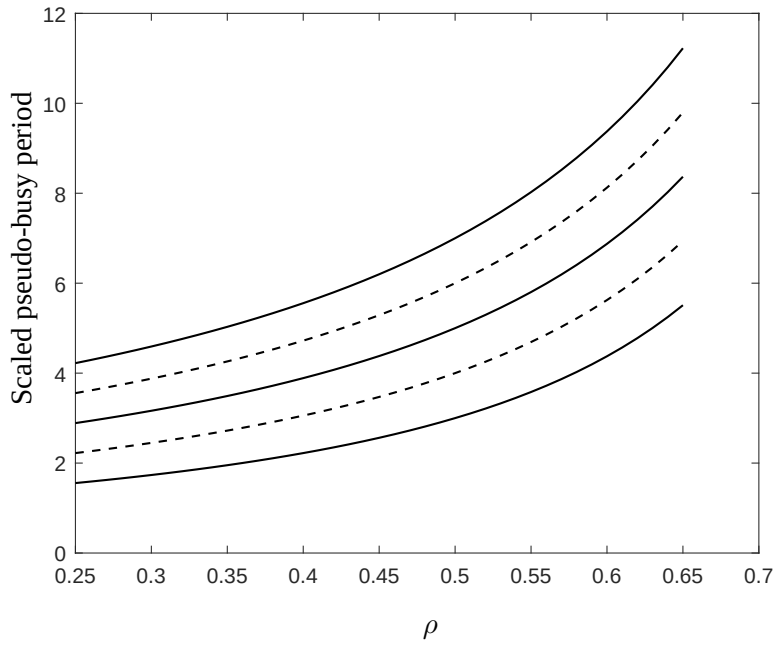

Figure 5: Scaled expectation of the pseudo-busy period of the second server as a function of ρ. For (bottom-up) $K = 2, 3, 4, 5$ and 6.

customers in the queue. Let $\mathbb{E}[N_w]$ denotes this expectation. The expression of $\mathbb{E}[N_w]$ is deduced from eqn (19) :

$$\mathbb{E}[N_w] = \mathbb{E}[N] - 2\rho \ . \tag{25}$$

Then, depending on the value K, the function to minimize corresponds to the expected total variable cost per time unit, and is given by :

$$C(K) = c_0[\mathbb{E}[T_A] + \mathbb{E}[T_P]]^{-1} + c_1 S_1 + c_2 \mathbb{E}[N_w] \ , \tag{26}$$

where here also, S_1 denotes the sum $\displaystyle\sum_{i=1}^{K-1} \pi_{i,1} + \sum_{i=K}^{\infty} \pi_i$. Note that this sum of probabilities S_1 is nothing but the mean time per time unit spent by the second server in the front office.

When the variable K is increased, the first two terms are decreasing while the term $c_2\mathbb{E}[N_w]$ is increasing. More precisely, considering a cycle of the intermittent server, we start from the relation :

$$S_1 = \frac{\mathbb{E}[T_P]}{(\mathbb{E}[T_A] + \mathbb{E}[T_P])} = \frac{1}{1 + \frac{\mathbb{E}[T_A]}{\mathbb{E}[T_P]}} \ . \tag{27}$$

Considering equations (22) and (23) we deduce that, when K tends to infinity, the two expectations tend to infinity. Considering now the ratio $\dfrac{\mathbb{E}[T_A]}{\mathbb{E}[T_P]}$ when K tends to infinity, since ϕ satisfies $\phi > 1$, the limit of this ratio is the same as the limit of the following ratio :

$$\lim_{K \longrightarrow +\infty} \frac{\mathbb{E}[T_A]}{\mathbb{E}[T_P]} = \lim_{K \longrightarrow +\infty} \frac{(2\phi - 1)}{(1-\phi)^2} \frac{\phi^K}{K} = +\infty \ . \tag{28}$$

Therefore, the first two terms of the cost function tends asymptotically to zero when K tends to infinity while the term $c_2\mathbb{E}[N_w]$ is increasing (from $c_2 2\rho^3/(1-\rho^2)$ when $K = 2$ to the asymptotic value

$c_2 4\rho^2/(1-2\rho)$ 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 $(\lambda/\mu - 1)$ in order that the system admits 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 $c_2\mathbb{E}[N_w]$ 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.

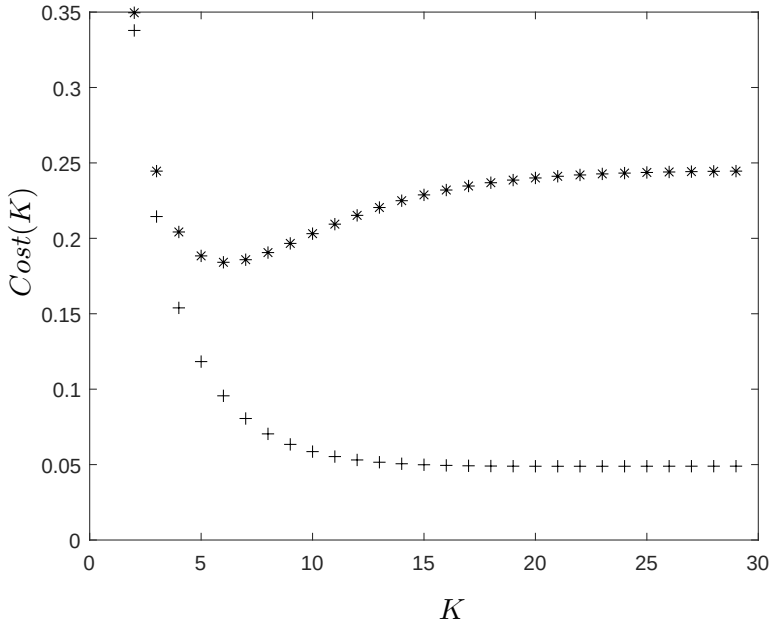

Figure 6: Variable cost function, with $\rho = 0.35$, $c_0 = 0.5$, $c_1 = 1$. Case 1 (stars): $c_2 = 0.15$. Case 2 (sign +): $c_2 = 0.03$.

In Figure 6, we have plotted two sets of values of $C(K)$ when $\rho = 0.35$ (alone the permanent server queue would have a utilization factor of 0.7), for $c_0 = 0.5$, $c_1 = 1$. Case $c_2 = 0.15$ (noted with stars) gives an optimal $K^* = 5$. From Figures (4) and (5), 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 5 times the mean service time. But case $c_2 = 0.03$ (noted with sign +) gives a decreasing cost for $K \in [2, 30]$.

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 quite large telecommunication switches where we have ”guard” processors to help the congested input queues on demand. It may also help in the context of network function virtualization (NFV) in which a service might be deployed on demand to face a

transient congestion. 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 Appendix 1: Determination of eqn. (12)

Starting from the normalizing equation :

$$S_0 + S_1 = 1 \ , \tag{29}$$

where $S_0 = \pi_0 + \sum_{i=1}^{K-1} \pi_{i,0}$, and $S_1 = \pi_0 + \sum_{i=1}^{K-1} \pi_{i,1} + \sum_{i=K}^{\infty} \pi_i$,

we first consider the partial sum S_0 :

$$\begin{aligned} S_0 &= \pi_0 + \sum_{i=1}^{K-1} \pi_{i,0} = \pi_0 + \sum_{i=1}^{K-1} \pi_{K-i,0} \ , \\ &= \pi_0 + \pi_{K-1,0} \sum_{i=1}^{K-1} (\sum_{j=0}^{i-1} \phi^j) = \pi_0 + \pi_{K-1,0} \sum_{i=1}^{K-1} (K-i)\phi^{i-1} \ , \\ &= \pi_0 + \pi_{K-1,0} \left(K \sum_{i=1}^{K-1} \phi^{i-1} - \sum_{i=1}^{K-1} i\phi^{i-1} \right) \ , \end{aligned}$$

or (if $\phi \neq 1$) :

$$\begin{aligned} S_0 &= \pi_0 + \pi_{K-1,0} \left(K \frac{(1-\phi^{K-1})}{(1-\phi)} - \frac{(1 - K\phi^{K-1} + (K-1)\phi^K)}{(1-\phi)^2} \right) \ , \\ &= \pi_0 + \frac{\pi_{K-1,0}}{(1-\phi)} \ \frac{K(1-\phi) - (1-\phi^K)}{(1-\phi)} \ , \\ &= \pi_0 + \pi_0 \frac{(1+2\phi)}{D_0} \ \frac{K(1-\phi) - (1-\phi^K)}{(1-\phi)} = \pi_0 \left(1 + \frac{(1+2\phi)(K(1-\phi) - (1-\phi^K))}{(1-\phi)D_0} \right) \ , \\ &= \frac{\pi_0}{(1-\phi)D_0} \left(\phi(1-\phi)^2 + (1+2\phi)[(K-1)(1-\phi) - \phi^2(1-\phi^{K-1})] \right) \ . \end{aligned}$$

Considering now the partial sum S_1, *i.e.*, the steady state probability that the back-office server is helping the front-office server, we have :

$$\begin{aligned} S_1 &= \sum_{i=1}^{K-1} \pi_{i,1} + \sum_{i=K}^{\infty} \pi_i \ , \\ &= \pi_{1,1} \sum_{i=1}^{K-1} \frac{(1+\rho) - 2\rho^i}{(1-\rho)} + \pi_K \sum_{i=K}^{\infty} \rho^{i-K} \ , \\ &= \pi_{1,1} \frac{(K-1)(1+\rho)}{(1-\rho)} - 2\pi_{1,1} \frac{1}{(1-\rho)} \sum_{i=1}^{K-1} \rho^i + \pi_K \frac{1}{(1-\rho)} \ , \\ &= \pi_{1,1} \frac{(K-1)(1+\rho)}{(1-\rho)} - \pi_{1,1} \frac{2}{(1-\rho)} \frac{(1-\rho^K)}{(1-\rho)} + \pi_{1,1} \frac{(1+\rho) - 2\rho^K}{(1-\rho)} \frac{1}{(1-\rho)} \ , \\ &= \pi_{1,1} \frac{(K-1)(1+\rho)}{(1-\rho)} - \pi_{1,1} \frac{1}{(1-\rho)} = \pi_{1,1} \frac{1 + (K-1)(1+\rho)}{(1-\rho)} \ , \\ &= \pi_0 \quad \frac{(1-\phi)[K + \rho(K-1)]}{(1-\rho)D_0} \ . \end{aligned}$$

Using the normalizing equation, *i.e.*, $S_0 + S_1 = 1$, we get the expression of probability π_0 when $\phi \neq 1$:

$$\pi_0 = \quad \frac{(1-\rho)(1-\phi)D_0}{D_1} \ , \tag{30}$$

where

$$D_1 = (1-\rho)\{\phi(1-\phi)^2 + (1+2\phi)[(K-1)(1-\phi) - \phi^2(1-\phi^{K-1})]\} + \\ +(1-\phi)^2[K+\rho(K-1)] \, .$$

For the special case where $\phi = 1$, it is not difficult, starting from the specific relations between probabilities given at the end of Section 3.1, to find the following expressions :

$$S_0 = \frac{3K(K+1)-4}{2(3K-2)}\pi_0 \, , \quad S_1 = \frac{(3K-1)}{(3K-2)}\pi_0 \, , \quad \pi_0 = \frac{2(3K-2)}{3(K(K+3)-2)} \, . \tag{31}$$

8 Appendix 2: Determination of mean number of customers

In order to obtain the expression, let us start by computing two partial sums (B_0 and B_1), under the condition $\phi \neq 1$:

$$\begin{aligned}
B_0 &= \sum_{i=1}^{K-1} i\pi_{i,0} = \sum_{i=1}^{K-1}(K-i)\pi_{K-i,0} = \pi_{K-1,0}\sum_{i=1}^{K-1}(K-i)(\sum_{j=0}^{i-1}\phi^j) \, , \\
&= \pi_{K-1,0}\sum_{i=1}^{K-1}(K-i)\frac{(1-\phi^i)}{(1-\phi)} = \frac{\pi_{K-1,0}}{(1-\phi)}\left(\sum_{i=1}^{K-1} i - K\sum_{i=0}^{K-1}\phi^i + K + \phi\sum_{i=1}^{K-1} i\phi^{i-1}\right) \, , \\
&= \frac{\pi_{K-1,0}}{(1-\phi)}\left(\frac{K(K+1)}{2} - K\sum_{i=0}^{K-1}\phi^i + \phi\sum_{i=1}^{K-1} i\phi^{i-1}\right) \, , \\
&= \frac{\pi_{K-1,0}}{(1-\phi)}\left(\frac{K(K+1)}{2} - \frac{K}{(1-\phi)} + \frac{\phi(1-\phi^K)}{(1-\phi)^2}\right) \, , \\
&= \pi_0 \quad \frac{(1+2\phi)}{D_0}\left(\frac{K(K+1)}{2} - \frac{K}{(1-\phi)} + \frac{\phi(1-\phi^K)}{(1-\phi)^2}\right) \, , \\
&= \frac{(1-\rho)(1-\phi)(1+2\phi)}{D_1}\left(\frac{K(K+1)}{2} - \frac{K}{(1-\phi)} + \frac{\phi(1-\phi^K)}{(1-\phi)^2}\right) \, ,
\end{aligned}$$

and secondly :

$$\begin{aligned}
B_1 &= \sum_{i=1}^{K-1} i\pi_{i,1} + \sum_{i=K}^{\infty} i\pi_i = \sum_{i=1}^{K-1} i\pi_{i,1} + \pi_K\sum_{i=K}^{\infty} i\rho^{i-K} \, , \\
&= \pi_{1,1}\left(\sum_{i=1}^{K-1} i\frac{(1+\rho)}{(1-\rho)} - \sum_{i=1}^{K-1}\frac{2i\rho^i}{(1-\rho)}\right) + \pi_{1,1}\frac{(1+\rho)-2\rho^K}{(1-\rho)}\sum_{i=K}^{\infty} i\rho^{i-K} \, , \\
&= \pi_{1,1}\frac{(1+\rho)}{(1-\rho)}\sum_{i=1}^{K-1} i - \frac{2\pi_{1,1}}{(1-\rho)}\sum_{i=1}^{\infty} i\rho^i + \pi_{1,1}\frac{(1+\rho)}{(1-\rho)}\sum_{i=0}^{\infty}(K+i)\rho^i \, , \\
&= \frac{\pi_{1,1}}{(1-\rho)}\left((1+\rho)\frac{K(K-1)}{2} - \frac{\rho}{(1-\rho)} + K\frac{(1+\rho)}{(1-\rho)}\right) \, , \\
&= \pi_0 \quad \frac{(1-\phi)}{(1-\rho)D_0}\left((1+\rho)\frac{K(K-1)}{2} + \frac{K+\rho(K-1)}{(1-\rho)}\right) \\
&= \frac{(1-\phi)^2}{D_1}\left((1+\rho)\frac{K(K-1)}{2} + \frac{K+\rho(K-1)}{(1-\rho)}\right) \, .
\end{aligned}$$

From that we get the expression of the expectation of the number of customers :

$$\begin{aligned}
\mathbb{E}[N] &= \sum_{i=1}^{K-1} i\pi_{i,0} + \sum_{i=1}^{K-1} i\pi_{i,1} + \sum_{i=K}^{\infty} i\pi_i \ , \\
&= \frac{(1-\rho)(1-\phi)(1+2\phi)}{D_1}\left(\frac{K(K+1)}{2} - \frac{(K)}{(1-\phi)} + \frac{\phi(1-\phi^K)}{(1-\phi)^2}\right) + \\
&+ \frac{(1-\phi)^2}{D_1}\left((1+\rho)\frac{K(K-1)}{2} + \frac{K+\rho(K-1)}{(1-\rho)}\right) \ , \\
&= \frac{(1-\phi)}{D_1}\left\{(1-\rho)(1+2\phi)\left(\frac{K(K+1)}{2} - \frac{K}{(1-\phi)} + \frac{\phi(1-\phi^K)}{(1-\phi)^2}\right) + \right. \\
&\left. +(1-\phi)\left((1+\rho)\frac{K(K-1)}{2} + \frac{K+\rho(K-1)}{(1-\rho)}\right)\right\} \ .
\end{aligned}$$

This last result corresponds to the expression presented in Section 3.2.

In the special situation where $\phi = 1$, let us first consider the sum B_0. Starting from the equality obtained above

$$B_0 = \pi_{K-1,0} \sum_{i=1}^{K-1} (K-i)(\sum_{j=0}^{i-1} \phi^j),$$

we get :

$$\begin{aligned}
B_0 &= \pi_{K-1,0} \sum_{i=1}^{K-1} (K-i)i \ , \\
&= \pi_{K-1,0}\left(K\sum_{i=1}^{K-1} i - \sum_{i=1}^{K-1} i^2\right) = \pi_{K-1,0}\left(K\frac{K(K+1)}{2} - \frac{(K-1)K(2K-1)}{6}\right) \ , \\
&= \pi_{K-1,0}\left(\frac{(K-1)K(K+1)}{6}\right) = \pi_0 \frac{3}{3K-2}\frac{(K-1)K(K+1)}{6} \ , \\
&= \pi_0 \frac{(K-1)K(K+1)}{2(3K-2)}.
\end{aligned}$$

Let us now consider the sum B_1. We may start from the following equality obtained above

$$B_1 = \frac{\pi_{1,1}}{(1-\rho)}\left((1+\rho)\frac{K(K-1)}{2} - \frac{\rho}{(1-\rho)} + K\frac{(1+\rho)}{(1-\rho)}\right) \ ,$$

and since here $\rho = 1/2$, we get :

$$B_1 = \frac{\pi_{1,1}}{2}\left(3K(K+3) - 4\right) = \frac{\pi_0}{2(3K-2)}\left(3K(K+3) - 4\right) \ .$$

After summation of B_0 and B_1 and use of the expression of π_0 given by relation 31, we are able to exhibit the following expression :

$$\mathbb{E}[N] = \frac{K(K(K+3)+8) - 4}{(K(K+3) - 2)} \ .$$

References

[1] Joseph 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] Peter G. Harrison and Naresh M. Patel. *Performance Modelling of Communication Networks and Computer Architecture*. Addison-Wesley, Reading, Mass., 1993.

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

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

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

[7] R. A. Sahner, K. S. Trivedi, and A. Puliafito. *Performance and Reliability Analysis of Computer Systems: An Example-Based Approach Using the SHARPE Software Package*. Kluwer Academic Publishers, 1996.