Industrial Engineering
& Management Systems
Vol 15, No 2, June 2016, pp.123-130
ISSN 1598-7248 | EISSN 2234-6473 |
http://dx.doi.org/10.7232/iems.2016.15.2.123

Control of $G/M^X/1$ Queueing System with *N*-Policy and Customer Impatience

Si-Yeong Lim, Sun Hur*
Department of Industrial and Management Engineering, Hanyang University, Ansan, Korea

(Received: November 9, 2015 / Revised: April 8, 2016 / Accepted: April 24, 2016)

ABSTRACT

We introduce a queueing system with general arrival stream and exponential service time under the *N*-policy, where customers may renege during idle period and arrival rates may vary according to the server's status. Probability distributions of the lengths of idle period and busy period are derived using absorbing Markov chain approach and a method to obtain the optimal control policy that minimizes long-run expected operating cost per unit time is provided. Numerical analysis is done to illustrate and characterize the method.

Keywords: Impatience, Control Policy, N-Policy, General Arrival

* Corresponding Author, E-mail: hursun@hanyang.ac.kr

1. INTRODUCTION

We consider a queueing system with a control policy under steady-state conditions. Especially, whenever the system is empty, the server becomes idle and resumes service again when there are *N* customers waiting in the system. This is called *N*-policy, where *N* is the threshold from which the server begins its service. Services are provided exhaustively, that is, continue until there remains no customer in the system.

Numerous papers have dealt with queueing models adopting *N*-policy since Yadin and Naor (1963). For studies on *N*-policy and related work, readers are advised to see Takagi (1991). While most of the queueing models with control policies studied adopt Poisson input, research on the general input queueing models with control policy is relatively scarce. Ke (2003) and Zhang and Tian (2004) introduce *N*-policy in the G/M/1 queueing systems to obtain the steady-state probability distributions of the number of customers in the system. Lee and Ahn (2002) and Lee and Park (2004) studied MAP/G/1 and BMAP/G/1 queues under *N*-policy, but optimal policies are not discussed due to the complexity of the performance measures. Chae and Lee (2005) consider GI/M/1 queues with generalized vacations including *N*-policy to obtain stationary probabilities of the embedded Markov chains by an absorbing Markov chain approach.

There is growing interest in the analysis of queueing systems with impatient customers due to their potential application in communication systems, call centers, production-inventory systems and many other related areas (see Akcan (2013), e.g., for the application to the inventory control system). As for the instance of queueing systems with impatient customers, see Benjaafar *et al*. (2010). Blackburn (1972) considers an M/G/1 queueing system with customer balking and/or reneging and gives a finite algorithm to compute the stationary optimal policy that maximizes the expected discounted reward under some conditions on the system parameters and unit costs. Altman and Yechiali (2006) presents a comprehensive analysis of the M/M/c and M/G/1 queues, where customers' impatience is due to the absence of servers upon arrival for both the multiple and the single-vacation cases, and obtain various closed-form results. More recently, Yue *et al*. (2011) consider a two-phase queueing system with impatient customers and multiple vacations with Poisson arrivals. They derive the closed-form expressions for various performance measures including the mean system sizes for various states of the server, the average rate of balking, the average rate of

reneging, and the average rate of loss.

Swensen (1986) utilizes the remaining workload to derive the asymptotic distributions of the actual waiting time and the virtual waiting time of G/M/c queue with impatient customers when the interarrival times follow the Coxian distribution. Choi *et al.* (2004) computes many performance measures in MAP/M/c queue with impatience. Bae and Kim (2010) derive the stationary distribution of the workload of the server of G/M/1 queue with constant patience time by the level crossing argument. Mandelbaum and Momčilović (2012) consider G/GI/N+GI queueing systems with impatience in the quality-and efficiency-domain (QED) regime. Diffusion approximation is used here to derive queue length and virtual waiting time processes.

Tadj and Choudhury (2005) perform extensive surveys on the optimal design and control of queueing systems with control policies. They describe the different kinds of threshold (policy) models available in the literature, especially of *N*-, *T*-, and *D*-policies along with various combined policies, but no queueing model with impatient customers under *N*-policy has been found in their review. Other assumptions, such as server setup, nonlinear cost structure, finite system capacity, server vacation, priority among customers, and so forth, are introduced in many literatures combined with *N*-policy and analyzed. To the best of our knowledge, however, there is no literature considering customers' impatience for general input and exponential server queueing system with *N*-policy. Recently, Chae and Kim (2007) and Kim and Yang (2011) derive the probability distribution of busy period of G/M/1 queues under server vacations and, especially, Chae and Lim (2008) introduces *N*-policy. Their results, however, are not computationally tractable even for the moderate value of *N*, because one should derive the *N*th derivative of a very complex formula to obtain the probability distribution of busy period.

In this paper, we introduce a queueing system with general arrival stream and exponential service time under the *N*-policy, where customers may renege during idle period and arrival rates may vary according to the server's status. In contrast to the previous study, our method utilizes a sample-path approach and is computationally applicable for a large value of *N*. Our results are useful for, e.g., the production-inventory system that handles perishable goods and the external orders arrive at the system in batches whose sizes are random. Because today's production facility is mostly involved in multiple tasks, it produces specific products with its full capacity and stores them in the inventory until the amount of inventory reach a predetermined level. In this case, the proper level of inventory from which the facility can share its production capability with other goods should be determined to minimize the overall cost. Our approach can be applied to this kind of make-to-order type production-inventory system.

Probability distributions of the lengths of idle and busy periods are derived by means of absorbing Markov chain approach and a method to obtain the optimal control policy that minimizes long-run expected operating cost per unit time is provided. Numerical examples are given to illustrate the method.

The paper is organized as follows: Section 2 introduces the model considered in the paper and notations are explained. Sections 3 and 4 derive the probability distributions and expected values of the idle and busy periods, respectively, on which the computation of the operating cost function in Section 5 is based. Section 6 provides numerical examples and Section 7 concludes the paper.

2. MODEL AND NOTATIONS

The following notations are used throughout the paper.

N:	threshold level, $N \geq 2$,
I, B:	random variables of lengths of idle and busy periods, respectively,
A_I, $A_I(\cdot)$:	random variable, distribution function of interarrival time during idle period, respectively
A_B, $A_B(\cdot)$:	random variable, distribution function of interarrival time during busy period, respectively
λ_I, λ_B:	arrival rates in idle and busy periods, respectively, where $\lambda_I = 1/E(A_I)$, $\lambda_B = 1/E(A_B)$,
$A_I^*(\theta)$, $A_B^*(\theta)$:	Laplace transforms of $A_I(\cdot)$ and $A_B(\cdot)$, respectively,
η:	rate of reneging,
S:	random variable of service time to a batch of customers,
μ:	service rate, where $\mu = 1/E(S)$
G:	random variable of the number of customers in each batch,
g_i:	probability that the batch size is i, $i = 1, 2, \cdots$
τ_n^-:	epoch of right before the n^{th} arrival, $n = 1, 2, \cdots$,
$N_I(t)$, $N_B(t)$:	numbers of customers in the system at time t during the idle and busy periods, respectively,
C_s:	setup cost incurred when the server initiates service,
C_h:	holding cost per customer per unit time in the system.

We consider a G/M^X/1 queueing system with a removable server under steady-state condition. When the system becomes empty, the server remains idle until *N* customers are waiting in the system (*N*-policy). Once the number of customers waiting in the system reaches to *N*, the server starts providing service exhaustively and server's idle period and server's busy period repeat over and over again. Customers, however, who have arrived during server idle period and are waiting for the server to start service, are impatient and renege. As a result,

they may leave the system without being served before the service commences. Therefore, the actual number of customers arrived during idle period may be greater than *N*. Once the server's busy period begins, however, no reneging is occurred. We assume $N \geq 2$ for simplicity but a slight modification can accommodate the case $N = 1$, as discussed in Section 4.

Customers' arrival tendency may become different according to the server status. For example, customers are reluctant to join the queue if the server is currently idle, while they tend to come faster to the system once the server starts service. Therefore, we assume the arrival rates are dependent on the server's state in this paper.

During the idle period, the inter-arrival time of customers, A_I, is generally distributed with distribution function $A_I(\cdot)$. Since the customers arrived during this period may be impatient and reneging, we assume the time between two consecutive reneging is exponentially distributed with mean $1/\eta$ and only one reneging occurs at a time. Subscript *I* is replaced with *B* in case of busy period, that is, A_B, $A_B(\cdot)$, and $\lambda_B = 1/E(A_B)$. Services are provided in batches whose size is random and service times given to each batch are exponentially distributed. Numbers of customers included in each batch are independent of each other with common probability distribution, $g_i \equiv \Pr(G = i)$, $i = 1, 2, \cdots$. Especially, we assume that at the moment of the end of busy period the whole system restarts, i.e., it forgets the elapsed time since the previous arrival and the next customer in idle period will arrive after the random time having distribution $A_I(\cdot)$. In the following two subsections, we derive the probability distributions of the lengths *I* of idle period and *B* of busy period when the threshold is *N*.

3. DERIVATION OF THE LENGTH OF IDLE PERIOD

The random process $\{N_I(\tau_n^-), n = 1, 2, \cdots\}$ with state space $\{0, 1, 2, \cdots, N\}$ is a discrete-time absorbing Markov chain because

$$N_I(\tau_{n+1}^-) = N_I(\tau_n^-) + 1 - D_{I,n}, \tag{1}$$

w here $D_{I,n}$ is total number of reneging customers during $A_{I,n} \equiv \tau_{n+1}^- - \tau_n^-$, the time between n^{th} and $(n+1)^{th}$ arrivals, and $D_{I,n} \leq N_I(\tau_n^-) + 1$, $N_I(\tau_n^-) \geq 0$. We use hereafter the generic random variables D_I and A_I instead of $D_{I,n}$ and $A_{I,n}$, $n = 1, 2, \cdots$. The probability generating function (pgf) $D_I(z)$ of D_I is given by

$$D_I(z) = \sum_{m=0}^{\infty} z^m \int_0^{\infty} \frac{(\eta x)^m e^{-\eta x}}{m!} dA_I(x) = A_I^*(\eta - \eta z), \tag{2}$$

where $A_I^*(\theta)$ is the Laplace transform (LT) of $A_I(\cdot)$. Then the probability distribution d_k of D_I is given by:

$$d_k \equiv P[D_I = k] = \frac{1}{k!} \frac{d^k}{dz^k} D_I(z)|_{z=0} \tag{3}$$

Now, we can build one-step transition probability matrix $\mathbf{P}_I$ of the absorbing discrete-time Markov chain $\{N_I(\tau_n^-), n = 1, 2, \cdots\}$, where $0, 1, 2, \cdots, N-1$ are transient states and *N* is the absorbing state, as following:

$$\mathbf{P}_I = \begin{bmatrix} \mathbf{B}_I & \mathbf{B}_I^0 \\ \mathbf{0} & 1 \end{bmatrix}, \quad \mathbf{B}_I = \begin{bmatrix} 1 - \sum_{i=0}^{0} d_i & d_0 & 0 & 0 & \cdots & 0 \\ 1 - \sum_{i=0}^{1} d_i & d_1 & d_0 & 0 & \cdots & 0 \\ 1 - \sum_{i=0}^{2} d_i & d_2 & d_1 & d_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 - \sum_{i=0}^{N-1} d_i & d_{N-1} & d_{N-2} & d_{N-3} & \cdots & d_1 \end{bmatrix},$$

$$\mathbf{B}_I^0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ d_0 \end{bmatrix}^T. \tag{4}$$

$\mathbf{B}_I$ is $N \times N$ transition probability matrix of the transitions among transient states, whereas $\mathbf{B}_I^0$ is the *N*-dimensional vector of transition probabilities from transient states to absorbing state. Initially, the system begins with state 0 and therefore the initial state probability vector $\boldsymbol{\alpha}$ is given by $\boldsymbol{\alpha} = (1, 0, \cdots, 0)$. Probability distribution of the total number X_I of steps for the Markov chain to be absorbed to the state *N* can then be calculated using the following formula (see, e.g., Kao, 1997),

$$P(X_I = k) = \boldsymbol{\alpha} \mathbf{B}_I^{k-1} \mathbf{B}_I^0, \quad k = N, N+1, \cdots, \tag{5}$$

and its pgf and expected value are given by:

$$X_I(z) = z\boldsymbol{\alpha}(\mathbf{I} - z\mathbf{B}_I)^{-1}\mathbf{B}_I^0, \tag{6}$$

$$E(X_I) = \boldsymbol{\alpha}(\mathbf{I} - \mathbf{B}_I)^{-1}\mathbf{e}. \tag{7}$$

Since the total elapsed time to enter the state *N* from the beginning of the system is $A_{I,1} + A_{I,2} + \cdots + A_{I,X_I}$, we finally obtain the LT $I^*(\theta)$ and expected value of *I*, length of idle period, as following:

$$I^*(\theta) = X_I(A_I^*(\theta)), \tag{8}$$

$$E(I) = E(X_I)E(A_I). \tag{9}$$

4. DERIVATION OF THE LENGTH OF BUSY PERIOD

Busy period begins right after the system size reaches

to N. Times between arrivals may be different during this period from the idle period. The server continues to provide the service exhaustively, that is, until there remains no customer in the system. In this section, we obtain the probability distribution of the length B of busy period which is the elapsed time taken for the number of customers to fall to zero from N for the first time.

Similarly to the case of idle period, we define D_B as the total number of customers departed from the system after service completion during A_B. Then the probability generating function $D_B(z)$ of D_B is given by:

$$D_B(z) = A_B^*(\mu - \mu G(z)), \tag{10}$$

and the probability distribution δ_k of D_B is:

$$\delta_k \equiv P[D_B = k] = \frac{1}{k!}\frac{d^k}{dz^k} D_B(z)|_{z=0}. \tag{11}$$

In order to derive the probability distribution of B, we take different steps from the idle period case. The random process $\{N_B(\tau_n^-),\ n=1, 2, \cdots\}$ with state space $\{1, 2, 3, \cdots\}$ can also be modeled as a discrete-time absorbing Markov chain. The imbedded points of this chain are the time epochs of customer arrivals (i.e., upward increases in system size) and this chain starts at the state N and absorbs to the state 1, not to the state 0. This is because the chain does not recognize when actually the number of customers falls to zero until there is an increase of system size to one. As one can see in the following Figure 1, the chain's absorbing epoch to the state 1 (time point ⓑ) does not agree with the instant that the busy period terminates (time point ⓓ), and therefore, we should consider the number of steps that the chain takes until one step before it absorbs to state 1 (time point ⓒ), and then add to it the time that the system size falls to zero (length of interval between ⓒ and ⓓ). The Figure 1 shows a typical sample path during the busy period explaining the above reasoning.

We compute the probability distribution of B by taking the following four steps:

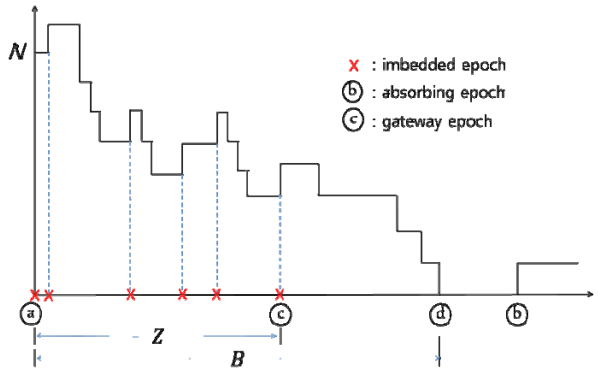


Figure 1. A typical sample path during busy period.

(i) To find the probability distributions of Y, number of imbedded epochs before the chain enters to the state 1 for the first time (Y = number of 'x' marks in Figure 1).
(ii) To derive the probability distributions of the "gateway state" and of the time Z to reach it. The state at the step Y is called gateway state, to which the chain visitsat one step before absorbing (state at ⓒ in Figure 1).
(iii) To calculate the number of batches served until the system size falls below 0 (i.e., number of downward decreases in system size between ⓒ and ⓓ in Figure 1) by conditioning the gateway state.
(iv) To obtain the time between ⓒ and ⓓ by summing the service times of the batches in (iii).

We explain the above steps in detail as following:
(i) Probability distribution of Y:
One-step transition probability matrix $\mathbf{P}_B$ of Markov chain $\{N_B(\tau_n),\ n=1, 2, \cdots\}$ with state space $\{1, 2, 3, \cdots\}$ is as follows:

$$\mathbf{P}_B = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{B}_B^0 & \mathbf{B}_B \end{bmatrix},\ \mathbf{B}_B = \begin{bmatrix} \delta_1 & \delta_0 & 0 & 0 & \cdots \\ \delta_2 & \delta_1 & \delta_0 & 0 & \cdots \\ \delta_3 & \delta_2 & \delta_1 & \delta_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

$$\mathbf{B}_B^0 = \begin{bmatrix} \sum_{i=2}^{\infty}\delta_i \\ \sum_{i=3}^{\infty}\delta_i \\ \sum_{i=4}^{\infty}\delta_i \\ \vdots \end{bmatrix}. \tag{12}$$

Then the number, Y, of steps before absorbing has the following probability distribution, pgf, and expected value (see e.g., Kao, 1997):

$$P(Y=i) = \boldsymbol{\beta}\mathbf{B}_B^{i-1}\mathbf{B}_B^0,\ \ i = 1, 2, 3, \cdots, \tag{13}$$

$$E(Y) = \boldsymbol{\beta}(\mathbf{I} - \mathbf{B}_B)^{-1}\mathbf{e}, \tag{14}$$

where $\boldsymbol{\beta} = (0, \cdots, 0, 1, 0, \cdots)$ and 1 occurs at the $(N\text{-}1)^{\text{th}}$ position.

(ii) Probability distributions of gateway state and time to gateway:
There are $\tilde{Y} \equiv Y - 1$ intervals and pgf $\tilde{Y}(z)$ of Y is

$$\tilde{Y}(z) = \frac{1}{z}Y(z) = \boldsymbol{\beta}(\mathbf{I} - z\mathbf{B}_B)^{-1}\mathbf{B}_B^0, \tag{15}$$

and expected value of Y is given by

$$E(\tilde{Y}) = E(Y) - 1. \tag{16}$$

Now, we can derive the LT $Z^*(\theta)$ of the time Z to the gateway state using the random sum $Z = A_{B,1} + A_{B,2} + \cdots + A_{B,\tilde{Y}}$,

$$Z^*(\theta) = \tilde{Y}(z)|_{z=A_B^*(\theta)} = \tilde{Y}(A_B^*(\theta)), \quad (17)$$

and expected value

$$E(Z) = E(\tilde{Y})E(A_B). \quad (18)$$

Since the gateway state is the number of customers at step Y, it is $N_B(\tau_Y)$. Denote ϕ_{ij} ($i,\ j = 2, 3, \cdots$) by the probability that the gateway state is j when starting at i, i.e., $\phi_{ij} \equiv P(N_B(\tau_Y) = j \mid N_B(0) = i)$. Then we have the following:

$$\phi_{ij} = \begin{cases} \sum_{k=2}^{i+1} (\mathbf{B}_B)_{ik}\phi_{kj}, & i \neq j, \\ (\mathbf{B}_B^0)_i + \sum_{k=2}^{i+1} (\mathbf{B}_B)_{ik}\phi_{kj}, & i = j. \end{cases} \quad (19)$$

By solving the above simultaneous linear equations with respect to ϕ_{ij}, we can get $\phi_{N,j}$ for $j = 2, 3, \cdots$, which is the probability distribution of gateway state. There are, however, infinitely many states and therefore it is impossible to get an exact solution in a closed form of (19). Instead, noticing that we only need $\phi_{N,j}$ for $j = 2, 3, \cdots$, not for all ϕ_{ij} ($i,\ j = 2, 3, \cdots$), and all entries after $(k+2)^{th}$ columns in k^{th} row are zeroes, we consider $(N+2)\times(N+2)$ submatrix of $\mathbf{B}_B$, and the first $N+2$ entries of $\mathbf{B}_B^0$. This makes (19) finite system of simultaneous equations and an approximation solution could be obtained.

(iii) Number of batches served until the system becomes empty by conditioning the gateway state:
Let L_j be the number of batches served until the system size drops to zero, assuming that the gateway state is j. Denoting $G_i \geq 1$ ($i = 1, 2, \cdots$) by the size of the i^{th} batch, it is given by:

$$L_j = \min\{k : G_1 + \cdots + G_{k-1} < j,\ G_1 + \cdots + G_k \geq j\}. \quad (20)$$

Then the probability distribution of L_j is, for $k = 1, 2, \cdots, j$, as follows:

$$\begin{aligned} P(L_j = k) &= P(G_1 + \cdots + G_{k-1} < j,\ G_1 + \cdots + G_k \geq j) \\ &= P(G_1 + \cdots + G_{k-1} \leq j-1) - P(G_1 + \cdots + G_k \leq j-1) \\ &= G^{(k-1)}(j-1) - G^{(k)}(j-1). \end{aligned} \quad (21)$$

where $G^{(i)}(\cdot)$ denotes the i-fold convolution of the distribution of G. Then we obtain the pgf $L_j(z)$ of L_j as following:

$$\begin{aligned} L_j(z) &= \sum_{k=1}^{j} z^k P(L_j = k) = \sum_{k=1}^{j} z^k [G^{(k-1)}(j-1) - G^{(k)}(j-1)] \\ &= z\sum_{k=1}^{j} z^k G^{(k)}(j-1) - \{\sum_{k=1}^{j} z^k G^{(k)}(j-1) - z^0 G^{(0)}(j-1)\} \\ &= 1 - (1-z)\sum_{k=1}^{j} z^k G^{(k)}(j-1), \end{aligned} \quad (22)$$

and expected value of L_j is

$$E(L_j) = L_j'(1) = \sum_{k=1}^{j} G^{(k)}(j-1). \quad (23)$$

(iv) Time for the system to be empty from gateway:
$S_1 + S_2 + \cdots + S_{L_j}$ is the time for the number of customers in the system to reach to zero from gateway state j, and the LT and mean of it are, $L_j(z)|_{z=S^*(\theta)} = L_j\left(\frac{\mu}{\mu+\theta}\right)$ and $(1/\mu)E(L_j)$, because S_k 's are iid exponential with mean $1/\mu$. Now, we finally obtain the LT $B^*(\theta)$ and mean of B as follows:

$$B^*(\theta) = Z^*(\theta) \cdot \sum_{j=2}^{\infty} \phi_{N,j} \cdot L_j\left(\frac{\mu}{\mu+\theta}\right), \quad (24)$$

$$E(B) = E(Z) + \frac{1}{\mu}\sum_{j=2}^{\infty} \phi_{N,j} \cdot E(L_j). \quad (25)$$

Remark 1: In case of $N = 1$, we need a slight modification because the starting state of the Markov chain is $N(=1)$, which is absorbing state. Therefore, we divide the cases according to whether the first transition is toward the state 1 or 2. The former happens when the inter arrival time is greater than the service time, and then the busy period is just the length of one single service time. For the latter, we can apply the procedure discussed above with $N = 2$.

Remark 2: We verify Eq. (25) by matching it to the existing results as special cases of our model. Expected value of busy period of M/M/1 system with N policy, which is a special case of our model with Poisson arrivals and single batch $(G = 1)$, is known to be $E(B) = N/(\mu - \lambda)$. Let $\mu = 1$ and λ be the arrival rate, then $\delta_k = \left(\frac{1}{\lambda+1}\right)^k \left(\frac{\lambda}{\lambda+1}\right)$, $k = 0, 1, 2, \cdots$. After building the transition probability matrix $\mathbf{P}_B$ in (12), incorporating with $E(L_j) = j$, we can obtain $E(B)$ by means of (14), (16), (17), (19), and (25). The following Table 1 shows matching result for $N = 5,\ 20$ and $\lambda = 0.5,\ 0.8$ compared to the known result.

Table1. Matching results of $E(B)$ to M/M/1 N policy

		$E(B) = N/(\mu - \lambda)$	$E(B)$ of our method
$N = 5$	$\lambda = 0.5$	10	10.32
	$\lambda = 0.8$	25	25.34
$N = 20$	$\lambda = 0.5$	40	40.15
	$\lambda = 0.8$	100	99.59

Another matching is done for the general arrival model. Literatures (e.g., Chae and Kim, 2007; Kim and

Yang, 2011; Chae and Lim, 2008) provide analytic results on the busy period of G/M/1 with N policy. According to (Chae and Lim, 2008), however, even for the simplest deterministic arrival case, one should differentiate N times the complex function to get the LT of busy period, which is almost intractable when N is not small. Instead, we consider D/M/1 with $N=1$ and compare $E(B)$ calculated by our method with the known result $E(B)=1/(1-z_0)$ (e.g., Takacs, 1962), where z_0 is the solution of $z=A_B^*(\mu-\mu z)=e^{-a(\mu-\mu z)}$ and a is interarrival time (constant). In this case, derivation of our $E(B)$ is modified to $E(B)=(1-e^{-a})\times(1/\mu)+e^{-a}\times E(B\mid N=2)$ as indicated in Remark 1. Table 2 summarizes the comparison result when $\mu=1$:

Table 2. Matching results of $E(B)$ to D/M/1 ($N=1$)

	$E(B)=1/(1-z_0)$	$E(B)$ of our method
$a=2$	1.256	1.255
$a=1.25$	2.703	2.834
$a=1.15$	4.082	4.183

From Table 1 and Table 2, we conclude our computational results matches well to the existing results.

5. COST FUNCTION

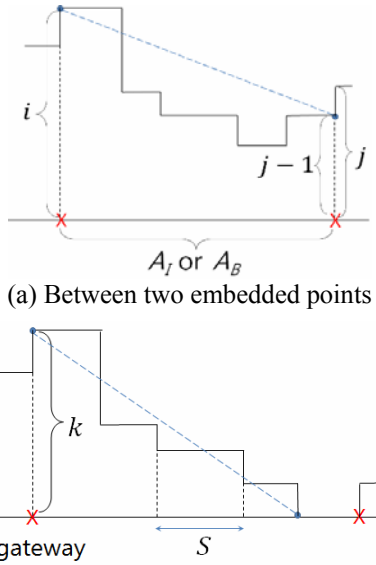
In this section, we minimize the total cost per unit time involved in a cycle. Let C_s be the setup cost incurred when the server initiates service because the number of waiting customers once reaches N. And let C_h be the holding cost per customer per unit time in the system. As the threshold N gets bigger, the length of a cycle becomes longer and setup cost per unit time gets lower, while the total customer holding cost becomes larger. Therefore, a proper level of N should be determined to trade these two costs off to achieve minimum overall cost. Expected setup cost per unit time is given by:

$$E(\text{Setup cost per unit time})=\frac{C_s}{E(I)+E(B)}. \tag{26}$$

Expected holding cost during idle period is as following: expected number of visits to a given state i before the chain absorbs to state N starting from state 0 is the $(0,i)$ th element of the fundamental matrix

$$\mathbf{I}+\mathbf{B}_I+\mathbf{B}_I^2+\cdots=(\mathbf{I}-\mathbf{B}_I)^{-1} \tag{27}$$

of the chain. If the chain is in state i, it stays there for, on average, $E(A_I)$ and moves to the state j with probability $(\mathbf{B}_I)_{ij}$, $j=0,1,\cdots,i+1$. Therefore, the average area of the parallelogram (see Figure 2(a)) is given by $E(A_I)\times(i+j-1)/2$ and by adding these areas up for all $i=0,1,\cdots,N-1$ and $j=0,1,\cdots,i+1$ we obtain the expected



(a) Between two embedded points

(b) Between gateway and absorbing points

Figure 2. Average number of customers.

number of customers during the idle period as follows:

$$E(\text{Number of customers during idle period})\equiv L^I$$
$$=\sum_{i=0}^{N-1}\left[\left((\mathbf{I}-\mathbf{B}_I)^{-1}\right)_{0i}\times\left\{\sum_{j=0}^{i+1}(\mathbf{B}_I)_{ij}\frac{i+j-1}{2}\times E(A_I)\right\}\right]. \tag{28}$$

Similarly, the expected number of customers during the busy period is given by the following:

$$E(\text{Number of customers during busy period})\equiv L^B$$
$$=\sum_{i=2}^{\infty}\left[\left((\mathbf{I}-\mathbf{B}_B)^{-1}\right)_{N+1,i}\times\left\{\sum_{j=2}^{i+1}(\mathbf{B}_B)_{ij}\frac{i+j-1}{2}E(A_B)\right\}\right]$$
$$+\sum_{k=2}^{\infty}\phi_{N+1,k}\frac{k}{2}E(L_k)E(S) \tag{29}$$

The first term of Eq. (29) is the expected total area up to the gateway state starting from N, and the second term is the expected total area from the gateway state to the state 0 (see Figure 2(b)). Finally, the expected total holding cost per unit time is given by:

$$E(\text{Total holding cost cost per unit time})=\frac{L^I+L^B}{E(I)+E(B)}\times C_h. \tag{30}$$

Finally, by adding the Eqs. (26) and (30) we have the total expected cost per unit time, and the optimal value N^* can be obtained by minimizing it.

6. NUMERICAL EXAMPLES

Some numerical examples are introduced in this section to illustrate our model. The inter arrival times of

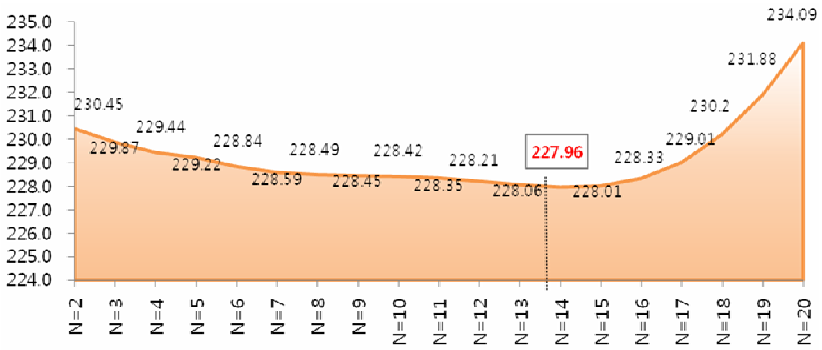


Figure 3. Graph of the total expected cost per unit time as *N* changes.

customers during idle and busy periods are assumed to be constant, $A_I = 0.1$ and $A_B = 10$, and the reneging rate $\eta = 1$ per unit time. We also assume the probability distribution of the number of customers in each batch is given by: $g_1 = 0.6$, $g_2 = 0.4$ and thus $E(G) = 1.4$. Service rate of each batch is $\mu = 1/E(S) = 0.14$ per unit time. Setup cost and holding cost are 1200 and 1per unit time, respectively $(C_s = 1200,\ C_h = 1)$. Threshold level, *N*, is varied from 5 to 20 to see the change of total expected cost per unit time, which is depicted in the Figure 3. In this setting, the optimal value of *N* is 14 and the minimum value of the total expected cost per unit time is 227.96. Since the optimal value 227.96 is less than 1% better than the values for $N = 3, 4, \cdots, 18$ in this specific example, a more flexible choice can be done among these values for the final result.

The following Figure 4 shows, for fixed N =15 and different values of A_B / A_I, the change of total cost as the cost ratio C_s / C_h increases. It is obvious, from Eqs. (26) and (30), that the total cost is a linear function of C_s / C_h, as seen in Figure 4, where the slope is $1/E$ (cycle). If the ratio A_B / A_I gets bigger, it implies either less customers arrive during busy period or more customers arrive during idle period. In the former case, busy period becomes shorter, while in the latter case idle period gets shorter. Length of cycle becomes smaller and therefore the slope of the line becomes steeper.

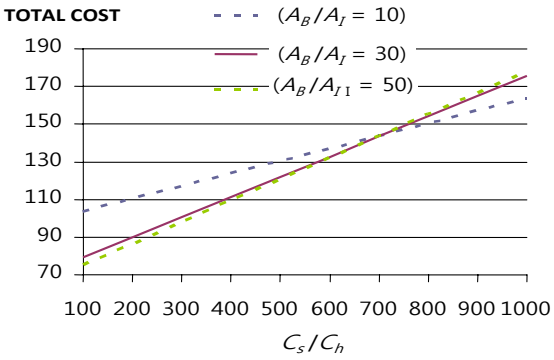


Figure 4. Graph of the total expected cost per unit time as C_s / C_h increases for fixed $N = 15$ and different values of A_B / A_I.

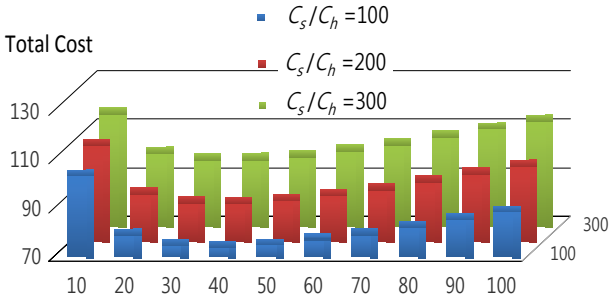


Figure 5. Graph of the total expected cost per unit time as A_B / A_I increases for fixed $N = 15$ and different values of C_s / C_h.

Now, in the Figure 5, we fix N = 15 and see the change of total cost as the inter arrival time ratio A_B / A_I varies, instead of C_s / C_h. For given C_s / C_h, total holding cost and setup cost are compensated of each other and we can obtain the optimal value of A_B / A_I which attains minimum total cost per unit time. As an example, when $C_s / C_h = 300$, the total cost is minimized by controlling the arrival rate during busy period to be 1/30 of that during idle period.

7. SUMMARY AND CONCLUSION

In this paper, we considered a general arrival and exponential service queueing system with *N*-policy. Customers arrived while the server is idle are impatient and may leave the system without being served after exponential time sojourn in the system. The arrival rates during idle and busy periods are different from each other and the server provides service in batches. We divided a cycle into an idle period and a busy period, of which the probability distributions of lengths were obtained based on the absorbing Markov chain approach. A method to find the optimal *N* value was provided when there exist setup cost and waiting cost of customers in the system to minimize overall expected cost per unit time.

By means of extensive numerical illustrations, the relationship among parameters such as C_s / C_h, A_B / A_I and

N are revealed, from which one can deduce the best operating policy that minimizes the overall cost of the considered queueing system.

REFERENCES

Akcan, S. (2013), A New Approximation for Inventory Control System with Decision Variable Lead-Time and Stochastic Demand, *International Journal of Industrial Engineering: Theory, Applications and Practice*, **20**(3/4).

Altman, E. and Yechiali, U. (2006), Analysis of Customers' Impatience in Queues with Server Vacations, *Queueing Systems*, **52**, 261-279.

Bae, J. and Kim, S. (2010), The Stationary Workload of the G/M/1 Queue with Impatient Customers, *Queueing Systems*, **64**, 253-265.

Benjaafar, S., Gayon, J., and Tepe, S. (2010), Optimal Control of a Production-Inventory System with Customer Impatience, *Operations Research Letters*, **38**, 267-272.

Blackburn, J. D. (1972), Optimal Control of a Single Server Queue with Balking and Reneging, *Management Science*, **19**, 297-313.

Chae, K. C. and Kim, S. J. (2007), Busy Period Analysis for the GI/M/1 Queue with Exponential Vacations, *Operations Research Letters*, **35**(1), 114-118.

Chae, K. C. and Lee, S. M. (2005), An Absorbing Markov Chain Approach to GI/M/1 Queues with Generalized Vacations, *Asia Pacific Management Review*, **10**, 163-167.

Chae, K. C. and Lim, D. E. (2008), Busy period analysis for the n-policy GI/M/c queue, *Journal of the Korean Statistical Society*, **37**(3), 285-290.

Choi, B. D., Kim, B., and Zhu, D. (2004), MAP/M/c Queue with Constant Impatience Time, *Mathematics of Operations Research*, **29**, 309-325.

Kao, P. C. (1997), *An Introduction to Stochastic Processes*, Duxbury Press, Belmont, California.

Ke, J.-C. (2003), The Analysis of a General Input Queue with N Policy and Exponential Vacations, *Queueing Systems*, **45**, 135-160.

Kim, K. and Yang, W. S. (2011), Busy Period Analysis for the GI/M/1 Queue with Phase-Type Vacations, *Journal of the Korean Statistical Society*, **40**(1), 55-62.

Lee, H. W. and Ahn, B. Y. (2002), Operational Behavior of the MAP/G/1 Queue under N-Policy with Single Vacation and Set-Up, *Journal of Applied Mathematics and Stochastic Analysis*, **15**, 167-196.

Lee, H. W. and Park, N. I. (2004), Using Factorization for Waiting Times in BMAP/G/1 Queues with N-Policy and Vacations, *Stochastic Analysis and Applications*, **22**, 755-773.

Mandelbaum, A. and Momčilović, P. (2012), Queues with Many Servers and Impatient Customers, *Mathematics of Operations Research*, **37**, 41-65.

Swensen, A. R. (1986), On a GI/M/c Queue with Bounded Waiting Times, *Operations Research*, **34**, 895-908.

Tadj, L. and Choudhury, G. (2005), Optimal Design and Control of Queues, *Sociedad de Estadística e Investigación Operativa, Top*, **13**, 359-412.

Takacs, L. (1962), *Theory of queues*, Oxford: Oxford University Press, reprinted in 1982 by Greenwood Press, Westport, CT.

Takagi H. (1991), *Queueing Analysis: A Foundation of Performance Evaluation*, North-Holland, **1**.

Yadin, M. and Naor, P. (1963), Queueing Systems with a Removable Service Station, *Operational Research Quarterly*, **14**, 393-405.

Yue, D., Yue, W., and Li, X. (2011), Analysis of a Two-Phase Queueing System with Impatient Customers and Multiple Vacations, *The Tenth International Symposium on Operations Research and Its Applications* (*ISORA 2011*), Dunhuang, China, 292-298.

Zhe George Zhang, Z. G. and Tian, N. (2004), The N threshold policy for the GI/M/1 queue, *Operations Research Letters*, **32**, 77-84.