M/M/1 Queueing System with Delayed Controlled Vacation

Yonglu Deng[*], Zhongshan University
W. John Braun[†], University of Winnipeg
Yiqiang Q. Zhao[†], University of Winnipeg

Abstract

An M/M/1 queue with delayed vacation is studied. If the server has been idle for a period of time (called the delay time), the server begins an exponentially distributed vacation which is repeated as long as the number of customers in the system remains less than some number K. For the cases of exponential and deterministic delay time, exact expressions for the steady state probability distribution are obtained, together with associated performance measures. System optimization is also considered; values of K are given which minimize the average total cost per unit time, and it is shown that the optimal delay period is either 0 (no delay) or infinite (no vacation), in case of Poisson arrivals.

Key words: Queueing model, stationary distribution, delayed vacation.

[*] Department of Mathematics, Zhongshan University, Guangzhou, 510275, China

[†] Department of Mathematics and Statistics, University of Winnipeg, Winnipeg, Manitoba, Canada, R3B 2E9.

1 Introduction

Queueing systems with server's vacations have been studied extensively. A comprehensive review on vacation models, methods and results up to 1991 can be found in either Doshi [6] or Takagi [11]. In the last few years, increasing interest in studying queueing systems with various rules of vacation has led to many extensions of previously existing results. For example, a batch arrival model with a finite capacity for the buffer size can be used to model some telecommunications systems using a time division multiple access (TDMA) scheme (Frey and Takahashi, [7]). Researchers have also done performance analysis on systems where probability distributions of the variables are more general and closer to reality (e.g. Chao and Zhao [2]).

In this paper, we consider vacation models which are extended in two different ways. The first one is the queueing system in which the length of the vacations can be controlled by means of the number of customers K arriving to the system during the vacation, and the level K may be chosen according to the arrival rate and the service rate, the cost per unit of waiting time and the cost of the server being transferred from vacation to work.

Secondly, we allow for a delay time before a vacation begins. During the delay time, the server is situated in warm standby state and it will start service immediately upon arrival of a customer to the system. It seems that the consideration of such a delayed vacation might be reasonable if the cost of a server's warm switch-on is lower than that of a server's cold switch-on.

The case of delayed vacation has also very recently been studied by Frey and Takahashi [8] and Sakai et al [9] where the term close-down time is used. They consider general service, delay and vacation times with Poisson (batch) arrivals, but must rely on numerical examples to gain insights into their results. We focus on a simpler model with exponential service and vacation which allows us to gain greater insight into the system with less effort.

In particular, we are able to consider optimization of the system with respect to the control parameter and the delay parameter. We show that for Poisson arrivals and exponential service and vacation, there is no nontrivial optimal delay period. However, the use of the control parameter to extend the vacation period until K customers arrive in the system can result in improved performance of the system.

2 Description and Analysis of the Model

We make the following assumptions.

1. Customers arrive at the system according to a Poisson process with intensity λ and there is one server in the system. The queue discipline is FCFS.

2. Service times are assumed to be exponentially distributed with mean $1/\mu$. The traffic intensity is $\rho = \lambda/\mu < 1$.

3. A period of time called the delay time occurs before the server goes on vacation. The delay can be interrupted by the arrival of a customer in which case the server resumes service. Thus, the length of the delay time is

$$\min(X, Y)$$

where X is exponentially distributed with mean $1/\lambda$, and Y is a random variable with mean $1/\theta_1$. In this paper, we assume that Y is either exponentially distributed or deterministic.

4. If the delay time is completed before the arrival of a customer, the server begins a vacation whose length is exponentially distributed with mean $1/\theta_2$. Upon completion of a vacation, the server resumes service if K (≥ 1) or more customers are in the system. Otherwise, it takes another vacation having length independent of and identically distributed to the preceding one.

5. All aforementioned random variables are independent of each other.

Let $S(t) = 0$, $S(t) = 1$, and $S(t) = 2$ denote the events that the server is busy, in delay period, and on vacation at epoch t, respectively. Define

$$\begin{aligned} p_{0j}(t) &= P(N(t) = j, S(t) = 0) \qquad (j = 1, 2, 3, \ldots) \\ p_1(t) &= P(N(t) = 0, S(t) = 1) \\ p_{2j}(t) &= P(N(t) = j, S(t) = 2) \qquad (j = 0, 1, 2, \ldots) \end{aligned}$$

$N(t)$ will denote the number of customers in the system at time t.

2.1 Exponential Delay Case

When the delay time is exponentially distributed, $\{(N(t),S(t))\,,t\geq 0\}$ is a standard continuous-time Markov chain. From the theory of Markov chains (e.g. [3]), it follows that $\{(N(t),S(t))\,,t\geq 0\}$ has a unique equilibrium distribution which satisfies the following family of equations.

Setting

$$\begin{aligned}
p_{0j} &= \lim_{t\to\infty} p_{0j}(t) \qquad (j=1,2,3,\ldots)\\
p_{1} &= \lim_{t\to\infty} p_{1}(t)\\
p_{2j} &= \lim_{t\to\infty} p_{2j}(t) \qquad (j=0,1,2,\ldots)
\end{aligned}$$

we have

$$\begin{aligned}
\lambda p_{20} &= \theta_1 p_1\\
\lambda p_{2j} &= \lambda p_{2,j-1} \qquad (j=1,2,3,\ldots,K-1)\\
(\lambda+\theta_2)p_{2j} &= \lambda p_{2,j-1} \qquad (j=K,K+1,K+2,\ldots)\\
(\lambda+\theta_1)p_{1} &= \mu p_{01} \qquad\qquad (1)\\
(\lambda+\mu)p_{01} &= \lambda p_{10} + \mu p_{02}\\
(\lambda+\mu)p_{0j} &= \lambda p_{0,j-1} + \mu p_{0,j+1} \qquad (j=2,3,4,\ldots,K-1)\\
(\lambda+\mu)p_{0j} &= \lambda p_{0,j-1} + \mu p_{0,j+1} + \theta_2 p_{2j} \qquad (j=K,K+1,K+2,\ldots)
\end{aligned}$$

$$\sum_{j=1}^{\infty} p_{0j} + p_1 + \sum_{j=0}^{\infty} p_{2j} = 1$$

The solution of these equations is as follows (see Appendix for a derivation):

$$p_1 = \frac{\lambda\theta_2(1-\rho)}{\lambda\theta_1+\lambda\theta_2+K\theta_1\theta_2} \qquad (2)$$

$$p_{2j} = \begin{cases} \frac{\theta_1}{\lambda}p_1 & (j=0,1,2,\ldots,K-1)\\ \left(\frac{\lambda}{\lambda+\theta_2}\right)^{j-k+1} p_{20} & (j=K,K+1,K+2,\ldots)\end{cases} \qquad (3)$$

$$p_{0j} = p_1\rho^j + \sum_{i=0}^{j-1}\rho^{j-i}p_{2i} \qquad (j=1,2,3,\ldots) \qquad (4)$$

2.2 Deterministic Delay Case

When the delay time is deterministic, $\{(N(t), S(t)), t \geq 0\}$ is not a Markov chain, but it can be extended to a continuous-state Markov chain with the use of the supplementary variable technique (e.g. [5], [4]). In particular, we introduce the random variable $X(t)$ which denotes the elapsed delay time at time t. Setting

$$\begin{aligned}
p_{0j} &= \lim_{t\to\infty} p_{0j}(t) \qquad (j = 1, 2, 3, \ldots) \\
p_1(x)dx &= \lim_{t\to\infty} P(x \leq X(t) < x + dx, S(t) = 1), \qquad 0 < x < 1/\theta_1 \\
p_{2j} &= \lim_{t\to\infty} p_{2j}(t) \qquad (j = 0, 1, 2, \ldots)
\end{aligned}$$

we have

$$\begin{aligned}
\lambda p_{20} &= p_1(1/\theta_1) \\
\lambda p_{2j} &= \lambda p_{2,j-1} \qquad (j = 1, 2, 3, \ldots, K-1) \\
(\lambda + \theta_2) p_{2j} &= \lambda p_{2,j-1} \qquad (j = K, K+1, K+2, \ldots) \\
p_1(0) &= \mu p_{01} \\
(\lambda + \mu) p_{01} &= \lambda \int_0^{1/\theta_1} p_1(x)dx + \mu p_{02} \\
(\lambda + \mu) p_{0j} &= \lambda p_{0,j-1} + \mu p_{0,j+1} \qquad (j = 2, 3, 4, \ldots, K-1) \\
(\lambda + \mu) p_{0j} &= \lambda p_{0,j-1} + \mu p_{0,j+1} + \theta_2 p_{2j} \qquad (j = K, K+1, K+2, \ldots) \\
\frac{d}{dx} p_1(x) &= -\lambda p_1(x), \qquad 0 < x < 1/\theta_1
\end{aligned} \tag{5}$$

$$\sum_{j=1}^{\infty} p_{0j} + \int_0^{1/\theta_1} p_1(x)dx + \sum_{j=0}^{\infty} p_{2j} = 1$$

The solution in this case is as follows:

$$p_1 = \frac{(e^{\lambda/\theta_1} - 1)(1-\rho)\theta_2}{\lambda - \theta_2 - e^{\lambda/\theta_1}\theta_2 - K\theta_2}$$

$$p_{2j} = \begin{cases} \frac{(1-\rho)\theta_2}{\lambda - \theta_2 + e^{\lambda/\theta_1}\theta_2 + K\theta_2} & (j = 0, 1, 2, \ldots, K-1) \\ \left(\frac{\lambda}{\lambda+\theta_2}\right)^{j-k+1} p_{20} & (j = K, K+1, K+2, \ldots) \end{cases}$$

$$p_{0j} = p_1 \rho^j + \sum_{i=0}^{j-1} \rho^{j-i} p_{2i} \qquad (j = 1, 2, 3, \ldots)$$

2.3 Mean Queue Length

We are now in a position to deduce the mean numbers of customers in the system and in the queue. Let L denote the mean number of customers in the system, and let L_q denote the mean queue length. Then, we have

$$L = \frac{\rho}{1-\rho} + \frac{\theta_1\theta_2}{\lambda\theta_1 + \lambda\theta_2 + K\theta_1\theta_2}\left(\frac{K(K-1)}{2} + \left(\frac{\lambda}{\theta_2}\right)^2 + \frac{K\lambda}{\theta_2}\right) \tag{6}$$

and

$$L_q = \frac{\rho^2}{1-\rho} + \frac{\theta_1\theta_2}{\lambda\theta_1 + \lambda\theta_2 + K\theta_1\theta_2}\left(\frac{K(K-1)}{2} + \left(\frac{\lambda}{\theta_2}\right)^2 + \frac{K\lambda}{\theta_2}\right) \tag{7}$$

in the exponential delay case. In the deterministic delay case, we have

$$L = \frac{\rho}{1-\rho} + \frac{2\,\lambda^2 + 2\,K\,\lambda\,\theta_2 - K\,{\theta_2}^2 + K^2\,{\theta_2}^2}{2\,\theta_2\,\left(\lambda - \theta_2 + e^{\frac{\lambda}{\theta_1}}\,\theta_2 + K\,\theta_2\right)} \tag{8}$$

and

$$L_q = \frac{\rho^2}{1-\rho} + \frac{2\,\lambda^2 + 2\,K\,\lambda\,\theta_2 - K\,{\theta_2}^2 + K^2\,{\theta_2}^2}{2\,\theta_2\,\left(\lambda - \theta_2 + e^{\frac{\lambda}{\theta_1}}\,\theta_2 + K\,\theta_2\right)} \tag{9}$$

2.4 Waiting Time

Let W_q denote the virtual waiting time of customers arriving in the system. Then

$$E[W_q] = -{W^*}'(0) = \frac{p_{20}}{(\mu-\lambda)\theta_2}\left[\frac{\lambda\mu}{\theta_2} + \frac{\lambda^2}{\mu-\lambda} + K\mu\right] + \frac{K(K-1)\mu p_{20}}{2\lambda(\mu-\lambda)} + \frac{\lambda(p_1 + p_{20}K)}{(\mu-\lambda)^2} \tag{10}$$

and

$$E[W_q|W_q>0] = \frac{1}{1-p_1}\left(\frac{p_{20}}{(\mu-\lambda)\theta_2}\left[\frac{\lambda\mu}{\theta_2} + \frac{\lambda^2}{\mu-\lambda} + K\mu\right] + \frac{K(K-1)\mu p_{20}}{2\lambda(\mu-\lambda)} + \frac{\lambda(p_1 + p_{20}K)}{(\mu-\lambda)^2}\right) \tag{11}$$

We also note that Little's formula also holds for this model.

2.5 Mean Length of Operational Period

The period of time from the beginning of a delay to the end of the subsequent delay will be defined as an operational period for the system; we denote its length by A.

By the model assumptions, a busy period which begins with one customer in the system, or a vacation of length V_K with a control level K will begin as soon as the delay terminates depending on whether a customer arrives or the server goes on vacation. In the first case, the busy period of length B_1 will be followed (eventually) by another delay of length D and then the above process will be repeated. In the latter case, a busy period of length B_R will follow the vacation of length V_K where R is a random variable with values in $\{K, K+1, \ldots\}$. Another delay of length D follows B_R so that the above process will then be repeated.

From the above considerations, it is clear that an operational period consists of a delay D, a random number (say, M, a non-negative integer-valued random variable) of successive $B_1 + D$ events, a vacation of length V_K, and a busy period of length B_R. In particular, we have

$$E[A] = E[D] + E[M](E[B_1] + E[D]) + E[V_K] + E[B_R] \tag{12}$$

The mean length of a delay is easily found. That is,

$$D = \min(X, Y)$$

where X is an exponential random variable with mean $1/\lambda$, and Y is a random variable with mean $1/\theta_1$.

If Y is exponential, then D is an exponential random variable with mean

$$E[D] = 1/(\lambda + \theta_1) \tag{13}$$

If Y is deterministic, then D is a truncated exponential random variable with mean

$$E[D] = (1 - e^{-\lambda/\theta_1})/\lambda \tag{14}$$

Letting M denote the number of $B_1 + D$ events before the server takes a vacation, and using the memoryless property of the exponential distribution, it is seen that these events are independent, and they end with a vacation with a certain probability p_0.

Therefore,

$$P(M = m) = p_0(1 - p_0)^m \qquad (m = 0, 1, 2, \ldots)$$

so that

$$E[M] = \frac{1 - p_0}{p_0} \tag{15}$$

When the delay time Y is exponential,

$$p_0 = P(X > Y) = \frac{\theta_1}{\lambda + \theta_1}$$

and when Y is deterministic,

$$p_0 = P(X > Y) = e^{-\lambda/\theta_1}$$

It is also clear that a busy period beginning with r customer in the system has mean

$$E[B_r] = \frac{r}{\mu - \lambda} \qquad (r = 1, 2, 3, \ldots) \tag{16}$$

It is shown in the appendix that the mean busy period beginning with a random number (R) of customers is given by

$$E[B_R] = \frac{\lambda + K\theta_2}{(\mu - \lambda)\theta_2} \tag{17}$$

A vacation period with control parameter K has mean length given by

$$E[V_K] = \frac{K}{\lambda} + \frac{1}{\theta_2} \tag{18}$$

Substituting (13), (15), (18), and (17) into (12) gives, upon simplification,

$$E[A] = \frac{\lambda(\theta_1 + \theta_2) + K\theta_1\theta_2}{(1 - \rho)\lambda\theta_1\theta_2} \tag{19}$$

in the exponential case, and using (14) in place of (13) gives

$$E[A] = \frac{\theta_2 e^{\lambda/\theta_1} + (K - 1)\theta_2 + \lambda}{\lambda(1 - \rho)\theta_2} \tag{20}$$

in the deterministic case.

3 Optimization of the Control Parameter

Suppose that the cost per unit time is

- $C_0 \leq 0$ for server vacation
- $C_1 \geq 0$ for delay (warm standby)

- $C_2 \geq 0$ for normal service $\quad (C_2 \geq C_1$ generally)
- $C_4 \geq 0$ per customer waiting in the system

In addition, $C_3 \geq 0$ denotes the cost due to switching the server from vacation to normal service.

For an operational period, the expected cost due to vacation is given by

$$C_0(K/\lambda + 1/\theta_2)$$

The expected cost due to delay is

$$C_1 E[D]/p_0$$

The expected cost while the server is busy is

$$C_2\left(\frac{\lambda+\theta_2 K}{(\mu-\lambda)\theta_2}\right)+\frac{C_2}{\mu-\lambda}\left(\frac{1}{p_0}-1\right)$$

From the previous section, the expected length of an operational period is

$$E[A]=\frac{K}{\lambda}+\frac{1}{\theta_2}+\frac{E[D]}{p_0}+\frac{\lambda+\theta_2 K}{(\mu-\lambda)\theta_2}+\frac{1}{\mu-\lambda}\left(\frac{1}{p_0}-1\right)$$

In the case of exponential delay, we then find that the cost for an operational period is given by

$$C=C_0\left(\frac{K}{\lambda}+\frac{1}{\theta_2}\right)+\frac{C_1}{\theta_1}+C_2\frac{\rho\lambda(\theta_1+\theta_2)+K\theta_1\theta_2}{(1-\rho)\lambda\theta_1\theta_2}+C_3$$

so that the average cost per unit time is

$$\frac{C}{E[A]}=\frac{C_0(1-\rho)\theta_1(K\theta_2+\lambda)+C_1(1-\rho)\lambda\theta_2+C_2\left(\rho\lambda(\theta_1+\theta_2)+K\rho\theta_1\theta_2\right)+C_3(1-\rho)\lambda\theta_1\theta_2}{\lambda(\theta_1+\theta_2)+K\theta_1\theta_2}$$

The average total cost per unit time is then

$$G=\frac{C}{E[A]}+C_4 L$$

Setting $dG/dK=0$, we obtain

$$\frac{C_4\theta_1\theta_2 K^2}{2}+C_4\lambda(\theta_1+\theta_2)K+C_0(1-\rho)\lambda\theta_2-C_1(1-\rho)\lambda\theta_2-C_3(1-\rho)\lambda\theta_1\theta_2+C_4\lambda\left(\lambda-\frac{\theta_1+\theta_2}{2}\right)=0$$

The nonnegative solution of this is given by

$$K^*=(\theta_1\theta_2)^{-1}\left\{-\lambda(\theta_1+\theta_2)+\sqrt{\lambda^2(\theta_1+\theta_2)^2-2\frac{\theta_1\theta_2 T}{C_4}}\right\}$$

where

$$T = C_0(1-\rho)\theta_2 - C_1(1-\rho)\theta_2 - C_3(1-\rho)\theta_1\theta_2 + C_4\lambda\left(\lambda - \frac{\theta_1+\theta_2}{2}\right)$$

K^* is nonnegative if $T \leq 0$, and the second derivative d^2G/dK^2 is nonnegative in this case. Therefore, if $K_{\min}$ is the nearest integer to K^*, then $K_{\min}$ will minimize cost (assuming all other quantities are held constant).

A similar analysis applies In the case of deterministic delay; again, optimization with respect to K gives rise to a quadratic equation in K for which the nonnegative solution is given by

$$K^* = 1 - e^{\frac{\lambda}{\theta_1}} - \frac{\lambda}{\theta_2} + \frac{\sqrt{C_4(\lambda^2\,\mu^2 + \lambda\,\mu^2\,\theta_2) + 2\,T + C_4\,e^{\frac{2\,\lambda}{\theta_1}}\;\mu^2\,{\theta_2}^2 + 2\,C_3\,\lambda\,\mu^2\,{\theta_2}^2}}{\sqrt{C_4}\,\mu\,\theta_2}$$

where

$$\begin{aligned} T = {} & \lambda^2\,{\theta_2}^2 - e^{\frac{\lambda}{\theta_1}}\,\lambda^2\,{\theta_2}^2 - \lambda\,\mu\,{\theta_2}^2 - C_0\,\lambda\,\mu\,{\theta_2}^2 + e^{\frac{\lambda}{\theta_1}}\,\lambda\,\mu\,{\theta_2}^2 + C_0\,e^{\frac{\lambda}{\theta_1}}\,\lambda\,\mu\,{\theta_2}^2 - C_3\,\lambda^2\,\mu\,{\theta_2}^2 \\ & + C_0\,\mu^2\,{\theta_2}^2 - C_0\,e^{\frac{\lambda}{\theta_1}}\,\mu^2\,{\theta_2}^2 - C_4\,e^{\frac{\lambda}{\mathit{theta1}}}\,\mu^2\,{\theta_2}^2 \end{aligned}$$

Remark: A similar treatment will yield optimality conditions on θ_1 and θ_2. We note that the cost function is monotonic in θ_1, so that the optimal delay period for this system is either infinite (no vacation) or 0 (no delay). We conjecture that this behaviour is a result of the memoryless property of the arrival process. On the basis of a simulation study, we have observed that, for non-Poisson arrivals, a nontrivial delay period will be optimal.

Appendix

A.1 Derivation of Stationary Queue Length Distribution

To solve the family of equations (1), we introduce the following generating functions

$$G_0(s) = \sum_{j=1}^{\infty} p_{0j}s^j \qquad (|s| \leq 1)$$

and

$$G_2(s) = \sum_{j=0}^{\infty} p_{2j}s^j \qquad (|s| \leq 1)$$

From the first three equations of (1), it is clear that

$$p_{20} = p_{21} = \cdots = p_{2,K-1} \tag{21}$$

$$p_{20} = \frac{\theta_1}{\lambda} p_1 \tag{22}$$

and

$$p_{2j} = \left(\frac{\lambda}{\lambda + \theta_2}\right)^{j-k+1} p_{20} \qquad (j = K, K+1, K+2, \ldots) \tag{23}$$

from which it follows that we can write

$$G_2(s) = \frac{p_{20}}{\lambda + \theta_2 - s\lambda}\left(\lambda + \frac{\theta_2(1-s^k)}{1-s}\right) \tag{24}$$

Multiplying the 5th, 6th and 7th equations of (1) by s^j and summing over $j \in \{1, 2, 3, \ldots\}$ gives

$$(\lambda + \mu)\sum_{j=1}^{\infty} p_{0j} = s(\lambda p_1 + \mu p_{02}) + \sum_{j=2}^{\infty} s^j(\lambda p_{0,j-1} + \mu p_{0,j+1}) + \sum_{j=K}^{\infty} s^j \theta_2 p_{2j}$$

or

$$(\lambda + \mu)G_0(s) = (\lambda s + \mu/s)G_0(s) + s\lambda p_1 - \mu p_{01} + \theta_2 \sum_{j=K}^{\infty} s^j p_{2j}$$

This, together with (23) and the 4th and 1st equations of (1), gives

$$\begin{aligned}
(\lambda + \mu - \lambda s - \mu/s)G_0(s) &= s\lambda p_{10} - (\lambda + \theta_1)p_1 + \frac{\theta_2 p_{20} s^K \lambda}{\lambda + \theta_2 - s\lambda} \\
&= (s-1)\lambda p_1 - \lambda p_{20} + \frac{\theta_2 p_{20} s^K \lambda}{\lambda + \theta_2 - s\lambda} \\
&= (s-1)\lambda p_1 - \frac{\lambda p_{20}}{\lambda + \theta_2 - s\lambda}\left(\lambda + \theta_2 - s\lambda - \theta_2 s^K\right) \\
&= (s-1)\left(\lambda p_1 + \frac{\lambda p_{20}}{\lambda + \theta_2 - s\lambda}\left(\lambda + \theta_2 \frac{1-s^K}{1-s}\right)\right)
\end{aligned}$$

Dividing through by $(s-1)$ and using (24), we can write

$$G_0(s)(\mu/s - \lambda) = \lambda\left(p_1 + G_2(s)\right)$$

so that

$$G_0(s) = \frac{\lambda s\left(p_1 + G_2(s)\right)}{\mu - s\lambda} \tag{25}$$

In terms of the traffic intensity, this can be written as

$$G_0(s) = \frac{\rho s\left(p_1 + G_2(s)\right)}{1 - s\rho}$$

Expanding in geometric series gives

$$G_0(s) = \sum_{j=1}^{\infty}(\rho s)^j p_1 + G_2(s)\sum_{j=1}^{\infty}(s\rho)^j$$

and expanding $G_2(s)$ further gives

$$G_0(s) = \sum_{j=1}^{\infty}(\rho s)^j p_1 + \sum_{j=1}^{\infty}\sum_{i=0}^{\infty} s^{i+j}\rho^j p_{2i}$$

from which we can see that

$$p_{0j} = p_1\rho^j + \sum_{i=0}^{j-1}\rho^{j-i}p_{2i} \qquad (j = 1, 2, 3, \ldots) \tag{26}$$

It remains for us only to determine p_1. Letting $s \to 1$ in (24) and (25), we obtain

$$\theta_2 G_2(1) = (\lambda + K\theta_2)p_{20}$$

and

$$(\mu - \lambda)G_0(1) = \lambda G_2(1) + \lambda p_1$$

By using the 1st and 8th equations of (1), it follows that

$$p_1 = \frac{\lambda\theta_2(1-\rho)}{\lambda\theta_1 + \lambda\theta_2 + K\theta_1\theta_2} \tag{27}$$

Thus, the equilibrium distribution of $\{(N(t), S(t)), t \geq 0\}$ is given by (21), (22), (23), (26) and (27).

A.2 Queue Length Distribution in the Deterministic Case

Defining $G_0(s)$ and $G_2(s)$ as for the exponential case, we can again obtain (25) and (24) where we now define

$$p_1 = \int_0^{1/\theta_1} p_1(x)dx$$

These relations determine $G_0(s)$ and $G_2(s)$ in terms of p_{20} and p_1. Thus, (21), (23) and (26) again hold, but with

$$p_{20} = \frac{(1-\rho)\theta_2}{\lambda - \theta_2 + e^{\lambda/\theta_1}\theta_2 + K\theta_2}$$

and

$$p_1 = \frac{(e^{\lambda/\theta_1} - 1)(1-\rho)\theta_2}{\lambda - \theta_2 - e^{\lambda/\theta_1}\theta_2 - K\theta_2}$$

A.3 Derivation of Mean Queue Length

Let N and Q denote the numbers of customers in the system and in the queue, respectively, assuming the system is in equilibrium. Also, let $G_N(s)$ and $G_Q(s)$ be the probability generating functions of N and Q, respectively.

Clearly,

$$G_N(s) = G_0(s) + G_2(s) + p_1 \qquad (|s| \leq 1) \tag{28}$$

It is also easy to see that

$$G_Q(s) = G_2(s) + p_1 - p_{01} + G_0(s)/s$$

Using (25) and the 4th equation of (1), we have

$$G_Q(s) = G_2(s) + p_1 + \frac{\lambda s}{s(\mu - \lambda s)}\left(G_2(s) + p_1\right) - \frac{\lambda + \theta_1}{\mu} p_1$$

which, upon rearrangement, yields

$$G_Q(s) = \frac{\mu + \lambda - \lambda s}{\mu - \lambda s} G_2(s) + p_1 \left(\frac{\rho}{1 - \rho s} + \frac{\mu - \lambda - \theta_1}{\mu} \right) \tag{29}$$

The means $L = E[N]$ and $L_q = E[Q]$ are obtained by evaluating the derivatives $G'_N(1)$ and $G'_Q(1)$. To do this, we note that (24) can be differentiated easily to give

$$G'_2(1) = p_{20} \left(\frac{K(K-1)}{2} + \frac{\lambda^2 + K\lambda\theta_2}{\theta_2^2} \right) \tag{30}$$

Using (25), we can readily obtain

$$G'_0(1) = \frac{\lambda\mu}{(\mu - \lambda)^2}\left(G_2(1) + p_{10}\right) + \frac{\lambda}{\mu - \lambda} G'_2(1) \tag{31}$$

Since

$$L = G'_0(1) + G'_2(1),$$

we can use (30) and (31) to obtain

$$L = \frac{\lambda\mu}{(\mu - \lambda)^2} \left(p_1 + \frac{(\lambda + K\theta_2)p_{20}}{\theta_2} \right) + \frac{\mu p_{20}}{\mu - \lambda} \left(\frac{K(K-1)}{2} + \frac{\lambda^2 + K\lambda\theta_2}{\theta_2^2} \right)$$

From this, we can obtain the special cases of exponential and deterministic delay displayed in equations (6), (7), (8) and (9).

A.4 Waiting Time Analysis

Let W_q denote the virtual waiting time of customers arriving in the system, assuming the system is in equilibrium, and let $W_q(x)$ denote the corresponding distribution function. Then

$$W_q(x) = P(W_q \leq x, S = 2, N < K) + P(W_q \leq x, S = 2, N \geq K) + P(W_q \leq x, S = 0) + p_1 \tag{32}$$

Denote the first 3 terms on the right hand side by $W_{21}(x)$, $W_{22}(x)$ and $W_{01}(x)$, respectively. We will evaluate $E[W_q]$ using

$$E[W_q] = -W^{*\prime}(0) = -\left(W_{21}^{*\prime}(0) + W_{22}^{*\prime}(0) + W_{10}^{*\prime}(0)\right)$$

where

$$W^{*\prime}(0) = \lim_{s \to 0} W^{*\prime}(s)$$

and $W^*(s)$ denotes the Laplace-Stieltjes (L-S) transform of $W(x)$.

If a customer arrives at the system during the server's vacation and sees $N = j$ $(< K)$ customers in the system, then the waiting time in the queue will be the sum of $K - (j+1)$ interarrivals, a residual vacation and j service times. Because of independence and the memoryless property of the exponential distribution, we have

$$\begin{aligned}
W_{21}^{*\prime}(s) &= \int_0^\infty e^{-sx} dW_{21}(x) \\
&= \sum_{j=0}^{K-1} p_{2j} \left(\frac{\lambda}{\lambda + s}\right)^{K-(j+1)} \left(\frac{\theta_2}{\theta_2 + s}\right) \left(\frac{\mu}{\mu + s}\right)^j \\
&= p_{20} \left(\frac{\theta_2}{\theta_2 + s}\right) \sum_{j=0}^{K-1} \left(\frac{\lambda}{\lambda + s}\right)^{K-(j+1)} \left(\frac{\mu}{\mu + s}\right)^j \\
&= p_{20} \left(\frac{\theta_2}{\theta_2 + s}\right) \frac{(\lambda + s)(\mu + s)}{s(\lambda - \mu)} \left(\left(\frac{\lambda}{\lambda + s}\right)^K - \left(\frac{\mu}{\mu + s}\right)^K\right) \qquad (33)
\end{aligned}$$

If a customer arrives in the system during vacation and sees $N = j$ $(\geq K)$ customers in the system, then the waiting time in the queue is the sum of a residual vacation and j service times. Thus,

$$W_{22}^{*\prime}(s) = \int_0^\infty e^{-sx} dW_{22}(x)$$

$$\begin{aligned} &= \sum_{j=K}^{\infty} p_{2j}\left(\frac{\theta_2}{\theta_2+s}\right)\left(\frac{\mu}{\mu+s}\right)^j \\ &= p_{20}\left(\frac{\theta_2}{\theta_2+s}\right)\sum_{j=K}^{\infty}\left(\frac{\lambda}{\lambda+\theta_2}\right)^{(j-K+1)}\left(\frac{\mu}{\mu+s}\right)^j \\ &= p_{20}\left(\frac{\theta_2}{\theta_2+s}\right)\left(\frac{\mu}{\mu+s}\right)^K\frac{1}{\frac{\lambda+\theta_2}{\lambda}-\frac{\mu}{\mu+s}} \qquad (34)\end{aligned}$$

Similarly, if a customer arrives in the system during a busy period and encounters $N = j(\geq 1)$ customers in the system, then the waiting time in the queue is the sum of a residual service time and $j-1$ service times. It then follows that

$$W_{01}^*(s) = \int_0^\infty e^{-sx}dW_{01}(x) = \sum_{j=1}^{\infty} p_{0j}\left(\frac{\mu}{\mu+s}\right)^j$$

Using (26) we see that

$$W_{01}^*(s) = \sum_{j=1}^{\infty}\{p_1\rho^j + \sum_{i=0}^{j-1} p_{2i}\rho^{j-i}\}\left(\frac{\mu}{\mu+s}\right)^j$$

Using (22) and (23) and several algebraic manipulations gives

$$W_{01}^*(s) = p_{20}\frac{\lambda}{\mu+s-\lambda}\left\{\frac{p_1}{p_{20}} + \frac{1-\left(\frac{\mu}{\mu+s}\right)^K}{1-\left(\frac{\mu}{\mu+s}\right)} + \left(\frac{\mu}{\mu+s}\right)^K\frac{\lambda(\mu+s)}{(\lambda+\theta_2)(\mu+s)-\lambda\mu}\right\} \qquad (35)$$

Combining (33), (34) and (35), we obtain

$$\begin{aligned} W^*(s) = p_{20}&\left\{\frac{\theta_2}{\theta_2+s}\left(\frac{\left(\frac{\lambda}{\lambda+s}\right)^K-\left(\frac{\mu}{\mu+s}\right)^K}{\frac{\lambda}{\lambda+s}-\frac{\mu}{\mu+s}} + \frac{\left(\frac{\mu}{\mu+s}\right)^K}{\frac{\lambda+\theta_2}{\lambda}-\frac{\mu}{\mu+s}}\right)\right. \\ &\left. + \frac{\lambda}{\mu+s-\lambda}\left(\frac{p_1}{p_{20}} + \frac{1-\left(\frac{\mu}{\mu+s}\right)^K}{1-\frac{\mu}{\mu+s}} + \frac{\left(\frac{\mu}{\mu+s}\right)^K}{\frac{\lambda+\theta_2}{\lambda}-\frac{\mu}{\mu+s}}\right) + \frac{p_1}{p_{20}}\right\} \qquad (36)\end{aligned}$$

Hence, we have (10).

The probability that a customer arriving in the system has to wait for service is given by

$$1 - p_1 \qquad (37)$$

Hence, given $W_q > 0$, the conditional distribution function $W_q(x)$ is given by

$$P(W_q \leq x | W_q > 0) = \frac{(W_{21}(x) + W_{22}(x) + W_{01}(x)}{1-p_1} \qquad (38)$$

Then, in view of (32) and (10), we obtain (11).

A.4 Analysis of the Operational Period

We consider the problem of obtaining the equation (17) for $E[B_R]$, the expected length of a busy period beginning with R customers in the system. First, note that

$$P(R = r) = \left(\sum_{j=K}^{\infty} p_{2j} \right)^{-1} p_{2r} \qquad (r = K, K+1, K+2, \ldots)$$

In view of (23), we have

$$P(R = r) = \frac{\theta_2}{\lambda} \left(\frac{\lambda}{\lambda + \theta_2} \right)^{r-K+1} \tag{39}$$

Because of (16), we can condition on R, and write

$$\begin{aligned} E[B_R] &= \sum_{r=K}^{\infty} E[B_R | R = r] P(R = r) \\ &= \sum_{r=K}^{\infty} E[B_r] P(R = r) \\ &= \sum_{r=K}^{\infty} \frac{r}{\mu - \lambda} \frac{\theta_2}{\lambda} \left(\frac{\lambda}{\lambda + \theta_2} \right)^{r-K+1} \end{aligned}$$

from which (17) follows immediately.

We turn next to the derivation of the equation (18) for $E[V_K]$, the expected length of vacation.

Let X, Z and S be random variables which are independent and exponentially distributed with mean $1/\lambda$, $1/\theta_2$ and $1/\mu$, respectively.

We use a recurrence method as follows:

$$\begin{aligned} E[V_K] &= E[V_K | Z < X] P(Z < X) + E[V_K | X \leq Z] P(X \leq Z) \\ &= E[Z + V'_K | Z < X] P(Z < X) + E[X + V'_{K-1} | X \leq Z] P(X \leq Z) \\ &= E[Z | Z < X] P(Z < X) + E[X | X \leq Z] P(X \leq Z) \\ &\quad + E[V'_K] P(Z < X) + E[V'_{K-1}] P(X \leq Z) \\ &= \frac{1}{\lambda + \theta_2} + \frac{\theta_2}{\lambda + \theta_2} E[V_K] + \frac{\lambda}{\lambda + \theta_2} E[V_{K-1}] \end{aligned} \tag{40}$$

where V'_K and V'_{K-1} have the same distributions as V_K and V_{K-1}, respectively, but they are independent of X and Z. We can then write

$$E[V_K] = \frac{1}{\lambda} + E[V_{K-1}] \tag{41}$$

To obtain $E[V_1]$, we note that

$$\begin{aligned} E[V_1] &= E[V_1|Z < X]P(Z < X) + E[V_1|X \leq Z]P(X \leq Z) \\ &= E[Z + V_1'|Z < X]P(Z < X) + E[Z|X \leq Z]P(X \leq Z) \\ &= E[Z] + E[V_1]\frac{\theta_2}{\lambda + \theta_2} \\ &= \frac{1}{\lambda} + \frac{1}{\theta_2} \qquad (42) \end{aligned}$$

Combining (41) and (42), we obtain (18).

Acknowledgement

This work was supported by research grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

[1] Alam, S.S., Acharya, D., and Rao, V.P. (1986) M/M/1 queue with server's vacation. *Asia-Pacific J. Oper. Res.* **3** 21–26.

[2] Chao, X. and Zhao, Y.Q. (1997) Analysis of multi-server queues with station and server vacations. (to appear in EJOR).

[3] Chung, K.L. (1967) *Markov Chains with Stationary Transition Probabilities.* 2nd edition. Springer-Verlag, Berlin.

[4] Chaudhry, M.L. and Templeton, J.G.C. (1983) *A First Course in Bulk Queues.* Wiley, New York.

[5] Cox, D. R. (1955) The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Proceedings of the Cambridge Philosophical Society* **51** 433–441.

[6] Doshi, B.T. (1986) Queueing system with vacations - a survey. *Queueing Systems* **1** 29–66.

[7] Frey, A. and Takahashi, Y. (1997) A Note on an M/GI/1/N Queue with Vacation Time and Exhaustive Service Discipline. (to appear in *Operations Research Letters*).

[8] Frey, A. and Takahashi, Y. (1997) An M^X/GI/1/N Queue with Close-down and Vacation Times. (submitted for publication)

[9] Sakai, Y, Takahashi, Y., Takahashi, Y., and Hasegawa, T. (1997) A Composite Queue with Vacation/Set-up/Close-down Times for SVCC in IP over ATM Networks. (submitted for publication)

[10] Shi, D.H. and Zhang, W.G. (1994) Analysis of the repairable queueing system M^X/G(M/G)/1(M/G) with multiple delay vacations. *Acta Mathematica Applicatae Sinica* **17** 201–214.

[11] Takagi, H. (1991) *Queueing Analysis Vol. 1 Vacation and Priority Systems.* North-Holland, Amsterdam.

[12] Tijms, H. (1994) *Stochastic Modelling – An Algorithmic Approach.*