Hindawi Publishing Corporation
International Journal of Stochastic Analysis
Volume 2012, Article ID 145867, 16 pages
doi:10.1155/2012/145867

Research Article

An $M/M/2$ Queueing System with Heterogeneous Servers Including One with Working Vacation

A. Krishnamoorthy[1] and C. Sreenivasan[2]

[1] *Department of Mathematics, Cochin University of Science and Technology, Cochin 682022, India*
[2] *Department of Mathematics, Government College, Chittur, Palakkad 678104, India*

Correspondence should be addressed to A. Krishnamoorthy, achyuthacusat@gmail.com

Received 24 February 2012; Accepted 13 May 2012

Academic Editor: Ho Lee

This paper analyzes an $M/M/2$ queueing system with two heterogeneous servers, one of which is always available but the other goes on vacation in the absence of customers waiting for service. The vacationing server, however, returns to serve at a low rate as an arrival finds the other server busy. The system is analyzed in the steady state using matrix geometric method. Busy period of the system is analyzed and mean waiting time in the stationary regime computed. Conditional stochastic decomposition of stationary queue length is obtained. An illustrative example is also provided.

1. Introduction

Queueing models with vacation have gained significance in the last three decades due to their wide range of applications, especially in the communication and the manufacturing systems. Doshi [1] provides an excellent survey of related works prior to 1986. Takagi [2] and Tian and Zhang [3] provide a good account of developments in this field since then. The literature on the vacation queueing models is growing rapidly.

In multiserver queueing models, we come across two classes of vacation mechanisms: station vacation and server vacation. In the first case, all servers take vacation simultaneously whenever the system becomes empty and they return to the system all together. Thus, station vacation is group vacation for all servers. For example, when a system consists of a number of machines operated by a single individual this scenario occurs. In such a situation, the whole station needs to be treated as a single unit for vacation, when the system is utilized for a secondary task. In the second case, each server takes its own vacation whenever it completes a service and finds no customers waiting for service. This phenomenon also occurs in practice. For example, in a post office or bank, when a clerk completes a service and finds no customer waiting, he or she might go to attend another task. This is what we refer to as the server vacation model. Analysis of a server vacation model is more complicated than that of a station vacation model. This is because at any time point, in the latter, we may have any number of servers between 0 and c on vacation. We need to track individual servers going on vacation and completing their vacation. Upon returning from a vacation some servers may find no customers waiting for service. These servers take another vacation. But if any server finds a waiting customer on returning from a vacation, it immediately starts service. For further details on queues with station and server vacations, we refer the reader to Chao and Zhao [4].

Most of the earlier work on multiserver queueing models deal with homogeneous servers; that is, the individual service rates are same for all servers in the system. But this assumption may be valid only when the service process is mechanically or electronically controlled. In a queueing system with human servers the above assumption is highly unrealistic. Often servers providing identical service, serves at different rates. This motivated the researchers to study multiserver queueing system with heterogeneous servers. Several authors have analyzed multiserver queues with vacation. Levy and Yechiali [5], Vinod [6], and Kao and Narayanan [7] discuss asynchronous multiple vacation models with exponentially distributed vacation times. They study the scenario, where any server goes on vacation whenever there are no customers waiting in the system at a service completion epoch. At a vacation termination instant, if there are no waiting customers, the server takes another vacation; if there is a customer waiting for service, the server resumes service. It should be remarked that Vinod [6] was the first to use matrix geometric solutions method to analyze multiserver vacation models. But none of the above models deal with heterogeneous servers.

Servi and Finn [8] introduced a working vacation model with the idea of offering services but at a lower rate, whenever the server is on vacation. Their model was generalized to the case of $M/G/1$ in [9, 10] and to $GI/M/1$ model in [11]. A survey of working vacation models with emphasis on the use of matrix analytic methods is given in Tian et al. [12]. Working vacation models have a number of applications in practice. Two such examples are given in [12]. Recently, Li and Tian [13] studied an $M/M/1$ queue with working vacations in which vacationing server offers services at a lower rate, for the first customer arriving during a vacation. Very recently, Zhang and Hou [14] studied a $MAP/G/1$ queue with working vacations and vacation interruptions using supplementary variable method. In this model, the authors assume that the vacation times are exponentially distributed and that the server gets back to normal service mode, when at a service (offered during a vacation) completion, the system has at least one customer waiting in the queue. The server is allowed to take multiple vacations. However, no work on multiserver working vacation model has come to our notice.

Neuts and Takahashi [15] observed that for queueing systems with more than two heterogeneous servers analytical results are intractable and only algorithmic approach could be used to study the steady state behavior of the system. Based on this observation, Krishna Kumar and Pavai Madheswari [16] analyzed $M/M/2$ queueing system with heterogeneous servers, where the servers go on vacation in the absence of customers waiting for service. In this paper, we discuss an $M/M/2$ queueing model with heterogeneous servers where one server remains idle but the other goes on working vacation in the absence of waiting customers.

This paper is organized as follows. In Section 2, model description is provided. In Section 3, the steady state analysis of the model is presented. In Section 4, we discuss an illustrative example.

2. Mathematical Model

We consider an $M/M/2$ queueing model with heterogeneous servers, server 1 and server 2. Server 1 is always available, whereas server 2 goes on vacation whenever there are no customers waiting for service. Let the service rates of servers 1 and 2 be μ_1 and μ_2, respectively, where $\mu_1 \neq \mu_2$. Customers arrive to the system according to a Poisson process with parameter λ. The duration of vacation is exponentially distributed with parameter η. At the end of a vacation, service commences if there is a customer waiting for service. Otherwise the server goes on another vacation. During vacation, if an arrival finds server 1 busy, server 2 returns to serve the customer but at a lower rate. To be precise, server 2 serves this customer at the rate $\theta\mu_2$, $0 < \theta \leq 1$. As this vacation gets over, server 2 instantaneously switches over to the normal service rate μ_2, if there is at least one customer waiting for service. Upon completion of a service at low rate, the server will (a) continue the current vacation if it is not finished and no customer is waiting for service; (b) continue the slow service if the vacation has not expired and if there is at least one customer waiting for service. For clarity, we make it clear that if an arriving customer finds a free server, he enters service immediately. Else he joins the queue.

2.1. The QBD Process

The model discussed in Section 2 can be studied as a quasi-birth-and-death (QBD) process. First, we set up the necessary notations.

At time t, let $N(t)$ be the number of customers in the system and

$$J(t) = \begin{cases} 0, & \text{if the server 2 is on vacation,} \\ 1, & \text{if the server 2 is working in vacation mode,} \\ 2, & \text{if the server 2 is working in normal mode,} \end{cases} \tag{2.1}$$

Let $X(t) = (N(t), J(t))$. Then $(X(t) : t \geq 0)$ is a continuous time Markov chain (CTMC) with states space

$$\Omega = \{(0,0),(1,0),(1,1),(1,2)\} \bigcup \bigcup_{i=2}^{\infty} l(i), \tag{2.2}$$

where

$$l(i) = \{(i,j) : i \geq 2,\ j = 1 \text{ or } 2\}. \tag{2.3}$$

Note that when $N(t) = 0$, the only possible value of $J(t)$ is 0 and when $N(t) = 1$, $J(t)$ has three possible values 0, 1, and 2.

The infinitesimal generator matrix Q of this Markov chain is given by

$$Q = \begin{bmatrix} B_{00} & B_{01} & & & & \\ B_{10} & B_{11} & B_{12} & & & \\ & B_{21} & A_1 & A_0 & & \\ & & A_2 & A_1 & A_0 & \\ & & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{2.4}$$

where the block matrices appearing in Q are as follows:

$$B_{00} = -\lambda, \qquad B_{01} = \begin{bmatrix} \lambda & 0 & 0 \end{bmatrix},$$

$$B_{10} = \begin{bmatrix} \mu_1 \\ \theta\mu_2 \\ \mu_2 \end{bmatrix}, \qquad B_{11} = \begin{bmatrix} -\lambda-\mu_1 & 0 & 0 \\ 0 & -\lambda-\theta\mu_2-\eta & \eta \\ 0 & 0 & -\lambda-\mu_2 \end{bmatrix}, \qquad B_{12} = \begin{bmatrix} \lambda & 0 \\ \lambda & 0 \\ 0 & \lambda \end{bmatrix},$$

$$B_{21} = \begin{bmatrix} \theta\mu_2 & \mu_1 & 0 \\ \mu_2 & 0 & \mu_1 \end{bmatrix}, \qquad A_0 = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \tag{2.5}$$

$$A_1 = \begin{bmatrix} -\lambda-\mu_1-\theta\mu_2-\eta & \eta \\ 0 & -\lambda-\mu_1-\mu_2 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} \mu_1+\theta\mu_2 & 0 \\ 0 & \mu_1+\mu_2 \end{bmatrix}.$$

3. Steady State Analysis

In this section, we will discuss the steady state analysis of the model under study.

3.1. Stability Condition

Theorem 3.1. *The queueing system described above is stable if and only if $\rho < 1$ where $\rho = \lambda/(\mu_1 + \mu_2)$.*

Proof. To establish the stability condition we use Pakes' lemma (see [17]). Let N_i be the number of customers in the system immediately after the departure of the ith customer. Then $\{N_i : i \in \mathbb{N}\}$ satisfies the equation

$$N_i = \begin{cases} N_{i-1} - 1 + V_i & \text{if } N_{i-1} \geq 1 \\ V_i & \text{if } N_{i-1} = 0, \end{cases} \tag{3.1}$$

where V_i is the number of arrivals during the service of ith customer. Clearly $\{N_i : i \in \mathbb{N}\}$ is an irreducible aperiodic Markov chain. Pakes' lemma asserts that an aperiodic irreducible Markov chain is ergodic, if there exists an $\epsilon > 0$ such that the mean drift

$$\phi_j = E\left[(N_{i+1} - N_i) \mid N_i = j\right] \tag{3.2}$$

is finite for all $j \in \mathbb{N}$ and $\phi_j \le -\epsilon$ for all $j \in \mathbb{N}$ except perhaps for a finite number. In the present model, value of the mean drift is

$$\phi_j = \begin{cases} -1+\rho & \text{if } j \ge 1 \\ \rho & \text{if } j = 0. \end{cases} \tag{3.3}$$

Thus, if $\rho < 1$, the Markov chain $\{N_i : i \in \mathbb{N}\}$ is ergodic and hence the condition is sufficient.

To prove the necessity of the condition, assume that $\rho \ge 1$. We use Theorem 3.1 in Sennot et al. [18], which states that $\{N_i : i \in \mathbb{N}\}$ is nonergodic if it satisfies Kaplan's condition, $\phi_j < \infty$, for $j \ge 0$ and there is a j_0 such that $\phi_j \ge 0$, for $j \ge j_0$. When $\rho \ge 1$ Kaplan's condition is readily satisfied. Hence, the Markov chain is not ergodic. □

3.2. Steady State Probability Vector

Let $\mathbf{x}$, partitioned as $\mathbf{x} = (x_0, \mathbf{x}_1, \mathbf{x}_2, \ldots)$, be the steady state probability vector of Q. Note that x_0 is a scalar, $\mathbf{x}_1 = (x_{10}, x_{11}, x_{12})$ and $\mathbf{x}_i = (x_{i1}, x_{i2})$ for $i \ge 2$. The vector $\mathbf{x}$ satisfies the condition $\mathbf{x}Q = \mathbf{0}$ and $\mathbf{xe} = 1$, where $\mathbf{e}$ is a column vector of 1's of appropriate dimension. Apparently when the stability condition is satisfied, the subvectors of $\mathbf{x}$, corresponding to the different levels, are given by the equation $\mathbf{x}_j = \mathbf{x}_2 R^{j-2}$, $j \ge 3$, where R is the minimal nonnegative solution of the matrix quadratic equation (see [20])

$$R^2 A_2 + RA_1 + A_0 = 0. \tag{3.4}$$

Knowing the matrix R, x_0, $\mathbf{x}_1$ and $\mathbf{x}_2$ are obtained by solving the equations

$$x_0 B_{00} + \mathbf{x}_1 B_{10} = 0, \tag{3.5}$$

$$x_0 B_{01} + \mathbf{x}_1 B_{11} + \mathbf{x}_2 B_{21} = 0, \tag{3.6}$$

$$\mathbf{x}_1 B_{12} + \mathbf{x}_2 (A_1 + RA_2) = 0, \tag{3.7}$$

subject to the normalizing condition

$$x_0 + \mathbf{x}_1\mathbf{e} + \mathbf{x}_2 (I - R)^{-1}\mathbf{e} = 1, \tag{3.8}$$

Theorem 3.2. *The matrix R of* (3.4) *is given by $R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$, where $R_{11} = (\lambda + \mu_1 + \theta\mu_2 + \eta - \sqrt{(\lambda + \mu_1 + \theta\mu_2 + \eta)^2 - 4\lambda(\mu_1 + \theta\mu_2)})/2(\mu_1 + \theta\mu_2)$, $R_{12} = \rho - (\mu_1 + \theta\mu_2)R_{11}/(\mu_1 + \mu_2)$ and $R_{22} = \rho$.*

Proof. Since A_0, A_1, and A_2 are upper triangular, R is essentially an upper triangular matrix. The value of R_{11}, follows from the assertion that R is the minimal nonnegative solution of (3.4). The rest of the proof is an easy consequence of the condition $RA_2\mathbf{e} = A_0\mathbf{e}$. □

Remark 3.3. Though R has a nice structure which enables us to make use of the properties like $R^k = \begin{bmatrix} R_{11}^k & R_{12}\sum_{j=0}^{k-1} R_{11}^j R_{22}^{k-j-1} \\ 0 & R_{22}^k \end{bmatrix}$, for $k \ge 1$, due to the form of the expression for R_{11} it may not be easy to carry out the computations required in the forthcoming discussions. Hence, we explore the possibility of algorithmic computation of R. The computation of R matrix can be carried out using a number of well-known methods such as logarithmic reduction. We list here only the main steps involved in logarithmic reduction algorithm for computation of R. For full details on the logarithmic reduction algorithm we refer the reader to [19].

Logarithmic Reduction Algorithm for R:

Step 0. $H \leftarrow (-A_1)^{-1}A_0$, $L \leftarrow (-A_1)^{-1}A_2$, $G = L$, and $T = H$.

Step 1.

$$\begin{gathered} U = HL + LH \\ M = H^2 \\ H \longleftarrow (I - U)^{-1}M \\ M \longleftarrow L^2 \\ L \longleftarrow (I - U)^{-1}M \\ G \longleftarrow G + TL \\ T \longleftarrow TH, \end{gathered} \tag{3.9}$$

continue Step 1 until $\|\mathbf{e} - G\mathbf{e}\|_\infty < \epsilon$.

Step 2. $R = -A_0(A_1 + A_0G)^{-1}$.

3.3. Busy Period Analysis

For the system under study, busy period is the interval between arrival of a customer to the empty system and the first epoch thereafter when the system becomes empty again. Thus, it is precisely the first passage time from the state $(1, 0)$ to the state $(0, 0)$. For the working vacation model, busy cycle for the system is the time interval between two successive departures, which leave the system empty. Thus, the busy cycle is the first return time to state $(0, 0)$ with at least one visit to any other state. But before analyzing the busy period structure we need to introduce the notion of fundamental period. For the QBD process under consideration, it is the first passage time from level i, where $i \geq 3$, to the level $i - 1$. The cases $i = 2$, $i = 1$, and $i = 0$, corresponding to the boundary states, need to be discussed separately. It should be noted that due to the structure of the QBD process, the distribution of the first passage time is invariant in i.

Let $G_{jj'}(k, x)$ denote the conditional probability that a QBD process starting in the state (i, j) at time $t = 0$ reaches the level $i - 1$ for the first time no later than time x, after exactly k transitions to the left and does so by entering the state $(i-1, j')$. For convenience we introduce the joint transform

$$\tilde{G}_{jj'}(z, s) = \sum_{k=1}^{\infty} z^k \int_0^{\infty} e^{-sx} \mathrm{d}G_{jj'}(k, x); \quad |z| \leq 1,\ \mathrm{Re}(s) \geq 0, \tag{3.10}$$

and the matrix

$$\tilde{G}(z,s) = \left(\tilde{G}_{jj'}(z,s)\right). \tag{3.11}$$

The matrix $\tilde{G}(z,s)$ is the unique solution to the equation (see [20])

$$\tilde{G}(z,s) = z(sI - A_1)^{-1} A_2 + (sI - A_1)^{-1} A_0 \tilde{G}^2(z,s). \tag{3.12}$$

The matrix $G = \tilde{G}(1,0)$ takes care of the first passage times, except for the boundary states. If we know the R matrix then G matrix can be computed using the result (see [19])

$$G = -(A_1 + RA_2)^{-1} A_2. \tag{3.13}$$

Otherwise, we may use logarithmic reduction method to compute G. For the boundary level states 2, 1, and 0, let $G_{jj'}^{(2,1)}(k,x)$, $G_{jj'}^{(1,0)}(k,x)$, and $G_{jj'}^{(0,0)}(k,x)$ be the conditional probability discussed above for the first passage times from level 2 to level 1, level 1 to level 0, and the first return time to the level 0, respectively. Then as in (3.12) we get

$$\tilde{G}^{(2,1)}(z,s) = z(sI - A_1)^{-1} B_{21} + (sI - A_1)^{-1} A_0 \tilde{G}(z,s) \tilde{G}^{(2,1)}(z,s), \tag{3.14}$$

$$\tilde{G}^{(1,0)}(z,s) = z(sI - B_{11})^{-1} B_{10} + (sI - B_{11})^{-1} B_{12} \tilde{G}^{(2,1)}(z,s) \tilde{G}^{(1,0)}(z,s), \tag{3.15}$$

$$\tilde{G}^{(0,0)}(z,s) = \left[\frac{\lambda}{s+\lambda} \quad 0 \quad 0\right] \tilde{G}^{(1,0)}(z,s). \tag{3.16}$$

Note that $\tilde{G}^{(1,0)}(z,s)$ is a 3×1 matrix. Thus, the Laplace Stieltjes transform (LST) of the busy period is the first element of $\tilde{G}^{(1,0)}(1,s)$. For convenience, we use the notations

$$G_{21} = \tilde{G}^{(2,1)}(1,0), \qquad G_{10} = \tilde{G}^{(1,0)}(1,0), \qquad G_{00} = \tilde{G}^{(0,0)}(1,0). \tag{3.17}$$

Due to the positive recurrence of the QBD process, matrices G, G_{21}, G_{10}, and G_{00} are all stochastic. If we let

$$C_0 = (-A_1)^{-1} A_2, \qquad C_2 = (-A_1)^{-1} A_0, \tag{3.18}$$

then G is the minimal nonnegative solution (see [20]) to the matrix equation

$$G = C_0 + C_2 G^2. \tag{3.19}$$

From (3.14), (3.15), and (3.16), we get

$$\begin{gathered}
G_{21} = -(A_1 + A_0 G)^{-1} B_{21}, \\
G_{10} = -(B_{11} + B_{12} G_{21})^{-1} B_{10}, \\
G_{00} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} G_{10}
\end{gathered} \tag{3.20}$$

respectively. Equation (3.12) is equivalent to

$$zA_2 - (sI - A_1)\widetilde{G}(z,s) + A_0 \widetilde{G}^2(z,s) = 0. \tag{3.21}$$

Let

$$\begin{gathered}
M = -\left.\frac{\partial \widetilde{G}(z,s)}{\partial s}\right|_{z=1,s=0}, \\
\widetilde{M} = \left.\frac{\partial \widetilde{G}(z,s)}{\partial z}\right|_{z=1,s=0}.
\end{gathered} \tag{3.22}$$

Differentiation of (3.21) with respect to s and z followed by setting $z = 1$ and $s = 0$ leads to (see [20])

$$\begin{gathered}
M = -A_1^{-1} G + C_2 (GM + MG), \\
\widetilde{M} = C_0 + C_2 \left(G\widetilde{M} + \widetilde{M} G \right),
\end{gathered} \tag{3.23}$$

with $\mathbf{0}$ as starting value for M and $\widetilde{M}$, successive substitutions in the above equations yield the values of M and $\widetilde{M}$. Applying an exactly similar reasoning to (3.14), (3.15), and (3.16), we get

$$\begin{gathered}
M_{21} = -(A_1 + A_0 G)^{-1} (I + A_0 M) G_{21}, \\
M_{10} = (B_{11} + B_{12} G_{21})^{-1} (I + B_{12} M_{21}) G_{10}, \\
M_{00} = \begin{bmatrix} \frac{1}{\lambda} & 0 & 0 \end{bmatrix} G_{10} + \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} M_{10},
\end{gathered} \tag{3.24}$$

where

$$M_{21} = -\frac{\partial \widetilde{G}^{(2,1)}(z,s)}{\partial s}\bigg|_{z=1,s=0},$$
$$M_{10} = -\frac{\partial \widetilde{G}^{(1,0)}(z,s)}{\partial s}\bigg|_{z=1,s=0}, \tag{3.25}$$
$$M_{00} = -\frac{\partial \widetilde{G}^{(0,0)}(z,s)}{\partial s}\bigg|_{z=1,s=0}.$$

Note that M_{10} is a 3×1 matrix and M_{00} is a scalar. The first element of the matrix M_{10} and M_{00} are mean lengths of a busy period and a busy cycle, respectively. The second and third elements of the matrix M_{10} are the first passage times to the state $(0,0)$ from $(1,1)$ and $(1,2)$, respectively. With the notations

$$\widetilde{M}_{21} = \frac{\partial \widetilde{G}^{(2,1)}(z,s)}{\partial z}\bigg|_{z=1,s=0},$$
$$\widetilde{M}_{10} = \frac{\partial \widetilde{G}^{(1,0)}(z,s)}{\partial z}\bigg|_{z=1,s=0}, \tag{3.26}$$

it follows from (3.14) and (3.15) that

$$\widetilde{M}_{21} = -(A_1 + A_0 G)^{-1}(B_{21} + A_0 M G_{21}),$$
$$\widetilde{M}_{10} = -(B_{11} + B_{12} G_{21})^{-1}(B_{10} + B_{12} M_{21} G_{10}). \tag{3.27}$$

The first component of the vector $\widetilde{M}_{10}$ is the mean number of service completions in a busy period.

3.4. Stationary Waiting Time Distribution in the Queue

Let $W(t)$ be the distribution function for the waiting time in the queue of an arriving (tagged) customer. Note that if there is no customer in the system, the arrival receives service immediately. If either of the two servers is not busy then also there would be no delay in getting service. Thus, the probability that the customer gets his service without waiting is $x_0 + x_{10} + x_{11} + x_{12}$. Hence, with probability $1 - x_0 - x_{10} - x_{11} - x_{12}$, the customer has to wait before getting the service. The waiting time may be viewed as the time until absorption in a Markov chain with state space

$$\Omega_1 = \{*\} \bigcup \{\mathbf{2}, \mathbf{3}, \ldots\}. \tag{3.28}$$

Here $*$ is the absorbing state, which corresponds to taking the tagged customer into service and is obtained by lumping together the level states $\mathbf{0} = \{(0,0)\}$ and $\mathbf{1} = \{(1,0),(1,1),(1,2)\}$.

For $i \geq 2$, the level $\mathbf{i}$ is given by $\mathbf{i} = \{(i,j),\ j = 1 \text{ or } 2\}$. The states other than the absorbing state correspond to the number of customers present in the system as the tagged customer arrives. Once the tagged customer joins the queue, the subsequent arrivals will not affect his waiting time in the queue. Hence the parameter λ does not show up in the generator matrix $\widetilde{Q}$ of this Markov process, given by

$$\widetilde{Q} = \begin{array}{c} \\ * \\ 2 \\ 3 \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccc} * & 2 & 3 & \cdots \end{array} \\ \begin{pmatrix} & & & \\ A_2\mathbf{e} & D & & \\ & A_2 & D & \\ & & \ddots & \ddots \end{pmatrix} \end{array}, \quad \text{where } D = \begin{bmatrix} -\mu_1 - \theta\mu_2 - \eta & \eta \\ 0 & -\mu_1 - \mu_2 \end{bmatrix}. \tag{3.29}$$

Define vector

$$\mathbf{y}(t) = (y_*(t), \mathbf{y}_2(t), \mathbf{y}_3(t), \ldots), \tag{3.30}$$

where

$$\mathbf{y}_i(t) = (y_{i1}(t), y_{i2}(t)), \quad \text{for } i \geq 2. \tag{3.31}$$

The components of the $\mathbf{y}_i(t)$ are the probabilities that at time t, the CTMC with generator $\widetilde{Q}$ is in the respective states of level i. Note that the scalar $y_*(t)$ is the probability that the process is in the absorbing state at time t. By the PASTA property, we get

$$\mathbf{y}(0) = (x_0 + x_{11} + x_{10} + x_{12}, \mathbf{x}_2, \mathbf{x}_3, \ldots). \tag{3.32}$$

Clearly

$$W(t) = y_*(t), \quad \text{for } t \geq 0. \tag{3.33}$$

The LST of $W(t)$ is given by (see [20])

$$\widetilde{W}(s) = \sum_{i=2}^{\infty} \mathbf{y}_i(0) \left[(sI - D)^{-1} A_2 \right]^{i-2} (sI - D)^{-1} A_2 \mathbf{e}. \tag{3.34}$$

The mean waiting time can be obtained from $\widetilde{W}(s)$ as

$$E(W) = -\widetilde{W}'(0) = \sum_{i=1}^{\infty} \mathbf{x}_{2+i} \sum_{j=0}^{i-1} U^j (-D)^{-1} U^{i-j} U \mathbf{e} + \sum_{i=0}^{\infty} \mathbf{x}_{2+i} U^i (-D)^{-2} A_2 \mathbf{e}, \tag{3.35}$$

where $U = (-D)^{-1}A_2$ is a stochastic matrix. Hence, (3.34) can be simplified as

$$E(W) = -\widetilde{W}'(0) = \sum_{i=1}^{\infty} \mathbf{x}_{2+i} \sum_{j=0}^{i-1} U^j (-D)^{-1}\mathbf{e} + \sum_{i=0}^{\infty} \mathbf{x}_{2+i} U^i (-D)^{-1}\mathbf{e}. \tag{3.36}$$

Let

$$H = \sum_{i=0}^{\infty} \mathbf{x}_{2+i} U^i, \tag{3.37}$$

Since U is stochastic, we get

$$H\mathbf{e} = \mathbf{x}_2 (I-R)^{-1}\mathbf{e} = 1 - x_0 - x_{10} - x_{11} - x_{12} \tag{3.38}$$

This result can be used to find an approximate value of H and hence that of the second term in (3.36) to any desired degree of accuracy. Thus, only the first term in (3.36) demands serious computation. For this we make use of the ideas in [15, 16, 21].

Now consider the matrix

$$U_2 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \tag{3.39}$$

which has the property that

$$UU_2 = U_2 U = U_2. \tag{3.40}$$

Then we get

$$\sum_{j=0}^{i-1} U^j (I - U + U_2) = I - U^i + iU_2, \quad \text{for } i \geq 1. \tag{3.41}$$

By the classical theorem on finite Markov chains, the matrix $(I - U + U_2)$ is nonsingular (see [22]). In view of the last equation, the first term in (3.36) becomes

$$\left[\sum_{i=1}^{\infty} \mathbf{x}_{2+i} \left(I - U^i + iU_2\right)\right] (I - U + U_2)^{-1} (-D)^{-1}\mathbf{e}. \tag{3.42}$$

With this simplification, we get

$$\begin{aligned} E(W) = {} & \left[\mathbf{x}_2 \left(R(I-R)^{-1} + I + R(I-R)^{-2}U_2\right) - H\right] (I - U + U_2)^{-1} (-D)^{-1}\mathbf{e} \\ & + H(-D)^{-1}\mathbf{e}. \end{aligned} \tag{3.43}$$

3.5. Conditional Stochastic Decomposition of Queue Length

In this section, we provide a stochastic decomposition of queue length in the stationary regime, subject to the condition that both servers are busy. Note that from (3.5)–(3.8) we get x_0, x_{10}, x_{11}, x_{12}, x_{21}, and x_{22}. Let Q_v be the queue length of the vacation model under study, subject to the condition that both servers are busy. Then we have the following.

Theorem 3.4. *If $\rho < 1$, then $Q_v = Q_0 + Q_d$, where Q_0 and Q_d are two independent random variables. Q_0 is the queue length of the M/M/2 queueing model with heterogeneous servers without vacation and Q_d can be interpreted as the additional queue length due to vacation and slow service, subject to the condition that both servers are busy.*

Proof. Let P_b denote the Probability that both servers are busy. Then

$$\begin{aligned}
P_b &= \sum_{n=2}^{\infty} x_{n2} = \sum_{n=2}^{\infty} x_{22}\rho^{n-2} + \sum_{n=3}^{\infty} x_{21}R_{12}\sum_{j=0}^{n-3} {R_{11}}^{j}\rho^{n-j-3} \\
&= x_{22}\rho\sum_{k=0}^{\infty}\rho^{k} + x_{21}R_{12}\sum_{k=0}^{\infty}{R_{11}}^{k}\sum_{k=0}^{\infty}\rho^{k}; \quad k = n-3 \\
&= (1-\rho)^{-1}\left(x_{22}\rho + x_{21}R_{12}(1-R_{11})^{-1}\right),
\end{aligned} \tag{3.44}$$

so that

$$\frac{1}{P_b} = (1-\rho)\left(x_{22}\rho + x_{21}R_{12}(1-R_{11})^{-1}\right)^{-1} = (1-\rho)\delta, \tag{3.45}$$

where

$$\delta = \left(x_{22}\rho + x_{21}R_{12}(1-R_{11})^{-1}\right)^{-1}, \tag{3.46}$$

$Q_v(z)$, the generating function of the queue length subject to the condition that both servers are busy, is given by

$$\begin{aligned}
Q_v(z) &= \frac{1}{P_b}\sum_{n=2}^{\infty} x_{n1}z^{n-2} = \frac{1}{P_b}\sum_{n=2}^{\infty} x_{22}\rho^{n-2}z^{n-2} \\
&\quad + \frac{1}{P_b}\sum_{n=3}^{\infty}\left(x_{21}R_{12}\sum_{j=0}^{n-3}{R_{11}}^{j}\rho^{n-j-3}\right)z^{n-3}.
\end{aligned} \tag{3.47}$$

By following a computational procedure similar to that of P_b, we arrive at

$$\begin{aligned}
Q_v(z) &= \frac{1-\rho}{1-\rho z}\left\{\delta\left(x_{22}\rho z + \frac{x_{21}\mathrm{R}_{12}}{1-R_{11}z}\right)\right\} \\
&= Q_0(z)Q_d(z),
\end{aligned} \tag{3.48}$$

where

$$Q_0(z) = \frac{1-\rho}{1-\rho z}, \tag{3.49}$$

$$Q_d(z) = \delta\left(x_{22}\rho z + \frac{x_{21}\mathrm{R}_{12}}{1-R_{11}z}\right). \tag{3.50}$$

From (3.49) it follows that $Q_0(z)$ is the generating function of an $M/M/2$ heterogeneous queuing model without vacations, which is precisely the case $\beta = 1$ in [23]. Equation (3.50) suggests that Q_d has a geometric distribution with parameter $1 - R_{11}$. □

Remark 3.5. Due to the algorithmic approach used in the derivation stationary waiting time distribution, a similar decomposition result for the waiting time distribution is far from reality.

3.6. Key System Performance Measures

In this section, we list a number of key system performance measures along with their formulae in addition to the busy period structure and the mean waiting time discussed above.

(1) The probability that the system is empty: $P_{\mathrm{EMP}} = x_0$.

(2) The probability that the server 1 is idle: $P_{\mathrm{IDL}} = x_0 + x_{11} + x_{12}$.

(3) The probability that the server 2 is on vacation: $P_{\mathrm{VAC}} = x_0 + x_{10}$.

(4) The probability that the server 2 is working in vacation mode: $P_{\mathrm{SLOW}} = \sum_{j=1}^{\infty} x_{j1} = x_{11} + x_{21}/(1 - R_{11})$.

(5) The probability that the server 2 is working in normal mode: $P_{\mathrm{NORM}} = 1 - x_0 - P_{\mathrm{SLOW}}$.

(6) The mean number of customers in the system: $\mu_{NS} = \sum_{j=1}^{\infty} j\mathbf{x}_j\mathbf{e} = x_{10} + x_{11} + x_{12} + \mathbf{x}_2(I - R)^{-2}R^{-1}\mathbf{e} - \mathbf{x}_2R^{-1}\mathbf{e}$.

4. Numerical Results

4.1. Illustrative Example

We analyze the effect of the parameters λ, η and θ on the key performance measures. Table 1 analyzes the effect of λ, Table 2 explains the effect of η, and Table 3 examines the effect of θ on the performance measures. To this end, we use the following abbreviations in addition to the notations used in Section 3.6.

μ_{WTQ}: Mean waiting time in the queue.

μ_{LBP}: Mean length of a busy period.

μ_{LBC}: Mean length of a busy cycle.

μ_{NSBP}: Mean number of service completions in a busy period.

(i) Since μ_1 and μ_2 are fixed, the traffic intensity ρ increases with λ. Due to this P_{NORM}, μ_{NS} and μ_{WTQ} increase and P_{VAC} and P_{IDL} decrease as λ increases. Note that the busy period starts with the Markov chain in the state $(1, 0)$; that is, with server 2

Table 1: Case A: $\mu_1 = 10$, $\mu_2 = 5$, $\eta = 1$ and $\theta = 0.6$. Case B: $\mu_1 = 5$, $\mu_2 = 10$, $\eta = 1$ and $\theta = 0.6$.

λ	A/B	P_{IDL}	P_{VAC}	P_{SLOW}	P_{NORM}	μ_{NS}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
2	A	0.829	0.913	0.065	0.022	1.676	0.002	0.080	0.496	1.167
	B	0.707	0.916	0.070	0.014	2.176	0.005	0.125	0.482	1.17
4	A	0.687	0.745	0.168	0.087	1.934	0.009	0.067	0.246	1.128
	B	0.520	0.755	0.188	0.057	2.301	0.016	0.097	0.236	1.156
6	A	0.550	0.568	0.246	0.186	2.189	0.017	0.062	0.166	1.121
	B	0.374	0.574	0.296	0.130	2.60	0.031	0.088	0.164	1.203
8	A	0.414	0.405	0.285	0.311	2.547	0.027	0.062	0.132	1.172
	B	0.251	0.399	0.363	0.238	3.178	0.047	0.094	0.143	1.362
10	A	0.282	0.262	0.278	0.460	3.193	0.036	0.072	0.122	1.340
	B	0.150	0.245	0.365	0.390	4.282	0.062	0.120	0.153	1.755
12	A	0.158	0.141	0.218	0.641	4.702	0.043	0.106	0.144	1.849
	B	0.074	0.123	0.282	0.595	6.626	0.069	0.191	0.216	2.818
14	A	0.049	0.042	0.092	0.866	11.885	0.047	0.294	0.324	4.659
	B	0.020	0.034	0.113	0.853	16.043	0.064	0.566	0.585	8.418

Table 2: Case A: $\lambda = 12$, $\mu_1 = 10$, $\mu_2 = 5$ and $\theta = 0.6$. Case B: $\lambda = 12$, $\mu_1 = 5$, $\mu_2 = 10$ and $\theta = 0.6$.

η	A/B	P_{IDL}	P_{VAC}	P_{SLOW}	P_{NORM}	μ_{NS}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
0.1	A	0.112	0.092	0.333	0.575	5.575	0.059	0.137	0.174	2.20
	B	0.028	0.044	0.451	0.506	14.857	0.209	0.536	0.56	6.952
0.2	A	0.125	0.104	0.306	0.589	5.314	0.054	0127	0.165	2.086
	B	0.04	0.064	0.408	0.528	10.582	0.137	0.365	0.39	4.907
0.3	A	0.133	0.113	0.287	0.600	5.169	0.051	0.121	0.159	2.019
	B	0.048	0.078	0.379	0.543	9.044	0.111	0.301	0.326	4.137
0.4	A	0.140	0.119	0.272	0.608	5.066	0.049	0.117	0.155	1.974
	B	0.054	0.088	0.357	0.555	8.232	0.097	0.266	0.290	3.714
0.5	A	0.144	0.125	0.26	0.615	4.983	0.047	0.114	0.152	1.941
	B	0.059	0.097	0.340	0.564	7.724	0.088	0.243	0.268	3.442
0.6	A	0.148	0.129	0.250	0.622	4.914	0.046	0.112	0.150	1.915
	B	0.063	0.103	0.325	0.572	7.374	0.082	0.227	0.252	3.248

on vacation. Hence, initially P_{SLOW} increases with λ. For this reason μ_{LBP}, μ_{LBC}, and μ_{NSBP} show an early downward trend. But as λ further increases P_{SLOW} declines as expected due to the high traffic intensity. Hence, μ_{LBP} and μ_{NSBP} reverse the direction of change. Due to the effect of P_{VAC} and P_{IDL}, this reversal occurs only at a later stage for μ_{LBC}. It is worth comparing the values of the measures in cases A and B. Even though the net service rate $\mu_1 + \mu_2 = 15$ in both cases, the effect of the vacation parameter η becomes more predominant when $\mu_1 < \mu_2$. Due to this the measures P_{VAC} and P_{IDL} take smaller values and the measures μ_{NS}, μ_{WTQ}, μ_{LBP}, and μ_{NSBP} take larger values in case B, compared to their values in case A.

(ii) As η increases, the mean duration of vacation $1/\eta$ decreases. But as the mean duration of vacation decreases, the probability of the expiry of the vacation without initiating the slow service increases. The chance of an early expiry of vacation always results in an increase in P_{NORM} and P_{VAC} and a decrease in P_{SLOW}. Note that $P_{\text{VAC}} + P_{\text{SLOW}}$ decreases as η increases and $P_{\text{VAC}} + P_{\text{SLOW}} < P_{\text{NORM}}$ for any value of η in the given range. So P_{IDL} increases with η. Thus, the proportion of time in which

Table 3: Case A: $\lambda = 12$, $\mu_1 = 10$, $\mu_2 = 5$ and $\eta = 1$. Case B: $\lambda = 12$, $\mu_1 = 5$, $\mu_2 = 10$ and $\eta = 1$.

θ	A/B	P_{IDL}	P_{VAC}	P_{SLOW}	P_{NORM}	μ_{NS}	μ_{WTQ}	μ_{LBP}	μ_{LBC}	μ_{NSBP}
0.1	A	0.112	0.073	0.167	0.761	4.402	0.044	0.158	0.195	2.439
	B	0.034	0.042	0.188	0.770	8.45	0.105	0.507	0.531	6.601
0.2	A	0.121	0.085	0.178	0.737	4.504	0.044	0.144	0.182	2.286
	B	0.039	0.053	0.207	0.740	8.126	0.098	0.420	0.444	5.556
0.3	A	0.13	0.098	0.190	0.713	4.581	0.044	0.133	0.171	2.152
	B	0.046	0.066	0.226	0.708	7.762	0.091	0.345	0.370	4.664
0.4	A	0.139	0.112	0.200	0.689	4.638	0.044	0.123	0.161	2.036
	B	0.054	0.082	0.246	0.673	7.379	0.084	0.283	0.308	3.918
0.5	A	0.149	0.126	0.209	0.664	4.678	0.044	0.114	0.152	1.936
	B	0.063	0.101	0.265	0.635	6.995	0.076	0.232	0.257	3.307
0.6	A	0.158	0.141	0.218	0.641	4.702	0.043	0.106	0.144	1.849
	B	0.074	0.123	0.282	0.595	6.626	0.069	0.191	0.216	2.818

both servers work at the normal rate increases as η increases. Hence, the measures μ_{NS}, μ_{WTQ}, μ_{LBP}, μ_{LBC}, and μ_{NSBP} decrease as η increases. The argument given the last paragraph holds good for the difference in magnitude of the measures in cases A and B.

(iii) As θ increases, the service rate $\theta\mu_2$ of the second server during vacation mode of service increases. As a result, server 2 clears out customers at an increased rate in slow service mode. This produces an increase in P_{VAC}, P_{SLOW}, and P_{IDL} and a decrease in P_{NORM} as expected. Consequently, μ_{LBP}, μ_{LBC} and μ_{NSBP} decrease as θ increases. The huge difference in the value of net service rate $\mu_1 + \theta\mu_2$ between cases A and B, during vacation mode of service, is the reason for the pattern of behavior of μ_{NS} in these two cases. Increase in θ does not affect μ_{WTQ} significantly in case A but it affects the measure in case B. This is because the effect of θ becomes significant only when μ_2 is large compared to μ_1.

5. Concluding Remarks

In this paper, we studied an $M/M/2$ queueing model with heterogeneous servers. One server follows multiple vacation policy. But this server offers service at a lower rate during vacation if customers arrive. The other server remains in the system even when it is empty. The busy period of the system was analyzed in the stationary regime. Mean waiting time of a customer has been computed. Conditional stochastic decomposition of queue length has been derived. An illustrative numerical example to bring out the qualitative nature of the model has been presented.

Acknowledgments

The authors would like to thank the anonymous referees for their valuable suggestions and comments which greatly improved the presentation of the paper. C. Sreenivasan's research is supported by the University Grants Commission, Government of India, under Faculty Development Programme (Grant no. F.FIP/11th Plan/KLCA042TF02).

References

[1] B. T. Doshi, "Queueing systems with vacations—a survey," *Queueing Systems*, vol. 1, no. 1, pp. 29–66, 1986.

[2] H. Takagi, *Queueing Analysis Volume 1: Vacations and Priority System*, North-Holland, Amsterdam, The Netherlands, 1991.

[3] N. Tian and Z. G. Zhang, *Vacation Queueing Models*, International Series in Operations Research & Management Science no. 93, Springer, New York, NY, USA, 2006.

[4] X. Chao and Y. Q. Zhao, "Analysis of multi-server queues with station and server vacations," *European Journal of Operational Research*, vol. 110, no. 2, pp. 392–406, 1998.

[5] Y. Levy and U. Yechiali, "An $M/M/s$ queue with servers' vacations," *Information Systems and Operational Research*, vol. 14, no. 2, pp. 153–163, 1976.

[6] B. Vinod, "Exponential queues with servers' vacations," *Journal of the Operational Research Society*, vol. 37, no. 10, pp. 1007–1014, 1986.

[7] E. P. C. Kao and K. S. Narayanan, "Analyses of an M/M/N/ queue with server's vacations," *European Journal of Operational Research*, vol. 54, no. 2, pp. 256–266, 1991.

[8] L. Servi and S. Finn, "M/M/1 queue with working vacations (M/M/1/WV)," *Performance*, vol. 50, pp. 41–52, 2002.

[9] J. Kim, D. Choi, and K. Chae, "Analysis of queue-length distribution of the M/G/1with working vacations," in *Proceedings of the International Conference on Statistics and Related*, Honolulu, Hawaii, USA, 2003.

[10] D. A. Wu and H. Takagi, "M/G/1 queue with multiple working vacations," *Performance Evaluation*, vol. 63, no. 7, pp. 654–681, 2006.

[11] Y. Baba, "Analysis of a GI/M/1 queue with multiple working vacations," *Operations Research Letters*, vol. 33, no. 2, pp. 201–209, 2005.

[12] N.-S. Tian, J.-H. Li, and Z. G. Zhang, "Matrix analytic method and working vacation queues—a survey," *International Journal of Information and Management Sciences*, vol. 20, no. 4, pp. 603–633, 2009.

[13] J. Li and N. Tian, "The M/M/1 queue with working vacations and vacation interruptions," *Journal of Systems Science and Systems Engineering*, vol. 16, no. 1, pp. 121–127, 2007.

[14] M. Zhang and Z. Hou, "Performance analysis of $MAP/G/1$ queue with working vacations and vacation interruption," *Applied Mathematical Modelling*, vol. 35, no. 4, pp. 1551–1560, 2011.

[15] M. F. Neuts and Y. Takahashi, "Asymptotic behavior of the stationary distributions in the $GI/PH/c$ queue with heterogeneous servers," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 4, pp. 441–452, 1981.

[16] B. Krishna Kumar and S. Pavai Madheswari, "An $M/M/2$ queueing system with heterogeneous servers and multiple vacations," *Mathematical and Computer Modelling*, vol. 41, no. 13, pp. 1415–1429, 2005.

[17] W. Szpankowski, "Stability conditions for multidimensional queueing systems with computer applications," *Operations Research*, vol. 36, no. 6, pp. 944–957, 1988.

[18] L. I. Sennott, P. A. Humblet, and R. L. Tweedie, "Mean drifts and the nonergodicity of Markov chains," *Operations Research*, vol. 31, no. 4, pp. 783–789, 1983.

[19] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*, ASA-SIAM Series on Statistics and Applied Probability, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa, USA, 1999.

[20] M. F. Neuts, *Matrix-Geometric Solutions in Stochastic Models*, vol. 2 of *Johns Hopkins Series in the Mathematical Sciences*, Johns Hopkins University Press, Baltimore, Md, USA, 1981, [1994 version is Dover Edition].

[21] M. F. Neuts and D. M. Lucantoni, "A Markovian queue with N servers subject to breakdowns and repairs," *Management Science*, vol. 25, no. 9, pp. 849–861, 1979.

[22] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, The University Series in Undergraduate Mathematics, D. Van Nostrand, Princeton, NJ, USA, 1960.

[23] V. P. Singh, "Two-server Markovian queues with balking: heterogeneous versus homogeneous servers," *Operations Research*, vol. 18, no. 1, pp. 145–159, 1970.