The Eighth International Symposium on Operations Research and Its Applications (ISORA'09)
Zhangjiajie, China, September 20–22, 2009

A Heterogeneous Two-Server Queueing System with Balking and Server Breakdowns

Dequan Yue[1] Wuyi Yue[2] Jun Yu[1] Ruiling Tian[1]

[1]Department of Statistics, College of Science, Yanshan University, Qinhuangdao 066004, China
[2]Department of Intelligence and Informatics, Konan University, Kobe 658-8501, Japan

Abstract In this paper, we study an M/M/2 queueing system with balking and two heterogeneous servers, Server 1 and Server 2. Customers arrive according to a Poisson process and form a single waiting line. Two parallel servers provide heterogeneous exponential service to customers on a first-come first-served basis. It is assumed that Server 1 is perfectly reliable and Server 2 is subject to breakdowns. For this system, we obtain the stationary condition where the system reaches a steady state. We also obtain the steady-state probabilities in a matrix form by using a matrix-geometric solution method. Finally, we produce explicit expressions of some performance measures such as the mean system size, the average balking rate and the probabilities that Server 2 is in various states. Numerical illustrations are also provided.

Keywords Queueing system; balking; breakdowns; matrix-geometric solution; mean system size

1 Introduction

Recent decades have seen an increasing interest in queueing systems with server breakdowns. This has been due to their applications in manufacturing systems, service systems, telecommunications and computer systems. In many practical queueing systems, situations often occur where servers are subject to breakdowns. For example, in a machine processing center, machine breakdowns may occur due to factors such as power failure, lack of preventive maintenance, or the use of inferior row materials. Other examples of queueing systems with server breakdowns can be found in service systems, computer systems and telecommunication systems.

Single server queueing systems with server breakdowns have been studied by many researchers including Federgruen and Green [1], Li *et al.* [2], Tang [3], Nakdimon and Yechiali [4], Wang *et al.* [5], Wang *et al.* [6], Choudhury and Tadj [7], to mention a few. Multi-server queueing systems with server breakdowns are more flexible and applicable in practice than single server counterparts. However, due to their analytical complexity, there have been only a few studies carried out on multi-server queueing system with server breakdowns. Mitrany and Avi-Itzhak [8] studied an M/M/N queue with server breakdowns and ample repair capacity. In their study, the moment generating function of the queue size is obtained by using the transformation method. Vinod [9] considered the same model using the matrix-geometric solution method. For $N = 1$, Vinod [9] imposed some restrictions on the server down-periods (either independent of the queue length or only occurring when the server is active). Neuts and Lucantoni [10] and Wartenhosrt [11] extended the models studied in [8] and [9] by considering a limited repair capacity. Neuts and Lucantoni [10] considered a single queue of customers, each served by one of N parallel servers. Wartenhosrt [11] considered N single-server queueing stations, each serving its own stream of customers. Wang and Chang [12] studied an M/M/R/N queue with balking, reneging and server breakdowns from the viewpoint of queueing. They solved the steady-state probability equations iteratively and derived the steady-state probabilities in a matrix form.

The models mentioned above all assumed the servers to be homogeneous, where the individual service rates were the same for all the servers in the system. This assumption may be valid only when the service process is highly mechanically or electronically controlled. In a queueing system with human servers, we can not expect work to be carried out at the same rate. We face situations of this kind in our everyday life, e.g., at checkout counters in department stores, in banks, in hospitals, etc.

Singh [13] studied an M/M/2 queueing system with balking and heterogenous servers. In [13], the author obtained the stationary queue length distribution and the mean queue length and also compared the model with heterogenous servers and the model with homogenous servers. Kumar and Madheswari [14] studied an M/M/2 queueing system with heterogenous servers and multiple vacations by using the matrix-geometric solution method. They studied the stationary queue length distribution and waiting time distribution along with their means via the rate matrix. Yue *et al.* [15] further considered the model in [14]. They obtained the explicit expression of the rate matrix and proved the conditional stochastic decomposition results for the stationary queue length and waiting time. Madan *et al.* [16] studied a two-server queue with Bernoulli schedules and a single vacation policy where the two servers provide heterogenous exponential service to customers. They obtained steady-state probability generating functions of the system size for various states of the servers.

The models studied in [13]-[16] all assume that the servers are reliable. However, we know that there are many practical queueing situations where the servers are subject to lengthy and unpredictable breakdowns. For example, in a machine processing system, machines may subject to breakdowns during production or when the system is idle. Failures occurred during production may be due to power failure or the use of inferior row materials, while failures occurred when the system is idle may be due to server's vacations or planed maintenances. Therefore, failures occurred during production may be different from failures when the system is idle. For this, in this paper, we consider a system with unreliable servers by extending the system model presented in [13]. We model the system as an M/M/2 queueing system with balking and two heterogeneous servers, where Server 1 is perfectly reliable but Server 2 is subject to two types of breakdowns, Type 1 breakdowns and Type 2 breakdowns. Type 1 breakdowns occur only in an idle period of Server 2, while Type 2 breakdowns occur only in a working period of Server 2. In our model, the balking probability depends on the states of servers. If an arriving customer finds at least one server is free and available (i.e., Server 1 is free, or Server 1 is busy while Server 2 is free and available), then the customer joins the system. If the customer finds both servers are busy, then the customer joins the system with probability b_0 $(0 \leq b_0 \leq 1)$, and balks with probability $1 - b_0$. If the customer finds Server 1 is busy while Server 2 is unavailable, then the customer joins the system with probability b_1 $(0 \leq b_1 \leq 1)$, and balks with probability $1-b_1$.

The rest of the paper is organized as follows. In Section 2, the model description and a quasi-birth-and-death (QBD) model formulation are presented. In Section 3, the stationary condition is derived. The explicit expressions of the steady-state probabilities in the matrix form and some performance measures are obtained. In Section 4, numerical illustrations are provided to highlight the effect of some parameters on the mean system size. Conclusions are given in Section 5.

2 Model Formulation

In this paper, we consider an M/M/2 queueing system with balking and server breakdowns, where the two servers have different service rates.

2.1 Model Assumptions

The assumptions of the system model are given as follows:

(a) Arrivals of customers follow a Poisson process with arrival rate λ. Arriving customers form a single waiting line based on the order of their arrivals.

(b) There are two servers in the system. Server 1 is perfectly reliable, while Server 2 is subject to two types of breakdowns. Type 1 breakdowns occur only in an idle period of Server 2, while Type 2 breakdowns occur only in a working period of Server 2. It is assumed that Server 2 has an exponentially distributed lifetime with different failure rates α_0 (≥ 0) for Type 1 and α (≥ 0) for Type 2, respectively. Whenever Server 2 breaks down, it is immediately repaired by a repairman. The repaired server is as good as a new one. The customer being serviced just before server breakdown needs to be serviced repeatedly and the elapsed service time is not available. The repair times of Server 2 are assumed to follow an another exponential distribution with repair rate β $(\beta \geq 0)$.

(c) If an arriving customer finds at least one server is free and available (i.e., Server 1 is free, or Server 1 is busy while Server 2 is free and available), then the customer joins the system. If the customer finds both servers are busy, then the customer joins the system with probability b_0 $(0 \leq b_0 \leq 1)$, and balks with probability $1-b_0$. If the customer finds Server 1 is busy while Server 2 is unavailable, then the customer joins the system with probability b_1 $(0 \leq b_1 \leq 1)$, and balks with probability $1-b_1$.

(d) If a customer arrives to find both servers free and available, the customer chooses Server 1 with probability p $(p \geq 0)$ and Server 2 with probability $1-p$.

(e) The two servers provide heterogeneous exponential service to customers on a first-come first-serviced (FCFS) basis with service rates μ_1 and μ_2 for Server 1 and Server 2, respectively.

(f) All stochastic processes involved in the system are independent of each other.

2.2 QBD Process

Let $L(t)$ be the number of customers in the system at time t, and let $J(t)$ be the status of Server 2 at time t, defined as follows:

$$J(t) = \begin{cases} 0, & \text{Server 2 is busy at time } t \\ 1, & \text{Server 2 is free at time } t \\ 2, & \text{Server 2 is broken down at time } t. \end{cases}$$

We define the system state by $L(t)$ and $J(t)$. Then $\{(L(t),J(t)),t\geq 0\}$ is a Markovian process with a state space Ω as follows:

$$\Omega=\{(0,0),(0,2)\}\cup\{(1,j),j=0,1,2\}\cup\{(i,j),i\geq 2,j=1,2\}.$$

Define the levels **0**, **1**, **2**, ... as the sets of the system states, $\mathbf{0}=\{(0,0),(0,2)\}$, $\mathbf{1}=\{(1,0),\ (1,1),\ (1,2)\}$, and $\boldsymbol{i}=\{(i,1),(i,2)\}$ if $i\geq 2$, where the elements of the sets are arranged in lexicographical order. Using elementary arguments, the process $\{(L(t),J(t)),t\geq 0\}$ has a transition rate matrix $\boldsymbol{Q}$ which has a block-tridiagonal structure given by

$$\boldsymbol{Q}=\begin{bmatrix} \boldsymbol{B}_{00} & \boldsymbol{B}_{01} & & & & & \\ \boldsymbol{B}_{10} & \boldsymbol{B}_{11} & \boldsymbol{B}_{12} & & & & \\ & \boldsymbol{B}_{21} & \boldsymbol{A}_1 & \boldsymbol{A}_0 & & & \\ & & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & & \\ & & & \boldsymbol{A}_2 & \boldsymbol{A}_1 & \boldsymbol{A}_0 & \\ & & & & \cdot & \cdot & \cdot \\ & & & & & \cdot & \cdot\ \cdot \end{bmatrix}.$$

Matrix $\boldsymbol{Q}$ is an infinitesimal generator of the Markov process $\{(L(t),J(t)),t\geq 0\}$ and is in the format of a quasi-birth-and-death (QBD) process. The sub-matrices $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, and $\boldsymbol{A}_2$ are square matrices of order 2, respectively and are given by

$$\boldsymbol{A}_0=\begin{bmatrix} \lambda b_0 & 0 \\ 0 & \lambda b_1 \end{bmatrix},\quad \boldsymbol{A}_1=\begin{bmatrix} -(\lambda b_0+\alpha+\mu_1+\mu_2) & \alpha \\ \beta & -(\lambda b_1+\beta+\mu_1) \end{bmatrix},$$

$$\boldsymbol{A}_2=\begin{bmatrix} \mu_1+\mu_2 & 0 \\ 0 & \mu_1 \end{bmatrix}.$$

The boundary matrices are defined by

$$\boldsymbol{B}_{00}=\begin{bmatrix} -(\lambda+\alpha_0) & \alpha_0 \\ \beta & -(\lambda+\beta) \end{bmatrix},\quad \boldsymbol{B}_{01}=\begin{bmatrix} \lambda p & \lambda(1-p) & 0 \\ 0 & 0 & \lambda \end{bmatrix},$$

$$\boldsymbol{B}_{11}=\begin{bmatrix} -(\lambda+\alpha_0+\mu_1) & 0 & \alpha_0 \\ 0 & -(\lambda+\alpha+\mu_2) & \alpha \\ \beta & 0 & -(\lambda b_1+\beta+\mu_1) \end{bmatrix},$$

$$\boldsymbol{B}_{10}=\begin{bmatrix} \mu_1 & 0 \\ \mu_2 & 0 \\ 0 & \mu_1 \end{bmatrix},\quad \boldsymbol{B}_{12}=\begin{bmatrix} \lambda & 0 \\ \lambda & 0 \\ 0 & \lambda b_1 \end{bmatrix},\quad \boldsymbol{B}_{21}=\begin{bmatrix} \mu_2 & \mu_1 & 0 \\ 0 & 0 & \mu_1 \end{bmatrix}.$$

3 Steady-state Analysis

In this section, we first derive the condition for the system to reach a steady state. Then, we derive the steady-state probabilities of the system by using a matrix-geometric solution method. The computations of the rate matrix and the boundary probability vectors are also discussed. We finally derive some performance measures of the system by using the steady-state probabilities.

3.1 Stationary Condition

We now derive the condition for the system to reach a steady state. Define matrix $\boldsymbol{A} = \boldsymbol{A}_0 + \boldsymbol{A}_1 + \boldsymbol{A}_2$. Then, matrix $\boldsymbol{A}$ can be written as

$$\boldsymbol{A} = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix}.$$

It is readily known that $\boldsymbol{A}$ is an irreducible generator of a Markov process. Let $\boldsymbol{\pi} = (\pi_0, \pi_1)$ be a stationary probability vector of this Markov process. Then, $\boldsymbol{\pi}$ satisfies the linear equations: $(\pi_0, \pi_1)\boldsymbol{A} = 0$ and $\pi_0 + \pi_1 = 1$. Solving these equations, we have

$$\pi_0 = \frac{\beta}{\alpha+\beta}, \quad \pi_1 = \frac{\alpha}{\alpha+\beta}.$$

By Theorem 3.1.1 in [17], the condition $\boldsymbol{\pi A}_0\boldsymbol{e} < \boldsymbol{\pi A}_2\boldsymbol{e}$ is the necessary and sufficient condition for stability of the QBD process, where $\boldsymbol{e}$ is a column vector of order 2 with all the elements equal to one. After some routine manipulation, the stationary condition turns out to be

$$\rho = \frac{\lambda(\alpha b_1 + \beta b_0)}{(\alpha+\beta)\mu_1 + \beta\mu_2} < 1. \tag{1}$$

Remark. If we let $b_0 = b_1 = b$ and $\alpha = 0$ in Eq. (1), then the stationary condition $\rho < 1$ becomes $b\lambda < \mu_1 + \mu_2$, which is the stationary condition obtained by Singh in [13]. When there is no balking, we let $b_0 = b_1 = 1$ in Eq. (1), then the stationary condition $\rho < 1$ becomes

$$\rho = \frac{\lambda(\alpha+\beta)}{(\alpha+\beta)\mu_1 + \beta\mu_2} < 1.$$

This is the stationary condition obtained by Yu *et al.* in [18].

3.2 Matrix-geometric Solution

Let L and J be the stationary random variables for the number of customers in the system and the status of Server 2. We denote the stationary probability by

$$P_{ij} = \{L = i, J = j\} = \lim_{t\to\infty} P\{L(t) = i, J(t) = j\}, \quad (i,j) \in \Omega$$

where $i = 1, 2, ..., j = 0, 1, 2$. Under the stationary condition $\rho < 1$, the stationary probability vector $\boldsymbol{P}$ of the generator $\boldsymbol{Q}$ exists. This stationary probability vector $\boldsymbol{P}$ is partitioned as $\boldsymbol{P} = (\boldsymbol{P}_0, \boldsymbol{P}_1, \boldsymbol{P}_2, ...)$, where $\boldsymbol{P}_0 = (P_{00}, P_{02})$, $\boldsymbol{P}_1 = (P_{10}, P_{11}, P_{12})$, and $\boldsymbol{P}_i = (P_{i0}, P_{i2})$ for $i \geq 2$.

Based on the matrix-geometric solution method in [17], the stationary probability vector $\boldsymbol{P}$ is given by

$$\boldsymbol{P}_0\boldsymbol{B}_{00} + \boldsymbol{P}_1\boldsymbol{B}_{10} = 0, \tag{2}$$

$$\boldsymbol{P}_0\boldsymbol{B}_{01} + \boldsymbol{P}_1\boldsymbol{B}_{11} + \boldsymbol{P}_2\boldsymbol{B}_{21} = 0, \tag{3}$$

$$\boldsymbol{P}_1\boldsymbol{B}_{12} + \boldsymbol{P}_2(\boldsymbol{A}_1 + \boldsymbol{R}\boldsymbol{A}_2) = 0, \tag{4}$$

$$\boldsymbol{P}_i = \boldsymbol{P}_2\boldsymbol{R}^{i-2}, \quad i = 3, 4, 5, ... \tag{5}$$

and the normalizing equation

$$\boldsymbol{P}_0\boldsymbol{e}+\boldsymbol{P}_1\boldsymbol{e}_1+\boldsymbol{P}_2(\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e}=1 \tag{6}$$

where $\boldsymbol{I}$ is an identity matrix of order 2, $\boldsymbol{e}_1$ is a column vector of order 3 with all the elements equal to one, and $\boldsymbol{R}$, called the rate matrix, is the minimal non-negative solution with a spectral radius of less than one, of the matrix quadratic equation as follows:

$$\boldsymbol{R}^2\boldsymbol{A}_2+\boldsymbol{R}\boldsymbol{A}_1+\boldsymbol{A}_0=\mathbf{0}. \tag{7}$$

In order to obtain the rate matrix $\boldsymbol{R}$, we need to solve Eq. (7). We accomplish this by letting

$$\boldsymbol{R}=\begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} \tag{8}$$

and using the following relation (see [17, p. 83]), $\boldsymbol{R}\boldsymbol{A}_2\boldsymbol{e}=\boldsymbol{A}_0\boldsymbol{e}$, so we get

$$r_{12}=\frac{\lambda b_0}{\mu_1}-(1+\frac{\mu_2}{\mu_1})r_{11}, \tag{9}$$

$$r_{21}=\frac{\lambda b_1}{\mu_1+\mu_2}-\frac{\mu_1}{\mu_1+\mu_2}r_{22} \tag{10}$$

where $\boldsymbol{A}_0$, $\boldsymbol{A}_1$, $\boldsymbol{A}_2$, $\boldsymbol{B}_{00}$, $\boldsymbol{B}_{01}$, $\boldsymbol{B}_{10}$, $\boldsymbol{B}_{11}$, $\boldsymbol{B}_{12}$ and $\boldsymbol{B}_{21}$ are given in Subsection 2.2.

Substituting $\boldsymbol{R}$ into Eq. (7) and using the Eqs. (9) and (10), after some routine manipulations, we get the following equations:

$$r_{11}^2+r_{11}r_{22}-f_0r_{11}-f_1r_{22}+f_1f_3=0, \tag{11}$$

$$r_{22}^2+r_{11}r_{22}-f_0r_{22}-f_2r_{11}+f_2f_4=0 \tag{12}$$

where

$$f_0=1+\frac{1}{\mu_1}(\lambda b_1+\beta)+\frac{\lambda b_0+\alpha}{\mu_1+\mu_2},\quad f_1=\frac{\lambda b_0}{\mu_1+\mu_2},$$
$$f_2=\frac{\lambda b_1}{\mu_1},\quad f_3=1+\frac{1}{\mu_1}(\lambda b_0+\beta),\quad f_4=\frac{\lambda b_0+\alpha}{\mu_1+\mu_2}.$$

From Eq. (11), we get

$$r_{22}=\frac{1}{f_1-r_{11}}(r_{11}^2-f_0r_{11}+f_1f_3). \tag{13}$$

Substituting Eq. (13) into Eq. (12), we get the third-degree equation containing one variable r_{11} as follows:

$$ar_{11}^3+br_{11}^2+cr_{11}+d=0 \tag{14}$$

where

$$a = f_1 - f_2,$$
$$b = f_1(f_3 - 2f_0) + f_2(f_4 + 2f_1),$$
$$c = f_1\left[f_1(f_3 - f_2) - f_0(f_3 - f_0) - 2f_2 f_4\right],$$
$$d = f_1^2\left[f_3(f_3 - f_0) + f_2 f_4\right].$$

The explicit expression of r_{11} can be obtained by using the root formula of the third-degree equation. Then, we can obtain r_{22} from Eq. (13), r_{12} from Eq. (9) and r_{21} from Eq. (10). The details are omitted since the exact solution of Eq. (14) obtained by using the root formula method is lengthy and tedious. However, we can calculate the rate matrix $\boldsymbol{R}$ approximately by using the following simple iterative method. We know that, from Eq. (7), we have

$$\boldsymbol{R} = -\left[\boldsymbol{A}_0 + \boldsymbol{R}^2\boldsymbol{A}_2\right]\boldsymbol{A}_1^{-1}. \tag{15}$$

Taking the initial value of $\boldsymbol{R} = \boldsymbol{0}$, we can iteratively solve for $\boldsymbol{R}$ and can check the accuracy of this approximation by using the equality $\boldsymbol{R}\boldsymbol{A}_2\boldsymbol{e} = \boldsymbol{A}_0\boldsymbol{e}$. The value of $\boldsymbol{R}$ will converge since $-\boldsymbol{A}_1^{-1}$ and $\boldsymbol{A}_0 + \boldsymbol{R}^2\boldsymbol{A}_2$ are positive. Hence, after each iteration, the elements of $\boldsymbol{R}$ will increase monotonically.

We compare the two methods of computing matrix $\boldsymbol{R}$ by a numerical example. The comparison results are given in Table 1, where the values of parameters of the system are given as follows: $\lambda = 10$, $\mu_1 = 10$, $\mu_2 = 15$, $b_0 = 0.9$, $b_1 = 0.4$, $p = 0.7$, $\beta = 18$, $\alpha_0 = 1$ and $\alpha = 18$.

Table 1. Comparison between the root formula method and the iterative method.

r_{ij}	Iterative method	Root formula	Error
r_{11}	0.2847946242	0.2847946248	6×10^{-10}
r_{12}	0.1880134376	0.1880134381	5×10^{-10}
r_{21}	0.0845427109	0.0845427111	2×10^{-10}
r_{22}	0.1886432220	0.1886432222	2×10^{-10}

Table 1 shows that the iterative method for computing the rate matrix is very simple and accurate. The error is within 10^{-10}. We will use this iterative method to compute the rate matrix for performing numerical experiments in Section 4.

3.3 Boundary Probability Vectors

In order to obtain the stationary boundary probability vectors $\boldsymbol{P}_0$, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$, we need to solve Eqs. (2)-(4) and Eq. (6). To accomplish this, we first prove that the matrix $\boldsymbol{D} = \boldsymbol{B}_{11} - \boldsymbol{B}_{10}\boldsymbol{B}_{00}^{-1}\boldsymbol{B}_{01}$ is a invertible matrix, and then derive the reverse of the matrix $\boldsymbol{D}$. The following lemma will be needed in our proof.

Lemma 1. Let $\boldsymbol{E} = (e_{ij})$ be a $n \times n$ square matrix defined on a real field. If

$$|e_{ii}| > \sum_{j=1}^{i-1} |e_{ij}| + \sum_{j=i+1}^{n} |e_{ij}|, \quad i = 1, 2, ..., n \tag{16}$$

where the empty summations $\sum_{j=1}^{0}$ and $\sum_{j=n+1}^{n}$ are defined to be zero, then $\boldsymbol{E}$ is invertible.
Proof. The proof is simple and thus is omitted.

The following symbols will be used in the following derivation:

$$x_1 = \frac{\beta}{\lambda+\alpha_0+\beta}, \quad x_2 = \frac{\alpha_0}{\lambda+\alpha_0+\beta}, \quad x_3 = \frac{\lambda+\beta}{\lambda+\alpha_0+\beta}.$$

Lemma 2. The matrix $\boldsymbol{D}$ is invertible and its inverse $\boldsymbol{D}^{-1}$ is given by

$$\boldsymbol{D}^{-1} = \frac{1}{|\boldsymbol{D}|}\begin{pmatrix} D_{11} & D_{21} & D_{31} \\ D_{12} & D_{22} & D_{32} \\ D_{13} & D_{23} & D_{33} \end{pmatrix} \tag{17}$$

where

$$D_{11} = x_1\mu_1[\lambda+p(\mu_2+\alpha)] - x_3\mu_2(1-p)(\lambda b_1+\beta) + (\lambda+\mu_2+\alpha)(\lambda b_1+\beta),$$

$$D_{12} = x_1\mu_1 p(\mu_2+\alpha) + (x_2\mu_2+\alpha)\beta + x_3\mu_2 p(\lambda b_1+\beta),$$

$$D_{13} = (x_1\mu_1 p+\beta)(\lambda+\mu_2+\alpha) - x_3\mu_2\beta(1-p),$$

$$D_{21} = x_1\mu_1(1-p)(\mu_1+\alpha_0) + x_3\mu_1(1-p)(\lambda b_1+\beta),$$

$$D_{22} = x_1\mu_1[\lambda+(1-p)(\mu_1+\alpha_0)] - x_2\mu_1\beta + (\lambda+\mu_1-x_3\mu_1 p)(\lambda b_1+\beta) + \lambda\alpha_0 b_1,$$

$$D_{23} = x_1\mu_1(1-p)(\lambda+\alpha_0+\mu_1) + x_3\mu_1(1-p)\beta,$$

$$D_{31} = x_2\mu_1(\lambda+\mu_2) - x_3[\alpha\mu_1 p+\alpha_0\mu_2(1-p)] + \alpha_0(\lambda+\mu_2+\alpha) + \alpha\mu_1,$$

$$D_{32} = x_2\mu_2(\lambda+\mu_1) - x_3[\alpha\mu_1 p+\alpha_0\mu_2(1-p)] + \alpha(\lambda+\mu_1+\alpha_0) + \alpha_0\mu_2,$$

$$D_{33} = x_2\mu_1\mu_2 - x_3[(\lambda+\alpha)p\mu_1+(\lambda+\alpha_0)\mu_2(1-p)] + (\lambda+\alpha_0)(\lambda+\mu_2\alpha) + (\lambda+\alpha)\mu_1$$

and the determinant

$$|\boldsymbol{D}| = -[(\lambda+\alpha_0)+(1-x_3 p)\mu_1]D_{11} + x_3(1-p)\mu_1 D_{12} + (\alpha_0+x_2\mu_1)D_{13}.$$

Proof. Let d_{ij} represent the (i,j)th element of matrix $\boldsymbol{D}$, $i,\ j=1,2,3$. Then, by matrix manipulation, we have

$$d_{11} = -(\lambda+\alpha_0)-(1-x_3p)\mu_1,\ d_{12} = x_3(1-p)\mu_1,\ d_{13} = \alpha_0+x_2\mu_1,$$

$$d_{21} = x_3p\mu_2,\ d_{22} = -(\lambda+\alpha)-[1-x_3(1-p)]\mu_2,\ d_{23} = \alpha+x_2\mu_2,$$

$$d_{31} = \beta+x_1p\mu_1,\ d_{32} = x_1(1-p)\mu_1,\ d_{33} = -(\lambda b_1+\beta)-x_1\mu_1.$$

It is easy to verify that

$$|d_{11}|-(|d_{12}|+|d_{13}|) = \lambda > 0,\ |d_{22}|-(|d_{21}|+|d_{23}|) = \lambda > 0,$$

$$|d_{33}|-(|d_{31}|+|d_{32}|) = \lambda b_1 > 0.$$

By Lemma 1, $\boldsymbol{D}$ is invertible. It is not difficult to obtain the co-factor D_{ij} of d_{ij} and the determinate $|\boldsymbol{D}|$ of $\boldsymbol{D}$ which are as given in Lemma 2. Thus the inverse $\boldsymbol{D}^{-1}$ of matrix $\boldsymbol{D}$ is obtained.

Next, we solve Eqs. (2)-(4) and (6). We define the following matrices:

$$\boldsymbol{F}_0 = \begin{pmatrix} (\lambda+\beta)\mu_1 & \alpha_0\mu_1 \\ (\lambda+\beta)\mu_2 & \alpha_0\mu_2 \\ \beta\mu_1 & (\lambda+\alpha_0)\mu_1 \end{pmatrix}, \tag{18}$$

$$\boldsymbol{F}_1 = \begin{pmatrix} D_{11}\mu_2 + D_{12}\mu_1 & D_{21}\mu_2 + D_{22}\mu_1 & D_{31}\mu_2 + D_{32}\mu_1 \\ D_{13}\mu_1 & D_{23}\mu_1 & D_{33}\mu_1 \end{pmatrix} \tag{19}$$

where D_{ij}, for $i,\ j = 1,2,3$, is given by Lemma 2. The following Theorem 1 gives the boundary probability vectors $\boldsymbol{P}_0$, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$.

Theorem 1. The boundary probability vectors are given by

$$\boldsymbol{P}_0 = \boldsymbol{P}_2\boldsymbol{M}_0, \tag{20}$$

$$\boldsymbol{P}_1 = \boldsymbol{P}_2\boldsymbol{M}_1, \tag{21}$$

and $\boldsymbol{P}_2$ is determined by the following equations:

$$\begin{cases} \boldsymbol{P}_2(\boldsymbol{M}_1\boldsymbol{B}_{12} + \boldsymbol{A}_1 + \boldsymbol{R}\boldsymbol{A}_2) = 0 \\ \boldsymbol{P}_2\left[\boldsymbol{M}_0\boldsymbol{e} + \boldsymbol{M}_1\boldsymbol{e}_1 + (\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e}\right] = 1 \end{cases} \tag{22}$$

where

$$\boldsymbol{M}_1 = -\frac{1}{|\boldsymbol{D}|}\boldsymbol{F}_1, \tag{23}$$

$$\boldsymbol{M}_0 = -\frac{1}{\lambda(\lambda+\alpha_0+\beta)|\boldsymbol{D}|}\boldsymbol{F}_1\boldsymbol{F}_0. \tag{24}$$

Proof. Note that $\boldsymbol{B}_{00}$ is invertible, and from Eq. (2) we get that

$$\boldsymbol{P}_0 = -\boldsymbol{P}_1\boldsymbol{B}_{10}\boldsymbol{B}_{00}^{-1}. \tag{25}$$

Substituting Eq. (25) into Eq. (3), we get

$$\boldsymbol{P}_1(\boldsymbol{B}_{11} - \boldsymbol{B}_{10}\boldsymbol{B}_{00}^{-1}\boldsymbol{B}_{01}) = -\boldsymbol{P}_2\boldsymbol{B}_{21}. \tag{26}$$

From Lemma 2, we know that $\boldsymbol{D} = \boldsymbol{B}_{11} - \boldsymbol{B}_{10}\boldsymbol{B}_{00}^{-1}\boldsymbol{B}_{01}$ is invertible. Thus, Eq. (26) yields that

$$\boldsymbol{P}_1 = -\boldsymbol{P}_2\boldsymbol{B}_{21}\boldsymbol{D}^{-1}. \tag{27}$$

Substituting Eq. (27) into Eq. (25), we get

$$\boldsymbol{P}_0 = \boldsymbol{P}_2\boldsymbol{B}_{21}\boldsymbol{D}^{-1}\boldsymbol{B}_{10}\boldsymbol{B}_{00}^{-1}. \tag{28}$$

After simple manipulations, we have

$$-\boldsymbol{B}_{21}\boldsymbol{D}^{-1} = -\frac{1}{|\boldsymbol{D}|}\boldsymbol{F}_1, \tag{29}$$

$$\boldsymbol{B}_{21}\boldsymbol{D}^{-1}\boldsymbol{B}_{10}\boldsymbol{B}_{00}^{-1} = -\frac{1}{\lambda(\lambda+\alpha_0+\beta)|\boldsymbol{D}|}\boldsymbol{F}_1\boldsymbol{F}_0. \tag{30}$$

Substitute Eqs. (29) and (30) into Eqs. (27) and (28), we obtain $\boldsymbol{P}_0$ and $\boldsymbol{P}_1$ which are given by Eqs. (20) and (21). Substitute Eqs. (20) and (21) into Eqs. (5) and (6), we get Eq. (22). This proves Theorem 1.

3.4 Performance Measures

The boundary probabilities $\boldsymbol{P}_0$, $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and probabilities $\boldsymbol{P}_i$ for $i \geq 3$ can be used to find the stationary distribution of the number of customers in the system. Let N denote the number of customers in the system at an arbitrary time under the stationary condition.

Theorem 2. The stationary distribution of the number of customers in the system is given by

$$P\{N=0\} = \boldsymbol{P}_0\boldsymbol{e}, \tag{31}$$

$$P\{N=1\} = \boldsymbol{P}_1\boldsymbol{e}_1, \tag{32}$$

$$P\{N=i\} = \boldsymbol{P}_2\boldsymbol{R}^{i-2}\boldsymbol{e}, \quad i \geq 2 \tag{33}$$

where the boundary probabilities $\boldsymbol{P}_0$, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are given by Theorem 1 in Subsection 3.3.

Proof. Note that

$$P\{N=0\} = P_{00} + P_{02} = \boldsymbol{P}_0\boldsymbol{e},$$

$$P\{N=1\} = P_{10} + P_{11} + P_{12} = \boldsymbol{p}_1\boldsymbol{e}_1,$$

we obtain Eqs. (31) and (32). Eq. (33) is obtained by noting that

$$P\{N=i\} = P_{i0} + P_{i2} = \boldsymbol{P}_i\boldsymbol{e}, \quad i \geq 2,$$

and using Eq. (5). This proves Theorem 2.

Let $\boldsymbol{\varepsilon}_1 = (1,0)^T$ and $\boldsymbol{\varepsilon}_2 = (0,1)^T$ represent two identity column vectors of order 2, and let $\boldsymbol{\delta}_1 = (1,0,0)^T$, $\boldsymbol{\delta}_2 = (0,1,0)^T$ and $\boldsymbol{\delta}_3 = (0,0,1)^T$ represent three identity column vectors of order 3. From Theorem 1 and Theorem 2, we can obtain some other performance measures which are given by the following corollary. Since the proof is simple, the details of the proof are omitted.

Corollary.
(a) The mean system size $E[N]$ is as follows:

$$E[N] = \boldsymbol{P}_2\left[\boldsymbol{M}_1\boldsymbol{e}_1 + (\boldsymbol{I}-\boldsymbol{R})^{-1}\boldsymbol{e} + (\boldsymbol{I}-\boldsymbol{R})^{-2}\boldsymbol{e}\right]. \tag{34}$$

(b) The mean number $E[N_q]$ of customers in the queue is as follows:

$$E[N_q] = \boldsymbol{P}_2 (\boldsymbol{I} - \boldsymbol{R})^{-2} \boldsymbol{e}. \tag{35}$$

(c) The average balking rate B_r of customers is as follows:

$$B_r = \lambda \boldsymbol{P}_2 \left\{ (\boldsymbol{I} - \boldsymbol{R})^{-1} [(1-b_0)\boldsymbol{\varepsilon}_1 + (1-b_1)\boldsymbol{\varepsilon}_2] + (1-b_1)\boldsymbol{M}_1 \boldsymbol{\delta}_3 \right\}. \tag{36}$$

(d) The probability P_f that Server 2 is free is given by

$$P_f = \boldsymbol{P}_2 (\boldsymbol{M}_0 \boldsymbol{\varepsilon}_1 + \boldsymbol{M}_1 \boldsymbol{\delta}_1). \tag{37}$$

(e) The probability P_b that Server 2 is busy is given by

$$P_b = \boldsymbol{P}_2 \left[\boldsymbol{M}_1 \boldsymbol{\delta}_2 + (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{\varepsilon}_1 \right]. \tag{38}$$

(f) The probability P_d that Server 2 is broken down is given by

$$P_d = \boldsymbol{P}_2 \left[\boldsymbol{M}_0 \boldsymbol{\varepsilon}_2 + \boldsymbol{M}_1 \boldsymbol{\delta}_3 + (\boldsymbol{I} - \boldsymbol{R})^{-1} \boldsymbol{\varepsilon}_2 \right]. \tag{39}$$

4 Numerical Illustrations

In order to explore the effect of various system parameters on the mean system size, some numerical experiments are performed and the results are displayed by graphs. MATLAB software was used to develop the computer program.

In Figs. 1-3, the mean system size $E[N]$ is plotted against the arrival rate λ for chosen values of the service rates μ_1, μ_2, the failure rates α_0, α, and with the joining probabilities b_0, b_1 satisfying the stationary condition $\rho < 1$.

In Fig. 1, we fix $b_0 = 0.9$, $b_1 = 0.5$, $\alpha_0 = 6$, $\alpha = 8$, $\beta = 10$, and $p = 0.7$. The mean system size $E[N]$ is plotted against the arrival rate λ for chosen values of μ_1 and μ_2. Figure 1 shows that the mean system size for the case of $\mu_1 = 15$ and $\mu_2 = 15$ is the smallest among all five cases. This is because that the total service rate $\mu_1 + \mu_2 = 30$ for this case is the largest among all five cases. We observed from Fig. 1 that the mean system size for the case of $\mu_1 = 5$ and $\mu_2 = 15$ is much larger than that for the case of $\mu_1 = 15$ and $\mu_2 = 5$ although the total service rate for each of the two cases is the same $\mu_1 + \mu_2 = 20$. This is because when a customer arrives to find both servers free and available, the customer chooses the faster server, Server 2, in the first case with probability 0.3 and the faster server, Server 1, in the second case with probability 0.7. This is why the mean system size for the first case is larger than that for the second case. As expected, the mean system size increases with the increasing of the arrival rate λ, while it decreases with the increasing of the service rate μ_1 or μ_2 for each server.

In Fig. 2, we fix $\mu_1 = 15$, $\mu_2 = 10$, $\alpha_0 = 6$, $\alpha = 8$, $\beta = 10$, and $p = 0.7$. The mean system size $E[N]$ is plotted against the arrival rate λ for chosen values of b_0 and b_1. Figure 2 shows that the mean system size increases with the increasing of the joining probability b_0 or b_1. This is because the larger the probability b_0 or b_1 is, the more customers are allowed to join the system, which results in the increasing of the mean system size. The graphs of the five cases indicate that the differences in the mean system size steadily

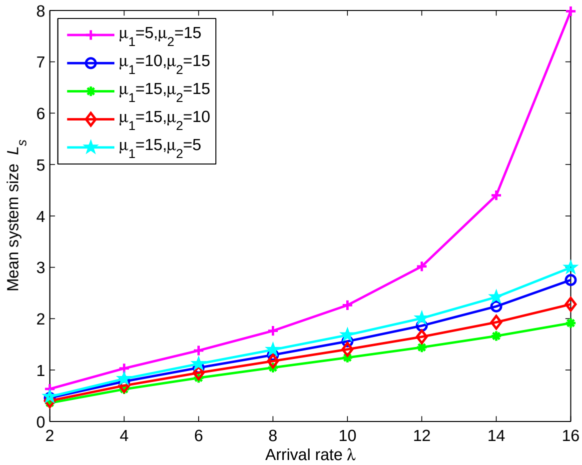

Figure 1: Mean system size $E[N]$ versus arrival rate λ for different values of μ_1 and μ_2.

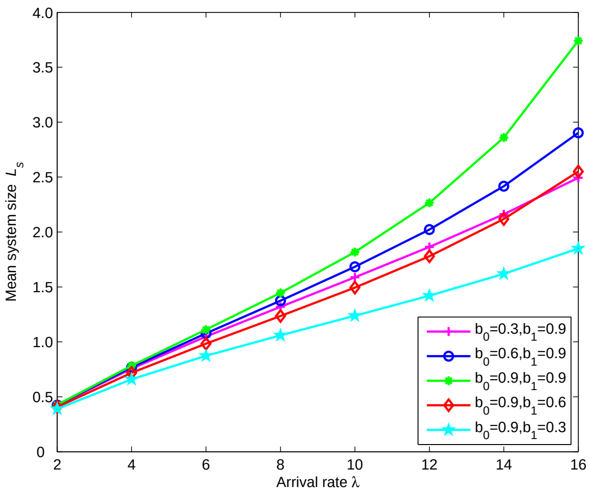

Figure 2: Mean system size $E[N]$ versus arrival rate λ for different values of b_0 and b_1.

decreases with a decreasing arrival rate. This can be explained by noting the fact that for most of the time when the servers are free, the arrival rate is much less than the service rate. Therefore, the mean system size will be very small when the arrival rate is small enough, which results in a decrease in the differences among the mean system sizes for decreasing arrival rates.

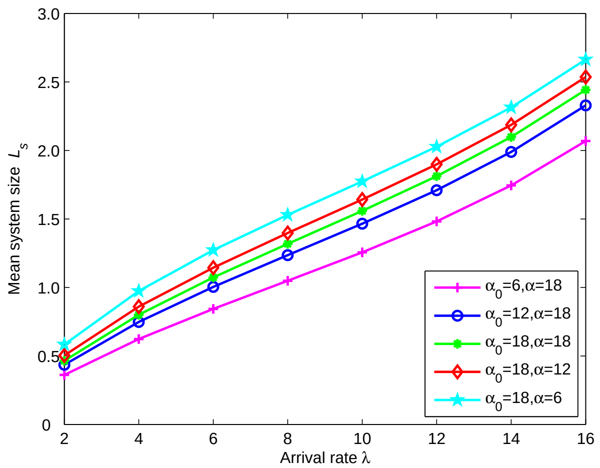

Figure 3: Mean system size $E[N]$ versus arrival rate λ for different values of α_0 and α.

In Fig. 3, we fix $\mu_1 = 15$, $\mu_2 = 10$, $b_0 = 0.9$, $b_1 = 0.5$, $\beta = 10$, and $p = 0.7$. The mean system size $E[N]$ is plotted against the arrival rate λ for chosen values of α_0 and α. It is observed from Fig. 3 that the mean system size increases with the increasing of the failure rate α_0 and decreases with the increasing of the failure rate α. This can be explained by noting the fact that α_0 and α are the different failure rates for Server 2 in its idle time and its busy time, respectively. On one hand, in the busy time of Server 2, the larger the failure rate α is, the smaller the availability of Server 2 is. This results in an increase in the mean system size. On the other hand, the more customers balk (are not allowed to join the system) the more the mean system size increases. This results in a decreasing mean system size. When the increasing mean system size due to the increasing failure rate α is less than the decreasing mean system size due to the balking of customers, the mean system size will decrease as the failure rate α increases. However, this is not so for the case of failure rate α_0. In the idle time of Server 2, the larger the failure rate α_0 is, the smaller the availability of Server 2 is. Also, the arriving customers will not balk, since there is at least one free and available server in the idle time of Server 2. This results in an increase in the mean system size. This is why the mean system size increases as the failure rate α_0 increases.

5 Conclusions

In this paper, a heterogeneous two-server queueing system with balking and server breakdowns was studied. We extended the model in [13] by considering server breakdowns. The model investigated in this paper is more realistic for modeling queueing situations where the server may experience many types of breakdowns which can be realized in manufacturing or production systems. The matrix-geometric solution method has been used in this paper for obtaining the stationary condition and some performance measures such as the stationary distribution of the number of customers in the system and the mean system size. We finally performed numerical experiments to explore the effect of various system parameters on the mean system size.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 70671088), and was supported in part by GRANT-IN-AID FOR SCIENCE RESEARCH (No. 21500086) and the Hirao Taro Foundation of KUAAR, Japan.

References

[1] A. Federgruen and L. Green, Queueing systems with service interruptions, Operations Research, 34 (1986), 752-768.

[2] W. Li, D. Shi and X. Chao, Reliability analysis of M/G/1 queueing systems with server breakdowns and vacations, Journal of Applied Probability, 34 (1997), 546-555.

[3] Y. Tang, Single-server M/G/1 queueing system subject to breakdowns -some reliability and queueing problems, Microelectronics Reliability, 37 (1997), 315-321.

[4] O. Nakdimon and U. Yechiali, Polling systems with breakdowns and repairs, European Journal of Operational Research, 149 (2003), 588-613.

[5] K. Wang, T. Wang and W. Pearn, Optimal control of the N policy M/G/1 queueing system with server breakdowns and general startup times, Applied Mathematical Modelling, 31 (2007), 2199-2212.

[6] J. Wang, B. Liu and J. Li, Transient analysis of an M/G/1 retrial queue subject to disasters and server failures, European Journal of Operational Research, 189 (2008), 1118-1132.

[7] G. Choudhury and L. Tadj, An M/G/1 queue with two phases of service subject to the server breakdown and delayed repair, Applied Mathematical Modelling, 33 (2009), 2699-2709.

[8] I. L. Mitrany and B. Avi-Itzhak, A many-server queue with service interruptions, Operations Research, 16 (1968), 628-638.

[9] B. Vinod, Unreliable queueing systems, Computers and Operations Research, 12 (1985), 322-340.

[10] M. F. Neuts and D. M. Lucantoni, A Markovian queue with N servers subject to breakdowns and repairs, Management Science, 25 (1979), 849-861.

[11] P. Wartenhosrt, *N* parallel queueing systems with server breakdown and repair, European Jopurnal of Operational Research, 82 (1995), 302-322.

[12] K. H. Wang and Y. C. Chang, Cost analysis of a finite M/M/R queueing system with balking, reneging and server breakdowns, Mathematical Methods of Operations Research, 56 (2002), 169-180.

[13] V. P. Singh, Two-server Markovian queues with balking: heterogeneous vs. homogeneous servers, Operations Research, 18 (1970), 145-159.

[14] B. K. Kumar and S. P. Madheswari, An M/M/2 queueing system with heterogeneous servers and multiple vacations, Mathematical and Computer Modelling, 41 (2005), 1415-1429.

[15] D. Yue, J. Yu and W. Yue, A Markovian queue with two heterogeneous servers and multiple vacations, Journal of Industrial and Management Optimization, 5 (2009), 453-465.

[16] K. C. Mandan, W. Abu-Dayyeh and F. Taiyyan, A two server queue with Bernoulli schedules and a single vacation policy, Applied Mathematics and Computation, 145 (2003), 59-71.

[17] M. F. Neuts, Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach, Baltimore, MD: Johns Hopkins University Press, 1981.

[18] J. Yu, D. Yue and R. Tian, Matrix-geometric solution of a repairable queueing system with two heterogeneous servers, Acta Mathematicae Applicate Sinica (in Chinese), 31 (2008), 894-900.