ECONSTOR
Make Your Publications Visible.

A Service of
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Chakravarthy, Srinivas R.; Shruti; Kulshrestha, Rakhee

Article

A queueing model with server breakdowns, repairs, vacations, and backup server

Operations Research Perspectives

Provided in Cooperation with:
Elsevier

Suggested Citation: Chakravarthy, Srinivas R.; Shruti; Kulshrestha, Rakhee (2020) : A queueing model with server breakdowns, repairs, vacations, and backup server, Operations Research Perspectives, ISSN 2214-7160, Elsevier, Amsterdam, Vol. 7, pp. 1-13, https://doi.org/10.1016/j.orp.2019.100131

This Version is available at:
https://hdl.handle.net/10419/246405

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by-nc-nd/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

Operations Research Perspectives 7 (2020) 100131

Contents lists available at ScienceDirect

Operations Research Perspectives

journal homepage: www.elsevier.com/locate/orp

A queueing model with server breakdowns, repairs, vacations, and backup server

Srinivas R. Chakravarthy[*,a], Shruti[b], Rakhee Kulshrestha[b]

[a] Departments of Industrial and Manufacturing Engineering & Mathematics Kettering University, Flint MI-48504, USA
[b] Department of Mathematics Birla Institute of Technology and Science, Pilani, 333031 India

ARTICLE INFO

Keywords:
Markovian arrival process
Working vacation
Server breakdown and repairs
Phase type distribution
Matrix analytical method

ABSTRACT

Queueing models wherein having a backup server during the absences (caused by vacations and breakdowns) of the main server have found many applications in practice. Such services offered by a backup server can be viewed as the (main) server working during a vacation or during a breakdown period. A backup server offering services can be thought of as the main server working (at a reduced rate) on vacations/repairs. Using Neuts' versatile point process for the arrivals and modeling the service times with phase type distributions, we propose a model that generalizes some of the previously published ones on working-vacation-breakdown-repair queues. We carry out the steady-state analysis and report interesting illustrative numerical examples. We also prove decomposition results for the rate matrix and the mean number in the system under some special cases.

1. Introduction

Queueing models where the server goes on vacations have applications in many areas, such as production and manufacturing systems, telecommunications, and service industries. The server can take vacations after being busy for a certain period of time or after serving a (fixed or random) number of customers/packets. The vacations have many purposes including catering to personal needs or other work-related activities. Some of the work-related activities include helping fellow co-workers and attending to clear overhead work left behind by the departed customers. Queueing models which consider the server vacations have been studied considerably by several authors [5,6,8,13,18,23,25,50].

The practical examples of vacation queueing models as well as queueing models where the server is subject to breakdowns and repairs have been well established in survey papers (see, e.g., [13]) and in the book ([50] - see Section 1.1). For example, in service sectors, it is common for the service provider to breakdown for many reasons among which is the failure to access their database to provide solutions to the customer's requests. During such times (which correspond to failures), the service provider might try to access a backup system and or collect more information from the customer. This duration, corresponding to bringing the database to accessible level, referred to as the repair time, the services are at a lower level. Our model is also applicable in a manufacturing, as the machines are subject to breakdowns and or under preventive maintenance schedule. These events can be considered as vacation times for the machines. At the same time the production cannot be stopped completely, as it may result in a delay in completing the orders; therefore, machines or operators with lower processing rate can work during such vacations. Thus, the model studied in our paper is very generic and can be applied in manufacturing and service sectors.

An example of a vacation model in telecommunication applications involve the ever-increasing use of wireless cellular networks, which triggers a huge consumption of energy. In order to develop energy efficient wireless cellular networks, researchers proposed hibernation (or sleeping) of a base station (*BS*) in the absence of an active user in the network. A sleeping *BS* is similar to a server on a vacation. There are a few vacation based *BS* sleeping models available in the literature (see, e.g. [41,46,51]). The study of these models has proven that one can save a significant amount of energy through such vacations. The authors in [52] studied a *BS* sleep model by imposing a condition that the system should have N or more customers in order for the vacation-returned server to offer services. This restriction increases the average amount of delay for the customers. The server-working-vacation models were introduced in order to avoid excessive delays in vacation models.

In [56], the queue length distribution for $MAP/G/1$ working-vacation-interruption queue was derived. Therein, the authors considered the working vacation to end when the current service gets over, and at that time the system has one or more customers waiting; otherwise, the server will continue to offer services during a vacation period. In the

* Corresponding author.
E-mail address: schakrav@kettering.edu (S.R. Chakravarthy).

https://doi.org/10.1016/j.orp.2019.100131
Received 7 August 2019; Received in revised form 21 November 2019; Accepted 25 November 2019
Available online 27 November 2019
2214-7160/ © 2019 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/BY-NC-ND/4.0/).

context of $GI/M/1$ queue with single working vacation, the waiting time is analyzed in [30]. The $GI/M/1/N-$type model with finite buffer and single working vacation policy is investigated in [26]. The transient queue-size distribution is derived there. The authors in [47] investigated an $MAP/PH/1-$type queue by incorporating an N-policy for vacations. Further, a working vacation model wherein there are two types of priority customers, a possibility of interrupting the vacations, and with services following a phase type ($PH-$) distribution was studied in [15]. A survey article, wherein a separate section dealing with Markovian arrival process (*MAP*), on *working vacation models* can be found in [11]. Later, Goswami and Selvarju [16] analyzed a $PH/M/c$ queue, where servers can go for multiple working vacations. The customers arriving during vacation periods become impatient as their waiting times in the queue exceed a pre-determined threshold, and hence leave the system. Recently, Ozkar and Kocer [43] presented an $M/C_k/1$ model with multiple working vacations.

The assumption that the service environment to be one hundred percent reliable can be impractical, especially in production and manufacturing, and communications systems. The service area can break down due to external shocks leading to the study of stochastic models with server breakdowns and repairs ([21,22,24]). In [53], the authors investigated an $M/M/1$ queueing system with working vacation and server breakdown under an N-policy. They considered different rates of breakdowns during working vacations, busy and idle periods. The $BMAP/PH/N-$type unreliable queue with infinite buffer was analyzed in [28]. Therein, breakdown occurrence moments are defined by the *MAP*, repair times follow $PH-$distribution and stationary characteristics are derived then. The model in which the server, after a shock, may fail partially, and thus offer services at a reduced rate up until restored to normalcy was studied in [48]. In [20], the term *working breakdown* was coined to describe this phenomenon. Further, the authors considered the provision of a backup server to serve at a reduced rate when the main server is unavailable. The authors in [31] analyzed equilibrium strategies for an $M/M/1$ queue with working breakdowns. A queue with working breakdowns and disasters was studied in [27]. The authors in [19] looked at a finite capacity queue with working breakdown and Bernoulli feedback. A finite capacity working-breakdown-reneging queueing system along with the retention of impatient customers was investigated in [54]. Later, Ye and Liu [55] studied $MAP/M/1-$type queue with working breakdowns and repairs. In [12], the steady-state probabilities for a finite working-vacation-working-breakdown queue were derived. There are some recent queueing models that consider *MAP* arrivals, $PH-$services in the context of vacations, optional services, and crowdsourcing (see, e.g. [9,10]).

Our working-vacation-breakdown-repair queueing model in this paper is applicable to cellular networks, as we know that in cellular network each cell has a base station that controls the call admissions and the quality of service of the network. If we want to model the base station properly and adequately, then we should consider the possibility of many users (customers) accessing the Internet on their mobiles at the same time. This creates the traffic in the network to be bursty at times, and the use of *MAP* to model the arrivals makes significant sense. The services provided by the base station controller can be modeled as $PH-$services. This choice is due to the fact that when the caller requests for a connection, the base station controller needs to (a) verify the caller; (b) check the current location of the caller; (c) search for a route to be assigned; and after these phases, the connection is made. As any other electronic component, the base station controller is also exposed to risks due to external shocks, and therefore subject to breakdowns. But at the same time, the services to mobile users are very important. Hence, the service providers cannot afford full interruptions in their services leading to backup servers being relied upon to provide services at reduced rates whenever the main sever is under repair. This is referred as working breakdowns. The base stations consume a large amount of energy for serving their customers. In order to save significant amount of energy, a base station can go on a vacation (i.e, a sleep mode of the base station) when there is no call waiting in the network. In such situations a backup server needs to be initiated (possibly serving at reduced rates) to serve arriving new calls and this is known as working vacations. While there are many variants of the arrival/service systems studied thus far, to the best of our knowledge, models using the Markovian arrivals and $PH-$services with working vacations and working breakdowns are studied for the first time in this paper.

In this paper, we look at a $MAP/PH/1$ type queues wherein the server (i) takes vacations (whenever the system becomes empty), (ii) breaks down (due to external shocks) and (iii) gets repaired. The backup server takes over serving the customers at a reduced rate whenever the main server is either on vacation or under repair. Upon the availability of the main server (either through returning from a vacation or a repair completion), the existing service rate will be set back to the normal rate. This model generalizes the recently studied one in [55]. The use of $PH-$distributions enables one to model a variety of distributions for the services. For example, some classical distributions such as generalized Erlang and finite mixtures of exponentials among others are very special cases of $PH-$ distributions. The durations of vacations are modeled with exponential distribution. It is possible for the (main) server to take multiple vacations due to no customers seen waiting at the time of the server returning from a vacation. The effect of the variation in the services in the context of vacation/working vacation models has not been fully explored and we focus on this aspect also in this paper.

The layout of the paper is described now. The model considered here is described in full detail in Section 2. The analysis in steady-state of the proposed model is presented in Section 3. The stationary sojourn time distribution and the effective service time of a customer are, respectively, obtained in Sections 4 and 5. A discussion of a few other performance measures is presented in Section 6 and Section 7 deals with some special cases. Section 8 is devoted to the decomposition results for the case of the $M/M/1$ model and are shown to be in agreement with the previously published ones. Some illustrative examples to point out the effects of having a backup server are presented in Section 9, and a few concluding remarks are given in Section 10.

For the rest of the paper, we adopt the following standard notation.

- The symbol $'$ stands for the transpose notation.
- The notations $\otimes$ and $\oplus$ stand for the Kronecker product and Kronecker sum, respectively (see, eg., [35,49]).
- $\mathbf{e}' = (1, 1, ..., 1)$, whose dimension should be clear in the context. Where more clarity is needed, the dimension will be mentioned, e.g., $\mathbf{e}(m)$ is a column vector of 1's of dimension m.
- $\mathbf{e}_i' = (0, 0, ..., 1, 0, ..., 0)$, where 1 is in the ith position.
- I denotes an identity matrix, whose dimension is dictated by the context.

2. Model description

The model considered in this paper is outlined in this section. The customers enter singly into the system according to Neuts' versatile point process [37], namely *MAP*, which is described by two matrices, D_0 and D_1. Thus, (D_0, D_1) of order m represents the *MAP* to describe the arrivals. While the matrix D_0 is to model the transitions pointing to no arrivals, the transitions within D_1 point to the arrivals. Assume $D = D_0 + D_1$ to be an irreducible matrix with an stationary vector $\boldsymbol{\delta}$. That is,

$$\boldsymbol{\delta} D = \mathbf{0}, \boldsymbol{\delta}\mathbf{e} = 1. \tag{1}$$

The average rate of arrivals is $\lambda = \boldsymbol{\delta} D_1 \mathbf{e}$. For full details on the rich class of point processes introduced with simpler notation in [34], we refer the reader to [2,4,7,33,38,39]. It is worth pointing out that among many well-known processes that are special cases of *MAP*, the Markov-modulated Poisson process (*MMPP*) stands out due to its ability to

model "burstiness" in arrival processes, especially in teletraffic modeling.

The system has a single server, which will be referred to as the main server. In the sequel, we interchangeably use the terms server and main server to refer to this server. The server is exposed to external shocks that occur according to a Poisson process with rate γ. These shocks will have an effect on the server only when the server is busy serving a customer. When the shock has an effect, the server breaks down and is sent for repair immediately, and at that moment the backup server instantaneously takes over the serving the customer but at a reduced rate. The repair times are exponentially distributed with rate ξ. Upon finding the system to be empty, the main server goes on a vacation. It is possible for the server to go on multiple vacations. The duration of each vacation is modeled with an exponential distribution with rate η. Note that the backup server is not subject to breakdowns. The backup server is available on a standby basis to take over serving the customers whenever the main server is on a vacation or under repair. Upon completion of a repair or a vacation, the main server either takes over the service from the backup server instantaneously or goes on a vacation due to the system being empty.

The service times of the main server as well as the backup server are modeled using $PH-$ distributions with representations, respectively, $(\boldsymbol{\beta}, S)$ and $(\boldsymbol{\beta}, \theta S)$, of order n, where $0 < \theta < 1$. Note that the service rates of the main and the backup servers are given by μ and $\theta\mu$, respectively, where $\mu = [\boldsymbol{\beta}(-S)^{-1}\mathbf{e}]^{-1}$. Moreover, we denote $\mathbf{S}^0$ as the column vector satisfying $S\mathbf{e} + \mathbf{S}^0 = \mathbf{0}$. The underlying random variables governing the durations of the inter-arrivals, the services, the vacations, the repairs, and the shocks are all assumed to be mutually independent of each other.

2.1. The Markov process

We now define a number of variables needed to describe the model under study. Let

- $N(t)$: number of customers in the system;
- $J_1(t)$: server status;
- $J_2(t)$: service phase, if any;
- $J_3(t)$: arrival phase;

at time t, where

$$J_1(t) = \begin{cases} 0; & \text{if main server is busy,} \\ 1; & \text{if main server is vacationing,} \\ 2; & \text{if main server is under repair.} \end{cases}$$

It is easy to verify that the process $\{N(t), J_1(t), J_2(t), J_3(t): t \geq 0\}$ is a Markov process possessing the *QBD*-structure on the state space $\Omega = \bigcup_{i=0}^{\infty} r(i)$, where $r(0) = \{(0, j_1, j_3), 1 \leq j_1 \leq 2, 1 \leq j_3 \leq m\}$ and $r(i) = \{(i, j_1, j_2, j_3): j_1 = 0, 1 \text{ or } 2, 1 \leq j_2 \leq n, 1 \leq j_3 \leq m\}$, $i \geq 1$. The model described above is studied as a quasi-birth-and-death process (*QBD*).

By level $\underline{\mathbf{i}}$, for $i > 0$, we denote the set of $3mn$ states given by $\underline{\mathbf{i}} = \{(i, j_1, j_2, j_3): j_1 = 0, 1 \text{ or } 2, 1 \leq j_2 \leq n, 1 \leq j_3 \leq m\}$. Note that when the system is empty (i.e., when $N(t) = 0$), $J_2(t)$ is undefined. A pictorial description of the transition diagram of the system is shown in Fig. 1.

The *QBD*−process has the (infinitesimal) generator, Q, given by

$$Q = \begin{pmatrix} B_1 & B_0 & & & \\ B_2 & A_1 & A_0 & & \\ & A_2 & A_1 & A_0 & \\ & & A_2 & A_1 & A_0 \\ & & & \ddots & \ddots & \ddots \end{pmatrix}, \tag{2}$$

where the (block) entries in Q are as follows.

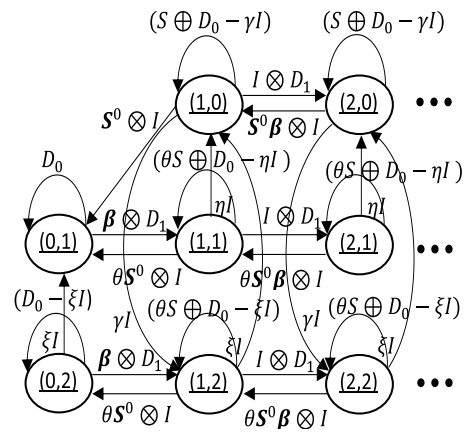

Fig. 1. State transition diagram of the system.

$$B_1 = \begin{bmatrix} D_0 & 0 \\ \xi I_m & D_0 - \xi I_m \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 & \boldsymbol{\beta} \otimes D_1 & 0 \\ 0 & 0 & \boldsymbol{\beta} \otimes D_1 \end{bmatrix}, \tag{3}$$

$$B_2 = \begin{bmatrix} \mathbf{S}^0 \otimes I_m & 0 \\ \theta\mathbf{S}^0 \otimes I_m & 0 \\ 0 & \theta\mathbf{S}^0 \otimes I_m \end{bmatrix},$$

$$A_2 = \begin{bmatrix} \mathbf{S}^0\boldsymbol{\beta} \otimes I_m & 0 & 0 \\ 0 & \theta\mathbf{S}^0\boldsymbol{\beta} \otimes I_m & 0 \\ 0 & 0 & \theta\mathbf{S}^0\boldsymbol{\beta} \otimes I_m \end{bmatrix}, \tag{4}$$

$$A_1 = \begin{bmatrix} S \oplus D_0 - \gamma I_{mn} & 0 & \gamma I_{mn} \\ \eta I_{mn} & \theta S \oplus D_0 - \eta I_{mn} & 0 \\ \xi I_{mn} & 0 & \theta S \oplus D_0 - \xi I_{mn} \end{bmatrix},$$

$$A_0 = I_{3n} \otimes D_1, \tag{5}$$

where B_1 is a square matrix of dimension 2m; B_0 and B_2 are, respectively, rectangular matrices of dimension 2m × 3mn and 3mn × 2m; and A_0, A_1 and A_2 are of dimension 3mn.

Lemma 1. *The stationary probability vector, $\boldsymbol{\pi}$ partitioned as $\boldsymbol{\pi} = (\boldsymbol{\pi}_0, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$, of the (reducible) generator $A = A_0 + A_1 + A_2$ is given by*

$$\boldsymbol{\pi}_0 = \frac{\xi\mu}{\xi + \gamma}(\boldsymbol{\beta}(-S)^{-1} \otimes \boldsymbol{\delta}), \quad \boldsymbol{\pi}_1 = \mathbf{0}, \quad \boldsymbol{\pi}_2 = \frac{\gamma\mu}{\xi + \gamma}(\boldsymbol{\beta}(-S)^{-1} \otimes \boldsymbol{\delta}), \tag{6}$$

where $\boldsymbol{\delta}$ is the stationary vector of D given in (1).

Proof. First note that the matrix A is given by

$$A = \begin{bmatrix} (S + \mathbf{S}^0\boldsymbol{\beta}) \oplus D - \gamma I_{mn} & 0 & \gamma I_{mn} \\ \eta I_{mn} & (\theta[S + \mathbf{S}^0\boldsymbol{\beta}]) \oplus D - \eta I_{mn} & 0 \\ \xi I_{mn} & 0 & (\theta[S + \mathbf{S}^0\boldsymbol{\beta}]) \oplus D - \xi I_{mn} \end{bmatrix}. \tag{7}$$

The steady-state equations corresponding to $\boldsymbol{\pi}A = 0$ and $\boldsymbol{\pi}\mathbf{e} = 1$, can be written as

$$\begin{aligned} \boldsymbol{\pi}_0[(S + \mathbf{S}^0\boldsymbol{\beta}) \oplus D - \gamma I_{mn}] + \eta\boldsymbol{\pi}_1 + \xi\boldsymbol{\pi}_2 &= \mathbf{0}, \\ \boldsymbol{\pi}_1[\theta(S + \mathbf{S}^0\boldsymbol{\beta}) \oplus D - \eta I_{mn}] &= \mathbf{0}, \\ \gamma\boldsymbol{\pi}_0 + \boldsymbol{\pi}_2[\theta(S + \mathbf{S}^0\boldsymbol{\beta}) \oplus D - \xi I_{mn}] &= \mathbf{0}, \\ \boldsymbol{\pi}_0\mathbf{e} + \boldsymbol{\pi}_1\mathbf{e} + \boldsymbol{\pi}_2\mathbf{e} &= 1. \end{aligned} \tag{8}$$

It is clear from the second equation in (8) that

$$\boldsymbol{\pi}_1 = \mathbf{0}, \tag{9}$$

and using this fact and adding the equations in (8), we get

$$\boldsymbol{\pi}_0[(S + \mathbf{S}^0\boldsymbol{\beta}) \oplus D] + \boldsymbol{\pi}_2[\theta(S + \mathbf{S}^0\boldsymbol{\beta}) \oplus D] = \mathbf{0}. \tag{10}$$

Again, it is clear from (10) that π_0 and π_2 are of the form

$$\pi_0 = c\mu(\boldsymbol{\beta}(-S)^{-1} \otimes \boldsymbol{\delta}), \quad \pi_2 = d\mu(\boldsymbol{\beta}(-S)^{-1} \otimes \boldsymbol{\delta}), \tag{11}$$

where c and d satisfy

$$\gamma c = \xi d \quad and \quad c + d = 1. \tag{12}$$

The claimed result in (6) follows from (12). □

Now we are ready to prove the stability condition of the working-vacation-breakdown-repair queueing model under study in Theorem 1 below.

Theorem 1. *The working-vacation-breakdown-repair queueing model with the generator given in (2) is stable if and only if*

$$\lambda < \mu\left[\frac{\xi + \theta\gamma}{\xi + \gamma}\right]. \tag{13}$$

Proof. The proof follows immediately on applying the stability condition, $\pi A_0 \mathbf{e} < \pi A_2 \mathbf{e}$ (see, e.g., [40]), for the $QBD-$process in (2). □

The above stability condition can be explained intuitively. Towards this end, we rewrite the equation in (13) as

$$\lambda < \frac{\mu\xi}{\xi + \gamma} + \frac{\theta\mu\gamma}{\xi + \gamma}. \tag{14}$$

The effective service rate is the quantity on the right-hand side of (14), which is obtained as the sum of the rates of the services offered by the main and the backup servers. Obviously, the arrival rate has to be less than the effective service rate in order for the queue to be stable.

3. The stationary probability vector

Let $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots)$, where $\mathbf{x}_0$ is of dimension $2m$ and $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are of dimension $3mn$, satisfy

$$\mathbf{x}Q = 0, \quad \mathbf{x}\mathbf{e} = 1. \tag{15}$$

When the working-vacation-breakdown-repair queueing under study is stable (i.e., condition in (13) is satisfied), the stationary probability vector $\mathbf{x}$ is derived as (see, e.g., [40]):

$$\mathbf{x}_i = \mathbf{x}_1 R^{i-1}, \; i \geq 1. \tag{16}$$

The matrix R is the minimal nonnegative solution to:

$$R^2 A_2 + RA_1 + A_0 = 0, \tag{17}$$

and $\mathbf{x}_0$ and $\mathbf{x}_1$ are found by solving

$$\begin{aligned} &\mathbf{x}_0 B_1 + \mathbf{x}_1 B_2 = \mathbf{0}, \\ &\mathbf{x}_0 B_0 + \mathbf{x}_1 (A_1 + RA_2) = \mathbf{0}, \end{aligned} \tag{18}$$

subject to the (normalizing) condition

$$\mathbf{x}_0 \mathbf{e} + \mathbf{x}_1 (I_{3mn} - R)^{-1} \mathbf{e} = 1. \tag{19}$$

For later use, we partition $\mathbf{x}_0 = (\mathbf{v}_0, \mathbf{w}_0)$, and $\mathbf{x}_i = (\mathbf{u}_i, \mathbf{v}_i, \mathbf{w}_i)$, $i \geq 1$, such that $\mathbf{v}_0$ and $\mathbf{w}_0$ are of dimension m and $\mathbf{u}_i$, $\mathbf{v}_i$ and $\mathbf{w}_i$ are of dimension mn.

The following are the interpretations of the vectors in steady-state.

- $\mathbf{v}_0$: the main server is on vacation with no one in the system and the arrival is in one of m phases.
- $\mathbf{w}_0$: the main server is under repair with no one in the system and the arrival is in one of m phases.
- $\mathbf{u}_i$: the system has exactly i customers with the main server busy serving with the arrival and services are in various phases.
- $\mathbf{v}_i$: the system has exactly i customers with the main server on vacation with the arrival and services are in various phases.
- $\mathbf{w}_i$: the system has exactly i customers with the main server under repair with the arrival and services are in various phases.

The steady state equations given in (15) can be rewritten as

$$\mathbf{v}_0 D_0 + \xi \mathbf{w}_0 + \mathbf{u}_1(\mathbf{S}^0 \otimes I_m) + \mathbf{v}_1(\theta \mathbf{S}^0 \otimes I_m) = \mathbf{0}, \tag{20}$$

$$\mathbf{w}_0(D_0 - \xi I_m) + \mathbf{w}_1(\theta \mathbf{S}^0 \otimes I_m) = \mathbf{0}, \tag{21}$$

$$\mathbf{u}_1(S \oplus D_0 - \gamma I_{mn}) + \eta \mathbf{v}_1 + \xi \mathbf{w}_1 + \mathbf{u}_2(\mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) = \mathbf{0}, \tag{22}$$

$$\mathbf{v}_0(\boldsymbol{\beta} \otimes D_1) + \mathbf{v}_1(\theta S \oplus D_0 - \eta I_{mn}) + \mathbf{v}_2(\theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) = \mathbf{0}, \tag{23}$$

$$\mathbf{w}_0(\boldsymbol{\beta} \otimes D_1) + \gamma \mathbf{u}_1 + \mathbf{w}_1(\theta S \oplus D_0 - \xi I_{mn}) + \mathbf{w}_2(\theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) = \mathbf{0}, \tag{24}$$

$$\mathbf{u}_{i-1} D_1 + \mathbf{u}_i(S \oplus D_0 - \gamma I_{mn}) + \eta \mathbf{v}_i + \xi \mathbf{w}_i + \mathbf{u}_{i+1}(\mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) = \mathbf{0}, \quad i \geq 2, \tag{25}$$

$$\mathbf{v}_{i-1} D_1 + \mathbf{v}_i(\theta S \oplus D_0 - \eta I_{mn}) + \mathbf{v}_{i+1}(\theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) = \mathbf{0}, \quad i \geq 2, \tag{26}$$

$$\mathbf{w}_{i-1} D_1 + \gamma \mathbf{u}_i + \mathbf{w}_i(\theta \mathbf{S}^0 \oplus D_0 - \xi I_{mn}) + \mathbf{w}_{i+1}(\theta \mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) = \mathbf{0}, \quad i \geq 2, \tag{27}$$

along with the (normalizing) restriction

$$(\mathbf{u}_0 + \mathbf{v}_0)\mathbf{e} + \sum_{i=1}^{\infty} (\mathbf{u}_i + \mathbf{v}_i + \mathbf{w}_i)\mathbf{e} = 1. \tag{28}$$

The rate matrix R can be computed using a number of well-known techniques in the queueing literature. For example, with m and n to be relatively reasonable, R can be obtained using, say, logarithmic reduction (see e.g., [29]). Otherwise, one should employ (block) Gauss-Siedel iteration. For this, it may be worth to exploit the special structure of the coefficient matrices A_0, A_1, and A_2, considering that these are of dimension $3mn$. For example, the structure of the matrix R as presented in the following theorem can be exploited when m and n are reasonably large.

Theorem 2. *The structure of R is such that*

$$R = \begin{bmatrix} R_{11} & 0 & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & 0 & R_{33} \end{bmatrix}. \tag{29}$$

Proof. From the probabilistic interpretation of R matrix, it is clear that R should have the form as given in (29). For example, noting that away from the boundary states (which is what one is interested when computing R), the (main) server when busy serving customers cannot go on vacation without visiting the boundary states. However, the (main) server can go under repair through an arrival of a shock but in this case, the server has to get back to being busy through repair completion. This results as zero in the intersection of the first (block) row and second (block) column. Similarly one sees the other zero block. However, here we will give another proof that is constructive in nature. First, we rewrite Eq. (17) as

$$R = (R^2 A_2 + A_0)(-A_1)^{-1}. \tag{30}$$

From the structure of A_1 (see Eq. (5)), it is easy to verify that

$$(-A_1)^{-1} = \begin{bmatrix} C_{11} & 0 & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & 0 & C_{33} \end{bmatrix}, \tag{31}$$

where

$$\begin{aligned} C_{11} &= [\gamma I_{mn} - S \oplus D_0 - \xi\gamma(\xi I_{mn} - \theta S \oplus D_0)^{-1}]^{-1}, \\ C_{22} &= (\eta I_{mn} - \theta S \oplus D_0)^{-1}, \\ C_{33} &= [\xi I_{mn} - \theta S \oplus D_0 - \xi\gamma(\gamma I_{mn} - S \oplus D_0)^{-1}]^{-1}, \\ C_{13} &= \gamma(\gamma I_{mn} - S \oplus D_0)^{-1} C_{33}, \quad C_{21} = \eta(\eta I_{mn} - \theta S \oplus D_0)^{-1} C_{11}, \\ C_{23} &= \eta(\eta I_{mn} - \theta S \oplus D_0)^{-1} C_{13}, \quad C_{31} = \xi(\xi I_{mn} - \theta S \oplus D_0)^{-1} C_{11}. \end{aligned} \tag{32}$$

Note that pre-multiplying $(-A_1)^{-1}$ by a diagonal (block) matrix will not destroy the structure as seen in (31). Thus, the matrices $A_2(-A_1)^{-1}$ and $A_0(-A_1)^{-1}$ will have the same structure as $(-A_1)^{-1}$. Also, the structure is

retained by matrix powers. That is, if F has that structure, then $F^n, n \geq 1$, also retains that structure. These are the keys to the proof of this theorem.

It is well-known (see [40]) that the sequence $\{R^{(k)}, k = 0, 1, 2, \ldots\}$ defined by

$$R^{(k+1)} = [(R^{(k)})^2 A_2 + A_0](-A_1)^{-1}, \quad k = 0, 1, 2, \ldots,$$

with $R^{(0)} = 0$ monotonically converges to the minimal non-negative solution to (17).

Noting that at each iterate the structure of $R^{(k)}$ remains same as that of $-(A_1)^{-1}$, the stated result is clear. □

Suppose, we reorder the state space such that the first two sets of states (of order $2mn$) correspond to the server is either serving or under repair, and the last set of states (of order mn) corresponds to the server being on vacation. Accordingly, we rearrange the entries of A_0, A_1, A_2, and R. Note that the entries of A_0 and A_2 remain the same, while R and A_1 are to be rearranged.

Suppose, we rewrite R based on the reordering of the states and if $\tilde{R}$ denotes the rate matrix of the reordered one. That is,

$$\tilde{R} = \begin{bmatrix} R_1 & 0 \\ R_2 & R_{22} \end{bmatrix}, \text{ with } R_1 = \begin{bmatrix} R_{11} & R_{13} \\ R_{31} & R_{33} \end{bmatrix}, \text{ and } R_2 = [R_{21} \quad R_{23}]. \tag{33}$$

Similarly, if $\tilde{A}_1$ denotes the rearranged A_1, such that

$$\tilde{A}_1 = \begin{bmatrix} \tilde{A}_{11} & 0 \\ \tilde{A}_{12} & \theta S \oplus D_0 - \eta I_{mn} \end{bmatrix},$$
$$\text{with } \tilde{A}_{11} = \begin{bmatrix} S \oplus D_0 - \gamma I_{mn} & \gamma I_{mn} \\ \xi I_{mn} & \theta S \oplus D_0 - \xi I_{mn} \end{bmatrix}, \text{ and } \tilde{A}_{12} = [\eta I_{mn} \quad 0]. \tag{34}$$

In the following theorem, we will show that R can be decomposed into three components. One component corresponds to the rate matrix, say, R_1, for the queueing model in which the server is busy serving at a normal rate or the server is under repair in which case the backup server is serving at a lower rate; the second component, say, R_2, corresponds to connecting R_1 and R_{22} matrices; and finally, the third component, R_{22}, satisfies the matrix-quadratic equation of the corresponding to vacation model.

Theorem 3. *The rate matrix R has the following decomposition. The matrices R_1 and R_{22} are obtained as solutions to matrix-quadratic equations and R_2 is obtained explicitly in terms of R_1 and R_{22} as follows.*

$$\begin{aligned} R_1^2 \tilde{A}_{21} + R_1 \tilde{A}_{11} + I_{2n} \otimes D_1 &= 0, \\ \theta R_{22}^2 (\mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) + R_{22}(\theta S \oplus D_0 - \eta I_{mn}) + (I_n \otimes D_1) &= 0, \\ \tau(R_2) = -\tau(R_{22}\tilde{A}_{12})[(I_{mn} \otimes R_1 \tilde{A}_{21}) + (R'_{22} \otimes \tilde{A}_{21}) + (I_{mn} \otimes \tilde{A}_{11})]^{-1}, \end{aligned} \tag{35}$$

where $\tau(B)$ denotes the direct sum of the rows B (see, e.g. [17]).

Proof. First note that we can rewrite (17) as

$$\begin{aligned} R_1^2 \tilde{A}_{21} + R_1 \tilde{A}_{11} + I_{2n} \otimes D_1 &= 0, \\ (R_2 R_1 + R_{22} R_2)\tilde{A}_{21} + R_2 \tilde{A}_{11} + R_{22}\tilde{A}_{12} &= 0, \\ \theta R_{22}^2 (\mathbf{S}^0 \boldsymbol{\beta} \otimes I_m) + R_{22}(\theta S \oplus D_0 - \eta I_{mn}) + (I_n \otimes D_1) &= 0. \end{aligned} \tag{36}$$

With the knowledge of R_1 and R_{22}, the matrix R_2 is explicitly obtained using the Sylvester matrix equation of the form $BXC + EXF + H = 0$, where only the matrix X is unknown. The stated result follows immediately by applying the result $\tau(BXC) = \tau(X)(B' \otimes C)$. □

Note: (a) It is worth pointing out that such a decomposition result for R has been reported in the past (see e.g., [1,3]).

(b) In view of the decomposition result, the computation of R may be reduced to dealing with matrices of smaller dimension. This is accomplished by first computing the submatrices R_1 and R_{22}, and then obtaining R_2 through taking advantage of the sparsity of the coefficient matrices.

The results in the following three lemmas have nice probabilistic interpretations. Further, they help to verify the accuracy in numerical computation.

Lemma 2. *We have*

$$\mathbf{v}_0 + \mathbf{w}_0 + \left[\sum_{i=1}^{\infty} (\mathbf{u}_i + \mathbf{v}_i + \mathbf{w}_i)(\mathbf{e} \otimes I_m) \right] = \boldsymbol{\delta}, \tag{37}$$

where $\boldsymbol{\delta}$ is given in (1).

Proof. Post-multiply each one of the equations in (20)–(27) by $\mathbf{e} \otimes I_m$ and verify (after some elementary manipulations) that

$$\begin{aligned} &\mathbf{v}_0 D_0 + \mathbf{w}_0 D_0 + \mathbf{u}_1(\mathbf{S}^0 \otimes I_m) + \theta \mathbf{v}_1(\mathbf{S}^0 \otimes I_m) + \theta \mathbf{w}_1(\mathbf{S}^0 \otimes I_m) = \mathbf{0}, \\ &\mathbf{v}_0 D_1 + \mathbf{w}_0 D_1 - \mathbf{u}_1(\mathbf{S}^0 \otimes I_m) + \mathbf{u}_1(\mathbf{e} \otimes D_0) - \theta \mathbf{v}_1(\mathbf{S}^0 \otimes I_m) \\ &\quad + \mathbf{v}_1(\mathbf{e} \otimes D_0) - \theta \mathbf{w}_1(\mathbf{S}^0 \otimes I_m) \\ &\quad + \mathbf{w}_1(\mathbf{e} \otimes D_0) + \mathbf{u}_2(\mathbf{S}^0 \otimes I_m) + \theta \mathbf{v}_2(\mathbf{S}^0 \otimes I_m) + \theta \mathbf{w}_2(\mathbf{S}^0 \otimes I_m) = \mathbf{0}, \\ &\mathbf{u}_{i-1}(\mathbf{e} \otimes D_1) + \mathbf{v}_{i-1}(\mathbf{e} \otimes D_1) + \mathbf{w}_{i-1}(\mathbf{e} \otimes D_1) - \mathbf{u}_i(\mathbf{S}^0 \otimes I_m) \\ &\quad + \mathbf{u}_i(\mathbf{e} \otimes D_0) - \theta \mathbf{v}_i(\mathbf{S}^0 \otimes I_m) \\ &\quad + \mathbf{v}_i(\mathbf{e} \otimes D_0) - \theta \mathbf{w}_i(\mathbf{S}^0 \otimes I_m) + \mathbf{w}_i(\mathbf{e} \otimes D_0) + \mathbf{u}_{i+1}(\mathbf{S}^0 \otimes I_m) \\ &\quad + \theta \mathbf{v}_{i+1}(\mathbf{S}^0 \otimes I_m) + \theta \mathbf{w}_{i+1}(\mathbf{S}^0 \otimes I_m) = \mathbf{0}, \quad i \geq 2, \end{aligned} \tag{38}$$

from which we immediately obtain

$$\left[\mathbf{v}_0 + \mathbf{w}_0 + \left(\sum_{i=1}^{\infty} (\mathbf{u}_i + \mathbf{v}_i + \mathbf{w}_i)(\mathbf{e} \otimes I_m) \right) \right] (D_0 + D_1) = \mathbf{0}. \tag{39}$$

The above equation coupled with the uniqueness of $\boldsymbol{\delta}$ yields the claimed result. □

Lemma 3. *We have*

$$\sum_{i=1}^{\infty} (\mathbf{u}_i + \theta \mathbf{v}_i + \theta \mathbf{w}_i)(\mathbf{S}^0 \otimes \mathbf{e}) = \lambda, \tag{40}$$

and hence

$$\sum_{i=1}^{\infty} (\mathbf{u}_i + \theta \mathbf{v}_i + \theta \mathbf{w}_i)(I_n \otimes \mathbf{e}) = \lambda \boldsymbol{\beta}(-\mathbf{S})^{-1}. \tag{41}$$

Proof. The result (40) is intuitively obvious since in steady-state the rate at which the customers depart the system should be equal to the arrival rate. Now, post-multiply each one of the equations in (20) through (27) by $I \otimes \mathbf{e}$ and verify, after some simple manipulations, that

$$\begin{aligned} &\mathbf{v}_0 D_0 \mathbf{e} + \mathbf{w}_0 D_0 \mathbf{e} + \mathbf{u}_1(\mathbf{S}^0 \otimes \mathbf{e}) + \theta \mathbf{v}_1(\mathbf{S}^0 \otimes \mathbf{e}) + \theta \mathbf{w}_1(\mathbf{S}^0 \otimes \mathbf{e}) = \mathbf{0}, \\ &\mathbf{v}_0(\boldsymbol{\beta} \otimes D_1 \mathbf{e}) + \mathbf{w}_0(\boldsymbol{\beta} \otimes D_1 \mathbf{e}) + \mathbf{u}_1(S \otimes \mathbf{e}) + \theta \mathbf{v}_1(S \otimes \mathbf{e}) \\ &\quad + \theta \mathbf{w}_1(S \otimes \mathbf{e}) + \mathbf{u}_1(I_n \otimes D_0 \mathbf{e}) \\ &\quad + \mathbf{v}_1(I_n \otimes D_0 \mathbf{e}) + \mathbf{w}_1(I_n \otimes D_0 \mathbf{e}) + \mathbf{u}_2(\mathbf{S}^0 \boldsymbol{\beta} \otimes \mathbf{e}) + \theta \mathbf{v}_2(\mathbf{S}^0 \boldsymbol{\beta} \otimes \mathbf{e}) \\ &\quad + \theta \mathbf{w}_2(\mathbf{S}^0 \boldsymbol{\beta} \otimes \mathbf{e}) = \mathbf{0}, \\ &(\mathbf{u}_{i-1} + \mathbf{v}_{i-1} + \mathbf{w}_{i-1})(I_n \otimes D_1 \mathbf{e}) + (\mathbf{u}_i + \theta \mathbf{v}_i + \theta \mathbf{w}_i)(S \otimes \mathbf{e}) \\ &\qquad + (\mathbf{u}_i + \mathbf{v}_i + \mathbf{w}_i)(I_n \otimes D_0 \mathbf{e}) \\ &\quad + (\mathbf{u}_{i+1} + \theta \mathbf{v}_{i+1} + \theta \mathbf{w}_{i+1})(\mathbf{S}^0 \boldsymbol{\beta} \otimes \mathbf{e}) = \mathbf{0}, \quad i \geq 2. \end{aligned} \tag{42}$$

Adding the above two equations and noting that

$$(\mathbf{v}_0 + \mathbf{w}_0)(\boldsymbol{\beta} \otimes D_1 \mathbf{e}) = (\mathbf{u}_1 + \theta \mathbf{v}_1 + \theta \mathbf{w}_1)(\mathbf{S}^0 \boldsymbol{\beta} \otimes \mathbf{e}), \tag{43}$$

we get

$$\sum_{i=1}^{\infty} (\mathbf{u}_i + \theta \mathbf{v}_i + \theta \mathbf{w}_i)(\mathbf{S}^0 \boldsymbol{\beta} \otimes \mathbf{e}) = \sum_{i=1}^{\infty} (\mathbf{u}_i + \theta \mathbf{v}_i + \theta \mathbf{w}_i)(-S \otimes \mathbf{e}). \tag{44}$$

The claimed result in (41) follows by using (40). □

4. Sojourn time distribution in steady-state

Let Y denote the sojourn time of a tagged customer at an arrival epoch. Here, we will derive an expression for $P(Y > t)$ in steady-state by applying the result of [42]. Note that the customers are served according to *FIFO* discipline. Suppose that $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, ...)$ denotes the stationary probability vector of the number in the system immediately after an arrival epoch. Note that the vectors, $\mathbf{y}_i$, $i \geq 1$, are of dimension $3mn$. The following theorem determines an expression for $\mathbf{y}$.

Theorem 4. *The stationary probability vector at an arrival epoch is given by*

$$\mathbf{y}_i = \frac{1}{\lambda}\mathbf{x}_0 B_0 \hat{R}^{i-1}, \quad i \geq 1, \tag{45}$$

where $\hat{R}$ is given by

$$\hat{R} = -(A_1 + RA_2)^{-1}A_0. \tag{46}$$

Proof. It follows from (17) that

$$R = A_0[-(A_1 + RA_2)^{-1}], \tag{47}$$

which can be rewritten as $R = A_0 N$. Let $\hat{R} = NA_0$, then $\hat{R} = -(A_1 + RA_2)^{-1}A_0$.

We note that

$$\mathbf{y}_1 = \frac{1}{\lambda}(0, \mathbf{x}_0(I_2 \otimes \boldsymbol{\beta} \otimes D_1)) = \frac{1}{\lambda}\mathbf{x}_0 B_0. \tag{48}$$

Also, we have

$$\begin{aligned}\mathbf{y}_{i+1} &= \frac{1}{\lambda}\mathbf{x}_i(I_{3n} \otimes D_1) = \frac{1}{\lambda}\mathbf{x}_1 R^{i-1}(I_{3n} \otimes D_1) \\ &= \frac{1}{\lambda}\mathbf{x}_1(I_{3n} \otimes D_1)\hat{R}^{i-1}, \quad i \geq 1,\end{aligned} \tag{49}$$

which along with the normalizing condition for the steady-state probability vector at arrivals, yields the stated result. □

Suppose that $\boldsymbol{\zeta} = \frac{1}{\lambda}\mathbf{x}_0 B_0 (I_{3mn} - \hat{R})^{-1}$. It is easy to see that $\boldsymbol{\zeta}$ gives the (joint) distribution of the phases of the arrival and service processes (in steady-state) soon after an arrival occurs. Theorem 5 below gives an expression for the complementary distribution of Y.

Theorem 5. *We have, from the definition of Y, that*

$$P(Y > t) = (\mathbf{e}' \otimes \boldsymbol{\zeta}) exp([(A_0 + A_1)' \otimes I_{3mn} + A_2' \otimes \hat{R}]t)\boldsymbol{\phi}, \quad t \geq 0, \tag{50}$$

where the column vector, $\boldsymbol{\phi}$, of dimension $(3mn)^2$ is given by

$$\boldsymbol{\phi} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_{3mn} \end{pmatrix}. \tag{51}$$

Proof. Follows immediately by applying Theorem 4 and Ozawas's result (see [42]). □

Corollary. The mean, μ_Y, of Y is

$$\mu_Y = (\mathbf{e}' \otimes \boldsymbol{\zeta})(-[(A_0 + A_1)' \otimes I_{3mn} + A_2' \otimes \hat{R}])^{-1}\boldsymbol{\phi}. \tag{52}$$

Note: The classical Little's law holds good here and is used as another key accuracy check in numerical computation.

5. Distribution of the effective service

In the working-vacation-breakdown-repair queueing model under study in this paper, a customer can be served by the main server or by the backup server or by both, depending on various scenarios. Hence, it is of interest to see the distribution of the effective service time of a customer. Noting that there is no need to track the arrival process once the service begins, the following theorem establishes an explicit result for this distribution.

Theorem 6. *The distribution of the effective service time of a customer is of phase type with representation $(\boldsymbol{\alpha}, T)$ of order 3m, where*

$$\boldsymbol{\alpha} = (a_1, a_2, a_3) \otimes \boldsymbol{\beta}, \quad T = \begin{bmatrix} S - \gamma I_m & 0 & \gamma I_m \\ \eta I_m & \theta S - \eta I_m & 0 \\ \xi I_m & 0 & \theta S - \xi I_m \end{bmatrix}, \tag{53}$$

where

$$a_j = \begin{cases} \frac{1}{\lambda}\sum_{i=2}^{\infty}\mathbf{u}_i(\mathbf{S}^0 \otimes \mathbf{e}), \ j = 1, \\ \frac{1}{\lambda}[(\mathbf{u}_1 + \theta\sum_{i=1}^{\infty}\mathbf{v}_i)(\mathbf{S}^0 \otimes \mathbf{e}) + \theta\xi\mathbf{w}_1(\mathbf{S}^0 \otimes (\xi I_m - D_0)^{-1}\mathbf{e})], \ j = 2, \\ \frac{\theta}{\lambda}[\sum_{i=2}^{\infty}\mathbf{w}_i(\mathbf{S}^0 \otimes \mathbf{e}) + \mathbf{w}_1(\mathbf{S}^0 \otimes (\xi I_m - D_0)^{-1}D_1\mathbf{e})], \ j = 3. \end{cases} \tag{54}$$

Proof. First, observe that the initial probability vector, $\boldsymbol{\alpha}$, has three components, each of dimension n. The quantities, a_1, a_2, and a_3, respectively, give the probabilities that the service is initiated by the main server, by the backup server while the server is on vacation, and by the backup server while the main server is under repair. These probabilities are multiplied by $\boldsymbol{\beta}$ to initiate a service in one of n phases. Now looking at the various possibilities of a customer to begin a service the claimed result follows. □

Corollary. (a) The mean, μ_{ST}, of the effective service time of a customer is calculated as

$$\mu_{ST} = \boldsymbol{\alpha}(-T)^{-1}\mathbf{e}. \tag{55}$$

(b) The mean, μ_{MST}, time spent by a customer with the (main) server is calculated as

$$\mu_{MST} = \boldsymbol{\alpha}(-T)^{-1}(\mathbf{e_1}(3) \otimes \mathbf{e}(mn)). \tag{56}$$

(c) The mean, μ_{BST}, time spent by a customer with the (backup) server is given by

$$\mu_{BST} = \boldsymbol{\alpha}(-T)^{-1}(\mathbf{e} - \mathbf{e_1}(3) \otimes \mathbf{e}(mn)) = \mu_{ST} - \mu_{MST}. \tag{57}$$

(d) The fraction, f_{MS}, of time a customer is with the main server is given by

$$f_{MS} = \frac{\mu_{MST}}{\mu_{ST}}. \tag{58}$$

6. Performance measures of the system

Qualitative analysis of any stochastic model requires establishing key system performance measures. In this section, we list a few in addition to the ones pointed out earlier.

1. P_{idle}^{System}: This is the probability of the system having no customers and is calculated as

   $$P_{idle}^{System} = \mathrm{x}_0\mathrm{e}.$$

2. P_{NB}: This is the probability that the main server is busy serving and is calculated as

   $$P_{NB} = \sum_{i=1}^{\infty}\mathrm{u}_i\mathrm{e}.$$

3. P_{WV}: This is the probability that the backup server is busy while the (main) server is on vacation and is obtained as

   $$P_{WV} = \sum_{i=1}^{\infty}\mathrm{v}_i\mathrm{e}.$$

4. P_{WB}: This is the probability that the backup server is busy while the (main) server is under repair and is obtained as

$$P_{WB} = \sum_{i=1}^{\infty} \mathrm{w}_i \mathrm{e}.$$

5. μ_{NS}: The average number in the system is derived as

$$\mu_{NS} = \sum_{i=1}^{\infty} i(\mathrm{u}_i + \mathrm{v}_i + \mathrm{w}_i)\ \mathrm{e} = \mathrm{x}_1(I - R)^{-2}\mathrm{e}.$$

7. Special cases

The purpose of this section it to point out a number of models in the literature can be obtained as special cases of our model in this paper.

7.1. M/M/1 model

For an $M/M/1$-type working-vacation-breakdown-repair queues, we set in our model: $m = n = 1$, $D_0 = -\lambda$, $D_1 = \lambda$, $\boldsymbol{\beta} = 1$, and $S = -\mu$. This reduces R to be a 3×3 matrix and the scalar quantity R_{22} (see the third equation in (36)) is the unique solution in (0,1):

$$R_{22} = \frac{\lambda + \eta + \theta\mu - \sqrt{(\theta\mu + \lambda + \eta)^2 - 4\lambda\theta\mu}}{2\theta\mu}. \tag{59}$$

7.1.1. $\eta \to \infty$

In this limiting case, there is no vacation for the server and hence η plays no role. This case reduces to the classical $M/M/1$ queue with working breakdowns studied in [20].

7.1.2. $\xi \to \infty$

In this limiting case, there are no breakdowns and hence no repairs, and thus both ξ and γ play no role. This case reduces to the classical $M/M/1$ queue with working vacations studied in [44].

7.1.3. $\theta = 0$

This case corresponds to no backup server during a vacation or a repair. Thus, the cases in 7.1.1 and 7.1.2 reduce, respectively, to the classical $M/M/1$−type vacation-breakdown-repair studied extensively (see, e.g., [45]).

7.2. MAP/M/1

By letting $\eta \to \infty$ and $n = 1$, $\beta = 1$, $S = -\mu$, the model analyzed here reduces to the one studied in [55].

7.3. MAP/PH/1: $\eta \to \infty$, $\theta = 0$

In this case, the server never takes a vacation and also there is no backup server to offer services during the times the main server is incapable to serve. This case reduces to the classical $MAP/PH/1$−type breakdown-repair queues. It is necessary and sufficient that $\lambda < \mu$ in this case for the stability of the queue.

8. Decomposition results for M/M/1 case

Here, we will prove decomposition results for μ_{NS} in the special case of $M/M/1$−type working-vacation-breakdown-repair queues. Hereafter, $L(z)$ will stand for the probability generating function (*PGF*) of the number in the system. Then, using (16), it is easy to see that

$$L(z) = x_0 + \sum_{k=1}^{\infty} z^k \mathbf{x}_k \mathbf{e} = x_0 + z\mathbf{x}_1(I - zR)^{-1}\mathbf{e}, \quad |z| < 1. \tag{60}$$

Case 1: The M/M/1 queue with exponential (non-working) vacations

If there is no backup server during vacation mode and no breakdown or repair, then by letting $\gamma = 0$, $\theta = 0$, the elements in Q (see (2)) are:

$$B_0 = [0 \quad \lambda], \quad B_1 = [-\lambda], \quad B_2 = \begin{bmatrix} \mu \\ 0 \end{bmatrix}, \tag{61}$$

$$A_0 = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \quad A_1 = \begin{bmatrix} -\lambda - \mu & 0 \\ \eta & -\lambda - \eta \end{bmatrix}, \quad A_2 = \begin{bmatrix} \mu & 0 \\ 0 & 0 \end{bmatrix}. \tag{62}$$

When the stability condition, $\rho = \frac{\lambda}{\mu} < 1$, holds good, then R (see (17)) is given by

$$R = \begin{bmatrix} \frac{\lambda}{\mu} & 0 \\ \frac{\lambda}{\mu} & \frac{\lambda}{\lambda + \eta} \end{bmatrix}, \tag{63}$$

from which the solution to the Eqs. (18) and (19) is obtained as

$$u_1 = \rho x_0, \quad v_1 = \frac{\lambda}{\lambda + \eta} x_0, \quad x_0 = \frac{(1 - \rho)\eta}{\lambda + \eta}. \tag{64}$$

The special structure of R and $(I - zR)^{-1}$ for this case yields

$$\begin{aligned} L(z) &= \frac{(1 - \rho)\eta}{\lambda + \eta}\left[1 + \frac{\rho z}{1 - \rho z} + \frac{\rho\lambda z^2}{(1 - \rho z)(\lambda + \eta - \lambda z)} + \frac{\lambda z}{\lambda + \eta - \lambda z}\right] \\ &= \frac{1 - \rho}{1 - \rho z} \cdot \frac{\eta}{\lambda + \eta - \lambda z} \\ &= L_0(z) L_d(z), \end{aligned} \tag{65}$$

where $L_0(z)$ is the *PGF* of the stationary system size in corresponding classical $M/M/1$ queue without vacation and $L_d(z)$ is the *PGF* of the additional system size during a vacation. It follows that μ_{NS} is decomposed as the sum of the mean number in the system in the corresponding classical $M/M/1$ queue and the mean number of arrivals during a vacation. That is,

$$\mu_{NS} = \frac{\rho}{1 - \rho} + \frac{\lambda}{\eta}, \tag{66}$$

which is in agreement with that obtained in [14].

Case 2: The M/M/1 queue with exponential (non-working) vacations and breakdowns

For this case, let $\theta = 0$ and verify that the (block) matrices in (2) are given by

$$B_0 = \begin{bmatrix} 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \quad B_1 = \begin{bmatrix} -\lambda & 0 \\ \xi & -\lambda - \xi \end{bmatrix}, \quad B_2 = \begin{bmatrix} \mu & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \tag{67}$$

$$A_0 = \begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}, \quad A_1 = \begin{bmatrix} -\lambda - \mu - \gamma & 0 & \gamma \\ \eta & -\lambda - \eta & 0 \\ \xi & 0 & -\lambda - \xi \end{bmatrix},$$
$$A_2 = \begin{bmatrix} \mu & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \tag{68}$$

Using the facts that (a) $RA_2\mathbf{e} = \lambda\mathbf{e}$ and (b) the entries of R are scalars, the structure of R in (29) yields $R_{11} = R_{21} = R_{31}$. Hence, under the condition that $\lambda < \frac{\mu\xi}{\gamma + \xi}$, the entries of R are given by

$$R_{11} = R_{21} = R_{31} = \frac{\lambda}{\mu}, \ R_{13} = R_{23} = \frac{\lambda\gamma}{\mu(\lambda + \xi)}, \ R_{22} = \frac{\lambda}{\lambda + \eta},$$
$$R_{33} = \frac{\lambda(\gamma + \mu)}{\mu(\lambda + \xi)}. \tag{69}$$

Verify that the solution to the Eqs. (18) and (19) is

$$w_0 = 0, \quad u_1 = \frac{\lambda}{\mu}v_0, \quad v_1 = \frac{\lambda}{\lambda+\eta}v_0, \quad w_1 = \frac{\gamma\lambda}{\mu(\lambda+\xi)}v_0,$$
$$v_0 = \frac{\eta(\mu\xi - \lambda\xi - \lambda\gamma)}{\mu\xi(\lambda+\eta)}. \tag{70}$$

Once again, the special structure $(I - zR)^{-1}$ for this case also yields the *PGF* in a compact form as

$$L(z) = \frac{\mathbf{v}_0(1 + zR_{23} - zR_{33})}{[1 - zR_{22}][1 - zR_{11}(1 + zR_{23} - zR_{33}) - zR_{33}]}. \tag{71}$$

Observe that μ_{NS} is calculated as

$$\mu_{NS} = \frac{dL(z)}{dz}|_{z=1} = l_1 + l_2 + l_3 \cdot l_4, \tag{72}$$

where

$$l_1 = \frac{\lambda}{\eta}, \quad l_2 = \frac{\rho}{1-\rho}, \quad l_3 = \frac{\lambda}{\xi}, \quad l_4 = \frac{\rho - (\lambda/\mu)}{1-\rho}, \quad \rho = \frac{\lambda(\xi+\gamma)}{\mu\xi}. \tag{73}$$

It is worth pointing out that μ_{NS} is now decomposed as the sum of (i) the mean number (l_2) in the system of the corresponding classical $M/M/1$ queue; (ii) the mean number (l_1) in the system due to a vacation; and (iii) an additional quantity ($l_3 \cdot l_4$) obtained as the product of the mean number (l_3) arriving during a repair and the mean number (l_4) that arrived during a non-service time. This decomposition is in agreement with the one given in [36] for the single arrival case.

Case 3: The M/M/1 queue with multiple working vacations

For this case, we set $\gamma = 0$, $\xi \to \infty$. Verify the elements in (2) are obtained as

$$B_0 = [0 \quad \lambda], \quad B_1 = [-\lambda], \quad B_2 = \begin{bmatrix} \mu \\ \theta\mu \end{bmatrix}, \tag{74}$$

$$A_0 = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \quad A_1 = \begin{bmatrix} -\lambda-\mu & 0 \\ \eta & -\lambda-\theta\mu-\eta \end{bmatrix}, \quad A_2 = \begin{bmatrix} \mu & 0 \\ 0 & \theta\mu \end{bmatrix}. \tag{75}$$

If $\rho = \frac{\lambda}{\mu} < 1$, then from Theorem 2, R (see (17)) is of the form

$$R = \begin{bmatrix} r_1 & 0 \\ r_2 & r_3 \end{bmatrix}, \tag{76}$$

where $r_1 = \rho$, $r_2 = \rho - \theta r$, $r_3 = r$, and r, the unique root in (0,1) of the equation

$$\theta\mu r^2 - (\lambda + \theta\mu + \eta)r + \lambda = 0, \tag{77}$$

is obtained as

$$r = \frac{1}{2\theta\mu}\left(\lambda + \theta\mu + \eta - \sqrt{(\lambda+\theta\mu+\eta)^2 - 4\lambda\theta\mu}\right). \tag{78}$$

The solution to the equations (18) and (19), is given by

$$u_1 = r_2 x_0, \quad v_1 = r_3 x_0, \quad x_0 = \frac{(1-r_1)(1-r_3)}{1-r_1+r_2}. \tag{79}$$

The mean, μ_{NS}, is given by

$$\mu_{NS} = \mathbf{x}_1(I-R)^{-2}\mathbf{e} = \frac{r_1}{1-r_1} + \frac{r_2+r_3-r_1}{(1-r_3)(1-r_1+r_2)}, \tag{80}$$

which can be further simplified as

$$\mu_{NS} = \frac{\rho}{1-\rho} + \frac{r_3(1-\theta)}{(1-r_3)(1-\theta r_3)}. \tag{81}$$

The above shows that the decomposition is obtained as the sum of the mean number in the system of the corresponding classical $M/M/1$ queue and the mean number in the system due to the server on vacation. This again coincides with the results of [32].

Case 4: The M/M/1 queue with (non-working) breakdowns

By letting $\theta = 0$ and $\eta \to \infty$, the (block) matrices in (2) are given by

$$B_0 = [\lambda \quad 0], \quad B_1 = [-\lambda], \quad B_2 = \begin{bmatrix} \mu \\ 0 \end{bmatrix}, \tag{82}$$

$$A_0 = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}, \quad A_1 = \begin{bmatrix} -\lambda-\mu-\gamma & \gamma \\ \xi & -\lambda-\xi \end{bmatrix}, \quad A_2 = \begin{bmatrix} \mu & 0 \\ 0 & 0 \end{bmatrix}. \tag{83}$$

Assuming the stability condition, $\lambda < \frac{\mu\xi}{\xi+\gamma}$, holds good, we have with the fact that $RA_2\mathbf{e} = \lambda\mathbf{e}$, and with Theorem 2, R is of the form

$$R = \begin{bmatrix} r_1 & r_2 \\ r_1 & r_3 \end{bmatrix}. \tag{84}$$

Verify that the solution (see (17)) yields the entries of R to be

$$r_1 = \frac{\lambda}{\mu}, \; r_2 = \frac{\lambda\gamma}{(\lambda+\xi)\mu}, \; r_3 = \frac{\lambda(\gamma+\mu)}{\mu(\lambda+\xi)}. \tag{85}$$

The solution to the Eqs. (18) and (19), is then given by

$$u_1 = r_1 x_0, \quad v_1 = r_2 x_0, \quad x_0 = \frac{d}{1+r_2-r_3}, \tag{86}$$

where $d = (1 - r_1 - r_3 - r_1(r_2 - r_3)) = \frac{\mu\xi - \lambda\xi - \lambda\gamma}{\mu(\lambda+\xi)}$.

It is easy to verify for this case, the *PGF* is given by

$$L(z) = \frac{d}{1+r_2-r_3} \cdot \frac{1+zr_2-zr_3}{1-zr_1(1+zr_2-zr_3)-zr_3}. \tag{87}$$

Thus, we have

$$\mu_{NS} = \frac{dL(z)}{dz}|_{z=1} = \frac{\rho}{1-\rho} + \frac{\lambda}{\xi}\frac{\rho-(\lambda/\mu)}{1-\rho}, \tag{88}$$

where $\rho = \frac{\lambda(\xi+\gamma)}{\mu\xi}$. Note that this decomposition result agrees with the one in [20].

9. Illustrative numerical examples

In this section, to understand the qualitative aspects of the working-vacation-breakdown-repair queueing model under study, we illustrate a few numerical examples. We analyze different scenarios by varying the parameters of the model including the arrival processes and service time distributions.

For the arrival process, we consider the following five sets of values for D_0 and D_1, which are considered as input data in many published works in the literature (see, e.g., [3,5–7]). For sake of completeness, we display them here.

ERLA: This is Erlang distribution of order 2.

$$D_0 = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}, \; D_1 = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}.$$

EXPA: This is the classical exponential distribution.

$$D_0 = (-1), \; D_1 = (1).$$

HEXA: This is the mixture of two exponentials.

$$D_0 = \begin{pmatrix} -1.90 & 0 \\ 0 & -0.19 \end{pmatrix}, \; D_1 = \begin{pmatrix} 1.71 & 0.19 \\ 0.171 & 0.019 \end{pmatrix}.$$

MNA: This one has a negative correlation (with a value of -0.4889) between two successive inter-arrival times.

$$D_0 = \begin{pmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{pmatrix}, \; D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.01002 & 0 & 0.9922 \\ 223.4925 & 0 & 2.2575 \end{pmatrix}.$$

MPA: This one has a positive correlation (with a value of 0.4889) between two successive inter-arrival times.

$$D_0 = \begin{pmatrix} -1.00222 & 1.00222 & 0 \\ 0 & -1.00222 & 0 \\ 0 & 0 & -225.75 \end{pmatrix}, D_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0.9922 & 0 & 0.01002 \\ 2.2575 & 0 & 223.4925 \end{pmatrix}.$$

Note that *ERLA, EXPA*, and *HEXA*, correspond to renewal processes. All five arrival processes are qualitatively different. Verify that the ratio of the standard deviations of the times between two successive inter-arrival times of these five processes with respect to the Erlang one are, respectively, 1, 1.4142, 3.1745, 1.9933, and 1.9933.

For the service times with representation $(\boldsymbol{\beta}, S)$, we consider the following three $PH-$distributions. Normalization of these representations will be done to get a desired μ.

ERLS: This is Erlang of order 2.

$$\boldsymbol{\beta} = (1, 0), \quad S = \begin{pmatrix} -2 & 2 \\ 0 & -2 \end{pmatrix}.$$

EXPS: This is an exponential one.

$$\boldsymbol{\beta} = 1, \quad S = (-1).$$

HEXS: This is the mixture of two exponentials.

$$\boldsymbol{\beta} = (0.9, 0.1), \quad S = \begin{pmatrix} -10 & 0 \\ 0 & -1 \end{pmatrix}.$$

Irrespective of how these are normalized, the standard deviation of *EXPS* and *HEXS* are, respectively, about 1.4142 and 3.1745 of that of *ERLS*.

Next, we define

$$\rho = \frac{\lambda(\xi + \gamma)}{\mu(\xi + \theta\gamma)}. \tag{89}$$

Example 1. Here, we fix $\lambda = 1$, $\mu = 1.149$ and $\theta = 0.8$, and investigate the behavior of μ_Y as a function of ξ, γ and η under different scenarios. Table 1 lists the values of μ_Y. Some key observations are in order.

- As expected, μ_Y, is a non-increasing function of η (for fixed ξ and γ), ξ (for fixed γ and η) and γ (for fixed ξ and η). This is the case for the types of arrival and service processes considered here. This is due to the fact that an increase in ξ or η, decreases the duration of the slow service (backup server) mode.
- We noticed that μ_Y is significantly large for positively correlated arrivals compared to other arrival processes. Noticing that *HEXA* has the largest variability in the inter-arrival times, for *MPA* to yield the largest μ_Y indicates the (significant) role played by positive correlation in the inter-arrival times.
- For all scenarios involving renewal arrivals, we see that a higher variation in the successive inter-arrival times yields a higher value for μ_Y. Moreover, the rate of increase is more significant for *HEXA*.
- We also observe that *HEXS* produce a larger value for μ_Y as compared to *ERLS*. This is the case for the five arrival processes.

Example 2. The effect of the vacation rate, η, on two performance measures, f_{MS} and $P_{BB} = P_{WB} + P_{WV}$ is of focus here. Recall that f_{MS} gives the fraction of time a customer is served by the main server, and P_{BB} is the probability that the backup server is busy serving. We fix $\lambda = 1$, $\mu = 1.372$, $\theta = 0.8$, $\xi = 0.5$, $\gamma = 0.4$ and vary η. In Fig. 2, we display the two selected measures under various scenarios. Analyzing the graphs in the figure, we notice the following.

- An increase in η causes an increase in f_{MS}, which is to be expected. However, the rate of increase is higher for small values of η. The type of arrival processes and the service times play a role in the rate of increase.
- Observe that among the renewal arrivals, the *HEXA* process produces the highest value for f_{MS}. Further, the difference (as compared to *ERLA* and *EXPA* cases) in f_{MS} widens as η is decreased. Among the correlated arrivals, f_{MS} as a function of η is large initially for *MPA* (as compared to *MNA*) and then as η takes a higher value, the roles are reversed; that is, *MNA* has a larger value for f_{MS} compared to that of *MPA*. This points to an interesting trend in f_{MS} among correlated arrivals.
- Also, we observe that *HEXS* yield a larger value for f_{MS} as compared to *ERLS*.
- The measure, P_{BB}, appears to decrease as a function of η. This is to be expected as an increase in η will decrease the duration of the vacation mode; thus, resulting in a smaller value for this probability.
- A decrease in the variation of the inter-arrival time (or service time) causes P_{BB} to increase. Thus, a decrease in the variation necessitates more usage of the backup server for all values of η.

Example 3. The focus here is to see how the tail probabilities of the sojourn time, $P(Y > t)$, behave under various scenarios. We fix $\lambda = 0.75$, $\theta = 0.5$, $\gamma = 1$, $\xi = 1$, $\eta = 2$, and consider two values for ρ, namely $\rho = 0.8$ and 0.99. Some key observations based on Fig. 3 are given below.

Table 1
The mean sojourn time.

ξ	γ	η	Erlang services (*ERLS*)					Hyperexponential services (*HEXS*)				
			ERLA	*EXPA*	*HEXA*	*MNA*	*MPA*	*ERLA*	*EXPA*	*HEXA*	*MNA*	*MPA*
0.5	0.1	0.5	5.17	7.33	23.61	7.58	448.80	23.52	25.71	42.46	25.89	467.14
		1	5.03	7.19	23.45	7.43	448.64	23.38	25.57	42.30	25.74	466.98
		1.5	4.98	7.14	23.40	7.38	448.59	23.33	25.52	42.25	25.69	466.92
		2	4.96	7.11	23.37	7.36	448.56	23.30	25.49	42.22	25.67	466.90
		3	4.94	7.09	23.34	7.33	448.54	23.28	25.47	42.20	25.64	466.87
1	0.2	0.5	5.15	7.31	23.59	7.56	448.78	23.50	25.69	42.43	25.86	467.11
		1	5.00	7.16	23.42	7.41	448.62	23.35	25.54	42.27	25.72	466.95
		1.5	4.96	7.11	23.37	7.36	448.56	23.30	25.49	42.22	25.66	466.90
		2	4.93	7.08	23.34	7.33	448.54	23.27	25.46	42.19	25.64	466.87
		3	4.90	7.06	23.31	7.30	448.51	23.25	25.44	42.17	25.61	466.84
1.5	0.3	0.5	5.14	7.30	23.58	7.55	448.77	23.49	25.68	42.42	25.86	467.11
		1	5.00	7.15	23.42	7.40	448.61	23.34	25.53	42.27	25.71	466.94
		1.5	4.95	7.10	23.36	7.35	448.56	23.29	25.48	42.21	25.66	466.89
		2	4.92	7.07	23.33	7.32	448.53	23.27	25.45	42.19	25.63	466.86
		3	4.90	7.05	23.31	7.29	448.50	23.24	25.43	42.16	25.60	466.84
2	0.4	0.5	5.14	7.30	23.57	7.55	448.77	23.49	25.68	42.42	25.85	467.10
		1	4.99	7.15	23.41	7.40	448.61	23.34	25.53	42.26	25.70	466.94
		1.5	4.94	7.10	23.36	7.34	448.55	23.29	25.48	42.21	25.65	466.89
		2	4.92	7.07	23.33	7.32	448.52	23.26	25.45	42.18	25.63	466.86
		3	4.89	7.04	23.30	7.29	448.50	23.24	25.42	42.15	25.60	466.83

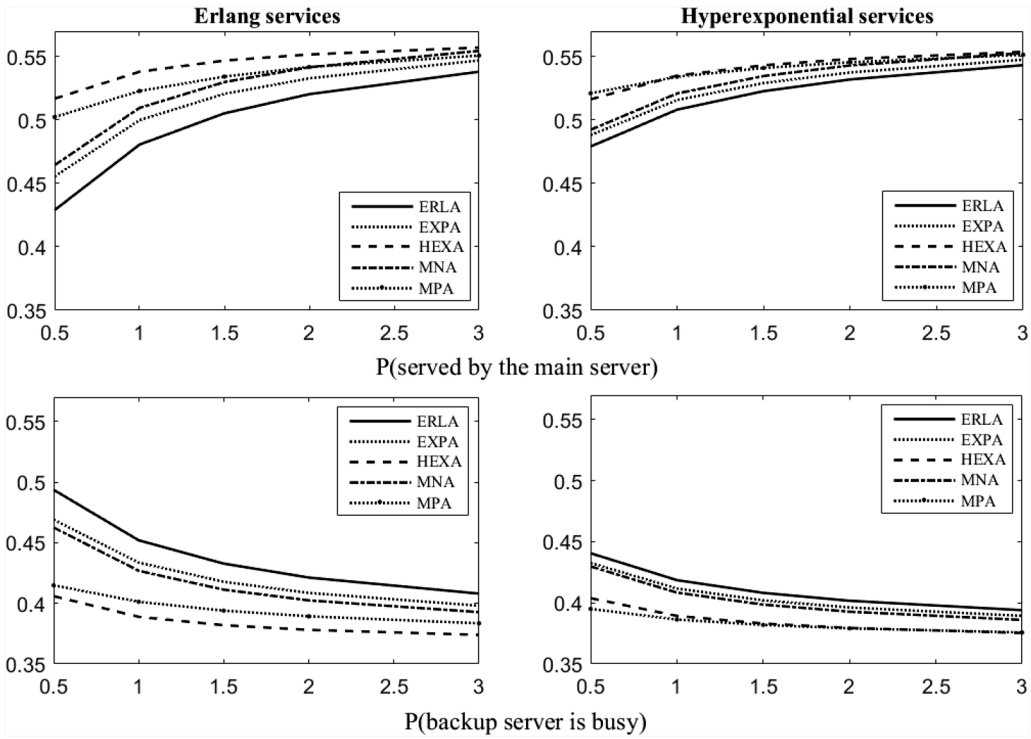

Fig. 2. Selected performance measures for *ERLS* and *HEXS* under different scenarios.

- As t increases, $P(Y > t)$ decreases as it should be. However, the rate of decrease is less significant for *HEXA* and *MPA* as compared to the other arrival processes. For small t, $P(Y > t)$ appears to be insignificant among different scenarios; however, when t is relatively large, we notice that $P(Y > t)$ significantly depends on the type of arrival processes while decreasing at a faster rate.

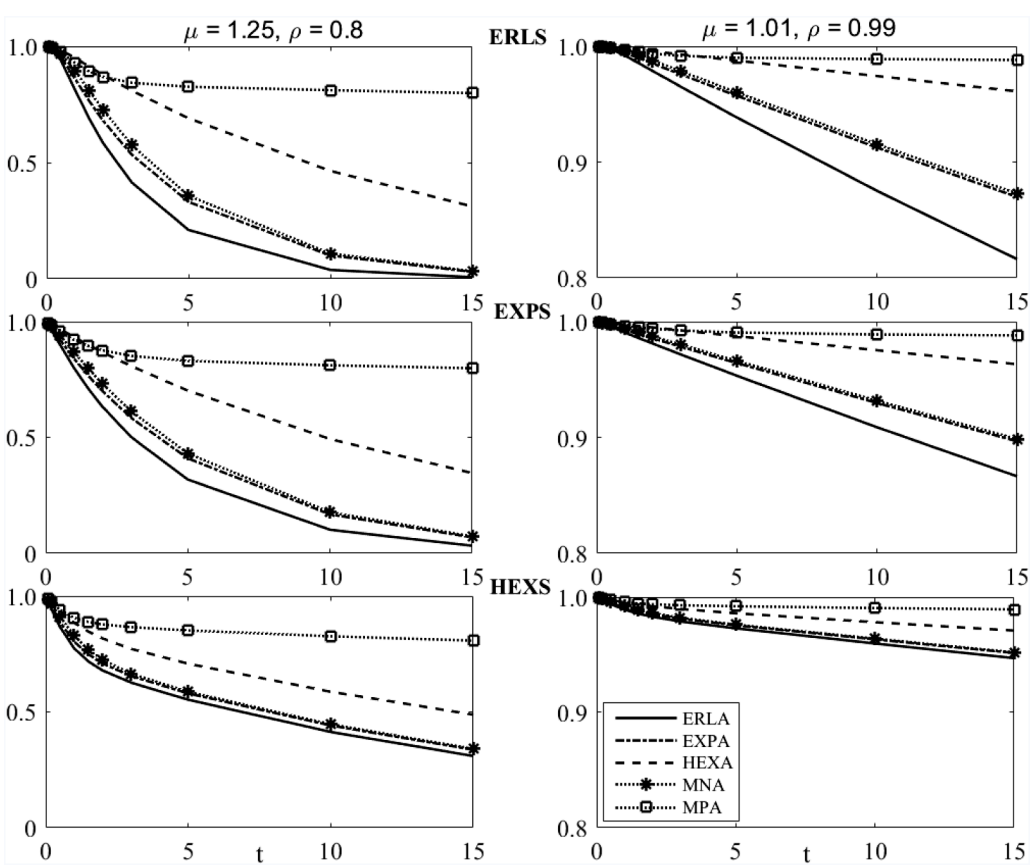

Fig. 3. Tail probabilities of the sojourn time for *MAP/PH/*1.

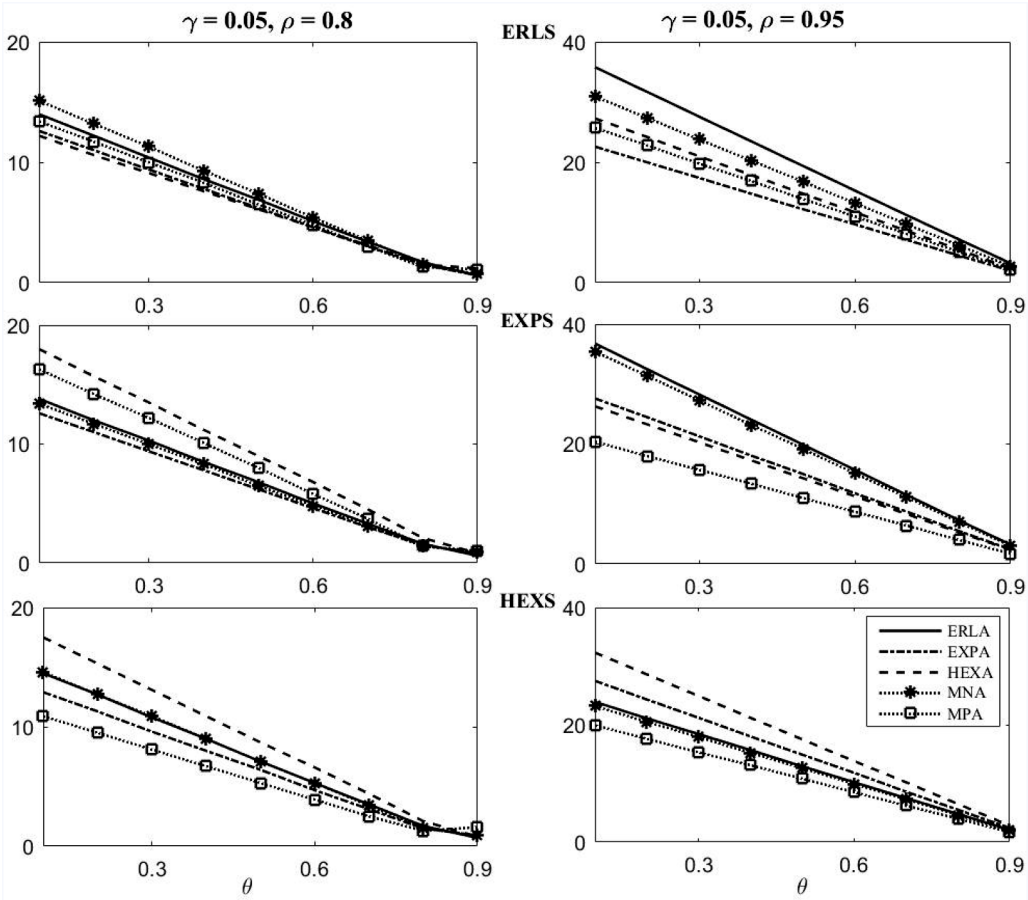

Fig. 4. Optimal values of the repair rate (ξ) under various scenarios.

- An increase in ρ leads to a significant increase $P(Y > t)$ (especially when t is relatively large). Among the renewal arrivals it appears that $P(Y > t)$ is largest for *HEXA*. Further, $P(Y > t)$ is significantly large for *MPA* in contrast to all other arrivals. This once again confirms the role of correlation in queueing.
- We also observed an interesting pattern. For all the scenarios, a higher variation in the service times produces a smaller value for $P(Y > t)$ initially and then as t becomes reasonably large, the roles are reversed. Moreover, the degree of sensitivity to the variation in the service times is higher for *ERLA* in comparison to the other arrival processes. This holds for both the values of ρ considered.

Example 4. The focus of this example is to look for a minimum value, say ξ^*, of the repair rate, ξ, (when all other parameters are fixed) such that the average numbers in the queue for the model under study (without vacation) and for the classical *MAP/PH/*1 queue do not differ by more than 10^{-3}. Here, we fix $\lambda = 1.0$. Considering two values of $\rho = 0.8, 0.95$, two values for $\gamma = 0.05, 1.0$ and varying θ, we display ξ^* in Figs. 4 and 5 under different scenarios. Note that μ will vary as we vary ρ, ξ, and θ (see (89)).

Examining the figures reveals the following:

- As θ is increased, ξ^* appears to decrease under all scenarios. Intuitively this happens because of the fact that a higher service rate (of the backup server) requires a lower repair rate to achieve the desired mean number of customers in the queue.
- As ρ increases, we notice a significant increase in the values of ξ^* for all cases considered.
- Among the renewal arrivals, we notice a different pattern. When $\gamma = 0.05$, $\rho = 0.8$, for *ERLS, ERLA* produces the largest value whereas for *EXPS* and *HEXS, HEXA* produces the largest value. However, for $\gamma = 0.05, 1.0$ with $\rho = 0.95$, we observe that for *ERLS* and *EXPS, ERLA* produces the largest value whereas for *HEXS*, again *HEXA* produces the largest value.
- It appears that for *MNA* arrivals, a larger ξ^* is required (under all scenarios considered except for *EXPS* with $\rho = 0.8$) as compared to *MPA*.
- The sensitivity to the type of arrivals and services in the values of ξ^*, appears to decrease when θ is increased. This seems to be the case for both values of γ when $\rho = 0.95$.
- Another interesting observation is that for fixed values of θ and ρ, ξ^* increases with an increase in the value of γ for all the arrival and service processes considered.

10. Concluding remarks

We studied a queueing model of the type *MAP/PH/*1 in the context of the server going on vacations, the server breaking down due to external shocks, the server repairs, and the server being helped by a backup server during vacations and repairs. A qualitative study of the model in steady-state through a number of system performance measures including the tail probabilities of the sojourn time for various scenarios is performed. The model can be generalized to multi-server case, but the dimension of the problem will increase significantly. One can also relax exponential assumptions for the breakdown, repair, and vacation times by assuming them to be of phase type. Once again, the dimension of the problem will considerably increase. These extensions can be explored in the future.

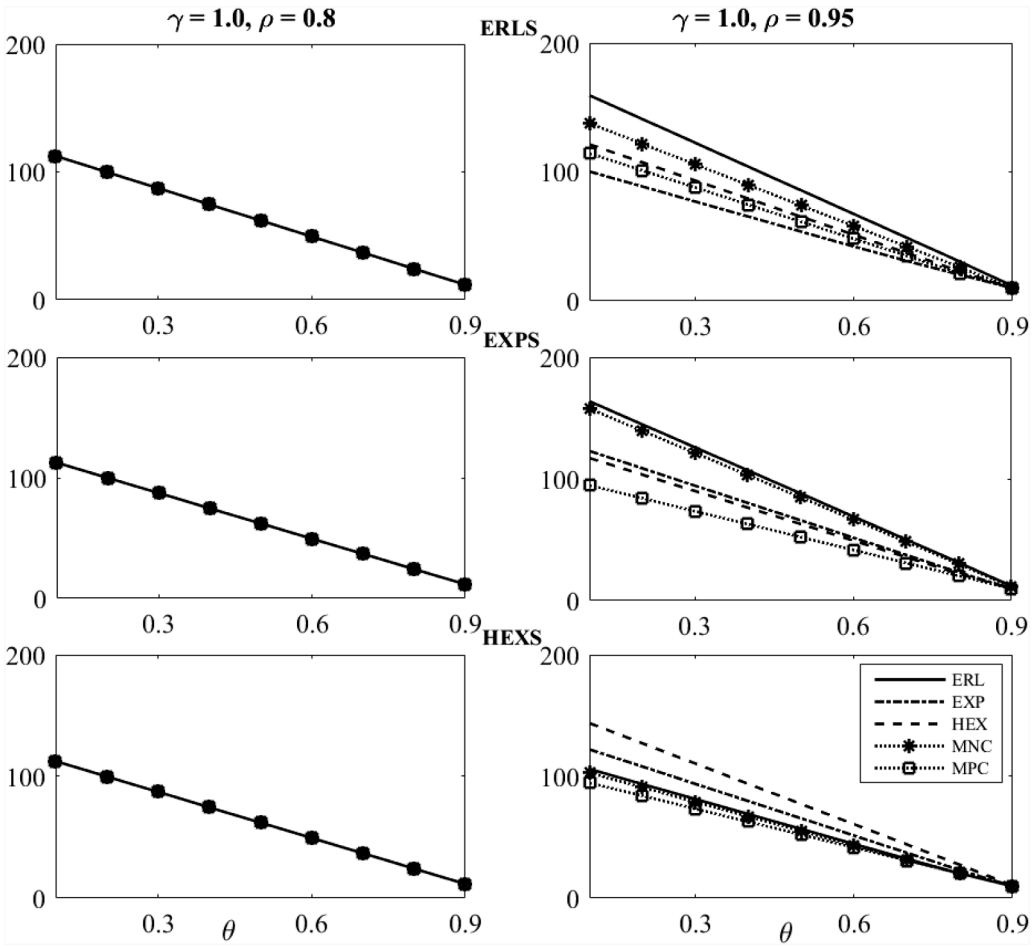

Fig. 5. Optimal values of the repair rate (ξ) under various scenarios.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors are thankful to the referees for their constructive suggestions which helped to improve the presentation of this paper.

References

[1] Alfa A. Some decomposition results for a class of vacation queues. Operat. Res. Lett. 2014;42:140–4.
[2] Artalejo JR, Gómez-Corral A. Markovian arrivals in stochastic modelling: a survey and some new results (invited article with discussion: rafael pérez-ocón, miklos telek and yiqiang q. zhao). SORT-Stat. Operat. Res. Trans. 2010;34:101–56.
[3] Chakravarthy S, Krishnamoorthy A, Joshua V. Analysis of a multi-server retrial queue with search of customers from the orbit. Perform. Evaluat. 2006;63:776–98.
[4] Chakravarthy SR. The batch Markovian arrival process: A review and future work. In: Krishnamoorthy A, editor. Advances in probability Theory and Stochastic Processes. NJ: Notable Publications Inc.; 2001. p. 21–39.
[5] Chakravarthy SR. A multi-server synchronous vacation model with thresholds and a probabilistic decision rule. Eur. J. Operat. Res. 2007;182:305–20.
[6] Chakravarthy SR. Analysis of a multi-server queue with Markovian arrivals and synchronous phase type vacations. Asia-Pacific J. Operat. Res. 2009;26:85–113.
[7] Chakravarthy SR. Markovian arrival process. Wiley Encycloped. Operat. Res. Manag. Sci. 2010.
[8] Chakravarthy SR. Analysis of a multi-server queueing model with vacations and optional secondary services. Math. Appl. 2013;41:127–51.
[9] Chakravarthy SR. Analysis of MAP/PH_1, $PH_2/1$ queue with vacations and optional secondary services. Appl. Math. Model. 2013;37:8886–902.
[10] Chakravarthy SR, Ozkar S. $MAP/PH/1$ Queueing model with working vacation and crowdsourcing. Math. Appl. 2016;44:263–94.
[11] Chandrasekaran V, Indhira K, Saravanarajan M, Rajadurai P. A survey on working vacation queueing models. Int. J. Pure Appl. Math. 2016;106:33–41.
[12] Deepa B, Kalidass K. An $M/M/1/N$ queue with working breakdowns and vacations. Int. J. Pure Appl. Math. 2018;119:859–73.
[13] Doshi BT. Queueing systems with vacations: a survey. Queue. Syst. 1986;1:29–66.
[14] Fuhrmann SW, Cooper RB. Stochastic decompositions in the $M/G/1$ queue with generalized vacations. Operat. Res. 1985;33:1117–29.
[15] Goswami C, Selvaraju N. A working vacation queue with priority customers and vacation interruptions. Int. J. Operat. Res. 2013;17:311–32.
[16] Goswami C, Selvaraju N. Phase-type arrivals and impatient customers in multi-server queue with multiple working vacations. Adv. Operat. Res. 2016;2016:1–17.
[17] Horn RA, Johnson CR. Topics in Matrix Analysis. Cambridge, UK: Cambridge University Press; 1991.
[18] Jacob V, Chakravarthy SR, Krishnamoorthy A. On a customer-induced interruption in a service system. Stoch. Anal. Appl. 2012;30:949–62.
[19] Kalidas K, Pavithra K. An $M/M/1/N$ queue with working breakdowns and bernoulli feedbacks. Int. J. Appl. Math. 2016;6:113–8.
[20] Kalidass K, Kasturi R. A queue with working breakdowns. Comput. Ind. Eng. 2012;63:779–83.
[21] Ke JC. Bi-level control for batch arrival queues with an early startup and unreliable server. Appl. Math. Model. 2004;28:469–85.
[22] Ke JC. Modified T vacation policy for an $M/M/1/N$ queueing system with an unreliable server and startup. Math. Comput. Model. 2005;41:1267–77.
[23] Ke JC. On $M/G/1$ system under NT policies with breakdowns, startup and close-down. Appl. Math. Model. 2006;30:49–66.
[24] Ke J-C, Lin C-H, Yang J-Y, Zhang ZG. Optimal (d, c) vacation policy for a finite buffer $M/M/C$ queue with unreliable servers and repairs. Appl. Math. Model. 2009;33:3949–62.
[25] Kempa WM. Transient workload distribution in the $M/G/1$ finite-buffer queue with single and multiple vacations. Ann. Operat. Res. 2016;239:381–400.
[26] Kempa WM, Kobielnik M. Transient solution for the queue-size distribution in a finite-buffer model with general independent input stream and single working vacation policy. Appl. Math. Model. 2018;59:614–28.
[27] Kim BK, Lee DH. The $M/G/1$ queue with disasters and working breakdowns. Appl. Math. Model. 2014;38:1788–98.
[28] Kim C, Klimenok VI, Dudin AN. Analysis of unreliable $BMAP/PH/n$ type queue with markovian flow of breakdowns. Appl. Math. Comput. 2017;314:154–72.

[29] Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. SIAM 1999.
[30] Li J, Tian N. Performance analysis of a $GI/M/1$ queue with single working vacation. Appl. Math. Comput. 2011;217:4960–71.
[31] Li L, Wang J, Zhang F. Equilibrium customer strategies in Markovian queues with partial breakdowns. Comput. Ind. Eng. 2013;66:751–7.
[32] Liu W-Y, Xu X-L, Tian NS. Stochastic decompositions in the $M/M/1$ queue with working vacations. Operat. Res. Lett. 2007;35:595–600.
[33] Lucantoni DM. New results on the single server queue with a batch Markovian arrival process. Commun. Stat. Stochast. Models 1991;7:1–46.
[34] Lucantoni DM, Meier-Hellstern KS, Neuts MF. A single-server queue with server vacations and a class of non-renewal arrival processes. Adv. Appl. Probab. 1990;22:676–705.
[35] Marcus M, Minc H. A survey of matrix theory and matrix inequalities. Courier Corporat. 1992.
[36] Mary KJR, Remona JM, Rajalakshmi R. Analysis of $M^x/M/1/MWV/BD$ queuing systems. Int. J. Comput. Appl. 2016;141:1–4.
[37] Neuts MF. A versatile Markovian point process. J. Appl. Probab. 1979;16:764–79.
[38] Neuts MF. Structured Stochastic Matrices of $M/G/1$ Type and Their Applications. New York: Marcel Dekker; 1989.
[39] Neuts MF. Models based on the Markovian arrival process. IEICE Trans. Commun. 1992;75:1255–65.
[40] Neuts MF. Matrix-geometric solutions in stochastic models: an algorithmic approach. Courier Corporat. 1994.
[41] Niu Z. TANGO: Traffic-aware network planning and green operation. IEEE Wireless Commun. 2011;18:25–9.
[42] Ozawa T. Sojourn time distributions in the queue defined by a general QBD process. Queueing Syst. 2006;53:203–11.
[43] Ozkar S, Kocer UU. $M/C_k/1$ Queue model with multiple working vacations. Int. J. Appl. Comput. Math. 2017;3:2729–44.
[44] Servi LD, Finn SG. $M/M/1$ Queues with working vacations $M/M/1/WV$. Perf. Evaluat. 2002;50:41–52.
[45] Shoukry E, Salwa MA, Boshra AS. Matrix geometric method for $M/M/1$ queueing model with and without breakdown ATM machines. Am. J. Eng. Res. (AJER) 2018;7:246–52.
[46] Son K, Kim H, Yi Y, Krishnamachari B. Base station operation and user association mechanisms for energy-delay tradeoffs in green cellular networks. IEEE J. Selected Areas Commun. 2011;29:1525–36.
[47] Sreenivasan C, Chakravarthy SR, Krishnamoorthy A. $M/PH/1$ Queue with working vacations, vacation interruptions and N policy. Appl. Math. Model. 2013;37:3879–93.
[48] Sridharan V, Jayashree P. Some characteristics on a finite queue with normal partial and total failures. Microelectron. Reliab. 1996;36:265–7.
[49] Steeb W-H, Hardy Y. Matrix Calculus and Kronecker Product: A Practical Approach to Linear and Multilinear Algebra Second Edition. World Scientific Publishing Company; 2011.
[50] Tian NS, Zhang ZG. Vacation Queueing Models: Theory and Applications. New York: Springer; 2006.
[51] Wu J, Wu Y, Zhou S, Niu Z. Traffic-aware power adaptation and base station sleep control for energy-delay tradeoffs in green cellular networks. Proceedings of the Global Communications Conference (GLOBECOM). IEEE; 2012. 3171–3176
[52] Wu J, Zhou S, Niu Z. Traffic-aware base station sleeping control and power matching for energy-delay tradeoffs in green cellular networks. IEEE Trans. Wireless Commun. 2013;12:4196–209.
[53] Yang D-Y, Wu CH. Cost-minimization analysis of a working vacation queue with N-policy and server breakdowns. Comput. Ind. Eng. 2015;82:151–8.
[54] Yang D-Y, Wu YY. Analysis of a finite-capacity system with working breakdowns and retention of impatient customers. J. Manufactur. Syst. 2017;44:207–16.
[55] Ye Q, Liu L. Analysis of $MAP/M/1$ queue with working breakdowns. Commun. Stat. Theory Methods 2018;47:3073–84.
[56] Zhang M, Hou Z. Performance analysis of $MAP/G/1$ queue with working vacations and vacation interruption. Appl. Math. Model. 2011;35:155–1560.