



Analysis of a renewal batch arrival queue with a fault-tolerant server using shift operator method

Miaomiao Yu^{1,2} · Yinghui Tang¹

Received: 14 April 2020 / Revised: 2 March 2021 / Accepted: 16 March 2021 / Published online: 2 April 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

Motivated by the fault tolerance for manufacturing, we investigate a renewal input bulk arrival queue with a fault-tolerant server, in which the server can keep working with a low service rate even if the partial failure occurs. Only when there are no customers in the system, the partial failure can be removed. To explore the performance measures of the queue, a more generic and simpler algorithm based on the right shift operator method for solving difference equations is employed to obtain the queue-length distributions at different time epochs. The significant feature of this algorithm lies in that it does not require the derivation of the transition probability matrix for the corresponding embedded Markov chain. Furthermore, we can resort to the queue-length distribution at the pre-arrival epoch to quickly get the expected sojourn time for an arbitrary customer. Finally, with the help of Padé approximation, several representative numerical examples are illustrated in tables and graphs, under which we show how to verify the correctness of our theoretical results through Little's law.

Keywords Fault tolerance · Batch arrival queue · Renewal input · Shift operator · Difference equation · Roots

Mathematics Subject Classification 60K25 · 68M20 · 90B22

✉ Miaomiao Yu
mmyu75@163.com

Yinghui Tang
tangyh@sicnu.edu.cn

¹ School of Mathematical Science, Sichuan Normal University, Chengdu 610066, Sichuan, China

² School of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong 643000, Sichuan, China

1 Introduction

A large number of papers in the field of queueing theory have appeared to analyze the waiting line system with an unreliable server. Among some earlier works in this area, Gaver (1962) studied the waiting line with interrupted service, including priorities. Avi-Itzhak and Naor (1963) and Thiruvengadam (1963) investigated some fundamental queueing problems with service facilities subject to breakdowns. Mitraný and Avi-Itzhak (1968) extended the analysis to the multi-server queue with service interruptions. Later, this model was revisited by Neuts and Lucantoni (1979) under the assumption that failed servers are repaired by one of the c repairmen. Over the past thirty years, authors like Sengupta (1990), Takine and Sengupta (1997), Tang (1997), Li et al. (1997), Madan (2003), Ke (2003, 2006), Wang (2004), Gray et al. (2004), Wang et al. (2010), Choudhury and Deka (2012), Jain et al. (2013), Kumar et al. (2020), Gorbunova and Lebedev (2020), and Kumar et al. (2021) have studied some single arrival queueing systems with an unreliable server. A common assumption in the above literature is that as soon as the server fails, it instantaneously undergoes repairs. However, in real-life situations, due to some unavoidable reasons, the repair of the failed server may be delayed. This phenomenon stimulates Choudhury and Tadj (2009), Choudhury and Kalita (2018) investigated the steady-state behavior of the unreliable M/G/1 queue with delay repair. On the other hand, from the mathematical and practical points of view, the case of batch arrival is more general, and also more challenging to handle. Many authors have contributed to the theory of batch arrival queues subject to unpredictable server breakdowns. Some notable works in this direction can be found in Ke and Lin (2006), Ke and Huang (2010, 2012), Choudhury and Tadj (2011), Singh et al. (2018), Choudhury and Deka (2018), Saggou et al. (2019), Jain and Kaur (2020). To get a more comprehensive summary of the recent research work on the topic of the unreliable queue, interested readers may refer to the review paper done by Jain et al. (2019a, 2019b).

It is worth noting that in the queueing systems mentioned above, the status of a server is usually modeled with two extreme states, working and failed. A working server is capable of serving customers at any instant of time. On the contrary, when the server fails, it is totally incapable of servicing and requires renewal or repair. However, such an assumption is not always true in reality. For example, a service facility or a manufacturing system built with fault tolerance capabilities will manage to keep operating (perhaps at a degraded level) in the presence of partial failure. Here, fault tolerance refers to the ability of a system to continue working without interruption when a partial failure occurs, wherein partial failures are often caused by the wear and tear of equipment components or the wrong action by an operator. The objective of creating a fault-tolerant system is to prevent disruptions arising in the service or manufacturing process, ensuring the high availability and business continuity of a stochastic service system. To theoretically analyze such phenomenon, Kalidass and Kasturi (2012) developed a new concept of working breakdowns in queueing theory. In a working breakdown queue, the server works at a lower service rate rather than completely stopping

service during the breakdown period. Many researchers motivated by their work have extended this type of queue in different frameworks. Kim and Lee (2014) further studied an M/G/1 queue with disasters and working breakdowns, in which the system is equipped with a substitute server for providing the working breakdown services to arriving customers. Liu and Song (2014) extended the idea in Kalidass and Kasturi (2012) to the Markovian batch arrival queue. Applying the matrix-geometric method (Neuts 1981), Liou (2015) also investigated M/M/1 unreliable queue subject to working breakdowns and impatient customers. Other recent works on queues with working breakdowns can be found in Li et al. (2013), Yen et al. (2016), Chen (2018), Ye and Liu (2018), Jain et al. (2019a, 2019b), Jiang and Xin (2019), and Gao et al. (2019). It is, however, slightly regrettable that very few authors have studied the working breakdown queue under general renewal arrivals. To date, only a small amount of literature has covered this topic. Using the embedded Markov chain and matrix analytic approach, Jiang and Liu (2017) considered the GI/M/1 queue with disasters and working breakdowns in a multi-phase service environment. Utilizing the supplementary variable technique (Cox 1955), Yang and Cho (2019) presented a recursive algorithm for computing the stationary queue-length distribution in N-policy GI/M/1/K queue with modified working breakdowns. Here, Yang and Cho did not assume that service and maintenance can be performed simultaneously. In their model assumptions, the repairman repairs the partially failed service facility when there are no customers queueing up for service. We think this assumption may be valid in some cases. For example, as the COVID-19 pandemic has spread across the globe, the severe shortage of medical masks caused by the health crisis quickly becomes a significant issue of the pandemic. To address the growing demand for masks during the outbreak of coronavirus disease, Honeywell, a leading manufacturer of personal protective equipment, quickly ramps up production by reducing regular equipment maintenance frequency. Such maintenance schedule adjustment can be regarded as a fault-tolerance mechanism in the production process that simultaneously provides a timely response to customer needs. This example also gives us a realistic background to study the queueing system with fault-tolerant operations.

It is an indisputable fact that queueing models with customers arriving in batches rather than singly have many applications in practice, for example, in flexible manufacturing systems. However, based on the above literature review, we may see that current research has not paid much attention to the renewal batch arrival queue with working breakdowns or with fault tolerance characteristics. Such a situation prompted us to study a bulk queue with renewal input and fault-tolerant operating modes. Also, to simplify the analytical study of the model, an alternative yet simple method with resorting to the supplementary variable technique and the shift operator technique in solving difference equations is employed to analyze our model. We note that in the past one year, Barbhuiya and Gupta (2019a, 2019b, 2020) used this unique method to reconsider the algorithm for computing the queue-length distributions in some renewal input bulk arrival queues. From their pioneering work, it was revealed that the algorithm based on the combination of these two techniques is easy to understand, and can be conveniently implemented by applying a suitable software package such as Maple, Mathematica or Matlab. More importantly, this

method requires neither derivation of the transition probability matrix of the Markov chain embedded at arrival instants nor the inversion of any probability generating function. We quickly realized that a more generic version of this method could be developed and applied to analyze our queueing model presented in this work.

The remaining part of our paper is organized as follows. We first describe the system and provide basic assumptions in Sect. 2. The differential-difference equations governing the queueing model are framed by the supplementary variable technique in Sect. 3. In Sect. 4, with the help of the theory of difference equation, we provide the procedure to analyze the steady-state queue-length distributions at different epochs of our model. Moreover, using the queue-length distribution at the pre-arrival epoch, we also evaluate the sojourn time of an arbitrary customer in Sect. 5. To validate the correctness of our theoretical results, we present some numerical examples in Sect. 6. Section 7 concludes the paper and directs possible further studies.

2 Model formulation and preliminaries

The mathematical model for this work describes as follows:

- Consider $GI^X/M/1$ queue with a fault-tolerant server wherein batches of customers arrive at epochs $0 = \tau_0, \tau_1, \tau_2, \dots, \tau_n, \dots$. The number of arrivals at each epoch is given by a random variable X having general distribution $g_i = \Pr\{X = i\}$, $i = 1, 2, \dots$, and probability generating function (p.g.f.) $G(z) = \sum_{i=1}^{\infty} g_i z^i$, $|z| \leq 1$. In most real-life situations, batch sizes are not infinite but finite. Thus, from a realistic and computational point of view, we suppose that the maximum batch size is b throughout this paper. As a result of this assumption, the p.g.f. and mean of the random variable X are given by $G(z) = \sum_{i=1}^b g_i z^i$ and $\bar{g} = \sum_{i=1}^b i g_i$, respectively.
- The inter-arrival times $\tau_{n+1} - \tau_n > 0$, $n = 0, 1, 2, \dots$, are independent and identically distributed (i.i.d.) random variables with common distribution function $A(t)$ and probability density function (p.d.f.) $a(t)$. Let the Laplace-Stieltjes transform of this distribution be denoted by $a^*(s) = \int_0^{\infty} e^{-st} dA(t)$ and let the mean inter-arrival time be denoted by $1/\lambda$, where $0 < 1/\lambda = -\left. \frac{d}{ds} a^*(s) \right|_{s=0} < \infty$.
- The customers are served individually by a single server on a first-come, first-served basis. If S_n is the service time of the n th customer to be served in normal state, then it is assumed that $\{S_n, n = 1, 2, 3, \dots\}$ is a sequence of positive i.i.d. random variables with the common exponential p.d.f. $\mu_0 e^{-\mu_0 t}$, $t > 0$.
- A partial failure of the server could occur during the normal busy period, and the time until the random partial breakdown of the server, denoted by L , is assumed to be exponentially distributed with rate η .
- When a partial failure occurs, the server continues working in degraded mode at a reduced service rate rather than entirely halting service. Let \tilde{S}_n represent the service time of the n th customer to be served in defective state, and we also assume that the sequence $\{\tilde{S}_n, n = 1, 2, 3, \dots\}$ are i.i.d. positive random variables following exponential distribution with parameter μ_1 ($\mu_1 < \mu_0$).

- Given customers’ sensitivity and emotional reaction to delay, and also to ensure continuity and efficient service process, the defective server is repaired after the system becomes empty. In other words, when there are customers present in the queue, the server cannot be repaired, even if the server is undergoing repairs. The repair time of the defective server is exponentially distributed with a mean rate β .
- It is further supposed that various stochastic processes involved in the queueing system are mutually independent. Additionally, for the stationary analysis of the model, we demands that $\rho = \bar{g}\lambda/\mu_1 < 1$ (see the proof of Theorems 1 and 2).

3 Governing equations of the system

Let $N(t)$ and $\xi(t)$ indicate the number of customers in the system (including the one being served) and the state of the server at time t , respectively. Here, $\xi(t)$ is a binary random variable. $\xi(t) = 0$ represents that the server is in a normal state, while $\xi(t) = 1$ represents that the server is defective. Moreover, the supplementary variable $R(t)$ corresponding to the remaining inter-batch arrival time at time t is used, which provides Chapman–Kolmogorov equations governing our current model. Thus, the state of the system at time t could be described by a multivariate stochastic process $\Theta(t) = \{N(t), \xi(t), R(t), t \geq 0\}$. In this process, the non-Markovian queueing process becomes Markovian by having the necessary information so that the future of the process depends only on its current state. To ease us into the analysis of such a queue, we define the joint probabilities as follows

$$\pi_{n,m}(x, t)dx = \Pr \{N(t) = n, \xi(t) = m, x < R(t) \leq x + dx\},$$

$$x \geq 0, \quad n = 0, 1, 2, \dots, \quad m = 0, 1.$$

In steady-state, let us further define

$$Q_{n,m}(x) = \lim_{t \rightarrow \infty} \pi_{n,m}(x, t), \quad n = 0, 1, 2, \dots, \quad m = 0, 1;$$

$$Q_{n,m} = \int_0^\infty Q_{n,m}(x)dx, \quad n = 0, 1, 2, \dots, \quad m = 0, 1.$$

To obtain the steady-state probabilities of the queue size $Q_{n,0}$ and $Q_{n,1}$, we first construct the differential-difference equations for describing the evolution of the model by observing the state of the queue at two consecutive time epochs t and $t + \Delta t$. Using the probabilistic argument, we can derive the following set of equations:

$$-\frac{d}{dx}Q_{0,0}(x) = \mu_0Q_{1,0}(x) + \beta Q_{0,1}(x), \tag{1}$$

$$-\frac{d}{dx}Q_{n,0}(x) = -(\mu_0 + \eta)Q_{n,0}(x) + \mu_0Q_{n+1,0}(x)$$

$$+ a(x) \sum_{i=1}^n g_i Q_{n-i,0}(0), \quad 1 \leq n \leq b-1, \tag{2}$$

$$\begin{aligned}
 -\frac{d}{dx}Q_{n,0}(x) &= -(\mu_0 + \eta)Q_{n,0}(x) + \mu_0Q_{n+1,0}(x) \\
 &\quad + a(x) \sum_{i=1}^b g_i Q_{n-i,0}(0), \quad n \geq b,
 \end{aligned} \tag{3}$$

$$-\frac{d}{dx}Q_{0,1}(x) = -\beta Q_{0,1}(x) + \mu_1 Q_{1,1}(x), \tag{4}$$

$$\begin{aligned}
 -\frac{d}{dx}Q_{n,1}(x) &= -\mu_1 Q_{n,1}(x) + \mu_1 Q_{n+1,1}(x) + \eta Q_{n,0}(x) \\
 &\quad + a(x) \sum_{i=1}^n g_i Q_{n-i,1}(0), \quad 1 \leq n \leq b-1,
 \end{aligned} \tag{5}$$

$$\begin{aligned}
 -\frac{d}{dx}Q_{n,1}(x) &= -\mu_1 Q_{n,1}(x) + \mu_1 Q_{n+1,1}(x) + \eta Q_{n,0}(x) \\
 &\quad + a(x) \sum_{i=1}^b g_i Q_{n-i,1}(0), \quad n \geq b.
 \end{aligned} \tag{6}$$

Since the Laplace transform are extremely useful in the solution of complicated differential-difference equations presented above, we define the Laplace transform of $Q_{n,m}(x)$ as

$$Q_{n,m}^*(s) = \int_0^\infty e^{-sx} Q_{n,m}(x) dx, \quad n = 0, 1, 2, \dots, \quad m = 0, 1.$$

Meanwhile, we also notice that $Q_{n,m}^*(0) = Q_{n,m}$. Then, taking the Laplace transform on both sides of Eqs. (1)–(6) and using the relation $-\int_0^\infty e^{-sx} \frac{d}{dx} Q_{n,m}(x) dx = Q_{n,m}(0) - sQ_{n,m}^*(s)$, we thus obtain the transformed equations as below

$$-sQ_{0,0}^*(s) = -Q_{0,0}(0) + \mu_0 Q_{1,0}^*(s) + \beta Q_{0,1}^*(s), \tag{7}$$

$$\begin{aligned}
 (\mu_0 + \eta - s)Q_{n,0}^*(s) &= -Q_{n,0}(0) + \mu_0 Q_{n+1,0}^*(s) \\
 &\quad + a^*(s) \sum_{i=1}^n g_i Q_{n-i,0}(0), \quad 1 \leq n \leq b-1,
 \end{aligned} \tag{8}$$

$$\begin{aligned}
 (\mu_0 + \eta - s)Q_{n,0}^*(s) &= -Q_{n,0}(0) + \mu_0 Q_{n+1,0}^*(s) \\
 &\quad + a^*(s) \sum_{i=1}^b g_i Q_{n-i,0}(0), \quad n \geq b,
 \end{aligned} \tag{9}$$

$$(\beta - s)Q_{0,1}^*(s) = -Q_{0,1}(0) + \mu_1 Q_{1,1}^*(s), \tag{10}$$

$$(\mu_1 - s)Q_{n,1}^*(s) = -Q_{n,1}(0) + \mu_1 Q_{n+1,1}^*(s) + \eta Q_{n,0}^*(s) + a^*(s) \sum_{i=1}^n g_i Q_{n-i,1}(0), \quad 1 \leq n \leq b - 1, \tag{11}$$

$$(\mu_1 - s)Q_{n,1}^*(s) = -Q_{n,1}(0) + \mu_1 Q_{n+1,1}^*(s) + \eta Q_{n,0}^*(s) + a^*(s) \sum_{i=1}^b g_i Q_{n-i,1}(0), \quad n \geq b. \tag{12}$$

Adding Eqs. (7)–(12), it yields

$$\sum_{n=0}^{\infty} Q_{n,0}^*(s) + \sum_{n=0}^{\infty} Q_{n,1}^*(s) = \frac{1 - a^*(s)}{s} \left(\sum_{n=0}^{\infty} Q_{n,0}(0) + \sum_{n=0}^{\infty} Q_{n,1}(0) \right). \tag{13}$$

Taking the limit as $s \rightarrow 0$, and using L'Hôpital's rule one time in the above Eq. (13), we have

$$\sum_{n=0}^{\infty} Q_{n,0}^*(0) + \sum_{n=0}^{\infty} Q_{n,1}^*(0) = \frac{1}{\lambda} \left(\sum_{n=0}^{\infty} Q_{n,0}(0) + \sum_{n=0}^{\infty} Q_{n,1}(0) \right). \tag{14}$$

Here, the normalization condition implies that

$$\sum_{n=0}^{\infty} Q_{n,0}^*(0) + \sum_{n=0}^{\infty} Q_{n,1}^*(0) = 1. \tag{15}$$

Substituting Eq. (15) into Eq. (14), we get

$$\sum_{n=0}^{\infty} Q_{n,0}(0) + \sum_{n=0}^{\infty} Q_{n,1}(0) = \lambda. \tag{16}$$

Employing Eq. (16), we may show that if $Q_{n,m}^-$ represents the steady-state probability of having n ($n \geq 0$) customers in the system and the server being in state m ($m = 0, 1$) immediately before batch arrival, then $Q_{n,m}^- = \frac{1}{\lambda} Q_{n,m}(0)$. To show this, it should be noted that $Q_{n,m}^-$ is proportional to $Q_{n,m}(0)$ and $\sum_{n=0}^{\infty} Q_{n,0}(0) + \sum_{n=0}^{\infty} Q_{n,1}(0)$, which gives

$$Q_{n,m}^- = \frac{Q_{n,m}(0)}{\sum_{n=0}^{\infty} Q_{n,0}(0) + \sum_{n=0}^{\infty} Q_{n,1}(0)} = \frac{1}{\lambda} Q_{n,m}(0). \tag{17}$$

Now, we have already seen that once $Q_{n,m}(0)$ and $Q_{n,m}^*(s)$ are obtained, we can further use them to get the queue-length distributions at the instant of a batch arrival

$(Q_{n,m}^-)$ and arbitrary epoch $(Q_{n,m})$. We will address how to deal with this problem in the next section.

4 Queue-length distributions at pre-arrival and arbitrary epochs

For analysis purposes, the discrete variable n will be considered to be the independent variable, while the function values $Q_{n,m}(0)$ and $Q_{n,m}^*(s)$ ($m = 0, 1$) will be the dependent variables. For the sequences $\{Q_{n,m}(0), n \geq 0\}$ and $\{Q_{n,m}^*(s), n \geq 0\}$, we define a right shift operator \mathbb{T} on the above two sequences and set $\mathbb{T}^j Q_{n,m}(0) = Q_{n+j,m}(0)$ and $\mathbb{T}^j Q_{n,m}^*(s) = Q_{n+j,m}^*(s)$, $j \geq 1$. Thus, Eq. (9) can now be written in terms of operator \mathbb{T} as follows

$$(s - \eta - \mu_0 + \mu_0 \mathbb{T})Q_{n,0}^*(s) = \left(\mathbb{T}^b - a^*(s) \sum_{i=1}^b g_i \mathbb{T}^{b-i} \right) Q_{n-b,0}(0), \quad n \geq b. \tag{18}$$

By setting $n \geq 0$ instead of $n \geq b$ and substituting $s = \eta + \mu_0 - \mu_0 \mathbb{T}$, Eq. (18) reduces to the below homogeneous difference equation with constant coefficients

$$\left(\mathbb{T}^b - a^*(\eta + \mu_0 - \mu_0 \mathbb{T}) \sum_{i=1}^b g_i \mathbb{T}^{b-i} \right) Q_{n,0}(0) = 0, \quad n \geq 0. \tag{19}$$

According to the fundamental theory of difference equation, the corresponding auxiliary equation

$$z^b - a^*(\eta + \mu_0 - \mu_0 z) \sum_{i=1}^b g_i z^{b-i} = 0 \tag{20}$$

is said to be the characteristic equation of Eq. (19). Next, we will apply Rouché’s theorem to find the number of roots of Eq. (20).

Theorem 1 *When $\frac{\bar{g}\lambda}{\mu_0} < 1$, the characteristic equation $z^b - a^*(\eta + \mu_0 - \mu_0 z) \sum_{i=1}^b g_i z^{b-i} = 0$ has exactly b roots inside the unit disk.*

Proof Consider two complex-valued functions $f(z) = z^b$ and $g(z) = -a^*(\eta + \mu_0 - \mu_0 z) \sum_{i=1}^b g_i z^{b-i}$. Let $H(z) = a^*(\eta + \mu_0 - \mu_0 z)$. Clearly, for a sufficiently small $\delta > 0$, $H(z)$ is holomorphic inside and on the closed contour $|z| = 1 + \delta$. Thus, according to the Taylor’s theorem for analytic complex function there exists a power series $\sum_{k=0}^\infty h_k (z - 1)^k$ which converges to $H(z)$. The coefficients h_k ($k = 0, 1, 2, \dots$) are given by $h_k = \frac{1}{2\pi i} \oint_C \frac{H(z)}{(z-1)^{k+1}} dz = \frac{H^{(k)}(1)}{k!}$, where $\mathbf{i} = \sqrt{-1}$ and C is any closed contour around 1 and lying completely inside $|z| \leq 1 + \delta$. Then, employing the Taylor expansion for $H(z)$, we may estimate $|f(z)|$ and $|g(z)|$ on the simple closed curve $|z| = 1 - \epsilon$, where $\epsilon > 0$ and is also a sufficiently small quantity. Thus, on $|z| = 1 - \epsilon$

$$\begin{aligned}
 |f(z)| &= (1 - \epsilon)^b = 1 - b\epsilon + o(\epsilon), \\
 |g(z)| &= |H(z)| \left| \sum_{i=1}^b g_i z^{b-i} \right| \leq H(|z|) \sum_{i=1}^b g_i |z|^{b-i} = H(1 - \epsilon) \sum_{i=1}^b g_i (1 - \epsilon)^{b-i} \\
 &= \left[H(1) + \frac{H'(1)}{1!} (1 - \epsilon - 1) + \sum_{k=2}^{\infty} \frac{H^{(k)}(1)}{k!} (1 - \epsilon - 1)^k \right] \sum_{i=1}^b g_i [1 - (b - i)\epsilon + o(\epsilon)] \\
 &= \left[\int_0^{\infty} (1 - \epsilon \mu_0 t) e^{-\eta t} dA(t) + o(\epsilon) \right] \sum_{i=1}^b g_i [1 - (b - i)\epsilon + o(\epsilon)] \\
 &\leq \left[\int_0^{\infty} (1 - \epsilon \mu_0 t) dA(t) + o(\epsilon) \right] \sum_{i=1}^b g_i [1 - (b - i)\epsilon + o(\epsilon)] \\
 &= \left[1 - \frac{\mu_0}{\lambda} \epsilon + o(\epsilon) \right] [1 + \bar{g}\epsilon - b\epsilon + o(\epsilon)] \\
 &= 1 - b\epsilon + \left(\bar{g} - \frac{\mu_0}{\lambda} \right) \epsilon + o(\epsilon) \leq |f(z)|
 \end{aligned}$$

Consequently, as all the conditions of Rouché’s theorem (see Klimenok 2001) are satisfied, $f(z) + g(z)$ has exactly b zeros inside the unit circle, since $f(z)$ has b . □

It has been shown by many authors that in queueing theory, roots of the characteristic equation are well structured, and they are generally distinct, see, e.g., Tijms (2003) and Chaudhry et al. (1990). Therefore, if we assume that the roots of Eq. (20) are distinct, denoted by $r_j, j = 1, 2, \dots, b$, the general solution of Eq. (19) can be expressed in the form

$$Q_{n,0}(0) = \sum_{j=1}^b \phi_j r_j^n, \quad n \geq 0, \tag{21}$$

where $\phi_j (j = 1, \dots, b)$ are real or complex constants to be determined. Substituting Eq. (21) into Eq. (18) yields

$$(s - \eta - \mu_0 + \mu_0 \mathbb{T}) Q_{n,0}^*(s) = \left(\mathbb{T}^b - a^*(s) \sum_{i=1}^b g_i \mathbb{T}^{b-i} \right) \sum_{j=1}^b \phi_j r_j^{n-b}, \quad n \geq b. \tag{22}$$

It is now necessary to turn to determine the unknown function $Q_{n,0}^*(s)$ from the above equation. Treating s as a fixed constant, Eq. (22) can be regarded as a first order non-homogeneous difference equation with constant coefficients and its general solution consists of two parts: solution to the corresponding homogeneous equation plus a particular solution to the nonhomogeneous equation. For Eq. (22), the general solution of the homogeneous part is written as $Q_{n,0}^{*(\text{hom})}(s) = D_1 \left(1 + \frac{\eta - s}{\mu_0} \right)^n$, where D_1 is an arbitrary constant. On the other hand, we note that an appropriate trial solution for $(s - \eta - \mu_0 + \mu_0 \mathbb{T}) Q_{n,0}^*(s) = \left(r_j^{-b} \mathbb{T}^b - a^*(s) \sum_{i=1}^b g_i r_j^{-b} \mathbb{T}^{b-i} \right) \phi_j r_j^n$ is $Q_{n,0}^*(s) = d_j r_j^n, j = 1, 2, \dots, b$. Thus, a particular solution of the nonhomogeneous Eq. (22) can be given as follows

If the above results hold true, for any $n \geq 1$

$$Q_{n,0}^*(s) = \sum_{j=1}^b \frac{\phi_j \left[r_j^n - a^*(s) \sum_{i=1}^b g_i r_j^{n-i} \right]}{s - \eta - \mu_0(1 - r_j)}. \tag{28}$$

Next, let us once again apply the right shift operator \mathbb{T} to Eq. (12). Then, we have

$$(s - \mu_1 + \mu_1 \mathbb{T})Q_{n,1}^*(s) = \left(\mathbb{T}^b - a^*(s) \sum_{i=1}^b g_i \mathbb{T}^{b-i} \right) Q_{n-b,1}(0) - \eta Q_{n,0}^*(s), \quad n \geq b. \tag{29}$$

By setting $s = \mu_1 - \mu_1 \mathbb{T}$ in Eq. (29) and substituting Eq. (28) into Eq. (29) yields

$$\begin{aligned} & \left[\mathbb{T}^b - a^*(\mu_1 - \mu_1 \mathbb{T}) \sum_{i=1}^b g_i \mathbb{T}^{b-i} \right] Q_{n,1}(0) = \eta Q_{n+b,0}^*(\mu_1 - \mu_1 \mathbb{T}) \\ & = \eta \sum_{j=1}^b \frac{\phi_j \left[r_j^b - a^*(\mu_1 - \mu_1 \mathbb{T}) \sum_{i=1}^b g_i r_j^{b-i} \right]}{\mu_1 - \mu_1 \mathbb{T} - \eta - \mu_0(1 - r_j)} r_j^n, \quad n \geq 0. \end{aligned} \tag{30}$$

Adopting Rouché’s theorem similar to that used for the previous Theorem 1, the following Theorem 2 shows that under certain conditions, the characteristic equation of the above difference equation also has exactly b roots inside the unit circle $|z| = 1$. We shall use the letters $\omega_1, \omega_2, \dots, \omega_b$ to denote these b roots.

Theorem 2 *When $\frac{\bar{g}\lambda}{\mu_1} < 1$, the characteristic equation $z^b - a^*(\mu_1 - \mu_1 z) \sum_{i=1}^b g_i z^{b-i} = 0$ has exactly b roots inside the unit disk.*

Furthermore, combining the results of Theorems 1 and Theorem 2, the obvious stability condition of the current queueing system is that $\frac{\bar{g}\lambda}{\mu_1} < 1$. By analogy with the corresponding procedure to compute $Q_{n,0}^*(s)$, the general solution of the non-homogeneous Eq. (30) is given by

$$Q_{n,1}(0) = \sum_{j=1}^b k_j \omega_j^n + \eta \sum_{j=1}^b \frac{\phi_j}{(\mu_1 - \mu_0)(1 - r_j) - \eta} r_j^n, \quad n \geq 0. \tag{31}$$

Here, the first term in the right-hand side of Eq. (31) is a solution to the associated homogeneous equation of Eq. (30), and k_1, k_2, \dots, k_b are the arbitrary constants that can be obtained from later analysis. On the other hand, the second term is a particular solution of Eq. (30). Substituting Eqs. (28) and (31) into the right-hand side of Eq. (29), we have

$$\begin{aligned}
 (s - \mu_1 + \mu_1 \mathbb{T})Q_{n,1}^*(s) &= \left(\mathbb{T}^b - a^*(s) \sum_{i=1}^b g_i \mathbb{T}^{b-i} \right) \left(\sum_{j=1}^b k_j \omega_j^{n-b} \right. \\
 &\quad \left. + \eta \sum_{j=1}^b \frac{\phi_j}{(\mu_1 - \mu_0)(1 - r_j) - \eta} r_j^{n-b} \right) \\
 &\quad - \eta \sum_{j=1}^b \frac{\phi_j \left[r_j^n - a^*(s) \sum_{i=1}^b g_i r_j^{n-i} \right]}{s - \eta - \mu_0(1 - r_j)}, \\
 &= \sum_{j=1}^b k_j \omega_j^n - a^*(s) \sum_{j=1}^b k_j \sum_{i=1}^b g_i \omega_j^{n-i} \\
 &\quad + \eta \sum_{j=1}^b \frac{\phi_j \left(r_j^n - a^*(s) \sum_{i=1}^b g_i r_j^{n-i} \right)}{(\mu_1 - \mu_0)(1 - r_j) - \eta} \\
 &\quad - \eta \sum_{j=1}^b \frac{\phi_j \left[r_j^n - a^*(s) \sum_{i=1}^b g_i r_j^{n-i} \right]}{s - \eta - \mu_0(1 - r_j)}, \quad n \geq b,
 \end{aligned} \tag{32}$$

which is also a nonhomogeneous difference equation with constant coefficients. Similar to the discussions surrounding Eq. (22), we of course first find the general solution of the associated homogeneous equation of Eq. (32). It has the form $Q_{n,1}^{*(\text{hom})}(s) = D_2 \left(1 - \frac{s}{\mu_1} \right)^n$, where D_2 is an undetermined constant. Then, utilizing the Table 2.3 presented in the monograph by Elaydi (2005), a particular solution of Eq. (32) can be given by

$$\begin{aligned}
 Q_{n,1}^{*(\text{part})}(s) &= \sum_{j=1}^b \frac{k_j \left(\omega_j^n - a^*(s) \sum_{i=1}^b g_i \omega_j^{n-i} \right)}{s - \mu_1(1 - \omega_j)} + \sum_{j=1}^b \frac{\phi_j \eta \left(r_j^n - a^*(s) \sum_{i=1}^b g_i r_j^{n-i} \right)}{s - \mu_1(1 - r_j)} \\
 &\quad \left[\frac{1}{(\mu_1 - \mu_0)(1 - r_j) - \eta} - \frac{1}{s - \eta - \mu_0(1 - r_j)} \right] \\
 &= \sum_{j=1}^b \frac{k_j \left(\omega_j^n - a^*(s) \sum_{i=1}^b g_i \omega_j^{n-i} \right)}{s - \mu_1(1 - \omega_j)} \\
 &\quad + \sum_{j=1}^b \frac{\phi_j \eta \left(r_j^n - a^*(s) \sum_{i=1}^b g_i r_j^{n-i} \right)}{[(\mu_1 - \mu_0)(1 - r_j) - \eta] [s - \eta - \mu_0(1 - r_j)]}.
 \end{aligned} \tag{33}$$

Thus, for $n \geq b$, the general solution of Eq. (32) can be written as $Q_{n,1}^*(s) = Q_{n,1}^{*(\text{part})}(s) + Q_{n,1}^{*(\text{hom})}(s)$. Summing over all permissible n from b to ∞ and taking the limit as $s \rightarrow 0$, the formula $\sum_{n=b}^{\infty} Q_{n,1}^*(0) = \sum_{n=b}^{\infty} Q_{n,1} \leq 1$ clearly holds. This means that the undetermined constant $D_2 = 0$. Otherwise, the limit of $\sum_{n=b}^{\infty} D_2 \left(1 - \frac{s}{\mu_2} \right)^n$, as s approaches 0, is infinity. Therefore, the solution of Eq. (32) takes the below form

$$\sum_{j=1}^b k_j + \eta \sum_{j=1}^b \frac{\phi_j}{(\mu_1 - \mu_0)(1 - r_j) - \eta} - \mu_1 \sum_{j=1}^b \frac{k_j (1 - a^*(\beta)G(\omega_j^{-1}))\omega_j}{\beta - \mu_1(1 - \omega_j)} - \mu_1 \sum_{j=1}^b \frac{\phi_j \eta (1 - a^*(\beta)G(r_j^{-1}))r_j}{[(\mu_1 - \mu_0)(1 - r_j) - \eta][\beta - \eta - \mu_0(1 - r_j)]} = 0. \tag{38}$$

By summing over all possible n from 0 to ∞ in Eq. (16) and employing Eqs. (21) and (31), we can further derive a relationship for ϕ_j and k_j , $j = 1, 2, \dots, b$.

$$\sum_{j=1}^b \frac{\phi_j(\mu_1 - \mu_0)}{(\mu_1 - \mu_0)(1 - r_j) - \eta} + \sum_{j=1}^b \frac{k_j}{1 - \omega_j} = \lambda. \tag{39}$$

So far, by the use of Eqs. (27), (36), (38) and (39), we have established a system of $2b$ linear equations in $2b$ variables which can be solved to obtain the constants ϕ_j 's and k_j 's, $j = 1, 2, \dots, b$. Once the arbitrary constants are determined by solving the linear algebraic equations, the stationary queue-length distributions at pre-arrival and arbitrary epochs are given, respectively, by

$$\begin{cases} Q_{n,0}^- = \frac{1}{\lambda} Q_{n,0}(0) = \frac{1}{\lambda} \sum_{j=1}^b \phi_j r_j^n, & n \geq 0, \\ Q_{n,1}^- = \frac{1}{\lambda} Q_{n,1}(0) = \frac{1}{\lambda} \left(\sum_{j=1}^b k_j \omega_j^n + \eta \sum_{j=1}^b \frac{\phi_j}{(\mu_1 - \mu_0)(1 - r_j) - \eta} r_j^n \right), & n \geq 0, \end{cases} \tag{40}$$

$$\begin{cases} Q_{n,0} = Q_{n,0}^*(0) = \sum_{j=1}^b \frac{\phi_j [G(r_j^{-1}) - 1]}{\eta + \mu_0(1 - r_j)} r_j^n, & n \geq 1, \\ Q_{n,1} = Q_{n,1}^*(0) = \sum_{j=1}^b \frac{k_j [G(\omega_j^{-1}) - 1] \omega_j^n}{\mu_1(1 - \omega_j)} + \sum_{j=1}^b \frac{\phi_j \eta [G(r_j^{-1}) - 1] r_j^n}{[(\mu_1 - \mu_0)(1 - r_j) - \eta][\eta + \mu_0(1 - r_j)]}, & n \geq 1. \end{cases} \tag{41}$$

Remark 1 Our queueing model is a generalization of the classical M/M/1 queue. More precisely, such a model reduces to M/M/1 queue when $\eta \rightarrow 0$, $\beta \rightarrow \infty$, $\mu_0 = \mu_1$, $g_1 = 1$ and $a^*(s) = \frac{\lambda}{s + \lambda}$. For this case $b = 1$, thus the single characteristic root inside the unit circle is $r_1 = \omega_1 = \lambda/\mu_1$. Hence, we may compute the probability that an arriving customer finds an empty system (denoted as Q_0^-) from Eq. (40) and obtain $Q_0^- = Q_{0,0}^- + Q_{0,1}^- = \frac{\phi_1 + k_1}{\lambda}$. On the other hand, substituting $r_1 = \omega_1 = \lambda/\mu_1$ into Eq. (39), we can find that $\frac{\phi_1 + k_1}{\lambda}$ satisfies the relationship $\frac{\phi_1 + k_1}{\lambda} = \left(1 - \frac{\lambda}{\mu_1}\right) = (1 - \rho) = Q_0^-$. The relationship also implies that $Q_n^- = Q_{n,0}^- + Q_{n,1}^- = \frac{\phi_1 + k_1}{\lambda} r_1^n = (1 - \rho)\rho^n$, $n = 1, 2, \dots$, where Q_n^- represents the probability that there are n customers in the system just before the arrival of a customer. Therefore, according to the PASTA property (Poisson Arrivals See Time Averages), we get exactly the same results that have been reported in the existing literature on queueing theory (see Gross and Harris 1985).

Furthermore, setting $s = 0$ in Eq. (10), we obtain

$$\begin{aligned}
 Q_{0,1} &= Q_{0,1}^*(0) = \frac{1}{\beta} \left[\mu_1 Q_{1,1}^*(0) - Q_{0,1}(0) \right] \\
 &= \frac{\mu_1}{\beta} \left(\sum_{j=1}^b \frac{k_j [G(\omega_j^{-1}) - 1] \omega_j}{\mu_1(1 - \omega_j)} + \sum_{j=1}^b \frac{\phi_j \eta [G(r_j^{-1}) - 1] r_j}{[(\mu_1 - \mu_0)(1 - r_j) - \eta] [\eta + \mu_0(1 - r_j)]} \right) \\
 &\quad - \frac{1}{\beta} \left(\sum_{j=1}^b k_j + \eta \sum_{j=1}^b \frac{\phi_j}{(\mu_1 - \mu_0)(1 - r_j) - \eta} \right).
 \end{aligned}
 \tag{42}$$

With the help of the normalizing condition $\sum_{n=0}^{\infty} Q_{n,0} + \sum_{n=0}^{\infty} Q_{n,1} = 1$, we may further obtain

$$\begin{aligned}
 Q_{0,0} &= 1 - \sum_{j=1}^b \frac{\phi_j r_j [G(r_j^{-1}) - 1]}{\eta(1 - r_j) + \mu_0(1 - r_j)^2} - \sum_{j=1}^b \frac{k_j \omega_j [G(\omega_j^{-1}) - 1]}{\mu_1(1 - \omega_j)^2} \\
 &\quad - \sum_{j=1}^b \frac{\phi_j r_j \eta [G(r_j^{-1}) - 1]}{[(\mu_1 - \mu_0)(1 - r_j)^2 - \eta(1 - r_j)] [\eta + \mu_0(1 - r_j)]} \\
 &\quad - \frac{\mu_1}{\beta} \left(\sum_{j=1}^b \frac{k_j [G(\omega_j^{-1}) - 1] \omega_j}{\mu_1(1 - \omega_j)} \right. \\
 &\quad \left. + \sum_{j=1}^b \frac{\phi_j \eta [G(r_j^{-1}) - 1] r_j}{[(\mu_1 - \mu_0)(1 - r_j) - \eta] [\eta + \mu_0(1 - r_j)]} \right) \\
 &\quad + \frac{1}{\beta} \left(\sum_{j=1}^b k_j + \eta \sum_{j=1}^b \frac{\phi_j}{(\mu_1 - \mu_0)(1 - r_j) - \eta} \right)
 \end{aligned}
 \tag{43}$$

This completes the analysis of queue-length distribution at different epochs.

5 Sojourn time for an arbitrary customer

The sojourn time T_A is the total time that an arbitrary customer in an arriving batch spends in the system until it departs from the system. Using the pre-arrival epoch probabilities $Q_{n,0}^-$ and $Q_{n,1}^-$ that we have derived in Sect. 4, we will investigate the Laplace–Stieltjes transform and the expectation of an arbitrary customer’s sojourn time in this section. For this purpose, we introduce a tagged customer who may arrive at a random position in the queue. Clearly, it may be seen that a tagged customer arrival may belong to one of the following cases:

Case 1. A batch containing the tagged customer arrives at the system during the normal working period and finds n customers already present in the system. Meanwhile, the number of customers that arrive in the same batch as the tagged customer, but enter service before the tagged customer is l ($l = 0, 1, \dots, b - 1$). To discuss the sojourn time of the tagged customer, this case can further be divided into two subcases: (i) The time until the random partial breakdown of the server is no less than the total service time of $n + l + 1$ customers. That is to say, the tagged customer completes its service and leaves the system in the normal working period; (ii) There are only m ($m = 0, 1, \dots, n + l$) service completions during the normal working period. In other words, the tagged customer leaves the system when the server undergoes in the defective state.

Case 2. An arriving batch sees the state of the system is $(n, 1)$, $n \geq 0$, and the $(l + 1)$ th customer in this batch is the tagged customer. Then, the sojourn time of the tagged customer equals the total customers' service times ahead of him plus his service time. Since the server is defective, these services are only provided with the low service rate μ_1 .

Let $W_A(t)$ be defined as the probability distribution function of T_A , and g_l^- denotes the probability of l number of customers ahead of a randomly selected tagged customer within the batch. From Burke (1975) and Chaudhry and Templeton (1983), g_l^- is given by $g_l^- = \frac{1}{g} \sum_{j=l+1}^{\infty} g_j$. Based on the above different cases, we have, from the theorem of total probability,

$$\begin{aligned}
 W_A(t) &= \Pr \{ T_A \leq t \} \\
 &= \sum_{n=0}^{\infty} Q_{n,0}^- \left[\sum_{l=0}^{b-1} g_l^- \left(\Pr \left\{ T_A = \sum_{h=1}^{n+l+1} S_h \leq t \mid \sum_{h=1}^{n+l+1} S_h \leq L \right\} \Pr \left\{ \sum_{h=1}^{n+l+1} S_h \leq L \right\} \right. \right. \\
 &\quad \left. \left. + \sum_{m=0}^{n+l} \Pr \left\{ T_A = L + \sum_{h=m+1}^{n+l+1} \tilde{S}_h \leq t \mid \sum_{h=1}^m S_h \leq L < \sum_{h=1}^{m+1} S_h \right\} \right. \right. \\
 &\quad \left. \left. \Pr \left\{ \sum_{h=1}^m S_h \leq L < \sum_{h=1}^{m+1} S_h \right\} \right) \right] \\
 &\quad + \sum_{n=0}^{\infty} Q_{n,1}^- \sum_{l=0}^{b-1} g_l^- \Pr \left\{ T_A = \sum_{h=1}^{n+l+1} \tilde{S}_h \leq t \right\} \\
 &= \sum_{n=0}^{\infty} Q_{n,0}^- \left[\sum_{l=0}^{b-1} g_l^- \left(\int_0^t \frac{\mu_0(\mu_0 x)^{n+l}}{(n+l)!} e^{-(\mu_0+\eta)x} dx \right. \right. \\
 &\quad \left. \left. + \sum_{m=0}^{n+l} \int_0^t \frac{(\mu_0 x)^m}{m!} e^{-\mu_0 x} \eta e^{-\eta x} \left(1 - e^{-\mu_1(t-x)} \sum_{i=0}^{n+l-m} \frac{(\mu_1(t-x))^i}{i!} \right) dx \right) \right] \\
 &\quad + \sum_{n=0}^{\infty} Q_{n,1}^- \sum_{l=0}^{b-1} g_l^- \left[1 - e^{-\mu_1 t} \sum_{i=0}^{n+l} \frac{(\mu_1 t)^i}{i!} \right].
 \end{aligned}$$

(44)

The Laplace–Stieltjes transform of the distribution of T_A is given by

$$\begin{aligned}
 W_A^*(s) &= \int_0^\infty e^{-st} dW_A(t) \\
 &= \sum_{n=0}^\infty Q_{n,0}^- \left[\sum_{l=0}^{b-1} g_l^- \left(\left(\frac{\mu_0}{s + \mu_0 + \eta} \right)^{n+l+1} \right. \right. \\
 &\quad \left. \left. + \sum_{m=0}^{n+l} \frac{\eta \mu_0^m}{(s + \mu_0 + \eta)^{m+1}} \left(\frac{\mu_1}{s + \mu_1} \right)^{n+l+1-m} \right) \right] \\
 &\quad + \sum_{n=0}^\infty Q_{n,1}^- \sum_{l=0}^{b-1} g_l^- \left(\frac{\mu_1}{s + \mu_1} \right)^{n+l+1}.
 \end{aligned} \tag{45}$$

Hence, the first and second moments of T_A may be found from the Laplace–Stieltjes transform as

$$\begin{aligned}
 E[T_A] &= -\frac{d}{ds} W_A^*(s) \Big|_{s=0} \\
 &= \sum_{n=0}^\infty Q_{n,0}^- \left[\sum_{l=0}^{b-1} g_l^- \left(\frac{(n+l+1)\mu_0^{n+l+1}}{(\mu_0 + \eta)^{n+l+2}} \right. \right. \\
 &\quad \left. \left. + \sum_{m=0}^{n+l} \frac{(m+1)\eta\mu_0^m}{(\mu_0 + \eta)^{m+2}} + \frac{\eta\mu_0^m}{(\mu_0 + \eta)^{m+1}} \frac{(n+l+1-m)}{\mu_1} \right) \right] \\
 &\quad + \sum_{n=0}^\infty Q_{n,1}^- \sum_{l=0}^{b-1} g_l^- \frac{(n+l+1)}{\mu_1},
 \end{aligned} \tag{46}$$

$$\begin{aligned}
 E[T_A^2] &= \frac{d^2}{ds^2} W_A^*(s) \Big|_{s=0} \\
 &= \sum_{n=0}^\infty Q_{n,0}^- \left[\sum_{l=0}^{b-1} g_l^- \left(\frac{(n+l+1)(n+l+2)\mu_0^{n+l+1}}{(\mu_0 + \eta)^{n+l+3}} \right. \right. \\
 &\quad \left. \left. + \sum_{m=0}^{n+l} \left(\frac{(m+1)(m+2)\eta\mu_0^m}{(\mu_0 + \eta)^{m+3}} \right. \right. \right. \\
 &\quad \left. \left. \left. + \frac{2(m+1)(n+l+1-m)\eta\mu_0^m}{(\mu_0 + \eta)^{m+2}\mu_1} + \frac{(n+l+1-m)(n+l+2-m)}{(\mu_0 + \eta)^{m+1}\mu_1^2} \right) \right) \right] \\
 &\quad + \sum_{n=0}^\infty Q_{n,1}^- \sum_{l=0}^{b-1} g_l^- \frac{(n+l+1)(n+l+2)}{\mu_1^2}.
 \end{aligned} \tag{47}$$

From the numerical examples presented in the next section, we will see that Eq. (46) can provide us an effective way to validate the correctness of our theoretical analysis results. Moreover, using Eqs. (46) and (47) we can compute the variance of T_A as $\text{Var}(T_A) = E[T_A^2] - E^2[T_A]$.

6 Numerical illustrations

To demonstrate the working schemes of the difference equation approach based on the right shift operator, we first describe the solution algorithm for calculating the steady-state probabilities $Q_{n,0}^-$, $Q_{n,1}^-$, $Q_{n,0}$ and $Q_{n,1}$, for $n \geq 0$. Given the values of μ_0 , μ_1 , η , β , the Laplace–Stieltjes transform expression of the inter-batch arrival time $a^*(s)$ and the probability mass function $g_i = \Pr\{X = i\}$, $i = 1, 2, \dots, b$, the steps of the solution algorithm are stated as follows:

- **Step 1:** Find the roots of the following two characteristic equations $z^b - a^*(\eta + \mu_0 - \mu_0 z) \sum_{i=1}^b g_i z^{b-i} = 0$ and $z^b - a^*(\mu_1 - \mu_1 z) \sum_{i=1}^b g_i z^{b-i} = 0$ inside the unit circle, respectively. Denote these roots as r_j and ω_j , $j = 1, 2, \dots, b$.
- **Step 2:** Insert these roots directly into Eqs. (27), (36), (38) and (39), and solve a system of linear equations to find the values of the unknown constants ϕ_j and k_j , $j = 1, 2, \dots, b$.
- **Step 3:** Substituting the values of ϕ_j and k_j into Eq. (40), find $Q_{n,0}^-$ and $Q_{n,1}^-$, for $n \geq 0$.
- **Step 4:** Inserting the values of ϕ_j and k_j into Eqs. (41), (42) and (43), compute $Q_{n,0}$ and $Q_{n,1}$, for $n \geq 0$.

All the calculations are performed on a PC having Corei7 processor at 3.20 gigahertz with 16 gigabytes RAM using Mathematica and Matlab software packages. We use Mathematica software to find the roots of the associated characteristic equations, and then write a Matlab code to solve a linear system of equations in $2b$ variables. Next, to illustrate the solution algorithm, we provide three numerical examples where the inter-batch arrival time distributions are 2-stage Erlang, uniform and deterministic, respectively. A variety of numerical results have been presented in self-explanatory tables and graphs. The notations used in these tables are the same as those defined earlier in the previous sections except $E[T_A]_{\text{Little}}$, which denotes the average sojourn time of an arbitrary customer evaluated through Little's formula.

Example 1 Consider the $E_2^X/M/1$ queueing system with fault tolerance capabilities. The 2-stage Erlang distribution is made up of two independent and identical exponential stages, each with mean 2.5. In this case, we have $a^*(s) = \left(\frac{0.4}{s+0.4}\right)^2$. The number of customers (X) belonging to each arrival has the following probability mass function $g_i = \Pr\{X = i\} = \frac{0.55(1-0.55)^{i-1}}{1-(1-0.55)^{10}}$ ($i = 1, 2, \dots, 10$) with mean value $\bar{g} = 1.814776$. For computation purpose, we fix the other parameters as $\mu_0 = 0.7$, $\mu_1 = 0.4$, $\eta = 0.005$ and $\beta = 0.2$. Table 1 displays the the roots of the characteristic equations inside the unit circle, where i is the imaginary unit. The corresponding constants ϕ_j and k_j for $j = 1, 2, \dots, 10$ are determined by solving a systems of linear equations with twenty variables in computing software. The calculation results are reported in Table 2.

With the known values of ϕ_j , k_j , r_j and ω_j , Table 3 also gives a few queue-length distributions at pre-arrival and arbitrary epochs. Utilizing the data presented in

Table 3, an effective approach is provided to verify the correctness of our numerical as well as analytical results. Actually, there are two different ways to get the average sojourn time for an arbitrary customer. One way is to substitute $Q_{n,0}^-$ and $Q_{n,1}^-$ into Eq. (46) and calculate $E[T_A]$ directly. The other is to apply Little’s formula to obtain $E[T_A] = L_s/\bar{g}\lambda$, where $L_s = \sum_{n=1}^{\infty} n(Q_{n,0} + Q_{n,1})$. The results from numerical computation (see the bottom of Table 3) indicate that $E[T_A]$ evaluated through Eq. (46) exactly matches with the one obtained from Little’s formula. It also implies that the theoretical analysis and numerical experiments performed in this paper is valid and reliable.

In the first example, we consider the case when inter-arrival times of groups have Erlangian distribution of order 2. We note that such an arrival process is a particular case of the batch Markovian arrival process (BMAP). The corresponding numerical results can be much easier obtained by using the well-known matrix geometric method. To further demonstrate the universality of the proposed method, some other examples with an inter-batch arrival time that does not belong to the class of PH distribution are given below.

Example 2 Consider the $U^X/M/1$ queue with fault tolerance capabilities, where the inter-batch arrival times are independent random variables, each distributed uniformly on interval 0 to $2/\lambda$. If we set $\lambda = 0.2$, the uniform distribution has the Laplace–Stieltjes transform $a^*(s) = \frac{1}{10s}(1 - e^{-10s})$. The maximum batch size is $b = 12$, and take the probability mass function of batch size to be $g_i = \Pr\{X = i\} = \frac{e^{-22i}}{i! \sum_{n=1}^b \frac{e^{-22n}}{n!}}$, $i = 1, 2, \dots, 12$ with mean value $\bar{g} = 1.930825$. Since $a^*(s)$ is a transcendental function, the corresponding characteristic equations cannot be directly solved by using the standard Mathematica commands. For the purpose of finding the roots of the characteristic equations, we wish to approximate $a^*(s)$ by means of a rational approximation $\mathbb{P}_1(s)/\mathbb{P}_2(s)$, where $\mathbb{P}_1(s)$ and $\mathbb{P}_2(s)$ are polynomials of degree m_1 and m_2 respectively. It is well known that the Padé approximation is a particular and classical type of rational fraction approximation. Many practical applications have proven that it is the best approximation of a function by a rational function of a given order. In this example, we shall use the so-called Padé rational approximation of degree (15, 16) to approximate $a^*(s)$

$$\begin{aligned}
 a^*(s) &= \frac{\mathbb{P}_1(s)}{\mathbb{P}_2(s)} \\
 &= \frac{1 + \frac{350s^2}{93} + \frac{3250s^4}{899} + \frac{687500s^6}{509733} + \frac{343750s^8}{1529199} + \frac{312500s^{10}}{18423207} + \frac{7812500s^{12}}{15088606533} + \frac{78125000s^{14}}{18061062020001}}{\left(1 + 5s + \frac{375s^2}{31} + \frac{1750s^3}{93} + \frac{56875s^4}{2697} + \frac{16250s^5}{899} + \frac{893750s^6}{72819} + \frac{3437500s^7}{509733} \right. \\
 &\quad \left. + \frac{171875s^8}{1718750s^9} + \frac{1718750s^{10}}{1529199} + \frac{5024511}{18423207} + \frac{1562500s^{11}}{1160662041} + \frac{19531250s^{12}}{244140625s^{16}} \right) \\
 &\quad \left. + \frac{39062500s^{13}}{15088606533} + \frac{1529199}{668928222963} + \frac{5024511}{18061062020001} + \frac{1160662041}{307038054340017} \right)
 \end{aligned}$$

Here, calculations with Padé approximant of $a^*(s)$ are straightforward and can be performed with Mathematica command “PadeApproximant”. How the choice of Padé (m_1, m_2) affects the accuracy of numerical results and how one can trade off between computation-time and accuracy is discussed in detail in the work of Singh

et al. (2014). By the fixed values of certain parameters μ_0, μ_1, η and β as 0.75, 0.5, 0.003 and 0.1 respectively, we present the numerical results for the roots of the characteristic equations inside the unit circle in Table 4.

Substituting these roots into Eqs. (27), (36), (38) and (39), and solving a simultaneous set of twenty-four linear equations, Table 5 gives the numerical results of ϕ_j and k_j for $j = 1, 2, \dots, 12$.

The summary of the calculations for the queue-length distributions at pre-arrival and arbitrary epochs is shown in Table 6. At the same time, the expected sojourn time for an arbitrary customer estimated in two different ways is also summarized at the bottom of Table 6. We may see that the numerical experiments verify our theoretical results and show their correctness once again.

Example 3 Consider the $D^X/M/1$ queue with a fault-tolerant server, in which the batch arrivals are equally spaced in time. For this inter-batch arrival time distribution $a^*(s) = e^{-\frac{s}{\lambda}}$, here we take $\lambda = 0.2$. Now, let us conduct the numerical experiment with the following parameters: $\mu_0 = 0.8, \mu_1 = 0.6, \eta = 0.0025$ and $\beta = 0.15$. Additionally, the corresponding numerical results were obtained by assuming that the batch size distribution has a p.g.f. $G(z) = 0.25z + 0.5z^2 + 0.1z^4 + 0.1z^6 + 0.05z^8$. This suggests that the batch size can be either 1, 2, 4, 6, or 8 with 25%, 50%, 10%, 10% and 5% probabilities, respectively. Since the deterministic inter-batch arrival time does not have a rational Laplace–Stieltjes transform, Mathematica software package cannot solve the characteristic equations (transcendental equations) directly. Through the Padé approximation of degree (8, 9), we also approximate the Laplace–Stieltjes transform of the inter-batch arrival time distribution e^{-5s} with a rational function of the type $\mathbb{P}_1(s)/\mathbb{P}_2(s)$:

$$e^{-5s} = \frac{\mathbb{P}_1(s)}{\mathbb{P}_2(s)} = \frac{1 - \frac{40s}{17} + \frac{175s^2}{68} - \frac{175s^3}{102} + \frac{625s^4}{816} - \frac{625s^5}{2652} + \frac{3125s^6}{63648} - \frac{15625s^7}{2450448} + \frac{15625s^8}{39207168}}{1 + \frac{45s}{17} + \frac{225s^2}{68} + \frac{175s^3}{68} + \frac{375s^4}{272} + \frac{1875s^5}{3536} + \frac{3125s^6}{21216} + \frac{15625s^7}{544544} + \frac{15625s^8}{4356352} + \frac{78125s^9}{352864512}}.$$

Letting $s = 0.8025 - 0.8z$ and $s = 0.6 - 0.6z$ in the above expression, respectively, and putting them into the characteristic equations, Table 7 gives all distinct roots that are found within the contour of a unit circle $|z| = 1$.

Table 1 The roots of the characteristic equations with modulus less than one in Example 1

| j | r_j | j | r_j | j | ω_j | j | ω_j |
|-----|----------------------|-----|-----------------------|-----|----------------------|-----|-----------------------|
| 1 | 0.656426 | 6 | -0.135281 - 0.323903i | 1 | 0.932260 | 6 | -0.140912 - 0.345224i |
| 2 | 0.294296 - 0.270739i | 7 | -0.135281 + 0.323903i | 2 | 0.314966 - 0.276982i | 7 | -0.140912 + 0.345224i |
| 3 | 0.294296 + 0.270739i | 8 | -0.284697 - 0.188870i | 3 | 0.314966 + 0.276982i | 8 | -0.302061 - 0.202056i |
| 4 | 0.074476 - 0.361137i | 9 | -0.284697 + 0.188870i | 4 | 0.084063 - 0.381289i | 9 | -0.302061 + 0.202056i |
| 5 | 0.074476 + 0.361137i | 10 | -0.338701 | 5 | 0.084063 + 0.381289i | 10 | -0.360414 |

Table 2 The values of ϕ_j and k_j for $j = 1, 2, \dots, 10$ in Example 1

| j | ϕ_j | j | ϕ_j | j | k_j | j | k_j |
|-----|------------------------|-----|------------------------|-----|------------------------|-----|------------------------|
| 1 | 0.032427 | 6 | $0.006740 + 0.000394i$ | 1 | 0.003565 | 6 | $0.000159 + 0.000020i$ |
| 2 | $0.006332 + 0.001687i$ | 7 | $0.006740 - 0.000394i$ | 2 | $0.000124 + 0.000069i$ | 7 | $0.000159 - 0.000020i$ |
| 3 | $0.006332 - 0.001687i$ | 8 | $0.006745 + 0.000173i$ | 3 | $0.000124 - 0.000069i$ | 8 | $0.000161 + 0.000009i$ |
| 4 | $0.006695 + 0.000769i$ | 9 | $0.006745 - 0.000173i$ | 4 | $0.000153 + 0.000038i$ | 9 | $0.000161 - 0.000009i$ |
| 5 | $0.006695 - 0.000769i$ | 10 | 0.006745 | 5 | $0.000153 - 0.000038i$ | 10 | 0.000161 |

Table 3 Steady-state queue-length distributions at pre-arrival and arbitrary epochs in Example 1

| n | $Q_{n,0}^-$ | $Q_{n,1}^-$ | $Q_{n,0}$ | $Q_{n,1}$ |
|-----|-------------|-------------|-----------|-----------|
| 0 | 0.460979 | 0.012428 | 0.371024 | 0.009600 |
| 1 | 0.099255 | 0.011868 | 0.128965 | 0.011014 |
| 2 | 0.066009 | 0.012336 | 0.085781 | 0.011917 |
| 3 | 0.043884 | 0.012349 | 0.057042 | 0.012210 |
| 4 | 0.029159 | 0.012078 | 0.037916 | 0.012116 |
| 5 | 0.019360 | 0.011636 | 0.025187 | 0.011784 |
| 6 | 0.012839 | 0.011099 | 0.016716 | 0.011311 |
| 7 | 0.008499 | 0.010514 | 0.011079 | 0.010761 |
| 8 | 0.005611 | 0.009911 | 0.007326 | 0.010176 |
| 9 | 0.003691 | 0.009312 | 0.004829 | 0.009581 |
| 10 | 0.002418 | 0.008728 | 0.003167 | 0.008994 |
| 11 | 0.001579 | 0.008168 | 0.002063 | 0.008424 |
| 12 | 0.001037 | 0.007634 | 0.001355 | 0.007880 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 70 | 0.000000 | 0.000000 | 0.000000 | 0.000136 |
| 71 | 0.000000 | 0.000000 | 0.000000 | 0.000127 |
| 72 | 0.000000 | 0.000000 | 0.000000 | 0.000118 |
| 73 | 0.000000 | 0.000000 | 0.000000 | 0.000110 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Sum | 0.756301 | 0.243699 | 0.755040 | 0.244960 |

$$E[T_A] = 13.304375, E[T_A]_{\text{Little}} = 13.304375$$

The unknown constants ϕ_j and k_j for $j = 1, 2, \dots, 8$ can be found in the same manner as described earlier. Table 8 displays all the calculation results.

Using Eqs. (40), (41), (42) and (43), we compute the probability distributions of the queue length at two different epochs, and the numerical results are listed in Table 9. It is also found that $E[T_A]$ matches exactly with the mean sojourn time calculated using Little’s law.

Example 4 In this example, the computer software, e.g., Mathematica and Matlab, are used to compare three configurations in terms of their $E[T_A]$ and L_s for three different inter-batch arrival time distributions: 2-stage Erlang, uniform and deterministic. We first perform a comparison for the average sojourn time of an arbitrary

customer using the assumption that these distributions have the same mean 5 but different standard deviations. We choose $\mu_0 = 0.9$, $\mu_1 = 0.75$, $\eta = 0.0025$, and suppose X is a discrete uniform random variable on the consecutive integers 1, 2, ..., 6, so that the traffic intensity $\rho = 0.883333$. At the same time, to evaluate the impact of the repair rate β on the value of $E[T_A]$, we vary the values of β from 0.05 to 0.25 and draw the plot of $E[T_A]$ as a function of β . Figure 1(a) depicts that $E[T_A]$ decreases as β increases. That is to say, moderate shortening the sojourn time of an arbitrary customer can be achieved by choosing a higher repair rate. On the other hand, we also observe that the expected sojourn time is larger for 2-stage Erlang distribution as compared to uniform and deterministic distributions. We think in most of the cases, by comparing $E[T_A]$ in terms of three different distributions of the inter-batch arrival time usually yields $E[T_A]_{E_2} > E[T_A]_U > E[T_A]_D$. This is due to the fact that the standard deviation of the deterministic distribution equals zero. In Figure 1(b), under the same parameter settings, it can be seen that the average queue length in the case of 2-stage Erlang inter-batch arrival time distribution with a higher standard deviation is larger than the one with lower standard deviations. Again deterministic distribution with zero standard deviation yields the lowest average queue length. Such results indicate that the inter-batch arrival time plays a significant role in determining the performance of the queueing system.

Example 5 In this example, we perform a comparative analysis on $E[T_A]$ and L_s based on changes in specific random arrival batch size distributions and assume that the inter-batch arrival times have a common 3-phase hyperexponential distribution, whose density and Laplace transform are given by

$$a(t) = 0.2 \cdot \frac{1}{10} e^{-\frac{1}{10}t} + 0.5 \cdot \frac{1}{5} e^{-\frac{1}{5}t} + 0.3 \cdot \frac{2}{5} e^{-\frac{2}{5}t}, \quad t > 0,$$

$$a^*(s) = \left(\frac{0.2 \times 0.1}{0.1 + s} \right) + \left(\frac{0.5 \times 0.2}{0.2 + s} \right) + \left(\frac{0.3 \times 0.4}{0.4 + s} \right).$$

A k -phase hyperexponential distribution is frequently used in queueing theory to model the distribution of the superposition of k independent events. For instance, the arrival of different types of customer to a single queueing station is often modeled as a hyperexponential distribution. To demonstrate the performance evaluation, we consider the following three batch size distributions:

Case 1: Batch size X follows the discrete uniform distribution on the interval $[1, 2, \dots, 16]$.

Case 2: The batch size distribution has the probability mass function: $g_4 = 0.25$, $g_8 = 0.25$, $g_{10} = 0.25$, and $g_{12} = 0.25$.

Case 3: The arriving batch size follows a 6-8-12 distribution with probability mass function $g_6 = 0.25$, $g_8 = 0.5$, and $g_{12} = 0.25$.

All these distributions are characterized by the average batch size equal to 8.5. However, these are qualitatively different in that they have different variances, the variances of above probability distributions are 21.25, 8.75 and 4.75, respectively. The numerical experiment has been done by fixing other system

Table 4 The roots of the characteristic equations with modulus less than one in Example 2

| j | r_j | j | r_j | j | ω_j | j | ω_j |
|-----|-----------------------|-----|-----------------------|-----|-----------------------|-----|-----------------------|
| 1 | 0.706120 | 7 | 0.157883 + 0.199331i | 1 | 0.939953 | 7 | -0.144906 + 0.215858i |
| 2 | 0.104655 - 0.252812i | 8 | -0.140320 - 0.209351i | 2 | 0.105721 - 0.264803i | 8 | -0.226770 - 0.120915i |
| 3 | 0.104655 + 0.252812i | 9 | -0.140320 + 0.209351i | 3 | 0.105721 + 0.264803i | 9 | -0.226770 + 0.120915i |
| 4 | -0.025770 - 0.255885i | 10 | -0.219786 - 0.117292i | 4 | -0.026809 - 0.264007i | 10 | -0.256024 |
| 5 | -0.025770 + 0.255885i | 11 | -0.219786 + 0.117292i | 5 | -0.026809 + 0.264007i | 11 | 0.154847 - 0.200497i |
| 6 | 0.157883 - 0.199331i | 12 | -0.248186 | 6 | -0.144906 - 0.215858i | 12 | 0.154847 + 0.200497i |

Table 5 The values of ϕ_j and k_j for $j = 1, 2, \dots, 12$ in Example 2

| j | ϕ_j | j | ϕ_j | j | k_j | j | k_j |
|-----|-----------------------|-----|-----------------------|-----|----------------------|-----|-----------------------|
| 1 | 0.030493 | 7 | -0.000330 + 0.002237i | 1 | 0.003538 | 7 | 0.000224 - 0.000005i |
| 2 | 0.006432 - 0.001180i | 8 | 0.004912 + 0.000029i | 2 | 0.000278 - 0.000086i | 8 | 0.000227 + 0.000003i |
| 3 | 0.006432 + 0.001180i | 9 | 0.004912 - 0.000029i | 3 | 0.000278 + 0.000086i | 9 | 0.000227 - 0.000003i |
| 4 | 0.004860 - 0.000060i | 10 | 0.004955 + 0.000017i | 4 | 0.000221 + 0.000009i | 10 | 0.000228 |
| 5 | 0.004860 + 0.000060i | 11 | 0.004955 - 0.000017i | 5 | 0.000221 - 0.000009i | 11 | -0.000025 - 0.000066i |
| 6 | -0.000330 - 0.002237i | 12 | 0.004968 | 6 | 0.000224 + 0.000005i | 12 | -0.000025 + 0.000066i |

Table 6 Steady-state queue-length distributions at pre-arrival and arbitrary epochs in Example 2

| n | $Q_{n,0}^-$ | $Q_{n,1}^-$ | $Q_{n,0}$ | $Q_{n,1}$ |
|-----|-------------|-------------|-----------|-----------|
| 0 | 0.385597 | 0.019724 | 0.280157 | 0.013109 |
| 1 | 0.081461 | 0.011536 | 0.101078 | 0.010511 |
| 2 | 0.069503 | 0.012420 | 0.091017 | 0.012049 |
| 3 | 0.053002 | 0.012565 | 0.070927 | 0.012557 |
| 4 | 0.038103 | 0.012330 | 0.051284 | 0.012512 |
| 5 | 0.026865 | 0.011933 | 0.036161 | 0.012221 |
| 6 | 0.018913 | 0.011458 | 0.025438 | 0.011813 |
| 7 | 0.013342 | 0.010943 | 0.017939 | 0.011336 |
| 8 | 0.009421 | 0.010409 | 0.012668 | 0.010820 |
| 9 | 0.006654 | 0.009870 | 0.008947 | 0.010286 |
| 10 | 0.004699 | 0.009339 | 0.006318 | 0.009751 |
| 11 | 0.003318 | 0.008821 | 0.004461 | 0.009223 |
| 12 | 0.002343 | 0.008321 | 0.003150 | 0.008710 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 70 | 0.000000 | 0.000232 | 0.000000 | 0.000243 |
| 71 | 0.000000 | 0.000218 | 0.000000 | 0.000229 |
| 72 | 0.000000 | 0.000205 | 0.000000 | 0.000215 |
| 73 | 0.000000 | 0.000193 | 0.000000 | 0.000202 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| sum | 0.718849 | 0.281151 | 0.717115 | 0.282885 |

$$E[T_A] = 13.797044, E[T_A]_{\text{Little}} = 13.797044$$

Table 7 The roots of the characteristic equations with modulus less than one in Example 3

| j | r_j | j | r_j | j | ω_j | j | ω_j |
|-----|-----------------------|-----|-----------------------|-----|-----------------------|-----|-----------------------|
| 1 | 0.718566 | 5 | -0.067063 + 0.382496i | 1 | 0.903423 | 5 | -0.060858 + 0.435105i |
| 2 | 0.240016 - 0.397509i | 6 | -0.278315 - 0.234259i | 2 | 0.284431 - 0.440947i | 6 | -0.423080 |
| 3 | 0.240016 + 0.397509i | 7 | -0.278315 + 0.234259i | 3 | 0.284431 + 0.440947i | 7 | -0.315605 - 0.277986i |
| 4 | -0.067063 - 0.382496i | 8 | -0.359817 | 4 | -0.060858 - 0.435105i | 8 | -0.315605 + 0.277986i |

Table 8 The values of ϕ_j and k_j for $j = 1, 2, \dots, 8$ in Example 3

| j | ϕ_j | j | ϕ_j | j | k_j | j | k_j |
|-----|----------------------|-----|----------------------|-----|----------------------|-----|----------------------|
| 1 | 0.030899 | 5 | 0.009947 + 0.000445i | 1 | 0.003109 | 5 | 0.000332 - 0.000013i |
| 2 | 0.011785 - 0.001844i | 6 | 0.009897 - 0.000170i | 2 | 0.000430 - 0.000026i | 6 | 0.000368 |
| 3 | 0.011785 + 0.001844i | 7 | 0.009897 + 0.000170i | 3 | 0.000430 + 0.000026i | 7 | 0.000340 + 0.000005i |
| 4 | 0.009947 - 0.000445i | 8 | 0.010227 | 4 | 0.000332 + 0.000013i | 8 | 0.000340 - 0.000005i |

Table 9 Steady-state queue-length distributions at pre-arrival and arbitrary epochs in Example 3

| n | $Q_{n,0}^-$ | $Q_{n,1}^-$ | $Q_{n,0}$ | $Q_{n,1}$ |
|-----|-------------|-------------|-----------|-----------|
| 0 | 0.521915 | 0.018047 | 0.293656 | 0.009315 |
| 1 | 0.077257 | 0.009211 | 0.128732 | 0.008344 |
| 2 | 0.059618 | 0.008867 | 0.115829 | 0.009375 |
| 3 | 0.048505 | 0.008618 | 0.061027 | 0.008072 |
| 4 | 0.039700 | 0.008460 | 0.059961 | 0.008416 |
| 5 | 0.030600 | 0.008109 | 0.046542 | 0.008189 |
| 6 | 0.022298 | 0.007609 | 0.043862 | 0.008250 |
| 7 | 0.015623 | 0.007020 | 0.028160 | 0.007620 |
| 8 | 0.010897 | 0.006437 | 0.023791 | 0.007263 |
| 9 | 0.007735 | 0.005883 | 0.013819 | 0.006578 |
| 10 | 0.005647 | 0.005384 | 0.010219 | 0.006064 |
| 11 | 0.004088 | 0.004914 | 0.007521 | 0.005570 |
| 12 | 0.002939 | 0.004473 | 0.005485 | 0.005099 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 70 | 0.000000 | 0.000013 | 0.000000 | 0.000015 |
| 71 | 0.000000 | 0.000011 | 0.000000 | 0.000013 |
| 72 | 0.000000 | 0.000010 | 0.000000 | 0.000012 |
| 73 | 0.000000 | 0.000009 | 0.000000 | 0.000011 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| Sum | 0.854297 | 0.145703 | 0.852568 | 0.147432 |

$$E[T_A] = 7.346311, E[T_A]_{\text{Little}} = 7.346311$$

parameters as $\mu_0 = 2.5$, $\mu_1 = 2$, $\eta = 0.001$ and $\beta = 0.08$. We observe that $E[T_A]$ and L_s decrease as the variance of the batch size distribution decreases, namely $E[T_A]_{(\text{case } 1)} = 9.384734 > E[T_A]_{(\text{case } 2)} = 8.329050 > E[T_A]_{(\text{case } 3)} = 7.990203$ and $L_{s(\text{case } 1)} = 15.194332 > L_{s(\text{case } 2)} = 13.485130 > L_{s(\text{case } 3)} = 12.936519$. So, we can

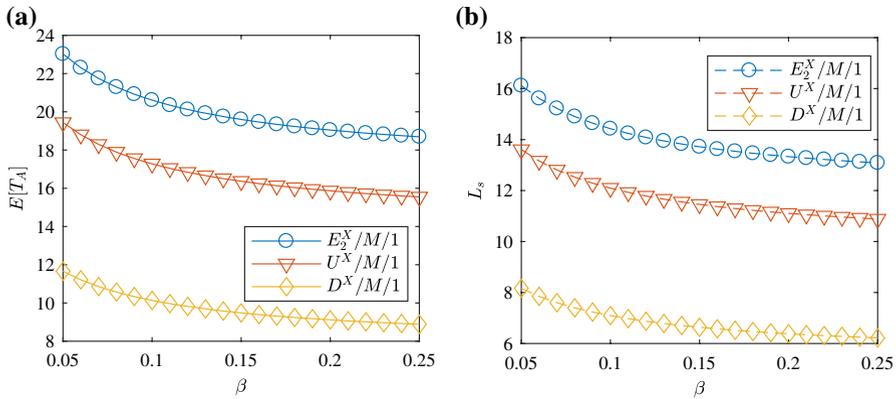


Fig. 1 Effect of β on $E[T_A]$ and L_s for various inter-batch arrival time distributions

conclude that the variance of batch size has obvious impact on performance metrics like average system size and average sojourn time.

7 Conclusions

In order to mitigate the severe potential consequences of system failure, fault-tolerant operating has become a critical attribute of the manufacturing system. In this paper, we have developed a $GI^X/M/1$ queue with a fault-tolerant server to study the behavior of key performance indicators for such a system. The mathematical analysis of the queueing model is carried out by employing a combination of two methods. One is the well-known supplementary variable technique, and the other is the shift operator method in difference equations theory. The most important reason why we adopt the above approach is that we want to avoid discussing the transition probability matrix associated with the embedded Markov chain. Because with the increase in the complexity involved in the queue, the derivation of the transition probability matrix becomes extremely difficult, and many practitioners cannot fully master these complicated and tedious probabilistic arguments. In essence, it is our intention to solve a practical queueing problem in the most simple and pragmatic way using the roots of the model's characteristic equation. From the algorithmic steps and numerical experiments presented in our work, we may also see that the methodology is nicely implemented in flexible scientific computing software that is easy to use even for users who are not professional mathematicians. A feasible plan for further study is to consider the discrete-time counterparts of this queue. Since several possible events may occur simultaneously in discrete-time queues, the analysis and computation will become more elaborate under such a situation.

Acknowledgements The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of this paper. This research was supported by the National

Natural Science Foundation of China (Grant No. 71571127), and the National Office for Philosophy and Social Sciences of P. R. China (Grant No. 20BGL109).

References

- Avi-Itzhak B, Naor P (1963) Some queueing problems with the service station subject to breakdowns. *Oper. Res.* 11:303–320. <https://doi.org/10.1287/opre.11.3.303>
- Barbhuiya FP, Gupta UC (2019a) A difference equation approach for analysing a batch service queue with the batch renewal arrival process. *J. Differ. Equ. Appl.* 25:233–242. <https://doi.org/10.1080/10236198.2019.1567723>
- Barbhuiya FP, Gupta UC (2019b) Discrete-time queue with batch renewal input and random serving capacity rule: $GI^X/Geo^Y/1$. *Queueing Syst.* 91:347–365. <https://doi.org/10.1007/s11134-019-09600-7>
- Barbhuiya FP, Gupta UC (2020) Analytical and computational aspects of the infinite buffer single server N policy queue with batch renewal input. *Comput. Oper. Res.* 118. <https://doi.org/10.1016/j.cor.2020.104916>
- Burke PJ (1975) Delays in single-server queues with batch input. *Oper. Res.* 23:830–832. <https://doi.org/10.1287/opre.23.4.830>
- Chaudhry ML, Harris CM, Marchal WG (1990) Robustness of rootfinding in single-server queueing models. *INFORMS J. Comput.* 2:273–286. <https://doi.org/10.1287/ijoc.2.3.273>
- Chaudhry ML, Templeton JGC (1983) *A First Course in Bulk Queues*. Wiley, New York
- Chen WL (2018) System reliability analysis of retrial machine repair systems with warm standbys and a single server of working breakdown and recovery policy. *Syst. Eng.* 21:59–69. <https://doi.org/10.1002/sys.21420>
- Choudhury G, Deka M (2012) A single server queueing system with two phases of service subject to server breakdown and Bernoulli vacation. *Appl. Math. Model.* 36:6050–6060. <https://doi.org/10.1016/j.apm.2012.01.047>
- Choudhury G, Deka M (2018) A batch arrival unreliable server delaying repair queue with two phases of service and Bernoulli vacation under multiple vacation policy. *Qual. Technol. Quant. Manag.* 15:157–186. <https://doi.org/10.1080/16843703.2016.1208934>
- Choudhury G, Kalita CR (2018) An M/G/1 queue with two types of general heterogeneous service and optional repeated service subject to server's breakdown and delayed repair. *Qual. Technol. Quant. Manag.* 15:622–654. <https://doi.org/10.1080/16843703.2017.1331499>
- Choudhury G, Tadj L (2009) An M/G/1 queue with two phases of service subject to the server breakdown and delayed repair. *Appl. Math. Model.* 33:2699–2709. <https://doi.org/10.1016/j.apm.2008.08.006>
- Choudhury G, Tadj L (2011) The optimal control of an $M^X/G/1$ unreliable server queue with two phases of service and Bernoulli vacation schedule. *Math. Comput. Model.* 54:673–688. <https://doi.org/10.1016/j.mcm.2011.03.010>
- Cox DR (1955) The analysis of non-Markovian stochastic processes by the inclusion of supplementary variables. *Math. Proc. Camb. Philos. Soc.* 51:433–441. <https://doi.org/10.1017/S0305004100030437>
- Elaydi S (2005) *An Introduction to Difference Equations*. Springer, New York
- Gao S, Wang JT, Do TV (2019) Analysis of a discrete-time repairable queue with disasters and working breakdowns. *RAIRO-Oper. Res.* 53:1197–1216. <https://doi.org/10.1051/ro/2018057>
- Gaver DP (1962) A waiting line with interrupted service including priorities. *J. R. Stat. Soc. Ser. B* 24:73–90. <https://doi.org/10.1111/j.2517-6161.1962.tb00438.x>
- Gorbanova A, Lebedev A (2020) Queueing system with two input flows, preemptive priority, and stochastic dropping. *Autom. Remote Control* 81:2230–2243. <https://doi.org/10.1134/S0005117920120073>
- Gray WJ, Wang PP, Scott M (2004) A queueing model with multiple types of server breakdowns. *Qual. Technol. Quant. Manag.* 2:245–255. <https://doi.org/10.1080/16843703.2004.11673076>
- Gross D, Harris CM (1985) *Fundamentals of Queueing Theory*, 2nd edn. Wiley, New York
- Jain M, Kaur S, Singh P (2019) Supplementary variable technique (SVT) for non-Markovian single server queue with service interruption (QSI). *Oper. Res.* <https://doi.org/10.1007/s12351-019-00519-8>

- Jain M, Kaur S (2020) (p, N)-Policy for unreliable server bulk queue with Bernoulli feedback. *Int. J. Appl. Comput. Math.* 6:170. <https://doi.org/10.1007/s40819-020-00912-4>
- Jain M, Sharma GC, Sharma R (2013) Unreliable server M/G/1 queue with multi-optional services and multi-optional vacations. *Int. J. Math. Oper. Res.* 5:145–169. <https://doi.org/10.1504/IJMOR.2013.052458>
- Jain M, Sharma R, Meena RK (2019) Performance modeling of fault-tolerant machining system with working vacation and working breakdown. *Arab. J. Sci. Eng.* 44:2825–2836. <https://doi.org/10.1007/s13369-018-3591-6>
- Jiang T, Liu LW (2017) The GI/M/1 queue in a multi-phase service environment with disasters and working breakdowns. *Int. J. Comput. Math.* 94:707–726. <https://doi.org/10.1080/00207160.2015.1128531>
- Jiang T, Xin BG (2019) Computational analysis of the queue with working breakdowns and delaying repair under a Bernoulli-schedule-controlled policy. *Commun. Stat.-Theory Methods* 48:926–941. <https://doi.org/10.1080/03610926.2017.1422756>
- Kalidass K, Kasturi R (2012) A queue with working breakdowns. *Comput. Ind. Eng.* 63:779–783. <https://doi.org/10.1016/j.cie.2012.04.018>
- Ke JC (2003) The optimal control of an M/G/1 queueing system with server vacations, startup and breakdowns. *Comput. Ind. Eng.* 44:567–579. [https://doi.org/10.1016/S0360-8352\(02\)00235-8](https://doi.org/10.1016/S0360-8352(02)00235-8)
- Ke JC (2006) An M/G/1 queue under hysteretic vacation policy with an early startup and un-reliable server. *Math. Methods Oper. Res.* 63:357–369. <https://doi.org/10.1007/s00186-005-0046-0>
- Ke JC, Huang KB (2010) Analysis of an unreliable server $M^{[X]}/G/1$ system with a randomized vacation policy and delayed repair. *Stoch. Models* 26:212–241. <https://doi.org/10.1080/15326341003756262>
- Ke JC, Huang KB (2012) Analysis of batch arrival queue with randomized vacation policy and an unreliable server. *J. Syst. Sci. Complex.* 25:759–777. <https://doi.org/10.1007/s11424-012-9154-0>
- Ke JC, Lin CH (2006) Maximum entropy solutions for batch arrival queue with an un-reliable server and delaying vacations. *Appl. Math. Comput.* 183:1328–1340. <https://doi.org/10.1016/j.amc.2006.05.174>
- Kim BK, Lee DH (2014) The M/G/1 queue with disasters and working breakdowns. *Appl. Math. Model.* 38:1788–1798. <https://doi.org/10.1016/j.apm.2013.09.016>
- Klimenok V (2001) On the modification of Rouche's theorem for the queueing theory problems. *Queueing Syst.* 38:431–434. <https://doi.org/10.1023/A:1010999928701>
- Kumar BK, Sankar R, Krishnan RN, Rukmani R (2021) Performance analysis of multi-processor two-stage tandem call center retrieval queues with non-reliable processors. *Methodol Comput Appl Probab.* <https://doi.org/10.1007/s11009-020-09842-6>
- Kumar MS, Dadlani A, Kim K (2020) Performance analysis of an unreliable M/G/1 retrieval queue with two-way communication. *Oper. Res.* 20:2267–2280. <https://doi.org/10.1007/s12351-018-0417-y>
- Li W, Shi DH, Chao XL (1997) Reliability analysis of M/G/1 queueing systems with server breakdowns and vacations. *J. Appl. Probab.* 34:546–555. <https://doi.org/10.2307/3215393>
- Li L, Wang JT, Zhang F (2013) Equilibrium customer strategies in Markovian queues with partial breakdowns. *Comput. Ind. Eng.* 66:751–757. <https://doi.org/10.1016/j.cie.2013.09.023>
- Liou CD (2015) Markovian queue optimisation analysis with an unreliable server subject to working breakdowns and impatient customers. *Int. J. Syst. Sci.* 46:2165–2182. <https://doi.org/10.1080/0020721.2013.859326>
- Liu ZM, Song Y (2014) The $M^X/M/1$ queue with working breakdown. *RAIRO-Oper. Res.* 48:399–413. <https://doi.org/10.1051/ro/2014014>
- Madan KC (2003) An M/G/1 type queue with time-homogeneous breakdowns and deterministic repair times. *Soochow J. Math.* 29:103–110
- Mitrany IL, Avi-Itzhak B (1968) A many-server queue with service interruptions. *Oper. Res.* 16:628–638. <https://doi.org/10.1287/opre.16.3.628>
- Neuts MF (1981) *Matrix-Geometric Solutions in Stochastic Models*. Johns Hopkins University Press, Baltimore
- Neuts MF, Lucantoni DM (1979) A Markovian queue with N servers subject to breakdowns and repairs. *Manag. Sci.* 25:849–861. <https://doi.org/10.1287/mnsc.25.9.849>
- Saggou H, Sadeg I, Ourbih-Tari M, Bourennane EB (2019) The analysis of unreliable $M^{[X]}/G/1$ queueing system with loss, vacation and two delays of verification. *Commun. Stat.-Simul. Comput.* 48:1366–1381. <https://doi.org/10.1080/03610918.2017.1414245>
- Sengupta B (1990) A queue with service interruptions in an alternating random environment. *Oper. Res.* 38:308–318. <https://doi.org/10.1287/opre.38.2.308>

- Singh CJ, Jain M, Kaur S (2018) Performance analysis of bulk arrival queue with balking, optional service, delayed repair and multi-phase repair. *Ain Shams Eng. J.* 9:2067–2077. <https://doi.org/10.1016/j.asej.2016.08.025>
- Singh G, Gupta UC, Chaudhry ML (2014) Analysis of queueing time distributions for MAP/D_N/1 queue. *Int. J. Comput. Math.* 91:1911–1930. <https://doi.org/10.1080/00207160.2013.867021>
- Takine T, Sengupta B (1997) A single server queue with service interruptions. *Queueing Syst.* 26:285–300. <https://doi.org/10.1023/A:1019189326131>
- Tang YH (1997) A single server M/G/1 queueing system subject to breakdowns-some reliability and queueing problems. *Microelectron. Reliab.* 37:315–321. [https://doi.org/10.1016/S0026-2714\(96\)00018-2](https://doi.org/10.1016/S0026-2714(96)00018-2)
- Thiruvengadam K (1963) Queueing with breakdowns. *Oper. Res.* 11:62–71. <https://doi.org/10.1287/opre.11.1.62>
- Tijms HC (2003) *A First Course in Stochastic Models*. Wiley, Chichester
- Wang JT (2004) An M/G/1 queue with second optional service and server breakdowns. *Comput. Math. Appl.* 47:1713–1723. <https://doi.org/10.1016/j.camwa.2004.06.024>
- Wang KH, Yang DY, Pearn WL (2010) Comparison of two randomized policy M/G/1 queues with second optional service, server breakdown and startup. *J. Comput. Appl. Math.* 234:812–824. <https://doi.org/10.1016/j.cam.2010.01.045>
- Yang DY, Cho YC (2019) Analysis of the N-Policy GI/M/1/K queueing systems with working breakdowns and repair. *Comput. J.* 62:130–143. <https://doi.org/10.1093/comjnl/bxy051>
- Ye QQ, Liu LW (2018) Analysis of MAP/M/1 queue with working breakdowns. *Commun. Stat.-Theory Methods* 47:3073–3084. <https://doi.org/10.1080/03610926.2017.1346808>
- Yen TC, Chen WL, Chen JY (2016) Reliability and sensitivity analysis of the controllable repair system with warm standbys and working breakdown. *Comput. Ind. Eng.* 97:84–92. <https://doi.org/10.1016/j.cie.2016.04.019>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.