European Journal of Operational Research 261 (2017) 613–625

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Stochastics and Statistics

A control-chart-based queueing approach for service facility maintenance with energy-delay tradeoff

Wenhui Zhou, Zhibin Zheng*, Wei Xie

School of Business Administration, South China University of Technology, Guangzhou 510000, Guangdong, China

ARTICLE INFO

Article history:
Received 5 May 2016
Accepted 13 March 2017
Available online 18 March 2017

Keywords:
OR in energy
Energy-delay tradeoff
Queueing system
Maintenance
Control chart

ABSTRACT

Maintenance planning and energy consumption control are critical issues in facility operations management. In practice, the energy consumption of a facility, which will be affected by the operation condition, is closely connected with the associated maintenance policy. Specifically, for an energy-consuming service system, though a frequent maintenance activity can keep the facility in a good condition with low energy consumption, it makes the delay time longer and leads to a poor customer experience. In this paper, we study a single-server queueing system with different energy consumption levels in the associated running states to address the conflict between energy consumption and customer delay. Two types of maintenance activities are implemented for the server, i.e., the planned maintenance and the reactive maintenance. The planned maintenance is adopted based on a frequency parameter at the beginning of an idle period, and the reactive maintenance is initialized by the Shewhart's individual control chart (condition-based maintenance). To capture the energy-delay tradeoff, our objective is to develop an optimal maintenance policy that minimizes the long-run expected total cost of the system under a customer waiting time constraint. Numerical experiments are conducted to analyze the problem, in which useful managerial insights are obtained for the optimal maintenance policy. The results demonstrate the robustness of the proposed maintenance model, its advantage over the model without control chart, and its applicability in general situations.

1. Introduction

Carbon emission has been one of the most critical environmental issues for years. The governments successively roll out the related policies for enterprises to achieve possible improvement. As a response, the enterprises start to recognize the importance and urgency of energy conservation and emission reduction. Service industry, without exception, is among this revolution-required group, especially the high energy consumption firms. For instance, for large server farms, the energy expense is one of the dominant operating costs. According to Heo, Henriksson, Liu, and Abdelzaher (2007), around 23%–50% of the revenue is spent on the power sources for running the systems. For the major commercial airlines, the statistics recorded by IATA reveal that the expenditure on fuel accounted for 25.3%–36.7% of the operating costs in 2008. Besides the energy consumption part, maintenance is another vital issue for these service firms, because the services are provided by large and expensive facilities. In practice, energy consumption and maintenance are two interconnected elements in facility management. The reason is that the energy consumption level of the facility highly depends on its operating condition, which will change based on the associated degradation process and maintenance activity. Generally, the energy consumption levels of abnormal states will be higher than that of a normal state (Alsyouf, 2006; Ang & Fwa, 1989). For examples, a worn bearing will increase the energy consumption level by aggravating the rolling friction in mechanical equipment, while, for electronic devices, the malfunction of cooling system will lead to a high-temperature working condition which also raises the energy consumption to a higher level. Therefore, a properly scheduled maintenance plays an important role in improving the system performance as well as reducing the energy consumption.

As a major component in operations management, maintenance has been extensively studied in the literature. Traditional maintenance models only focus on the "technical" state of the system, e.g., machine failure mechanisms, which typically do not include some important characteristics of the "operating" state, such as customers, workload, inventory, etc. (Kaufman & Lewis, 2007). However, for a service system, the servers are facing the customers directly. Because the customer satisfaction is an important factor,

* Corresponding author.
E-mail address: zhibzheng@gmail.com (Z. Zheng).

http://dx.doi.org/10.1016/j.ejor.2017.03.026
0377-2217/

which can be reflected by the customer delay, for the service firms, they often compete on the customer waiting time (Allon & Federgruen, 2007). Hence, when making the maintenance decisions, customer delay is treated as a crucial operational index of the service system. In previous research, by considering the customer holding cost, a group of scholars have discussed the queuing systems with repairable servers. The objective of their model is to find a corrective maintenance policy that minimizes the long-run average operating costs (customer holding cost is included). Lam, Zhang, and Liu (2006) present a geometric process model for M/M/1 queueing system with a repairable service station, in which the replacement policy for the station is optimized. Kaufman and Lewis (2007) model a single-server queue via a semi-Markov decision process. Both repair and replacement models are adopted in their study. When the server has a deteriorating property, Yang, Lim, and Chae (2009) introduce the maintenance policy for the single-server queues with random shocks. Recently, Xie, Liao, and Jin (2014) propose a queueing system to characterize the repair-by-replacement action for a modular equipment to address the redundancy allocation and spare parts inventory issues. Taleb and Aissani (2016) consider both corrective maintenance and preventive maintenance in an unreliable retrial queue with persistent and impatient customers. However, most of the above-mentioned research assumes the exponential arrival/service time, which may not be applied to the cases with general distributions. To address this issue, anther stream of research is conducted to study the maintenance problem in more complex service systems. For instances, Federgruen and So (1990) consider a single-server queueing system with Poisson arrivals and general service time. Wartenhorst (1995) presents an exact (matrix-geometric) solution and a simple approximation (stochastic decomposition) to study a multi-server and multi-repairman system. Li, Ying, and Zhao (2006) consider a BMAP/G/1 retrial queue with a repairable server, where the server's life time is exponentially distributed and the repair time is general. Delia and Rafael (2008) present two types of repairs (depending on the system's deteriorating level), i.e., minimal and perfect repairs. In their model, the duration times of different repairs follow different Phase-Type (PH) distributions, and the failures and inspections are characterized by different Markovian arrival processes (MAP). Montoro-Cazorla, Pérez-Ocón, and del Carmen Segovia (2009) examine the replacement policy for a system suffered the MAP shocks. Due to the high-dimension state space and complexity of the system, their results are obtained by applying the Matrix-analytic method, which is first introduced and studied by Neuts (1981). This method can be employed for steady state analysis of a certain class of continuous time Markov processes and is still being developed and improved to solve the quasi-birth-death (QBD) problems for different queuing systems, such as PH queues (Latouche & Ramaswami, 1999), Markovian queues with marked transition (He & Neuts, 1998), multi-server retrial queues (Artalejo, Gómez-Corral, & Neuts, 2001), fluid queues (Dzial, Breuer, Soares, Latouche, & Remiche, 2005 and Soares & Ana, 2006), etc.

In aforementioned maintenance models, the maintenance action is carried out by schedule or when the system stops or fails. While, in some systems, the equipment may gradually deteriorate and enter intermediate operating states prior to failure. If such intermediate states can be, directly or indirectly, identified, a more efficient maintenance method called condition-based maintenance (CBM) can be adopted to handle this situation. Under the CBM, maintenance action is called by monitoring the operating condition of the facility. The literature of CBM is vast. Rahim (1994) jointly optimizes the parameters of the $\overline{X}$ control chart and the inspection schedule for an imperfect production system. Ben-Daya and Rahim (2000) attempt to integrate the preventive maintenance and $\overline{X}$ control chart, where the in-control period follows a general probability distribution with an increasing hazard rate. Linderman, McKone-Sweet, and Anderson (2005) propose a generalized model to coordinate the statistical process control and the planned maintenance. Carnero (2005) introduces a decision-making model based on the selection of diagnostic techniques and instrumentations for the predictive maintenance programs. In addition, Zhou and Zhu (2008) develop an integrated model of control chart and maintenance management, which optimally determines four policy variables and minimizes the hourly costs. Bana e Costa, Carnero, and Duarte (2012) develop a multi-criteria model to audit a predictive maintenance program that is implemented in the General Hospital of Ciudad Real in Spain. Liu, Yu, Ma, and Tu (2013) apply CBM to study the $\overline{X}$ control chart for the series systems with two identical units. Recently, Yin, Zhang, Zhu, Deng, and He (2015) proposed an integrated model of statistical process control and maintenance to study the delayed monitoring policy and derive an economic model. Peng and van Houtum (2016) combined CBM and economic manufacturing quantity to evaluate the average long-run cost rate of a degrading manufacturing system via the renewal theory. Keizer, Teunter and Veldman (2017) considered the joint optimization of the CBM and spares planning for multi-component systems by using Markov decision process. The interested readers are referred to Ben-Daya (1999), Cassady, Bowden, Liew, and Pohl (2000), Kuo (2006), Panagiotidou and (2007), Panagiotidou and Nenes (2009), Pandey, Kulkarni, and Vrat (2010), Wang (2012), Liu et al. (2013) and Jafari and Makis (2016). Essentially, almost all of the CBM-related research is focused on the manufacturing industry and employs the production quality as the intermediate to monitor the equipment condition and schedule the maintenance. However, it is not appropriate and feasible to examine the same characteristics for the service systems, because the service quality of the facility is difficult to measure and cannot reflect the associated operating conditions. As an alternative, the energy consumption level of the server is a good candidate as it is tightly linked with the operating states. Since the energy consumption level is higher in the abnormal states, it is reasonable to utilize the state of energy consumption as the intermediate to monitor the server's operating condition.

It is obvious that there is a clear tradeoff between energy consumption and customer delay in the high-energy-consumption service systems (Gandhi, Gupta, Harchol-Balter, & Kozuch, 2010). For example, a frequent maintenance action can keep the facility in a good condition with low energy consumption, but the customer waiting time will be increased. This will have a negative impact on the customer satisfaction, especially for the time-sensitive customers. Thus, in the service system, it is not sufficient to only consider the corresponding energy costs. Instead, a more appropriate maintenance policy should be implemented to take the energy-delay tradeoff into consideration. However, previous research mainly considers the energy-delay tradeoff in the computer applications (Gonzalez & Horowitz, 1996; Juang, Wu, Peh, Martonosi, & Clark, 2005; Kang, Abbaspour, & Pedram, 2003; Kin, Gupta; Stan & Skadron, 2003), most of the authors employ the metric of Energy-Response time Product (ERP), also known as Energy-Delay Product, to capture the tradeoff between energy and delay time. Specifically, for a policy π, the ERP is given by $ERP^{\pi} = E[P^{\pi}] \cdot E[T^{\pi}]$, where $E[P^{\pi}]$ is the long-run average power consumed and $E[T^{\pi}]$ is the mean customer delay time in the system under the control policy π (Gandhi et al., 2010). However, practically, it is difficult for the server providers to eliminate the customer delay because of the associated marginal cost. Most of the servers can only promise to complete the service within a period of time, e.g., the examination in a hospital. This service time can be used to represent the service level of the firm. Because the ERP does not include the maintenance cost, we adopt the service level, defined by the delay time, as a constraint to construct the associated maintenance policy.

In this paper, a single-server queueing system with different energy consumption levels is analyzed. To avoid the high energy consumption in an abnormal state, two types of the maintenance activities, i.e., the planned maintenance and the reactive maintenance, are applied to improve the availability and performance of the server. We combine queueing theory and control chart to model the customer delay time and the CBM. The planned maintenance is implemented at the beginning of the idle period with a given frequency, and the reactive maintenance is initialized by the Shewhart's individual control chart, which monitors the energy consumption level and will generate the alert signals when the system shifts to the out-of-control state. Finally, to capture the tradeoff between energy consumption and customer delay, we develop an optimal maintenance policy to minimize the long-run expected cost of the system under the service level constraint. Due to the complexity in investigating the optimal policy analytically, we conduct numerical experiments to demonstrate the performance of our model. The result shows that the optimal policy becomes sensitive when the service level increases (with a more strict the sojourn time requirement), and the policy-changing trend provide useful reference to the decision maker. In addition, compared to the system without individual control, the proposed model shows its advantage in cost saving. Finally, the applicability of our model is validated in general cases.

The rest of the paper is organized as follows. Section 2 briefly describes the problem setting. The proposed problem, including both special and general cases, is analyzed in Section 3. Section 4 develops an optimization model to obtain the optimal maintenance policy. Numerical experiments are conducted to analyze the optimal policy in Section 5. Finally, Section 6 concludes the work.

2. Model description

Consider a single service facility (server), e.g., a server farm, which takes care of a customer with a random service time. Customer arrival process is also random and the customers will wait in a single queue if the server is busy. The facility requires power to serve the customers, and the energy consumption per unit time e, called marginal energy consumption, can be measured and monitored right after the epoch of service completion. In practice, the energy consumption will be affected by the condition of the server, the operation of the staff, the working environment, etc. For instance, the efficiency and fuel consumption of a bus are highly dependent on the vehicle and engine attributes, the passenger load, the travel speed, the number of stops and the road grade (Ang & Fwa, 1989). Thus, due to the randomness of energy consumption, we assume that the marginal energy consumption e follows the normal distribution. In particular, we assume that the mean energy consumption level of the server varies in different operating conditions (states) (Kaufman & Lewis, 2007). Two types of states are considered, i.e., low energy consumption state 0 and high energy consumption state 1 (more condition states can be considered in general). When the server is in state i ($i = 0, 1$), e follows normal distribution with mean e_i and variance σ^2, i.e., $e \sim N(e_i, \sigma^2)$, where $e_0 < e_1$. To avoid the high setup cost and keep the response speed, the server is assumed to be on standby during the idle time. Because the energy consumption in an idle state is lower than that in an operating state, when the server is idle in state i ($i = 0, 1$), we assume the marginal energy consumption e^I follows normal distribution with mean e_i^I and variance σ^2. i.e., $e^I \sim N(e_i^I, \sigma^2)$. Thus, it reasonable to have $e_0^I < e_1^I < e_1$ and $e_0^I < e_0$.

While the server consumes more energy in state 1, to reduce the energy consumption, maintenance action can be implemented to keep the server in state 0. Usually, the switch of the server's state does not have the self-announcing property, thus, a process monitoring tool is needed to detect the state change. We adopt the Shewhart's individual control chart shown in Fig. 1, where the central line (CL) is e_0, the upper control limit (UCL) is $e_0 + k\sigma$ and the lower control limit (LCL) is $e_0 - k\sigma$, respectively. Hence, the individual control chart in Fig. 1 is divided into two regions, i.e., I_1: $e \in (LCL, UCL)$ (in-control region) and $I_2 : e \in (-\infty, LCL] \cup [UCL, +\infty)$ (signal region). The marginal energy consumption e will be measured when the service is completed. If e drops in region I_1, then no action is needed for the server and the first customer in the queue is served immediately. Otherwise, the control chart will signal the "out-of-control" message. As a result, the reactive maintenance (RM) will be triggered by the "out-of-control" signal. Assume the RM is perfect, regardless of the previous state, the server will be restored to state 0. Let α_i be the probability that, given the server is in state i, the control chart signals the "out-of-control" message, one has

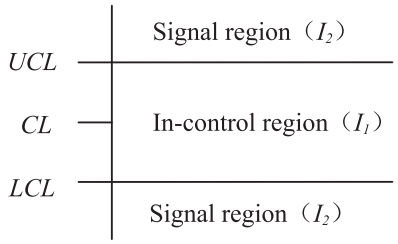

Fig. 1. Individual control chart

$$\begin{aligned}\alpha_0 &= P\{e \in I_2 | i = 0\} \\ &= P\{e \geq e_0 + k\sigma | i = 0\} + P\{e \leq e_0 - k\sigma | i = 0\} \\ &= P\left\{\frac{e - e_0}{\sigma} \geq k | i = 0\right\} + P\left\{\frac{e - e_0}{\sigma} \leq -k | i = 0\right\} \\ &= 1 - \Phi(k) + \Phi(-k),\end{aligned}$$

and

$$\begin{aligned}\alpha_1 &= P\{e \in I_2 | i = 1\} \\ &= P\left\{\frac{e - e_1}{\sigma} \geq \frac{e_0 - e_1 + k\sigma}{\sigma} | i = 1\right\} \\ &\quad + P\left\{\frac{e - e_1}{\sigma} \geq \frac{e_0 - e_1 - k\sigma}{\sigma} | i = 1\right\} \\ &= 1 - \Phi\left(\frac{e_0 - e_1 + k\sigma}{\sigma}\right) + \Phi\left(\frac{e_0 - e_1 - k\sigma}{\sigma}\right),\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

Besides the RM, planned maintenance (PM) policy is also adopted to guarantee the system health. However, in real-world situation, to save the maintenance costs, the PM may not be scheduled for every idle period. The frequency of PM is determined by the decision maker, which can be represented by a proportion parameter p_m, e.g., $p_m = 0.25$ means that the PM is implemented every 4 idle periods. For convenience, we assume that when the service is completed and no customer is in the waiting line, the PM will be taken with $p_m \times 100\%$ chance. The maintenance management framework for the single-server queueing system is shown in Fig. 2.

Although the maintenance activities can reduce the energy consumption of the server, but the associated PM and RM will deactivate the service facility and make the customer waiting time longer, which will lead to complaints. Therefore, customer delay is a crucial factor in the decision-making process of maintenance. To capture the effect of maintenance policy on the customer waiting time and average marginal energy consumption, we first analyze the performance of the queueing system in the following section.

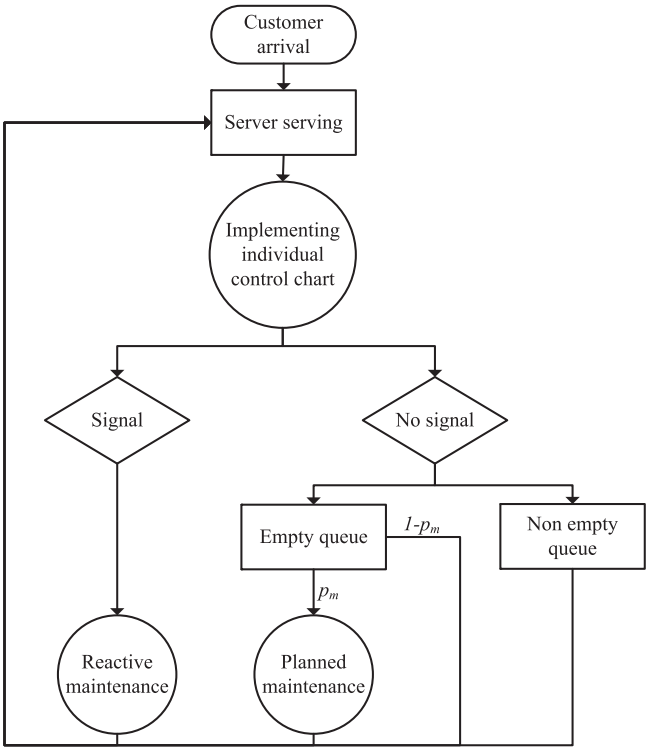

Fig. 2. The framework of maintenance model.

3. Model analysis

3.1. Special case: M/M/1 queueing service system

In this section, to deliver some analytically tractable results for gaining more insights of our problem, we first study energy-delay tradeoff maintenance problem based on an M/M/1 queueing service system, in which the customers are served based on the first-come, first-served (FCFS) discipline. The customer arrival process is a Poisson process with rate λ, and the service time is an independent and identically distributed (*i.i.d*) random variable that following the exponential distribution with rate μ. In addition, the deterioration process of the server is also assumed to be the Poisson process with rate β (i.e., the mean time from state 0 to state 1 is $1/\beta$), and the maintenance time of the system is an *i.i.d* and exponentially distributed random variable with rate γ.

We analyze the steady-state performance of this queueing system by using the probability generating function (PGF) method (the steady-state performance is valuable for many real-world applications, e.g., a server farm). Based on the assumptions, we can construct a two-dimensional continuous-time Markov chain (CTMC) $\{N_1(t), N_2(t), t \geq 0\}$ to describe the state transition of the system. $N_1(t)$ represents the number of customers in the system (including the customers in service), and $N_2(t)$ is the condition state of the server which is defined as

$$N_2(t) = \begin{cases} 0, & \text{low energy consumption state,} \\ 1, & \text{high energy consumption state,} \\ 2, & \text{maintenance state.} \end{cases}$$

Then, the state space of the CTMC is $\{0, 1, 2, 3, \ldots\} \times \{0, 1, 2\}$ and the state transition diagram is presented in Fig. 3. Define $\pi_{i,j}$ ($i \in \{0, 1, 2, \ldots\}$, $j \in \{0, 1, 2\}$) as the steady-state probability of the CTMC. By solving the following balance equations

$$(\lambda+\beta)\pi_{0,0} = (1-p_m)(1-\alpha_0)\mu\pi_{1,0} + \gamma\pi_{0,2}, \tag{1}$$

$$\lambda\pi_{0,1} = \beta\pi_{0,0} + (1-p_m)(1-\alpha_1)\mu\pi_{1,1}, \tag{2}$$

$$(\gamma+\lambda)\pi_{0,2} = [\alpha_0\mu + p_m(1-\alpha_0)\mu]\pi_{1,0} + [\alpha_1\mu + p_m(1-\alpha_1)\mu]\pi_{1,1}, \tag{3}$$

$$(\gamma+\lambda)\pi_{i,2} = \lambda\pi_{i-1,2} + \alpha_0\mu\pi_{i+1,0} + \alpha_1\mu\pi_{i+1,1}, \quad (i \geq 1), \tag{4}$$

$$(\lambda+\beta+\mu)\pi_{i+1,0} = \lambda\pi_{i,0} + (1-\alpha_0)\mu\pi_{i+2,0} + \gamma\pi_{i+1,2}, \quad (i \geq 0), \tag{5}$$

$$(\lambda+\mu)\pi_{i+1,1} = \lambda\pi_{i,1} + (1-\alpha_1)\mu\pi_{i+2,1} + \beta\pi_{i+1,0}, \quad (i \geq 0), \tag{6}$$

we can obtain the steady-state probabilities of the system. Before quantifying the expected waiting time, the ergodic condition of the CTMC should be investigated, which is given by the following lemma.

Lemma 1. *The proposed CTMC $\{N_1(t), N_2(t), t \geq 0\}$ is ergodic if and only if*

$$\frac{\lambda\mu}{\mu+\beta}\left(\frac{1-\alpha_0}{\mu} + \frac{\alpha_0}{\mu} + \frac{\alpha_0}{\gamma}\right) + \frac{\lambda\beta}{\mu+\beta}\left(\frac{1-\alpha_1}{\mu} + \frac{\alpha_1}{\mu} + \frac{\alpha_1}{\gamma}\right) < 1. \tag{7}$$

Proof. Let $\{t_n, n \in N\}$ be the sequence of epochs when the services are completed or the customers leave the system. The sequence of random variables $\{Y_n = N(t_n)\}$ forms an irreducible and aperiodic embedded Markov chain. To prove the sufficient condition, we first define the mean drift $\chi_i = E[f(Y_{n+1}) - f(Y_n) \mid Y_n = i]$. Consider $f(x) = x$ yields

$$\begin{aligned}\chi_i &= E[N(t_{n+1}) - N(t_n) \mid N(t_n) = i] \\ &= \frac{\lambda\mu}{\mu+\beta}\left(\frac{1-\alpha_0}{\mu} + \frac{\alpha_0}{\mu} + \frac{\alpha_0}{\gamma}\right) \\ &\quad + \frac{\lambda\beta}{\mu+\beta}\left(\frac{1-\alpha_1}{\mu} + \frac{\alpha_1}{\mu} + \frac{\alpha_1}{\gamma}\right) - 1.\end{aligned}$$

Clearly, if $\frac{\lambda\mu}{\mu+\beta}(\frac{1-\alpha_0}{\mu} + \frac{\alpha_0}{\mu} + \frac{\alpha_0}{\gamma}) + \frac{\lambda\beta}{\mu+\beta}(\frac{1-\alpha_1}{\mu} + \frac{\alpha_1}{\mu} + \frac{\alpha_1}{\gamma}) < 1$, one has $|\chi_i| < \infty$ for all i and $\lim_{i\to\infty} \sup \chi_i < 0$. Based on the Foster's criterion, the embedded Markov chain $\{Y_n, n \geq 0\}$ is ergodic. This finishes the proof. □

Because the queueing system is stable when the proposed CTMC is ergodic, Lemma 1 actually shows the stability condition of the system. In a broad sense, $(1-\alpha_0)/\mu + \alpha_0/\mu + \alpha_0/\gamma$ and $(1-\alpha_1)/\mu + \alpha_1/\mu + \alpha_1/\gamma$ in Eq. (7) are the service times in states 0 and state 1, respectively. It can be seen that the service time includes the real service time $1/\mu$ and the possible maintenance time $1/\gamma$ (if the control chart sends out the "out-of-control" signal).

Assume that the ergodic condition in Lemma 1 holds, the expected waiting time of the queueing system can be derived by using the probability generating function (PGF). The PGFs for the steady state probability $\pi_{i,j}$ are

$$P_0(z) = \sum_{i=0}^{\infty}\pi_{i,0}z^i, \quad P_1(z) = \sum_{i=0}^{\infty}\pi_{i,1}z^i, \quad P_2(z) = \sum_{i=0}^{\infty}\pi_{i,2}z^i.$$

Due to the complexity of the CTMC, to make the associated formulas concise, we define the following equations.

$$\begin{aligned} A_i(z) &= (\lambda+\beta+\mu)z - \lambda z^2 - (1-\alpha_i)\mu, \\ B(z) &= (\lambda+\mu)z - \lambda z^2 - (1-\alpha_1)\mu, \\ C_i(z) &= \mu z - (1-\alpha_i)\mu, \quad D_i = p_m(1-\alpha_i)\mu, \\ E(z) &= \lambda z - \mu, \quad H_i = \alpha_i\mu + \beta, \end{aligned}$$

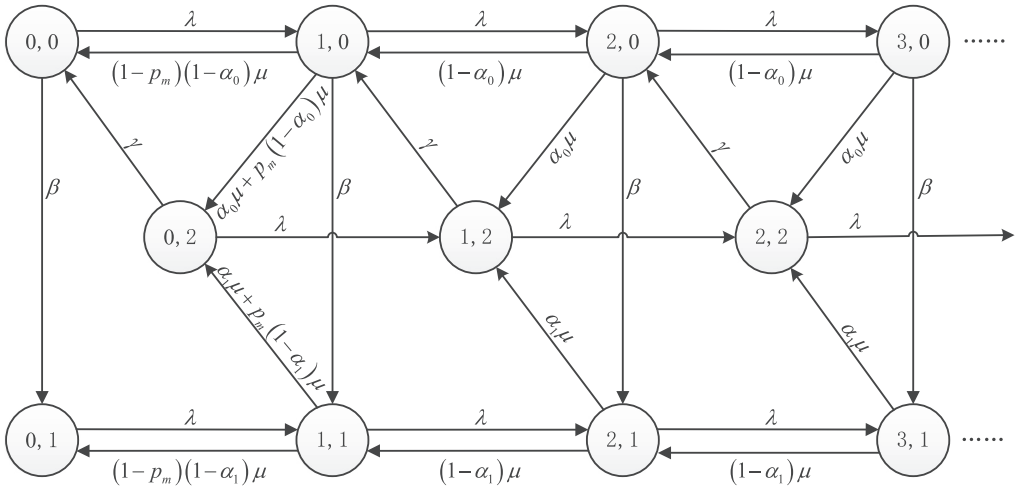

Fig. 3. The state transition diagram of M/M/1 case.

$$
\begin{aligned}
F(z) &= \gamma + (1-z)\lambda, \quad G(z) = A_1(z)E(z)\gamma + A_0(z)B(z)\lambda,\\
I_i &= (1-p_m)(1-\alpha_i)\mu, \quad E_1 = \lambda - \mu,\\
G_1 &= E_1H_1\gamma + \alpha_1\mu\lambda H_0, \quad B_0 = (\lambda+\mu)z_0 - \lambda z_0^2 - (1-\alpha_1)\mu,\\
K_{00} &= \alpha_1\mu^2(\alpha_0\lambda-\gamma)(\alpha_1\mu\gamma+\beta\gamma+\alpha_1\mu H_0) - \alpha_0\alpha_1\mu^2G_1,\\
K_{01} &= \alpha_1\mu\gamma(G_1 - \lambda\beta\gamma - \alpha_1\mu\lambda\gamma - \alpha_1\mu\lambda H_0),\\
K_{10} &= \alpha_1\mu\gamma D_0(E_1H_1 - \alpha_1\mu\lambda - \lambda\beta), \quad K_{11} = -\alpha_1\mu^2\gamma D_1H_0,\\
a_1 &= (\lambda K_{00} + \beta K_{01})(\mu\gamma + I_0\lambda) + \lambda(\lambda+\beta)(\lambda+\gamma)K_{10},\\
a_2 &= (I_1K_{01} + \lambda K_{11})(\mu\gamma + I_0\lambda) - \lambda\gamma(\alpha_1\mu + D_1)K_{10},\\
a_3 &= \mu\lambda(\alpha_0\lambda - F_0)(\mu\gamma + I_0\lambda)B_0 - \alpha_1\mu\lambda\gamma\beta z_0(\mu\gamma + I_0\lambda)\\
&\quad - (\lambda+\beta)(\lambda+\gamma)\lambda^2B_0D_0z_0,\\
a_4 &= (\alpha_1\mu + D_1)\lambda^2\gamma B_0D_0z_0 + \gamma\lambda(\mu\gamma + I_0\lambda)D_1E_0z_0\\
&\quad - \alpha_1\mu\lambda\gamma z_0I_1(\mu\gamma + I_0\lambda),
\end{aligned}
$$

where $i \in \{0, 1\}$ and z_0 is the positive root of $G(z_0) = 0$.

Theorem 1. *When the ergodic condition in Lemma 1 holds, the PGFs are*

$$
\begin{cases}
P_0(z) = \frac{\mu[\alpha_0\lambda - F(z)]B(z)\pi_{0,0} - \alpha_1\mu\lambda\gamma z\pi_{0,1} - \lambda B(z)D_0z\pi_{1,0} + \gamma D_1E(z)z\pi_{1,1}}{G(z)},\\
P_1(z) = \frac{\mu\beta[\alpha_0\lambda - F(z)]B(z)z\pi_{0,0} + [C_1(z)G(z) - \alpha_1\mu\lambda\beta\gamma z^2]\pi_{0,1}}{B(z)G(z)}\\
\qquad - \frac{\lambda\beta B(z)D_0z^2\pi_{1,0} - D_1[\beta\gamma E(z)z - G(z)]z\pi_{1,1}}{B(z)G(z)},\\
P_2(z) = \frac{[\mu(\alpha_0\lambda - F(z))A_0(z)B(z) - C_0(z)G(z)]\pi_{0,0} - \alpha_1\mu\lambda\gamma A_0(z)z\pi_{0,1}}{\gamma zG(z)}\\
\qquad + \frac{\gamma A_1(z)D_0E(z)z\pi_{1,0} + \gamma A_0(z)D_1E(z)z\pi_{1,1}}{\gamma zG(z)},
\end{cases}
$$

where

$$\pi_{0,0} = \frac{a_4(\alpha_1\mu\lambda\gamma(\mu\gamma + I_0\lambda)G_1)}{a_1a_4 - a_2a_3}, \tag{8}$$

$$\pi_{0,1} = \frac{(\beta a_4 - a_3I_1)(\alpha_1\mu\lambda\gamma(\mu\gamma + I_0\lambda)G_1)}{\lambda(a_1a_4 - a_2a_3)}, \tag{9}$$

$$\pi_{1,0} = \frac{[(\lambda+\beta)(\lambda+\gamma)a_4 + (\alpha_1\mu + D_1)\gamma a_3](\alpha_1\mu\lambda\gamma(\mu\gamma + I_0\lambda)G_1)}{(\mu\gamma + I_0\lambda)(a_1a_4 - a_2a_3)}, \tag{10}$$

$$\pi_{1,1} = \frac{-a_3(\alpha_1\mu\lambda\gamma(\mu\gamma + I_0\lambda)G_1)}{a_1a_4 - a_2a_3}. \tag{11}$$

Proof. See the Appendix. □

Define p_i, $i \in \{0, 1, 2\}$, as the steady-state probability that the server is in state i. According to Theorem 1, it is easy to obtain the close-form expressions of these steady-state probabilities in the following proposition.

Proposition 1. *If the CTMC is ergodic, the steady-state probabilities that the server is in states 0, 1 and 2 are given by*

$$
\begin{cases}
p_0 = \frac{\alpha_1\mu^2(\alpha_0\lambda-\gamma)\pi_{0,0} - \alpha_1\mu\lambda\gamma\pi_{0,1} - \alpha_1\mu\lambda D_0\pi_{1,0} + \gamma D_1E_1\pi_{1,1}}{G_1},\\
p_1 = \frac{\alpha_1\mu^2\beta(\alpha_0\lambda-\gamma)\pi_{0,0} + \alpha_1\mu(G_1 - \lambda\beta\gamma)\pi_{0,1} - \alpha_1\mu\lambda\beta D_0\pi_{1,0} + D_1(\beta\gamma E_1 - G_1)\pi_{1,1}}{\alpha_1\mu G_1},\\
p_2 = \frac{[\alpha_1\mu^2(\alpha_0\lambda-\gamma)H_0 - \alpha_0\mu G_1]\pi_{0,0} - \alpha_1\mu\lambda\gamma H_0\pi_{0,1} + \gamma D_0E_1H_1\pi_{1,0} + \gamma D_1E_1H_0\pi_{1,1}}{\gamma G_1},
\end{cases}
$$

where $\pi_{0,0}$, $\pi_{0,1}$, $\pi_{1,0}$, $\pi_{1,1}$ *are given by Eqs. (8)–(10).*

Proof. See the Appendix. □

Let W be the expected sojourn time of a customer in the system, the following proposition provides its closed-form solution.

Proposition 2. *If the CTMC is ergodic, the expected sojourn time of the customer is*

$$
\begin{aligned}
W &= \frac{f_0'G_1 - f_0G_1'}{\lambda G_1^2} + \frac{\alpha_1\mu f_1'G_1 - f_1[(\mu-\lambda)G_1 + \alpha_1\mu G_1']}{\lambda\alpha_1^2\mu^2G_1^2}\\
&\quad + \frac{f_2'G_1 - f_2(G_1 + G_1')}{\lambda\gamma G_1^2},
\end{aligned} \tag{12}
$$

where

$$
\begin{aligned}
G_1' &= (\beta - E_1)(E_1\gamma + \alpha_1\mu\lambda) + \lambda(H_1\gamma - H_0E_1),\\
f_0 &= \alpha_1\mu^2(\alpha_0\lambda - \gamma)\pi_{0,0} - \alpha_1\mu\lambda\gamma\pi_{0,1}\\
&\quad - \alpha_1\mu\lambda D_0\pi_{1,0} + \gamma D_1E_1\pi_{1,1},\\
f_1 &= \alpha_1\mu^2\beta(\alpha_0\lambda - \gamma)\pi_{0,0} + \alpha_1\mu(G_1 - \lambda\beta\gamma)\pi_{0,1}\\
&\quad - \alpha_1\mu\lambda\beta D_0\pi_{1,0} + D_1(\beta\gamma E_1 - G_1)\pi_{1,1},\\
f_2 &= [\alpha_1\mu^2(\alpha_0\lambda - \gamma)H_0 - \alpha_0\mu G_1]\pi_{0,0} - \alpha_1\mu\lambda\gamma H_0\pi_{0,1}\\
&\quad + \gamma D_0E_1H_1\pi_{1,0} + \gamma D_1E_1H_0\pi_{1,1},\\
f_0' &= \mu[\alpha_1\mu\lambda - (\alpha_0\lambda - \gamma)E_1]\pi_{00} - \alpha_1\mu\lambda\gamma\pi_{01}\\
&\quad - \lambda(\alpha_1\mu - E_1)D_0\pi_{10} + \gamma(\lambda + E_1)D_1\pi_{11},\\
f_1' &= \alpha_1\mu^2\beta[\alpha_1\mu\lambda - (\alpha_0\lambda - \gamma)E_1]\pi_{00}\\
&\quad + \alpha_1\mu^2(G_1 + \alpha_1G_1' - 2\alpha_1\lambda\beta\gamma)\pi_{01}\\
&\quad - \alpha_1\mu\lambda\beta(2\alpha_1\mu - E_1)D_0\pi_{10}\\
&\quad + \alpha_1\mu D_1[\beta\gamma(\lambda + 2E_1) - G_1 - G_1']\pi_{11},\\
f_2' &= \mu^2\gamma[\alpha_1\lambda H_0 - G_1 - \alpha_0G_1']\pi_{00}\\
&\quad + \mu\gamma(\alpha_0\lambda - \gamma)[\alpha_1\mu(\beta - E_1) - H_0E_1]\pi_{00}
\end{aligned}
$$

$$-\alpha_1\mu\lambda\gamma^2(\beta-E_1+H_0)\pi_{01}+\gamma^2D_0[(\beta-E_1)E_1 +H_1(\lambda+E_1)]\pi_{10}+\gamma^2D_1[(\beta-E_1)E_1+H_0(\lambda+E_1)]\pi_{11},$$

where $\pi_{0,0}$, $\pi_{0,1}$, $\pi_{1,0}$, $\pi_{1,1}$ *are given by Eqs. (8)–(10).*

Proof. According to Theorem 1, the PGF of $N_1(t)$ is

$$L(z)=P_0(z)+P_1(z)+P_2(z).$$

Hence, the expected number of customers in the system can be calculated as

$$L=\lim_{z\to1}\frac{\partial P_0(z)}{\partial z}+\lim_{z\to1}\frac{\partial P_1(z)}{\partial z}+\lim_{z\to1}\frac{\partial P_2(z)}{\partial z} =\left(\frac{f_0'G_1-f_0G_1'}{G_1^2}+\frac{\alpha_1\mu f_1'G_1-f_1[(\mu-\lambda)G_1+\alpha_1\mu G_1']}{\alpha_1^2\mu^2G_1^2} +\frac{f_2'G_1-f_2(G_1+G_1')}{\gamma G_1^2}\right).$$

Following the Little's formula, i.e., $W=\frac{1}{\lambda}L$, the expected waiting time in the system (the expected sojourn time) is yielded. This finishes the proof. □

Though the notation used in Proposition 2 looks a bit complicated, the result in Eq. (12) is intuitive. The first (second or third) term just represents that the expected sojourn time when server is in the low energy consumption (high energy consumption or maintenance) state.

3.2. General case: MAP/PH/1 queueing service system

Though we can obtain mathematically tractable results based on the special case, the assumptions of Possion arrivals and exponential service time make the problem restrictive and not adaptive to the real-world situations. To overcome those limitations, in this subsection, we extend our basic maintenance model to a more general case, i.e., MAP/PH/1 queueing service system. In this system, the customers are still served based on the first-come, first-served (FCFS) discipline, but the customer arrival process is generalized to an MAP with an infinitesimal generator $S=S^0+S^1$ in the state space {1, ..., m}, where $S^0=(S^0_{i,j})_{m\times m}$ and $S^1=(S^1_{i,j})_{m\times m}$. Specifically, the off-diagonal elements in S^0 and all the elements in S^1 are nonnegative, and all of the diagonal entries of S^0 are nonpositive. A single customer only arrives at each Type-1 transition epoch. Assume the underlying Markov chain S possesses the irreducibility and let $\boldsymbol{\chi}$ be its stationary probability vector. Then, $\boldsymbol{\chi}$ is uniquely determined by $\boldsymbol{\chi}\cdot(S^0+S^1)=\mathbf{0}$ and $\boldsymbol{\chi}\cdot\mathbf{1}=1$, and the mean arrival rate of the MAP is $\widehat{\lambda}=\boldsymbol{\chi}S^1\mathbf{1}$, where $\mathbf{1}$ is an all-ones vector. In addition, the variance ν of customer inter-arrival time can be calculated as $\nu=2\widehat{\lambda}^{-1}\boldsymbol{\chi}(-S^0)^{-1}\mathbf{1}-\widehat{\lambda}^{-2}$, and the squared coefficient of variation (SCV) is given by $c_\nu=2\widehat{\lambda}\boldsymbol{\chi}(-S^0)^{-1}\mathbf{1}-1$. For further properties of the MAP, we refer reader to Asmussen and Koole (1993) and Latouche and Ramaswami (1999).

On the other hand, the service time is assumed to follow a PH distribution with representation (θ, T) of order n, which indicates the number of phases of the server is n and $T^0=-T\mathbf{1}$. With the PH distribution, the service time can be quantified as the time required for the underlying Markov process (with finite states $\{1,2,\ldots,n,n+1\}$) to reach the single absorbing state $n+1$, conditioned on the fact that the initial state of this process starts from one of the states {1, 2, ..., n} (according to initial probability vector α). Then, the matrix T can be interpreted as the transition rate matrix for the transient states, while T^0 represents the column vector of absorbing rates. Hence, The mean and variance of the service time can be calculated by $\widehat{\mu}=-\theta T^{-1}\mathbf{1}$ and $\widehat{\nu}=2\theta(-T)^{-2}\mathbf{1}-(-\theta T^{-1}\mathbf{1})^2$, and the SCV is $\widehat{c_\nu}=(2\theta(-T)^{-2}\mathbf{1})/\widehat{\mu}^2-1$. The readers are referred to Neuts (1981) for more information about PH distributions.

Note that, by using the MAP and PH distribution will significantly increase the dimension of state space and complexity of the system. To address these difficulties, we need to apply the matrix-analytic methods (Neuts, 1981) to study the steady-state performance of the associated queueing system. Let $M_1(t)$ be the number of customers in the system (including the ones in service), $M_2(t)$ be the condition state of the server, which is defined as

$$M_2(t)=\begin{cases}0, & \text{low energy consumption state,}\\ 1, & \text{high energy consumption state,}\\ 2, & \text{maintenance state,}\end{cases}$$

$M_3(t)$ be the phase of the service process at time t, and $M_4(t)$ be the phase of the arrival process at time t.

Under the proposed assumptions, we can construct a QBD process

$$\{M_1(t),M_2(t),M_3(t),M_4(t),t\geq0\}$$

with state space

$$\Omega=\bigcup_{i=0}^{\infty}l(i),$$

where

$$l(0)=\{(0,0,1),\ldots,(0,0,m),\ldots,(0,1,m),\ldots,(0,2,m)\},$$

and

$$l(i)=\{(i,j,k,h):j=0,1,2,1\leq k\leq n,1\leq h\leq m\},i\geq1.$$

The states are arranged in the standard ascending order as follows:

Level 0: (0,0,1),...,(0,0,m),...,(0,1,m),...,(0,2,m),
Level 1: (1,0,1,1),...,(1,0,1,m),...,(1,0,2,m),...,(1,0,n, m),...,(1,1,n, m),..., (1,2,n, m),
Level 2: (2,0,1,1),...,(2,0,0,m),...,(2,0,2,m),...,(2,0,n, m),...,(2,1,n, m), ...,(2,2,n, m), ..., ..., ...

Particularly, when $M_1(t)=0$, $M_3(t)$ does not play any role in the system and will not be tracked. Thus, in this case, one only need to take care of states $M_2(t)$ and $M_4(t)$.

Then, denote I as the identity matrix, the transition matrix generator Q of the QBD process can be written as:

$$Q=\begin{pmatrix}K_0 & J_0 & & & \\ Y_0 & K_1 & J & & \\ & Y & K_1 & J & \\ & & Y & K_1 & J \\ & & & \ddots & \ddots & \ddots\end{pmatrix}, \tag{13}$$

where

$$K_0=\begin{pmatrix}S^0-\beta I_m & \beta I_m & 0\\ 0 & S^0 & 0\\ \gamma I_m & 0 & S^0-\gamma I_m\end{pmatrix}_{3m\times3m},$$

$$J_0=\begin{pmatrix}\theta\otimes S^1 & 0 & 0\\ 0 & \theta\otimes S^1 & 0\\ 0 & 0 & \theta\otimes S^1\end{pmatrix}_{3m\times3mn},$$

$$J=\begin{pmatrix}I_n\otimes S^1 & 0 & 0\\ 0 & I_n\otimes S^1 & 0\\ 0 & 0 & I_n\otimes S^1\end{pmatrix}_{3mn\times3mn},$$

$$K_1=\begin{pmatrix}T\oplus S^0-\beta I_{mn} & \beta I_{mn} & 0\\ 0 & T\oplus S^0 & 0\\ \gamma I_{mn} & 0 & I_n\otimes S^0-\gamma I_{mn}\end{pmatrix}_{3mn\times3mn},$$

$$Y = \begin{pmatrix} (1-\alpha_0)\theta T^0 \otimes I_{mn} & 0 & \alpha_0\theta T^0 \otimes I_{mn} \\ 0 & (1-\alpha_1)\theta T^0 \otimes I_{mn} & \alpha_1\theta T^0 \otimes I_{mn} \\ 0 & 0 & 0 \end{pmatrix}_{3mn\times 3mn},$$

$$Y_0 = \begin{pmatrix} (1-p_m)(1-\alpha_0)T^0 \otimes I_m & 0 & \{\alpha_0 + p_m(1-\alpha_0)\}T^0 \otimes I_m \\ 0 & (1-p_m)(1-\alpha_1)T^0 \otimes I_m & \{\alpha_1 + p_m(1-\alpha_1)\}T^0 \otimes I_m \\ 0 & 0 & 0 \end{pmatrix}_{3mn\times 3m}.$$

Let $\pi_{0,j,h} = \lim_{t\to\infty}\pi_{0,j,h}(t)$ and $\pi_{i,j,k,h} = \lim_{t\to\infty}\pi_{i,j,k,h}(t)(i \geq 1)$ be the steady-state probability of the QBD, and denote

$$\begin{aligned}\Pi_0 &= (\pi_{0,0,1}, \ldots, \pi_{0,0,m}, \ldots, \pi_{0,1,m}, \ldots, \pi_{0,2,m}),\\ \Pi_i &= (\pi_{i,0,1,1}, \ldots, \pi_{i,0,1,m}, \pi_{i,0,2,1}, \ldots, \pi_{i,0,2,m}, \ldots,\\ &\quad \times \pi_{i,0,n,m}, \ldots, \pi_{i,1,n,m}, \ldots, \pi_{i,2,n,m}),\\ \Pi &= (\Pi_0,\ \Pi_1,\ \Pi_2,\ \ldots),\end{aligned}$$

where Π_0 has a dimension of $3m$, and Π_i $(i = 1, 2, 3, \ldots)$ have a dimension of $3mn$. By implementing the matrix geometric method from Neuts (1981), it is easy to obtain that

$$\Pi_{i+1} = \Pi_i R = \Pi_1 R^{i-1},\ i \geq 1.$$

Because $J + RK_1 + R^2Y = 0$, R can be calculated by the following iterative approach:

$$R(n+1) = -(J + R(n)^2 Y)K_1^{-1}.$$

According to the simultaneous linear equations $\Pi Q = 0$ and $\Pi\mathbf{1} = 1$, the boundary vectors Π_0 and Π_1 can be computed from

$$\begin{cases} \Pi_0 K_0 + \Pi_1 Y_0 = 0, \\ \Pi_0 J_0 + \Pi_1 K_1 + \Pi_2 Y = \Pi_0 J_0 + \Pi_1(K_1 + RY) = 0, \\ \sum_{i=0}^{\infty}\Pi_i = \Pi_0\mathbf{1} + \Pi_1\sum_{i=1}^{\infty}R^{i-1}\mathbf{1} = \Pi_0\mathbf{1} + \Pi_1(I-R)^{-1}\mathbf{1} = 1. \end{cases}$$

Define $\widehat{p}_j$, $j \in \{0, 1, 2\}$, as the steady-state probability that the server is in state j, the following proposition gives these steady-state probabilities.

Proposition 3. *The steady-state probability $\widehat{p}_j$, $j \in \{0, 1, 2\}$, that the server is in state j is given by*

$$\widehat{p}_j = \sum_{h=1}^{m}\pi_{0,j,h} + \sum_{i=1}^{\infty}\sum_{k=1}^{n}\sum_{h=1}^{m}\pi_{i,j,k,h}.$$

Proof. It can be obtained by definition. □

Let $\widehat{W}$ be the expected sojourn time of a customer in the system, it can be solved by the following proposition.

Proposition 4. *The expected sojourn time of the customer in the system is*

$$E[\widehat{W}] = \frac{\Pi_1(I-R)^{-2}\mathbf{1}}{\widehat{\lambda}}.$$

Proof. The expected number of customers in the system $E[\widehat{L}]$ can be expressed as

$$E[\widehat{L}] = \sum_{i=0}^{\infty} i\Pi_i\mathbf{1} = \sum_{i=1}^{\infty} i\Pi_1 R^{i-1}\mathbf{1} = \Pi_1(I-R)^{-2}\mathbf{1}.$$

Thus, following the Little's formula, the expected sojourn time of the customer in the system can be easily obtained. □

3.3. Extension: multi-energy-consumption-level service system

In practice, the main causes of the system degeneration are the glitches or wears of some components. Different causes may result in different degradation levels, and the system health states will have more than two levels. For example, if only one of the cooling fans of the CPU fails, there may not have significant influence on power consumption of server farm. However, if the whole cooling system of the server farm is down, the power consumption level will sharply increase. Thus, two energy consumption states are not enough to deal with more complex service systems. Based on the previously constructed MAP/PH/1 model, we can extend the two-energy-consumption-level case to a multi-level case by adding states in the transition diagram. Take a three-energy-consumption-level server as example, i.e., low energy consumption state 0, high energy consumption state 1, and medium energy consumption state 3, we assume $e \sim N(e_i, \sigma^2)$ and $e^I \sim N(e_i^I, \sigma^2)$ $(i = 0, 1, 3)$ when the server is busy and idle in state i, respectively. Thus, it reasonable to have $e_0 < e_3 < e_1$, $e_0^I < e_3^I < e_1^I$, and $e_i^I < e_i$. In addition, the deterioration process of the server follows the Poisson process with rate β_j $(j = 1, 2, 3)$, that is, the mean time from state 0 to state 1, from state 0 to state 3, and from state 3 to state 1 are $1/\beta_1$, $1/\beta_2$, and $1/\beta_3$. Similarly, the probability α_2 that, given the server is in state 3, the control chart signals the "out-of-control" message can be obtained as

$$\begin{aligned}\alpha_2 &= P\{e \in I_2 | i = 3\}\\ &= P\left\{\frac{e-e_3}{\sigma} \geq \frac{e_0 - e_3 + k\sigma}{\sigma} | i = 3\right\}\\ &\quad + P\left\{\frac{e-e_3}{\sigma} \geq \frac{e_0 - e_3 - k\sigma}{\sigma} | i = 3\right\}\\ &= 1 - \Phi\left(\frac{e_0 - e_3 + k\sigma}{\sigma}\right) + \Phi\left(\frac{e_0 - e_3 - k\sigma}{\sigma}\right).\end{aligned}$$

Let $\overline{M}_1(t)$ be the number of customers in the system (including the ones in service), $\overline{M}_2(t)$ be the condition state of the server, which is defined as

$$\overline{M}_2(t) = \begin{cases} 0, & \text{low energy consumption state,} \\ 1, & \text{high energy consumption state,} \\ 2, & \text{maintenance state,} \\ 3, & \text{medium energy consumption state,} \end{cases}$$

$\overline{M}_3(t)$ be the phase of the service process at time t, and $\overline{M}_4(t)$ be the phase of the arrival process at time t. Under the proposed assumptions, we can similarly construct a QBD process as

$$\{\overline{M}_1(t), \overline{M}_2(t), \overline{M}_3(t), \overline{M}_4(t), t \geq 0\}.$$

Although, the system becomes more complex when the number of energy-consumption states increases, the associated steady-state performance can still be studied by employing the matrix-analytic methods. As an illustrative purpose, we only consider two energy-consumption levels for the system throughout the rest of this paper.

4. Optimal maintenance policy

In this section, we can develop the corresponding maintenance policy based on the above theoretical results. From a long-run perspective, we have investigated the steady-state behavior of the system. In this section, we attempt to develop the associated optimal

maintenance policy to balance the tradeoffs in this problem. Two types of maintenance activities are considered, i.e., the PM and the RM. Because the PM is controlled by the frequency parameter p_m and the RM relies on the control limit parameter k, our objective is to jointly optimize these two decision variables (p_m, k) to improve the system's performance.

For a high-energy-consumption service facility, the tradeoff between energy consumption and maintenance cost, which significantly affects the maintenance decisions, has been extensively studied. Most of the existing research (see Ben-Daya & Rahim, 2000; Lam et al., 2006; Panagiotidou and, 2007; Zhou & Zhu, 2008; Liao, Xie, & Jin, 2013, etc.) balances the tradeoff by minimizing the long-run expected cost per unit time. However, this is insufficient for a service system, in which the customer delay is one of the most important indexes. To improve the customer satisfaction, the service system needs to reduce the delay time as well. Thus, customer delay should be also considered in the decision-making process of the maintenance policy. In practice, many services have waiting time limitations (e.g., the examination in a hospital), which can be used as the service level constraint in our model. Therefore, in this work, unlike the previous research, we optimize the maintenance policy by considering the energy-delay tradeoff.

Let $\widehat{TC}$ be the long-run expected cost per unit time of the service system. Based on the above discussion, $\widehat{TC}$ consists of two parts of costs, i.e., the energy consumption cost and the maintenance cost. Note that, different types of states will have different levels of energy consumption. Thus, we need to consider the costs of idle and busy states. In addition, the maintenance cost is split into the PM cost and the RM cost. Let C_e be the cost of per unit energy, C_{pm} and C_{rm} be the costs of realizing the PM and the RM per unit time, respectively. Then, under the general MAP/HP/1 situation proposed in Section 3.2, $\widehat{TC}$ can be quantified by

$$\begin{aligned}\widehat{TC} = C_e[e_0^l\widehat{\pi}_{0,0} + e_1^l\widehat{\pi}_{0,1} + e_0(\widehat{p}_0 - \widehat{\pi}_{0,0}) + e_1(\widehat{p}_1 - \widehat{\pi}_{0,1})] \\ + C_{pm}\widehat{\pi}_{pm} + C_{rm}(\widehat{p}_2 - \widehat{\pi}_{pm}),\end{aligned} \tag{14}$$

where the steady-state probabilities $\widehat{p}_j$ $(j = 0, 1, 2)$ are shown in Proposition 3 and the boundary probabilities can be calculated as

$$\begin{cases}\widehat{\pi}_{0,0} = \sum_{h=1}^{m}\pi_{0,0,h}, \\ \widehat{\pi}_{0,1} = \sum_{h=1}^{m}\pi_{0,1,h}.\end{cases}$$

Furthermore, $\widehat{\pi}_{pm}$ can be obtained by

$$\begin{aligned}\widehat{\pi}_{pm} &= P\{PMstate|maintenance\ state\} \\ &= \sum_{i=0}^{\infty}\sum_{k=1}^{n}\sum_{h=1}^{m}P\{PMstate|(i,j,k,h) = (i,2,k,h)\} \\ &= \sum_{h=1}^{m}\widetilde{\pi}_{0,2,h} + \sum_{i=1}^{\infty}\sum_{k=1}^{n}\sum_{h=1}^{m}\widetilde{\pi}_{i,2,k,h},\end{aligned}$$

where

$$\begin{cases}\widetilde{\pi}_{0,2,h} = \pi_{0,2,h}\frac{\pi_{1,0,k,h}[p_m(1-\alpha_0)T^0\otimes I_m] + \pi_{1,1,k,h}[p_m(1-\alpha_1)T^0\otimes I_m]}{\pi_{1,0,k,h}[(\alpha_0+p_m(1-\alpha_0))T^0\otimes I_m] + \pi_{1,1,k,h}[(\alpha_1+p_m(1-\alpha_1))T^0\otimes I_m]}, \\ \widetilde{\pi}_{1,2,k,h} = \pi_{1,2,k,h}\frac{\widetilde{\pi}_{0,2,h}(\theta\otimes S^1)}{\pi_{0,2,h}(\theta\otimes S^1) + \pi_{2,0,k,h}[\alpha_0\theta T^0\otimes I_{mn}] + \pi_{2,1,k,h}[\alpha_1\theta T^0\otimes I_{mn}]}, \\ \widetilde{\pi}_{i,2,k,h} = \pi_{i,2,k,h}\frac{\widetilde{\pi}_{i-1,2,k,h}(I_n\otimes S^1)}{\pi_{i-1,2,k,h}(I_n\otimes S^1) + \pi_{i+1,0,k,h}[\alpha_0\theta T^0\otimes I_{mn}] + \pi_{i+1,1,k,h}[\alpha_1\theta T^0\otimes I_{mn}]}, \\ \quad i \geq 2.\end{cases}$$

To guarantee the service quality, the waiting time limitation is set as w_0. Meanwhile, the expected customer waiting time should not exceed w_0. Because we do not consider the service capacity design, there exists a lower bound $\underline{w}_0 = \inf w_0 = \lim_{p_m,\alpha_0,\alpha_1\to 0}\widehat{W}$ for w_0. In other words, the lower bound is exact the expected waiting time of the system without the RM and PM activities (both of RM and PM will increase the expected customer waiting time). Then, the optimization problem is formulated as

$$\begin{aligned}\min_{p_m,k}\quad & \widehat{TC} = C_e[e_0^l\widehat{\pi}_{0,0} + e_1^l\widehat{\pi}_{0,1} + e_0(\widehat{p}_0 - \widehat{\pi}_{0,0}) \\ & \qquad + e_1(\widehat{p}_1 - \widehat{\pi}_{0,1})] \\ & \qquad + C_{pm}\widehat{\pi}_{pm} + C_{rm}(\widehat{p}_2 - \widehat{\pi}_{pm}) \\ \text{subject to:}\quad & \widehat{W} \leq w_0, \\ & 0 \leq p_m \leq 1, k \geq 0.\end{aligned} \tag{15}$$

We aim at developing an optimal maintenance policy (p_m, k) to minimize the long-run expected energy consumption and maintenance cost per unit time of the system. Note that the models studied in Section 3 can be readily reduced to the one without control chart by setting $k = \infty$ ($\alpha_i = 0$).

Remark 1. For the M/M/1 situation proposed in Section 3.1, the long-run expected cost per unit time of the service system TC is similar to Eq. (14) as

$$\begin{aligned}TC = C_e[e_0^l\pi_{0,0} + e_1^l\pi_{0,1} + e_0(p_0 - \pi_{0,0}) + e_1(p_1 - \pi_{0,1})] \\ + C_{pm}\pi_{pm} + C_{rm}(p_2 - \pi_{pm}),\end{aligned}$$

where π_{pm} is the probability when the system is in the PM state. Then, π_{pm} can be obtained by

$$\begin{aligned}\pi_{pm} &= P\{PMstate|maintenance\ state\} \\ &= \sum_{i=0}^{\infty}P\{PMstate|(i,j) = (i,2)\} \\ &= \sum_{i=0}^{\infty}\widetilde{\pi}_{i,2},\end{aligned}$$

where

$$\begin{cases}\widetilde{\pi}_{0,2} = \pi_{0,2}\frac{p_m(1-\alpha_0)\mu\pi_{1,0} + p_m(1-\alpha_1)\mu\pi_{1,1}}{[\alpha_0\mu + p_m(1-\alpha_0)\mu]\pi_{1,0} + [\alpha_1\mu + p_m(1-\alpha_1)\mu]\pi_{1,1}}, \\ \widetilde{\pi}_{i,2} = \pi_{i,2}\frac{\lambda\widetilde{\pi}_{i-1,2}}{\lambda\pi_{i-1,2} + \alpha_0\mu\pi_{i+1,0} + \alpha_1\mu\pi_{i+1,1}}, \qquad i \geq 1, \\ \pi_{0,2} = \frac{a_4(\lambda+\beta)(\alpha_1\mu\lambda\gamma(\mu\gamma+l_0\lambda)G_1)}{\gamma(a_1a_4 - a_2a_3)} \\ \qquad - \frac{l_0[(\lambda+\beta)(\lambda+\gamma)a_4 + (\alpha_1\mu + D_1)\gamma a_3](\alpha_1\mu\lambda\gamma(\mu\gamma+l_0\lambda)G_1)}{\gamma(\mu\gamma+l_0\lambda)(a_1a_4 - a_2a_3)}.\end{cases}$$

5. Numerical experiments

5.1. Application to a healthcare equipment

Computed Tomography (CT) scan is an efficient and important testing tool in large general hospitals. The CT scanner is one of the most expensive medical equipment (for example, a 320-slice CT scanner costs about 0.8–1.5 million dollars). Thus, not all of the hospitals have the ability to install and operate the machine. As a result, patients may have to make appointments in advance and face a long waiting time. Besides, the CT scanner is a high-power electrical appliance that consists of different energy-consuming systems, such as the X-ray tubes, cooling system, gradient system, computer system, etc. Therefore, an appropriate maintenance policy should be developed for the CT scanner to improve the reliability, reduce the energy consumption and guarantee the service level.

We take the CT scanner as an illustrative example and apply the proposed model to obtain the optimal maintenance policy (p_m, k) for the hospital. All the required data of the model can be collected from practice, for example, for the M/M/1 system, we can collect the arrival time of each patient and construct a data set of inter-arrival time. This data set can be used to estimate the arrival rate λ. In addition, we can further record the starting time and ending time of the scanning process for each patient. Then, we can calculate the scanning time of each patient and use these data to estimate the service rate μ. All these estimates can be done by Maximum Likelihood Estimation. In this example, for model exhibit purpose, we just assume the parameter settings. Because the

Table 1
Control-chart-based optimal policy and the corresponding costs.

Parameters	(p_m^*, k^*)	W^*	TC^*	$\pi_{0,0}^*$	π_{01}^*	π_{pm}^*	p_0^*	p_1^*	p_2^*
Optimal value	(0.52,2.73)	6.37	79.83	0.07	0.01	0.17	0.66	0.10	0.24

Table 2
Optimal policy (p_m^*, k^*) for different customer arrival processes.

Input	Mean $\widehat{\lambda}$	Variance ν	SCV c_ν	(p_m^*, k^*)	TC^*
Poisson	1	1	1	(0.49,3.44)	77.31
Er1	1	0.3	0.3	(0.66,2.92)	77.26
MAP	1	4	4	(1,3.02)	77.45

proposed model is adaptive to different situations, once the associated data and/or parameters are available, we can apply the solution procedure in the same way.

The system parameters are set as: the customers' arrival rate $\lambda = 0.65$, the service rate is $\mu = 1$, the occurrence rate of glitches is $\beta = 0.04$, the repair rate is $\gamma = 0.35$, and the waiting time limitation is $w_0 = 10$ hours. The mean energy consumption levels with the same standard deviation adopted are $e_0 = 65$ kilowatt, $e_1 = 130$ kilowatt, $e_0^I = 50$ kilowatt, $e_1^I = 100$ kilowatt, and $\sigma = 30$ kilowatt. The average cost per unit energy is $C_e = \$1$/kilowatt hour, the average costs per unit time of the PM and the RM are $C_{pm} = \$85$/hour and $C_{rm} = \$150$/hour, respectively.

First, we discuss the case when the control chart is used. Due to the analytical complexity, to solve the optimization problem in Eq. (15), a grid search method is employed to obtain an approximated optimal maintenance policy (p_m, k). We search k from 0 to 6 and p_m from 0 to 1 with the same step size 0.01. The corresponding results are presented in Table 1.

One can see that, from Table 1, the optimal maintenance policy is $(p_m^*, k^*) = (0.52, 2.73)$ and the minimum long-run expected cost per unit time $TC^* = \$79.83$/hour.

Before examining the model sensitivity and effects of parameters, to validate the applicability of the proposed model, we further show the optimal maintenance policies for the CT scanner problem with different arrival processes and service time distributions. The same setting of the basic level is remained except the arrival process and the service time distribution. We first construct the following three MAPs with the same mean value, but different variances, to examine the effects of various arrival processes.

1. Poisson with mean 1.
2. Erlang-3 (Er1)

$$S^0 = \begin{pmatrix} -3 & 3 & 0 \\ 0 & -3 & 3 \\ 0 & 0 & -3 \end{pmatrix}, \quad S^1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 3 & 0 & 0 \end{pmatrix}.$$

3. Markovian arrival process (MAP)

$$S^0 = \begin{pmatrix} -2.66 & 0.12 & 0.12 \\ 0.13 & -0.5 & 0.08 \\ 0.14 & 0.08 & -0.32 \end{pmatrix},$$

$$S^1 = \begin{pmatrix} 2.3 & 0.08 & 0.04 \\ 0.09 & 0.18 & 0.02 \\ 0.05 & 0.01 & 0.04 \end{pmatrix}.$$

The corresponding results are shown in the Table 2. From the table, one can see that our model works for different arrival processes, and the optimal p_m^*, k^*, and TC^* for can be obtained accordingly.

Similarly, for the service time, we also present three different PH distributions to show the model applicability.

1. Exponential with mean 2.

Table 3
Optimal policy (p_m^*, k^*) for different service distributions.

Input	Mean $\widehat{\mu}$	Variance $\widehat{\nu}$	SCV $\widehat{c_\nu}$	(p_m^*, k^*)	TC^*
Exponential	2	4	1	(0.49,3.44)	77.31
Er2	2	2	0.50	(0.96,4.10)	75.85
Er3	2	1.5	0.375	(0.97,4.26)	75.71

Table 4
Experimental data set and range of TC.

Parameters	−10%	Basic	+10%	TC^*			
				−10%	Basic	+10%	Range (%)
λ	0.585	0.650	0.715	79.41	79.83	80.23	1.02%
μ	0.900	1.000	1.100	80.62	79.83	79.23	0.76%
β	0.036	0.040	0.044	78.90	79.83	80.74	1.13%
γ	0.315	0.350	0.385	80.55	79.83	79.24	0.75%
e_0	58.5	65	71.5	75.44	79.83	84.18	5.44%
e_1	117	130	143	79.49	79.83	79.91	0.09%
e_0^I	45	50	55	79.45	79.49	80.18	0.88%
e_1^I	90	100	110	79.75	79.83	79.92	0.10%
C_{pm}	76.5	85	93.5	78.49	79.83	81.18	1.68%
C_{rm}	135	150	165	78.64	79.83	81.03	1.50%

2. Erlang-2 (Er2)

$$\theta = (1, 0), T = \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix}, \quad T^0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

3. Erlang-3 (Er3)

$$\theta = (1, 0, 0), \quad T = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -2 & 2 \\ 0 & 0 & -2 \end{pmatrix}, \quad T^0 = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}.$$

The corresponding results are shown in the Table 3.

From Tables 2 and 3, for different situations, though the expected customer arrival rates (service distributions) are the same, the optimal maintenance policies and costs may vary due to different variations or the SCVs.

5.2. Sensitivity analysis

In practice, the estimation of model parameters involves errors, thus, it is necessary to study the robustness of the proposed method. To demonstrate how the solution technique works, a systematic sensitivity analysis is conducted in this subsection. The analysis is performed based on the illustrative example in Section 5.1. In Table 4, basic level represents the result in Table 1. Based on the basic level, the results from inaccurate cases of the basic one are obtained by varying the values of the parameters by −10% and +10%, respectively. Suppose the same maintenance policy (0.52, 2.73) is applied for different levels (we adopt the analysis method used in Makis (2008)). Define that

$$range(\%) = \frac{max(-10\%, Basic, +10\%) - min(-10\%, Basic, +10\%)}{Basic} \times 100\%.$$

Then, with the optimal policy of the basic case, the *range* can be used as the volatility index of the long-run expected cost per unit time TC for the system using three levels of parameters.

As shown in Table 4, most of the *ranges* are well-controlled within 2% except for the *range* of e_0 reaching to 5.44%. The reason is that the steady-state probability that the server is in state 0 (p_0) is much bigger, compared to states 1 (p_1) and 2 (p_2) in this example. However, in practice, the value of e_0 can be measure and estimate easily, for example, the power consumption of CT scanner can be easily measure from the electricity meter. Therefore, the established maintenance model performs well, even if the parameter estimation is not accurate, and indicates a good applicability in practice.

Table 5
The effect of parameters and comparison of two maintenance models when $\omega_0 = 10$.

Parameter	Value	(p_m^*, k^*)	TC^*	W^*	$\widetilde{p_m}^*$	$\widetilde{TC}^*$	$\widetilde{W}^*$	Δ
λ	0.75	(0.20,2.95)	79.07	9.87	0.99	90.47	6.54	−12.60%
	0.70	(0.48,2.81)	79.49	7.73	0.95	86.81	5.80	−8.44%
	0.65	(0.52,2.73)	79.83	6.37	0.90	84.00	5.24	−4.95%
	0.60	(0.57,2.67)	80.03	5.50	0.86	81.82	4.81	−2.18%
	0.55	(0.63,2.58)	80.10	4.89	0.83	80.13	4.46	−0.04%
μ	1.10	(0.52,2.86)	79.18	5.05	0.80	81.22	4.49	−2.51%
	1.05	(0.52,2.86)	79.50	5.61	0.86	82.43	4.83	−3.56%
	1.00	(0.52,2.73)	79.83	6.37	0.90	84.00	5.24	−4.95%
	0.95	(0.53,2.67)	80.18	7.49	0.95	86.03	5.77	−6.81%
	0.90	(0.53,2.61)	80.53	9.26	1.00	88.75	6.50	−9.26%
β	0.06	(0.90,2.51)	83.85	7.60	1.00	87.67	5.32	−4.35%
	0.05	(0.73,2.59)	81.96	7.01	0.99	85.99	5.32	−4.68%
	0.04	(0.52,2.73)	79.83	6.37	0.90	84.00	5.24	−4.95%
	0.03	(0.44,2.79)	77.02	5.24	0.79	81.54	5.13	−5.55%
	0.02	(0.31,2.82)	73.25	4.62	0.64	78.39	4.95	−6.55%
γ	0.45	(0.60,2.64)	78.32	5.40	0.94	81.27	4.68	−3.62%
	0.40	(0.56,2.68)	79.00	5.81	0.92	82.29	4.93	−4.00%
	0.35	(0.52,2.73)	79.83	6.37	0.90	84.00	5.24	−4.95%
	0.30	(0.49,2.78)	80.89	7.20	0.89	85.87	5.67	−5.80%
	0.25	(0.47,2.84)	82.28	8.49	0.87	87.98	6.28	−6.48%
C_{rm}	200	(0.84,3.18)	83.03	6.22	0.90	84.00	5.24	−1.15%
	175	(0.73,2.94)	81.60	6.36	0.90	84.00	5.24	−2.85%
	150	(0.52,2.73)	79.83	6.37	0.90	84.00	5.24	−4.95%
	125	(0,2.35)	77.08	6.16	0.90	84.00	5.24	−8.23%
	100	(0,2.19)	73.71	6.80	0.90	84.00	5.24	−12.25%

Table 6
The effect of parameters and comparison of two maintenance models when $\omega_0 = 4.4$.

Parameter	Value	(p_m^*, k^*)	TC^*	W^*	$\widetilde{p_m}^*$	$\widetilde{TC}^*$	$\widetilde{W}^*$	Δ
λ	0.75	(0.37,5.10)	117.89	4.40	0.02	118.21	4.40	−0.27%
	0.7	(0.51,4.62)	96.64	4.40	0.16	96.79	4.39	−0.16%
	0.65	(0.69,4.09)	87.25	4.40	0.33	87.34	4.39	−0.11%
	0.6	(0.73,3.95)	82.54	4.40	0.53	82.60	4.39	−0.08%
	0.55	(0.08,3.20)	76.53	4.40	0.76	80.16	4.39	−4.53%
μ	1.1	(0.04,2.99)	77.78	4.40	0.72	81.26	4.39	−4.28%
	1.05	(0.12,3.47)	81.03	4.40	0.52	83.20	4.40	−2.61%
	1	(0.69,4.09)	87.25	4.40	0.33	87.34	4.39	−0.11%
	0.95	(0.62,4.83)	96.44	4.40	0.16	96.52	4.39	−0.09%
	0.9	(0.56,5.12)	120.43	4.40	0.01	120.49	4.39	−0.05%
β	0.06	(1,5.07)	92.17	4.40	0.33	92.22	4.38	−0.06%
	0.05	(0.82,4.65)	89.97	4.40	0.33	90.01	4.39	−0.05%
	0.04	(0.69,4.09)	87.25	4.40	0.33	87.34	4.39	−0.11%
	0.03	(0.09,3.47)	81.08	4.40	0.33	84.03	4.39	−3.51%
	0.02	(0.02,3.29)	74.76	4.40	0.32	79.79	4.40	−6.30%
γ	0.45	(0.05,3.13)	78.42	4.40	0.61	83.29	4.39	−5.84%
	0.4	(0.11,3.84)	83.32	4.40	0.45	84.96	4.38	−1.93%
	0.35	(0.69,4.09)	87.25	4.40	0.33	87.34	4.39	−0.11%
	0.3	(0.57,4.93)	90.75	4.40	0.23	90.84	4.38	−0.10%
	0.25	(0.51,5.15)	95.65	4.40	0.15	95.73	4.35	−0.08%
C_{rm}	200	(0.98,4.83)	87.29	4.40	0.33	87.34	4.39	−0.06%
	175	(0.75,4.47)	87.27	4.40	0.33	87.34	4.39	−0.08%
	150	(0.69,4.09)	87.25	4.40	0.33	87.34	4.39	−0.11%
	125	(0.27,3.84)	84.44	4.40	0.33	87.34	4.39	−3.32%
	100	(0,3.56)	82.25	4.40	0.33	87.34	4.39	−5.83%

5.3. Effects of parameters

In this subsection, we first examine the effects of parameters on the optimal maintenance policy for the CT scanner problem and demonstrate the efficiency of the control chart. The same setting of the basic level is remained. For each parameter, we will change its value multiple times while remaining other parameters as constants to evaluate the associated optimal decisions. To measure the effectiveness of the control chart, the following efficiency improvement index should be defined first.

$$\Delta\ (\%) = \frac{TC - \widetilde{TC}}{\widetilde{TC}} \times 100\%,$$

in which $\widetilde{TC}$ is the long-run expected cost per unit time under the maintenance model without control chart which can be readily reduced from the model studied in Section 3 by setting $k = \infty$ ($\alpha_i = 0$).

Obviously, compared to the case without control chart, Δ represents the percentage change in long-run expected cost per unit time when the control chart is implemented (under the same service level requirement). Moreover, to discuss the problem under different scenarios, two sojourn time limitations are selected, i.e., $\omega_0 = 10$ and $\omega_0 = 4.4$, to analyze the problem. The corresponding computational results are presented in Tables 5 and 6, respectively. $\widetilde{p_m}^*$ and $\widetilde{W}^*$ are the optimal maintenance policy under the model without control chart($k = \infty$) and corresponding expected sojourn time, respectively.

CASE I: When the service level constraint $\omega_0 = 10$, from Table 5, we can observe the following trends.

- p_m^* is nonincreasing in λ and μ, while k^* is nondecreasing in λ and μ. The reason is that when the control chart is implemented in the service system, the number of sampling per unit time will depend on the customer arrival rate and the service rate. The increase in customer arrival rate or service rate will lead to a higher sampling frequency (this increases the outlier detection rate). Thus, to reduce the maintenance cost and shorten the customer delay time, the control limits of the control chart should be expanded and the frequency of PM should be properly lowered. In addition, it can be seen that TC^* decreases as λ or μ increases.
- Inversely, p_m^* is nondecreasing in β and γ, while k^* is nonincreasing in β and γ. However, TC^* increases (decreases) as β (γ) increases. The result is intuitive, because the frequencies of both the PM and the RM should be increased if the system is more easy to occur glitches. One the other hand, to guarantee the service level, the maintenance frequency needs to be reduced when the expected maintenance time becomes longer.
- It can be seen that all of p_m^*, k^* and TC^* increase as C_{rm} increases. Because when the RM cost gets high, to balance the maintenance cost and the energy cost, one needs to reduce the frequency of the RM while increasing that of the PM. The increased RM cost also has a direct impact on the long-run expected cost per unit time.
- Notice that the improvement index Δ is always negative for all the cases in Table 5. This means that, with the same service level constraint, the operational cost can achieve a significant reduction by implementing the control chart to monitor the service facility. Therefore the control chart indeed plays an important role in maintenance when the traffic intensity of system is large (see the left-hand side of Eq. (7)).

CASE II: If service level constraint becomes more strict, i.e., $\omega_0 = 4.4$, from Table 6, different observations can be found as follows.

- As λ or β increases (μ or γ decreases) and passes a threshold value, p_m^* will suddenly jump up to a relative large level (e.g., from 0.01 to 0.1). Because, when the traffic intensity is small, the maintenance activity mainly relies on the RM and the need of the PM is not high. However, compared to the PM, the high-frequency RM will contribute a much longer waiting time to the system. Thus, when the traffic intensity becomes large, to satisfy the small sojourn time requirement, the system has to uplift the PM frequency while reducing the RM frequency. Meanwhile, the PM should be called in a rush to substitute the time-consuming RM.
- k^* is increasing in λ, β and C_{rm} while decreasing in μ and γ. For parameters β, μ and γ, the changing tendencies of k^* are in

the opposite direction, compared to that in Table 5. The reason is similar to the above observation. When the traffic intensity of system becomes large enough, to satisfy the small sojourn time limit, the frequency of the RM should be sharply decreased.

- Again, the improvement index Δ still remains negative in Table 6. However, the changing trend of Δ is in the opposite direction (except C_{rm}), compared to that in Table 5 (i.e., the bigger traffic intensity is, the smaller Δ will be). To satisfy the small sojourn time limitation, the system has to sacrifice the operation cost.

6. Discussion and concluding remarks

In this paper, a queueing system is developed to characterize a single server with different energy consumption levels in different running states. Two types of maintenance activities are implemented for the server, i.e., the planned maintenance and the reactive maintenance. The frequency of planned maintenance is indicated by a proportion parameter at the beginning of each idle period, and the reactive maintenance is called by the Shewhart's individual control chart. To capture the energy-delay tradeoff, we introduce an optimal maintenance policy to minimize the long-run expected energy consumption and maintenance cost per unit time of the system under a service level constraint.

The proposed model is analyzed for a special case first to gain more insights of the problem. Then, it is extended to more general situations, i.e., multi-state degradation process and general arrival processes and/or service time distributions. However, the optimal maintenance policy is difficult, if not impossible, to be obtained in closed-form expressions. Instead, we conduct numerical experiments to investigate the optimal policy. The results demonstrate that the proposed maintenance model is robust and performs well even the parameters are not accurately estimated, and is superior to the system without monitoring process. Moreover, several managerial insights regarding how to determine an appropriate maintenance policy are investigated. When the customers are not sensitive to the waiting time, as the customer arrival rate and service rate increase, the system will only increase the optimal policy (p_m^*, k^*) within small ranges to balance the energy-delay tradeoff. In addition, if the glitches occur not so frequent and the expected maintenance time is long, the optimal policy suggests a low p_m^* and a high k^* to guarantee the service efficiency. On the opposite side, if the customers are very sensitive to the delay, the values of p_m^* and k^* show enlarging trends to balance the energy-delay tradeoff. Finally, when the sojourn time constraint is loose, the control chart plays an important role in maintenance when the traffic intensity is large, but becomes trivial when the sojourn time requirement is strict.

Finally, several extensions of the proposed work are desired to be further investigated. The steady-state-performance measures used in our paper cannot describe the time-varying parameters throughout the day. Thus, these measures are inappropriate for some systems which may not reach the stable states, e.g., elevators. For the future research, we believe that employing transient performance measures to investigate the maintenance of unstable systems will be an interesting and practical topic. In addition, this study assumes that the energy consumption e of the system is independent of the service rate μ. However, under some circumstances, e may depend on the service rate, e.g., the train consumes more power when its speed is high. One can relax this assumption by treating e as a function of the service rate μ, and adopt μ as another decision variable. Last but not least, extending the model to a two-component series or parallel system is a promising direction as well.

Acknowledgement

The authors would like to thank the Associate Editor and two anonymous reviewers for their valuable comments and suggestions that helped considerably improve the quality of the manuscript. The research of Wenhui Zhou and Zhibin Zheng is supported by the NSFC under 71571070 and 71271089 and by the Natural Science Foundation of Guangdong Province under 2015A030311032; the research of Wei Xie is supported by the NSFC under 71601079 and by the Natural Science Foundation of Guangdong Province under 2016A030310415.

Appendix

Proof of Theorem 1. To solve the balance equations in Eqs. (1)–(6), we define the PGFs for the stationary probability $\pi_{i,j}$ as

$$P_0(z)=\sum_{i=0}^{\infty}\pi_{i,0}z^i,\quad P_1(z)=\sum_{i=0}^{\infty}\pi_{i,1}z^i,\quad P_2(z)=\sum_{i=0}^{\infty}\pi_{i,2}z^i.$$

According to Eqs. (4), (5) and (6) (multiplied by z^i and do the summation for all i), we have

$$\begin{aligned}&[(\gamma+\lambda)z-\lambda z^2]P_2(z)=\alpha_0\mu P_0(z)+\alpha_1\mu P_1(z)+(\gamma+\lambda)z\pi_{0,2}\\&\quad-\alpha_0\mu z\pi_{1,0}-\alpha_0\mu\pi_{0,0}-\alpha_1\mu z\pi_{1,1}-\alpha_1\mu\pi_{0,1},\end{aligned}\tag{16}$$

$$\begin{aligned}&[(\lambda+\beta+\mu)z-\lambda z^2-(1-\alpha_0)\mu]P_0(z)\\&\quad=\gamma zP_2(z)+[(\lambda+\beta+\mu)z-(1-\alpha_0)\mu]\pi_{0,0}\\&\quad-(1-\alpha_0)\mu z\pi_{1,0}-\gamma z\pi_{0,2},\end{aligned}\tag{17}$$

$$\begin{aligned}&[(\lambda+\mu)z-\lambda z^2-(1-\alpha_0)\mu]P_1(z)\\&\quad=\beta zP_0(z)+[(\lambda+\mu)z-(1-\alpha_1)\mu]\pi_{0,1}\\&\quad-(1-\alpha_1)\mu z\pi_{1,1}-\beta z\pi_{0,0}.\end{aligned}\tag{18}$$

Then, substitute Eqs. (1) and (2) into Eqs. (17) and (18), after some algebra, one has

$$P_2(z)=\frac{A_0(z)P_0(z)-C_0(z)\pi_{0,0}+D_0z\pi_{1,0}}{\gamma z},\tag{19}$$

$$P_1(z)=\frac{\beta zP_0(z)+C_1(z)\pi_{0,1}-D_1z\pi_{1,1}}{B(z)}.\tag{20}$$

Solving Eqs. (16), (19) and (20), we can obtain

$$P_0(z)=\frac{\mu[\alpha_0\lambda-F(z)]B(z)\pi_{0,0}-\alpha_1\mu\lambda\gamma z\pi_{0,1}-\lambda B(z)D_0z\pi_{1,0}+\gamma D_1E(z)z\pi_{1,1}}{G(z)},\tag{21}$$

$$\begin{aligned}P_1(z)=&\frac{\mu\beta[\alpha_0\lambda-F(z)]B(z)z\pi_{0,0}+[C_1(z)G(z)-\alpha_1\mu\lambda\beta\gamma z^2]\pi_{0,1}}{B(z)G(z)}\\&-\frac{\lambda\beta B(z)D_0z^2\pi_{1,0}-D_1[\beta\gamma E(z)z-G(z)]z\pi_{1,1}}{B(z)G(z)},\end{aligned}\tag{22}$$

$$\begin{aligned}P_2(z)=&\frac{[\mu(\alpha_0\lambda-F(z))A_0(z)B(z)-C_0(z)G(z)]\pi_{0,0}-\alpha_1\mu\lambda\gamma A_0(z)z\pi_{0,1}}{\gamma zG(z)}\\&+\frac{\gamma A_1(z)D_0E(z)z\pi_{1,0}+\gamma A_0(z)D_1E(z)z\pi_{1,1}}{\gamma zG(z)}.\end{aligned}\tag{23}$$

It can be seen that the expressions of $P_0(z)$, $P_1(z)$ and $P_2(z)$ only contain four unknown probabilities, i.e., $\pi_{0,0}$, $\pi_{0,1}$, $\pi_{1,0}$ and $\pi_{1,1}$, thus, we only need to find four equations to determine them.

To find the first equation, we can examine the stability of the system. According to the L'Hospital's rule, the probabilities that the system is in the normal state, the failure state and the

maintenance state can be computed as

$$p_0 = \sum_{i=0}^{\infty} \pi_{i,0} = \lim_{z\to 1} P_0(z), \quad p_1 = \sum_{i=0}^{\infty} \pi_{i,1} = \lim_{z\to 1} P_1(z),$$

$$p_2 = \sum_{i=0}^{\infty} \pi_{i,1} = \lim_{z\to 1} P_2(z),$$

which yield

$$p_0 = \frac{\alpha_1\mu^2(\alpha_0\lambda-\gamma)\pi_{0,0} - \alpha_1\mu\lambda\gamma\pi_{0,1} - \alpha_1\mu\lambda D_0\pi_{1,0} + \gamma D_1E_1\pi_{1,1}}{G_1}, \tag{24}$$

$$p_1 = \frac{\alpha_1\mu^2\beta(\alpha_0\lambda-\gamma)\pi_{0,0} + \alpha_1\mu(G_1-\lambda\beta\gamma)\pi_{0,1} - \alpha_1\mu\lambda\beta D_0\pi_{1,0} + D_1(\beta\gamma E_1 - G_1)\pi_{1,1}}{\alpha_1\mu G_1}, \tag{25}$$

$$p_2 = \frac{[\alpha_1\mu^2(\alpha_0\lambda-\gamma)H_0 - \alpha_0\mu G_1]\pi_{0,0} - \alpha_1\mu\lambda\gamma H_0\pi_{0,1} + \gamma D_0E_1H_1\pi_{1,0} + \gamma D_1E_1H_0\pi_{1,1}}{\gamma G_1}, \tag{26}$$

where

$$H_i = \alpha_i\mu + \beta, (i = 0, 1),$$
$$G_1 = G(1) = E_1H_1\gamma + \alpha_1\mu\lambda H_0,$$
$$E_1 = E(1) = \lambda - \mu.$$

Because $p_0 + p_1 + p_2 = 1$, the first equation is obtained as

$$K_{00}\pi_{0,0} + K_{01}\pi_{0,1} + K_{10}\pi_{1,0} + K_{11}\pi_{1,1} = \alpha_1\mu\gamma G_1. \tag{27}$$

When $z \le 1$, the generating functions $p_0(z)$, $p_1(z)$ and $p_2(z)$ are convergent. The numerator of $L(Z)$ must be equal to zero when $G(z_0) = 0$. Thus, the second equation can be formulated as

$$\mu[\alpha_0\lambda - F(z_0)]B(z_0)\pi_{0,0} - \alpha_1\mu\lambda\gamma z_0\pi_{0,1} - \lambda B(z_0)D_0z_0\pi_{1,0} + \gamma D_1E(z_0)z_0\pi_{1,1} = 0. \tag{28}$$

Furthermore, based on Eqs. (1), (2) and (3), it is easy to obtain the third and fourth equations

$$\pi_{1,0} = \frac{(\lambda+\beta)(\lambda+\gamma)\pi_{0,0} - (\alpha_1\mu + D_1)\gamma\pi_{1,1}}{(\mu\gamma + I_0\lambda)}, \tag{29}$$

$$\pi_{0,1} = \frac{\beta\pi_{0,0} + I_1\pi_{1,1}}{\lambda}. \tag{30}$$

Then, substitute Eqs. (29) and (30) into Eqs. (27) and (28), we have

$$a_1\pi_{0,0} + a_2\pi_{1,1} = \alpha_1\mu\lambda\gamma(\mu\gamma + I_0\lambda)G_1, \tag{31}$$

$$a_3\pi_{0,0} + a_4\pi_{1,1} = 0. \tag{32}$$

Solving Eqs. (29)–(32) yields the probabilities $\pi_{0,0}$, $\pi_{0,1}$, $\pi_{1,0}$ and $\pi_{1,1}$ in Eqs. (8)–(11). Finally, substitute Eqs. (8)–(11) into Eqs. (21)–(23), one can have the PGFs $P_0(z)$, $P_1(z)$ and $P_2(z)$. This completes the proof. □

Proof of Proposition 1. To prove this proposition, we just need to substitute Eqs. (8)–(11) into Eqs. (24)–(26). This finishes the proof. □

References

Allon, G., & Federgruen, A. (2007). Competition in service industries. *Operations Research, 55*(1), 37–55.

Alsyouf, I. (2006). Measuring maintenance performance using a balanced scorecard approach. *Journal of Quality in Maintenance Engineering, 12*(2), 133–149.

Ang, B. W., & Fwa, T. F. (1989). A study on the fuel-consumption characteristics of public buses. *Energy, 14*(12), 797C803.

Artalejo, J. R., Gómez-Corral, A., & Neuts, M. F. (2001). Analysis of multiserver queues with constant retrial rate. *European Journal of Operational Research, 135*(3), 569–581.

Asmussen, S., & Koole, G. (1993). Marked point processes as limits of Markovian arrival streams. *Journal of Applied Probability, 30*(2), 365–372.

Ben-Daya, M. (1999). Integrated production maintenance and quality model for imperfect processes. *IIE Transactions, 31*(6), 491–501.

Ben-Daya, M., & Rahim, M. (2000). Effect of maintenance on the economic design of x-control chart. *European Journal of Operational Research, 120*(1), 131–143.

Carnero, M. C. (2005). Selection of diagnostic techniques and instrumentation in a predictive maintenance program. a case study. *Decision Support Systems, 38*(4), 539–555.

Cassady, C. R., Bowden, O. R., Liew, L., & Pohl, E. A. (2005). Combining preventive maintenance and statistical process control: a preliminary investigation. *IIE Transactions, 32*(6), 471–478.

Bana e Costa, C. A., Carnero, M. C., & Duarte, M. (2005). A multi-criteria model for auditing a predictive maintenance programme. *European Journal of Operational Research, 217*(2), 381–393.

Delia, M.-C., & Rafael, P.-O. (2008). A maintenance model with failures and inspection following Markovian arrival processes and two repair modes. *European Journal of Operational Research, 186*(2), 694–707.

Dzial, T., Breuer, L., Soares, A. d. S., Latouche, G., & Remiche, M.-A. (2005). Fluid queues to solve jump processes. *Performance Evaluation, 62*(1), 132–146.

Federgruen, A., & So, K. C. (1990). Optimal maintenance policies for single-server queueing systems subject to breakdowns. *Operations Research, 38*(2), 330–343.

Gandhi, A., Gupta, V., Harchol-Balter, M., & Kozuch, M. A. (2010). Optimality analysis of energy-performance trade-off for server farm management. *Performance Evaluation, 67*(11), 1155–1171.

Gonzalez, R., & Horowitz, M. (1996). Energy dissipation in general purpose microprocessors. *IEEE Journal of Solid-State Circuits, 31*(9), 1277–1284.

He, Q.-M., & Neuts, M. F. (1998). Markov chains with marked transitions. *Stochastic Processes and Their Applications, 74*(1), 37–52.

Heo, J., Henriksson, D., Liu, X., & Abdelzaher, T. (2007). Integrating adaptive components: An emerging challenge in performance-adaptive systems and a server farm case-study. In *28th IEEE international real-time systems symposium, 2007. RTSS 2007* (pp. 227–238). IEEE.

Jafari, L., & Makis, V. (2016). Optimal lot-sizing and maintenance policy for a partially observable production system. *Computers & Industrial Engineering, 93*, 88–98.

Juang, P., Wu, Q., Peh, L. S., Martonosi, M., & Clark, D. W. (2005). Coordinated, distributed, formal energy management of chip multiprocessors. *Proceedings of the 2005 international symposium on low power electronics and design ISLPED '05* (pp. 127–130).

Kang, C. W., Abbaspour, S., & Pedram, M. (2003). Buffer sizing for minimum energy-delay product by using an approximating polynomial. *ACM Great Lakes Symposium on VLSI* (pp. 112–115).

Kaufman, D. L., & Lewis, M. E. (2007). Machine maintenance with workload considerations. *Naval Research Logistics, 54*(7), 750–766.

Kin, J., Gupta, M., & Mangione-Smith, W. H. (1997). The filter cache: an energy efficient memory structure. *2012 45th Annual IEEE/ACM international symposium on microarchitecture* p. 184.

Kuo, Y. (2006). Optimal adaptive control policy for joint machine maintenance and product quality control. *European Journal of Operational Research, 171*(2), 586–597.

Lam, Y., Zhang, Y. L., & Liu, Q. (2006). A geometric process model for M / M /1 queueing system with a repairable service station. *European Journal of Operational Research, 168*(1), 100–121.

Latouche, G., & Ramaswami, V. (1999). *Introduction to matrix analytic methods in stochastic modeling*: vol.5. SIAM.

Li, Q. L., Ying, Y., & Zhao, Y. Q. (2006). A BMAP/G/1 retrial queue with a server subject to breakdowns and repairs. *Annals of Operations Research, 141*(1), 233–270.

Liao, H., Xie, W., & Jin, T. (2013). Managing operational availability via integrated redundancy allocation and spare parts provisioning. In *IIE annual conference proceedings* (p. 3225). Institute of Industrial Engineers-Publisher.

Linderman, K., McKone-Sweet, K. E., & Anderson, J. C. (2005). An integrated systems approach to process control and maintenance. *European Journal of Operational Research, 164*(2), 324–340.

Liu, L., Yu, M., Ma, Y., & Tu, Y. (2013). Economic and economic-statistical designs of an control chart for two-unit series systems with condition-based maintenance. *European Journal of Operational Research, 226*(3), 491–499.

Makis, V. (2008). Multivariate Bayesian control chart. *Operations Research, 56*(2), 487–496.

Montoro-Cazorla, D., Pérez-Ocón, R., & del Carmen Segovia, M. (2009). Replacement policy in a system under shocks following a Markovian arrival process. *Reliability Engineering & System Safety, 94*(2), 497–502.

Neuts, M. F. (1981). *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Courier Corporation.

Panagiotidou, S., & Nenes, G. (2009). An economically designed, integrated quality and maintenance model using an adaptive Shewhart chart. *Reliability Engineering & System Safety, 94*(3), 732–741.

Panagiotidou, S., & Tagaras, G. (2007). Optimal preventive maintenance for equipment with two quality states and general failure time distributions. *European Journal of Operational Research, 180*(1), 329–353.

Pandey, D., Kulkarni, M. S., & Vrat, P. (2010). Joint consideration of production scheduling, maintenance and quality policies: a review and conceptual framework. *International Journal of Advanced Operations Management, 2*(1), 1–24.

Peng, H., & van Houtum, G.-J. (2016). Joint optimization of condition-based maintenance and production lot-sizing. *European Journal of Operational Research, 253*(1), 94–107.

Rahim, M. (1994). Joint determination of production quantity, inspection schedule, and control chart design. *IIE Transactions, 26*(6), 2–11.
Soares, d. S., & Ana, G. L. (2006). Matrix-analytic methods for fluid queues with finite buffers. *Performance Evaluation, 63*(4), 295–314.
Stan, M. R., & Skadron, K. (2003). Guest editors' introduction: Power-aware computing. *Computer, 36*(12), 35–38.
Taleb, S., & Aissani, A. (2016). Preventive maintenance in an unreliable m/g/1 retrial queue with persistent and impatient customers. *Annals of Operations Research, 247*(1), 291–317.
Keizer, M. C. A. O., Teunter, R. H., & Veldman, J. (2017). Joint condition-based maintenance and inventory optimization for systems with multiple components. *European Journal of Operational Research, 257*(1), 209–222.
Wang, W. (2012). A simulation-based multivariate Bayesian control chart for real time condition-based maintenance of complex systems. *European Journal of Operational Research, 218*(3), 726–734.
Wartenhorst, P. (1995). N parallel queueing systems with server breakdown and repair. *European Journal of Operational Research, 82*(2), 302–322.
Xie, W., Liao, H., & Jin, T. (2014). Maximizing system availability through joint decision on component redundancy and spares inventory. *European Journal of Operational Research, 237*(1), 164–176.
Yang, W. S., Lim, D. E., & Chae, K. C. (2009). Maintenance of deteriorating single server queues with random shocks. *Computers & Industrial Engineering, 57*(4), 1404–1406.
Yin, H., Zhang, G., Zhu, H., Deng, Y., & He, F. (2015). An integrated model of statistical process control and maintenance based on the delayed monitoring. *Reliability Engineering & System Safety, 133*, 323–333.
Zhou, W.-H., & Zhu, G.-L. (2008). Economic design of integrated model of control chart and maintenance management. *Mathematical and Computer Modelling, 47*(11), 1389–1395.