Queueing Syst (2017) 86:241–260
DOI 10.1007/s11134-017-9533-2

Batch arrival single-server queue with variable service speed and setup time

Moeko Yajima[1] · **Tuan Phung-Duc**[2]

Received: 29 September 2016 / Revised: 15 May 2017 / Published online: 5 June 2017

Abstract In this paper, we consider an M^X/M/1/SET-VARI queue which has batch arrivals, variable service speed and setup time. Our model is motivated by power-aware servers in data centers where dynamic scaling techniques are used. The service speed of the server is proportional to the number of jobs in the system. The contribution of our paper is threefold. First, we obtain the necessary and sufficient condition for the stability of the system. Second, we derive an expression for the probability generating function of the number of jobs in the system. Third, our main contribution is the derivation of the Laplace–Stieltjes transform (LST) of the sojourn time distribution, which is obtained in series form involving infinite-dimensional matrices. In this model, since the service speed varies upon arrivals and departures of jobs, the sojourn time of a tagged job is affected by the batches that arrive after it. This makes the derivation of the LST of the sojourn time complex and challenging. In addition, we present some numerical examples to show the trade-off between the mean sojourn time (response time) and the energy consumption. Using the numerical inverse Laplace–Stieltjes transform, we also obtain the sojourn time distribution, which can be used for setting the service-level agreement in data centers.

✉ Moeko Yajima
yajima.m.ad@m.titech.ac.jp

Tuan Phung-Duc
tuan@sk.tsukuba.ac.jp

[1] Department of Mathematical and Computing Sciences, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8552, Japan

[2] Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Keywords Batch arrival · Variable service speed · Setup time · State dependent · Sojourn time · Stability condition

Mathematics Subject Classification 60K25

1 Introduction

In this paper, we consider a single-server queue with batch Poisson arrivals, variable service speed and setup time. Our model is motivated by power-aware servers in data centers [9,10,17]. The CPU of a server is able to process at multiple speeds by using either frequency scaling [13] or dynamic voltage and frequency scaling (DVFS) techniques [11,15]. In recent years, CPUs with variable speed have become popular because they can save energy consumption while keeping acceptable response time for jobs. The server can automatically adjust its speed according to the workload in the system. By doing so, the power consumption is small at low workload and is large at high workload.

In this paper, we assume that jobs arrive at the system in batches according to a Poisson process and that the arrival process is independent of the state of the system. The service requirement of each job in a batch is independently and identically distributed (i.i.d.) with an exponential distribution. The service speed of the server is instantaneously adapted according to the number of jobs in the system. In particular, the service rate of the server is proportional to the number of jobs in the system. Furthermore, the server is turned off immediately after becoming empty in order to save energy consumption. At the moment when a batch arrives at an empty system, the OFF server is turned on. However, some exponentially distributed setup time is needed in order to reactivate the OFF server. We call the above queueing model the M^X/M/1/SET-VARI queue where SET and VARI stand for setup and variable service rate, respectively.

The contribution of this paper is threefold. First, we obtain the necessary and sufficient condition for the existence of the unique stationary queue length distribution, which we call the stability condition hereafter. We show that the stability condition of our model is that the logarithmic moment of the batch size is finite. Interestingly, the system can be stable even if the mean batch size is infinite. Second, we derive the probability generating function (PGF) of the number of jobs in the system. It should be noted that the number of jobs in the system of our model is identical to that of the M^X/M/∞ queue with setup time, which to the best of our knowledge has not been investigated in the literature. Third, we derive the Laplace–Stieltjes transform (LST) of the sojourn time distribution, which is obtained in series form involving infinite-dimensional matrices. The derivation of the sojourn time distribution is challenging because the sojourn time of a tagged job depends on not only the state of the system upon arrival but also on the batches arriving after it. Therefore, the sojourn time distribution cannot be obtained directly from the PGF of the queue length distribution via the distributional Little's law [8].

Our model extends the one proposed by Lu et al. [9] in which an M/M/1/SET-VARI queue was considered. In [9], the solution in terms of infinite series was presented for the stationary queue length distribution. From the queue length distribution, the mean

response time is obtained via Little's law and the mean power consumption is obtained. These metrics are used in [9] to find the energy-response trade-off. However, the sojourn time distribution was not considered in [9]. Baba [2] considered the M^X/M/1 queue with setup time where the processing speed of the server is fixed. He derived the PGF of the number of jobs in the system and the LST of the sojourn time distribution. Adan and D'Auria [1] considered a single-server queueing system where jobs arrive according to a Poisson arrival stream, the service requirements of jobs follow the exponential distribution with mean 1 and the service rate of the server is controlled by a threshold. They derived the stationary distribution of the number of jobs in the system and the LST of the sojourn time distribution in explicit form. The sojourn time distribution of our model is derived using first-step analysis, which is also adopted by Adan and D' Auria [1]. The difference is that the underlying Markov chain in Adan and D' Auria [1] is homogeneous after a threshold, while our underlying Markov chain is spatially nonhomogeneous. As a result, the former allows explicit expression while our formulae involve inverse mappings of infinite matrices.

The remainder of this paper is organized as follows: In Sect. 2, we describe the M^X/M/1/SET-VARI queue in detail. In Sect. 3, we derive the stability condition. In Sect. 4, we derive the PGF of the number of jobs in the system in an integral but computable form. In Sect. 5, we derive the LST of the sojourn time distribution. In Sect. 6, we present numerical experiments showing the energy-performance trade-off and the sojourn time distribution by numerically inverting the Laplace–Stieltjes transform. Finally, in Sect. 7, we present the conclusion of this paper and future work.

2 Model

In this section, we describe our queueing model, the M^X/M/1/SET-VARI queue, in detail. The M^X/M/1/SET-VARI queue has a single-server operating under the FCFS (First Come First Served) service discipline and an infinite buffer space. Batches of jobs arrive at the system according to a Poisson process with rate λ. The numbers of jobs in batches are i.i.d., where X is the batch size with distribution $x_i = \mathsf{P}(X = i)$ for $i \in \mathbb{N} := \{1, 2, \dots\}$ and the PGF (Probability Generating Function) is denoted by $\widehat{X}(z) := \sum_{k\geq 1} x_k z^k$.

The special feature of our model is that the service speed of the server is proportional to the number of jobs in the system. In particular, the service rate is $n\mu$, provided that the number of jobs in the system is n. This is equivalent to the following assumption: The service requirements of jobs are i.i.d. with exponential distribution with mean 1. The basic speed of the server when there is one job in the system is given by $\mu \in (0, \infty)$. When there are n jobs in the system, the speed of the server is scaled up to $n\mu$. Thus, when there are n jobs in the system, the residual sojourn time of the ongoing job follows the exponential distribution with mean $(n\mu)^{-1}$ due to the memoryless property of exponential distributions.

In order to save energy, the server is turned off immediately if the system becomes empty upon a service completion. Furthermore, when a batch arrives at the empty system, the server is turned on. However, the server needs some setup time before it can process jobs. Therefore, if a batch arrives at the empty system, it has to wait

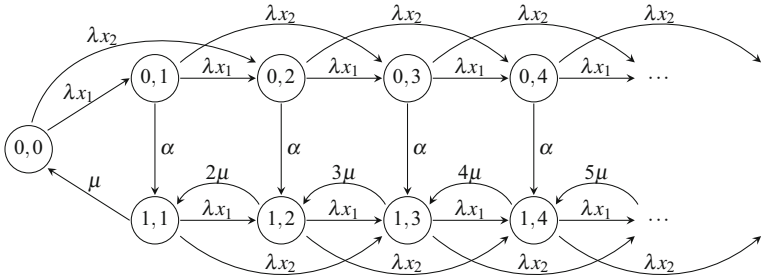

Fig. 1 Transition diagram of the case $x_1 + x_2 = 1$

until the setup time finishes. We assume that the setup time follows the exponential distribution with mean α^{-1}. During the setup time, the server cannot serve a job but consumes energy.

Let $I(t)$ and $N(t)$ denote the state of the server and the number of jobs in the system, respectively, at time t. When the server is off or in the setup process, $I(t) = 0$, and when the server is processing a job, $I(t) = 1$. In the current setting, the joint stochastic process $\{Z(t) := (I(t), N(t)); t \geq 0\}$ is an irreducible continuous-time Markov chain with state space $\mathscr{S} = \{(0, j);\ j \in \mathbb{Z}_+\} \cup \{(1, j);\ j \in \mathbb{N}\}$, where $\mathbb{Z}_+ := \{0\} \cup \mathbb{N}$. We assume that $x_1 > 0$ so that the Markov chain $\{Z(t)\}$ is irreducible. Figure 1 shows the state transition diagram of this Markov chain for a special case where the maximum batch size is two.

Remark 1 As mentioned in Sect. 1, the PGF of the number of jobs in the system of the M^X/M/1/SET-VARI queue is identical to that of the M^X/M/∞ queue with setup time. However, the sojourn time distributions of these two models may be different because the sojourn time distribution of a tagged job of the latter is determined upon its arrival, while that of the former is affected by future arrivals. Some researchers have studied the M^X/M/∞ queue without setup time. For example, Shanbhag [14] derived moment generating functions of some performance measures, for example, the number of jobs in the system and the sojourn time. Cong [3] derived the stability condition.

3 The stability condition

In this section, we derive the stability condition of our model. Because $\{Z(t); t \geq 0\}$ is an irreducible and regular Markov chain, $\{Z(t)\}$ is positive recurrent if and only if a unique stationary distribution exists.

It follows from Theorem 1 that the stability condition of the M^X/M/1/SET-VARI queue is that the logarithmic moment of the batch size is finite. Cong [3] derived the same stability condition for the special case of the M^X/M/∞ queue without setup time. The addition of the setup time does not change the stability of the system. This is intuitively clear because the effects of setup times disappear when the system is under heavy load (i.e., large number of jobs present).

It should be noted that the proof of Cong [3] is based on the transient solution, which is derived using the method of collective marks. Here, we prove the stability condition for a more general model than the one in Cong [3] using alternative methods.

Theorem 1 $\{Z(t); t \geq 0\}$ *has a unique stationary distribution if and only if*

$$\mathsf{E}[\log(X+1)] < \infty. \tag{3.1}$$

Proof It should be noted that $\{Z(t)\}$ is positive recurrent if and only if there exists an invariant measure $\boldsymbol{\xi} := (\xi_{i,j})_{(i,j)\in\mathscr{S}}$ of $\{Z(t)\}$ such that $\xi_{i,j} > 0$ for $(i,j) \in \mathscr{S}$ and $\sum_{(i,j)\in\mathscr{S}} \xi_{i,j} < \infty$ [12]. We define the generating functions $\widehat{\xi}_0(z)$ and $\widehat{\xi}_1(z)$ as follows:

$$\widehat{\xi}_0(z) = \sum_{j=0}^{\infty} \xi_{0,j} z^j, \quad \widehat{\xi}_1(z) = \sum_{j=1}^{\infty} \xi_{1,j} z^j.$$

The invariant measure $\boldsymbol{\xi}$ satisfies the following balance equations:

$$\lambda \xi_{0,0} = \mu \xi_{1,1}, \tag{3.2}$$

$$(\lambda+\alpha)\xi_{0,j} = \lambda \sum_{k=1}^{j} x_k \xi_{0,j-k}, \quad j = 1, 2, \ldots, \tag{3.3}$$

$$(\lambda+\mu)\xi_{1,1} = \alpha \xi_{0,1} + 2\mu \xi_{1,2}, \tag{3.4}$$

$$(\lambda + j\mu)\xi_{1,j} = \alpha \xi_{0,j} + (1+j)\mu \xi_{1,j+1} + \lambda \sum_{k=1}^{j-1} x_k \xi_{1,j-k}, \quad j = 2, 3, \ldots. \tag{3.5}$$

Multiplying (3.3) by z^j, taking the sum over $j \in \mathbb{N}$, and rearranging the result, we obtain

$$\widehat{\xi}_0(z) = \frac{(\lambda+\alpha)\xi_{0,0}}{\lambda + \alpha - \lambda \widehat{X}(z)}. \tag{3.6}$$

Multiplying (3.4) by z and (3.5) by z^j and taking the sum over $j \geq 2$ yields

$$\begin{aligned} &\lambda \sum_{j=1}^{\infty} \xi_{1,j} z^j + \mu z \sum_{j=1}^{\infty} \xi_{1,j} \left(z^j\right)' \\ &= \alpha \sum_{j=0}^{\infty} \xi_{0,j} z^j - \alpha \xi_{0,0} + \mu \sum_{j=1}^{\infty} \xi_{1,j} \left(z^j\right)' - \mu \xi_{1,1} + \lambda \sum_{k=1}^{\infty} x_k z^k \sum_{j=1}^{\infty} \xi_{1,j} z^j. \end{aligned}$$

Rearranging the above equation, we find that

$$\frac{\mathrm{d}}{\mathrm{d}z} \widehat{\xi}_1(z) = \frac{\lambda}{\mu} q(z) \widehat{\xi}_1(z) + \frac{\lambda}{\mu} q(z) \widehat{\xi}_0(z), \tag{3.7}$$

where $q(z) := (1 - \widehat{X}(z))/(1-z)$. We define $Q(z)$ as the primitive function of $q(z)$ such that $Q(0) = 0$. The solution of (3.7) is given by

$$\widehat{\xi}_1(z) = H(z)\exp\Big(\frac{\lambda}{\mu}Q(z)\Big), \tag{3.8}$$

where $H(z)$ is some function which will be determined later. Differentiating (3.8) and substituting the result into (3.7), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}z}H(z) = \exp\Big(-\frac{\lambda}{\mu}Q(z)\Big)\frac{\lambda}{\mu}q(z)\widehat{\xi}_0(z).$$

It follows from $\widehat{\xi}_1(0) = 0$ and $Q(0) = 0$ that $H(0) = 0$. Therefore, we have

$$H(z) = \int_0^z \exp\Big(-\frac{\lambda}{\mu}Q(u)\Big)\frac{\lambda}{\mu}q(u)\widehat{\xi}_0(u)\mathrm{d}u.$$

Substituting this equation into (3.8), we obtain

$$\widehat{\xi}_1(z) = \exp\Big(\frac{\lambda}{\mu}Q(z)\Big)\Big\{\int_0^z \exp\Big(-\frac{\lambda}{\mu}Q(u)\Big)\frac{\lambda}{\mu}q(u)\widehat{\xi}_0(u)\mathrm{d}u\Big\}. \tag{3.9}$$

It should be noted that (3.6) and (3.9) are equivalent to the system of balance equations (3.2)–(3.6).

Assuming that $\{Z(t)\}$ is positive recurrent, we will prove that (3.1) holds. Thus, there exists an invariant measure $\boldsymbol{\xi}$ such that $\xi_{i,j} > 0$ for $(i, j) \in \mathscr{S}$ and $\sum_{(i,j)\in\mathscr{S}} \xi_{i,j} < \infty$. Therefore, we have

$$\xi_{0,0} \le \widehat{\xi}_0(z), \quad \text{for all } z \in [0, 1], \tag{3.10}$$

$$\widehat{\xi}_0(z) + \widehat{\xi}_1(z) \le \sum_{(i,j)\in\mathscr{S}} \xi_{i,j}, \quad \text{for all } z \in [0, 1]. \tag{3.11}$$

From (3.9) and (3.10), we also have

$$\begin{aligned}\widehat{\xi}_0(z) + \widehat{\xi}_1(z) &\ge \xi_{0,0} + \xi_{0,0}\exp\Big(\frac{\lambda}{\mu}Q(z)\Big)\int_0^z \exp\Big(-\frac{\lambda}{\mu}Q(u)\Big)\frac{\lambda}{\mu}q(u)\mathrm{d}u\\ &= \xi_{0,0} + \xi_{0,0}\exp\Big(\frac{\lambda}{\mu}Q(z)\Big)\int_0^z \frac{\mathrm{d}}{\mathrm{d}u}\Big\{-\exp\Big(-\frac{\lambda}{\mu}Q(u)\Big)\Big\}\mathrm{d}u\\ &= \xi_{0,0}\exp\Big(\frac{\lambda}{\mu}Q(z)\Big), \quad \text{for all } z \in [0, 1].\end{aligned} \tag{3.12}$$

Because of (3.11), $\xi_{i,j} > 0$, $(i, j) \in \mathscr{S}$, and $\sum_{(i,j)\in\mathscr{S}} \xi_{i,j} < \infty$. We can take the limit of (3.12) as $z \uparrow 1$ in order to obtain

$$\sum_{(i,j)\in\mathscr{S}}^{\infty} \xi_{i,j} \ge \xi_{0,0}\exp\Big(\frac{\lambda}{\mu}Q(1)\Big). \tag{3.13}$$

Using $Q(z) = \sum_{j\geq 1} \mathsf{P}[X \geq j]z^j/j$, we can show the following inequality:

$$\begin{aligned} Q(1) &= \sum_{k=1}^{\infty} x_k \left\{ \sum_{j=1}^{k+1} \frac{1}{j} - \log(k+1) \right\} + \mathsf{E}\Big[- \frac{1}{X+1} + \log(X+1) \Big] \\ &> \gamma - \mathsf{E}\Big[\frac{1}{X+1}\Big] + \mathsf{E}[\log(X+1)] \\ &\geq \gamma - 1/2 + \mathsf{E}[\log(X+1)], \end{aligned} \tag{3.14}$$

where γ is Euler's constant [16]. The first inequality in (3.14) is due to

$$\sum_{j=1}^{k+1} \frac{1}{j} - \log(k+1) > \gamma,$$

while the second inequality in (3.14) is because $X \geq 1$. Therefore, it follows from (3.13), (3.18) and $\sum_{(i,j)\in\mathscr{S}} \xi_{i,j} < \infty$ that $\mathsf{E}[\log(X+1)] < \infty$.

Now, assuming that $\mathsf{E}[\log(X+1)] < \infty$, we will prove the existence of a positive invariant measure $\boldsymbol{\xi}$ such that $\sum_{(i,j)\in\mathscr{S}} \xi_{i,j} < \infty$. We select an arbitrary $\xi_{0,0} > 0$. First, we prove that $\xi_{i,j} > 0$ for any $(i, j) \in \mathscr{S}$. We can recursively prove that $\xi_{0,j} > 0$ for $j \in \mathbb{N}$ by using (3.3), and it follows from (3.2) that $\xi_{1,1} > 0$. In addition, comparing the coefficients of z^j on both sides of (3.7), we have, for $j \in \mathbb{N}$,

$$(j+1)\xi_{1,j+1} = \frac{\lambda}{\mu} \sum_{k=0}^{j} \mathsf{P}[X > j-k]\xi_{0,k} + \frac{\lambda}{\mu} \sum_{k=1}^{j} \mathsf{P}[X > j-k]\xi_{1,k}, \tag{3.15}$$

where we have used $q(z) = \sum_{j\geq 0} \mathsf{P}[X > j]z^j$. Due to $\xi_{0,0} > 0$, we can also prove that $\xi_{1,j} > 0$ for $j \geq 2$ by using the recursive formula (3.15). Thus, under the assumption that $\xi_{0,0} > 0$, it follows that $\xi_{i,j} > 0$ for any $(i, j) \in \mathscr{S}$.

Next, we prove that $\sum_{(i,j)\in\mathscr{S}} \xi_{i,j} < \infty$. From (3.6), we have

$$\widehat{\xi}_0(z) \leq \frac{\lambda+\alpha}{\lambda+\alpha-\lambda\widehat{X}(1)}\xi_{0,0} \leq \frac{\lambda+\alpha}{\alpha}\xi_{0,0}, \quad \text{for all } z \in [0,1], \tag{3.16}$$

where the second inequality is to cover the case $\mathsf{P}[X < \infty] < 1$. From (3.9) and (3.16), we also have

$$\begin{aligned} &\widehat{\xi}_0(z) + \widehat{\xi}_1(z) \\ &\quad \leq \xi_{0,0}\frac{\lambda+\alpha}{\alpha} + \xi_{0,0}\frac{\lambda+\alpha}{\alpha} \exp\Big(\frac{\lambda}{\mu}Q(z)\Big) \int_0^z \exp\Big(-\frac{\lambda}{\mu}Q(u)\Big)\frac{\lambda}{\mu}q(u)\mathrm{d}u \\ &\quad = \xi_{0,0}\frac{\lambda+\alpha}{\alpha} + \xi_{0,0}\frac{\lambda+\alpha}{\alpha} \exp\Big(\frac{\lambda}{\mu}Q(z)\Big) \int_0^z \frac{\mathrm{d}}{\mathrm{d}u}\Big\{-\exp\Big(-\frac{\lambda}{\mu}Q(u)\Big)\Big\}\mathrm{d}u \\ &\quad = \xi_{0,0}\frac{\lambda+\alpha}{\alpha} \exp\Big(\frac{\lambda}{\mu}Q(z)\Big), \quad \text{for all } z \in [0,1]. \end{aligned} \tag{3.17}$$

Using $Q(z) = \sum_{j\geq 1} \mathsf{P}[X \geq j] z^j / j$, we can show the following inequality:

$$\begin{aligned} Q(z) &\leq \sum_{j\geq 1} \mathsf{P}[X \geq j] \frac{1}{j} \\ &= \sum_{k=1}^{\infty} x_k \left\{ \sum_{j=1}^{k} \frac{1}{j} - \log(k+1) \right\} + \mathsf{E}[\log(X+1)] \\ &< \gamma + \mathsf{E}[\log(X+1)], \quad \text{for all } z \in [0,1], \end{aligned} \tag{3.18}$$

where the inequality in (3.18) is due to

$$\sum_{j=1}^{k} \frac{1}{j} - \log(k+1) < \gamma.$$

Taking the limit of (3.17) as $z \uparrow 1$, it follows from (3.18) that

$$\lim_{z\uparrow 1} \left\{ \widehat{\xi}_0(z) + \widehat{\xi}_1(z) \right\} \leq \xi_{0,0} \frac{\lambda+\alpha}{\alpha} \exp\left(\frac{\lambda}{\mu} \{\gamma + \mathsf{E}[\log(X+1)]\} \right). \tag{3.19}$$

Note that

$$\begin{aligned} \lim_{z\uparrow 1} \left\{ \widehat{\xi}_0(z) + \widehat{\xi}_1(z) \right\} &= \lim_{z\uparrow 1} \sum_{j=0}^{\infty} \xi_{0,j} z^j + \lim_{z\uparrow 1} \sum_{j=1}^{\infty} \xi_{1,j} z^j \\ &= \sum_{j=0}^{\infty} \lim_{z\uparrow 1} \xi_{0,j} z^j + \sum_{j=1}^{\infty} \lim_{z\uparrow 1} \xi_{1,j} z^j \\ &= \sum_{j=0}^{\infty} \xi_{0,j} + \sum_{j=1}^{\infty} \xi_{1,j}, \end{aligned} \tag{3.20}$$

where the second equation holds because $\xi_{i,j} > 0$ for $(i,j) \in \mathscr{S}$. It follows from (3.19), (3.20), $\xi_{0,0} > 0$ and $\mathsf{E}[\log(X+1)] < \infty$ that there exists an invariant measure $\boldsymbol{\xi}$ such that $\xi_{i,j} > 0$ for $(i,j) \in \mathscr{S}$ and $\sum_{(i,j)\in\mathscr{S}} \xi_{i,j} < \infty$. Therefore $\{Z(t)\}$ is positive recurrent. □

4 The number of jobs in the system

In this section, we consider the number of jobs in the system in steady state, i.e., assuming $\mathsf{E}[\log(X+1)] < \infty$. From Theorem 1, there exists the unique stationary distribution $\boldsymbol{\pi} = (\pi_{i,j})_{i,j\in\mathscr{S}}$, where $\pi_{i,j} := \lim_{t\to\infty} \mathsf{P}[\, I(t) = i, N(t) = j \,]$ for $(i,j) \in \mathscr{S}$. We define the number of jobs in the system in steady state as N, and its PGF as $\widehat{\pi}(z) := \sum_{j=0}^{\infty} \pi_{0,j} z^j + \sum_{j=1}^{\infty} \pi_{1,j} z^j$. The PGF $\widehat{\pi}(z)$ is given by Theorem 2.

Theorem 2 *$\widehat{\pi}(z)$ is given as follows:*

$$\widehat{\pi}(z) = \frac{\exp\left(\frac{\lambda}{\mu}Q(z)\right) + \int_0^z \exp\left(\frac{\lambda}{\mu}\{Q(z) - Q(u)\}\right)\frac{\lambda(\lambda+\alpha)\widehat{X}'(u)}{\{\lambda+\alpha-\lambda\widehat{X}(u)\}^2}\mathrm{d}u}{\exp\left(\frac{\lambda}{\mu}Q(1)\right) + \int_0^1 \exp\left(\frac{\lambda}{\mu}\{Q(1) - Q(u)\}\right)\frac{\lambda(\lambda+\alpha)\widehat{X}'(u)}{\{\lambda+\alpha-\lambda\widehat{X}(u)\}^2}\mathrm{d}u}. \quad (4.1)$$

Proof The stationary distribution $\boldsymbol{\pi}$ is the positive invariant measure which satisfies the normalizing condition, i.e., $\widehat{\pi}(1) = 1$. From (3.6) and (3.9), we have

$$\widehat{\pi}(z) = \frac{\lambda+\alpha}{\lambda+\alpha-\lambda\widehat{X}(z)}\pi_{0,0} + \int_0^z \exp\Big(-\frac{\lambda}{\mu}\big\{Q(z) - Q(u)\big\}\Big)\frac{\lambda}{\mu}q(u)\frac{\lambda+\alpha}{\lambda+\alpha-\lambda\widehat{X}(z)}\pi_{0,0}\mathrm{d}u. \quad (4.2)$$

By partial integration, we obtain

$$\widehat{\pi}(z) = \pi_{0,0}\exp\left(\frac{\lambda}{\mu}Q(z)\right) + \pi_{0,0}\int_0^z \exp\left(\frac{\lambda}{\mu}\big\{Q(z) - Q(u)\big\}\right)\frac{\lambda(\lambda+\alpha)\widehat{X}'(z)}{\{\lambda+\alpha-\lambda\widehat{X}(z)\}^2}\mathrm{d}u. \quad (4.3)$$

From (4.3) and $\widehat{\pi}(1) = 1$, we obtain $\pi_{0,0}$ as follows:

$$\pi_{0,0}^{-1} = \exp\left(\frac{\lambda}{\mu}Q(1)\right) + \int_0^1 \exp\left(\frac{\lambda}{\mu}\big\{Q(1) - Q(u)\big\}\right)\frac{\lambda(\lambda+\alpha)\widehat{X}'(z)}{\{\lambda+\alpha-\lambda\widehat{X}(z)\}^2}\mathrm{d}u. \quad (4.4)$$

Substituting (4.4) into (4.3), we obtain Theorem 2. □

Remark 2 As mentioned in Sect. 1, the number of jobs in the system in our model is identical to that in the $\mathrm{M}^X/\mathrm{M}/\infty$ queue with setup time. Simplifying equation (2.9) in Shanbhag [14] for the system without setup time, we know that the PGF of the number of customers for that system, denoted by $\widehat{\pi}^*(z)$, can be obtained as follows:

$$\widehat{\pi}^*(z) = \exp\left(\frac{\lambda}{\mu}\{Q(z) - Q(1)\}\right). \quad (4.5)$$

It is easy to see that (4.1) tends to (4.5) as $\alpha \to \infty$.

We can easily obtain the average number of jobs in the system.

Corollary 1 *Assuming that $\mathsf{E}[X] < \infty$, $\mathsf{E}[N]$ is given as follows:*

$$\mathsf{E}[N] = \frac{\lambda\mathsf{E}[X]}{\mu}\left\{\frac{\lambda+\alpha}{\alpha}\frac{\mu}{\alpha}\pi_{0,0} + 1\right\},$$

where $\pi_{0,0}$ is given by (4.4).

Proof Differentiating (4.1), we obtain

$$\widehat{\pi}'(z) = \frac{\lambda(\lambda+\alpha)\widehat{X}'(z)}{(\lambda+\alpha-\lambda\widehat{X}(z))^2}\pi_{0,0} + \frac{\lambda}{\mu}q(z)\widehat{\pi}(z).$$

Taking the limit as $z \uparrow 1$ in the above equation yields

$$\begin{aligned}\widehat{\pi}'(1) &= \lim_{z\uparrow 1}\widehat{\pi}'(z)\\ &= \frac{\lambda(\lambda+\alpha)\mathsf{E}[X]}{\alpha^2}\pi_{0,0} + \frac{\lambda}{\mu}\lim_{z\uparrow 1}\frac{1-\widehat{X}(z)}{1-z}\\ &= \frac{\lambda(\lambda+\alpha)\mathsf{E}[X]}{\alpha^2}\pi_{0,0} + \frac{\lambda}{\mu}\mathsf{E}[X],\end{aligned}$$

where the third equality is due to L'Hospital's rule. From the relation $\mathsf{E}[N] = \widehat{\pi}'(1)$, we obtain Corollary 1. □

5 Sojourn time distribution

In this section, we derive the LST of the sojourn time distribution. Note that the LST of a distribution function $F(t)$ is defined as $F^*(s) := \int_{t\geq 0} \mathrm{e}^{-st}F(\mathrm{d}t)$. We assume that $\mathsf{E}[X] < \infty$ in this section for the existence of the equilibrium distribution.

In the M^X/M/1/SET-VARI queue, the server changes the speed upon arrivals and departures of jobs. Therefore, the sojourn time distribution of a tagged job is affected by the batches that arrive after it. This makes the derivation of the sojourn time distribution complex and challenging. We first derive the conditional LST for the sojourn time distribution. Then, combining with the queue length distribution, we obtain the unconditional LST for the sojourn time distribution.

First, we consider the case where the server is processing a job when the tagged job arrives. Let $S_1(n, m)$ denote the residual sojourn time of the tagged job, given that it is in the mth position and the system state is $(1, n)$. Conditioning on the first-step transitions, we have, for $m \in \mathbb{N}$ and $n \geq m$,

$$S_1(n,m) = \frac{Y}{\lambda+n\mu} + \begin{cases} S_1(n-1,m-1), & \text{w.p. } \dfrac{n\mu}{\lambda+n\mu},\\ S_1(n+k,m), & \text{w.p. } \dfrac{\lambda x_k}{\lambda+n\mu}, \quad k\in\mathbb{N}, \end{cases} \tag{5.1}$$

where Y denotes the exponential random variable with mean 1, and $S_1(n, 0) = 0$ for $n \in \mathbb{N}$. Furthermore, let $\psi_1(n, m, s)$ denote the LST of $S_1(n, m)$. Taking the LST of both sides of (5.1), we obtain, for $m \in \mathbb{N}$ and $n \geq m$,

$$\psi_1(n,m,s) = \frac{n\mu}{s+\lambda+n\mu}\psi_1(n-1,m-1,s) + \frac{\lambda}{s+\lambda+n\mu}\sum_{k=1}^{\infty} x_k \psi_1(n+k,m,s), \tag{5.2}$$

where $\psi_1(n,0,s) = 1$. We use the convention that $\psi_1(n,m,s) = 0$ for $n < m$. Furthermore, we define the infinite column vector $\boldsymbol{\psi}_1(m,s)$ as

$$\boldsymbol{\psi}_1(m,s) = (\psi_1(0,m,s), \ldots, \psi_1(m-1,m,s), \psi_1(m,m,s), \psi_1(m+1,m,s), \ldots)^\top,$$

where $\boldsymbol{a}^\top$ denotes the transposed vector of $\boldsymbol{a}$. Note that the nth element of $\boldsymbol{\psi}_1(m,s)$ is $\psi_1(n-1,m,s)$ for $n \geq m+1$.

We define infinite matrices $\boldsymbol{\Lambda}^{(1)}$ and $\boldsymbol{M}$ as

$$\boldsymbol{\Lambda}^{(1)} = \left(\Lambda_{ij}^{(1)}\right)_{(i,j)\in\mathbb{N}\times\mathbb{N}}, \tag{5.3}$$

$$\Lambda_{ij}^{(1)} = \begin{cases} \dfrac{\lambda x_{j-i}}{s+\lambda+(i-1)\mu}, & 1 < i < j, \\ 0, & \text{otherwise}, \end{cases}$$

$$\boldsymbol{M} = \left(M_{ij}\right)_{(i,j)\in\mathbb{N}\times\mathbb{N}}, \tag{5.4}$$

$$M_{ij} = \begin{cases} \dfrac{(i-1)\mu}{s+\lambda+(i-1)\mu}, & 1 < i = j+1, \\ 0, & \text{otherwise}. \end{cases}$$

Rearranging (5.2) by using these matrices, we obtain

$$\boldsymbol{\psi}_1(m,s) = \boldsymbol{M}\boldsymbol{\psi}_1(m-1,s) + \boldsymbol{\Lambda}^{(1)}\boldsymbol{\psi}_1(m,s), \qquad m \in \mathbb{N}. \tag{5.5}$$

From $\boldsymbol{\psi}_1(n,0,s) = 1$ for any $n \in \mathbb{N}$, we have

$$\boldsymbol{\psi}_1(0,s) = \mathbf{1},$$

where $\mathbf{1}$ is the infinite column vector whose elements are all equal to 1. Let $\| z \|_2$ denote the Euclidean norm of the vector z. We prove that the operator norm of the infinite matrix $\boldsymbol{\Lambda}^{(1)}$, $\| \boldsymbol{\Lambda}^{(1)} \| = \sup_{\|z\|_2=1} \| \boldsymbol{\Lambda}^{(1)} z \|_2$, is strictly smaller than 1. Indeed, for all $z = (z_1, z_2, \ldots)^\top$ such that $\| z \|_2 = 1$, we have

$$\begin{aligned} \| \boldsymbol{\Lambda}^{(1)} z \|_2 &= \left(\sum_{i>1} \left(\sum_{j>i} \frac{\lambda x_{j-i}}{s+\lambda+(i-1)\mu} z_j \right)^2 \right)^{1/2} \\ &= \left(\sum_{i>1} \frac{\lambda}{s+\lambda+(i-1)\mu} \left(\sum_{j>i} x_{j-i} z_j \right)^2 \right)^{1/2} \\ &< \frac{\lambda}{\lambda+s} \left(\sum_{i>1} \left(\sum_{j\geq 1} x_j z_{j+i} \right)^2 \right)^{1/2} \end{aligned}$$

$$\leq \frac{\lambda}{\lambda+s}\Big(\sum_{i>1}\sum_{j\geq 1} x_j\Big(z_{j+i}\Big)^2\Big)^{1/2}$$
$$\leq \frac{\lambda}{\lambda+s}\Big(\sum_{j\geq 1} x_j||\boldsymbol{z}||_2^2\Big)^{1/2}$$
$$= \frac{\lambda}{\lambda+s},$$

where the second inequality holds because of Jensen's inequality. Thus,

$$\| \boldsymbol{\Lambda}^{(1)} \| \leq \frac{\lambda}{\lambda+s} < 1.$$

Because $\| \boldsymbol{\Lambda}^{(1)} \|$ is strictly smaller than 1, $(\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})$ has an inverse mapping, where $\boldsymbol{I}$ is the infinite identity matrix [6, Section 29, Theorem 8]. Therefore, from (5.5), we obtain the following recurrence equations, for $m \in \mathbb{N}$:

$$\boldsymbol{\psi}_1(m,s) = (\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})^{-1}\boldsymbol{M}\boldsymbol{\psi}_1(m-1,s).$$

Solving this equation, we obtain

$$\boldsymbol{\psi}_1(m,s) = \{(\boldsymbol{I} - \boldsymbol{\Lambda}^{(1)})^{-1}\boldsymbol{M}\}^m\mathbf{1}. \tag{5.6}$$

Next, we consider the case where the server is not processing a job when the tagged job arrives. Let $S_0(n,m)$ denote the residual sojourn time of the tagged job, given that the tagged job is in the mth position and the system state is $(0,n)$. Let $\psi_0(n,m,s)$ denote the LST of $S_0(n,m)$ for $n \in \mathbb{N}$ and $m \leq n$. In addition, we define the infinite column vector $\boldsymbol{\psi}_0(m,s)$ as

$$\boldsymbol{\psi}_0(m,s) = (\psi_0(0,m,s),\ldots,\psi_0(m-1,m,s),\psi_0(m,m,s),\psi_0(m+1,m,s),\ldots)^\top,$$

where $\psi_0(n,0,s) = 1, n \in \mathbb{Z}_+$, and $\psi_0(n,m,s) = 0, n \in \mathbb{Z}_+$ and $m > n$. We use the convention that $\psi_0(n,m,s) = 0$ for $n < m$. As with the analysis for $\boldsymbol{\psi}_1(m,s)$, i.e., (5.1)–(5.6), we obtain

$$\boldsymbol{\psi}_0(m,s) = (\boldsymbol{I} - \boldsymbol{\Lambda}^{(0)})^{-1}\boldsymbol{A}\boldsymbol{\psi}_1(m,s), \tag{5.7}$$

where the infinite matrices $\boldsymbol{\Lambda}^{(0)}$ and $\boldsymbol{A}$ are defined as follows:

$$\boldsymbol{\Lambda}^{(0)} = \big(\Lambda_{ij}^{(0)}\big)_{(i,j)\in\mathbb{N}\times\mathbb{N}}, \qquad \Lambda_{ij}^{(0)} = \begin{cases} \dfrac{\lambda x_{j-i}}{s+\lambda+\alpha}, & 1 < i < j, \\ 0, & \text{otherwise}, \end{cases} \tag{5.8}$$

$$\boldsymbol{A} = \big(A_{ij}\big)_{(i,j)\in\mathbb{N}\times\mathbb{N}}, \qquad A_{ij} = \begin{cases} \dfrac{\alpha}{s+\lambda+\alpha}, & 1 < i = j, \\ 0, & \text{otherwise}. \end{cases} \tag{5.9}$$

Note that $(\boldsymbol{I} - \boldsymbol{\Lambda}^{(0)})$ has an inverse mapping, which can be proved similarly to the analysis for $\boldsymbol{\Lambda}^{(1)}$.

Next, we derive the unconditional LST of the sojourn time distribution. To this end, we define $\tau(i, n, m)$ as the probability that the tagged job is located in the mth position and the state of the system becomes (i, n) immediately after its arrival. Let $(\mathscr{I}, \mathscr{L}_p)$ denote the state of the system just before the tagged job arrives at the system. Let $\mathscr{P}$ denote the position at which the tagged job is located immediately after it enters the system. Let $\widetilde{X}$ denote the number of jobs in the batch to which the tagged job belongs. Under the assumption that $\mathsf{E}[X] < \infty$, the distribution of $\widetilde{X}$ is the equilibrium distribution of X, given in [4], for $k \in \mathbb{N}$,

$$\mathsf{P}[\widetilde{X} = k] = \frac{k x_k}{\mathsf{E}[X]}.$$

In addition, from PASTA [18], we have, for $(i, n) \in \mathscr{S}$,

$$\mathsf{P}[\mathscr{I} = i,\ \mathscr{L}_p = n] = \pi_{i,n}.$$

Let $\mathscr{X}(n, m) := \{k\,;\ x_k > 0,\ n \le k \le m\}$. We obtain, for $n \in \mathbb{N}$ and $n \ge m$,

$$\begin{aligned}
&\tau(1, n, m)\\
&= \sum_{k \in \mathscr{X}(n-m+1, n-1)} \mathsf{P}[\mathscr{I} = 1,\ \mathscr{L}_p = n - k,\ \widetilde{X} = k,\ \mathscr{P} = m]\\
&= \sum_{k \in \mathscr{X}(n-m+1, n-1)} \mathsf{P}[\mathscr{P} = m | \mathscr{I} = 1,\ \mathscr{L}_p = n - k,\ \widetilde{X} = k]\\
&\quad \times \mathsf{P}[\mathscr{I} = 1,\ \mathscr{L}_p = n - k,\ \widetilde{X} = k]\\
&= \sum_{k \in \mathscr{X}(n-m+1, n-1)} \frac{1}{k} \mathsf{P}[\tilde{X} = k | \mathscr{I} = 1,\ \mathscr{L}_p = n - k] \mathsf{P}[\mathscr{I} = 1,\ \mathscr{L}_p = n - k]\\
&= \sum_{k \in \mathscr{X}(n-m+1, n-1)} \frac{1}{k} \frac{k x_k}{\mathsf{E}[X]} \pi_{1, n-k}\\
&= \sum_{k=n-m+1}^{n-1} \pi_{1, n-k} \frac{x_k}{\mathsf{E}[X]}.
\end{aligned}$$

Similarly to the above, we obtain, for $n \in \mathbb{Z}_+$ and $n \ge m$,

$$\tau(0, n, m) = \sum_{k=n-m+1}^{n} \pi_{0, n-k} \frac{x_k}{\mathsf{E}[X]}.$$

Since we use the convention that $\psi_i(n, m, s) = 0$ for $n < m$, the LST of the sojourn time distribution, denoted by $\psi(s)$, can be expressed as follows by using $\psi_0(n, m, s)$ and $\psi_1(n, m, s)$:

$$\psi(s) = \sum_{n=1}^{\infty}\sum_{m=1}^{n}\tau(0,n,m)\psi_0(n,m,s) + \sum_{n=2}^{\infty}\sum_{m=2}^{n}\tau(1,n,m)\psi_1(n,m,s)$$
$$= \sum_{m=1}^{\infty}\sum_{n=m}^{\infty}\sum_{k=n-m+1}^{n}\pi_{0,n-k}\frac{x_k}{\mathsf{E}[X]}\psi_0(n,m,s)$$
$$+ \sum_{m=2}^{\infty}\sum_{n=m}^{\infty}\sum_{k=n-m+1}^{n-1}\pi_{1,n-k}\frac{x_k}{\mathsf{E}[X]}\psi_1(n,m,s). \tag{5.10}$$

It is obvious that the infinite series included in $\psi(s)$ converges. The reason is that $\sum_{n=1}^{\infty}\sum_{m=1}^{n}\tau(0,n,m) + \sum_{n=2}^{\infty}\sum_{m=2}^{n}\tau(1,n,m) = 1$ and $0 \le \psi_i(n,m,s) \le 1$ for $i = 1, 2$, $n \in \mathbb{N}$ and $1 \le m \le n$.

For a compact expression of (5.10), we define the infinite matrices $\boldsymbol{I}_m$, for $m \in \mathbb{N}$, and $\boldsymbol{B}$ as

$$\boldsymbol{I}_m = \left(I_{ij}\right)_{(i,j)\in\mathbb{N}\times\mathbb{N}}, \qquad I_{ij} = \begin{cases} 1, & 1 \le i = j \le m, \\ 0, & \text{otherwise}, \end{cases} \tag{5.11}$$

$$\boldsymbol{B} = \left(B_{ij}\right)_{(i,j)\in\mathbb{N}\times\mathbb{N}}, \qquad B_{ij} = \begin{cases} x_{j-i}, & 1 \le i < j, \\ 0, & \text{otherwise}. \end{cases} \tag{5.12}$$

In addition, we define the infinite row vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ as

$$\boldsymbol{\pi}_0 = (\pi_{0,0}, \pi_{0,1}, \pi_{1,2}, \ldots), \quad \boldsymbol{\pi}_1 = (0, \pi_{1,1}, \pi_{1,2}, \ldots).$$

Rearranging (5.10) by using these matrices and vectors, we obtain

$$\psi(s) = \sum_{m=1}^{\infty}\frac{\boldsymbol{\pi}_0}{\mathsf{E}[X]}\boldsymbol{I}_m\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{I}_m)\boldsymbol{\psi}_0(m,s)$$
$$+ \sum_{m=1}^{\infty}\frac{\boldsymbol{\pi}_1}{\mathsf{E}[X]}\boldsymbol{I}_m\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{I}_m)\boldsymbol{\psi}_1(m,s). \tag{5.13}$$

From (5.6), (5.7) and (5.13), we obtain the LST of the sojourn time distribution as follows:

Theorem 3 *The LST of the sojourn time distribution, $\psi(s)$, is given as follows:*

$$\psi(s) = \sum_{m=1}^{\infty}\frac{1}{\mathsf{E}[X]}\left[\boldsymbol{\pi}_0\boldsymbol{I}_m\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{I}_m)(\boldsymbol{I}-\boldsymbol{\Lambda}^{(0)})^{-1}\boldsymbol{A}\right]\left[(\boldsymbol{I}-\boldsymbol{\Lambda}^{(1)})^{-1}\boldsymbol{M}\right]^m\mathbf{1}$$
$$+ \sum_{m=1}^{\infty}\frac{1}{\mathsf{E}[X]}\left[\boldsymbol{\pi}_1\boldsymbol{I}_m\boldsymbol{B}(\boldsymbol{I}-\boldsymbol{I}_m)\right]\left[(\boldsymbol{I}-\boldsymbol{\Lambda}^{(1)})^{-1}\boldsymbol{M}\right]^m\mathbf{1},$$

where $\boldsymbol{\Lambda}^{(1)}$, $\boldsymbol{M}$, $\boldsymbol{\Lambda}^{(0)}$, $\boldsymbol{A}$, $\boldsymbol{I}_m$ and $\boldsymbol{B}$ are given by (5.3), (5.4), (5.8), (5.9), (5.11 *and* (5.12, *respectively.*

Remark 3 The LST of the sojourn time distribution given in Theorem 3 is in series form involving infinite-dimensional matrices. Therefore, an approximation is necessary for numerical calculation. In Sect. 6.2, for numerical experiments, we present a method to approximate $\psi(s)$. However, we have not yet been able to find a bound for the error. It is important future work to find an approximation method with guaranteed accuracy.

6 Numerical results

In this section, we present some numerical results for the M^X/M/1/SET-VARI queue. We consider three types of distribution for X: the binomial distribution with parameters $n \in \mathbb{N}$ and $0 < p < 1$, denoted by Binom(n, p), the discrete uniform distribution with parameters $a, b \in \mathbb{N}$ $(a \leq b)$, denoted by Unif$\{a, b\}$, and the geometric distribution with parameter $0 < p < 1$, denoted by Geo(p).

6.1 Energy consumption and response time trade-off

We consider the trade-off between the average energy consumption and the average sojourn time. In order to compare the variable speed CPU with the fixed speed CPU, we also consider the M^X/M/1/SET-FIX queue where the service speed is fixed, while other settings are kept the same as the M^X/M/1/SET-VARI queue. We use the PGF of the number of jobs in the system derived for the M^X/M/1/SET-FIX queue in [2]. The assumptions regarding energy consumption per unit time for each state are given in Table 1.

Note that the constants K_{service} and K_{set} from Table 1 depend on the particular system. Let $\mathsf{E}[W_{\text{v}}]$ and $\mathsf{E}[W_{\text{f}}]$ denote the average sojourn time of the variable speed queue and that of the fixed speed queue, respectively. In addition, let $\mathsf{E}[P_{\text{v}}]$ and $\mathsf{E}[P_{\text{f}}]$ denote the average energy consumption of the variable speed queue and the fixed speed queue, respectively. $\mathsf{E}[P_{\text{v}}]$ and $\mathsf{E}[P_{\text{f}}]$ can be expressed as follows:

$$\mathsf{E}[P_{\text{v}}] = \mu^2\{\widehat{\pi}_0(1) - \pi_{0,0} + \widehat{\pi}_1''(1) + \widehat{\pi}_1'(1)\},$$
$$\mathsf{E}[P_{\text{f}}] = \frac{\mu^2}{(\lambda+\alpha)(\lambda+\alpha\lambda\mathsf{E}[X]/\mu)}.$$

We will explore the relationship between the average sojourn time and the average power consumption. In addition, we will compare $\mathsf{E}[W_{\text{v}}]$ with $\mathsf{E}[W_{\text{f}}]$ under the condition that $\mathsf{E}[P_{\text{v}}] = \mathsf{E}[P_{\text{f}}]$. In what follows, we assume that $\alpha = 0.1$ and $\lambda\mathsf{E}[X] = 1$.

Table 1 Energy consumption per unit time at each state [7,9]

State		Variable	Fixed
Service	$(1, j)$ $j \geq 1$	$K_{\text{service}} \times (j\mu)^2$	$K_{\text{service}} \times \mu^2$
Setup	$(0, j)$ $j \geq 1$	$K_{\text{set}} \times \mu^2$	
Idle	$(0, 0)$	0	

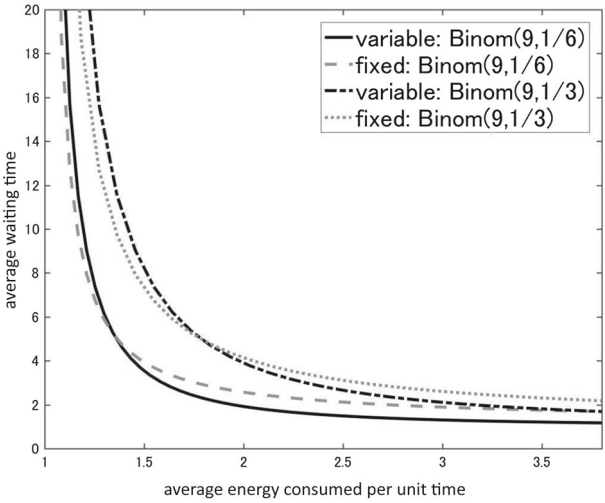

Fig. 2 $\lambda \mathsf{E}[X] = 1.0,\ \alpha = 1.0$. X follows the binomial distribution

Note that $\lambda \mathsf{E}[X]$ is the mean number of jobs arriving per unit time. In the numerical experiments, the value of λ and the distribution of X change while keeping $\lambda \mathsf{E}[X] = 1$.

The procedure of numerical experiments is as follows. First, fixing the value of $\mu \in (0, 3]$, we compute the average energy consumption per unit time of the M^X/M/1/SET-VARI queue, denoted by A_{v}, and the average sojourn time. Let $\mu_{\mathrm{f}}(A)$ denote the unique service rate which realizes the average energy consumption A in the fixed queue. Next, we compute $\mu_{\mathrm{f}}(A_{\mathrm{v}})$ by

$$\mu_{\mathrm{f}}(A_{\mathrm{v}}) = \big(-\alpha\lambda \mathsf{E}[X] + \sqrt{(\alpha\lambda \mathsf{E}[X])^2 + 4\lambda A_{\mathrm{v}}(\lambda + \alpha)}\,\big)/(2\lambda).$$

As a result, the energy consumption for both models is kept the same. Finally, we compute the average sojourn time of the M^X/M/1/SET-FIX queue under the parameter $\mu_{\mathrm{f}}(A_{\mathrm{v}})$. In this numerical experiment, we compute the average sojourn time from the average number of jobs in the system using Little's formula [8].

In Fig. 2, we present the results when X is Binom$(9, 1/6)$ (and $\lambda = 0.4$) and when X is Binom$(9, 1/3)$ (and $\lambda = 0.25$). Jobs in the case of Binom$(9, 1/3)$ are more likely to arrive in larger batches than those in the case of Binom$(9, 1/6)$. In Fig. 3, we present the results when X is Unif$\{1, 4\}$ (and $\lambda = 0.4$) and when X is Unif$\{1, 7\}$ (and $\lambda = 0.25$). Similarly, jobs in the case of Unif$\{1, 7\}$ are likely to arrive in larger batches than those in the case of Unif$\{1, 4\}$.

In all the cases in Figs. 2 and 3, we observe the trade-off between the average sojourn time and the average power consumption, i.e., the average sojourn time is smaller when the average power consumption is larger. In addition, keeping the average number of arrivals per unit time ($\lambda \mathsf{E}[X]$) and the average power consumption the same, arriving in larger batches results in a smaller average sojourn time. This result seems intuitively true and might be proven.

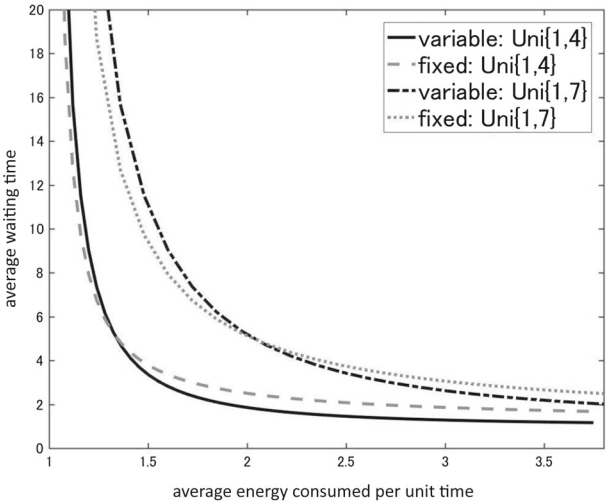

Fig. 3 $\lambda \mathsf{E}[X] = 1.0,\ \alpha = 1.0$. X follows the discrete uniform distribution

Let's compare the variable speed queue with the fixed queue. Our numerical results show that for high-performance applications, in which delays must be kept small, having variable speed can result in both shorter delays and lower energy than having fixed speed, while the opposite is true for applications where energy usage is more important than delay performance.

These observations could be explained as follows: Energy consumption in the variable service queue monotonically increases with the number of jobs in the system while that in fixed service queue is constant whenever there are jobs in the system.

6.2 The sojourn time distribution

In Figs. 4 and 5, we present the probability densities of the sojourn time distributions computed by numerically inverting the Laplace–Stieltjes transform. In what follows, we assume that $\alpha = 0.1$ and $\lambda \mathsf{E}[X] = 1$. The LST of the sojourn time distribution is given in Theorem 3, but it is in series form involving infinite-dimensional matrices. Therefore, as mentioned in Remark 3, approximation is necessary for numerical calculation.

We present a procedure to compute the LST of the sojourn time distribution, $\psi(s)$. First, we truncate the infinite vectors $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_1$ to the vectors of their first $(N^* + 1)$ elements, where the constant N^* is determined by

$$N^* = \inf\left\{n \in \mathbb{N};\ 1 - \sum\nolimits_{j=0}^{n} \pi_{0,j} - \sum\nolimits_{j=1}^{n} \pi_{1,j} < 10^{-4}\right\}.$$

This is equivalent to disregarding the states with more than N^* jobs in the system whose probability is 10^{-4}. We compute $\pi_{0,0}$ by (4.4) and $\pi_{i,j}$, $i = 0, 1$, $j = 1, \ldots, N^*$, by (3.3), (3.4) and (3.5). In addition, we truncate the infinite matrices appearing in $\psi(s)$ to their $(N^* + 1) \times (N^* + 1)$ north-west corner matrices. We compute each element

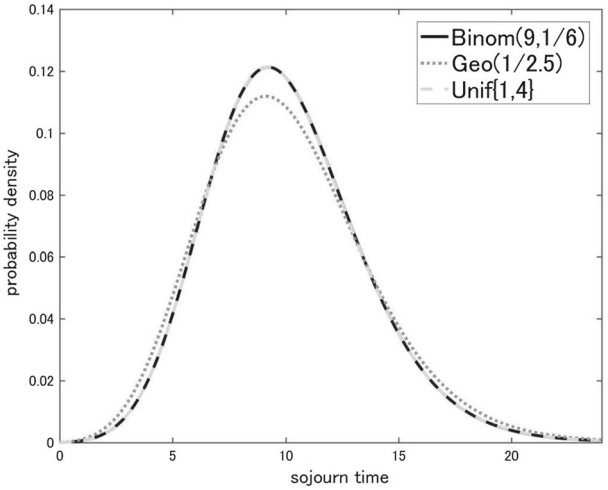

Fig. 4 Comparison of the cases in which X follows different distributions. $\lambda = 0.4$, $\alpha = 0.1$, $\mu = 0.1$, $\mathsf{E}[X] = 2.5$

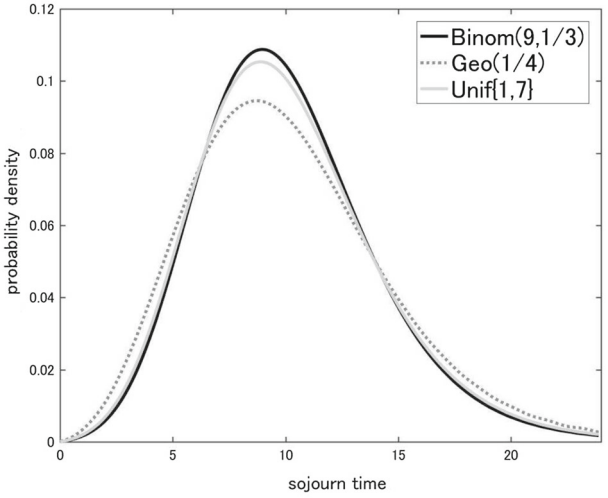

Fig. 5 Comparison of the cases in which X follows different distributions. $\lambda = 0.25$, $\alpha = 0.1$, $\mu = 0.1$, $\mathsf{E}[X] = 4.0$

of the infinite matrices by (5.3), (5.4), (5.8), (5.9), (5.11) and (5.12). Let $\psi^*(s)$ denote the function computed by Theorem 3 using the truncated matrices and vectors. In our numerical experiments, we use the value of $\psi^*(s)$ as an approximation to the LST of the sojourn time distribution. It is important future work to estimate the error of this approximation. Our extensive numerical experiments show that the approximation is fairly accurate in the sense that the final results do not change much as N^* is increased.

Next, we present the procedure to compute the value of the sojourn time distribution for $t \in [0, T]$ by numerically inverting the Laplace–Stieltjes transform [5] for fixed $T > 0$. The function $f^{(K)}(t)$ and the constant $K^*(t)$ are defined as follows:

$$f^{(K)}(t) = \frac{h}{\pi}\exp\left(\frac{6}{T}t\right)\left\{\frac{\psi(6/T)}{2} + \sum_{k=1}^{K}\mathrm{Re}\left[\psi(6/T + ikh)\exp(ikht)\right]\right\},$$
$$K^*(t) = \inf\left\{K \in \mathbb{N};\ \left|f^{(K)}(t) - f^{(K-1)}(t)\right| < 10^{-4}\right\},$$

where $i = \sqrt{-1}$ and $\mathrm{Re}(a + ib) = a$. In our numerical experiments, we use the value of $f^{(K^*(t))}(t)$ as the sojourn time distribution. The constant h is the step size; we set $h = 1/100$.

In Figs. 4 and 5, we investigate the impact of the batch size distribution on the sojourn time distribution. Figure 4 presents the sojourn time distribution for $\lambda = 0.4$, $\mathsf{E}[X] = 2.5$ and $\mu = 0.1$, while Fig. 5 shows that for $\lambda = 0.25$, $\mathsf{E}[X] = 4$ and $\mu = 0.1$. Note that in both figures $\lambda\mathsf{E}[X] = 1$. We observe that the curves of Binom(9, 1/6) and Uni(1, 4) almost coincide. The values of second, third and fourth moments are 7.5, 25.8 and 99.2 for Binom(9, 1/6), and 7.5, 25.0 and 113.5 for Uni(1, 4). On the other hand, the values of second, third and fourth moments are 10.0, 58.8 and 480.0 for Geo(1/2.5). This suggests that high-order moments (roughly fourth or higher) have less influence in the sojourn time distribution.

Compared with Fig. 4, the curves for the binomial distribution, uniform distribution and geometric distribution are different in Fig. 5. The second moments are 18.0 for Binom(9, 1/3), 20.0 for Uni(1, 4) and 38.0 for Geo(1/2.5). This suggests that the second moment of the batch size has a significant impact on the sojourn time distribution.

7 Conclusion

In this paper, we have studied the M^X/M/1/SET-VARI queue. We have derived the PGF of the number of jobs in the system in an integral form. Furthermore, we have derived the LST of the sojourn time distribution, which is obtained in series form involving infinite-dimensional matrices. Through numerical experiments, we have been able to observe some insights into the sojourn time distribution and the average energy consumption. One remark is that the stationary queue length distribution and the sojourn time distribution of the finite buffer version can be obtained using almost the same procedure as for the infinite buffer model, so we have omitted that analysis here. Furthermore, the finite buffer is easier in the sense that it is always stable and the sojourn time distribution does not involve infinite matrices. As future work, we plan to consider the model where the service rate is an arbitrary function of the number of jobs in the system. Models with general setup time and service time distributions may also be investigated somewhere else.

Acknowledgements We would like to thank the guest editors and two anonymous referees for their constructive comments which significantly improved the presentation of the paper. The research of TP was partially supported by JSPS KAKENHI Grant Number 26730011.

References

1. Adan, I., D'Auria, B.: Sojourn time in a single-server queue with threshold service rate control. SIAM J. Appl. Math. **76**(1), 197–216 (2016)
2. Baba, Y.: The M^X/M/1 queue with multiple working vacation. Am. J. Oper. Res. **2**(2), 217–224 (2012)

3. Cong, T.D.: On the $M^X/G/\infty$ queue with heterogeneous customers in a batch. J. Appl. Probab. **31**(1), 280–286 (1994)
4. Downton, F.: Waiting time in bulk service queues. J. R. Stat. Soc. **17**(2), 256–261 (1955)
5. Durbin, F.: Numerical inversion of Laplace transforms: an efficient improvement to Dubner and Abate's method. Comput. J. **17**(4), 371–376 (1974)
6. Fomin, S.V., et al.: Elements of the Theory of Functions and Functional Analysis, vol. 1. Courier Corporation, North Chelmsford (1999)
7. Gandhi, A., Harchol-Balter, M., Adan, I.: Server farms with setup costs. Perform. Eval. **67**(11), 1123–1138 (2010)
8. Keilson, J., Servi, L.D.: A distributional form of Little's law. Oper. Res. Lett. **7**(5), 223–227 (1988)
9. Lu, X., Aalto, S., Lassila, P.: Performance-energy trade-off in data centers: Impact of switching delay. In: Proceedings of 22nd IEEE ITC Specialist Seminar on Energy Efficient and Green Networking (SSEEGN), pp. 50–55 (2013)
10. Maccio, V.J., Down, D.G.: On optimal policies for energy-aware servers. Perform. Eval. **90**, 36–52 (2015)
11. Mittal, S.: A survey of techniques for improving energy efficiency in embedded computing systems. Int. J. Comput. Aided Eng. Technol. **6**, 440–459 (2014)
12. Norris, J.R.: Markov Chains. Cambridge University Press, Cambridge (1998)
13. Rohde, U.L.: Digital PLL Frequency Synthesizers: Theory and Design. Prentice-Hall, Englewood Cliffs, NJ (1983)
14. Shanbhag, D.N.: On infinite server queues with batch arrivals. J. Appl. Probab. **3**(1), 274–279 (1966)
15. Sueur, E.L., Heiser, G.: Dynamic voltage and frequency scaling: the laws of diminishing returns. In: Proceedings of the 2010 International Conference on Power Aware Computing and Systems, pp. 1–8 (2010)
16. Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis. Cambridge University Press, Cambridge (1996)
17. Wierman, A., Andrew, L., Tang, A.: Power-aware speed scaling in processor sharing systems: optimality and robustness. Perform. Eval. **69**, 601–622 (2012)
18. Wolft, R.W.: Poisson arrivals see time average. Oper. Res. **30**, 223–231 (1982)