EURANDOM PREPRINT SERIES

2020-012

September 11, 2020

An M/PH/1 queue with workload-dependent processing speed and vacations

Y. Sukama, O. Boxma, T. Phung-Duc
ISSN 1389-2355

Edited by Y.S. on September 11, 2020

An M/PH/1 queue with workload-dependent processing speed and vacations

Yutaka Sakuma[1*], Onno Boxma[2] and Tuan Phung-Duc[3]

[1*] Department of Computer Science, National Defense Academy of Japan, Yokosuka, Japan, Corresponding author, sakuma@nda.ac.jp

[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands; o.j.boxma@tue.nl

[3] Department of Policy and Planning Sciences, Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan; tuan@sk.tsukuba.ac.jp

Abstract

Motivated by the trade-off issue between delay performance and energy consumption in modern computer and communication systems, we consider a single-server queue with Phase-type service requirements and with the following two special features. Firstly, the service speed is a, piecewise constant, function of the workload. Secondly, the server switches off when the system becomes empty, only to be activated again when the workload reaches a certain threshold. For this system, we obtain the steady-state workload distribution and its moments of any order. We use this result to choose the activation threshold such that a certain cost function, involving processing costs, activation costs and mean workload, is minimized. In the case of exponential service requirements, we also derive the Laplace-Stieltjes transform of the length of the active period of the server.

Keywords: Workload-dependent service; Phase-type demand; Matrix exponential solution; Autoscaling

Mathematics Subject Classification: 60K25; 60K05; 90B22.

1 Introduction

We consider an $M/PH/1$ queue with two special features. Firstly, the service speed is a piecewise constant function of the workload. Secondly, the server switches off when the system becomes empty, and it is re-activated when the workload reaches a certain threshold. In the remainder of this section we discuss the motivation for our study (Subsection 1.1), mention related literature (Subsection 1.2) and outline the rest of the paper (Subsection 1.3).

1.1 Motivation

Cloud services have become ubiquitous in our modern information society. Cloud services are even more and more important in the current era where we are facing the Covid-19 pandemic for which remote work is strongly recommended to decrease the spread of the virus. Nowadays, most Internet users are familiar with some cloud-based storage services such as Dropbox, Google drive etc. and with video-conferencing services such as Zoom, Microsoft Teams etc. These systems are supported by large-scale data centers where thousands of servers are available, consuming a huge amount of energy. Thus, it is crucial to have mechanisms that can balance energy consumption and performance.

While energy saving is indispensable to reduce the CO_2, most data centers are still designed for peak traffic. As a result, many servers are idle in off-peak periods – but they may still consume about 60% of their peak energy consumption [15, 21]. Autoscaling techniques are proposed to solve the trade-off between energy consumption and delay performance. In particular, an autoscaling algorithm adjusts the processing speed according to its workload. At the data center level, an autoscaling algorithm controls the number of active servers in the system in response to the workload [15, 21, 24, 27]. Furthermore, at the individual computer level, a CPU is also able to adjust the processing speed by either dynamic frequency scaling or dynamic voltage scaling techniques [35, 36, 37]. The processing speed is scaled up when the workload is high and scaled down under a low workload. As a result, less energy is consumed under off-peak periods while acceptable delay can be achieved in a heavy-load situation [13].

Queues with changeable service rate also fit the autoscaling mechanisms in 5G networks. In 5G networks, the key technology is network function virtualization (VNF) in which a physical resource can be virtualized to network functions. The operator can dynamically add or release these network functions to optimally construct cost-effective systems in response to their workload. As a result, these systems can be modeled using queueing models with changeable service speeds [26, 27].

Apart from the interest in power-saving computer and communication systems, queues with changeable service speed also naturally arise in service systems with human servers. In particular, in many real-world service systems such as call centers, staffs are scheduled to meet the demands of customers. Also, a human server may serve at high speed when the workload is high, and may spend more time on a job when the workload is low [11].

1.2 Related literature and our contribution

The topic of speed scaling in data centers and power-saving CPU and autoscaling of 5G networks provides motivation for our study. See [35] for an insightful discussion of speed scaling and [26, 27] for queueing analysis of an autoscaling algorithm in 5G networks. Recent papers which consider single server queues with speed scaling where the speed of the server is proportional to the number of jobs in the system are, e.g., [25, 36, 37]. Multiserver queues with ON-OFF control (turning idle servers off) have been extensively studied [15, 21, 24].

Our model is related to the literature on queues and dams with a level-dependent outflow rate. Influential early papers are [16, 17]; we refer to [4] for some more recent results and further references.

Our model is also related to the rich vacation literature: the server takes a vacation when the workload is zero, and returns to service when the workload reaches or exceeds a certain level. Such a D-policy has been extensively studied for the classical $M/G/1$ queue. See [14] for references and, in particular, for an optimality proof. For the case of switching costs and running costs, and with a holding cost per time unit which is a non-negative decreasing right-continuous function of the current workload, Feinberg and Kella [14] prove that D-policies are optimal for the average-cost-per-time-unit criterion: there is an optimal policy that either runs the server all the time or switches the server off when the system becomes empty and switches it on when the workload reaches or exceeds some threshold D.

Markov modulated queues with workload dependent service rates have been extensively studied in the literature [10, 12, 18, 22, 33, 38]. All these studies first reduce the models to fluid models for which matrix analytic methods, spectral methods, and Schur decomposition methods are applied to derive numerical solutions for the distribution of the workload processes (See da Silva Soares and Latouche [10], Mandjes et al. [22] and Kankaya and Akar [18] for the methodologies). In contrast to the above, motivated by power-saving in modern computer and communication systems, we consider a model with Poisson input, Phase-type distribution, and vacation. For this model, using the level-crossing method and renewal theory, we obtain a direct solution for the workload distribution and its moments of any order. Our solution is more direct in the sense that the workload distribution is expressed in terms of matrix exponentials whose components are explicitly written in terms of given parameters. These results are then used for an optimization problem balancing performance and energy-consumption tradeoff.

We finally would like to point out the relation to our recent paper [29]. There we propose and analyze an M/G/1-type queueing model that also features two power-saving mechanisms. The speed of the server is scaled according to the workload in the system. Moreover, the server is turned off when the system is empty and is activated again once the workload reaches a certain threshold. In the case of arbitrarily distributed service time and general service speed function, the stationary workload is expressed in terms of a series whose terms are recursively obtained by an integral formula. While the distribution of the workload can in principle be evaluated for this general case, the computation is highly complex. Simpler expressions are obtained for the case of exponential service requirement and the case where the service speed is a linear function of the workload.

In the present paper our aim is to derive a computable solution for a relatively general model with Phase-type service requirements and a piecewise constant service rate. Phase-type distributions lie dense in the class of distributions with positive support and thus can approximate any service requirement distribution with any accuracy. Furthermore, piecewise constant functions can also be used to approximate an arbitrary function. In addition, from a practical point of view, it is natural that the service rate is switched at discrete points [13].

1.3 Structure of the paper

The remainder of the paper is organized as follows. In Section 2 we derive the stationary workload distribution for the case of an M/PH/1 queue with vacations and with only two processing speeds. We first focus on this case, before tackling the case of an arbitrary number of different processing speeds in Section 3, because the analysis is quite technical; in this way, we improve the readability of the paper. We also obtain a computable form for the moments of any order for the stationary workload. In Section 4, we demonstrate the analysis of the active period for an M/M/1 with two processing speeds. An optimization problem is formulated in Section 5 and finally numerical examples are presented in Section 6.

2 M/PH/1 queue with two processing speeds and vacations

In this section we consider the special case of a FCFS queue with a single server who, when active, works at one of two possible speeds: when the workload is below a threshold d_1 it works at speed r_1, and above it at speed r_2. Furthermore, when the system becomes empty, the server is switched off, only to be activated again when the workload exceeds some threshold level L. We assume that $L = d_1$ to keep the model for the moment as simple as possible while retaining its essential elements. In Section 3, we shall consider the more general case of piecewise constant service speed with K_0 different speed values, and in which L not necessarily coincides with one of the thresholds at which also the service speed in an active period changes. However, the analysis of that case is quite involved, and consideration of the simpler case of the present section will help the reader get acquainted with our approach.

The remainder of the section is organized as follows. Subsection 2.1 contains a more detailed model description, as well as a lemma about the computation of the convolution of matrix exponentials that will play a key role in the remainder of the paper. In Subsection 2.2, we derive the stationary workload density when the server is inactive. In Subsections 2.3 and 2.4, we successively determine the stationary workload density when the server is active while the workload is above, respectively below, d_1. The mean active period features in many of the formulas; we compute it explicitly in Subsection 2.5.

2.1 Model description

We consider a single-server FCFS queue, where the server has a single waiting line with infinite capacity. Customers arrive according to a Poisson process with rate $\lambda > 0$. The service requirements of the customers are independent and identically distributed (i.i.d.), generically indicated by B, with the following phase-type distribution (see, e.g. [20])

$$B(x) = 1 - \boldsymbol{\tau} \exp(\boldsymbol{T}x)\mathbf{1}, \quad x \geq 0, \tag{2.1}$$

where $\boldsymbol{\tau}$ is a $(1 \times N)$ probability row vector, and $\boldsymbol{T}$ is an $(N \times N)$ defective transition rate matrix, where N is a positive integer, and $\mathbf{1}$ is a column vector with ones whose size will be

determined in the context. The tail of $B(\cdot)$ is denoted by $\overline{B}(x) := 1 - B(x)$.

Let Z_t, where $t \in \mathbb{R}_+ := [0, \infty)$, denote the unfinished workload (workload, for short) in the system at time t. According to the value or history of the workload, the server is assumed to alternate between "inactive" and "active" states, which are referred to as modes 0 and 1, respectively. Let $S_t \in \{0, 1\}$ denote the mode of the server at time t, which is defined as follows.

(i) When the workload Z_t hits 0 at a time t, the server enters mode 0, i.e., $S_t = 0$. After that, the server remains in mode 0 until the workload exceeds threshold $L = d_1 > 0$, i.e., $S_u = 0$ as long as $0 \le Z_u \le d_1$ for $u \ge t$.

(ii) When the workload Z_t exceeds d_1 at a time t while the mode was 0 at time $t-$, the server changes its mode from 0 to 1, i.e., $S_t = 1$. The server remains in that mode until the workload hits 0 again, i.e., $S_u = 1$ as long as $0 < Z_u < \infty$ for $u \ge t$.

The processing speed of the server is assumed to depend on both the server's mode and workload in the following way. Let $s_i(x)$, where $i = 0, 1$ and $x \ge 0$, denote the processing speed of the server when $(Z_t, S_t) = (x, i)$; it is defined as follows:

$$s_0(x) = 0, \quad 0 \le x \le d_1, \tag{2.2}$$

$$s_1(x) = \begin{cases} r_1, & 0 < x \le d_1, \\ r_2, & d_1 < x < \infty, \end{cases} \tag{2.3}$$

where r_1 and r_2 are positive numbers. $\{(Z_t, S_t); t \ge 0\}$ is a Markov process. Figure 1 shows a sample path of this Markov process.

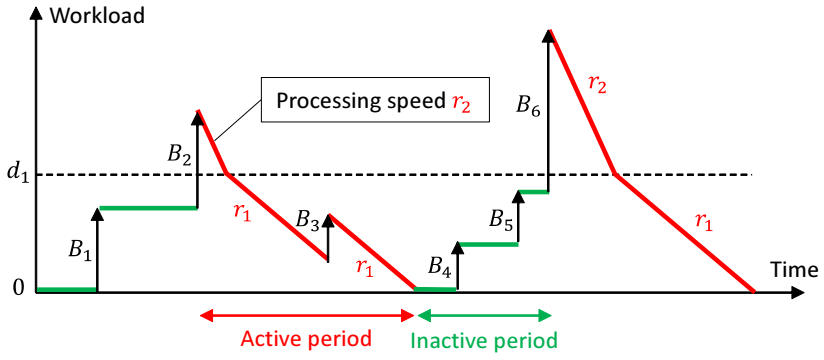

Figure 1: The workload process for the case of two processing speeds.

We assume that it is positive-recurrent, and denote its replica in steady state by (Z, S). The stability condition of this Markov process is given as follows.

$$\lambda \boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbf{1} < r_2, \tag{2.4}$$

where the lefthand side is the mean amount of work arriving in a unit of time, and the righthand side is the processing rate of the server when the workload is greater than d_1.

We introduce some notations. For $i = 1, 2$, let n_i and k_i be positive integers such that $k_i \leq n_i$. For an $n_1 \times n_2$ matrix $\boldsymbol{A} := (a_{ij}; 1 \leq i \leq n_1, 1 \leq j \leq n_2)$, we denote the $k_1 \times k_2$ northeast sub-matrix of $\boldsymbol{A}$ by $[\boldsymbol{A}]^{(k_1,k_2)}$, i.e.,

$$[\boldsymbol{A}]^{(k_1,k_2)} = (a_{ij}; 1 \leq i \leq k_1, n_2 - k_2 < j \leq n_2). \tag{2.5}$$

For a positive integer n, let $\boldsymbol{I}_n$ denote the $n \times n$ identity matrix, and $\boldsymbol{O}$ denote a zero matrix whose size will be determined in the context. Throughout this paper, the next lemma is useful when we compute the convolution of matrix exponentials.

Lemma 2.1 *For positive integers n_1 and n_2, let $\boldsymbol{\alpha}$, $\boldsymbol{X}$, $\boldsymbol{b}$, and $\boldsymbol{Y}$ be $1 \times n_1$, $n_1 \times n_1$, $n_1 \times 1$, and $n_2 \times n_2$ matrices, respectively. For $x \geq 0$, the convolution of matrix exponentials*

$$\int_0^x \boldsymbol{\alpha} \exp(\boldsymbol{X}u)\boldsymbol{b} \exp(\boldsymbol{Y}(x-u))\mathrm{d}u \tag{2.6}$$

is computed as follows. Let $\boldsymbol{M}_{11} := \boldsymbol{I}_{n_2} \otimes \boldsymbol{X}$, $\boldsymbol{M}_{12} := \boldsymbol{I}_{n_2} \otimes \boldsymbol{b}$, and $\boldsymbol{M}_{22} := \boldsymbol{Y}$, where $\otimes$ is the Kronecker product, then (2.6) is given by

$$\int_0^x \boldsymbol{\alpha} \exp(\boldsymbol{X}u)\boldsymbol{b} \exp(\boldsymbol{Y}(x-u))\mathrm{d}u = (\boldsymbol{I}_{n_2} \otimes \boldsymbol{\alpha})[\exp(\boldsymbol{M}x)]^{(n_1 n_2, n_2)}, \tag{2.7}$$

where

$$\boldsymbol{M} := \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{O} & \boldsymbol{M}_{22} \end{pmatrix}. \tag{2.8}$$

Proof. The proof is similar to that of Theorem 1 of [34]. According to the partition in (2.8), we denote $\boldsymbol{F}(x) := \exp(\boldsymbol{M}x)$ by

$$\boldsymbol{F}(x) = \begin{pmatrix} \boldsymbol{F}_{11}(x) & \boldsymbol{F}_{12}(x) \\ \boldsymbol{O} & \boldsymbol{F}_{22}(x) \end{pmatrix}. \tag{2.9}$$

We note that $\boldsymbol{F}'(x) := \frac{\mathrm{d}}{\mathrm{d}x}\boldsymbol{F}(x) = \boldsymbol{M}\boldsymbol{F}(x)$, i.e.,

$$\boldsymbol{F}'_{11}(x) = \boldsymbol{M}_{11}\boldsymbol{F}_{11}(x), \tag{2.10}$$

$$\boldsymbol{F}'_{12}(x) = \boldsymbol{M}_{11}\boldsymbol{F}_{12}(x) + \boldsymbol{M}_{12}\boldsymbol{F}_{22}(x), \tag{2.11}$$

$$\boldsymbol{F}'_{22}(x) = \boldsymbol{M}_{22}\boldsymbol{F}_{22}(x). \tag{2.12}$$

From (2.10), (2.12), and $\boldsymbol{F}(0) = \boldsymbol{I}_{n_1 n_2 + n_2}$, we have

$$\boldsymbol{F}_{11}(x) = \exp(\boldsymbol{M}_{11}x), \quad \boldsymbol{F}_{22}(x) = \exp(\boldsymbol{M}_{22}x). \tag{2.13}$$

From (2.11) and the second equation of (2.13), we have

$$\boldsymbol{F}'_{12}(x) = \boldsymbol{M}_{11}\boldsymbol{F}_{12}(x) + \boldsymbol{M}_{12}\exp(\boldsymbol{M}_{22}x). \tag{2.14}$$

It is readily seen that the solution of the differential equation (2.14) is given by

$$\boldsymbol{F}_{12}(x) = \int_0^x \exp(\boldsymbol{M}_{11}(x-u))\boldsymbol{M}_{12}\exp(\boldsymbol{M}_{22}u)\mathrm{d}u \tag{2.15}$$

$$= \int_0^x (\boldsymbol{I}_{n_2} \otimes \exp(\boldsymbol{X}(x-u)))(\boldsymbol{I}_{n_2} \otimes \boldsymbol{b})\exp(\boldsymbol{Y}u)\mathrm{d}u. \tag{2.16}$$

By pre-multiplying $\boldsymbol{I}_{n_2} \otimes \boldsymbol{\alpha}$ to (2.16), we obtain (2.7). □

By choosing $\boldsymbol{\alpha} := 1$, $\boldsymbol{X} := 0$, and $\boldsymbol{b} := 1$ in Lemma 2.1, we obtain the next result.

Corollary 2.1 *For an $n_2 \times n_2$ matrix $\boldsymbol{Y}$, we have*

$$\int_0^x \exp(\boldsymbol{Y}u)\mathrm{d}u = \left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_{n_2} \\ \boldsymbol{O} & \boldsymbol{Y} \end{pmatrix} x\right)\right]^{(n_2,n_2)}. \tag{2.17}$$

In the subsequent subsections, we present a computational procedure for the stationary density of the workload.

2.2 Stationary density in mode 0

In this subsection we determine the steady-state density of the workload during the times in which the server is in mode 0. Such an inactive period lasts from the instant the system becomes empty until the next instant in which the workload exceeds a certain threshold level d_1. Assume that $Z_0 = 0$, i.e., there is no workload in the system and the server is in mode 0 at time 0. For $i \in \mathbb{N}_0 := \{0, 1, 2, \ldots\}$, let θ_i denote the i-th arrival time after time 0, where $\theta_0 := 0$. For $x \geq 0$, let

$$\nu(x) := \sup\{i \in \mathbb{N}_0; Z_{\theta_i} \leq x\}, \tag{2.18}$$

which is the number of customer arrivals until the workload becomes larger than x. We note that $m(x) := \mathsf{E}[\nu(x)]$ is the renewal function with the renewal interval distribution $B(x)$ $(x \geq 0)$ in (2.1). From Theorem 3.1.2 in [20], we have for $0 \leq x < d_1$,

$$m(x) = \int_0^x \boldsymbol{\tau}\exp(\boldsymbol{D}u)\boldsymbol{t}\mathrm{d}u = \boldsymbol{\tau}\left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_N \\ \boldsymbol{O} & \boldsymbol{D} \end{pmatrix} x\right)\right]^{(N,N)}\boldsymbol{t}, \tag{2.19}$$

$$m'(x) = \boldsymbol{\tau}\exp(\boldsymbol{D}x)\boldsymbol{t}, \tag{2.20}$$

where $\boldsymbol{t} := -\boldsymbol{T}\boldsymbol{1}$ and $\boldsymbol{D} := \boldsymbol{T} + \boldsymbol{t}\boldsymbol{\tau}$, and the second equation in (2.19) follows from Corollary 2.1.

Remark 2.1 By Theorem 3.1.4 in [20], the renewal function and its derivative are given in more explicit formulas as follows: for $x \geq 0$,

$$m(x) = \eta^{-1}x + \boldsymbol{\tau}(\exp(\boldsymbol{D}x) - \boldsymbol{I})(\boldsymbol{D} - \boldsymbol{\Pi})^{-1}\boldsymbol{t}, \tag{2.21}$$

$$m'(x) = \eta^{-1} + \boldsymbol{\tau}\boldsymbol{D}\exp(\boldsymbol{D}x)(\boldsymbol{D} - \boldsymbol{\Pi})^{-1}\boldsymbol{t}, \tag{2.22}$$

where $\mathbf{\Pi} := \mathbf{1}\boldsymbol{\pi}$ and $\boldsymbol{\pi}$ is the stationary distribution of $\boldsymbol{D}$, i.e., $\boldsymbol{\pi}$ is the nonnegative solution of $\boldsymbol{\pi}\boldsymbol{D} = \mathbf{0}$ and $\boldsymbol{\pi}\mathbf{1} = 1$. In this paper, we need to compute some convolutions including $m'(x)$. Therefore, we use (2.19) and (2.20) to obtain our formulas in a simpler form.

We are ready to consider the stationary density of the workload when the server is in mode 0. Let m_i ($i = 0, 1$) be the mean length of a mode-i interval. From Subsection 2.1 of [29] we have

$$m_0 = \lambda^{-1}(1 + m(d_1)). \tag{2.23}$$

Let $v(x|S = 0)$, where $x > 0$, denote the conditional stationary density of the workload when the server is in mode 0, which is given by (cf. Subsection 2.1 of [29])

$$v(x|S = 0) = \begin{cases} \frac{\lambda^{-1}m'(x)}{m_0}, & 0 < x \le d_1, \\ 0, & x > d_1. \end{cases} \tag{2.24}$$

From (2.23) and (2.24), we have

$$v(x|S = 0) = \begin{cases} \frac{m'(x)}{1+m(d_1)}, & 0 < x \le d_1, \\ 0, & x > d_1. \end{cases} \tag{2.25}$$

For $i = 0, 1$, let $p_i := \Pr(S = i)$, which is the marginal distribution when the server is in mode i, and it is readily seen that $p_i = m_i/(m_0 + m_1)$, i.e.,

$$p_0 = \frac{1 + m(d_1)}{1 + m(d_1) + \lambda m_1}, \quad p_1 = \frac{\lambda m_1}{1 + m(d_1) + \lambda m_1}, \tag{2.26}$$

where m_1 later will be determined by using a normalizing condition. Let $v_0(x)$, $0 < x \le d_1$, denote the unconditional stationary density of the workload when the server is in mode 0, i.e., $v_0(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Pr(Z \le x, S = 0)$, then we have

$$v_0(x) = \begin{cases} p_0 \cdot v(x|S = 0), & 0 < x \le d_1, \\ 0, & x > d_1. \end{cases} \tag{2.27}$$

Furthermore, let $V(0) := \Pr(Z = 0, S = 0)$, which is the marginal probability that there are no customers in the system, and which is given by $V(0) = \lambda^{-1}/(m_0 + m_1)$, i.e.,

$$V(0) = \frac{1}{1 + m(d_1) + \lambda m_1}. \tag{2.28}$$

Combining (2.25), (2.26) and (2.27), the results in this subsection can be summarized as follows.

Lemma 2.2 *The stationary density of the workload when the server is in mode* 0 *equals*

$$v_0(x) = \begin{cases} \frac{m'(x)}{1+m(d_1)+\lambda m_1}, & 0 < x \le d_1, \\ 0, & x > d_1, \end{cases} \tag{2.29}$$

where $m(x)$ is given by (2.19) and $m'(x)$ by (2.20); the probability that there is no customer in the system is given by (2.28).

In the next two subsections we determine the stationary density in mode 1, distinguishing between the cases that the workload is above d_1 (Subsection 2.3) and below d_1 (Subsection 2.4).

2.3 Stationary density in mode 1 when the workload is above d_1

Let $v_1(x)$, where $x > 0$, denote the stationary density of the workload when the server is in mode 1, i.e., $v_1(x) = \frac{\mathrm{d}}{\mathrm{d}x}\Pr(Z \le x, S = 1)$. We use the level crossing technique (see, e.g., [5]), which states that, in steady state, each workload level is crossed just as often from above and from below. Hence the stationary density of the workload, denoted by $v(x) := v_0(x) + v_1(x)$, where $x > 0$, satisfies the following relations.

$$r_1(v(x) - v_0(x)) = \lambda V(0)\overline{B}(x) + \lambda \int_0^x \overline{B}(x-y)v(y)\mathrm{d}y, \quad 0 < x \le d_1, \tag{2.30}$$

$$r_2 v(x) = \lambda V(0)\overline{B}(x) + \lambda \int_0^x \overline{B}(x-y)v(y)\mathrm{d}y, \quad d_1 < x < \infty, \tag{2.31}$$

where $\overline{B}(x) = \boldsymbol{\tau}\exp(\boldsymbol{T}x)\mathbf{1}$ for $x \ge 0$. We first solve (2.31), and show that the solution is given by a matrix exponential form. To this end, for $x > d_1$, we consider a $(1 \times N)$ row vector, which is denoted by $\boldsymbol{v}(x)$, and satisfies the following equation:

$$r_2\boldsymbol{v}(x) = \lambda V(0)\boldsymbol{\tau}\exp(\boldsymbol{T}x) + \lambda \int_0^x \boldsymbol{v}(y)\mathbf{1}\boldsymbol{\tau}\exp(\boldsymbol{T}(x-y))\mathrm{d}y, \quad d_1 < x < \infty. \tag{2.32}$$

We note that $v_1(x) = v(x) = \boldsymbol{v}(x)\mathbf{1}$ for $x > d_1$ from (2.31) and (2.32). Similar to [30], by taking the derivative of (2.32) with respect to x, we have

$$\begin{aligned} r_2\boldsymbol{v}'(x) &= \lambda V(0)\boldsymbol{\tau}\exp(\boldsymbol{T}x)\boldsymbol{T} + \lambda\boldsymbol{v}(x)\mathbf{1}\boldsymbol{\tau} + \lambda\int_0^x \boldsymbol{v}(y)\mathbf{1}\boldsymbol{\tau}\exp(\boldsymbol{T}(x-y))\mathrm{d}y\cdot\boldsymbol{T} \\ &= r_2\boldsymbol{v}(x)(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}), \quad d_1 < x < \infty, \end{aligned} \tag{2.33}$$

where the last equality follows by applying (2.32) to the last term of the first line of (2.33). The solution of (2.33) is given by the following matrix exponential form:

$$\boldsymbol{v}(x) = \tilde{\boldsymbol{u}}\exp((\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})(x - d_1)), \quad d_1 < x < \infty, \tag{2.34}$$

where $\tilde{\boldsymbol{u}}$ is a $(1 \times N)$ row vector, which will be determined later (see (2.48)).

The results in this subsection are summarized as follows.

Lemma 2.3 *For $x > d_1$, the stationary density of the workload when the server is in mode 1, is given by $v_1(x) = \boldsymbol{v}(x)\mathbf{1}$, where*

$$\boldsymbol{v}(x) = \tilde{\boldsymbol{u}}\exp((\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})(x - d_1)), \quad x > d_1, \tag{2.35}$$

where $\tilde{\boldsymbol{u}}$ is given by (2.48).

2.4 Stationary density in mode 1 when the workload is below d_1

We next solve (2.30). Since $v(x) = v_0(x) + v_1(x)$, (2.30) is rewritten into

$$r_1 v_1(x) = \lambda V(0)\overline{B}(x) + \lambda \int_0^x (v_0(y) + v_1(y))\overline{B}(x-y)\mathrm{d}y, \quad 0 < x \le d_1. \tag{2.36}$$

To solve (2.36), we consider $(1 \times N)$ row vectors $\boldsymbol{v}_0(x)$ and $\boldsymbol{v}_1(x)$ satisfying the following equation: for $0 < x \le d_1$,

$$r_1 \boldsymbol{v}_1(x) = \lambda V(0)\boldsymbol{\tau}\exp(\boldsymbol{T}x) + \lambda \int_0^x (\boldsymbol{v}_0(y) + \boldsymbol{v}_1(y))\mathbf{1}\boldsymbol{\tau}\exp(\boldsymbol{T}(x-y))\mathrm{d}y. \tag{2.37}$$

We note that $v_0(x) = \boldsymbol{v}_0(x)\mathbf{1}$ and $v_1(x) = \boldsymbol{v}_1(x)\mathbf{1}$ for $0 < x \le d_1$. Similar to (2.33), by taking the derivative of (2.37) with respect to x, and applying (2.37) to the derivative, we have

$$\boldsymbol{v}_1'(x) = \boldsymbol{v}_1(x)(\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}) + \lambda r_1^{-1} v_0(x)\boldsymbol{\tau}, \quad 0 < x \le d_1. \tag{2.38}$$

Suppose that the solution of (2.38) is given by the following form:

$$\boldsymbol{v}_1(x) = \boldsymbol{w}(x)\exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})x), \quad 0 < x \le d_1, \tag{2.39}$$

where $\boldsymbol{w}(x)$ is a $(1 \times N)$ row vector and is determined as follows. By taking the derivative of (2.39) with respect to x, we have

$$\boldsymbol{v}_1'(x) = \boldsymbol{v}_1(x)(\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}) + \boldsymbol{w}'(x)\exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})x), \quad 0 < x \le d_1. \tag{2.40}$$

By comparing (2.38) with (2.40), we have

$$\boldsymbol{w}'(x) = \lambda r_1^{-1} v_0(x)\boldsymbol{\tau}\exp(-(\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})x), \quad 0 < x \le d_1. \tag{2.41}$$

We then have

$$\boldsymbol{w}(x) = \int_0^x \lambda r_1^{-1} v_0(u)\boldsymbol{\tau}\exp(-(\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})u)\mathrm{d}u + \boldsymbol{c}, \quad 0 < x \le d_1, \tag{2.42}$$

where $\boldsymbol{c}$ is a $(1 \times N)$ row vector and is determined as follows (see (2.45) below). By taking the limit $x \downarrow 0$ of (2.37), we have

$$r_1 \boldsymbol{v}_1(0+) = \lambda V(0)\boldsymbol{\tau} = \frac{\lambda\boldsymbol{\tau}}{1 + m(d_1) + \lambda m_1}, \tag{2.43}$$

where the last equality follows from (2.28). From (2.39) and (2.42), we have

$$\boldsymbol{v}_1(0+) = \boldsymbol{c}. \tag{2.44}$$

From (2.43) and (2.44), we obtain

$$\boldsymbol{c} = \frac{\lambda r_1^{-1}\boldsymbol{\tau}}{1 + m(d_1) + \lambda m_1}. \tag{2.45}$$

We are now ready to formulate the following lemma.

Lemma 2.4 *For $0 < x \leq d_1$, the stationary density $v_1(x)$ when the server is in mode 1 is given by $v_1(x) = \boldsymbol{v}_1(x)\mathbf{1}$, where*

$$\boldsymbol{v}_1(x) = \frac{\lambda r_1^{-1}\boldsymbol{\tau}}{1 + m(d_1) + \lambda m_1}\left\{(\boldsymbol{I}_N \otimes \boldsymbol{\tau})\left[\exp\left(\overline{\boldsymbol{M}}x\right)\right]^{(N^2,N)} + \exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})x)\right\}, \quad (2.46)$$

with

$$\overline{\boldsymbol{M}} := \begin{pmatrix} \boldsymbol{I}_N \otimes \boldsymbol{D} & \boldsymbol{I}_N \otimes \boldsymbol{t} \\ \boldsymbol{O} & \lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T} \end{pmatrix}. \quad (2.47)$$

Furthermore,

$$\tilde{\boldsymbol{u}} = \frac{\lambda r_2^{-1}\boldsymbol{\tau}}{1 + m(d_1) + \lambda m_1}\left\{(\boldsymbol{I}_N \otimes \boldsymbol{\tau})\left[\exp\left(\overline{\boldsymbol{M}}d_1\right)\right]^{(N^2,N)} + \exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})d_1)\right\}. \quad (2.48)$$

Proof. From (2.29), (2.39), and (2.42) (see also (2.45)), $\boldsymbol{v}_1(x)$ for $0 < x \leq d_1$ is given by the sum of the following two terms,

$$\frac{\lambda r_1^{-1}\boldsymbol{\tau}}{1 + m(d_1) + \lambda m_1}\int_0^x \boldsymbol{\tau}\exp(\boldsymbol{D}u)\boldsymbol{t}\exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})(x-u))\mathrm{d}u \quad (2.49)$$

and

$$\frac{\lambda r_1^{-1}\boldsymbol{\tau}}{1 + m(d_1) + \lambda m_1}\exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})x). \quad (2.50)$$

From Lemma 2.1, the integral in (2.49) is given by

$$(\boldsymbol{I}_N \otimes \boldsymbol{\tau})\left[\exp\left(\overline{\boldsymbol{M}}x\right)\right]^{(N^2,N)}. \quad (2.51)$$

From (2.49), (2.50), and (2.51), we obtain (2.46). Using (2.34) and (2.46), and observing that $\boldsymbol{v}_1(d_1+) = r_1 r_2^{-1}\boldsymbol{v}_1(d_1-)$ from (2.32) and (2.37), we obtain (2.48). □

2.5 Computation of the mean active period

In this subsection, we compute the mean active period m_1 from the normalizing condition, i.e.,

$$V(0) + \int_0^{d_1} v_0(x)\mathrm{d}x + \int_0^{d_1} v_1(x)\mathrm{d}x + \int_{d_1}^{\infty} v_1(x)\mathrm{d}x = 1. \quad (2.52)$$

The three integrals are successively obtained in Lemmas 2.5, 2.6 and 2.7. We first use Lemma 2.2 to obtain $\int_0^{d_1} v_0(x)\mathrm{d}x$.

Lemma 2.5

$$\int_0^{d_1} v_0(x)\mathrm{d}x = \frac{m(d_1)}{1 + m(d_1) + \lambda m_1}, \quad (2.53)$$

where

$$m(d_1) = \boldsymbol{\tau}\left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_N \\ \boldsymbol{O} & \boldsymbol{D} \end{pmatrix} d_1\right)\right]^{(N,N)}\boldsymbol{t}. \quad (2.54)$$

We next consider $\int_0^{d_1} v_1(x)\mathrm{d}x$. From Corollary 2.1, we have

$$\begin{aligned}\int_0^{d_1} \left[\exp\left(\overline{\boldsymbol{M}}x\right)\right]^{(N^2,N)} \mathrm{d}x &= \left[\int_0^{d_1} \exp\left(\overline{\boldsymbol{M}}x\right) \mathrm{d}x\right]^{(N^2,N)} \\ &= \left[\left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_{N^2+N} \\ \boldsymbol{O} & \overline{\boldsymbol{M}} \end{pmatrix} d_1\right)\right]^{(N^2+N,N^2+N)}\right]^{(N^2,N)} \\ &= \left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_{N^2+N} \\ \boldsymbol{O} & \overline{\boldsymbol{M}} \end{pmatrix} d_1\right)\right]^{(N^2,N)}, \end{aligned} \tag{2.55}$$

and

$$\int_0^{d_1} \exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})x)\mathrm{d}x = \left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_N \\ \boldsymbol{O} & \lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T} \end{pmatrix} d_1\right)\right]^{(N,N)}. \tag{2.56}$$

The next lemma immediately follows from Lemma 2.4, (2.55) and (2.56).

Lemma 2.6 *The second integral in the lefthand side of (2.52) is given by*

$$\begin{aligned}\int_0^{d_1} v_1(x)\mathrm{d}x = \frac{\lambda r_1^{-1}\boldsymbol{\tau}}{1 + m(d_1) + \lambda m_1}\Bigg\{ &(\boldsymbol{I}_N \otimes \boldsymbol{\tau}) \left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_{N^2+N} \\ \boldsymbol{O} & \overline{\boldsymbol{M}} \end{pmatrix} d_1\right)\right]^{(N^2,N)} \\ &+ \left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_N \\ \boldsymbol{O} & \lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T} \end{pmatrix} d_1\right)\right]^{(N,N)} \Bigg\}\mathbf{1}, \end{aligned} \tag{2.57}$$

where $\overline{\boldsymbol{M}}$ is given by (2.47).

We finally compute $\int_{d_1}^{\infty} v_1(x)\mathrm{d}x$ as follows.

Lemma 2.7 *From Lemma 2.3 and the stability condition (2.4), we have*

$$\int_{d_1}^{\infty} v_1(x)\mathrm{d}x = \tilde{\boldsymbol{u}}\left(-(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})\right)^{-1}\mathbf{1}, \tag{2.58}$$

where $\tilde{\boldsymbol{u}}$ is given by (2.48).

Proof. We first show that $\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}$ is invertible, i.e., $(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})^{-1}$ exists. To this end, we consider the following equation

$$\boldsymbol{x}(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}) = \mathbf{0}, \tag{2.59}$$

where $\boldsymbol{x}$ is a $(1 \times N)$ row vector. By post-multiplying the lefthand side of (2.59) by $r_2(-\boldsymbol{T})^{-1}\mathbf{1}$, we have

$$\boldsymbol{x}\mathbf{1}\{\lambda\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbf{1} - r_2\} = 0, \tag{2.60}$$

and from the stability condition (2.4), we have

$$\boldsymbol{x}\mathbf{1} = 0. \tag{2.61}$$

By applying (2.61) to (2.59) and noting the existence of $(-\boldsymbol{T})^{-1}$, we obtain $\boldsymbol{x} = \mathbf{0}$, which implies that $(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})^{-1}$ exists.

We next show (2.58). From Lemma 2.3 and the existence of $(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})^{-1}$ we have

$$\begin{aligned}\int_{d_1}^{\infty} v_1(x)\mathrm{d}x &= \tilde{\boldsymbol{u}}\left\{\lim_{y\to\infty}\int_0^y \exp((\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})x)\mathrm{d}x\right\}\mathbf{1} \\ &= \tilde{\boldsymbol{u}}\left(-(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})\right)^{-1}\left\{\boldsymbol{I}_N - \lim_{y\to\infty}\exp((\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})y)\right\}\mathbf{1}.\end{aligned} \tag{2.62}$$

It remains to show that

$$\lim_{y\to\infty}\exp((\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})y) = \boldsymbol{O}. \tag{2.63}$$

Noting that $\boldsymbol{\pi}$ is the stationary distribution of $\boldsymbol{T} + \boldsymbol{t}\boldsymbol{\tau}$, i.e., $\boldsymbol{\pi}\mathbf{1} = 1$ and $\boldsymbol{\pi}\boldsymbol{T} = -\boldsymbol{\pi}\boldsymbol{t}\boldsymbol{\tau}$, we have

$$\begin{aligned}\boldsymbol{\pi}(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}) &= (\lambda r_2^{-1} - \boldsymbol{\pi}\boldsymbol{t})\boldsymbol{\tau} \\ &= \left(\lambda r_2^{-1} - \frac{1}{\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbf{1}}\right)\boldsymbol{\tau},\end{aligned} \tag{2.64}$$

where the last equation follows from $\boldsymbol{\pi}\boldsymbol{T} = -\boldsymbol{\pi}\boldsymbol{t}\boldsymbol{\tau}$, i.e., $1 = \boldsymbol{\pi}\mathbf{1} = \boldsymbol{\pi}\boldsymbol{t}\boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbf{1}$. Since $\boldsymbol{\tau}$ is a nonnegative and nonzero vector, (2.64) and the stability condition (2.4) imply that $\boldsymbol{\pi}(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}) \le \mathbf{0}$ and $\boldsymbol{\pi}(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}) \neq \mathbf{0}$. From Theorem 1.6 (b) in [31], the Perron-Frobenius eigenvalue of $\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}$ is negative, which implies (2.63). □

From Lemmas 2.5–2.7, the mean active period is given as follows.

Theorem 2.1 *The mean active period for the case of two processing speeds is given by*

$$\begin{aligned}m_1 = r_1^{-1}\boldsymbol{\tau}&\left\{(\boldsymbol{I}_N \otimes \boldsymbol{\tau})\left[\exp\left(\begin{pmatrix}\boldsymbol{O} & \boldsymbol{I}_{N^2+N} \\ \boldsymbol{O} & \overline{\boldsymbol{M}}\end{pmatrix} d_1\right)\right]^{(N^2,N)} + \left[\exp\left(\begin{pmatrix}\boldsymbol{O} & \boldsymbol{I}_N \\ \boldsymbol{O} & \lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}\end{pmatrix} d_1\right)\right]^{(N,N)}\right\}\mathbf{1} \\ &+ r_2^{-1}\boldsymbol{\tau}\left\{(\boldsymbol{I}_N \otimes \boldsymbol{\tau})\left[\exp\left(\overline{\boldsymbol{M}}d_1\right)\right]^{(N^2,N)} + \exp((\lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})d_1)\right\}\left(-(\lambda r_2^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})\right)^{-1}\mathbf{1}.\end{aligned} \tag{2.65}$$

3 Extension to the case of multiple processing speeds

In this section, we consider an extension of the model of the previous section to the case of an arbitrary number of different (constant) processing speeds. Subsection 3.1 provides a detailed model description. In Subsection 3.2 we consider the stationarity workload density when the server is inactive (briefly, as this is very similar to the result in Subsection 2.2). Subsections 3.3 and 3.4 are successively devoted to the stationary workload density when the server is active and the workload is above, respectively below, the threshold level d_K.

3.1 Model description

In this section we extend the assumptions (i), (ii), and the service processing speed $s_i(x)$ ($i = 0, 1$ and $x \geq 0$) in Section 2 as follows. Let $\{d_k; k = 0, 1, 2, \ldots, K_0\}$ be an increasing sequence such that $d_0 := 0$.

(i)' When the workload Z_t hits $d_0 (= 0)$ at time t, the server enters mode 0, i.e., $S_t = 0$. After that, the server remains in mode 0 until the workload exceeds a threshold $d_K > 0$, where $K \leq K_0$ is a positive integer, i.e., $S_u = 0$ as long as $Z_u < d_K$ for $u > t$.

(ii)' When the workload Z_t exceeds d_K at time t, the server changes its mode from 0 to 1, i.e., $S_t = 1$. The server remains in that mode until the workload hits 0 again, i.e., $S_u = 1$ as long as $Z_u > 0$ for $u > t$.

The processing speed of the server depends on both the server's mode and workload in the following way. Let $s_i(x)$, where $i = 0, 1$ and $x \geq 0$, denote the processing speed of the server when $(Z_t, S_t) = (x, i)$; it is defined as follows:

$$s_0(x) = 0, \quad 0 \leq x \leq d_K, \tag{3.1}$$

$$s_1(x) = r_k, \quad x \in J_k := (d_{k-1}, d_k], \ k = 1, 2, 3, \ldots, K_0, \tag{3.2}$$

where $\{r_k; k = 1, 2, \ldots, K_0\}$ is a positive valued sequence and $d_{K_0} = \infty$. Note that $\{s_0(x); x > d_K\}$ and $s_1(0)$ need not be specified in view of the definition of modes 0 and 1. Also note that d_K coincides with one of the switching levels d_k of mode 1. However, this is no restriction, as one could take $r_K = r_{K+1}$. A sample path of the workload is presented in Figure 2.

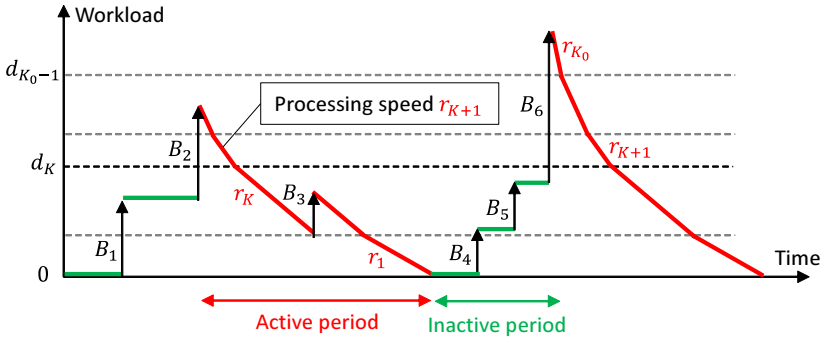

Figure 2: The workload process for the case of two processing speeds.

We assume that the Markov process (Z_t, S_t) is positive-recurrent, and denote its replica in steady state by (Z, S). Similar to (2.4), the stability condition of this Markov process is given by

$$\frac{\lambda \boldsymbol{\tau}(-\boldsymbol{T})^{-1}\mathbf{1}}{r_{K_0}} < 1. \tag{3.3}$$

In the subsequent subsections, we present a computational procedure for the stationary density of the workload.

3.2 Stationary density in mode 0

Similar to Lemma 2.2, the stationary density in mode 0, and the marginal probability that there is no customer in the system, are given as follows.

Lemma 3.1 *The stationary density of the workload when the server is in mode* 0 *is given by*

$$v_0(x) = \begin{cases} \frac{m'(x)}{1+m(d_K)+\lambda m_1}, & 0 < x \le d_K, \\ 0, & x > d_K, \end{cases} \tag{3.4}$$

where $m(x)$ *is given by (2.19) and* $m'(x)$ *by (2.20), and* m_1 *is the mean active period which will be determined later (see (3.48)). The marginal probability that there is no customer in the system is given by*

$$V(0) = \frac{1}{1 + m(d_K) + \lambda m_1}. \tag{3.5}$$

In the next two subsections we determine the stationary density in mode 1, distinguishing between the cases that the workload is above d_K (Subsection 3.3) and below d_K (Subsection 3.4).

3.3 Stationary density in mode 1 when the workload is above d_K

Similar to (2.30) and (2.31), the stationary density of the workload satisfies the following relations.

$$r_k(v(x) - v_0(x)) = \lambda V(0)\overline{B}(x) + \lambda \int_0^x \overline{B}(x-y)v(y)\mathrm{d}y, \quad x \in J_k,\ 1 \le k \le K, \tag{3.6}$$

$$r_k v(x) = \lambda V(0)\overline{B}(x) + \lambda \int_0^x \overline{B}(x-y)v(y)\mathrm{d}y, \quad x \in J_k,\ K+1 \le k \le K_0. \tag{3.7}$$

To find a solution of (3.7), we consider the following equation:

$$r_k \boldsymbol{v}(x) = \lambda V(0)\boldsymbol{\tau}\exp(\boldsymbol{T}x) + \lambda \int_0^x \boldsymbol{v}(y)\mathbf{1}\boldsymbol{\tau}\exp(\boldsymbol{T}(x-y))\mathrm{d}y, \quad x \in J_k,\ K+1 \le k \le K_0. \tag{3.8}$$

We note that $v_1(x) = v(x) = \boldsymbol{v}(x)\mathbf{1}$ for $x > d_K$ by (3.7) and (3.8). Similar to (2.34), by taking the derivative of (3.8) with respect to x, and applying (3.8) to the derivative, we have

$$\boldsymbol{v}'(x) = \boldsymbol{v}(x)(\lambda r_k^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}), \quad x \in J_k,\ K+1 \le k \le K_0. \tag{3.9}$$

The solution of (3.9) is given by the following matrix exponential form:

$$\boldsymbol{v}(x) = \tilde{\boldsymbol{u}}_k \boldsymbol{U}_k(x), \quad x \in J_k,\ K+1 \le k \le K_0, \tag{3.10}$$

where $\{\tilde{\boldsymbol{u}}_k; K+1 \le k \le K_0\}$ is a set of $(1 \times N)$ row vectors, which will be determined later (see (3.16)), and

$$\boldsymbol{U}_k(x) := \exp((\lambda r_k^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})(x - d_{k-1})), \quad x \in J_k,\ 1 \le k \le K_0. \tag{3.11}$$

We note that

$$\lim_{x \downarrow d_{k-1}} \boldsymbol{U}_k(x) = \boldsymbol{I}_N, \quad 1 \le k \le K_0. \tag{3.12}$$

The sequence $\{\tilde{\boldsymbol{u}}_k; K+1 \le k \le K_0\}$ is recursively determined as follows. By taking the two limits $x \uparrow d_k$ and $x \downarrow d_k$ of (3.8), we have

$$r_k \boldsymbol{v}(d_k-) = r_{k+1}\boldsymbol{v}(d_k+), \quad K+1 \le k \le K_0 - 1. \tag{3.13}$$

From (3.10) and (3.12), we have, for $K+1 \le k \le K_0 - 1$,

$$\boldsymbol{v}(d_k-) = \tilde{\boldsymbol{u}}_k \boldsymbol{U}_k(d_k), \quad \boldsymbol{v}(d_k+) = \tilde{\boldsymbol{u}}_{k+1}. \tag{3.14}$$

From (3.13) and (3.14), we have

$$\tilde{\boldsymbol{u}}_{k+1} = \frac{r_k}{r_{k+1}} \tilde{\boldsymbol{u}}_k \boldsymbol{U}_k(d_k), \quad K+1 \le k \le K_0 - 1, \tag{3.15}$$

which yields

$$\tilde{\boldsymbol{u}}_k = \frac{r_{K+1}}{r_k} \tilde{\boldsymbol{u}}_{K+1} \left\{\Pi_{i=K+1}^{k-1} \boldsymbol{U}_i(d_i)\right\}, \quad K+1 \le k \le K_0, \tag{3.16}$$

where $\tilde{\boldsymbol{u}}_{K+1}$ will be determined later (see (3.39)).

In what follows, we summarize the results in this subsection. To this end, we introduce the following notation:

$$\hat{\boldsymbol{U}}_{k,l} := \Pi_{i=k}^{l} \boldsymbol{U}_i(d_i), \quad 1 \le k, l \le K_0, \tag{3.17}$$

where the empty product is an identity matrix, i.e., $\hat{\boldsymbol{U}}_{k,l} = \boldsymbol{I}_N$ for $k > l$.

Lemma 3.2 *For $x > d_K$, the stationary density of the workload when the server is in mode 1, is given by $v_1(x) = \boldsymbol{v}(x)\mathbf{1}$, where*

$$\boldsymbol{v}(x) = \frac{r_{K+1}}{r_k} \tilde{\boldsymbol{u}}_{K+1} \hat{\boldsymbol{U}}_{K+1,k-1} \boldsymbol{U}_k(x), \tag{3.18}$$

$$\boldsymbol{U}_k(x) = \exp((\lambda r_k^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})(x - d_{k-1})), \tag{3.19}$$

for $x \in J_k$ and $K+1 \le k \le K_0$, and $\tilde{\boldsymbol{u}}_{K+1}$ is given by (3.39).

3.4 Stationary density in mode 1 when the workload is below d_K

To solve (3.6), we consider $(1 \times N)$ row vectors $\boldsymbol{v}_0(x)$ and $\boldsymbol{v}_1(x)$ satisfying the following equation: for $x \in J_k,\ 1 \le k \le K$,

$$r_k \boldsymbol{v}_1(x) = \lambda V(0) \boldsymbol{\tau} \exp(\boldsymbol{T}x) + \lambda \int_0^x (\boldsymbol{v}_0(y) + \boldsymbol{v}_1(y)) \mathbf{1} \boldsymbol{\tau} \exp(\boldsymbol{T}(x-y)) \mathrm{d}y. \tag{3.20}$$

We note that $v_1(x) = \boldsymbol{v}_1(x)\mathbf{1}$ for $0 < x \le d_K$. Similar to (2.38), by taking the derivative of (3.20) with respect to x, we have

$$\boldsymbol{v}_1'(x) = \boldsymbol{v}_1(x)(\lambda r_k^{-1} \mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}) + \lambda r_k^{-1} v_0(x) \boldsymbol{\tau}, \quad x \in J_k,\ 1 \le k \le K. \tag{3.21}$$

Suppose that the solution of (3.21) is given by the following form:

$$\boldsymbol{v}_1(x) = \boldsymbol{w}_k(x) \boldsymbol{U}_k(x), \quad x \in J_k,\ 1 \le k \le K, \tag{3.22}$$

where $\boldsymbol{U}_k(x)$ is given by (3.11), and $\{\boldsymbol{w}_k(x); 1 \le k \le K\}$ is a set of $(1 \times N)$ row vectors. By taking the derivative in (3.22) and comparing with (3.21), we have

$$\boldsymbol{w}_k'(x) = \lambda r_k^{-1} v_0(x) \boldsymbol{\tau} \exp(-(\lambda r_k^{-1} \mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})(x - d_{k-1})), \quad x \in J_k,\ 1 \le k \le K. \tag{3.23}$$

From (3.23), we have for $x \in J_k,\ 1 \le k \le K$,

$$\boldsymbol{w}_k(x) = \boldsymbol{g}_k(x) + \boldsymbol{c}_k, \tag{3.24}$$

$$\text{with } \boldsymbol{g}_k(x) := \int_{d_{k-1}}^x \lambda r_k^{-1} v_0(u) \boldsymbol{\tau} \exp(-(\lambda r_k^{-1} \mathbf{1}\boldsymbol{\tau} + \boldsymbol{T})(u - d_{k-1})) \mathrm{d}u, \tag{3.25}$$

where $\{\boldsymbol{c}_k; 1 \le k \le K\}$ is a set of $(1 \times N)$ row vectors, which will be determined below (see (3.30) and (3.31)). We note that

$$\boldsymbol{g}_k(d_{k-1}) = \mathbf{0}, \quad 1 \le k \le K. \tag{3.26}$$

In what follows, we summarize the results in this subsection. For simplicity of the exposition, let

$$\boldsymbol{W}_k(x) := (\boldsymbol{I}_N \otimes \boldsymbol{\tau}_k) \left[\exp\left(\overline{\boldsymbol{M}}_k(x - d_{k-1})\right)\right]^{(N^2,N)}, \quad x \in J_k,\ 1 \le k \le K, \tag{3.27}$$

where

$$\boldsymbol{\tau}_k := \boldsymbol{\tau} \exp(\boldsymbol{D} d_{k-1}), \quad \overline{\boldsymbol{M}}_k := \begin{pmatrix} \boldsymbol{I}_N \otimes \boldsymbol{D} & \boldsymbol{I}_N \otimes \boldsymbol{t} \\ \boldsymbol{O} & \lambda r_k^{-1} \mathbf{1}\boldsymbol{\tau} + \boldsymbol{T} \end{pmatrix}. \tag{3.28}$$

Lemma 3.3 *For $0 < x \le d_K$, the stationary density $v_1(x)$ when the server is in mode 1 is given by $v_1(x) = \boldsymbol{v}_1(x)\mathbf{1}$, where for $x \in J_k,\ 1 \le k \le K$,*

$$\boldsymbol{v}_1(x) = \frac{\lambda r_k^{-1} \boldsymbol{\tau}}{1 + m(d_K) + \lambda m_1} \boldsymbol{W}_k(x) + \boldsymbol{c}_k \boldsymbol{U}_k(x), \tag{3.29}$$

where $\{\boldsymbol{c}_k; 1 \le k \le K\}$ *is given by*

$$\boldsymbol{c}_k = \frac{\lambda r_k^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\left\{\sum_{i=1}^{\min\{k-1,K\}} \boldsymbol{W}_i(d_i)\hat{\boldsymbol{U}}_{i+1,k-1} + \hat{\boldsymbol{U}}_{1,k-1}\right\}. \tag{3.30}$$

In particular, $\boldsymbol{c}_1$ *is given by*

$$\boldsymbol{c}_1 = \frac{\lambda r_1^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}. \tag{3.31}$$

Proof. From (3.22) and (3.24), we have

$$\boldsymbol{v}_1(x) = \boldsymbol{g}_k(x)\boldsymbol{U}_k(x) + \boldsymbol{c}_k\boldsymbol{U}_k(x), \quad x \in J_k,\ 1 \le k \le K. \tag{3.32}$$

From (3.4), (3.11), and (3.25), the first term in the righthand side of (3.32) is calculated as follows:

$$\begin{aligned}\boldsymbol{g}_k(x)\boldsymbol{U}_k(x) &= \frac{\lambda r_k^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\int_0^{x-d_{k-1}} \boldsymbol{\tau}\exp(\boldsymbol{D}d_{k-1})\exp(\boldsymbol{D}u)\boldsymbol{t} \\ &\qquad \times \exp((\lambda r_k^{-1}\mathbf{1}\boldsymbol{\tau}+\boldsymbol{T})(x-d_{k-1}-u))\mathrm{d}u \\ &= \frac{\lambda r_k^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\boldsymbol{W}_k(x), \end{aligned} \tag{3.33}$$

where the last equation follows from Lemma 2.1, and then (3.29) is obtained.

It remains to show that $\{\boldsymbol{c}_k; 1 \le k \le K\}$ is given by (3.30). By taking the limit $x \downarrow 0$ of (3.20) and applying (3.5) to it, we have

$$\boldsymbol{v}_1(0+) = \frac{\lambda r_1^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}. \tag{3.34}$$

From (3.11) and (3.27), we have $\boldsymbol{U}_1(0+) = \boldsymbol{I}_N$ and $\boldsymbol{W}_1(0+) = \boldsymbol{O}$, respectively, which imply $\boldsymbol{v}_1(0+) = \boldsymbol{c}_1$ from (3.34). Then we have (3.31).

Similar to (3.13), by taking the two limits $x \uparrow d_k$ and $x \downarrow d_k$ of (3.20) for $1 \le k \le K-1$, we have

$$r_k\boldsymbol{v}_1(d_k-) = r_{k+1}\boldsymbol{v}_1(d_k+), \quad 1 \le k \le K-1. \tag{3.35}$$

For $1 \le k \le K-1$, by taking limits $x \uparrow d_k$ and $x \downarrow d_k$ of (3.29), we then have

$$\boldsymbol{v}_1(d_k-) = (\boldsymbol{g}_k(d_k)+\boldsymbol{c}_k)\boldsymbol{U}_k(d_k), \tag{3.36}$$

$$\begin{aligned}\boldsymbol{v}_1(d_k+) &= (\boldsymbol{g}_{k+1}(d_k)+\boldsymbol{c}_{k+1})\boldsymbol{U}_{k+1}(d_k) \\ &= \boldsymbol{c}_{k+1}, \end{aligned} \tag{3.37}$$

where the last equality follows from (3.12) and (3.26). From (3.35), (3.36), and (3.37), we have

$$\begin{aligned}\boldsymbol{c}_{k+1} &= r_{k+1}^{-1}\left\{r_k\boldsymbol{g}_k(d_k)\boldsymbol{U}_k(d_k) + r_k\boldsymbol{c}_k\boldsymbol{U}_k(d_k)\right\} \\ &= \frac{\lambda r_{k+1}^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\boldsymbol{W}_k(d_k) + r_{k+1}^{-1}r_k\boldsymbol{c}_k\hat{\boldsymbol{U}}_{k,k}, \end{aligned} \tag{3.38}$$

for $1 \le k \le K-1$, where the second equality sign follows from (3.33) and (3.17). By recursively applying (3.38) to itself, we obtain (3.30). □

We now show that $\tilde{\boldsymbol{u}}_{K+1}$ in (3.18) is equal to $\boldsymbol{c}_{K+1}$ (here we extend the definition of $\boldsymbol{c}_k$ in (3.30) to the case $k = K+1$).

Lemma 3.4 *We have $\tilde{\boldsymbol{u}}_{K+1} = \boldsymbol{c}_{K+1}$, i.e.,*

$$\tilde{\boldsymbol{u}}_{K+1} = \frac{\lambda r_{K+1}^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\left\{\sum_{i=1}^{K}\boldsymbol{W}_i(d_i)\hat{\boldsymbol{U}}_{i+1,K}+\hat{\boldsymbol{U}}_{1,K}\right\}. \tag{3.39}$$

Proof. Similar to (3.13), from (3.8) and (3.20), we have

$$r_K\boldsymbol{v}_1(d_K-) = r_{K+1}\boldsymbol{v}_1(d_K+). \tag{3.40}$$

From (3.29) and (3.18) (see also (3.12)), we have

$$\boldsymbol{v}_1(d_K-) = \frac{\lambda r_K^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\boldsymbol{W}_K(d_K)+\boldsymbol{c}_K\boldsymbol{U}_K(d_K), \tag{3.41}$$

$$\boldsymbol{v}_1(d_K+) = \tilde{\boldsymbol{u}}_{K+1}. \tag{3.42}$$

The preceding three equations and (3.30) imply that $\tilde{\boldsymbol{u}}_{K+1} = \boldsymbol{c}_{K+1}$. □

3.5 Computation of the mean active period

We compute the mean active period m_1 by the normalizing condition, i.e.,

$$V(0)+\int_0^{d_K}(v_0(x)+\boldsymbol{v}_1(x)\mathbf{1})\mathrm{d}x+\int_{d_K}^{\infty}\boldsymbol{v}(x)\mathbf{1}\mathrm{d}x = 1. \tag{3.43}$$

From Lemma 3.1, we have

$$V(0)+\int_0^{d_K}v_0(x)\mathrm{d}x = \frac{1+m(d_K)}{1+m(d_K)+\lambda m_1}. \tag{3.44}$$

Note that from Lemma 3.2 and (3.39), for $x \in J_k$ and $K+1 \le k \le K_0$, we have

$$\begin{aligned}\boldsymbol{v}(x) &= \frac{r_{K+1}}{r_k}\frac{\lambda r_{K+1}^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\left\{\sum_{i=1}^{K}\boldsymbol{W}_i(d_i)\hat{\boldsymbol{U}}_{i+1,K}+\hat{\boldsymbol{U}}_{1,K}\right\}\hat{\boldsymbol{U}}_{K+1,k-1}\boldsymbol{U}_k(x)\\ &= \frac{\lambda r_k^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\left\{\sum_{i=1}^{K}\boldsymbol{W}_i(d_i)\hat{\boldsymbol{U}}_{i+1,k-1}+\hat{\boldsymbol{U}}_{1,k-1}\right\}\boldsymbol{U}_k(x)\\ &= \boldsymbol{c}_k\boldsymbol{U}_k(x),\end{aligned} \tag{3.45}$$

where the second and third equations follow from (3.17) and (3.30), respectively. From Lemmas 3.2 and 3.3, and (3.45), we have

$$\int_0^{d_K} \boldsymbol{v}_1(x)\mathbf{1}\mathrm{d}x + \int_{d_K}^{\infty} \boldsymbol{v}(x)\mathbf{1}\mathrm{d}x$$
$$= \frac{\lambda\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\left\{\sum_{k=1}^{K} r_k^{-1}\int_{d_{k-1}}^{d_k}\boldsymbol{W}_k(x)\mathrm{d}x + \sum_{k=1}^{K_0} r_k^{-1}\boldsymbol{C}_k\int_{d_{k-1}}^{d_k}\boldsymbol{U}_k(x)\mathrm{d}x\right\}\mathbf{1}, \quad (3.46)$$

where

$$\boldsymbol{C}_k := \sum_{i=1}^{\min\{k-1,K\}} \boldsymbol{W}_i(d_i)\hat{\boldsymbol{U}}_{i+1,k-1} + \hat{\boldsymbol{U}}_{1,k-1}, \quad 1 \le k \le K_0. \quad (3.47)$$

Therefore, the mean active period is given as follows.

Theorem 3.1 *From (3.43), (3.44), and (3.46), the mean active period is given by*

$$m_1 = \boldsymbol{\tau}\left\{\sum_{k=1}^{K} r_k^{-1}\int_{d_{k-1}}^{d_k}\boldsymbol{W}_k(x)\mathrm{d}x + \sum_{k=1}^{K_0} r_k^{-1}\boldsymbol{C}_k\int_{d_{k-1}}^{d_k}\boldsymbol{U}_k(x)\mathrm{d}x\right\}\mathbf{1}, \quad (3.48)$$

where the integrals are calculated in the same way as (2.55) and (2.56), i.e.,

$$\int_{d_{k-1}}^{d_k}\boldsymbol{W}_k(x)\mathrm{d}x = (\boldsymbol{I}_N\otimes\boldsymbol{\tau}_k)\left[\exp\left(\begin{pmatrix}\boldsymbol{O} & \boldsymbol{I}_{N^2+N}\\ \boldsymbol{O} & \overline{\boldsymbol{M}}_k\end{pmatrix}(d_k-d_{k-1})\right)\right]^{(N^2,N)}, \quad 1\le k\le K, \quad (3.49)$$

$$\int_{d_{k-1}}^{d_k}\boldsymbol{U}_k(x)\mathrm{d}x = \left[\exp\left(\begin{pmatrix}\boldsymbol{O} & \boldsymbol{I}_N\\ \boldsymbol{O} & \lambda r_k^{-1}\mathbf{1}\boldsymbol{\tau}+\boldsymbol{T}\end{pmatrix}(d_k-d_{k-1})\right)\right]^{(N,N)}, \quad 1\le k\le K_0. \quad (3.50)$$

For $k = K_0$, the last integral is given by $\left(-(\lambda r_{K_0}^{-1}\mathbf{1}\boldsymbol{\tau}+\boldsymbol{T})\right)^{-1}$.

3.6 Moments of the workload

In what follows, we show a computational procedure for the n-th moment of the workload, where n is a positive integer. To this end, we introduce some new notations. Let $\mathscr{M}$ be the set of all square matrices. For $\boldsymbol{X}\in\mathscr{M}$, let $\sigma(\boldsymbol{X})$ denote the order of $\boldsymbol{X}$. For a positive integer n, let $\boldsymbol{O}_n$ and $\boldsymbol{I}_n$ denote the $n\times n$ zero and identity matrices, respectively. Let $\Psi:\mathscr{M}\to\mathscr{M}$ be a mapping defined as follows:

$$\Psi(\boldsymbol{X}) := \begin{pmatrix}\boldsymbol{O}_{\sigma(\boldsymbol{X})} & \boldsymbol{I}_{\sigma(\boldsymbol{X})}\\ \boldsymbol{O}_{\sigma(\boldsymbol{X})} & \boldsymbol{X}\end{pmatrix}, \quad \boldsymbol{X}\in\mathscr{M}. \quad (3.51)$$

From Corollary 2.1, we have for $0\le x\le y$,

$$\int_x^y \exp(\boldsymbol{X}u)\mathrm{d}u = [\exp(\Psi(\boldsymbol{X})y) - \exp(\Psi(\boldsymbol{X})x)]^{(\sigma(\boldsymbol{X}),\sigma(\boldsymbol{X}))}. \quad (3.52)$$

Lemma 3.5 *For a positive integer n, $\boldsymbol{X} \in \mathscr{M}$ and $d \geq 0$, we have*

$$\int_0^d u^n \exp(\boldsymbol{X}u)\mathrm{d}u = \sum_{i=0}^{n}(-1)^i \frac{n!}{(n-i)!} d^{n-i}[\exp(\Psi^{i+1}(\boldsymbol{X})d)]^{(\sigma(\boldsymbol{X}),\sigma(\boldsymbol{X}))}, \tag{3.53}$$

where Ψ^{i+1} is the $(i+1)$-th iterate of Ψ.

Proof. Noting that $u^n = \int_0^u n y^{n-1}\mathrm{d}y$, we have

$$\begin{aligned}\int_0^d u^n \exp(\boldsymbol{X}u)\mathrm{d}u &= \int_0^d \left\{\int_y^d \exp(\boldsymbol{X}u)\mathrm{d}u\right\} n y^{n-1}\mathrm{d}y \\ &= \int_0^d n y^{n-1}\mathrm{d}y \times [\exp(\Psi(\boldsymbol{X})d)]^{(\sigma(\boldsymbol{X}),\sigma(\boldsymbol{X}))} \\ &\quad - n\left[\int_0^d y^{n-1}\exp(\Psi(\boldsymbol{X})y)\mathrm{d}y\right]^{(\sigma(\boldsymbol{X}),\sigma(\boldsymbol{X}))},\end{aligned} \tag{3.54}$$

where the last equality follows from (3.52). We then obtain the following recurrence formula:

$$\int_0^d u^n \exp(\boldsymbol{X}u)\mathrm{d}u = d^n[\exp(\Psi(\boldsymbol{X})d)]^{(\sigma(\boldsymbol{X}),\sigma(\boldsymbol{X}))} - n\left[\int_0^d u^{n-1}\exp(\Psi(\boldsymbol{X})u)\mathrm{d}u\right]^{(\sigma(\boldsymbol{X}),\sigma(\boldsymbol{X}))}, \tag{3.55}$$

which implies (3.53). □

From Lemmas 3.1–3.3, we have

$$\mathsf{E}[Z^n] = \frac{\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\tilde{\boldsymbol{D}}_K(n)\boldsymbol{t} + \sum_{1\leq k\leq K}\frac{\lambda r_k^{-1}\boldsymbol{\tau}}{1+m(d_K)+\lambda m_1}\tilde{\boldsymbol{W}}_k(n)\mathbf{1} + \sum_{1\leq k\leq K_0}\boldsymbol{c}_k\tilde{\boldsymbol{U}}_k(n)\mathbf{1}, \tag{3.56}$$

where

$$\tilde{\boldsymbol{D}}_K(n) := \int_0^{d_K} x^n \exp(\boldsymbol{D}x)\mathrm{d}x, \tag{3.57}$$

$$\tilde{\boldsymbol{W}}_k(n) := \int_{x\in J_k} x^n \boldsymbol{W}_k(x)\mathrm{d}x, \quad 1\leq k\leq K, \tag{3.58}$$

$$\tilde{\boldsymbol{U}}_k(n) := \int_{x\in J_k} x^n \boldsymbol{U}_k(x)\mathrm{d}x, \quad 1\leq k\leq K_0. \tag{3.59}$$

By applying Lemma 3.5 to (3.56), a computable formula for the n-th moment of the workload is given as follows.

Theorem 3.2 *For a positive integer n, the n-th moment of the workload is given by (3.56), where*

$$\tilde{\boldsymbol{D}}_K(n) = \sum_{i=0}^{n}(-1)^i\frac{n!}{(n-i)!}(d_K)^{n-i}[\exp(\Psi^{i+1}(\boldsymbol{D})d_K)]^{(N,N)}, \tag{3.60}$$

$$\begin{aligned}\tilde{\boldsymbol{W}}_k(n) = (\boldsymbol{I}_N \otimes \boldsymbol{\tau}_k)\sum_{j=0}^{n}\binom{n}{j}(d_{k-1})^{n-j}\sum_{i=0}^{j}(-1)^i\frac{j!}{(j-i)!}(d_k-d_{k-1})^{j-i}\\ \times[\exp(\Psi^{i+1}(\overline{\boldsymbol{M}}_k)(d_k-d_{k-1}))]^{(N^2,N)}, \quad 1\le k\le K,\end{aligned} \tag{3.61}$$

$$\begin{aligned}\tilde{\boldsymbol{U}}_k(n) = \sum_{j=0}^{n}\binom{n}{j}(d_{k-1})^{n-j}\sum_{i=0}^{j}(-1)^i\frac{j!}{(j-i)!}(d_k-d_{k-1})^{j-i}\\ \times[\exp(\Psi^{i+1}(\lambda r_k^{-1}\mathbf{1}\boldsymbol{\tau}+\boldsymbol{T})(d_k-d_{k-1}))]^{(N,N)}, \quad 1\le k\le K_0-1,\end{aligned} \tag{3.62}$$

$$\tilde{\boldsymbol{U}}_{K_0}(n) = \sum_{i=0}^{n}\frac{n!}{(n-i)!}(d_{K_0-1})^{n-i}(-\lambda r_{K_0}^{-1}\mathbf{1}\boldsymbol{\tau}-\boldsymbol{T})^{-1-i}. \tag{3.63}$$

Proof. Equation (3.60) is obtained by applying Lemma 3.5 to (3.57). Since $(-\lambda r_{K_0}^{-1}\mathbf{1}\boldsymbol{\tau}-\boldsymbol{T})^{-1}$ exists, Equation (3.63) is obtained by partially integrating Equation (3.59) with $k := K_0$. Because the computation for Equation (3.62) is similar to that of Equation (3.61), we only show the derivation of the latter one. From Equations (3.58) and (3.27), we have

$$\begin{aligned}\tilde{\boldsymbol{W}}_k(n) &= (\boldsymbol{I}_N \otimes \boldsymbol{\tau}_k)\int_0^{d_k-d_{k-1}}(x+d_{k-1})^n[\exp(\overline{\boldsymbol{M}}_k x)]^{(N^2,N)}\mathrm{d}x\\ &= (\boldsymbol{I}_N \otimes \boldsymbol{\tau}_k)\sum_{j=0}^{n}\binom{n}{j}(d_{k-1})^{n-j}\left[\int_0^{d_k-d_{k-1}}x^j\exp(\overline{\boldsymbol{M}}_k x)\mathrm{d}x\right]^{(N^2,N)},\end{aligned} \tag{3.64}$$

where the last equality follows from the binomial theorem. We then obtain Equation (3.61) by applying Lemma 3.5 to Equation (3.64). □

4 Example: The active period of an $M/M/1$ queue with two processing speeds and vacations

In previous sections, we have obtained a computable form for the mean active period. In this section, we further study the Laplace-Stieltjes transform (LST) of the active period for the special case with exponential service demand and multiple thresholds. From the LST, we derive the mean active period for an $M/M/1$ queue with two processing speeds and vacations. We also simplify the mean active period obtained in Theorem 2.1 and confirm that it is identical to the one obtained via the LST.

4.1 LST of the active period for M/M/1 with multiple service speeds and vacations

In this subsection, we study the LST of the active period, denoted by A, of the queueing model in Section 3, where the service requirements are assumed to be exponentially distributed with mean μ^{-1}. For $s \geq 0$, let $\varphi(s) := \mathsf{E}[e^{-sA}]$ denote the LST of the active period, and $\varphi(s,x) := \mathsf{E}[e^{-sA}|\text{initial workload is } x]$ $(x > 0)$ denote the LST of the active period when it starts with a workload at level x.

For $x \in J_k$, where $1 \leq k \leq K_0$, and for $\Delta \downarrow 0$, distinguishing the two possibilities of having no arrival or one arrival in the next Δ gives

$$\varphi(s, x + r_k\Delta) = (1 - \lambda\Delta)e^{-s\Delta}\varphi(s,x) + \lambda\Delta e^{-s\Delta}\int_{y=0}^{\infty} \mu e^{-\mu y}\varphi(s, x+y)\mathrm{d}y + o(\Delta),$$

which readily leads to the integro-differential equation

$$\varphi_x(s,x) = -\frac{\lambda + s}{r_k}\varphi(s,x) + \frac{\lambda}{r_k}e^{\mu x}\int_{z=x}^{\infty}\mu e^{-\mu z}\varphi(s,z)\mathrm{d}z, \quad x \in J_k,\ 1 \leq k \leq K_0. \tag{4.1}$$

The solution of this differential equation is given by the following lemma.

Lemma 4.1 *For $x \in J_k$ and $1 \leq k \leq K_0$, the LST of the active period starting at level x is given by*

$$\varphi(s,x) = A_k(s)e^{\alpha_k(s)x} + B_k(s)e^{\beta_k(s)x}, \tag{4.2}$$

where

$$\alpha_k(s) = \frac{1}{2}\left\{-\left(\frac{\lambda+s}{r_k} - \mu\right) + \sqrt{\left(\frac{\lambda+s}{r_k} - \mu\right)^2 + 4\frac{\mu s}{r_k}}\right\} > 0, \tag{4.3}$$

$$\beta_k(s) = \frac{1}{2}\left\{-\left(\frac{\lambda+s}{r_k} - \mu\right) - \sqrt{\left(\frac{\lambda+s}{r_k} - \mu\right)^2 + 4\frac{\mu s}{r_k}}\right\} < 0, \tag{4.4}$$

and $\{A_k(s), B_k(s); 1 \leq k \leq K_0\}$ are constants which are determined by the following $2K_0$ equations:

$$A_{K_0}(s) = 0, \quad A_1(s) + B_1(s) = 1, \tag{4.5}$$

and for $1 \leq k \leq K_0 - 1$,

$$A_k(s)e^{\alpha_k(s)d_k} + B_k(s)e^{\beta_k(s)d_k} = A_{k+1}(s)e^{\alpha_{k+1}(s)d_k} + B_{k+1}(s)e^{\beta_{k+1}(s)d_k}, \tag{4.6}$$

and

$$\begin{aligned} r_k\left\{A_k(s)\alpha_k(s)e^{\alpha_k(s)d_k} + B_k(s)\beta_k(s)e^{\beta_k(s)d_k}\right\} = \\ r_{k+1}\left\{A_{k+1}(s)\alpha_{k+1}(s)e^{\alpha_{k+1}(s)d_k} + B_{k+1}(s)\beta_{k+1}(s)e^{\beta_{k+1}(s)d_k}\right\}. \end{aligned} \tag{4.7}$$

Proof. By differentiating the terms of equation (4.1) w.r.t. x and eliminating the integral, we obtain the following second-order differential equation:

$$\varphi_{xx}(s,x) + \left(\frac{\lambda+s}{r_k} - \mu\right)\varphi_x(s,x) - \frac{\mu s}{r_k}\varphi(s,x) = 0, \quad x \in J_k,\ 1 \le k \le K_0. \tag{4.8}$$

The general solution of (4.8) is given by (4.2). We next find $2K_0$ equations to determine $\{A_k(s), B_k(s); 1 \le k \le K_0\}$. From the definition of the LST, we have

$$\lim_{x\to\infty} \varphi(s,x) = 0, \quad \lim_{x\to 0} \varphi(s,x) = 1, \tag{4.9}$$

and

$$\varphi(s,d_k-) = \varphi(s,d_k+), \quad 1 \le k \le K_0 - 1. \tag{4.10}$$

Equations (4.2) and (4.9) imply (4.5), and equations (4.2) and (4.10) imply (4.6). For $1 \le k \le K_0 - 1$, by taking $x \uparrow d_k$ and $x \downarrow d_k$ of (4.1), respectively, and noting (4.10), we have

$$r_k\varphi_x(s,d_k-) = r_{k+1}\varphi_x(s,d_k+), \quad 1 \le k \le K_0 - 1. \tag{4.11}$$

By differentiating (4.2) w.r.t. x, and substituting the result into (4.11), we have (4.7). □

We next calculate $\{A_k(s), B_k(s); 1 \le k \le K_0\}$ by (4.5), (4.6), and (4.7). For ease of notation, for $1 \le k \le K_0 - 1$ and $i = 0, 1$, let

$$a_{k+i,k} := e^{\alpha_{k+i}(s)d_k}, \quad b_{k+i,k} := e^{\beta_{k+i}(s)d_k}, \tag{4.12}$$

$$a'_{k+i,k} := \alpha_{k+i}(s)e^{\alpha_{k+i}(s)d_k}, \quad b'_{k+i,k} := \beta_{k+i}(s)e^{\beta_{k+i}(s)d_k}. \tag{4.13}$$

For $1 \le k \le K_0 - 1$, by eliminating $B_k(s)$ from equations (4.6) and (4.7), we have

$$A_k(s) = c_{00}^{(k)}A_{k+1}(s) + c_{01}^{(k)}B_{k+1}(s), \tag{4.14}$$

where

$$c_{00}^{(k)} := \frac{a'_{k+1,k}b_{k,k}r_{k+1} - a_{k+1,k}b'_{k,k}r_k}{(a'_{k,k}b_{k,k} - a_{k,k}b'_{k,k})r_k}, \quad c_{01}^{(k)} := \frac{b_{k,k}b'_{k+1,k}r_{k+1} - b_{k+1,k}b'_{k,k}r_k}{(a'_{k,k}b_{k,k} - a_{k,k}b'_{k,k})r_k}. \tag{4.15}$$

Similarly, we have for $1 \le k \le K_0 - 1$,

$$B_k(s) = c_{10}^{(k)}A_{k+1}(s) + c_{11}^{(k)}B_{k+1}(s), \tag{4.16}$$

where

$$c_{10}^{(k)} := \frac{a_{k,k}a'_{k+1,k}r_{k+1} - a_{k+1,k}a'_{k,k}r_k}{(a_{k,k}b'_{k,k} - a'_{k,k}b_{k,k})r_k}, \quad c_{11}^{(k)} := \frac{b'_{k+1,k}a_{k,k}r_{k+1} - b_{k+1,k}a'_{k,k}r_k}{(a_{k,k}b'_{k,k} - a'_{k,k}b_{k,k})r_k}. \tag{4.17}$$

From (4.14) and (4.16), we have for $1 \le k \le K_0 - 1$,

$$\begin{pmatrix} A_k(s) \\ B_k(s) \end{pmatrix} = \boldsymbol{C}_k \begin{pmatrix} A_{k+1}(s) \\ B_{k+1}(s) \end{pmatrix} = \overline{\boldsymbol{C}}_k \begin{pmatrix} A_{K_0}(s) \\ B_{K_0}(s) \end{pmatrix} = \overline{\boldsymbol{C}}_k \begin{pmatrix} 0 \\ B_{K_0}(s) \end{pmatrix}, \tag{4.18}$$

where

$$\boldsymbol{C}_k := \begin{pmatrix} c_{00}^{(k)} & c_{01}^{(k)} \\ c_{10}^{(k)} & c_{11}^{(k)} \end{pmatrix}, \qquad \overline{\boldsymbol{C}}_k := \begin{pmatrix} \overline{c}_{00}^{(k)} & \overline{c}_{01}^{(k)} \\ \overline{c}_{10}^{(k)} & \overline{c}_{11}^{(k)} \end{pmatrix} = \boldsymbol{C}_k \boldsymbol{C}_{k+1} \cdots \boldsymbol{C}_{K_0-1}, \tag{4.19}$$

and the last equation in (4.18) follows from (4.5). From (4.5) and (4.18), $\{A_k(s), B_k(s); 1 \le k \le K_0\}$ are given as follows.

Lemma 4.2 *For* $1 \le k \le K_0 - 1$,

$$A_k(s) = \frac{\overline{c}_{01}^{(k)}}{\overline{c}_{01}^{(1)} + \overline{c}_{11}^{(1)}}, \quad B_k(s) = \frac{\overline{c}_{11}^{(k)}}{\overline{c}_{01}^{(1)} + \overline{c}_{11}^{(1)}}, \tag{4.20}$$

and

$$A_{K_0}(s) = 0, \quad B_{K_0}(s) = \frac{1}{\overline{c}_{01}^{(1)} + \overline{c}_{11}^{(1)}}. \tag{4.21}$$

4.2 Mean active period in case of two service speeds, Approach 1: LST

In this section, we apply Lemma 4.1 to the case of two service speeds, i.e., $K_0 := 2$, and derive the mean active period for the $M/M/1$ queue with two processing speeds and vacations. To this end, we introduce some new notations. For $x \in \mathbb{R}$, let $\langle x \rangle^{-1} = \min(0, x)$ and $\langle x \rangle^{+1} = \max(0, x)$. We denote the traffic intensities when the workload is in $[0, d_1)$ and $[d_1, \infty)$ by $\rho_1 := \lambda/(r_1\mu)$ and $\rho_2 := \lambda/(r_2\mu)$, respectively. The stability condition (see also (2.4)) is rewritten by

$$\rho_2 < 1. \tag{4.22}$$

By conditioning on the initial fluid level when the active period starts, and from the memoryless property of the exponential distribution, the LST of the active period is given as follows: for $s \ge 0$, we have

$$\begin{aligned} \varphi(s) &= \int_0^\infty \varphi(s, d_1 + x) \mu e^{-\mu x} \mathrm{d}x \\ &= \int_0^\infty B_2(s) e^{\beta_2(s)(d_1+x)} \mu e^{-\mu x} \mathrm{d}x \\ &= \frac{\mu e^{\beta_2(s) d_1}}{\mu - \beta_2(s)} B_2(s), \end{aligned} \tag{4.23}$$

where the second equation follows from Lemma 4.1, and the coefficients $\beta_2(s)$ and $B_2(s)$ are given as follows (see also (4.4) and (4.21)):

$$\beta_2(s) = \frac{1}{2} \left\{ -\left(\frac{\lambda + s}{r_2} - \mu \right) - \sqrt{\left(\frac{\lambda + s}{r_2} - \mu \right)^2 + 4 \frac{\mu s}{r_2}} \right\}, \tag{4.24}$$

$$B_2(s) = \frac{1}{c_{01}^{(1)} + c_{11}^{(1)}}. \tag{4.25}$$

Note that from Equations (4.23) and (4.24), we have

$$\beta_2(0) = 0, \quad \varphi(0) = B_2(0) = 1. \tag{4.26}$$

From (4.23) and (4.26), the mean active period m_1 is calculated as follows.

$$m_1 = -\varphi'(0) = -B_2'(0) - \frac{1+\mu d_1}{\mu}\beta_2'(0), \tag{4.27}$$

where

$$\beta_2'(0) = \frac{-1}{r_2(1-\rho_2)}. \tag{4.28}$$

It remains to show the computation of $B_2'(0)$. From (4.25) and (4.26), we have

$$B_2(s)^{-1}\xi(s) = \zeta(s), \tag{4.29}$$

where

$$\xi(s) := r_1(\alpha_1(s) - \beta_1(s))e^{(\alpha_1(s)+\beta_1(s))d_1}, \tag{4.30}$$

$$\zeta(s) := r_2(e^{\beta_1(s)d_1} - e^{\alpha_1(s)d_1})\beta_2(s)e^{\beta_2(s)d_1} - r_1(\beta_1(s)e^{\beta_1(s)d_1} - \alpha_1(s)e^{\alpha_1(s)d_1})e^{\beta_2(s)d_1}. \tag{4.31}$$

By differentiating (4.29) w.r.t. s and then taking $s \to 0$, we obtain

$$B_2'(0)\xi(0) = \xi'(0) - \zeta'(0), \tag{4.32}$$

where

$$\xi(0) = \mu r_1 e^{(1-\rho_1)\mu d_1}|1-\rho_1|, \tag{4.33}$$

$$\xi'(0) = \left(-\mu d_1|1-\rho_1| + \frac{1+\rho_1}{|1-\rho_1|}\right)e^{(1-\rho_1)\mu d_1}. \tag{4.34}$$

From (4.26) and (4.28), and noting that

$$\alpha_1(0) = \mu\langle 1-\rho_1\rangle^+, \quad \alpha_1'(0) = \frac{1+\rho_1-|1-\rho_1|}{2r_1|1-\rho_1|}, \tag{4.35}$$

$$\beta_1(0) = \mu\langle 1-\rho_1\rangle^-, \quad \beta_1'(0) = -\frac{1+\rho_1+|1-\rho_1|}{2r_1|1-\rho_1|}, \tag{4.36}$$

then we obtain the derivative $\zeta'(0)$ in (4.32) as follows:

$$\zeta'(0) = \sum_{k=\pm 1} e^{\langle 1-\rho_1\rangle^k \mu d_1}\left\{\frac{k}{r_2(1-\rho_2)}(r_2 - \langle 1-\rho_1\rangle^k \mu r_1 d_1) + \frac{1+\rho_1-k|1-\rho_1|}{2|1-\rho_1|}(1+\langle 1-\rho_1\rangle^k \mu d_1)\right\}. \tag{4.37}$$

From Equations (4.32) – (4.34) and (4.37), we have for $\rho_1 \neq 1$,

$$B_2'(0) = \frac{1}{1-\rho_1}\left\{-\frac{d_1}{r_1} + \frac{1-e^{-(1-\rho_1)\mu d_1}}{(1-\rho_1)\mu r_1}\right\} + \frac{1}{1-\rho_2}\left\{\frac{d_1}{r_2} - \frac{1-e^{-(1-\rho_1)\mu d_1}}{(1-\rho_1)\mu r_1}\right\}. \tag{4.38}$$

From (4.27), (4.28), and (4.38), the mean active period is given as follows.

Corollary 4.1 *For $\rho_1 \neq 1$, we have*

$$m_1 = \frac{\rho_1}{\lambda}\frac{(1-\rho_1)\mu d_1 - (1-e^{-(1-\rho_1)\mu d_1})\rho_1}{(1-\rho_1)^2} + \frac{1-\rho_1 e^{-(1-\rho_1)\mu d_1}}{\lambda(1-\rho_1)}\frac{\rho_2}{1-\rho_2}. \tag{4.39}$$

On the other hand, for $\rho_1 = 1$, we have

$$m_1 = \frac{\mu d_1(2+\mu d_1)}{2\lambda} + \frac{1+\mu d_1}{\lambda}\frac{\rho_2}{1-\rho_2}. \tag{4.40}$$

Proof. By substituting (4.28) and (4.38) into (4.27), we have for $\rho_1 \neq 1$,

$$\begin{aligned} m_1 &= \frac{1}{1-\rho_1}\left\{\frac{d_1}{r_1} - \frac{1-e^{-(1-\rho_1)\mu d_1}}{(1-\rho_1)\mu r_1}\right\} + \frac{1}{1-\rho_2}\left\{-\frac{d_1}{r_2} + \frac{1-e^{-(1-\rho_1)\mu d_1}}{(1-\rho_1)\mu r_1}\right\} \\ &\quad + \frac{1+\mu d_1}{\mu r_2}\frac{1}{1-\rho_2} \\ &= \frac{1}{\lambda(1-\rho_1)}\left\{\frac{\rho_1(1-\rho_1)\mu d_1 - \rho_1(1-e^{-(1-\rho_1)\mu d_1})}{1-\rho_1}\right. \\ &\quad \left. + \frac{(1-\rho_1 e^{-(1-\rho_1)\mu d_1})\rho_2 + (1-e^{-(1-\rho_1)\mu d_1})\rho_1(1-\rho_2)}{1-\rho_2}\right\}, \end{aligned} \tag{4.41}$$

which implies (4.40). Since the mean active period is a monotonic function of ρ_1, we obtain (4.40) by letting $\rho_1 \to 1$ in (4.39). □

4.3 Mean active period in case of two service speeds, Approach 2: Application of Theorem 2.1

We apply Theorem 2.1 to the $M/M/1$ queue with two processing speeds and vacations. Since the service time is assumed to be exponentially distributed with mean μ^{-1}, we put

$$N := 1, \quad \boldsymbol{\tau} := 1, \quad \boldsymbol{T} := -\mu, \quad \boldsymbol{t} := \mu, \quad \boldsymbol{D} := 0, \tag{4.42}$$

and apply

$$\boldsymbol{I}_N \otimes \boldsymbol{\tau} := 1, \quad \boldsymbol{I}_{N^2+N} := \boldsymbol{I}_2, \quad \lambda r_1^{-1}\mathbf{1}\boldsymbol{\tau} + \boldsymbol{T} := -(1-\rho_1)\mu, \quad \overline{\boldsymbol{M}} := \begin{pmatrix} 0 & \mu \\ 0 & -(1-\rho_1)\mu \end{pmatrix} \tag{4.43}$$

to Theorem 2.1. We then obtain the same formulas given in Corollary 4.1 for the mean active period.

Corollary 4.2 *For $\rho_1 \neq 1$ and $\rho_1 = 1$, the mean active period is given by (4.39) and (4.40), respectively.*

Proof. It is sufficient to consider the first case, i.e., we consider the case of $\rho_1 \neq 1$. Under (4.42) and (4.43), we have

$$\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_{N^2+N} \\ \boldsymbol{O} & \overline{\boldsymbol{M}} \end{pmatrix} d_1\right) = \begin{pmatrix} \boldsymbol{I}_2 & \sum_{k=1}^{\infty} \frac{\overline{\boldsymbol{M}}^{k-1} d_1^k}{k!} \\ \boldsymbol{O} & \boldsymbol{I}_2 + \sum_{k=1}^{\infty} \frac{\overline{\boldsymbol{M}}^{k} d_1^k}{k!} \end{pmatrix}, \tag{4.44}$$

which implies that

$$\begin{aligned} \left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_{N^2+N} \\ \boldsymbol{O} & \overline{\boldsymbol{M}} \end{pmatrix} d_1\right)\right]^{(N^2,N)} &= \left[\sum_{k=1}^{\infty} \frac{\overline{\boldsymbol{M}}^{k-1} d_1^k}{k!}\right]^{(1,1)} \\ &= \sum_{k=2}^{\infty} \frac{\mu(-(1-\rho_1)\mu)^{k-2} d_1^k}{k!}, \end{aligned} \tag{4.45}$$

where the last equation follows from

$$\sum_{k=1}^{\infty} \frac{\overline{\boldsymbol{M}}^{k-1} d_1^k}{k!} = \begin{pmatrix} d_1 & \sum_{k=2}^{\infty} \frac{\mu(-(1-\rho_1)\mu)^{k-2} d_1^k}{k!} \\ 0 & d_1 + \sum_{k=2}^{\infty} \frac{(-(1-\rho_1)\mu)^{k-1} d_1^k}{k!} \end{pmatrix}. \tag{4.46}$$

Since $\rho_1 \neq 1$, we have

$$\left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_{N^2+N} \\ \boldsymbol{O} & \overline{\boldsymbol{M}} \end{pmatrix} d_1\right)\right]^{(N^2,N)} = \frac{(1-\rho_1)\mu d_1 - (1 - e^{-(1-\rho_1)\mu d_1})}{\mu(1-\rho_1)^2}. \tag{4.47}$$

Similar to (4.47), we have

$$\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_N \\ \boldsymbol{O} & \lambda r_1^{-1} \mathbf{1}\boldsymbol{\tau} + \boldsymbol{T} \end{pmatrix} d_1\right) = \begin{pmatrix} 1 & \sum_{k=1}^{\infty} \frac{(-(1-\rho_1)\mu)^{k-1} d_1^k}{k!} \\ 0 & 1 + \sum_{k=1}^{\infty} \frac{(-(1-\rho_1)\mu)^{k} d_1^k}{k!} \end{pmatrix}, \tag{4.48}$$

which implies that since $\rho_1 \neq 1$,

$$\left[\exp\left(\begin{pmatrix} \boldsymbol{O} & \boldsymbol{I}_N \\ \boldsymbol{O} & \lambda r_1^{-1} \mathbf{1}\boldsymbol{\tau} + \boldsymbol{T} \end{pmatrix} d_1\right)\right]^{(N,N)} = \frac{1 - e^{-(1-\rho_1)\mu d_1}}{\mu(1-\rho_1)}. \tag{4.49}$$

Furthermore, we have

$$\exp(\overline{\boldsymbol{M}} d_1) = \begin{pmatrix} 1 & \sum_{k=1}^{\infty} \frac{\mu(-(1-\rho_1)\mu)^{k-1} d_1^k}{k!} \\ 0 & 1 + \sum_{k=1}^{\infty} \frac{(-(1-\rho_1)\mu)^{k} d_1^k}{k!} \end{pmatrix}, \tag{4.50}$$

which implies that since $\rho_1 \neq 1$,

$$[\exp(\overline{\boldsymbol{M}} d_1)]^{(N^2,N)} = \frac{1 - e^{-(1-\rho_1)\mu d_1}}{1-\rho_1}. \tag{4.51}$$

Under the stability condition (4.22), we have

$$(-(\lambda r_2^{-1} \mathbf{1}\boldsymbol{\tau} + \boldsymbol{T}))^{-1} = \frac{1}{(1-\rho_2)\mu}. \tag{4.52}$$

By substituting (4.47), (4.49), (4.51), and (4.52) into (2.65) of Theorem 2.1, we obtain (4.39). □

5 Optimization problem

In this section we formulate a cost minimization problem. That problem is numerically studied in the next section.

We consider the following three types of costs per unit of time:

- $c_{p,k}$ $(1 \leq k \leq K_0)$: A cost related to the power consumption when the server is in mode 1 and processes the workload with rate r_k.
- $c_{p,0}$: A cost related to the power consumption when the server is in mode 0.
- c_s : A cost when the server switches from mode 0 to 1.

For $1 \leq k \leq K_0$, let $m_{1,k}$ denote the mean active period when the workload is in J_k. The mean power consumption cost (energy consumption cost per time unit) is given by

$$\sum_{k=1}^{K_0} c_{p,k} \frac{m_{1,k}}{m_0 + m_1} + c_{p,0} \frac{m_0}{m_0 + m_1} + c_s \frac{1}{m_0 + m_1}, \tag{5.1}$$

where $m_0 = \lambda^{-1}(1 + m(d_K))$ (see Lemma 3.1) and m_1 is given by (3.48). From Theorem 3.1, $m_{1,k}$ $(1 \leq k \leq K_0)$ is given as follows.

$$m_{1,k} = \boldsymbol{\tau} \left\{ \left(r_k^{-1} \int_{d_{k-1}}^{d_k} \boldsymbol{W}_k(x) \mathrm{d}x \right) \mathbb{1}(k \leq K) + r_k^{-1} \boldsymbol{C}_k \int_{d_{k-1}}^{d_k} \boldsymbol{U}_k(x) \mathrm{d}x \right\} \mathbf{1}. \tag{5.2}$$

We note that each of m_0, m_1, and $m_{1,k}$ $(1 \leq k \leq K_0)$ is given by a matrix exponential form (see Lemma 3.1 and Theorem 3.1), therefore it is easy to implement (5.1) in a numerical calculation.

In the cost function above, the cost $c_{p,k}$ will be set to $c_{p,k} = c_p r_k^2$. There is evidence that power consumption is a convex function of the processing speed and it is reasonable to set it as a quadratic function of the speed [23].

Also taking into account the performance, we consider the following cost function for our system.

$$Cost = c_h \mathrm{E}[Z] + c_p \left(\sum_{k=1}^{K_0} r_k^2 \frac{m_{1,k}}{m_0 + m_1} \right) + c_{p,0} \frac{m_0}{m_0 + m_1} + c_s \frac{1}{m_0 + m_1}. \tag{5.3}$$

In equation (5.3), the first term is related to the performance, i.e., the smaller the mean workload, the smaller the response time for jobs is. The second term (the summation) is related to the power consumption of the server in the active period. The smaller the processing speed r_k, the smaller the power consumption is; but it leads to a bigger $\mathrm{E}[Z]$. The third term is the holding cost (power consumption when the server is inactive). It should be noted that it is reasonable to set $c_{p,0} = 0.6 c_p$ as the server in inactive mode can consume about 60% power, compared to when it is busy processing a job. The last term is related to the switching cost.

It should be noted that a CPU instantaneously consumes a large amount of energy once it is switched on.

We will consider the optimization problem for minimizing the cost function. In particular, we will find the threshold d_K which minimizes the cost function. Furthermore, we also investigate the service curve $r(x)$ which minimizes the cost function.

6 Numerical results

In this section, we consider the effect of the service rate function and the threshold on the cost function. To this end, we fix the arrival rate and the job size distribution. In particular, $\lambda = 1$ and the job sizes follow a two-stage Erlang distribution with mean 2. The coefficients in the cost function are set as follows: $c_h = 0.1, c_s = 30, c_p = 1, c_{p,0} = 0.6 \times c_p$. We consider the service rate function in the form $r(x) = r_1 x^\alpha + r_0$, where we restrict the parameters as follows: $r_0 = 1$ and $r_1 = 0.1, 1, 10, 100$. The service rate function is approximated by the step function with step size 0.1 and $d_{K_0-1} = 20$ and for $x \geq 20$ the service rate is approximated by $r(20)$. We first consider the cost function against d_K for some special service rate functions: $r(x) = r_1 x + r_0$, $r(x) = r_1\sqrt{x} + r_0$ and $r(x) = r_1 x^2 + r_0$, where we fix $r_0 = 1$ and consider $r_1 = 0.1, 1, 10, 100$.

Figure 3 shows the cost function against the threshold d_K for $r(x) = r_1\sqrt{x} + r_0$. In the case $r_1 = 0.1$, the stability condition is violated. We observe that the curves for $r_1 = 1$ and 10 are convex, implying the existence of a threshold that minimizes the cost function. In the third curve with $r_1 = 100$, the cost function monotonically increases implying that $d_K = 0$ is the optimal threshold. These results suggest that when the service rate is large enough, it is optimal to switch the server on as soon as a job is available.

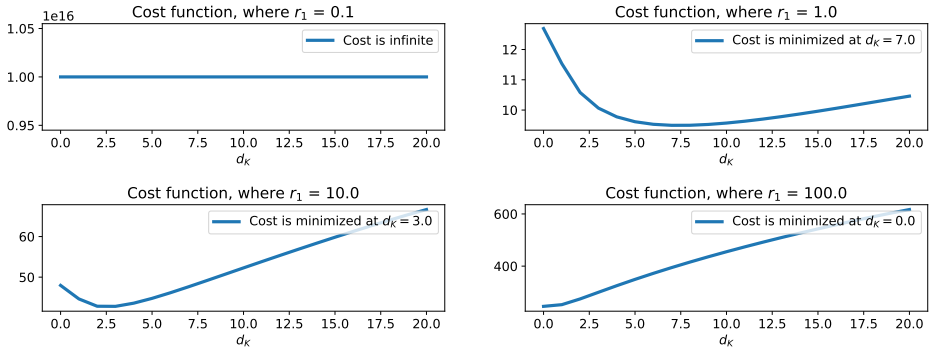

Figure 3: Cost function against threshold ($r(x) = r_1\sqrt{x} + r_0$).

Figure 4 shows the cost function against the threshold d_K for $r(x) = r_1 x + r_0$. This figure shows that the cost function is minimized at positive values of d_K for $r_1 = 0.1, 1$ and 10 while it is minimized at $d_K = 0$ for $r_1 = 100$.

Figure 5 shows the cost function against the threshold d_K for $r(x) = r_1 x^2 + r_0$. We observe that the cost function is minimized at a positive value of d_K for $r_1 = 0.1$ and 1 while it is

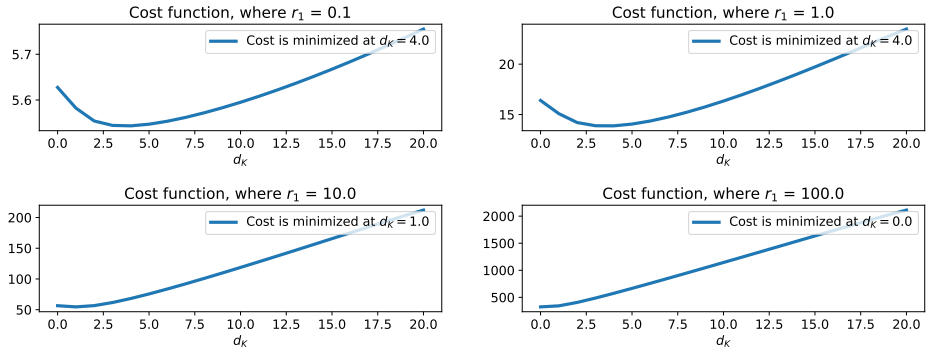

Figure 4: Cost function against threshold ($r(x) = r_1 x + r_0$).

minimized at $d_K = 0$ for $r_1 = 10$ and 100.

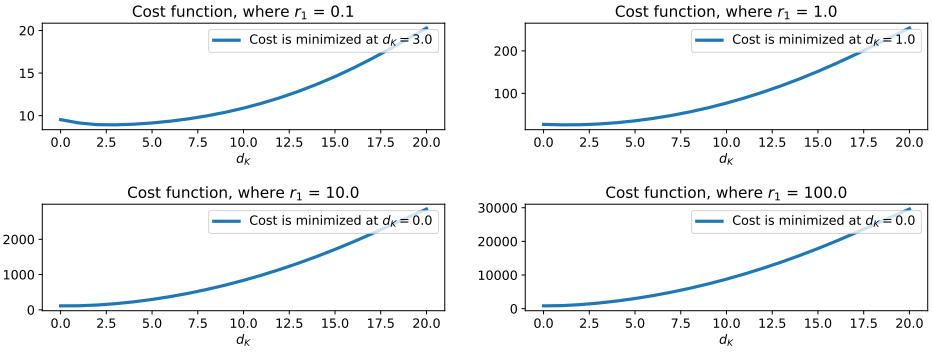

Figure 5: Cost function against threshold ($r(x) = r_1 x^2 + r_0$).

Observing all the graphs above, there exists a threshold d_K which minimizes the cost function. Some time, the cost function is minimized at $d_K = 0$ and in some other cases, it is minimized at a non-trivial $d_K > 0$. The common trend is that for a fast service rate (larger r_1 and/or α), the cost function is likely minimized at $d_K = 0$ whereas if the service rate is relatively small, the cost function is likely minimized at some positive value of the threshold d_K. For the case $r_1 = 100$, all the curves show that the cost function is minimized at $d_K = 0$.

This motivates us to have a closer look at the minimal cost for each fixed $\alpha \in [0.1, 3]$, where the service curve is given by $r(x) = r_1 x^\alpha + r_0$. Figures 6–8 show the optimal threshold d_K (on the right y-axis) and the corresponding cost (on the left y-axis) against α. We also find that the optimized cost function is minimized at some positive α for $r_1 = 0.1$ while it monotonically increases with α for $r_1 = 1, 10$. This implies that for relatively large r_1, the value of α should be small so that the service rate is not too large to balance the power consumption.

From Figures 6–8, we observe a general rule that in most of the cases ($\alpha \geq 0.2$), the optimal threshold d_K decreases with increasing α. This suggests that for a fast service rate, it is better to set a low threshold.

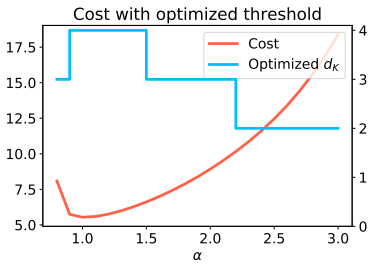

Figure 6: Case of $r_1 = 0.1$.

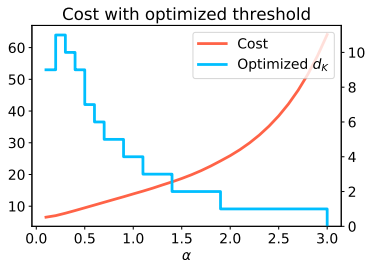

Figure 7: Case of $r_1 = 1.0$.

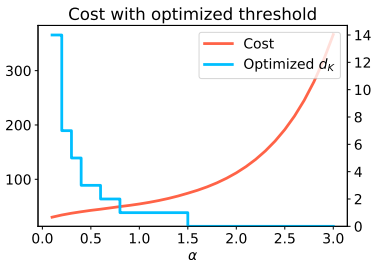

Figure 8: Case of $r_1 = 10.0$.

Acknowledgments

The research of YS and TP was supported in part by JSPS Kakenhi Grant Numbers JP18K11186 and JP18K18006, respectively. The research of OB was funded by the NWO Gravitation Program NETWORKS, Grant Number 024.002.003.

References

[1] M. Abramowitz and I.A. Stegun (1965). *Handbook of Mathematical Functions*. Dover Publications, Inc., New York.

[2] S. Asmussen (2003). *Applied Probability and Queues*, 2nd edition, Springer-Verlag, New York.

[3] S. M. Baik and and Y. M. Ko (2020). A QoS-aware workload routing and server speed scaling policy for energy-efficient data centers: a robust queueing theoretic approach. arXiv:1912.09870v1

[4] R. Bekker, S.C. Borst, O.J. Boxma and O. Kella (2004). Queues with workload-dependent arrival and service rates. *Queueing Systems* **46**, 537-556.

[5] P.H. Brill and M.J.M. Posner (1977). Level crossing in point processes applied to queues: Single server case. *Operations Research* **25(4)**, 662-674.

[6] S. Browne and K. Sigman (1992). Workload-modulated queues with application to storage processes. *Journal of Applied Probability* **29(3)**, 699-712.

[7] L. Chen and N. Li (2015). On the interaction between load balancing and speed scaling. *IEEE Journal on Selected Areas in Communications* **33(12)**, 2567-2578.

[8] J.W. Cohen (1977). On up- and downcrossings. *Journal of Applied Probability* **14(2)**, 405-410.

[9] J.W. Cohen (1982). *The Single Server Queue*. Second edition. North-Holland Publ. Cy., Amsterdam.

[10] A. da Silva Soares and G. Latouche (2009). Fluid queues with level dependent evolution. *European Journal of Operational Research* **196(3)**, 1041-1048.

[11] M. Delasay, A. Ingolfsson and B. Kolfal (2016). Modeling load and overwork effects in queueing systems with adaptive service rates. *Operations Research* **64(4)**, 867-885.

[12] T. Dzial, L. Breuer, A. da Silva Soares, G. Latouche and M.A. Remiche (2005). Fluid queues to solve jump processes. *Performance Evaluation* **62(1-4)**, 132-146.

[13] M. Elahi and C. Williamson (2018). On Saturation Effects in Coupled Speed Scaling. In: McIver A., Horvath A. (eds.) Quantitative Evaluation of Systems. QEST 2018. *Lecture Notes in Computer Science*, vol 11024. Springer, Cham.

[14] E.A. Feinberg and O. Kella (2002). Optimality of D-policies for an $M/G/1$ queue with a removable server. *Queueing Systems* **42**, 355-376.

[15] A. Gandhi, S. Doroudi, M. Harchol-Balter and A. Scheller-Wolf (2013). Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *ACM SIGMETRICS Performance Evaluation Review* **41**, 153-166.

[16] D.P. Gaver and R.G. Miller (1962). Limiting distributions for some storage problems. In: K.J. Arrow, S. Karlin and H. Scarf, eds., *Studies in Applied Probability and Management Science*, 110-126. Stanford University Press, Stanford, CA.

[17] J.M. Harrison and S.I. Resnick (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Mathematics of Operations Research* **1(4)**, 347-358.

[18] H.E. Kankaya and N. Akar (2008). Solving multi-regime feedback fluid queues. *Stochastic Models* **24(3)**, 425-450.

[19] D. Koops, O.J. Boxma and M.R.H. Mandjes (2017). Networks of $\cdot/G/\infty$ queues with shot-noise-driven arrival intensities. *Queueing Systems* **86**, 301-325.

[20] G. Latouche and V. Ramaswami (1999). *Introduction to Matrix Analytic Methods in Stochastic Modeling*. SIAM Press, Philadelphia.

[21] V. J. Maccio and D. G. Down (2018). Structural properties and exact analysis of energy-aware multiserver queueing systems with setup times. *Performance Evaluation* **121**, 48-66.

[22] M. Mandjes, D. Mitra and W. Scheinhardt (2003). Models of network access using feedback fluid queues. *Queueing Systems* **44(4)**, 365-398.

[23] A. Marin, I. Mitrani, M. Elahi and C. Williamson (2018). Control and optimization of the SRPT service policy by frequency scaling. In: McIver A., Horvath A. (eds) Quantitative Evaluation of Systems. QEST 2018. *Lecture Notes in Computer Science* **11024**. Springer, Cham.

[24] T. Phung-Duc (2017). Exact solutions for M/M/c/setup queues. *Telecommunication Systems* **64**, 309-324.

[25] T. Phung-Duc, W. Rogiest and S. Wittevrongeli (2017). Single server retrial queues with speed scaling: Analysis and performance evaluation. *Journal of Industrial and Management Optimization* **13**, 1927-1943.

[26] Y. Ren, T. Phung-Duc, J.C. Chen and Z.W. Yu (2016). Dynamic auto scaling algorithm (DASA) for 5G mobile networks. In *Proceedings of 2016 IEEE Global Communications Conference (GLOBECOM)* 1-6, IEEE.

[27] Y. Ren, T. Phung-Duc and J.C. Chen (2017). Design and Analysis of Dynamic Auto Scaling Algorithm (DASA) for virtual EPC (vEPC) in 5G Networks. arXiv:1604.05803v3.

[28] S.M. Ross (1983). *Stochastic Processes*. Wiley, New York.

[29] Y. Sakuma, O.J. Boxma and T. Phung-Duc (2019). A single-server queue with workload-dependent service speed and vacations. In: T. Phung-Duc, S. Kasahara and S. Wittevrongel (eds.) Queueing Theory and Network Applications, Proc. QTNA2019. *Lecture notes in Computer Science* **11688**. Springer, Cham, pp. 112-127.

[30] Y. Sakuma and T. Takine (2017). Multi-class $M/PH/1$ queues with deterministic impatience times. *Stochastic Models* **33**, 1-29.

[31] E. Seneta (1981). *Non-negative Matrices and Markov Chains*. Springer, New York.

[32] F.G. Tricomi (1957). *Integral Equations*. Interscience Publishers, New York; reprinted by Dover, 1985.

[33] B. Van Houdt (2012). Analysis of the adaptive MMAP [K]/PH [K]/1 queue: a multi-type queue with adaptive arrivals and general impatience. *European Journal of Operational Research* **220(3)**, 695-704.

[34] C. Van Loan (1978). Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control* **23**, 395-404.

[35] A. Wierman, L.L.H. Andrew and M. Lin (2011). Speed scaling: An algorithmic perspective. In: Handbook of Energy-Aware and Green Computing. Chapman & Hall / CRC Computing and Information Science Series.

[36] M. Yajima and T. Phung-Duc (2017). Batch arrival single-server queue with variable service speed and setup time. *Queueing Systems* **86**, 241-260.

[37] M. Yajima and T. Phung-Duc (2020). Analysis of a variable service speed single server queue with batch arrivals and general setup time, *Performance Evaluation* **138**, Article no. 102082.

[38] M.A. Yazici and N. Akar (2013). The workload-dependent MAP/PH/1 queue with infinite/finite workload capacity. *Performance Evaluation* **70(12)**, 1047-1058.