EURANDOM PREPRINT SERIES

2019-001

February 26, 2019

A single server queue with workload-dependent service speed and vacations

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

A single server queue with workload-dependent service speed and vacations

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

February 26, 2019

[1*] Department of Computer Science, National Defense Academy of Japan, Yokosuka, Japan, Corresponding author, sakuma@nda.ac.jp, supported by JSPS KAKENHI Grant No. JP16K21704.
[2] Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands; o.j.boxma@tue.nl; Research funded by the NWO Gravitation Program NETWORKS, Grant Number 024.002.003.
[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

In modern data centers, the trade-off between processing speed and energy consumption is an important issue. Motivated by this, we consider a queueing system in which the service speed is a function of the workload, and in which 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. We use this result to choose the activation threshold such that a certain cost function, involving holding costs and activation costs, is minimized.

1 Introduction

In this paper we consider an $M/G/1$-type queueing system with the following two special features: (i) the service speed is not constant, but a function of the workload, and (ii) the server switches off when the system becomes empty, only to be activated again when the workload reaches a certain threshold. In the remainder of this introduction we successively provide a motivation for this study, present a detailed model description, discuss related literature and give an overview of the rest of the paper.

1.1 Motivation

Cloud service has become ubiquitous in our modern information society. Most Internet users are familiar with some cloud service such as Dropbox, Slack, Google drive etc. These services are supported by data centers where thousands of servers are available, consuming a large amount of energy. Thus, it is crucial to have mechanisms balancing energy consumption and performance for users. While energy saving is very important, most data centers are still designed for peak traffic of users. As a result, in the off-peak period, most servers are idle but still consume about 60% of their peak energy consumption [10, 14]. One simple idea is to use an ON-OFF control that automatically adjusts the number of active servers according to the workload. In addition, dynamic scaling techniques such as frequency scaling or voltage scaling enable individual computers to adjust their processing speed in accordance with their workload.

These automatic scaling techniques have the advantage of balancing performance and energy consumption. Because the energy consumption is a monotonic function of the processing speed, less energy is consumed when the system is less congested. When the workload is high, the processing speed is scaled up and thus, the delay performance is improved. At the single computer (server, CPU) level, on the other hand, energy could be saved by adjusting the processing speed of a server according to its own workload. These considerations, featuring the important trade-off between processing speed and energy consumption, motivate the analysis and optimization of queueing systems where the server capacity is dynamically changed according to the workload.

Apart from the interest in power-saving computer systems, queues with variable service speed also naturally arise in service systems with human servers. In particular, in service systems such as call centers, staff numbers are scheduled to meet the demands of customers. Also a human server may speed up when the workload is large, and may spend more time on a job when the workload is small.

In this paper, we propose and analyze a queueing model that 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. We obtain the distribution for the stationary workload in the system and its mean. We also formulate an optimization problem.

1.2 Model description

The model under consideration is an $M/G/1$ queue with two special features (cf. Figure 1): (i) when the server is active and the amount of work present equals $x > 0$, the server works at speed $r(x)$, and (ii) when the workload has dropped to zero, the server becomes inactive ("takes a vacation") and remains inactive until the workload has reached some level $M > 0$, after which it immediately resumes service. We denote the rate of the Poisson arrival process by λ, and the i.i.d. (independent, identically distributed) service requirements by $B_1, B_2, \ldots$, with distribution $B(\cdot)$ and Laplace-Stieltjes transform (LST) $\beta(s)$. B will denote a generic service requirement. For much of the paper, we shall assume

that $B(x) = 1 - \mathrm{e}^{-\mu x}$, $x \geq 0$.

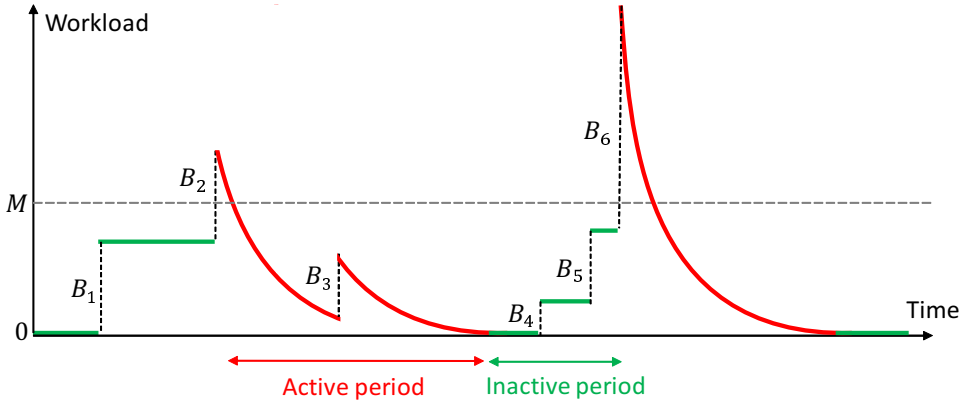

Figure 1: The workload process.

The case without vacations has been the subject of several studies (see, e.g., [3] and its references). The stability condition for that case is that (cf. [5, 6, 11]),

$$\mathrm{limsup}_{x\to\infty} \frac{\lambda \mathbb{E}(B)}{r(x)} < 1. \tag{1.1}$$

Clearly, the same condition should hold in case the server takes a vacation until workload level M is reached. From now on we assume that (1.1) holds. Below we focus on the steady-state workload distribution $V(\cdot)$ and its density $v(\cdot)$. We also need to take into account the steady-state workload distribution $V_I(\cdot)$ and its density $v_I(\cdot)$ during inactive (vacation) periods of the server; by $p_I := V_I(\infty)$ we denote the probability that the server is inactive.

Define

$$R(x,z) := \int_z^x \frac{1}{r(y)} \mathrm{d}y, \quad 0 \leq z < x < \infty; \tag{1.2}$$

$R(x,z)$ represents the time required to move from level x down to level z in the absence of any arrivals. In particular, $R(x) := R(x,0)$ denotes the time required to empty the system when starting at level x, in the absence of any arrivals. We assume in the sequel that $R(x) < \infty$ for $x < \infty$; notice that this excludes the choice $r(x) = rx$, which is sometimes termed the shotnoise case [13].

1.3 Related literature

Our model is related to several topics in the queueing literature. First of all, it is a special example of a queue with vacations: the server takes a vacation when the system becomes empty, and resumes service when the workload reaches or exceeds a certain level. In the classical $M/G/1$ setting, such a *D-policy* has been extensively studied. We refer to [8] 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 [8] prove that D-policies are optimal for the average-cost-per-time-unit criterion. This means that 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.

Secondly, our model touches upon the topic of speed scaling. We refer to [19] for an insightful discussion of speed scaling. Recent papers which consider single server queues with speed scaling where the speed of the server is adjusted according to the number of jobs in the system are, e.g., [16, 20]. Multiserver queues with ON-OFF control have been extensively studied [9, 10, 14, 15]. In the models in those papers, each server is turned off once it has no jobs to process and is turned on again when jobs are waiting.

Thirdly, there is an extensive literature on queues and dams with a level-dependent outflow rate. We mention the pioneering papers [11, 12] and refer to [3] for some more recent results and further references.

1.4 The structure of the paper

Section 2 is devoted to a study of the steady-state workload distribution. A cost minimization problem is considered in Section 3, where also various numerical results are shown. Section 4 contains some suggestions for further research.

2 The workload

In this section we first present integral equations for the steady-state workload density $v(\cdot)$ (Subsection 2.1), while already deriving the workload density during inactive periods; then we formally solve those integral equations (Subsection 2.2), and finally we present a detailed solution for two special cases: exponentially distributed service requirements (Subsection 2.3) and generally distributed service requirements with $r(x) = r_1 x + r_0$ (Subsection 2.4).

2.1 Integral equations for the workload density

We use the Level Crossing Technique (LCT), cf. [4, 6], which is based on the principle that, in steady state, each level x is crossed just as often from above and from below. We need to distinguish between $x < M$ and $x \geq M$. When $x \geq M$, we have, with $V(0) = \mathbb{P}(V = 0)$ (see also Figure 1):

$$r(x)v(x) = \lambda \int_{y=0}^{x} \mathbb{P}(B > x - y)v(y)\mathrm{d}y + \lambda \mathbb{P}(B > x)V(0), \tag{2.1}$$

and when $x < M$ then

$$r(x)(v(x) - v_I(x)) = \lambda \int_{y=0}^{x} \mathbb{P}(B > x - y)v(y)\mathrm{d}y + \lambda \mathbb{P}(B > x)V(0). \tag{2.2}$$

In both cases, the righthand side represents the upcrossing rate, which seems self-explanatory (see also Sections II.4.5 and III.5.10 of [7] for a similar integral equation for, respectively, the ordinary $M/G/1$ queue and the $M/G/1$ queue with service speed $r(x)$). The lefthand side gives the downcrossing rate. Here one has to realize that for $x \in (0, M)$ there can only be a downcrossing when the server is active; hence the term $v(x) - v_I(x)$ for $x \in (0, M)$. Let us now first determine $v_I(x)$ for $x \in (0, M)$.

The density $v_I(x)$.
One can write

$$\begin{aligned} v_I(x) &= v(x|\text{server inactive})\mathbb{P}(\text{server inactive}), \quad 0 < x < M, \\ &= 0, \quad x \geq M. \end{aligned} \tag{2.3}$$

The probability p_I that the server is inactive equals the fraction of time that the server is inactive; hence, with m_0 and m_1 the mean of an inactive and of an active period, we have

$$p_I = \frac{m_0}{m_0 + m_1}. \tag{2.4}$$

It is easy to determine m_0. Obviously,

$$m_0 = 1/\lambda \times (1 + m(M)), \tag{2.5}$$

where $m(x)$ is the renewal function, defined as $m(x) := \mathbb{E}N(x)$, with $\{N(x) := \sup\{n : B_1 + \cdots + B_n \leq x\}$ (cf. Chapter 3 of [17]). The conditional workload density given that the server is inactive also follows from renewal theory, and turns out to be closely related to the renewal function. Introducing the density

$$y(x) := v(x|\text{server inactive}), \tag{2.6}$$

with distribution $Y(\cdot)$, we shall prove the following.

Theorem 1.

$$Y(x) = \frac{1 + m(x)}{1 + m(M)}, \quad 0 \leq x \leq M. \tag{2.7}$$

Proof. Remove all the active periods, to obtain a sequence of successive inactive periods. Applying LCT to the thus obtained process, equating the numbers of workload downcrossings and upcrossings of any level $x \in [0, M]$ we obtain:

$$\lambda \int_0^x \mathbb{P}(B > x - u)\mathrm{d}Y(u) = \frac{1}{m_0}, \quad 0 \leq x \leq M. \tag{2.8}$$

The righthand side of this equation reflects the event in which level M is *upcrossed*, which instantaneously (because we have omitted the active periods) is followed by a jump from above M to level 0 – and hence a *downcrossing* of each level $x \in [0, M]$. This happens once

per inactive period; hence the term $\frac{1}{m_0}$. Divide both sides of (2.8) by λ and observe that (e.g., using (2.8) for $x = 0$)

$$Y(0) = \mathbb{P}(V = 0|\text{server inactive}) = \frac{1}{\lambda m_0}. \tag{2.9}$$

Rewrite (2.8) into

$$Y(x)-Y(0) = \int_0^x \mathbb{P}(B < x-u)\mathrm{d}Y(u) = \mathbb{P}(B < x)Y(0)+\int_0^x \mathbb{P}(B < x-u)y(u)\mathrm{d}u, \quad 0 \le x \le M, \tag{2.10}$$

and subsequently into

$$\frac{Y(x) - Y(0)}{Y(0)} = \mathbb{P}(B < x) + \int_0^x \mathbb{P}(B < x - u)\mathrm{d}\frac{Y(u) - Y(0)}{Y(0)}, \quad 0 \le x \le M. \tag{2.11}$$

Comparison with the well-known renewal equation (cf. Chapter 3 of [17])

$$m(x) = \mathbb{P}(B < x) + \int_0^x \mathbb{P}(B < x - u)dm(u), \tag{2.12}$$

shows that $\frac{Y(x)-Y(0)}{Y(0)} = m(x)$ and hence $Y(x) = Y(0)(1 + m(x))$. Finally use the fact that $Y(M) = 1$.

Remark 2.1. In the special case in which $B \sim \exp(\mu)$, one has $m(x) = \mu x$, and hence $y(x) = \frac{\mu}{1+\mu M}$; the workload during an inactive period, when positive, is uniformly distributed on $(0, M)$.

Remark 2.2. For future use we observe that $v(\cdot)$ has a discontinuity in $x = M$, as revealed by (2.1) and (2.2):

$$v(M) - v(M-) = -v_I(M-). \tag{2.13}$$

Remark 2.3. We close this subsection by pointing out that, in all model variants to be studied in this paper, we have the following relation:

$$\frac{1}{\lambda V(0)} = m_0 + m_1. \tag{2.14}$$

Indeed, $\lambda V(0)$ is the rate of a customer arriving in an empty system, and hence $\frac{1}{\lambda V(0)}$ is the mean cycle time, viz., the sum of the means of an inactive period and an active period. Since m_0 is known, Formula (2.14) constitutes a relation between two important quantities: the probability $V(0)$ of an empty system, and the mean m_1 of an active period. These quantities will appear in most of the key workload formulas to be discussed in the sequel. Notice in particular, combining (2.3), (2.4), (2.5), (2.7) and (2.14), that

$$v_I(x) = V(0)m'(x), \quad 0 \le x < M. \tag{2.15}$$

2.2 Solution of the integral equations

In this subsection we present a formal solution of the integral equations (2.1) and (2.2). First rewrite these two equations into one integral equation:

$$v(x) = \int_{y=0}^{x} K(x,y)v(y)\mathrm{d}y + L(x), \tag{2.16}$$

where

$$K(x,y) := \frac{\lambda \mathbb{P}(B > x-y)}{r(x)}, \quad 0 \le y < x, \tag{2.17}$$

and (using (2.14) to express the unknown constant m_1 into $V(0)$):

$$\begin{aligned} L(x) &:= V(0)K(x,0), \quad x \ge M, \\ L(x) &:= V(0)K(x,0) + v_I(x) = V(0)[K(x,0) + m_0\lambda y(x)], \quad x < M, \end{aligned} \tag{2.18}$$

where the last equality follows from (2.5), (2.7) and (2.15). Integral equation (2.16) is a Volterra integral equation of the second kind. The classical Picard iteration ([18], Chapter I) results in the following formal solution in terms of an infinite series of convolutions. Define recursively

$$K_n(x,y) := \int_{y}^{x} K(x,z)K_{n-1}(z,y)\mathrm{d}z, \quad 0 < y < x, \quad n = 2,3,\ldots,$$

where $K_1(x,y) := K(x,y)$. Then the Picard iteration applied to (2.16) yields

$$\begin{aligned} v(x) &= L(x) + \int_{y=0}^{x} K(x,y)[L(y) + \int_{z=0}^{y} K(y,z)v(z)\mathrm{d}z]\mathrm{d}y \\ &= \ldots = L(x) + \sum_{n=1}^{\infty} \int_{0}^{x} K_n(x,y)L(y)\mathrm{d}y. \end{aligned} \tag{2.19}$$

One can follow the approach of [12] and use the bound $K(x,y) \le \frac{\lambda}{r(x)}$ to show inductively that $K_{n+1}(x,y) \le \frac{(\int_y^x \frac{\lambda}{r(u)}\mathrm{d}u)^n}{n!}\frac{\lambda}{r(x)}$. That implies the convergence of the infinite sum in (2.19).

What remains to be done is to find the unknown constant $V(0)$. This can be done by using the normalizing condition $\int_0^\infty v(x)\mathrm{d}x + V(0) = 1$.

Although one thus in principle obtains an expression for $v(\cdot)$, this solution is a rather formal one, expressed in terms of an infinite sum of non-explicit convolutions. Therefore we restrict ourselves in the next subsections to two special cases, for which we aim to derive more explicit expressions for $v(\cdot)$, viz., (i) the case of exponentially distributed service times, and (ii) the case of $r(x) = r_1 x + r_0$.

2.3 Solution of the integral equations in the case of exponentially distributed service requirements

In this subsection we assume that $B(x) = 1 - \mathrm{e}^{-\mu x}$. After multiplication by $\mathrm{e}^{\mu x}$, Formula (2.1) reduces to

$$r(x)\mathrm{e}^{\mu x}v(x) = \lambda \int_{y=0}^{x} \mathrm{e}^{\mu y}v(y)\mathrm{d}y + \lambda V(0), \quad x \geq M, \tag{2.20}$$

which after differentiation and straightforward calculations yields:

$$v'(x) = \frac{\lambda - \mu r(x) - r'(x)}{r(x)}v(x), \quad x \geq M. \tag{2.21}$$

Hence, remembering that $R(x) = \int_0^x \frac{1}{r(y)}\mathrm{d}y$, and introducing the yet unknown constant C:

$$v(x) = C\frac{\mathrm{e}^{\lambda R(x) - \mu x}}{r(x)}, \quad x \geq M. \tag{2.22}$$

We now turn to (2.2). In the case of exponentially distributed service requirements, we already observed in Subsection 2.1 that $v(x|\text{server inactive})$ is constant. Hence also $v_I(x)$ is constant: $v_I(x) = v_I(0)$, $0 \leq x < M$. After multiplication by $\mathrm{e}^{\mu x}$, Formula (2.2) reduces to

$$r(x)\mathrm{e}^{\mu x}v(x) = \lambda \int_{y=0}^{x} \mathrm{e}^{\mu y}v(y)\mathrm{d}y + \lambda V(0) + r(x)\mathrm{e}^{\mu x}v_I(0), \quad x < M, \tag{2.23}$$

which after differentiation and straightforward calculations yields:

$$v'(x) = \frac{\lambda - \mu r(x) - r'(x)}{r(x)}v(x) + v_I(0)(\frac{r'(x)}{r(x)} + \mu), \quad x < M. \tag{2.24}$$

Using variation of constants to solve this inhomogeneous first-order differential equation, we obtain for $x < M$:

$$\begin{aligned} v(x) &= C^*\frac{\mathrm{e}^{\lambda R(x) - \mu x}}{r(x)} + v_I(0)\int_{y=0}^{x}(\frac{r'(y)}{r(y)} + \mu)\mathrm{e}^{\lambda R(x,y) - \mu(x-y)}\frac{r(y)}{r(x)}\mathrm{d}y \\ &= \frac{\mathrm{e}^{\lambda R(x) - \mu x}}{r(x)}[C^* + v_I(0)\int_{y=0}^{x}(r'(y) + \mu r(y))\mathrm{e}^{-\lambda R(y) + \mu y}\mathrm{d}y]. \end{aligned} \tag{2.25}$$

We still need to determine several unknown constants: $V(0)$, $v_I(0)$ and the two constants C and C^*. For this, we have the following equations:
(i) The normalizing condition: $\int_0^\infty v(x)\mathrm{d}x + V(0) = 1$.
(ii) Formula (2.15) for $x = 0$ yields $v_I(0) = \mu V(0)$.
(iii) It follows from (2.2) for $x = 0$ that $r(0)[v(0) - v_I(0)] = \lambda V(0)$, while (2.25) implies that $r(0)v(0) = C^*$; hence

$$C^* = \lambda V(0) + r(0)v_I(0) = V(0)[\lambda + \mu r(0)]. \tag{2.26}$$

(iv) Finally we use the discontinuity of $v(\cdot)$ in M, as described in Remark 2.2. After a lengthy calculation, C follows from (2.13), (2.22) en (2.25):

$$C = V(0)[\lambda + \lambda\mu \int_0^M \mathrm{e}^{-\lambda R(y)+\mu y}\mathrm{d}y]. \tag{2.27}$$

The fact that $v(x)$ is both for $x < M$ and $x > M$ linearly expressed in $V(0)$ makes it relatively straightforward to determine that remaining unknown $V(0)$ from the normalizing condition.

The following theorem summarizes our results of this subsection. The expression for $v(x)$, $x < M$ was obtained by using (2.25) and (2.26) and performing a partial integration.

Theorem 2.

$$v(x) = \mu V(0) + V(0)\frac{\mathrm{e}^{\lambda R(x)-\mu x}}{r(x)}\lambda(1 + \mu \int_{y=0}^{x} \mathrm{e}^{-\lambda R(y)+\mu y}\mathrm{d}y), \quad x < M, \tag{2.28}$$

$$v(x) = V(0)\frac{\mathrm{e}^{\lambda R(x)-\mu x}}{r(x)}\lambda(1 + \mu \int_{y=0}^{M} \mathrm{e}^{-\lambda R(y)+\mu y}\mathrm{d}y), \quad x \geq M, \tag{2.29}$$

with

$$\begin{aligned} V(0)^{-1} &= 1 + \mu M + \int_{x=0}^{M} \frac{\mathrm{e}^{\lambda R(x)-\mu x}}{r(x)}\lambda(1 + \mu \int_{y=0}^{x} \mathrm{e}^{-\lambda R(y)+\mu y}\mathrm{d}y)\mathrm{d}x \\ &+ \int_{x=M}^{\infty} \frac{\mathrm{e}^{\lambda R(x)-\mu x}}{r(x)}\lambda(1 + \mu \int_{y=0}^{M} \mathrm{e}^{-\lambda R(y)+\mu y}\mathrm{d}y)\mathrm{d}x. \end{aligned} \tag{2.30}$$

2.4 Solution of the integral equations in the case of linear service speed

In this subsection we allow the service requirements to be generally distributed, but we assume the service speed to be linear: $r(x) = r_1 x + r_0$, where $r_0, r_1 > 0$. Notice that the stability condition (1.1) is always fulfilled, and that the condition that $R(x) < \infty$ for all finite x is also fulfilled. We apply Laplace transformation to (2.1) and (2.2), introducing

$$\begin{aligned} \phi(s) &:= \int_{x=0}^{\infty} \mathrm{e}^{-sx} v(x)\mathrm{d}x \qquad (2.31) \\ &= \int_{x=0}^{M-} \mathrm{e}^{-sx} v(x)\mathrm{d}x + \mathrm{e}^{-sM}(v(M) - v(M-)) + \int_{x=M}^{\infty} \mathrm{e}^{-sx} v(x)\mathrm{d}x, \quad \text{Re } s \geq 0. \end{aligned}$$

It follows from (2.1) and (2.2) that

$$-r_1 \frac{\mathrm{d}}{\mathrm{d}s}\phi(s) + r_0\phi(s) = \lambda\frac{1-\beta(s)}{s}\phi(s) + \lambda\frac{1-\beta(s)}{s}V(0) + \gamma(s), \tag{2.32}$$

where we introduce $\gamma(s) := \int_{x=0}^{M} \mathrm{e}^{-sx}(r_1 x + r_0) v_I(x) \mathrm{d}x$. According to (2.15) we have $v_I(x) = V(0)m'(x)$. Hence $\gamma(s)$ is known up to the yet unknown $V(0)$:

$$\gamma(s) = V(0) \int_{x=0}^{M} \mathrm{e}^{-sx}(r_1 x + r_0) m'(x) \mathrm{d}x =: V(0)\delta(s). \tag{2.33}$$

Solving the inhomogeneous first-order differential equation (2.32) yields, with D a yet unknown constant:

$$\phi(s) = \mathrm{e}^{\frac{r_0}{r_1}s - \frac{\lambda}{r_1}\int_0^s \frac{1-\beta(u)}{u}\mathrm{d}u}[D - V(0)\int_{v=0}^{s}[\frac{\lambda}{r_1}\frac{1-\beta(v)}{v} + \frac{1}{r_1}\delta(v)]\mathrm{e}^{-\frac{r_0}{r_1}v + \frac{\lambda}{r_1}\int_0^v \frac{1-\beta(u)}{u}\mathrm{d}u}\mathrm{d}v]. \tag{2.34}$$

We still need to determine two unknown constants: $V(0)$ and D. Noticing that $\lim_{s\to\infty}\phi(s) = 0$ gives

$$D = V(0)\int_{v=0}^{\infty}[\frac{\lambda}{r_1}\frac{1-\beta(v)}{v} + \frac{1}{r_1}\delta(v)]\mathrm{e}^{-\frac{r_0}{r_1}v + \frac{\lambda}{r_1}\int_0^v \frac{1-\beta(u)}{u}\mathrm{d}u}\mathrm{d}v. \tag{2.35}$$

Indeed, it is easy to see that the exponential in (2.34), $\mathrm{e}^{\frac{r_0}{r_1}s - \frac{\lambda}{r_1}\int_0^s \frac{1-\beta(u)}{u}\mathrm{d}u}$, tends to ∞ for $s \to \infty$, because the $\frac{r_0}{r_1}s$ term dominates for large s:

$$|\int_0^s \frac{1-\beta(u)}{u}\mathrm{d}u| \le \int_0^1 |\frac{\beta(u)-\beta(0)}{u}|\mathrm{d}u + \int_1^s \frac{1}{u}\mathrm{d}u \le \mathbb{E}B + \ln(s).$$

The normalizing condition states that $\phi(0) + V(0) = 1$, and hence

$$D = 1 - V(0). \tag{2.36}$$

We thus obtain one linear equation in the remaining unknown $V(0)$. The following theorem summarizes our results of this subsection.

Theorem 3.

$$\phi(s) = \mathrm{e}^{\frac{r_0}{r_1}s - \frac{\lambda}{r_1}\int_0^s \frac{1-\beta(u)}{u}\mathrm{d}u}V(0)\int_{v=s}^{\infty}[\frac{\lambda}{r_1}\frac{1-\beta(v)}{v} + \frac{1}{r_1}\delta(v)]\mathrm{e}^{-\frac{r_0}{r_1}v + \frac{\lambda}{r_1}\int_0^v \frac{1-\beta(u)}{u}\mathrm{d}u}\mathrm{d}v], \tag{2.37}$$

with

$$V(0)^{-1} = 1 + \int_{v=0}^{\infty}[\frac{\lambda}{r_1}\frac{1-\beta(v)}{v} + \frac{1}{r_1}\delta(v)]\mathrm{e}^{-\frac{r_0}{r_1}v + \frac{\lambda}{r_1}\int_0^v \frac{1-\beta(u)}{u}\mathrm{d}u}\mathrm{d}v]. \tag{2.38}$$

Remark 2.4. If $r_1 = 0$, our system reduces to an ordinary $M/G/1$ queue with a server which is switched off when the system becomes empty and gets activated again when the workload reaches a certain threshold (D-policy, cf. [8]). It readily follows from (2.32) (where the first term disappears for $r_1 = 0$) that $\phi(s)$ now becomes the product of the workload LST in an ordinary $M/G/1$ queue and an additional term that relates to the off-periods; such decomposition results are well-known in the literature of queues with vacations.

Remark 2.5. A tedious but straightforward calculation verifies that the results of Theorems 2 and 3 agree when $r(x) = r_1 x + r_0$ and $B(x) = 1 - \mathrm{e}^{-\mu x}$. One first takes Laplace transforms in (2.28) and (2.29), obtaining

$$\begin{aligned}\frac{\phi(s)}{V(0)} &= \int_{x=0}^{M} \mu \mathrm{e}^{-sx}\mathrm{d}x + \lambda \int_{x=0}^{\infty} \mathrm{e}^{-(s+\mu)x} \frac{\left(\frac{r_1 x + r_0}{r_0}\right)^{\frac{\lambda}{r_1}}}{r_1 x + r_0}\mathrm{d}x \\ &+ \lambda\mu \int_{y=0}^{M} \mathrm{e}^{\mu y}\left(\frac{r_1 y + r_0}{r_0}\right)^{-\frac{\lambda}{r_1}}\mathrm{d}y \int_{x=y}^{\infty} \mathrm{e}^{-(s+\mu)x} \frac{\left(\frac{r_1 x + r_0}{r_0}\right)^{\frac{\lambda}{r_1}}}{r_1 x + r_0}\mathrm{d}x. \qquad (2.39)\end{aligned}$$

One partial integration in the last integral of (2.39) gives a cancellation against the first term in the righthand side. Subsequently the transformation $\frac{r_1 x + r_0}{r_1 y + r_0} = \frac{v+\mu}{s+\mu}$ leads to the expression in (2.37).

Remark 2.6. From (2.34), using that $\mathbb{E}V = -\phi'(s)|_{s=0}$, it follows that

$$\begin{aligned}\mathbb{E}V &= -\frac{r_0}{r_1}(1 - V(0)) + \frac{\lambda \mathbb{E}B}{r_1}(1 - V(0)) + \frac{\lambda \mathbb{E}B}{r_1}V(0) + \frac{1}{r_1}\int_{x=0}^{M}(r_1 x + r_0)v_I(x)\mathrm{d}x \\ &= \frac{\lambda \mathbb{E}B - r_0}{r_1} + V(0)[\frac{r_0}{r_1} + \frac{1}{r_1}\int_{x=0}^{M}(r_1 x + r_0)m'(x)\mathrm{d}x]. \qquad (2.40)\end{aligned}$$

In the special case of $\exp(\mu)$ service times, $m'(x) = \mu$ and we have:

$$\mathbb{E}V = \frac{\lambda - \mu r_0}{\mu r_1} + V(0)[\frac{r_0}{r_1}(1 + \mu M) + \frac{1}{2}\mu M^2]. \qquad (2.41)$$

3 Cost optimization

Suppose that two types of costs are involved in the operation of the system: holding costs c_h per time unit for each unit of work present in the system, and switching costs c_s for each time the server is switched on. We are interested in choosing M such that the system costs are minimized. Hence we consider the following minimization problem (cf. (2.14)):

$$\mathrm{Minimize}_{\mathrm{M}} \qquad c_h \mathbb{E}V + c_s \frac{1}{m_0 + m_1} = c_h \mathbb{E}V + c_s \lambda V(0). \qquad (3.1)$$

In addition, the system might receive profits from each amount of work that is being served. However, we can ignore that profit, as it does not depend on the choice of M.

We focus on the case, studied in Subsection 2.4, in which $r(x) = r_1 x + r_0$. It follows from (2.41) that our optimization problem becomes:

$$\mathrm{Min}_{\mathrm{M}} \;\; c_h \frac{\lambda \mathbb{E}B - r_0}{r_1} + c_h V(0)[\frac{r_0}{r_1} + \frac{1}{r_1}\int_{x=0}^{M}(r_1 x + r_0)m'(x)\mathrm{d}x] + c_s \lambda V(0), \qquad (3.2)$$

which amounts to minimizing, w.r.t. M, the function

$$f(M) := V(0)[\frac{c_h r_0}{r_1} + \frac{c_h}{r_1}\int_{x=0}^{M}(r_1 x + r_0)m'(x)\mathrm{d}x + c_s\lambda]; \tag{3.3}$$

here $V(0)$ depends on M, and is given by (2.38).

The derivative of $f(M)$ w.r.t. M should be zero, and hence M should satisfy

$$\begin{aligned}\frac{c_h}{r_1}(r_1 M + r_0)m'(M) &= V(0)[c_s\lambda + \frac{c_h r_0}{r_1} + \frac{c_h}{r_1}\int_0^M (r_1 x + r_0)m'(x)\mathrm{d}x] \\ &\times \int_0^\infty \frac{1}{r_1}\mathrm{e}^{-vM}(r_1 M + r_0)m'(M)\mathrm{e}^{-\frac{r_0}{r_1}v + \frac{\lambda}{r_1}\int_0^v \frac{1-\beta(u)}{u}\mathrm{d}u}\mathrm{d}v. \end{aligned}\tag{3.4}$$

Let us now restrict ourselves to the case of $\exp(\mu)$ service times. Then (3.4) reduces to

$$\begin{aligned}\frac{c_h\mu}{r_1}(r_1 M + r_0) &= V(0)[c_s\lambda + c_h\frac{r_0}{r_1} + \frac{c_h\mu}{r_1}(\frac{r_1}{2}M^2 + r_0 M)] \\ &\times (M + \frac{r_0}{r_1})\mu\int_0^\infty \mathrm{e}^{-vM}\mathrm{e}^{-\frac{r_0}{r_1}v}(\frac{\mu+v}{\mu})^{\frac{\lambda}{r_1}}\mathrm{d}v. \end{aligned}\tag{3.5}$$

Here $1/V(0)$ simplifies to

$$\frac{1}{V(0)} = 1 + \int_0^\infty \left(\frac{\lambda}{r_1}\frac{1}{\mu+v} + \frac{\mu}{r_1}\int_0^M \mathrm{e}^{-vx}(r_1 x + r_0)\mathrm{d}x\right)\mathrm{e}^{-\frac{r_0}{r_1}v}(\frac{\mu+v}{\mu})^{\frac{\lambda}{r_1}}\mathrm{d}v. \tag{3.6}$$

Remark 4.1. Matters simplify further if we assume that

$$\lambda = r_1. \tag{3.7}$$

By interchanging the two integrals in (3.6), we then obtain an explicit expression for $V(0)$:

$$\frac{1}{V(0)} = 1 + \frac{r_1}{\mu r_0} + \mu M + \ln\frac{\frac{r_0}{r_1} + M}{\frac{r_0}{r_1}}. \tag{3.8}$$

Remark 4.2. It should be observed that, if we take $r_0 = 0$, then the first integral in the right-hand side of (3.6) diverges, giving $V(0) = 0$. The explanation is that, when the service speed is $r_1 x$, the system never becomes zero.

3.1 Numerical examples

We plot some graphs to show the behavior of the cost function as a function of the threshold M.

In our numerical experiments, we fix the arrival rate: $\lambda = 1$ and show the effect of other parameters on the cost function. Intuitively, on the one hand, a large threshold M leads to a larger workload in the system since the inactive period is longer. On the other hand, a large threshold may prevent frequent switching and thus may reduce the switching cost.

Thus, it is expected to have an optimal M which balances the two types of costs. In all our numerical experiments, the cost function was convex, and we found a unique optimal M. However, we have not yet been able to analytically show convexity of the cost function in M.

Case 1: Cost function vs. M for various r_0
Figure 2 displays the cost function against the threshold M for several values of r_0; $r_0 = 1, 5, 10$. Other parameters are as follows: $c_h = 0.1, c_s = 1, \mu = 1, r_1 = 10$. Notice the above-mentioned convexity of the curves, guaranteeing that there is an optimal M that minimizes the cost function. We also observe that the optimal M is almost insensitive to r_0 in this case. A close inspection shows that the optimal M slightly increases with r_0.

Case 2: Cost function vs. M for various r_1
Figure 3 displays the cost function against the threshold M for several values of r_1; $r_1 = 0.1, 0.5, 1$. Parameters are fixed as follows: $c_h = 0.1, c_s = 1, \mu = 1, r_0 = 1$. The optimal value of M for a larger r_1 is seen to be bigger than that for a smaller r_1.

Case 3: Cost function vs. M for various μ
In this case, we display the cost function against the threshold M for various values of μ. Fixed parameters are as follows: $c_h = 0.1, c_s = 1, r_0 = 1$. Figure 4 is for the case $r_1 = 0.1$ while Figure 5 is for the case $r_1 = 10$. We observe from Figures 4 and 5 that the optimal value of M increases with μ.

Numerical insights. Extensive numerical results suggest that the cost function is a convex function of M, so that there exists an optimal value of the threshold M. Furthermore, not surprisingly, the optimal threshold increases with r_0, r_1 and μ. A rigorous proof of the convexity of the cost function in the threshold M is left for future work.

4 Conclusion and suggestions for further research

Motivated by the trade-off issue between processing speed and energy consumption in data centers, we have studied a queueing system in which the service speed is a function of the workload, and in which the server switches off when becoming idle, only to be activated again when the workload reaches a certain threshold. We have derived the (LST of the) workload distribution, and we have used an expression for its mean to determine the threshold level that minimizes a certain cost function.

Topics on our research agenda include:
(i) A further study of the cost minimization problem, in which we also would like to tackle the question whether the cost function is convex. We furthermore wish to extend our cost function, taking power consumption as a function of processing speed into account.
(ii) A study of the active period distribution and the distribution of a full cycle, consisting of an inactive and subsequent active period. It should be observed that the length of an active period depends on the length of the preceding inactive period, but that the length

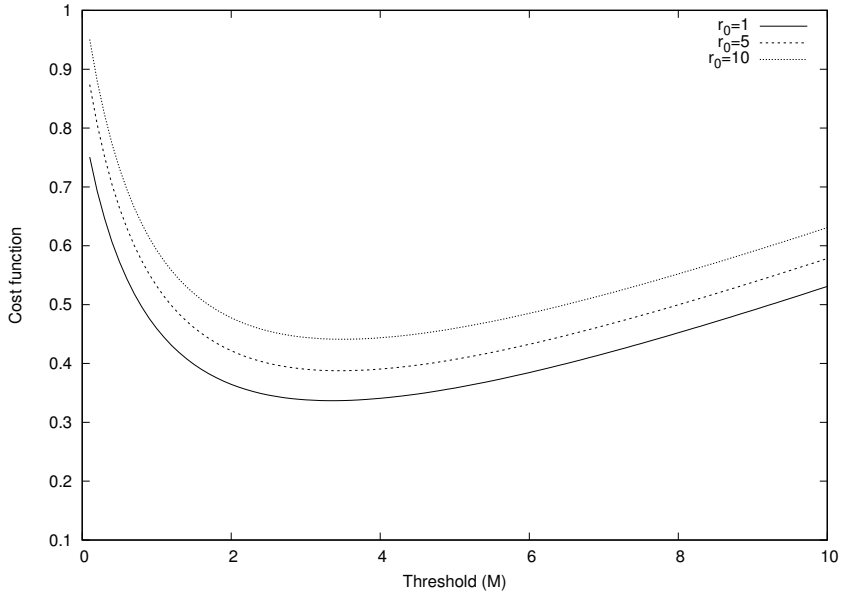

Figure 2: Cost function for Case 1: $c_h = 0.1, c_s = 1$; various r_0.

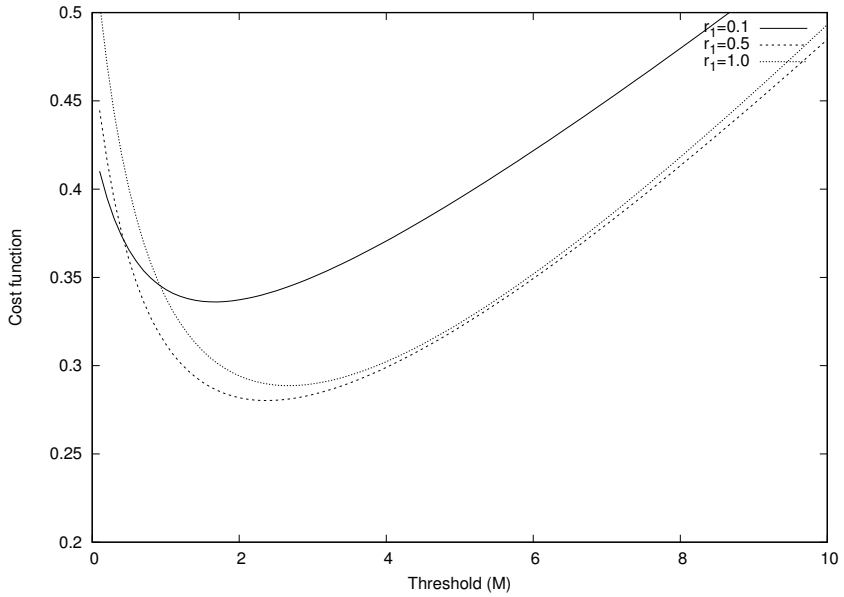

Figure 3: Cost function for Case 2: $c_h = 0.1, c_s = 1$; various r_1.

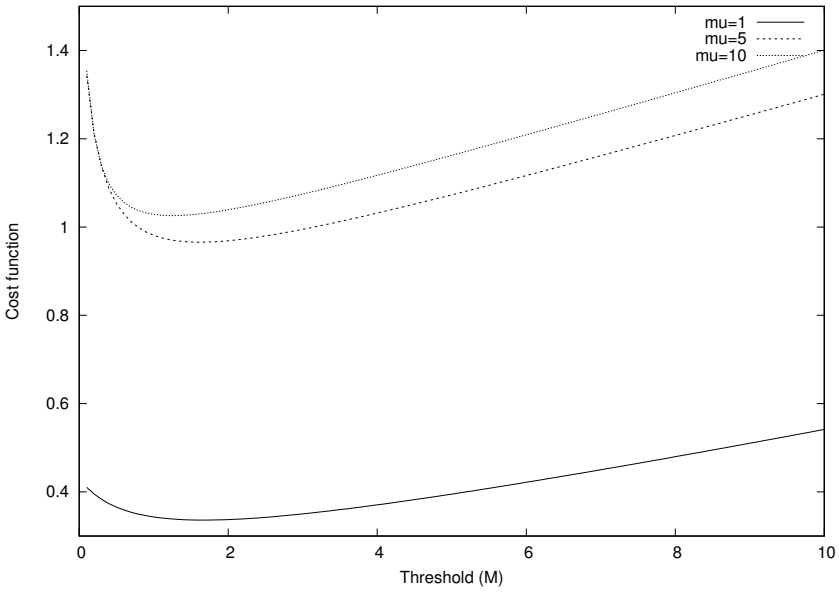

Figure 4: Cost function for Case 3: $c_h = 0.1, c_s = 1, r_1 = 0.1$; various μ.

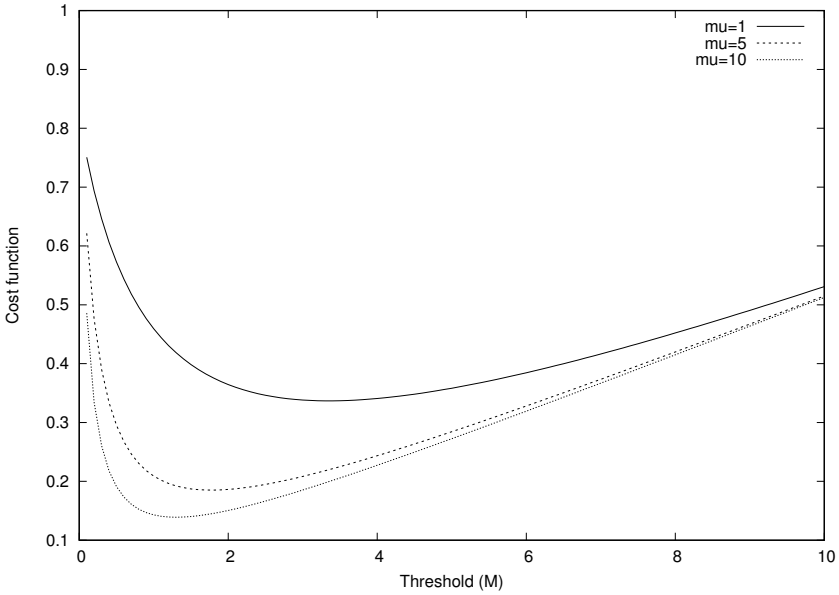

Figure 5: Cost function for Case 3: $c_h = 0.1, c_s = 1, r_1 = 10$; various μ.

of an inactive period does not depend on the length of the preceding active period; hence the distribution of the sum of the lengths of an inactive and subsequent active period in general differs from the distribution of the sum of the lengths of an active and subsequent inactive period.
(iii) We are presently analyzing the model variant in which the processing speed $r(x)$ is piecewise constant ($r(x) = r_i$ when the workload is lying in an interval J_i, $i = 1, 2, \dots$), and in which the service requirement distribution $B(\cdot)$ is phase-type. This is a case for which it seems to be possible to obtain quite explicit results.

5 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] 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.

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

[5] S. Browne and K. Sigman (1992). Workload-modulated queues with application to storage processes. *J. Appl. Probab.* **29**, 699-712.

[6] J.W. Cohen (1977). On up- and downcrossings. *J. Appl. Probab.* **14**, 405-410.

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

[8] 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.

[9] A. Gandhi, M. Harchol-Balter and I. Adan (2010). Server farms with setup costs. *Performance Evaluation* **67**, 1123-1138.

[10] 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.

[11] 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.

[12] J.M. Harrison and S.I. Resnick (1976). The stationary distribution and first exit probabilities of a storage process with general release rule. *Math. Oper. Res.* **1**, 347-358.

[13] 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.

[14] 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.

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

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

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

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

[19] 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.

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