Centrum voor Wiskunde en Informatica

REPORT*RAPPORT*

PNA

Probability, Networks and Algorithms

Probability, Networks and Algorithms

On a queueing model with service interruption

O.J. Boxma, M.R.H. Mandjes, O. Kella

REPORT PNA-R0610 SEPTEMBER 2006

Centrum voor Wiskunde en Informatica (CWI) is the national research institute for Mathematics and Computer Science. It is sponsored by the Netherlands Organisation for Scientific Research (NWO).
CWI is a founding member of ERCIM, the European Research Consortium for Informatics and Mathematics.

CWI's research has a theme-oriented structure and is grouped into four clusters. Listed below are the names of the clusters and in parentheses their acronyms.

Probability, Networks and Algorithms (PNA)

Software Engineering (SEN)

Modelling, Analysis and Simulation (MAS)

Information Systems (INS)

P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

ISSN 1386-3711

On a queueing model with service interruption

ABSTRACT
Single-server queues in which the server takes vacations arise naturally as models for a wide range of computer-, communication- and production systems. In almost all studies on vacation models, the vacation lengths are assumed to be independent of the arrival, service, workload and queue length processes. In the present study we allow the length of a vacation to depend on the length of the previous active period, viz., the period since the previous vacation. Under rather general assumptions regarding the offered work during active periods and vacations, we determine the steady-state workload distribution. We conclude by discussing several special cases including polling models, and relate our findings to results obtained earlier.

2000 Mathematics Subject Classification: 60K25
Keywords and Phrases: Lévy process, storage process, busy and idle periods, queues with server vacations
Note: Part of this work was done while the third author visited EURANDOM and CWI as a Stieltjes visiting professor. The third author is also supported by grant 964/06 from the Israel Science Foundation. The research of the first and second was done within the framework of the BRICKS project and the European Network of Excellence Euro-NGI.

On a queueing model with service interruptions

Onno Boxma* Michel Mandjes† Offer Kella‡

Abstract

Single-server queues in which the server takes vacations arise naturally as models for a wide range of computer-, communication- and production systems. In almost all studies on vacation models, the vacation lengths are assumed to be independent of the arrival, service, workload and queue length processes. In the present study we allow the length of a vacation to depend on the length of the previous active period, viz., the period since the previous vacation. Under rather general assumptions regarding the offered work during active periods and vacations, we determine the steady-state workload distribution. We conclude by discussing several special cases including polling models, and relate our findings to results obtained earlier.

Keywords: Lévy process, storage process, busy and idle periods, queues with server vacations
AMS Subject Classification: Primary 60K05; Secondary 60K25

*EURANDOM and Department of Mathematics and Computer Science; Eindhoven University of Technology; P.O. Box 513; 5600 MB Eindhoven; The Netherlands (boxma@win.tue.nl).

†Korteweg-de Vries Institute for Mathematics; The University of Amsterdam; Plantage Muidergracht 24; 1018 TV Amsterdam; The Netherlands, and CWI; P.O. Box 94079; 1090 GB Amsterdam; The Netherlands (mmandjes@science.uva.nl).

‡Department of Statistics; The Hebrew University of Jerusalem; Mount Scopus, Jerusalem 91905; Israel (Offer.Kella@huji.ac.il).

1 Introduction

In traditional queueing models a stream of customers arrives at a service station according to some stochastic process. These customers enter a waiting line when they find the server busy, where they wait until they can be processed. Although it is commonly assumed that the server is always available, an interesting alternative model is a system in which the server takes *vacations*, i.e., it alternates between an active and inactive mode. During the active times it is working at full speed, whereas it is not processing any work during the vacations.

Systems with server vacations arise naturally as models for an important class of computer, communication, and production systems. The vacations may, e.g., represent server breakdowns, or periods in which the server processes work generated by an other class of customers. Static priority queues and polling models constitute important classes of models for the latter situation. Performance measures like queue lengths, waiting times and workloads have been intensively studied for many types of vacation models. For extensive surveys on vacation queues we refer to [4], its update [6], the paper [19] that focuses on control aspects, and the book [16]. See also [17, 18] for surveys on polling models. In almost all studies on vacation and polling models, the vacation lengths (or switchover times) are assumed to be independent of other vacations (switchover times) and of the arrival, service, workload, and queue length processes. A notable vacation exception is the paper of Harris and Marchal [11], in which the probability of the server taking a vacation after a service completion, and the length of the vacation, depend on the number of customers present when a service ends. In polling, Altman [1], Groenevelt and Altman [10] and Eliazar [8] study models in which there is interdependence between switchover times. The latter paper also considers a generalization from the common compound Poisson input processes to input according to general Lévy subordinator processes.

Contribution. The main goal of the present paper is to extend earlier results to the situation that allows the length of a vacation to depend on the length of the previous *active* period, viz., the period since the previous vacation. It is noted that in many applications, such a dependence is very natural. For example, in polling models a relatively long visit of the server to a tagged queue probably leads to a substantial accumulation of work in subsequent queues, and hence to a relatively long *intervisit* time of the tagged queue. In-

deed, we shall see that particular model choices give rise to polling systems. We remark that our findings may be viewed as queueing-theoretic results. We have chosen, however, to put them into the more general framework of Lévy processes, and they can be interpreted as results on storage processes, cf. Prabhu [12]. Let us now give a brief model description.

- During *active periods*, work is generated according to a Lévy process $X^{\mathrm{D}}(\cdot)$ with negative drift, until the workload reaches zero (i.e., the storage reservoir is empty).
- From then on, the storage level behaves according to a second Lévy process $X^{\mathrm{U}}(\cdot)$, which we assume to be non-decreasing. As during this period work accumulates in the queue, we may interpret it as a vacation; it lasts $aI+bV$, where I is a function of the length of the preceding active period, and V is an independent vacation time, and a and b are given non-negative scalars.

  The case in which the workload is still zero after $aI + bV$, has to be treated separately: then the vacation period is extended until work is generated by $X^{\mathrm{U}}(\cdot)$.
- Subsequently a new active period starts; etc.

In this way the stochastic storage process alternatingly experiences active and passive (vacation) periods. Observe that the classical M/G/1 queue with single vacations is a special case of our model; it is obtained by taking $a = 0$, and by $X^{\mathrm{U}}(\cdot)$ being a compound Poisson process, and $X^{\mathrm{D}}(\cdot)$ exactly the same compound Poisson process but now decreased by a linear drift term.

Methodology. A few words on the methodology, and the nature of the obtained results. We concentrate on the derivation of the steady-state workload distribution for the above-described model. We do so by first considering the distribution at the embedded epochs in which the server switches between a vacation and an active period. We express the state of the system at such an embedded epoch in terms of the system at the previous embedded epoch, and then find the Laplace-Stieltjes transform of the workload by iteration.
Then we use the result for the embedded epochs to characterize the distribution of the workload at *arbitrary* epochs, relying on martingale techniques and renewal arguments. For the classical M/G/1 queue with server vacations, Doshi [5] presents a related analysis, specializing to the distributions during active periods and vacations.

Organization. The paper is organized as follows. Section 2 contains a detailed model description and some preliminary results. In Section 3 we determine the distribution of the workload at epochs in which the process switches from *passive* to *active*. In Section 4 this result is exploited to obtain the steady-state workload distribution during active periods, during passive periods, and overall. Section 5 considers special cases and ramifications.

2 Model description and some preliminaries

In this section, we formally introduce the model of our storage system, that can alternatively be interpreted as a queue with service interruptions. We also present some preliminaries that we frequently use in our analysis.
The dynamic behavior of the storage system consists (alternatingly) of service periods (or: active periods) and interruptions (or: passive periods, vacations), as follows.

- Suppose there is $z \geq 0$ present in the storage system at the beginning of a service period. The storage level evolves according to a Lévy process $X^{\mathrm{D}}(\cdot)$ until the storage level reaches 0. Let the drift $\varrho^{\mathrm{D}} := \mathbb{E}[X^{\mathrm{D}}(1)]$ be negative (but finite). Throughout it is assumed that $X^{\mathrm{D}}(\cdot)$ has no negative jumps; as an immediate consequence, level 0 is actually attained, say at time $\tau(z)$:

  $$\tau(z) := \inf\left\{t \geq 0 : z + X^{\mathrm{D}}(t) = 0\right\}.$$

  We define the *Laplace exponent* of $X^{\mathrm{D}}(\cdot)$ by $\varphi^{\mathrm{D}}(\alpha) := \log \mathbb{E}[\exp(-\alpha X^{\mathrm{D}}(1))]$, and hence $\mathbb{E}[\exp(-\alpha X^{\mathrm{D}}(t))] = e^{\varphi^{\mathrm{D}}(\alpha)t}$. Also, $\psi^{\mathrm{D}}(\cdot)$ denotes the inverse of $\varphi^{\mathrm{D}}(\cdot)$. It is well-known that for any Lévy process that has no negative jumps, $\tau(\cdot)$ is a Lévy process itself, with Laplace exponent $-\psi^{\mathrm{D}}(\alpha)$, that is

  $$\mathbb{E}[e^{-\alpha\tau(z)}] = e^{-\psi^{\mathrm{D}}(\alpha)z}, \tag{1}$$

  see, for instance, Thm. 46.3 in [13]. Notice that, for $\alpha \geq 0$, $\psi^{\mathrm{D}}(\alpha)$ is uniquely defined as the inverse of $\varphi^{\mathrm{D}}(\cdot)$, as $\varphi^{\mathrm{D}}(\cdot)$ increases on $[0, \infty)$.

- At the moment that the storage system hits 0, a service interruption starts. From then on, the storage level evolves according to a second Lévy process $X^{\mathrm{U}}(\cdot)$, which

is assumed to be non-decreasing; in other words: $X^{\rm U}(\cdot)$ is a *subordinator*. This service interruption lasts for a time $\sigma(z)$ that equals $aI + bV$, where I denotes the amount of work generated by an *external* Lévy process (also a subordinator) $X^{\rm E}(\cdot)$ during $\tau(z)$, and V an independent vacation time (the vacations constitute a sequence of i.i.d. non-negative random variables); a and b are given non-negative scalars. Let $\varphi^{\rm U}(\cdot)$ be the Laplace exponent belonging to $X^{\rm U}(\cdot)$, and $\varphi^{\rm E}(\cdot)$ the one belonging to $X^{\rm E}(\cdot)$. Finally, we let $\varrho^{\rm U} := \mathbb{E}[X^{\rm U}(1)] < \infty$ and $\varrho^{\rm E} := \mathbb{E}[X^{\rm E}(1)] < \infty$ be the corresponding (positive) drifts.

Eqn. (1) immediately yields the transform of the duration of the service interruption, if the preceding service period started off at storage level z:

$$\mathbb{E}[e^{-\alpha\sigma(z)}] = \mathbb{E}[e^{\varphi^{\rm E}(\alpha a)\tau(z)}]\mathbb{E}[e^{-\alpha bV}] = e^{-\psi^{\rm D}(-\varphi^{\rm E}(\alpha a))z}\mathbb{E}[e^{-\alpha bV}]. \tag{2}$$

The case that the queue is still empty after $aI + bV$ (i.e., at time $\tau(z) + \sigma(z)$) should be handled separately. In this situation we extend the vacation period until work is generated by $X^{\rm U}(\cdot)$. We denote this additional period by the random variable L, whereas B stands for the storage level as soon as it becomes positive.

- Then a new service period starts again (i.e., the Lévy process $X^{\rm D}(\cdot)$ becomes active again), etc.

3 Equilibrium distribution of embedded process

In this section we concentrate on the distribution of the storage level at embedded epochs, viz. *the epochs at which an active period starts*; it is directly verified that these storage levels constitute a discrete-time Markov process on $[0, \infty)$. To characterize the equilibrium distribution of the storage level at the embedded epochs, we apply a two-step procedure. Throughout, we let $Z(t)$ denote the storage level at time t.

In the first step we let z be some given initial storage level at time 0 (assuming that an active period starts at time 0). Then we compute the transform of the storage level at the epoch at which the next active period starts. More precisely: we give an explicit expression for $\mathbb{E}[e^{-\alpha Z(\vartheta(z))}]$; here $\vartheta(z)$ is defined as $\tau(z) + \sigma(z)$ if the storage level at $\tau(z) + \sigma(z)$ is positive, and as $\tau(z) + \sigma(z) + L$ otherwise.

In the second step we find the 'invariant' distribution of the initial storage level, i.e., the distribution of Z_0 such that

$$\mathbb{E}[e^{-\alpha Z_0}] = \mathbb{E}[e^{-\alpha Z(\vartheta(Z_0))}].$$

As a by-product, we also derive the stability condition of the storage model. In the next section, we use these results to find the steady-state distribution of the storage level at an arbitrary point in time.

Recursion. Assume that there is some initial storage level z at time 0. We first suppose that the interruption period equals $aI + bV = t$. Define p_t as the probability that the subordinator $X^{\mathrm{U}}(\cdot)$ still has value 0 at time $t \geq 0$, i.e., $p_t := \mathbb{P}[X^{\mathrm{U}}(t) = 0]$; realize that $p_t = \lim_{\alpha\to\infty} e^{\varphi^{\mathrm{U}}(\alpha)t}$. Let $\beta(\alpha)$ denote the transform of the storage level as soon as it becomes positive, given that it was still zero at time $\tau(z) + \sigma(z)$, i.e., $\beta(\alpha) := \mathbb{E}[e^{-\alpha B}]$. With $1\{A\}$ the indicator function of the event A, we obtain, noticing that $\vartheta(z) = X^{\mathrm{U}}(t)$ if $X^{\mathrm{U}}(t) > 0$ and B otherwise,

$$\begin{aligned}\mathbb{E}[e^{-\alpha Z(\vartheta(z))}] &= \mathbb{E}[e^{-\alpha X^{\mathrm{U}}(t)} 1\{X^{\mathrm{U}}(t) > 0\}] + p_t\beta(\alpha) \\ &= e^{\varphi^{\mathrm{U}}(\alpha)t} + p_t(\beta(\alpha) - 1).\end{aligned}$$

Now a direct deconditioning argument yields

$$\mathbb{E}[e^{-\alpha Z(\vartheta(z))}] = \mathbb{E}[e^{\varphi^{\mathrm{U}}(\alpha)aI}]\mathbb{E}[e^{\varphi^{\mathrm{U}}(\alpha)bV}] + \mathbb{P}[X^{\mathrm{U}}(aI + bV) = 0]\,(\beta(\alpha) - 1).$$

Here $\mathbb{E}[e^{-\alpha I}]$ can be further evaluated, cf. (2):

$$\mathbb{E}[e^{-\alpha I}] = \mathbb{E}[e^{\varphi^{\mathrm{E}}(\alpha)\tau(z)}] = e^{-\psi^{\mathrm{D}}(-\varphi^{\mathrm{E}}(\alpha))z}.$$

Also, as an immediate consequence of (2),

$$\begin{aligned}\mathbb{P}[X^{\mathrm{U}}(aI + bV) = 0] &= \lim_{\alpha\to\infty} \mathbb{E}[e^{\varphi^{\mathrm{U}}(\alpha)(aI+bV)}] \\ &= \lim_{\alpha\to\infty} e^{-\psi^{\mathrm{D}}(-\varphi^{\mathrm{E}}(-\varphi^{\mathrm{U}}(\alpha)a))z}\mathbb{E}[e^{\varphi^{\mathrm{U}}(\alpha)bV}].\end{aligned}$$

For simplicity we abbreviate $h(\alpha) := \psi^{\mathrm{D}}(-\varphi^{\mathrm{E}}(-\varphi^{\mathrm{U}}(\alpha)a))$ and $g(\alpha) := -\varphi^{\mathrm{U}}(\alpha)b$, and note that both $h(\cdot)$ and $g(\cdot)$ map $[0,\infty)$ on $[0,\infty)$. We obtain for the transform of $Z(\vartheta(z))$ of the storage level at the second embedded epoch, given that the level at the first embedded epoch was z:

$$\mathbb{E}[e^{-\alpha Z(\vartheta(z))}] = e^{-h(\alpha)z}\,\mathbb{E}[e^{-g(\alpha)V}] + e^{-h(\infty)z}\,\mathbb{E}[e^{-g(\infty)V}](\beta(\alpha) - 1). \tag{3}$$

Stability constraint. We now give an intuitive argument that yields the stability condition; we later make this argument rigorous. Let z be the initial storage level. Then it is expected that the buffer is empty at time $-z/\varrho^{\mathrm{D}}$ (recall that ϱ^{D} is the drift of Lévy process X^{D}, i.e., a negative number). Then $\mathbb{E}\sigma(z)$ equals $-za\varrho^{\mathrm{E}}/\varrho^{\mathrm{D}} + b\mathbb{E}[V]$, and

$$\mathbb{E}[Z(\vartheta(z))] = -za\frac{\varrho^{\mathrm{U}}\varrho^{\mathrm{E}}}{\varrho^{\mathrm{D}}} + b\varrho^{\mathrm{U}}\mathbb{E}[V].$$

The stability requirement is that this be smaller than z (for z large). We obtain the condition $a\varrho^{\mathrm{U}}\varrho^{\mathrm{E}} < -\varrho^{\mathrm{D}}$, irrespective of b and $\mathbb{E}[V]$.

Equilibrium distribution. To find the equilibrium distribution of Z at these embedded epochs, Eqn. (3) yields that we have to equate

$$\mathbb{E}[e^{-\alpha Z}] = \mathbb{E}[e^{-h(\alpha)Z}]\mathbb{E}[e^{-g(\alpha)V}] + \mathbb{E}[e^{-h(\infty)Z}]\mathbb{E}[e^{-g(\infty)V}](\beta(\alpha) - 1). \tag{4}$$

This equation can be solved by iteration, yielding the following result.

Theorem 1 *The Laplace-Stieltjes transform of Z at the embedded points equals*

$$\mathbb{E}[e^{-\alpha Z}] = \prod_{m=0}^{\infty} \mathbb{E}[e^{-g(h^{(m)}(\alpha))V}] - \sum_{j=0}^{\infty} \mathbb{E}[e^{-g(\infty)V}]\mathbb{E}[e^{-h(\infty)Z}](1 - \beta(h^{(j)}(\alpha))) \prod_{m=0}^{j-1} \mathbb{E}[e^{-g(h^{(m)}(\alpha))V}], \tag{5}$$

where the empty product is defined as 1*. Here*

$$\mathbb{E}[e^{-h(\infty)Z}] = \frac{\prod_{m=0}^{\infty} \mathbb{E}[e^{-g(h^{(m+1)}(\infty))V}]}{1 + \sum_{j=0}^{\infty} \mathbb{E}[e^{-g(\infty)V}](1 - \beta(h^{(j+1)}(\infty))) \prod_{m=0}^{j-1} \mathbb{E}[e^{-g(h^{(m+1)}(\infty))V}]}.$$

Notice that the expression for $\mathbb{E}[e^{-h(\infty)Z}]$ was found by inserting $\alpha = h(\infty)$ in (5).
Thus we have found an explicit expression for $\mathbb{E}[e^{-\alpha Z}]$, provided that the infinite product converges to a positive value. To find a condition under which this is the case, first observe that $h(\cdot)$ is increasing and concave, as it is a composition of the increasing and concave functions $\psi^{\mathrm{D}}(\cdot)$, $-\varphi^{\mathrm{E}}(\cdot)$ and $-\varphi^{\mathrm{U}}(\cdot)$ that map $[0,\infty)$ on $[0,\infty)$; also $h(0) = 0$. We therefore have that, for all $\alpha \geq 0$,

$$h'(\alpha) \leq h'(0) = \left[(\psi^{\mathrm{D}})'(0)\right] \left[(\varphi^{\mathrm{E}})'(0)\right] \left[(\varphi^{\mathrm{U}})'(0)\right] a.$$

Now suppose that $\gamma := h'(0) < 1$; then Banach's contraction theorem implies that $h^{(m+1)}(\infty) \leq \gamma^m h(\infty)$. Also, for m large,

$$\mathbb{E}[e^{-g(\gamma^m h(\infty))V}] \sim \mathbb{E}[e^{-g'(0)\gamma^m h(\infty)V}] \sim e^{-\varrho^{\mathrm{U}}\gamma^m h(\infty)\mathbb{E}V},$$

and, as $\gamma < 1$,

$$\prod_{m=1}^{\infty} e^{-\varrho^{\mathrm{U}}\gamma^m h(\infty)\mathbb{E}V} > 0.$$

In other words, the infinite product converges if $\gamma < 1$. The equilibrium condition follows by realizing that $\gamma < 1$ reduces to $a\varrho^{\mathrm{U}}\varrho^{\mathrm{E}} < -\varrho^{\mathrm{D}}$, as before.

Mean workload; correlation structure. From (4) we can compute the mean workload at the embedded epochs. Differentiating both sides with respect to α and inserting $\alpha = 0$, we obtain

$$-\mathbb{E}[Z] = -h'(0)\mathbb{E}[Z] - g'(0)\mathbb{E}[V] + \mathbb{E}[e^{-h(\infty)Z}]\mathbb{E}[e^{-g(\infty)V}]\beta'(\alpha),$$

or

$$\begin{aligned}\mathbb{E}[Z] &= (1-h'(0))^{-1}\left(g'(0)\mathbb{E}[V] - \mathbb{E}[e^{-h(\infty)Z}]\mathbb{E}[e^{-g(\infty)V}]\,\beta'(0)\right)\\ &= \left(\frac{\varrho^{\mathrm{D}}}{a\varrho^{\mathrm{U}}\varrho^{\mathrm{E}}+\varrho^{\mathrm{D}}}\right)\left(b\varrho^{\mathrm{U}}\mathbb{E}[V] + \mathbb{E}[e^{-h(\infty)Z}]\mathbb{E}[e^{-g(\infty)V}]\,\mathbb{E}[B]\right).\end{aligned}$$

Also the correlation structure can be characterized. To this end, suppose that Z_n is the storage level at the n-th embedded epoch. Directly from (3),

$$\begin{aligned}\mathbb{E}[e^{-\alpha_0 Z_0 - \alpha_1 Z_1}] &= \mathbb{E}[e^{-(\alpha_0 + h(\alpha_1)Z_0}]\mathbb{E}[e^{-g(\alpha_1)V}]\\ &\quad + \mathbb{E}[e^{-(\alpha_0+h(\infty)Z_0}]\mathbb{E}[e^{-g(\infty)V}](\beta(\alpha_1)-1).\end{aligned}$$

It can be checked that

$$\mathbb{E}[Z_0Z_1] = -\frac{a\varrho^{\mathrm{E}}\varrho^{\mathrm{U}}}{\varrho^{\mathrm{D}}}\mathbb{E}[Z^2] + b\varrho^{\mathrm{U}}\mathbb{E}[Z]\mathbb{E}[V] + \mathbb{E}[Ze^{-h(\infty)Z}]\mathbb{E}[e^{-g(\infty)V}]\mathbb{E}[B];$$

here $\mathbb{E}[Z^2]$ and $\mathbb{E}[Ze^{-h(\infty)Z}]$ can be derived from Thm. 1. We have thus found an explicit expression for $\mathbb{C}\mathrm{ov}(Z_0, Z_1) = \mathbb{E}[Z_0Z_1] - (\mathbb{E}[Z])^2$.

4 Equilibrium distribution of the workload

In the previous section we have analyzed the transform of the equilibrium distribution of the storage level at the start of the active period. The present section translates this into the transform of the equilibrium distribution at an *arbitrary* instant in time. The procedure followed uses a decoupling of the active periods and the interruptions; the desired transform follows by weighing these in an appropriate way.

Active periods. First concentrate on the active periods. Consider the martingale

$$e^{-\alpha X^{\mathrm{D}}(t)} - e^{-\alpha X^{\mathrm{D}}(0)} - \varphi^{\mathrm{D}}(\alpha) \int_0^t e^{-\alpha X^{\mathrm{D}}(s)} \mathrm{d}s,$$

with stopping time $\tau(z)$. Using the fact that $X^{\mathrm{D}}(\cdot)$ has no negative jumps, we derive the identity, by applying 'optional stopping',

$$\mathbb{E}\left[\int_0^{\tau(z)} e^{-\alpha X^{\mathrm{D}}(s)} \mathrm{d}s\right] = \frac{e^{\alpha z} - 1}{\varphi^{\mathrm{D}}(\alpha)}.$$

Recalling that, on $[0, \tau(z)]$, it holds that $z + X^{\mathrm{D}}(s) = Z(s)$, this immediately yields

$$\mathcal{L}(\alpha) := \mathbb{E}\left[\int_0^{\tau(Z)} e^{-\alpha Z(s)} \mathrm{d}s\right] = \frac{1 - \mathbb{E}[e^{-\alpha Z}]}{\varphi^{\mathrm{D}}(\alpha)},$$

where the random variable Z denotes the storage level at the beginning of an active period (such that $\mathbb{E}[e^{-\alpha Z}]$ can be computed as in the previous section). This expression can be interpreted by using the integrated tail distribution Z^{res} of Z, characterized by the transform

$$\mathbb{E}[e^{-\alpha Z^{\mathrm{res}}}] = \frac{1}{\alpha \mathbb{E}[Z]}(1 - \mathbb{E}[e^{-\alpha Z}]).$$

Now $\mathcal{L}(\alpha)$ can be expressed in terms of the distribution of Z^{res}:

$$\mathcal{L}(\alpha) = \frac{\alpha \mathbb{E}[e^{-\alpha Z^{\mathrm{res}}}]}{\varphi^{\mathrm{D}}(\alpha)} \mathbb{E}[Z]. \tag{6}$$

Division by $\mathbb{E}[\tau(Z)] = \mathbb{E}[Z]/(-\varrho^{\mathrm{D}})$ yields an expression for the steady-state workload during active periods,

$$\frac{-\varrho^{\mathrm{D}} \alpha \mathbb{E}[e^{-\alpha Z^{\mathrm{res}}}]}{\varphi^{\mathrm{D}}(\alpha)};$$

notice the similarity with the celebrated Pollaczek-Khinchine formula.

Interruptions. Now concentrate on the service interruptions. The time average distribution during these intervals is characterized by the ratio of

$$\mathbb{E}\left[\int_0^{\sigma(Z)} e^{-\alpha X^{\mathrm{U}}(s)}\mathrm{d}s\, 1\{X^{\mathrm{U}}(\sigma(Z))>0\}\right] + \mathbb{E}\left[\int_0^{\sigma(Z)+L} e^{-\alpha X^{\mathrm{U}}(s)}\mathrm{d}s\, 1\{X^{\mathrm{U}}(\sigma(Z))=0\}\right] \tag{7}$$

and

$$\mathbb{E}\left[\sigma(Z)\, 1\{X^{\mathrm{U}}(\sigma(Z))>0\}\right] + \mathbb{E}\left[(\sigma(Z)+L)\, 1\{X^{\mathrm{U}}(\sigma(Z))=0\}\right]. \tag{8}$$

Let us start by considering the numerator (7), which can be rewritten as

$$\mathcal{N}(\alpha) := \mathbb{E}\left[\int_0^{\sigma(Z)} e^{-\alpha X^{\mathrm{U}}(s)}\mathrm{d}s\right] + \mathbb{E}[p_{\sigma(Z)}]\mathbb{E}[L];$$

use that $X^{\mathrm{U}}(s)=0$ until the end of the period L. One readily verifies that

$$\mathbb{E}\left[\int_0^{\sigma(z)} e^{-\alpha X^{\mathrm{U}}(s)}\mathrm{d}s\right] = \int_0^{\infty}\int_0^{x} e^{\varphi^{\mathrm{U}}(\alpha)s}\mathrm{d}s\mathrm{d}\mathbb{P}(\sigma(z)\le x) = \frac{\mathbb{E}\left[e^{\varphi^{\mathrm{U}}(\alpha)\sigma(z)}\right]-1}{\varphi^{\mathrm{U}}(\alpha)};$$

recall that both numerator and denominator of the last expression are negative (for positive α). We conclude that (7) reduces to

$$\mathcal{N}(\alpha) = \frac{\mathbb{E}[e^{-h(\alpha)Z}]\mathbb{E}[e^{-g(\alpha)V}]-1}{\varphi^{\mathrm{U}}(\alpha)} + \mathbb{E}[e^{-h(\infty)Z}]\mathbb{E}[e^{-g(\infty)V}]\mathbb{E}[L]. \tag{9}$$

Likewise, the denominator (8) equals

$$\begin{aligned}\mathcal{D} &:= \mathbb{E}[\sigma(Z)] + \mathbb{E}[L\, 1\{X^{\mathrm{U}}(\sigma(Z))=0\}]\\ &= -\mathbb{E}[Z]a\frac{\varrho^{\mathrm{E}}}{\varrho^{\mathrm{D}}} + b\mathbb{E}[V] + \mathbb{E}[e^{-h(\infty)Z}]\mathbb{E}[e^{-g(\infty)V}]\mathbb{E}[L].\end{aligned} \tag{10}$$

We obtain the following result.

Theorem 2 *The Laplace-Stieltjes transform of the steady-state storage level W is given by*

$$\mathbb{E}[e^{-\alpha W}] = \frac{\mathcal{L}(\alpha)+\mathcal{N}(\alpha)}{\mathcal{C}+\mathcal{D}},$$

where $\mathcal{C} := \mathbb{E}[\tau(Z)] = \mathbb{E}[Z]/(-\varrho^{\mathrm{D}})$ and where $\mathcal{L}(\alpha)$, $\mathcal{N}(\alpha)$ and $\mathcal{D}$ are given by (6), (9) and (10), respectively.

5 Special cases and ramifications

In this section we consider three special cases of the general model considered so far. Subsequently we discuss two model variants which can also be analysed in detail.

Example 1. The main feature of the general model under consideration is the dependence between the length of a passive period and the length of the preceding active period. This dependence is eliminated by taking $a = 0$, yielding $h(\alpha) \equiv 0$. Consider the classical M/G/1 vacation queue 'with single vacations', viz., an M/G/1 queue in which the server goes on vacation when the system has become idle, and when finding the system empty upon returning, the server waits until the arrival of the first customer.
The following notation is used. The customers arrive according to a Poisson process with rate λ. The amounts of work brought along by the customers are an i.i.d. sequence of random variables (where the size of such a job has transform $\beta(\cdot)$). It is assumed that $\varrho := -\lambda\beta'(0) < 1$.
It is easily seen that this model is a special case of our model; take

$$a = 0, \quad b = 1, \quad \varphi^{\mathrm{U}}(\alpha) = -\lambda(1 - \beta(\alpha)), \quad \varphi^{\mathrm{D}}(\alpha) = -\lambda(1 - \beta(\alpha)) + \alpha.$$

Notice that we could have taken other functions than $\beta(\alpha)$, thus allowing for different service time distributions of customers who arrive during an active period, a passive period and at the end of L. Also, by taking $\lambda^\star$ rather than λ in the definition of $\varphi^{\mathrm{U}}(\cdot)$, we could have allowed for a different arrival rate during passive periods. Shanthikumar [14] and Doshi [7] study vacation models in which the arrival rate changes per period (active/vacation).
Below we specify the Laplace-Stieltjes transform (LST) of Z, from Theorem 1:

$$\mathbb{E}[e^{-\alpha Z}] = \mathbb{E}[e^{-\lambda(1-\beta(\alpha))V}] - (1 - \beta(\alpha))\mathbb{E}[e^{-\lambda V}];$$

notice that in the first product in the right-hand side of (5) only the $(m = 0)$-factor is not equal to 1, and in the sum only the $(j = 0)$-term is not equal to 0. The four elements of the expression for the LST of the storage level W in Thm. 2 (observe that $\varrho^{\mathrm{D}} = \lambda\mathbb{E}[B] - 1 = \varrho - 1$):

$$\mathcal{C} = \mathbb{E}[\tau(Z)] = \frac{\mathbb{E}[Z]}{1 - \varrho}, \qquad \mathcal{D} = \mathbb{E}[V] + \mathbb{E}[e^{-\lambda V}]\frac{1}{\lambda},$$

$$\mathcal{L}(\alpha) = \frac{(1-\varrho)\alpha}{\alpha - \lambda(1-\beta(\alpha))} \mathbb{E}\left[e^{-\alpha Z^{\rm res}}\right] \mathbb{E}[\tau(Z)],$$

$$\mathcal{N}(\alpha) = \frac{1 - \mathbb{E}[e^{-\lambda(1-\beta(\alpha))V}]}{\lambda(1-\beta(\alpha))} + \mathbb{E}[e^{-\lambda V}]\frac{1}{\lambda}.$$

We refer to Takagi [17] for an extensive discussion of the M/G/1 queue with single vacations. The above expression for $\mathcal{L}(\alpha)$ can be found in Eqn. (2.28a) on p. 126 of [17] (Takagi presents the waiting time transform of a customer who arrives during an active period; PASTA implies that this is also the conditional workload transform).

Remark 1 It should be observed that in this model Z can be thought of as the workload that has accumulated during the passive period. The structure of $\mathcal{L}(\alpha)$ reveals a decomposition of the workload during active periods into an M/G/1 workload and an independent additional term. Such decomposition results play a central role in the literature on vacation models. Indeed, Fuhrmann and Cooper [9] prove that, for a large class of M/G/1-type vacation queues, a decomposition of the following type holds: the steady-state queue length in the vacation model equals, in distribution, the sum of two independent quantities, viz., the queue length in the corresponding model without vacations and a term representing the effect of the vacations. Similar decompositions have been obtained for waiting times and workloads. See [2, 3] for such a workload decomposition in single-server queues with multiple customer classes, like polling models.

Example 2. Another extreme case is $b = 0$, $\varphi^{\rm E}(\alpha) \equiv \alpha$. The length of a passive period now equals a times the length of the preceding active period. The fact that $g(\alpha) \equiv 0$ leads to some simplifications in Thms. 1 and 2.

Example 3. Let us consider a two-queue polling model, with exhaustive service at both queues Q_1 and Q_2, and with independent Poisson arrival processes with rates $\lambda_1,\ \lambda_2$ and service time LSTs $\beta_1(\alpha),\ \beta_2(\alpha)$ (with means μ_1 and μ_2, respectively). Compared to the classical 2-queue polling model, cf. [15], we introduce one slight adaptation: When the server leaves a queue and *both* queues are empty, the server waits for the first arrival at Q_1 (instead of waiting for the first arrival at *any* of the queues). One may adapt the definition of the period L from Section 2 to retrieve the classical polling model.

Using the results of the previous two sections, the workload distribution in Q_1 is obtained by making the following choices:

$$a = 1, \quad b = 0, \quad \varphi^{\mathrm{D}}(\alpha) = -\lambda_1(1 - \beta_1(\alpha)) + \alpha,$$
$$\varphi^{\mathrm{U}}(\alpha) = -\lambda_1(1 - \beta_1(\alpha)), \quad \varphi^{\mathrm{E}}(\alpha) = -\lambda_2(1 - \gamma_2(\alpha)).$$

Here $\gamma_2(\alpha)$ is the transform of the busy period distribution in an M/G/1 queue with arrival rate λ_2 and service time transform $\beta_2(\alpha)$. This choice of $\varphi^{\mathrm{E}}(\alpha)$ accomplishes the following: During an active period of Q_1, customers arrive at Q_2 according to a Poisson process with rate λ_2. During the subsequent passive period of Q_1, each of those, say, N_2 customers at Q_2 is served, along with all those arriving in Q_2 during their service time, etc. This amounts to N_2 busy periods at Q_2, reflecting exhaustive service at Q_2.
It is readily checked that the stability constraint $a\varrho^{\mathrm{U}}\varrho^{\mathrm{E}} < -\varrho^{\mathrm{D}}$ of this model translates to

$$\lambda_1\mu_1 \left(\frac{\lambda_2\mu_2}{1 - \lambda_2\mu_2} \right) < 1 - \lambda_1\mu_1.$$

Defining $\varrho_i := \lambda_i\mu_i$, this reduces to the familiar $\varrho_1 + \varrho_2 < 1$.

A crucial feature of the analysis of the previous sections was that $\mathbb{E}[\mathrm{e}^{-\alpha\tau(z)}] = \mathrm{e}^{-f(\alpha)z}$ for some function $f(\cdot)$. This is an important property of Lévy processes, but can also be enforced by appropriate different choices of $\tau(z)$. We consider two such choices.

Ramification 1. Assume that an active period lasts until the workload has been reduced to a fraction c of its value at the beginning of the active period (the situation described in the previous sections corresponds to $c = 0$). So

$$\tau(z) := \inf\{t \geq 0 : z + X^{\mathrm{D}}(t) = cz\}, \;\; 0 \leq c < 1,$$

which is in distribution equal to $\inf\{t \geq 0 : (1 - c)z + X^{\mathrm{D}}(t) = 0\}$. Hence

$$\mathbb{E}[e^{-\alpha\tau(z)}] = e^{-\psi^{\mathrm{D}}(\alpha)(1-c)z}.$$

The case in which $c \in (0, 1)$ is relatively easy, since the system never empties. Then Eqn. (4) is modified into the following form:

$$\mathbb{E}[e^{-\alpha Z}] = \mathbb{E}[e^{-k(\alpha)Z}]\mathbb{E}[e^{-g(\alpha)V}], \tag{11}$$

with $k(\alpha) := \psi^{\mathrm{D}}(-\varphi^{\mathrm{E}}(\varphi^{\mathrm{U}}(-\alpha)a))(1 - c) = (1 - c)h(\alpha)$ and, as before, $g(\alpha) = -\varphi^{\mathrm{U}}(\alpha)b$. The stability condition is easily seen to be $a\varrho^{\mathrm{U}}\varrho^{\mathrm{E}}(1 - c) < -\varrho^{\mathrm{D}}(1 - c)$, which reduces to $a\varrho^{\mathrm{U}}\varrho^{\mathrm{E}} < -\varrho^{\mathrm{D}}$.

It is not surprising that this criterion is independent of c. During the active periods the storage level decreases by an amount $z(1-c)$, if the initial storage level is z. During the vacations, the storage level increases on average by $az(1-c)\varrho^{\mathrm{E}}\varrho^{\mathrm{U}}/(-\varrho^{\mathrm{D}})$. To ensure stability we should have that the average increase is smaller than the decrease $z(1-c)$. We see that the factor $1-c$ cancels.
If the stability condition (which is equivalent with $k'(0) < 1$) holds, then (11) yields the following expression for the steady-state workload Z at the embedded points of beginnings of activity periods:

$$\mathbb{E}[e^{-\alpha Z}] = \prod_{m=0}^{\infty} \mathbb{E}[e^{-g(k^{(m)}(\alpha))V}].$$

Ramification 2. Let us consider a two-queue polling model with gated service at queue Q_1 and exhaustive service at queue Q_2, and with independent Poisson arrival processes with rates λ_1, λ_2 and service time transform $\beta_1(\alpha)$, $\beta_2(\alpha)$. The gated service policy amounts to the following: When the server visits a queue, it serves exactly all the work (customers) present upon arrival, and then moves on to the next queue. Just like in Example 3 above, we introduce one slight adaptation to the classical 2-queue gated/exhaustive polling model: When the server leaves Q_2 and Q_1 is empty, the server waits for the first arrival at Q_1. If we replace the definition of $\tau(z)$ in Section 2 by $\tau(z) := z$, and we choose $\varphi^{\mathrm{D}}(\alpha) := -\lambda_1(1-\beta_1(\alpha)) + \alpha$, then an activity period may be viewed as the visit period of an M/G/1 queue Q_1 with arrival rate λ_1 and service time LST $\beta_1(\alpha)$, operating under the gated service policy. Using the results of the previous two sections, the workload distribution in Q_1 is obtained by making the following choices:

$$a = 1, \quad b = 0, \quad \varphi^{\mathrm{D}}(\alpha) = -\lambda_1(1-\beta_1(\alpha)) + \alpha,$$
$$\varphi^{\mathrm{U}}(\alpha) = -\lambda_1(1-\beta_1(\alpha)), \quad \varphi^{\mathrm{E}}(\alpha) = -\lambda_2(1-\gamma_2(\alpha)).$$

Here $\gamma_2(\alpha)$ is defined as in Example 3.

Remark 2 It should now also be clear how to model polling models with exhaustive service at one queue and gated at another. Furthermore, choosing $b > 0$ allows for switch-over times between queues.

Acknowledgments

Part of this work was done while the third author visited EURANDOM and CWI as a Stieltjes visiting professor. The third author is also supported by grant 964/06 from the Israel Science Foundation. The research of the first and second was done within the framework of the BRICKS project and the European Network of Excellence Euro-NGI.

References

[1] E. Altman (2002). Stochastic recursive equations with applications to queues with dependent vacations. *Annals of Operations Research* **112**, 43-61.

[2] O.J. Boxma and W.P. Groenendijk (1987). Pseudo-conservation laws in cyclic-service systems. *J. Appl. Probab.* **24**, 949-964.

[3] O.J. Boxma (1989). Workloads and waiting times in single-server systems with multiple customer classes. *Queueing Systems* **5**, 185-214.

[4] B.T. Doshi (1986). Queueing systems with vacations – a survey. *Queueing Systems* **1**, 29-66.

[5] B.T. Doshi (1989). Conditional and unconditional distributions for M/G/1 type queues with server vacations. *Queueing Systems* **7**, 229-251.

[6] B.T. Doshi (1990). Single server queues with vacations. In: *Stochastic Analysis of Computer and Communication Systems.* H. Takagi (ed.). North-Holland Publ. Co., Amsterdam, pp. 217-265.

[7] B.T. Doshi (1990). Generalizations of the stochastic decomposition results for single server queues with vacations. *Stochastic Models* **6**, 307-333.

[8] I. Eliazar (2005). Gated polling systems with Lévy inflow and inter-dependent switchover times: A dynamical-systems approach. *Queueing Systems* **49**, 49-72.

[9] S.W. Fuhrmann and R.B. Cooper (1985). Stochastic decompositions in the M/G/1 queue with generalized vacations. *Oper. Res.* **33**, 1117-1129.

[10] R. Groenevelt and E. Altman (2005). Analysis of alternating-priority queueing models with (cross) correlated switchover times. *Queueing Systems* **51**, 199-247.

[11] C.M. Harris and W.G. Marchal (1988). State dependence in M/G/1 server-vacation models. *Oper. Res.* **36**, 560-565.

[12] N.U. Prabhu (1998). *Stochastic Storage Processes.* Springer, New York, 1998.

[13] K. Sato (1999). *Lévy processes and infinitely divisible distributions*, Cambridge University Press, Cambridge.

[14] J.G. Shanthikumar (1988). On stochastic decomposition in M/G/1 type queues with generalized server vacations. *Oper. Res.* **36**, 566-569.

[15] L. Takács (1968). Two queues attended by a single server. *Oper. Res.* **16**, 639-650.

[16] H. Takagi (1990). Queueing analysis of polling models: An update. In: *Stochastic Analysis of Computer and Communication Systems.* H. Takagi (ed.). North-Holland Publ. Co., Amsterdam, pp. 267-318.

[17] H. Takagi (1991). *Queueing Analysis. Vol. 1: Vacation and Priority Systems, Part 1.* North-Holland Publ. Co., Amsterdam.

[18] H. Takagi (1997). Queueing analysis of polling models: Progress in 1990-1994. In: *Frontiers in Queueing.* J.H. Dshalalow (ed.). CRC Press, Boca Raton (Fl.), pp. 119-146.

[19] J. Teghem, Jr. (1986). Control of the service process in a queueing system. *Eur. J. Oper. Res.* **23**, 141-158.