

# A queueing system with vacations after a random amount of work

Ivo Adan<sup>\*</sup>, Onno Boxma<sup>†</sup>, Dieter Claeys<sup>‡</sup>, and Offer Kella<sup>§</sup>

## Abstract

This paper considers an  $M/G/1$  queue with the following vacation discipline. The server takes a vacation as soon as it has served a certain amount of work since the end of the previous vacation. If the system becomes empty before the server has completed this amount of work, then it stays idle until the next customer arrival and then becomes active again. Such a vacation discipline arises, for example, in the maintenance of production systems, where machines or equipment mainly degrade while being operational.

We derive an explicit expression for the distribution of the time it takes until the prespecified amount of work has been served. For the case the total amount of work till vacation is exponentially distributed, we derive the transforms of the steady-state workload at various epochs, busy period, waiting time, sojourn time, and queue length distributions.

## 1 Introduction

Queueing systems with vacations have been studied extensively [7, 20, 21] and have applications in a wide range of areas. Many vacation policies have been studied in literature, such as exhaustive, gated, number-limited and time-limited service. In the exhaustive case, the server initiates a vacation when the system becomes empty, whereas in the gated case, it starts a new vacation when all customers that were present at the end of the previous vacation have been served. In number-limited systems, a vacation is initiated when the server has served a predetermined number of customers *or* when the system becomes empty. Similarly, in time-limited systems [12, 16, 17], a vacation starts when the server has served a predetermined amount of time *or* when the system becomes empty.

In queueing theory, the objective is to characterize performance measures of a system in terms of descriptive parameters. In the context of manufacturing systems, queueing models are often used to predict the work in process (WIP) and lead time in terms of the customer demand (arrival rate), processing speed (service rate), etc. (see e.g. [4, 10, 19]). In this case vacations typically describe sharing of resources (machines, tools, operators, etc.) among production processes. For example, from the point of view of the production process of a specific part type, a vacation of a machine may correspond to, e.g., the processing of another part type. Several production scheduling policies exist and these are often modelled by exhaustive, gated, number-limited or time-limited service.

---

<sup>\*</sup>Department of Industrial Engineering and Innovation Sciences, and Department of Mechanical Engineering, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (i.adan@tue.nl)

<sup>†</sup>EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (o.j.boxma@tue.nl); research done in the framework of the IAP BESTCOM project, funded by the Belgian government

<sup>‡</sup>Department of Industrial Systems Engineering and Product Design, Ghent University, Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium (dieter.claeys@UGent.be); Flanders Make, Oude Diestersebaan 133, 3920 Lommel, Belgium

<sup>§</sup>Department of Statistics, The Hebrew University of Jerusalem, Mount Scopus, Jerusalem 91905, Israel (offer.kella@huji.ac.il); supported in part by grant 1647/17 from the Israel Science Foundation and the Vigevani Chair in Statistics

Vacations can also be used to model maintenance activities in manufacturing systems. During such a vacation, the machine is not working, because it is being maintained, either preventively or correctively. Maintenance activities in manufacturing systems are usually modelled as single vacations after time-limited service: after some time of processing has elapsed, maintenance is carried out, either because the machine has failed or because preventive maintenance is planned (see e.g. [6, 8, 15, 22]). However, time-limited service is not always an adequate model to describe maintenance activities. Stoppages resulting in corrective maintenance may only occur when the machine is running. In addition, preventive maintenance policies can also be usage-based, because machine tools deteriorate while being used and not while being idle. Boring equipment and roller bearings, for instance, degrade while being operational, due to friction forces. Usage-based preventive maintenance policies schedule maintenance after the machine has been used for some time (see e.g. [23, 11, 18, 2]). Such policies keep track of the operational hours, and maintenance is carried as soon as the threshold for the operational hours has been reached.

To the best of our knowledge, no vacation queueing models exist that adequately model maintenance activities that are not triggered by time but by actual usage of a machine. The objective of this paper is to develop such a model. To incorporate this usage-based maintenance feature, we introduce a new vacation queueing system with *modified* time-limited service. We study a system where the server starts a vacation when a predetermined amount of work distributed like  $A$  has been completed since the most recent service initiation. If the system becomes empty beforehand, then no vacation is initiated, that is, the server remains idle and begins serving the next arriving customer as soon as it arrives. Hence, the system alternates between single vacations and visit periods during which an amount of work distributed like  $A$  is completed.

This paper is in some respects a companion paper of [1], where the server takes a vacation if it has served a random number  $N$  of customers, staying idle in the queue when the system becomes empty before  $N$  customers have been served. In [1], the following expressions were established: the joint transform of the length of a visit period and the number of customers in the system at the end of that period, the generating function of the number of customers at a random instant and the Laplace-Stieltjes transform of the delay of a customer. The key idea was to exploit a link with a result from Cohen [5] about the transient behavior of the queue length at customer departure epochs in an ordinary  $M/G/1$  queue and to apply contour integration and the Fuhrmann-Cooper decomposition. In the present paper, however, the server goes on vacation after having served an amount of work. This is better suited for modelling the work in process at a machine that is maintained after having worked a prespecified amount of time.

We consider various performance measures for the vacation model under consideration. We determine the distribution of the total time until a certain amount of work  $x$  has been served, first when starting in an empty system and subsequently when starting from a workload level  $z > 0$  (this total time may also include idle periods). We then use this result to determine the total time until a random amount of work distributed like  $A$  has been served, starting from some workload level  $z$ .

From then on, we focus on the case where  $A \sim \exp(\mu)$ . The memoryless property of the exponential distribution allows us to establish a link between the vacation model under consideration and two other  $M/G/1$  queues (referred to as *Model I* and *Model II*). Model I is an  $M/G/1$  queue with two types of customers; the first type corresponds to the customers in the vacation queue under consideration, and the second type arrives according to a Poisson process with rate  $\mu$  and has service requirements corresponding to the total amount of work that arrives during a vacation of the original system. Model II has the same Poisson arrival process as the vacation model under consideration, but *extended* service requirements: an extended service time consists of an ordinary service time of the vacation model plus all the vacations which interrupt that service time. The

relation with Model I allows us to determine the steady-state workload distribution of the original model. The relation with Model II allows us to determine the steady-state busy period, waiting time, sojourn time and queue length distribution. In the case of workload, we also distinguish several time epochs, deriving the workload transform at the beginning of a vacation, at the end of a vacation, and at an arbitrary epoch.

The paper is organized as follows. In Section 2 we provide a detailed model description. Section 3 focuses on the time it takes to serve a fixed amount of work  $x$ , or a random amount of work distributed like  $A$ . We use some of the ideas of that section in Section 4 to derive an expression for the Laplace-Stieltjes transform of the workload at the beginning of a vacation. Triggered by the remarkable and quite suggestive form of this transform, we relate our queuing model to Model I mentioned above, with two types of customers. The steady-state workload at an arbitrary epoch can also be obtained via this analogy, but this requires a less straightforward approach; it is presented in Section 5. In Section 6 we outline the relation to Model II and exploit it to derive exact expressions for the transforms of busy period, waiting time, sojourn time and queue length. Finally, Section 7 concludes the paper and sketches a framework for formulating optimization problems.

## 2 Model description

We consider an  $M/G/1$  queue with arrival rate  $\lambda$  and service requirements  $B_1, B_2, \dots$  which are independent and identically distributed (i.i.d.). The special feature of the model is its vacation mechanism. The server takes its  $i$ th vacation as soon as it has served (i.e., has been active) exactly  $A_i$  amount of work since the end of the previous vacation. These active periods  $A_i$ ,  $i = 1, 2, \dots$ , are i.i.d. We also assume successive vacation lengths  $V_1, V_2, \dots$  to be i.i.d. If the system becomes empty between two successive vacations, the server stays in the system, remaining idle until another customer arrives. Furthermore, when the server returns from a vacation and finds the system empty, it also waits until a customer arrives. The server hence alternates between vacations and periods which we call *visit* periods, during which it serves a random amount of work. We assume independence between interarrival times, service times, active periods and vacation times. Finally,  $A, B, V$  denote generic active periods, service times and vacations, and the distribution of a random variable  $X$  will be denoted by  $F_X(\cdot)$ , while  $\bar{F}_X := 1 - F_X$ . The behaviour of the system is illustrated in Figure 1.

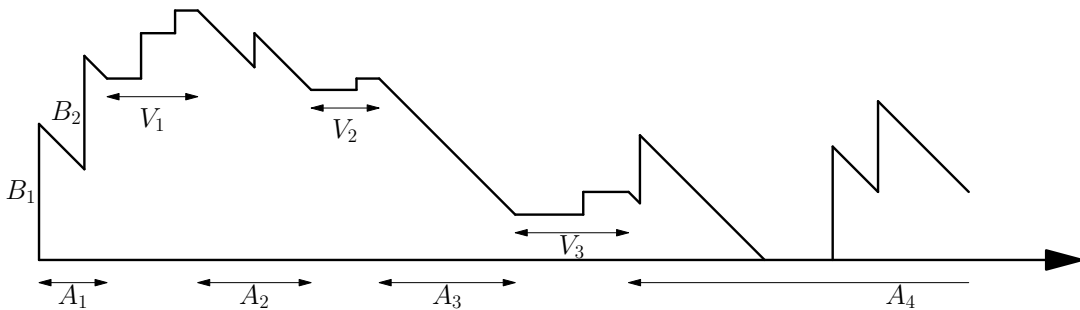


Figure 1: Illustration of the behaviour of the system.

### 3 Time until the cumulative busy time reaches some level

We start by assuming that the system is initially empty (this will be relaxed soon). Let  $\theta, \theta_1, \theta_2, \dots$  be i.i.d. random variables distributed like a busy period in an  $M/G/1$  queue and let  $N_\theta(\cdot)$  be the associated renewal counting process. Then  $N_\theta(x)$  is distributed like the number of busy periods completed until the server has been busy  $x$  time units. If  $e, e_1, e_2, \dots$  are i.i.d. with  $e \sim \exp(\lambda)$ , independent of the busy period and  $E_n := \sum_{i=1}^n e_i$  (with  $E_0 = 0$ ) then, as initially there is an idle time (since the system starts empty), the total idle time until the server was busy for  $x$  units of time is distributed like  $E_{N_\theta(x)+1}$  and thus has the transform

$$\mathbb{E}e^{-\alpha E_{N_\theta(x)+1}} = \mathbb{E} \left( \frac{\lambda}{\lambda + \alpha} \right)^{N_\theta(x)+1}. \quad (1)$$

Since  $P(N_\theta(x) = n) = F_\theta^{*n}(x) - F_\theta^{*(n+1)}(x)$ , where  $F_\theta(\cdot)$  has the distribution of  $\theta$  and  $F_\theta^{*n}(\cdot)$  is its  $n$ th-fold convolution, it is easy to check that for any  $0 < u < 1$  we have that (as for any renewal counting process),

$$\mathbb{E}u^{N_\theta(x)+1} = \sum_{n=0}^{\infty} u^{n+1} (F_\theta^{*n}(x) - F_\theta^{*(n+1)}(x)) = 1 - \sum_{n=0}^{\infty} (1-u)u^n F_\theta^{*n}(x).$$

It is interesting to note that (again, as for any renewal process) if  $\Theta_n := \sum_{i=1}^n \theta_i$  (with  $\Theta_0 = 0$ ) and  $\nu(u) + 1 \sim \text{Geometric}(1-u)$  and is independent of everything else, then the right side can be written as  $P(\Theta_{\nu(u)} > x)$ .

If  $A \sim \exp(\beta)$  and is independent of everything else, then clearly  $\mathbb{E}F_\theta^{*n}(A) = P(\Theta_n \leq A) = \mathbb{E}e^{-\beta\Theta_n} = (\mathbb{E}e^{-\beta\theta})^n$  so that

$$\mathbb{E}u^{N_\theta(A)+1} = 1 - \frac{1-u}{1-u\mathbb{E}e^{-\beta\theta}}.$$

When we insert  $u = \frac{\lambda}{\lambda+\alpha}$  we have, due to (1), that

$$\mathbb{E}e^{-\alpha E_{N_\theta(A)+1}} = \frac{\lambda(1 - \mathbb{E}e^{-\beta\theta})}{\lambda(1 - \mathbb{E}e^{-\beta\theta}) + \alpha}, \quad (2)$$

so that  $E_{N_\theta(A)+1} \sim \exp(\lambda(1 - \mathbb{E}e^{-\beta\theta}))$ . If this is all we wanted to obtain then we could have deduced this from the following basic argument. Due to the memoryless property  $N_\theta(A) + 1 \sim \text{Geometric}(P(A \leq \theta))$  and thus  $E_{N_\theta(A)+1} \sim \exp(\lambda P(A \leq \theta))$  where we finally note that  $P(A \leq \theta) = 1 - \mathbb{E}e^{-\beta\theta}$ . Note that we now have an explicit formula for the double transform

$$\int_0^\infty e^{-\beta x - \alpha E_{N_\theta(x)+1}} dx = \frac{\mathbb{E}e^{-\alpha E_{N_\theta(A)+1}}}{\beta}.$$

If the initial workload is some  $z > 0$  rather than zero, then there is no initial idle time and we replace  $N_\theta(\cdot)$  by a delayed renewal counting process  $N_\theta^z(x)$  where  $\theta_1$  is distributed like  $\theta^z$ : the time until the system empties and  $\theta_2, \theta_3, \dots$  are distributed like  $\theta$ . It is well known that

$$\mathbb{E}e^{-\alpha\theta^z} = e^{-(\alpha + \lambda(1 - \mathbb{E}e^{-\alpha\theta}))z},$$

and in particular if  $B$  is distributed like the service time and is independent of everything else, then  $\theta^B \sim \theta$ . Here the total idle time until the server has been busy for  $x$  time units is distributed like  $E_{N_\theta^z(x)}$  and similar arguments lead to

$$\mathbb{E}u^{E_{N_\theta^z(x)}} = 1 - \sum_{n=0}^{\infty} (1-u)u^n F_\theta^{*n} * F_{\theta^z}(x).$$

As for the case where  $z = 0$ , we can either directly or by applying the memoryless property obtain that when  $A \sim \exp(\beta)$  then

$$\mathbb{E}e^{-\alpha E_{N_\theta^z(A)}} = 1 - \mathbb{E}e^{-\beta\theta^z} + \mathbb{E}e^{-\beta\theta^z} \frac{\lambda(1 - \mathbb{E}e^{-\beta\theta})}{\lambda(1 - \mathbb{E}e^{-\beta\theta}) + \alpha}. \quad (3)$$

That is, with probability  $P(\theta^z \geq A) = 1 - \mathbb{E}e^{-\beta\theta^z}$  there are no idle times and hence the conditional transform is 1 and with probability  $P(\theta^z < A) = \mathbb{E}e^{-\beta\theta^z}$  the conditional transform of the total idle time is like the one starting from an empty system. Note that when  $z = 0$  then (3) reduces to (2) and that also here we have an explicit expression for the double transform

$$\int_0^\infty e^{-\beta x - \alpha E_{N_\theta^z(x)}} dx = \frac{\mathbb{E}e^{-\alpha E_{N_\theta^z(A)}}}{\beta}.$$

Since the time until the server is busy for  $x$  time units is distributed like  $\tau_x^z = x + E_{N_\theta^z(x)+1}$ , we can insert  $\alpha + \beta$  instead of  $\beta$  to obtain the double transform of this time resulting in

$$\int_0^\infty e^{-\beta x} \mathbb{E}e^{-\alpha\tau_x^z} dx = \frac{1}{\alpha + \beta} \left( 1 - \mathbb{E}e^{-(\beta+\alpha)\theta^z} + \mathbb{E}e^{-(\beta+\alpha)\theta^z} \frac{\lambda(1 - \mathbb{E}e^{-(\beta+\alpha)\theta})}{\lambda(1 - \mathbb{E}e^{-(\beta+\alpha)\theta}) + \alpha} \right), \quad (4)$$

or equivalently, if  $A \sim \exp(\beta)$  then

$$\mathbb{E}e^{-\alpha\tau_A^z} = \frac{\beta}{\alpha + \beta} \left( 1 - \mathbb{E}e^{-(\beta+\alpha)\theta^z} + \mathbb{E}e^{-(\beta+\alpha)\theta^z} \frac{\lambda(1 - \mathbb{E}e^{-(\beta+\alpha)\theta})}{\lambda(1 - \mathbb{E}e^{-(\beta+\alpha)\theta}) + \alpha} \right). \quad (5)$$

**Remark 1.** If  $A \sim H(p_1, \dots, p_K, \mu_1, \dots, \mu_K)$  (hyper-exponential), then one can take a weighted sum of terms as appearing in (5), with  $\beta$  replaced by  $\mu_i$ . If  $A \sim \text{Erlang}(n, \mu)$ , then

$$\mathbb{E}e^{-\alpha\tau_A^z} = \frac{\mu^n}{(n-1)!} \int_0^\infty x^{n-1} e^{-\mu x} \mathbb{E}e^{-\alpha\tau_x^z} dx = \frac{(-1)^{n-1}}{(n-1)!} \mu^n \frac{d^{n-1}}{d\mu^{n-1}} \int_0^\infty e^{-\mu x} \mathbb{E}e^{-\alpha\tau_x^z} dx$$

and in principle one can obtain an expression for  $\mathbb{E}e^{-\alpha\tau_A^z}$  by differentiating (4).

## 4 The workload at the end of a visit period

In this section we consider  $Z_+$ , the workload at the end of a visit period, that is the workload at the beginning of a vacation. We assume from now on that  $A \sim \exp(\mu)$ .

We first consider the case where the visit starts at workload level  $z$ . Let  $W_t^z$  be the workload at time  $t$  when the process starts from  $z$  at time 0 and  $X_t := \sum_{i=1}^{N(t)} B_i - t$  is the net input process. Recall the definition of  $\tau_x^z$  and  $\theta^z$  from Section 3. Then  $\tau_A^z$  is the length of the first visit.

If  $A > \theta^z$ , then at time  $\theta^z$  the workload hits zero and, after an exponentially distributed idle time, jumps by a random amount distributed like  $B$ . Due to the memoryless property of  $A$  and the Markov property of  $W_t^z$ , the conditional distribution of  $W_{\tau_A^z}$  given that  $A > \theta^z$  is the same as the distribution of  $Y^B = W_{\tau_A^B}^B$ .

If  $A \leq \theta^z$ , then the conditional distribution of  $W_{\tau_A^z}$  is the same as the conditional distribution of  $z + X_A$  given that  $A \leq \theta^z$ . Let  $Y^z$  have this distribution. Then,

$$W_{\tau_A^z}^z \sim (1 - I)Y^z + IW_{\theta_A^B}^B, \quad (6)$$

where  $I, Y^z, Y^B$  are independent random variables with  $I \sim 1_{\{A > \theta^z\}}$ . By substituting  $z = B$  in (6), it is an easy exercise to show that  $W_{\theta_A^B}^B$  has the conditional distribution of  $B + X_A$  given that  $A \leq \theta = \theta^B$ . Namely it is distributed like  $Y^B$  and (6) becomes

$$W_{\tau_A^z}^z \sim (1 - I)Y^z + IY^B .$$

Thus, we wish to identify

$$\mathbb{E}[e^{-\alpha Y^z}] = \mathbb{E}[e^{-\alpha(z+X_A)} | A \leq \theta^z] .$$

First, note that

$$\mathbb{P}(I = 1) = \mathbb{P}(A > \theta^z) = \mathbb{E}e^{-\mu\theta^z} = e^{-(\mu+\lambda(1-\mathbb{E}e^{-\mu\theta}))z} .$$

Denote

$$\varphi(\alpha) = \log \mathbb{E}e^{-\alpha X_1} = \alpha - \lambda(1 - \mathbb{E}e^{-\alpha B}) ,$$

then, it is well known and easy to check that  $\varphi(\mu + \lambda(1 - \mathbb{E}e^{-\mu\theta})) = \mu$ , so that  $\mu + \lambda(1 - \mathbb{E}e^{-\mu\theta}) = \varphi^{-1}(\mu)$  and in particular, that  $\mathbb{E}e^{-\mu\theta} = \mathbb{E}e^{-\varphi^{-1}(\mu)B}$ . Moreover,

$$\mathbb{E}e^{-\alpha(z+X_A)} = e^{-\alpha z} \mathbb{E}e^{\varphi(\alpha)A} = e^{-\alpha z} \frac{\mu}{\mu - \varphi(\alpha)} .$$

Due to the memoryless property of  $A$  and the Markov property of  $X$ ,

$$\mathbb{E} \left[ e^{-\alpha(z+X_A)} | A > \theta^z \right] = \mathbb{E}e^{-\alpha X_A} = \frac{\mu}{\mu - \varphi(\alpha)} .$$

Clearly,

$$\begin{aligned} \mathbb{E}e^{-\alpha(z+X_A)} &= \mathbb{P}(I = 0) \mathbb{E} \left[ e^{-\alpha(z+X_A)} | A \leq \theta^z \right] \\ &\quad + \mathbb{P}(I = 1) \mathbb{E} \left[ e^{-\alpha(z+X_A)} | A > \theta^z \right] , \end{aligned}$$

and thus

$$\mathbb{E}e^{-\alpha Y^z} = \mathbb{E} \left[ e^{-\alpha(z+X_A)} | A \leq \theta^z \right] = \frac{e^{-\alpha z} - e^{-\varphi^{-1}(\mu)z}}{1 - e^{-\varphi^{-1}(\mu)z}} \frac{\mu}{\mu - \varphi(\alpha)} ,$$

and similarly

$$\mathbb{E}e^{-\alpha Y^B} = \mathbb{E} \left[ e^{-\alpha(B+X_A)} | A \leq \theta \right] = \frac{\mathbb{E}e^{-\alpha B} - \mathbb{E}e^{-\varphi^{-1}(\mu)B}}{1 - \mathbb{E}e^{-\varphi^{-1}(\mu)B}} \frac{\mu}{\mu - \varphi(\alpha)} ,$$

for  $0 \leq \alpha < \varphi^{-1}(\mu)$ . Thus, using (6),

$$\begin{aligned} \mathbb{E}e^{-\alpha W_{\tau_A^z}^z} &= \left( e^{-\alpha z} - e^{-\varphi^{-1}(\mu)z} \right) \frac{\mu}{\mu - \varphi(\alpha)} \\ &\quad + e^{-\varphi^{-1}(\mu)z} \frac{\mathbb{E}e^{-\alpha B} - \mathbb{E}e^{-\varphi^{-1}(\mu)B}}{1 - \mathbb{E}e^{-\varphi^{-1}(\mu)B}} \frac{\mu}{\mu - \varphi(\alpha)} \\ &= \left( e^{-\alpha z} - e^{-\varphi^{-1}(\mu)z} \frac{1 - \mathbb{E}e^{-\alpha B}}{1 - \mathbb{E}e^{-\varphi^{-1}(\mu)B}} \right) \frac{\mu}{\mu - \varphi(\alpha)} \\ &= \left( e^{-\alpha z} - e^{-\varphi^{-1}(\mu)z} \frac{\alpha}{\varphi^{-1}(\mu)} \frac{\mathbb{E}e^{-\alpha B_e}}{\mathbb{E}e^{-\varphi^{-1}(\mu)B_e}} \right) \frac{\mu}{\mu - \varphi(\alpha)} , \end{aligned} \tag{7}$$

where  $B_e$  has the stationary residual life distribution associated with  $B$ , that is, with density  $(1 - F_B(\cdot))/EB$ .

We have thus obtained the LST of the workload at the end of a visit period, when starting that visit period with an amount of work  $z$ . By adding the amount of work that enters in the subsequent vacation, we express the LST of the workload at the start of a visit period in terms of the workload at the start of the previous visit period. We can thus derive the steady-state workload distribution in our  $M/G/1$  queue with vacations, at the beginning of an arbitrary vacation (i.e., the end of a visit period), and then also at the end of a vacation (i.e., the beginning of a visit period). Let  $Z_+$  be the steady-state workload at the beginning of an arbitrary vacation, and  $Z_-$  the steady-state workload at the end of an arbitrary vacation. We make the following two observations. First,

$$Z_- \sim Z_+ + U ,$$

where  $Z_+, U$  are independent and

$$U = \sum_{i=1}^{N_\lambda(V)} B_i \quad (8)$$

denotes the amount of work that enters in an arbitrary vacation, with  $N_\lambda(V)$  being the number of (Poisson with rate  $\lambda$ ) arrivals during a vacation  $V$ , with an empty sum being equal to zero. Therefore,

$$\mathbb{E}e^{-\alpha Z_-} = \mathbb{E}e^{-\alpha Z_+} \mathbb{E}e^{-\lambda(1-\mathbb{E}e^{-\alpha B})V} . \quad (9)$$

Second, in steady state the workloads at two successive vacation beginnings should have the same distribution:

$$\mathbb{E}e^{-\alpha Z_+} = \mathbb{E} \exp \left( -\alpha W_{\tau_A}^{Z_-} \right) .$$

In combination with (7) and (9) this yields a relation involving  $\mathbb{E}e^{-\alpha Z_+}$  on both sides:

$$\mathbb{E}e^{-\alpha Z_+} = \left( \mathbb{E}e^{-\lambda(1-\mathbb{E}e^{-\alpha B})V} \mathbb{E}e^{-\alpha Z_+} - \mathbb{E}e^{-\lambda(1-\mathbb{E}e^{-\varphi^{-1}(\mu)B})V} \mathbb{E}e^{-\varphi^{-1}(\mu)Z_+} \frac{\alpha}{\varphi^{-1}(\mu)} \frac{\mathbb{E}e^{-\alpha B_e}}{\mathbb{E}e^{-\varphi^{-1}(\mu)B_e}} \right) \frac{\mu}{\mu - \varphi(\alpha)} ,$$

and thus

$$\begin{aligned} & \mathbb{E}e^{-\alpha Z_+} \left[ 1 - \frac{\mu}{\mu - \varphi(\alpha)} \mathbb{E}e^{-\lambda(1-\mathbb{E}e^{-\alpha B})V} \right] \\ &= -\mathbb{E}e^{-\lambda(1-\mathbb{E}e^{-\varphi^{-1}(\mu)B})V} \mathbb{E}e^{-\varphi^{-1}(\mu)Z_+} \frac{\mu}{\mu - \varphi(\alpha)} \frac{\alpha}{\varphi^{-1}(\mu)} \frac{\mathbb{E}e^{-\alpha B_e}}{\mathbb{E}e^{-\varphi^{-1}(\mu)B_e}} . \end{aligned}$$

Hence, with  $C$  some constant,

$$\mathbb{E}e^{-\alpha Z_+} = C \frac{\alpha}{\varphi(\alpha) - \mu(1 - \mathbb{E}e^{-\lambda(1-\mathbb{E}e^{-\alpha B})V})} \mathbb{E}e^{-\alpha B_e} .$$

Letting  $\alpha \downarrow 0$  determines the normalizing constant:  $C = \varphi'(0) + \mu \mathbb{E}V \lambda \mathbb{E}B$ , so

$$\mathbb{E}e^{-\alpha Z_+} = \frac{(1 - \rho(1 + \mu \mathbb{E}V))\alpha}{\varphi(\alpha) - \mu(1 - \mathbb{E}e^{-\lambda(1-\mathbb{E}e^{-\alpha B})V})} \mathbb{E}e^{-\alpha B_e} . \quad (10)$$

Finally the LST of the steady-state workload at the end of a vacation,  $Z_-$ , follows from (10) and (9).

The form of the LST of  $Z_+$  in (10) is quite interesting. It is the product of two LSTs of positive random variables. The second one obviously is a residual service time  $B_e$ . To analyze the first LST, we rewrite it as follows, introducing  $\rho_u := \mu \mathbb{E}U = \mu \lambda \mathbb{E}B \mathbb{E}V$ :

$$\frac{(1 - \rho(1 + \mu \mathbb{E}V))\alpha}{\varphi(\alpha) - \mu(1 - \mathbb{E}e^{-\lambda(1 - \mathbb{E}e^{-\alpha B})V})} = \frac{(1 - \rho - \rho_u)\alpha}{\alpha - \lambda - \mu + (\lambda + \mu)\left(\frac{\lambda}{\lambda + \mu} \mathbb{E}e^{-\alpha B} + \frac{\mu}{\lambda + \mu} \mathbb{E}e^{-\lambda(1 - \mathbb{E}e^{-\alpha B})V}\right)} .$$

This is the Pollaczek-Khintchine formula for the LST of the waiting time, or workload, in an ordinary  $M/G/1$  queue where the input is the sum of two compound Poisson processes. The first is the original compound Poisson process with arrival rate  $\lambda$  and jumps distributed like  $B$ . The second has arrival rate  $\mu$  and jumps distributed like  $U$  (cf. (8)). Put differently, this is the input process of some  $M/G/1$  queue with arrival rate  $\lambda + \mu$  and service time distribution

$$\frac{\lambda}{\lambda + \mu} F_B(\cdot) + \frac{\mu}{\lambda + \mu} F_U(\cdot) .$$

In Section 1 we referred to this  $M/G/1$  queue as Model I. In conclusion, we have

$$Z_+ \sim Z_I + B_e ,$$

with  $Z_I$  the steady-state workload in model I.

## 5 The workload at an arbitrary epoch

In this section we consider the steady-state workload, at an *arbitrary* epoch, for the  $M/G/1$  queue with vacations and  $\exp(\mu)$ -distributed active periods. Our approach will be to take a weighted average of the workloads during vacations, busy periods and idle periods (the workload then trivially is zero).

(i) *The workload process  $Z_v(t)$  restricted to vacations*

It can be readily observed that the steady-state workload  $Z_v$  restricted to vacations is distributed like

$$Z_v \sim Z_+ + U_e , \tag{11}$$

with  $U_e$  the amount of work that has arrived during the elapsed vacation time, that is

$$U_e := \sum_{i=1}^{N_\lambda(V_e)} B_i ,$$

with  $V_e$  the elapsed vacation time. The LST of  $U_e$  reads (see e.g. Theorem 3.2 of [13]):

$$\mathbb{E}e^{-\alpha U_e} = \frac{1 - \mathbb{E}e^{-\lambda(1 - \mathbb{E}e^{-\alpha B})V}}{\lambda(1 - \mathbb{E}e^{-\alpha B})\mathbb{E}V} = \mathbb{E}e^{-\lambda(1 - \mathbb{E}e^{-\alpha B})V_e} . \tag{12}$$

The combination of (11) and (12) yields

$$\mathbb{E}e^{-\alpha Z_v} = \mathbb{E}e^{-\alpha Z_+} \frac{1 - \mathbb{E}e^{-\lambda(1 - \mathbb{E}e^{-\alpha B})V}}{\lambda(1 - \mathbb{E}e^{-\alpha B})\mathbb{E}V} . \tag{13}$$



(ii) *The workload process  $Z_b(t)$  restricted to busy periods*

As during busy periods vacations are generated according to a Poisson process with rate  $\mu$ , the PASTA property states that the workload at the start of a vacation has the same distribution as the workload at an arbitrary epoch of a busy period:

$$Z_b \sim Z_+ . \quad (14)$$

**Remark 2.** *The distribution of  $Z_b$  can also be established alternatively by linking  $Z_b(t)$  to the virtual waiting time in model I. This is elaborated upon in Appendix A.*

The stability condition for this  $M/G/1$  queue without vacations and idle periods is  $\lambda\mathbb{E}B + \mu\mathbb{E}U = \rho + \rho_u < 1$ , as this is the condition that the workload hits zero after a finite expected amount of time starting from any initial level having a finite expectation (this is in fact true in general for Lévy processes with no negative jumps and a negative mean). Therefore, this is also the stability condition for the  $M/G/1$  queue under consideration in this paper, viz., the  $M/G/1$  queue with vacations and  $\exp(\mu)$  active periods.

(iii) *The workload process restricted to idle periods*

During idle periods the workload is zero.

In order to identify the complete stationary distribution  $Z$  we need to compute the fractions of time the process spends in each state (busy/vacation/idle) and take the corresponding mixture. First, a standard balancing argument gives

$$\rho = \mathbb{P}[\text{busy}] = \mathbb{P}[\text{busy}|\text{not idle}] (1 - \mathbb{P}[\text{idle}]) . \quad (15)$$

The process restricted to not being idle alternates between busy (mean length  $1/\mu$ ) and vacation (mean length  $\mathbb{E}[V]$ ). Hence,

$$\mathbb{P}[\text{busy}|\text{not idle}] = \frac{\frac{1}{\mu}}{\frac{1}{\mu} + \mathbb{E}V} = \frac{1}{1 + \mu\mathbb{E}[V]} . \quad (16)$$

Combining (15) and (16) yields

$$\begin{aligned} \mathbb{P}[\text{idle}] &= 1 - \rho(1 + \mu\mathbb{E}[V]) \\ &= 1 - \rho - \mu\lambda\mathbb{E}[B]\mathbb{E}[V] \\ &= 1 - \rho - \mu\mathbb{E}[U] \\ &= 1 - \rho - \rho_u . \end{aligned} \quad (17)$$

Finally,

$$\mathbb{P}[\text{vacation}] = 1 - (1 - \rho - \rho_u) - \rho = \rho_u . \quad (18)$$

Combining (15), (17), (18), (13), (14) and (10) we find after some straightforward manipulations:

**Theorem 1.** *The steady-state workload  $Z$  in the  $M/G/1$  queue with vacations and  $\exp(\mu)$ -distributed active periods has the following LST:*

$$\begin{aligned} \mathbb{E}e^{-\alpha Z} &= (1 - \rho - \rho_u) \left[ 1 + \frac{\rho\mathbb{E}e^{-\alpha B_e} + \rho_u\mathbb{E}e^{-\alpha(B_e+U_e)}}{1 - (\rho + \rho_u)\mathbb{E}e^{-\alpha G_e}} \right] \\ &= \frac{1 - \rho - \rho_u}{1 - (\rho + \rho_u)\mathbb{E}e^{-\alpha G_e}} \left[ 1 - \rho_u\mathbb{E}e^{-\alpha U_e}(1 - \mathbb{E}e^{-\alpha B_e}) \right] , \end{aligned}$$

where,

$$\mathbb{E}e^{-\alpha G_e} := \frac{1 - \mathbb{E}e^{-\alpha G}}{\alpha \mathbb{E}G} = \frac{\rho \mathbb{E}e^{-\alpha B_e} + \rho_u \mathbb{E}e^{-\alpha U_e}}{\rho + \rho_u} ,$$

with

$$\mathbb{E}Z = \frac{\rho + \rho_u}{1 - \rho - \rho_u} \mathbb{E}G_e + \rho_u \mathbb{E}B_e ,$$

and

$$\mathbb{E}G_e = \frac{\rho}{\rho + \rho_u} \mathbb{E}B_e + \frac{\rho_u}{\rho + \rho_u} \lambda \mathbb{E}B \mathbb{E}V_e = \frac{\rho \mathbb{E}B_e + \rho \rho_u \mathbb{E}V_e}{\rho + \rho_u} .$$

## 6 Waiting time, queue length and busy period

In this section we shall derive the (transform of the) steady-state busy period, waiting time, sojourn time and queue length in the  $M/G/1$  queue with  $\exp(\mu)$  active periods and vacations. In doing this, we rely on a relation between that  $M/G/1$  queue and the following  $M/G/1$  queue *without* vacations, but with extended service times; we refer to the latter system as Model II (cf. Section 1) or as the *extended system*. We should add that the relation crucially depends on the fact that vacations (ends of active periods) occur according to a Poisson process.

The extended system is an  $M/G/1$  queue with arrival rate  $\lambda$  and with service times

$$B_{ext} \sim B + \sum_{i=1}^{N_\mu(B)} V_i .$$

The interpretation is the following. During the service time  $B$  of a customer, vacations arrive according to a Poisson process with rate  $\mu$ . The  $i$ th such vacation extends  $B$  by its length  $V_i$ . All customer interarrival times, service times,  $\exp(\mu)$  interarrival times of vacations and all vacation times are assumed to be independent, and we have the usual i.i.d. assumptions. It is easily verified that

$$\mathbb{E}e^{-\alpha B_{ext}} = \mathbb{E}e^{-(\alpha + \mu(1 - \mathbb{E}e^{-\alpha V}))B} ,$$

and

$$\mathbb{E}B_{ext} = \mathbb{E}B(1 + \mu \mathbb{E}V) ,$$

yielding a total load  $\lambda \mathbb{E}B_{ext} = \rho + \rho_u$  in the extended system.

Let us now couple both  $M/G/1$  queues, in the sense that arrivals occur at identical moments in both systems, and the service requirement  $B_{ext,i}$  of the  $i$ th customer in the extended system is chosen exactly equal to the sum of the service requirement  $B_i$  in the other system, *plus* the lengths of all vacations that occur during its service. Figure 2 presents a realization of the workload in the original  $M/G/1$  queue with vacations and in the extended system. It reveals something that, after some thinking, appears to be obvious: When defining a busy period in both systems as the time from the arrival of a customer into an empty system until the first departure thereafter of a customer who leaves the system behind empty, the busy periods of both systems have identical lengths. In fact, each arriving customer spends exactly the same time in both systems. Indeed, the arrivals in both systems coincide, but also the time that a customer in the extended system spends in service is identical to the time its counterpart in the original system spends either in service or in an interruption mode in which the server has taken a vacation. Also, the time a

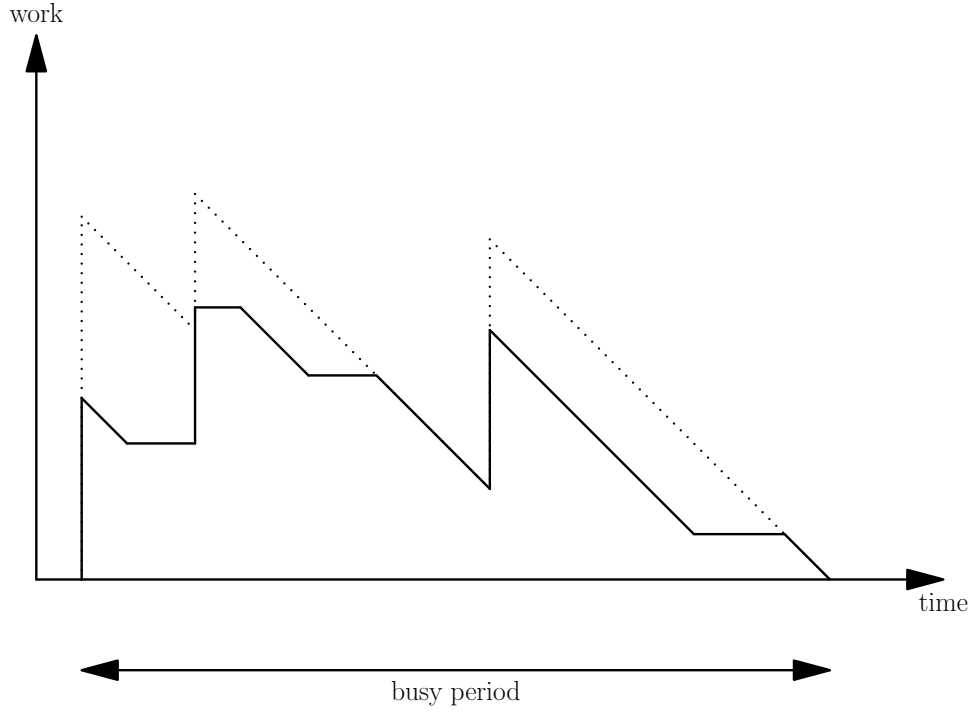


Figure 2: Work in the vacation system (solid line) and work in the  $M/G/1$  system with extended service times (dotted lines): the lengths of the busy periods are equal.

customer in the extended system spends waiting until its service begins is identical to the time a customer in the original system spends waiting until its service begins *for the first time*, because those times are determined by the arrival intervals (which are the same in both systems) and the previous service times (in the extended system), respectively, the service plus interruption times (in the original system). Similarly, the sojourn times in both systems coincide. Finally, since not only the arrival times but also the departure times of any customer in both systems agree, also the system content (those in service included) distributions in both systems are identical. Relying on well-known  $M/G/1$  results as can be found, e.g., in Chapter II.4 of [5], we may now conclude the following.

**Theorem 2.** *Let  $\Theta$ ,  $M$ ,  $W$ ,  $D$  and  $H$  denote, respectively, the steady-state busy period, number served in a busy period, waiting time, sojourn time (service time included) and system content in the  $M/G/1$  queue with  $\exp(\mu)$ -distributed active periods and vacations. Then  $\mathbb{E}z^M e^{-\alpha\Theta}$  is the unique solution in the unit circle  $|x| \leq 1$  of the equation*

$$x = z\mathbb{E}e^{-(\alpha+\lambda(1-x))B} ;$$

$$\mathbb{E}e^{-\alpha W} = \frac{(1-\rho-\rho_u)\alpha}{\alpha-\lambda(1-\mathbb{E}e^{-\alpha B_{ext}})} ; \quad (19)$$

$$\mathbb{E}e^{-\alpha D} = \mathbb{E}e^{-\alpha W} \mathbb{E}e^{-\alpha B_{ext}} ; \quad (20)$$

$$\mathbb{E}z^H = \frac{(1-\rho-\rho_u)(1-z)\mathbb{E}e^{-\lambda(1-z)B_{ext}}}{\mathbb{E}e^{-\lambda(1-z)B_{ext}} - z} . \quad (21)$$

**Remark 3.** *It should be observed that the relation between the original system and Model II does not allow us to derive the workload distribution in the original system. For that purpose, Model I is the right choice.*

**Remark 4.** *Equation (21) can be rewritten as*

$$\mathbb{E}z^H = \chi(z) \frac{(1 - \rho) \mathbb{E}e^{-\lambda(1-z)B} (1 - z)}{\mathbb{E}e^{-\lambda(1-z)B} - z}, \quad (22)$$

where

$$\chi(z) = \frac{(1 - \rho - \rho_u) (\mathbb{E}e^{-\lambda(1-z)B} - z) \mathbb{E}e^{-\lambda(1-z)B_{ext}}}{(1 - \rho) (\mathbb{E}e^{-\lambda(1-z)B_{ext}} - z) \mathbb{E}e^{-\lambda(1-z)B}}.$$

By virtue of Fuhrmann-Cooper decomposition [9] and the quotient in (22) being the pgf of the system content in the classic  $M/G/1$  queue,  $\chi(z)$  is the pgf of the system content at a random moment in vacation. In addition, Fuhrmann-Cooper decomposition [9] also gives the following relation between the sojourn time  $D$  in the system and the sojourn time  $D_{M/G/1}$  in the classic  $M/G/1$  system:

$$\mathbb{E}e^{-\alpha D} = \mathbb{E}e^{-\alpha D_{M/G/1}} \chi(1 - \alpha/\lambda). \quad (23)$$

After some calculations (23) can be rewritten as

$$\mathbb{E}e^{-\alpha D} = \frac{(1 - \rho - \rho_u) \alpha}{\alpha - \lambda (1 - \mathbb{E}e^{-\alpha B_{ext}})} \mathbb{E}e^{-\alpha B_{ext}},$$

which is consistent with (19) and (20).

## 7 Summary and concluding remarks

In this paper, we have studied an  $M/G/1$  queueing model with a *modified* time-limited vacation mechanism. Contrary to the traditional time-limited vacation mechanism, the server does not switch to vacation mode after a predefined time or when the system is empty, but the server switches to a vacation only after a predefined amount of *work* has been processed. The main contribution of this paper is the mathematical analysis of this model. More specifically, we have derived an explicit expression for the distribution of the time it takes until the prespecified amount of work has been served. For the case the total amount of work till vacation is exponentially distributed, we have derived the transforms of the steady-state workload at various epochs, busy period, waiting time, sojourn time, and queue length distributions.

Depending on the practical context, this model could be used as a framework for formulating and solving optimization problems. For example, in the context of sharing resources among production processes, keeping the work in process (WIP) low while avoiding high opportunity costs of frequently keeping the server idle is the trade-off to tackle. Therefore, the optimization problem may be defined as minimizing the weighted sum of the costs associated with the mean workload ( $\mathbb{E}[Z]$ ) and the costs of keeping the server idle ( $\mathbb{P}[idle]$ ). In the context of maintenance, the challenge is to determine the optimal timing of preventive maintenance such that on one hand, the WIP level is low, while on the other hand, unscheduled breakdowns are avoided. Our results enable to express WIP as  $\mathbb{E}[Z]$  and the risk of failure can be expressed as the probability that the usage time to failure is smaller than the amount of work  $A$  to be executed, which in case of  $A$  exponentially distributed with parameter  $\mu$  can be readily expressed as  $F^*(\mu)$  with  $F^*(\cdot)$  the Laplace transform of the usage

time at failure of the machine. Formulating and solving optimization problems in detail is left as future work as the scope of this paper is to characterize the performance measures of this modified time-limited vacation system ( $\mathbb{E}[Z]$ ,  $\mathbb{P}[\text{idle}]$ , et cetera) in terms of its input parameters (time  $A$  until vacation, arrival rate, service rate, et cetera).

## A Alternative approach for establishing $Z_b(t)$

In the process  $Z_b(t)$  restricted to busy periods, all idle periods are removed and the vacations are condensed to instantaneous arrivals of work  $U$ , the total amount of work that accumulates during the vacation. Hence, instead of vacations  $V_i$  occurring at rate  $\mu$  in the original process, service requirements  $U_i$  occur at rate  $\mu$  in  $Z_b(t)$ ; those service requirements  $U_i$  are i.i.d. with common random variable  $U$ , which is distributed as

$$U \sim \sum_{i=1}^{N_\lambda(V)} B_i .$$

The (restricted) workload process  $Z_b(t)$  corresponding to the (unrestricted) workload process from Figure 1 is illustrated in Figure 3.

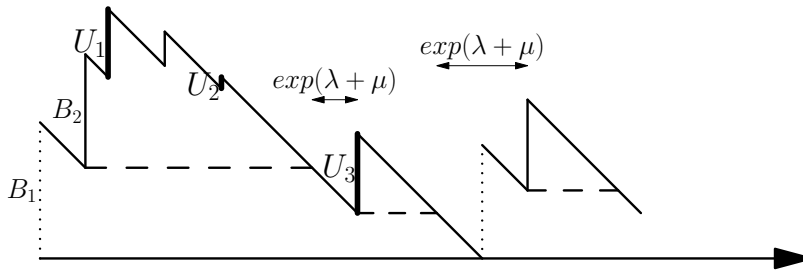


Figure 3: Restricted workload process  $Z_b(t)$  corresponding to the unrestricted workload process depicted in Figure 1; dotted lines illustrate secondary jumps, bold lines show arrivals of work  $U$ .

The workload process  $Z_b(t)$  in fact corresponds to the workload process in an  $M/G/1$  queue with arrival rate  $\lambda + \mu$  and service time distribution

$$\frac{\lambda}{\lambda + \mu} F_B(\cdot) + \frac{\mu}{\lambda + \mu} F_U(\cdot) .$$

In Section 1 we referred to this  $M/G/1$  queue as Model I. In addition, every time the workload process  $Z_b(t)$  hits zero there is an instantaneous jump up distributed like  $B$  (the service time that appears at the end of an idle period). These instantaneous jumps are referred to as secondary jumps and are illustrated as dotted lines in Figure 1. All other jumps, both those indicated with normal and bold lines in Figure 1, are referred to as non-secondary jumps.

Note that the part above the dashed lines (starting at the first non-secondary jump after the previous dashed period has finished and ends when the workload level reaches the same level as just before that jump) corresponds to the workload process during a busy period in Model I. In addition, the part below the dashed lines is distributed as a remaining or elapsed service time  $B_e$ . In the remainder, we prove that the limiting distribution  $Z_b$  of the restricted workload process is

that of an independent sum of two random variables. The first has the stationary distribution  $Z_I$  of the workload in Model I and the second has the distribution of  $B_e$ :

$$Z_b \sim Z_I + B_e . \quad (24)$$

The key is to note that  $Z_b(t)$  corresponds to the *virtual waiting time* in model I with the extra feature of having vacations distributed as  $B$  whenever the system becomes empty. Due to the Poisson arrival process, the PASTA property holds, and thus this virtual waiting time is distributed as the steady-state waiting time. Next, application of the Fuhrmann-Cooper decomposition [9] yields

$$\mathbb{E}e^{-\alpha Z_b} = \mathbb{E}e^{-\alpha Z_I} \chi_I (1 - \alpha/\lambda) , \quad (25)$$

with  $\chi_I(z)$  the pgf of the system content at a random moment in a vacation in model I with vacations. As a vacation starts only when the system becomes empty in that system, it holds that

$$\chi_I(z) = \mathbb{E}e^{-\lambda(1-z)B_e} ,$$

and thus that

$$\chi_I (1 - \alpha/\lambda) = \mathbb{E}e^{-\alpha B_e} . \quad (26)$$

The combination of (25) and (26) yields (24).

**Remark 5.** (24) can also be proved by [14] or [3], which is valid for a more general Lévy setting.

## References

- [1] O.J. Boxma, D. Claeys, L. Gulikers, O. Kella (2015). A queueing system with vacations after  $N$  services, *Naval Research Logistics* 62, 646-658.
- [2] M. Ben-Daya, U. Kumar, D.N.P. Murthy (1993). *Maintenance Engineering*, Wiley.
- [3] O. Boxma, O. Kella (2014). Decomposition results for stochastic storage processes and queues with alternating Lévy inputs, *Queueing Systems* 77, 97-112.
- [4] J.A. Buzacott, J.G. Shanthikumar (1993). *Stochastic Models of Manufacturing Systems*, Prentice Hall, Englewood Cliffs.
- [5] J.W. Cohen (1982). *The Single Server Queue*, North-Holland, Amsterdam.
- [6] T. Demoor (2010). The preemptive repeat hybrid server interruption model, *Analytical and Stochastic Modeling Techniques and Applications. ASMTA 2010. Lecture Notes in Computer Science* 6148, 59-71.
- [7] B.T. Doshi (1986). Queueing systems with vacations – A survey, *Queueing Systems* 1(1), 29-66.
- [8] D. Fiems, T. Maertens, H. Bruneel (2008). Queueing systems with different types of interruptions, *European Journal of Operational Research* 188, 838-845.
- [9] S.W. Fuhrmann, R.B. Cooper (1985). Stochastic decompositions in the  $M/G/1$  queue with generalized vacations, *Operations Research* 33(5), 1117-1129.
- [10] W.J. Hopp, M.L. Spearman (2012). *Factory Physics*, 3rd ed., McGraw-Hill.

- [11] A.K.S. Jardine, A.H.C. Tsang (2006). *Maintenance, Replacement, and Reliability. Theory and Applications*. Dekker Mechanical Engineering. CRC Press.
- [12] T. Katayama (2001). Waiting time analysis for a queueing system with time-limited service and exponential timer, *Naval Research Logistics* 48, 638-651.
- [13] O. Kella (1998). An exhaustive Lévy storage process with intermittent output, *Stochastic Models* 14, 979-992.
- [14] O. Kella, W. Whitt (1991). Queues with server vacations and Lévy processes with secondary jump input, *Annals of Applied Probability* 1, 104-117.
- [15] D. Lee (1997). Analysis of a single server queue with semi-Markovian service interruption, *Queueing Systems* 27(1-2), 153-178.
- [16] K.K. Leung, M. Eisenberg (1997). A single-server queue with vacations and gated time-limited service, *IEEE Transactions on Communications* 38, 1454-1462.
- [17] K.K. Leung, M. Eisenberg (1991). A single-server queue with vacations and non-gated time-limited service, *Performance Evaluation* 12, 115-125.
- [18] X.S. Si, W. Wang, C.H. Hu, and D.H. Zhou (2011). Remaining useful life estimation - A review on the statistical data driven approaches. *European Journal of Operational Research* 213, 1-14.
- [19] R. Suri (1998). *Quick Response Manufacturing*, Taylor&Francis Inc.
- [20] H. Takagi (1991). *Queueing Analysis: A Foundation of Performance Evaluation, volume 1: Vacation and Priority Systems, Part 1*, North-Holland, Amsterdam.
- [21] N. Tian, Z.G. Zhang (2006). *Vacation Queueing Models*, Springer, New York.
- [22] F. Van der Duyn Schouten, S. Vanneste (1995). Maintenance optimization of a production system with buffer capacity, *European Journal of Operational Research* 82, 323-338.
- [23] H.Z. Wang. A survey of maintenance policies of deteriorating systems (2002). *European Journal of Operational Research* 139, 469-489.