EURANDOM PREPRINT SERIES

2019-005

June 3, 2019

The (S-1,S) inventory model and its counterparts in queueing theory

O. Boxma, D. Perry, W. Stadje
ISSN 1389-2355

The $(S-1,S)$ inventory model and its counterparts in queueing theory

Onno J. Boxma[a,1], David Perry[b,1], Wolfgang Stadje[c,1]

[a] *Department of Mathematics and Computer Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*
[b] *Holon Institute of Technology, P.O. Box 305, Holon 5810201, Israel*
[c] *Institute of Mathematics, University of Osnabrück, Osnabrück 49069, Germany*

Abstract

We explore the relationship between the $(S-1,S)$ inventory model and three well-known queueing models: the Erlang loss system, the machine-repair model and a two-node Jackson network. Exploiting this relationship allows us to obtain key performance measures of the $(S-1,S)$ model, like the so-called virtual outdating time, the number of items on the shelf in steady state, the long-run rate of unsatisfied demands and the distribution of the empty shelf period.

Keywords: $(S-1,S)$ inventory model; Erlang loss system; machine-repair model; Jackson network; virtual outdating time.

1. Introduction

The $(S-1,S)$ inventory system [13, 14] is a closed cyclic model of S items of which there are $j \in \{0,1,...,S\}$ items waiting 'on a shelf' (in inventory), while $S-j$ other items are on their way to the shelf after having been ordered. In the basic model, demands for items arrive according to a Poisson process. If a demand for an item arrives while there is at least one item on the shelf, this demand is immediately satisfied and under the FIFO issuing policy that means that the oldest item is removed from the shelf. Immediately after that removal a new item is ordered; ordered items enter the shelf after independent and identically distributed leadtimes. If a demand arrives with no items on the shelf, then this demand leaves unsatisfied. Natural extensions of the basic model described above allow removals from the shelf to occur not only due to demands, but also due to perishability, obsolescence or "sudden deaths" of items. Another extension is that demands which can not immediately be satisfied have some patience, i.e., are willing to wait a certain amount of time.

Our $(S-1,S)$ analysis is based on the observation that the basic $(S-1,S)$ model can be mapped one-to-one on the machine-repair model (also called machine interference or computer-terminal model; cf. Section 4.11 of [8] or Section 3.9 of [9], and to a product-form queueing network [1] with one single-server node and one infinite-server node. The latter queueing model is also known to be equivalent to the well-known Erlang loss $(M/G/S/S)$ queueing system.

Our main goals in this paper are (i) to emphasize the relation between the $(S-1,S)$ inventory model and a number of important queueing models, in particular the relation between the $(S-1,S)$ model and the Erlang loss system and (ii) to exploit the latter relation to obtain results for key performance measures of the $(S-1,S)$ model: the so-called *Virtual Outdating Time* (VOT), the number of items on the shelf in steady state, the long-run rate of unsatisfied demands and the distribution of the empty shelf period.
The first of these performance measures was studied in [11, 12, 13] for the case of *constant* leadtimes.

The paper is organized as follows. In Section 2 we present a model description of the $(S-1,S)$ inventory system, and we also briefly review the $M/G/S/S$ model, the machine-repair model and the two-node closed queueing network with one single-server node and one infinite-server node; subsequently we outline their relations. Section 3 is devoted to a discussion of the Virtual Outdating Time process $\{V(t), t \geq 0\}$ in the $(S-1,S)$ model. We derive the steady-state density of $V(t)$ by exploiting the relation between the $(S-1,S)$ model and the Erlang loss model, using a particular result from Section 2.3 of [4] for the Erlang

Email addresses: o.j.boxma@tue.nl (Onno J. Boxma), davidper@hit.ac.il (David Perry), wstadje@uos.de (Wolfgang Stadje)

loss model. In Section 4 we discuss a variant of the basic $(S-1,S)$ model: items on the shelf can suddenly become worthless and have to be removed from the shelf. Section 5 contains some conclusions and suggestions for further research.

2. Preliminaries

In this section we briefly review the four classical operations research models mentioned above, viz., the $(S-1,S)$ inventory system with Poisson(λ) demand arrival process and generally distributed leadtimes with distribution function $G(\cdot)$; the Erlang loss system with Poisson(λ) arrival process and service times $\sim G(\cdot)$; the machine-repair model (MRM) with lifetimes $\sim G(\cdot)$ and one repairman with exp(λ) repair times; and the two-node Jackson product-form network with one exp(λ) single server node and one infinite-server node with generally distributed service times $\sim G(\cdot)$. The key observation is that, in their standard form as sketched above, these four systems *are basically equivalent*; we can map them one-to-one onto each other. For the MRM and the two-node Jackson network this is not difficult to see, and for the Erlang loss system and the MRM this is well-known; for $(S-1,S)$ and Erlang loss the relation is also straightforward, but it has received somewhat less attention in the literature [10, 15]. We discuss the relations between these four models and their functionals in some more detail in Subsection 2.5. We close the section by mentioning variants and extensions of the standard $(S-1,S)$ inventory system and their equivalent counterparts in the Erlang loss system, the MRM and the two-node Jackson product-form network. The main reasons for discussing these relationships are (i) they will enable us to obtain a new result for the $(S-1,S)$ system, and (ii) we feel that relationships between $(S-1,S)$ (and its extensions) and one or more of the three queueing models (and their corresponding extensions) have considerable potential.

2.1. The $(S-1,S)$ inventory system

The $(S-1,S)$ inventory system is a closed storage system with $S \in \mathbb{N}$ items, either on the shelf or on their way to the shelf, after having been ordered. Demands for items arrive according to a Poisson process with rate λ. If there is no item on the shelf, then an arriving demand leaves unsatisfied and is lost. If at least one item is present on the shelf, then an arriving demand is immediately satisfied; furthermore, instantaneously a new item is ordered. Under the FIFO issuing policy any removal from the shelf is that of the oldest item. The leadtimes of ordered items (times between issuing the order and arrival on the shelf) are independent, identically distributed (i.i.d.) random variables, with general distribution function $G(\cdot)$ and mean τ.

A key performance measure is the steady-state distribution of the number of items $\tilde{K}$ on the shelf: $Pr(\tilde{K}=n),\ n=0,1,\ldots,S$. For reasons that soon will become clear, we focus on $K=S-\tilde{K}$, the number of items *not* on the shelf.

Another important performance measure is the Virtual Outdating Time process $\mathbf{V}=\{V(t),t\geq 0\}$. Let us consider this VOT process in more detail. We construct the VOT from the age processes $\mathbf{A_i}=\{A_i(t):t\geq 0\}$ for $i=1,2,...,S$, where $\mathbf{A_i}$ is defined such that $A_i(t)=x<0$ if $-x$ is the age of the ith item on the shelf (provided there are at least i items present) and if $A_i(t)=x\geq 0$ then the ith item has been ordered but not yet arrived at the shelf and x is its time until arrival.

The VOT process $\mathbf{V}$ is formally defined by

$$V(t)=\min_{i=1,2,\ldots,S}\{A_i(t)\};\quad 0\leq t<\infty.$$

Note that this definition is independent of the numbering of the S items. In words, $V(t)$ indicates the age of the oldest item in the system (among all S items) in the sense that when $V(t)=x<0$ then $-x$ is the age of the oldest item present on the shelf and when $V(t)=x\geq 0$ then the shelf is empty and x is the time until the next arrival. In Figure 1 the process $\mathbf{V}$ is emphasized by the thick line.

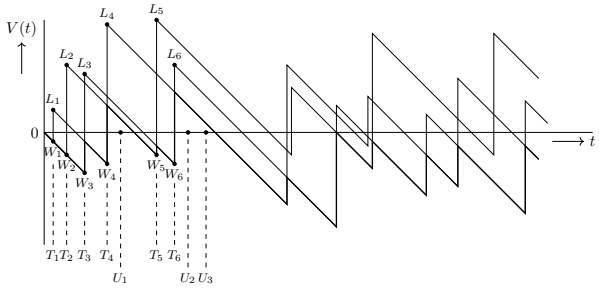

Figure 1: VOT for the case of $S=3$ with general leadtime τ.
L_i denote leadtimes; T_i (U_i) arrival epochs of satisfied (unsatisfied) demands;
$-W_i$ = age of the oldest item on the shelf at epochs of satisfied demands.
Initial condition: Just before $t=0$, the shelf was empty, but at $t=0$ an item arrived on the shelf.

Remark 1.
At first sight, the VOT behaves quite similar to the workload process in the $M/G/1$ queue (apart from becoming negative). However, it should be noted that $\mathbf{V}$ is not a Markov process, since in the

general leadtime case some of the jump sizes may be equal to an interarrival time that occurred in the past. In particular, in the special case of constant leadtimes each jump size (ignoring the first few) is equal to a certain interarrival time that occurred in the past (see also Remark 5 below).

2.2. The $M/G/S/S$ queue

The $M/G/S/S$ queue, also called Erlang loss system [7], is a queueing system with Poisson(λ) arrival process, general service time distribution function $G(\cdot)$ with mean τ, S servers and no waiting room. An arriving customer who finds all servers busy is lost. All interarrival and service times are independent. Let $\mathbf{K} = \{K(t) : t \geq 0\}$ be the process representing the number of customers in the system, let K be this number in steady state and let $p_n := \Pr(K = n)$ denote the steady-state probability of the system having n customers. Also let $G_{r1}, ..., G_{rn}$ be the remaining service times in steady state. Finally, we use the notation $G_e(\cdot)$ for the equilibrium distribution function of $G(\cdot)$, i.e.,

$$G_e(x) = \int_0^x \frac{1 - G(y)}{\tau} \mathrm{d}y.$$

For the special case of the $M/M/S/S$ queue, $\{p_n\}$ satisfies the recursion

$$\lambda p_n = \frac{n+1}{\tau} p_{n+1}, \qquad n = 0, ..., S-1$$

whose solution is easily seen to be given by

$$p_n = \frac{(\lambda\tau)^n/n!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}. \tag{1}$$

For future use, we mention two important facts regarding the Erlang loss system with general service time distribution (see [16], Section 5.11).

1. *Insensitivity.* p_n is insensitive to $G(\cdot)$ in the sense that the number of customers in the system depends on $G(\cdot)$ only through its mean; i.e., formula (1) remains valid for $M/G/S/S$.
2. *Conditional independence.*

$$\Pr[K = n; G_{ri} \leq x_i, i = 1, ..., n] = p_n \prod_{i=1}^{n} G_e(x_i). \tag{2}$$

Thus, given $K = n$, the remaining service times $G_{r1}, ..., G_{rn}$ are conditionally identically distributed and independent; moreover, each remaining service time has the equilibrium distribution.

The insensitivity property (1) with respect to $G(\cdot)$ depends on having Poisson arrivals, but this result can be generalized to a 'Poisson-like' arrival process in which the arrival rate λ_n is a function of the number of customers in system, $n = 0, 1, ..., S$. This is the arrival process for the 'birth' part of a birth and death process. In this case, the above insensitivity property and Formula (2) still hold (see [16] p. 273), while p_n now satisfies the recursion

$$\lambda_n p_n = \frac{n+1}{\tau} p_{n+1}, \qquad n = 0, 1, ..., S-1.$$

We return to this extension in Section 4.

2.3. The machine-repair model

The MRM, also called the machine interference model or the computer-terminal model (cf. [9], Section 3.9), describes a system of S machines which are subject to breakdown and repair. A machine is operative during a random time with distribution $G(\cdot)$ and mean τ, but then it breaks down and is sent to a repairshop. This repairshop has R repairmen, who individually serve broken machines FCFS, with exp(λ)-distributed service times. A repaired machine immediately goes back into operation. All involved operative and repair time periods are independent. In the classical case, $R = 1$. Then the steady-state distribution of the number of operative machines is given by (1), regardless of the distribution function $G(\cdot)$. One way to see this is to view the machine-repair model with one repairman as a two-queue closed Jackson network, with a single server queue (the repair queue) and an infinite-server queue, and use classical product-form results [1, 6]; see the next subsection. Another special case is $R = S$; in that case, the machine-repair model may obviously be viewed as a two-queue closed Jackson network with two infinite-server queues.

2.4. The two-node Jackson network

Consider a closed queueing network consisting of two nodes and with a fixed number of S customers. Node 1 has a single server, who is serving customers in order of arrival, with i.i.d. exp(λ) service times. Node 2 is an infinite-server system with generally distributed service times $\sim G(\cdot)$ with mean τ. It is well-known [1] that the steady-state probability $p(S-n, n)$ of having $S-n$ customers in node 1 and (hence) n customers in node 2 is given by the product form, for $n = 0, 1, \ldots, S$,

$$p(S-n, n) = \frac{(\frac{1}{\lambda})^{S-n}\frac{\tau^n}{n!}}{\sum_{i=0}^{S}(\frac{1}{\lambda})^{S-i}\frac{\tau^i}{i!}} = \frac{(\lambda\tau)^n/n!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}. \tag{3}$$

Once again the insensitivity with respect to the distribution function $G(\cdot)$, apart from its mean, manifests itself.

2.5. Relations between the models

In this subsection we summarize the translations which have to be made to relate the $(S-1,S)$ system to the three fundamental queueing models $M/G/S/S$, MRM and the 2-node closed Jackson network with a single-server node and an infinite-server node. Crucial are the following observations.

I. MRM and the two-queue closed Jackson network with a single exponential server and an infinite-server (or S-server) queue with general service times are equivalent; each lifetime in the MRM can be viewed as a service time in the infinite-server node.

II. As is well-known, the $M/G/S/S$ Erlang loss system with Poisson(λ) arrival process and general $\sim G(\cdot)$ service times can be mapped one-to-one onto the machine-repair model with S machines, general $\sim G(\cdot)$ lifetimes and one repairman with exp(λ) repair times. One can identify the service times in $M/G/S/S$ with the lifetimes in MRM, and then it is seen in particular that the Markov process represented by the vector of the number of busy servers together with the residual service times in $M/G/S/S$ has exactly the same steady-state distribution as the vector of the number of working machines together with the residual lifetimes in MRM. That joint steady-state distribution is given in (2).

III. In the same way one can map the $(S-1,S)$ inventory system with general $\sim G(\cdot)$ leadtimes and Poisson(λ) demand one-to-one onto the MRM: The oldest item on the shelf in $(S-1,S)$ is taken by the next arriving demand after exp(λ)-distributed amount of time, corresponding to the machine that broke down the longest time ago being repaired after the same exp(λ)-distributed amount of time; and an item returns to the shelf after a leadtime $\sim G(\cdot)$, corresponding to a lifetime of a machine.

The above constructions imply the equivalence of all four models. Hence, the joint steady-state distribution of the number of items *not* on the shelf and their residual leadtimes is also given by (2).

Below we explicitly relate key quantities of the four models to one another, paying in particular attention to the VOT process $\mathbf{V}$. Table 1 contains an overview of these relations.

The first two rows of the table speak for themselves: these concern the generally distributed service/life/leadtimes and the exponential rates. The third row concerns a quantity $K \in \{0,1,\ldots,S\}$: this is the number of items *not* on the shelf for $(S-1,S)$, the number of busy servers in $M/G/S/S$, the number of working machines in MRM and the number of customers at the infinite-server queue in the two-node Jackson network. The last two rows are somewhat less straightforward. They concern in particular the Virtual Outdating Time of $(S-1,S)$, i.e., the age of the oldest item on the shelf when $K<S$ and the time until the first item arrives on the shelf when the shelf is empty $(K=S)$. The table gives the translation of these quantities to $M/G/S/S$, MRM and the two-node Jackson network.

Remark 2

One could also consider generalizations of the $(S-1,S)$ model, like patience of the customers who demand items, and perishability of items on the shelf; these translate into generalizations of the $M/G/S/S$ model and of the MRM model (and the 2-node Jackson network, which we disregard for the moment).

In Section 4 we shall study the case of *sudden deaths* in $(S-1,S)$. This corresponds in $M/G/S/S$ to the case that an idle server receives a job after exp(ξ) (next to the ordinary Poisson arrival process). In MRM any machine in the repairshop leaves the shop after exp(ξ), even if it has not yet been repaired; a new machine is instantaneously bought.

Outdating (perishability) in $(S-1,S)$ corresponds in $M/G/S/S$ to the case in which a server who has been idle for longer than a given (random or deterministic) time is instantaneously given a job to do. In MRM a machine has a finite "patience" in the repairshop; if that patience is exceeded, it leaves the repairshop and instantaneously a new machine is bought.

Patience of demands in $(S-1,S)$ corresponds in $M/G/S/S$ to the case that an arriving customer who finds all servers busy is willing to wait a while (his patience time); this becomes the $M/G/S+G$ model. In MRM, there seems to be no natural counterpart to this.

3. The VOT process

In this section we formulate and prove the main theorem of the paper; it gives the steady-state density $f(\cdot)$ of the VOT process $\{V(t), t \geq 0\}$ that was introduced in Subsection 2.1. From the VOT all important performance measures of the $(S-1,S)$ inventory model can be derived. As seen in Subsection 2.1, we need to distinguish between the cases $K=S$ and $K<S$.

	$(S-1,S)$	$M/G/S/S$	MRM	closed SSQ + ISQ
$G(\cdot)$	leadtime	service time	life time	service time ISQ
λ	demand rate	arrival rate	repair rate	service rate SSQ
K	# items not on the shelf	# busy servers	# working machines	# in ISQ
$-V$ if $K<S$	age oldest item on shelf	time longest idle server has been idle	time job in repair has been in shop	past sojourn time of customer in service in SSQ
V if $K=S$	time till next arrival on shelf	time till next service completion	time till next breakdown	time till next service completion in ISQ

Table 1: Relations among the models

Theorem 1. *The steady-state density $f(\cdot)$ of the VOT is given by*

$$f(x) = \frac{(\lambda\tau)^S/S!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!} S(1-G_e(x))^{S-1}\frac{1-G(x)}{\tau}, \quad x>0,$$

$$f(x) = \frac{1}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}\lambda \mathrm{e}^{\lambda x}\frac{(\lambda(\tau-x))^{S-1}}{(S-1)!}, \quad x<0. \tag{4}$$

Proof. Let us first determine $f(x)$ when $x>0$, i.e., when $K=S$. Recall that, if $K(t)=S$ (no items on the shelf at time t) then $V(t)$ is the time until the next arrival at the shelf, so it is equal to the minimum of the residual leadtimes of all the S items which are presently on their way to the shelf. It follows from Formula (2) for the joint distribution of the number of customers in the $M/G/S/S$ system and their residual service times, in combination with the equivalence results between $(S-1,S)$ and $M/G/S/S$ as discussed in the previous section, that

$$\Pr(V>x|K=S) = (1-G_e(x))^S, \quad x>0, \tag{5}$$

and hence

$$f(x|K=S) = S(1-G_e(x))^{S-1}\frac{1-G(x)}{\tau}, \quad x>0. \tag{6}$$

Multiplying by $\Pr(K=S)$, as given in the Erlang loss system, yields the first assertion of the theorem.

Let us now determine $f(x)$ when $x<0$, i.e., when $K<S$. If $K(t)<S$ then $-V(t)$ is the age of the oldest item on the shelf. We again rely on the equivalence between $(S-1,S)$ and $M/G/S/S$. In the $M/G/S/S$ terminology, $-V$ is the time the longest idle server has been idle. The crucial observation now is the following. As proven by Cohen [4] (see in particular p. 70) for the $M/G/S/S$ model, the process $\{(K(t),\zeta^{(i)}(t),\eta^{(i)}(t)), i=0,1,\ldots,K(t), t\in(-\infty,\infty)\}$, with $\zeta^{(i)}(t)$ and $\eta^{(i)}(t)$ denoting the residual and past service time of the ith customer in service, is *reversible* when it is stationary. If we now reverse time in $M/G/S/S$, then the joint distribution of the number of busy servers and the time the longest idle server has been idle becomes the joint distribution of the number of busy servers and the time until the last idle server becomes busy. The latter time is Erlang$(S-k,\lambda)$ distributed if there are k busy servers. Here we assume that an arriving customer is taken into service by the server who has been idle the longest; just as, in $(S-1,S)$, a demand takes the oldest item on the shelf. Using (2) and time reversal we now obtain, for $k=0,1,\ldots,S-1$ and $x<0$:

$$\Pr(K=k, V>x) = \frac{(\lambda\tau)^k/k!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}[1-\sum_{j=0}^{S-k-1}\mathrm{e}^{\lambda x}\frac{(-\lambda x)^j}{j!}]. \tag{7}$$

Accordingly, for $k=0,1,\ldots,S-1$ and $x<0$, with $f_k(x)$ denoting the joint density of the VOT process and number not on the shelf,

$$f_k(x) = \frac{(\lambda\tau)^k/k!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}\ \lambda\mathrm{e}^{\lambda x}\ \frac{(-\lambda x)^{S-k-1}}{(S-k-1)!}, \tag{8}$$

and

$$\sum_{k=0}^{S-1} f_k(x) = \sum_{k=0}^{S-1}\frac{(\lambda\tau)^k/k!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}\ \lambda\mathrm{e}^{\lambda x}\ \frac{(-\lambda x)^{S-k-1}}{(S-k-1)!} = \frac{1}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}\ \lambda\mathrm{e}^{\lambda x}\ \frac{(\lambda(\tau-x))^{S-1}}{(S-1)!}. \tag{9}$$

■

Remark 3
The reversibility result from Cohen [4] that was used in the proof actually is a subresult, part of his lengthy but essentially elementary proof of the important insensitivity result for the number of customers in $M/G/S/S$. Cohen ends the corresponding chapter with the statement "It is of great interest to investigate whether the approach used here in the investigation of the $M/G/K$ loss system can be used also for those other models." $(S-1,S)$ turns out to be such a model.

Remark 4
It is easily checked that $f(\cdot)$ as given in Theorem 1 is a density. Indeed $f(x)>0$ for all x,

while integration over $x > 0$ gives $\frac{(\lambda\tau)^S/S!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}$; and integration over $x < 0$ gives (after recognizing the Erlang(S,λ) distribution): $\frac{\sum_{i=0}^{S-1}((\lambda\tau)^i/i!}{\sum_{i=0}^{S}(\lambda\tau)^i/i!}$. Hence $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$. It should also be noticed that plugging in $x = 0$ in either of the two equations of (4) gives the same result. Indeed, $f(0-) = f(0+)$, since 0 is a point of continuity of the VOT.

Remark 5
It may be tempting to consider the following alternative derivation of the density $f(\cdot)$ of the VOT. We give the argument for the case of constant leadtimes, referring to Figure 2.

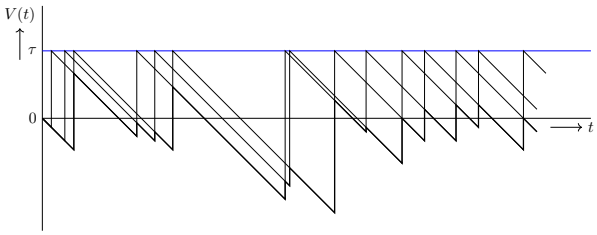

Figure 2: VOT for the case of $S = 3$ with constant leadtime τ

Apply the so-called Level-Crossing Technique (LCT), cf. [3, 5], to the $V(t)$ process. According to LCT, the long-run average number of downcrossings of any level x equals the long-run average number of upcrossings of x. Our claim is that this leads to the following identity:

$$f(x) = \int_{-\infty}^{x} \lambda(w)(\frac{\tau - x}{\tau - w})^{S-1} f(w)\mathrm{d}w, \quad -\infty < x < \tau, \tag{10}$$

where $\lambda(w) = \lambda$ for $w < 0$ and $\lambda(w) = 0$ otherwise. Indeed, the lefthand side of (10) is the downcrossing rate of level x. We now explain the righthand side. First of all, if $V = w > 0$ then the shelf is empty and there can be no upcrossing: $\lambda(w) = 0$. Now assume the shelf is not empty. When the items are ordered according to their remaining shelf lifes we have $A_1(t) < \cdots < A_S(t)$, where $A_1(t) = V(t)$. Suppose that t is a demand arrival time. By conditioning on the age of the oldest item on the shelf ($V(t-) = w < 0$), we know that there were exactly $S-1$ arrivals in the interval $(t-(\tau-w), t)$. Hence, when $V(t-) = w$, the upcrossing rate of level x is $\lambda(w) = \lambda$ times the conditional probability that the first arrival time in $(t-(\tau-w), t)$ is larger than $t-(x-w)$ given that the S-th arrival takes place at time t. Now suppose for a moment that these arrival times follow a Poisson process. Then we can use a well-known property of the Poisson process: the $S-1$ arrival epochs in the interval $(t-(\tau-w), t)$ are uniformly distributed on that interval. Hence the probability that the first of these arrival epochs occurs after $t-(x-w)$ is given by

$$(\frac{\tau - x}{\tau - w})^{S-1}.$$

Moreover, PASTA (Poisson Arrivals See Time Averages) would allow us to assume that the workload at these jump epochs has density $f(\cdot)$. This results in (10).

However, one has to be very careful with this reasoning. In particular, observe that there are $S-1$ *admitted* demand arrivals in an interval of length $\tau - w$, but we know that all arrivals in the last w time units were admitted, because $V(t) < 0$ during those time units – whereas we do not know whether all arrivals in the earlier part of the time interval were admitted. Furthermore, the VOT is not a Markov process; every jump size (after the third arrival in Figure 2) is an interarrival time that occurred in the past. This suggests that a more delicate argument is required to use the LCT here. It seems vaguely reminiscent of phenomena in the above-discussed related queueing models. For example, for the Erlang loss model Bonald [2] has shown that the insensitivity w.r.t. the service time distribution beyond its mean even holds when the assumption of a Poisson arrival process is extended in the following way: users generate *sessions* according to a Poisson process, and each session is composed of a random finite number of calls and idle periods. Furthermore, in product-form Jackson networks with feedback, nodes have been shown to behave like $M/M/c$ queues *even though their arrival process is not Poisson* [6].

Note that, when following the above argument, the LCT equation (10) can be extended to the case of general functions $\lambda(w)$. It then takes the form

$$f(x) = \int_{-\infty}^{\min(x,0)} \lambda(w)(1 - \frac{\Lambda(\tau - x)}{\Lambda(\tau - w)})^{S-1} f(w)\mathrm{d}w, \tag{11}$$

where $\Lambda(x) = \int_{-\infty}^{x} \lambda(w)\mathrm{d}w$.

Remark 6
Despite what was observed at the end of Remark 5, the LCT relation (10) appears to be valid. Indeed, first consider this equation for $x < 0$. Introducing

$$g(x) := \frac{f(x)}{(\tau - x)^{S-1}},$$

equation (10) is translated into

$$g(x) = \lambda \int_{-\infty}^{x} g(w)\mathrm{d}w, \quad x < 0, \tag{12}$$

which after differentiation with respect to x yields $g(x) = C_1 e^{\lambda x}$ for some constant C_1, when $x < 0$. We thus easily retrieve expression (4) for $f(x)$, $x < 0$.
Next consider (10) for $x > 0$. Since $\lambda(w) = 0$ for $w > 0$, we obtain

$$g(x) = \lambda \int_{-\infty}^{0} g(w)\mathrm{d}w, \quad x > 0, \qquad (13)$$

yielding $g(x) = C_2$ and $f(x) = C_2(\tau - x)^{S-1}$ for some constant C_2, when $x > 0$. Realizing that, in this case of constant leadtimes, $G_e(x) = x/\tau$, $0 < x < \tau$, we easily retrieve expression (4) for $f(x)$, $x > 0$.

There are other relevant measures and functionals for $(S-1, S)$ than the law of the number of items on the shelf and the VOT. Among them are the rate of the unsatisfied demand λ^* and the distribution of the emptiness period I, i.e., the time period that the shelf is empty. We consider them below.

The rate of the unsatisfied demand. Recall that the event $\{V > 0\}$ is equivalent to the event $\{K = S\}$ and λ^*/λ is the long-run proportion of unsatisfied demands. However, a demand is unsatisfied if and only if it arrives when the shelf is empty. Thus the long run average proportion of unsatisfied demands is equal to the steady-state probability that the shelf is empty. By the equivalence with the Erlang loss system, that probability is given by p_S in (1). We thus obtain by PASTA:

$$\lambda^* = \lambda \frac{(\lambda\tau)^S}{S!} \left[\sum_{j=0}^{S} \frac{(\lambda\tau)^j}{j!} \right]^{-1}. \qquad (14)$$

Note that $1/\lambda^*$ is the expected length of the time between successive unsatisfied demands.

The distribution of the emptiness period. Again exploiting the relation between $(S-1, S)$ and the $M/G/S/S$ system, we easily determine a – new, to the best of our knowledge – result for the distribution function $U_I(\cdot)$ of the emptiness period.

Theorem 2.

$$U_I(x) = 1 - (1 - G_e(x))^{S-1}(1 - G(x)), \quad x > 0.$$

Proof. The equivalence between $(S-1, S)$ and $M/G/S/S$ implies that $U_I(x)$ is the distribution of the uninterrupted time all S servers are busy in $M/G/S/S$. When a customer arriving at an $M/G/S/S$ system finds only one server idle, a period starts in which all S servers are busy. The time this period lasts is the minimum of that customer's full service time and the residual service times of all other $S-1$ customers. Using PASTA and the conditional independence of the residual service times and of the number of busy servers, cf. (2), the result follows. ■

4. Sudden deaths

In this section we consider the $(S-1, S)$ inventory system with general leadtime distribution, with the following special feature: items leave the shelf either by demand satisfaction (Poisson(λ)), *or* due to a sudden death (exp(ξ) per item).

Let us first restrict ourselves to the case with only sudden death, so without demand. The $(S-1, S)$ model now is equivalent with a machine-repair model with S (i.e., ample) repairmen, each with repair rate ξ. This MRM, in its turn, is obviously equivalent to a closed queueing network consisting of two infinite-server nodes. The distribution of the number of working repairmen clearly is binomial: we have S independent alternating renewal processes, with alternating general lifetime and exponential repair phases. This is a special case of a product-form queueing network, which is known to be insensitive [1]. If one also adds the Poisson demand feature, it becomes a Jackson network with one infinite server with general service time distribution function $G(\cdot)$, and one single server with state-dependent service rate $n\xi + \lambda$ when there are n customers at the single server. This two-node closed queueing network also is a special case of a product-form queueing network, and again it is insensitive [1]; see also Theorem 15 on p. 323 of [16].

The equivalence between the $(S-1, S)$ inventory system and the MRM with ample repairmen, and with the two-node queueing network, combined with the above-mentioned insensitivity, immediately leads to the steady-state distribution of the number of items on the shelf in $(S-1, S)$. For the case without demand, we get this via the argument of having insensitivity and then writing down the following balance equations for the case of *exponential* leadtimes:
the probabilities $p_n := \Pr(K = n)$, with K the number *not* on the shelf, are the solution of the balance equations

$$(S-n)\xi p_n = \frac{n+1}{\tau} p_{n+1}, \qquad n = 0, 1, ..., S-1. \qquad (15)$$

Solving (15) with the normalizing condition $\sum_{n=0}^{S} p_n = 1$ we indeed get the binomial distri-

bution: for $n = 0, 1, \ldots, S$,

$$p_n = \left(\begin{array}{c} S \\ n \end{array} \right) \left(\frac{\xi\tau}{\xi\tau + 1} \right)^n \left(\frac{1}{\xi\tau + 1} \right)^{S-n}. \quad (16)$$

Note that we could also have used the above argument of having S independent alternating renewal processes.

If one allows both sudden death and demand, the balance equations in the exponential case become

$$((S-n)\xi+\lambda)p_n = \frac{n+1}{\tau} p_{n+1}, \quad n = 0, 1, ..., S-1, \quad (17)$$

whose solution is given by

$$p_n = \tau^n p_0 \prod_{i=1}^{n} \frac{(S-n+i)\xi + \lambda}{i},$$

where

$$p_0 = \left[\sum_{n=0}^{S} \frac{\tau^n}{n!} \prod_{i=1}^{n} ((S-n+i)\xi + \lambda) \right]^{-1}.$$

However, due to the insensitivity property the above result still holds if the leadtimes have a general distribution with mean τ.

5. Conclusions and suggestions for further research

In this paper we have explored, and exploited, the relationship between the $(S-1, S)$ inventory model and three well-known queueing models: the Erlang loss system, the machine-repair model and a two-node Jackson network. This relationship allowed us to obtain performance measures like the density of the virtual outdating time and the distribution of the empty shelf period.

It would be interesting to consider variants and generalizations of $(S-1, S)$, and to investigate whether one can again exploit the relation to queueing models. Examples are (i) perishability or outdating of items, (ii) patience of demands, and (iii) *value of the system* – here one takes the sum of the values of the items according to their ages, young items being more worth than old items. We already briefly commented on (i) and (ii) in Remark 2. The most interesting avenue to explore might be the case of patience of demands in $(S-1, S)$, as it is a very relevant case while there is a rich literature for its multiserver counterpart, the $M/G/S + G$ queue. Another fascinating topic for further research is the applicability of the Level Crossing Technique, as briefly discussed in Remark 5.

Acknowledgment. The authors gratefully acknowledge contributions from Marko Boon and Geert-Jan van Houtum. The research of Onno Boxma was partly funded by the NWO Gravitation Programme NETWORKS, Grant Number 024.002.003.

References

[1] F. Baskett, K.M. Chandy, R.R. Muntz and F.G. Palacios (1975). Open, closed and mixed networks of queues with different classes of customers. *Journal of the Association for Computing Machinery* **22**, 248-260.

[2] T. Bonald (2006). The Erlang model with non-Poisson call arrivals. *Performance Evaluation Review* **34**, 276-286 (Proc. SIGMETRICS '06/Performance '06).

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

[4] J.W. Cohen (1976). *On Regenerative Processes in Queueing Theory*. Springer.

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

[6] F.P. Kelly (2011). *Reversibility and Stochastic Networks*. Cambridge University Press.

[7] J.P. Kharoufeh (2010). The $M/G/s/s$ Queue. *Wiley Encyclopedia of Operations Research and Management Science*.

[8] L. Kleinrock (1976). *Queueing Systems*, Volume 2. Wiley.

[9] H. Kobayashi (1978). *Modeling and Analysis. An Introduction to System Performance Evaluation Methodology*. Addison-Wesley.

[10] A.A. Kranenburg and G.J. van Houtum (2007). Cost optimization in the $(S-1, S)$ lost sales inventory model with multiple demand classes. *Operations Research Letters* **35**, 493-502.

[11] S. Nahmias, D. Perry and W. Stadje (2004). Actuarial valuation of perishable inventory system. *Probability in the Engineering and Informational Sciences* **18**, 219-232.

[12] D. Perry (1999). Analysis of a sampling control scheme for a perishable inventory system. *Operations Research* **47(6)**, 966-973.

[13] D. Perry and M.J.M. Posner (1998). An $(S-1, S)$ inventory system with fixed shelflife and constant leadtimes. *Operations Research* **46**, S65-S71.

[14] C.P. Schmidt and S. Nahmias (1985). $(S-1, S)$ policies for perishable inventory. *Management Science* **31**, 719-728.

[15] S.A. Smith (1977). Optimal inventories for an $(S-1, S)$ system with no backorders. *Management Science* **23**, 522-528.

[16] R.W. Wolff (1989). *Stochastic Modeling and the Theory of Queues*. Prentice Hall.