A queueing model with dependence between service and interarrival times

Eindhoven University
of Technology

Department of Mathematics
and Computing Sciences

Memorandum COSOR 98-27

A queueing model with dependence between service and interarrival times

O.J. Boxma
D. Perry

Eindhoven, November 1998
The Netherlands

A Queueing Model with Dependence between Service and Interarrival Times

O.J. Boxma[1,2] and D. Perry[3]
[1] *Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands*
[2] *CWI, P.O. Box 94079, 1090 GB Amsterdam, The Netherlands*
[3] *Dept. of Statistics, University of Haifa, 31999 Haifa, Israel*

Abstract

We consider a storage model that can be either interpreted as a certain queueing model with dependence between a service request and the subsequent interarrival time, or as a fluid production/inventory model with a two-state random environment. We establish a direct link between the workload distributions of the queueing model and the production/inventory model, and we present a detailed analysis of the workload and waiting time process of the queueing system.

1991 Mathematics Subject Classification: 60K25, 90B22.
Keywords & Phrases: $M/G/1$ queue, workload, dependence between service and interarrival times.

1 Introduction

In this paper we consider a storage model that can be interpreted either as a certain queueing model with dependence between a service request and the subsequent interarrival time, or as a fluid production/inventory model with a two-state random environment. The two models are directly related in the sense that the steady-state law of the workload in the queueing interpretation can be expressed in terms of the steady-state law of the buffer content in the fluid interpretation. As a stochastic model, the fluid interpretation may be more natural than the queueing one; however, the queueing interpretation enables to locate the problem in the general setting of queueing models, and to use well-established tools and results from queueing theory for the solution.

Under the fluid interpretation, we consider a manufacturing problem incorporating machine reliability and maintenance. We assume that items are

produced continuously and uniformly by a single machine that is subject to breakdown. During a machine ON time, there is a deterministic net flow into the inventory buffer at rate $\alpha > 0$, where α is the production rate minus the demand rate. If a failure occurs before some time T has elapsed, then a machine repair operation will start; this results in an OFF period of type 0. However, if a failure does not occur before T, then the controller stops production to initiate a preventive maintenance action; this results in an OFF period of type 1. During an OFF period of either type 0 or type 1, there is a deterministic demand generating a linear outflow from the inventory buffer at a constant rate that, w.l.o.g., is taken to be 1. It is assumed that negative inventory is not allowed, so there is no outflow whenever the buffer is empty.

Perry and Posner [13] have studied the special case in which the threshold T is a constant and the OFF periods of both types are exponentially distributed. Their analysis is based on the fact (cf. Kella and Whitt [9]) that the conditional steady-state buffer content distribution, given that it is positive, is independent of the inflow rate α. Obviously, the validity of this fact holds for the more general case - to be considered in the present paper - in which the threshold T is a generally distributed random variable. By setting $\alpha = \infty$, we generate a sample path in which ON periods are deleted and are represented by upward jumps, while the OFF periods are being glued together.

This brings us to the above-mentioned queueing interpretation, as the resulting process can be interpreted as the workload process of the following queueing model. Customers arrive with a service request at a single server. Service requests of successive customers are independent, identically distributed random variables B_i, $i = 1, 2, \ldots$. Upon arrival, the service request is registered. If the service request is less than a threshold T_i, then the next interarrival interval is exponentially distributed with rate λ_0 (this corresponds to an OFF period of type 0); otherwise, the service time becomes exactly equal to T_i (is cut off at T_i), and the next interarrival interval is exponentially distributed with rate λ_1.

In this paper we concentrate on the queueing interpretation. We present a detailed analysis of the joint distribution of workload and 'state of the arrival process', as well as of the waiting time distribution. We refer to Kella and Whitt [9] for a discussion of the equivalence relation between the workload process of the $GI/G/1$ queue and the buffer content process of a fluid model with linear flow and a two-state random environment. See [8]

for an extension to the case of non-linear flow, and [3] for several equivalence relationships in the case of a three-state random environment. For related work on production and manufacturing fluid models, in which equivalence relations with queueing models are studied, we refer to Meyer, Rothkopf and Smith [11, 12], Chen and Yao [4], and Vanneste and Van der Duyn Schouten [15].

According to the original production/inventory interpretation, our model is motivated as a replacement model with preventive maintenance. For a detailed survey on replacement policies and preventive maintenance we refer to Beichelt [1] and Voldes-Flores and Feldman [16].

According to the queueing interpretation, our model is a correlated single server queue with Markov arrivals. There are quite a few studies on single server queues in which the service time of a customer depends on the previous interarrival time; see, e.g., Borst et al. [2] that also contains an extensive list of references. But in the present case, as in [13] mentioned above, it is the interarrival time that depends on the service request of the previous arrival. In principle, this kind of dependence is allowed in the Lindley approach to the waiting time process in the $G/G/1$ queue (see Cohen [6], Section II.6.3). However, the resulting Wiener-Hopf decomposition in general does not yield a very explicit solution – whereas in the present case such an explicit solution is indeed obtained. See Cidon et al. [5] for a detailed analysis of a $\cdot/M/1$ queue in which the interarrival time depends *linearly* on the service time of the previous customer.

The paper is organized as follows. In Section 2 we present the queueing model in detail. The workload process is analysed in Section 3, and the waiting time process in Section 4. Section 5 contains conclusions and some suggestions for further research.

2 Model formulation

We consider the following queueing model. Customers arrive with a service request at a single server. Service requests of successive customers are independent, identically distributed (i.i.d.) random variables B_i, $i = 1, 2, \ldots$ with distribution $B(\cdot)$, mean β and Laplace-Stieltjes transform (LST) $\beta(\cdot)$. Upon arrival, the service request is registered. If the service request B_i is less than a threshold T_i, then the next interarrival interval is exponentially distributed

with rate λ_0; otherwise, the service *time* becomes exactly equal to T_i (is cut off at T_i), and the next interarrival interval is exponentially distributed with rate λ_1.

Perry and Posner [13] have studied this model in the case of a deterministic threshold T. In the production/inventory application that motivated their study, the threshold is not necessarily deterministic. In the present study, we assume the threshold T_i to be i.i.d. random variables with general distribution $T(\cdot)$ that has mean τ and LST $\tau(\cdot)$. In the sequel, B (T) shall denote a generic service request (threshold) with distribution $B(\cdot)$ ($T(\cdot)$). It will turn out that a detailed analysis of the steady-state workload process, and of the waiting time, is possible. The workload analysis will be provided in Section 3, and the waiting time analysis in Section 4. We close the present section with some useful additional notation, and two observations. Define, for Re $s \geq 0$:

$$\chi(s) := \mathrm{E}[\mathrm{e}^{-sB}(B < T)] = \int_{x=0}^{\infty} \mathrm{e}^{-sx}(1 - T(x))\mathrm{d}B(x), \tag{2.1}$$

$$\psi(s) := \mathrm{E}[\mathrm{e}^{-sT}(T < B)] = \int_{x=0}^{\infty} \mathrm{e}^{-sx}(1 - B(x))\mathrm{d}T(x). \tag{2.2}$$

Note that

$$\chi(s) + \psi(s) = \mathrm{E}[\mathrm{e}^{-s\min(B,T)}].$$

It immediately follows that

$$\mathrm{E}[\min(B,T)] = -\chi'(0) - \psi'(0).$$

Our first observation concerns the ergodicity condition. This condition is simply that $\mathrm{E}[\min(B,T)]$ is less than the mean interarrival time $\mathrm{E}A$, where A denotes a generic interarrival time:

$$-\chi'(0) - \psi'(0) < \frac{1}{\lambda_0}\chi(0) + \frac{1}{\lambda_1}\psi(0). \tag{2.3}$$

For the waiting time, this follows immediately from Lindley's theorem for the ordinary single server queue ([10], cf. also [6], Section II.1.3), as this theorem applies even if an interarrival time depends on the previous service time.

Our second observation is, that interarrival times are negatively (positively) correlated with the previous service request if $\lambda_1 > (<)\lambda_0$, as one would expect:

$$\mathrm{E}[BA] = \frac{\beta}{\lambda_1} + (\frac{1}{\lambda_1} - \frac{1}{\lambda_0})\chi'(0),$$

and

$$\mathrm{E}BEA = \beta[\frac{1}{\lambda_0}\chi(0) + \frac{1}{\lambda_1}\psi(0)] \;=\; \beta[\frac{1}{\lambda_1} + (\frac{1}{\lambda_0} - \frac{1}{\lambda_1})\chi(0)],$$

and hence

$$\mathrm{cov}(B, A) \;=\; (\frac{1}{\lambda_1} - \frac{1}{\lambda_0})(\chi(0)\beta + \chi'(0)). \tag{2.4}$$

The result follows since $\chi(0)\beta + \chi'(0) = \mathrm{P}\{B < t\}[\mathrm{E}B - \mathrm{E}[B|B < T]] > 0$.

3 The workload process

For the queueing model described in the previous section, consider the stochastic process $((V(t), J(t)); t \geq 0)$, with $V(t)$ denoting the workload at time t and $J(t)$ denoting the status of the interarrival interval at time t: $J(t) = i$ if the interarrival interval is exponentially distributed with rate λ_i, $i = 0, 1$. Note that the above stochastic process is a Markov process. We assume that the ergodicity condition (2.3) is satisfied. Consider the steady-state distribution

$$F_i(x) \;:=\; \lim_{t\to\infty}\mathrm{P}\{V(t) < x, J(t) = i\}, \quad x \geq 0,\; i = 0, 1. \tag{3.1}$$

Using level crossing theory [7], or using the integro-differential method of Tákacs [14] (cf. also [6], Section II.4.5), we obtain for $x > 0$:

$$\begin{aligned}\frac{dF_0(x)}{dx} \;=\; \lambda_0 F_0(x) \;&-\; \lambda_0 \int_{y=0}^{x}\int_{b=0}^{x-y}(1 - T(b))\mathrm{d}B(b)\mathrm{d}F_0(y) \\ &-\; \lambda_1 \int_{y=0}^{x}\int_{b=0}^{x-y}(1 - T(b))\mathrm{d}B(b)\mathrm{d}F_1(y),\end{aligned} \tag{3.2}$$

$$\begin{aligned}\frac{dF_1(x)}{dx} \;=\; \lambda_1 F_1(x) \;&-\; \lambda_0 \int_{y=0}^{x}\int_{z=0}^{x-y}(1 - B(z))\mathrm{d}T(z)\mathrm{d}F_0(y) \\ &-\; \lambda_1 \int_{y=0}^{x}\int_{z=0}^{x-y}(1 - B(z))\mathrm{d}T(z)\mathrm{d}F_1(y).\end{aligned} \tag{3.3}$$

Let

$$\phi_i(s) \;:=\; \int_{0-}^{\infty} \mathrm{e}^{-sx}\mathrm{d}F_i(x), \quad \mathrm{Re}\; s \geq 0,\; i = 0, 1. \tag{3.4}$$

It follows from (3.2) and (3.3) that, for Re $s \geq 0$ (note that we assume that $\mathrm{P}(B > 0) = \mathrm{P}(T > 0) = 1$; otherwise a minor change is required),

$$\begin{aligned}
\phi_0(s) - F_0(0+) &= \lambda_0 \frac{\phi_0(s)}{s} - \lambda_0 \frac{\phi_0(s)}{s}\chi(s) - \lambda_1 \frac{\phi_1(s)}{s}\chi(s), \\
\phi_1(s) - F_1(0+) &= \lambda_1 \frac{\phi_1(s)}{s} - \lambda_0 \frac{\phi_0(s)}{s}\psi(s) - \lambda_1 \frac{\phi_1(s)}{s}\psi(s).
\end{aligned}$$

Eliminating $\phi_1(s)$ from these two equations, it follows that

$$\begin{aligned}
\phi_0(s)[s - \lambda_0 + \lambda_0\chi(s)] &= sF_0(0+) - \\
&- \lambda_1\chi(s)\frac{sF_1(0+) - \lambda_0\psi(s)\phi_0(s)}{s - \lambda_1 + \lambda_1\psi(s)},
\end{aligned} \tag{3.5}$$

or, for Re $s \geq 0$:

$$\begin{aligned}
&\phi_0(s)[s - \lambda_0 + \lambda_0\chi(s) - \lambda_0\lambda_1\frac{\chi(s)\psi(s)}{s - \lambda_1 + \lambda_1\psi(s)}] \\
&= sF_0(0+) - \frac{\lambda_1\chi(s)sF_1(0+)}{s - \lambda_1 + \lambda_1\psi(s)}.
\end{aligned} \tag{3.6}$$

Similarly, for Re $s \geq 0$:

$$\begin{aligned}
&\phi_1(s)[s - \lambda_1 + \lambda_1\psi(s) - \lambda_0\lambda_1\frac{\chi(s)\psi(s)}{s - \lambda_0 + \lambda_0\chi(s)}] \\
&= sF_1(0+) - \frac{\lambda_0\psi(s)sF_0(0+)}{s - \lambda_0 + \lambda_0\chi(s)}.
\end{aligned} \tag{3.7}$$

After some calculations, in which s is factored out in numerator and denominator, we rewrite (3.6) and (3.7) into: For Re $s \geq 0$,

$$\phi_0(s) = \frac{(s - \lambda_1 + \lambda_1\psi(s))F_0(0+) - \lambda_1\chi(s)F_1(0+)}{s - \lambda_0 - \lambda_1 + \lambda_0\chi(s) + \lambda_1\psi(s) + \lambda_0\lambda_1\frac{1-\chi(s)-\psi(s)}{s}}, \tag{3.8}$$

$$\phi_1(s) = \frac{(s - \lambda_0 + \lambda_0\chi(s))F_1(0+) - \lambda_0\psi(s)F_0(0+)}{s - \lambda_0 - \lambda_1 + \lambda_0\chi(s) + \lambda_1\psi(s) + \lambda_0\lambda_1\frac{1-\chi(s)-\psi(s)}{s}}. \tag{3.9}$$

Let us now determine the constants $F_0(0+)$ and $F_1(0+)$. Obviously, we have the normalizing equation

$$F_0(\infty) + F_1(\infty) = 1, \tag{3.10}$$

i.e.,

$$\phi_0(0) + \phi_1(0) = 1. \tag{3.11}$$

Substitution of $s = 0$ in (3.8) and (3.9) yields, after a straightforward calculation:

$$F_0(0+) + F_1(0+) = \frac{\lambda_0\psi(0) + \lambda_1\chi(0) + \lambda_0\lambda_1[\chi'(0) + \psi'(0)]}{\lambda_0\psi(0) + \lambda_1\chi(0)}. \tag{3.12}$$

Remark 3.1. Note that $F_0(0+) + F_1(0+)$ is positive iff the ergodicity condition (2.3) holds. In fact, dividing both numerator and denominator of the righthand side of (3.12) by $\lambda_0\lambda_1$ shows that $F_0(0+) + F_1(0+) = 1 - \mathrm{E}[\min(B,T)]/\mathrm{E}A$.

A second equation for $F_0(0+)$ and $F_1(0+)$ is obtained by observing that the denominator of the righthand side of (3.8) (which coincides with the denominator of the righthand side of (3.9)) has exactly one zero in Re $s > 0$ (we outline the proof in Remark 3.2 below). Call this zero σ. Since the LST's $\phi_0(s)$ and $\phi_1(s)$ are analytic functions for Re $s > 0$, the numerators of the righthand sides of (3.8) and (3.9) must also be zero for $s = \sigma$. In both cases, this yields the same relation between $F_0(0+)$ and $F_1(0+)$ (the equality of these two relations is most easily seen by observing that the lefthand side of (3.6), and of (3.7), is zero for $s = \sigma$):

$$F_0(0+) = \frac{\sigma - \lambda_0 + \lambda_0\chi(\sigma)}{\lambda_0\psi(\sigma)}F_1(0+) = \frac{\lambda_1\chi(\sigma)}{\sigma - \lambda_1 + \lambda_1\psi(\sigma)}F_1(0+). \tag{3.13}$$

This completes the determination of $\phi_i(s)$, $i = 0, 1$.

Remark 3.2. The fact that the denominator of the righthand side of (3.8), to be called $h(s)$, has exactly one zero in Re $s > 0$ can be shown by application of Rouché's theorem to $h(s)$. Write $h(s) = h_1(s)+h_2(s)$, with $h_1(s) := s-\lambda_0-\lambda_1$ and $h_2(s) := \lambda_0\chi(s) + \lambda_1\psi(s) + \lambda_0\lambda_1\frac{1-\chi(s)-\psi(s)}{s}$. It is useful to observe that $\frac{1-\chi(s)-\psi(s)}{s\mathrm{E}[\min(B,T)]}$ is the LST of the *residual* service time, where service time is $\min(B,T)$. Take R to be a closed contour, consisting of the imaginary axis from $-ir$ to $+ir$ and a semi-circle in the right halfplane with radius r and origin O; we'll let $r \to \infty$. We observe that $h_1(s)$ and $h_2(s)$ are analytic inside R, and that $h_1(s)$ has exactly one zero inside R for r large enough.

On the boundary, $|h_1(s)| > |h_2(s)|$. This is obviously true on the semi-circle; on the imaginary axis, one has $|h_1(s)| \geq \lambda_0 + \lambda_1$ and

$$\begin{aligned} |h_2(s)| &\leq |\lambda_0\chi(0)| + |\lambda_1\psi(0)| + \lambda_0\lambda_1 \mathrm{E}[\min(B,T)] \\ &< \lambda_0\chi(0) + \lambda_1\psi(0) + (\lambda_1\chi(0) + \lambda_0\psi(0)) = \lambda_0 + \lambda_1, \end{aligned}$$

the last inequality following from the ergodicity condition (2.3). Rouché's theorem now implies that $h(s)$ has for Re $s > 0$ just as many zeros as $h_1(s)$, viz., one zero. Note that, on the imaginary axis, $s = 0$ is a removable singularity.

The unique zero σ is real and lies between λ_0 and λ_1. Indeed, substitution of $s = \lambda_0$ and $s = \lambda_1$ in $h(s)$ immediately shows that, since $h(\lambda_0) = (\lambda_0 - \lambda_1)\chi(\lambda_0)$ and $h(\lambda_1) = (\lambda_1 - \lambda_0)\psi(\lambda_1)$:

$$\min(\lambda_0, \lambda_1) \leq \sigma \leq \max(\lambda_0, \lambda_1).$$

Remark 3.3. If the threshold is infinite, then $\psi(s) \equiv 0$ and $\chi(s) \equiv \beta(s)$. It is not hard to check, using (3.9), that then $F_1(0+) = 0$ (e.g., by substituting $s = \sigma$). So $\phi_1(s) = 0$, and $\phi_0(s)$ reduces to the Pollaczek-Khintchine expression $sF_0(0+)/(s - \lambda_0 + \lambda_0\beta(s))$. Hence the system now behaves like an $M/G/1$ queue with arrival rate λ_0 and service time distribution $B(\cdot)$ – as it should.

Summation of the two expressions (3.8) and (3.9) for $\phi_0(s)$ and $\phi_1(s)$ gives the LST of the distribution of the steady-state workload V. We present this result, plus an expression for the mean workload, in the next theorem. Remember that $h(s)$ is the denominator of (3.8) and (3.9).

Theorem 3.1. For Re $s \geq 0$,

$$\mathrm{E}[\mathrm{e}^{-sV}] = \frac{[s - \lambda_1 + (\lambda_1 - \lambda_0)\psi(s)]F_0(0+) + [s - \lambda_0 + (\lambda_0 - \lambda_1)\chi(s)]F_1(0+)}{s - \lambda_0 - \lambda_1 + \lambda_0\chi(s) + \lambda_1\psi(s) + \lambda_0\lambda_1\frac{1-\chi(s)-\psi(s)}{s}}; \tag{3.14}$$

$$\mathrm{E}[V] = -\frac{[1 + (\lambda_1 - \lambda_0)\psi'(0)]F_0(0+) + [1 + (\lambda_0 - \lambda_1)\chi'(0)]F_1(0+)}{h(0)} + \frac{h'(0)}{h(0)}. \tag{3.15}$$

Remark 3.4. For $\lambda_0 = \lambda_1$, it is easily checked that (3.14) reduces to the LST of the steady-state workload distribution in the $M/G/1$ queue with arrival rate $\lambda_0 = \lambda_1$ and service time LST $\chi(s) + \psi(s)$, which is the LST of

$\min(B,T)$.
Note that one can easily express $\mathrm{E}[V]$ into the model parameters. For example,

$$\begin{aligned} h(0) &= -\lambda_0 - \lambda_1 + \lambda_0\chi(0) + \lambda_1\psi(0) + \lambda_0\lambda_1\mathrm{E}[\min(B,T)] \\ &= \lambda_0\lambda_1[\mathrm{E}[\min(B,T)] - \mathrm{E}A]. \end{aligned} \tag{3.16}$$

Remark 3.5. As mentioned in the Introduction, the workload process of the queueing model and the content process of the production/inventory fluid model have the same conditional steady-state law, given that the processes are positive. It is also easy to see that both processes are regenerative processes. This enables one to express the steady-state law of each process in terms of the steady-state law of the other process. More specifically, let $F_W(\cdot)$ and $F_C(\cdot)$ denote the steady-state distributions of the workload process and the content process, respectively. Then

$$F_W(x) = \pi_W + \int_0^x f_W(y)\mathrm{d}y, \quad F_C(x) = \pi_C + \int_0^x f_C(y)\mathrm{d}y, \tag{3.17}$$

where π_W and π_C are the respective atoms at 0 of the workload and the content process, and $f_W(\cdot)$ and $f_C(\cdot)$ are the respective absolutely continuous densities. The fact that the conditional steady-state laws are the same simply says that for all $x > 0$:

$$\frac{f_W(x)}{1-\pi_W} = \frac{f_C(x)}{1-\pi_C}. \tag{3.18}$$

Since both processes are regenerative, we have:

$$\pi_W = \frac{\mathrm{E}[\text{idle period}]}{\mathrm{E}[\text{idle period}] + \mathrm{E}[\text{busy period}]},$$

$$\pi_C = \frac{\mathrm{E}[\text{silence period}]}{\mathrm{E}[\text{silence period}] + \mathrm{E}[\text{activity period}]},$$

where by 'idle period' and 'busy period' we mean that the queueing system is empty and not empty, respectively. Similarly, the 'silence period' and the 'activity period' are the time periods in which the production/inventory system is empty and not empty, respectively.

By the construction of the workoad process from the content process,

$$\mathrm{E}[\text{idle period}] = \mathrm{E}[\text{silence period}],$$

and (remember that α is the net inflow rate during machine ON times)

$$(1+\alpha)\mathrm{E}[\text{busy period}] \;=\; \mathrm{E}[\text{activity period}],$$

so that

$$f_C(x) \;=\; f_W(x)(1+\alpha)\frac{\mathrm{E}[\text{idle period}] \;+\; \mathrm{E}[\text{busy period}]}{\mathrm{E}[\text{idle period}] \;+\; (1+\alpha)\mathrm{E}[\text{busy period}]}. \qquad (3.19)$$

4 The waiting time

Let W_n denote the waiting time of the nth arriving customer, $n = 1, 2, \ldots$. In the present section we determine the steady-state waiting time distribution, by using a Wiener-Hopf approach. This is a classical approach for the $G/G/1$ queue, cf. Lindley [10]; see also Section II.6.3 of [6]. A nice feature of the model under consideration is, that the Wiener-Hopf decomposition can be worked out in detail here, leading to quite explicit results.

Starting-point is the following recurrence relation (below, $[x]^+$ denotes $\max(0,x)$ and $[x]^-$ denotes $\min(0,x)$):

$$\begin{aligned} W_{n+1} &= [W_n + B_n - A_{n+1}^{(0)}]^+ \quad \text{if } B_n < T_n, \\ &= [W_n + T_n - A_{n+1}^{(1)}]^+ \quad \text{if } B_n \geq T_n. \end{aligned} \qquad (4.1)$$

Here $A_{n+1}^{(i)}$ denotes an interarrival interval that is exponentially distributed with rate λ_i, $i = 0, 1$. Taking LST's and using the identity

$$\mathrm{e}^{-s[x]^+} + \mathrm{e}^{-s[x]^-} \;=\; 1 + \mathrm{e}^{-sx},$$

we obtain for $\mathrm{Re}\, s = 0$:

$$\begin{aligned} &\mathrm{E}[\mathrm{e}^{-sW_{n+1}}] = 1 - \mathrm{E}[\mathrm{e}^{-s[W_n+B_n-A_{n+1}^{(0)}]^-}(B_n < T_n)] \\ -\;&\mathrm{E}[\mathrm{e}^{-s[W_n+T_n-A_{n+1}^{(1)}]^-}(B_n \geq T_n)] \\ +\;&\mathrm{E}[\mathrm{e}^{-s(W_n+B_n-A_{n+1}^{(0)})}(B_n < T_n)] + \mathrm{E}[\mathrm{e}^{-s(W_n+T_n-A_{n+1}^{(1)})}(B_n \geq T_n)]. \end{aligned} \qquad (4.2)$$

Now we assume that the ergodicity condition (2.3) holds, and we restrict ourselves to the steady-state waiting time distribution. W shall denote a

random variable with this steady-state distribution. After a brief calculation, Formula (4.2) leads to the following identity: For Re $s = 0$,

$$\mathrm{E}[\mathrm{e}^{-sW}] \;=\; \zeta^{-}(s) + \mathrm{E}[\mathrm{e}^{-sW}]\frac{\lambda_0}{\lambda_0 - s}\chi(s) + \mathrm{E}[\mathrm{e}^{-sW}]\frac{\lambda_1}{\lambda_1 - s}\psi(s), \qquad (4.3)$$

where

$$\zeta^{-}(s) \;:=\; 1 - \mathrm{E}[\mathrm{e}^{-s[W+B-A^{(0)}]^{-}}(B < T)] - \mathrm{E}[\mathrm{e}^{-s[W+T-A^{(1)}]^{-}}(B \geq T)].$$

Note that $\zeta^{-}(s)$ is analytic in Re $s \leq 0$.

Remembering the definition of $h(s)$ in Remark 3.2 – note that $h(s)$ is the denominator of the righthand side of both (3.8) and (3.9) – we can rewrite (4.3) into:

$$h(s)\mathrm{E}[\mathrm{e}^{-sW}] \;=\; \frac{(s-\lambda_0)(s-\lambda_1)}{s}\zeta^{-}(s), \quad \text{Re } s = 0. \qquad (4.4)$$

We now have the appropriate formulation for a Wiener-Hopf boundary value problem. The lefthand side of (4.4) is bounded and analytic in Re $s > 0$, and the righthand side of (4.4) is bounded and analytic in Re $s < 0$, and both sides are equal on the boundary Re $s = 0$. Note that $s = 0$ is a removable singularity of the righthand side. In view of the linear behaviour in s at infinity, Liouville's theorem implies that the lefthand side should equal a first-order polynomial, say $as + b$, for Re $s \geq 0$. In fact the same holds for the righthand side in Re $s \leq 0$, but we are not interested in determining $\zeta^{-}(s)$. Remembering that $s = \sigma$ is the only zero of $h(s)$ in Re $s > 0$, cf. Remark 3.2, it follows that $b = -a\sigma$. Substitution of $s = 0$ yields that

$$h(0).1 \;=\; -a\sigma,$$

so (see (3.16))

$$a \;=\; \frac{\lambda_0\lambda_1}{\sigma}(\mathrm{E}A - \mathrm{E}[\min(B,T)]). \qquad (4.5)$$

Finally we obtain the following expression for the LST of the steady-state waiting time distribution:

Theorem 4.1. For Re $s \geq 0$,

$$\mathrm{E}[\mathrm{e}^{-sW}] \;=\; \frac{\frac{\lambda_0\lambda_1}{\sigma}(\mathrm{E}A - \mathrm{E}[\min(B,T)])(s-\sigma)}{s - \lambda_0 - \lambda_1 + \lambda_0\chi(s) + \lambda_1\psi(s) + \lambda_0\lambda_1\frac{1-\chi(s)-\psi(s)}{s}}. \qquad (4.6)$$

$$EW = \frac{-a}{h(0)} + \frac{h'(0)}{h(0)} = \frac{1}{\sigma} + \frac{h'(0)}{h(0)}. \tag{4.7}$$

Remark 4.1. Note that the steady-state distributions of V and W differ; although the arrivals occur at exponentially distributed intervals, PASTA does not hold.

Remark 4.2. The LST's of V and W appear to be relatively simple functions of the LST's $\chi(s)$ and $\psi(s)$. The latter functions are easily expressed in the LST of T when service requests are exponentially distributed, and in the LST of B when thresholds are exponentially distributed. In the former case,

$$\chi(s) = \frac{1 - \tau(s + 1/\beta)}{1 + \beta s}, \quad \psi(s) = \tau(s + 1/\beta), \quad \text{Re } s \geq 0,$$

with $\tau(\cdot)$ the LST of the threshold distribution; in the latter case,

$$\chi(s) = \beta(s + 1/\tau), \quad \psi(s) = \frac{1 - \beta(s + 1/\tau)}{1 + \tau s}, \quad \text{Re } s \geq 0,$$

with $\beta(\cdot)$ the LST of the service request distribution.

5 Conclusions and suggestions for further research

In this paper we have presented a detailed analysis of the workload and waiting time process of a queueing model with dependence between a service request and the subsequent interarrival time. We have also established a link between this model and a certain production/inventory model, that provided the initial motivation for the present study.

We have taken exponentially distributed interarrival intervals, with rate depending on the previous service request. It should be possible to extend the analysis of Sections 3 and 4 to the case of interarrival intervals with distributions that have a rational LST; cf. the analysis of the $K_m/G/1$ queue in Section II.5.11 of [6]. For example, in (4.3) the terms $\lambda_i/(\lambda_i - s)$ then have

to be replaced by more complicated quotients of polynomials, after which all the zeros of the resulting new function $h(s)$ in Re $s > 0$ must be determined.

The results of the present study might be used for optimization purposes. In the production/inventory version of the model, e.g., the goal could be to choose the production rate $1 + \alpha$ such that some cost function is minimized. E.g., with respect to the choice of α there should be a trade-off between holding costs (which are linear in the mean buffer content and then increasing in α) and unsatisfied demand costs (which are linear in the probability π_C of having an empty buffer, and then decreasing in α).

Acknowledgement. Financial support of the Dutch research organization NWO, that has enabled the visit of David Perry to The Netherlands, is gratefully acknowledged. The authors are indebted to Professor J.W. Cohen for a very interesting discussion.

References

[1] F. Beichelt (1993). *A unifying treatment of replacement policies with minimal repair*, Naval Res. Logist. Quarterly **40**, 51-67.

[2] S.C. Borst, O.J. Boxma and M.B. Combé (1993). *An $M/G/1$ queue with customer collection*, Stochastic Models **9**, 341-371.

[3] O.J. Boxma, D. Perry and F.A. van der Duyn Schouten (1998). *Fluid queues and mountain processes*, Report in preparation.

[4] H. Chen and D.D. Yao (1992). *A fluid model for systems with random disruptions*, Operations Research **S40**, S239-S247.

[5] I. Cidon, R. Guérin, A. Khamisy and M. Sidi (1991). *On queues with inter-arrival times proportional to service times*, Tech. Rep. Technion EE PUB 811.

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

[7] B.T. Doshi (1992). *Level-crossing analysis of queues*, In: U.N. Bhat and I.V. Basawa (eds), *Queueing and Related Models* (Clarendon Press, Oxford) pp. 3-33.

[8] H. Kaspi, O. Kella and D. Perry (1996). *Dam processes with state dependent batch sizes and intermittent production processes with state dependent rates*, Queueing Systems **24**, 37-57.

[9] O. Kella and W. Whitt (1992). *A storage model with a two-state random environment*, Operations Research **S40**, S257-S262.

[10] D.V. Lindley (1952). *The theory of queues with a single server*, Proc. Cambridge Phil. Soc. **48**, 277-289.

[11] R.R. Meyer, M.H. Rothkopf and S.A. Smith (1979). *Reliability and inventory in a production-storage system*, Management Science **25**, 799-807.

[12] R.R. Meyer, M.H. Rothkopf and S.A. Smith (1983). *Erratum to reliability and inventory in a production-storage system*, Management Science **29**, 1346.

[13] D. Perry and M.J.M. Posner (1997). *A production/inventory model with random repair and maintenance modes*, Working paper, submitted for publication.

[14] L. Tákacs (1962). *Introduction to the Theory of Queues* (Oxford University Press, New York).

[15] S.G. Vanneste and F.A. van der Duyn Schouten (1995). *Maintenance optimization of a production system with buffer capacity*, Eur. J. Oper. Res. **32**, 323-338.

[16] C. Voldes-Flores and R.M. Feldman (1989). *A survey of preventive maintenance models for stochastically deteriorating single-unit systems*, Naval Res. Logist. Quarterly **36**, 419-446.