Queues with workload-dependent arrival and service rates

TU/e technische universiteit eindhoven

SPOR-Report 2003-11

Queues with workload-dependent arrival and service rates

R. Bekker
S.C. Borst
O.J. Boxma
O. Kella

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven, April 2003
The Netherlands

/department of mathematics and computing science

SPOR-Report
Reports in Statistics, Probability and Operations Research

Eindhoven University of Technology
Department of Mathematics and Computing Science
Probability theory, Statistics and Operations research
P.O. Box 513
5600 MB Eindhoven - The Netherlands

Secretariat: Main Building 9.10
Telephone: + 31 40 247 3130
E-mail: wscosor@win.tue.nl
Internet: http://www.win.tue.nl/math/bs/cosor.html

ISSN 1567-5211

Queues with Workload-Dependent Arrival and Service Rates

R. Bekker [a,b*], S.C. Borst [a,b,c], O.J. Boxma [a,b], O. Kella [d,†]

[a] CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands

[b] Department of Mathematics & Computer Science
Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands

[c] Bell Laboratories, Lucent Technologies
P.O. Box 636, Murray Hill, NJ 07974, USA

[d] Department of Statistics
The Hebrew University of Jerusalem
Mount Scopus, Jerusalem 91905, Israel

Abstract

We consider two types of queues with workload-dependent *arrival rate* and *service speed*. Our study is motivated by queueing scenarios where the arrival rate and/or speed of the server depends on the amount of work present, like production systems and the Internet.

First, in the $M/G/1$ case, we compare the steady-state distribution of the workload (both at arbitrary epochs and at arrival instants) in two models, in which the ratio of arrival rate and service speed is equal. Applying level crossing arguments, we show that the steady-state distributions are proportional. Second, we consider a $G/G/1$-type queue with workload-dependent interarrival times and service speed. Using a stochastic mean value approach, several well-known relations for the workload at various epochs in the ordinary $G/G/1$ queue are generalized.

*Supported by a research grant from Philips Electronics.

†Supported by grant 935/0 from the Israel Science Foundation. The authors also acknowledge the support of EURANDOM and of INTAS under project M-265

1 Introduction

In queueing systems, the speed of the server often depends on the amount of work present. This is particularly true if the server does not represent a machine, but rather a human being. For example, Bertrand and Van Ooijen [3] describe a production system where the speed of the server is relatively low when there is much work (stress) or when there is very little work (laziness). In addition, the rate at which jobs arrive at the service system may also depend on the amount of work present. In the human-server example, we may try to control the arrival of jobs to optimize server performance. In packet-switched communication systems, the transmission rate of data connections may be dynamically adapted based on the buffer content, see for instance [10, 11, 18, 21]. In particular, feedback information on the buffer state provides the basis for the Transmission Control Protocol (TCP) to carefully regulate the transmission rate of Internet flows.

These considerations led us to study single-server queues with state-dependent interarrival times and general (state-dependent) service speed. In the first part of this paper, customers are assumed to arrive at the queueing system according to a Poisson process, where the arrival rate depends on the workload. The service requirement of a customer is generally distributed, and work is depleted according to a general release rate function that also depends on the workload. In the second part, the Markovian case is extended to the regenerative case: we consider a similar model, however with general interarrival times, which may depend on the amount of work present.

In classical queueing systems, the speed of the server and the arrival rate of customers are usually assumed to be constant. In such systems, the Markovian case amounts to an ordinary $M/G/1$ queue, whereas the regenerative case represents the classical $G/G/1$ queue. Furthermore, the workload process in the Markovian model with general release rule constitutes a dam process. In fact, dams with general release rate function were studied before queues with general service speed drew attention, see Prabhu [20] for an overview of dam studies up to 1965. Dams with general release and Poisson inputs were studied by Asmussen [1], Cohen [6], Gaver and Miller [12], Harrison and Resnick [13], and many others. For dams with more general input processes, see Cohen and Rubinovitch [7], Kaspi *et al.* [15], and references therein.

The two main goals of our study are the following. (i) To establish relationships between two queueing models with arrival rates $\lambda_i(x)$ and release rates $r_i(x)$, $i = 1, 2$, for which $\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}$, $\forall x > 0$. Such relationships will allow us to obtain results for a whole class of models from the analysis of one particular model. (ii) To extend relations between the steady-state workload and the workload at arrival times (waiting times) for the $GI/G/1$ queue to queues with workload-dependent arrival rates and service speeds. We now discuss these two aspects in slightly greater detail.

Ad (i). We consider two related dams, or $M/G/1$ queues, with general (state-dependent) arrival rate and service speed. We show that the workload distributions are proportional and argue that the difference between the two models is just a rescaling of time. A similar result holds for the workload just before arrival instants

- a quantity that does not necessarily equal the waiting time when the service speed is workload-dependent. The derivation of the proportionality relations is partly based on level crossing arguments that lead to a Volterra integral equation of the second kind. These insights also provide an important tool in determining the steady-state workload distribution in an individual model. It turns out that a release rate function $r(\cdot)$ allowing the possibility of an empty system plays a crucial role.
Ad (ii). The $G/G/1$ queue with state-dependent release requires a different method. Using a Palm theoretic approach, we establish some general relations between the workload just before arrival instants and the workload at arbitrary time epochs. In the case of Poisson arrivals, we generalize the PASTA property for a continuous-state Markov process. Moreover, various well-known relations for ordinary $G/G/1$-type queues are extended to queues with general release rate.
This paper is organized as follows. In Section 2, we introduce the $M/G/1$-type model with state-dependent arrival rate and service speed and consider the level crossing equations. In Section 3, we present the proportionality relations with respect to the workload process when we consider two related $M/G/1$-type queues. The steady-state densities of some special cases are determined explicitly in Section 4. In Section 5, we present several relations between the steady-state workload and the workload just before arrival instants for the $G/G/1$ queue with state-dependent release. Finally, Section 6 contains conclusions and suggestions for further research.

2 Model description and preliminaries

In this section, we introduce some notation for the $M/G/1$-type queue with general state-dependent arrival rate and service speed. The level crossing equation is stated for this particular model and some special attention is paid to the case that the workload process has no atom at state 0.

Model description
Consider a Markovian workload process with the following dynamics. Between arrivals, the server serves according to some workload-dependent service rate function $r(x)$. Arrivals are governed by a workload-dependent rate function $\lambda(x)$. More precisely, let V_t be the workload at time t and W_n be the workload immediately before the n-th arrival epoch. Given that the workload at time t_0 is w and the next arrival is at time $t_1 > t_0$, the workload process during the interval (t_0, t_1) behaves as $V_{t_0+t} = w - \int_{t_0}^{t_0+t} r(V_s)\mathrm{d}s$ (a deterministic process). If A is distributed like the time until the next arrival (starting from initial workload w), then $\mathbb{P}_w(A > t) = e^{-\int_0^t \lambda(V_s)\mathrm{d}s}$, meaning that the hazard rate function of A (its density divided by the tail) at t is given by $\lambda(V_t)$. We assume that $\lambda(\cdot)$ is nonnegative, left-continuous, and has a right limit on $[0, \infty)$. Also, we assume that $r(0) = 0$ and that $r(\cdot)$ is strictly positive, left-continuous, and has a strictly positive right limit on $(0, \infty)$. Each arrival increases the workload by some positive amount (job size), where these amounts form a sequence of i.i.d. random variables $B_1, B_2, \ldots$. The random variables B_i have

distribution function $B(\cdot)$, with mean β, and Laplace-Stieltjes Transform (LST) $\beta(\cdot)$.
Throughout, we assume that the workload process is ergodic and has a stationary distribution. In order to prevent a general drift to infinity, the rate functions must satisfy $\limsup_{x\to\infty} \beta\frac{\lambda(x)}{r(x)} < 1$ (see also Cohen [5] and Gaver and Miller [12]). Next, let the steady-state random variables of V_t and W_n be denoted by V and W, and let $v(\cdot)$, $w(\cdot)$ denote their densities.

Define

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

representing the time required to move from state x down to state z in the absence of any arrivals. Of particular interest is $R(x) := R(x,0)$ representing the time required for a workload x to drain, again in the absence of arrivals. A related quantity is

$$\tilde{R}(x,z) := \int_z^x \frac{\lambda(y)}{r(y)}\mathrm{d}y, \qquad 0 \le z < x < \infty.$$

In particular, $\tilde{R}(x) := \tilde{R}(x,0)$ determines whether or not the workload process has an atom at 0 (see also Asmussen [1], p. 288, in case $\lambda(\cdot)$ is fixed). The case $\tilde{R}(x) < \infty$, for all $0 < x < \infty$, represents the situation that the workload process has an atom at state 0, whereas $\tilde{R}(x) = \infty$, for some $0 < x < \infty$ (and then for all) identifies the case that state 0 cannot be reached by the workload process. We assume that $\int_0^x \lambda(y)\mathrm{d}y$ and $\int_0^x r(y)^{-1}\mathrm{d}y$ cannot be both infinite.

Level crossings
Taking $r(x) \equiv 1$ and $\lambda(x) \equiv \lambda$ results in the ordinary $M/G/1$ queue. The level crossing identity for the workload is well-known in this case, see e.g. Cohen [4, 6]. In $M/G/1$-type queues with time-varying arrival rate, the workload level crossing identity has been obtained by Takács [23], while Hasofer [14] shows some additional properties. The proof proposed by Takács may be extended in a rather straightforward way to queues with workload-dependent service and arrival rates. This results in the following theorem:

Theorem 2.1. *The workload density $v(x)$ exists and satisfies the equation*

$$r(x)v(x) = \lambda(0)V(0)(1-B(x)) + \int_{y=0^+}^{x} (1-B(x-y))\lambda(y)v(y)\mathrm{d}y, \qquad x > 0. \tag{2}$$

This integro-differential equation has the following interpretation. The left-hand side of the equation corresponds to the downcrossing rate through level x, while the right-hand side represents the long-run average number of upcrossings through x from the workload level 0 and workload levels between 0 and x respectively. If the workload process has an atom at state 0, it is obvious that $\{V_t, t \ge 0\}$ is a regenerative process, with arrivals of customers in an empty system as regeneration points. Under the assumption of an ergodic process, the expected cycle length is

finite and it follows by level crossing theory that the workload density is well-defined. With some modification, the result can be extended to workload processes that do not reach state 0 (see e.g. [5] for details). Note that if $\tilde{R}(x) = \infty$, then $V(0) = 0$. However, the level crossing equation still holds, and the first term on the right-hand side of (2) just disappears.

3 Relations between two $M/G/1$ queues

In this section we consider two isolated $M/G/1$ queues with arrival rates $\lambda_i(\cdot)$, release rates $r_i(\cdot)$ and service requirements B_n^i for the n-th customers ($i = 1, 2$). Let $B_1^i, B_2^i, \ldots$ be independent and identically distributed random variables with distribution function $B(\cdot)$, and let $r_i(\cdot), \lambda_i(\cdot)$ have the same analytical properties as $r(\cdot), \lambda(\cdot)$ specified in Section 2. Furthermore, define $\tilde{R}_i(\cdot,\cdot), \tilde{R}_i(\cdot), V_i(\cdot), v_i(\cdot)$, and $w_i(\cdot)$ in a similar way as we defined $\tilde{R}(\cdot,\cdot), \tilde{R}(\cdot), V(\cdot), v(\cdot)$, and $w(\cdot)$ in Section 2. We assume that the two queueing models, to be denoted as Models 1 and 2, are related in the following way:

$$\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}, \qquad \forall x > 0. \tag{3}$$

Note that $\tilde{R}_1(x)$ thus equals $\tilde{R}_2(x)$. As a consequence, the workload process in both models either has an atom at state 0, or does not hit state 0 at all.
We now state the three theorems of this section.

Theorem 3.1. *For all $x > 0$,*

$$\frac{v_1(x)}{v_2(x)} = C\frac{r_2(x)}{r_1(x)},$$

with $C = \frac{\lambda_1(0)V_1(0)}{\lambda_2(0)V_2(0)}$ if $\tilde{R}_i(x) < \infty$ for all $0 < x < \infty$, and $C = 1$ if $\tilde{R}_i(x) = \infty$ for some $0 < x < \infty$.

We now turn to the density $w(\cdot)$. Without proof, we claim that it exists just like $v(\cdot)$ (see Theorem 2 and the end of this section).

Theorem 3.2. *$W_i(0) = \lambda_i(0)V_i(0)/\bar{\lambda}_i, i = 1, 2$, with $\bar{\lambda}_i := \int_{0^+}^{\infty} \lambda_i(x)v_i(x)\mathrm{d}x + \lambda_i(0)V_i(0)$, and for all $x > 0$,*

$$w_i(x) = \frac{1}{\bar{\lambda}_i}\lambda_i(x)v_i(x), \qquad i = 1, 2.$$

Theorem 3.3. *$W_1(0) = W_2(0)$, and for all $x > 0$,*

$$w_1(x) = w_2(x).$$

Remark 3.1. In principle Theorem 3.3 (Theorem 3.1) can be derived from Theorems 3.1 and 3.2 (3.3 and 3.2). To give more insight into the underlying similarities between the two models, we prove each of the three theorems separately.

Remark 3.2. Note that $\lambda_i(x) \equiv \lambda$ would yield the PASTA result that the workload at an arbitrary time and the workload at an arrival epoch have the same distribution. Theorem 3.2 may thus be viewed as a generalization of the PASTA result, however, under the assumption of a continuous-state stationary workload process.

Proof of Theorem 3.1. We apply the level crossing identity to Model i and define $z_i(x) := r_i(x)v_i(x)$, $i = 1, 2$. Then (2) reduces to

$$z_i(x) = \lambda_i(0)V_i(0)(1-B(x)) + \int_{y=0^+}^{x} (1-B(x-y))\frac{\lambda_i(y)}{r_i(y)} z_i(y)\mathrm{d}y, \qquad x > 0. \quad (4)$$

If $\tilde{R}_i(x) = \infty$ for some $0 < x < \infty$, then $V_i(0) = 0$ and the result follows easily. So assume that $\tilde{R}_i(x) < \infty$ for all $0 < x < \infty$.
Observe that (4) is a Volterra integral equation of the second kind. Let its kernel be $K^{(i)}(x, y) := (1 - B(x - y))\frac{\lambda_i(y)}{r_i(y)}$ for $0 < y < x < \infty$ and $K_*^{(i)}(x, 0) := 1 - B(x)$ for $0 < x < \infty$. Notice that due to (3) the kernels of both models are the same and we may drop the index i from our notation. Now define recursively

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

and

$$K_{n*}(x, 0) := \int_{0^+}^{x} K_n(x, y)K_*(y, 0)\mathrm{d}y, \qquad 0 < x < \infty,\ n = 1, 2, \ldots,$$

where $K_1(x, y) := K(x, y)$ and $K_{0*}(x, 0) := K_*(x, 0)$. So the classical successive substitution method for Volterra integral equations gives:

$$\begin{aligned} z_i(x) &= \lambda_i(0)V_i(0)K_*(x, 0) + \int_{0^+}^{x} K(x, y)K_*(y, 0)\lambda_i(0)V_i(0)\mathrm{d}y + \ldots \\ &= K_{0*}(x, 0)\lambda_i(0)V_i(0) + K_{1*}(x, 0)\lambda_i(0)V_i(0) + \ldots \\ &= \lambda_i(0)V_i(0)\sum_{n=0}^{\infty} K_{n*}(x, 0). \end{aligned} \quad (5)$$

Dividing $z_1(x)$ by $z_2(x)$ and substituting $z_i(x) = r_i(x)v_i(x), i = 1, 2$, gives

$$\frac{v_1(x)}{v_2(x)} = \frac{r_2(x)}{r_1(x)}\frac{\lambda_1(0)V_1(0)}{\lambda_2(0)V_2(0)},$$

and we have shown the result. □

The Volterra approach provides a useful tool for determining the workload densities. If $\tilde{R}_i(x) < \infty$ for all $0 < x < \infty$, we can follow an idea of Harrison and Resnick [13], and use the bound $K(x, y) \leq \frac{\lambda(y)}{r(y)}$ to show inductively that $K_{(n+1)*}(x, y) \leq \frac{(\tilde{R}(x,y))^n}{n!}\frac{\lambda(y)}{r(y)}$. Note that the sum in (5) is well-defined and we have a closed-form expression for $z_i(x)$ and thus for $v_i(x)$.

However, if $\tilde{R}_i(x) = \infty$ for some $0 < x < \infty$, then the workload process approaches the state 0, but never reaches it. In this case the integrated kernel, $\int_0^x K(y,0)\mathrm{d}y$, is unbounded and equation (4) is often referred to as a singular integral equation. Solving these equations is very hard in general, and goes beyond the scope of this research (see Linz [17], Ch. 1 and 3.5, Mikhlin [19] Ch. 1 and 3, and Zabreyko *et al.* [24], Ch. 1,6, and 9 for a more detailed discussion). In Section 4 we give the steady-state workload distribution for some special cases.

Let us now give an intuitive explanation of Theorem 3.2, based on a Bayesian argument. In Section 5 we derive a more general result, from which the theorem follows as a special case.

Consider either of the two models and drop the index i from the notation. The probability of having two or more arrivals in a small time interval $(t, t+\Delta)$ is of order $o(\Delta)$. Then, by simple conditioning arguments we have $\mathbb{P}(\text{arrival in } (t, t+\Delta)) = \int_{y=0+}^{\infty} \lambda(y)\Delta v(y)\mathrm{d}y + \lambda(0)\Delta V(0) + o(\Delta)$ and $\mathbb{P}(\text{arrival in } (t, t+\Delta)|V_t > x)\mathbb{P}(V_t > x) = \int_{y=x}^{\infty} \lambda(y)\Delta v(y)\mathrm{d}y + o(\Delta)$. Let us consider the excess probability of the workload at jump epochs,

$$\begin{aligned}
\mathbb{P}(W > x) &= \lim_{\Delta\to 0} \mathbb{P}(V_t > x | \text{arrival in } (t, t+\Delta)) \\
&= \lim_{\Delta\to 0} \frac{\mathbb{P}(\text{arrival in } (t, t+\Delta)|V_t > x)\mathbb{P}(V_t > x)}{\mathbb{P}(\text{arrival in } (t, t+\Delta))} \\
&= \frac{1}{\bar{\lambda}} \int_{y=x}^{\infty} \lambda(y)v(y)\mathrm{d}y, \qquad\qquad x \geq 0,
\end{aligned}$$

where $\bar{\lambda} = \int_{y=0+}^{\infty} \lambda(y)v(y)\mathrm{d}y + \lambda(0)V(0)$. Differentiating results in the intuitive explanation of the theorem:

$$w(x) = \frac{\mathrm{d}}{\mathrm{d}x}(1 - \mathbb{P}(W > x)) = \frac{1}{\bar{\lambda}} v(x)\lambda(x).$$

We now turn to Theorem 3.2. Let us consider either of the two models and define

$$\begin{aligned}
A^w_{n,x} &= \text{workload decrement between arrivals } n \text{ and } n+1, \\
&\quad \text{when workload immediately } \textit{after } n\text{-th arrival epoch is } x, \qquad n \in \mathbb{N}, x > 0.
\end{aligned}$$

Observe that $A^w_{n,x}$ may be interpreted as some kind of interarrival time between the n-th and $(n+1)$-st customer. While the interarrival time is usually expressed in terms of time, $A^w_{n,x}$ represents the workload decrement between two successive arrivals. Remember that a similar argument holds for the service requirement, which in general does not equal the service time, and the workload at jump epochs, which in general differs from the waiting time. This demonstrates that the following well-known relation, usually interpreted in terms of time, holds again in terms of workload:

$$W_{n+1} = (W_n + B_n - A^w_{n,W_n+B_n})^+, \qquad n = 1, 2, \ldots. \tag{6}$$

If we omitted the times between successive arrivals, we would have a system of only upward (arrival of a customer) and downward jumps (workload decrement during an interarrival interval). The distribution of the workload at arrival epochs only depends on the sizes of these jumps, as can be concluded from (6). Hence, the workload at jump epochs only depends indirectly on the time between two successive arrivals. The distribution of the service requirements (upward jumps) is by assumption identical for Models 1 and 2. In order to prove Theorem 3.3, it suffices to show that the sizes of the downward jumps, i.e. the workload decrements during an interarrival interval, are identically distributed for Models 1 and 2.

Thus let us consider the interarrival time and the workload decrement in an interarrival interval of either of the two models. Assume that at time 0 a customer arrives and the workload just after the arrival is $W + B$, with realization $W + B = y$. Denote by A_y^t the conditional interarrival time and by A_y^w the conditional workload decrement during the interarrival time, i.e., the event $\{A_y^w > v\}$ represents the situation that when the next customer arrives the workload is smaller than $y - v$. If we let t_v be the conditional time required for a workload decrease of v, then the events $\{A_y^w > v\}$ and $\{A_y^t > t_v\}$ are identical.
Next, use an alternative definition of a Poisson arrival process with rate $\lambda(x)$ when the workload equals x, to determine the excess probability of the conditional interarrival time,

$$\mathbb{P}(A_y^t > t_v) = e^{-\int_{t=0}^{t_v} \lambda(V_t)\mathrm{d}t}, \qquad y > v. \tag{7}$$

Recall that $r(x)$ is the rate of decline at time t if the workload V_t equals x. Hence, between successive arrivals the workload process satisfies (see, e.g., [1, 5, 12, 13])

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = -r(V_t). \tag{8}$$

Since the amount of work at $t = 0$ equals y, and t_v is defined such that $\int_{y-v}^{y} \frac{1}{r(x)}\mathrm{d}x = t_v$, the next proposition follows easily from (7) and (8).

Proposition 3.1. *Assume that the workload just after an arrival is y ($W + B = y$), then, for $y > v$:*

$$\mathbb{P}(A_y^w > v) = e^{-\int_{u=y-v}^{y} \frac{\lambda(u)}{r(u)}\mathrm{d}u}. \tag{9}$$

Differentiation of (9) yields the conditional density:

$$\frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(A_{W+B}^w \le x | W + B = y) = \frac{\lambda(x)}{r(x)} e^{-\int_{z=x}^{y} \frac{\lambda(z)}{r(z)}\mathrm{d}z}, \qquad 0 < x < y. \tag{10}$$

Proof of Theorem 3.3. Since $A_{n,x}^w$ depends on $W_n + B_n$, (6) leads to:

$$\mathbb{P}(W_{n+1} > z) = \int_{u=z}^{\infty} \mathbb{P}(A_{n,u}^w < u - z \mid W_n + B_n = u)\mathrm{d}\mathbb{P}(W_n + B_n \le u), \qquad n = 1, 2, \ldots.$$

Notice that the distribution of $A_{n,x}^w$ (cf. (9)) depends only on the ratio $\frac{\lambda(\cdot)}{r(\cdot)}$ and hence is the same in both models. Since the distribution of the service requirements is the same by assumption, we can use a stochastic coupling argument to complete the proof. □

From Theorem 3.1 and the discussion of Theorem 3.3 we can conclude that changing between Model 1 and Model 2 is just a rescaling of time. If we consider the workload process of Model i, with the speed of time equal to $\frac{1}{r_i(x)}$ when the workload is x, then Models 1 and 2 are equivalent. The special case with $r_1(x) \equiv r_1$ and $r_2(x) \equiv r_2$ can thus be interpreted as observing the workload at two different (but constant) time scales, which clearly does not affect the workload distribution. It obviously follows from the rescaling arguments and the existence of a workload density at arrival instants in the ordinary $M/G/1$ queue, that the density $w(\cdot)$ in the more general model is also well-defined.

4 Special cases

The main results of Section 3 provide us with a tool to translate known results for a particular model to a whole class of related models. In this section we consider several examples. Throughout the section we use the notation $f(x) \propto g(x)$ if $f(x) = cg(x)$ for all $x > 0$ and some constant c.

Case (i): Arrival control follows service rate.
We start with an $M/G/1$ queue with $\lambda(x) = Cr(x)$. Note that this special case might be applicable to queues where the arrival control must follow the service rate, see for instance [3]:

- high workload: panicking server, reduce $\lambda(\cdot)$
- medium workload: fast server, send more work
- small workload: lazy server, send less work

Note that the third case does not seem desirable. However, in practical situations the main focus will be on a server with a relatively high workload and regimes with a small workload are usually of very limited interest.

It is obvious that $\frac{\lambda(x)}{r(x)} = \frac{C}{1}$. Apply Theorems 3.1 and 3.3 to see that $r(x)v(x) \propto v_{M/G/1}(x)$ and $w(x) = w_{M/G/1}(x)$, where $v_{M/G/1}(\cdot)$ and $w_{M/G/1}(\cdot)$ are the workload densities of the ordinary $M/G/1$ queue with arrival rate C and service speed 1, at arbitrary and arrival epochs, respectively. Hence the workload process in the $M/G/1$ queue with general service speed $r(x)$ and arrival rate $\lambda(x) = Cr(x)$ can be analyzed in detail.

Case (ii): Shot noise.
Another well-known example is the shot noise model, i.e., $\lambda(x) \equiv \lambda$ and $r(x) = rx$, cf. [16], [22], p. 393, or [6], p. 558. First notice that for the shot noise model $\tilde{R}(x) = \infty$ for $x > 0$ and the Volterra successive iteration approach cannot (at least not directly) be applied. However, it is still possible to analyze this model due to

the special form of $\lambda(\cdot)$ and $r(\cdot)$. First consider the level crossing identity (2) for this case,

$$rxv(x) = \lambda \int_{y=0^+}^{x} (1 - B(x-y))v(y)\mathrm{d}y, \qquad x > 0.$$

Let $\phi(s) := \int_0^\infty e^{-sx}v(x)\mathrm{d}x$ be the Laplace Transform (LT) of the workload density, then $-\frac{\mathrm{d}}{\mathrm{d}s}\phi(s) = \int_0^\infty e^{-sx}xv(x)\mathrm{d}x$ and we have,

$$-\frac{\mathrm{d}}{\mathrm{d}s}\phi(s) = \frac{\lambda}{r}\frac{1-\beta(s)}{s}\phi(s).$$

Solving this differential equation yields (cf. [16], [22], p. 393, or [6], p. 558)

$$\phi(s) = \exp\left\{-\frac{\lambda}{r}\int_0^s \frac{1-\beta(u)}{u}\mathrm{d}u\right\}. \tag{11}$$

Because of the PASTA property, the LT of the workload density at arrival epochs also equals $\phi(\cdot)$.

Since the shot noise model is thus solved, we immediately obtain the LT of the workload density at arbitrary and at arrival epochs of models with $\lambda(x) = f(x)x^\alpha$ and $r(x) = f(x)x^{\alpha+1}$, where $\int_0^x f(y)y^\alpha\mathrm{d}y$ and $\int_0^x f(y)^{-1}y^{-(\alpha+1)}\mathrm{d}y$ are not both infinite (use Theorems 3.1-3.3).

Furthermore, let us consider some special properties of the shot noise model. Take $\lambda \equiv r \equiv 1$, and use (9) to observe that, given $W_n + B_n$, W_{n+1} is uniformly distributed on the stochastic interval $(0, W_n + B_n)$. This means that the steady-state distribution W must satisfy

$$W =^d U(0, W + B), \tag{12}$$

with $U(x, y)$ denoting the uniform distribution on the interval (x, y). But if W is $U(0, W + B)$ distributed, then also B must be $U(0, W + B)$ distributed. Hence, in a model with exponentially distributed service requirements, W must have the same exponential distribution as the service requirements B.

Remark 4.1. Using LST's it can also be readily shown that equation (12), with B $\exp(\mu)$ distributed, implies that W is $\exp(\mu)$ distributed. Indeed, (12) implies that the LST $\psi(s)$ of W satisfies

$$\psi(s) = \frac{1}{s}\int_{\sigma=0}^{s} \psi(\sigma)\frac{\mu}{\mu+\sigma}\mathrm{d}\sigma,$$

with boundary condition $\psi(0) = 1$, which after differentiation yields $\psi(s) = \frac{\mu}{\mu+s}$.

Another special case arises when the service requirements are exponential random variables with rate μ, while λ, r are not necessarily identically 1. Substituting $\beta(s) = \frac{\mu}{\mu+s}$ in (11) yields

$$\phi(s) = \left(\frac{\mu}{\mu+s}\right)^{\lambda/r}.$$

This is the LST of the Gamma distribution, i.e., $v(\cdot) \propto \text{Gamma}(\frac{\lambda}{r}, \mu)$ and due to the PASTA property also $w(\cdot) \propto \text{Gamma}(\frac{\lambda}{r}, \mu)$. Furthermore, if we consider the model with $\lambda(x) = \frac{\lambda}{x}$ and $r(x) \equiv r$, it follows that $w(\cdot) \propto \text{Gamma}(\frac{\lambda}{r}, \mu)$ and $v(x)$ is proportional to $xw(x)$, hence, $v(\cdot) \propto \text{Gamma}(\frac{\lambda}{r} + 1, \mu)$. Note that taking $\lambda \equiv r \equiv 1$ indeed gives that $w(\cdot)$ is exponentially distributed with parameter μ.

Case (iii): Exponential service times.
Consider the $M/M/1$ queue with general $\lambda(\cdot)$ and $r(\cdot)$ functions. Substituting $B(x) = 1 - e^{-\mu x}$ in (2) gives

$$r(x)v(x) = \lambda(0)V(0)e^{-\mu x} + \int_{y=0^+}^{x} \lambda(y)v(y)e^{-\mu(x-y)}\mathrm{d}y, \qquad x > 0. \tag{13}$$

Multiply by $e^{\mu x}$, define $f(x) := e^{\mu x}r(x)v(x)$ and differentiate to obtain

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x) = f(x)\frac{\lambda(x)}{r(x)}.$$

The solution of this differential equation is unique up to a constant and may be written in the form

$$f(x) = C\exp\left\{\tilde{R}(x,1)\right\}, \qquad x > 0. \tag{14}$$

Using a straightforward extension of Asmussen [1], p. 295, it follows that we have positive recurrence (and hence, C can be determined by normalization) if and only if

$$\alpha := \int_0^\infty \frac{1}{r(x)}\exp\left\{\tilde{R}(x,1) - \mu x\right\}\mathrm{d}x < \infty.$$

If $V(0) = 0$, then $C = \alpha^{-1}$. However, if $V(0) > 0$, observe from (13) that $\lim_{x\downarrow 0} r(x)v(x) = \lambda(0)V(0)$ and use the straightforward extension $\tilde{R}(x,y) = -\tilde{R}(y,x)$ for $0 < x < y < \infty$ to see that $C = \lambda(0)V(0)\exp\{\tilde{R}(1,0)\}$. Hence, if $V(0) > 0$, then

$$v(x) = \frac{\lambda(0)V(0)}{r(x)}\exp\left\{\int_0^x \left(\frac{\lambda(y)}{r(y)} - \mu\right)\mathrm{d}y\right\}.$$

From this solution it becomes immediately clear that for two models with $\lambda_i(\cdot)$ and $r_i(\cdot)$ $(i = 1, 2)$ satisfying (3), Theorem 3.1 holds in the $M/M/1$ case.

Case (iv): Some other models.
Next, consider the $M/G/1$ queue with $\lambda(x) \equiv \lambda$ and $r(x) = x^2$. Noting that $\int_0^\infty x^2 v(x)e^{-sx}\mathrm{d}x = \phi''(s)$ and using the level crossing identity (2) yields

$$\phi''(s) = \lambda\frac{1-\beta(s)}{s}\phi(s). \tag{15}$$

Denoting $g(s) := \lambda\frac{1-\beta(s)}{s}$ and using the transformations $\phi(s) = e^{f(s)}$ and $h(s) = f'(s)$ gives

$$h'(s) + \left[h(s)\right]^2 = g(s).$$

This non-linear first-order differential equation is in general very difficult to solve. Note that $\tilde{R}(x) = \infty$, for all $x > 0$, and the workload process can thus not reach state 0. However, for some specific choices of the service requirement we obtain an expression for the LT. For instance, the LT of $v(x)$ in the $M/M/1$ case is given by (use (14))

$$\phi(s) = G \int_0^\infty \frac{e^{-(s+\mu)x}}{x^2} e^{-\frac{\lambda}{x}} \mathrm{d}x,$$

with G some normalizing constant. Integrating $\phi''(s)$ by parts shows that (15) is satisfied for the case of exponential service requirements.

If we let the service requirement be Erlang$(2, \mu)$ distributed, then substituting $\lambda(x) = \lambda, r(x) = x^2$ into (2), defining $z(x) := e^{\mu x} v(x)$, and differentiating twice gives

$$x^2 z''(x) + (4x - \lambda) z'(x) + (2 - \lambda\mu) z(x) = 0. \tag{16}$$

This linear second-order differential equation can be transformed into a differential equation of Bessel type. Furthermore, define $\psi(s) := \int_0^\infty e^{-sx} z(x) \mathrm{d}x$, take the LT in (16), and observe that $\int_0^\infty e^{-sx} z''(x) \mathrm{d}x = s \int_0^\infty e^{-sx} z'(x) \mathrm{d}x - z'(0) = s^2 \psi(s) - z'(0) - z(0)$ and $\phi(s) = \psi(s + \mu)$ to show that (15) is satisfied. We conclude with the observation that if we can calculate $v(\cdot)$ and/or $w(\cdot)$ this immediately gives results for $M/G/1$ models with $\lambda(x) = f(x)x^\alpha$ and $r(x) = f(x)x^{\alpha+2}$, where $\int_0^x f(y) y^\alpha \mathrm{d}y$ and $\int_0^x f(y)^{-1} y^{-(\alpha+2)} \mathrm{d}y$ are not both infinite (again use Theorems 3.1-3.3).

Remark 4.2. If $\lambda(x) \equiv \lambda$ and $r(x) = e^{ax}$, then $R(x) = \int_0^x e^{-ay} \mathrm{d}y < \infty$ for $0 < x < \infty, |a| < \infty$. In this case, we can follow [13], using the bound $K(x, y) \leq \frac{\lambda}{r(y)}$ and show inductively that

$$K_{n+1}(x, y) \leq \frac{\lambda^{n+1} (R(x) - R(y))^n}{r(y) n!} = \frac{\lambda^{n+1} e^{-ay} (e^{-ay} - e^{-ax})^n}{a^n n!}, \qquad 0 < y < x < \infty.$$

Thus the sum $\sum_{n=1}^\infty K_n(x, y)$ is well-defined, and hence, $\sum_{n=0}^\infty K_{n*}(x, y)$ is well-defined as well. So the steady-state workload density is given by (5):

$$v(x) = \lambda V(0) e^{-ax} \sum_{n=0}^\infty K_{n*}(x, 0), \qquad x > 0,$$

with $V(0)$ determined via normalization:

$$V(0) = \left[1 + \int_{0+}^\infty \lambda e^{-ax} \sum_{n=0}^\infty K_{n*}(x, 0) \mathrm{d}x \right]^{-1}.$$

Of course, this can be extended to models with $\lambda(x) = f(x)e^{bx}$ and $r(x) = f(x)e^{ax}$, where $\int_0^x f(y) e^{by} \mathrm{d}y$ and $\int_0^x f(y)^{-1} e^{-ay} \mathrm{d}y$ are not both infinite.

5 Palm theoretic approach

So far, we considered $M/G/1$-type queues with general arrival and service rate, $\lambda(\cdot)$, $r(\cdot)$, depending on the amount of work in the system. Recall that B_n denotes the service requirement of the n-th customer and A_n denotes the interarrival time between the n-th and $(n+1)$-th customer, $n = 0, 1, \ldots$, where B and A are their steady-state random variables. Furthermore, let V_t again denote the workload at time t and let W_n again be the workload immediately before the n-th jump epoch, with steady-state random variables V and W, respectively.
In this section, we first continue the study of this Markovian case. Using Palm theoretic principles, we establish a general relation between V and W, or rather $f(V)$ and $f(W)$. Some specific well-chosen $f(\cdot)$-functions yield convenient relations for, e.g., the excess probability, the expectation, or the LST of the considered random variables. In addition, Theorem 3.2 follows easily as a consequence. We proceed by allowing general renewal arrival processes and again establish a general relation between V and W. Some examples show that the relations may be viewed as extensions of some well-known relations for ordinary $G/G/1$ queues. Furthermore, in case of Poisson arrivals, the level crossing equations are derived in an alternative way. We conclude with an extension of the dynamics driving the workload process in the ordinary $G/G/1$ queue to similar dynamics in $G/G/1$-type queues with general release rate.
In Sections 3 and 4 the arrivals followed a Poisson process. We applied the level crossing equations (2) to determine the limiting distribution and to show some equivalence properties. Although level crossing arguments still hold for non-Poissonian arrival processes (see Doshi [8]), their applicability, however, is limited. In order to handle the general renewal nature of the input process, we adopt a totally different approach based on Palm theoretic principles, see for instance [2], Section 1.3. Specifically, we express $\mathbb{E}f(V)$ as a stochastic mean value over one arbitrary interarrival interval. Let $W+B$ and W_A denote the workload at the beginning and end, respectively, of the (arbitrary) interarrival interval A. If we assume that the function $f(\cdot)$ is such that the considered expectations exist and are finite, then

$$\mathbb{E}f(V) = \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{t=0}^{A} f(V_t)\mathrm{d}t\right]. \tag{17}$$

First, let us consider the Markov case.

Theorem 5.1. *Let $f(\cdot)$ be such that $\mathbb{E}f(V)$ exists and is finite, then*

$$\mathbb{E}f(V) = \mathbb{E}\left[\frac{f(W)}{\lambda(W)}\right]\frac{1}{\mathbb{E}\frac{1}{\lambda(W)}}. \tag{18}$$

Proof. Starting with the stochastic mean-value result (17) and introducing the in-

dicator function $I(\cdot)$, we have

$$\begin{aligned}\mathbb{E}[f(V)] &= \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{t=0}^{A} f(V_t)\mathrm{d}t\right] \\ &= \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{t=0}^{\infty} f(V_t)I(A>t)\mathrm{d}t\right].\end{aligned}$$

Note that $\mathbb{E}(I(A>t)|W+B) = \mathbb{P}(A>t|W+B) = \mathbb{P}(W_A < V_t|W+B)$, and use (8) to see that

$$\begin{aligned}\mathbb{E}[f(V)] &= \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{x=W+B}^{0} f(x)\mathbb{P}(W_A<x|W_0=W+B)\frac{\mathrm{d}x}{-r(x)}\right] \\ &= \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{x=0}^{W+B}\frac{f(x)}{\lambda(x)}\mathrm{d}\mathbb{P}(W_A\le x|W_0=W+B)\right] \\ &= \frac{1}{\mathbb{E}A}\mathbb{E}\left[\frac{f(W_A)}{\lambda(W_A)}\right]. \qquad (19)\end{aligned}$$

The second equality sign follows by combining (9) and (10). Notice that W_A and W have the same distribution. Furthermore, taking $f(x) \equiv 1$ yields

$$\mathbb{E}A = \mathbb{E}\frac{1}{\lambda(W)}$$

which completes the proof. □

Let us briefly consider some special cases of $f(\cdot)$. Taking $f(x) = x$, respectively $f(x) = e^{-sx}$, gives a relation between $\mathbb{E}V$ and $\mathbb{E}W$, respectively a relation between the LST's of V and W. Furthermore, taking $f(x) = I(x>v)$ expresses the steady-state excess probability of the workload in terms of $\lambda(\cdot)$ and the steady-state workload density at arrival epochs. Taking $f(x) = \lambda(x)g(x)$ yields

$$\mathbb{E}[\lambda(V)g(V)] = \frac{\mathbb{E}g(W)}{\mathbb{E}\frac{1}{\lambda(W)}} = \mathbb{E}\lambda(V)\mathbb{E}g(W), \qquad (20)$$

or equivalently

$$\mathbb{E}g(W) = \frac{\mathbb{E}[\lambda(V)g(V)]}{\mathbb{E}\lambda(V)},$$

where the second equality sign in (20) follows from taking $f(x) = \lambda(x)$ in (18). Now taking $g(x) = e^{-sx}$ yields

$$\mathbb{E}\left[e^{-sW}\right] = \frac{\mathbb{E}\left[\lambda(V)e^{-sV}\right]}{\mathbb{E}\lambda(V)}. \qquad (21)$$

Because of the one-to-one correspondence between an LST and its inverse, (21) implies that the steady-state workload density at arrival epochs $w(x)$ is proportional

to the product of $\lambda(x)$ and the steady-state workload density $v(x)$. Note that we have just proven Theorem 3.2.

Now let us consider a generalization of the above-described $M/G/1$-type model, by allowing *generally* distributed interarrival times, which may depend on the workload W found upon arrival according to some distribution $\mathbb{P}(A < x|W = w)$. We again derive a relation between V and W by starting from the stochastic mean-value result (17). Let B^e denote the residual service requirement, then its density is $\frac{1-B(\cdot)}{\mathbb{E}B}$.

Theorem 5.2. *Let $f(\cdot)$ be such that $\mathbb{E}f(V)$ exists and is finite, then*

$$\mathbb{E}[f(V)|V > 0] = \mathbb{E}[r(V)|V > 0]\mathbb{E}\left[\frac{f(W + B^e)}{r(W + B^e)}\right]. \tag{22}$$

Proof. First define $g(w, z) := \int_0^z f(w + u)\mathrm{d}u$ and consider

$$\begin{aligned}
\mathbb{E}\left[\int_0^B f(w + x)\mathrm{d}x\right] &= \mathbb{E}[g(w, B)] \\
&= \int_0^\infty g'(w, x)\mathbb{P}(B > x)\mathrm{d}x + g(w, 0) \\
&= \int_0^\infty f(w + x)\mathbb{P}(B > x)\mathrm{d}x \\
&= \mathbb{E}B\mathbb{E}[f(w + B^e)].
\end{aligned} \tag{23}$$

Starting with the stochastic mean-value result (17), making the substitution $u = V_t$ and using (8) yields

$$\mathbb{E}[f(V)] = \frac{1}{\mathbb{E}A}\left(\mathbb{E}\left[\int_{u=W+B}^{W} f(u)\frac{\mathrm{d}u}{-r(u)}\right] + f(0)\mathbb{E}(A - \tau)^+\right), \tag{24}$$

where $x^+ = \max(0, x)$ and $\tau := \inf\{t > 0 : V_t = 0\}$. As $V = 0$ might be a special point we focus on $V > 0$, resulting in:

$$\mathbb{E}[f(V)|V > 0]\mathbb{P}(V > 0) = \frac{1}{\mathbb{E}A}\mathbb{E}\left[\int_{u=W}^{W+B} \frac{f(u)}{r(u)}\mathrm{d}u\right]. \tag{25}$$

Since W_{n+1} depends on $W_n + B_n$ the boundaries in $\int_W^{W+B}$ really are dependent, as they represent the workload at two successive arrival epochs. But we can rewrite $\mathbb{E}\int_W^{W+B} = \mathbb{E}[\int_0^{W+B} - \int_0^W]$ and observe that both W_n and W_{n+1} have the same steady-state distribution as W. Thus we can rewrite (25) into

$$\begin{aligned}
\mathbb{E}[f(V)|V > 0]\mathbb{P}(V > 0) &= \frac{1}{\mathbb{E}A}\mathbb{E}_B\left[\int_{w=0}^{\infty} \mathrm{d}\mathbb{P}(W \le w)\int_{u=w}^{w+B} \frac{f(u)}{r(u)}\mathrm{d}u\right] \\
&= \frac{1}{\mathbb{E}A}\int_{w=0}^{\infty} \mathrm{d}\mathbb{P}(W \le w)\mathbb{E}B\mathbb{E}\left[\frac{f(w + B^e)}{r(w + B^e)}\right] \\
&= \frac{\mathbb{E}B}{\mathbb{E}A}\mathbb{E}\left[\frac{f(W + B^e)}{r(W + B^e)}\right],
\end{aligned} \tag{26}$$

where we have used (23) in the second equality. The Theorem follows by taking $f(x) = r(x)$, leading to $\mathbb{E}[r(V)|V > 0]\mathbb{P}(V > 0) = \frac{\mathbb{E}B}{\mathbb{E}A}$. □

Again, taking respectively $f(x) = x$, $f(x) = e^{-sx}$, and $f(x) = I(x > v)$ gives a relation between the workload at arbitrary epochs V and the workload at arrival epochs W, for respectively the expectation, the LST, and the excess probabilities. Taking $f(x) = r(x)g(x)$ yields

$$\mathbb{E}[r(V)g(V)|V > 0] = \mathbb{E}[r(V)|V > 0]\mathbb{E}[g(W + B^e)],$$

and in particular

$$\mathbb{E}\left[r(V)e^{-sV}|V > 0\right] = \mathbb{E}\left[r(V)|V > 0\right]\mathbb{E}\left[e^{-s(W+B^e)}\right]. \tag{27}$$

The latter relation implies that the steady-state density of $W + B^e$ is proportional to the product of $r(\cdot)$ and the conditional steady-state density of V.

Using (27) we obtain an alternative proof of the level crossing identity (2). To show this, we let the arrival process be Poisson with intensity $\lambda(x)$ when the workload equals x. Note that the interarrival time and workload at arrival epochs are dependent; however, we can still apply Theorem 5.2 and (24). Furthermore, observe that for the choice of $f(x) = r(x)e^{-sx}$, we have that $f(0) = 0$, since we assumed that $r(0) = 0$. Then, by conditioning, it follows directly from (26) that $\mathbb{E}r(V) = \frac{\mathbb{E}B}{\mathbb{E}A}$, and similarly, we can rewrite (27) into

$$\mathbb{E}\left[r(V)e^{-sV}\right] = \frac{\mathbb{E}B}{\mathbb{E}A}\mathbb{E}\left[e^{-s(W+B^e)}\right].$$

Using the one-to-one correspondence between an LST and its inverse again, yields

$$r(x)v(x) = \frac{\mathbb{E}B}{\mathbb{E}A}\int_{y=0^-}^{x} \frac{1 - B(x-y)}{\mathbb{E}B} w(y)\mathrm{d}y,$$

where the 0^- in the integral denotes the inclusion of the (possibly exceptional) point 0. Furthermore, we had proven that $w(y) = \lambda(y)v(y)/\bar{\lambda}$, with $\bar{\lambda} = \int_{0^-}^{\infty} \lambda(y)v(y)\mathrm{d}y$ (see for instance Theorem 3.2). Take $f(x) = \lambda(x)$ in (19) to see that $1/\bar{\lambda} = \mathbb{E}A$ and the constants cancel,

$$r(x)v(x) = \int_{y=0^-}^{x} (1 - B(x-y))\lambda(y)v(y)\mathrm{d}y, \qquad x > 0.$$

Hence, we have shown the level crossing identity (2) in an alternative way.

Remark 5.1. Formula (25) is also valid when the n-th service time B_n is dependent on the workload W_n at its arrival.

Remark 5.2. It should be noted that taking $r(x) \equiv 1$ in (27) results in a well-known result for the $GI/G/1$ queue, cf. [6], p. 296 or [1], p. 189:

$$V|V > 0 =^d W + B^e.$$

Another interesting relation between V and W in the ordinary $GI/G/1$ queue is presented in Asmussen [1], p. 189:

$$V =^d (W + B - A^e)^+, \tag{28}$$

where A^e denotes a residual interarrival time. We now generalize (28) to a $G/G/1$ queue with general service rate $r(x)$ when the workload equals x. By the stochastic mean-value result (17) and some similar manipulations as we did proving Theorems 5.1 and 5.2, one can find the following relation:

$$\begin{aligned}
\mathbb{E}[f(V)] &= \frac{1}{\mathbb{E}A}\mathbb{E}_{A,V_t}\left[\int_{t=0}^{\infty} f(V_t)I(A > t)\mathrm{d}t\right] \\
&= \mathbb{E}_{V_t}\left[\int_{t=0}^{\infty} f(V_t)\frac{\mathbb{P}(A > t)}{\mathbb{E}A}\mathrm{d}t\right] \\
&= \mathbb{E}f(V_{A^e}).
\end{aligned}$$

Take $f(x) = I(x > v)$, where $I(\cdot)$ is the indicator function, then it follows that $\mathbb{P}(V > v) = \mathbb{P}(V_{A^e} > v)$. The latter probability equals the probability that A^e is less than the time required for a process, which decreases according to the function $r(\cdot)$, to go from $W + B$ (workload at $t = 0$) to v. Recall that $R(x)$ (see (1)) represents the time required for a workload x to drain in the absence of any arrivals. Moreover, the time required to go from $W + B$ to v according to the described process equals $R(W + B) - R(v) = R(W + B, v)$. Hence, we obtain

$$\mathbb{P}(V > v) = \mathbb{P}(R(W + B) - R(v) > A^e), \qquad v \geq 0.$$

When $r(x) \equiv 1$, this indeed yields (28).

In the remainder of this section we assume that the workload process $\{V_t, t \geq 0\}$ has an atom at zero, or equivalently, that $R(x) < \infty$. Our goal is to study the process $\{R(V_t), t \geq 0\}$. Note that $R(x)$ (like $R(w + x, w)$) is strictly increasing in x so we can speak unambiguously of $R^{-1}(t)$. We are interested in the service (release) process, and assume for the moment that the arrival process is shut off. Besides the time required for a workload u to go down to $x, 0 \leq x \leq u$, in the absence of any arrivals (which equals $R(u) - R(x)$), we are interested in, for instance, the workload level at time $t > 0$ when $V_0 = u$. It is well-known that the latter expression equals $R^{-1}(R(u) - t)$ [13]. So, in principle it is possible to switch from the workload interpretation to the time interpretation and vice versa. However, there does not seem to be much hope for convenient expressions.

Using the definition of $R(\cdot)$ and the transformation property (8) between workload and time, it is easy to see that $R(V)$ transforms the workload V into the time required to finish the work in the system when no arrivals occur. This means that as long as there are no jumps, the process $R(V_t)$ decreases linearly with slope -1 until $R(V_t) = 0$ and then remains 0 until the next arrival. To get some feeling for the $R(V_t)$ process, it is easiest to think of its graphical representation: we have rescaled the workload axis such that in each time interval of length Δx where no arrival occurs, the decrement of the function $R(\cdot)$ is Δx. This means that every very small workload interval $(x, x + r(x)\Delta x)$ of the V_t process is compressed (or expanded) to an interval $(x, x + \Delta x)$. Since $r(\cdot)$ is left-continuous and has right-hand limits, it is bounded on closed intervals and hence the time required to move down from level $x + r(x)\Delta x$ to level x is indeed $r(x)\frac{\Delta x}{r(x)} = \Delta x$ for Δx small enough.

The jump sizes of the $R(V_t)$ process consist of the differences of the time required to empty the system just before, and just after, the arrival epoch. Hence, in steady state this service requirement equals $R(W + B) - R(W)$, or alternatively $R(W + B, W)$. Thus the $R(V_t)$ process behaves like an ordinary $GI/G/1$ queue with workload-dependent service requirements, and follows the same sample path as V_t if we transform the jump size distribution according to the above integral. From the arguments above we can observe that for the $G/G/1$ queue with service rate $r(x)$ when the workload equals x, we have the following relations between V and W:

Theorem 5.3. *If $R(x) < \infty$ for all $0 < x < \infty$, then*

$$R(V) \stackrel{d}{=} \left(R(W + B) - A^e\right)^+, \tag{29}$$

$$R(W) \stackrel{d}{=} \left(R(W + B) - A\right)^+. \tag{30}$$

Furthermore, if we were able to solve the stationary distribution (density) of the $R(V_t)$ process, denoted by $V^R(\cdot)$ ($v^R(\cdot)$), we would have the stationary distribution (density) of V_t, since $v(x) = \frac{v^R(x)}{r(x)}$.

Remark 5.3. Taking $r(x) \equiv 1$ in (29) and (30) results respectively in (28) and the well-known relation for the $GI/G/1$ queue (see for instance [6], p. 167).

Remark 5.4. In a similar way like (22) we can derive that the expected jump size of the $R(V_t)$ process equals $\mathbb{E}B\mathbb{E}\frac{1}{r(W+B^e)}$.

Hence, the transformation from a $G/G/1$ queue with general service rate function $r(\cdot)$ to an ordinary $G/G/1$ queue (with a server working at unit speed) can be interpreted as a rescaling of the service requirement. In the transformed $R(V_t)$ model, the amount of work a customer brings upon arrival takes into account the time required to finish this additional workload. If, for instance, $r(x) \equiv r$, we rescale the service requirement by the factor r^{-1} to take into account that the server would have been working at speed r in the V_t process. Note that, in the absence of any arrivals, the time required to finish a workload B at speed r indeed equals the time required to finish a workload $r^{-1}B$ at unit speed.

6 Conclusions and research topics

We studied single-server queues with state-dependent interarrival times and service speed. The two main contributions of this paper may be summarized as follows. Firstly, in the case of Poisson arrivals, we derived proportionality relations between the workload distribution of two queues that have the same ratio of arrival rate and service speed. Such relationships allow us to obtain results for a whole class of models from the analysis of one particular model. Secondly, we analyzed $G/G/1$-type queues with workload-dependent service speed and interarrival times. Using a Palm theoretic approach, several well-known relations for the workload at various epochs in the ordinary $G/G/1$ queue were generalized. Moreover, an extension of the PASTA result to $M/G/1$ queues with state-dependent arrival rate followed as a by-product.
Finally, we mention three topics for further research.

(i) Extension of the results for $M/G/1$-type queues with general (workload-dependent) arrival and service rate to finite buffers. This is the subject of a forthcoming study.

(ii) Derivation of the steady-state workload distribution is subject to some complications if the workload process has no atom at state 0. In the case of Poisson arrivals, we observed that level crossing arguments lead to an integral equation that is very difficult to solve in general. The steady-state behavior of a workload process without an atom at state 0 may be a subject of further exploration.

(iii) In production systems, for example, workload management may be realized by controlling the arrival rate of new jobs, or by regulating the speed of the server. Given a target steady-state behavior of the workload, an important issue is the design of the system such that this target is met. This so-called *reverse engineering* (cf. [9]) is left for a subsequent investigation.

References

[1] Asmussen, S. (1987). *Applied Probability and Queues.* Wiley, New York.

[2] Baccelli, F., P. Brémaud (2003). *Elements of Queueing Theory*, Second Edition. Springer-Verlag, Berlin.

[3] Bertrand, J.W.M., H.P.G. van Ooijen (2000). *Workload based order release and productivity: a missing link.* Working paper, Department of Technology Management, Eindhoven University of Technology, to appear in Production Planning & Control.

[4] Cohen, J.W. (1976). On Regenerative Processes in Queueing Theory. *Lecture Notes in Economics and Mathematical Systems* **121**. Springer-Verlag, Berlin.

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

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

[7] Cohen, J.W., M. Rubinovitch (1977). On level crossings and cycles in dam processes. *Math. Oper. Res.* **2**, 297–310.

[8] Doshi, B.T. (1992). Level-crossing analysis of queues. In: Bhat, U.N., Basawa I.V. (editors). *Queueing and Related Models.* Oxford Statistical Science Series, Oxford Univ. Press, New York, 3–33.

[9] Eliazar, I., Klafter, J. (2002). *Lévy-driven Langevin systems: targeted stochasticity.* Technical report, Recanaty Faculty of Management and Sackler Faculty of Exact Sciences, Tel Aviv University, Israel.

[10] Elwalid, A., D. Mitra (1991). Analysis and design of rate-based congestion control of high-speed networks, I: stochastic fluid models, access regulation. *Queueing Systems* **9**, 29–64.

[11] Elwalid, A., D. Mitra (1994). Statistical multiplexing with loss priorities in rate-based congestion control of high-speed networks. *IEEE Trans. Commun.* **42**, 2989–3002.

[12] Gaver, D.P., Jr., R.G. Miller, Jr. (1962). Limiting distributions for some storage problems. In: Arrow, K.J., Karlin, S., Scarf, H. (editors). *Studies in Applied Probability and Management Science.* Stanford University Press, Stanford, California, 110–126.

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

[14] Hasofer, A.M. (1963). On the integrability, continuity and differentiability of a family of functions introduced by L. Takács. *Ann. Math. Statist.* **34**, 1045–1049.

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

[16] Keilson, J., N.D. Mermin (1959). The second-order distribution of integrated shot noise. *IRE Transactions on Information Theory* **5**, 75–77.

[17] Linz, P. (1985). *Analytical and Numerical Methods for Volterra Equations.* SIAM Studies in Applied Mathematics **7**. SIAM, Philadelphia.

[18] Mandjes, M., D. Mitra, W.R.W. Scheinhardt (2002). A simple model of network access: feedback adaptation of rates and admission control. In: *Proceedings of Infocom 2002*, 3-12.

[19] Mikhlin, S.G. (1957). *Integral Equations.* International series of monographs on pure and applied mathematics **4**. Pergamon Press, London.

[20] Prabhu, N.U. (1965). *Queues and Inventories.* Wiley, London.

[21] Ramanan, K., A. Weiss (1997). Sharing bandwidth in ATM. In: *Proceedings of the Allerton Conference*, 732-740.

[22] Ross, S.M. (1996). *Stochastic Processes*, Second Edition. Wiley, New York.

[23] Takács, L. (1955). Investigation of waiting time problems by reduction to Markov processes. *Acta Math. Acad. Sci. Hungar.* **6**, 101–129.

[24] Zabreiko, P.P, A.I. Koshelev, M.A. Krasnosel'skii; transl. and ed. by T.O. Shaposhnikova, R.S. Anderssen and S.G. Mikhlin (1975). *Integral Equations: a Reference Text.* Monographs and textbooks on pure and applied mathematics. Noordhoff, Leiden.