Finite-buffer queues with workload-dependent service and arrival rates

Finite-Buffer Queues with Workload-Dependent Service and Arrival Rates

René Bekker*

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

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

Abstract

We consider M/G/1 queues with workload-dependent *arrival rate*, *service speed*, and *restricted accessibility*. The admittance of customers typically depends on the amount of work found upon arrival in addition to its own service requirement. Typical examples are the finite dam, systems with customer impatience and queues regulated by the complete rejection discipline. 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, we compare the steady-state distribution of the workload in two finite-buffer models, in which the ratio of arrival and service speed is equal. Second, we find an explicit expression for the cycle maximum in an M/G/1 queue with workload-dependent arrival and service rate. And third, we derive a formal solution for the steady-state workload density in case of restricted accessibility. The proportionality relation between some finite and infinite-buffer queues is extended. Level crossings and Volterra integral equations play a key role in our approach.

Keywords: restricted accessibility, finite buffer, impatience, state-dependent rates, workload, cycle maximum, level crossings, Volterra integral equation.

1 Introduction

Queueing systems where the service speed is workload-dependent are well-known, specifically in the studies of dams and storage processes, see e.g. [3, 8, 12, 17, 19, 22]. Also, in production systems the speed of the server often depends on the amount of work. This is particularly true if the server is not being represented by a machine, but rather by a human being; see [6, 27] for an example where the speed of the server is relatively low when there is much work (stress) or when there is little work (laziness).

*Supported by a research grant from Philips Electronics

In addition to general service speeds, the rate at which jobs arrive at the system may also depend on the amount of work present. In the human-server production system, we may try to control the arrival rate 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 [15, 16, 26, 29]. 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 where the arrival rate and service speed depend on the amount of work present (see also [4]). Specifically, we consider M/G/1-systems with restricted accessibility. In this paper, the admittance of customers typically depends on the amount of work upon arrival in addition to its own service requirement. In such systems, we may distinguish three main rejection disciplines: (i) the finite dam, governed by the partial rejection rule (scenario f), (ii) systems with impatience of customers depending on the amount of work found upon arrival (scenario i), and (iii) queues regulated by the complete rejection discipline (scenario c).

The three main goals of our study are the following. First, we establish relationships between two queueing models with arrival rates $\lambda_i(x)$ and service speeds $r_i(x), i = 1, 2$, for which $\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}, \forall x > 0$. Hereby, we extend results from [4] to queues with restricted accessibility. These relationships between two queueing systems will allow us to obtain results for a whole class of models from the analysis of one particular model. This is particularly useful in considering performance measures such as steady-state workload densities and loss probabilities.

Turning to our second goal, we obtain an explicit (formal) expression for the cycle maximum in an M/G/1 queue with workload-dependent service and arrival rate. This may be an important tool in determining the maximum buffer content. Exact results for such systems are hardly known; we refer to Asmussen [2] for an overview on cycle maxima.

Third, we derive a formal solution for the steady-state workload density in finite-buffer M/G/1 systems. Often, the density may be expressed as the solution of a Volterra integral equation. In some special cases, this reduces to an analytically tractable expression. Otherwise, numerical methods are widely available, see e.g. [21, 24]. Another tool to solve the workload density is the proportionality of the workload distribution between systems with finite and infinite-buffer capacities. This relation is well-known for some traditional queueing models (where work is depleted at unit rate), see [11, 12, 20]. Using a similar sample path approach as in [20, 31], the proportionality relation is extended to similar systems with workload-dependent arrival and service rate.

In classical queueing systems, the workload just before a customer arrival represents a waiting time, and the workload right after an arrival instant may be identified with a sojourn time. For such models, the rejection rules have a direct interpretation. Our first discipline, the finite dam (scenario f), represents a system where every customer has a bounded sojourn time; a rejected customer may enter the system but his sojourn time is restricted. This model is also frequently used in the context of inventory and storage processes. Due to the above mentioned proportionality, the finite dam is closely related to the infinite-buffer version of the model [11, 12], and has thus been analyzed in detail, see e.g. [12].

The second rejection discipline, scenario i, reflects the fact that impatient customers are only willing to wait a limited amount of time. Results are also well-known for these traditional queueing models, see e.g. [7, 10, 13, 23]. In queues with general service speeds,

the workload found upon arrival does in general not equal the waiting time. However, these two quantities are closely related and the admittance may depend on the workload upon arrival.
Finally, the third discipline, scenario c, also exhibits the case in which customers have a restricted sojourn time. In contrast with scenario f, rejected customers are completely rejected and do not join the queue. This scenario is probably one of the more difficult to analyze. Results are only known for the M/M/1 and M/D/1 case (see e.g. [10, 18]), and the Ph/Ph/1 case [25]. Asymptotics for more general models are obtained in [32].
This paper is organized as follows. In Section 2 we introduce the general model. The relations between two finite-buffer queues are given in Section 3. In Section 4, the finite dam (scenario f) is studied and the proportionality relation between finite and infinite-buffer systems is presented. First-exit probabilities and cycle maxima are considered in Section 5. Section 6 examines scenarios i and c, and we conclude with some examples in Section 7.

2 Model description and preliminaries

In this section, we introduce the general model and obtain some preliminary results. Some examples of typical finite-buffer models are given at the end of the section.
We first describe the general system. Customers arrive at a queueing system according to a Poisson process with arrival rate $\lambda(x)$ when the workload equals $x, x \geq 0$; in other words, the probability of an arrival in some interval $(t, t+h)$ equals $\lambda(x)h + o(h)$ for $h \downarrow 0$ when the work present at time t equals x. We assume that $\lambda(\cdot)$ is nonnegative, left-continuous and has a right limit on $[0, \infty)$. The service requirement of customer n is denoted by $B_n, n = 1, 2, \ldots$, whereas $B_1, B_2, \ldots$ are assumed to be independent, identically distributed with distribution $B(\cdot)$.
Depending on the service requirement and the amount of work found upon arrival, customers may or may not be (fully) accepted. In particular, if the workload just before arrival equals w, and the service requirement is b, then the amount of work right after the arrival instant is $g(w, b, K)$. We assume that $w \leq g(w, b, K) \leq w + b$, where K represents a potential maximum buffer size (see the end of this section for some examples).
We allow the server to operate according to a general service rate (speed) function, a function of the amount of work present. We denote the service rate function by $r : [0, \infty) \to [0, \infty)$, assume that $r(0) = 0$ and that $r(\cdot)$ is strictly positive, left-continuous, and has a right limit on $(0, \infty)$.
In the general model, we define $V^g(t)$ as the workload at time t and let W_n^g be the workload immediately before the n-th arrival epoch. We denote the steady-state random variables of $V^g(t)$ and W_n^g by V^g and W^g, and let $V^g(\cdot)$ and $W^g(\cdot)$ denote their distributions, and $v^g(\cdot)$ and $w^g(\cdot)$ their densities. In the sequel, it is assumed that $\lambda(\cdot), r(\cdot), B(\cdot)$ are chosen such that the steady-state distribution of the infinite-buffer version, that is, for $g(w, b, K) = w + b$, exists (and then for all $g(\cdot, \cdot, \cdot)$). For details on stability and existence of steady-state distributions, we refer to [8, 9].
Define

$$R(x) := \int_0^x \frac{1}{r(y)} \mathrm{d}y, \qquad 0 < x < \infty,$$

representing the time required for the system to become empty in the absence of any arrivals, starting with workload x. Note that $R(x) < \infty$, for all $x > 0$, means that state

zero can be reached in a finite amount of time from any state $x > 0$. A related quantity is

$$\Lambda(x) := \int_0^x \frac{\lambda(y)}{r(y)} \mathrm{d}y, \qquad 0 < x < \infty,$$

which determines whether the workload process of the infinite buffer queue has an atom at state zero. In case of finite buffers, some modification is required to regulate the workload behavior for states that can not be attained. Specifically, set $r(x) \equiv 1$ and $\lambda(x) \equiv \lambda$ for all $x > 0$ for which $\mathbb{P}(g(y, B, K) > x) = 0)$, for all $0 \leq y < x$. Then the workload process has indeed an atom at state zero if and only if $\Lambda(x) < \infty$ for all $0 < x < \infty$, as in the infinite-buffer queue. In case $\lambda(\cdot)$ is fixed ($\lambda(x) \equiv \lambda$), $\Lambda(x) = \lambda R(x)$ and we refer to Asmussen [3, Ch. XIV] and Brockwell *et al.* [8] for more details.
Furthermore, consider the interarrival time and its corresponding workload decrement, i.e., the amount of work finished during the interarrival time. Denote by A_y the conditional workload decrement during the interarrival interval starting with workload y, i.e., the event $\{A_y > v\}$ means that the workload is smaller than $y - v$ upon the arrival of the next customer. Note that the time required to move from y down to v in the absence of any arrivals equals $R(y) - R(v)$. Since $r(x) > 0$ for all $x > 0$, it follows that $R(\cdot)$ is strictly increasing, which implies a one-to-one correspondence between the interarrival time and its corresponding workload decrement.
The conditional distribution of the workload decrement during an interarrival interval was already obtained in [4]:

Proposition 2.1. *Let the workload just after an arrival be y ($g(w, b, K) = y$); then, for $y > v$,*

$$\mathbb{P}(A_y > v) = e^{-\int_{u=y-v}^{y} \frac{\lambda(u)}{r(u)} \mathrm{d}u}. \tag{1}$$

Turning back to the workload process $\{V^g(t), t \geq 0\}$, we may define the process right before jump epochs recursively, by

$$W_{n+1}^g = \max(g(W_n^g, B_n, K) - A_{n, g(W_n^g, B_n, K)}, 0), \tag{2}$$

where $A_{n,g(\cdot,\cdot,\cdot)}$ is the interarrival time between the n-th and $(n+1)$-th customer, depending on the workload right after the n-th jump epoch. In between jumps, the workload process is governed by the input rate function, and the process satisfies

$$\frac{\mathrm{d}V^g(t)}{\mathrm{d}t} = r(V^g(t)).$$

This concludes the description of the dynamics of the system. We refer to Harrison and Resnick [19] for a further discussion.

Special cases
As mentioned in Section 1, four important special cases of the general setting are queues with infinite buffers, finite-buffer dams (scenario f), systems with customer impatience (scenario i), and queues regulated by the complete rejection discipline (scenario c).
The model with infinite buffer size is discussed in [4] and is simply the model where every customer is completely accepted. The finite-buffer dam - regulated by the partial rejection discipline - originates from the study of water dams. The content of a dam is finite and additional water just overflows. In the context of queueing, this implies that the n-th

arriving customer is admitted to the system if and only if $W_n + B_n \leq K$. However, a partially rejected (not fully accepted) customer may enter the system, but with restricted service requirement $K - W_n$.
Models with customer impatience stem from classical queueing systems, with a server working at unit speed. In that case, the workload upon arrival identifies a waiting time and the impatience is represented by the fact that customers are willing to wait a limited amount of time K. In case of general service speeds, the n-th arriving customer is accepted if $W_n \leq K$ and fully rejected otherwise (see [7] for some potential applications). Finally, the system with complete rejections is probably one of the more difficult to analyze, although the rejection rule is simple: the n-th customer is admitted if $W_n + B_n \leq K$, and totally rejected otherwise.
Summarizing, these four scenarios may be represented as follows:

$$g(w,b,K) = \begin{cases} w+b, & \text{infinite-buffer queue,} \\ \min(w+b,K), & \text{scenario } f\text{; finite dam,} \\ w+bI(w\leq K), & \text{scenario } i\text{; customer impatience,} \\ w+bI(w+b\leq K), & \text{scenario } c\text{; complete rejection discipline.} \end{cases}$$

Here $I(\cdot)$ denotes the indicator function. Finally, we indicate the notational conventions arising from the models. If we consider an arbitrary $g(\cdot,\cdot,\cdot)$, we add an index g. The infinite-buffer system is denoted by just omitting the g from the definitions of the general model. The finite-buffer dam may be obtained by substituting K for g. The models with customer impatience and complete rejections are given by writing K,i and K,c for g, respectively.

3 Relations between two finite-buffer queues

In this section, we analyze the workload relations between two (general) finite-buffer queues that have the same ratio between arrival and service rate. It turns out that a similar relation as for infinite-buffer queues [4, Theorem 3.1] still holds (in fact, the infinite-buffer queue is just a special case of the general setting studied here: Choose $g(w,b,K) = w+b$, for all $w,b \geq 0$). In addition, the formal solution of the steady-state workload density is considered. However, we start with studying the relation between workloads at arrival instants and arbitrary epochs.
In view of loss probabilities, the relation between the workload at jump epochs and arbitrary epochs is an important one. The following theorem extends results for infinite-buffer queues [4, Theorem 3.2] to our setting.

Theorem 3.1. *$W^g(0) = \lambda(0)V^g(0)/\bar{\lambda}^g$, with $\bar{\lambda}^g := \int_{0^+}^{\infty} \lambda(x)v^g(x)\mathrm{d}x + \lambda(0)V^g(0)$, and for all $x > 0$,*

$$w^g(x) = \frac{1}{\bar{\lambda}^g}\lambda(x)v^g(x).$$

Proof. Observe that $g(w,b,K) \leq w+b$ ensures that the expected cycle length is finite and the workload process is thus ergodic. By level crossing theory, it then follows that the workload density is well-defined. Moreover, $g(w,b,K) \geq w$ rules out scenarios of work removal. Now, substitute $g(W^g,B,K)$ for every $W+B$ in [4, Theorem 5.1] and the results follow easily. □

Note that $\lambda(x) \equiv \lambda$ would yield the PASTA property, which states that the workload at an arbitrary time and the workload at an arrival epoch have the same distribution. With respect to workload as the key performance measure, Thereom 3.1 may be viewed as a generalization of the PASTA result.
In the infinite-buffer system, two models, with identical ratio between arrival and release rate, can be related (see [4, Theorem 3.1 and 3.3]). This relation between two different M/G/1 queues can be extended to the general model, as presented in the next theorem.

Theorem 3.2. *Consider two queues of the model of Section 2, to be denoted as Models* 1 *and* 2*, such that* $\lambda_1(x)/r_1(x) = \lambda_2(x)/r_2(x)$*, for all* $x > 0$*. Then,* $W_1^g(0) = W_2^g(0)$*, and for all* $x > 0$*,*

$$w_1^g(x) = w_2^g(x). \tag{3}$$

Also,

$$\frac{v_1^g(x)}{v_2^g(x)} = C\frac{r_2(x)}{r_1(x)}, \tag{4}$$

with $C = \frac{\lambda_1(0)V_1^g(0)}{\lambda_2(0)V_2^g(0)}$ *if* $\Lambda_i(x) < \infty$ *for all* $0 < x < \infty$*, and* $C = 1$ *if* $\Lambda_i(x) = \infty$ *for some* $0 < x < \infty$*.*

Before we prove the above theorem, we first derive the steady-state workload density. Besides that the formal solution of this density is a slight extension of infinite-buffer results, it turns out to be a useful tool to express the workload density in a more elegant form in some special cases. Moreover, Equation (4) follows then directly by division.
Now, consider either of the two models and assume for the moment that the workload process has an atom at state zero. We start by considering the level crossing equations (see [14] for a survey on level crossings),

$$r(x)v^g(x) = \lambda(0)V^g(0)\mathbb{P}(g(0,B,K) > x) + \int_{0^+}^{x} \lambda(y)v^g(y)\mathbb{P}(g(y,B,K) > x)\mathrm{d}y. \tag{5}$$

This equation reflects the fact that the rate of crossing level x from above should equal, in steady-state, the rate of crossing level x from below.
Note that in many finite-buffer systems, the workload is bounded by the capacity K, as in scenarios f and c. In that case, $\mathbb{P}(g(y,B,K) > K) = 0$, and we only have to consider $0 < x \leq K$. In, for example, scenario i, cases with workloads above K may exist. However, jumps occur only from workloads smaller than K, and the range of integration can be modified by $(0, \min(x,K)]$.
Define $z(y) := \lambda(y)v^g(y)$ and multiply both sides of (5) by $\lambda(x)/r(x)$. We then obtain

$$z(x) = \frac{\lambda(x)}{r(x)}\lambda(0)V^g(0)\mathbb{P}(g(0,B,K) > x) + \int_{0^+}^{x} \frac{\lambda(x)}{r(x)}z(y)\mathbb{P}(g(y,B,K) > x)\mathrm{d}y. \tag{6}$$

We now proceed as in Harrison & Resnick [19]: define the kernel $K^g(x,y) := K_1^g(x,y) := \mathbb{P}(g(y,B,K) > x)\lambda(x)/r(x), 0 \leq y < x < \infty$, and let its iterates be recursively defined by

$$K_{n+1}^g(x,y) := \int_y^x K^g(x,z)K_n^g(z,y)\mathrm{d}z. \tag{7}$$

Note that in, for instance, the infinite-buffer system, $\mathbb{P}(g(y,B,K)>x)=1-B(x-y)$. Moreover, observe that (6) is a Volterra integral equation of the second kind, and rewrite it as

$$z(x)=\lambda(0)V^g(0)K^g(x,0)+\int_{0^+}^{x} z(y)K^g(x,y)\mathrm{d}y. \tag{8}$$

Iterate this relation $N-1$ times (see [19]):

$$z(x)=\lambda(0)V^g(0)\sum_{n=1}^{N}K_n^g(x,0)+\int_{0^+}^{x} z(y)K_N^g(x,y)\mathrm{d}y.$$

Finally, define

$$K^{g,*}(x,y):=\sum_{n=1}^{\infty}K_n^g(x,y). \tag{9}$$

If this sum is well defined, we have $z(x)=\lambda(0)V^g(0)K^{g,*}(x,0)$. However, we may use the obvious bound $K^g(x,y)\leq\lambda(x)/r(x)$ to show inductively that $K_{n+1}(x,y)\leq(\Lambda(x,y))^n\lambda(x)/(r(x)n!)$. Hence, since $K_{n+1}(x,y)\to 0$, as $n\to\infty$ and the kernels are bounded for all $0<y<x<\infty$, the infinite sum is indeed well defined. Now, use the definition of $z(\cdot)$, to obtain

$$v^g(x)=\frac{\lambda(0)V^g(0)K^{g,*}(x,0)}{\lambda(x)}, \tag{10}$$

where $V^g(0)$ follows from normalization. The steady-state workload density is presented in the following lemma.

Lemma 3.1. *If $\Lambda(x)<\infty$ for all $0<x<\infty$, then*

$$v^g(x)=\frac{\lambda(0)V^g(0)K^{g,*}(x,0)}{\lambda(x)}, \tag{11}$$

where $V^g(0)=\left[1+\lambda(0)\int_0^\infty\frac{K^{g,*}(x,0)}{\lambda(x)}\mathrm{d}x\right]^{-1}$.

In Sections 4 and 6 it is indicated how this general approach may be applied to some (in)finite-buffer queues. Next, in Section 7, the infinite sum of Volterra kernels is explicitly calculated for some special cases. But, we first use this Lemma to derive Equation (4).

Proof of Theorem 3.2. Observe that, by (1) and (2), the dynamics of both systems are equivalent when $\frac{\lambda_1(x)}{r_1(x)}=\frac{\lambda_2(x)}{r_2(x)}$. Hence, using a stochastic coupling argument, the first part of the Theorem, that is (3), follows easily.
We now turn to (4). Note that $\Lambda_1(x)=\Lambda_2(x)$ implying that either the workload processes in both systems have an atom at state zero, or not. If $\Lambda_i(x)=\infty$ $(i=1,2)$ for some $x>0$, then $V_i^g(0)=0$ and (4) follows directly from (6) and the definition of $z(\cdot)$. So, assume that $V_i^g(0)>0$. We use the derivation of the steady-state workload density as described above. First, observe that the kernels $K^g(x,y)=\mathbb{P}(g(y,B,K)>x)\lambda(x)/r(x)$ are the same in both models, and hence, the iterated kernels and their infinite sums are equal. Now, use (11) and divide $v_1^g(x)$ by $v_2^g(x)$ to obtain

$$\frac{v_1^g(x)}{v_2^g(x)}=\frac{\lambda_1(0)V_1^g(0)}{\lambda_2(0)V_2^g(0)}\frac{\lambda_2(x)}{\lambda_1(x)},\qquad x>0.$$

Substituting $\lambda_2(x)/\lambda_1(x)=r_2(x)/r_1(x)$ completes the proof. □

Remark 3.1. *There are alternative ways to solve (5). We may also divide by $r(x)$ on both sides of Equation (5), or define $z(x) := r(x)v^g(x)$. The technique to solve the integral equation remains the same, however, with slightly different kernels.*

4 Finite-buffer dam

In this section, we study the steady-state workload distribution in the finite-buffer M/G/1 dam with general arrival rate and service speed (scenario f). The dynamics of this model are as in Section 2, where $g(w, b, K) = \min(w + b, K)$. As indicated, the steady-state workload random variables, distributions, and densities are denoted by substituting K for g in Section 2, i.e., $V^K, V^K(\cdot), v^K(\cdot)$ at arbitrary times and $W^K, W^K(\cdot), w^K(\cdot)$ just before arrival epochs. The model with infinite buffer is denoted by just omitting the g from the notation. For convenience, we refer to scenario f as the finite-buffer queue or dam in this section.

First, we show that the steady-state workload distributions in the finite and infinite-buffer dam are proportional. For instance, Hooghiemstra [20] based his proof for the classical M/G/1 queue on the idea that the finite and infinite-buffer queue behave according to similar sample paths below workload level K. He argued that at a downcrossing of level K in the infinite-buffer queue, the time until the next arrival epoch is independent of the previous arrival, and hence, the residual interarrival time behaves like an ordinary one. As required in the sample path comparison, we show that this lack of memory also holds for general M/G/1-type queues with state-dependent arrival and service rates. After making some comments about regenerative properties, the result of Hooghiemstra is extended to our system, following similar arguments. Moreover, the steady-state workload distribution at arrival epochs is considered. This is no longer necessarily equal to the workload distribution at arbitrary epochs, since the classical PASTA property no longer holds.

Second, as an example of the general case and as a prelude to Section 5, we derive the steady-state workload density for the finite dam, using level crossing arguments and the Volterra successive substitution method. And third, we consider the long-run fraction of not fully accepted customers, denoting this performance measure by P_K.

The next preparatory lemma presents the lack-of-memory property of the workload decrement during an interarrival interval.

Lemma 4.1. (**Memoryless property**). *The residual workload decrement at a downcrossing of level x in an M/G/1 queue with arrival rate $\lambda(\cdot)$ and service rate $r(\cdot)$ is independent of the finished amount of work during the elapsed interarrival time, i.e.,*

$$\mathbb{P}(A_{x+y} > y + v | A_{x+y} > y) = \mathbb{P}(A_x > v), \qquad x, y, v > 0, \quad x > v$$

Proof. Using a simple conditioning argument and Proposition 2.1, it follows that

$$\begin{aligned} \mathbb{P}(A_{x+y} > y + v | A_{x+y} > y) &= e^{-\int_{x-v}^{x+y} \frac{\lambda(u)}{r(u)} \mathrm{d}u} e^{\int_{x}^{x+y} \frac{\lambda(u)}{r(u)} \mathrm{d}u} \\ &= e^{-\int_{x-v}^{x} \frac{\lambda(u)}{r(u)} \mathrm{d}u} \\ &= \mathbb{P}(A_x > v). \end{aligned}$$

Notice that $\mathbb{P}(A_{x+y} > y+v | A_{x+y} > y)$ is independent of y, representing the lack-of-memory property. □

Next, we state our main proportionality result.

Theorem 4.1. *For* $0 \leq x \leq K$,

$$\mathbb{P}(V^K \leq x) = \frac{\mathbb{P}(V \leq x)}{\mathbb{P}(V \leq K)}, \tag{12}$$

while at arrival epochs,

$$\mathbb{P}(W^K \leq x) = \frac{\mathbb{P}(W \leq x)}{\mathbb{P}(W \leq K)}.$$

Before we prove the theorem, we first make some general remarks about regenerative processes. Instead of applying level crossing arguments and using Lemma 3.1, it is also possible to make a direct comparison between the finite and infinite-buffer queue. We apply the latter approach. Following Asmussen [3], we exploit the regenerative character of the workload process and let the points with workload level K be its regeneration points. Note that this is possible due to the memoryless property (Lemma 4.1). Furthermore, this choice allows queueing systems where empty queues cannot occur. Denote the length of a regeneration cycle in the finite and infinite-buffer queue and the number of arrivals during this cycle, by τ^K, τ, N^K, and N, respectively. Then, the distributions of V and V^K are given by, cf. [11],

$$\mathbb{P}(V \leq x) = \frac{1}{\mathbb{E}\tau}\mathbb{E}\left[\int_0^{\tau} I(V(t) \leq x)\mathrm{d}t\right], \tag{13}$$

$$\mathbb{P}(V^K \leq x) = \frac{1}{\mathbb{E}\tau^K}\mathbb{E}\left[\int_0^{\tau^K} I(V^K(t) \leq x)\mathrm{d}t\right]. \tag{14}$$

The distributions of W and W^K can be obtained in a similar fashion, cf. [11],

$$\mathbb{P}(W \leq x) = \frac{1}{\mathbb{E}N}\mathbb{E}\left[\sum_{i=1}^{N} I(W_i \leq x)\right],$$

$$\mathbb{P}(W^K \leq x) = \frac{1}{\mathbb{E}N^K}\mathbb{E}\left[\sum_{i=1}^{N^K} I(W_i \leq x)\right].$$

We are now ready to prove our main theorem.

Proof of Theorem 4.1. Consider the stochastic process $\{V(t), t \geq 0\}$. We construct a stochastic process $\hat{V}^K(t)$ directly from $V(t)$ and show that $\hat{V}^K(t)$ and $V^K(t)$ are driven by the same dynamics. First, take an arbitrary sample path of $V(t)$. We leave the parts below level K unchanged and cut out the parts of the sample path between each upcrossing and a consecutive downcrossing of level K. Connecting the remaining parts, we obtain the process $\hat{V}^K(t)$. By the memoryless property, the workload decrement of $\hat{V}^K(t)$ after hitting K behaves like an ordinary workload decrement. Thus $\hat{V}^K(t)$ and $V^K(t)$ are driven by the same dynamics and we may simplify notation by identifying the process $\{V^K(t), t \geq 0\}$ with $V^K(t) := \hat{V}^K(t)$, $t \geq 0$.
Clearly, $\mathbb{E}\left[\int_0^{\tau} I(V(t) \leq x)\mathrm{d}t\right]$ and $\mathbb{E}\left[\int_0^{\tau^K} I(V^K(t) \leq x)\mathrm{d}t\right]$ are equal. Observe that $\frac{\mathbb{E}\tau^K}{\mathbb{E}\tau}$ represents the long-run fraction of time that the workload process of the infinite-buffer

queue is below level K and, by (13) and (14), we have shown the first part of the theorem. The second part follows directly from the same sample path construction and the observation that $\frac{\mathbb{E}N^K}{\mathbb{E}N}$ equals the long-run fraction of arrivals finding the workload below level K. This concludes our proof. □

Remark 4.1. *Theorem 4.1 remains valid for any other model where the virtual waiting time process remains unchanged below level K (Hooghiemstra [20] noted this already for the classical M/G/1 queue). Specifically, the theorem applies for any function $g(w, b, K)$, if for all $w, b, K \geq 0$ the function satisfies*

$$\begin{array}{ll} g(w,b,K) = w+b, & \text{if } w+b \leq K, \\ g(w,b,K) \geq K, & \text{if } w+b > K. \end{array} \tag{15}$$

We give another example of such a system in Section 6.

In the remainder of the paper, we assume that the workload process has an atom at state zero, i.e., $\Lambda(x) < \infty$, for all $0 < x < \infty$.
Both as an example and as a prelude to Section 5, we now further examine the steady-state workload density $v^K(\cdot)$. This density is well-known in case of general service rate functions and a constant arrival rate, see e.g. [19]. A similar technique - also used in Section 3 - may be used in case of a general arrival rate (see also [4]). We briefly indicate how Lemma 3.1 may be used in scenario f.
We start by considering the level crossing equations,

$$r(x)v^K(x) = \lambda(0)V^K(0)(1-B(x)) + \int_{0^+}^{x} \lambda(y)v^K(y)(1-B(x-y))\mathrm{d}y, \qquad 0 < x \leq K. \tag{16}$$

This equation is well-known and follows directly from an up- and downcrossing argument. Also, (16) may be straightforwardly obtained from the general case (16) by observing that $\mathbb{P}(g(y, B, K) > x) = 1 - B(x - y)$ for $0 \leq y < x < K$ and 0 otherwise. Now, it is evident that we may define the kernel (as in [19]) by $K(x, y) := (1 - B(x - y))\lambda(x)/r(x), 0 \leq y < x < K$. In the remainder, we refer to this kernel as the basic kernel as it appears in many queueing systems. The iterates of $K(x, y)$ and the infinite sum are defined as in Section 3, i.e., (7) and (9) respectively.
Now, applying Lemma 3.1 yields the well-known representation of the workload density for scenario f:

$$v^K(x) = \frac{\lambda(0)V^K(0)K^*(x,0)}{\lambda(x)}, \qquad 0 < x \leq K, \tag{17}$$

where $V^K(0)$ follows from normalization. Note that (12) can thus also be derived from the formal solution of the density. However, we believe that the derivation of Theorem 4.1 is especially insightful as it brings out the typical sample-path relation between the infinite buffer queue and the finite dam (scenario f).
We conclude this section by analyzing the probability that a customer cannot be completely accepted, also referred to as loss probability. It follows directly from a regenerative argument, see also [31], that

$$P_K = \mathbb{P}(W^K + B > K).$$

Condition on W^K and apply Theorem 3.1 to obtain the following corollary:

Corollary 4.1. *For the loss probability in scenario f, we have*

$$P_K = \frac{1}{\overline{\lambda}^K}\left[\lambda(0)V^K(0)(1-B(K)) + \int_{0^+}^{K} \lambda(y)v^K(y)(1-B(K-y))\mathrm{d}y\right], \qquad (18)$$

with $v^K(\cdot)$ given in (17) and $V^K(0)$ equal to $1 - \int_0^K v^K(x)\mathrm{d}x$.

Note that other performance measures may also be directly obtained from the workload density and Theorem 3.1.

5 First exit probabilities and cycle maxima

In this section, we focus on queues with infinite buffer capacity and determine first exit probabilities and the distribution of the cycle maximum. To do so, we use the finite-buffer dam, analyzed in Section 4, to a large extent. Moreover, we show that first-exit probabilities are related to the dual of a finite dam. Also observe that, for well-chosen K, first-exit probabilities are the same for a range of finite-buffer models, such as the scenario f (use Remark 4.1).

Consider the model with arrival rate 1, and release rate $\hat{r}(x) := \frac{r(x)}{\lambda(x)}$ when the workload equals x. Theorem 3.1 shows that both models have the same workload density at arrival epochs, $w(\cdot)$. As a consequence, the amounts of work just after an arrival instant follow the same distribution as well. Also, observe that the workload process $\{V(t), t \geq 0\}$ attains local minima just before a jump and local maxima right after a jump. Considering first-exit probabilities, it then easily follows that we may consider (without loss of generality) a model with arrival rate 1, and release rate $\hat{r}(x)$. In fact, the same argument holds for cycle maxima, as it may be considered to be a special case of a first-exit probability. So, in this section we often assume, without loss of generality, that the arrival rate equals 1.

Starting with first-exit probabilities, we assume that $0 \leq a < b < \infty$, and let $\tau(a) := \inf\{t > 0 | V(t) \leq a\}$ and $\tau(b) := \inf\{t > 0 | V(t) \geq b\}$ correspond to the first-exit times[1]. Starting from x, we denote the probability that the workload process hits state b before state a by $U(x)$, i.e., $U(x) := \mathbb{P}_x(\tau(b) < \tau(a))$. Now, the first-exit probabilities can be obtained from those in models with constant arrival rate (in particular [19]) and the observation above. Define

$$\alpha(a,b) := \left[1 + \frac{r(b)}{\lambda(b)}\int_a^b \frac{\lambda(x)}{r(x)}K^*(b,x)\mathrm{d}x\right]^{-1}. \qquad (19)$$

We obtain the following lemma:

Lemma 5.1. *We have,*

$$U(x) = \begin{cases} 0, & \text{if } 0 \leq x \leq a, \\ \int_a^x u(y)\mathrm{d}y, & \text{if } a < x \leq b, \\ 1, & \text{if } x > b, \end{cases}$$

where $u(x) = \alpha(a,b)r(b)\lambda(x)K^*(b,x)/(\lambda(b)r(x))$ for $x \in (a,b)$.

[1]Note that we use b here in a different fashion as in Sections 2-4. An alternative notation for b could be K, but we decided to follow the literature on first-exit probabilities and use b in the context of hitting times.

Proof. Apply [19, Theorem 3] to the dam with release rate $\hat{r}(\cdot)$. □

Remark 5.1. *In fact, first-exit probabilities with $a > 0$ may be reduced to a similar first exit probability with $a = 0$. Modify the system to a finite-buffer dam of capacity $b - a$ and with release rate $\breve{r}(x) := r(x + a)$ when the workload equals x. Denote the modified first hitting times by $\breve{\tau}(0)$ and $\breve{\tau}(b - a)$, and let, for $x \in (0, b - a]$, $\breve{U}(x) := \mathbb{P}_x(\breve{\tau}(b - a) < \breve{\tau}(0))$ be the probability that the modified system hits state $b - a$ before state 0, starting from x. Then, apply Lemma 5.1 to the modified system (thus with release rate $\breve{r}(\cdot)$). Note that $\breve{K}(x, y) = (1-B(x-y))/\breve{r}(x) = K(x+a, y+a)$, and it can be easily shown (by induction) that $\breve{K}_n(x, y) = K_n(x + a, y + a)$. Now, it is just straightforward calculation to show that $\breve{U}(x - a) = U(x)$.*

Concerning cycle maxima, we assume that at time 0 a customer enters an empty system and define $C_{\max} := \sup\{0 \le t \le \tau(0) : V(t)\}$. Denote $\tilde{r}(x) := \hat{r}(b - x) = r(b - x)$ and let $P_b^{\tilde{r}(\cdot)}$ be the loss probability in a finite dam (scenario f) with release rate $\tilde{r}(\cdot)$. The following relation between cycle maxima and loss probabilities has been obtained in [5]:

Lemma 5.2. *We have,*

$$\mathbb{P}(C_{\max} \ge b) = P_b^{\tilde{r}(\cdot)}.$$

Motivated by this relation, we first analyze scenario f with arrival rate 1 and release rate $\tilde{r}(\cdot)$ in more detail. This turns out to be a useful tool to determine the distribution of the cycle maximum in general terms.

Let $\tilde{v}(\cdot)$ denote the steady-state workload density of the model with arrival rate 1 and release rate $\tilde{r}(\cdot)$. Using level crossing arguments, we have, for $0 < x < b$,

$$\tilde{r}(x)\tilde{v}(x) = \tilde{V}(0)(1 - B(x)) + \int_{0^+}^{x} \tilde{v}(y)(1 - B(x - y))\mathrm{d}y.$$

Define $z(x) := \tilde{r}(x)\tilde{v}(x)$, and Volterra kernel $\tilde{K}(x, y) := (1 - B(x - y))/\tilde{r}(y)$, for $0 < y < x < b$, and $\tilde{K}(x, 0) := 1 - B(x)$ for $0 < x < b$. Observe that we can relate $\tilde{K}(x, y)$ to the basic kernel in Section 4. Specifically, for $0 < y < x < b$,

$$\tilde{K}(x, y) = (1 - B(b - y - (b - x)))/r(b - y) = K(b - y, b - x),$$

and for $0 < x < b$,

$$\tilde{K}(x, 0) = 1 - B(b - (b - x)) = K(b, b - x)r(b).$$

Now, using the successive substitution method for Volterra kernels as in Section 3, yields (for $0 < x < b$),

$$\begin{aligned}
z(x) &= \tilde{V}(0)\tilde{K}(x, 0) + \int_{0^+}^{x} z(y)\tilde{K}(x, y)\mathrm{d}y \\
&= \tilde{V}(0)K(b, b - x)r(b) + \int_{0^+}^{x} z(y)K(b - y, b - x)\mathrm{d}y \\
&= \tilde{V}(0)K(b, b - x)r(b) + \int_{0^+}^{x} [K(b, b - y)K(b - y, b - x)r(b)\tilde{V}(0) \\
&\quad + \int_{0}^{y} K(b - u, b - y)K(b - y, b - x)z(u)\mathrm{d}u]\mathrm{d}y \\
&= K_1(b, b - x)r(b)\tilde{V}(0) + K_2(b, b - x)r(b)\tilde{V}(0) + \int_{0}^{x} z(u)K_2(b - u, b - x)\mathrm{d}u,
\end{aligned}$$

where the last equality follows from Fubini's theorem and

$$\begin{aligned}\int_u^x K(b-u,b-z)K_n(b-z,b-x)\mathrm{d}z &= \int_{b-x}^{b-u} K(b-u,z)K_n(z,b-x)\mathrm{d}z \\ &= K_{n+1}(b-u,b-x).\end{aligned}$$

Iterating this argument gives

$$z(x) = r(b)\tilde{V}(0)K^*(b,b-x).$$

Finally, use the definition of $z(\cdot)$ to express the steady-state density of the model with release rate $\tilde{r}(\cdot)$ into the original model (with release rate $r(\cdot)$):

$$\tilde{v}(x) = \frac{\tilde{V}(0)r(b)K^*(b,b-x)}{r(b-x)}, \tag{20}$$

where $\tilde{V}(0)$ follows from normalization.
Returning to the cycle maximum of our original model, we have, by Lemma 5.2:

Theorem 5.1. *For the cycle maximum in an M/G/1-type dam, with arrival rate $\lambda(\cdot)$ and release rate $r(\cdot)$, we have*

$$\mathbb{P}(C_{\max} \geq b) = \tilde{V}(0)(1-B(b)) + \int_0^b \tilde{v}(x)(1-B(b-x))\mathrm{d}x, \tag{21}$$

where

$$\tilde{v}(x) = \frac{\tilde{V}(0)r(b)K^*(b,b-x)\lambda(b-x)}{\lambda(b)r(b-x)}, \tag{22}$$

and $\tilde{V}(0) = [1 + \int_0^b r(b)K^*(b,b-x)\lambda(b-x)(\lambda(b)r(b-x))^{-1}\mathrm{d}x]^{-1}$.

We give two different proofs of the above theorem; the first one uses the equivalence between cycle maxima and loss probabilities, and the second exploits knowledge of first-exit probabilities.

Proof I (via $P_b^{\tilde{r}(\cdot)}$). To prove Theorem 5.1, we use the relation between loss probabilities and cycle maxima, Lemma 5.2. We already analyzed $P_b^{r(\cdot)}$ in Section 4 (Corollary 4.1). Use the fact that the cycle maximum only depends on $\lambda(\cdot)$ and $r(\cdot)$ via their ratio and note that the steady-state density for the model with release rate $\tilde{r}(x) = \frac{r(b-x)}{\lambda(b-x)}$ are then given by (20). Applying (18) to the model with $\lambda = 1$ and release rate $\tilde{r}(\cdot)$ gives the result. □

Proof II (via $U(x)$). First note that $\alpha(0,b) = \tilde{V}(0)$, by substituting $u = b - x$ in (19). Then, given the service requirement of a customer entering an empty system, the cycle maximum may be rewritten as a first exit probability. Specifically, condition on the service requirement of the "first customer", and use Lemma 5.1 in the third equality:

$$\begin{aligned}\mathbb{P}(C_{\max} \geq b) &= \int_{x=0}^{\infty} U(x)\mathrm{d}B(x) \\ &= \int_{x=0}^{b}\int_{y=0}^{x} u(y)\mathrm{d}y\mathrm{d}B(x) + (1-B(b)) \\ &= \int_{y=0}^{b} \alpha(0,b)\frac{r(b)\lambda(y)K^*(b,y)}{r(y)\lambda(b)}\int_{x=y}^{b}\mathrm{d}B(x)\mathrm{d}y + (1-B(b)) \\ &= \int_{y=0}^{b} \alpha(0,b)\frac{r(b)\lambda(y)K^*(b,y)}{r(y)\lambda(b)}(1-B(y))\mathrm{d}y + \alpha(0,b)(1-B(b)),\end{aligned}$$

where the final step follows from (cf. (19))

$$1 = \alpha(0,b) + \int_0^b \alpha(0,b)\frac{r(b)\lambda(y)K^*(b,y)}{r(y)\lambda(b)}\mathrm{d}y.$$

The theorem now follows by straightforward substitution. □

Alternatively, the first-exit probabilities, given by Harrison and Resnick [19], may also be related to a finite dam with release rate $\tilde{r}(x) = \hat{r}(b-x)$ when the workload equals x. For $a = 0$ and $0 \le x \le b$, we use the steady-state workload density (22) directly:

$$\begin{aligned} 1 - \tilde{V}(x) &= \int_{y=x}^b \tilde{v}(y)\mathrm{d}y \\ &= \int_{y=x}^b \alpha(0,b)\frac{r(b)K^*(b,b-y)\lambda(b-y)}{\lambda(b)r(b-y)}\mathrm{d}y \\ &= \int_{u=0}^{b-x} \alpha(0,b)\frac{r(b)K^*(b,u)\lambda(u)}{\lambda(b)r(u)}\mathrm{d}u = U(b-x), \end{aligned} \tag{23}$$

where $V(0) = \alpha(0,b)$ follows from Lemma 5.1 and Theorem 5.1. Using Remark 5.1, we may generalize this equivalence relation to cases with $a > 0$.

Lemma 5.3. *Let $\tilde{V}(\cdot)$ be the workload distribution of the finite-buffer system (scenario f) of capacity $b-a$ and release rate $\tilde{r}(\cdot)$. Then, for $x \in [0, b-a]$,*

$$1 - \tilde{V}(x) = U(b-x).$$

Proof. Consider the system with finite buffer $b-a$ and release rate $\tilde{r}(x) = r(b-x)$ when the workload equals x. Note that a modification of $\tilde{r}(\cdot)$ to the case $a = 0$ (as in Remark 5.1) is not required, since $\tilde{r}(x) = \hat{r}((b-a-x)+a)$.
Although the workload is upperbounded by $b-a$, we can use exactly the same analysis as if it was bounded by b, and express the steady-state workload density as follows:

$$\tilde{v}(x) = \frac{\tilde{V}(0)\hat{r}(b)K^*(b,b-x)}{\hat{r}(b-x)},$$

where

$$\tilde{V}(0) = \left[1 + \int_{x=0}^{b-a} \frac{\hat{r}(b)K^*(b,b-x)}{\hat{r}(b-x)}\mathrm{d}x\right]^{-1}, \tag{24}$$

follows from normalization. Substitute $y = b-x$ and $\hat{r}(x) = \frac{r(x)}{\lambda(x)}$ in (24), to see that

$$\tilde{V}(0) = \left[1 + \int_{y=a}^{b} \frac{r(b)K^*(b,y)\lambda(y)}{\lambda(b)r(y)}\mathrm{d}y\right]^{-1} = \alpha(a,b).$$

Now, using the same argument as in (23), with substitution $\hat{r}(y) = r(y)/\lambda(y)$ and $u = b-y$, completes the proof. □

Remark 5.2. *We conjecture that (a modified) Lemma 5.3 also holds in case of general i.i.d. interarrival times and a general service rate $r(\cdot)$ if we start with a regular interarrival interval. Denote by $\tilde{W}(\cdot)$ the workload distribution right before an arrival instant in the finite-buffer system of capacity $b-a$ and service rate $\tilde{r}(\cdot)$. Then, using the machinery of monotone stochastic recursions [1] and a similar construction as in [5], we may show that*

$$1 - \tilde{W}(x) = U(b-x).$$

6 Other finite-buffer systems

There are many finite-buffer systems, which may be distinguished by the rejection discipline of arriving customers. Among the most important ones is the finite-buffer dam regulated by the partial rejection discipline (scenario f), see Section 4. Two other finite-buffer systems of importance are models with customer impatience (scenario i) and queues governed by the complete rejection discipline (scenario c). These models are investigated in this section.

We begin this section by examining scenario i and observing that for workloads less than K the proportionality result (Theorem 4.1) holds. To determine the density of workloads larger than K, in addition to the normalizing constant, we apply level crossings and the successive substitution method for Volterra integral equations. We conclude the study of scenario i by considering the loss probability. Turning to the second model, scenario c, we derive a formal solution for the steady-state workload density using similar techniques as for scenario i. Finally the loss probability in scenario c is considered.

We start with scenario i, or equivalently, let $g(w,b,K) = w + bI(w < K)$ in the general set-up. Note that there exist $w, b, K \geq 0$ such that $g(w,b,K) > K$, implying that workload levels above K may occur. In particular, only customers arriving at the system while the workload just before arrival is larger than K are rejected. This resembles the impatience of arriving customers: they are only willing to wait a maximum (stochastic) amount of time. Moreover, as noted in [7, 20], the virtual waiting time process below level K remains unchanged for this model. This intuitive statement can be made rigorous by observing that for all $0 \leq w \leq K$, $g(w,b,K) = w + b$ and thus (15) is satisfied. We consequently have the following (see also Remark 4.1):

Corollary 6.1. *For $0 \leq x \leq K$, we have,*

$$\mathbb{P}(V^{K,i} \leq x) = c^{K,i}\mathbb{P}(V \leq x), \tag{25}$$

with $c^{K,i}$ some normalizing constant, $\mathbb{P}(V \leq K) \leq (c^{K,i})^{-1} \leq 1$, while at arrival epochs of accepted customers (thus given $W^{K,i} \leq K$),

$$\mathbb{P}(W^{K,i} \leq x | W^{K,i} \leq K) = \frac{\mathbb{P}(W \leq x)}{\mathbb{P}(W \leq K)}.$$

Remark 6.1. *By a simple division and using (25) and (12) twice, we may alternatively write, for $0 \leq x, y \leq K$,*

$$\frac{\mathbb{P}(V \leq x)}{\mathbb{P}(V \leq y)} = \frac{\mathbb{P}(V^{K,i} \leq x)}{\mathbb{P}(V^{K,i} \leq y)} = \frac{\mathbb{P}(V^{K} \leq x)}{\mathbb{P}(V^{K} \leq y)}. \tag{26}$$

In fact, by (25), the workload distribution in the infinite-buffer case does not completely determine the workload distribution in scenario i. The normalizing constant can only be determined by knowledge of the workload behavior on all possible levels of the workload process. For $x > K$, however, there seems to be no direct link to the infinite-buffer queue. Level crossing arguments, i.e., equivalence between up- and downcrossings of a fixed level $x > 0$, provide a tool to resolve this issue.

Next, we derive the steady-state workload distribution for all $x \geq 0$ in scenario i, using the general approach described in Section 3. If the workload upon arrival is below level K, thus $0 \leq w \leq K$, then we just have $\mathbb{P}(g(w,B,K) > x) = 1 - B(x-w)$, while

$\mathbb{P}(g(w,B,K) > w) = 0$ otherwise. The general level crossing equations can now be rewritten into a more appealing expression: For $0 < x \leq K$, we have

$$r(x)v^{K,i}(x) = \lambda(0)V^{K,i}(0)(1-B(x)) + \int_{0^+}^{x} \lambda(y)v^{K,i}(y)(1-B(x-y))\mathrm{d}y,$$

and for $x > K$,

$$r(x)v^{K,i}(x) = \lambda(0)V^{K,i}(0)(1-B(x)) + \int_{0^+}^{K} \lambda(y)v^{K,i}(y)(1-B(x-y))\mathrm{d}y. \tag{27}$$

The level crossing equations can be solved using Lemma 3.1, by defining the kernel $K^i(x,y) := I(y < K)(1-B(x-y))\lambda(x)/r(x)$, for $0 \leq y < x < \infty$. In case $0 < y \leq K$, we just obtain our basic kernel $K(x,y)$ of Section 4. By Lemma 3.1, it is thus evident that, for $0 \leq x \leq K$,

$$z^{K,i}(x) = \lambda(0)V^{K,i}(0)K^*(x,0), \tag{28}$$

where $z^{K,i}(x) := \lambda(x)v^{K,i}(x)$. The same result can be deduced from Theorem 4.1 and (17).
The case $x > K$ may be derived in a slightly more elegant fashion; rewrite (27) into

$$z^{K,i}(x) = \lambda(0)V^{K,i}(0)K(x,0) + \int_{0^+}^{K} z^{K,i}(y)K(x,y)\mathrm{d}y. \tag{29}$$

Using the result of $z^{K,i}(y)$ for $y \leq K$ and substituting this in (29), we have

$$z^{K,i}(x) = \lambda(0)V^{K,i}(0)\left[K(x,0) + \int_{0}^{K} K(x,y)K^*(y,0)\mathrm{d}y\right],$$

after which $v^{K,i}(x) = z^{K,i}(x)/\lambda(x)$ and $V^{K,i}(0)$ can be determined by normalization.
For completeness, we give the resulting normalizing constant in general terms (take $y = 0$ in (26)):

$$c^{K,i} = \frac{1 + \int_0^\infty \frac{\lambda(0)}{\lambda(x)}K^*(x,0)\mathrm{d}x}{1 + \int_{x=0}^{K} \frac{\lambda(0)}{\lambda(x)}K^*(x,0)\mathrm{d}x + \int_{x=K}^{\infty}\left[\frac{\lambda(0)}{r(x)}(1-B(x)) + \int_{y=0}^{K}\frac{\lambda(0)}{r(x)}(1-B(x-y))K^*(y,0)\mathrm{d}y\right]\mathrm{d}x}.$$

Remark 6.2. *The cases $0 < x \leq K$ and $x > K$ may be combined by writing*

$$v^{K,i}(x) = \frac{\lambda(0)V^{K,i}(0)}{\lambda(x)}\left[K(x,0) + \int_{0}^{x\wedge K} K(x,y)K^*(y,0)\mathrm{d}y\right]. \tag{30}$$

Equation (28) can then be recovered by using $K^*(y,0) = \sum_{n=1}^{\infty} K_n(y,0)$ and interchanging integral and sum.

Finally, it is an easy exercise to determine the long-run fraction of rejected customers P_K^i. After all, the customers that are rejected are just those that arrive while the workload is above level K, or more formally $P_K^i = \mathbb{P}(W^{K,i} > K)$. Apply Theorem 3.1 to see that

$$\begin{aligned} P_K^i &= \int_K^\infty w^{K,i}(x)\mathrm{d}x \\ &= \frac{1}{\bar{\lambda}^{K,i}}\int_K^\infty \lambda(x)v^{K,i}(x)\mathrm{d}x. \end{aligned} \tag{31}$$

We now turn to scenario c. This system is also a special case of the general set-up and can be obtained by taking $g(w,b,K) = w+bI(w+b \leq K)$. Note that there is no $w \in [0,K]$ and $b, K \geq 0$ such that $g(w,b,K) > K$. This implies that, starting from initial workload below K, the workload process is bounded by its buffer content and we only have to analyze workloads below K.

The proportionality result, as presented in Theorem 4.1, does not hold for this scenario. Combined with Remark 4.1, this is obvious from the fact that $g(w,b,K) = w < K$ for $w \in [0,K)$ with $w + b > K$. Intuitively, the workload process below level K is indeed affected if a customer arrives that would cause a workload above the buffer content (in which case that customer is completely rejected). However, we can still solve the level crossing equations to determine the steady-state workload density.

Denote the steady-state workload density by $v^{K,c}(\cdot)$. Observe that an upcrossing of level x occurs, if at levels $y < x$ a customer arrives that has a service requirement larger than $x-y$, but smaller than $K-y$. Specifically, $\mathbb{P}(g(y,B,K) > x) = 1-B(x-y)-(1-B(K-y)) = B(K-y)-B(x-y)$. The level crossing equation may then be rewritten as follows. For $0 < x < K$,

$$r(x)v^{K,c}(x) = \lambda(0)V^{K,c}(0)(B(K)-B(x)) + \int_{0^+}^{x} \lambda(y)v^{K,c}(y)(B(K-y)-B(x-y))\mathrm{d}y. \quad (32)$$

In view of (8), we define the Volterra kernel as $K^c(x,y) := (B(K-y)-B(x-y))\frac{\lambda(x)}{r(x)}$, $0 \leq y < x < K$. Using Lemma 3.1 (with respective iterates and infinite sum), we can directly write

$$v^{K,c}(x) = \frac{\lambda(0)V^{K,c}(0)K^{c,*}(x,0)}{\lambda(x)}, \qquad 0 < x < K.$$

Determining $V^{K,c}(0)$ by normalization concludes the derivation of the steady-state workload distribution.

Finally, we focus on the long-run fraction of rejected customers, P_K^c. By definition, a customer is rejected if, upon arrival, the workload present in addition to the service requirement exceeds the buffer capacity K. Then, conditioning on the workload just before a customer arrival and using Theorem 3.1 in the second equation, we have

$$\begin{aligned} P_K^c &= W^{K,c}(0)(1-B(K)) + \int_{0^+}^{K} w^{K,c}(x)(1-B(K-x))\mathrm{d}x \quad (33) \\ &= \frac{1}{\bar{\lambda}^{K,c}}\left[\lambda(0)V^{K,c}(0)(1-B(K)) + \int_{0^+}^{K} \lambda(x)v^{K,c}(x)(1-B(K-x))\mathrm{d}x\right]. \end{aligned}$$

Remark 6.3. *Note that the loss probabilities P_K^i and P_K^c only depend on the ratio between $\lambda(\cdot)$ and $r(\cdot)$. This is a direct consequence of Equations (31) and (33) in addition to Theorem 3.2. At the intuitive level, this is evident from the fact that changing between Models 1 and 2 (in which $\lambda_1(x)/r_1(x) = \lambda_2(x)/r_2(x)$, for all $x > 0$) is just a rescaling of time. This property was obtained in [4, 28] and applied in Section 5. So, to avoid explicit calculations of, for instance, $\bar{\lambda}^{K,i}$ and $\bar{\lambda}^{K,c}$, we may assume (without loss of generality) that the arrival rate is fixed.*

7 Some examples

In Sections 2-6 we expressed the steady-state workload densities, first-exit probabilities, and cycle maxima in terms of an infinite sum of Volterra kernels. Numerical methods to compute these sums are widely available, see for example [21, 24]. Since we obtained closed-form expressions for the performance measures of interest, we are done from a practical point of view. However, for some special cases, the Volterra integral equations reduce to an analytically tractable expression.
In this section, we discuss some special cases and show that several known results can be recovered from the Volterra kernels. In addition, we derive some results that appear to be new. We first discuss the case of constant arrival and service rate and then continue with the case of exponential service requirements. We conclude with a remark on the extension to rejection rules based on a stochastic barricade.

7.1 Constant arrival and service rate

Suppose that $r(x) \equiv r > 0$ and $\lambda(x) \equiv \lambda r > 0$. Observe that, using Theorems 3.1 and 3.2, we may assume that $r(x) \equiv 1$ and the model thus reduces to an ordinary M/G/1 queue. Denote the arrival rate by λ, the mean service requirement by β, and let $\rho := \lambda\beta$ be the load of the system. Also, let

$$H(x) := \beta^{-1} \int_0^x (1 - B(y))\mathrm{d}y,$$

denote the stationary residual service requirement distribution with density $h(\cdot)$.
In the M/G/1 case, the basic kernel $K(x, y)$ reduces to $\lambda(1-B(x-y))$ and it is well-known, see for instance [19], that,

$$K^*(x, y) = \sum_{n=1}^{\infty} \rho^n h_n(x - y). \tag{34}$$

Here, $h_n(\cdot)$ is the density of the n-fold convolution $H_n(\cdot)$. Now, combine Lemma 3.1 with (34) and take Laplace Transforms to obtain the famous Pollaczek-Khinchine formula. The finite-dam is just the truncated version (use Theorem 4.1).
Turning to the model with customer impatience (scenario i), the normalizing constant in Corollary 6.1 may be determined by (26) and an application of Little's law. First apply (26) with $y = 0$:

$$\mathbb{P}(V^{K,i} \leq x) = \mathbb{P}(V \leq x)\frac{V^{K,i}(0)}{V(0)}.$$

Then, use Little in the first and (31) and PASTA in the second equation (see also [23]), to obtain,

$$V^{K,i}(0) = 1 - \rho(1 - P_K^i) = 1 - \rho V^{K,i}(K).$$

Apply (26) again to $V^{K,i}(K)$ (and use $V(0) = 1 - \rho$), then, after some rewriting, we may express the steady-state workload density of scenario i in terms of the classical M/G/1 queue (see also [7, 23]):

$$V^{K,i}(x) = \frac{V(x)}{1 - \rho + \rho V(K)}.$$

Finally, the first-exit probabilities follow from a direct computation, see [19]. Also, Takács' formula for cycle maxima [30] may be easily recovered from Theorem 5.1 and the truncation property for finite dams (Theorem 4.1).

7.2 Exponential service requirements

Suppose that $1 - B(x) = e^{-\mu x}$, meaning that the service requirements are exponentially distributed with mean $1/\mu$. For the basic kernel, we then may write $K(x, y) = e^{-\mu(x-y)}\lambda(x)/r(x)$, and we can explicitly compute (similar to [19])

$$K^*(x, y) = \frac{\lambda(x)}{r(x)} \exp\{-\mu(x - y) + \Lambda(x) - \Lambda(y)\}. \tag{35}$$

Using Lemma 3.1, the familiar steady-state workload density in the infinite-buffer queue directly appears (see e.g. [4, 8, 19], or [3], p. 388):

$$v(x) = \frac{\lambda(0)V(0)}{r(x)} \exp\{-\mu x + \Lambda(x)\}. \tag{36}$$

The explicit form in (35) also allows us to evaluate (30). After lengthy calculations, we deduce the following:

Corollary 7.1. *For the M/M/1 queue with customer impatience (scenario i), arrival rate $\lambda(\cdot)$, and service rate $r(\cdot)$, we have*

$$v^{K,i}(x) = \frac{\lambda(0)V^{K,i}(0)}{r(x)} \exp\{-\mu x + \Lambda(x \wedge K)\},$$

where $V^{K,i}(0)$ follows by normalization.

Turning to scenario c (complete rejections), we obtain the following corollary:

Corollary 7.2. *For the M/M/1 queue with complete rejections (scenario c), arrival rate $\lambda(\cdot)$, and service rate $r(\cdot)$, we have*

$$v^{K,c}(x) = \frac{V^{K,c}(0)\lambda(0)(1 - e^{-\mu(K-x)})}{r(x)} \exp\{-\mu x + \Lambda^c(x)\},$$

where $\Lambda^c(x) = \int_0^x \frac{\lambda(y)(1-e^{-\mu(K-y)})}{r(y)}\mathrm{d}y$ and $V^{K,c}(0)$ follows by normalization.

Proof. Note that, by conditioning on $B > x - y$, $\mathbb{P}(g(y, B, K) > x)$ may be rewritten as $e^{-\mu(x-y)}(1 - e^{-\mu(K-x)})$. Thus, by substitution in (5), (or (32)), we have

$$r(x)v^g(x) = \lambda(0)V^g(0)e^{-\mu x}(1 - e^{-\mu(K-x)}) + \int_{0^+}^{x} \lambda(y)v^g(y)e^{-\mu(x-y)}(1 - e^{-\mu(K-x)})\mathrm{d}y.$$

Multiply both sides by $(1 - e^{-\mu(K-x)})^{-1}$. Then, comparing with (16), it follows that scenario c is equivalent to scenario f, but with $r(x)$ replaced by $r^c(x) := r(x)(1-e^{-\mu(K-x)})^{-1}$. Appropriately adjusting $\Lambda(\cdot)$, resulting in $\Lambda^c(\cdot)$, and applying (36) gives the result. □

The result for the classical M/M/1-queue with complete rejections [18] can now easily be recovered from our corollary.
The first-exit probabilities may be deduced from Lemma 5.1. Alternatively, the first-exit probabilities may also be obtained from the steady-state workload density in the finite dam with arrival and release rate $\lambda(b - x)$ and $r(b - x)$ when the workload equals x. The cycle maximum can be derived in the same way.

Corollary 7.3. *For the cycle maximum in an M/M/1 queue, with arrival rate $\lambda(\cdot)$ and service rate $r(\cdot)$, we have*

$$\mathbb{P}(C_{\max} > x) = \tilde{V}(0) \exp\{\Lambda(x) - \mu x\},$$

where $\tilde{V}(0) = \left[1 + \int_0^b \frac{\lambda(x)}{r(x)} \exp\{-\mu(b-x) + \Lambda(b) - \Lambda(x)\} \mathrm{d}x\right]^{-1}$.

Proof. Combining (22) with (35) and some rewriting yields

$$\tilde{v}(x) = \tilde{V}(0) \frac{\lambda(x)}{r(x)} \exp\{-\mu x + \Lambda(b) - \Lambda(b-x)\}.$$

$\tilde{V}(0)$ now follows directly by normalization. Moreover, use (21),

$$\begin{aligned}
\mathbb{P}(C_{\max} > b) &= \tilde{V}(0) e^{-\mu b} + \int_0^b \tilde{V}(0) e^{-\mu b} \frac{\lambda(b-x)}{r(b-x)} e^{\Lambda(b) - \Lambda(b-x)} \mathrm{d}x \\
&= \tilde{V}(0) e^{-\mu b} \left[1 + e^{\Lambda(b)} \int_0^b \frac{\lambda(x)}{r(x)} e^{-\Lambda(x)} \mathrm{d}x\right] \\
&= \tilde{V}(0) e^{-\mu b} \left[1 + e^{\Lambda(b)} \left(1 - e^{-\Lambda(b)}\right)\right] \\
&= \tilde{V}(0) e^{-\mu b + \Lambda(b)},
\end{aligned}$$

completing the proof. □

7.3 Stochastic barricade

In this paper we considered an M/G/1-type model with restricted accessibility, in which the rejection rule is based on a deterministic barricade. This may be extended by replacing K by a stochastic variable, see for instance [13, 28]. This extension can easily be included into our framework. Replace at the n-th arrival epoch K by the random variable U_n, with distribution $F_U(\cdot)$ (independent of the service and arrival processes). The acceptance of the n-th customer in the scenarios of Section 2 is now determined as follows, see also [28]:

$$g(W_n, B_n, U_n) = \begin{cases} W_n + \min(W_n + B_n, (U_n - W_n)^+), & \text{scenario } f, \\ W_n + B_n I(W_n < U_n), & \text{scenario } i, \\ W_n + B_n I(W_n + B_n \leq U_n), & \text{scenario } c. \end{cases}$$

Note that in case $\lambda(\cdot)$ and $r(\cdot)$ are fixed, U_n represents the maximal waiting time (scenario i), or sojourn time (scenarios f and c). This model with stochastic impatience is well-known and studied in, e.g., [13, 28].
Also in case of a random barricade, we again obtain a Volterra integral equation of the second kind. For the given examples, we have the following Volterra kernels, where $0 \leq y < x < \infty$ (see [28]),

$$K^g(x,y) = \begin{cases} (1 - B(x-y))(1 - F_U(x))\lambda(x)/r(x), & \text{scenario } f, \\ (1 - B(x-y))(1 - F_U(y))\lambda(x)/r(x), & \text{scenario } i, \\ \frac{\lambda(x)}{r(x)} \int_x^\infty (B(z-y) - B(x-y)) \mathrm{d}F_U(z), & \text{scenario } c. \end{cases}$$

Even though these kernels might be difficult to determine in general, we may express the steady-state workload density in terms of these kernels, see Lemma 3.1. Some examples

can be found in [28] in case of exponential service requirements and either exponential or deterministic barricades.
Finally, consider two (general) finite-buffer queues governed by the same distributions $B(\cdot)$ and $F_U(\cdot)$, but with arrival and service rates $\lambda_i(\cdot)$ and $r_i(\cdot)$, $i = 1, 2$, such that $\frac{\lambda_1(x)}{r_1(x)} = \frac{\lambda_2(x)}{r_2(x)}$, for all $x > 0$. The queues are then related in the same way as the queues in Section 3. From the discussion of Volterra kernels above, it is evident that (4) still holds in this broader context. Note that the dynamics in both systems are still equal, resulting in the generalization of (3). More general, we state that Theorems 3.1 and 3.2 also hold in this framework.

Acknowledgement

The author is grateful to Onno Boxma and Bert Zwart for comments on an earlier draft of this paper and some useful remarks.

References

[1] Asmussen, S., K. Sigman (1996). Monotone stochastic recursions and their duals. *Probability in the Engineering and Informational Sciences* **10**, 1–20.

[2] Asmussen, S. (1998). Extreme value theory for queues via cycle maxima. *Extremes* **2**, 137–168.

[3] Asmussen, S. (2003). *Applied Probability and Queues*, Second Edition. Springer, New York.

[4] Bekker, R., S.C. Borst, O.J. Boxma, O. Kella (2003). Queues with workload-dependent arrival and service rates. To appear in *Queueing Systems*.

[5] Bekker, R., A.P. Zwart (2003). On an equivalence between loss rates and cycle maxima in queues and dams. SPOR-Report 2003-17, Eindhoven University of Technology.

[6] Bertrand, J.W.M., H.P.G. van Ooijen (2002). Workload based order release and productivity: a missing link. *Production Planning & Control* **13**, 665–678.

[7] Boots, N.K., H.C. Tijms (1999). A multiserver queueing system with impatient customers. *Management Science* **45**, 444-448.

[8] Brockwell, P.J., S.I. Resnick, R.L. Tweedie (1982). Storage processes with general release rule and additive inputs. *Advances in Applied Probability* **14**, 392–433.

[9] Browne, S., K. Sigman (1992). Work-modulated queues with applications to storage processes. *Journal of Applied Probability* **29**, 699–712.

[10] Cohen, J.W. (1969). Single-server queues with restricted accessibility. *Journal of Engineering Mathematics* **3**, 265–284.

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

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

[13] Daley, D.J. (1965). General customer impatience in the queue $GI/G/1$. *Journal of Applied Probability* **2**, 186–205.

[14] 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.

[15] 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.

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

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

[18] Gavish, B., P. Schweitzer (1977). The markovian queue with bounded waiting time. *Management Science* **23**, 1349–1357.

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

[20] Hooghiemstra, G. (1987). A path construction for the virtual waiting time of an M/G/1 queue. *Statistica Neerlandica* **41**, 175–181.

[21] Jagerman, D. (1985). Certain Volterra integral equations arising in queueing. *Stochastic Models* **1**, 239–256.

[22] 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.

[23] Kok, A.G. de, H.C. Tijms (1985). A queueing system with impatient customers. *Journal of Applied Probability* **22**, 688–696.

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

[25] Loris-Teghem, J. (1972). On the waiting time distribution in a generalized queueing system with uniformly bounded sojourn times. *Journal of Applied Probability* **9**, 642–649.

[26] 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.

[27] Van Ooijen, H.P.G., J.W.M. Bertrand (2003). The effects of a simple arrival rate control policy on throughput and work-in-process in production systems with workload dependent processing rates. *International Journal of Production Economics* **85**, 61–68.

[28] Perry, D., S. Asmussen (1995). Rejection rules in the M/G/1 queue. *Queueing Systems* **19**, 105–130.

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

[30] Takács, L. (1965). Application of Ballot theorems in the theory of queues. In: Smith, W.L., W.E. Wilkinson (editors). *Proceedings of the Symposium on Congestion Theory*, 337–398. University of North Carolina Press, Chapel Hill.

[31] Zwart, A.P. (2000). A fluid queue with a finite buffer and subexponential input. *Advances in Applied Probability* **32**, 221–243.

[32] Zwart, A.P. (2003). Loss rates in the M/G/1 queue with complete rejection. SPOR-Report 2003-27, Eindhoven University of Technology.