ON THE CYCLE MAXIMUM OF MOUNTAINS, DAMS AND QUEUES

Onno J. Boxma(*) and David Perry(**)

(*) EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology; boxma@win.tue.nl

(**) Department of Statistics, University of Haifa; dperry@stat.haifa.ac.il

ABSTRACT

We determine the distribution of the maximum level of the workload in some queueing, dam and storage processes. The models under consideration are the following. (i) The Markov mountain: a storage or dam model that alternates between exponentially distributed ON and OFF periods. The buffer content increases (decreases) at some state-dependent rate when ON (OFF). (ii) The semi-Markov mountain: as (i), but with generally distributed ON periods. (iii) The $M/G/1$ queue with various forms of customer impatience.

In: Communications in Statistics - part A. Theory and Methods 38 (2009) 2706-2720.
Special issue in honour of S. Zacks.

1 INTRODUCTION

This paper is dedicated to Shelley Zacks, in friendship and admiration. Both authors have the pleasure of collaborating with Shelley on queueing-theoretic problems closely related to the present paper. For him this topic is just one out of many which have received his attention. Accordingly, the range of papers and the readership of this special issue will be broad. Hence we shall start leisurely, by introducing some classical concepts from queueing theory.

The classical single server queue (cf. Cohen (1982)) is a model of a service facility with a single server, at which customers arrive with a service request. It is usually assumed that the server works at unit speed, serving customers according to some service discipline like First-Come-First-Served (FCFS). It is also typically assumed that successive interarrival times of customers are independent, identically distributed (i.i.d.) random variables, and that the service times of successive customers are also i.i.d., and independent of the arrival process. In the case of general interarrival and service time distributions, this model is termed the $G/G/1$ queue; the $M/G/1$ queue has exponentially distributed interarrival times, and the $G/M/1$ queue has exponentially distributed service times (M for Memoryless or Markovian).

The most important performance measures under consideration in queueing theory are the queue length distribution (a system-oriented performance measure) and the waiting time distribution (a customer-oriented performance measure). Other important performance measures are the distribution of the busy period, i.e., the time during which the server is uninterruptedly working, and the distribution of the cycle maximum, i.e., the largest amount of work present in the system during such a busy period. The cycle maximum is an interesting quantity because, in practice, buffer sizes are finite and the cycle maximum gives an indication for the probability that that maximum buffer size is needed or exceeded. For example, if M denotes the maximal workload in a busy period of the $M/G/1$ queue with unlimited capacity, then $\Pr(M < K)$ denotes the probability that no overflow occurs during a busy period of the $M/G/1$ queue with finite capacity K.

In the sequel we shall not only consider queues, but also dams and mountains. While

the workload in a queueing model increases instantaneously when a customer arrives, the workload (content) in the *mountain process* as defined in Boxma et al. (2005) both decreases *and* increases gradually: during OFF (ON) periods it decreases (increases) with some general rate which depends on the current level. See Figure 1 in Section 2 for an example. If one deletes the ON periods from the mountain process, one gets a so-called *dam process*, but with jumps upward which are not i.i.d. (because of the above-mentioned level-dependent increment rate). If one deletes the OFF periods from the mountain process, one gets a general risk (or inventory/production) process, but with jumps downward which are not i.i.d.

This paper is devoted to the study of cycle maxima of some queueing, dam and mountain processes. In all these processes, the cycle maximum is the maximal workload in the system during a busy period. It is not only a key performance measure for queues but also for dams and storage processes, due to its importance in extreme value theory (cf. Asmussen (1998) or Asmussen (2003), pp. 298-301, pp. 368; here special attention is given to the probability that the cycle maximum exceeds K, with K tending to infinity) and due to its importance for the study of systems with a finite capacity (cf. Cohen (1968)).

The cycle maximum in the $M/G/1$ queue

For future reference we mention two expressions for the distribution of M in the $M/G/1$ queue with arrival rate λ, generic service time G with distribution $G(\cdot)$, and with steady-state workload distribution $V(\cdot)$ with density $v(\cdot)$; the latter steady-state workload distribution exists iff the offered load $\rho := \lambda E[G] < 1$. These expressions were independently discovered by Takács (1967), Section 29, and Cohen (1968). For $x > 0$:

$$\Pr(M < x) = 1 - \frac{1}{\lambda}\frac{d}{dx}\ln V(x) = 1 - \frac{1}{\lambda}\frac{v(x)}{V(x)}; \tag{1}$$

$$\Pr(M < x) = \frac{V * G(x)}{V(x)}. \tag{2}$$

Here $*$ denotes a convolution. Because of PASTA, the workload distribution equals the waiting time distribution. Hence one can write, with W and Q denoting a generic waiting

and sojourn time:

$$\Pr(M < x) = \frac{\Pr(Q < x)}{\Pr(W < x)}. \tag{3}$$

Remark 1 *It readily follows from (1) that (cf. Cohen (1982), pp. 618):*

$$E[M] = \frac{1}{\lambda} \ln \frac{1}{1-\rho}. \tag{4}$$

This is a remarkably simple expression; notice that only the mean *service time plays a role here.*

if $\rho = 1$ then the busy period is finite with probability one, but its mean is infinite. If $\rho > 1$ then there is a positive probability that a busy period is not finite. Hence one then cannot speak of the cycle maximum of an arbitrary busy period. We refer to Cohen (1976), Section 3.3, for the conditional cycle maximum *given* the busy period is finite, in the case that $\rho > 1$. See Cohen (1968) for the $G/M/1$ case with $\rho < 1$, and Adan et al. (2005) for the $G/M/1$ case with $\rho > 1$. In Asmussen and Perry (1992), the distribution of M is derived for a single server queue with a general Markovian arrival process. See Asmussen (1998) for a survey on cycle maxima.

A recent paper of Albrecher et al. (2009) presents a new proof of (1), as well as providing additional motivation for studying cycle maxima. That paper is devoted to (a generalization of) the classical Cramér-Lundberg risk model. Here

$$R(t) = x + ct - \sum_{i=1}^{N(t)} G_i \tag{5}$$

describes the surplus at time t of an insurance portfolio, where x is the initial capital, c is a constant premium intensity (we choose $c = 1$) and $\{G_i\}_{i \geq 1}$ is a sequence of independent and identically distributed positive random variables which denote claims. The claim number process $N(t)$ is assumed to be a homogeneous Poisson process with intensity λ. A crucial quantity in risk theory is the infinite-time survival probability

$$\phi(x) = \Pr(R(t) \geq 0 \; \forall \; t \geq 0 \,|\, R(0) = x).$$

The following lemma and its proof, from Albrecher et al. (2009), provide a direct link between the infinite-time survival probability of the Cramér-Lundberg insurance risk process $R(t)$ and the maximum workload M of a busy period in the corresponding $M/G/1$ queue with arrival rate λ and with generic service time G:

Lemma 1 *Under the net profit condition $\lambda\, E(G) < 1$ we have for every $x \geq 0$:*

$$\Pr(M < x) = 1 - \frac{1}{\lambda}\,\frac{d}{dx}\,\log\phi(x). \tag{6}$$

Proof. The risk process $R(t)$ starting in x can only survive if after the first claim, occurring at some surplus level y, the level y will be reached again before ruin occurs. This is equivalent to: the maximum workload M of the corresponding $M/G/1$ queue does not exceed y. Since we are only concerned about eventual survival, the part until the process returns to y (which is a busy period of the corresponding $M/G/1$ queue, turned upside down) can be cut out. So the survival probability $\phi(x)$ simply is the probability of *zero* events during $[x, \infty)$ of an inhomogeneous Poisson process with time-dependent rate $\lambda(t) = \lambda \Pr(M > t)$, which implies

$$\phi(x) = \exp\left(-\int_x^\infty \lambda(t)\,dt\right) = \exp\left(-\lambda\int_x^\infty \Pr(M > t)\,dt\right), \tag{7}$$

yielding (6). It is well known (see pp. 30-32 of Asmussen (2000)) that the survival probability $\phi(x)$ in the Cramér-Lundberg risk model equals the steady-state workload distribution $V(x)$ in the corresponding $M/G/1$ queue. (1) now immediately follows from (6).

The models to be considered in this paper

In Boxma et al. (1999) a fluid queue or dam was studied with a buffer content that varies linearly during periods that are generated by a three-state semi-Markov process. Two cases were distinguished for this "mountain process": (i) two upward slopes and one downward slope, and (ii) one upward slope and two downward slopes. In both cases, the length of at least one of the three periods has a general distribution. The cycle maximum or top of the mountain, i.e., the maximal buffer content during a busy period, was one of the quantities which were studied in Boxma et al. (1999) for these two cases.

In the present paper, we consider the top of the mountain for another mountain process: a storage or dam model that alternates between ON and OFF periods. When ON, the buffer content increases *at some state-dependent rate* $\alpha(x)$. When OFF, the buffer content decreases *at some state-dependent rate* $\beta(x)$, $x > 0$; when it reaches zero, it stays at zero until another ON period begins. The steady-state buffer content of this process, as well as conditions for its existence and uniqueness, were studied in Boxma et al. (2005). In the present paper, we focus on the top of the mountain, for the special case that the ON and OFF periods are all independent and exponentially distributed. We call this model the *Markov mountain*. We obtain very explicit results for the hazard rate of the cycle maximum (top of the mountain), and for its distribution. We subsequently extend this model to the case in which the ON periods have a general distribution, the *semi-Markov mountain*. The distribution of the cycle maximum is expressed in the workload distribution in two different ways. The results are less explicit than for the Markov mountain, in the sense that the workload distribution is only determined as the solution of some Volterra integral equation.

We also consider the cycle maximum for $M/G/1$-type queues with restricted accessibility. We distinguish three cases: (i) customers are not admitted (become impatient) when their waiting time exceeds a certain patience time (*Model 1*; the patience time is the time a customer is willing to wait before becoming impatient and leaving the system); (ii) customers are partially admitted when their waiting time is less than a certain patience time but their waiting plus service time exceeds that patience time (*Model 2*); (iii) customers are not admitted when their waiting plus service time exceeds a certain patience time. In each case it is assumed that the waiting and service time are observable at the moment of arrival.

The rest of the paper is organized as follows. The Markov mountain is studied in Section 2, and the semi-Markov mountain in Section 3. Section 4 is devoted to the study of the cycle maximum for the three $M/G/1$-type queues with restricted accessibility.

2 THE MARKOV MOUNTAIN

Consider a special case of the mountain process as introduced in Boxma et al. (2005). That is, we assume a mountain process with intermittent ON and OFF periods. This intermittence of ON and OFF periods generates an alternating renewal process such that within each ON-OFF cycle the ON period and the OFF period are independent and exponentially(μ), respectively exponentially(λ), distributed. The mountain process $\mathbf{X} = \{X(t) : t \geq 0\}$ is a regenerative nonnegative process whose cycle starts at an ON period with $X(0) = 0$. The *hilly period* of the mountains (uninterrupted period with a positive level) is equivalent to the *wet period* of dams, to the *production period* for inventories and to the *busy period* in queues. Similarly, the *sea level period* of the mountains is equivalent to the *dry period* of dams, to the *unsatisfied demand period* for inventories and to the *idle period* in queues.

Formally, we define the end of a hilly period by $\tau = \inf\{t > 0 : X(t) = 0\}$ and the end of a sea level period I by $\tau + I$, where $I = \inf\{t > 0 : X(\tau + t) > 0\}$. Then, $T = \tau + I$ is the length of a cycle. In this study we consider a special case of Boxma et al. (2005) in the sense that only the increase rate $\alpha(x)$ during ON periods and the decrease rate $\beta(x)$ during OFF periods are state dependent (in Boxma et al. (2005) the lengths of the ON and OFF periods are also state dependent). A typical realization of the mountain process is depicted in Figure 1.

We focus on the analysis of $M = max\{X(t) : 0 < t < T\}$ which is the *top of the mountain* during one cycle. To this end we construct the dam process $\mathbf{D} = \{D(t) : t \geq 0\}$ by deleting the ON periods from $\mathbf{X}$ and gluing together the OFF periods. Note that $\mathbf{D}$ is a special Markov dam with Poisson arrivals at rate λ. However, the jumps are state dependent; they are neither independent nor identically distributed. In any case, by construction, M is the cycle maximum of both $\mathbf{X}$ and $\mathbf{D}$. We therefore apply the stochastic analysis of M associated with the dam $\mathbf{D}$ and the fact that the ON periods of $\mathbf{X}$ are $exp(\mu)$ leads to the following obvious

Criterion 1 *The hazard rate function of the jump size at any level x in $\mathbf{D}$ is $\mu/\alpha(x)$.*

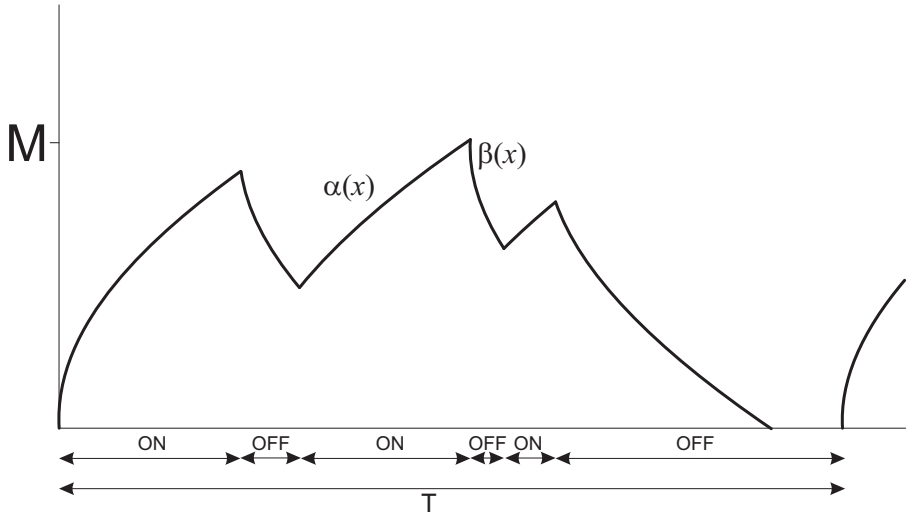

Figure 1: The Markov mountain

It is clear that the cycle maximum of **D** occurs at a moment of a jump and it must be a record value. Furthermore, in order that M is the cycle maximum of **D** it must be the last record value during the *cycle* of **D** and thus also during the *wet period* of **D**. In other words, level x is the cycle maximum of **D** if and only if it is a record value such that after reaching level x the dam **D** will reach level 0 before upcrossing level x again. Let $\theta(x)$ be the probability of the latter event. In the proof of the next theorem we use an argument similar to that used in the proof of Theorem 5 of Adan et al. (2005).

Theorem 1 *Let $r_M(x)$ be the hazard rate function of M at x. Then*

$$r_M(x) = \frac{\mu}{\alpha(x)}\theta(x).$$

Proof. Since the ON times in the mountain **X** are $exp(\mu)$ distributed, $\mu dx/\alpha(x)$ is the infinitesimal probability that an arbitrary record value of **D** lands in $[x, x+dx)$. But $M \in [x, x+dx)$ if and only if the latter record value is the last record value in the busy period and the probability of the latter event is $\theta(x)$. By the strong Markov property, we find $r_M(x)dx$ by taking the product of $\mu dx/\alpha(x)$ and $\theta(x)$. ■

To compute $\theta(x)$ we use an argument similar to that used in Adan et al. (2005) and Asmussen and Perry (1992). In the following calculations we omit $o(dx)$ terms. First note

that, due to the fact that $\mathbf{D}$ is a Markov process, we have for all $x > 0$ the equation

$$\theta(x+dx) = \left[1 - \frac{\lambda dx}{\beta(x)}\right] \left[\theta(x) + (1-\theta(x)) \frac{\mu dx}{\alpha(x)} \theta(x)\right]. \tag{8}$$

To better understand the right hand side of (8) note that the paths with arrivals in $[0, \frac{dx}{\beta(x)})$ do not provide a contribution to the content level of $\mathbf{D}$ since they have probability λdx and will upcross level $x + dx$ unless the further event of service termination in $[0, \frac{dx}{\beta(x)})$ (having probability $\frac{\mu dx}{\alpha(x)}$) occurs. The term $\theta(x)$ in (8) then corresponds to paths which downcross level x and do not upcross again. The term $(1-\theta(x)) \frac{\mu dx}{\alpha(x)} \theta(x)$ corresponds to paths which downcross level x and upcross again before hitting level 0 with jump terminating at $u \in (x, x+dx]$, where the value of u does not matter since $\theta(u) = \theta(x) + O(dx)$. Hence, from (8) we get

$$\theta'(x) = -\frac{\lambda \theta(x)}{\beta(x)} + \frac{\mu \theta(x)}{\alpha(x)} - \frac{\mu \theta^2(x)}{\alpha(x)}. \tag{9}$$

To simplify (9) substitute

$$\eta(x) = 1/\theta(x).$$

We get after some elementary algebra

$$\eta'(x) + \eta(x) \left(\frac{\mu}{\alpha(x)} - \frac{\lambda}{\beta(x)}\right) = \frac{\mu}{\alpha(x)}. \tag{10}$$

Let

$$A(x) = \int_0^x [1/\alpha(y)] dy, \quad \text{and} \quad B(x) = \int_0^x [1/\beta(y)] dy. \tag{11}$$

By multiplying both sides of (10) by $e^{\mu A(x) - \lambda B(x)}$ we get

$$e^{\mu A(x) - \lambda B(x)} \left[\eta'(x) + \eta(x) \left(\frac{\mu}{\alpha(x)} - \frac{\lambda}{\beta(x)}\right)\right] = e^{\mu A(x) - \lambda B(x)} \frac{\mu}{\alpha(x)}. \tag{12}$$

Solving for $\eta(x)$ in (12) we get

$$\eta(x) = H(x) e^{-\mu A(x) + \lambda B(x)} + k_0 e^{-\mu A(x) + \lambda B(x)},$$

where $H(x) = \int_0^x e^{\mu A(y) - \lambda B(y)} \frac{\mu}{\alpha(y)} dy$ and k_0 is a constant. Obviously, $\eta(0) = 1$ so that $k_0 = 1$. We thus get

$$\eta(x) = [H(x) + 1] e^{-\mu A(x) + \lambda B(x)}. \tag{13}$$

Substituting (13) in Theorem 1 we obtain

$$r_M(x) = \frac{\mu e^{\mu A(x) - \lambda B(x)}}{\alpha(x)[H(x)+1]}.$$

From Theorem 1 we conclude, using the familiar relation $\Pr(M > x) = e^{-\int_0^x r_M(y)dy}$ between a distribution and its hazard rate function:

Theorem 2

$$\Pr(M > x) = e^{-\int_0^x \frac{\mu e^{\mu A(y) - \lambda B(y)}}{\alpha(y)[H(y)+1]} dy}. \tag{14}$$

Theorems 1 and 2 show (i) that the distribution of the cycle maximum is directly related to the probability $\theta(x)$ of reaching level 0 starting from x, before reaching level x again, and (ii) how this distribution can be expressed in the rate $\alpha(x)$ and the integrals $A(x) = \int_0^x [1/\alpha(y)]dy$ and $B(x) = \int_0^x [1/\beta(y)]dy$.

3 THE SEMI-MARKOV MOUNTAIN

In this section we extend the model introduced in Section 2 to the case of a semi-Markov mountain. We assume that the ON period is not necessarily exponentially distributed. The trade-off that stems from this generalization is clear due to the type of the final results in the generalized model. That is, while the distribution of M is given explicitly in (14) in the Markov case, for the semi-Markov case we only manage to express $P(M > x)$ in terms of $\theta(x)$ and $f(x)$, the steady-state density of the dam associated with the mountain. The latter density is the solution of a Volterra-type integral equation. We have not obtained an expression for $\theta(x)$ in the general case. In the special case $\alpha(x) \equiv 1$, $\theta(x)$ and hence $P(M > x)$ are known; cf. Remark 2 below.

In our general model we assume that the distribution of the ON period, $G(\cdot)$, is absolutely continuous with density $g(\cdot)$, while the OFF period is still $exp(\lambda)$. We also assume that

$$\int_0^x \frac{1}{\beta(y)} dy < \infty$$

so that level 0 can be reached with probability 1 from any level x and without getting into details we assume that the necessary and also sufficient conditions for stability hold so that the mountain is a regenerative process.

For this general case we apply the same approach as in Section 2 and construct the Dam **D** from the Mountain **X** by deleting the ON periods and gluing the OFF periods together. The release function of the resulting dam is $\beta(\cdot)$ but the jumps are neither independent nor identically distributed. In fact, the conditional probability that the process after the jump stays below x given it starts at level w is $G(A(x) - A(w))$ where $A(\cdot)$ is defined in (11).

Let $F(\cdot)$ be the steady-state distribution of the content level of **D**. By Level Crossing Theory (LCT), $F(\cdot)$ is an absolutely continuous distribution with density $f(\cdot)$ and the balance equation of Pollaczek-Khintchine type is given by

$$\begin{aligned} \beta(x)f(x) &= \lambda \textstyle\int_0^x [1 - G(A(x) - A(w))]dF(w) \\ &= \lambda \textstyle\int_0^x [1 - G(A(x) - A(w))]f(w)dw + \lambda\pi[1 - G(A(x))], \end{aligned} \tag{15}$$

where

$$\pi = \lim_{t\to\infty} \Pr(D(t) = 0).$$

For future reference we shall rewrite the first equality in (15) as

$$\beta(x)f(x) = \lambda[F(x) - (F \circledast G)(x)]. \tag{16}$$

The balance equation (15) is a Volterra-type equation which is known to be uniquely solvable by a Neumann series in the space of continuous functions (for example, see Harrison and Resnick (1976)). To solve for $f(\cdot)$ we use the notation

$$f(x) = \int_0^x f(w)Q(w, x)dw + \pi Q(0, x),$$

where

$$Q(w, x) := Q^{*1}(w, x) := \frac{\lambda[1 - G(A(x) - A(w))]}{\beta(x)}.$$

Define

$$Q^{*(n+1)}(0, x) := \int_0^x Q^{*n}(0, y)Q(y, x)dy.$$

The Neumann series yields:

$$f(x) = \pi \sum_{n=1}^{\infty} Q^{*n}(0, x), \tag{17}$$

where π can be calculated from the normalizing condition

$$\int_0^{\infty} f(x)dx = 1 - \pi.$$

We thus obtain

$$f(x) = \frac{\sum_{n=1}^{\infty} Q^{*n}(0, x)}{1 + \int_0^{\infty} \sum_{n=1}^{\infty} Q^{*n}(0, y)dy}. \tag{18}$$

We are now in position to introduce a fundamental *proportionality result based on Level Crossing Theory*. To this end we define $D_x(T)$ as the number of downcrossings of level x during the cycle T, where T is defined just like in Section 2.

By conditioning on whether the event $\{M \leq x\}$ occurred or not we get by the law of total probability:

$$ED_x(T) = E[D_x(T) \mid M \leq x] \Pr(M \leq x) + E[D_x(T) \mid M > x] \Pr(M > x).$$

Clearly, $E[D_x(T) \mid M \leq x] = 0$ and by the strong Markov property the conditional number of downcrossings of level x given the event $\{M > x\}$ occurred is geometrically $(1 - \theta(x))$ distributed so that

$$E[D_x(T) \mid M > x] = \frac{1}{\theta(x)},$$

and hence

$$\Pr(M > x) = \theta(x) ED_x(T). \tag{19}$$

Let D_x be the long-run average number of downcrossings of level $x > 0$ by $\mathbf{D}$. On the one hand we get by the renewal reward theorem:

$$D_x = \frac{ED_x(T)}{ET}. \tag{20}$$

On the other hand, by LCT,

$$D_x = \beta(x) f(x). \tag{21}$$

Combining (16), (19), (20) and (21), and using that $\Pr(M > 0) = 1$, we get the following theorem, in which $P(M > x)$ is expressed in terms of $\theta(x)$ and the steady-state density $f(x)$ of the workload of the dam associated with the mountain:

Theorem 3

$$\begin{aligned} \Pr(M > x) &= \theta(x)\frac{\beta(x)f(x)}{\beta(0)f(0)} \\ &= \frac{\lambda}{\beta(0)f(0)}\theta(x)[F(x) - (F \circledast G)(x)]. \end{aligned} \tag{22}$$

Remark 2 *In general it is difficult to determine $\theta(x)$. $\theta(x)$ is known in the special case that $\alpha(x) \equiv 1$, so $A(x) = x$; cf. Formula (14) of Lee (2007):*

$$\theta(x) = \frac{1}{1 + \int_0^x \sum_{n=1}^{\infty} Q^{*n}(0, y)dy}. \tag{23}$$

Comparison of this formula and (17) reveals that, in this case,

$$\theta(x) = \frac{\pi}{\pi + \int_0^x f(y)dy} = \frac{F(0)}{F(x)}. \tag{24}$$

Hence in this case of the $M/G/1$ queue with release rate $\beta(x)$, we retrieve a formula that was already obtained by Bekker and Zwart (2005) (notice that $F \circledast G$ reduces to $F * G$ when $A(x) = x$):

$$\Pr(M > x) = \frac{\lambda F(0)}{\beta(0)f(0)}\frac{F(x) - (F * G)(x)}{F(x)}. \tag{25}$$

Since $\lambda F(0) = \beta(0)f(0)$, cf. (16), we have

$$\Pr(M < x) = \frac{(F * G)(x)}{F(x)}. \tag{26}$$

(26) should be compared with (2).

Furthermore, combination of the first line of (22) and (17) shows that

$$\Pr(M > x) = \frac{\beta(x)f(x)F(0)}{\beta(0)f(0)F(x)} = \frac{\beta(x)}{\lambda}\frac{f(x)}{F(x)} = \frac{\beta(x)}{\lambda}\frac{d}{dx}\ln F(x). \tag{27}$$

(27) should be compared with (1).

4 QUEUES WITH RESTRICTED ACCESSIBILITY

We focus on the three prototype models as described in earlier works (see for example Asmussen and Perry (1992), Boxma et al. (2009), Cohen (1969) and Perry et al. (2000)). The above models are those of single server queues of the $M/G/1$ type and the restricted accessibility is based on either the waiting times or the sojourn times, but not based on the number of waiting customers. Also, it is assumed that the service requirements, as well as the patience of the customers, are observable. That means that customers *know* the waiting times, the sojourn times and the patience at their arrivals so that they do not waste time in line. In other words, if the waiting time (or the sojourn time) is greater than the patience, the customer is not admitted to the system.

For all three models we designate the waiting time of the nth arriving customer by W_n, the sojourn time by Q_n, the service requirement by S_n and the patience by Y_n. We assume that $S_1, S_2, ...$ and $Y_1, Y_2, ...$ respectively, are sequences of i.i.d. random variables, which are also independent of each other.

In Model 1 the accessibility is applied only to the waiting time but the service requirements are also the actual services.

Formally, we have for **Model 1**:

$$Q_n = (W_n + S_n) \cdot \mathbf{1}_{\{W_n \leq Y_n\}}; \tag{28}$$

if W_n would exceed Y_n, then the nth arriving customer will not enter the system.
In Model 2 and Model 3 the accessibility is applied to the sojourn times. In Model 2 the service might be truncated if the sojourn time is greater than the patience.

Formally, we have for **Model 2**:

$$Q_n = (W_n + S_n) \cdot \mathbf{1}_{\{W_n + S_n \leq Y_n\}} + Y_n \cdot \mathbf{1}_{\{W_n \leq Y_n \leq W_n + S_n\}}. \tag{29}$$

In Model 3 a customer is admitted to the system only if the service requirement is also the actual service.

Formally, we have for **Model 3**:

$$Q_n = (W_n + S_n) \cdot \mathbf{1}_{\{W_n+S_n \le Y_n\}}. \tag{30}$$

In Model 1 below we express the law of M in terms of LST (see Theorem 4). Then for Models 2 and 3 we apply Theorem 1 and compute separately the values of $r_M(x)$ and $\theta(x)$.

4.1 Model 1

In the language of queueing theory and in case that the interarrival times, the service requirements and the patience are general, this model is known as the $G/G/1+G$ (see Baccelli et al. (1984)) queueing system. In this study we restrict the attention to the $M/M/1+M$ special case: We assume Poisson arrivals with rate λ, $exp(\mu)$ service requirements and independent $exp(\xi)$ patience. Let $\mathbf{V} = \{V(t) : t \ge 0\}$ be the work process of the above queue and as before let τ be the busy period. Then $M = \max_{0\le t\le\tau} V(t)$ is the cycle maximum. Let X_j be the jth record value in the cycle for $j = 1, 2, ..., N$, where N is the last record value in the cycle. Then, $M = X_N$. Let L_j be the jth record time (with $L_1 = 0$ and $X_0 = 0$). In a similar notation to that used in Section 2, we define

$$U^-_{j+1} = \inf\{t > 0 : V(L_j + t) = 0\}, \quad U^+_{j+1} = \inf\{t > 0 : V(L_j + t) = X_{j+1}\},$$

and let

$$\gamma(x) = \Pr(U^-_j < U^+_j \mid X_{j-1} = x), \tag{31}$$

where $x = 0$ whenever $j = 1$ (note that the conditional probability (31) is independent of j since by the strong Markov property $\Pr(U^-_j < U^+_j \mid X_{j-1} = x) = \Pr(U^-_j - L_{j-1} < U^+_j - L_{j-1} \mid X_{j-1} = x)$). Define

$$\gamma_j = \Pr(U^-_j < U^+_j),$$

and observe that, because of the memoryless property of the jumps, X_j has an Erlang(j, μ) distribution for $j = 1, 2, \ldots$ (see Figure 2). Hence

$$\gamma_j = \int_0^\infty \frac{e^{-\mu x}(\mu x)^{j-1}\mu}{(j-1)!}\gamma(x)dx. \tag{32}$$

To compute $\gamma(x)$ we use an argument similar to that in (8). For small dx we have the equation (again ignoring $o(dx)$ terms in the sequel)

$$\gamma(x+dx) = \left[1-\lambda e^{-\xi x}dx\right]\left[\gamma(x)+(1-\gamma(x))\,\mu dx\gamma(x)\right]. \tag{33}$$

By definition of (31), x is a record value and $\gamma(x)$ is the probability to reach level 0 before reaching a new record value. After the record value x is reached it is a necessary condition that arrivals in $[x, x+dx)$ will not be admitted to the system. The probability of the latter event is $\left[1-\lambda e^{-\xi x}dx\right]+o(dx)$. The interpretation of the expression is the same as that of (8). This leads to the differential equation

$$\gamma'(x) = -\lambda \mathrm{e}^{-\xi x}\gamma(x) + \mu(1-\gamma(x))\gamma(x).$$

Introducing $\nu(x) := \frac{1}{\gamma(x)}$, we have

$$\nu'(x) = \nu(x)\left(\lambda e^{-\xi x}-\mu\right)+\mu. \tag{34}$$

Solving for $\nu(x)$ in (34) we get, using that $\nu(0)=\gamma(0)=1$:

$$\nu(x) = \mathrm{e}^{-\mu x}\mathrm{e}^{\frac{\lambda}{\xi}(1-\mathrm{e}^{-\xi x})} + \mu\int_0^x \mathrm{e}^{\mu(y-x)}\mathrm{e}^{\frac{\lambda}{\xi}(\mathrm{e}^{-\xi y}-\mathrm{e}^{-\xi x})}\mathrm{d}y, \tag{35}$$

so

$$\gamma(x) = \frac{\mathrm{e}^{\mu x}\mathrm{e}^{\frac{-\lambda}{\xi}(1-\mathrm{e}^{-\xi x})}}{1+\mu\int_0^x \mathrm{e}^{\mu y}\mathrm{e}^{\frac{-\lambda}{\xi}(1-\mathrm{e}^{-\xi y})}\mathrm{d}y}. \tag{36}$$

Finally, in the next theorem we give the law of M in terms of Laplace-Stieltjes transforms (LST); see also Figure 2. It is expressed in the γ_j, which are given by (32) with $\gamma(x)$ being determined by (36).

Theorem 4

$$Ee^{-\alpha M} = \sum_{n=1}^{\infty}\left(\frac{\mu}{\mu+\alpha}\right)^n (1-\gamma_1)\cdots(1-\gamma_{n-1})\gamma_n. \tag{37}$$

Proof. As observed before, by the lack of memory property of the jumps, X_n is Erlang(n,μ) distributed, so that the LST of X_n is $\left(\frac{\mu}{\mu+\alpha}\right)^n$. By the strong Markov property the conditional probability that $N=n$ given there are at least $n-1$ record values is γ_n. Multiplying, the result follows. ■

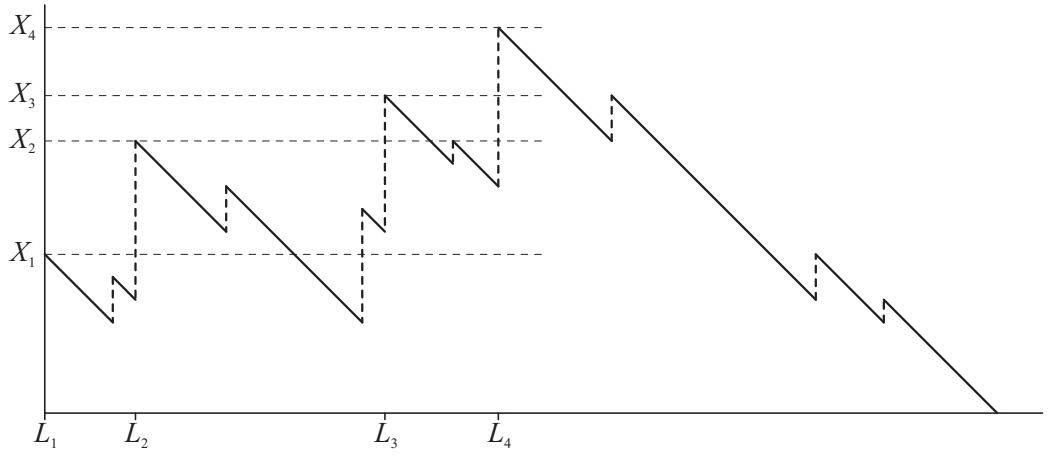

Figure 2: Model 1

Remark 3 *The LST of M is introduced in terms of an infinite sum. However, the RHS of (37) clearly converges. This fact allows us to approximate the LST of M by summing the RHS of (37) up to a predetermined number of terms.*

4.2 Model 2

As in Theorem 2, we use that $P(M > x) = \mathrm{e}^{-\int_0^x r_M(y)dy}$. In Model 2,

$$r_M(x) = (\mu + \xi)\gamma(x).$$

Similar to Model 1 we have for small dx, ignoring $o(dx)$ terms:

$$\gamma(x + dx) = \left[1 - \lambda e^{-\xi x} dx\right] \left[\gamma(x) + (1 - \gamma(x))\,(\mu + \xi)dx\gamma(x)\right].$$

This equation is the *same* as Equation (33) for Model 1, with μ replaced by $\mu + \xi$. Hence $\gamma(x)$ is as given in (36), with μ replaced by $\mu + \xi$. The hazard rate function $r_M(x)$, and hence $P(M > x)$, are now also specified.

4.3 Model 3

In a similar manner to that of Subsection 4.2, we introduce the hazard rate function

$$r_M(x) = \mu e^{-\xi x}\gamma(x),$$

where in this case

$$\gamma(x+dx) = \left[1 - \lambda e^{-\xi x}\frac{\mu}{\mu+\xi}dx\right]\left[\gamma(x) + (1-\gamma(x))\,\mu dx\gamma(x)\right]. \tag{38}$$

To better understand (38) note that by the memoryless property of the patience, $\lambda e^{-\xi x}\frac{\mu}{\mu+\xi}dx$ is the probability that a customer arrives and is admitted when the work is at level $x+dx$. Now notice that (38) is the *same* as Equation (33) for Model 1, with λ replaced by $\lambda\frac{\mu}{\mu+\xi}$. Hence $\gamma(x)$ is as given in (36), with λ replaced by $\lambda\frac{\mu}{\mu+\xi}$. Having found $r_M(x)$, $P(M>x)$ finally follows as before.

5 CONCLUSIONS

In this paper we have studied the cycle maximum of the following models: (i) The Markov mountain: a storage or dam model that alternates between exponentially distributed ON and OFF periods; the buffer content increases (decreases) at some state-dependent rate when ON (OFF). (ii) The semi-Markov mountain: as (i), but with generally distributed ON periods. (iii) The $M/G/1$ queue with three different forms of customer impatience.

We see the following problem as the main topic for further research. In the analysis of the semi-Markov mountain, we managed to obtain the workload density $f(\cdot)$ (see (18)), and we have expressed $\Pr(M>x)$ into $\theta(x)$ (see Theorem 3), but we did not yet manage to determine an explicit expression for that $\theta(x)$ except for the special case $\alpha(x)\equiv 1$. It is tempting to speculate that, for more general $\alpha(x)$, Formula (26) still holds, but with $(F*G)(x)$ being replaced by $(F\circledast G)(x)$.

Acknowledgment. We gratefully acknowledge insightful comments by a referee which have improved the presentation.

References

[1] I. Adan, O.J. Boxma and D. Perry (2005). "The $G/M/1$ Queue Revisited". *Mathematical Methods of Operations Research* 58(2), 437-452.

[2] H. Albrecher, S.C. Borst, O.J. Boxma and J.A.C. Resing (2009). "The Tax Identity in Risk Theory – A Simple Proof and an Extension". To appear in *Insurance Mathematics and Economics*.

[3] S. Asmussen (1998). "Extreme Value Theory for Queues Via Cycle Maxima". *Extremes* 1:2, 137-168.

[4] S. Asmussen (2000). *Ruin Probabilities*. World Scientific, Singapore.

[5] S. Asmussen (2003). *Applied Probability and Queues*. Wiley, New York. NOT CITED??

[6] S. Asmussen and D. Perry (1992). "On Cycle Maxima, First Passage Problems and Extreme Value Theory for Queues". *Stochastic Models* 8, 421-458.

[7] F. Baccelli, P. Boyer and G. Hebuterne (1984). "Single-Server Queues with Impatient Customers". *Advances in Applied Probability* 16, 887-905.

[8] R. Bekker and A.P. Zwart (2005). "On an Equivalence between Loss Rates and Cycle Maxima in Queues and Dams". *Probability in the Engineering and Informational Sciences* 19, 241-255.

[9] O.J. Boxma, D. Perry and F.A. van der Duyn Schouten (1999). "Fluid Queues and Mountain Processes". *Probability in the Engineering and Informational Sciences* 13, 407-427.

[10] Onno J. Boxma, Haya Kaspi, Offer Kella and David Perry (2005). "ON/OFF Storage Systems with State Dependent Input, Output and Switching Rates". *Probability in the Engineering and Informational Sciences* 19, 1-14.

[11] O.J. Boxma, D. Perry, W. Stadje and S. Zacks (2009). "The $M/G/1$ Queue with Quasi Restricted Accessibility". To appear in *Stochastic Models*.

[12] J.W. Cohen (1968). "Extreme Value Distributions for the $M/G/1$ and the $G/M/1$ Queueing Systems". *Ann. Inst. H. Poincaré Sect.* B 4, 83-98.

[13] J.W. Cohen (1969). "Single Server Queues with Restricted Accessibility". *Journal of Engineering Mathematics* 3, 265-285.

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

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

[16] J. M. Harrison and S.I. Resnick (1976). "The Stationary Distribution and First Exit Probabilities of a Storage Process with General Release Rule". *Math. Oper. Res.* 1(4), 347–358.

[17] J. Lee (2007). "First Exit Times of Compound Poisson Dams with a General Release Rule". *Math. Meth. Oper. Res.* 65, 169-178.

[18] David Perry, Wolfgang Stadje and Shelemyahu Zacks (2000). "Busy Period Analysis for $M/G/1$ and $G/M/1$-Type Queues with Restricted Accessibility". *Operations Research Letters* 27(4), 163-174.

[19] L. Takács (1967). *Combinatorial Methods in the Theory of Stochastic Processes*. Wiley, New York.