EURANDOM PREPRINT SERIES
2010-002

The $M/G/1+G$ queue revisited

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

The $M/G/1+G$ queue revisited

Onno Boxma[*] David Perry[†] and Wolfgang Stadje[‡]

Abstract

We consider an $M/G/1$ queue with the following form of customer impatience: an arriving customer balks or reneges when its virtual waiting time, i.e., the amount of work seen upon arrival, is larger than a certain random patience time. We consider the number of customers in the system, the maximum workload during a busy period, and the length of a busy period. We also briefly treat the analogous model in which any customer enters the system and leaves at the end of his patience time or at the end of his virtual sojourn time, whichever occurs first.
Key words. Single-server queue; impatience; balking; reneging; workload; number of customers; busy period; cycle maximum.

1 Introduction

Queueing systems with customer impatience represent a wide range of service systems in which customers may become impatient when they do not receive service fast enough. One may think of customers at call centers, or of customers representing perishable goods, like blood samples which wait to be tested and become obsolete after a certain due date. Depending on the application, patience may refer to waiting time or to sojourn (= waiting plus service) time. In the Kendall notation a single server queue with impatience is usually denoted by $G/G/1+G$; we would like to propose writing $G/G/1+G^w$ if patience refers to waiting time, and $G/G/1+G^s$ if it refers to sojourn time. We shall assume in the latter case that an impatient customer gets served until his patience runs out (partial rejection; his real sojourn time is the minimum of his patience and his waiting plus service time).

A pioneering paper on queueing models with impatience is [4]; it studies the queue length distribution in the $M/M/s+D^w$ model (where D indicates deterministic patience) and the $M/M/1+D^s$ model with partial rejection. Gavish and Schweitzer [9] consider the workload (virtual waiting time) and several other performance measures in the $M/M/1+D^s$ queue in the case of full rejection. In [1, 2] necessary and sufficient conditions for the existence of the steady-state virtual waiting-time distribution in the $G/G/1+G^w$ were obtained. The latter distribution was subsequently obtained for $M/G/1+M^w$ and $M/G/1+E_k^w$; see [3] for $M/G/1+D^w$, and see [10] who consider all three impatience/rejection rules. Finch [8] has derived the waiting-time distribution in the $G/M/1+D^w$ queue. Stanford [12] relates the waiting-time distribution of the (successful) customers and the workload seen by an arbitrary arrival in $G/G/1+G^w$. See Stanford [13] for a brief literature review. An important study regarding the busy period for models with customer impatience is

[*]EURANDOM and Department of Mathematics and Computer Science, Eindhoven University of Technology, HG 9.14, P.O. Box 513, 5600 MB Eindhoven, The Netherlands (boxma@win.tue.nl)

[†]Department of Statistics, University of Haifa, Haifa 31909 Israel (dperry@haifa.ac.il)

[‡]Department of Mathematics and Computer Science, University of Osnabrück, 49069 Osnabrück, Germany (wolfgang@mathematik.uni-osnabrueck.de)

Subba Rao [15]. He considers the $M/G/1 + M^w$ model, and derives the joint distribution of number of customers served in a busy period and length of that period, given that it starts with $i+1$ customers. In [14] he studies the combined effects of customer impatience and balking. Perry et al. [11] study the busy period for the $M/G/1$ and the $G/M/1$ model with restricted accessibility: a customer is partially rejected if the workload at his arrival is below a certain fixed threshold. For other variants of $M/G/1$ models with restricted accessibility see [10, 6]. In a recent study [7] the busy period distribution in $M/G/1 + G^w$ is obtained for various choices of the patience time distribution. The main cases under consideration are exponential patience and a discrete patience distribution.

The present paper aims to present some new approaches, novel views and results for $M/G/1+G^w$ and $M/G/1 + G^s$. The paper is organized as follows. Section 2 contains the model description and briefly reviews a result from [2] for the workload distribution in the $M/G/1 + G^w$ queue, which is needed in the remainder of the paper. Section 3 considers the steady-state distribution of the number of customers, and in Section 4 we derive the distribution of the maximum workload in a busy period for the case of *exponential* service times. Finally, Section 5 presents a new view on the busy period distribution. In each section we mainly focus on the G^w case, but we also present analogous results for the G^s case.

For $M/G/1 + G^w$ one can distinguish two submodels: (i) the *balking* case, which we also indicate as the *observable* case: every customer knows his own (potential) and all prior service requirements and patience times upon his arrival so that he either leaves immediately or stays and will eventually receive service; (ii) the *reneging* case, which we also indicate as the *unobservable* case: every customer enters the system and leaves only when his patience runs out or his service terminates. Note that neither the waiting times (of eventually served customers) nor the busy periods would be affected if customers stayed until their patience is running out; the same holds for the workload process provided that only the work that is really carried out is taken into account. Of course, the numbers of customers present will be different in the two submodels.

2 The waiting time in steady state

Consider an $M/G/1 + G^w$ queue with arrival rate λ, service time distribution G and patience distribution H. If the patience Y_i of the ith customer exceeds the workload present upon his arrival, he leaves immediately. By PASTA, the steady-state workload distribution equals the steady-state waiting-time distribution. Denote their density by f; a level crossing argument readily implies that it exists. Following [2], we briefly review what is known about f. By level crossing, f satisfies the integral equation

$$f(x) = \lambda \int_0^x [1 - G(x-w)][1 - H(w)] f(w)\ \mathrm{d}w + \lambda\pi[1 - G(x)], \tag{2.1}$$

where π is the steady-state probability that the system is empty. Define

$$K(x,w) = \lambda[1 - G(x-w)][1 - H(w)], \quad x \geq 0, \tag{2.2}$$

and

$$K_{n+1}(x,y) = \int_y^x K_n(x,w)K(w,y)\ \mathrm{d}w = \int_y^x K(x,w)K_n(w,y)\ \mathrm{d}w, \quad x \geq y,\ n \geq 1.$$

Then (2.1) takes the form

$$f(x) = \int_0^x K(x,w)f(w)\ \mathrm{d}w + \pi K(x,0)$$

and by setting $K^0(x,y) = \sum_{n=1}^{\infty} K_n(x,y)$, we get

$$f(x) = \pi \sum_{n=1}^{\infty} K_n(x,0) = \pi K^0(x,0). \tag{2.3}$$

The normalizing condition yields

$$\pi \int_0^{\infty} K^0(x,0)\ \mathrm{d}x = 1 - \pi$$

and

$$\pi = \left[1 + \int_0^{\infty} K^0(x,0)\ \mathrm{d}x\right]^{-1}.$$

In the $M/G/1 + G^s$ (with partial rejection, as mentioned in the introduction) the balance equation for the workload density in steady state is

$$f(x) = \lambda \int_0^x [1 - G(x-w)][1 - H(x)]f(w)\ \mathrm{d}w + \lambda\pi[1 - G(x)][1 - H(x)], \quad x \geq 0, \tag{2.4}$$

so that the solution of f in (2.4) is the same as that given in (2.3) with the kernel of (2.2) replaced by the kernel

$$K(x,w) = \lambda[1 - G(x-w)][1 - H(x)], \quad x \geq 0.$$

3 The number of customers in the system

3.1 The $M/G/1 + G^w$ model

We start by considering the $M/G/1+G^w$ model for the unobservable case (reneging). In this model every customer enters the system and only leaves the waiting line before service when his patience runs out. We now derive the steady-state distribution of the number of customers present.

Let W_n denote the current workload when the nth customer arrives. The sojourn time of this customer, D_n, is given by

$$D_n = W_n + S_n + [Y_n - (W_n + S_n)] \cdot 1_{\{Y_n < W_n\}}$$

where Y_n is his patience time and S_n is his service requirement.

In the ordinary $M/G/1$ FCFS system there is a simple relation between the sojourn time and the number of customers, because the customers left behind by a departing customer are exactly the ones that arrive during his sojourn time. However, the latter argument cannot be applied in the $M/G/1 + G^w$ system. To see this, consider a departure due to impatience of an arbitrary customer who was not first in line. Obviously, the customers who are in front of him in line arrived before his arrival and thus not during his sojourn time. To deal with this difficulty we construct a modified $M/G/1 + G^w$-type system based on the same collections of random variables S_n and Y_n and the same arrival process. In this modified system there are no departures due to impatience;

its evolution proceeds as follows. All arriving customers enter the system and receive service but each customer who reaches the end of his patience time while waiting in line only gets a service time ε when his *waiting* time has elapsed, for some prespecified small $\varepsilon > 0$ (that will later tend to zero). The service times of the other customers are unchanged. In this system all departures are due to service terminations, and any departing customer leaves behind only the customers that arrived during his sojourn time. Note that the sojourn times in the original $M/G/1 + G^w$ system and the sojourn times in the modified one are not the same. But the number of customers, at an arbitrary time, in the original $M/G/1 + G^w$ system and the number of customers in the modified system whose service times are not (yet) changed *are* the same. Therefore, we will now derive the steady-state law of the number of customers in the modified $M/G/1 + G^w$ system whose service times have not yet been changed.

To this end let W_ε and D_ε^{mod} be the workload as seen upon arrival and the sojourn time in the modified system in steady state. It should be noted that, for $\varepsilon \to 0$, W_ε will converge to the steady-state workload W whose density was given in (2.3); however, D_ε^{mod} does *not* converge to the sojourn time in the original system – hence the superscript mod. We have

$$\mathbb{E}e^{-\alpha D_\varepsilon^{mod}} = \mathbb{E}\left(e^{-\alpha D_\varepsilon^{mod}} \mid Y < W_\varepsilon\right)\mathbb{P}(Y < W_\varepsilon) + \mathbb{E}\left(e^{-\alpha D_\varepsilon^{mod}} \mid Y \geq W_\varepsilon\right)\mathbb{P}(Y \geq W_\varepsilon).$$

Clearly, $Y < W_\varepsilon$ implies that $D_\varepsilon^{mod} = W_\varepsilon + \varepsilon$ and $Y \geq W_\varepsilon$ implies that $D_\varepsilon^{mod} = W_\varepsilon + S$, where W_ε, Y and S represent generic random variables for the waiting time in steady state, the patience time and the service time, respectively, and are independent. Thus,

$$\begin{aligned} \mathbb{E}e^{-\alpha D_\varepsilon^{mod}} &= \mathbb{E}e^{-\alpha(W_\varepsilon+\varepsilon)}1_{\{Y<W_\varepsilon\}} + \mathbb{E}e^{-\alpha(W_\varepsilon+S)}1_{\{Y\geq W_\varepsilon\}} \\ &= e^{-\alpha\varepsilon}\mathbb{E}[e^{-\alpha W_\varepsilon}H(W_\varepsilon)] + G^*(\alpha)\mathbb{E}[e^{-\alpha W_\varepsilon}(1-H(W_\varepsilon))]. \end{aligned} \tag{3.1}$$

In the case that $1 - H(x) = e^{-\xi x}$ we get in (3.1)

$$\mathbb{E}e^{-\alpha D_\varepsilon^{mod}} = e^{-\alpha\varepsilon}[\mathbb{E}[e^{-\alpha W_\varepsilon}] - \mathbb{E}[e^{(\alpha+\xi)W_\varepsilon}]] + G^*(\alpha)\mathbb{E}[e^{-(\xi+\alpha)W_\varepsilon}].$$

At any moment of departure from the modified system two types of customers are left behind. We call customers of type 1 those whose service time has not (yet) been changed and customers of type 2 those whose service time has been changed to ε. Let θ_n^ε be the probability that n waiting customers of type 1 are left behind at a moment of departure. Then

$$\begin{aligned} \theta_n^\varepsilon &= \int_0^\infty \sum_{k=n}^\infty \binom{k}{n} p_t^n q_t^{k-n} \frac{e^{-\lambda t}(\lambda t)^k}{k!} \mathrm{d}F_{D_\varepsilon^{mod}}(t) \\ &= \int_0^\infty \frac{e^{-\lambda p_t t}(\lambda p_t t)^n}{n!} \mathrm{d}F_{D_\varepsilon^{mod}}(t) \end{aligned} \tag{3.2}$$

where $F_{D_\varepsilon^{mod}}$ is the distribution function of D_ε^{mod} (whose Laplace-Stieltjes transform is given in (3.1)), p_t is the probability that an arbitrary customer who arrived during a time interval of length t is of type 1 and $q_t = 1 - p_t$. Given the number of customer arrivals during a time interval of length t is k, their arrival times have the same joint distribution as the order statistics taken from a uniform sample of size $k \geq n$. Thus,

$$p_t = \frac{1}{t}\int_0^t [1 - H(t-s)]\ \mathrm{d}s = \frac{1}{t}\int_0^t [1 - H(s)]\ \mathrm{d}s.$$

Therefore

$$\theta_n^\varepsilon = \int_0^\infty \frac{e^{-\lambda \int_0^t [1-H(s)]\ \mathrm{d}s} (\lambda \int_0^t [1-H(s)]\ \mathrm{d}s)^n}{n!} \mathrm{d}F_{D_\varepsilon^{mod}}(t). \tag{3.3}$$

Computing the generating function $\Theta_\varepsilon(z) = \sum_{n=0}^\infty \theta_n^\varepsilon z^n$ from (3.3) we get

$$\Theta_\varepsilon(z) = \int_0^\infty \exp\left\{-\lambda(1-z)\int_0^t [1-H(s)]\ \mathrm{d}s\right\} \mathrm{d}F_{D_\varepsilon^{mod}}(t).$$

Now let $\varepsilon \to 0$. Then D_ε^{mod} converges in distribution to the steady-state sojourn time of the modified $M/G/1+G^w$ system and, by (3.3) and the fact that the integrand in (3.3) is a bounded continuous function, θ_n^ε converges to the steady-state probability θ_n that in the original system a departing customer leaves n customers behind. In particular, with D^{mod} denoting the steady-state sojourn time and $F_{D^{mod}}$ its distribution function,

$$\Theta(z) = \lim_{\varepsilon\to 0} \Theta_\varepsilon(z) = \int_0^\infty \exp\left\{-\lambda(1-z)\int_0^t [1-H(s)]\ \mathrm{d}s\right\} \mathrm{d}F_{D^{mod}}(t) \tag{3.4}$$

is the generating function of the number of customers left behind in the original $M/G/1+G^w$ system. By ASTA, $\Theta(z)$ is also the generating function of the number of waiting customers in steady state. Note that the LST of D^{mod} can be expressed in terms of the steady-state waiting time W in the form

$$\mathbb{E}e^{-\alpha D^{mod}} = \mathbb{E}[e^{-\alpha W} H(W)] + G^*(\alpha)\mathbb{E}[e^{-\alpha W}(1-H(W))]. \tag{3.5}$$

The density f of W is given by (2.3).

We now turn to the case of balking (observable case). Again, we construct a modified system in which every customer enters the system but, in contrast to the "non-observable case", the service requirement of a customer of type 2 is changed to ε *upon his arrival*. Then, a customer is of type 1 if and only if his patience is greater than his waiting time. Accordingly, the number of customers of type 1 left behind at a departure (and thus also in steady state) is given by

$$\theta_n^\varepsilon = \int_0^\infty \frac{e^{-\lambda t \mathbb{P}(Y>W_\varepsilon)} (\lambda t \mathbb{P}(Y>W_\varepsilon))^n}{n!} \mathrm{d}F_{D_\varepsilon^{mod}}(t),$$

with generating function

$$\Theta^\varepsilon(z) = \int_0^\infty \mathrm{e}^{-\lambda \mathbb{P}(Y>W)(1-z)t} \mathrm{d}F_{D_\varepsilon^{mod}}(t).$$

Notice that $\mathbb{E}e^{-\alpha D_\varepsilon^{mod}}$ is still given by (3.1). For $\varepsilon \to 0$ we get, using a similar continuity argument as for the unobservable case:

$$\Theta(z) = \lim_{\varepsilon\to 0} \Theta_\varepsilon(z) = \int_0^\infty \mathrm{e}^{-\lambda \mathbb{P}(Y>W)(1-z)t} \mathrm{d}F_{D^{mod}}(t).$$

In case that Y is $\exp(\xi)$-distributed we get

$$\Theta(z) = \int_0^\infty \mathrm{e}^{-\lambda W^*(\xi)(1-z)t} \mathrm{d}F_{D^{mod}}(t), \tag{3.6}$$

where $W^*(\xi)$ is the LST of the steady-state workload W.

3.2 The $M/G/1 + G^s$ model

We now turn to the case of patience referring to sojourn time, allowing partial rejection. We restrict ourselves to the unobservable variant, so

$$D_n = \min(Y_n, W_n + S_n).$$

We use the same arguments as above, featuring customers of type 1 and customers of type 2. The formula for the generating function $\Theta(z)$ of the number of customers in steady state as given in (3.4) still holds but the distribution $F_D^{mod}(t)$ has to be modified. As before, in the modified system a customer with patience less than his waiting time gets a service time ε when his waiting time has elapsed; eventually $\epsilon \downarrow 0$. Distinguishing between the events $\{Y < W\}$, $\{W \leq Y < W + S\}$ and $Y \geq W + S\}$, it is seen that F_D^{mod} has LST

$$\begin{aligned} \mathbb{E}e^{-\alpha D^{mod}} &= \mathbb{E}[e^{-\alpha W} H(W)] + \mathbb{E}[\int_W^{W+S} e^{-\alpha y} \mathrm{d}H(y)] \\ &\quad + \mathbb{E}[e^{-\alpha(W+S)}(1 - H(W+S))]. \end{aligned} \tag{3.7}$$

Remark.
For the special cases $M/M/1 + M^w$ and $M/M/1 + M^s$ it is possible to obtain the law of the number of customers in steady state in a more straightforward manner. In both cases the number of customers constitutes a Markov process. Thus, for the special case $M/M/1+M^w$ with impatience rate η we have the balance equations

$$\lambda p_n = (\mu + n\eta)p_{n+1}, \quad n = 0, 1, \ldots$$

and for $M/M/1 + M^s$ we get

$$\lambda p_n = [\mu + (n+1)\eta]p_{n+1}, \quad n = 0, 1, \ldots$$

These recursive equations can be easily solved. Note that it is enough to solve only the $M/M/1+M^w$ case, since the solution for $M/M/1 + M^s$ is the same as that for the $M/M/1 + M^w$ case in which μ is replaced by $\mu + \eta$.

Remark.
¿From (3.4) and (3.6) it is possible to derive the appropriate Little's law for the $M/G/1 + G^w$ and the $M/G/1 + G^s$ systems, respectively. Let X be the number of customers in steady state. Then by taking derivatives in (3.4) and setting $z = 1$ we obtain for $M/G/1 + G^w$ in the reneging case

$$\mathbb{E}X = \lambda \int_0^\infty [1 - H(s)][1 - F_{D^{mod}}(s)] \, \mathrm{d}s.$$

Similarly, for $M/G/1 + G^w$ in the balking case

$$\mathbb{E}X = \lambda \mathbb{E}D^{mod} \cdot \mathbb{P}(Y > W),$$

where Y is the generic patience and W is the steady state waiting time. The same expressions hold for the $M/G/1 + G^s$ system, but then D^{mod} is the sojourn time whose LST is given in (3.7).

Remark.
One can derive a simple (though important) conservation law holding for the $M/G/1 + G^w$ and the $M/G/1 + G^s$ systems. Let ξ be the rejection rate of the customers who eventually do not receive service in the $M/G/1 + G^w$ system. Then clearly

$$\lambda = \xi + \frac{1 - \pi}{\mathbb{E}S},$$

where π is the steady-state probability of an empty system and S is a generic service time.

4 The maximum workload during a busy cycle

4.1 The $M/M/1+G^w$ model

Assume that at time zero a new customer enters an empty system. Let $V(t)$ denote the workload at time t. We want to compute the LST of $M = \max_{0\le t\le B} V(t)$, the maximum workload during a busy period B. The problem can be 'discretized' as follows. Let $X_1 = V(0), X_2, X_3, \ldots$ be the sequence of successive record values of the process $\{V(t) : t \ge 0\}$. Then $M = X_N$ where N is the number of record values in the first busy period. But $N = j$ means that for $i = 1, \ldots, j-1$ in the time between the ith and the $(i+1)$st record, $V(t)$ does not reach zero but it becomes zero between the jth and the $(j+1)$st record. This problem is solvable explicitly in the case $M/M/1+G^w$ of $\exp(\mu)$-distributed service times (and general patience times); in general it seems analytically intractable because the segments of the sample path in the above decomposition are not independent.

By the lack-of-memory property of the $\exp(\mu)$ distribution of the upward jumps of workload $\mathbf{V}$ for $M/M/1+G^w$, the jth record value exceeds the $(j-1)$th one by an $\exp(\mu)$-distributed 'overshoot' independent of the past.

Now let $\gamma(x)$ be the conditional probability that the system will become empty between the ith and the $(i+1)$st record time given that $X_i = x$ (clearly $\gamma(x)$ does not depend on i). Let us determine the probability to hit zero before reaching a new record value when starting from a record value at level $x + \mathrm{d}x$. For small $\mathrm{d}x$ we have the equation

$$\gamma(x+\mathrm{d}x) = [1-\lambda(1-H(x))\ \mathrm{d}x]\,[\gamma(x) + (1-\gamma(x))\,\mu \mathrm{d}x\gamma(x)] + o(\mathrm{d}x). \tag{4.1}$$

Indeed, except for terms of order $o(\mathrm{d}x)$ one has to go down during the first dx time units without new arrivals entering the system (the probability of the latter event is $1-\lambda(1-H(x))\ \mathrm{d}x + o(\mathrm{d}x)$), then continue from level x and either (i) hit zero before exceeding x (probability $\gamma(x)$) or (ii) not hit zero before jumping back to the interval $(x, x+\mathrm{d}x]$ and then continue from there and hit zero before the next record (probability $(1-\gamma(x))\mu\ \mathrm{d}x\ \gamma(x) + o(\mathrm{d}x)$).

This leads to the Bernoulli differential equation

$$\gamma'(x) = -\lambda\,[1-H(x)]\,\gamma(x) + \mu(1-\gamma(x))\gamma(x).$$

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

$$\nu'(x) = \nu(x)\,(\lambda\,[1-H(x)] - \mu) + \mu. \tag{4.2}$$

Solving for $\nu(x)$ in (4.2) we get, using that $\nu(0) = \gamma(0) = 1$:

$$\gamma(x) = \left[\mu \int_0^x \exp\left\{\int_v^x [\lambda(1-H(u)) - \mu]\ \mathrm{d}u\right\}\ \mathrm{d}v + \exp\left\{\int_0^x [\lambda(1-H(u)) - \mu]\ \mathrm{d}u\right\}\right]^{-1}. \tag{4.3}$$

Theorem 1.

$$\mathbb{P}(M > x) = \exp\left\{-\mu \int_0^x \gamma(u)\ \mathrm{d}u\right\}.$$

Proof. The hazard rate $h(x)\ \mathrm{d}x = \Pr(M \in (x, x+\mathrm{d}x] \mid M > x)$ is given by

$$h(x)\ \mathrm{d}x = \mu\gamma(x)\ \mathrm{d}x + o(\mathrm{d}x).$$

Indeed, conditional on $M > x$ the cycle maximum falls in $(x, x+\mathrm{d}x]$ if and only if there is a record value in this interval (this has probability $\mu\ \mathrm{d}x + o(\mathrm{d}x)$) such that before the next record value is reached the system becomes empty (which has probability $\gamma(x) + O(\mathrm{d}x)$). □

By the above structural results we can also write down explicit expressions for certain functionals of the cycle records, for example

$$\mathbb{E}e^{-\alpha M}1_{\{N=j\}} = \mu^j \int_0^\infty \dots \int_0^\infty e^{-(\alpha+\mu)(u_1+\dots+u_j)}(1-\gamma(u_1))(1-\gamma(u_1+u_2)) \\ \dots(1-\gamma(u_1+\dots+u_{j-1}))\gamma(u_1+\dots+u_j)\ \mathrm{d}u_1 \dots \mathrm{d}u_j.$$

Remark.
In the special case of infinite patience, it follows from (4.3) that

$$\gamma(x) = \frac{1-\rho}{1-\rho\mathrm{e}^{-\mu(1-\rho)x}}. \tag{4.4}$$

It is well-known (cf. Section 29 of [16]) that, for the $M/G/1$ queue,

$$\Pr(M < x) = \frac{\Pr(W+S<x)}{\Pr(W<x)},$$

where W denotes waiting time or workload and S denotes service time. Hence for $M/M/1$

$$\Pr(M < x) = \frac{1-\mathrm{e}^{-\mu(1-\rho)x}}{1-\rho\mathrm{e}^{-\mu(1-\rho)x}}.$$

Now it is easy to verify that the failure rate of M, i.e., $\frac{d}{dx}\mathbb{P}(M<x)/\mathbb{P}(M \geq x)$, equals $\mu\gamma(x)$ as given in (4.4), in agreement with Theorem 1.

Remark.
It should be noticed that these results for M correct Theorem 4 of [5], where in $M = X_1 + \dots + X_N$ the dependence of N and the X_j was ignored.

4.2 The $M/M/1+M^s$ model

Now let us again consider the system with sojourn time impatience. For the steady-state law of the cycle maximum we restrict the attention to the $M/M/1+M^s$ system (in the case of *sojourn time* impatience, non-exponential impatience destroys the lack-of-memory property which we used to derive (4.1)) and apply Section 4.2 of [5]. We obtain the differential equation (4.1) with $1-H(x) = e^{-\xi x}$ and with $\mu+\xi$ replacing μ. That is, we get

$$\gamma(x+\mathrm{d}x) = \left[1-\lambda e^{-\xi x}\ \mathrm{d}x\right]\left[\gamma(x) + (1-\gamma(x))\,(\mu+\xi)\ \mathrm{d}x\ \gamma(x)\right] + o(\mathrm{d}x). \tag{4.5}$$

The solution is given in (4.3) but with $\mu+\xi$ replacing μ and with $1-H(x) = e^{-\xi x}$. Then, the hazard rate function is $r_M(x) = (\mu+\xi)\gamma(x)$ and

$$\mathbb{P}(M > x) = \exp\left\{-(\mu+\xi)\int_0^x \gamma(u)\ \mathrm{d}u\right\}.$$

5 The busy period

5.1 The $M/G/1+G^w$ model

Let B be the duration of a busy period, initiated by a customer arriving at an empty system, of the $M/G/1+G^w$ queue under consideration. In this section we present a novel, algorithmic, method for obtaining the LST of B (see [15, 7] for other studies of the busy period in $M/G/1+G^w$) and some more general level crossing results. Let $\{V(t): t \geq 0\}$ be the virtual waiting-time process and define the stopping time

$$T_x = \inf\{t : V(t) = x\}$$

and

$$\Gamma(\alpha; u, x) = \mathbb{E}(e^{-\alpha T_x} \mid V(0) = u), \quad u > x \geq 0,$$

which is the conditional LST of the time it takes the workload to move from level u down to level x. We also need the conditional LST

$$\Psi(\alpha; x) = \int_0^\infty \Gamma(\alpha; x+y, x)\ \mathrm{d}G(y). \tag{5.1}$$

Lemma 1. *We have*

$$\Psi(\alpha, x)\ = \int_0^\infty \exp\left\{-\left[\alpha y + \lambda \int_0^y [1-H(x+u)][1-\Psi(\alpha; x+u)]\ \mathrm{d}u\right]\right\}\ \mathrm{d}G(y). \tag{5.2}$$

Proof. Considering what can happen in an infinitesimally small interval $(x+y, x+y+\mathrm{d}y)$ we obtain

$$\begin{aligned}\Gamma(\alpha; x+y+\mathrm{d}y, x) \ &= \ [1-\lambda(1-H(x+y))\mathrm{d}y][1-\alpha\ \mathrm{d}y]\Gamma(\alpha; x+y, x) \\ &\quad +\lambda(1-H(x+y))\Psi(\alpha; x+y)\Gamma(\alpha; x+y, x)\mathrm{d}y + o(\mathrm{d}y).\end{aligned}$$

This equation can be written as

$$d\Gamma(\alpha; x+y, x)/dy)/\Gamma(\alpha; x+y, x) = -\alpha - \lambda[1-H(x+y)][1-\Psi(\alpha; x+y)].$$

Integrating it from 0 to y and using $\Gamma(\alpha; x, x) = 1$ yields:

$$\Gamma(\alpha; x+y, x) = e^{-\left[\alpha y + \lambda \int_0^y [1-H(x+u)][1-\Psi(\alpha; x+u)]\ \mathrm{d}u\right]},$$

and the proof is completed by substituting $\Gamma(\alpha; x+y, x)$ in (5.1). □

The lemma provides a complicated functional equation for the function $x \mapsto \Psi(\alpha; x)$ for any fixed $\alpha > 0$. This equation uniquely determines $\Psi(\alpha; \cdot)$ for every sufficiently large α. To see this, let $\mathcal{B}$ be the set of bounded functions in $C[0, \infty)$ which, endowed with the supremum norm, is a Banach space. For fixed $\alpha > 0$ define the operator $I_\alpha : \mathcal{B} \to \mathcal{B}$ by

$$(I_\alpha f)(x) = \int_0^\infty e^{-\left[\alpha y + \lambda \int_0^y [1-H(x+u)][1-f(x+u)]\mathrm{d}u\right]} \mathrm{d}G(y), \quad x \in [0, \infty),\ f \in \mathcal{B}.$$

If α is large enough, this operator is a contraction because for $f, \tilde{f} \in \mathcal{B}$,

$$\begin{aligned}
& |(I_\alpha f)(x) - (I_\alpha \tilde{f})(x)| \\
\leq & \int_0^\infty e^{-\alpha y} \left| e^{-\lambda \int_0^y [1-H(x+u)][1-f(x+u)]\ \mathrm{d}u} - e^{-\lambda \int_0^y [1-H(x+u)][1-\tilde{f}(x+u)]\ \mathrm{d}u} \right| \mathrm{d}G(y) \\
& \leq \int_0^\infty e^{-\alpha y} \lambda \left| \int_0^y [1-H(x+u)][1-f(x+u)]\ \mathrm{d}u \right. \\
& \qquad \left. - \int_0^y [1-H(x+u)][1-\tilde{f}(x+u)]\ \mathrm{d}u \right| \mathrm{d}G(y) \\
& \leq \lambda \int_0^\infty y e^{-\alpha y} ||f - \tilde{f}||_\infty \mathrm{d}G(y),
\end{aligned}$$

yielding

$$||I_\alpha f - I_\alpha \tilde{f}||_\infty \leq \lambda J(\alpha) ||f - \tilde{f}||_\infty,$$

where $J(\alpha) = \int_0^\infty y e^{-\alpha y} \mathrm{d}G(y)$. Clearly $J(\alpha) \to 0$ as $\alpha \to \infty$. It follows that I_α is a contraction if $\lambda J(\alpha) < 1$, which holds if

$$\alpha > \alpha_0 = \sup\{\beta \geq 0 \mid J(\beta) \geq 1/\lambda\}.$$

In this case I_α has a unique fixed point in $\mathcal{B}$, which is of course given by $\Psi(\alpha; \cdot)$. Moreover, for arbitrary $f_0 \in \mathcal{B}$ the sequence $(I_\alpha)^n f_0$, $n \geq 1$, converges to $\Psi(\alpha; \cdot)$ uniformly on $[0, \infty)$ at an exponential rate.

Now setting $x = 0$ in (5.1) we get

$$\Psi(\alpha; 0) = \mathbb{E} e^{-\alpha B}.$$

Hence, (5.2) yields

Theorem 2. *The LST of B is given by*

$$\mathbb{E} e^{-\alpha B} = \int_0^\infty \exp\left\{ - \left[\alpha y + \lambda \int_0^y [1-H(u)][1-\Psi(\alpha; u)]\ \mathrm{d}u \right] \right\}\ \mathrm{d}G(y).$$

Thus, we have a method to compute $\mathbb{E} e^{-\alpha B}$ for $\alpha > \alpha_0$. Note that the distribution of B is uniquely determined by its LST on the interval (α_0, ∞).

5.2 The $M/G/1 + G^s$ model

For the busy period of the sojourn time impatience model corresponding to $M/G/1 + G^s$ we use the same arguments as for $M/G/1 + G^w$. First define

$$\Phi(\alpha; x) = \int_0^\infty \Gamma(\alpha; x+v, x) \left\{ (1 - H(x+v)) \mathrm{d}G(v) + (1 - G(v)) \mathrm{d}H(x+v) \right\}. \tag{5.3}$$

Notice that the busy period LST is obtained as $\mathbb{E} \mathrm{e}^{-\alpha B} = \Phi(\alpha; 0)$.

Lemma 2. *We have*

$$\begin{aligned}
\Phi(\alpha, x) \quad &= \quad \int_0^\infty \exp\left\{ - \left[\alpha y + \lambda \int_0^y [1-H(x+u)]\ \mathrm{d}u - \lambda \int_0^y \Phi(\alpha; x+u)\ \mathrm{d}u \right] \right\} \\
&\times \quad \left\{ (1 - H(x+y)) \mathrm{d}G(y) + (1 - G(y)) \mathrm{d}H(x+y) \right\}.
\end{aligned} \tag{5.4}$$

Proof. Starting from

$$\begin{aligned}\Gamma(\alpha; x+y+\mathrm{d}y, x) &= [1-\lambda(1-H(x+y))\mathrm{d}y][1-\alpha\ \mathrm{d}y]\Gamma(\alpha; x+y, x) \\ &\quad +\lambda\Phi(\alpha; x+y)\Gamma(\alpha; x+y, x)\mathrm{d}y + o(\mathrm{d}y),\end{aligned}$$

we get

$$\Gamma(\alpha; x+y, x) = \exp\left\{-\left[\alpha y + \lambda\int_0^y [1-H(x+u)]\ \mathrm{d}u - \lambda\int_0^y \Phi(\alpha; x+u)\ \mathrm{d}u\right]\right\}.$$

Now use (5.3). □

This lemma provides a complicated functional equation for the function $x \mapsto \Phi(\alpha; x)$ for any fixed $\alpha > 0$. This equation uniquely determines $\Psi(\alpha; \cdot)$ for every sufficiently large α. Notice the similarity to the functional equation for the busy period LST in an ordinary $M/G/1$ queue.

Acknowledgements. The research of O.J. Boxma was done within the framework of the BRICKS project and the European Network of Excellence Euro-NF. D. Perry gratefully acknowledges a visitor grant from the Netherlands Organisation for Scientific Research NWO. W. Stadje was supported by the Deutsche Forschungsgemeinschaft.

References

[1] F. Baccelli and G. Hébuterne (1981). On queues with impatient customers. In: F.J. Kylstra (ed.). *Performance '81* (North-Holland Publ. Cy., Amsterdam), pp. 159-179.

[2] F. Baccelli, P. Boyer and G. Hébuterne (1984). Single-server queues with impatient customers. *Adv. Appl. Probab.* **16**, 887-905.

[3] J. Bae, S. Kim and E.Y. Lee (2001). The virtual waiting time of the $M/G/1$ queue with impatient customers. *Queueing Systems* **38**, 485-494.

[4] D.Y. Barrer (1957). Queuing with impatient customers and ordered service. *Oper. Res.* **5**, 650-656.

[5] O.J. Boxma and D. Perry (2009). On the cycle maximum of mountains, dams and queues. *Comm. Stat., Theory & Methods* **38**, 2706-2720 (special issue in honor of Shelley Zacks).

[6] O.J. Boxma, D. Perry, W. Stadje and S. Zacks (2009). The $M/G/1$ queue with quasi-restricted accessibility. *Stochastic Models* **25**, 151-196.

[7] O.J. Boxma, D. Perry, W. Stadje and S. Zacks (2009). The busy period of an $M/G/1$ queue with customer impatience. *EURANDOM Report 2009-012*. To appear in: *J. Appl. Prob.*

[8] P.D. Finch (1960). Deterministic customer impatience in the queueing system $GI/M/1$. *Biometrika* **47**, 45-52.

[9] B. Gavish and P.J. Schweitzer (1977). The Markovian queue with bounded waiting time. *Management Science* **23**, 1349-1357.

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

[11] D. Perry, W. Stadje and S. Zacks (2000). Busy period analysis for $M/G/1$ and $G/M/1$ type queues with restricted accessibility. *Oper. Res. Letters* **27**, 163-174.

[12] R.E. Stanford (1979). Reneging phenomena in single channel queues. *Math. Oper. Res.* **4**, 162-178.

[13] R.E. Stanford (1990). On queues with impatience. *Adv. Appl. Probab.* **22**, 768-769.

[14] S. Subba Rao (1967). Queuing models with balking and reneging. *Ann. Inst. Math. (Japan)* **19**, 55-71.

[15] S. Subba Rao (1967/1968). Queuing with balking and reneging in $M/G/1$ systems. *Metrika* **6**, 173-188.

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