A Sample Path Analysis of M/GI/1 Queues with Workload Restrictions

Jian-Qiang Hu
Dept. of Manufacturing Engineering
Boston University
Boston, MA 02215

Michael A. Zazanis *
Dept. of Industrial Engineering
Northwestern University
Evanston, IL 60208

Abstract

A simple random time change is used to analyze M/GI/1 queues with workload restrictions. The types of restrictions considered include workload bounds and rejection of jobs whose waiting times exceed a (possibly random) threshold. Load dependent service rates and vacations are also allowed and in each case the steady state distribution of the workload process for the system with workload restrictions is obtained in terms of that of the corresponding M/GI/1 queue without restrictions. The novel sample path arguments used simplify and generalize previous results.
Keywords: Workload Restrictions, Finite Dams, Balking
Short Title: M/GI/1 Queues with Restricted Workload

1 Introduction

We consider queueing systems with compound Poisson input process, a load dependent server whose behavior may include vacations, and workload restrictions. Typical forms of restrictions are:

i) A bound b on the workload. A finite dam where the overflow is lost is the classic example. Another example is a *processor sharing* queue where jobs have a fixed time limit b in the system. If a job has not finished service by the time the limit b is reached, it leaves without completing its service requirement. Here the workload at time t, X_t, represents the "real" load on the processor; if no further work arrives after t and the processor works at unit rate, the queue would become idle at $t + X_t$.

ii) Balking in FCFS queues when waiting (or sojourn) times exceed a certain (possibly random) threshold a. We may assume either that a job does not join the queue at all if its waiting time exceeds a or that it joins the queue but is subsequently "timed out" if its waiting time

*Supported in part by N.S.F. Grant DDM-8905638

exceeds a. The two models are equivalent from the point of view of real workload but differ of course if one considers the number of jobs in the system. The more general case where a job upon arrival at time t joins the queue with probability $p(X_t)$ is also considered.

An extensive literature exists on such systems. The steady state distribution for the finite dam with compound Poisson input and a constant release rule was first obtained by Gani and Prabhu [7] under the assumption of deterministic inflows. Markovian queues with balking based on the workload were first considered in Barrer [2]. For an overview of early results we refer the reader to Gnedenko and Kovalenko [10], Cohen [5], and the references therein.

We shall illustrate our results via the M/GI/1 queue with workload bounded above by b. Let the steady state distribution for the workload process be $F_b(x)$, assume that the unrestricted system is stable, and denote by $F(x)$ the steady state distribution of the unrestricted workload process. Then the workload process in the restricted system with bound b has steady state distribution function given by

$$F_b(x) \;=\; \frac{F(x)}{F(b)} \;, \qquad \forall\, x \in [0,b] \;. \tag{1}$$

This was established by Takács [15]. Earlier results include Ghosal [8] and Takács [14]. Franken et al. [6, pp. 155–157] obtain the same result for M/GI/1 queues with bounded workload and "warm-up periods" (vacations, in more standard terminology). Their approach is based on Miyazawa's conservation principle and the PASTA property of the arrival process.

Asmussen [1, pp. 297–298] sketches a proof of the above result using essentially a "cut and paste" argument. Choosing regenerative cycles for the unrestricted system starting at b, he divides the regenerative cycle into two parts, the first entirely below b and the second above it. Discarding the second part of the cycle immediately leads to cycles for the bounded workload system and to (1). The argument is presented in detail in section 6 where a result for the transient distribution is also given.

In this paper we extend this "cut and paste" argument to include more general random time changes and apply the technique to queues with more general restrictions where the simplicity and effectiveness of this approach become apparent. Before closing this section, we point out that the "cut and paste" idea has been used in other applications as well. For example it was used for sensitivity analysis via simulation in Ho and Li [11] and Cassandras and Strickland [4]. Glasserman and Gong [9] recently apply it to the M/GI/1/K queue to derive the proportionality result on the ergodic queue length probabilities.

2 Random time changes for regenerative systems

Random time changes have been used extensively in the analysis of stochastic differential equations (e.g. see [12]) as well as point processes [3]. Our goal is to use simple random time change arguments to obtain the steady state distribution of "complicated" queueing models in terms of the known steady state distribution of simpler models. To be more specific, let $\{X_t; t \geq 0\}$ be a real-valued process, adapted to a history $\{\mathcal{F}_t; t \geq 0\}$. Let $\{\alpha_t; t \geq 0\}$ be a real, *nonnegative* process also adapted to $\mathcal{F}_t$, representing the *inverse* of the rate at which a "random clock" is running. We will assume that $\alpha_t \leq \alpha$ for all $t \geq 0$ w.p.1 for some $\alpha > 0$. Both $\{X_t\}$ and $\{\alpha_t\}$ have *left continuous* sample paths with right limits w.p.1.

We will assume that the vector process $\{(X_t, \alpha_t); t \geq 0\}$ is regenerative in the classical sense (see Asmussen [1, p.125],) with respect to an ordinary, non-lattice renewal process $\{S_n; n = 0, 1, \ldots\}$, with $E[S_1 - S_0] \leq \infty$. Finally let $(X_\infty, \alpha_\infty)$ be a random vector distributed according to the limiting distribution of the process $\{(X_t, \alpha_t)\}$. (In the queueing examples we give subsequently, $\{X_t\}$ will be the workload process of the M/GI/1 queue.) Let

$$\varphi(t) \;=\; \int_0^t \alpha_s ds \;, \tag{2}$$

and define the random time change $\tau = \varphi^{-1}$ by means of $\tau(t) = \inf\{u : \varphi(u) > t\}$. Finally define the time-changed process $\{Y_t\}$ by means of $Y_t = X_{\tau(t)}$.

Lemma 1 *The point process $\{\varphi(S_n); n = 0, 1, \ldots\}$ is renewal with $E[\varphi(S_{n+1}) - \varphi(S_n)] < \infty$ and the time-changed process $\{Y_t\}$ defined above is regenerative with respect to it.*

Proof: Define $\Delta S_n \stackrel{\text{def}}{=} S_{n+1} - S_n$. A moment's reflection shows that the sequence of random variables

$$\varphi(S_{n+1}) - \varphi(S_n) \;=\; \int_{S_n}^{S_{n+1}} \alpha_s ds \;=\; \int_0^{\Delta S_{n+1}} \alpha_{S_n+s} ds \;, \quad n = 0, 1, \ldots,$$

is i.i.d. in view of the fact that $\{\alpha_t\}$ is regenerative. Also, $E[\varphi(S_{n+1}) - \varphi(S_n)] = E[\int_{S_n}^{S_{n+1}} \alpha_s ds] \leq \alpha E[S_{n+1} - S_n] < \infty$.

It remains to show that $\{X_{\tau(\varphi(S_n)+s)}; s \geq 0\}$ is independent of $\{\varphi(S_0), \varphi(S_1), \ldots \varphi(S_n); \{X_s; 0 \leq s \leq \varphi(S_n)\}\}$. Equivalently, it is enough to show that, for any k and any $t_i > 0$, $i = 1, 2, \ldots, k$, $(X_{\tau(\varphi(S_n)+t_1)}, \ldots, X_{\tau(\varphi(S_n)+t_k)})$ is independent of $\{\varphi(S_0), \varphi(S_1), \ldots \varphi(S_n); \{X_s; 0 \leq s \leq \varphi(S_n)\}\}$. Let $v_i = \inf\{x : \int_{S_n}^x \alpha_u du > t_i\}$. Then $\tau(\varphi(S_n) + t_i) = S_n + v_i$ and $(v_1, \ldots v_k)$ are independent of $\{\varphi(S_0), \varphi(S_1), \ldots \varphi(S_n); \{X_s; 0 \leq s \leq \varphi(S_n)\}\}$ because $\{(X_t, \alpha_t)\}$ is regenerative. Hence,

$(X_{S_n+v_1}, \ldots, X_{S_n+v_k})$ is independent of $\{S_0, S_1, \ldots, S_n; \{(X_s, \alpha_s); 0 \leq s \leq S_n\}\}$ and the proof is complete. $\square$

Theorem 1 *In addition to the above assumptions, suppose that the random variable* $\int_{S_n}^{S_{n+1}} \alpha_s ds$ *is non-lattice. Then the limiting distribution of the time-changed process* $\{Y_t\}$ *exists and is given by*

$$P(Y_\infty \leq x) \;=\; \frac{E[\int_{S_0}^{S_1} \mathbf{1}(X_s \leq x)\alpha_s ds]}{E[\varphi(S_1) - \varphi(S_0)]} \;, \tag{3}$$

or equivalently by

$$P(Y_\infty \leq x) \;=\; \frac{E[\mathbf{1}(X_\infty \leq x)\alpha_\infty]}{E[\alpha_\infty]} \;. \tag{4}$$

Proof: From lemma 1 follows that

$$P(Y_\infty \leq x) \;=\; \frac{E[\int_{\varphi(S_0)}^{\varphi(S_1)} \mathbf{1}(Y_s \leq x) ds]}{E[\varphi(S_1) - \varphi(S_0)]} \;. \tag{5}$$

Since $Y_s = X_{\tau(s)}$, the numerator in the rhs of (5) can be written as $E[\int_{\varphi(S_0)}^{\varphi(S_1)} \mathbf{1}(X_{\tau(s)} \leq x) ds]$ which with the change of variable $u = \tau(s)$ becomes $E[\int_{S_0}^{S_1} \mathbf{1}(X_u \leq x)\alpha_u du]$, and this establishes (3). (4) follows immediately from our assumption that $\{(X_t, \alpha_t)\}$ is a regenerative process with respect to the renewal process $\{S_n\}$ which implies that $E[\varphi(S_1) - \varphi(S_0)] = E[\int_{S_0}^{S_1} \alpha_s ds] = E[\alpha_\infty]E[S_1 - S_0]$, and similarly that $E[\int_{S_0}^{S_1} \mathbf{1}(X_s \leq x)\alpha_s ds] = E[\mathbf{1}(X_\infty \leq x)\alpha_\infty]E[S_1 - S_0]$. $\square$

3 The M/GI/1 queue with bounded workload process

Consider a single server queue in which the arrival epochs $\{A_n\}$ are Poisson with rate λ and the sequence of service requirements $\{\sigma_n\}$ is i.i.d. and independent of the Poisson process. We assume that the server performs work at the rate of $r(x)$ work units per time unit when the workload is equal to x ($r(0) = 0$). Though for simplicity we will focus on queues without vacations in the rest of this paper, our analysis as well as our results remain the same for queues with vacations.

Define the *unbounded* workload process $\{X_t \,;\, t \geq 0\}$ to have *left-continuous* sample paths such that w.p.1 $A_0 = 0$, $X_{A_0} = 0$, $X_{A_0+} = \sigma_0$, $\frac{d}{dt}X_t = -r(X_t)$ for $t \in (A_n, A_{n+1})$, and $X_{A_n+} = X_{A_n} + \sigma_n$. We assume that $\lambda E\sigma_1 \leq \liminf_{x\to\infty} r(x)$ which ensures the stability of the unbounded system (e.g. see Franken et al. [6, p.154]).

Define next the *bounded* workload process $\{X_t^b \,;\, t \geq 0\}$ to have *left-continuous* sample paths such that w.p.1 $A_0 = 0$, $X_{A_0}^b = 0$, $X_{A_0+}^b = \sigma_0 \wedge b$, $\frac{d}{dt}X_t^b = -r(X_t^b)$ for $t \in (A_n, A_{n+1})$, and $X_{A_n+}^b = (X_{A_n-}^b + \sigma_n) \wedge b$. (Without loss of generality we have assumed that both systems are initially empty.) Finally, let $\alpha_t = \mathbf{1}(X_t \leq b)$.

Remark. The process $\{Y_t\ ;\ t \geq 0\}$ defined as in section 2 has the same statistics as the process $\{X_t^b\ ;\ t \geq 0\}$. This can be verified easily by a "cut and paste" argument. In particular, $P(Y_\infty \leq x) = P(X_\infty^b \leq x)$. (Figure 1 illustrates the case $r(X_t) = \mathbf{1}(X_t > 0)$.)

We can easily verify that $\{(X_t, \alpha_t)\}$ is regenerative with respect to $\{S_n\}$, the arrival epochs of those jobs in the unbounded system that initiate busy periods. Also, $\alpha_t \leq 1\ \forall t$ w.p.1 and $\int_{S_n}^{S_{n+1}} \alpha_s ds$ is non-lattice. In view of our remark and Theorem 1, the steady state distribution of the workload process X_t^b is given by

$$P(X_\infty^b \leq x) \ = \ \frac{E[\mathbf{1}(X_\infty \leq x)\mathbf{1}(X_\infty \leq b)]}{E[\mathbf{1}(X_\infty \leq b)]} \ = \ \frac{F(x)}{F(b)} \ , \quad \forall\ x \in [0, b] \ .$$

An argument similar to the above gives the distribution of the workload in a system with bound b in terms of the distribution of the workload in the same system with bound $c > b$:

$$F_b(x) \ = \ \frac{F_c(x)}{F_c(b)} \ , \quad \forall\ x \in [0, b] \ . \tag{6}$$

4 M/GI/1 queues with balking

With the same notation as in section 3, consider again the *left continuous* M/GI/1 workload process (without restrictions) $\{X_t; t \geq 0\}$ and the process $\{X_t^a; t \geq 0\}$ obtained by assuming that jobs whose waiting times would be more than a do not join the system. Formally, w.p.1 $A_0 = 0$, $X_{A_0}^a = 0$, $X_{A_0+}^a = \sigma_0$, $\frac{d}{dt} X_t^a = -r(X_t^a)$ for $t \in (A_n, A_{n+1})$, and $X_{A_n+}^a = X_{A_n}^a + \sigma_n \mathbf{1}(X_{A_n} \leq a)$. In the sequel we describe a cut and paste procedure which will enable us to construct a sample path of the system with balking from the sample path of the original M/GI/1 queue.

We start by defining two sequences of stopping times as follows: $Q_1 = 0$, $V_n = \sup\{t > Q_n\ :\ X_t > a\}$, $n = 1, 2, \ldots$, $Q_n = \inf\{A_i > V_{n-1}\ :\ X_{A_i} > a\}$, $n = 2, 3, \ldots$. Define also the process $\{\alpha_t\}$ by $\alpha_0 = 1$ w.p.1 and $\alpha_t = \sum_{n=1}^\infty \mathbf{1}(Q_n < t \leq V_n)$. ($\alpha_t$ equals 0 on the intervals to be deleted and 1 on the intervals to be retained and is also left-continuous.) Define $\varphi(t)$, $\tau(t)$, and $\{Y_t; t \geq 0\}$ as in section 2. Then, it is easily verified that $\{Y_t; t \geq 0\}$ has the same statistics as $\{X_t^a; t \geq 0\}$. (Figure 2 illustrates the case $r(X_t) = \mathbf{1}(X_t > 0)$.) $\{S_n; n = 0, 1, \ldots\}$ again denotes the renewal process corresponding to the arrival epochs of those jobs in the unrestricted process $\{X_t\}$ that initiate busy periods. Clearly $\{(X_t, \alpha_t)\}$ is regenerative with respect to $\{S_n\}$ and the analysis in section 2 applies.

Let $F_a(x)$ be the steady state distribution of X_t^a, $f_a(x)$ its density, and $\overline{F}_a(x) = 1 - F_a(x)$ the corresponding survivor function. Also, let $B(x) = P(\sigma_1 \leq x)$ be the service time distribution, $m = E[\sigma_1]$ its expected value, and $\rho = \lambda m$. We will assume that the corresponding M/GI/1

queue without balking is stable and we will denote the steady state distribution of the workload by F and its density by f. When the rate of the server does not depend on the load, i.e. when $r(y) = \mathbf{1}(y > 0)$, F is given by the Pollaczek-Khintchine formula:

$$F(x) = (1-\rho)\sum_{k=0}^{\infty} \rho^k B_e^{*k}(x) \quad \text{for } x \geq 0,$$

where B_e denotes the equilibrium residual life distribution of B and B_e^{*k} the k-fold convolution of B_e with itself with the convention that B_e^{*0} is the Dirac distribution at 0.

Lemma 2 *Let $U_a(X)$ be the number of x-upcrossings of the process $\{Y_t\}$ in $[\varphi(S_0), \varphi(S_1))$. Then,*

$$U_a(x) \;=\; \sum_{\{S_0 \leq A_i < S_1\}} \mathbf{1}(Y_{A_i} \leq x)\, \mathbf{1}(Y_{A_i+} > x)\, \mathbf{1}(\alpha_{A_i} = 1)\;, \tag{7}$$

$$f_a(x) \;=\; r(x)^{-1} \frac{E[U_a(x)]}{E[\varphi(S_1) - \varphi(S_0)]}\;. \tag{8}$$

Proof: As it becomes clear from Figure 2, the time changed process $\{Y_t\}$ will have the same number of x-upcrossings in a busy period as $\{X_t\}$ if $x \leq a$. If $x > a$ then only upcrossings satisfying $X_{A_i} \leq a$ and $X_{A_i+} > x$ correspond to undeleted segments of the sample path of $\{X_t\}$ and therefore to upcrossings of $\{Y_t\}$. Note that $\{\alpha_t\}$ has left-continuous sample paths, and that $\{\alpha_{A_i} = 1\} = \{X_{A_i} \leq a\}$. Hence $\mathbf{1}(Y_{A_i} \leq x)\mathbf{1}(Y_{A_i+} > x)\mathbf{1}(\alpha_{A_i} = 1)$ equals one only for arrival epochs A_i corresponding to "undeleted" upcrossings. This establishes (7). (8) is a special case of a well-known result [13]. □

Theorem 2 *The steady state distribution of the workload process of the M/G/1 queue with balking is given by*

$$F_a(x) \;=\; F_a(0) \;+\; \int_0^x f_a(y)dy,$$

with

$$f_a(x) \;=\; \frac{\lambda r(x)^{-1}}{C} \int_{0-}^{x\wedge a} \overline{B}(x-y)F(dy)\;, \tag{9}$$

$$F_a(0) \;=\; \frac{1}{C}F(0), \tag{10}$$

where C is a normalization constant given by

$$C \;=\; 1 \;-\; \lambda \int_{y=a}^{\infty}\int_{t=0}^{\infty} [\, R(t+y) \;-\; R(y)\,]B(dt)F(dy)\;, \tag{11}$$

with $R(x) = \int_0^x r(y)^{-1}dy$. In particular when the service rate is constant, i.e. $r(y) = \mathbf{1}(y > 0)$, we have $C = 1 - \rho(1 - F(a))$.

Proof: From $\{\alpha_{A_i} = 1\} = \{X_{A_i} \le a\}$, (7), and (8) we obtain

$$f_a(x) \;=\; \frac{r(x)^{-1}}{E[\varphi(S_1) - \varphi(S_0)]}\, E[\sum_{\{S_0 \le A_i < S_1\}} \mathbf{1}(X_{A_i} \le x)\, \mathbf{1}(X_{A_i+} > x)\, \mathbf{1}(X_{A_i} \le a)\,]. \tag{12}$$

Let

$$C \;=\; \frac{E[\varphi(S_1) - \varphi(S_0)]}{E[S_1 - S_0]}\,, \tag{13}$$

and $N_{[S_0,S_1)}$ be the number of Poisson arrivals during $[S_0, S_1)$ which is also the number of jobs in the first busy period of the unrestricted system. From Wald's lemma it follows that $EN_{[S_0,S_1)} = \lambda E[S_1 - S_0]$. Multiplying and dividing the lhs of (12) with $EN_{[S_0,S_1)}$, we obtain

$$\begin{aligned} f_a(x) &= r(x)^{-1} \frac{EN_{[S_0,S_1)}}{E[\varphi(S_1) - \varphi(S_0)]}\, \frac{1}{EN_{[S_0,S_1)}} E[\sum_{\{S_0 \le A_i < S_1\}} \mathbf{1}(X_{A_i} \le x \wedge a)\, \mathbf{1}(X_{A_i+} > x)] \\ &= \frac{1}{C} \lambda r(x)^{-1} \int_{0-}^{x \wedge a} \overline{B}(x-y) F(dy)\,. \end{aligned} \tag{14}$$

To compute the size of the atom at zero, $F_a(0)$, we note that our cut-and-paste time transformation never eliminates idle periods, i.e. that $X_t = 0$ implies $a_t = 1$. Hence from (3) we have

$$F_a(0) \;=\; \frac{E[\int_{S_0}^{S_1} \mathbf{1}(X_s = 0)\alpha_s ds]}{E[\varphi(S_1) - \varphi(S_0)]} \;=\; \frac{1}{C} \frac{E[\int_{S_0}^{S_1} \mathbf{1}(X_s = 0) ds]}{E[S_1 - S_0]} \;=\; \frac{1}{C} F(0)\,. \tag{15}$$

The value of C can be obtained by the normalization relation

$$\begin{aligned} 1 - F_a(0) &= \frac{\lambda}{C} \int_0^\infty r(x)^{-1} \int_{0-}^{x \wedge a} \overline{B}(x-y) F(dy) dx \\ &= \frac{\lambda}{C} \left(\int_0^\infty r(x)^{-1} \int_{0-}^{x} \overline{B}(x-y) F(dy) dx - \int_a^\infty r(x)^{-1} \int_a^x \overline{B}(x-y) F(dy) dx \right). \end{aligned} \tag{16}$$

The steady state distribution and density of the workload process of the unrestricted M/GI/1 queue satisfy the relationship

$$f(x) \;=\; \lambda r(x)^{-1} \int_{0-}^{x} \overline{B}(x-y) F(dy)\,. \tag{17}$$

This is a consequence of Takács formula and PASTA (see Franken et al. [6, 129]). Combining (15), (16), and (17) we obtain

$$C \;=\; 1 \;-\; \lambda \int_a^\infty r(x)^{-1} \int_a^x \overline{B}(x-y) F(dy) dx\,, \tag{18}$$

or equivalently

$$C \;=\; 1 \;-\; \lambda \int_{y=a}^\infty \int_{t=0}^\infty [\, R(t+y) \;-\; R(y)\,] B(dt) F(dy)\,, \tag{19}$$

where $R(x) = \int_0^x r(y)^{-1} dy$. When $r(y) = \mathbf{1}(y > 0)$, we have $C = 1 - \rho(1 - F(a))$.

Finally, for $x < a$ (14) and (17) give

$$f_a(x) \;=\; \frac{\lambda r(x)^{-1}}{C} \int_{0-}^{x} \overline{B}(x-y) F(dy) \;=\; \frac{1}{C} f(x)\,. \qquad \square$$

5 Load-dependent arrival rates

We now extend the results in section 4 to allow arrival rates to be load-dependent in the following sense. Suppose that $g : \mathbf{R}_0^+ \to \mathbf{R}_0^+$ is a nonnegative real function with $g(x) \leq \lambda$ for all $x \in \mathbf{R}_0^+$. Assume that jobs arrive according to a Poisson process with rate λ and a job decides to join the queue with probability $p(x) = g(x)/\lambda$, independently of anything else, when x is the workload the job observes upon its arrival.

As before, let $\{X_t\}$ be the workload process of the unrestricted M/GI/1 queue. Let $\{\xi_i\}$ be an i.i.d. sequence of random variables uniformly distributed in $[0, 1]$ and independent of $\{X_t; t \geq 0\}$. Define the following sequence of stopping times: $Q_1 = \min\{A_i \ : \ p(X_{A_i}) < \xi_i\}$, $V_m = \sup\{t > Q_m \ : \ X_t = X_{Q_m}\}$, $m = 1, 2, \ldots$, $Q_m = \min\{A_i > V_{m-1} \ : \ p(X_{A_i}) > \xi_i\}$. An argument along the lines of the analysis in section 4 gives the following expression for the density of the steady state distribution of the workload process:

$$f_p(x) \ = \ \frac{\lambda r(x)^{-1}}{F(0) + \lambda \int_0^\infty r(x)^{-1} \int_{0-}^\infty p(y)\overline{B}(x-y)F(dy)dx} \ \int_{0-}^x p(y)\overline{B}(x-y)F(dy) \tag{20}$$

As a special case, consider a FCFS queue where the i'th job decides to join the queue only if its waiting time is less than ζ_i where $\{\zeta_i\}$ is an i.i.d. sequence of nonnegative random variables with distribution $P(\zeta_1 \leq x) = H(x)$. This corresponds to the above model with $p(x) = H(x)$. Balking when the *sojourn* time of the i'th job exceeds ζ_i corresponds to $p(x) = \int_0^\infty H(x + y)dB(y)$.

6 Transient behavior of the M/GI/1 queue with bounded workload

With the same notation as in section 3, let $X_0 = b$, and $X_0^b = b$. Takács [15] established the following result concerning the transient behavior of queues with bounded workload:

$$\gamma \int_0^\infty P_b(X_t^b \leq x)e^{-\gamma t}dt \ = \ \frac{\int_0^\infty P_b(X_t \leq x)e^{-\gamma t}dt}{\int_0^\infty P_b(X_t \leq b)e^{-\gamma t}dt} \ , \tag{21}$$

where $P_b(X_t \leq x) = P(X_t \leq x | X_0 = b)$ and $\gamma > 0$. A simple proof of this result can be obtained as follows. Assume that T is an exponentially distributed r.v. with rate γ, independent of the process $\{X_t\}$. Then the above relation can be written as

$$P_b(X_T^b \leq x) \ = \ \frac{P_b(X_T \leq x)}{P_b(X_T \leq b)} \ . \tag{22}$$

Let $Q_1 = \inf\{t > 0 \ : \ X_t > b\}$ and $V_1 = \inf\{t > Q_1 \ : \ X_t = b\}$, i.e. Q_1 is the first b-upcrossing and V_1 the first b-downcrossing. Since $\tilde{X}_0 = \tilde{X}_{Q_1} = b$, the processes $\{\tilde{X}_t; t \geq 0\}$ and $\{\tilde{X}_{Q_1+t}; t \geq 0\}$ have

the same statistics. By the same token, $\{X_t; t \geq 0\}$ and $\{X_{V_1+t}; t \geq 0\}$ also have the same statistics. On the other hand, both T and $(T - Q_1 | T \geq Q_1)$ have the same distribution due to the memoryless property of the exponential random variable T and they are independent of the processes $\{\tilde{X}_t; t \geq 0\}$ and $\{\tilde{X}_{Q_1+t}; t \geq 0\}$ respectively. It then follows that $\tilde{X}_T$ and $(\tilde{X}_{Q_1+(T-Q_1)} | T \geq Q_1) = (\tilde{X}_T | T \geq Q_1)$ have the same distribution, i.e., $P_b(\tilde{X}_T \leq x) = P_b(\tilde{X}_T \leq x | T \geq Q_1)$. Also note that if $T \leq Q_1$ then $\tilde{X}_T = X_T$. Therefore,

$$\begin{aligned} P_b(X_T^b \leq x) &= P_b(\tilde{X}_T \leq x) \\ &= P_b(\tilde{X}_T \leq x, T \leq Q_1) + P_b(\tilde{X}_T \leq x | T \geq Q_1) P(T \geq Q_1) \\ &= P_b(X_T \leq x, T \leq Q_1) + P_b(\tilde{X}_T \leq x) P(T \geq Q_1) \ , \end{aligned}$$

which leads to

$$P_b(X_T^b \leq x) P(T \leq Q_1) = P_b(X_T \leq x, T \leq Q_1) \ . \tag{23}$$

Similar to (23) we can show that

$$P_b(X_T \leq x) P(T \leq V_1) = P_b(X_T \leq x, T \leq V_1) = P_b(X_T \leq x, T \leq Q_1) \ , \tag{24}$$

since $P_b(X_T \leq x, Q_1 \leq T \leq V_1) = 0$. (22) follows immediately from (23) and (24).

References

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

[2] Barrer, D.V., (1957). "Queueing with impatient customers and indifferent clerks", *Opns. Res.*, **5**, 650-656.

[3] Brémaud, P., (1981). *Point Processes and Queues*, Springer-Verlag.

[4] Cassandras, C.G. and S.G. Strickland, (1989). "On-Line sensitivity analysis of Markov chains," *IEEE Trans. Autom. Control*, **34**, 76-86.

[5] Cohen, J.W., (1968). "Single server queue with uniformly bounded virtual waiting time," *J. Appl. Prob.* **5**, 93-122.

[6] Franken, P., D. Köning, U. Arndt and V. Schmidt, (1982). *Queues and Point Processes*, John Wiley, New York.

[7] Gani, J. and N.U. Prabhu (1958). "Continuous time treatment of a storage problem," *Nature*, **182**, 39-40.

[8] Ghosal, A. (1963). "Queues with finite waiting time," *Opns Res.* **11**, 919-921.

[9] Glasserman, P. and W.B. Gong (1991). "Time-changing and truncating K-capacity queues from one K to another," *J. Appl. Prob.* **28**, 647-655.

[10] Gnedenko, B.V., and I.N. Kovalenko, (1989). *Introduction to Queueing Theory*, Second Edition, Birkhauser, Boston.

[11] Ho, Y.C. and S. Li, (1988). "Extensions of infinitesimal perturbation analysis," *IEEE Transactions on Automatic Control*, **AC-33**, 427-438.

[12] McKean, H.P., Jr., (1969). *Stochastic Integrals*, Academic Press.

[13] Miyazawa, M. (1985). "The intensity conservation law for queues with randomly changed service rate," *J. Appl. Prob.* **22**, 408-418.

[14] Takács, L., (1967). "The distribution of the content of finite dams," *J. Appl. Prob.* **4**, 151-161.

[15] Takács, L., (1974). "A single server queue with limited virtual waiting time", *J. Appl. Prob.* **11**, 612-617.