Probability in the Engineering and Informational Sciences, **10**, 1996, 261–277. Printed in the U.S.A.

THE WORKLOAD IN THE M/G/1 QUEUE WITH WORK REMOVAL

RICHARD J. BOUCHERIE
Department of Econometrics
University of Amsterdam
Roetersstraat 11, 1018 WB Amsterdam, The Netherlands

ONNO J. BOXMA
CWI, P.O. Box 94079
1090 GB Amsterdam, The Netherlands
and
Tilburg University, Faculty of Economics
P.O. Box 90153
5000 LE Tilburg, The Netherlands

We consider an M/G/1 queue with the special feature of additional negative customers, who arrive according to a Poisson process. Negative customers require no service, but at their arrival a stochastic amount of work is instantaneously removed from the system. We show that the workload distribution in this M/G/1 queue with negative customers equals the waiting time distribution in a GI/G/1 queue with ordinary customers only; the effect of the negative customers is incorporated in the new arrival process.

1. INTRODUCTION

Queueing systems with negative customers were recently introduced by Gelenbe [11]. In contrast with the ordinary customers, upon arrival to a queue a negative customer removes one ordinary customer from the queue. Negative customers

The research of R. J. Boucherie has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. The research of O. J. Boxma was supported in part by the European Grant BRA-QMIPS of CEC-DG XIII.

© 1996 Cambridge University Press 0269-9648/96 $11.00 + .10

can, for example, be interpreted as work removal signals in production networks, as inhibitor signals in neural networks, and as synchronization signals in parallel computation. In the latter application, negative customers may also indicate the breakdown of a processor and the resulting destruction of work.

Since their introduction, negative customers have been studied by many authors. Product form results for the equilibrium distribution of the number of ordinary customers in networks of queues with exponential service and negative customers were, among others, derived by Gelenbe [11], Boucherie and van Dijk [2], Chao and Pinedo [6], and Henderson [17]. These results were extended to generally distributed service times in Chao [5]. Ergodicity is addressed in Gelenbe, Glynn, and Sigman [12] for the M/G/1 queue and for product form networks with exponential service in Gelenbe and Schassberger [13]. For the M/G/1 queue with negative customers, the queue length distribution is analyzed by Harrison and Pitel [16]. Their analysis of the generating function for the equilibrium queue length distribution leads to a Fredholm integral equation of the first kind, which must be solved numerically. This integral equation gives rise to intricate numerical difficulties, as its numerical solution is a so-called ill-posed problem. A more tractable M/G/1 system results when a negative customer removes all the work (hence all the customers) from the system; this model is analyzed by Jain and Sigman [18].

In the present paper, we analyze a generalization of the model of Jain and Sigman [18]: the M/G/1 queue with negative customers, in which a negative customer removes a *random amount of work* that does not necessarily correspond to an integer number of customers. Independently, the same model is studied by Jain and Sigman [19]. Their analysis uses a rate conservation approach and leads to a Pollaczek-Khintchine formula. Our analysis leads to a Wiener-Hopf equation for the Laplace-Stieltjes transform of the equilibrium distribution of the workload in the queue. The Wiener-Hopf equation can be analyzed using standard methods. In addition to the direct analysis leading to a Wiener-Hopf equation, we also show that, for the analysis of the workload, the M/G/1 queue with negative customers can be transformed into a GI/G/1 queue with only ordinary customers. The effect of the negative customers is incorporated in the new *arrival* process. This is rather surprising, because intuitively the effect of negative customers corresponds to a change in *service* effort.

The mathematical accessibility of our model, compared to that of Harrison and Pitel [16], and the resulting new insight into the latter model, represents part of the motivation for the study of the amount of *work* in the system. Furthermore, the model is related to, and of relevance for, inventory theory and risk theory (cf. Prabhu [20]). In inventory systems, an instantaneous removal of inventory occurs—for example, when perishable goods are stored; such a removal usually depends on the length of time that the goods have been in the system. In risk theory, the net gain of an insurance company increases linearly (contrary to the linearly *decreasing* workload in a queue), with downward jumps due to claims and upward jumps due to the death of a life-insurance policy-

holder. The ruin problem, which is central to risk theory, also gives rise to a Wiener–Hopf problem (cf. Cramér [9]).

The model with removal of a random amount of work according to a Poisson process is a special case of the M/G/1 queue, or dam, with as output process a process with stationary independent increments. The latter M/G/1 queue was studied by Grinstein and Rubinovitch [14]; see also Gani and Pyke [10], and and the fundamental studies of Rogozin concerning processes with independent increments (see, e.g., Rogozin [21]), which was kindly mentioned to us by A. A. Borovkov). Boxma [3] considered the generalization of the model of Grinstein and Rubinovitch [14] to the GI/G/1 case. The model presently under consideration is sufficiently simple to allow a much more detailed analysis than the preceding references.

The paper is organized as follows. In Section 2 we present the model, and in Section 3 we use level crossing arguments to derive a Wiener–Hopf equation for the Laplace–Stieltjes transform of the work in the system. In Section 4, we introduce the transformation of the M/G/1 queue with negative customers into a standard GI/G/1 queue. Via PASTA the workload at arrival epochs for this GI/G/1 queue determines the workload at arbitrary times in the M/G/1 queue. Generally distributed interarrival times are addressed in Remarks 4.1–4.3. Section 5 contains several examples, including the case in which a negative customer removes *all* the work from the queue [18] and the case in which a negative customer removes an amount of work corresponding to the service requirement of a customer [16]. The sojourn time distribution is also analyzed in Section 5.

2. THE MODEL

Consider the M/G/1 queue. Customers arrive to the queue according to a Poisson process with rate λ^+. We shall refer to these customers as ordinary or positive customers. Their service requirement has an absolutely continuous distribution $B(\cdot)$, with finite mean β, $B(0+) = 0$, and Laplace–Stieltjes transform $\beta(s) := \int_0^\infty e^{-sx} dB(x)$, which exists for $\operatorname{Re} s \geq 0$.

The server works at unit rate when customers are present. In addition to the ordinary customers, negative customers arrive to the queue with Poisson(λ^-) arrival rate. These negative customers reduce the amount of work in the system according to an absolutely continuous distribution $C(\cdot)$, with mean γ, $C(0+) = 0$, and Laplace–Stieltjes transform $\gamma(s) := \int_0^\infty e^{-sx} dC(x)$, which exists for $\operatorname{Re} s \geq 0$. Assume that $\lambda^+\beta < 1 + \lambda^-\gamma$; this is the necessary and sufficient *ergodicity condition* for the queue (cf. Lemma 4.1, later).

Let v_t denote the workload in the system at time t, that is, the sum of the remaining service time of the customer in service and the (possibly reduced) service times of the waiting customers. Clearly, v_t is independent of the service discipline, as well as of the discipline used to reduce the amount of work upon arrival of a negative customer. Let $V_t(x,v) := \mathbf{P}(v_t < x \mid v_0 = v)$, $x \geq 0$, $v \geq 0$, be the distribution of the workload in the queue at time t. If $\lambda^+\beta < 1 + \lambda^-\gamma$,

then the workload distribution $V_t(x,v)$ approaches the—unique—equilibrium version $V(x)$ for $t \to \infty$ (cf. Lemma 4.2, later).

In this paper, we establish a relation for $V(\cdot)$. To this end, we use two methods. The first method uses level crossing arguments (cf. Brill and Posner [4] and Cohen [7]) to establish a Wiener-Hopf equation for $V(\cdot)$. This Wiener-Hopf equation can be analyzed following standard methods. The second method is based on an interpretation of the arrival process of positive and negative customers. The M/G/1 queue with negative customers is transformed into an equivalent GI/G/1 queue with only positive customers. This allows us to analyze the M/G/1 queue with negative customers via standard arguments for the GI/G/1 queue, as presented in Cohen [8].

3. DERIVATION OF THE WIENER-HOPF EQUATION USING LEVEL CROSSINGS

We shall derive a Wiener-Hopf integral equation for the workload distribution, using a level crossing argument (cf. Brill and Posner [4]). Upcrossings of level x occur when a positive customer arrives at the queue. Due to the assumption of Poisson arrivals and stationarity, the long-run average rate of jumps starting at level y is $\lambda^+ dV(y)$. The proportion of jumps starting at level $y < x$ that end above level x is $1 - B(x-y)$, the probability that the required service exceeds $x - y$. Thus, the average rate of upcrossings of level x is $\lambda^+ \int_{0^-}^{x} (1 - B(x-y))dV(y)$. Similarly, the average rate of downcrossings of level x due to jumps is $\lambda^- \int_x^\infty (1 - C(y-x))dV(y)$. Let $v(x)$ denote the stationary density function of the workload ($v(x)$ can be shown to exist for $x > 0$). Then, $v(x)$ equals the average rate of downcrossings of the level x at points of continuity of the sample functions of v_t. Collecting terms, and equating the rate of up- and downcrossings, gives the following Wiener-Hopf integral equation for the workload:

$$v(x) = \int_{0^-}^{\infty} k(x-y)dV(y), \qquad x > 0, \tag{3.1}$$

with

$$k(x) := \lambda^+(1 - B(x))\mathbf{1}(x > 0) - \lambda^-(1 - C(-x))\mathbf{1}(x \le 0), \qquad -\infty < x < \infty,$$

($\mathbf{1}(\cdot)$ denoting the indicator function), and with normalization condition

$$\int_{0^-}^{\infty} dV(y) = V(0+) + \int_0^\infty v(y)dy = 1.$$

Equation (3.1) is the Wiener-Hopf integral equation for the workload. It is in the standard Wiener-Hopf form, which can be analyzed following standard methods as described in the literature. The usual approach is to extend Eq. (3.1) to $x < 0$ and to subsequently apply Fourier or Laplace-Stieltjes transforms. To this end, define

$$\varphi_+(s) := \int_{0^-}^{\infty} e^{-sx} dV(x), \qquad \text{Re } s \geq 0,$$

$$\varphi_-(s) := \int_{-\infty}^{0} e^{-sx} v_-(x) dx, \qquad \text{Re } s \leq 0,$$

where $v_-(x) := \int_{0^-}^{\infty} k(x-y) dV(y)$, $x \leq 0$. It can easily be verified that $\varphi_+(s)$ and $\varphi_-(-s)$ are bounded and analytic for $\text{Re } s > 0$ and continuous for $\text{Re } s \geq 0$. From Eq. (3.1), we obtain that $\varphi_+(s) - V(0+) = K(s)\varphi_+(s) - \varphi_-(s)$, or

$$(1 - K(s))\varphi_+(s) = V(0+) - \varphi_-(s), \qquad \text{Re } s = 0, \tag{3.2}$$

where

$$K(s) := \int_{-\infty}^{\infty} k(x) e^{-sx} dx = \lambda^+ \frac{(1 - \beta(s))}{s} - \lambda^- \frac{(1 - \gamma(-s))}{-s}, \qquad \text{Re } s = 0;$$

the equality is obtained from the definition of $k(\cdot)$.

We have to construct two functions $\varphi_+(s)$ and $\varphi_-(s)$ that satisfy Eq. (3.2) as well as the boundedness and analyticity conditions already formulated, and

$$\lim_{s \downarrow 0} \varphi_+(s) = 1, \qquad \lim_{\substack{|s| \to \infty \\ \arg s = 0}} \varphi_+(s) = V(0+), \qquad \lim_{\substack{|s| \to \infty \\ \arg s = \pi}} \varphi_-(s) = 0.$$

Thus, we have to consider the Wiener-Hopf factorization of the kernel $1 - K(s)$. This problem can be solved formally using the Wiener-Hopf theory (cf. Cohen [8, Sect. I.6.6 or II.6.3]). The idea is to factorize: $1 - K(s) = G_+(s)/G_-(s)$, where $G_+(s)$ is analytic in the right half plane and $G_-(s)$ is analytic in the left half plane. Then,

$$G_+(s)\varphi_+(s) = G_-(s)[V(0+) - \varphi_-(s)], \qquad \text{Re } s = 0, \tag{3.3}$$

where the left-hand side is analytic and bounded for $\text{Re } s > 0$ and the right-hand side is analytic and bounded for $\text{Re } s < 0$, and the sides are equal for $\text{Re } s = 0$. Thus, the left-hand side and right-hand side are analytic continuations of each other. From Liouville's theorem and the behavior of both sides at infinity, both sides are determined and, hence, so is $\varphi_+(s)$. Explicit expressions for the preceding factorization, and, hence, for $\varphi_+(s)$ and $\varphi_-(s)$, are obtained when either $\beta(s)$ or $\gamma(s)$ is a rational function of s. In Section 5.1, we work out the details for the latter case. But first we show, in Section 4, that—for the analysis of the workload—the M/G/1 queue with negative customers can be transformed into a GI/G/1 queue with only ordinary customers.

4. TRANSFORMATION INTO AN EQUIVALENT GI/G/1 QUEUE

For the M/G/1 queue Poisson arrivals see time averages. Because the arrival process of positive customers is a Poisson process that is independent of the state of the queue and of the arrival process of negative customers, this prop-

erty remains valid for the M/G/1 queue with negative customers (cf. Wolff [22, p. 294]). As a consequence, the amount of work found by a positive customer arriving to the queue equals in distribution the steady-state amount of work in the system. The amount of work found by such customers can be analyzed via a transformation of the M/G/1 queue into a GI/G/1 queue with positive customers only. This transformation is based on the observation that the influence of negative customers can be seen as a lengthening of the interarrival times for positive customers.

Let τ_n denote the interarrival time between the nth and $(n+1)$st positive customers in the original M/G/1 queue, and let σ_n be the required amount of service of positive customer n. Then, $\{\tau_n\}_n$ and $\{\sigma_n\}_n$ are independent sequences of random variables (r.v.'s), and each of these sequences consists of independent and identically distributed (i.i.d.) r.v.'s. Let w_n denote the amount of work in the system found by customer n. The amount of work, w_{n+1}, found by the next arriving positive customer, is then given by

$$w_{n+1} = \max(w_n + \sigma_n - \tau_n - d_n, 0), \tag{4.1}$$

where d_n is the amount of work destroyed by the negative customers that arrive during τ_n.

The arrival processes of the positive and negative customers are independent Poisson processes with parameters λ^+ and λ^-, respectively. Thus, the probability that exactly k negative customers arrive during an interarrival time of positive customers is $(1-p)p^k$, where $p := \lambda^-/(\lambda^+ + \lambda^-)$, $k = 0,1,2,\ldots$. Let K_n denote the number of negative customers arriving during the interarrival time τ_n. The amount of work destroyed by the jth negative customer arriving during this interarrival time is denoted by C_j^n and has distribution $C(\cdot)$. As a consequence, $\{d_n\}_n$ is an i.i.d. sequence, where d_n has the same distribution as $\sum_{j=1}^{K_n} C_j^n$. It is obvious that $\{d_n\}_n$ is independent of $\{\sigma_n\}_n$ but that τ_n and d_n are dependent.

From Eq. (4.1), we obtain that the amount of work found in the queue by an arriving positive customer in the M/G/1 queue with negative customers corresponds to the waiting time of a customer arriving in a GI/G/1 queue with required service times $\{\sigma_n\}_n$ and interarrival times $\{\tau_n^*\}_n$, where

$$\tau_n^* := \tau_n + d_n = \tau_n + \sum_{j=1}^{K_n} C_j^n. \tag{4.2}$$

The r.v.'s K_n and C_j^n are independent. Therefore, $\mathbf{E}\tau_n^* = \mathbf{E}\tau_n + \mathbf{E}K_n\mathbf{E}C_j^n = (1/\lambda^+) + (\lambda^-/\lambda^+)\gamma$. The ergodicity condition for the GI/G/1 queue is $\mathbf{E}\tau_n^* < \mathbf{E}\sigma_n$.

The following lemma formulates the equivalence and presents the ergodicity condition for the GI/G/1 queue, and therefore also for the M/G/1 queue with negative customers.

LEMMA 4.1: *The amount of work found by arriving customers in the* M/G/1 *queue with Poisson* (λ^+) *arrival rate of positive customers with service requirements* σ_n *and Poisson* (λ^-) *arrival rate of negative customers, which destroy an amount of work according to the distribution* $C(\cdot)$, *has the same distribution as the amount of work found by arrivals in the* GI/G/1 *queue with interarrival times* τ_n^* *described in Eq.* (4.2) *and service requirements* σ_n. *This queue is ergodic if and only if* $\lambda^+\beta < 1 + \lambda^-\gamma$.

Lemma 4.1 shows that the number of customers served in a busy cycle for the M/G/1 queue with negative customers is distributed as the same quantity for the GI/G/1 queue with interarrival times τ_n^* and service requirements σ_n. Cohen [8, Sect. II.5.5] analyzes the generating function of the number of customers served in a busy cycle for the GI/G/1 queue. The stability condition $\lambda^+\beta < 1 + \lambda^-\gamma$ guarantees that the number of customers served in a busy cycle is finite a.s. and has finite first moment [8]. It also follows that the length of a busy cycle for the M/G/1 queue with negative customers is finite a.s. and has finite first moment.

The following result follows easily from standard results on regenerative processes (e.g., Asmussen [1, Sect. V.1]) and the PASTA property (cf. Wolff [22, p. 294]). It justifies using the analysis of the workload found by arrivals to the GI/G/1 queue to determine the steady-state workload for the M/G/1 queue with negative customers.

LEMMA 4.2: *Under the ergodicity condition* $\lambda^+\beta < 1 + \lambda^-\gamma$ *the workload* v_t *converges in distribution to the a.s. finite r.v. that is distributed as the stationary workload found by customers arriving to the* GI/G/1 *queue with interarrival times* τ_n^* *and service requirements* σ_n.

The waiting time in the stationary GI/G/1 queue can be analyzed via the Wiener–Hopf technique, as described in Cohen [8, p. 338]. Recall that w_n is distributed as the waiting time of customer n arriving in the GI/G/1 queue with interarrival times $\{\tau_n^*\}_n$ and service times $\{\sigma_n\}_n$.

Following Cohen, we define $W(x) := \lim_{n\to\infty} \mathbf{P}(w_n < x \mid w_1 = w)$, which is independent of w, and

$$W_+(x) := W(x), \qquad W_-(x) := 0, \qquad x \geq 0,$$

$$W_+(x) := 0, \qquad W_-(x) := \int_{-\infty}^{x} W(x-u)dU(u), \qquad x < 0,$$

where $U(\cdot)$ is the distribution of $\sigma_n - \tau_n^*$. The Laplace–Stieltjes transforms

$$\omega_+(s) := \int_{0^-}^{\infty} e^{-sx} dW_+(x), \qquad \operatorname{Re} s \geq 0,$$

$$\omega_-(s) := \int_{-\infty}^{0-} e^{-sx} dW_-(x), \qquad \operatorname{Re} s \leq 0,$$

exist for $\operatorname{Re} s = 0$, and $\omega_+(s)$ and $\omega_-(-s)$ are bounded and analytic for $\operatorname{Re} s > 0$ and continuous for $\operatorname{Re} s \geq 0$. The problem is reduced to the construction of two such functions $\omega_+(s)$, $\omega_-(s)$ that satisfy

$$\omega_+(s)\{1 - \beta(s)\alpha(-s)\} = -\omega_-(s), \qquad \operatorname{Re} s = 0, \tag{4.3}$$

and

$$\lim_{s \downarrow 0} \omega_+(s) = 1, \qquad \lim_{\substack{|s| \to \infty \\ \arg s = 0}} \omega_+(s) = W_+(0), \qquad \lim_{\substack{|s| \to \infty \\ \arg s = \pi}} \omega_-(s) = 0, \tag{4.4}$$

where

$$\alpha(s) := \int_{0^-}^{\infty} e^{-st} d\mathbf{P}(\tau_n^* < t), \qquad \operatorname{Re} s \geq 0.$$

As a consequence, we have to consider the Wiener-Hopf decomposition of the kernel $1 - \beta(s)\alpha(-s)$.

For *generally* distributed interarrival times τ_n, we now proceed with the computation of $\alpha(\cdot)$. To this end, let $F(\cdot)$ be the distribution of the interarrival times, with Laplace-Stieltjes transform $\tau(\cdot)$. Then,

$$\begin{aligned}
\alpha(s) &= \sum_{k=0}^{\infty} \int_{t=0^-}^{\infty} e^{-st} \int_{x=0}^{t} dF(x) d\mathbf{P}(\tau_n^* < t, K_n = k \mid \tau_n = x) \\
&= \sum_{k=0}^{\infty} \int_{x=0^-}^{\infty} e^{-sx} dF(x) \int_{t=x}^{\infty} e^{-s(t-x)} d\mathbf{P}\left(\sum_{j=1}^{k} C_j^n < t - x \right) \mathbf{P}(K_n = k \mid \tau_n = x) \\
&= \tau(s + \lambda^-(1 - \gamma(s))), \qquad \operatorname{Re} s \geq 0.
\end{aligned} \tag{4.5}$$

For the M/G/1 queue with Poisson(λ^+) arrivals of positive customers, we have $\tau(s) = \lambda^+/(\lambda^+ + s)$. This gives

$$\alpha(s) = \frac{\lambda^+}{\lambda^+ + s + \lambda^-(1 - \gamma(s))}, \qquad \operatorname{Re} s \geq 0. \tag{4.6}$$

The explicit expression for $\alpha(s)$ allows us to obtain the Wiener-Hopf decomposition of the kernel $1 - \beta(s)\alpha(-s)$ explicitly under the assumption that either $\gamma(\cdot)$ or $\beta(\cdot)$ is rational. This is illustrated in the examples in the next section.

We end the present section with some remarks. Remarks 4.1–4.3 consider a queue with *generally* distributed interarrival times and negative customers. The first remark considers the standard GI/G/1 queue with an additional Poisson stream of negative customers. Remarks 4.2 and 4.3 consider a queue with i.i.d. interarrival times; customers are declared positive or negative upon arrival. In Remark 4.2, the effect of negative customers is incorporated in the interarrival times of positive customers, whereas in Remark 4.3 this effect is incorporated in the service times. Finally, Remark 4.4 compares the different approaches and shows that all methods lead to the same expression for the workload.

Remark 4.1: The transformation to a standard GI/G/1 queue makes use of the memoryless property of the interarrival times of *negative* customers only. As a consequence, this transformation can be extended to analyze the amount of work found by ordinary customers arriving to a GI/G/1 queue where additional negative customers arrive according to a Poisson(λ^-) process. Let $\{\tau_n\}_n$ be the interarrival times of positive customers. The r.v.'s K_n and C_j^n can be defined as before, and the r.v. d_n can be computed for given τ_n. Equations (4.2) and (4.5) remain valid. We have the following generalization of Lemma 4.1:

The amount of work found by arriving customers in the GI/G/1 *queue with interarrival times τ_n for positive customers with service requirements σ_n and Poisson (λ^-) arrival rate of negative customers, which destroy an amount of work according to the distribution $C(\cdot)$, has the same distribution as the amount of work found by arrivals (the waiting time) in the* GI/G/1 *queue with interarrival times τ_n^* described in Eq.* (4.2) *and service requirements σ_n. This queue is ergodic if and only if $\lambda^+\beta < 1 + \lambda^-\gamma$.*

The result of Lemma 4.2 cannot be extended to the GI/G/1 queue with negative customers, because PASTA does not hold for this model.

Remark 4.2: In the extension of Remark 4.1, we have used the memoryless property of the interarrival times of negative customers. The following extension does not make use of this memoryless property. Let customers arrive to the queue with i.i.d. interarrival times $\{\tau_n\}_n$. Upon arrival to the queue, a customer is declared to be positive with probability p and negative with probability $1 - p$. A positive customer has service requirement σ_n with distribution $B(\cdot)$, and a negative customer removes an amount of work with distribution $C(\cdot)$. The amount of work found by a customer that is declared positive can again be related to the amount of work found by a customer arriving to a GI/G/1 queue. To this end, let customer N be declared positive. Let K_N be the number of customers that is declared negative before a customer is declared positive. Then, $\mathbf{P}(K_N = k) = (1-p)^k p$, $k = 0,1,2,\ldots$. Let C_j^N be the amount of work that is removed by the jth negative customer; then, the next positive customer finds an amount of work $w_{N+1} = \max(w_N + \sigma_N - \tau_N^*, 0)$, where the distribution of τ_N^* has Laplace-Stieltjes transform $(1-p)\tau(s)/[1 - p\tau(s)\gamma(s)]$ (in the case of a Poisson($\lambda^+ + \lambda^-$) arrival process, this coincides with formula (4.6)). The analysis proceeds as before, but the result of Lemma 4.2 cannot be obtained here.

Remark 4.3: The model of Remark 4.2 can be slightly reformulated, by considering just *one* type of customer, the nth customer having service requirement $\hat{\sigma}_n$. With probability p, customer n has a positive service requirement with distribution $B(\cdot)$, and with probability $1 - p$ a negative service requirement with distribution $C(\cdot)$. Clearly, $\hat{\beta}(s) := \mathbf{E}[e^{-s\hat{\sigma}_n}] = p\beta(s) + (1-p)\gamma(-s)$, $\operatorname{Re} s = 0$. Let $\{\tau_n\}_n$ be the i.i.d. interarrival times, and let $\tau(s)$ denote the Laplace-Stieltjes transform of their distribution. The workload w_{n+1} found by customer $n + 1$

is then given by $w_{n+1} = \max(w_n + \hat{o}_n - \tau_n, 0)$. Now follow the derivation of the Laplace-Stieltjes transform for the waiting time in the GI/G/1 queue (cf. Cohen [8]; there service times are positive). Use the identity $e^{-s[x]^+} - e^{-sx} = 1 - e^{-s[x]^-}$, with $[x]^+ := \max(0,x)$ and $[x]^- := \min(0,x)$. Substituting $x = w_n + \hat{o}_n - \tau_n$ and taking expectations, we find in the stationary case:

$$\mathbf{E}[e^{-sw_n}][1 - \hat{\beta}(s)\tau(-s)] = 1 - \mathbf{E}[e^{-s[w_n+\hat{o}_n-\tau_n]^-}], \qquad \operatorname{Re} s = 0. \tag{4.7}$$

Hence, we have to consider the Wiener-Hopf factorization of

$$1 - \hat{\beta}(s)\tau(-s) = 1 - \tau(-s)\{p\beta(s) + (1-p)\gamma(-s)\}, \qquad \operatorname{Re} s = 0. \tag{4.8}$$

Remark 4.4: Let us now restrict ourselves to the case of exponentially distributed interarrival times and relate the kernel in Eq. (4.8) to the kernels appearing in Eqs. (3.2) and (4.3). We have $\tau(s) = (\lambda^+ + \lambda^-)/(\lambda^+ + \lambda^- + s)$, and $p = \lambda^+/(\lambda^+ + \lambda^-)$. This gives, with $\operatorname{Re} s = 0$,

$$1 - \hat{\beta}(s)\tau(-s) = \frac{\lambda^+(1-\beta(s)) + \lambda^-(1-\gamma(-s)) - s}{\lambda^+ + \lambda^- - s} = -\frac{s(1-K(s))}{\lambda^+ + \lambda^- - s}.$$

Now note that a factorization $1 - K(s) = G_+(s)/G_-(s)$, as briefly discussed at the end of Section 3, immediately yields the factorization

$$1 - \hat{\beta}(s)\tau(-s) = \frac{sG_+(s)}{G_-(s)(s - \lambda^+ - \lambda^-)},$$

and hence, see Eq. (4.7),

$$sG_+(s)\mathbf{E}[e^{-sw_n}] = G_-(s)(s - \lambda^+ - \lambda^-)(1 - \mathbf{E}[e^{-s[w_n+\hat{o}_n-\tau_n]^-}]).$$

We have essentially the same Wiener-Hopf problem as at the end of Section 3; application of Liouville's theorem will confirm the equality of $\mathbf{E}[e^{-sw_n}]$ and $\varphi_+(s)$. Hence, the workload at arrival epochs for the model of Remark 4.3 (with positive and negative service requirements) has the same equilibrium distribution as the workload in the model of Section 3.

Similarly, the kernel $1 - K(s)$ appearing in Eq. (3.2) can be related to the kernel $1 - \beta(s)\alpha(-s)$ that appears in Eq. (4.3) for the GI/G/1 model with extended interarrival times:

$$1 - \beta(s)\alpha(-s) = \frac{s(1-K(s))}{s - \lambda^+ - \lambda^-(1-\gamma(-s))} = -\frac{s(1-K(s))\alpha(-s)}{\lambda^+}. \tag{4.9}$$

This leads to the factorization

$$sG_+(s)\omega_+(s) = \lambda^+ \frac{G_-(s)}{\alpha(-s)}\omega_-(s), \qquad \operatorname{Re} s = 0.$$

Note that $\alpha(-s)$, as given in Eq. (4.6), cannot become zero for finite s, $\operatorname{Re} s < 0$, because $\gamma(s)$ is the Laplace-Stieltjes transform of a probability distribution. Solving the Wiener-Hopf problem then confirms what has already been stated

in Lemma 4.2: $\varphi_+(s) = \omega_+(s)$, $\operatorname{Re} s \geq 0$, and thus the equilibrium distribution $V(\cdot)$ of the workload analyzed in Section 3 equals the equilibrium distribution $W(\cdot)$ of the waiting time of a customer in the equivalent GI/G/1 queue.

5. EXAMPLES

The above-indicated equivalence among $\varphi_+(s)$, $\omega_+(s)$, and $\mathbf{E}[e^{-sW_n}]$ allows us to arbitrarily consider any one of them. Let us concentrate for the moment on $\omega_+(s)$, hence on the GI/G/1 queue with extended interarrival times. For rational $\beta(\cdot)$, the analysis follows the standard lines for the G/K$_n$/1 queue as presented in Cohen [8, Sect. II.5.10]. In this section, we present several examples for which the zeros of the kernel of Eq. (4.3) can be found. It is shown in Section 5.1 that for work removal ("killing") distributions with rational Laplace–Stieltjes transform the analysis of the M/G/1 queue with negative customers follows the lines of the K$_m$/G/1 queue. Section 5.2 studies systems with complete breakdowns that remove all the work from the system. Here the results of Jain and Sigman [18] are reproduced. In Section 5.3, removal of customer equivalents is discussed. It is shown that the choice $\gamma(\cdot) = \beta(\cdot)$ (killing = service) generally does *not* lead to a good correspondence between the model with work removal and the model with customer removal. Section 5.4 studies the time it takes before a certain amount of work is removed from the system by service and killing.

5.1. Killing Distributions with Rational Transform

Rationality of $\gamma(s)$ implies rationality of $\alpha(s)$ (cf. Eq. (4.6)). Suppose that $\gamma(s) = \gamma_1(s)/\gamma_2(s)$, where $\gamma_2(\cdot)$ is a polynomial of degree m, and $\gamma_1(\cdot)$ a polynomial of degree less than m, such that $\gamma_1(\cdot)$ and $\gamma_2(\cdot)$ have no common zeros; normalize $\gamma_2(s)$ such that $\gamma_2(0) = 1$. Then,

$$\alpha(s) = \frac{\lambda^+ \gamma_2(s)}{(\lambda^+ + s)\gamma_2(s) + \lambda^-(\gamma_2(s) - \gamma_1(s))}, \qquad \operatorname{Re} s \geq 0.$$

Thus, $\alpha(s) = \alpha_1(s)/\alpha_2(s)$, where $\alpha_1(\cdot)$ is a polynomial of degree m and $\alpha_2(\cdot)$ is a polynomial of degree $m + 1$. $\alpha(-s)$ is analytic for $\operatorname{Re} s < 0$. This implies that $\alpha_2(-s)$ has no zeros for $\operatorname{Re} s < 0$, that is, $\alpha_2(-s)$ has $m + 1$ zeros, counted according to their multiplicity, for $\operatorname{Re} s \geq 0$. The kernel can be written as

$$1 - \beta(s)\alpha(-s) = \frac{1}{\alpha_2(-s)}\,[\alpha_2(-s) - \beta(s)\alpha_1(-s)], \qquad \operatorname{Re} s = 0.$$

Apply Rouché's theorem (Cohen [8, App. 6]) to this kernel, on the closed contour Ω consisting of the line from $-\epsilon - i\infty$ to $-\epsilon + i\infty$ and the closing semicircle in the right half plane, for ϵ small enough. It then follows that, when the ergodicity condition holds, $\alpha_2(-s) - \beta(s)\alpha_1(-s)$ has exactly m zeros δ_i, $i = 1, \ldots, m$,

with $\operatorname{Re}\delta_i > 0$, and one zero $\delta_0 = 0$. This implies (Cohen [8, pp. 329–330]) that Eq. (4.3) can be written as

$$\omega_+(s)\frac{[\alpha_2(-s)-\beta(s)\alpha_1(-s)]}{s\prod_{i=1}^{m}(\delta_i - s)} = -\omega_-(s)\frac{\alpha_2(-s)}{s\prod_{i=1}^{m}(\delta_i - s)}, \qquad \operatorname{Re} s = 0,$$

where the left-hand side is analytic and bounded for $\operatorname{Re} s > 0$ and the right-hand side is analytic and bounded for $\operatorname{Re} s < 0$, and the sides are equal for $\operatorname{Re} s = 0$. Thus, the left-hand side and right-hand side are analytic continuations of each other. From Liouville's theorem and the bounded behavior of both sides at infinity (cf. also Eq. (4.4)), it follows that both sides are equal to the same constant c. The constant c is determined by taking the limit $s \downarrow 0$. This gives

$$\omega_+(s) = \frac{\lambda^+\beta - 1 - \lambda^-\gamma}{\alpha_2(-s) - \beta(s)\alpha_1(-s)}\, s\prod_{i=1}^{m}\frac{\delta_i - s}{\delta_i}, \qquad \operatorname{Re} s \geq 0.$$

The case of exponential killing is a special case of the preceding result. We now have $\gamma(s) = 1/(1+\gamma s)$. The kernel has one zero $\delta_0 = 0$ and one zero δ_1 in the right half plane. If, in addition, we assume that the service requirements are exponentially distributed, then we can compute the value of δ_1: it is the unique positive root (use the ergodicity condition) of

$$\gamma\beta\delta_1^2 + (\gamma - \beta - \lambda^+\gamma\beta - \lambda^-\gamma\beta)\delta_1 + (-1 - \lambda^-\gamma + \beta\lambda^+) = 0. \tag{5.1}$$

In this example,

$$\omega_+(s) = (1+\beta s)\frac{\zeta}{\zeta + s}, \qquad \operatorname{Re} s \geq 0,$$

where $\zeta = (1 + \lambda^-\gamma - \lambda^+\beta)/(\beta\gamma\delta_1)$ is minus the (real) negative root of Eq. (5.1). Note that the Laplace-Stieltjes transform of the workload, $\varphi_+(s)$, equals $\omega_+(s)$. Hence,

$$V(x) * B(x) = 1 - e^{-\zeta x}, \qquad x \geq 0. \tag{5.2}$$

5.2. System Breakdown

In this example, we assume that a negative customer removes all work in the system upon arrival. A negative customer can be seen as a complete breakdown of the system removing all work or as a reset of the system. In addition, it is possible for the breakdown to add a subsequent repair period during which all arriving positive customers are lost. This will not affect the analysis.

For this example, we have that $C(x) = 0$, $x < \infty$, with a point mass at infinity; hence, $\gamma = \infty$. Following the analysis presented in Section 4, we obtain that the queue is ergodic for all values of $\lambda^+\beta$ if $\lambda^- > 0$. The Laplace-Stieltjes trans-

form of $C(\cdot)$ is $\gamma(s) = 0$ for all $s < \infty$. Insertion of this expression into Eq. (4.6) gives

$$\alpha(s) = \lambda^+/(\lambda^+ + \lambda^- + s), \qquad \operatorname{Re} s \geq 0,$$

which corresponds to the expression obtained for rational $\gamma(\cdot)$ with $\gamma_1(\cdot) \equiv 0$, $\gamma_2(\cdot) \equiv 1$. The analysis for rational $\gamma(\cdot)$ of the previous subsection cannot be followed further.

The kernel now reads

$$1 - \beta(s)\alpha(-s) = \frac{\lambda^+ + \lambda^- - s - \beta(s)\lambda^+}{\lambda^+ + \lambda^- - s}, \qquad \operatorname{Re} s = 0.$$

The numerator of this expression has one zero δ with $\operatorname{Re}\delta > 0$; the denominator has one zero. Formula (4.3) can be written as

$$\omega_+(s)\frac{\lambda^+ + \lambda^- - s - \beta(s)\lambda^+}{s - \delta} = -\omega_-(s)\frac{\lambda^+ + \lambda^- - s}{s - \delta}, \qquad \operatorname{Re} s = 0.$$

From Liouville's theorem and the bounded behavior of both sides at infinity, it follows that both sides are equal to the same constant c; we thus obtain that

$$\omega_+(s) = c\frac{s - \delta}{\lambda^+ + \lambda^- - s - \beta(s)\lambda^+}, \qquad \operatorname{Re} s \geq 0, \tag{5.3}$$

where c can be determined from Eq. (4.4). Taking $s = 0$ gives $c = -\lambda^-/\delta$. Taking the limit $s \to \infty$, we obtain that $\delta W(0) = \lambda^-$.

The model with complete breakdown corresponds to the system with disasters analyzed in Jain and Sigman [18]. In particular, Eq. (5.3) is obtained via a rate conservation approach in Proposition 1 of that reference.

5.3. Removal of Customer Equivalents

A direct relation between the results for work removal as presented above, and customer removal as presented in Harrison and Pitel [16], seems to be rather involved when generally distributed service requirements are allowed. For some special cases, a direct comparison is possible.

An easy example for which both models can be made equivalent is given by the M/D/1 queue with negative customers removing *customers* at the end of the queue. In the model of the present paper, we can set $C(\cdot) = B(\cdot)$. However, this example seems to be the only example of such simplicity. The problem with comparing the two models already becomes apparent when we compare the stability conditions. Harrison and Pitel [16] present the stability conditions for their model with several disciplines for customer removals. These—for some disciplines quite involved—conditions do not resemble our simple stability condition, except in special cases. One such case is the M/M/1 queue with negative customers and removal of the customer in service.

The equilibrium queue length distribution in the model with customer-in-service removal is $\pi(n) = (1 - \rho)\rho^n$, where $\rho = \lambda^+\beta/(1 + \lambda^-\beta)$, as can easily be seen when the effect of negative customers is incorporated in the service requirements: the resulting effectively obtained service time is exponentially distributed with mean $\beta/(1 + \lambda^-\beta)$. The stability condition is $\rho < 1$, that is, $\lambda^+\beta < 1 + \lambda^-\beta$. Let V_c denote the workload in the model with customer-in-service removal; then,

$$\mathbf{E}V_c = \frac{\lambda^+\beta^2}{1 + \lambda^-\beta - \lambda^+\beta}. \tag{5.4}$$

For the M/M/1 queue with work removal, we have shown in Section 4 that the amount of work in the system is distributed as the amount of work found by customers arriving to a GI/M/1 queue. From Cohen [8, p. 230] and Lemma 4.2, we obtain that

$$\mathbf{E}V = \frac{\nu\beta}{1 - \nu}, \tag{5.5}$$

where $\nu = 1 - \zeta\beta$, and ζ is minus the unique negative root of Eq. (5.1). Comparison of the average workload $\mathbf{E}V$ for the model with work removal, and $\mathbf{E}V_c$ for the model with customer-in-service removal shows that equality occurs iff $\nu = \rho$. Formula (5.1) shows that $\nu = \rho$ iff

$$\gamma = (1 + \lambda^-\beta)/\lambda^+ = \beta/\rho. \tag{5.6}$$

This value of γ is such that the average number of negative customers needed to remove one positive customer equals 1. Thus, for the M/M/1 queue, by a choice of γ that is somewhat larger than the average service requirement β we obtain that on the average a negative customer removes one positive customer from the queue, and $\mathbf{E}V_c = \mathbf{E}V$. Observe that $\gamma > \beta$ is not required for the general model, as can be seen from the M/D/1 queue. This suggests that a general rule for the comparison of the model with customer removal and the model with work removal might be difficult.

5.4. The Sojourn Time

To study the sojourn time of a customer in our model with FIFO service and work removal from the head of the queue, define $\mathbf{T}(x)$ to be the time it takes before an amount of work x has disappeared from the system. Obviously, $\mathbf{P}(\mathbf{T}(x) > x) = 0$, whereas

$$\mathbf{P}(\mathbf{T}(x) = x) = e^{-\lambda^- x}, \qquad x \geq 0;$$

this corresponds to no negative arrival during x. When there *are* negative arrivals the amount of work x will disappear in less than x time units. Conditioning on the number of negative arrivals, we can write

$$\mathbf{P}(\mathbf{T}(x) > z) = \sum_{j=0}^{\infty} e^{-\lambda^- z} \frac{(\lambda^- z)^j}{j!} C^{j*}(x - z), \qquad 0 \le z < x. \tag{5.7}$$

Hence,

$$\mathbf{ET}(x) = \int_{z=0}^{x} \mathbf{P}(\mathbf{T}(x) > z)dz = \int_{z=0}^{x} \sum_{j=0}^{\infty} e^{-\lambda^- z} \frac{(\lambda^- z)^j}{j!} C^{j*}(x - z)dz.$$

The Laplace-Stieltjes transform of $\mathbf{ET}(x)$ readily follows from this relation:

$$\psi(s) := \int_{x=0}^{\infty} e^{-sx} d[\mathbf{ET}(x)] = \frac{1}{s + \lambda^- - \lambda^- \gamma(s)}, \qquad \operatorname{Re} s > 0. \tag{5.8}$$

In the case of exponential killing, so $\gamma(s) = 1/(1 + s\gamma)$, we have

$$\psi(s) = \frac{s + 1/\gamma}{s(s + \lambda^- + 1/\gamma)}, \qquad \operatorname{Re} s > 0,$$

and hence

$$\mathbf{ET}(x) = \frac{x}{1 + \lambda^- \gamma} + \frac{\lambda^- \gamma^2}{(1 + \lambda^- \gamma)^2} (1 - e^{-(\lambda^- + 1/\gamma)x}), \qquad x \ge 0. \tag{5.9}$$

For large x, the factor $x/(1 + \lambda^- \gamma)$ dominates. For a general killing distribution the application of a Tauber theorem for Laplace-Stieltjes transforms (cf. Cohen [8, App. 4]; note that $\mathbf{ET}(x)$ is nondecreasing in x) to Eq. (5.8) also shows that, for $x \to \infty$, $\mathbf{ET}(x) \approx x/(1 + \lambda^- \gamma)$. The obvious interpretation is that on the average $\lambda^- \gamma$ work is killed per unit of time, when there is a large enough amount of work present.

Note that the study of the decrement of the workload from x to 0 amounts to the study of the increment, from 0 to x, of a process with stationary independent increments. The nice structure of that process in our model leads to relatively simple expressions. It also enables us to analyze the sojourn time $\mathbf{T}$ of an arbitrary customer K, for the case that work is removed from the *head* of the queue. Because of the PASTA property, the amount of work found by K upon its arrival has the distribution of the steady-state workload distribution $V(\cdot)$. $\mathbf{T}$ equals the time it takes before that workload, plus K's service request, has disappeared. So

$$\mathbf{E}[e^{-s\mathbf{T}}] = \int_{z=0}^{\infty} e^{-sz} \int_{x=z}^{\infty} d(V * B)(x) d_z \mathbf{P}(\mathbf{T}(x) < z), \qquad \operatorname{Re} s \ge 0; \tag{5.10}$$

partial integration and Eq. (5.7) lead for $\operatorname{Re} s \ge 0$ to

$$\mathbf{E}[e^{-s\mathbf{T}}] = 1 - s \int_{x=0}^{\infty} d(V * B)(x) \int_{z=0}^{x} e^{-sz} \sum_{j=0}^{\infty} e^{-\lambda^- z} \frac{(\lambda^- z)^j}{j!} C^{j*}(x - z)dz. \tag{5.11}$$

We now restrict ourselves to the case of exponential service and work removal distributions. It follows from Eq. (5.2) that $V(x)*B(x) = 1 - \exp[-\zeta x], x \geq 0$. After some arithmetic, we find

$$\mathbf{E}[e^{-s\mathbf{T}}] = \frac{\zeta[1 + \gamma\zeta + \gamma\lambda^-]}{s[1 + \gamma\zeta] + \zeta[1 + \gamma\zeta + \gamma\lambda^-]}, \qquad \text{Re}\, s \geq 0. \tag{5.12}$$

Apparently, for this case in which all four interarrival, service, and work removal distributions are memoryless, the sojourn time $\mathbf{T}$ is also exponentially distributed; whereas $V(\cdot)*B(\cdot)$ has mean $1/\zeta$, the mean sojourn time equals

$$\mathbf{ET} = \frac{1}{\zeta}\,\frac{1 + \gamma\zeta}{1 + \gamma\zeta + \gamma\lambda^-}. \tag{5.13}$$

For the case of an exponential killing distribution but general service time distribution, it follows from Eq. (5.9) that the mean sojourn time equals

$$\int_{x=0}^{\infty} \mathbf{ET}(x)\,d(V*B)(x)$$

$$= \frac{\mathbf{E}V + \beta}{1 + \lambda^-\gamma} + \frac{\lambda^-\gamma^2}{(1 + \lambda^-\gamma)^2}\,[1 - \omega_+(\lambda^- + 1/\gamma)\beta(\lambda^- + 1/\gamma)].$$

Acknowledgments

The authors wish to thank F. Baccelli for suggesting to use a transformation to incorporate the effect of negative customers, J. W. Cohen for valuable discussions, and K. Sigman for sending an unpublished version of the report by Jain and Sigman [18].

References

1. Asmussen, S. (1987). *Applied probability and queues.* New York: Wiley.
2. Boucherie, R.J. & van Dijk, N.M. (1994). Local balance in queueing networks with positive and negative customers. *Annals of Operations Research* 48: 463–492.
3. Boxma, O.J. (1975). The single-server queue with random service output. *Journal of Applied Probability* 12: 763–778.
4. Brill, P.H. & Posner, M.J.M. (1977). Level crossings in point processes applied to queues: Single-server case. *Operations Research* 25: 662–674.
5. Chao, X. (1995). Networks of queues with customers, signals and arbitrary service. *Operations Research* 43: 537–544.
6. Chao, X. & Pinedo, M. (1993). On generalized networks of queues with positive and negative arrivals. *Probability in the Engineering and Informational Sciences* 7: 301–334.
7. Cohen, J.W. (1977). On up- and downcrossings. *Journal of Applied Probability* 14: 405–410.
8. Cohen, J.W. (1982). *The single server queue.* Amsterdam: North-Holland.
9. Cramér, H. (1955). *Collective risk theory.* Reprinted from the Jubilee Volume of Skandia Insurance Company. Stockholm: Esselte.
10. Gani, J. & Pyke, R. (1960). The content of a dam as the supremum of an infinitely divisible process. *Journal of Mathematics and Mechanics* 9: 639–651.
11. Gelenbe, E. (1991). Product-form queueing networks with negative and positive customers. *Journal of Applied Probability* 28: 656–663.

12. Gelenbe, E., Glynn, P., & Sigman, K. (1991). Queues with negative arrivals. *Journal of Applied Probability* 28: 245–250.
13. Gelenbe, E. & Schassberger, R. (1992). Stability of product form G-networks. *Probability in the Engineering and Informational Sciences* 6: 271–276.
14. Grinstein, J. & Rubinovitch, M. (1974). Queues with random service output: The case of Poisson arrivals. *Journal of Applied Probability* 11: 771–784.
15. Harrison, P.G. & Pitel, E. (1993). Sojourn times in single-server queues with negative customers. *Journal of Applied Probability* 30: 943–963.
16. Harrison, P.G. & Pitel, E. (1994). The M/G/1 queue with negative customers. Research Report, Imperial College, London (to appear in *Journal of Applied Probability*).
17. Henderson, W. (1993). Queueing networks with negative customers and negative queue lengths. *Journal of Applied Probability* 30: 931–942.
18. Jain, G. & Sigman, K. (1994). A Pollaczek-Khintchine formulation for M/G/1 queues with disasters. Research Report, Columbia University, New York, July (to appear in *Journal of Applied Probability*).
19. Jain, G. & Sigman, K. (1995). A generalization of Pollaczek-Khintchine formula to account for negative arrivals. Research Report, Columbia University, New York, August.
20. Prabhu, N.U. (1980). *Stochastic storage processes*. New York: Springer-Verlag.
21. Rogozin, B.A. (1966). On the distribution of functionals related to boundary problems for processes with independent increments. *Theory of Probability and Its Applications* 11: 580–591.
22. Wolff, R.W. (1989). *Stochastic modeling and the theory of queues*. Englewood Cliffs, NJ: Prentice-Hall.