Martingale methods for analysing single–server queues

M. Roughan*† and C.E.M. Pearce‡

May 14, 2002

Abstract

In this paper we presents a martingale method for analysing queues of $M/G/1$ type, which have been generalised so that the system passes through a series of phases on which the service behaviour may differ. The analysis uses the process embedded at departures to create a martingale, which makes possible the calculation of the probability generating function of the stationary occupancy distribution. Specific examples are given, for instance a model of an unreliable queueing system, and an example of a queue–length–threshold overload–control system.

keywords: $M/G/1$ queue, martingale, queueing, phase, overload control

1 Introduction

The $M/G/1$ queue is one of the fundamental models of queueing theory. Many computing, communication and manufacturing systems can be modelled by $M/G/1$ models (see, for example, Cohen [6], Cooper [7], Kleinrock [16]). The $M/G/1$ queue has numerous variants tailored to particular applications. A number of methodologies have be presented for analysing these, such as the highly successful matrix–analytic methodology of Neuts [24].

In this paper we develop an alternative method which uses martingale and renewal methods to find solutions for a general class of $M/G/1$–type queues. Such an approach was originally applied to queueing problems for which the solutions were already known (see Baccelli and Makowski [2]–[5] and Park [26]). In Roughan [30] and Roughan and Pearce [31, 32] the method was extended to a queueing system with history–dependent behaviour. This paper details the extension of this theory to a general class of systems. Also presented are examples of queueing systems to which this model can be applied.

The class of systems consists of those where a busy period can be decomposed into a series of phases[1]. This paper extends the types of transitions between phases – previous papers all considered fixed thresholds based on queue length. Some new alternatives are based on the length of the current busy period, or other random variables. The behaviour of the server can be different in each phase; for instance, the service-time distribution or the service discipline may change between phases, or the arrival rate to the system could change between phases.

The systems encompassed include cases in which the service times or arrival rates are state–dependent, as in Dshalalow [10], Gong *et al.* [12], Ibe and Trivedi [13], Kijima and Makimoto [14, 15], LaMaire [17], Morrison [21], Schormans *et al.* [34] and T. Takine *et al.* [36]. The method also extends to other models where the system's behaviour can depend on the history of the system. Standard methods of analysing such sytems introduce a supplementary variable to incorporate the historical information in the current state. We do not need to do this, potentially reducing the complexity of the required computations dramatically.

*M. Roughan is at AT&T Labs – Research, 180 Park Av. Florham Park, NJ 07932 USA. Email: *roughan@research.att.com.*

†The majority of this work was done while M. Roughan was a PhD student at the University of Adelaide, supported by an Australian Priority Postgraduate Award and a scholarship from the Teletraffic Research Centre in the Department of Applied Mathematics.

‡C.E.M. Pearce is in the Department of Applied Mathematics, University of Adelaide, Adelaide 5005, AUSTRALIA. Email: *cpearce@maths.adelaide.edu.au*

[1]This usage is distinct from that of phase–type distributions used by Neuts [24] and others.

The method described here uses Doob's Optional Stopping Theorem on a martingale formed from the process embedded at service completion times. This theorem allows one to relate the behaviour of the queue at stopping times, the ends of busy periods in Baccelli and Makowski's methodology, to the behaviour of a renewal process embedded at these stopping times.

Our method allows other time points, the ends of phases, to have significance in the queueing process, even though these epochs may not form a standard renewal process. Standard renewal theory has been extended [27] to provide analogous results for the case where the renewal process has structure within each renewal period.

The paper is organised as follows. Section 2 presents the precise set of models covered and defines the stopping times to be used, the martingale from which we derive our results and the condition for regularity of the stopping times. The section also defines the multi–phase renewal process and presents the results that we use. The main results are presented in Section 3, in particular Theorem 3.5 of Subsection 3.2. Section 4 discusses extensions to the theory, including how it might be applied to different types of server and arrival–process behaviour. Section 5 considers some examples, including simple previously known cases to verify the method, and some more complex cases including a model of an overload control. One of the results shown is that the solutions to these examples all take a general form which looks like the solution to the $M/G/1$ queue with the complexities of the phases structure bundled into one correction term. Section 6 lists some other results (besides the probability generating function of the occupancy distribution) which may be derived from the martingale approach. Some concluding remarks are made in Section 7. There are also two appendices, the first of which supplies proofs of the conditions for regularity of the stopping times and the second proofs of results used with the examples.

2 Preliminaries

The *multi–phase M/G/1 queue* is a variant of the $M/G/1$ queueing process in which each service time is a non–negative random variable with probability distribution function chosen from $\{A^j(\cdot)\}_{j=1}^N$. An interval during which $A^j(\cdot)$ operates is called *phase* j.

We assume that the phase changes occur at the ends of services and are stopping times with respect to the filtration generated by the queueing process, that is, a decision to change phase at time T is based only on information about the system up to time T and not on any information about the future behaviour of the queue. We assume also that times spent in two phases are independent if the two phases are not in the same busy period. These limitations are necessary for the analysis to follow and are realistic for many systems.

The motivating case is that in which the phases occur in some specific pre–defined order. We shall label the phases in the order in which they occur and call one transition through all of the phases a *cycle*.

It will also be convenient to consider each cycle of transitions between phases to correspond to a busy period. That is, we start the cycle in phase 1 when an arriving customer finds the server unoccupied, and the end of the busy period corresponds to the end of phase N. If the cycle is not complete by the time the system is again empty we deem the process to have spent zero time in the remaining phases. Thus each phase is entered exactly once during each busy period. This and the fact that the times spent in two phases in different busy periods are independent mean that the ends of busy periods are still renewal points of the process.

As in a standard analysis of the $M/G/1$ queue, we consider the embedded, discrete–time process formed by observing the number of customers in the system after departures. However, unlike the standard $M/G/1$ system, this embedded process does not form a Markov chain without the additional complexity of a supplementary variable to describe the current phase. We follow the approach of Baccelli and Makowski [3, 4] in defining a martingale with respect to the embedded process, and from this establish a relationship between the forward recurrence times in a multi–phase Markov Renewal Process (MRP) and the system size – the fundamental relationship of this paper, Theorem 3.2. From an analysis of the multi–phase MRP (as in [27]), the limiting probability generating function for the distribution of the number of customers in the system may be expressed in terms of the state sojourn times of the multi–phase MRP. Helpful results for the calculation of these probability generating functions are found in Corollaries 3.3 and 3.4 using the martingale once again. A general form for the probability generating function corresponding to stationary occupancy can then be found. This is expressed in Theorem 3.5.

Note that in this section the exact nature of the transitions between phases is not specified. We provide the constraints on what types of transitions are allowed but say nothing about the actual way in which the transitions are governed. The theory allows a quite general approach to these transitions. In Sections 5.2-5.5 we consider a number of possible cases. For example in Section 5.2 we consider a process in which the transition occurs when some threshold is crossed. This threshold could be a physical limit on the queue size or a limit on the number of customers served after the busy period begins. We shall often refer to the points in the process at which transitions occur as a threshold, though the scope of what is meant by a threshold is quite large.

2.1 The formal model

The formal model used here is a generalisation of the $M/G/1$ process incorporating phases. Arrivals form a Poisson process with rate λ. The service–time distribution depends on the phase of the system at the beginning of the service, the distribution function $A^j(\cdot)$ applying when the phase is $j \in \mathcal{S} := \{1, 2, \ldots, N\}$

If A_n^j denotes the number of arrivals during the n–th service time, given the service starts with the system in phase j, the sequences $(A_n^j)_{n\geq 0}$ $(j \in \mathcal{S})$ then constitute N independent sequences, each of which is formed from independentally and identically distributed random variables. We may define a filtration $(\mathcal{F}_n)$ by $\mathcal{F}_n = \sigma(A_m^j \,|\, 0 \leq m \leq n,\ j \in \mathcal{S})$ and put $\mathcal{F} := \bigcup_{n=1}^{\infty} \mathcal{F}_n$.

Changes of phase obey the following rules.

1. The phase can change only on service completion.
2. The times at which phase changes occur are stopping times with respect to $(\mathcal{F}_n)$.
3. The times spent in two phases in different busy periods are independent.
4. The busy period starts in phase 1.
5. The phases occur in order, so phase i is followed by phase $i+1$ (mod N).
6. Each phase is entered exactly once during a busy period.

The time from the beginning of phase 1 to the subsequent end of phase N is referred to as a *cycle*. A cycle thus corresponds to a busy period. We say $\phi(n) = j$ if, after the nth departure, the system is in phase j.

Denote by (X_n) the discrete–time embedded process, where X_n is the number of customers seen by the nth departing customer, who leaves at time t_n, say. We assume that $X_0 = 0$.

When the i–th transition out of phase j occurs at the epoch t_n in a realisation ω of the process, we write $T_j^i(\omega) = n$. With a slight abuse of terminology, we may also in that situation speak of the epoch T_j^i as being at time n (in the embedded process). For convenience we set $T_0^i = T_N^{i-1}$. We define

$$C_n^j := \bigcup_{i \in I\!N} \{\omega \,|\, T_{j-1}^i(\omega) \leq n < T_j^i(\omega)\},$$

so that C_n^j is the event that at time t_n the system is in phase j. In the usual indicator notation

$$I(C_n^j) = \begin{cases} 1 & \phi(n) = j, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the above does not preclude zero time being spent in a phase, as T_{j-1}^i might equal T_j^i. In this case we still say that T_{j-1}^i 'occurs before' T_j^i.

We define $a_i^j = P(A_1^j = i)$ and form the probability generating functions

$$a_j(z) = E\left[z^{A_1^j}\right] = \sum_{i=0}^{\infty} a_i^j \, z^i = A^{j*}(\lambda - \lambda z)$$

in terms of the Laplace–Stieltjes transform A^{j*} of A^j. The *traffic intensity* during phase j is defined as $\rho_j = a_j'(1)$, the mean number of arrivals during a single service in phase j. We have $\rho_j = \lambda/\mu_j$, where $1/\mu_j$ is the mean service time during phase j. It is convenient to introduce $\xi_j(z) = z/a_j(z)$.

In a given realisation ω, X_n is determined by A^j_m and $I(C^j_m)$ at times $m \le n$. Since the ends of phases are at stopping times, $\{T^i_j \le n\} \in \mathcal{F}_n$ for $j \in \mathcal{S}$ and $i \in I\!N$ and so $C^j_n = \bigcup_{i \in I\!N} \{T^i_{j-1} \le n < T^i_j\} \in \mathcal{F}_n$ and $I(C^j_n)$ is $\mathcal{F}_n$–measurable. Thus X_n is $\mathcal{F}_n$-measurable and indeed for $m \le n$, X_m, $I(C^j_m)$ and A^j_m are all $\mathcal{F}_n$-measurable. We can define $N+2$ sequences of stopping times $(\tau_i(n))$ $(0 \le i \le N)$ and $(\tau(n))$ for $n \in \mathbb{Z}^+$ by

$$\begin{aligned} \tau(n) &= \begin{cases} \inf\{m > n \mid X_m = 0\}, & \text{when the set concerned is non–empty,} \\ \infty, & \text{otherwise,} \end{cases} && (1) \\ \tau_j(n) &= \tau(n) \wedge \inf\{m \ge n \mid \phi(m) > j\} \\ &= \tau(n) \wedge \inf\left\{ m \ge n \,\middle|\, \sum_{i=j+1}^{N} I(C^i_m) = 1 \right\}. && (2) \end{aligned}$$

When $j = N$, the sum is empty and so $\tau_N(n) = \tau(n)$. Thus

$$n = \tau_0(n) \le \tau_1(n) \le \cdots \le \tau_i(n) \le \cdots \le \tau_N(n) = \tau(n) \ a.s.$$

The random variable $\tau(n)$ thus represents the epoch of the end of the busy period current at time t_n. We have $\tau_j(n) = n$ if the process has already been through phase j in the current busy period; otherwise it is the time of the next transition out of phase j. When the busy period ends we deem the process to pass through any remaining phases, spending zero time in each. That $\tau(n)$ and $\tau_j(n)$ are stopping times comes directly from the fact that phase transitions are allowed only at stopping times.

We can define also the times

$$\begin{aligned} \nu_j(n) &= \tau_j(n) - \tau_{j-1}(n), \\ \mu_j(n) &= \begin{cases} \nu_j(n), & X_n \neq 0, \\ 0, & X_n = 0, \end{cases} \end{aligned} \tag{3}$$

for $j \in \mathcal{S}$. We can allow a dummy service completion at time zero, with X_0 taken to be some random variable Ξ. For our purposes it is convenient to take $X_0 = 0$ a.s. and correspondingly $\phi(0) = 1$. Hence $T^0_0 = 0$ and

$$\mu_j(\tau_{l-1}(0)) = \begin{cases} \nu_j(0), & j \ge l, \\ 0, & j < l. \end{cases}$$

We now introduce the martingale which will provide the majority of our results.

Theorem 2.1 *For each $z \in (0,1]$ define $M_0(z) = 1$ and*

$$M_n(z) = z^{X_n} \prod_{k=0}^{n-1} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C^j_k) a_j(z)} \right), \quad n \ge 1.$$

Then $\big(M_n(z)\big)$ is a non–negative integrable martingale with respect to $(\mathcal{F}_n)$.

Proof: From the $\mathcal{F}_n$-measurability of C^j_k and $I(X_k \neq 0)$ for $0 \le k \le n$ we have

$$\begin{aligned} E\left[M_{n+1}(z) \mid \mathcal{F}_n\right] &= E\left[z^{X_{n+1}} \prod_{k=0}^{n} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C^j_k) a_j(z)} \right) \middle| \mathcal{F}_n \right] \\ &= \left\{ \prod_{k=0}^{n} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C^j_k) a_j(z)} \right) \right\} E\left[z^{X_{n+1}} \middle| \mathcal{F}_n \right] \ a.s. \end{aligned}$$

A comparison of the numbers of customers left by successive departures leads to the recurrence relation

$$X_{n+1} = X_n - I(X_n \neq 0) + \sum_{j=1}^{N} I(C_n^j) A_{n+1}^j,$$

whence

$$E\left[M_{n+1}(z) \middle| \mathcal{F}_n\right] = \left\{ \prod_{k=0}^{n} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C_k^j) a_j(z)} \right) \right\} z^{X_n - I(X_n \neq 0)} E\left[z^{\sum_{j=1}^{N} I(C_n^j) A_{n+1}^j} \middle| \mathcal{F}_n \right] \quad a.s.$$

Since C_n^j $(j \in \mathcal{S})$ is a disjoint, complete set, we derive

$$E\left[z^{\sum_{j=1}^{N} I(C_n^j) A_{n+1}^j} \middle| \mathcal{F}_n \right] = \sum_{i=1}^{N} I(C_n^j) a_j(z) \quad a.s.$$

and so

$$E\left[M_{n+1}(z) \middle| \mathcal{F}_n\right] = z^{X_n} \prod_{k=0}^{n-1} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C_k^j) a_j(z)} \right) = M_n(z) \quad a.s.$$

The non–negativity of $M_n(z)$ is immediate, so that $E\left[|M_n(z)|\right] = E\left[M_n(z)\right] < \infty$ and the martingale is integrable. □

2.2 Stability and regularity

For applications to equilibrium results the processes involved must be stable and recurrent. Stability is usually easy to demonstrate for $M/G/1$ processes, though care must be taken here to avoid traps. However, stability is not sufficient for the results herein. We also require *regularity* of stopping times with respect to the martingale above. A sufficient condition for stability and regularity turns out to be given by the following, as we shall establish in Appendix A.

Definition. Let $\mathcal{S}^* \subset \mathcal{S}$ be the set of j with $\rho_j > 1$. We say Condition $(*)$ holds if

$$E\left[\prod_{j \in \mathcal{S}^*} \xi_j(z)^{\tau_j(0) - \tau_{j-1}(0)} \right] < \infty$$

for all $z \in [0, 1]$.

To apply Doob's Optional Sampling Theorem, we must demonstrate that the stopping times involved are regular for the martingale. This is trivial when the martingale is uniformly integrable (see Neveu, IV-**3**-14), but this need not always occur. This difficulty is addressed by Theorem 2.2 below, which is proved in Appendix A. In the theorem the stopping time γ refers to any of the stopping times defined in the previous subsection (for instance $\tau_i(n)$ for some i and n), and also to the stopping times $\tau_0(\gamma), \ldots, \tau_N(\gamma)$ and $\tau(\gamma)$ which are defined by extension of (1) and (2) to n a random variable in place of a constant.

Theorem 2.2 *If $(*)$ is satisfied, then the stopping times $\tau_0(\gamma), \ldots, \tau_N(\gamma)$ and $\tau(\gamma)$ are regular for the martingale $(M_n(z))$, $z \in [0, 1]$. Furthermore when $\tau(\gamma) = \infty$, we have $M_{\tau(\gamma)}(z) = 0$ a.s.*

When $\rho_j < 1$ for all j the system is stable and $\mathcal{S}^* = \phi$, so that Condition $(*)$ is automatically satisfied. Also, when $\mathcal{S}^* = \{i\}$, the condition becomes

$$E\left[\alpha^{\tau_i(0) - \tau_{i-1}(0)} \right] < \infty,$$

where $\alpha = \sup_{z \in [0,1]} \xi_i(z)$.

2.3 Multi–phase renewal processes

Renewal processes and MRP (Pyke [28, 29]) have been used heavily in stochastic modelling. Nakagawa and Osaki [22] proposed a new type of renewal process in which not every state transition occurs at a renewal point. Such processes are covered by a general result of Marlow and Tortorella [20], which considers a reliability process in which the operating and repair times are not necessarily independent. One example proposed by Nakagawa and Osaki (Type 1–MRP) is a modified MRP with N states entered sequentially of which $N-1$ are non–regeneration states. We refer to this process as a *Multi–Phase Renewal Process* (MPRP).

We assume here that the renewal process is lattice and aperiodic.

Define $S_l^m \in \mathbb{Z}^+$ to be the epoch of the mth entry into state l. We call the period $[S_0^m, S_N^m)$ the mth cycle. We define the state sojourn lifetimes i_j of the mth cycle by

$$i_j = S_j^m - S_{j-1}^m \qquad \text{for } 1 \le j \le N,\ \ m \ge 1.$$

Here we identify the mth entry into state N with the $(m+1)$th entry into state 0, that is, $S_N^m = S_0^{m+1}$.

We identify the times S and T in the MPRP and the multi–phase $M/G/1$ queue respectively. That is, we associate the times at which phase changes occur in the queue with the times in the MPRP. To this end we allow inter-dependence between state sojourn lifetimes within a cycle but assume independence between cycles. This is essential to the idea that the entry epochs of the states 1 to $N-1$ are non–regeneration points while the entry epochs of state 0 are regeneration points. We denote by $i_1, i_2, \ldots, i_N$ the state sojourn lifetimes and take $i_1 + i_2 + \cdots + i_N \ge 1$ with probability 1, and further define $m = E\,(i_1 + \cdots + i_N)$.

If the process is in the mth cycle and state k at time n, the residual sojourn time r_j for state j is defined as

$$r_j \;=\; \begin{cases} 0 & j \le k \\ S_j^m - n & j = k+1 \\ S_j^m - S_{j-1}^m & j > k+1 \end{cases}$$

for $1 \le j \le N$. Note that for $j > k+1$ the residual sojourn time is just the state sojourn lifetime, that is, $i_j = r_j$. We define (when they exist) the probability generating functions

$$Q^*(x_1, x_2, \ldots, x_N) \;=\; E\left[\prod_{k=1}^{N} x_k^{r_k}\right], \tag{4}$$

$$F_l^*(x_{l+1}, x_{l+2}, \ldots, x_N) \;=\; E\left[\prod_{k=l+1}^{N} x_k^{i_k}\right] \tag{5}$$

for $0 \le l < N$ and $F_N^* = 1$. Thus Q^* is the probability generating function for the residual sojourn times and F_l^* that of the final $N-l$ state sojourn lifetimes. Note that

$$F_l^*(x_{l+1}, x_{l+2}, \ldots, x_N) = F_0^*(1, \ldots, 1, x_{l+1}, \ldots, x_N).$$

Theorem 2.3 *If $x_1, x_2, \ldots, x_N \in [0,1]$, then*

$$Q^*(x_1, x_2, \ldots, x_N) = \frac{1}{m} \sum_{l=1}^{N} \frac{F_l^*(x_{l+1}, \ldots, x_N) - F_{l-1}^*(x_l, x_{l+1}, \ldots, x_N)}{1 - x_l}. \tag{6}$$

Further, if the series defined by Q^* converges for some $x_1, x_2, \ldots, x_N$ not all in the interval $[0,1]$, then it converges to the right–hand side of (6).

Proof: See [27, Theorem 2]. □

3 Results

3.1 Optional stopping

We now move on to the main results of this paper which are derived using the Optional Stopping Theorem.

Theorem 3.1 *Suppose $z \in [0,1)$ and $(*)$ is satisfied. Then*

$$E\left[\prod_{j=1}^{N} \xi_j(z)^{\nu_j(\gamma)} \middle| \mathcal{F}_\gamma\right] \quad = \quad z^{X_\gamma} z^{I(X_\gamma=0)} \quad a.s. \tag{7}$$

Proof: The stopping times γ and $\tau(\gamma)$ satisfy $\gamma \leq \tau(\gamma)$ a.s. and under Condition (*) Theorem 2.2 demonstrates the regularity of these stopping times. Hence we may apply Doob's Optional Sampling Theorem to obtain $E\left[M_{\tau(\gamma)}(z)\middle|\mathcal{F}_\gamma\right] = M_\gamma(z)$ a.s., which may be rewritten

$$E\left[z^{X_{\tau(\gamma)}} \prod_{k=0}^{\tau(\gamma)-1} \frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C_k^j) a_j(z)} \middle| \mathcal{F}_\gamma\right] = z^{X_\gamma} \prod_{k=0}^{\gamma-1} \frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C_k^j) a_j(z)} \quad a.s. \tag{8}$$

The product on the right–hand side of (8) is $\mathcal{F}_\gamma$-measurable and may therefore be cancelled from both sides. Further, if $\tau(\gamma) = \infty$, then Theorem 2.2 implies that $M_{\tau(\gamma)}(z) = 0$ and so this case makes no contribution to the expectation above. If $\tau(\gamma) < \infty$ then $X_{\tau(\gamma)} = 0$ and therefore (8) reduces to

$$E\left[\prod_{k=\gamma}^{\tau(\gamma)-1} \frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{N} I(C_k^j) a_j(z)} \middle| \mathcal{F}_\gamma\right] \quad = \quad z^{X_\gamma} \quad a.s.$$

By definition $X_k \neq 0$ for $k = \gamma + 1$ to $\tau(\gamma) - 1$, so the left–hand side becomes

$$E\left[\frac{z^{I(X_\gamma \neq 0)}}{\sum_{j=1}^{N} I(C_\gamma^j) a_j(z)} \prod_{k=\gamma+1}^{\tau(\gamma)-1} \frac{z}{\sum_{j=1}^{N} I(C_k^j) a_j(z)} \middle| \mathcal{F}_\gamma\right].$$

As $z^{I(X_\gamma=0)}$ is also $\mathcal{F}_\gamma$–measurable, we may rearrange to obtain

$$E\left[\prod_{k=\gamma}^{\tau(\gamma)-1} \frac{z}{\sum_{j=1}^{N} I(C_k^j) a_j(z)} \middle| \mathcal{F}_\gamma\right] \quad = \quad z^{I(X_\gamma=0)} z^{X_\gamma} \quad a.s.$$

Since

$$I(C_k^j) = \begin{cases} 1, & \tau_{i-1}(\gamma) \leq k < \tau_i(\gamma), \\ 0, & \text{otherwise,} \end{cases}$$

we may make a further simplification to

$$E\left[\prod_{j=1}^{N} \left(\frac{z}{a_j(z)}\right)^{\tau_j(\gamma)-\tau_{j-1}(\gamma)} \middle| \mathcal{F}_\gamma\right] \quad = \quad z^{I(X_\gamma=0)} z^{X_\gamma} \quad a.s.$$

Substitution of the definitions of $\nu_j(\gamma)$ and $\xi_j(z)$ gives the required result. □

Theorem 3.2 *Suppose $z \in [0,1)$ and Condition $(*)$ is satisfied. Then*

$$E\left[z^{X_\gamma}\right] \quad = \quad E\left[\prod_{j=1}^{N} \xi_j(z)^{\mu_j(\gamma)}\right].$$

Proof: Multiplying (7) by $I(X_\gamma \neq 0)$ and noting that $I(X_\gamma \neq 0)$ is $\mathcal{F}_\gamma$-measurable leads to

$$I(X_\gamma \neq 0) z^{X_\gamma} \quad = \quad E\left[I(X_\gamma \neq 0) \prod_{j=1}^{N} \xi_j(z)^{\nu_j(\gamma)} \middle| \mathcal{F}_\gamma \right] \quad a.s.$$

Taking expectations yields

$$E\left[I(X_\gamma \neq 0) z^{X_\gamma} \right] \quad = \quad E\left[I(X_\gamma \neq 0) \prod_{j=1}^{N} \xi_j(z)^{\nu_j(\gamma)} \right].$$

Since $X_\gamma \neq 0$ implies $\mu_j(\gamma) = \nu_j(\gamma)$ for $j \in \mathcal{S}$, it follows that

$$E\left[z^{X_\gamma} \right] \quad = \quad P(X_\gamma = 0) + E\left[I(X_\gamma \neq 0) \prod_{j=1}^{N} \xi_j(z)^{\mu_j(\gamma)} \right].$$

The events $[X_\gamma = 0]$ and $[\mu_j(0) = 0,\ j \in \mathcal{S}]$ are equivalent by (3) and so

$$[X_\gamma \neq 0] = [\mu_j(0) \neq 0 \text{ for some } j],$$

which implies

$$E\left[z^{X_\gamma} \right] = P(\mu_j(n) = 0,\ \forall j) + E\left[I\{\exists j:\ \mu_j(n) \neq 0\} \prod_{j=1}^{N} \xi_j(z)^{\mu_j(\gamma)} \right],$$

whence the desired result $\square$

Corollary 3.3 *If $z \in [0,1)$, Condition $(*)$ is satisfied and $l > 1$, then*

$$E\left[\prod_{j=l}^{N} \xi_j(z)^{\nu_j(0)} \right] \quad = \quad E\left[z^{X_{\tau_{l-1}(0)}} \right].$$

Proof: This follows by setting $\gamma = \tau_{l-1}(0)$ in the preceding theorem and noting that the assumption $X_0 = 0$ entails

$$\mu_j(\tau_{l-1}(0)) = \begin{cases} \nu_j(0), & j \geq l, \\ 0, & j < l. \end{cases}$$

$\square$

Remark: The case $l = 1$ is excluded, since when $l = 1$, we have $\tau_{l-1}(0) = \tau_0(0) = 0$ and $\mu_j(0) = 0$ because $X_0 = 0$. Thus the result would not hold. In its place we have the following.

Corollary 3.4 *If $z \in [0,1)$ and Condition $(*)$ is satisfied, then*

$$E\left[\prod_{j=1}^{N} \xi_j(z)^{\nu_j(0)} \right] \quad = \quad z.$$

Proof: Putting $\gamma = 0$ in Theorem 3.1 and taking expectations gives

$$E\left[\prod_{j=1}^{N} \xi_j(z)^{\nu_j(0)} \right] \quad = \quad E\left[z^{X_0} z^{I\{X_0 = 0\}} \right],$$

which, since we assume $X_0 = 0$, yields the stated result. $\square$

3.2 Relationship with the MPRP

The connection between the MPRP and the multi–phase $M/G/1$ queueing process is now clear. Each phase of the queueing process is associated with a state in the MPRP. Transitions between states in the MPRP occur at the same times as changes in phase in the queueing process. Because phases change only at the end of services we use the embedded process and consequently the MPRP has discrete or lattice time. We must resort to a generalised MPRP because there is no requirement that the time spent in each phase be independent of the times spent in each of the other phases. However, we do assume that the times when the system is left empty are renewal epochs. This means the times spent in phases during different busy periods must be independent. Because we assume the queue begins with a dummy service leaving it empty ($X_0 = 0$), we have a non–delayed renewal process.

Given this description we see that the $\mu_j(n)$ are forward recurrence times in the MPRP and the $\nu_j(0)$ sojourn lifetimes in the MPRP. Therefore

$$\begin{aligned} Q^*(\xi_1(z), \xi_2(z), \ldots, \xi_N(z)) &= \lim_{n\to\infty} E\left[\prod_{j=1}^{N} \xi_j(z)^{\mu_j(n)}\right], \\ F_l^*(\xi_l(z), \xi_{l+2}(z), \ldots, \xi_N(z)) &= E\left[\prod_{j=l}^{N} \xi_j(z)^{\nu_j(0)}\right], \end{aligned}$$

where Q^* and F_l^* are defined in (4) and (5), respectively. Thus Theorems 3.2 and 3.4 and Corollary 3.3 imply respectively that

$$\begin{aligned} Q^*(\xi_1(z), \xi_2(z), \ldots, \xi_N(z)) &= \lim_{n\to\infty} E\left[z^{X_n}\right], & (9)\\ F_0^*(\xi_1(z), \xi_2(z), \ldots, \xi_N(z)) &= z, & (10)\\ F_l^*(\xi_{l+1}(z), \ldots, \xi_N(z)) &= E\left[z^{X_{\tau_l(0)}}\right]. & (11) \end{aligned}$$

From these we deduce the following.

Theorem 3.5 *For $z \in [0,1)$ and Condition $(*)$ satisfied,*

$$E\left[z^X\right] = \frac{1}{m}\left[\sum_{l=2}^{N} \frac{E\left[z^{X_{\tau_l(0)}}\right] - E\left[z^{X_{\tau_{l-1}(0)}}\right]}{1-\xi_l(z)} + \frac{E\left[z^{X_{\tau_1(0)}}\right] - z}{1-\xi_1(z)}\right],$$

where m acts as a normalising constant and $X(t) \to X$ a.s. as $t \to \infty$.

Proof: Substituting (9) in Theorem 2.3 gives

$$\lim_{n\to\infty} E\left[z^{X_n}\right] = \frac{1}{m}\sum_{l=0}^{N-1}\left[\frac{F_{l+1}^*(\xi_{l+2}(z), \ldots, \xi_N(z)) - F_l^*(\xi_{l+1}(z), \ldots, \xi_N(z))}{1-\xi_l(z)}\right].$$

Under steady–state conditions arrivals, the distributions of the number of customers seen by arrivals and left by departures are the same (see [7, pp. 154–5]). In the event of Poisson arrivals, this common distribution agrees with that of the number present at a randomly chosen instant (the 'PASTA' principle, see [37]). Hence by the dominated convergence theorem,

$$\lim_{n\to\infty} E\left[z^{X_n}\right] = \lim_{t\to\infty} E\left[z^{X(t)}\right].$$

Substitution of (10) and (11) into the above then gives the required result. □

4 Extensions

4.1 Server behaviour

The above results may be extended *via* generalisations of server behaviour during a phase. We note several examples.
(i) Waiting for service: During phase j a customer waits a random period of time before beginning service. If the extra time has probability distribution function $B^j(\cdot)$, then

$$a_j(z) = A^{j*}(\lambda(1-z))B^{j*}(\lambda(1-z)).$$

This type of behaviour is a simple example of a different service–time distribution, but is included in order to show how the method described here may be applied to standard models such as a queue with generalised vacations.
(ii) $\mathbf{M}_j$-policy queue: During phase j a customer waits until M_j arrivals have occurred before commencing service. In this case

$$a_j(z) = z^{M_j} A^{j*}(\lambda(1-z)).$$

This might occur if the server has some overhead associated with starting and stopping service. To reduce the overhead the busy period must be increased in length, which can be done by increasing the first service time as above, giving what Neuts [24] refers to as the N–policy queue.
(iii) Other service discplines: The server discipline has not been mentioned in the previous discussion, as the results apply to all non–preemptive server disciplines. In fact the service discipline may even change from phase to phase through non–preemptive server disciplines.

4.2 Arrival process behaviour

The above examples extend the server behaviour during a phase. The arrival process may also be modified between phases. The only limitation here is that the PASTA results may no longer apply and therefore we may, in some cases, obtain only the probability distribution relating to system occupancy as seen by arriving customers. In many cases this may be of more interest. Following are three examples of processes with modifications to the arrival process during different phases.
(i) Partial blocking: During phase j arrivals are blocked with probability p_j. The arrivals are Poisson with new rate λp_j and so

$$a_j(z) = A^{j*}(\lambda p_j(1-z)).$$

This situation could occur as part of a control strategy. For example if the arrival process is the superposition of several independent Poisson streams with component rates λ_i, it will itself be a Poisson stream, with rate $\sum \lambda_i$. If these streams are then assigned different priorities and are blocked on the basis of their priority during certain phases j, the congestion of the higher priority arrivals may be controlled.
(ii) Batch arrival processes: The Poisson arrivals may occur in batches whose size is given by a random variable B. Put $b_i := P(B = i)$. Jun Huek Park [26] extends Baccelli and Makowski's work to the $M^B/G/1$ queue, showing the probability generating function of the number of arrivals during a service to be

$$a_j(z) = A^{j*}(\lambda[1 - B(z)]),$$

where $B(z)$ is the generating function for the batch size distribution. Park's extension may be applied immediately here.
(iii) Limited–$\mathbf{N}_j$ services: During phase j only the first N_j arrivals during any particular service are allowed to join the queue, any further arrivals being blocked. Here

$$a_j(z) = \sum_{i=0}^{N_j - 1} a_i^j \left(z^i - z^{N_j}\right) + z^{N_j},$$

where a_i^j is the probability of i arrivals occurring during a service time in phase j. Note that the arrival process is no longer independent of the service times, but if $N_j \neq 1$ the arrivals (and departures) still see the distributions found above.

4.3 Generalisations of the phases

We have imposed quite specific limitations on the completion times of phases. To contemplate possible generalisations of the phase structure we must first consider the restrictions on phase structure prescribed in Section 2.1. Rule 1 is a result of treating the embedded process. The restriction is to ensure that the behaviour of the embedded process characterises that of the queueing process. Rule 2 guarantees that Doob's Optional Sampling Theorem can be applied at the relevant time points. Rules 3 and 4 ensure that the ends of busy periods are renewal points of the process. Finally Rules 5 and 6 are required so that the structure of the renewal process we consider is that of an MPRP.

Rules 1 and 2 are therefore crucial to the whole martingale approach. Rules 3 and 4 are requirements for the Markov renewal results.

Rules 5 and 6 are required to enforce the multi-phase nature of the renewal process. If we could obtain a result equivalent to Theorem 2.3 for generalised Markov renewal processes not of multi–phase form, these two rules could be modified or removed. However they are not as restrictive as they might seem. For instance,

(a) if a particular phase is skipped over, we may insert a transition through the missing phase which takes zero time;

(b) if a phase is visited more than once during a cycle, we may consider second and subsequent entries to that phase to be into new phases.

4.4 Infinitely many phases

We now consider reasons for allowing an infinite number of phases. Some cases with more complex structure require conceptual modifications to fit the form of sequential phases, each occurring once during a busy period. For example, one or more phases may not necessarily occur during every busy period. As noted in (a) above, we may deal with this by inserting transitions of zero duration through the missed phases. An infinite number of phases can easily be dealt with in this case so long as with probability one the original process has only a finite number of phases of positive length occurring during a busy period.

Likwise under scenario (b) two phases may alternate arbitrarily often during a busy period, resulting in there being an infinite number of phases. In Section 5.5 we consider such an example. The two phases alternate only finitely often with probability one, so recurrence of the process is not affected.

With this procedure $a_j(z)$ (and so also $\xi_j(z)$) will be the same for a number of the new phases, which results in considerable simplifications.

In all of these cases we simply generalise our results by replacing N with infinity. Condition $(*)$ ensures continued validity of the results.

4.5 Blocking versus zero service time

In some cases we may choose $\rho_j = 0$. This implies that $\lambda \int_0^\infty (1 - A^j(t))dt = 0$, which may occur only if service times are instantaneous or the arrival rate is zero. We refer to the former as discarding a customer, while the latter occurs if all the arriving customers are blocked. Both have the consequence $a_j(z) = 1$ and so lead to the same stationary distribution. An apparent difference is that customers who are discarded enter service for zero time, and therefore wait in the queue until they receive service, while those who are blocked never enter the system and therefore do not contribute to the number of customers in the queue.

To resolve this apparent paradox we need to note that because the traffic intensity can change only at the end of a service time, the blocking of customers begins only at the end of the service time and therefore in terms of queue length seen by arriving customers is the same as the case of discarding customers. As the arrival rate is changing, PASTA no longer applies, so that the time–averaged behaviour of the system is no longer necessarily the same as that seen by arrivals.

5 Examples

5.1 A single–phase example

This section provides a number of examples to demonstrate the power of the results above. First we consider the simplest example of this type, the $M/G/1$ queue. This is one of the problems to which this technique was originally applied by Baccelli and Makowski [4] and the results here differ from theirs only in the notation.

The queue–length distribution in the ergodic $M/G/1$ queue is well–known and can be derived in various ways (see Cooper [7]). It is given by

$$E\left[z^X\right] = (1-\rho)\frac{a(z)(1-z)}{a(z)-z},$$

where $a(z)$ is the probability generating function for the distribution of the number of arrivals during a service. In our terminology this is a single-phase $M/G/1$ queue and the result is obtained directly from Theorem 3.5 with $N = 1$ as

$$E\left[z^X\right] = \frac{1}{m}\left[\frac{1-z}{1-\xi_1(z)}\right] = \frac{1}{m}\left[\frac{a_1(z)(1-z)}{a_1(z)-z}\right],$$

where $m = 1/(1-\rho_1)$. Theorem 3.4 gives

$$F^*(\xi_1(z)) = z,$$

which may then be used both to calculate m and to provide the probability generating function for the distribution of the number of customers served during a busy period (Baccelli and Makowski [4]). Namely, for each $\xi \in [0,1)$, the equation $z = \xi a(z)$ has a unique solution $z = Z(\xi)$ in the interval $[0,1]$. From (2.14) in Baccelli and Makowski [4]

$$F^*(y) = Z(y).$$

5.2 Two-phase examples

The next simplest case has two phases. Each occurs exactly once during a busy period and always in the same order. The server starts a busy period with service–time distribution $A^1(\cdot)$ and switches to distribution $A^2(\cdot)$ when some 'threshold' is reached. It switches back to A^1 when the system again becomes empty.

For notational convenience, and to match the notation of [30], we write A and B for A^1 and A^2 with the corresponding changes in notation:

$$\begin{aligned}
A(t) &= A^1(t), & B(t) &= A^2(t),\\
A_n &= A^1_n, & B_n &= A^2_n,\\
a(z) &= a_1(z), & b(z) &= a_2(z),\\
\xi_a(z) &= \xi_1(z), & \xi_b(z) &= \xi_2(z),\\
a_i &= a^1_i, & b_i &= a^2_i,\\
\rho_a &= \rho_1, & \rho_b &= \rho_2.
\end{aligned} \tag{12}$$

We consider four examples of thresholds corresponding to different stopping times at which the switch from A to B occurs.

(i) We speak of a *fixed upward threshold* if the phase change occurs at the first time a departing customer leaves more than k customers behind. This example is treated *in extenso* in [30] so we shall not address it in full detail here. To avoid triviality in this case we assume that $\rho_a > 0$.

(ii) We refer to a *geometrically–distributed random–time threshold* when the phase change occurs after a random number of customers, given by a geometric distribution with parameter p, has been served in the busy period.

(iii) A *fixed-time threshold* occurs when the phase change occurs after a set number S of customers is served during the busy period.

(iv) A *general random-time threshold* occurs after a random number of customers, determined by a probability distribution $h(\cdot)$, have been served.

A broad practical motivation for studying such modifications of the $M/G/1$ system arises from various questions of control. Thus if the number of customers in the system becomes too large (and so waiting times too long) it may be desired to switch to a faster service rate to remove the excess customers. In this case the queue is cleared before returning to the slower service rate. A variant in which the faster service continues only until enough customers are removed from the system is considered in Section 5.5. For an example of such stochastic control over a queue, see Dshalalow [10]. The novelty of the present model is that the phase can depend on the history of the process, not merely the present state.

The fixed–time threshold corresponds to a cruder type of control. For some related ideas see Neuts [24] and the E–limited service discipline of LaMaire [17].

A diametrically–opposite motivation presents itself for random–time thresholds. Here the threshold can correspond to a switch to slower service representing some partial breakdown of the service facility which can be repaired quickly but only when the system is free of customers. This example is further extended in Section 5.4.

Theorem 5.1 *The following results hold for the relevant thresholds described above.*

(i) For a fixed upward threshold, Condition $(*)$ *holds for all* $k \in I\!N$ *if* $\rho_b \leq 1$.

(ii) For a geometrically–distributed random–time threshold, Condition $(*)$ *holds for* $p \in (1 - 1/\alpha, 1)$ *and* $\rho_b \leq 1$, *where* $\alpha = \sup_{z\in[0,1)} \xi_a(z)$.

(iii) For a fixed–time threshold, Condition $(*)$ *holds for all* $S \in I\!N$ *if* $\rho_b \leq 1$.

(iv) For a general random–time threshold, Condition $(*)$ *holds if* $\rho_b \leq 1$ *and there exists an* N *such that* $h(n) = 0$ *for all* $n > N$, *though this is not a necessary condition.*

Proof: Since $\rho_b < 1$, the condition is trivial if $\rho_a \leq 1$, while if $\rho_a > 1$ it reduces to

$$E\left[\alpha^{\tau_1(0)}\right] < \infty.$$

The cases are dealt with separately:

(i) in Lemma 3.2 in [30];

(ii) in Lemma 3 in Section B.1;

(iii) in Section B.2;

(iv) the given condition is sufficient because it implies that the random–time threshold, and hence $\tau_1(0)$, is bounded above by N. Thus Condition $(*)$ is satisified. □

For the rest of this section we shall make use of the $k \times k$ matrix ($k \in I\!N$) defined by

$$\mathbf{P}_k = \begin{pmatrix} a_1 & a_2 & a_3 & \cdots & a_{k-1} & a_k \\ a_0 & a_1 & a_2 & \cdots & a_{k-2} & a_{k-1} \\ 0 & a_0 & a_1 & \cdots & a_{k-3} & a_{k-2} \\ & & & \vdots & & \\ 0 & 0 & 0 & \cdots & a_0 & a_1 \end{pmatrix}.$$

Theorem 5.2 *Given the conditions in Theorem 5.1 for the respective types of threshold, we have the following.*

(a) For $\rho_b > 1$ *the queue is transient.*

(b) For $\rho_b = 1$ *the queue is null recurrent.*

(c) For $\rho_b < 1$ *the queue is positive recurrent and the probability generating function for the equilibrium occupancy distribution is given by*

$$E\left[z^X\right] = \frac{1}{m}\left[\frac{b(z)(1-z) + \{b(z) - a(z)\}\, zR_q^{(F)}(z)}{b(z) - z}\right] \tag{13}$$

for $z \in [0,1)$. *The mean length* m *of the busy period is given by*

$$m = \left[\frac{1 + \{\rho_a - \rho_b\} R_q^{(F)}(1)}{1 - \rho_b} \right],$$

where $R_q^{(F)}(z)$ *is a non–negative function bounded above on the interval* $[0,1]$ *and determined by the specific type* F *of threshold applying and a parameter* q *associated with that type. Thus*

$$F = \begin{cases} U, & \textit{fixed upward threshold at } k, \quad q = k \in I\!N, \\ G, & \textit{geometric random-time,} \quad q = p \in (1 - 1/\alpha, 1), \\ T, & \textit{fixed-time threshold at } S, \quad q = S \in I\!N. \\ R, & \textit{general random-time,} \quad q = h : I\!N \to I\!R. \end{cases}$$

The functional forms are given by

$$R_k^{(U)}(z) = \frac{1}{z} \mathbf{e}_1 (\mathbf{I} - \mathbf{P}_k)^{-1} \mathbf{z}^t,$$

where $\mathbf{z}^t = (z, z^2, \ldots, z^k)$,

$$R_p^{(G)}(z) = \frac{z - F^*(1-p)}{z - a(z)(1-p)},$$

where $F^*(z)$ *is the probabiliy generating function for the distribution of the number of customers served during a busy period of the M/G/1 queue with service–time distribution* $A(\cdot)$ *(see [6, 7]),*

$$R_S^{(T)}(z) = \left[\frac{\xi_a(z)^S - 1}{\xi_a(z)^{S-1}(\xi_a(z) - 1)} \right] - \frac{(1 - \delta_{S1})}{z} \sum_{i=1}^{S-1} \left[\frac{\xi_a(z)^{S-i} - 1}{\xi_a(z)^{S-i-1}(\xi_a(z) - 1)} \right] a^{(i)}, \qquad (14)$$

where

$$a^{(i)} = \int_0^\infty \frac{e^{-\lambda x}(\lambda x)^{i-1}}{i!} dA^{(i)}(x),$$

$A^{(i)}(\cdot)$ *being the* i*-fold convolution of* $A(\cdot)$, *and*

$$R_h^{(R)}(z) = \sum_{n=1}^{\infty} h(n) R_n^{(T)}(z).$$

Proof: With the notation of (12), the martingale from Theorem 2.1 becomes

$$\begin{aligned} M_0(z) &= 1, \\ M_n(z) &= z^{X_n} \prod_{k=0}^{n-1} \left(\frac{z^{I(X_k \neq 0)}}{I(C_k^1)a(z) + I(C_k^2)b(z)} \right). \end{aligned}$$

If $\rho_a > 0$ and $\rho_b > 1$, Lemma 2 of Appendix A demonstrates that $\tau(0) = \infty$ with positive probability. Thus the queue is unstable in the sense that $X_n \to \infty$ as $n \to \infty$. The lemma shows also that the system is null–recurrent when $\rho_b = 1$.

Rearranging Theorem 3.5 gives the probability generating function for the distribution of the number of customers in the system at equilibrium to be

$$E\left[z^X\right] = \frac{1}{m} \left[\frac{b(z)(1-z)}{b(z) - z} + \frac{\{b(z) - a(z)\} z \left[E\left[z^{X_{\tau_1(0)}}\right] - z \right]}{(a(z) - z)(b(z) - z)} \right].$$

The term $E\left[z^{X_{\tau_1(0)}}\right] - z$, is different for each threshold (the derivation of the value in each case is given in Appendix B), but note that in each case the result is of the form

$$E\left[z^{X_{\tau_1(0)}}\right] - z = (a(z) - z) R(z), \qquad (15)$$

for some function $R(z)$ bounded on $[0,1]$. The reason for this can be found in [32, Appendix A], based on the recurrence relation for the embedded process. We can therefore write the solution as (13).

Since m is the mean number of customers served in a busy period, it can be calculated through use of the probability generating function $F^*(x_1, x_2)$. A simpler alternative is to let $z \uparrow 1$ in (13). The left–hand side is a probability generating function and so tends to unity. Since $a(1) = 1$, $a'(1) = \rho_a$, $b(1) = 1$ and $b'(1) = \rho_b$, the expression in brackets on the right may be evaluated by L'Hôpital's rule. Equating the two expressions yields

$$m = \left[\frac{1 + \{\rho_a - \rho_b\} R_q^{(F)}(1)}{1 - \rho_b}\right].$$

□

Remarks: (i) The form of the solution is that of the Pollaczek–Khintchine formula [7] for the probability generating function of the distribution of the number of customers in an $M/G/1$ queue with traffic intensity ρ_b, plus a correction term which takes into account the altered behaviour of the queue in phase A.
(ii) The value of m depends on the distribution $B(\cdot)$ only through ρ_b.
(iii) The solution for the queue–length threshold requires a matrix inversion, but the matrix $\mathbf{I} - \mathbf{P}_{K_o}$ to be inverted is already in upper–Hessenberg form [11] and the inversion is therefore easily performed, even for quite large matrices.
(iv) The solution for the fixed–time threshold can be found through a set of discrete convolutions using the fact that $a^{(k)}$ is $1/k$ times the probability that there are $k-1$ arrivals during k services so that

$$\begin{aligned} a^{(1)} &= a_0, \\ a^{(2)} &= a_1 a_0, \\ a^{(3)} &= a_1^2 a_0 + a_2 a_1 a_0^2, \\ &\vdots \end{aligned}$$

(v) The geometrically–distributed random threshold relies on the probability generating function $F^*(z)$ of the distribution of the number of customers served during a busy period of a standard $M/G/1$ queue. Although this may be hard to obtain in closed form, it can be easily calculated numerically through iteration to the fixed point of a simple equation, see for example [9].

5.3 The M/G/1 queue with generalised vacations

The solution above is similar to that for the $M/G/1$ queue with generalised vacations, also called the $M/G/1$ queue with an exceptional first service, where only the first arrival to an empty system notices altered behaviour. This is a special case of both

- the fixed upward threshold with $k = 0$,
- the fixed time threshold with $S = 1$.

In either case $R(z) = 1$ and

$$E\left[z^X\right] = \frac{1}{m}\frac{b(z)(1-z) + z(b(z) - a(z))}{b(z) - z},$$

which matches the result

$$E\left[z^X\right] = \frac{1}{m}\frac{b(z) - za(z)}{b(z) - z}$$

found for this system by Yeo [38].

5.4 The unreliable $M/G/1$ queue

Consider a queue which may break down at some stage during operation. Assuming the breakdown time occurs after a geometrically–distributed random number of customers have been served, we can model the system as a type (ii) 2–phase queue from the previous section. We assume that in standard operation the queue is stable, so that $\rho_a < 1$. Furthermore we assume that the repair time has probability distribution function $R(\cdot)$ and that any customers in the system at breakdown are lost, along with any that arrive during repair, so that $\rho_b = 0$ and $b(z) = 1$.

The probability generating function of the stationary distribution of the number of customers in this system is given by Theorem 5.2 as

$$E\left[z^X\right] \;=\; \frac{1}{m}\left[\frac{(1-z)+\{1-a(z)\}\,zR_p^{(G)}(z)}{1-z}\right]$$

for $z \in [0,1)$, with the mean number m of customers served in the busy period being given by

$$m \;=\; 1+\rho_a R_p^{(G)}(1)$$

and

$$R_p^{(G)}(z) = \frac{z-F^*(1-p)}{z-a(z)(1-p)},$$

where $F^*(z)$ is the probability generating function for the distribution of the number of customers served during a busy period of the $M/G/1$ queue with service–time distribution $A(\cdot)$.

5.5 Queue length overload control of the $M/G/1$ queue

In this example we consider a generalisation of the standard queue–length–threshold overload control in which there are two thresholds, one K_o for the onset of congestion and a second K_a for the abatement of congestion. This is a generalisation of the previous fixed–threshold queue, where $K_a = 0$. This type of system has been numerically analysed using the block matrix methodology of Neuts [24] in Neuts [23] and Li [18]. The present methodology allows us to derive closed forms of the probability generating functions.

If we assume that service times during congestion are reduced, possibly by discarding low–priority traffic, then we have traffic intensities ρ_j in the two regimes $j = u$(ncongested) $j = c$(ongested), and probability generating functions $a_j(z) = \sum_{i=1}^{\infty} a_i^j z^i = \tilde{G}_j(\lambda[1-z])$ $(j = u, c)$.

The hysteretic overload control of a queue is extensively considered in [31, 32], so the results described here are brief. Proofs may be found in [32]. Simply, the odd (even) phases in this system are the phases when the system enters the congested (uncongested) regime, listed in order from the beginning of the busy period so that $a_{2j+1} = a_u$ and $a_{2j} = a_c$. The results then require the use of an infinite sequence of phases because the queue may move from uncongested to congested an unbounded number of times before the end of a busy period. However, the probability of an infinite number of transitions is zero and all the relevant sums converge to give the result below.

Theorem 5.3 *For the process described above, when $\rho_u > 0$ and $\rho_c < 1$ the probability generating function for the distribution of the number of customers in the system seen by an arriving customer in the steady state is given by*

$$E\left[z^X\right] = \frac{1}{m}\left\{\frac{a_c(z)(1-z)+\{a_c(z)-a_u(z)\}R_{K_oK_a}(z)}{a_c(z)-z}\right\},$$

for $z \in [0,1)$. Here

$$\begin{aligned}
R_{K_oK_a}(z) &= \left(\mathbf{e}_1^T + \left(\frac{h_1}{1-h}\right)\mathbf{e}_{K_a}^T\right)(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{z},\\
h &= 1-a_0^u\,{\mathbf{e}_{K_a}}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{e}_1\,,\\
h_1 &= 1-a_0^u\,{\mathbf{e}_1}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{e}_1\,,
\end{aligned}$$

where $\mathbf{P}_{K_o}$ *is defined in terms of* a_i^u *by*

$$\mathbf{P}_{K_o} = \begin{pmatrix} a_1^u & a_2^u & a_3^u & \cdots & a_{K_o-1}^u & a_{K_o}^u \\ a_0^u & a_1^u & a_2^u & \cdots & a_{K_o-2}^u & a_{K_o-1}^u \\ 0 & a_0^u & a_1^u & \cdots & a_{K_o-3}^u & a_{K_o-2}^u \\ & & & \vdots & & \\ 0 & 0 & 0 & \cdots & a_0^u & a_1^u \end{pmatrix}$$

and we define $\mathbf{e}_i = (\delta_{1i}, \delta_{2i}, \ldots, \delta_{K_o i})^T$ *and* $\mathbf{z} = (z, z^2, \ldots, z^{K_o})^T$*. The mean number* m *of customers served in a busy period is given by*

$$m = \left[\frac{1 + \{\rho_u - \rho_c\} R_{K_o K_a}(1)}{1 - \rho_c}\right].$$

The reader will note the similarity of this result to that presented above for the two–phase system with a single upwards threshold. The only difference is in the function $R(z)$, which has an extra term. Of interest is the fact that these results show this form again and again for different thresholds, and phases structures. Such a simple form hints that more complex thresholding schemes could be dealt with in the same way, with similar results.

It is sometimes felt that probability generating functions are unsatisfactory for providing numerical results for distributions. In fact it is not hard to invert such a function to obtain the distribution. To illustrate the practical utility of the results we demonstrate in one case using the method of Daigle [8] to perform the inversion. Similarly, we could invert any of the probability generating functions presented above to obtain numerical results. An alternative method for performing the inversion is given in [1].

The scenario for which we produce numerical results has threshold values of $K_a = 50, K_o = 62$ taken from realistic values given in Rumsewicz and Smith [33]. Note that these are not trivially small and require inversion of a 62×62 array, but this presents no numerical difficulties. The service times are exponential (though we examine more complex cases equally easily in [32]) and during congestion the service rate is doubled to reduce the congestion. *Figure 1* shows five cases, one with a standard load, one at the edge of being overloaded, and three overload scenarios. The log–probability graph clearly shows that the overload control has the desirable property of preventing excursions to long queue lengths for the overload scenarios without substantially effecting the behaviour when there is no overload.

6 Other Results

In this section we examine some other results that arise as part of our analysis of the examples presented. In particular we focus on the two–phase case discussed above, though similar results can be obtained even for the infinite–phase case [32].

6.1 The probability of a given phase

For some applications, the probability of being in a given phase is of interest. For example, we may wish to minimise the cost of running the queue, given the costs for running it in phases 1 and 2. Alternatively, in the breakdown model, we may want to know the proportion of customers affected by breakdowns. In this section we shall use Little's law to calculate these probabilities. Little's Law [19]

$$L = \lambda W \tag{16}$$

links the average number L of customers in a system, the arrival rate λ to the system and the average time W spent by a customer in the system. When this is applied to the server taken as a system in its own right, L is the probability that there is a customer in the system and W the average service time, so

$$\begin{aligned} L &= P(X \neq 0) = 1 - 1/m, \\ W &= \frac{1-\phi}{\mu_a} + \frac{\phi}{\mu_b}, \end{aligned}$$

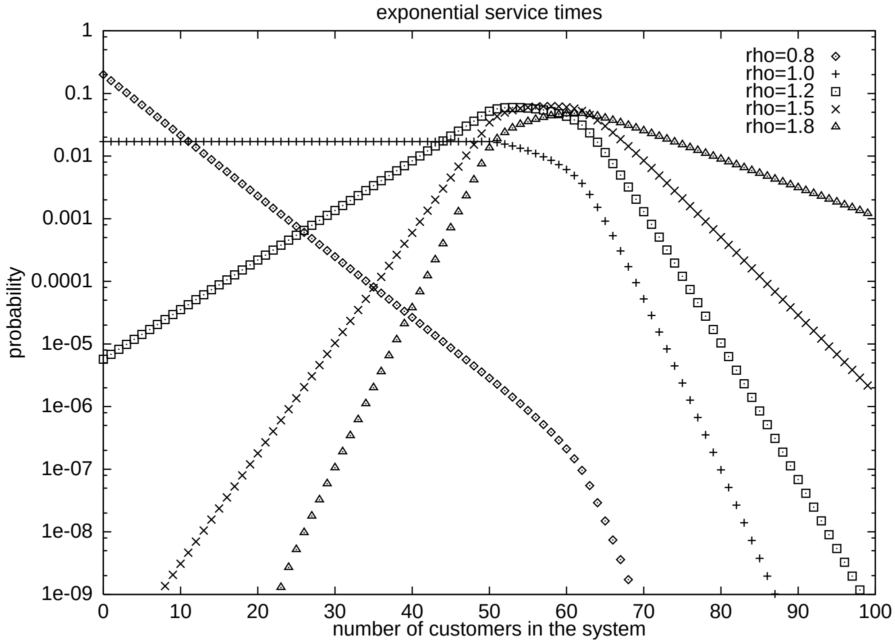

Figure 1: The queue–length distribution with $K_a = 50, K_o = 62$ and exponential service times on a log–probability graph.

where ϕ is the probability the system is in phase 2. We have from Theorem 5.2 that

$$m = \left[\frac{1 + \{\rho_a - \rho_b\}\, R_q^{(F)}(1)}{1 - \rho_b}\right],$$

where $R_q^{(F)}(z)$ is determined by the specific type of threshold between the phases. Thus

$$\begin{aligned} L &= \frac{\rho_b + (\rho_a - \rho_b) R_q^{(F)}(1)}{1 + (\rho_a - \rho_b) R_q^{(F)}(1)}, \\ \lambda W &= \phi(\rho_b - \rho_a) + \rho_b. \end{aligned}$$

Substitution in (16) yields an equation for ϕ when $\rho_a \neq \rho_b$, namely

$$\phi = \frac{1 + (\rho_a - 1) R_q^{(F)}(1)}{1 + (\rho_a - \rho_b) R_q^{(F)}(1)}. \tag{17}$$

Thus the probability of being in phase 2 is insensitive to the form of the distribution $B(\cdot)$, depending only on ρ_b, the traffic intensity during phase 2.

As an example, consider the random threshold model, where

$$R_q^{(F)}(z) = R_p^{(R)}(z) = \frac{z - F^*(1-p)}{z - (1-p)a(z)}.$$

Taking $z = 1$ gives

$$R_q^{(F)}(1) = \frac{1}{p}[1 - F^*(1-p)].$$

Substitution back into (17) yields

$$\phi = \frac{p + (\rho_a - 1)[1 - F^*(1-p)]}{p + (\rho_a - \rho_b)[1 - F^*(1-p)]}.$$

In the breakdown model where all of the customers in the system at the time of a breakdown are discarded, $\rho_b = 0$. In this case ϕ is the probability that a customer is discarded because of a breakdown and is given by

$$\phi = \frac{p + (\rho_a - 1)[1 - F^*(1-p)]}{p + \rho_a[1 - F^*(1-p)]} \quad = \quad 1 - \frac{1 - F^*(1-p)}{p + \rho_a[1 - F^*(1-p)]}.$$

6.2 The length of the phases

Another question of interest is the time spent in each of the two phases. As we have seen in the previous sections, the probability generating functions $E\left[z^{\tau_1(0)}\right]$ are complicated and depend on the specific threshold. We shall restrict our attention to averages, which are relatively easy to calculate. The numbers of customers served respectively during phases 1 and 2 of a busy period are given by $\nu_1(0)$ and $\nu_2(0)$. We calculate the expected values as follows. We have

$$\begin{aligned} \frac{d}{dz}\left[F_0^*\left(\frac{z}{a(z)}, \frac{z}{b(z)}\right)\right] &= \left(\frac{a(z) - za'(z)}{a(z)^2}\right) E\left[\nu_1(0)\left(\frac{z}{a(z)}\right)^{\nu_1(0)-1}\left(\frac{z}{b(z)}\right)^{\nu_2(0)}\right] \\ &\quad + \left(\frac{b(z) - zb'(z)}{b(z)^2}\right) E\left[\nu_2(0)\left(\frac{z}{a(z)}\right)^{\nu_1(0)}\left(\frac{z}{b(z)}\right)^{\nu_2(0)-1}\right], \\ \frac{d}{dz}\left[F_1^*\left(\frac{z}{b(z)}\right)\right] &= \left(\frac{b(z) - zb'(z)}{b(z)^2}\right) E\left[\nu_2(0)\left(\frac{z}{b(z)}\right)^{\nu_2(0)-1}\right], \end{aligned}$$

from which we derive on substituting $z = 1$ that

$$\frac{d}{dz}\left[F_0^*\left(\frac{z}{a(z)}, \frac{z}{b(z)}\right)\right]_{z=1} = (1-\rho_a)\, E\left[\nu_1(0)\right] + (1-\rho_b)\, E\left[\nu_2(0)\right], \tag{18}$$

$$\frac{d}{dz}\left[F_1^*\left(\frac{z}{b(z)}\right)\right]_{z=1} = (1-\rho_b)\, E\left[\nu_2(0)\right]. \tag{19}$$

From (10), (11) and (15) we deduce that

$$F_0^*\left(\frac{z}{a(z)}, \frac{z}{b(z)}\right) = z, \tag{20}$$

$$F_1^*\left(\frac{z}{b(z)}\right) = [a(z) - z]\, R_q^{(F)}(z) + z. \tag{21}$$

On taking derivatives in (20) and (21) (again putting $z = 1$) and comparing with (18) and (19), we get

$$\begin{aligned} (1-\rho_a)E[\nu_1(0)] + (1-\rho_b)E[\nu_2(0)], &= 1, \\ (1-\rho_b)E[\nu_2(0)] &= (\rho_a - 1)R_q^{(F)}(1) + 1, \end{aligned}$$

whence

$$E[\nu_1(0)] = R_q^{(F)}(1), \tag{22}$$

$$E[\nu_2(0)] = \frac{1 + (\rho_a - 1)R_q^{(F)}(1)}{1 - \rho_b}. \tag{23}$$

Note that by adding the corresponding sides of (22) and 23) we obtain

$$E[\nu(0)] = \frac{1 + (\rho_a - \rho_b)R_q^{(F)}(1)}{1 - \rho_b},$$

which agrees with the value of m derived above.

7 Conclusion

This paper has presented a generalisation of a method for solving a class of queueing problems of $M/G/1$ type. The method is not limited to the class of problems considered here, but could be further extended. In [4] the authors use similar techniques to derive transient results for the $M/G/1$ queueing system, and the same could be done for the present class of queueing systems.

We present a number of examples, and give illustrative numerical results for an example which models a hysteretic overload control.

There are still open problems. For instance, Condition $(*)$ for regularity of the stopping times used in the proofs can be seen (with other minor conditions) to be a sufficient condition for recurrence of the queueing system, and hence stability. It would be somewhat easier to use the results described herein if the reverse were true, that is, if stability of the queueing system of interest implied the regularity of the stopping times involved. While this is the case in each specific example we have examined in detail, a general proof has remained elusive.

References

[1] Joseph Abate and Ward Whitt. Numerical inversion of probability generating functions. *Operations Research Letters*, 12:245–251, 1992.

[2] F. Baccelli. Exponential martingales and wald's formula for two-queue networks. *J.Appl.Probab.*, 23:812–819, 1986.

[3] F. Baccelli and A.M. Makowski. Direct martingale arguments for stability: the M/GI/1 case. *Systems Control Letters*, 6:181–186, 1985.

[4] F. Baccelli and A.M. Makowski. Dynamic,transient and stationary behaviour of the M/GI/1 queue via martingales. *Annals of Probability*, 17(4):1691–1699, 1989.

[5] F. Baccelli and A.M. Makowski. Martingale relations for the M/GI/1 queue with markov modulated poisson input. *Stochastic Processes and their Applications*, 38:99–133, 1991.

[6] J.W. Cohen. *The Single Server Queue.* North-Holland, Amsterdam, 1969.

[7] R.B. Cooper. *Introduction to Queueing Theory.* The Macmillian Company, 1972.

[8] John N. Daigle. Queue length distributions from probability generating functions via discrete Fourier transforms. *Operations Research Letters*, 8:229–236, August 1989.

[9] John N. Daigle and Matthew Roughan. Queue-length distributions for multi-priority queueing systems. IEEE INFOCOM'99, 1999.

[10] J. Dshalalow. Multi-channel queueing systems with infinite waiting rooms and stochastic control. *J.Appl.Probab.*, 26:345–362, 1989.

[11] G.H. Golub and C.F. van Loan. *Matrix Computations.* North Oxford Academic, 1983.

[12] Wei-Bo Gong, Anlu Yan, and Christos G. Cassandras. The M/G/1 queue with queue-length dependent arrival rate. *Commun. Statist.-Stochastic Models*, 8(4):733–741, 1992.

[13] O.C. Ibe and K.S. Trivedi. Two queues with alternating service and server breakdown. *Queueing Systems*, 7:253–268, 1990.

[14] M. Kijima and N. Makimoto. Computation of the quasi-stationary distributions in m(n)/gi/1/k and gi/m(n)/1/k queues. *Queueing Systems*, 11:255–272, 1992.

[15] M. Kijima and N. Makimoto. A unifed approach to gi/m(n)/1/k and m(n)/g/1/k queues via finite quasi-birth-death processes. *Commun. Statist.-Stochastic Models*, 8(2):269–288, 1992.

[16] L. Kleinrock. *Queueing Systems*, volume I: Theory. John Wiley and Sons, 1975.

[17] R.O. LaMaire. M/G/1/N vacation model with varying e-limited service discipline. *Queueing Systems*, 11:357–375, 1992.

[18] San-Qi Li. Overload control in a finite message storage buffer. *IEEE Transactions on Communications*, 37(12):1330–1338, 1989.

[19] J.D.C. Little. A proof of the queueing formula: L = λ w. *Operations Research*, 9(3):383–87, 1961.

[20] N.A. Marlow and M. Tortorella. Some general characteristics of two-state reliability models for maintained systems. *Journal of Applied Probability*, 32:805–820, 1995.

[21] J.A. Morrison. Two-server queue with one server idle below a threshold. *Queueing systems*, 7:325–336, 1990.

[22] T. Nakagawa and S. Osaki. Markov renewal processes with some non-regeneration points and their applications to reliability theory. *Microelectronics and Reliability*, 15:633–636, 1976.

[23] Marcel F. Neuts. A queueing model for a storage buffer in which the arrival rate is controlled by a switch with random delay. *Performance Evaluation*, 5:243–256, 1985.

[24] M.F. Neuts. *Structured stochastic matrices of M/G/1 type and their applications.* Marcel Dekker, 1989.

[25] J. Neveu. *Discrete-Time Martingales* North-Holland, Amsterdam, 1975.

[26] J.H. Park. The analysis of the M^X/G/1 queue by a martingale method. Master Thesis, Korea Advanced Institute of Science and Technology, 1990.

[27] C.E.M. Pearce and M. Roughan. Forward delay times in multi-phase discrete-time renewal processes. *Asia-Pacific Journal of Operations Research*, 14(1):1–10, 1997.

[28] R. Pyke. Markov renewal processes: Definitions and preliminary properties. *Ann. Math. Stat.*, 32:1231–1242, 1961.

[29] R. Pyke. Markov renewal processes with finitely many states. *Ann. Math. Stat.*, 32:1242–1259, 1961.

[30] M. Roughan. An analysis of a modified M/G/1 queue using a martingale technique. *Journal of Applied Probability*, 33:224–238, March 1996.

[31] M. Roughan and C.E.M. Pearce. Analysis of a hysteretic overload control. In *ITC-16, Edinburgh*, pages 293–302, 1999.

[32] M. Roughan and C.E.M. Pearce. A martingale analysis of hysteretic overload control. *Advances in Performance Analysis*, 3(1):1–30, 2000.

[33] Michael P. Rumsewicz and Donald E. Smith. A comparison of SS7 congestion control options during mass call-in situations. *IEEE/ACM Transactions on Networks*, 3(1):1–9, Feb 1995.

[34] J.A. Schormans, J.M. Pitts, and E.M. Scharf. A priority queue with superimposed geometric batch arrivals. *Communications in Statistics-Stochastic Models*, 9(1):105–122, 1993.

[35] L. Takács. *Introduction to the Theory of Queues.* Oxford University Press, 1962.

[36] H. Takagi. Time dependent process of M/G/1 vacation models with exhaustive service. *J.Appl.Prob.*, 29:418–429, 1992.

[37] R.W. Wolff. Poisson arrivals see time averages. *Opns. Res.*, 30:223–231, 1982.

[38] G.F. Yeo. Single server queue with modified service mechanisms. *Journal of the Australian Mathematical Society*, 2:499–507, 1962.

A Stability and regularity

In order for stationary distributions to exist the processes involved must be stable and recurrent. Lemma 2 provides a useful result based on the stability of the $M/G/1$ queue and Condition $(*)$, presented in the following section, a sufficient condition for stability. The process should also be irreducible, that is, there should be no more than one communicating class, a possibility unlikely in the models considered here, but possible in certain pathological examples.

The obvious criterion is simply that $\tau(n)$ be a.s. finite for all $n \in \mathbb{Z}^+$, that is, that state 0 will recur a.s. within a finite time. Lemma 1 shows that it is sufficient to look at $\tau(0)$.

At the completion of a busy period, the process enters phase one, which it cannot leave until there has been at least one service completion and so at least one arrival. Hence we shall require $\rho_1 > 0$.

Lemma 1: *Suppose $X_0 = 0$ and $\tau(0)$ is a.s. finite. Then $\tau(n)$ and $\tau_i(n)$ are also a.s. finite for $i \in \mathcal{S}$ and all $n \in \mathbb{Z}^+$.*

Proof: When $X_0 = 0$, $\tau(0) < \infty$ a.s. implies that the busy period is a.s. finite and so $\tau(n) < \infty$ a.s. Hence $\tau_i(n) < \infty$ a.s., since $\tau_i(n) \leq \tau(n)$ a.s. □

Lemma 2: *If $P(\tau_N(0) - \tau_{N-1}(0) > 0) > 0$ and $\rho_N > 1$, then the process is unstable, while if $\rho_N = 1$, the system will be null-recurrent.*

Proof: It is implicit that $\tau_{N-1}(0) < \infty$. In this case we have $P(\tau_N(0) - \tau_{N-1}(0) > 0) > 0$, so over a busy period there is a positive probability of spending time in phase N. While in phase N the queue behaves as an $M/G/1$ queue and the stability conditions of the $M/G/1$ queue apply. Hence for $\rho_N > 1$, $P(\tau_N(0) < \infty) < 1$ and the queue is unstable, and for $\rho_N = 1$ the system is null–recurrent. □

We now turn our attention to the primary question of regularity of stopping times. To apply Doob's Optional Sampling Theorem we must demonstrate that the stopping times involved are regular for the martingale. This is trivial when the martingale is uniformly integrable (Neveu, IV-**3**-14), but the latter need not be the case. Condition (*) provides a sufficient condition for regularity, as is shown below.

Theorem A.1 *If Condition $(*)$ is satisfied, then the stopping times $\tau_0(n), \ldots \tau_N(n)$ and $\tau(n)$ are regular for the martingale $(M_n(z))$, $z \in [0,1]$, $n \in \mathbb{Z}^+$. Furthermore, when $\tau(n) = \infty$,*

$$M_{\tau(n)}(z) = 0.$$

Proof: First we consider the case with $\mathcal{S}^* = \phi$. In this case Takács's lemma ([35, page 47]) implies that $\xi_j(z) \leq 1$ for all $j \in \mathcal{S}$. Hence $|M_n(z)| \leq 1$ and so $\Big(M_n(z)\Big)$ is uniformly integrable. As $\Big(M_n(z)\Big)$ is a positive integrable martingale, condition (a) of Neveu IV-**3**-14 is automatically satisfied. Since $M_n(z)$ is uniformly integrable, so too is $M_n(z)\, I(\tau > n)$ for all stopping times τ and so condition (b) of Neveu IV-**3**-14 is also satisfied. Hence any stopping time is regular in this case.

Next we show that $\tau(0)$ is regular for the case when $\mathcal{S}^* = \{i\}$ using Neveu IV-**3**-16 which has two necessary conditions. Condition (1) of this proposition,

$$\int_{\{\tau(0)<\infty\}} |M_{\tau(0)}(z)| dP < \infty,$$

is automatically satisfied for our martingale. Condition (2),

$$\lim_{n\to\infty} \int_{\{\tau(0)>n\}} |M_n(z)| dP = 0,$$

can be seen to hold as follows. Noting that the martingale is non–negative, we start with $n > 0$ from

$$|M_n(z)| = z^{X_n} \prod_{k=0}^{n-1} \frac{z^{I\{X_k \neq 0\}}}{\sum_{j=1}^N I(C_k^j) a_j(z)},$$

which because $\tau(0) > n$ gives

$$\begin{aligned}
|M_n(z)| &\leq \left(\prod_{k=0}^{(\tau_1(0)\wedge n)-1} \xi_1(z) \right) \left(\prod_{k=\tau_1(0)\wedge n}^{(\tau_2(0)\wedge n)-1} \xi_2(z) \right) \cdots \left(\prod_{k=\tau_{N-1}(0)\wedge n}^{(\tau_N(0)\wedge n)-1} \xi_N(z) \right) \\
&= \xi_1(z)^{(\tau_1(0)\wedge n)} \xi_2(z)^{(\tau_2(0)\wedge n)-(\tau_1(0)\wedge n)} \cdots \xi_N(z)^{n-(\tau_{N-1}(0)\wedge n)} \\
&\leq \alpha^{(\tau_i(0)\wedge n)-(\tau_{i-1}(0)\wedge n)} \\
&\leq \alpha^{\tau_i(0)-\tau_{i-1}(0)} \ a.s., \qquad (24)
\end{aligned}$$

as $\alpha = \sup_{z\in[0,1]} \xi_i(z) > 1$. By the a.s. finiteness of $\tau(0)$ implied by $(*)$,

$$\lim_{n\to\infty} I(\tau(0) > n) = 0 \ a.s. \qquad (25)$$

Thus (24) and (25) imply $|M_n(z)|\, I(\tau(0) > n) \to 0$ a.s. as $n \to \infty$. Also from (24) we obtain

$$|M_n(z)|\, I(\tau(0) > n) \leq \alpha^{\tau_i(0)-\tau_{i-1}(0)} \ a.s.,$$

the right–hand side of which has finite expectation by assumption. Thus by dominated convergence

$$\lim_{n\to\infty} E\left[|M_n(z)|\, I(\tau(0) > n)\right] = E\left[\lim_{n\to\infty} |M_n(z)|\, I(\tau(0) > n)\right] = 0,$$

which satisfies Condition (2). Thus $\tau(0)$ is regular for $(M_n(z))$ with $z \in [0,1]$. The regularity of $\tau(0)$ for general $\mathcal{S}^*$ can be derived in a similar manner.

We now show that the regularity of $\tau(0)$ implies that of $\tau(n)$ and $\tau_i(n)$ for $i \in \mathcal{S}$ and all $n \in \mathbb{N}$. Given $\tau(n)$ is regular, Neveu IV-**3**-13 implies that $\tau_i(n)$ must also be regular, so it remains to show that the regularity of $\tau(0)$ implies the regularity of $\tau(n)$. Once again Condition (1) of Neveu IV-**3**-16 is automatically satisfied, so we need only show Condition (2), which we do as follows. We have

$$M_{\tau(n)}(z) = z^{X_{\tau(n)}} \left\{ \prod_{k=0}^{\eta(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \right\} \prod_{k=\eta(n)}^{\tau(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right),$$

where $\eta(n) = \sup\{m \leq n | X_m = 0\}$, the epoch of the beginning of the current busy period. The last product is over one busy period. Due to the regenerative nature of this process at time $\eta(n)$, for $m \geq n$

$$\begin{aligned}
&E\left[I(\tau(n) > m)\, z^{X_{\tau(n)}} \prod_{k=\eta(n)}^{\tau(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \right] \\
&= E\left[I(\tau(0) > m - \eta(n))\, z^{X_{\tau(0)}} \prod_{k=0}^{\tau(0)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \right].
\end{aligned}$$

Also the events

$$\prod_{k=0}^{\eta(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \text{ and } I(\tau(n) > m)\, z^{X_{\tau(n)}} \prod_{k=\eta(n)}^{\tau(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right)$$

are independent. Thus for $m \geq n$ we can write

$$\begin{aligned}
&E\left[|M_{\tau(n)}(z)| I(\tau(n) > m) \right] \\
&= E\left[\prod_{k=0}^{\eta(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \right] E\left[I(\tau(n) > m)\, z^{X_{\tau(n)}} \prod_{k=\eta(n)}^{\tau(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \right] \\
&= E\left[\prod_{k=0}^{\eta(n)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \right] E\left[I(\tau(0) > m - \eta(n))\, z^{X_{\tau(0)}} \prod_{k=0}^{\tau(0)-1} \left(\frac{z^{I(X_k\neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)} \right) \right].
\end{aligned}$$

For fixed, finite n, the first expectation is clearly finite, and as $n \to \infty$ the second tends to

$$\lim_{n\to\infty} E\left[I(\tau(0) > m) M_{\tau(0)}(z)\right]$$

which tends to zero as $m \to \infty$ since $\tau(0)$ is regular, as was shown above. Thus, once again, we have Condition (2) of Neveu IV-**3**-16, leading to regularity of the stopping times. $\square$

Henceforth we shall refer to the stopping time γ generically for any one of n, $\tau_0(n), \ldots \tau_N(n)$ or $\tau(n)$ for $n \in \mathbb{Z}^+$.

Theorem 2.2 *If Condition $(*)$ is satisfied, then the stopping times $\tau_0(\gamma), \ldots, \tau_N(\gamma)$ and $\tau(\gamma)$ are regular for the martingale $(M_n(z))$, $z \in [0,1]$. Furthermore when $\tau(\gamma) = \infty$, we have $M_{\tau(\gamma)}(z) = 0$ a.s.*

Proof: We proceed with the application of Neveu's Proposition IV-**3**-16 as in the previous theorem. As before Condition (1) is automatically satisfied. Condition (2),

$$\lim_{n\to\infty} \int_{\{\tau(\gamma)>n\}} |M_n(z)|\, dP = 0,$$

can be verified as follows. We have

$$\begin{aligned}
\lim_{n\to\infty} \int_{\{\tau(\gamma)>n\}} |M_n(z)|\, dP &= \lim_{n\to\infty} \int_{\{\tau(\gamma)>n\}} \left[I(\gamma \le n) + I(\gamma > n)\right] M_n(z)\, dP \\
&= \lim_{n\to\infty} \int_{\{\tau(\gamma)>n\}} I(\gamma \le n) M_n(z)\, dP + \lim_{n\to\infty} \int_{\{\gamma>n\}} M_n(z)\, dP. \qquad (26)
\end{aligned}$$

When $\gamma = \tau(m)$, the second term on the right in (26) becomes

$$\lim_{n\to\infty} \int_{\{\tau(m)>n\}} M_n(z)\, dP \quad = \quad 0,$$

since by Theorem A.1, $\tau(m)$ is regular for the martingale and satisfies condition (2) of Neveu IV-**3**-16. Hence

$$\begin{aligned}
\lim_{n\to\infty} \int_{\{\tau(\gamma)>n\}} |M_n(z)|\, dP &= \lim_{n\to\infty} \int_{\{\tau(\gamma)>n\}} I(\gamma \le n) M_n(z)\, dP \\
&= \lim_{n\to\infty} \int_{\{\gamma \le n < \tau(\gamma)\}} M_n(z)\, dP,
\end{aligned}$$

when $\gamma = \tau(m)$. In this integral $\tau(m) \le n$ and so we can write

$$M_n(z) \quad = \quad \left[\prod_{k=0}^{\tau(m)-1} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)}\right)\right] \left[z^{X_n} \prod_{k=\tau(m)}^{n-1} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^N I(C_k^j) a_j(z)}\right)\right].$$

As in the previous theorem the two parts of this product are independent due to the regenerative nature of the process. Furthermore the regularity of $\tau(m)$ implies that the former term has finite expectation. The expectation of the second part can be shown to tend to zero as $n \to \infty$ as before.

Thus the theorem holds for $\gamma = \tau(m)$. When $\gamma = \tau_i(m)$ we have $\tau(\tau(m)) \ge \tau(\tau_i(m))$ a.s. and so from Neveu IV-**3**-13 we get the regularity of all $\tau(\tau_i(m))$, the result we need. The latter part of the theorem again comes directly from Neveu IV-**3**-16. $\square$

B Two phase example - proofs

Proofs for the fixed upward queue–length threshold appear in [30], so we need consider only the other three cases.

B.1 The geometrically–distributed random–time threshold

This applies when the threshold occurs at a random time R with the geometric distribution $P(R = i) = (1-p)^{i-1}p$. The parameter p $(0 < p < 1)$ gives the probability of the switch occurring at the end of a service, that is, given $\phi(n) = 1$ and $X_{n+1} \neq 0$, we have $P(\tau_1(n) = n + 1) = p$. Thus

$$\begin{aligned} \tau_1(n) &= \begin{cases} \tau(n) \wedge (n + R), & \text{if } \phi(n) = 1 \\ n, & \text{if } \phi(n) = 2 \end{cases} \\ &= I(C_n^1)\,[\tau(n) \wedge (n + R)] + I(C_n^2)n. \end{aligned}$$

From this we have

$$\tau_1(0) \quad = \quad \tau(0) \wedge R.$$

The following lemma proves that Condition $(*)$ is satisfied.
Lemma 3: *When $\rho_a > 0$, $\rho_b \leq 1$ and $p > 1 - 1/\alpha$, Condition $(*)$ is satisfied.*
Proof: When $\rho_a \leq 1$ and hence $\alpha = 1$, Condition $(*)$ is trivially true. When $\rho_a > 1$ and hence $\alpha > 1$, then $\alpha^{\tau_1(0)} \leq \alpha^R$ *a.s.* and so

$$E\left[\alpha^{\tau_1(0)}\right] \leq E\left[\alpha^R\right] = \sum_{i=1}^{\infty} \alpha^i P(R = i) = p\alpha \sum_{i=1}^{\infty} \alpha^{i-1}(1-p)^{i-1}.$$

This converges to $p\alpha[1 - \alpha(1-p)]^{-1}$ when $\alpha(1-p) < 1$, that is, when $p > 1 - 1/\alpha$.

We next proceed to find the value of $R_p^{(G)}(z)$.

Theorem B.1 *For $\rho_a > 0$, $p > 1 - 1/\alpha$ with $\alpha = \sup_{z\in[0,1]} \xi_1(z)$ and $z \in [0, 1)$,*

$$E\left[z^{X_{\tau_1(0)}}\right] - z \quad = \quad [a(z) - z]\, R_p^{(G)}(z),$$

where

$$R_p^{(G)}(z) \quad = \quad \frac{z - F^*(1-p)}{z - a(z)(1-p)}$$

and $F^*(\xi)$ is the unique solution $z = Z(\xi)$ to $z = \xi a(z)$ for $z \in [0, 1)$.

Proof: There are two cases to consider:

1. the busy period ends before we hit the random threshold and therefore $X_{\tau_1(0)} = 0$;

2. the busy period ends after the transition into phase 2.

Decomposing into the two cases we get

$$E\left[z^{X_{\tau_1(0)}}\right] \quad = \quad P(\tau(0) \leq R) + E\left[I(\tau(0) > R)z^{X_R}\right], \tag{27}$$

where R is the geometric random variable.

The first term is given by

$$\frac{F^*(1-p)}{1-p}, \tag{28}$$

where F^* is the PGF of the number of customers served in the busy period of a simple $M/G/1$ queue with service–time distribution $A(\cdot)$. $F^*(\xi)$ is the unique solution $z = Z(\xi)$ to $z = \xi a(z)$ for $z \in [0, 1)$.

The second term can be derived by noting that the behaviour of the system in phase 1 will be just that of a simple $M/G/1$ queue with service–time distribution $A(\cdot)$. Define

$$g(y, n) = E\left[y^{\mu_1(n)} I(\tau(0) > n)\right]$$

for the single–phase $M/G/1$ queue. We can use a slight modification of the proof of Theorem 3.2 to show that

$$E\left[I(\tau(0)>n)z^{X_n}\right]=g(\xi_a(z),n).$$

Baccelli and Makowski [4, 2.30] show that for the simple $M/G/1$ queue

$$\sum_{n=1}^{\infty} g(y,n)t^n=\frac{tF^*(y)-yF^*(t)}{y-t}.$$

Hence

$$\begin{aligned} E\left[I(\tau(0)>R)z^{X_R}\right] &= \sum_{n=1}^{\infty} E\left[I(\tau(0)>n)z^{X_n}\,\middle|\,R=n\right]p\{R=n\}, \\ &= \frac{p}{1-p}.\frac{(1-p)z-\xi_a(z)F^*(1-p)}{\xi_a(z)-(1-p)}. \end{aligned} \tag{29}$$

Combining (28) and (29) in (27) gives the required result. □

B.2 Fixed–time threshold

In this case the threshold is passed after a fixed number S of customers have been served during a busy period. If the system clears before this threshold is reached, then it is reset when the next busy period begins, as times spent in phases in different busy periods are independent. Denote by $\chi(n)$ the number of customers served since the beginning of the current busy period. Then

$$\begin{aligned} \tau_1(n) &= \begin{cases} \tau(n)\wedge\Big(n+S-\chi(n)\Big) & \text{if } \phi(n)=1 \\ n & \text{if } \phi(n)=2 \end{cases} \\ &= \tau(n)\wedge\left[n+(S-\chi(n))^+\right]. \end{aligned}$$

More simply, when we start from an empty system we get

$$\tau_1(0) \quad = \quad \tau(0)\wedge S.$$

Condition $(*)$ is trivially satisfied for this type of threshold, since $\alpha>1$ implies $\alpha^S\geq\alpha^{\tau_1(0)}$ a.s.

Theorem B.2 *For $\rho_a>0$, $S\in I\!N$ and $z\in[0,1]$,*

$$E\left[z^{X_{\tau_1(0)}}\right]-z=[a(z)-z]R_S^{(T)}(z), \tag{30}$$

where $R_S^{(T)}(z)$ is defined recursively by $R_1^{(T)}(z)=1$ and

$$R_{S+1}^{(T)}(z)=\frac{1}{z}\left\{R_S^{(T)}(z)a(z)-a_0R_S^{(T)}(0)\right\}+1.$$

Proof: Define the stopped process (Z_n) by $Z_n=X_{n\wedge\tau(0)}$. As $E\left[z^{X_{\tau_1(0)}}\right]=E\left[z^{Z_S}\right]$, we consider Z_S rather than $X_{\tau_1(0)}$ throughout the proof. When $S=1$

$$E\left[z^{X_{\tau_1(0)}}\right]-z=E\left[z^{Z_1}\right]-z=E\left[z^{A_1}\right]-z=a(z)-z,$$

where A_1 is the number of arrivals during the first service. This gives (30) for $S=1$, providing the basis for proving (30) by induction. Now $Z_{n+1}=Z_n+A_{n+1}-1$ for $S=n+1$ and so

$$E\left[z^{Z_{n+1}}\,\middle|\,Z_n\right] \quad = \quad \begin{cases} z^{Z_n}\frac{a(z)}{z}, & Z_n>0, \\ 1, & Z_n=0. \end{cases}$$

Hence

$$E\left[z^{Z_{n+1}}\right]-z \;=\; P(Z_n=0)+\frac{a(z)}{z}\left\{E\left[z^{Z_n}\right]-P(Z_n=0)\right\}-z.$$

For the inductive step, assume (30) is true for $S=n$. Since $P(Z_n=0)=a_0R_n^{(T)}(0)$

$$\begin{aligned}E\left[z^{Z_{n+1}}\right]-z &= a_0R_n^{(T)}(0)+\frac{a(z)}{z}\left\{[a(z)-z]R_n^{(T)}(z)+z-a_0R_n^{(T)}(0)\right\}-z\\ &= [a(z)-z]\left\{\frac{R_n^{(T)}(z)a(z)}{z}-\frac{a_0R_n^{(T)}(0)}{z}+1\right\},\end{aligned}$$

which with the given recursion for $R_S^T(z)$ gives (30) for $S=n+1$. Therefore by induction the theorem is true for all $S\in I\!N$. □

The previous theorem does not provide a complete solution as it does not give $R_n^{(T)}(0)$, nor does it give the solution in closed form, as required for Theorem 5.2 equation (14). The following theorem provides these missing results.

Theorem B.3 *For the process discussed above*

$$R_S^{(T)}(z) \;=\; \left[\frac{1-\xi_a(z)^S}{\xi_a(z)^{S-1}(1-\xi_a(z))}\right]-\frac{(1-\delta_{S1})}{z}\sum_{k=1}^{S-1}\left[\frac{1-\xi_a(z)^{S-k}}{\xi_a(z)^{S-k-1}(1-\xi_a(z))}\right]a^{(k)}\quad (z\neq 0)$$

and

$$R_S^{(T)}(0) \;=\; \frac{1}{a_0}\sum_{k=1}^{S}a^{(k)},$$

where

$$a^{(k)}=\int_0^\infty\frac{e^{-\lambda x}(\lambda x)^{k-1}}{k!}dA^{(k)}(x).$$

Proof: First note that $E\left[z^{X_{\tau_1(0)}}\right]-z$ for any random time threshold can be formed from a sum of the fixed time threshold value weighted by the probability distribution. In the case of the geometrically distributed random time threshold, we have

$$R_p^{(G)}(z) \;=\; \frac{p}{1-p}\,[a(z)-z]\sum_{n=1}^{\infty}R_n^{(T)}(z)(1-p)^n \tag{31}$$

We have already derived $R_p^{(G)}(z)$ in Theorem B.1. Hence the problem is one of finding a Taylor series expansion for $R_p^{(G)}(z)$ in $1-p$ about 0 and equating coefficients. To do so we use Takács's lemma,

$$F^*(p)=\sum_{k=1}^{\infty}p^ka^{(k)},$$

and some extensive rearrangement to get

$$R_n^{(T)}(z) \;=\; \left[\frac{1-\left(\frac{a(z)}{z}\right)^n}{1-\frac{a(z)}{z}}\right]-\frac{(1-\delta_{n,1})}{z}\sum_{k=1}^{n-1}\left[\frac{1-\left(\frac{a(z)}{z}\right)^{n-k}}{1-\frac{a(z)}{z}}\right]a^{(k)}$$

for $z\in(z_0,1]$, where $pa(z_0)/z_0=1$. In fact $R_n^{(T)}(z)$ is independent of p, so we can take the limit as $p\to 0$. Thus the above is true for $z\in(0,1]$. $R_n^{(T)}(0)$ may be found by substituting $z=0$ in (31), performing the same Taylor series expansion and equating coefficients once again. □