A Martingale Analysis of Hysteretic Overload Control.

M. Roughan
Department of Electrical Engineering,
University of Melbourne,
Victoria 3010, AUSTRALIA.
m.roughan@ee.mu.oz.au

C.E.M. Pearce
Department of Applied Mathematics,
University of Adelaide,
Adelaide 5005, AUSTRALIA.
cpearce@maths.adelaide.edu.au

Abstract: Overload control is critical in preventing congestion in modern switching networks. One method, hysteretic overload control, uses two thresholds, a congestion onset and a congestion abatement threshold, to detect congestion. Variations of this method of overload control have been used in the Signaling System Number 7 (SS7) protocol specified by the ITU-T (International Telecommunications Union, Telecommunications Standardization Sector) and also proposed for use in broadband networks. We provide an analytic technique for investigating the performance of such congestion controls and thence setting parameters such as the threshold levels. The technique relies on a martingale–based relationship between a queueing process and an embedded renewal process.

1 Introduction

How does one protect a modern switching network from overload? Answering this question has become critical to the reliable operation of modern switching networks, because of the increase in services with unpredictable traffic loads. An example is provided by the Common Channel Signaling traffic associated with Intelligent Network services such as 'televoting'. In essence, there are two related questions: how to detect or measure the congestion caused by an overload and how to mitigate it. A simple and intuitively appealing mechanism to detect congestion is a queue–length threshold. The purpose of this paper is to examine the behavior of systems that use two distinct queue–length thresholds to detect congestion. Such a technique has been recommended for the Signaling

System Number 7 (SS7) protocol [1, p. 313], [2], and proposed for application in broadband networks [3].

Congestion is detected *via* a pair of queue–length thresholds, a congestion onset K_o and a congestion abatement threshold K_a. For example, in the SS7 protocol, a link is considered congested if the number of messages in the Signaling Transfer Point (STP) link transmit buffer exceeds the onset threshold, and the link returns to the uncongested state only when the number of messages in the buffer falls to the congestion abatement threshold or below. The two thresholds typically are chosen so that $K_a < K_o$, leading to a hysteretic effect, described below.

When congestion occurs the control acts to reduce the input traffic by discarding some of the input packets. In broadband networks selective discard of packets [4] discards low priority packets to minimize the impact on perceived quality of service. The model used here for the discard strategy is Percentage Throttling (PT), where some percentage of the originating traffic is randomly blocked at the source. In this model we assume that blocked traffic is lost from the system, that is, customers do not retry at a later time, or alternatively packets are not retransmitted.

Rumsewicz and Smith [2] used simulations to compare a realistic implementation of this overload control with others used in SS7. Their results indicated that a simple system as described above (though with more than one level of throttling) was preferable to more complex systems that use multiple thresholds for different priority messages.

There are a number of mathematical analyses of various overload control systems in which $K_a = K_o$. For instance, Morrison [5] investigated a system in which a second server is added when congestion is detected. Gong and Cassandras [6] considered a system in which the arrival rate is dependent on the number of customers in the queue. Both these examples are limited to systems in which service times are exponentially distributed. Perry and Asmussen [7] considered a queue with generally–distributed service times and an admission policy based on either the workload in the queue or the sojourn time of a customer in the queue. More recently Leung [3] considered a system with the service–time distribution dependent on the workload in the system.

These examples do not allow for the two distinct thresholds that lead to the hysteretic effect in which the queue exhibits different behavior when the load increases, from that as it decreases. Hysteresis has been suggested as a mechanism to reduce the number of times the congestion status switches state [1], reducing any cost associated with this switching.

The block–matrix methodology of Neuts [8] has been used by Neuts [9] and Li [10] to derive numerical results for systems with hysteretic thresholds. In this paper we use an analytic form for the generating function of the number of messages in the buffer, found using an elegant martingale–based methodology. The closed–form result requires little computation to evaluate the queue–length distribution and thence the queue utilization and the blocking probability in the finite–buffer case. Further, the method allows the derivation of critical features of the overload control, such as the time between the onset and abatement of congestion.

The buffer is modeled using a variant of the $M/G/1$ queue in which the queue state is separated into two regimes, congested and uncongested, each with its own arrival rate. The technique relies on a martingale analysis based on the work of Rosenkrantz [11] and Baccelli and Makowski [12, 13] and extended by Roughan in [14] and [15]. Perry and Asmussen [7] have used similar arguments.

Our main result, Theorem 3, which gives the probability generating function for the distribution of the number of customers in the system (as seen by an arriving customer), was conjectured in [15]. We provide a proof of this conjecture through Theorem 2, which demonstrates the conditions required by the conjecture. We derive a number of quantities including

- the probability the queue is congested (in Section 3.5),
- the traffic load accepted by the system (Section 3.5) and
- the time between onset and abatement of congestion (Section 3.6).

We present examples of numerical results for each of these performance measures which verify quantitatively the intuition about the effects of hysteretic overload controls.

The paper is organized as follows. Section 2 describes our overload control model and Section 3 provides a mathematical analysis of this model, including stability results, the derivation of the generating function and the derivation of the time spent between switching congestion status. Section 4 provides numerical results for the performance measures listed above, as well as the queue–length distribution. Section 5 suggests some extensions to the work and summarizes our key results. An appendix gives the derivation of a technical result useful in calculating the generating functions used here.

2 The model

This section provides a definition of our model of a buffer which uses hysteretic overload control. The model is a generalization of the $M/G/1$ queue, a simple queue with Poisson arrivals, generally-distributed service times, a single server, and an infinite waiting room. The queue represents customers' messages waiting to be processed in the buffer of some processor. The $M/G/1$ queue is generalized to model the overload control by separating its behavior into two regimes of operation, congested and uncongested. The PT source overload control changes the uncongested arrival rate λ_u to λ_c during the congested regime.

An alternative to source overload control is to alter the service–time distribution of the process, for instance, by stripping the headers to find the message priority and discarding those of low priority, resulting in a short service time for these messages. If the service time for the discarded packages were zero, this model would be essentially the same as the source control model described above. In reality it takes some processing time even to discard a message. Furthermore, in practice retrials may result in significant problems for this type of control. Therefore source control, as considered below, is preferable.

The regime changes from uncongested to congested when, after completion of a service (the processing of a message in the buffer), the number of messages in the system is greater than the congestion onset threshold K_o. The regime changes from congested to uncongested when the number of messages in the buffer falls to the congestion abatement threshold K_a or below. Typically $K_a < K_o$, resulting in hysteretic behavior. The case $K_a = K_o$ is included in the analysis described here but $K_a > K_o$ makes little sense and is not.

The process is modeled as follows. Take the number of customers in the system at time t to be $X(t)$ and the service completion epochs to be $t_1 < t_2 < \cdots$, where t_n is the departure time of the nth customer. We consider the process embedded at customer departure epochs, that is, (X_n), where $X_n = X(t_n+)$, the number of customers in the system as seen by the nth departing customer. Cooper [16, pp. 154] shows that the arriving customers see the same queue–length distribution as the departures. Note that, in practice the distribution seen by the arrivals is or equal or greater importance than the stationary distribution. Furthermore, in the model described above, the congestion status may be changed only at the completion of a service and therefore depends only on the embedded queueing process (X_n).

We assume the process begins at time zero with a dummy departure leaving the queue empty, that is, $t_0 = 0$ and $X_0 = 0$. The assumption is convenient, and does not effect our results as we are concerned here with equilibrium behavior.

Arrivals to the process are Poisson with rates λ_u and λ_c depending on the current congestion status. Service times are independently and identically distributed with probability distribution function $G(\cdot)$ and mean $1/\mu$. The traffic intensities ρ_s are given by $\rho_s = \lambda_s/\mu$ for $s = u, c$.

We model the arrivals using two distinct sequences of independent identically distributed random variables A_n^s ($s = u, c$ and $n = 1, 2, \ldots$). Here A_n^u and A_n^c are respectively the numbers of customers to arrive during the nth service given that during this service the queue is uncongested or congested. The probability generating function for the number of arrivals during a service is $a_s(z) = \sum_{i=1}^{\infty} a_i^s z^i = \tilde{G}(\lambda_s[1-z])$ $(s = u, c)$, where $a_i^s = \mathrm{prob}\{A_1^s = i\}$ and $\tilde{G}(\cdot)$ is the Laplace–Stieltjes transform of the service–time distribution function G (see [16]).

3 The martingale analysis

The model defined above is specified on a probability space $(\Omega, \mathcal{F}, P)$ generated by the congestion status and number of customers in the system. We define the filtration $(\mathcal{F}_n)$ by

$$\mathcal{F}_n = \sigma(X_0, A_m^s | \, 1 \le m \le n, s = u, c).$$

This contains the history of the queueing process, including the congestion status, up to time n.

3.1 Phases and stopping times

The process can be modeled using the analysis of [14] by regarding each busy cycle as a sequence of phases. We denote $P_n = 1, 2, \ldots.$ the phase at time $n \in \mathbb{Z}^+$. A phase ends when the queue changes congestion status and the cycle of phases restarts when the busy cycle ends, that is, when the system becomes empty. Let C_n denote the congestion status (u or c) at time n. We have the following rules:

```
if $X_n = 0$ then $P_n = 1$,
else if $C_n = C_{n-1}$ then $P_n = P_{n-1}$,
else if $C_n \neq C_{n-1}$ then $P_n = P_{n-1} + 1$.
```

Odd–numbered phases then correspond to periods when the queue status is uncongested, even–numbered to congested periods. We may define a_i^j for $j = 1, 2, \ldots$ to be a_i^u for j odd and a_i^c for j even and thence define $a_j(z)$ and ρ_j for $j = 1, 2, \ldots.$. We employ the usual indicator notation

$$I(A) = \begin{cases} 1, & \text{when event } A \text{ occurs,} \\ 0, & \text{otherwise.} \end{cases}$$

The assumption that a dummy service completion occurs at time zero leaving the system empty implies $P_0 = 1$. If $X_0 = 0$, then X_n and $I(P_m = j)$ are determined purely by A_m^j at times $m \le n$ and hence the ends of the phases are stopping times.

We define, for each time $n \in \mathbb{Z}^+$, a stopping time $\tau(n)$ which is either the end of the current busy period or, if the queue is empty at time n, the end of the next busy period. We define also the stopping time $\tau_j(n)$ ($j = 0, 1, 2, \ldots$) to be the maximum of n and the end of phase j in the current busy cycle. When the busy period ends, we consider the process to go through the remaining phases, spending zero time in each, and so we may formally define

$$\begin{aligned}
\tau(n) &= \begin{cases} \inf\{m > n \mid X_m = 0\}, & \text{if the set is non-empty,} \\ \infty, & \text{otherwise,} \end{cases} \\
\tau_j(n) &= \tau(n) \wedge \inf\{m \geq n \mid P_m > j\},
\end{aligned}$$

where as usual $x \wedge y :=\min(x, y)$. We have $n = \tau_0(n) \leq \tau_1(n) \leq \tau_2(n) \leq \cdots \leq \tau(n)$. As the ends of phases correspond to changes in the congestion status, an intuitive recursive definition of the stopping times $\tau_j(n)$ is

$$\begin{aligned}
\tau_{2j+1}(n) &= \tau(n) \wedge \inf\{m \geq \tau_{2j}(n) \mid X_m > K_o\}, \\
\tau_{2j+2}(n) &= \tau(n) \wedge \inf\{m \geq \tau_{2j+1}(n) \mid X_m \leq K_a\}.
\end{aligned}$$

We can also define, for $j = 1, 2, \ldots$, the times

$$\nu_j(n) = \tau_j(n) - \tau_{j-1}(n), \tag{1}$$

$$\mu_j(n) = \begin{cases} \nu_j(n), & X_n \neq 0, \\ 0, & X_n = 0. \end{cases} \tag{2}$$

We can interpret $\mu_j(n)$ as the forward recurrence time at time n of phase $j+1$ in a cyclical generalized Markov renewal process [17].

3.2 The martingale

We now define the martingale which will provide the majority of our results.
Theorem 1: *If*

$$\begin{aligned}
M_0(z) &= 1, \\
M_n(z) &= z^{X_n} \prod_{k=0}^{n-1} \left(\frac{z^{I(X_k \neq 0)}}{\sum_{j=1}^{\infty} I(P_k = j) a_j(z)} \right), \quad n \geq 1,
\end{aligned}$$

then $(M_n(z))_{n \geq 0}$ *is a nonnegative integrable martingale for* $z \in (0, 1]$.

Proof: The result is a straightforward extension of those in [14] and [15], which use the recurrence relation

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

to demonstrate directly the martingale condition $E\left[M_{n+1}(z)|\,\mathcal{F}_n\right] = M_n(z)$ a.s. □

3.3 Stability and regularity

Of obvious interest are the conditions for stability of the queue. These are established in [15]. Simply stated, the queue is stable if and only if $0 \leq \rho_c < 1$. It is null-recurrent for $\rho_c = 1$ and transient for $\rho_c > 1$. A desirable consequence is that stability is independent of the uncongested traffic intensity, and hence an overloaded queue will be stable so long as the originating traffic is sufficiently throttled.

The result can be understood intuitively by noting that, when congested, the queue behaves as if it were a standard $M/G/1$ queue with traffic intensity ρ_c. This queue is always considered congested when there are more than K_o customers in the buffer. Hence regardless of its uncongested behavior, the queue reverts to the standard stability behavior of the $M/G/1$ queue whenever there are more than K_o customers in the buffer.

The following analysis makes use of the optional sampling theorem [18], which requires that the stopping times involved be regular for the martingale. In general stronger conditions than stability are required for this to apply. The following theorem proves that in the present context stability is a sufficient condition for regularity.

Theorem 2: *The stopping times $\tau_i(n)$ and $\tau(n)$ $(i, n \in \mathbb{Z}^+)$ are regular for the martingale $(M_n(z))_0^\infty$.*

Proof: Theorem 3.3 of [15] shows that establishing regularity reduces to demonstrating a condition referred to as (*). The condition is that

$$E\left[\prod_{j \in S^*} \xi_j(z)^{\nu_j(0)}\right] < \infty$$

for all $z \in [0, 1]$, where $\xi_j(z) = z/a_j(z)$, and S^* is the set of all indices j such that the traffic intensity ρ_j during phase j exceeds unity. For a stable queue $\rho_c < 1$ (although ρ_u may be greater than one). Therefore the condition is

$$E\left[\prod_{j \text{ odd}} \xi_u(z)^{\nu_j(0)}\right] < \infty.$$

The random variable X_n can decrease by a maximum of 1 at each time step, and therefore at congestion abatement times τ_{2n} the process is always exactly at the congestion abatement threshold, that is, $X_{\tau_{2n}} = K_a$. Consequently the process exhibits renewals at times τ_{2i}, that is, the behavior of the system before and after τ_{2i} is independent. Therefore the times $\nu_j(0)$ for j odd are independent. Thus we may reduce the condition to

$$\prod_{j \text{ odd}} E\left[\xi_u(z)^{\nu_j(0)}\right] < \infty.$$

The renewals at congestion abatement times imply the probabilities $p\{\nu_i = k\}$ for $i > 1$ satisfy the recurrence

$$\begin{aligned} p\{\nu_{2i+1} = k\} &= hp\{\nu_{2i-1} = k\}, & \text{for } k > 0, \\ p\{\nu_{2i+1} = 0\} &= p\{\nu_{2i-1} = 0\} + (1-h)p\{\nu_{2i-1} > 0\}, \end{aligned}$$

where $h \in (0,1)$ is the probability that the process returns to the congested state from the abatement threshold before the busy period terminates. Define

$$\nu_{2i+1}(z) = \sum_{k=0}^{\infty} \xi_u(z)^k p\{\nu_{2i+1} = k\}.$$

We can multiply the recurrence relation above by $\xi_u(z)^k$ and sum over k to get

$$\begin{aligned} \nu_{2i+1}(z) &= h\sum_{k=1}^{\infty} \xi_u(z)^k p\{\nu_{2i-1} = k\} + p\{\nu_{2i-1} = 0\} + (1-h)p\{\nu_{2i-1} > 0\} \\ &= h\sum_{k=0}^{\infty} \xi_u(z)^k p\{\nu_{2i-1} = k\} + 1 - h, \end{aligned}$$

when this exists. For $i > 1$ this recurrence relation has solution

$$\nu_{2i+1}(z) = h^{i-1}\nu_3(z) + 1$$

wherever $\nu_3(z)$ exists. In [15, Lemma 4.3.1] it was shown that $\nu_1(z)$ exists for $z \in [0,1]$. A minor modification of the lemma shows that $\nu_3(z)$ also exists.

Therefore

$$\prod_{j \text{ odd}} E\left[\xi_u(z)^{\nu_j(0)}\right] = \prod_{i=0}^{\infty} \nu_{2i+1}(z) = \nu_1(z)\nu_3(z) \prod_{i=2}^{\infty} (1 + h^{i-1}\nu_3(z))$$

when this exists. From Gradshteyn and Ryzhik [19, 0.252] a necessary and sufficient condition for the last product to converge is that $\sum_{i=2}^{\infty} h^{i-1}\nu_3(z)$ converge. This is automatic since $h \in (0,1)$, so we are done. □

3.4 Equilibrium results

Theorem 3: *Define*

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

and set $\mathbf{e}_i = (\delta_{1i}, \delta_{2i}, \ldots, \delta_{K_o i})^T$ *and* $\mathbf{z} = (z, z^2, \ldots, z^{K_o})^T$. *Then if* $\rho_u > 0$ *and* $\rho_c < 1$, *the probability generating function for the equilibrium number of customers in the system (as seen by arriving customers) 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 $z \in [0,1)$, *where*

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

and the mean number m *of customers served in a busy period is given by*

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

Proof: Doob's Optional Sampling Theorem [18, Proposition IV-3-12] states that for stopping times S, T satisfying $S \leq T$ a.s. and which are regular for the martingale (M_n),

$$E\left[M_T \middle| \mathcal{F}_S\right] = M_S, \ \ a.s.$$

A consequence is Theorem 3.10 of [15], which states that the probability generating function of the number of customers in the queue is given by

$$\begin{aligned} E\left[z^X\right] &= \frac{1}{m}\left\{\frac{E\left[z^{X_{\tau_1(0)}}\right] - z}{1 - \xi_u(z)} + \frac{E\left[z^{X_{\tau_2(0)}}\right] - E\left[z^{X_{\tau_1(0)}}\right]}{1 - \xi_c(z)}\right\} \\ &+ \frac{1}{m}\sum_{n=1}^{\infty}\left\{\frac{E\left[z^{X_{\tau_{2n+1}(0)}}\right] - E\left[z^{X_{\tau_{2n}(0)}}\right]}{1 - \xi_u(z)}\right\} \\ &+ \frac{1}{m}\sum_{n=1}^{\infty}\left\{\frac{E\left[z^{X_{\tau_{2n+2}(0)}}\right] - E\left[z^{X_{\tau_{2n+1}(0)}}\right]}{1 - \xi_c(z)}\right\}. \end{aligned} \tag{4}$$

We define $h_n = p\{\tau_n(0) < \tau(0)\}$, the probability that phase $n+1$ is reached *before* the end of the first busy period. If the system moves into the congested state by passing above the congestion onset threshold, it must, at some time before the end of the busy period, pass below the congestion abatement threshold. Hence $h_{2n} = h_{2n-1}$. Also, as noted above, when the process drops to the abatement threshold there is a renewal in the sense that the future behavior of the queue is independent of the past behavior of the queue. Hence $h_{2n+2} = h_{2n}h$ for $n \in I\!N$. From these two relationships we derive

$$\begin{aligned} h_{2n+1} &= h_1 h^n, && (5)\\ h_{2n+2} &= h_1 h^n. && (6)\end{aligned}$$

If $\tau_n(0) < \tau(0)$ then $X_{\tau_n(0)} \neq 0$, and so for $n > 1$ we get

$$\begin{aligned} E\left[z^{X_{\tau_n(0)}}\right] &= E\left[z^{X_{\tau_n(0)}} I(X_{\tau_{n-1}(0)} = 0)\right] + E\left[z^{X_{\tau_n(0)}} I(X_{\tau_{n-1}(0)} \neq 0)\right] \\ &= p\{X_{\tau_{n-1}(0)} = 0\} + p\{X_{\tau_{n-1}(0)} \neq 0\} E\left[z^{X_{\tau_n(0)}} \middle| X_{\tau_{n-1}(0)} \neq 0\right] \\ &= 1 - h_{n-1} + h_{n-1} E\left[z^{X_{\tau_n(0)}} \middle| X_{\tau_{n-1}(0)} > 0\right]. \qquad (7)\end{aligned}$$

For $n > 0$, (7) gives

$$\begin{aligned} E\left[z^{X_{\tau_{2n}(0)}}\right] &= 1 - h_{2n-1} + h_{2n-1} z^{K_a}, \\ E\left[z^{X_{\tau_{2n+1}(0)}}\right] &= 1 - h_{2n} + h_{2n} E\left[z^{X_{\tau_{2n+1}(0)}} \middle| X_{\tau_{2n}(0)} = K_a\right]. \end{aligned}$$

We may set

$$r(z) = E\left[z^{X_{\tau_{2n+1}(0)}} \middle| X_{\tau_{2n}(0)} = K_a\right] \qquad (8)$$

for $n > 0$, since this expression is independent of n. Then

$$\begin{aligned} E\left[z^{X_{\tau_2(0)}}\right] - E\left[z^{X_{\tau_1(0)}}\right] &= 1 - h_1 + h_1 z^{K_a} - E\left[z^{X_{\tau_1(0)}}\right], && (9)\\ E\left[z^{X_{\tau_{2n+1}(0)}}\right] - E\left[z^{X_{\tau_{2n}(0)}}\right] &= h_1 h^{n-1}(r(z) - z^{K_a}), && (10)\\ E\left[z^{X_{\tau_{2n+2}(0)}}\right] - E\left[z^{X_{\tau_{2n+1}(0)}}\right] &= h_1 h^{n-1}\left(1 - h + hz^{K_a} - r(z)\right). && (11)\end{aligned}$$

Substitution from (9), (10) and (11) into (4) gives

$$\begin{aligned} E\left[z^X\right] &= \frac{1}{m}\left\{ \frac{E\left[z^{X_{\tau_1(0)}}\right] - z}{1 - \xi_u(z)} + \frac{1 - E\left[z^{X_{\tau_1(0)}}\right]}{1 - \xi_c(z)} \right. \\ &\qquad \left. + \frac{h_1}{(1-h)}\left[\frac{(r(z) - z^{K_a})(\xi_u(z) - \xi_c(z))}{(1 - \xi_u(z))(1 - \xi_c(z))}\right]\right\}. \qquad (12)\end{aligned}$$

We now calculate $r(z) - z^{K_a} = E\left[z^{X_{\tau_{2n+1}(0)}}\middle| X_{\tau_{2n}(0)} = K_a\right] - z^{K_a}$, which (21) in the Appendix gives as

$$r(z) - z^{K_a} = \frac{a_u(z)}{z}\left[1 - \xi_u(z)\right]\sum_{n=0}^{\infty} \mathbf{g}_n^T \mathbf{z},$$

where $\mathbf{g}_n = (g_n(1), g_n(2), \cdots, g_n(K_o))^T$ and

$$g_n(m) = p\{\tau_{2i+1}(0) > \tau_{2i}(0) + n, X_{\tau_{2i}(0)+n} = m | X_{\tau_{2i}(0)} = K_a\}.$$

It is evident that $\mathbf{g}_n^T = \mathbf{e}_{K_a}^T \mathbf{P}_{K_o}^{\;n}$. The sum over n of $\mathbf{P}_{K_o}^{\;n}$ has been shown to converge ([14, Lemma 3.2]), giving $\sum_{n=0}^{\infty} \mathbf{g}_n^T = \mathbf{e}_{K_a}^T (\mathbf{I} - \mathbf{P}_{K_o})^{-1}$. Therefore

$$r(z) - z^{K_a} = \frac{a_u(z)}{z}\left[1 - \xi_u(z)\right] \mathbf{e}_{K_a}^T (\mathbf{I} - \mathbf{P}_{K_o})^{-1} \mathbf{z}. \qquad (13)$$

Substitution from (13) into (12) gives

$$E\left[z^X\right] = \frac{1}{m}\left\{ \frac{E\left[z^{X_{\tau_1(0)}}\right] - z}{1 - \xi_u(z)} + \frac{1 - E\left[z^{X_{\tau_1(0)}}\right]}{1 - \xi_c(z)} + \frac{h_1}{1-h}\left[\frac{\mathbf{e}_{K_a}^T (\mathbf{I} - \mathbf{P}_{K_o})^{-1} \mathbf{z}\left(1 - \frac{a_u(z)}{a_c(z)}\right)}{(1 - \xi_c(z))}\right]\right\}.$$

The first two terms of the right–hand side of this equation appeared in [14] where a specific case of this queue, the $M/G/1$ queue with the abatement threshold set to zero, was considered. Note that the terms depend only on $\tau_1(0)$, which is unaffected by the abatement threshold. These terms were shown in in [14] to provide

$$\frac{E\left[z^{X_{\tau_1(0)}}\right] - z}{1 - \xi_u(z)} + \frac{1 - E\left[z^{X_{\tau_1(0)}}\right]}{1 - \xi_c(z)} = \frac{a_c(z)(1-z) + \{a_c(z) - a_u(z)\}\, \mathbf{e}_1 (\mathbf{I} - \mathbf{P}_{K_o})^{-1} \mathbf{z}^t}{a_c(z) - z}.$$

Rearranging slightly, we get

$$E\left[z^X\right] = \frac{1}{m}\left\{ \frac{a_c(z)(1-z) + \{a_c(z) - a_u(z)\}\left[\left(\mathbf{e}_1 + \frac{h_1}{1-h}\mathbf{e}_{K_a}\right)(\mathbf{I} - \mathbf{P}_{K_o})^{-1}\mathbf{z}^t\right]}{a_c(z) - z} \right\}.$$

We now calculate m using $E\left[z^X\right]_{z=1} = 1$. Taking the limit as $z \uparrow 1$ using L'Hôpital's rule yields that

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

As $E\left[z^X\right]_{z=0} = 1/m$, the probability that the queue is empty is $1/m$ and hence m is the mean number of customers served in a busy period. From the definitions of h and h_1, we get

$$\begin{aligned} h_1 &= 1 - E\left[z^{X_{\tau_1(0)}} \middle| X_0 = 0 \right]_{z=0}, \\ h &= 1 - E\left[z^{X_{\tau_{2n+1}(0)}} \middle| X_{\tau_{2n}(0)} = K_a \right]_{z=0}, \end{aligned}$$

which can be calculated from (13) and the similar expression in [14] to be

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

the desired result. □

Remark 1: The form of the solution is that of the Pollaczek–Khintchine Equation [16] for the probability generating function of the stationary number of customers in the $M/G/1$ queue with traffic intensity ρ_c, plus a correction term which takes into account the altered behavior of the queue in the uncongested regime. The solution, though more complicated, is very similar to that for the $M/G/1$ queue with generalized vacations where only the first arrival to an empty system notices altered behavior.

Remark 2: The solution requires a matrix inversion. The matrix $(\mathbf{I} - \mathbf{P}_{K_o})$ to be inverted is already in upper–Hessenberg form [20] and the inversion is therefore easily performed, even for quite large matrices.

Remark 3: The theorem has been described in terms of a source control model, but applies equally well to packet discard models where the service-time distribution of discarded packets is changed. In this case $a_j(z) = \tilde{G}_j(\lambda[1 - z])$, where $G_j(\cdot)$ is the service–time distribution during the congested phase. Furthermore, in this case the arrivals form a homogeneous Poisson process, and therefore PASTA [21] (Poisson Arrivals See Time Averages) implies that the arriving customers see the time–averaged behavior of the system. Hence our result gives the stationary queue–length distribution.

3.5 Simple performance estimates

First we introduce some terminology. The *offered load* ρ_u refers to the load offered to the system prior to any overload control. The *accepted load* ρ_a is that part of the load accepted by the system after application of overload controls. The *rejected load* ρ_r refers to traffic blocked by the overload control, not by overflowing a finite buffer.

To calculate the accepted load we apply Little's law $L = \lambda W$ to the processor, rather than the queue, so that L is the average work in the system, namely the processor utilization, while λ is the arrival rate to the system and W the mean service time. The processor utilization is one minus the probability $1/m$ of the system being empty. The arrival rate times the mean service time is the accepted load ρ_a. Thus

$$\rho_a = 1 - \frac{1}{m} = \frac{\rho_c + (\rho_u - \rho_c) R_{K_o K_a}(1)}{1 + (\rho_u - \rho_c) R_{K_o K_a}(1)}. \tag{14}$$

To calculate the proportion of time the system spends in the congested state we note that PT is applied during congestion reducing the load on the system from ρ_u to ρ_c. The accepted load on the system is thus $\rho_a = (1-\psi)\rho_u + \psi\rho_c$, where ψ is the proportion of time the queue spends congested. In conjunction with (14), this expression yields

$$\psi = \frac{1 + (\rho_u - 1) R_{K_o K_a}(1)}{1 + (\rho_u - \rho_c) R_{K_o K_a}(1)}. \tag{15}$$

The rejected traffic is just $\rho_r = (\rho_u - \rho_c)\psi = \rho_u - \rho_a$, and therefore the probability of being blocked by the overload control is $p_B = \rho_r/\rho_u$.

3.6 The time spent in the congested region

One of the principal reasons for introducing the hysteretic effect into this type of threshold–based overload control is to limit the oscillatory behavior that can occur for a single fixed threshold. In order to measure this behavior we must calculate the time spent before switching regimes.

Theorem 4: *For $n \geq 1$, $\rho_c < 1$ and $\rho_u > 1$ we have*

$$E\left[\nu_{2n+1}(0) \middle| \nu_{2n+1}(0) > 0\right] = \mathbf{e}_{K_a}^T (\mathbf{I} - \mathbf{P}_{K_o})^{-1} \mathbf{1}, \tag{16}$$

$$E\left[\nu_{2n+2}(0) \middle| \nu_{2n+2}(0) > 0\right] = \frac{K_a(1-h) + (\rho_u - 1)\, \mathbf{e}_{K_a}^T (\mathbf{I} - \mathbf{P}_{K_o})^{-1} \mathbf{1}}{(1-\rho_c) h}. \tag{17}$$

Proof: Corollary 3.7 of [15] with N set equal to ∞ states that

$$E\left[\prod_{j=i+1}^{\infty} \xi_j(z)^{\nu_j(0)}\right] = E\left[z^{X_{\tau_i(0)}}\right].$$

Taking the difference for consecutive values of i supplies

$$E\left[\left(1-\xi_{i+1}(z)^{\nu_{i+1}(0)}\right)\prod_{j=i+2}^{\infty} \xi_j(z)^{\nu_j(0)}\right] \quad = \quad E\left[z^{X_{\tau_{i+1}(0)}}\right] - E\left[z^{X_{\tau_i(0)}}\right]. \quad (18)$$

Since we are interested only in limiting behavior, we consider the cases $i > 1$. The right–hand side of (18) is given in these two cases i even and i odd by (10) and (11) which state, *via* (13), that

$$\begin{aligned} E\left[z^{X_{\tau_{2n+1}(0)}}\right] - E\left[z^{X_{\tau_{2n}(0)}}\right] &= h_1 h^{n-1}\left[\frac{a_u(z)}{z} - 1\right]\mathbf{e}_{K_a}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{z}, \\ E\left[z^{X_{\tau_{2n+2}(0)}}\right] - E\left[z^{X_{\tau_{2n+1}(0)}}\right] &= \\ h_1 h^{n-1}&\left((1-h)(1-z^{K_a}) - \left[\frac{a_u(z)}{z} - 1\right]\mathbf{e}_{K_a}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{z}\right). \end{aligned}$$

The derivative of the left–hand side at $z = 1$ is

$$\frac{d}{dz}\left[E\left[\left(1-\xi_{i+1}(z)^{\nu_{i+1}(0)}\right)\prod_{j=i+2}^{\infty} \xi_j(z)^{\nu_j(0)}\right]\right]_{z=1} = (\rho_{i+1}-1)E\left[\nu_{i+1}(0)\right].$$

Similarly the right–hand side leads to

$$\begin{aligned} \frac{d}{dz}\left[E\left[z^{X_{\tau_{2n+1}(0)}}\right] - E\left[z^{X_{\tau_{2n}(0)}}\right]\right]_{z=1} &= h_1 h^{n-1}\left[\rho_u - 1\right]\mathbf{e}_{K_a}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{1}, \\ \frac{d}{dz}\left[E\left[z^{X_{\tau_{2n+2}(0)}}\right] - E\left[z^{X_{\tau_{2n+1}(0)}}\right]\right]_{z=1} &= \\ h_1 h^{n-1}&\left(-(1-h)K_a - \left[\rho_u - 1\right]\mathbf{e}_{K_a}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{1}\right). \end{aligned}$$

Equating gives

$$\begin{aligned} E\left[\nu_{2n+1}(0)\right] &= h_1 h^{n-1}\mathbf{e}_{K_a}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{1}, \\ E\left[\nu_{2n+2}(0)\right] &= \frac{h_1 h^{n-1}}{1-\rho_c}\left((1-h)K_a + \left[\rho_u - 1\right]\mathbf{e}_{K_a}^T(\mathbf{I}-\mathbf{P}_{K_o})^{-1}\mathbf{1}\right). \end{aligned}$$

From (5) and (6), we have

$$\begin{array}{rclclcl} p\{\nu_{2n+1}(0) > 0\} & = & p\{X_{\tau_{2n}(0)} > 0\} & = & h_{2n} & = & h_1 h^{n-1}, \\ p\{\nu_{2n+2}(0) > 0\} & = & p\{X_{\tau_{2n+1}(0)} > 0\} & = & h_{2n+1} & = & h_1 h^{n}. \end{array}$$

Since $E\left[X|\, X > 0\right] = E\left[X\right]/p\{X > 0\}$ for a nonnegative random variable, we get

$$\begin{aligned} E\left[\nu_{2n+1}(0)|\, \nu_{2n+1}(0) > 0\right] &= \mathbf{e}_{K_a}^T (\mathbf{I} - \mathbf{P}_{K_o})^{-1}\mathbf{1}, \\ E\left[\nu_{2n+2}(0)|\, \nu_{2n+2}(0) > 0\right] &= \frac{(1-h)K_a + (\rho_u - 1)\,\mathbf{e}_{K_a}^T (\mathbf{I} - \mathbf{P}_{K_o})^{-1}\mathbf{1}}{(1-\rho_c)h}, \end{aligned}$$

as required. □

Remark 4: The expression $E\left[\nu_{2n+2}(0)|\, \nu_{2n+2}(0) > 0\right]$ in (17) is the average number of customers served between the onset and abatement of congestion. Equation (16) includes the possibility that the phase ends because the busy cycle has ended. The two conditional expectations cannot be added directly to obtain a measure of the cycle length (the total time between an onset and the following onset) because of the different conditionings. We now address this question.

Theorem 5: *The mean number of customers served in a cycle through two consecutive phases (congested and uncongested) is given by*

$$E\left[\nu\right] = \frac{m(1-h)}{h_1},$$

where $\nu = \nu_{2n+1} + \nu_{2n+2}$ for some $n \geq 1$.

Proof: The busy period is divided into a number of cycles through pairs of phases. We can calculate the average time of a cycle by calculating the mean number of customers served in a busy period and dividing by the average number of cycles occurring during one busy period. The average number of cycles occurring is simply the probability h_1 of at least one cycle occurring times the average number $1/(1-h)$ of cycles occurring. Putting these components together gives the required result. □

4 Numerical results

We now describe some examples and provide numerical results. We begin by presenting a method for inverting generating functions to find queue–length distributions and then describe the examples to be considered. The section then provides numerical results relating to the queue–length distribution and other performance measures. Note that although we derive the queue–length distribution for the infinite–buffer case, loss probabilities for the finite–buffer case can be derived from the infinite–buffer distribution.

4.1 Inverting the generating function

Daigle [22] has demonstrated an efficient method for calculating the probabilities p_n from a generating function $F^*(z) = \sum_{i=0}^{\infty} p_n z^n$ for variants of the $M/G/1$ queueing process. Daigle's method uses the discrete Fourier transform as follows. The characteristic function of the queue–length distribution can be expressed in terms of the generating function of a complex argument as the complex Fourier series

$$\Phi(\alpha) = F^*\left(e^{-i2\pi\alpha}\right) = \sum_{n=0}^{\infty} p_n e^{-i2\pi\alpha n},$$

with basis set

$$\phi_n(\alpha) = e^{-i2\pi\alpha n}, n = 0, \pm 1, \pm 2, \ldots.$$

Applying the inverse Fourier transform gives

$$p_n = \int_0^1 \Phi(\alpha)\overline{\phi_n}(\alpha)d\alpha.$$

Numerically this can be performed by calculating $\Phi(\alpha)$ at $L+1$ equi–spaced intervals of the interval $[0,1)$ and applying the inverse discrete Fourier transform to these values, resulting in

$$c_{n,L} = \frac{1}{L+1}\sum_{l=0}^{L} F^*\left(e^{-i2\pi\alpha l/(L+1)}\right) e^{i2\pi\alpha nl/(L+1)},$$

for $n \leq L$. Daigle [22] showed that

$$c_{n,L} = p_n + \sum_{m=1}^{\infty} p_{n+m(L+1)},$$

the non–equality of $c_{n,L}$ and p_n being referred to as 'aliasing'.

In principle this property can be used to approximate the probabilities p_n by increasing L until the tail probabilities are small enough. However round–off errors become important for large L, restricting the usefulness of the approximation, in particular for this application, where aliasing can have serious side–effects. Daigle's method relies on the property that the tail probabilities of the queue decrease geometrically for $M/G/1$ queueing systems, that is, for each $\varepsilon > 0$ there exists an N such that for all $n > N_\varepsilon$

$$|p_n - p_N * r^{n-N}| < \varepsilon.$$

With computational accuracy ε, choose $L > N_\varepsilon$. Daigle showed that

$$\begin{aligned} r_0 &= \frac{c_{0,L} - p_0}{c_{L,L}} = \frac{c_{0,L} - 1/m}{c_{L,L}}, \\ p_n &= \begin{cases} c_{n,L} - (c_{0,L} - p_0)\, r_0^n, & 1 \le n \le L, \\ p_K r_0^{n-K}, & n > L. \end{cases} \end{aligned}$$

Daigle provided a simple method for choosing L by calculating

$$\begin{aligned} r_{n,L} &= \frac{c_{n,L}}{c_{n-1,L}}, \qquad \forall n : N_\varepsilon < n \le L \\ a_L &= \max_{N_\varepsilon < n \le L} \left| \frac{r_0 - r_{n,L}}{r_0} \right|. \end{aligned}$$

The calculations of queue–length distributions based on this method were written in C++ using a free matrix library called `NEWMAT` [23], which included Fast Fourier Transform code, and the code used for matrix inversion.

4.2 The scenarios

The threshold values used, taken from realistic values given in Rumsewicz and Smith [2], are shown in Table 1. These two sets of thresholds require inversion of 62×62 and 100×100, arrays respectively. The inversion can be made significantly easier by taking advantage of the upper Hessenberg [20] form of the matrix $\mathbf{I} - \mathbf{P}_{K_o}$, but for our purposes, computation time being relatively unimportant, it was sufficient to use a standard inversion routine based on the QR decomposition of the matrix.

Threshold	Set 1	Set 2
Abatement	50	90
Onset	62	100

Table 1: Congestion Threshold Settings.

Both overload, and standard load scenarios were investigated in order to compare the behavior of the overload scheme. In all cases the throttling factor was 50%, that is, when in the congested regime the arrival rate was decreased by 50%. Four service time distributions:

1. the negative exponential distribution,
2. the Erlang-5 distribution,
3. the Erlang-20 distribution,
4. the deterministic distribution,

were examined and compared. For simplicity unit mean service times were used in all cases.

4.3 The number of messages in the buffer

Figures 1(a) and (b) show the results of applying the algorithms with the first set of thresholds from Table 1 for the three overload scenarios $\rho_u = 1.2$, 1.5 and 1.8 and the two non–overload cases $\rho_u = 0.8$ and 1.0, with exponential service times. Figure 2 shows what happens when the second set of thresholds are used for the overload cases with $\rho_u = 1.2$ and 1.8.

The effect of applying the overload control to the standard 0.8 load scenario is negligible. The net result of applying the overload control to the overload scenarios is to isolate the probability mass between the two thresholds, with a geometric drop off outside the immediate region surrounding the thresholds. This behavior exactly matches what one might expect - Remark 1 notes the similarity of the generating function being investigated to that of the standard $M/G/1$ queue which exhibits this sort of geometric tail. The fact that tail behavior of the queue is similar to that in the $M/G/1$ queue makes setting the size of the buffer, in the finite–state case, a reasonably simple task.

Furthermore, the behavior of queue under this type of control matches the requirements of such a control, namely

- it does not effect normal performance significantly and
- under overload it limits the extent of excursions to large queue sizes.

Figure 3 compares the behavior of the queue when the service–time distribution varies through exponential, Erlang–5 and Erlang–20 to a deterministic distribution, whilst keeping the mean service time constant. As $n \to \infty$ the Erlang–n distribution approaches the deterministic distribution, a fact illustrated in the figure. Furthermore the results demonstrate the applicability of the methods for distributions other than exponential.

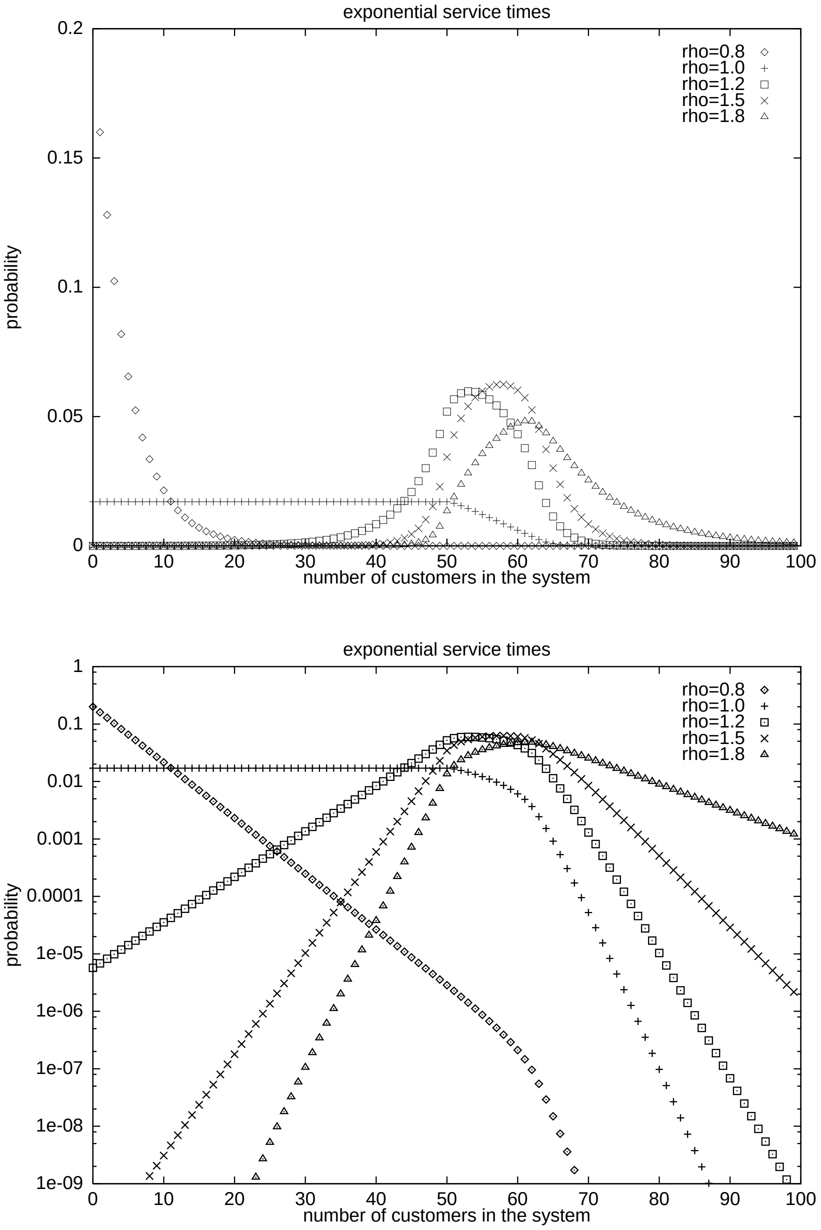


Figure 1: The queue–length distribution with $K_a = 50, K_o = 62$ and exponential service times. The first graph shows the probabilities on a linear axis, the second a log probability graph.

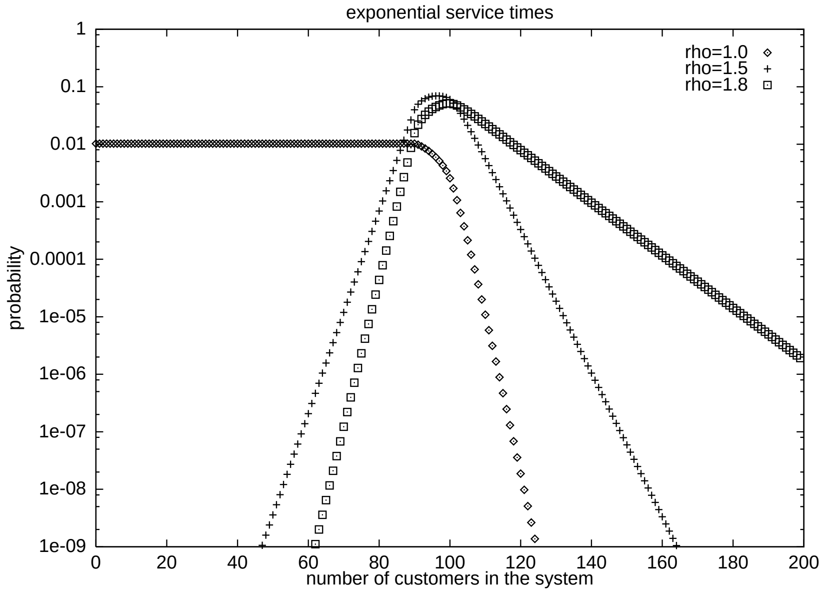


Figure 2: The queue–length distribution with $K_a = 90, K_o = 100$ and exponential service times.

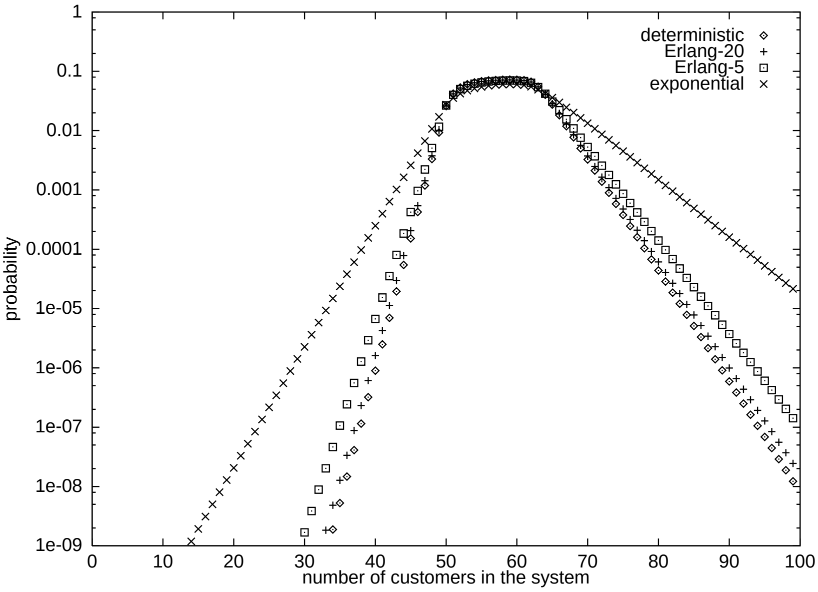


Figure 3: A comparison between the queue–length distributions for exponential, Erlang–5, Erlang–20 and deterministic service times, with $K_a = 50, K_o = 62$, $\rho_u = 1.6$.

4.4 Simple performance measures

As noted in Sections 3.4 and 3.5 there are several simple performance measures which may be used to assess the behavior of the queueing system. Two reciprocal measures are the probability $p(0)$ that the system is empty and the mean number m of customers served in one busy period, large values of m corresponding to high system utilizations. Figure 4 shows $\log(m)$ for a range of scenarios. The independent variable chosen here was the abatement threshold, given a constant onset threshold. Note the large values of m, leading to $p(0)$ being very small (< 0.03).

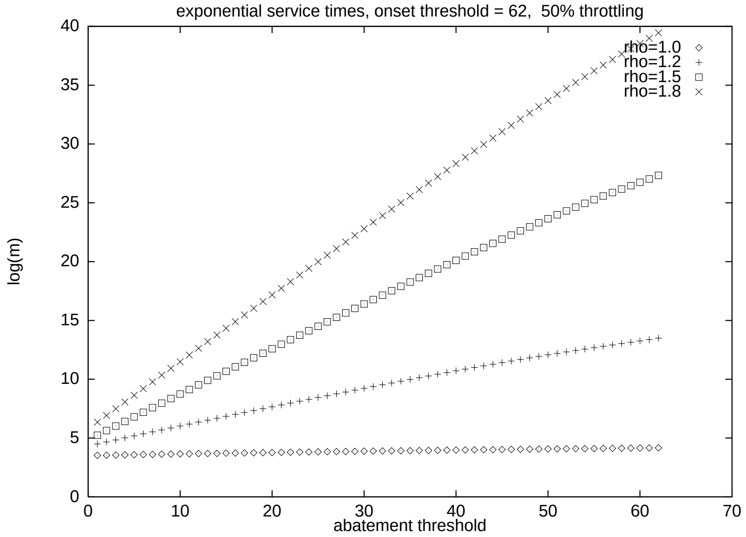


Figure 4: The log of the average length of the busy period for different abatement threshold values.

The accepted traffic load is given by (14). Figure 5 shows this performance measure. Notably it is near unity for all but the smallest values of K_a in all the overload scenarios. Hence nearly the maximum possible number of messages is being accepted by the server, a desirable result.

We use (15) to calculate the probability of the queue being congested. Figure 6 illustrates its value over a range of abatement thresholds and displays marked insensitivity to the abatement threshold.

The insensitivity of these results to the value of K_a is important, because it means that K_a can be set to achieve other performance goals, such as minimizing the number of congestion status switching events, with almost no cost in terms of increased loss rates, or a larger proportion of time spent in congestion.

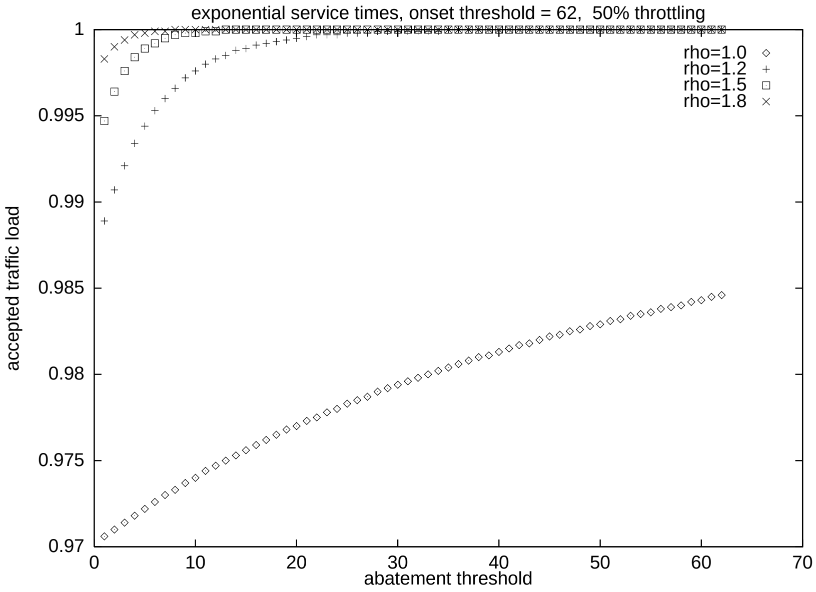


Figure 5: The accepted traffic load for different abatement threshold values.

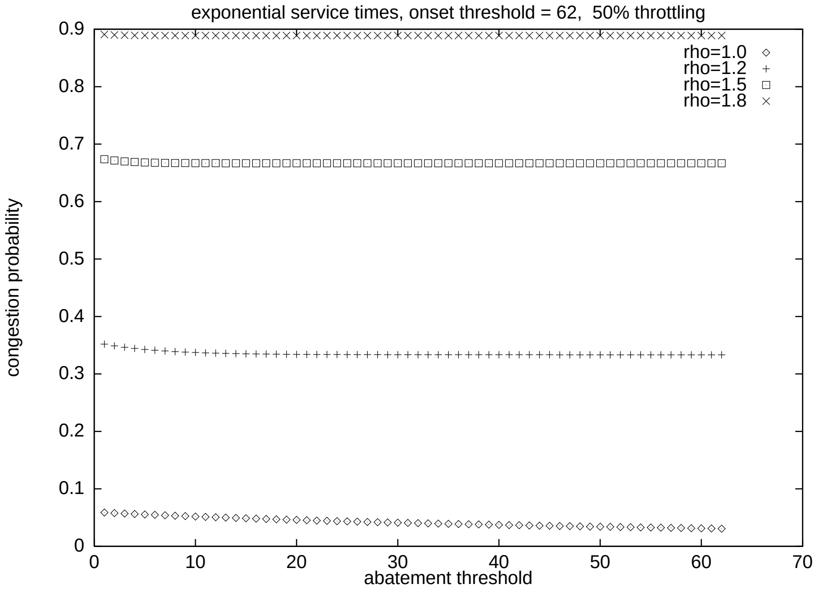


Figure 6: The probability of the queue being in the congested regime for different abatement threshold values.

4.5 The time between switching regimes

Section 3.6 provides two ways of estimating the length cycle between uncongested and congested regimes. The first is to estimate the mean time from crossing the congestion onset threshold until recrossing the congestion abatement threshold. The estimate given by (16) is illustrated in Figure 7 for a number of the scenarios described above. The principal feature is that the value is almost linearly dependent on the congestion abatement threshold. Hence the time between switching can be increased by decreasing the congestion abatement threshold. The slope is determined by the traffic intensity in the congested regime, and seems to be insensitive to the service-time distribution, as displayed by the similarity between Figures (a) and (b). In fact in the examples displayed the slopes can be neatly approximated by

$$s = \frac{1}{1 - \rho_c}.$$

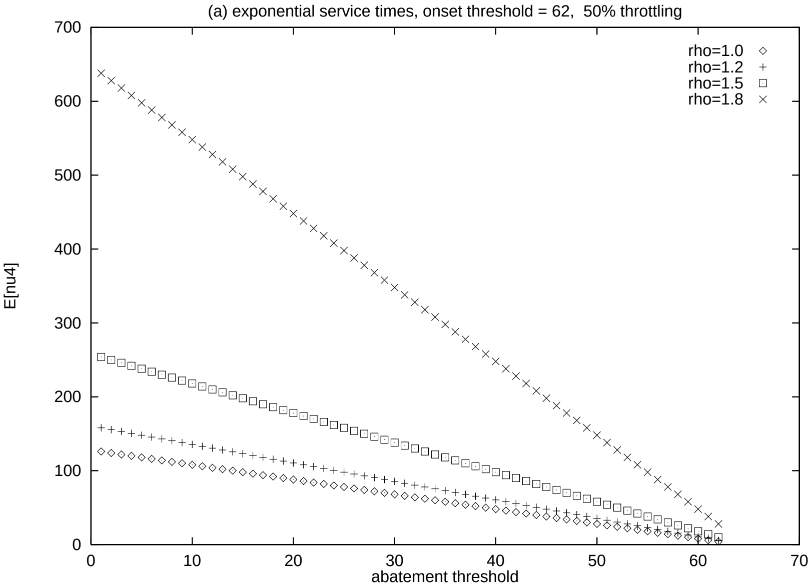
The second estimate of the cycle time is given by $E[\nu]$ which directly estimates the mean cycle time. The result given in Theorem 5 is illustrated in Figure 8 for the same range of scenarios as shown in Figure 7. Again cycle time increases with decrease in abatement threshold, but in this case the increase is only linear for the overload scenarios. The scenario with offered load $\rho_u = 1.0$ has a long cycle time that is not linearly dependent on the threshold because its behavior during the uncongested phase is that of a mean–zero random walk, while in the overload scenarios the behavior is that of a random walk with drift.

Again the behavior seems to be insensitive to the service–time distribution. We should however note that the order of the overload scenarios, in terms of cycle length, is different for this statistic.

5 Conclusion

Obviously the model analyzed here does not encapsulate all of the features used in overload controls, and in particular SS7 congestion controls, nor is it intended to. The aim was to study the behavior of the hysteretic overload control mechanism. Such controls are of recent interest [3] due to the need to provide overload controls in broadband networks. This paper provides some key results describing the behavior of a queue using this control: the PGF of the queue–length distribution, the probability of the queue being congested, the traffic load accepted by the system and the time between onset and abatement of congestion.

These results have been used to show quantitatively that the control behaves as desired – limiting excursions to long queue lengths during overloads with little impact under normal loads.



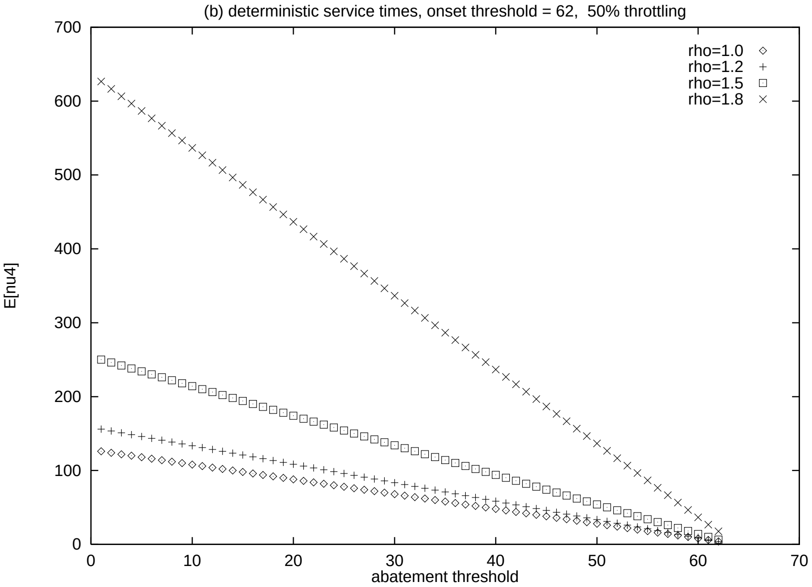


Figure 7: The mean time between onset and abatement of congestion.

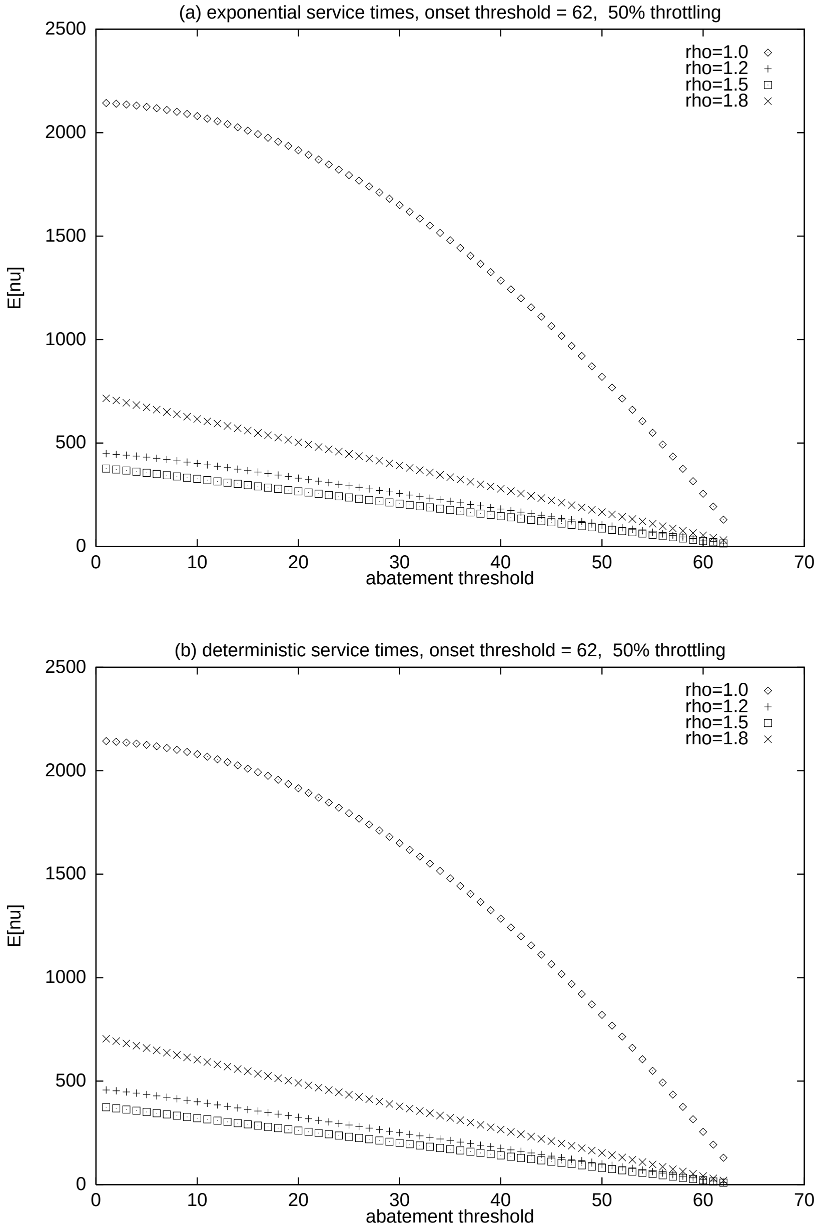


Figure 8: The mean cycle length for a cycle from congestion back to a new phase of congestion.

Intuitively, the reason for introducing a second distinct threshold for measuring the abatement of congestion separately from the onset of congestion is that the congestion cycle time will increase with increasing separation between the onset and abatement thresholds. This paper demonstrates that this is indeed the case, and provides a direct method for estimating the increase in cycle time.

The closed–form nature of the results makes them applicable to finding optimal threshold settings. Additionally, the results are also applicable to so called heavy–tailed distributions such as the Pareto distribution which have been receiving recent interest [24] for modeling packet traffic. These distributions may have infinite variance making many methods for calculating solutions inappropriate. Future work will examine these extensions.

Acknowledgment

M. Roughan conducted the majority of this work during his PhD at the University of Adelaide, supported by an Australian Postgraduate Research Award and a scholarship from the Teletraffic Research Centre at the University of Adelaide.

REFERENCES

[1] W. Stallings, *ISDN and Broadband ISDN with Frame Relay and ATM.* Prentice Hall, third ed., 1995.

[2] M. P. Rumsewicz and D. E. Smith, “A comparison of SS7 congestion control options during mass call-in situations,” *IEEE/ACM Transactions on Networks*, vol. 3, pp. 1–9, Feb 1995.

[3] K. K. Leung, “Load-dependent service queues with applications to congestion control in broadband networks,” in *IEEE Global Telecomunications Conference, GLOBECOM’97*, pp. 1674–79, 1997.

[4] N. Yin, S. Li, and T. Stern, “Congestion control for packet voice by selective packet discarding,” in *IEEE Global Telecomunications Conference, GLOBECOM’87*, (Tokyo, Japan), pp. 1782–1786, 1987.

[5] J. Morrison, “Two-server queue with one server idle below a threshold,” *Queueing systems*, vol. 7, pp. 325–336, 1990.

[6] W.-B. Gong, A. Yan, and C. G. Cassandras, “The M/G/1 queue with queue-length dependent arrival rate,” *Commun. Statist.-Stochastic Models*, vol. 8, no. 4, pp. 733–741, 1992.

[7] D. Perry and S. Asmussen, "Rejection rules in the M/G/1 queue," *Queueing Systems*, vol. 19, pp. pp 105–130, 1995.

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

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

[10] S.-Q. Li, "Overload control in a finite message storage buffer," *IEEE Transactions on Communications*, vol. 37, no. 12, pp. 1330–1338, 1989.

[11] W. Rosenkrantz, "Calculation of the Laplace transform of the length of the busy period for the M/G/1 queue via martingales," *Annals of Probability*, vol. 11, no. 3, pp. 817–818, 1983.

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

[13] F. Baccelli and A. Makowski, "Dynamic,transient and stationary behaviour of the M/GI/1 queue via martingales," *Annals of Probability*, vol. 17, no. 4, pp. 1691–1699, 1989.

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

[15] M. Roughan, *An application of martingales to queueing theory.* PhD thesis, University of Adelaide, Department of Applied Mathematics, 1994.

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

[17] C. Pearce and M. Roughan, "Forward delay times in multi-phase discrete-time renewal processes," *Asia-Pacific Journal of Operations Research*, vol. 14, no. 1, pp. 1–10, 1997.

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

[19] I. Gradshteyn and I. Ryzhik, *Table of integrals, series and products.* Academic Press, corrected and enlarged ed., 1980.

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

[21] R. Wolff, “Poisson arrivals see time averages,” *Opns. Res.*, vol. 30, pp. 223–231, 1982.

[22] J. N. Daigle, “Queue length distributions from probability generating functions via discrete Fourier transforms,” *Operations Research Letters*, vol. 8, pp. 229–236, August 1989.

[23] R. Davies, *Documentation for NEWMAT08A, A Matrix Library in C++.* robert.davies@vuw.ac.nz, 1995.

[24] M. E. Crovella and A. Bestavros, “Self-similarity in World Wide Web traffic: Evidence and possible causes,” *IEEE/ACM Transactions on Networking*, vol. 5, December 1997.

Appendix A Random walks

We consider a random walk $(X_n)_0^\infty$ on the nonnegative integers and associated stopping time

$$T = \inf\{n > 0 | X_n \in A^c\},$$

where A is a proper subset of the positive integers. The walk is prescribed by

$$X_n = X_0 + \sum_{i=1}^{n} Y_i, \ (n \geq 0)$$

with Y_i independent, integer–valued random variables given by

$$\begin{aligned} p\{Y_1 = m\} &= f_1(m), \text{ for } m \geq -1, \\ p\{Y_i = m\} &= f(m), \text{for } i > 1 \text{ and } m \geq -1. \end{aligned}$$

To ensure that the random walk makes no excursions to the negative integers before it is stopped, we assume that one of the regimes
(a) $X_0 \geq 0$ a.s. and $Y_1 \geq 0$ a.s.,
(b) $X_0 > 0$ a.s.
applies. We set

$$\begin{aligned} g_n(m) &= p\{T > n, X_n = m\}, \ (n, m \geq 0) \\ h_n(m) &= p\{T = n, X_T = m\}, \ (n > 0, \ m \geq 0). \end{aligned}$$

By the definition of T we have the boundary conditions

$$\begin{aligned} g_n(m) &= \begin{cases} p\{X_0 = m\}, & \text{if } n = 0 \\ 0, & \text{if } m \in A^c, n > 0, \end{cases} \\ h_n(m) &= 0, \text{ if } m \in A \quad (n \geq 1). \end{aligned}$$

We have also for $n > 0$ the recurrences

$$\begin{aligned} h_1(m) + g_1(m) &= \sum_k g_0(k) f_1(m-k), \\ h_{n+1}(m) + g_{n+1}(m) &= \sum_k g_n(k) f(m-k). \end{aligned}$$

For $|z| \leq 1$ we define the generating functions

$$\begin{aligned} F(z) &:= \sum_{k=0}^{\infty} f(k-1) z^k, \\ G_n(z) &:= \sum_{k=1}^{\infty} g_n(k) z^{k-1}, \quad (n > 0), \\ H_n(z) &= \sum_{k=0}^{\infty} h_n(k) z^k, \quad (n > 0). \end{aligned}$$

If regime (a) applies, we put (again for $|z| \leq 1$)

$$\begin{aligned} F_1(z) &:= \sum_{k=0}^{\infty} f_1(k) z^k, \\ G_0(z) &:= \sum_{k=0}^{\infty} g_0(k) z^k; \end{aligned}$$

otherwise if regime (b) applies, we put

$$\begin{aligned} F_1(z) &:= \sum_{k=-1}^{\infty} f_1(k) z^{k+1}, \\ G_0(z) &:= \sum_{k=1}^{\infty} g_0(k) z^{k-1}. \end{aligned}$$

Then under both regimes, our recurrence relations provide

$$H_1(z) + zG_1(z) = G_0(z) F_1(z),$$
$$H_{n+1}(z) + zG_{n+1}(z) = G_n(z) F(z), \text{for } n > 0.$$

Finally, for $|w| \leq 1$, set

$$\begin{aligned} G(w,z) &:= \sum_{n=0}^{\infty} G_n(z) w^n, \\ H(w,z) &:= \sum_{n=1}^{\infty} H_n(z) w^n. \end{aligned}$$

Forming generating functions again we derive

$$H(w,z) \quad = \quad G(w,z)\,[wF(z)-z] + zG_0(z) + w\,[F_1(z)-F(z)]\,G_0(z). \tag{19}$$

It is readily verified that the double series for $G(w,z)$ is absolutely convergent for $|z| \leq 1$, $|w| \leq 1$. The series for $H(w,z)$ is too, provided $E(T) < \infty$. We assume that this condition is satisfied.

We may choose our walk to represent a stable $M/G/1$–type system. In this context we have $E(Y_i) < 0$ for $i > 1$ and so $F'(1) < 1$. Since $F(1) = 1$, and $F(\cdot)$ is convex, we have $z/F(z) \in [0,1]$ for $z \in [0,1]$. If also $F_1(z) = F(z)$ then for $z \in [0,1]$ we can set $w = z/F(z)$ in (19) to derive

$$H(z/F(z),z) = zG_0(z). \tag{20}$$

When modeling the first phase of a busy period, we take $X_0 = 0$ so $G_0(z) = 1$, and $A = \{1,2,\ldots,K_o\}$ so that $T = \tau_1(0)$. Regime (a) applies with $F_1(z) = F(z) = a_u(z)$. Hence we derive $H(z/a_u(z),z) = z$, and so, given that by definition $H(w,z) = E\left[z^{X_T}w^T\right]$,

$$E\left[z^{X_T}(z/a_u(z))^T\right] = z.$$

This is an extension of the standard busy period result

$$E\left[(z/a(z))^T\right] = z,$$

where T is the time the busy period ends and hence $A = \{0\}$ and $X_T = 0$ a.s.

We may also model subsequent odd–numbered phases, that is, phases starting when the queue reaches the abatement threshold K_a. In this case we assume that the random walk begins at time $\tau_{2i}(0)$ in state $X_{\tau_{2i}(0)} = K_a$. Again, $A = \{1,2,\ldots,K_o\}$ so that $T = \tau_{2i+1}(0)$, the time when the onset threshold is exceeded or the busy period ends. As before $F_1(z) = F(z) = a_u(z)$, but now we are interested in $r(z)$ as defined in (8). In the present context this is

$$H(1,z) \quad = \quad G(1,z)\,[F(z)-z] + zG_0(z).$$

Regime (b) holds so that $G_0(z) = z^{K_a-1}$ and therefore

$$r(z) \;=\; E\left[z^{X_{\tau_{2i+1}(0)}}|X_{\tau_{2i}(0)} = K_a\right] \;=\; G(1,z)\,[a_u(z)-z] + z^{K_a}.$$

Here $G(1,z) = \sum_{n=0}^{\infty}\sum_{m=1}^{\infty} g_n(m)z^{m-1} = (1/z)\sum_{n=0}^{\infty}\mathbf{g}_n^T\mathbf{z}$, with

$$g_n(m) = p\{\tau_{2i+1}(0) > \tau_{2i}(0)+n, X_{\tau_{2i}(0)+n} = m|X_{\tau_{2i}(0)} = K_a\}.$$

Therefore

$$r(z) - z^{K_a} \quad = \quad \frac{a_u(z)}{z}\left[1-\frac{z}{a_u(z)}\right]\sum_{n=0}^{\infty}\mathbf{g}_n^T\mathbf{z}. \tag{21}$$