Analysis of a Hysteretic Overload Control

M. Roughan[a] * and C.E.M. Pearce[b]

[a]Software Engineering Research Centre, RMIT University,
Level 3, 110 Victoria St, Carlton, Vic 3053, Australia
E-mail: matt@serc.rmit.edu.au.

[b]Department of Applied Mathematics, University of Adelaide,
Adelaide 5005, AUSTRALIA.
E-mail: cpearce@maths.adelaide.edu.au.

Overload control is critical in preventing congestion of modern switching networks. One method, hysteretic overload control, uses two thresholds, a congestion onset, and a congestion abatement threshold to determine congestion status. 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. This paper provides an analytic technique for investigating the performance of such congestion controls, and thence setting the key 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 do you protect a modern switching network from overload? Answering this question has become critical to the reliable operation of a modern switching network, due to the increase in services with unpredictable traffic loads; for example, the Common Channel Signaling traffic associated with Intelligent Network services such as "Televoting". In essence, there are two related questions: how do you detect or measure the congestion caused by an overload, and how do you mitigate the effects of the overload. 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

*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.

considered congested if the number of messages in the Signaling Transfer Point (STP) link transmit buffer exceeds the onset threshold, and the link only returns to the uncongested state 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 date, 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, such as that 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] investigates 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. However, both of these examples are limited to systems in which the service times are exponentially distributed. Perry and Asmussen [7] consider 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, and more recently Leung [3] considers a system with the service-time distribution dependent on the workload in the system.

The papers above do not allow for two distinct thresholds leading to the hysteretic effect where the queue exhibits different behavior as the load increases from that as the load 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 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 a different arrival rate. The technique used to produce the results 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] also use similar arguments.

The principal contribution of this work derives from the result described in Theorem 1 which gives the probability generating function for the distribution of the number of customers in the system (as seen by an arriving customer). This paper makes a number of other original contributions, the chief of these being estimates of

- the probability of a customer arriving to a congested queue,
- the traffic load accepted by the system, and
- the time between onset and abatement of congestion.

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

This paper is organized as follows. Section II describes our overload control model, and provides analytical results. Section III provides numerical results for the performance measures listed above. Section IV summarizes the key results of the paper and suggests an extension to the work.

2. The model

This section provides a definition of our model of a buffer which uses hysteretic overload control. The model is a variant 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 M/G/1 queue is generalized to model the overload control by separating the behavior of the queue 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 times for the discarded packages were zero, then this model would be essentially the same as the source control model described above (from the point of view of arrivals). 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. Note that the case $K_a = K_o$ is included in the analysis described here, but that the case with $K_a > K_o$ makes little sense, and is not included.

Formally, 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, the process $X_n = X(t_n+)$, the number of customers in the system as seen by the nth departing customer. Cooper [17, pp. 154] shows that the arriving customers see the same queue length distribution as the departures. Note that, in practise 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 only be changed at the completion of a service and therefore depends only on the embedded queueing process X_n.

Arrivals are Poisson with rates λ_u and λ_c depending on the current congestion status. The service times are Independent Identically Distributed (IID) Random Variables (RV) with probability distribution function $G(\cdot)$, and mean $1/\mu$. The traffic intensities ρ_j are given by $\rho_j = \lambda_j/\mu$ for $j = u, c$.

We model the arrivals using two sequences of IID RV, A_n^j, $j = u, c$ and $n = 1, 2, \ldots$. Here A_n^u and A_n^c are respectively the number of customers to arrive during the nth service given that during this service the queue is considered to be uncongested or congested. The probability generating function for the number of arrivals during a service is $a_j(z) = \sum_{i=1}^{\infty} a_i^j z^i = \tilde{G}(\lambda_j[1-z])$, $j = u, c$, where $a_i^j = \text{prob}\{A_1^j = i\}$ and $\tilde{G}(\cdot)$ is the Laplace-Stieltjes transform of the service-time distribution [17].

2.1. Stability

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

For a proof of the stability conditions see [15]. The result can be easily understood 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 the behavior of the queue when uncongested, 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.

2.2. The queue-length distribution

We now provide the result which will be used to examine the behavior of the overload control considered here.

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

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

The mean number m, of customers served in one busy period, is given by

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

and the column vectors $\mathbf{e}_i = (\delta_{1i}, \delta_{2i}, \ldots, \delta_{K_o i})^T$ *and* $\mathbf{z} = (z, z^2, \ldots, z^{K_o})^T$.
Proof: For brevity the proof, which appears elsewhere [15] and [18], is omitted. □

Remark 1: The form of the solution is that of the Pollaczek-Khintchine Equation [17] for the probability generating function of the 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, but the matrix to be inverted, $(\mathbf{I} - \mathbf{P}_{K_o})$, is already in upper-Hessenberg form [19] 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 it applies equally well to packet discard models where the service-time distribution of discarded packets is changed. In which case $a_j(z) = \tilde{G}_j(\lambda[1 - z])$, where $G_j(\cdot)$ is the service-time distribution during the congested phase.

2.3. Simple performance estimates

In this analysis we make some simple definitions. The *offered load* refers to the load offered to the system prior to any overload control, and is denoted by ρ_u. The *accepted load* ρ_a, is that part of the load accepted by the system after application of overload controls.

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 which is given by the processor utilization, while λ is the arrival rate to the system, and W is the mean service time. The processor utilization is one minus $1/m$, the probability of the system being empty. The arrival rate times the mean service time is the accepted load ρ_a. Thus

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

To calculate the proportion of time that the system spends in the congested state we note that when PT is applied the load on the system is reduced from ρ_u to ρ_c. Thus the accepted load on the system is $\rho_a = (1 - \psi)\rho_u + \psi\rho_c$, where ψ is the proportion of time the queue is seen (by arrivals) to be congested. In conjunction with Equation (1) this expression yields

$$\psi = \frac{1 + (\rho_u - 1) R_{K_o K_a}(1)}{1 + (\rho_u - \rho_c) R_{K_o K_a}(1)}. \tag{2}$$

2.4. 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. A simple estimate of the rate of oscillation is given by the number of customers served during a cycle from the start of a congested phase to the start of the next congested phase, whose mean value is given in the following theorem.

Theorem 2: *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 ν is the cycle time, and Theorem 1 defines m, h and h_1.

Proof: Again, due to space limitations, the proof is omitted; it appears in both [15] and [18]. □

3. Numerical results

This part of the paper describes some examples, and provides numerical results for these. The chief result given above is in the form of a probability generating function, and to obtain queue length distributions this must be inverted. Daigle [20] demonstrates an efficient method for inverting generating functions for variants of the M/G/1 queueing process using the Discrete Fourier Transform. Our calculations were written in C++ using a free matrix library called `NEWMAT` [21], which included Fast Fourier Transform code, and the code used for matrix inversion. 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.

The threshold values used, taken from realistic values given by Rumsewicz and Smith [2], are shown in Table 1.

Table 1
Congestion Threshold Settings.

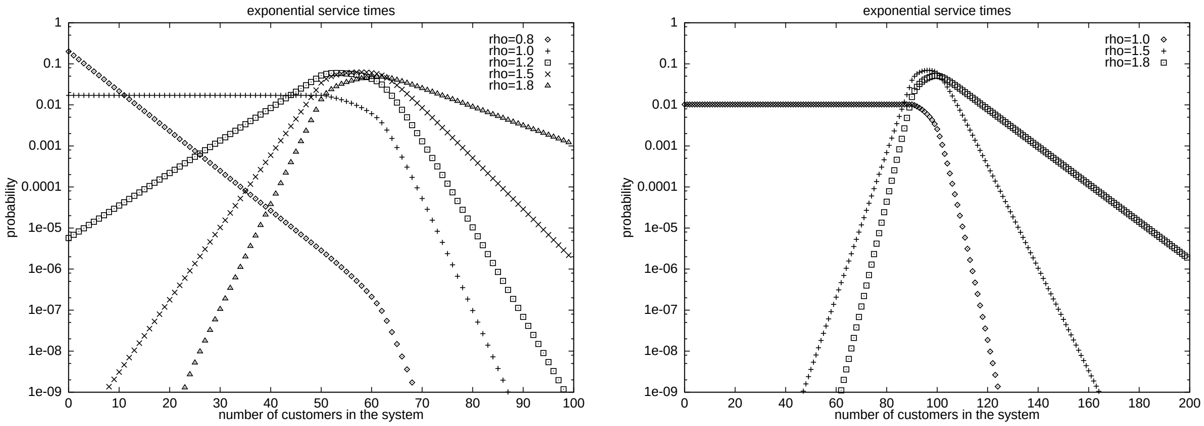
Threshold	Set 1	Set 2
Abatement	50	90
Onset	62	100

3.1. The number of messages in the buffer

The probability distribution for the number of messages in the buffer may be calculated using the method above applied to the probability generating function given in Theorem 1.

Figure 1(a) shows the results of applying the overload control 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 scenarios $\rho_u = 0.8$ and $\rho_u = 1.0$, with exponential service times, and 50% throttling in the congested regime. The two cases without overload provide a comparison to the overload cases. *Figure 1*(b) shows what happens when the second set of thresholds are used in the cases with $\rho_u = 1.0$, 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 this 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 geometric drop off can be seen in the figure by the straight line asymptotes of the curves. This behavior exactly matches what you 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 the tail behavior of the queue is similar to that in the M/G/1 queue makes setting the size of the buffer a reasonably simple task in the finite state case.



(a) $K_a = 50, K_o = 62$

(b) $K_a = 90, K_o = 100$

Figure 1. The queue length distribution with exponential service times.

Note that, importantly, the behavior of a queue under this type of control matches the requirements of such a control, namely,

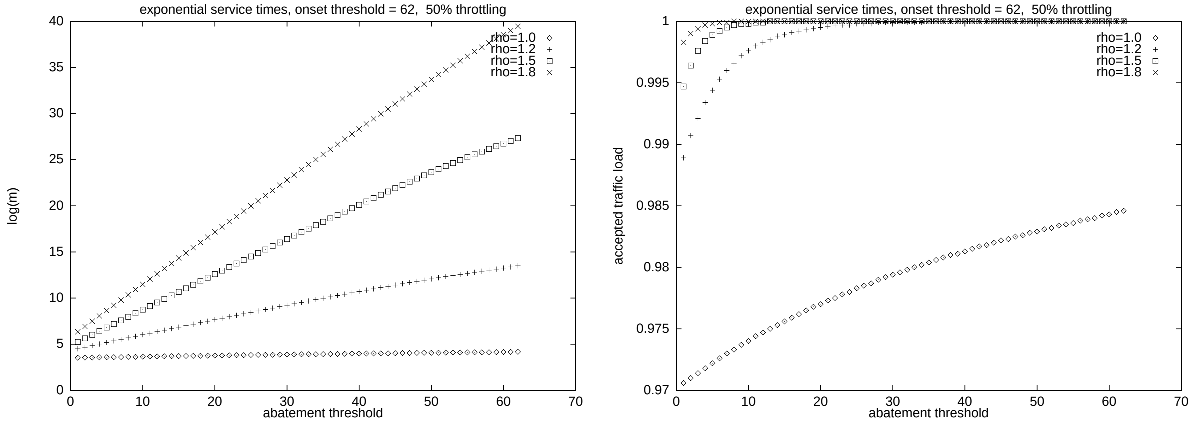
- it does not significantly effect normal performance,
- it limits excursions to large queue sizes.

We have also studied the queues behavior with more complex service-time distributions such as the Erlang distribution with similar observations (though space prevents presentation of these results).

3.2. Other performance measures

As noted in Sections 2.2 and 2.3 there are several simple performance measures which may be used to assess the behavior of the queueing system. Two reciprocal measures are the probability that the system is empty $p(0)$ and the mean number of customers m served in one busy period. *Figure 2*(a) shows the values of $\log(m)$ for offered loads $\rho_u = 1.0, 1.2, 1.5$ and 1.8, exponential service times, 50% throttling and an onset threshold set to 62. The independent variable chosen here was the abatement threshold, given a constant onset threshold; the reasons for choosing this will become clear. The large values of m correspond to very high server utilization.

Equation (1) gives the accepted traffic load. *Figure 2*(b) shows this performance measure. Notably this performance measure is near 1 for all of the overload scenarios. Hence, nearly the maximum possible number of messages are being accepted by the server, a desirable result. The insensitivity of this result to the value of K_a is also important, because



(a) The log of the average length of the busy period for different abatement threshold values.

(b) The average accepted traffic load for different abatement threshold values.

Figure 2. Simple performance measures

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.

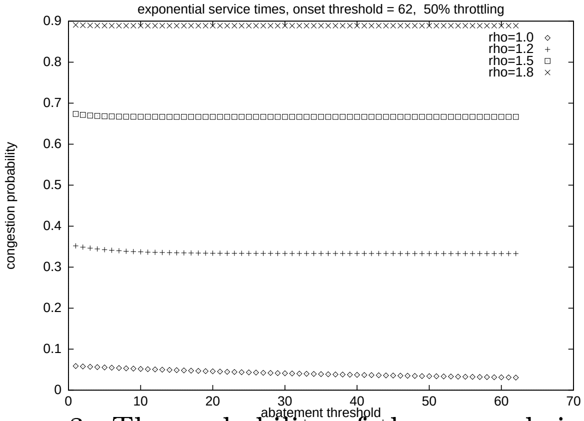


Figure 3: The probability of the queue being in the congested regime for different abatement threshold values.

Equation 2 gives an expression for the probability of the queue being congested. *Figure 3* illustrates this probability over a range of abatement thresholds. We can note that for the overload scenarios, the results are quite insensitive to the value of the abatement threshold.

Section 2.4 provides an estimate of the cycle length between the uncongested and congested regimes. The estimate of the cycle time is given by $E[\nu]$ which directly estimates the mean cycle time. The result given in Theorem 2 is illustrated in *Figure 4*. Note in particular a linear increase in cycle time with decrease in abatement threshold for the overload scenarios. In the case with offered load $\rho_u = 1.0$ the cycle time is not linearly dependent on the threshold because the 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.

From comparison between *Figures 4* (a) and (b) which show the same scenarios for exponential, and deterministic service times respectively, we can note that the mean cycle time seems to be insensitive to the service-time distribution.

4. Conclusion

Obviously the model analyzed here does not encapsulate all of the features used in overload controls, in particular SS7 congestion controls; nor is it intended to. The aim was to study the hysteretic overload control mechanism. Such controls are of recent

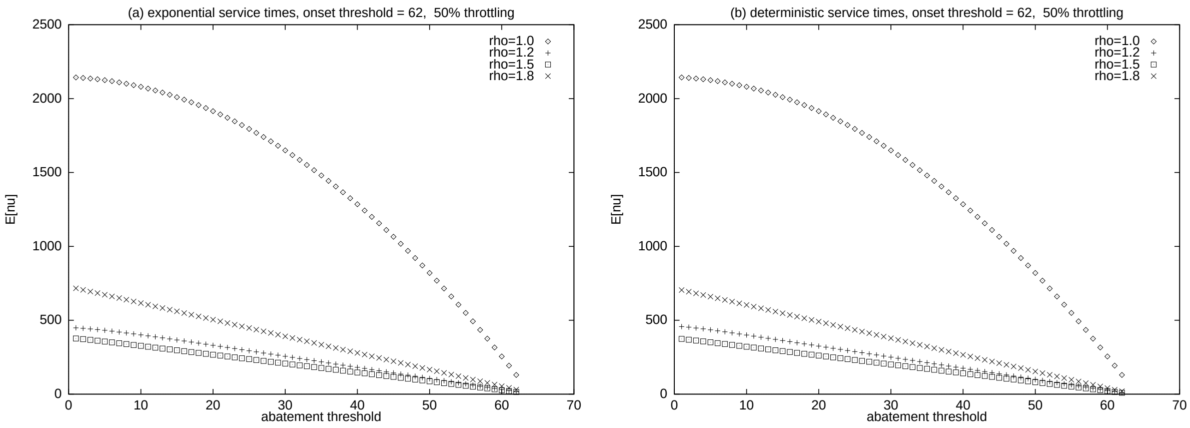


Figure 4. The mean cycle length for a cycle from congestion, back to a new phase of congestion.

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 (as seen by an arriving customer), the traffic load accepted by the system, and the time between onset and abatement of congestion.

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

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 [22] for modeling packet traffic. These distributions may have infinite variance making many methods for calculating solutions inappropriate. Future work is intended to examine these extensions.

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. A. Y. Wei-Bo Gong 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. Wolff, "Poisson arrivals see time averages," *Opns. Res.*, vol. 30, pp. 223–231, 1982.
17. R. Cooper, *Introduction to Queueing Theory.* The Macmillian Company, 1972.
18. M. Roughan and C. Pearce, "A martingale analysis of hysteretic overload control applied to the M/G/1 queue.," Tech. Rep. SERC-0059, Software Engineering Research Centre, RMIT University, Level 2, 723 Swanston St., Carlton, Vic. 3053, AUSTRLIA, February 1998.
19. G. Golub and C. van Loan, *Matrix Computations.* North Oxford Academic, 1983.
20. J. N. Daigle, "Queue length distributions from probability generating functions via discrete Fourier transforms," *Operations Research Letters*, vol. 8, pp. 229–236, August 1989.
21. R. Davies, *Documentation for NEWMAT08A, A Matrix Library in C++.* robert.davies@vuw.ac.nz, 1995.
22. 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.