Proceedings of ICON '99, Brisbane, Queensland, Australia, 28 September – 1 October, 1999, pp 53 – 58.

Application of the M/Pareto Process to Modeling Broadband Traffic Streams

Timothy D. Neame, Moshe Zukerman
The University of Melbourne
Parkville, Vic. 3052, Australia.
Email: {t.neame, m.zukerman}@ee.mu.oz.au

Ronald G. Addie
University of Southern Queensland
Toowoomba, Qld. 4350, Australia.
Email: addie@usq.edu.au

Abstract

In this paper we examine the usefulness of the M/Pareto process as a model for broadband traffic. We show that the queueing performance of the M/Pareto process depends upon the level of aggregation in the process. When the level of aggregation is high, the M/Pareto converges to a long range dependent Gaussian process. For lower levels of aggregation, the M/Pareto is capable of modeling the queueing performance of real broadband traffic traces.

1 Introduction

Broadband traffic types, such as VBR video and data streams, have only appeared relatively recently. As yet, no consensus has been reached regarding how we should model these traffic types. It is not reasonable to attempt to accurately represent all the intricacies of these traffic types in a single model. What is required is a model that reasonably represents real traffic, in some practical sense, that will be widely acceptable and will provide consistency in performance comparisons of real networks and systems. For a model which is to be used as this type of tool, ease of use is more important than complete accuracy. For most engineering purposes, a matching in terms of queueing performance is sufficient.

Once a consensus is reached on such a model, it could be used in a wide variety of network planning, dimensioning and management related applications. Depending upon the complexity of the model, it could be applied to tasks including call admission control (CAC), network dimensioning, performance evaluation and as an input to business decisions. Such a model would also simplify the evaluation of new equipment or protocols.

The process of determining the most appropriate model for broadband traffic is complicated by recent studies of broadband traffic traces which have shown that broadband traffic is often long range dependent (LRD) in nature. This property manifests itself as long bursts of higher than average activity. Long range dependence has been shown in both data streams [14, 16] and VBR video streams [10]. This fact tells us that traditional Markovian models cannot be extended into the broadband domain, as they are not able to accurately reflect the behaviour of LRD traffic streams. In response, many new models have been proposed [5, 12, 13, 18, 20] for LRD traffic.

In this paper we examine the ability of the M/Pareto process to meet our needs as a model for broadband traffic streams. In Section 2 we explain our technique for evaluating a traffic model, and define the queueing framework used throughout this paper. In Section 3 we give a brief description of the M/Pareto model used. In Section 4 we describe methods which can be used to generate multiple M/Pareto processes all with the same mean, variance and Hurst parameter, but with differing levels of aggregation, and which yield different queueing results. In Section 5 we demonstrate that the behaviour of an M/Pareto process converges to that of a fractal Gaussian process for increasing levels of aggregation. Section 6 presents results showing that, where the level of aggregation is chosen appropriately, the M/Pareto model can accurately model realistic broadband traffic sources. In Section 7 we show that the correct level of aggregation required for the M/Pareto model to match the modeled traffic appears to depend on the service rate being considered as well as the properties of the traffic stream being modeled.

2 Modeling approach

There are two broad categories of traffic models. One alternative is to consider each packet individually. The random process used as a traffic model generates values representing the arrival times of individual packets (or equivalently the inter-arrival times between successive packets). The other alternative is to consider the rate at which work arrives at the buffer. Although there is a healthy body of work which considers this arrival process within a continuous time framework, (see [19] and references therein) we focus on the discrete time case. Time is broken into fixed length intervals, and we model the number of packets arriving in each interval.

We choose a discrete time arrival process because it makes it possible to carry out longer simulations than would be possible using inter-arrival times. Considering this arrival process does result in a loss of resolution, as it becomes impossible for us to identify precisely when a packet arrived, but provided the interval size is not too large, the resulting errors are generally small.

The usefulness of a given model is judged based on its ability to characterise the buffer overflow probability of measured traffic in a single server queue (SSQ). The overflow probability is the probability that the length of the queue of unfinished work in an infinite buffer SSQ exceeds a given threshold. In monitoring of network performance the cell loss ratio (CLR) is more commonly used to measure queueing performance. CLR is the proportion of the cells arriving at a finite buffer queue which are discarded due to the buffer being full. Overflow probability is easier to deal with analytically, while CLR is simpler to measure in practical circumstances. Although the two values are not identical, overflow probability gives a good estimate of CLR in most cases.

To calculate the overflow probabilities for a given traffic stream, we consider a FIFO single server queue with an infinite buffer. We consider time to be divided into fixed length sampling intervals. The model allows arbitrary choice of interval length.

Let A_n be a continuous random variable representing the amount of work entering the system during the nth sampling interval. We assume that the process $\{A_n\}$ is both stationary and ergodic. We define τ to be the constant service rate, i.e. it is a fixed number representing the amount of work which can be processed by the server per sampling interval. We assume for simplicity that the service takes place at the end of the interval. Let the mean of A_n be denoted by μ and its variance by σ^2.

Let the sequence of continuous random variables Y_n be the net input process defined as

$$Y_n = A_n - \tau, \qquad n \geq 0$$

and let m be the mean of the net input, that is

$$m = \mathrm{E}(A_n) - \tau = \mu - \tau.$$

A necessary and sufficient condition for queueing stability is $m < 0$. Since τ is constant, the variance of the unfinished work process is equal to that of the arrival process. That is: $\sigma^2 = \mathrm{Var}(A_n) = \mathrm{Var}(Y_n)$.

Let V_n be the unfinished work at the beginning of the nth sampling interval. Using the above notation, the system unfinished work process, for the case of an infinite buffer, satisfies Lindley's recurrence equation:

$$V_{n+1} = (V_n + Y_n)^+, \qquad n \geq 0, \qquad (1)$$

where $V_0 = 0$ and where $X^+ = \max(0, X)$. The steady state buffer overflow probability, $\Pr(Q > t)$, is simply the probability that the amount of unfinished work in the queue exceeds a given threshold t, i.e $\Pr(Q > t) = \Pr(V_\infty > t)$.

The correlations in an LRD arrival process are represented using the Hurst parameter, H. The Hurst parameter is related to the rate at which the correlations decrease with increasing lag. The higher the value of H, the higher the level of correlation and consequently the worse the queueing performance. $H \in (0.5, 1.0]$ in positively correlated LRD traffic streams. $H = 0.5$ for traffic without LRD (short range dependent, or SRD traffic). Several methods for evaluating H in a given traffic stream are given in [9].

3 The M/Pareto model

It is now widely accepted that LRD traffic forms a significant part of the traffic to be carried over broadband networks. LRD traffic, regardless of its source, is characterised by significant long bursts (see [19] and references therein). It is therefore appealing to model LRD traffic with a model which involves long bursts. The M/Pareto model is just such a process, generating an arrival process based on overlapping bursts. The M/Pareto model described below is closely related to that given in [15], and is one of a family of such processes which form a sub-group of the more general $M/G/\infty$ models explored in [13, 18].

M/Pareto traffic is composed of a number of overlapping bursts. Bursts arrive according to a Poisson process with rate λ. The duration of each burst is random, and chosen from a Pareto distribution. The complementary distribution function for a Pareto-distributed random variable is given by

$$\Pr\{X > x\} = \begin{cases} \left(\frac{x}{\delta}\right)^{-\gamma}, & x \geq \delta, \\ 1, & \text{otherwise,} \end{cases}$$

$1 < \gamma < 2, \delta > 0$. The mean of X is $\frac{\delta\gamma}{(\gamma-1)}$ and the variance of X is infinite.

The rate of the Poisson process, λ, controls the frequency with which new bursts commence. The superposition of two independent M/Pareto processes with identical burst length distributions will itself be an M/Pareto process with Poisson arrival rate equal to the sum of the arrival rates of the two constituent processes. Thus, increasing λ can be considered to represent an increase in the number of sources which make up an M/Pareto stream.

The cell arrival process for each burst is constant for the duration of that burst, and has rate r. All bursts generate cells at the same rate r. Thus the mean number of cells within one burst is: $\frac{r\delta\gamma}{\gamma-1}$. The mean amount of work arriving within an interval of length t in the M/Pareto traffic model is $\frac{\lambda t r\delta\gamma}{(\gamma-1)}$.

Although the Pareto process has infinite variance, the variance of the M/Pareto process is finite. In [19] the term "Poisson burst process" was used to refer to processes, such as the M/Pareto process, where i.i.d. bursts of fixed rate start according to a Poisson process. For a Poisson burst process the variance function is given by repeatedly integrating the distribution function, according to

$$\sigma^2(t) = 2\lambda r^2 \int_0^t dt \int_0^u du \int_v^\infty dx \Pr\left\{X > x\right\}$$

Calculating for Pareto distributed burst durations gives

$$\sigma^2(t) = \begin{cases} 2r^2\lambda t^2 \left(\frac{\delta}{2}\left(1 - \frac{1}{1-\gamma}\right) - \frac{t}{6}\right), & 0 \le t \le \delta \\ 2r^2\lambda \left\{\delta^3 \left(\frac{1}{3} - \frac{1}{2-2\gamma} \right.\right. & \\ \qquad \left. + \frac{1}{(1-\gamma)(2-\gamma)(3-\gamma)}\right) & \\ \quad +\delta^2 \left(\frac{1}{2} - \frac{1}{1-\gamma} \right. & \\ \qquad \left. + \frac{1}{(1-\gamma)(2-\gamma)}\right)(t-\delta) & \\ \quad \left. - \frac{t^{3-\gamma}}{\delta^{-\gamma}(1-\gamma)(2-\gamma)(3-\gamma)}\right\}, & t > \delta \end{cases} \tag{2}$$

This corresponds with the variance function for processes of this type given in [2]. It represents a correction to the variance function quoted in [3, 7, 17].

Examining the expression for the variance, we see that for large t, the dominant term is $2r^2\lambda\frac{\delta^\gamma t^{3-\gamma}}{(1-\gamma)(2-\gamma)(3-\gamma)}$. If we define $H = \frac{3-\gamma}{2}$ then we can observe that for increasing t the growth of this function is proportional to t^{2H}. This implies that this model is *asymptotically self similar* with Hurst parameter $H = \frac{3-\gamma}{2}$.

4 Changing the level of aggregation

We have just seen that the form of the M/Pareto model we have chosen has four parameters:

- the Poisson arrival rate, λ, which controls the arrival of bursts,
- the arrival rate within an active burst, r,
- the rate of decrease of the Pareto tail, γ, and
- the starting point of the Pareto tail, δ.

If we attempt to create an M/Pareto process which produces given values for the mean arrival rate, μ, variance, σ^2 and Hurst parameter, H, we will need to choose one of the four parameters of the M/Pareto arbitrarily. This suggests that there will be infinitely many M/Pareto processes which produce the same values of μ, σ^2 and H. Each of these M/Pareto processes will potentially produce different queueing performance.

Throughout this paper we examine families of M/Pareto processes with different Poisson arrival rates, but identical values of μ, σ^2 and H. There are a number of ways we could achieve this. In this work we limit ourselves to trading off between the Poisson arrival rate for bursts, λ, and the cell arrival rate contributed by each burst, r. The parameters controlling the burst duration (δ and γ) are held fixed. This method allows us to model a situation where an aggregated traffic stream has an increasing number of sources (represented by increasing λ) each contributing a smaller proportion of the overall traffic load (represented by decreasing r), but where the burst durations remain unchanged.

If we restrict ourselves to changing only λ and r, the Hurst parameter, H, will be unaffected by any changes. In order to maintain a constant value for the variance we utilise the relationship given in Equation (2), and so if λ is multiplied by a factor n, then the transmission rate for each burst is reduced by dividing r by $\sqrt{n}$. Making these changes to λ and r will increase the mean arrival rate of the M/Pareto process, unless we permit ourselves to alter the burst length distribution.

We choose not to alter the burst length distribution, and look for other ways to deal with this change to the mean. When we examine queueing performance we can essentially ignore the problem, and simply alter the service rate so that m is still matched. The queueing performance will not be altered, as the overflow probabilities are determined by the mean of the net arrival rate $m = \mu - \tau$, not by the mean of the actual arrival process. Alternatively, if the mean of the M/Pareto process is important, we can introduce a constant bit rate (CBR) component, κ, to maintain a matching between the mean arrival rate of the M/Pareto process and the mean arrival rate of the modeled stream. The addition of κ cells per interval to every arrival interval will not affect the values of σ^2 or H, nor the queueing performance of the process.

5 A Gaussian future?

If traffic is Gaussian, a number of advantages become apparent. Most importantly bandwidth usage can be made more efficient. As [4] shows, significant multiplexing gains are possible with Gaussian traffic. Another key advantage is the wide range of analytical results which exist for Gaussian traffic [2, 5, 6, 11]. Appendix A gives analytic expressions for the overflow probabilities for both SRD and LRD Gaussian processes. Although a Gaussian model is attractive, previous measurements of network traffic [14, 16] have shown little to encourage the belief that real network traffic is Gaussian.

However, we have reason to believe that traffic may become more Gaussian in the future. As was suggested in [1], according to the central limit theorem, as the number of independent sources contributing to an aggregate flow increases, the total amount of traffic arriving in a fixed length

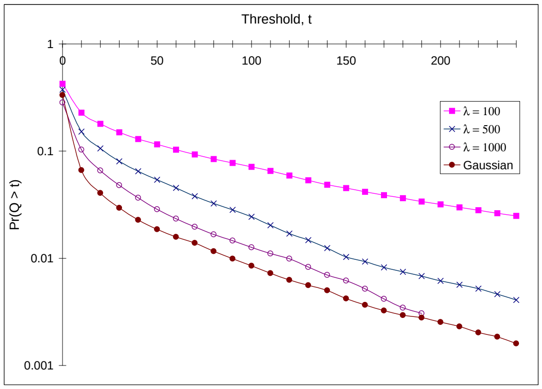

Figure 1. Gaussian convergence for increasing aggregation

interval will tend towards a Gaussian random variable. We can use the M/Pareto process to demonstrate this effect.

In Section 3 we related the number of independent sources contributing to an aggregated traffic stream to the Poisson arrival rate λ in the M/Pareto model. The central limit theorem predicts that as the level of aggregation (i.e. λ) increases, the behaviour of the M/Pareto process will approach that of a Gaussian process.

We showed in [7] that this Gaussian convergence does occur. Figure 1 shows an example in which we see a family of M/Pareto processes, all with the same values of m, σ^2 and H, but with differing levels of aggregation. Also shown is a Gaussian process, with the same values of m, σ^2 and H as the M/Pareto processes. As the value of λ increases the queueing performance improves, until a good approximation of Gaussian performance is achieved. Along the way lower values of λ produce different queueing performance results for M/Pareto processes with the same values of m, σ^2 and H.

6 Comparison with real traffic

Past attempts to match the mean, variance and Hurst parameter of a model like Gaussian fractal or M/Pareto [16] to that of the real traffic, and, using this matched process, attempting to predict the queueing curve (overflow probability versus threshold) have had only limited success. However, the previous section has shown that there are many M/Pareto processes which have the same mean, variance and Hurst parameter, but which yield different queueing curves. At most one of these processes will match the queueing curve of the real traffic. The differentiating factor is the value of the Poisson arrival rate, λ, which we have termed the "level of aggregation" in the process. Without

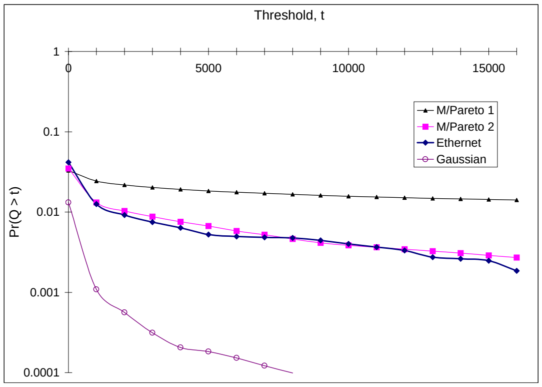

Figure 2. Matching with an Ethernet trace

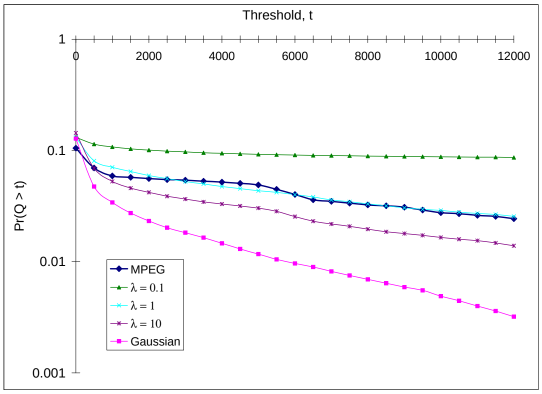

Figure 3. Matching with a VBR video trace

determining the correct value of this parameter, we cannot expect to achieve a match between the queueing curve of the M/Pareto process and that of the modeled traffic.

In [3] we showed that the M/Pareto model can accurately model the performance of an Ethernet trace provided that the right value for the parameter λ is selected. Figure 2 shows an example of this. In Figure 2, the curve designated as *M/Pareto 1* represents an M/Pareto process matched to the three parameters (m, σ^2 and H) of the real traffic, with an arbitrary choice of λ. The curve designated as *M/Pareto 2* represents a process which also produces the same values of m, σ^2 and H, but with a more careful choice of λ. The *Gaussian* curve in the figure shows that the Ethernet traffic stream cannot be modeled by a Gaussian process.

In [17] we showed that matching the queueing curves of VBR video streams is also possible using the M/Pareto process. Figure 3 shows an example in which we fit an M/Pareto process to a VBR MPEG sequence generated by Rose and analysed in [20]. Even though the modeled traffic is not an aggregated stream, the "level of aggregation"

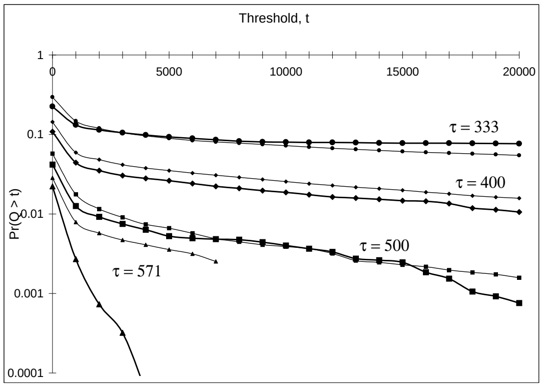

Figure 4. Altering the service rate

parameter, λ, must still be correctly chosen for an accurate modeling of the stream to be possible.

The M/Pareto model *is* capable of predicting the queueing curves of both aggregated data traffic and VBR video streams. Accurate fitting is achieved only when the level of aggregation parameter, λ is correctly assigned. Merely matching the mean, variance and Hurst parameter of the M/Pareto process to that of the original traffic stream is not sufficient accurately predict the queueing curve.

7 Model limitations

We have seen that choosing the right value for λ is vital in creating an M/Pareto process capable of matching the queueing curves of real traffic streams. However the choice of λ is complicated by the fact that the correct value of λ differs depending on the service rate τ (or equivalently the value of m).

In Figure 4, we show the queueing curves produced when a pair of traffic streams are fed into SSQs with a variety of services rates. The figure shows four pairs of curves. In each pair the heavier line represents the queueing performance of the Ethernet trace when fed into an SSQ with service rate τ. The lighter line represents the performance of an M/Pareto process matched to the properties of the Ethernet trace in an identical SSQ. The M/Pareto process used has λ chosen so as to provide a good fit with the Ethernet traffic when $\tau = 500$. As Figure 4 shows, while a given value of λ may give an acceptable fitting for a range of service rates, in general a different value of λ must be determined for each service rate considered.

Choosing the correct value for λ is not a trivial exercise. As yet we have no systematic method for determining the value of λ to be used in modeling a given traffic stream. Trial and error must be used for each different traffic source, and for each different service rate. In every case we have considered so far it has been possible to find a value of λ which is appropriate, but a systematic method would greatly accelerate this process. Until a heuristic for determining λ is developed, this will limit the practical usefulness of the M/Pareto process.

8 Conclusions

This paper has presented an examination of the M/Pareto process. We have shown that, provided we assign the correct values to all the parameters of the M/Pareto model, we can use an M/Pareto process to accurately predict the queueing performance of an arbitrary broadband traffic stream. For most practical purposes, this makes the M/Pareto process a good candidate as a model of broadband traffic streams. The M/Pareto model still has some limitations, but it has cleared the first and most vital hurdle, and is worthy of further consideration.

Appendix A: Gaussian Formulae

Consider a FIFO single server queue, with all the assumptions, definitions and notation of the discrete time queueing model described in Section 2. We now add a further condition and suppose that the arrival process, $\{A_n\}$, is not only stationary and ergodic but also Gaussian. We allow for any autocorrelation function for the arrival process. The process may be either SRD or LRD, but these cases are handled slightly differently.

A.1 The Short Range Dependent Case

For the SRD case, the tail of the overflow probability (or the unfinished work distribution) is exponential. By [5], it can be approximated by

$$\Pr\{V_\infty > t\} \approx \tilde{c}e^{s^*t}, t > 0, \tag{3}$$

where $s^* = \frac{2m}{v}$, $\tilde{c} = -s^*\psi(-m,\sigma)/\mathrm{erf}\big(-m\sigma/(v\sqrt{2})\big)$, and

$$\psi(x,\sigma) = \frac{\sigma}{\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}} - \frac{x}{2}\mathrm{erfc}\left(\frac{x}{\sigma\sqrt{2}}\right). \tag{4}$$

v is the asymptotic variance rate (AVR), given by

$$v = \lim_{k\to\infty}\frac{\mathrm{Var}\left\{\sum_{n=1}^{k} Y_n\right\}}{k}.$$

The result obtained for the rate of the tail s^* is exact while the result obtained for weight of the tail is an approximation.

A.2 The Long Range Dependent Case

For an LRD Gaussian process v is not finite and the unfinished work distribution does not have a dominant exponential tail. Fortunately, as shown in [6] (except that a blunder introduced an incorrect factor in that paper) we can apply the SRD results to the LRD case and obtain the following approximation for the overflow probability:

$$\Pr\{V_\infty > t\} \approx \frac{\sqrt{2\pi}}{\sigma}\psi(-m,\sigma)e^{s^*(t)t} \qquad (5)$$

in which

$$s^*(t) = -\frac{1}{2\sigma^2}|1-H|^{-2}\left(\frac{H}{|(1-H)m|}\right)^{-2H} t^{1-2H}.$$

Formula (5) is quite accurate so long as the Hurst parameter takes on values larger than 0.5.

References

[1] R. G. Addie. On the Weak Convergence of Long Range Dependent Traffic Processes. In *Proceedings of the International Workshop on Long Range Dependence*, January 1997. Accepted for publication in *Journal of Statistical Planning and Inference*.

[2] R. Addie, P. Mannersalo and I. Norros. Performance Formulae for Queues with Gaussian Input. In *Proceedings of ITC 16*, June 1999, pp 1169–1178.

[3] R. G. Addie, T. D. Neame and M. Zukerman. Modeling Superposition of Many Sources Generating Self Similar Traffic. In *Proceedings of ICC '99*, June 1999.

[4] R. G. Addie and M. Zukerman. Queues with Total Recall – Application to the B-ISDN. In Proceedings of ITC 14, 1994, pp 45–54.

[5] R. G. Addie and M. Zukerman. An Approximation for Performance Evaluation of Stationary Single Server Queues. *IEEE Transactions on Communications*, December 1994.

[6] R. G. Addie, M. Zukerman, and T. D. Neame. Fractal Traffic: Measurements, Modelling and Performance Evaluation. In *Proceedings of Infocom '95*, April 1995.

[7] R. G. Addie, M. Zukerman and T. D. Neame. Broadband Traffic Modeling: Simple Solutions to Hard Problems. *IEEE Communications Magazine*, Vol. 36, No. 8, August 1998.

[8] A. Arvidsson and P. Karlsson. On Traffic Models for TCP/IP. In *Proceedings of ITC 16*, June 1999, pp 457–466.

[9] J. Beran. *Statistics for Long-Memory Processes*. Chapman & Hall, New York, 1994.

[10] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger. Long-Range-Dependence in Variable-Bit-Rate Video Traffic *IEEE Transactions on Communications*, Vol. 43, No. 2/3/4, February/March/April 1995, pp 1566–1579.

[11] J. Choe and N. B. Shroff. New Bounds and Approximations using Extreme Value Theory for the Queue Length Distribution in High-Speed Networks. In *Proceedings of IEEE Infocom '98*, San Francisco, 1998, pp. 364–371.

[12] A. Karasiridis and D. Hatzinakos. A Non-Gaussian Self-Similar Process for Broadband Heavy Traffic Modeling. In *Proceedings of Globecom '98*, December 1998.

[13] M. M. Krunz and A. M. Makowski. Modeling Video Traffic Using $M/G/\infty$ Input Processes: A Compromise Between Markovian and LRD Models. *IEEE Journal on Selected Areas in Communications*, Vol. 16, No. 5, June 1998.

[14] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the Self-Similar Nature of Ethernet Traffic (Extended Version). *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, 1994, pp 1–15.

[15] N. Likhanov, B. Tsybakov and N. D. Georganas. Analysis of an ATM Buffer with Self-Similar ("Fractal") Input Traffic. In *Proceedings of Infocom '95*, April 1995.

[16] T. D. Neame, R. G. Addie, M. Zukerman, and F. Huebner. Investigation of Traffic Models for High Speed Data Networks. In *Proceedings of ATNAC '95*, December 1995.

[17] T. D. Neame, M. Zukerman and R. G. Addie. Applying Multiplexing Characterization to VBR Video Traffic. In *Proceedings of ITC 16*, June 1999.

[18] M. Parulekar and A. M. Makowski. Tail Probabilities for a Multiplexer with Self-Similar Traffic. In *Proceedings of Infocom '96*, 1996.

[19] J. Roberts, U. Mocci, and J. Virtamo. *Broadband Network Teletraffic, Final Report of Action COST 242*. Springer, 1996.

[20] O. Rose. Statistical Properties of MPEG Video Traffic and Their Impact on Traffic Modeling in ATM Systems. In *Proceedings of the 20th Annual Conference on Local Computer Networks*, October 1995.