

Darryl Veitch

Telecom Research Laboratories
770 Blackburn Rd., Clayton Vic. 3168, Australia

ABSTRACT

We provide models which are capable of describing the long term correlations and self similar burstiness structure found in recent measurements of packet networks and VBR video.

Two families of “fractal” arrival processes are presented which capture these features extremely compactly. We show the equivalence of one of these to processes with unsummable auto-correlation functions used recently [16] to describe long term correlation and burstiness. Our approach however has advantages. The other (1 parameter) family generates burstiness on all time scales. It shows how blocking can occur even for arrival streams with zero arrival rate. This illustrates how parameters describing scaling of burstiness and correlation must replace useless long terms averages.

1 Introduction

Despite the best efforts of network designers to predict and regulate the behaviour of sources in broadband networks (eg through parameter policing), the reality is that until real measurements can be made, the detailed nature of broadband traffic will remain largely unknown. In this partial vacuum, models will inevitably be prejudiced by the intuition which arises from the well established models, and known behaviour, of traditional teletraffic: voice and narrowband data. For example, the models employed to date to describe broadband traffic: the batch Poisson process, the Markov modulated Poisson process, layered Markov models, the general Gaussian model, and others (see [1], [3], [11], [12], [22], [13], [14], [21], [23] and [24]), assume finite mean and variance of all arrival quantities, and incorporate only short term correlations. These fundamental assumptions need not necessarily apply in a real broadband environment, and indeed recent studies (described below) show convincingly that they do not hold true for traffic types which will be important in future multiservice networks. These considerations motivate the present study of alternative arrival processes.

The need for alternative models has been appreciated before. Apart from the authors of the measurement studies described below, Erramilli and Singh [8] have considered the possibility of employing chaotic deterministic maps with intermittency properties, to generate arrival streams with long range dependence and inter-burst gaps of all sizes.

The present models have not yet been refined to the point

where one would expect good quantitative agreement with real broadband traffic. It is important to point out however, that in the absence of testing against real data, there is no reason to be confident that the quantitative accuracy of any traditional model will be any better. Our models, at the very least, highlight possibilities which as yet have received very little attention, and have the advantage of being simpler. In a “pre-measurement age”, the concern should be to have a wide range of tools available, rather than only generating models which behave according to preconceived ideas which cannot be tested. This establishes the *prima facie* importance of the models we study here. Now that some relevant data has become available, it is apparent that a broadening in new directions is indeed necessary.

2 The Evidence

Kathy Meier-Hellstern and Pat Wirth et al [20] from AT&T Bell laboratories measured ISDN D-channel packet data from an office automation environment. They discovered an enormous variability which could not be well explained by traditional models. However they successfully modelled the data by decomposing the traffic into three different types: active typing, irregular activity, and machine generated packets. The latter type was unremarkable, however of the other two, which both arise from human-machine interaction, the distribution of the number of arrivals within the state had *infinite* variance. In addition, the irregular state, which encompassed such diverse activities as thinking time and coffee breaks, had an *infinite* average interarrival time.

Will Leland and Dan Wilson from Bellcore [15] recorded LAN traffic continuously over several Ethernets in great detail, over periods of weeks. Within the LAN's they found extreme variability, and indices of dispersion (a sample size dependent variance to mean ratio) of the arrival rate which did not converge to steady values with increasing sample size. They also found burstiness on all timescales over a range six orders of magnitude wide: milliseconds to days. This means that there was no natural scale on which bursts were organised, rather each burst could be resolved into groups of smaller bursts, right down to the level of packets. They argued decisively that any feasible variant of traditional models would not show such features. The results were similar across the different LAN's, over widely varying load conditions, and through network reconfigurations.

Erramilli and Willinger [9] and Leyland et al. [16] analyse the above data sets, as well as VBR video sequences compiled by Beran et al [4]. Several quantities characterising the

correlation structure of the data sets were measured, and in each case results indicating long term correlations, slowly decaying variances, and self similarity were found. These are all related and strongly at variance with the characteristics of models previously considered.

These studies are inspiration for an investigation of truly new traffic models. In particular, they encourage the use of stochastic processes which possess scaling properties, and which employ distributions with infinite moments. They are also suggestive of self similar structures, and hence of fractals (see [18] for a detailed introduction to fractals and their applications).

In the next section we describe the foundations on which we build our models.

3 Background

There are innumerable possible non-standard models one could choose to investigate. We make a beginning within the domain of renewal processes (RP's). That is, interarrival times are drawn directly from a Probability Distribution Function (PDF) $A(t)$ of a positive random variable X (real or discrete as appropriate), in a succession of independent trials, to generate a set of arrival instances. The discrete counting process $N(t)$ gives the number of arrivals up to time t .

The Renewal Function $m(t) = E[N(t)]$ is a central object of study. Let μ denote the expectation $E[X]$. The Elementary Renewal Theorem states that the arrival rate $\lambda = \lim_{t \rightarrow \infty} m(t)/t$ is well defined and equals $1/\mu$. In the case of $\mu = \infty$ we have $\lambda = 0$. Note that $m(t)$ represents a *sample length dependent average arrival rate*.

The above describes an Ordinary Renewal Process (ORP). If instead of $A(t)$ we choose $A_e(t)$ for the first interarrival time, where $A_e(t) = \int_0^t (1 - A(x))/\mu dx$ is the residual life distribution, we obtain a Stationary or Equilibrium Renewal Process (ERP). Such a process induces a stationary point process on the time line, and $m(t)/t = 1/\mu$ for all t . Only in the case of the Poisson process is the ORP equal to its corresponding ERP.

Within this standard framework we consider two non-standard cases of an ORP. The first is when the expectation (and hence all higher moments) of X is infinite. The second is when μ is finite, but $Var[X]$ remains infinite. They are explored in Sections 4 and 5 respectively.

In the first "fractal" case $A_e(t)$ is not defined, and so we cannot consider the corresponding ERP, or rather, it does not exist except in the form where 'nothing ever arrives'. In other words, the arrival stream is fundamentally transient, it has no non-trivial steady state component on which the usual equilibrium-based theory can act. We are therefore constrained to study the ORP, where, because of the implicit arrival at the origin, $m(t)$ has the interpretation of being the average arrival rate in an interval of length t , *conditional on the fact that something has already arrived*. The work here owes much to the model of errors in transmission

channels of Berger and Mandelbrot [6], and Mandelbrot [17].

In the second case $A_e(t)$ is well defined (though it itself has infinite expectation). Hence we study the ERP and consider the connection between it and the covariant stationary processes (CoVSP's) with unsummable auto-covariance functions used in [16] to describe long term correlation and burstiness.

We now introduce some distributions with infinite moments, both because we will have need of them, and in the hope of removing the aura of unapplicability which the reader may feel surrounds them. This must begin with the observation that there is nothing inherently "strange" in the PDF's of such random variables, they are well defined and normalisable in the usual way. Their characteristic feature is typically a density with a "fat" tail(s), which approaches zero at a slower than exponential rate.

Although less well known in applied areas than to mathematicians, there are many concrete problems where they occur naturally. Two well known examples are the distribution of first passage times (time to first reach a given point) for one dimensional Brownian motion, and the Cauchy distribution, neither of which possesses an expectation. This latter "unusual" distribution corresponds to the $n = 1$ case of the student's t distribution.

The Pareto distribution [7], has been used widely in economic theory. There are several forms of it, the simplest of which is the "classical" Pareto with the PDF: $F_x(x) = 1 - (x/\epsilon)^{-D}$, $x \geq \epsilon$, $\epsilon, D > 0$. The n^{th} moment exists only if $n < D$, and is given by $D\epsilon^n/(D - n)$.

The Zeta distribution [7], is commonly quoted as the discrete version of the Pareto (these were the distributions used by Meier-Hellstern et al. [20]). It is often easier however to work with the similar "discrete hyperbolic" distribution, so called because of the hyperbolic form: $P\{N \geq n\} = n^{-D}$, $D > 0$, of its survival function. Like the Zeta distribution, it has infinite mean for $D \in (0, 1]$, and infinite variance for $D \in (0, 2]$.

Before moving on to a description of our models, we make a further comment on the acceptability of infinite moments. Consider the sample variance of some distribution. Even when it is the case that, for physical reasons, the actual variance must be finite, it may still be best to model it as infinite. As eloquently argued by Mandelbrot (see [18] p337-9), this arises when the actual, finite value, is very large and difficult to measure. In such a case, the choice of infinite variance is much less arbitrary than the injection of a largely unmeasurable new parameter. Of course in other branches of science, the modelling of large quantities as infinite is standard practice.

4 "Fractal" Arrival Processes

In this section we consider the case where $\mu = E[X]$ is infinite, and we study $m(t)$ for the ORP.

4.1 Discrete time

Much of the recent work in broadband modelling has involved queuing analysis in discrete time. This is motivated by the discrete 53-octet cell structure of ATM, one of the main technologies aiming to underpin multimedia broadband communications. Consider then the following specification of $A(t)$ in discrete time, $t = 1, 2, \dots$, measured in cell length *slots*. The arrival time of a cell is associated with the beginning of its header.

$$A(t) = P\{X \leq t\} = 1 - (t + 1)^{-D}, \quad D \in (0, 1) \quad (1)$$

that is, $A(t)$ is given by a discrete hyperbolic distribution with infinite expectation.

Since λ is zero, by Little's Law, the average number of customers in a simple queuing system (assume GI/G/1) fed by such a traffic source, is zero also. The seeming uselessness of this result at first glance should not deter us. In this context, zero arrival rate does *not* imply at all that no data arrives, indeed, $N(\infty) = \infty$, as for any RP. The reason for this is the intrinsically conditioned nature of each sample path of an ORP as discussed earlier. Physically, this corresponds to the certainty that *some* cells will be sent, so we will not find ourselves in the middle of an indefinitely long gap in practice.

It is well known that for highly bursty arrival streams, the average arrival rate drops dramatically as the averaging interval is increased from very small values through to moderate values (*ms* to *seconds* or even *minutes*). An arrival rate of zero simply idealises the assumption that arrivals are very scarce over extremely long timescales. As explained in section 4.2, this is completely consistent with arbitrarily high *conditional arrival rates over finite intervals*. To further justify the assumption of a very low (zero) absolute arrival rate, we turn again to the authority of measurement: the work of Meier-Hellstern et al. demonstrates that non-standard features must be considered even for individual traffic *components*.

The leftmost graph in Figure 1 shows a realisation of $N(t)$ generated using equation 1 with $D = 0.6$. $N(t)$ is plotted against t , for t up to 100000 slots. The graphs on the right are successive enlargements about the origin. They look much like the original: bursts followed by long gaps. They reveal that a nested burst structure continues to lower scales, with no preferred timescale. It is important to note that this *clustering effect is a natural product of the model*, despite its simplicity, and is not due to any correlation structure designed into it (see [17], [18]). This is in sharp contrast to layered Markov models (for example [11]). Not only do these models require complicated state spaces to generate hierarchical clustering, but they suffer from unavoidable arbitrariness, both in the choice of the number of different time scales, and their actual values. They also have far too many parameters to be practical.

Despite the infinite expectation of $A(t)$ and the resulting zero arrival rate, all quantities measured over *finite* intervals

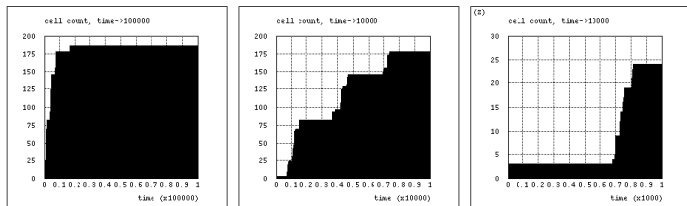


Figure 1: $N(t)$ for the discrete arrival process.

of time (and of course all sample statistics), are themselves finite and non-zero. Indeed it is trivial to see that, given that a cell arrives at $t = 0$, that the expectation of the number of cells to arrive in the next n slots is finite and non-zero for all n . It can be shown [17] that this conditional expectation goes as n^D for large n , and thus conditionally:

$$E[N(t)/t] \sim t^{D-1}$$

which goes to zero in the limit since $D < 1$, as required. (The same is true for *any* interval of length n , conditional on it containing at least one cell.)

Now consider for example, a deterministic queue with a service rate less than one, fed by one of these sources representing highly bursty (though sparse on a long time scale) multiplexed traffic. Despite the total arrival rate, and hence the expected waiting time of an arbitrary packet, being zero, from reasoning similar to that above it is clear that the conditional waiting time for the n^{th} arrival, given at least one arrival has been served, will not be. Thus, with a finite buffer, blocking will still occur. Its probability is a function of the service rate which can be chosen to satisfy a specified grade of service, given D , and the time scale to which the grade of service refers.

This shows that the meaning of “maximum capacity” needs to be redefined for bursty traffic. A channel can be at near maximum capacity almost independent of the value of the absolute arrival rate, because of extreme burstiness.

The non-exponential growth law for the conditional expectation over a finite interval is related to “mass” laws for random fractals on the line, hence we refer to this model as a “fractal arrival process”. Because of the huge range of timescales involved in a high speed network, it is valuable to keep fractals in mind, as simple, natural models of burst within burst behaviour which continue over a very wide range of scales. There will of course be a lower time scale where the fractal like behaviour will cease.

4.2 Continuous time

Because of the small size of cells and the high bandwidths of ATM, it may be prudent to ignore the finite granularity of cells, and model arrivals by a continuous time process. This conveniently allows us to scale the interarrival distribution $A(t)$ continuously. There is no problem for instance, in adjusting the position of the 95th percentile to any desired point. For example, one can set $Pr\{\text{interarrival time} < 1 \text{ ms}\} = A(0.000001) = .95$, with the average interarrival time remaining unbounded.

Consider a continuous positive random variable X with a tail density which goes as t^{-D-1} , $D < 1$, for large t , for example the Classical Pareto with $D < 1$. Such a variable has infinite expectation, so again $\lambda = 0$.

Let $A_n(t)$ be the PDF of the time to the n^{th} arrival. From standard renewal theory [5],

$$m(t) = E[N(t)] = \sum_{n=1}^{\infty} P\{N(t) \geq n\} = \sum_{n=1}^{\infty} A_n(t)$$

Using a Tauberian theorem ([10], p446), it can be shown that for any choice of $A(t)$ with a tail density of this form, $m(t) \sim t^D$. Thus, just as in the discrete case we have $m(t)/t \sim t^{D-1}$, $t \gg 1$. The same intrinsic burstiness on all timescales (above some lower limit), conditional blocking etc, are also present in this continuous time example.

Mandelbrot [17] also obtains this result using the classical Pareto distribution. He goes further by considering the limit as $\epsilon \rightarrow 0$ to generate what can be regarded as truly ‘‘fractal’’ arrivals. What this latter term means is that the corresponding processes, with probability one, will generate arrivals which as sets on the time line will have a fractal dimension equal to D . An actual fractal however, possesses an uncountable infinity of points, and thus cannot be generated by any probabilistic process which is normalizable, such as the one we have described here, or the Pareto for any $\epsilon > 0$. For these, $P\{N(t) \geq 1 | \text{an arrival at } t = 0\}$ tends to 0, not ∞ , as $t \rightarrow 0$. Because of the self similarity of the present model however, which derives from the tail behaviour of $A(t)$, the arrivals generated will be fractal-like beyond some finite length-scale, which can be chosen as small as desired. If $D > 1$, the arrival rate will become greater than zero, and so scaling of this type will not be possible. The interpretation of D as an approximate fractal dimension would then be restricted to a range of scales with an upper as well as a lower bound. It is these processes we now consider.

5 Covariant Stationary Processes

An ERP can be viewed as a stationary point process, and also as the stationary discrete time process $Y = \{Y_i\}$, formed by dividing time into blocks of width τ and defining $Y_i = N[i\tau] - N[(i-1)\tau]$, $i = 1, 2, \dots$. Evidently $E[Y] = \tau/\mu$, $\nu = \text{Var}[Y]$ and $\xi_{i-j} = \text{Cov}[Y_i, Y_j]$ are all finite, so Y is also a CoVSP.

In [16], the observed properties of LAN traffic were convincingly accounted for in a CoVSP framework where the autocorrelation function $\{\rho_k = \xi_k/\nu\}$ was unsummable: $\sum_0^{\infty} \rho_k$ does not converge. Specifically, a power law decrease in ρ_k with k was assumed, rather than the usual geometric decay. This implies the presence of long term correlation, a feature absent in all normal traffic models. This correlation structure is equivalent to the ‘‘slowly decaying variance’’ property, ie $\text{Var}[(Y_i + Y_{i+1} + \dots + Y_{i+k-1})/k] \sim k^{-\beta}$ for k large, $\beta \in (0, 1)$ (for normal models $\beta = 1$). In the present context, the left hand side of this expression is equivalent to $\text{Var}[N[k\tau]/k\tau]$, and we wish to discover if $\text{Var}[X] = \infty$

implies the right hand side. We again assume a tail density $\sim t^{-D-1}$, now with $D \in [1, 2)$.

$\text{Var}[N[t]]$ is just $\frac{2}{\mu}(\int_0^t m_o(u) du + t/2 - t^2/2\mu)$, where m_o is the renewal function of the ORP (see [5]). Thus we now require the asymptotic behaviour of the ‘transient’ remainder $m_o(t) - t/\mu$. It is well known that this goes as $(\sigma^2 - \mu^2)/2\mu^2$ in the case of finite variance σ ; in our case the remainder diverges. Thus it is reasonable to assume that $m_o(t) - t/\mu \sim at^{2-D}$, $(2-D) \in (0, 1)$ for large t , and substitution in the above equation yields $\text{Var}[N[t]] \sim a't^{3-D}$, $t \gg 1$

$$\text{Thus } \text{Var}[N[k\tau]/k\tau] \sim a'k^{1-D}\tau^{1-D}$$

in agreement with the above with $D = 1 + \beta$. This shows the equivalence of the two approaches. The ERP formulation however, is much simpler, and more attractive from the point of view of artificial traffic stream generation.

6 Conclusion and Future Work

As long as broadband models are not tested against detailed data, they will be vulnerable to the danger of being based on intuitions which are no longer valid. The analysis of new measurements of ISDN, LAN and VBR Video data, has suggested that the traditional assumptions and concepts of teletraffic modelling are not adequate for the study of these important source types, and hence not for the future networks which will interconnect them.

The non-standard renewal processes we have studied here make a good beginning in capturing and characterizing the long term correlations, burstiness on all scales, and sample size dependent statistics, observed in these real traffic streams. The models need refinement before use in quantitative network design, however they are at least consistent with the main measured features, whereas conventional models do not capture them at all.

The ‘‘fractal’’ processes have only one parameter, and yet display burstiness on a continuum of scales, a feature which nested Markov models would require many parameters to artificially estimate. They dramatically illustrate the fact that the arrival rate is an average concept which is no longer of relevance: since blocking occurs despite $\lambda = 0$. An expression is given for conditional arrival rates as a function of the length of the measurement interval.

The two parameter (D and μ) ERP processes were shown to be equivalent to the special covariant stationary models used elsewhere [16] to describe long term correlation and burstiness. The ERP formulation shares the properties of these models, but is simpler and more useful for the generation of artificial traffic streams.

In each model there is a parameter D which measures, not an absolute arrival rate or a fixed burst scale, but the way in which measured rates and burstiness varies *with* time scale. In this way ‘‘complex’’ traffic is treated naturally and with economy. The crucial role of the size of the measurement interval to measurement is now well recognized. It is not feasible however to continually monitor traffic streams in real

time at small timescales. Hence, models are required which can infer the results of measurements on all scales, from measurement over a small range of medium scales. Scaling/fractal models are precisely those which have the characteristics to allow this. Thus this work has much to say on the difficult issue (see [2], [19], [25]) of real time parameter estimation, as well as the characterization of burstiness. It also shows how long term correlation need not be irrelevant for cell level queuing.

The avenues for further investigation are too extensive to map here. Some of the more obvious questions however concern the connection between ERP's and the fractal models in the case of an upper bound to scaling, and also further details of the connection between the ERP's and CoVSP's. Statistical issues regarding the estimation of D are crucial and need addressing (some techniques already exist, see [16]). The chaotic maps approach needs to be developed and related to the work here. Different objects could be the focus of fractal attention. For instance the specification of a "net input process" which would enable the *stable distributions* [10], to be profitably employed. There is of course enormous scope for the development of queueing theory with non-standard arrival processes, with a focus on time-scale dependent conditional statistics. Finally, there is a whole range of questions to be explored regarding the superposition of non-standard processes, both with themselves, and more conventional ones.

References

- [1] R. Addie and M. Zukerman, "A Gaussian Traffic Model for Statistical Multiplexers," Proceedings, *ABSSS '92*, Clayton, Victoria, Australia, July 1992.
- [2] A. Avidsson and R. Harris, "Performance comparison of models of individual and merged bursty traffics," Proceedings, *6th ATRS*, Wollongong, Australia, November 1991.
- [3] B. Bensaou, J. Guibert and J. W. Roberts, "Fluid approximation for a superposition of on/off sources," Proceedings, *ITC Seminar*, Morristown, NJ, Oct. 1990.
- [4] J. Beran, R. Sherman, M. Taqqu and W. Willinger, "Variable-Bit-Rate Video Traffic and Long Range Dependence," Bellcore TM-ARH-020766, Feb. 1992.
- [5] D. R. Cox, "Renewal Theory," *Chapman and Hall* 1962.
- [6] J. M. Berger, B. B. Mandelbrot, "A New Model for Error Clustering in Telephone Circuits" *IBM Journal*, Vol 7, No. 3, July 1963.
- [7] Encyclopedia of Statistical Sciences, Vol 6 p569, Vol S p672, *John Wiley & Sons*, 1987.
- [8] A. Erramilli and R. P. Singh, "The Application of Deterministic Chaotic Maps to Characterize Traffic in Broadband Packet Networks," Proceedings, *ITC Specialist Seminar*, 1990.
- [9] A. Erramilli and W. Willinger, "Fractal Properties in Packet Traffic Measurements," Proc. ITC Seminar, St. Petersburg, 1993.
- [10] W. Feller, "An Introduction to Probability Theory and Its Applications," Volume II, *Wiley* 1966.
- [11] O. Gihl and P. Tran-Gia, "A layered description of ATM cell traffic streams and correlation analysis," *ATR*, vol. 24, no. 2, pp. 9-18, 1990.
- [12] J. J. Gordon, "Modelling bursty traffic with two-state sources," *ATR*, vol. 24, no. 2, pp. 51-63, 1990.
- [13] O. Hashida, Y. Takahashi and B. Sengupta, "Switched batch Bernoulli process (SBBP) and the discrete-time SBBP/G/1 queue with application to statistical multiplexer performance," *IEEE JSAC*, vol. 9, pp. 394-401, April 1991.
- [14] H. Hefes and D. Lucantoni, "A Markov modulated characterization of voice and data traffic and related statistical multiplexer performance," *IEEE JSAC*, vol. SAC-4, pp. 856-867, Sept. 1986.
- [15] W. E. Leyland, D. V. Wilson, "High Time Resolution Measurements and Analysis of LAN Traffic: Implications for LAN Interconnection," Proc. *IEEE Infocom* 1991.
- [16] W. E. Leyland, M. Taqqu W. Willinger and D. V. Wilson, "On the Self Similar Nature of Ethernet Traffic," Preprint, Feb. 1993.
- [17] B. B. Mandelbrot, "Self Similar Error Clusters in Communications Systems and the Concept of Conditional Stationarity" *IEEE Transactions on Communications Technology*, Volume 13, pp71-90, 1965.
- [18] B. B. Mandelbrot, "The Fractal Geometry of Nature," *Freeman*, 1983.
- [19] K. S. Meier-Hellstern, "A fitting algorithm for Markov-modulated Poisson process having two arrival rates," *European Journal of Operational Research*, vol. 29, pp. 370-377, 1987.
- [20] K. S. Meier-Hellstern, P. E. Wirth, Y-L. Yan, and D. A. Hoefflin. "Traffic Models for ISDN Data Users: Office Automation Application," Proceedings, *ITC13*, pp 167, Copenhagen 1991.
- [21] I. Norros, J. W. Roberts, A. Simonian and J. T. Virtamo, "The superposition of variable bit rate sources in an ATM multiplexer," *IEEE JSAC*, vol. 9, pp. 378-387, April 1991.
- [22] M. H. Rossiter, "A switched Poisson model for data traffic," *ATR*, vol. 21, no. 1, 1987.
- [23] H. Saito, M. Kawarasaki and H. Yamada, "An analysis of statistical multiplexing in an ATM transport network," *IEEE J. Select. Areas Commun.*, vol. 9, pp. 359-367, April 1991.
- [24] K. Sriram and W. Whitt, "Characterizing superposition arrival processes in packet multiplexers for voice and data," *IEEE JSAC*, vol. SAC-4, pp. 833-846, Sept. 1986.
- [25] B. Warfield and S. Chan, "Real-Time Traffic Estimation in B-ISDN," Proceedings of the 7th Australian Teletraffic Research Seminar, Mannum, South Australia, Nov. 1992