Modelling LRD and SRD Traffic with the Batch Renewal Process: Review of Results and Open Issues

Rod J. Fretwell and Demetres D. Kouvatsos

Networks and Performance Engineering Research Group
Informatics Research Institute, University of Bradford,
Bradford BD7 1DP, United Kingdom
{R.J.Fretwell,D.Kouvatsos}@Bradford.ac.uk

Abstract. The batch renewal process is the least biased choice of a process given only the measures of count and interval correlations at all lags. This article reviews the batch renewal process for modelling both LRD (long range dependent) and SRD (short range dependent) traffic flows. The exposition focuses mainly in the discrete-space discrete-time domain and in the wider context of general traffic in that domain. However, corresponding results in the continuous-time domain are also presented. Moreover, some applications of the batch renewal process in simple queues and in queueing network models are undertaken and associated analytic performance results are devised. The article concludes with open research problems and issues relating to the batch renewal process.

1 Introduction

Over the past two decades there has been great interest in (auto)correlated traffic because of its adverse impact upon performance of high speed telecommunications systems by buffer congestion and blocking or packet loss, transmission delay and jitter (delay variability).

In 1986 Sriram and Whitt [19] considered the effect on a multiplexer of superposition of a number of identical renewal processes (modelling telephony talkspurts). They observed that "the aggregate arrival process possesses exceptional long-term positive dependence" and reported the adverse effect on performance of "dependence among interarrival times" in terms of congestion in the queue, delay and blocking probability, provided that the buffer were sufficiently large that "...many interarrivals times interact in the queue." However, when the buffer was small the impact of traffic correlation was restrained: the queue behaved more like one fed by a renewal process.

Gusella [8] collected traces of traffic in a large Ethernet and in 1991 proposed that traffic correlation be characterized by the indices of dispersion. He illustrated his proposal by computing the sample IDC's (indices of dispersion for counts) and IDI's (indices of dispersion for intervals) for measurements of traffic

D. Kouvatsos (Ed.): Next Generation Internet, LNCS 5233, pp. 141–173, 2011.
© Springer-Verlag Berlin Heidelberg 2011

generated by each of six workstations and gave a procedure for fitting the indices of dispersion to the parameters of a 2-phase MMPP (Markov modulated Poisson process). It is of interest to note that Sriram and Whitt [19] used an approximation based on fitting a 2-phase MMPP as also did Heffes and Lucantoni [9] (for the same model as in [19]) and that the recommended fitting procedure is different in all three papers.

Generally the models used most commonly for early investigations into the impact of correlated traffic were simple forms of the Neuts process [18], predominantly MMPP's with small numbers of phases. This class of models can address only short-range dependent (SRD) traffic.

By using indices of dispersion, Gusella implicitly assumed SRD traffic and one of his concerns was for the possible non-stationarity in the traffic over the longer periods of time. However, the lengths of his traces were short relative to the extensive, precise traces of Bellcore LAN traffic which were collected subsequently. From analysis of those data first Fowler and Leland [4] (1991) reported LAN traffic with unbounded IDC and, in 1994, Leland, Taqqu, Willinger and Wilson discerned "the self-similar nature of Ethernet traffic" [15].

Similar effects have been reported subsequently by many researchers, from simulation studies and analysis of a variety of models, and have led to the present consensus that, in general terms, traffic correlation adversely affects queue congestion, waiting times and blocking probabilities and that long term positive correlation can have significant impact, even when the magnitude of the correlation is relatively low. Consequently there has been more interest in models, such as fractional Brownian motion (fBM), and in Pareto distributions of inter-arrival times [7], which can capture the asymptotically 'hyperbolic' decline in covariances for long-range dependent (LRD) processes.

The popular models for SRD traffic can be fitted tolerably well to covariances in measure traffic at small lags but are limited necessarily to geometrically declining covariances at long lags. Contrarily, the popular models of LRD traffic can be fitted precisely to the (asymptotic) decline in covariances with increasing long lags for measured traffic but do not provide for matching covariances at shorter lags. However, the batch renewal process can match both correlation of counts and correlation of intervals at all lags.

The observation of that property of the batch renewal process (first reported in [11]), derived from consideration of the duality implicit in Gusella's argument (in [8]) for equality $I_\infty = J_\infty$ of the limits I_∞ for the IDC and J_∞ for the IDI as lags tend to infinity in a wide sense stationary process. The duality is most readily apparent in the discrete space discrete time domain. Section 2 of this article addresses different views of general discrete-space discrete-time traffic and shows that one of those views leads naturally to introducing the batch renewal process. An essentially similar argument for the continuous-time domain is given by Li [16].

The next three sections focus upon the batch renewal process itself. Section 3 shows how to construct a batch renewal process which matches measured correlation, whether LRD or SRD and to arbitrary accuracy. Section 4 presents solution methods for simple single-server queues fed by general batch renewal processes and closed-form results for the sGGeo [12] in particular. The sGGeo is a batch renewal process which has proved useful as an investigative tool (a role in which the sGGeo features in Section 8). Section 5 shows how burst structure is induced in the deparures from a finite-buffer queue fed by correlated traffic.

Before illustrating other applications of the batch renewal process, Section 6 gives consideration to a more general class of traffic processes and to the ways such processes might be modelled. Then, because the batch renewal process is *the least biased choice of process given only measures of correlation* [12], it has application as the standard for reference in comparison with other correlated traffic processes. An example of such usage is provided in Section 7 which reports an investigation into the effect of bias consequent upon chosing some other process to capture traffic correlation. Section 8 shows an application in which the sGGeo is used in examining the impact of SRD traffic correlation upon the accuracy of a fast algorithm for approximate analysis of networks.

The article concludes with a review of some open problems and research topics in Section 9.

2 External Views of Traffic and Traffic Processes

In classic queueing theory, traffic is described as the sequence of instants at which customers arrive to the queue system or, usually, as the sequence of interarrival times (the intervals between successive arrivals). In this view of traffic, we number the customers consecutively, in order of arrival instant, and then define the n^{th} interarrival time x_n to be the time between the instant of the $(n-1)^{\text{th}}$ arrival and that of the n^{th} arrival.

In digital computer systems and telecommunications systems there is usually a *natural* unit of time. For example, in an output buffer of an ATM switch, the output port transmits an ATM cell at regular intervals at at rate determined by the output line transmission speed; the (fixed) time to transmit one cell is the natural unit of time in this case. As far as buffer performance is concerned, those cells that arrived during one transmission period might just as well have arrived all together at the start of that period.

Each time period is called a *slot* and the instant that marks the end of one slot (and the begining of the next) is a called an *epoch*. In discrete time models, events are deemed to occur at epochs only.

So, in digital systems, there is another natural way of viewing traffic: that is, in terms of the numbers of arrivals at successive epochs.

Usually, when discussing models of traffic processes, we are concerned with the *internal* representation of the process. For example, we may describe a DBMAP (discrete time batch Markovian arrival process) as a traffic process in which there is an underlying Markov chain over a countable space J such that whenever the process be in phase $i \in J$ there are n arrivals generated and a transition to phase $j \in J$ with probability $d_{ij}(n)$, $n \in \mathbb{N}_0$. That description gives an internal representation of the process because it describes how the traffic is generated, not what the traffic is.

On the other hand, when we are dealing with observed traffic we are concerned with the *external* view of the traffic process, without necessarily knowing what internal representation might have generated the traffic.

When we record, for example, the size c_t of a message segment detected at time t or the number c_t of individual ATM cells that arrive at an output port buffer during the t^{th} transmission slot, we are, in effect, viewing the traffic as a sequence $\{c_0, c_1, \ldots, c_T\}$ of counts at the successive epochs numbered $0, 1, \ldots, T$. We may then choose to regard that sequence $\{c_0, c_1, \ldots, c_T\}$ as if it were a finite subsequence of the infinite sequence $\{c_t : t \in \mathbb{Z},\ c_t \in \mathbb{N}_0\}$ which, in turn, we may choose to regard as being a possible realization of a count process $\{c(t) : t \in \mathbb{Z},\ c(t) \in \mathbb{N}_0\}$.

Alternatively, we may choose to regard the traffic as a sequence of interarrival times. We may number the customers consecutively, in order of arrival (applying arbitrary ordering on simultaneous arrivals), and then define the n^{th} interarrival time x_n to be the number of slots between the epoch of the $(n-1)^{\text{th}}$ arrival and that of the n^{th} arrival. Just as for the counts view of traffic, we may choose to regard a sequence of observed interarrival times as comprising some finite subsequence from a realization of a persistent interarrival time process $\{x(n) : n \in \mathbb{Z},\ x(n) \in \mathbb{N}_0\}$, i.e. the random function $x(n)$ is the duration of the interval between the n^{th} individual arrival and the $(n+1)^{\text{th}}$ arrival.

Count processes and interarrival processes are equivalent in the sense that for every realization of a count process we can construct an equivalent realization of a corresponding interarrival process and *vice versa*.

There is some symmetry in the duality between these two views of traffic.

- Interarrival times greater than zero correspond to intervals between successive points (in the count process) at which the counts are greater than zero.
- Counts greater than zero correspond to intervals between successive points (in the interarrival process) at which the interarrival times are greater than zero.
- When interarrival times are *iid* (independent and identically distributed), as from a renewal process, the corresponding count process is covariance stationary.
- When counts are *iid*, as from a batch Bernoulli process, the corresponding interarrival process is covariance stationary.

If the two views of traffic, as a counts process or as an interarrivals time process, which are presented above, are perceived as being in opposition to each other then there is an intermediate, more symmetric view of traffic. In this view the traffic is described as an alternating process of (non-empty) batches and intervals (at least one slot long) between batches. An equivalent description is a 2-dimensional process $\{\xi(s), \kappa(s) : s \in \mathbb{Z},\ \xi(s), \kappa(s) \in \mathbb{N}_0\}$ in which the component $\kappa(s)$ represents the number of simultaneous arrivals in the s^{th} (non-empty) batch and $\xi(s)$ represents the interval between the $(s-1)^{\text{th}}$ batch and the s^{th} batch. Figure 1 illustrates the relation between this 2-dimensional process and the counts process and the interarrival time process.

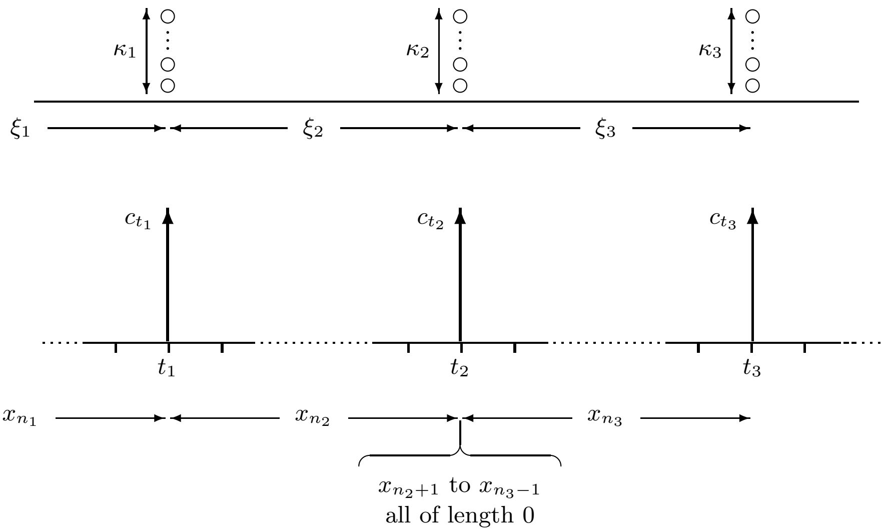

Fig. 1. Relationships between a realization $\{\ldots,(\xi_1,\kappa_1),(\xi_2,\kappa_2),(\xi_3,\kappa_3),\ldots\}$ of the process $\{\xi(s),\kappa(s)\}$ and realizations of the counts process $\{c(t)\}$ and of the interarrival times process $\{x(n)\}$. In this illustration $n_2 = n_1 + c_{t_1}, \ldots,$ $t_2 = t_1 + x_{n_2}, \ldots$

The features of the duality between the two previous views of traffic extend to this 2-dimensional view.

- When the batch sizes $\kappa(\cdot)$ are *iid* and the intervals $\xi(\cdot)$ between batches are *iid* the 2-dimensional process is a batch renewal process. Then the corresponding count process is covariance stationary and the corresponding interarrival process is covariance stationary.

To describe more precisely the relationship between the sequence $\{\xi(s),\kappa(s)\}$ and the sequence $\{x(n)\}$, let n_s be the number of the interval between the last individual arrival of batch s and the first of batch $s+1$: equivalently, let the individual arrivals be numbered such that arrival n_s be the last member of batch s, arrivals $n_s{+}1$, $n_s{+}2$, ..., n_{s+1} be the the first, second, ..., last member (respectively) of batch $s{+}1$. Then $x(n_s) = \xi(s)$ and $n_{s+1} = n_s + \kappa(s{+}1)$. The $\kappa(s{+}1)$ members of batch $s{+}1$ arrive simultaneously: the intervals between them are each of zero duration; so $x(n) = 0$ for $n_s < n < n_{s+1}$.

Obviously, each of the sequences $\{\xi(s),\kappa(s)\}$, $\{c(t)\}$ and $\{x(n)\}$ contains the same information (although in a different form) about the traffic process. Each of the sequences can be derived from either of the other two.

The usual measures of traffic correlation assume that the sequences $\{c(t)\}$ and $\{x(n)\}$ be wide sense stationary.

Definition. A random sequence $\{x(n) : n = \ldots,-2,-1,0,1,2,\ldots\}$ is stationary in the wide sense (equivalently, stationary in Khinchin's sense) if

- the random function $x(n)$ has finite mean $\mathsf{E}\left[x(n)\right] = x$ which is constant (independent of n) and
- the correlation function $\mathsf{Cov}\left[x(n), x(m)\right] \overset{\Delta}{=} \mathsf{E}\left[(x(n) - x)(x(m) - x)\right]$ is finite and depends on the lag $n-m$ only.

Observe that $\mathsf{Cov}\left[x(n), x(n+\ell)\right] = \mathsf{Cov}\left[x(n+\ell), x(n)\right]$, by symmetry of the definition, and that $\mathsf{Cov}\left[x(n+\ell), x(n)\right] = \mathsf{Cov}\left[x(n), x(n-\ell)\right]$, by change of variable n to $n-\ell$. Consequently, $\mathsf{Cov}\left[x(n), x(n+\ell)\right] = \mathsf{Cov}\left[x(n), x(n-\ell)\right]$. Only the magnitude of the lag is significant and it is therefore necessary to consider positive lags only.

Traffic correlation is customarily expressed either as the correlation functions on $\{c(t)\}$ and $\{x(n)\}$ or as the indices of dispersion. The index of dispersion for counts is defined to be the sequence $\{I_t : \ t = 1, 2, \ldots\}$ where

$$I_t = \frac{\mathsf{Var}\left[c(i+1) + \cdots + c(i+t)\right]}{\mathsf{E}\left[c(i+1) + \cdots + c(i+t)\right]} = \frac{\mathsf{Var}\left[c(i+1) + \cdots + c(i+t)\right]}{t\,\mathsf{E}\left[c(i)\right]} . \tag{1}$$

The index of dispersion for intervals is defined to be the sequence $\{J_n : \ n = 1, 2, \ldots\}$ where

$$J_n = \frac{\mathsf{Var}\left[x(i+1) + \cdots + x(i+n)\right]}{\mathsf{E}\left[x(i+1) + \cdots + x(i+n)\right]^2/n} = \frac{\mathsf{Var}\left[x(i+1) + \cdots + x(i+n)\right]}{n\,\mathsf{E}\left[x(i)\right]^2} . \tag{2}$$

Observe that, if λ be the intensity of the traffic, $\mathsf{E}\left[c(t)\right] = c = \lambda$ and $\mathsf{E}\left[x(n)\right] = x = 1/\lambda$.

The indices of dispersion and the correlation functions contain exactly the same information and are related in a simple way.

$$t\,I_t = \sum_{i=1}^{t} i\,K_{t-i} \qquad \text{and} \qquad n\,J_n = \sum_{j=1}^{n} j\,L_{n-j} \tag{3}$$

where

$$K_\ell = \begin{cases} \dfrac{1}{\lambda}\,\mathsf{Var}\left[c(t)\right] & \ell = 0 \\ 2\,\dfrac{1}{\lambda}\,\mathsf{Cov}\left[c(t), c(t+\ell)\right] & \ell = 1, 2, \ldots \end{cases} \tag{4}$$

and

$$L_\ell = \begin{cases} \lambda^2\,\mathsf{Var}\left[x(n)\right] & \ell = 0 \\ 2\,\lambda^2\,\mathsf{Cov}\left[x(n), x(n+\ell)\right] & \ell = 1, 2, \ldots \end{cases} . \tag{5}$$

In particular, J_1 is the square coefficient of variation C_x^2 of the intervals $x(n)$ between successive individual arrivals and, for bounded indices of dispersion,

$$J_\infty = I_\infty . \tag{6}$$

3 The Batch Renewal Process That Matches Measured Correlation

The batch renewal process is the *least biased choice* of process given only the count covariances $\{\mathsf{Cov}\left[c(t), c(t{+}\ell)\right] : t, \ell \in \mathbb{Z}\}$ and the interarrival covariances $\{\mathsf{Cov}\left[x(n), x(n{+}\ell)\right] : n, \ell \in \mathbb{Z}\}$ [12]. This section shows how to identify the batch renewal process that matches exactly the given covariances. The approach is first to derive covariance generating functions for a general batch renewal process and then to solve the corresponding equations to express the batch renewal process probability generating functions in terms of the covariance generating functions.

The exposition is in two parts. The first is applicable when both the count covariances are summable and also the interval covariances are summable (Short Range Dependent processes). The second subsection applies to cases in which either the count covariances are not summable or the interval covariances are not summable (Long Range Dependent processes).

The following notation is used.

$\{a(t) = \mathsf{P}\left[\xi(s) = t\right] : t = 1, 2, \ldots\}$ the *pmf* (probability mass function) of the interval between successive batches
$\{b(n) = \mathsf{P}\left[\kappa(s) = n\right] : n = 1, 2, \ldots\}$ the *pmf* of the batch size
$a = \mathsf{E}\left[\xi(s)\right]$, C_a^2 the mean and SCV (squared coefficient of variation) of the interval between successive batches
$b = \mathsf{E}\left[\kappa(s)\right]$, C_b^2 the mean and SCV of the batch size
$\lambda = {}^{b}\!/_{a}$ the mean arrival rate
$A(\omega) = \sum_{t=1}^{\infty} a(t)\,\omega^t$ the *pgf* (probability generating function) of $\{a(t)\}$
$B(z) = \sum_{n=1}^{\infty} b(n) z^n$ the *pgf* of $\{b(n)\}$

3.1 Short Range Dependent Processes

We shall say that a process is short range dependent if both the count covariances are summable and the interval covariances are summable and shall see that condition, in the case of the batch renewal process, is equivalent to the variances of counts and of intervals both being finite. Finite variances are assumed in this sub-section.

Calculation of the various expectations (mean, variance and covariances) is greatly facilitated by exploiting conditional independence.

- Independence of batch size $\{\kappa(s)\}$ implies conditional independence of the count $c(t)$ at epoch $t \in \mathbb{Z}$ given only that $c(t) > 0$.
- Independence of intervals $\{\xi(s)\}$ between batches implies conditional independence of the interval $x(n)$ between individual arrivals ($n \in \mathbb{Z}$) given only that the interval $x(n) > 0$.

But only random variables with values greater than zero contribute to expectations.

$$\begin{aligned}\lambda = \mathsf{E}\left[c(t)\right] &= \mathsf{P}\left[c(t) > 0\right]\mathsf{E}\left[c(t) \mid c(t) > 0\right] + \mathsf{P}\left[c(t) = 0\right]\mathsf{E}\left[c(t) \mid c(t) = 0\right] \\ &= \frac{1}{\mathsf{E}\left[\xi(s)\right]}\,\mathsf{E}\left[\kappa(s)\right] + 0 \\ &= \frac{1}{a}\,b = {}^{b}\!/_{a} \end{aligned} \tag{7}$$

$$\begin{aligned}\mathsf{Var}\left[c(t)\right] = \mathsf{E}\left[c(t)^2\right] - \mathsf{E}\left[c(t)\right]^2 &= \mathsf{P}\left[c(t) > 0\right]\mathsf{E}\left[c(t)^2 \mid c(t) > 0\right] - \mathsf{E}\left[c(t)\right]^2 \\ &= \frac{1}{\mathsf{E}\left[\xi(s)\right]}\,\mathsf{E}\left[\kappa(s)^2\right] - \mathsf{E}\left[c(t)\right]^2 \\ &= \frac{1}{\mathsf{E}\left[\xi(s)\right]}\big(\mathsf{Var}\left[\kappa(s)\right] + \mathsf{E}\left[\kappa(s)\right]^2\big) - \mathsf{E}\left[c(t)\right]^2 \\ &= \frac{1}{a}\,b^2\big(C_b^2 + 1\big) - \left(\frac{b}{a}\right)^2 \\ &= \frac{b^2}{a}\left(C_b^2 + 1 - \frac{1}{a}\right) \end{aligned} \tag{8}$$

and, for $\ell = 1, 2, \ldots,$

$$\begin{aligned}&\mathsf{P}\left[c(t) = n, c(t{+}\ell) = k, n > 0, k > 0\right] \\ &\qquad = \mathsf{P}\left[c(t) > 0\right] \\ &\qquad \times \mathsf{P}\left[c(t) = n \mid c(t) > 0\right] \\ &\qquad \times \mathsf{P}\left[c(t{+}\ell) > 0 \mid c(t) = n,\ c(t) > 0\right] \\ &\qquad \times \mathsf{P}\left[c(t{+}\ell) = k \mid c(t{+}\ell) > 0,\ c(t) = n,\ c(t) > 0\right] \\ &\qquad = \mathsf{P}\left[c(t) > 0\right] \\ &\qquad \times \mathsf{P}\left[c(t) = n \mid c(t) > 0\right] \\ &\qquad \times \mathsf{P}\left[c(t{+}\ell) > 0 \mid c(t) > 0\right] \\ &\qquad \times \mathsf{P}\left[c(t{+}\ell) = k \mid c(t{+}\ell) > 0\right] \\ &\qquad = \frac{1}{a}\,b(n)\phi_\ell b(k) \end{aligned} \tag{9}$$

so that

$$\mathsf{E}\left[c(t)c(t{+}\ell)\right] = \frac{b^2}{a}\,\phi_\ell \tag{10}$$

where $\phi_\ell = \mathsf{P}\left[c(t{+}\ell) > 0 \mid c(t) > 0\right]$ is the probability that some integral number of intervals between batches be exactly ℓ slots long. Clearly ϕ_ℓ must satisfy

$$\phi_\ell = \begin{cases} 1 & \ell = 0 \\ \displaystyle\sum_{t=1}^{\ell} a(t)\,\phi_{\ell-t} & \ell = 1, 2, \ldots \end{cases} \tag{11}$$

and so is generated by

$$\begin{aligned} \sum_{\ell=0}^{\infty} \phi_\ell\,\omega^\ell &= 1 + \sum_{\ell=1}^{\infty}\sum_{t=1}^{\ell} a(t)\,\phi_{\ell-t}\omega^\ell \\ &= 1 + \sum_{t=1}^{\infty}\sum_{\ell=t}^{\infty} a(t)\,\phi_{\ell-t}\omega^\ell \\ &= 1 + A(\omega)\sum_{\ell=0}^{\infty}\phi_\ell\,\omega^\ell \\ &= \frac{1}{1-A(\omega)} \end{aligned} \tag{12}$$

The structure of the batch renewal process makes it relatively simple to derive the form of covariances by exploiting conditional independence. For example, the distribution of count $c(t)$ at epoch t depends only upon the condition $c(t) > 0$. Also, calculation of the covariances makes obvious that they are stationary. Then, assuming that the variances $\mathsf{Var}\left[\xi(s)\right]$ and $\mathsf{Var}\left[\kappa(s)\right]$ are finite, it can be seen that the covariances are given by the generating functions

$$\begin{aligned} K(\omega) = \sum_{\ell=0}^{\infty} K_\ell\omega^\ell &= {}^1\!/\!_\lambda\Big(\mathsf{Var}\left[c(t)\right] + 2\sum_{\ell=1}^{\infty}\mathsf{Cov}\left[c(t), c(t{+}\ell)\right]\omega^\ell\Big) \\ &= b\left(C_b^2 + \frac{1+A(\omega)}{1-A(\omega)} - \frac{1}{a}\,\frac{1+\omega}{1-\omega}\right) \end{aligned} \tag{13}$$

and

$$\begin{aligned} L(z) = \sum_{\ell=0}^{\infty} L_\ell z^\ell &= \lambda^2\Big(\mathsf{Var}\left[x(n)\right] + 2\sum_{\ell=1}^{\infty}\mathsf{Cov}\left[x(n), x(n{+}\ell)\right]z^\ell\Big) \\ &= b\left(C_a^2 + \frac{1+B(z)}{1-B(z)} - \frac{1}{b}\,\frac{1+z}{1-z}\right) \end{aligned} \tag{14}$$

Under the assumption that the analyst has been able to express the covariances in the form of the generating functions $K(\omega)$ and $L(z)$, construction of the corresponding batch renewal process reduces to solving equations (13) and (14) for $A(\omega)$ and $B(z)$, as follows.

Setting $\omega = 0$ in (13) and $z = 0$ in (14) immediately yields

$$K(0) = b\left(C_b^2 + 1 - {}^1\!/\!_a\right) \tag{15}$$

$$L(0) = b\left(C_a^2 + 1 - {}^1\!/\!_b\right) \tag{16}$$

and, by considering limits as $\omega \to 1-$ in (13) and $z \to 1-$ in (14),

$$K(1-) = b\left(C_a^2 + C_b^2\right) \qquad \text{and} \qquad L(1-) = b\left(C_a^2 + C_b^2\right) \ . \tag{17}$$

It may be observed that the existence of those limits is equivalent to saying that the covariances be summable by the method of Abel. Furthermore, $K(1-) = I_\infty$ and $L(1-) = J_\infty$, where I_∞ and J_∞ are the limits for the IDC and IDI as lags tend to infinity.

Then, from equations (15), (16) and (17), b must satisfy

$$K(0) + L(0) - K(1-) = 2b - 1 - \lambda \tag{18}$$

and, using (18), equations (13) and (14) can be manipulated to give

$$A(\omega) = 1 - \frac{K(0) + L(0) - K(1-) + 1 + \lambda}{K(\omega) + L(0) - K(1-) + 1 + \lambda\dfrac{1+\omega}{1-\omega}} \tag{19}$$

and

$$B(z) = 1 - \frac{K(0) + L(0) - K(1-) + 1 + \lambda}{K(0) + L(z) - K(1-) + \dfrac{1+z}{1-z} + \lambda} \ . \tag{20}$$

Constructing the Covariance Generating Functions $K(\omega)$ and $L(z)$. For the analyst to have decided that the process be short range dependent, it is likely that the graph of log covariance against lag is (approximately) piece-wise linear — which is equivalent to saying that the correlation function is (approximated by) the weighted sum of geometric terms or that the generating function ($K(\omega)$ or $L(z)$) is a rational function (of ω or of z, respectively).

In that case, the geometric components may be extracted progressively, begining with line segment for the longest lags, until adequate fit with the data be obtained.

Direct Numerical Solution. The main objection to direct calculation of the component distributions $a(t)$ and $b(n)$ is that, for fixed precision arithmetic, rounding errors accumulate and are likely to become significant when dealing with covariances at the longer lags.

Where the analyst has algebraic expressions for the measures of correlation (such that $L(1) \equiv I_\infty = K(1) \equiv J_\infty$) equations (19) and (20) can be employed directly to produce the *pgf*'s of the component distributions of the appropriate batch renewal process.

In considering the case of measurements of the correlation of actual traffic it is apparent that there are fundamental problems to construction of a general numerical algorithm to determine the corresponding batch renewal process. For example, equation (19) gives the recurrence relationship

$$a(t) = \frac{K_t + 2\lambda - \sum_{\ell=1}^{t-1} a(\ell)(K_{t-\ell} + 2\lambda)}{K_0 + L_0 - K(1-) + 1 + \lambda} \qquad \text{for } t = 2, 3, \ldots,$$

which calculation requires the difference between numbers of similar magnitude.

Firstly there is the (lesser) difficulty of estimating $K(1) = L(1)$, which is equivalent to estimating geometric tails to complement the truncated sets of measurements. Secondly the form of the recurrence relationship suggests that the effect of rounding errors might accumulate rapidly. This difficulty is inherent. By defining ϕ_ℓ by its generating function

$$\sum_{\ell=0}^{\infty} \phi_\ell \, \omega^\ell \overset{\Delta}{=} \frac{1}{1 - A(\omega)}$$

the essence of the recurrence relation is seen to be

$$a(\ell) = \phi_\ell - \sum_{t=1}^{\ell} a(t)\phi_{\ell-t} \qquad \ell = 1, 2, \cdots$$

where

$$\phi_0 = 1 \quad \text{and} \quad \phi_\ell = \frac{K_\ell + 2\lambda}{K_0 + L_0 - K(1) + 1 + \lambda} \quad \ell = 1, 2, \cdots .$$

Consequently, actual traffic measurements should be converted to an algebraic representation and then the algebraic method be used.

Generally, when the logarithm of the measured correlation be plotted (with error bars) against the corresponding lag, the resulting graph may be (or may be approximated by) a series of straight line segments — which is equivalent to saying that the correlation function is (approximated by) the weighted sum of geometric terms or that the generating function ($K(\omega)$ or $L(z)$) is a rational function (of ω or of z, respectively).

The simplest form of the graphs of $\log K_\ell$ against ℓ and $\log L_\ell$ against ℓ is when both are straight line graphs. This case may arise naturally, because of the characteristics of the traffic source, or may arise from the practicalities in actual traffic measurements. The size of the data sets may be limited by the time period for which the traffic process may be regarded as being wide sense stationary. Then the practical recourse is to fit a straight line to the data points. When $\log K_\ell$ and $\log L_\ell$ are linear in ℓ the corresponding batch renewal process is of the simplest non-trivial form. It is the form which is used for the arrival process to the queue in sections 4.3.

SRD Batch Renewal process in Continuous Time. The results for the batch renewal process in continuous time are similar to those for the discrete-time domain. Corresponding to (13) and (14) we have[16]

$$K(\theta) = b\left(C_b^2 + \frac{1+A(\theta)}{1-A(\theta)} - \frac{1}{a}\frac{2}{\theta}\right) \tag{21}$$

and

$$L(z) = b\left(C_a^2 + \frac{1+B(z)}{1-B(z)} - \frac{1}{b}\frac{1+z}{1-z}\right) \tag{22}$$

where $A(\theta)$ is now the Laplace transform of the density of intervals between batches and $K(\theta)$ generates the count covariances.

By considering the limits

$$\begin{aligned} K(0) &= \lim_{\theta\to 0} K(\theta) = b\left(C_a^2 + C_b^2\right) \\ L(1-) &= \lim_{z\to 1-} L(z) = b\left(C_a^2 + C_b^2\right) \\ K(\infty) &= \lim_{\theta\to\infty} = b\left(C_a^2 + 1\right) \\ L(0) &= \lim_{z\to 0} L(z) = b\left(C_a^2 + 1 - \frac{1}{b}\right) \end{aligned}$$

we obtain[16]

$$A(\theta) = 1 - \frac{K(\infty) + L(0) - K(0) + 1}{K(\theta) + L(0) - K(0) + 1 + \lambda\dfrac{2}{\omega}} \tag{23}$$

and

$$B(z) = 1 - \frac{K(\infty) + L(0) - K(0) + 1}{K(\infty) + L(z) - K(0) + \dfrac{1+z}{1-z}}\,. \tag{24}$$

3.2 Long Range Dependent Processes

We shall say that a process is long range dependent if either the count covariances are not summable or the interval covariances are not summable.

This subsection addresses the case in which either the count covariances are not summable or the interval covariances are not summable. For illustration, consider the case when the interval covariances are not summable but the count covariances are summable. In this case, by taking limits as $z \to 1-$ in equation (14), $L(1-) = \infty$ and the SCV C_b^2 of batch size is infinite — an instance of "the infinite variance syndrome". Consequently, even though the sample variance of counts is finite (necessarily), we have to treat the counts as arising from a process with infinite variance. Thus, the generating function $K(\omega)$ cannot be used unmodified. Instead, define $K_+(\omega)$ by

$$K_+(\omega) = \sum_{\ell=1}^{\infty} K_\ell\, \omega^\ell = {}^2\!/\!_\lambda \sum_{\ell=1}^{\infty} \mathsf{Cov}\left[c(t), c(t+\ell)\right] \omega^\ell \tag{25}$$

and then the analysis of the preceding sub-section can be adapted to yield

$$A(\omega) = 1 - \frac{L(0) - K_+(1-) + 1 + \lambda}{K_+(\omega) + L(0) - K_+(1-) + 1 + \lambda \dfrac{1+\omega}{1-\omega}} \tag{26}$$

and

$$B(z) = 1 - \frac{L(0) - K_+(1-) + 1 + \lambda}{L(z) - K_+(1-) + \dfrac{1+z}{1-z} + \lambda}\,. \tag{27}$$

For the analyst to have decided that the process be long range dependent, it is likely that the graph of log covariance against log lag would be asymptotically linear. If the asymptotic slope be $-s$ then $K(\omega)$ can be represented as the sum of two terms, with one term having the form $C\big((1-\omega)^{-1-s} - 1\big)$. Components may be extracted progressively until adequate fit with the data be obtained.

3.3 Improper Batch Renewal Processes

For some correlation structures the corresponding batch renewal process is improper, i.e. the constituent distributions contain negative probabilities. The cause can be seen by considering equation (18) for SRD processes or the corresponding relation (such as)

$$L(0) - K_+(1-) = 2b - 1 - \lambda$$

for LRD processes. In each case, the left hand side of the equation may be so small that the mean batch size b does not exceed 1: indeed b may be negative. Whereas, from the formulation of the batch renewal process, b is the expected size of a non-empty batch.

The question then arises as to whether such improper batch renewal process may be used for performance prediction of (for example) a buffer fed by the traffic. Possible approaches are discussed in Section 9.

4 Simple Queues Fed by Batch Renewal Process Traffic

4.1 GI$^\mathrm{G}$/D/1/N

Consider the discrete-time GI$^\mathrm{G}$/D/1/N censored queue under DF.

- The arrivals are from the batch renewal process.
- The service time is fixed at one slot.
- The queue capacity is N, including the customer in service.
- At an epoch, an arrival may take the place released by a departure.
- When the system becomes full, other customers in the arriving batch are lost.

Events (arrivals and departures) occur at discrete points in time (epochs) only. The intervals between epochs are called *slots* and, without loss of generality, may be regarded as being of constant duration. At an epoch at which both arrivals and departures occur, the departing customers release the places, which they had been occupying, to be available to arriving customers (*departures first* memory management policy). The service time for a customer is one slot and the first customer arriving to an empty system (after any departures) receives service and departs at the end of the slot in which it arrived (*immediate service* policy). By GI^{G} arrivals process is meant the intervals between batches are independent and of general distribution and the batch size distribution is general (batch renewal process).

Let the state of the queue be the number of customers in the queue (buffered or receiving service). Because the transitions at epochs are deemed to be instantaneous the *pmf* for the stationary distribution of queue length is simply the time average probability for each state observed during slots only.

The solution described in this section is based upon the observation that, between arrival epochs, the state in each slot is determined completely by the state in the previous slot. (By 'arrival epoch' is meant an epoch at which there is a batch of arrivals). The number of customers in the queue is reduced by one departure at each epoch until either the queue becomes empty or an arrival epoch is reached. Thus, given the state in the slot immediately following an arrivals epoch, the evolution of the queue until the next arrivals epoch depends only upon the interval between the two batches of arrivals. But, at the arrival epoch, the change in state (after accounting for any departure at that epoch) depends only upon the size of the arriving batch.

The steady state behaviour of the queue may be solved by considering the state at points immediately before and immediately after each batch of arrivals. It is apparent that each point is an embedding point for a Markov chain.

Consider two (related) Markov chains embedded at arrival epochs.

- For the first chain (chain 'A'), the state is the number of customers in the queue after allowing for any departure at that epoch but discounting the new arrivals at that epoch. Let $p_N^A(n)$ be the probability that the state be n, $n = 0, \ldots, N{-}1$ (where N is the capacity of the queue).
- For the second chain (chain 'D'), the state is the number of customers in the queue after allowing for any departure at that epoch but including the new arrivals. Let $p_N^D(n)$ be the probability that the state be n, $n = 1, \ldots, N$.

Equivalently, one might treat the departures as *actually* occuring before arrivals and focus on the two points 1) at which the departures have already gone and the arrivals have not yet come, 2) immediately after the arrivals.

To see the relation between the two Markov chains, first consider the state of each chain at an arrival epoch. Chain 'D' may be in state n, $n = 1, \ldots, N{-}1$, when chain 'A' is in state k, $k = 0, \ldots, n{-}1$, and there be just $n{-}k$ arrivals in the batch. Alternatively, chain 'D' may be in state N when chain 'A' is in state k, $k = 0, \ldots, N{-}1$, and there be at least $N{-}k$ arrivals in the batch. Therefore

$$p^D_N(n) = \begin{cases} \displaystyle\sum_{k=0}^{n-1} p^A_N(k)\, b(n{-}k) & n = 1, \ldots, N{-}1 \\ \displaystyle\sum_{k=0}^{N-1} p^A_N(k) \sum_{r=N-k}^{\infty} b(r) & n = N \end{cases} \tag{28}$$

Next, consider the state of each chain at successive arrival epochs. At the later epoch the chain 'A' may be in state n, $n = 1, \ldots, N{-}1$, when chain 'D' is in state k, $k = n{+}1, \ldots, N$, at the earlier arrival epoch and there be just $k{-}n$ departures in the interval between the two arrival epochs, i.e. the interval is $k{-}n$ slots long. Alternatively, at the later epoch the chain 'A' may be in state 0 when chain 'D' is in state k, $k = 1, \ldots, N$, at the earlier epoch and the interval is at least k slots long. Therefore

$$p^A_N(n) = \begin{cases} \displaystyle\sum_{k=1}^{N} p^D_N(k) \sum_{t=k}^{\infty} a(t) & n = 0 \\ \displaystyle\sum_{k=n+1}^{N} p^D_N(k)\, a(k{-}n) & n = 1, \ldots, N{-}1 \end{cases} \tag{29}$$

Performance statistics and measures of interest are obtainable, in obvious ways, in terms of the solutions to equations (28) and (29) for the two Markov chains.

Queue Length Distribution. If the second Markov chain (chain 'D') be in state k at an arrival epoch then, in each successive slot of the interval to the next arrival epoch, the queue will be in state k, $k{-}1$, etc. until either the queue becomes empty or the next batch arrives.

If the interval to the next arrivals epoch be t slots then, if $t \leq k$, the queue visits states $k, \ldots, k{-}t{+}1$ for one slot each but, if $t > k$, the queue visits states $k, \ldots, 1$ for one slot each and then remains in state 0 for the remaining $t{-}k$ slots (see Figure 2).

Thus, the time average probability $p_N(n)$ that the queue be in state n is given by

$$p_N(n) = \begin{cases} \displaystyle\frac{1}{a} \sum_{k=1}^{N} p^D_N(k) \sum_{t=k+1}^{\infty} (t - k)\, a(t) & n = 0 \\ \displaystyle\frac{1}{a} \sum_{k=n}^{N} p^D_N(k) \sum_{t=k-n+1}^{\infty} a(t) & n = 1, \ldots, N \end{cases} \tag{30}$$

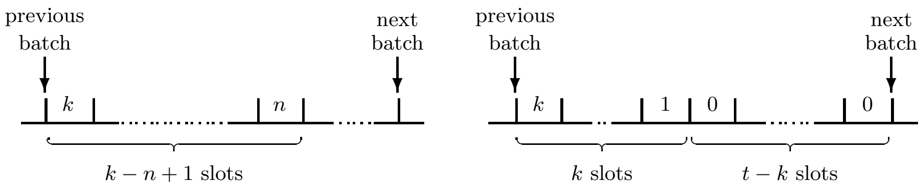

Fig. 2. Ways in which queue length n ($n > 0$) and queue length 0 may be reached during an interval between batches

in which $a(t)/a$ is the probability that an arbitrary slot be at any given position within an interval of t slots between batches.

Blocking Probability. If chain 'A' be in state k at an arrival epoch there are $N{-}k$ places available in the queue to the arriving batch. So, if the batch contain $N{-}k{+}r$ arrivals, r of those arrivals are blocked.

For an arbitrary arrival, the probability that it be in a batch of size n is $nb(n)/b$ and the probability that it be in any given position in the batch is $1/n$. Therefore the marginal probability π_N^B that any individual arrival be turned away is

$$\pi_N^B = \sum_{k=0}^{N-1} p_N^A(k) \sum_{r=1}^{\infty} \frac{r}{b}\, b(N{-}k{+}r) \tag{31}$$

which, by reference to (28–30), can be seen to satisfy the flow balance equation

$$\lambda(1-\pi_N^B) = 1 - p_N(0)\,. \tag{32}$$

Waiting Time. Because service time is one slot per customer the waiting time of an arrival is given by its position in the queue at the instant of its arrival, given that the arrival enter the queue and not be blocked. If there be k in the queue (i.e. Markov chain 'A' be in state k at that arrival epoch) then the arrival in position $t{-}k$ of the batch will enter the queue provided that $t \le N$ and will then remain in the queue for t slots. Thus,

$$\begin{aligned}
&\mathsf{P}\left[\text{waiting time } {=}t \;\middle|\; k \text{ in queue, arrival not blocked}\right]\\
&\quad = \frac{\mathsf{P}\left[\text{customer in position } t{-}k \text{ of batch, } k < t \le N\right]}{\mathsf{P}\left[\text{arrival not blocked}\right]}\\
&\quad = \frac{\sum_{r=t-k}^{\infty} \dfrac{rb(r)}{b}\dfrac{1}{r}}{1-\pi_N^B} = \frac{\sum_{n=t}^{\infty} b(n{-}k)}{b(1-\pi_N^B)}
\end{aligned}$$

Therefore the conditional probability $w_{_N}(t)$ that an arbitrary arrival spend t slots in the queue, given that the arrival not be blocked, is given by

$$w_{_N}(t) = \frac{\displaystyle\sum_{n=t}^{\infty}\sum_{k=0}^{t-1} p^A_{_N}(k)b(n-k)}{b(1-\pi^B_{_N})} = \frac{\displaystyle\sum_{n=t}^{\infty}\sum_{k=0}^{t-1} p^A_{_N}(k)b(n-k)}{\displaystyle\sum_{t=1}^{N}\sum_{n=t}^{\infty}\sum_{k=0}^{t-1} p^A_{_N}(k)b(n-k)} \tag{33a}$$

which may be manipulated, using equations (28–30) to show that

$$w_{_N}(t) = \frac{p_{_N}(t)}{1-p_{_N}(0)} \qquad \text{for } t = 1, \ldots, N. \tag{33b}$$

It may be observed that the relation (33b), between waiting time and stationary queue length, is what should be expected when the service time is deterministic at one customer per slot [21].

Example. This section shows some numerical results for the batch renewal process with LRD counts. In the chosen case, the inter-batch *pgf* is of the form

$$A(\omega) = 1 - a + a\omega + (a-1)(1-\omega)^{2-s} \tag{34}$$

where $0 < s < 1$ and, for a proper *pmf*, $1 < a < \dfrac{2-s}{1-s}$, and the batch size *pmf* is of the form

$$b(n) = \begin{cases} 1-\eta & n = 1 \\ \eta\nu(1-\nu)^{n-2} & n = 2, 3, \ldots \end{cases} \tag{35}$$

For the graphs in Figures 3 and 4 the following parameter values were used for three values of s.

- $a = 2$.
- $\eta = 1/8$, $\nu = 3/8$ (giving $b = 4/3$, $\lambda = 2/3$).
- Queue capacity $N = 100$.

For Figure 3 the calculation is the recursive relation derived by inverting equation 25 of Section 3.2. The graphs show the positive count covariances; at small lags some covariances are zero or negative in the cases shown. It is clear that asymptotic slopes of the graphs approach $-s$ rapidly.

For Figure 4 the algorithm is the general method given in Section 4.1.

4.2 Continuous-Time GI$^{\mathrm{G}}$/G/1/N Queues

The approach, taken in the previous section, of considering two Markov chains embedded immediately before and immediately after each batch of arrivals can

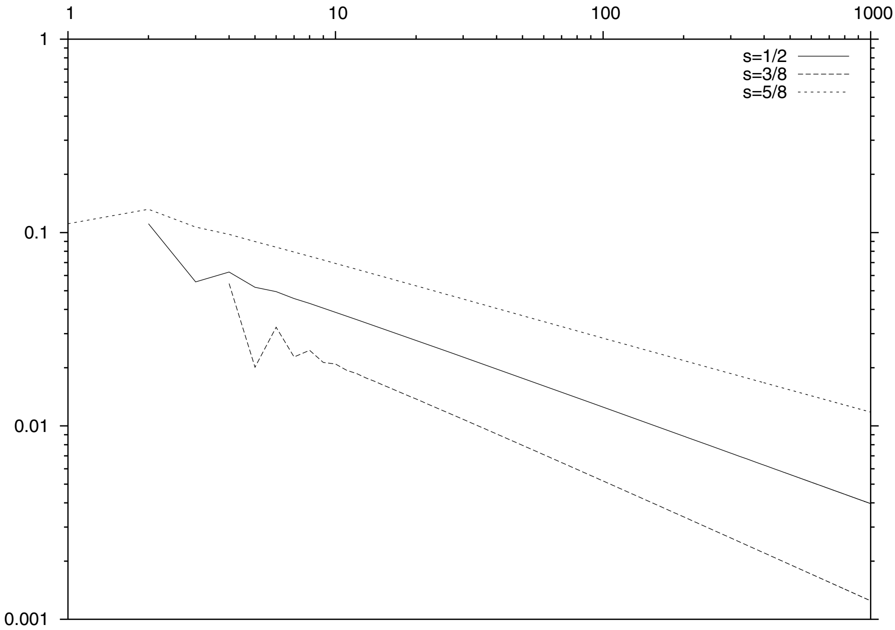

Fig. 3. Count covariances against lags (logarithmic scales)

be extended to a queue with general service time distribution. Li obtained the following general results in the continuous-time domain[16] in terms of the density $g(k,t)$ of the probability that k customers can complete service in time t (i.e. k depart provided that there are at least k in the system, otherwise all customers depart).

Relationship betwwen the chains. c.f. (28) and (29),

$$p_N^D(n) = \begin{cases} \displaystyle\sum_{k=0}^{n-1} p_N^A(k)\, b(n-k) & n = 1,\ldots,N-1 \\ \displaystyle\sum_{k=0}^{N-1} p_N^A(k) \sum_{r=N-k}^{\infty} b(r) & n = N \end{cases} \tag{36}$$

$$p_N^A(n) = \begin{cases} \displaystyle\sum_{k=1}^{N} p_N^D(k) \int_0^{\infty} \sum_{r=k}^{\infty} g(r,t)\, a(t)\, dt & n = 0 \\ \displaystyle\sum_{k=n}^{N} p_N^D(k) \int_0^{\infty} g(k-n,t)\, a(t)\, dt & n = 1,\ldots,N \end{cases} \tag{37}$$

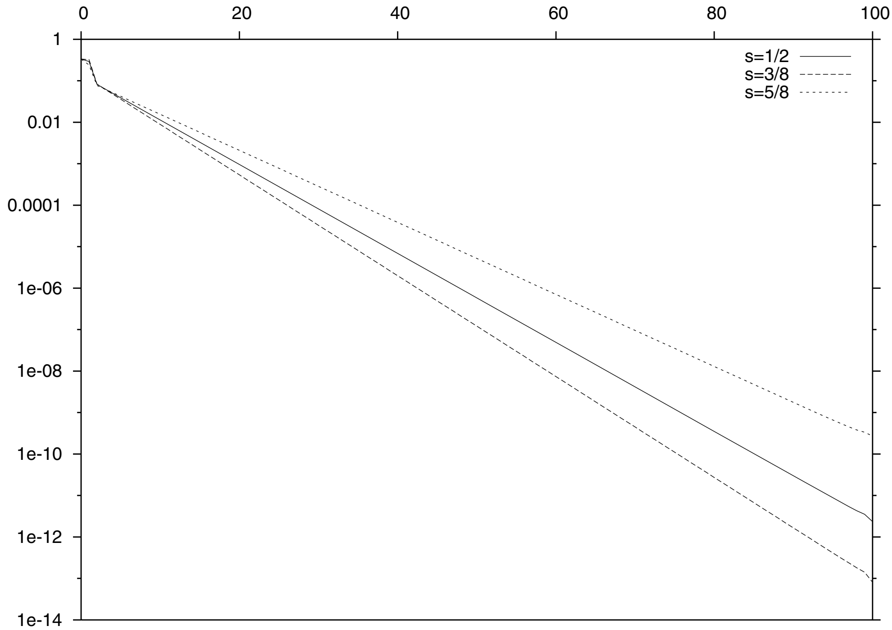

Fig. 4. Queue length distribution (logarithmic scale)

Queue Length Distribution. c.f. (30)

$$p_{_N}(n) = \begin{cases} \dfrac{1}{a}\displaystyle\sum_{k=1}^{N} p^{D}_{N}(k) \int_0^\infty \int_0^t \sum_{r=k}^{\infty} g(r,s)\, ds\, a(t)\, dt & n = 0 \\ \dfrac{1}{a}\displaystyle\sum_{k=n}^{N} p^{D}_{N}(k) \int_0^\infty \int_0^t g(k-n,s)\, ds\, a(t)\, dt & n = 1,\ldots,N \end{cases} \tag{38}$$

Blocking Probability. c.f. (31)

$$\pi^{B}_{N} = \sum_{k=0}^{N} p^{A}_{N}(k) \sum_{r=1}^{\infty} \frac{r}{b}\, b(N-k+r) \tag{39}$$

Waiting Time Density. c.f. (33a)

$$w_{_N}(t) = \frac{\displaystyle\sum_{k=0}^{N-1} p^{A}_{N}(k) \sum_{i=1}^{N-k} g(k+i,t) \sum_{r=i}^{\infty} b(r)}{b(1-\pi^{B}_{N})} \tag{40}$$

4.3 The sGGeo/D/1/N Queue

The sGGeo[1] is the simplest batch renewal process in which there is both count correlation and interval correlation.

$$a(t) = \begin{cases} 1-\sigma & t=1 \\ \sigma\tau(1-\tau)^{t-2} & t=2,3,\ldots \end{cases} \qquad b(n) = \begin{cases} 1-\eta & n=1 \\ \eta\nu(1-\nu)^{n-2} & n=2,3,\ldots \end{cases} \tag{41}$$

For the sGGeo, the covariances of counts and the covariances of intervals (between individual arrivals) both decline geometrically, viz.

$$\begin{aligned} \mathsf{Cov}\left[c(t), c(t+\ell)\right] &= \lambda^2(a-1){\beta_a}^{\ell} && \text{where } \beta_a = 1-\sigma-\tau \\ \mathsf{Cov}\left[x(n), x(n+\ell)\right] &= \frac{1}{\lambda^2}(b-1){\beta_b}^{\ell} && \text{where } \beta_b = 1-\eta-\nu \end{aligned} \tag{42}$$

Remarks. *The sGGeo may be appropriate to model a traffic source for which only the first two moments of message size and of intervals between messages are known. It is also the appropriate model of measured traffic when either the decline in covariances is geometric (c.f. equation 42) or there be so few measurements that the best procedure is to fit a straight line to the plot of the logarithms of measured covariances against lags.*

The sGGeo/D/1/N Queue Length Distribution. By using the particular forms (41) in application of the general methods of Section 4.1 it is seen that the sGGeo/D/1/N queue length distribution has the form

$$p_{_N}(n) = \begin{cases} \dfrac{1}{Z_{_N}}(1-\lambda) & n=0 \\ \dfrac{1}{Z_{_N}}\lambda(1-y) & n=1 \\ \dfrac{1}{Z_{_N}}\lambda y(1-x)x^{n-2} & n=2,\ldots,N-1 \\ \dfrac{1}{Z_{_N}}\lambda y(1-x)x^{N-2}\dfrac{1}{1-\beta_a x} & n=N \end{cases} \tag{43}$$

where $\beta_a = 1-\sigma-\tau$ and $\beta_b = 1-\eta-\nu$ are as defined at (42) and x and y are given by

$$1-y = \frac{1-x}{1-\beta_b}, \quad x = \frac{\sigma(1-\eta-\nu)+\eta}{\sigma+(1-\sigma-\tau)\eta}$$

and the normalizing constant $Z_{_N}$ may be written

$$Z_{_N} = 1 - \lambda y \frac{1-\beta_a}{1-\beta_a x} x^{N-1} \tag{44}$$

[1] The sGGeo process is so named because both the constituent distributions (i.e. of the batch sizes and of the intervals between batches) have the form of a Generalized Geometric (GGeo) shifted by one.

The sGGeo/D/1/N Mean Queue Length

$$L_N = \frac{1}{Z_N}\left(\lambda + \lambda\frac{b-1}{1-\lambda}\frac{1-\beta_a\beta_b}{(1-\beta_a)(1-\beta_b)}(1-x^{N-1}) \right.$$
$$\left. +N\frac{b-1}{a-1}\frac{\beta_a}{1-\beta_a}x^{N-1}\right) \qquad (45)$$

Figure 5 shows the effect of correlation on mean queue length and that the effect is constrained for small values of buffer capacity N.

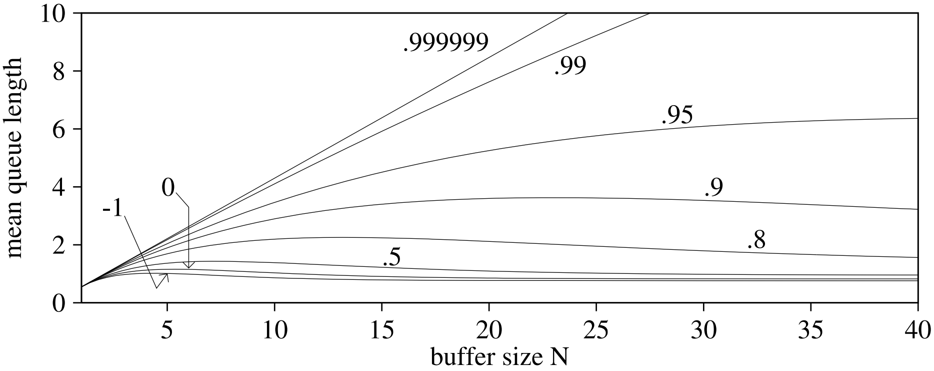

Fig. 5. Mean queue length against buffer size N for mean batch size $b = 1.5$, mean interval $a = 7.5$ slots between batches, intensity $\lambda = 0.2$, $\beta_a = 0.8$ and various values of β_b

The sGGeo/D/1/N Blocking Probability. Because the probability π^B_N, that any individual arrival be turned away, satisfies the flow balance equation $\lambda(1-\pi^B_N) = 1 - p_N(0)$, it follows from equation (43) that

$$\pi^B_N = \frac{1-\lambda}{\lambda}\frac{1-Z_N}{Z_N} = \frac{(1-\lambda)y\dfrac{1-\beta_a}{1-\beta_a x}x^{N-1}}{1-\lambda y\dfrac{1-\beta_a}{1-\beta_a x}x^{N-1}}\,. \qquad (46)$$

This relation shows that the asymptotic behaviour of π^B_N with increasing buffer size N is log-linear:

$$\frac{\pi^B_{N+1}}{\pi^B_N} \longrightarrow x \quad \text{as } N \longrightarrow \infty \qquad (47)$$

Indeed π^B_N may approach its asymptote for relatively small values of N, as is illustrated by the graphs in Figure 6. Expressions (47) and (44) also show that

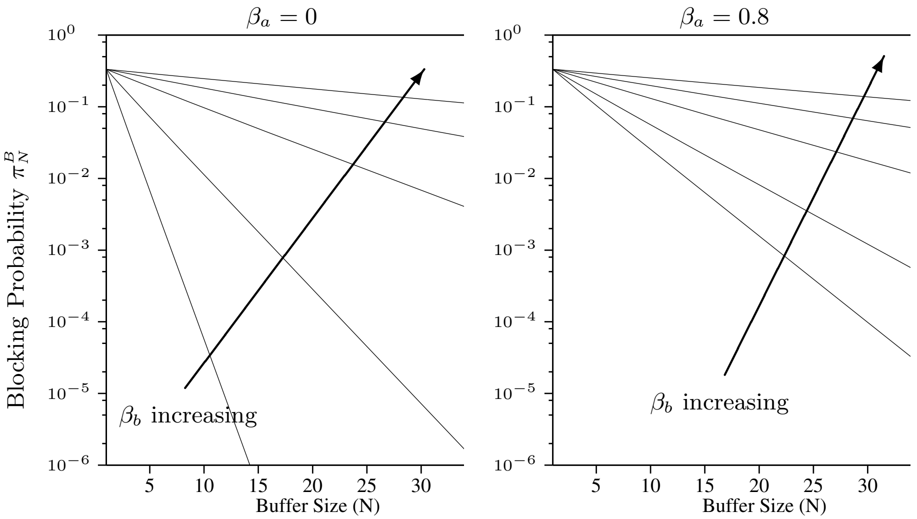

Fig. 6. Blocking probability against buffer size for mean batch size $b = 1.5$, mean interval $a = 7.5$ slots between batches, intensity $\lambda = 0.2$ with $\beta_b = 0, 0.5, 0.8, 0.9, 0.95$

$$\pi_N^B \to \pi_1^B = 1 - \frac{1}{b} \quad \text{as } x \to 1 \tag{48}$$

i.e. as $\beta_a \to 1$ or as $\beta_b \to 1$.

5 Effect of a Queue on Correlation—Creation of Burst Structure

The departure process from a GI$^{\mathrm{G}}$/D/1/N queue is determined by the cycle of busy period followed by idle period. For each slot the server is busy there is a departure. Consecutive departures constitute a burst. Because the intervals between batches are independent each cycle of busy period followed by idle period is independent of other busy/idle cycles. The distribution of one burst length (busy period) and successive silence (idle) period is governed by the following relationships.

$$busy(n,i) = \sum_{k=1}^{\min(N,n)} busy(n,i;k\,)b_N(k) \tag{49}$$

$$busy(n,i;k) = \begin{cases} a(n{+}i) & n = k \\ \displaystyle\sum_{\ell=0}^{k-1} \sum_{q=\ell+1}^{\min(N,n-k+\ell)} a(k{-}\ell)\, b_{N-\ell}(q{-}\ell) busy(n{-}k{+}\ell, i; q) & n > k \end{cases} \tag{50}$$

where $busy(n, i)$ is the marginal probability that the server be busy for n slots and idle for i slots, $busy(n, i; k)$ is the conditional probability that the server be busy for n slots and idle for i slots given that the busy period begin with k customers in the queue and where $b_k(n)$ is the probability that just n arrivals join the queue from a batch when there be k spaces in the queue.

Observe that, for $n < N$, both $busy(n, i; k)$ and $busy(n, i)$ take the same values in the finite buffer system as they do in the infinite buffer system. Observe further that

$$busy(n{+}1, i; 1) = \sum_{q=1}^{\min(N,n)} a(1)\, b_N(q)\, busy(n, i; q) = a(1)\, busy(n, i) \qquad (51)$$

and that, when the idle period is independent of the busy period, the probability that the idle period be i slots is

$$\frac{a(i{+}!1)}{1 - a(1)} \qquad (52)$$

Example

When the batch renewal process has both batch sizes and intervals between batches distributed as shifted Generalized Geometric (as in the example of Section 4.3) the idle periods are independent of the busy periods and are distributed geometrically. Thus only the busy period distribution needs to be considered. A typical form is shown in figure 7.

In departures from an infinite buffer the burst length is distributed as the sum of two geometrics. For moderate values of β_a (correlation of counts) there is a marked knee in the graph.

For finite buffers the form of the burst length distribution is more complex. Two features are obvious in figure 7.

First, there is a ‘hump’ or accumulation of mass at burst lengths just longer than the buffer size N. The reason is intuitively obvious because, on the one hand, the probability of any busy periods less than N slots is the same for both finite and infinite buffer queues but, on the other hand, in comparison to the infinite buffer the finite buffer reduces the probability of longer busy periods.

Secondly, the tail of the distribution depends upon the location of the knee. This is most readily explicable in terms of the limited ‘memory’ of the finite buffer queue: at any time the state of a queue of capacity N and deterministic service time of one slot is independent of its state at any time which is more than N slots earlier. If the knee occurs after the finite buffer distribution separates from the infinite buffer distribution (at burst length N) then the queue ‘memory’ includes the knee, which appears as waviness in the tail. Whereas, if the finite buffer distribution does not include the knee the tail is relatively straight.

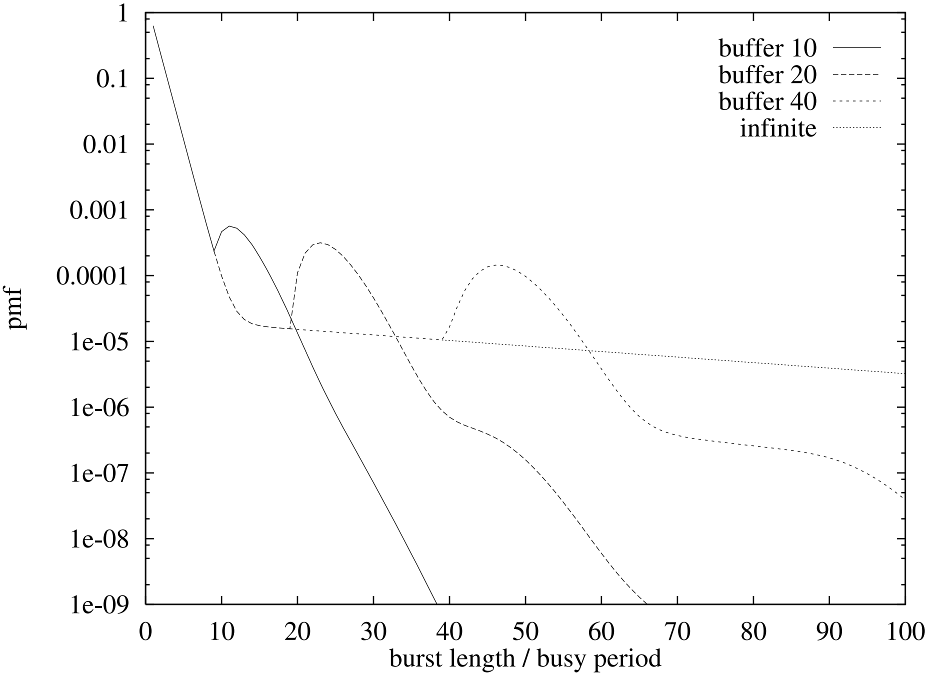

Fig. 7. *Pmf* of departure process burst length (busy period) for mean batch size $b = 1.25$, mean interval $a = 6.25$ slots between batches, intensity $\lambda = 0.25$, $\beta_a = 0.25$ and $\beta_b = 0.99$ for finite buffers of size 10, 20 and 40 and for infinite buffer

6 Equivalence in Discrete Space Discrete Time Processes

In this section we consider a large class of internal models of discrete time traffic processes. The discussion has relevance to the design of the experiment that is described in section 7 and also to points raised in Section 9. The batch renewal process is always representable in the class, e.g. as a trivial batch Markov renewal process that has one phase only.

Correspondences between some representations of processes are well known, for example between the semi-Markov and Markov renewal processes [2] and between the MAP and Neuts process [17]. In the context of discrete time processes, correspondence between other representations can be seen and this observation leads to the notion of a class of processes in which each member admits a variety of representations, so that the class might equally well be defined in terms of any of the representations. The class that is considered here is that of *all processes that admit representation as MMBBP's over countable phase spaces.* Figure 8 shows some relationships between three representations of that class: batch Markov renewal process or SMP; DBMAP; MMBBP.

1. An arbitrary MMBBP may be described as a batch Markov renewal process in which the sojourn (in a phase between two points of the batch Markov

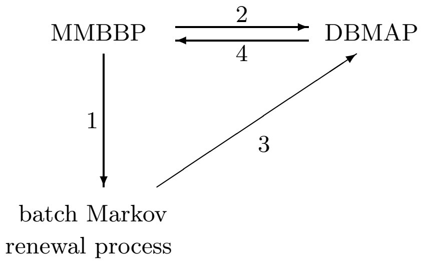

Fig. 8. Equivalence of discrete time models. The arrows show subset relations, e.g. arrow 1 shows that the Markov modulated processes are special cases of the batch Markov renewal process.

renewal process) is always one slot long. Therefore the set of MMBBP's is a subset of the set of the batch Markov renewal processes.

2. An arbitrary MMBBP may be described as a DBMAP in which the number of arrivals generated at an epoch is conditionally independent of the phase in the next slot given the current phase of the DBMAP. Therefore the set of MMBBP's is a subset of the set of the DBMAP's.

Then, by virtue of the transitivity of the subset relation, it is sufficient to show two further subset relations, such as those labelled 3 and 4 in Figure 8, to establish equivalence between all three representations.

3. For an arbitrary discrete time batch Markov renewal process there may be constructed a DBMAP that is equivalent to the MMBBP in the sense that, at every point in the evolution of the batch Markov renewal process, the same behaviour of the constructed DBMAP is exactly the same as that of the batch Markov renewal process. So, each batch Markov renewal process is representable as a DBMAP. Figure 9 illustrates the essential feature in the construction, which is that the DBMAP should contain a phase j_t for each phase j of the Markov renewal process and each possible sojourn t slots in phase j of the Markov renewal process. Whenever, in the Markov renewal process, there are k arrivals at point n together with a transition from phase i to phase j for a sojourn of t slots there should be correspondingly at epoch τ_n in the DBMAP k arrivals together with a transition from phase i_1 to

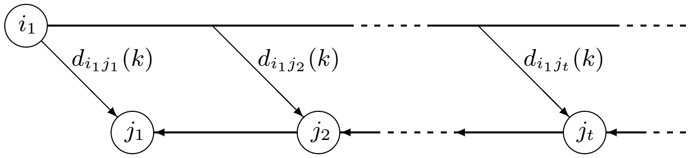

Fig. 9. DBMAP phases and transitions corresponding to Markov renewal process phase transition $i \rightarrow j$

DBMAP phase ϕ		i		j		k	
MMBBP phase ϕ'		(i, j)		(j, k)		(k, ℓ)	
slot number		t		$t+1$		$t+2$	
epoch number	$t-1$		t		$t+1$		$t+2$

Fig. 10. Relation between phase $\phi(\cdot)$ of DBMAP and phase $\phi'(\cdot)$ of equivalent MMBBP: $\phi'(t) = (i, j)$ whenever $\phi(t) = i$ and $\phi(t+1) = j$

phase j_t followed by successive transitions to phases $j_{t-1}, \ldots, j_1$ at the $t-1$ successive epochs $\tau_n+1, \ldots, \tau_n+t-1 = \tau_{n+1}-1$.

4. For an arbitrary DBMAP an equivalent MMBBP may be constructed and so, in that sense, each DBMAP is representable as a MMBBP. Figure 10 illustrates the essential feature in the construction, which is that the MMBBP should contain a phase (i, j) for each phase transition $i \to j$ of the DBMAP. Whenever in the DBMAP there be n arrivals generated at an epoch t together with a transition from phase i to phase j and followed (at epoch $t+1$) by a transition to phase k, there should be correspondingly at epoch t in the constructed process n arrivals generated and transition from phase (i, j) to phase (j, k).

The equivalence between the three representations of traffic models may suffice to show why the MMBBP may be used for the experiment that is described in Section 7. However, restricting consideration to just three types of internal model does seem arbitrary. That the equivalence might be more general provokes the conjecture that the class (of *all processes that admit representation as MMBBP's over countable phase spaces*) might properly suffice for internal models of all realizable (discrete time) traffic.

7 Biased Results from Other Models

The batch renewal process is, in information theoretic terms, the least biased choice of traffic process given only the customary measures of correlation (e.g. indices of dispersion) [12]. In the batch renewal process there is no semblance of burst structure or, indeed, of any other feature other than correlation. The batch renewal process may be described fairly as "pure correlation". The question then arises as to what is the effect upon (say) queue performance caused by the bias of chosing some other model for traffic that is characterised by correlation? This section describes an experiment designed to provide some insight.

The reference model chosen was the MMBBP/D/1 queue. Because nothing was known about the impact of chosing some process other than the batch renewal process, it was desirable to choose a form of arrivals process that readily permitted extremes of behaviour. The arrivals process was chosen to be a

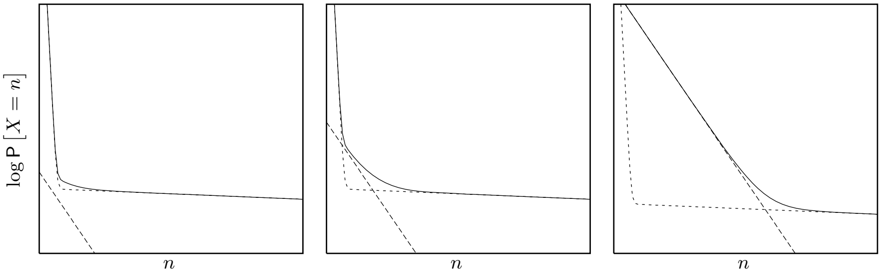

Fig. 11. Effect of adding a third phase to a 2-phase distribution $\mathsf{P}\left[X=n\right]=(1-\alpha)\big(A(1-x_1)x_1{}^n+(1-A)(1-x_2)x_2{}^n\big)+\alpha(1-x_3)x_3{}^n$ with $A=0.998$, $x_1=0.01$, $x_2=0.99$, $x_3=0.7$ and $\alpha=10^{-4},10^{-2},0.5$. Solid line: 3-phase distribution $\mathsf{P}\left[X=n\right]$; dotted line: 2-phase distribution, as with $\alpha=0$; dashed line: contribution $\alpha(1-x_3)x_3{}^n$ of third phase.

2-phase MMBBP in which the distributions of counts in each phase were GGeo (an extremal 2-phase distribution). Figure 11 is intended to show why it was thought that two phases might provide extreme behaviour by illustrating how often, in a distribution composed of the weighted sum of geometric terms, either two phases dominate or the distribution is 'smoothed'. The other choices — deterministic service, with one customer served per slot, and infinite capacity — were made to avoid effects not directly related to the arrivals process.

The first part of the experiment was conducted with high intensity traffic ($\lambda=0.9$ customer/slot), mean intensity 0.8 in one phase and 1.1 in the other but while varying mean sojourn in the phases and variance of counts in each phase. For each set of MMBBP parameters, the corresponding batch renewal process (i.e. that with counts and intervals covariances identical to those of the MMBBP) was determined as described in Section 3, the queue length distributions were computed for both the MMBBP/D/1 and the $\cdot$/D/1 queue fed by the corresponding batch renewal process. Figure 12 shows a typical result. The smaller geometric rate (steeper first segment) in the MMBBP/D/1 queue length distribution clearly implies lower waiting time, less jitter and, extrapolating to the finite buffer case, lower cell loss rate as compared with the distribution attributable to the correlation alone. In other words, the MMBBP yields optimistic results in the cases considered.

That observation from the first part was formulated as *the proposition*

> the smallest geometric rate (steepest segment) in the MMBBP/D/1 queue length distribution is less than that of the queue fed by the corresponding batch renewal process

The experiment was then extended to randomly generated MMBBP's. For each of 2–, 3– and 6–phase models, 4000 MMBBP's were generated randomly: for each MMBBP the phase transition matrix entries were taken randomly from a uniform distribution and then each row of the matrix was normalised; for each MMBBP the mean intensity for each phase was taken randomly from a uniform

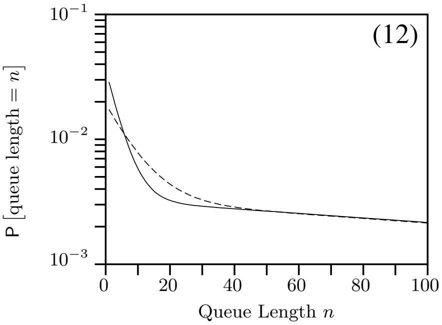

Fig. 12. Typical result from first part of experiment: solid line for MMBBP; dashed line for batch renewal process

distribution and scaled (5 times) to give overall mean rates (i.e. with respect to the stationary phase distribution for the particular MMBBP) of 0.1, 0.3, 0.5, 0.7 and 0.9. For each MMBBP with each value of mean rate, the largest poles of the MMBBP/D/1 and corresponding batch renewal process queue length distribution generating functions were computed and compared. Then all the (60 000) computations and comparisons were repeated after biasing each transition matrix to ensure diagonal dominance: the bias was applied to each row by halving each off-diagonal entry and increasing the diagonal entry in compensation.

Table 1 shows the proportion of cases in which *the proposition* was true. The difference between the "unbiased" and "biased" columns shows the significance of diagonal dominance in the phase transition matrix, especially for the 6-phase models, and that the effect is increased somewhat by high intensity. However

Table 1. Proportion of cases supporting *the proposition*

degree	λ	unbiased	biased
2	0.1	81%	>99%
	0.3	81%	>99%
	0.5	81%	>99%
	0.7	81%	>99%
	0.9	81%	>99%
3	0.1	80%	>99%
	0.3	80%	>99%
	0.5	80%	>99%
	0.7	80%	>99%
	0.9	80%	>99%
6	0.1	46%	88%
	0.3	46%	88%
	0.5	46%	89%
	0.7	46%	93%
	0.9	45%	94%

diagonal dominance alone is not sufficient (as shown by the "biased" column) nor is it necessary: in the "unbiased" column the proportion of results that support *the proposition* is far greater than that of diagonally dominant matrices among randomly generated matrices. The results strongly suggest that the MMBBP is likely to yield optimistic results, under conditions of positive traffic correlation of counts and intervals arising from long sojourn in each phase, c.f. the "biased" column of Table 1.

8 Cost Effective Approximation for Queueing Network Analysis

Exact analysis of queueing network models of communications systems can be intractable. So it is important to know in what circumstances a simpler process can approximate to tolerable accuracy the behaviour of the more complex process.

The formulation of the batch renewal process makes clear (c.f. equations 13 and 14) why correlation in SRD processes has much the same impact on measures of queue performance as does variability of interarrival time in renewal processes and variability of counts from batch Bernoulli processes. This observation leads to the notion that the effect of both counts and interarrival correlation in SRD traffic input to a queue might be captured (to some tolerable approximation) by variability in either counts alone (i.e. by an 'equivalent' batch Bernoulli process) or interarrival times (i.e. by an 'equivalent' renewal process). The accuracy on measures of queue performance of substituting some 'equivalent' processes for SRD traffic processes has been investigated in [1,3].

There had previously been some indication that such 'equivalent' processes provided tolerable accuracy in analysis of queueing networks. Typically, at each queue in the network the arrivals traffic is a superposition of departures from other queues and of external traffic. Clearly such traffic is correlated. However, in a number of fast algorithms based upon entropy maximisation and queue-by-queue decomposition (such as given in [10]), the input to each queue is treated *as if* it were completely free of correlation (c.f. Jackson networks). In [10] and similar algorithms, the input to each queue is, in effect, replaced by an 'equivalent' process with the same mean and variance of interarrival time. Nevertheless, the algorithms typically give good accuracy, in comparison with simulation results, when applied to networks of arbitrary topology and complexity. They had previously been used for networks for which input traffic was uncorrelated.

In a particular application of queue-by-queue decomposition, Kouvatsos et al.[13] considered networks fed by SRD traffic represented by the sGGeo process. In this context, the existing algorithm devised in [10], based upon the principle of maximum entropy, was extended to treat input sGGeo traffic as if substituted, with a tolerable accuracy, by uncorrelated traffic represented by an ordinary GGeo [10,12] for which the count distribution had the same mean and variance as those of the sGGeo.

Another queue, that might be used as a building block in analysis of an open queueing network of nodes with multiple servers and with correlated traffic is the GI^{G}/Geo/c, analyzed by Writtevrongel, Bruneel and Vinck [20]. This queue has a general batch renewal arrival process, infinite buffer and c servers with independent geometrically distributed service times. The analysis of this queue was based on the use of generating functions in conjunction with complex analysis and contour integration. Consequently, new analytic expressions for the generating functions of the system contents during an arrival slot were obtained as well as at an arbitrary slot. Moreover, the delay analysis the queue under the first-come-first-served discipline was presented.

An alternate approach, for each queue in a network with correlated input, is that followed by Laevens [14] to relate the output process of each queue to its input.

9 Open Problems and Research Topics

1. For given correlation, the batch renewal process is known to be the least biased choice of all possible processes which exhibit that correlation. However there are some patterns of correlation for which the corresponding batch renewal process is improper. So it might be as well to add to the previous statement the rider "... provided that the batch renewal process be proper". The question then arises as to what might be the least biased choice of process
   (a) when the correlation be such that the batch renewal process not be proper,
   (b) when other constraints (in addition to correlation) be applicable; perhaps the most interesting is that of the least biased choice of process for given count and intervals correlation given also that the traffic is on a line i.e. no simultaneous arrivals.
2. Given that there are some patterns of correlation for which the corresponding batch renewal process is improper, the question then arises as to whether such improper batch renewal process may be used for performance prediction of (for example) a buffer fed by the traffic.
   (a) For finite buffer queues, numeric methods may be derived, possibly based upon the general relations given in Section 4. In effect, that would be to attempt to find the stationary vector for a transition matrix which has some entries negative. Clearly, a naïve implementation would be unstable.
   (b) For batch renewal processes having especially simple forms (such as the sGGeo example used in Section 4) it may be possible to derive explicit closed form solutions (assuming that the parameters corresponded to a proper batch renewal process) and then apply the explicit form (with the parameters of the improper batch renewal process). However, except for the simplest forms, this approach is unworkable.
   (c) A potentially rewarding approach is to view the observed traffic as having resulted from a stream from which customers have been removed.

That is equivalent to saying that the traffic contains 'negative customers' [6]. Then the observed traffic might be modelled as the merge of traffic from a (proper) batch renewal process plus randomly inserted negative customers. This approach appears feasible because, for any traffic stream, injecting (positive) customers randomly affects the generating functions such that $K(0) + L(0)$ increases faster than $K(1-)$ and $L(0)$ increases faster than $K_+(1-)$.

3. The demonstration in Section 6 between some representations of traffic processes provokes further questions. For example, the constructions used to demonstrate relations 3 and 4 of Figure 8 depend upon the phase space being discrete but none of the relations depend in any way upon the form of distributions of counts at each point of the process nor, indeed, upon the such distibutions being discrete.

   The equivalence argument works just as well when the counts have continuous distribution. So, could the duality between counts and interarrival times be exploited to show equivalence between various forms of (discrete space) continuous time models?
4. It would be very useful to know if the conjecture (that is offered at the end of Section 6) were true. Then, when attempting to match some measured traffic, there would be no inherent bias in seeking an appropriate model only amongst the most convenient representation.

   However, the conjecture might be not testable. A conjecture of equal practical value might be expressed as *there is no test that in finite time discriminates between traffic generated by process X and that from some MMBBP over a countable phase space* for every process X that is not known to be a member of the class.

   Again, that second conjecture might not be testable. But it might be possible to make some progress towards proving or disproving the second conjecture for some characteristics of traffic or for some interesting subclasses of traffic.
5. It is clearly important to have fast algorithms, such as those mentioned in Section 8, that provide reliably accurate approximations for network performance. Our confidence in their results is based upon the empirical evidence that the results have been good for all cases – so far.
   (a) More experiments are needed on networks with correlated inputs.
   (b) The bases of the implicit assumptions and approximations inherent in the algorithms need to be examined – both to discover simple expressions for the accuracy of the results (or bounds on the errors) and also to see whether the accuracy could be improved without degrading the speed of the algorithms.

References

1. Dimakopoulos, G.A.: On the Approximation of Complex Traffic Models on ATM Networks, M.Phil. Dissertation. Postgraduate School of Computing and Mathematics, University of Bradford (2000)

2. Disney, R.L., Kiessler, P.C.: Traffic Processes in Queueing Networks: A Markov Renewal Approach. The John Hopkins University Press, Baltimore (1987)
3. Fretwell, R.J., Dimakopoulos, G.A., Kouvatsos, D.D.: Ignoring Count Correlation in SRD Traffic. In: Bradley, J.T., Davies, N.J. (eds.) Proc. 15th UK Perf. Eng. Workshop, pp. 285–294. UK Performance Engineering Workshop Publishers (1999)
4. Fowler, H.J., Leland, W.E.: Local Area Network Traffic Characteristics, with Implications for Broadband Network Congestion Management. IEEE JSAC 9(7), 1139–1149 (1991)
5. Andrade, J., Martinez-Pascua, M.J.: Use of the IDC to Characterize LAN Traffic. In: Kouvatsos, D. (ed.) Proc. 2nd. Workshop on Performance Modelling and Evaluation of ATM Networks, pp. 15/1–15/12 (1994)
6. Gelenbe, E.: Random Neural Networks with Positive and Negative Signals and Product Form Solution. Neural Computation 1(4), 502–510 (1989)
7. Gordon, J.J.: Pareto Process as a Model of Self-Similar Packet Traffic. In: Proc. Globecom 1995, Singapore, pp. 2232–2236 (1995)
8. Gusella, R.: Characterizing the Variability of Arrival Processes with Indexes of Dispersion. IEEE JSAC 9(2), 203–211 (1991)
9. Heffes, H., Lucantoni, D.M.: A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance. IEEE JSAC 4(6), 856–868 (1986)
10. Kouvatsos, D.D., Tabel-Aouel, N.M., Denazis, S.G.: Approximate Analysis of Discrete-time Networks with or without Blocking. In: Perros, H.G., Viniotis, Y. (eds.) High Speed Networks and their Performance, vol. C-21, pp. 399–424. North-Holland, Amsterdam (1994)
11. Kouvatsos, D.D., Fretwell, R.: Discrete Time Batch Renewal Processes with Application to ATM Switch Performance. In: Hillston, J., et al. (eds.) Proc. 10th. UK Computer and Telecomms. Performance Eng. Workshop, September 1994, pp. 187–192. Edinburgh University Press, Edinburgh (1994)
12. Kouvatsos, D., Fretwell, R.: Closed Form Performance Distributions of a Discrete Time $\mathrm{GI}^{\mathrm{G}}/\mathrm{D}/1/N$ Queue with Correlated Traffic. In: Fdida, S., Onvural, R.O. (eds.) Enabling High Speed Networks, October 1995, pp. 141–163. IFIP Publication, Chapman and Hall (1995)
13. Kouvatsos, D.D., Awan, I.U., Fretwell, R., Dimakopoulos, G.: A Cost-effective Approximation for SRD Traffic in Arbitrary Multi-buffered Networks. Computer Networks 34, 97–113 (2000)
14. Laevens, K.: The Output Process of a Discrete Time $\mathrm{GI}^{\mathrm{G}}/\mathrm{D}/1$ Queue. In: Proc. 6th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, Research Papers, pp. 20/1–20/10 (July 1998)
15. Leland, W.E., Taqqu, M.S., Willinger, W., Wilson, D.V.: On the Self-Similar Nature of Ethernet Traffic (Extended Version). IEEE/ACM Transactions on Networking 2(1), 1–14 (1994)
16. Li, W.: Performance Analysis of Queues with Correlated Traffic, PhD Thesis (University of Bradford) (2007)
17. Lucantoni, D.M.: The BMAP/G/1 Queue: A Tutorial. In: Donatiello, L., Nelson, R. (eds.) SIGMETRICS 1993 and Performance 1993. LNCS, vol. 729, pp. 330–358. Springer, Heidelberg (1993)
18. Neuts, M.F.: A Versatile Markovian Point Process. J. Appl. Prob. 16, 764–779 (1979)

19. Sriram, K., Whitt, W.: Characterizing Superposition Arrival Processes in Packet Multiplexers for Voice and Data. IEEE JSAC 4(6), 833–846 (1986)
20. Wittevrongel, S., Bruneel, H., Vinck, B.: Analysis of the discrete-time $G^{(G)}$/Geom/c queueing model. In: Gregori, E., Conti, M., Campbell, A.T., Omidyar, G., Zukerman, M. (eds.) NETWORKING 2002. LNCS, vol. 2345, pp. 757–768. Springer, Heidelberg (2002)
21. Xiong, Y., Bruneel, H.: Buffer Contents and Delay for Statistical Multiplexers with Fixed Length Packet Train Arrivals. Performance Evaluation 17(1), 31–42 (1993)