Modeling of Correlated Arrival Processes in the Internet

Muckai K Girish
SBC Technology Resources, Inc.
4698 Willow Road
Pleasanton, CA 94588
mgirish@tri.sbc.com

Jian-Qiang Hu *
Dept. of Manufacturing Engineering
Boston University
15 St. Mary’s Street
Boston, MA 02215.
hu@enga.bu.edu

Abstract

The correlations in the arrival processes in the Internet are very significant since voice, video and data are transported simultaneously over high speed links. Detailed knowledge of the behavior of the arrival processes is of paramount importance in traffic engineering and network optimization. In this paper, we propose to use Markov-modulated arrival processes to approximate the correlated arrival processes in the Internet. Two approximate procedures are developed to evaluate the parameters of the underlying Markov chain and the conditional density functions given a few moments and the lag-1 autocorrelation of the arrival process. The resulting queueing system can be analyzed using the existing techniques for the G/G/1 queue with Markov-modulated arrival processes. We also show how our model can be used in connection admission control (CAC).

Keywords: Internet; ATM Networks; Autocorrelations; Markov-Modulated Arrival Processes, CAC.

1 Introduction

The field of communication systems has been witnessing tremendous growth with the development of many complex devices and mechanisms that have helped the explosive growth of the Internet. Integrated communication systems have been designed to facilitate the transport of video and voice along with data and messages. High-speed routers, Frame Relay and Asynchronous Transfer Mode (ATM) are some examples. The arrival distributions in these networks are typically bursty and highly correlated. The analysis of queues with correlations in the input distributions have posed serious difficulties to the researchers and the practitioners alike. In spite of the remarkable progress achieved in this area, these systems are still not understood completely. Ignoring correlations often leads to high levels of inaccuracy in the performance measures (see Girish and Hu 1996a and Girish 1996). A simulation study reported in Livny et al. (1993) considered a single server queue with correlations. Their results point to the fact that the average waiting time deviates considerably from that of estimations by ignoring the correlations.

Traffic management in the Internet backbone nodes is a complex task given the large number of switched and permanent virtual circuits that demand differing Quality of Service requirements from the network. In order to have effective network design, operation and control, one needs to be able to have accurate performance evaluation mechanisms in place. The typical performance measures of interest are the distributions of the congestion measures such as the waiting time, buffer occupancy and buffer overflow probabilities.

One of the few models used to study correlated queues is the G/G/1 queue with Markov-modulated arrivals (MAP). For a review on Markov-modulated queues, see Prabhu and Zhu (1989) and the references therein. The MacLaurin series for the moments of the waiting and system time for this system were derived in Zhu and Li (1993) and for the moments and the lag-1 autocorrelation of the interdeparture time were derived in Girish and Hu (1996b,c). It was shown that Padé approximation is very effective in estimating these performance measures from their MacLaurin series. Refer to Baker (1975) and Petrushev and Popov (1987) for a comprehensive treatment of the theory of Padé approximation. Most of these models assume that the interarrival and service time distributions are known and aid in performance evaluation of the resulting queue. But, in practice, these input distributions need to be approximated from some measurable quantities. Moreover, very little work has been done in the area of modeling correlated arrival processes. It is the intent of this paper to develop simple schemes to approximately characterize the arrival processes given some specific information about them.

The dependence properties of the arrival processes in packet-switched networks have been stud-
parameterized nodes are then analyzed based on these two parameters and the first two moments of the service time distribution. Though QNA works very well under a variety of network topology and traffic loads, it gives rise to large errors when the correlations in the interarrival time random variables are significant since it does not take into account the autocorrelations. This can also be observed in other approximation techniques that consider higher orders of the moments of the arrival and service time distributions as well (see Girish and Hu 1996a and Girish 1996). Therefore the lack of results for taking correlations in the analysis has necessitated the need for modified parametric decomposition methods. The results of this paper, thus, can be used to consider the autocorrelations along with the moments in characterizing the interactions between nodes in general queueing networks.

Many Internet backbone networks consist of overlay networks that include ATM networks as well. ATM networks have been designed to guarantee Quality of Service (QoS) for the connections that are admitted. Different types of traffic require different service qualities. Voice and real-time video connections require very low delay, delay variations and cell loss; whereas regular data connections can tolerate higher delays. Presently, most ATM networks support the following QoS classes: CBR (constant bit rate), VBR (real time and non-real time variable bit rates), ABR (available bit rate) and UBR (unspecified bit rate). A Connection Admission Control (CAC) algorithm determines if a call is admissible by computing the equivalent bandwidth required by the connection and comparing it to the available bandwidth in the network. We show how the results on modeling of the arrival processes developed in this paper can be applied to a popular connection admission control algorithm. One of the pioneering works in connection admission control can be found in Guerin et al. (1991). They considered a two-state fluid-flow model to approximate the arrival processes. This process can be completely described by three parameters: the peak cell rate and the means of the burst and idle periods (which are assumed to be exponentially distributed).

The remainder of this paper is organized as follows. The general problem of modeling a Markov-modulated arrival process given the moments and lag-1 autocorrelation is presented in Section 2. A special case with modeling Poisson conditional arrival distribution is discussed in Section 3. The case with general conditional arrival processes is studied in Section 4. The application of this approach in Connection Admission Control is outlined in Section 5 and this paper is concluded in Section 6.
2 The Problem

Let $A$ denote the generic random variable representing the correlated arrival process in the Internet and let $A_k; k \geq 1$ be the interarrival time between arrivals $k - 1$ and $k$, where $A_0 = 0$. In this section, we outline the problem where the first $n$ non-central moments of the arrival process $(M_1, \ldots, M_n)$ are known as well as its lag-1 autocorrelation $(\chi)$. Recall that

$$M_n = E[A^n] \forall n \geq 1,$$

$$\chi = E[A_1 A_2].$$

The objective is to identify a Markov chain, its transition and stationary probabilities and the moments of the conditional distributions. Let $\{J_n; n \geq 0\}$ be the underlying Markov chain, $E$ be the state space of the Markov chain, $\mathcal{P} = (p_{ij}), i, j \in E$ be the transition probability matrix and let $\pi_i, i \in E$ be the stationary probability of being in state $i$. Further, we denote the conditional non-central moments as:

$$M_{ij} = E[A_i^1|J_0 = i, J_{n+1} = j]$$

The problem is to determine $E$, $\pi_i$, $M_{ij}$, $M_{ij}$, ..., $M_{ij}$ from the following set of equations:

$$\sum_i \sum_j \sum_{k} \pi_{ip_{ij}} p_{jk} M_{ij1} M_{jk1} = \chi$$

$$\sum_i \sum_j \pi_{ip_{ij}} = M_{11}$$

$$\sum_i \sum_j \pi_{ip_{ij}} M_{ij2} = M_{22}$$

$$\cdots$$

$$\sum_i \sum_j \pi_{ip_{ij}} M_{ijn} = M_{nn}$$

Since this general problem is not easily solvable, in the following two sections, we consider two special cases.

3 Matching Two Moments and the Lag-1 Autocorrelation

We look at the case when the first two moments and the lag-1 autocorrelations of a distribution are known and our objective is to find a Markov-modulated Poisson process (MMPP) that fits these parameters. Note that a MMPP has Poisson conditional arrival processes. Recall that the known moments are denoted by $M_1$, $M_2$ and the lag-1 autocorrelation by $\chi$. We now consider a Markov chain which is irreducible, aperiodic and recurrent in a state space $E$. We assume that this Markov chain has only two states ($E = \{1, 2\}$) and the doubly-stochastic transition probability matrix is given by:

$$\mathcal{P} = \begin{pmatrix} 1 - p_1 & p_1 \\ p_2 & 1 - p_2 \end{pmatrix}$$

Then the steady state probabilities can be immediately determined to be: $\pi_1 = p_2/(p_1 + p_2)$, $\pi_2 = p_1/(p_1 + p_2)$ by solving the equations, $\pi \mathcal{P} = \pi$ and $\sum_{i \in E} \pi_i = 1$. We now add restrictions on the conditional probability distributions of the interarrival time given $J_n = i$, for $i = 1, 2$ are exponential with mean $x_i$ i.e.,

$$F_i(t) = P\{A_n \leq t|J_n = i\} = 1 - \exp(-\frac{t}{x_i}), \quad (1)$$

where $F_i(\cdot)$ represents the cumulative distribution function of the conditional arrival distribution. What remain to be determined are $p_1, p_2, x_1, x_2$ from the following set of equations which can be readily established by applying the simplified assumptions in this section to the general problem formulation in Section 2:

$$p_2(1 - p_1)x_1^2 + 2p_1p_2x_1x_2 + p_1(1 - p_2)x_2^2 = 2\chi$$

$$p_2x_1 + p_1x_2 = (p_1 + p_2)M_1 \quad (2)$$

$$p_2x_1^2 + p_1x_2^2 = \frac{1}{2}(p_1 + p_2)M_2$$

Let us denote the squared coefficient of variation as, $c^2 = (M_2/M_1^2) - 1$. Solving (2), we get

$$p_1 + p_2 = \frac{M_2 - 2\chi}{M_2 - 2M_1^2} \quad (3)$$

$$x_1 = M_1(1 \pm \sqrt{\frac{p_1(c^2 - 1)}{2p_2}}) \quad (4)$$

$$x_2 = M_1(1 \pm \sqrt{\frac{p_2(c^2 - 1)}{2p_1}}) \quad (5)$$

We note that we have one degree of freedom in determining parameters $p_1$ and $p_2$.

4 Matching Three Moments and the Lag-1 Autocorrelation

Given the first three moments and the lag-1 autocorrelation of the interarrival time random variable $A$, we address the problem of finding a Markov-modulated arrival process (MAP) that fits these parameters exactly in this section. The dependencies will be captured by the Markov chain. Recall that we denoted the moments of $A$ by $M_1, M_2, M_3$ and the lag-1 autocorrelation by $\chi$. We now look at a Markov chain which is irreducible, aperiodic and recurrent in a state space $E$. We again assume that this Markov chain has only two states ($E = \{1, 2\}$) but with probability transition matrix

$$\mathcal{P} = \begin{pmatrix} p & 1 - p \\ 1 - p & p \end{pmatrix}$$
For this Markov chain, its steady state probabilities are given by \( \pi_1 = \pi_2 = 0.5 \). The lag-1 correlation coefficient, \( \delta \) is defined as:

\[
\delta = \frac{X - M_1^2}{M_2 - M_1^2} \tag{6}
\]

We now add restrictions on the conditional probability distributions of the above Markov chain. The conditional interarrival distributions given \( J_n = i \), for \( i = 1, 2 \) have \( x_i, y_i, z_i \) as their first three non-central moments, respectively, i.e., for \( i = 1, 2 \),

\[
\begin{align*}
E[A_n | J_n = i] &= x_i \\
E[A_n^2 | J_n = i] &= y_i \\
E[A_n^3 | J_n = i] &= z_i
\end{align*}
\]

What remain to be determined are \( p, x_1, x_2, y_1, y_2, z_1, z_2 \) from the following set of equations:

\[
p x_1^2 + 2(1-p) x_1 x_2 + p x_2^2 = 2X
\]

\[
x_1 + x_2 = 2M_1 \\
y_1 + y_2 = 2M_2 \\
z_1 + z_2 = 2M_3
\]

The above equations can now be solved. Solving the first two equations for \( x_1 \) and \( x_2 \) we get:

\[
x_1 = M_1 + \sqrt{\frac{X - M_1^2}{2p - 1}} \tag{7}
\]

\[
x_2 = M_1 - \sqrt{\frac{X - M_1^2}{2p - 1}} \tag{8}
\]

If the correlation is positive, then

\[
\frac{X}{2M_1^2} \leq p \leq 1
\]

Now back to solving for the higher conditional moments, we have the following fundamental relationships that the moments of any probability distribution have to satisfy:

\[
\begin{align*}
M_1 &\geq 0 \\
M_2 - M_1^2 &\geq 0 \\
M_1 M_3 - M_2^2 &\geq 0
\end{align*}
\]

Since the conditional moments have to satisfy the above too, we get

\[
\begin{align*}
x_1^2 &\leq y_1 \leq 2M_2 - x_2^2 \\
y_1^2 &\leq z_1 \leq 2M_3 - \frac{y_2^2}{x_2}
\end{align*}
\]

Therefore, we define two more parameters denoted by \( \beta \) and \( \gamma \) that satisfy the constraints: \( 0 \leq \beta \leq 1 \) and \( 0 \leq \gamma \leq 1 \), such that we can express the higher moments as:

\[
\begin{align*}
y_1 &= \beta x_1^2 + (1 - \beta)(2M_2 - x_2^2) \tag{9} \\
y_2 &= 2M_2 - y_1 \tag{10} \\
z_1 &= \gamma \frac{y_1^2}{x_1} + (1 - \gamma)(2M_3 - \frac{y_2^2}{x_2}) \tag{11} \\
z_2 &= 2M_3 - z_1 \tag{12}
\end{align*}
\]

Hence \( \beta \) and \( \gamma \) can be chosen such that the second and third conditional moments satisfy the above constraints. After determining these, the first three conditional moments can now be matched to a probability distribution such as a mixture of Erlang distributions (see Johnson and Taaffe 1989). It is now easy to evaluate the performance of the resulting queueing system using any of the known approximation techniques (see Prabhu and Zhu 1989, Zhu and Li 1993, Girish and Hu 1996, and the references therein). We summarize the procedure for obtaining the parameters of the Markov-modulated arrival process below:

**Algorithm for Modeling MAP:**

1. Choose a value for \( p \).
2. Determine \( x_1 \) and \( x_2 \) from Equations (7) and (8).
3. Choose \( 0 \leq \beta \leq 1 \) and evaluate \( y_1 \) and \( y_2 \) from Equations (9) and (10).
4. Choose \( 0 \leq \gamma \leq 1 \) and evaluate \( z_1 \) and \( z_2 \) from Equations (11) and (12).
5. Match the conditional moments to a mixed Erlang distribution using the results of Johnson and Taaffe (1989).
6. Use this Markov-modulated process as the inter-arrival distribution and analyze the queue using any of the existing techniques.

One of the main issues in the selection of the parameters, \( \alpha, \beta, \gamma \) is the order of the mixed Erlang distributions \( n \) that are obtained by matching the conditional moments. The procedure of Johnson and Taaffe (1989) for matching the first three moments to a mixed Erlang distribution of common order involves the following two inequalities that \( n \) has to satisfy:

\[
n \geq M_1^2 \tag{13}
\]

\[
n \geq \frac{2(M_2 - M_1^2)^2 + M_1^4 M_2 - M_1 M_3}{M_1 M_3 - (M_2 - M_1^2)(M_2 - 2M_1^2)} \tag{14}
\]
In order to avoid numerical errors in the analysis, we would like to keep \( n \) as small as possible. Our numerical experience provides a few guidelines as to the selection of the parameters. In general, \( \alpha \) should be as close to zero as possible. We also found that \( \gamma = 5 \) and \( 3 \leq \beta \leq 7 \), in general, give good results. More research is needed to optimize the parameter values for any system. One possible way is to express \( n \) given in (13) and (14) in terms of \( \alpha, \beta, \gamma \) and use classical optimization techniques to minimize \( n \), thereby, obtaining the optimal parameter values.

5 The Connection Admission Control Algorithm

An important feature of ATM networks is the concept of guarantees for the Quality of Service (QoS). A call requesting connection set up requests a certain QoS which, if approved, can be guaranteed for the lifetime of the connection. Currently there are four types of QoS classes supported by most ATM networks: CBR (Constant Bit Rate), VBR (Real time and Non-Real time Variable Bit Rates), ABR (Available Bit Rate) and UBR (Unspecified Bit Rate). From the perspective of an ATM network, a connection is determined to be admissible by implementing a CAC algorithm which determines the equivalent capacity or bandwidth required for the connection. This equivalent bandwidth is compared with the available bandwidth and if sufficient capacity remain, then the connection is admitted, otherwise it is rejected. Due to the nature of statistical multiplexing of ATM networks and the varying requirements of the different QoS classes, determination of the equivalent capacity or bandwidth is a non-trivial problem and has been studied by several researchers and practitioners.

A pioneering work in the determination of bandwidth allocation in packet switching networks can be found in Guerin et al. (1991). They develop an approximate technique to estimate the equivalent capacity for high-speed packet switching networks. In this section, we study the relationship of their approach to the modeling of correlated arrival processes that was developed in this paper. The model considered in Guerin et al. (1991) is a two-state fluid-flow model in which the arrivals occur at a peak rate of \( R \) per second or zero. The busy and idle periods are distributed exponentially with means \( B \) and \( I \), respectively. The three parameters, \((R, B, I)\) completely specify the arrival process, where \( \rho \) is the utilization. During the busy period, the interarrival time is \( 1/R \) and during the idle period, the interarrival time is \( I \). Also, the average number of arrivals during one cycle of busy and idle periods is \( BR \). Incorporating this, we get:

\[
M_1 = \frac{I + (RB/R)}{RB},
\]

\[
M_2 = \frac{1 + (RB/R)}{RB},
\]

\[
\chi = \frac{(2I/R) + (2RB/R^2)}{RB}.
\]

Noting that \( \rho = B/(B + I) \), the above expressions can be simplified as:

\[
M_1 = \frac{B + I}{RB} = \frac{1}{\rho R},
\]

\[
M_2 = \frac{\rho^2 + 1}{R E} = \frac{(1 - \rho)^2 B + 1}{\rho^2 R} + \frac{1}{RE},
\]

\[
\chi = \frac{2(B + I)}{RE B} = \frac{2}{\rho RE}.
\]

Another way of looking at this problem is to transform the Markov-modulated arrival process characteristics to that of the fluid-flow model. Given a MAP, the first two moments and the lag-1 autocorrelation can be determined easily and the three parameters of the fluid model can be evaluated from the following:

\[
\rho = \frac{\chi}{2M_1^2},
\]

\[
R = \frac{2M_1}{\chi},
\]

\[
B = \frac{\chi(4M_1^2M_2 - \chi^2)}{2M_1(2M_1^2 - \chi^2)}.
\]

The set of Equations (15) forms a useful way to use MAP modeled from a variety of situations in applying the CAC algorithm for ATM networks.

6 Conclusions

The arrival processes in the Internet backbone nodes are typically highly correlated since they are constituted by the superposition of voice, video and data traffic streams. In order to facilitate the modeling of such systems given a few parameters of the arrival process, we considered two models that approximate the arrival processes by Markov-modulated processes (MAP). Both these models consider a two-state markov chain and one case with Poisson conditional arrival processes and the other case with general conditional arrival distributions. We derived methods to calculate the parameters of these distributions. This modeling technique has significance in the analysis of many communication systems which encounter bursty traffic. These models can also be used to approximate the interarrival time distributions at the downstream node in tandem queueing networks in which the interdeparture time from the upstream nodes have significant correlations. We also showed how our results can be applied in connection admission control.
An interesting direction for future research is to consider the long range dependence in our model. This can then ensure the capture of the long range dependence characteristics in the approximating arrival processes. It would also be useful to generalize the results of this paper by matching more moments and increasing the state space of the underlying Markov chain.

References


