Markov Modulated Periodic Arrival Process Offered to an ATM Multiplexer

Nail Akar and Erdal Arıkan
Electrical and Electronics Eng. Dept.
Bilkent University, 06533 Ankara, Turkey

ABSTRACT When a superposition of on/off sources is offered to a deterministic server, a particular queueing system arises whose analysis has a significant role in ATM based networks. Periodic cell generation during active times is a major feature of these sources. In this paper a new analytical method is provided to solve for this queueing system via an approximation to the transient behavior of the $nD/D/1$ queue. The solution to the queue length distribution is given in terms of a solution to a linear differential equation with variable coefficients. The technique proposed here has close similarities with the fluid flow approximations and is amenable to extension for more complicated queueing systems with such correlated arrival processes. A numerical example for a packetized voice multiplexer is finally given to demonstrate our results.

Introduction

The Asynchronous Transfer Mode (ATM) is the preferred transfer mode for the Broadband ISDN (B-ISDN). The core of an ATM network is "asynchronous multiplexing" on the basis of which transmission links and switching devices are shared by different virtual connections. Information is transmitted in the form of constant length packets, called "cells". Since ATM has the potential to improve bandwidth efficiency via the use of statistical multiplexing of variable bit-rate sources, characterization of a traffic stream belonging to a particular connection turns out to have an important role. In fixed bit rate coding schemes, sources emit cells periodically with a frequency determined by their bit rate. On/off sources emit cells periodically during activity (on) times alternating with silence (off) times during which there is no cell generation. These two periods are in general of variable length. In this paper, we focus on a queueing system in which several on/off sources with an identical period share a buffer of infinite size. Given the number of sources and the associated traffic parameters, we are interested in the probability distribution function of the buffer content. A 2-state continuous-time Markov chain model (see Figure 1) will be used to describe the aforementioned traffic stream. In this model, the silence times and the activity times are exponentially distributed with means $1/\lambda$ and $1/\mu$, respectively. This 2-state model can be extended to construct an N-state Markov chain to describe the superposition process of N on/off sources (see Figure 2). The state of the Markov chain is defined to be the number of active sources. In an arbitrary state, say n, of the Markov chain whose state holding time is exponentially distributed with parameter $\sigma_n = (N-n)\lambda+n\mu$, n sources independently transmit cells with an identical period. In general, we call the arrival process as a Markov modulated periodic arrival process.

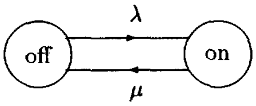

Figure 1: 2-state Markov model for an on/off source

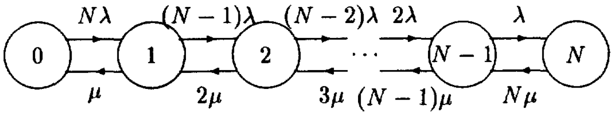

Figure 2: Birth-death model for the superposition of N on/off sources

The arrival process associated with the superposition possesses two kinds of correlations:

- negative correlations of arrivals in successive time slots due to the periodic nature of cell transmissions,
- positive correlations among the average arrival rates in successive periods of length greater than the intercell times of the multiplexed sources.

There are various approaches proposed in the literature which take account of these correlation effects in the performance analysis of the queueing system [1]. One basic approach is using fluid flow models. These models approximate the cell arrival and service process by continuous arrival and departure of a fluid. The superposition of a finite number of on/off fluid sources is considered in [2] for which the authors give a computationally efficient algorithm to evaluate the buffer occupancy distribution. However, the model does not give accurate results for low to moderate traffic when cell layer contention dominates over burst layer contention. This is because the first type of correlations cannot be captured by fluid flow models. The model and technique proposed in [2] is further applied to the finite buffer case in [3] to solve for the cell loss rate which is a critical performance measure in ATM networks.

For an accurate analysis of an ATM multiplexer, the negative correlation between cell interarrival times should also be taken into consideration. Actually, when the instantaneous arrival rate is less than the link rate, the queueing system

0-7803-0917-0/93$03.00 © 1993 IEEE

behaves like the so-called $nD/D/1$ queue: a superposition of independent periodic sources (n sources) with an identical period but with random phase feeds a constant service time buffer. This queue is investigated in [4], [5] to find the steady-state distribution of the queue length. Different periods are also allowed in [6] where accurate approximate formulas for the queue length distribution are derived.

In our queueing model that considers a Markov modulated periodic arrival process as input to the multiplexer, the transient behavior of the $nD/D/1$ queue has a significant role. The focus of this paper is on the derivation of a relationship between fluid sources and periodic sources, arrival rates of which are Markov modulated in the same manner, through an approximation of the transient behavior of the $nD/D/1$ queue. This approximation is mainly based on an interpolation of the queue length whose distribution is known at certain epochs. The solution for the overall problem is then reduced to the solution of a linear differential equation with variable coefficients whereas in fluid flow approximations the corresponding equation is simply linear with constant coefficients. A numerical example is finally given in the context of a packetized voice multiplexer.

Problem Formulation and Analysis

The method used in solving for the steady-state distribution of the queue length for the Markov modulated periodic arrival case is composed of two main stages. The first stage consists of an approximation to the transient behavior of the $nD/D/1$ queue in a continuous-time framework. In the second stage, we extend our results for the $nD/D/1$ queue to solve for the continuous-time Markov model which characterizes the input traffic.

Let us first consider the case when the number of active sources (n) is fixed. In our queueing model, we assume that the time axis is slotted where each time slot is as long as the transmission time of a cell. The cells arriving to the queue are served on a first-come-first-serve basis and the queue has infinite size. The n active sources each transmit fixed length cells with a period of R slots, independently of each other. In an arbitrary frame of R slots, each input source's cell can be in any of these R slots with equal probability. The peak source rate in cells/sec is denoted by P and the service rate of the buffer is denoted by C, which actually equals to PR cells/sec. Without loss of generality, we assume that the departures take place at the beginning of slots, and arrivals during slots. Let us assume a stable queue ($n < R$) for the time being and define the following random variables

$$Q_k \;=\; \text{queue length at the end of } k^{th} \text{ slot},$$
$$a_k \;=\; \text{number of arrivals in the } k^{th} \text{ slot}.$$

The queueing strategy is the following:

$$Q_k = \begin{cases} Q_0 & \text{if } k = 0 \\ \max(Q_{k-1} - 1, 0) + a_k & \text{if } k > 0 \end{cases}$$

By iteration on k, one can check using algebraic manipulations that

$$Q_R = \max(\tilde{Q}_n, Q_0 + n - R) \tag{1}$$

where the random variable $\tilde{Q}_n$ is defined via

$$\tilde{Q}_n = \max_{0 \le j < R}\left(\sum_{l=R-j}^{R} a_l - j\right). \tag{2}$$

The cumulative distribution function for the random variable $\tilde{Q}_n$ is expressed by the following summation [5]

$$\tilde{Q}_n(q) \overset{\Delta}{=} Pr\{\tilde{Q}_n \le q\}$$

$$= 1 - \sum_{x=1}^{n-\bar{q}} \frac{R-n+\bar{q}}{R-x} C(n, \bar{q}+x)\left(\frac{x}{R}\right)^{\bar{q}+x}\left(1-\frac{x}{R}\right)^{n-\bar{q}-x}, \tag{3}$$

where $\bar{q}$ is the largest integer smaller than q and $q \le n-1$. In order to obtain the queue length evolution equations for $n < R$, we iterate on equation (1) on an R-slot basis so that by periodicity of arrivals we have

$$Q_{kR} = \max(\tilde{Q}_n, Q_0 + k(n-R)), \quad k = 1, 2, \ldots \tag{4}$$

There is, in fact, a strong interconnection between periodic models and fluid flow models. In the latter models, information is assumed to arrive uniformly to the multiplexer and the server similarly removes information from the queue, in a continuous manner. The computational tractability and buffer size independent solvability of fluid flow approximation techniques suggest a further study of this interconnection.

If we define $Q(t)$ as the queue length at time t, the fluid flow approximations suggest that [2]:

$$Q(t) = \max(0, Q_0 + (Pn - C)t). \tag{5}$$

Note the noninteger values that $Q(t)$ may take due to the absence of the concept of packetization in fluid models.

There are two major differences between the expressions (4) and (5). The first term associated with the short term fluctuations of the queue length is the random variable $\tilde{Q}_n$ in the periodic model whereas it equals zero in the fluid model. This is why the fluid flow models do not give accurate results in light to moderate traffic when several on/off sources are multiplexed on a common link. This deficiency belonging to fluid models has been mentioned by several authors [7], [8]. The second term associated with the dynamical behavior of the queue length in (5) is just a linear interpolation of the corresponding term in (4).

For the overload states, since the probability that the queue length is zero at some time epoch is negligible, fluid flow approximation gives accurate results in the analysis of the transient response of the queue. Taking (4) as our key equality, our approach is mainly based on interpolating the second term as in (5) while preserving the first term, $\tilde{Q}_n$, which captures the short term fluctuations in the cell layer. In regard of these observations, we approximate $Q(t)$ by

$$Q(t) = \begin{cases} \max(\tilde{Q}_n, Q_0 + (Pn - C)t), & n < R \\ Q_0 + (Pn - C)t, & n \ge R \end{cases} \tag{6}$$

The accuracy of this approximation for the average queue length in an $nD/D/1$ queue is examined in Figure 3 and compared with simulation results and fluid flow approximations.

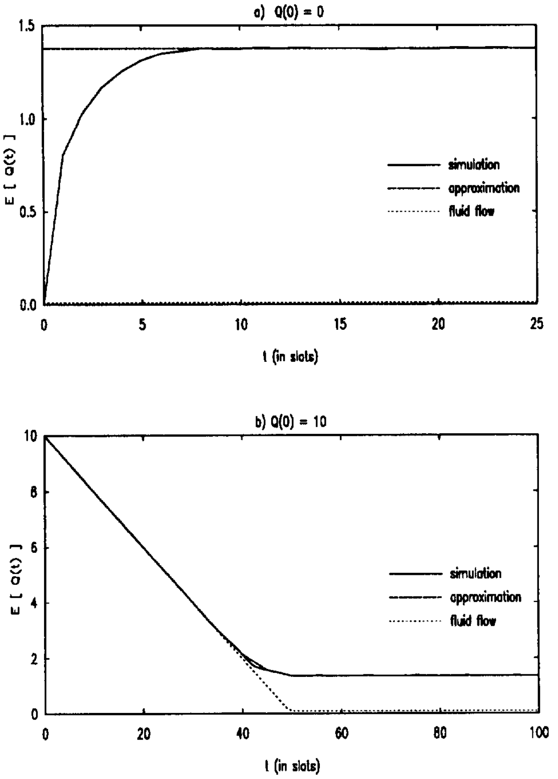

Figure 3: Comparison of approximations for the expected value of the queue length for the case $R = 10$ and $n = 8$.

For the case $R = 10$ and $n = 8$, we allow the queue start with two different initial levels.

The major observation is that, the approximation (6) is accurate when the initial buffer content is not in the vicinity of zero. Even in this case, the approximation is able to track the simulation curve after the queue length reaches its steady-state distribution. Better results compared with fluid flow approximations are obtained irrespective of the initial buffer content. The fundamental approximation in (6) will now be used to derive formulas for the queue length distribution when the number of active sources is changed according to the state of the birth-death model.

Let us now consider the traffic model in Figure 2 and concentrate on a particular state (n, $0 \le n \le N$) of the Markov chain. Let $X(t)$ be the buffer content and $S(t)$ be the state of the Markov chain at time t. Also let π_n be the stationary probability of n sources being active. We then define the following stationary probabilities (as $t \to \infty, \Delta t \to 0^+$):

$$\bar{F}_b(n,x) = Pr\{X(t) \le x | S(t+\Delta t) = n, S(t) \ne S(t+\Delta t)\}$$

and

$$\bar{F}_e(n,x) = Pr\{X(t) \le x | S(t+\Delta t) \ne S(t), S(t) = n\}.$$

Note that $S(t)$ is the state of a continuous-time Markov chain; given $S(t)$, the buffer content $X(t)$ will be independent of $S(t+\Delta t)$. This fact yields

$$\bar{F}_e(n,x) = Pr\{X(t) \le x | S(t) = n\}.$$

To interpret, $\bar{F}_b(n,x)$ is the equilibrium probability that the queue length is less than x given that a state transition to state n is about to occur. Similarly, $\bar{F}_e(n,x)$ is the stationary probability that the queue length is less than x given that a state transition from state n is about to occur. In other words, we observe the queue length at the time epochs when state transitions occur and henceforth define the corresponding random variables. We then define $F_b(n,x) \triangleq \pi_n \bar{F}_b(n,x)$ and $F_e(n,x) \triangleq \pi_n \bar{F}_e(n,x)$.

Each time the Markov system changes a state we assume a complete phase randomization of all the sources whereas for the original system an active source's phase is independent of the other sources' state transitions. With this assumption, the stationary queue length at the moment of state transition to n and $\tilde{Q}_n$ become independent. By exploiting the approximation (6) and with the above assumption one obtains

$$F_e(n,x) = \begin{cases} \tilde{Q}_n(x) F_f(n,x), & n < R \\ F_f(n,x), & n \ge R \end{cases} \tag{7}$$

where

$$F_f(n,x) \triangleq \left(\int_0^\infty F_b(n, x + (C - Pn)t) \sigma_n \exp(-\sigma_n t) dt \right) u(x),$$

the subscript f denotes the fluid flow term and $u(\cdot)$ is the unit step function. We can then write down the differential equation that governs $F_f(n,x)$ for $x \ge 0$:

$$\frac{d}{dx} F_f(n,x) = \frac{\sigma_n}{C - Pn} F_f(n,x) + \frac{\sigma_n}{Pn - C} F_b(n,x), \; n \ne R. \tag{8}$$

Now letting $p(m,n)$ be the state transition rate from state m to state n, we relate $F_b(n,x)$'s to $F_e(\cdot,x)$'s. It is not difficult to show by using the balance equations of the Markov chain that

$$\sigma_n F_b(n,x) = \sum_{m \ne n} p(m,n) F_e(m,x). \tag{9}$$

Combining (7), (8), and (9), we finally obtain the following differential equations for $F_f(n,x)$'s:

$$\frac{d}{dx} F_f(n,x) = \frac{\sigma_n}{C - Pn} F_f(n,x)$$
$$+ \frac{1}{Pn - C} \sum_{m \ne n} p(m,n) \tilde{Q}_m(x) F_f(m,x), \; n \ne R \tag{10}$$

and

$$F_f(R,x) = \frac{1}{\sigma_R} \sum_{m \ne R} p(m,R) \tilde{Q}_m(x) F_f(m,x).$$

In the above equations, $\tilde{Q}_m(x) \triangleq 1$, $\forall x \ge 0, m \ge R$. If the term $\tilde{Q}_n(x)$ is taken as unity $\forall n$, $n = 0, 1, \ldots, N$, then the above equations are equivalent to the fluid flow equations [2].

Eliminating the algebraic equation pertaining to $F_f(R,x)$ and defining

$$F_f(x) = \begin{bmatrix} F_f(0,x) \\ F_f(1,x) \\ \vdots \\ F_f(R-1,x) \\ F_f(R+1,x) \\ \vdots \\ F_f(N,x) \end{bmatrix},$$

we finally have

$$\frac{d}{dx}F_f(x) = A(x)F_f(x), \; x \geq 0, \tag{11}$$

where the $N \times N$ matrix $A(x)$ is determined through a suitable arrangement of the differential equations in (10). Actually,

$$A(x) = A_i, \; x \in [i, i+1), \; i \in \mathcal{Z}_+, \; 0 \leq i \leq R-2,$$

and

$$A(x) = A, \; x \in [R-1, \infty)$$

for some appropriate constant matrices A_i's and A, due to the piecewise constant structure of the distributions $\tilde{Q}_n(\cdot)$'s. Given the initial condition $F_f(0)$, the differential equation (11) has a unique continuous solution described by

$$F_f(x) = \exp(A_i(x-i))F_f(i), \; x \in [i, i+1], \; 0 \leq i \leq R-2, \tag{12}$$

and

$$F_f(x) = \exp(A(x-(R-1)))F_f(R-1), \; x \geq R-1. \tag{13}$$

In order to find the initial condition, we make use of the following observations:

1 For $n > R$, the queue is always increasing, so the queue length cannot be zero. Therefore, $F_f(n,0) = 0$ for $n > R$.

2 The matrix A is, in fact, equivalent to the state matrix in fluid flow models, therefore it is known to have $R-1$ positive real eigenvalues, $N-R$ negative real eigenvalues and an eigenvalue at the origin. In order for the solution not to blow up as $x \rightarrow \infty$, no positive (unstable) modes of A should be excited by the choice of $F_f(0)$.

3 The behavior of $F_f(n,x)$ as $x \rightarrow \infty$ is easy to write:

$$F_f(n, \infty) = \pi_n, \; \forall n, \; 0 \leq n \leq N.$$

Now, let z_i be a stable eigenvalue of A and ϕ_i be its corresponding right eigenvector. Then, by observation 2 and (13), the solution to $F_f(x)$ can be written in the form

$$F_f(x) = F_f(\infty) + \sum_{i=1}^{N-R} \exp(z_i(x-R+1))\mu_i\phi_i, \; x \geq R-1$$

which yields

$$F_f(R-1) = F_f(\infty) + \sum_{i=1}^{N-R} \mu_i\phi_i, \tag{14}$$

where μ_i's are coefficients to be determined. The relationship between $F_f(0)$ and $F_f(R-1)$ now needs to be established. Using (12), one can write

$$F_f(R-1) = ZF_f(0) \triangleq \left(\prod_{i=0}^{R-2} \exp(A_i)\right) F_f(0). \tag{15}$$

Besides, by observation 1, $F_f(0)$ is in the form

$$F_f(0) = \begin{bmatrix} f \\ 0 \end{bmatrix},$$

where f is of size $R \times 1$. Combining (14) and (15), one can solve for μ_i's and f, and thus the initial condition $F_f(0)$ through a linear matrix equation of size N. Having found the initial condition, the solutions given in (12) and (13) complete our description of the queue length distribution through (7). The essential difference between the method presented here and computations encountered in solving the fluid flow models is the calculation of the linear operator Z in (15).

The overall cdf of queue length is the sum of the individual elements $F_e(n,x)$:

$$Pr(\text{queue length} \leq x) = \sum_{n=0}^{N} F_e(n,x).$$

Numerical Example

We consider a packetized voice system with voice peak rate 32 Kbits/sec., $R = 10$, mean active period 353 ms. and mean silent period 650 ms. The packets are 64 bytes and the packet transmission time is 1.6 ms. Within an active period, cells from an individual voice source are transmitted in a periodic manner, each source's phase being uniform between 0 and 9. In Table 1, the mean waiting time in the queue with respect to the number of voice sources by our analysis method and the fluid flow approximations is given and these values are compared with the simulation results. The analysis method proposed in this paper gives highly accurate results independent of the degree of utilization in the system whereas fluid flow approximation is only satisfactory in the heavy load regime. Figure 4 is devoted to the queue length survivor function which is obtained for the cases $N = 15$ and 20, respectively. In both cases, the method we propose is able to capture the simulation curve for the buffer survivor function accurately.

Conclusions

In the present paper, a new theory for the approximation of the queue length distribution for the Markov modulated periodic arrival process is presented. This method is a natural extension and generalization of fluid flow models which are commonly used in the communications literature. From a multi-layer concept, the technique is capable of capturing the short term fluctuations of the queue length at the cell layer. Therefore, accurate results are obtained in the analysis of a packetized voice multiplexer for different possible loads.

No.	simulation	% 95 conf.	approximations [ms]	
sources	res. (ms.)	interval	analysis	fluid flow
4	0.0929	±0.0021	0.0948	0.00
6	0.1638	±0.003	0.1591	0.00
8	0.2474	±0.003	0.2383	0.00
12	0.4716	±0.0035	0.4813	0.0023
14	0.6474	±0.0065	0.6918	0.0383
16	1.044	±0.03	1.136	0.269
18	2.205	±0.04	2.311	1.199
20	5.32	±0.26	5.46	4.09
22	13.64	±0.38	13.61	12.02
24	35.53	±0.96	35.16	33.40
25	61.6	±2.0	58.8	57.0
26	111.0	±3.5	105.6	103.8
27	258.1	±8.7	224.6	222.9

Table 1: Comparison of approximations of the mean waiting time with the simulation results for the case R = 10.

Except for the determination of the linear operator Z defined in (15), numerical procedures are the same as the ones used in solving the fluid models. One may propose many approximative schemes for determining Z (e.g., trapezoidal approximation) so that a computationally tractable algorithm is provided. Use of the same underlying mathematical framework provides an easy generalization of this idea for more complicated queueing problems for which fluid flow techniques are successfully applied. We believe that the method demonstrated here can be used to develop techniques for the performance evaluation of typical traffic control schemes proposed for ATM networks.

The methodology developed here is valid for discrete-time queueing schemes where the modulating process is a continuous-time Markov chain. This choice is due to the discrete-time operation of ATM multiplexers and the continuous-time nature of the fluid flow approximations on the basis of which we make the performance comparisons. The framework presented here can readily be reformulated to cover other models (e.g., both the multiplexer and the chain work in continuous-time (or in discrete-time)). These extensions and the computational aspects of the method need to be investigated. One other future work is to develop performance analysis schemes in the case of multi-class traffic which, in this framework, needs an accurate approximation to the transient response of the $\sum D_i/D/1$ queue where multiple periods are allowed.

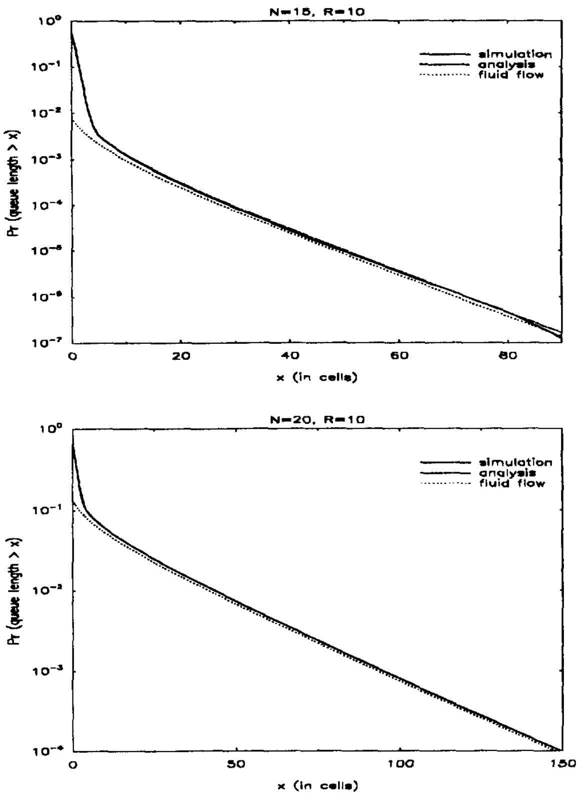

Figure 4: Comparison of the queue length survivor function for our proposed method with simulation results and the fluid flow approximations.

References

[1] I. W. Habib and T. N. Saadawi "Multimedia traffic characteristics in broadband networks," *IEEE Commun. Magazine*, vol. 30, pp. 48–54, 1992.

[2] D. Anick, D. Mitra, and M. M. Sondhi "Stochastic theory of a data handling system with multiple sources," *Bell Syst. Tech. Jour.*, vol. 61, pp. 1871–1894, 1982.

[3] R. C. F. Tucker "Accurate method for analysis of a packet-speech multiplexer with limited delay," *IEEE Trans. Commun.*, vol. 36, pp. 479–483, 1988.

[4] A. E. Eckberg "The single server queue with periodic arrival process and deterministic service time," *IEEE Trans. Commun.*, vol. 27, pp. 556–562, 1979.

[5] A. Bhargava, P. Humblet, and M. G. Hluchyj "Queueing analysis of continuous bit-stream transport in packet networks," in *GLOBECOM*, 1989.

[6] J. W. Roberts and J. T. Virtamo "The superposition of periodic cell arrival processes in an ATM multiplexer," *IEEE Trans. Commun.*, vol. 39, pp. 298–303, 1991.

[7] K. Liao and L. G. Mason "A heuristic approach for performance analysis of ATM systems," in *GLOBECOM*, pp. 1931–1935, 1990.

[8] J. N. Daigle and J. D. Langford "Models for analysis of packet voice communication systems," *IEEE JSAC*, vol. 4, pp. 1293–1297, 1986.