Markov modulated periodic arrival process offered to an ATM multiplexer

Nail Akar *, Erdal Arıkan

Electrical and Electronics Engineering Department, Bilkent University, 06533 Ankara, Turkey

Received July 1993; revised November 1993

Abstract

When a superposition of on/off sources is offered to a deterministic server, we are faced with a particular queueing system, the analysis of which has a significant role in ATM networks. Periodic cell generation during active times is a major feature of these sources. We provide an analytical approach 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 approximation 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.

Keywords: ATM networks; Fluid-flow models; $nD/D/1$ queue

1. Introduction

The Asynchronous Transfer Mode (ATM) is the preferred transfer mode for the Broadband ISDN (B-ISDN). The B-ISDN will be required to support a wide mix of services (e.g., voice, video, high-speed data) over a common ATM transport network. 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 independent of connections' traffic characteristics and performance requirements. 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
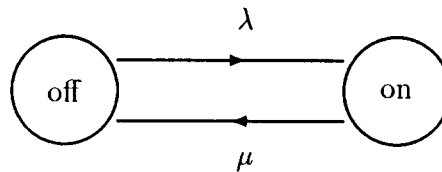
---

* Corresponding author.

Fig. 1. 2-state Markov model for an on-off source.

belonging to a particular connection turns out to have a critical 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 Fig. 1) will be used to describe the above-mentioned 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 easily be extended to construct an N-state Markov chain to describe the superposition process of N on/off sources (Fig. 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. Our approach has a significant feature of combining the discrete-time nature of periodic arrivals in a slotted system and the continuous-time nature of the underlying Markov chain. In general, we call the arrival process as a Markov modulated periodic arrival process.

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 into account of these correlation effects in the performance analysis of the queueing system. A promising approach is to approximate the superposition nonrenewal point process by a renewal process [1] in which positive correlations are accounted for by the choice of the second moment of the cell
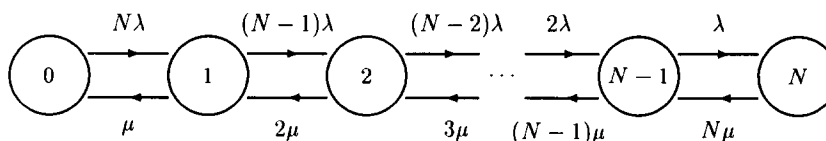

Fig. 2. Birth-death model for the superposition of N on-off sources.

interarrival time distribution. Another common approach is to approximate the superposition by fitting certain parameters of the original process to another stochastic process, the Markov modulated Poisson process (MMPP). The MMPP is a nonrenewal doubly stochastic process where the Poisson arrival intensity is varied according to the state of a continuous-time Markov chain. A 2-state MMPP is proposed in [2] where certain four parameters are chosen to match four arrival process characteristics which is then used to evaluate the average delay for an infinite buffer voice multiplexer with good accuracy. Other choices for the four fitting parameters have also been proposed to yield more accurate results [3,4]. In [5], superposition of N on/off sources is modelled by an N-state MMPP where the arrival intensity is proportional to the number of active sources. This model, however, overestimated the queue lengths due to discard of negative correlation effects. Matrix geometric methods [2] and spectral expansion techniques [6] are proposed computational methods to solve for queueing systems for which the input is an MMPP. These models do not generally give sufficiently accurate results since they do not capture the short-term effects, besides, the extendibility of these models for the performance analysis of more complicated queueing systems (e.g., queues with overload control) is not well understood.

Fluid flow models have attracted the attention of many researchers in the telecommunications literature due to their simplicity. 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 [7] where the arrival rate is modulated with respect to the state of a Markov chain as in MMPP. The authors [7] 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. The model and technique proposed in [7] is generalized for the finite buffer case in [8] to solve for the cell loss rate, a critical value in ATM networks. In spite of inaccuracies encountered in input traffic modeling, many extensions of fluid flow models have been proposed to analyze certain congestion control schemes (e.g., call admission control [4], access regulator [9], cell discarding mechanisms [10]) which are of fundamental importance in ATM networks. It is actually the ease of computation in fluid flow approximations which makes these extensions and generalizations possible.

For an accurate analysis of an ATM multiplexer, the negative correlation between cell interarrival times should be taken into consideration. Actually, when the instantaneous arrival rate is less than the link rate, the queueing system behaves like the so-called $n\mathrm{D}/\mathrm{D}/1$ 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 [11,12] to find the steady-state distribution of the queue length. Different periods are also allowed in [13] where accurate approximate formulas for the queue length distribution are derived.

In the Markov modulated periodic arrivals case, the transient behavior of the $n\mathrm{D}/\mathrm{D}/1$ queue turns out to have a significant role. The focus of the present paper is the derivation of a relationship between fluid sources and periodic sources, arrival rates of which are Markov modulated, through an approximation of the transient behavior of the $n\mathrm{D}/\mathrm{D}/1$ queue. This approximation is mainly based on an interpolation of the queue length whose distribution is exactly 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.

2. 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.

2.1. Transient analysis of the $nD/D/1$ queue

In our queueing model, 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/s in denoted by P and the service rate of the buffer is denoted by C, which actually equals to PR cells/s. 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 = queue length at the end of kth slot,

a_k = number of arrivals in the kth 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

$$\begin{aligned} Q_1 &= \max(a_1, Q_0 + a_1 - 1), \\ Q_2 &= \max(a_2, a_1 + a_2 - 1, Q_0 + a_1 + a_2 - 2) \\ &\vdots \\ Q_R &= \max(\tilde{Q}_n, Q_0 + n - R) \end{aligned} \tag{1}$$

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 [12,14]:

$$\tilde{Q}_n(q) \triangleq \Pr\left(\tilde{Q}_n \leq q\right)$$

$$= \begin{cases} 1 - \displaystyle\sum_{x=1}^{n-\bar{q}} \frac{R-n+\bar{q}}{R-x} C(n, \bar{q}+x)\left(\frac{x}{R}\right)^{\bar{q}+x}\left(1-\frac{x}{R}\right)^{n-\bar{q}-x}, & 0 \leq q \leq n-1, \\ 1, & q > n-1 \end{cases} \tag{3}$$

where $\bar{q}$ is the largest integer smaller than q. In [12], the change in the number of active sources, n, is assumed to happen slowly. In this case, the queue length reaches its steady-state distribution $\tilde{Q}_n(\cdot)$ whenever $n < R$. Equation (3) is then sufficient to compute the distribution of the queue length and the queueing delay assuming that n does not exceed R so that the queueing system is stable. Considering the Markov model which modulates the number of incoming active sources, we have two significant observations: (1) there are possible overload states $(n > R)$ in which case there is no limiting distribution, (2) even for the underload states $(n < R)$, the state holding time is not generally long enough for the queue length to reach its steady-state distribution $\tilde{Q}_n(\cdot)$ before the Markov chain makes a transition to another state. In regard of these observations, the transient analysis of the $n\mathrm{D}/\mathrm{D}/1$ queue becomes the major issue for our purposes.

In order to obtain the queue length evolution equations for $n < R$, we iterate on Eq. (1) on an R-slot basis so that by periodicity of arrivals we have

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

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 [7]

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

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 in fact 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 [15,16]. 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

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

The analysis of the problem can also be made in a discrete-time framework but due to a large deal of literature on fluid flow approximations, continuous-time methods have been preferred to make comparisons clear both in formulation and analysis stages. In the next section, the fundamental approximation in (6) will 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.

2.2. Markov modulated periodic arrivals

Let us now consider the traffic model in Fig. 2 and concentrate on a particular state $(n, 0 \leq n \leq N)$ 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 $t \to \infty$, $\Delta t \to 0$):

$$
F_b(n, x) = \Pr\{S(t) = n\}\overline{F}_b(n, x),
\tag{7}
$$

where

$$
\overline{F}_b(n, x) = \Pr\{X(t) \leq x \mid S(t + \Delta t) = n, S(t) \neq S(t + \Delta t)\}
$$

and

$$
F_e(n, x) = \Pr\{S(t) = n\}\overline{F}_e(n, x),
\tag{8}
$$

where

$$
\overline{F}_e(n, x) = \Pr\{X(t) \leq x \mid S(t + \Delta t) \neq S(t), S(t) = n\}.
$$

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

$$
\overline{F}_e(n, x) = \Pr\{X(t) \leq x \mid S(t) = n\},
$$

and we therefore write

$$
F_e(n, x) = \Pr\{X(t) \leq x, S(t) = n\},
\tag{9}
$$

To interpret, $\overline{F}_b(n, x)$ 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, $\overline{F}_e(n, x)$ 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. Recall that, the state holding time at state n is exponentially distributed with parameter σ_n which is the total

probability flow rate out of state n. Each time the Markov system changes a state, a complete phase randomization of all the sources is assumed to occur 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

$$
F_e(n, x) = \begin{cases} \bar{Q}_n(x) F_f(n, x), & n < R, \\ F_f(n, x), & n \geq R \end{cases} \tag{10}
$$

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 therefore write

$$
F_f(n, x) = \begin{cases} F_b(n, x) * \left(\dfrac{-\sigma_n}{Pn - C} \exp\left(\dfrac{\sigma_n x}{C - Pn} \right) u(-x) \right), & x \geq 0,\ n < R, \\[2mm] F_b(n, x), & n = R, \\[2mm] F_b(n, x) * \left(\dfrac{\sigma_n}{Pn - C} \exp\left(\dfrac{\sigma_n x}{C - Pn} \right) u(x) \right), & x \geq 0,\ n > R. \end{cases} \tag{11}
$$

Here, $*$ is the convolution operator. Note the analogy between the above expression and the output equation (pertaining to $F_f(n, x)$) of a first order linear system with the input $F_b(n, x)$. This analogous linear system is anti-causal in the case $n < R$ whereas it is causal when $n > R$. Writing down the state equations of this system, we now have

$$
\frac{d}{dx} F_f(n, x) = \frac{\sigma_n}{C - Pn} F_f(n, x) + \frac{\sigma_n}{Pn - C} F_b(n, x), \quad x \geq 0,\ n \neq R. \tag{12}
$$

Let $p(m, n)$ be the state transition rate from state m to state n. We now relate $F_b(n, x)$'s to $F_e(\cdot, x)$'s. For this purpose, we rewrite $\bar{F}_b(n, x)$ in definition (7) as $t \to \infty$, $\Delta t \to 0$:

$$
\bar{F}_b(n, x) = \frac{\displaystyle\sum_{m \neq n} \Pr\{X(t) \leq x,\ S(t + \Delta t) = n,\ S(t) = m\}}{\displaystyle\sum_{m \neq n} \Pr\{S(t + \Delta t) = n \mid S(t) = m\} \Pr\{S(t) = m\}}
$$

$$
= \frac{\displaystyle\sum_{m \neq n} \Pr\{X(t) \leq x \mid S(t + \Delta t) = n,\ S(t) = m\} \Pr\{S(t + \Delta t) = n \mid S(t) = m\} \pi_m}{\displaystyle\sum_{m \neq n} p(m, n) \pi_m \Delta t}
$$

$$
= \frac{\displaystyle\sum_{m \neq n} \bar{F}_e(m, x) p(m, n) \pi_m}{\displaystyle\sum_{m \neq n} p(m, n) \pi_m}.
$$

Multiplying the last equality by π_n and recalling the balance equations of the Markov process,

$$\pi_n \sigma_n = \sum_{m \neq n} p(m, n)\pi_m,$$

we have

$$\sigma_n F_b(n, x) = \sum_{m \neq n} p(m, n)F_e(m, x). \tag{13}$$

For our birth-death model,

$$p(n, n + 1) = (N - n)\lambda, \quad n = 0, 1, \ldots, N - 1,$$

$$p(n, n - 1) = n\mu, \quad n = 1, 2, \ldots, N.$$

Combining (10), (12) and (13), we finally obtain the following differential equations for $F_f(n, x)$'s:

$$\frac{\mathrm{d}}{\mathrm{d}x}F_f(n, x) = \frac{\sigma_n}{C - Pn}F_f(n, x) + \frac{1}{Pn - C}\sum_{m \neq n} p(m, n)\tilde{Q}_m(x)F_f(m, x), \quad n \neq R, \tag{14}$$

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

In the above equations, $\tilde{Q}_m(x) \triangleq 1$, $\forall x \geq 0$, $m \geq R$. 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 [7]. Eliminating the algebraic equation pertaining to $F_f(R, x)$ and defining

$$F_f(x) = \left[F_f(0, x) \; F_f(1, x) \cdots F_f(R - 1, x) \; F_f(R + 1, x) \cdots F_f(N, x) \right]^{\mathrm{T}},$$

we finally have

$$\frac{\mathrm{d}}{\mathrm{d}x}F_f(x) = A(x)F_f(x), \quad x \geq 0, \tag{15}$$

where the $N \times N$ matrix $A(x)$ is determined through a suitable arrangement of the differential equations in (14). Actually,

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

and

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

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 (15) has a unique continuous solution described by

$$F_f(x) = \exp(A_i(x - i))F_f(i), \quad x \in [i, i + 1], 0 \leq i \leq R - 2, \tag{16}$$

and

$$F_f(x) = \exp(A(x - (R - 1)))F_f(R - 1), \quad x \geq R - 1. \tag{17}$$

In order to find the initial condition, we make use of the following observations:
(a) For $n > R$, the queue is always increasing, so the queue length cannot be zero. Therefore, $F_f(n, 0) = 0$ for $n > R$.
(b) 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 $x \to \infty$, no positive (unstable) modes of A should be excited by the choice of $F_f(0)$.
(c) The behavior of $F_f(n, x)$ as $x \to \infty$ is easy to write;

$$F_f(n, \infty) = \pi_n, \quad \forall n, \, 0 \le n \le N.$$

Now, let z_i be a stable eigenvalue of A and ϕ_i be its corresponding right eigenvector. Then, by observation (b) and (17), the solution to $F_f(x)$ can be written in the form

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

which yields

$$F_f(R - 1) = F_f(\infty) + \sum_{i=1}^{N-R} \mu_i\phi_i, \tag{18}$$

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 (16), one can write

$$F_f(R - 1) = ZF_f(0) \triangleq \left(\prod_{i=0}^{R-2} \exp(A_i) \right) F_f(0). \tag{19}$$

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

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

where f is of size $R \times 1$. Combining (18) and (19), 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 (16) and (17) complete our description of the queue length distribution at the state transition instants through the equation (10). 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 (19).

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

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

One other aim is actually finding the distribution of the queueing delay rather than the queue length. Queue length can easily be converted to queueing delay by substituting Ct for x.

Table 1
Comparison of approximations of the mean waiting time in the queue ($R = 10$)

No. voice sources	Simulation results (ms.)	% 95 conf. interval	Approximations [ms]	
			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

However, to form the cdf of the queueing delay, each $F_e(n, x)$ should be weighted before summation [8,15]:

$$\Pr(\text{delay} \le t \text{ s}) = \frac{1}{\alpha N} \sum_{n=0}^{N} n F_e(n, Ct)$$

where

$$\alpha = \frac{\lambda}{\lambda + \mu}$$

is the average fraction of sources being active.

2.3. Numerical examples

We first consider a packetized voice system with line speed 320 Kbits/s, voice peak rate 32 Kbits/s, mean active period 353 ms and mean silent period 650 ms. The mean number of cells in a talkspurt is approximately 22. The cells are 64 bytes and the cell transmission time is 1.6 ms. The simulation results are obtained based on the discrete-time queueing scheme as described in Section 2.1. 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 $R - 1$ ($R = 10$ for this example). 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 approximation 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. Figs. 3 and 4 are devoted to the queue length survivor function, which are obtained for the cases $N = 15$ and 20,
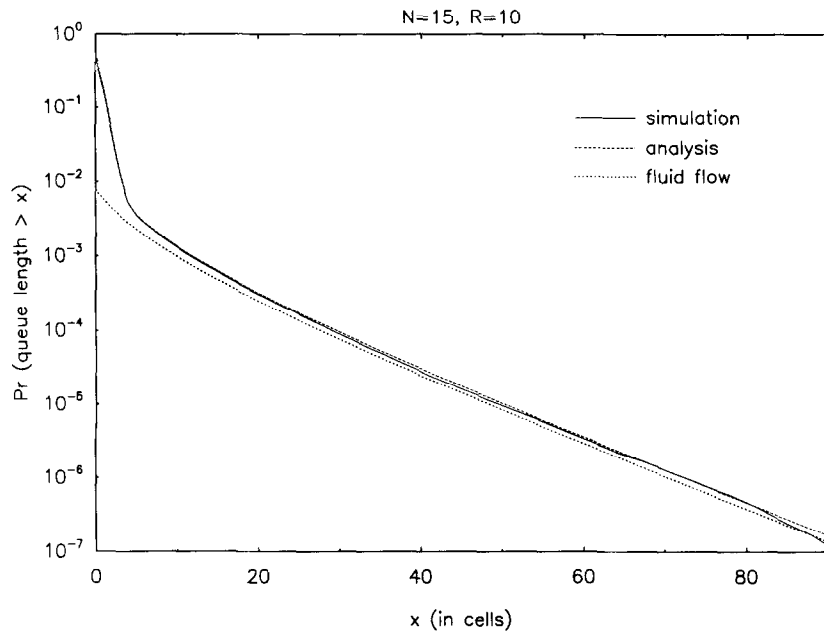
Fig. 3. Comparison of the queue length survivor function for our proposed method with simulation results and the fluid flow approximations ($N = 15$, $R = 10$, utilization = 0.52).
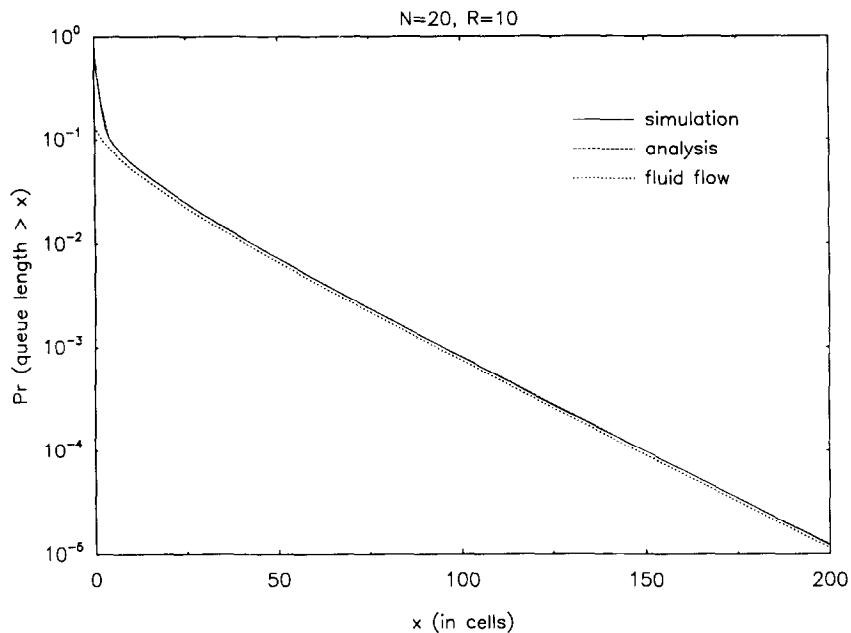

Fig. 4. Comparison of the queue length survivor function for our proposed method with simulation results and the fluid flow approximations ($N = 20$, $R = 10$, utilization = 0.70).

Table 2
Comparison of approximations of the mean waiting time in the queue ($R = 48$)

No. voice sources	Simulation results (ms.)	% 95 conf. interval	Approximations [ms]	
			Analysis	Fluid flow
60	0.1243	± 0.0004	0.0774	0.00
80	0.222	± 0.001	0.186	0.00
90	0.298	± 0.003	0.290	0.00
100	0.431	± 0.016	0.482	0.02
110	0.976	± 0.077	1.096	0.36
120	4.52	± 0.2	4.72	3.49
125	12.51	± 0.7	12.20	10.14
130	36.91	± 1.15	36.20	32.60
132	66.0	± 5.2	64.26	57.90
134	151.5	± 12.7	142.7	124.6

respectively. In both cases, the method we propose is able to capture the simulation curve for the buffer survivor function very accurately.

We then extend the example to the case where the link speed = 1.536 Mbits/s and the cell transmission time = 0.333 ms ($R = 48$ in this case). This is actually the classical packetized voice example found in the literature [8,15]. To avoid numerical inaccuracies in the calculation of the initial condition $F_j(0)$, we approximate A_i by A for $i \geq 9$. Actually, a closer study of $\tilde{Q}_n(\cdot)$
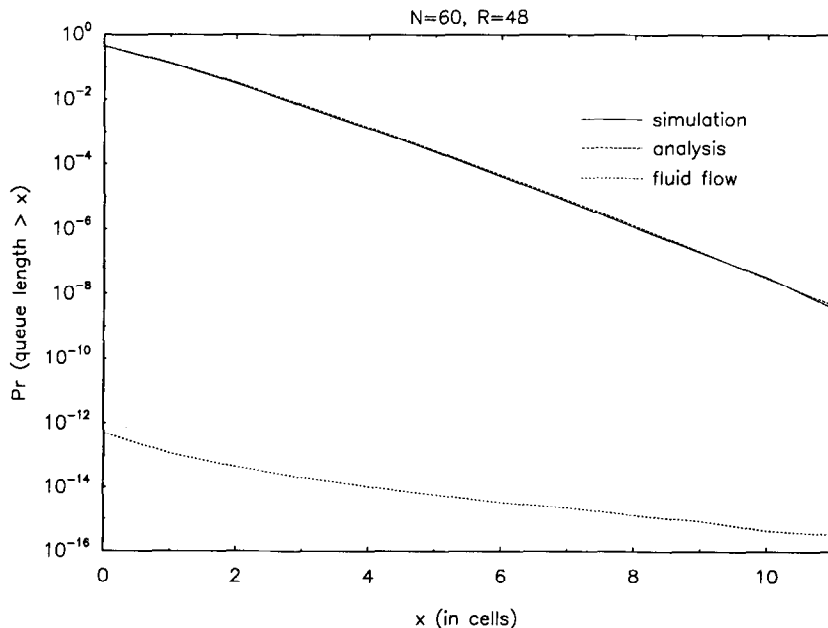

Fig. 5. Comparison of the queue length survivor function for our proposed method with simulation results and the fluid flow approximations ($N = 60$, $R = 48$, utilization = 0.44).
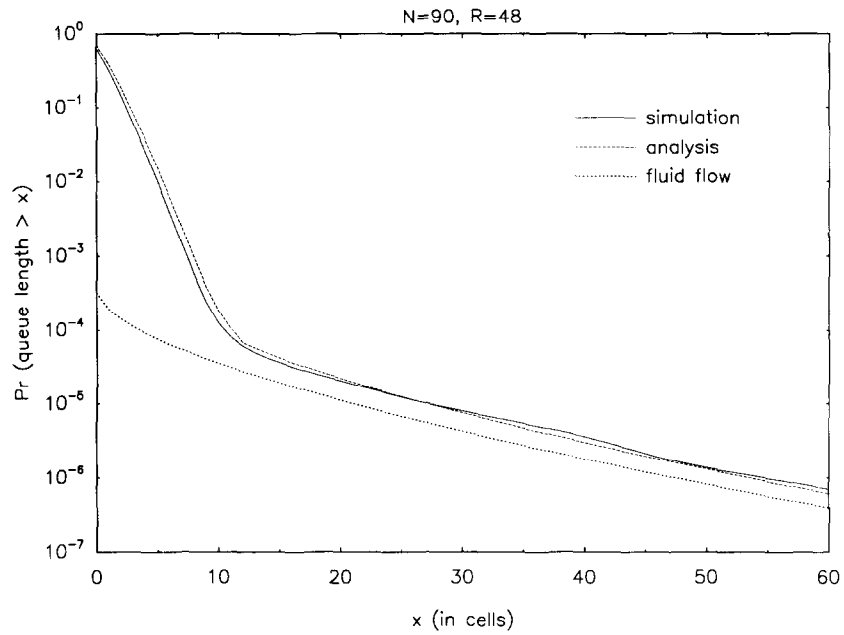
Fig. 6. Comparison of the queue length survivor function for our proposed method with simulation results and the fluid flow approximations ($N = 90$, $R = 48$, utilization = 0.66).
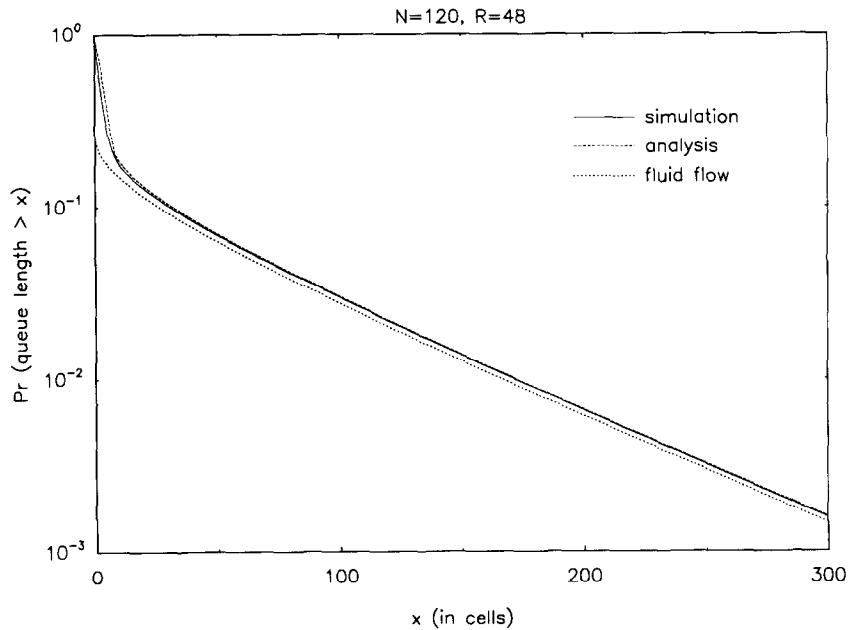

Fig. 7. Comparison of the queue length survivor function for our proposed method with simulation results and the fluid flow approximations ($N = 120$, $R = 48$, utilization = 0.88).

shows that this leads to an error of no more than 2% for each entry of A_i. We note that, as can be verified easily, fluid flow approximation is equivalent to setting

$$A_i = A \quad \forall i \geq 0.$$

We further use the trapezoidal approximation [17]:

$$\exp(A_i) \approx (I - A_i/2)^{-1}(I + A_i/2)$$

to avoid eigensystem calculations. The results associated with the mean waiting time in the queue is given in Table 2.

The queue length survivor functions, for the cases $N = 60$, 90, and 120 are presented in Figs. 5–7, respectively. Remarkably accurate results are obtained for all the cases compared with the fluid flow approximations in spite of the employment of the above-mentioned approximations. The numerical results provided here demonstrate three significant aspects of our proposed method:

- mathematical formulation is simple and similar to fluid flow models and yields better results,
- the method provides satisfactory results through all traffic regimes,
- besides the averaged performance criteria (e.g., mean queueing delay, mean buffer size), the method is able to capture the cdf of the queue length.

3. 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, highly accurate results are obtained in the analysis of a packetized voice multiplexer for all possible loads.

Except for the determination of the linear operator Z defined in (19), 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, moment matching techniques) so that a computationally tractable algorithm is proposed. Usage 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 $\Sigma D_i/D/1$ queue where multiple periods are allowed.

Acknowledgement

References

[1] K. Sriram and W. Whitt, Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE JSAC* **4** (1986) 833–846.

[2] H. Heffes and D.M. Lucantoni, A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, *IEEE JSAC* **4** (1986) 856–868.

[3] K. Liao and L.G. Mason, A discrete-time single server queue with a two-level modulated input and its applications, *Proc. GLOBECOM,* 1989, pp. 913–918.

[4] R. Gusella, Characterizing the variability of arrival processes with indexes of dispersion, *IEEE JSAC* **9** (1991) 203–211.

[5] I. Ide, Superposition of interrupted poisson processes and its application to packetized voice multiplexers, *Proc. ITC-12,* 1988.

[6] A.I. Elwalid, D. Mitra and T.E. Stern, A theory of statistical multiplexing of markovian sources: spectral expansions and algorithms, *Proc. 1st Int. Workshop Numer. Solut. Markov Chains,* 1990.

[7] D. Anick, D. Mitra and M.M. Sondhi, Stochastic theory of a data handling system with multiple sources, *Bell Syst. Tech. J.* **61** (1982) 1871–1894.

[8] R.C.F. Tucker, Accurate method for analysis of a packet-speech multiplexer with limited delay, *IEEE Trans. Comm.* **36** (1988) 479–483.

[9] A.I. Elwalid and D. Mitra, Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation, *Queueing Systems* **9** (1991) 29–64.

[10] A.I. Elwalid and D. Mitra, Analysis and design of rate-based congestion control of high speed networks, II: statistical multiplexing, unpublished notes, 1992.

[11] A.E. Eckberg, The single server queue with periodic arrival process and deterministic service time, *IEEE Trans. Comm.* **27** (1979) 556–562.

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

[13] J.W. Roberts and J.T. Virtamo, The superposition of periodic cell arrival processes in an ATM multiplexer, *IEEE Trans. Comm.* **39** (1991) 298–303.

[14] I. Norros, J.W. Roberts, A. Simonian and J.T. Virtamo, The superposition of variable bit rate sources in an ATM multiplexer, *IEEE JSAC* **9** (1991) 378–387.

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

[16] J.N. Daigle and J.D. Langford, Models for analysis of packet voice communication systems, *IEEE JSAC* **4** (1986) 1293–1297.

[17] J. Vlach and K. Singhal, *Computer Methods for Circuit Analysis and Design,* Van Nostrand Reinhold, New York (1983).

Nail Akar received the B.S. degree from Middle East Technical University, Ankara, and the M.S. degree from Bilkent University, Ankara, in 1987 and 1989, respectively, both in Electrical and Electronics Engineering. Currently, he is working toward the Ph.D. degree at Bilkent University, Ankara. His research interests include performance evaluation and traffic control in ATM networks and applications of linear systems theory and queueing theory to communication networking problems.

Erdal Arıkan received the B.S. degree from Caltech in 1981, the M.S. and Ph.D. degrees from MIT in 1982 and 1985, all in Electrical Engineering. He was an Assistant Professor at the Electrical and Computer Engineering Department and the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign from January 1986 until August 1987. At present, he is an Associate Professor at the Electrical and Electronics Engineering Department of Bilkent University, Ankara, Turkey. His primary research interests are in information theory, coding, and communication networks.