

Tractable Analysis of a Finite Capacity Polling System under Bursty and Correlated Arrivals

Y. Frank Jou

Arne A. Nilsson

Fuyung Lai

Center for Communications and Signal Processing
Department of Electrical and Computer Engineering
North Carolina State University

TR-92/17
September 1992

Tractable Analysis of a Finite Capacity Polling System under ATM Bursty and Correlated Arrivals

Y. Frank Jou and Arne A. Nilsson

Center for Communications and Signal Processing
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, N.C. 27695-7914
U. S. A.

Fuyung Lai

IBM, V57/B660 P.O. Box 12195
Research Triangle Park, N.C. 27709
U. S. A.

Abstract

This paper is concerned with the mean delay and the probability of cell loss that bursty arrivals incur in an ATM switching system which can be modeled as a finite capacity polling system with nonexhaustive cyclic service. The arrival process to each input port of the system is modeled by a Markov Modulated Bernoulli Process (MMBP) which is able to describe the bursty and correlated nature of the ATM traffic. A practical polling system with finite capacity, as the one we deal with here, does not lend itself to an exact solution. In this paper, we introduce a tractable approach to provide an analytical approximation. This approach is validated extensively by comparing it against simulation results under different configurations. It is shown that both the mean delays and the cell loss probabilities obtained from this analysis provide highly accurate estimates.

1 Introduction

Numerous high speed networking approaches have been proposed to meet the stringent requirements of the broadband integrated networks. Among these approaches, the Asynchronous Transfer Mode (ATM) technology has been selected by international standards bodies as the basis for future broadband ISDN facilities [1], [2]. The ATM is a packet oriented transfer mode based on statistical multiplexing, in which the information is transported in short, fixed length packets referred to as cells. ATM provides the means to transporting different types of highly bursty and correlated traffic such as voice, data, animated images and multimedia. The bandwidth flexibility, the capability to handle all services in a uniform way, and the possible use of statistical multiplexing are advantageous features of ATM.

There has been many ATM switch architectures proposed in the literature (see for instance [3] - [6]). The majority of these switch architectures are based on multi-stage interconnection networks [7]. In this paper, we consider the mean delay and cell loss probability that bursty and correlated arrivals incur in an ATM switch architecture as shown in figure 1 [8]. This ATM switch architecture is constructed by connecting self-routing switching modules (SRMs) in a three-stage link configuration which is called a multi-stage self-routing network (MSRN). Each stage of MSRN consists of eight self-routing switching modules. Each module is an 8×8 crossbar switch which has a finite buffer associated with each crosspoint. The cells in each buffer are transmitted in a cyclic order.

To study the performance of this SRM, we model it by a polling system with cyclic service. In the literature, multiqueue systems served by a single server have been the subject of numerous investigations (see [9] - [12], and references therein). Various polling strategies like cyclic or priority service and different types of service disciplines, e.g. exhaustive, gated, or limited service, have been considered. In most of these investigations, the input processes are assumed to be Poisson, and the queues of the polling system are assumed to have infinite capacity. In order to include more realistic modeling elements in the class of polling systems, we consider bursty and correlated arrival processes as inputs into the polling system, which

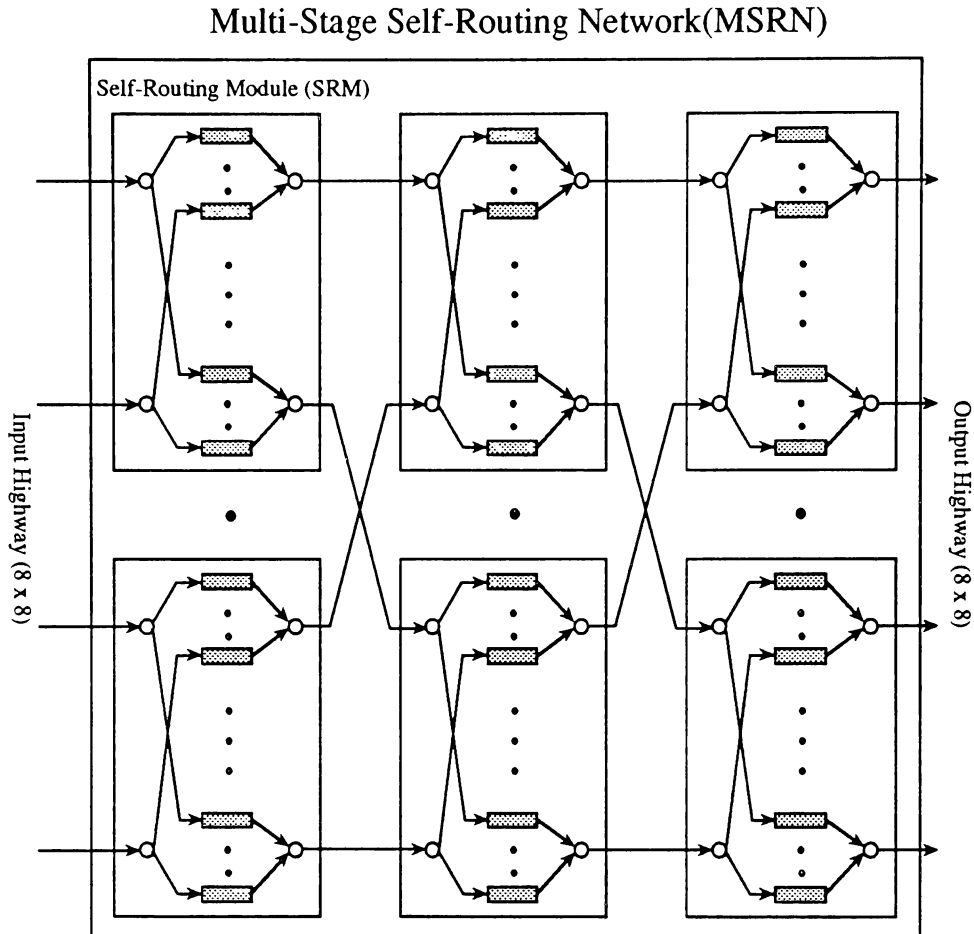


Figure 1: ATM Switching Architecture.

has a finite buffer capacity.

In this paper, we present a queueing model to compute the mean delay and cell loss probability that cells incur in the finite capacity polling system. This model will assume symmetric traffic load, zero switchover time, and ‘limited - 1’ service [13]. In Section 2 we describe in detail the model we propose. This model will require analysis of the queue length distribution of the polling system and a multiple urn model with uniform occupancy probability which are presented in Sections 3 and 4, respectively. In Section 4, we assume that each queue (urn) can accommodate any number of cells up to its capacity with equal probability. The cell loss probability obtained from this approach overestimates, and the

mean delay underestimates the simulation result. To have better accuracy, we use the queue length distribution of a single queue in the polling system as the probability for a particular queue to accommodate a certain number of cells. By assuming the independence among the queues, we then take this probability distribution as a basis to compute the occupancy probability of all the possible configurations. In Section 5, we solve an *MMBP/D/1/L* with vacation queueing model and present an algorithm which originated from the closed queueing model to compute the normalization constant. By incorporating the results from these two subsections into the setup of Sections 3 and 4, we are able to provide highly accurate performance measures. Extensive numerical results validated by computer simulations are given in Section 6. Finally, Section 7 presents our conclusions.

2 Model description

In this section we describe in detail the arrival process and the queueing models which we propose.

2.1 Arrival process

Since most of the traffic sources that an ATM network supports are bursty and correlated, a Poisson process may no longer be suitable for describing the network traffic. For instance, interactive data and a compressed video generate cells at a near-peak rate for a very short period of time. Immediately following a near-peak rate, such a source may become inactive, thus generating no cells. With this scenario, the usual approximation of arrival process by a Poisson process will fail to capture the bursty nature of input traffic and may result in a quite dramatic error in the performance estimation. Kuehn[14] has shown that the system behavior is much more sensitive to arrival processes than to service processes. Knowing the bursty and correlated nature in the ATM traffic, we propose to model the arrival process by a Markov Modulated Bernoulli Process(MMBP) to have a better description of the traffic behavior.

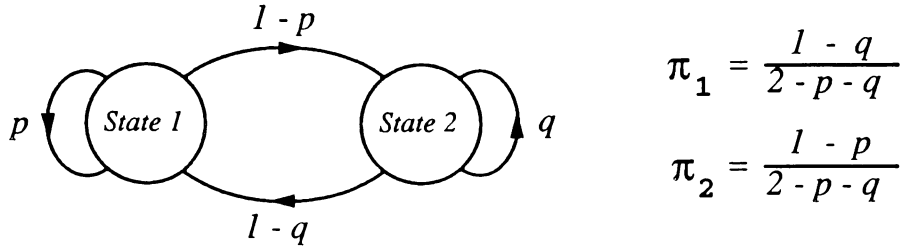


Figure 2: The Markov chain of a two-state MMBP

2.1.1 Generating function of the interarrival time distribution

A general MMBP is a variation of a Bernoulli process where the arrival rate varies according to an m -state Markov chain. This variation enables an MMBP to capture the notation of burstiness and correlation in the arrival traffic. In this paper, we specifically consider a two-state MMBP which is characterized by the transition probability matrix \mathbf{P}_t and the arrival rate matrix $\mathbf{\Lambda}$ defined as the following:

$$\mathbf{P}_t = \begin{bmatrix} p & 1-p \\ 1-q & q \end{bmatrix} \quad \text{and} \quad \mathbf{\Lambda} = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}.$$

The duration for a two-state MMBP to stay in either state is geometrically distributed. Arrivals occur in a Bernoulli fashion with parameters α and β when the process is in states 1 and 2, respectively. Given that the process is in state 1 (or state 2) at slot i , it will remain in the same state in the next slot $i+1$ with probability p (or q), or will change to state 2 (or state 1) with probability $1-p$ (or $1-q$). The transitions between these two states are shown in figure 2, where π_1 and π_2 are the probabilities that the Markov chain is in states 1 and 2, respectively.

Let t be the interarrival time between two successive arrivals and t_i , $i = 1$ or 2 , be the time interval from the moment when the process is in state i until the instant when an arrival occurs. We have

$$t = \begin{cases} t_1 & \text{w.p. } \frac{\alpha(1-q)}{\alpha(1-q) + \beta(1-p)} \\ t_2 & \text{w.p. } \frac{\beta(1-p)}{\alpha(1-q) + \beta(1-p)}. \end{cases}$$

The time intervals t_1 and t_2 can be expressed as the following:

$$t_1 = \begin{cases} 1 & w.p. \quad \alpha p + \beta(1-p) \\ 1 + t_1 & w.p. \quad (1-\alpha)p \\ 1 + t_2 & w.p. \quad (1-\beta)(1-p), \end{cases} \quad (1)$$

and

$$t_2 = \begin{cases} 1 & w.p. \quad \beta q + \alpha(1-q) \\ 1 + t_1 & w.p. \quad (1-\alpha)(1-q) \\ 1 + t_2 & w.p. \quad (1-\beta)q. \end{cases} \quad (2)$$

By taking the z -transform of t_1 and t_2 , we get

$$A_1(z) \equiv E(z^{t_1}) = \frac{\alpha(1-\beta)(1-p-q)z^2 + [\alpha p + \beta(1-p)]z}{(1-\alpha)(1-\beta)(p+q-1)z^2 - [(1-\beta)q + (1-\alpha)p]z + 1}, \quad (3)$$

$$A_2(z) \equiv E(z^{t_2}) = \frac{(1-\alpha)\beta(1-p-q)z^2 + [\alpha(1-q) + \beta q]z}{(1-\alpha)(1-\beta)(p+q-1)z^2 - [(1-\beta)q + (1-\alpha)p]z + 1}. \quad (4)$$

Hence, the generating function of the probability distribution of the interarrival time $A(z) \equiv E\{z^t\}$ is

$$A(z) = \frac{\alpha(1-q)A_1(z) + \beta(1-p)A_2(z)}{\alpha(1-q) + \beta(1-p)}.$$

It can be shown that the mean interarrival time $E\{t\}$ and the squared coefficient of variation of the time between successive arrivals C^2 are as follows:

$$\begin{aligned} E\{t\} &= \frac{2-p-q}{\alpha(1-q) + \beta(1-p)}, \\ C^2 &= \frac{Var(t)}{E\{t\}^2} \\ &= \frac{2[\alpha(1-q) + \beta(1-p)]}{\alpha(1-q) + \beta(1-p) + \alpha\beta(p+q-1)} - \frac{\alpha(1-q) + \beta(1-p)}{2-p-q} \\ &\quad + \frac{2[\alpha(1-p) + \beta(1-q)][\alpha(1-q) + \beta(1-p)](p+q-1)}{(2-p-q)^2[\alpha(1-q) + \beta(1-p) + \alpha\beta(p+q-1)]} - 1. \end{aligned}$$

The average arrival rate, i.e. the probability that a slot contains a cell, λ is

$$\lambda = \frac{\alpha(1 - q) + \beta(1 - p)}{2 - p - q}.$$

In this paper, we assume α equals 1 which will generate the most bursty traffic. By varying p , q , and β , we can have different traffic loads and at the same time change the burstiness and correlation of the arrival process.

2.1.2 Autocorrelation of the interarrival time of an MMBP

Following the formulation in [15], we define t_{ij} as the time interval starting from a particular slot when the arrival process is in state i and ending at a slot when the next arrival occurs and the arrival process is in state j . Therefore,

$$t_{11} = \begin{cases} 1 & w.p. \ \alpha p \\ 1 + t_{11} & w.p. \ (1 - \alpha)p \\ 1 + t_{21} & w.p. \ (1 - \beta)(1 - p), \end{cases}$$

$$t_{21} = \begin{cases} 1 & w.p. \ \alpha(1 - q) \\ 1 + t_{11} & w.p. \ (1 - \alpha)(1 - q) \\ 1 + t_{21} & w.p. \ (1 - \beta)q, \end{cases}$$

$$t_{12} = \begin{cases} 1 & w.p. \ \beta(1 - p) \\ 1 + t_{12} & w.p. \ (1 - \alpha)p \\ 1 + t_{22} & w.p. \ (1 - \beta)(1 - p), \end{cases}$$

$$t_{22} = \begin{cases} 1 & w.p. \ \beta q \\ 1 + t_{12} & w.p. \ (1 - \alpha)(1 - q) \\ 1 + t_{22} & w.p. \ (1 - \beta)(1 - q). \end{cases}$$

Let S_n denote the state of the arrival process when the n^{th} arrival occurs. Also, define T_n as the interarrival time between the $(n - 1)^{\text{th}}$ and n^{th} arrivals, and let $T_{n,j}$ be the interarrival time between the $(n - 1)^{\text{th}}$ and n^{th} arrivals while the n^{th} arrival occurs in state j . If we define

$$A_{ij} \equiv E[z^{T_{n,j}} | S_{n-1} = i],$$

then from the definition of t_{ij} and $T_{n,j}$ we have

$$A_{ij} = E[z^{t_{ij}}], \quad \text{where } 1 \leq i, j \leq 2.$$

Therefore,

$$A_{11}(z) = \alpha pz + (1 - \alpha)pzA_{11}(z) + (1 - \beta)(1 - p)zA_{21}(z), \quad (5)$$

$$A_{21}(z) = \alpha(1 - q)z + (1 - \alpha)(1 - q)zA_{11}(z) + (1 - \beta)qzA_{21}(z), \quad (6)$$

$$A_{12}(z) = \beta(1 - p)z + (1 - \alpha)pzA_{12}(z) + (1 - \beta)(1 - p)zA_{22}(z), \quad (7)$$

$$A_{22}(z) = \beta qz + (1 - \alpha)(1 - q)zA_{12}(z) + (1 - \beta)qzA_{22}(z). \quad (8)$$

Let

$$B_j(z) \equiv E[z^{T_n} | S_{n-1} = j], \quad \text{and} \quad C_i(z_1, z_2) \equiv E[z_1^{T_{n-1}} z_2^{T_n} | S_{n-2} = i]. \quad (9)$$

Hence,

$$C_i(z_1, z_2) = \sum_{j=1}^2 A_{ij}(z_1)B_j(z_2).$$

Define

$$A(z) \equiv \begin{bmatrix} A_{11}(z) & A_{12}(z) \\ A_{21}(z) & A_{22}(z) \end{bmatrix}, \quad B(z) \equiv \begin{bmatrix} B_1(z) \\ B_2(z) \end{bmatrix}, \quad \text{and} \quad C(z_1, z_2) \equiv \begin{bmatrix} C_1(z_1, z_2) \\ C_2(z_1, z_2) \end{bmatrix};$$

we have

$$C(z_1, z_2) = A(z_1)B(z_2).$$

Equations (5) to (8) we rewrite in a matrix form as follows:

$$\begin{bmatrix} 1 - (1 - \alpha)pz & -(1 - \beta)(1 - p)z \\ -(1 - \alpha)(1 - q)z & 1 - (1 - \beta)qz \end{bmatrix} \begin{bmatrix} A_{11}(z) & A_{12}(z) \\ A_{21}(z) & A_{22}(z) \end{bmatrix} = \begin{bmatrix} \alpha pz & \beta(1 - p)z \\ \alpha(1 - q)z & \beta qz \end{bmatrix};$$

therefore,

$$[I - z\mathbf{P}_t(I - \Lambda)]A(z) = \mathbf{P}_t\Lambda z. \quad (10)$$

The term $\mathbf{P}_t(I - \Lambda)$ in equation (10) represents a transition without an arrival and will be denoted as $\mathbf{P}_{t\text{woa}}$. Similarly, the term $\mathbf{P}_t\Lambda$ represents a transition with an arrival and will be denoted as $\mathbf{P}_{t\text{wa}}$. Hence, $A(z)$ can be expressed as

$$A(z) = [I - z\mathbf{P}_{t\text{woa}}]^{-1}\mathbf{P}_{t\text{wa}}z.$$

Notice that

$$B_i(z) = A_{i1}(z) + A_{i2}(z).$$

Therefore, we have

$$\begin{aligned} B(z) &= [I - z\mathbf{P}_t(I - \Lambda)]^{-1}\mathbf{P}_t\vec{\lambda}z \\ &= [I - z\mathbf{P}_{t\text{woa}}]^{-1}\mathbf{P}_t\vec{\lambda}z, \quad \text{where } \vec{\lambda} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \end{aligned}$$

From the definition of $C_i(z_1, z_2)$ in equation (9), we get

$$\begin{aligned} E[z_1^{T_{n-1}} z_2^{T_n}] &= \begin{bmatrix} P(S_{n-2} = 1) & P(S_{n-2} = 2) \end{bmatrix} \begin{bmatrix} C_1(z_1, z_2) \\ C_2(z_1, z_2) \end{bmatrix} \\ &= P_\alpha A(z_1)B(z_2), \end{aligned}$$

where

$$P_\alpha \equiv \begin{bmatrix} P(S_{n-2} = 1) & P(S_{n-2} = 2) \end{bmatrix} = \begin{bmatrix} \frac{\alpha(1-q)}{(\alpha(1-q)+\beta(1-p))} & \frac{\beta(1-p)}{(\alpha(1-q)+\beta(1-p))} \end{bmatrix}.$$

Finally, $E[T_{n-1}T_n]$ is readily obtained through $E[z_1^{T_{n-1}} z_2^{T_n}]$ as

$$\begin{aligned} E[T_{n-1}T_n] &= P_\alpha \frac{dA(z_1)}{dz_1} \frac{dB(z_2)}{dz_2} \Big|_{z_1=1, z_2=1} \\ &= P_\alpha (I - \mathbf{P}_{t\text{woa}})^{-2} \mathbf{P}_{t\text{wa}} (I - \mathbf{P}_{t\text{woa}})^{-2} \mathbf{P}_t \vec{\lambda}. \end{aligned} \quad (11)$$

The autocorrelation coefficient of the interarrival time of MMBP with lag 1 is given by

$$\begin{aligned} \psi_1 &= \frac{\text{Cov}(T_{n-1}, T_n)}{\text{Var}(T_n)} \\ &= \frac{E[T_{n-1}T_n] - E(T_{n-1})E(T_n)}{\text{Var}(T_n)} \\ &= \frac{\alpha\beta(\alpha - \beta)^2(1-p)(1-q)(p+q-1)^2}{C^2(2-p-q)^2[\alpha(1-q) + \beta(1-p) + \alpha\beta(p+q-1)]^2}. \end{aligned} \quad (12)$$

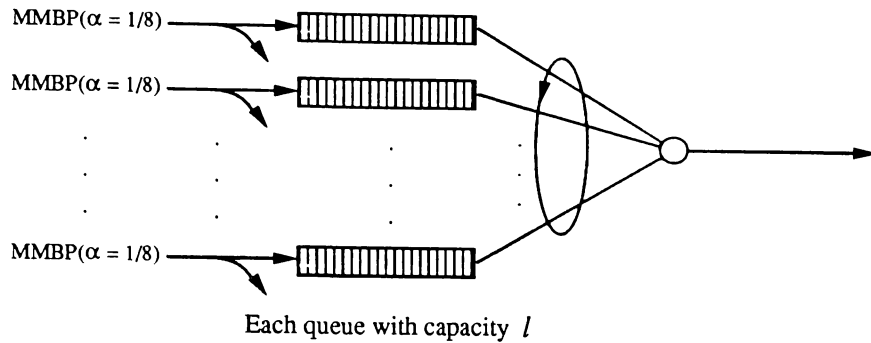


Figure 3: Multiqueue System Served by a Single Server.

As mentioned earlier in this section, an MMBP is a generalization of a Bernoulli process. In fact, an MMBP has two special cases. When α equals β , the process, in essence, only has one single state and becomes a Bernoulli process. If we let either α or β be zero, an MMBP is degenerated to an Interrupted Bernoulli Process (IBP). In both special cases, the autocorrelation of the interarrival time is zero which can be easily validated in equation (12).

2.2 Queueing models

Based on the structure of the SRM, we evaluate its performance by using a multiqueue system as shown in figure 3. Under the assumption of symmetric traffic, the arrival processes to the multiqueue system will be characterized also as MMBP's with the same parameters of p and q as in the original arrival processes. However, since every original arrival process branches to eight possible outlets, the parameters α and β of the arrival processes to the polling system will only be one eighth of the original process.

Instead of considering the queue lengths of the multiqueue system individually, we will call the distribution of the total number of cells in the polling system as the aggregate queue length distribution. In order to obtain this aggregate queue length distribution, it is necessary to consider the blocking effect due to the finite buffer space.

We model the queues in the multiqueue system as multiple urns which have the same limited capacity. Given the number of cells waiting in the system, it is assumed that the

occupancy of each queue is independent from every other queue and the cells are uniformly distributed in any queue, i.e., each position in a queue is equally likely to be occupied. With this model, we can compute the weighting of the cell occupancy configuration which could cause cell loss and then establish the transition matrix which allows us to compute the aggregate queue length distribution. After this distribution is obtained, the mean delay and cell loss probability follow readily.

Notice that given R cells in the system, the occupancy of these cells in reality will more likely be evenly distributed among these queues because in general the server will more frequently visit the queues with longer queue sizes than the queues with shorter queue sizes. Therefore, given the number of cells in the polling system exceeding a single queue capacity, the occupancy configuration which has at least one full queue is less likely to occur. Hence, the assumption of uniform occupancy will give us a conservative estimate which can serve as an upper bound for the cell loss probability of the polling system.

In order to provide better approximation, we refine the uniform assumption such that the probability of a certain queue position to be occupied is according to the queue length distribution of a single queue. Based upon this refined assumption, we propose a queueing model, $MMBP/D/1/L$ with vacation, to obtain the queue length distribution of a particular queue in the polling system. This queue length distribution will serve as a basis for computing the weighting of all the possible configurations. The total weighting of all the cell occupancy configurations can be found by a technique originating from the closed queueing model. We use the sum of these weightings as a normalization factor to modify the weighting of the configurations which could cause cell loss. This refined weighting is to be incorporated in the steady state equations to compute the aggregate queue length distribution which provides us with highly accurate performance measures.

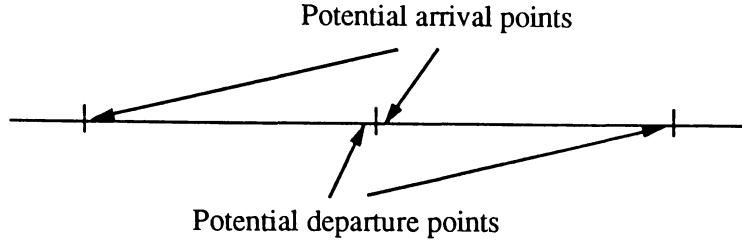


Figure 4: Potential arrival and departure points.

3 Aggregate queue length distribution

In our model, we assume that arrivals can only occur at the beginning of each slot and that departures leave the system at the end of each slot. This arrangement, as illustrated in figure 4, is called an early arrival system according to Hunter [16]. During a slot period, one cell may arrive on each input link, and one cell may be transmitted, given that the system is not empty. The state change, from state 1 to state 2 or vice versa, only occurs at the slot point.

Next, the cell arrival process is analyzed for the cells arriving from all inputs in a slot time. We assume that there are N queues in the multiqueue system. The number of input links in state 1, K , and arrival cells, M , in a unit time can both vary from zero to N . The probability that the number of input links in state 1 is K is

$$P_K(k) = \binom{N}{k} \pi_1^k \pi_2^{N-k},$$

where π_1 and π_2 denote the probabilities that an input link is in state 1 or state 2, respectively.

If the number of input links staying in state 1 is k , then m cells will arrive in a unit time with the probability

$$P_{M|K}(m|k) = \sum_{i=0}^k \binom{k}{i} \alpha^i (1-\alpha)^{k-i} \binom{N-k}{m-i} \beta^{m-i} (1-\beta)^{N-k-m+i}.$$

The state transition probability of having k' lines in state 1 in a slot given k lines in state 1

in the previous slot is given by

$$P_{K'|K}(k'|k) = \sum_{j=0}^k \binom{k}{j} p^j (1-p)^{k-j} \binom{N-k}{k'-j} q^{N-k-k'+j} (1-q)^{k'-j}. \quad (13)$$

In order to describe the blocking effect, we extend the concept introduced in [17] and define a conditional probability $P(m''|m', qsize)$ as

$P\{ m'' \text{ cells accepted} \mid m' \text{ cells arrived and } qsize \text{ cells in the system before arrivals} \}$.

This probability will be further described and computed by a multiple urn model which is discussed in the next section.

Next, we define a two dimensional state variable (K, Q) such that the queue length becomes Q as the result of having M' cells arrive and M'' cells accepted in a slot, given that K input lines are in state 1. The state probability $P_{K,Q}(k, qsize)$ can be obtained by a numerical solution of the following steady state equations:

$$P_{K,Q}(k', qsize') = \sum_{k=0}^N \sum_{m'=0}^{k'} \sum_{qsize=0}^{Q_m} P_{K,Q}(k, qsize) P_{K'|K}(k'|k) P_{M|K}(m'|k') P(m''|m', qsize''), \quad (14)$$

$$\sum_{k=0}^N \sum_{qsize=0}^{Q_m} P_{K,Q}(k, qsize) = 1,$$

where Q_m is the total capacity of the multiqueue system. In equation (14), $qsize''$ and $qsize'$ are given by $qsize'' = \max(qsize - 1, 0)$ and $qsize' = m'' + qsize''$, respectively. From $P_{K,Q}(k, qsize)$, we can sum over K and find the queue length distribution $P_Q(qsize)$. Since we have assumed zero switchover time in the system, the mean output rate λ_{out} can be determined as

$$\begin{aligned} \lambda_{out} &= 1 - P_Q(0) \\ &= \lambda_{in} (1 - P_{loss}). \end{aligned}$$

Therefore, the cell loss probability P_{loss} is obtained as

$$P_{loss} = 1 - \frac{\lambda_{out}}{\lambda_{in}}.$$

After we compute the mean queue length L , the mean delay W can be determined by using Little's result as

$$W = \frac{L}{\lambda_{out}}.$$

4 Multiple urn model

In order to compute the conditional probability $P(m^n | m', qsize)$ defined in the last section, we use a multiple urn model to find $T_{n,r,l}$, the total number of ways to place r indistinguishable balls into n distinguishable boxes, given that the capacity of each box is limited to l . We have

$$T_{n,r,l} = \sum_{i=0}^n (-1)^i \binom{n}{i} \binom{n+r-i(l+1)-1}{n-1}.$$

Also, define $C_{n,r,l}$ as the number of ways of having at least one full box for this multiple urn model. If the total number of balls r is less than l , $C_{n,r,l}$ equals zero. For $kl \leq r < (k+1)l$, it can be shown that

$$C_{n,r,l} = \sum_{i=1}^k \binom{n}{i} (T_{n-i,r-i*l,l} - C_{n-i,r-i*l,l}), \quad (15)$$

where $C_{n,r,l}$ is obtained recursively. Notice that the expression in the summation in equation (15) denotes the number of ways to have exactly i full boxes. For further derivation, the readers are referred to [18].

According to the definition of $P(m^n | m', qsize)$, we now obtain the conditional probability

as

$$P(m''|m', r) = \frac{\sum_{i=0}^{\lfloor \frac{r}{l} \rfloor} \frac{\binom{N-i}{m''} \binom{i}{m'-m''}}{\binom{N}{m'}} \binom{N}{i} (T_{N-i, r-i+l, l} - C_{N-i, r-i+l, l})}{T_{N, r, l}}.$$

5 Refinement of the Approximation

From the computations in Sections 3 and 4, we find that the cell loss probabilities obtained from the analytical model overestimate the results obtained from simulation. On the other hand, the mean delays tend to underestimate. These phenomena are due to the assumption of equal probability among the configurations of occupancy in the multiple urn model. Based upon this observation, we solve an *MMBP/D/1/L* with vacation queueing model and formulate the derivation to compute the normalization factor in order to refine the approximation as discussed in Section 2.2.

5.1 *MMBP/D/1/L* with Vacation

If we focus on a specific queue in the polling system, the queueing model of this particular queue can be identified as an *MMBP/D/1/L* with vacation, where L denotes the capacity of this queue. The vacation refers to the time interval that the server takes to serve the rest of the queues in the system. When the server finds an empty system, the server waits one slot to start the next service cycle. To solve this queueing model, we use the algorithm implemented in Section 3 to obtain the aggregate queue length distribution of the polling system in order to characterize the vacation time distribution of this single queue. To better describe the vacation time distribution, we divide it into two cases. When the server sees an empty queue, the vacation time distribution is denoted by $v(0, k)$, where $1 \leq k \leq N - 1$. Otherwise, we denote the distribution with a service as $v(1, k)$, where $0 \leq k \leq N - 1$. These distributions can be computed by the equations in Section 4 as follows.

Define $D_N(i, j)$ as the number of configurations which have i empty queues, given that there are j cells in the system of N queues. Or equivalently, we can define $D_N(i, j)$ as the number of configurations which have i queues full of holes, given that there are $N \times L - j$ holes in the system. Therefore, from equation (15), $D_N(i, j)$ can be expressed as

$$D_N(i, j) = \binom{N}{i} (T_{N-i, (N-i) \times L - j, L} - C_{N-i, (N-i) \times L - j, L}).$$

The vacation time distribution without a service $v(0, k)$ is given by

$$v(0, k) = \sum_{j=1}^{(N-1) \times L} \frac{D_{N-1}(N-1-k, j)}{T_{N-1, j, L}} \frac{P_Q(j)}{\sum_{i=0}^{(N-1) \times L} P_Q(i)} + \mathbf{1}_{\{k=1\}} \frac{P_Q(0)}{\sum_{i=0}^{(N-1) \times L} P_Q(i)},$$

where $1 \leq k \leq N-1$. Similarly, $v(1, k)$ can be obtained as

$$v(1, k) = \sum_{j=1}^{(N-1) \times L} \frac{\sum_{i=1}^{\min\{j, L\}} D_{N-1}(N-1-k, j-i)}{T_{N, j, L} - T_{N-1, j, L}} \frac{P_Q(j)}{1 - P_Q(0)} + \sum_{j=(N-1) \times L + 1}^{N \times L} \mathbf{1}_{\{k=N-1\}} \frac{P_Q(j)}{1 - P_Q(0)},$$

where $0 \leq k \leq N-1$.

We now proceed to solve the queueing model of $MMBP/D/1/L$ with vacation. The state of the system is defined by the variables (a, n, v) , where a represents the state of the arrival process, n denotes the number of cells in this single queue system, and v indicates the residual vacation time of the server or the number of time slots before next service. The system is observed at a point slightly after the potential arrival point. Once we generate the rate matrix \mathbf{Q} , the stationary probability vector \mathbf{x} is obtained by solving the linear equations $\mathbf{x}\mathbf{Q} = \mathbf{0}$. The time average queue length distribution $p_t(n)$ follows by summing vector \mathbf{x} over all possible values of a and v .

5.2 Computing the Normalization Factor

We recognize that the discrepancy of the results obtained from Sections 3 and 4 is due to the assumption that each position in a queue has the same probability of occupation. In

other words, the queue length distribution under this assumption is $1/L$ uniformly. This is, of course, not the case in reality. In order to refine this assumption, we use the queue length distribution $p_t(r)$ found in Section 5.1 as the probability for a particular queue to accommodate r cells. Given the R cells in the polling system, the stationary distribution of the system state $\mathbf{r} = (r_1, r_2, \dots, r_N)$ is equal to

$$\begin{aligned} P(\mathbf{r}) &= P(r_1, r_2, \dots, r_N) \\ &= \frac{1}{G} \prod_{i=1}^N P_i(r_i), \quad \sum_{i=1}^N r_i = R, \end{aligned} \quad (16)$$

where the states of the various queues are assumed to be independent. The probability $P_i(r_i)$ in equation (16) is given by $p_t(r)$. This setup is exactly the same as the case in the closed queueing network except that we have finite capacity in each queue, i.e. $r_i \leq L$, rather than $r_i \leq R$ in the general cases. Following the same spirit, we extend the derivation given in [19] to compute the normalization factor $G_N(R)$ recursively. In particular, we have

$$\begin{aligned} G_1(R) &= P_1(R) && 0 \leq R \leq L, \\ G_2(R) &= \sum_{r_2=\max(0, R-L)}^{\min(L, R)} P_2(r_2)G_1(R - r_2) && 0 \leq R \leq 2L, \\ &\vdots \\ G_N(R) &= \sum_{r_N=\max(0, R-(N-1)\times L)}^{\min(L, R)} P_N(r_N)G_{N-1}(R - r_N) && 0 \leq R \leq N \times L. \end{aligned}$$

This recurrence relation allows $G_N(R)$ to be computed in $O(NR^2)$ steps.

The above equations are subject to the constraint of having L spaces in each queue. In order to refine the conditional blocking probability in Section 3, we need another normalization factor $\overline{G}_N(R)$ which follows the same derivation, but is subject to the constraint of $L - 1$ buffer spaces. When $R > N \times (L - 1)$, we simply let $\overline{G}_N(R)$ be zero.

Define $P_{full}(K, R)$ as the probability of having R cells in the polling system and K out

of N queues full. Therefore, we have

$$\begin{aligned}
 P_{full}(0, R) &= \frac{\overline{G_N}(R)}{G_N(R)}, \\
 P_{full}(K, R) &= \frac{\binom{N}{K} G_K(K \times L) \overline{G_{N-K}}(R - K \times L)}{G_N(R)}, \quad 1 \leq K \leq N.
 \end{aligned}$$

The refined conditional blocking probability $P(m''|m', R)$ can be expressed as

$$P(m''|m', R) = \sum_{K=0}^{\lfloor \frac{R}{L} \rfloor} \frac{\binom{K}{m' - m''} \binom{N - K}{m''} P_{full}(K, R)}{\binom{N}{m'}}.$$

This blocking probability can be inserted into equation (14) in Section 3 to obtain the desired performance measures.

5.3 Further refinement

Notice that the aggregate queue length distribution and vacation time distributions used to solve the $MMBP/D/1/L$ with vacation queueing model are obtained under the assumption that each position in a queue has the same probability of occupation. After finding both the single and aggregate queue length distributions from Sections 5.1 and 5.2, we need to revise the vacation time distributions and run another iteration to compute the mean delay and cell loss probability. From the experience of our experiment, the convergence is very fast. Normally, the mean delay is quite stable throughout the computation. For loss probability, it usually takes two or three iterations to converge below one percent.

To revise the vacation time distributions, we follow the concept and notation employed in Sections 5.1 and 5.2. Given that there are j cells in the N queues, we compute $D_N(i, j)$, the probability which has i empty queues. Since we focus on the occupancy of the holes instead of the cells, the probability $P_i(r_i)$ in equation (16) is given by $p_i(L - r)$, rather than $p_i(r)$. To make the expression clear, we use $H_N(R)$ to denote the normalization factor when

R refers to the number of holes in the system. Also, $\overline{H_N(R)}$ is defined similarly to $\overline{G_N(R)}$.

Therefore,

$$D_N(i, j) = \binom{N}{i} H_i(i \times L) \overline{H_{N-i}}[(N-i) \times L - j],$$

where $0 \leq i \leq N-1$, and $1 \leq j \leq (N-i) \times L$. Hence, the vacation time distribution without a service $v(0, k)$ can be obtained as

$$v(0, k) = \sum_{j=1}^{(N-1) \times L} \frac{D_{N-1}(N-1-k, j)}{G_{N-1}(j)} \frac{P_Q(j)}{\sum_{i=0}^{(N-1) \times L} P_Q(i)} + \mathbf{1}_{\{k=1\}} \frac{P_Q(0)}{\sum_{i=0}^{(N-1) \times L} P_Q(i)},$$

where $1 \leq k \leq N-1$. Also, $v(1, k)$ is given by

$$\begin{aligned} v(1, k) &= \sum_{j=1}^{(N-1) \times L} \frac{\sum_{i=1}^{\min[j, L]} p_t(i) D_{N-1}(N-1-k, j-i)}{G_N(j) - G_{N-1}(j) p_t(0)} \frac{P_Q(j)}{1 - P_Q(0)} \\ &+ \sum_{j=(N-1) \times L + 1}^{N \times L} \mathbf{1}_{\{k=N-1\}} \frac{P_Q(j)}{1 - P_Q(0)}, \end{aligned}$$

where $0 \leq k \leq N-1$.

6 Numerical results

In this section, we examine several configurations where the presented approximations are compared against the simulation results. The performance measures, mean delay, and cell loss probability are affected by the burstiness and correlation of the arrival processes as well as the buffer capacity of the multiqueue system. In terms of queue capacity, we present two cases where the buffer sizes are 4 and 8, respectively.

In order to show the effect of the arrival burstiness, we vary the squared coefficient of variation (C^2) of the arrival processes from 1, to 20, to 200. These three kinds of burstiness represent three typical cases. When C^2 equals 1, we can regard this arrival process as being smooth. The burstiness of voice is represented by the case where C^2 equals 20. We use $C^2 =$

200 for the burstiness of data traffic. Also, to illustrate the impact of the autocorrelation of the arrival processes, we vary the autocorrelation coefficient of the interarrival time from 0 to 0.4.

Figures 5 and 6 show the mean delay times that the arrivals incur under different burstinesses when the queue capacities are 4 and 8, respectively. The worst case is where C^2 equals 200 and the arrival rate is 0.9. In this worst case, the relative errors between the analytical results and the simulations are 4.2% and 7.0%, when the queue capacities are 4 and 8, respectively. It is clearly shown that the analysis follows the simulation closely. From these figures, we see that for the larger queue capacity, the bursty effect becomes more obvious.

The impact of the burstiness toward the cell loss probability can be seen in figures 7 and 8 where the buffer capacities vary from 4 to 8. From these figures, we see that the analysis traces the simulation nicely when the queue capacity is 4. With the buffer capacity of 8, the analytic results could be twice as high as the simulations. However, a benefit of this analysis is that it overestimates the cell loss probability and provides us with a conservative approximation. Figures 9 to 10 compare the cell loss probabilities under different queue capacities when C^2 is fixed at 20 and autocorrelation coefficients of the interarrival time vary from 0 to 0.4. This comparison clearly illustrates that the impact of the autocorrelation is very significant.

Due to the complexity of the analysis of polling systems, there are some approximations in the literature, which model a polling system by a single queue with vacation. Of course, the difference between these two systems is due to the autocorrelation of the interscanning time of the server. The service mechanism in the polling system is correlated, whereas the vacation time in the single queue system is assumed to be independently and identically distributed. Figures 11 to 12 demonstrate the impact of the autocorrelation of the service process. The performance measures of the single queue model are solved exactly where the vacation time distributions are obtained from the simulation of polling systems. As far as the mean delay is concerned, the largest discrepancies are 30% and 39% when the arrival processes are IBP's with C^2 's equal to 20 and 200, respectively. For cell loss probabilities,

the differences vary from one to four orders of magnitude. These figures demonstrate that the single queue with vacation system is not able to describe the complexity of the finite capacity polling system. Also, these comparisons show the significance of the approach we presented.

7 Conclusion

A realistic polling system with finite capacity does not lend itself to an exact analysis. In this paper, we have presented an effective approach to provide an analytical approximation. For the arrival processes, we take into account the effect of bursty and correlated arrivals which is an essential feature in the ATM environment.

It is shown that the analytical model works well with different burstiness and traffic loads of the arrival process as well as the queue capacity of the polling system. The analysis is also computationally effective. The major portion of the computation time of this analysis is devoted to solving the steady state equations in order to obtain the aggregate queue length distribution. It is our experience that the speedup of the analysis over simulation on average is more than two orders of magnitude.

The strength of this analysis is that it is able to provide accurate performance measures in a short period of time. Therefore, it can be applied to decide system specifications in order to meet the performance requirements under the ATM environment without running the simulations.

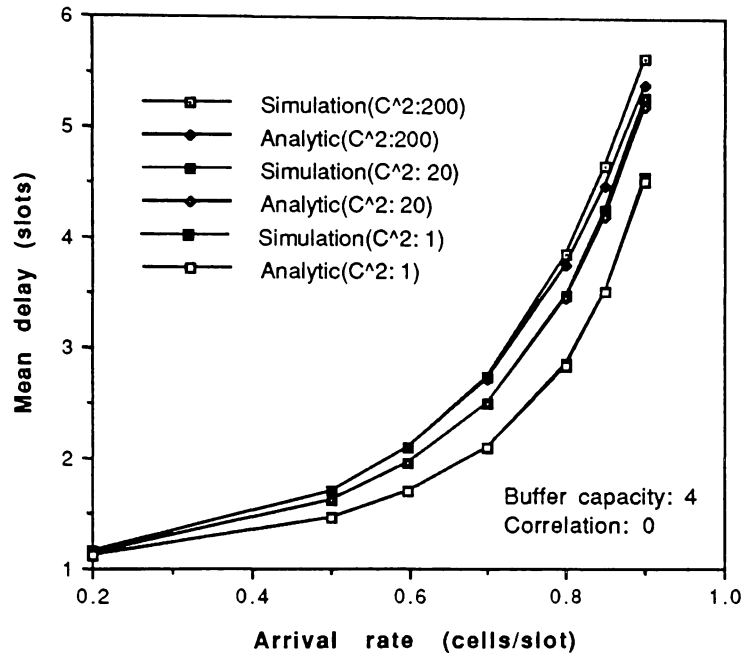


Figure 5: Mean delays incurred in the polling system when queue capacity = 4.

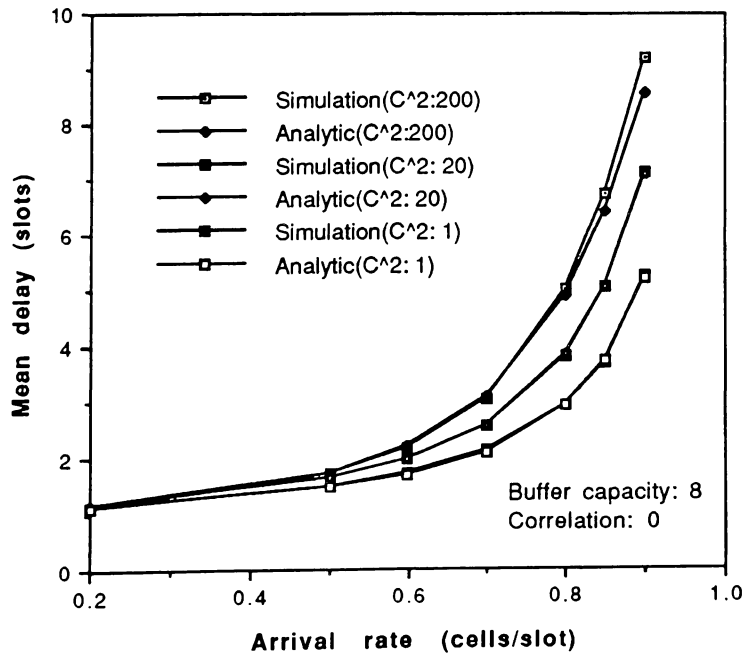


Figure 6: Mean delays incurred in the polling system when queue capacity = 8.

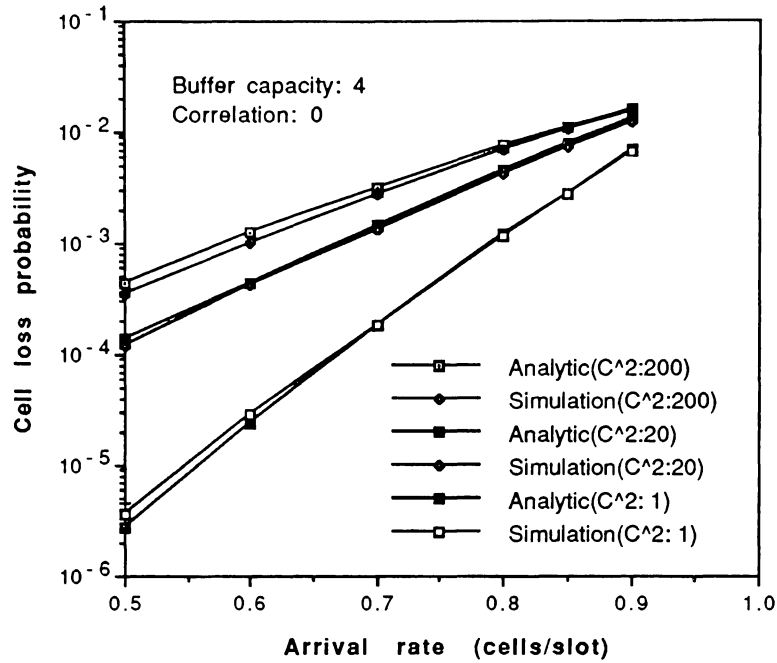


Figure 7: Cell loss probability incurred in the polling system when queue capacity = 4.

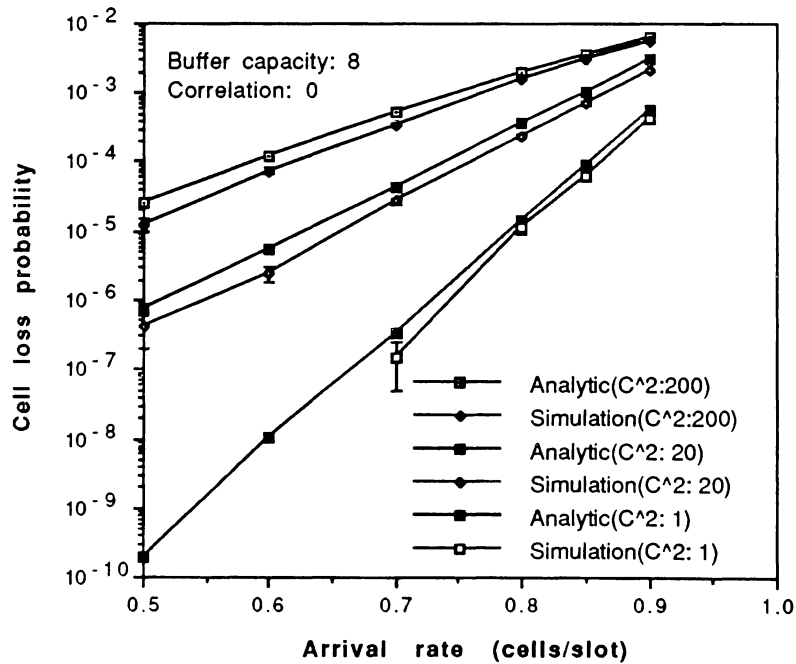


Figure 8: Cell loss probability incurred in the polling system when queue capacity = 8.

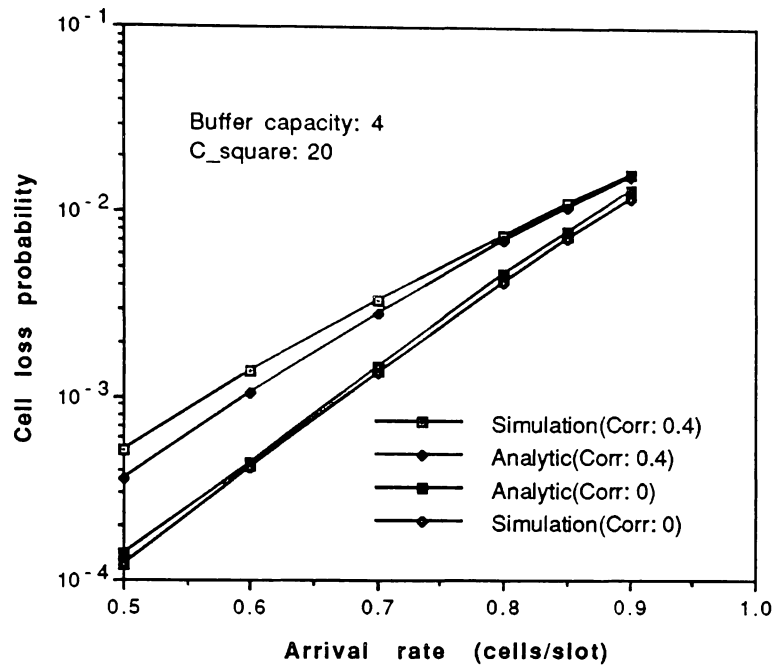


Figure 9: Cell loss probability incurred in a polling system with different arrival correlations when the queue capacity = 4.

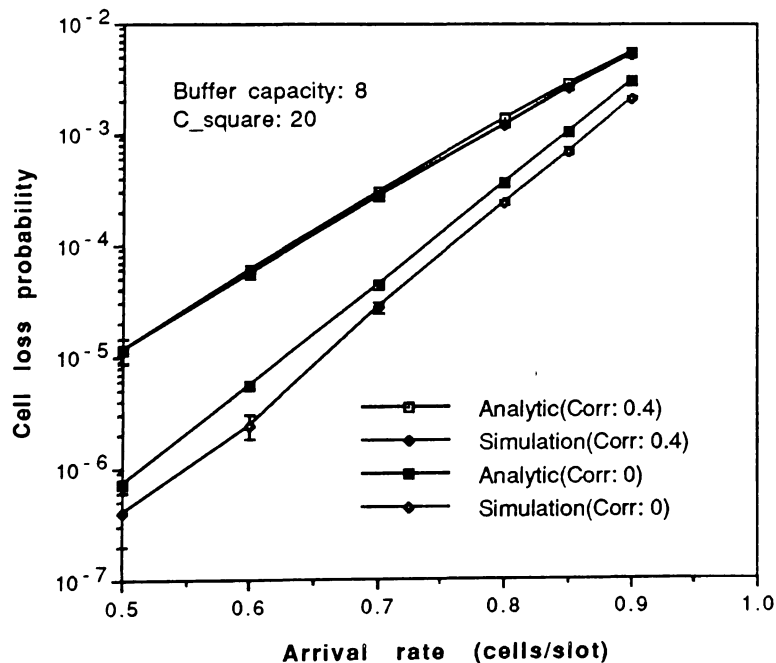


Figure 10: Cell loss probability incurred in a polling system with different arrival correlations when the queue capacity = 8.

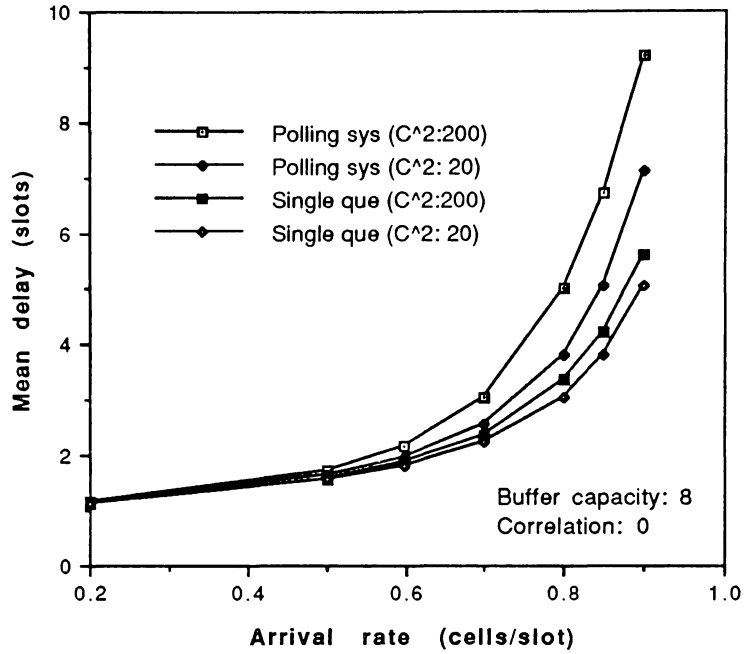


Figure 11: Mean delays incurred in polling and single queue systems.

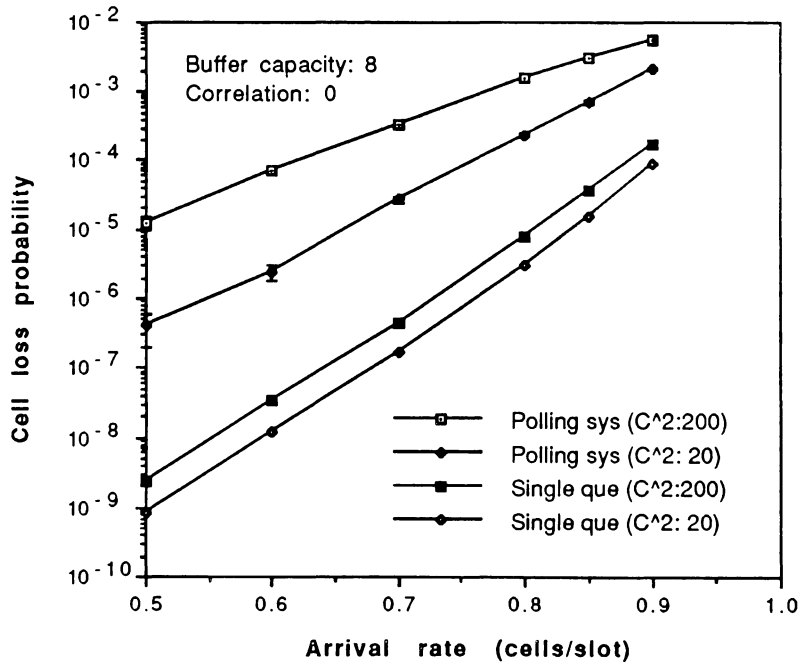


Figure 12: Cell loss probability incurred in polling and single queue systems.

REFERENCES

- [1] "Recommendations drafted by Working Party XVIII/8 (General B-ISDN aspects) to be approved in 1991," *CCITT Report R45*, November 1990.
- [2] "Draft Recommendation I.121: Broadband Aspects of ISDN," *CCITT SG XVIII. Report R34*, June 1990.
- [3] A. A. Nilsson, F.-Y. Lai, and H. G. Perros, "An approximate analysis of a bufferless $N \times N$ synchronous Clos ATM switch," *Proceedings, Canadian Conference on Electrical and Computer Engineering*, pp. 39.1.1–39.1.4, 1990.
- [4] I. Cidon, *et al*, "Real-time packet switching: a performance analysis," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1576–1586, 1988.
- [5] H. Suzuki, *et al*, "Output-buffer switch architecture for asynchronous transfer mode," *IEEE International Conference on Communications*, vol. 1, pp. 99–103, 1989.
- [6] H. Kuwahara, N. Endo, M. Ogino, and T. Kozaki, "Shared buffer memory switch for an ATM exchange," *IEEE International Conference on Communications*, vol. 1, pp. 118–122, 1989.
- [7] F. A. Tobagi, "Fast packet switch architectures for broadband integrated services digital networks," *Proceedings of the IEEE*, vol. 78, no. 1, pp. 133–167, 1990.
- [8] K. Hajikano, K. Murakami, E. Iwabuchi, O. Isono, and T. Kobayashi, "Asynchronous transfer mode switching architecture for broadband ISDN - multistage self-routing switching," *IEEE International Conference on Communications*, vol. 2, pp. 911–915, 1988.
- [9] O.J. Boxma and W P. Groenendijk, "Waiting times in discrete-time cyclic-service systems," *IEEE Transactions on Communications*, vol. 36, no. 2, pp. 164–170, 1988.

- [10] W. Bux, "Token-ring local-area networks and their performance," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 238–256, 1989.
- [11] P. Tran-Gia, "Analysis of polling systems with general input process and finite capacity," *IEEE Transactions on Communications*, vol. 40, no. 2, pp. 337–344, 1992.
- [12] H. Takagi, "Queueing analysis of polling models: an update," *Stochastic Analysis of Computer and Communication Systems*, pp. 267–318, H. Takagi (editor), Elsevier Science Publishers B.V. (North-Holland) Amsterdam, 1990.
- [13] M. Lang and M. Bosch, "Performance analysis of finite capacity polling systems with limited-m service," *Proc. ITC-13*, vol. 14, pp. 731–735, 1991.
- [14] P. J. Kuehn, "Multiqueue systems with nonexhaustive cyclic service," *B. S. T. J.*, vol. 58, no. 3, pp. 671–698, 1979.
- [15] A. A. Nilsson, "Notes on the autocorrelation of the interarrival time of MMBP and MMPP," *Unpublished work*.
- [16] J. J. Hunter, *Mathematical Techniques of Applied Probability, Vol. 2, Ch. 9*. New York, NY: Academic Press, Inc., 1983.
- [17] A. Ganz and I. Chlamtac, "Tractable analytical models of demand assignment protocols in networks with arbitrary buffer capacity," *IEEE Transactions on Communications*, vol. 40, no. 5, pp. 926–939, 1992.
- [18] Y. F. Jou, A. A. Nilsson, and F.-Y. Lai, "The upper bounds for performance measures of a finite capacity polling system under bursty arrivals," *Proceedings of the Second International Conference on QUEUEING NETWORKS WITH FINITE CAPACITY*, Edited by R. O. Onvural and I. F. Akyildiz, Elsevier Science Publishers B.V. (North-Holland) Amsterdam, May 1992.
- [19] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems, Ch. 3*. New York, NY: Academic Press, Inc., 1980.