Discrete-Time Queueing Analysis of a Finite Capacity Polling System with Limited Cyclic-Service Discipline

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-93/4
February 1993

Discrete-Time Queueing Analysis of a Finite Capacity Polling System with Limited Cyclic-Service Discipline

Y. Frank Jou and Arne A. Nilsson
Center for Communications and Signal Processing
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC 27695-7914

Fuyung Lai
IBM, V57/B660 P.O. Box 12195
Research Triangle Park, N.C. 27709

Abstract

This paper is concerned with the mean delay and the loss probability that bursty and correlated arrivals incur in a discrete-time polling system with limited-M cyclic-service where the server serves only a maximum of M customers during each visit to a certain queue. 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 traffic in high-speed communication networks. 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 providing 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 loss probabilities obtained from this analysis provide accurate estimates.

1 Introduction

In the literature, multiqueue systems served by a single server have been the subject of numerous investigations (see [1] [6], and references therein). Such considerable research attention is due to the wide applicability of these models in communication, computer, and production systems. 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 has a finite buffer capacity.

In recent years, the need for performance evaluation of Asynchronous Transfer Mode (ATM) networks has generated a widespread interest in the analysis of discrete-time queueing systems. In an ATM network the size of a packet (commonly known as *cell*) is constant, and therefore the transmission (service) time is constant and defined over a slotted time axis. In view of this application, we present a discrete queueing model to compute the mean delay and loss probability that arrivals incur in the finite capacity polling system. This model will assume symmetric traffic load, zero switchover time, and limited-M cyclic-service where the server serves only a maximum of M customers during each visit to a certain queue. In accordance with the terminology of ATM networks, we shall refer to a customer as a cell.

In Section 2 we describe in detail the model we propose. This model will require analysis of the aggregate queue (or system) length distribution of the polling system which is presented in Sections 3. To compute this queue length distribution, it is necessary to take the blocking effect into account which will be described in Section 4. In Section 4, we treat the polling system as a multiple urn model and compute the blocking probability by two assumptions. It is assumed that, first, the occupancy of each queue is independent from that of every other queue and, second, each queue can accommodate any number of cells up to its capacity with equal probability. The cell loss probability obtained from the second assumption overestimates, and the mean delay underestimates the simulation results. 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. In Section 5, we obtain the vacation time distribution to solve an $MMBP/G/1/L$ queue with vacation time and limited-M service discipline. The obtained single queue length distribution is incorporated into the setup in Sections 3 and 4 to refine the desired performance measures. Extensive numerical results validated by computer simulations are given in Section 6. It is shown that both the mean delays and the loss probabilities obtained from this analysis provide accurate estimates. Finally, Section 7 presents our conclusions.

2 Model description

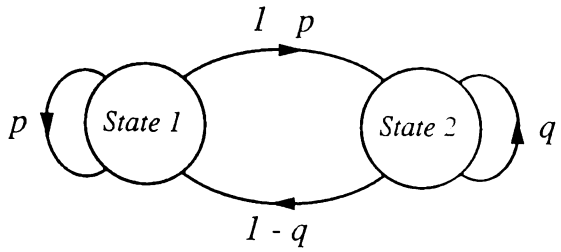
In this section we describe in detail the arrival process and the queueing models which we propose.

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 [8] has shown that the system behavior is much more sensitive to arrival processes than to service processes. Knowing the bursty and correlated traffic nature in ATM networks, we propose to model the arrival process by a Markov Modulated Bernoulli Process (MMBP) or Switched Bernoulli Process (SBP) to have a better description of the traffic behavior.

2.1.1 Generating function of the interarrival time distribution

Hashida *et al.* [9] have characterized the counting process of the Switched Batch Bernoulli Process (SBBP) and derived the probability generating function of the cumulative number of arriving cells during $(0, t]$, which was used to evaluate several important statistical characteristics. In this paper, we provide another approach to characterize the autocorrelation of the



$$\pi_1 = \frac{1 - q}{2 - p - q}$$

$$\pi_2 = \frac{1 \quad p}{2 - p - q}$$

Figure 1: The Markov chain of a two-state MMBP

MMBP. We specifically consider a two-state MMBP which is characterized by the transition probability matrix $\mathbf{P_t}$ and the arrival rate matrix Λ defined as the following:

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

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 1, 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

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

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

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

and

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

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

$$A_1(z) \equiv E(z^{t_1}) = z[\alpha p + \beta(1-p) + (1-\alpha)pA_1(z) + (1-\beta)(1-p)A_2(z)], \tag{3}$$

$$A_2(z) \equiv E(z^{t_2}) = z[\beta q + \alpha(1-q) + (1-\alpha)(1-q)A_1(z) + (1-\beta)qA_2(z)]. \tag{4}$$

Hence, the generating function of the probability distribution of the interarrival time is

$$A(z) \equiv E(z^t) = \frac{\alpha(1-q)A_1(z) + \beta(1-p)A_2(z)}{\alpha(1-q) + \beta(1-p)} = \underline{\mathbf{P}}_s \begin{bmatrix} A_1(z) \\ A_2(z) \end{bmatrix},$$

where $\underline{\mathbf{P}}_s$ is a vector of probabilities that an arrival occurs in state 1 and state 2, respectively. We rewrite equations (3) and (4) in a matrix form as follows:

$$\begin{bmatrix} 1-(1-\alpha)pz & -(1-\beta)(1-p)z \\ -(1-\alpha)(1-q)z & 1-(1-\beta)qz \end{bmatrix} \begin{bmatrix} A_1(z) \\ A_2(z) \end{bmatrix} = z \begin{bmatrix} \alpha p + \beta(1-p) \\ \alpha(1-q) + \beta q \end{bmatrix};$$

therefore,

$$[I - z\mathbf{P_t}(I-\Lambda)] \begin{bmatrix} A_1(z) \\ A_2(z) \end{bmatrix} = z\mathbf{P_t}\vec{\lambda}, \qquad \text{where } \vec{\lambda} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \tag{5}$$

Nilsson and Cui [10] have shown that by differentiating equation (5) and taking $z = 1$, we can easily compute the derivatives of $A(z)$ in a recursive manner , i.e.,

$$\begin{bmatrix} A_1^{(1)}(1) \\ A_2^{(1)}(1) \end{bmatrix} = [I - \mathbf{P_t}(I-\Lambda)]^{-1} \begin{bmatrix} A_1(1) \\ A_2(1) \end{bmatrix} = [I - \mathbf{P_t}(I-\Lambda)]^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{6}$$

and

$$\begin{bmatrix} A_1^{(k)}(1) \\ A_2^{(k)}(1) \end{bmatrix} = k\{[I - \mathbf{P_t}(I-\Lambda)]^{-1} - I\} \begin{bmatrix} A_1^{(k-1)}(1) \\ A_2^{(k-1)}(1) \end{bmatrix}, \qquad \text{where } k \geq 2. \tag{7}$$

After we obtain the derivatives of $A(z)$, it is a simple routine to compute the moments and the squared coefficient of variation (C^2) of the interarrival time.

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

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,

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

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

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

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

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 $(n-1)^{th}$ and n^{th} arrivals, and let $T_{n,j}$ be the interarrival time between the $(n-1)^{th}$ 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

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

Therefore,

$$A_{11}(z) = \alpha pz + (1-\alpha)pzA_{11}(z) + (1-\beta)(1-p)zA_{21}(z), \tag{8}$$

$$A_{21}(z) = \alpha(1-q)z + (1-\alpha)(1-q)zA_{11}(z) + (1-\beta)qzA_{21}(z), \tag{9}$$

$$A_{12}(z) = \beta(1-p)z + (1-\alpha)pzA_{12}(z) + (1-\beta)(1-p)zA_{22}(z), \tag{10}$$

$$A_{22}(z) = \beta qz + (1-\alpha)(1-q)zA_{12}(z) + (1-\beta)qzA_{22}(z). \tag{11}$$

Let

$$B_j(z) \equiv E[\, z^{T_n} | S_{n-1} = j], \quad \text{and} \quad C_i(z_1, z_2) \equiv E[\, z_1^{T_{n-1}} \, z_2^{T_n} \mid S_{n-2} = i]. \tag{12}$$

Hence,

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

Define

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

we have

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

Equations (8) to (11) 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,

$$[I - z\mathbf{P_t}(I - \Lambda)]A(z) \;=\; \mathbf{P_t}\Lambda z. \tag{13}$$

The term $\mathbf{P_t}(I - \Lambda)$ in equation (13) represents a transition without an arrival and will be denoted as $\mathbf{P_{twoa}}$. Similarly, the term $\mathbf{P_t}\Lambda$ represents a transition with an arrival and will be denoted as $\mathbf{P_{twa}}$. Hence, $A(z)$ can be expressed as

$$A(z) = [I - z\mathbf{P_{twoa}}]^{-1}\mathbf{P_{twa}}z.$$

Notice that

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

Therefore, we have

$$\begin{aligned} B(z) \;&=\; [I - z\mathbf{P_t}(I - \Lambda)]^{-1}\mathbf{P_t}\vec{\lambda}z \\ &=\; [I - z\mathbf{P_{twoa}}]^{-1}\mathbf{P_t}\vec{\lambda}z, \qquad\qquad \text{where } \vec{\lambda} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}. \end{aligned}$$

From the definition of $C_i(z_1, z_2)$ in equation (12), we get

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

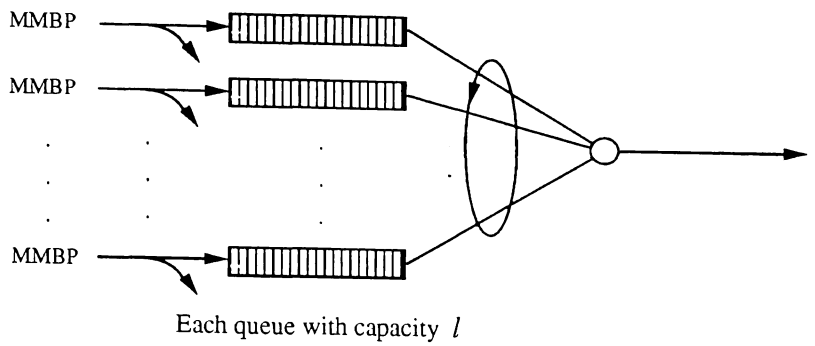


Figure 2: Multiqueue System Served by a Single Server.

where

$$P_a \equiv \left[\begin{array}{cc} P(S_{n-2}=1) & P(S_{n-2}=2) \end{array} \right] = \left[\begin{array}{cc} \frac{\alpha(1-q)}{(\alpha(1-q)+\beta(1-p)} & \frac{\beta(1-p)}{(\alpha(1-q)+\beta(1-p)} \end{array} \right].$$

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

$$\begin{aligned} E[T_{n-1}T_n] &= P_a \frac{dA(z_1)}{dz_1}\frac{dB(z_2)}{dz_2}\,|_{z_1=1,z_2=1} \\ &= P_a(I-\mathbf{P_{twoa}})^{-2}\mathbf{P_{twa}}(I-\mathbf{P_{twoa}})^{-2}\mathbf{P_t}\vec{\lambda}. \end{aligned} \tag{14}$$

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

$$\begin{aligned} \psi_1 &= \frac{Cov(T_{n-1}\ T_n)}{Var(T_n)} \\ &= \frac{E[T_{n-1}\ T_n]-E(T_{n-1})E(T_n)}{Var(T_n)} \\ &= \frac{\alpha\beta(\alpha-\beta)^2(1-p)(1-q)(p+q-1)^2}{C^2(2-p-q)^2[\alpha(1-q)+\beta(1-p)+\alpha\beta(p+q-1)]^2}. \end{aligned} \tag{15}$$

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 (15).

2.2 Queueing model

The polling system we study is shown in figure 2. 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 that of every other queue and the cells are uniformly distributed in any queue, i.e., each queue is equally likely to accommodate any number of cells, from 0 up to its full capacity. With this model, we can compute the weighting of the cell occupancy configuration which might 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. This result 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 for a certain queue to accommodate a certain number of cells is according to the queue length distribution of a single queue in the polling system. Based upon this refined assumption, we propose a queueing model, $MMBP/G/1/L$ queue with vacation time and limited-M service discipline, 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 a 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 enables us to obtain highly accurate performance measures.

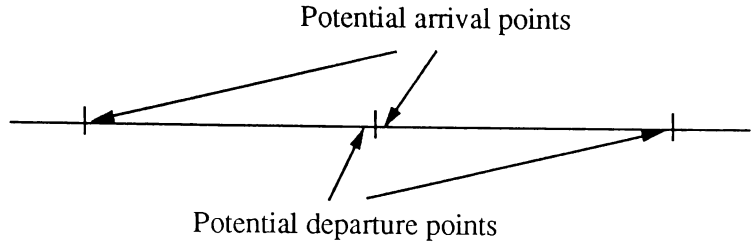


Figure 3: Potential arrival and departure points.

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 3, is called an early arrival system according to Hunter [11]. 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

$$P_{K'|K}(k'|k) = \sum_{j=0}^{k} \binom{k}{j} p^j (1-p)^{k-j} \binom{N-k}{k'-j} q^{N-k-k'+j} (1-q)^{k'-j}. \tag{16}$$

In order to describe the blocking effect, we extend the concept introduced in [12] and define a conditional probability $P(m"|m', qsize)$ as

$P\{$ *m" cells accepted* | *m' cells arrived and qsize 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:

$$P_{K,Q}(k', qsize') = \sum_{k=0}^{N} \sum_{m'=0}^{k'} \sum_{qsize=0}^{Q_m} P_{K,Q}(k, qsize) P_{K'|K}(k'|k) P_{M|K}(m'|k') P(m"|m', qsize"), \quad (17)$$

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

where Q_m is the total capacity of the multiqueue system. In equation (17), $qsize"$ 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 Conditional Blocking Probability

The conditional blocking probability $P(m"|m', qsize)$ defined in the last section can be computed through a multiple urn model which was proposed in [13]. This conditional probability can also be obtained through a more general algorithm by using the concept originating from a closed queueing model (see [14]).

Given R cells in a polling system with N queues, the stationary distribution of the system state $\underline{r} = (r_1, r_2, \cdots, r_N)$ is equal to

$$\begin{aligned} P(\underline{r}) &= P(r_1, r_2, \cdots, r_N) \\ &= \frac{1}{G} \prod_{i=1}^{N} P_q(r_i), \qquad \sum_{i=1}^{N} r_i = R, \end{aligned} \tag{18}$$

where the states of the various queues are assumed to be independent. $P_q(r_i)$ in equation (18) denotes the probability of having r_i cells in the i^{th} queue. This setup is exactly the same as the case in a 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.

Define $G_Z^{X,Y}(R)$ as the normalization factor when there are Z queues in the system, and each queue contains at most X and at least Y cells with a total of R cells present in the system. Following the same form, we extend the derivation given in [14] to compute the normalization factor $G_N^{L,0}(R)$ recursively. In particular, we first assume that each queue can accommodate up to L cells with equal probability, i.e. $P_q(k) = \frac{1}{1+L}, 0 \leq k \leq L$. Hence, we have

$$G_1^{L,0}(R) = P_q(R) \qquad 0 \leq R \leq L, \tag{19}$$

$$G_2^{L,0}(R) = \sum_{r_2=max(0,R-L)}^{min(L,R)} P_q(r_2) G_1^{L,0}(R - r_2) \qquad 0 \leq R \leq 2L, \tag{20}$$

$$\vdots$$

$$G_N^{L,0}(R) = \sum_{r_N=max[0,R-(N-1)\times L]}^{min(L,R)} P_q(r_N) G_{N-1}^{L,0}(R - r_N) \qquad 0 \leq R \leq N \times L. \tag{21}$$

This recurrence relation allows $G_N^{L,0}(R)$ to be computed in $O(NR^2)$ steps.

Define $P_{full}(k, R)$ as the probability of having R cells in the polling system and k out of N queues full. Therefore, it follows

$$P_{full}(0, R) = \frac{G_N^{L-1,0}(R)}{G_N^{L,0}(R)},$$

$$P_{full}(k,R) = \frac{\binom{N}{k} G_k^{L,0}(k \times L) G_{N-k}^{L-1,0}(R - k \times L)}{G_N^{L,0}(R)}, \qquad 1 \le k \le N.$$

Notice that $G_Z^{X,Y}(R) = 0$ if $R \ge X \times Z$, or $R \le Y \times Z$. The conditional blocking probability $P(m"|m', R)$ can be expressed as

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

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 polling system. Based upon this observation, we solve an $MMBP/G/1/L$ queue with vacation time and limited service discipline. This queue length distribution will be used in equations (19) to (21) in Section 4 to refine the conditional blocking probability. In turn, this new conditional blocking probability will be used in equation (17) to refine the desired performance measures.

5.1 $MMBP/G/1/L$ queue with vacation time and limited service discipline

If we focus on a specific queue in the polling system, the queueing model of this particular queue can be identified as an $MMBP/G/1/L$ with vacation time and limited service discipline, 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 aggregate queue length distribution of the polling system 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 $1 \le k \le M \times (N-1)$.

Otherwise, we denote the distribution with a service as $v_1(k)$, where $0 \leq k \leq M \times (N-1)$. The vacation time in each service cycle is of course varying dynamically. Here we approximate it by taking its average in a static fashion. When the server is about to leave a certain queue (service or not), it takes a snapshot of the other $N-1$ queues and decides the duration of its vacation before the next service cycle.

To obtain the vacation time distributions, we can exhaust every possible occupancy configuration and decide its respective vacation time. However, the complexity of this approach is at the order of $O(N!)$ and is cost-prohibitive to implement. The other way of obtaining these distributions is to approach from the reverse direction, namely, for each $v_0(k)$ or $v_1(k)$ we exhaust all possible configurations which can contribute to these vacation time distributions. This second approach can be realized through a similiar computation as given in Section 4 as follows, where $1 \leq M \leq L$,

$$\begin{aligned} v_0(k) &= \sum_{i=k}^{UB_1} \sum_{j=LB_2}^{\lfloor \frac{k}{M} \rfloor} \frac{\binom{N-1}{j} G_j^{L,M}(i-k+j\times M) G_{N-j-1}^{M-1,0}(k-j\times M)}{G_{N-1}^{L,0}(i)} \frac{P_Q(i)}{\sum_{n=0}^{(N-1)\times L} P_Q(n)} \\ &+ \mathbf{1}_{\{k=1\}} \frac{P_Q(0)}{\sum_{n=0}^{(N-1)\times L} P_Q(n)}, \qquad 1 \leq k \leq M \times (N-1), \end{aligned} \tag{22}$$

and

$$\begin{aligned} v_1(k) &= \sum_{m=1}^{L} P_q(m) \sum_{i=k}^{UB_1} \sum_{j=LB_2}^{\lfloor \frac{k}{M} \rfloor} \frac{\binom{N-1}{j} G_j^{L,M}(i-k+j\times M) G_{N-j-1}^{M-1,0}(k-j\times M)}{G_N^{L,0}(i+m) - G_{N-1}^{L,0}(i+m) P_q(0)} \\ &\times \frac{P_Q(i+m)}{1-P_Q(0)}, \qquad 0 \leq k \leq M \times (N-1), \end{aligned} \tag{23}$$

where

$$\begin{aligned} UB_1 &= \left\lfloor \frac{k}{M} \right\rfloor \times L + k \bmod M, \\ LB_2 &= max(\,0,\, \left\lceil \frac{i-(N-1)\times(M-1)}{L-(M-1)} \right\rceil), \\ \mathbf{1}_{\{k=1\}} &= \begin{cases} 1 & k=1, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Notice that we have the initial conditions $G_0^{M-1,0}(0) = 1$, and $G_0^{L,M}(0) = 1$. UB_1 is the upper bound or the maximum number of cells which are present in the rest of $N-1$ queues to cause

k slots of vacation time before the server starts its next service cycle. LB_2 is the minimum number of queues that each one of them has buffered at least M cells.

To understand equation (22), we assume that the server finds i cells in the rest of $N-1$ queues after it leaves a typical queue, say queue T, without service and has to take k slots before it comes back to visit queue T. If there are j queues, each accommodating at least M cells, then these j queues will account for $j \times M$ slots of vacation time. The server will spend $k - j \times M$ slots to serve those remaining $N - j - 1$ queues, each accommodating at most $M-1$ cells. Through simple arithmetics, we know that there are $i - (k - j \times M)$ cells buffered in those j queues. The second line of equation (22) explains one of the operational assumptions that when the server finds an empty system, it waits one slot to serve queue T which starts another service cycle. Equation (23) can be understood by the same arguments.

We now proceed to solve the queueing model of $MMBP/G/1/L$ with vacation time and limited-M service discipline. The basic techniques used in this derivation are similar to those used in [15]. The steady state distribution of queue length is computed by the embedded Markov chain approach. Cells arrive at the system in accordance with an MMBP with parameters α, β, p, and q. There is a single server whose service time on a specific queue can be up to M slots. If an arriving cell finds exactly L cells in the system, then this cell is lost without being served.

The system will be examined at time epochs $\{t_0, t_1, \cdots\}$ of service completion or vacation termination. The state space of the system is $\{\eta_i, \xi_i, R_i\}$, defined as follows. The state of the arrival process at the embedded point t_i is denoted by η_i, where η_i is 1 if the MMBP is in its state 1, or 2 if it is in state 2. If the point t_i is a vacation termination instant, i.e., the point when the server starts to scan this particular queue, then $\xi_i = 0$; otherwise $\xi_i = m$, $m = 1, \cdots, M$, indicating that the point t_i is the service completion instant of the m^{th} cells in the present busy period. R_i is the number of cells in the system at the embedded point t_i. The state transitions occur at cell departure instants or vacation termination instants. Let $X_{j,m,l}$ and $Y_{j,l}$ represent the limiting probability distributions defined as follows

$$\begin{aligned} X_{j,m,l} &\equiv \lim_{i\to\infty} P_r\{\eta_i = j, \xi_i = m, R_i = l\} \qquad 0 \le l \le L-1, \\ Y_{j,l} &\equiv \lim_{i\to\infty} P_r\{\eta_i = j, \xi_i = 0, R_i = l\} \qquad 0 \le l \le L. \end{aligned}$$

When the system is under steady state, these limiting probability distributions satisfy the following equations:

$$X_{j,1,l-1} = Y_{j,l}, \quad j = 1,2; \quad 1 \le l \le L, \tag{24}$$

$$X_{j,m,l} = \sum_{k=l}^{l+1} \sum_{i=1}^{2} c_{i,j,l-k+1} X_{i,m-1,k}, \quad j = 1,2; \quad 2 \le m \le M; \quad 0 \le l \le L-2, \tag{25}$$

$$X_{j,m,L-1} = \sum_{i=1}^{2} c_{i,j,1} X_{i,m-1,L-1}, \quad j = 1,2; \quad 2 \le m \le M, \tag{26}$$

$$Y_{j,l} = \sum_{i=1}^{2} g_{i,j,l} \sum_{m=1}^{M-1} X_{i,m,0} + \sum_{i=1}^{2} \sum_{k=0}^{l} g_{i,j,l-k} X_{i,M,k} + \sum_{i=1}^{2} h_{i,j,l} Y_{i,0}, \quad j = 1,2; \quad 0 \le l \le L-1, \tag{27}$$

$$Y_{j,L} = \sum_{i=1}^{2} \left[g^c_{i,j,L} \sum_{m=1}^{M-1} X_{i,m,0} + \sum_{k=0}^{L-1} g^c_{i,j,L-k} X_{i,M,k} + h^c_{i,j,L} Y_{j,0} \right], \quad j = 1,2, \tag{28}$$

where $c_{i,j,k}$ is a one slot transition probability that the arrival process starts at state i and ends at state j and there are k cells, either 0 or 1, arriving during this one slot transition. Parameters $g_{i,j,k}$ and $h_{i,j,k}$ denote the transition probabilities that k cells arrive during a vacation time after the server visits queue T with a service and without a service, respectively. The transition probabilities $g^c_{i,j,k}$ and $h^c_{i,j,k}$ are given as $g^c_{i,j,k} = \sum_{m=k}^{(N-1)\times M} g_{i,j,m}$, and $h^c_{i,j,k} = \sum_{m=k}^{(N-1)\times M} h_{i,j,m}$. We have

$$\begin{aligned} c_{1,1,0} &= p(1-\alpha), & c_{1,1,1} &= p\alpha, \\ c_{1,2,0} &= (1-p)(1-\beta), & c_{1,2,1} &= (1-p)\beta, \\ c_{2,1,0} &= (1-q)(1-\alpha), & c_{2,1,1} &= (1-q)\alpha, \\ c_{2,2,0} &= q(1-\beta), & c_{2,2,1} &= q\beta. \end{aligned}$$

The transition probabilities $g_{i,j,k}$ and $h_{i,j,k}$ can be obtained as follows. Define $f(j,n|i,k)$ as the probability of having n slots in state 1 among k transition slots with the arrival process starting at state i and ending at state j between two immediate vacation termination instants. This probability can be obtained recursively by the following equations:

$$\begin{aligned} f(1,n|1,k) &= (1-q)\, f(2,n-1|1,k-1) + p\, f(1,n-1|1,k-1), \\ f(2,n|1,k) &= q\, f(2,n|1,k-1) + (1-p)\, f(1,n|1,k-1), \\ f(1,n|2,k) &= (1-q)\, f(2,n-1|2,k-1) + p\, f(1,n-1|2,k-1), \\ f(2,n|2,k) &= q\, f(2,n|2,k-1) + (1-p)\, f(1,n|2,k-1), \end{aligned}$$

where the initial conditions are given by

$$\begin{aligned} f(1,0|1,1) &= 0, & f(1,1|1,1) &= p, \\ f(2,0|1,1) &= 1-p, & f(2,1|1,1) &= 0, \\ f(1,0|2,1) &= 0, & f(1,1|2,1) &= 1-q, \\ f(2,0|2,1) &= q, & f(2,1|2,1) &= 0. \end{aligned}$$

We further define $F(j,l|i,k)$ as the probability of having l cells arriving in k transition slots with the arrival process starting at state i and ending at state j during a vacation period. The probability $F(j,l|i,k)$ can be expressed in terms of $f(j,n|i,k)$ as

$$F(j,l|i,k) = \sum_{n=0}^{k} \sum_{m=0}^{n} \binom{n}{m} \alpha^m (1-\alpha)^{n-m} \binom{k-n}{l-m} \beta^{l-m} (1-\beta)^{k-n-l+m} f(j,n|i,k).$$

From the definition of $g_{i,j,l}$ and $h_{i,j,l}$, we have

$$\begin{aligned} g_{i,j,l} &= \sum_{k=1}^{(N-1)\times M} F(j,l|i,k) v_1'(k), \\ h_{i,j,l} &= \sum_{k=1}^{(N-1)\times M} F(j,l|i,k) v_0(k), \end{aligned}$$

where $v_1'(1) = v_1(0) + v_1(1)$, and $v_1'(k) = v_1(k)$, $2 \le k \le (N-1) \times M$. Combining the steady state equations (24) and (28) together with the normalization equation

$$\sum_{i=1}^{2} \left(\sum_{j=1}^{M} \sum_{k=0}^{L-1} X_{i,j,k} + \sum_{k=0}^{L} Y_{i,k} \right) = 1,$$

we can easily solve for $X_{i,j,k}$ and $Y_{i,k}$. The single queue length distribution P_{MMBP_Q} is readily found to be

$$P_{MMBP\_Q}(k) = \sum_{i=1}^{2} \left(\sum_{j=1}^{M} X_{i,j,k} + Y_{i,k} \right), \qquad 0 \le k \le L.$$

(For simplity, we assume $X_{i,j,L} = 0$.) This queue length distribution $P_{MMBP\_Q}(k)$ is to be used in Section 4 as $P_q(k)$ to refine the conditional blocking probability.

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 polling system. It is assumed that there are eight queues in the multiqueue system.

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 the 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. In terms of queue capacity, we present two cases where the buffer sizes are 4 and 8, respectively.

Figures 4 and 5 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 1.1% and 3.0%, 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 6 and 7 where the buffer capacities vary from 4 to 8. From these figures, we see that the analysis traces the simulation nicely when C^2s are 20 and 200.

Figures 8 to 9 compare the mean cell delay and loss probability when queue capacity is 8, C^2 is fixed at 20 and autocorrelation coefficient of the interarrival time is either 0 or 0.4. This comparison clearly illustrates that the impact of the autocorrelation is very significant.

When we vary the service discipline from limited-1 to limited-3, we see that generally the cell loss probability becomes higher and mean delay becomes smaller. However, the differences are not significant, especially when the queue capacity is increased. The results of this experiment are shown in Figures 10 and 11 when the buffer capacity is 4. The approximation tends to overestimate the cell loss probabilities when the value of M is increased. Generally, our analysis provides very accurate estimates in mean cell delays.

7 Conclusion

Due to the practical applicability, polling systems have been the subject of numerous investigations in the literature. Many exact analyses have been reported when the queue capacity is assumed to be infinite. However, 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 analytical approximations. For the arrival processes, we take into account the effect of bursty and correlated arrivals which is an essential feature in ATM networks. Also, this paper provides an analytical structure to consider limited service discipline.

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.

References

[1] 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.

[2] W. Bux, "Token-ring local-area networks and their performance," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 238–256, 1989.

[3] K. K. Leung, "Cyclic-service systems with probabilistically-limited service," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 2, pp. 185–193, 1991.

[4] 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.

[5] 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.

[6] P. Tran-Gia and T. Raith, "Performance analysis of finite capacity polling systems with nonexhaustive service," *Performance Evaluation*, vol. 9, pp. 1–16, 1988.

[7] G. Kimura and Y. Takahashi, "Traffic analysis for token ring systems with limited service," *Electronics and Communications in Japan, Part I*, vol. 71, no. 8, pp. 80–90, 1988.

[8] P. J. Kuehn, "Multiqueue systems with nonexhaustive cyclic service," *B.S.T.J.*, vol. 58, no. 3, pp. 671–698, 1979.

[9] O. Hashida, Y. Takahashi, and S. Shimogawa, "Switched batch bernoulli process (SBBP) and the discrete-time SBBP/G/1 queue with application to statistical multiplexer performance," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 3, pp. 394–401, 1991.

[10] Z. Cui and A. A. Nilsson, "The departure process of a discrete-time finite capacity system with correlated arrivals," *submitted for publication.*

[11] J. J. Hunter, *Mathematical Techniques of Applied Probability, Vol. 2, Ch. 9.* New York, NY: Academic Press, Inc., 1983.

[12] 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.

[13] 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.

[14] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems, Ch. 3.* New York, NY: Academic Press, Inc., 1980.

[15] Tony T. Lee, "M/G/1/N queue with vacation time and limited service discipline," *Performance Evaluation*, vol. 9, pp. 181–190, 1989.

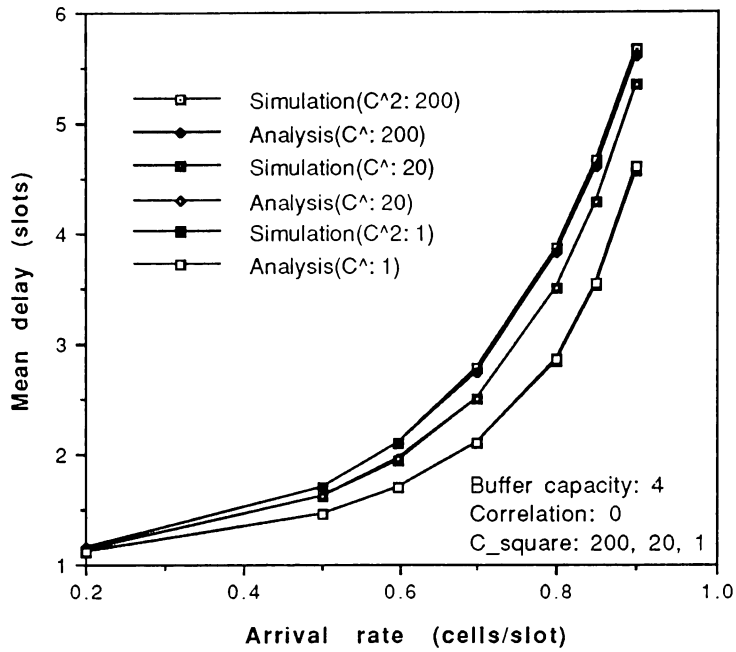


Figure 4: Mean delays incurred in the polling system when queue capacity = 4.

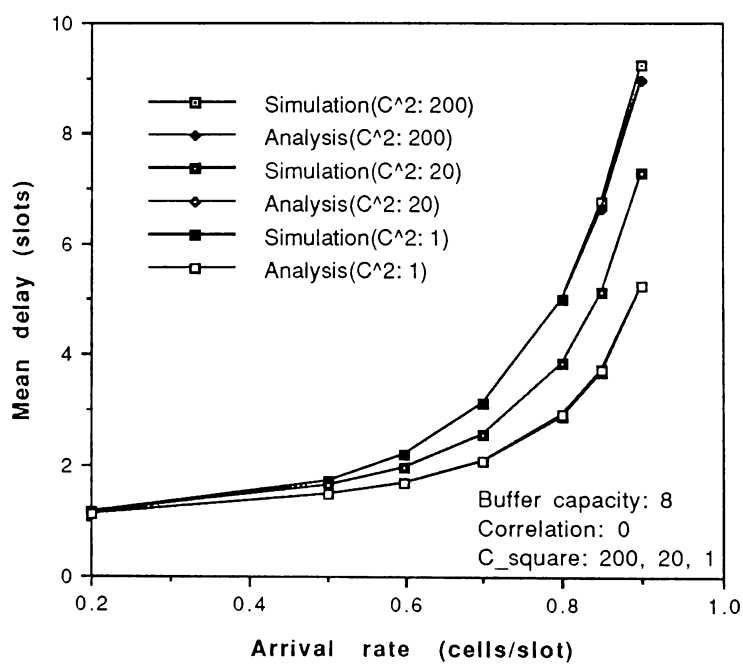


Figure 5: Mean delays incurred in the polling system when queue capacity = 8.

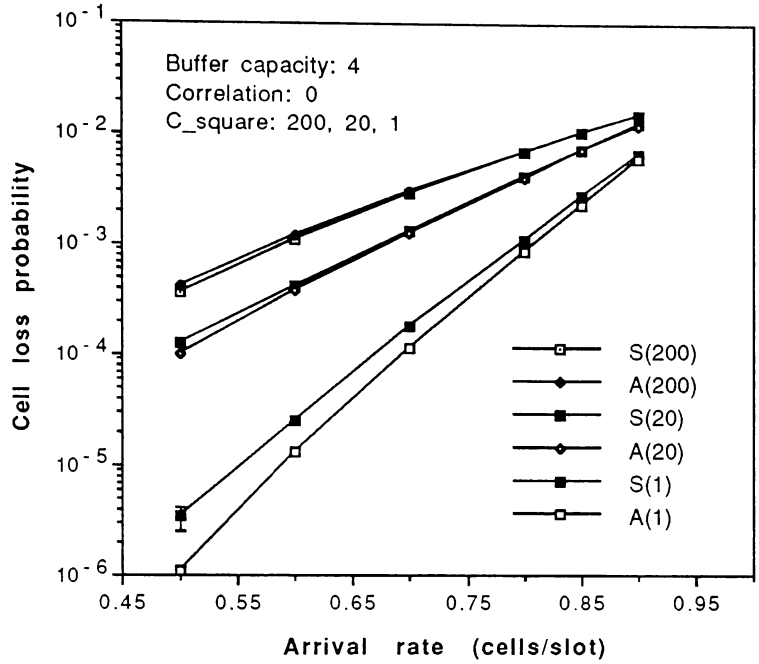


Figure 6: Cell loss probability incurred in the polling system when queue capacity = 4.

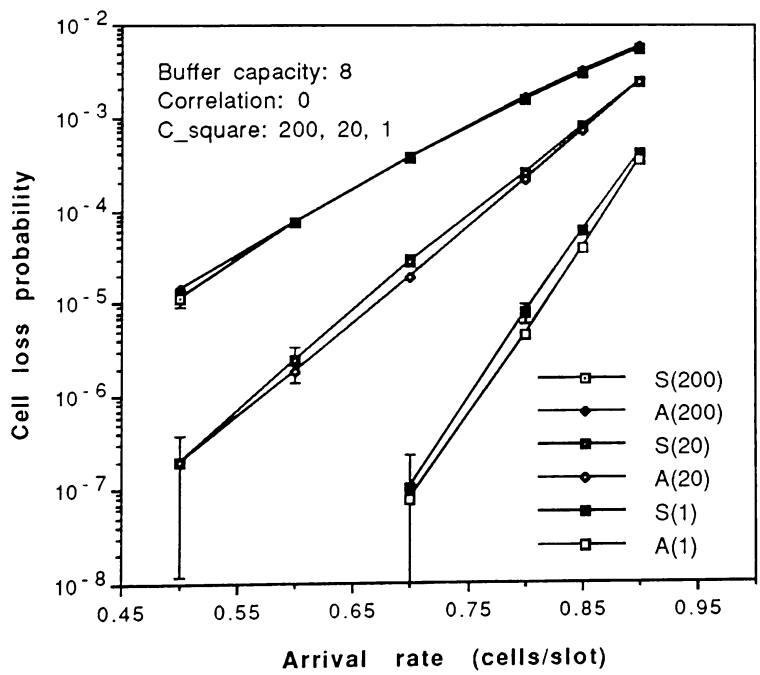


Figure 7: Cell loss probability incurred in the polling system when queue capacity = 8.

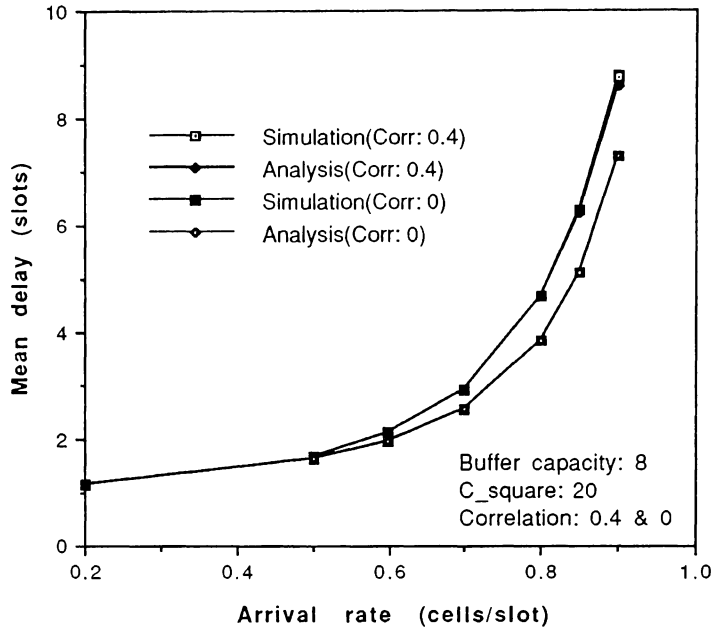


Figure 8: Mean delays incurred in a polling system with different arrival correlations when the queue capacity = 8.

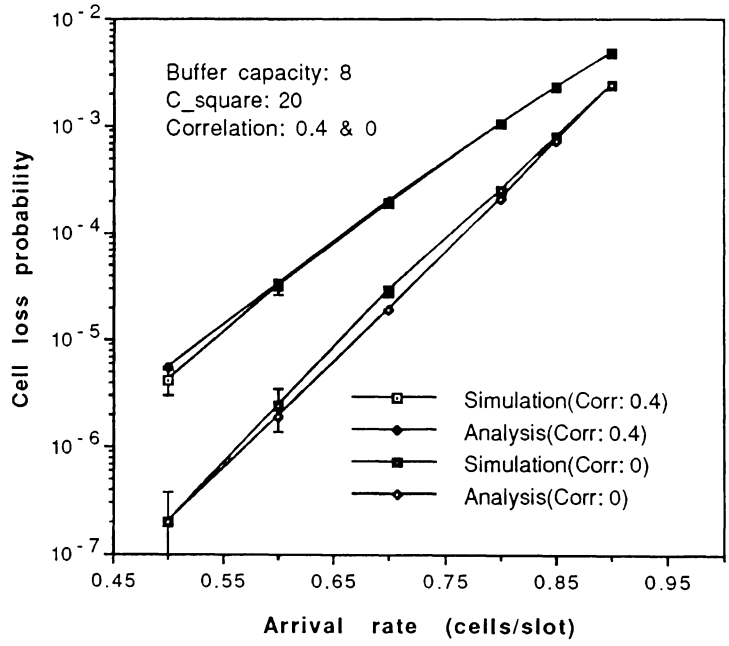


Figure 9: Cell loss probability incurred in a polling system with different arrival correlations when the queue capacity = 8.

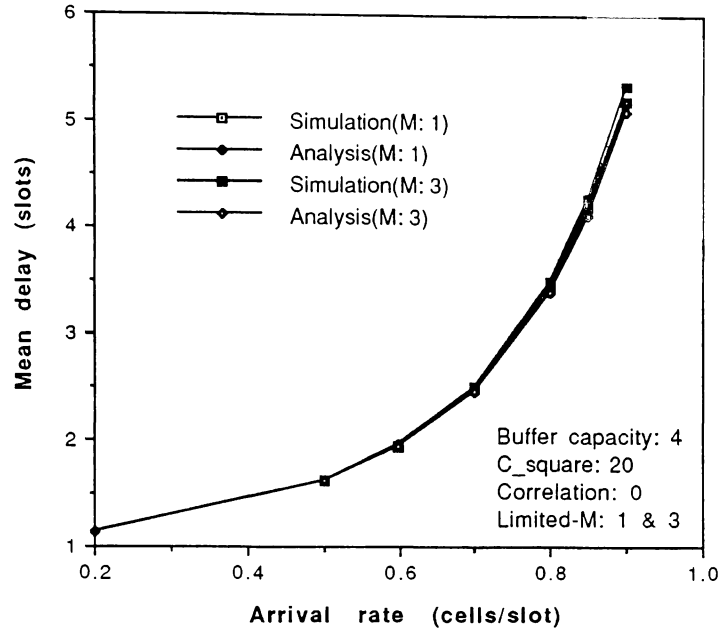


Figure 10: Mean delays incurred in the polling system when queue capacity = 4, and service up to 1 and 3 cells, respectively.

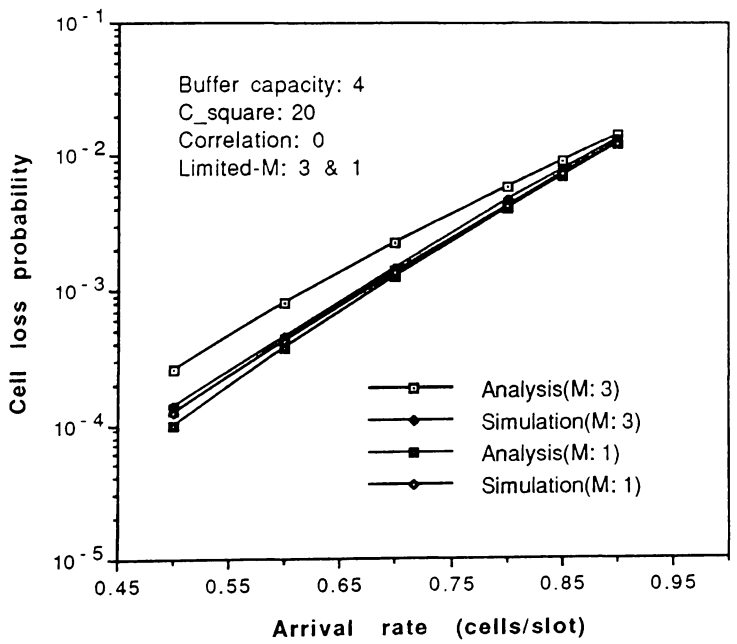


Figure 11: Cell loss probability incurred in the polling system when queue capacity = 4, and service up to 1 and 3 cells, respectively.