Simple Discrete Time Models for Performance Parameters of Multiplexer with Homogeneous ON-OFF Sources *

Khalid Begain[1], László Jereb[2], Miklós Telek[2], Tien Van Do[2]

[1] Department of Computer Science
Mu'tah University, 61710 Mu'tah, Jordan

[2] Department of Telecommunications
Technical University of Budapest, 1521 Budapest, Hungary

Abstract

The paper addresses the analysis of a single multiplexing node in ATM networks. It presents analytical models for evaluating the performance parameters of a multiplexer that has N identical ON-OFF type input sources and an output channel with finite buffer. The channel speed is assumed to be an integer times of the source speed in ON state. A two dimensional Discrete Time Markov Chain is introduced where the two dimensions describe the number of ON sources and the number of cells in the finite buffer at a given time. Two time scales are defined in order to ensure accurate results in calculating the performance parameters, e.g. cell loss and cell delay. Three alternative models of the cell arrival process are discussed and the performance parameters are derived.

Key words: Discrete-Time Markov Chain Models, ATM Multiplexer, Buffer Dimensioning, Performance Evaluation.

1 Introduction

Broadband ISDN (B-ISDN) is the network planned to carry different types of information including voice, video, and data. The CCITT has adopted the Asynchronous Transfer Mode (ATM) as the switching technique for the future high speed network due to its flexible and effective utilization of network resources. Since then ATM has become an intensive research area and the main interest has been devoted to the development of methods in order to ensure Quality of Service requirements (throughput, cell loss, delay, etc) for each data type.

*This work was performed with the support of the European Community through the contract COPERNICUS-1463.

The ATM is a packet-like switching and multiplexing technique in which messages are split into short fixed-length (53 Bytes) packets called cells. Cells may be lost or may suffer delay for different reasons, while they are transmitted from the source to the destination. The buffer overflow in an intermediate switching or multiplexing node can be one of the reasons of the loss or delay. The tolerance for cell loss or delay varies with the type of carried traffic. For example, packetized voice traffic allows relatively high cell loss probability but it has little tolerance to the delay while data can tolerate some delay but they are very sensitive to the cell loss.

In this paper, the problem of multiplexing is addressed. Namely, the special case of N identical ON-OFF sources with one high speed output. This problem has been studied in many papers providing both analytical and simulation results, however, most of them assume a continuous time or fluid flow model of the system which is only an approximation of the real situation.

Anick et al [1] considered the general data handling problem assuming continuous time model with exponential distributions for both the ON and the OFF intervals. The time unit was taken as the average of the ON intervals. The information unit was taken as the incoming information per time unit. In their model the server capacity was a given (not integer) value and the buffer size was infinite. The equilibrium buffer content distribution and its moments were derived for the model. Numerical results for the overflow probability of a predefined buffer backlog were presented as well.

Kosten [7] studied a similar model assuming a finite number of different groups of ON-OFF sources, but the time unit was not defined. Similar achievements were presented (eigenvalues-eigenvectors, buffer content distribution, …) and numerical results for two groups of sources were given.

Daigle and Landford [3] used the same model of [1] to study the problem of packet voice communication system. While Halfin [4] modified the previous model and allowed finite and infinite number of ON-OFF sources and a state-dependent Poisson message arrival with general packet length distribution. The results were provided with Laplace-Stieljes transforms for the equilibrium buffer content distribution, from which the moments and the delay characteristics were also derived. Some computational experiences and comparisons in case of Poisson input processes were summarized as well.

Mitra [10] used a similar model for production machines with service and failed states.

In addition to the traditional results implemented on these kinds of problems, he gave a detailed study on stable and unstable systems. Tucker [13] also used a model similar to [1], but the server capacity was defined as an integer number of information units and the buffer size was finite. Results on buffer content distribution, cell loss, and delay were given with simulation comparisons.

Li [9] assumed a finite number of ON-OFF sources, with exponential distributions for both the ON and the OFF intervals, a finite buffer size and a non-integer channel (server) capacity. He defined the group of states where packets are blocked and derived the mean holding time and the initial distribution for this group. He also gave results concerning the mean duration of the overloaded periods. An embedded Markov chain at the end of the overloaded periods was introduced which describes the worst case for the arriving packet concerning the delay. The paper presented many simulation results and approximations for integrated voice and data examples.

Baiocchi et al. [2] extended the previous idea by defining an overloaded and underloaded intervals and fitted the two-level model parameters (e.g., cell generation rate) to the initial multi-level description. Results were obtained for the two-level model and system parameters with comparison with simulations.

The common feature of the above mentioned models is that all of them have assumed a continuous time distribution for the ON and OFF intervals. On the other hand, Li [8] introduced a discrete time model assuming finite number of ON-OFF sources and geometrical distributions for the ON and OFF intervals. He fixed that in one time unit only one ON and/or OFF source can change state. The channel capacity was assumed to be an integer number of sources and the buffer size could be either zero (burst switching-clipping case) or infinite (packet switching case). Similarly, he defined overloaded and underloaded intervals and derived the mean cumulative time spent by a given (underloaded or overloaded) process in a given state. For the case of burst switching, he determined the bit clipping rate (BCR) and the mean length of the overloaded (clipping) periods with numerical results. Meanwhile for the packet switching case, he defined an embedded Markov chain such as that in [9] that was described before.

Finally, Hubner and Tran-Gia [6] used similar model, but in their model the server capacity was given as a non-integer number of information units and the buffer size was finite. They defined three cases:

- fixed number of ON sources (Quasi-stationary analysis) for which steady-state probabilities and cell blocking probabilities were given,
- fixed number of ON-OFF sources for which approximations based on the quasi-stationary results were given,
- model for call admission control (CAC) where blocking probabilities were calculated.

In this paper, a discrete time model based on a finite number of sources and a finite size buffer is introduced from which results on cell loss, average buffer length, and delay are given. In Section 2, the main assumptions for this model are given, and the mathematical model is derived in a step-by-step manner. Section 3 defines three different model alternatives. Section 4 gives the steady-state solution from which the most important expressions on performance parameters are obtained. Section 5 shows the numerical example and, finally, some conclusions complete the paper.

2 Model Assumptions

Physical model

Consider a multiplexing node with the following features :

- N identical sources with two states (ON, OFF).
- Sources in the ON state generate cells with rate v_s, where the time unit is taken so that v_s=1 [cell/time unit] holds.
- Sources in the OFF state do not generate any cells.
- There is one output transmission link with the transmission rate $v_l = C$ [cell/time unit].
- If more cells arrive than the link capacity, the extra cells are stored in a buffer of length L.
- Cells arriving when the buffer is full are lost.

The system is studied in order to find analytical results on the expected cell loss, the cell delay, and the average buffer content.

Source process

Assume that the behavior of a source can be described by a discrete-time Markov chain (DTMC) with two states (ON-OFF). The distribution of the length of the ON periods is assumed to be of geometrical with parameter β, while the OFF periods are also geometrical with parameter α. The transition probabilities of the DTMC are :

$$Pr[OFF \to ON] = \alpha$$
$$Pr[ON \to OFF] = \beta \tag{1}$$

Let us define now ξ_n denoting the number of sources in ON state at time n. It is obvious that this process is also a DTMC with state space $\Omega = \{0, 1, ..., N\}$ and the state transition probabilities can be written as:

$$p_{ij} = \sum_{k=\max(i+j-N,0)}^{\min(i,j)} \begin{pmatrix} i \\ k \end{pmatrix} (1-\beta)^k \beta^{i-k} \begin{pmatrix} N-i \\ j-k \end{pmatrix} \alpha^{j-k} (1-\alpha)^{N-i-j+k} \tag{2}$$

This expression of the transition probabilities takes into account that the transition from state i to state j may occur if k out of the i ON sources ($0 \leq k \leq i$) stay in the ON state and $(j-k)$ other sources turn from the OFF to the ON state.

Let $\mathbf{P}^{(n)} = \{P_i^{(n)}\}$ ($P_i^{(n)} = Pr\{\xi_n = i\}$) denotes the state probability vector of process ξ_n at time n and $\mathbf{p} = \{p_i, i = 1, \ldots, N\}$ denotes the steady state probability vector of ξ_n.

Buffer content

The process describing the number of cells in the buffer plays an essential role in evaluating the performance parameters mentioned before. Let η_n denote this process with state space $\Phi = \{0, 1, \ldots, L\}$, where L is the size of the buffer. The state transition probabilities of η_n are dependent on the state of process ξ_n, therefore one should study the two processes together.

The global model

By these assumptions, we define the compound process (ξ_n, η_n) with the states (i, j), where $i = 0, 1, \ldots, N$ and $j = 0, 1, \ldots, L$ and the state transition probabilities as follows:

$$p_{i,j,u,v} = Pr[u \text{ ON source}, v \text{ cells in buffer at time } (n+1) | i, j \text{ at time } n]$$

and the state probability matrix can be written as $\mathbf{\Pi}^{(n)} = \{\Pi^{(n)}(i,j)\} = Pr\{\xi_n = i, \eta_n = j\}$ and the steady-state probability matrix is denoted by $\pi(i,j)$.

In order to determine these probabilities, it is necessary to state that since the channel speed (capacity) is assumed to be an integer number of cells per unit time, a micro slot can be also defined as the time necessary to transmit one cell on the output channel. Although, one can suppose more reasonable to choose the micro slot as the time unit, the time unit was chosen to ensure the Markov (memoryless) property for processes ξ_n and (ξ_n, η_n).

This fact and the fact that i cells are generated during a time unit when i sources are in the ON state imply that the state transition probabilities of the compound process (ξ_n, η_n) vary depending on the arrival process. In the paper, we study three different situations of the arrival process as the most informative and useful cases for determining the performance parameters.

3 Model Alternatives

Model 1.: Arrivals occur at the beginning of the time slot

In this case we assume that one cell arrives from every On source at the beginning of any time slot, so that the buffer content will be $\min(i+j, L)$ cells , where i is the number of ON sources and j is the number of cells in the buffer at the end of the previous time slot. Thus, the number of cells that will be found in the buffer at the end of the time slot can be written as:

$$\eta_{n+1} = \max(\min(j+i, L) - C, 0) \tag{3}$$

Using the above approach, the number of cells $c_{i,j}$ and the total delay of cells $d_{i,j}$ in state (i,j) can be expressed in the following form:

$$c_{i,j} = \max(i + j - L, 0) \tag{4}$$

$$d_{i,j} = \sum_{l=j}^{\min(i+j,L)-1} l, \tag{5}$$

where the delay is measured in the micro slot unit.

Model 2.: Late arrivals with delayed access [11]

The arrival process is considered to be continuous in time, but cells arriving during the time slot must wait until the beginning of the next slot. In this case, the number of cells that will be found in the buffer at the beginning of the next time slot is written as:

$$\eta_{n+1} = \min(i + \max(j - C, 0), L), \tag{6}$$

while the cell loss and the cell delay are obtained from the following expressions:

$$c_{i,j} = \max(i + \max(j - C, 0) - L, 0) \tag{7}$$

$$d_{i,j} = \sum_{l=0}^{\min(i, L-\max(j-C,0))-1} \frac{C}{2} + \max(j - C, 0) + l \tag{8}$$

Model 1 and 2 provide identical cell loss, while Model 2 shows longer cell delay. It seems to be clear that there are other cases resulting in lower values for both the cell loss and the cell delay. In the sequel, Model 3 is discussed, which can be considered as an optimistic case. Although Model 2 seems to be the worst case, neither the pessimum nuture of Model 2 nor the optimum nature of Model 3 have not been proved so far.

Model 3.: Cells arrive one-by-one in the micro slot starting when the buffer becomes empty, and the remaining cells (if any) arrive at the end of time slot

For state (i, j)

$$\eta_{n+1} = \min(\max(j + i - C, 0), L), \tag{9}$$

$$c_{i,j} = \max(i + j - C - L, 0), \tag{10}$$

and

$$d_{i,j} = \sum_{l=0}^{\min(i-\max(C-j,0), L-\max(j-C,0))-1} \max(j - C, 0) + l \tag{11}$$

hold, where $\max(C - j, 0)$ is the number of empty micro slots after all the cells being served when $C > j$ and $\max(j - C, 0)$ gives the number of cell remaining at the end of the slot. It is obvious, that only one of the above quantities can take positive value at the same time.

4 Performance Parameters

Taking into account the model alternatives used to describe the arrival procedure for process (ξ_n, η_n), it can be seen that, for any time instant n, (ξ_{n+1}, η_{n+1}) depends only on (ξ_n, η_n), which means it is a DTMC with transition probabilities $p_{i,j,u,v}$ defined as follows:

$$p_{i,j,u,v} = \begin{cases} p_{i,u} & \text{if } \eta_n = j \text{ and } \eta_{n+1} = v \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

where $p_{i,u}$ is the transition probability of process ξ_n and v is calculated based on the above model alternatives.

With these transition probabilities, the steady-state probabilities $\pi = \{\pi(i,j)\}$ of the compound process (ξ_n, η_n) can be obtained from the well-known DTMC equations [5]. Then, the main performance parameters for the system can be given as follows:

- **The average cell loss**

$$Cl = \frac{\sum_{i=0}^{N}\sum_{j=0}^{L} \pi(i,j) \cdot c_{i,j}}{\sum_{i=0}^{N} i \cdot p_i} \tag{13}$$

  where p_i denotes the steady state probability of state i of process ξ_n, and the denominator gives the average number of the arrived cells.

- **The average cell delay**

$$D = \frac{\sum_{i=0}^{N}\sum_{j=0}^{L} \pi(i,j) \cdot d_{i,j}}{\sum_{i=0}^{N}\sum_{j=0}^{L} (i - c_{i,j}) \cdot \pi(i,j)} \tag{14}$$

  where the denominator gives the average number of transmitted cells.

- **The average buffer content**

$$B = \sum_{j=0}^{L} j \cdot \left(\sum_{i=0}^{N} \pi(i,j) \right) \tag{15}$$

In the next section, we give a numerical example on the model presented in the paper for all models and provide diagrams on the performance parameters.

5 Numerical Example

The models are demonstrated on an ATM multiplexer with ON-OFF type voice input channels with mean talkspurt duration β^{-1}=352ms and mean silence duration α^{-1}=650ms [12]. The output channel is assumed to be T1 line with a rate of 1.536Mbps. Taking into account that each ON source generates data with rate 64kbps and each 47 Byte should be encapsulated into 53 Byte cell by AAL1, the source speed will be 72,17kbps and thus C=21. With these values the time unit is 5.875ms and the micro time slot is 279.76μs.

Figures 1-4 show the calculated results for the expected cell loss and average cell delay versus buffer size with N=30 (Figures 1), 2) and with N=40 (Figures 3,4) respectively, while Figures 5-6 show the same parameters versus the number of sources.

It can be observed that Model 1 and Model 2 provide identical results for the average cell loss while Model 3 gives lower values. For the average delay, the figures show the highest values while Model 3 gives the lowest possible among the investigated models. It can also be seen (Figures 5,6) that the difference between Model 1 and Model 3 becomes less as the number of sources increases. This is due to the fact that Model 3 will behave like Model 1 in the heavy load situations.

Figures 1 and 3 show that the increment of N affects strongly the cell loss and an acceptable cell loss value can be reached only with high buffer size, while Figures 2

and 4 show the increment of buffer size increases the average cell delay very slightly since the cell loss will be less as well.

For the case N = 40, the difference between the values obtained for Models 1,2 and 3 are relatively small which allows some design decisions to be considered about the size of buffer based on other aspects. On the other hand, Figure 2 shows that this difference is rather large in the case of N = 30, which means that more realistic cell arrival distributions should be considered.

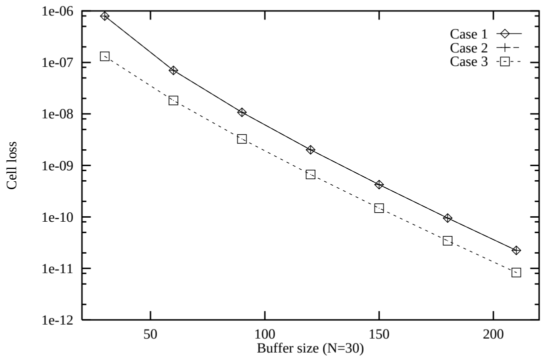

Figure 1: Cell loss probability

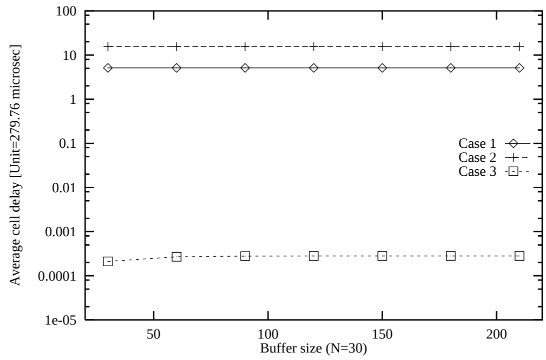

Figure 2: Average cell delay

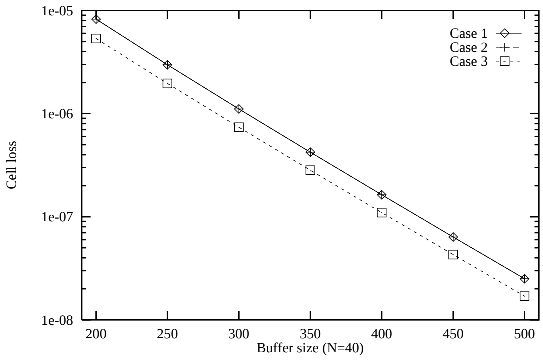

Figure 3: Cell loss probability

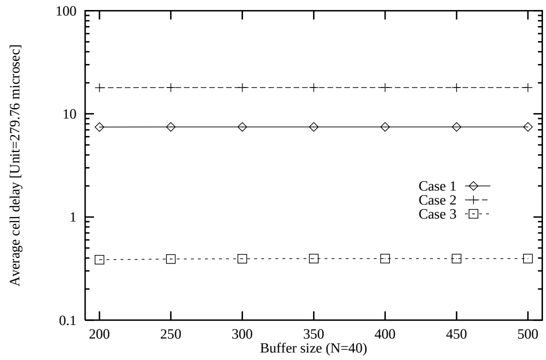

Figure 4: Average cell delay

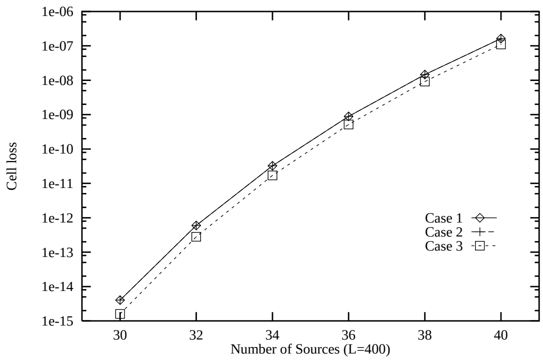

Figure 5: Cell loss probability

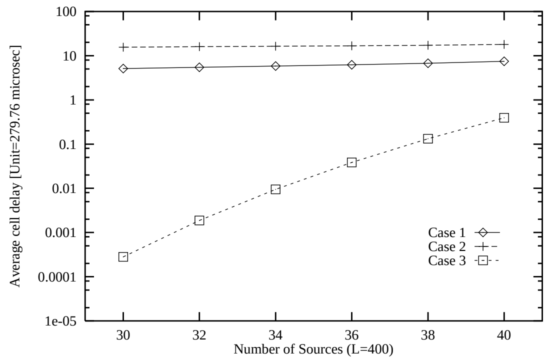

Figure 6: Average cell delay

6 Conclusions

The paper presents performance models of an ATM multiplexing node for which explicit expressions are given for the most important performance parameters; cell loss, cell delay and buffer content.

The models are based on the two-dimensional DTMC with the number of ON sources and the number of cells in the buffer.

The models are implemented on an IBM RISC-6000 Model-570 machine and they provide relatively short execution time even for higher N and L values (e.g., for $N = 40$ and $L = 500$ the execution time was 258 seconds).

In addition, the numerical example showed some interesting results on the three different model alternatives introduced in the paper giving some indications on the upper and lower limits for all the performance parameters mentioned before.

References

[1] D. Anick, D. Mitra, and M. M. Sondhi. Stochastic theory of a data handling systems with multiple resources. *The Bell Systems Technical Journal*, 61:1871–1894, 1981.

[2] A. Baiocchi, N. B. Melazzi, M. Listanti, A. Rovari, and R. Winkler. Loss performance analysis of an atm loaded with high speed on-off sources. *IEEE J. Selected Areas on Communications*, SAC-9:388–393, 1991.

[3] J. N Diagle and J. D. Landford. Models for the analysis of packet voice communication systems. *IEEE J. Selected Areas on Communications*, SAC-4(6):847–855, 1986.

[4] S. Halfin. The backlog of data in buffers with variable input and output rates. In H. Rudin and W. Bux, editors, *Performance of Computer-Communication Systems*, pages 307–319, Amsterdam, 1984. Elsevier.

[5] R.A. Howard. *Dynamic Probabilistic Systems, Volume II: Semi-Markov and Decision Processes.* John Wiley and Sons, New York, 1971.

[6] F. Hubner and P. Tran-Gia. Performance of the finite capacity asynchronous multiplexer with modulated input. *Telecommunication Systems*, 1:263–278, 1993.

[7] L. Kosten. Theory of data-handling with groups of multiple resourses. In H. Rudin and W. Bux, editors, *Performance of Computer-Communication Systems*, pages 321–331, Amsterdam, 1984. Elsevier.

[8] S. Q. Li. A new performance measurement for voice transmission in burst and packet switching. *IEEE Transactions on Communications*, COM-35:1083–1094, Nov 1987.

[9] S. Q. Li. Stude on information loss in packet voice systems. *IEEE Transactions on Communications*, COM-37:1193–1202, Nov 1989.

[10] D. Mitra. Theory of a fluid model of producers and consumers coupled by a buffer. *Advanced Applied Probability*, 20:646–676, 1988.

[11] G. Pujolle and H. G. Perros. Queueing systems for modeling ATM networks. In T. Hasegawa, H. Takagi, and Y. Takahashi, editors, *Performance of Distributed Systems and Integrated Communication Networks*, pages 301–322. Elsevier Science Publishers (North-Holland), 1992.

[12] K. Sriram and W. Whitt. Characterizing superposition arrival processes in packet multiplexers for voice and data. *IEEE J. Selected Areas on Communications*, SAC-4(6):833–846, 1986.

[13] R. C. Tucker. Accurate method for analysis of packet speech multiplexer with limited delay. *IEEE Transactions on Communications*, COM-36:479–483, April 1988.