2

# Using Maximum Entropy Principle for Output Burst Characterization of an ATM Switch

T. Srinivasa Roa, Sanjay K. Bose, K.R. Srivathsan

*E-Mail: tsr, skb, krsr@iitk.ernet.in Dept. of Elec. Engg.I.I.T. Kanpur - 208 01 INDIA*
*ph.: +91 512 250697*

**Abstract**

Maximum Entropy Principle is used in deriving an approximate expression for the burst length of a tagged call at the output of an ATM switch. The statistical multiplexer is approximated as a variable server, infinite buffer queuing system with only cells from the tagged call as clients where each incoming cell sees the server in randomly variable vacations. Numerical experiments are carried out and compared with the simulation results.

**Keywords**

Statistical Multiplexer, ON-OFF source, Instantaneous bandwidth available, Maximum Entropy principle

## 1 INTRODUCTION

Asynchronous Transfer Mode (ATM) is expected to be the carrier mode for Broadband Integrated Services Data Networks (B-ISDN). In B-ISDN, different calls will have different call characteristics, like peak rate, average rate, etc. Also different calls will have different QOS requirements, like packet loss, packet delay, etc. The optical fibre communication, perceived to be a suitable media for B-ISDN applications, provides Bit Error Rate (BER) as low as $10^{-9}$ - $10^{-10}$. Hence ATM which provides cell-based connection oriented network service, is an ideal transport for B-ISDN services on low error fibre optic media. Connection-oriented network service is preferred over connection-less network service because the former demands less processing overhead at intermediate switches than the latter.

Due to the "bursty" nature of B-ISDN applications, statistical multiplexing of calls is preferred for its effective utilization of bandwidth and buffer resources. Statistical multiplexing, however causes degradation of QOS parameters like average and standard deviation of cell delay and cell loss due to congestion at intermediate ATM switches. Reactive controls, like end-to-end flow control, are commonly used in low speed networks like X.25. In the ATM environment, reactive congestion controls may not be effective because of the large Bandwidth-distance product. Preventive congestion controls like Call Admission Controls (CAC) and User Parameter Control (UPC) are proposed for avoiding congestion in ATM networks. With Call Admission Controls in place, each intermediate switch in the pre-determined path of the call, is required to determine whether the incomming call can be served with the demanded QOS parameters without effecting the QOS of existing calls. If the call can be accepted, the switch forwards the "call request" to the next switch; otherwise the switch sends "call reject" back to the source, in which case the source may hunt for another route for the call.

Most of the literature in performance modeling and evaluation of ATM networks deal with a single link or an isolated switching node. The end-to-end performance analyses of large-scale Broadband Integrated networks

D. D. Kouvatsos (ed.), *ATM Networks*

is essential not only for implementation of Call Admission Control procedures but also for understanding the efficiency and financial viability of the network as a whole. As in any interconnected network, the output of the upstream node will be the input for the next node and hence knowledge of the output characterstics of ATM switch is essential. The exact characterization of the output process of an ATM switch is complex and intractable due to statistical multiplexing of various classes of multimedia traffic. Moreover, some intermediate nodes may be fed by output streams of more than one upstream node.

Most approaches for characterizing the output processes, proposed in the literature are approximations and have only limited applicablity in call admission controls and end-to-end performance analyses. Y. Ohba, et al [ohba 91] consider an ATM switch in the presence of three kinds of traffic, GI-stream, Batch arrivals and a set of IPP sources. A transient expression for the queue length distribution at the arrival instants of cells from the GI-stream, is developed. Using that queue length distribution, the waiting time distribution and inter-departure time distributions of cells from a GI-stream are obtained. Even though, in principle, the same transient expression can be used iteratively for obtaining steady state queue length distribution, it may not be practical for larger systems. I. Stavrakakis [stav 91] developed models for bursty traffic when they undergo splitting and merging. Specifically, three different models were proposed and compared for bursty traffic when it is splitted and cells routed into the tagged direction with probability p, and diverted away from the tagged direction with probability (1-p). The merging of bursty traffic is characterised as another bursty process in terms of the probabilities of the queue being empty and the queue being not empty. This also analyzes the output processes at intermediate switches in a system of inter-connected switches, with the following assumptions – the input at any switch is only a fraction, p of the output from a previous switch. i.e. the cell will be sent to the targeted direction with probability p. In case of bursty traffic, the above assumption may not be valid.

In certain B-ISDN applications, jitter is one of the QOS parameters. Specifically for real-time applicaions like audio, jitter is required to be low, so that proper replay of audio is possible at the destination. W.Matragi, et al [mat 94-I, mat 94-II] modeled the jitter of a call at the output as the difference in queue lengths at the departure instants of consecutive cells. They considered the jitter process for a GI-stream of customers in the presence of a batch arrival process. In [mat 94-I], the Z-transform of the jitter of a GI-stream at the output of a single node is obtained. This is extended in [mat 94-II], for the estimation of end-to-end jitter incurred by a periodic traffic in an ATM network. In [rob 92, boy 92], the influence of jitter on peak rate enforcement and user parameter control algorithms is studied. Due to intermittent clumping of cells, user parameter control algorithms need to be more complex. I. Cidon, et al [cid 94], obtained analytical expressions for messages, maximum cell delay in a message and the number of cells in a message whose delay exceeded pre-specified time thresholds. The analytical expressions obtained here can be solved recursively.

In [wan 93], J.L. Wang, et al considered a two queue priority system, where real time traffic is given high priority over non-real time traffic. The probability distributions for inter-departure times of cells from each queue are obtained.

In order to overcome the difficulties in output characterization of ATM switch, almost all the call admission control procedures and performance analyses reported in the literature, assume "Node Decomposition". To use this to determine whether to accept a call, intermediate switches in the path use the call characteristics as they appear at source; this in effect assumes that the characteristics will not be effected by the upstream switches. However there has been little research, (except [lau 93]) in validating this assumption. In [lau 93], the authors attempted the problem of validation of "Nodal decomposition" approach through extensive simulations. Both homogeneous as well as heterogeneous ON-OFF sources are considered to study the input-to-output distortion in individual traffic source as a function of peak rate of each source and overall load factor. The authors also studied the cross-correlations amongs the output sources. This paper summarily reports two conditions under which nodal decomposition can be applied in network-wide performance modeling. These are – *1. If the peak access rate of each source does not exceed 5 % of the total link capacity, source distortion will be negligible. 2. Should no more than 10 % of the departing sources go to the same immediate downstream link, inter source cross-correlation will have negligible effects on the queuing performance of the downstream nodes.*

From the congestion control and call management points of view, *burstiness* is one of the important properties of traffic whose knowledge enables us in designing call admission control procedures with better utilization of

buffer and bandwidth resources. Friesen and Wong [frie 93] considered interconnection of user nodes which are fed by multiple traffic sources and switch nodes. In the presence of bursty traffic, they analyzed mean queue lengths, mean delays at every user node and switch node. It is observed that mean queue length and mean delay are larger at the user node than at the first switch node. Similarly, the average queue length and mean delay are larger at the first switch node than at the second switch. Smoothening or Burst reduction of bursty sources is claimed as the reason for this. The smoothening effect increases for higher load. S.Low and P.Varaiya [low 91, low 93] defined burstiness of traffic in terms of the buffer required at the server for the given service rate. Using a deterministic fluid flow model, they show that both fixed rate and leaky bucket servers are burst reducing.

We consider a queue with $N$ ON-OFF sources as input, served by a slotted channel. Given the characteristics of each call at the input side, we obtain expressions for the density function of its burst length at the output side. The inter-departure time between cells of a call within a burst, and hence the length of a burst at the output side, depends not only on the instantaneous queue length, but also on the instantaneous states of all the calls at the input side. This is modeled using the Maximum Entropy principle. The queue length distribution can be obtained by approximating the multiplexed traffic at the input to the queue as a 2- state Markov Modulated Poisson Process (MMPP) [hef 86].

The problem attempted in this paper is different from the earlier literature [mat 94-I, mat 94-II, ohba 91, stav 91, wan 93]. W. Matragi, et al [mat 94-I, mat 94-II] and Ohba, et al [ohba 91] considered only GI- stream in the presence of batch traffic. I. Stavrakakis [stav 91] and J.L. Wang, et al [wan 93] considered only combined output characteristics. The output characterization of individual ON-OFF sources is considered important for obvious applications in telephone, data networks, etc. Also to the best knowledge of the authors, usage of the Maximum Entropy principle for estimation of the service time density function is new.

In this paper, we analyze the distribution of burst length of the tagged ON-OFF source at the output of a multiplexer with infinite buffer. The input to the multiplexer is a set of heterogeneous or homogeneous ON-OFF sources. Section 2 presents the model as an infinite buffered queue fed by arrivals from a number of ON-OFF sources. The effect of other sources on the output characteristics of the tagged call is twofold. The inter-cell departure time of two successive cells within a burst of the tagged call depends on the number of sources that are in ON state at that instant. Section 3 introduces the notion of instananeous bandwidth available to the tagged call which models the number of sources that are in ON state at that instant. We also present in this section the usage of Maximum Entropy principle to estimate the density function of the instantaneous bandwidth. The inter-cell departure time of successive cells within a burst of the tagged call also depends on the queue length distribution which in turn depends on the state of other sources. In Section 4, a modified queue model with variable server is presented. The input to this queue is cells from the tagged call. The variable service time of the server is to model the instantaneous bandwidth available to the tagged call. Also the server is assumed to go on vacation at the beginning of ON state which will model the dependence in the queue length distribution. In Section 5, density function of the output burst length is analysed. Some numerical examples are presented in Section 6, and compared with simulation results. Section 7 gives the concluding remarks.

## 2 MODEL DESCRIPTION

We consider an ATM statistical multiplexer with an infinite buffer serving $N$ ON-OFF sources each generating cells of constant size. The multiplexer is served by a single channel with capacity $C$ bits/sec. The channel is slotted with slot size equal to the service service time of a cell. This multiplexer buffer can be modeled as a discrete-time single server system.

Each ON-FF source alternates between ON and OFF states. During the ON state, source $i$ ($i = 1, .., N$) generates traffic at a constant rate $R_i$ bits/sec. Without loss of generality, we consider the size of a cell to be 53 bytes (ATM standard). Each source is modeled as a discrete source such that it can be described completely at time instants $\tau_0, \tau_1, ..., \tau_{j-1}, \tau_j, \tau_{j+1}, ...$, where $a_i \equiv \tau_{n-1} - \tau_n = 53 \times 8/R_i$ sec., for all n. At an arbitrary instant $\tau_n$,

if the source $i$ is in ON state, the source will continue to be in the ON state at time instant $\tau_{n+1}$ with probability $\alpha_i$ and with probability $(1 - \alpha_i)$, the source will switch to OFF state at $\tau_{n+1}$. Similarly if the source is in OFF state at the instant $\tau_n$, it will continue to be in OFF state at the instant $\tau_{n+1}$ with probability $\beta_i$ and switch to ON state with probability $(1 - \beta_i)$. The source will emit a cell of size 53 bytes at the time instant $\tau_n$, if it is in the ON state at that instant. Let $\theta^i_{\text{ON}}$ be the average number of cells emitted by source $i$ during an ON period and $\theta^i_{\text{OFF}}$ be the average length of OFF period in units of cell times. Then we get

$$\theta^i_{\text{ON}} = \frac{1}{1-\alpha_i}, \qquad \theta^i_{\text{OFF}} = \frac{1}{1-\beta_i}$$

So the average traffic load of source $i$, $R^i_{avg}$ is given by,

$$R^i_{avg} = \frac{\theta^i_{\text{ON}}}{\theta^i_{\text{ON}} + \theta^i_{\text{OFF}}}$$

Consider now, the intercell-departure time for cells belonging to the same ON period of source $i$. This intercell-departure time depends not only on the number of cells belonging to other sources, served in between two cells of source $i$ but also on the queue length at the departure time of the first cell of the tagged cell pair. The number of cells belonging to other sources, that are served in between the tagged cell pair, is a random variable and depends on the number of other sources that are in the ON state at that instant. Using this, the statistical multiplexer can be approximated as an infinite buffered queue (with only the tagged ON-OFF source $i$ as the input), which is being served by a server with a random service rate $u$; the server is also assumed to go on vacation before starting service to a cell. Thus the server with a variable service rate takes into account the fact that the effective instantaneous bandwidth available to the cells of the tagged source is variable and depends on the number of ON-OFF sources that are in ON state at that instant. The vacation period of the server is also a random variable and takes into account the fact that before commencement of service to a cell, the cells that are waiting in the multiplexer, need to be served.

# 3 INSTANTANEOUS BANDWIDTH

In the statistical multiplexer, we consider the service of cells belonging to the same ON state of source $i$. Specifically, between two cells of the same ON state of source $i$, depending on the states of other sources, cells belong to other sources will also get served. If the number of cells belonging to other sources present in between two cells belonging to the source $i$ is large enough so that the service time for all those cells is more than $a_i$, then the instantaneous inter-departure time between the cells of source $i$ is more than $a_i$ and is equal to the total service time of the cells that are queued in between those two cells. If this number is small enough so that the total service time for all those cells is smaller than $a_i$, then there are two possible cases:

- If the next cell of source $i$ has arrived before the departure of the previous cell, then the interdeparture time between the cells of source $i$ will be equal to the service time of cells belonging to the other sources.
- Otherwise, the inter-departure time between the cells of source $i$ will be equal to $a_i$.

The number of cells of other sources arriving between the cells of source $i$ depends on which sources are ON at that instant and their peak rates. Consider a specific situation when all $N$ sources are in the ON state. Here in between two cells belonging to source $i$, the average number of cells of other sources, that are queued up, can be calculated as follows:

The average number of cells belonging to other sources

$$= \frac{R_1}{R_i} + \frac{R_2}{R_i} + \ldots + \frac{R_{i-1}}{R_i} + \frac{R_{i+1}}{R_i} + \ldots + \frac{R_N}{R_i}$$
$$= \frac{1}{R_i} \sum_{j=1, j \neq i}^{N} R_j$$

The service time required to serve these cells

$$= \frac{1}{R_i} \left[ \sum_{j=1, j \neq i}^{N} R_j \right] \frac{53 \times 8}{C}$$

So, the average inter-departure time between the cells belonging to source $i$

$$= \frac{53 \times 8}{C} \left[ 1 + \frac{1}{R_i} \sum_{j=1, j \neq i}^{N} R_j \right]$$

$$= \frac{53 \times 8}{u_i}$$

where $u_i = \frac{C.R_i}{\sum_{j=1}^{N} R_j}$ is the instantaneous channel bandwidth of source $i$.

Consider another situation where only source $i$ is ON. Then no other cells will be queued in between two cells of source $i$. In this case, the instantaneous channel bandwidth of source $i$, will be $u_i = C$.

The instantaneous bandwidth $u_i$ available to source $i$ can be defined as the state of the system with respect to source $i$, and is a discrete random variable which can take upto $(2^{N-1})$ values. Analysis involving a discrete random variable with such a large state space may not be practical. When the rate $\frac{C}{R_j}$, for all $j$, is large enough, the instantaneous bandwidth, $u_i$ can be approximated to be a continuous random variable. If the distribution of $u_i$, is known, it means that the effect of all other sources on source $i$ has been characterized.

Approximating a discrete random variable as a continuous random variable involves obtaining density function of a continuous random variable with point probabilities as constraints. In principle, this can be formulated as a Maximum Entropy problem with the density function as the optimizing variable and the point probabilities of the discrete random variable as the constraints. Due to the large size of the constraint set, this problem is complex; we simplify this by considering only a fixed and small set of constraints.

## 3.1 Maximum Entropy Principle

Consider the instantaneous bandwidth available to source $i$ as the state of the system. Dropping the subscript $i$, we denote this as $u$. Assume that we know only its minimum, maximum and the average. Given this, the Maximum Entropy principle [shor 80, jay 57, wil 70, fer 70, kou 94], can be used to estimate the density function of $u$. For the last four decades, Maximum Entropy principle is being used in various engineering fields like Operation Research, Transportation, Queueing theory, etc for estimation of the state probability distribution in the absence of complete information about the state of the system. Of late Maximum Entropy principle found

applications in the area of ATM networks as well. Kouvatsos, et al [kou 94] has used Maximum Entropy principle for estimating the queue length distribution of the statistical multiplexer. In this paper, we use Maximum Entropy principle to estimate density function of the instantaneous bandwidth available when we know only its average. The Maximum Entropy principle will "choose" the density function such that the entropy is maximized with the given information as constraints. In other words, if $p(u)$ is the density function of $u$, we find $p(u)$ by maximizing

$$\text{Entropy, } H(u) = -\int_{u_1}^{u_2} p(u) \ln p(u) du \tag{1}$$

such that,

$$\int_{u_1}^{u_2} p(u) du = 1 \tag{2}$$

$$\int_{u_1}^{u_2} up(u) du = \overline{u} \tag{3}$$

where,
$u_1$ = Minimum value of that $u$ can attain
$u_2$ = Maximum value of that $u$ can attain
Here $u_2 = C$, channel capacity.
$\overline{u}$ = Average of $u$

Clearly, $p(u)$, that can be obtained from above, may not be true density function of $u$. Also the available information about $u$ may not be sufficient to obtain the actual density function of $u$. Maximum Entropy principle will estimate the density function which satisfies the given information, but mostly non-commital about whatever not known. Also if we re-estimate the density function with additional information, the so-obtained density function may be different from that obtained previously. Hence the density function obtained from this Maximum Entropy principle will be an approximation to the true density of the system state $u$.
The solution for the above set of equations is discussed in Appendix and is given by,

$$p(u) = e^{\lambda_1 - 1} e^{\lambda_2 u}$$

where $\lambda_1, \lambda_2$ can be obtained from,

$$e^{\lambda_1 - 1} \left( e^{\lambda_2 u_2} - e^{\lambda_2 u_1} \right) = \lambda_2 \tag{4}$$

$$\frac{u_2 e^{\lambda_2 u_2} - u_1 e^{\lambda_2 u_1}}{e^{\lambda_2 u_2} - e^{\lambda_2 u_1}} - \frac{1}{\lambda_2} = \overline{u} \tag{5}$$

It is argued in the Appendix that Eq. ( 5) has unique solution for $\lambda_2$. It can also be observed that for moderate to high load conditions ($0.4 \leq \rho \leq 0.99$), where the average instantaneous bandwidth $\overline{u} \leq \frac{u_1 + u_2}{2}$, the solution $\lambda_2$ is -ve. Since, as reported in the literature, at low load conditions, the input-output distortion is negligible, we consider here only the case, $\lambda_2 \leq 0$.
Now rewriting Eq. ( 5), we get

$$\frac{u_2 - u_1}{u_2 - \overline{u} - \frac{1}{\lambda_2}} = 1 - e^{\lambda_2 (u_2 - u_1)} \tag{6}$$

Since $\lambda_2$ is -ve,

$$e^{\lambda_2(u_2-u_1)} \leq 1 \tag{7}$$

Assuming that the term, $e^{\lambda_2(u_2-u_1)}$ in Eq. ( 6) is negligibly small, the solution for Eq. ( 6) is given by –

$$\lambda_2 \approx \frac{1}{u_1 - \overline{u}} \tag{8}$$

It is observed that the above assumption is valid with varying accuracies in many examples we considered. The ratio of channel capacity and peak rate of the call is one of the factors which effect the validity of the assumption. Although, exact condition for the validity is yet to be derived, an empirical condition can be arrived at by conditioning that $e^{\lambda_2(u_2-u_1)}$ is negligible.
For $e^{\lambda_2(u_2-u_1)}$ to be negligibly small,

$$|\lambda_2(u_2 - u_1)| \geq 10$$

But in this case, $\lambda_2$ is given by Eq. ( 8). Hence the empirical condition for the validity of the above assumption is given by –

$$\left|\frac{u_2 - u_1}{u_1 - \overline{u}}\right| \geq 10 \tag{9}$$

In case the above condition is not satisfied, Eq ( 5) can be solved numerically for $\lambda_2$. The following successive approximation algorithm is used to evaluate $\lambda_2$ iteratively with initial guess is given by Eq. ( 8). The main advantage of this algorithm is its insensitity to the initial guess. The neccessary condition for this algorithm to converge is given by –

$$e^{\lambda_2(u_2-u_1)} \leq \left(\frac{u_1 - \overline{u}}{u_2 - \overline{u}}\right)^2$$

which can be satisfied for all moderate to heavy load condistions.
So

$$u_1 = \frac{C.R_i}{\sum_{j=1}^{N} R_j}$$

Similarly maximum, $u_2$ of the state of the system is the maximum possible share of the channel bandwidth for source $i$ as it occurs when the instantaneous load is minimum possible, i.e. all the sources, except the tagged source $i$, are in OFF state.

Let $\nu = (x_1, x_2, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_N)$ be the combined state of all sources given that the source $i$ is in ON state, where

$$x_j = \begin{cases} 0 & \text{if the source } j \text{ is in OFF state} \\ 1 & \textit{Otherwise} \end{cases}$$

Also let $u_i(\nu)$ and $p(\nu)$ be instantaneous bandwidth available for source $i$ when the sources are in state $\nu$ and joint probability that the sources are in state $\nu$, respectively. Then the expected value of instantaneous bandwidth available, $\overline{u}$ for source $i$ when it is ON state can be obtained as –

$$\overline{u} = \sum u_i(\nu).p(\nu)$$

Since all the sources are independent of each other, we can write,

$$p(\nu) = \prod_{j=1, j\neq i}^{N} p(x_j)$$

where $p(x_j)$ is the probability that source $j$ is in ON state, if $x_j = 1$ or in OFF state if $x_j = 0$
Also it can be easily shown that

$$p(x_j = 1) = \frac{(1-\beta_j)}{(1-\alpha_j)+(1-\beta_j)}$$

and

$$p(x_j = 0) = \frac{(1-\alpha_j)}{(1-\alpha_j)+(1-\beta_j)}$$

# 4 INTER-CELL DEPARTURE TIME

We now consider the infinite buffer queue served by a single server with capacity, $u$ where $u$ is a random variable with density function $p(u)$. The customers to this queue are the cells belonging to source $i$. We also assume that the server will be on vacation at the time of arrival of each cell. The vacation period is a random variable, $v$ ($\geq 0$) . Let $b = \frac{53\times 8}{u}$ be the service time of a cell in this queueing system, with $f_b(b)$ and $B^*(s)$ as the density function and Laplace Transform of $b$ respectively.

The vacation period seen by an arriving cell of soure $i$ will be the time required to serve the cells that are waiting in queue at the arrival instant of this cell. The inter-cell departure time at the output of this queue is equivalent to the inter-departure time of cells belonging to source $i$, from this multiplexer.

We consider two cases. When the vacation period for the cell is so large that before the start of service of this cell, next cell of the same ON period (or burst) has arrived into the queue, then the instantaneous inter-cell departure time between the present cell and the next, is equal to the service time of the cell in the above queueing system. Let us define $d_1$ as inter-departure time given that new cell has arrived before the service of previous cell started. Then $d_1 = b$ and $f_{d_1}$ and $D_1^*(s)$ are the density function and Laplace Transform for $d_1$, respectively.

In the other situation, the vacation is small enough so that the service of the cell starts before the arrival of the next cell of the same burst. Let us define $d_2$ as the inter-departure time between the cells in this case. Then we get $d_2 = max(b, a)$, where $a$ is the interarrival time of cells of source $i$ within a burst (subscript $i$ removed for simplification).

Since $d_2$ is random variable, let us define $f_{d_2}(d)$ and $D_2^*(s)$ as the density function and Laplace Transform of $d_2$, respectively. This yields

$$f_{d_2}(d) = f_b(b)F_a(d) + F_b(d)f_a(d)$$

where, $F_b(.)$ is the distribution function of $b$ and $f_a(.)$ and $F_a(.)$ are the density and distribution functions of $a$, respectively.

Since $a$ is constant,

$$f_a(d) = \delta(d - a)$$

where $\delta$ is dirac delta function.

$$F_a(d) = \begin{cases} 1 & \text{if } d \geq a \\ 0 & otherwise \end{cases}$$

So rewriting,

$$f_{d_2} = f_b(d)F_a(d) + F_b(d)\delta(d - a)$$

Then

$$\begin{aligned} D_2^*(s) &= \int_0^\infty f_{d_2}(d)e^{-ds}dd \\ &= \int_0^\infty f_b(b)e^{-bs}db + F_b(a)e^{-as} \end{aligned}$$

## 4.1 Vacation Period And Queue Length Distribution

In the previous section, we considered the server with vacations, where the vacation period is equivalent to the service time required to serve all the cells ahead of tagged cell of the tagged source in the multiplexer buffer. The vacation period at any arbitrary time instant is the time required to serve a cell at the channel rate $C$ times the queue length at that instant. Therefore, the queue length distribution of the multiplexer will be needed to find the vacation period distribution.

The statistical multiplexing of $N$ ON-OFF sources can be approximated to be a 2-state MMPP as proposed in Heffes, et al [hef 86]. The queue length distribution of the $MMPP \mid D \mid 1$ infinite buffer queue may be obtained as proposed by Ramaswami [ram 80, ram 88] and Lucantoni [luc 91]. We define $q$ as the queue length of the multiplexer at the cell departure instants and $q(n)$ is the steady state queue length distribution.

Consider the probability that at the start of the service time of a cell of source $i$ in the multiplexer, the next cell of the same ON state is also waiting. Consider the instant when the service of a cell belonging to source $i$ has started. Let the queue length at that instant be given by $q'$, with distribution, $q'(n) = q(n - 1)$, for n = 1, 2, ...

Let $S$ be the set of all possible combined states of all sources and $R_\nu$ be the total arrival rate of cells into the multiplexer when the combined state is $\nu$, i.e

$$R_\nu = \sum_{j=1}^{N} x_j R_j, \qquad x_j \epsilon s$$

Between two cell arrivals of source $i$, the number of cells of other sources that can arrive is given by $n_\nu = aR_\nu$.

If the queue length $q'$ is greater than $n_\nu$ when service to a cell of source $i$ starts then another cell of the same ON state is also waiting. Let $p_\nu$ denote the probability of this event given that the combined state of all sources is $\nu$.

$$p_\nu = \sum_{n=n_\nu+1}^{\infty} q'(n)$$

Since $n_\nu$ is real number, it can be written as –

$$n_\nu = n_I + n_f = n_I(1 - n_f) + (n_I + 1)n_f$$

where $n_I$ and $n_f$ are integral and fractional part of $n_\nu$, respectively.
Then

$$p_\nu = (1 - n_f) \times q'(n_I + 1) + \sum_{n=n_I+2}^{\infty} q'(n)$$

Let $p$ denote the probability averaged over state $\nu$ that at the start of service of a cell belonging to source $i$, the next cell of the same ON state has also arrived. Then

$$p = \sum_{\nu \epsilon S} p_\nu p(\nu)$$

Now let us define $d$ as the inter-departure time of cells of the same ON state of source $i$. Then $d$ is given by,

$$d = d_1 p + (1 - p)d_2$$

Let $f_d(d)$ and $D^*(s)$ be the density function and Laplace Transform of $d$, respectively where $D^*(s)$ is given by

$$D^*(s) = D_1^*(sp)D_2^*(s(1 - p))$$

# 5 BURST LENGTH AT THE OUTPUT

We define the burst length of source $i$ at the output of the multiplexer as the time difference between the start of service of first cell of the burst at the input to the departure of the last cell of that burst. Let $br_n$ denote the burst length at the output when there are $n$ cells in the corresponding burst at the input side. Assuming that within an ON state of tagged source, variations in both the instantaneous channel bandwidth as well as vacation periods are negligible, $br_n$ can be approximated as

$$br_n = (n - 1)d$$

if $Br_n^*(s)$ is the Laplace transform of $br_n$,

$$Br_n^*(s) = D_1^*((n - 1)ps)D_2^*((1 - p)(n - 1)s) \tag{10}$$

But the probability that there are $n$ cells in the burst at the input is $\alpha^{n-1}(1-\alpha)$, for $n = 1, 2, \ldots$. Let $br$ denote burst length at the output averaged over $n$ and $Br^*(s)$ is the Laplace transform of $br$; where

$$br = \sum_{n=1}^{\infty} \alpha^{n-1}(1-\alpha) br_n$$

Then

$$Br^*(s) = \prod_{n=1}^{\infty} Br_n^*(\alpha_{n-1}(1-\alpha)s) \tag{11}$$

## 5.1 Average of Burst Length

Differentiating Eq. ( 11) w.r.t. $s$,

$$Br^{*\prime}(s) = \sum_{j=1}^{\infty} Br_j^{*\prime}(\alpha^{j-1}(1-\alpha)s)\alpha^{j-1}(1-\alpha) \prod_{i=1, i\neq j}^{\infty} Br_i^*(\alpha^{i-1}(1-\alpha)s)$$

Substituting $s = 0$, we get

$$Br^{*\prime}(0) = (1-\alpha)\sum_{j=1}^{\infty} \alpha^{j-1} Br_j^{*\prime}(0) \tag{12}$$

Differentiating Eq. ( 10) w.r.t. $s$, and substituting $s = 0$, we get

$$Br_n^{*\prime}(0) = (n-1)pD_1^{*\prime}(0) + (1-p)(n-1)D_2^{*\prime}(0) \tag{13}$$

Substituting Eq. ( 13) in Eq. ( 12),

$$Br^{*\prime}(0) = (1-\alpha)\sum_{j=1}^{infty} \alpha^{j-1}(j-1)\left[pD_1^{*\prime}(0) + (1-p)D_2^{*\prime}(0)\right]$$

Now

$$D_1^{*\prime}(s) = B^{*\prime}(s) = -\int_0^{\infty} b f_b(b) e^{-bs} db$$

Substituting $s = 0$ in the above equation, we get

$$\begin{aligned} D_1^{*\prime}(0) &= -\int_0^{\infty} b f_b(b) db \\ &= -\int_{u_1}^{u_2} \frac{53 \times 8}{u} \frac{\lambda_2}{[e^{\lambda_2 u_2} - e^{\lambda_2 u_1}]} e^{\lambda_2 u} du \end{aligned}$$

Similarly,

$$D_2^{*\prime}(s) = -\int_a^{\infty} bf_b(b)e^{-bs}db - aF_b(a)e^{-as}$$
$$= -\int_{u_1}^{R_i} \frac{53\times 8}{u}\frac{\lambda_2}{[e^{\lambda_2 u_2} - e^{\lambda_2 u_1}]}e^{\lambda_2 u}du - aF_b(a)$$

The average of burst length, $\overline{br}$

$$= (\alpha - 1)\sum_{j=1}^{\infty}\alpha^{j-1}(j-1)\,[pD_1^{*\prime}(0) + (1-p)D_2^{*\prime}(0)] \tag{14}$$

where,

$$D_1^{*\prime}(0) = -\int_{u_1}^{u_2} \frac{53\times 8}{u}\frac{\lambda_2}{[e^{\lambda_2 u_2} - e^{\lambda_2 u_1}]}e^{\lambda_2 u du} \tag{15}$$

$$D_2^{*\prime}(0) = -\int_{u_1}^{R_i} \frac{53\times 8}{u}\frac{\lambda_2}{[e^{\lambda_2 u_2} - e^{\lambda_2 u_1}]}e^{\lambda_2 u}du - aF_b(a) \tag{16}$$

# 6 NUMERICAL RESULTS AND DISCUSSION

In this section, we discuss two set of numerical experiments that were made to gauge the accuracy of the expressions derived in previous sections. The average burst length at the output side of the multiplexer is calculated and compared with simulation results. In these two experiments, we consider channel bandwidth of 155 Mbits/sec and cell size of 53 bytes. In both the experiments, calls with same characteristics (i.e. homogeneous calls) were considered.

In the first experiment, each call is described by, Peak Rate, $R_i = 20$ MBits/sec., $\alpha_i = 0.95$ and Average/Peak Rate ratio = 0.4576. The experiment was conducted with 3 different load factors, where $\rho$ is defined as

$$\rho = \frac{C}{\sum_{i=1}^{N} R_i}$$

| No. of Calls | Load Factor $\rho$ | Burst Length at the output in sec. | | |
|---|---|---|---|---|
| | | Simulation | Calculated with QLDs from Sim. | Calculated with QLDs from appr. |
| 13 | 0.767 | 410 | 411 | 428 |
| 15 | 0.8856 | 428 | 438 | 456 |
| 16 | 0.944 | 434 | 454 | 468 |

In the second experiment, we consider each call with Peak Rate, $R_i$ = 10 MBits/sec., $\alpha_i$ = 0.95 and Average/Peak rate ratio = 0.4576.

| No. of Calls | Load Factor $\rho$ | Burst Length at the output in sec. | | |
|---|---|---|---|---|
| | | Simulation | Calculated with QLDs from Sim. | Calculated with QLDs from appr. |
| 25 | 0.738 | 812 | 828 | 848 |
| 27 | 0.797 | 814 | 858 | 883 |
| 30 | 0.8856 | 824 | 896 | 934 |

It can be observed from above tables that the percentage of error in the burst length calculated with Queue Length Distribution (QLD) obtained from 2- state MMPP approximation is more than in those calculated using the QLDs obtained from the simulations. This may be due to fact that the probabilities of higher order queue lengths are underestimated in the 2- state MMPP approximation. The QLDs calculated with approximations using higher order MMPP (MMPP with more than 2 states) may improve the percentage of error.

It is also observed that as the number of calls increases, the percentage of error in burst lengths also increases. This may be attributed to the loss of information about the higher order moments of instantaneous bandwidth available to the tagged source in the Maximum Entropy approximation. To be more clear, let us consider "occupied channel bandwidth" in the homogeneous case, which is the sum of the peak rates of those sources which are in ON state. The occupied channel bandwidth is a random variable which depends on another random variable, number of sources that are in ON state at that instant. The second moment of the occupied channel bandwidth depends on the second moment of the number of sources that are in ON state whose dependence on the total number of sources is second order polynomial. So any increase in the total number of sources, would cause the second moment of the occupied channel bandwidth to increase by a second order polynomial factor. Hence as the total number of sources increases, the difference between the exact second moment and the estimated second moment from Maximum Entropy principle with only first moment as the constraint, increases by a second order polynomial factor. Since the occupied channel bandwidth and instantaneous bandwidth available to the tagged call are closely related, same arguements holds good for instantaneous bandwidth as well. Hence as the number of sources increases, the loss of information about higher order moments is higher in the Maximum Entropy approach.

## 7 CONCLUSION

An approximate expression is derived for the burst length of a tagged call at the output of an ATM switch by approximating the statistical multiplexer as a single variable server infinite buffer queuing system with only cells from the tagged call as customers. Each incoming cell also sees the server in randomly variable vacation periods. The density function of the service rate of the server is approximated using Maximum Entropy Principle. Two numerical examples are presented to gauge the accuracy/inaccuracy of the approximation. Considering the fact that only first moment is used as constraint, the accuracy of the results is impressive. In the authors' opinion, the main contributions of the paper are 1) introduction of Maximum Entropy principle for the estimation service time density function and 2) modeling of the statistical multiplexer as a variable server queuing systems with server vacations. This approach can be extended further by including more constraints for better estimation of the density function of the instantaneous bandwidth available.

## APPENDIX 1 SOLUTION OF EQ. (1), EQ. (2) AND EQ. (3):

Using Langrangian principle,

$$F(p) = -\int_{u_1}^{u_2} p(u)\ln p(u)du + \lambda_1\left[\int_{u_1}^{u_2} p(u)du - 1\right] + \lambda_2\left[\int_{u_1}^{u_2} up(u)du - \overline{u}\right] \tag{17}$$

Where $\lambda_1, \lambda_2$ are Langrangian Coefficients.

Differentiate Eq. ( 17) with respect to $p(u)$ and equate it to 0.

$$\frac{dF}{dp} = -\int_{u_1}^{u_2}\left[\ln p(u)du + p(u).\frac{1}{p(u)}du\right] + \lambda_1\left[\int_{u_1}^{u_2} du\right] + \lambda_2\left[\int_{u_1}^{u_2} udu\right] = 0$$

Rewriting,

$$\int_{u_1}^{u_2}[-\ln p(u) - 1 + \lambda_1 + \lambda_2 u]\,du = 0$$

After simplification, We get

$$p(u) = e^{\lambda_1 - 1}.e^{\lambda_2 u} \tag{18}$$

Substituting Eq. ( 18) in Eq. ( 2),

$$\int_{u_1}^{u_2} e^{\lambda_1 - 1}.e^{\lambda_2 u}du = 1$$

We get,

$$e^{\lambda_1 - 1}\left(e^{\lambda_2 u_2} - e^{\lambda_2 u_1}\right) = \lambda_2 \tag{19}$$

Substituting Eq. ( 18) in Eq. ( 3),

$$\int_{u_1}^{u_2} ue^{\lambda_1 - 1}.e^{\lambda_2 u}du = \overline{u}$$

After simplification and substituting Eq. ( 19), we get,

$$\frac{u_2 e^{\lambda_2 u_2} - u_1 e^{\lambda_2 u_1}}{e^{\lambda_2 u_2} - e^{\lambda_2 u_1}} - \frac{1}{\lambda_2} = \overline{u}$$

So the probability density function estimated from Maximum Entropy principle is given by,

$$p(u) = e^{\lambda_1 - 1}e^{\lambda_2 u}$$

Where $\lambda_1, \lambda_2$ can be obtained from,

$$e^{\lambda_1 - 1}\left(e^{\lambda_2 u_2} - e^{\lambda_2 u_1}\right) = \lambda_2$$

$$\frac{u_2 e^{\lambda_2 u_2} - u_1 e^{\lambda_2 u_1}}{e^{\lambda_2 u_2} - e^{\lambda_2 u_1}} - \frac{1}{\lambda_2} = \overline{u} \tag{20}$$

**Evaluation of Lagrangian Coefficients**
Define,

$$f = \frac{u_2 e^{\lambda_2 u_2} - u_1 e^{\lambda_2 u_1}}{e^{\lambda_2 u_2} - e^{\lambda_2 u_1}} - \frac{1}{\lambda_2} \tag{21}$$

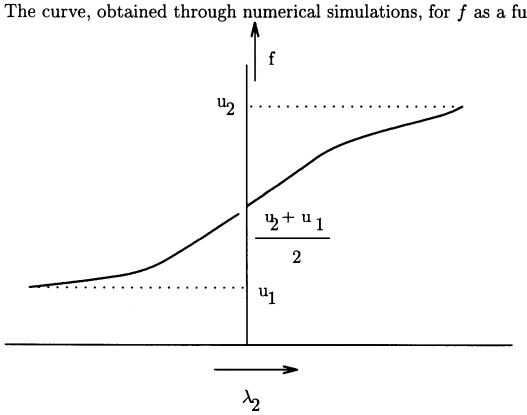
We can easily show that –

$$\lim_{\lambda_2 \to 0} f = \frac{u_1 + u_2}{2}$$

$$\lim_{\lambda_2 \to -\infty} f = u_1$$

$$\lim_{\lambda_2 \to +\infty} f = u_2$$

The curve, obtained through numerical simulations, for $f$ as a function of $\lambda_2$ is shown in the figure below.




From the above figure and the limits of $f$, we can conclude that Eq. ( 20) has unique solution $\lambda_2$.

## REFERENCES

[ohba 91] Y. Ohba. M. Murata, H. Miyahar, *Analysis of Interdeparture Processes for Bursty Traffic in ATM networks*, IEEE J-SAC, April 1991, Vol. 9, No. 3, P.No. 468-476.

[stav 91] I.Stavrakakis, *Efficient Modeling of Merging and Splitting processes in Large Networking Structures*, IEEE J-SAC, Oct 1991, Vol. 9, No. 8, P. No. 1336-1347.

[frie 93] V.J.Friesen, J.W.Wong, *The Effect of Multiplexing, Switching and Other Factors on the Performance of Broadband Networks*, Proc. of IEEE INFOCOM '93, P.No. 1194-1203.

[low 91] S. Low, P. Varaiya, *A Simple Theory of Traffic and Resource Allocation in ATM*, Proc. of IEEE GLOBECOM '91, P.No. 1633-1637.

[low 93] S. Low, P. Varaiya, *Burstiness Bounds for Some Burst Reducing Servers*, Proc. of IEEE INFOCOM '93, P.No. 1a.1.1-1a.1.8.

[mat 94-I] W. Matragi, C. Bisdikian, K. Sohraby, *Jitter Calculus in ATM Networks; Single Node Case*, Proc. of IEEE INFOCOM '94, P.No. 232-241.

[mat 94-II] W. Matragi, C. Bisdikian, K. Sohraby, *Jitter Calculus in ATM Networks; Multi Node Case*, Proc. of IEEE INFOCOM '94, P.No. 242-251.

[rob 92] J.Roberts, F. Guillemin, *Jitter in ATM Networks and its Impact on Peak Rate Enforcement*, Performance Evaluation, 1992. Vol. 16, P.No. 35-48.

[boy 92] P.E.Boyer, F.M.Guillemin, M.J. Servel, J.P.Coudreuse, *Spacing Cells Protects and Enhances Utilization of ATM Network Links*, IEEE Network, Sep 1992, P.No. 38-49.

[cid 94] I. Cidon, A. Khamisy, M. Sidi, *Dispersed Messages in Discrete-Time Queues: Delay, Jitter and threshold Crossing*, Proc. of IEEE INFOCOM '94, P.No. 218-223.

[wan 93] J.L. Wang, J.P. Zhou, C.Wang, Y.H. Fan, *Interdeparture processes of Traffic from ATM Networks*, Proc. of IEEE INFOCOM '93, P.No. 1337-1341.

[hef 86] H. Heffes, D.M. Lucantoni, *A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance*, IEEE J-SAC, Sep 1986, Vol. SAC-4, No. 6. P. No. 856-868.

[shor 80] J.E. Shore. R.W. Johnson, *Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy*, IEEE Trans. on Info. Theory, Jan 1980, Vol. IT-26, No. 1, P. No. 26-36.

[jay 57] E.T. Jaynes. *Information Theory and Statistical Mechanics*, Physical Review, May 1957. Vol. 106, No. 4, P.No. 620-630.

[wil 70] A.G.Wilson. *The Use of the Concept of Entropy in System Modeling*, Operational Res. Quarterly, Vol. 21, 1970, No. 2, P.No. 247-265.

[fer 70] A.E. Ferdinand, *A Statistical Mechanical Approach to System Analysis*, IBM J. Research, 1970, P.No 539-547.

[kou 94] D.D. Kouvatsos, N.M. Tabet-Aouel, S.G. Denazis, *ME-based Approxiamtions for Geneard Discrete-time Queueing Models*, Performance Eval., 1994, Vol. 21, P.No. 81-109.

[ram 80] V. Ramaswami, *The N/G/1 Queue and its Detailed Analysis*, Adv. Appl. Probability, 1980, Vol. 12, P.No. 222-261.

[ram 88] V. Ramaswami, *A Stable Recursion for the Steady State Vector in Markov Chains of M/G/1 Type*, Comm. Stat.- Stochastic Models, 1988, Vol. 4(1), P.No. 183-188.

[luc 91] D.M. Lucantoni, *New Results on the Single Server Queue with a Batch Markovian Arrival Process*, Comm. Stat.- Stochastic Models, 1991, Vol. 7(1), P.No. 1-46.

[lau 93] Wing-cheong Lau, San-qi Li, *Traffic Analysis in Large-Scale High-Speed Integrated Networks: Validation of Nodal Decomposition Approach*, Proceedings of IEEE INFOCOM '93, 1993.

## BIOGRAPHY

Bose, Sanjay K.: Prof. Bose did his Ph.D. from the State University of New York, Stobny Brook in 1980. He was with the Corporate RD of the General Electric Co. at Schenectady, N.Y during 1980-82. Since 1982 he has been on the faculty of the Indian Institute of Technology, Kanpur where he is currently a Professor in the

Department of Electrical Engineering. Prof. Bose has held visiting appointments at the University of Adelaide, Queensland University of Technology and the University of Pretoria. His research interests are in performance evaluation of computer and telecommunication networks. Prof. Bose is a memeber of Eta Kappa Nu, Sigma Xi and a Senior Member of IEEE.

Srivathsan K.R.: Prof. Srivathsan did his Ph.D. from Queen's University, Canada in 1981. He has been on the faculty of the Indian Institute of Technology, Kanpur since 1982 where he is currently a Professor in the Department of Electrical Engineering. He has been active in the area of Computer Networks and related applications and is one of the Coordinators of the ERNET project providing network facilities to academic and research institutions in India.