PERFORMANCE ANALYSIS OF FINITE CAPACITY QUEUES WITH COMPLEX BUFFER MANAGEMENT SCHEME FOR MULTIPLE CLASS TRAFFIC WITH QOS CONSTRAINTS

Shakeel Ahmad and Bashir Ahmad
Institute of Computing and Information Technology, Gomal University, Dera Ismail Khan (NWFP) Pakistan

ABSTRACT
Finite buffer queues under complex management schemes are of great importance towards effective congestion control and quality of service (QoS) provision in the telecommunication networks. Performance modelling of such queues has gained significant importance in order to satisfy various QoS constraints posed by several new mostly multimedia services. This paper analyses a single server GE/GE/1/**N** censored queue with R(R = 2) traffic classes under complete buffer partitioning scheme and first come first served (FCFS) service discipline. A finite capacity vector, **N**, represents the capacitiy of partitions for each class (N_1, N_2, … , N_R). The forms of the state probabilities, as well as basic performance measures such as blocking probability are analytically established at equilibrium via appropriate mean value constraints based on the principle of maximum entropy (ME). Typical numerical experiments are included to illustrate the credibility of the proposed analysis in the context of Generalised Exponential (GE)-type queues.

KEYWORDS
Complete partitioning queues, bursty traffic, performance modelling.

INTRODUCTION

A stable GE/GE/1/**N** censored queue with a single server, finite capacity and multiple class traffic under complex buffer management schemes is an important building block in the performance of computer systems and communication networks. The analysis of such queue is very difficult to tackle using the classical queueing theory. Traffic generated by the real time services such as voice over IP, video streaming, etc is generally very sensitive to the transmission. End-to-end delay and delay jitter are usually introduced due to random queueing in the network routers. Traditionally finite capacity queues with tail drop (TD) mechanisms have been employed in the network routers. Such queues temporarily accommodate the arriving packets when the server is busy. The arriving packets are dropped when the queue reaches its maximum capacity. Although this technique is simple, it suffers various problems, e.g., lock out, global synchronisation and full queue (Floyd 1993). The main problem among these is the full queue which causes longer delays and makes this technique an inappropriate choice for time sensitive applications.

In order to support such real time services along with traditional non-real-time services such as data transfer and emails, traditional queuemanagement schemes need replacing with sophisticated and effective mechanisms. The use of buffer partitioning for controlling congestion in communication buffers is an effective solution in current Internet routers. Under this partitioning scheme, packets from each class will occupy the pre-allocated spaces.The partitioning of the buffer can define a specific trade-off between packet delay and packet loss which can be adjusted to suit a particular type of service and its quality of service (QoS) requirements. This technique maintains a small size steady state queue, thus results in reduced packet loss, decreased endto-end delay and the avoidance of lock out behaviour thus using the network resources more efficiently.

Many queueing systems with priorities have been explored by (Cohen, 1969) and various applications of the analytical results of priority queues to data communication systems are surveyed by (Moraes, 1990). A stable infinite capacity G/G/1 queue with a single server and priority classes under either Preemptive-Resume (PR) or Head-of-the-Line (HOL) scheduling disciplines has been analysed in [Kouvatsos and Aouel, 1989) by applying the method of entropy maximisation (MEM). MEMhas also been used in (Kouvatsos, 1986) to study a stable single class G/G/1/N censored queue with a single server and First-Come-First-Served (FCFS) scheduling discipline. All these system use a

complete sharing of buffer which can cause starvation for lower priority class traffic.

This paper presents a new analytical solution for a finite capacity queueing system with complete buffer partitioning scheme (CBP), bursty external multiple traffic. Numerical results show the effectiveness of buffer partitioning scheme for various traffic loads.

The paper is organised as follows: The maximum entropy (ME) solution for a stable GE/GE/1/**N** censored queue with CBP scheme is characterised in Section 2. State probabilities are presented in Section 3. Numerical results involving Generalised Exponential (GE) interarrival and service time distributions, are included in Section 4. Section 5 includes conclusions.

PRELIMINARIES

The Principle of ME

The principle of ME (Jaynes, 1957) provides a self-consistent method of inference for characterising an unknown but true probability distribution, subject to known (or known to exist) mean value constraints. The ME solution can be expressed in terms of a normalising constant and a product of Lagrangian coefficients corresponding to the constraints. In an information theoretic context (Jaynes, 1957) the ME solution corresponds to the maximum disorder of system states and, thus, is considered to be the least biased distribution estimate of all solutions that satisfy the system's constraints.

In sampling terms, it has been shown (Jaynes, 1957) that, given the imposed constraints, the ME solution can be experimentally realised in overwhelmingly more ways than any other distribution. Major discrepancies between the ME distribution and the experimentally observed distribution indicate that important physical constraints have been overlooked. Conversely, experimental agreement with the ME solution represents evidence that the constraints of the system have been properly identified.

More details on entropy maximisation and its applications can be found in (Kouvatsos and Awan, 1998).

The GE Distribution

The GE distribution is an interevent-time distribution of the form (c.f., Fig. 1)

$$F(t) = P(W \le t) = 1 - \tau e^{-\sigma t}, \quad t \ge 0 \qquad (1)$$

$$\tau = 2/(C^2+1) \qquad (2)$$

$$\sigma = \tau\nu, \qquad (3)$$

where W is a mixed-time random variable (rv) of the interevent-time, whilst $(1/\nu, C^2)$ are the mean and squared coefficient of variation (SCV) of rv W. The GE distribution versatile, possessing pseudo memoryless properties which makes the solution of many GE-type queueing systems and networks analytically tractable (Kouvatsos and Awan, 1998).

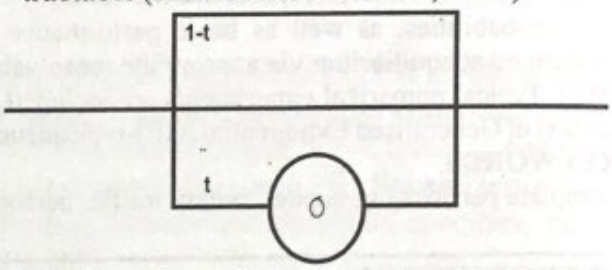


Fig.1. The GE Distribution with parameters τ and σ ($0 \le \tau \le 1$)

The choice of the GE distribution is further motivated by the fact that measurements of actual interarrival or service times may be generally limited and so only few parameters can be computed reliably. Typically, only the mean and variancemay be relied upon, and thus, a choice of a distribution which implies least bias (i.e., introduction of arbitrary and, therefore, false assumptions) is that of GE-type distribution. For example, in the context of ATM networks, this model is particularly applicable in cases of traffic with low level of correlation or where smoothing schemes are introduced at the adaptation level (e.g., for a stored video source) with the objective of minimising or even eliminating the problem of traffic correlation (Ball, 1996).Moreover, under renewality assumptions, the GE distribution is most appropriate to model simultaneous job arrivals at output port queues generated by different bursty sources (e.g., voice or high resolution video) with known first two moments. In this context, the burstiness of the arrival process is characterised by the SCV of the interarrival-time or, equivalently, the size of the incoming bulk.

The Complete Buffer Partitioning Scheme

Under a CBP management scheme, buffer is

divided into partitions equal to the number of traffic classes. This partitioning can be equal or based on the spefic requirements of each traffic class. packets of any class can join the finite capacity queue as long as there is space for their partition.

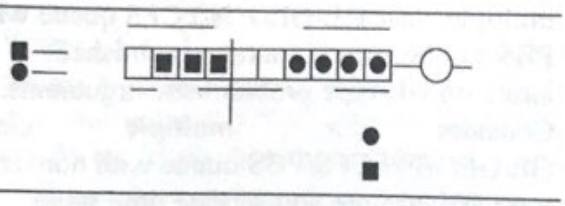


Fig.2. The CBP management scheme with traffic classes

ME ANALYSIS OF GE/GE/1/N QUEUE WITH COMPLETE BUFFER PARTITIONING SCHEME

This section presents the analysis of a single server GE/GE/1/**N** system to model a finite capacity queue with buffer partitioning. The analysis models the bursty external traffic with compound poisson process (CPP) and the transmission times of this traffic is represented by the GE distribution under FCFS service discipline. The total buffer capacity is N (N>2) and the vector **N** represents buffer partitioning $\{(N_1, N_2, \ldots, N_R)\}$ to give separate buffer to different classes of traffic in order to control the delay and delay jitter by reducing the queue length.

Notations

For each class i (i=1,2,…,R), let λ_i be the mean arrival rate, C^2_{ai} be the inter-arrival time SCV, μ_i be the mean service rate and C^2_{si} be the service time SCV.

Focusing on a stable GE/GE/1/**N**/FCFS queue, let at any given time n_i $n_i \le N_i$ be the number of class i packets in the queue (waiting and/or receiving service).

$\mathbf{S}=(n_1, n_2, \ldots, n_R)$ be a joint queue state,

$$\text{where} \sum_{i=1}^{R} n_i \le N$$

Q be the set of all feasible states **S**

$\mathbf{n}=(n_1, n_2, \ldots, n_R)$ be an aggregate joint queue state (n.b., $\mathbf{0}=0, \ldots, 0$)

Ω be the set of all feasible states **n.**

Remarks

The arrival process for each class i (i=1,2,…,R) is assumed to be censored, i.e., a packet of class i will be lost if on arrival it finds N_i (i=1,2,…,R) packets at the queue.

For exploration purposes, the analysis that follows focuses on the FCFS with CBP is applicable in the performance modelling of networks for the effective mechanism of traffic congestion control and also for providing various QoS demands by different multimedia services.

Prior Information

For each state **S**, $\mathbf{S} \in \mathbf{Q}$ and class i (i=1,2,…,R) the following auxiliary functions are defined:

$n_i(\mathbf{S})$ = the number of class i packets present in state **S**,

$$s_i(S) = \begin{cases} 1, & \text{if class } i \text{ packet is in service} \\ 0, & \text{otherwise,} \end{cases}$$

$$f_i(S) = \begin{cases} 1, & \text{if } \sum_{i=1}^{R} n_i(S) = N_i \;\&\; s_i(S) = 1 \\ 0, & \text{otherwise,} \end{cases}$$

Suppose what is known about the state probabilities $\{P(\mathbf{S})\}$ is that they satisfy the

- Normalisation constraint

$$\sum_{S \in Q} P(S) = 1, \qquad (4)$$

and that the following marginal mean value constraints per class I exist:

- Server utilisation, U_i, $(0<U_i<1)$,

$$\sum_{S \in Q} s_i(S) P(S) = U_i, i = 1,2,\ldots,R; \qquad (5)$$

Mean queue length, L_i $(U_i \le L_i \le N_i)$,

$$\sum_{S \in Q} n_i(S) P(S) = L_i, i = 1,2,\ldots,R; \qquad (6)$$

- Full buffer state probability, ϕ_i $(0<\phi_i<1)$,

$$\sum_{S \in Q} f_i(S) P(S) = \phi_i, i = 1,2,\ldots,R; \qquad (7)$$

satisfying the flow balance equations, namely

$$\lambda_i (1-\pi_i) = \mu_i U_i \quad i=1,2,3,\ldots\ldots R; \qquad (8)$$

where π_i is the blocking probability that an arriving packet of class i finds N_i (i=1,…,R) packets in the queue (waiting or receiving

service).

The choice of mean value constraints (4) - (7) is based on the type of constraints used for the ME analysis of stable multiple class queue without space priorities(Kouvatsos and Denazis , 1993). Note that if additional constraints are used, it is no longer feasible to capture a computationally efficient ME solution in closed-form. Conversely, the removal of one or more constraints from the set (4) - (7) will result into an ME solution of reduced accuracy.

A Universal Maximum Entropy Solution

A universal form of the state probability distribution P(**S**), **S** $\in$ **Q** can be characterised by maximising the entropy functional

$$H(P) = -\sum_{S} P(S) \log P(S), \qquad (9)$$

subject to constraints (4) - (7). By employing Lagrange's method of undetermined multipliers, the ME solution is expressed by

$$P(S) = \frac{1}{Z} \prod_{i=1}^{R} g_i^{s_i(S)} x_i^{n_i(S)} y_i^{f_i(S)}, \forall S \in Q; \quad (10)$$

Where Z, the normalising constant, is clearly given by

$$Z = \sum_{S \in Q} \left(\prod_{i=1}^{R} g_i^{s_i(S)} x_i^{n_i(S)} y_i^{f_i(S)} \right), \qquad (11)$$

and $\{g_i, x_i, y_i, i=1,2, \ldots ,R\}$ are the Lagrangian coefficients corresponding to constraints (5) - (7), respectively.

Remarks

Although constraints (5) - (7) are not known priori, nevertheless it is assumed that these constraints exist. This information, therefore, has been incorporated into the ME formalism (4) - (9) in order to characterise the form of the joint state probability (10).

Aggregating (10) over all feasible states **S** $\in$ **Q** and after some manipulation, the joint aggregate ME queue length distribution $\{P(\mathbf{n}), \mathbf{n} \in \Omega\}$ is given by:

$$P(0) = \frac{1}{Z} \qquad (12)$$

$$P(k) = \sum_{i=1}^{R} \Pr ob(Q_{i,k})$$

$$= \frac{1}{Z} \left(R \prod_{j=1} x_j^k \right) \sum_{j=1}^{R} k_j g_j y_j^{\delta(k)} \left(\frac{N_j!}{\prod_{i=1}^{R} (k_i - N_j)} \right) \qquad (13)$$

where, $\delta(k) = 1$, if $\sum_j k_j = N_i$ & $s_i(k) = 1$, or 0, otherwise; and N_i are the threshold values for each class j, j=1,2, … ,R.

Blocking Probability

A universal form for the marginal blocking probabilities $\{\pi_i, i=1,2, \ldots ,R\}$ of a stable multiple class GE/GE/1/**N**/FCFS queue with PBS can be approximately established, based on GE-type probabilistic arguments. Consider a multiple class GE/GE/1/**N**/FCFS/PBS queue with non-zero inter-arrival time and service time stage selection probabilities $\sigma_i = \left(C_{ai}^2 + 1\right)/2$ and $r_i = \left(C_{si}^2 + 1\right)/2$, respectively. Each arriving bulk of class i(i=1,2, … ,R) joins th e queue at Poisson arriving instants and finds the same aggregate number of packets as a random observer (n.b., this assumption is strictly true if the SCVs of the GE -type inter-arrival times per class are equal). Let us focus on a tagged packet within an arriving bulk of class i (i=1,2, … ,R) which finds the queue in state $\mathbf{n}_j = (0,\ldots,0, n_j, n_{j+1}, \ldots ,n_R)$ where $n_k = 0$, k = 1,2,…,j-1. Clearly, the total number of packets in the queue is $v = \sum_{k=j}^{R} n_k$ and the number of available buffer spaces is equal to N-v.

Using the probabilistic arguments, the blocking probabilities can be approximated as follows:

$$\pi_i = \sum_{k=0}^{N} \delta_i(k) (1-\sigma_i)^{[N-k]^+} P(k) \qquad (14)$$

where

$$\delta_i(k) = \begin{cases} \dfrac{r_i}{r_i(1-\sigma_i)+\sigma_i}, & k = 0 \\ 1, & \text{otherwise} \end{cases} \qquad (15)$$

Lagrangian Coefficients

The Lagrangian coefficients x_i and g_i can be approximated analytically by making asymptotic connections to an infinite capacity queues. Assuming x_i and g_i are invariant to the buffer capacity of size N, it can be established that

$$x_i = \frac{L_i - \rho_i}{L} \qquad (16)$$

$$g_i = \frac{(1-X)\rho_i}{(1-\rho)x_i} \qquad (17)$$

Where $X = \sum_{i=1}^{R} x_i, L = \sum_{i=1}^{R} L_i$ and L_i is the asymptotic marginal mean queue length of a multi-class GE/GE/1 queue (Kouvatsos and Awan, 1998). Note that statistics L_i, i=1,2, ...,R can be determined by (Kouvatsos and Denazis, 1993).

$$L_i = \frac{\rho_i}{2}\left(C_{ai}^2 + 1\right) + \frac{1}{2(1-\rho)}\sum_{j=1}^{R}\frac{\Lambda_i}{\Lambda_j}\rho_j^2\left(C_{ai}^2 + C_{sj}^2\right) \qquad (18)$$

where $\rho_i = \Lambda_i / \mu_i, \rho = \sum_{i=1}^{R} \rho_i$

By substituting the value of aggregate probabilities, $P_N(n)$, n=0,1, ... ,N, and blocking probabilities, π_i, i=1,2, ... ,R, into the flow balance condition (8), the Lagrangian coefficients{y_i, i=1,2, ... ,R} can be easily derived.

NUMERICAL RESULTS

This section presents numerical results which have been conducted using the proposed analytical model and simulation (based on QNAP-2 at 95% confidence interval (Veran, 1985) to evaluate the effectiveness of the buffer partitioning scheme for multiple traffic classes. In these experiments, Poisson distribution and Generalised Exponential (GE) - type traffic has been used which can be represented by the first two moments and exhibits the burstiness property of the traffic. These experiments use two heterogeneous video applications, representing a delay sensitive live video with 2.0 Mbps arrival rate and delay tolerant video streaming with 2.5 Mbps arrival rate. The service rate (or the link capacity) remains same as 4.0 Mbps.The service rate (or the link capacity) is 4.0 Mbps. For a total capacity of ten packets for each class, Figures 3 and 4 shows the server utilization for each class of traffic under various traffic loads, respectively. It can be observed from these figures that under a wide range of input data, the ME solutions are very comparable to those obtained by corresponding simulation models.

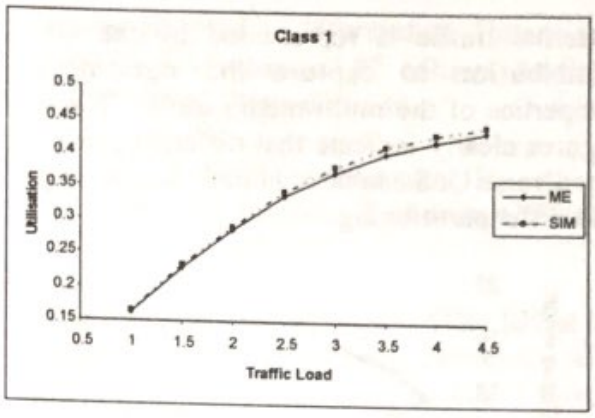


Fig.3. Marginal Utilizations for Class-1 traffic

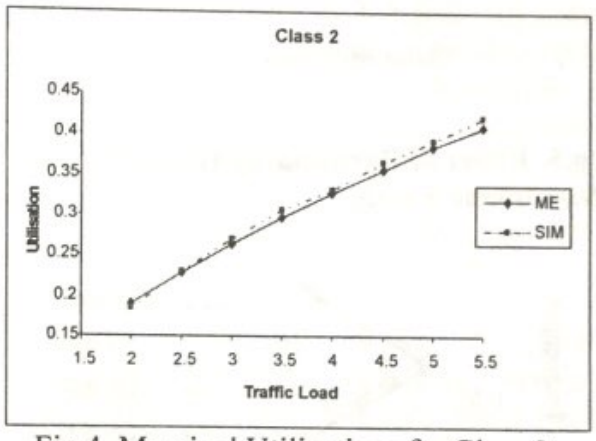


Fig.4. Marginal Utilizations for Class-2 traffic

Furthermore, Figures 5-7 show the effectiveness of using different partitions under CBP scheme in order to provide various grades of services to different classes of traffic when the arrival process for both streams is according to Poisson distribution. It is very much clear in all these figures that changing the portioning values generates significant impact on different performance measures including mean queue length, throughput and blocking probabilities per class. In Figure 5, increasing the partitioning for class1 reduces the space in the buffer for class 2 traffic. Consequently leading to high mean queue length for class1 traffic and lower mean queue length for class 2 traffic. This causes high blocking probability for class 2 as compared to class1 traffic. Similar is the effect of partitioning on the throughput which increase for class 1 by reducing the space for class 2.

Similar to above experiment, Figures 8-10 show the effect of buffer partitioning on various performance measures when the

external traffic is represented by the GE distribution to capture the burstiness properties of the multi-media traffic. These figures clearly indicate that different grades for diverse QoS can be obtained by adjusting the buffer partitioning.

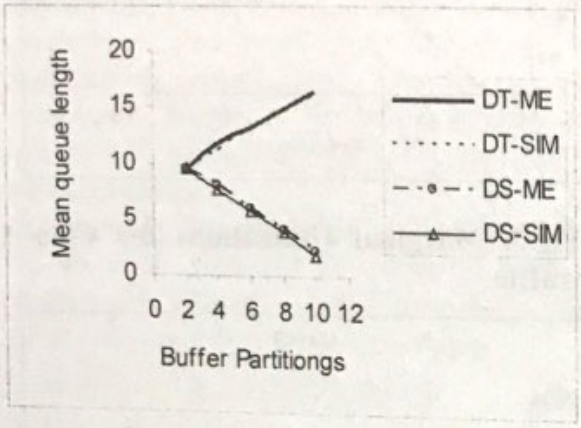


Fig.5. Effect of Partitioning Difference on Mean queue length

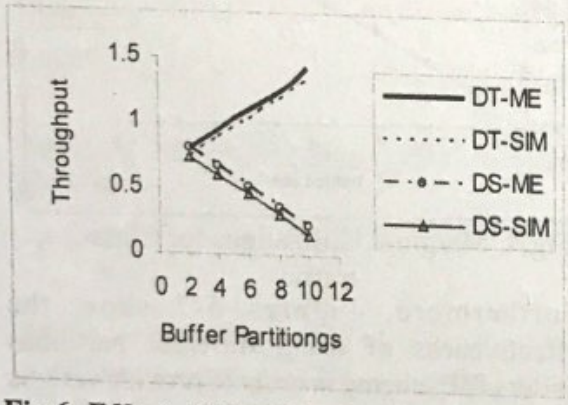


Fig.6. Effect of Partitioning Difference on Throughput

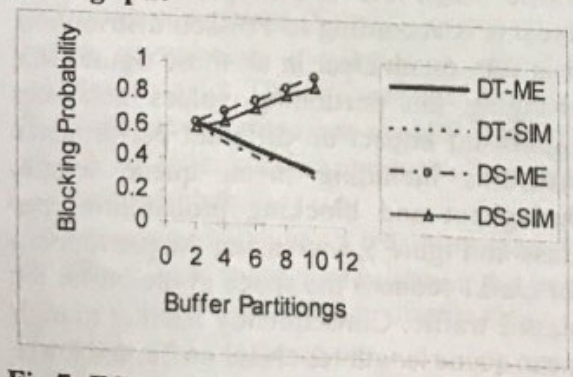
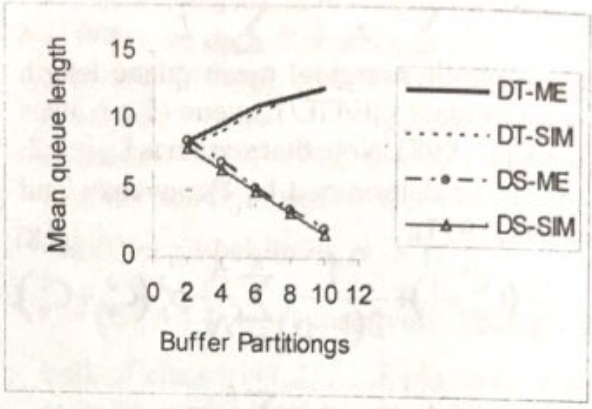


Fig.7. Effect of Partitioning Difference on blocking Probabilities



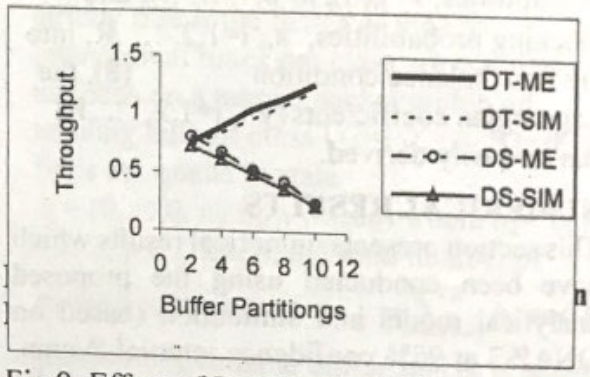


Fig.9. Effect of Partitioning Difference on Throughput

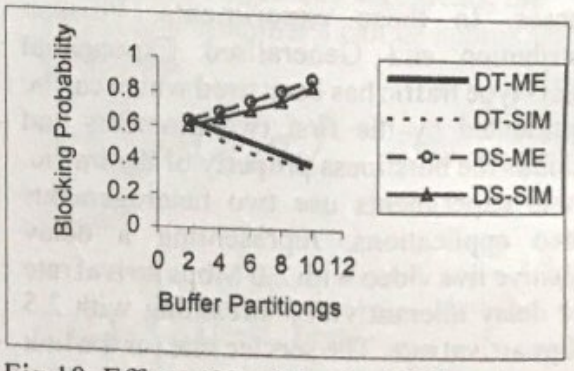


Fig.10. Effect of Partitioning Difference on blocking Probabilities

CONCLUSIONS

Analysis of a complete buffer partitioning based queue for brsty external traffic has been presented. This model can be used to reduce end-to-end delay for traffic generated by these applications. In this context, Product-form approximation, based on the principle of ME, for a stable GE/GE/1/**N** queue with FCFS scheduling discipline under CBP scheme has been proposed as a useful performance evaluation tool. This scheme effectively controls the allocation of buffer to various traffic classes according to their delay constraints. Closed form analytical

expressions for state probabilities have been derived. The proposed model has been implemented using the GE-type external traffic represent the bursty nature of the multimedia traffic. Typical numerical examples have been included to show the impact of buffer partitioning on various performance metrics for delay sensitive video streams and delay tolerant data packets.

REFERENCES

Floyd S and Jacobson V(1993). Random Early Detection Gateways for Congestion Avoidance, IEEE/ACM Transaction on Networking, 11(4): pp.397-413

Cohen JW (1969). The Single Server Queue. Revised edition, North-Holland Publishing Company, Amsterdam, Chap-3, pp. 113-119

Moraes DLFM (1990) .Priority Scheduling in Multiaccess Comm: Stochastic Analysis of Computer and Comm: Systems, H. Takagi (ed.), Elsevier Sc. Publishers (North-Holland), Amsterdam, Chap-7, pp. 699-732

Kouvatsos DD and Aouel TNM (1989). A Maximum Entropy Priority Approximation for a Stable G/G/1 Queue, Acta Informatica 27:247-286.

Kouvatsos DD (1986). Maximum Entropy and the G/G/1/N Queue, Acta Informatica, 23: pp. 545-565.

Jaynes ET (1957). Information Theory and Statistical Mechanics, 2nd edition, Addison Wisely. Chap11, pp. 620-630.

Jaynes ET(1957). Information Theory and Statistical Mechanics, II, Addison Wisely. Chap-6, pp.171-190.

Kouvatsos DD and Awan IU(1998). MEM for Arbitrary Closed Queueing Networks with RS-Blocking and Multiple Job Classes, Elsevier International Journal of Annals of Operations Research 79: pp.231-269.

Ball FD, Hutchinson, Kouvatsos VBR (1996) Video Traffic Smoothing at the AAL SAR Level, Proc. of 4th IFIP Workshop on Performance Modelling and Evaluation of ATMNetworks, D.D.

Kouvatsos (Ed.), Ilkley, (1996), pp. 28/1-28/10.

Kouvatsos DD and Denazis SG (1993). Entropy Maximised Queueing Networks with Blocking and Multiple Job Classes, Elsevier International Journal of Performance Evaluation, 17: 89-205.

Veran M and Potier D (1985). A Portable Environment for Queueing Network Modeling Techniques and Tools for Performance Analysis, D.Potier (ed.), North Holland, Chap-2, pp. 25-63