Available online at www.sciencedirect.com

ScienceDirect

Journal of Computer and System Sciences 74 (2008) 232–242

JOURNAL OF COMPUTER AND SYSTEM SCIENCES

www.elsevier.com/locate/jcss

Performance analysis of multimedia based web traffic with QoS constraints

Irfan Awan [a,*], Shakeel Ahmad [b], Bashir Ahmad [b]

[a] *Mobile Computing and Networks Research Group, Department of Computing, University of Bradford, Bradford BD7 1DP, UK*
[b] *Institute of Computing and Information Technology, Gomal University, D.I. Khan, NWFP, Pakistan*

Received 15 June 2006; received in revised form 31 October 2006

Available online 24 April 2007

Abstract

During the recent years, there has been a tremendous growth in the development and deployment of multimedia based networked applications such as video streaming, IP telephony, interactive games, among others. These applications, in contrast to elastic applications such as email and data sharing, are delay and delay jitter sensitive but can tolerate certain level of packet loss. A vital element of end-to-end delay and delay jitter is the random queueing delays in network switches and routers. Analysis of robust mechanisms for buffer management at network routers needs to be carried out in order to reduce end-to-end delay for traffic generated by multimedia applications. In this context, a threshold based buffer management scheme for accommodating multiple class multimedia traffic in network routers has been analysed. This technique effectively controls the allocation of buffer to various traffic classes according to their delay constraints. The forms of the joint state probabilities, as well as basic performance measures such as blocking probabilities are analytically established at equilibrium. Typical numerical experiments are included to illustrate the credibility of the proposed mechanism in the context of different quality of service (QoS) grades for various network traffic classes. This model, therefore, can be used as a powerful tool to provide a required grade of service to a particular class of multimedia based web traffic in any heterogeneous network.

Keywords: Multimedia applications; Queue thresholds; Performance analysis

1. Introduction

During the recent years real-time applications such as video streaming, interactive games and voice over IP have become increasingly popular among computer users. These applications are generally delay sensitive and need preferential treatment in order to satisfy a desired level of Quality of Service (QoS) constraints. Many enterprises demand applications development using software which integrates the support for real-time applications with the support for conventional computations. Such demands have posed various challenges to network community for robust design of the communication infrastructure. In that, significant developments have been made to design networks with the ability to guarantee the QoS for the real time data [1].

[*] Corresponding author.
E-mail addresses: i.u.awan@bradford.ac.uk (I. Awan), shakeel_1965@yahoo.com (S. Ahmad), bashahmad@gmail.com (B. Ahmad).

0022-0000/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2007.04.004

Traffic generated by the multimedia applications 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 [2]. The main problem among these is the full queue which causes longer delays and makes this technique an inappropriate choice for time sensitive applications.

To support these multimedia applications along with traditional non-real-time services such as data transfer and emails, traditional queue management schemes need replacing with sophisticated and effective mechanisms. The use of thresholds for controlling congestion in communication buffers is well known and is used, for example, active queue management (AQM) in current Internet routers. The basic principle is a simple one: if the mean queue length exceeds a pre-determined threshold, the arriving packets are dropped or marked with a specific probability. Then the position of the threshold and the value of the drop probability 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 end-to-end delay and the avoidance of lock out behaviour thus using the network resources more efficiently.

A number of studies have been reported in the literature to implement AQM. These include random early detection (RED) [3], random early marking (REM) [4,5], a virtual queue based scheme where the virtual queue is adaptive [6–8] and a proportional integral controller mechanism [9], among others. Of the above schemes to implement AQM, the RED mechanism is the one recommended by The Internet Society in [2]: Quote: Unless a developer has reasons to provide another equivalent mechanism we recommend that RED be used. This mechanism has the potential to overcome some of the problems discovered in drop tail mechanisms which are specific to the Internet traffic, such as synchronisation of TCP flows and correlation of the drop events (multiple packets dropped in sequence) within a TCP flow and it is therefore this mechanism that we will focus on in this paper.

The main aim of this paper is to formulate such a model with multiple queue thresholds and examine the queueing behaviour for multimedia type traffic under first come first served (FCFS) service discipline. To facilitate this aim, we study a stable GE/GE/1/**N** censored queue with a single server, finite capacity and multiple classes under a complex but effective buffer management scheme. The external bursty traffic and service time have been modelled using the generalised exponential (GE) distribution. The analysis has been carried out using the principle of maximum entropy (ME).

The remaining paper is organised as follows: Section 2 presents related work. Some preliminaries are outlined in Section 3. The ME solution for a stable GE/GE/1/**N** censored queue with buffer thresholds and FCFS scheduling discipline is characterised in Section 4. Numerical results, involving GE interarrival and service time distributions, are included in Section 5. Section 6 finally concludes the paper.

2. Related work

Traffic congestion in the Internet routers occurs when the aggregate demand exceeds the available capacity of resources. Performance modelling techniques help to design effective mechanisms capable of providing interactive services (such as voice, data and video) by developing performance models. The rapid development of communication networks and technologies in recent years has imposed great challenges for network support due to introduction of various multimedia type applications. As most current networking protocols and congestion control techniques were designed mainly for delay tolerant or elastic services, many buffer management schemes that make them efficient for these services are no longer true for new delay sensitive applications and cause severe performance degradation problems.

Feedback is a traditional technique for indicating the status of congestion within the network. Helali et al. [10] presented a multi-profile communication environment to ensure end-to-end QoS management. It supports dynamic assignment of application requirements to the network resources. The main contribution lies in the prediction of network congestion using feedback control algorithm to avoid overloading with streaming multimedia traffic.

Issues related to QoS and reliability design of packet networks have been addressed by mapping the end-user performance constraints into transport-layer performance constraints and then into network-layer performance constraints [11]. A collection of heuristic algorithms have also been presented and their performance has been validated against simulation. However, their analysis completely ignores employment of buffer management policies in the network routers.

Manvi and Venkataram [12] proposed an agent based scheme for adaptive bandwidth allocation when congestion occurs. This scheme functions at the network nodes and adaptively finds an alternative path for every congested or failed link and reallocates the bandwidth for the affected multimedia. The analysis of the proposed model is based on the simple assumption of modelling of constant bit rate (CBR) and variable bit rate (VBR) traffic sources using uniform distribution.

Kusmeirek and Du have simulated the problem of monitoring the available network resources by modelling the network behaviour and identifying the factors of vital importance [13]. Their approach relies on simple early congestion notification (ECN)-based mechanism to obtain a feedback from the network and end-point observations to determine available bandwidth and adjust the transmission rate using 3-rate adaptation mechanism. Their study considers only CBR-based traffic sources.

Most of the current studies are either based simple traffic assumptions or do not consider any buffer management for effectively dealing with the delay and delay jitter caused by the network queueing. This paper analyses a threshold based partial buffer sharing scheme which gives preferential treatment to delay sensitive traffic over the delay tolerant traffic in order to reduce the queueing delay. It models the external bursty traffic using GE distribution to closely match the real video streams.

3. Preliminaries

This section presents a brief overview of the principle of ME, one of the strong methodologies, for establishing product-form solutions of complex but realistic queueing systems, the GE distribution to model the external bursty traffic and service times and finally illustrates PBS scheme.

3.1. The principle of ME

The principle of ME [14,15] 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 the information theoretic context [14], 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 [15] 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 [16].

3.2. The GE distribution

The interevent-time distribution used will be the GE distribution (cf., Fig. 1), defined as:

$$F(t) = P(W \leqslant t) = 1 - \tau e^{-\sigma t}, \quad t \geqslant 0, \tag{1}$$

$$\tau = 2/\left(C^2 + 1\right), \tag{2}$$

$$\sigma = \tau \nu, \tag{3}$$

where W is a mixed-type 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 is versatile, possessing pseudo-memoryless properties which makes the solution of many GE-type queueing systems and networks analytically tractable [16].

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 variance may be relied upon, and thus, a choice of a distribution which implies least bias (i.e., introduction

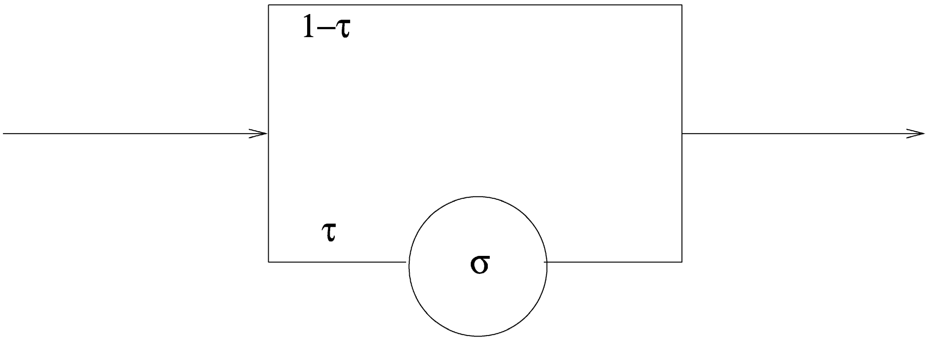

Fig. 1. The GE distribution with parameters τ and σ $(0 \leqslant \tau \leqslant 1)$.

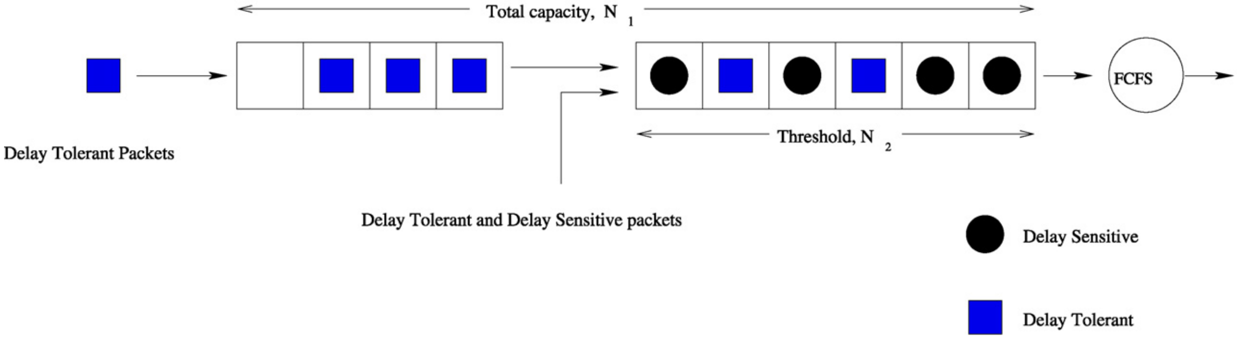

Fig. 2. The PBS management scheme with two job video streams under FCFS.

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 [17]. 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.

3.3. The PBS management scheme

The PBS management scheme works by setting a descending sequence of thresholds N_i $(N_i > 0, i = 1, 2, \ldots, R)$ corresponding to R priority classes of a single server queue with finite capacity N_1. Different packet loss and delay requirements under various load conditions can be met by adjusting the threshold values. The highest priority packets of class 1 can join the queue simply if there is space. However, lower priority packets of class i $(i = 2, \ldots, R)$ can only join the queue if the total number of packets in the queue is less than a threshold value N_i $(N_i \leqslant N_{i-1})$. Once the number of packets waiting for service reaches N_i, all lower priority packets of class k $(k = i + 1, \ldots, R)$ will be lost on arrival but higher priority packets of class ℓ $(\ell = 1, \ldots, i - 1)$ will continue to join the queue until it reaches threshold value, N_ℓ $(\ell = 1, \ldots, i - 1)$ (cf. Fig. 2 for PBS with two packet classes, respectively). The employment of single server finite capacity queues under a PBS scheme and applications into the performance evaluation of high speed networks can be seen in [18–21].

4. ME analysis of GE/GE/1/N/FCFS queue with buffer thresholds

This section presents the analysis of a single server GE/GE/1/**N** system to model a finite capacity queue with buffer thresholds. The analysis models the bursty external traffic with compound Poisson process (CPP) and the transmission times of this traffic are represented by the GE distribution under FCFS service discipline. The total buffer capacity is N_1 $(N_1 > 0)$ and the vector $\mathbf{N}$ represents a sequence of thresholds $\{(N_1, N_2, \ldots, N_R),\ 0 < N_i \leqslant N_{i-1},\ i = 2, \ldots, R\}$ to give space priorities to different classes of multimedia type traffic in order to control the delay and delay jitter by reducing the queue length.

Notation. For each class i $(i = 1, 2, \ldots, R)$, let λ_i be the mean arrival rate, C_{ai}^2 be the interarrival time SCV, μ_i be the mean service rate and C_{si}^2 be the service time SCV.

Focusing on a stable GE/GE/1/**N**/FCFS queue, let at any given time

n_i $(0 \leqslant n_i \leqslant N_i)$ be the number of class i $(i = 1, 2, \ldots, R)$ packets in the queue (waiting and/or receiving service),
$\mathbf{S} = (n_1, n_2, \ldots, n_R)$ be a joint queue state, where $\sum_{i=1}^{R} n_i \leqslant N_1$ (n.b., for an idle queue $\mathbf{S} \equiv \mathbf{0}$ with $\omega = 0$),
$\mathbf{Q}$ be the set of all feasible states $\mathbf{S}$,
$\mathbf{n} = (n_1, n_2, \ldots, n_R)$ be an aggregate joint queue state (n.b., $\mathbf{0} = (0, \ldots, 0)$),
Ω be the set of all feasible states $\mathbf{n}$.

Remarks.

- The arrival process for each class i $(i = 1, 2, \ldots, R)$ is assumed to be censored, i.e., a packet of class i will be lost if on arrival it finds N_i $(i = 1, 2, \ldots, R)$ packets at the queue.
- For exploration purposes, the analysis that follows focuses on the FCFS with PBS, 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.

4.1. Prior information

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

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

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

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

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

- normalisation constraint

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

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

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

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

- mean queue length, L_i $(U_i \leqslant L_i < N_1)$,

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

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

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

satisfying the flow balance equations, namely

$$\lambda_i(1-\pi_i)=\mu_i U_i, \quad i=1,2,\dots,R; \tag{8}$$

where π_i is the blocking probability that an arriving packet of class i finds N_i $(i=1,\dots,R)$ 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 (cf., [22]). 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.

4.2. A universal maximum entropy solution

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

$$H(P)=-\sum_{s}P(\mathbf{S})\log P(\mathbf{S}), \tag{9}$$

subject to constraints (4)–(7). By employing Lagrange's method of undetermined multipliers [12], the ME solution is expressed by

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

where Z, the normalising constant, is clearly given by

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

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

Remarks. Although constraints (5)–(7) are not known a 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). An efficient computational implementation of the ME solution (10), however, requires the prior estimation of the Lagrangian coefficients. This can be achieved by making GE-type buffer size invariance assumptions with regard to Lagrangian coefficients $\{g_i, x_i,\ i=1,2,\dots,R\}$ together with asymptotic connections to an infinite capacity GE/GE/1 queue (cf., [16]).

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

$$P(\mathbf{0})=\frac{1}{Z}, \tag{12}$$

$$P(\mathbf{k})=\sum_{i=1}^{R}\mathrm{Prob}(Q_{i;\mathbf{k}})=\frac{1}{Z}\left(\prod_{j=1}^{R}x_j^{k_j}\right)\sum_{j=1}^{R}k_j\left(\frac{(\sum_{i=1}^{R}k_i-N_j)!}{\prod_{i=1}^{R}(k_i-N_j)!}\right)g_j y_j^{\delta(\mathbf{k})}, \tag{13}$$

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

A universal form for the marginal blocking probabilities $\{\pi_i,\ i=1,2,\dots,R\}$ 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 interarrival time and service time stage selection probabilities $\sigma_i=(C_{ai}^2+1)/2$ and $r_i=(C_{si}^2+1)/2$, respectively. Each arriving bulk of class i $(i=1,2,\dots,R)$ joins the 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 interarrival times per class are equal). Let us focus on a tagged packet within an arriving bulk of class i $(i = 1, 2, \ldots, R)$ which finds the queue in state $\mathbf{n}_j = (0, \ldots, 0, n_j, n_{j+1}, \ldots, n_R)$ where $n_k = 0$, $k = 1, 2, \ldots, j-1$. 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 $N_1 - v$.

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

$$\pi_i = \sum_{k=0}^{N} \delta_i(k)(1-\sigma_i)^{[N_1-k]^+} P_{N_1}(k), \tag{14}$$

where

$$\delta_i(k) = \begin{cases} \frac{r_i}{r_i(1-\sigma_i)+\sigma_i}, & k = 0, \\ 1, & ow. \end{cases} \tag{15}$$

4.3. The 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 size N_1, it can be established that

$$x_i = \frac{\langle n_i \rangle - \rho_i}{\langle n \rangle}, \tag{16}$$

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

where $X = \sum_{i=1}^{R} x_i$, $\langle n \rangle = \sum_{i=1}^{R} \langle n_i \rangle$ and $\langle n_i \rangle$ is the asymptotic marginal mean queue length of a multi-class GE/GE/1 queue. Note that statistics $\langle n_i \rangle$, $i = 1, 2, \ldots, R$ can be determined by (cf., Kouvatsos and Denazis [22])

$$\langle n_i \rangle = \frac{\rho_i}{2}\left(C_{ai}^2 + 1\right) + \frac{1}{2(1-\rho)} \sum_{j=1}^{R} \frac{\Lambda_i}{\Lambda_j} \rho_j^2 \left(C_{aj}^2 + C_{sj}^2\right), \tag{18}$$

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

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

5. Numerical results

A number of experiments have been conducted using the proposed analytical model and simulation (based on QNAP-2 at 95% confidence interval [23]) to evaluate the buffer thresholds based mechanism for delay sensitive and delay tolerant multimedia traffic. The objective of these experiments is to inspect the effectiveness of buffer thresholds for reducing the system response time for delay sensitive video and with acceptable level of delay for delay tolerant stored video streaming. Furthermore, status of the buffer occupancy can be used to generate the feedback decision for the sending source to possibly avoid network bottleneck and packet loss.

As far the modelling is concerned, the representation of the video traffic inside the Internet is a sensitive issue. Traditionally Poisson distribution has been used to model the Internet traffic. But due to the bursty nature of multimedia services, traffic generated by live video and video streaming do not follow Poisson process [24]. In these experiments, 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.

In the first experiment, we are using two homogeneous video applications, representing a delay sensitive live video and delay tolerant video streaming, which generate traffic with the same arrival rates and SCV values. We have encoded both types of traffic as MPEG-2 at a bit rate of 1.75 Mbps. The overall data rate for each application, after taking the header parts into account is around 2.0 Mbps. The service rate (or the link capacity) is 4.0 Mbps.

For a total capacity of ten packets, Fig. 3 shows that by increasing the threshold value, the mean response time for the delay sensitive applications reduces whilst it increases for the delay tolerant applications. This is because more

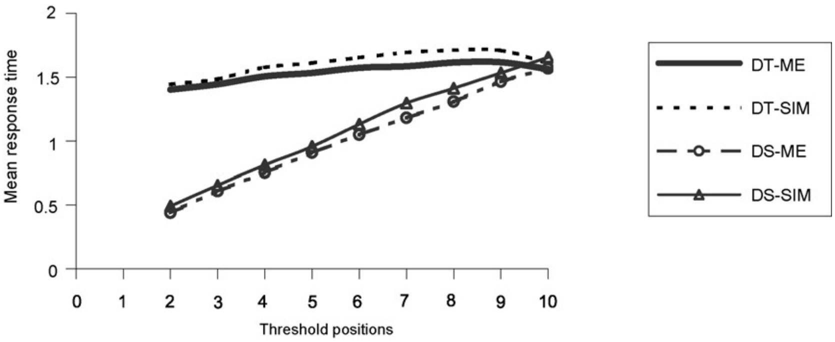

Fig. 3. Response times.

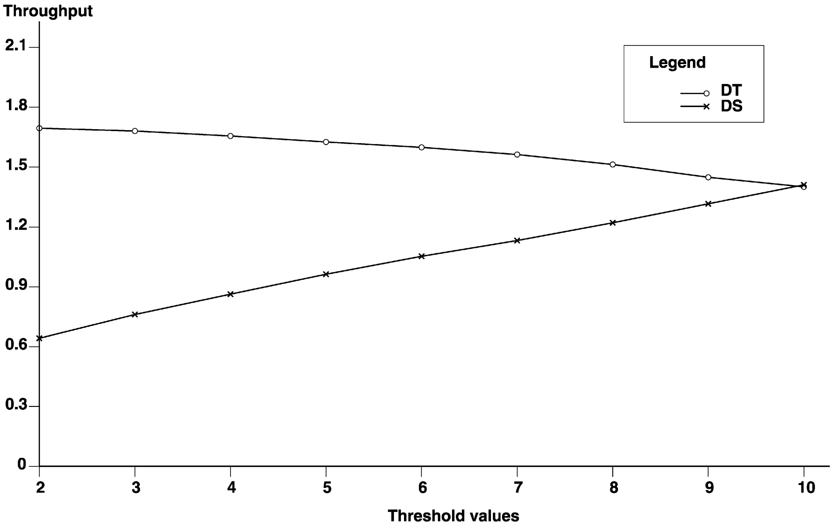

Fig. 4. Throughput.

space will be shared with packets from delay tolerant applications whose packets are arriving with the same rate. It is clearly seen that when the entire buffer becomes a shared region, both streams, being homogeneous, will exhibit the same mean response time.

Figures 4 and 5 show a similar effect of threshold position settings on blocking probability and throughput values for both types of video streams. Increasing values of buffer threshold will increase the throughput for delay tolerant streams whilst decreasing that of delay sensitive traffic. Blocking probability will also decrease for delay tolerant traffic by making more buffer space available.

It is interesting to note in Figs. 3–5 that all curves coincide at one point when the threshold gets its maximum value demonstrating that the entire buffer is being commonly shared by both homogeneous video streams arriving with the same arrival rate. At this stage both streams, delay sensitive and delay tolerant, will be equally treated by the system. This shows the strengths of buffer thresholds to meet diverse QoS requirement by different streams.

In the second experiment, we are using 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 the same as 4.0 Mbps. We are also using higher values of SCV for delay tolerant video streaming to show its high bursty nature.

Figures 6 and 7 show that mean response time and throughput values for delay sensitive traffic and delay tolerant streams vary similar to homogeneous streams by increasing the threshold position. The main difference can be noticed that both curves coincide before the entire buffer becomes shared (i.e., before the threshold gains its maximum value). This is because the traffic is more bursty and arriving with different rates.

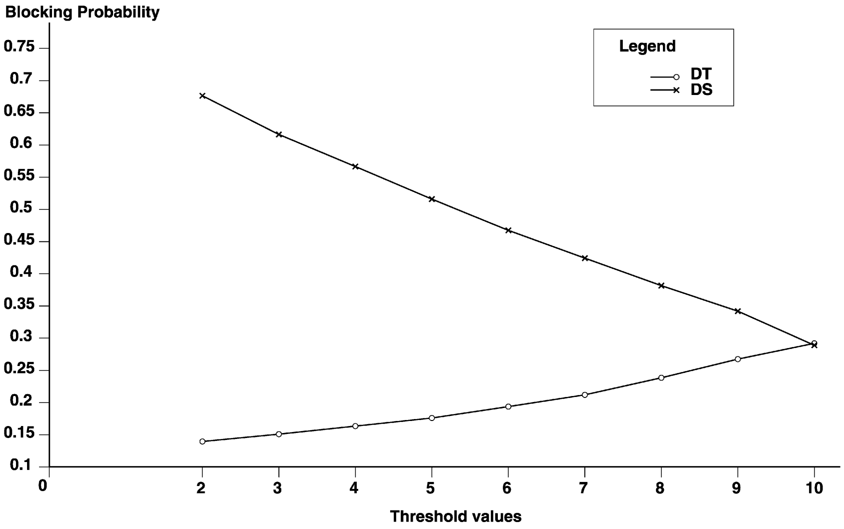

Fig. 5. Blocking probability.

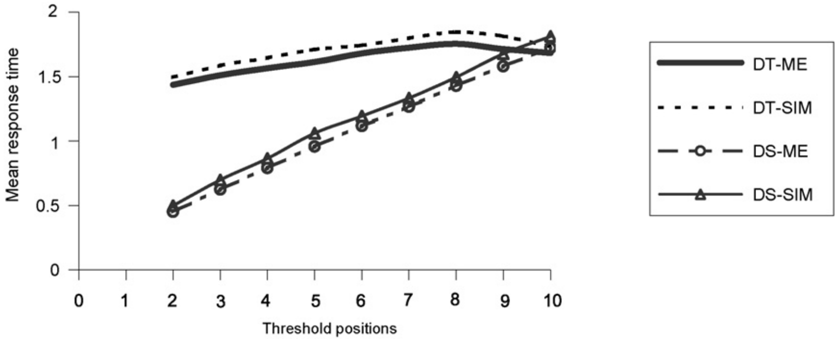

Fig. 6. Response times.

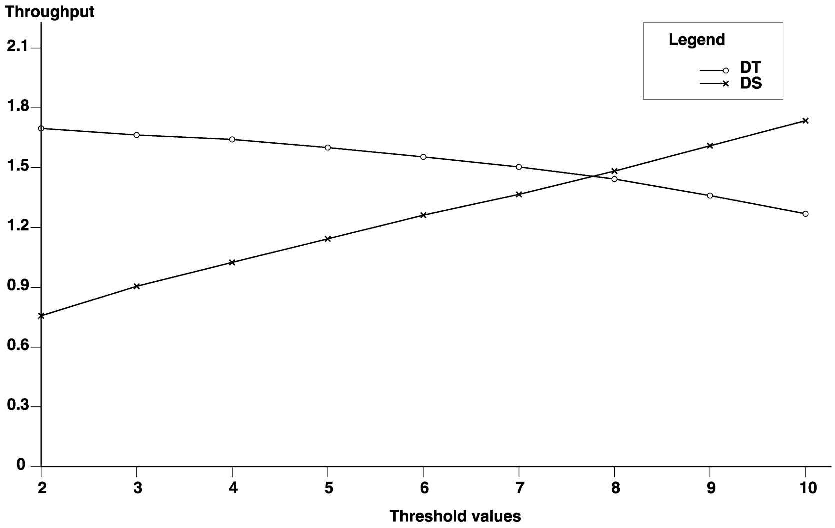

Fig. 7. Throughput.

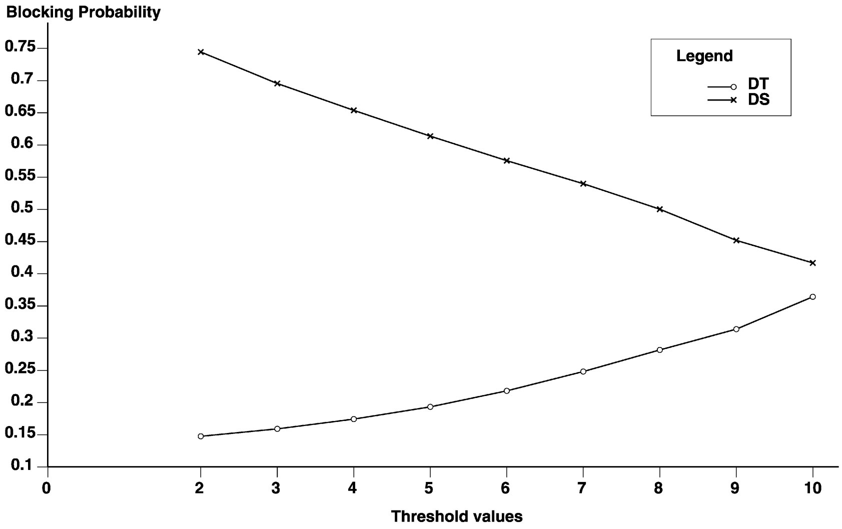

Fig. 8. Blocking probability.

Furthermore, Fig. 8 shows the effectiveness of using thresholds with PBS scheme under FCFS in order to provide various grades of service to different classes of video streams. It has been clearly shown from Fig. 8, that by increasing the position of threshold the packet loss probability for high priority decreases slightly but that of low priority traffic increases. When the position of threshold reaches the total capacity, both traffic classes show the slightly different loss probabilities which is due to the different bursty properties of the two heterogeneous streams. It means a packet of any class will be lost with different rate when it finds the queue full upon its arrival. This model, therefore, can be used as a powerful tool to provide a required grade of service to a particular class of traffic in any heterogeneous distributed network.

6. Conclusions

This paper presents analysis of a threshold based queue for multimedia based networked applications in order 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 PBS buffer management 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 and blocking 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 threshold settings on various performance metrics for delay tolerant and delay sensitive video streams.

References

[1] D.C.A. Bulteman, SMIL 2.0.2. examples and comparisons, IEEE Multimedia 9 (1) (2002) 74–84.

[2] B. Braden et al., Recommendations on queue management and congestion avoidance in the Internet, in: IETF RFC, 2309, April 1998.

[3] S. Floyd, V. Jacobson, Random early detection gateways for congestion avoidance, IEEE/ACM Trans. Netw. 1 (4) (1993) 397–413.

[4] D. Lapsley, S. Low, Ransom early marking for Internet congestion control, in: Proceeding of GLOBECOM ’99, December 1999, pp. 1747–1752.

[5] S. Athuraliya, D. Lapsley, S. Low, An enhanced random early marking algorithm for Internet flow control, in: INFOCOM 2000, pp. 1425–1434.

[6] R. Gibbens, F. Kelly, Distributed connection acceptance control for a connectionless network, in: Proceeding of the 16th Intl. Teletraffic Congress, Edinburgh, Scotland, June 1999.

[7] S. Kunniyur, R. Srikant, End-to-end congestion control: Utility function, random losses and ECN marks, in: Proceeding of INFOCOM 2000, Tel Aviv, Israel, March 2000.

[8] S. Kunniyur, R. Srikant, A time-scale decomposition approach to adaptive ECN marking, in: Proceeding of INFOCOM 2001, Alaska, Anchorage, April 2001.

[9] C. Hoolot, V. Misra, D. Towlsey, W. Gong, On designing improved controllers for AQM routers supporting TCP flows, UMass CMPSCI Technical Report 00-42, 2000.

[10] A. Helali, A. Soudani, S. Nasri, T. Divoux, An approach for end-to-end QoS and network resources management, Comput. Standards Interfaces 28 (1) (2005) 93–108.

[11] E. Wille, M. Mehhia, E. Leonardi, M. Marsan, Algorithm for IP network design with end-to-end QoS constraints, Comput. Netw. 50 (2006) 1086–1103.

[12] S.S. Manvi, P. Venkataram, An agent based adaptive bandwidth allocation scheme for multimedia applications, J. Systems Softw. 75 (2005) 305–318.

[13] E. Kusmeirek, D. Du, Streaming video delivery over Internet with adaptive end-to-end QoS, J. Systems Softw. 75 (2005) 237–252.

[14] E.T. Jaynes, Information theory and statistical mechanics, Phys. Rev. 106 (1957) 620–630.

[15] E.T. Jaynes, Information theory and statistical mechanics, II, Phys. Rev. 108 (1957) 171–190.

[16] D.D. Kouvatsos, I.U. Awan, MEM for arbitrary closed queueing networks with RS-blocking and multiple job classes, Ann. Oper. Res. 79 (1998) 231–269.

[17] F. Ball, D. Hutchinson, D.D. Kouvatsos, VBR video traffic smoothing at the AAL SAR level, in: D.D. Kouvatsos (Ed.), Proc. of 4th IFIP Workshop on Performance Modelling and Evaluation of ATM Networks, Ilkley, 1996, pp. 28/1–28/10.

[18] S.Q. Li, Overload control in a finite message storage buffer, IEEE Trans. Commun. 37 (12) (1989) 1330–1338.

[19] N. Yin, S.Q. Li, T.E. Stern, Congestion control for packet voice by selective packet discarding, IEEE Trans. Commun. 38 (5) (1990) 674–683.

[20] H. Kroner, Comparative performance study of space priority mechanisms for ATM networks, in: Proc. IEEE INFOCOM ’90, 1989, pp. 1136–1143.

[21] D. Hong, T. Suda, Congestion control and prevention in ATM networks, IEEE Network Magazine, 1991, pp. 10–16.

[22] D.D. Kouvatsos, S.G. Denazis, Entropy maximised queueing networks with blocking and multiple job classes, Perform. Eval. 17 (1993) 189–205.

[23] M. Veran, D. Potier, QNAP-2: A portable environment for queueing network modelling techniques and tools for performance analysis, in: D. Potier (Ed.), North-Holland, 1985, pp. 25–63.

[24] V. Paxson, S. Floyd, Wide-area traffic: The failure of Poisson modelling, IEEE/ACM Trans. Netw. 3 (3) (1995) 226–244.