PERFORMANCE MODELLING OF A MULTIPLE THRESHOLD RED MECHANISM FOR BURSTY AND CORRELATED INTERNET TRAFFIC WITH MMPP ARRIVAL PROCESS*

ASFAND-E-YAR, I.U.AWAN and M.E.WOODWARD

Networks and Performance Engineering Research Group
Department of Computing, School of Informatics
University of Bradford, Bradford,
BD7 1DP Bradford, West Yorkshire, England, UK
E-mail : {A.Yar, I.U.Awan, M.E.Woodward }@Bradford.ac.uk,

Abstract. Access to the large web content hosted all over the world by users of the Internet engage many hosts, routers/switches and faster links. They challenge the internet backbone to operate at its capacity to assure efficient content access. This may result in congestion and raises concerns over various Quality of Service (QoS) issues like high delays, high packet loss and low throughput of the system for various Internet applications. Thus, there is a need to develop effective congestion control mechanisms in order to meet various Quality of Service (QoS) related performance parameters. In this paper, our emphasis is on the Active Queue Management (AQM) mechanisms, particularly Random Early Detection (RED). We propose a threshold based novel analytical model based on standard RED mechanism. Various numerical examples are presented for Internet traffic scenarios containing both the burstiness and correlation properties of the network traffic.

Key Words: Congestion Control, Active Queue Management (AQM) mechanisms, Random Early Detection (RED), Continous Time Markov Modulated Poisson Process (CT-MMPP)

1 INTRODUCTION

In recent years, there are two major technological forces that have driven the evolutionary era in communication. They are: (i) Wireless Technology evolution and (ii) the Internet. As these forces converge, the demand for new services, increasing bandwidth and ubiquitous connectivity will grow continually. Increasing traffic due to various services will cause traffic congestion at input and output ports of Internet routers and switches. Thus, the problem of congested networks occurring when the aggregate demand exceeds the available capacity of resources will always be a challenge for network operators, consultants and researchers. Congestion results in delay, packet loss or incomplete information access and results in overflowing queues. The total (end-to-end) network delay is essentially the sum of the queuing delay in routers and switches and propagation delay. Currently queuing delay dominates most Round Trip Times (RTT). The goal should be to reduce the network delay close to the propagation delay.

The cost effective way of evaluating network performance with regard to Quality of Service (QoS) issues i.e, speedy connectivity and negligible packet loss with high bulk throughput is to use traffic models and performance modelling techniques. Over a number of years traffic modelling has received considerable attention from the networking community. A good traffic model should be accurate enough to capture the statistical characteristics of actual traffic and at the same time should be computationally efficient. Along with Traffic modelling, the performance modelling techniques help to determine the efficacy of telecommunication networks capable of providing interactive services (such as voice, data and video) by developing good models. Thus, in

*Supported by Engineering and Physical Sciences Research Council (EPSRC), UK, under grant GR/S01658/01.

this paper we use traffic and performance modelling techniques for the analysis of our proposed model for the Internet traffic containing both burstiness and correlation properties of the network traffic.

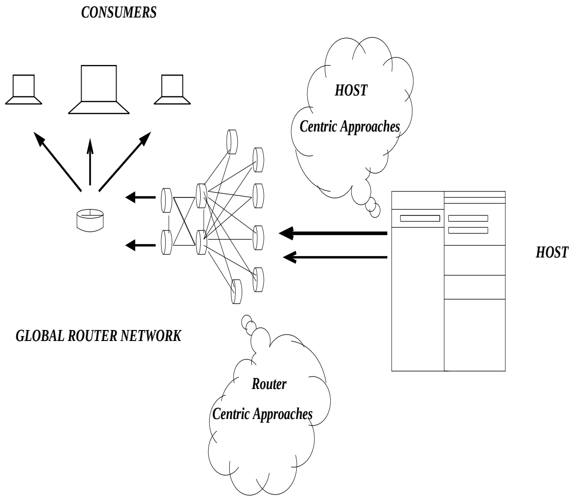

Fig. 1. Complex Internet Setup.

Over the past few years researchers have been working on various methodologies to overcome the Quality of Service (QoS) issues, particularly *Congestion Control* [1] and *Congestion Avoidance* [2] faced by Internet Traffic due to increasing number of users around the globe. As a result, there are mechanisms which operate from the edge of the network called HOST-Centric approaches [3] for congestion control and avoidance, but they are not enough to coupe for congestion in all circumstances because of the limit to how much control can be exercised from the edge of the network i.e. HOST. Thus, in order to complement the HOST-Centric approaches, there are mechanisms devised to work inside the network known as ROUTER centric approaches for congestion control and avoidance.(c.f., Fig. 1). Thus, each Internet router can be modeled as a queuing system with certain arrival source (HOST), a finite capacity Queue and a server (Router Network), serving packets under certain service disciplines.(c.f., Fig. 2)
There are two main approaches for congestion avoidance at routers. They are: (a) *Scheduling Algorithms*, which regulate the allocation of bandwidth and packet selection on the basis of available resources and (b) *Queue Management*, which manages queue length by dropping packets when necessary. Furthermore, there are two classical Queue Management techniques.

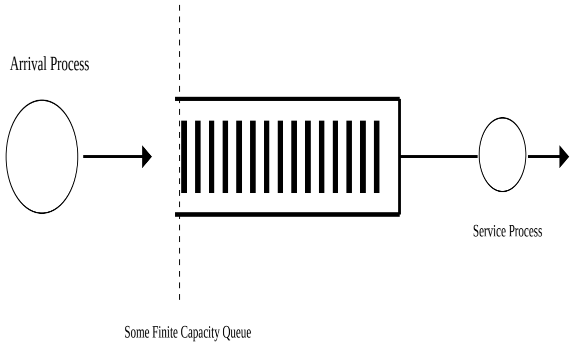

Fig. 2. A simple Queuing System .

Firstly, *Drop Tail*, where a maximum queue length is set for each queue at the router and accepts the packets until the maximum queue length is reached. Once the maximum queue length is achieved, the algorithm drops packets until queue length is again below the maximum set value. Thus, Drop Tail has serious limitations to its use like:(i) It is not Suited for Interactive Networks such as voice-video sessions and web transfers requiring low end-to-end delay and jitter. Because the drop tail queues are always full or close to full for long periods of time and packets are continuously dropped when the queue reaches its maximum length resulting in large delays which makes interactive applications (voice, video etc) untenable. (ii) Global synchronization; arises because the full queue length is unable to absorb bursty packet arrivals and thus many packets are dropped resulting in global synchronization. Thus, we may also say that the main reason of global synchronization in drop tail queues is its lockout behavior where the queue is monopolized by some flows and other connections may not easily use the queue.

[1] [2, 5], It comes into play after the congestion at a network has occurred and the network is overloaded.

[2] preventive technique [1 - 4], which comes into play before network is congested by overloading. A congestion avoidance scheme maintains the network in a region of low delay and high through put by keeping the average queue size low, such that fluctuations in the actual queue size should be allowed to accommodate bursty traffic and transient congestion.

[3] In 1988, Jacobson [3], pioneered the concept of TCP congestion control and avoidance which forms the basis of today's HOST Centric congestion control approaches and mechanisms. TCP Congestion Control algorithms like slow start, congestion avoidance, fast recovery and fast retransmission; and TCP variants like Tahoe: The sender implements fast retransmission only, Reno: The sender implements both fast retransmission and fast recovery, Modified New – Reno: The sender retransmits one lost packet per round-trip time (RTT) upon receiving partial Acknowledgements (ACKs) and terminates the recovery phase when the whole window is acknowledged, and SACK.

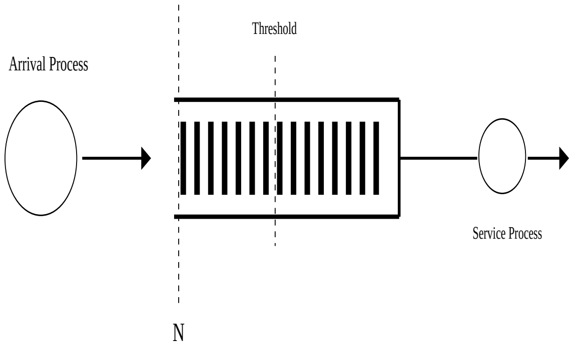

Fig. 3. Active Queue Management: Second Traditional Queue Management Technique.

Secondly, *Active Queue Management* (AQM) (c.f., Fig. 3) where packets are dropped before the queue becomes full by using some sort of threshold. AQM [10, 11], provides a solution to overcome the demerits of the drop tail scheme. It maintains a small size steady state queue, thus resulting in reduced packet loss, decreased end-to-end delay, and the avoidance of lock out behavior. In this way the network resources are used more efficiently. Also, by keeping the average queue size small results in the efficient use of bandwidth and increases the link utilization by avoiding global synchronization. Furthermore, due to availability of extra queue space, packet bursts will be absorbed as well. Finally, the bias of routers against flows that use small bandwidth due to monopolize flows will be prevented, which will result in prevention of lockout behavior.

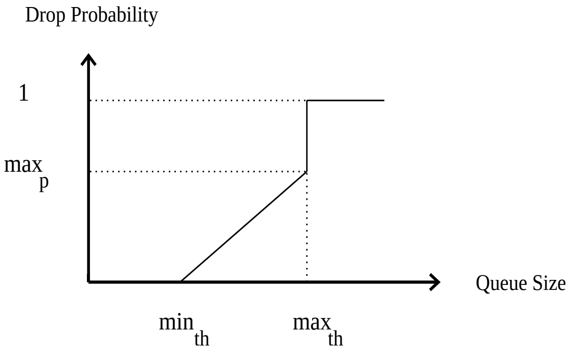

Fig. 4. RED's Packet Drop Probability.

The AQM techniques to maintain Acceptable Level of Service (ALoS) is Random Early Detection (RED) recommended by Internet Engineering Task Force (IETF) in RFC 2309 [6, 7], (c.f., Fig. 4 and 5). The basic idea is to sense congestion and try to provide feedback to the sources by either marking or dropping their packets, even if space is still available in the queue. Therefore, RED design objectives are to: (i) minimize packet loss, (ii) minimize queuing delay, (iii) maintain high link utilization (iv) remove biases against bursty sources and (v) avoid global synchronization.

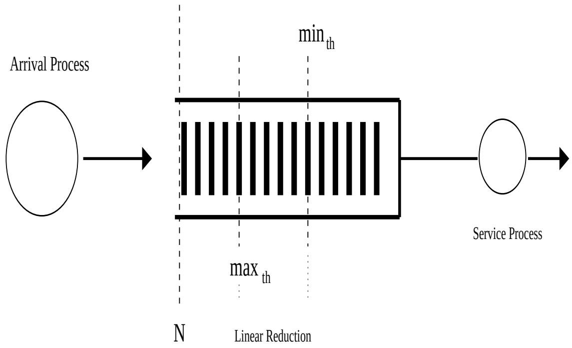

Fig. 5. RED Queue.

The remainder of the paper is as follows: Section 2 describes briefly an overview of the related work on RED as an AQM mechanism and while doing so tries to highlight the limitations of RED. Section 3 elaborates the proposed analytical performance model, where a Continous Time Markov Modulated Poisson Process (CT-MMPP) is used under a sampling approach to model input network traffic. Section 4 discusses results obtained by using some of the numerical examples and shows the effect of traffic burstiness and correlation. We conclude the paper in Section 5 with some future directions based on the proposed model in this paper.

2 RELATED WORK

Many AQM schemes have been proposed [8], yet no single one can achieve optimal performance in all network and traffic scenarios. An AQM scheme without adaptability can hardly meet the QoS requirements under dynamic network and traffic conditions. Also, most AQM schemes are not efficient in providing fairness for users with different congestion sensitivities, and thus are not efficient to control non-responsive traffic. So, we may say that all the existing AQM schemes cannot be directly adapted when applied, for example, to wireless networks. If it is applied in wireless networks it will not be able to absorb bursts of packets as allocated buffer capacity may have already achieved its maximum threshold allocation over a period of time. Now, till it re-averages a significant amount of traffic will continue to be lost. On the other hand if we use instantaneous queue length where the buffer accommodates itself with the changing characteristics of input traffic, the buffer is less likely to be full since the initial packet losses as the instantaneous queue increases will give immediate reduction of arrival rate and the network will not experience overloading. This may be achieved by time stamping the data arriving in the buffer at every instance and storing this information. As the time progress and packets are still arriving the buffer adjusts its occupancy by keeping the buffer thresholds moving to utilize the resource more effectively. Thus, we consider the use of instantaneous queue length to be more appropriate in the analysis if the proposed model.

Sally Floyd and Van Jacobson proposed the RED algorithm for RED Gateways [7], which calculates the average queue size, using a low pass filter with an exponential weighted moving average (EWMA). The average queue size is compared to two thresholds, a minimum threshold and a maximum threshold. When the average size of queue is less than the minimum threshold, no packet is dropped but if the average queue size is greater than the minimum threshold, every arriving packet is dropped with certain probability P_a till it reaches the maximum threshold. After maximum threshold, all arriving packets are dropped. Thus, we can summarize some of the key parameters for RED as: (i) maximum threshold $((\max_{Th}))$, (ii) minimum threshold $((\min_{Th}))$, (iii) queue length ($qLen$), (iv) weighted factor for average queue length comparison (W_q), (v) the maximum probability of early drop $((\max_P)$. Fig. 5 shows RED's packet drop probability distribution and Fig. 6 shows a generic representation of a RED queue.

Bonal et al [9] proposed a model that uses classical queueing theory (M/M/1/K Model and Drop function for RED based on its algorithm) to evaluate RED performance and quantify the benefits (or lack thereof) brought about by RED. Basically, three major aspects of the RED scheme, namely the bias against bursty traffic, synchronization of TCP flows and a queuing delay study were presented in details and compared with those of Drop Tail scheme to evaluate and validate the performance of RED. Firoiu and Borden [10], were the followers by introducing the flow feedback control model established to analyze the stability of the RED control system. A number of studies have been reported [10 - 13] in the literature for deploying RED, or Adaptive Virtual Queue (AVQ) when deployed as a congestion control mechanism on TCP/IP routers. Almost exclusively these studies are based on simulation and mostly consider the best case scenarios for their simulations in which a constant number of TCP connections, each sending continuously, share a queue on a bottleneck link. More recently, Mikkel et al [14] presented a simulation based performance study for the worst case situation in which there are a dynamically changing number of TCP connections with highly variable lifetimes. However, the study focussed mainly on a very special scenario where a link is considered as carrying only Web-like traffic and is specific to HTTP 1.0 protocols.

In spite of the fact that probably RED is the most promising AQM [8, 10 - 15] scheme for congestion avoidance and control, research has shown that the performance of RED is highly dependent upon the way its parameters are tuned and the environment where it is used. Variants of RED [8] try to solve some of the problems. One of the main weaknesses pointed out in the RED variants [8] is the variation of the average queue size of RED with the level of congestion and parameter settings. When $\max_p$ is large and/or network is lightly congested, the average queue size is near $\min_{th}$; conversely when $\max_p$ is small and/or the network is heavily congested, the average queue size is close to $\max_{th}$. Thus, the queuing delays at the routers cannot be easily estimated because the changes in the average queue size vary widely according to the parameters and congestion in the network. Unfortunately, a major consideration of QoS is the delay, and this means network providers will not be able to have an accurate estimate of the queuing delay in network. Some researchers [9, 14] even claim that there is no advantage in using RED over Drop Tail as in their opinion throughput may be very poor for RED if the parameters are not tuned properly. Consequently, the number of consecutively dropped packets may be larger than those of Drop Tail so that RED may not help the global synchronization problem as well.

Thus, generally speaking, all queue management mechanisms perform well in certain scenarios but are sub-optimal in others. Thus development of new AQM mechanisms and evaluation of their performance can still be considered as a challenge for the research community [8]. Although research on RED applications and its variants [8] seem to be very elaborate and promising, still further investigations are necessary. Bartek Wydrowski et al. in [16] stresses the need of improved AQM mechanisms for today's best-effort networks to deliver the performance required for a wide range of interactive and multimedia applications that have demanding delay and bandwidth requirements.

However, despite the recommendation [6] by the Internet Engineering Task Force (IETF) and RED being widely deployed, the fact is that RED is not thoroughly understood. There is little operational experience of RED in large scale networks

- one of the few published measurement study is limited in scope because it only considers the router performance as opposed to the end-to-end performance and it does not clearly describe the measurement settings and the exact information being measured [15]. It is not quite clear how to choose RED parameters whereas the recommended values in [6] have changed over time and there are few published analytical models of RED that would, for example, allow us to quantify the impact of parameter settings on performance or the impact of different parameter values taken by different ISPs in a large network. Hence, we can say that so far no analytical solution for the performance analysis of RED is available to help understand its specific operational functions. This is the motivation for this paper.

3 PROPOSED MODEL

The Markov Modulated Poisson Process (MMPP) has been extensively used to model B-ISDN sources [17], such as voice and video, as well as characterizing the superposed traffic. It captures the burstiness and correlation properties of the network traffic. In addition to characterizing the desired properties of B-ISDN applications, these models are analytically tractable and produce results that are acceptable approximations to reality.

An MMPP is a doubly stochastic Poisson Process [17]. The arrivals occur in a Poisson manner with a rate that varies according to a k-state Markov chain, which is independent of the arrival process. Accordingly, an MMPP is characterized by the transition rate matrix of its underlying Markov chain and arrival rates. Let i be the state of the Markov chain, σ_{ij} be the transition rate from state i to state j, $i \neq j$ and λ_i the arrival rate when the Markov chain is in state i, $\lambda_i > 0$, and $i \in \{1, ..., k\}$. Define:

$$\sigma_i = -\sum_{j=1; i\neq j}^{k} \sigma_{ik}$$

In matrix form, we have

$$Q = \begin{bmatrix} -\sigma_1 & \sigma_{12} & .... & \sigma_{1k} \\ \sigma_{21} & -\sigma_2 & .... & \sigma_{2k} \\ . & . & . & . \\ \sigma_{k1} & \sigma_{k2} & .... & -\sigma_k \end{bmatrix} \text{ and}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & .... & 0 \\ 0 & \lambda_2 & .... & 0 \\ . & . & . & . \\ 0 & 0 & .... & \lambda_k \end{bmatrix}$$

Assuming that Q does not depend on time t, the steady-state probability vector, π, of Q is the solution of the following system of equations:

$$\pi Q = 0; \quad \sum_{i=1}^{k} \pi_j = 1$$

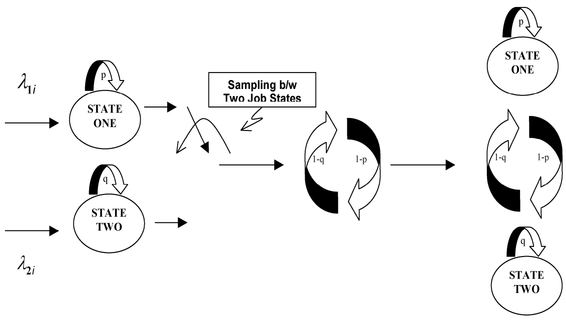

Fig. 6. Sampling concept for proposed model.

A k-state MMPP can be viewed in two ways: (i) As k independent traffic streams, each having a different Poisson arrival rate, which are randomly sampled at the input. Each stream might thus represent a different class of traffic. (ii) As a single traffic stream with correlated inter-arrival times and a variable squared co-efficient of variation of the inter-arrival times (and thus a variable burstiness). Fig. 6 shows the sampling process for our analysis where two independent Poisson streams are randomly sampled with exponentially distributed inter-switching time.

In our model we use a pair of thresholds for each traffic class. Fig. 7 shows the positions of thresholds for two traffic classes, in a finite capacity buffer where traffic arrives using a 2-state Markov Modulated Poisson Process (MMPP) distribution.

We use a conventional matrices approach to solve two-dimensional CT-MMPP Markov chains [18-21]. In order to do so we create a Z generator matrix based on a system state transition diagram (c.f., Fig. 9) comprising the input parameters λ_{ij} – arrival rate, μ – service rate, $p = R$ – Rate of transition from one state to the second state and $q = \acute{R}$ – Rate of transition from second state to the first state. The procedure of solving the matrix is the same as depicted in [18]. Figs. 7 - 8, show the positions of thresholds and regions for the linear reduction in arrival rate. We assume that N_{21} - N_{11} = N_{22} - N_{12}, such that N_{12} - N_{11} = N_{22} - N_{21} and $N_{21} > N_{12}$.

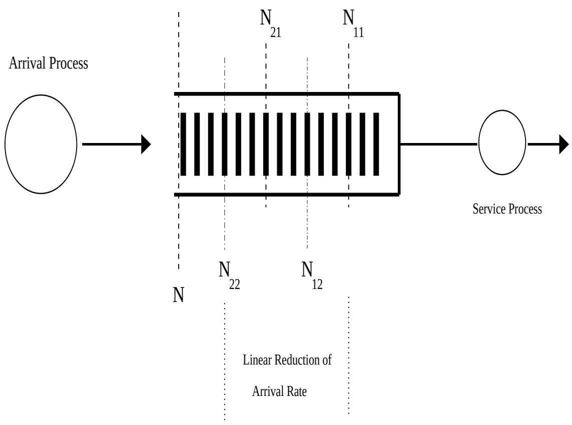

Fig. 7. RED Mechanism for AQM: Single buffer with two thresholds per traffic class. (N_{1i}- min and N_{2i}- max threshold per class where: i = 1,2, number of class.

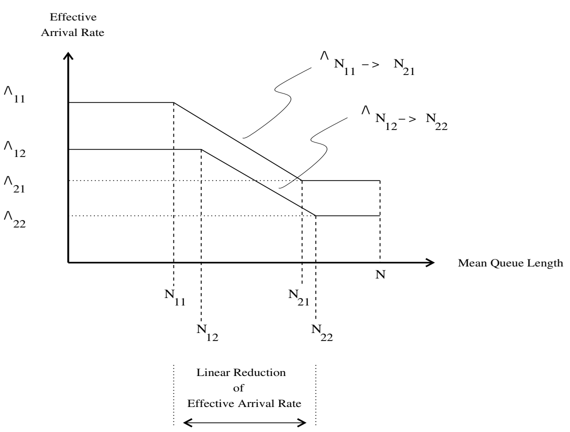

Fig. 8. Application of Standard RED Algorithm to the arrival process(N_{1i}- min and N_{2i}- max threshold where: i = 1,2, number of job class.

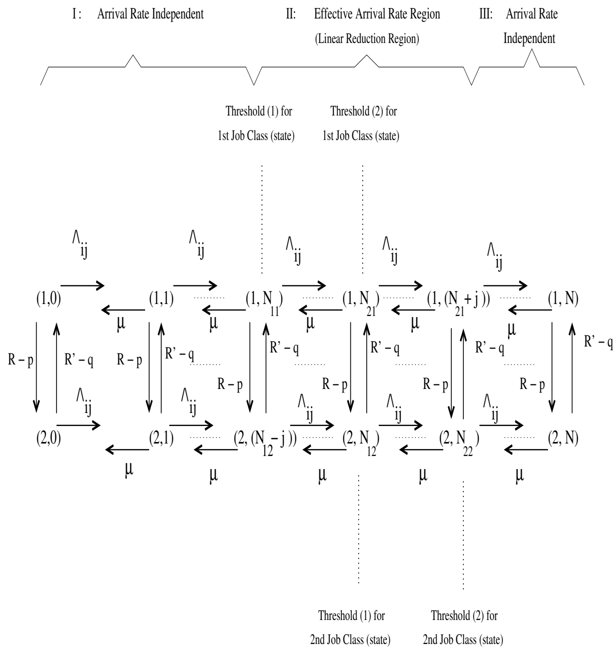

Fig. 9. State Transition Diagram for the Proposed Model based on RED Mechanism and CT-MMPP.

The arrival rate λ_{ij} are all independent of the state before N_{1i} or after N_{2i} and depend on the state between N_{1i} and N_{2i}, where i = 1,2, and j = 0, 1, 2, ... N, N being full queue capacity, i.e. each arrival rate is different with each state and will be linearly reduced by dropping packets in a region between thresholds N_{11}, N_{21}, and thresholds N_{12}, N_{22}, for first and second traffic class respectively. These arrival rates can be obtained as:

$$\lambda_{ij} = \begin{cases} \lambda_{1i}, \\ for, i = 1,2 \;\; and \;\; 0 \leq j \leq N_{1i} \\ \lambda_j = \lambda_{1i} - (N_{1i} - j)\frac{\lambda_{2i} - \lambda_{1i}}{N_{2i} - N_{1i}}, \\ for, i = 1,2 \;\; and \;\; N_{1i} + 1 \leq j \leq N_{2i} \\ \lambda_{2i}, \\ for, i = 1,2 \;\; and \;\; N_{2i} + 1 \leq j \leq N \end{cases} \tag{1}$$

In order to perform the steady state analysis of the system, we use algorithm in [18] to solve the joint steady state probability vector $\mathbf{P} = \mathbf{P}_j$ (0≤ j ≤N) in the two-dimensional Markov chain, which satisfies the following equations:

$$\mathbf{PZ} = 0 \qquad and \qquad \mathbf{Pe} = 1 \tag{2}$$

where **Z** is the generator matrix based on the system state transition diagram (c.f., Fig. 9). We have grouped states (c.f., Fig. 9) according to the total number of customers in the queue (where: j = states = 0, 1, 2, ... N, for i = 1,2, number of job class). We order states lexigraphically, i.e. (1,0), (2,0), (1,1), (2,1), (1,2), (2,2)...(1,N), (2,N) where, N is the maximum queue length (full buffer capacity). Thus, the generator matrix **Z** is given by:

$$Z = \begin{pmatrix} A_{ij} & E_{ij} & 0 & 0 & . & . & . \\ C_{ij} & D_{ij} & E_{ij} & 0 & . & . & . \\ 0 & C_{ij} & D_{ij} & E_{ij} & 0 & . & . \\ 0 & 0 & C_{ij} & D_{ij} & E_{ij} & 0 & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & C_{ij} & B_{ij} \end{pmatrix} \tag{3}$$

Where:

$$A_{ij} = \begin{pmatrix} -(p+\lambda_{ij}) & p \\ q & -(q+\lambda_{ij}) \end{pmatrix}$$

$$B_{ij} = \begin{pmatrix} -(\mu+p) & p \\ q & -(\mu+q) \end{pmatrix}$$

$$E_{ij} = \begin{pmatrix} \lambda_{ij} & 0 \\ 0 & \lambda_{ij} \end{pmatrix}$$

$$C_{ij} = \begin{pmatrix} \mu & 0 \\ 0 & \mu \end{pmatrix}$$

and

$$D_{ij} = \begin{pmatrix} -(\mu+p+\lambda_{ij}) & p \\ q & -(\mu+q+\lambda_{ij}) \end{pmatrix}$$

Solving the generator matrix **Z** yield the steady state vector as [18-23]:

$$\mathbf{P} = \mathbf{u}(\mathbf{I} - \mathbf{Q} \ + \mathbf{eu})^{-1} \tag{4}$$

where $\mathbf{Q} = \mathbf{I} + \mathbf{Z}/\min\{\mathbf{Z}_{i,i}\}$ and $\mathbf{u}$ is an arbitrary row vector of $\mathbf{Q}$. $\mathbf{I}$ is an N×N identity matrix and $\mathbf{e} = (1, 1, \ldots, 1)^T$ is a unit column vector of length N.

After the equilibrium probabilities $\mathbf{P}_j$ ($0 \leq$ j $\leq$N) are found, we can evaluate system performance metrics such as mean system occupancy, mean packet waiting time, system throughput and packet dropping probabilities. The average buffer occupancy or Mean Queue Length (MQL) can be expressed from the equilibrium probabilities $\mathbf{P}_j$ as:

$$L = \sum_{j=0}^{N} j\mathrm{P}_j \tag{5}$$

Using Little's law, the delay for this finite capacity queue can be obtained as:

$$W = \frac{L}{S} \tag{6}$$

where S is the mean throughput of the continous-time finite capacity queue given by the fraction of time the server is busy and is given as follows:

$$S = (1 - \mathrm{P}_0) \times \beta \tag{7}$$

Similarly, we calculate the aggregate probability of packet Loss (as a measure of blocking probability) for the system by using the following relation:

$$P_B = \sum_j \left[1 - \tfrac{\lambda_j}{\lambda_1}\right] P(j) = \sum_j P(j) - \tfrac{1}{\lambda_1} \sum_j \lambda_j P(j)$$
$$\Rightarrow P_B = 1 - \tfrac{1}{\lambda_1} \sum_j \lambda_j P(j)$$
$$\Rightarrow P_B = 1 - \tfrac{1}{\lambda_1} \sum_j \lambda_j \left[P_{1j} + P_{2j}\right]$$

$$\Rightarrow P_B = 1 - \left[\frac{1}{(\sum_{i=1}^{2} \lambda_{1i})} \left(\sum_{j=0}^{N} \sum_{i=1}^{2} \lambda_{ij} P_{ij} \right) \right] \tag{8}$$

4 NUMERICAL RESULTS

This section presents typical numerical results obtained from the proposed analytical model for Internet traffic congestion control to demonstrate its credibility subject to various parameter settings. These results show the effectiveness of the proposed technique for bursty external traffic at routers in a network. These results demonstrate the effect of threshold settings on various performance measures including mean queue length, throughput, delay, and packet loss probability distribution. This section also includes some experiments to show the impact of traffic burstiness and correlated properties on the system performance.

4.1 System Performance

Figs. 10 - 13 present MQL, throughput, delay and packet loss probability against input parameters and threshold variations respectively. In our numerical examples for the fixed input parameters, we vary N_{12} and N_{22}, where: N_{21} - N_{11} = N_{22} - N_{12}, and N_{12} - N_{11} = N_{22} - N_{21}, such that $N_{21} > N_{12}$. From the results it is clear that as we increase the threshold values, the mean queue length increases, which in turn increases the utilization of the system resulting in high throughput which is of the fact that even though the queue is accumulating packets, they are being served efficiently at the same time.

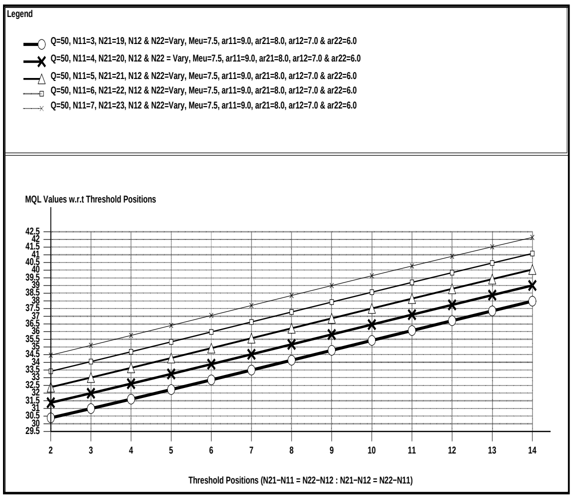

Fig. 10. Effects of Threshold on Mean Queue Length.

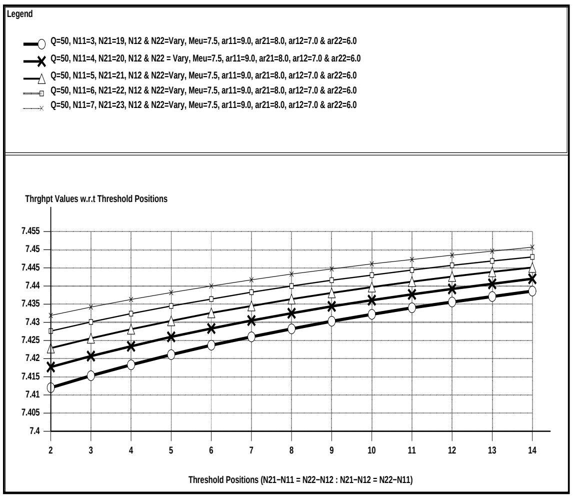

Fig. 11. Effect of Threshold on Throughput.

Also, we may see from figs. 12 and 13 that even though the delay is increasing it can be seen as trade off to the packet loss which in turn is decreasing. Similarly, with increase in threshold a linear reduction is achieved for the number of jobs in the system with a certain probability, which is RED standard. Also, it is evident from results that the mean queue length can be maintained by setting the threshold value in order to prevent congestion.

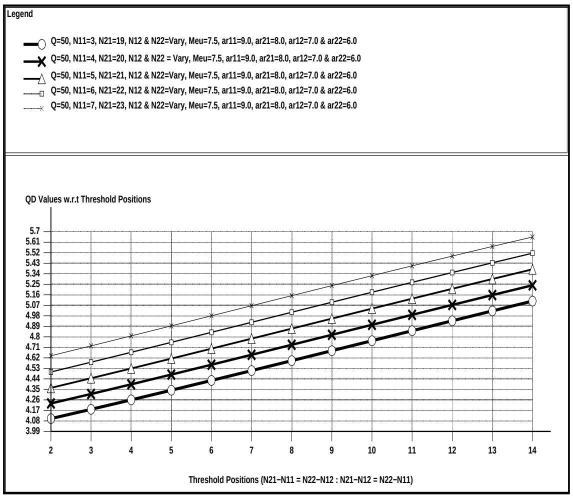

Fig. 12. Effects of Threshold on Delay.

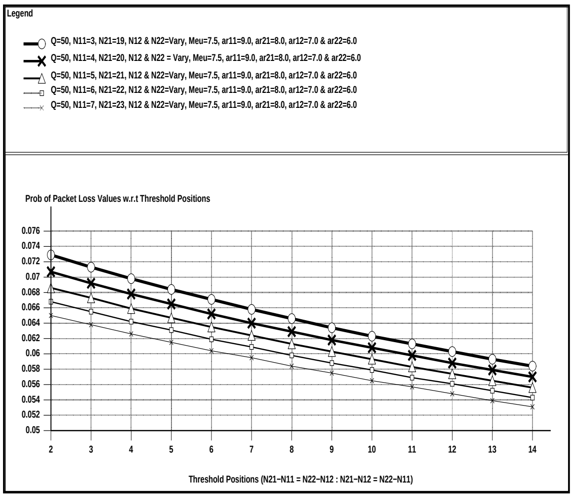

Fig. 13. Effect of Threshold on Probability of Packet Loss.

4.2 Effect of Traffic Burstiness and Autocorrelation Coefficient

The Square Coefficient of Variation (SCV) is a measure of the variability associated with the interarrival and processing times of the system. The variation in SCV value (a function of arrival rate and transition probability in each state) greatly affects the performance of a switch/router and is an important measure of the degree of traffic burstiness in the CT-MMPP traffic source. In our model, the SCV, c^2, of the interarrival time of a two-state CT-MMPP for the packets arrival process is given by the following expression [17]:

$$c^2 = 1 + \frac{2pq\left(\lambda_1 - \lambda_2\right)^2}{\left(q\lambda_1 + p\lambda_2 + \lambda_1\lambda_2\right)\left(p+q\right)^2} \quad (9)$$

The autocorrelation coefficient of the interarrival times and the number of arrivals are the two important measures of interest [17]. The autocorrelation function of the interarrival time of packets with lag 1, *C(1)*, is given by:

$$C(1) = \frac{\lambda_1\lambda_2\left(\lambda_1 - \lambda_2\right)pq}{c^2\left\{p+q\right\}^2\left\{\lambda_1\lambda_2 + \lambda_1 q + \lambda_2 p\right\}^2} \quad (10)$$

Fig 14 shows the effect of load variation (input arrival rate) on SCV of interarrival times. Fig 15 shows the comparison results for the autocorrelation coefficient of the interarrival times with lag 1.

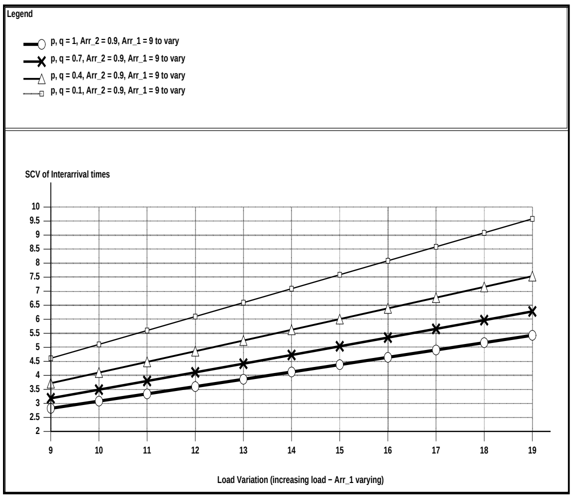

Fig. 14. Effects of load variation on SCV.

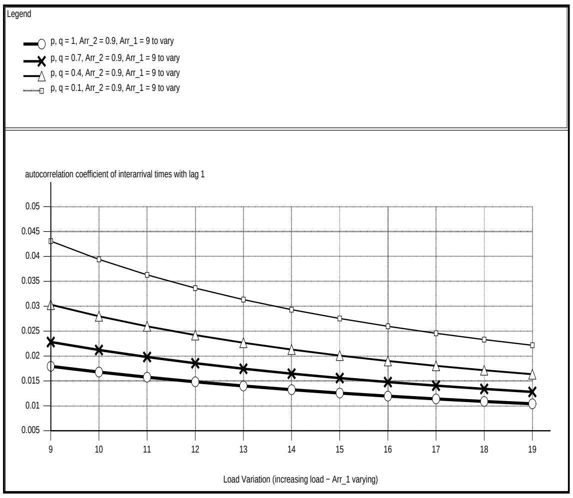

Fig. 15. Autocorrelation comparison for different values of SCV with Lag 1.

Figs. 16 - 18 show that the effect of varying queue size on MQL, throughput and mean delay for different values of SCV. The results show higher burstiness traffic causing higher system MQL,higher mean delay and also higher throughput for the same threshold settings based on the assumption that the total amount of traffic remains constant.

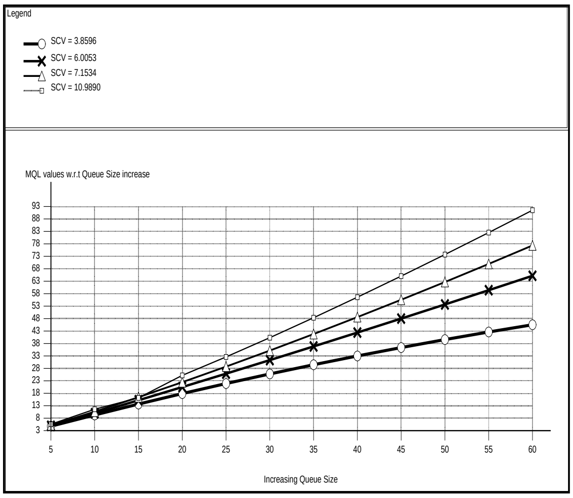

Fig. 16. Effects on MQL by varying Queue Size for different value of SCV.

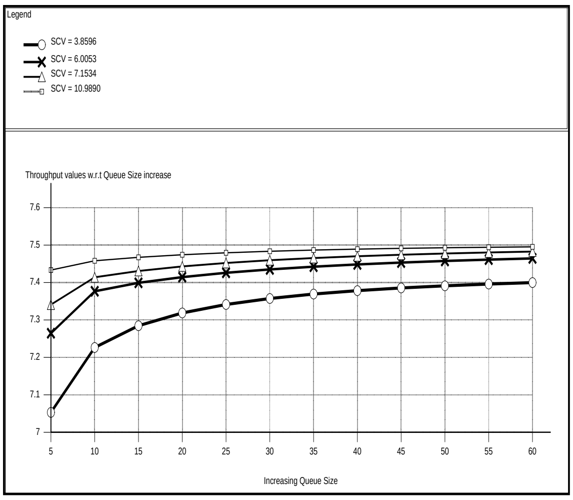

Fig. 17. Effects on Throughput by varying Queue Size for different value of SCV.

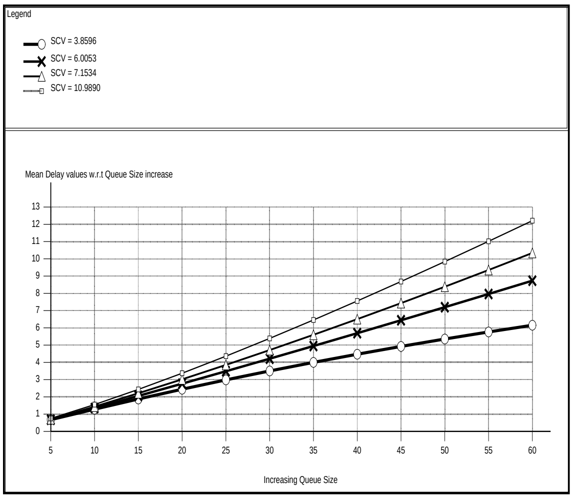

Fig. 18. Effects on Mean Delay by varying Queue Size for different value of SCV.

Figs. 19 - 21 show that the higher burstiness traffic causes higher system MQL and higher mean delay but lower loss probability for the same threshold settings for different SCV values.

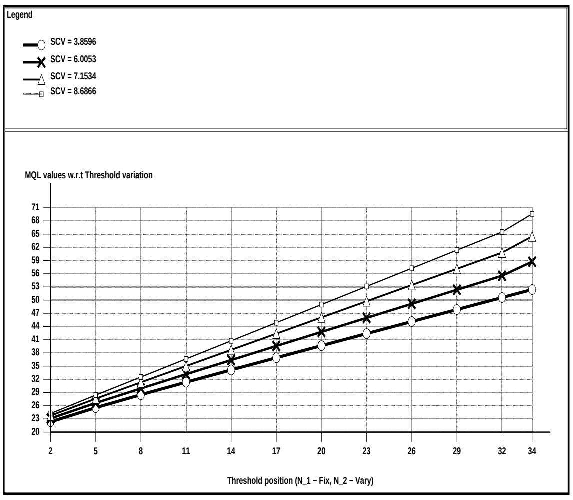

Fig. 19. Effects on MQL by varying threshold for different value of SCV.

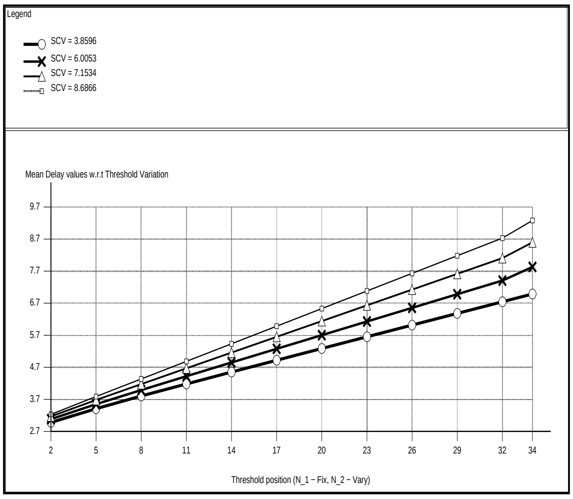

Fig. 20. Effects on Delay by varying threshold for different value of SCV.

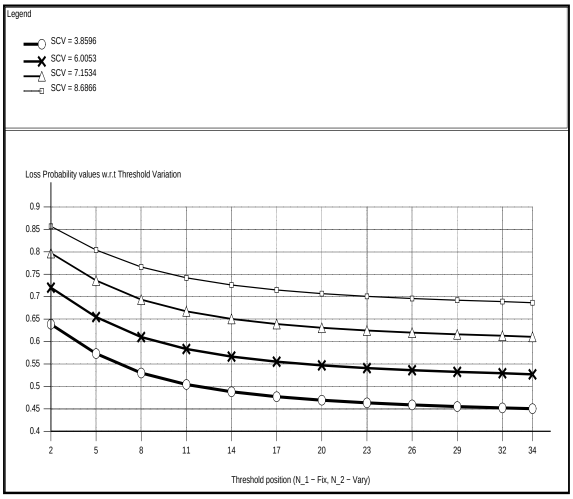

Fig. 21. Effects on Probability of Packet Loss by varying threshold for different value of SCV.

Figs. 22 - 25 show that the higher burstiness traffic causes higher system MQL, higher throughput and higher mean delay but trades offs as reduced loss probability for the same threshold settings variation over different SCV values.

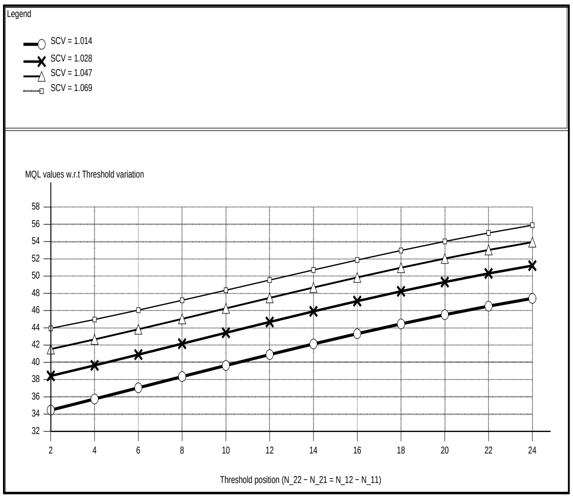

Fig. 22. Effects on MQL by using multiple thresholds for different value of SCV.

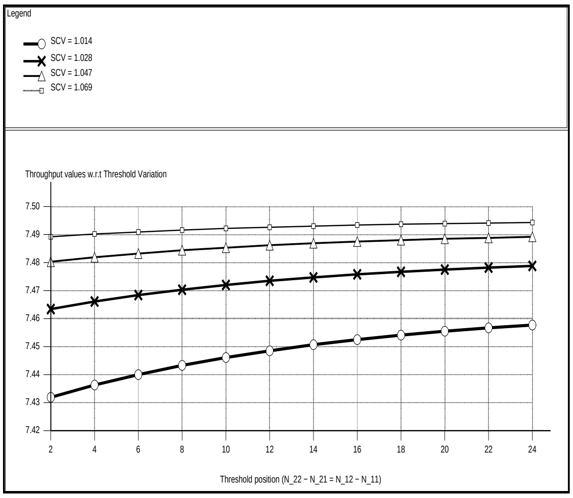

Fig. 23. Effects on Throughput by using multiple thresholds for different value of SCV.

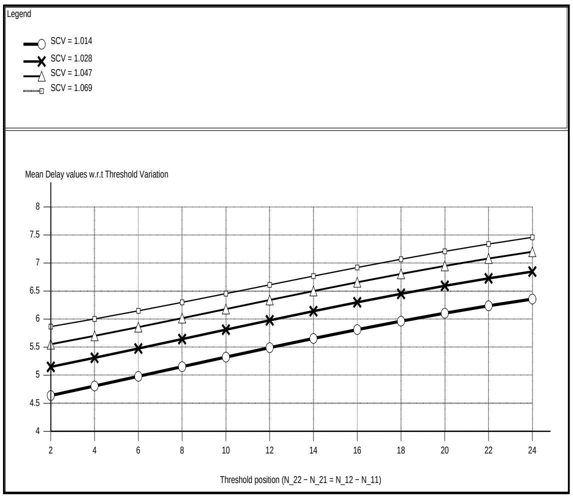

Fig. 24. Effects on Delay by using multiple thresholds for different value of SCV.

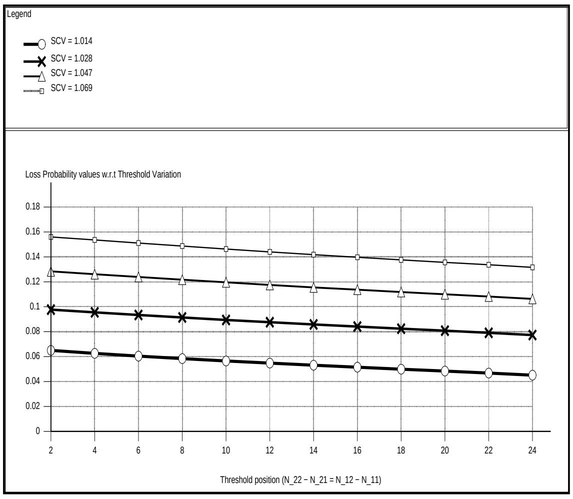

Fig. 25. Effects on Loss Probability by using multiple thresholds for different value of SCV.

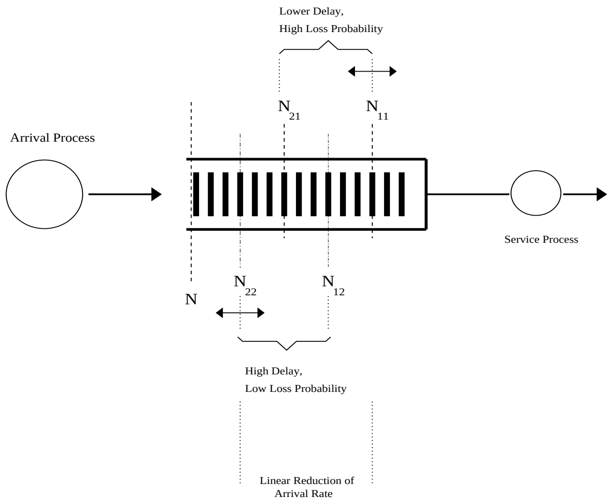

Fig. 26. An example of multiple threshold RED mechanism for marginal effects of threshold variation on QoS for SCV.

4.3 Marginal System Performance

Similarly, some examples are carried out to demonstrate the effect of thresholds on the marginal QoS measurements. Fig 26 demonstrates one such example where we consider four thresholds, two for each job class. In this example, the thresholds N_{11} - min and N_{21} - max correspond to the arrivals of first job class, and N_{12} - min and N_{22} - max correspond to the arrivals of second job class, where N correspond to the full buffer capacity, respectively. We consider two scenarios for same SCV. The full queue capacity is fix at $N = 50$ for both scenarios. Now, we first consider that the threshold N_{12} and N_{22} vary from 11 - 22 and 27 - 38 position while keeping the threshold, $N_{11} = 9$ and $N_{21} = 25$, fixed. Secondly, we fix the threshold N_{12} and N_{22} at 29 and 45 position while varying the threshold $N_{11} = 16 - 27$ and $N_{21} = 32 - 43$ position with the condition that $N_{22} - N_{21} = N_{12} - N_{11}$ and $N_{21} - N_{11} = N_{22} - N_{12}$ where: $N_{21} > N_{12}$ (for both scenarios) . Then the results of marginal effects of threshold positions obtained for most important QoS measures i.e., Delay and Loss Probability are as shown in Figs. 27 and 28.

In Fig 27, we see that as the difference between the threshold values increases the marginal delay increases for the second job class whereas it decreases for the first job class and vice versa. It is because of the fact that as the threshold difference increases for a job class more space is available for the arrivals of a particular job class in the queue and thus packets of a job class will be served more resulting in high throughput for a job class. Similarly, Fig 28 provides the results for the effects of threshold variation on marginal loss probability which decreases for second job class because marginal delay increases for second job class. Consequently, marginal loss probability increases for the first job class because marginal delay decreases for first job class. Thus, we may say that by tuning the threshold position we make an effective use of available resource for bursty and correlated traffic streams containing data, real-time etc as multiple arrivals. Moreover, increase in delay can be seen as a trade off for the reduced packet loss.

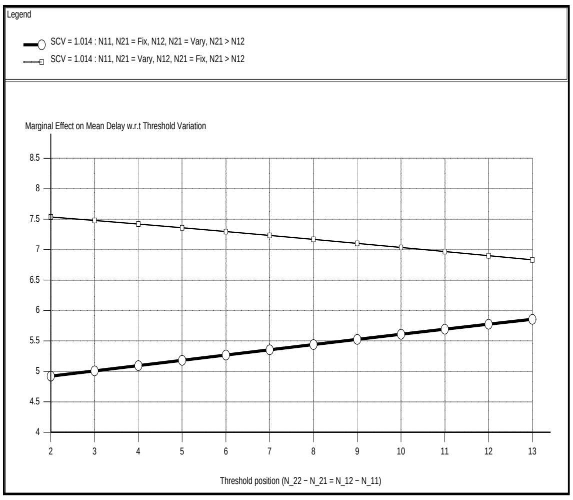

Fig. 27. Effect of Threshold Variation on Marginal Delay for a fixed SCV.

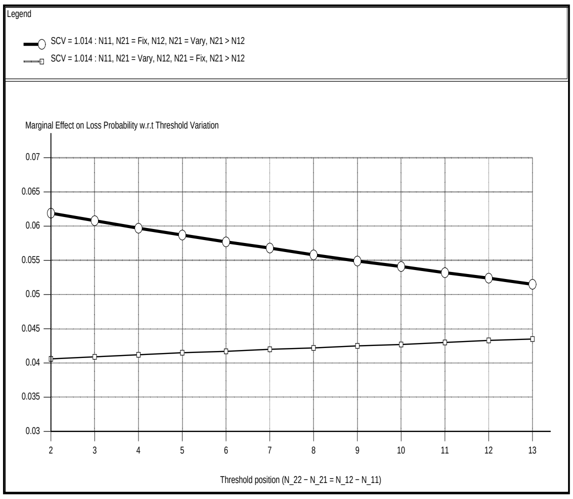

Fig. 28. Effect of Threshold variation on Marginal Loss Probability for a fixed SCV.

5 CONCLUSIONS and FUTURE WORK

A novel approximate analytical performance model of a multiple threshold Random Early Detection (RED) congestion control mechanism for implementing the AQM scheme is presented in this paper. The analysis of finite capacity queue based on Two-state CT-MMPP distribution has been proposed to model the bursty Internet traffic. The traffic source slows down the arrival process once the queue size reaches the maximum threshold, jobs are blocked. Different job loss and QoS requirements under various load conditions can be met by adjusting the threshold values. Also, the effect of threshold based queue on correlated traffic scenarios by introducing correlation to our CT-MMPP model of bursty traffic is demonstrated. Typical numerical examples are included to demonstrate the effects of threshold variation on QoS measures and correlation function of a system.
The future work will involve the application of the proposed model in a simulator environment containing real-time 3G traffic [24] as an input source and evaluating its impact on QoS measures while using a model based on the one proposed in this paper.

References

[1] Jain R and Ramakrishnan, K.K,"Congestion avoidance in Computer Networks with a connectionless Network Layer: Concepts, Goals & Methodology", Proc. IEEE Computer

Networking Symp. Washington D C, April 1988, pp 134-143

[2] Seungwan RYU, Christopher RUMP, Chunming QIAO, "Advances in Internet Congestion Control", IEEE Communications Survey, Third Quarter 2003, Volume 5, No.1, http://www.comsoc.org/pubs/surveys

[3] Jacobson, V., "Congestion Avoidance and Control", In Proceedings of SIGICOMM, Volume 1101, pages 34-45, Stanford, CA, USA, Mar 27-28 1988.

[4] Stevens, W.R., "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", RFC 2001.

[5] J. Nagle, "Congestion Control in IP/TCP Internet works", January 1984-1989, RFC 896.

[6] B. Braden et al, Recommendations on Queue Management and Congestion in the Internet, IETF RFC 2309 April 1998.

[7] S. Floyd and V. Jacobson, Random Early Detection Gateways for Congestion Avoidance, IEEE/ACM Transaction on Networking, Vol 1, No 4, pp 397-413, August 1993.

[8] Asfand-E-Yar, I U Awan and M E Woodward (2004): "Performance Modelling of Congestion Control Mechanism", Research Report RS-10-04, Computing Department, University of Bradford, United Kingdom.

[9] T.Bonal,M.May, and J.C.Bolot, "Analytical Evaluation of RED performance", IEEE INFOCOM 2000.

[10] V. Firoiu and M. Borden, "A Study of Active Queue Management for Congestion Control", http://www.ieee-infocom.org/2000/papers/405.pdf.

[11] S. Kunniyur and R. Srikant, Analysis and Design of an Adaptive Virtual Queue (AVQ) Algorithm for Active Queue Management (AQM), SIGCOMM 2001. pp. 123-134, Aug 2001.

[12] S. Kunniyur and R. Srikant, End-to-End Congestion Control: Utility Function, Random Losses and ECN Marks, Proceedings of Infocom 2000, Tel Aviv, Israel, March 2000.

[13] S. Floyd, R. Gummadi and S. Shenker, "Adaptive RED: An algorithm for Increasing the Robustness of REDs Active Queue Management".

[14] M. Christiansen, Kevin Jeffary, david Ott, F.Donelson Smith, "Tunning RED for Web Traffic", IEEE/ACM Transaction on Networking, Volume 9, No. 3,June 2001, Pager No: 249 - 264.

[15] S. Doran, Interface Graphs, http://adm.ebone.net/sdm/red-1.html

[16] Bartek Wydrowski and Moshe Zukerman, "QoS in Best-Effort Networks", IEEE Communication Magazine, December 2002, 0163-6804/02.

[17] Raif O. Onvural,"Asynchronous Transfer Mode Networks - Performance Issues", Second Edition, ISBN: 0-89006-804-6, Artech House Publishers.

[18] Wolfgang Fischer and Kathleen Meier-Hellstern, "The Markov-modulated Poisson process (MMPP) cookbook", Performance Evaluation, 18(1992) 149-171.

[19] Lin Guan, Irfan-Ullah Awan, Michael E. Woodward, "Stochastic modelling of random early detection based congestion control mechanism for bursty and correlated traffic", IEE Proceedings - Software 151(5): 240-247 (2004).

[20] Asfand-E-Yar, I U Awan and M E Woodward (2005): "Performance Analysis of Congestion Control for Bursty Internet Traffic", 8th International Conference on Computer Modelling and Simulation, United Kingdom Simulation Society (UKSIM 2005), Co-sponsored by the IEEE UK-RI Computer Chapter, 6th - 8th April 2005, St John's College, ISBN: 1-84233-111-6, Oxford, United Kingdom.

[21] Asfand-E-Yar, I U Awan and M E Woodward (2005): "Analysis of Threshold Based Queue with MMPP Arrival Process", Systems Communications 2005 (ICW 2005, ICHSN 2005, ICMCS 2005, SENET 2005), Aug 14-17, 2005, Montreal, Canada, IEEE Computer Society, ISBN: 0-7695-2422-2.

[22] Marcel F. Neuts, "Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach", ISBN: 0-486-68342-7 (1994).

[23] Randolph Nelson, ""Matrix Geometric Solutions in Markov Models: A Mathematical Tutorial", Research Report, IBM Research Division, RC 16777 (74293) 4/22/91, Computer Science, 24 Pages.

[24] European Telecommunications Standards Institute (ETSI), Universal Mobile Telecommunications System (UMTS): Selection Procedures for the choice of radio transmission technologies of the UMTS (UMTS 30.03 Version 3.1.0), Technical Report, TR 101 112 V3.1.0 (1997 - 11), DTR.SMG-023003U.

BIOGRAPHY

Mr Asfand-E-Yar graduated in Electrical Engineering (Telecommunication) in 1999 from N.W.F.P University of Engineering and Technology (UET) Peshawar (PAKISTAN) in First Division (HONOURS). Mr Yar was awarded degree of MSc in "Communication and Real-time Electronic Systems" with major in Mobile Communications on 6 Dec 2001, at University of Bradford, UK. Over a number of years, Mr Yar has been involved working on various projects in the Telecom Industry such as Traffic Modelling, Quality of Service (QoS) issues in High speed Networks and Mobile Networks, Mobile Network Planning and Optimization, Network Management etc. Also, he has been working in the Telecom Regulatory environment where he was involved with De-regulation Policy and Regulations frame work. In August 2003, Mr Asfand-E-Yar joined the University of Bradford, Computing Department as a Research Assistant to work on an EPSRC funded project. Later, in April 2005, he was appointed as Research Fellow and continued his work towards the award of a Doctorate in Telecommunication Networks and Performance Engineering. Mr Asfand-E-Yar, is a Member of Institute of Engineers Pakistan (IEP), Member of IEEE, IEE and British Computer Society (BCS).

Irfan Awan received his BSc from Gomal University, Pakistan (1986), MSc (Computer Science) from Qauid-e-Azam University, Pakistan (1990). He served for three years as a lecturer in the Computer Science Department, BZ University Pakistan. Irfan Awan received his PhD degree ('97) in Performance Analysis of Queueing Network Models with priorities and blocking from the University of Bradford - UK. He is a senior lecturer in the Department of Computing, University of Bradford which he joined in 1999. He is member of the Mobile Computing and Networks Research Group and co-tutor for the MSc Mobile Computing Course. Dr. Awan's research has mainly focussed on developing cost effective analytical models for measuring the performance of complex queueing networks with finite capacities and priorities. He has produced over 100 publications and edited proceedings of the 20th UK Performance Engineering Workshop. He has also authored several special issues of the international journals and is a member of various programme committees and steering committees for International conferences. He is a member of IEEE, ILT, BCS and UK Simulation Society.

Professor Mike Woodward graduated with a first class honours degree in Electronic and Electrical Engineering from the University of Nottingham in 1967 and received a PhD degree from the same institution in 1971 for research into the decomposition of sequential logic systems.

In 1970 he joined the staff of the Department of Electronic and Electrical Engineering at Loughborough University as a lecturer, being promoted to Senior Lecturer in 1980 and Reader in Stochastic Modelling in 1995. He remained at Loughborough until 1998 when he was appointed to the Chair in Telecommunications at the University of Bradford where he also became the Director of the Telecommunications Research Center. He is currently the Head of the Department of Computing at the University of Bradford. His current research interests include queueing networks, telecommunications traffic modelling, quality of service routing and mobile communications systems and he is the author of two books and over one hundred research papers on the above and related topics. He is supervisor to fourteen full time research students. Professor Woodward is a Fellow of the Institute of Mathematics and its Applications (FIMA) and is a Chartered Mathematician (CMath), Chartered Scientist (CSci) and a Chartered Engineer (CEng).