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).

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