UNIVERSITI PUTRA MALAYSIA

PREDICTIVE QUALITY OF SERVICE SCHEMES FOR REAL-TIME MULTIMEDIA APPLICATIONS IN COMMUNICATIONS

SHAMALA SUBRAMANIAM

FSKTM 1999 5

PREDICTIVE QUALITY OF SERVICE SCHEMES FOR REAL-TIME MULTIMEDIA APPLICATIONS IN COMMUNICATIONS

By

SHAMALA SUBRAMANIAM

Thesis Submitted in Fulfilment of the Requirements for the Degree of Master of Science in the Faculty of Computer Science and Information Technology Universiti Putra Malaysia

March 1999

This thesis is dedicated to
My dearest Amma, Aka and Sue .....

ACKNOWLEDGEMENTS

It is my pleasure to extend my heartfelt gratitude to my supervisor Prof. Madya Dr. Ashwani Kumar Ramani. He has blessed upon me and many others the desire and appreciation towards acquiring knowledge. His dedication, concern, patience and wisdom towards his profession have given light to many individuals life. His relentless pursue towards perfection has produced research of absolute quality. It is truly his guidance and dedication that has enabled me to complete my work.

I am also grateful to Dr Mohd Yazid Saman and Miss Shyamala Doraisamy my co-supervisors for their active encouragement. They provided valuable advises and insight for my research. I would also like to thank Dr. Abd. Azim Abdul Ghani, Dean, Faculty of Computer Science and Information Technology and other faculty members for the support given to complete my research.

I would like to thank my mother who has constantly given me the confidence to pursue my dreams. She has been a wonderful mother and her dreams to ensure her children's success has always wanted me to strive for the best. I would also like to thank my two sisters, who have always given me the joy of being the youngest.

Date : February 1999

Shamala Subramaniam

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI	Average Interval
AR	Average Rate
ARPA	Advanced Research Projects Agency
ATM	Asynchronous Transfer Mode
CSE	Current Status Environment
DMS	Distributed Multimedia Systems
DTE	Data Terminal Exchange
FCFS	First-Come-First-Serve
FIFO+	First-In-First-Out+
GUI	Graphical User Interface
HoL	Head-of_Line
HoL-PJ	Head-of-Line with Priority Jumping
ISO	International Standard Organisation
LAN	Local Area Network
MLT	Minimum Laxity
OCP_A	OCcuPancy_Adjusting
PSE	Packet Switching Exchange
QoS	Quality of Service
TDM	Time Division Multiplexing
VBR	Variable Bit Rate

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirements for the degree of Master of Science.

PREDICTIVE QUALITY OF SERVICE SCHEMES FOR REAL-TIME MULTIMEDIA APPLICATIONS IN COMMUNICATIONS

By

SHAMALA SUBRAMANIAM

March 1999

Chairman : Associate Professor Ashwani Kumar Ramani, Ph.D.

Faculty : Computer Science and Information Technology

In guaranteed QoS resources are reserved based on the worst-case analysis. This scheme guarantees QoS but results in low resource utilisation, subsequently depriving other users from acquiring service. In contrary, predictive QoS allocates an initial amount of resource, whereby QoS is guaranteed with a certain probability of degradation. This scheme is tailored for applications that are adaptive and robust towards sudden fluctuations in the service provided. The nature of multimedia data such as the variable bit rate has encouraged the implementation of predictive Quality of Service (QoS) as compared to guaranteed QoS.

In a typical QoS scheme, two factors contribute to the computation of packet loss, which are : (i) a new packet dropped due to buffer overflow and (ii) a buffered packet dropped due to expired delay. The buffer resource is increased when the observed packet loss ratio has violated the requested level of packet loss ratio. The limitations in this scheme is caused by the admission of all packets into

the buffer, inclusive of packets that will eventually be dropped due to expired delay. Subsequently, this results in poor resource management.

In this research, two pro-active dynamic QoS control schemes are designed, the dynamic QoS control scheme with delay estimation, and the hybrid dynamic QoS control scheme. In both schemes, every new packet arrival is compared against the estimated delay it will experience, before being admitted into the buffer. If the estimated delay expires the requested delay bound, then the packet is dropped. In the hybrid scheme, every packet is checked before being admitted into the buffer, and also, the packets successfully admitted into the buffer are evaluated on the actual delay experienced before being transmitted to the receiver.

The results obtained through the simulation models have shown that two schemes have significantly improved the average delay for different traffic patterns. In addition to improving the average delay in delay sensitive traffic, improvement is seen in the average packet loss ratio, and subsequently increasing the throughput for delay sensitive traffic. However, in packet loss sensitive traffic, the old scheme remains beneficial. The proposed scheme can be adopted for multimedia applications to enhance the QoS in terms of better delay and improved utilisation.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

SKEMA KUALITI PERKHIDMATAN JANGKAAN UNTUK APLIKASI MULTIMEDIA DALAM KOMUNIKASI

Oleh

SHAMALA SUBRAMANIAM

Mac 1999

Pengerusi : Profesor Madya Ashwani Kumar Ramani, Ph.D.

Fakulti : Fakulti Sains Komputer dan Teknologi Maklumat

Ciri-ciri yang dikaitkan dengan data multimedia, terutamanya ketidakstabilan dari segi transmisi, telah menggalakan penggunaan skema kualiti jangkaan berbanding dengan skema kualiti yang terjamin. Dalam skema kualiti terjamin, peruntukan sumber adalah berdasarkan kes yang memerlukan jumlah sumber yang terbanyak. Namun begitu, jumlah sebenar sumber yang digunakan adalah kurang daripada jumlah yang telah diperuntukan. Justerunya, peruntukan sumber seumpama ini, akan menjejaskan penggunaan sumber secara efektif. Sebaliknya, skema kualiti jangkaan, akan mengaggihkan jumlah minima sumber yang diperlukan untuk memenuhi permintaan pengguna terhadap kualiti transmisi.

Skema kualiti jangkaan adalah sesuai untuk applikasi multimedia yang dapat menanggani kemerosotan dalam kualiti perkhidmatan. Memandangkan, bahawa keadaan rangkaian adalah tidak stabil dan sukar untuk dijamin, skema jangkaan secara dinamik diperlukan. Dalam skema ini, dua faktor mempengaruhi

perangkaan paket yang disingkirkan. Faktor pertama adalah apabila paket yang baru perlu disingkirkan kerana ruang memori yang penuh dan faktor kedua adalah apabila sesebuah paket yang perlu di storankan sebelum ditransmisi kepada penerima telah melebihi jumlah kelambatan penghantaran yang ditetapkan. Dalam skema ini, didapati bahawa penggunaan ruang memori adalah tidak berkesan. Ini disebabkan tidak terdapat kawalan terhadap paket yang dimasukkan ke dalam memori. Dalam kajian ini, masalah ini telah diatasi, dengan dua skema jangkaan dinamik yang pro-aktif. Skema pertama adalah skema dinamik dengan jangkaan kelambatan penghantaran dan skema dua adalah model hibrid. Dalam kedua-dua skema ini, sebelum setiap paket dimasukkan ke dalam memori, setiap paket perlu melalui satu ujian jangkaan kelambatan penghantaran. Sekiranya, jumlah ini melebihi jumlah yang ditetapkan, maka paket tersebut disingkirkan.

Hasil kajian yang dijalankan melalui kaedah simulasi, telah membuktikan bahawa kedua model pro-aktif yang disarankan, telah meghasilkan bukan sahaja pengurangan dari segi kelambatan penghantaran tetapi telah memperbaiki penggunaan memori secara berkesan. Maka, skema yang telah dicadangkan amat bersesuaian untuk data multimedia yang peka terhadap kelambatan penghantaran.

CHAPTER I

INTRODUCTION

Trends such as "Information Superhighway" and " Information Society" have emerged in the past decade [1]. Such trends have encouraged research developments in various IT fields especially multimedia-based projects. These projects are conducted by the industry, academia and government [2]. The synergy between the advances of computer and communication technologies will lay a foundation for designing various multimedia systems [3]. The simultaneous needs of workgroups in Local Area Network (LAN) have resulted in the evolution of Distributed Multimedia Systems (DMS). DMS is a generic term that describes a model of information, where data sources are located separately from data sinks [4]. The integration of DMS in conventional networks presents many challenges. Two of which are the data communications networking infrastructure and the higher importance placed on the issue of Quality of Service (QoS). The subsequent sections present a detail discussion on these issues.

Data Communications Networking Infrastructure

Data communications networks constitute of two main types: packet switching networks and circuit switching networks. In circuit switching networks, each connection results in a physical communication channel being set up through the network from calling to the called subscriber equipment [5]. The resulting connection provides a path with a fixed data rate at which both subscribers must operate depending upon the nature of the equipment used in each exchange. An important feature of circuit switching networks is the need to set-up an end-to-end path before any data can be sent. Thus, prior to data transmission the called signal must propagate all the way to the destination and be acknowledged. The elapsed time between the end of dialling and the start of ringing can consume a substantial amount of time, subsequently, resulting in long set-up times. Many of the real time applications find this long set-up times undesirable [6].

In contrast to circuit switching, packet switching network permits two communicating subscribers to operate at two varying data rates. The term packet was created by the Advanced Research Projects Agency (ARPA) research team to distinguish between long messages that form the input into a network and the blocks that they are broken into prior to transmission across a network. This expression in turn led to the term packet switching being given to a particular type of communication processor [7]. In packet switching networks all the data transmitted is first assembled into one or more message units (packets) by the source data terminal. These packets comprise of the source and destination Data

Terminal Exchange (DTE's) network addresses. Then they are passed bit serially by the source DTE to it's local Packet Switching Exchange (PSE). The PSE, upon receipt of each packet, first stores the packet and then checks the destination address it contains [5]. Each of these PSE, contains a routing directory specifying the outgoing link(s) to be used for each network address. The PSE forwards the packet on the appropriate link at the maximum available bit rate. This mode of operation is referred to as packet store-and-forward.

An important feature of loading data into packets, is the ability for different traffic flows belonging to the different end-to-end sessions to statistically share a common communication line. Typically, packets are queued and scheduled for transmission may be on a First-Come-First-Serve (FCFS) basis. Packet switching is a network extension of statistical multiplexing, with data from individual sources segmented into small packets that are stored and forwarded from switch to switch towards their destination [8]. The statistical smoothing of variable rate sources is achieved as packets are buffered in the network switches. The consequences of the buffering stages result in variable delays being introduced. In addition, data losses may be introduced due to buffer overflow. Thus, it is important to select parameters to minimise these effects while maintaining the efficiency of statistical averaging [5]. The introduction of packet switching in the early 1970s, was mainly for the bursty computer communications that could tolerate delay, and was also tailored for specific types of traffic [9]. However, with the growing popularity of the Internet, there has been a rapid rise in the integration of different traffic types, such as real-time audio and video traffic etc.

Even though circuit switching can be used to support performance guarantees associated with real-time data, the current technology trend is towards the use of packet switching for all types of traffic [10]. An example is in the Asynchronous Transfer Mode (ATM) networks, where video, audio and data are all carried in 53 byte cells. The conventional data architectures were well suited for low-speed networks that employed traditional data applications such as file transfer, e-mail and Telnet, which required services with scalable (elastic) delay requirements. However, the applications with real-time properties proved that the conventional architectures were unable to support them. Thus, there is requirement to redesign the architectures that can manage real-time data, as per the constraints.

This chapter is organised further in the following order. The next section presents the design and features of conventional data architectures. The properties of real-time data are discussed in subsequent section along with the features of an enhanced architecture to support these real-time properties. One of the important elements of our research is QoS, and the next section is dedicated to introduce and discuss the QoS concept. The objective of the research and the thesis organisation subsequently follow this section.

Conventional Data Architectures

Conventional data architectures were designed for traditional data applications that could permit variable delays. In packet switching networks, each communication channel is shared among many traffic flows. Packets are queued and scheduled for transmission generally in a FCFS basis. The service scheme in traditional architectures employs best-effort service. In best-effort service [11], the service scheduler tries to maximise the amount of traffic, which can meet the delay requirements. However, under high load, the traffic unable to be scheduled on time is dropped. The reason that contributes to the packet drop in best-effort services is the limitations in applications to specify the minimal amount of resources needed to satisfy a given quality of service. The ideal operation desired by resource reservation schemes is to allocate the minimum resources to an application that is needed to satisfy its QoS requirements. However, it is very difficult and wasteful to achieve such reservations. Thus, the Internet and other packet switched networks offer best-effort services based upon total resource sharing. In this case a source does not need to notify the network on its QoS requirement and no resources are reserved for any particular flow. This causes a wide variation in the actual delay and bandwidth to be experienced by the applications. The result of implementing no admission control increases the possibility of the occurrence of congestion. The responsibility of overcoming the congestion is given to the end systems. Although, best-effort services are unable to provide fully guaranteed services to flows but they have supported high resource utilisation.

Data applications require error free transmission emphasising less on stringent timing requirements. Low speed packet switched networks have employed the window –based flow control mechanisms for traffic control [12]. The window -based flow control system uses returned acknowledgements both to initiate loss recovery and to regulate further data transmissions. These acknowledgements incur a rather large and variable delay.

The integration and execution of multimedia applications in the Internet have discovered that the conventional data architecture is unable to support and facilitate these newly emerging applications. Much of these limitations are attributed to the properties associated to the real-time traffic.

Multimedia Real Time Traffic

The deployment of multimedia applications presents new challenges to the current networking technologies and data service architectures. In contrast to data traffic that is typically sporadic and unpredictable, multimedia real time applications impose stringent performance requirements. Also the various DMS applications may have differing requirements. Thus, the services in the multimedia systems must be parameterised rather than performed on a best-effort basis. Parameterisation allows the flexibility and customisation of the services, so that each new application does not require implementing a new set of services. Most of the current QoS parameters differ from the parameters described by the International Standard Organisation (ISO) because of the variety of applications, media sent and the quality of networks and the end-systems. This section describes the parameters that constitute in DMS: delay, jitter, synchronisation, packet loss ratio and throughput.

Delay

Digital audio and video streams consist of sequence of media quartes, such as video frames or audio samples, which convey message only when presented continuously in time. The presence of large packet delay variations causes playback points to be postponed, resulting in loss of fidelity. Different applications have varying resilience towards delay. Most multimedia applications require an explicit delay bound on their transmitted packets because of delay sensitivity. Delay in the network is due to several causes. Delay caused by the propagation of the packet at the speed of light and the delay of transmission at each switch point waiting for the entire packet to arrive before commencing the next stage of transmission, are classified as the fixed component associated to delay. Queuing is the fundamental consequence of the statistical sharing that occurs in packet networks [13]. The variable time spent by each packet in the service queue in the switches define the variable component associated to delay. An important factor of delay is that the delay bound varies from one application to another application. Delay bound is defined as the maximum network delay experienced by the packets of an application. The management of packets for meeting respective delay bounds is an important issue.

Network clients may differ between one another. One way to distinguish these clients is the manner in which delay is addressed. There are two possible ways in which delay is managed [10]. First, applications utilise the priori delay bound advertised by the network to set the playback point. This point is fixed, regardless of the actual delay experienced. In contrast, for adaptive applications, the system will measure the network delay experienced by arriving packets and then tune the playback point to the minimal delay that still produces sufficiently low loss rate[14]. Another method to distinguish network clients is the level of tolerance towards brief interruptions in service. Clients can be tolerant towards such irregularities such as in the case of a non-interactive application as compared to a tele-robotics application that is intolerant, as such fluctuations causing the malfunctioning of a remote robot may be detrimental towards the operation. Based upon these two distinctions, network clients can be classified into two main groups: rigid or adaptive and tolerant or intolerant as indicated in Figure 1 [15].

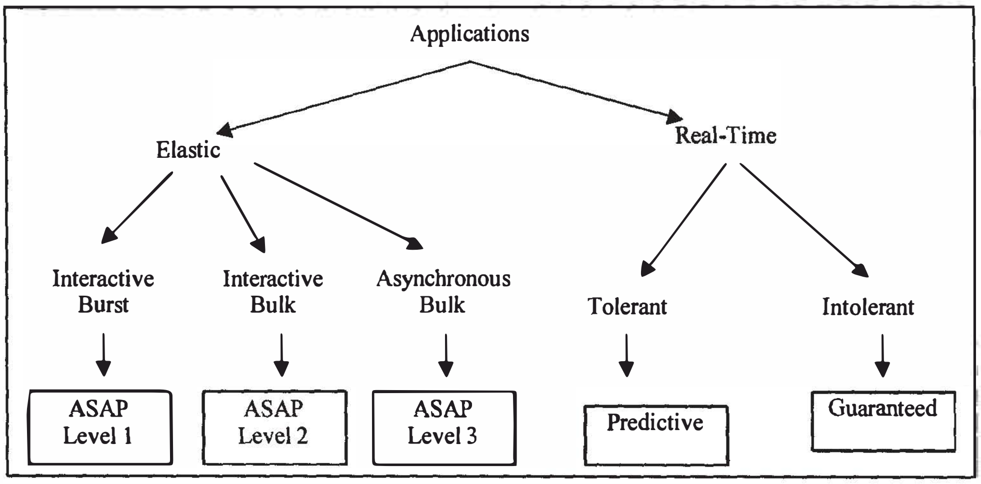

Figure 1.1: Classification of Network Clients Based on Delay Management

These characteristics associated to the network clients determine the type of service rendered to the client. There are four different types of services: guaranteed, statistical, predictive and best-effort services. Table 1.1 [16] displays the degrees of QoS and the respective resource utilisation and QoS reliability. Guaranteed services provide QoS guarantees (Table 1.1), as specified through the

QoS parameter value (bounds) in either a deterministic or statistical representation. The deterministic bound may be given through a single value (e.g. average value, contractual value, threshold value, target value), a pair of values (e.g. minimum and average value, lowest quality and target quality) or an interval of values (e.g. minimum and average value, lowest quality and target quality). Guaranteed services may also deal with statistical bounds of QoS parameters [3], such as statistical bound on error rate. Worst -case traffic analysis and receiving sufficient resources to provide requested QoS under the worst network condition is practised in guaranteed services. An example is reserving resources according to the peak rate of the source can provide guaranteed QoS. The trade-off in guaranteed service is between high QoS reliability and low resource utilisation.

The second degree of QoS in Table 1.1, is the statistical QoS and is achieved by studying the statistical behaviour of the traffic by means of formal analysis. It prevents the QoS degradation from falling below a deterministic level. Realisation of guaranteed or statistical QoS requires modelling of the sources hence present limitations as to the supported source types. In statistical QoS, deadlines are guaranteed to be met with a certain probability. To provide these guarantees, admission control algorithms must consider the stochastic behaviour of the system while admitting new clients [15].

A predictive service (history service) is based on the past network behaviour. Thus, the QoS parameters are estimates of past behaviour, which the service tries to match. This service does not require detailed modelling of the