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

<table>
<thead>
<tr>
<th>ACKNOWLEDGEMENTS</th>
<th>iii</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS</td>
<td>xi</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>xii</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>xiv</td>
</tr>
</tbody>
</table>

## CHAPTER

### I

#### INTRODUCTION

1. Data Communications Networking Infrastructure ............. 2
2. Conventional Data Architectures .................................. 4
3. Multimedia Real Time Traffic .................................. 6
   - Delay ........................................................................... 7
   - Packet Loss Ratio .................................................. 10
   - Synchronisation ..................................................... 11
   - Jitter ......................................................................... 12
   - Throughput ............................................................ 12
4. QoS Concept ............................................................ 14
5. Resource Concept ..................................................... 15
6. The Flow Concept ..................................................... 16
7. Research Objectives .................................................. 17
8. Organisation of Thesis .............................................. 19

### II

#### LITERATURE REVIEW

1. QoS Issues ............................................................... 23
2. QoS Specification ...................................................... 24
3. QoS Mechanisms .......................................................... 26
4. QoS Management Mechanisms ........................................ 33
5. QoS Control Mechanisms ............................................. 35
6. QoS Control Based Scheduling Algorithms ...................... 37
7. Dynamic QoS Control Scheme ....................................... 44
8. Summary ......................................................................... 47

### III

#### METHODOLOGY AND SIMULATION DESIGN

1. Simulation and Modelling ............................................. 53
2. Discrete Event Simulation ............................................ 53
   - Model Formulation .................................................. 55
   - Data Preparation and Model Translation ....................... 56
   - Event Scheduler Routine .......................................... 57
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Different Degrees of QoS</td>
<td>10</td>
</tr>
<tr>
<td>3.1 Advantages and Disadvantages of General Purpose Languages and Special Purpose Languages</td>
<td>57</td>
</tr>
<tr>
<td>4.1 QoS Tuple Configurations</td>
<td>79</td>
</tr>
<tr>
<td>4.2 List Of Variables Used In The Resource Sub-Module</td>
<td>82</td>
</tr>
<tr>
<td>5.1 Traffic Pattern Configurations</td>
<td>95</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Classification of Network Clients Based on Delay Management</td>
<td>8</td>
</tr>
<tr>
<td>1.2</td>
<td>QoS-Layered Model for the DMS</td>
<td>15</td>
</tr>
<tr>
<td>2.1</td>
<td>QoS Parameters derived during QoS Mapping</td>
<td>31</td>
</tr>
<tr>
<td>2.2</td>
<td>QoS Control based Scheduling Algorithms</td>
<td>39</td>
</tr>
<tr>
<td>2.3</td>
<td>Single Flow OCCupancy_Adjusting (OCP_A) Algorithm</td>
<td>47</td>
</tr>
<tr>
<td>3.1</td>
<td>Steps to Ensure a Systematic Approach to Performance Evaluation</td>
<td>50</td>
</tr>
<tr>
<td>3.2</td>
<td>The Simulation Steps and Inter-Relationship</td>
<td>54</td>
</tr>
<tr>
<td>3.3</td>
<td>The Event Scheduler Routine</td>
<td>58</td>
</tr>
<tr>
<td>3.4</td>
<td>The Simulation Clock Advancing Mechanism</td>
<td>60</td>
</tr>
<tr>
<td>3.5</td>
<td>The Arrival Event Routine</td>
<td>62</td>
</tr>
<tr>
<td>3.6</td>
<td>The Departure Event Routine</td>
<td>62</td>
</tr>
<tr>
<td>4.1</td>
<td>The Dynamic QoS Control Algorithm</td>
<td>71</td>
</tr>
<tr>
<td>4.2</td>
<td>Delay Estimation Algorithm</td>
<td>73</td>
</tr>
<tr>
<td>4.3</td>
<td>The Dynamic QoS Control Algorithm with Delay Estimation</td>
<td>74</td>
</tr>
<tr>
<td>4.4</td>
<td>The Hybrid Dynamic QoS Control Algorithm</td>
<td>76</td>
</tr>
<tr>
<td>4.5</td>
<td>Simulation Model</td>
<td>78</td>
</tr>
<tr>
<td>5.1</td>
<td>Average Delay against Arrival Rate for Model 1</td>
<td>101</td>
</tr>
<tr>
<td>5.2</td>
<td>Average Allocated Buffer against Arrival Rate for Model 1</td>
<td>101</td>
</tr>
<tr>
<td>5.3</td>
<td>Average Packet Loss Ratio against Arrival Rate for Model 1</td>
<td>102</td>
</tr>
<tr>
<td>5.4</td>
<td>Throughput against Arrival Rate for Model 1</td>
<td>102</td>
</tr>
</tbody>
</table>
5.5 Average Delay against Arrival Rate for Model 2

5.6 Average Allocated Buffer against Arrival Rate for Model 2

5.7 Average Packet Loss Ratio against Arrival Rate for Model 2

5.8 Throughput against Arrival Rate for Model 2

5.9 Average Delay against Arrival Rate for Model 3

5.10 Average Allocated Buffer against Arrival Rate for Model 3

5.11 Average Packet Loss Ratio against Arrival Rate for Model 3

5.12 Throughput against Arrival Rate for Model 3

5.13 Average Delay versus Arrival Rate for Traffic Pattern 1

5.14 Average Delay against Arrival Rate for Traffic Pattern 2

5.15 Average Delay against Arrival Rate for Traffic Pattern 3

5.16 Average Delay against Arrival Rate for Traffic Pattern 4

5.17 Average Packet Loss Ratio against Arrival Rate for Traffic Pattern 1

5.18 Average Packet Loss Ratio against Arrival Rate for Traffic Pattern 1 for Model 1

5.19 Average Packet Loss Ratio against Arrival Rate for Traffic Pattern 1 for Model 2

5.20 Average Packet Loss Ratio against Arrival Rate for Traffic Pattern 1 for Model 3

5.21 Average Packet Loss Ratio against Arrival Rate for Traffic Pattern 2

5.22 Average Packet Loss Ratio against Arrival Rate for Traffic Pattern 3

5.23 Average Packet Loss Ratio against Arrival Rate for Traffic Pattern 4
5.24 Average Allocated Buffer against Arrival Rate for Traffic Pattern 1
5.25 Average Allocated Buffer against Arrival Rate for Traffic Pattern 2
5.26 Average Allocated Buffer against Arrival Rate for Traffic Pattern 3
5.27 Average Allocated Buffer against Arrival Rate for Traffic Pattern 4
5.28 Average Allocated Buffer against Arrival Rate for Traffic Pattern 1
5.29 Average Allocated Buffer against Arrival Rate for Traffic Pattern 2
5.30 Average Allocated Buffer against Arrival Rate for Traffic Pattern 3
5.31 Average Allocated Buffer against Arrival Rate for Traffic Pattern 4
5.32 Throughput against Arrival Rate for Traffic Pattern 1
5.33 Throughput against Arrival Rate for Traffic Pattern 2
5.34 Throughput against Arrival Rate for Traffic Pattern 3
5.35 Throughput against Arrival Rate for Traffic Pattern 4
# LIST OF ABBREVIATIONS

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

This thesis presents an analysis of Predictive Quality of Service (QoS) schemes for real-time multimedia applications in communications. The thesis is structured as follows:

1. **Introduction**: This section provides an overview of the research problem and the objectives of the study.
2. **Literature Review**: A comprehensive review of existing research on QoS in communications, focusing on multimedia applications.
3. **Methodology**: Details of the methodology used in the study, including the experimental setup and data collection process.
4. **Results**: Presentation of the findings and analysis of the data collected.
5. **Discussion**: Interpretation of the results and their implications for real-time multimedia applications.
6. **Conclusion**: Summary of the research findings and their significance, along with recommendations for future work.

The thesis concludes with a discussion on the implications of the findings for real-time multimedia applications, highlighting the potential benefits of Predictive QoS schemes over Guaranteed QoS schemes. The research contributes to the understanding of QoS management in multimedia communications, providing insights for improving the efficiency and reliability of real-time services.
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.

Skema kualiti jangkaan adalah sesuai untuk aplikasi 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

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 its 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 PSEs 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 service. The service scheduler 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](image)

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