A Quality of Service Monitoring System for Service Level Agreement Verification

Xiaoyuan Ta

A thesis submitted in fulfilment of the
requirements for the award of the degree of

MASTER OF ENGINEERING BY RESEARCH

SCHOOL OF ELECTRICAL AND INFORMATION ENGINEERING
THE UNIVERSITY OF SYDNEY

March 2006

Abstract

Service-level-agreement (SLA) monitoring measures network Quality-of-Service (QoS) parameters to evaluate whether the service performance complies with the SLAs. It is becoming increasingly important for both Internet service providers (ISPs) and their customers. However, the rapid expansion of the Internet makes SLA monitoring a challenging task. As an efficient method to reduce both complexity and overheads for QoS measurements, sampling techniques have been used in SLA monitoring systems.

In this thesis, I conduct a comprehensive study of sampling methods for network QoS measurements. I develop an efficient sampling strategy, which makes the measurements less intrusive and more efficient, and I design a network performance monitoring software, which monitors such QoS parameters as packet delay, packet loss and jitter for SLA monitoring and verification.

The thesis starts with a discussion on the characteristics of QoS metrics related to the design of the monitoring system and the challenges in monitoring these metrics. Major measurement methodologies for monitoring these metrics are introduced. Existing monitoring systems can be broadly classified into two categories: active and passive measurements. The advantages and disadvantages of both methodologies are discussed and an active measurement methodology is chosen to realise the monitoring system.

Secondly, the thesis describes the most common sampling techniques, such as systematic sampling, Poisson sampling and stratified random sampling. Theoretical analysis is performed on the fundamental limits of sampling accuracy. Theoretical

analysis is also conducted on the performance of the sampling techniques, which is validated using simulation with real traffic. Both theoretical analysis and simulation results show that the stratified random sampling with optimum allocation achieves the best performance, compared with the other sampling methods. However, stratified sampling with optimum allocation requires extra statistics from the parent traffic traces, which cannot be obtained in real applications. In order to overcome this shortcoming, a novel adaptive stratified sampling strategy is proposed, based on stratified sampling with optimum allocation. A least-mean-square (LMS) linear prediction algorithm is employed to predict the required statistics from the past observations. Simulation results show that the proposed adaptive stratified sampling method closely approaches the performance of the stratified sampling with optimum allocation.

Finally, a detailed introduction to the SLA monitoring software design is presented. Measurement results are displayed which calibrate systematic error in the measurements. Measurements between various remote sites have demonstrated impressively good QoS provided by Australian ISPs for premium services.

Acknowledgments

First and foremost, I would like to express my appreciation to my supervisor, Dr Guoqiang Mao, for his kind advice and support during my study. He has not only supplied an interesting research topic but also made many contributions and valuable comments. Moreover, he patiently read and marked up my draft thesis and gave me valuable suggestions that significantly improved the quality of this thesis. I have learned a lot from him, especially on how to be a good researcher—being rigorous, energetic and dedicated to research.

I would also like to thank Mr Lixiang Xiong and Mr Zhuo Chen, for their kind help and suggestions for revising this thesis. I would also like to thank people in the telecommunication research group, for their constant and ongoing generous help.

I would like to express my gratitude to my family. My parents and my sister always give me support, care and love. My wife's parents always give me understanding and encouragement. Their passion for careers and their attitudes to life are always role models for me. I owe a special gratitude to my wife Ying Shan. Her constant love, support and inspiration have made my Master's study such an enjoyable and rewarding experience.

The work presented in this thesis has been supported by the contracted research project "BLO No. 7260" from Optus. I am grateful to Optus for providing financial support for my study.

Contents

Chapter 1

Introduction

1.1 Background

Internet Service Providers (ISPs) now offer service level agreements (SLAs) routinely to their customers. Management needs contractual guarantees that business objectives are met, and end-users demand assurance that their critical network applications and services are available when needed. The availability of SLAs and a means to validate them gives management the confidence to move ahead. The wide adoption of the E-business model has made it essential that service-providers deliver on SLAs in a quantitative and qualitative manner. This has driven the service-providers to seek consistent testing and measurement methods that make real sense of customer network performance.

An SLA is defined by the International Telecommunications Union (ITU) as "a negotiated agreement between a customer and the service provider on levels of service characteristics and the associated set of metrics. The content of SLAs varies depending on the service offering and includes the attributes required for the negotiated agreement" [1]. The Internet Engineering Task Force (IETF) defines SLAs in a similar way [2]. Figure 1.1 shows the main features of the SLAs.

Generally speaking, a good SLA should include these three key aspects:

- Service level objectives: encompass Quality-of-Service (QoS) parameters or

class of service provided, service availability and reliability, authentication issues, SLA expiry date, and so on.

- Service measuring components: specify the way of measuring service quality and other parameters used to assess whether the service complies with the SLA.

- Financial compensation components: include billing options, penalties for breaking the contract, and so forth.

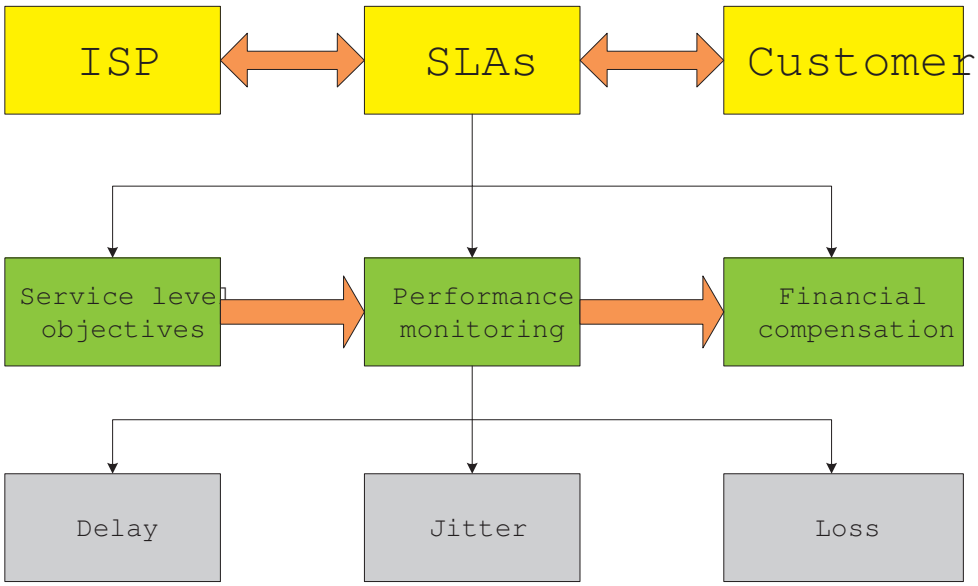

Figure 1.1 Structure of service-level-agreements

SLA monitoring is an important part of SLA management. It is useful for both network operators and individual customers, who want to check whether the service performance indeed complies with the SLAs. Moreover, the ability to measure against key performance indicators facilitates the continuous quality improvement process. It helps the ISPs to locate the bottleneck in their networks. A service performance problem becomes an opportunity to structurally improve overall service quality and customer satisfaction.

1.2 Research Motivation and Contribution

SLA monitoring is about collecting statistical metrics about network performance to evaluate whether the provider complies with the level of QoS that the customer expects [3]. Therefore, accurate measurement and estimation of network performance becomes a key challenge in SLA monitoring. However, the implementation of measurement becomes increasingly difficult and complex due to the rapid expansion of the Internet. Traditional measurement tools, such as “ping”, cannot satisfy the measurement requirements nowadays. Moreover, the dramatic increase in the speed of wide area backbones presents obstacles to complete statistics collection. The enormous amount of measurement data may significantly increase the cost and resource usage [4].

In order to solve these problems, sampling techniques are employed in SLA monitoring systems to reduce the quantity of control data and resources required to process it, and finally to reduce the measurement complexity and cost. Systematic sampling and random sampling are two widely used methods in existing monitoring systems, but both of them have severe limitations. Stratified random sampling can achieve higher estimation accuracy, but its high complexity may compromise its advantages.

The aim of this research project, which has been funded by Optus through the research contract “BLO No. 7260”, is to develop an efficient sampling strategy to make the measurement less intrusive and more efficient. Then a network performance monitoring software, which monitors such QoS parameters as packet delay, packet loss and jitter for SLA monitoring and verification, and which uses the proposed sampling strategy, needs to be designed. These objectives have been fully achieved. Firstly, a theoretical analysis of the performance of different sampling techniques (both count-based and timer-based) is presented. Secondly, a novel adaptive stratified sampling strategy is developed and validated. Finally, QoS monitoring software is delivered at the end of the project, which has been highly rated by Optus. This thesis provides a comprehensive summary of the outcome of the project.

1.3 Thesis Outline

This thesis consists of seven chapters, the rest of which are organised as follows:

Chapter 2 presents a comprehensive review of related work. Firstly, I describe the main usages of Internet measurement, and the standard metrics for measurement as defined by the IETF's IP Performance Metrics Working Group (IPPM). Secondly, I discuss in detail the characteristics of QoS metrics related to the design of the monitoring system in this project, i.e., packet delay, packet loss and jitter, and challenges in monitoring these metrics. Thirdly, I introduce the major methodologies of network performance measurement, including both passive measurement and active measurement, as well as their advantages and disadvantages.

Chapter 3 describes major sampling techniques that can be used in the sampling-based monitoring system, such as systematic sampling, random sampling, stratified random sampling and adaptive sampling. Discussion of the fundamental limit (i.e., minimum sample size required for a given confidence level and an error bound) of the accuracy of sampling techniques is then presented.

Chapter 4 presents a theoretical analysis of the performance of two fundamental sampling techniques, i.e., systematic sampling and random sampling, and compares their performance. Autocorrelation ρ of packet delay of the parent delay trace is used as a factor in the performance comparison between time-based systematic sampling and time-based Poisson sampling. ρ is also used to determine the stratification boundaries for stratified sampling. Simulation results using real traffic trace provided by the WAND group is presented to validate the theoretical analysis.

Chapter 5 proposes an adaptive stratified sampling strategy for SLA monitoring, which is based on the stratified sampling with optimum allocation discussed in Chapter 4.4.2. Although stratified sampling with optimum allocation can achieve a satisfactory accuracy of estimation, it has severe imitations. The stratified sampling with optimum allocation requires extra statistics (e.g., standard deviation of packet delay, total number of packets) of the parent trace to determine the stratum sample size. In real applications, these statistics are not known *a priori*. To address the challenge,

a novel adaptive sampling method is proposed, which employs a least-mean-square (LMS) algorithm to predict the standard deviation of packet delay from past observations. The sample size for the next stratum is calculated from the predicted standard deviation. Sampling results that show good performance are presented.

Chapter 6 provides a detailed introduction to the monitoring software design. I start with an introduction to the software environment and functionality. A description of the procedure of the TCP measurement, UDP measurement and ICMP measurement is then presented. The systematic error of the software is calibrated. Finally, I introduce the software's graphic-user-interface (GUI) design and demonstrate several test results in real networks.

Chapter 7 concludes this thesis by providing a summary of my major contributions. The direction for future study is also discussed.

Chapter 2

Literature Review

2.1 Characteristics of QoS Metrics

2.1.1 Main Usages of Internet Measurements

2.1.2 Standard Metrics

2.1.3 Packet Delay

2.1.3.1 One-way Delay Measurement

2.1.3.2 Round-trip Delay Measurement

2.1.4 Jitter

2.1.5 Packet Loss

2.1.5.1 Bernoulli Loss Model

2.1.5.2 Two-state Markov Chain Model

2.1.5.3 n-th order Markov Chain Model

2.1.5.4 Extended Gilbert Model

2.2 Network Measurement Methodology

2.2.1 Passive Measurement

2.2.2 Active Measurement

2.3 Summary

Chapter 3

Sampling Techniques

3.1 Introduction

3.2 Sampling Techniques

3.2.1 Systematic Sampling

3.2.2 Random Sampling

3.2.3 Stratified Sampling

3.2.4 Adaptive Sampling

3.2.5 Sampling Trigger

3.3 Accuracy of Sampling

3.4 Summary

Chapter 4

Performance Comparison of Different Sampling Techniques

4.1 Introduction

4.2 Delay Traffic Trace

4.3 Systematic Sampling *vs* Random Sampling

4.3.1 Comparison between Count-based Systematic Sampling and Count-based Simple Random Sampling

4.3.1.1 Simulation Results

4.3.2 Comparison between Timer-based Systematic Sampling and Timer-based Poisson Sampling

4.3.2.1 Simulation Results

4.4 Stratified Sampling *vs* Simple Random Sampling

4.4.1 Stratified Sampling with Proportional Allocation

4.4.2 Stratified Sampling with Optimum Allocation

4.5 Impact of Packet Size on Delay Measurements

4.6 Summary

Chapter 5

Adaptive Stratified Sampling

5.1 Introduction

5.2 Least-mean-square Algorithm

5.3 Adaptive Stratified Sampling Algorithm

5.3.1 Prediction of Sample Size within Strata

5.3.2 Prediction Error

5.3.3 Stratification Boundaries

5.4 Simulation Results

5.5 Summary

Chapter 6

Monitoring Software Design

6.1 Introduction

6.2 Software Environment

6.2.1 IP Precedence Setting

6.2.2 Software Platform

6.2.3 Network Programming Interface

6.3 Software Functionality

6.4 Measurement Using ICMP, UDP and TCP Protocols

6.4.1 ICMP Measurement

6.4.2 UDP Measurement

6.4.3 TCP Measurement

6.5 Accuracy of Time Measurement

6.6 Multi-thread and File Management

6.7 GUI Design

6.8 Summary

Chapter 7

Conclusion

7.1 Future Study

Bibliography

[1] ITU-T, "Support of ip-based services using ip transfer capabilities," *Tech. Rep. Rec. Y.1241*, 2001.

[2] S. Blake, D. Black, M. Carlson, E. Divies, Z. Wang, and W. Weiss, "An architecture for differentiated services," *IETF RFC 2475*, 1998.

[3] C. Molina-Jimenez, S. Shrivastava, J. Crowcroft, and P. Gevros, "On the monitoring of contractual service level agreements," in *Proceedings of the First International Workshop on Electronic Contracting (WEC'04)*, April, 2004.

[4] K. Claffy, G. Polyzos, and H.-W. Braun, "Application of sampling methodologies to network traffic characterization," *ACM SIGCOMM Computer Communication Review*, vol. 23, no. 4, pp. 194–203, 1993.

Appendix A

Mathematical Derivation

A.1 Derivation of PDF of the sum of m consecutive inter-arrival time slots of the Poisson process

A.2 Derivation of comparison between the variance of the systematic sample mean and the variance of the Poisson sample mean

A.2.1 Sufficient condition

A.2.2 Necessary condition

A.3 Derivation of the transformation of σ^2 in terms of the autocorrelation function

A.4 Derivation of the difference of the variance of the sample mean between optimum allocation and proportional allocation

A.5 Derivation of the relative error between $Var_{opt}(\bar{y})$ and $Var_{act}(\bar{y})$