Using Network Application Behavior to Predict Performance

by

Chunling Ma

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

April 2008

APPROVED:

Professor Craig E. Wills, Thesis Advisor

Professor Robert E. Kinicki, Thesis Reader

Abstract

Today's continuously growing Internet requires users and network applications to have knowledge of network metrics. This knowledge is critical for decision making during the usage of network applications. This thesis studies application related network metrics. The major approach in this work is to examine the traffic between a simulated user and network applications. We use the historical data collected from previous usage of network applications to make predictions for future usage of those applications. We also use the historical data obtained from a given application to make predictions about another application. Prediction mechanisms require us to make parameter choices so that certain weights can be placed on historical data versus current data. We study these different choices and use the values from our best experimental results. From these studies we conclude that our data prediction is quite accurate and remains stable over a range of parameter choices. The use of shared routing paths between users and network applications are explored in the performance prediction of applications. Only some servers at the same locations show similar prediction results. The network applications studied are also varied, including web, streaming, DNS, etc. We see whether sharing information obtained from different applications can be used to make predictions of application performance. However, we observe limited success in predictions across applications.

Acknowledgments

I would like to extend my sincere gratitude to my thesis advisor Professor Craig E. Wills. He has helped and inspired me all along in completing my thesis work. He has also exemplified to me what entails to do academic research — hard work, open-mindedness, scientific attitude, and perseverance. These have become invaluable assets in my life. His technical advice and insights have been always helpful. I thank him for all the guidance and especially for showing great patience with me.

I thank Professor Robert E. Kinicki for being the reader for my thesis and providing useful comments.

I would like to express my cordial thanks to my friend Abhishek Kumar from WPI for his long-term support both technically and mentally. He helped me in searching clusters with special types of servers all over the United States. Without his help, I would not have collected the data used for this work. He has also continuously encouraged me whenever I was frustrated and intimidated, and helped me overcome many technical obstacles in finishing this work.

I also thank my follow PEDS friends, Hao Shang, Mingzhe Li, and Feng Li, for supplying me the right tools, good advice and thought-provoking conversations.

My savior Jesus Christ has been always with me in the ups and downs of my life. I thank Him for His unchanging love and miraculous healing so that I could come back soon to continue to finish this work. My sincere thanks also go out to many prayers from sisters and brothers at the Chinese Gospel Church in Worcester, MA and at the Grace Brethren Church in Mansfield, OH.

I dedicate my work to my late parents, whose never-say-die spirit has been a

tremendous inspiration for me to finish this elongated work. They had personally set a great example for me. I thank them for their unconditional love and support, and all the precious values they taught me that have become part of my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As the Internet continues to grow today, more users and applications are involved. Users and applications need to make decisions regarding expected performance of the network. Accurate prediction of metrics is critical for the decision making in these network-based applications. For example, in the selection of a peer server during FTP or web-based file transfer, overall throughput is a major selection criterion. Round trip time (RTT), available bandwidth and rates of packet loss are all useful in a streaming media application when deciding the quality of video or audio to be sent to the user. In various degrees, machines at a local cluster can share network performance information obtained from a remote server, different servers in a remote cluster, or even different clusters. For a local cluster of many users, it can be useful to build a database to store network information inferred from previous access of network applications and use this information to make predictions for performance of future user access of network applications.

There are various factors that affect the accuracy of prediction. Overall, we wish to answer the following questions in this research:

1. How well can previous measurements be used for predicting future measurements? How frequently do these measurements need to be taken in order to make accurate predictions?

2. How can topological similarities between network paths to servers be used in making predictions?

3. How effectively can information inferred from one network application be used to predict application performance of another?

In order to fully explore the answers to these questions, we set up experiments to periodically send out active probes to network applications. By analyzing the received packet trace of each connection, we are able to infer the RTT, available bandwidth, potential packet loss, overall throughput, and generate a summary for each connection.

To answer the first question, we need to find a way of utilizing information obtained from old connections to network application servers. Exponential decay predictors are used to make predictions based on historical data, since they are computationally inexpensive and achieve good prediction results. The factors of data recency and quantity are considered. For example, what weight should be assigned to a network metric inferred from a connection fifteen minutes ago, or two hours or even eight hours ago, as well as how to combine them to achieve the best prediction accuracy? We use different time intervals in our exponential predictor, to identify the significance of data of different ages for the prediction accuracy. Choices of parameters in the exponential decay predictors are investigated and the best experimental values are used for each network metric.

To answer the second question, we study the variation in the metrics of the same

kind of applications at different locations. The correlation between these variations tells us how effectively shared topologies can be used for predictions.

To answer the third question, different network applications are exploited. They include the most common Internet usages: web application, streaming media application, large FTP transfers, DNS requests, ping and traceroute. Information inferred from a connection set up to one network application is studied to see how much of it can be shared by another application which is of the same kind or even different.

The rest of the thesis is organized as follows: Chapter 2 is related work; Chapter 3 talks about the background of this work; In Chapter 4 we discusses preliminary work that was done for this research; Chapter 5 discusses the new experiment environment and data collection; as well as studies the variation of the measurements collected; In Chapters 6 to 8 we analyze the results obtained from these experiments to show how predictions can be made; Chapter 9 summarizes conclusions and Chapter 10 points at future work.

Chapter 2

Related Work

In this chapter we discuss some of the work done by other researchers related to that of this thesis. Related work has been done in the area of performance prediction as well as performance modeling and inference. In Section 2.1 we discuss different approaches that have been developed to make predictions of network metrics. In Section2.2 we briefly discuss mathematical models used in performance prediction.

2.1 Performance Prediction

Part of the inspiration of doing this research came from [13, 14]. [13] proposed the idea of using shared passive probes to make predictions of network performance. The design architecture of a system to implement this idea was also proposed. It was suggested to build a database to collect historical data and use it for predicting future performance. [14] was its follow-up work which mainly studied HTTP object downloading, and used average values to make predictions of further network accesses. The proposed approach in this work is also mainly passive. However, [13, 14]

did not try to understand the accuracy of predictions and the factors that could affect this accuracy. Their work did not take into account the age of measurements. It also did not consider the use of shared routing paths to predict application performance. Their work only used the Web and not other network applications. All this is taken into consideration in our approach of making predictions.

[4] digitalized network performance into 0's or 1's, signifying degradation or non-degradation, for a small time period. It made predictions based on digitalized values and proposed exponential decay, polynomial-decay, VW-cover and hidden-markov predictors as the mathematical models for prediction. We borrowed the exponential decay model in our predictor model. There were also other works studying predictions for one single network metric. For example, [12] used mathematical models ARMA and MMPP stochastic processes to predict bandwidth. [6] used the Amherst model to predict throughput given average RTT, time-out duration, and loss rate. [9] also used a passive approach to estimate TCP RTTs.

2.2 Performance Modeling and Inference

A great amount of study has been done in performance modeling and inference. In our work, we integrated some of the useful results from previous research. [10] proposed a mathematical model presenting the correlation between throughput and other metrics including round trip times and loss rate. [5] modeled round trip times in different phases of a typical TCP connection. [11] inferred the TCP version used in a TCP connection by analyzing the packets sent and received. Different TCP versions can have different congestion control behavior, which in turn affects the measurement and prediction of throughput and available bandwidth. [8] modeled

TCP behavior through passive measurements, and [7] built a model showing the correlations between bandwidth and throughput.

2.3 Summary

In this chapter we saw the attempts by earlier researchers to make network metric predictions. We also saw the work of other researchers related to performance modeling and inference. We use some of this work as a basis for this thesis. At the same time we address the aspects that were not covered in these works. In the next chapter we discuss the background of our work. We take a close look at TCP windowing mechanism and explain our choice of predictor model.

Chapter 3

Background

In this section, we discuss background for the methodology used to collect the data and the model of the predictor we use for predictions.

3.1 TCP Windowing

Before we start collecting the data, it is important to understand the inner workings of the TCP protocol. One of the important aspects of TCP that interests us is the congestion avoidance algorithm.

TCP uses a congestion window (cwnd) in the sender side to do congestion avoidance. The congestion window indicates the maximum amount of data that can be sent out on a connection without being acknowledged. The sender is allowed to increase the congestion window either according to the Slow Start algorithm, that is, by one segment for each incoming acknowledgment (ACK), or according to Congestion Avoidance, at a rate of one segment in a round-trip time. The slow

start threshold (ssthresh) is used to determine whether to use Slow Start or Congestion Avoidance algorithm. When the connection first starts, TCP performs Slow Start, doubles its cwnd in each round trip time. When cwnd reaches ssthresh, TCP performs Congestion Avoidance. Most servers use TCP NewReno nowadays. In NewReno, when the sender receives three duplicate ACKs, it performs fast retransmission by sending the packet before a time-out. At the same time, the sender halves its cwnd and ssthresh, and starts performing Congestion Avoidance. When the sender experiences a time-out for a packet sent, its ssthresh is halved but its cwnd is reduced to one and performs Slow Start like the connection just started.

Since TCP uses the above mentioned congestion avoidance mechanism, packets are observed to be sent in bursts in each round trip time. That is, only the number of cwnd packets can be sent before an acknowledgment is received, and the server stalls for the rest of the time. The server stalls after cwnd packets are sent in an RTT because usually it takes shorter time to send cwnd packets than the round trip time provided the bandwidth is relatively large and the congestion window size is relatively small. In this work, we call this observed burst of sending packets in groups a round. Usually separate rounds are clearly observed at the beginning of a connection when the cwnd has not grown too large in Slow Start.

In this work, most of the applications we use are web applications. Therefore, we can only examine the incoming packets at a receiver. Since packets are sent in rounds, they are received in rounds as well. Separating rounds of received packets gives us RTT measurements using the difference between the time stamp of the first packet in each nth round and that in the (n+1)th round. We can also obtain the available bandwidth in each round using pairs of consecutively received packets. This can be significantly larger than the overall throughput of the connection since

the sender has various stall times in rounds due to the TCP windowing. Another benefit of knowing the inner working of TCP windowing is that we can use the initial value of cwnd, ssthresh, and available bandwidth and RTT that we measured to predict how long a transfer takes, given an object size.

3.2 Predictor Model

Exponential predictors have been the most used predictors since they are easy to implement and do not require extra state keeping. So in our work, we use exponential predictors to predict RTTs and throughput. It is shown as below:

$$M_{pred}(n+1) = \lambda M_{pred}(n) + (1-\lambda)M(n)$$

where $M(n)$is the nth measurement we obtained from the data, $M_{pred}(n)$is the predicted value for the nth measurement, and the parameter $0 < \lambda < 1$. In our algorithm, we set $M_{pred}(1) = M(1)$, and predictions are valid starting from $M_{pred}(2)$. The value of λ gives weights for previous measurements, and it is additive over all previous measurements. A predictor with λ closer to 0 puts more emphasis on the most recent measurements, and a predictor with λ closer to 1 emphasizes the quantity of previous measurements. The best value of λ depends on the data and can be determined experimentally.

3.3 Summary

In this chapter we took a close look at the TCP windowing mechanism. We saw how it influenced our methods of measuring network metrics such as RTT and available

bandwidth. For our prediction methods we decided to use the exponential predictor. Now that we have our methods in place, we discuss the experiments conducted using these methods in the following chapters.

Chapter 4

Preliminary Work

This chapter describes the preliminary work carried out in the summer of 2004. The data was mainly collected for web applications.

4.1 Experimental Setup

In the summer of 2004, we used 33 web servers from 13 clusters, located all over the continental United States. See Table 4.1. They were mostly the servers of popular news channels, popular newspapers and state governments. Some web sites had their DNS servers and web servers in the same cluster, while some did not.

In this preliminary work, only web retrievals were made from a Linux machine located at Worcester Polytechnic Institute at Worcester, MA to each of the web servers at a five-minute interval for a time period of seven days. The machine had SuSE Linux 2.4 kernel. TCP stacks were modified at the kernel level, and fields were added to the system call *getsockopt()* to track out-of-order received packets,

Table 4.1: 33 Web Servers at 13 Remote Clusters

Cluster Name	Location	DNS Server	Web Servers
Boston Globe	Boston, MA	N/A	www.boston.com weather.boston.com realestate.boston.com
MBTA	Boston, MA	ns1.itg.net	eagles.mbta.com csisw1.mbta.com
NY Times	New York, NY	ns1t.nytimes.com	www.nytimes.com movies.nytimes.com www.nytco.com
MSN	Redmond, WA	N/A	www.msn.com shopping.msn.com mobile.msn.com groups.msn.com
DJC	Seattle, WA	lp1.djc.com	www.djc.com
CNN	Atlanta, GA	N/A	www.cnn.com edition.cnn.com si.cnn.com money.cnn.com
Augusta Chronicle	Atlanta, GA	znet.groupz.net	www.augustachronicle.com ap.augustachronicle.com
IL Government	Springfield, IL	ns1.state.il.us	www.illinois.gov www.dnr.state.il.us www.kidcareillinois.com
Ameriquest	Los Angeles, CA	ns1.accads.com	www.ameriquestmortgage.com careers.ameriquest.com customers.ameriquest.com
San Francisco	Freemont, CA	ns1.blvds.com	www.sanfrancisco.com mail.sanfrancisco.com
CA Government	Sacramento, CA	N/A	democrats.assembly.ca.gov republican.assembly.ca.gov capitolmuseum.ca.gov
Dallas News	Dallas, TX	ns1.belo.com	www.dallasnews.com signin.dallasnews.com
LA Times	Log Angels, CA	N/A	www.latimes.com

potential packet losses, and duplicate ACKs sent. *getsockopt()* was the main system call used to collect statistics at the packet level for each web retrieval.

4.2 Variation of Network Metrics

In this section, we show the variations of different network metrics over time obtained from web-based applications. The metrics used were Round Trip Time(RTT), potential packet loss, available bandwidth and throughput.

We found that for the same server, both minimum RTTs and median RTTs within a connection, varied in different degrees depending on the particular server of the connection. Server processing times, available bandwidth, and overall throughput also varied from server to server. Some connections tended to be more stable than others, which means old RTTs from previous access to a server can provide different accuracies for network application performance prediction.

4.2.1 RTTs

We show the time series graphs and corresponding CDF (Cumulative Distribution Function) graphs of minimum RTTs of six remote clusters in Figures 4.1 through 4.12. These clusters include www.boston.com cluster (Boston, MA), www.nytimes.com cluster (New York City), www.cnn.com(Atlanta, GA), www.msn.com(Redmond, WA), www.illinois.gov (Springfield, IL), and www.ameriquestmortgage.com cluster (Los Angeles, CA).

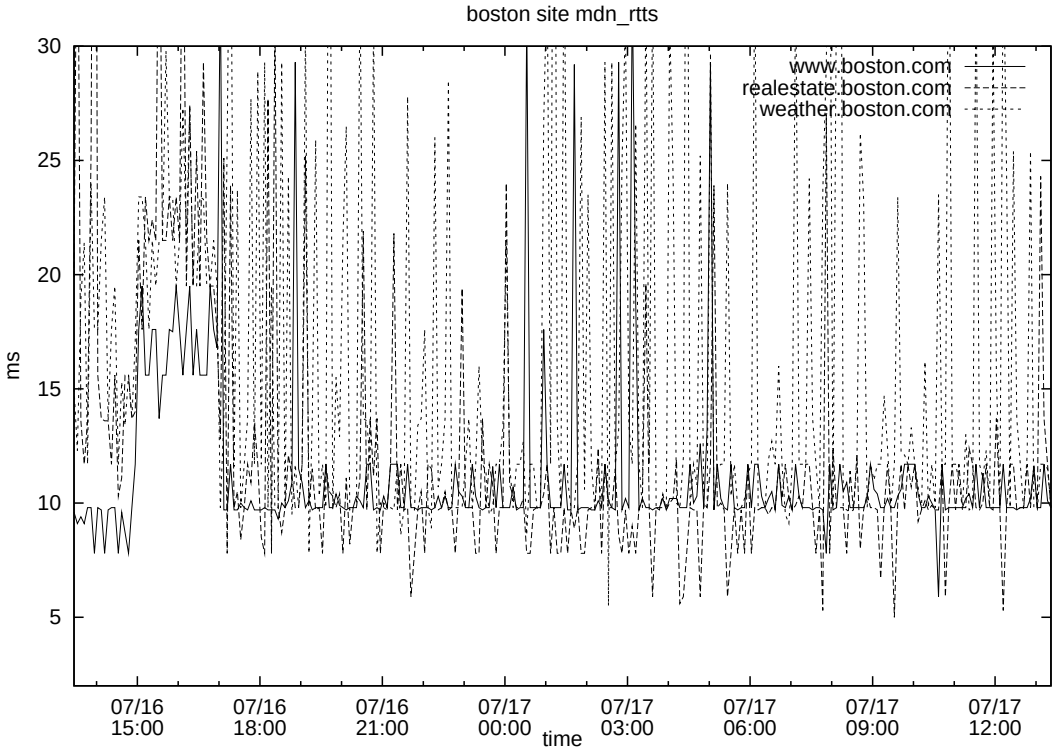

Figure 4.1: Median RTTs of Boston Globe Cluster

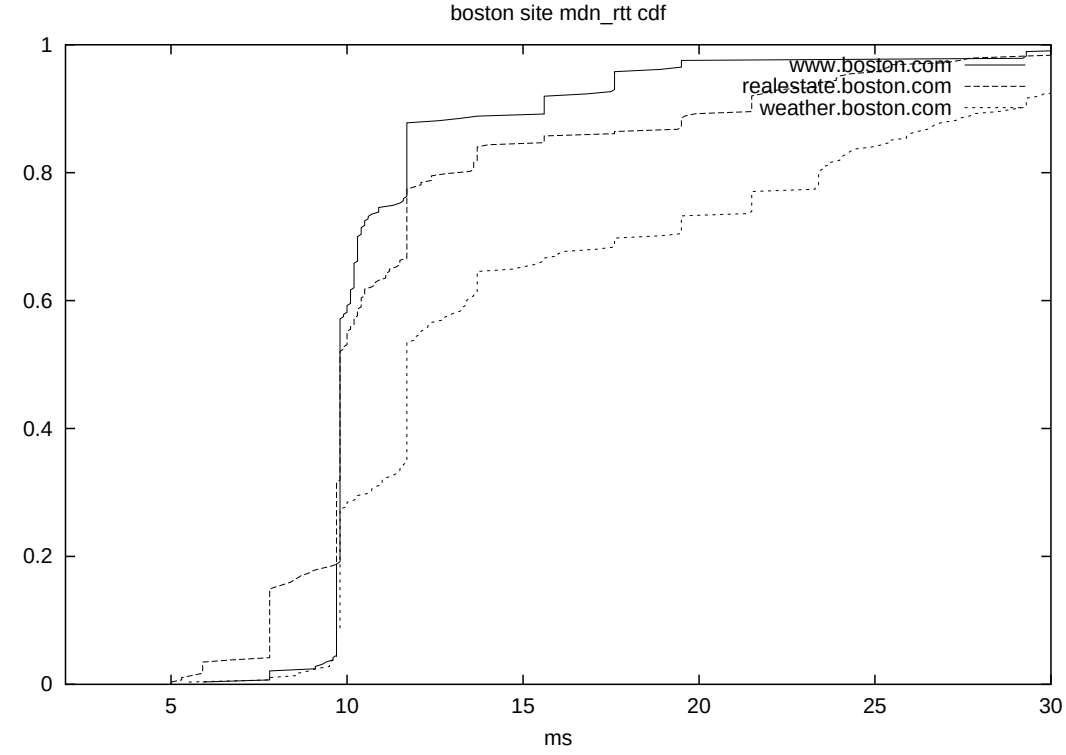

Figure 4.2: CDF of Median RTTs of Boston Globe Cluster

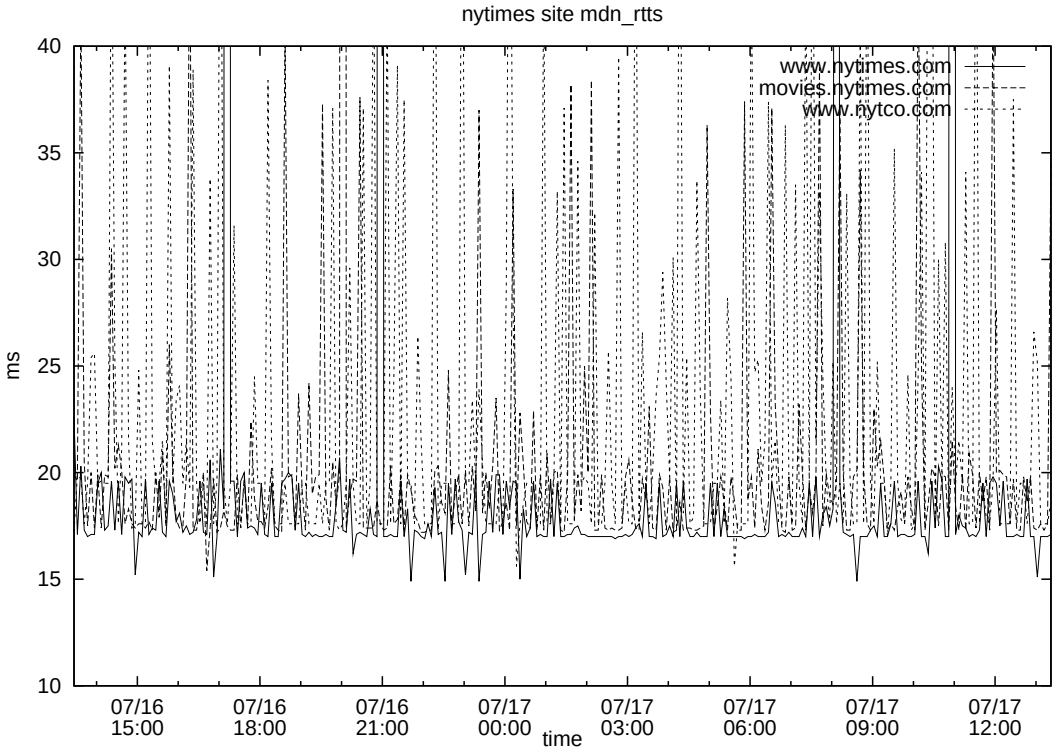

Figure 4.3: Median RTTs of NYTimes Cluster

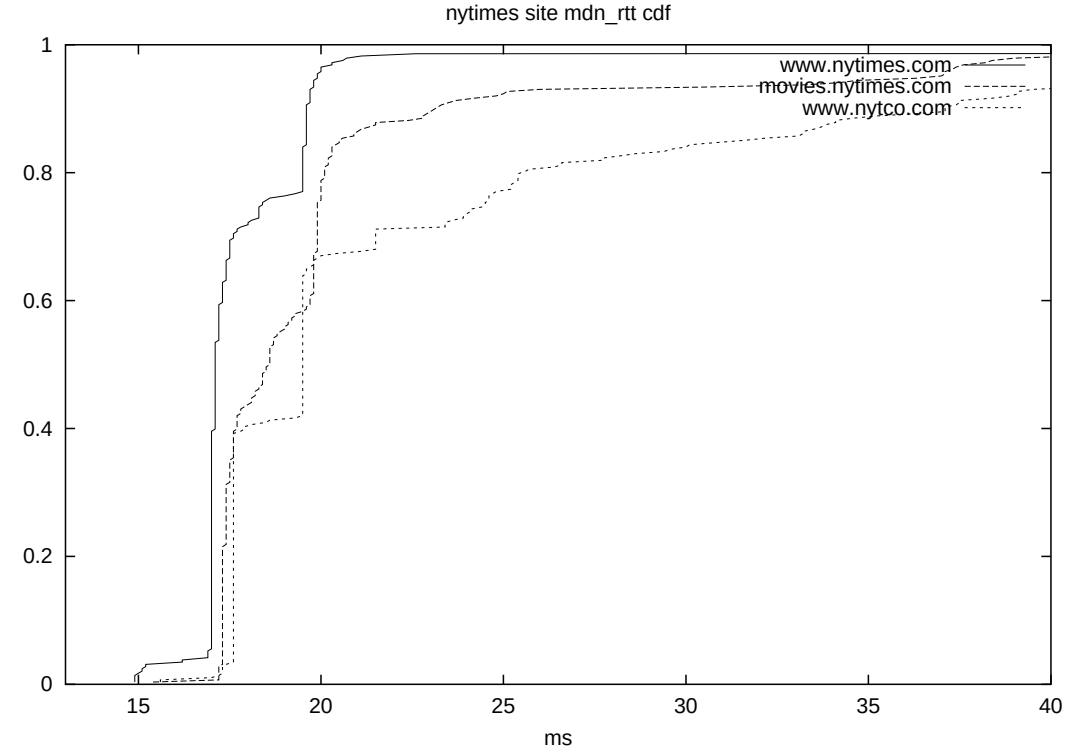

Figure 4.4: CDF of Median RTTs of NYTimes Cluster

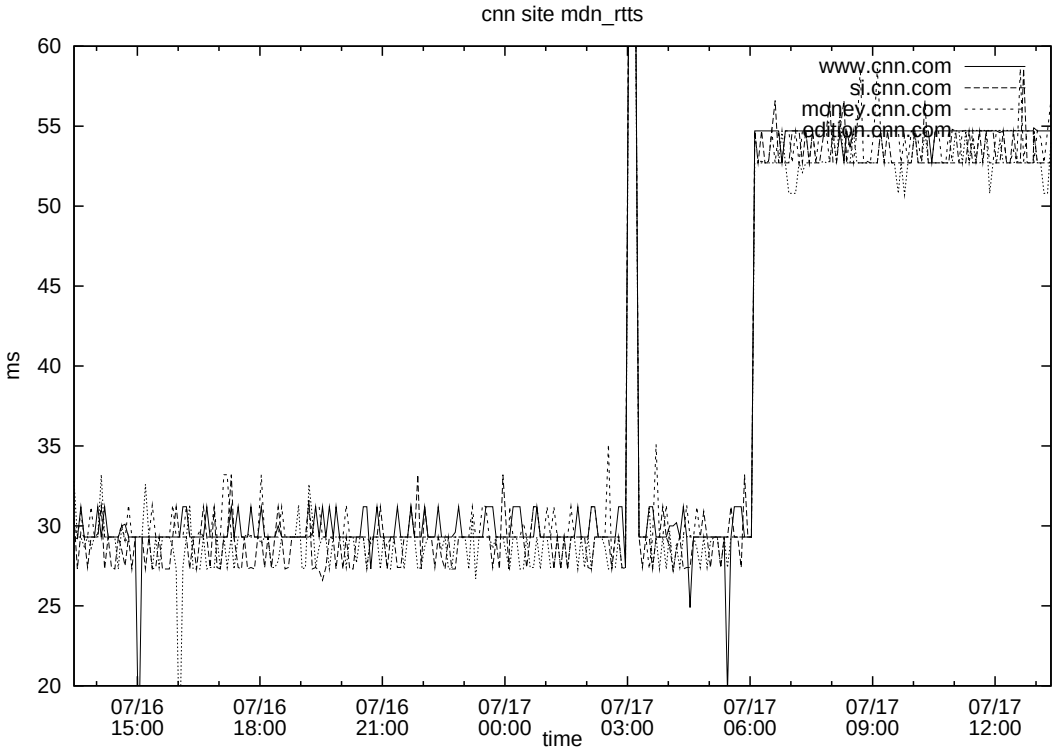

Figure 4.5: Median RTTs of CNN Cluster

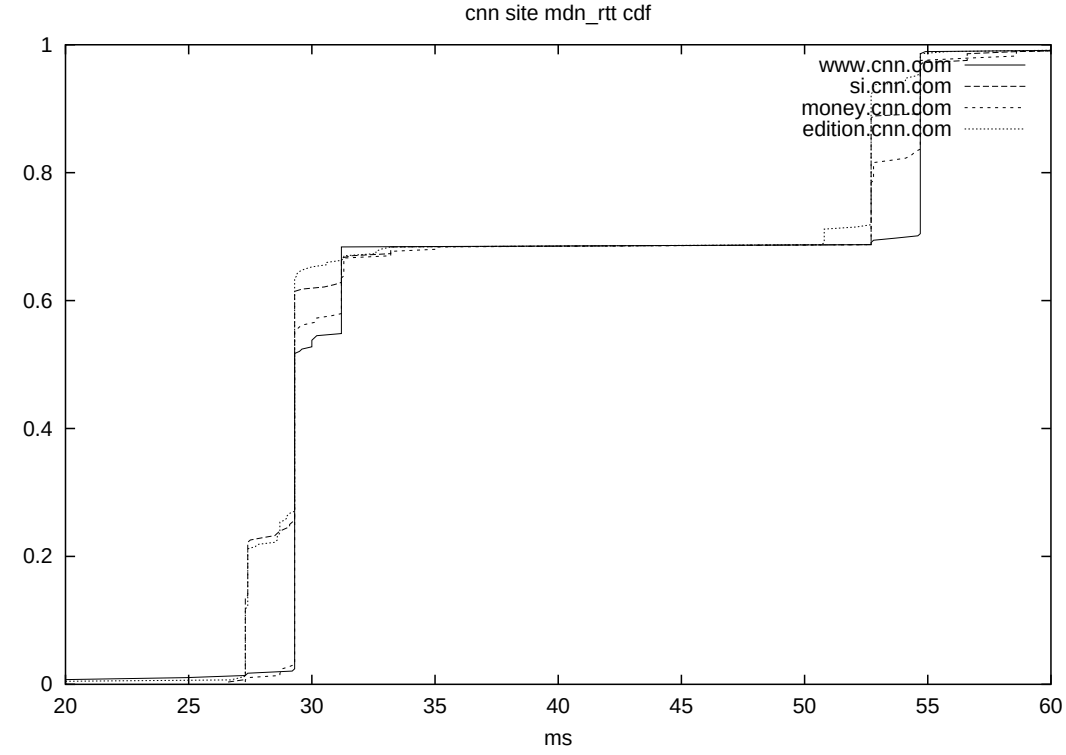

Figure 4.6: CDF of Median RTTs of CNN Cluster

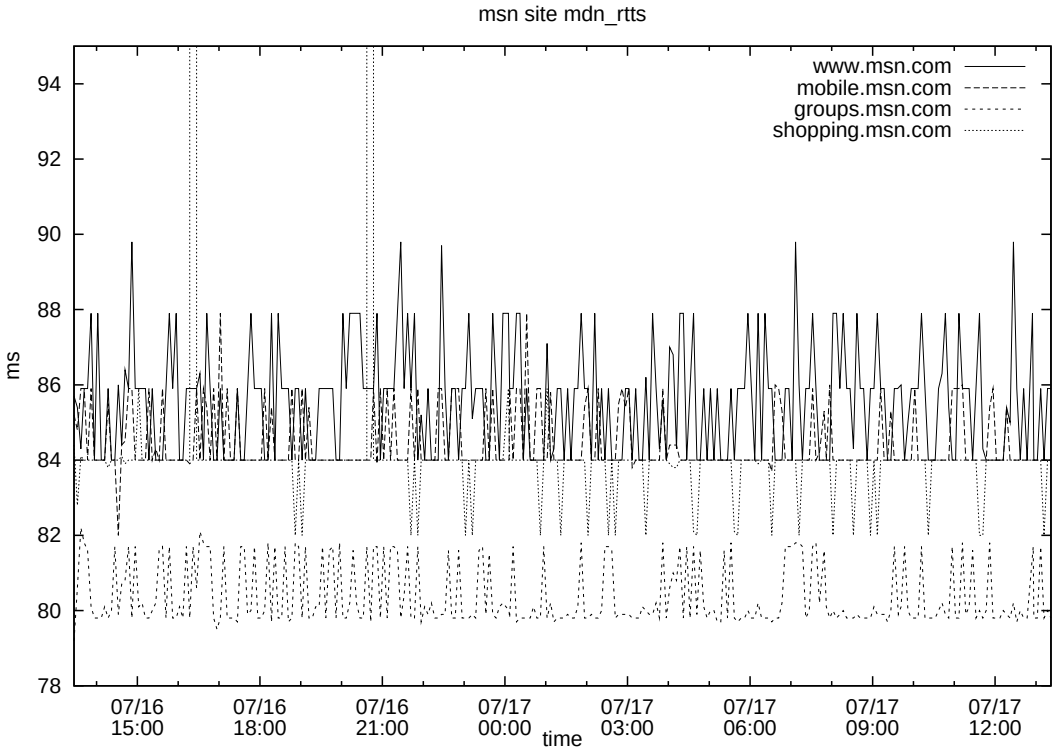

Figure 4.7: Median RTTs of MSN Cluster

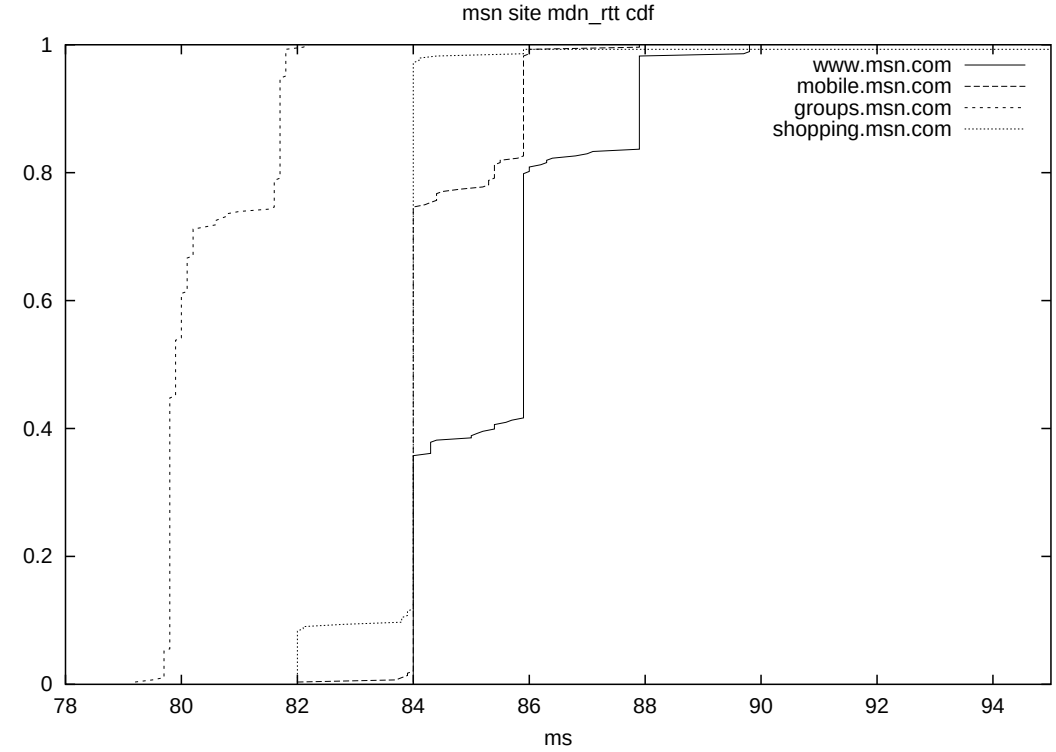

Figure 4.8: CDF of Median RTTs of MSN Cluster

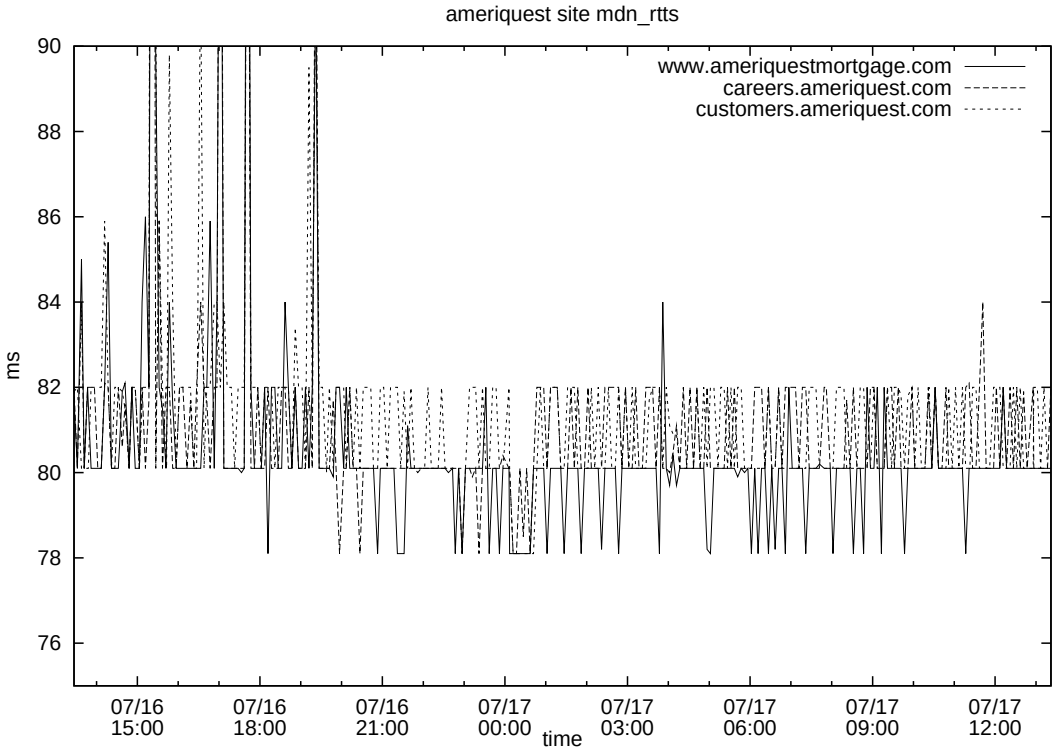

Figure 4.9: Median RTTs of Ameriquest Cluster

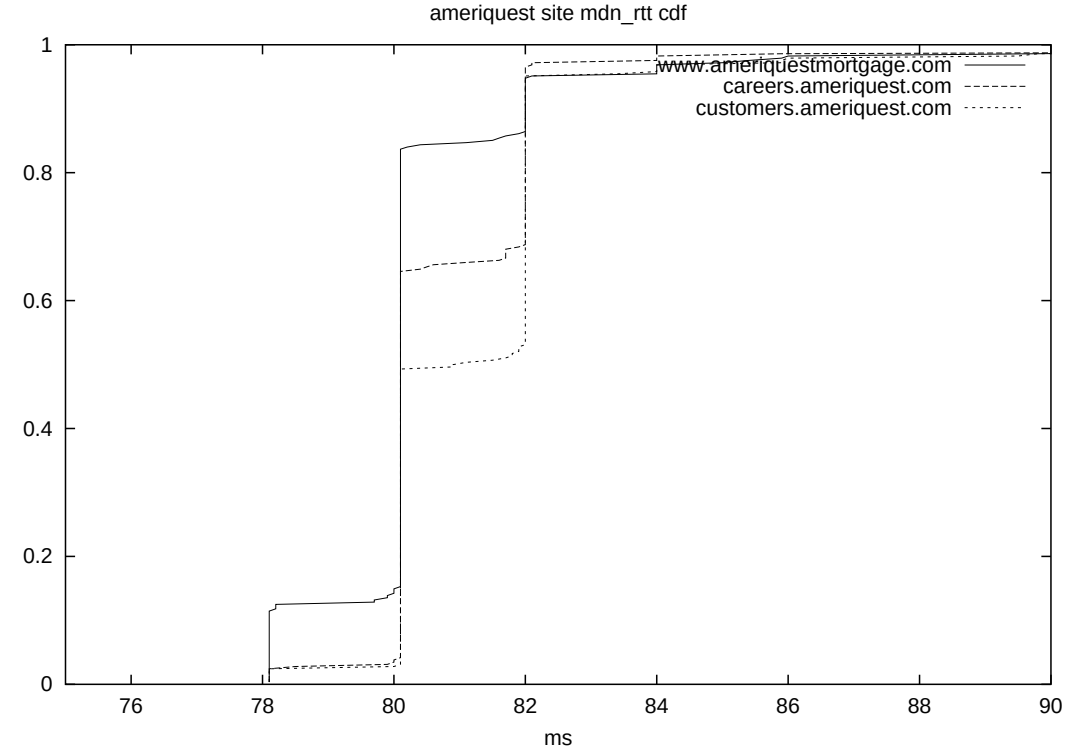

Figure 4.10: CDF of Median RTTs of Ameriquest Cluster

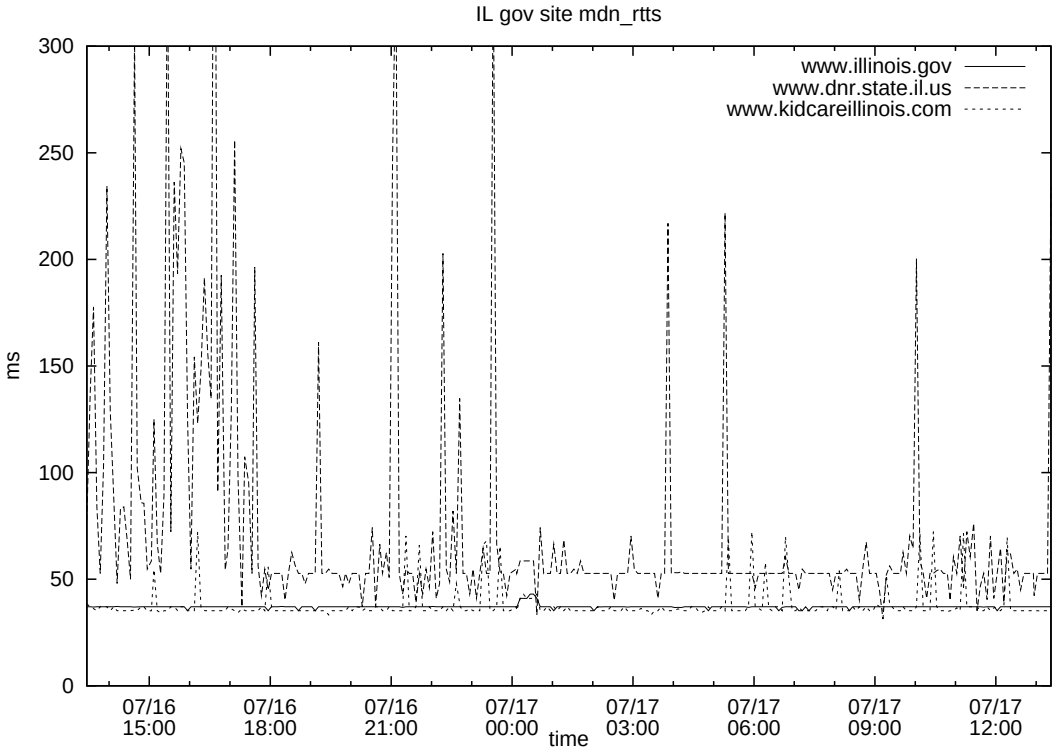

Figure 4.11: Median RTTs of IL Government Cluster

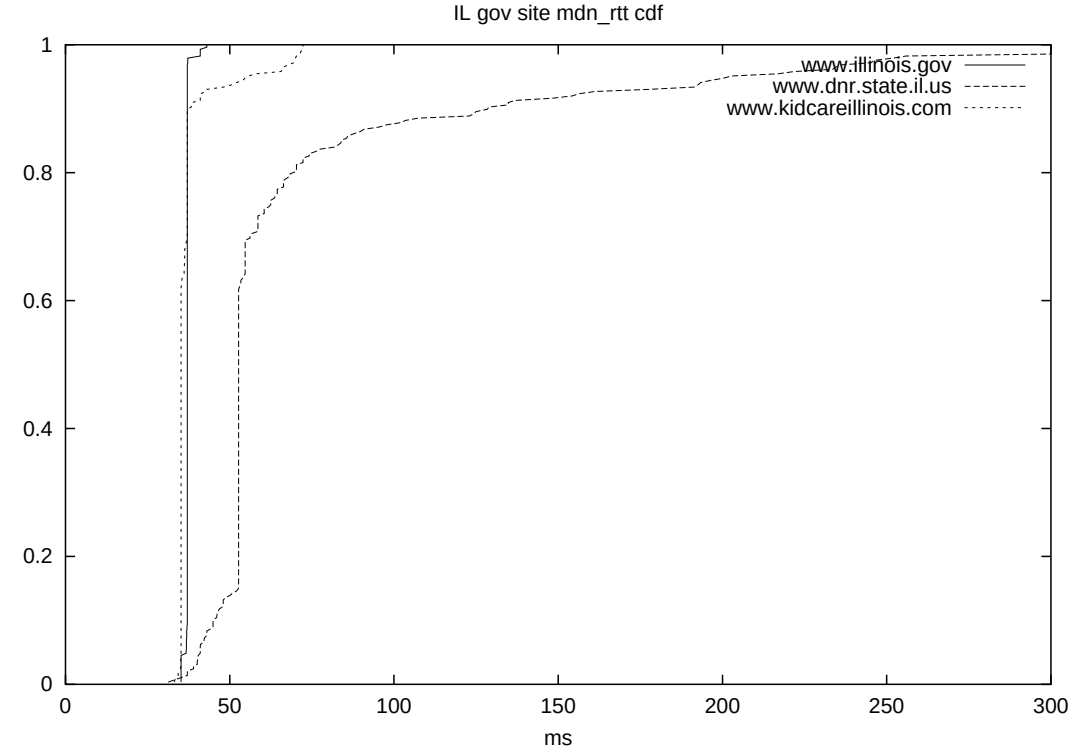

Figure 4.12: CDF of Median RTTs of IL Government Cluster

Some servers had consistent round trip times from the WPI cluster with little variation over time. For example, the three nytimes servers, servers www.msn.com, shopping.msn.com, and groups.msn.com at the MSN cluster, and servers www.illinois.gov and www.kidcareillinois.com at the IL Government cluster, all fall into this category. This consistency can also be observed at two separate time frames at the cnn servers and ameriquest servers, except that there was a clear jump or drop in the median RTT value at a certain time. We suspected this could be caused by a route change on the network since we did not observe significant change of available bandwidth during this period. Three servers at the Boston Globe cluster have shown large variation of their median RTTs.

4.2.2 Median Available Bandwidth

In our data processing, we used packet pairs in each round of a connection, and calculated available bandwidth for each packet pair. We drew the CDF graph of the median available bandwidth of a connection. Figure 4.13 through Figure 4.16 are respectively such CDF graphs for six remote clusters. Surprisingly, we found that the median available bandwidth of most of servers have similar distributions. They all concentrate around 700K bytes/sec, except www.nytimes.com and www.nytco.com which have higher available bandwidth around 1200K bytes/sec. We suspect this is due to a bottleneck link close to WPI which most of the connections share.

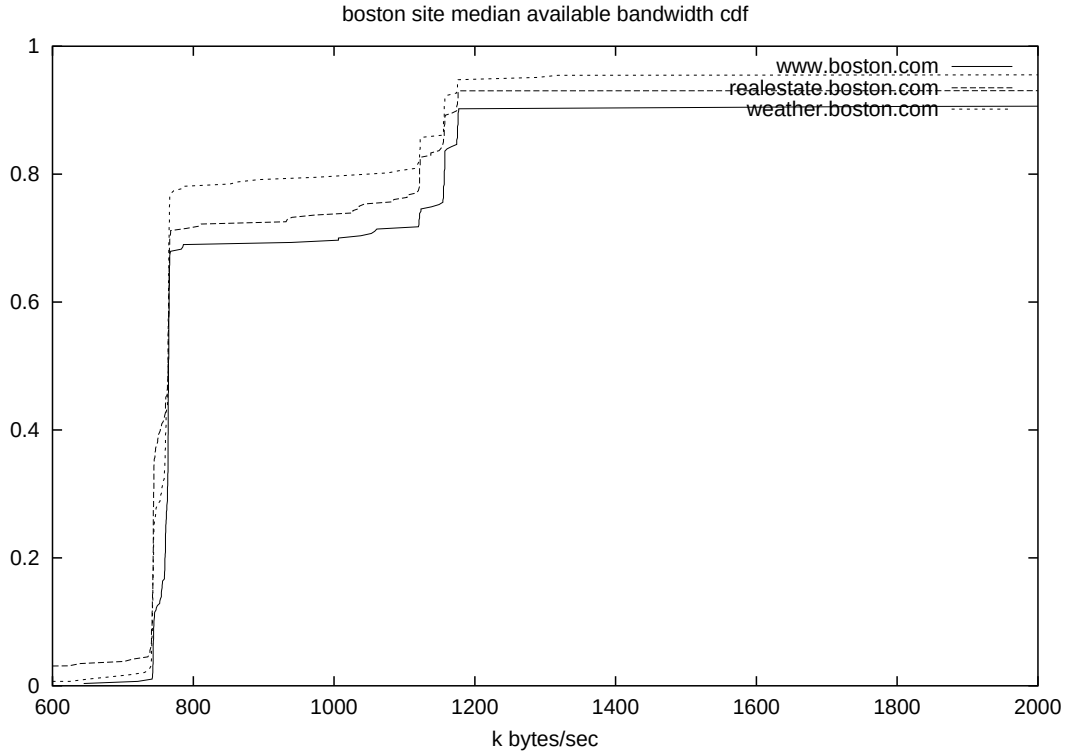

Figure 4.13: CDF of Median Available Bandwidth of Boston Globe Cluster

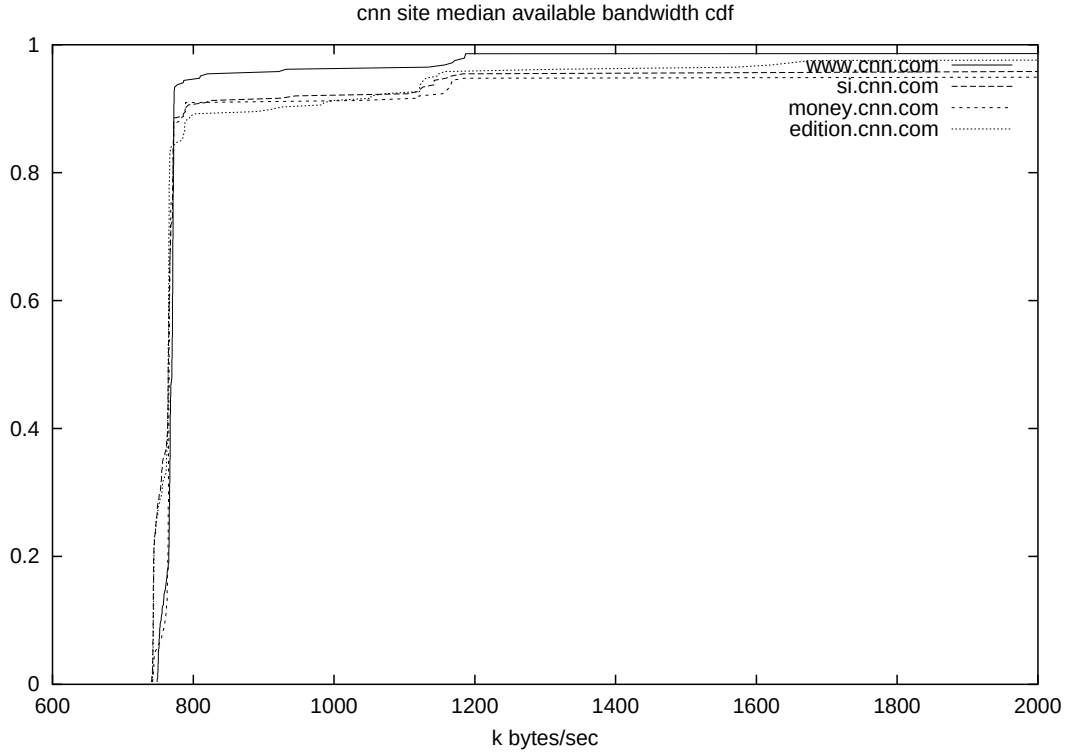

Figure 4.14: CDF of Median Available Bandwidth of CNN Cluster

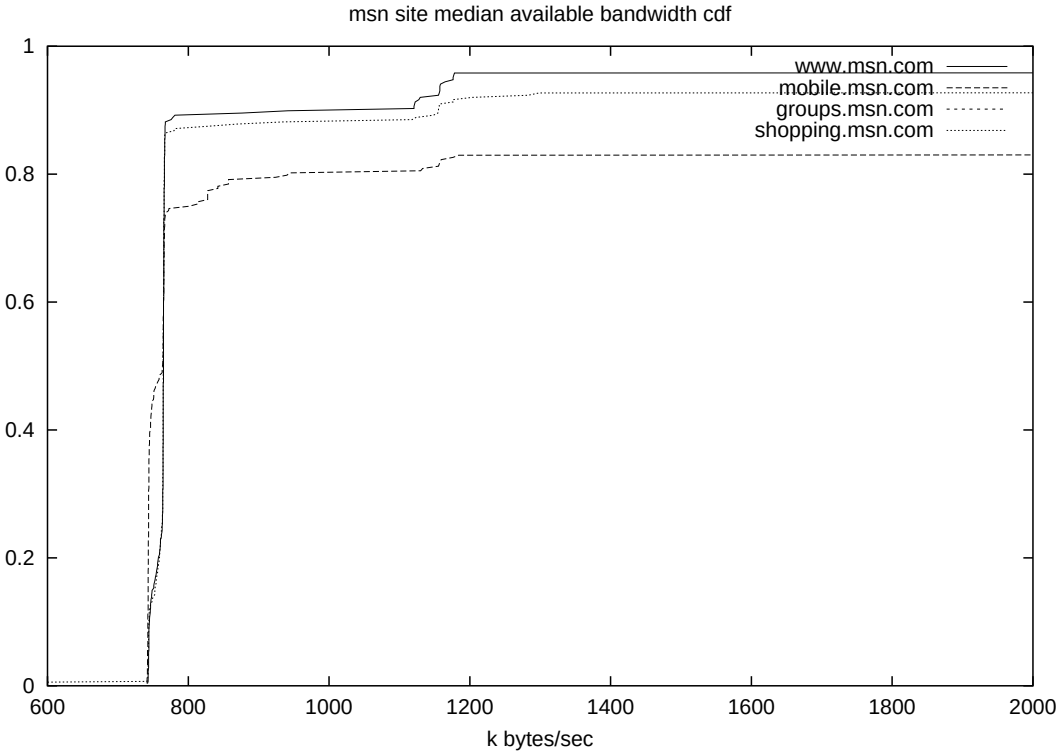

Figure 4.15: CDF of Median Available Bandwidth of MSN Cluster

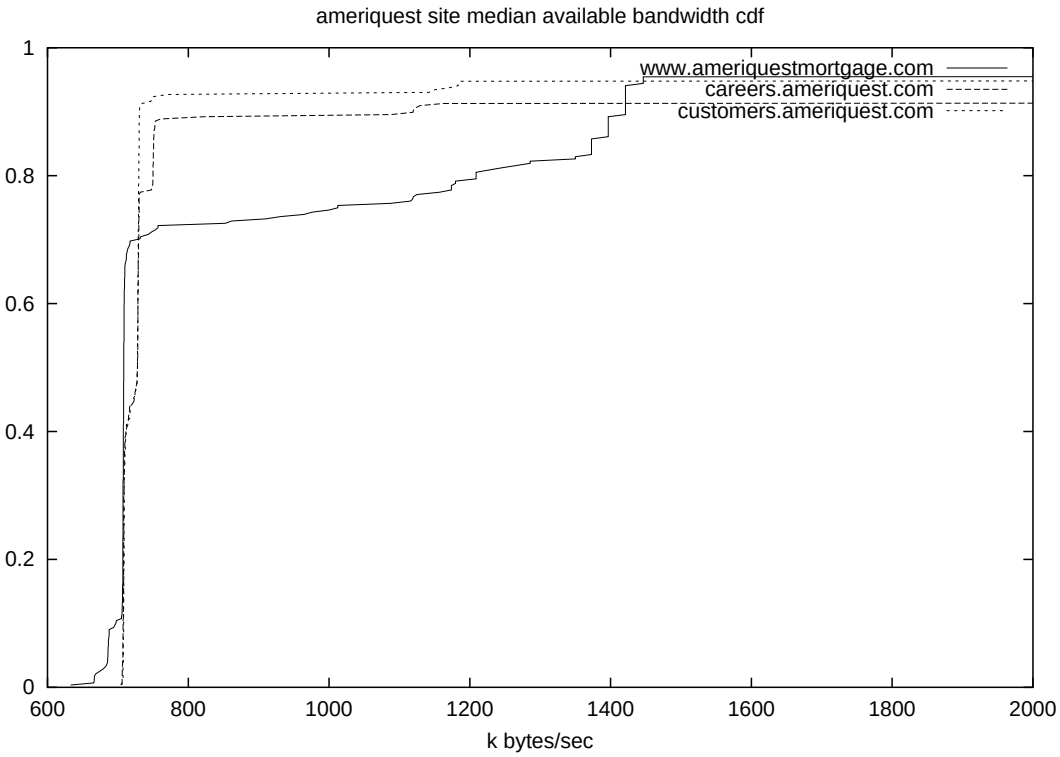

Figure 4.16: CDF of Median Available Bandwidth of Ameriquest Cluster

4.2.3 Connection Throughput

Connection throughput in this work is defined by the number of bytes transferred in the connection divided by the time taken to transfer them. As we have previously analyzed the TCP inner workings, the overall throughput for a server having a long RTT from WPI depends more on the RTT value than the bandwidth in the early rounds of a connection. This is because the server has to stall in each round to wait for ACKs from the client. The object sizes at the servers we studied were mostly 50k - 100k and connections took only a few rounds to finish. Server stalling dominated most of the connection, especially for the farther clusters. So we expect to see connection throughput increase as RTT decrease, and vice versa. This has been shown in Figure 4.17 through Figure 4.20.

4.3 Predictions

Making time-based prediction on this data set is illustrated in this section. In our experiment, we found by comparing with other more complex mathematical predictors such as polynomial decay predictor, that the exponential decay predictor achieved good prediction accuracy with simple computation.

We tried different λ values that could possibly affect the prediction accuracy. We used $\lambda = 0, 0.3, 0.65, 0.8, 0.95$. The smaller λ is, the more importance is placed on the recent accesses. We found that in predicting RTTs, as the value of λ became smaller, the performance accuracy became higher; while the opposite was true with the throughput prediction. However, we saw that overall the choices of different values of λ did not make a significant difference in terms of the prediction accuracy.

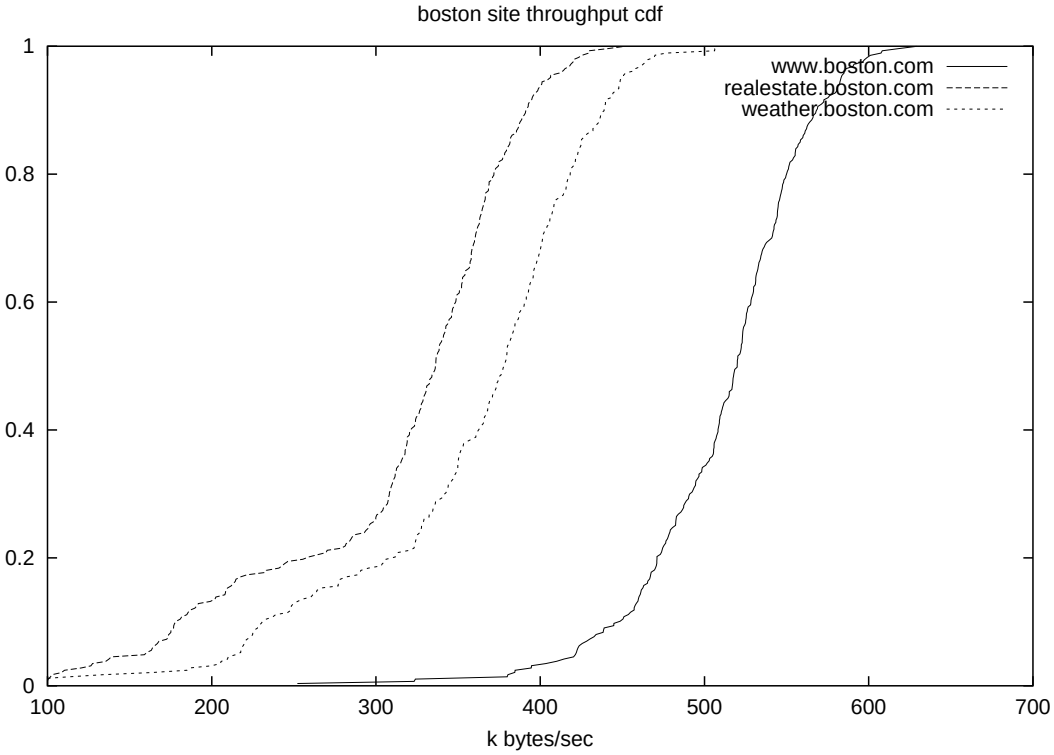

Figure 4.17: CDF of Throughput of Boston Globe Cluster

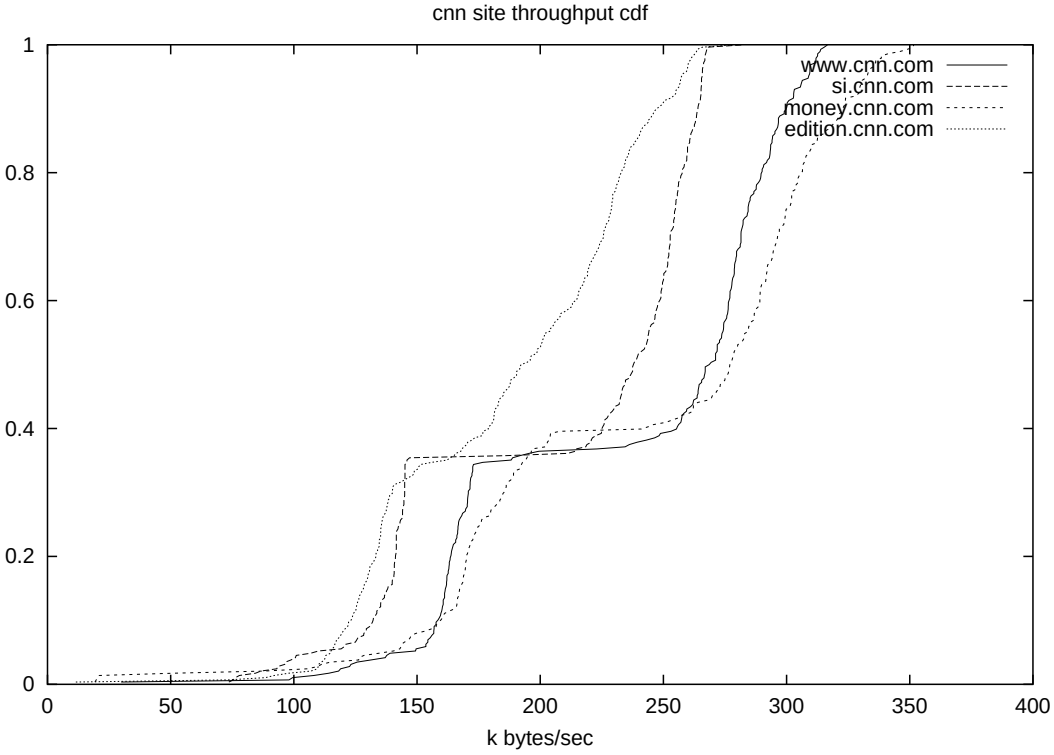

Figure 4.18: CDF of Throughput of CNN Cluster

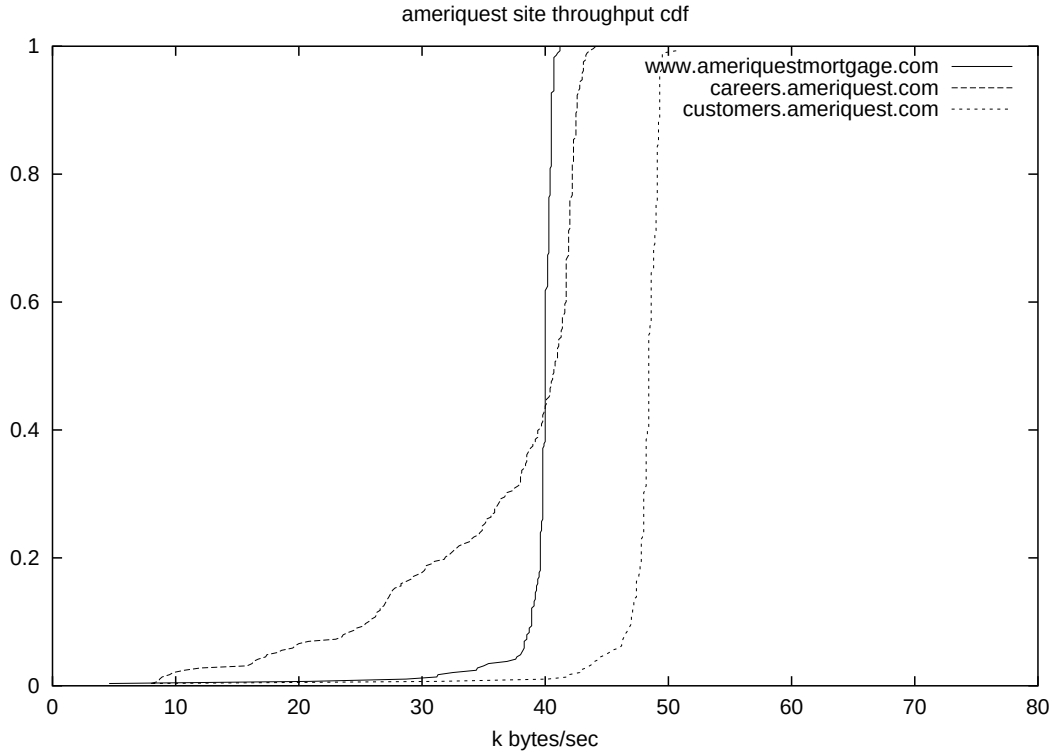

Figure 4.19: CDF of Throughput of Ameriquest Cluster

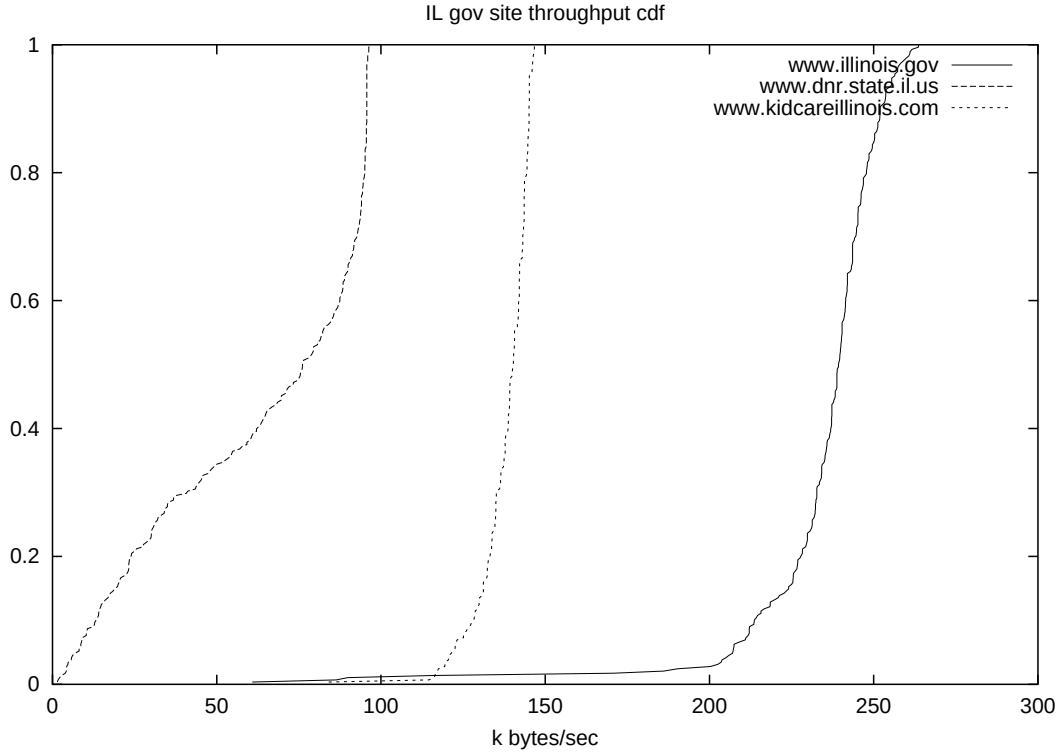

Figure 4.20: CDF of Throughput of IL Government Cluster

4.4 Summary

In this chapter we presented some of our preliminary work. We set up connections to various web servers geographically spread across the U.S. We outlined our data collection methods and presented some of our results. We saw that metrics remained mostly stable though they did show a tendency to shift over time. Also within the same cluster, different web servers had different measurements. We also experimented with different parameter values for our exponential predictor. However there were shortcomings with the work done so far. First of all we had only run the experiments with web applications. Second, the data was collected three years earlier. It is possible that during this period the routes would have changed. Hence we felt the need to conduct more experiments. These are described in the next chapter.

Chapter 5

New Experiments

This chapter describes the new experiments that were carried out to obtain network performance metrics in Jan 2007. We also describe how the data was processed so that it can be utilized to make predictions. This chapter describes the preliminary work carried out in the summer of 2004. The data was mainly collected for web applications.

5.1 Experimental Setup

In our experiment, we used 20 DNS servers, 62 web servers, 10 real streaming servers and tracerouted to 20 servers, located at 20 remote clusters from 8 geographical locations all over the continental United States, as show in Tables 5.1 and 5.2.

They were mostly the servers of popular news channels, popular newspapers and state governments. All the clusters that we selected had the DNS server, web servers, and the real streaming server at the same location. We make the assumption that

Table 5.1: Network Application Servers at Eight Geographical Locations in the U.S.

Location	Cluster Name	DNS Server	Web Servers	Real Streaming Server	Traceroute to Server
Boston, MA	Boston Globe	ns-a.pnap.net	www.boston.com weather.boston.com www.explorenewengland.com	N/A	www.boston.com
	MBTA	ns1.itg.net	www.mbta.com trip.mbta.com	N/A	www.mbta.com
	Web Hosting	ns2.cwwebs.com	www.aviationdisasterlawyers.com www.asbestoslaw.info www.pharmaceuticallawyers.com	www.consultwebs.com	www.consultwebs.com
New York, NY	NY Times	ns1t.nytimes.com	www.nytimes.com movies.nytimes.com homefinance.nytimes.com query.nytimes.com	N/A	www.nytimes.com
	UN	ns.undp.org	www.undp.org www.rbas.undp.org www.dz.undp.org google.undp.org	www.undp.org	www.undp.org
Atlanta, GA	CNN	twdns-04.ns.aol.com	www.cnn.com edition.cnn.com si.cnn.com money.cnn.com	N/A	www.cnn.com
	Weather.com	dns2.weather.com	www.weather.com forgetaway.weather.com ktopfw.weather.com br.weather.com	N/A	www.weather.com
	GA Government	ns3.state.ga.us	www.georgia.gov www.files.georgia.gov oca.awe.gta.ga.gov www.gov.state.ga.us	www.georgia.gov	www.georgia.gov
Springfield, IL	IL Government	ns1.state.il.us	www.dnr.state.il.us www.illinois.gov www.allkidscovered.com	www.illinois.gov	www.dnr.state.il.us
	IL Education	ns1.illinois.net	www.isbe.net	www.isbe.net	www.isbe.net

Table 5.2: Network Application Servers at Eight Geographical Locations in the U.S. (Continued)

Location	Cluster Name	DNS Server	Web Servers	Real Streaming Server	Traceroute to Server
Raymond, WA	MSN	ns1.msft.net	www.msn.com entertainment.msn.com music.msn.com weather.msn.com	N/A	www.msn.com
	Real	ns1.real.com	www.realnetworks.com brasil.real.com musicstore.real.com	rxns-rbn-sea10.rbn.com	www.realnetworks.com
Los Angeles, CA	Ameriquest	ns1.accads.com	www.ameriquestmortgage.com careers.ameriquest.com www.ameriquestracing.com	N/A	www.ameriquestmortgage.com
	City of LA	citylans1.lacity.org	www.lacity.org eng.lacity.org publiccsd.lacity.org parc1.lacity.org www.griffithobservatory.org	realav.lacity.org	www.lacity.org
San Francisco, CA	San Francisco	ns1.blvds.com	sanfrancisco.com www.santa-clara.com www.santacruz.com www.oakland.com	N/A	sanfrancisco.com
	CA Government	ns1.net.ca.gov	democrats.assembly.ca.gov www.legislature.ca.gov republican.assembly.ca.gov	N/A	democrats.assembly.ca.gov
	City of Davis	wheel.dcn.davis.ca.us	www.city.davis.ca.us events.dcn.org www.dcn.org	media.city.davis.ca.us	www.city.davis.ca.us
Dallas, TX	Dallas News	ns1.belo.com	www.dallasnews.com www.cowboysplus.com www.guidelive.com	N/A	www.dallasnews.com
	City of Irving	sob.ci.irving.tx.us	www.ci.irving.tx.us	www.ci.irving.tx.us	www.ci.irving.tx.us
	Online Video	ns.rackspace.com	www.lapdonline.org	www.lapdonline.org	www.lapdonline.org

if the first two bytes of server IP addresses are the same, the servers are at the same location. Each location would include at least one streaming server.

In identifying the location of a server, we normally use traceroute to observe the intermediate routers' names and make a proper guess. For example, for www.msn.com, we have the following traceroute results shown in Table 5.3. As we can see in hop 13 "microsoft-1-lo-jmb-706.sttlwa.pacificwave.net", "sttlwa" means Seattle in WA. From the RTT at this hop, we know this hop is close to the final destination of www.msn.com, and the rest on the route could be on MSN's internal network. Besides traceroute, we also use AntiOnline [1] and Geobytes IP locator [2] to verify the location we inferred from traceroute.

> traceroute www.msn.com
traceroute to www.msn.com (207.68.173.76), 30 hops max, 40 byte packets
1 RTR-PHSR1-FULLER.INF.WPI.EDU (130.215.24.3) 0.334 ms 0.230 ms 0.207 ms
2 RTR-GPOP1-BACKBONE.INF.WPI.EDU (130.215.0.131) 0.728 ms 0.565 ms 0.522 ms
3 WPI-GODDARD.GODDARD.GIGAPOP.NET (130.215.7.17) 0.695 ms 0.652 ms 0.666 ms
4 WORCESTER-BOSTON.GODDARD.GIGAPOP.NET (130.215.6.2) 1.620 ms 1.632 ms 1.628 ms
5 nox1sumgw1-Vl-591-NoX-WPI.nox.org (192.5.89.41) 1.723 ms 1.619 ms 1.643 ms
6 nox300gw1-Vl-803-NoX.nox.org (192.5.89.238) 1.836 ms 1.798 ms 1.648 ms
7 nox300gw1-PEER-NoX-INTERNET2-192-5-89-222.nox.org (192.5.89.222) 6.739 ms 6.694 ms 6.659 ms
8 so-0-0-0.0.rtr.wash.net.internet2.edu (64.57.28.11) 35.503 ms 43.231 ms 39.360 ms
9 so-0-2-0.0.rtr.chic.net.internet2.edu (64.57.28.12) 28.453 ms 28.244 ms 28.321 ms
10 so-4-3-0.0.rtr.kans.net.internet2.edu (64.57.28.36) 38.840 ms 38.770 ms 38.789 ms
11 so-0-0-0.0.rtr.salt.net.internet2.edu (64.57.28.24) 63.572 ms 63.430 ms 63.400 ms
12 so-0-0-0.0.rtr.seat.net.internet2.edu (64.57.28.26) 80.023 ms 79.780 ms 79.801 ms
13 microsoft-1-lo-jmb-706.sttlwa.pacificwave.net (207.231.240.7) 80.076 ms 80.031 ms 79.976 ms
14 ge-7-3-0-58.wst-64cb-1a.ntwk.msn.net (207.46.36.177) 80.112 ms 79.897 ms 79.937 ms
15 ge-7-0-0-0.wst-64cb-1b.ntwk.msn.net (207.46.34.122) 80.076 ms 79.987 ms 80.015 ms
16 ge-6-1-0-0.tuk-64cb-1b.ntwk.msn.net (207.46.35.33) 80.412 ms 80.419 ms 80.354 ms
17 ten1-2.tuk-76c-1a.ntwk.msn.net (207.46.44.50) 80.219 ms 80.126 ms 80.310 ms
18 * * *
19 207.68.173.76 80.548 ms 80.590 ms 80.581 ms

Table 5.3: Traceroute to www.msn.com

Four Internet applications were used in obtaining the network metrics. They were DNS requests, web page retrievals, real streaming object downloading, and

traceroute. All requests were sent from a machine located at Worcester Polytechnic Institute at Worcester, MA to each of the servers at a certain time interval for a time period of 21 days in Jan 2007. DNS requests, web page retrievals, and real streaming were run every 10 minutes, and traceroutes were sent every three hours to show if there was any route change. Each stream was run for 15 seconds.

The machine used to send requests were running SuSE 10.1 with a 2.6 Linux kernel. Tcpdump was used to capture the packets received from the kernel level at the client side. Tcpdump recorded the time stamp when each data packet and acknowledgments were sent and received. For each data packet, we recorded its sequence number, its packet size, and the time stamp when it was received. For each acknowledgment, we recorded the sequence number it acknowledges and the time stamp it was sent out. We also kept track of the number of duplicate ACKs sent out for each connection.

5.2 Measurement Mechanism

For each web page retrieval, we inferred network metrics from it by studying the TCP packets received. We summarize the metrics for each retrieval to give the connection a health rating of good, bad or medium. The network metrics used included round trip time, available bandwidth, overall throughput, out of order received packets, and duplicate acknowledgments sent.

As mentioned earlier, received packets were grouped together in analyzing performance of one web retrieval. Figures 5.1 through 5.3 show the patterns of packets received. They are xplot graphs from tcptrace results by processing the original tcpdump data. Figures 5.1 through 5.3 are connections with health ratings of 2, 1,

and 0 respectively, which will be discussed in more detail in Subsection 5.2.7.

```
r_pkt_grp = 1;
foreach received packet r_pkt do
    if r_pkt is a RST or duplicate SYN packet then
        rst_dupsyn = TRUE;
        break;                                  /* stop packet grouping */
    end
    if r_pkt is the first received packet then
        set syn_rtt;
        add packet r_pkt to packet group r_pkt_grp;
    else                                        /* not the first packet */
        r_pkt_dist ← time interval between r_pkt and pre_r_pkt;
        if r_pkt_dist < 0.35 × syn_rtt then              /* same group */
            add packet r_pkt to current packet group r_pkt_grp;
        else                                      /* not the same group */
            grp_dist ← time interval between r_pkt and 1st packet of r_pkt_grp;
            if 0.75 × syn_rtt < grp_dist < 1.3 × syn_rtt then   // new group
                r_pkt_grp + +;                 /* start a new packet group */
                add packet r_pkt to packet group r_pkt_grp;
            else
                /* stop packet grouping when the group pattern
                   becomes unclear                                        */
                break;
            end
        end
    end
    pre_r_pkt = r_pkt;
end
```

Algorithm 1: Grouping Received Packets in a TCP Connection

We separated the received packets into groups so that the time intervals between separated groups are approximately one round trip time. We could obtain the initial window size of cwnd, which is the number of packets in the first group. We could also infer the pattern of cwnd growing, either linearly or exponentially by observing the number of packets in each group. When the pattern of groups is not clear any more as the number of rounds goes up or there is potential packet loss, we stop

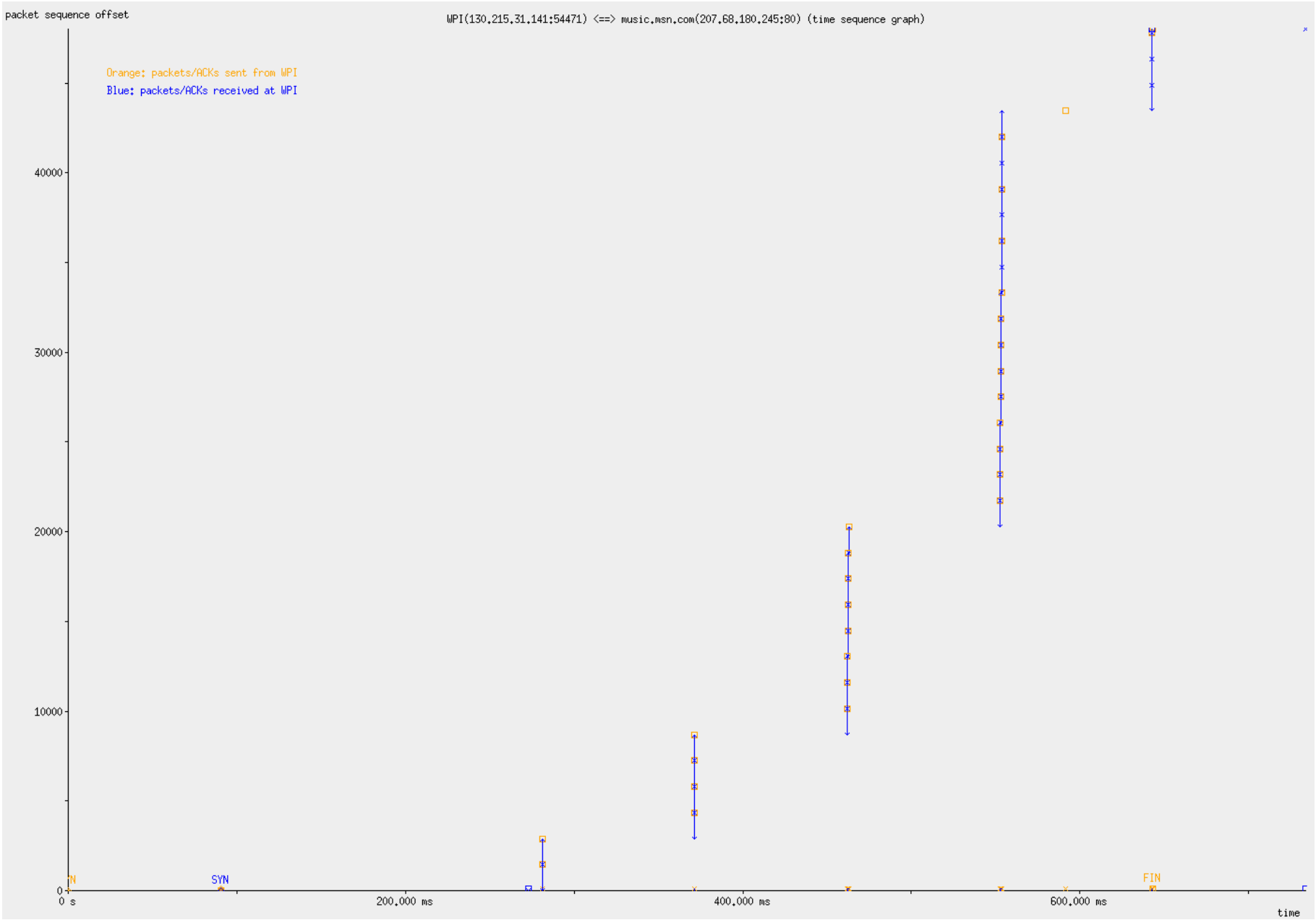

Figure 5.1: Packet Time Sequence for WPI client and Web Server music.msn.com, Connection Health Rating=2

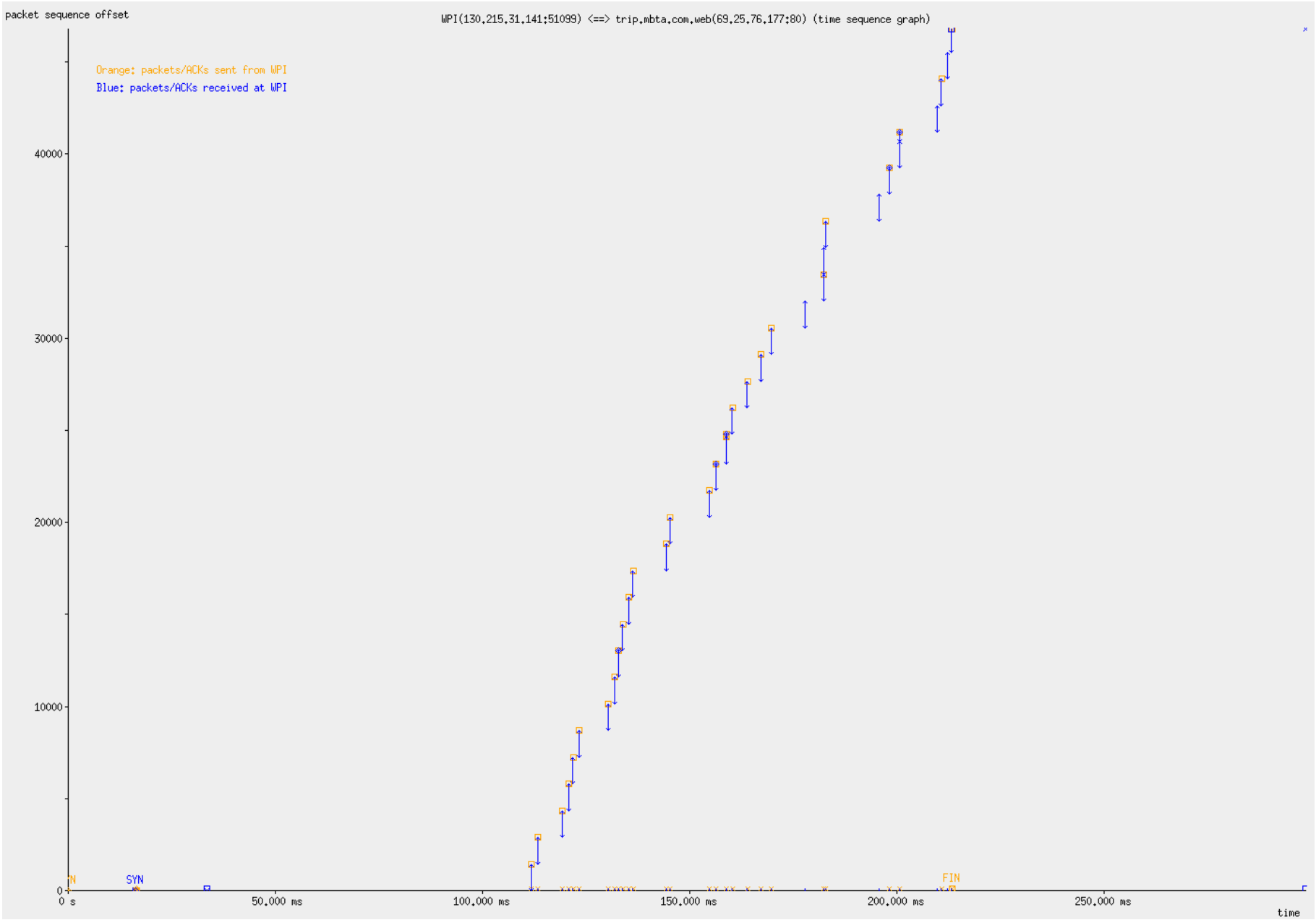

Figure 5.2: Packet Time Sequence for WPI client and Web Server trip.mbta.com, Connection Health Rating=1

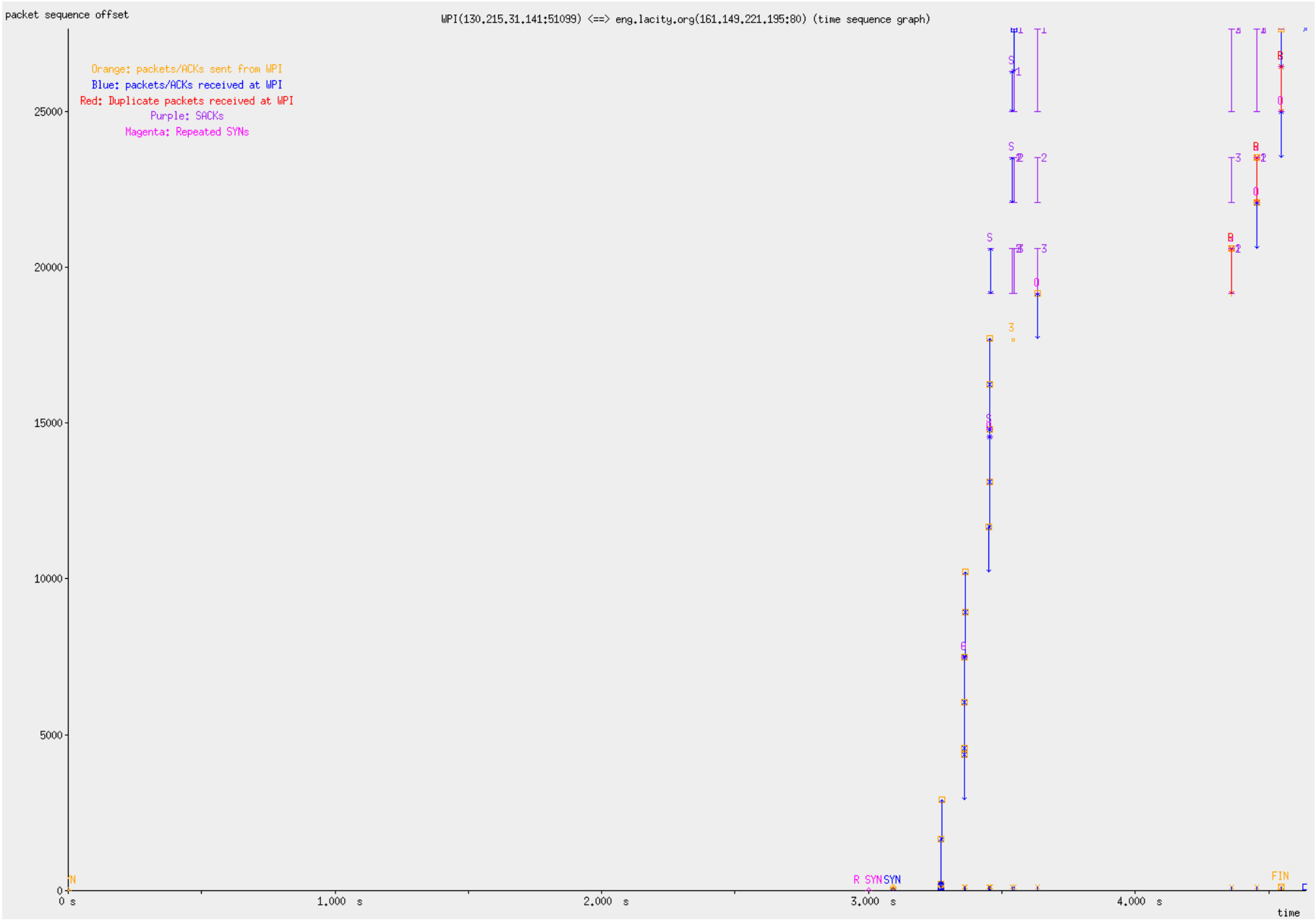

Figure 5.3: Packet Time Sequence for WPI client and Web Server eng.lacity.org, Connection Health Rating=0

grouping. Most of the network metrics are inferred from the grouping information.

The detailed algorithm used to group TCP received packets is illustrated in Algorithm 1. In this algorithm we first measure the RTT using the SYN packets. Then we measure the time difference of each received packet from the previous packet and based on this distance we make one of the following three decisions:

1. add the packet to the current group,

2. start a new group, and

3. stop the process of grouping.

We make the decision of adding a packet to the current group if its time distance from the previously received packet is less than 0.35 times the RTT. If this distance is greater than 0.35 the RTT, then we measure the distance of this packet from the first packet of the current group. If this time distance is between 0.75 times the RTT and 1.3 times the RTT then a new group is started with the current packet at the head. Any distance outside this range is an indication that grouping pattern is not clear and the algorithm terminates.

Our approach of grouping packets and therefore inferring metrics is different than tcptrace [3]. Tcptrace can only infer the RTT at the sender side upon its receiving the ACK for the packet it just sent. This approach is only effective at the sender side. Since we are mostly passively receiving data in accessing a web application at the client side, this approach becomes ineffective in inferring the RTT as well as other metrics. The only RTT measurement tcptrace can give is the first SYN RTT. In contrast, our approach only uses received packets and separates them into groups, where the group distance in a well-behaved TCP connection would

be roughly an RTT measure. Our approach increases more data points than what tcptrace can infer, while it is at the risk of wrong measurements in using wrong grouping information. However, we checked the grouping results using Algorithm 1, and most of the grouping results coincide with visual confirmation. For example, Fig 5.1 shows a perfectly well-behaved connection, where packets arrive in clear groups and the group distances are RTTs. The algorithm generated the same results and inferred five RTT measurements from it. Fig 5.2 shows a medium connection, where packet groups are clear, and we only use the SYN RTT measurement in that connection.

5.2.1 Metrics Used

In this section we discuss the metrics used to infer overall connection health ratings and to make predictions; and how they were obtained from the network trace. These metrics include round trip time(RTT), available bandwidth, overall throughput, out-of-order received packets, and duplicate acknowledgments sent.

5.2.2 RTTs

As mentioned in previous section, we group the packets by analyzing the packet receiving pattern so that time interval between groups were approximately one RTT. Our collected RTT measurements include the time intervals between packet groups and the first SYN-ACK round trip time. Average RTT and RTT standard deviation are calculated based on these RTT data points.

5.2.3 Available Bandwidth

We define available bandwidth as the highest rate at which data can be transferred between the client and the server. Our calculations are based on observations made at the receiver. For any packet group that contains n packets (r_pkt[0] through r_pkt[n-1]) and $n \geq 4$, we can infer one measurement for available bandwidth. We collect one data point per group. The calculation of the available bandwidth can be expressed as:

$$available\ bandwidth = \frac{sum\ of\ packet\ sizes\ of\ r\_pkt[1]\ through\ r\_pkt[n\text{-}1]]}{time\ interval\ between\ packet\ r\_pkt[0]\ and\ r\_pkt[n\text{-}1]} \quad (5.1)$$

If we are able to separate the packets into a few groups whose size is greater than four, we obtain a few data points for available bandwidth from that connection. We use a minimum of four packets per group to avoid outliers. Only when there are at least four packets in a group, do we use it to calculate one data point of available bandwidth.

5.2.4 Overall Throughput

Overall throughput is defined as the data received over the entire connection divided by the time elapsed, in a particular network application. This metric can only be measured more accurately in web and streaming applications compared with others, since there is more data pumped through in these two cases.

5.2.5 Out-of-Order Received Packets

Packets received out of order can be an indication of a potentially degraded network connection. Therefore, the number of out-of-order received packets is also considered as a measurement in evaluating one connection. Although it is not used as a direct prediction metric, it is used to rate the health level of the entire connection.

5.2.6 Duplicate/Triple-Duplicate ACKs Sent

As we mentioned in Section 3.1 "TCP Windowing", both three duplicate ACKS received for a TCP version using fast retransmission and a time-out at the sender side can indicate a potential packet loss, or at least some potential performance degradation. In most cases a network application client cannot know for sure if a packet sent from the server was lost. Since we can only infer the performance at the client side, we consider a packet loss event when more than three duplicate ACKs are sent. At the same time we also keep track of the number of duplicate ACKs sent.

5.2.7 Connection Health Ratings

As mentioned earlier, we give a health rating to each run of a network application as a summary to indicate if it is a good, bad or medium connection.

The detailed algorithm used to rate a connection is illustrated in Algorithm 2. This algorithm reads the packet patterns and empirically rates each connection as 0 (bad), 1(medium) or 2(good). If in a connection we see triple duplicate ACKs or

duplicate received packets then it is a bad connection. It is also a bad connection if we see a connection reset or duplicate SYN packets. If total bytes received are 0 then obviously it is a bad connection. To differentiate between a good and medium connection we use the conditions listed below. If all of these conditions are satisfied then it is a good connection, else it is a medium one:

Output:
0: bad connection;
1: medium connection;
2: good connection

```
if rst_dupsyn == TRUE and
   num_of_triple_duplicate_sent_acks >= 1 and
   num_of_duplicate_received_packets >= 1 and
   total_received_bytes == 0
then
    return 0 ;                                    /* bad connection */
else                                  /* good or medium connection */
    bndwdth_data_point_prcnt =
      num_of_bandwidth_data_points / total_received_packets;
    rtt_stddev_avg_ratio = average_rtt / rtt_standard_deviation;
    out_of_order_rcv_pkt_prcnt =
      num_of_out_of_order_received_packets / total_received_packets;
    dup_ack_snt_prnt =
      num_of_duplicate_acks_sent / total_received_packets;
    if bndwdth_data_point_prcnt >= 0.6 or
       rtt_stddev_avg_ratio <= 0.2 or
       out_of_order_rcv_pkt_prcnt <= 0.2 or
       dup_ack_snt_prnt <= 0.2 or
    then
        return 2 ;                                 /* good connection */
    else
        return 1 ;                               /* medium connection */
    end
end
```

Algorithm 2: Rating Connection Health

1. Grouped packet percentage: This is the percentage of total received packets that could be placed into groups as determined by algorithm 1. At least sixty percent of total packets should be groupable for a good connection.

2. Mean to Standard Deviation ratio for RTT: This ratio should not be greater than 0.2.

3. Percentage of packets received out of order: This value should be less than twenty percent.

4. Percentage of total packets that had duplicate ACKs: This value should be less than twenty percent.

As we presented at the beginning of Section 5.2, Figures 5.1 through 5.3 show connections rated as 2, 1, and 0 respectively, using Algorithm 2.

5.3 Summary

In this section we showed our expanded set of experiments. Not only were the experiments run for a longer time, but also on a larger range of network applications. We collected data for a larger set of metrics. Our algorithms for grouping received packets and rating the health of TCP connections have been described. In the following chapters we analyze this data and see how predictions can be made using observed network metrics.

Chapter 6

Time-Based Prediction

In this chapter, we discuss the variations of different network metrics over time. We study how historical data can be used to make predictions for future accesses. Only web-based applications have been studied in the experiments for this chapter.

6.1 Actual Results of Measurements

We have found that for the same server, connection health ratings, RTTs, available bandwidth, and overall throughput varied in different degrees, depending on the particular connection. Some connections tend to be more stable than others for certain metrics. This means, historic metrics from previous accesses to a server can provide different accuracies for prediction.

6.1.1 Round Trip Time (RTT)

For all sixty-two web servers studied in our experiment, their average RTTs remained relatively stable over time. Figures 6.1 through 6.6 show a few examples of the CDF

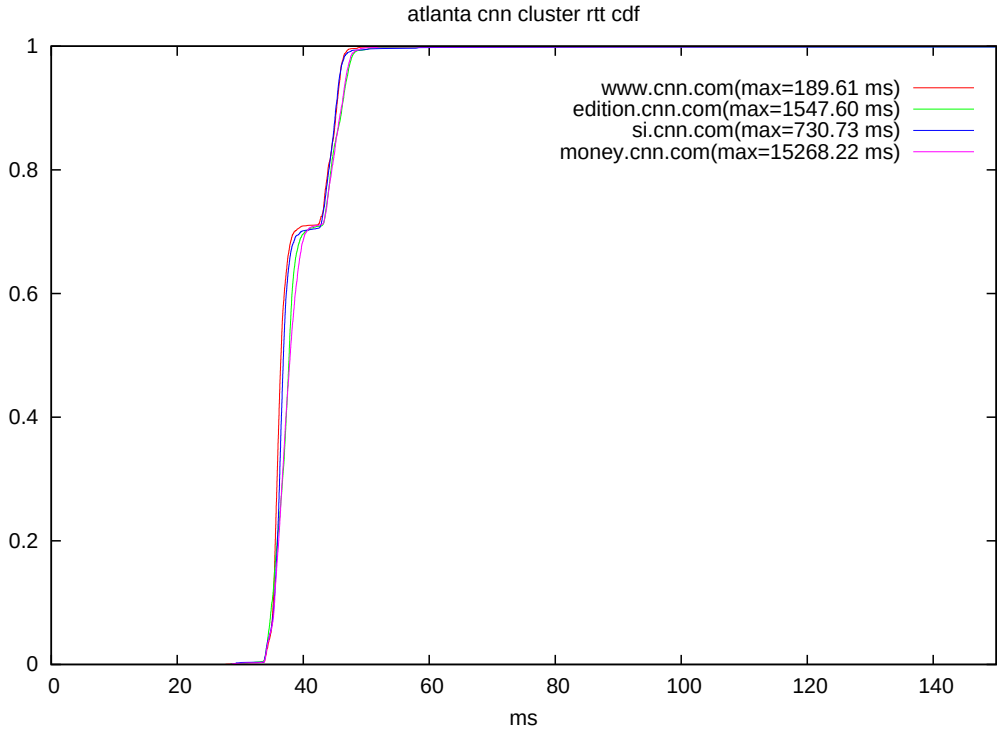

Figure 6.1: CDF of RTTs of Cluster CNN at Atlanta, GA

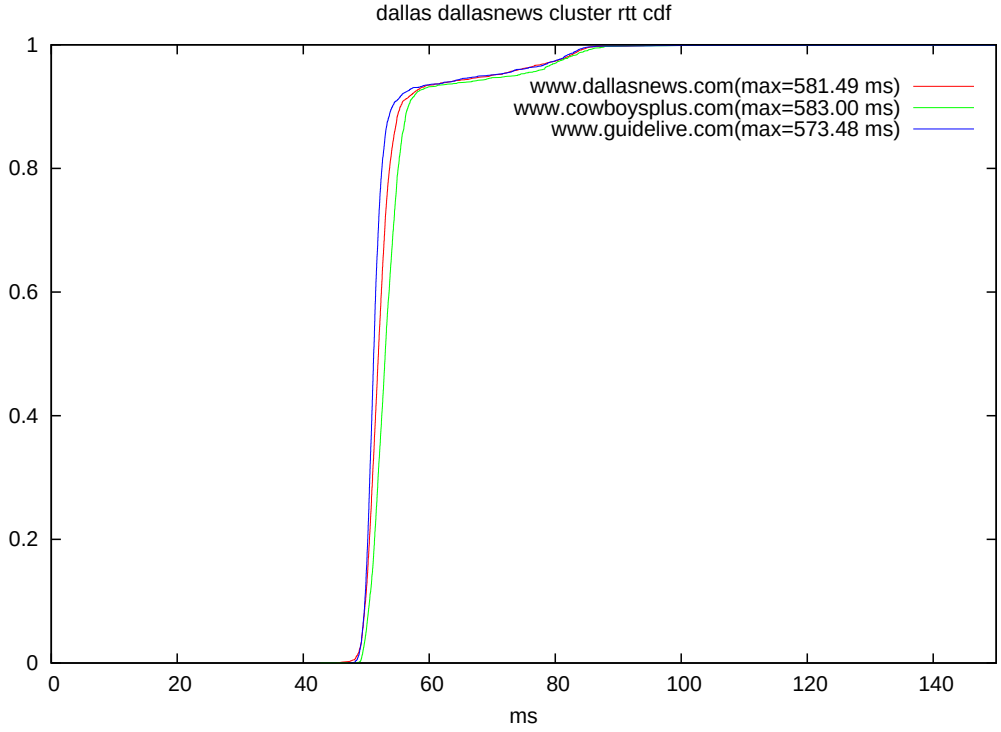

Figure 6.2: CDF of RTTs of Cluster Dallas News at Dallas, TX

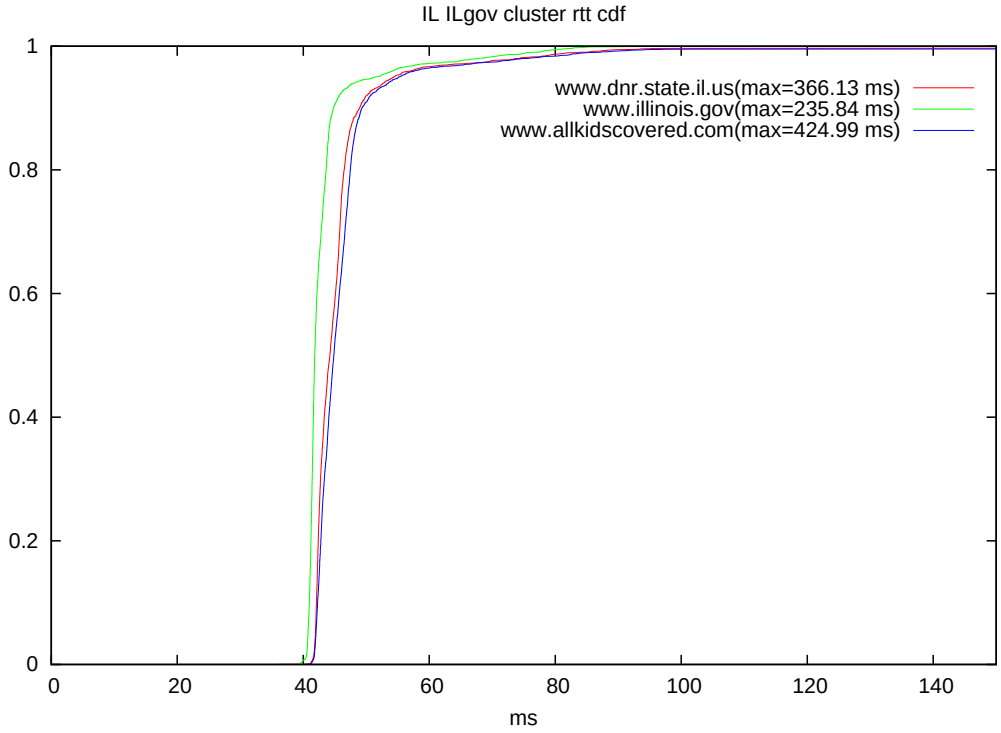

Figure 6.3: CDF of RTTs of Cluster IL Government at Springfield, IL

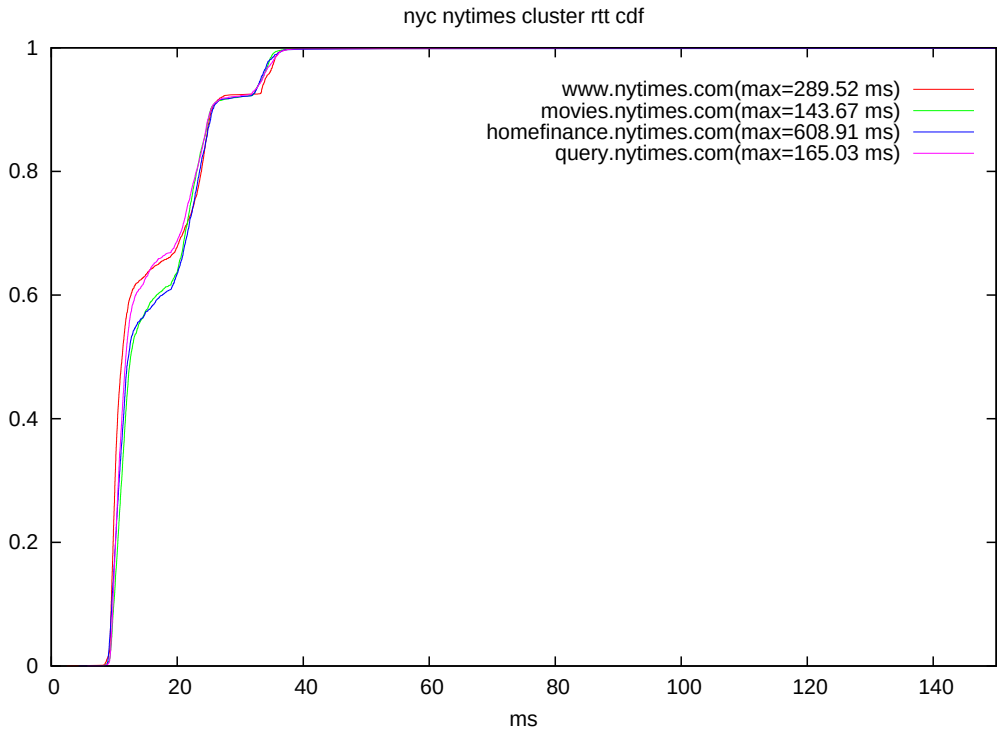

Figure 6.4: CDF of RTTs of Cluster NYTimes at New York, NY

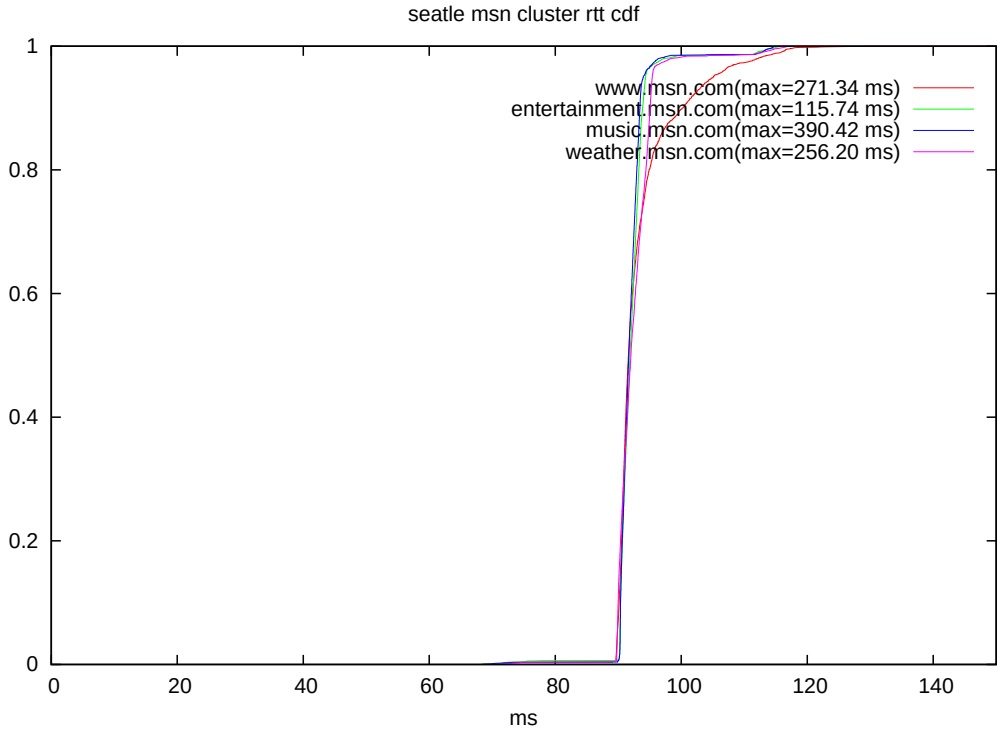

Figure 6.5: CDF of RTTs of Cluster MSN at Seattle, WA

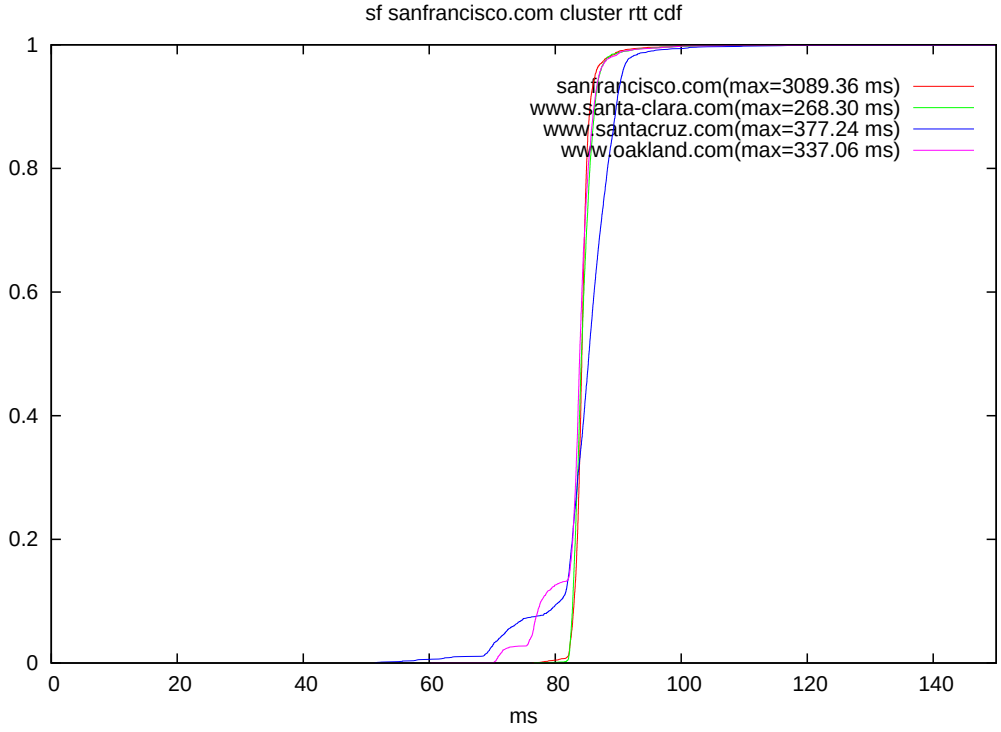

Figure 6.6: CDF of RTTs of Cluster SanFrancisco.com at San Francisco, CA

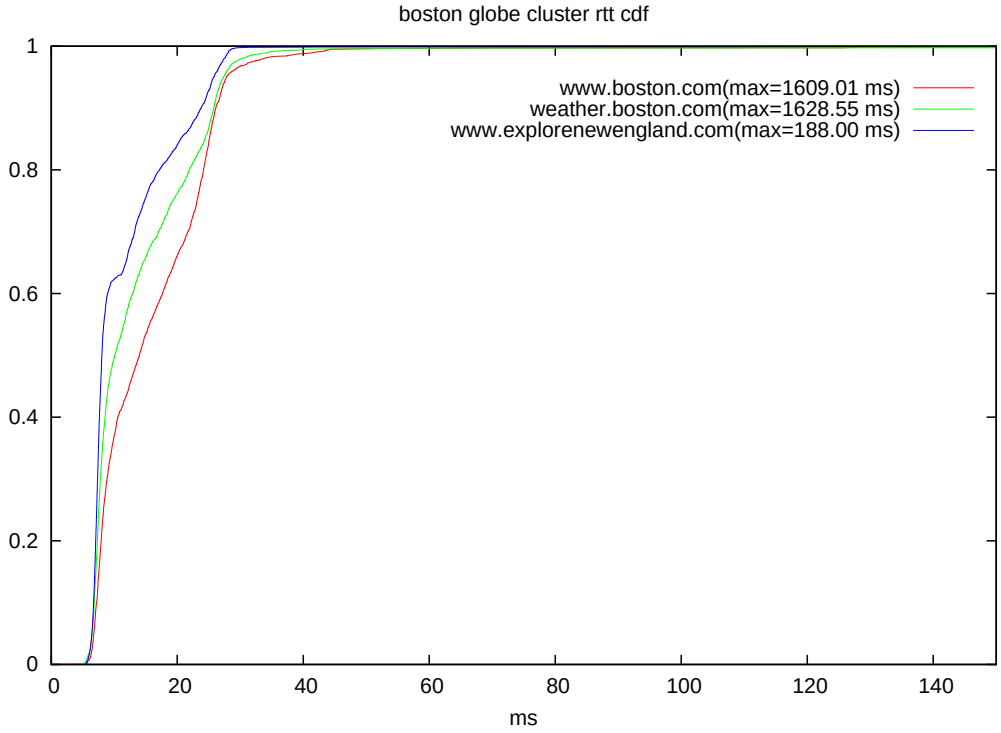

Figure 6.7: CDF of RTTs of Cluster Boston Globe at Boston, MA

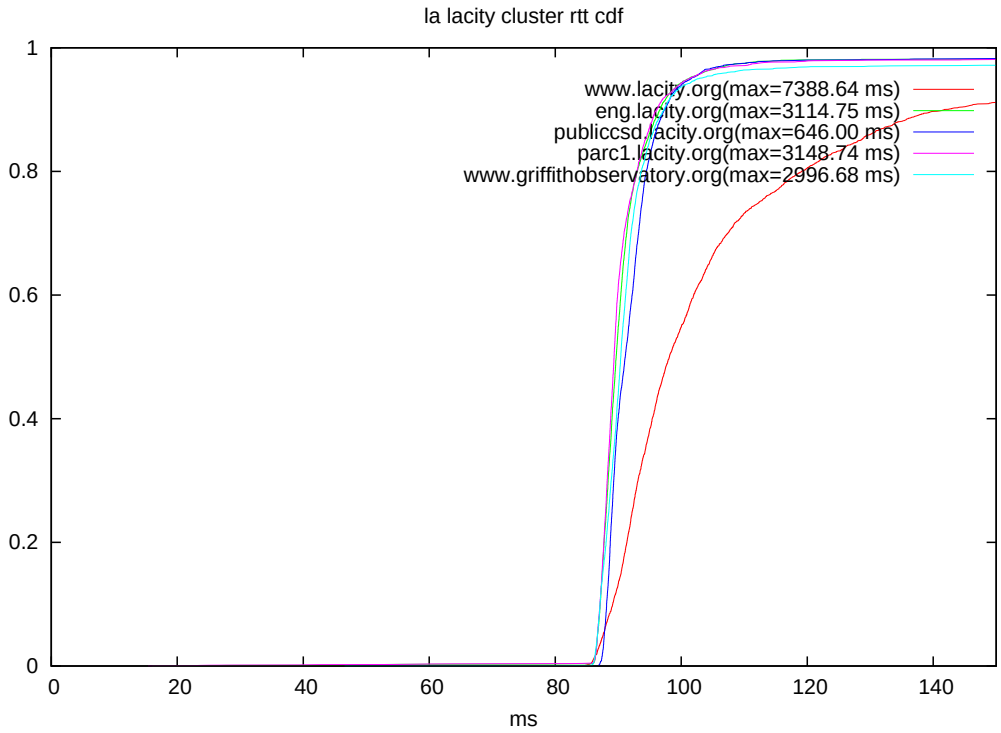

Figure 6.8: CDF of RTTs of Cluster LA city at Los Angeles, CA

(Cumulative Distribution Function) graphs of clusters with their servers all plotted in the same graphs. These clusters have shown rather good stability of RTT values, as we can see each CDF is roughly concentrated on a single value.

Figures 6.7 through 6.8 show two clusters with relatively more varied RTTs. The variance is only slightly larger than that in Figures 6.1 through 6.6.

6.1.2 Available Bandwidth

In our data processing, we obtain one data point of available bandwidth for each group, and use the median value of all data points as available bandwidth observed for the entire connection. We also find the connection level available bandwidth for each web server remains relatively stable over time. See Figures 6.9 through 6.12.

6.1.3 Connection Throughput

Connection throughput exhibits more variance than RTTs. Some of the clusters and servers show good stability over the entire collection period, while some have rather large variation. The variation exists either over one server or over multiple servers from the same cluster.

Figures 6.13 through 6.16 show the clusters with little variance on overall throughput. We can see all the servers within the cluster center around one value in the CDF graphs.

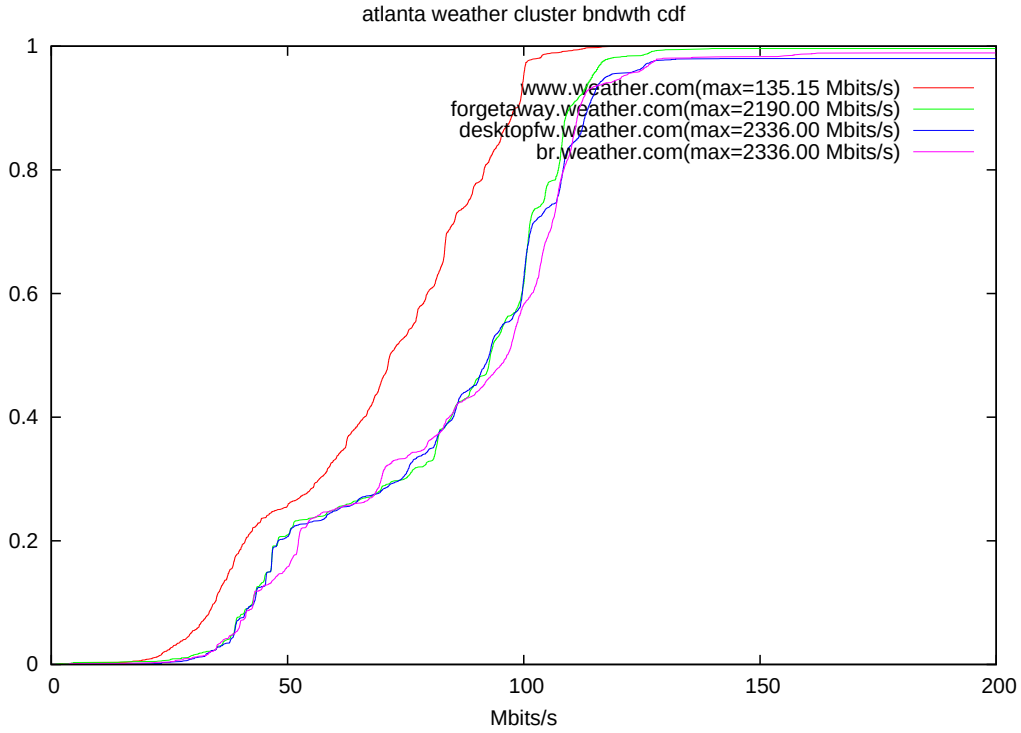

Figure 6.9: CDF of Available Bandwidth of Cluster weather.com at Atlanta, GA

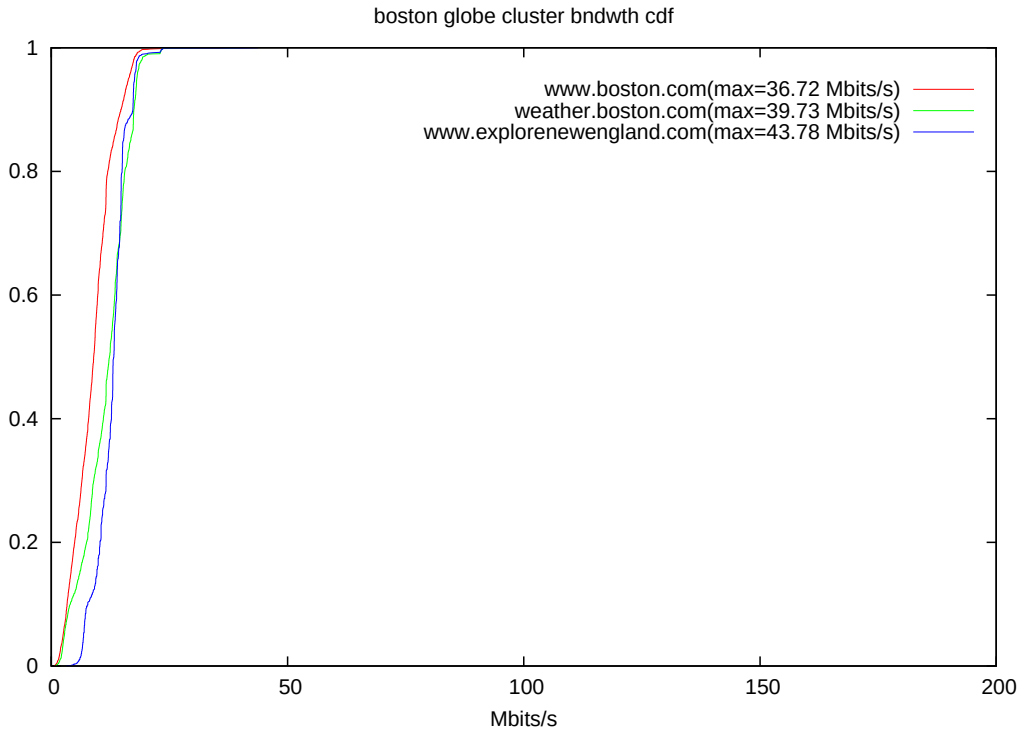

Figure 6.10: CDF of Available Bandwidth of Cluster Boston Globe at Boston, MA

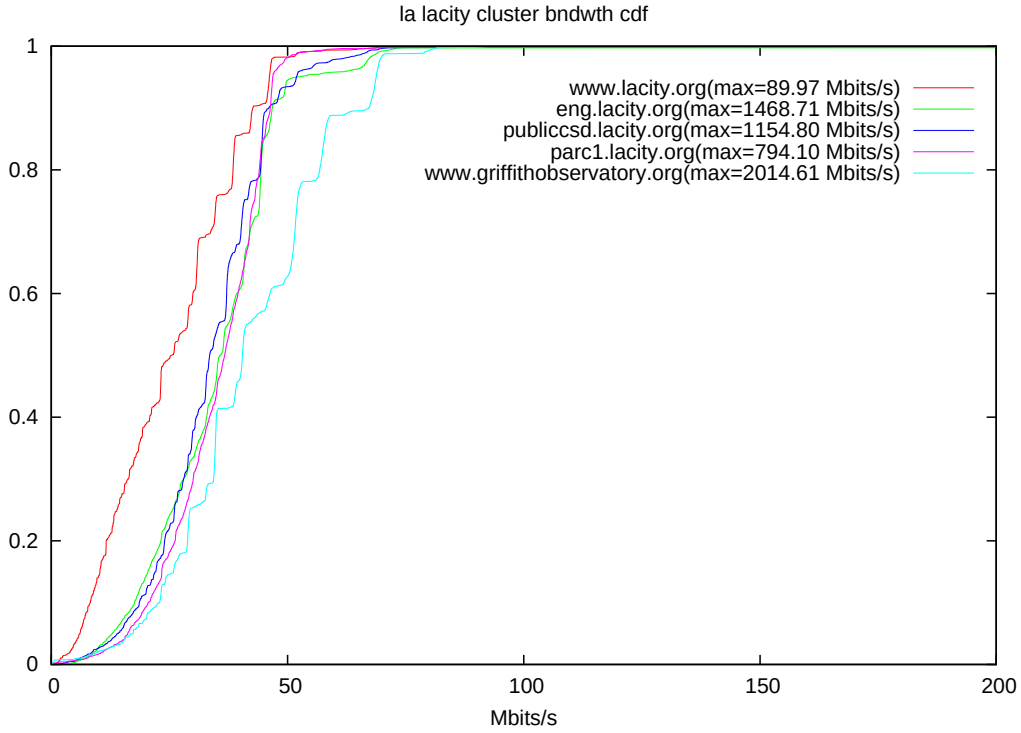

Figure 6.11: CDF of Available Bandwidth of Cluster LA city at Los Angeles, CA

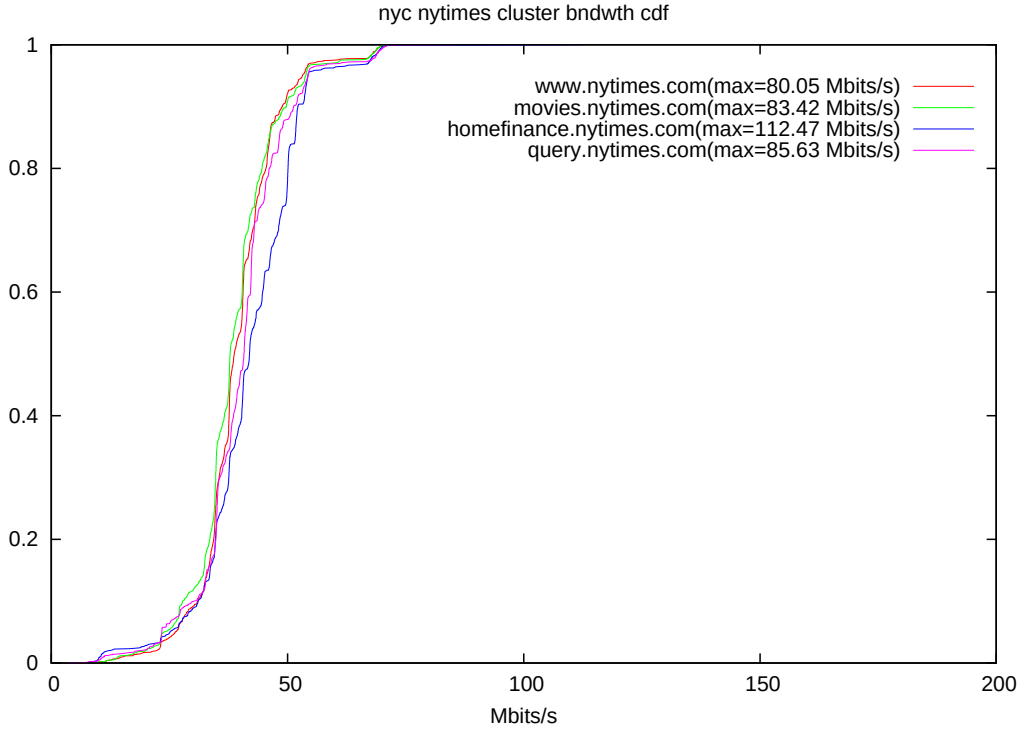

Figure 6.12: CDF of Available Bandwidth of Cluster NYTimes at New York, NY

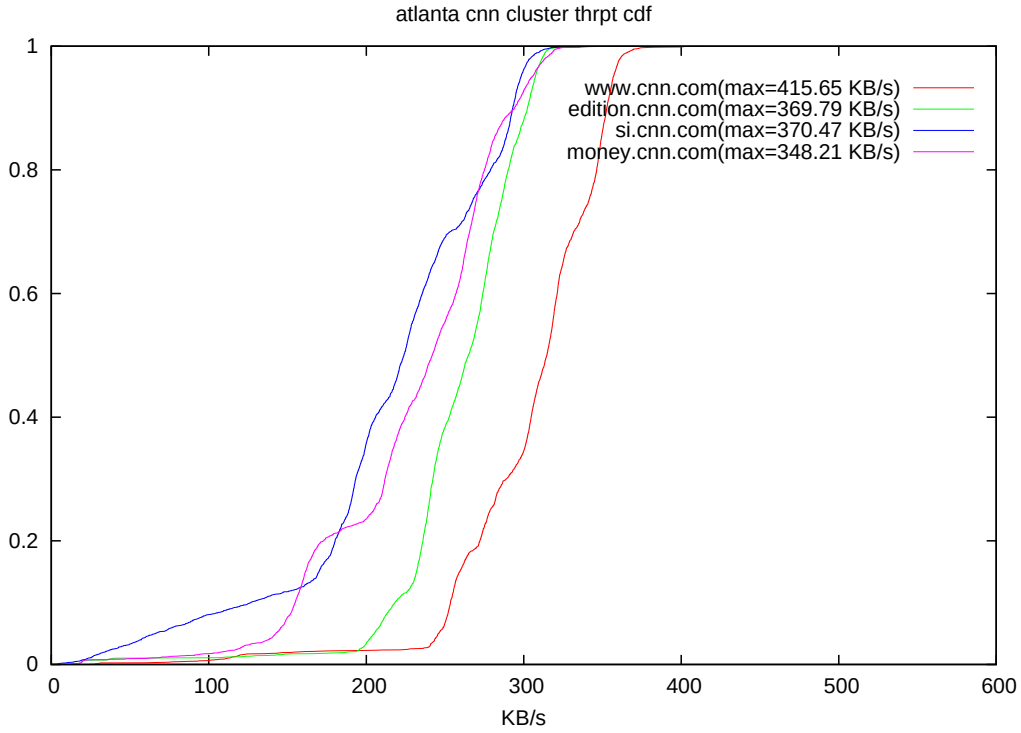

Figure 6.13: CDF of Throughput of Cluster CNN at Atlanta, GA

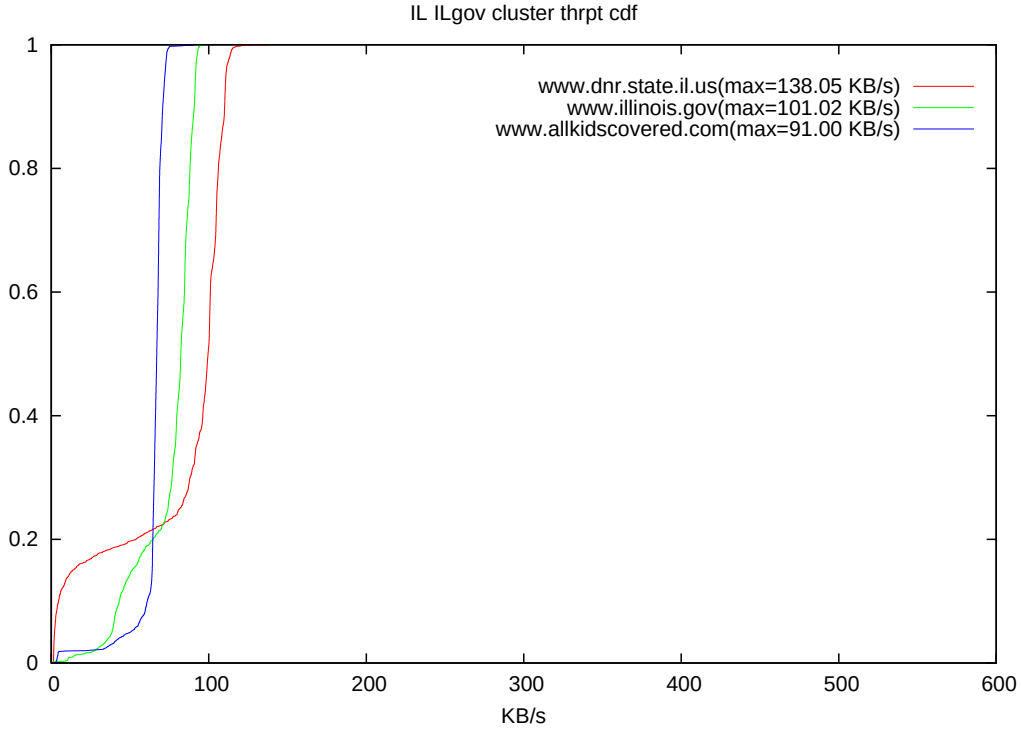

Figure 6.14: CDF of Throughput of Cluster IL Government at Springfield, IL

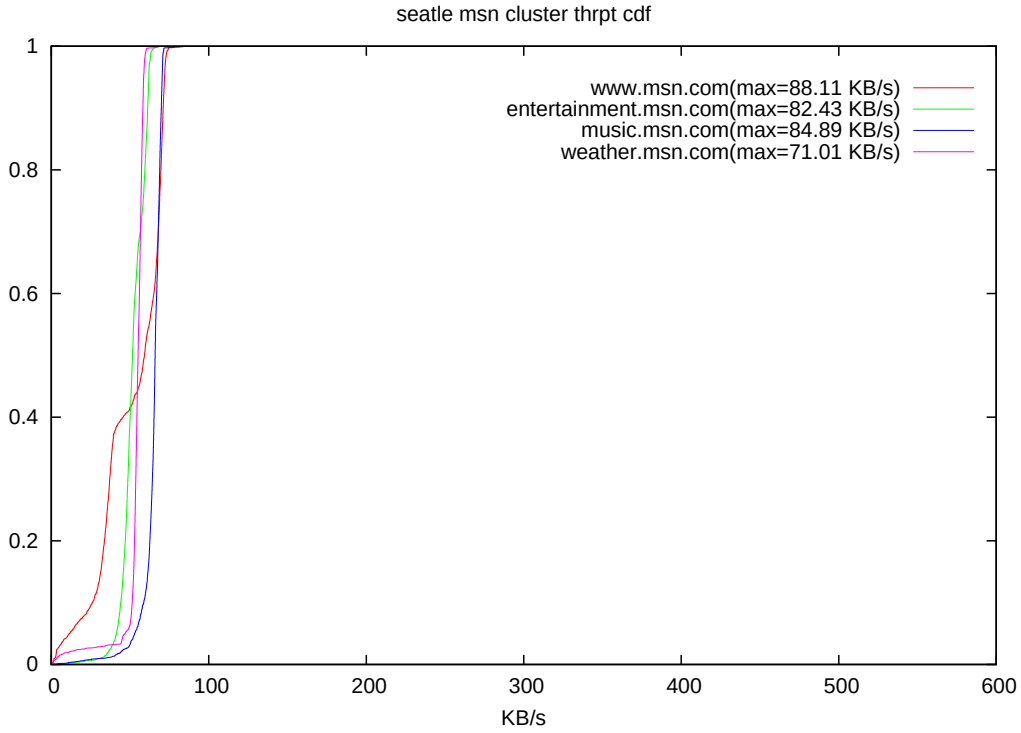

Figure 6.15: CDF of Throughput of Cluster MSN at Seattle, WA

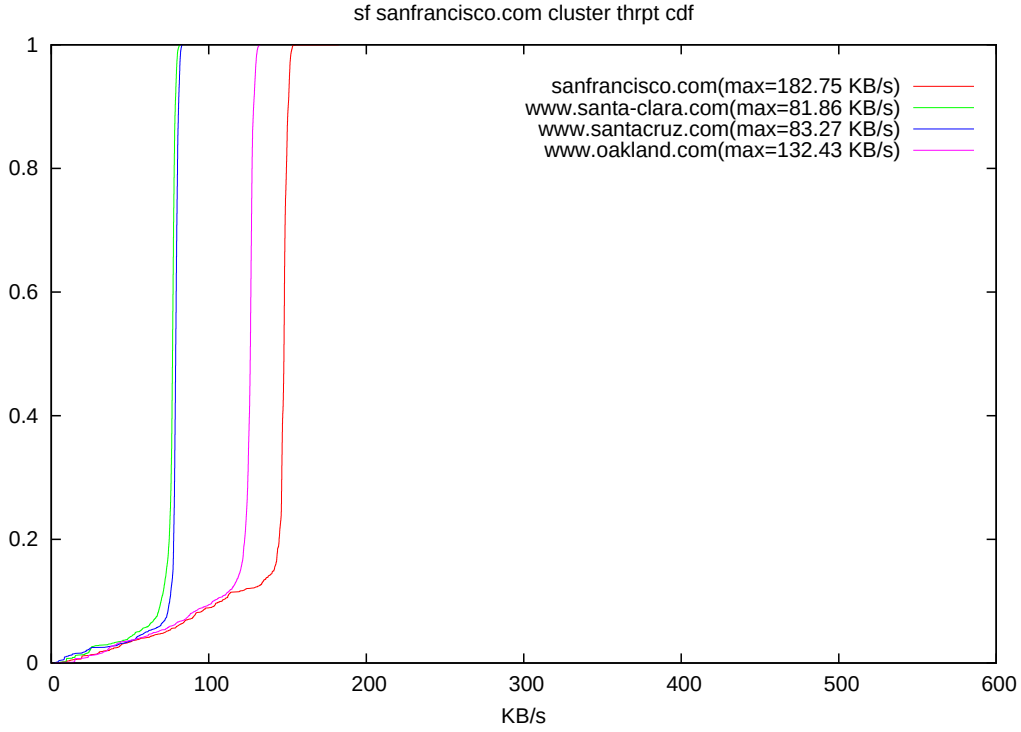

Figure 6.16: CDF of Throughput of Cluster SanFrancisco.com at San Francisco, CA

In contrast, Figures 6.17 through 6.20 show more distribution variance. The variance is either exhibited over the data set for one server, as in server www.boston.com in cluster Boston Globe, or among multiple servers belonging to one cluster, as in cluster Dallas News.

As we have analyzed the TCP inner workings previously, the overall throughput for a server having a long RTT from WPI, depends more on the RTT value than the bandwidth in the early rounds of a connection, since the server has to stall in each round to wait for ACKs from the client. The object sizes at the servers we study are mostly 50KB - 100KB. A connection only took a few RTTs to finish. Server stalling dominated most of the connection, especially for servers with longer RTTs. So we expect to see connection throughput increase as RTT decreases, and vice versa.

We can basically observe this behavior in the RTT and throughput graphs. For example, servers at cluster CNN have a relatively shorter RTT than those at cluster MSN, as shown in Figures 6.1 (RTT 30ms), 6.5 (RTT 90ms), 6.13 (throughput 250KB/s) and 6.15(throughput 50KB/s).

6.1.4 Connection Health Ratings

Since the health rating of a connection only takes three possible values, where 2 means 'good', 1 'medium', and 0 'bad', we use histograms instead of CDF graphs to present its distribution in this section. We find from the histograms that most of the clusters and the servers have a health rating of 2, and the health rating remains

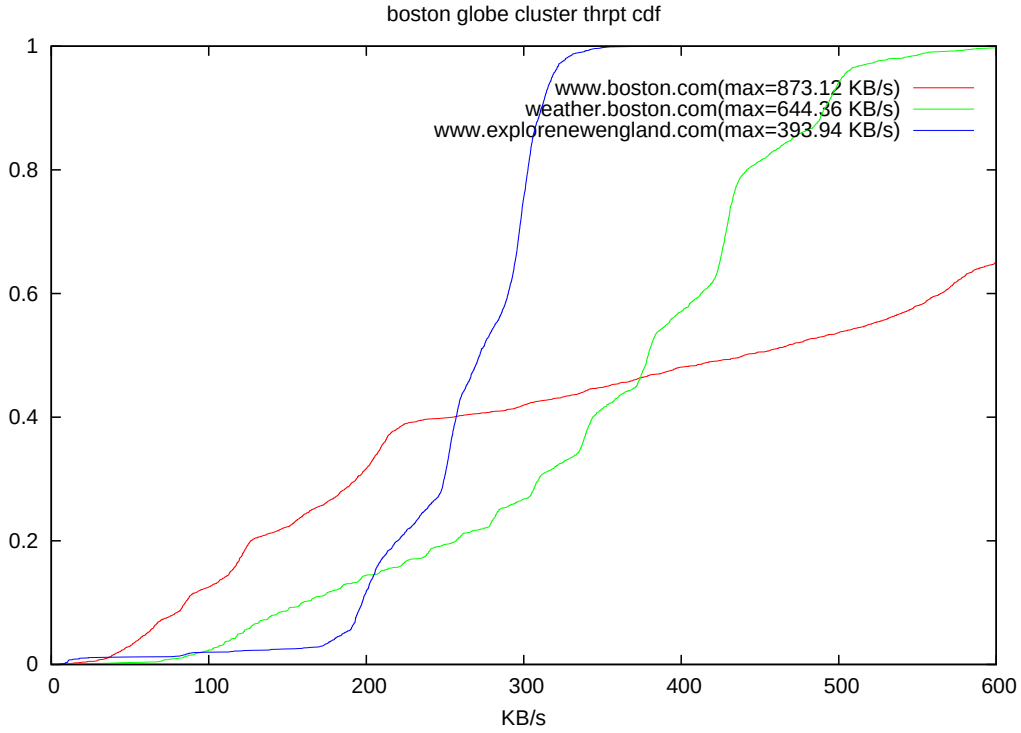

Figure 6.17: CDF of Throughput of Cluster Boston Globe at Boston, MA

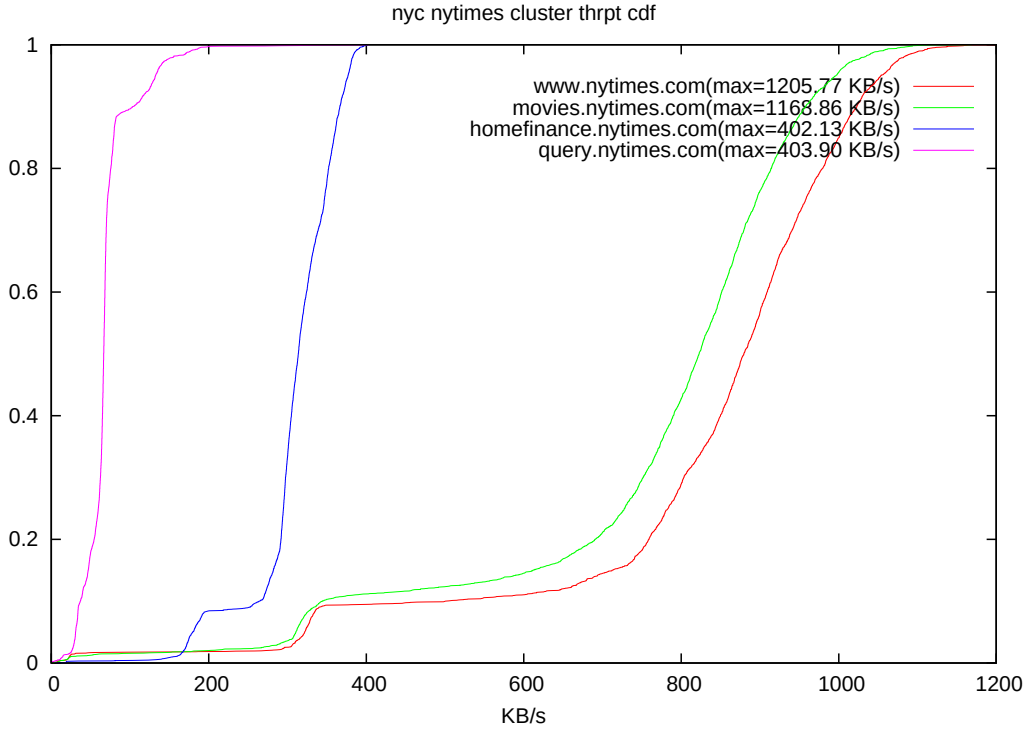

Figure 6.18: CDF of Throughput of Cluster NYTimes at New York, NY

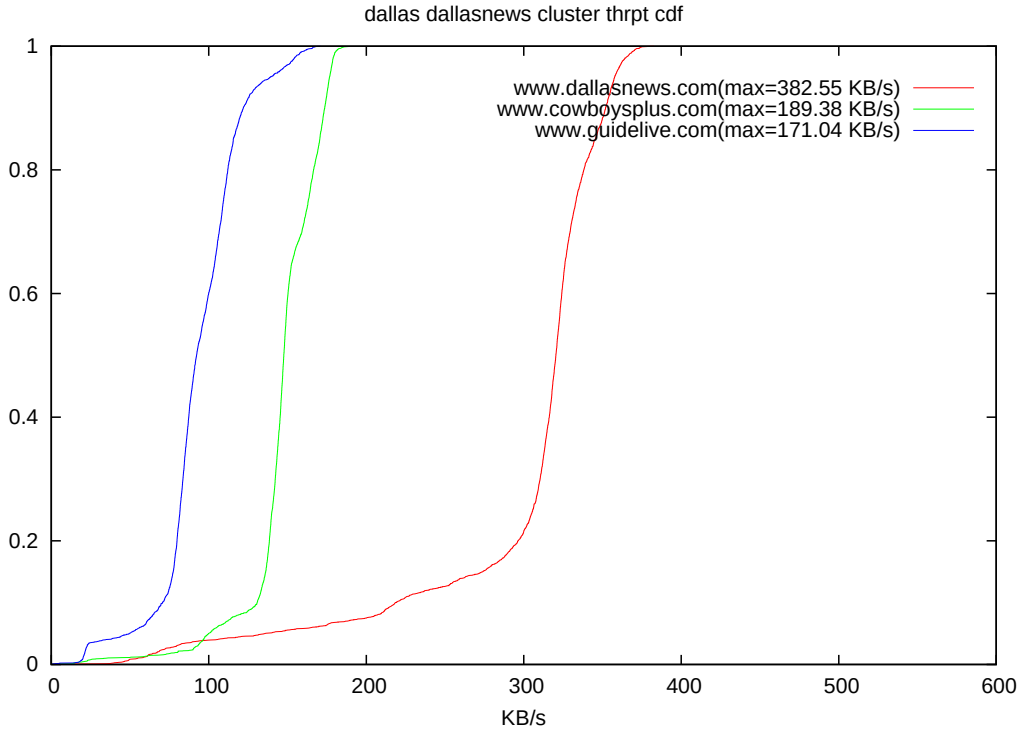

Figure 6.19: CDF of Throughput of Cluster Dallas News at Dallas, TX

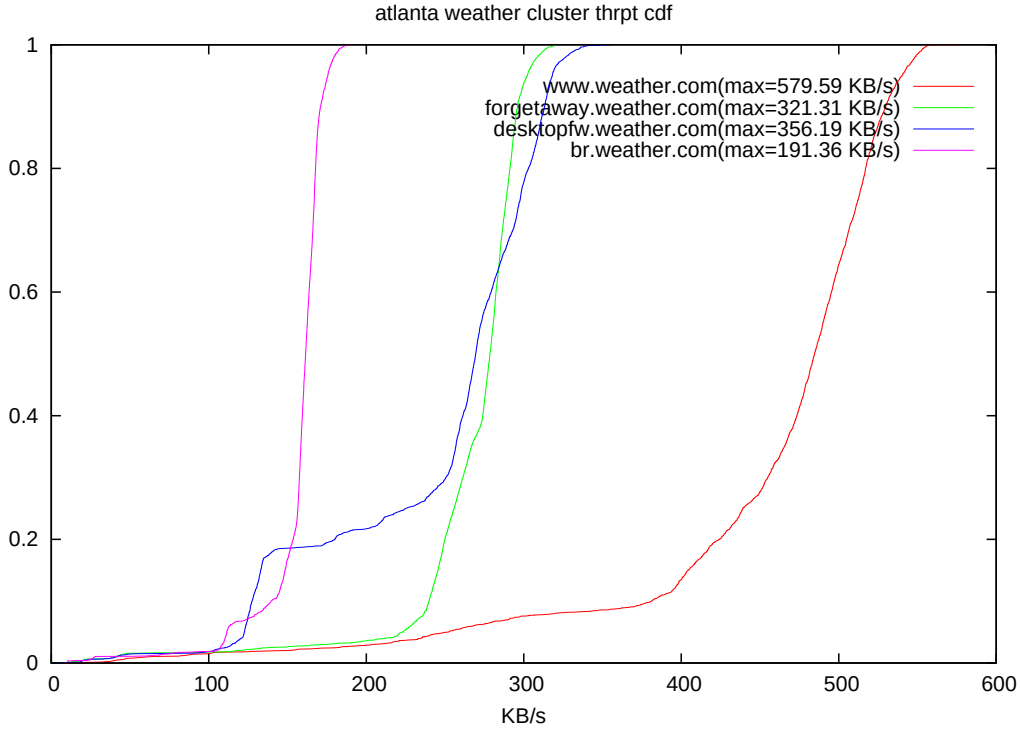

Figure 6.20: CDF of Throughput of Cluster weather.com at Atlanta, GA

stable over time. See Figures 6.21 through 6.26.

We can also see that the clusters with a health rating of 2 have good and concentrated distribution of RTTs and connection throughput, which further validates the connection health rating algorithm. For example, cluster IL Government, in Figures 6.3, 6.14 and 6.24.

Figures 6.27 and 6.28 are two clusters where servers' connection health ratings varied between two values instead of having one concentrated value. If we look at their RTTs and throughput, there are also large variance in these two metrics, which coincides with the connection health rating distribution.

6.2 Predictions

In this section, we discuss how well predictions can be made using the exponential predictor model proposed in the background.

6.2.1 Choices of Parameters

In our experiments, we find that exponential decay predictor achieves good prediction accuracies while keeping the computation simple. Therefore, we use exponential decay predictor for our metric prediction.

We use different λ values that can possibly affect the prediction accuracy. From our preliminary work we determined that lambda did not affect overall prediction. Hence this time we select fewer values with a more even distribution. The λ values we use are $\lambda = 0, 0.25, 0.5, 0.75$. In theory, the smaller λ is, the more importance

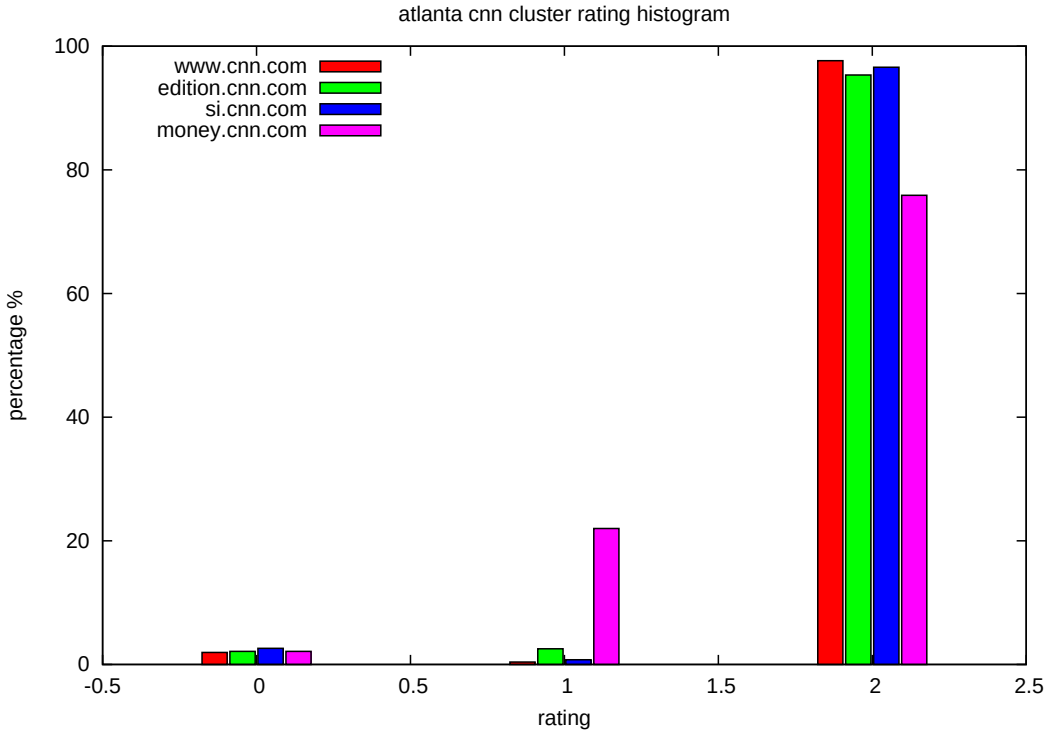

Figure 6.21: Health Ratings of Cluster CNN at Atlanta, GA

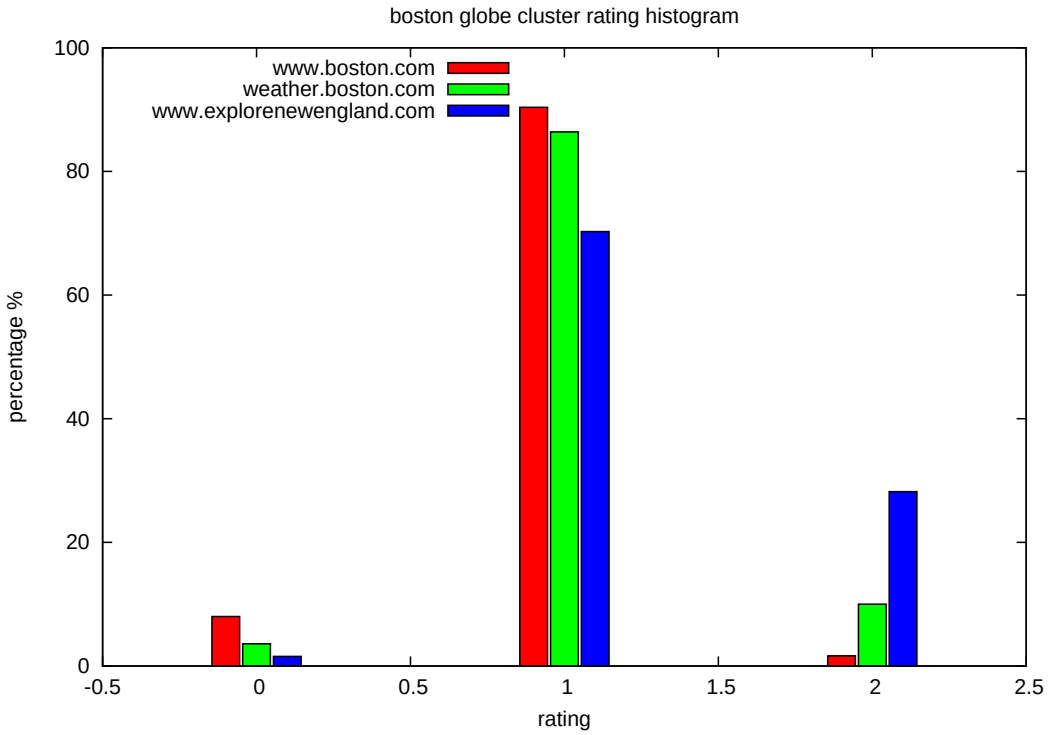

Figure 6.22: Health Ratings of Cluster Boston Globe at Boston, MA

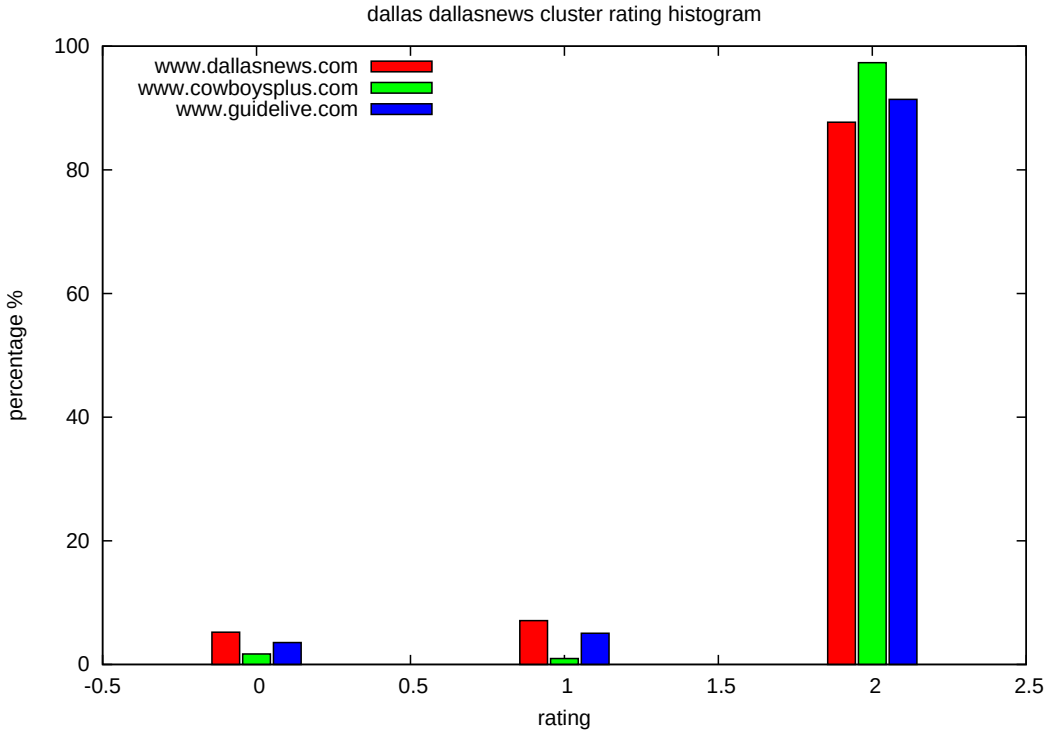

Figure 6.23: Health Ratings of Cluster Dallas News at Dallas, TX

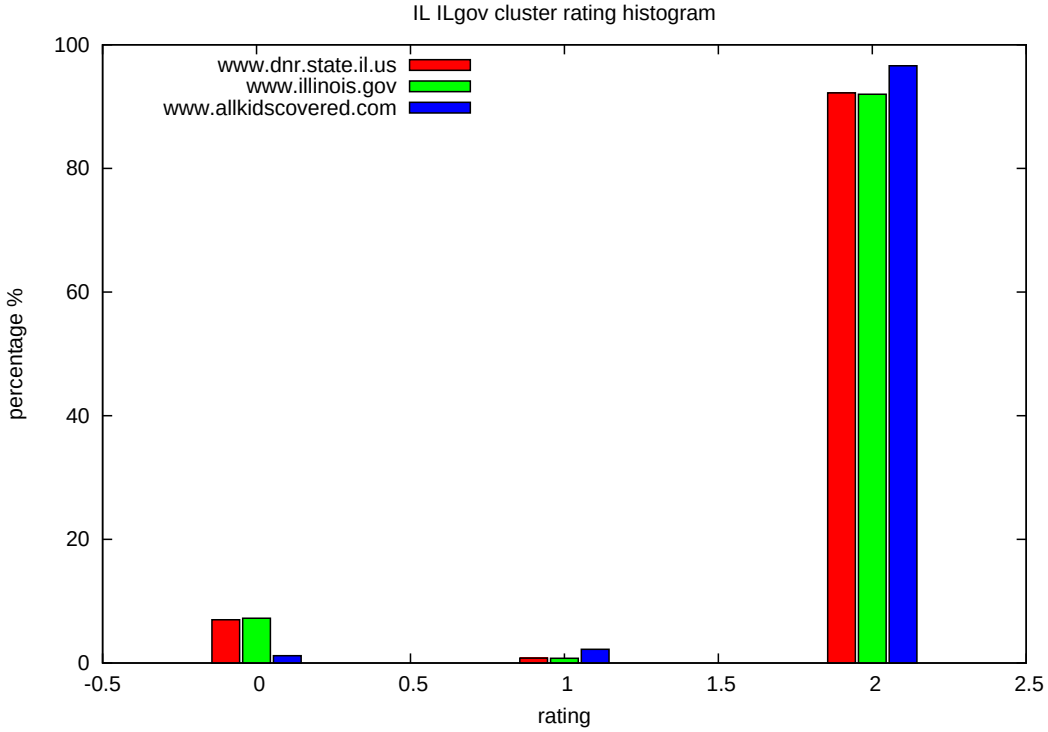

Figure 6.24: Health Ratings of Cluster IL Government at Springfield, IL

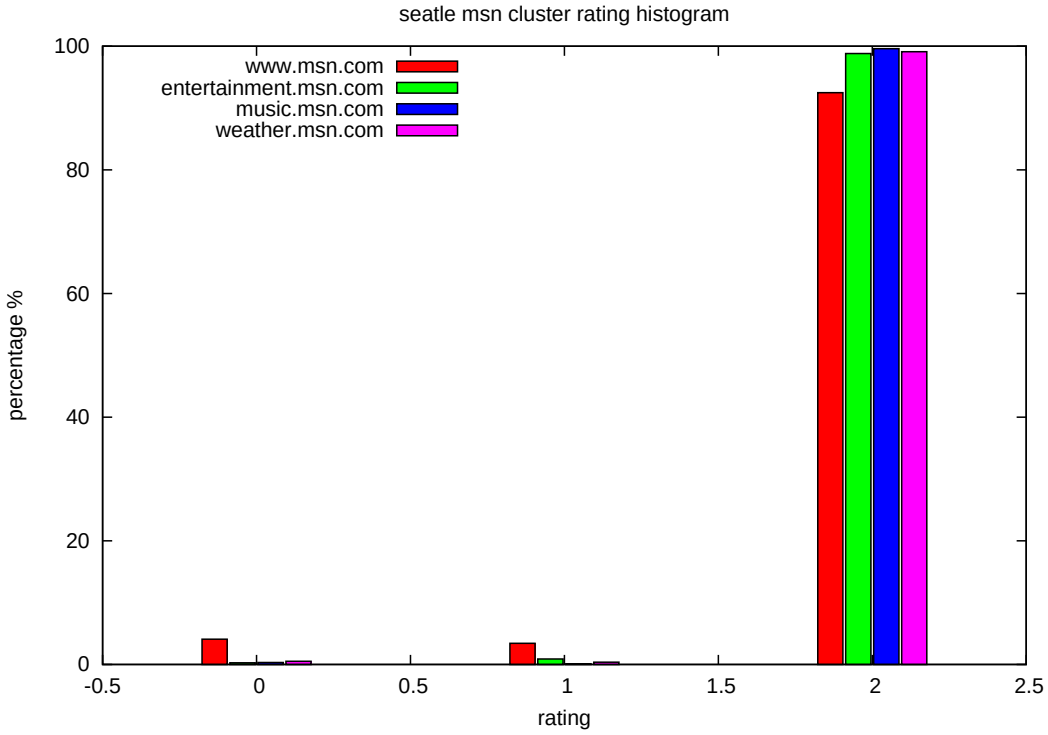

Figure 6.25: Health Ratings of Cluster MSN at Seattle, WA

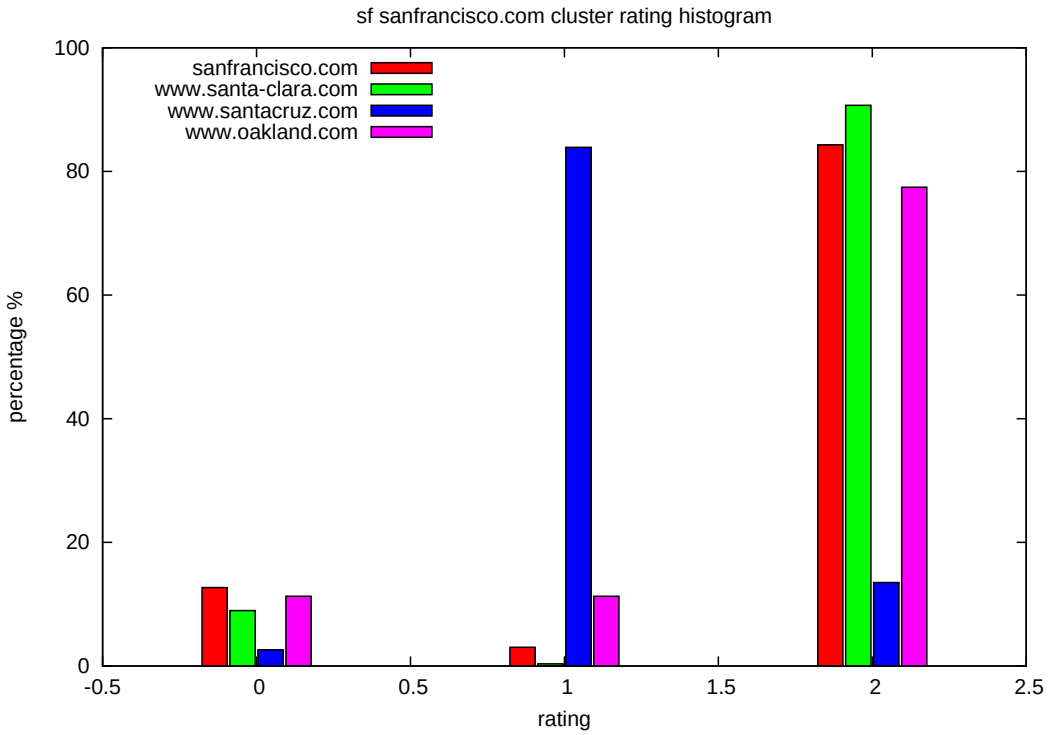

Figure 6.26: Health Ratings of Cluster SanFrancisco.com at San Francisco, CA

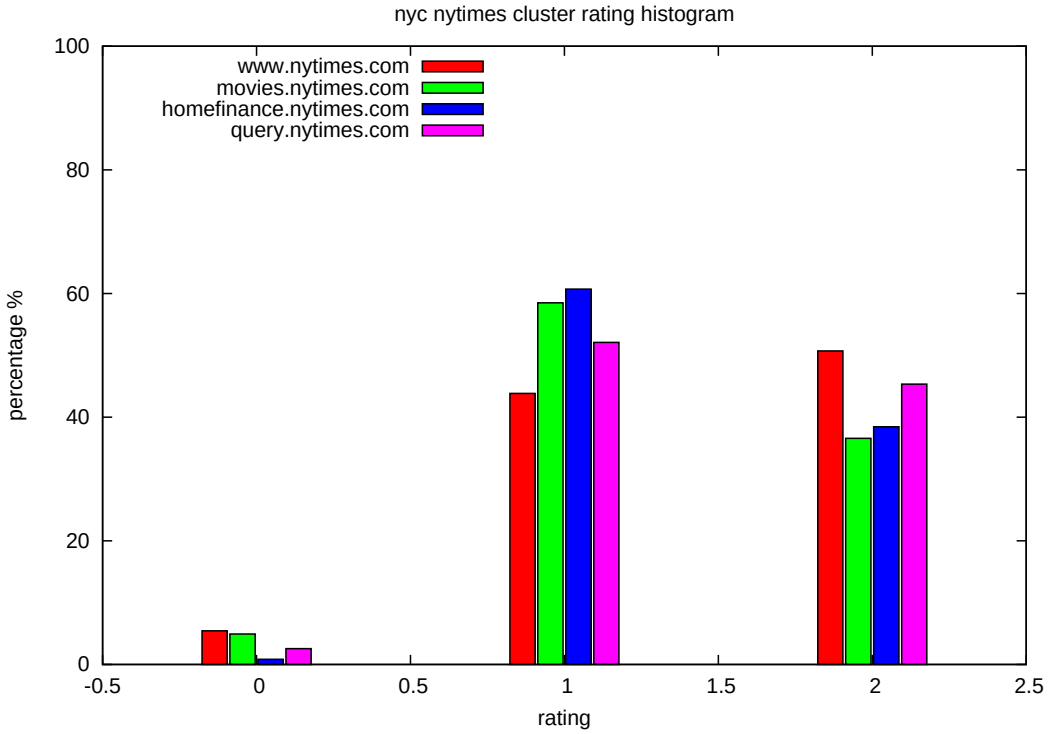

Figure 6.27: Health Ratings of Cluster NYTimes at New York, NY

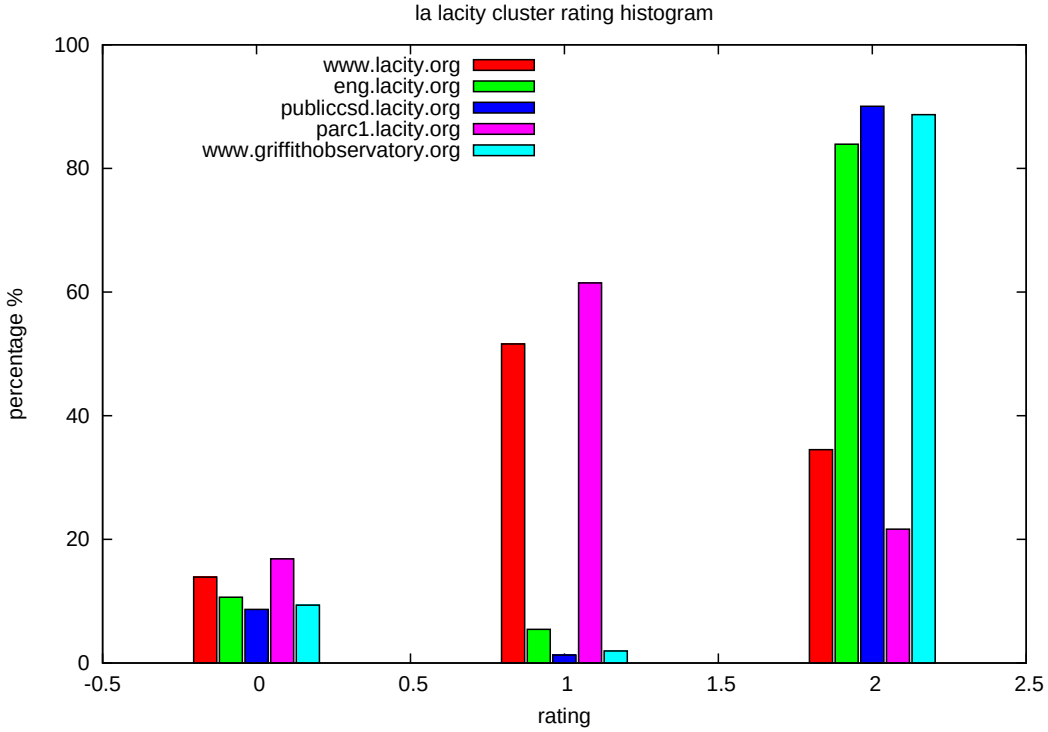

Figure 6.28: Health Ratings of Cluster LA city at Los Angeles, CA

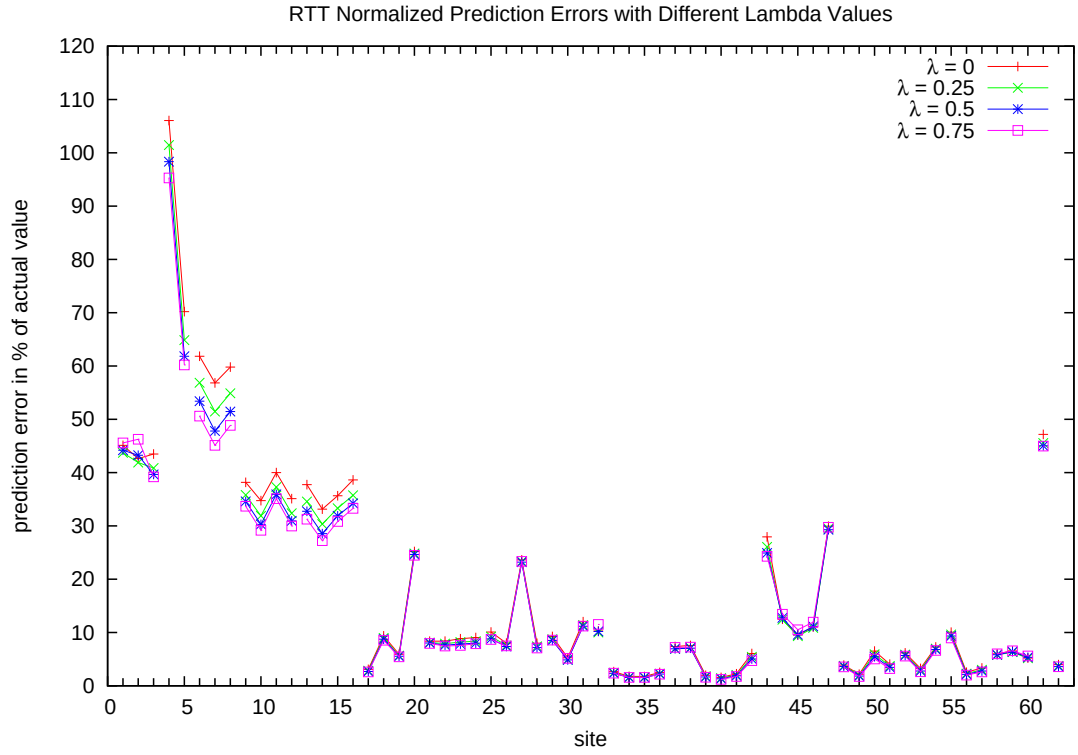

Figure 6.29: RTT Normalized Prediction Errors with Different Lambda Values

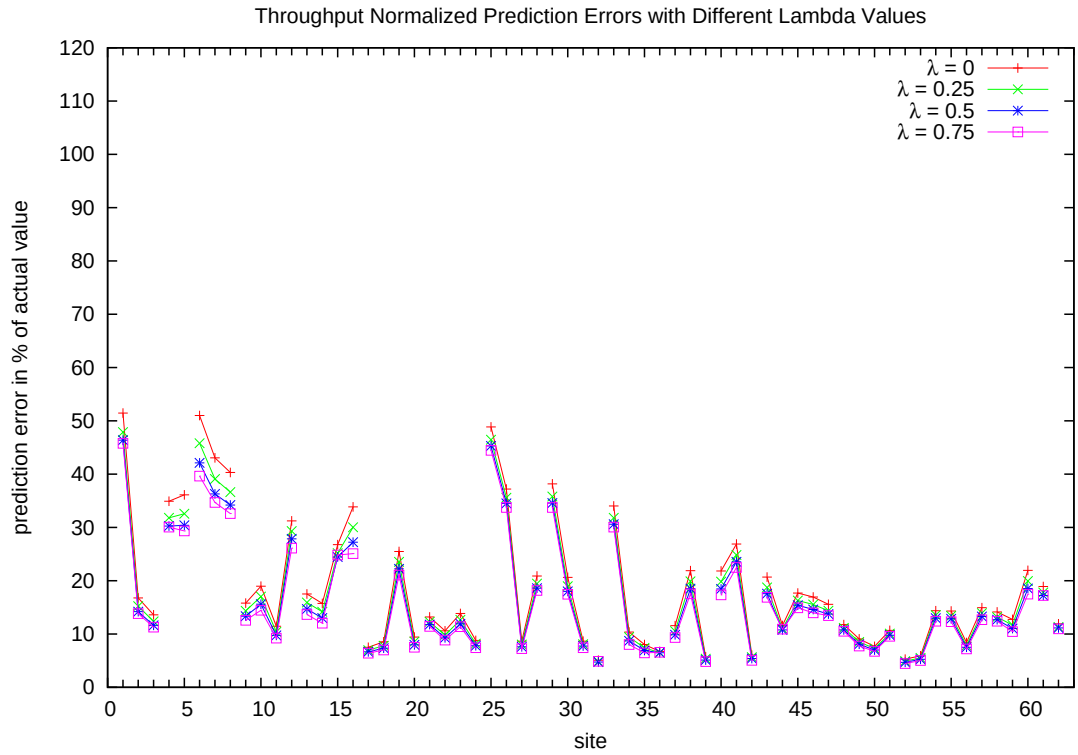

Figure 6.30: Connection Throughput Normalized Prediction Errors with Different Lambda Values

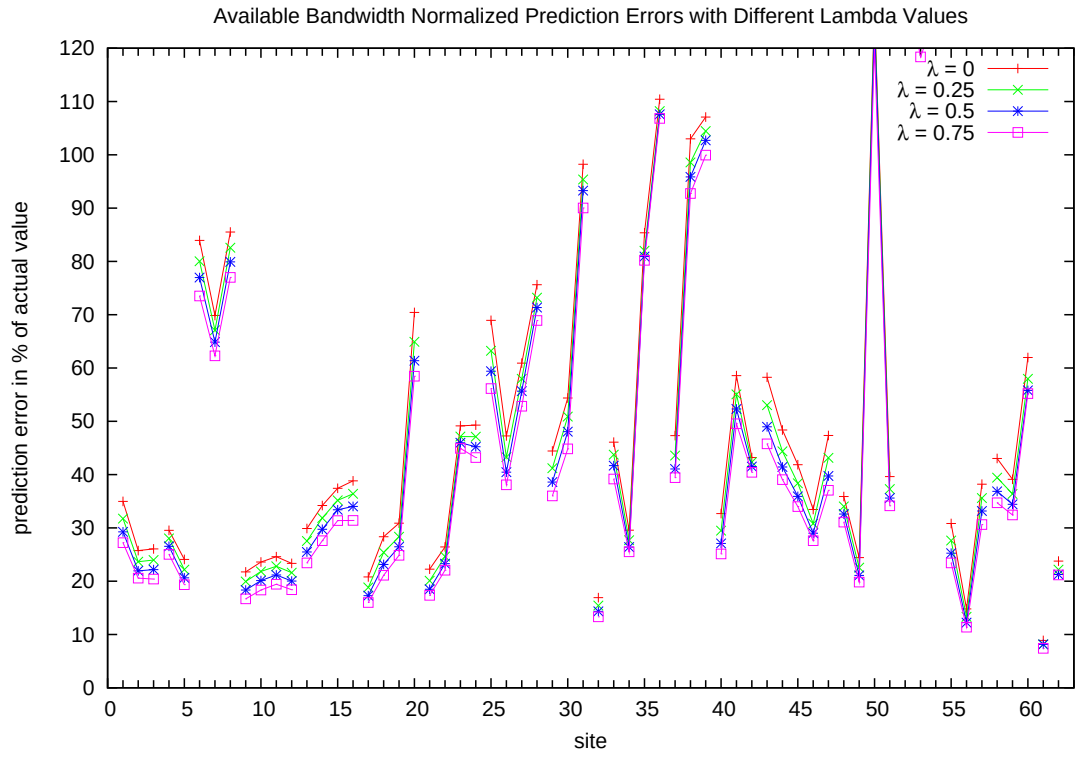

Figure 6.31: Available Bandwidth Normalized Prediction Errors with Different Lambda Values

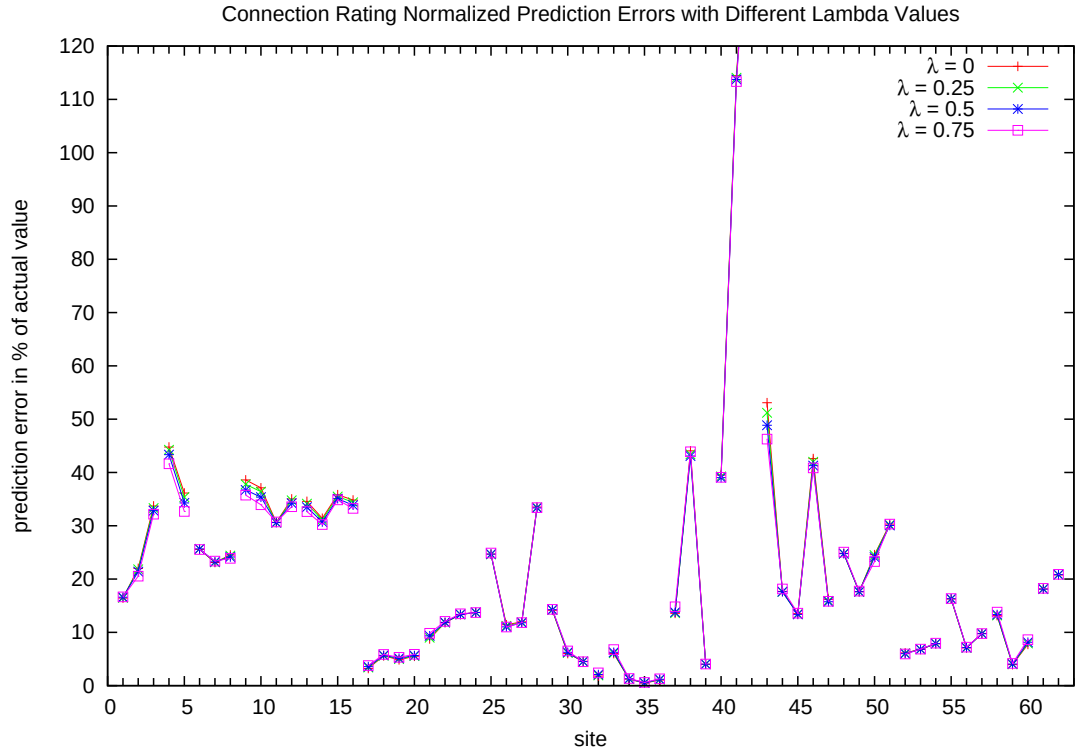

Figure 6.32: Connection Health Rating Normalized Prediction Errors with Different Lambda Values

is placed on the more recent accesses; the larger λ is, the more important is placed on a weighted average over all the previous accesses. However, in our experiment, we find that prediction accuracy does not rely much on the choice of λ, but on the variance of the data itself.

In Appendix A , Tables A.1 through A.8 show the detailed prediction results with different λ values, at each web server for RTT, connection throughput, available bandwidth and connection health ratings respectively. The prediction errors, i.e., the difference between the actual values and the predicted values, are expressed in the absolute values in each column for the corresponding λ values. In order to make prediction errors comparable, we also put in the tables their normalized values in the parenthesis after the absolute ones. The normalized value is expressed in the form of percentage and calculated as the ratio of the prediction error to the actual value. The web servers are numbered from 1 to 62 as shown in the tables. In Figures 6.29 through 6.32, we use these web server numbers and plot the normalized prediction errors in the percentage of the actual values for each λ values for different metrics. Web servers that belong to the same cluster are connected by lines in the figures.

As we can see, most prediction errors fall in the range of 10-20%(bandwidth is 20-30%), which is good or acceptable for some applications. Those which have higher prediction errors are normally servers with more variation. For example, server www.boston.com had large variation of throughput in Figure 6.17 and of bandwidth in Figure 6.10, and the prediction for both metrics have large errors too in Figures 6.30 and 6.31. We can also see servers from the same cluster do not necessarily have similar prediction errors and can vary significantly. We believe this

is also due to difference in individual server behavior and metric variance over time.

In most of the cases, λ values do not make a significant difference, except for a few web servers and clusters such as "MBTA" and "NYTimes" in RTT and bandwidth prediction, where larger λ values achieve smaller prediction errors. As we can see in Fig 6.29 - 6.32, most lines with different λ values coincide with each other. We believe this is due to the stability of the data set itself, therefore the λ value used is not critical to prediction accuracy.

We also plotted the absolute values of prediction errors shown in Figures 6.33 through 6.36 in comparison to the normalized ones. As we can see some servers could have similar values of absolute prediction errors, while the normalized prediction errors in terms of the ratio of prediction errors to the actual values could be different. For example, in Figure 6.36, server 1-3, 6-16 and 21-24 all have similar absolute prediction errors, but in Figure 6.29, their normalized prediction errors are more varied. This discrepancy between actual and normalized errors is because their actual values are different and the same amount of data change do not have the same affect on different servers.

In the cases where we see large prediction errors for certain metrics for some servers, we show here some CDF graphs for normalized prediction errors to have a closer look at whether the choice of λ makes a difference for connections with large variation. Figure 6.37 shows the RTT normalized prediction errors for www.mbta.com (server 2 in Figure 6.29), where we do not see significant difference in the overall distribution of normalize prediction errors. So is the case with other metrics and other servers. Figure 6.38 shows another example, the bandwidth normalized prediction

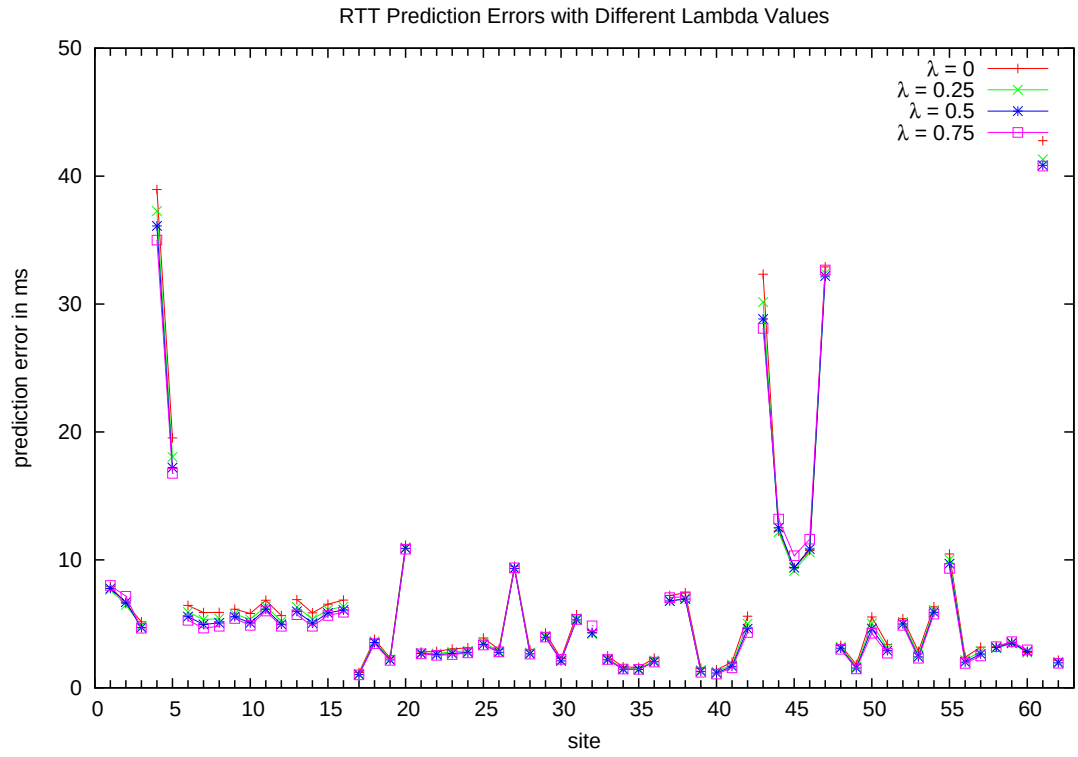

Figure 6.33: RTT Prediction Errors with Different Lambda Values

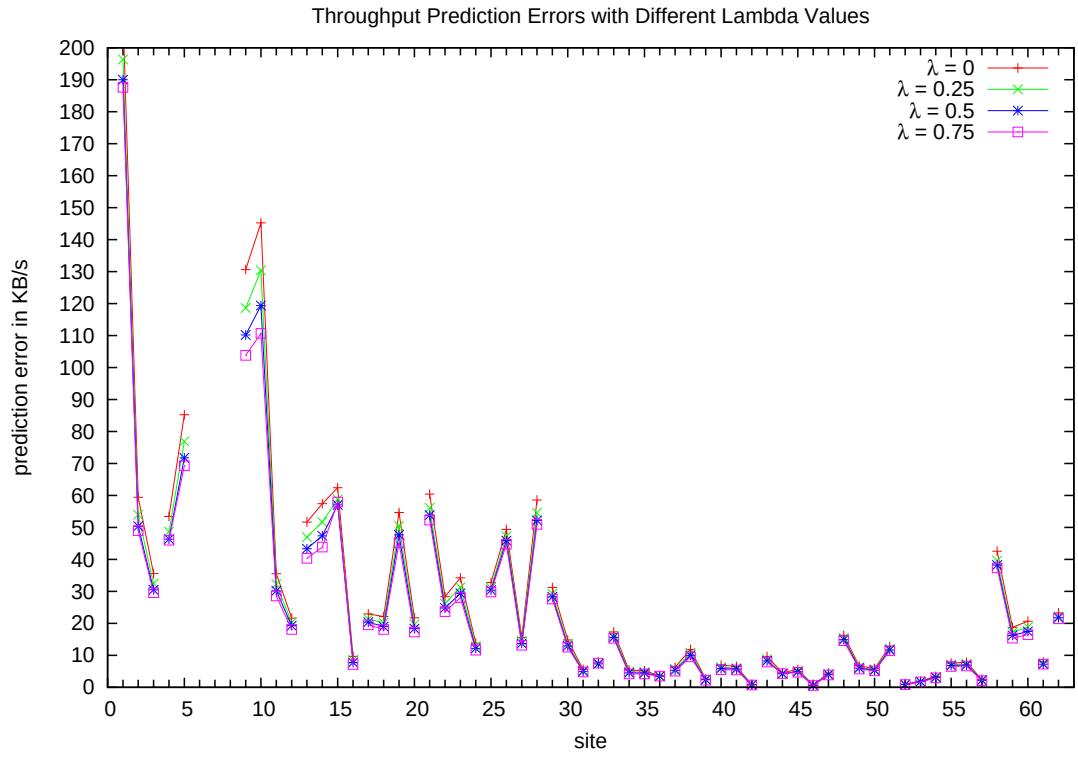

Figure 6.34: Connection Throughput Prediction Errors with Different Lambda Values

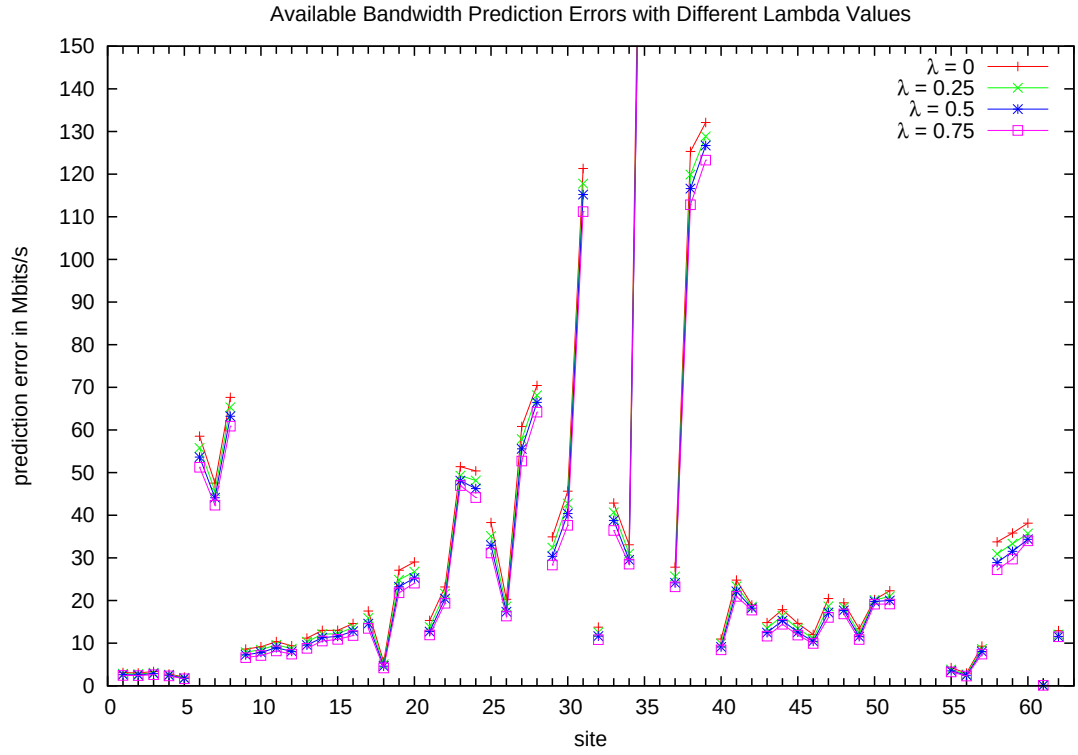

Figure 6.35: Available Bandwidth Prediction Errors with Different Lambda Values

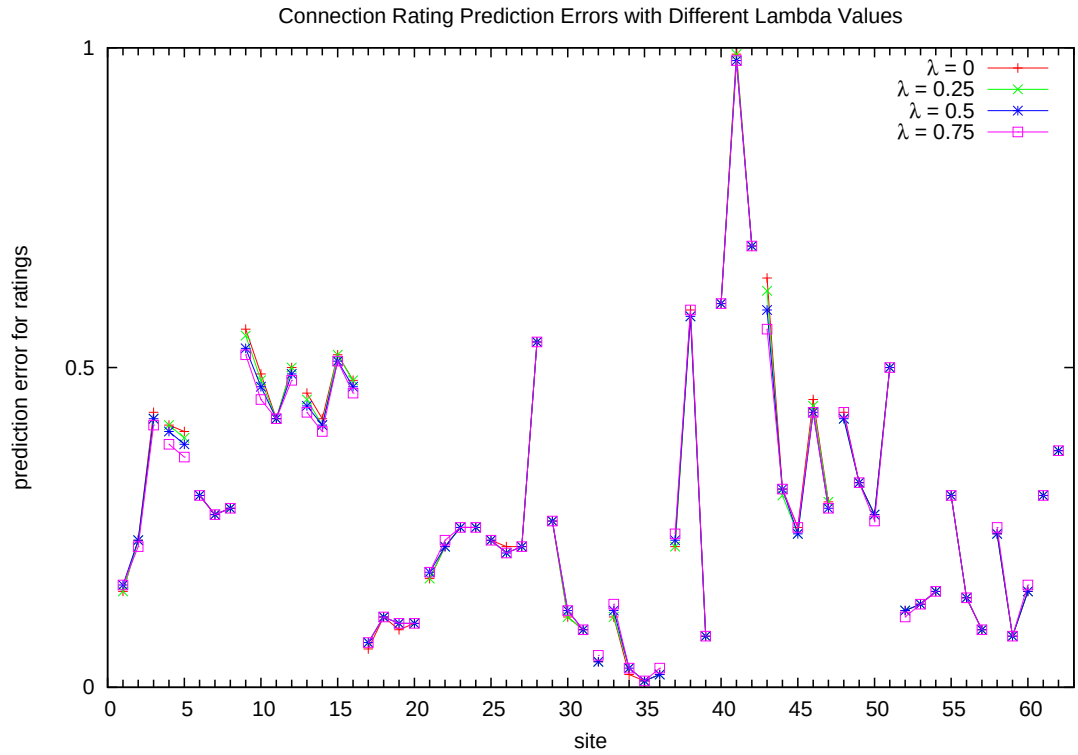

Figure 6.36: Connection Health Rating Prediction Errors with Different Lambda Values

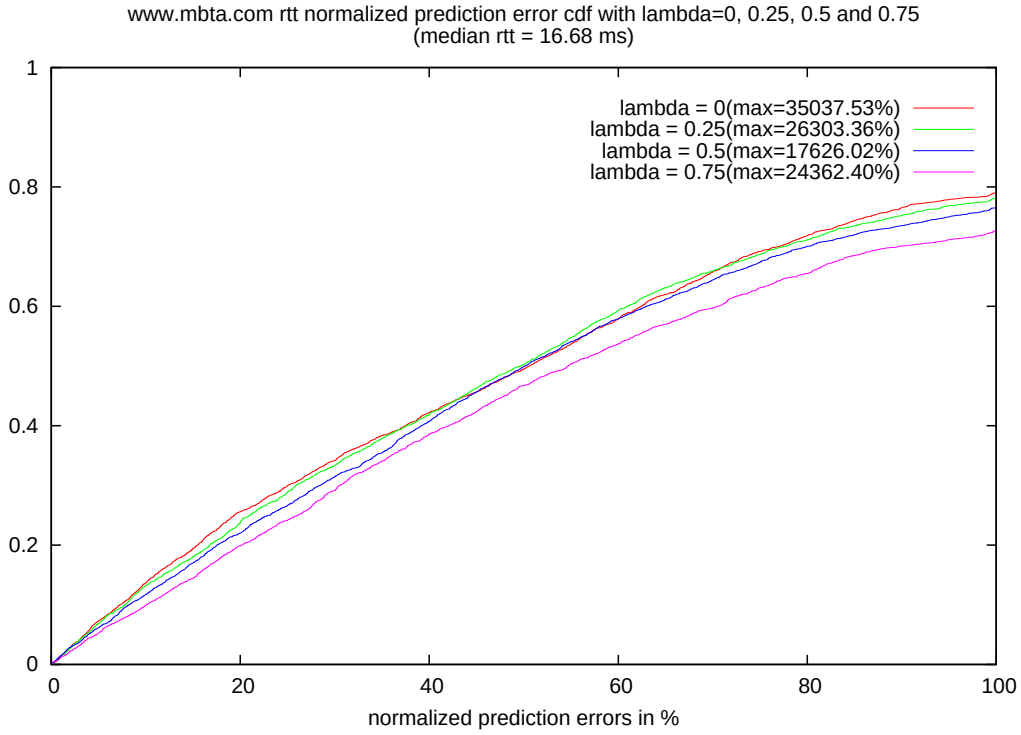

Figure 6.37: CDF of RTT Normalized Prediction Errors for www.mbta.com with Different Lambda Values

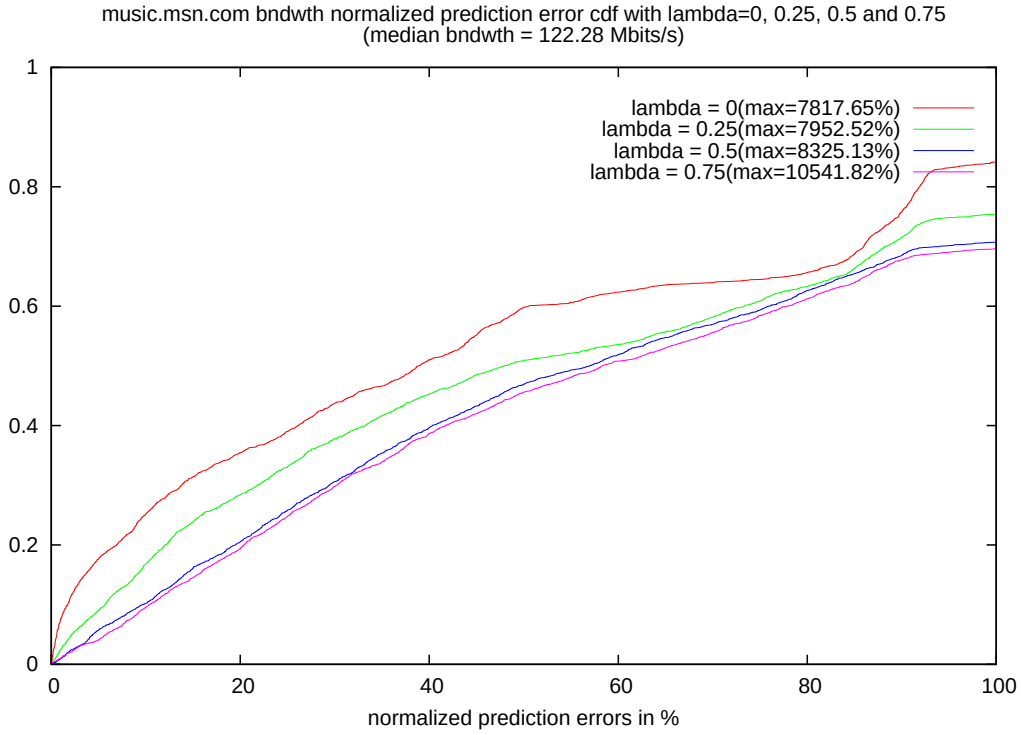

Figure 6.38: CDF of Bandwidth Normalized Prediction Errors for music.msn.com with Different Lambda Values

errors for music.msn.com (server 35 in Figure 6.31).

6.2.2 Choices of Data Collection Intervals

Another aspect of the prediction is to investigate whether the frequency of accesses significantly affects the prediction accuracy. Our default collection interval for web servers was ten minutes. We extract from the original data set data points of every one hour, every two hours and every four hours respectively, and use them together with the original data set to do prediction. We compare the prediction accuracy for the different collection intervals. In all the cases here, we fix λ value as $\lambda = 0.5$. The detailed prediction results with different collection intervals for RTT, connection throughput, available bandwidth and connection health ratings are listed in Table A.9 through A.16 in Appendix A. In the same way for different λ values, we plot the normalized prediction errors in the percentage of the actual values for each collection interval for each metric, with servers from the same clusters connected by lines. See Figures 6.39 through 6.42.

Interestingly enough, as we can see from these figures, using large collection intervals, for example four hours, does not degrade prediction accuracy in most cases. However, there are a few exceptions where larger collection intervals do cause degradation of prediction accuracy, for example, cluster "Boston Globe" and cluster "MBTA" in RTT, throughput, and available bandwidth. In both cases, the variance of the data itself determines the prediction accuracy in reaction to the choice of the collection intervals. For a relatively stable set of data, use of a large collection interval does not impact prediction accuracy, while it does for a more varied data

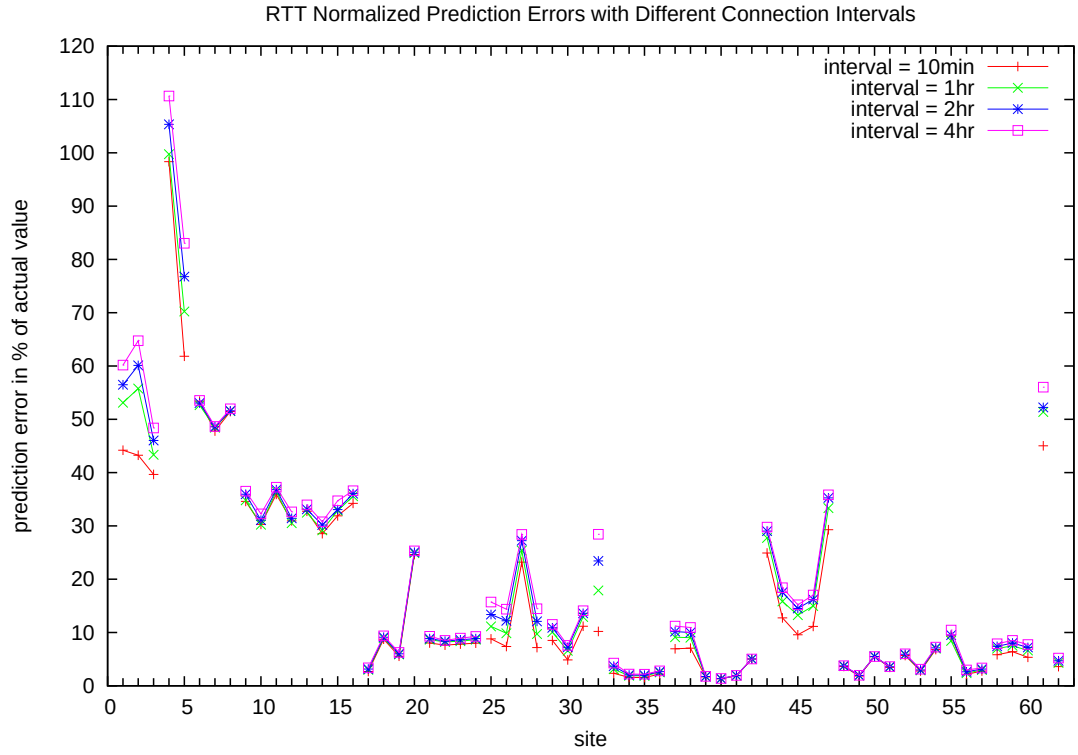

Figure 6.39: RTT Normalized Prediction Errors with Different Connection Intervals

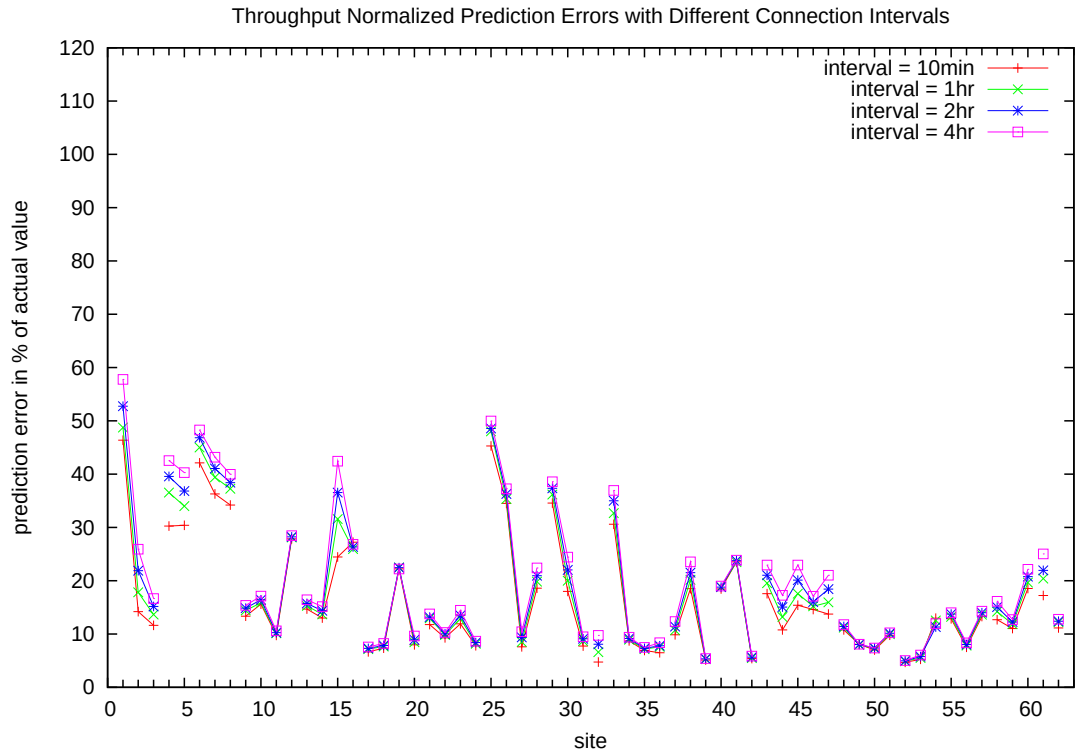

Figure 6.40: Connection Throughput Normalized Prediction Errors with Different Connection Intervals

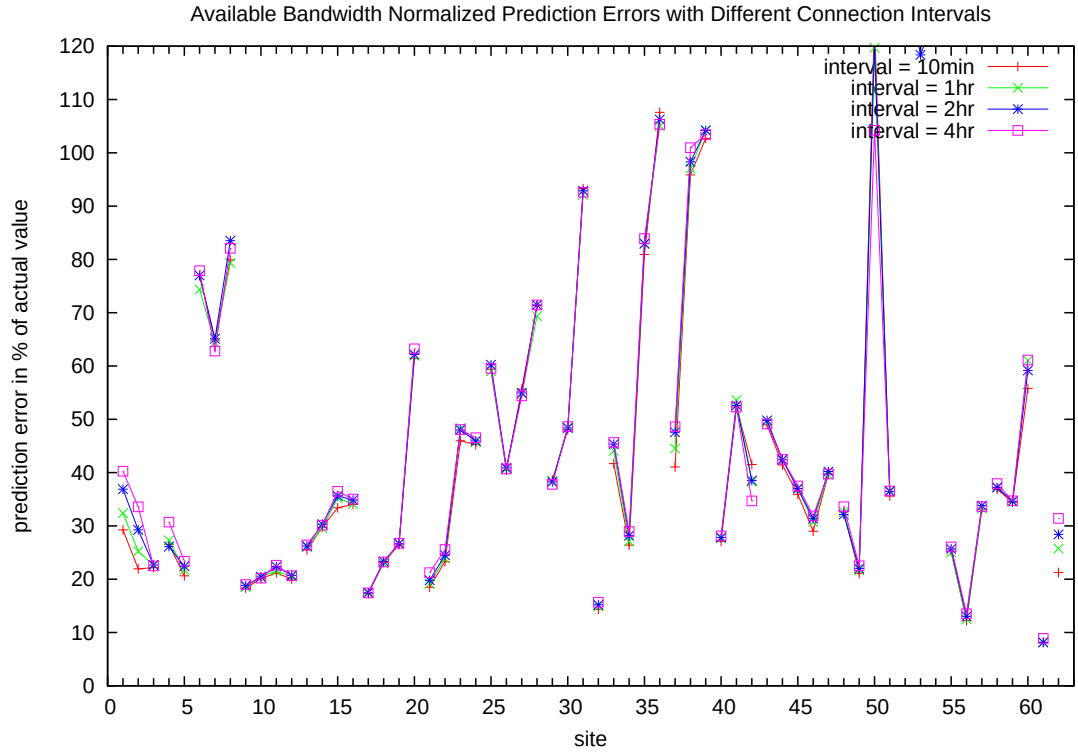

Figure 6.41: Available Bandwidth Normalized Prediction Errors with Different Connection Intervals

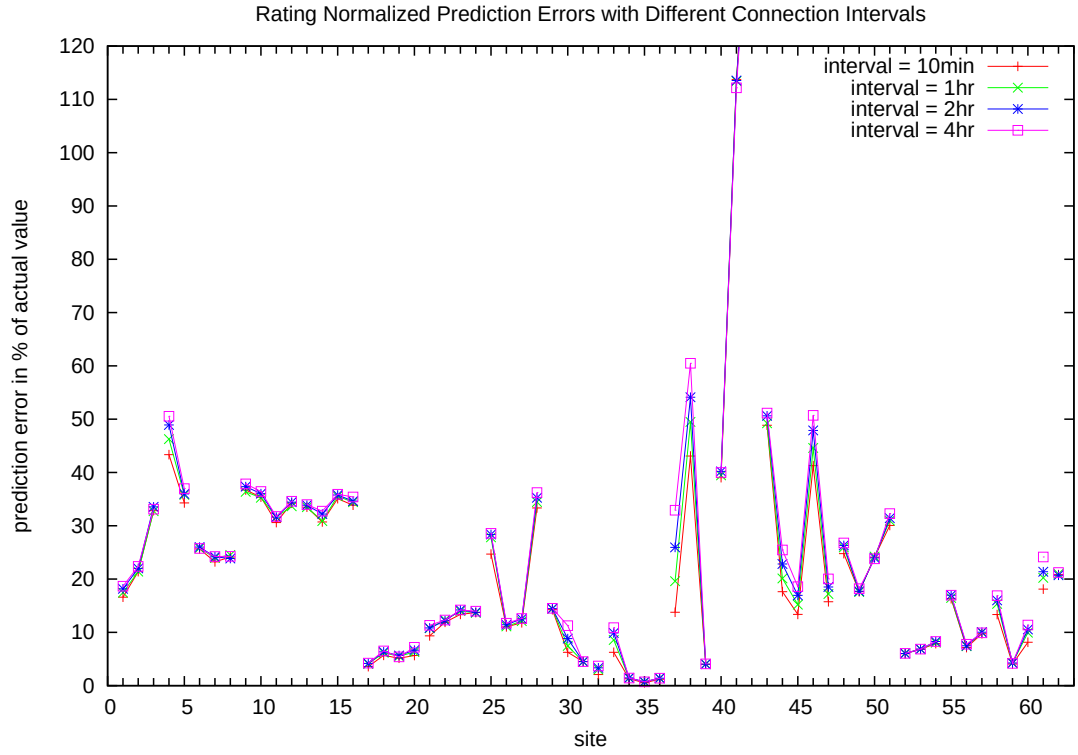

Figure 6.42: Connection Health Rating Normalized Prediction Errors with Different Connection Intervals

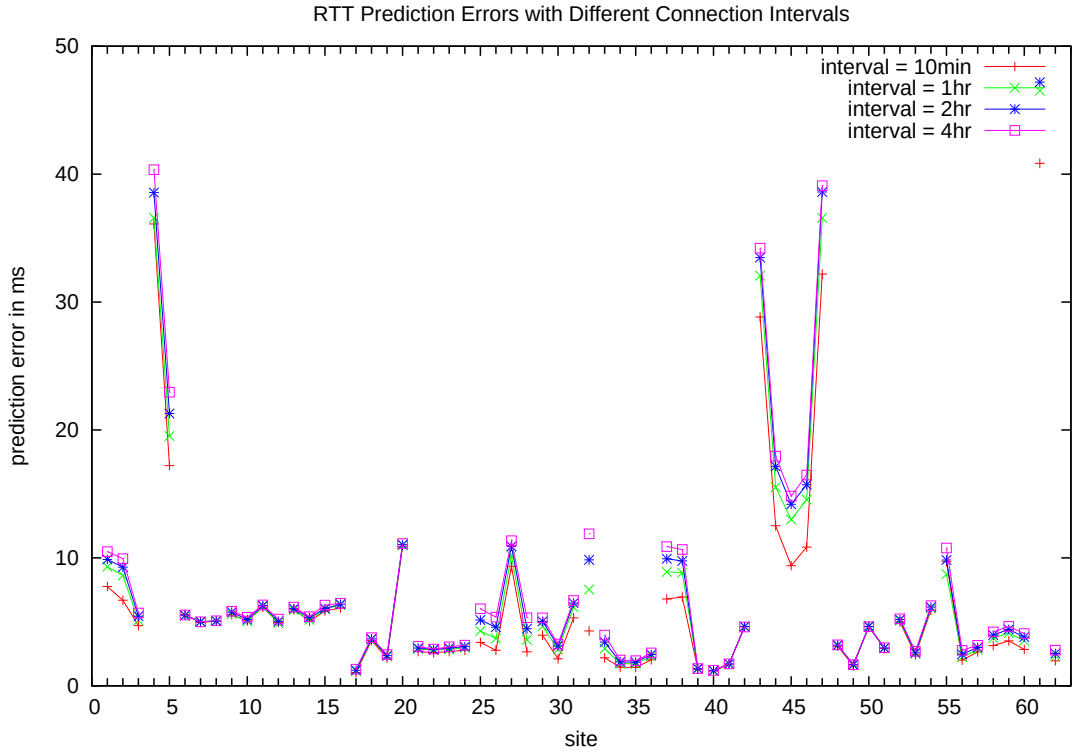

Figure 6.43: RTT Prediction Errors with Different Connection Intervals

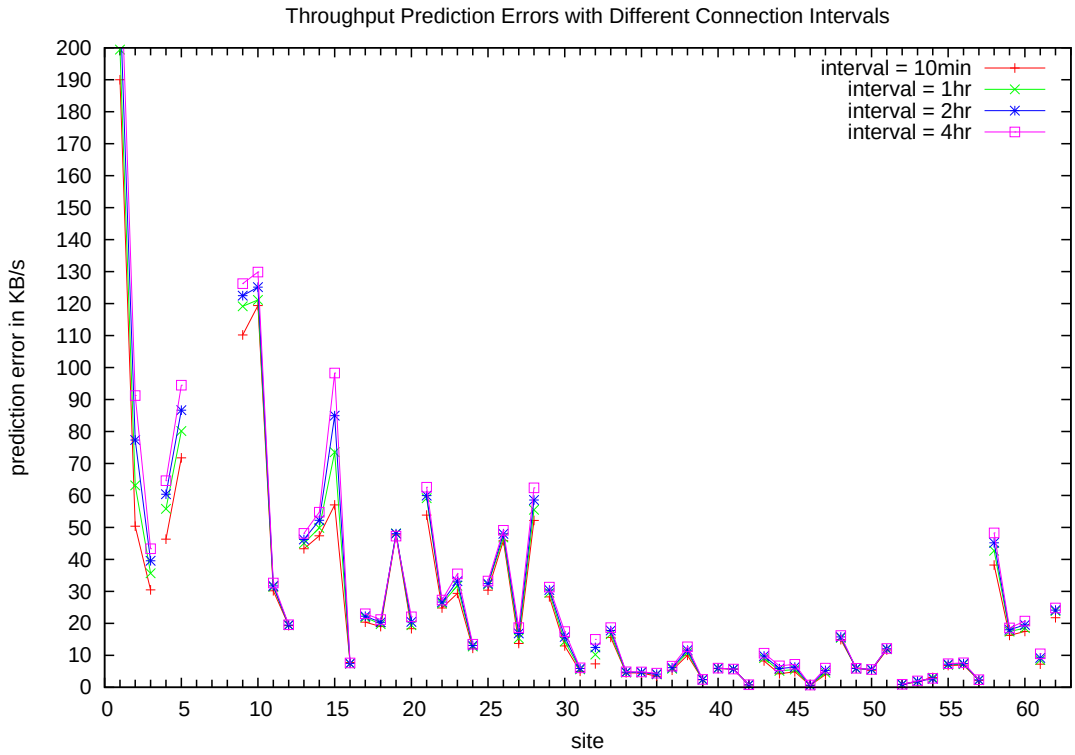

Figure 6.44: Connection Throughput Prediction Errors with Different Connection Intervals

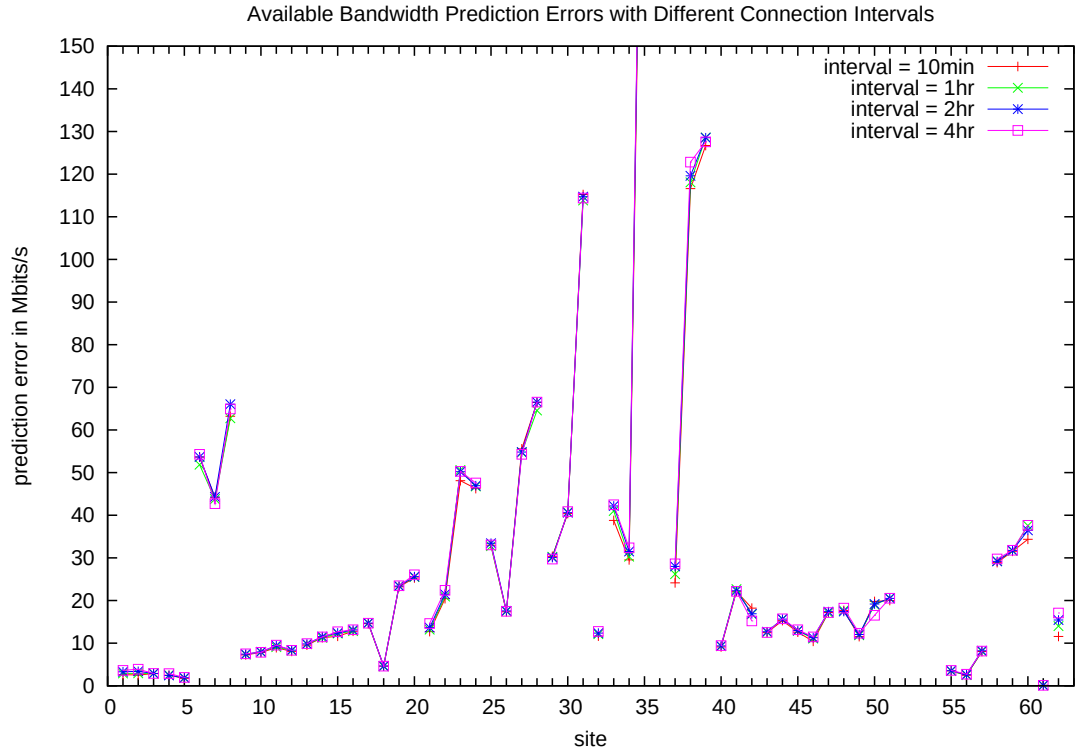

Figure 6.45: Available Bandwidth Prediction Errors with Different Connection Intervals

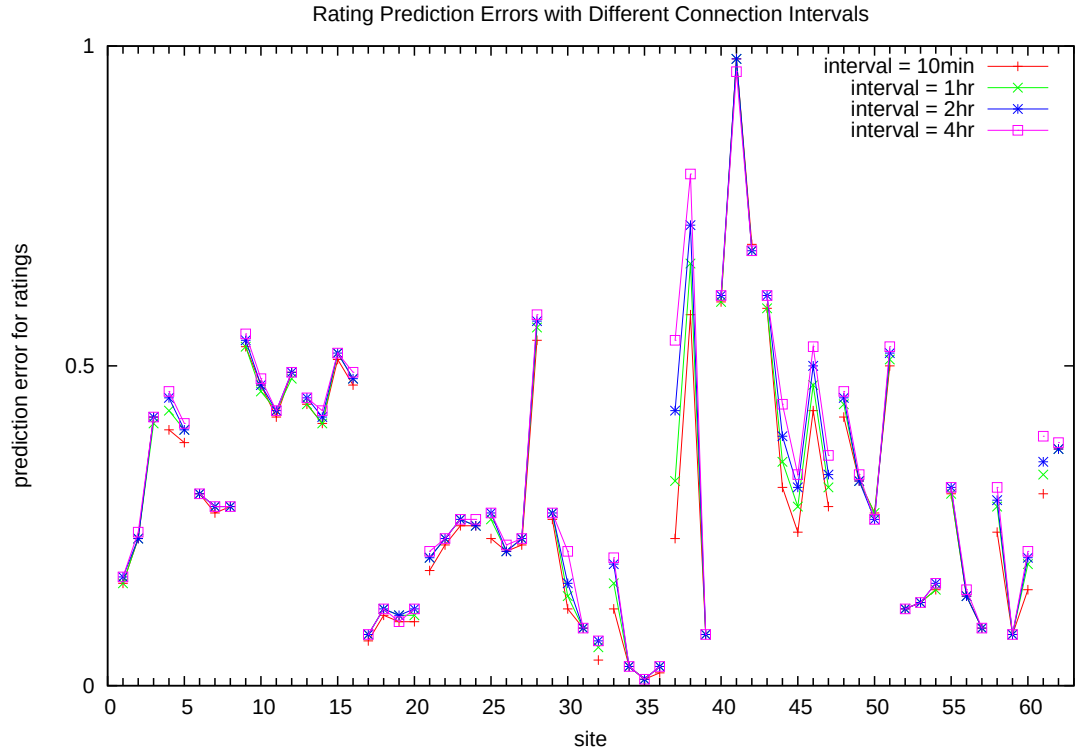

Figure 6.46: Connection Health Rating Prediction Errors with Different Connection Intervals

set.

Similarly as for the choice of λ values, Figures 6.43 through 6.46, show the absolute prediction errors for different collection intervals.

6.3 Summary

The work in this section covered two main topics. We first studied the variations in network metrics using the data collected in Chapter 5. We can see the different behavior of these metrics, e.g. RTT is stable, but throughput varies a lot with object sizes.

We also saw that time-based predictions using the exponential predictor gave good prediction results in most cases. Using normalized prediction values we were able to study the accuracy of these variations. In the following chapters we see some more predictions and their analysis.

Chapter 7

Topology-Based Prediction

In this chapter, we discuss how well network metrics' prediction can be made using data from other web servers or other clusters. We want to find out if the change in data of one metric at one server is also reflected on another server.

As we mentioned in earlier chapters, a database at a local cluster can be built to make predictions about the next user access, using information on previous accesses to the same server as well as to other server locations. The choice of λ values for the predictor and time intervals of using has been discussed in the previous chapter. The task for this chapter is to investigate the possibility of utilizing accesses to other servers to make prediction for a given server. In order to verify this idea, we want to find out if there is such correlation of network metrics among servers.

7.1 Correlation of Actual Values of Network Metrics Among Web Servers

Before studying the correlation of data variance of network metrics among web servers, we want to investigate the correlation of the actual data itself. For each of

the network metrics: RTT, connection throughput, available bandwidth and connection health rating, we calculate the correlation coefficient for every web server pair. For correlation coefficients, a commonly accepted mathematical understanding of their values is as follows,

$$\begin{cases} 0 \leq |r| < 0.5 & \textit{weak correlation} \\ 0.5 \leq |r| < 0.8 & \textit{moderate correlation} \\ 0.8 \leq |r| \leq 1 & \textit{strong correlation} \end{cases}$$

Therefore, we only examine correlation coefficients greater than 0.5. Among all the network metrics, servers demonstrate moderate to strong correlation only on their RTT values. The reason being that the calculation of other network metrics such as connection throughput and available bandwidth are sensitive to object sizes which can vary from server to server. Therefore, we have only observed correlation for the RTT metric.

Table 7.1: Correlation Coefficients of RTT Actual Values

location	cluster name	web server 1	web server 2	correlation coefficient
Boston, MA	Boston Globe	www.boston.com	weather.boston.com	**0.97**
Atlanta, GA	Weather.com	www.weather.com	forgetaway.weather.com	0.78
	GA Government	www.georgia.gov	www.gov.state.ga.us	**0.80**
		www.georgia.gov	www.files.georgia.gov	0.67
		www.gov.state.ga.us	www.files.georgia.gov	0.68
Raymond, WA	MSN	www.msn.com	entertainment.msn.com	0.59
		www.msn.com	weather.msn.com	0.56
		entertainment.msn.com	weather.msn.com	0.76
		entertainment.msn.com	music.msn.com	0.62
		weather.msn.com	music.msn.com	**0.88**
	Real	www.realnetworks.com	brasil.real.com	**0.82**
Los Angeles, CA	City of LA	publiccsd.lacity.org	parc1.lacity.org	0.67
Dallas, TX	Dallas News	www.dallasnews.com	www.cowboysplus.com	**0.86**
		www.dallasnews.com	www.guidelive.com	**0.85**
		www.cowboysplus.com	www.guidelive.com	**0.95**

Note: Correlation coefficients greater than 0.8 are in bold font.

Table 7.1 shows all the web server RTT correlation coefficients with a value greater than 0.5, where values greater than 0.8 are marked with bold. As we can see, all the server pairs that have moderate to strong correlation belong to the same cluster, but not all clusters or all servers in a cluster are present in the table. That all correlated servers belong to the same cluster can be explained by the fact that these servers share most of their routes except the last one or two hops. The clusters that do not exhibit server correlation could be because data variation mostly dominates server performance instead of network condition. We can also see in Table 7.1, some server pairs can have higher correlation coefficient than others. We suspect the difference among server pair correlation be due to the two servers' host configuration or implementation and path difference on last hops.

7.2 Correlation of Prediction Errors of Network Metrics among Web Servers

In order to use data collected from one server to make prediction for another, we want to find out whether there is a correlation of data change between the two servers. In other words, if data change at server 1 is $\Delta 1$ (the difference between two measurements obtained at t and t-Δt at server 1) and data change at server 2 is $\Delta 2$ (the difference between two measurements obtained at t and t-Δt at server 2), we want to find out whether $\Delta 1$ is linearly correlated to $\Delta 2$. This idea is based on the hypothesis that performance change on the shared route of the two servers is the cause of the network metric change on both servers. In order to verify this

hypothesis, data variation correlation is studied. Particularly, we use prediction errors as an expression of data change.

When we discussed the choice of λ values in chapter 6, we used $\lambda = 0$, 0.25, 0.5 and 0.75, respectively. When $\lambda = 0$, the previous data point is used as the predicted value, so the prediction error (the difference between the actual values and the predicted values) in that case reflects change of the data set itself, i.e., the variation of two consecutive data points. When λ takes on other non-zero values, the only difference is that the prediction error is a variation measurement for the current data point over all the previous ones with a weight parameter set to decide the emphasis to be put on previous accesses. With this in mind, we examined the correlation coefficients of the prediction errors for every web server pair, to demonstrate if data change at one server results in similar change on the other. We calculated the correlation coefficients for all λ values ($\lambda = 0$, 0.25, 0.5 and 0.75), and the results were similar. Therefore, we only show the prediction error correlation coefficients with $\lambda = 0.75$ in Table 7.2, where all the values greater than 0.5 are shown and those greater than 0.8 are marked with bold.

As we can see, Table 7.2 has the same set of locations of server pairs that have moderate/strong correlation as Table 7.1 does. Not surprisingly, most of clusters and servers present in Table 7.1 are also present in Table 7.2. Table 7.2 contains more clusters and server pairs, however, than Table 7.1. However, in Table 7.2, many clusters in the second locations are not from the same locations as the first ones, and they two do not even seem to share much of the common route. For example, www.aviationdisasterlawyers.com is located in Boston, and all three servers that have moderate correlation to it on prediction errors are from three different locations

Table 7.2: Correlation Coefficients of RTT Prediction Errors at $\lambda = 0.75$

location 1	cluster 1	web server 1	location 2	cluster 2	web server 2	correlation coefficient
Boston, MA	Boston Globe	www.boston.com	-	-	weather.boston.com	**0.98**
		www.explorenewengland.com	San Francisco, CA	City of Davis	www.dcn.org	0.58
	Web Hosting	www.aviationdisasterlawyers.com	Raymond, WA	MSN	www.msn.com	0.53
			Los Angeles, CA	Ameriquest	careers.ameriquest.com	0.58
			San Francisco, CA	Sanfrancisco	www.legislature.ca.gov	0.70
		www.asbestoslaw.info	Raymond, WA	MSN	music.msn.com	0.72
					weather.msn.com	0.57
Atlanta, GA	CNN	www.cnn.com	-	Weather.com	br.weather.com	0.52
			Springfield, IL	IL Government	www.illinois.gov	0.62
			Dallas, TX	Online Video	www.lapdonline.org	0.58
	Weather.com	www.weather.com	-	-	forgetaway.weather.com	**0.80**
		br.weather.com	San Francisco, CA	City of Davis	www.dcn.org	0.51
			Dallas, TX	Online Video	www.lapdonline.org	0.58
Raymond, WA	MSN	www.msn.com	Los Angeles, CA	Ameriquest	careers.ameriquest.com	0.57
			-	-	entertainment.msn.com	0.57
		entertainment.msn.com	-	-	music.msn.com	0.51
					weather.msn.com	0.68
		music.msn.com	-	-	weather.msn.com	**0.91**
	Real	www.realnetworks.com	-	-	brasil.real.com	0.78
Los Angeles, CA	City of LA	publiccsd.lacity.org	-	-	parc1.lacity.org	0.62
Dallas, TX	Dallas News	www.dallasnews.com	-	-	www.cowboysplus.com	**0.85**
					www.guidelive.com	**0.85**
		www.cowboysplus.com	-	-	www.guidelive.com	**0.96**

Note: "-" means location 2 or cluster 2 is the same as location 1 or cluster 1. Correlation coefficients greater than 0.8 are in bold font.

(www.msn.com in Raymond, WA, careers.ameriquest.com in Los Angeles, CA, and www.legislature.ca.gov in San Francisco, CA) and have little shared route with www.boston.com. The reason why this is happening is unclear.

Overall, we can conclude that using performance change at other servers sharing some common route may not be an effective approach in predicting a given server's performance change. We suspect that the insufficiency is due to performance variation being largely caused by variations occurring on the non-shared routes or different server behavior. Different server behavior can include various user loads at the different servers, server configuration and implementation difference, and other known server-related reasons.

7.3 Summary

For topology-based predictions we examined the possibility of predicting the metrics of a given web server using data from another web server, by studying the correlation of data collected for different servers. Then we examined our hypothesis that network variation is mostly introduced in the shared paths, by calculating the correlation of prediction errors for different servers. In the following chapter we predict the metrics of one application using the data of another application.

Chapter 8

Prediction Across Applications

In the previous chapters, we have only studied one network application – web retrievals. In this chapter, network metrics prediction across different applications is examined. Besides web retrievals, two more applications are studied in predicting performance for each other. They are DNS requests and real video streaming.

8.1 Motivation

There are a variety of popular Internet applications, for example, web retrievals, streaming media application, gaming application, large FTP transfers, and DNS requests. Internet accesses to one application can be used to predict performance of another at the same cluster. To find out how well we can possibly utilize this kind of information is the motivation of studying cross-application metrics for performance prediction.

At the same time, we also know different applications can have varied require-

ments for different metrics. For example, a streaming application is more sensitive to jitter in the RTT and bandwidth than a file transfer application. Similarly a gaming application is more sensitive to the change of RTTs than that of bandwidth, since most of the packets are in small sizes. Data collected at a short HTTP connection may not be useful to make predictions of available bandwidth for a streaming application. All of these issues become a challenge for making predictions for an application using the measured data for another. Therefore, identifying the factors that significantly affect the performance for each individual network application is an aspect that should be taken into consideration as well.

Let us assume we already know what the QoS demands for each application are. The next task would be to know how confidently data obtained for a given metric from one application can be used for another. Can the available bandwidth inferred from the web access of a small page be used by a streaming media application to select a low or high bandwidth version?

8.2 DNS Requests vs. Web Retrievals

As we have mentioned in Chapter 5, in Tables 5.1 and 5.2, we collected traces of DNS requests to 20 DNS servers from 20 different clusters at 8 geographical locations. Since we can only infer RTT measurement from a DNS request, it is the only metric to be studied for cross application prediction between DNS requests and web retrievals.

8.2.1 Data Distribution

We put DNS request data together with that of web retrievals at the same cluster. First, we want to look at the distribution of the raw data for the two applications at each cluster. Most of the DNS requests have their RTTs' distributions coinciding with other web servers from the same cluster, as shown in Figures 8.1 through 8.4. In each figure, DNS data is represented in red. In some of the clusters, DNS RTT measurements lag behind those obtained from web servers. See Figures 8.5 through 8.6.

8.2.2 Correlation Coefficients

Next, we want to find out if there is any moderate to strong correlation between metrics collected from the DNS server and those from any of the web servers in the same cluster. If there is, using across application information for performance prediction is applicable to those servers, and vice versa. Here we use the same idea mentioned in Chapter 7, investigating the correlation between the two applications for both the actual values and their prediction errors. Table 8.1 shows all the DNS servers that have RTT correlation coefficients of 0.5 or greater with web servers at the same cluster. Values greater than 0.8 are marked with bold. Correlation coefficients for both the actual values and prediction errors at $\lambda = 0.75$ are shown in Table 8.1. As we can see, in only a few clusters, the DNS servers show moderate to strong correlation to some of the web servers from the same cluster. For these web servers, we can use RTTs collected from DNS requests to make predictions. We can see that for those servers whose actual RTT measurements have correlation

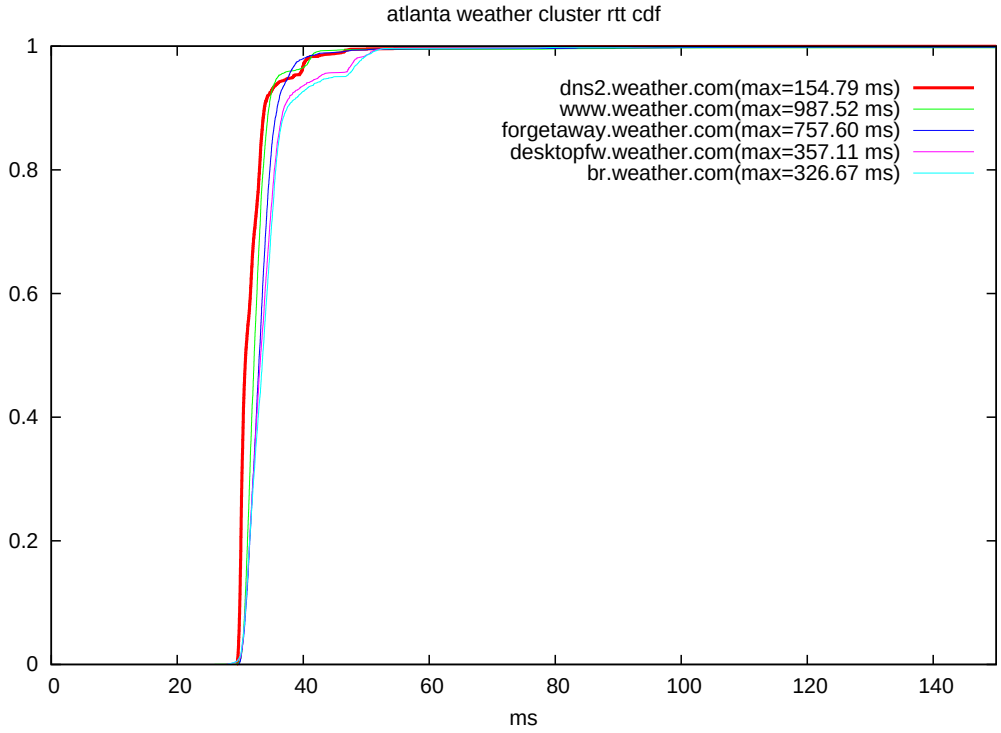

Figure 8.1: RTTs of DNS and Web Servers at Cluster Weather.com at Atlanta, GA

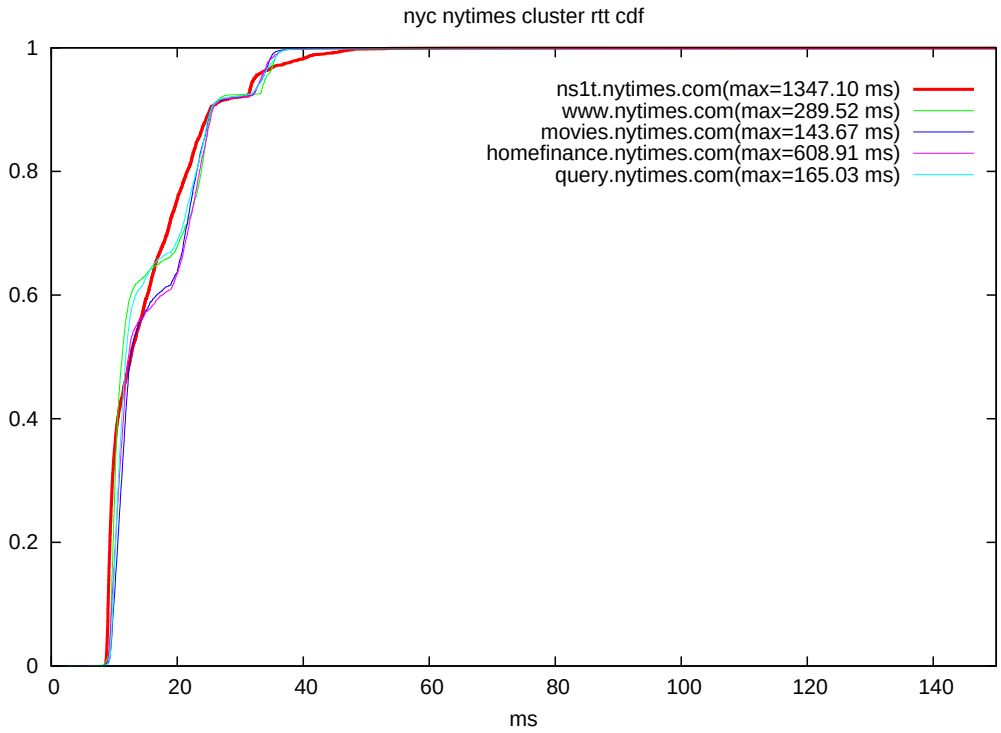

Figure 8.2: RTTs of DNS and Web Servers at Cluster NYTimes at New York, NY

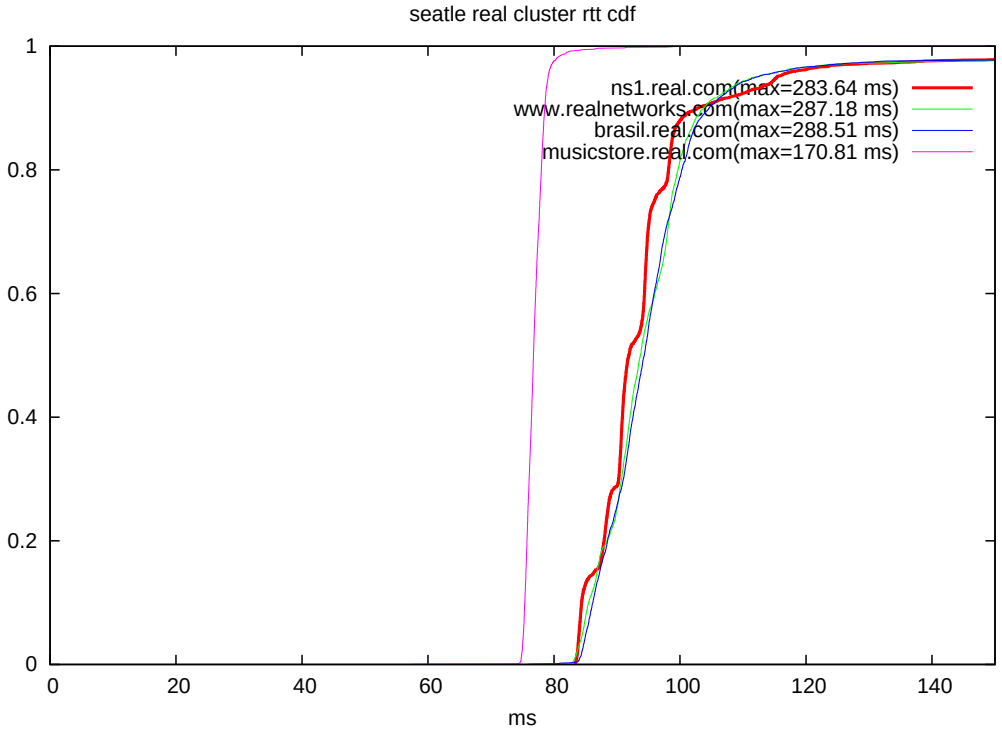

Figure 8.3: RTTs of DNS and Web Servers at Cluster Real at Raymond, WA

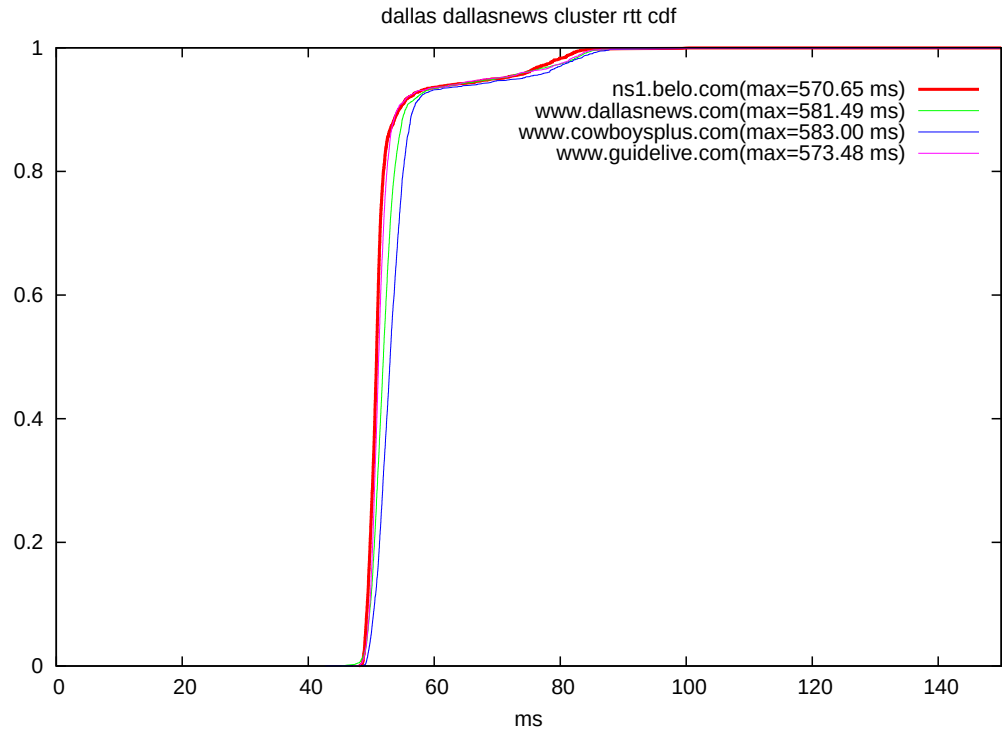

Figure 8.4: RTTs of DNS and Web Servers at Cluster Dallas News at Dallas, TX

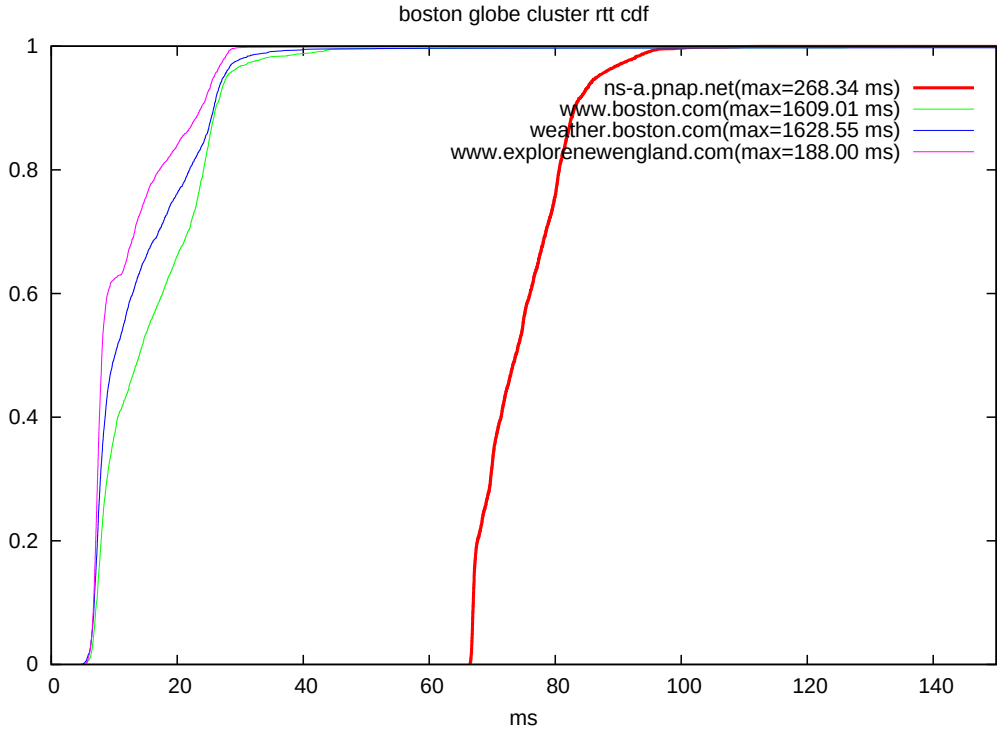

Figure 8.5: RTTs of DNS and Web Servers at Cluster Boston Globe at Boston, MA

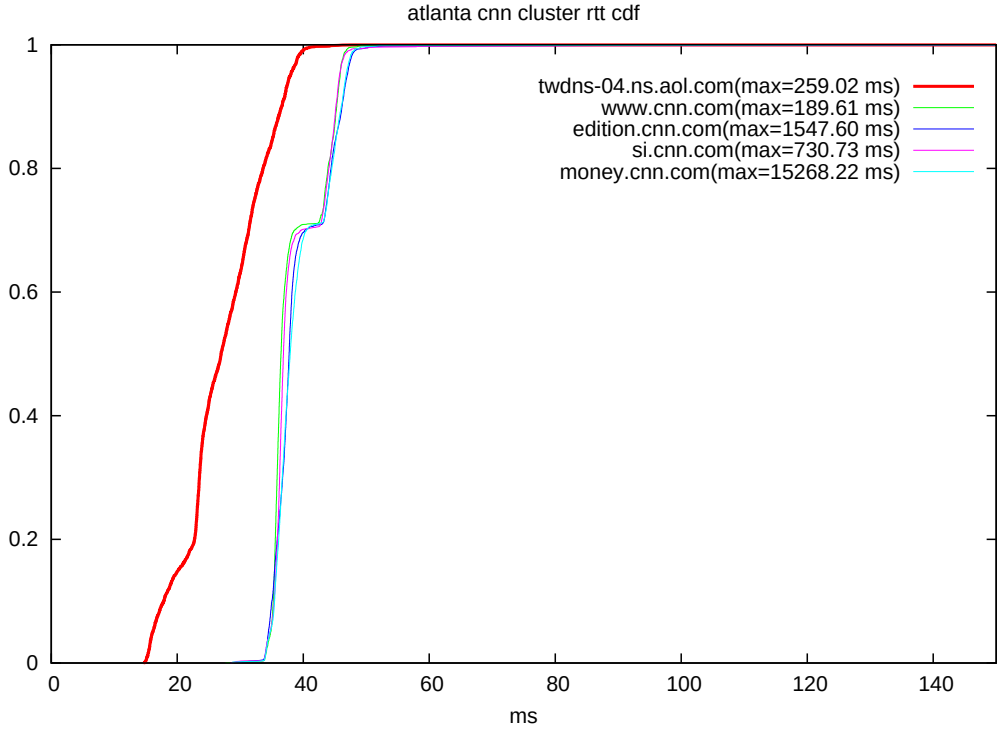

Figure 8.6: RTTs of DNS and Web Servers at Cluster CNN at Atlanta, GA

coefficients greater than 0.5 with web servers, coincide with their prediction errors as well.

Table 8.1: Correlation Coefficients of RTT Measurements from DNS Servers vs. Web Servers (Actual Values / Prediction Errors at $\lambda = 0.75$)

Location	Cluster Name	DNS Server	Web Servers	Correlation coefficients
Springfield, IL	IL Government	ns1.state.il.us	www.illinois.gov	0.62/0.60
	IL Education	ns1.illinois.net	www.isbe.net	0.86/0.61
Raymond, WA	Real	ns1.real.com	www.realnetworks.com	0.79/0.79
			brasil.real.com	**0.81**/0.79
Los Angeles, CA	City of LA	citylans1.lacity.org	publiccsd.lacity.org	**0.88/0.85**
			parc1.lacity.org	0.60/0.53
Dallas, TX	Dallas News	ns1.belo.com	www.dallasnews.com	**0.92/0.91**
			www.cowboysplus.com	**0.87/0.85**
			www.guidelive.com	**0.91/0.92**

Note: Correlation coefficients greater than 0.8 are in bold font.

8.3 Real Streaming Videos vs. DNS Requests and Web Retrievals

Besides the DNS application, another type of network application we study is the real streaming videos. The streaming servers used in our experiments are listed in Tables 5.1 and 5.2 in Chapter 5. Only ten clusters have real streaming servers, with one server in each, but they cover all the eight locations listed in the tables. We want to find out if metrics inferred from real streaming accesses can be used to predict performance of DNS and web applications located at the same cluster.

8.3.1 Metrics Inferred from Real Streaming

A typical real streaming access normally involves two flows: one TCP flow of RTSP messages between the the client and the server to set up the data transfer connection

and exchange streaming related parameters; one UDP flow for data transfer of real streaming videos. In the TCP flow, the client initiates packets and receives replies along with ACKs. Such a scenario enables us to infer an RTT measurement once we see the client receives an immediate ACK to the packet it just sent. In the UDP flow, we observe real streaming data is usually received by the client at an even rate. This rate is normally negotiated between the client and the server at the initial connection setup stage in the TCP flow. In the following discussion of Section 8.3, we call this receiving rate as throughput for the real streaming access, a reflection of how fast data is pushed through the connection. Therefore, we use both RTTs and throughput as the metrics studied for the streaming application.

8.3.2 Data Distribution

First, we want to look at the distribution of the raw data for web, DNS, and streaming applications. We find that all the streaming servers have their RTT measurements coincide with those obtained from the DNS and web servers at the same clusters. Figures 8.7 through 8.12 show the distributions of RTTs obtained from web, DNS and streaming applications at some clusters. In each figure, DNS data is represented in red, and streaming data in light green.

As we mentioned earlier in Subsection 8.3.1, the throughput we define for the streaming application is actually the receiving rate of the real streaming data. We also plot the distributions of the raw data of it measured from streaming and web applications. Most throughput measurements obtained from streaming applications have comparable distribution to those from web applications, as shown in Figures 8.13 through 8.16. Again, in each of the figures, streaming data is represented

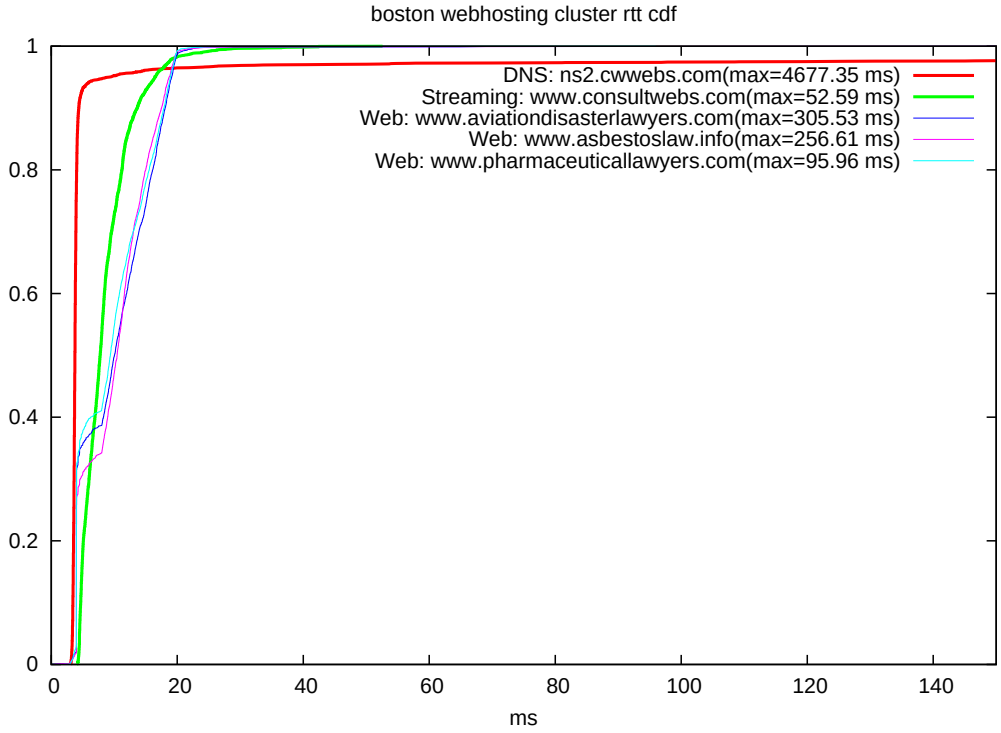

Figure 8.7: RTTs of DNS, Streaming and Web Servers at Cluster Web Hosting at Boston, MA

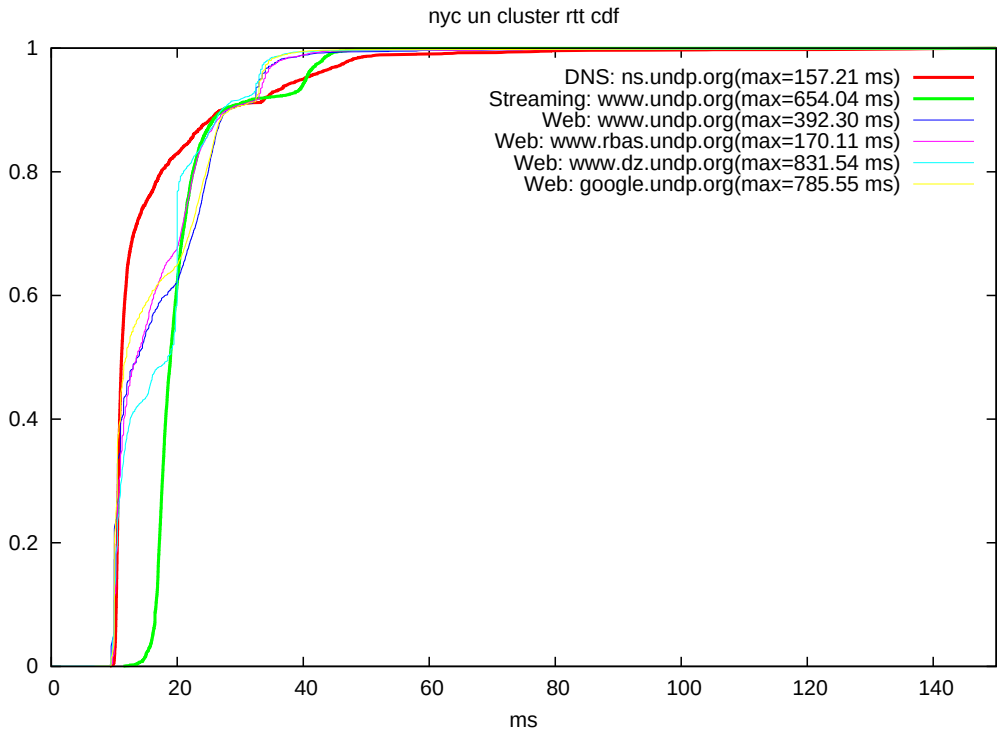

Figure 8.8: RTTs of DNS, Streaming and Web Servers at Cluster NYTimes at New York, NY

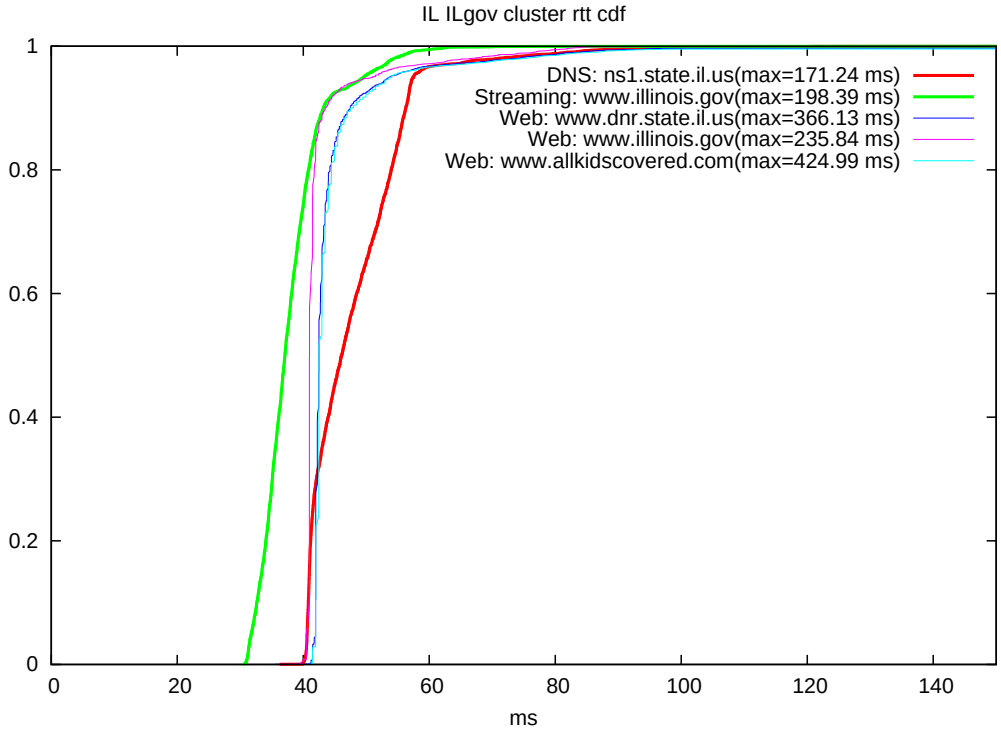

Figure 8.9: RTTs of DNS, Streaming and Web Servers at Cluster IL Government at Springfield, IL

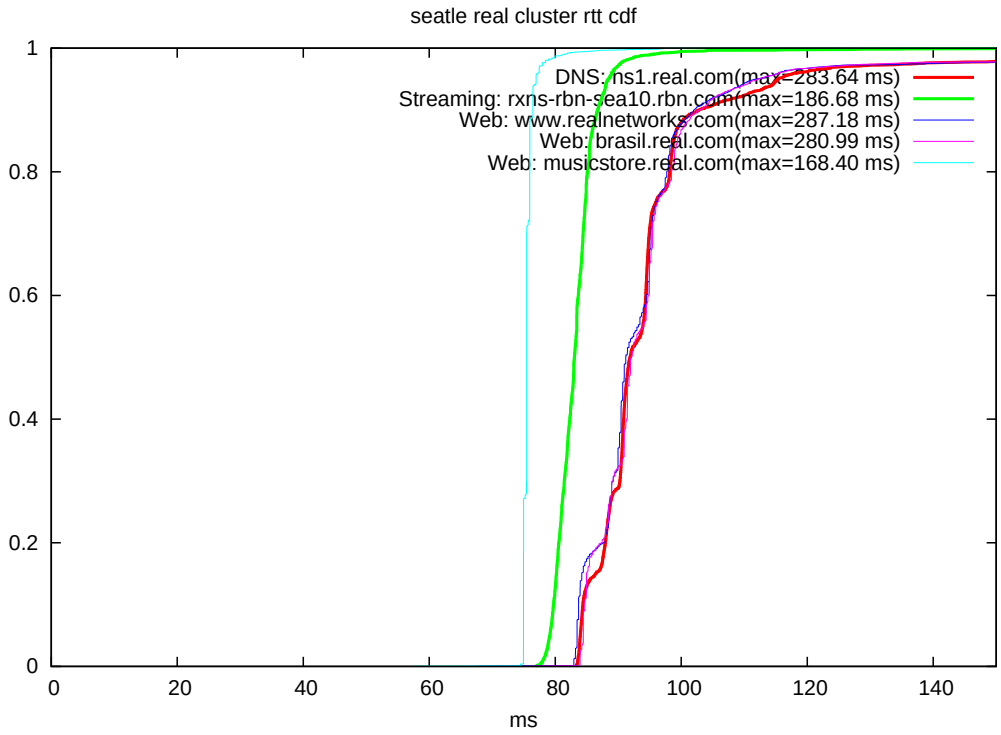

Figure 8.10: RTTs of DNS, Streaming and Web Servers at Cluster Real at Seattle, WA

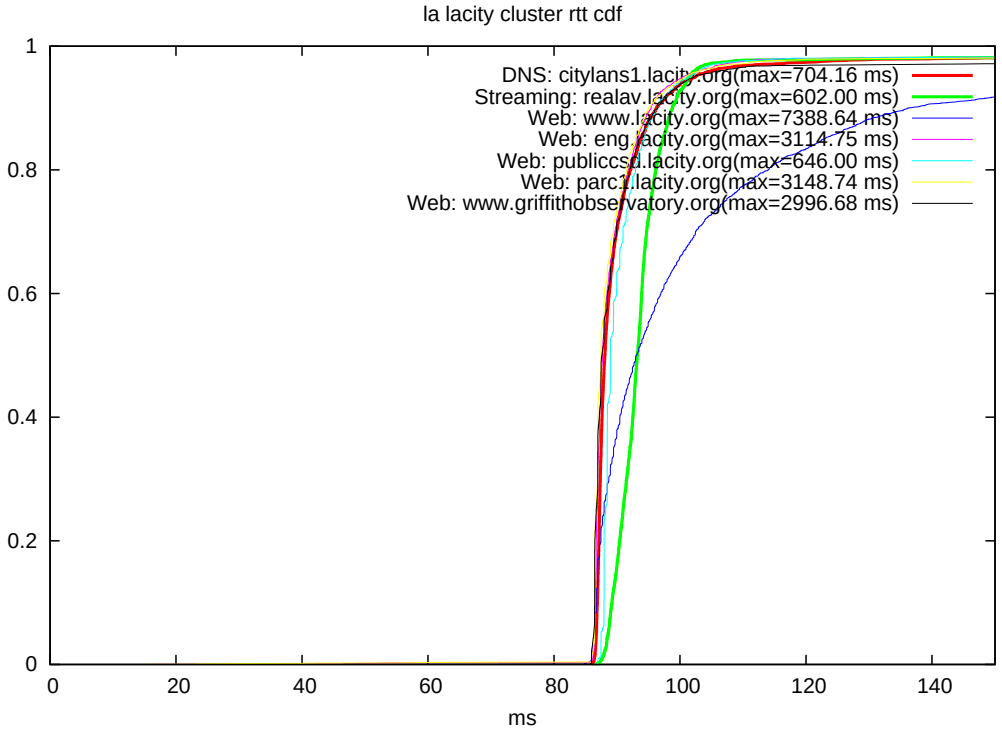

Figure 8.11: RTTs of DNS, Streaming and Web Servers at Cluster City of LA at Los Angeles, CA

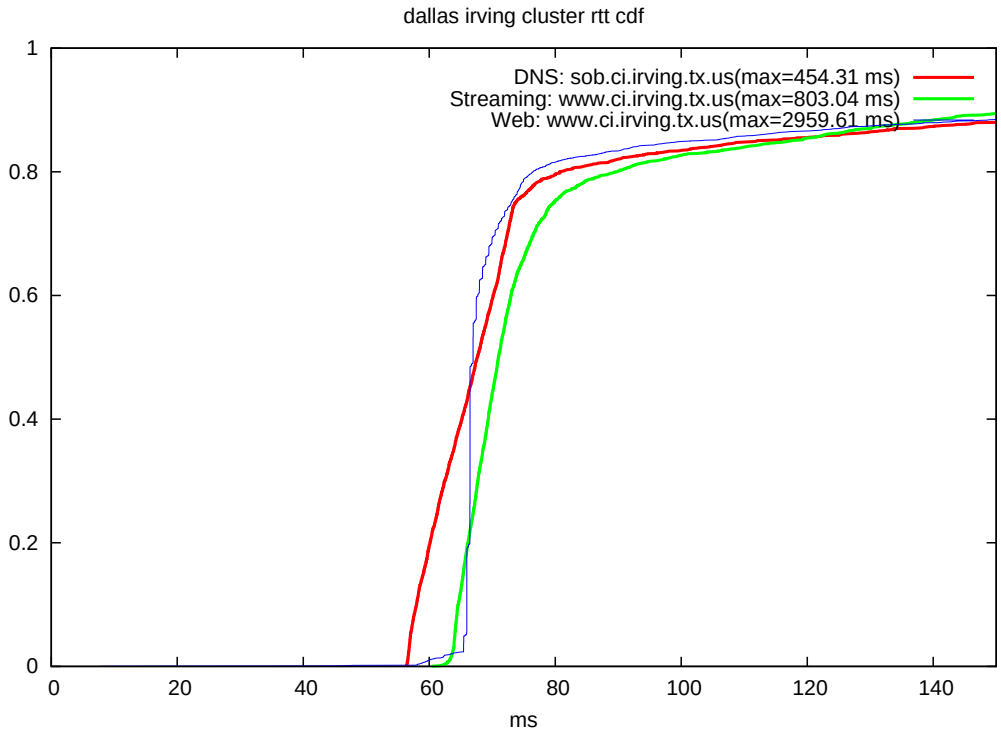

Figure 8.12: RTTs of DNS, Streaming and Web Servers at Cluster City of Irving at Dallas, TX

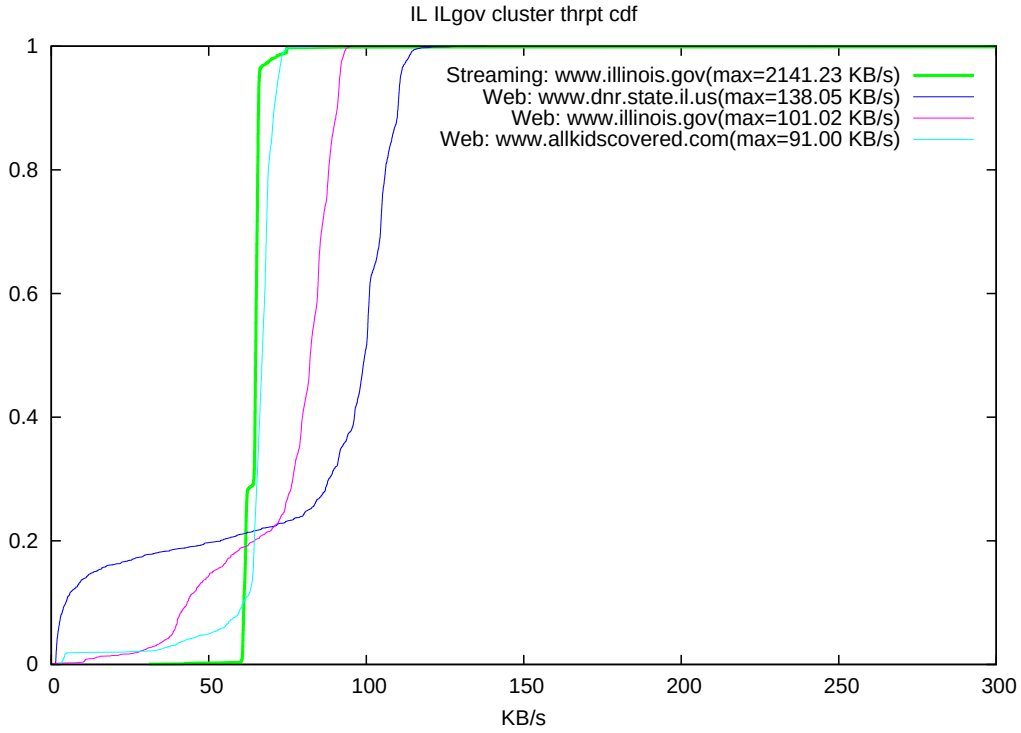

Figure 8.13: Throughput of DNS, Streaming and Web Servers at Cluster IL Government at Springfield, IL

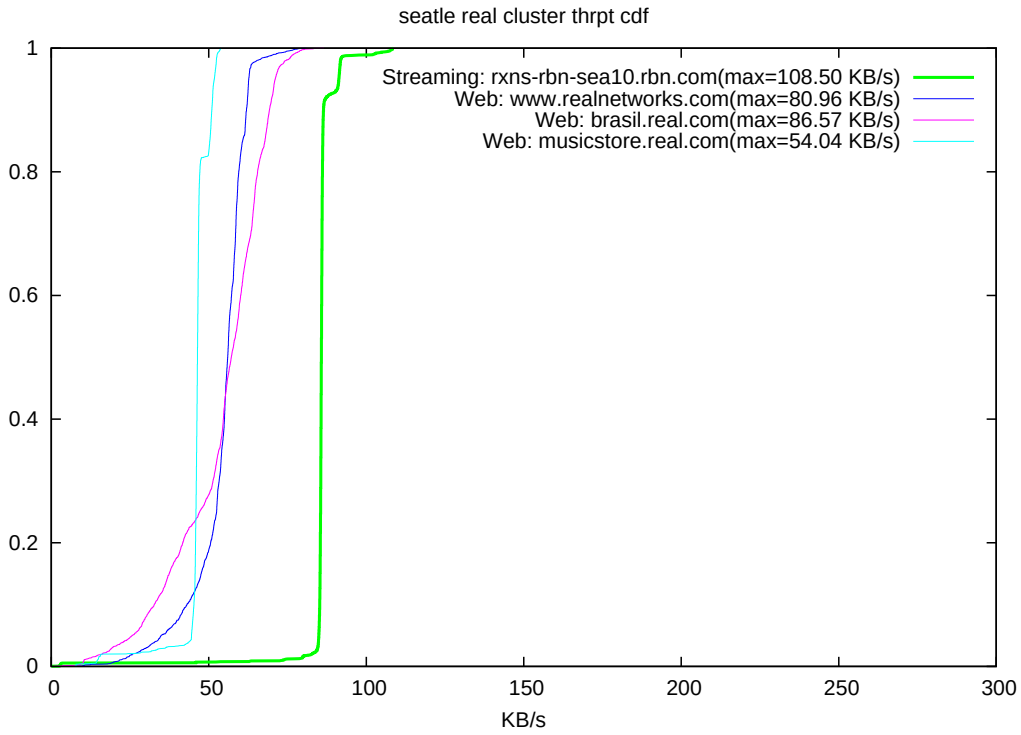

Figure 8.14: Throughput of DNS, Streaming and Web Servers at Cluster NYTimes at Seattle, WA

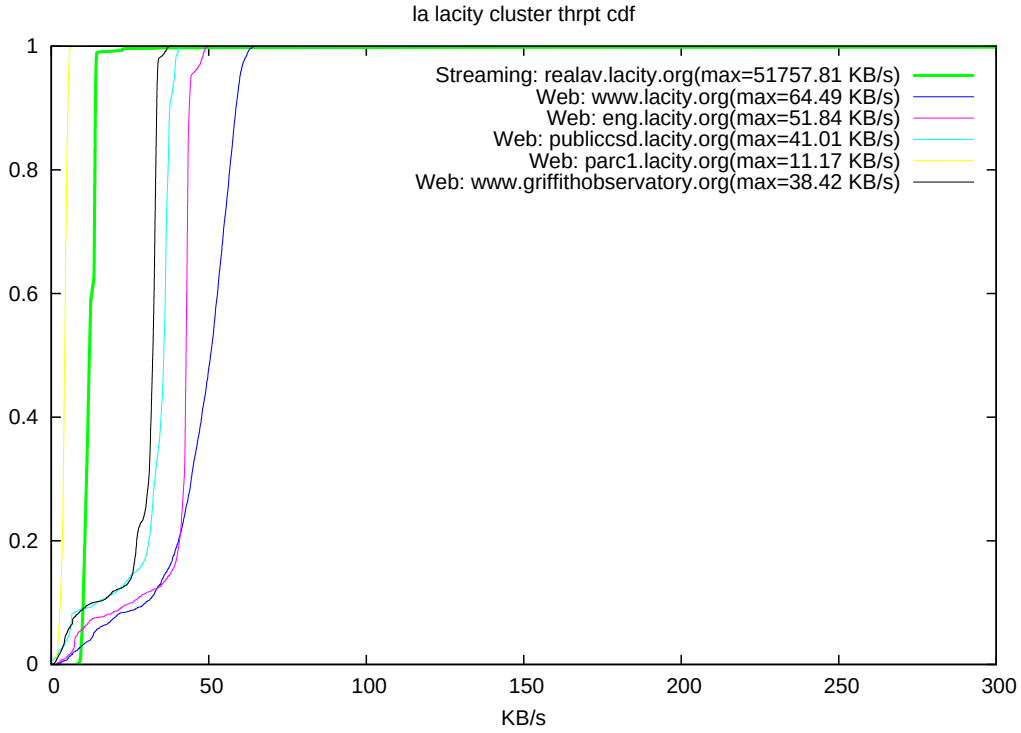

Figure 8.15: Throughput of DNS, Streaming and Web Servers at Cluster City of LA at Los Angeles, CA

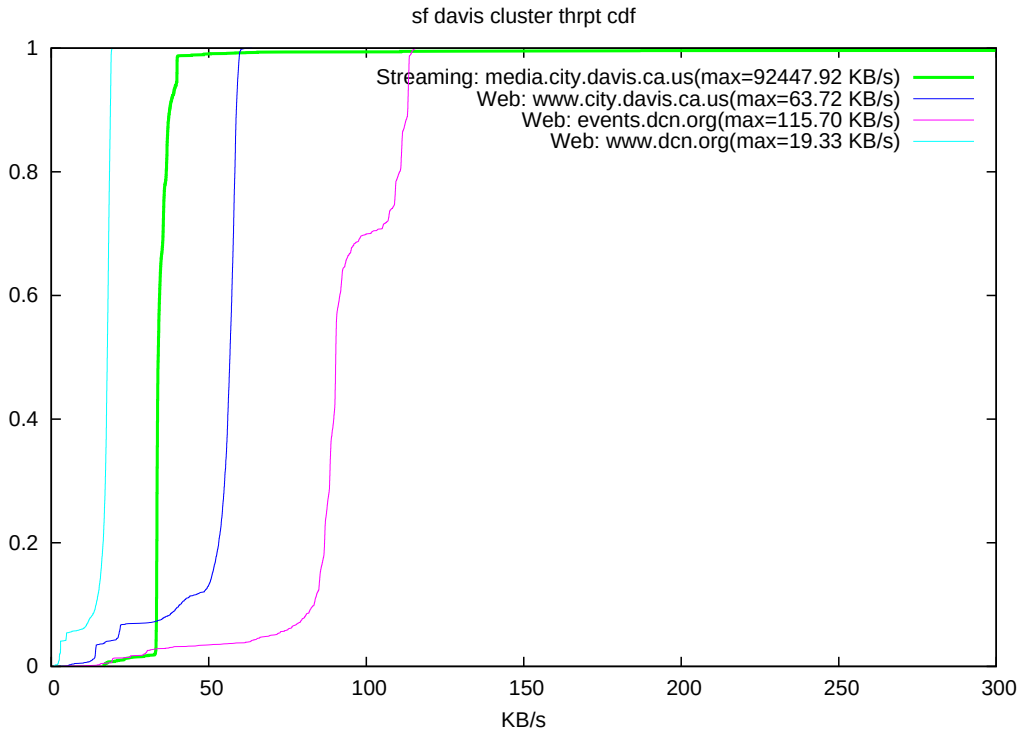

Figure 8.16: Throughput of DNS, Streaming and Web Servers at Cluster City of Davis at Davis, CA

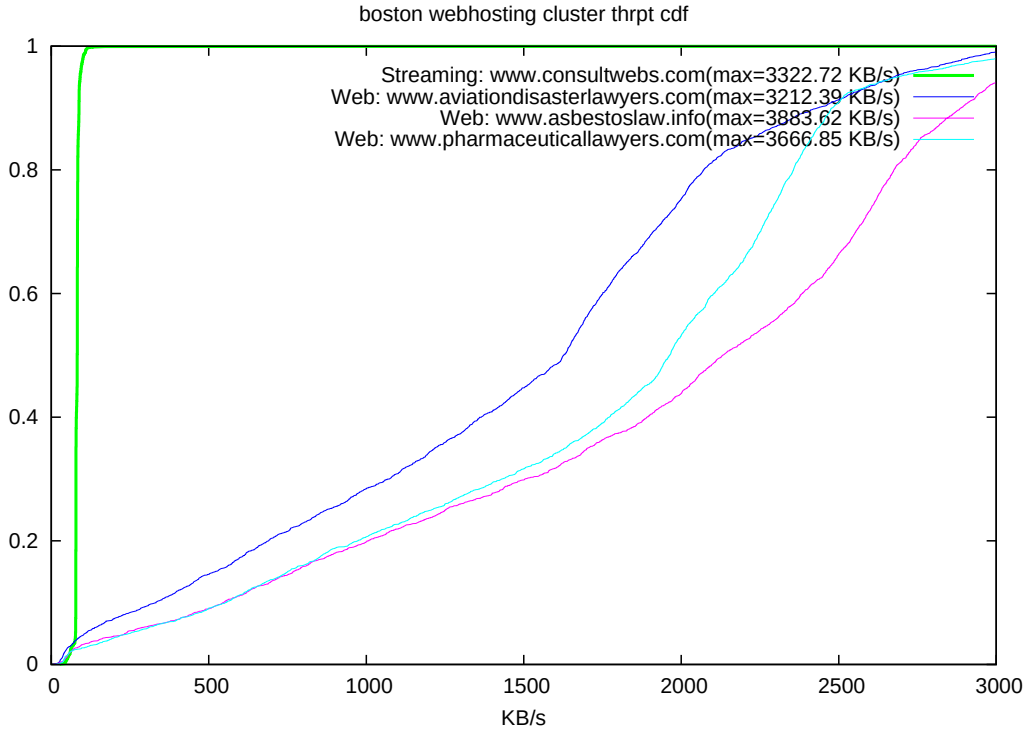

Figure 8.17: Throughput of DNS, Streaming and Web Servers at Cluster Boston Globe at Boston, MA

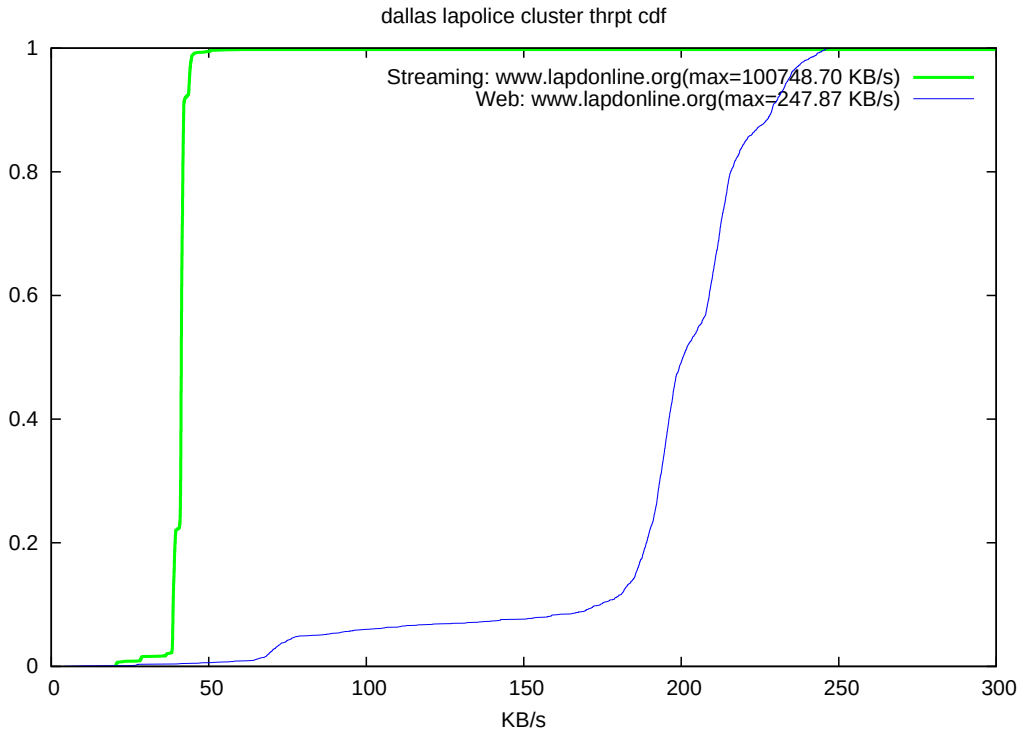

Figure 8.18: Throughput of DNS, Streaming and Web Servers at Cluster Online Video at Dallas, TX

in light green. However, we also observe that at some of the clusters, streaming throughput measurements are much smaller than those obtained from web servers. See Figures 8.17 through 8.18.

8.3.3 Correlation Coefficients

Second, we want to find out if there is any moderate to strong correlation between metrics collected from the streaming server and the DNS server, as well as between the streaming server and web servers in the same cluster. Again, we investigate the correlation between two applications for both the actual values and the prediction errors.

Table 8.2 shows all the streaming servers that have RTT correlation coefficients of 0.5 or greater with either a DNS server or a web server at the same cluster. Values greater than 0.8 are marked with bold. Correlation coefficients for both actual values and prediction errors at $\lambda = 0.75$ are shown in Table 8.2. As we can see, six out of ten streaming servers studied show up in the table. There are stronger correlations between the actual values than between the prediction errors. The streaming servers show moderate to strong correlation to some (not necessarily all) of the web servers at the same clusters, and so does it to the DNS servers. For these clusters and servers that show correlation, we use RTTs collected from one application to make prediction for another.

We also study the correlation for throughput between the streaming servers and any of the web servers at the same clusters. However, we do not find any moderate to strong correlation between any of the pairs of a streaming server and a web server at the same cluster. We believe that because the receiving rate for a streaming

Table 8.2: Correlation Coefficients of RTT Measurements from Streaming Servers vs. DNS and Web Servers (Actual Values / Prediction Errors at $\lambda = 0.75$)

Location	Cluster Name	Streaming Server	DNS/Web Servers	Correlation coefficients
New York, NY	UN	www.dz.undp.org	DNS: ns.undp.org Web: google.undp.org	0.61/- **0.84**/**0.89**
Springfield, IL	GA Government	www.georgia.gov	Web: www.georgia.gov Web: www.files.georgia.gov Web: www.gov.state.ga.us	**0.83**/- 0.67/- **0.83**/0.55
	IL Education	www.isbe.net	DNS: ns1.illinois.net Web: www.isbe.net	**0.85**/0.63 **0.83**/0.75
Los Angeles, CA	City of LA	realav.lacity.org	DNS: citylans1.lacity.org Web: publiccsd.lacity.org Web: parc1.lacity.org	**0.80**/0.77 **0.85**/**0.82** 0.58/-
Dallas, TX	Dallas News	www.ci.irving.tx.us	DNS: sob.ci.irving.tx.us	0.73/0.68
	Online Video	www.lapdonline.org	Web: www.lapdonline.org	0.71/0.64

Note: Correlation coefficients greater than 0.8 are in bold font.
A value less than 0.5 is shown as "-".

application is negotiated between the client and the server at the initial connection setup, it may not be good reflection of the throughput of a web access. We also investigate such correlation between receiving rates from streaming applications and measured bandwidth from web retrievals at the same cluster. Again, for the same reason, we do not see any moderate to strong correlation between the two metrics collected from the two applications.

8.4 Summary

In this section we examined the possibility of using network metrics from one application for prediction of another application. The different applications studied here were web transfers, DNS lookups and Real streaming videos. We compared the measured RTTs across these applications by studying the correlation coefficients. We saw good correlation of RTT between some web retrievals and DNS requests. We studied the working of the real time streaming protocol (RTSP). We defined

our method to measure RTTs and receiving rates for real streaming videos. We observed good correlation for RTT measurements between streaming servers and web servers for some locations. The same behavior was observed for streaming servers and DNS servers. However we did not observe good correlation for receiving rate measurements between real streaming servers and web servers.

Chapter 9

Conclusions

Looking back at the original questions we raised in Chapter 1 "Introduction", we want to summarize our findings in the form of the answers to those questions.

The first set of questions were "How well can previous measurements be used for predicting future measurements?" and "How frequently do these measurements need to be taken in order to make accurate predictions?" In our study of time-based predictions we can conclude that the exponential predictor provides good accuracy in using previous data points to predict future network accesses. This is also due to the stability of the data set itself, because our experiments were run over a good network connection from WPI. We also find in most cases the value of the parameter to the exponential predictor (λ) is not significant. Using λ=0.75 generally provides good results. In exploring different collection frequencies, we find that the choice of collection time intervals is not critical. The data remains relatively stable when we experiment with different time intervals. However large intervals do cause significant prediction errors for some servers.

The second question was "How can topological similarities between network

paths to servers be used in making predictions?". We studied topology-based predictions, i.e. predicting performance among servers that share certain common paths from WPI. At some clusters, we see prediction results show good correlation among some web servers in the same cluster. However not all servers in the same cluster show correlation, and not all the clusters have servers whose prediction results correlation. Therefore, using topology commonality in server performance prediction has to be analyzed on a case by case basis.

The third question was "How effectively can information inferred from one network application be used to predict application performance of another?". We look at the feasibility of making predictions across applications e.g. using RTT measured from DNS to make predictions for web servers. The findings are similar here to the topology-based prediction. We can see that correlation between DNS requests and web servers is observed in a few clusters but not all. Again, using across application performance data for prediction should be evaluated depending on the individual servers involved. We look at real streaming videos in this context as well. We can see that for RTT measurements there is some correlation between streaming videos and web retrievals. We also see that there is RTT correlation between streaming videos and DNS lookups. However when we do the same for throughput measurements, we see no correlation between real streaming servers and web servers.

Chapter 10

Future Work

As an extension of this work we would like to analyze additional kinds of applications and the correlation of different metrics. For example we could study real streaming media and study correlation of its bandwidth with web retrievals. As a result, we will have more datasets to analyze feasibility of prediction.

In our current work, we studied time-based, topology-based and cross-application prediction mechanism. However we studied them in isolation. In the future we intend to build a more comprehensive model which will integrate these three techniques. One way would be to build a local database at a cluster storing metrics from all user accesses. This database can then be accessed by different servers running different applications. If this approach shows some success then a mechanism to share information among different clusters through their databases can be built as well.

The experiments in this work were all run from the network at Worcester Polytechnic Institute. Hence all our design decisions were influenced by the well connected OC1 link. As a variation the experiments can be done on networks with

different bandwidths e.g. a DSL/cable home link. It would be interesting to study the behavior of network metrics over the range of parameters on such relatively slower networks. Having data collected from different networks will provide us a better insight into the behavior of the various parameters used in our work. It may also help to generalize the design choices such as with a higher degree of confidence For example the choice of λ, measurement intervals etc. could differ on a DSL/cable link.

Appendix A

Time-Based Prediction Errors

Table A.1: RTT Prediction Results with Different λ Values

location	cluster name	web servers	average actual values (ms)	λ=0 mean prediction errors (ms)	λ=0.25 mean prediction errors (ms)	λ=0.5 mean prediction errors (ms)	λ=0.75 mean prediction errors (ms)
Boston, MA	Boston Globe	1: www.boston.com	17.56	7.92 (45.1%)	7.67 (43.7%)	7.76 (44.19%)	8.00 (45.58%)
		2: weather.boston.com	15.48	6.60 (42.64%)	6.48 (41.88%)	6.69 (43.25%)	7.16 (46.25%)
		3: www.explorenewengland.com	11.86	5.16 (43.48%)	4.85 (40.86%)	4.70 (39.65%)	4.65 (39.2%)
	MBTA	4: www.mbta.com	36.74	38.96 (106.05%)	37.28 (101.47%)	36.12 (98.33%)	35.01 (95.29%)
		5: trip.mbta.com	27.85	19.55 (70.2%)	18.06 (64.86%)	17.22 (61.84%)	16.76 (60.2%)
	Web Hosting	6: www.aviationdisasterlawyers.com	10.42	6.44 (61.83%)	5.93 (56.87%)	5.56 (53.4%)	5.27 (50.61%)
		7: www.asbestoslaw.info	10.36	5.89 (56.82%)	5.33 (51.45%)	4.95 (47.81%)	4.67 (45.12%)
		8: www.pharmaceuticallawyers.com	9.85	5.89 (59.79%)	5.41 (54.88%)	5.07 (51.46%)	4.81 (48.87%)
New York, NY	NYTimes	9: www.nytimes.com	16.09	6.14 (38.18%)	5.76 (35.81%)	5.56 (34.59%)	5.42 (33.69%)
		10: movies.nytimes.com	16.74	5.81 (34.73%)	5.34 (31.9%)	5.07 (30.28%)	4.88 (29.16%)
		11: homefinance.nytimes.com	17.10	6.84 (40.01%)	6.37 (37.28%)	6.14 (35.91%)	6.00 (35.12%)
		12: query.nytimes.com	16.07	5.64 (35.1%)	5.21 (32.38%)	4.97 (30.93%)	4.82 (29.96%)
	UN	13: www.undp.org	18.27	6.90 (37.74%)	6.32 (34.56%)	5.98 (32.73%)	5.71 (31.27%)
		14: www.rbas.undp.org	17.69	5.86 (33.15%)	5.37 (30.35%)	5.05 (28.53%)	4.82 (27.22%)
		15: www.dz.undp.org	18.31	6.53 (35.67%)	6.10 (33.31%)	5.84 (31.88%)	5.65 (30.84%)
		16: google.undp.org	17.76	6.86 (38.64%)	6.35 (35.74%)	6.08 (34.23%)	5.91 (33.3%)
Atlanta, GA	CNN	17: www.cnn.com	38.65	1.19 (3.08%)	1.09 (2.83%)	1.04 (2.69%)	1.01 (2.61%)
		18: edition.cnn.com	40.41	3.79 (9.37%)	3.63 (8.98%)	3.53 (8.73%)	3.45 (8.55%)
		19: si.cnn.com	39.36	2.33 (5.92%)	2.23 (5.67%)	2.18 (5.54%)	2.14 (5.43%)
		20: money.cnn.com	44.24	11.15 (25.2%)	10.99 (24.83%)	10.90 (24.64%)	10.83 (24.49%)
	Weather.com	21: www.weather.com	33.45	2.80 (8.38%)	2.72 (8.12%)	2.68 (8.02%)	2.66 (7.95%)
		22: forgetaway.weather.com	33.90	2.84 (8.38%)	2.68 (7.9%)	2.59 (7.63%)	2.53 (7.45%)
		23: desktopfw.weather.com	34.28	3.03 (8.84%)	2.82 (8.21%)	2.68 (7.83%)	2.60 (7.58%)
		24: br.weather.com	34.62	3.13 (9.05%)	2.90 (8.37%)	2.77 (8.01%)	2.73 (7.88%)
	GA gov	25: www.georgia.gov	38.57	3.90 (10.1%)	3.56 (9.24%)	3.39 (8.79%)	3.35 (8.69%)
		26: www.files.georgia.gov	37.61	3.01 (8.01%)	2.85 (7.58%)	2.77 (7.38%)	2.81 (7.48%)
		27: oca.awe.gta.ga.gov	40.22	9.51 (23.64%)	9.37 (23.31%)	9.33 (23.2%)	9.38 (23.33%)
		28: www.gov.state.ga.us	37.08	2.96 (7.97%)	2.75 (7.41%)	2.66 (7.18%)	2.63 (7.1%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	46.41	4.29 (9.25%)	4.06 (8.74%)	3.95 (8.51%)	3.97 (8.56%)
		30: www.illinois.gov	43.39	2.28 (5.26%)	2.14 (4.93%)	2.11 (4.87%)	2.23 (5.14%)
		31: www.allkidscovered.com	47.64	5.73 (12.02%)	5.44 (11.41%)	5.32 (11.16%)	5.34 (11.21%)
	IL Education	32: www.isbe.net	42.12	4.33 (10.29%)	4.21 (10%)	4.29 (10.19%)	4.85 (11.51%)

Table A.2: RTT Prediction Results with Different λ Values (Continued)

location	cluster name	web servers	average actual values (ms)	λ=0 mean prediction errors (ms)	λ=0.25 mean prediction errors (ms)	λ=0.5 mean prediction errors (ms)	λ=0.75 mean prediction errors (ms)
Raymond, WA	MSN	33: www.msn.com	93.43	2.49 (2.66%)	2.28 (2.45%)	2.19 (2.34%)	2.22 (2.38%)
		34: entertainment.msn.com	92.10	1.63 (1.77%)	1.50 (1.63%)	1.45 (1.58%)	1.46 (1.58%)
		35: music.msn.com	92.04	1.59 (1.73%)	1.48 (1.61%)	1.43 (1.55%)	1.44 (1.57%)
		36: weather.msn.com	92.48	2.32 (2.51%)	2.13 (2.3%)	2.03 (2.2%)	2.01 (2.17%)
	Real	37: www.realnetworks.com	97.85	7.21 (7.37%)	6.79 (6.94%)	6.78 (6.93%)	7.05 (7.2%)
		38: brasil.real.com	98.08	7.43 (7.58%)	7.01 (7.14%)	6.94 (7.07%)	7.12 (7.26%)
		39: musicstore.real.com	76.98	1.52 (1.98%)	1.38 (1.8%)	1.29 (1.67%)	1.22 (1.59%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	89.04	1.40 (1.58%)	1.26 (1.42%)	1.16 (1.3%)	1.07 (1.2%)
		41: careers.ameriquest.com	89.02	2.01 (2.26%)	1.82 (2.04%)	1.68 (1.88%)	1.56 (1.75%)
		42: www.ameriquestracing.com	92.11	5.59 (6.07%)	5.01 (5.44%)	4.62 (5.02%)	4.33 (4.7%)
	City of LA	43: www.lacity.org	115.76	32.34 (27.94%)	30.16 (26.06%)	28.83 (24.91%)	28.10 (24.28%)
		44: eng.lacity.org	98.29	12.26 (12.48%)	12.18 (12.39%)	12.51 (12.73%)	13.20 (13.43%)
		45: publiccsd.lacity.org	98.26	9.37 (9.53%)	9.14 (9.31%)	9.39 (9.56%)	10.32 (10.5%)
		46: parc1.lacity.org	97.47	10.74 (11.02%)	10.58 (10.86%)	10.85 (11.13%)	11.63 (11.93%)
		47: www.griffithobservatory.org	109.91	32.92 (29.95%)	32.41 (29.48%)	32.20 (29.29%)	32.68 (29.73%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	85.39	3.31 (3.88%)	3.19 (3.74%)	3.10 (3.63%)	3.01 (3.53%)
		49: www.santa-clara.com	84.60	1.82 (2.16%)	1.66 (1.96%)	1.55 (1.83%)	1.48 (1.75%)
		50: www.santacruz.com	84.96	5.53 (6.51%)	5.02 (5.91%)	4.63 (5.45%)	4.26 (5.02%)
		51: www.oakland.com	83.56	3.39 (4.06%)	3.11 (3.72%)	2.90 (3.47%)	2.70 (3.23%)
	CA gov	52: democrats.assembly.ca.gov	87.78	5.41 (6.17%)	5.15 (5.87%)	4.99 (5.68%)	4.88 (5.56%)
		53: www.legislature.ca.gov	87.20	2.84 (3.25%)	2.59 (2.97%)	2.42 (2.78%)	2.31 (2.64%)
		54: republican.assembly.ca.gov	86.96	6.33 (7.28%)	6.07 (6.98%)	5.90 (6.79%)	5.76 (6.63%)
	City of Davis	55: www.city.davis.ca.us	103.86	10.46 (10.07%)	10.01 (9.64%)	9.71 (9.35%)	9.33 (8.98%)
		56: events.dcn.org	94.20	2.39 (2.54%)	2.16 (2.3%)	2.01 (2.13%)	1.90 (2.01%)
		57: www.dcn.org	95.51	3.18 (3.33%)	2.86 (2.99%)	2.64 (2.76%)	2.49 (2.61%)
Dallas, TX	Dallas News	58: www.dallasnews.com	53.72	3.22 (5.99%)	3.11 (5.8%)	3.14 (5.84%)	3.22 (6%)
		59: www.cowboysplus.com	54.90	3.58 (6.53%)	3.47 (6.32%)	3.51 (6.4%)	3.61 (6.57%)
		60: www.guidelive.com	53.12	2.76 (5.19%)	2.75 (5.17%)	2.84 (5.34%)	2.97 (5.59%)
	City of Irving	61: www.ci.irving.tx.us	90.70	42.79 (47.17%)	41.32 (45.55%)	40.85 (45.04%)	40.79 (44.97%)
	Online Video	62: www.lapdonline.org	54.00	2.15 (3.98%)	2.03 (3.75%)	1.95 (3.61%)	1.92 (3.56%)

Table A.3: Connection Throughput Prediction Results with Different λ Values

location	cluster name	web servers	average actual values (KB/s)	λ=0 mean prediction errors (KB/s)	λ=0.25 mean prediction errors (KB/s)	λ=0.5 mean prediction errors (KB/s)	λ=0.75 mean prediction errors (KB/s)
Boston, MA	Boston Globe	1: www.boston.com	410.03	210.96 (51.45%)	196.40 (47.9%)	190.08 (46.36%)	187.70 (45.78%)
		2: weather.boston.com	354.59	59.43 (16.76%)	53.96 (15.22%)	50.39 (14.21%)	49.02 (13.82%)
		3: www.explorenewengland.com	261.94	35.59 (13.59%)	32.41 (12.37%)	30.47 (11.63%)	29.57 (11.29%)
	MBTA	4: www.mbta.com	153.09	53.45 (34.91%)	48.68 (31.8%)	46.33 (30.27%)	46.01 (30.06%)
		5: trip.mbta.com	236.20	85.28 (36.1%)	76.94 (32.57%)	71.79 (30.39%)	69.29 (29.34%)
	Web Hosting	6: www.aviationdisasterlawyers.com	1466.44	747.67 (50.99%)	671.83 (45.81%)	617.69 (42.12%)	581.14 (39.63%)
		7: www.asbestoslaw.info	1926.40	829.69 (43.07%)	752.82 (39.08%)	699.25 (36.3%)	668.33 (34.69%)
		8: www.pharmaceuticallawyers.com	1738.13	700.67 (40.31%)	636.84 (36.64%)	594.55 (34.21%)	566.82 (32.61%)
New York, NY	NYTimes	9: www.nytimes.com	826.75	130.65 (15.8%)	118.67 (14.35%)	110.23 (13.33%)	103.86 (12.56%)
		10: movies.nytimes.com	765.26	145.28 (18.98%)	130.44 (17.04%)	119.40 (15.6%)	110.70 (14.47%)
		11: homefinance.nytimes.com	310.01	35.54 (11.46%)	32.28 (10.41%)	30.18 (9.73%)	28.58 (9.22%)
		12: query.nytimes.com	69.27	21.63 (31.23%)	20.30 (29.3%)	19.30 (27.86%)	18.08 (26.11%)
	UN	13: www.undp.org	295.06	51.69 (17.52%)	46.96 (15.91%)	43.36 (14.7%)	40.33 (13.67%)
		14: www.rbas.undp.org	364.88	57.44 (15.74%)	51.73 (14.18%)	47.41 (12.99%)	43.89 (12.03%)
		15: www.dz.undp.org	233.30	62.44 (26.76%)	58.75 (25.18%)	57.09 (24.47%)	57.95 (24.84%)
		16: google.undp.org	28.54	9.66 (33.86%)	8.57 (30.04%)	7.77 (27.24%)	7.16 (25.09%)
Atlanta, GA	CNN	17: www.cnn.com	305.56	22.97 (7.52%)	21.34 (6.98%)	20.37 (6.67%)	19.54 (6.39%)
		18: edition.cnn.com	259.17	22.16 (8.55%)	20.20 (7.79%)	19.00 (7.33%)	18.11 (6.99%)
		19: si.cnn.com	214.58	54.70 (25.49%)	50.52 (23.54%)	47.68 (22.22%)	45.16 (21.05%)
		20: money.cnn.com	230.45	21.71 (9.42%)	19.73 (8.56%)	18.35 (7.96%)	17.27 (7.5%)
	Weather.com	21: www.weather.com	457.36	60.44 (13.21%)	56.17 (12.28%)	53.90 (11.78%)	52.29 (11.43%)
		22: forgetaway.weather.com	267.24	28.34 (10.6%)	26.17 (9.79%)	24.80 (9.28%)	23.65 (8.85%)
		23: desktopfw.weather.com	246.97	34.26 (13.87%)	31.28 (12.67%)	29.41 (11.91%)	28.06 (11.36%)
		24: br.weather.com	156.74	13.80 (8.81%)	12.86 (8.2%)	12.20 (7.79%)	11.63 (7.42%)
	GA gov	25: www.georgia.gov	67.05	32.76 (48.86%)	31.16 (46.47%)	30.37 (45.29%)	29.82 (44.47%)
		26: www.files.georgia.gov	132.81	49.42 (37.21%)	47.25 (35.58%)	45.85 (34.52%)	44.79 (33.73%)
		27: oca.awe.gta.ga.gov	180.52	15.61 (8.65%)	14.49 (8.03%)	13.73 (7.61%)	13.12 (7.27%)
		28: www.gov.state.ga.us	280.68	58.61 (20.88%)	54.62 (19.46%)	52.22 (18.61%)	50.96 (18.16%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	81.86	31.25 (38.18%)	29.30 (35.8%)	28.30 (34.57%)	27.62 (33.74%)
		30: www.illinois.gov	71.90	14.84 (20.63%)	13.64 (18.97%)	12.94 (18%)	12.55 (17.46%)
		31: www.allkidscovered.com	64.90	5.62 (8.66%)	5.25 (8.09%)	5.01 (7.73%)	4.83 (7.44%)
	IL Education	32: www.isbe.net	155.14	7.87 (5.08%)	7.49 (4.83%)	7.35 (4.74%)	7.57 (4.88%)

Table A.4: Connection Throughput Prediction Results with Different λ Values (Continued)

location	cluster name	web servers	average actual values (KB/s)	λ=0 mean prediction errors (KB/s)	λ=0.25 mean prediction errors (KB/s)	λ=0.5 mean prediction errors (KB/s)	λ=0.75 mean prediction errors (KB/s)
Raymond, WA	MSN	33: www.msn.com	50.97	17.34 (34.02%)	16.20 (31.79%)	15.60 (30.61%)	15.32 (30.06%)
		34: entertainment.msn.com	51.96	5.40 (10.38%)	4.90 (9.44%)	4.53 (8.72%)	4.17 (8.03%)
		35: music.msn.com	64.52	5.19 (8.04%)	4.77 (7.39%)	4.46 (6.92%)	4.17 (6.46%)
		36: weather.msn.com	53.40	3.60 (6.74%)	3.47 (6.5%)	3.46 (6.49%)	3.51 (6.58%)
	Real	37: www.realnetworks.com	54.13	6.27 (11.58%)	5.71 (10.55%)	5.33 (9.85%)	5.05 (9.33%)
		38: brasil.real.com	54.13	11.85 (21.9%)	10.77 (19.9%)	10.02 (18.5%)	9.50 (17.55%)
		39: musicstore.real.com	46.25	2.64 (5.71%)	2.48 (5.36%)	2.36 (5.11%)	2.24 (4.84%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	31.79	6.94 (21.84%)	6.30 (19.83%)	5.87 (18.47%)	5.51 (17.34%)
		41: careers.ameriquest.com	24.21	6.51 (26.88%)	6.01 (24.82%)	5.70 (23.55%)	5.44 (22.48%)
		42: www.ameriquestracing.com	13.73	0.83 (6.02%)	0.78 (5.67%)	0.74 (5.39%)	0.70 (5.07%)
	City of LA	43: www.lacity.org	46.79	9.68 (20.7%)	8.77 (18.75%)	8.23 (17.58%)	7.90 (16.89%)
		44: eng.lacity.org	39.16	4.55 (11.61%)	4.31 (10.99%)	4.22 (10.78%)	4.28 (10.92%)
		45: publiccsd.lacity.org	31.67	5.60 (17.69%)	5.16 (16.29%)	4.88 (15.41%)	4.72 (14.9%)
		46: parc1.lacity.org	4.13	0.70 (16.94%)	0.64 (15.55%)	0.60 (14.61%)	0.58 (13.96%)
		47: www.griffithobservatory.org	28.77	4.48 (15.56%)	4.13 (14.37%)	3.95 (13.74%)	3.87 (13.46%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	138.60	16.30 (11.76%)	15.46 (11.15%)	14.93 (10.77%)	14.57 (10.52%)
		49: www.santa-clara.com	73.96	6.73 (9.09%)	6.30 (8.51%)	5.99 (8.1%)	5.71 (7.73%)
		50: www.santacruz.com	76.46	5.88 (7.69%)	5.57 (7.28%)	5.35 (7%)	5.15 (6.73%)
		51: www.oakland.com	119.42	12.76 (10.68%)	12.03 (10.08%)	11.68 (9.78%)	11.44 (9.58%)
	CA gov	52: democrats.assembly.ca.gov	18.76	0.99 (5.28%)	0.92 (4.9%)	0.88 (4.68%)	0.84 (4.45%)
		53: www.legislature.ca.gov	32.92	1.95 (5.92%)	1.82 (5.52%)	1.74 (5.27%)	1.65 (5.02%)
		54: republican.assembly.ca.gov	23.84	3.42 (14.36%)	3.23 (13.55%)	3.09 (12.96%)	2.96 (12.41%)
	City of Davis	55: www.city.davis.ca.us	52.92	7.57 (14.31%)	7.10 (13.41%)	6.80 (12.85%)	6.52 (12.32%)
		56: events.dcn.org	92.78	7.79 (8.4%)	7.30 (7.87%)	6.96 (7.51%)	6.67 (7.19%)
		57: www.dcn.org	16.65	2.49 (14.93%)	2.32 (13.91%)	2.21 (13.29%)	2.11 (12.68%)
Dallas, TX	Dallas News	58: www.dallasnews.com	301.15	42.57 (14.14%)	39.70 (13.18%)	38.25 (12.7%)	37.34 (12.4%)
		59: www.cowboysplus.com	147.28	18.76 (12.74%)	17.24 (11.7%)	16.23 (11.02%)	15.38 (10.45%)
		60: www.guidelive.com	94.31	20.69 (21.94%)	18.82 (19.95%)	17.49 (18.55%)	16.50 (17.49%)
	City of Irving	61: www.ci.irving.tx.us	42.13	7.97 (18.91%)	7.49 (17.79%)	7.26 (17.23%)	7.26 (17.24%)
	Online Video	62: www.lapdonline.org	195.61	23.28 (11.9%)	22.13 (11.31%)	21.77 (11.13%)	21.46 (10.97%)

Table A.5: Available Bandwidth Prediction Results with Different λ Values

location	cluster name	web servers	average actual values (Mbits/s)	λ=0 mean prediction errors (Mbits/s)	λ=0.25 mean prediction errors (Mbits/s)	λ=0.5 mean prediction errors (Mbits/s)	λ=0.75 mean prediction errors (Mbits/s)
Boston, MA	Boston Globe	1: www.boston.com	802.28	667.59 (83.21%)	614.89 (76.64%)	578.04 (72.05%)	554.05 (69.06%)
		2: weather.boston.com	809.11	714.26 (88.28%)	685.27 (84.69%)	666.06 (82.32%)	649.99 (80.33%)
		3: www.explorenewengland.com	539.48	694.43 (128.72%)	683.14 (126.63%)	680.56 (126.15%)	679.63 (125.98%)
	MBTA	4: www.mbta.com	240.09	274.94 (114.51%)	266.82 (111.13%)	260.83 (108.64%)	256.66 (106.9%)
		5: trip.mbta.com	272.37	316.56 (116.22%)	302.03 (110.89%)	290.70 (106.73%)	279.24 (102.52%)
	Web Hosting	6: www.aviationdisasterlawyers.com	3029.81	937.00 (30.93%)	890.34 (29.39%)	855.44 (28.23%)	833.53 (27.51%)
		7: www.asbestoslaw.info	2991.38	994.73 (33.25%)	948.13 (31.7%)	914.02 (30.56%)	890.13 (29.76%)
		8: www.pharmaceuticallawyers.com	2961.10	1018.14 (34.38%)	958.73 (32.38%)	913.90 (30.86%)	884.39 (29.87%)
New York, NY	NYTimes	9: www.nytimes.com	2158.96	1035.36 (47.96%)	955.27 (44.25%)	882.57 (40.88%)	817.04 (37.84%)
		10: movies.nytimes.com	2118.18	1118.28 (52.79%)	1022.22 (48.26%)	933.98 (44.09%)	864.52 (40.81%)
		11: homefinance.nytimes.com	1879.20	1018.38 (54.19%)	932.35 (49.61%)	853.69 (45.43%)	791.33 (42.11%)
		12: query.nytimes.com	1858.99	974.36 (52.41%)	892.91 (48.03%)	814.88 (43.83%)	748.42 (40.26%)
	UN	13: www.undp.org	2203.93	785.93 (35.66%)	717.39 (32.55%)	661.15 (30%)	619.71 (28.12%)
		14: www.rbas.undp.org	2430.16	946.66 (38.95%)	887.61 (36.52%)	827.48 (34.05%)	774.73 (31.88%)
		15: www.dz.undp.org	2134.97	658.81 (30.86%)	602.96 (28.24%)	556.59 (26.07%)	518.64 (24.29%)
		16: google.undp.org	2164.12	1124.75 (51.97%)	1033.55 (47.76%)	950.92 (43.94%)	896.68 (41.43%)
Atlanta, GA	CNN	17: www.cnn.com	3588.08	217.15 (6.05%)	215.63 (6.01%)	214.12 (5.97%)	213.00 (5.94%)
		18: edition.cnn.com	3358.65	426.84 (12.71%)	422.92 (12.59%)	418.34 (12.46%)	413.76 (12.32%)
		19: si.cnn.com	3474.39	359.14 (10.34%)	354.83 (10.21%)	350.42 (10.09%)	347.09 (9.99%)
		20: money.cnn.com	3323.31	444.64 (13.38%)	434.04 (13.06%)	428.98 (12.91%)	427.78 (12.87%)
	Weather.com	21: www.weather.com	3451.94	254.81 (7.38%)	242.81 (7.03%)	231.77 (6.71%)	222.21 (6.44%)
		22: forgetaway.weather.com	3226.51	420.86 (13.04%)	413.40 (12.81%)	407.00 (12.61%)	402.33 (12.47%)
		23: desktopfw.weather.com	3187.47	431.39 (13.53%)	420.47 (13.19%)	410.96 (12.89%)	401.59 (12.6%)
		24: br.weather.com	3065.98	500.85 (16.34%)	489.74 (15.97%)	479.84 (15.65%)	473.58 (15.45%)
	GA gov	25: www.georgia.gov	2324.80	621.08 (26.72%)	584.12 (25.13%)	550.46 (23.68%)	521.28 (22.42%)
		26: www.files.georgia.gov	2884.89	474.01 (16.43%)	464.42 (16.1%)	455.18 (15.78%)	446.20 (15.47%)
		27: oca.awe.gta.ga.gov	3176.62	403.33 (12.7%)	398.81 (12.55%)	394.19 (12.41%)	391.06 (12.31%)
		28: www.gov.state.ga.us	3190.60	492.54 (15.44%)	473.79 (14.85%)	455.86 (14.29%)	441.43 (13.84%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	2959.29	452.84 (15.3%)	451.80 (15.27%)	448.53 (15.16%)	441.15 (14.91%)
		30: www.illinois.gov	3100.57	447.81 (14.44%)	440.66 (14.21%)	435.24 (14.04%)	430.81 (13.89%)
		31: www.allkidscovered.com	2645.86	373.11 (14.1%)	361.04 (13.65%)	342.82 (12.96%)	312.41 (11.81%)
	IL Education	32: www.isbe.net	3093.16	431.13 (13.94%)	430.43 (13.92%)	430.05 (13.9%)	430.41 (13.91%)

Table A.6: Available Bandwidth Prediction Results with Different λ Values (Continued)

location	cluster name	web servers	average actual values (Mbits/s)	λ=0 mean prediction errors (Mbits/s)	λ=0.25 mean prediction errors (Mbits/s)	λ=0.5 mean prediction errors (Mbits/s)	λ=0.75 mean prediction errors (Mbits/s)
Raymond, WA	MSN	33: www.msn.com	3211.42	506.86 (15.78%)	501.78 (15.63%)	495.31 (15.42%)	487.29 (15.17%)
		34: entertainment.msn.com	3306.96	451.37 (13.65%)	451.15 (13.64%)	450.58 (13.63%)	448.95 (13.58%)
		35: music.msn.com	3329.37	429.87 (12.91%)	433.85 (13.03%)	434.23 (13.04%)	431.68 (12.97%)
		36: weather.msn.com	3136.28	430.85 (13.74%)	429.89 (13.71%)	430.01 (13.71%)	430.59 (13.73%)
	Real	37: www.realnetworks.com	2965.63	577.68 (19.48%)	563.20 (18.99%)	549.11 (18.52%)	539.36 (18.19%)
		38: brasil.real.com	2881.89	607.19 (21.07%)	586.05 (20.34%)	566.31 (19.65%)	554.67 (19.25%)
		39: musicstore.real.com	2817.36	501.55 (17.8%)	481.97 (17.11%)	455.43 (16.17%)	415.41 (14.74%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	2580.57	470.15 (18.22%)	436.67 (16.92%)	401.37 (15.55%)	360.02 (13.95%)
		41: careers.ameriquest.com	1913.84	686.36 (35.86%)	644.22 (33.66%)	615.67 (32.17%)	599.53 (31.33%)
		42: www.ameriquestracing.com	704.36	678.48 (96.33%)	645.13 (91.59%)	622.75 (88.41%)	609.54 (86.54%)
	City of LA	43: www.lacity.org	2013.87	712.23 (35.37%)	649.83 (32.27%)	602.75 (29.93%)	568.98 (28.25%)
		44: eng.lacity.org	2524.40	612.83 (24.28%)	577.01 (22.86%)	540.53 (21.41%)	502.00 (19.89%)
		45: publiccsd.lacity.org	2217.67	707.25 (31.89%)	660.05 (29.76%)	619.89 (27.95%)	587.34 (26.48%)
		46: parc1.lacity.org	3012.80	571.28 (18.96%)	556.38 (18.47%)	539.60 (17.91%)	521.67 (17.32%)
		47: www.griffithobservatory.org	1722.67	550.78 (31.97%)	508.75 (29.53%)	473.34 (27.48%)	439.57 (25.52%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	3144.97	455.60 (14.49%)	441.48 (14.04%)	426.65 (13.57%)	407.55 (12.96%)
		49: www.santa-clara.com	2597.90	578.36 (22.26%)	547.64 (21.08%)	522.97 (20.13%)	505.18 (19.45%)
		50: www.santacruz.com	1096.01	877.24 (80.04%)	812.50 (74.13%)	757.42 (69.11%)	696.55 (63.55%)
		51: www.oakland.com	2938.73	523.51 (17.81%)	502.98 (17.12%)	479.12 (16.3%)	448.43 (15.26%)
	CA gov	52: democrats.assembly.ca.gov	1853.14	722.08 (38.97%)	668.66 (36.08%)	628.59 (33.92%)	598.39 (32.29%)
		53: www.legislature.ca.gov	2312.80	686.32 (29.67%)	638.13 (27.59%)	598.88 (25.89%)	568.21 (24.57%)
		54: republican.assembly.ca.gov	1953.91	747.22 (38.24%)	688.85 (35.26%)	646.73 (33.1%)	618.79 (31.67%)
	City of Davis	55: www.city.davis.ca.us	1661.41	397.25 (23.91%)	368.86 (22.2%)	341.84 (20.58%)	312.01 (18.78%)
		56: events.dcn.org	1772.14	365.25 (20.61%)	339.57 (19.16%)	314.95 (17.77%)	288.08 (16.26%)
		57: www.dcn.org	915.14	636.52 (69.55%)	593.40 (64.84%)	558.84 (61.07%)	527.98 (57.69%)
Dallas, TX	Dallas News	58: www.dallasnews.com	3635.06	162.60 (4.47%)	161.38 (4.44%)	160.03 (4.4%)	158.29 (4.35%)
		59: www.cowboysplus.com	3371.22	449.23 (13.33%)	446.05 (13.23%)	442.57 (13.13%)	438.70 (13.01%)
		60: www.guidelive.com	3371.00	420.36 (12.47%)	420.64 (12.48%)	419.51 (12.44%)	418.75 (12.42%)
	City of Irving	61: www.ci.irving.tx.us	42.02	37.78 (89.92%)	36.59 (87.08%)	36.21 (86.17%)	36.24 (86.26%)
	Online Video	62: www.lapdonline.org	2853.44	489.08 (17.14%)	472.76 (16.57%)	461.83 (16.19%)	463.53 (16.24%)

Table A.7: Connection Ratings Prediction Results with Different λ Values

location	cluster name	web servers	average actual values	λ=0 mean prediction errors	λ=0.25 mean prediction errors	λ=0.5 mean prediction errors	λ=0.75 mean prediction errors
Boston, MA	Boston Globe	1: www.boston.com	0.94	0.15 (16.42%)	0.15 (16.49%)	0.16 (16.59%)	0.16 (16.66%)
		2: weather.boston.com	1.06	0.23 (22.01%)	0.23 (21.83%)	0.23 (21.4%)	0.22 (20.55%)
		3: www.explorenewengland.com	1.27	0.43 (33.72%)	0.42 (33.32%)	0.42 (32.82%)	0.41 (32.17%)
	MBTA	4: www.mbta.com	0.92	0.41 (44.78%)	0.41 (44.26%)	0.40 (43.35%)	0.39 (41.64%)
		5: trip.mbta.com	1.11	0.40 (36.13%)	0.39 (35.38%)	0.38 (34.29%)	0.36 (32.69%)
	Web Hosting	6: www.aviationdisasterlawyers.com	1.18	0.30 (25.57%)	0.30 (25.66%)	0.30 (25.66%)	0.30 (25.54%)
		7: www.asbestoslaw.info	1.15	0.27 (23.1%)	0.27 (23.13%)	0.27 (23.23%)	0.27 (23.36%)
		8: www.pharmaceuticallawyers.com	1.16	0.28 (24.52%)	0.28 (24.39%)	0.28 (24.14%)	0.28 (23.89%)
New York, NY	NYTimes	9: www.nytimes.com	1.45	0.56 (38.61%)	0.55 (37.76%)	0.53 (36.7%)	0.52 (35.75%)
		10: movies.nytimes.com	1.32	0.49 (37.1%)	0.48 (36.4%)	0.47 (35.37%)	0.45 (33.94%)
		11: homefinance.nytimes.com	1.38	0.42 (30.66%)	0.42 (30.58%)	0.42 (30.6%)	0.42 (30.72%)
		12: query.nytimes.com	1.43	0.50 (35.03%)	0.50 (34.77%)	0.49 (34.24%)	0.48 (33.57%)
	UN	13: www.undp.org	1.32	0.46 (34.58%)	0.45 (34.14%)	0.44 (33.48%)	0.43 (32.66%)
		14: www.rbas.undp.org	1.32	0.42 (31.41%)	0.41 (31.12%)	0.41 (30.71%)	0.40 (30.23%)
		15: www.dz.undp.org	1.46	0.52 (35.84%)	0.52 (35.48%)	0.51 (35.09%)	0.51 (34.85%)
		16: google.undp.org	1.39	0.49 (34.83%)	0.48 (34.41%)	0.47 (33.86%)	0.46 (33.25%)
Atlanta, GA	CNN	17: www.cnn.com	1.96	0.06 (3.24%)	0.07 (3.41%)	0.07 (3.58%)	0.07 (3.78%)
		18: edition.cnn.com	1.93	0.11 (5.61%)	0.11 (5.63%)	0.11 (5.68%)	0.11 (5.86%)
		19: si.cnn.com	1.94	0.09 (4.86%)	0.10 (5%)	0.10 (5.14%)	0.10 (5.32%)
		20: money.cnn.com	1.74	0.10 (5.52%)	0.10 (5.56%)	0.10 (5.67%)	0.10 (5.88%)
	Weather.com	21: www.weather.com	1.88	0.17 (8.77%)	0.17 (9%)	0.18 (9.34%)	0.18 (9.82%)
		22: forgetaway.weather.com	1.87	0.22 (11.84%)	0.22 (11.82%)	0.22 (11.88%)	0.23 (12.06%)
		23: desktopfw.weather.com	1.85	0.25 (13.34%)	0.25 (13.36%)	0.25 (13.38%)	0.25 (13.49%)
		24: br.weather.com	1.85	0.25 (13.74%)	0.25 (13.74%)	0.25 (13.72%)	0.25 (13.71%)
	GA gov	25: www.georgia.gov	0.94	0.23 (24.68%)	0.23 (24.63%)	0.23 (24.68%)	0.23 (24.88%)
		26: www.files.georgia.gov	1.88	0.22 (11.43%)	0.21 (11.22%)	0.21 (11.04%)	0.21 (11.01%)
		27: oca.awe.gta.ga.gov	1.87	0.22 (11.98%)	0.22 (11.88%)	0.22 (11.79%)	0.22 (11.84%)
		28: www.gov.state.ga.us	1.62	0.54 (33.52%)	0.54 (33.43%)	0.54 (33.34%)	0.54 (33.39%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	1.85	0.26 (14.14%)	0.26 (14.21%)	0.26 (14.25%)	0.27 (14.31%)
		30: www.illinois.gov	1.85	0.11 (6.04%)	0.11 (6.14%)	0.12 (6.26%)	0.12 (6.5%)
		31: www.allkidscovered.com	1.95	0.09 (4.57%)	0.09 (4.54%)	0.09 (4.52%)	0.09 (4.49%)
	IL Education	32: www.isbe.net	1.96	0.04 (1.79%)	0.04 (1.91%)	0.04 (2.09%)	0.05 (2.4%)

Table A.8: Connection Ratings Prediction Results with Different λ Values

location	cluster name	web servers	average actual values	λ=0 mean prediction errors	λ=0.25 mean prediction errors	λ=0.5 mean prediction errors	λ=0.75 mean prediction errors
Raymond, WA	MSN	33: www.msn.com	1.88	0.11 (6.02%)	0.11 (6.09%)	0.12 (6.25%)	0.13 (6.76%)
		34: entertainment.msn.com	1.99	0.02 (1.24%)	0.03 (1.26%)	0.03 (1.29%)	0.03 (1.34%)
		35: music.msn.com	1.99	0.01 (0.49%)	0.01 (0.51%)	0.01 (0.54%)	0.01 (0.6%)
		36: weather.msn.com	1.99	0.02 (1.06%)	0.02 (1.12%)	0.02 (1.19%)	0.03 (1.29%)
	Real	37: www.realnetworks.com	1.65	0.22 (13.56%)	0.22 (13.59%)	0.23 (13.77%)	0.24 (14.81%)
		38: brasil.real.com	1.34	0.59 (44.06%)	0.58 (43.41%)	0.58 (43.09%)	0.59 (43.89%)
		39: musicstore.real.com	1.96	0.08 (4.04%)	0.08 (4.05%)	0.08 (4.04%)	0.08 (4.06%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	1.52	0.60 (39.52%)	0.60 (39.26%)	0.60 (39.04%)	0.60 (39.09%)
		41: careers.ameriquest.com	0.86	0.99 (114.47%)	0.99 (114.05%)	0.98 (113.74%)	0.98 (113.34%)
		42: www.ameriquestracing.com	0.44	0.69 (157.02%)	0.69 (156.95%)	0.69 (156.52%)	0.69 (155.88%)
	City of LA	43: www.lacity.org	1.21	0.64 (53.09%)	0.62 (51.21%)	0.59 (48.87%)	0.56 (46.25%)
		44: eng.lacity.org	1.73	0.31 (17.68%)	0.30 (17.59%)	0.31 (17.63%)	0.31 (18.13%)
		45: publiccsd.lacity.org	1.81	0.25 (13.52%)	0.24 (13.45%)	0.24 (13.37%)	0.25 (13.61%)
		46: parc1.lacity.org	1.05	0.45 (42.61%)	0.44 (42.01%)	0.43 (41.28%)	0.43 (40.86%)
		47: www.griffithobservatory.org	1.79	0.29 (16.19%)	0.29 (15.91%)	0.28 (15.73%)	0.28 (15.82%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	1.72	0.43 (24.82%)	0.42 (24.77%)	0.42 (24.76%)	0.43 (25.05%)
		49: www.santa-clara.com	1.82	0.32 (17.64%)	0.32 (17.6%)	0.32 (17.64%)	0.32 (17.75%)
		50: www.santacruz.com	1.11	0.27 (24.57%)	0.27 (24.49%)	0.27 (24.12%)	0.26 (23.28%)
		51: www.oakland.com	1.66	0.50 (30.3%)	0.50 (30.16%)	0.50 (30.06%)	0.50 (30.31%)
	CA gov	52: democrats.assembly.ca.gov	1.93	0.12 (6.15%)	0.12 (6.09%)	0.12 (6%)	0.11 (5.94%)
		53: www.legislature.ca.gov	1.93	0.13 (6.79%)	0.13 (6.82%)	0.13 (6.84%)	0.13 (6.87%)
		54: republican.assembly.ca.gov	1.91	0.15 (7.87%)	0.15 (7.9%)	0.15 (7.89%)	0.15 (7.96%)
	City of Davis	55: www.city.davis.ca.us	1.83	0.30 (16.25%)	0.30 (16.27%)	0.30 (16.31%)	0.30 (16.33%)
		56: events.dcn.org	1.92	0.14 (7.08%)	0.14 (7.08%)	0.14 (7.13%)	0.14 (7.22%)
		57: www.dcn.org	0.95	0.09 (9.81%)	0.09 (9.77%)	0.09 (9.75%)	0.09 (9.76%)
Dallas, TX	Dallas News	58: www.dallasnews.com	1.83	0.24 (13.14%)	0.24 (13.16%)	0.24 (13.33%)	0.25 (13.8%)
		59: www.cowboysplus.com	1.96	0.08 (3.94%)	0.08 (4%)	0.08 (4.07%)	0.08 (4.16%)
		60: www.guidelive.com	1.88	0.15 (7.74%)	0.15 (7.9%)	0.15 (8.15%)	0.16 (8.63%)
	City of Irving	61: www.ci.irving.tx.us	1.64	0.30 (18.15%)	0.30 (18.17%)	0.30 (18.12%)	0.30 (18.26%)
	Online Video	62: www.lapdonline.org	1.78	0.37 (20.76%)	0.37 (20.8%)	0.37 (20.82%)	0.37 (20.89%)

Table A.9: RTT Prediction Results with Different Collection Intervals when $\lambda = 0.5$

location	cluster name	web servers	average actual values (ms)	interval=10min mean prediction errors (ms)	interval=1hr mean prediction errors (ms)	interval=2hr mean prediction errors (ms)	interval=4hr mean prediction errors (ms)
Boston, MA	Boston Globe	1: www.boston.com	17.56	7.76 (44.19%)	9.33 (53.12%)	9.92 (56.48%)	10.57 (60.17%)
		2: weather.boston.com	15.48	6.69 (43.25%)	8.63 (55.78%)	9.30 (60.12%)	10.02 (64.73%)
		3: www.explorenewengland.com	11.86	4.70 (39.65%)	5.14 (43.35%)	5.46 (46.02%)	5.73 (48.36%)
	MBTA	4: www.mbta.com	36.74	36.12 (98.33%)	36.65 (99.75%)	38.70 (105.34%)	40.65 (110.65%)
		5: trip.mbta.com	27.85	17.22 (61.84%)	19.55 (70.22%)	21.37 (76.76%)	23.12 (83.02%)
	Web Hosting	6: www.aviationdisasterlawyers.com	10.42	5.56 (53.4%)	5.49 (52.65%)	5.53 (53.07%)	5.58 (53.55%)
		7: www.asbestoslaw.info	10.36	4.95 (47.81%)	5.02 (48.44%)	5.03 (48.59%)	5.04 (48.66%)
		8: www.pharmaceuticallawyers.com	9.85	5.07 (51.46%)	5.08 (51.55%)	5.08 (51.59%)	5.12 (51.98%)
New York, NY	NYTimes	9: www.nytimes.com	16.09	5.56 (34.59%)	5.60 (34.8%)	5.78 (35.94%)	5.87 (36.5%)
		10: movies.nytimes.com	16.74	5.07 (30.28%)	5.06 (30.22%)	5.20 (31.08%)	5.40 (32.26%)
		11: homefinance.nytimes.com	17.10	6.14 (35.91%)	6.23 (36.46%)	6.29 (36.81%)	6.37 (37.25%)
		12: query.nytimes.com	16.07	4.97 (30.93%)	4.90 (30.5%)	5.05 (31.43%)	5.25 (32.64%)
	UN	13: www.undp.org	18.27	5.98 (32.73%)	5.94 (32.49%)	6.05 (33.1%)	6.20 (33.95%)
		14: www.rbas.undp.org	17.69	5.05 (28.53%)	5.16 (29.16%)	5.34 (30.17%)	5.44 (30.78%)
		15: www.dz.undp.org	18.31	5.84 (31.88%)	6.02 (32.87%)	6.05 (33.06%)	6.36 (34.71%)
		16: google.undp.org	17.76	6.08 (34.23%)	6.32 (35.58%)	6.40 (36.06%)	6.50 (36.6%)
Atlanta, GA	CNN	17: www.cnn.com	38.65	1.04 (2.69%)	1.15 (2.99%)	1.21 (3.13%)	1.31 (3.38%)
		18: edition.cnn.com	40.41	3.53 (8.73%)	3.62 (8.95%)	3.68 (9.11%)	3.79 (9.38%)
		19: si.cnn.com	39.36	2.18 (5.54%)	2.32 (5.89%)	2.36 (6.01%)	2.46 (6.25%)
		20: money.cnn.com	44.24	10.90 (24.64%)	11.01 (24.89%)	11.08 (25.05%)	11.20 (25.32%)
	Weather.com	21: www.weather.com	33.45	2.68 (8.02%)	2.89 (8.65%)	2.96 (8.86%)	3.11 (9.28%)
		22: forgetaway.weather.com	33.90	2.59 (7.63%)	2.79 (8.22%)	2.84 (8.37%)	2.89 (8.52%)
		23: desktopfw.weather.com	34.28	2.68 (7.83%)	2.87 (8.37%)	2.96 (8.62%)	3.07 (8.95%)
		24: br.weather.com	34.62	2.77 (8.01%)	3.00 (8.66%)	3.07 (8.87%)	3.20 (9.26%)
	GA gov	25: www.georgia.gov	38.57	3.39 (8.79%)	4.30 (11.15%)	5.16 (13.38%)	6.06 (15.72%)
		26: www.files.georgia.gov	37.61	2.77 (7.38%)	3.71 (9.86%)	4.59 (12.21%)	5.41 (14.37%)
		27: oca.awe.gta.ga.gov	40.22	9.33 (23.2%)	10.20 (25.37%)	10.95 (27.22%)	11.42 (28.4%)
		28: www.gov.state.ga.us	37.08	2.66 (7.18%)	3.61 (9.72%)	4.49 (12.1%)	5.36 (14.45%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	46.41	3.95 (8.51%)	4.67 (10.07%)	5.06 (10.89%)	5.35 (11.52%)
		30: www.illinois.gov	43.39	2.11 (4.87%)	2.85 (6.57%)	3.13 (7.22%)	3.29 (7.57%)
		31: www.allkidscovered.com	47.64	5.32 (11.16%)	6.15 (12.91%)	6.48 (13.61%)	6.72 (14.1%)
	IL Education	32: www.isbe.net	42.12	4.29 (10.19%)	7.53 (17.88%)	9.87 (23.44%)	11.97 (28.43%)

Table A.10: RTT Prediction Results with Different Collection Intervals when $\lambda = 0.5$ (Continued)

location	cluster name	web servers	average actual values (ms)	interval=10min mean prediction errors (ms)	interval=1hr mean prediction errors (ms)	interval=2hr mean prediction errors (ms)	interval=4hr mean prediction errors (ms)
Raymond, WA	MSN	33: www.msn.com	93.43	2.19 (2.34%)	2.92 (3.12%)	3.44 (3.68%)	3.99 (4.27%)
		34: entertainment.msn.com	92.10	1.45 (1.58%)	1.75 (1.9%)	1.88 (2.04%)	2.02 (2.2%)
		35: music.msn.com	92.04	1.43 (1.55%)	1.71 (1.86%)	1.85 (2.01%)	1.99 (2.16%)
		36: weather.msn.com	92.48	2.03 (2.2%)	2.31 (2.5%)	2.46 (2.66%)	2.59 (2.8%)
	Real	37: www.realnetworks.com	97.85	6.78 (6.93%)	8.93 (9.12%)	9.96 (10.18%)	10.96 (11.2%)
		38: brasil.real.com	98.08	6.94 (7.07%)	8.86 (9.04%)	9.80 (9.99%)	10.74 (10.95%)
		39: musicstore.real.com	76.98	1.29 (1.67%)	1.34 (1.74%)	1.34 (1.74%)	1.36 (1.77%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	89.04	1.16 (1.3%)	1.17 (1.31%)	1.18 (1.33%)	1.21 (1.36%)
		41: careers.ameriquest.com	89.02	1.68 (1.88%)	1.70 (1.91%)	1.72 (1.94%)	1.72 (1.93%)
		42: www.ameriquestracing.com	92.11	4.62 (5.02%)	4.65 (5.04%)	4.64 (5.04%)	4.64 (5.03%)
	City of LA	43: www.lacity.org	115.76	28.83 (24.91%)	32.12 (27.75%)	33.60 (29.03%)	34.47 (29.78%)
		44: eng.lacity.org	98.29	12.51 (12.73%)	15.55 (15.82%)	17.23 (17.53%)	18.09 (18.4%)
		45: publiccsd.lacity.org	98.26	9.39 (9.56%)	13.03 (13.26%)	14.24 (14.5%)	14.94 (15.21%)
		46: parc1.lacity.org	97.47	10.85 (11.13%)	14.58 (14.96%)	15.78 (16.19%)	16.60 (17.03%)
		47: www.griffithobservatory.org	109.91	32.20 (29.29%)	36.65 (33.35%)	38.75 (35.26%)	39.38 (35.83%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	85.39	3.10 (3.63%)	3.15 (3.69%)	3.19 (3.74%)	3.22 (3.77%)
		49: www.santa-clara.com	84.60	1.55 (1.83%)	1.63 (1.93%)	1.65 (1.95%)	1.66 (1.97%)
		50: www.santacruz.com	84.96	4.63 (5.45%)	4.70 (5.54%)	4.65 (5.48%)	4.65 (5.48%)
		51: www.oakland.com	83.56	2.90 (3.47%)	2.94 (3.52%)	2.98 (3.57%)	3.00 (3.59%)
	CA gov	52: democrats.assembly.ca.gov	87.78	4.99 (5.68%)	5.11 (5.83%)	5.20 (5.92%)	5.28 (6.01%)
		53: www.legislature.ca.gov	87.20	2.42 (2.78%)	2.51 (2.88%)	2.58 (2.96%)	2.71 (3.11%)
		54: republican.assembly.ca.gov	86.96	5.90 (6.79%)	6.06 (6.97%)	6.18 (7.11%)	6.31 (7.25%)
	City of Davis	55: www.city.davis.ca.us	103.86	9.71 (9.35%)	8.75 (8.42%)	9.87 (9.5%)	10.85 (10.45%)
		56: events.dcn.org	94.20	2.01 (2.13%)	2.24 (2.38%)	2.48 (2.63%)	2.79 (2.96%)
		57: www.dcn.org	95.51	2.64 (2.76%)	2.84 (2.97%)	3.01 (3.15%)	3.19 (3.34%)
Dallas, TX	Dallas News	58: www.dallasnews.com	53.72	3.14 (5.84%)	3.75 (6.98%)	3.98 (7.41%)	4.24 (7.89%)
		59: www.cowboysplus.com	54.90	3.51 (6.4%)	4.13 (7.52%)	4.41 (8.04%)	4.67 (8.52%)
		60: www.guidelive.com	53.12	2.84 (5.34%)	3.52 (6.63%)	3.83 (7.21%)	4.11 (7.73%)
	City of Irving	61: www.ci.irving.tx.us	90.70	40.85 (45.04%)	46.60 (51.38%)	47.36 (52.22%)	50.81 (56.02%)
	Online Video	62: www.lapdonline.org	54.00	1.95 (3.61%)	2.33 (4.31%)	2.53 (4.68%)	2.81 (5.2%)

Table A.11: Connection Throughput Prediction Results with Different Collection Intervals when $\lambda = 0.5$

location	cluster name	web servers	average actual values (KB/s)	interval=10min mean prediction errors (KB/s)	interval=1hr mean prediction errors (KB/s)	interval=2hr mean prediction errors (KB/s)	interval=4hr mean prediction errors (KB/s)
Boston, MA	Boston Globe	1: www.boston.com	410.03	190.08 (46.36%)	199.81 (48.73%)	216.24 (52.74%)	236.88 (57.77%)
		2: weather.boston.com	354.59	50.39 (14.21%)	63.24 (17.84%)	77.63 (21.89%)	91.90 (25.92%)
		3: www.explorenewengland.com	261.94	30.47 (11.63%)	35.76 (13.65%)	39.71 (15.16%)	43.67 (16.67%)
	MBTA	4: www.mbta.com	153.09	46.33 (30.27%)	55.92 (36.53%)	60.58 (39.57%)	65.11 (42.53%)
		5: trip.mbta.com	236.20	71.79 (30.39%)	80.31 (34%)	86.97 (36.82%)	95.17 (40.29%)
	Web Hosting	6: www.aviationdisasterlawyers.com	1466.44	617.69 (42.12%)	659.87 (45%)	686.93 (46.84%)	708.24 (48.3%)
		7: www.asbestoslaw.info	1926.40	699.25 (36.3%)	759.18 (39.41%)	790.91 (41.06%)	831.88 (43.18%)
		8: www.pharmaceuticallawyers.com	1738.13	594.55 (34.21%)	647.43 (37.25%)	667.58 (38.41%)	694.79 (39.97%)
New York, NY	NYTimes	9: www.nytimes.com	826.75	110.23 (13.33%)	119.34 (14.43%)	122.93 (14.87%)	127.18 (15.38%)
		10: movies.nytimes.com	765.26	119.40 (15.6%)	121.37 (15.86%)	125.63 (16.42%)	130.84 (17.1%)
		11: homefinance.nytimes.com	310.01	30.18 (9.73%)	31.46 (10.15%)	31.83 (10.27%)	32.88 (10.61%)
		12: query.nytimes.com	69.27	19.30 (27.86%)	19.47 (28.1%)	19.56 (28.24%)	19.71 (28.45%)
	UN	13: www.undp.org	295.06	43.36 (14.7%)	45.12 (15.29%)	46.42 (15.73%)	48.50 (16.44%)
		14: www.rbas.undp.org	364.88	47.41 (12.99%)	49.98 (13.7%)	52.43 (14.37%)	55.19 (15.13%)
		15: www.dz.undp.org	233.30	57.09 (24.47%)	73.64 (31.56%)	85.28 (36.55%)	99.02 (42.44%)
		16: google.undp.org	28.54	7.77 (27.24%)	7.40 (25.93%)	7.52 (26.35%)	7.65 (26.82%)
Atlanta, GA	CNN	17: www.cnn.com	305.56	20.37 (6.67%)	21.92 (7.17%)	22.33 (7.31%)	23.16 (7.58%)
		18: edition.cnn.com	259.17	19.00 (7.33%)	19.82 (7.65%)	20.40 (7.87%)	21.36 (8.24%)
		19: si.cnn.com	214.58	47.68 (22.22%)	48.24 (22.48%)	48.22 (22.47%)	47.65 (22.21%)
		20: money.cnn.com	230.45	18.35 (7.96%)	19.47 (8.45%)	20.56 (8.92%)	22.22 (9.64%)
	Weather.com	21: www.weather.com	457.36	53.90 (11.78%)	59.24 (12.95%)	60.18 (13.16%)	63.03 (13.78%)
		22: forgetaway.weather.com	267.24	24.80 (9.28%)	26.30 (9.84%)	26.79 (10.02%)	27.48 (10.28%)
		23: desktopfw.weather.com	246.97	29.41 (11.91%)	31.83 (12.89%)	33.29 (13.48%)	35.73 (14.47%)
		24: br.weather.com	156.74	12.20 (7.79%)	12.79 (8.16%)	13.15 (8.39%)	13.59 (8.67%)
	GA gov	25: www.georgia.gov	67.05	30.37 (45.29%)	32.21 (48.03%)	32.57 (48.58%)	33.51 (49.97%)
		26: www.files.georgia.gov	132.81	45.85 (34.52%)	46.93 (35.34%)	48.18 (36.28%)	49.47 (37.25%)
		27: oca.awe.gta.ga.gov	180.52	13.73 (7.61%)	15.35 (8.5%)	16.90 (9.36%)	18.81 (10.42%)
		28: www.gov.state.ga.us	280.68	52.22 (18.61%)	55.62 (19.82%)	58.76 (20.94%)	62.84 (22.39%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	81.86	28.30 (34.57%)	29.61 (36.17%)	30.56 (37.33%)	31.58 (38.58%)
		30: www.illinois.gov	71.90	12.94 (18%)	14.37 (19.99%)	15.84 (22.03%)	17.57 (24.44%)
		31: www.allkidscovered.com	64.90	5.01 (7.73%)	5.64 (8.68%)	5.92 (9.11%)	6.16 (9.49%)
	IL Education	32: www.isbe.net	155.14	7.35 (4.74%)	10.23 (6.6%)	12.54 (8.08%)	15.10 (9.73%)

Table A.12: Connection Throughput Prediction Results with Different Collection Intervals when $\lambda = 0.5$ (Continued)

location	cluster name	web servers	average actual values (KB/s)	interval=10min mean prediction errors (KB/s)	interval=1hr mean prediction errors (KB/s)	interval=2hr mean prediction errors (KB/s)	interval=4hr mean prediction errors (KB/s)
Raymond, WA	MSN	33: www.msn.com	50.97	15.60 (30.61%)	16.66 (32.69%)	17.84 (34.99%)	18.82 (36.93%)
		34: entertainment.msn.com	51.96	4.53 (8.72%)	4.64 (8.94%)	4.77 (9.18%)	4.88 (9.39%)
		35: music.msn.com	64.52	4.46 (6.92%)	4.62 (7.16%)	4.72 (7.32%)	4.83 (7.49%)
		36: weather.msn.com	53.40	3.46 (6.49%)	4.05 (7.59%)	4.17 (7.81%)	4.47 (8.37%)
	Real	37: www.realnetworks.com	54.13	5.33 (9.85%)	5.80 (10.71%)	6.16 (11.39%)	6.68 (12.35%)
		38: brasil.real.com	54.13	10.02 (18.5%)	10.92 (20.16%)	11.64 (21.5%)	12.75 (23.55%)
		39: musicstore.real.com	46.25	2.36 (5.11%)	2.45 (5.29%)	2.46 (5.33%)	2.49 (5.39%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	31.79	5.87 (18.47%)	5.99 (18.83%)	5.98 (18.81%)	6.04 (19.01%)
		41: careers.ameriquest.com	24.21	5.70 (23.55%)	5.68 (23.48%)	5.75 (23.76%)	5.78 (23.86%)
		42: www.ameriquestracing.com	13.73	0.74 (5.39%)	0.75 (5.48%)	0.78 (5.65%)	0.81 (5.87%)
	City of LA	43: www.lacity.org	46.79	8.23 (17.58%)	9.14 (19.54%)	9.84 (21.04%)	10.73 (22.94%)
		44: eng.lacity.org	39.16	4.22 (10.78%)	5.17 (13.2%)	5.89 (15.05%)	6.76 (17.26%)
		45: publiccsd.lacity.org	31.67	4.88 (15.41%)	5.57 (17.57%)	6.37 (20.11%)	7.27 (22.95%)
		46: parc1.lacity.org	4.13	0.60 (14.61%)	0.63 (15.2%)	0.66 (16%)	0.71 (17.08%)
		47: www.griffithobservatory.org	28.77	3.95 (13.74%)	4.59 (15.94%)	5.30 (18.43%)	6.05 (21.04%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	138.60	14.93 (10.77%)	15.51 (11.19%)	15.83 (11.42%)	16.31 (11.77%)
		49: www.santa-clara.com	73.96	5.99 (8.1%)	5.89 (7.97%)	5.93 (8.02%)	5.99 (8.09%)
		50: www.santacruz.com	76.46	5.35 (7%)	5.58 (7.3%)	5.65 (7.38%)	5.63 (7.36%)
		51: www.oakland.com	119.42	11.68 (9.78%)	11.92 (9.98%)	12.12 (10.15%)	12.22 (10.24%)
	CA gov	52: democrats.assembly.ca.gov	18.76	0.88 (4.68%)	0.92 (4.89%)	0.93 (4.96%)	0.95 (5.04%)
		53: www.legislature.ca.gov	32.92	1.74 (5.27%)	1.81 (5.5%)	1.90 (5.77%)	2.00 (6.07%)
		54: republican.assembly.ca.gov	23.84	3.09 (12.96%)	2.99 (12.55%)	2.70 (11.32%)	2.85 (11.94%)
	City of Davis	55: www.city.davis.ca.us	52.92	6.80 (12.85%)	7.01 (13.24%)	7.27 (13.75%)	7.41 (14%)
		56: events.dcn.org	92.78	6.96 (7.51%)	7.26 (7.83%)	7.46 (8.04%)	7.74 (8.34%)
		57: www.dcn.org	16.65	2.21 (13.29%)	2.27 (13.62%)	2.33 (13.99%)	2.38 (14.29%)
Dallas, TX	Dallas News	58: www.dallasnews.com	301.15	38.25 (12.7%)	42.79 (14.21%)	45.36 (15.06%)	48.63 (16.15%)
		59: www.cowboysplus.com	147.28	16.23 (11.02%)	17.48 (11.87%)	18.05 (12.26%)	18.72 (12.71%)
		60: www.guidelive.com	94.31	17.49 (18.55%)	18.54 (19.66%)	19.59 (20.77%)	20.87 (22.13%)
	City of Irving	61: www.ci.irving.tx.us	42.13	7.26 (17.23%)	8.60 (20.41%)	9.24 (21.93%)	10.55 (25.03%)
	Online Video	62: www.lapdonline.org	195.61	21.77 (11.13%)	23.92 (12.23%)	24.24 (12.39%)	25.00 (12.78%)

Table A.13: Available Bandwidth Prediction Results with Different Collection Intervals when $\lambda = 0.5$

location	cluster name	web servers	average actual values (Mbits/s)	interval=10min mean prediction errors (Mbits/s)	interval=1hr mean prediction errors (Mbits/s)	interval=2hr mean prediction errors (Mbits/s)	interval=4hr mean prediction errors (Mbits/s)
Boston, MA	Boston Globe	1: www.boston.com	802.28	578.04 (72.05%)	606.59 (75.61%)	626.59 (78.1%)	654.42 (81.57%)
		2: weather.boston.com	809.11	666.06 (82.32%)	686.95 (84.9%)	711.96 (87.99%)	744.06 (91.96%)
		3: www.explorenewengland.com	539.48	680.56 (126.15%)	688.51 (127.63%)	686.94 (127.34%)	696.42 (129.09%)
	MBTA	4: www.mbta.com	240.09	260.83 (108.64%)	262.09 (109.16%)	272.16 (113.36%)	281.75 (117.35%)
		5: trip.mbta.com	272.37	290.70 (106.73%)	291.10 (106.88%)	295.31 (108.42%)	297.88 (109.37%)
	Web Hosting	6: www.aviationdisasterlawyers.com	3029.81	855.44 (28.23%)	860.53 (28.4%)	879.89 (29.04%)	877.98 (28.98%)
		7: www.asbestoslaw.info	2991.38	914.02 (30.56%)	920.46 (30.77%)	911.35 (30.47%)	913.57 (30.54%)
		8: www.pharmaceuticallawyers.com	2961.10	913.90 (30.86%)	929.13 (31.38%)	916.53 (30.95%)	917.72 (30.99%)
New York, NY	NYTimes	9: www.nytimes.com	2158.96	882.57 (40.88%)	885.48 (41.01%)	886.91 (41.08%)	879.20 (40.72%)
		10: movies.nytimes.com	2118.18	933.98 (44.09%)	929.90 (43.9%)	934.73 (44.13%)	946.27 (44.67%)
		11: homefinance.nytimes.com	1879.20	853.69 (45.43%)	852.54 (45.37%)	845.20 (44.98%)	850.15 (45.24%)
		12: query.nytimes.com	1858.99	814.88 (43.83%)	802.45 (43.17%)	811.67 (43.66%)	823.17 (44.28%)
	UN	13: www.undp.org	2203.93	661.15 (30%)	659.06 (29.9%)	660.71 (29.98%)	669.90 (30.4%)
		14: www.rbas.undp.org	2430.16	827.48 (34.05%)	834.19 (34.33%)	845.58 (34.8%)	846.79 (34.85%)
		15: www.dz.undp.org	2134.97	556.59 (26.07%)	573.00 (26.84%)	579.85 (27.16%)	583.69 (27.34%)
		16: google.undp.org	2164.12	950.92 (43.94%)	974.87 (45.05%)	980.64 (45.31%)	980.79 (45.32%)
Atlanta, GA	CNN	17: www.cnn.com	3588.08	214.12 (5.97%)	215.46 (6%)	215.70 (6.01%)	219.65 (6.12%)
		18: edition.cnn.com	3358.65	418.34 (12.46%)	423.22 (12.6%)	424.54 (12.64%)	425.77 (12.68%)
		19: si.cnn.com	3474.39	350.42 (10.09%)	349.92 (10.07%)	354.40 (10.2%)	353.39 (10.17%)
		20: money.cnn.com	3323.31	428.98 (12.91%)	433.16 (13.03%)	427.31 (12.86%)	426.66 (12.84%)
	Weather.com	21: www.weather.com	3451.94	231.77 (6.71%)	244.81 (7.09%)	248.79 (7.21%)	268.97 (7.79%)
		22: forgetaway.weather.com	3226.51	407.00 (12.61%)	411.62 (12.76%)	418.52 (12.97%)	441.53 (13.68%)
		23: desktopfw.weather.com	3187.47	410.96 (12.89%)	420.52 (13.19%)	430.43 (13.5%)	446.48 (14.01%)
		24: br.weather.com	3065.98	479.84 (15.65%)	485.69 (15.84%)	489.44 (15.96%)	504.67 (16.46%)
	GA gov	25: www.georgia.gov	2324.80	550.46 (23.68%)	566.93 (24.39%)	564.91 (24.3%)	567.87 (24.43%)
		26: www.files.georgia.gov	2884.89	455.18 (15.78%)	452.98 (15.7%)	459.75 (15.94%)	462.80 (16.04%)
		27: oca.awe.gta.ga.gov	3176.62	394.19 (12.41%)	400.99 (12.62%)	403.59 (12.7%)	397.88 (12.53%)
		28: www.gov.state.ga.us	3190.60	455.86 (14.29%)	463.74 (14.53%)	464.80 (14.57%)	471.86 (14.79%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	2959.29	448.53 (15.16%)	447.32 (15.12%)	446.16 (15.08%)	436.21 (14.74%)
		30: www.illinois.gov	3100.57	435.24 (14.04%)	436.35 (14.07%)	434.21 (14%)	435.42 (14.04%)
		31: www.allkidscovered.com	2645.86	342.82 (12.96%)	339.99 (12.85%)	342.06 (12.93%)	345.56 (13.06%)
	IL Education	32: www.isbe.net	3093.16	430.05 (13.9%)	431.24 (13.94%)	434.59 (14.05%)	436.95 (14.13%)

Table A.14: Available Bandwidth Prediction Results with Different Collection Intervals when $\lambda = 0.5$ (Continued)

location	cluster name	web servers	average actual values (Mbits/s)	interval=10min mean prediction errors (Mbits/s)	interval=1hr mean prediction errors (Mbits/s)	interval=2hr mean prediction errors (Mbits/s)	interval=4hr mean prediction errors (Mbits/s)
Raymond, WA	MSN	33: www.msn.com	3211.42	495.31 (15.42%)	503.97 (15.69%)	507.58 (15.81%)	514.83 (16.03%)
		34: entertainment.msn.com	3306.96	450.58 (13.63%)	456.52 (13.8%)	452.11 (13.67%)	450.76 (13.63%)
		35: music.msn.com	3329.37	434.23 (13.04%)	429.23 (12.89%)	428.72 (12.88%)	434.06 (13.04%)
		36: weather.msn.com	3136.28	430.01 (13.71%)	430.03 (13.71%)	437.82 (13.96%)	439.23 (14%)
	Real	37: www.realnetworks.com	2965.63	549.11 (18.52%)	609.97 (20.57%)	659.67 (22.24%)	704.38 (23.75%)
		38: brasil.real.com	2881.89	566.31 (19.65%)	623.51 (21.64%)	670.16 (23.25%)	717.97 (24.91%)
		39: musicstore.real.com	2817.36	455.43 (16.17%)	453.18 (16.09%)	454.81 (16.14%)	456.29 (16.2%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	2580.57	401.37 (15.55%)	400.34 (15.51%)	399.67 (15.49%)	403.79 (15.65%)
		41: careers.ameriquest.com	1913.84	615.67 (32.17%)	628.70 (32.85%)	630.34 (32.94%)	630.28 (32.93%)
		42: www.ameriquestracing.com	704.36	622.75 (88.41%)	621.46 (88.23%)	623.99 (88.59%)	634.64 (90.1%)
	City of LA	43: www.lacity.org	2013.87	602.75 (29.93%)	597.58 (29.67%)	599.75 (29.78%)	602.13 (29.9%)
		44: eng.lacity.org	2524.40	540.53 (21.41%)	546.56 (21.65%)	551.09 (21.83%)	551.99 (21.87%)
		45: publiccsd.lacity.org	2217.67	619.89 (27.95%)	628.12 (28.32%)	624.10 (28.14%)	626.85 (28.27%)
		46: parc1.lacity.org	3012.80	539.60 (17.91%)	554.31 (18.4%)	558.45 (18.54%)	566.00 (18.79%)
		47: www.griffithobservatory.org	1722.67	473.34 (27.48%)	479.71 (27.85%)	476.45 (27.66%)	471.61 (27.38%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	3144.97	426.65 (13.57%)	423.22 (13.46%)	428.13 (13.61%)	431.53 (13.72%)
		49: www.santa-clara.com	2597.90	522.97 (20.13%)	528.84 (20.36%)	527.81 (20.32%)	533.67 (20.54%)
		50: www.santacruz.com	1096.01	757.42 (69.11%)	753.43 (68.74%)	748.75 (68.32%)	748.46 (68.29%)
		51: www.oakland.com	2938.73	479.12 (16.3%)	483.37 (16.45%)	481.78 (16.39%)	486.45 (16.55%)
	CA gov	52: democrats.assembly.ca.gov	1853.14	628.59 (33.92%)	621.95 (33.56%)	618.22 (33.36%)	618.76 (33.39%)
		53: www.legislature.ca.gov	2312.80	598.88 (25.89%)	602.94 (26.07%)	599.17 (25.91%)	613.50 (26.53%)
		54: republican.assembly.ca.gov	1953.91	646.73 (33.1%)	642.89 (32.9%)	644.72 (33%)	644.47 (32.98%)
	City of Davis	55: www.city.davis.ca.us	1661.41	341.84 (20.58%)	338.90 (20.4%)	338.95 (20.4%)	340.07 (20.47%)
		56: events.dcn.org	1772.14	314.95 (17.77%)	314.24 (17.73%)	312.47 (17.63%)	313.49 (17.69%)
		57: www.dcn.org	915.14	558.84 (61.07%)	553.32 (60.46%)	553.94 (60.53%)	553.75 (60.51%)
Dallas, TX	Dallas News	58: www.dallasnews.com	3635.06	160.03 (4.4%)	163.99 (4.51%)	165.20 (4.54%)	167.30 (4.6%)
		59: www.cowboysplus.com	3371.22	442.57 (13.13%)	440.91 (13.08%)	445.07 (13.2%)	443.65 (13.16%)
		60: www.guidelive.com	3371.00	419.51 (12.44%)	442.59 (13.13%)	446.68 (13.25%)	449.59 (13.34%)
	City of Irving	61: www.ci.irving.tx.us	42.02	36.21 (86.17%)	36.72 (87.39%)	37.16 (88.43%)	38.56 (91.78%)
	Online Video	62: www.lapdonline.org	2853.44	461.83 (16.19%)	531.40 (18.62%)	564.30 (19.78%)	618.81 (21.69%)

Table A.15: Connection Ratings Prediction Results with Different Collection Intervals when $\lambda = 0.5$

location	cluster name	web servers	average actual values	interval=10min mean prediction errors	interval=1hr mean prediction errors	interval=2hr mean prediction errors	interval=4hr mean prediction errors
Boston, MA	Boston Globe	1: www.boston.com	0.94	0.16 (16.59%)	0.16 (17.46%)	0.17 (18.18%)	0.17 (18.65%)
		2: weather.boston.com	1.06	0.23 (21.4%)	0.23 (21.4%)	0.23 (22.01%)	0.24 (22.37%)
		3: www.explorenewengland.com	1.27	0.42 (32.82%)	0.41 (32.73%)	0.42 (33.54%)	0.42 (33.11%)
	MBTA	4: www.mbta.com	0.92	0.40 (43.35%)	0.43 (46.27%)	0.45 (48.91%)	0.47 (50.54%)
		5: trip.mbta.com	1.11	0.38 (34.29%)	0.40 (35.81%)	0.40 (35.97%)	0.41 (36.97%)
	Web Hosting	6: www.aviationdisasterlawyers.com	1.18	0.30 (25.66%)	0.30 (25.79%)	0.31 (26%)	0.30 (25.75%)
		7: www.asbestoslaw.info	1.15	0.27 (23.23%)	0.28 (23.97%)	0.28 (24.17%)	0.28 (24.24%)
		8: www.pharmaceuticallawyers.com	1.16	0.28 (24.14%)	0.28 (24.52%)	0.28 (23.85%)	0.28 (24.26%)
New York, NY	NYTimes	9: www.nytimes.com	1.45	0.53 (36.7%)	0.53 (36.37%)	0.54 (37.33%)	0.55 (37.87%)
		10: movies.nytimes.com	1.32	0.47 (35.37%)	0.46 (35.26%)	0.47 (36.03%)	0.48 (36.42%)
		11: homefinance.nytimes.com	1.38	0.42 (30.6%)	0.44 (31.64%)	0.44 (31.64%)	0.44 (31.74%)
		12: query.nytimes.com	1.43	0.49 (34.24%)	0.48 (33.65%)	0.49 (34.43%)	0.49 (34.6%)
	UN	13: www.undp.org	1.32	0.44 (33.48%)	0.44 (33.49%)	0.45 (33.87%)	0.45 (33.97%)
		14: www.rbas.undp.org	1.32	0.41 (30.71%)	0.41 (30.9%)	0.43 (32.21%)	0.43 (32.74%)
		15: www.dz.undp.org	1.46	0.51 (35.09%)	0.52 (35.53%)	0.52 (35.77%)	0.52 (35.91%)
		16: google.undp.org	1.39	0.47 (33.86%)	0.48 (34.42%)	0.48 (34.62%)	0.49 (35.39%)
Atlanta, GA	CNN	17: www.cnn.com	1.96	0.07 (3.58%)	0.08 (4.09%)	0.08 (4.17%)	0.08 (4.24%)
		18: edition.cnn.com	1.93	0.11 (5.68%)	0.12 (6.17%)	0.12 (6.34%)	0.13 (6.51%)
		19: si.cnn.com	1.94	0.10 (5.14%)	0.11 (5.52%)	0.11 (5.65%)	0.10 (5.4%)
		20: money.cnn.com	1.74	0.10 (5.67%)	0.11 (6.31%)	0.12 (6.65%)	0.13 (7.21%)
	Weather.com	21: www.weather.com	1.88	0.18 (9.34%)	0.20 (10.77%)	0.20 (10.82%)	0.21 (11.31%)
		22: forgetaway.weather.com	1.87	0.22 (11.88%)	0.23 (12.34%)	0.23 (12.17%)	0.23 (12.31%)
		23: desktopfw.weather.com	1.85	0.25 (13.38%)	0.26 (14.01%)	0.26 (14.2%)	0.26 (14.21%)
		24: br.weather.com	1.85	0.25 (13.72%)	0.25 (13.66%)	0.25 (13.78%)	0.26 (13.97%)
	GA gov	25: www.georgia.gov	0.94	0.23 (24.68%)	0.26 (27.83%)	0.27 (28.36%)	0.27 (28.58%)
		26: www.files.georgia.gov	1.88	0.21 (11.04%)	0.21 (11.12%)	0.22 (11.44%)	0.22 (11.72%)
		27: oca.awe.gta.ga.gov	1.87	0.22 (11.79%)	0.23 (12.11%)	0.23 (12.53%)	0.24 (12.61%)
		28: www.gov.state.ga.us	1.62	0.54 (33.34%)	0.56 (34.45%)	0.57 (35.26%)	0.59 (36.24%)
Springfield, IL	IL gov	29: www.dnr.state.il.us	1.85	0.26 (14.25%)	0.27 (14.42%)	0.27 (14.47%)	0.27 (14.51%)
		30: www.illinois.gov	1.85	0.12 (6.26%)	0.14 (7.45%)	0.16 (8.84%)	0.21 (11.27%)
		31: www.allkidscovered.com	1.95	0.09 (4.52%)	0.09 (4.48%)	0.09 (4.52%)	0.09 (4.5%)
	IL Education	32: www.isbe.net	1.96	0.04 (2.09%)	0.06 (3.02%)	0.07 (3.35%)	0.07 (3.69%)

Table A.16: Connection Ratings Prediction Results with Different Collection Intervals when $\lambda = 0.5$ (Continued)

location	cluster name	web servers	average actual values	interval=10min mean prediction errors	interval=1hr mean prediction errors	interval=2hr mean prediction errors	interval=4hr mean prediction errors
Raymond, WA	MSN	33: www.msn.com	1.88	0.12 (6.25%)	0.16 (8.56%)	0.19 (9.89%)	0.21 (10.89%)
		34: entertainment.msn.com	1.99	0.03 (1.29%)	0.03 (1.44%)	0.03 (1.45%)	0.03 (1.44%)
		35: music.msn.com	1.99	0.01 (0.54%)	0.01 (0.71%)	0.01 (0.71%)	0.01 (0.72%)
		36: weather.msn.com	1.99	0.02 (1.19%)	0.03 (1.41%)	0.03 (1.4%)	0.03 (1.4%)
	Real	37: www.realnetworks.com	1.65	0.23 (13.77%)	0.32 (19.59%)	0.43 (25.93%)	0.54 (32.92%)
		38: brasil.real.com	1.34	0.58 (43.09%)	0.66 (49.49%)	0.72 (54.11%)	0.81 (60.48%)
		39: musicstore.real.com	1.96	0.08 (4.04%)	0.08 (4.07%)	0.08 (4.03%)	0.08 (4.1%)
Los Angeles, CA	Ameriquest	40: www.ameriquestmortgage.com	1.52	0.60 (39.04%)	0.60 (39.46%)	0.61 (40.19%)	0.61 (39.99%)
		41: careers.ameriquest.com	0.86	0.98 (113.74%)	0.98 (113.43%)	0.98 (113.55%)	0.97 (112.18%)
		42: www.ameriquestracing.com	0.44	0.69 (156.52%)	0.69 (155.86%)	0.68 (154.47%)	0.69 (156.67%)
	City of LA	43: www.lacity.org	1.21	0.59 (48.87%)	0.59 (49.22%)	0.61 (50.63%)	0.62 (51.13%)
		44: eng.lacity.org	1.73	0.31 (17.63%)	0.35 (20.1%)	0.40 (22.83%)	0.44 (25.47%)
		45: publiccsd.lacity.org	1.81	0.24 (13.37%)	0.28 (15.19%)	0.31 (16.93%)	0.34 (18.57%)
		46: parc1.lacity.org	1.05	0.43 (41.28%)	0.47 (44.59%)	0.50 (47.89%)	0.53 (50.7%)
		47: www.griffithobservatory.org	1.79	0.28 (15.73%)	0.31 (17.21%)	0.33 (18.51%)	0.36 (20.03%)
San Francisco, CA	Sanfrancisco	48: sanfrancisco.com	1.72	0.42 (24.76%)	0.45 (25.95%)	0.45 (26.31%)	0.46 (26.76%)
		49: www.santa-clara.com	1.82	0.32 (17.64%)	0.32 (17.57%)	0.32 (17.72%)	0.33 (18.2%)
		50: www.santacruz.com	1.11	0.27 (24.12%)	0.27 (24.22%)	0.27 (23.98%)	0.26 (23.81%)
		51: www.oakland.com	1.66	0.50 (30.06%)	0.51 (30.96%)	0.52 (31.44%)	0.54 (32.29%)
	CA gov	52: democrats.assembly.ca.gov	1.93	0.12 (6%)	0.12 (6.08%)	0.12 (6.03%)	0.12 (6.06%)
		53: www.legislature.ca.gov	1.93	0.13 (6.84%)	0.13 (6.86%)	0.13 (6.86%)	0.13 (6.87%)
		54: republican.assembly.ca.gov	1.91	0.15 (7.89%)	0.15 (8.05%)	0.16 (8.29%)	0.16 (8.28%)
	City of Davis	55: www.city.davis.ca.us	1.83	0.30 (16.31%)	0.30 (16.47%)	0.31 (16.97%)	0.31 (16.96%)
		56: events.dcn.org	1.92	0.14 (7.13%)	0.14 (7.4%)	0.14 (7.49%)	0.15 (7.77%)
		57: www.dcn.org	0.95	0.09 (9.75%)	0.10 (10.02%)	0.10 (10.03%)	0.09 (9.9%)
Dallas, TX	Dallas News	58: www.dallasnews.com	1.83	0.24 (13.33%)	0.28 (15.34%)	0.29 (16.01%)	0.31 (16.9%)
		59: www.cowboysplus.com	1.96	0.08 (4.07%)	0.08 (4.27%)	0.08 (4.27%)	0.08 (4.17%)
		60: www.guidelive.com	1.88	0.15 (8.15%)	0.19 (9.94%)	0.20 (10.53%)	0.21 (11.36%)
	City of Irving	61: www.ci.irving.tx.us	1.64	0.30 (18.12%)	0.33 (20.24%)	0.35 (21.37%)	0.40 (24.14%)
	Online Video	62: www.lapdonline.org	1.78	0.37 (20.82%)	0.37 (20.96%)	0.37 (20.72%)	0.38 (21.24%)

Bibliography

[1] AntiOnline IP Locator. http://www.antionline.com/tools-and-toys/ip-locate.

[2] Geobytes IP Locator. http://www.geobytes.com/IpLocator.htm.

[3] tcptrace. http://www.tcptrace.org.

[4] Anat Bremler-Barr, Edith Cohen, Haim Kaplan, and Yishay Mansour. Predicting and Bypassing End-to-end Internet Service Degradations. In *ACM SIGCOMM Workshop on Internet Measurment*, Marseille, France, November 2002.

[5] Neal Cardwell, Stefan Savage, and Thomas Anderson. Modeling TCP Latency. In *IEEE Infocom*, Tel Aviv, Israel, March 2000.

[6] Mukul Goyal, Roch Guerin, and Raju Rajan. Predicting TCP Throughput From Non-invasive Network Sampling. In *IEEE Infocom*, New York, NY, U.S., June 2002.

[7] Manish Jain and Constantinos Dovrolis. End-to-end Available Bandwidth: Measurement Methodology, Dynamics, and Relation with TCP Throughput. In *ACM SIGCOMM*, Pittsburgh, PA, U.S., August 2002.

[8] Sharad Jaiswal, Gianluca Iannaccone, Christophe Diot, Jim Kurose, and Don Towsley. Inferring TCP Connection Characteristics Through Passive Measurements. In *IEEE Infocom*, Hongkong, China, March 2004.

[9] Hao Jiang and Constantinos Dovrolis. Passive Estimation of TCP Round-Trip Times. *ACM SIGCOMM Computer Communication Review*, 32(3):75–88, July 2002.

[10] Jitendra Padhye, Victor Firoiu, Don Towsley, and Jim Kurose. Modeling TCP Throughput: A Simple Model and its Empirical Validation. *ACM SIGCOMM Computer Communication Review*, 28(4):303–314, October 1998.

[11] Jitendra Pahdye and Sally Floyd. On Inferring TCP Behavior. In *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, San Diego, CA, U.S., August 2001. ACM.

[12] Aimin Sang and San qi Li. A Predictability Analysis of Network Traffic. In *IEEE Infocom*, Tel Aviv, Israel, March 2000.

[13] Srinivasan Seshan, Mark Stemm, and Randy H. Katz. SPAND: Shared Passive Network Performance Discovery. In *USENIX Symposium on Internet Technologies and Systems*, Monterey, CA, U.S., December 1997.

[14] Mark Stemm, Srinivasan Seshan, and Randy H. Katz. A Network Measurement Architecture for Adaptive Applications. In *IEEE Infocom*, Tel Aviv, Israel, March 2000.