



**UNIVERSITI PUTRA MALAYSIA**

**PERFORMANCE IMPROVEMENT STUDIES IN ACCESSING WEB  
DOCUMENTS**

**SAADIAH BT YAHYA**

**FSAS 1998 30**

**PERFORMANCE IMPROVEMENT STUDIES IN ACCESSING WEB  
DOCUMENTS**

**By**

**SAADIAH BT YAHYA**

**Thesis Submitted in Fulfilment of the Requirements for the  
Degree of Doctor of Philosophy in the Faculty of  
Science and Environmental Studies  
Universiti Putra Malaysia**

**August 1998**



## ACKNOWLEDGEMENTS

In the name of Allah, the Beneficent, the Merciful.

First and foremost, I thank Allah the All-Mighty for whatever I have now. It is with His ascendancy, I have completed this study. Most important, my family and I are happy and enjoy good virtues and health.

Next, I would like to express my earnest gratitude and heartfelt thanks to the chairman of my supervisory committee, Associate Professor Dr Abu Talib Othman, for his guidance, support and encouragement throughout my study. His comments and suggestions although sometimes are difficult to fulfil but were constructive. Without him, this thesis would not have been possible.

It is also a great honour and pleasure to acknowledge Dr Md Yazid Mohd Saman for pursuing the role of the chairman of supervisory committee. My gratitude also goes to Dr Ali Mamat, a member of the supervising committee and Dr (Prof.) Ramani A. K. for their support, helpful suggestions and making corrections to this thesis. To all lecturers of Computer Science Department of UPM especially Dr (Hajjah) Fatimah, my thanks for their support and guidance. To all Post Graduate



Office staff especially Pn Faridah and Pn Arbaiah, thank for all your advice and help pertaining to rules and regulations of submitting thesis.

A very special thank is due to Mr Ken-chi Chinen from NAIST Japan and Mr Nigel Smith from Hensa Unix, who have guided and encouraged me in encountering many problems pertaining to prefetching. I am indebted to a number of people for their assistance, specially Mr Chuck Neerdaels an engineering manager of Netscape Communications Corporation for his encouragement and prompt delivery of the version 2.5 proxy; Encik Azizi Ngah Tasir the Dean of the Faculty of Information Technology and Quantitative Science (FTMSK), Mara Institute of Technology (ITM), whose support is really enchanting; to Salmah Abdul Aziz whom together we explore the world of caching; Rima Kartikasari for doing lot of communications with the Netscape Communications Corporation, and other staff of the FTMSK ITM Computer Laboratory.

To all my friends, specially Noreha Hussin, Zainab Abu Bakar, Noor Laila Mohd Noor, Syed Ahmad, Adnan Ahmad (Dr), Sharifah Sakinah (Dr), Noor Habibah, Farok Azamat and Isahak Kassim (Dr), I would like to express my thanks for your ideas, help and encouragement. I really appreciate it. I am also deeply grateful to all those who use to share their experiences and problems in the Netscape's snews (netscape.server.proxy) during the period of my study; Pn Nurah



the proxy maintainer from JARING and Mr Ismail Ali the executive on network operations of TMnet Malaysia.

I must mention the unwavering support I received from my loving husband Mohd Toib is never forgotten. He stood beside me when no one wants to be there, I really treasure his sacrifice, courage and support especially on those hardship moments. To all my four children Farah Sakinah, Ashrafu Anuar, Mohamad Farhan and Amirah Diyana who endure infinite patience and understanding of their mummy's work. It is to them that I dedicate the work.

Last but not least, I would like to take this opportunity to thank ITM for generously sponsoring me the scholarship. I cherish their support and prompt payment of the allowances.



## TABLE OF CONTENT

	<b>Page</b>
<b>ACKNOWLEDGEMENT</b> .....	ii
<b>LIST OF TABLES</b> .....	xi
<b>LIST OF FIGURES</b> .....	xiv
<b>LIST OF ABBREVIATIONS</b> .....	xvi
<b>ABSTRACT</b> .....	xx
<b>ABSTRAK</b> .....	x xiii

### CHAPTER

<b>I INTRODUCTION</b> .....	<b>1</b>
BACKGROUND.....	1
OBJECTIVES OF THE STUDY .....	6
PROBLEMS STATEMENT.....	6
IMPORTANCE OF THE STUDY .....	8
OVERVIEW OF THESIS.....	10
<b>II LITERATURE REVIEW</b> .....	<b>12</b>
CACHING PROXY.....	13
<i>Corporate and Organisational Caching</i> .....	13
<i>National Caching Service</i> .....	14
<i>Hierarchical Caching</i> .....	16



<i>Push-Caching</i> .....	19
CACHE CONSISTENCY .....	21
PREFETCHING.....	23
<i>Client-initiated Prefetching</i> .....	24
<i>Server-initiated Prefetching</i> .....	24
<i>An Interactive Prefetching Proxy Server</i> .....	27
CONCLUSION.....	29
<b>III EXPERIMENTAL DESIGN AND DEVELOPMENT .....</b>	<b>32</b>
INTRODUCTION.....	32
SOFTWARE SELECTION .....	33
INSTALLING, CONFIGURING AND MANAGING THE NETSCAPE PROXY SERVER ....	35
<i>Configuring Client Browsers</i> .....	37
<i>Monitoring and Configuring the Proxy Server</i> .....	39
<i>Configuring System Setting</i> .....	39
<i>Enabling the Cache</i> .....	45
<i>Creating a Cache Working Directory</i> .....	46
<i>Recording URLs</i> .....	46
<i>Setting the Cache Size</i> .....	47
<i>Editing the Cache Capacity</i> .....	48
<i>Caching HTTP and FTP Documents</i> .....	48
MONITORING THE SERVER'S STATUS .....	52
<i>Monitor the Server using the Server Manager</i> .....	52
<i>Working with Log Files</i> .....	53
<i>Setting log preferences</i> .....	58
<i>Working with the log analyser</i> .....	58
<i>Archiving log files</i> .....	59



THE LOG FILE ANALYSIS.....	59
<i>Archiving Access Log Files from the Server Manager</i> .....	60
<i>Running the Log Analyser from the Server Manager</i> .....	61
CONCLUSION.....	63

#### **IV STRATEGIES OF REFRESHING WEB DOCUMENTS.....64**

INTRODUCTION.....	64
ILLUSTRATION OF THE PROPOSED PREFETCHING TECHNIQUES .....	65
<i>Proxy Users</i> .....	65
<i>Description of the Cache Setting</i> .....	66
<i>Prefetching Strategies</i> .....	71
<i>The Batch Update Configuration File</i> .....	72
APPLICATION OF THE PROPOSED REFRESHING TECHNIQUES .....	73
EXPLANATION OF THE OUTPUT OF NETSCAPE PROXY SERVER'S ANALYSER....	75
<i>Summary of Totals Server Statistics</i> .....	75
<i>The Most Commonly Accessed URLs</i> .....	77
<i>Client Hosts Most Often Accessing The Proxy Server</i> .....	77
<i>Status Code Report</i> .....	80
<i>Data Flow Report</i> .....	80
<i>Requests and Connections</i> .....	81
<i>Cache Performance Report</i> .....	82
<i>Transfer Time Report</i> .....	84
DESCRIPTION OF THE FURTHER ANALYSIS TREATMENT .....	85
SELECTIVE PREFETCHING UPDATE .....	86
<i>The Frequency of Access (F) Made within the Specified Interval</i> .....	87
<i>The Recency Time (R) of a Requested Document</i> .....	87
<i>Size (S) of the Document</i> .....	88





	<i>Programs for the Selective Prefetching Update</i> .....	90
	<i>The Selective Prefetching Implementation</i> .....	95
	CONCLUSION .....	98
<b>V</b>	<b>STATISTICAL ANALYSIS AND CACHE PERFORMANCE.....</b>	<b>100</b>
	INTRODUCTION.....	100
	STATISTICAL ANALYSIS OF THE PROXY SERVER.....	101
	<i>Hourly Activity</i> .....	102
	<i>Requests and Connections</i> .....	104
	<i>Cache Performance</i> .....	106
	<i>Total Kilobytes Transferred</i> .....	110
	<i>The Not-Modified (304) Response Code</i> .....	112
	ANALYSIS OF VARIANCE OF CACHE PERFORMANCES .....	113
	<i>Avoided Remote Connections</i> .....	115
	<i>Average Transfer Time Improvement</i> .....	120
	<i>Average Transfer Time with Caching without Errors</i> .....	124
	<i>Average Transaction Time</i> .....	128
	<i>Total Proxy Hits</i> .....	136
	<i>Hourly Activity</i> .....	140
	CONCLUSION.....	142
<b>VI</b>	<b>CONCLUSIONS AND RECOMMENDATIONS .....</b>	<b>145</b>
	CONCLUSIONS .....	145
	FUTURE RESEARCH.....	151
	<b>BIBLIOGRAPHY: .....</b>	<b>154</b>



<b>APPENDIX .....</b>	<b>161</b>
AI	The Distribution of Users for the Proxy Server 2.0 .....162
AII	The Distribution of Additional Users for the Proxy Server 2.5 .....163
B	Common and Extended Log File Format ..... 164
C	The Copy of The Posted Email to Proxy Users ..... 167
D	The analysis Report for access.28Aug-07Am File ..... 170
E	Makefile Codes for Adapting the Server to the Webcopy ..... 178
F	Usage of Webcopy ..... 180
GI	Statistical Analysis for the Six Months ..... 183
GII	Statistical Analysis for the Four Caching ..... 184
HI	Variables selected for the M1 ..... 192
HII	Variables Selected for the M2 ..... 193
HIII	Variables Selected for the M3 ..... 194
HIV	Variables Selected for the M4..... 195
I	Four Groups Anova Results..... 196
JI	LSD Approach for %Avoided Remote Connections..... 197
JII	LSD Approach for % Average Transfer Time Improvement..... 198
JIII	LSD Approach for % Average Transfer with Caching without Error..... 199
JIV	LSD Approach for Average Transaction Time ..... 200
JV	LSD Approach for % Not Modified Responses (304's) ..... 201
JVI	LSD Approach for % Total Proxy Hits ..... 202
KI	Scheffe 's Test for %Avoided Remote Connections..... 203
KII	Scheffe 's Test for % Average Transfer Time Improvement..... 204
KIII	Scheffe 's Test for % Average Transfer with Caching without Error..... 205
KIV	Scheffe 's Test for Average Transaction Time ..... 206
KV	Scheffe 's Test for % Not-Modified Responses (304's)..... 207



**KVI Scheffe 's Test for % Total Proxy Hits ..... 208**

**BIBLIOGRAPHICAL SKETCH..... 209**



## LIST OF TABLES

Table	Page
1 Reserved Port Number for a Specific Service .....	38
2 Suggested Number of Processes Based on Average Request Service Time and Number of Requests .....	42
3 Using the Cache Expiration and Cache Refresh settings with HTTP .....	49
4 Description of Extended Log File Sample .....	57
5 The Proposed Caching Strategies .....	70
6 Random Schedule of the Caching Policies .....	71
7 Totals of Server Statistics .....	76
8 Five Most Commonly Accessed URL's for the Day .....	77
9 The Top Five Client Hosts Accessing the Server .....	78
10 Distribution by Service and Percentage Finished Time .....	79
11 Status Not Available Code .....	80
12 Data Flow Descriptions .....	81
13 Requests and Connection Statistics .....	82
14 Cache Performance .....	83
15 Transfer Time .....	85
16 Selected Performance Metrics .....	86
17 Example of Executable File Invocation .....	96
18 The Cron Tab File for Automating the Selective Refreshing .....	97
19 The Name for the Six Archived Log Files .....	101
20 Requests in the Most Active Hour .....	103
21 Requests in Second at the Most Active Hour .....	104
22 Requests and Connections Activity .....	105



23	Cache Hits .....	107
24	Percentage Cache Hits .....	109
25	Total Kilobytes Transferred .....	110
26	304's Responses Against the Total Hits .....	112
27	Percentage Avoided Remote Connections .....	116
28	Four Groups ANOVA Results: Percentage Avoided Remote Connections by Policy of Caching .....	119
29	Scheffe Test of Group Differences with Percentage Avoided Remote Connections .....	120
30	Percentage Average Transfer Time Improvement (second per request) .....	121
31	Four Groups ANOVA Results: % Average Transfer Time Improvement by Policy of Caching .....	123
32	Scheffe Test of Group Differences with Percentage Average Transfer Time Improvement .....	124
33	Percentage Average Transfer with Caching without Error .....	125
34	Four Groups ANOVA Results: Percentage Average Transfer Time Caching without Error by Policy of Caching .....	127
35	Scheffe Test of Group Differences with Percentage Average Transfer with caching without Errors .....	128
36	Average Transaction Time (second per request) .....	129
37	Four Groups ANOVA Results: Average Transaction Time by Policy of Caching .....	131
38	Scheffe Test of Group Differences with Average Transaction Time .....	132
39	Percentage of Not-Modified (304's) Responses .....	133
40	Four Groups ANOVA Results: Percentage of Not-Modified (304's) Responses by Policy of Caching .....	135
41	Scheffe Test of Group Differences with the Not-Modified (304's) Responses ..	136



42	Percentage Total Proxy Hits .....	137
43	Four Groups ANOVA Results: Percentage Total Proxy Hits by Policy of Caching .....	139
44	Scheffe Test of Group Differences with Percentage Total Proxy Hits .....	139



## LIST OF FIGURES

Figure	Page
1 From the Option bar, select Network Preferences and Proxies. ....	38
2 Only Proxying FTP and HTTP at Port 8080. No Proxy for Local Documents.....	38
3 System Specifics Form of the Administration Server .....	41
4 A Sample of an Error Log File .....	55
5 A sample of an Access Log File in the Extended Format .....	56
6 The Batch Refreshing Program .....	72
7 Segment of Transaction of Access File for 12/8/97 .....	74
8 A Histogram of Requests by Service and Percentage Finished Time .....	79
9 The Web Pages and the Associated In-line Images and Page References .....	93
10 Selective Batch Prefetching Update Programs and Interfaces.....	95
11 The Flow Diagram of Automated Selective Batch Prefetching Update.....	98
12 Graphical Representation of Requests in the Most Active Hour .....	103
13 Graphical Representation of Requests in Second at Most Active Hour .....	104
14 Graphical Representation of Requests and Connections .....	106
15 Graphical Representation of Cache Hits.....	108
16 Graphical Representation of Percentage Cache Hits.....	109
17 Graphical Representation of Total Kilobytes Transferred.....	111
18 Graphical Representation of 304's Responses Against the Total Hits.....	113
19 Percentage Avoided Remote Connections.....	117
20 Boxplots of Percentage Avoided Remote Connections for Four Groups of ANOVA.....	118
21 Percentage Average Transfer Time Improvement .....	121



22	Boxplots of Percentage Average Transfer Time Improvement for Four Groups of ANOVA.....	122
23	Average Transfer Time with Caching without Error .....	125
24	Boxplots of Percentage Average Transfer Time with Caching without Error for Four Groups of ANOVA.....	126
25	Average Transaction Time.....	130
26	Boxplots of Average Transaction Time for Four Groups of ANOVA .....	130
27	Percentage of Not-Modified (304's) Responses .....	133
28	Boxplots of Percentage Not-Modified (304's) Responses for Four Groups of ANOVA.....	134
29	Percentage Total Proxy Hits .....	137
30	Boxplots of Percentage Total Proxy Hits for Four Groups of ANOVA .....	138
31	Number of Request During Most Active Hour .....	140
32	Number of Requests at Most Active Second of the Most Active Hour .....	141





## LIST OF ABBREVIATIONS

<b>AFS</b>	-	Andrew File System
<b>ATM</b>	-	Asynchronous Transfer Mode
<b>CERN</b>	-	Centre Europeenne puorla Recherché Nucleaire (European Centre for Nuclear Research)
<b>CGI</b>	-	Common Gateway Interface
<b>COINS</b>	-	COrporate INformation Superhighway
<b>DEC</b>	-	Digital Equipment Corporation
<b>DNS</b>	-	Domain Name Server
<b>Email</b>	-	Electronic Mail
<b>EMIS</b>	-	European Mathematical Information Service
<b>ERCIM</b>	-	European Research Consortium for Informatics and Mathematics
<b>FTMSK</b>	-	Fakulti Teknologi Maklumat dan Sains Kuantitatif (Faculty of Information Technology and Quantitative Science)
<b>Hensa</b>	-	Higher Education National Software Archive
<b>FTP</b>	-	File Transfer Protocol
<b>GB</b>	-	Giga Byte
<b>Gbps</b>	-	Giga Bit per Second
<b>GIF</b>	-	Graphics Interchange Format
<b>HTML</b>	-	HyperText Markup Language



<b>HTTP</b>	-	HyperText Transfer Protocol
<b>HTTPD</b>	-	HyperText Transfer Protocol Daemon
<b>HTTPS</b>	-	HyperText Transfer Protocol Secured ( A secure version of HTTP)
<b>I/O</b>	-	Input/Output
<b>IP</b>	-	Internet Protocol
<b>ISP</b>	-	Internet Service Provider
<b>IT</b>	-	Information Technology
<b>ITM</b>	-	Institut Teknologi Mara (Mara Institute of Technology)
<b>JARING</b>	-	Joint Advanced Research Integrated NetworkinG
<b>Km</b>	-	Kilometer
<b>Kbps</b>	-	Kilo bit per second
<b>LAN</b>	-	Local Area Network
<b>MB</b>	-	Mega Bytes
<b>Mbps</b>	-	Mega bit per second
<b>MD5</b>	-	Message Digest 5
<b>MIME</b>	-	Multi-Purpose Internet Mail Extensions
<b>MIMOS</b>	-	Malaysian Institute of Microelectronics Systems
<b>MSC</b>	-	Multimedia Super Corridor
<b>NAIST</b>	-	Nara Institute of Technology
<b>NCSA</b>	-	National Computer System Association



<b>NFS</b>	-	Network File System
<b>NIS</b>	-	Network Information Service
<b>NLANR</b>	-	National Laboratory for Advanced Networking Research
<b>NSFNET</b>	-	National Science Foundation NETwork
<b>PC</b>	-	Personal Computer
<b>Perl</b>	-	Practical Extraction and Report Language
<b>RAM</b>	-	Random Access Memory
<b>SGML</b>	-	Standard Generic Markup Language
<b>SNMP</b>	-	Simple Network Management Protocol
<b>SPM</b>	-	Selective Prefetching Metric
<b>SSL</b>	-	Secure Sockets Layer
<b>TCP</b>	-	Transmission Control Protocol
<b>TCP/IP</b>	-	Transfer Control Protocol/Internet Protocol
<b>TM</b>	-	Telekom Malaysia
<b>TMnet</b>	-	Telekom Malaysia computer network
<b>TTL</b>	-	Time To Live
<b>URL</b>	-	Uniform Resource Locator
<b>UK</b>	-	United Kingdom
<b>UKC</b>	-	University of Kent Canterbury
<b>UPM</b>	-	Universiti Putra Malaysia
<b>USA</b>	-	United State of America



<b>WAIS</b>	-	<b>Wide Area Information Service</b>
<b>WAN</b>	-	<b>Wide Area Network</b>
<b>Wcol</b>	-	<b>WWW collector</b>
<b>WWW/Web/W3</b>	-	<b>World Wide Web</b>



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

**PERFORMANCE IMPROVEMENT STUDY IN ACCESSING WEB DOCUMENTS**

**By**

**SAADIAH BTE YAHYA**

**May 1998**

**Chairman: Ass. Prof Dr Abu Talib Othman**

**Faculty: Science and Environmental Studies**

As the World Wide Web has now become the standard interface for interactive information services over the Internet, the perceived latency in WWW interaction is becoming an important and crucial issue. Currently, Web users often experience response delay of several seconds or even longer to non-local Web sites especially when the pages they attempt to access are very popular. For WWW to be acceptable for general daily use, the response delay must be reduced.

The potential solutions to the problem lie in the extensive use of caching (disk based) and prefetching in WWW. Both caching and prefetching explore the patterns and knowledge in the Web accesses.



This thesis describes and tests the efficiency of a batch prefetching update (refreshing) in accessing HTTP and FTP documents on the global Internet. The update is scheduled to run at idle time when the traffic is less congested and the server activity is low. The batch refreshing effort would be fruitful when the refreshed documents are really requested before they turn stale again. The effectiveness of the batch refreshing is verified by running a statistical analysis of the access log files.

In the first part of the study, a Proxy Server at the LAN of FTMSK, ITM was set-up, configured and monitored for the use of 400 users. Access log files are collected and analysed for a period of six months. The analysis result would be a benchmark for the caching proxy with batch refreshing in the second part of the work.

The following areas are addressed: The maintenance of up-to-date cache data with minimal network overhead; The design of refreshing policies; The proposed algorithms and programs for the *Selective* prefetching update; and the analysis of variance to determine performance's improvement.

From the statistical analysis of access log files, it was found that: Cache performances would improve by increasing the refreshing interval of the cached



documents; and additional batch refreshing treatments could not further enhance performances.



Abstrak tesis dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk Ijazah Falsafah Kedoktoran.

## **KAJIAN PENINGKATAN PRESTASI BAGI CAPAIAN DOKUMEN WEB**

**Oleh**

**SAADIAH BTE YAHYA**

**Mei 1998**

**Pengerusi: Prof Madya Dr Abu Talib Othman**

**Faculty: Sains dan Alam Sekitar**

Apabila World-Wide Web menjadi antara-muka piawai bagi perkhidmatan maklumat interaktif Internet, *latency* yang dialami oleh interaksi WWW telah menjadi satu isu yang penting. Kini pengguna Web sering mengalami kelengahan sambutan beberapa saat atau mungkin lebih lama apabila menghubungi Web yang bukan berada di rangkaian setempat, lebih-lebih lagi sekiranya halaman yang mereka cuba mencapai adalah terlalu popular. Kelengahan sambutan perlu dikurangkan sebaiknya bagi menjamin WWW dapat diterima dalam penggunaan am harian.

Penggunaan cache (berasaskan cakera) dan prapapaian WWW secara meluas adalah penyelesaian kepada masalah yang telah dibincangkan. Cache dan prapapaian keduanya menggunakan corak dan pengetahuan capaian Web.





Tesis ini menerang dan menguji keberkesanan kemaskini pracaapaian berkelompok (penyegaran) bagi capaian dokumen HTTP dan FTP pada Internet secara global. Kemaskinian dikelompok untuk beroperasi pada waktu melahu ketika trafik kurang berpusu dan aktiviti pelayan adalah rendah. Usaha penyegaran berkelompok akan bermakna apabila halaman yang disegarkan benar-benar diminta sebelum ia menjadi basi semula. Keberkesanan penyegaran berkelompok akan ditentukan dengan ujian analisis statistik pada fail log capaian.

Di bahagian pertama penyelidikan, satu pelayan proxy pada LAN telah dibangunkan di FTMSK ITM, dikonfigurasi dan dimonitor untuk kegunaan 400 pengguna. Fail log capaian telah dikumpul dan dianalisis bagi tempoh enam bulan. Hasil analisis akan menjadi ukuran kepada proxy cache dengan penyegaran berkelompok di bahagian kedua penyelidikan.

Perkara berikut telah diberikan tumpuan: Penyelenggaraan kemas-kinian data cache dengan penggunaan rangkaian yang minima; Penciptaan polisi penyegaran; Pencadangan algoritma dan aturcara pracaapaian kemaskinian *berpilih*; dan penganalisan varian untuk menentukan peningkatan prestasi.

Dari analisis statistik fail log capaian, telah didapati: Prestasi cache boleh ditingkatkan hanya dengan menambah tempoh kesegaran dokumen; dan rawatan