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 commitee. 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, Ashraful 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

## LIST OF TABLES

**Table** **Page**

## LIST OF FIGURES

**Figure** **Page**

## LIST OF ABBREVIATIONS

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

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

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

| | | |
|---|---|---|
| **WAIS** | - | Wide Area Information Service |
| **WAN** | - | Wide Area Network |
| **Wcol** | - | WWW collector |
| **WWW/Web/W3** | - | World Wide Web |

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 pracapaian WWW secara meluas adalah penyelesaian kepada masalah yang telah dibincangkan. Cache dan pracapaian keduanya menggunakan corak dan pengetahuan capaian Web.

Tesis ini menerang dan menguji keberkesanan kemaskini pracapaian 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 pracapaian kemaskinian *berpilih*; dan penganalisisan 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