American Journal of Networks and Communications
2015; 4(3): 37-43
Published online April 24, 2015 (http://www.sciencepublishinggroup.com/j/ajnc)
doi: 10.11648/j.ajnc.20150403.13
ISSN: 2326-893X (Print); ISSN: 2326-8964 (Online)

Performance Analysis of Hybrid Web Caching Architecture

Ho Khanh Lam[1], Nguyen Xuan Truong[2]

[1]Faculty of Information Technology, Hung Yen University of Technology and Education, Hung Yen, Vietnam
[2]Training Department, Hung Yen University of Technology and Education, Hung Yen, Vietnam

Email addresses:
lamhokhanh@gmail.com (H. K. Lam), truongutehy@gmail.com (N. X. Truong)

To cite this article:
Ho Khanh Lam, Nguyen Xuan Truong. Performance Analysis of Hybrid Web Caching Architecture. *American Journal of Networks and Communications.* Vol. 4, No. 3, 2015, pp. 37-43. doi: 10.11648/j.ajnc.20150403.13

Abstract: The development of next-generation wireless networks combine with the radio network techniques which use technology such as GSM, GPRS, 3G (UMTS, CDMA2000), LTE, WLAN, and WiMAX. It requires the construction and expansion in time of high-speed telecommunication channels for the level of the internet service provider (ISP). In fact, in many developing countries and in Vietnam, the investment rate increased bandwidth capacity can not enough for the demand use the Internet as economic issues, investment procedures. Web caching architecture is one of the effective solutions to save bandwidth while ensuring to satisfy strong demand for internet access. Hybrid web caching architecture (hybrid web caching architecture) is a solution that is used by networks because it takes advantage of the strengths of the web caching architecture stratification and dispersion, reducing connection time and transmission time, helping internet service providers to plan and save network resources at each level in an optimal way. This paper propose a novel produce to hybrid web caching architecture based on the determined time at each level of web-winning network and web time overall winner of the ISP network with n-level network.

Keywords: Hybrid Web Caching Architecture, Performance Analysis

1. Introduction

The architecture of the Internet service provider (ISP) usually organized into four levels: access networks, Institutional networks, regional networks, national backbone, and International backbone. The layer up will have bandwidth communication greater. Institutional networks is access networks which are organized according to the POP point local (poit of presence). it include technologies, telecommunications equipment high speed and allows users to connect to the Internet through the ISP of them. An POP point has some unified address and a set of IP addresses for accessing the Internet from the user (the client). A fact a POP of ISP can stay inside the house telecommunications networks. Where will design and construction of POP and how much bandwidth should be based on the standard of living, population and literacy, the focus of the economic base of schools, etc .. and requires more detailed cost of telecommunication channel capacity, the capacity of access equipment (servers, routers, etc ..).

More than 80 percent of the traffic using the user's web access, via web ISPs provide multimedia services require high-speed, latency and ensure high quality of service (QoS). Thus, the ISP's POP from the outset to ensure the circulation of access for end users from the private LAN, or from the radio access network. Therefore, the solution to improve the performance of web services: web access latency reduction for the client reduce bandwidth costs, it has web caching architecture suitable for ISP network architecture. Based on internet architecture, there are three types of web caching architecture that most ISPs apply, such as: hierarchical web caching architecture, distributed web caching and hybrid web caching. Stratified web caching architecture allows requests from terminal users from network routers low level (level access network) to the higher-level network: Institutional Network, Regional network, national network of the ISP network, if all of the network-level ISP networks are no web content which client requests are required to be transferred to the international Internet. So this is the worst case, it has the largest network latency for a web content which is required by client. Also this architecture for cache hit rate is not high at each network level, but it requires large bandwidth among the network levels. Architecture distributed web caching system is ensure for the peer web cache associated with each other in the network level, so it will ensuring a high rate of secondary web network at each level, and thus saving bandwidth between levels network. However, this architecture requires a large investment costs for the system-level web cache in each

network: bandwidth and the web server.

Web caching system is a hybrid solution which usually used by ISP. it combines two types of architectural stratification and distribution. At each level of the network perform distributed caching web architecture, but not all nodes have the web cache system, because the reasons is ensuring cost savings, but only the nodes are high bandwidth requirements by there are large numbers of citizens using the Internet. And all web cache system such links in a peer to peer networks to increase the level of use of web caching system at each level of the network. Web cache architecture is ensure interconnect for these web cache system, so it is hit rate when request of the client moved up on the network layer and also save bandwidth between the network layer. Figure 1 is a diagram of the architecture of the hybrid web-based caching associated with 4 main level. Highest level of entire web caching architecture that combines are web caching system center (level 1) of the national Internet backbone network, ISP CC (Central Cache). At the ISP Regional Network system has the Web cache area (level 2), Regional Cache. Next Level (Level 3) is the web cache system of local networks, Institutional Cache. The client network-level access to the proxy server. We connect with the local access network. The telecommunications provinces buttons POP(Point of Presence Network) is the local node access the Internet. In the POP put the web cache system, IC. The client can be a PC, a mobile phone, directly or through LAN, connected to the Internet via POP in access networks such as ADSL, mobile network. The POP associated with the high-speed transmission with regional networks.

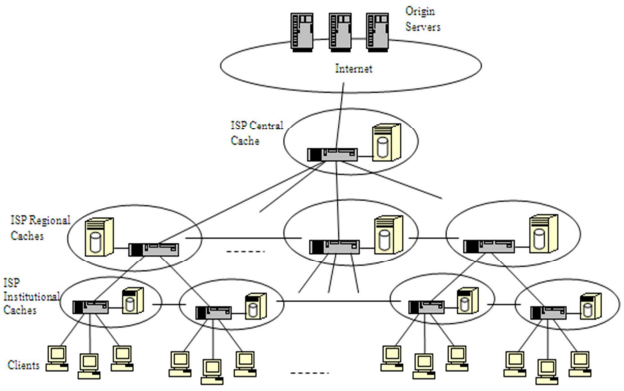

Figure 1. *Architecture of the Internet web caching stratified.*

The web browser client can directly or indirectly through local LAN proxy server sends requests for status online web access to the local network. In the case of a local proxy server does not have the content of the web page request, the proxy server forwards requests from the client to the IC system in the local POP. If the IC system has content client requests (winning IC), the IC transfer will transfer content requirements for the client (also local Proxy server saves the contents of this web page). In case of winning the IC hit time to the requirements of the client (or latency response) is the smallest. When IC miss means that the website content which the client requires not available in the system IC, so the IC system requirements are transferred to the network-level up network - area network.

At the local network, if the system requirements RC hit, it will transfers the contents to the System IC and IC transition forward to Proxy server and client. If the RC miss request of the client is transferred to the CC system of national networks. If the CC has system contents, CC hit is transferred to the RC system, and IC systems, and to the Proxy server, client. If the CC miss requests of the client is transferred to the international Internet, to web server root.

2. The Study of the Performance of Web Caching Architecture on the Internet

The author Pablo Rodriguez, Christian Spanner, ... [1] gave the model of the web caching architecture stratification and distribution, and analyze the performance of this architecture with connection time, transmission time, delay, and cache hit rate based on the theory of Markov queuing model M/D/1. However, in this study the authors suggest that the average connection time at the network layer is the same and only depend on the network layer. This is incorrect because the actual bandwidth in each different network layer, data transfer speeds are different, different delay systems and web cache different capacity, different level combination (cooperation) and in the lowest layer network - network access layer also depends very much on the access network technology, the concentration of population in the local node and the terminal network.

The author Guangwei Bai and Carey Williamson [2] analyzed the load characteristics of web caching architecture stratification, or those of the authors Balamash Abdullah and Marwan Krunz [3] when analyzing system for caching Web traffic. The study evaluated the performance of web caching system of the authors in the paper [4] [5] [6] [7] [8] [9] showed that at each level of the network must be investigated individual assessment by the difference in traffic. Queues and Markov chains are commonly used to evaluate the analytical performance of web caching architecture, web proxy server.

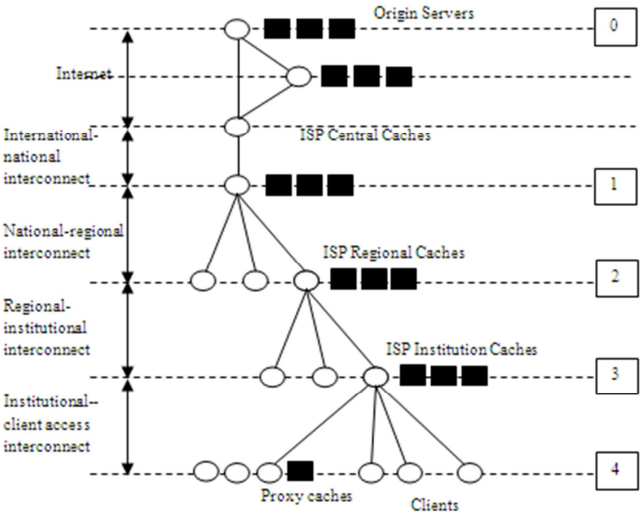

Figure 2. *Model tree of web caching architecture stratified.*

3. Solution Queue Model

Based on the research of the authors in the paper [1], in this proposed model tree diagram of web caching hybrid architecture (Figure 2). Layer 4: the network layer of the user terminal (the proxy server of the LAN, the client separately), Layer 3: local access network (radio access networks, ADSL, the local POP with the Institutional caches), layer 2 network with the regional caches, layer 1 national network with the central caches system , layer 0: international Internet to the origin servers. The highlight rectangle at each network layer represents the peer to peer (P2P) web caching systems.

The hybrid caching architecture makes web access latency of the content of the web client and it was hit ratio decreases if the web hit is high in each network layer. But if the web miss is occurs in multiple network layer the time delay will the greater for access web. To analyze the performance of web caching hybrid system, different methods of analysis of previous studies.

Because each network layer, the web server link peer to peer, so each web caching system at each level network can be considered multi-server connections in parallel and is represented by a model of queuing system type M/M/m/q with both loss and delay. Given the HTTP requests from web caching system independent of the client to each other, and the number of client HTTP requests generated unrestricted. The time between HTTP requests to the system and the service life of the system with exponential distribution. Web caching system has m the same server m = 1, 2, …. We have the capacity cache is limited by the model as the queue to receive HTTP requests with the length q.

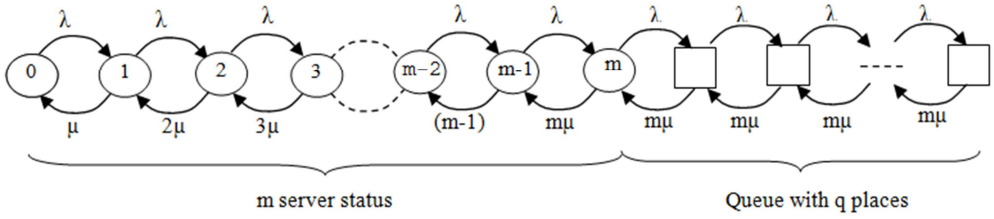

Figure 3. *Graph CTMC state's web cache system M/M/m/q.*

Because the transmission delay due to network environment at each level different networks so we assume that the average arrival rate of HTTP requests in each level is $\lambda_i; i = 0,1,2,...,n$ where n is the network level, and the average speed service of web caching system(server) at each level of the network is $\mu_i;\ i = 0,1,2,...,n$. The level of use of each web server is:

$$U = \frac{\lambda_i}{\mu_i}$$

If web caching system in each network layer consists of m parallel server connection and has q positon in queue, we have state of the CTMC shown in figure 3.

For this CTMC chain of flow balance equation is satisfied as follows:

+ where $0 \le k \le m$:

$$p_{ik} = \frac{U_i^k}{k!} p_{i0}; \forall k = 1,2,...,m \qquad (1)$$

where, p_{i0} is the probability that the client HTTP requests to the web-winning network level *i* and web cache system is idle, p_{ik} is the probability that the client HTTP requests to the web-winning web caching system and network level i is in state k (k serving HTTP requests).

+ when $m \le k$ but space in the queue from 1 to q:

$$\left.\begin{array}{l} \lambda_i p_{im} = m\mu_i p_{i(m+1)};\quad p_{i(m+1)} = \frac{U_i^m}{m!}\left(\frac{U_i}{m}\right) p_{i0} \\ \lambda_i p_{i(m+1)} = m\mu_i p_{i(m+2)}; p_{i(m+2)} = \frac{U_i^m}{m!}\left(\frac{U_i}{m}\right)^2 p_{i0} \\ \lambda_i p_{i(m+k-1)} = m\mu_i p_{i(m+k)}; p_{i(m+k)} = \frac{U_i^m}{m!}\left(\frac{U_i}{m}\right)^k p_{i0} \\ \forall k = 1,2,...,q \end{array}\right\} \Rightarrow p_{i(m+k)} = \frac{U_i^m}{m!}\left(\frac{U_i}{m}\right)^k p_{i0}$$

+ For model M/M/m/q normalize this condition must be finished, which is the sum of the probabilities must equal 1:

$$1 = \sum_{k=0}^{m+q} p_{ik} = p_{i0}\left(\sum_{k=0}^{m}\frac{U_i^k}{k!} + \frac{U_i^m}{m!}\sum_{k=1}^{q}\left(\frac{U_i}{m}\right)^k\right) = p_{i0}S;$$

$$S = \sum_{k=0}^{m}\frac{U_i^k}{k!} + \frac{U_i^m}{m!}\sum_{k=1}^{q}\left(\frac{U_i}{m}\right)^k$$

Suy ra:

$$p_{i0} = \frac{1}{S} = \left(\sum_{k=0}^{m}\frac{U_i^k}{k!} + \frac{U_i^m}{m!}\sum_{k=1}^{q}\left(\frac{U_i}{m}\right)^k\right)^{-1} \qquad (2)$$

+ When the entire m server and q positon in the queue of web caching system of tiered networks are busy, the new HTTP request to the system will be locked (not included in the queue). This is the case of congestion in the network layer i as the entire web caching system with m servers were overloaded. This state is determined by the probability of lock or probability of loss at the network level i and by $B_i = p_{i(m+q)}$:

$$B = p_{m+q} = \frac{U_i^m}{m!}\left(\frac{U_i}{m}\right)^q p_{i0} \quad (3)$$

Thus the probability of new customers must wait is the probability which the server is busy and the queue is vacancies $p_{i(m+k)}$ với $1 \le k \le q$.

Apply Zipf law and the Internet, we can determine the number of access on a large number of the local. Must have statistical forecasting residential access in each region, and based on these results build web caching systems to optimize server capacity (CPU and memory capacity) to ensure value for q is contained the maximum number of HTTP requests.

With model M/M/m/q We can be calculated performance parameters for web caching system for each network level i as follows:

1) The number of HTTP requests in the queue of the system's web-based caching i, $E[N_{iq}]$:

$$\begin{aligned} E[N_{iQ}] = \sum_{k=1}^{q} k p_{i(m+k)} = p_{i0}\frac{U_i^m}{m!}\left(\left(\frac{U_i}{m}\right) + 2\left(\frac{U_i}{m}\right)^2 + 3\left(\frac{U_i}{m}\right)^3 + ... + q\left(\frac{U_i}{m}\right)^q\right) \\ = p_{i0}\frac{U_i^m}{m!}\sum_{k=1}^{q} k\left(\frac{U_i}{m}\right)^k \end{aligned} \quad (4)$$

2) The average waiting time of an HTTP request in the system's web caching for layer network i to be serviced, $E[W_{iQ}]$ is determined by Little law:

$$E[W_{iQ}] = \frac{E[N_{iQ}]}{\lambda_i} = \frac{1}{\lambda_i}\left(p_{i0}\frac{U_i^m}{m!}\sum_{k=1}^{q} k\left(\frac{U_i}{m}\right)^k\right) \quad (5)$$

3) The average response time $E[C_i]$ web caching system in each layer network i :

This is the average time that a client's HTTP request (web content) is processed in the web cache system (including waiting time in the queue and the time serviced (web content is found):

$$E[C_i] = E[W_{iQ}] + E[S_i] = \frac{E[N_{iQ}]}{\lambda_i} + \frac{1}{\mu_i} \quad (6)$$

4) In general, if the architecture of a network web caching ISP's network, the client's HTTP request miss web in level network n, the hit web is web caching system at layer network i, where $n > i$, the response of web caching system at the layer network n for HTTP requests from the client by:

$$\begin{aligned} E[R_{WC}] = E[R_{nH}] + (Miss_n)(E[R_{(n-1)H}] + \\ (Miss_{n-2})(E[R_{(n-2)H}] + ... + (Miss_1)(E[R_{0H}])...) \end{aligned} \quad (7)$$

where, $E[R_i]-$ average response of the system's web caching of the network layer i when the web cache hit layer i; $Miss_n -$ rate cache miss in layer network n.

When cache miss in the local network (proxy server) and client cache hit at the local access network, the average response of web caching system will be:

$$E[R_{3H}] = (D_{4M} + D_{4REQ}) + E[C_3] + D_4 \quad (8)$$

where, D_{nM} - delays in the web miss layer network n; $D_{4M} -$ delay by the web miss in LAN proxy server; D_{nREQ} - Delay dependent bandwidth transmission channels which require the client HTTP transfer from layer network n to layer network $n-1$; $D_{4REQ} -$ A delay of a local proxy server to access the network via POP; D_n - delay reply of web content requires the client dependent transmission channel bandwidth from layer network $n-1$ to layer network n, and depending the size of web content; $D_4 -$ Delay reply of web content requests from the client caches intstitutional dependent transmission channel bandwidth from layer network $n-1$ to layer network n;

$$E[C_3] = E[W_{3Q}] + E[S_3] = \frac{E[N_{3Q}]}{\lambda_3} + \frac{1}{\mu_3}.$$

When cache miss in the local network (proxy server) client, cache miss at the local access network, the network-level cache hit, the average response of web caching system will be:

$$\begin{aligned} E[R_{2H}] = ((D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + \\ E[C_2] + D_3 + D_4) \end{aligned} \quad (9)$$

where $E[C_2] = E[W_{2Q}] + E[S_2] = \frac{E[N_{2Q}]}{\lambda_2} + \frac{1}{\mu_2}$

When cache miss in the local network (proxy server) client, cache miss at the local access network, the network-level cache slipped, hit central cache national network level, the average response of web caching system will are:

$$\begin{aligned} E[R_{1H}] = ((D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + \\ (D_{2M} + D_{2REQ}) + E[C_1] + D_2 + D_3 + D_4) \end{aligned} \quad (10)$$

When cache miss in the local network (proxy server) client, cache miss at the local access network, the network-level cache miss, miss central national network-level cache, the cache hit in the international Internet, the response average of web caching system will be

$$\begin{aligned} E[R_{0H}] = ((D_{4M} + D_{4REQ}) + (D_{3M} + D_{3REQ}) + \\ (D_{2M} + D_{2REQ}) + (D_{1M} + D_{1REQ}) + \\ E[C_0] + D_1 + D_2 + D_3 + D_4) \end{aligned} \quad (11)$$

Web caching architecture of a hybrid ISP 4 for average

response for Internet access are:

$$E[R_{WC}] = E[R_{3H}] + (Miss_3)(E[R_{2H}] + (Miss_2)(E[R_{1H}] + (Miss_1)(E[R_{0H}]))) \quad (12)$$

Thus, the worst case is not required web hit at all levels of the ISP network in the national network and only hit the web on an international level at the Internet web server resources. The equation (12) presents mean responses for internet access depending on web caching architecture of each network level, web caching organization, size of web cache, and bandwidth of communication channel (D_n), the protocols and web cache replacement algorithm, the rate cache miss levels ($Miss_n$). Internet network architecture 3 layer is very common: Institutional, regional, and national. However, to response the requirements of small delay for high-speed service and real-time equation(12) can be the basis for design calculations Internet and Web caching architectures suitable for each ISP.

Table 1. *The average response dependency ratios at the network level cache miss calculated (12) and that the value* $E[R_{3H}] = 5ms, E[R_{2H}] = 8ms, E[R_{1H}] = 12ms, E[R_{0H}] = 12ms$.

Only change $Miss_3$				**Only change** $Miss_2$				**Only change** $Miss_1$			
$Miss_3$	$Miss_2$	$Miss_1$	$E[R_{WC}]$	$Miss_3$	$Miss_2$	$Miss_1$	$E[R_{WC}]$	$Miss_3$	$Miss_2$	$Miss_1$	$E[R_{WC}]$
0.1	0.7	0.7	7.09	0.7	0.1	0.7	11.89	0.7	0.7	0.1	16.09
0.2	0.7	0.7	9.18	0.7	0.2	0.7	13.18	0.7	0.7	0.2	16.68
0.3	0.7	0.7	11.26	0.7	0.3	0.7	14.46	0.7	0.7	0.3	17.26
0.4	0.7	0.7	13.35	0.7	0.4	0.7	15.75	0.7	0.7	0.4	17.85
0.5	0.7	0.7	15.44	0.7	0.5	0.7	17.04	0.7	0.7	0.5	18.44
0.6	0.7	0.7	17.53	0.7	0.6	0.7	18.33	0.7	0.7	0.6	19.03
0.7	0.7	0.7	19.62	0.7	0.7	0.7	19.62	0.7	0.7	0.7	19.62
0.8	0.7	0.7	21.70	0.7	0.8	0.7	20.90	0.7	0.7	0.8	20.20
0.9	0.7	0.7	23.79	0.7	0.9	0.7	22.19	0.7	0.7	0.9	20.79

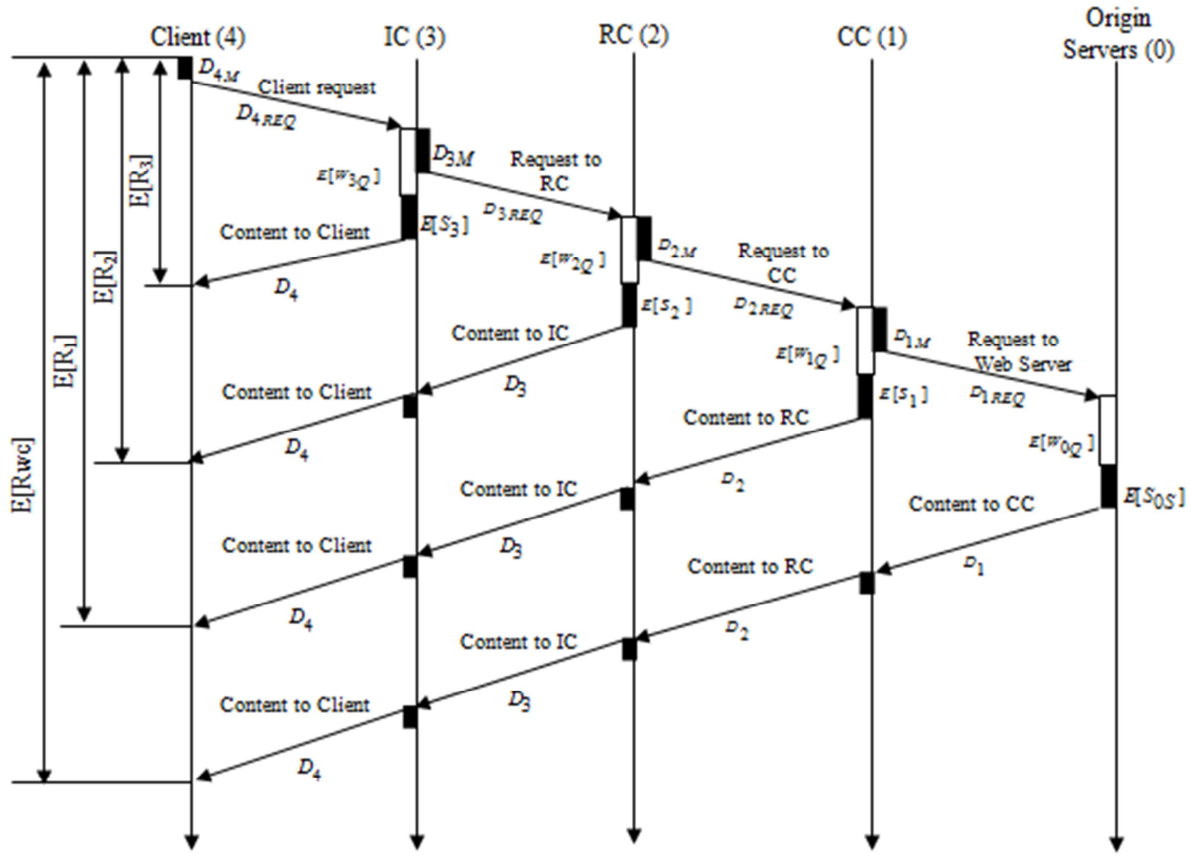

Figure 4. *Delay time chart of HTTP transaction for client on Internet with hybrid Web catching architecture.*

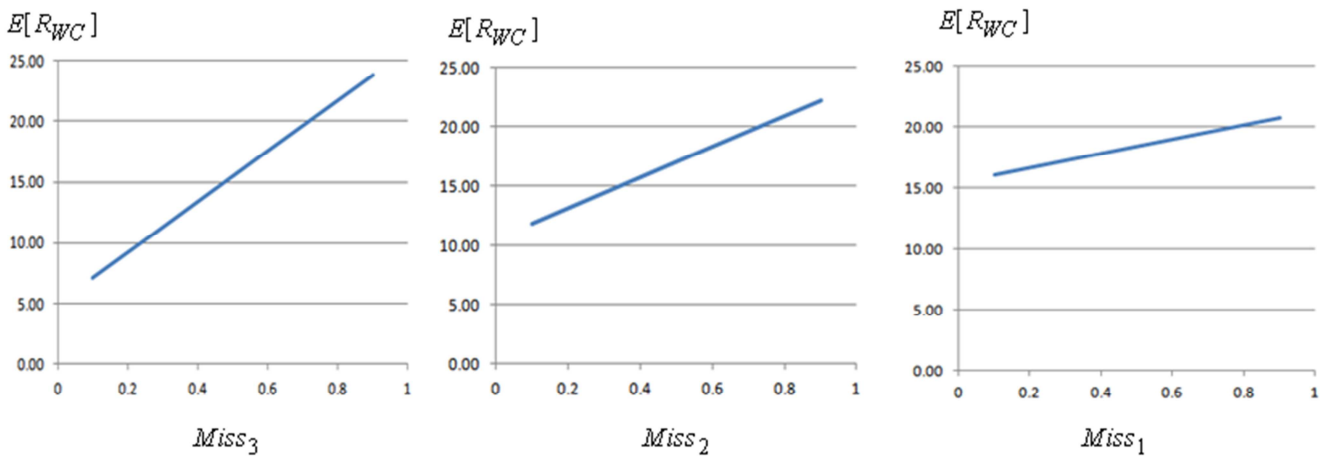

Figure 5. *Comparison of the effects of the ratio cache miss come* $E[R_{WC}]$.

4. Results and Discussion

Table 1 shows the results calculated average response of web caching architecture hybrid equation (12) according to the change of the ratio cache miss in the level Internet. Seen in Figure 5, the ratio cache miss at the local access network greatly affect local average response of web caching architecture. When $Miss_3 = 0.1$, then $E[R_{WC}] = 7.09\,ms$ is smallest, but when $Miss_3 = 0.9$ then $E[R_{WC}] = 23.79\,ms$ is largest with the changes of $Miss_2$, $Miss_1$. Thus, the solution build system Institutional caches in the local access network to ensure the performance of web caching architecture better than a lot of investment and costly network of regional and national level. Organizations in the caching system with POP application cache replacement algorithms and protocols as web caching solutions more economical and more efficient hybrid architecture for web caching. If additional systems of the LAN proxy servers attached to the end user, it also proved that the improved solution for the web proxy cache servers as well as the responsiveness of web caching architecture better.

Based on the formula (12) we can calculate the average response according to the dependence of the size or type of web services, the bandwidth of the access network, the transmission channel at the network level, etc .

5. Conclusions

In the present study, a hybrid web caching architecture for queue model has been suggested to estimate performance based on time at each level of web-winning network and web time overall winner of the ISP network with n-level network. The average response was calculated according to the dependence of the size or type of web services, the bandwidth of the access network, the transmission channel at the network level. Organizations in the caching system with POP application cache replacement algorithms and protocols as web caching solutions more economical and more efficient hybrid architecture for web caching. Moreover, the present solution is improved solution for the web proxy cache servers as well as the responsiveness of web caching architecture better if we have systems of the LAN proxy servers attached to the end user.

References

[1] Pablo Rodriguez, Christian Spanner, Ernst W.Biersack: Web Caching Architectures: Hierarchical and Distributed Caching. http://workshop99.ircache.net (4th International WWW Caching Workshop), Institut EUROCOM, france, 1999.

[2] Guangwei Bai, Carey Williamson: Workload Characterization in Web Caching Hierarchies. 10th IEEE International Symposium on Modeling, analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS'02), 2002.

[3] Abdullah Balamash, Marwan Krunz and Philippe Nain: Performance Analysis of a Client-Side Caching/Prefetching System for Web traffic. Computer Networks, Volume 51, Issue 13, 12 September 2007, Pages 3673-3692, Copyright @ 2007 Elsevier B.V. All rights reserved.

[4] Gunter Bolch, Stefan Greiner, Hermann de Meer, and Kishor S. Trivedi,"queueing Networks and Markov Chains Modeling and Performance Evaluation with Computer Science Applications". A Wiley-Interscience Publication Copyright © 1998 by John Wiley & Sons, Inc.

[5] Carey Williamson, Mudashiru Busari: Simulation Evaluation of Web Caching Architectures, M.Sc. Thesis, June 2000, Department of Computer science, University of Saskatchewan, http://www.cs.usask.ca/faculty/carey/.

[6] A. Rousskov, "On Performance of Caching Proxies", In ACM SIGMETRICS, Madison, USA, september 1998.

[7] C. Maltzahn, J.Richardson, "Performance Issues of Enterprise Level Web Proxies", 1998

[8] G.N.K.Suresh babu and S.K.Srivatsa "An Analysis of Web Caching Strategies to Improve Web Performance". International Journal of Software and Web Sciences (IJSWS) 13-270; © 2013. ISSN (Print): 2279-0063 ISSN (Online): 2279-0071.

[9] V. Padmapriya and K.Thenmozhi,"Web caching and response time optimization based on eviction method". International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, Issue 4, April 2013. ISSN: 2319-8753.

[10] Dr.K.Ramu1 and Dr.R.Sugumar,"Design and Implementation of Server Side Web Proxy Caching Algorithm" International Journal of Advanced Research in Computer and Communication Engineering VOL. 1, ISSUE 1, MARCH 2012. ISSN 2278 – 1021.