Middle-East Journal of Scientific Research 18 (10): 1517-1523, 2013
ISSN 1990-9233
© IDOSI Publications, 2013
DOI: 10.5829/idosi.mejsr.2013.18.10.12420

Network Services and Applications: Web Caching

[1]N.F. Dzulkhifli, [1]R. Alsaqour, [2]O. Alsaqour,
[1]H. Shaker and [3]R.A. Saeed

[1]School of Computer Science, Faculty of Information Science and Technology,
University Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
[2]Department of Computer Engineering, Faculty of Engineering and Technology,
The University of Jordan, 11942, Amman, Jordan
[3]Electronics Engineering Department, Sudan University of Science and Technology, Sudan

Abstract: As Internet content today grows and the number of user increase thus, the Internet traffic also increase. As one of the most popular applications running on Internet, the World Wide Web (WWW) experiences network congestion and overloading of the origin server. The existence of Web caches plays important roles in reducing network traffic between clients and server. In this paper, we discuss the overview of Web caching generally including the types of Web cache, the Web cache architecture, Internet Cache Protocols (ICP), Hyper Text Caching Protocols (HTCP), the advantages and disadvantages of Web caching and also the dynamic data caching.

Key words: Web caching · ICP · HTCP · Bandwidth · Network congestion

INTRODUCTION

The World Wide Web (WWW) started commercially in the early 1990 and shows the flooding of Internet users as times pass. This requires the high performance of Web system because of badly bus traffic and latency of data access occurs back day. Web traffic is caused by the interaction between many components such as clients, proxies and servers [1]. It result the Web pages available in a fraction form on WWW. Furthermore, when the clients ask a request from browser, the request has sent it to the server caused the waste of time. The declined of performance makes the clients more frustrated and moves on to another websites and this lead to more bigger and expensive connection to Internet for premium satisfaction.

Thus, the cache memory concept is implementing to Web because of the characteristic of the memory to allow fast access to Internet. Like disk-cache, it is used to store disk pages that often accessed for fast access (however, memory cache is faster than disk cache). Generally, we can say that Web caching is the storage or a folder full of Web objects that placed close to client for faster access and improved the performance of Web surfer [2]. It is stored for a period of times. The Web object refer to Web pages such as Hyper Text Markup Language (HTML), images, video or any file that can be retrieve from server or websites [3].

The rest of paper is organized as follows. We first discuss about types of caching that is commonly implements in server and its architecture. Then we introduced the Internet cache protocol and HTCP where it is responsible in exchange information about Web object cached by them. Lastly, we delve into the advantages and disadvantages of Web caching and then the existence dynamic Web caching in performance Web server.

MATERIALS AND METHODS

Types of Caching: There are several types of Web cache were introduced for Web objects which are to increase consistency in displays the information.

Corresponding Author: R. Alsaqour, School of Computer Science, Faculty of Information Science and Technology
University Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

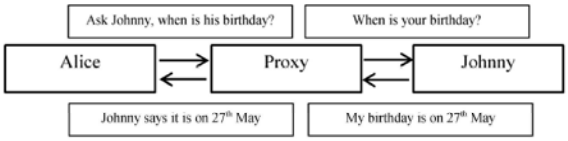

Fig. 1: The concept of proxy

Browser Cache: It lies on every popular Web cache and responsible as temporary storage in memory or disk that holds most recently downloaded pages. So, if user back to one of these pages, it just views it back without having to download the pages again. However, to ensure the pages are recently displayed, the cache compared the online version with the one in cache folder. If changed, the browser automatically downloads, display and cache otherwise, just display the cached page.

Proxy Cache: While a proxy cache is a shared network device that can undertake, Web transactions on behalf of clients or in other words the page can quickly retrieved by not only the same user but also the different user every time the page is requested. The word proxy itself means as "to act on someone behalf" which is means the clients and server thinks they are communicates to each other but actually they only dealing with the proxy server [4]. The server intended to simplify and controls the requested services as it also takes part in build firewall to prevent private connection from any attacks.

Figure 1 above is the example of the role of proxy which is it act as intermediary for request from clients seeking data or page from server. The page that has been retrieved saves on proxy server or servers' hard disk for faster retrieving rather than loads from Internet. It also caches the incoming Web pages. For examples, if the user jumps to a new page on the same sites like same images, the proxy cache them already for quicker browsing.

Reverse (Inverse) Proxy Cache: The reverse proxy is the router for all requests that have been retrieved from Internet to the Web servers. Then, the respond is returned as though it is originated from the Web server itself. It is installed in front of one or more Web server and contradict with forward proxy where serves only the restricted set of websites. Furthermore, it implements the load balancing by distributes towards various server where every server have to serve its own applications area [5]. It also can reduce load balance by caching static contents like pictures or other static graphical content as well as dynamic content and provide encryption also compression to speed up the loading times. No need to

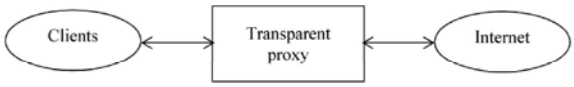

Fig. 2: Describes the location of transparent proxy cache

publish new Uniform Resources Locator (URLs) for public even though the administration changes by Web server behinds the firewalls. For better performance of intercepting the request to the Web server responds to the request rather providing priority to cached pages by reduce the amount of pages actually generated by the Web.

Transparent Proxy Cache: It also known as intercepting proxy, inline proxy or forced proxy which is have ability of interception between clients and servers at networks layer without being visible in other words clients does not have to bother about its existence.

Figure 2 above shows the transparent proxy cache location which is between clients and Internet and performing task as 'gateways' or routers. Thus, every request from Web wills automatically going through the proxy. It makes a note to which they forward it and not hide receiver Internet Protocol (IP) address. Internet Service Provider (ISPs) used bandwidth to retrieve Web page from server to send to client. It is commonly used them in some countries to save upstream bandwidth and cost. In addition, it also improved the performance of customer response by caching.

Server-side Caching: It is the Web based software component that allows saving and reading temporary information that took a long time to compute. It does not use in-memory cache techniques but store its cache on a disk in easy for understanding file structure. It speed up dynamic website or content data caching but it need to enable or equipped with some destructor so that it could deleting cache data that has been changed and if not it will prevent the data be displayed [6]. As an example Helicon Jet technology needs to set up with cache expiration timeout for updating information only for dynamic websites because for static files, Helicon Jet updates these files immediately after any modification involves to master file.

Web Cache Architecture: Basically Web cache is located between client and server, so that the process of retrieve the Web object to the client becomes much faster. Web should not have problem likes poor response times and system downs because of peak access. If not it will cause lost revenue [7].

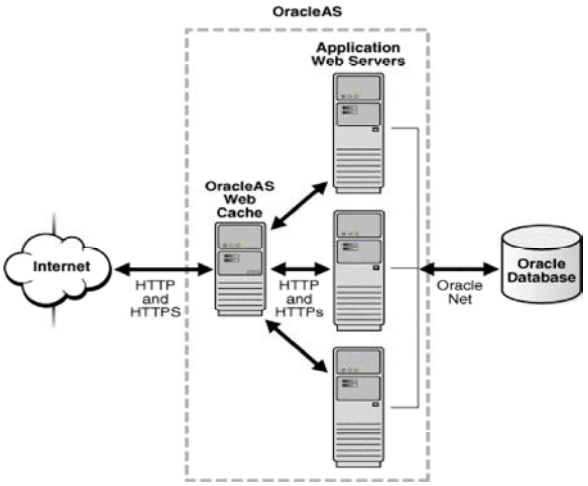

Fig. 3: Oracle Web cache architecture [5]

Thus, Oracle introduces its basic Web cache architecture as shown in Figure 3. The Oracle Web Cache (OWC) acts as virtual server for the application of Web Server. To enable the communication of Web browser with Oracle Web Cache when accessing a Web site is by configures the OWC with the same IP address that is registered for a sites domain name and the application Web Server's host name [7]. Like other Web cache, every request asks by clients send to cache memory. If the requested content is in its cache and sends it to browser, it is called as cache hits. Otherwise, it is called as cache miss. Then, the application web server sends the content through OWC to the client and makes a copy of the page in cache. Any invalid or outdated page will remove from cache.

The size of community connected to cache influence the performance of these cache as the hit rates can be increase because the more people accessing the same cache, the higher probability of the file present in the cache. For large scale cache, there are two common approaches have been implements to cooperates which are hierarchical and distributed caching. Hierarchical caching has been placed at different network level because it has several intermediates caches that communicate each other. All cache hierarchies recognize the concept parent and child [8]. The higher cache in hierarchy is known as parent cache for forward request, while a cache which is forward request to the parent cache known as a child cache.

The hierarchical cache, shown in Figure 4, works from the bottom level to the upper level or called as parent cache. Firstly, the request is served to the client cache (all arrow refer to clients) that is lie at the bottom level before redirected to institutional cache if does not hit the documents. If the document is not found there, the institutional cache would forward the request to the next level which is regional cache before forwards to national cache. If the request is a cache miss as goes down the hierarchy level, the national cache (parent cache) retrieves the content from origin server or in another cache. Once the cache hit the request, it travels down the hierarchy and leaving a copy of document on each intermediate cache. There are optimum number of caches that should be cooperates for each level before redirect the request to the parent cache in the hierarchy or to origin server. However there are several problem encounters with this hierarchical cache topology:

- Each level have own additional delays [9].
- The higher level caches can encounter bottlenecks and have long queuing delays.
- Redundant copies of same document at each cache levels.

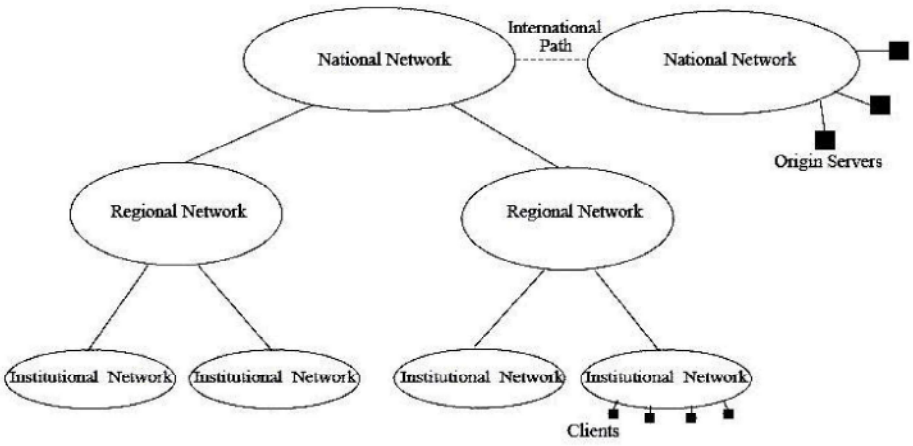

Fig. 4: Hierarchical caching topology

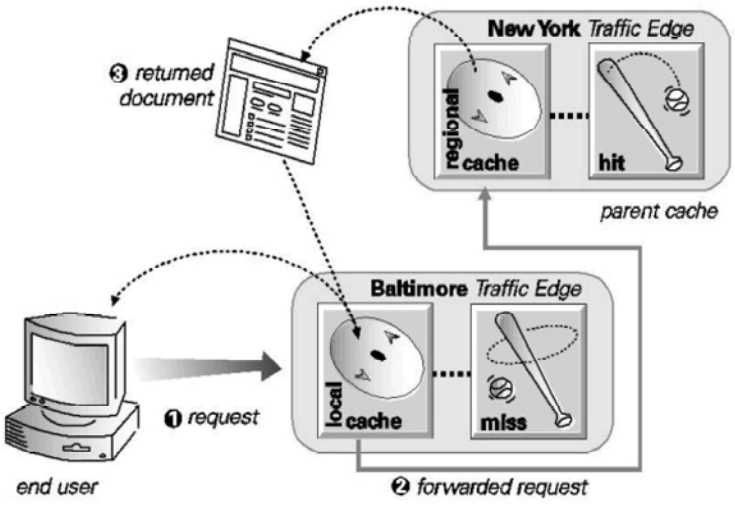

Fig. 5: Theillustration of simple cache hierarchy with a Traffic Server node configured to use a parent cache

Figure 5 shows that a client sends a request to a Traffic Server node or known as a child in the hierarchy cache. The request is a cache miss because it is configured to forward miss requests to a parent cache. Then, the request was forwarded to parent cache, where it is a cache hit. The parent sends a copy of the content to the Traffic Server, where it is cached and send to the client. Any future request for the same document now can be retrieved directly from the Traffic Server cache until the document is stale or expired.

While the distributed caching were introduce as no intermediates cache and only depend on institutional cache at the edge of network to communicate with each other. In order to store and decide from which institutional caches to retrieve a miss document, metadata information or location hints are used [9]. Result, that each cooperate institutional caches knows each other contents. For more efficient metadata information, a hierarchical caching of intermediates nodes can be used but it does not provide the documents but only the information of the location of documents. However by using distributed caching no additional delays are introduced and no redundant copies at intermediates caches thus, saving disk space.

The performance of hierarchical and distributed caching is measured by latency resulted to retrieve a document. Hierarchical distribution placing redundant copies in intermediates cache level because it has lower connection times. Besides that, for a distributed cache has lower transmission times than hierarchical since it distributes the traffic better using more bandwidth in lower network levels. However through hybrid scheme, by using the right number of cache cooperates at each network level can combine both advantages of hierarchical and distributed caching in term to reduce the connection time and the transmission time [9]. Thus, for minimization of the retrieval latency and the usage of bandwidth, it is depends on the traffic of network, the parent cache and the file's size.

Internet Cache Protocol: Internet Cache Protocol (ICP) is a lightweight message format that used for communicating among Web cache [10]. It is used by Web cache as exchange information about Web object cached by them. The information later is used to decide from where to retrieve the requested Web object. It also is used efficiently as possible to minimize the request towards originating server because it mainly goes to capture an object towards neighbouring cache.

ICP is used mainly in cache mesh to find the specific Web objects or the existence of URL in neighbouring cache by sending a query. Then, the neighbour replies back via ICP weather it is a miss or a hit. ICP transaction basically is done as follows [10]:

- Local cache receives an HTTP request from cache client.
- The local cache sends ICP queries to locate an object in neighbouring cache (or peer caches) in a time given.
- The peer caches receive the queries and send ICP replies. It examines and decides which reply message has been sent. It could be denied, miss, or hit.
- The local caches receive the ICP replies and decide where to forward the request.

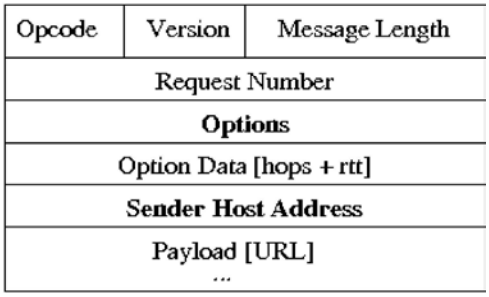

Fig. 6: ICP message format

ICP also function as sender of cache access policies. It gives a warning in advance about the result (hit or miss) to every subsequent HTTP request. So, we certainly would not send the request to sender as the result could always depends on other request field such as Cache Control. The ICP messages are generally very small [11]. It consists of two sections: header and data. Its format is consists of 20-octet binary header followed by URL string.

The Figure 6 is the format of ICP implements in Web caching. The features in format are described below:

Opcode: The operation code used in ICP listed in Table 1. Version: The ICP version number.

Message Length: Shows the sum of length (octets) of ICP message. However, it must not exceed 16,384 octets in length.

Request Number: As an opaque identifier. It needs to be including into the reply message.

Options: A 32-bit field of option flags that allows extension of this version of the protocol in certain, limited ways.

Option Data: A 4-octet field to support optional features.

Sender Host Address: The IPv4 address of the host sending the ICP message (practically not used because this field probably not be trusted.)

Payload: Vary depends on the Opcode (mostly often contains null-terminated URL string).

In addition, there are the differences between ICP and HTTP. ICP as we states before was designed to be simple, small and efficient otherwise; HTTP is designed to support a rich and sophisticated set of features. Other than that, any request and reply headers of HTTP consist of lines of American Standard Code for Information Interchange (ASCII) text by a Carriage Return and Line Feed (CRLF) pair while ICP only uses a fixed size header and represents number in binary.

HTCP is a protocol which is newer than ICP and is better at predicting hits. It have many common characteristics with ICP [12] even though it broader in scope and more complex. Both, HTCP and ICP used User Datagram Protocol (UDP) as transport and known as pre- request protocols. But HTCP deals with ICP a few problems such as:

- ICP query only contain a Uniform Resources Identifier (URI) but not a single request method, whereas HTCP contains full HTTP request header.
- ICP does not provide security. HTCP has optional message authentication via shared secret keys. However, both does not support encrypted message.
- As states before, ICP is simple, fixed-size binary message format and hard to expand while HTCP is complex and have variable-binary message format.

Table 1: The opcode‘s table.

Value	Name	Explanation
0	ICP_OP_INVALID	To detect zero-filled or malformed messages.
1	ICP_OP_QUERY	A query message
2	ICP_OP_HIT	Indicates the requested URL exist in the cache and allowed to retrieved
3	ICP_OP_MISS	Indicates the requested URL does not exist in the cache
4	ICP_OP_ERR	Indicates occur error in parsing the query message
5-9	UNUSED	-
10	ICP_OP_SECHO	Use in simulates a query to an origin server
11	ICP_OP_DECHO	Use in simulates a query to a cache which does not use ICP
12-20	UNUSED	-
21	ICP_OP_MISS_NOFETCH	Cannot fetch the URL from the replying cache
22	ICP_OP_DENIED	The query sites is not allowed to retrieves the URL from the cache
23	ICP_OP_HIT_OBJ	Same like ICP_OP_HIT but included in the reply message (query message).

The purposed using HTCP is to discover HTTP caches and cached data, handles the sets of HTTP caches, also to monitor cache activity [12]. It allows full request and response toward headers that can be used in cache management and extends the domain of cache management as including monitoring a remote cache‘s addition and deletion, asking immediate deletion and also send hints about Web objects like the locations of third party of cacheable objects or the uncacheability or unavailability of Web objects measurement. Transmission of all multi-octets HTCP protocol is in network byte order which is the byte of multi-byte number are transmitted on a network. However, it may or may not match the order of number that normally stored in memory for a particular processor.

The senders should set all the reserved fields to binary zero and left unexamined by receivers. While headers has to present with CRLF line termination like in HTTP. Like ICP, HTCP also need to support UDP as message is sent in form of UDP datagrams. It also need to act in useful way when there are no responses, delayed or damaged. The Internet Assigned Number Authority (IANA) already assigned port 4827 as the standard Transmission Control Protocol (TCP) and UDP port number of HTCP [12].

RESULTS AND DISCUSSION

The Advantages and Disadvantages of Web Caching:
The Advantages: Web cache is created to improve the performance in surfing Internet. Generally it gives many advantages to the surfer such as:

Web caching minimizes the workload off of the server by saving the requesting objects in Web cache which are weather at browser, proxy and server level. In addition, the request that gets object hit are directly sent to the users from the cache while object miss or the stale data is the only request from origin server.

It also can reduce the Internet bandwidth by utilizing a proxy cache thus; can decrease Web traffic and network congestion. Furthermore, an organization also can control the accessible of user in surfing Internet to avoid their employees wasting the bandwidth browsing non-related websites.

While the user can experiences the reduction of latency when accessing Web sites by using Web caches because the minimization of transmission delays and the reducing network traffic.

The enhancing of Web services is proven when the remote server’s gone down or network partitioning; the user still can get or view a copy of cached file at the proxy.

The Disadvantages: However there are a few problems regarding Web caches:

The most gaining issues are user could look the stale objects due to improper the proxy cache updating the objects.

In proxy caching, there lie a limit in how many user can access to avoid an increasing of latency to a desirable amount. So, to design a caching system, the cache hit should be maximizing while the cost of cache miss have to minimize.

For a single proxy cache, it should have limit to be set for the number of clients that it can serve to avoid bottleneck. Besides, the proxy server should be almost as efficient as user accessing the origin server.

Some of the origin server act to disable caching of their Web documents because caching reduces hits in their own servers.

Dynamic Data Caching: The existence of dynamic data always slows down the performance of Web servers. Typically high-performance of Web servers can deliver several hundred static files per seconds, contradict with the delivering of dynamic pages where the rate of delivery is often one or two order of magnitude slower [13]. Sites nowadays always disable caching of documents because it contains banner ads which is they make money from them where payments is based on the number of hits of documents. Thus, the server acts to disable caching function and for other reason is to gather information about user who is using the contents. There are at least two current dynamic data caching approaches: active cache and server accelerator.

Active cache provides for caching of dynamic documents by let server to furnish cache applet at Web proxies together with a documents. In other words, whenever there are caches hit to the documents, proxies has to fetch corresponding cache applet. Furthermore, when a user asked for the hits on cached copy, the proxy would like to service the demand; the proxy needs to invoke the cache applet with the user request and other information as arguments. So, it can act without interacting with the server. In reply, the cache applet then decide what would proxy send back to user either applet give proxy a new documents to send back to the user, or allow the proxy to send back the cached document or asked proxy to send request to the Web server. Cache applets allows server to perform variety of functions such as logging user access, rotates advertising banners, checking access permission, etc.

Web server accelerator is run in embedded operating system where it can enable to the Web server to serve a huge number of pages per seconds or in other words, to

speed up user accesses. It offers Application Program Interface (API) that allows addition, deletion and updating cached data by application programs. The existence of API allows the cache of both static and dynamic data. Data Update Propagation (DUP) algorithm is used to facilitate invalidated and updated data by maintains the data depends on information between cached data and underlying data in graph.

CONCLUSION

Avoiding the World Wide Web becomes World Wide Wait, the Web caching is created to be one of the effective to overcome networks traffic and server overloading. Caching has been a key of technology at solving server bottleneck, network congestion and minimizes the user access latency. In addition, this paper has discuss on how Web caching help to improves the performance of Web by go through the architecture of Web cache and the existence of ICP to communicates with all Web caches. As a future work, we will conduct a survey on analysing and evaluating the differences of Web caches embedded in different operating system and also the enhancement of Web caches encounter the problems that states in disadvantages sections.

ACKNOWLEDGMENT

This study was partially funded by the University Kebangsaan Malaysia under Grant Nos. UKM-GUP-2012-089 and FRGS/1/2012/SG05/UKM/02/7.

REFERENCES

1. Ihm, S., 2011. Understanding and Improving Modern Web Traffic Caching: Princeton University.
2. Mohan, C., 2001. Caching Technologies for Web Applications.
3. Korobkin, D.M., S.A. Fomenkov, S.G. Kolesnikov and Y.F. Voronin, 2013. System of Physical Effects Extraction from Natural Language Text in the Internet.
4. Bennett, F., T. Richardson and A. Harter, 1994. Teleporting-making applications mobile, in Mobile Computing Systems and Applications, 1994. WMCSA 1994. First Workshop on, pp: 82-84.
5. Sekar, K., K. Ravichandran and J. Sethuraman, 2013. A Mining Based Quality Evaluation and Prediction of Web Applications, World Applied Sciences Journal, 25: 1490-1501.
6. Barish, G. and K. Obraczke, 2000. World wide web caching: Trends and techniques," Communications Magazine, IEEE, 38: 178-184.
7. Anton, J., L. Jacobs, X. Liu, J. Parker, Z. Zeng and T. Zhong, 2002. Web caching for database applications with Oracle Web Cache, in Proceedings of the 2002. ACM SIGMOD international conference on Management of data, pp: 594-599.
8. Mahanti, A., C. Williamson and D. Eager, 2000. Traffic analysis of a web proxy caching hierarchy," Network, IEEE, 14: 16-23.
9. Rodriguez, P., C. Spanner and E. W. Biersack, 2001. Analysis of web caching architectures: hierarchical and distributed caching, Networking, IEEE/ACM Transactions on, 9: 404-418.
10. Wessels, D., 1997. Application of internet cache protocol (ICP), version, pp: 2.
11. Cooper, I., I. Melve and G. Tomlinson, 2001. Internet web replication and caching taxonomy.
12. Vixie, P. and D. Wessels, January, 2000. Hyper Text Caching Protocol (HTCP/0.0), RFC 2756.
13. Iyengar, A. and J. Challenger, 1997. Improving web server performance by caching dynamic data, in USENIX Symposium on Internet Technologies and Systems, pp: 49-60.