Volume 4, Issue 3, March 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Review on Web Caching Techniques

Mukesh Dawar[1]
Department of Computer Science & Engineering
Shivalik Institute of Engineering & Technology, Aliyaspur
India

Charanjit Singh[2]
Department of Computer Science & Engineering
RIMT Institute of Engineering & Technology,
Mandigobindgarh, India

Abstract: ***The concept of web caching has attained a prominent place in network architecture. A Web cache is a mechanism for the temporary storage (caching) of web documents, such as HTML pages and images, to reduce bandwidth usage, server load, and perceived lag. A web cache stores copies of documents passing through it; subsequent requests may be satisfied from the cache if certain conditions are met. The World Wide Web suffers from scaling and reliability problems due to overloaded and congested proxy servers. Caching at local proxy serves help, but cannot satisfy more than a third to half of requests; more requests are still sent to original remote servers. This paper discusses several web caching schemes such as Distributed Web Caching (DWC), Distributed Web Caching with Clustering (DWCC), Robust Distributed Web Caching (RDWC), Distributed Web Caching for Robustness, Low latency & Disconnection Handling (DWCRLD). Clustering improves the retrieval latency and also helps to provide load balancing in distributed environment. But this cannot ensure the scalability issues, easy handling of frequent disconnections of proxy servers and metadata management issues in the network.***

Keywords: Web caching, Metadata Server, Distributed Web Caching, Clustering, Latency, Robustness, Scalability, Disconnection Handling, Proxy server, clients.

I. INTRODUCTION

The surge in popularity of the World Wide Web (WWW) has introduced new issues such as Internet traffic and bandwidth consumption. Recently much research has focused on improving Web performance by reducing the bandwidth consumption and WWW traffic. It means that fewer requests and responses need to go over the network and fewer requests for a server to handle. A *Web cache* sits between one or more Web servers (also known as *origin servers*) and a client or many clients, and watches requests come by, saving copies of the responses like HTML pages, images and files (collectively known as *representations*) for itself. Then, if there is another request for the same URL, it can use the response that it has, instead of asking the origin server for it again. There are two main reasons that Web caches are used:

- To **reduce latency** — Because the request is satisfied from the cache (which is closer to the client) instead of the origin server, it takes less time for it to get the representation and display it. This makes the Web seem more responsive.
- To **reduce network traffic** — Because representations are reused, it reduces the amount of bandwidth used by a client. This saves money if the client is paying for traffic, and keeps their bandwidth requirements lower and more manageable.

Kinds of Web Caches

- **Browser Caches:** If you examine the preferences dialog of any modern Web browser (like Internet Explorer, Safari or Mozilla), you'll probably notice a "cache" setting. This lets you set aside a section of your computer's hard disk to store representations that you've seen, just for you. The browser cache works according to fairly simple rules. It will check to make sure that the representations are fresh, usually once a session (that is, the once in the current invocation of the browser).This cache is especially useful when users hit the "back" button or click a link to see a page they've just looked at. Also, if you use the same navigation images throughout your site, they'll be served from browsers' caches almost instantaneously.
- **Proxy Caches:** Web proxy caches work on the same principle, but a much larger scale. Proxies serve hundreds or thousands of users in the same way; large corporations and ISPs often set them up on their firewalls, or as standalone devices (also known as *intermediaries*).Because proxy caches aren't part of the client or the origin server, but instead are out on the network, requests have to be routed to them somehow. One way to do this is to use your browser's proxy setting to manually tell it what proxy to use; another is using interception. *Interception proxies* have Web requests redirected to them by the underlying network itself, so that clients don't need to be configured for them, or even know about them. Proxy caches are a type of *shared cache*; rather than just having one person using them, they usually have a large number of users, and because of this they are very good at reducing latency and network traffic. That's because popular representations are reused a number of times.
- **Gateway Caches :**Also known as "reverse proxy caches" or "surrogate caches," gateway caches are also intermediaries, but instead of being deployed by network administrators to save bandwidth, they're typically deployed by Webmasters themselves, to make their sites more scalable, reliable and better performing. Requests

can be routed to gateway caches by a number of methods, but typically some form of load balancer is used to make one or more of them look like the origin server to clients. *Content delivery networks* (CDNs) distribute gateway caches throughout the Internet (or a part of it) and sell caching to interested Web sites. Speedera and Akamai are examples of CDNs.

II. WEB CACHING TECHNIQUES

Web caching is the approach of temporary storage of web objects, such as HTML files, for later retrieval. Few approaches have been suggested for effective web caching schemes.

2.1 Proxy caching

A proxy cache server receives HTTP requests from clients for a web object and if it finds the requested object in its cache, it returns the object to the user without disturbing the upstream network connection or destination server. If it is not available in the cache, the proxy attempts to fetch the object directly from the object's home server. Finally the originating server, which has the object, gets it, possibly deposits it and returns the object to the user. The benefits of proxy caching are supposed to reduce network traffic and reduce average latency. Proxy caches are often located near network gateways to reduce the bandwidth required over expensive dedicated Internet connections.

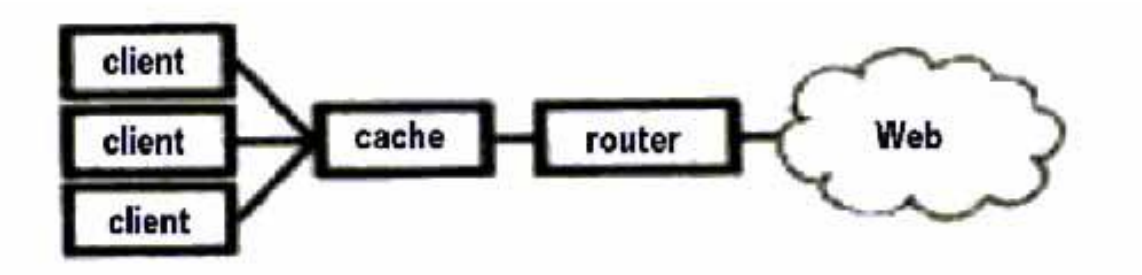

Figure: A standalone proxy configuration

2.1.1 Reverse Proxy Caching

An interesting variation to the proxy cache approach is the notion of reverse proxy caching, in which caches deployed near the servers, instead of near the clients. This is an attractive solution for servers that expect a high number of requests and want to assure a high level of quality of service (QoS). Reverse proxy caching is a useful mechanism when supporting virtual domains mapped to a single physical site, which is an popular service for many different service providers.

2.1.2 Transparent Caching

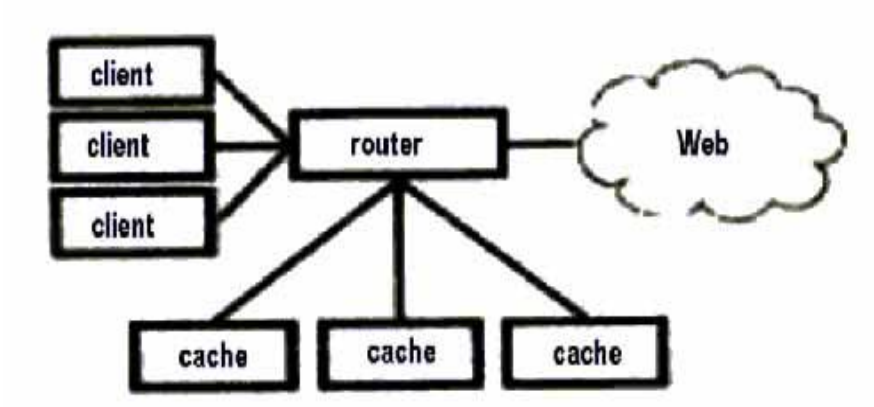

Figure 2: A router-transparent configuration.

One of the main drawbacks of the proxy server approach is the requirement to configure web browsers. The architecture of transparent caching eliminates this handicap. Transparent caches work by intercepting HTTP requests and redirecting them to web cache servers or clusters.

2.2 Adaptive web caching

The adaptive web caching system provides an effective evolutionary step towards the above goal. Adaptive caching consists of multiple, distributed caches which dynamically join and leave cache groups based on content demand. Adaptive caching uses the Cache Group Management Protocol (CGMP) and the Content Routing Protocol (CRP), CGMP specifies how meshes are formed and how individual caches join and leave that meshes.CRP is used to locate cached content from within the existing meshes.

2.3 Push Cashing

The idea of having a server decides when and where to cache its documents, was introduced as push cashing. The key idea behind this architecture is to keep cached data close to those clients requesting that information. Data is dynamically mirrored as the originating server identifies where requests originate.

2.4 Active caching

An active cache scheme is proposed in to support caching of dynamic contents at Web proxies. The growth of the Internet and the World Wide Web has significantly increased the amount of online information and services available to the general population of the society. The Active Cache is a scheme which migrates parts of server processing on each user request to the caching proxy in a flexible on demand fashion via "cache applets".

III. PREVIOUS WORK

An analysis of existing web caching schemes on the basis of connection time latencies, loads and transmission time latencies is done by Christian Spanner, Pablo Rodrigueand Ernst in [2pablo]. Some effective web caching architectures are Single and Multiple caches, Distributed Web Caching, Distributed Web Caching with Clustering, Hierarchical Caching, Cooperating Web Caching Scheme, Hybrid Web Caching and Robust Web Caching Schemes.

A. Hierarchical web caching scheme:

In hierarchical web architecture client's caches are at the bottom level and server caches are placed at the different levels of hierarchy. The request is redirected to the next level caches in hierarchy for every miss. The request is forwarded to the origin server if document is not found at any level and on reply copy is maintained at each intermediate server. But this scheme had problem of longer queries delay and redundancy of data at each level.

B. Distributed web caching scheme:

Internet Caching Protocol (ICP) and Hyper Text Caching Protocol (HTCP) were designed for distributed environment. They support management, discovery and retrieval of updated data from parent and neighboring caches as well. Cache Array Routing Protocol (CARP) is another approach for distributed caching. In this URL space is divided among an array whose elements are loosely coupled caches. Each document is hashed to a particular cache. Tewari proposes a scheme for fully distributed internet cache architecture, in which location hints are maintained and replicated at all local institutional caches. In Cache Digest and in the Relais Project [17] all caches maintain local directories of contents of other caches for ease of locating documents in other caches. Also caches keep exchanging messages with each other indicating their contents.

C. Hybrid web caching scheme:

In they have presented hybrid scheme, with the concept of caches cooperation at each level in hierarchy. Rabinovich have altered this scheme by limiting the cooperation between the neighboring caches. This scheme advantages by avoiding fetching of documents from the slower or distant caches if that document could be retrieved at a lower cost directly from the origin server.

D. Distributed Web Caching scheme with clustering:

Distributed Web Caching (DWC) scheme supports exchange of metadata among proxy servers periodically to maintain a complete metadata. Thus every proxy server maintains metadata of all other proxy servers along with its own metadata as shown in Figure 3. For every request proxy server firstly checks its own metadata for requested page id, if there is a hit the page is transmitted to the client otherwise it checks for page id in the whole metadata of other proxy servers and if found the request is forwarded to that server else to the origin server. Advantage of this scheme is high hit ratio. But if the size of metadata and number of proxy servers increases it becomes highly unmanageable. Even this scheme fails if any of the proxy servers gets disconnected.

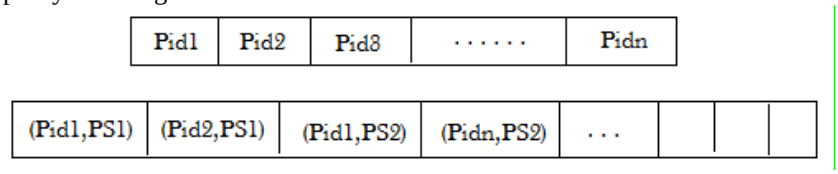

Figure 3: Metadata maintained at each proxy server

The Distributed Web Caching with clustering (DWCC) and Distributed Web Caching for Robustness, Low latency & Disconnection Handling (DWCRLD) are based on the geographical region based clustering. The author provides a solution for robustness and scalability problem in web caching due to heavy load. They have used the concept of clustering along with the feature of dynamic allocation of requests by maintaining metadata of neighboring clusters only not of all clusters. They also provide the concept of managing the load of overloaded server by transferring requests to less loaded proxy servers. The author has refined their scheme to handle more delays and frequent disconnections of proxy servers. This can result in fastest response to the clients and also provide load balancing. Even these schemes suffer from the scalability problem. If size of the cluster grows, size of metadata grows as well then metadata at every proxy servers can become unmanageable. This problem can be overcome by our enhanced proposed architecture.

IV. THE PROPOSED STRATEGY

This technique is based upon the Distributed Web Caching for clustering based environment in geographical region i.e DWCC (Distributed Web Caching with Clustering). The entire geographically together proxy servers are placed in the same cluster. This dynamic scheme provides easy management of metadata. Low latency by a factor of s/m can be achieved by this scheme. Unlike previous techniques there is no need of metadata exchange before serving any request. This scheme also limits the number of client's requests per cluster to avoid overloading of cluster's proxy servers. If any of the proxy servers becomes overloaded it will start dropping the further requests or lead to long delays. In this scheme

number of clients for every server and every cluster is maintaining a cluster queue length data structure to avoid this situation [3tiwari].

The proposed strategy includes origin servers, clusters of proxy servers and clients as shown in Figure 3. One extra node is added to every cluster that is Metadata Server (MDS). MDS's task is to maintain metadata of all proxy servers within own cluster and metadata of neighboring cluster. In previous strategy [3] every proxy server itself maintains metadata of its own cluster as well as of their neighboring clusters. So this strategy will reduce efforts and time of proxy servers.

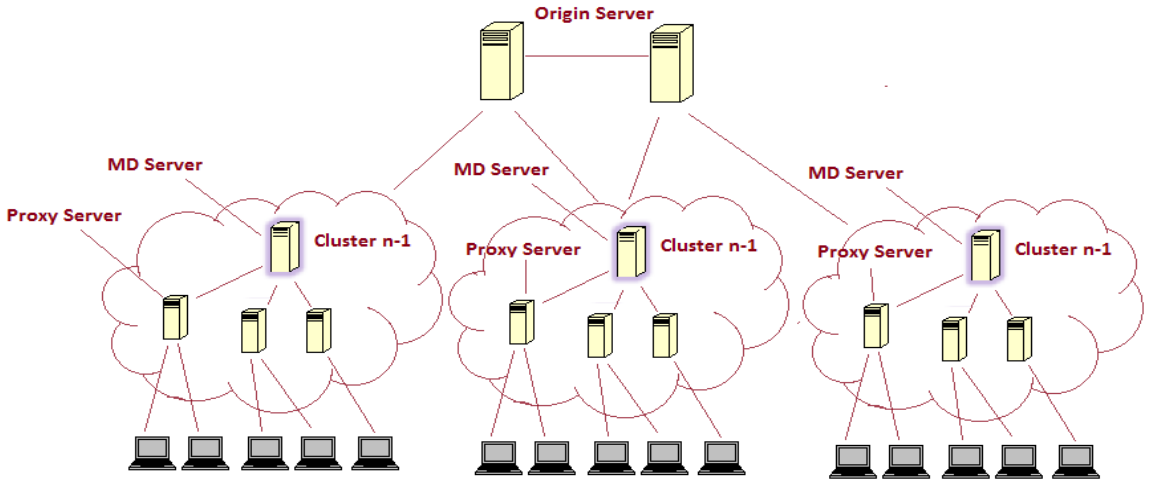

Figure 4: The Proposed Scheme's Architecture

V. CONCLUSIONS

Web caching is the best solution to reduce the internet traffic and bandwidth consumption. It is also a low cost technique for improving the Web latency. Now days, proxy caches are increasingly used around the world to reduce bandwidth and make less severe delays associated with delays. Web proxy servers sharing their cache directories through a common mapping service that can be queried with at most one message exchange. A number of caching schemes already exits. Some effective techniques such as DWC, DWCC, RDWC and DWCRLD schemes for distributed environment along with their limitations have been discussed in this paper. In this work, we have proposed a strategy called "*Improved Metadata Management & Scalability in Dynamic Distributed Web Caching*" that can be easily deployed in the future. This is based on the DWCRLD to improve the scalability and to alleviate extra overhead of metadata management of the proxy servers and also reduces the network traffic as well. This scheme also makes it easy to handle frequent disconnections in the network. By this even if the number of proxy servers grows in the network, metadata management will never be an issue. But this scheme is also having certain limitations such as hardware failure, cache routing, fault tolerance, proxy placement, security and dynamic data caching etc.

REFERENCES

[1] K. Claffy, H.W. Braun, *Web traffic characterization: An assessment of the impact of caching documents from NCSAs web server*, in Electronic Proc. 2nd World Wide Web Conf.94: Mosaic and the Web, 1994

[2] Pablo Rodriguez, Christian Spanner, and Ernst W. Biersack, *Analysis of Web Caching Architectures: Hierarchical and Distributed Caching*, IEEE/ACM Transactions On Networking, Vol. 9, NO. 4, AUG 2001

[3] Rajeev Tiwari, Neeraj Kumar, *Dynamic Web Caching: for Robustness, Low Latency & Disconnection Handling*, 2nd IEEE international conference on Parallel, Distributed and Grid Computing, 2012

[4] Chankhunthod et. al., *A hierarchical internet object cache*, in Proc. 1996 annual conference on USENIX Annual Technical Conference, San Diego, CA, Jan. 1996.

[5] Povey and J. Harrison, *A distributed Internet cache*, in Proc. 20th Australian Computer Science Conf., Sydney, Australia, Feb. 1997.

[6] V. Cardellini, M. Colajanni, P.S. Yu, *Geographic Load balancing for scalable distributed Web systems*, Proc. of MASCOTS'2000, IEEE Computer Society, San Francisco, CA, pp 20-27 Aug. 2000.

[7] D.Wessels and K. Claffy, *Application of Internet cache protocol (ICP)*, version 2, Internet Engineering Task Force, Internet Draft: draft-wessels-icp-v2-appl-00. Work in Progress., May 1997.

[8] V. Valloppillil and K. W. Ross., *Cache Array Routing Protocol*, v1.1. Internet draft. [Online]. Available: http://ds1.internic.net/internetdrafts/draft-vinod-carp-v1-03.txt, 1998

[9] Rajeev Tiwari, Lalit Garg, *Robust Distributed Web Caching Scheme: A Dynamic Clustering Approach*, in International Journal of Engineering Science and Technology in ISSN : 0975-5462 Vol. 3 No. 2 Feb 2011,pp 1069-1076.

[10] Adam Dingle, Tomáš Pártl, *Web Cache Coherence, Journal reference: Computer Networks and ISDN Systems*, Volume 28, issues 7–11, p. 907.

[11] S. Gadde, M. Rabinovich, and J. Chase, *Reduce, reuse, recycle: An approach to building large internet caches*, in *Proc. 6th Workshop on Hot Topics in Operating Systems (HotOS–VI)*, May 1997.
[12] P. Vixie and D. Wessels, *RFC 2756: Hyper text caching protocol*, (HTCP/0.0), Jan. 2000.
[13] R. Tewari, M. Dahlin, H. M. Vin, and J. S. Kay, *Beyond hierarchies: Design considerations for disturbed caching on the Internet*, in *Proc. ICDCS '99 Conf.*, Austin, TX, May 1999.
[14] A. Rousskov and D. Wessels, *Cache digest*, in *Proc. 3rd Int. WWW Caching Workshop*, June 1998, pp. 272–273.
[15] Makpangou, G. Pierre, C. Khoury, and N. Dorta, *Replicated directory service for weakly consistent replicated caches*, in *Proc. ICDCS'99 Conf.*, Austin, TX, May.
[16] Rabinovich, J. Chase, and S. Gadde, *Not all hits are created equal: Cooperative proxy caching over a wide-area network*, in *Proc. 3rd Int. WWW Caching Workshop*, Manchester, U.K., June 1998.
[17] Tiwari Rajeev and Khan Gulista, *Load Balancing in Distributed Web Caching: A Novel Clustering Approach*, Proc. of ICM2ST-1O,International Conference on Methods and models in science and technology pp. 341-345, November 6, 2010, vol.l324
[18] Rajeev Tiwari, Gulista khan, Load *Balancing through distributed Web Caching with clusters*, Proceeding of the CSNA 2010 Springer, pp 46-54, Chennai, India.

