Report on ‘A Survey of Web Caching Schemes for the Internet’ by Cheng Han

Summary: This paper presents the work researchers do when improving the Web performance. Web caching is an effective way to alleviate bottleneck and reduce network traffic. The paper discusses current Web caching schemes, but further works are needs in order to improve Web performance.

Background& Motivation: As the World Wide Web growing in size, network congestion and server overloading becomes a serious problem. Web caching can alleviate the service bottleneck and reduce the network traffic thereby minimize the user access latency. Hence the research in Web caching is required.

Main idea/solution: The paper first introduces the World Wide Web (WWW), and today’s main problems Web users are suffering from are the network congestion and server overload. Researchers work on solving these problems and caching popular objects at locations close to the clients is an effective solution. And in order to solve this problem, introducing the proxy. Part 2 of the paper asks some questions about how to make a WWW caching system work. There are some desirable properties include fast access, robustness, transparency, scalability, efficiency, adaptivity, stability, load balanced, ability to deal with heterogeneity, and simplicity the WWW caching system we want to have. Some caching architectures are then discussed in the paper. For hierarchical caching architecture, it can be explained as caches are placed at multiple levels of the network. When finding document, if the before level cannot find it, a deeper finding will continue until it gets the targeted document. And it travels down the hierarchy and leave a copy of the document in each of the intermediate caches along its path. But this architecture still has many problems therefore introducing distributed caching architecture. There are only caches at the bottom level. The hierarchy is only used to distribute directory information about the location of the document, not the copy of the document. Based on distributed caching, hybrid caching architecture is introduced which combine the advantages of both hierarchical and distributed caching. In order to solve the problem of how to quickly locate a cache containing the wanted document, we have the common approach which is to build a caching distribution tree away from each popular server towards sources of high demand and do cache resolution in two ways: caching routing table or hash functions. However, no matter what caching algorithm is used during the process, there is usually not more than half of the documents can be found. In order to get rid of this problem, prefetching, which can anticipate future document requests and preload or prefetch these documents in a local cache. This process can be applied in three ways in the Web contexts, which are between browser clients and Web servers, between proxies and Web servers and between browser clients and proxies. For prefetching between browser clients and Web servers, some studies were made but these studies came out too early that they do not consider or model caching proxies and fail to answer the question about performance of prefetching completely. And the rest two are proxies-based. The first two approaches however have the probabilities of increasing wide area network traffic, while the last one simply affects the traffic over the modems or the LANs, which is kind of advantage in this field. All of these approaches are trying to prefetch either documents that are considered as popular at servers or documents that are predicted to be accessed by user in the near future based on the access pattern. Then the paper considers the cache placement/replacement which attempt to minimize various cost metrics. They can be classified into three categories include traditional replacement policies and its direct extensions, key-based replacement policies and cost-based replacement policies. However, the performance of replacement policies depends highly on traffic characteristics of WWW accesses. Hence there is no policy that can suit every situation for all Web access patterns. Cache coherency is a widely concern problem since every Web cache must update pages in its cache so that it can give users pages which should be as fresh as possible. In this part we have HTTP commands that assist Web proxies in maintaining cache coherence and then introduces cache coherence mechanisms which provides two types of consistency: strong cache consistency and weak cache consistency. Caching contents include data cache, connection cache and computation cache. There are two approaches in predicting users' future requests. We also face the hot spots problem when a large number of client access data or get some services from a singe server. The solution is to store copies of hot pages/services through the internet and this can spread the work of serving a hot page/services across many servers. As mentioned before, only part of web data is cacheable and there are two approaches include active cache and web server accelerator can improve the Web performance. One thing should be mentioned is that the effectiveness of caching theory relies on the existing Web reference streams and the proper use of cache management policy.

Results: 1. Hybrid caching architecture is introduced which combine the advantages of both hierarchical and distributed caching. 2.Between browser clients and Web servers, between proxies and Web servers and between browser clients and proxies are three ways prefetching can do in the Web contexts. The first two approaches however have the probabilities of increasing wide area network traffic, while the last one simply affects the traffic over the modems or the LANs, which is kind of advantage in this field. All of these approaches are trying to prefetch either documents that are considered as popular at servers or documents that are predicted to be accessed by user in the near future based on the access pattern. 3. Traditional replacement policies and its direct extensions, key-based replacement policies and cost-based replacement policies are three categories of cache placement/replacement. The performance of replacement policies depends highly on traffic characteristics of WWW accesses. Hence there is no policy that can suit every situation for all Web access patterns. 4. About cache coherency, we have the strategies of HTTP commands that assist Web proxies in maintaining cache coherence and cache coherence mechanisms to provide users with fresh pages. 5. Cache proxy is an effective mechanism to improve the Web performance with proxy can serve in data cache, connection cache and computation cache. 6. There are two approaches in predicting users' future requests. 7. The solution to hot spots problem is to store copies of hot pages/services through the internet and this can spread the work of serving a hot page/services across many servers. 8. The placement of proxy is also important 9. Active cache and web server accelerator can improve the Web performance. 10. The effectiveness of caching theory relies on the existing Web reference streams and the proper use of cache management policy.

Conclusion: Users are facing the problem of network congestion and server overloading. The focus on improving Web performance never end. Web caching is an effective way to alleviate bottleneck and reduce network traffic, which can minimize the user access latency. Some current Web caching schemes are presented in this paper. However, problems still exist in Web caching and more future works should be made to improve the performance of Web.