Chapter 10

Performance-Enhanced Caching Scheme for Web Clusters for Dynamic Content

A. Raghunathan
Bharat Heavy Electricals Limited, India

K. Murugesan
National Institute of Technology, Tiruchirappalli, India

ABSTRACT

In order to improve the QoS of applications, clusters of web servers are increasingly used in web services. Caching helps improve performance in web servers, but is largely exploited only for static web content. With more web applications using backend databases today, caching of dynamic content has a crucial role in web performance. This paper presents a set of cache management schemes for handling dynamic data in web clusters by sharing cached contents. These schemes use either automatic or expiry-based cache validation, and work with any type of request distribution. The techniques improve response by utilizing the caches efficiently and reducing redundant database accesses by web servers while ensuring cache consistency. The authors present caching schemes for both horizontal and vertical cluster architectures. Simulations show an appreciable performance rise in response times of queries in clustered web servers.

INTRODUCTION

Web services have become an important component of business and enterprise operations. In order to meet the current need for their improved availability, scalability and performance, clusters of web servers are increasingly being employed in many installations (Hyun et al., 2003; Sharifian et al., 2009, 2010). Caching of data in a web server provides a further improvement in the application’s response time (Adams, 2004; Anbazhagan & Nagarajan, 2002) but till recently caching has
largely been exploited only for static web content. Today, however, most web applications serve up dynamic content to the clients by acquiring data from backend databases (Challenger et al., 2004). While caching of this dynamic data, wherever possible, does improve the response times of the servers (Holmedahl et al., 1998), the need for maintaining cache currency with the database is a problem that needs to be addressed, and its complexity increases when caching is employed in multiple servers within the cluster.

When requests from clients for web pages are distributed across a cluster, different requests for a given page may be directed to different web servers at different times. Each server uses its own cache to respond to this request, accessing the cache immediately without re-querying the database. In case of static content the data does not change and so the cached result would be valid to answer subsequent requests. But if the content is dynamic, it becomes necessary to maintain data consistency among all the caches in the cluster, since each server’s cache is independent, to ensure that the cluster members do not return different results for the same request. In addition, visiting the database by several servers to fetch the same data is inefficient. There is a wide scope for further work on caching of dynamic content in web clusters towards improving QoS of Web services. In this paper we present a set of cache management schemes to improve the overall performance of a web cluster by sharing cached contents with the help of the front-end web switch. Our methods differ from earlier shared caching approaches (Chen et al., 2003; Cuenca-Acuna et al., 2001; Holmedahl et al., 1998). Our schemes utilize the server caches efficiently, reducing redundant database accesses by web servers while also ensuring consistency of cache contents. The proposed techniques could be used with any algorithm used for request distribution within a cluster.

Two principal methods are generally used for cache validation of dynamic data: automatic and expiry-based. The former is preferred for frequently changing data while the latter is employed when the data does not change frequently or when data does not have to be refreshed to the client whenever it changes. We present our caching algorithms for both these methods.

Clusters are also classified in two ways: horizontal and vertical (Turaga et al., 2006). Constituent servers in a horizontal cluster reside on different physical machines; in a vertical cluster they reside on a single physical machine. We propose our caching schemes for both types of clusters. Our simulations show an appreciable performance increase in all cases.

CACHING OF DYNAMIC DATA IN WEB SERVERS

Cache Revalidation

Since dynamic web content is usually treated as non-cacheable, caching has largely been confined to static web content. However, accessing the backend servers is expensive in terms of time and resources and should be preferred only for frequently changing data. Even on many web sites that display dynamic content, the backend data either does not change often, or when it does, it may not be necessary to display the data on every change; some latency in data refreshes may be tolerable. Examples of such cases include data from product catalogs, vendor information, customer information, archived data, etc. In such cases it would make sense to respond with cached data rather than fetch it from the backend every time.

Web queries that fetch content from a backend data source are fired by the clients usually in the form of URLs with or without parameters either as hyperlinks or through forms. The web (application) server that receives a query re-transmits it to the backend, fetches the query results and stores the results in its cache (Figure 1). It then formats the results as a HTML web page and serves it to the client for display on the browser. When a client invokes the same query again, the (web) application server attempts to verify that