Chapter 96
Speeding up the Internet:
Exploiting Historical User Request Patterns for Web Caching

Chetan Kumar
California State University San Marcos, USA

INTRODUCTION

The Internet has witnessed a tremendous growth in the amount of available information, and this trend of increasing traffic is likely to continue. According to a Cisco Systems forecast report (2008) the growth in Internet traffic is to be driven by Web 2.0 technologies such as video and social networking and collaboration. Some excerpts of the Cisco forecast report (2008) are as follows.

- "Global Internet Protocol (IP) traffic will increase by a factor of six from 2007 to 2012, reaching 44 exabytes per month in 2012, compared to fewer than 7 exabytes per month in 2007.
- Total IP traffic for 2012 will amount to more than half a zettabyte (or 522 exabytes). A zettabyte is a trillion gigabytes.
- Monthly global IP traffic in December 2012 will be 11 exabytes higher than in December 2011, a single-year increase that will exceed the amount by which traffic increased in the eight years since 2000" (Cisco forecast report 2008).

Despite technological advances this traffic increase can lead to significant user delays in web access (Datta et al. 2003, Mookherjee and Tan 2002, Watson et al. 1999). Web caching is one approach to reduce such delays. Caching involves temporary storage of web object copies at locations that are relatively close to the end user. As a result user requests can be served faster than if they were served

DOI: 10.4018/978-1-61520-611-7.ch096

directly from the origin web server (Hosanagar and Tan 2004, Davison 2007).

Caching can be performed at different levels in a computer network. Proxy caches are situated at computer network access points for web users (Davison 2007). Other locations where caching may be performed include browser and web-server levels (Davison 2001, Kumar and Norris 2008). Proxy caches can store copies of web objects and directly serve requests for them in the network, consequently avoiding repeated requests to origin web servers. As a result there is reduced network traffic, load on web servers, and average delays experienced by web users (Cao and Irani 1997, Datta et al. 2003). Kumar (2009) illustrate the benefit of a network of proxy caches using an example of the IRCache network (www.ircache.net). Figure 1 shows how a network of proxy caches with nodes at three locations can reduce user delays. If the U.K. node has requests for web pages chrysler.com, ford.com, and mercedes-benz.com, that it has not cached, then these requests can be satisfied from the U.S. and Germany nodes. Therefore the U.K. node need not go to the origin web server to satisfy requests for objects it does not hold itself but are held by neighbor caches. Since origin server requests typically have the longest waiting times, by reducing them proxy caches can significantly reduce network delays (Kumar 2009). Proxy caching is widely used by computer network administrators and technology providers (Davison 2007). Examples include proxy caching solution providers such as Oracle (www.oracle.com/technology/products/ias/web_cache/index.html), content delivery network (CDN) firms such as Akamai (www.akamai.com), and Internet service providers (ISP) such as AT&T (www.att.com). The following are two illustrations, adapted from Davison (2007), of how some firms may practically benefit from caching. In one case a company such as Intel may employ a proxy cache near its network gateway to serve its many users (e.g., clients within Intel) with cached objects from many servers. As a result Intel reduces the bandwidth required over expensive dedicated Internet connections. In another scenario a content provider such as Yahoo can place a proxy cache directly in front of a particular server to reduce the number of requests that the server must handle. This service to speed up content delivery, also called reverse caching as a proxy node may cache objects for many clients but from usually only one server, is professionally provided by CDN firms such as Akamai. In both scenarios access delays are reduced thereby benefitting all Internet users (Davison 2007). Of course in choosing caching solutions, as in any IT investment decision, firms have to evaluate costs of an implementation versus its benefit, before deciding on the appropriate caching service. In this article we discuss some proxy caching approaches that exploit historical user request patterns to reduce user request delays (Kumar and Norris 2008, Zeng et al. 2004).

RELATED LITERATURE AND BACKGROUND

There is a growing interest in caching due to its application in reducing user delays while accessing the increasingly congested Internet (Datta et al. 2003, Davison 2007). Podlipnig and Boszormenyi (2003), Zeng et al. (2004), and Datta et al. (2003), provide an extensive survey of numerous caching techniques. These include popular cache replacement strategies such as least recently used (LRU), where the least recently requested object is evicted from the cache to make space for a new one, and their many extensions. While most caching studies focus on improving performance on measures such as user latency and bandwidth reduction, there have been relatively few studies that consider a data or model driven approach for managing caches. Cockburn and Mckenzie (2002) and Tauscher and Greenberg (1997) study client–side behavior on the Internet. They note that the