Integrating Caching Techniques in CDNs using a Classification Approach

George Pallis, University of Cyprus, Cyprus
Konstantinos Stamos, Aristotle University of Thessaloniki, Greece
Athena Vakali, Aristotle University of Thessaloniki, Greece
Charilaos Thomos, Aristotle University of Thessaloniki, Greece
George Andreadis, Aristotle University of Thessaloniki, Greece

ABSTRACT

Content Delivery Networks (CDNs) provide an efficient support for serving “resource-hungry” applications while minimizing the network impact of content delivery as well as shifting the traffic away from overloaded origin servers. However, their performance gain is limited since the storage space in CDN’s servers is not used optimally. In order to manage their storage capacity in an efficient way, we integrate caching techniques in CDNs. The challenge is to decide which objects would be devoted to caching so as the CDN’s server may be used both as a replicator and as a proxy server. In this article, we propose a nonlinear, nonparametric model which classifies the CDN’s server cache into two parts. Through a detailed simulation environment, we show that the proposed technique can yield significant reduction in user-perceived latency as compared with other heuristic schemes.

Keywords: Distributed Systems; Neural Networks; Web-Based Applications

INTRODUCTION

With the enormous growth of Web traffic on the Internet, it is essential that Web’s scalability and performance keep up with the increasing demands and expectations. On a daily basis, clients use the Internet for “resource-hungry” applications which involve content such as video, audio on-demand, and distributed data. For instance, the Internet video site YouTube hits more than 100 million videos per day. Estimations of Youtube’s bandwidth go from 25TB/day to 200TB/day. At the same time, more and more Web content servers are delivering greater volumes of content but with high
sensitivity to delays. For instance, a delay on a financial data-feed Web site (e.g., USD to EUR currency stock markets) may cause serious problems to the end-users. Thus, the efficient and scalable content delivery on the Web remains a challenge. Which technologies could meet the above challenge? The answer to this question lies in combining replica placement and caching techniques in Content Delivery Networks (CDNs) (Bakiras & Loukopoulos, 2005; Stamos, Pallis, & Vakali, 2006).

The key idea behind caching is to keep content close to the end-user according to a cache replacement policy. Specifically, the end-user’s request for an object is posed to a proxy server, which may contain a cached version of the object. If the proxy server contains a “fresh copy” of the requested object (cache hit) then the end-user receives it directly from the proxy cache, elsewhere (cache miss) the end-user is redirected to the origin server (where the Web site is located). Therefore, both the bandwidth consumption and the network traffic are reduced. Additionally, network availability is significantly improved since the end-user may receive a copy even if the origin server is unavailable. Another advantage of caching is that fresh content is added into the caches leading to better storage usage. A complementary to caching technique is prefetching (Sidiropoulos et al., 2008). Prefetching is proposed to find meaningful object access patterns in order to predict future requests. Therefore, objects may be transferred to the proxy server a priori (before they are even requested).

As far as the replication approach is concerned, its main idea is to bring static content replicas close to the end-user. This is currently applied in the CDNs (Bent, Rabinovich, Voelker, & Xiao, 2006; Sidiropoulos et al., 2008). A typical CDN is depicted in Figure 1. A CDN consists of a set of surrogate servers geographically distributed in the Web, which contain copies (replicas) of content belonging to the origin server (according to a specific storage capacity). Therefore, CDNs act as a network layer between the origin server and the end-users, for handling their requests. With this approach, content is located near to the end-user, yielding low response times and high content availability since many replicas are distributed. The origin server is “relieved” from requests since the majority of them are handled by the CDN, whereas Quality of Service (QoS) and efficiency are guaranteed in a scalable way. Finally, an important characteristic of the CDNs is the efficiency against flash crowd events. Specifically, a flash crowd event occurs when unpredictably numerous users access a Web site (i.e., Sept. 11th, Tsunamis, etc.).

Caching and replication deal with situations as separate approaches. Caching is mainly addressed to proxy servers, whereas replication is based on CDNs. However, the content replication practice of CDN includes inherent limitations. The major limitation is that CDN infrastructure does not manage in an efficient way the replicated content. The replicas’ placement is static for a considerable amount of time. The static nature of the outsourced content leads to inefficient storage capacity usage since the cache of surrogate servers after a period of time may contain unnecessary objects. As a result, if user access patterns change, the replicas in surrogate servers could not satisfy the future users’ requests. A solution to the above issue would be to integrate both caching and replication policies to the storage space of CDN surrogate servers. The experimental results in Stamos, Pallis, and Vakali (2006) showed that a scheme which combines caching and replication outperforms the stand-alone Web caching and static content replication implementations.

In this article, which extends an earlier work (Pallis et. al., 2007), we focus on deploying a new scheme for CDNs taking advantages of caching and replication. Specifically, we consider a CDN whose surrogate servers act simultaneously both as proxy servers and content replicators. The major contributions of this work are:

• Proposing a CDN framework where the surrogate servers act both as proxy caches
Opportunities and Limitations in M-Commerce
www.igi-global.com/chapter/opportunities-limitations-commerce/31436?camid=4v1a