Chapter 1

Introduction

1.1 Research Background
1.2 Thesis Contributions
1.3 Thesis Organization

1.1 Research Background

In recent years the Internet and the World Wide Web have exploded in popularity. The Internet has become an important source of both information and entertainment for people all over the world. The phenomenal growth has been spurred on by the increasing ease of use of browsers and more appealing content such as image, audio and video.

This ever-increasing demand brings increased pressure to the Internet to provide fast and reliable content delivery to Web users, as site performance can strongly impact a content provider's bottom line in which consumers are more likely to visit and/or purchase from sites that load quickly, reliably and consistently. Moreover, the increased use of rich-media content (consisting of image, audio and video) puts a huge load on the storage and network infrastructure. Therefore, how to efficiently deliver Web content from the original server to the end users has been a critical issue for the development of the Internet.

Web content delivery attracted many researchers and projects. Client-side proxy caching is a basic client-server model that locally installs caching proxies at ISPs or institutions. But it doesn’t allow the content provider any control over how the content would be cached. A new content distribution paradigm, Content Distribution Networks, emerged to remedy this problem and they have become an alternative content distribution method for some commercial Web content. Recently a Peer-to-Peer Network was also proposed where each peer in the network is both a client and a server. However, because rich-media content has their distinct statistical properties and user viewing patterns, current delivery and caching schemes for normal Web objects such as HTML files can still not be efficiently applied to Web contents such as image, audio and video.

1.2 Thesis Contributions

This thesis makes several contributions. The first contribution (Chapter 3) is about an integrated pre-fetching and replacing algorithm for Graceful Image Caching. As the Web workload characteristics show that more than 60% of network traffic is caused by image documents, how to efficiently deliver these images is one of the most critical issues. Proxy caches are commonly introduced between servers and clients to reduce data traffic and improve response time in the Internet. However, current system employs a “hard” caching strategy: an image is either stored in a cache or not ever if its size is very large. It has been shown that an image caching method (Graceful Caching) based on hierarchical coding format can obtain better performance than conventional caching schemes. However, as the capacity of the cache is limited, how to efficiently allocate the cache memory to achieve a minimum expected delay time is still a problem to be resolved. In chapter 3 we present an integrated caching algorithm to deal with the above problem for image databases, Web browsers, proxies and other similar applications in the Internet. By analyzing the Web request distribution of Graceful Caching, both replacement and pre-fetching algorithms are proposed. We also show that our proposal could be carried out based on information readily available in the proxy server. It flexibly adapts its parameters to the hit rates and access pattern of users’ requesting documents in the Graceful Caching. We quantitatively evaluate our proposal when the related parameters are changed, and compare our proposal with other existing ones. The results show that the proposed algorithm outperforms all of the previous ones.

The second contribution (Chapter 4) is a novel mechanism to deliver and cache streaming media by using segment-based caching and hierarchically distributed proxies. With the wide availability of high-speed networks, an increasing number of streaming media objects are being distributed over the Internet. Because of their distinct statistical properties and user viewing patterns, traditional delivery and caching schemes for normal Web objects such as HTML files or images can not be efficiently applied to streaming media such as audio and video. Content delivery networks (CDN) appeared recently and are being deployed quite rapidly. However, their concern is mainly placed on efficient delivery of normal Web content. Some CDN companies advocate their streaming caching support, but their technical details are not yet clarified nor verified. In chapter 4 we therefore propose an integrated caching scheme for streaming media

with segment-based caching and hierarchically distributed proxies. Firstly, because storing the entire stream in a single proxy cache is inefficient or even impossible due to its large size, different segment-based caching algorithms are proposed and compared. In our proposal, a part of the requested stream is cached in a local cache, and the remainder of the stream will be cached in an upper proxy cache. Secondly, because each different stream has a different popularity and each segment has different access patterns, we propose two kinds of replacing algorithms to decide which segments of which streams should be removed when the cache exceeds its limit. One of these two algorithms keeps the same relative length of each stream in the cache, while the other keeps the most accessed segments in the cache. Finally, how to coordinate the streaming caching with the current caching scheme for normal Web objects such as HTML files or images is considered. By introducing the EWMA (Exponential Weighted Moving Average) estimator, a Web-friendly caching scheme, which can improve the performance of the whole caching system, is proposed. We verify our proposals by simulations and the results show that the performance of contents delivery over the Internet can be effectively improved.

Our third contribution (Chapter 5) is a new replication algorithm for streaming media in CDN. As we all know, CDN improves end-user performance by replicating Web contents on a group of geographically distributed servers called content servers which are located closer to users. However, how to replicate Web contents to different content servers is still a main problem. Current replica strategies in CDN are to simply and repeatedly keep the same replicas of the original objects on many content servers. Disadvantages of the current method are as follows: on one hand, to repeatedly store the same large size objects such as streaming data into different content servers consumes too more server space since the stream has large size. On the other hand, as some content servers are not always requested by the clients, the waste of the storage cost is increased. It is more serious for replicating some large-sized objects such as streaming media, which are distributing over the Internet more and more. Therefore, to improve the user response time and storage cost, in chapter 5 we propose a new algorithm to replicate streaming media by using hierarchical streaming format. We give a mathematical analysis about the average number of the inter-AS hops that a request must traverse.Then, we propose a new replica algorithm which can efficiently decrease user response and storage cost. Simulation results show that our proposal achieved better performance than other previous ones.

1.3 Thesis Organization

This thesis consists of an introductory chapter (Chapter 2) which presents an overview of content distribution technologies, three research chapters (Chapter 3~5) and a concluding chapter (Chapter 6).

In Chapter 3, we focus on integrated pre-fetching and replacing algorithm for Graceful Image Caching. The theory of Web access is firstly presented. Then, the related replacing and pre-fetching algorithms are shown respectively. Finally, we introduce how to get the necessary parameters for our proposal and simulation results are also given.

A novel mechanism to deliver segmented streaming media over hierarchically distributed proxies is presented in Chapter 4. We first introduce three caching algorithms to just keep a part of each stream in the local proxies. Secondly, the corresponding caching algorithms are given to decide how to select one segment to be removed from a cache when the local cache is full. Finally, we show the proposed Web-friendly caching scheme and all simulation results.

In chapter 5 we talk about a new replication algorithm for streaming media in the CDN. The characteristics of Scalable Streaming are presented first. Then, a mathematics analysis of the traversed AS hops among the cooperative CDN servers is given. We present our proposed algorithm and evaluate its performance through extensive experiments at last.

We summarize the results of this thesis and future work in Chapter 6.