Chapter XXIII

Stream Dependent Caching

Introduction

Caching has been successfully implemented on the Internet to reduce workload on the content server and the Internet. We have seen in the last chapter how the cache replacement methods are adapted for multimedia objects in memory caching. In this chapter, we shall show how the caching is tailored to provide better performance for continuous request streams.

Even though caching reduces the access latency when there are cache hits, there are chances that cache misses occur. When cache misses occur, the request stream is sent through the network to the remote storage devices. The requests are then served at the remote storage devices. The requested multimedia objects are retrieved from the storage devices, delivered through the network to the client. The cache content will also be modified to store the accessed object.

Copyright © 2008, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
Multimedia data requests are continuously sent to the remote storage devices. Each request may ask for only a small part of data. The union of all the requested data is the entire object. In order to provide continuous display of media object for a period of time, the storage system needs to provide a guarantee on the continuous delivery of data (Chae et al., 2002; Chang & Hock, 2000).

Although caching increases the service rate of data requests, it is inevitable that some misses occur. When the cache hit ratios are low, the workload on the remote storage devices becomes heavy. When the workload on the storage device is too heavy, response time and access delay of the requests could increase indefinitely. As a result, the data cannot be retrieved within the guarantee time. This results in violations of the continuous display guarantee.

In order to provide continuous display guarantee of multimedia information, the requested multimedia data must be delivered continuously. However, this cannot be easily achieved on today’s Internet. Congestions in the network could also hinder the smooth delivery of data. Unfortunately, the Internet is designed and implemented in a way that congestions cannot be completely avoided. It may be a fact that congestions persistently occur when the stream is running for a long enough time.

Many methods to provide continuous multimedia streams have been proposed and investigated. However, the implementation of these techniques on the Internet still has some difficulties due to the presence of legacy routers.

As multimedia objects are large in size, the limited memory cache space can only store a few objects. If all the accessed objects are of the same size, the size aware cache replacement methods would not increase the number of objects being cached. In this situation, the cache hit ratio is still constrained by the size of the local cache.

As the multimedia objects are so large, it becomes necessary to create the cache level on local disks, instead of the random access memory. With a bigger cache space, the cache level on disks can reduce more capacity misses. However, the cache level on the disks must be created carefully. As the service time of disk requests is rather long, the disk throughput is limited. Thus, the disk throughput should be higher than the data rate of the objects so that the objects on the cache are accessible.

If the workload is too high for an individual disk, multiple disks or disk array may be used. In such condition, the workload of the disks should be well balanced. Balanced disk load can avoid bottlenecks to build up and overload individual disk.
Building Multi-Modal Relational Graphs for Multimedia Retrieval
www.igi-global.com/article/building-multi-modal-relational-graphs/54460?camid=4v1a