Chapter XXII

Memory Caching Methods

Introduction

The objective of data caching and object caching is to improve the performance in accessing multimedia objects from their storage. An efficient cache storage method can have many benefits:

First, caching can increase server capacity in serving more streams. Since the cache may satisfy some requests before they arrive at the server, the server will receive fewer requests to its data. Thus, the workload on the data server can be reduced. As a result, the server can serve more streams.

Second, it can reduce access latency when a recently accessed object is being accessed again. When an object is accessed again, the first copy of the accessed object in the cache can already satisfy the request. Since the cache can often access objects with smaller delay, the latency on accessing the objects is thus smaller.
Third, caching can also reduce network bandwidth need when a request can be served locally by the local cache. If the cache is on the hard disk within local area network, the request may be served without sending the requests outside the local area network. Thus, the network traffic may be reduced.

Fourth, caching may be used to balance the workload among cache devices by directing the request to different cache devices evenly. If the cache servers containing the same cache copies have different workloads, the request can be directed to the server with the lightest workload. Therefore, the workload among the cache devices can be balanced.

In order to achieve memory caching, there must be at least two or more different storage levels. Each storage level consists of storage devices that can store data for later retrievals. The local cache storage level is closer to the client or the source of request. The data at this cache level can be accessed faster or with lower delay. The remote cache storage level is closer to the destination or the source of data objects. It takes longer time to access data from this storage level than the local cache level.

When a request for data is being served, the local cache is first searched to find if a copy of the required data object exists. If such a data copy can be found, the validity of the data copy is then checked. If the data copy is valid, then the condition is considered a hit. The data copy in the local cache level is then delivered to satisfy the request.

If the data copy cannot be found or the data copy in the local cache is invalid, then a cache miss occurs. The request is then forwarded to the remote storage level to be served. The request is then served at the remote level and a data copy is returned; the data copy is then stored onto the local cache level according to the cache admission policy.
Related Content

A Taxonomy of Database Operations on Mobile Devices
www.igi-global.com/chapter/taxonomy-database-operations-mobile-devices/20957?camid=4v1a

www.igi-global.com/chapter/improved-spatial-temporal-moving-object-detection-method-resistant-to-noise/94124?camid=4v1a

Multi-Label Classification Method for Multimedia Tagging
Aiyesha Ma, Ishwar Sethi and Nilesh Patel (2010). International Journal of Multimedia Data Engineering and Management (pp. 57-75).
www.igi-global.com/article/multi-label-classification-method-multimedia/45755?camid=4v1a
A Real-Time 3D Visualization Framework for Multimedia Data Management, Simulation, and Prediction: Case Study in Geospatial-Temporal Biomedical Disease Surveillance Networks