Cache Invalidation in a Mobile Environment

Say Ying Lim
Monash University, Australia

INTRODUCTION

The rapid development, as well as recent advances in wireless network technologies, has led to the development of the concept of mobile computing. A mobile computing environment enables mobile users to query databases from their mobile devices over the wireless communication channels (Cai & Tan, 1999). The potential market for mobile computing applications is projected to increase over time by the currently increasingly mobile world, which enables a user to satisfy their needs by having the ability to access information anywhere, anytime. However, the typical nature of a mobile environment includes low bandwidth and low reliability of wireless channels, which causes frequent disconnection to the mobile users. Often, mobile devices are associated with low memory storage and low power computation and with a limited power supply (Myers & Beigl, 2003). Thus, for mobile computing to be widely deployed, it is important to cope with the current limitation of power conservation and low bandwidth of the wireless channel. These two issues create a great challenge for fellow researchers in the area of mobile computing.

By introducing data caching into the mobile environment, it is believed to be a very useful and effective method in conserving bandwidth and power consumptions. This is because, when the data item is cached, the mobile user can avoid requests for the same data if the data are valid. And this would lead to reduced transmissions, which implies better utilization of the nature of the wireless channel of limited bandwidth. The cached data are able to support disconnected or intermitted connected operations as well. In addition, this also leads to cost reduction if the billing is per KB data transfer (Lai, Tari, & Bertok, 2003). Caching has emerged as a fundamental technique especially in distributed systems, as it not only helps reduce communication costs but also offloads shared database servers. Generally, caching in a mobile environment is complicated by the fact that the caches need to be kept consistent at all time.

In this article, we describe the use of caching that allows coping with the characteristics of the mobile environment. We concentrate particularly on cache invalidation strategy, which is basically a type of caching strategy that is used to ensure that the data items that are cached in the mobile client are consistent in comparison to the ones that are stored on the server.

BACKGROUND

Caching at the mobile client helps in relieving the low bandwidth constraints imposed in the mobile environment (Kara & Edwards, 2003). Without the ability to cache data, there will be increased communication in the remote servers for data and this eventually leads to increased cost and, with the nature of an environment that is vulnerable to frequent disconnection, may also lead to higher costs (Leong & Si, 1997). However, the frequent disconnection and the mobility of clients complicate the issue of keeping the cache consistent with those that are stored in the servers (Chand, Joshi, & Misra, 2004).

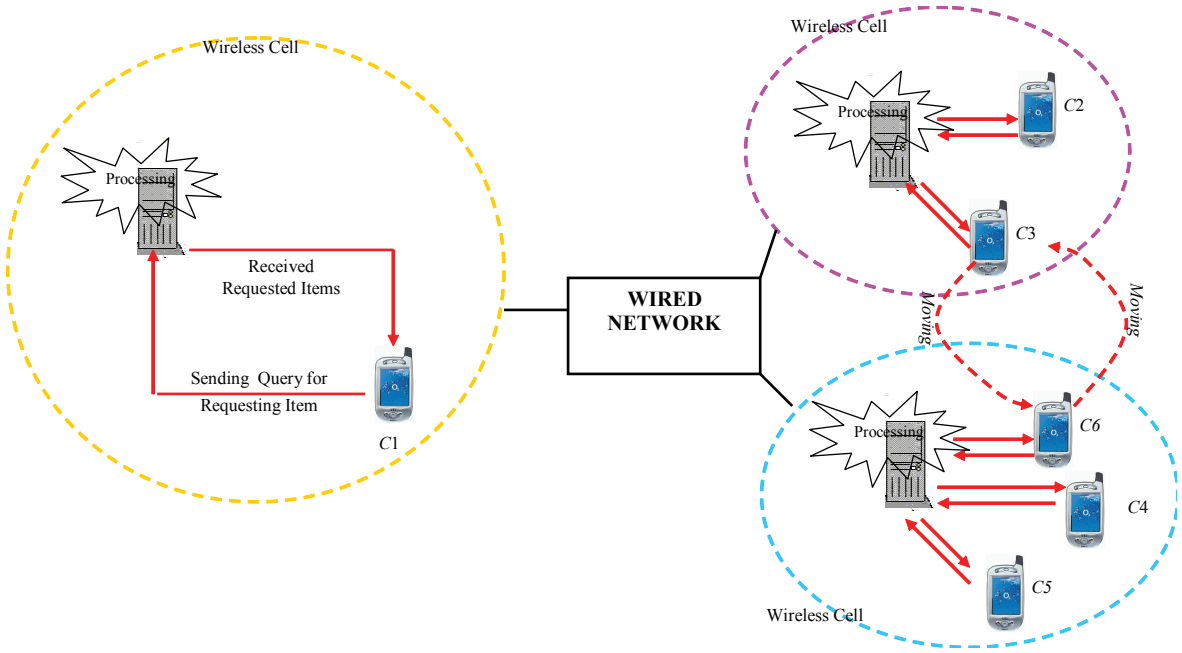
Thus, when caching is used, ensuring data consistency is an important issue that needs considerable attention at all times (Lao, Tari, & Bertok, 2003). This is because the data that has been cached may have been outdated and no longer valid in comparison to the data from the corresponding servers or broadcast channel.

Figure 1 shows an illustration of a typical mobile environment that consists of mobile clients and servers, which are also know as mobile host (MH) and mobile support system (MSS) respectively. The mobile clients and servers communicate via a wireless channel within a certain coverage, known as cell (Chand, Joshi, & Misra, 2003; Cai & Tan, 1999). There are two approaches for sending a query in a mobile environment, which are: (a) The mobile clients are free to request data directly from the server via the wireless channel and the server will process and pass the desired data items back and (b) the mobile clients can tune into the broadcast channel to obtain the desired data items and download it to his/her mobile device. This can be illustrated in Figure 1a and Figure 1b respectively. The assumption is that updates are only able to occur at the server side and mobile clients can only have a read only feature.

CACHE INVALIDATION

Due to the important issue in the mobile environment, which is the ability to maintain data consistency, cache invalidation strategy is of utmost significance to ensure that the data items cached in the mobile client are consistent with those that are stored on the server. In order to ensure that data that are about to be used is consistent, a client must validate its cache prior to using any data from it.

Figure 1. Mobile environment architecture



There are several distinctive and significant benefits that cache invalidation brings to a mobile computing environment. If cache data are not validated to check for consistency, it will become useless and out-of-date. However, if one can utilize the cache data then the benefits it may bring include energy savings—that is, by reducing the amount of data transfer—and in return result in cost savings.

Using Cache Invalidation in a Mobile Environment

This can be done by using the broadcasting concept in communicating cache validation information to mobile clients. The server broadcasts the cache information, which is known as cache invalidation report (IR), periodically on the air to help clients validate their cache to ensure they are still consistent and can be used. It appears that the broadcast mechanism is more appropriate for the mobile environment due to its characteristic of salability, which allows it to broadcast data to an arbitrary number of clients who can listen to the broadcast channel anytime (Lai, Tari, & Bertok, 2003). By using the broadcasting approach, whereby the server periodically broadcasts the IR to indicate the change data items, it eliminates the need to query directly to the server for a validation cache copies. The mobile clients would be able to listen to the broadcast channel on the IR and use them to validate their local cache respectively (Cao, 2002).

Although cache invalidation strategy is important in a mobile environment, it will be vulnerable to disconnection and the mobility of the clients. One of the main reasons that cause mobile clients frequent disconnection is the limited battery power, and that is why mobile clients often disconnect to conserve battery power. It may appear to be very expensive at times to validate the cache for clients that experience frequent disconnection, especially with narrow wireless links. Other drawbacks would include long query latency, which is associated with the need of the mobile client to listen to the channel for the next IR first before he is able to conclude whether the cache is valid or not before answering a query. Another major drawback is the unnecessary data items in the IR that the server keeps. This refers to data items that are not cached by any mobile clients. This is thereby wasting a significant amount of wireless bandwidth.

Example 1: A mobile client in a shopping complex denoted as *C*1 in Figure 2 wanted to know which store to visit by obtaining a store directory. The client has previously visited this store and already has a copy of the result in his cache. In order to answer a query, the client will listen to the IR that are broadcasted and use it for validation against its local cache to see if it is valid or not. If there is a valid cached copy that can be used in answering the query, which is getting the store directories, then the result will be returned immediately. Otherwise, if the store directories have changed and now contain new shops, then the invalid caches have to be refreshed via sending a query to the server (Elmagarmid et al., 2003). The server would keep track of the recently updated data and broadcast the up-to-date IR every now and