P

Peer-to-Peer Cooperative Caching in Mobile Environments

Chi-Yin Chow
University of Minnesota – Twin Cities, USA

Hong Va Leong
The Hong Kong Polytechnic University, Hong Kong

Alvin T. S. Chan
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

An infrastructure-based mobile environment is formed with a wireless network connecting mobile hosts (MHs) and mobile support stations (MSSs). MHs are clients equipped with portable devices, such as laptops, personal digital assistants, cellular phones, and so on, while MSSs are stationary servers providing information access for the MHs residing in their service areas. With the recent widespread deployment of contemporary peer-to-peer (known as P2P throughout this chapter) wireless communication technologies, such as IEEE 802.11 (IEEE Standard 802-11, 1997) and Bluetooth (Bluetooth SIG, 2004), coupled with the fact that the computation power and storage capacity of most portable devices have been improving at a fast pace, a new information sharing paradigm known as P2P information access has rapidly taken shape. The MHs can share information among themselves rather than having to rely solely on their connections to the MSS. This article reviews a hybrid communication framework—that is, mobile cooperative caching—which combines the P2P information access paradigm into the infrastructure-based mobile environment.

BACKGROUND

In mobile environments, there are two different types of communication architecture, *infrastructure* and *ad hoc based.* The infrastructure-based mobile communication architecture is formed with MHs and MSSs. The MHs can only retrieve their desired data items from MSSs, either by requesting them over shared point-to-point channels (*pull-based data dissemination model*) or catching them from scalable broadcast channels (*push-based data dissemination model*) or through the utilization of both types of channels (*hybrid data dissemination model*). This type of communication architecture is the most commonly deployed one in real life.

The emergence of the state-of-the-art P2P communication technologies leads to the development of an ad-hoc-based mobile communication architecture, also known as a *mobile ad-hoc network* (MANET). In MANETs, the MHs can share information among themselves without any help of MSSs. This kind of sharing paradigm is also referred to as a P2P *data dissemination model.*

In a pull-based environment, the MHs have to retrieve their desired data items from the MSS whenever they encounter local cache misses. Since the mobile environment is characterized by limited bandwidth, the communication channel between the MSS and the MHs would potentially become a scalability bottleneck in the system, as it serves an enormous number of MHs. Although push-based and hybrid data dissemination models are scalable, the MHs adopting these two models generally suffer from longer access latency and higher power consumption than those adopting the pull-based one, as they need to tune in to the broadcast channel and wait for the broadcast channel index or their desired data items to appear. Furthermore, since the data items are broadcast sequentially, the MHs experience longer access latency with an increasing number of data items being broadcast.

MANET is practical to a mobile system with no fixed infrastructure support, such as battlefield, rescue operations, and so on (Fife & Gruenwald, 2003). However, it is not suitable for commercial mobile applications. In MANETs, the MHs can rove freely and disconnect themselves from the network at any instant. These two particular characteristics lead to dynamic changes in the network topology. As a result, the MHs could suffer from long access latency or access failure when the peers holding the desired data items are far way or unreachable. The latter situation is caused by network partitioning (Wang & Li, 2002) or client disconnection.

The inherent shortcomings of the infrastructure- and ad-hoc-based communication architecture lead to a result that a mobile application adopting either one of these architectures alone would not be as appropriate in most real commercial settings. In reality, long access latency or access failure could possibly cause the abortion of valuable transactions or the suspension of critical activities, so that it is likely to reduce user satisfaction and loyalty, and potentially bring damages to the organization involved. The drawbacks of the existing mobile data dissemination models motivate researchers to develop a novel hybrid communication framework—mobile cooperative caching—in which a convectional infrastructure-based mobile communication framework is used in combination with a P2P data dissemination paradigm for deploying mobile information access applications in reality.

MOBILE COOPERATIVE CACHING

Recently, mobile cooperative caching has been drawing increasing attention. Several mobile cooperative caching schemes were proposed during the preceding years. These works can be divided into two major categories: *cooperative data dissemination* and *cooperative cache management.* The work of cooperative data dissemination (Lau, Kumar, & Venkatesh, 2002; Papadopouli & Schulzrinne, 2001; Sailhan & Issarny, 2003; Shen, Das, Kumar, & Wang, 2004) mainly focuses on designing protocols for the MHs to search their desired data items and forward the data items from source MHs or MSSs to them in a mobile environment. The work pertaining to cooperative cache management focuses on designing protocols and algorithms for the MHs to manage their cache space, not only with respect to themselves, but also with respect to their peers, in order to improve system performance along such design dimensions as *cooperative data replica allocation* (Hara, 2001, 2002a, 2002b; Hara, Loh, & Nishio, 2003), *cooperative cache invalidation* (Hayashi, Hara, & Nishio, 2003), and *cooperative cache admission control* and *cache replacement* (Lim, Lee, Cao, & Das, 2003; Chow, Leong, & Chan, 2004, 2005).

COOPERATIVE DATA DISSEMINATION

Sailhan and Issarny (2003) propose an intuitive cooperative data dissemination scheme for a MANET environment. If an MH can directly connect to an MSS, it would obtain the required data items from the MSS; otherwise, the MH has to enlist its peers at a distance less than the MSS for help to turn in the required data items. If no such peer caches the data items, the peers route the request to the nearest MSS. A local cache replacement strategy is also proposed for the MH based on the access probability and time-to-live of the cached data items, and the estimated energy cost of retrieving them.

A similar cooperative data dissemination scheme is designed to support continuous media access in MANETs (Lau et al., 2002). Two data location schemes, namely *cache-state* and *reactive,* are proposed for the MHs to determine the nearest data source that can be either the cache of their peers or the original servers to retrieve their desired multimedia objects. Cache-state is a proactive scheme, whereas reactive is an on-demand scheme. The performance evaluation result shows that the reactive scheme outperforms the cache-state one in terms of network traffic, quality of service (QoS), and access latency.

7DS (seven degrees of separation) (Papadopouli & Schulzrinne, 2001) is another cooperative data dissemination scheme that is used as a complementary component to the infrastructure support with power conservation. When an MH fails to connect to the MSS to retrieve its desired data items, it would attempt to search its neighboring 7DS peers for them. The power conservation scheme adjusts the MHs' degree of activity or participation in 7DS based on their available battery levels.

Shen et al. (2004) propose another cooperative data dissemination scheme with power conservation, called *energy-efficient cooperative caching with optimal radius* (ECOR), in a mobile environment. In ECOR, an optimal radius (in number of hops) is estimated by an analytical model that considers the MH's location, data access probability, and network density for each data item. The MHs exchange the cache content and the optimal radius of each cached data item among themselves. When an MH encounters a local cache miss, if it finds that any peers cache its desired data item, and the distance between the MH and the peer is within the optimal radius based on its local state, the MH sends a request message to the peer that is the closest to the selected holder of the data item. Otherwise, the MH obtains the data items from the MSS.

Yin and Cao (2004) propose three other cooperative caching schemes, called *CacheData, CachePath,* and *HybridCache.* The idea of CacheData is that an MH caches a passing-by data item, if the data item is popular and a condition that all requests for the data items are not originated by the same MH is satisfied. For CachePath, the MH caches path information of the passing-by data item instead of the data item. To conserve cache space, an MH does not cache path information of all passing-by data items. It only caches the path information of a data item if it is closer to the requesting MH than the MSS. HybridCache is a hybrid scheme that combines both CacheData and CachePath. An MH either applies CacheData or CachePath based on three factors: data item size, data item time-to-live, and the distance between the MH's distance to the data holder and the distance to the MSS.

COOPERATIVE CACHE MANAGEMENT

All literature related to cooperative cache management can be further divided into four sub-categories: *cooperative*