Peer-to-Peer Cooperative Caching in Mobile Environments

Chi-Yin Chow
Department of Computer Science and Engineering
University o Minnesota
4-192 EE/CSci Building
200 Union Street S.E.
Minneapolis, MN 55455, USA
email: chowx038@umn.edu

Hong Va Leong*
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
email: cshleong@comp.polyu.edu.hk

Alvin T. S. Chan
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
email: cstschan@comp.polyu.edu.hk

(* Corresponding author)

Peer-to-Peer Cooperative Caching in Mobile Environment

Chi-Yin Chow, University of Minnesota, Minneapolis, MN 55455

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. 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 chapter reviews a hybrid communication architecture, i.e., mobile cooperative caching, that combines 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*. An 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, i.e., *pull-based data dissemination model*, or catching them from scalable broadcast channels, i.e., *push-based data dissemination model*, or through the utilization of both types of channels, i.e., *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 that is also known as *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 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 architecture, i.e., mobile cooperative caching, in which a convectional infrastructure-based mobile communication architecture 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), *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 a 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 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, i.e., a data item is likely to be cached by which MHs, 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 which 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 data replica allocation*, *cooperative cache invalidation*, *cooperative cache admission control* and *cooperative cache replacement*.

COOPERATIVE DATA REPLICA ALLOCATION

Data replica allocation techniques (Hara, 2001, 2002a, 2002b; Hara et al., 2003) are adopted in mobile cooperative caching to improve data accessibility, in order to alleviate the network partitioning problem. Hara (2001) proposes three data replica allocation schemes: *SAF* (Static Access Frequency), *DAFN* (Dynamic Access Frequency and Neighborhood) and *DCG* (Dynamic Connectivity based Group). The MHs applying SAF only consider their own individual access probability to each data item. DAFN extends SAF to take the access probability to each data item of the MHs' connected neighborhoods into account. Finally, DCG groups the MHs with highly stable connection together. A group of MHs possesses a high connection stability, as they form a *biconnected component* in the network. DCG considers the access probability to each data item of all MHs in the same group. The performance evaluation result shows that DCG gives the highest data accessibility, but it incurs higher network traffic than the other two schemes. Thus, DCG can be considered as a scheme that trades network traffic for data accessibility.

These three data replica allocation schemes are then adopted to a push-based mobile environment (Hara, 2002a). Other than data access probability, the schemes also consider the latency on accessing data items from the peers and broadcast channel. Furthermore, the proposed replica allocation schemes are further extended, namely, Extended SAF (*E-SAF*), Extended DAFN (*E-DAFN*) and Extended DCG (*E-DCG*), to consider periodic data update by allowing the extended allocation schemes to consider the remaining time until next update of each data item (Hara, 2002b). In addition to the access probability, Hara et al. (2003) also consider the stability of radio links. The stability of a radio link is defined as the remaining time period that two MHs will still be connected to each other. The longer the time period indicates the higher the stability of a ratio link.

Huang, Chen, and Peng (2003) propose another distributed data replica allocation scheme in MANETs, called *DRAM*, to improve data accessibility and reduce network traffic pretending to the replication mechanism. DRAM extends E-DCG (Hara, 2002b) to consider group mobility pattern for data replica allocation. It is assumed that some MHs tend to roam together and they share a common access range. To discover the group mobility pattern among MHs, a distributed clustering algorithm is adopted to cluster several MHs who possess similar mobility pattern into a group. The clustering algorithm is executed periodically to adapt to the changes in network topology. Then, the data replicas are allocated to each group member based on group access probability to the data items, and the remaining time until the next update on them. DRAM is found to perform better than E-DCG in terms of data accessibility and network traffic.

COOPERATIVE CACHE INVALIDATION

Hayashi et al. (2003) propose two cache invalidation schemes, namely, *update broadcast* and *connection rebroadcast*. The former one is a straightforward, flooding-based scheme. An MH that caches an original copy of a data item broadcasts an invalidation report to other peers, when that MH updates the data item. The latter one can be referred to as a cooperative cache invalidation scheme. When two MHs are newly connected to each other, they broadcast their collected cache invalidation information to their connected peers. The newly connected MHs and other peers receiving their broadcast information update their own previously received cache invalidation information to identify any obsolete data items in their cache. The performance evaluation result shows that connection rebroadcast scheme reduces the number of accesses to invalid cached data items, but it incurs higher network traffic than update broadcast scheme.

COOPERATIVE CACHE ADMISSION CONTROL AND CACHE REPLACEMENT

Lim et al. (2003) propose a cooperative caching scheme for Internet-based MANETs, namely IMANET. In IMANET, a simple, flooding-based searching scheme is proposed for the MHs to search their desired data items in the network. IMANET also provides two cooperative data management protocols: *cooperative cache admission control* and *cache replacement*. For the cooperative cache admission control protocol, an MH determines whether to cache a data item based on the distance between itself and the data source that can be either other peers caching the data item or the MSS. For the cooperative cache replacement protocol, called *time and distance sensitive* (TDS), a victim data item is selected to be evicted from the cache by an MH based on two factors: the distance between itself and other peers caching the victim or the MSS, and the freshness of the distance information. The distance information is updated when the corresponding data item is accessed by other MHs. Since the network topology changes frequently, the distance information could become outdated, as it has not been updated for a long time.

There is a need for cooperating peers to cache useful data items together, so as to improve cache hit from peers. This could be realized by capturing the data requirement of individual peers in conjunction with their mobility patterns. Two group-based mobile cooperative caching schemes, namely, GroCoca (Chow et al., 2004) and DGCoca (Chow, Leong, & Chan, 2005), make use of a concept of a *tightly-coupled group* (TCG) that is defined as a group of MHs that are *geographically* and *operationally close*, i.e., sharing common mobility and data access patterns. Two MHs are considered to be geographically and operationally close based upon their locations and the set of data items they access respectively.

GroCoca is a *centralized* group-based mobile cooperative caching scheme, in which the MSS uses an incremental clustering algorithm to discover TCGs based on the weighted average distance and data access similarity of any two MHs. In GroCoca, when an MH encounters a local cache miss and its peer can turns in its desired data item to it, it only caches the data item, if its local cache has not been fully occupied or the peer is not belonging to its TCG. In other words, the MHs do not cache the data items that are provided by their TCG members, on the belief that the data items can be readily available from the peer if needed.

On the contrary, DGCoca is a *distributed* group-based mobile cooperative caching scheme, in which a stable neighbor discovery algorithm is proposed for the MHs to discovery their own TCG's members dynamically without any help of the MSS. The MHs adopting DGCoca not only make use of the cooperative cache admission control protocol proposed in GroCoca, but they also perform cooperative cache replacement to further improve data accessibility. The proposed cooperative cache replacement protocol possesses three important properties. First, the most valuable data items are always retained in the local cache. Second, in a local cache, a data item which has not been accessed for a long period will be replaced eventually. Third, in a TCG, a data item which "spawns" replica is first replaced in order to increase the effective cache size.

FUTURE TRENDS

Mobile cooperative caching is one of the most promising techniques to improve system performance in mobile environments. The future mobile cooperative caching scheme will focus on investigating security and privacy issues. Since an MH can cache outdated, counterfeit or harmful data items, the access to these data items potentially brings damages to other peers. Therefore, it is necessary to design a mechanism for an MH to ensure the freshness, reliability and safety of the data items retrieved from other peers. Also, when an MH accesses cached information from others, the MH may tell some personal information about them, e.g., personal data, preferences, location information, data access history, etc. To resolve this problem, there should be a permission control mechanism on accessing cached data item, so that the MHs can choose what kinds of information not to be disclosed. Also, an anonymous data sharing protocol should be developed to protect the privacy of the participating MHs in a mobile cooperative caching environment because their personal information may be revealed by other peers through their requests.

CONCLUSION

Mobile cooperative caching is a novel hybrid communication architecture that combines P2P data dissemination model into conventional infrastructure-based communication architecture. In mobile cooperative caching, the MHs can retrieve their desired data items not only from the MSS, but also from their peers. The MHs also manage their cache space with respect to themselves and their peers to improve system performance, such as cooperative data replica allocation, cooperative cache invalidation, cooperative cache admission control and cache replacement. The future trend of mobile cooperative caching will be focusing on how to resolve the security and privacy issues.

REFERENCES

Bluetooth SIG (2004). *Bluetooth specification v2.0*. [online] http://www.bluetooth.org.

Chow, C.-Y., Leong, H. V., & Chan, A. T. S. (2004). Group-based cooperative cache management for mobile clients in a mobile environment. *Proceedings of the 33rd International Conference on Parallel Processing (ICPP)*, pp. 83–90. Montreal, Canada.

Chow, C.-Y., Leong, H. V., & Chan, A. T. S. (2005). Distributed group-based cooperative caching in a mobile broadcast environment. *Proceedings of the Sixth International Conference on Mobile Data Management (MDM)*, pp. 97–106. Ayia Napa, Cyprus.

Fife, L. D., & Gruenwald, L. (2003). Research issues for data communication in mobile ad-hoc network database systems. *ACM SIGMOD Record, 32*(2), 42–47.

Hara, T. (2001). Effective replica allocation in ad hoc networks for improving data accessibility. *Pro-*

ceedings of the 20th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), pp. 1568–1576. Anchorage, Alaska, USA.

Hara, T. (2002). Cooperative caching by mobile clients in push-based information systems. *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM)*, pp. 186–193. McLean, Virginia, USA.

Hara, T. (2002). Replica allocation in ad hoc networks with periodic data update. *Proceedings of the 3rd Interntional Conference on Mobile Data Management (MDM)*, pp. 79–86. Singapore.

Hara, T., Loh, Y.-H., & Nishio, S. (2003). Data replication methods based on the stability of radio links in ad hoc networks. *Proceedings of the 6th International Workshop on Mobility in Databases and Distributed Systems (MDDS), in conjunction with the 14th International Conference on Database and Expert Systems Applications (DEXA)*, pp. 969–973. Prague, Czech Republic.

Hayashi, H., Hara, T., & Nishio, S. (2003). Cache invalidation for updated data in ad hoc networks. *Proceedings of the 11st International Conference on Cooperative Information Systems (CoopIS)*, pp. 516–535. Catania, Sicily, Italy.

Huang, J.-L., Chen, M.-S., & Peng, W.-C. (2003). Exploring group mobility for replica data allocation in a mobile environment. *Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM)*, pp. 161–168. New Orleans, Louisiana, USA.

IEEE Standard 802-11 (1997). *IEEE Standard for Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*.

Lau, W. H. O., Kumar, M., & Venkatesh, S. (2002). A cooperative cache architecture in support of caching multimedia objects in MANETs. *Proceedings of the 5th ACM International Workshop on Wireless Mobile Multimedia (WoWMoM), in conjunction with the 8th International Conference on Mobile Computing and Networking (MobiCom)*, pp. 56–63. Atlanta, Georgia, USA.

Lim, S., Lee, W.-C., Cao, G., & Das, C. R. (2003). A novel caching scheme for internet based mobile ad hoc networks. *Proceedings of the 12th IEEE International Conference on Computer Communications and Networks (ICCCN)*, pp. 38–43. Dallas, Texas, USA.

Papadopouli, M., & Schulzrinne, H. (2001). Effects of power conservation, wireless coverage and cooperation on data dissemination among mobile devices. *Proceedings of the 2nd ACM Interational*

Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc), pp. 117–127. Long Beach, California, USA.

Sailhan, F., & Issarny, V. (2003). Cooperative caching in ad hoc networks. *Proceedings of the 4th International Conference on Mobile Data Management (MDM)*, pp. 13–28. Melbourne, Australia.

Shen, H., Das, S. K., Kumar, M., & Wang, Z. (2004). Cooperative caching with optimal radius in hybrid wireless network. *Proceedings of the 3rd International IFIP-TC6 Networking Conference*, pp. 841–853. Athens, Greece.

Wang, K. H., & Li, B. (2002). Efficient and guaranteed service coverage in partitionable mobile ad-hoc networks. *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pp. 1089–1098. New York, USA.

Yin, L., & Cao, G. (2004). Supporting cooperative caching in ad hoc networks. *Proceedings of the 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pp. 2537–2547. Hong Kong.

TERMS AND THEIR DEFINITION

Mobile ad-hoc network (MANET): MANET is a peer-to-peer mobile communication architecture, in which the mobile clients can share information among themselves without any help of mobile support stations.

Network partition: There is no mobile client acting as a gateway between two groups of mobile clients in a mobile environment.

Data replica allocation: It is an algorithm for allocating some selected data items to a single mobile client or a group of mobile clients based on some prescribed criteria.

Cache invalidation: It is a mechanism for a mobile client to check whether its cached data items have been updated by someone else.

Cache admission control: It is a set of policies for a mobile client to decide on whether to cache a data item.

Cache replacement: When a cache has no room for storing a data item, the least valuable cached data item is literately removed from the cache until there is enough space for caching the required data item.

Time-to-live: It is the remaining time period of a data item to be updated or evicted from a cache.

Biconnected component: In graph theory, a biconnected component is a maximal subset of edges of a connected graph such that the corresponding induced subgraph cannot be disconnected by deleting any one vertex.

Tightly-coupled group: A group of mobile clients share common mobility pattern and data affinity.