Overload-Driven Mobility-Aware Cache Management in Wireless Environments

Humeyra Topcu-Altintas, Yun Huang and Nalini Venkatasubramanian
School of Information & Computer Sciences, University of California, Irvine
{htopcu, yunh, nalini}@ics.uci.edu

Abstract

In this paper, we propose a novel cache management strategy that uses the notion of "overload" where overload is defined as a situation when there is insufficient proxy cache for a new incoming user in a mobile region. In an overloaded situation, there will be increased network traffic (since the original server will need to be contacted) and increased service delays for mobile hosts (due to cache misses). The proposed techniques attempt to decrease the penalty of overloaded traffic and to reduce the number of remote accesses by increasing cache hit ratio. Our cache replacement algorithm is holistic in that it considers (i) mobility of the clients, (ii) predicted overloads, (iii) sizes of cached objects and (iv) their access frequencies in determining which object's cache (how much cache space) to be replaced, and when to replace. Performance results show that our overload-driven cache management strategy outperforms the existing popular policies.

1. Introduction

In infrastructure based wireless environments, a base station or access point provides connectivity between mobile hosts and remote servers. Placing a proxy cache at the base station is an effective way of reducing overall bandwidth consumption in the network; fair and efficient access to content and services for mobile hosts can be ensured by providing dedicated cache capacity to each mobile host in the service area of the base station. However, developing techniques for allocation and maintenance of the dedicated cache for each mobile host in the cell is not a trivial task given constant mobility of nodes, occurrence of non-uniform heavy traffic and unpredictable service access patterns. For instance, that a large number of Mobile Hosts (MHs) are moving along on freeways and transportation networks while some MHs are heading to specific destinations (a particular building) and other MHs are moving freely among the cells at different speeds cause non-uniform heavy traffic. The goal of this paper is to develop enabling techniques for providing continuous services to multiple mobile hosts at low cost.

However, user mobility introduces new challenges to cache management in addition to generic issues of maintaining fairness, ensuring high cache hit ratios and supporting a high level of user satisfaction. First, user mobility patterns can be unpredictable, making it difficult to estimate the service time for which the allocated cache is needed. Secondly, when a sudden increase of the network traffic causes an overload (i.e. the proxy cache is insufficient to accommodate the requests of all MHs in its region) achieving fairness is a challenging problem. Thirdly, mobile hosts may require caches for various types (and sizes) of objects, which in turn affects user satisfaction at different levels. One may consider mechanisms for pre-allocating cache to ensure continuous services for potentially incoming MHs; while this will enhance user satisfaction if the anticipated hosts arrive in the expected region, it may cause wasted resources especially when the current cache space is not sufficient for all existing users. Most of the previous studies on cache replacement [1,2,7], cache invalidation [5,6] and cache relocation [8, 9] have focused on cache replacement issues by considering cached object sizes, user access patterns, update frequencies, etc. To the best of our knowledge, none of these techniques apply non-uniform traffic information for making cache management decisions.

2. CacheRight adaptations

In this paper, we propose a novel cache management strategy, CacheRight that takes into account of mobility of users and characteristics of the

1-4244-0499-1/06/$20.00 ©2006 IEEE

cached objects. A key emphasis is to develop cache replacement and relocation techniques under different workloads, e.g. the case of heavy non-uniform traffic. In particular, we (a) define the notion of an "overload" to be the situation when there is insufficient proxy cache for a new incoming user and (b) develop intelligent cache management techniques based on this notion of "overload". The CacheRight includes two main techniques, a proactive cache adaptation technique that handles non-uniform traffic patterns, and a dynamic cache adaptation technique that addresses real time overloads.

Proactive CacheRight is executed periodically in order to avoid future overloads. More specifically, the proactive technique attempts to predict and estimate future cache overload conditions in a region and determines how cached objects can be effectively reconfigured to best accommodate the expected overloads. For instance, the number of overloads that may occur on each proxy for the next time period is predicted by applying overload records kept in previous periods using an Auto Regressive (AR) model [3]. Once the number of overloads that would happen in the next period of time has been predicted, the proactive CacheRight will then try to determine the size of cache space that should be freely available at the beginning of the next period of time in order to avoid the predicted overloads. If it requests an addition of extra cache space to the currently available free cache, it will release a certain amount of cache space by removing a number objects from the current proxy cache. The release procedure is carried out by a mobility-aware object ranking algorithm. In order to improve cache hit ratio, proactive CacheRight determines which objects can be removed from the cache by using a mobility-aware object ranking algorithm. The goal of the ranking algorithm is to reduce negative effect on cache hit ratio caused by removing cached data objects for future mobile hosts. For example, data objects that belong to a larger number of owners will receive higher priority; the mobile hosts who will leave this cell in the next period would possibly access their cached data objects less frequently than those mobile hosts that will keep staying in this cell.

Dynamic CacheRight determines proper cache space for a new cache request, when a new mobile host moves into a cell. Thus, dynamic CacheRight can handle unpredicted overload situations that may not be addressed by the periodic proactive CacheRight, e.g. significant workload variations caused by dynamic changes to user mobility and request access patterns.

Both techniques manage proxy cache for mobile hosts by taking into considerations of a number of influencing factors including: mobility of the clients, the predicted number of overloads, the sizes of currently cached objects and the access frequencies of cached objects. The combined effect of the two strategies working in tandem is to decrease the penalty of overloaded traffic on cache proxies, to reduce the number of remote accesses by increasing the cache hit ratio, and to improve user satisfaction.

3. Performance evaluations

We have evaluated our proposed CacheRight adaptations by a range of experiments under various configurations of the simulated system. With the basic configurations, we simulated a cellular network with 25 cells with a cache proxy located on each base station (BS), a main single server and mobile users that request different kinds of objects from the server. Each BS has the maximum cache capacity initially set as 40 MB. Requested cache space of mobile hosts varies from 1 MB to 3 MB. The main single server has a collection of 2000 data items. Data item sizes vary from 1KB to 100KB. Popularity of data objects follows three types of distributions, INCRT, DECRT and RAND [7]. Mobile requests to these objects are modeled by Zipf's distribution with skewness parameter $\theta=0.8$ [4]. We apply four mobility models to characterize their movements: *Random, Straight ahead, Regular* and *Combination* that combines the pervious three models. In the *Random* model, each mobile host can move freely to any direction. In the *Straight ahead* model, each mobile host has a specific moving direction to follow. In the *Regular* model, mobile hosts have a number of moving patterns.

Our simulation results show that CacheRight techniques decrease the number of overloads and increase the cache hit ratio under different traffic conditions, mobility patterns and proxy cache capacities. Given space limitation, we only illustrate the following results.

Figure 1 shows that the average number of overloads can be effectively reduced by proactive CacheRight adaptations under different mobility patterns. Figure 2 shows that compared with other cache replacement techniques, LRU and LRU-MIN [1], mobility-aware CacheRight adaptations can achieve higher cache hit ratio and is more resilient to increasing traffic loads, i.e. number of mobile hosts, under all types of object distribution patterns (namely INCRT, DECRT and RAND). In order to make fair comparisons, we keep other procedures (e.g. predicting overloads and determining cache size to release) of cache adaptations the same except the cache

replacement step (i.e. CacheRight takes mobility-aware cache replacement techniques). Note that Figure 2 shows results that are collected under *combination* mobility pattern, and results of other mobility patterns can lead to the same conclusion.

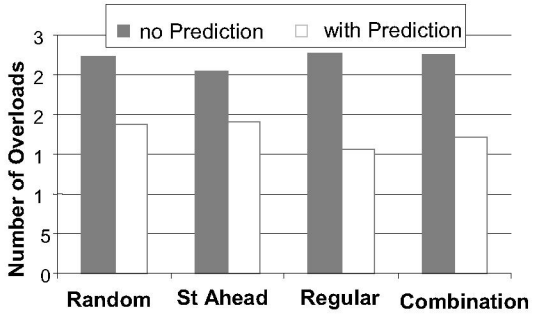
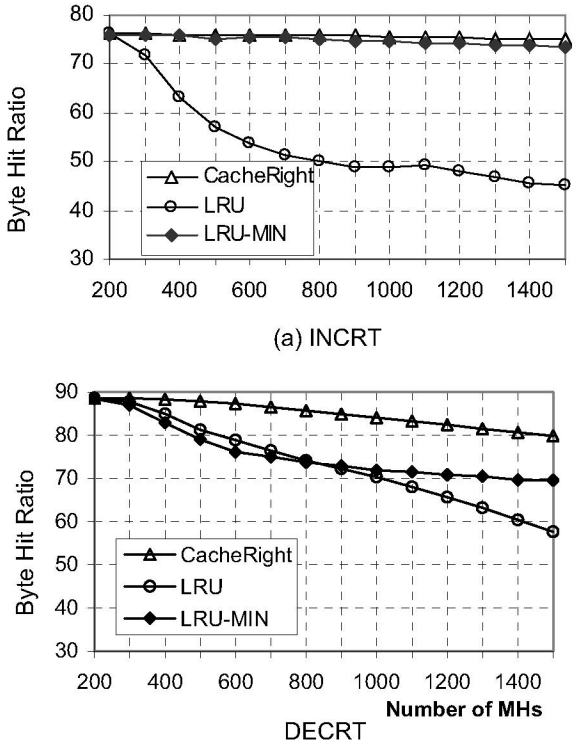


Figure 1: Impact of mobility patterns



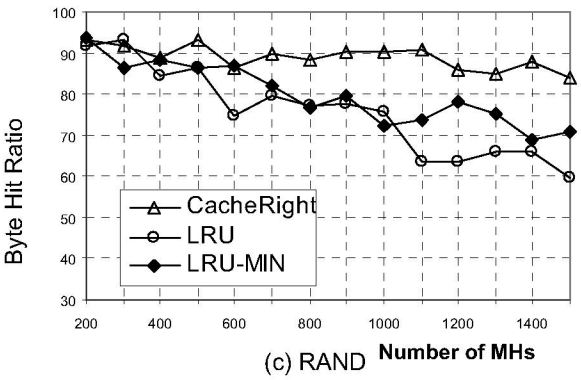


Figure 2: CacheRight performance under different data object distributions

4. References

[1] M. Abrams, et al., "Caching Proxies: Limitations and Potentials", Proc. 4th Int. WWW Conf., 1995.

[2] C. Aggarwal, J.L. Wolf, P.S. Yu, "Caching on the World Wide Web", IEEE Transactions on Knowledge and Data Engineering, Vol.11, No.1, pp. 94-107, 1999.

[3] C. Chatfield, *The Analysis of Time Series, An Introduction*, Chapman&Hall, pp.53-56, 1996.

[4] V. Almeida, et al., "Characterizing Reference Locality in the WWW", PDIS, 1996.

[5] G. Cao, "A Scalable Low Latency Cache Invalidation Strategy for Mobile Environments", IEEE Transactions on Knowledge and Data Engineering, Vol.15, No.5, 2003.

[6] S. Narayan, J. Pandya, P. Mohapatra, D.Ghosal, "Analysis of Windowing and Peering Schemes for Cache Coherency in Mobile Devices", NETWORKING 2005.

[7] J. Xu, et al., "Performance Evaluation of an Optimal Cache Replacement Policy for Wireless Data Dissemination under Cache Consistency", IEEE Transactions on Knowledge and Data Engineering, Vol.16, No.1, 2004.

[8] S. Hadjiefthymiades, L. Merakos, "Using Proxy Cache Relocation to Accelerate Web Browsing in Wireless/Mobile Communications", WWW10, 2001.

[9] J. Huang, M. Chen, "Intelligent Cache Capacity Allocation and Relocation Schemes in a Mobile Proxy", Proc. of the 3rd Int. MMCN, 2003.