Efficient Proactive Caching for Supporting Seamless Mobility

Vasilios A. Siris, Xenofon Vasilakos, and George C. Polyzos
Mobile Multimedia Laboratory, Department of Informatics
Athens University of Economics and Business, Greece
{vsiris, xvas, polyzos}@aueb.gr

***Abstract*—We present a distributed proactive caching approach that exploits user mobility information to decide where to proactively cache data to support seamless mobility, while efficiently utilizing cache storage using a congestion pricing scheme. The proposed approach is applicable to the case where objects have different sizes and to a two-level cache hierarchy, for both of which the proactive caching problem is hard. Additionally, our modeling framework considers the case where the delay is independent of the requested data object size and the case where the delay is a function of the object size. Our evaluation results show how various system parameters influence the delay gains of the proposed approach, which achieves robust and good performance relative to an oracle and an optimal scheme for a flat cache structure.**

I. Introduction

Proactively fetching data to reduce delay in mobility scenarios has been proposed within the context of publish-subscribe network architectures [1], [2], [3], for vehicular WiFi access [4], and more recently for cellular networks [5], [6]. Even earlier, prefetching has been proposed and investigated for file-system and Web access performance enhancement at least since the mid nineties. In this paper we focus on proactive caching related to mobility.

Proactive caching achieves gains by making data objects requested by a mobile user immediately available when the mobile moves to a new network attachment point, thus reducing the delay for obtaining the data compared to transferring it from the original source where the data is located; the higher delay for obtaining the data from the source can be due to the larger distance (number of hops), or because the path to the source includes low capacity links, such as a low capacity backhaul of femto/small cells or WiFi hotspots. Proactive caching can exploit mobility information, such as the probability of a mobile connecting to future attachment points. Exploiting such mobility information to undertake proactive actions has been applied to reduce handover delays in WiFi and cellular networks [7], [8], [9].

We propose a novel distributed approach for proactively fetching data that a mobile requests in caches located close the network attachment points where the mobile has some probability to connect to. The caches that should proactively fetch data objects are selected based on mobility information and the gains from reduced data transfer delay. The approach is applicable to the case where the requested data objects have different sizes and to a two-level cache hierarchy, which are both hard problems. Moreover, our modeling framework considers the case where a data object's transfer delay is independent of the object's size and the case where the transfer delay is a function of the object's size. Although our focus is on using caching to proactively fetch data based on mobility prediction, our framework can account for content popularity, which is a factor considered in traditional caching and data placement. Moreover, although our objective is to reduce the data transfer delay, the objective can include other forms of cost such as network or monetary.

A novel aspect of the proposed approach is the use of congestion pricing that considers the demand for caching and the available storage to efficiently utilize limited cache capacity, while reducing the delay for transferring the requested data to a mobile. A congestion pricing approach also allows us to solve the proactive caching problem in a distributed manner, by deciding independently for each data object where it should be cached. Congestion pricing has been extensively studied for flow and congestion control. To the best of our knowledge, the present work is the first to apply it to proactive caching.

The problem of selecting caches to proactively fetch data objects requested by mobiles is fundamentally different from both conventional caching and data (or content) placement. Both these problems involve caching data objects closer to potential requesters, based on the popularity of the objects. In contrast, the problem investigated in this paper involves proactively fetching data objects requested by mobiles based on their mobility; different mobiles, independent of the objects they request, can have different mobility patterns, which can result in different proactive caching decisions. A second difference is that, when a mobile eventually moves to a new attachment point, the cache space that was occupied by the data requested by the mobile is freed. Also, proactive caching involves selected caches pulling data from sources, whereas data placement involves pushing data to caches. Finally, our proactive caching approach does not perform eviction or replacement when a cache is full, as in conventional caching; rather, our approach uses congestion pricing to ensure that the objects proactively cached are those for which the highest delay gains are achieved. Interestingly, prior work has found that hierarchical or cooperative caching isn't helpful when the user population is above some relatively small threshold [10], which is due to the heavy-tailed object distributions. However, such results are not applicable to the two-level proactive cache hierarchy discussed in this paper, where the decision to cache data in leaf and mid-level caches depends on the mobility patterns and the gains from reduced transfer delay.

Proactive caching of data objects requested by a mobile in caches (or proxies) close to the mobile's future attachment points requires knowledge or prediction of mobile user requests. Knowledge of user requests is available *natively* in Information-Centric Networking (ICN) architectures, which employ a receiver-driven model where receivers request content by its name from one or more publishers, thus supporting

receiver mobility [11]. Alternatively, one can predict the data that a mobile is likely to request [12], [13]. Note that the focus of this paper is not on how mobility prediction can be performed, nor how the data objects requested are known or predicted. Rather, the paper focuses on developing and evaluating efficient schemes that exploit knowledge or prediction of mobility and data requests to proactively cache data to reduce the transfer delay in mobility scenarios.

In summary, our contributions are the following:

- We propose a distributed proactive caching approach that selects caches to proactively fetch data objects based on mobility information and uses congestion pricing to efficiently utilize cache storage.
- The proposed approach can be applied to the case where data objects have different sizes and to a two-level hierarchy of caches.
- Our modeling framework considers both the case where the delay does not depend on the size of the requested data objects and the case where the delay is a function of the size of the requested data objects.
- We present a comprehensive set of evaluation results that show how various system parameters influence the delay gains of the proposed approach, which achieves robust and good performance relative to an oracle and an optimal scheme for a flat cache structure.

The rest of the paper is structured as follows: In Section II we discuss related work, identifying where it differs from the work in this paper. In Section III we present the models and procedures for efficient proactive caching. Specifically, in Section III-A we consider a flat cache structure while in Section III-B we consider a two-level cache hierarchy; for both, we assume that delays are independent of the requested object sizes. The case where delays are a function of the requested object sizes is discussed in Section III-C. In Section IV we evaluate the proposed proactive caching scheme for scenarios where the delays experienced by all mobiles are the same and for scenarios that involve a scaled-down Internet topology, where different mobiles can experience different delays for transferring an object from its source. Finally, in Section V we conclude the paper identifying directions for future work.

II. Related work

In the context of publish-subscribe ICN networks, work related to proactive caching includes [1], [2]. The work in [1] considers proactive caching based on prediction, but does not investigate specific algorithms. The work in [2] proposes proactive caching in all caches that lie one-hop ahead, close to all network attachment points the mobile can attach to; such a solution will of course lead to inefficient utilization of caches close to attachment points that a mobile has a small probability to attach to. On the other hand, our approach involves selecting a subset of caches close to future network attachment points, based on the probabilities that the mobile moves to these points. Moveover, our investigations compare the proposed proactive caching approach with a naive approach that caches data objects in all one-hop caches, subject to the availability of cache storage. The work in this paper differs from our previous work on proactive caching in ICN networks [14], in that (1) the current paper proposes an approach based on congestion pricing to efficiently utilize cache storage, (2) we propose a proactive caching approach for a two-level cache hierarchy, and (3a) we consider the case where the delay is independent of the requested data object size and (3b) the case where the delay is a function of the object size.

The feasibility of using prediction together with prefetching for vehicular WiFi access is investigated in [4], which develops a prefetching protocol, but does not propose or evaluate specific prefetching algorithms. The work of [5] investigates proactive caching of video content in caches located in femtocells. Although the objective of [5] is similar to our objective, there are key differences in both the network model and the proposed solution: First, [5] considers a network of partially overlapping femtocells, which differs from the model considered in this paper. Second, we consider proactive caching of data requested by each mobile, based on the probability the mobile will move to different future attachment points; once the mobile moves, the data can be removed from all caches it was proactively fetched. On the other hand, the problem of [5] is essentially a data placement problem, which considers the popularity of the requested data [15]. Finally, we consider a decentralized approach using congestion pricing to efficiently utilize cache storage. On the other hand, [5] proposes a centralized greedy algorithm. The work of [6] considers proactive caching (or seeding) to minimize the peak of the total load in a cellular network, subject to user impatience constraints, which essentially impose a maximum delay for the content to be available to users requesting it. An offline water-filling algorithm that determines the content transfer schedules, initially assuming perfect knowledge of future user interests, is proposed; the algorithm is later adapted to handle uncertainty of user interests and traffic. Our work differs in that our constraint is the cache storage, rather than the network capacity as in [6]. Furthermore, [5], [6] do not consider the case where the transfer delay is a function of the object size.

III. Efficient proactive caching

Proactive caching is used to prefetch data requested by a mobile, so that it is immediately available when the mobile connects to its new network attachment point, thus reducing the transfer delay. We consider two cases for the transfer delay. In the first case, the transfer delay is independent of the size of the requested data objects, while in the second the transfer delay is a function of the data object size.

The transfer delay is independent of the requested data object size when the object size is small, e.g. fire/security alerts, thus the transfer delay is determined primarily by the propagation delay rather than the transmission delay. The delay for obtaining an object from its the original remote source is denoted D_{R}, whereas the delay for obtaining an object from the local cache is denoted D_{L}. These delays can depend on the distance to the source or cache, e.g. in number of hops. Note that, although we use the term delay, D_{R} and D_{L} can in general include the cost (e.g. network, monetary) for obtaining data from the source or a remote location, which is independent of the requested data object size. For the above assumption of the transfer delay, Section III-A describes the proactive caching approach for a flat set of caches, while Section III-B considers proactive caching in a two-level cache hierarchy.

The transfer delay will depend on the data object size when it is dominated by the transmission delay, e.g. for large objects such as video files. For example, when the mobile is connected to a WiFi hotspot (or 3G/4G small cell), the delay for transferring a data object from a local cache is proportional to the object size and inversely proportional to the WiFi (or 3G/4G) data rate. On the other hand, if the object is transferred from a remote location over a lower capacity backhaul link (e.g. xDSL) that connects the hotspot (or small cell) to the Internet, then the delay is inversely proportional to the backhaul rate, which is smaller than the WiFi (or 3G/4G) rate. As above, the delay can actually involve any cost that is proportional to the object size, e.g. cost per unit of transferred data in the case of data volume charging. Section III-C describes our proactive caching approach when the delay is a function of the requested object size.

A. *Proactive caching in a flat cache structure*

In this section we present our approach for selecting the caches that should proactively fetch data objects, in the case of a flat set of caches and when the transfer delay is independent of the size of the requested data objects. Under this assumption, as we will see, the data object is either fully cached or not cached. Our objective is to minimize the average delay across all requested objects, subject to the cache storage constraints. Note that in a flat cache structure, the caches are independent, hence prefetching decisions are also independent.

Let q_s^l denote the probability that the mobile requesting object s moves to cache l and B_l denote the maximum storage at cache l; we initially assume that all objects have the same size o. Note that the probability q_s^l can depend on the specific mobile, its current or past location, and the time instant, but for simplicity we do not make this dependence explicit in the notation. Also, let S_l be the set of objects requested by mobiles that have non-zero probability to move to cache l and L be the set of caches. We define the following optimization problem:

$$\min_{b_s^l} \quad \sum_{s \in S_l} \mathcal{D}_s \tag{1}$$

$$\text{subject to} \quad \sum_{s \in S_l} o \cdot b_s^l \leq B_l\,, \quad \forall l \in L, \tag{2}$$

where $\mathcal{D}_s = \sum_{l \in L} \mathcal{D}_s^l$ is the average delay for obtaining object s and b_s^l equals one if the object s is proactively fetched in cache l and zero if it is not proactively fetched in cache l. $\mathcal{D}_s^l$ is equal to $q_s^l D_{\text{R}}$ if the object is not in cache l ($b_s^l = 0$) and needs to be obtained from its original remote location and $q_s^l D_{\text{L}}$ if the object is stored in cache l ($b_s^l = 1$). The above optimization problem involves finding, Figure 1, for each data object s requested by a mobile, the subset $L' \in L$ of caches that will proactively fetch object s so that it is immediately available to the mobile when it connects to an attachment point close to the cache, in order to achieve the optimization target (1) while satisfying the cache storage constraint (2).

In order to efficiently utilize the cache storage, we introduce a congestion price p_l which is adapted based on the demand for caching and the available storage. Specifically,

$$p_l(t+1) = \left[p_l(t) + \gamma\left(o \cdot b^l(t) - B_l\right)\right]^+\,, \tag{3}$$

where $b^l(t)$ is the aggregate demand at cache l at time t and γ is the price update factor, which determines how quickly the cache congestion price adapts to changes of the demand for caching. For the evaluation in Section IV, the cache price is updated when a new cache request appears.

The decision to proactively fetch an object s at cache l is based on the following rule:

$$b_s^l = \begin{cases} 0 & \text{if} \quad q_s^l(D_{\text{R}} - D_{\text{L}}) < p_l \\ 1 & \text{if} \quad q_s^l(D_{\text{R}} - D_{\text{L}}) \geq p_l \end{cases} \tag{4}$$

The above decision rule provides a distributed and decentralized approach to decide, for each object s and for each cache l, whether the object should be prefetched. Of course, the object will actually be prefetched only if cache storage is available. Adjusting the cache price using (3) directs the system towards efficient use of cache storage, while achieving the optimization target (1). Specifically, when the cache is underutilized, i.e. cache storage is available, the congestion price decreases, thus allowing more objects to be proactively fetched in the cache based on the decision rule (4). On the contrary, when the cache demand is larger than the cache size, then the price increases which in turn, due to (4), reduces the number of objects that are requested to be cached. Furthermore, when the cache price is such that the amount of requested cache is equal to the cache storage, then due to the decision rule (4) we are certain that these requests correspond to the highest values of $q_s^l(D_{\text{R}} - D_{\text{L}})$, hence the minimum in (1) is achieved. Note that content popularity can be incorporated in the above model if we replace q_s^l with the sum $q_s^l + r_s$, where r_s is the popularity of object s.

Two practical issues related to the application of the decision procedure (4) include where the decision is taken and if the decision is to proactively cache an object, when should prefetching start. Regarding where the decision is taken, one option is for the mobile requesting object s, or some proxy on behalf of the mobile, to inform all caches located close to its possible future attachment points about its transition probability: When cache l learns the probability q_s^l, then together with the delays $D_{\text{L}}, D_{\text{R}}$ and the cache congestion price p_l it can apply the decision rule (4). Alternatively, the caching decision can be taken at the mobile, or its proxy, in which case it would need to learn the delays $D_{\text{L}}, D_{\text{R}}$ and the cache price p_l from all caches it has some probability to connect to. The second issue of when to start to prefetch a data object is related to the time interval after which the mobile connects to

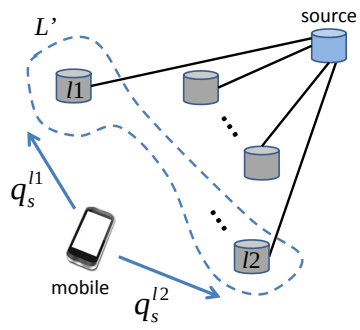


Fig. 1. The proactive caching problem involves selecting, based on the mobile transition probabilities, the set L' of caches to proactively cache object s, in order to achieve (1) while satisfying (2). In the proposed scheme the decision to proactively prefetch a data object is taken independently for each cache.

its next network attachment point and the time for the cache to download the requested object from its remote location.

When a mobile moves to its new attachment point, then it can directly receive the requested object from the local cache, if the cache had prefetched the requested object. Otherwise, the mobile obtains the requested object from the original source. When a mobile connects to its new attachment point, then the space occupied in all caches that proactively cached that mobile's data can be freed; of course, if the local cache at the mobile's new attachment point had prefetched the mobile's data, then the corresponding space will be freed after the data is transferred to the mobile. The above actions require communication and cooperation among caches.

The model presented above can be extended to objects with different sizes, by replacing the constraint in (2) with

$$\sum_{s \in S_l} o_s \cdot b_s^l \leq B_l \,,$$

where o_s is the size of object s. Additionally, the cache price p_l on the right side of the inequalities in (4) should be replaced with $o_s \cdot p_l$. For objects with different sizes, the optimization problem becomes identical to the *0/1 Knapsack Problem*, which falls within the class of NP-hard problems.

Another extension is when the remote and local delay is different for different data objects and caches, hence they can be denoted $D_{\text{R}}^{s,l}$, $D_{\text{L}}^{s,l}$ and the decision rule (4) can be adapted accordingly. Instead of maintaining different delays for different objects or mobiles, delays can be associated with object or mobile types, or can depend on the mobile's initial network attachment point. In all these cases, the actual values of the delay can be estimated in a measurement-based manner.

1) Optimal for equal-size objects: For a given set of cache requests, the optimal in the case of equal size data objects, can be obtained for a flat set of caches as follows: For each cache l, we order the cache requests in decreasing value of $q_s^l(D_{\text{R}} - D_{\text{L}})$. Then, starting from the request with the highest $q_s^l(D_{\text{R}} - D_{\text{L}})$, we fill the cache until the constraint B_l is reached.

The above procedure for obtaining the optimal is performed in rounds: in the beginning of each round we have a given set of cache requests. This is unlike the solution based on cache congestion pricing, where the decision of whether to cache an object is taken iteratively, based on (4), for each cache request, hence can be applied on-line. A practical issue with the optimal solution is the duration of each round, which determines the number of cache requests considered; this duration depends on the time interval after which a mobile connects to its new attachment point.

B. Proactive caching in a two-level hierarchy

Next we consider a two-level cache hierarchy. Each leaf node can be under only one mid-level cache, Figure 2, and an object can be proactively fetched at a leaf cache, at a mid-level cache, or both. The delay for obtaining an object from a mid-level cache is D_{M}, which satisfies $D_{\text{L}} < D_{\text{M}} < D_{\text{R}}$. At any time instance, there will be a given set of cache requests from the mobiles that are active at that time instant. For each such time instance, the proactive cache problem for a two-level cache hierarchy has similarities with the data placement problem [16], where the probability of an object being requested at a specific cache is given by the probability of the mobile moving to the corresponding network attachment point. The authors of [16] show that the data placement problem with different object sizes is NP-complete. Although there are cases where the placement problem in a hierarchical network with equal size objects can be solved in polynomial time [17], such solutions have a high polynomial degree and apply to an offline version of the problem.

In a two-level cache hierarchy, the leaf cache can correspond to caching that is performed at a local area network, such as femto/small cell, hotspot or home network, whereas the mid-level cache can correspond to caching at the ISP that connects the local network to the Internet.

Our approach to solve the proactive caching problem in a two-level cache hierarchy first considers two flat cache selection problems: one assuming that the object is proactively fetched in the mid-level cache and the other assuming that the object is not proactively fetched in the mid-level cache. Each of the aforementioned flat cache problems can be solved using the distributed approach presented in Section III-A, by having the mid-level cache send the leaf caches the delay D_{R}, which is the delay for obtaining the object from the remote source, and the delay D_{M}, which is the delay for obtaining the object from the mid-level cache. Each leaf cache, using (4), decides whether to cache the specific data object for each of the two problems: For the problem where the object is assumed not to be cached in the mid-level cache, the leaf cache uses (4) to decide if the object should be prefetched in the leaf cache. Whereas, for the problem where the object is assumed to be cached in the mid-level buffer, the leaf cache uses (4) replacing D_{R} with D_{M} to decide if the object should be prefetched in the leaf cache. Each leaf cache informs the mid-level cache about the resulting average object delay for each of the two problems: $q_s^l D_{\text{R}}$ or $q_s^l D_{\text{M}}$ if the decision is not to proactively fetch the object and $q_s^l D_{\text{L}}$ if the decision is to proactively fetch the object in the leaf cache.

After receiving from all leaf caches the delay for the two problems, the mid-level cache takes the sum of the delays for each problem: $\mathcal{D}_{\text{M}}^{\text{mid}}$ in the case the mid-level cache proactively fetches the data object and $\mathcal{D}_{\text{R}}^{\text{mid}}$ in the case the mid-level cache does not proactively fetch the data object. The decision of whether to cache or not cache a data object in the mid-level cache is determined based on a decision rule that resembles (4): If $\mathcal{D}_{\text{R}}^{\text{mid}} - \mathcal{D}_{\text{M}}^{\text{mid}} \geq p_{\text{mid}}$ then the object is proactively fetched in the mid-level cache, otherwise it is not fetched. The variable

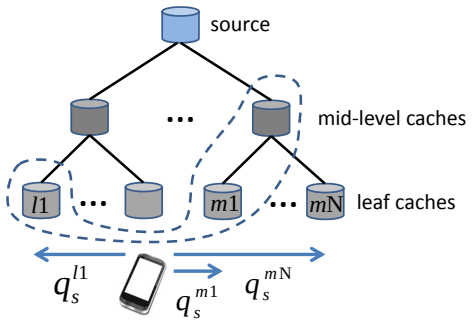


Fig. 2. In a two-level cache hierarchy, each mid-level cache cooperates with its leaf caches to decide which will proactively cache a data object. No such cooperation is necessary between mid-level caches.

p_{mid} is the congestion price for the mid-level cache, which is updated in a similar manner as the congestion price for the leaf caches (3), but based on the demand for caching in the mid-level cache and the storage available in the mid-level cache. Following its decision, the mid-level cache then informs the leaf caches which delay factor (D_{R} or D_{M}) they should use in (4) to eventually decide whether to cache a data object.

The above procedure is distributed and requires some cooperation between the mid-level cache and its leaf caches. Moreover, it can be applied to a hierarchy with more than one mid-level caches, as long as each leaf cache is a child of only one mid-level cache. The proactive caching decision for each mid-level cache and its leaf caches follows the above approach, and can be performed independently of other mid-level caches.

C. Proactive caching when delay is a function of object size

In Sections III-A and III-B we assumed that the transfer delays $D_{\text{L}}, D_{\text{M}}, D_{\text{R}}$ are independent of the data object size. In this section we consider the case where the transfer delay is a function of the data object size. As we will see, in this case there can be gains if a part of an object, and not necessarily the whole object, is proactively cached.

Let R_{L} be the rate for transferring to a mobile data stored locally at a cache close to the mobile's attachment point, and R_{R} be the rate for transferring to a mobile data from a remote source. We assume that $R_{\text{R}} < R_{\text{L}}$, which justifies why proactively caching can be beneficial. As an example, R_{L} can be the rate for transferring data across a WiFi or 3G/4G interface, while R_{R} can be the rate for transferring data across a hotspot or small cell's backhaul link, which is typically smaller than the rate of a WiFi or 3G/4G interface.

Assume that some part x_s^l of object s that has size o_s is proactively fetched to cache l, whereas the remaining part $o_s - x_s^l$ would need to be obtained from the object's original remote location. If the mobile requesting object s moves to an attachment point close to cache l, then the delay for transferring the whole object s is given by

$$\mathcal{D}_s(x_s^l) = \frac{x_s^l}{R_{\text{L}}} + \frac{o_s - x_s^l}{R_{\text{R}}} = \frac{o_s}{R_{\text{R}}} - \left(\frac{1}{R_{\text{R}}} - \frac{1}{R_{\text{L}}}\right) x_s^l \,. \quad (5)$$

In the last equation we have assumed that the rates $R_{\text{L}}, R_{\text{R}}$ are the same for all caches l. The results presented below are similar when the rates are different for different caches.

Consider a utility function $\mathcal{U}_s(d)$ that represents a mobile user's valuation for delay d in transferring object s. $\mathcal{U}_s(d)$ is a decreasing function, and possible shapes for it are shown in Figure 3. Figure 3(a) corresponds to the case where a user obtains no value when the delay is above some maximum threshold, and obtains a value that increases linearly as the delay approaches zero. The sigmoid utility in Figure 3(b) corresponds to the case where a user obtains maximum value when the delay is below some minimum threshold, while he obtains zero value when the delay is above some maximum threshold; such a curve approximates the exact step utility of hard real-time applications, with strict delay requirements.

We can define the utility $U_s^l(x_s^l) = \mathcal{U}_s(\mathcal{D}_s(x_s^l)/q_s^l)$, which is now a function of the part x_s^l of object s that is proactively fetched in cache l. Note that in the last equation we have added the factor $1/q_s^l$ to account for the transition probability

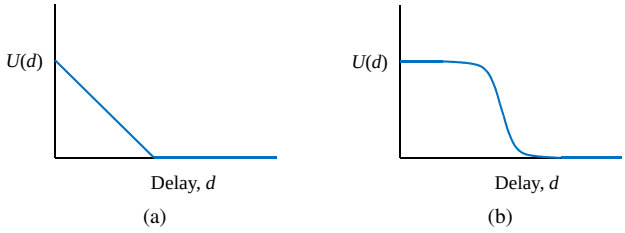


Fig. 3. Example utilities as a function of delay.

to cache l. Hence, if a mobile has a higher probability to move to cache l compared to some other cache, then it would need to proactively cache a larger amount of the data object to achieve the same utility. We assume that the utility function $U_s^l(x_s^l)$ is continuous and strictly increasing in the interval $[m_s^l, M_s^l]$, where $m_s^l \geq 0$ and $m_s^l < M_s^l \leq o_s$ are minimum and maximum values of x_s^l for which the following hold: $U_s^l(x_s^l) = U_s^l(m_s^l)$ for $x_s^l \leq m_s^l$ and $U_s^l(x_s^l) = U_s^l(M_s^l)$ for $x_s^l \geq M_s^l$. As an example, a utility function that corresponds to Figure 3(a) is $U_s^l(x_s^l) = \frac{\mathcal{R}}{q_s^l}(x_s^l - m_s^l)$, for $x_s^l \in [m_s^l, M_s^l]$, where $\mathcal{R} = \frac{1}{R_{\text{R}}} - \frac{1}{R_{\text{L}}}$.

As in the previous two subsections, we define a cache congestion price that is updated as in (3), with $o \cdot b^l(t)$ replaced by $x^l(t) = \sum_{s \in S_l} x_s^l(t)$, which gives the aggregate demand for cache storage at l, when data objects can be partially cached:

$$p_l(t+1) = \left[p_l(t) + \gamma\left(x^l(t) - B_l\right)\right]^+ , \quad (6)$$

where as before γ is a price update factor.

The framework below follows the model of [18]. We define the following decision rule for selecting the amount x_s^l of object s to be proactively fetched in cache l:

$$x_s^l = \begin{cases} m_s^l & \text{if } \frac{1}{U_s^l(m_s^l)} \leq p_l \\ {U_s^l}^{-1}\left(\frac{1}{p_l}\right) & \text{if } \frac{1}{U_s^l(M_s^l)} < p_l < \frac{1}{U_s^l(m_s^l)} \\ M_s^l & \text{if } p_l \leq \frac{1}{U_s^l(M_s^l)} \end{cases} \quad (7)$$

Unlike the binary decision rule (4), where a data object is either fully cached or is not cached, with (7) it may happen that only a part of object s is fetched at cache l: x_s^l is a continuous variable with values in $[m_s^l, M_s^l]$, where $m_s^l \leq 0$ and $m_s^l < M_s^l \leq o_s$.

Using the results from [18], it can be shown that a system where the amount x_s^l of object s to be proactively fetched in cache l is determined by (7) and the cache price is updated according to (6), with an appropriately small value γ, will converge and solve the following optimization problem:

$$\max_{m_s^l \leq x_s^l \leq M_s^l} \quad \sum_{l \in L} \sum_{s \in S} V_s(x_s^l)$$

$$\text{subject to} \quad \sum_{s \in S_l} x_s^l \leq B_l \,, \quad l \in L$$

where S is the set of all data objects, S_l is the set of objects cached at l, L is the set of all caches, and

$$V_s(x_s^l) = \int_{m_s^l}^{x_s^l} \frac{1}{U_s^l(y)} dy \,, \quad m_s^l \leq x_s^l \leq M_s^l \,.$$

As an example, for the utility $U_s^l(x_s^l) = \frac{\mathcal{R}}{q_s^l}(x_s^l - m_s^l)$, we have $V_s(x_s^l) = \frac{q_s^l}{\mathcal{R}} \log(x_s^l - m_s^l)$.

Moreover, at the optimum the requested amount of data to be proactively cached $\{x_s^{l\,*} : s \in S, l \in L\}$ achieves utility proportional fairness: for any other set of cached data sizes $\{x_s^l : m_s^l \leq x_s^l \leq M_s^l, s \in S, l \in L\}$,

$$\sum_{l \in L} \sum_{s \in S} \frac{x_s^l - x_s^{l\,*}}{U_s^l(x_s^{l\,*})} \leq 0\,.$$

Utility fairness can be seen as a way to allocate limited resources in a manner that is fair from an application perspective, since it takes into account the actual valuation (utility) for a specific amount of resources (cache storage in our case). In contrast, resource-oriented fairness definitions, such as proportional fairness and max-min fairness, seek to allocate resources in a fair manner from a resource-centric perspective, which does not necessarily reflect the requirements at the application level. Thus, considering utility fairness in the model of this section results in an application-oriented approach for performing proactive caching.

IV. Evaluation

In this section we evaluate the proposed Efficient Proactive Caching (EPC) scheme using the OMNeT++ simulation framework. We present results for a flat and a two-level cache hierarchy, for data objects that have the same size and when the delay is independent of the object size. Evaluation results for different object sizes and when the delay is a function of the object size will be included in a followup of this paper.

A. Simulation model

We consider scenarios where the delay for obtaining a data object is the same for all mobiles (referred to as fixed delay scenarios) and scenarios where the delay for obtaining a data object from the source is variable and depends on the number of hops between the source and the mobile (receiver); the latter scenario involves a scaled-down Internet topology containing 400 nodes, with each node representing an autonomous system (AS) [19]. The value of various system parameters considered in the simulations are shown in Table I. In the fixed delay scenarios, there are a total of 160 mobile users that issue requests for objects of the same size. The mobiles can move to 8 different network attachment points, where there is a corresponding local leaf cache. At a higher level, there is a mid-level cache. Whenever a mobile performs a handover, we assume that a new mobile enters the system, thus the total number of active mobiles in the system always remains 160. In the scenarios with scaled-down Internet topology, we conduct simulations over a set of 10 different neighborhoods, with each neighborhood having 160 mobile users, yielding a total of 1600 mobile users. Each neighborhood is composed by 8 ASes lying on the edge of the topology. The vast majority of such ASes are stub (access) networks to the Internet. For each neighborhood, we randomly select an initial stub AS and then form the neighborhood by selecting 8 stub ASes based on minimum hop distance from the initial stub AS. The underlying idea is that the selected nodes are neighbors of the initial stub AS, thus typically they should be a few hops away for mobile users hosted by the initial stub AS. Note that the neighbors are selected so there are no overlaps between different neighborhoods.

A mobile user can move with some transition probability to one of the 8 different attachment points, each with its own local cache. In the fixed delay scenarios there is one mid-level cache, whereas in the scaled-down Internet topology scenarios there is one mid-level cache in each neighborhood. We consider 4 different sets of mobile transition probabilities, Table I, with different *skewness*, where a higher *skewness* corresponds to a higher probability to move to a particular attachment point (equivalently, cache). We also assume that the destination network attachment point (equivalently, the destination cache) with the highest transition probability is different for different mobiles, such that the number of active mobiles moving to a specific cache is on average the same, and equal to 20 throughout the simulation. Finally, note that with the EPC and optimal schemes the transition probabilities used in the caching decisions are measured as the simulation progresses.

The performance of the EPC scheme depends only on the ratio of delays $D_{\text{R}}/D_{\text{L}}$ and $D_{\text{M}}/D_{\text{L}}$, because the decision rule for both the leaf cache (4) and the mid-level cache has a linear dependence on the delays and the congestion price is adjusted to achieve high utilization; hence we consider the delay ratios rather than the absolute delay values. Assuming that the leaf cache performs caching at a local area network (e.g. femto/small cell or hotspot) and that the mid-level cache performs caching at the ISP that connects the local network to the Internet, we have taken $D_{\text{M}}/D_{\text{L}} = 2$ and 5, which can be seen as the number of hops or the actual delay for obtaining a data object from an ISP cache relative to a local network cache. We have considered $D_{\text{R}}/D_{\text{L}} = 10$ and 18; together with the values of $D_{\text{M}}/D_{\text{L}}$, these give values of $D_{\text{R}}/D_{\text{M}}$ in the range $[2, 5]$; the lowest value $D_{\text{R}}/D_{\text{M}} = 2$ corresponds to the case where an ISP has a direct peering link with the content provider network (source for a data object), in which case their distance is two AS hops. At the other side, studies have shown that the average inter-AS path length has remained practically constant and equal to 4.2 over the last 12 years [20], and for this reason we have selected the highest value of $D_{\text{R}}/D_{\text{M}}$ to be 5. For the scaled-down Internet topology scenarios, we consider $D_{\text{M}}/D_{\text{L}} = 5$ and $D_{\text{R}}/D_{\text{L}} = \frac{9}{5}(\#\text{ hops} - 1) + 1$; the latter gives an average $D_{\text{R}}/D_{\text{L}}$ equal to 10, since the average number of hops between a source and receiver in the scaled-down Internet topology is 6.

The performance of the proposed EPC scheme is compared to the optimal scheme described in Section III-A1, to a naive scheme, and an oracle. It is important to note that the optimal scheme in the case of a flat cache structure is implemented such that the cache allocation is performed whenever cache storage is freed. In addition to being time consuming, the ability to implement frequent cache allocations can be constrained by the time for actually transferring the data objects to the caches where they are proactively fetched.

With the naive scheme, a mobile requests caching for all data objects, provided that storage is available. With the oracle, for each new cache request we assume that the data object is prefetched (provided there is cache space) by the cache located at the attachment point where the mobile will eventually connect to (hence the name oracle). Unlike the optimal scheme, which allocates cache storage in rounds considering the requests from active mobiles, the EPC, naive and oracle iteratively, for each new cache request, take caching decisions that do not change until the corresponding handoff is performed.

TABLE I. PARAMETER VALUES. THE VALUES DESIGNATED WITH * ARE DEFAULT VALUES, I.E. THE VALUES IF THE SPECIFIC EVALUATION SCENARIO DOES NOT INDICATE OTHERWISE.

Parameter	Values	
	Fixed delay	Scaled-down Internet topology
# of active mobiles	160	160 per neighborhood 10 neighborhoods
# of attachment points	8	8 per neighborhood
Avg. mobile trans probs	SKD50%: $50\%, 20\%, 10\%, 7.5\%, 5\%, 3 \times 2.5\%$ SKD70%*: $70\%, 2 \times 10\%, 3 \times 2.5\%, 2 \times 1.25\%$ SKD90%: $90\%, 3 \times 2\%, 4 \times 1\%$	
Std. dev. of trans probs	$5\%^*, 30\%$	
Delay	$D_M/D_L = 2, 5^*$ $D_R/D_L = 10^*, 18$	$D_M/D_L = 5$ $D_R/D_L = \frac{9}{5}$(#hops-1)+1
Total cache (leaf+mid)	$0 - 320$ objects, default:240^*	

B. Evaluation results

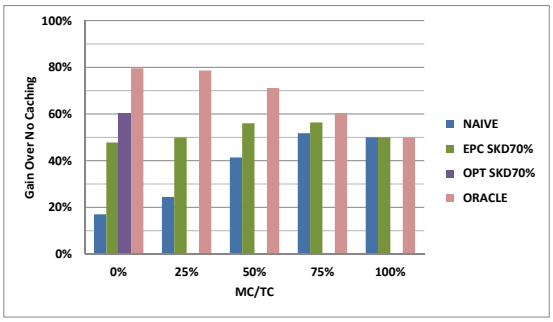
In this section we compare the various schemes in terms of the gains in reducing the average delay for the mobiles to obtain the requested data objects, compared to the delay if no caching is used, i.e. when all data objects are obtained from the original sources. Because the performance of the naive and oracle schemes showed very small dependence on the mobile transition probabilities, for these schemes we do not show results for different transition probabilities.

The results shown are the average of 10 runs for each scenario in the fixed delay case and 10 runs for each neighborhood in the case of the scaled-down Internet topology. Each run has a duration that corresponds to 10.000 handoffs. For these parameters, the 95% confidence interval was within 5% of the average values, hence they are not shown in the graphs.

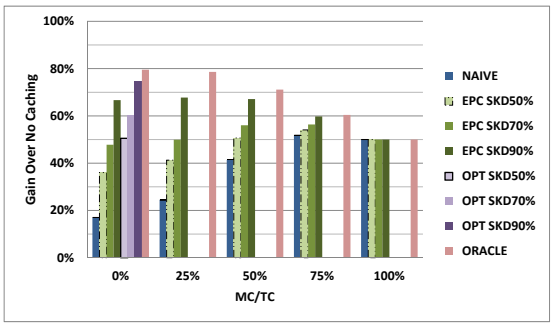
The conclusions from the evaluation are the following:

- The gains of EPC are higher when there is more mobility information, i.e. for a higher skewness of the mobile transition probabilities; these gains are close to those of the optimal scheme for a flat cache structure and the oracle.
- For a higher skewness of the mobile transition probabilities, EPC achieves higher gains when more storage is allocated to leaf caches. Moreover, the gains of EPC are significantly higher than the naive scheme when more storage is allocated to the leaf caches.
- Even for a relatively high variation of the mobile transition probabilities, EPC's gains are robust, and significantly higher than the naive scheme, when the skewness of transition probabilities is high and when more storage is allocated to leaf caches.

1) Comparison of Efficient Proactive Caching (EPC) with the naive, optimal, and oracle schemes: Figure 4(a) considers a fixed amount of total cache storage TC. The x-axis shows the percentage MC/TC of the total cache storage that is allocated to the mid-level cache; hence, 0% indicates that all cache storage is equally distributed to the leaf caches, whereas 100% indicates that all cache storage is allocated to the mid-level cache. Figure 4(a) shows that EPC achieves a higher gain than the naive scheme for small values of MC/TC. Moreover, EPC's gain is close to the optimal scheme, in the case of a flat cache structure ($MC/TC = 0$). On the other hand, for larger values of MC/TC the gains of all schemes are close, and become equal when all storage is allocated to the mid-level cache; this occurs because the capacity of the mid-level cache is larger



(a) SKD70%



(b) SKD50%,70%,90%

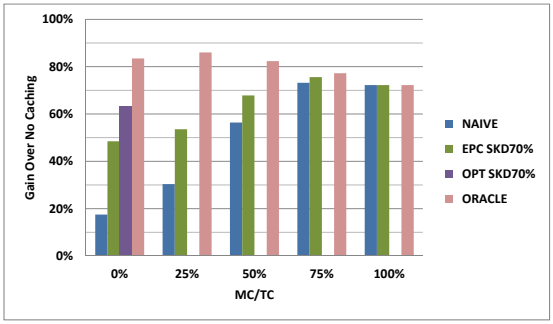
Fig. 4. Influence of mobile transition probabilities on gain. Fixed-delay, $D_R/D_L = 10$, $D_M/D_L = 5$, $TC = 240$ (total cache).

than the total demand and all requested objects can be cached, thus the gains for all schemes are equal and determined by the delay of the mid-level cache. Figure 4(a) also shows that the oracle achieves the best performance when all the storage is allocated to leaf caches, whereas the best performance of the EPC and naive schemes for the specific system parameters is achieved when $MC/TC = 75\%$.

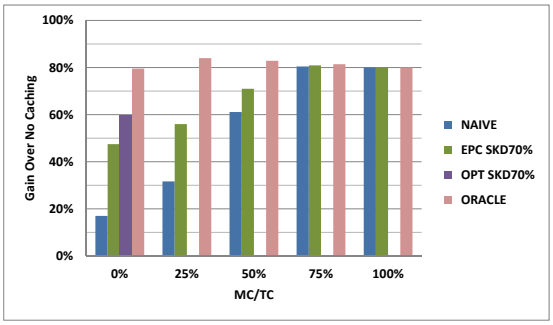
2) Influence of mobile transition probabilities: Figure 4(b) shows that when the *skewness* of the mobile transition probabilities increases, then the gains of EPC are higher when more storage is allocated to the leaf caches. Moreover, when the skewness is large (i.e. SKD90%), then EPC achieves more than 80% of the gains achieved by the oracle and almost 90% of the gains achieved by the optimal scheme for a flat cache structure. It is important to note that, for large skewness, the EPC scheme achieves a higher gain than the naive scheme even when the allocation of storage to leaf and mid-caches is selected *to achieve the best performance for each scheme:* Specifically, for $MC/TC = 25\%$ EPC achieves gain 68%, which is more than 30% higher than the highest gain achieved by the naive scheme, namely 52% for $MC/TC = 75\%$.

3) Influence of delay ratios: Comparing Figure 5(a) to 4(a) shows that when the delay for obtaining data objects from their original sources is higher, then the performance of all schemes is generally higher, especially when more storage is allocated to the mid-level cache (larger values of MC/TC). Figure 5(b) shows that when the delay for obtaining data from a mid-level cache is close to the delay for obtaining the data from a leaf cache ($D_M/D_L = 2$), then EPC achieves the highest gains when all storage is allocated to the mid-level cache.

4) Influence of total cache storage: Next we consider a scenario with the scaled-down Internet topology. Figure 6(a) shows the influence of the total cache storage on the performance in the case of a flat cache structure, while Figure 6(b)

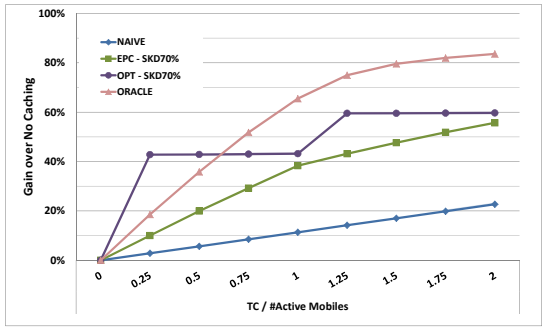


(a) $D_{\mathrm{R}}/D_{\mathrm{L}} = 18, D_{\mathrm{M}}/D_{\mathrm{L}} = 5$

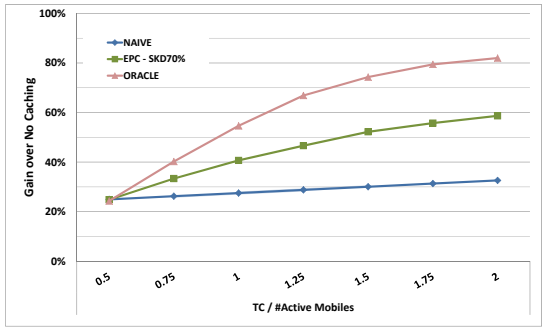


(b) $D_{\mathrm{R}}/D_{\mathrm{L}} = 10, D_{\mathrm{M}}/D_{\mathrm{L}} = 2$

Fig. 5. Influence of delay ratios on gain. Fixed delay, SKD70%, $TC = 240$.



(a) $MC = 0$ (flat cache structure)



(b) $MC = 80$ (mid-level cache)

Fig. 6. Influence of total cache size on gain. Scaled-down Internet topology.

shows the same influence when the mid-level cache of each neighborhood can store 80 objects. Observe that while the gain of the naive scheme increases linearly with the total cache, the gains for the EPC and oracle increase slower for larger values of the total cache. Furthermore, the optimal scheme exhibits a stepwise behavior, which is a result of its operation: the optimal scheme recomputes the cache allocation whenever cache storage is freed; this is also the reason that the optimal scheme has, for small values of cache storage, a higher gain than oracle: the oracle scheme decides where an object is cached when the corresponding request appears; this decision does not change until the corresponding handoff is performed. However, it may happen that when the decision was made there was no available storage at the cache where the mobile will eventually move to. On the other hand, the optimal scheme continuously recomputes the cache allocation when storage space becomes available, hence can take advantage of any free cache space right before a handover is performed.

5) Transient behavior: Figure 7 shows that when the system starts with no knowledge of the mobile transition probabilities, EPC's gain converges to its steady state value slower than the other schemes; this is due to the dynamic updating of cache congestion prices and the measurement-based estimation of mobile transition probabilities. Moreover, as expected, the convergence time is larger for smaller values of factor γ: after a change of the mobile transition probabilities from SKD70% to SKD90% at time 5000, the gain reaches 95% of its steady state value after approximately 750 handovers for $\gamma = 0.5$ and 490 handovers for $\gamma = 8$.

6) Influence of variation of mobile transition probabilities: Comparison of Figures 8 and 4(a) shows that EPC's gains are robust to the variation of the mobile transition probabilities: for values of MC that EPC achieves the highest gains (50% and 75%), an increase of the standard deviation of the transition probabilities from 5% (Figure 4(a)) to 30% (Figure 8) reduces the gains of EPC by less than 10%. Moreover, the higher variation of the transition probabilities appears to influence the oracle more than the other schemes. This is because the higher variation can lead the oracle to require more leaf-level caching than what is available, hence is forced to use the mid-level cache which achieves lower delay gains; this is verified by the higher use of the mid-level buffer by the oracle, for a higher variation of the transition probabilities.

7) Comparison of fixed delay with scaled-down Internet topology: Comparison of Figure 9, which is for the scaled-down Internet topology, and Figure 4(a), which is for fixed delay, shows that the variation due to the Internet topology has a larger influence on the oracle scheme; similar to the above explanation for the higher variation of the mobile transition probabilities, this is because the higher variation results in the oracle scheme requiring leaf-level caching more than the available storage, thus forcing it to use the mid-level cache for which the delay gains are lower.

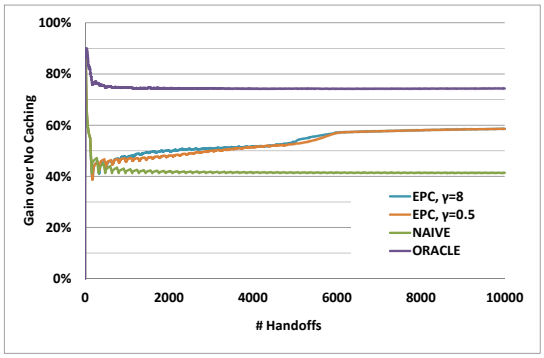


Fig. 7. Transient behavior. The EPC scheme converges slower for smaller values of the price update factor γ. At point 5000 the mobile transition probabilities change from SKD70% to SKD90%. Fixed delay.

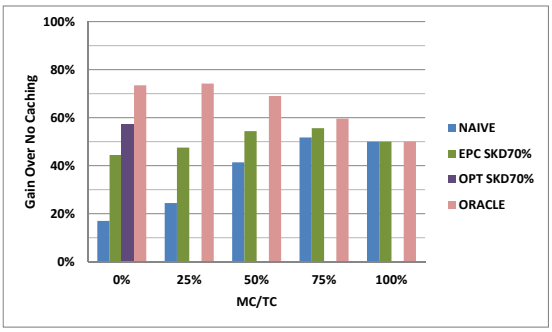


Fig. 8. Influence of variation of mobile transition probabilities. Fixed delay, standard deviation of mobile transition probabilities 30%.

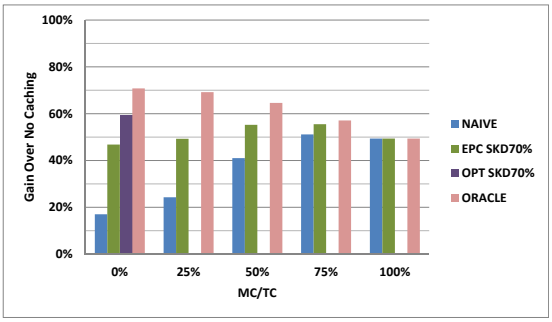


Fig. 9. Scaled-down Internet topology, $D_R/D_L = 10, D_M/D_L = 5$, $TC = 240$ (total cache).

V. Conclusions

We have presented and evaluated a proactive caching scheme for reducing the delay in mobile scenarios, which exploits mobility information and uses congestion pricing to efficiently utilize cache storage. Our modeling framework includes the case of a flat cache structure and a two-level cache hierarchy, and the case where the delay is independent of the size of the requested objects and where the delay is a function of the object sizes. Our evaluation results show how various parameters influence the delay gains of the proposed scheme, which achieves robust and good performance relative to a scheme which attempts to naively cache all data objects, an optimal scheme for a flat cache structure, and an oracle which knows a priori a mobile's future attachment point.

Ongoing and future work includes extending the proactive caching scheme to consider cases where the same data object is requested by more than one mobile user, thus exploiting mobility information together with object popularity. A second extension is to allow a mobile, after a cache miss, to obtain the requested data object not from the original source, but from another cache that proactively fetched the object and is closer to the mobile than the original source; this corresponds to a hierarchy with two or more levels of caches, which can be further motivated when inter-ISP charging costs are present. The proposed models can also be extended to include constraints on the available capacity, such as a low capacity backhaul in femto/small cells and WiFi hotspots, thus capturing in a uniform manner the constraints on cache storage and network capacity. Another direction for future work involves adapting the proposed proactive caching framework that uses congestion pricing to achieve efficient cache utilization to dense femto/small cells and WiFi hotspots with overlapping coverage. The content placement problem in such topologies is NP-hard [5], hence a solution like the one proposed in this paper that uses congestion pricing can have advantages. Finally, an interesting direction is the application of proactive caching and congested resource pricing in mobile software agent scenarios, where mobility involves software relocation rather than device/physical mobility. In this context, the constrained resources can be storage, for delay critical applications that process a large amount of data, or CPU processing.

References

[1] I. Burcea, H. Jacobsen, E. Lara, V. Muthusamy, and M. Petrovic, "Disconnected Operation in Publish/Subscribe Middleware," in *Proc. of IEEE Int'l Conference on Mobile Data Management (MDM)*, 2004.

[2] A. Gaddah and T. Kunz, "Extending mobility to publish/subscribe systems using a pro-active caching approach," *Mobile Information Systems*, vol. 6, no. 4, pp. 293–324, 2010.

[3] V. A. Siris, X. Vasilakos, and G. C. Polyzos, "A Selective Neighbor Caching Approach for Supporting Mobility in Publish/Subscribe Networks," in *Proc. of ERCIM Workshop on eMobility, WWIC*, 2011.

[4] P. Deshpande, A. Kashyap, C. Sung, and S. Das, "Predictive Methods for Improved Vehicular WiFi Access," in *Proc. of ACM MobiSys*, 2009.

[5] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless Video Content Delivery through Distributed Caching Helpers," in *Proc. of IEEE Infocom*, 2012.

[6] F. Malandrino, M. Kurant, A. Markopoulou, C. Westphal, and U. C. Kozat, "Proactive Seeding for Information Cascades in Cellular Networks," in *Proc. of IEEE Infocom*, 2012.

[7] M.-H. Chiu and M. A. Bassiouni, "Predictive Schemes for Handoff Prioritization in Cellular Networks Based on Mobile Positioning," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 510–522, March 2000.

[8] W.-S. Soh and H. S. Kim, "QoS provisioning in cellular networks based on mobility prediction techniques," *IEEE Communications Magazine*, vol. 41, no. 1, pp. 86–92, January 2003.

[9] S. Pack, H. Jung, T. Kwon, and Y. Choi, "SNC: A Selective Neighbor Caching Scheme for Fast Handoff in IEEE 802.11 Wireless Networks," *ACM Mobile Computing and Communications Review*, vol. 9, no. 4, pp. 39–49, October 2005.

[10] A. Wolman, M. Voelker, N. Sharma, N. Cardwell, A. Karlin, and H. Levy, "On the Scale and Performance of Cooperative Web Proxy Caching," in *Proc. of ACM Symp. on Op. Sys. Principles (SOSP)*, 1999.

[11] G. Xylomenos, X. Vasilakos, C. Tsilopoulos, V. A. Siris, and G. C. Polyzos, "Caching and Mobility Support in a Publish-Subscribe Internet Architecture," *IEEE Comm. Mag.*, vol. 50, no. 7, pp. 128–136, 2012.

[12] B. D. Higgins, J. Flinn, T. Giuli, B. Noble, C. Peplin, and D. Watson, "Informed Mobile Prefetching," in *Proc. of ACM MobiSys*, 2012.

[13] D. Lymberopoulos, O. Riva, K. Strauss, A. Mittal, and A. Ntoulas, "Informed Mobile Prefetching," in *Proc. of ACM Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2012.

[14] X. Vasilakos, V. A. Siris, G. C. Polyzos, and M. Pomonis, "Proactive Selective Neighbor Caching for Enhancing Mobility Support in Information-Centric Networks," in *Proc. of ACM SIGCOMM Workshop on Information-Centric Networking (ICN 2012)*, 2012.

[15] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, and K. Ramakrishnan, "Optimal content placement for a large-scale VoD system," in *Proc. of ACM Co-NEXT*, 2010.

[16] I. Baev, R. Rajaraman, and C. Swamy, "Approximation algorithms for data placement problems," *SIAM Journal on Computing*, vol. 38, no. 4, pp. 1411–1429, 2008.

[17] M. R. Korupolu, C. G. Plaxton, and R. Rajaraman, "Placement algorithms for hierarchical cooperative caching," *Journal of Algorithms*, vol. 38, pp. 260–302, 2001.

[18] W.-H. Wang, M. Palaniswami, and S. H. Low, "Application-oriented flow control: Fundamentals, algorithms and fairness," *IEEE/ACM Trans. on Networking*, vol. 14, no. 6, pp. 1282–1291, 2006.

[19] X. Dimitropoulos, D. Krioukov, A. Vahdat, and G. Riley, "Graph annotations in modeling comlex network topologies," *ACM Transactions on Modeling and CoMuter Simulation*, vol. 19, November 2009.

[20] A. Dhamdhere and C. Dovrolis, "Twelve years in the evolution of the internet ecosystem," *IEEE/ACM Trans. on Networking*, vol. 19, no. 5, pp. 1420–1433, 2011.