Proactive Content Distribution for Dynamic Content

John Tadrous, Atilla Eryilmaz and Hesham El Gamal

***Abstract*—We study the bounds and means of optimal caching in overlay Content Distribution Networks (CDN) that serve data with *dynamic* content to end-users who send random requests for the most up-to-date version of such content. Applications with such dynamic content are numerous, including daily news, weather conditions, stock market prices, social networking messages, etc. The service for such a dynamically changing content necessitates a fundamentally different approach than traditional *pull-based* (also called non-proactive) schemes. In particular, *proactive* caching is required to optimize the type and amount of content to be updated in the local servers of a CDN hence minimize the transmission and caching costs, subject to storage constraints.**

We study the metric of *cost reduction* achieved by proactive caching over non-proactive caching strategies. We introduce the notion of *popularity* to establish fundamental upper and lower bounds on cost reduction under different degrees of storage space constraints. We prove the lower bounds to achieve the optimal rate of increase achieved by the upper bounds as the database of items increases. In particular, for a general form of convex, super-linear and monotonically increasing cost functions, our results reveal that the optimal cost reduction scales as the cost function itself, or at least as its first derivative, depending on the number of *popular* data items, as well as the cache storage capacity.

I. Introduction

Content delivery networks offer fast and reliable means of communicating data content from service providers (SPs) to end-users that are sparsely located around the world [1], [2]. Most of CDN providers such as Akamai, and Limelight use an *overlay* structure for CDN organization [3], where the CDN consists of two types of servers: origin servers which contain root content and are updated by the service provider, and replica (or surrogate [2]) servers which contain replica of the root content and are responsible for communicating it to its associated end-users upon request.

Although CDNs improve the quality and speed of delivery as well as reduce the bottleneck demand at the origin servers, there are significant cost issues arising due to transmission of content from both origin and replica servers [1]. Moreover, in dynamic content environment, placement of the new data in the replica servers poses a concern about efficient approaches that would minimize the cost incurred due to unnecessary transmissions [3],[4].

In this paper, we investigate a scenario where data content is highly dynamic while end-users are interested in the most up-to-date version of it. Such content includes daily news, weather conditions, stock market prices, social networking messages, etc. The CDN has a hard delay constraint of supplying the requested content to the end-user at the same slot of demand. This requirement causes most of the user requests to be sent directly from the origin servers at a higher cost. However, based on the recent findings on human behavioral patterns reporting the human behavior to be up to 93% predictable [5], and the emerging work on proactive resource allocation for wireless networks [6]-[8], we propose and study a proactive content outsourcing scheme for urgent demand and dynamic content.

Over the past years there has been a growing research in the deployment of CDNs. It has addressed several aspects of CDN operation and performance ranging from choosing the best place to deploy origin and replica servers [9], [10] to caching and routing strategies of the new content [4], [11]. In [12], [13], different schemes for replica server content eviction have been investigated but neither of them considers a scenario where dynamic content is updated at a high rate. In fact, pull-based schemes [2], which respond reactively to end-user demand are no longer efficient and potentially cause extra undesired cost. Consequently, a better way to respond to such a high content update rate is to proactively (i.e., ahead of time) cache new data in replica servers to utilize the available storage capacity efficiently.

Existing work on proactive content distribution is found in [4], where the considered cost function is the number of hops that requested data has to traverse in order to reach a corresponding end-user in a model representing the CDN as a graph with its nodes being origin servers. A probabilistic model has been used to describe the demand on a certain content and heuristic greedy schemes have been proposed to minimize the expected time average delivery cost. However, the authors do not establish solid performance analysis for the proposed techniques.

In this paper, we consider an overlay CDN setup where the load at an origin server scales linearly with the amount of requests that have to be served, whereas it scales at most linearly with the number of requests at each replica server. The cost incurred due to content delivery is modeled as an increasing convex function in the total load served by each server. Assuming that there are M data items receiving updates at each new slot and that end-users are interested in the updated version of it, we define our cost reduction metric and prove the gains that can leveraged through proactive caching, by characterizing the asymptotic performance of cost reduction as M grows to infinity. We prove the cost reduction to scale as the cost function itself, and in the worst case scenario, it scales as the first derivative of the cost function.

The rest of this paper is organized as follows. In Section II, we layout the system model. In Section III, we formulate the proactive content caching problem. In Section IV, we

Authors are with the Department of Electrical and Computer Engineering at the Ohio State University, Columbus, USA. E-mail: {tadrousj,eryilmaz,helgamal}@ece.osu.edu. The work of John Tadrous is supported by QNRF grant number NPRP 09-1168-2-455. The work of Atilla Eryilmaz is also supported by NSF grants: CAREER-CNS-0953515 and CCF-0916664.

provide the cost reduction scaling results. We validate the analytical results through a numerical simulation in Section V, and conclude the paper in Section VI.

II. System Model

We consider an overlay Content Distribution Network (CDN) structure (e.g., see [2], [3]) whereby a single *origin server* is connected to D *replica servers*. Each replica server d has a storage capacity of size B_d and is serving a set of end-users.

Dynamic Data Content: We consider the service of M *data items* by the CDN, where each such data item represents a different type of dynamically changing information content, as observed in applications of news media, podcasts, online gaming, and social networks. We assume a slotted system operation, with the slot size set to the period at which the content is refreshed and the user requests are received, it can range from seconds to hours depending on the particular application domain. In slot t, each data item m receives a constant update of size $S > 0$ at the origin server[1], which must be supplied to the end-users at the next slot $t+1$, in case it is requested. If the future requests of data items were perfectly known, the optimal caching strategy would simply be the storing of those items with the greatest demand. However, the user demand for the data items in the subsequent slot are only *statistically* known, which requires a careful proactive caching strategy that we study in this paper.

Stochastic Requests: To characterize the stochastic user interests, we use a random process of $\{N_{m,d}^t\}_t$ to count the number of end-users at replica server d requesting data item m over time. In our study, the random variables $N_{m,d}^t$ are assumed to have a positive mean $\mathbb{E}[N_{m,d}^t] > 0$, and to be independent and identically distributed (i.i.d.) over time. For any replica server d, $N_{m,d}^t$ and $N_{j,d}^t$ are independent if $m \neq j$. We further assume a bounded number of requests to each data item. That is, $N_{m,d}^t \leq N$, for all m, d, t and for some finite N.

Cost of Service: We focus on the total cost incurred by the CDN to deliver new content to end-users. We take the cost of serving a given amount of data from a certain server and in a fixed duration to be a smooth, strictly convex, and increasing function of the load $C : \mathbb{R}_+ \to \mathbb{R}_+$.

Further, we assume that the origin server communicates data to the end-users through a *unicast* channel where the load scales linearly in the number of requests. While this assumption can be relaxed, it captures the resource consumption in current networks where service of requests are decoupled. On the other hand, we assume that replica servers can employ more sophisticated techniques, such as *network coding* (c.f. [14], [15]), that can reduce the cost incurred from sending the same content to the end-users. To capture this, we assume that if an amount A_d of data is to be sent from server d to N_d end-users, then the total load at the replica server d scales with N_d as $A_d \cdot g(N_d)$, where $g : \mathbb{N} \to \mathbb{R}_+$ is a non-decreasing and non-negative function with $g(K) \leq K$ on $\mathbb{N}$. This condition formalizes the fact that the replica server is not more costly than the origin server.

[1]We fix the size of the content update for simplicity of notation, the analysis will still hold under a different update size for each data item.

III. Problem Formulation

In this section, we describe the model of operation for both the *non-proactive* content caching scheme which is taken to be a baseline for performance analysis, and the proposed *proactive* content caching scheme.

A. Non-proactive Caching

Under non-proactive caching (known in [2], [16] as pull-based content outsourcing), the CDN servers reactively respond to the initial demand pattern for a newly refreshed content. They wait for users to generate requests and then they pull the new content from the origin server, supply it to the end-users, and cache it for future demand service. In this work, however, we focus on dynamic content updates that arrive at each new slot. Since 1) end-users' demand has to be supplied at the same slot of request emanation, and 2) users are always interested in the most recent version of the requested data items, the CDN will be obliged to serve the requests directly from the origin server, as transmission from origin server to replica servers then from replica server to end-users results in undesirable delay. The total load, therefore, at each replica server d in that case is **zero**.

Thus, the expected total cost incurred by the CDN in time slot t, is written as

$$C_t^{\mathcal{N}}(M) = \mathbb{E}\left[C\left(\sum_{d=1}^{D}\sum_{m=1}^{M} S \cdot N_{m,d}^t\right)\right], \quad \forall t \geq 0, \qquad (1)$$

where the superscript $\mathcal{N}$ in $C_t^{\mathcal{N}}(M)$ denotes the *non-proactive* operation of that system. We consider the pull-based scenario as our baseline for comparison with the proactive caching scheme proposed below.

B. Proactive Caching

The proactive caching scheme is motivated by the recent findings on the predictable human behavior [5] as well as the popularity modeling of web data content introduced in [17] which enables the CDN to construct a demand profile to each data item m, $m = 1, \cdots, M$.

In order to efficiently utilize the available storage space at each replica server in a time slot t, the origin server exploits the available statistics about the demand on each data item to *proactively* send a portion $x_{m,d}^{t+1}$ of each data item m to replica server d. Thus, at the next slot $t+1$, replica server d will be able to directly supply $x_{m,d}^{t+1}$ to the $N_{m,d}^{t+1}$ users requesting item m and the origin server will have to provide the rest of it which is $S - x_{m,d}^{t+1}$ directly to the end-users.

Hence, the total expected cost as a function of the number of data items M and the current slot t will be given by

$$\begin{aligned} C_t^{\mathcal{P}}(M) =& \mathbb{E}\left[C\left(\sum_{d=1}^{D}\sum_{m=1}^{M}(S - x_{m,d}^t)N_{m,d}^t + x_{m,d}^{t+1}\right)\right] \\ &+ \sum_{d=1}^{D}\mathbb{E}\left[C\left(\sum_{m=1}^{M} x_{m,d}^t \cdot g(N_{m,d}^t)\right)\right], \quad t \geq 0. \end{aligned} \qquad (2)$$

where the superscript $\mathcal{P}$ in $C_t^{\mathcal{P}}(M)$ denotes the *proactive* caching approach employed by the CDN.

C. Performance Metric

We consider the time **average expected cost reduction** as our performance metric to analyze the system gains.

First, note that the time average expected cost for the non-proactive scheme is given by

$$C^{\mathcal{N}}(M) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} C_t^{\mathcal{N}}(M) = C_1^{\mathcal{N}}(M), \quad M \in \mathbb{N}, \tag{3}$$

since $\{N_{m,d}^t\}_{t\geq 0}$ is an i.i.d. sequence for every (m,d) pair.

In comparison, the time average expected cost for the proactive scheme is given by

$$C^{\mathcal{P}}(M) = \underset{\{x_{m,d}^k\}_{k\geq 0},\forall m,d}{\text{minimize}} \quad \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} C_t^{\mathcal{P}}(M) \tag{4}$$

$$\text{subject to} \quad 0 \leq x_{m,d}^k \leq S, \quad \forall m,d,k, \quad \sum_{m=1}^{M} x_{m,d}^k \leq B_d, \quad \forall d,k$$

where $B_d = K_d S, \forall d$, with K_d being a positive integer. We consider a CDN scheduler that does not observe the instantaneous realization of the demand $N_{m,d}^t$, i.e., it operates based on the knowledge about the system statistics.

We note that the minimum of the above optimization problem exists and is unique. Existence follows since the objective function is convex; the composition in $x_{m,d}^t$ is linear $\forall m,d$, and the sum of strictly convex functions is strictly convex (c.f. Chapter 3 of [18]), and the constraint set is compact. Uniqueness follows from the strict convexity. Hence, for each (m,d) pair, there exists an optimal sequence $\{x^{*t}_{m,d}\}_{t\geq 0}$ that results in the smallest possible time average cost.

The next lemma is required to prove that $\limsup_{T\to\infty}$ in (4) can be replaced with $\lim_{T\to\infty}$ and hence the time average of proactive expected cost exists. Moreover, the same lemma will be used to prove other crucial results in this paper.

Lemma 1: Let $\{y_n\}_{n\geq 0}$ be a bounded sequence in $\mathbb{R}_+$, then the limit $\lim_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} y_i$ exists and is finite.

Proof. Proof idea is based on showing that $\left\{\frac{\sum_{i=1}^{n-1} y_i}{n}\right\}_n$ is a Cauchy sequence. We refer the interested reader to Appendix A in [19]. ■

Corollary 1: Fix $M \in \mathbb{N}$, then $C^{\mathcal{P}}(M)$ defined in (4) exists.

We are now ready to define the *cost reduction* metric as the time average difference between the non-proactive and proactive costs:

$$\Delta C(M) := C^{\mathcal{N}}(M) - C^{\mathcal{P}}(M), \quad M \in \mathbb{N}. \tag{5}$$

Note that, $C^{\mathcal{N}}(M) \geq C^{\mathcal{P}}(M)$, $\forall M$, as $C^{\mathcal{N}}(M)$ is obtained when $x_{m,d}^t = 0$, $\forall m,d,t$, thus $\Delta C(M) \geq 0$.

In Section IV-C, we analyze the asymptotic behavior of the cost reduction $\Delta C(M)$ as the number of data items M scales.

IV. Cost Reduction Analysis

In this section, we study the performance of optimal cost reduction under proactive content caching. To that end, we first establish the existence of a steady-state solution to (4).

A. Steady-State Solution

Let $\mathbf{x}^{*t}$ denote the vector of optimal proactive downloads $(x^{*t}_{m,d})_{m,d}$ at time t. That is, for a given M, we assume that $\{\mathbf{x}^{*t}\}_t$ is the optimal solution to (4).

Theorem 1: Fix $M \in \mathbb{N}$. Then, there exists a unique vector $\mathbf{x}^*$ such that

$$\lim_{t\to\infty} \mathbf{x}^{*t} = \mathbf{x}^*.$$

Proof. We apply Jensen's inequality to the objective function of (4), and then use Lemma 1 to establish the existence of a steady-state solution. Then, uniqueness follows from the strict convexity of the objective function. We refer interested reader to Appendix B in [19]. ■

From Theorem 1, it turns out that (4) is equivalent to

$$C^{\mathcal{P}}(M) = \min_{x_{m,d},\forall m,d} \mathbb{E}\left[C\left(\sum_{d,m}(S - x_{m,d})N_{m,d} + x_{m,d}\right)\right] + \sum_{d=1}^{D} \mathbb{E}\left[C\left(\sum_{m=1}^{M} x_{m,d}\cdot g(N_{m,d})\right)\right] \tag{6}$$

$$\text{subject to} \quad 0 \leq x_{m,d} \leq S, \quad \forall m,d, \sum_{m=1}^{M} x_{m,d} \leq B_d, \quad \forall d,$$

where we have omitted the time dependence since the random variables $N_{m,d}^1, N_{m,d}^2, \cdots$ are i.i.d. for every (m,d) pair. Thus the CDN needs only to solve the problem once, obtain an optimal solution $\mathbf{x}^* = (x_{m,d}^*)_{m,d}$, and apply it at every time slot in order to achieve the minimum cost.

B. Fundamental Bounds

Now we consider upper and lower bounds on $\Delta C(M)$ which will be used to derive the asymptotic results.

Definition 1 (Popular item): For each data item and replica server pair (m,d), where $m = 1,\cdots,M$, and $d = 1,\cdots,D$, define a marginal cost

$$\mu_{m,d} := \mathbb{E}\left[C'\left(\sum_{u=1}^{D}\sum_{j=1}^{M} SN_{j,u}\right)(N_{m,d} - 1) - C'(0)\cdot g(N_{m,d})\right]. \tag{7}$$

We say that data item m is **popular** with respect to server d if $\mu_{m,d} > 0$. Moreover, denote by $\mathcal{M}_d(M)$ the set of data items that are popular with respect to server d, $d = 1,\cdots,D$, and let $M_d^*(M) := |\mathcal{M}_d(M)|$, the cardinality of $\mathcal{M}_d(M)$. ◇

In the above definition, C' is the first derivative of C. The marginal cost $\mu_{m,d}$ captures the contribution of the data item m, requested by users of replica server d, to the total non-proactive cost. The data item is considered popular if $\mu_{m,d}$ is positive, as popular items qualify for content caching.

Remark 1: In some existing work (see e.g. [17]), the popularity of data items is captured through the probability of requesting each of them, where data items with high such probability are considered popular. In our definition, however, we formulate the popularity from the perspective of the CDN operator. Popular data items have high potential towards increasing the cost, and hence can be proactively cached in order to reduce it. Interestingly, both definitions are not conflicting under a constant data item size S, as data items with high probability of demand can be cached to enhance the cost reduction. Nevertheless, in the more general case,

when each data item has a different size, data items that yield high cost are not necessarily the ones having higher demand probabilities, as the size of the data item contributes substantially to the expected cost.

Now we establish fundamental upper and lower bounds on the cost reduction. Recall $x^*_{m,d}$ is the cached amount from item d in replica server d selected by the *optimal* steady-state proactive strategy.

Lemma 2 (Upper bound on cost reduction): Let $M \in \mathbb{N}$, and for[2] $d = 1, \cdots, D$,

$$\Delta C(M) \leq \mathbb{E}\left[C'\left(\sum_{u=1}^{D}\sum_{j=1}^{M} SN_{j,u}\right)\cdot \sum_{d=1}^{D}\sum_{m\in\mathcal{M}_d} x^*_{m,d}(N_{m,d}-1)\right] - \sum_{d=1}^{D}\mathbb{E}\left[C'(0)\cdot \sum_{m\in\mathcal{M}_d} x^*_{m,d} g(N_{m,d})\right]. \quad (8)$$

Proof. Idea is based on the use of the definition of popularity, the non-negativity of $x^*_{m,d}$ and the mean value theorem (MVT) for random variables. We refer the interested reader to the complete proof in Appendix C of [19]. ■

Now, in order to establish a lower bound on ΔC, we develop a proactive caching policy. As a first step to accomplish this, we define a set $\mathcal{K}_d(M)$ as

$$\mathcal{K}_d := \Big\{ m \in \mathcal{M}_d : \mu_{m,d} \geq \mu_{j,d}, \forall j \in \mathcal{M}_d \backslash \mathcal{K}_d, \; |\mathcal{K}_d(m)| = \min\{K_d, M^*_d\}\Big\}, \quad \forall d = 1, \cdots, D. \quad (9)$$

That is, $\mathcal{K}_d$ contains the $\min\{K_d, M^*_d\}$ data items with highest marginal cost if not cached in server d. Further, we introduce a parameter $\hat{x}$ as the solution to

$$\mathbb{E}\left[C'\left(\sum_{u=1}^{D}\sum_{j=1}^{M} SN_{j,u} + \sum_{j\in\mathcal{K}_u}\hat{x}(1-N_{j,u})\right)\cdot\sum_{d=1}^{D}\sum_{m\in\mathcal{K}_d}(N_{m,d}-1)\right] = \sum_{d=1}^{D}\mathbb{E}\left[C'\left(\sum_{m\in\mathcal{K}_d}\hat{x}\cdot g(N_{m,d})\right)\times\sum_{m\in\mathcal{K}_d} g(N_{m,d})\right] \quad (10)$$

or otherwise, if (10) does not hold for any positive $\hat{x} < S$, we set $\hat{x} = S$. Note that $\hat{x}$ can only be positive which follows from (7), noting that $\mathcal{K}_d$ contains popular items for all d, and C' is monotonically increasing.

Now we propose the following proactive content caching policy.

Definition 2 (Policy A): A proactive caching policy, Policy A, caches a portion $\tilde{x}_{m,d}$ of data item m in server d, where

$$\tilde{x}_{m,d} := \begin{cases} \tilde{x}, & \text{if } m \in \mathcal{K}_d, \\ 0, & \text{if } m \notin \mathcal{K}_d, \end{cases} \quad \forall m, d, \quad (11)$$

and $\tilde{x} = \hat{x} - r$ for some $r > 0$ chosen such that $\tilde{x} > 0$ and

$$\mathbb{E}\left[C'\left(\sum_{u,j} SN_{j,u} + \sum_{j\in\mathcal{K}_u}\tilde{x}(1-N_{j,u})\right)\cdot\sum_{d=1}^{D}\sum_{m\in\mathcal{K}_d}(N_{m,d}-1)\right] > \sum_{d=1}^{D}\mathbb{E}\left[C'\left(\sum_{m\in\mathcal{K}_d}\tilde{x}\cdot g(N_{m,d})\right)\cdot\sum_{m\in\mathcal{K}_d} g(N_{m,d})\right]. \diamond \quad (12)$$

[2]Note that we have omitted the dependence of $\mathcal{M}_d$ on M just for simplifying the notation.

Note that, $\tilde{x}$ exists by the monotonicity of C', and $\tilde{x}_{m,d}$ satisfies the constraints of the optimization (6) for all m, d.

Hence, we can establish the following lower bound on ΔC.

Lemma 3 (Lower bound on cost reduction): Let $M \in \mathbb{N}$, then under Policy A, the cost reduction satisfies

$$\Delta C(M) \geq \tilde{x}\cdot\mathbb{E}\left[C'\left(\sum_{u,j} SN_{j,u} + \tilde{x}\cdot\sum_{j\in\mathcal{K}_u}(1-N_{j,u})\right)\cdot\sum_{d,m\in\mathcal{K}_d}(N_{m,d}-1)\right] - \tilde{x}\cdot\sum_{d=1}^{D}\mathbb{E}\left[C'\left(\sum_{m\in\mathcal{K}_d}\tilde{x}\cdot g(N_{m,d})\right)\sum_{m\in\mathcal{K}_d} g(N_{m,d})\right]. \quad (13)$$

Proof. The proof follows by applying the MVT for random variables to $\Delta C(M)$ and noting that Policy A is not necessarily optimal. We refer the interested reader to Appendix D in [19]. ■

C. Asymptotic Performance

In this subsection, we investigate the *scaling* order of the cost reduction when the number of data items grows to infinity.

Theorem 2: For any strictly convex, and monotonically increasing cost function $C : \mathbb{R}_+ \to \mathbb{R}_+$ satisfying

$$\lim_{L\to\infty}\frac{L^\delta}{C'(L)} = 0, \quad \text{for some } \delta > 0, \quad (14)$$

suppose that there exists a non-negative and non-decreasing function h such that $h(M) \leq M$, $\forall M \in \mathbb{N}$,

$$\limsup_{M\to\infty}\frac{\min\{M^*_d(M), K_d(M)\}}{h(M)} < \infty, \quad \forall d, \quad (15)$$

and

$$\liminf_{M\to\infty}\frac{\min\{M^*_d(M), K_d(M)\}}{h(M)} > 0 \quad \text{for some } d, \quad (16)$$

where $M^*_d(M)$ and $K_d(M)$ are the number of popular data items, and the number of data items that can be stored in server d, respectively. Then:

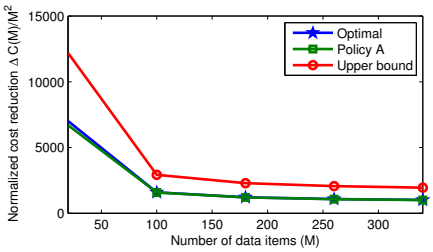
$$\Delta C(M) = \Theta(h(M)C'(\alpha M)), \text{for some } \alpha > 0. \quad (17)$$

Proof. The idea of the proof is based on showing that $\limsup_{M\to\infty}\frac{\Delta C(M)}{h(M)C'(\alpha_2 M)} < \infty$ for some $\alpha_2 > 0$. Then, we prove that $\liminf_{M\to\infty}\frac{\Delta C(M)}{h(M)C'(\alpha_1 M)} > 0$ for some $0 < \alpha_1 \leq \alpha_2$. Hence we conclude that there exists $\alpha \in [\alpha_1, \alpha_2]$ for which $\Delta C(M) = \Theta(h(M)C'(\alpha M))$. We refer the interested reader to the complete proof in Appendix E of [19]. ■

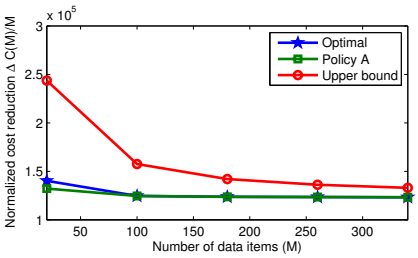
Theorem 2 highlights the asymptotic scaling order of the cost reduction with the number of data items. At this point, the following remarks can be made.

Remark 2: The cost reduction leveraged through proactive content caching grows unboundedly as the number of data items increases. Even if the available storage at the replica servers is finite, i.e., $h(M)$ is finite, still $C'(M)$ grows to infinity with M (from Conditions (14)).

Remark 3: If the number of popular data items and the available buffer storage scale linearly with M, that is $h(M) = M$, then the leveraged cost reduction scales as the cost function itself.



(a) $\Delta C(M)$ scales with M^2, each replica server can store all the popular data items.



(b) $\Delta C(M)$ scales with M, $K_d = 10$, $\forall M$.

Fig. 1: Scaling of $\Delta C(M)$ in different regimes.

Remark 4: Typical cost functions, such as those capturing delays and energy consumption, increase super-linearly in the total load. Therefore, they satisfy Condition (14).

Remark 5: To see the necessity of (14), consider the cost function $C(L) = L - \log(L+1)$, which is strictly convex and increasing, but does not satisfy (14). Assuming $D = 1$, and a deterministic number of requests $N > 1$ targeting each of M data items, then $\Delta C(M)$ does not scale as $\Theta(\frac{\alpha M^2}{1+\alpha M})$ for any $\alpha > 0$.

V. Numerical Results

In this section, the performance of cost reduction under proactive content caching is evaluated numerically. We consider a single origin server and $D = 4$ replica servers. We assume that there are $N = 100$ users covered by each replica server. The random variable $N_{m,d}$ has a binomial distribution with parameter $\phi_{m,d}$, which is a heavy tail distribution to ensure all data items being popular. We assume a quadratic cost function $C(L) = L^2$ and take all the data items to be *popular*, while varying the available buffer size. Throughout the simulation we take $g(N_{m,d}) = N_{m,d}$ which represents a worst-case load at the replica servers. The content update size is set to $S = 5$.

In Fig. 1, we compare the cost reduction scaling under abundant and finite buffer storage conditions. Fig. 1a depicts the cost reduction normalized by M^2 under the optimal policy and Policy A as well as the upper bound derived in (8).

Fig. 1b, on the other hand, shows the cost reduction as it scales with M under a finite buffer constraint. In both scenarios, one can observe the scaling order of the cost reduction, and notice that the optimal policy and Policy A coincide as the number of data items grows.

VI. Conclusion

In this work, we have studied a fundamental question of how to optimally utilize the predictability of a dynamic data content through a proactive content caching scheme. We have formulated and analyzed the problem of minimizing the time average expected cost at a content distribution network (CDN), while data items experience a high content update rate. We have introduced a metric for measuring the popularity of data items, from the CDN's perspective, and proved its efficiency in proactive caching strategies. In particular, under proactive content caching, we have shown that for any convex and super-linearly increasing cost function, unbounded cost reduction gain can be leveraged as the number of data items grows, even if the available storage capacity at replica servers is finite.

References

[1] A. Vakali, and G. Pallis, Content delivery networks: status and trends, *IEEE Internet Computing*, IEEE Computer Society, pp. 68-74, November-December 2003.

[2] G. Pallis, and A. Vakali, Insight and perspectives for content delivery networks, *Communications of the ACM*, vol. 49, vo. 1, ACM Press, NY, USA, pp. 101-106, January 2006.

[3] A. K. Pathan, and R. Buyya, A taxonomy and survey of content delivery networks, *Tech Report*, University of Melbourne, 2007.

[4] J. Kangasharju, J. Roberts, and K. Ross, "Object replication strategies in content distribution networks," *Computer Communications*, vol 25, no. 4, pp. 376383, March 2002.

[5] C. Song, Z. Qu, N. Blumm, and A. Barabas, "Limits of Predictability in Human Mobility", *Science*, vol. 327, pp. 1018-1021, Feb. 2010.

[6] H. El Gamal, J. Tadrous, and A. Eryilmaz, "Proactive resource allocation: Turning predictable behavior into spectral gain," *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, vol., no., pp.427-434, Sept. 29 2010-Oct. 1 2010.

[7] J. Tadrous, A. Eryilmaz, and Hesham El Gamal, "Proactive Multicasting with Predicable Demands," *IEEE International Symposium on Information Theory (ISIT) 2011*, vol., no., pp.239-243, Jul. 2011.

[8] J. Tadrous, A. Eryilmaz, H. El Gamal, and M. Nafie, "Proactive resource allocation in cognitive networks," *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, vol., no., pp.1425-1429, 6-9 Nov. 2011.

[9] L. Qiu, V. Padmanabhan, and G. Voelker, On the placement of web server replicas, *Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. INFOCOM 2001.*, vol.3, no., pp.1587-1596, 2001.

[10] L. Pangfeng, and W. Jan-Jan, "Optimal replica placement strategy for hierarchical data grid systems," *Sixth IEEE International Symposium on Cluster Computing and the Grid 2006. CCGRID 06*, vol.1, no., pp.416-420, 16-19 May 2006.

[11] Y. Chen, R.H. Katz, and J.D. Kubiatowicz, Dynamic replica placement for scalable content delivery, *International Workshop on Peer-to-Peer Systems (IPTPS 02)*, LNCS 2429, Springer-Verlag, pp. 306318, 2002.

[12] M.M. Amble, P. Parag, S. Shakkottai, and L. Ying, "Content-aware caching and traffic management in content distribution networks," *Proceedings IEEE INFOCOM 2011*, vol., no., pp.2858-2866, 10-15 April. 2011.

[13] K. Psounis, and B. Prabhakar,"Efficient randomized Web-cache replacement schemes using samples from past eviction times," *IEEE/ACM Transactions on Networking*, vol.10, no.4, pp. 441- 454, Aug. 2002.

[14] A. Eryilmaz, A. Ozdaglar, M. Medard, and E. Ahmed, "On the delay and throughput gains of coding in unreliable networks," *IEEE Transactions on Information Theory* , vol.54, no.12, pp.5511-5524, Dec. 2008.

[15] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. Wainwright, and K. Ramchandran, "Network Coding for Distributed Storage Systems," *IEEE Transactions on Information Theory*, vol. 56, no. 9, Sept. 2010.

[16] N. Fujita, Y. Ishikawa, A. Iwata, and R. Izmailov, "Coarse-grain replica management strategies for dynamic replication of Web contents," *Computer Networks*, vol. 45, no. 1, pp. 19-34, May 2004.

[17] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon, I tube, you tube, everybody tubes: analyzing the worlds largest user generated content video system, *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, vol., no., pp. 114, 2007.

[18] S. Boyd and L. Vandenberge,"Convex Optimization," *Cambridge University Press*, 2004.

[19] J. Tadrous, A. Eryilmaz, and H. El Gamal, " Proactive content distribution for dynamically changing content," *Technical report*, www2.ece.ohio-state.edu/~tadrousj/ProactiveCaching.pdf