

**Adaptive Caching Networks With Optimality Guarantees**

**Abstract:**

We study the optimal placement of content over a network of caches, a problem naturally arising in several networking applications. Given a demand of content request rates and paths followed, we wish to determine the content placement that maximizes the expected caching gain, i.e., the reduction of routing costs due to intermediate caching. The offline version of this problem is NP-hard and, in general, the demand and topology may be *a priori* unknown. Hence, a distributed, adaptive approximation algorithm for placing contents into caches is desired. We show that path replication, a simple algorithm frequently encountered in literature, can be arbitrarily suboptimal when combined with traditional eviction policies. We propose a distributed, adaptive algorithm that performs stochastic gradient ascent on a concave relaxation of the expected caching gain, and constructs a probabilistic content placement within a 1 *−* 1*/e* factor from the optimal, in expectation. Motivated by our analysis, we also propose a novel greedy eviction policy to be used with path replication, and show through numerical evaluations that both algorithms significantly outperform path replication with traditional eviction policies over a broad array of network topologies.

**Existing System:**

In hierarchical CDNs, requests for content can be served by intermediate caches placed at the network’s edge, e.g., within the same administrative domain (e.g., AS or ISP) as the originator of the request; if, however, content is not cached locally, the request can be forwarded to a core server, which acts as a cache of last resort. Similarly, in CCNs, named data items are stored at designated sources, and requests for named content are forwarded to these sources. Intermediate routers can cache items carried by responses, and subsequently serve future requests. Both settings directly map to the abstract problem we study here.

In these and many other applications, it is natural to assume that the demand, determined by the frequencies of requests and the paths they follow, is dynamic

and not a priori known. For this reason, we seek algorithms that are adaptive, i.e., (a) discover an optimal item placement without prior knowledge of this demand and (b) adapt to its changes. Moreover, collecting information at a single centralized location may be impractical in large networks consisting of different administrative domains. Distributed algorithms, in which a node’s caching decisions rely only on locally available information, allow the network to scale and are thus preferable.

**Proposed System:**

We prove that the ubiquitous path replication algorithm, combined with LRU, LFU, or FIFO eviction policies, leads to allocations that are *arbitrarily suboptimal.* Our result extends to any *myopic* strategy, that ignores costs incurred upstream due to cache misses.

We construct a distributed, adaptive algorithm that converges to a probabilistic allocation of items to caches that is within a 1 *−* 1*/e* factor from the optimal, *without* prior knowledge of the demand (i.e., items requested and routes followed) or the network’s topology. The algorithm performs a projected gradient ascent over a concave objective approximating the expected caching gain.

*•* Motivated by this construction, we also propose a new eviction policy to be used with path replication: whenever an item is back-propagated over a path, the nodes on the path have the opportunity to store it and evict an existing content, according to a greedy policy we design.

*•* We prove that the problem of deterministic caching gain maximization is equivalent to several probabilistic variants. This equivalence has surprising implications. For example, independent caches are as powerful as caches whose contents are coupled, as they attain the same maximal expected caching gain. Moreover, there is no advantage in satisfying relaxed capacity constraints only in expectation, when caches are independent.