





On formulation of online algorithm and framework of near-optimally tractable eviction for nonuniform caches

Thepparit Banditwattanawong ^a  , Masawee Masdisornchote ^b

Show more 

 Share  Cite

<https://doi.org/10.1016/j.comnet.2020.107332> 

[Get rights and content](#) 

Abstract

Distributed data sharing in Internet, social, and cloud computing paradigms incurs nonuniform costs such as consumed downstream bandwidth, downloading delays, and cloud-data-out monetary charges. To alleviate such costs, caches have been widely deployed and researched to find efficient online-cache-eviction algorithms. However, existing online algorithms have no offline foundation, thus they are far from a global optimum. This paper proposes a framework for developing an online cache eviction algorithm, which is grounded in a near-optimally and tractably offline algorithm namely Shortest Maximum Forward Distance (SMFD). On the formulation of the framework, the near-optimality and tractability properties of existing offline algorithms were evaluated through the formal evaluation of the optimality and computational tractability based upon variable object costs. Subsequently, the lowest-complexity suboptimal schemes were empirically investigated to seek a near-optimal offline one, which appears to be SMFD. Also, the results originally show that SMFD and its variant can practically achieve multiple conventionally upper bounds of limitless cache sizes in a stable and simultaneous manner. SMFD was then transformed into novel online SMFD by quantifying the offline property, maximum forward distance, of SMFD with cache-scope time-to-live approximation. To guarantee the effectiveness of online SMFD over evolving request streams, a safety bound guideline and a cache-scope time-to-live distribution model are also proposed. Finally, experience on the framework was gained via experiments based on deep-neural-network models.

Introduction

In modern computing paradigms such as Internet, social, cloud, and fog computing, both big and small data are shared in a distributed manner across the globe. The distributed data sharing incurs nonuniform costs, for instance, origin server loads, monetary server-data-out charges, downloading latencies, and consumed downstream bandwidth. To optimize such costs, the client initiated replication of data, aka caching, is widely adopted in various forms such as traditionally single caching proxies [1], cooperative caches [2], content-centric caches [3], and cloud

caches [4]. The focus of this paper is noncooperative caching in host centric networks as opposite to content centric networks.

Inside every caching system, when a cache space becomes insufficient to resolve a new cache miss, a cache eviction policy or algorithm takes care of cache space acquisition. The policy either makes room in a cache by mandatorily evicting in-cache objects or probably discards to store a requested missing object in the cache. The latter cache-miss resolution is referred to as *optional eviction* (aka bypassing [5]), which relies on the fact that caching the missing object that has a trivial profit or is even no longer requested is not cost-effective. Both mandatory and optional cache evictions are cache-admission control methods [6].

Besides cache admission control, researches on cache eviction can also be divided into offline and online categories. *Offline* cache eviction [5] utilizes the complete sequence of future object requests to search for the optimal cache-state transition. In practice, however, there is neither complete knowledge about future request streams nor a nondeterministic machine that always selects optimal cache state transition. Therefore, an *online* cache eviction policy, which processes a single object request at a time is mandated. Existing efforts [5], [7], [8], [9] studied offline policies to merely seek theoretical bounds for online performances. In fact, the offline policies not only have theoretical but also practical significance on which online policies, both mandatory and optional eviction, can be established as will be demonstrated herein.

Grounded in an offline algorithm namely *SMFD*, this paper devises an online cache eviction algorithm, so-called online *SMFD*, which is practically near-optimal for nonuniform object costs. To achieve this, *SMFD* together with its variant *SMFD** is firstly identified to be near-optimal and tractable through both formally and empirically evaluation of existing offline algorithms. *SMFD* is subsequently elaborated as online *SMFD* by quantifying the offline property called maximum forward distance of *SMFD* through the prediction of a novel concept, cache-scope time-to-live. However, it is natural that supervised-learning predictive models cannot maintain the even levels of prediction errors in the long term due to unpredictably-changing input data. This fact becomes the motivation of the paper to further propose a framework that is capable of ensuring online *SMFD*'s effectiveness against evolving object-request streams. The framework includes not only online *SMFD* but also a safety bound guideline, which enables the rapid and economical revision of artificial-neural-network (ANN)-based predictive models to fit changing target environments. The framework additionally consists of the distribution model of cache-scope time-to-lives used to support prediction method design and workload synthesis. Fig. 1 provides an overview of the framework formulation described above. We also employ Fig. 1 to improve the paper's readability. Since the paper focuses on the framework formulation, the theme of the paper is end-to-end process-centric and organized in a causality manner as depicted in the parentheses in Fig. 1.

The rest of this paper is continued with contribution and significance clarification. Section 2 describes a rationale for a performance metric appropriate for nonuniform-cost data caching within noncooperative and host-centric environments. The same section also extensively provides the optimality, based on such a metric, and tractability proofs of the existing offline policies to find the least-complex and suboptimal policies. Section 3 presents trace-driven analyses to seek near-optimal offline policies in the least-complex suboptimal ones. Section 4 explains the online framework based on a near-optimal offline policy and a new concept namely cache-scope time-to-live. Section 5 discusses the related state-of-the-art efforts. Section 6 concludes with vital findings and opens up problems in the field of study.

The key contributions along with significances of the paper are summarized in a presentation order as follows.

1. In present existence, only some offline cache eviction policies provided optimality proofs whereas the proofs made different and impractical assumptions and totally aimed for a conservative performance goal. Furthermore, merely some existent policies presented their computational

complexities, which were figured out from different scopes of algorithm definitions. This paper reveals the optimality and complexity properties of those existing offline algorithms within the same evaluation environment based on the practical assumption of nonuniform costs, a cost performance goal, and the equivalent scopes of complexity analyses.

The significance of this contribution is vistas for tackling cache eviction algorithms. Specifically, revealing mutually-comparable optimality properties provides an evaluating and classifying means and benchmarks for future policies. Moreover, knowing time complexities is not only crucial to simulation resource estimation but also spotlights tractable policies for practically online implementation.

2. Among the existing offline policies that belong to a suboptimal and tractable class, the paper identifies least-complexity and near-optimal policies, *SMFD* and *SMFD**, in terms of cost saving through trace driven simulation.

SMFD instructs online *SMFD* to lead to both considerable cost- and time-savings. Consider a public-cloud based enterprise utilizing 10-Gbps Internet link with 50% average utilization for 8 working hours a day and 260 workdays a year or 4570.31 TB per annum of data transferred out of the cloud. A near-optimal policy outperforms classical ones up to about 0.5% of network bandwidth saving (based on performance results in this paper). Such bandwidth saving can be translated as 23TB of annual data transfer or 2340 USD per year based on Amazon Web Services cloud data-out charges. Yet, the near-optimal policy saves data transfer time about 0.3% more than the classical policies (based on the performance results in the paper), which can be translated as 6.5 working hours a year.

3. That *SMFD* and *SMFD**, attain the performance upper bounds of limitless cache sizes not only in several performance metrics simultaneously but also in a stable manner across the wide range of cache capacity constraints is highly original.

To the best of our knowledge, there exists no previous research of cache eviction achieving the various performance upper bounds of limitless cache sizes concurrently and tolerating the varied limitations of cache capacities. Both *SMFD* and *SMFD** set breakthrough records and provide the offline world with newly challenging benchmarks.

4. The paper verifies the upper-bound properties of the existing offline policies by simulating them in conjunction with several well-known online policies. The results originally show that not all offline policies set performance upper bounds for online policies.

It has long been unknown whether every offline policy universally confines the performances of online policies. This paper originally verifies that there exist offline policies performing even worse than some online ones in a nonuniform cost environment. Such a finding encourages the more careful perception of an offline world as to its limitation and importantly avoids misleading bounds in future researches.

5. Recent effort claims that the conservative performance upper-bounds of infinitely large caches are unreasonable in both theoretical and practical studies. In contrast, this paper shows that *SMFD* and *SMFD** are able to achieve such upper bounds in several experiments. This finding substantiates the validity of the infinitely large caches for the first time and importantly upholds the engagement of the infinitely large caches in all past and forthcoming researches.

6. The paper originally demonstrates that some specific offline policies not only provide theoretical bounds but are also realizable as online policies. The paper elaborates a new concept cache-scope time-to-live to turn *SMFD* into novel online *SMFD*. The paper then proposes a framework to enable the flexible development of online *SMFD* with guaranteed performances.

With the current advancement of machine learning techniques and tools, this paper develops online *SMFD* by using a deep-neural-network technique. Unfortunately, by the nature of neural networks, online *SMFD* cannot remain effective in deployment environments that evolve unpredictably, thus mandating the development framework as aforementioned.

Access through your organization

Check access to the full text by signing in through your organization.

Access through your organization

Section snippets

Optimality and tractability evaluation

This section initially defines the optimality quality of cache eviction algorithms followed by the theoretical evaluation of existing offline algorithms.

Object hit rate (OHR) has long been used to quantify the spatiotemporal similarity (i.e., requests to objects in a cache indicate forthcoming requests to the objects in the cache) of in-cache objects. In the past World Wide Web (WWW), online cache eviction policies extensively optimized OHR (or its inversion, an object miss count) by favoring ...

Near-Optimal algorithm identification

There has been no insight as to which of the suboptimal and lowest complexity schemes performs best (i.e., near-optimal) in terms of CSR to serve online policy development. Identifying the near-optimal offline policies by engaging a formal approximation technique assuming IRM, unfortunately, disregards temporal and spatial localities [20], [21] and is neither practical nor the intent of this paper stemmed from Definition2. Instead, this section engages the conventional means of ...

Online framework development

Establishing an online policy grounded in an offline one avoids local optimum and brings about near-optimality. Among the offline policies of the same tractable class (Cf. Table 1), the benefits of *SMFD* and *SMFD** lie in their minimum number of parameters unlike profit-based schemes such as C_0 and C_0^* that incorporate several parameters. However, it is intuitive that converting *SMFD* and *SMFD** to online versions without the support of future access information tends to lead to inefficient results ...

Related work

Offline schemes are explored beyond those in Section2 as follows. Seeking an optimal offline approach was formulated as the shortest path problem and solved with an optimality principle in *Hosseini – Khayat's OPT*,

BL , and PE [8]. C_0 [12], originally named $OPT - C$ [34], and C_0^* relied on Markov Decision Process [12] and IRM. Mattson's OPT has long been known to be optimal for uniform objects [17], [35]. Longest forward distance-any ($LFD - A$ or $LFD - any$) [36] is a variant of Mattson's OPT for optional ...

Conclusion

This paper has enlightened for the first time the fair comparison of offline cache-eviction schemes and demonstrated that not all of them were unrealizable. In particular, existing offline policies including $SMFD$ and $SMFD^*$ have been classified into (sub)optimal and computationally (in)tractable policies. The tractable suboptimal policies have been further examined by trace-driven simulation that identified $SMFD$ and $SMFD^*$ as near-optimal policies for nonuniform object costs. Both $SMFD$ and $SMFD^*$...

CRedit authorship contribution statement

Thepparit Banditwattanawong: Conceptualization, Formal analysis, Methodology, Validation, Visualization, Writing - original draft, Writing - review & editing. **Masawee Masdisornchote:** Investigation, Resources, Writing - review & editing. ...

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. ...

Thepparit Banditwattanawong was born in Bangkok, Thailand. He received B.Eng. degree (Honors) in computer engineering from King Mongkut's Institute of Technology Ladkrabang, Thailand and M.Eng. degree from Asian Institute of Technology, Thailand. He obtained his Ph.D. degree in informatics from the National Institute of Informatics (NII), The Graduate University for Advanced Studies, Tokyo, Japan. He is currently an Assistant Professor with the Department of Computer Science, Kasetsart ...

...

...

[Recommended articles](#)

Research data for this article

 Data not available / The authors do not have permission to share data

Open Data

for download under the [CC BY licence](#) ↗

Supplementary Data S1

(XML, 260B)

Supplementary Raw Research Data. This is open data under the CC BY license <http://creativecommons.org/licenses/by/4.0/> ↗

[Download data](#)

[Further information on research data](#) ↗

References (54)

F. Khandaker *et al.*

[A functional taxonomy of caching schemes: towards guided designs in information-centric networks](#)

Comput. Networks (2019)

T. Banditwattanawong *et al.*

[Multi-provider cloud computing network infrastructure optimization](#)

Future Generation Computer Systems (2016)

D. Starobinski *et al.*

[Probabilistic methods for web caching](#)

Perform. Eval. (2001)

D.S. Berger *et al.*

[Exact analysis of TTL cache networks](#)

Performance Evaluation (2014)

E. Cohen *et al.*

[Performance aspects of distributed caches using TTL-based consistency](#)

Theor. Comput. Sci. (2005)

D. Wessels

[Squid: the definitive guide](#)

(2004)

D. Wessels *et al.*

[Icp and the squid web cache](#)

IEEE J. Sel. Areas Commun. (1998)

M. Brehob *et al.*

[Optimal replacement is NP-hard for nonstandard caches](#)

IEEE Trans. Comput. (2004)

C. Aggarwal *et al.*

[Caching on the world wide web](#)

IEEE Trans Knowl Data Eng (1999)

D.S. Berger *et al.*

[Practical bounds on optimal caching with variable object sizes](#)

Proc. ACM Meas. Anal. Comput. Syst. (2018)



View more references

Cited by (0)



Thepparit Banditwattanawong was born in Bangkok, Thailand. He received B.Eng. degree (Honors) in computer engineering from King Mongkut's Institute of Technology Ladkrabang, Thailand and M.Eng. degree from Asian Institute of Technology, Thailand. He obtained his Ph.D. degree in informatics from the National Institute of Informatics (NII), The Graduate University for Advanced Studies, Tokyo, Japan. He is currently an Assistant Professor with the Department of Computer Science, Kasetsart University, Bangkok, Thailand. His main areas of research interests include computer network optimization, distributed computing, and cloud computing.



Masawee Masdisornchote received B.S. degree from Silpakorn University, Thailand and M.S. degree from King Mongkut's Institute of Technology Ladkrabang, Thailand. She is currently a full-time assistant professor with Sripatum University, Bangkok, Thailand. Her main research areas include data mining.

[View full text](#)

© 2020 Elsevier B.V. All rights reserved.



All content on this site: Copyright © 2025 Elsevier B.V., its licensors, and contributors. All rights are reserved, including those for text and data mining, AI training, and similar technologies. For all open access content, the relevant licensing terms apply.

