Self-Adjusting Cache Advertisement and Selection

Itamar Cohen
Ariel University
Ariel, Israel
itamarc@ariel.ac.il

ABSTRACT

We present a lightweight, self-adjusting algorithm for cache-content advertisement and cache selection. Our algorithm increases the hit ratio and mitigates wasteful, unnecessary cache accesses and cache-content advertisements.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Access strategies, cooperative caching, replica selection.

1 PROBLEM

In distributed caching systems, a client that looks for data selects some of the caches and queries them. Each query is associated with an *access cost* representing, e.g., latency, energy, or bandwidth [1]. Failing to retrieve the requested datum from any of the queried caches incurs a high *miss penalty* representing, e.g., the cost of accessing a remote server. The goal is to minimize the *service cost*, defined as the overall caches' access costs, plus the overall miss penalties along the trace.

To minimize the service cost, the caches periodically advertise *indicators* (e.g., Bloom filters [4]) that summarize the list of cached items. However, due to space and bandwidth constraints, these indicators exhibit false indications [2–4]. State-of-the-art solutions [3] incur excessive overhead and are not responsive enough to dynamic changes in the workload.

2 SOLUTION

Our solution, Self-Adjusting Lightweight Selection and Advertisement Algorithm (SALSA2), works as follows. Each cache maintains hit/miss ratio statistics, and advertises an updated indicator only if (i) the observed current false indication ratio crosses a pre-defined threshold, or (ii) the ratio of new items in the cache crosses some threshold. Caches periodically update the client about the observed false indication ratio. The client injects these observed false indication ratios to the PGM cache selection algorithm [1] to cleverly select which caches to query.

Fig. 1 compares the performance of the state-of-the-art solution [3] and SALSA2 for varying cache traces and miss penalty values. The hit ratio captures the ratio of items retrieved from a cache. Service costs are normalized w.r.t. a hypothetical, optimal scheme with perfect knowledge about the data locations. The bandwidth is presented as the overall number of bits sent to advertise the caches' content over the number of requested items. The results show that our algorithm achieves higher hit ratios, comparable or lower service costs, and lower bandwidth across the board.

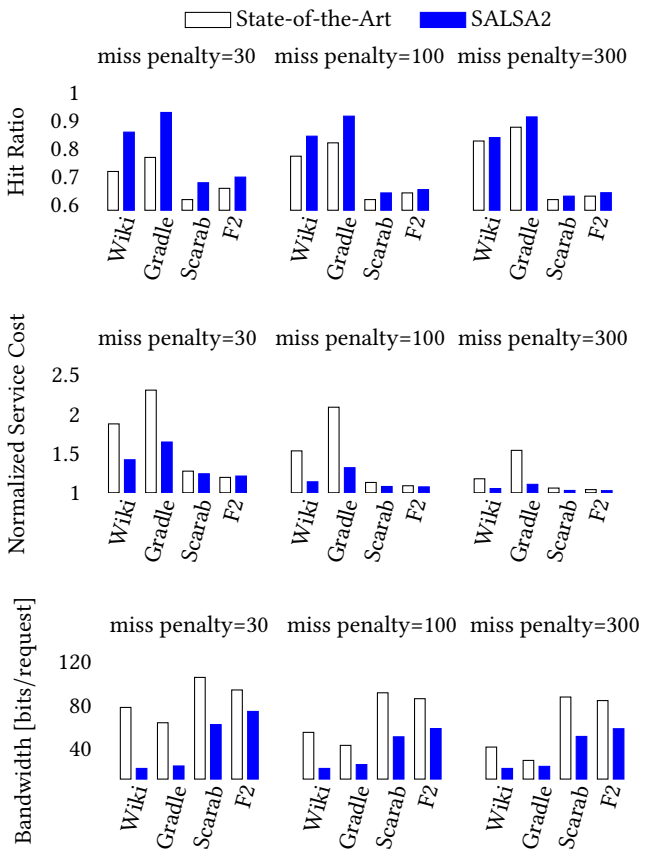

Figure 1: Hit ratio (top), normalized service cost (middle), and bandwidth consumption (bottom) comparison.

REFERENCES

[1] Itamar Cohen et al. 2021. Access Strategies for Network Caching. *IEEE/ACM ToN* 29 (2021), 609–622. Issue 2.
[2] Itamar Cohen, Gil Einziger, and Gabriel Scalosub. 2021. Self-adjusting advertisement of cache indicators with bandwidth constraints. In *IEEE INFOCOM*. 1–10.
[3] Itamar Cohen, Gil Einziger, and Gabriel Scalosub. 2022. False Negative Awareness in Indicator-Based Caching Systems. *IEEE/ACM ToN* 30, 6 (2022), 2674–2687.
[4] Sasu Tarkoma, Christian Esteve Rothenberg, and Eemil Lagerspetz. 2012. Theory and practice of Bloom filters for distributed systems. *IEEE Comm. Surv. & Tut.* 14, 1 (2012), 131–155.

SYSTOR '23, June 5–7, 2023, Haifa, Israel

ACM ISBN 978-1-4503-9962-3/23/06.
https://doi.org/10.1145/3579370.3594754