Delayed Hits in Multi-Level Caches

Benjamin Carleton[1], Nirav Atre[2], Justine Sherry[2], Weina Wang[2]
[1]*University of Rochester,* [2]*Carnegie Mellon University*

Background

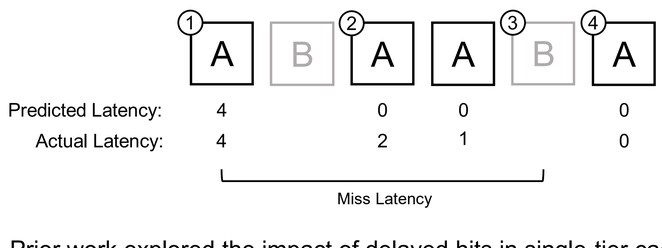
- Traditional caching models assume that outstanding requests are resolved before new requests arrive
- High-throughput systems violate this assumption

① A request arrives for object **A**, resulting in a cache miss. A fetch is sent to the backing store

② Another request for **A** arrives before the fetch returns, resulting in a *delayed hit*

③ **A** arrives in the cache and the requests are served

④ The next request for **A** results in a true hit



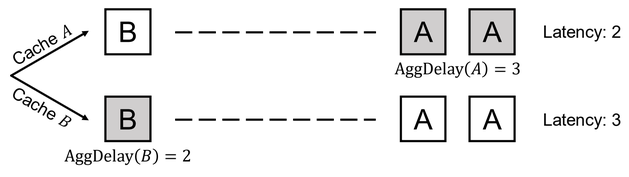
- Prior work explored the impact of delayed hits in single-tier caches and introduced MAD, a delayed-hits-aware caching algorithm capable of reducing CDN latencies by 12–18% [1]

Minimum Aggregate Delay (MAD)

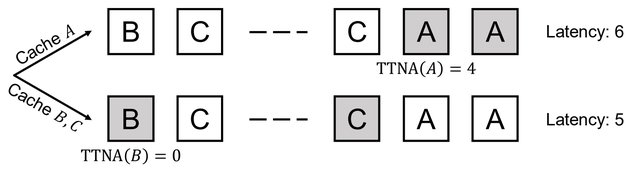
- An object's *aggregate delay* is the total latency incurred by a miss and all delayed hits that occur while the object would be fetched
- MAD: evict the object with the lowest rank:

$$\text{Rank}(x) = \frac{\text{AggDelay}(x)}{\text{TTNA}(x)}$$

The Need for Aggregate Delay:



The Need for Time to Next Access:



Motivation

- Prior investigation into the effects of delayed hits considered single-tier cache configurations; real systems such as CDNs often comprise multiple tiers
- Delayed hits cause real-world performance to diverge from the predictions of traditional caching models
- Policies that maximize hit rate may be suboptimal

Objectives

Characterize the effects of delayed hits in multi-tier caches:

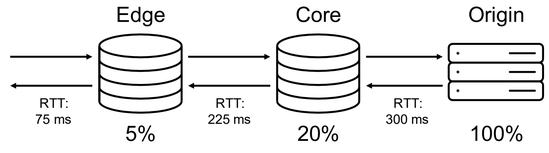
1. For multi-tier caches exhibiting delayed hits, to what extent do true latencies diverge from those predicted by traditional caching models?
2. Can extending delayed-hits-aware policies for use in multi-tier configurations yield improved latencies?

Methods

- Augment an existing caching simulator to accurately model the multi-tier caches used in real-world systems
- Simulate an empirical CDN cache configuration on a CDN trace with a high request rate

CDN Configuration

- Tier latencies based on Fastly's CDN [2]

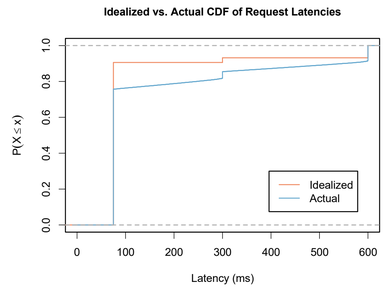


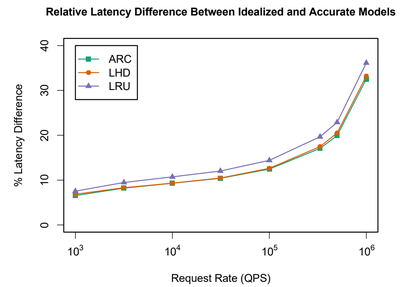
Extending MAD

- Aggregate delay windows must be sized according to the cache miss latency, but, in a multi-tier cache, the time to fetch an object following a cache miss is not fixed
- Multi-tier MAD: adaptively calculate parameters that cannot be determined statically
- Dynamic miss latencies are recorded at each cache tier, and aggregate delay windows are sized according to the cumulative average miss latency

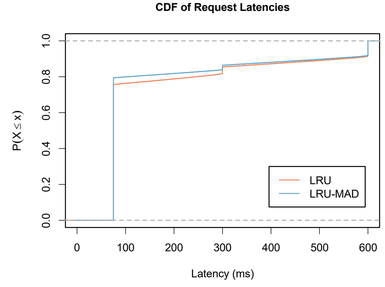
Evaluation

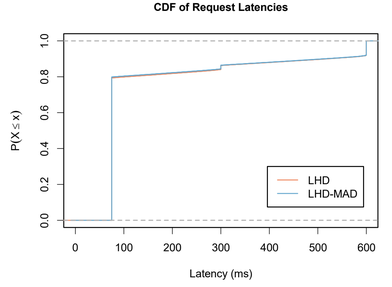
- At high request rates, true latencies diverge from predicted latencies by 32.53–36.13%

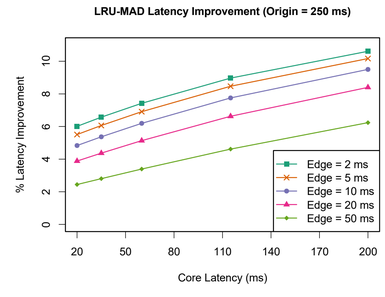


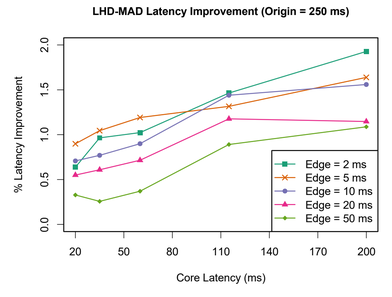


- MAD yields a latency improvement of 0.85–5.61% with our empirical cache configuration
- Synthetic configurations see improvements of 10.61%









Discussion

- Delayed hits can still be a prominent factor in the performance of multi-tier caches, although their effect in this setting is diminished in comparison to the single-tier setting

[1] Atre et al. 2020. Caching with Delayed Hits. *(SIGCOMM '20)*.
[2] Ghabashneh & Rao. 2020. Exploring the Interplay between CDN Caching and Video Streaming Performance. *(IEEE INFOCOM 2020)*.