On Resource Pooling and Separation for LRU Caching

Jian Tan, Guocong Quan, Kaiyi Ji, Ness Shroff
The Ohio State University, Columbus, OH, USA
{tan.252,quan.72,ji.367,shroff.11}@osu.edu

ABSTRACT

Caching systems using the Least Recently Used (LRU) principle have now become ubiquitous. A fundamental question for these systems is whether the cache space should be pooled together or divided to serve multiple flows of data item requests in order to minimize the miss probabilities. In this paper, we show that there is no straight yes or no answer to this question, and depends on complex combinations of critical factors, including, e.g., request rates, overlapped data items across different request flows, data item popularities and their sizes. To this end, we characterize the performance of multiple flows of data item requests under resource pooling and separation when the cache size is large.

Analytically we show that it is asymptotically optimal to jointly serve multiple flows if their data item sizes and popularity distributions are similar, and their arrival rates do not differ significantly; the self-organizing property of LRU caching automatically optimizes the resource allocation among them asymptotically. Otherwise, separating these flows could be better, e.g., when data sizes vary significantly. We also quantify critical points beyond which resource pooling is better than separation for each of the flows when the overlapped data items exceed certain levels. These results provide new insights on the performance of caching systems.

KEYWORDS

Caching algorithm; LRU; Miss probability; Memcached

ACM Reference Format:
Jian Tan, Guocong Quan, Kaiyi Ji, Ness Shroff. 2018. On Resource Pooling and Separation for LRU Caching. In *SIGMETRICS '18 Abstracts: ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems Abstracts, June 18–22, 2018, Irvine, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3219617.3219628

MAIN RESULT

The current engineering practice for deciding resource pooling or separation to optimize LRU caching presents a dilemma. On the one hand, a shared cache space that provides sufficiently high hit ratios for multiple flows of data item requests can improve the utilization of the system, especially when a data item brought into cache by one flow can be directly used by the other. On the other hand, resource separation facilitates capacity planning and ensures adequate quality of service for each of the flows. This dilemma only scratches the surface. In fact, four critical factors jointly impact the cache miss probabilities, including *request rates*, *overlapped data items* across different request flows, *data item popularities* and *their sizes*. Depending on the setting, it may be better to pool or separate the cache.

To systematically investigate this problem, we theoretically characterize the asymptotic overall miss ratios achieved by resource pooling and the optimal separation method under a large class of heavy-tailed popularity distributions, e.g., Zipf's distributions, when the cache sizes tend to infinity. Our results demonstrate the complexity of the optimal strategy: resource pooling can be asymptotically equal to, better or worse than separation, respectively. For practical insights, the following engineering implications can be derived for typical systems with large caching spaces under different settings, when reducing the overall miss ratio is the main goal.

1) Resource pooling and separation are asymptotically equal
The optimal resource separation scheme is believed to be better than pooling for caching systems. However, it is not clear whether the difference is significant or not when the cache size is large. For multiple disjoint flows of requests with (nearly) identical item sizes, we show that the overall miss ratio achieved by resource pooling is asymptotically equal to that achieved by the optimal separation method, due to the self-organizing behavior of LRU caching.

2) Separation is better than pooling
Interestingly, when the data item sizes are very different among multiple disjoint flows of data item requests, we show that the optimal resource separation can achieve lower overall miss ratios than resource pooling under certain conditions.

3) Pooling is better than separation
If the data item request rates or the popularities are time-varying, a static separation method cannot always achieve the optimal overall miss ratio. However, due to the self-organizing behavior, resource pooling adaptively adjusts to the optimal allocation, as long as the request rates and the popularities are relatively stable during the adjusting periods. Moreover, if there exist overlapped data items across these flows, the asymptotic miss ratio of each individual flow can be lower under resource pooling than under the optimal separation method when the popularities, the request rates, and the amount of overlapped data items satisfy specific conditions.

We conduct extensive simulations and successfully validate all of the theoretical results. These results provide new insights that can potentially further improve the performance of caching systems in real practice. The details are presented in the full paper [1].

This work was supported by the National Science Foundation under Grant No. 1717060 and No. 1719371.

SIGMETRICS '18 Abstracts, June 18–22, 2018, Irvine, CA, USA

ACM ISBN 978-1-4503-5846-0/18/06.
https://doi.org/10.1145/3219617.3219628

REFERENCES

[1] Jian Tan, Guocong Quan, Kaiyi Ji, and Ness Shroff. 2018. On resource pooling and separation for LRU caching. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 2, 1 (March 2018).