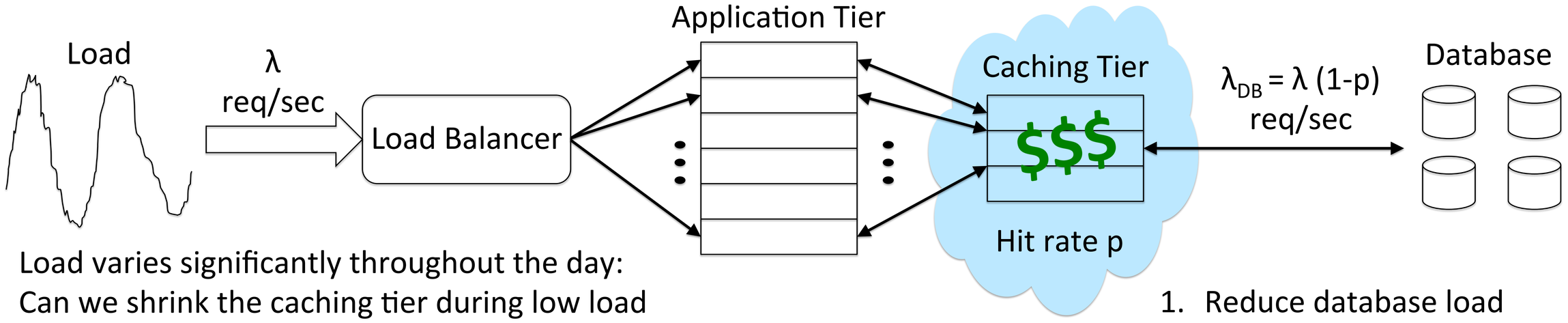
CacheScale: Saving Cash by Using Less Cache

Timothy Zhu, Anshul Gandhi, and Mor Harchol-Balter (Carnegie Mellon University); Michael A. Kozuch (Intel Labs)

Problem Overview

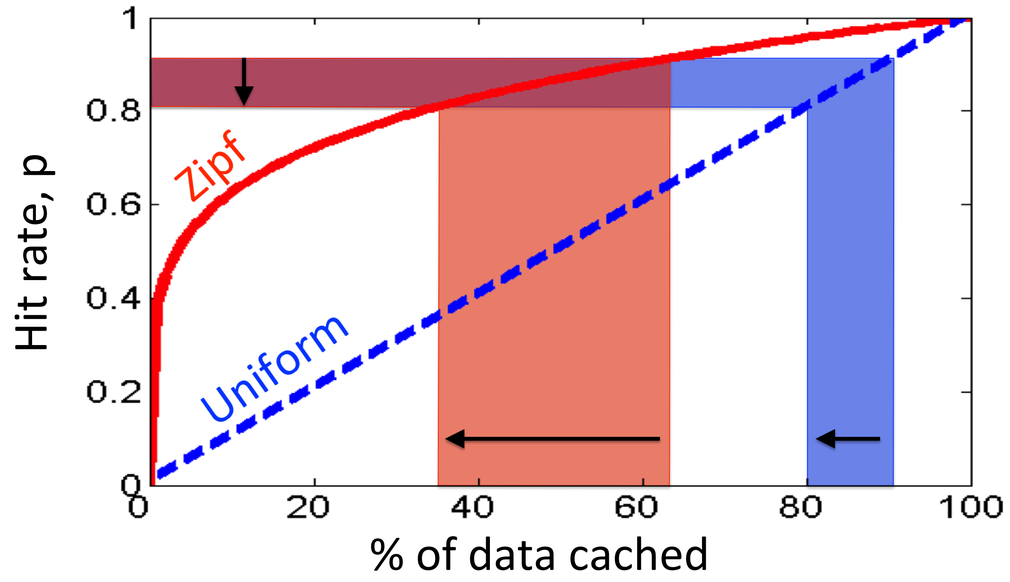


Load varies significantly throughout the day:
Can we shrink the caching tier during low load to reduce operational costs?

Caching tier helps to:

1. Reduce database load
2. Reduce data access latency

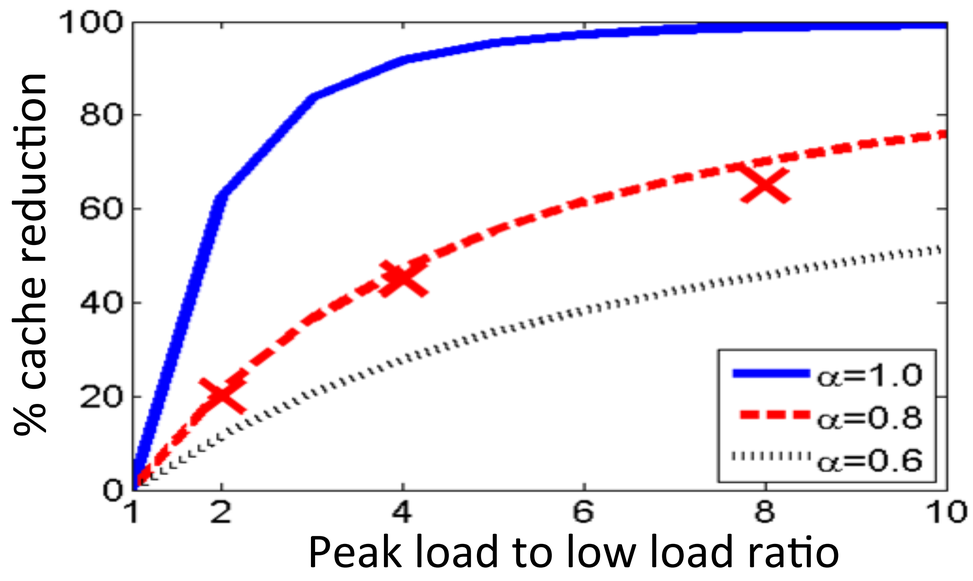
Popularity distribution



Small decrease in desired hit rate

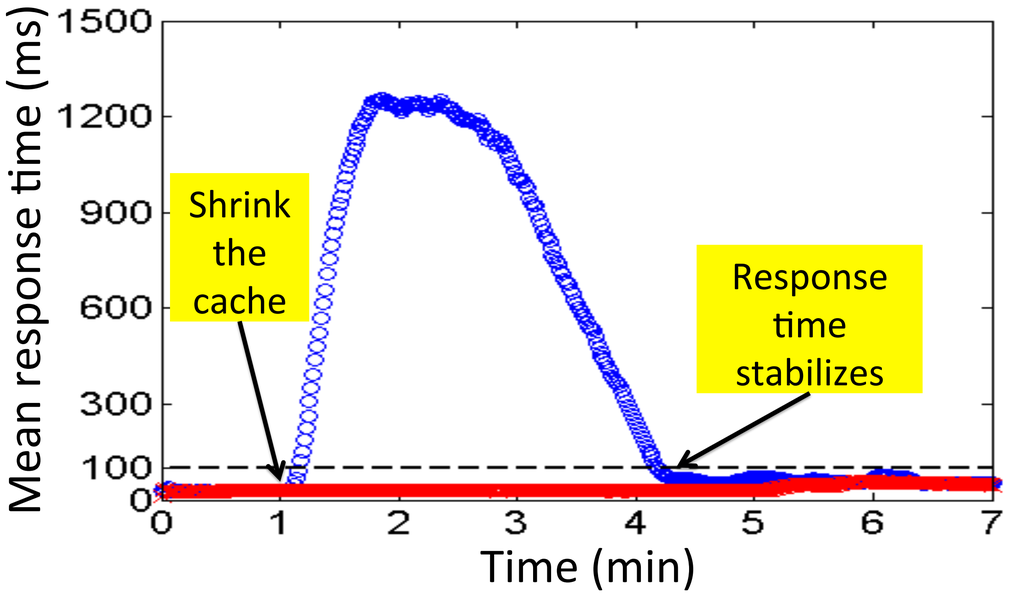
- Uniform → Small decrease in required cache size
- Zipf → **Large** decrease in required cache size

Results



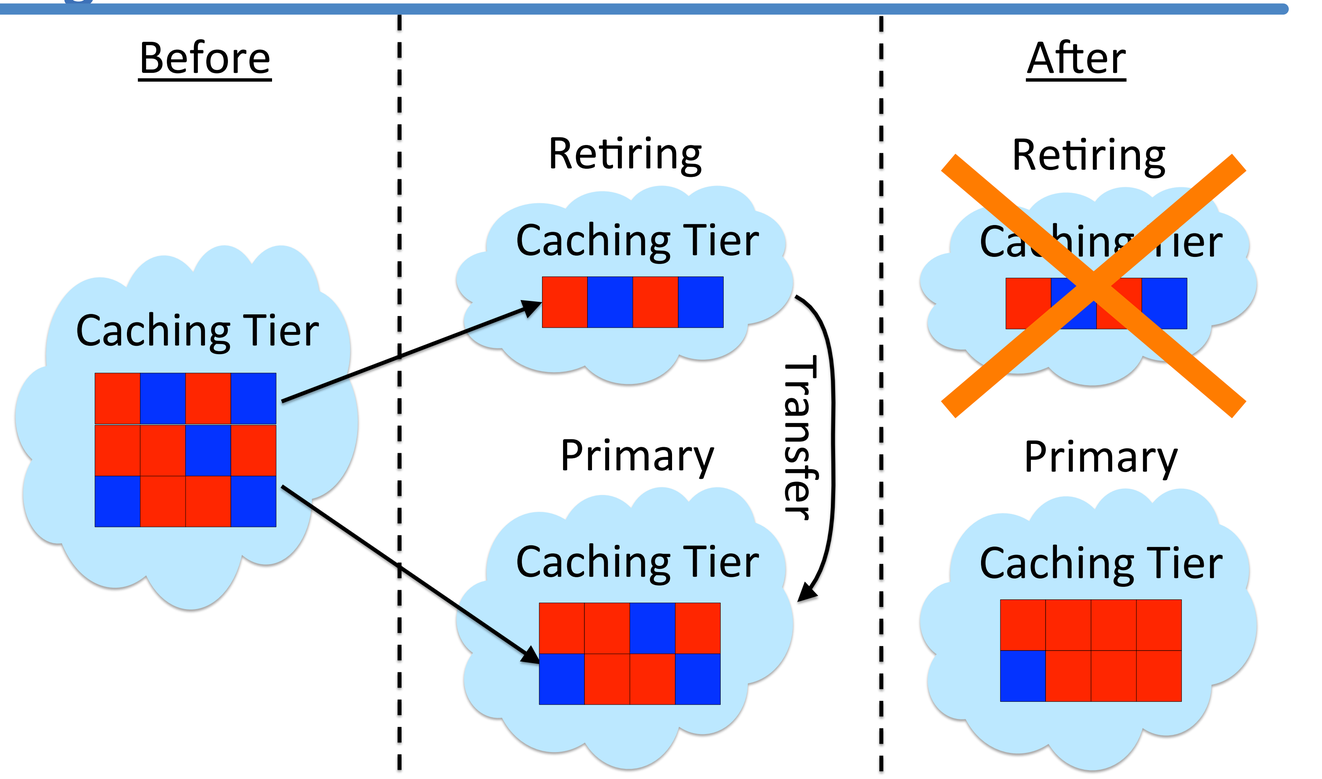
- Larger ratios between peak load and low load
  ➔ More potential for cache size reduction
- Substantial savings for a range of Zipf popularity distributions with varying skew parameters, α

Transferring "hot" data



Performance can temporarily suffer when shrinking the cache due to losing a lot of "hot" data

Transferring the hot data before shrinking the cache mitigates this problem



Divide cache instances into retiring group and primary group

If incoming request hits in retiring group, then transfer data to primary group

Intel Science & Technology Center for Cloud Computing

Carnegie Mellon University · Georgia Tech · intel · Princeton University · UC Berkeley