Saving Cash by Using Less (Mem)Cache

Timothy Zhu
Carnegie Mellon University

Anshul Gandhi, Mor Harchol-Balter
Carnegie Mellon University

Michael A. Kozuch
Intel Labs
Why have a caching tier?

1. Reduce database (DB) load \( (\lambda_{DB} << \lambda) \)
Why have a caching tier?

1. Reduce database (DB) load \((\lambda_{\text{DB}} \ll \lambda)\)
Why have a caching tier?

1. Reduce database (DB) load \( (\lambda_{DB} \ll \lambda) \)
2. Reduce latency
Application in the Cloud

Shrink your cache during low load

Load Balancer

Application Tier

Caching Tier

Database

$\lambda$ req/sec

$\lambda_{DB}$ req/sec

$\lambda_{req/sec}$

> 1/3 of the cost

[Krioukov'10]
[Chen'08]

[Ousterhout'10]
Key Questions

1. Will cache misses overwhelm the DB?  
   \( \lambda_{DB} \) too high?

2. Are the savings significant?

3. What about the “hot” data?
Key Questions

1. Will cache misses overwhelm the DB? 
   \( \lambda_{DB} \) too high?

2. Are the savings significant?

3. What about the “hot” data?
Will cache misses overwhelm the DB?

**Goal:** Keep $\lambda_{DB} = \lambda(1-p)$ low

If $\lambda$ drops, $(1-p)$ can be higher, $p$ can be lower, and **SAVE $$**
1. Will cache misses overwhelm the DB?
   No, we can afford a lower hit rate at low load

2. Are the savings significant?

3. What about the “hot” data?
1. Will cache misses overwhelm the DB?
   No, we can afford a lower hit rate at low load

2. Are the savings significant?

3. What about the “hot” data?
Are the savings significant?

- It depends on the popularity distribution

Small decrease in hit rate

Uniform

Small decrease in caching tier size
Are the savings significant?

- It depends on the popularity distribution

Small decrease in hit rate

Zipf

Large decrease in caching tier size
Savings

50% cache savings

Load

4x

% cache reduction

Peak to low load ratio

More skewed

Less skewed
1. Will cache misses overwhelm the DB?
   No, we can afford a lower hit rate at low load.

2. Are the savings significant?
   Small decrease in hit rate
   Large decrease in Zipf caching tier size

3. What about the “hot” data?
Key Questions

1. Will cache misses overwhelm the DB?
   No, we can afford a lower hit rate at low load

2. Are the savings significant?
   Small decrease in hit rate  Large decrease in caching tier size

3. What about the “hot” data?
   a. Is there a problem?
   b. What can we do about it?
Performance can temporarily suffer if we lose a lot of hot data.
What can we do about the hot data?

Start state

End state

Caching Tier
What can we do about the hot data?

We need to transfer the hot data before shrinking the cache.
Effect of transferring hot data

- Transferring the hot data before shrinking the cache eliminates performance degradation.

![Graph showing the effect of transferring hot data before shrinking the cache](image_url)

- Naïve solution
  - Mean response time (ms)
  - Shrink the cache
  - Response time stabilizes

- Our solution
1. Will cache misses overwhelm the DB?
   No, we can afford a lower hit rate at low load

2. Are the savings significant?
   Small decrease in hit rate \( \rightarrow \) Large decrease in Zipf caching tier size

3. What about the “hot” data?
   We need to transfer the hot data before shrinking the cache

Use less cache \( \rightarrow \) Low load \( \rightarrow \) Save $$$