When is the Cache Warm? Manufacturing a Rule of Thumb

Lei Zhang
Emory University

Juncheng Yang
Carnegie Mellon University

Anna Blasiak
Indigo Inc/ Akamai Inc

Mike McCall
Facebook Inc/ Akamai Inc

Ymir Vigfusson
Emory University

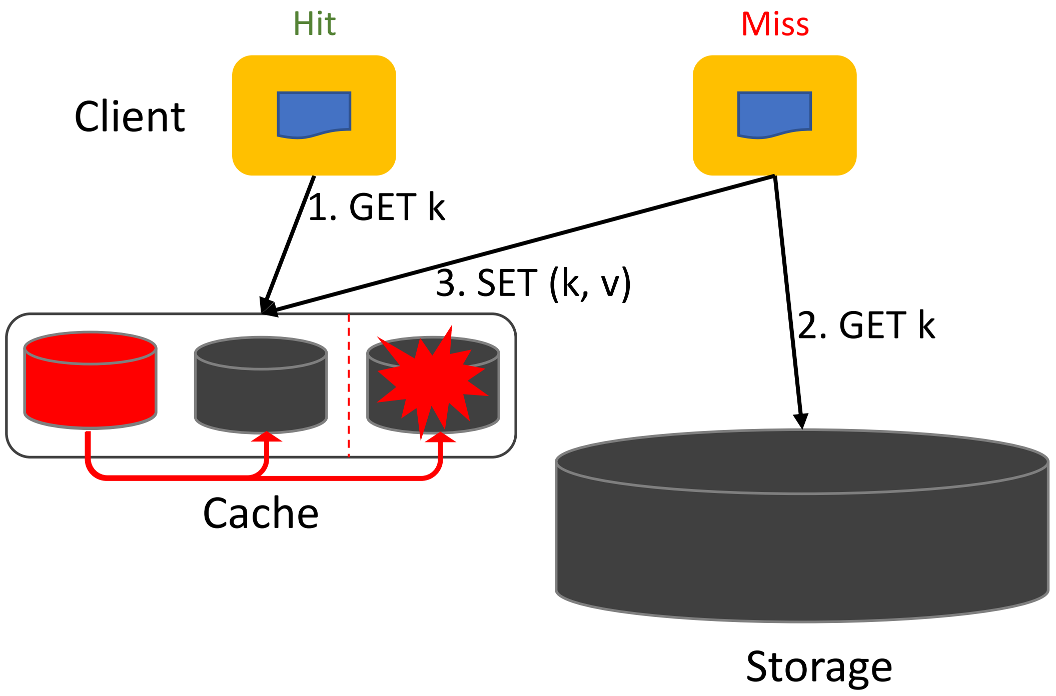
Distributed Caches are Dynamic

Example: Look-aside caches in web services

Various dynamic operations
- Cache partitioning, re-partitioning, load balancing
- Failure recovery

Cache server starts out ‘cold’ (or partly cold)

Warmup: Getting cache from ‘cold’ to ‘hot’



Understanding Cache Warmup

Imagine if you're operating some cache servers…

Caches are only useful when they contain useful data

- Cache misses = end-users get their data slower
- Cache misses = expensive load on storage servers

Cache has warmed up when it provides "sufficient" performance

- Considered by few recent works, but never carefully quantified
- <u>Implicit</u> in many designs (e.g. rate of cache repartitioning)

Challenging to define and calculate

- Warmup is a dynamic process
- Static metrics (Hit Ratio) are insufficient

Cache Dynamics

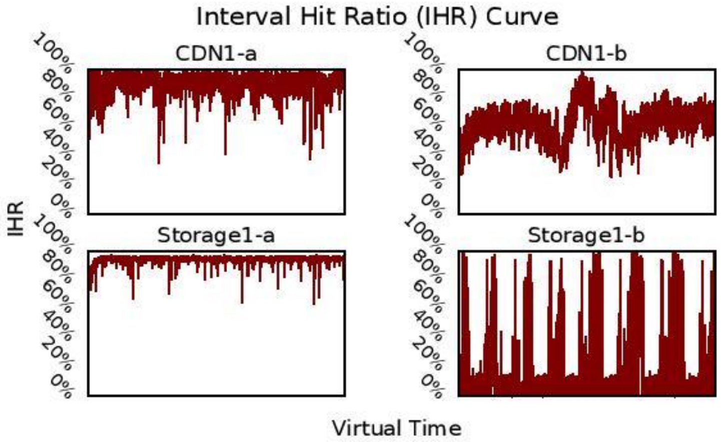
Cache performance depends fundamentally on workload dynamics

We capture cache dynamics through the *Interval Hit Ratio*

- Effectively a sliding window over hit rate.
- **Example:** LRU, cache size = 3

IHR = 0/3			IHR = 3/3			IHR = 1/3			IHR = 1/3			IHR = 3/3		
		C	C	C	C	C	C	C	C	C	C	C	C	C
	B	B	B	B	B	B	E	E	E	B	B	B	B	B
A	A	A	A	A	A	D	D	D	A	A	A	A	A	A
A	B	C	A	B	C	D	E	C	A	B	C	A	B	C

HR = 8/15



Defining Warmup

Natural definition: ‘converge to original’

Assume the operation started from beginning

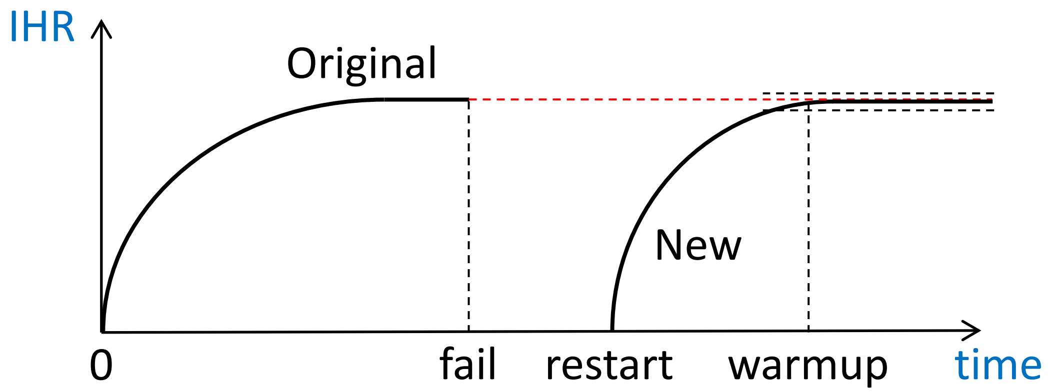
Beats the alternatives:

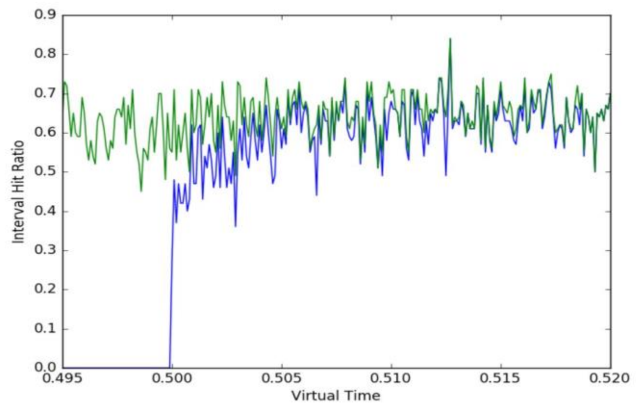
Arbitrary *Hit Ratio* threshold

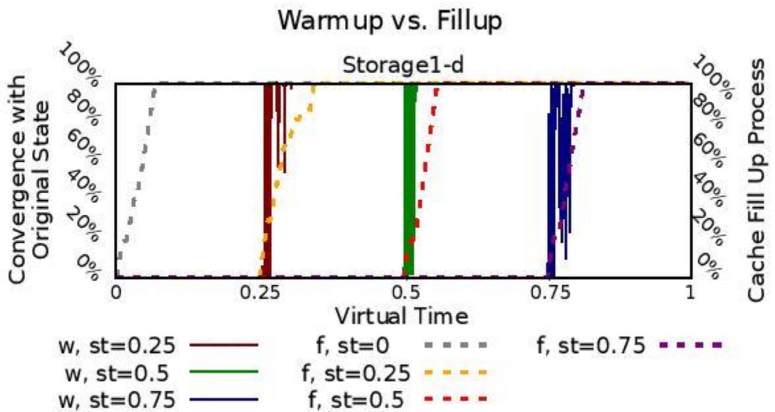
Arbitrary *Time* threshold

<u>Result:</u> Warmup is faster than fillup

- 16.6%-39.1%







Defining Warmup Time

For cache size s and tolerance level ϵ, a cache that recovers at time st is considered warmed up at time t if for any end time $et > t$, we have:

$$|IHR(0, et, s) - IHR(st, et, s)| < \epsilon.$$

Computing warmup time = offline analysis on IHR results

- Requires future knowledge of IHRs

How can we estimate warmup time in practice?

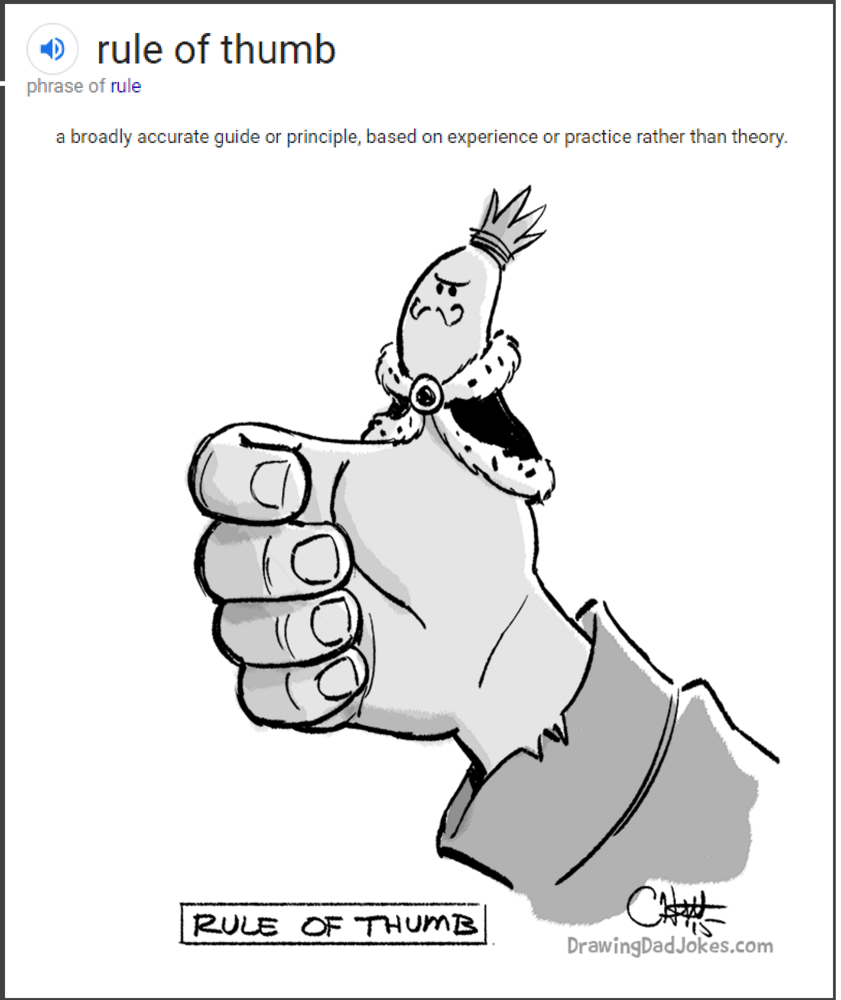
Solution: Rule of Thumb

Practical estimation of blackbox metrics

Goal: derive a rule of thumb formula for warmup time

- Make it **simple**
- Make it **accurate**
- Make it **general**

Estimates should fully consider cache dynamics



Deriving a Rule of Thumb

Compute offline warmup time as defined

Using spatially sampled workloads for efficiency

Relax the dynamic factors

Using maximum warmup time over all possible restart/recovery times

Approximate static factors

Cache size and tolerance level

Apply (log)-linear regression for warmup time and factors, discover relationships

Result: $\textbf{warmup-time}(\text{size}, \varepsilon) \propto \text{size}^{p_s} \cdot e^{-p_e \varepsilon}$

Extension: enlarging cache size, *e.g.* for cache partitioning (see paper)

Evaluating the rule

We used multiple types of workloads

Simplicity: ✓

Accuracy: R^2 likelihood test score

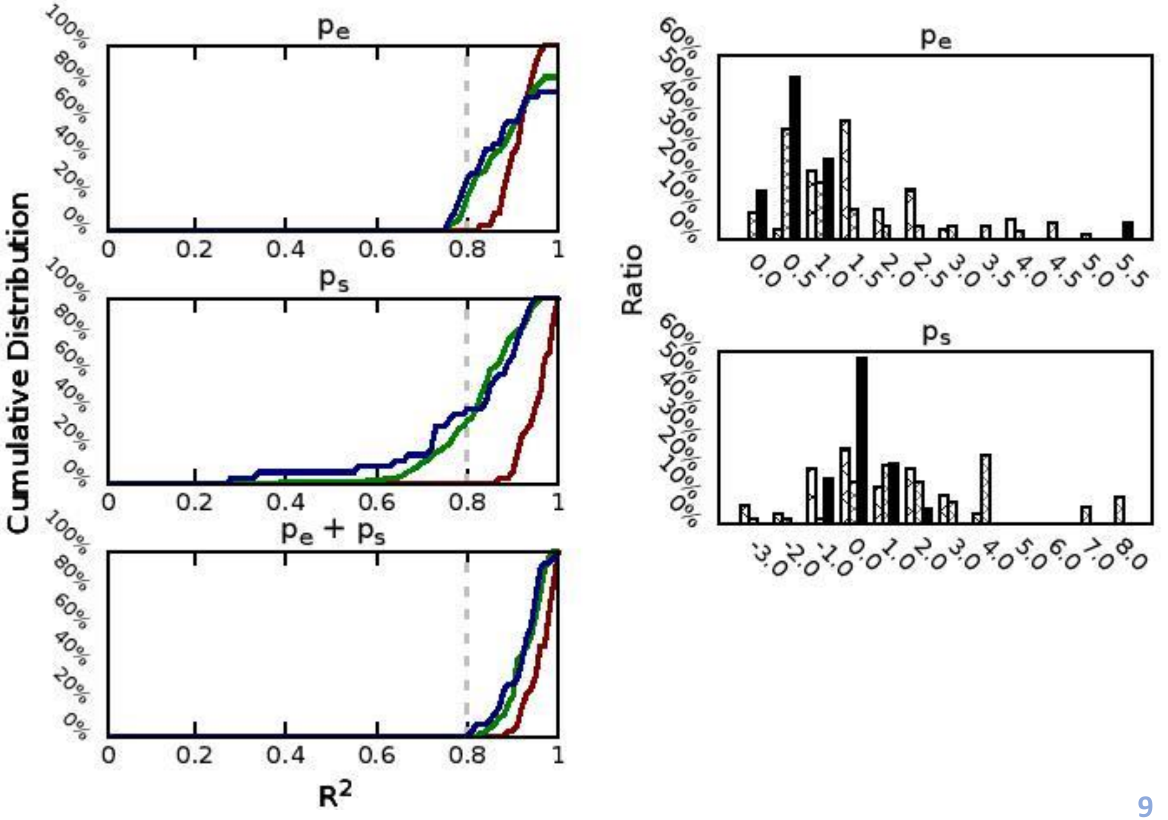
80% as threshold of a significance fit

More accurate with combined params

Generality: parameter range

Concentrate within each workload group

$$\textbf{warmup-time}(\text{size}, \varepsilon) = \boldsymbol{C} * \text{size}^{p_s} \cdot e^{-p_e \varepsilon}$$



Applying the Rule of Thumb

If your workload is similar to ours, use our formula.

Otherwise follow same process as how the formula was generated:

1. Get offline simulation results with workload(s) and cache parameters (s, ϵ)

offline-results = `SIMULATE(workloads, params)`

2. Get workload specific formula

warmup-time formula = `ANALYZE(offline-results, params)`

3. Use the formula for future operation decisions

Discussion

How to quantify the original cache state?

- Initial cache state (assumed to be stale or empty in the paper)
- When we reduce the cache size, what items are evicted?

Are our assumptions about cache dynamics justified in practice?

- Warmup time with different recovery/restart points
- Requires input from real systems

Conclusion

Warmup time matters in distributed caches, yet rarely studied

Use Interval Hit Ratio to capture cache dynamics

Nifty rule of thumb formula to use in your cache server operations

We plan to open source the warmup package!

Thank you!

Questions? geraldleizhang@gmail.com