Self-Optimizing Caches

Irfan Ahmad
CachePhysics
About

Irfan Ahmad
CachePhysics Cofounder
CloudPhysics Cofounder
VMware (Kernel, Resource Management), Transmeta, 40+ Patents
University of Waterloo
@virtualirfan

CachePhysics

Data Path Monitoring and Modeling Software
Real-time Predictive Modeling of Data Access Patterns
Increasing Performance & Cost Efficiency of Existing Caches
Powering Next-Generation Self-Learning Caches
Caches are Critical to Every Application

Yet…

Intelligent Cache Management is Non-Existent

- Is this performance good?
- Can performance be improved?
- How much Cache for App A vs B vs …?
- What happens if I add / remove DRAM?
- How much DRAM versus Flash?
- How to achieve 99%ile latency of X µs?
- What if I add / remove workloads?
- Is there cache thrashing / pollution?
- What if I change cache parameters?

Cache Performance

- Hit Ratio: 65%
- Cache Size: 128GB
Modeling Performance in Real-Time

- Learn performance model of applications and cache
- Predict the performance of workload as $f(\text{cache size}, \text{params})$

Cache Performance
- Hit Ratio: 65%
- Cache Size: 128GB

Latency (ms) vs. Cache Size (GB)

Lower is better
Understanding Cache Models

Models help decide useful increments of change.

![Graph showing cache size vs latency with x2, x4, x8 increments.](image-url)
Understanding Cache Models (2)

Lower is better

Often, most operating points are highly inefficient.
Sample Models From Production Workloads

![Graph showing miss ratio vs cache size for different models.](image)
Understanding Model-based Adaptation

Single Workload. Prediction of performance under different policies.
LIRS Adaptation Examples
2Q Adaptation Examples

![Graph showing miss ratio vs. cache size for 2Q - msr_web and 2Q - msr_src1, with lines representing Auto, Min Static, and Max Static.]
Cliff Removal: New Class of Acceleration

- Steer the curve?
  - Interpolate convex hull
  - Need Model (HPCA ’15)
- Shadow partitions $\alpha$, $\beta$
  - Steer different fractions of refs to each
  - Emulate cache sizes on convex hull via hashing
Cliff Reduction Results

69% of potential gain
Achieving Latency Targets

Client target 95th %ile latency is 7 ms

- Autoset cache partitions size to 16GB to guarantee avg latency SLOs

Latency Target (7 ms)

Cache Allocation (>16 GB)

* Throughput targets can be implemented similarly
Multi-Tier Sizing

* Can model network bandwidth as a function of cache misses from each tier.
Features

- Self-learn predictions for each app
- Alert, recommendations
- Recommendation/SLA API
- Capacity planning, what-ifs
Self-Optimizing Data Path

Monitoring

Auto-Select Policies (dynamic parameters)

Latency Reduction (Thrashing Remediation)

Results:
- Safely quantify impact of changes
- 50-150% efficiency improvements common
- Latency SLAs met
- Fewer production firefights
- Higher consolidation ratios
- Accurate Capacity Planning

Latency Guarantees

Accurate Tiering

Multi-Tenant Isolation

Miss Ratio

Cache Size (GB)

0.0 0.2 0.4 0.6 0.8

0 42 84 128 170

Lower is better

Miss Ratio

Cache Size (GB)

0.0 0.2 0.4 0.6 0.8

0 42 84 128 170

Lower is better

Tier 0 allocation for this client
Tier 1 allocation for this client
Remote Tier

Part. 0
Part. 1

Client 0
Client 1

2018 Storage Developer Conference. © CachePhysics, Inc. All Rights Reserved.
CachePhysics

irfan@cachephysics.com  650-417-8559  @virtualirfan