Self-Optimizing Caches

Irfan Ahmad

CachePhysics

About

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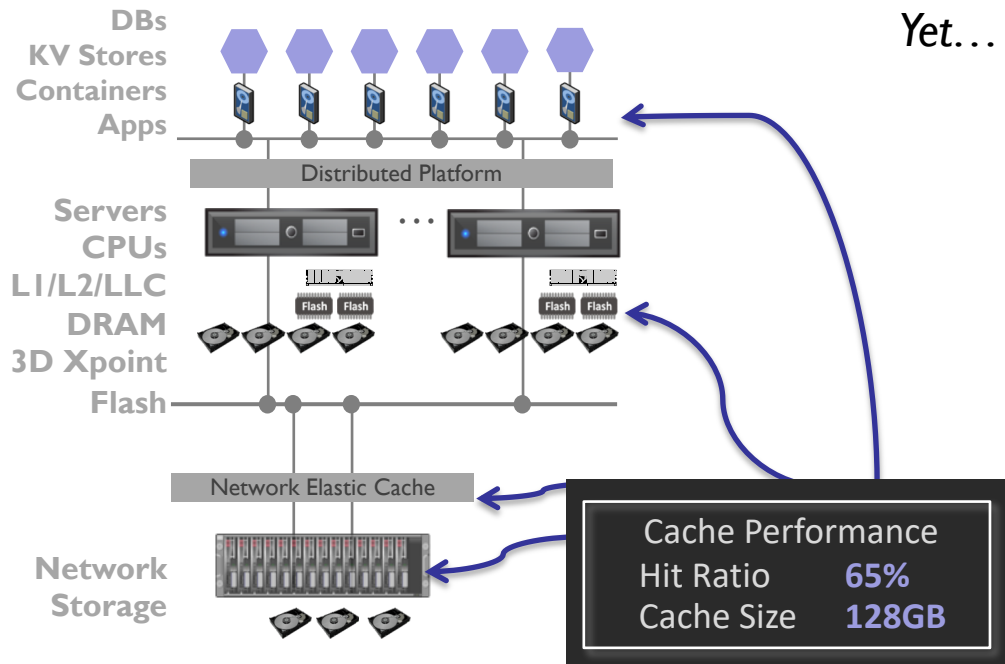@virtualirfan

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

Labels (left hierarchy):
- DBs
- KV Stores
- Containers
- Apps

Distributed Platform

- Servers
- CPUs
- L1/L2/LLC
- DRAM
- 3D Xpoint
- Flash

Flash   Flash   Flash   Flash

Network Elastic Cache

- Network
- Storage

Yet…

Cache Performance
Hit Ratio   **65%**
Cache Size   **128GB**

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?
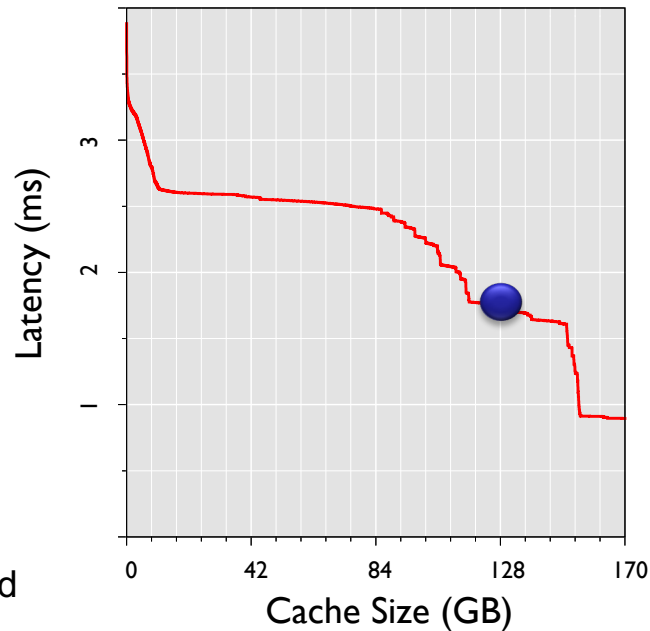- Is there cache thrashing / pollution?
- What if I change cache parameters?

SDC 18

Modeling Performance in Real-Time
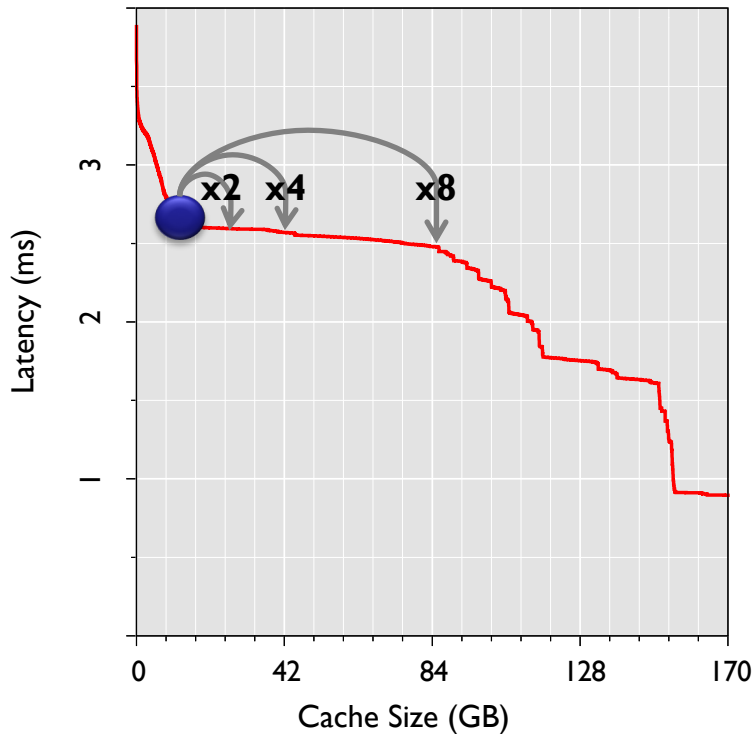
Lower is better

Cache Performance
Hit Ratio   **65%**
Cache Size   **128GB**

- Learn performance model of applications and cache
- Predict the performance of workload as *f(cache size, params)*

Latency (ms) vs Cache Size (GB)
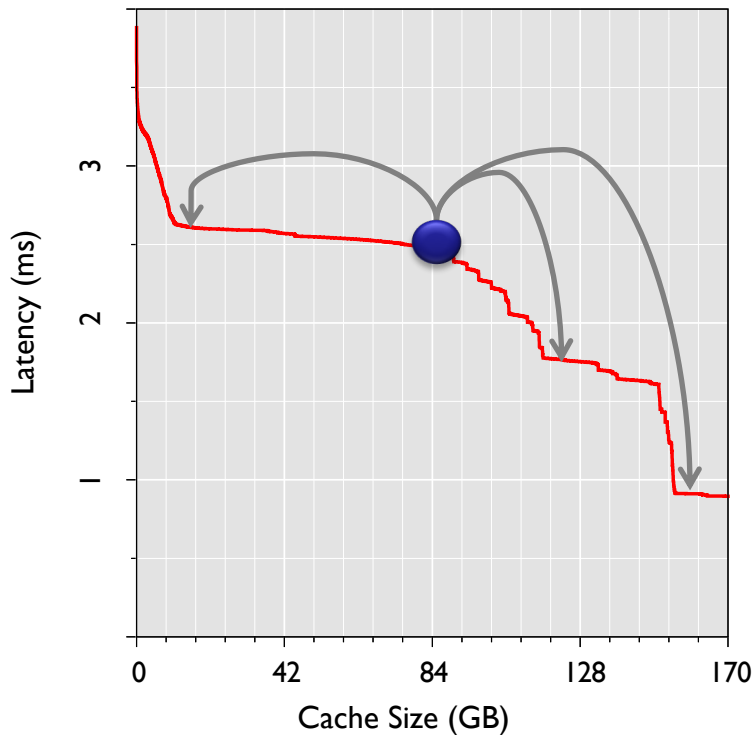
Understanding Cache Models

Lower is better

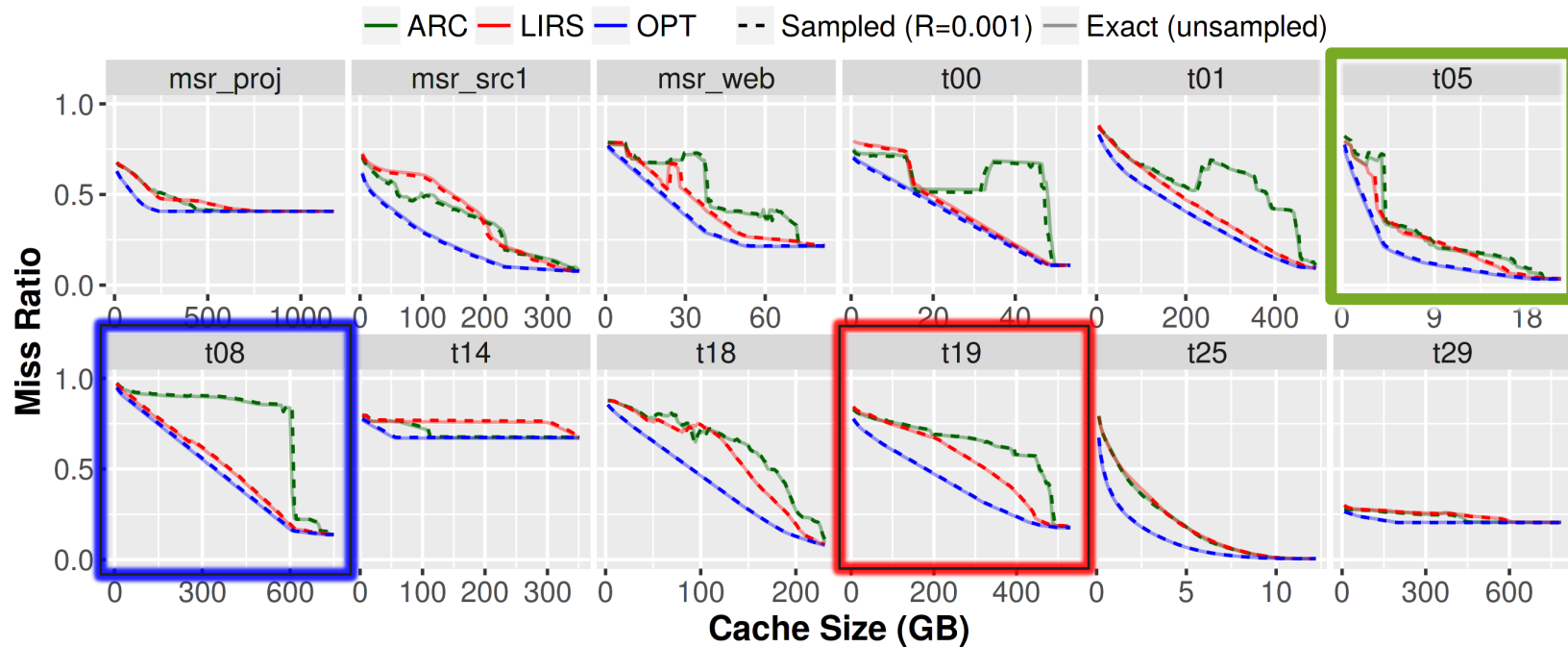

Models help decide useful increments of change.
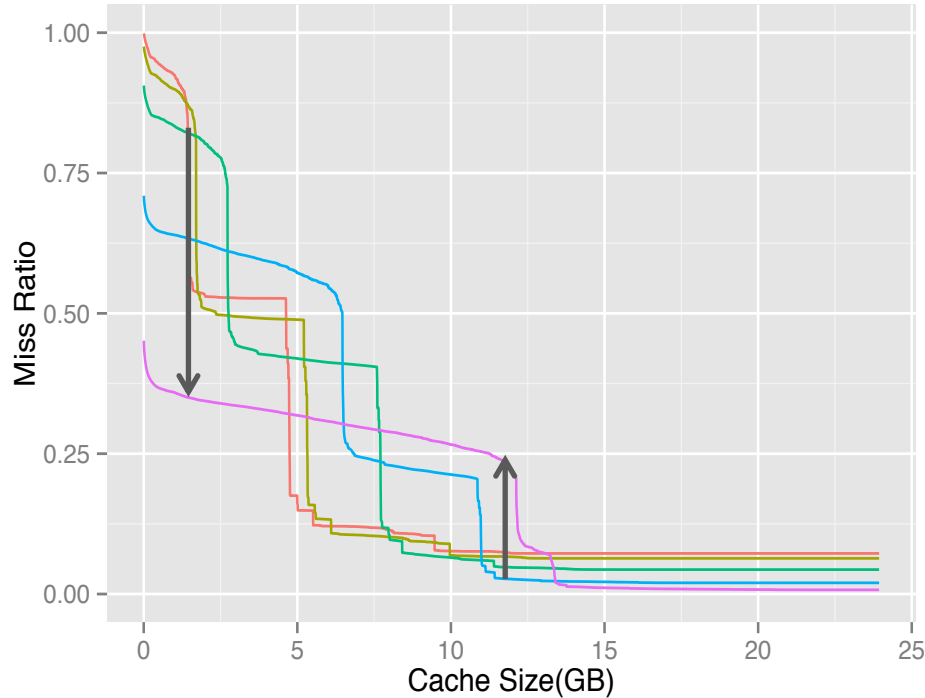
Understanding Cache Models (2)

Lower is better

Often, most operating points are highly inefficient.

Sample Models From Production Workloads

Legend: ARC — LIRS — OPT — Sampled (R=0.001) - - Exact (unsampled) —
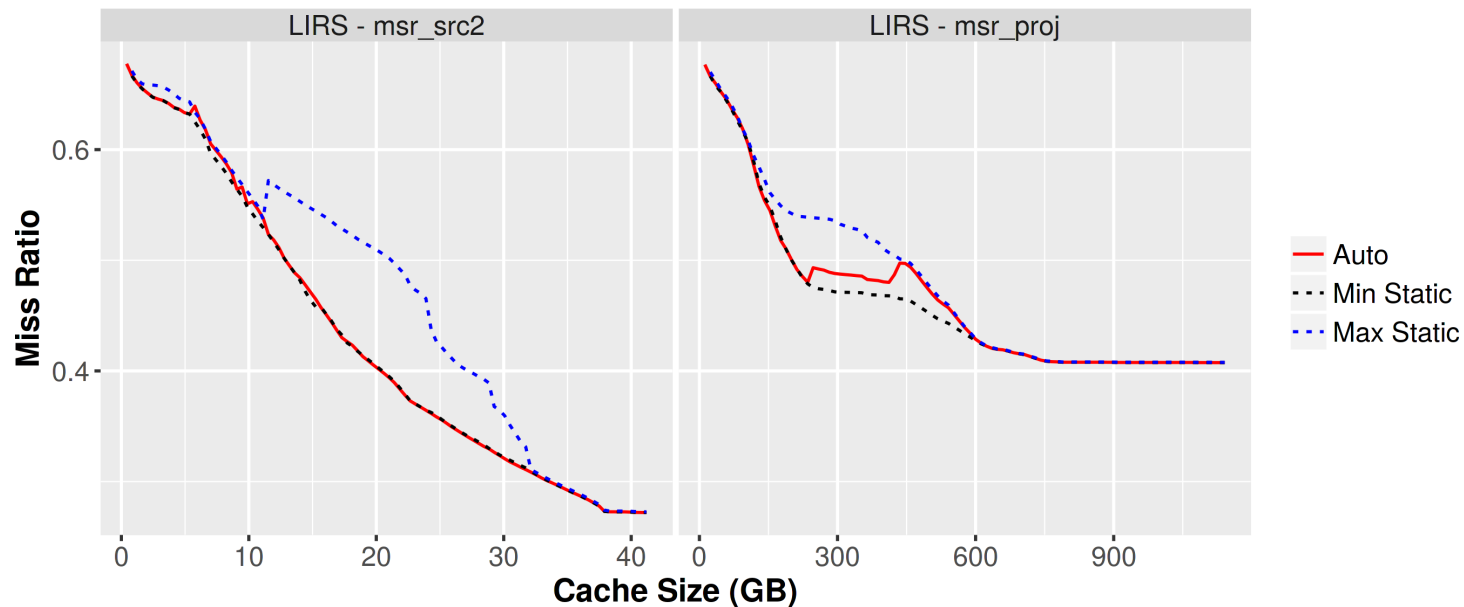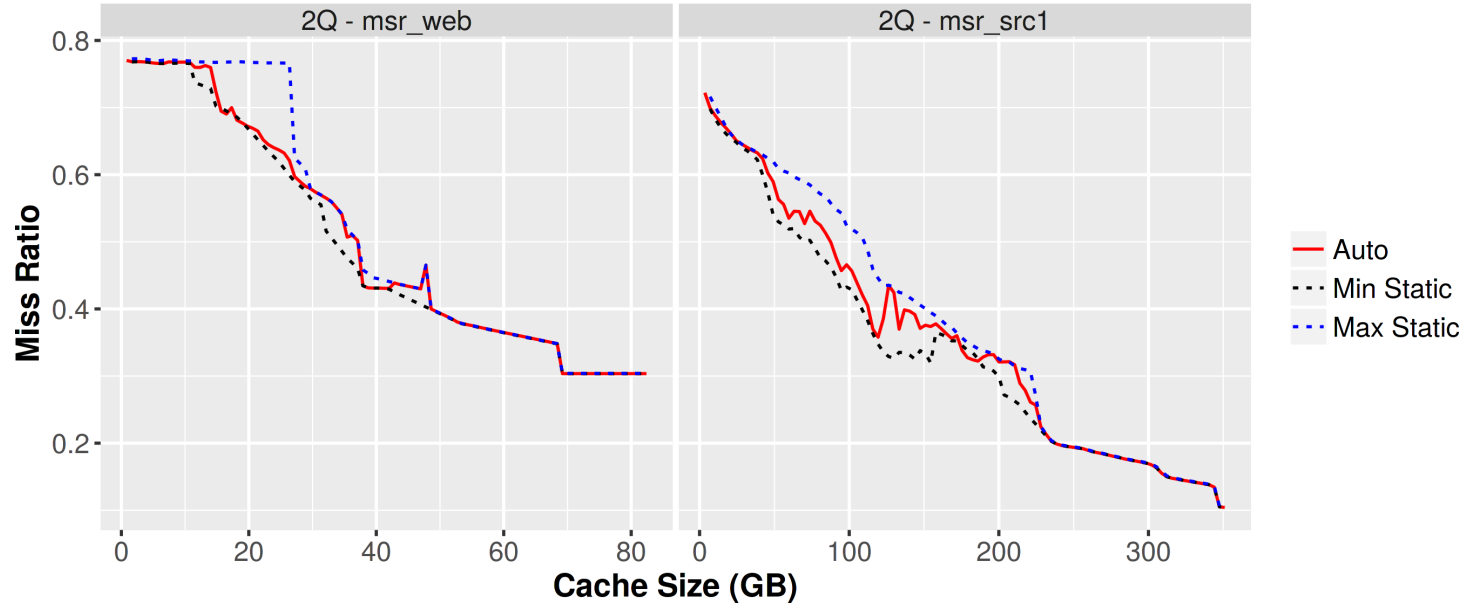
Understanding Model-based Adaptation

Single Workload.
Prediction of
performance
under different
policies.

LIRS Adaptation Examples
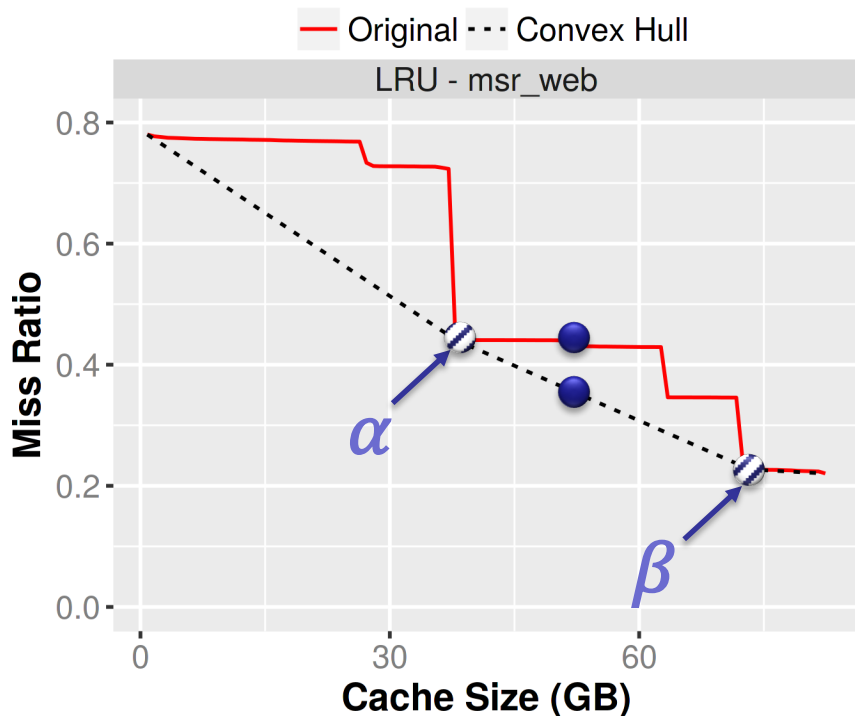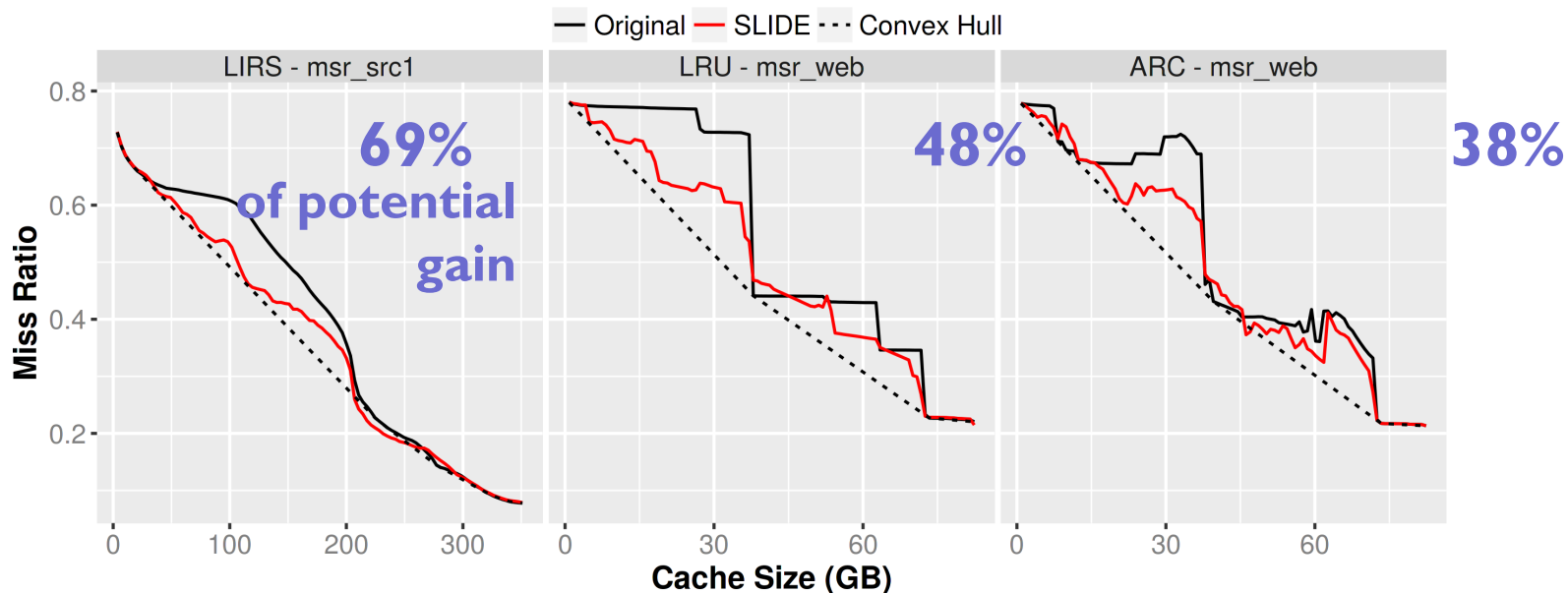
2Q Adaptation Examples

Cliff Removal: New Class of Acceleration

LRU - msr_web
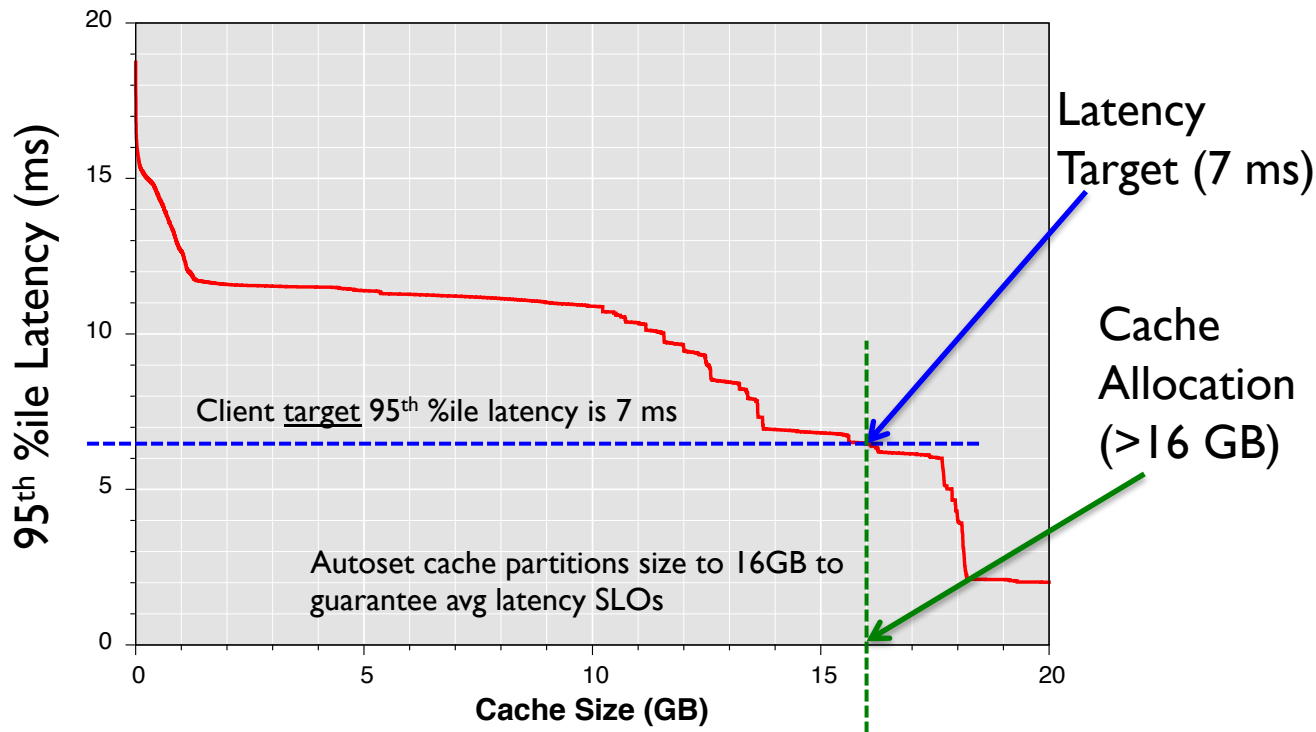
- Original --- Convex Hull

□ Steer the curve?
- □ Interpolate convex hull
- □ Need Model (HPCA '15)

□ Shadow partitions α, β
- □ Steer different fractions of refs to each
- □ Emulate cache sizes on convex hull via hashing

SDC 18

Cliff Reduction Results

Achieving Latency Targets

Latency Target (7 ms)

Cache Allocation (>16 GB)

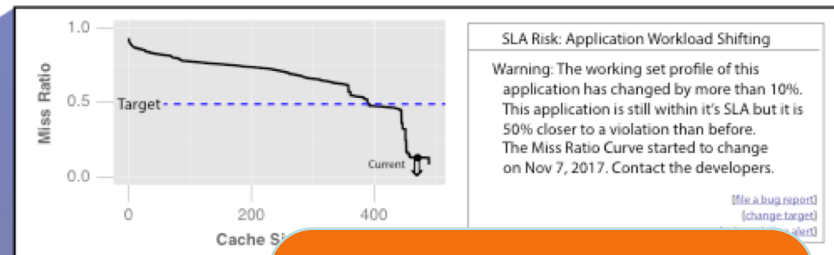Client $target$ 95th %ile latency is 7 ms

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

* Throughput targets can be implemented similarly

95th %ile Latency (ms)

Cache Size (GB)

Multi-Tier Sizing

Tier 0 (DRAM) allocation

Tier 1 (3D Xpoint)

Tier 2 (Local Flash)

Network Misses

Tier 3 (Remote Flash)
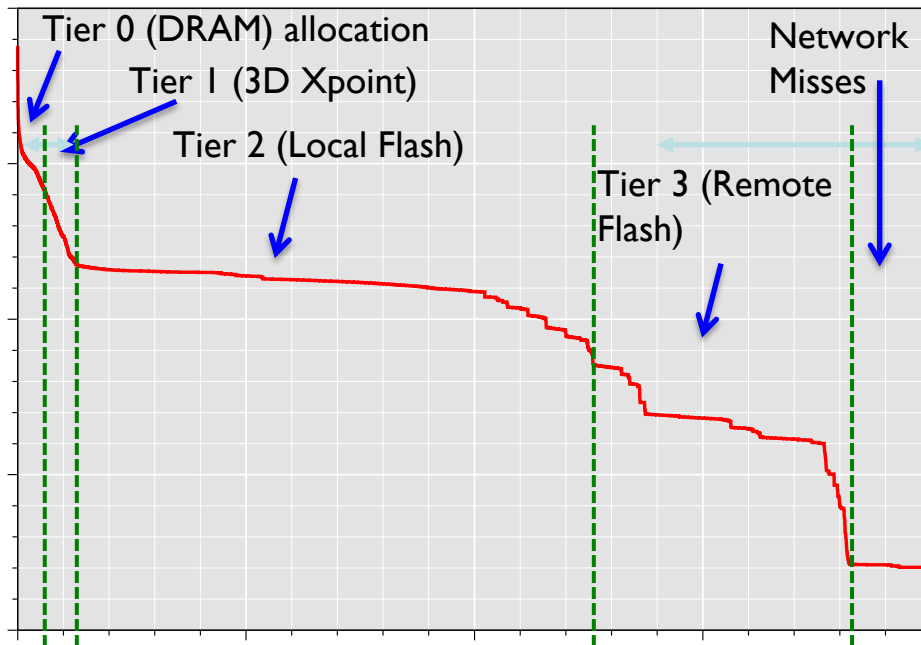
* Can model network bandwidth as a function of cache misses from each tier

SLA Violation

Recommendation: increase DRAM Memory
 This application has a performance SLA
 than can no longer be achieved due to a
 change in the workload.
 The current size of 325 GB should be
 increased to: 650 GB

[accept & schedule]
[change target]
[acknowledge alert]

SLA Risk: Application Workload Shifting

Warning: The working set profile of this
 application has changed by more than 10%.
 This application is still within it's SLA but it is
 50% closer to a violation than before.
 The Miss Ratio Curve started to change
 on Nov 7, 2017. Contact the developers.

[file a bug report]
[change target]

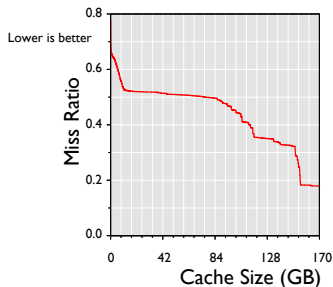**Features**
- Self-learn predictions for each app
- Alert, recommendations
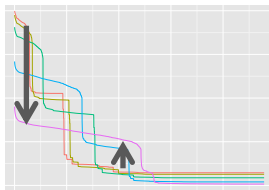- Recommendation/SLA API
- Capacity planning, what-ifs

Cost Saving Recommendation

Warning: This application is not benefitting
 from most of the DRAM allocated to it.
 Suggest reducing DRAM down from 350GB
 down to: 90 GB.
 The new size is within 5% of the performance
 but with 75% cheaper!

[accept & schedule]
[set target]
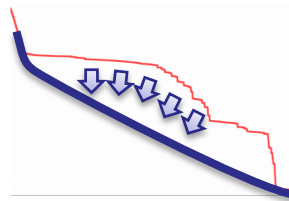[acknowledge alert]
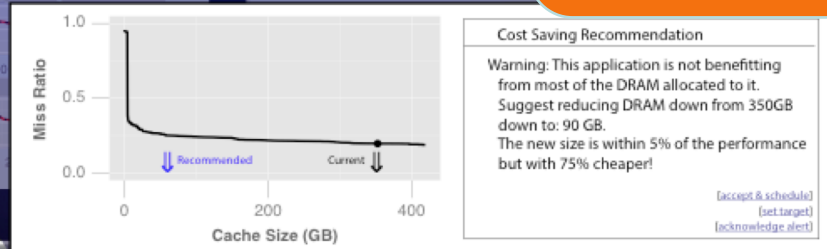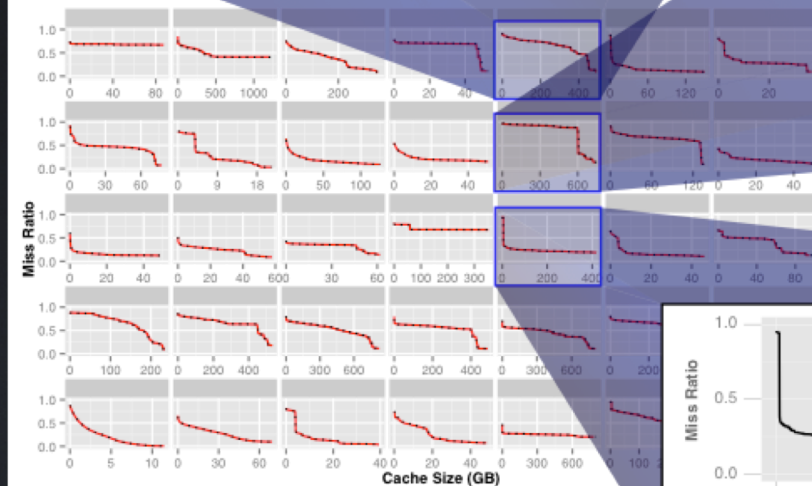
Self-Optimizing Data Path

Monitoring

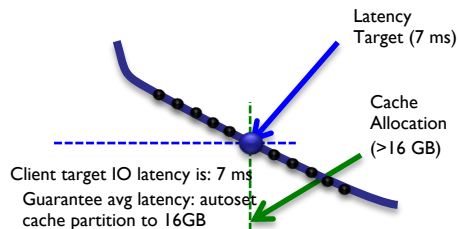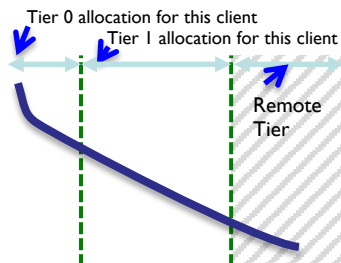### Auto-Select Policies
(dynamic parameters)
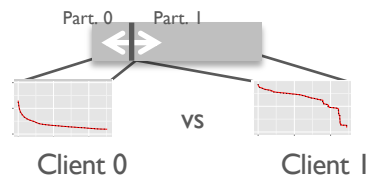

Latency Reduction
(Thrashing Remediation)

Latency Guarantees

Latency Target (7 ms)

Cache Allocation (>16 GB)

Client target IO latency is: 7 ms
Guarantee avg latency: autoset cache partition to 16GB

Accurate Tiering

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

Multi-Tenant Isolation

Part. 0   Part. 1

vs

Client 0   Client 1
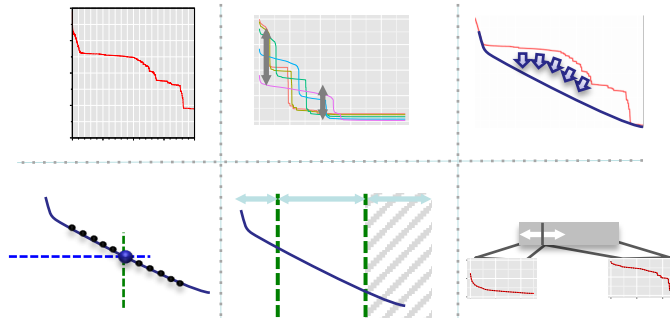
Results:

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

CachePhysics

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