CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture

Yanqi Zhou[1] Henry Hoffmann[2] David Wentzlaff[1]

1 PRINCETON UNIVERSITY
PRINCETON
School of Engineering and Applied Science

2 THE UNIVERSITY OF CHICAGO
Department of Computer Science

Web Services Have Latency Requirements

2 MINUTE READ

How One Second Could Cost Amazon $1.6 Billion In Sales

Econsultancy Subscriber Research & Data Blog Events Training

It's Official – 'Web Stress' is Bad for Business

A Limited Pallet of Choices in EC2

Region: US West (Oregon)

	Linux/UNIX Usage	Windows Usage
Standard On-Demand Instanc		
Small (Default)	065 per	r Hour
Medium	0 p	Hour
Large	0 p	Hour
Extra Large	320 per	er Hour
Micro On-Demand Instances		
Micro	$0.020 per Hour	$0.020 per Hour
High-Memory On-Demand Instances		
Extra Large		$0.570 per Hour
Double Extra Large		$1.140 per Hour
Quadruple Extra Large		$2.280 per Hour
High-CPU On-Demand Ins		
Medium		$0.285 per Hour
Extra Large		$1.140 per Hour
Cluster Compute Instance		
Eight Extra Large	$2.400 per Hour	$2.970 per Hour

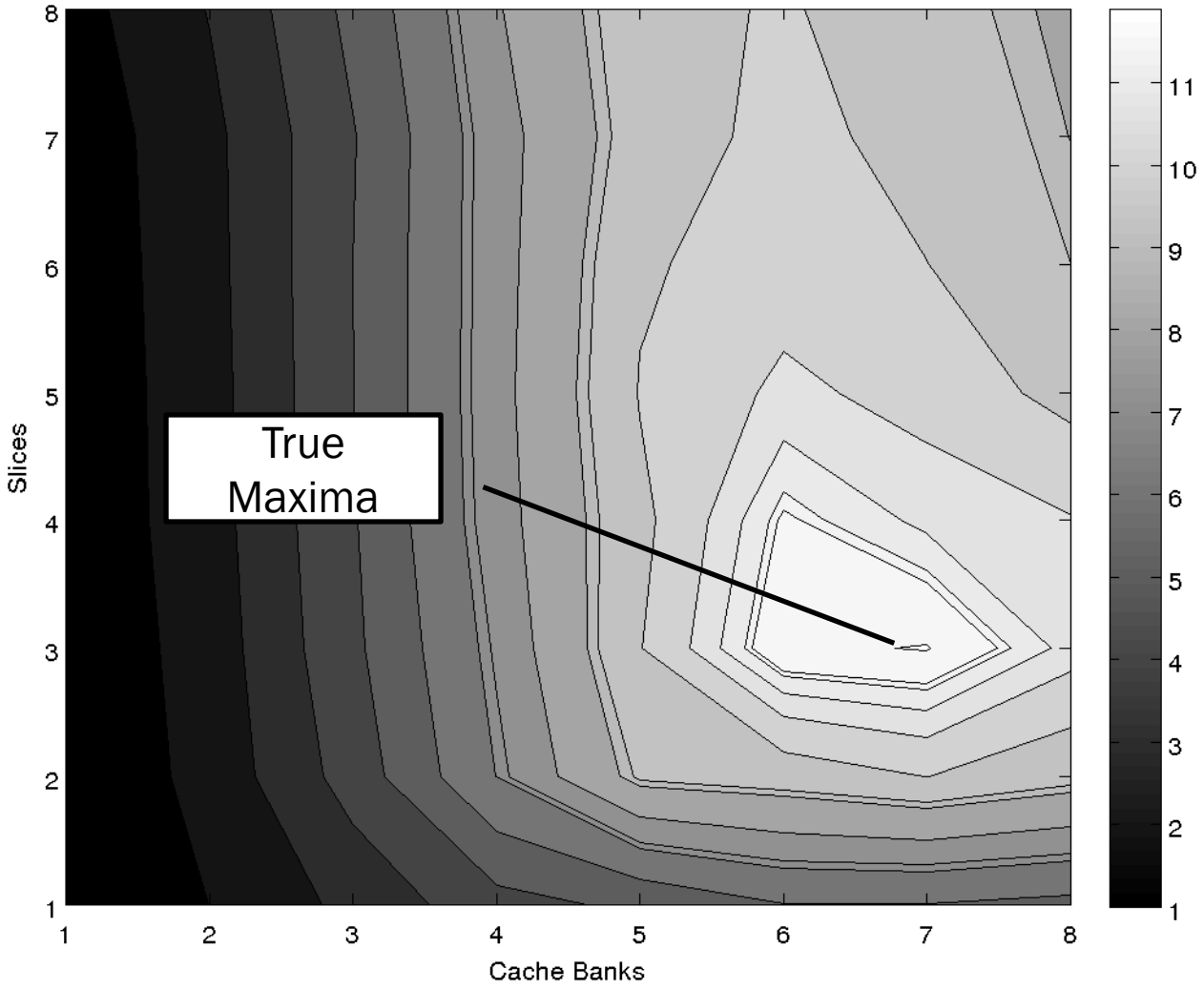
I need more performance

I need to reduce costs

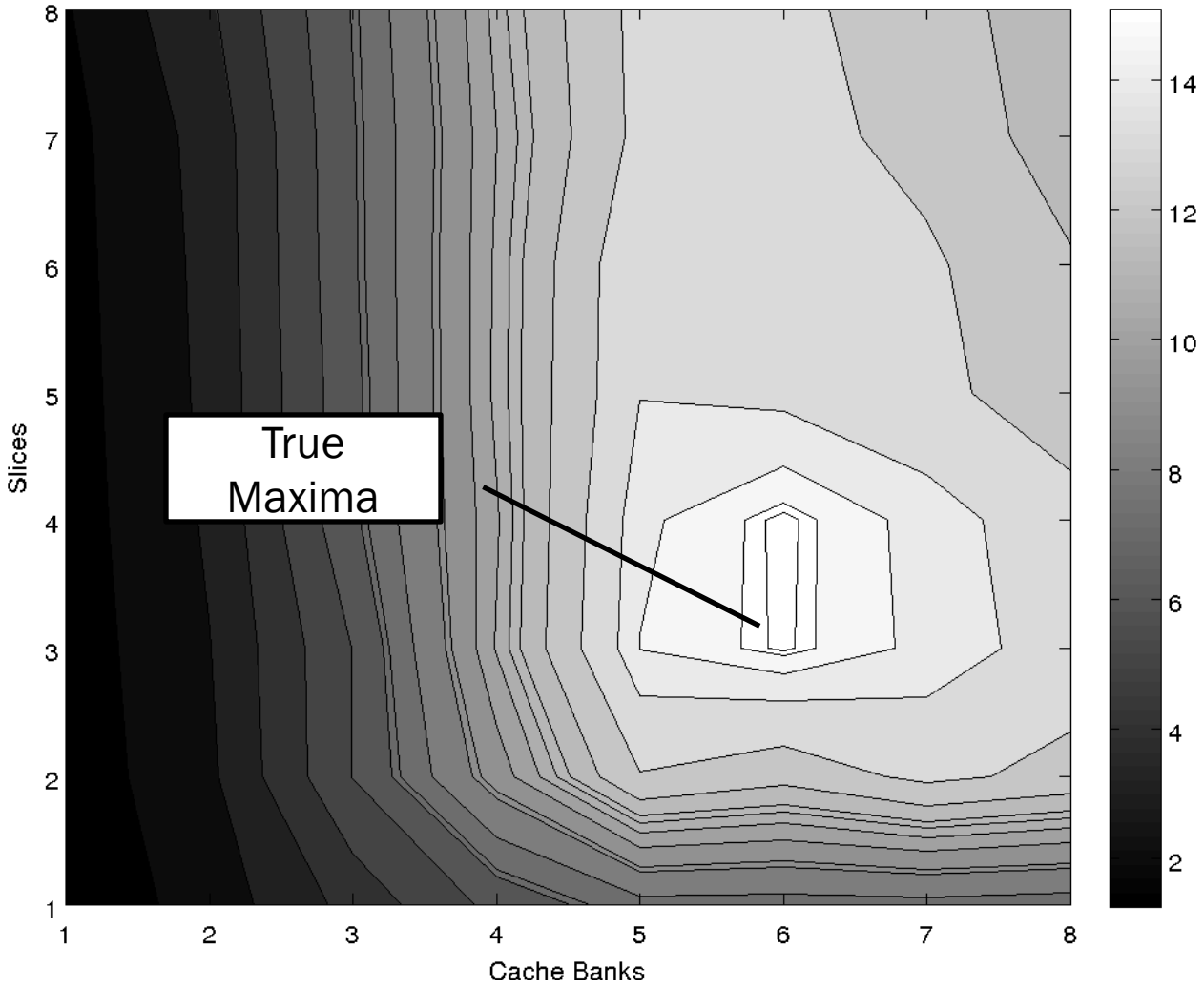
Image Credit: Amazon

Features	Distributed ILP	TRIPS, CLP	Core Fusion	WiDGET	Conjoined Cores	Clustered	big. LITTLE	Sharing Arch.
Scale up & down	✔	✔	✗	✔	✗	✗	✗	✔
Distributed/switched	✔	✔	✗	✗	✗	✗	✗	✔
Symmetric	✔	✔	✔	✔	✔	✔	✗	✔
Partition L2	✔	✔	✗	✗	✗	✗	✗	✔
Dynamic OoO	✗	✗	✔	✗	✔	✔	✔ ✗	✔
ISA Compatible	✔	✗	✔	✔	✔	✔	✔	✔
Multiple Metrics	✗	✔	✗	✗	✗	✗	✗	✔

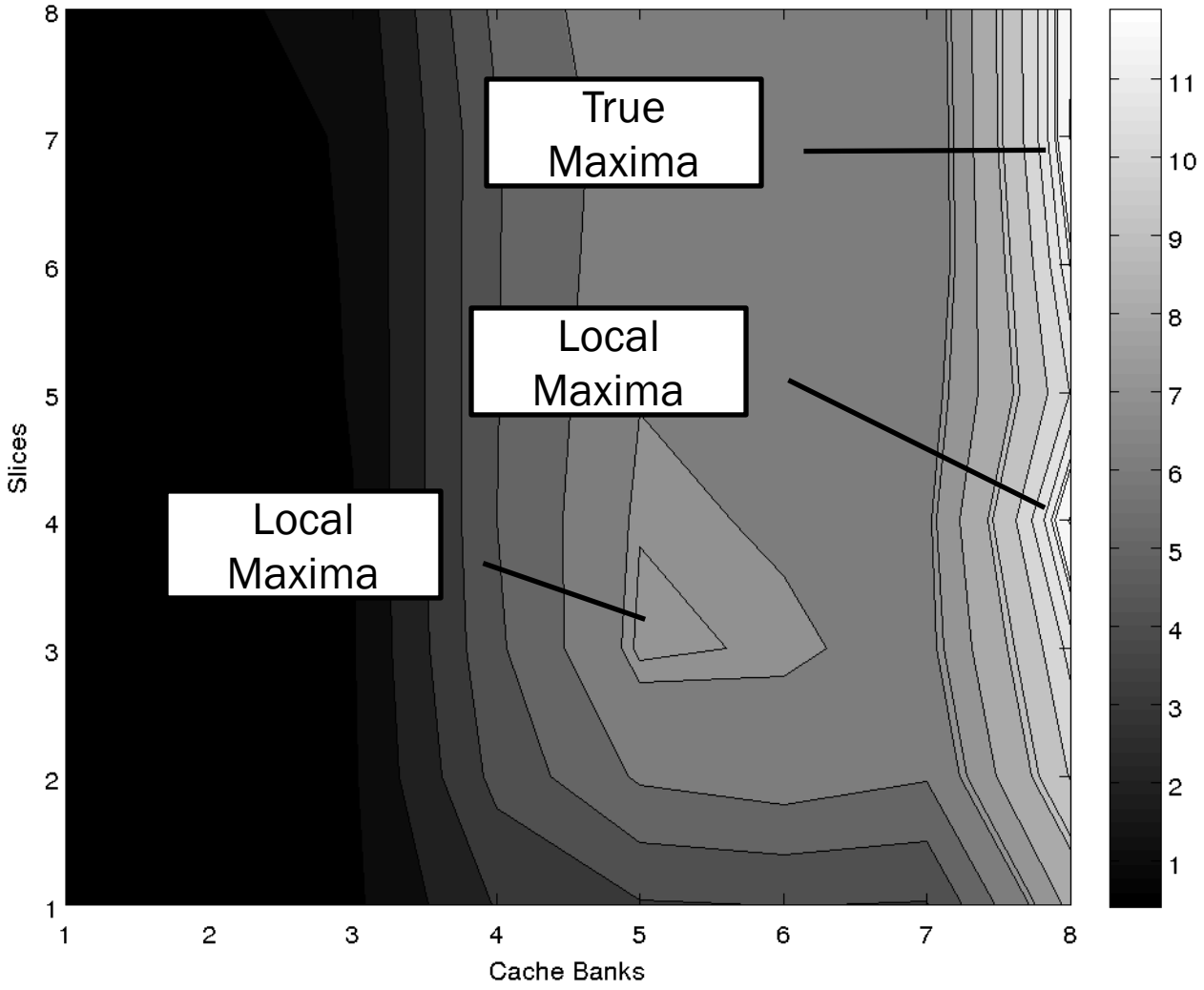
Managing Fine-grain Configurability is Hard



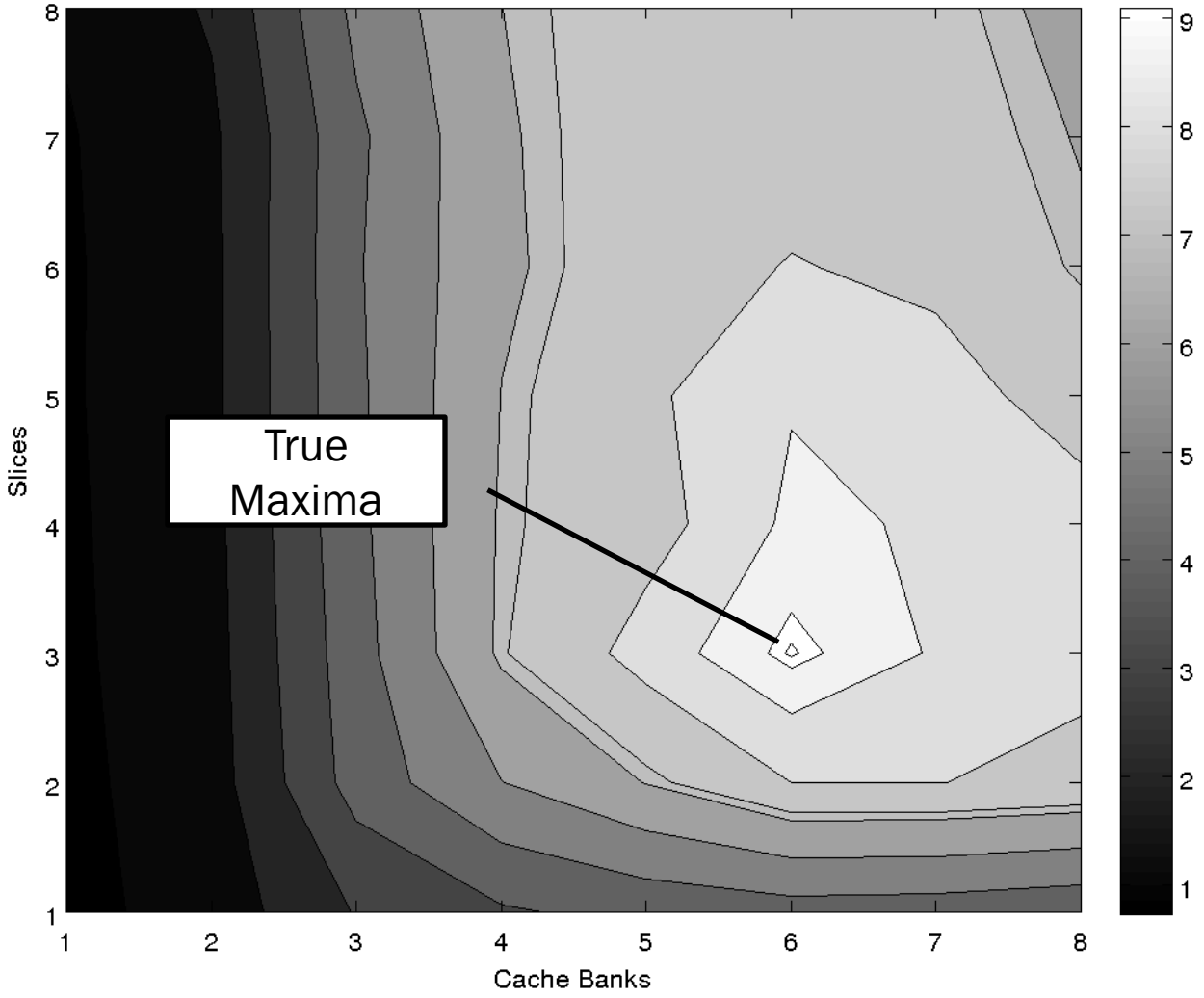
Managing Fine-grain Configurability is Hard



Managing Fine-grain Configurability is Hard

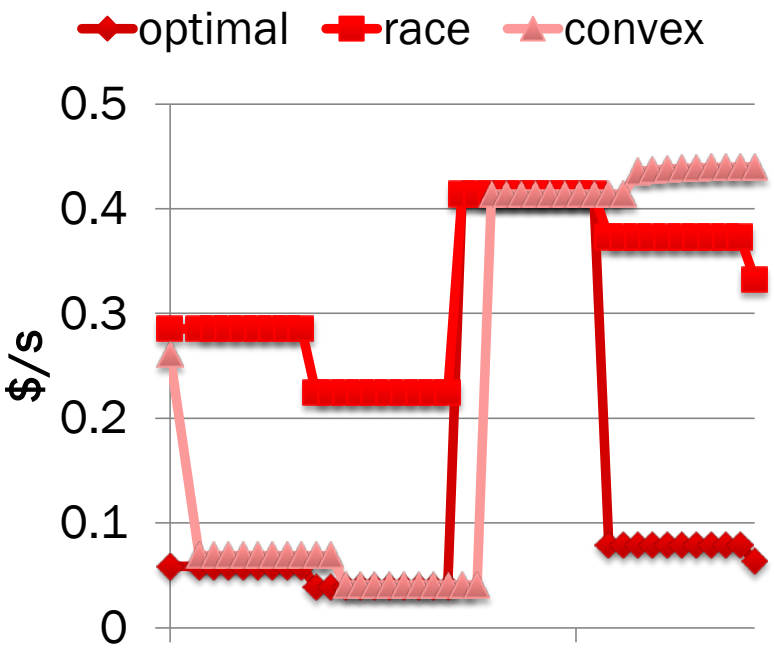


Managing Fine-grain Configurability is Hard

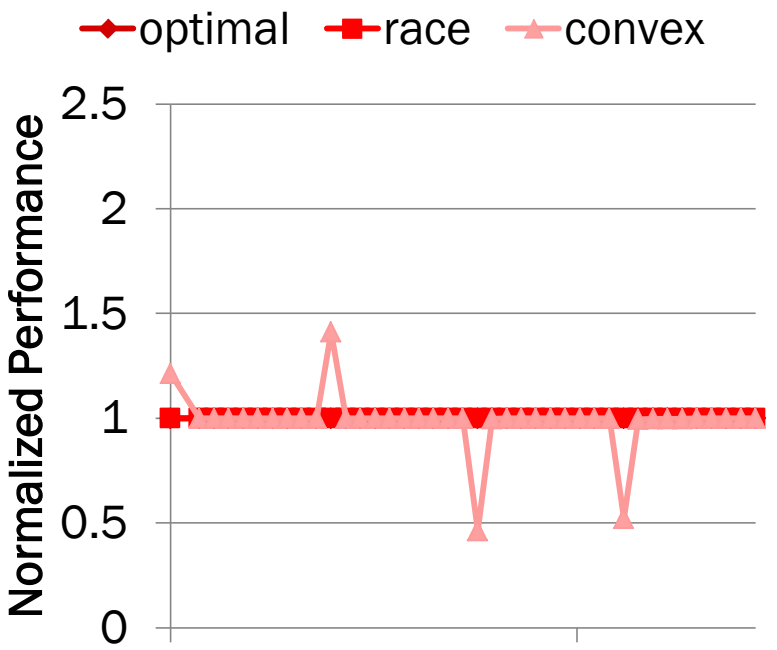


Benefit of Managing Fine-grain Configurability

Cost



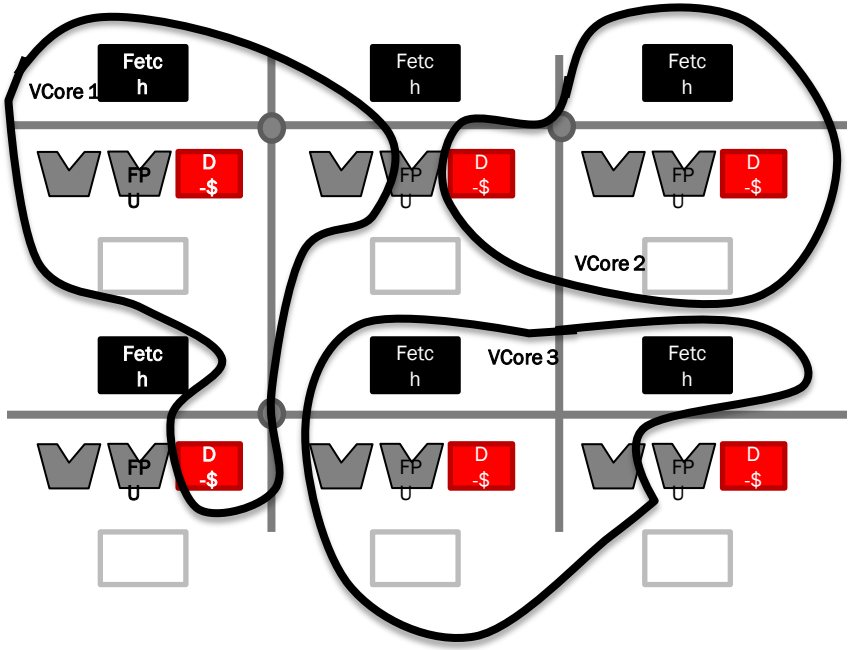
Performance

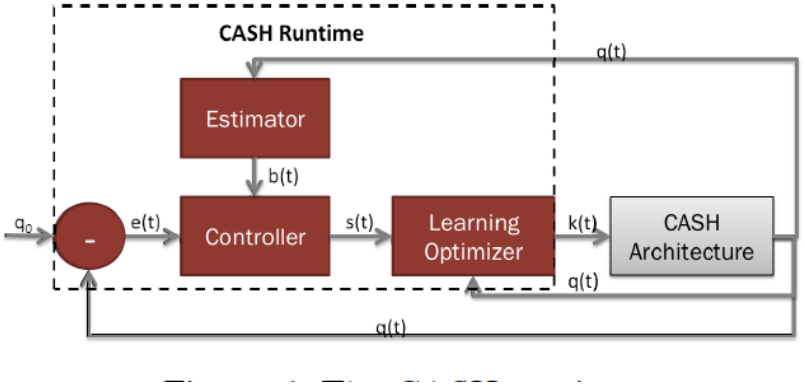


Observations:

- Fine-grain configurability is extremely helpful given current constraints
- Such configurability *should* produce non-convex optimization spaces
- Much research needs to be done on managing these non-convex spaces

CASH: Architecture and Runtime for Managing Fine-grain Configurability



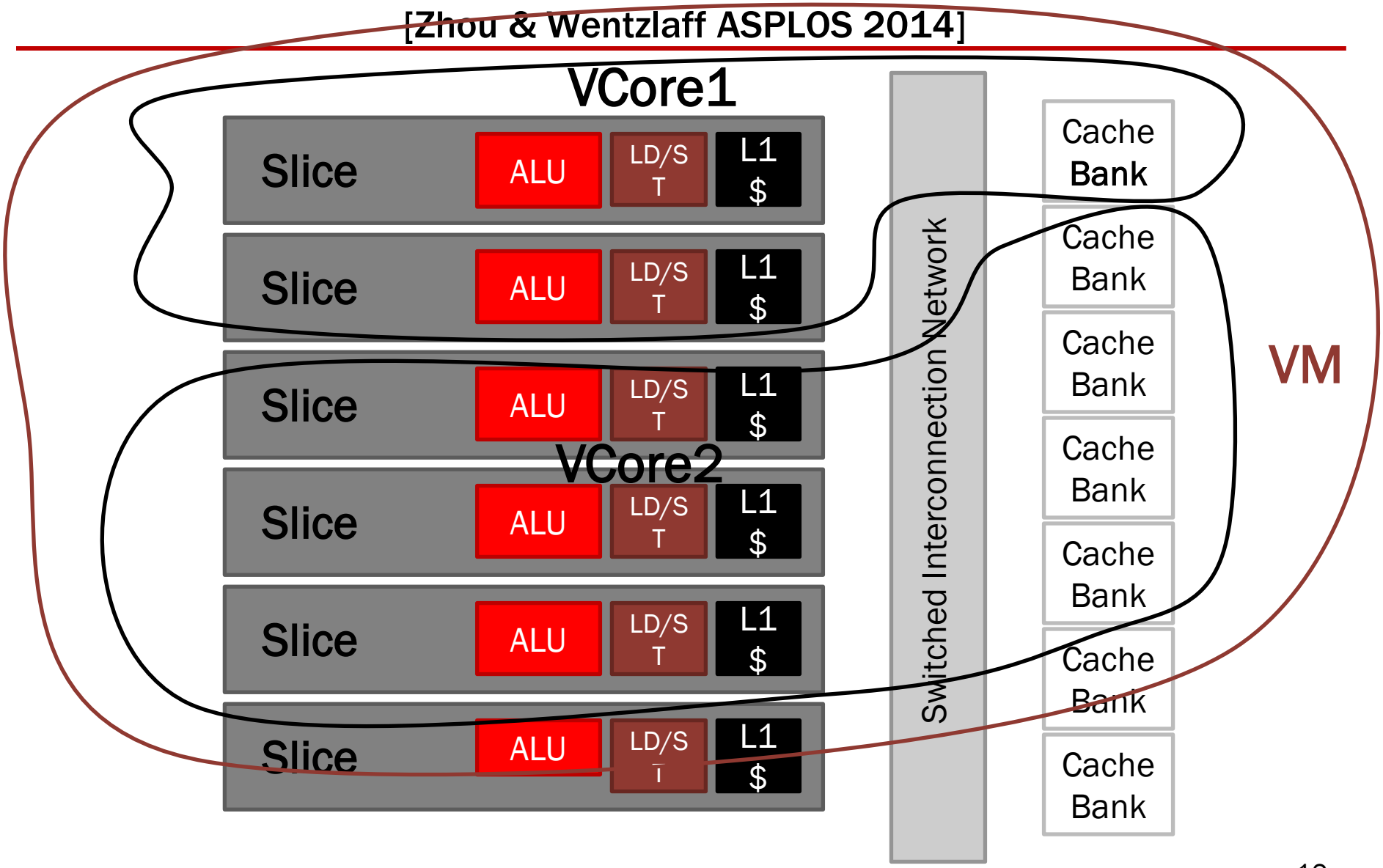


Solution:

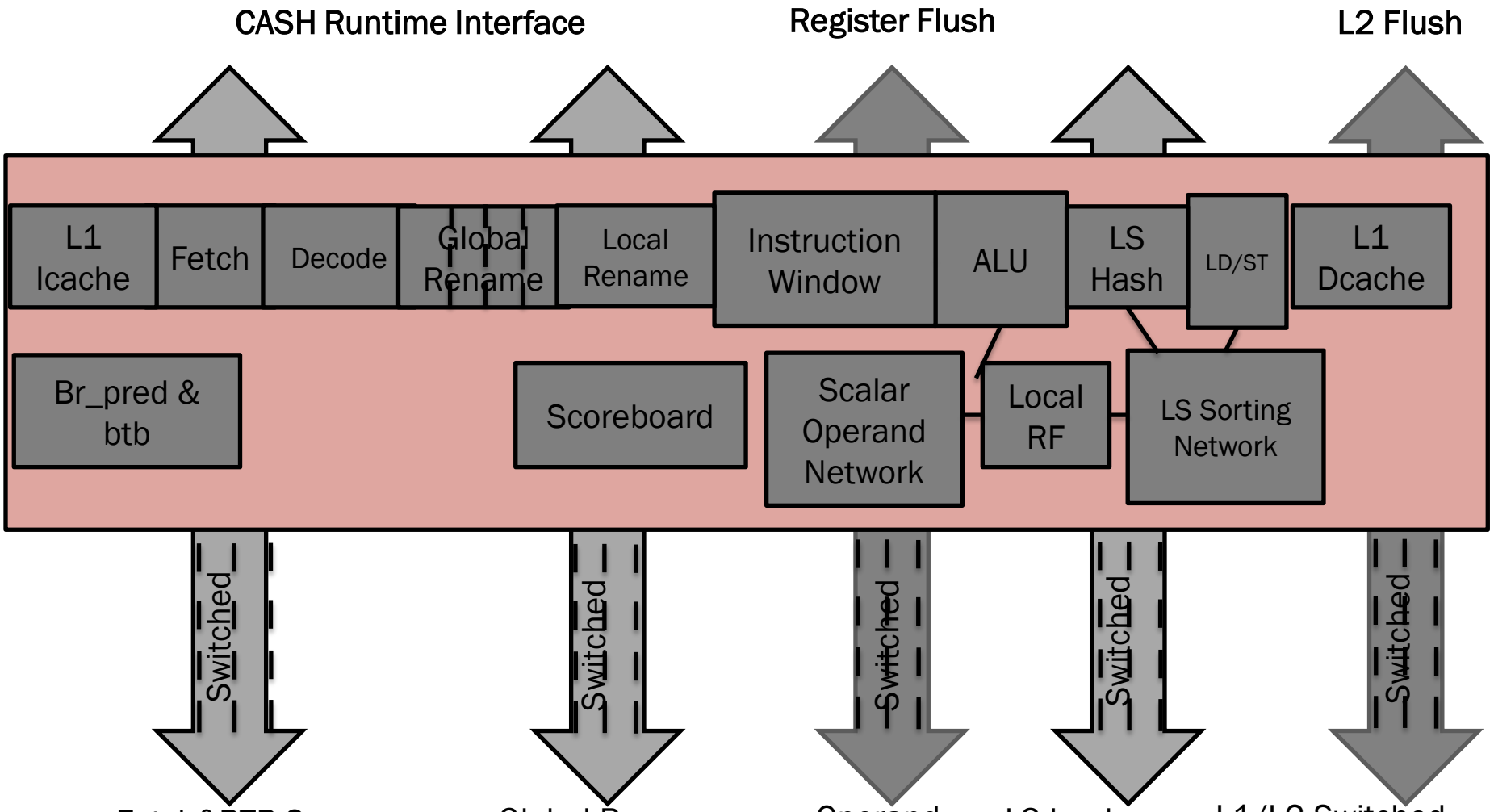
- Statically homogeneous, dynamically heterogeneous architecture
- Lightweight runtime system manages configurability online

Architecture: Sharing Architecture

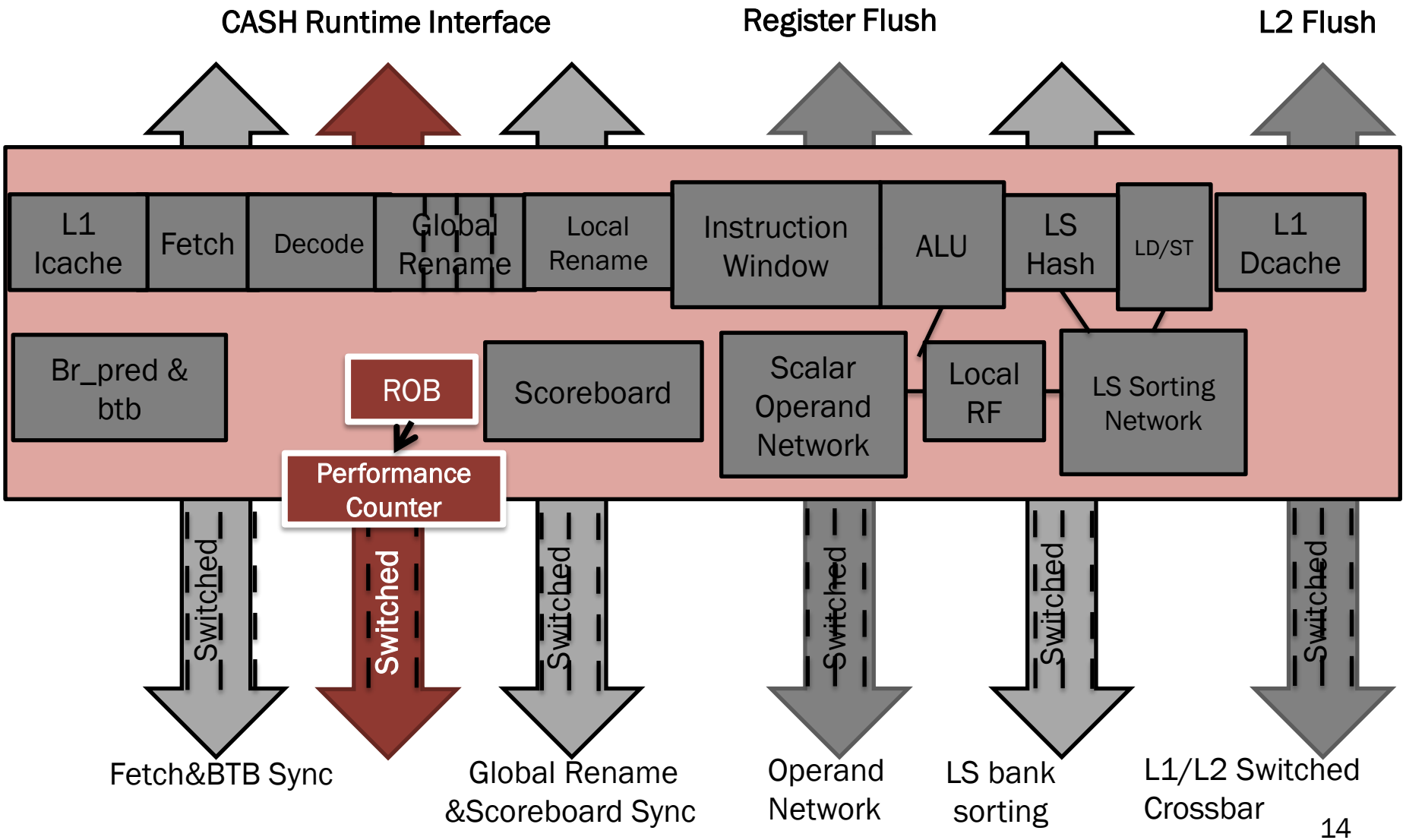
[Zhou & Wentzlaff ASPLOS 2014]



Slice Details



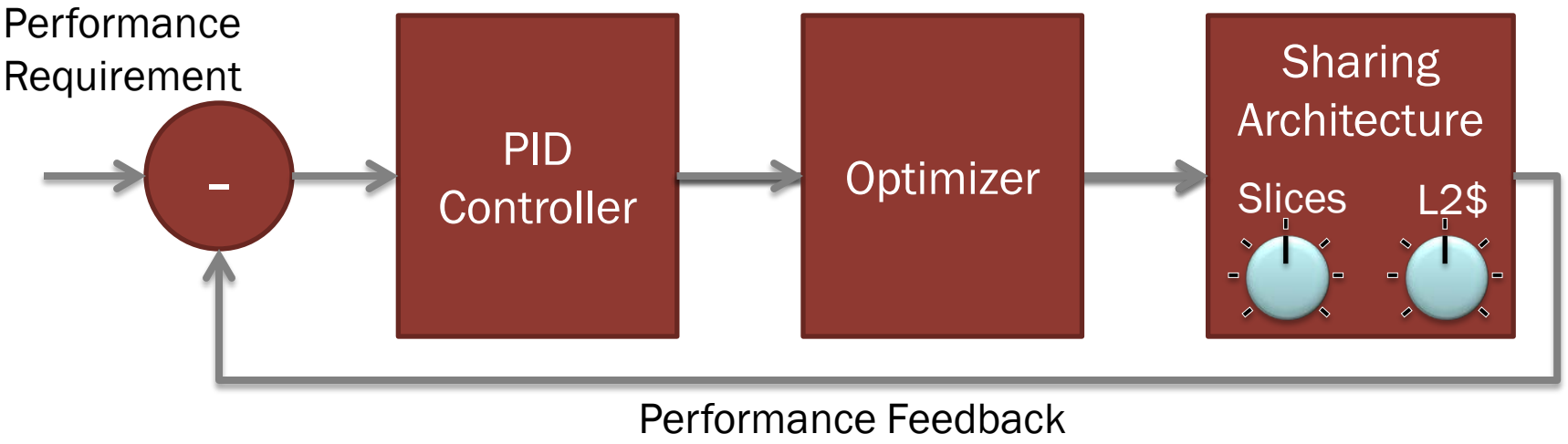
Aggregating Performance Data Across Slices



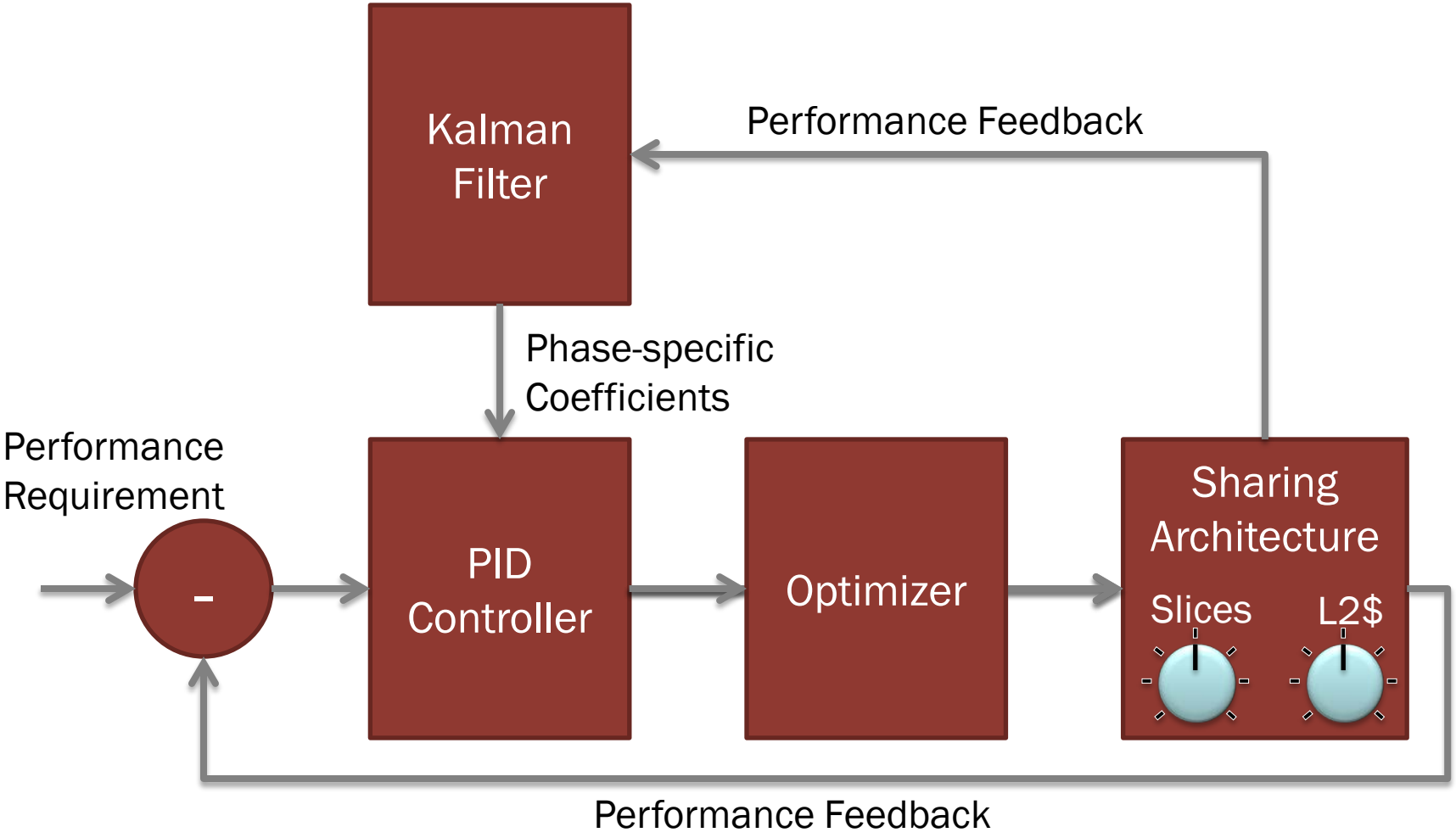
Controlling Application Performance

All problems in computer science can be solved by another level of indirection…
-- David Wheeler

Corollary: All optimization problems can be solved by another level of feedback…

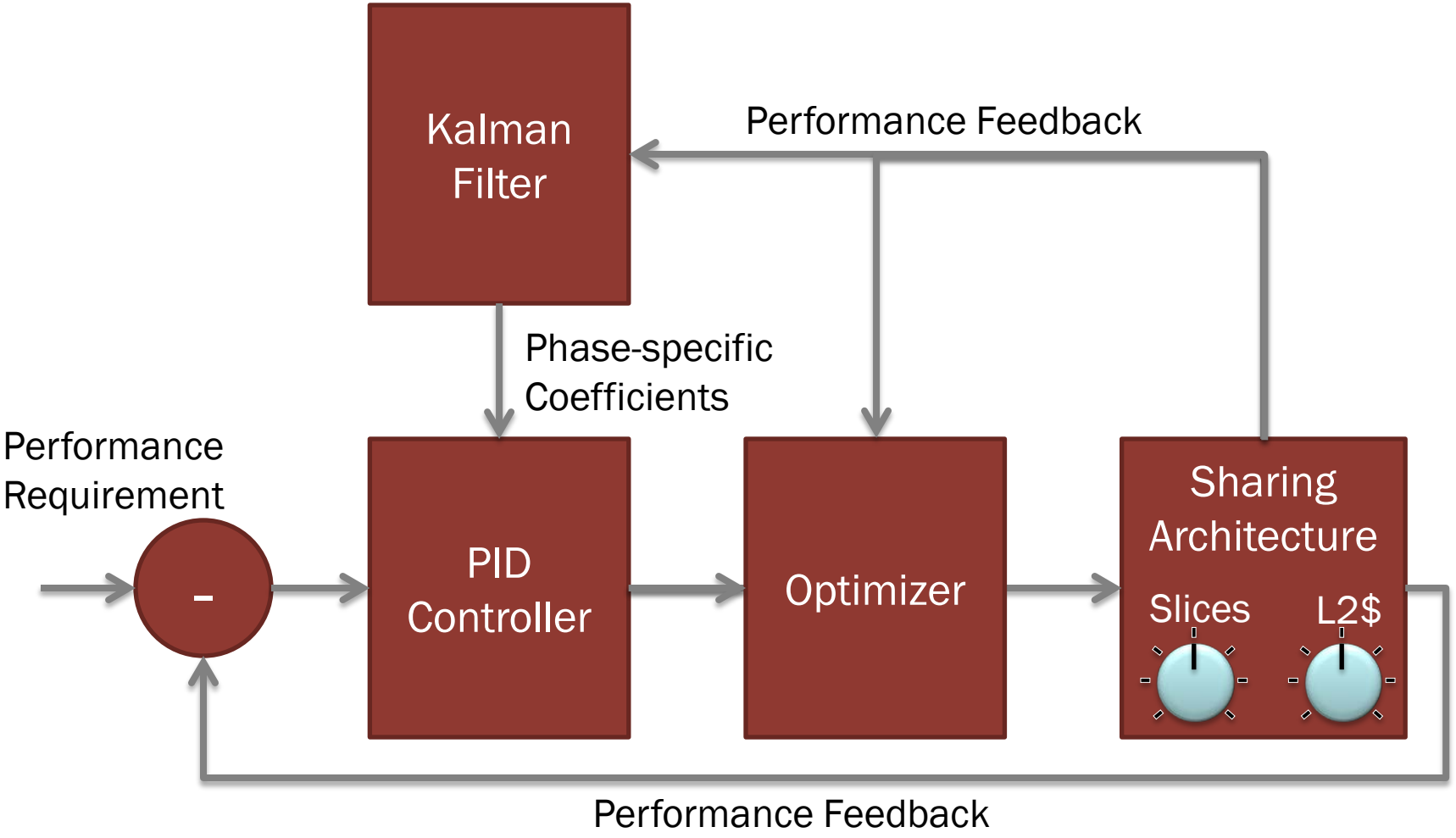


Controlling Application Performance
Adapting to Phases



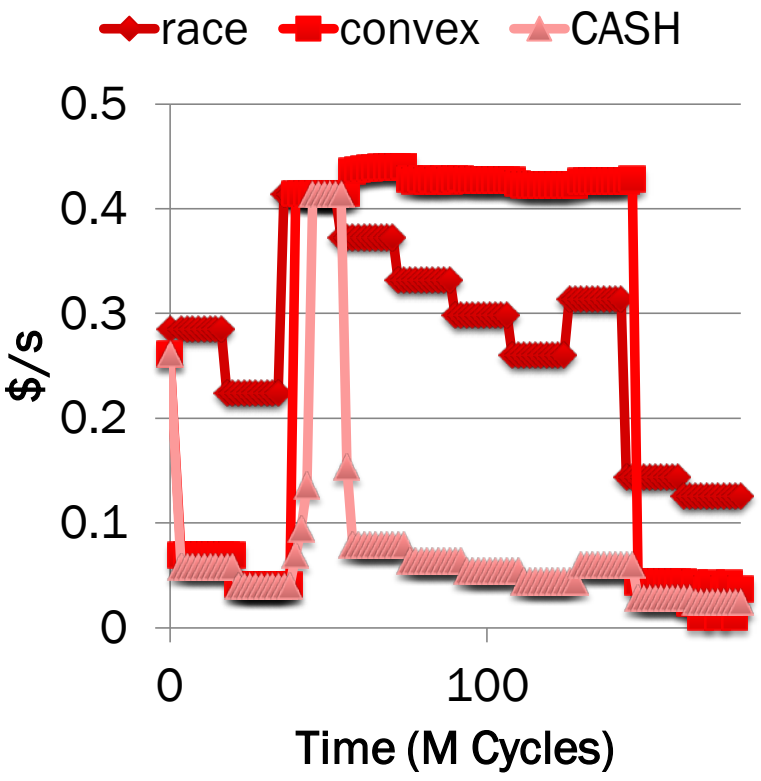
Controlling Application Performance

Learn Online

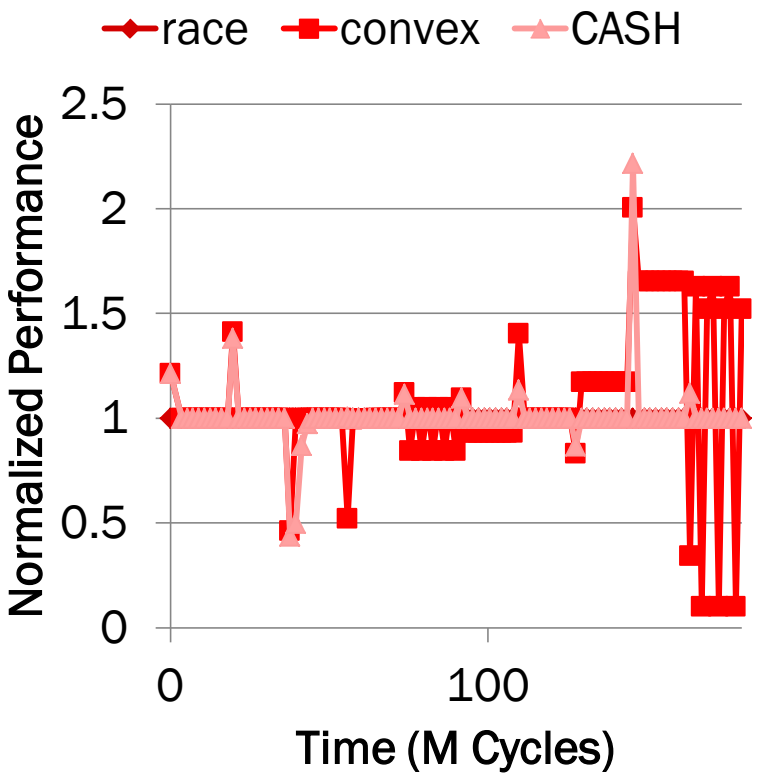


Benefit of Managing Fine-grain Configurability

Cost

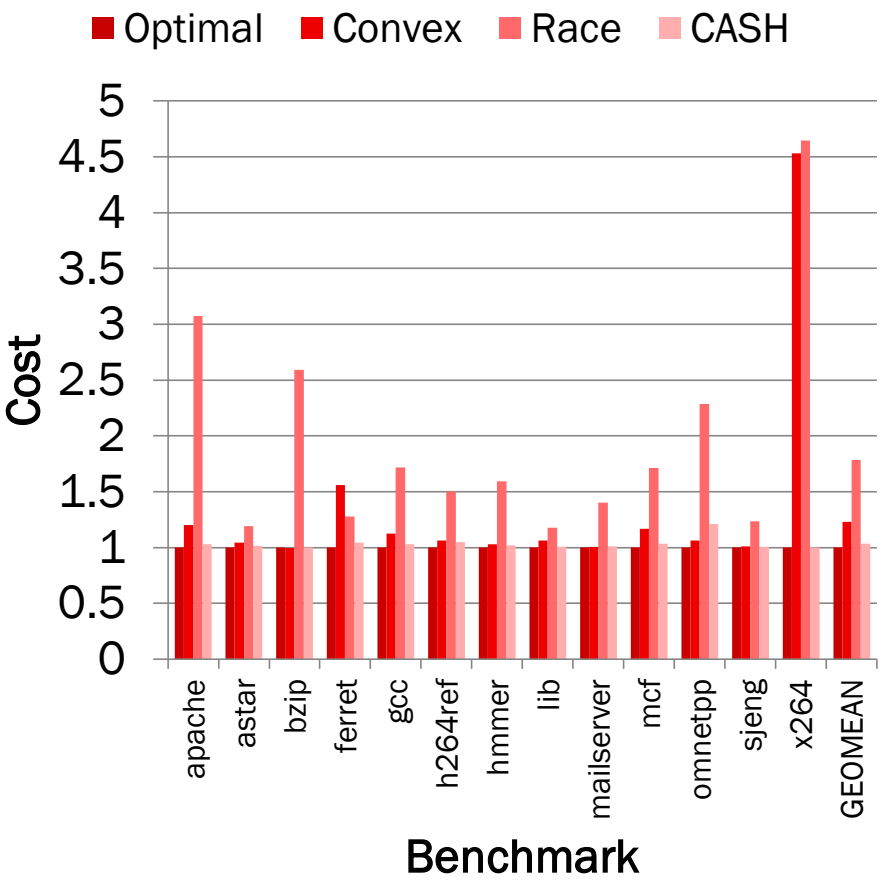


Performance

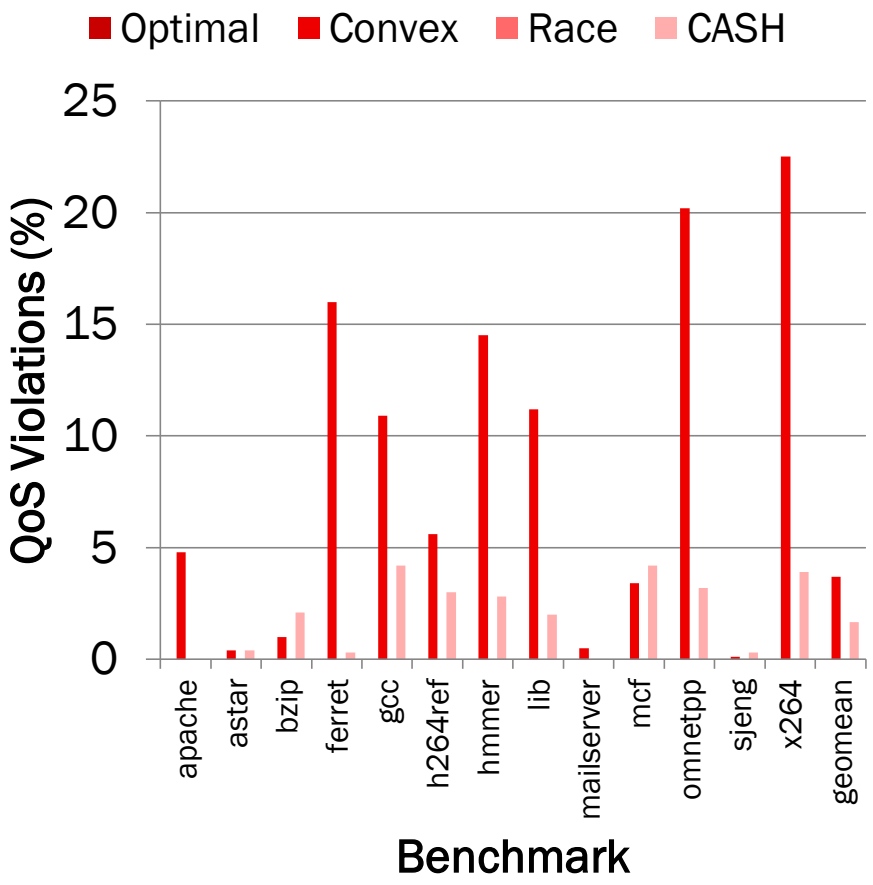


Comparison of Fine-grain Management Techniques

Cost

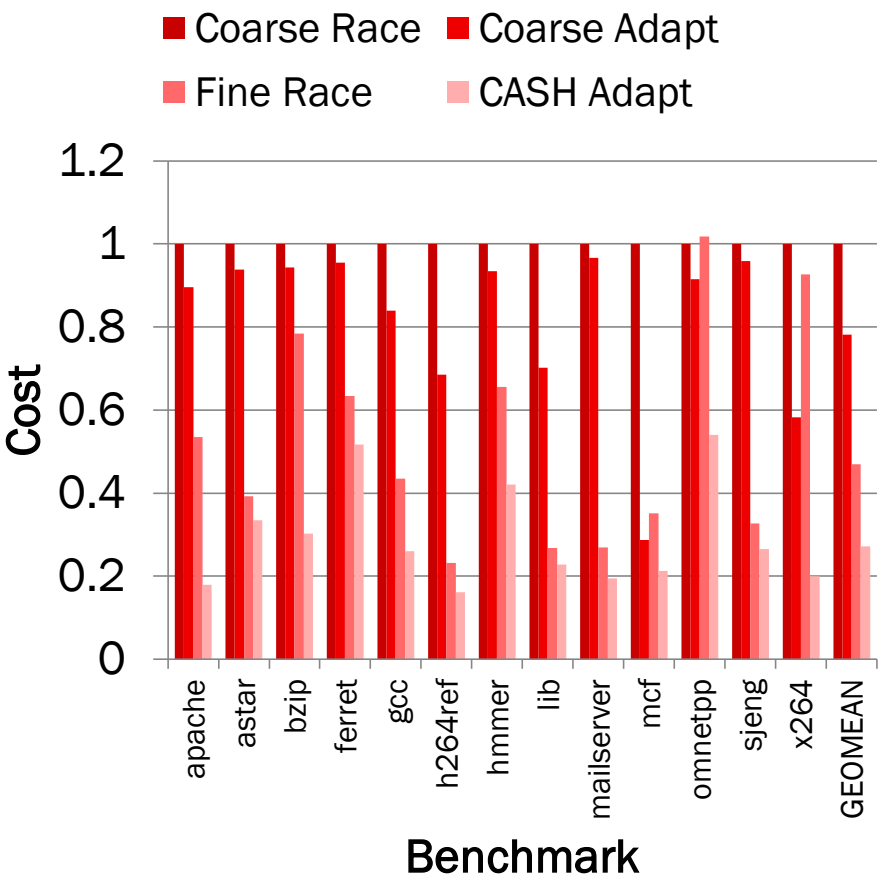


QoS Violations

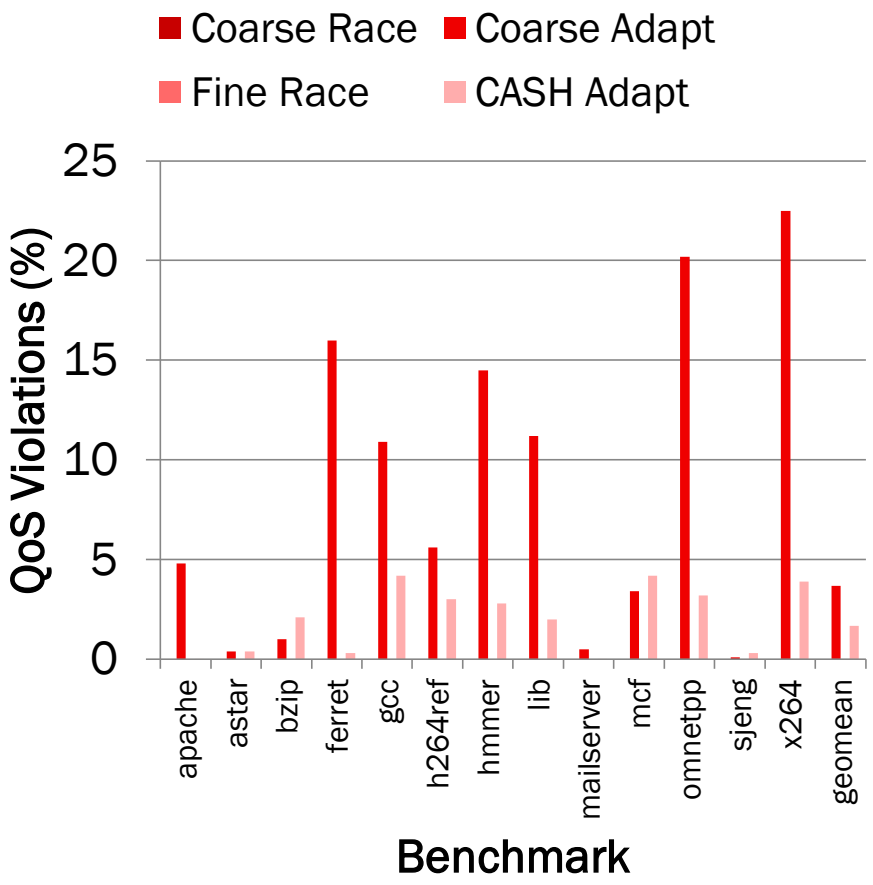


Comparisons to Coarse-grain Reconfigurable

Cost



QoS Violations



CASH: Supporting IaaS Customers with a Sub-core Configurable Architecture

PRINCETON UNIVERSITY
PRINCETON
School of Engineering and Applied Science

Yanqi Zhou Henry Hoffmann David Wentzlaff

THE UNIVERSITY OF CHICAGO
Department of Computer Science