Go further, faster®

From server-side to host-side:

Flash memory for enterprise storage

Jiri Schindler et al. (see credits)

Advanced Technology Group

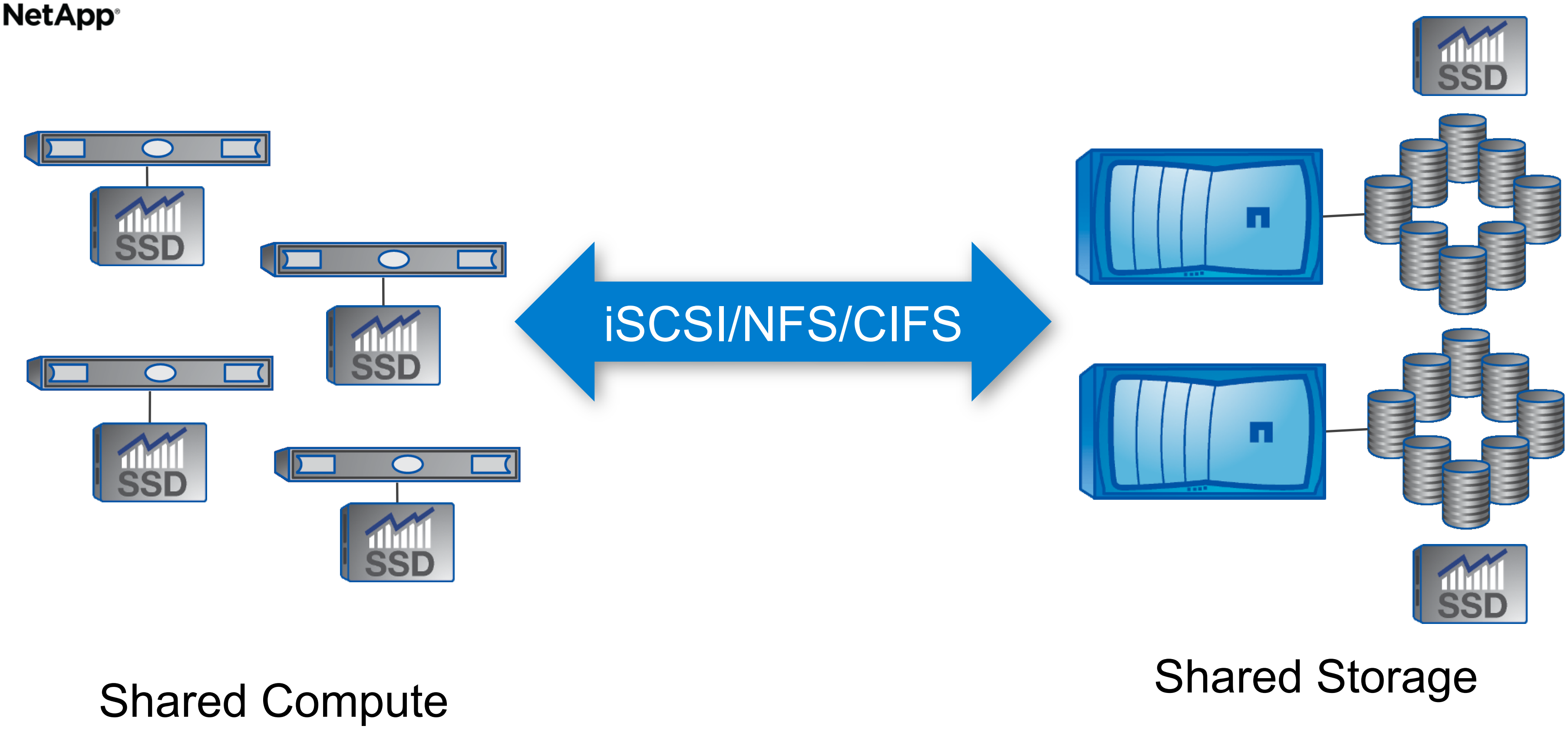
NetApp

May 9, 2012

v 1.0

© 2012 NetApp, Inc. All rights reserved.

Data Centers with Flash SSDs



How do we make effective use of flash SSDs while preserving the benefits of shared storage?

© 2012 NetApp, Inc. All rights reserved.

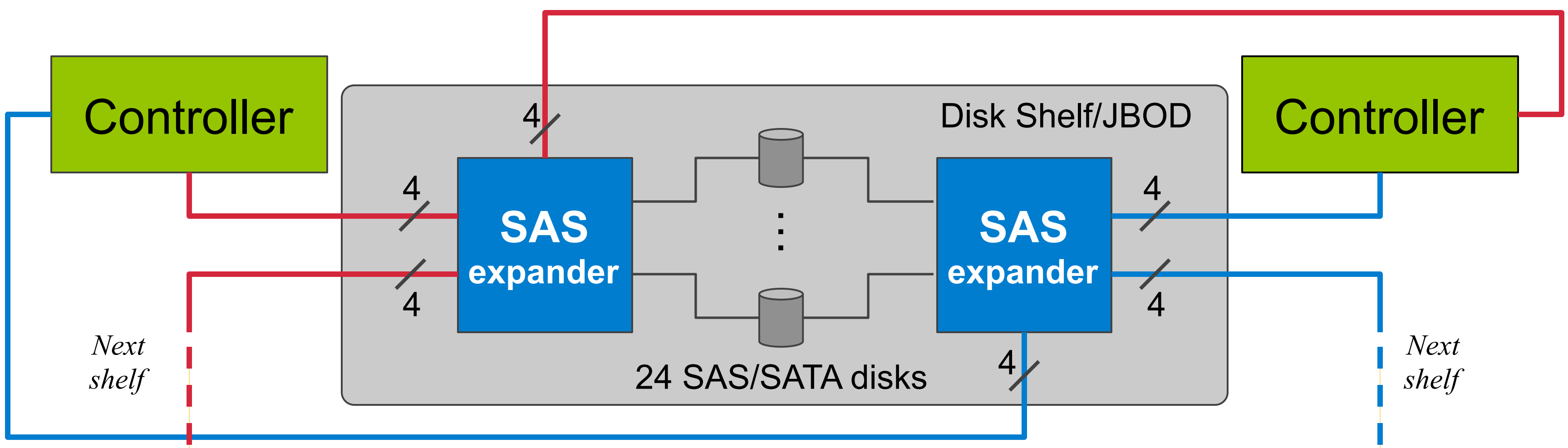
Step I. Replace HDDs with SSDs

© 2012 NetApp, Inc. All rights reserved.

SAS Disk Shelves

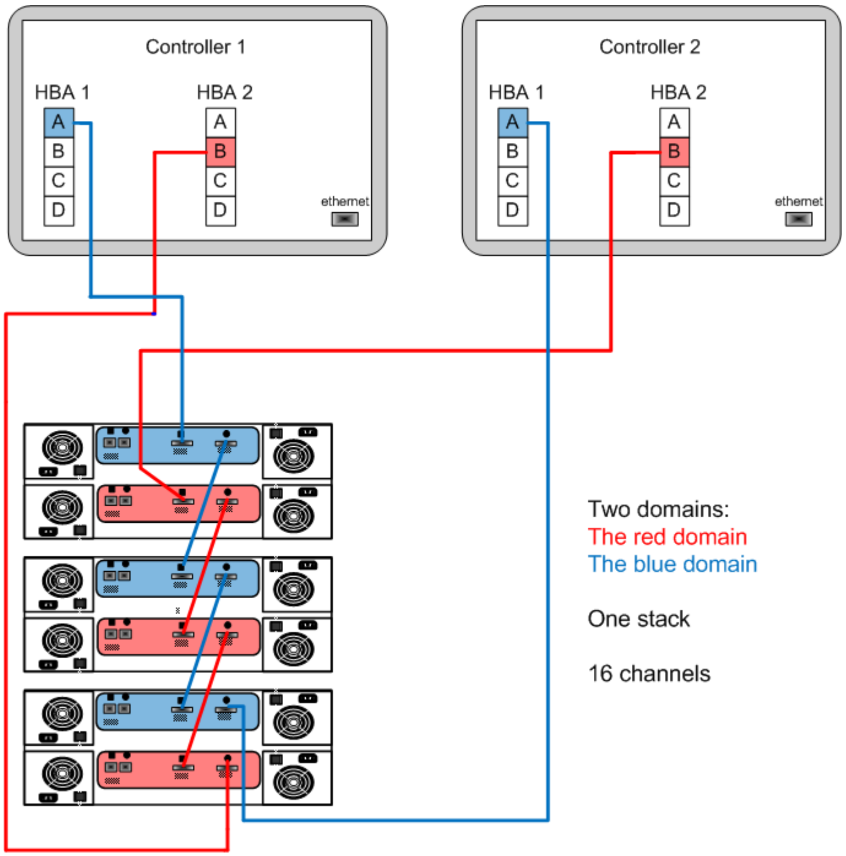
DS4243 Shelf: 24x 4U

- Each disk in a carrier
  - Hot-swappable
  - 3.5” or 2.5” form-factor
- Serial-attached SCSI expanders
  - 36-port cross-bar switch
  - Single link: 3 or 6 Gb/s, ~60-80K IOPS



© 2012 NetApp, Inc. All rights reserved.

Daisy-chaining Disk Shelves



- Single Flash SSD
  - 10-12K IOPS
  - ~125 MB/s
- Port-to-port links
  - Opened individually
- As chain grows:
  - IOPS diminished
  - BW limitations

SAS back-end (w/ many shelves) can be an IOPS bottleneck

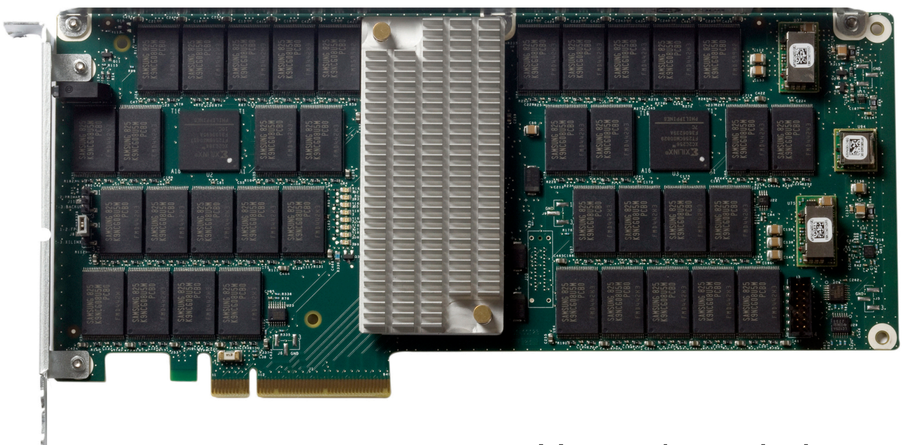
© 2012 NetApp, Inc. All rights reserved.

NetApp®

Step II. Optimize for ONTAP Data Path

© 2012 NetApp, Inc. All rights reserved.

Flash Cache (PAM-II Card) Overview



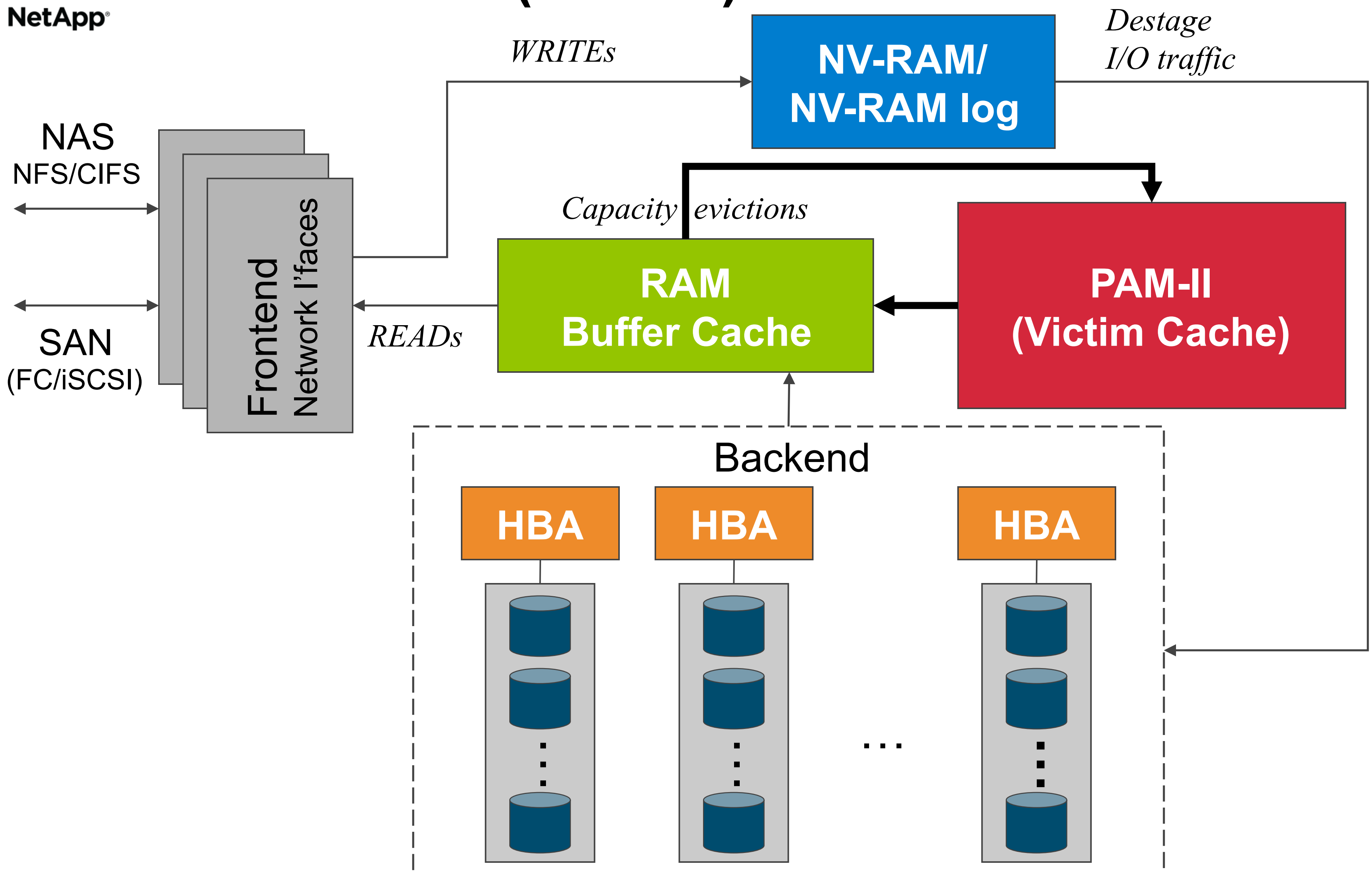
- NetApp-designed card
  - No COTS design existed
  - FPGA controller
  - 512GB SLC Flash

Up to 4 cards in a singe FAS controller (up to 8 in FAS6xx0 series)

- Specific to Data ONTAP® I/O data path
  - Read-only victim cache behind RAM buffer cache
- Minimal SW changes
  - Leverage existing RAM-based PAM card design
  - Buffer tags in RAM, simple FTL

© 2012 NetApp, Inc. All rights reserved.

Flash Cache (PAM-II)



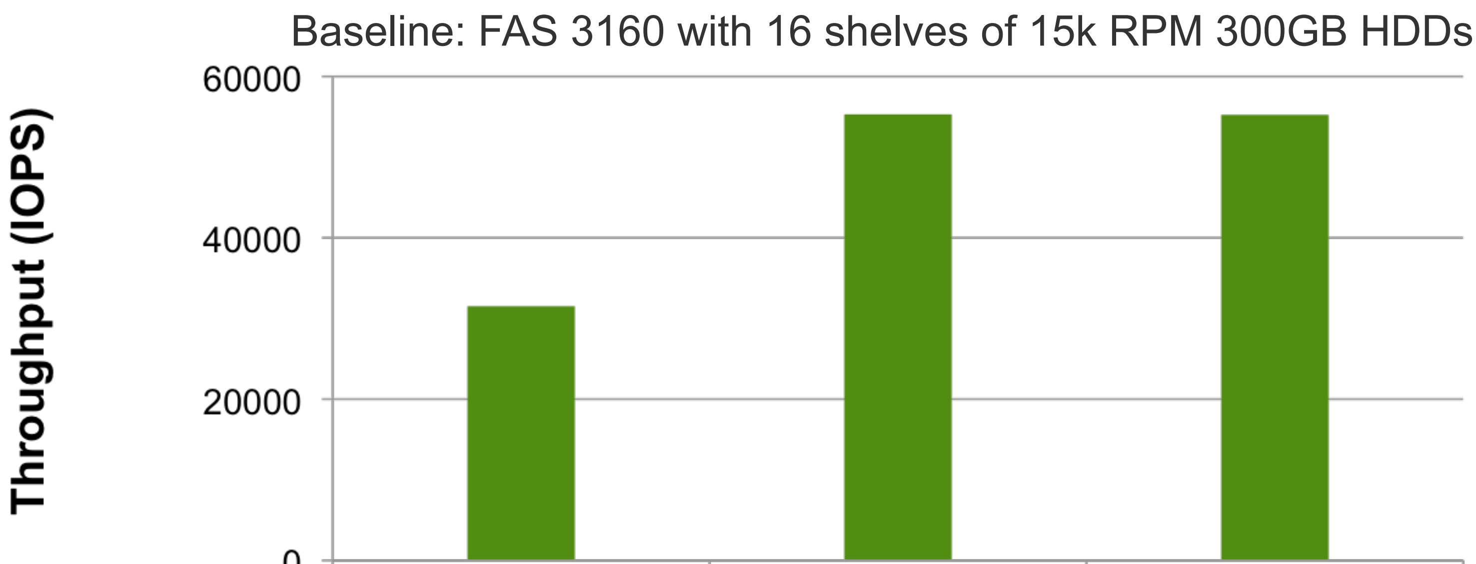
© 2012 NetApp, Inc. All rights reserved.

Managing Flash Cache

- Flash acts as buffer for read-only (clean) data
  - No “in-place” overwrites of cached data
  - Invalidation of existing mapping on new write
- Simple FTL
  - Circular buffer w/ generation garbage collector
    - implicit wear leveling
- Tag store for buffer headers in RAM
  - Takes away RAM buffer cache space for data
    - Non-trivial with 8 PAM-II cards/4TB of Flash

© 2012 NetApp, Inc. All rights reserved.

OLTP-like Workload Performance



		Baseline (84 HDDs)	Add Disks (140 HDDs)	Add PAM-II
1.8x	IOPS	31504	55297	55230
1.6x	Resp. Time (ms)	27.5	27.5	16.9
	Power (kW)	3.1	6.8	3.1
	Rack Space (U's)	24	54	24

Same operational costs, 30% COGS price reduction w/ PAM-II

Source: NetApp White Paper WP-7082-0809 http://media.netapp.com/documents/wp-7082.pdf

© 2012 NetApp, Inc. All rights reserved.

SPECsfs2008 (nfs.v3) Performance

Baseline: FAS 3160 with 16 shelves of 15k RPM 300GB HDDs

	Baseline (FC-AL HDDs)	FC-AL Disks w/ PAM II	SATA Disks w/ PAM-II
IOPS	60409	60507	60389
Resp. Time (ms)	2.18	1.58	2.18
Disks	224	56	96
Rack Space (U's)	35	11	17

Cost savings: replace FC-AL with fewer SATA HDDs & PAM-II

Source: http://www.spec.org/sfs2008/results/sfs2008nfs.html

© 2012 NetApp, Inc. All rights reserved.

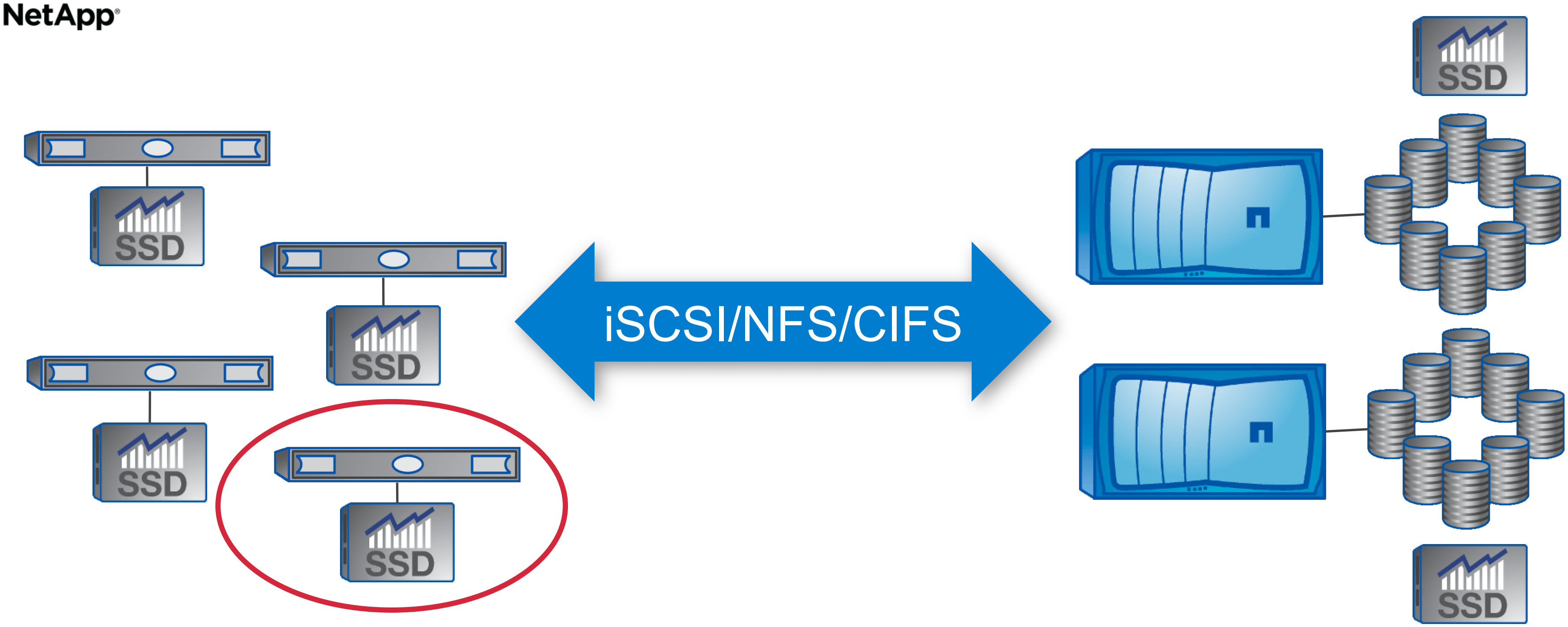
Step III. Combine with Host-side

Project Mercury

Presented at MSST '12 Conference

© 2012 NetApp, Inc. All rights reserved.

Data Centers with Flash SSDs



Shared (Virtualized) Compute

Shared Storage

Reliability and availability is different at host-side

© 2012 NetApp, Inc. All rights reserved.

Available and Durable

<u>Goal</u>

- Never lose data in any situation

<u>Consequence</u>

- Write-through policy

Chose common denominator policy
Other policies possible that leverage app's specifics

© 2012 NetApp, Inc. All rights reserved.

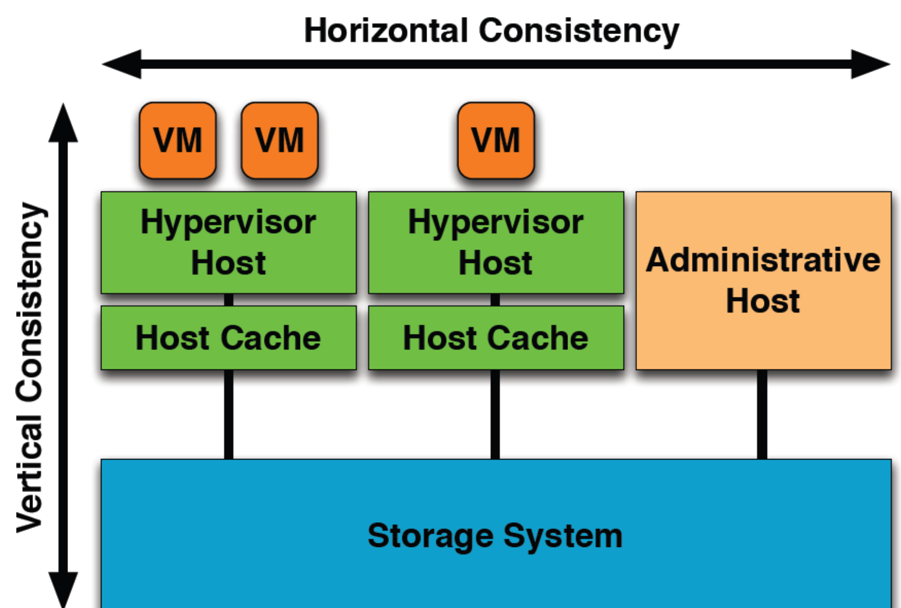
Correct and Consistent

Goals

- Consistency with storage array
- Consistent with peers

Consequences

- Cache non-shared objects
- Invalidate on migration, restore, etc.



© 2012 NetApp, Inc. All rights reserved.

Datacenter Management Integration

Goal

- Simple and transparent management

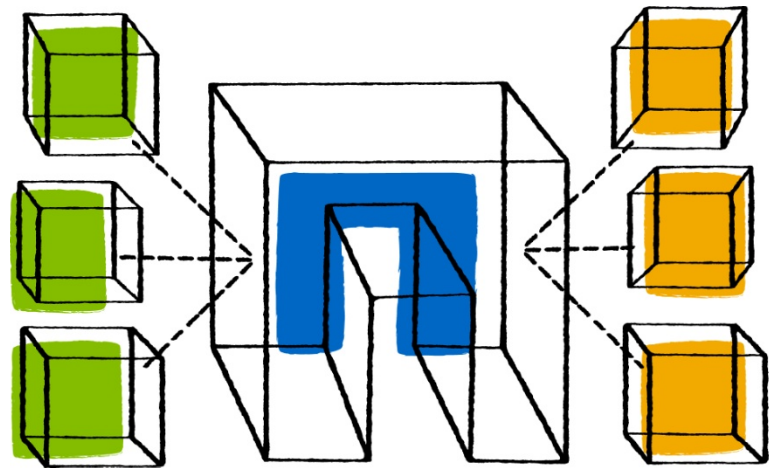
Consequence

- Hypervisor integration

Most important for the end-user deployment

© 2012 NetApp, Inc. All rights reserved.

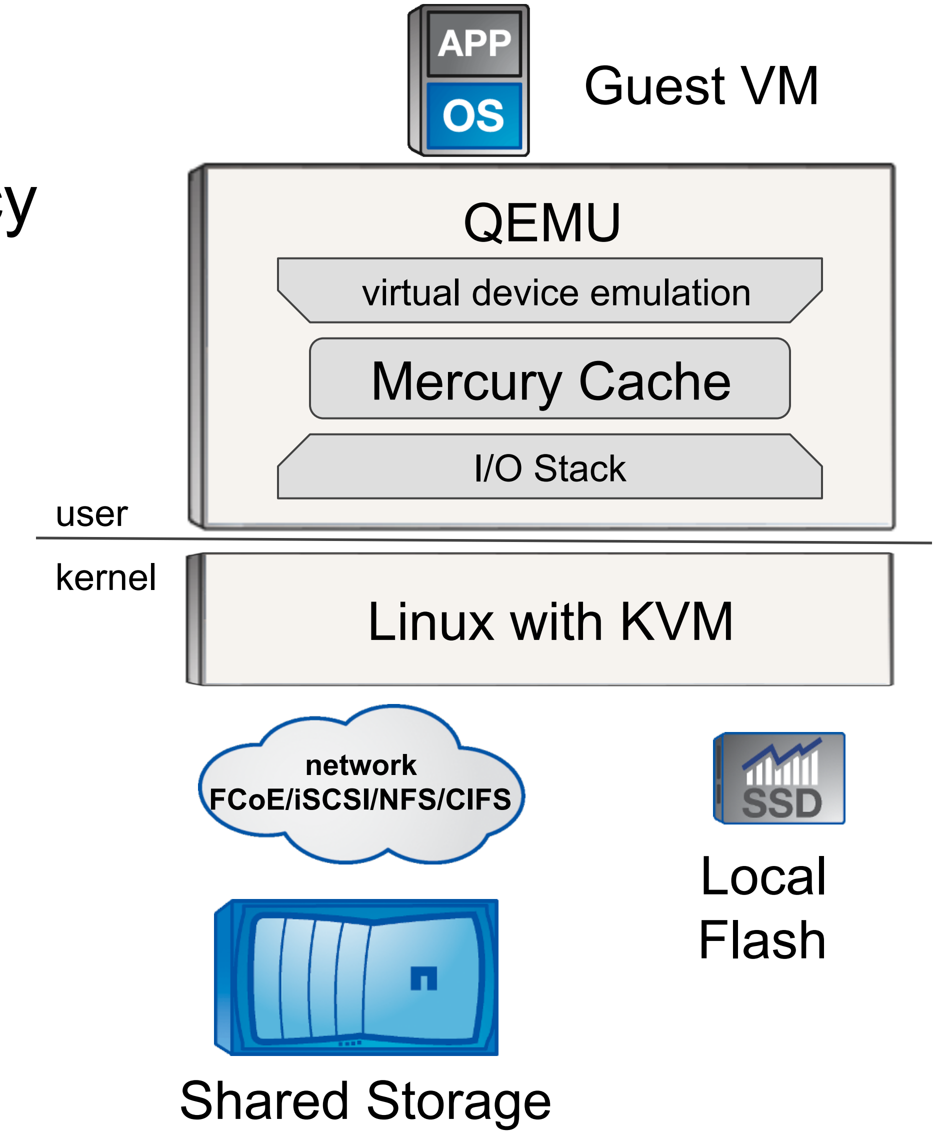
Design & Implementation



© 2012 NetApp, Inc. All rights reserved.

Prototype Implementation Overview

- Write-through
  - Simple cache consistency
- Persistent
  - Warm cache on restart
  - Cache durability after a crash is ongoing work



© 2012 NetApp, Inc. All rights reserved.

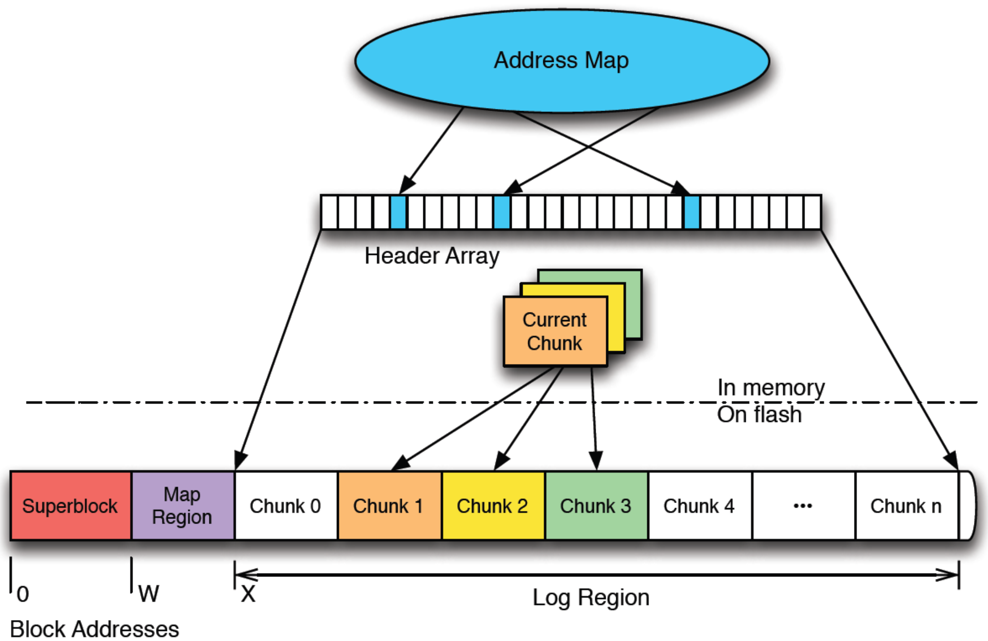
Detecting Cache Hits

- All cache metadata in RAM for speed.
  - Mercury is a second-level cache →
    modest hit rate →
    minimize cache overhead
  - Memory-to-cache ratio is 0.5%
    (e.g., 500 GB cache requires 2.5 GB of RAM)
- Cache headers
  - One header for each block in the cache
- Address Map
  - (primary storage, LBA) keys, header index values
  - Implemented with hash table, O(1) lookup time

© 2012 NetApp, Inc. All rights reserved.

Cache Insertion

- Specialize I/O access patterns for flash
  - LFS-style writes
    - Large chunks match erase (meta) block size
    - Minimizes cleaning/slowdown at the SSD FTL



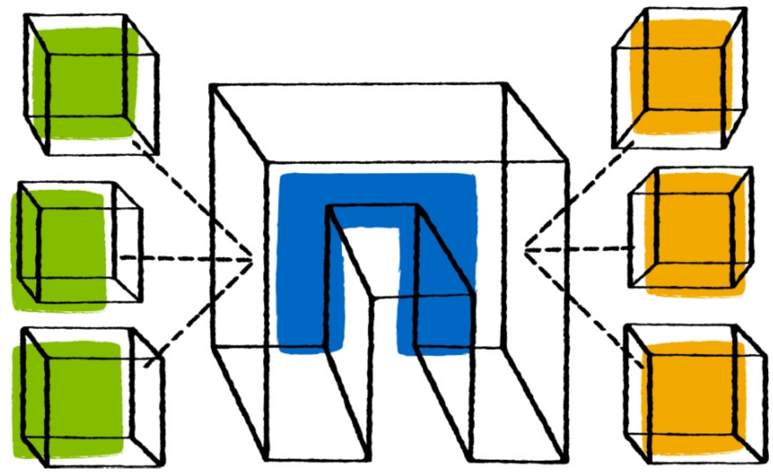
© 2012 NetApp, Inc. All rights reserved.

Admittance Policies

- Unrestricted (default)
  – All writes and read misses inserted in the cache
- Write-Around
  – writes skip the cache
- Sequential I/O Bypass (ongoing work)
  – Sequential reads, writes, or both skip the cache

© 2012 NetApp, Inc. All rights reserved.

Results



© 2012 NetApp, Inc. All rights reserved.

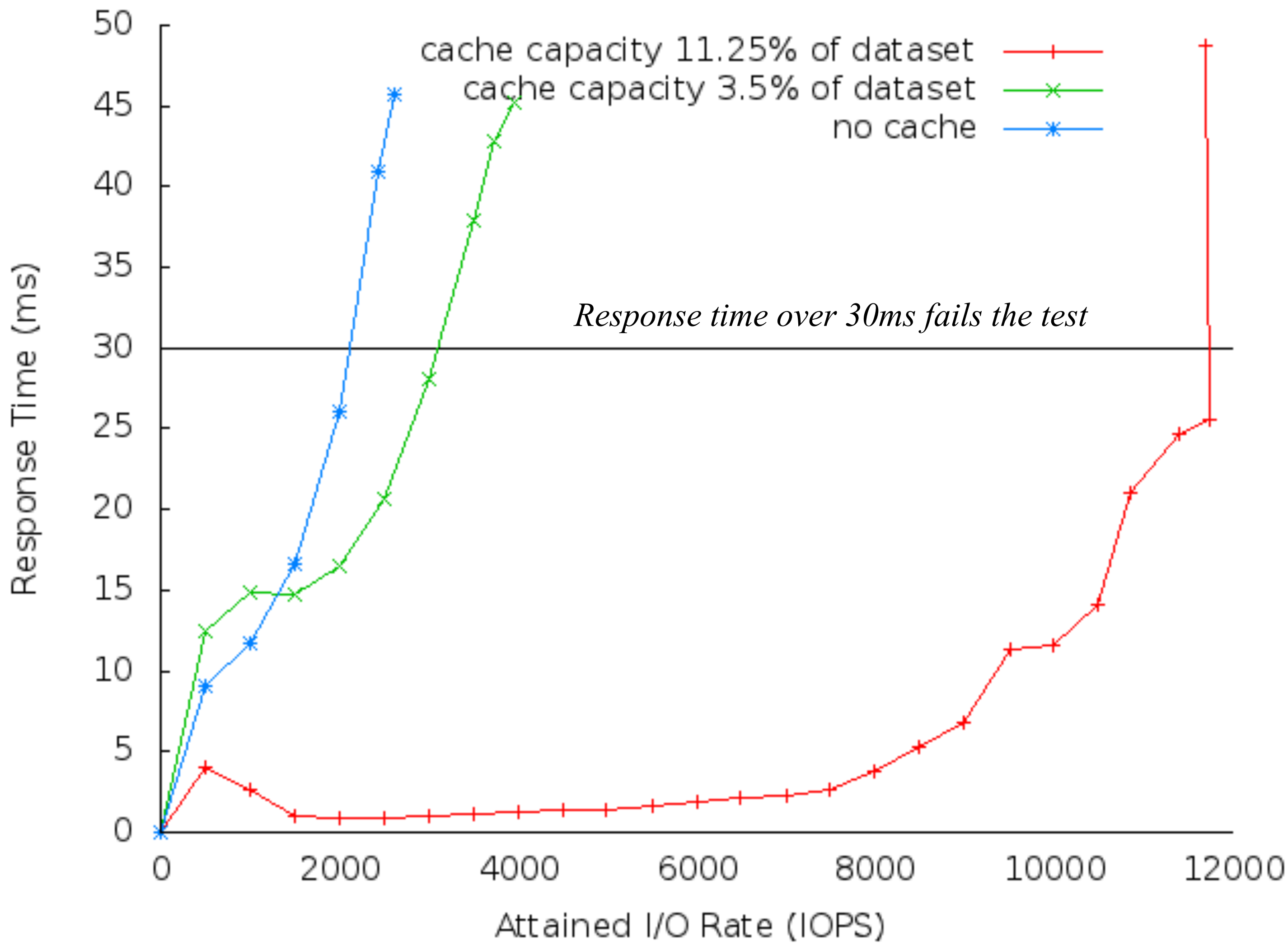
Evaluation Setup

- Two workloads:
  - Microsoft® Exchange Jetstress
  - NetApp® Enterprise Workload[1]
- Flash cache
  - PCIe device with SLC (single-level cell) flash
    - Paper contains SLC and MLC SSD results
- Other hardware
  - x86 Server with Linux, KVM/QEMU
  - NetApp FAS3270 with iSCSI LUN(s)

[1] S. Daniel et al., *A portable, open-source implementation of the SPC-1 workload*.

© 2012 NetApp, Inc. All rights reserved.

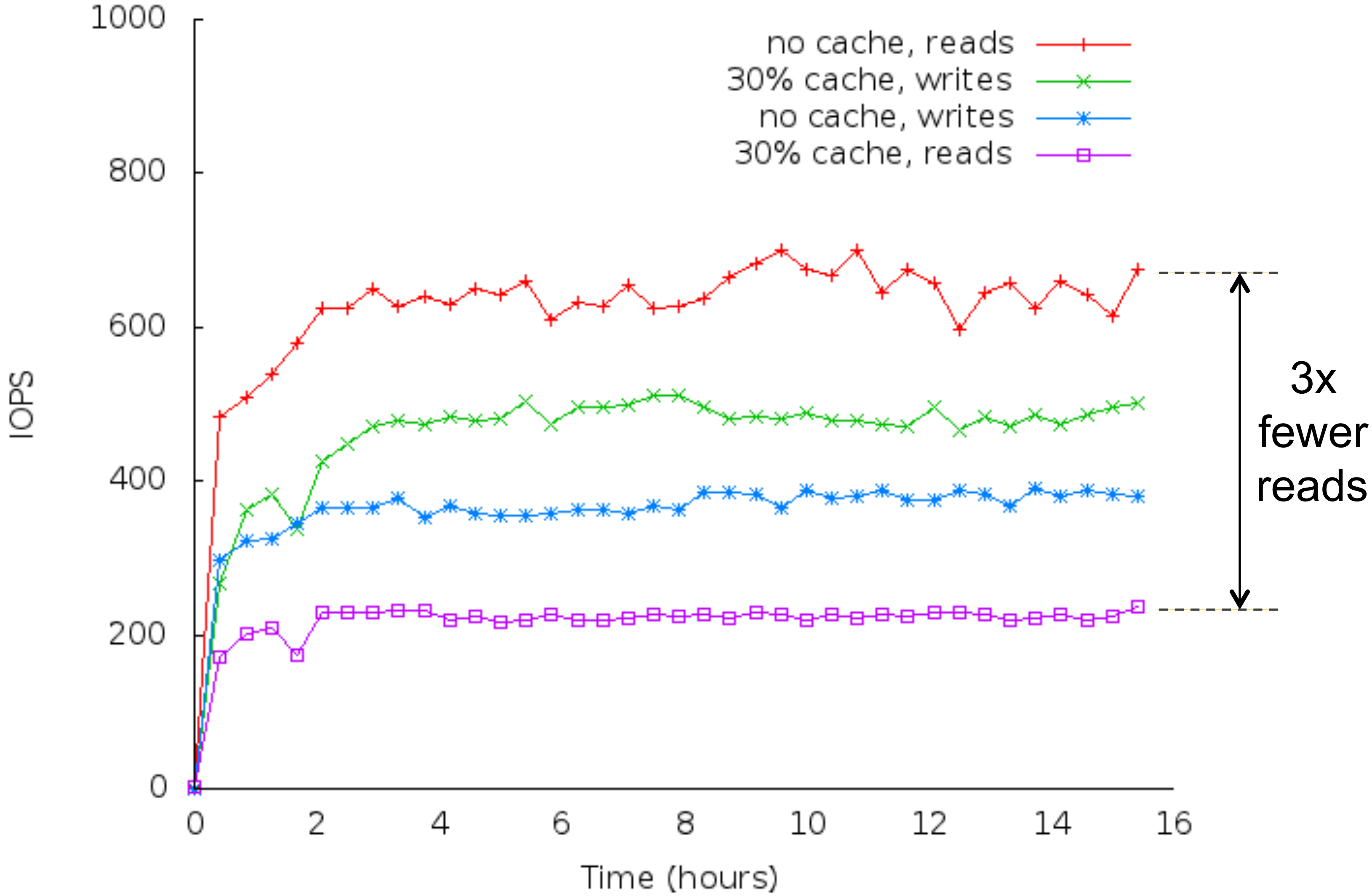
Significant Response Time Improvement



Enterprise workload. Unrestricted admittance policy. CLOCK eviction policy.

© 2012 NetApp, Inc. All rights reserved.

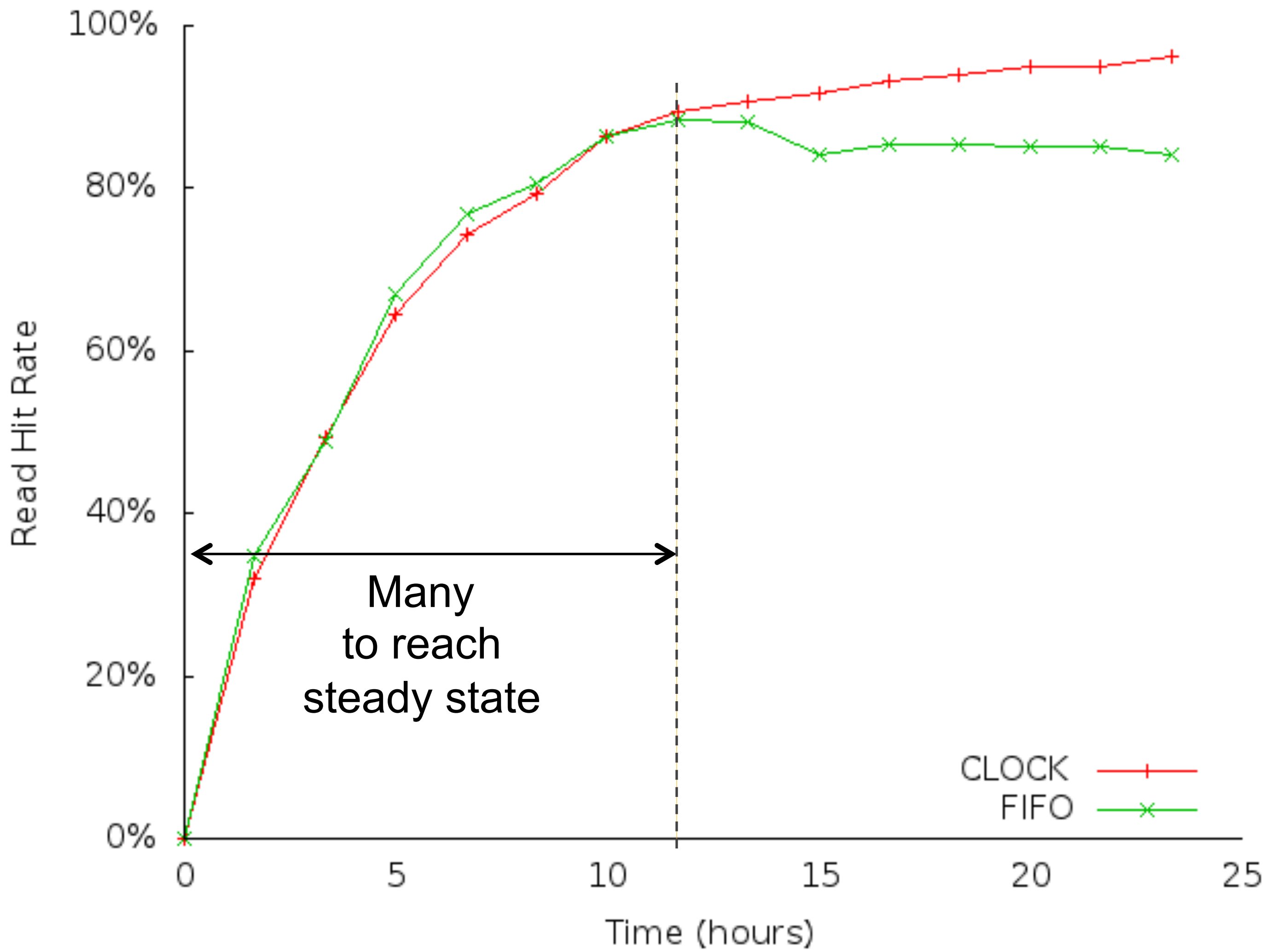
Reducing Access to Networked Storage



Jetstress workload. Unrestricted admittance policy. FIFO eviction policy.

© 2012 NetApp, Inc. All rights reserved.

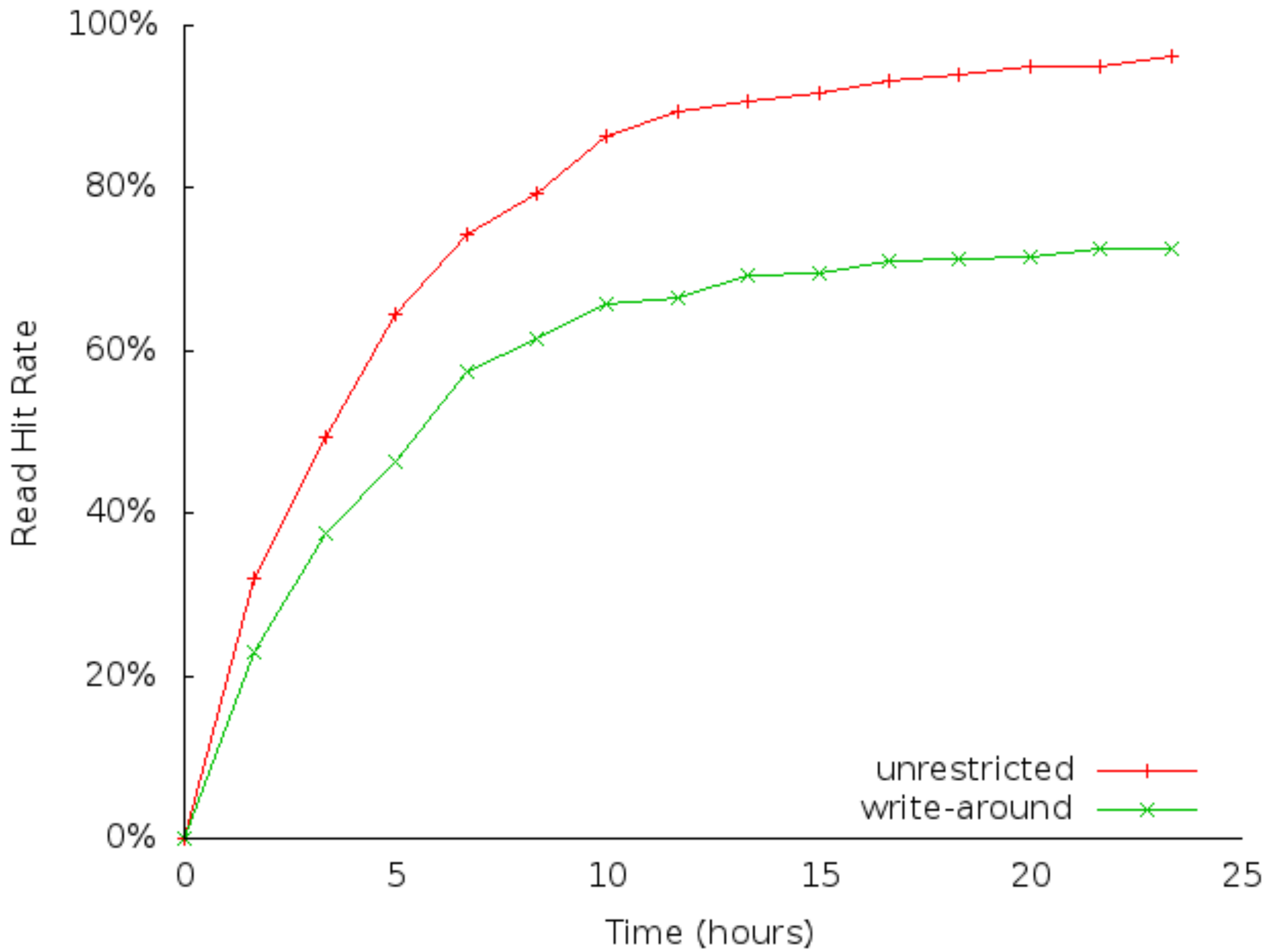
Warming Cache: Takes a Long Time



Enterprise workload. Unrestricted admittance policy. Flash capacity set to 11.25% of dataset.

© 2012 NetApp, Inc. All rights reserved.

Unrestricted Beats Write-Around



Enterprise workload. CLOCK eviction policy. Flash device capacity set to 11.25% of dataset.

© 2012 NetApp, Inc. All rights reserved.

Host-side Flash Summary

- **Host-side flash**
  - minimizes flash access latency
- **Hypervisor-based I/O cache**
  - simplifies deployment
- **Persistent**
  - cache is warm on a restart
- **Write-through**
  - consistent with primary storage

© 2012 NetApp, Inc. All rights reserved.

Concluding Remarks

- Working with real-world constraints
  – Timing is everything

- Design for the long haul
  – Deliver something useful fast

- Learn from the users
  – Collect field data

- Improve design in iterations over-time

© 2012 NetApp, Inc. All rights reserved.

Credits

- Efforts of many product engineers

- Project Mercury

Steve Byan	James Lentini
Anshul Madan	Luis Pabón
Michael Condict	Jeff Kimmel
Steve Kleiman	Christopher Small
Mark Storer	

Advanced Technology Group
NetApp

© 2012 NetApp, Inc. All rights reserved.

NetApp

Thank you

© 2012 NetApp, Inc. All rights reserved.