NetApp™

Go further, faster™

Re-designing Enterprise Storage Systems for FLASH Memory

Jiri Schindler

Advanced Technology Group

v. 1.2

NVRAMOS 2009, Jeju Island, S. Korea, October 19-20

Building Systems and Software

Scalable Family of Networked Storage Systems

Integrated Data Management Solutions

vmware Virtualized Datacenter Environments Windows Server 2008 Hyper-V

Databases ◆ Messaging (E-mail) ◆ Engineering Applications

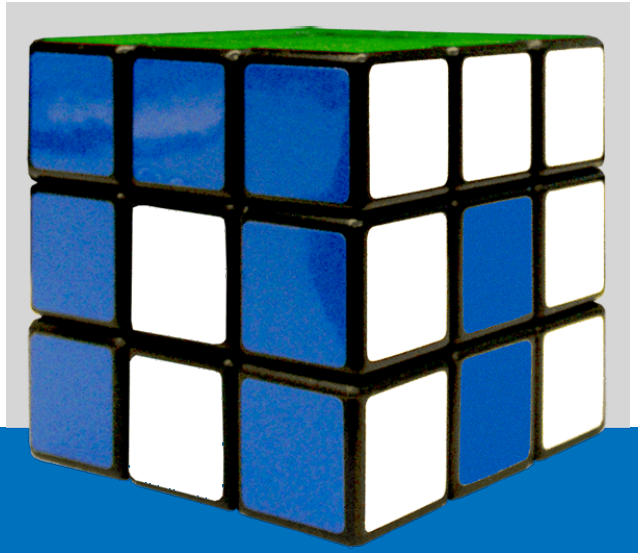
Data Protection & Retention Solutions

Backup ◆ Security ◆ DR ◆ Compliance ◆ Archive

Talk Outline

- Examples of Flash in the enterprise
- The future of ESS architectures w/ Flash
- Research challenge

Enterprise Storage Systems (ESS) Basics



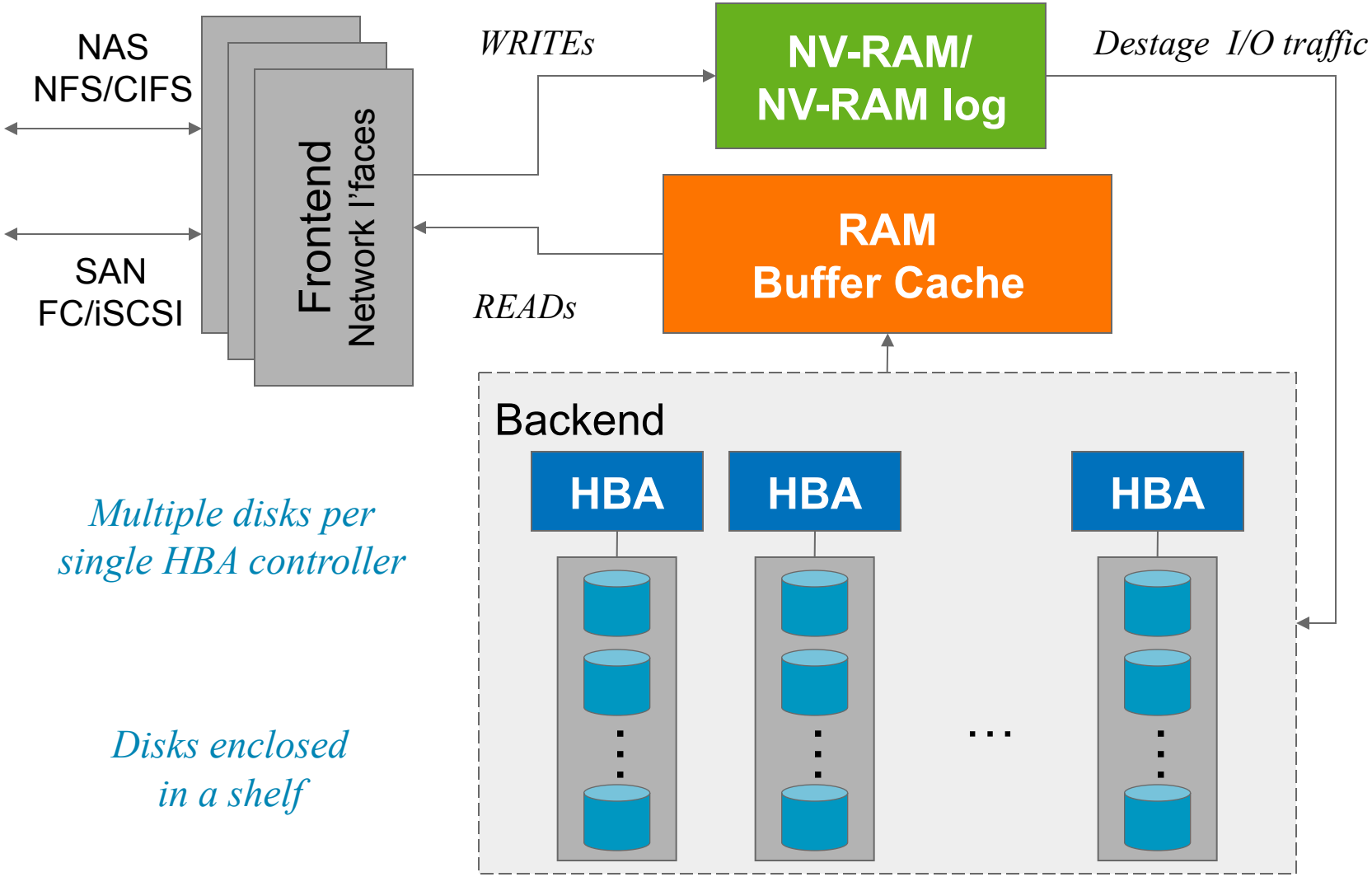
Datacenter Ecosystem

- Many clients
  - Host/client-side caching of reads in local RAM
- One shared/consolidated storage system
  - Ensures consistent view of shared data
- Consider the total cost of ownership (TCO)
  - Management/operating costs dominate
    - Push for lights-out design of data centers
    - Push for automation and easy management

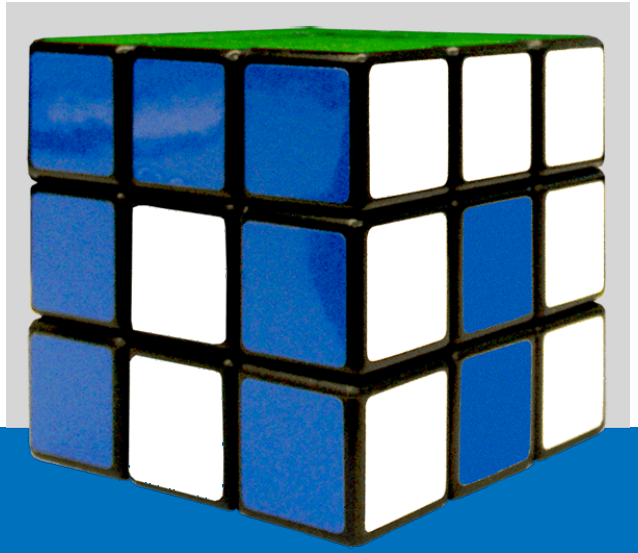
What is ESS?

- Purpose-built system
  - Commodity hardware components
    - Easier/faster integration of new technology
    - Too expensive to build custom silicone/ASIC
  - Customized software
    - Provides reliability & high data availability
      - Resiliency to component failures
    - Eases management of many components
      - One admin for PB (1000+) of disks
- Single architecture for a variety of workloads
  - Single OS: Data ONTAP® (for NetApp)

Enterprise Storage Systems (ESS)



Flash Presence in the Enterprise: Architectures & Trends



Host/client-side Flash memory

- Server-class motherboard chipsets
  - Flash memory attached to the server front side bus
- Client direct-attached storage systems
  - PCI card with OS driver
    - Fusion I/O SLC/MLC card

    http://www.fusionio.com/Products.aspx
- Drawbacks
  - No HA capabilities
  - Not amenable to shared access by multiple clients

Network-attached subsystems

- Storage interface
  - Texas Memory Systems RamSAN-500
    - FC storage system w/2TB, RAID config

    http://www.ramsan.com/products/ramsan-500.htm
  - Violin 1010 storage appliance
    - FC/Ethernet-based memory appliance, up to 4TB

    http://violin-memory.com/Flash

- Fast storage for “Cloud” applications
  - Schooner MEMCACHEd appliance
    - Specialized, higher-level protocol I’face

    http://www.schoonerinfotech.com/products/memcached-appliance.html

ESS Server-side Flash

- SSDs (covered in last year's NVRAMOS talk)
  - Replacement of 3.5" or 2.5" SFF HDDs disks

- PCI-based accelerator cards
  - NetApp® PAM-II Card
    - 256/512GB victim cache

    http://www.netapp.com/us/products/storage-systems/performance-acceleration-module/
  - Sun F20 controller
    - SLC-based 96GB cache for 8-disk SAS controller

    http://www.sun.com/storage/disk_systems/sss/f20/

Incremental Architectural Changes

- Basic premise
  - Hardware component change is easier than software change if it fits existing architecture
  - SW testing more complex than HW qualification
- Replace HDDs with SSDs

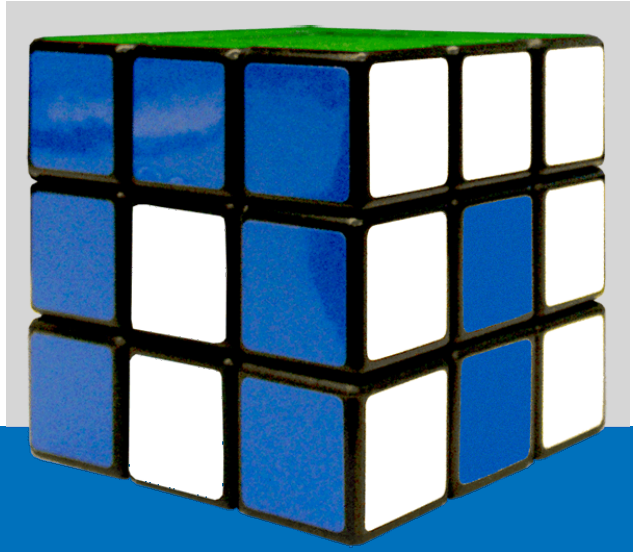
  Step 1: HW component substitution
  - No need for data path software changes
  - Improve (read) IO throughput

  Step 2: Architectural and/or SW changes
  - Differentiated or automatic tiering of data
    - Place FS metadata, hot-data, or working set into SSDs

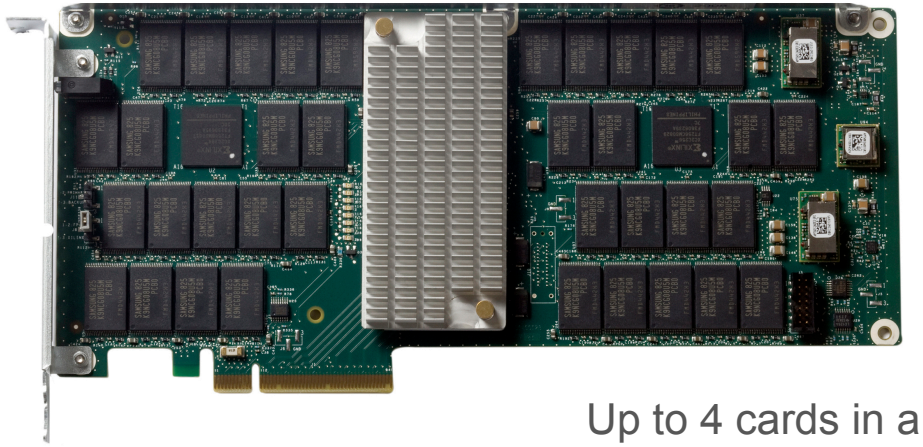
Incremental Architectural Changes cont'd

- Replace the system's NV-RAM implemented as battery-backed RAM with Flash memory
  - High write density not suited to Flash technology
    - All of system's writes go through NV-RAM
      - Throughput mismatch
      - Limited erase cycles of SLC & MLC
    - Flash has poor IOPS/GB performance
      - Larger NVRAM upsets the NVRAM:HDD balance
- Improvements driven by business reasons
  - Short time-to-market
  - Acquisition and operational cost efficiencies

Example of Incremental Architectural Changes: NetApp® PAM II Card



NetApp® PAM-II Overview

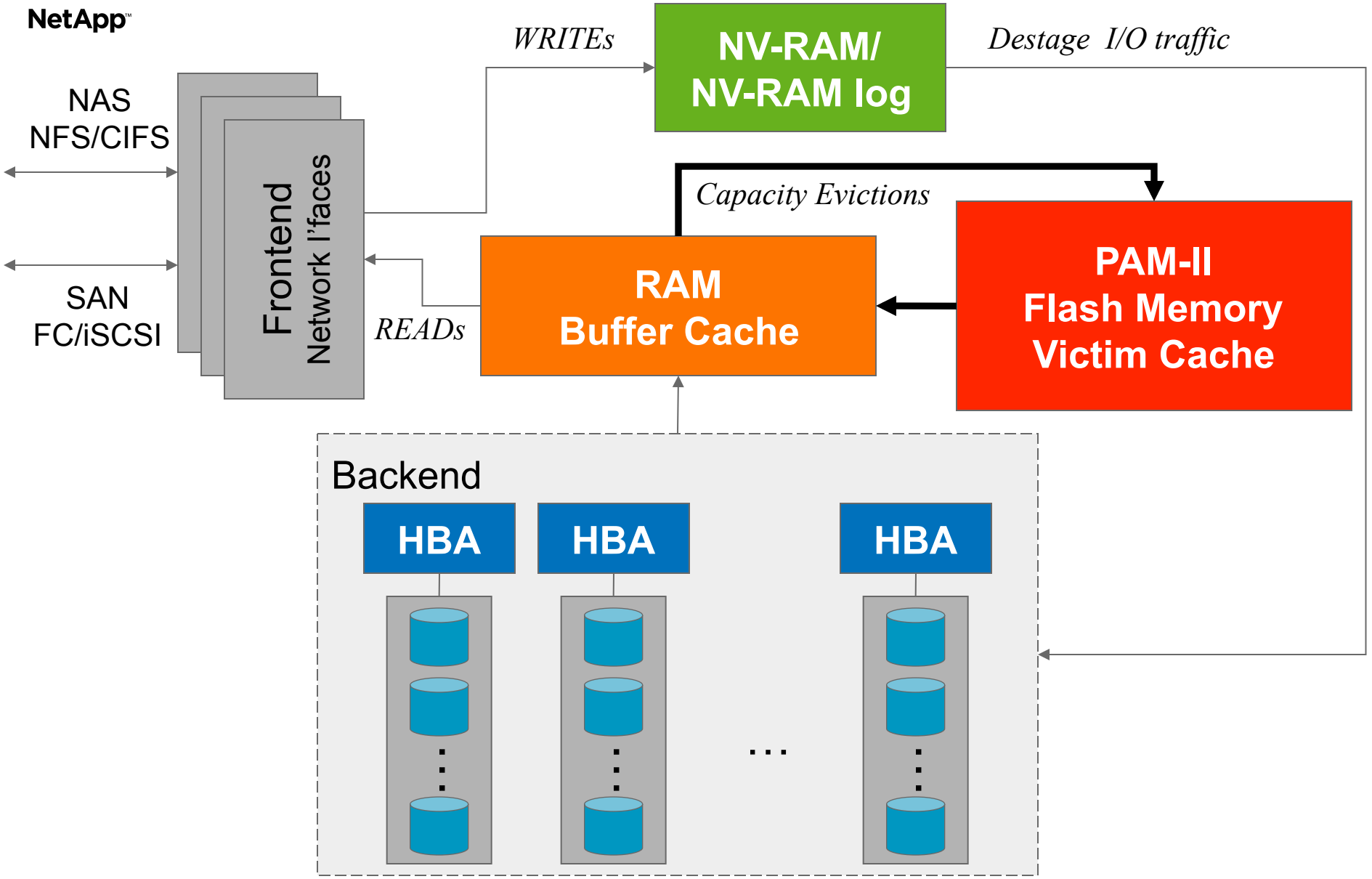


- NetApp-designed card
  - No COTS design existed
  - FPGA controller
  - 256/512GB SLC Flash

Up to 4 cards in a singe FAS controller (up to 8 in FAS60x0 series)

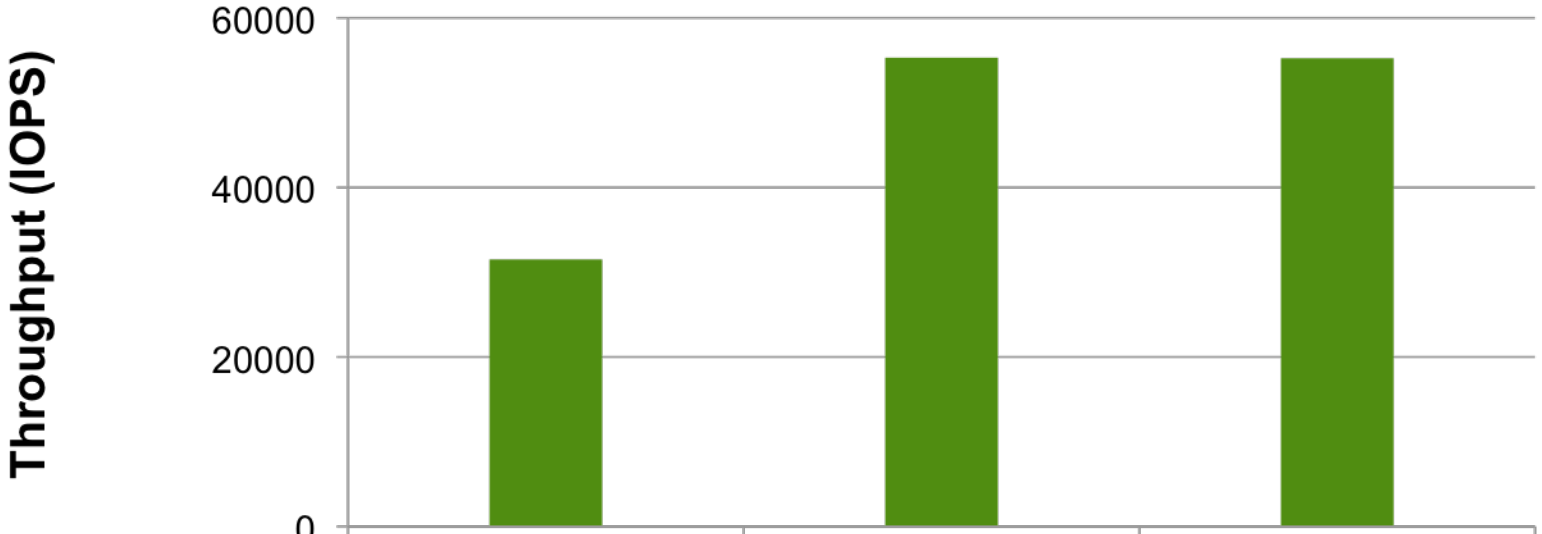
- Specific to Data ONTAP® I/O data path
  - Read-only victim cache placed between RAM buffer cache and back-end HDDs
- Minimal SW changes
  - Leverage existing RAM-based PAM card design
  - Buffer tags in RAM, simple FTL

PAM-II Victim Cache



OLTP-like Workload Performance

Baseline system: FAS 3160 with 6 shelves of 15k RPM 300GB HDDs



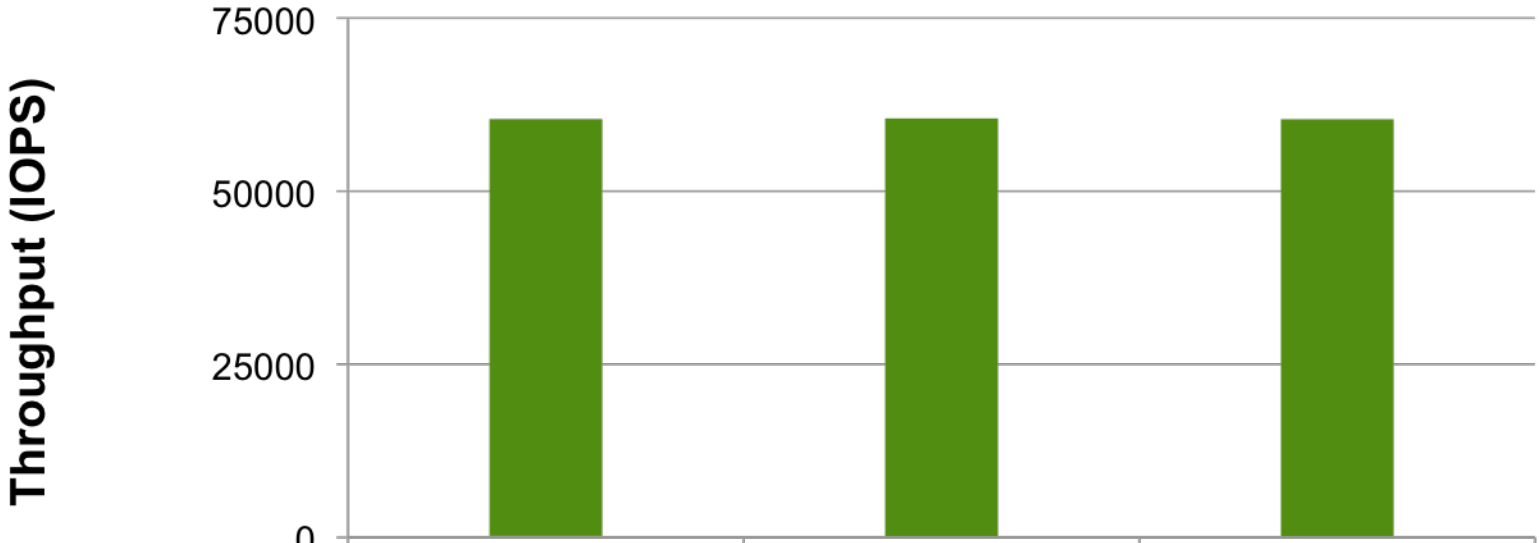
		Baseline (84 HDDs)	Add Disks (140 HDDs)	Add PAM-II
1.8x	IOPS	31504	55297	55230
1.6x	Resp. Time (ms)	27.5	27.5	16.9
	Power (kW)	3.1	6.8	3.1
	Rack Space (U's)	24	54	24

Same operational costs, 30% COGS price reduction

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

SPECsfs2008 (nfs.v3) Performance

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

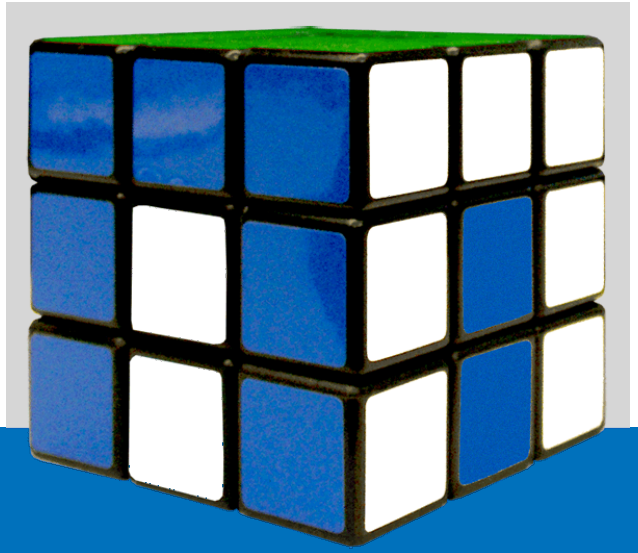


	Baseline (FC-AL HDDs)	FC-AL Disks w/ PAM II	SATA Disks w/ PAM-II
IOPS	60409	60507	60389
Resp. Time (ms)	2.18	1.58	2.18
Disks	224	56	96
Rack Space (U's)	35	11	17

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

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

Re-designing ESS



Does the "ESS architecture" picture hold?

- Balance of resources
  - Natural IOPS bottlenecks
    - Centralized NV-RAM
    - Back-end shared controllers & interconnect
- Workloads & cost considerations
  - $/IOPS analysis is not sufficient
    - Must consider capacity & power as well
    - Effective IOPS/GB
  - Cost of flash device vs. cost of infrastructure
    - Disk shelf slot tax

Big Picture Summons

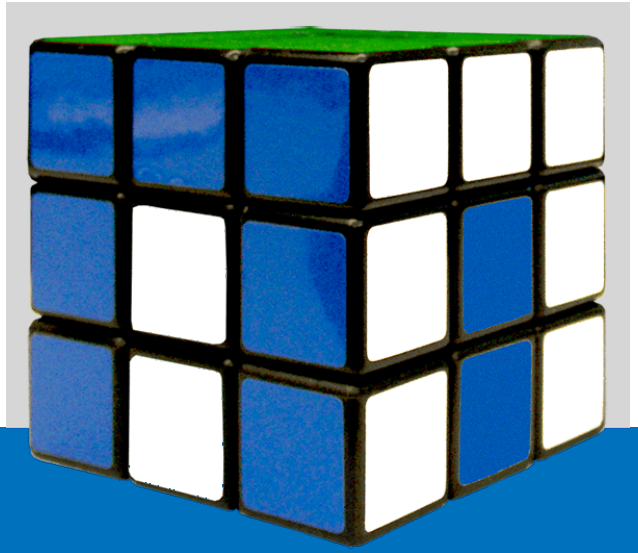
- Flash access latency similar to network hop
  - Durable data should reside at the client
    - Loose benefits when data is centrally stored
      - Efficient management of storage
      - Opportunity to single-instance data (e.g., VMDKs)
    - Resiliency to failures when sharing
    - Consistency semantics in shared environment

- Should clients do write-back or write-through to local Flash memory?

Write-back vs. Write-through

- Can network & ESS support the BW from 1000s of clients w/ write-through semantics?
  - Same is true about latency and OP throughput
- Do we need to build these pipes or can we avoid false communication?
  - Perhaps a change in application behavior
- What changes when using write-back caches?
  - ESSs provide & manage consistent view of data
  - Central authority must exist
    - … even if implemented as a distributed system

Concluding Remarks



ESS Architecture of the Future

- Migration towards two storage tiers
  - IOPS tier
  - Capacity tier
- Tighter integration of technologies
- Changing hardware/system boundaries
  - Software-managed client-side HW
- Successful architecture must work regardless of the implementation/packaging details

Academic Research Challenge

- Published works in Flash memory systems
  - Focus on a single device
    - algorithms & policies for writing/destaging
    - FTLs and file systems
  - Incremental
    - Put FLASH at the right memory hierarchy level
- Think big w/ the whole ecosystem in mind
  - Datacenter (PB+) scale w/ 1000s of clients
- Don’t be afraid to change/redefine architecture
  - Embrace bold and new approaches

Summary

- What's already here
  - Proliferation of FLASH on both end of the "wire"
  - Many packaging options exist today
    - SSD as HDD replacement with SAS/SATA i'face
    - Accelerator cards for servers & clients
    - Embedded flash memory chips
- What's next
  - Different architectures w/ more distributed data
  - The centralized management model will remain

Discussion