NetApp™

Go further, faster™

Flash memory SSDs in Enterprise Storage Systems

Jiri Schindler

Advanced Technology Group

v. 1.3

About …

- NetApp
  - maker of large-scale unified (NAS/SAN) storage systems
- Me
  - technical staff at the Advanced Technology Group
    - under the CTO office
    - explorations with 2-5 year product horizon
- Disclaimer
  - the views presented here are primarily my own; they should not be taken as the company's official views or positions on the subject matter discussed herein

© 2008 NetApp. All rights reserved.

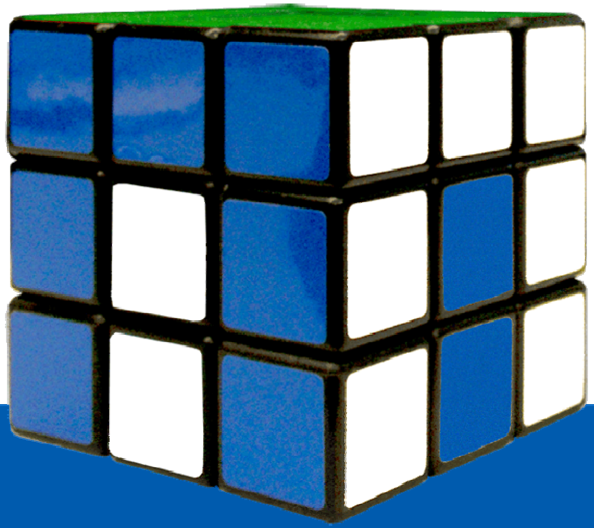
Outline

- Basics of Enterprise Storage Systems (ESS)
- Replacing disk drive with Flash-memory SSD
- Designing SSDs for ESS
- Concluding remarks & Design Challenge

© 2008 NetApp. All rights reserved.

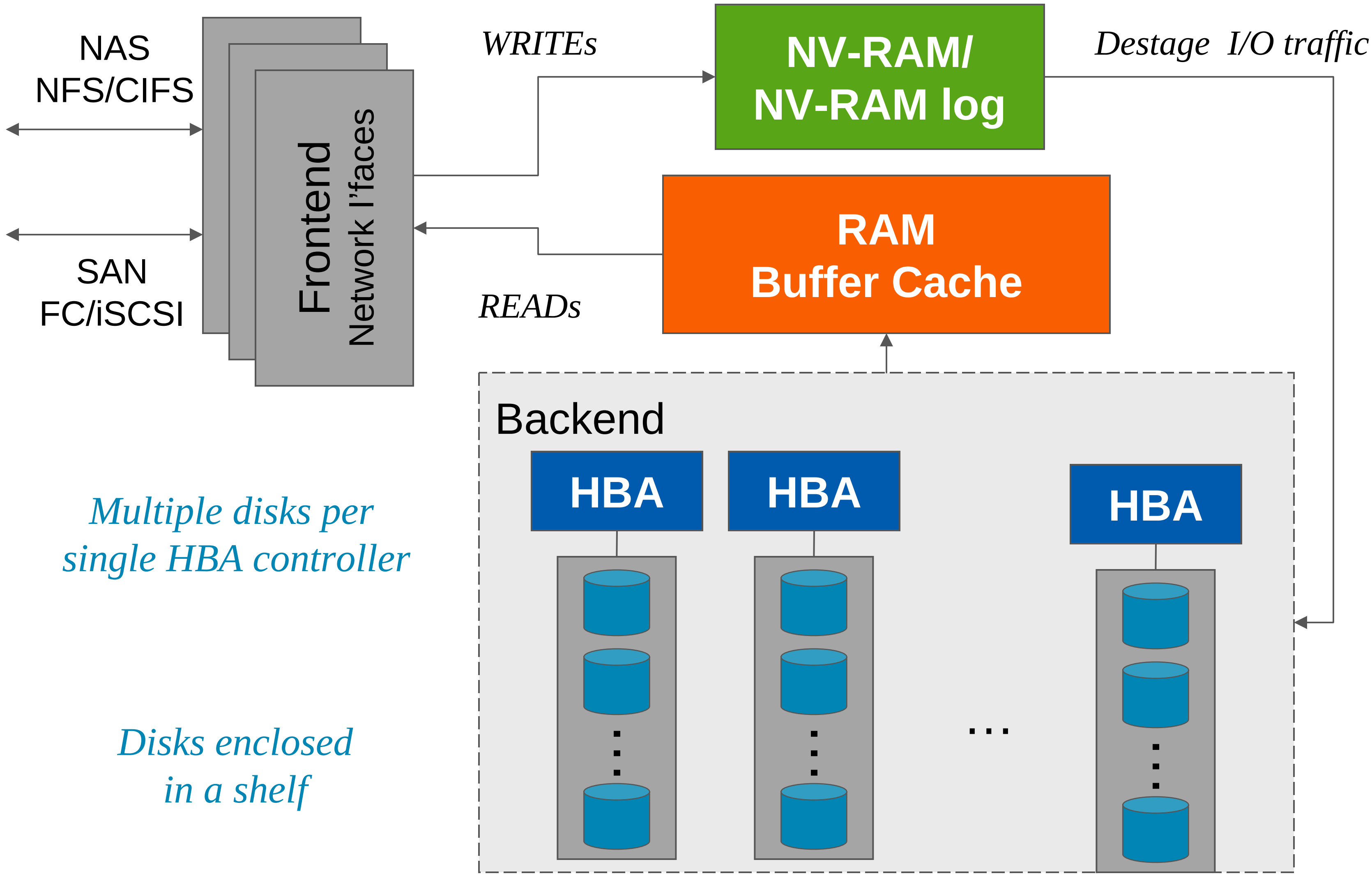
NetApp™

Enterprise Storage Systems Basics



© 2008 NetApp. All rights reserved.

Enterprise Storage Systems (ESS)



© 2008 NetApp. All rights reserved.

Enterprise vs. Desktop Environments

- Shared back-end components & interconnect
  - single HBA peaks around 15-20k IOPS
    - a single device can deliver 100-600 IOPS
    - limit the number of HDD devices
      - primary reason is fault isolation

back-end design typically balanced for HDD IOPS

  - back-end FC-AL interconnect
    - FC loop peaks at 8 Gbps ≅ 700-800 MB/s
    - a single HDD device can deliver >120 MB/s

back-end does not usually scale for aggregate BW

© 2008 NetApp. All rights reserved.

Enterprise vs. Desktop Environments

- Data durability
  - writes first end up in NV-RAM
    - different implementations & organizations
  - data sent to device must end up on the media to guarantee consistency & forward progress
    - systems typically disable HDD on-board caches
      - setup through SCSI mode pages/SATA commands
    - finer-grain control via specific commands
      - SCSI command to write-out a range of LBNs
      - SATA flush (whole) cache command
  - NV-RAM mostly solved (but expensive) problem
    - opportunities for FLASH memory

© 2008 NetApp. All rights reserved.

Enterprise vs. Desktop Environments

- Device writes exhibit different access patterns
  - bursts of high-write activity
    - NV-RAM accumulates data until a limit reached
      - capacity limits
      - write-pending limits
    - de-stage en-masse to back-end disks
  - reads and writes still intermixed
    - Read/Modify/Write for update in-place RAID
    - disk partitions/slices belong to different volumes
      - different volumes can be de-staged at different times

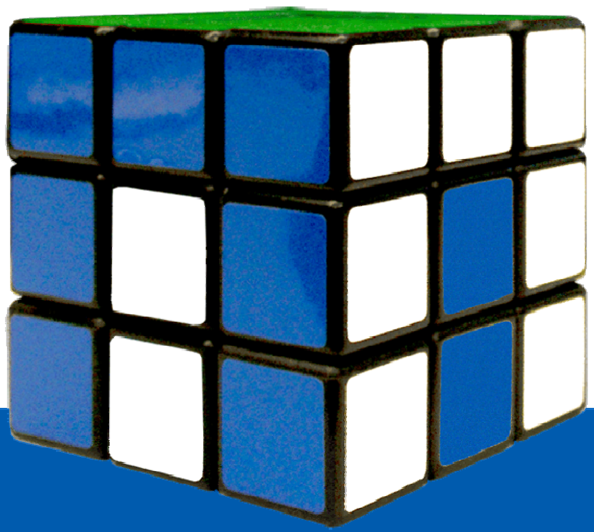
© 2008 NetApp. All rights reserved.

Additional Features

- Data scrubbing/verification
  - all blocks on the media are periodically read to verify that data is correct & device is responsive
    - SCSI command set support w/ READ VERIFY
      - media accessed, but no data xfer-ed on the bus
  - devices are usually not completely “idle”
- Storing extra info with data
  - 520 bytes per sector favored over 512
    - lost-write protection, checksum, context info …
  - T10 (SCSI Spec) DIF extensions
    - proposal to allow for end-to-end data checks

© 2008 NetApp. All rights reserved.

Replacing hard-disk drives with Flash memory-based SSDs



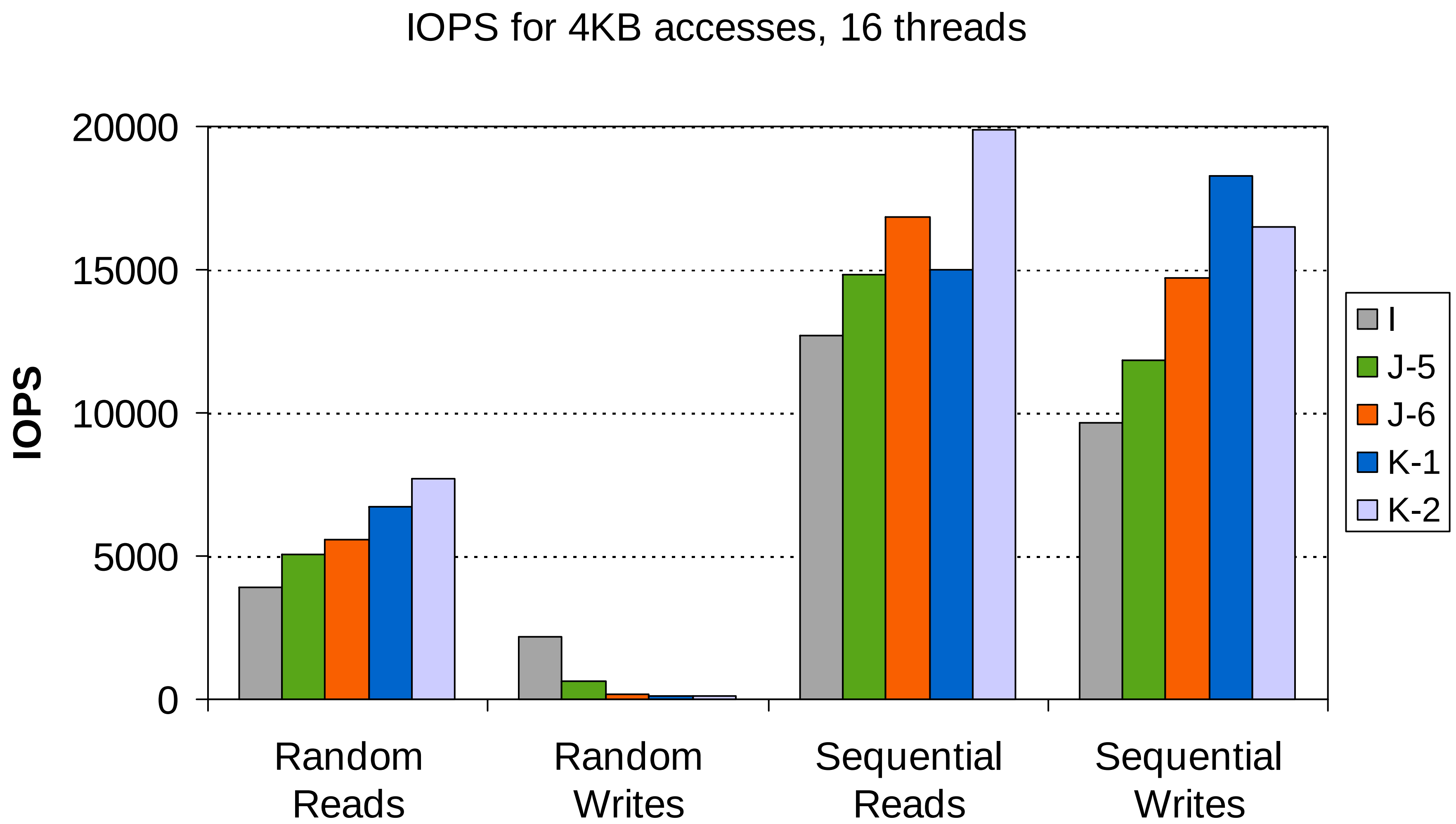
© 2008 NetApp. All rights reserved.

A Natural Fit?

- Interface and form-factor compatibility
  – SATA & FC today, SAS making in-roads
  – most storage systems today use 3.5” form-factor
  – 2.5” SFF likely to gain market-share in the near future
- Performance advantages over HDDs
  – $/(Read) IOPS, Watts/IOPS, …
- However, there are other system aspects to consider…
  – more on this later
- Comparison of cost-effective “enterprise” SSDs
  – 20-30 $/GB street-price
  – SLC NAND Flash memory, SATA interface, 2.5” SFF

© 2008 NetApp. All rights reserved.

SSD Performance Examples



© 2008 NetApp. All rights reserved.

Random 4/4.5KB Read IOPS, 16 Threads

- Recall, the additional per-block context information
  - 8 bytes for each 512 bytes
  - if 520 BPS unavailable, store context info in the 9th sector
  - other ways to reduce the wasted space exist

Device	4KB	4.5KB	Diff
I	3208	3214	0%
K-1	6723	6357	-5%
K-2	7695	7416	-4%
J-5	5056	4490	-11%
J-6	5548	4721	-15%

Non-power-of-two I/O size impacts READ performance

© 2008 NetApp. All rights reserved.

Random 4/4.5KB Write IOPS, 16 Threads

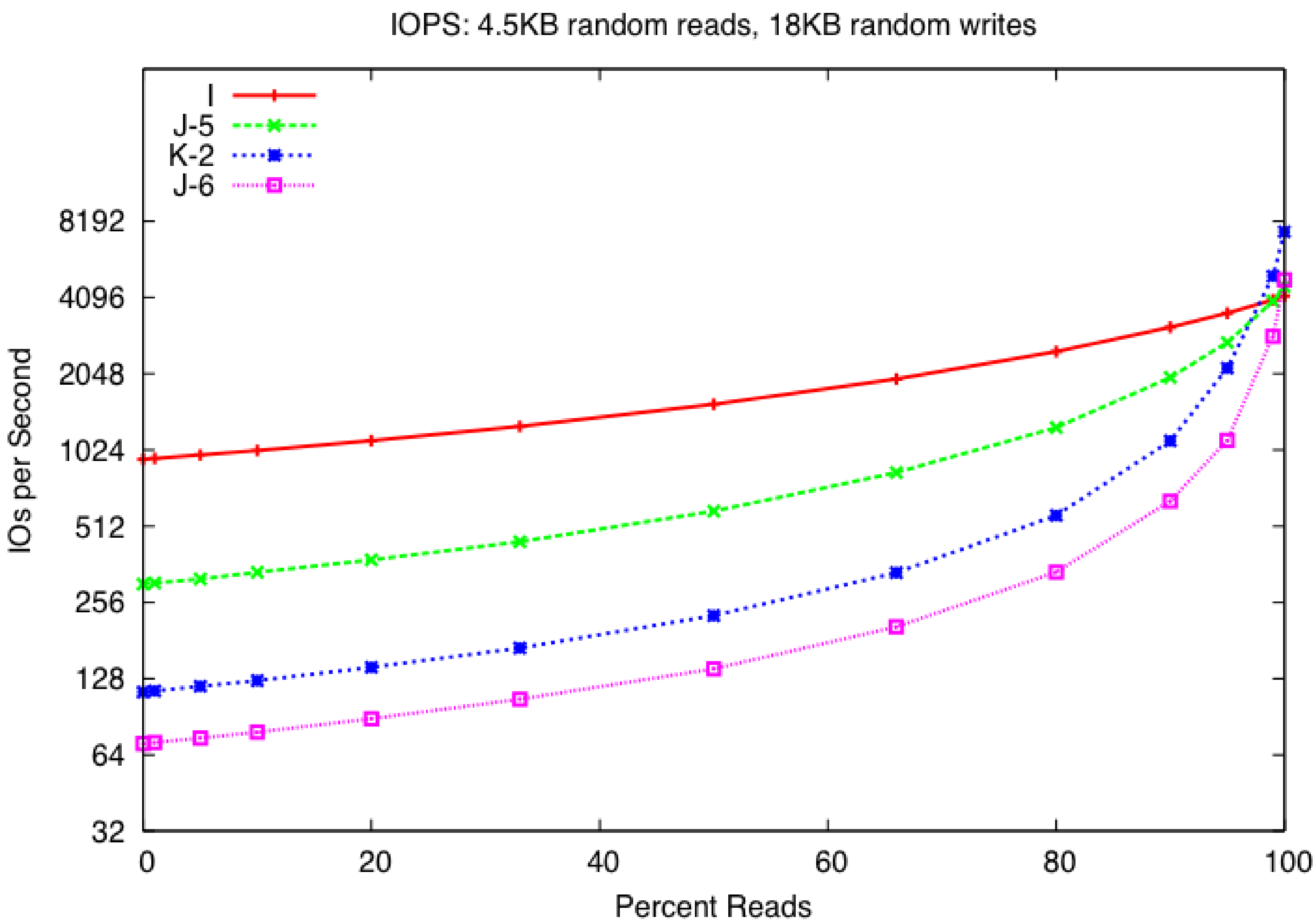
- Misaligned writes

Device	4KB	4.5KB	Diff
I	2207	2198	0%
K-1	128	127	0%
K-2	113	110	-3%
J-5	622	489	-21%
J-6	144	116	-19%

The impact on WRITEs can be bigger than for READs

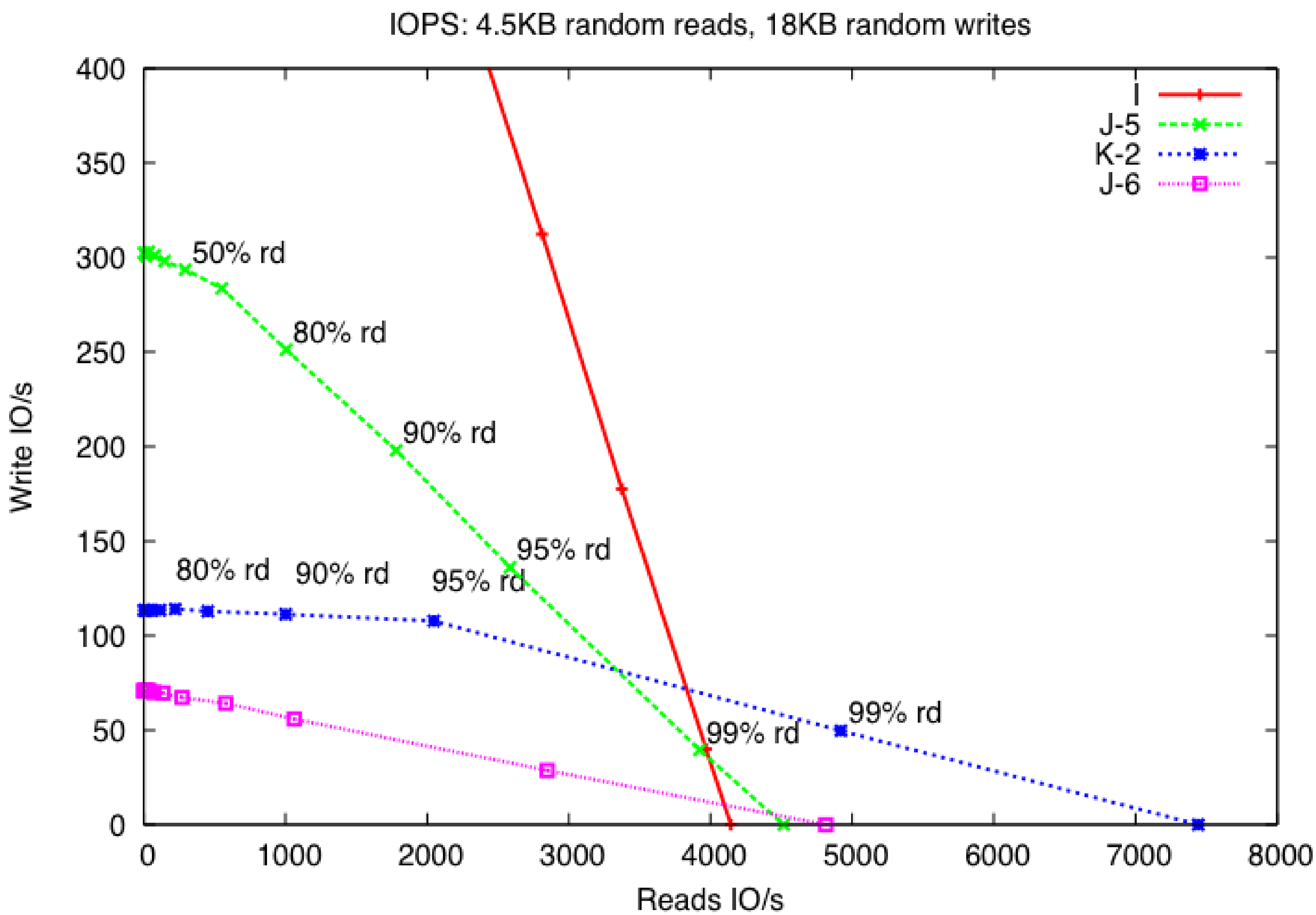
© 2008 NetApp. All rights reserved.

Reads vs. Writes Performance



© 2008 NetApp. All rights reserved.

Write Penalty



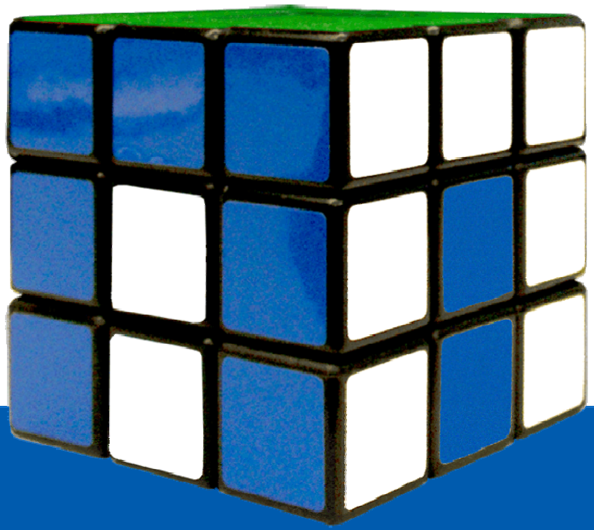
© 2008 NetApp. All rights reserved.

Summary of Observations

- Excellent Read IOPS
  - 10-100x better than HDD
- Large disparity between Read and Write IOPS
  - ratio ranges between 1.5 to ~8
- Few Writes have large impact on Read IOPS
  - pure read vs. read-mostly workloads
- Mis-alignment/non-power-of-two size
  - impact between 0% and up to 20 %

© 2008 NetApp. All rights reserved.

ESS-class SSDs



© 2008 NetApp. All rights reserved.

Readying SSDs for the ESS

- Basic premise
  - ESS HW architectures evolve slowly
    - easier to adapt individual devices
  - Flash-based SSDs are still a nascent technology
- Trade off read perf. for better write performance
  - large Read IOPS highlights interconnect limits
    - 4 SSDs can easily saturate a single HBA
  - better support for (potentially) bursty write behavior
    - faster write out of data will increase I/O throughput
- Limit impact of writes on read-mostly workloads
  - “adaptation” of Amdahl’s law

© 2008 NetApp. All rights reserved.

Readying SSDs for the ESS cont'd

- Full support of enterprise-class interconnects
  - features, not speed
  - SAS likely to displace FC as device connection
    - lower cost, switched architecture
- Ensure data stability for WRITEs
  - may design alternatives
    - no cache
    - write-through cache behavior
    - cache flush commands
    - stable cache
- Efficient request queueing support
  - ESS can keep device queues filled

© 2008 NetApp. All rights reserved.

Readying SSDs for the ESS cont'd

- Efficient handling of per-block context info
  - support for larger than 512 byte sector sizes
    - allow for more than 64 bytes per 2K page
    - expose additional bytes through the i'face
  - alternative: efficient handling of misaligned I/Os
    - if having additional per-page bytes is infeasible

- ESS systems can support larger sectors
  - HDDs want to go to 4KB native sector size
  - NetApp DataONTAP® uses 4KB sector
    - 4096 bytes of payload + 64 bytes of context info

© 2008 NetApp. All rights reserved.

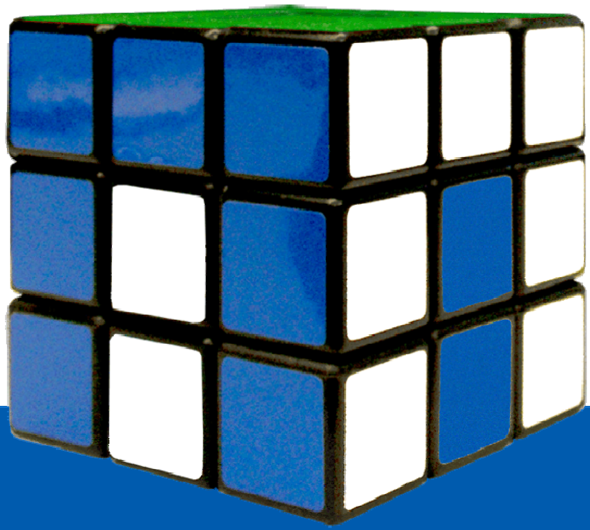
ESS Are All About Preventing Data Loss

- Device reliability
  - need to understand device failure characteristics
    - and express them in well-understood metrics
  - ESS know how to handle failures: RAID
    - the RAID-level trade-offs are well understood

- Anticipating looming device failure
  - early warning of device failure is important
    - disk S.M.A.R.T. is largely unsuccessful

© 2008 NetApp. All rights reserved.

Concluding Remarks



© 2008 NetApp. All rights reserved.

SSD Design Challenge: New FTL

- Lots of previous published work
  - mostly targeting desktops/PDAs
    - largely not applicable for a variety of reasons
- Enterprise workloads are different
  - bursty writes can streamline materializations
    - lends itself to log-structured FS organization
  - NetApp DataONTAP specific feature
    - data never overwritten
  - systems are rarely truly “idle”
- Fine-grain control over materialization
  - allows ESS systems to move forward

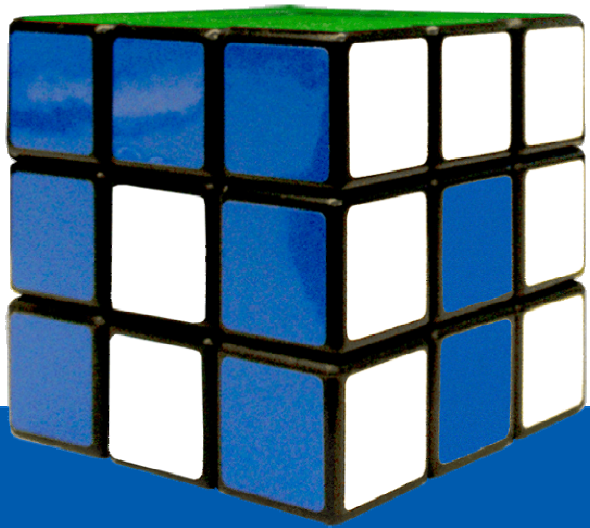
© 2008 NetApp. All rights reserved.

SSD Design Challenge: New Technology

- Power-loss protection
  - preserve non-materialized mappings
    - capacitor to power emergency writes to FLASH
    - store core info in non-volatile memory e.g., FRAM
- Address read-disturb issue & “bit rot”
  - do we need scrubbing or is it harmful?
- Provide additional per-page storage area
  - exposed through storage interface
- Expose information about device health
- Full support of storage protocol features

© 2008 NetApp. All rights reserved.

Discussion



© 2008 NetApp. All rights reserved.