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
Outline

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

NAS
NFS/CIFS

SAN
FC/iSCSI

Frontend Network Interfaces

NV-RAM/ NV-RAM log

RAM Buffer Cache

Destage I/O traffic

Multiple disks per single HBA controller

Disks enclosed in a shelf
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
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
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
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
Replacing hard-disk drives with Flash memory-based SSDs
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
SSD Performance Examples

IOPS for 4KB accesses, 16 threads

IOPS

Random Reads
Random Writes
Sequential Reads
Sequential Writes

© 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

<table>
<thead>
<tr>
<th>Device</th>
<th>4KB</th>
<th>4.5KB</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>3208</td>
<td>3214</td>
<td>0%</td>
</tr>
<tr>
<td>K-1</td>
<td>6723</td>
<td>6357</td>
<td>-5%</td>
</tr>
<tr>
<td>K-2</td>
<td>7695</td>
<td>7416</td>
<td>-4%</td>
</tr>
<tr>
<td>J-5</td>
<td>5056</td>
<td>4490</td>
<td>-11%</td>
</tr>
<tr>
<td>J-6</td>
<td>5548</td>
<td>4721</td>
<td>-15%</td>
</tr>
</tbody>
</table>

Non-power-of-two I/O size impacts READ performance
Random 4/4.5KB Write IOPS, 16 Threads

- Misaligned writes

<table>
<thead>
<tr>
<th>Device</th>
<th>4KB</th>
<th>4.5KB</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>l</td>
<td>2207</td>
<td>2198</td>
<td>0%</td>
</tr>
<tr>
<td>K-1</td>
<td>128</td>
<td>127</td>
<td>0%</td>
</tr>
<tr>
<td>K-2</td>
<td>113</td>
<td>110</td>
<td>-3%</td>
</tr>
<tr>
<td>J-5</td>
<td>622</td>
<td>489</td>
<td>-21%</td>
</tr>
<tr>
<td>J-6</td>
<td>144</td>
<td>116</td>
<td>-19%</td>
</tr>
</tbody>
</table>

The impact on WRITEs can be bigger than for READs
Reads vs. Writes Performance

IOPS: 4.5KB random reads, 18KB random writes

Percent Reads

IOs per Second

0 32 64 128 256 512 1024 2048 4096 8192

0 20 40 60 80 100

I
J-5
J-6
K-2
Write Penalty

IOPS: 4.5KB random reads, 18KB random writes

- 50% rd
- 80% rd
- 90% rd
- 95% rd
- 99% rd
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%
ESS-class SSDs
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
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 queuing support
  - ESS can keep device queues filled
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
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
Concluding Remarks
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
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
Discussion