Performance Analysis of Commodity and Enterprise Class Flash Devices

Neal M. Master, Matthew Andrews, Jason Hick, Shane Canon & Nicholas J. Wright
Data Trends at NERSC cont.

HPSS Total Number of Tapes

Date
8/1/09 9/1/09 10/1/09 11/1/09 12/1/09 1/1/10 2/1/10 3/1/10 4/1/10 5/1/10 6/1/10 7/1/10 8/1/10

Total Number of Tapes
0 5000 10000 15000 20000 25000 30000 35000 40000 45000

- 9840D
- T10KB
- T10KA
- 9940B
- 9840A
Memory Capacity Trends

- Technology trends:
  - Memory density 2X every 3 yrs; processor logic every 2
  - Storage costs ($/MB) drops more gradually than logic costs

![Evolution of memory density](image1)

![Cost of Computation vs. Memory](image2)

The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it.
I/O Performance Challenges

Performance Crisis #1
• Disks are outpaced by compute speed.
• To achieve reasonable aggregate bandwidth many spindles needed – $10^3$ spindles = 1PB but only ~0.1 TB/s!

Performance Crisis #2
Data Motion on an Exascale Machine
will be expensive – both in terms of energy & performance!
Flash – What is it good for?

• Fits nicely into latency gap between spinning disk and memory

• Lots of open Q’s:
  – PCI vs SATA vs ?
  – SLC vs MLC
  – Write requires block erase - performance dependent upon previous IO pattern
  – Correct algorithm in software at all levels
  – …. 
Devices Evaluated

- **3 PCI-e SLC**
  - Virident tachlOn 400GB 8x
  - FusionIO ioDrive Duo 2x 160GB 4x
  - Texas Memory Systems RamSan-20 450GB 4x

- **2 SATA MLC**
  - Intel X-25M 160GB
  - OCZ Colossus 250GB
IOZone Experiments

• Bandwidth
  – Vary block size: \(2^n\) KB, \(n = 2-8\)
  – Vary concurrency: \(2^n\) threads, \(n=0-7\) (1-128)
  – Vary IO Patterns: Sequential Write/Re-write, Sequential Read/Re-read, Random Write, Mixed Random Write/Read, Random Read

• IOPS
  – 4KB block size
  – Vary concurrency: \(2^n\) threads, \(n=0-7\) (1-128)
SATA Bandwidths

INTEL X25-M READ

Bandwidth (MB/s)

IO Block Size (KB)

Number of Threads
PCI-Bandwidths continued

Virident tachIOn READ

Number of Threads

IO Block Size (KB)

Bandwidth (MB/s)
Bandwidth Summary

- TMS RamSan 20 (450GB)
- Virident tachIOn (400GB)
- Fusion IO ioDrive Duo (Single Slot, 160GB)
- Intel X-25M (160GB)
- OCZ Colossus (250 GB)
IOPS - READ

- Virident tachIOn (400GB)
- TMS RamSan 20 (450GB)
- Fusion IO ioDrive Duo (Single Slot, 160GB)
- Intel X-25M (160GB)
- OCZ Colossus (250 GB)
Flash Device Evaluation - IOPS

- TMS RamSan 20 (450GB)
- Virident tachION (400GB)
- Fusion IO ioDrive Duo (Single Slot, 160GB)
- Intel X-25M (160GB)
- OCZ Colossus (250 GB)

Thousands IOPs

Peak Read  Peak Write
Degradation Experiment

- Create a file using
  - `Cat /dev/urandom | dd`
  - that fills X% of the drive X=30,50,70,90
- Using FIO randomly write to the file
  - Using 4KB blocks - IOPS
  - Using 128KB blocks - BW
Degradation - IOPS

- Virident tachlOn (400GB)
- TMS RamSan 20 (450GB)
- Fusion IO ioDrive Duo (Single Slot, 160GB)
- Intel X-25M (160GB)
- OCZ Colossus (250 GB)
Degradation – IOPS Summary

<table>
<thead>
<tr>
<th>Product</th>
<th>Capacity (GB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virident tachION</td>
<td>400 GB</td>
</tr>
<tr>
<td>TMS RamSan 20</td>
<td>450 GB</td>
</tr>
<tr>
<td>Fusion IO ioDrive Duo</td>
<td>Single Slot, 160 GB</td>
</tr>
<tr>
<td>Intel X-25M</td>
<td>160 GB</td>
</tr>
<tr>
<td>OCZ Colossus</td>
<td>250 GB</td>
</tr>
</tbody>
</table>

Percentage of Peak Write IO/s:
- 30%
- 50%
- 70%
- 90%
Degradation - Bandwidth

- Virident tachIOn (400GB)
- TMS RamSan 20 (450GB)
- Fusion IO ioDrive Duo (Single Slot, 160GB)
- Intel X-25M (160GB)
- OCZ Colossus (250 GB)
Degradation BW Summary

Percentage of Peak Write Bandwidth

- 30% Capacity
- 50% Capacity
- 70% Capacity
- 90% Capacity

<table>
<thead>
<tr>
<th>Storage Device</th>
<th>Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virident tachIOn (400GB)</td>
<td></td>
</tr>
<tr>
<td>TMS RamSan 20 (450GB)</td>
<td></td>
</tr>
<tr>
<td>Fusion IO ioDrive Duo (Single Slot, 160GB)</td>
<td></td>
</tr>
<tr>
<td>Intel X-25M (160GB)</td>
<td></td>
</tr>
<tr>
<td>OCZ Colossus (250 GB)</td>
<td></td>
</tr>
</tbody>
</table>
Summary

- PCI devices are much more capable than the SATA ones
- For PCI read ~ write for both sequential I/O and IOPS
- It is important to test for your workload each device
- The PCI devices especially can be difficult to use……
Future Work

• Testing Flash with Hadoop
• Evaluating various new storage technologies. PCM etc
• Explore other uses for flash
  – Metadata storage
Combining Flash with Hadoop

![Bar Chart]

- **Number of concurrent writers**: 1, 4, 8, 25
- **Throughput MB/s**
- **Comparisons**:
  - 10MB
  - 10GB
  - 10MB-SSD
  - 10GB-SSD