Innovative Solid State Storage Architecture

Michael Cornwell
Lead Technologist - Flash Memory
First Solid State Disk
1978

- StorageTek 4305, Solid State Disk
- 45MB, .6ms access time
- $8,800,000/GB
Flash Back

.. Two years ago

Non-volatile Shared Memory Systems

PCI-E Memory Arrays

SSD Aware Storage Arrays

NV-Memory Integration HSM

HDD Replacement

DIMMs

NVM-Boot

Systems

Storage

DIMMs

DIMMs

HDD Replacement

HDD Replacement

Storage Developer Conference 2009
Inside the HDD

- **Zoned Performance**
  - Outside Diameter is fastest, Inside 30% less

- **Static Addressing**
  - First block outside, Last block inside

- **Bad Block Management**
  - Maps out bad blocks (particles, etc.)

- **Wear-leveling**
  - Prevent user from harming devices
  - “Adjacent Track Interference”

- **RPM makes difference**
  - Rotational latency (4ms-16ms)

- **Single I/O Processing**
  - Single actuator per drive

- **Command Queuing**
  - Servo Seek Optimization
Building Blocks of Storage Systems

- Goal to mask spindle performance with DRAM and aggregation
Shrinking the IOP

**Hard Disk Assembly**
- 110MB/sec Read/Write
- 350 IOPS Read/Write
- 10 Watts
- 7502 mm²

**NAND Flash Die**
- 40MB/sec Read/Write
- 7000 IOPs Read/146 IOPs Write
- 50 mWatts
- 165 mm²
Solid State Disk

- **Asymmetric Performance**
  - Write Speed Slower than Read
  - Write chunk larger than Read

- **Dynamic Addressing**
  - Internal filesystem continually moves data

- **Bad Block Management**
  - Maps out bad blocks (bit disturbances, die failures)

- **Wear-leveling**
  - Prevent user from harming devices
  - “Cell burn out”

- **Parallel I/O Processing**
  - Multiple channels to NAND devices

- **Command Queuing**
  - Keep channels fully utilized
Lessons of CPU Parallelization

- Requires Re-write Of Application
  
  > Multithreaded Programming
  
  > Not all Applications are

- Speed still matters for Single threaded Legacy Applications

- Virtualization doesn’t solve Everything
  
  > Difficultly in Management

- Adoption Slow
Death by Queue

- Application Requests can get hopelessly stuck in queues
- Up 7,000 to 10,000 I/Os Per device
• Benefits of Latency lost in threaded workloads
Parallel Storage Processing

- Segmentation Device Queues
- Significant Latency Reduction
- Easier implementation then port virtualization
Data Alignment

<table>
<thead>
<tr>
<th>LBA</th>
<th>PBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>512B</td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

**Emulated 4K**
- Device still reports 512B sectors to the host.
- Physical size remains 4096B
- Maintains legacy performance at cost of performance
- Even or Old Alignment

<table>
<thead>
<tr>
<th>LBA</th>
<th>PBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4096B</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
</tbody>
</table>

**Native 4K**
- Device still reports 4096B sectors to the host.
- Physical size remains 4096B
- No backwards compatibility
Read/Modify/Write Penalty

**Aligned Writes**

Host

Write 4096B

Target

Write 4096B

**Un-Aligned Writes**

Host

Write 512B

Target

Store 512B

1

2

3 Modify Data

4

Write 4096B

Read 4096B
Open Flash Module
Next Generation in Server Storage

• First Server SSD designed for dense computing applications
• Minimum Size for Maximum Performance
• Leverages Existing Interfaces SATA/SAS
  > Reserved PCI-e in Pinout
• Released as an Open Standard to the Industry - JEDEC
• Supports optional backup power system
Evolution of Server-based Storage

- **2000**: 3.5” 15K HDD, 16 Watts
- **2004**: 2.5” 10K HDD, 8 Watts
- **2009**: Open Flash Module, 2 Watts
Modern Server Design
Pushing Thermal Limits to the Extreme

Airflow

CPUs

Fans

DRAM

I/O Expansion
New Server Architecture
Flash Integration

- Flash inside for storage performance
- HDD inside for storage capacity
NVM Host Controller Interface

The Standard for PCI-e Flash Connectivity

• Non-Volatile Memory Host Controller Interface (NVMHCI)
• Fresh take on interface for SSDs and caches
• Significantly lower latency than legacy disk interfaces
• Improved Queue Prioritization
• 40+ Member Companies
• Initial Revision Completed April 2008
NVMHCI - Enterprise Extensions
Data Center Features to PCI-e Flash

• True Replacement for SAS/Fiber Channel for Solid State
• Hot-plug Support
• Multi-path Interface
  > Active/Active failover
• Port Virtualization
• End to End Data Integrity
NVMHCI - Enterprise Extensions
Chassis Based Shared Memory

- PCI-e Interconnect
  - Multi-path
  - I/O Virtualization
- Tiered Memory Architecture
  - 100’s GB of Cache
  - 10’s TB of NAND
- Shared Programming Model exists Today
  - memcached, OpenMP, VMCI
- Based on Open Standards
Storage Programming

malloc()

open(), read(), write()

Will Solid State Storage ever move to a true memory model?
Non-Volatile DRAM
NAND backed DRAM Solutions

- Leverages Performance of DRAM with Non-volatility of NAND
- Attaches to existing high speed, low latency DRAM interfaces
- Applications Today
  > Industrial / Military Applications
  > RAID / Storage Controllers
- Requires software awareness of non-volatility
- Truly Tier System Memory Controllers
- Intermix Volatile and Non-Volatile DRAM
In Closing....

“Every Microsecond is Sacred”
Thank you

Michael Cornwell
michael.cornwell@sun.com