HPC @ Speed of Memory

Violin Memory – Redefining Storage Economics

Martin Coleman
Sales Engineer
mcoleman@vmem.com
www.violin-memory.com
High Performance Computing...

- Generally accepted as:
  - “... the aggregation of computing to deliver much higher performance than from a single compute node, in order to solve large problems in science, medicine, oil & gas exploration, engineering or business ...”
  - Specialised nodes, filesystems, metadata, interconnects (typically Infiniband)

- Storage: *latency* is king, not IOPS...
  - Capacity is abundant. Performance is not. This matters.

Storage is holding you back!
Spinning Disk: Why it’s a Problem…

- Mechanical Magnetic Storage Platters
  - Rotating at 7200, 10000 or 15000 RPM
- Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead
  - Seek time: 3.5ms – 7.5ms
  - Rotational Latency: 2ms
  - Typically >5ms
  - Then there’s fragmentation, and random IO
- What is the CPU doing while the disk is positioning heads to retrieve data?
  - Waiting…
    - Measured as “IO Wait,” and shows “Idle”
- Result:
  - Low CPU Utilisation
  - Long running jobs
  - Poorly utilised infrastructure
Traditional Storage Speed Lags CPU

- Moore’s Law - Processor and Network Performance/Price double every 18 months

- Disk capacity exceeds Moore’s Law

- Disk performance has not kept up causing under-utilization of CPU
  - Still rotating at only 15k RPM
  - Still < 200 random IOPs per spindle @ >5ms

- Flash performance has exceeded Moore’s Law
  - No mechanical components
  - Violin Plays Requiem of the dying disk drive
  - “The aggressive trend of the shrinking process design rule or technology node in NAND flash memory technology effectively accelerates Moore’s Law.”

### Graph

- Improvement (normalized such that 2000 = 1)
- Processor performance (Moore’s Law)
- Flash
- HDD Performance
- Cost per GB is becoming Cost per IOP

### Costs

- Cost per GB
- Cost per IOP

Transaction Wait
How Do You Make Storage Go FAST?

- Add expensive DRAM to legacy array
- Short stroking – more spindles
- Wide striping – more ports
- Add SSD to legacy array
- ‘Read-Only’ flash cache
- “FAST”
- “Easy Tier”

High Acquisition Costs
Higher Operational Costs
Enter Flash Memory…

- Advantages:
  - Extremely fast (15usec or less)
  - No mechanical components – not susceptible to vibration
  - Compact (Violin now using 19nm die) – unsurpassed density vs areal magnetic media

- Limitations:
  - Block erasure
  - Limited operations (read/write/erase)
    - Garbage Collection blocks Reads and Writes
  - Read disturb
  - Memory wear

- Flash is also used in Commodity SSD Drives – Micron, Intel, STEC, OCZ etc
  - SSD Write Cliffs…
  - Commodity SSD does not handle power failure very well
    - All too common for SSD to lose data in power failure tests

- Choose your flash wisely – ground up design vs commodity…
Enabling HPC @ Speed of Memory, vs Bound by Disk

CPU Cycle with Magnetic Disk:
- I/O Wait: 80% Wait, 20% Work
- t

CPU Cycle with Memory Storage:
- I/O Wait: 5% Wait, 95% Work
- t

Storage @ the Speed of Memory
Close the gap between CPU and Storage performance

Eliminate Latency ➔ More Work in the Same Time
Brief History of Flash

- In 1987, Toshiba invented NAND Flash
  - IEEE paper originally published in 1984

- What is Flash Memory?
  - A non-volatile computer memory that can be electrically erased and reprogrammed. A specific type of EEPROM, erased and programmed in large blocks
  - Flash memory is non-volatile, no power is needed to maintain the information stored in the chip.

- Evolving modern life:
  - Music: Vinyl -> Audio Cassette Tape -> Spinning Disk iPod -> Flash based iPod
  - Portable Storage: 5.25” floppy disks -> 3.5” floppy disks -> USB Thumb Drives

- We no longer use 35mm film for images – Kodak?
  - Mobile phones with 15MP cameras

- Smart phones: we store our lives on flash memory today
  - Calendar, contacts, email, pictures, web pages, documents
Architecture Matters - Flash Memory vs. SSD

Violin Memory Flash Array

- Built from ground-up
- Engineered for flash
- Memory-like performance
- Latency in microseconds

Everyone Else

- Legacy architecture
- SSD instead of HDD
- Disk-plus performance
- Latency still in milliseconds
Flash vs SSD vs Disk

- **3 Operations Permitted on Flash:**
  - Read – very fast, 15μsec (that’s 0.015ms)
  - Write – slower @ <1.5ms – but only to previously unwritten cells
  - Erase – very slow @ <5ms – garbage collection/grooming/”trim” (limited erase cycles)
    - Blocks all other operations leading to “write cliffs” once array is >60% written

- **Traditional Raid Unsuitable**
  - Use of distributed parity to overcome individual spindle hot-spots – not applicable to Flash
  - “Read-modify-Write” – would incur significant penalty due to erase cycles and slower writes
  - Legacy array with SSD is still slow...

- **Violin patented innovations: vRaid, VIMM’s, Switched Memory Fabric**
  - Ensure non-blocking reads and writes, even wear leveling and extended Flash life
    - No read or write operation will be blocked by an erase
  - Massive distributed parallel processing across patented Flash Translation Layers on each VIMM
  - Spike free latency – no write cliff

- **Violin shipping 5\textsuperscript{th} generation Flash**
Write Cliff Affects All Flash Solutions To Some Degree

- “… the effect where SSD performance drops off after all free Flash cells have been initially written to and the controller cannot provide enough free blocks to keep up with write requests…”
  - Up to 80% performance drop

- IO queued behind Erase operations (Garbage Collection)

- Real issue is that Erase operations also get in the way of Read operations

- Mitigating or eliminating the Write Cliff requires special flash management logic

 transient random write bandwidth degradation

Source: Nersc
How do Commodity SSD’s Try Delay the Write Cliff?

- Aka Write Amplification
- Commodity SSD controls garbage collection
  - Not the storage array
  - Array vendor dependent on middleman – limits control, increases cost
- Storage Array attempts to delay write cliff by:
  - Striping wide across all SSD’s
  - Short-provisioning the SSD – eg fixed 70% format rates, dictated by SSD drive vendor, not the array vendor…
- SSD Vendors use same old spinning disk legacy techniques
- Violin vRAID and VMOS control the garbage collection
  - Innovation from the ground up

The Violin Innovation Advantage

- Technological innovation at every layer from Hardware to Software
  - Intellectual Property (IP) aggregation resulting in a fundamentally unique solution
- Deep software and hardware integration
  - Toshiba partnership
  - Violin Switched memory architecture
  - vMOS™ - Violin Memory Operating System optimized for flash
  - vRAID™ - Flash optimized RAID
  - Four-level system architecture

No Spinning Disk
No SSD
- No Middle Man

Violin V6000

Raw High Performance Flash

Toshiba Flash

VIMM SLC/MLC 256GB/512GB/1TB

24/44/64 Groups

4/8/12 Groups

Up to 64TB in 3U
vMOS vRAID vs SSD in Legacy Array: Effect of Control of Garbage Collection

Latency vs. Time
10% Load

- Other SSDx4 RAID 0
- Other SSD - No RAID
- Violin Flash RAID

Latency vs. Time
90% Load

- Other SSDx4 RAID 0
- Other SSD - No RAID
- Violin Flash RAID

- Erase Spikes
- Software RAID Processing
- Software Striping
- Hardware Striping
- Blocking Erase Spikes
- Non-blocking Erases
## Violin Memory 6000 Series Models

<table>
<thead>
<tr>
<th></th>
<th>6212</th>
<th>6222</th>
<th>6232</th>
<th>6264</th>
<th>6606</th>
<th>6611</th>
<th>6616</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Form factor</strong></td>
<td>3U</td>
<td>3U</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Flash type</strong></td>
<td>Capacity (MLC)</td>
<td>Performance (SLC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Raw Capacity (TiB)</strong></td>
<td>12TiB</td>
<td>22TiB</td>
<td>32TiB</td>
<td>64TiB</td>
<td>6TiB</td>
<td>11TiB</td>
<td>16TiB</td>
</tr>
<tr>
<td><strong>Raw Capacity (TB)</strong></td>
<td>13.2TB</td>
<td>24.2TB</td>
<td>35.2TB</td>
<td>70.3TB</td>
<td>6.6TB</td>
<td>12.1TB</td>
<td>17.6TB</td>
</tr>
<tr>
<td><strong>I/O Connectivity</strong></td>
<td>8Gb FC, 10GbE iSCSI, 40 Gb IB, PCIe G2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Maximum 4KB IOPS (Mixed)</strong></td>
<td>200K IOPS</td>
<td>350K IOPS</td>
<td>500K IOPS</td>
<td>750K IOPS</td>
<td>450K IOPS</td>
<td>800K IOPS</td>
<td>1M IOPS</td>
</tr>
<tr>
<td><strong>Maximum Bandwidth (100% Reads)</strong></td>
<td>1.5GB/s</td>
<td>2.5GB/s</td>
<td>4GB/s</td>
<td>4GB/s</td>
<td>3GB/s</td>
<td>3.5GB/s</td>
<td>4GB/s</td>
</tr>
<tr>
<td><strong>Nominal Latency</strong></td>
<td>500 µsec (mixed)</td>
<td>250 µsec (mixed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Latency Comparison

- No seek times
- Non-volatile
- Extreme Performance

- 8,000µs (32 times the performance)
- 3,000µs (12 times the performance)
What this means to the HPC Community

- Research runs faster
  - Enables the drive to real time, interactive processing
  - More instances and deeper analytics can be run in the same time

- High concurrent loads can be run on the same data
The Storage of Choice for Performance Records

2010  2011  2012

- 6/21/10: TPC-E World Record (HP, SQL Server)
- 5/9/11: TPC-C World Record (HP, SQL Server)
- 5/23/11: TPC-C World Record (HP, SQL Server)
- 6/22/11: File System World Record (IBM)
- 12/8/11: TPC-C World Record (Oracle, CISCO)
- 9/11/12: VMmark 2.1 World Record (VMware, CISCO)
- 9/27/12: VMmark 2.1 World Record (VMware, Fujitsu)
- 10/2/12: VMmark 2.1 World Record (VMware, CISCO)
- 11/13/12: VMmark 2.1 World Record (VMware, Dell)
- 12/25/12: VMmark 2.1 World Record (VMware, Fujitsu)
IBM Smashes GPFS World Record by 37x

- Set using Violin Flash Storage, 2011
  - Scanned 10 Billion Files in 43 Minutes, Setting a New Standard for Big Data Applications
  - By using a small cluster of ten IBM xSeries servers, IBM's cluster file system (GPFS), and by placing file system metadata on a new solid-state storage appliance from Violin Memory, IBM Research demonstrated, for the first time, the ability to do policy-guided storage management (daily tasks such as file selection for backup, migration, etc.) for a 10-billion-file environment in 43 minutes. This new record shatters previous record by factor of 37.

- Tests used older generation Violin array with PCIe connect to X3650
  - Violin V6000 provides native Infiniband connection to server network
  - Use GPFS storage management for metadata placement
MetaData Latency is a Killer in HPC

- General file operations such as create, open, read, etc. require metadata lookups
- Typically metadata is 10-15% of main storage capacity
  - E.g. storage 250TB short-stroked SAS spindles, metadata can be ~20TB
  - Meta data typically on short-stroked commodity SSD shelves in a legacy storage cabinet
Accelerate Your Metadata - @ Speed of Memory…

- Metadata LUNs are presented to each node, or storage node (NSD), or dedicated metadata nodes, via Infiniband or FC SAN, depending on HPC topology
  - These LUNs are ideal for hosting on 3U Violin Memory arrays
- Use built-in GPFS or Lustre data management to identify the Violin LUNs as metadata stores and migrate metadata to Violin LUNs
- Accelerate your ingest (faster metadata index updates)
- Accelerate your analysis (faster metadata search)
Simple Management Operations

- **Provision storage and Go!**
  - Select LUN capacity and let vRAID automate placement
  - No tuning required
  - Hot swap for non disruptive operations

- **Seamlessly handle performance spikes**
  - Customer example:
    - Rogue full table scans in dba scripts
    - System handled the load spikes and still met core application SLAs

- **Advanced Graphical User Interface**
  - Fully customizable dashboard
  - Detailed performance statistics
  - Supported as a vCenter Plug-In
Violin Symphony: Manage PB’s @ Speed of Memory

- Manage 100’s of Violin flash Memory arrays through a single interface
- Enable multi-tenancy with role based access control and Smart Groups
- Share information through custom reports with up to 2 years of historic data
- Achieve pro-active wellness with advanced health & SLA monitoring
- Personalize visibility through fully customizable dashboards and gadgets
Transition from spinning to solid state storage already underway.

– STEVE O’DONNELL, ESG