Unblocking The I/O Bottleneck
RAID Algorithms for Flash Memory

(Making Flash Enterprise-grade and Cost-effective)

Morgan Littlewood

VP Product Management

Violin Memory, Inc

Mountain View, CA

littlewo@vmem.com
Evolving Flash Storage

PURPOSE-BUILT Memory Arrays

• Shared storage with integrated RAID
• Sustained R/W throughput
• 7x24x365 operation

1st Generation - SSDs

2nd Generation - PCIe cards

3rd Generation

• Workstation/Gaming
• Memory extension/cache
• 60+ vendors by year end

Direct drive replacement
Cost sensitive
100s of vendors
“Eventually virtualization will play a different role and completely disaggregate the server. Instead of having a physical box with storage, CPU, memory, etc. built into it, the virtualization will allow for the server to be made up of virtual components.”

Zeus Kerravala, SVP of Enterprise Research at Yankee Group
Violin 3000 Series

Flash Memory Arrays for the Next Gen Data Center

Flash VIMMs
- 10TB+ Density in 3U
- SLC, MLC and DRAM VIMMs
- Sustained Write IOPS
- Hot-swap capability

Flash RAID
- Spike-Free latency
- 80% Flash Efficient
  - vs. 50% for RAID-1
- Hot swap & Fail-in-place
- 99.999% Availability

Flash Networking
- Sub 100µsec latency
- Multi-host sharing
- PCIe x4/x8, 8 Gbit/s FC
- 10GbE: iSCSI & FCoE
Violin Flash Aggregation

Density & Efficiency reduces capital cost, space, power and operations costs. 50% lower TCO with 80% lower latency!

Violin Flash arrays are similar density & cost to 15Krpm HDD storage systems!
Why is RAID Needed?

- RAID = Redundant Array of Inexpensive Devices
  - Device failures happen
  - Devices need to be replaced

- Enterprise-grade = Real business
  - Data loss costs money (and jobs)
  - Down-time costs money (and customers)
  - IT Staff time cost money

- RAID side benefits
  - Increased bandwidth (RAID-0 is really AID-0)
  - Simpler software
RAID for HDDs

- Each Write is mirrored
- 50% space efficient
  - All data is mirrored
- 50% Bandwidth
- Preferred for writes

RAID 1

Disk 0

A1
A2
A3
A4

Disk 1

A1
A2
A3
A4

RAID 5

Disk 0

A1
B1
C1
Dp

Disk 1

A2
B2
Cp
D1

Disk 2

A3
Bp
C2
D2

Disk 3

Ap
B3
C3
D3

- Write = Read-Modify-Write
  - 2 Reads + 2 Writes
  - Data & Parity
- 80+% Space Efficient
- 30-70% Bandwidth
  - Poor for writes
## RAID Issues for Flash Memory

<table>
<thead>
<tr>
<th>Flash Issues</th>
<th>Consideration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erase Blocking</td>
<td>Erase times of 2 - 10ms - Reads are blocked</td>
</tr>
<tr>
<td>Slow Writes</td>
<td>Writes are slower than Reads</td>
</tr>
<tr>
<td>Write Amplification</td>
<td>Random writes require garbage collection and have write amplification. Slows the whole array down.</td>
</tr>
<tr>
<td>RAID Latency</td>
<td>Application latency is important. Needs to be competitive with cache latency.</td>
</tr>
<tr>
<td>Reliability</td>
<td>Thousands of devices per array &amp; hence much lower MTBF than HDD arrays.</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Flash GBs are more expensive than HDD GBs.</td>
</tr>
</tbody>
</table>
RAID 1 for Flash

- Double Flash costs
- 50% lower capacity per system
- 50% lower IOPS per system
- Latency is impacted by Erases and Writes

### Flash Issues vs RAID 1

<table>
<thead>
<tr>
<th>Flash Issues</th>
<th>RAID 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erase Blocking</td>
<td>Same blocking as single Flash SSD</td>
</tr>
<tr>
<td>Slow Writes</td>
<td>2x writes</td>
</tr>
<tr>
<td>Write Amplification</td>
<td>2 x writes = double write amplification</td>
</tr>
<tr>
<td>RAID Latency</td>
<td>Added latency from software RAID</td>
</tr>
<tr>
<td>Reliability</td>
<td>Need to replace whole SSD if single Flash device fails</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Uses 2x the number of Gbytes for 2x cost</td>
</tr>
</tbody>
</table>
RAID 5 for Flash

- Efficient GBs.. but Poor Performance
- 75% lower IOPS per system
- Latency is impacted by Erases & Writes
  - Worse because of Read-Modify-Write
  - Worst-case because of striping

<table>
<thead>
<tr>
<th>Flash Issues</th>
<th>RAID 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erase Blocking</td>
<td>Reads are blocked. Writes are also blocked because of Read-Modify-Write.</td>
</tr>
<tr>
<td>Slow Writes</td>
<td>2x writes and 2 additional reads</td>
</tr>
<tr>
<td>Write Amplification</td>
<td>2 x writes = double write amplification</td>
</tr>
<tr>
<td>RAID Latency</td>
<td>Added latency from software RAID and extra reads. Striping requires worst-case latency of SSD set.</td>
</tr>
<tr>
<td>Reliability</td>
<td>Need to replace whole SSD if single Flash device fails</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Uses 1.2x the number of Gbytes for 1.2x cost</td>
</tr>
</tbody>
</table>
Hardware Flash RAID

1st and Best RAID algorithm for Enterprise-Grade Flash

**Hardware**
- No software latency
- Extremely High IOPS

**Flash**
- Non-blocking Erases
- Wear leveling across system
- Handles die & block failures automatically

**RAID**
- Handles module failures
- Enables hot-swap
- 4+1Parity = 80% efficient
  - Capacity & IOPS
Fail-in-Place

- **Scenario 1: Flash device Fails**
  - User data rebuilt using RAID algorithm
  - Data is rebuilt into other devices on same VIMM
  - VIMM keeps on operating!

- **Scenario 2: Flash VIMM fails**
  - VIMM taken out of service < 1 second
  - Rebuild data into 1 of N spare VIMMs < 1 hour
  - Only 20% less bandwidth
  - VIMM can replaced at any time
    - Hot service while appliance is operating
    - or monthly maintenance window

No Operational Impact
No Service Impact
The Infamous SSD “Write Cliff”
RAID makes it worse

SSDs without RAID
RAID-5 would reduce by 75%

Violin’s Sustained Performance With RAID
240,000+ IOPS
<1,000+ IOPS

Source: AnandTech Labs
Violin Hardware Flash RAID

- Efficient GBs and Performance
- Only 20% lower IOPS per system
- Latency is not impacted by Erases
  - No Read-Modify-Write
  - Striping delays minimized

<table>
<thead>
<tr>
<th>Flash Issues</th>
<th>Violin Flash RAID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erase Blocking</td>
<td>No Erase blocking</td>
</tr>
<tr>
<td>Slow Writes</td>
<td>1.25x Writes</td>
</tr>
<tr>
<td>Write Amplication</td>
<td>1.25x writes = less write amplification</td>
</tr>
<tr>
<td>RAID Latency</td>
<td>Hardware-based RAID algorithm with no erase blocking</td>
</tr>
<tr>
<td>Reliability</td>
<td>Flash device failures handled by RAID algorithm without module replacement.</td>
</tr>
<tr>
<td>Efficiency</td>
<td>Uses 1.25x the number of Gbytes for 1.25x cost</td>
</tr>
</tbody>
</table>
Impact on Flash Algorithms and Firmware

- Enterprise/Data Center requirements are vastly different from laptops and PCs
  - Running Oracle/DB2/SQLServer to make money
  - 24x7 non-stop operation required
  - Data loss is catastrophic - not annoying
- System level solutions are required
  - RAID approaches are well accepted.
- Flash requires new RAID algorithms and approaches
  - Violin uses an Intelligent Memory Module concept
  - VIMM has a role to play in RAID algorithm
    - Minimize latency – Non-blocking Erases
    - Increase IOPS - Hardware based garbage collection
    - Increase Reliability – Fail-in-place capabilities
Violin Flash Networking

Direct Attach

PCI Express
- PCIe x8 20Gbit/s
- Dual PCIe x4 2x 10Gbit/s

Fibre Channel
- 2-8 ports 4 /8Gb FC

Network Attach

10Gb Ethernet
- 2-4 ports 10Gb iSCSI or FCoE

Memory Gateways
- LUN Management
- MPIO, Security
- NFS Cache

Maximum Flexibility: Connectivity can be changed during 10 Year product lifetime
Hosts and OSes can also be changed
Coexistence with SAN and NAS

- High-activity LUNs moved to Silicon Storage Array
  - Increased performance
  - Reduced cost & power

OLTP Database
Oracle, SQL Server

Data Warehousing
Oracle, Greenplum

Enterprise Apps
MS Exchange

Logs, Snapshots
Logs, Back-ups

SAN Infrastructure

Indexes, Data, Temp space
Active Mailboxes

Large-Capacity Storage
- Snapshots for back-up
- Log Files
- Long-term file storage

HDD Storage
- 3K IOPS per shelf
- 5,000 µsec latency
- 10 – 1,000 TByte

High-Performance Storage
- Indexes
- Data tables
- Temp space
- Snapshots for analytics

Silicon Storage
- 250K IOPS per shelf
- < 200 µsec latency
- 1 to 100s TByte
Thank You

Morgan Littlewood

VP Product Management
Violin Memory, Inc

Mountain View, CA
littlewo@vmem.com