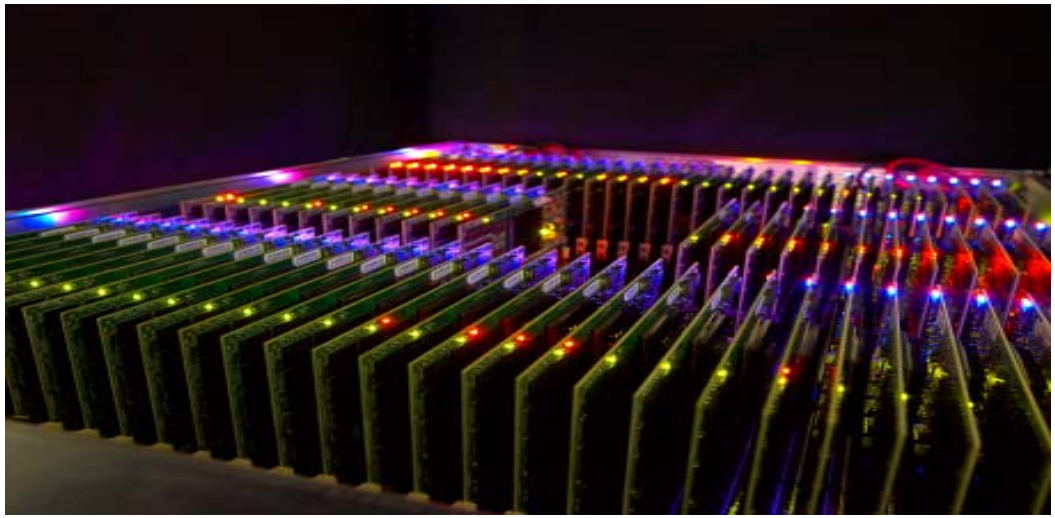

Violin
MEMORY
Unblocking
The I/O Bottleneck

RAID Algorithms for Flash Memory

(Making Flash Enterprise-grade and Cost-effective)

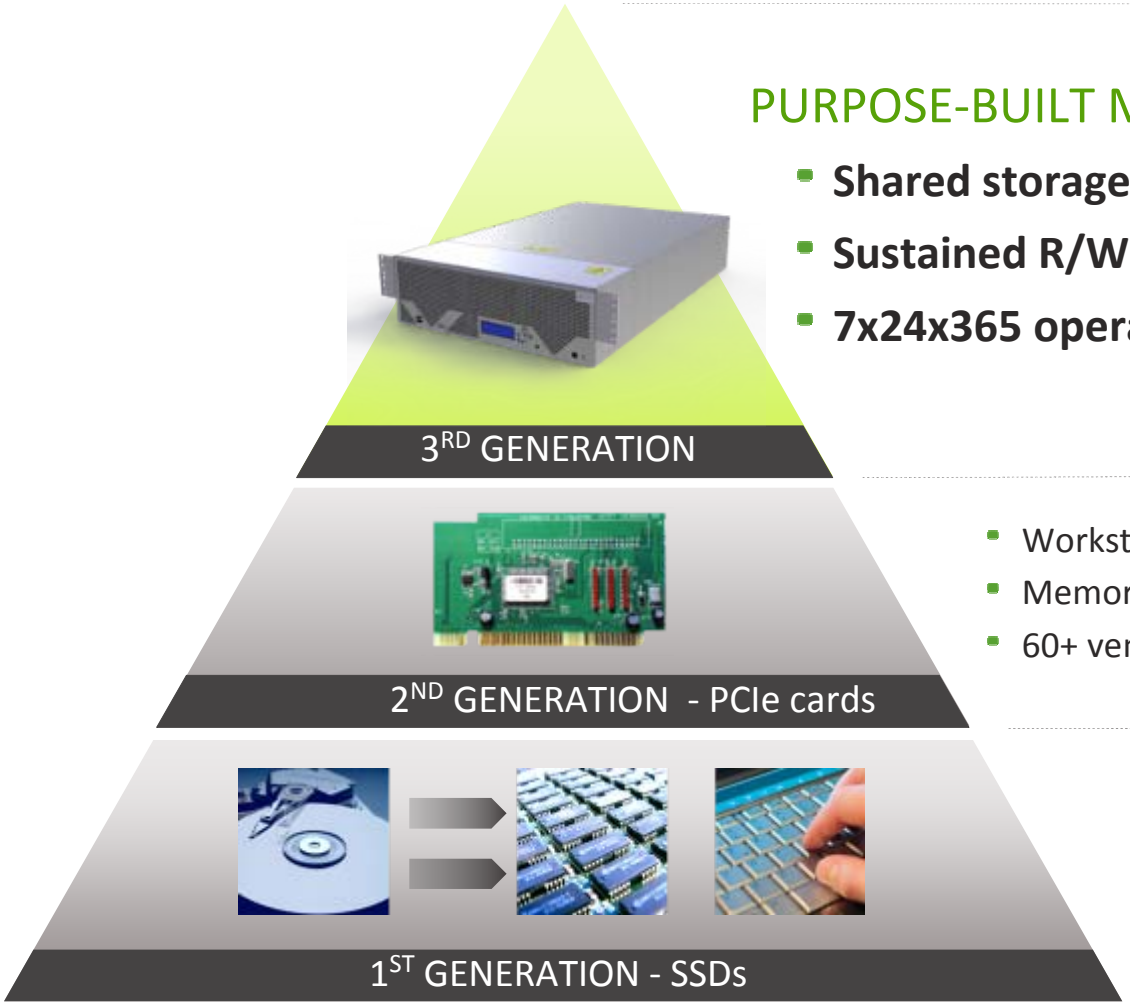


Morgan Littlewood

VP Product Management
Violin Memory, Inc

Mountain View, CA
littlewo@vmem.com

Evolving Flash Storage



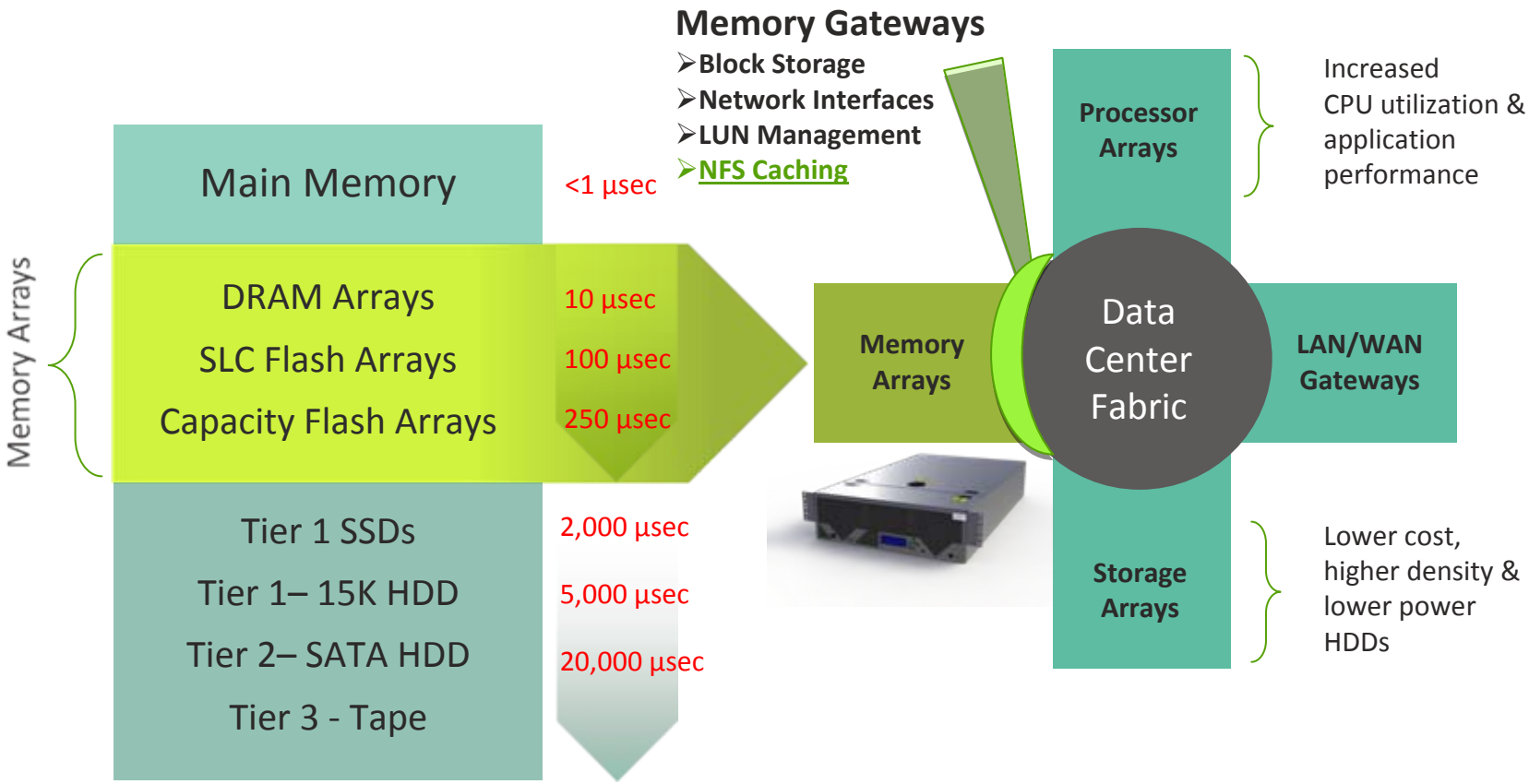
PURPOSE-BUILT Memory Arrays

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

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

- Direct drive replacement
- Cost sensitive
- 100s of vendors

Virtualized Data Center



"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

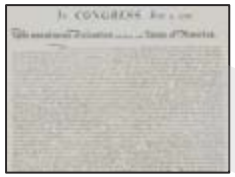
Flash Networking

- 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!

User block
4KB



2,000,000,000 Blocks

Flash Package
8GB

1344 Packages

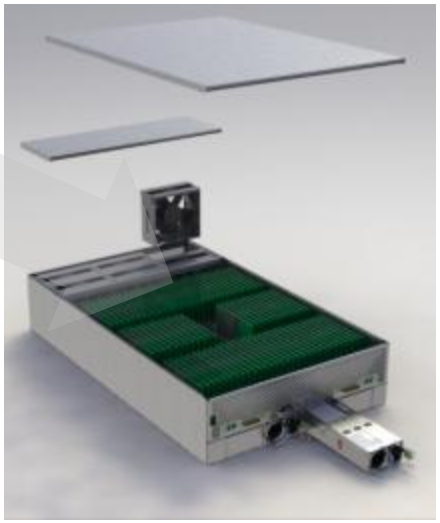
Flash VIMM
128GB

84 VIMMs

Flash RAID Group
600GB

16 Groups

Flash Systems
10TB



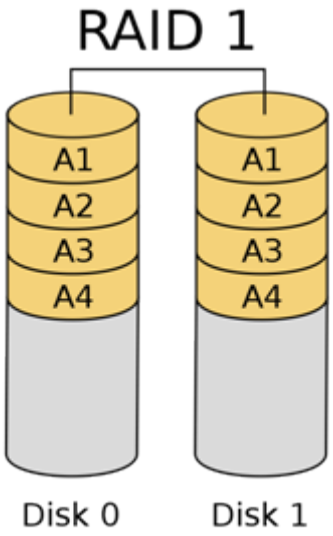
1 Violin Flash Array

Violin Flash arrays are similar density & cost to 15Krpm HDD storage systems!

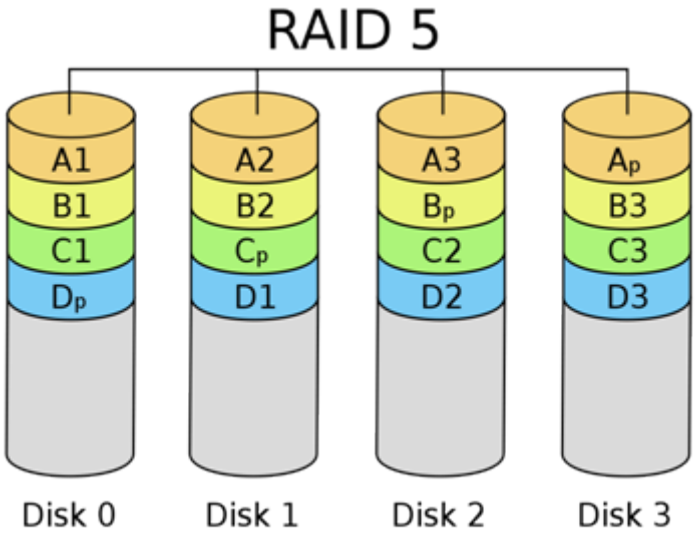
Why is RAID Needed?

- 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

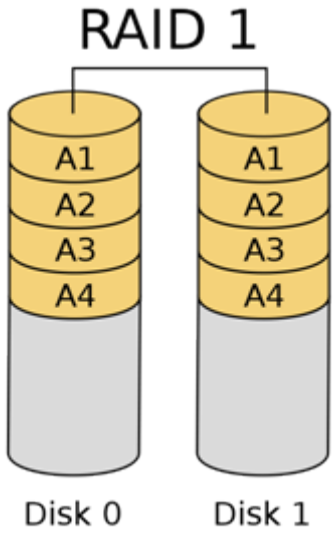


- Write = Read-Modify-Write
  - 2 Reads + 2 Writes
  - Data & Parity
- 80+% Space Efficient
- 30-70% Bandwidth
  - Poor for writes

RAID Issues for Flash Memory

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

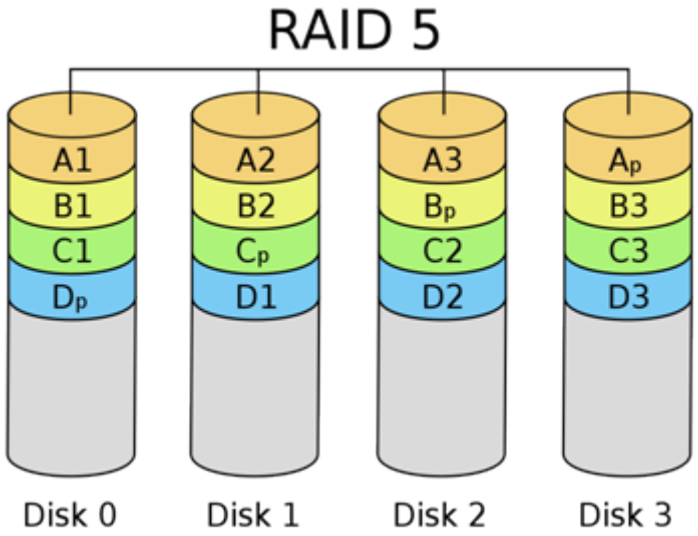
RAID 1 for Flash



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

Flash Issues	RAID 1
Erase Blocking	Same blocking as single Flash SSD
Slow Writes	2x writes
Write Amplication	2 x writes = double write amplification
RAID Latency	Added latency from software RAID
Reliability	Need to replace whole SSD if single Flash device fails
Efficiency	Uses 2x the number of Gbytes for 2x cost

RAID 5 for Flash

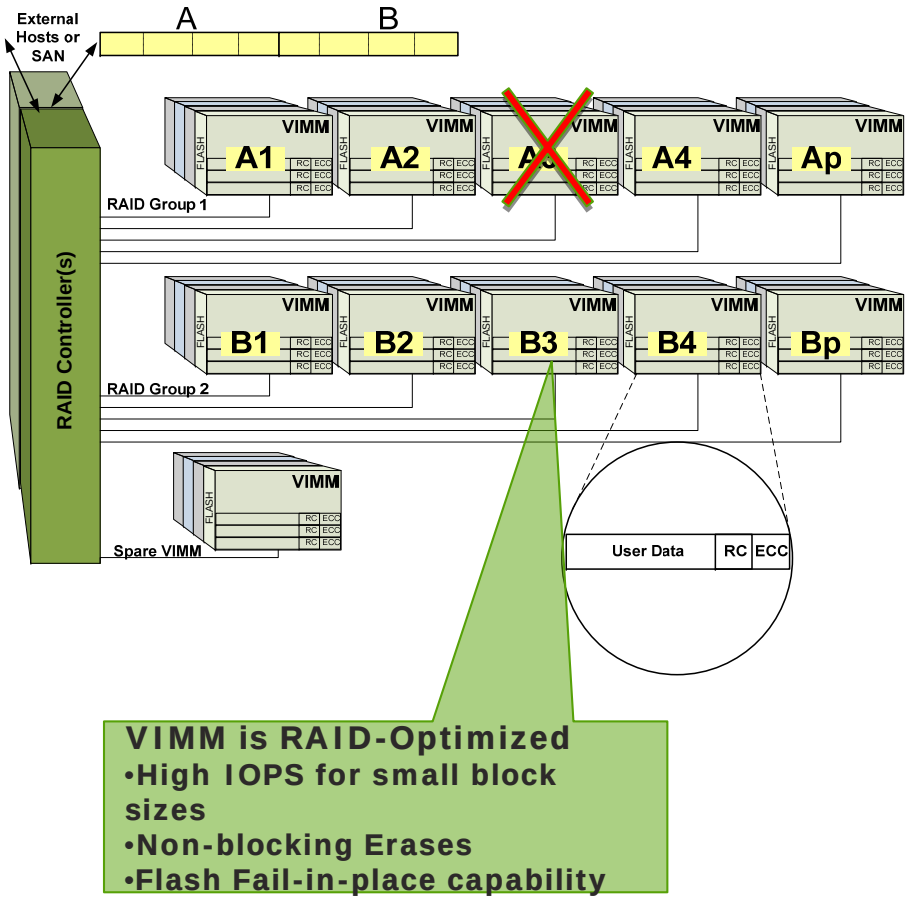


- Efficient GBs.. but Poor Performance
- 75% lower IOPS per system
- Latency is impacted by Erases & Writes
  - Worse because of Read-Modify-Write
  - Worst-case because of striping

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

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

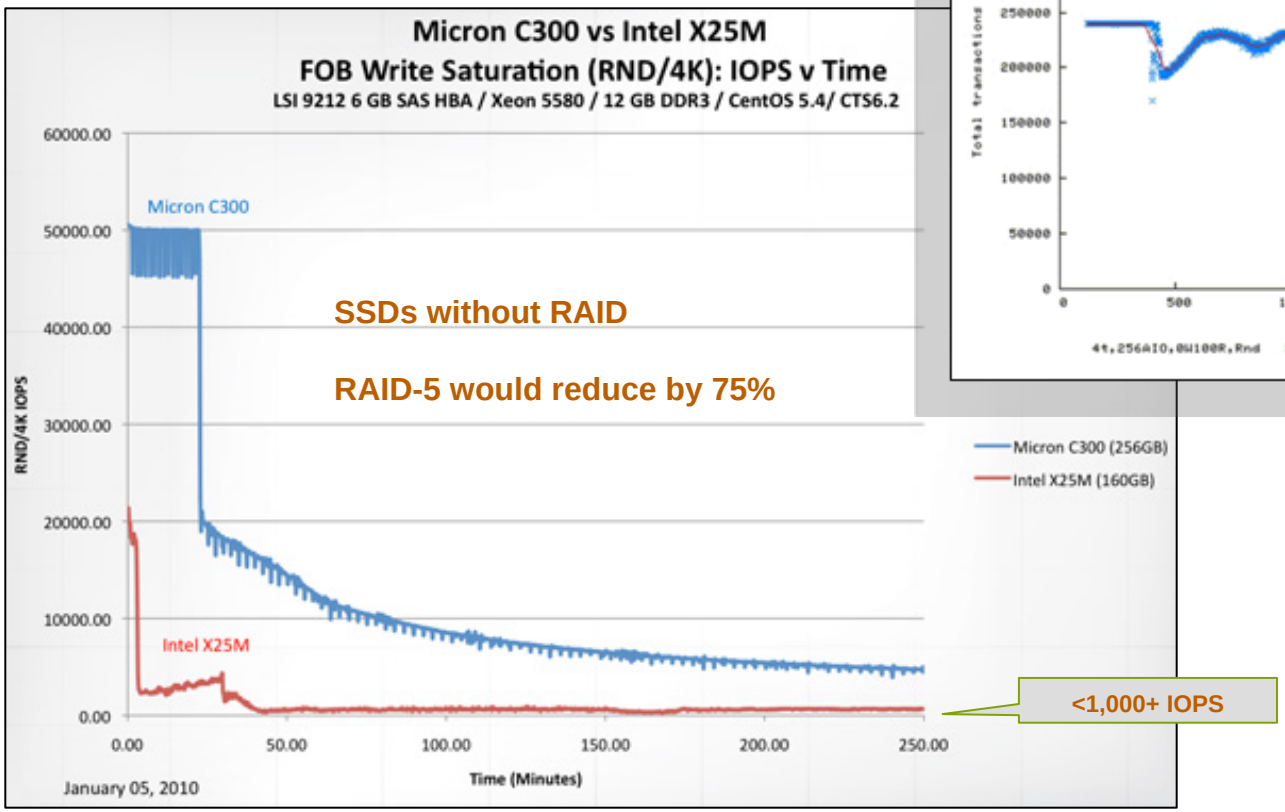
Fail-in-Place

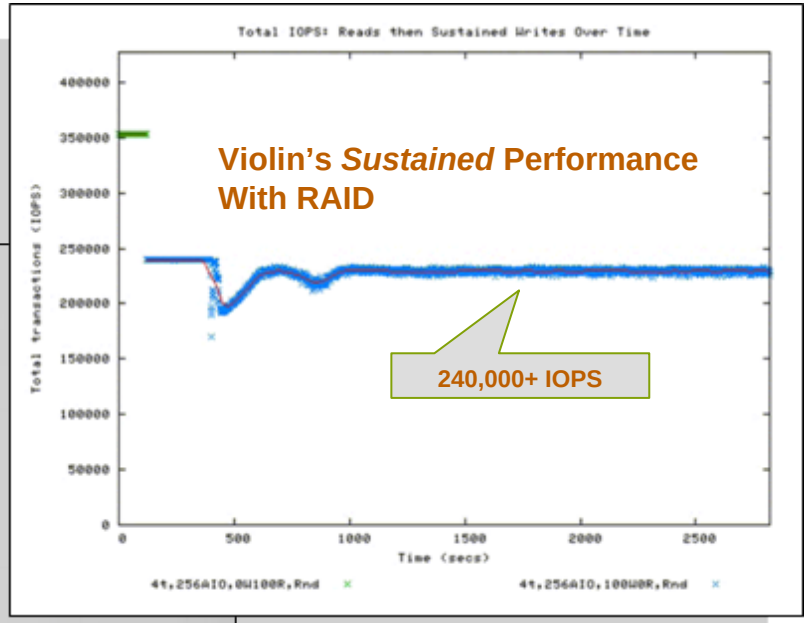
- **Scenario 1: Flash device Fails**
  - User data rebuilt using RAID algorithm
  - Data is rebuilt into other devices on same VIMM
  - VIMM keeps on operating!

- **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

The Infamous SSD "Write Cliff"

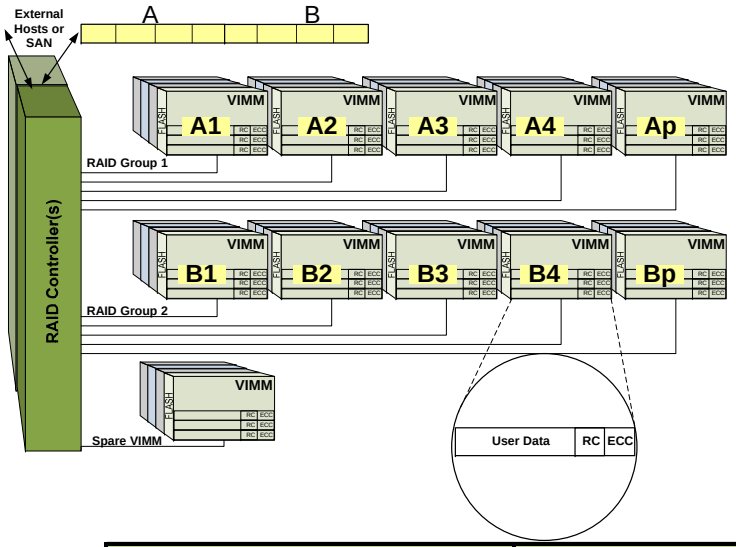
RAID makes it worse





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

Flash Issues	Violin Flash RAID
Erase Blocking	No Erase blocking
Slow Writes	1.25x Writes
Write Amplication	1.25x writes = less write amplification
RAID Latency	Hardware-based RAID algorithm with no erase blocking
Reliability	Flash device failures handled by RAID algorithm without module replacement.
Efficiency	Uses 1.25x the number of Gbytes for 1.25x cost

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

Server 2003
To Windows 7
WHQL

32 & 64bit
Open source driver

Solaris x86

Any Host
Any OS
Any Network

PCI Express

Fibre Channel

10Gb Ethernet

PCIe x8
20Gbit/s

Dual PCIe x4
2x 10Gbit/s

2 -8 ports
4 /8Gb FC

2 -4 ports
10Gb iSCSI or FCoE

Memory Gateways
LUN Management
MPIO, Security
NFS Cache

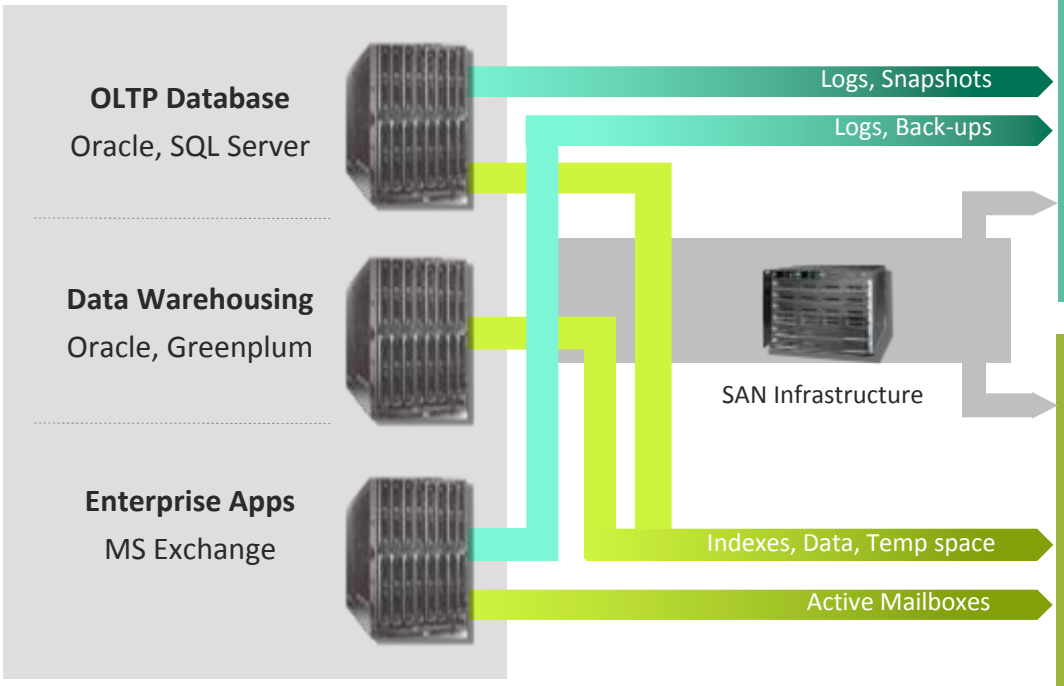
Direct Attach

Network Attach

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

Coexistence with SAN and NAS

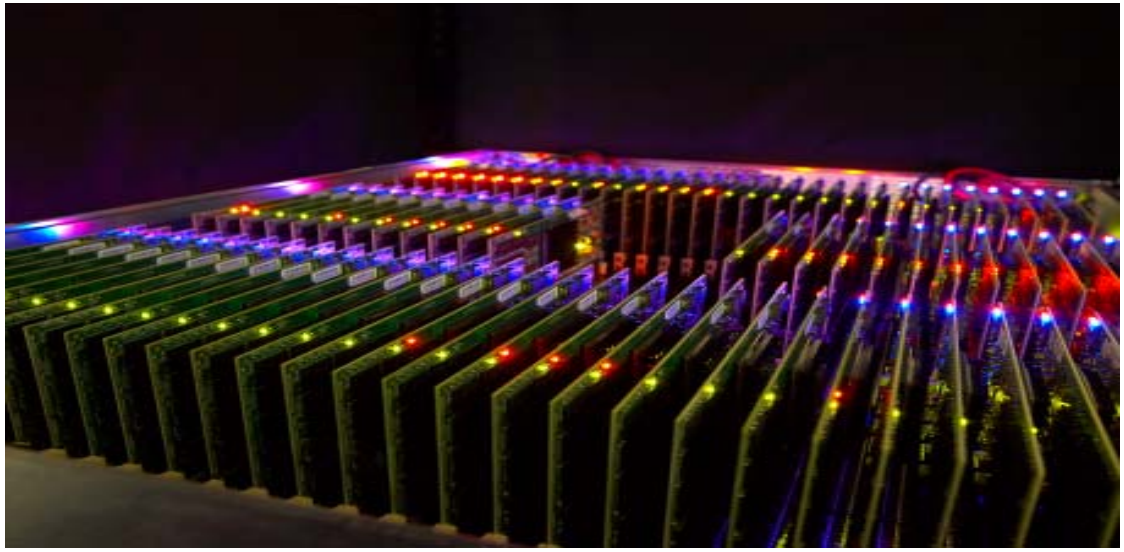
- High-activity LUNs moved to Silicon Storage Array
  - Increased performance
  - Reduced cost & power



- **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