I don't need that much performance and other fables from the world of storage

Jon Bennett
Founder and CTO
Violin Memory

Kevin Rowett
SVP of Engineering
A supercomputer is a device for turning compute-bound problems into I/O-bound problems

— Ken Batcher
What is a supercomputer?

Linpack Score

Cray 1 = 12 Mflops
My Phone = 20 Mflops
Feed me Seymour
No no, the other Seymour
Feed me Jon

For a little guy I am very hungry.....

4GB/s @ 1 Million IOPs please.... random not sequential
4GB/s & 500K IOPs looks like this

1,920 – 300 GB, 15K RPM FC disk drives

32 FC connections
4GB/s & 1M IOPs looks like this

64 -- 256GB SLC, $\text{VMC}^2$ VIMMs
High availability memory array

- Memory Gateway x2
- Network Interface x4
- Array Controller x2
- VIMMs x64
- vRAID Controller x4
- Power Controller x2
- Power Supplies x2
- Fans 3x2
40 Racks
Or One With Ten Times the Performance?

Today’s Data Center

Best the industry can offer
- 40 racks / 9,600 disks
- EMC Symmetrix*

*EMC 1 million IOPS benchmark
VMworld 2011

1 rack – 10 Violin Arrays

10 Million IOPS
40 GB/sec
I don't need that much performance and other fables from the world of storage

Jon Bennett
Founder and CTO
Violin Memory

Kevin Rowett
SVP of Engineering
• Paradigm Shift
• Metrics
• Performance comparisons
• VXN Architecture
• vRAID
• Flash realities
• Comments
The Woes of Revolutionary Technology

Hard to displace a mature incumbent technology

- Management APIs
- Application Plugins
- Snapshots
- Clones
- Replication
- Deduplication
- Backup
- Encryption
- TRIM
- VAAI
- VSA
- Thin Provisioning
- Migration
- Live Maintenance
- Non-disruptive Upgrades
- Space Heater
- Air Ship Ballast
- Etc
All those features.............

birthday cake granola bars on a stick!
Paradigm Shift

Diagram of a duck's head.
In the enterprise disk is dead.
$/GB the wrong metric

Scatter Plot: SPC-1 non-SSD/Flash (140GB or larger, disk drives only)
IOPS vs Number of Drives as of 26May2011

http://silvertonconsulting.com/blog/2012/01/19/latest-spc-1-results-iops-vs-drive-counts-chart-of-the-month/
Using the wrong metric is bad

**Challenge**
- Classroom training with 700 workstations
- Slow boot and application load time
  - Boot > 180 sec, load several minutes
  - Audio / video streaming jittery
- FC SAN w/ 300 short-stroked 15K rpm HDD

**Solution**
- Violin flash Memory Arrays
- Repurpose SAN array for archive

**Benefits**
- Higher end-user productivity
- Boot time reduced to 9 sec.
- Application load to under 30 sec
- Smooth audio/video streaming
• Lies, Damn Lies and Benchmarks
  – Exploitable benchmarks, compression, dedupe, limited address range, too much sequential access.

• New Benchmarks may be (are) wrong

• The best benchmark is application itself
  – What happens when the app is wrong?
### Current SPC-1 #2

<table>
<thead>
<tr>
<th>SPC-1 Reported Data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tested Storage Product (TSP) Name:</strong> HP P10000 3PAR V800 Storage System</td>
<td></td>
</tr>
<tr>
<td><strong>Metric</strong></td>
<td><strong>Reported Result</strong></td>
</tr>
<tr>
<td>SPC-1 IOPS™</td>
<td>450,212.66</td>
</tr>
<tr>
<td>SPC-1 Price-Performance™</td>
<td>$6.59/SPC-1 IOPS™</td>
</tr>
<tr>
<td>Total ASU Capacity</td>
<td>230,400,000 GB</td>
</tr>
<tr>
<td>Data Protection Level</td>
<td>Protected (Mirroring)</td>
</tr>
<tr>
<td>Total TSC Price (including three-year maintenance)</td>
<td>$2,965,892</td>
</tr>
</tbody>
</table>

![Server racks](image)
Latency

Average Response Time Distribution (Ramp_sust @9005 BSUs)
– Worst case, not average

– Its all random, anything trying to exploit access patterns is doomed to fail

– Except when it isn’t
  • databases
Hybrid HDD/SSD system

SPC-1 Reported Data

<table>
<thead>
<tr>
<th>Metric</th>
<th>Reported Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPC-1 IOPS™</td>
<td>137,066.20</td>
</tr>
<tr>
<td>SPC-1 Price-Performance™</td>
<td>$2.99/SPC-1 IOPS™</td>
</tr>
<tr>
<td>Total ASU Capacity</td>
<td>23,703.035 GB</td>
</tr>
<tr>
<td>Data Protection Level</td>
<td>Protected (Mirroring)</td>
</tr>
<tr>
<td>Total TSC Price (including three-year maintenance)</td>
<td>$409,933</td>
</tr>
</tbody>
</table>

Priced Storage Configuration Components:

- **Oracle Sun ZFS Storage 7420c Appliance**
  - 2 – Sun ZFS 7420 controllers (cluster configuration)
    - 512 GB cache/memory per controller (1024 GB total)
  - 12 – Sun StorageTek 8Gb Fibre Channel PCIe HBAs (includes SFPs)
  - 8 – dual-ported SAS-2 HBAs
  - 8 – 512 GB Solid State Devices (read cache SSDs, 4096 GB total)
  - 8 – 73 GB Solid State Devices (write cache SSDs, 584 GB total)
  - 24 – 8 Gb FC front-end connections (12 used)
  - 16 – SAS-2 backend connections (16 used)

- 12 – 2m LC to LC FC Optical Cables RoHS-6 compliant
- 8 – 2m, Mini, shielded, SAS cables
- 12 – Sun disk shelf: base chassis each with 2 SAS-2 IO modules, 2 AC PSUs and 2 cooling fans
- 280 – 300 GB 15K RPM SAS-2 disk drives
Average Response Time Distribution (Ramp_sust @2741 BSUs)

- All ASUs
- ASU1
- ASU2
- ASU3

Startup

Measurement Interval

Average Response Time (ms)

Test Run Minutes

Lat...... ??????
Something hidden in the corner
I see what you did there…
Just use SSDs

– Garbage collection
– Write cliff
– Have to update the meta data too
  • this matters a lot if there is no locality of access
– Too many constraints
  • Protocols
  • Form factors
  • Power (Loss)
  • General purpose
  • Cooling
  • Stand Alone Operation
Pure SSD system

Summary of Results

<table>
<thead>
<tr>
<th>Metric</th>
<th>Reported Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPC-1 IOPS™</td>
<td>20,003.03</td>
</tr>
<tr>
<td>SPC-1 Price-Performance</td>
<td>$6.55/SPC-1 IOPS™</td>
</tr>
<tr>
<td>Total ASU Capacity</td>
<td>515.397 GB</td>
</tr>
<tr>
<td>Data Protection Level</td>
<td>Protected (Mirroring)</td>
</tr>
<tr>
<td>Total TSC Price (including three-year maintenance)</td>
<td>$130,982.94</td>
</tr>
</tbody>
</table>

HP ProLiant BL460c G6 Server
HP BladeSystem c7000 Enclosure
2 – 24 Port 8 Gb zoned Blade Enclosure Switches
2 – Dual Port 8 Gb FC HBAs
2 – 8 Gb Shortwave SFPs
2 – LC-LC cables

HP P6500 Enterprise Virtual Array
Dual Controllers with 8 GB cache/controller
8 – 200 GB Solid State Devices (SSDs)
Latency

Average Response Time Distribution (Ramp_sust @400 BSUs)

Measurement Interval

Average Response Time (ms)

Test Run Minutes

- All ASUs
- ASU1
- ASU2
- ASU3
Nothing hiding this time
VXM Architecture
VXM Architecture
VXM Architecture

- ACM
- HBA
- MG (DB Server)
- Management Ports
- VCM
- VM
- DB Clients
- HBA
- Management Ports
The power supply
vRAID

So how does it work?
vRAID

4KB User Data

1KB
1KB
1KB
1KB
1KB

VIMM
VIMM
VIMM
VIMM
VIMM

E / W
R
R
R
R
R

R
R
R
E / W
R
R

R
R
R
E / W
R
R

Time

Note: conceptual ONLY - NOT actual implementation pattern!
How can it keep up while only writing one column at a time?
The write cliff

Degradation - Bandwidth

Fusion-io 160GB SLC card drops from 700 MB/s to 100MB/s with 128KB blocks.
- Garbage collection in software (host CPU)
- Tests at <50% usable capacity.
- 10K IOPS at 4K block.
- MLC cards are 70% worse.
The write cliff

Degradation - Bandwidth

- Virident tach1On (400GB)
- TMS RamSan 20 (450GB)
- Fusion IO ioDrive Duo (Single Slot, 160GB)
- Intel X-25M (160GB)
- OCZ Colossus (250 GB)

RAMSAN SLC card drops from 700 MB/s to 240MB/s with 128KB blocks.
- Tests at <50% usable capacity.
- <25K IOPS at 4K block.
V-3220

Performance Report

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
<th>VIMM Type</th>
<th>VIMM Number</th>
<th>Raw Capacity (GB)</th>
<th>Random IOPS</th>
<th>Bandwidth (128K) MB/s</th>
<th>Latency µs</th>
<th>Formatted Capacity (GB)</th>
<th>Power W</th>
<th>Operation</th>
<th>Times min</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-3220</td>
<td>V-3200 Memory Array with 21.4TB SLC Flash</td>
<td>SLC</td>
<td>84</td>
<td>21,504</td>
<td>Max. Read 347,000</td>
<td>1437</td>
<td>90</td>
<td>15,393@ 87%</td>
<td>856</td>
<td>Boot</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Hot-swappable Flash/Power/Fans</td>
<td>Size</td>
<td>256GB</td>
<td>64</td>
<td>Max. Write 192,000</td>
<td>762</td>
<td>20</td>
<td>11,545@ 65%</td>
<td>860</td>
<td>Format</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>vRAID with P-Clex4GT interface</td>
<td>Usable After RAID</td>
<td>18,384</td>
<td>17,592</td>
<td>Sust. Mixed 191,686</td>
<td>926</td>
<td>234</td>
<td>13,744@ 78%</td>
<td>940</td>
<td>Rebuild</td>
<td>128</td>
</tr>
</tbody>
</table>

Sustained Random IOPS (4K) vs. Read/Write Mix

Typical HDD Array Latency = 5,000 - 20,000 Microseconds

685 Clyde Ave, Mountain View, CA 94043 USA
33 Wood Avenue South, 3rd Floor, Iselin, NJ 08830 USA
Ph: 1-888-9VIOLIN (984-6546) Int +1 732 218 6076
Email: sales@violin-memory.com www.violin-memory.com
V6000: Sustained IOPS

Sustained Random IOPS with RAID

- **V6616**
  - Hardware-based GC
  - Distributed operation
  - vRAID: 4+1P
  - Striping across 64 Modules
  - 256GB SLC Modules
  - 900K Sustained Write IOPS

- **Disk Array (4 racks)**
  - 1000 15Krpm HDDs
  - RAID-1
  - 200K Sustained IOPS

- **PCIe card**
  - Software-based GC
  - No RAID across modules
  - 10K Sustained Write IOPS

V-6616 IOPS (4K) vs. Read/Write Mix

- Max. Random IOPS
- Sustained Random IOPS
V6000: Sustained IOPS (MLC)

Sustained Random IOPS with RAID

- **V6232**
  - Hardware-based GC
  - Distributed operation
  - vRAID: 4+1P
  - Striping across 64 Modules
  - 512GB MLC Modules
  - 500K Sustained Write IOPS

- **Disk Array**
  - 1000 15Krpm HDDs
  - RAID-1
  - 200K Sustained IOPS

- **PCIe card**
  - Software-based GC
  - No RAID across modules
  - 10K Sustained Write IOPS

V-6232 IOPS (4K) vs. Read/Write Mix

- Max. Random IOPS
- Sustained Random IOPS
Sustained random write
4 hour test (962MBps +/- 2%)
Ave Bandwidth vs R/W mix

Sum of For Pivot - average MBps

Read %

# Col13: Read weight (percentage)
Bandwidth vs. R/W mix

- Random workload, 256 threads, 4KB
- 1 hour, 40 minute (60 secs per %)

- Performance
  - Over 1GBps for most workload mixes
  - Variability @ high read % is due to aggressive preventative read-disturb data movement
Flash wears out

24nm MLC @10K accelerated P/E cycles

Recommended ECC of 40 bits @ 3K P/E cycles
Flash Realities

– “SSD” requirements are meaningless in the datacenter
– What does wear out really mean?
– Who cares about the data sheet?
– Don’t need 1 year, 1 month or even 1 week retention in the enterprise
– Enterprise scale flash doesn’t wear out
  • But PCIe cards used for swap space will.
MLC vs eMLC

- Binned parts
- Architecture changes
- More ECC
- Slower P/E
- FAB Experience
- Nothing at all
We “manage” the SSDs
It is memory not disk, use it as such

- A file system to named later
  - 30K untuned
  - 240K remove disk layout code
  - 500K change locking
  - 900K remove/relax ordering requirements
  - Hadoop needs to learn from this
Ricardo B (ricardo.b@xxxxx.xx) on 1/21/11 wrote:

> Can *anyone* point out those mythical SSDs

> - for which TRIM reduces performance, either in the short or long term.

Umm. It's more like the reverse.

The _common_ case is that trim reduces performance, because it makes file delete slower.

Yes, yes, so people are trying to come up with better ways, where you discard in batches, and where you just discard entirely off-line. As mentioned, my personal opinion is that on-line trim is just stupid, and doing it offline (when the system is quieter - so not necessarily _totally_ offline, but basically not during real ops) is the only sane approach.
But dammit, can you guys shut the f*ck up already? Admit that you were wrong, and trim isn't fast, and isn't a simple thing. Because IO patterns matter, and without ESP you can never know when it's a good idea to TRIM.

99% of the time, you're probably much better just reusing the block and letting the normal write activity be the "trim and overwrite".

I'm done arguing. Come back when you have re-acquainted yourself with reality. As it is, I'm just seeing a lot of "TRIM should be fast" noise, with no f*cking clue.

Linus
Need to revisit your assumptions

– Just assume you are doing everything wrong

– “Incremental” improvements not going to cut it.

– “best practices” that no longer are.

– Even legal issues need examination
Interesting areas

• OS stacks
• File Systems
• Databases
• Cloud-like systems
• Big Data
“Will no one rid me of this turbulent protocol?”

- Jon