



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 Batchner

What is a supercomputer?

violin
MEMORY



Linpack Score

Cray 1 = 12 Mflops

My Phone = 20 Mflops

Feed me Seymour

violin
MEMORY



No no, the *other* Seymour

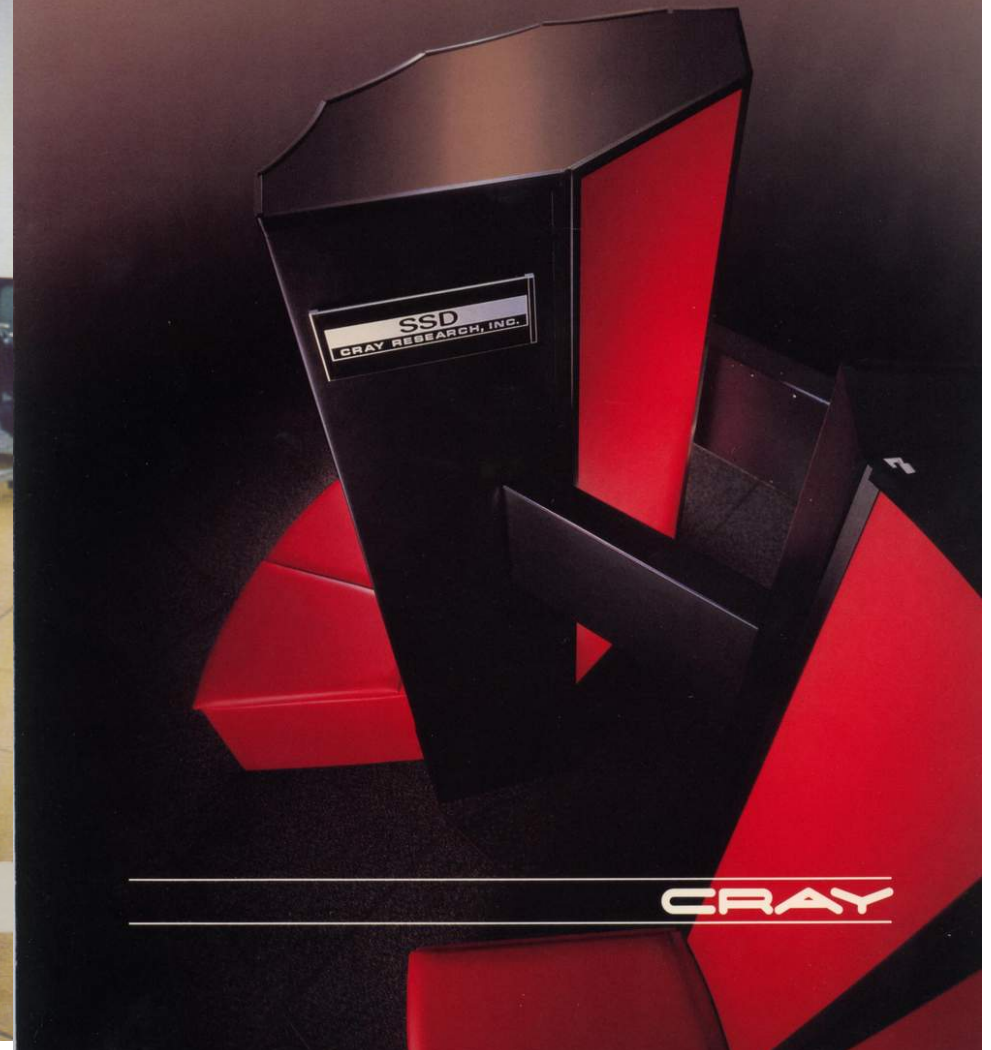
violin
MEMORY

Solid-state Storage Device (SSD)



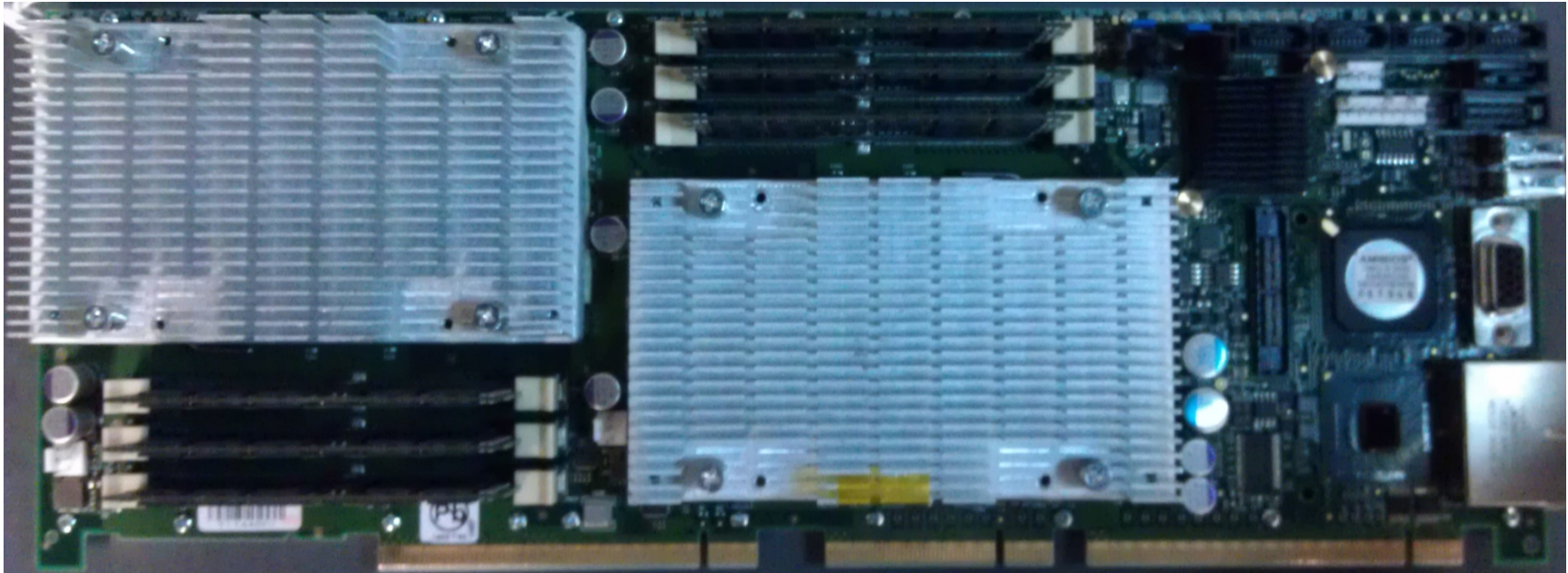
A Product Release
from Cray Research, Inc.

Solid-state Storage Device



Feed me Jon

violin
MEMORY



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

4GB/s @ 1 Million IOPs please....
random not sequential

4GB/s & 500K IOPs looks like this **violin** MEMORY

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



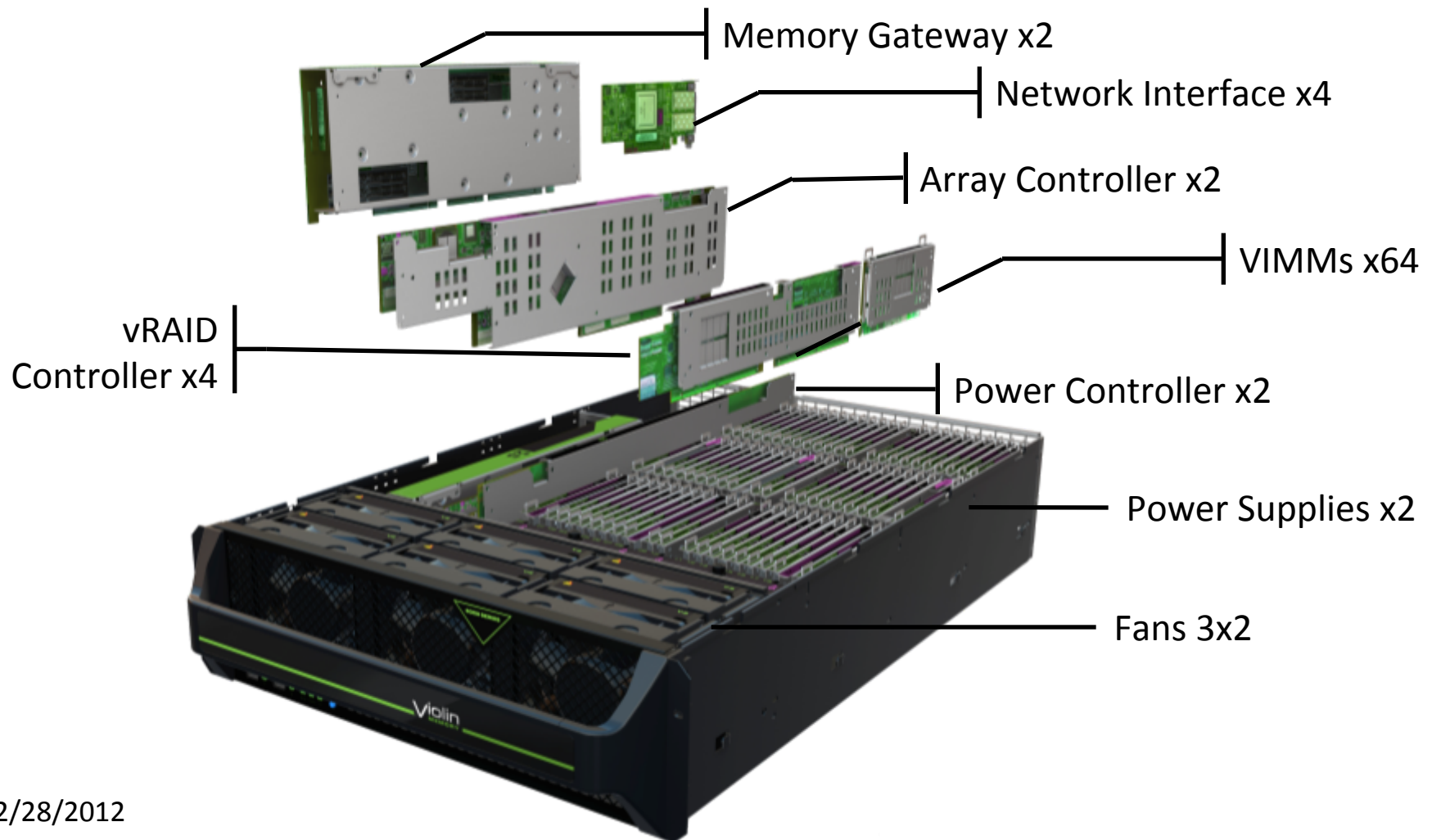
4GB/s & 1M IOPs looks like this

violin
MEMORY

64 -- 256GB SLC, VMC² VIMMs



High availability memory array



2/28/2012

40 Racks Or One With Ten Times the Performance?

violin
MEMORY

Today's Data Center



10 Million IOPS
40 GB/sec

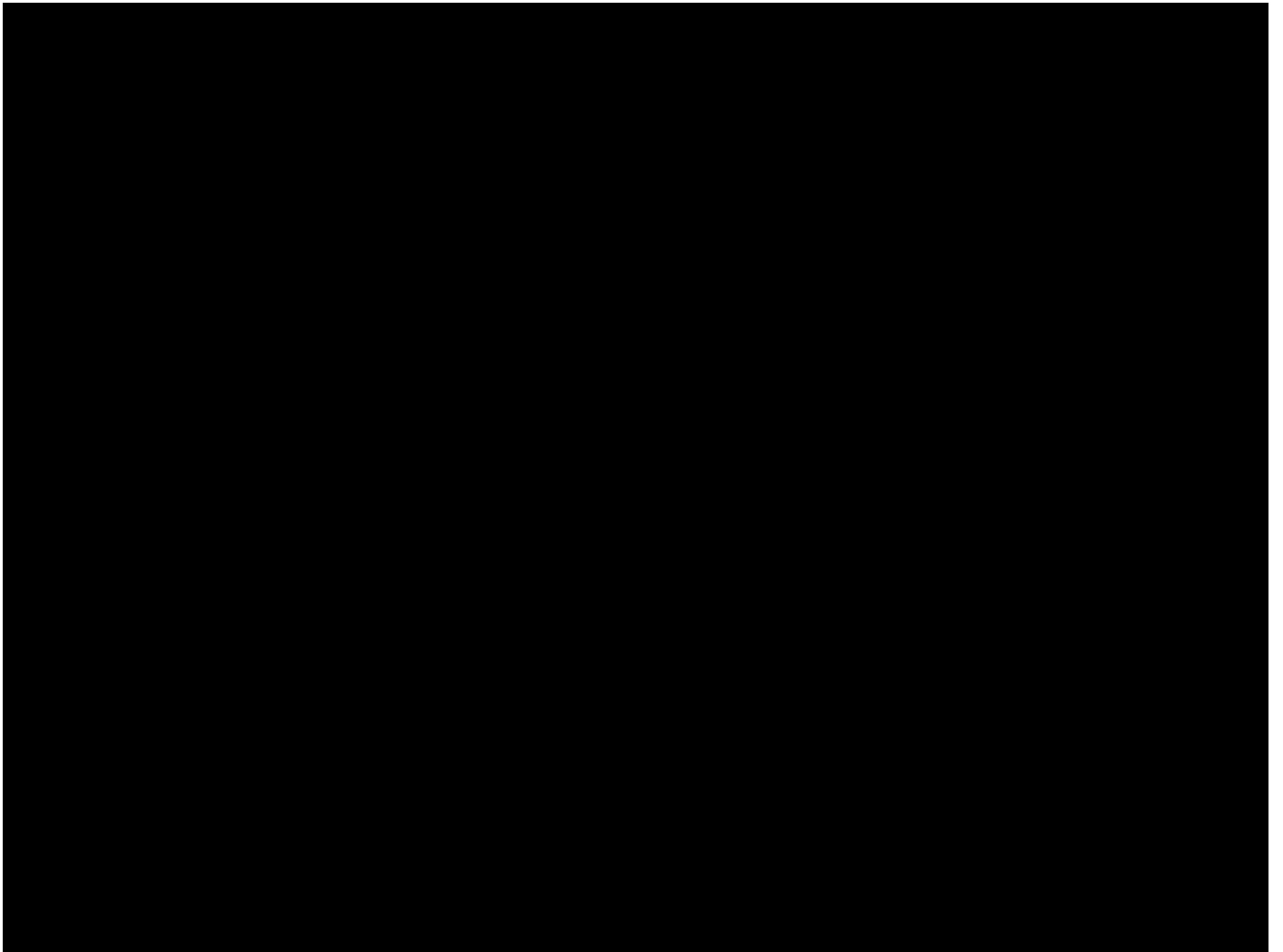
Best the industry can offer

- 40 racks / 9,600 disks
- EMC Symmetrix*

*EMC 1 million IOPS benchmark
[VMworld 2011](#)



1 rack – 10 Violin Arrays





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
- VXM 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.....

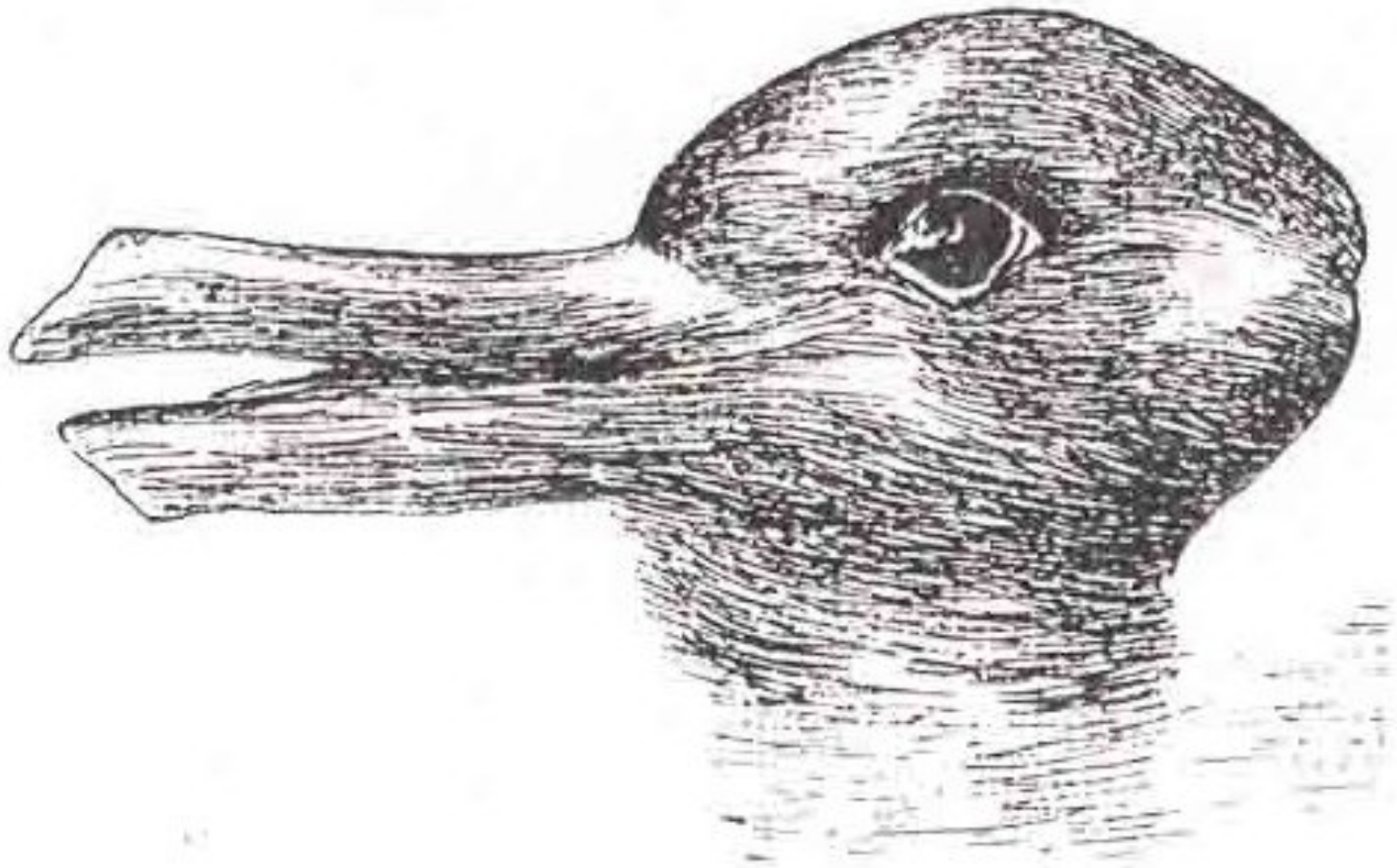
violin
MEMORY

birthday cake granola bars on a stick!



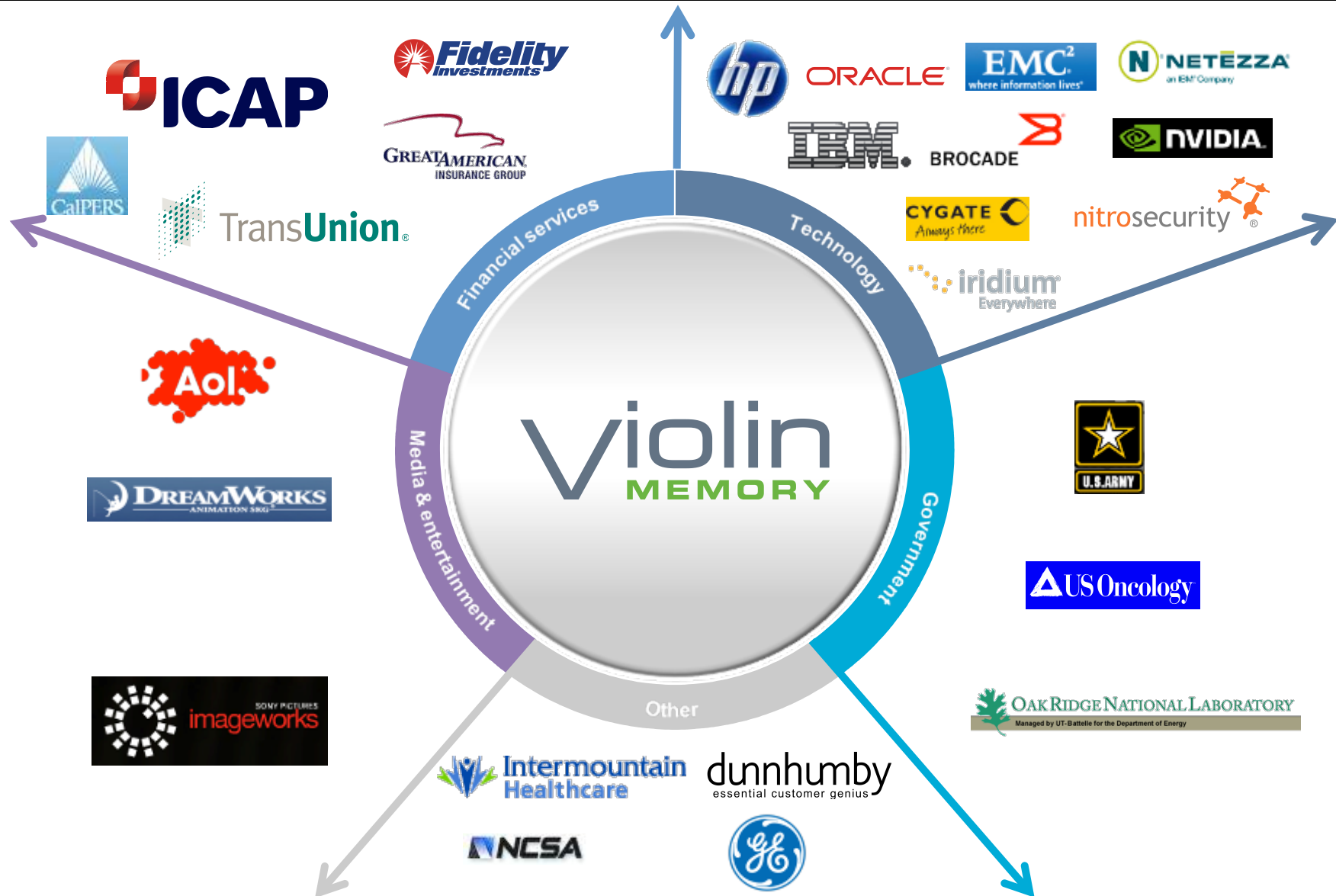
Paradigm Shift

violin
MEMORY



In the enterprise disk is dead.

violin
MEMORY

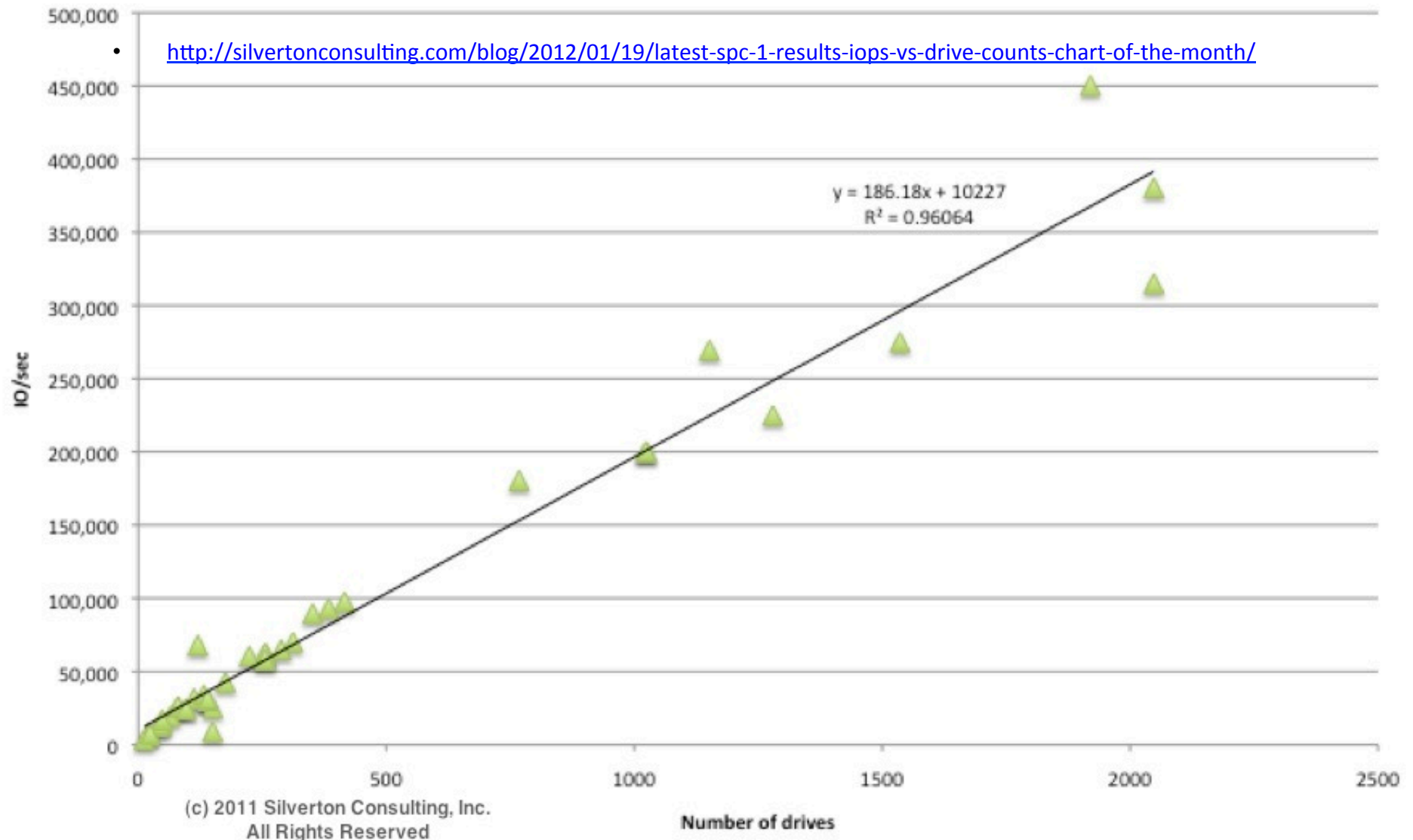


2/28/2012

\$/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**



Using the wrong metric is bad



Federal Customer

- Reduced boot time from 180+ sec. to 9 sec.
- Reduced application load time from minutes to under 30 sec.
- Repurposed SAN for archive

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?

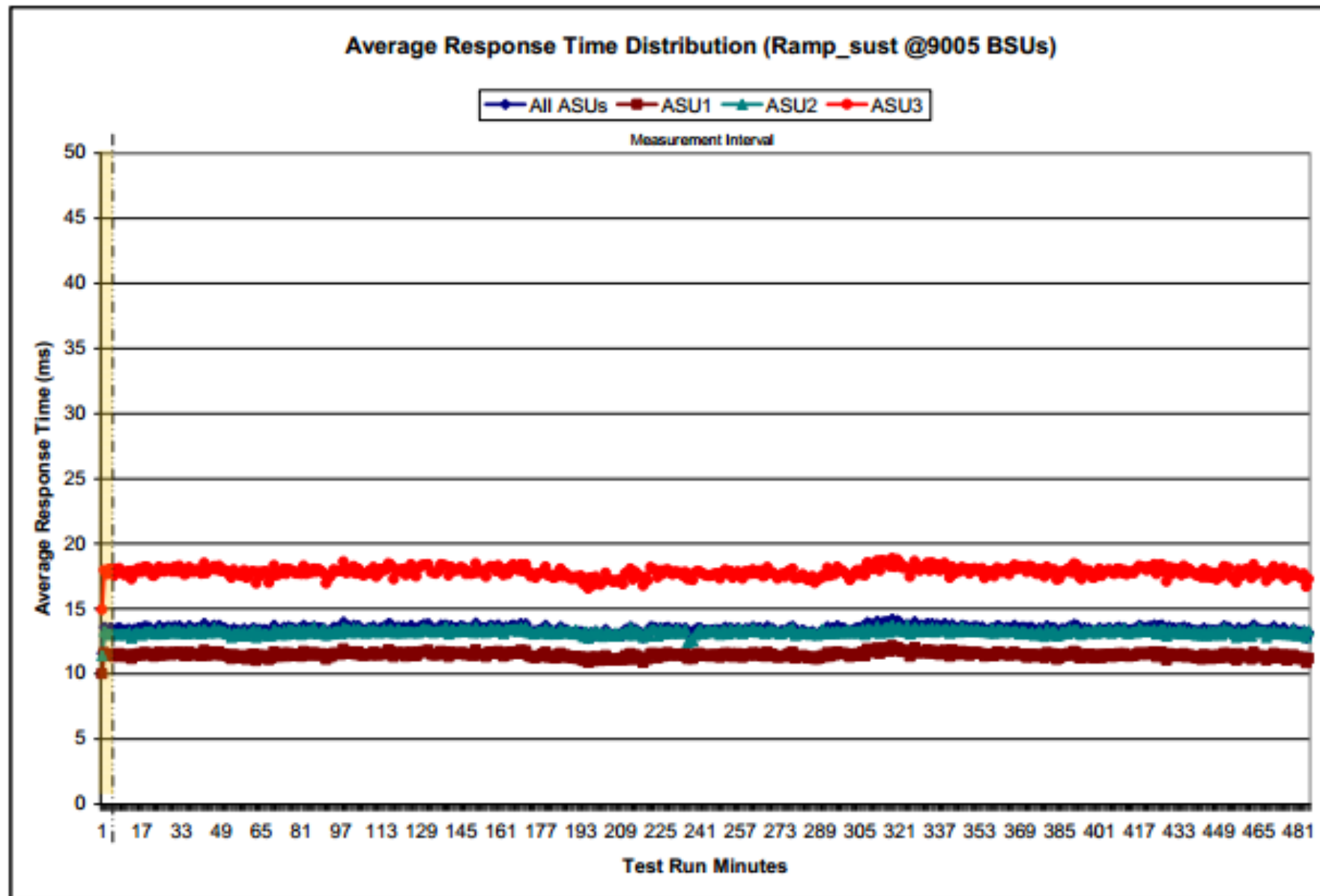
Current SPC-1 #2



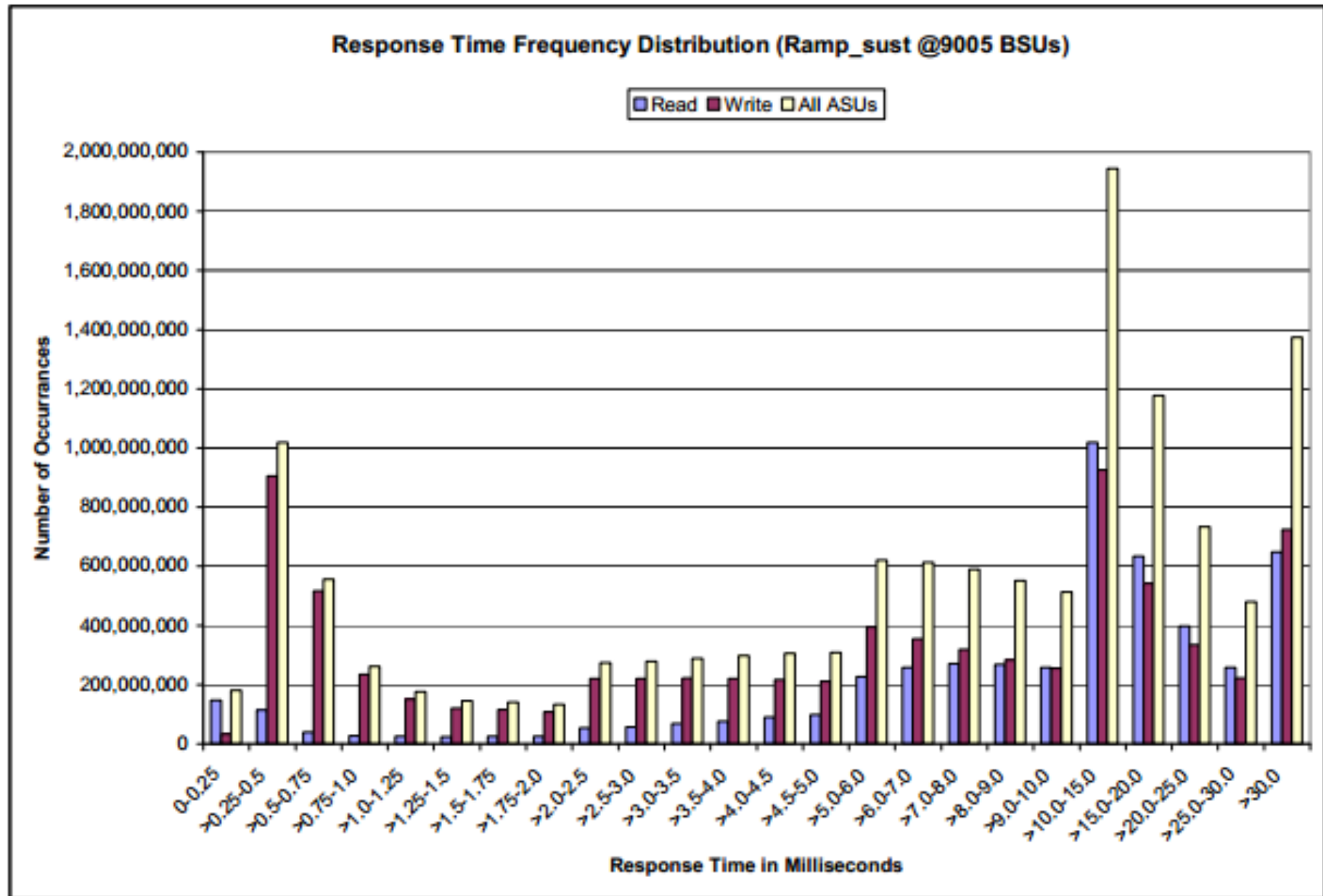
SPC-1 Reported Data	
Tested Storage Product (TSP) Name: HP P10000 3PAR V800 Storage System	
Metric	Reported Result
SPC-1 IOPS™	450,212.66
SPC-1 Price-Performance™	\$6.59/SPC-1 IOPS™
Total ASU Capacity	230,400.000 GB
Data Protection Level	Protected (<i>Mirroring</i>)
Total TSC Price (including three-year maintenance)	\$2,965,892



Latency



Latency



The “I/O Blender”



- 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



6 – dual-port 8 Gb FC HBAs



6 – 8Gb front-end connections per controller
(1 connection from each HBA to each controller)



Oracle Sun ZFS Storage 7420c Appliance

SPC-1 Reported Data

Tested Storage Product (TSP) Name: Oracle Sun ZFS Storage 7420c Appliance

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

Priced Storage Configuration Components:

6 – Sun StorageTek 8Gb Fibre Channel PCIe HBAs (includes SFPs)

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)

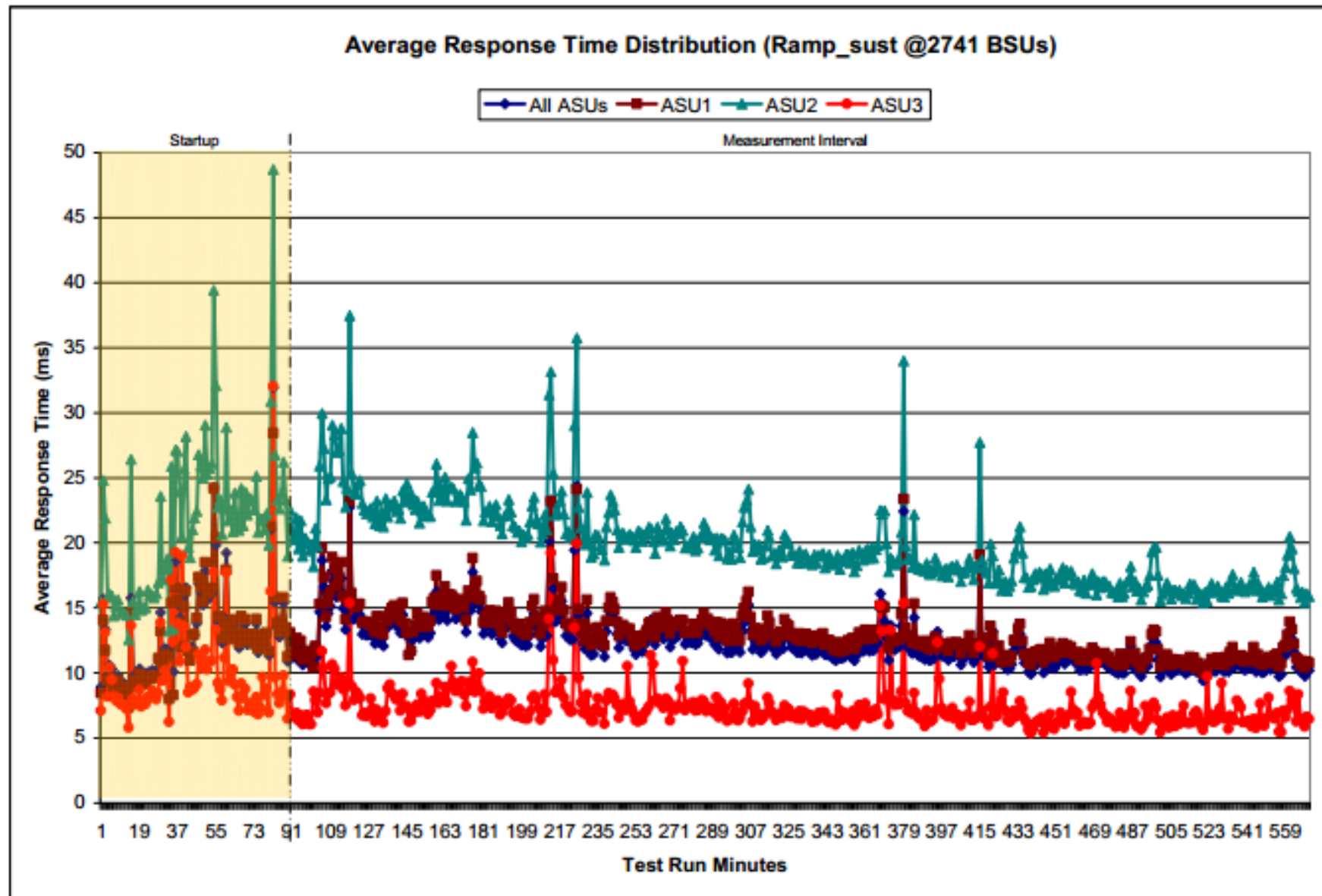
12 – 2m LC to LC FC Optical Cables RoHS-6 compliant

8 – 2m, Mini, shielded, SAS cables

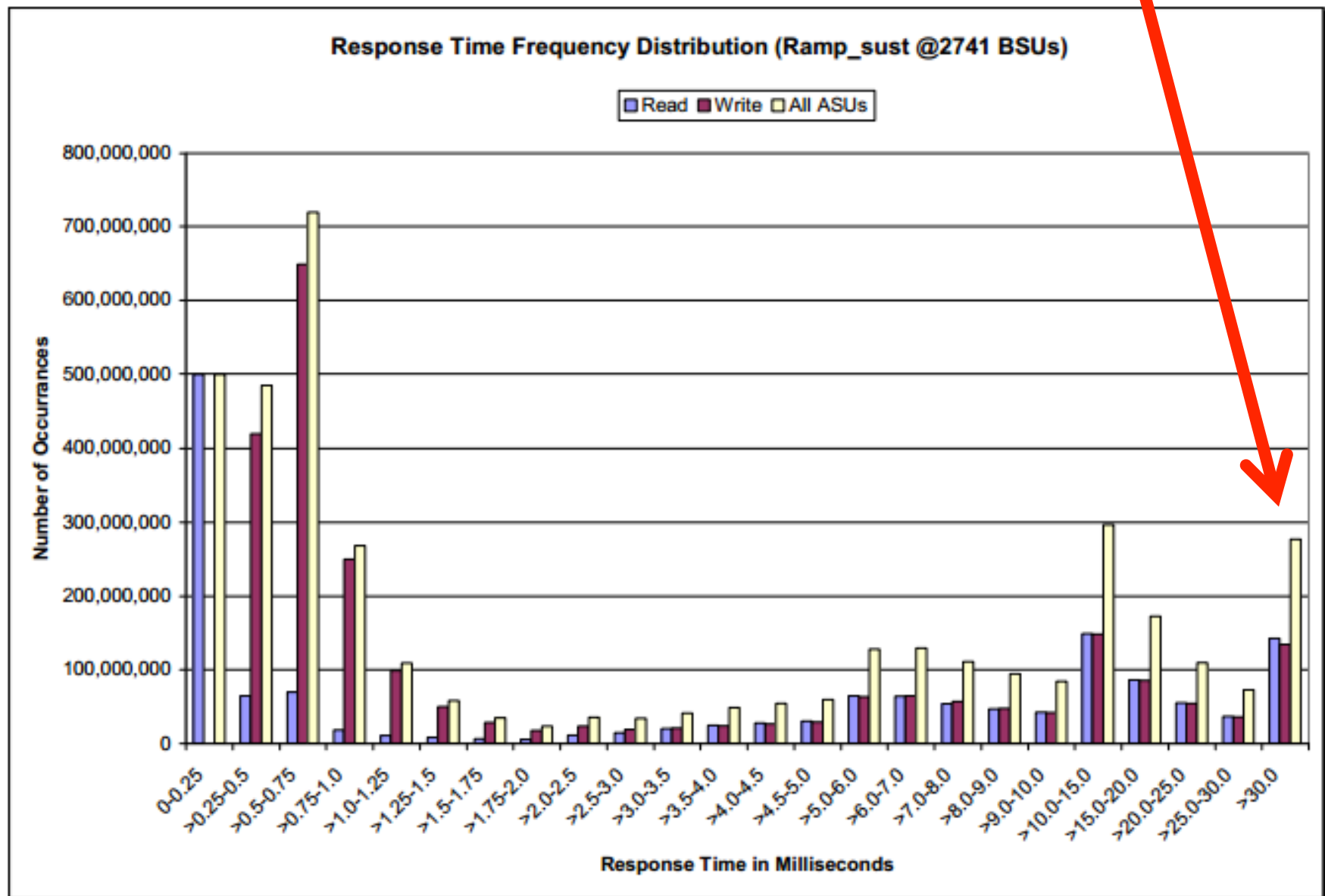
12 – Sun disk shelf: base chassis
each with 2 SAS-2 IO modules, 2 AC PSUs and 2 cooling fans

280 – 300 GB 15K RPM SAS-2 disk drives

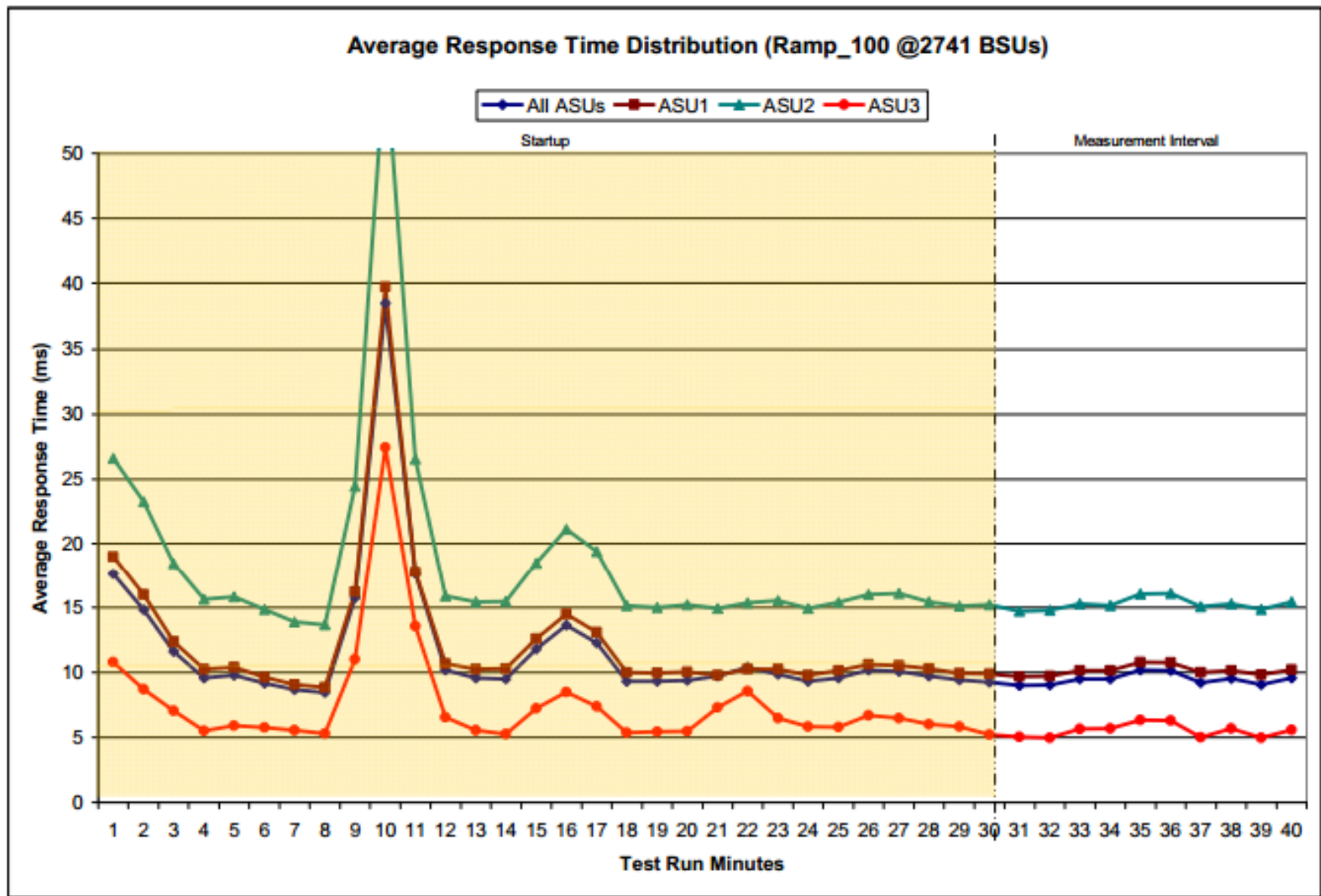
Lat..... ??????



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



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

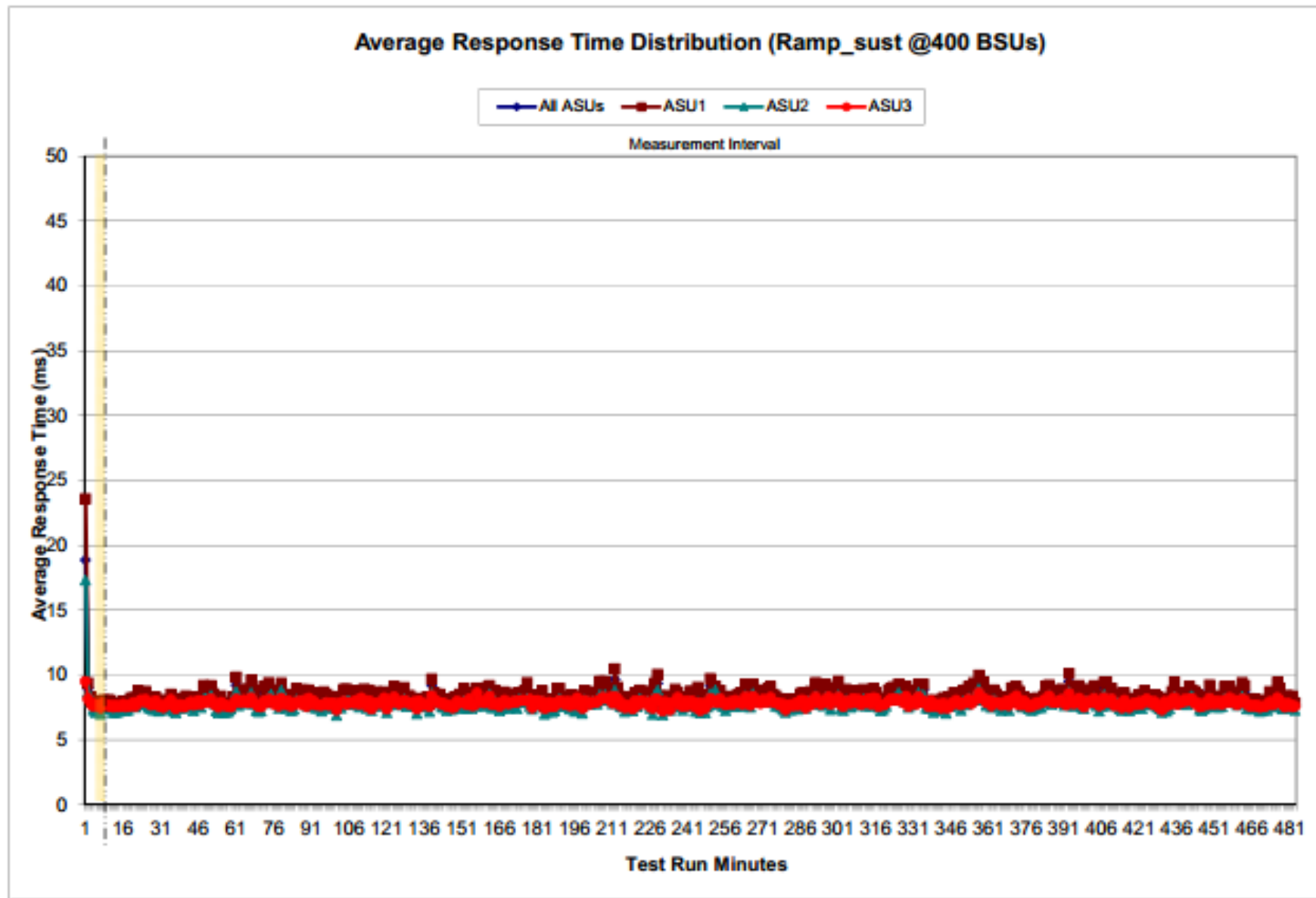
Dual Controllers with 8 GB cache/controller

8 –200 GB Solid State Devices (SSDs)

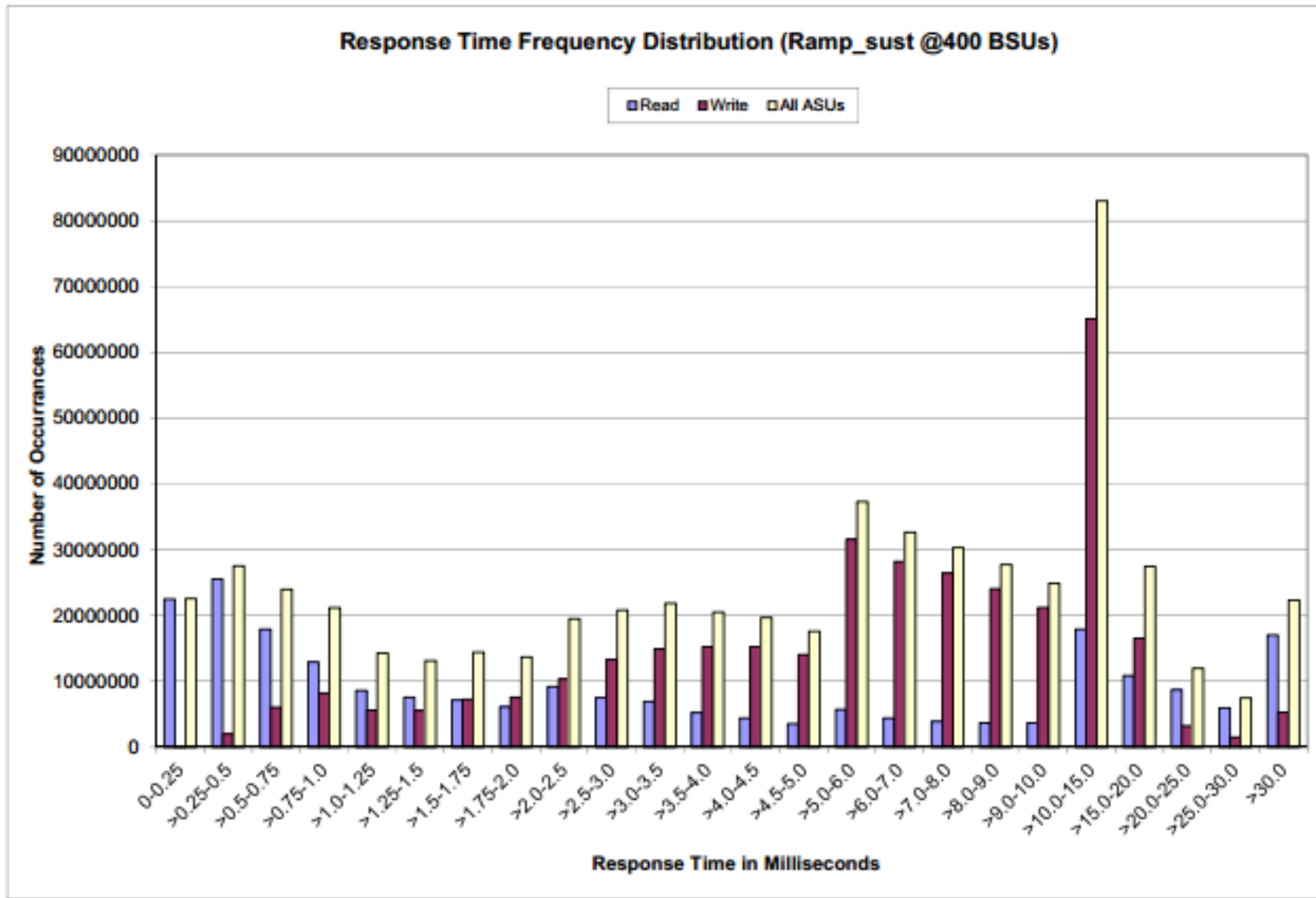
Summary of Results

SPC-1 Results	
Tested Storage Configuration (TSC) Name: HP P6500 Enterprise Virtual Array	
Metric	Reported Result
SPC-1 IOPS™	20,003.03
SPC-1 Price-Performance	\$6.55/SPC-1 IOPS™
Total ASU Capacity	515.397 GB
Data Protection Level	Protected (Mirroring)
Total TSC Price (including three-year maintenance)	\$130,982.94

Latency



Nothing hiding this time

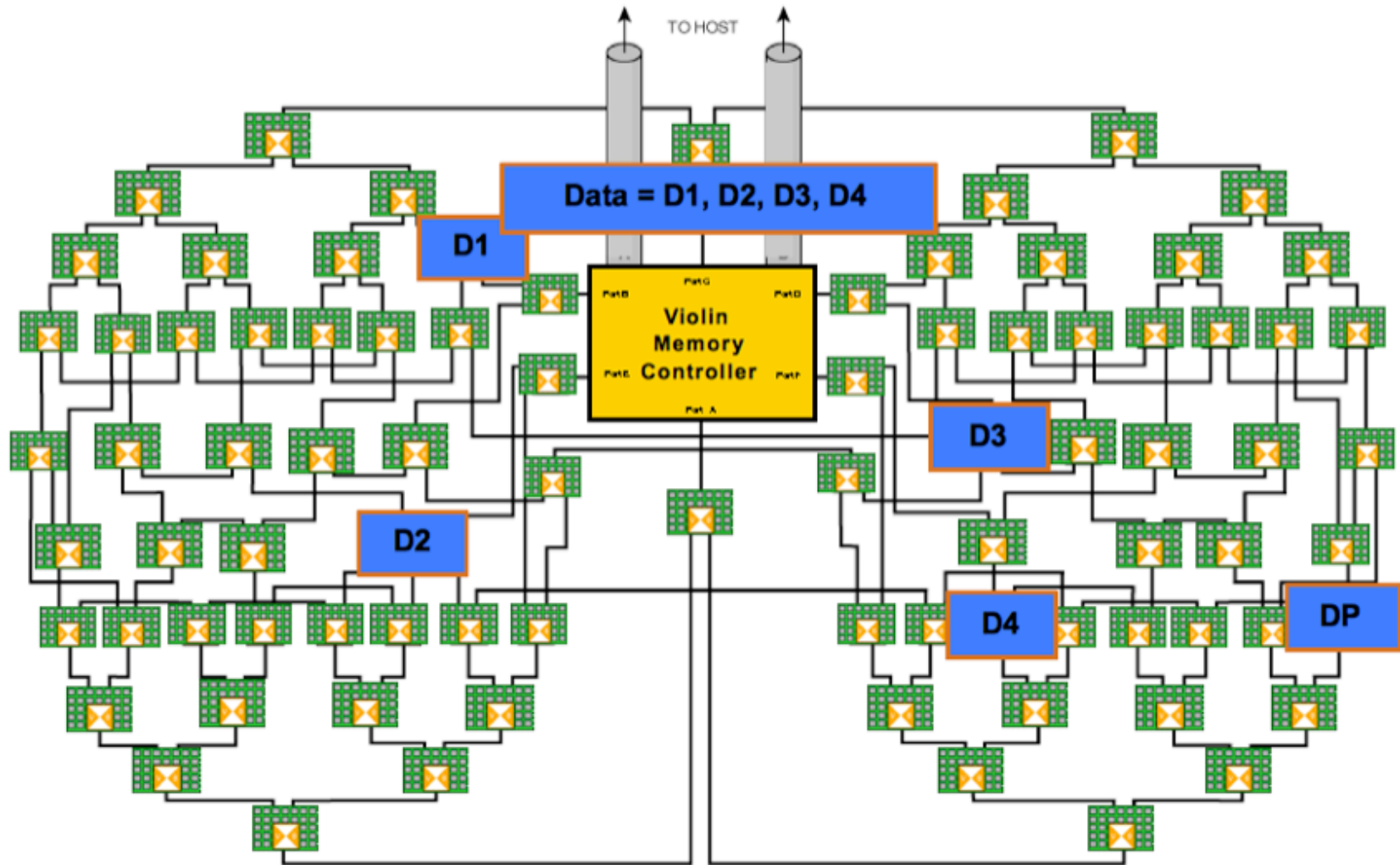


VXM Architecture

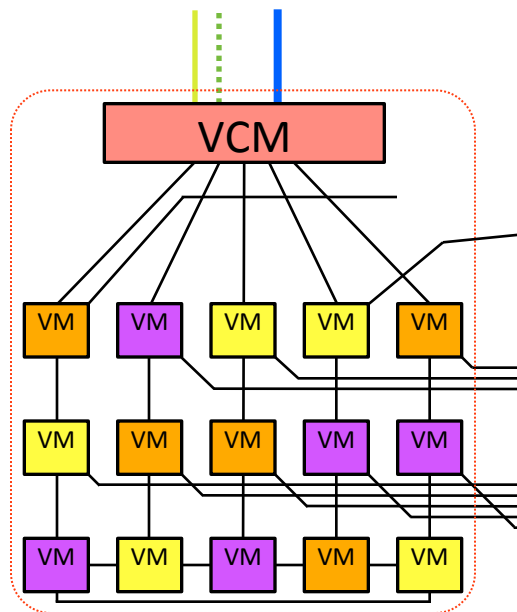


VXM Architecture

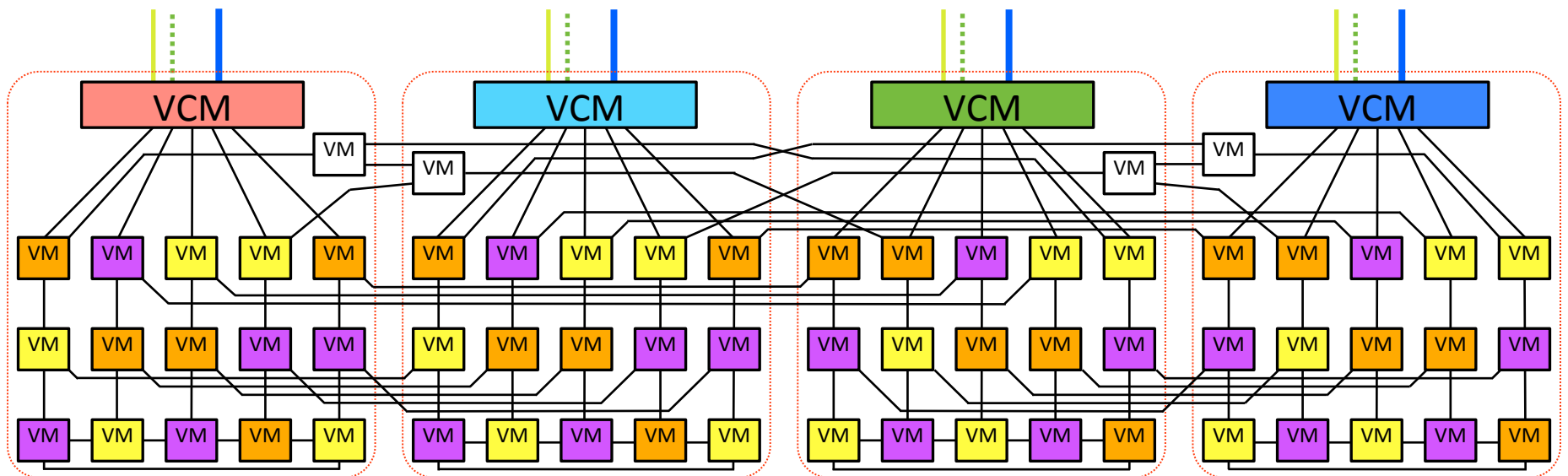
violin
MEMORY



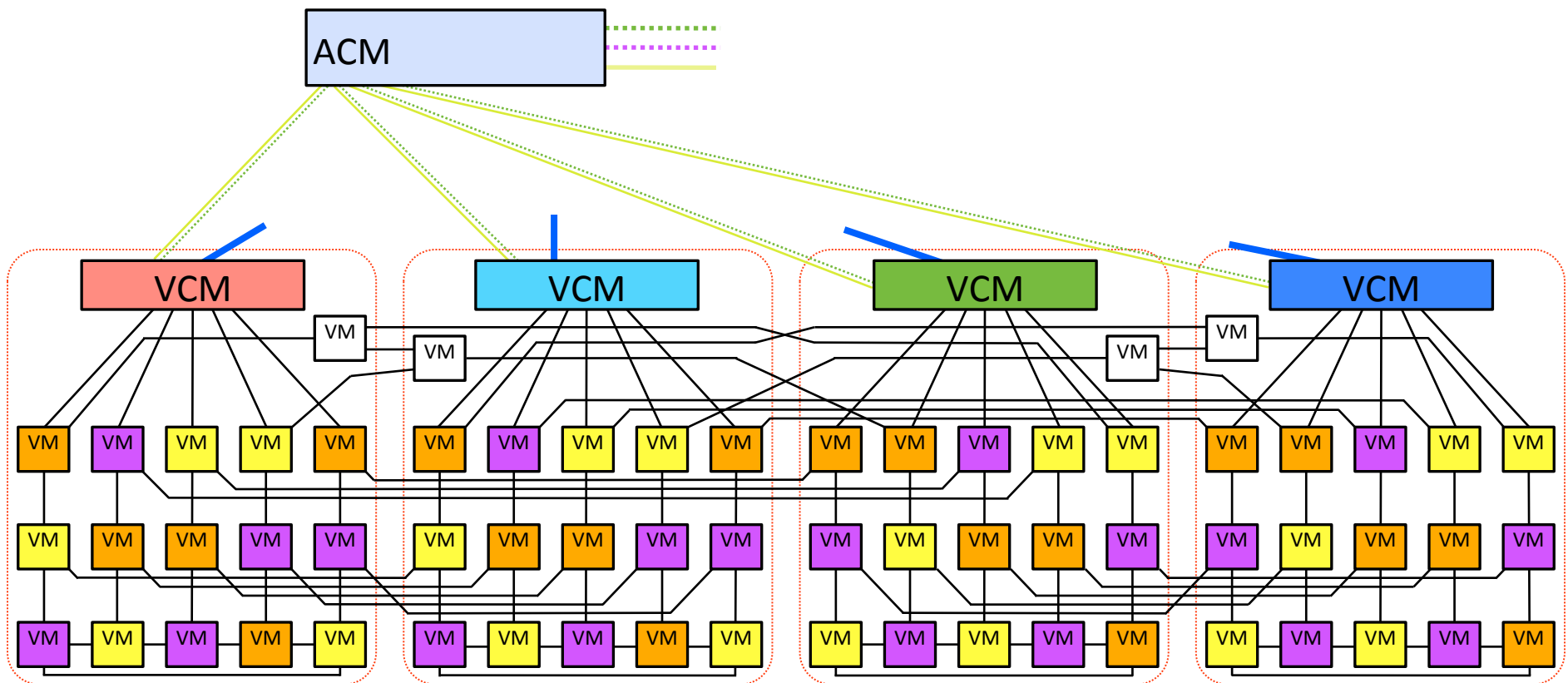
VXM Architecture



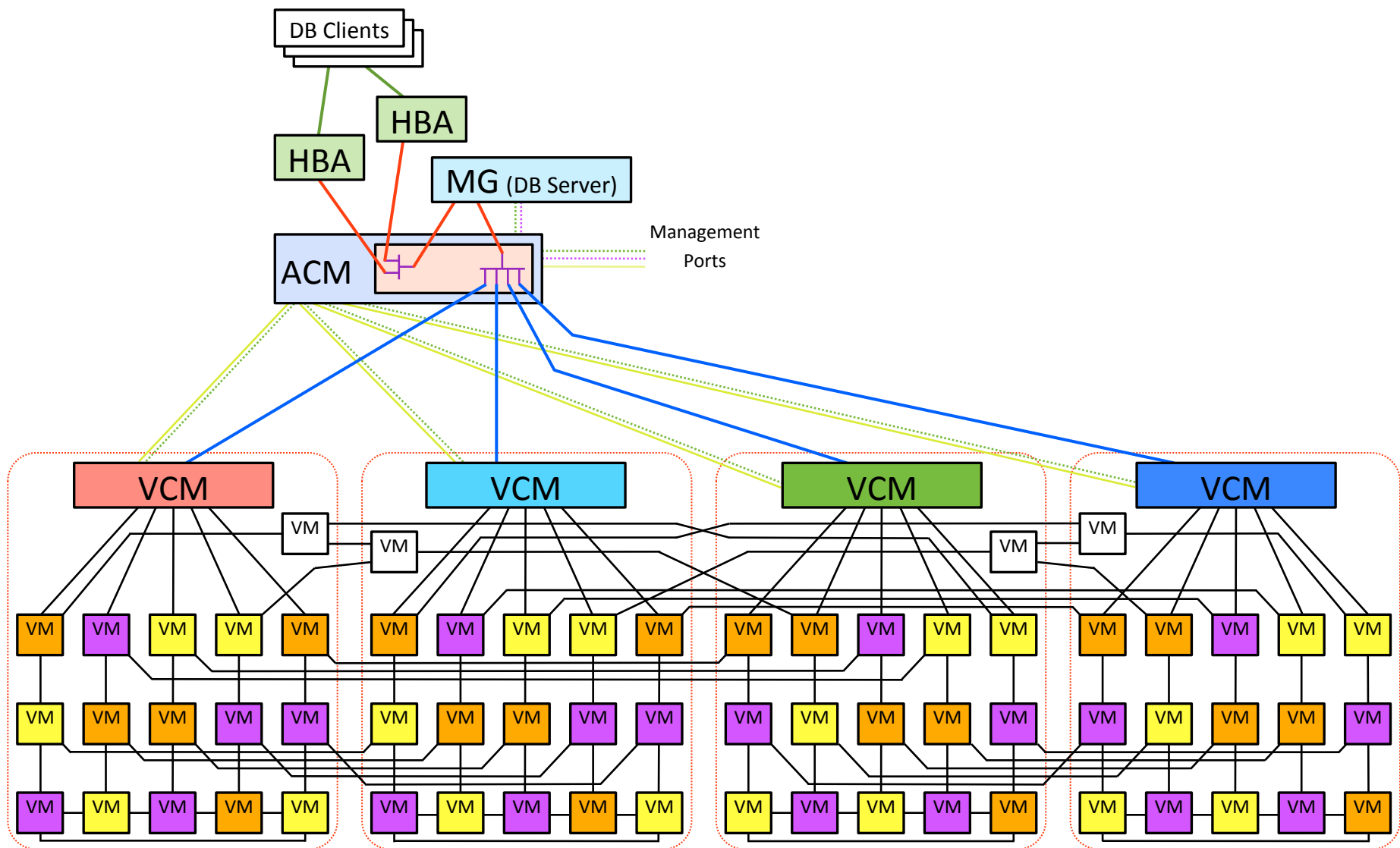
VXM Architecture



VXM Architecture



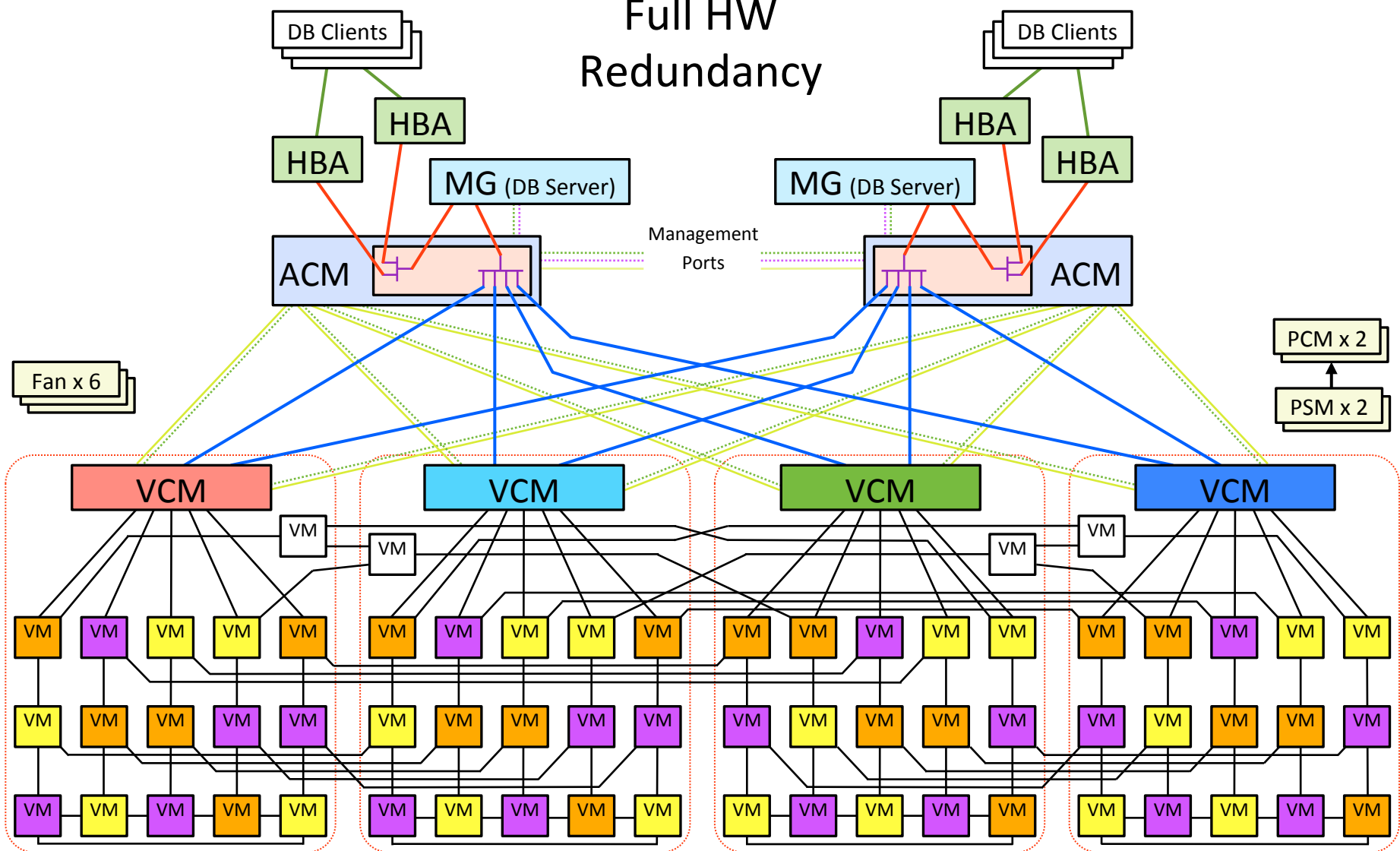
VXM Architecture



VXM Architecture



Full HW
Redundancy



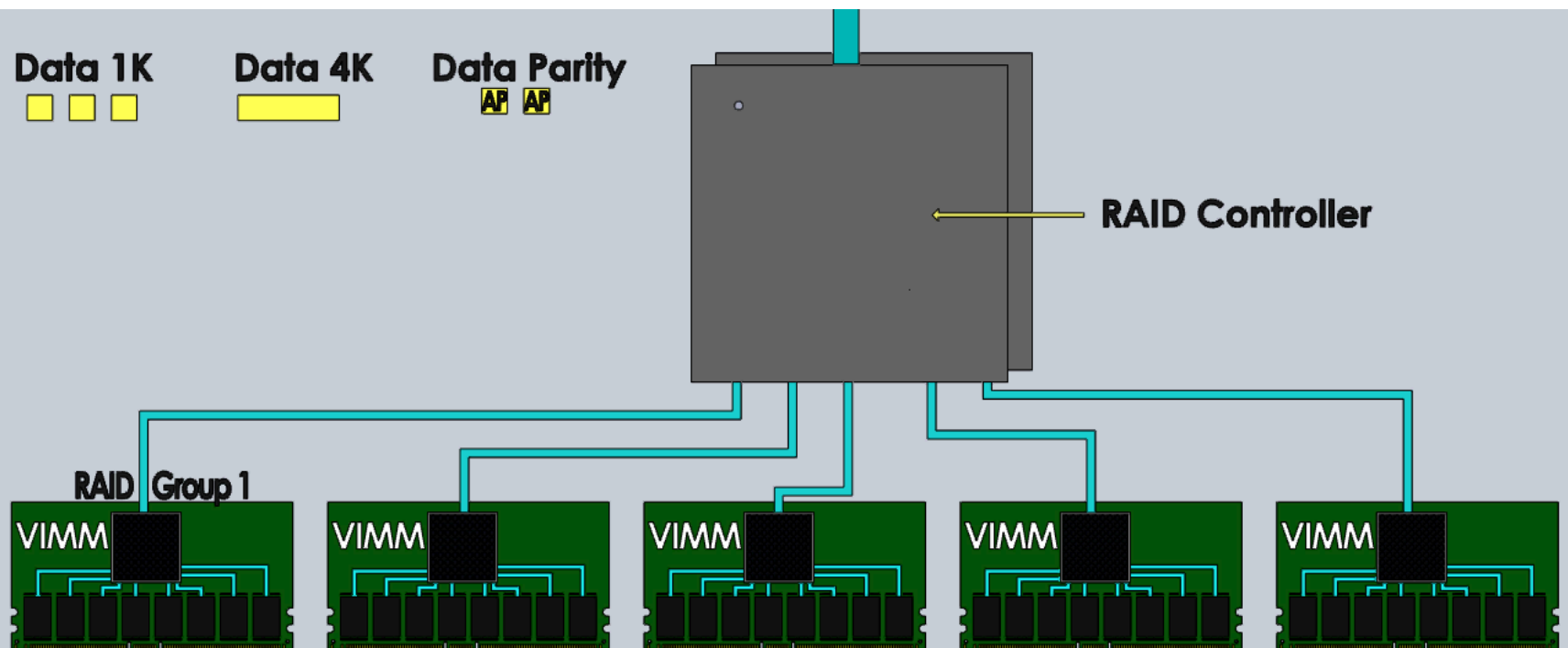
The power supply

violin
MEMORY

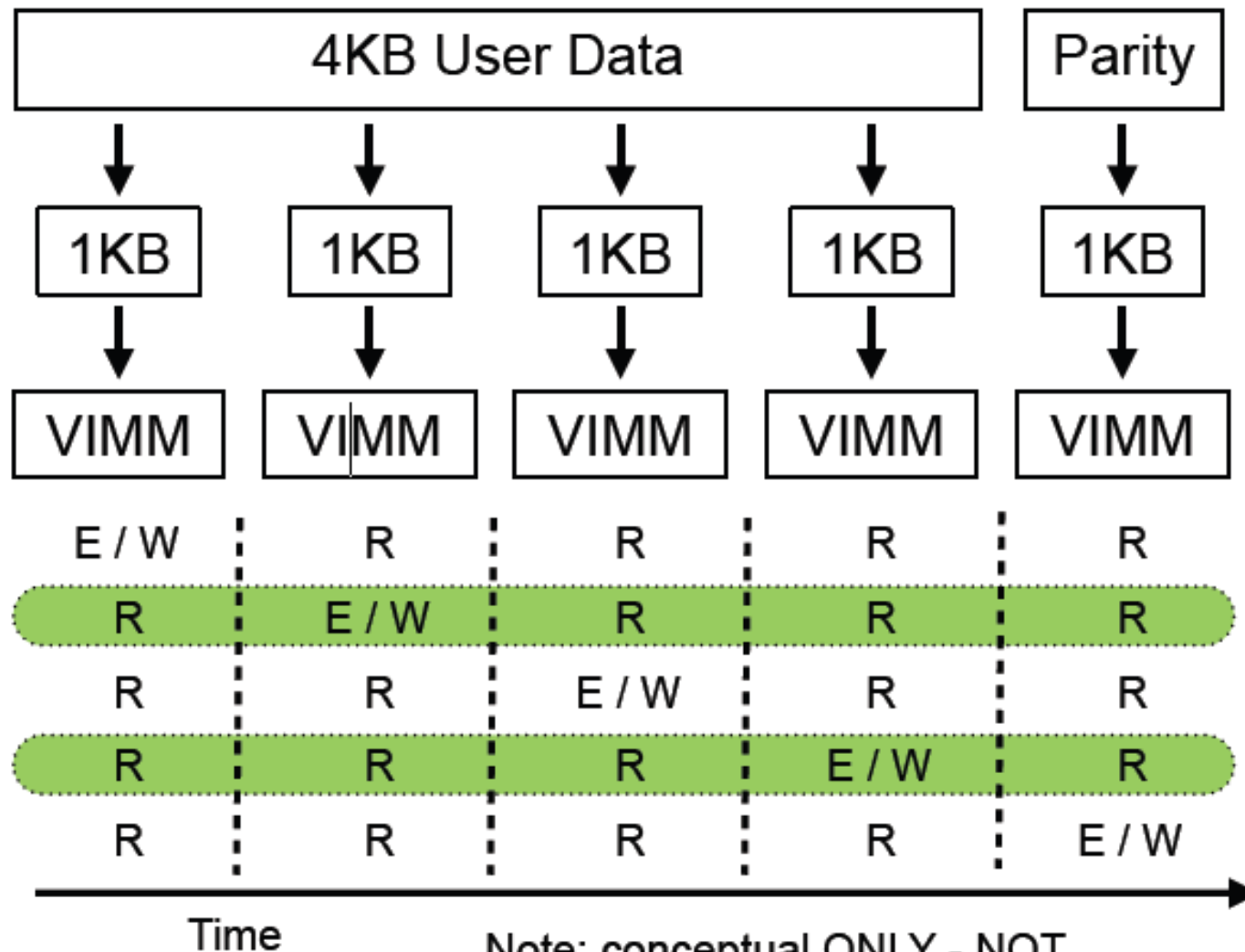


vRAID

So how does it work?

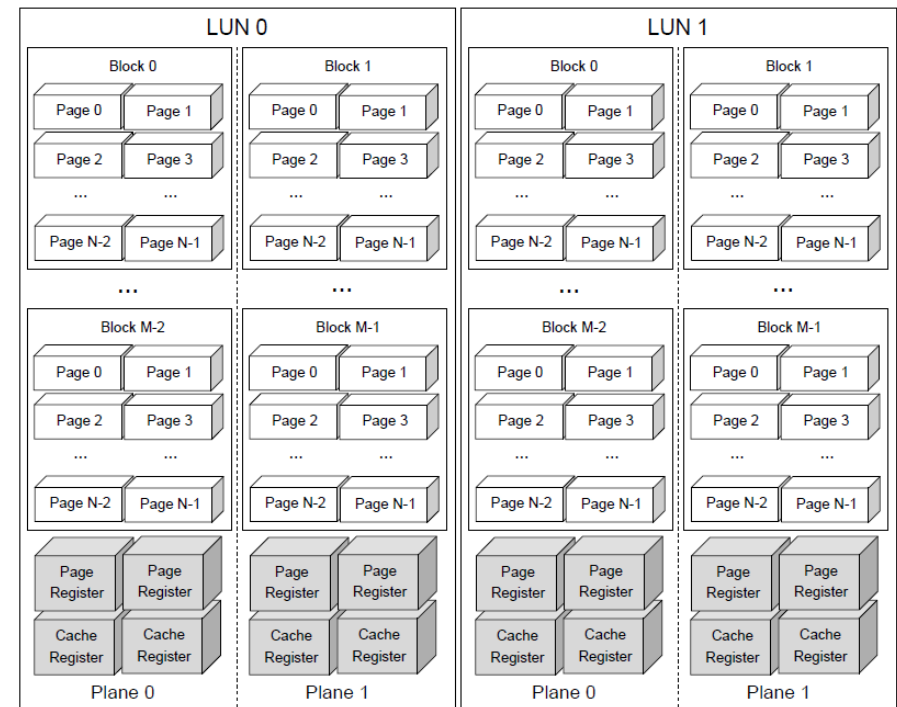
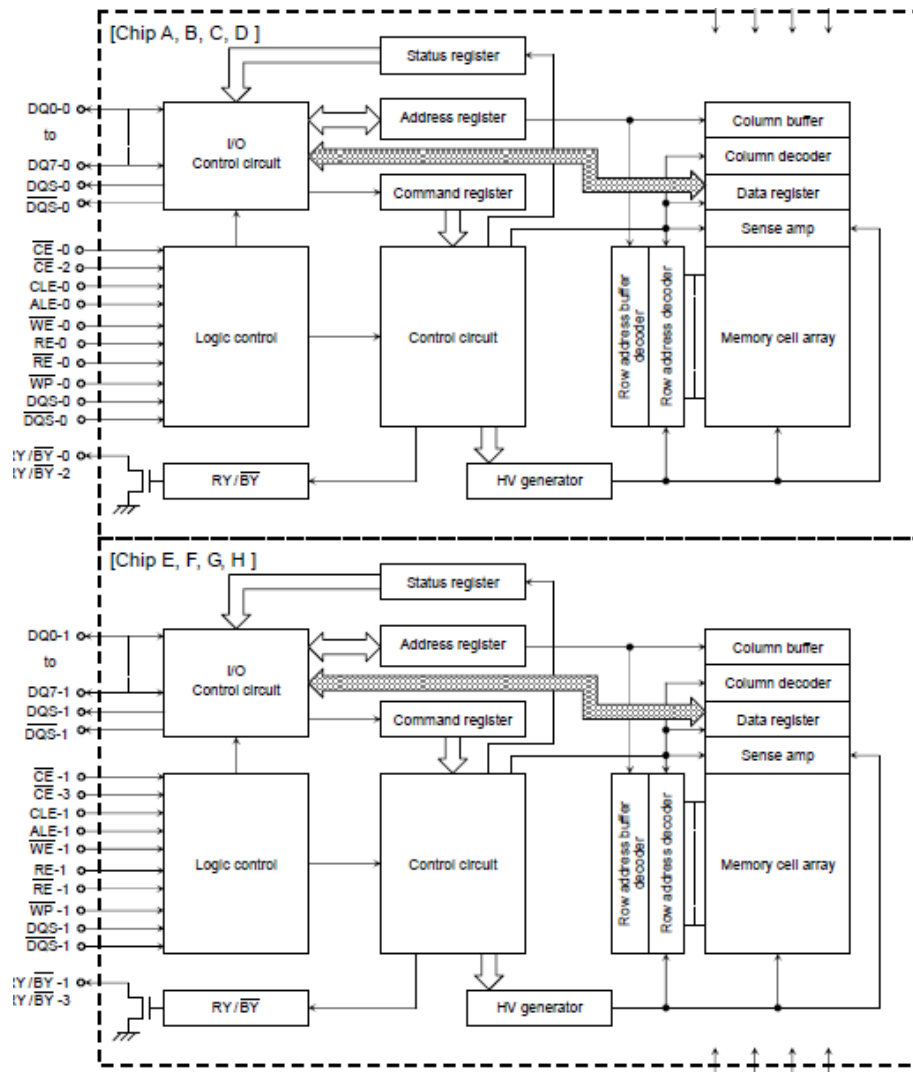


vRAID



Note: conceptual ONLY - NOT
actual implementation pattern!

How can it keep up while only writing one column at a time?

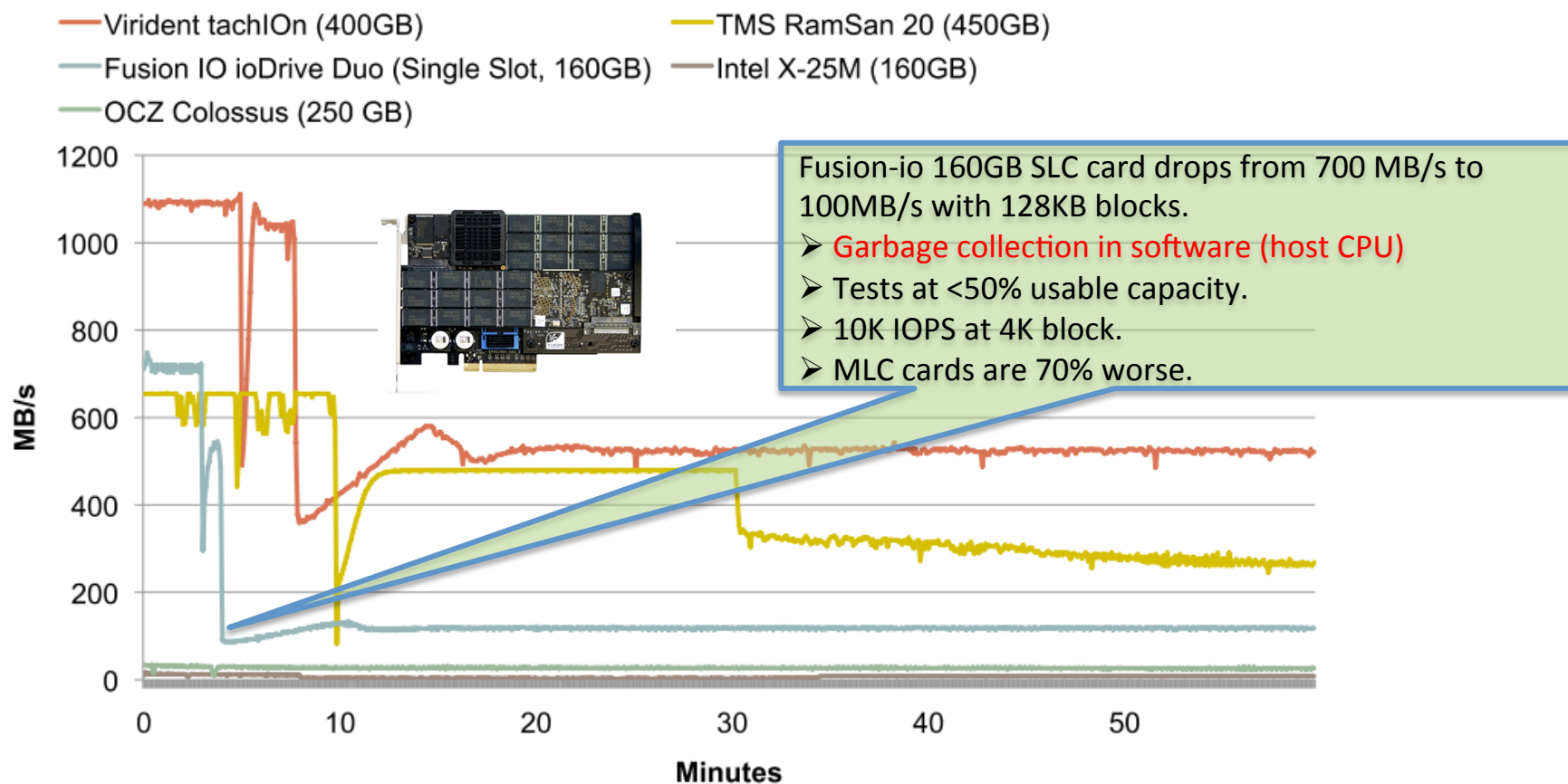


The write cliff

violin
MEMORY



Degradation - Bandwidth

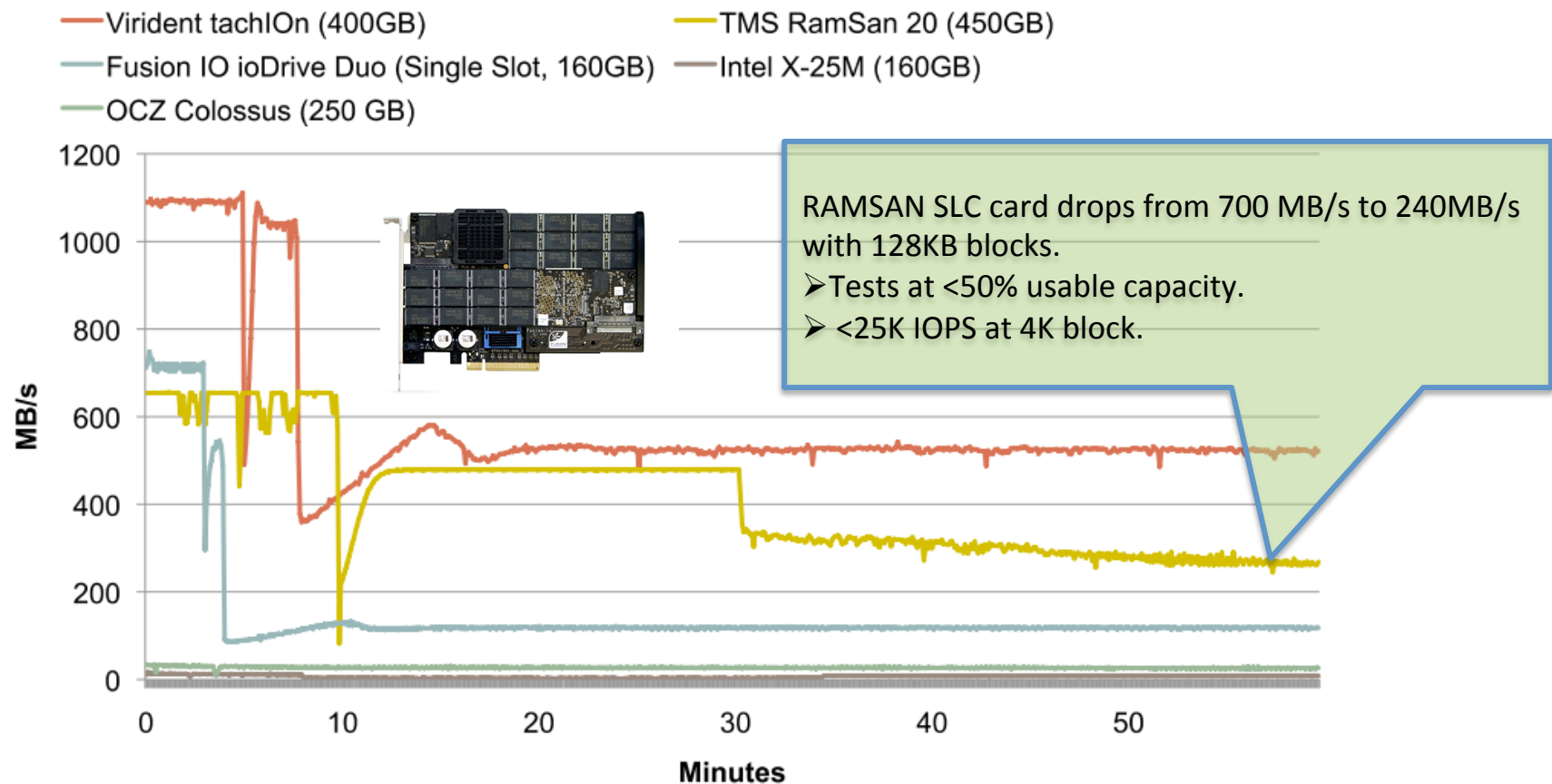


The write cliff

violin
MEMORY



Degradation - Bandwidth





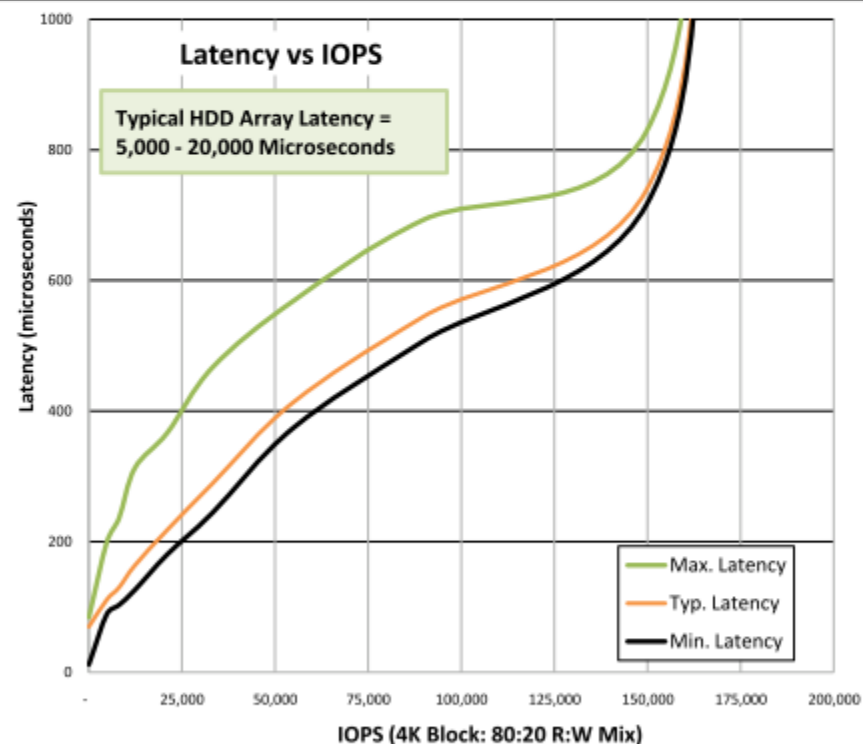
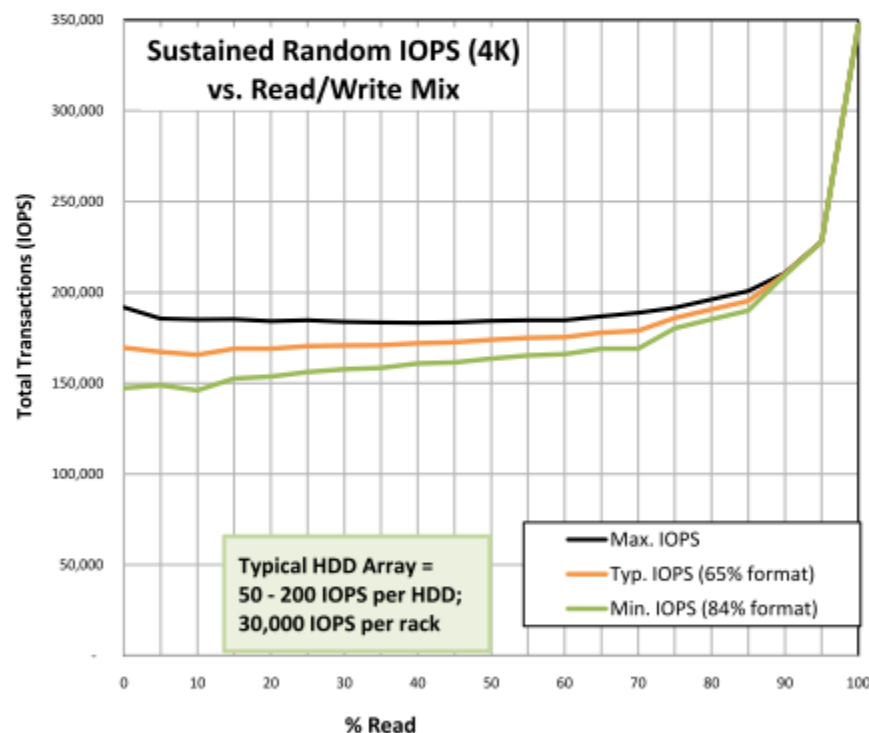
V-3220



Performance Report

18-Aug-11

System	Description	VIMM	VIMM	Raw Capacity		Random IOPS		Bandwidth	Latency	Formatted Capacity (GB)	Power W	Times	
		Type	Number	GiB(2 ³⁰)	GB(10 ⁹)	4K Block		(128K) MB/s	µs			Operation	min
V-3220	V-3200 Memory Array with 21.4TB SLC Flash Hot-swappable Flash/Power/Fans vRAID with PCIe8G1 Interface	SLC	84	21,504	23,090	Max. Read	347,000	1437	90	15,393@ 87%	856	Boot	9
		Size	Usable After RAID		Max. Write	192,000	782	20	11,545@ 65%	960	Format	60	
		256GB	64	16,384	17,592	Sust. Mixed	191,686	928	234	13,744@ 78%	940	Rebuild	128



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

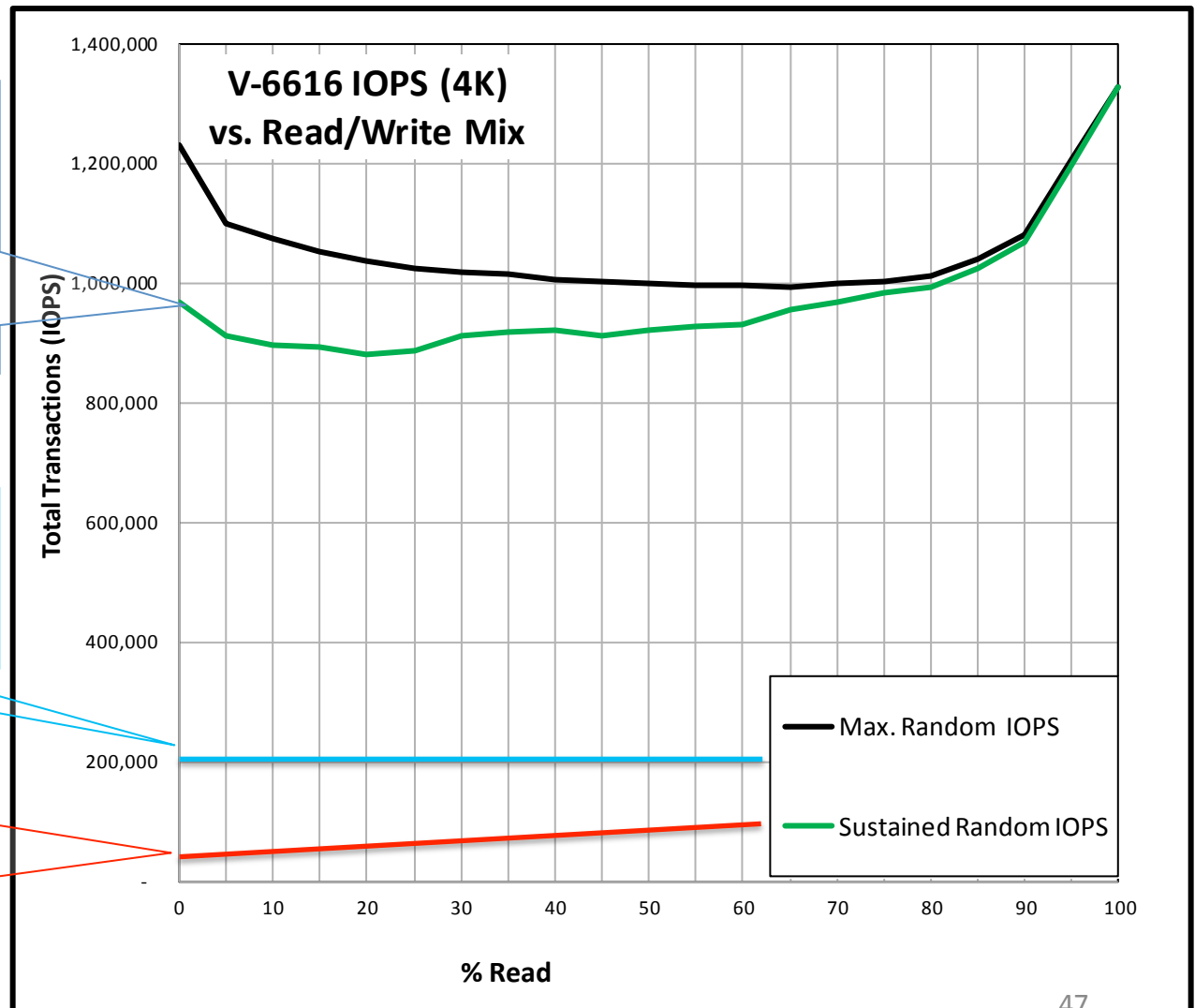
- Hardware-based GC
- Distributed operation
- vRAID: 4+1P
- Striping across 64 Modules
- 256GB SLC Modules
- ✓900K Sustained Write IOPS

•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



V6000: Sustained IOPS (MLC)



Sustained Random IOPS with RAID

•V6232

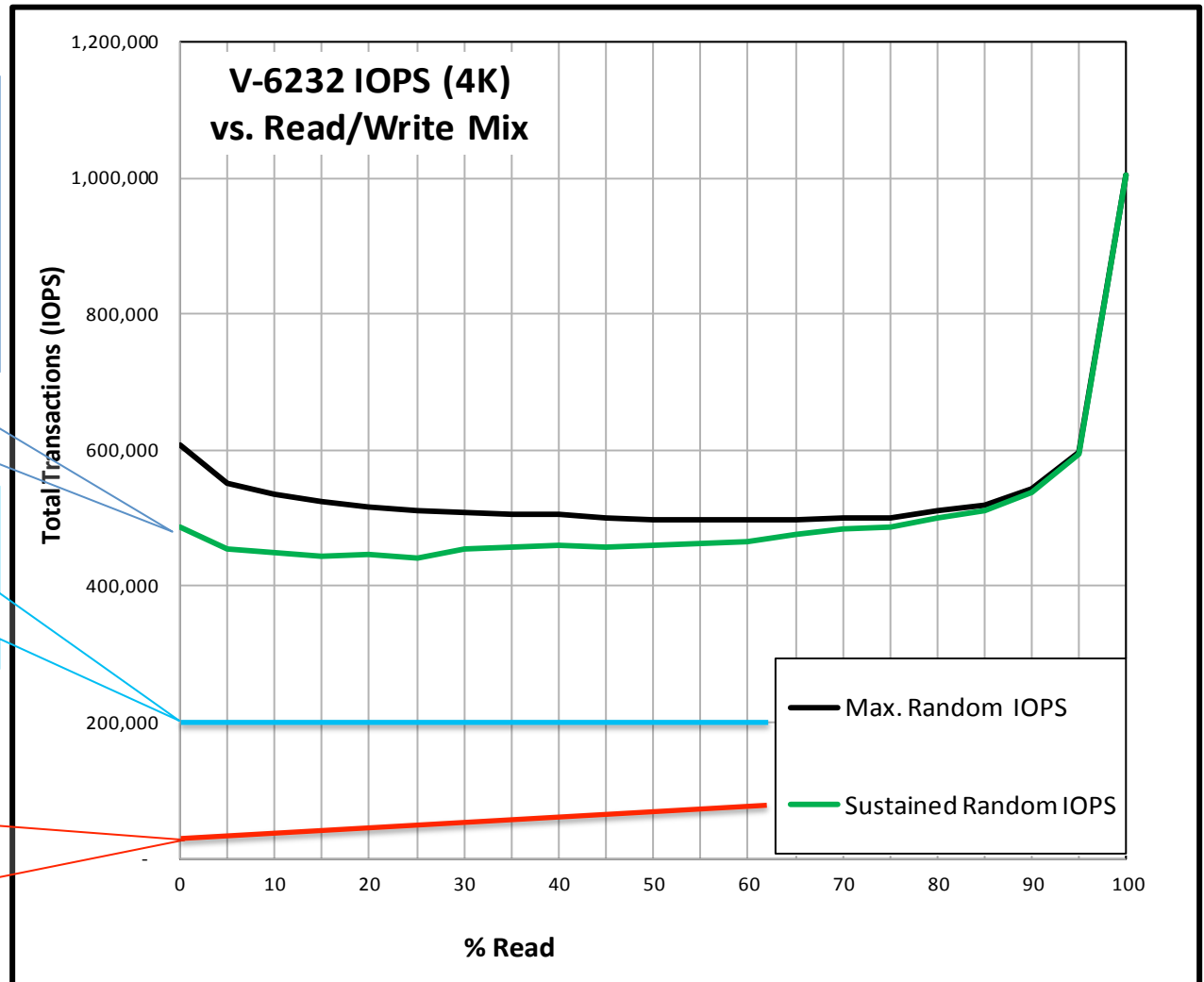
- Hardware-based GC
- Distributed operation
- vRAID: 4+1P
- Striping across 64 Modules
- 512GB MLC Modules
- ✓500K Sustained Write IOPS

•Disk Array

- 1000 15Krpm HDDs
- RAID-1
- ×200K Sustained IOPS

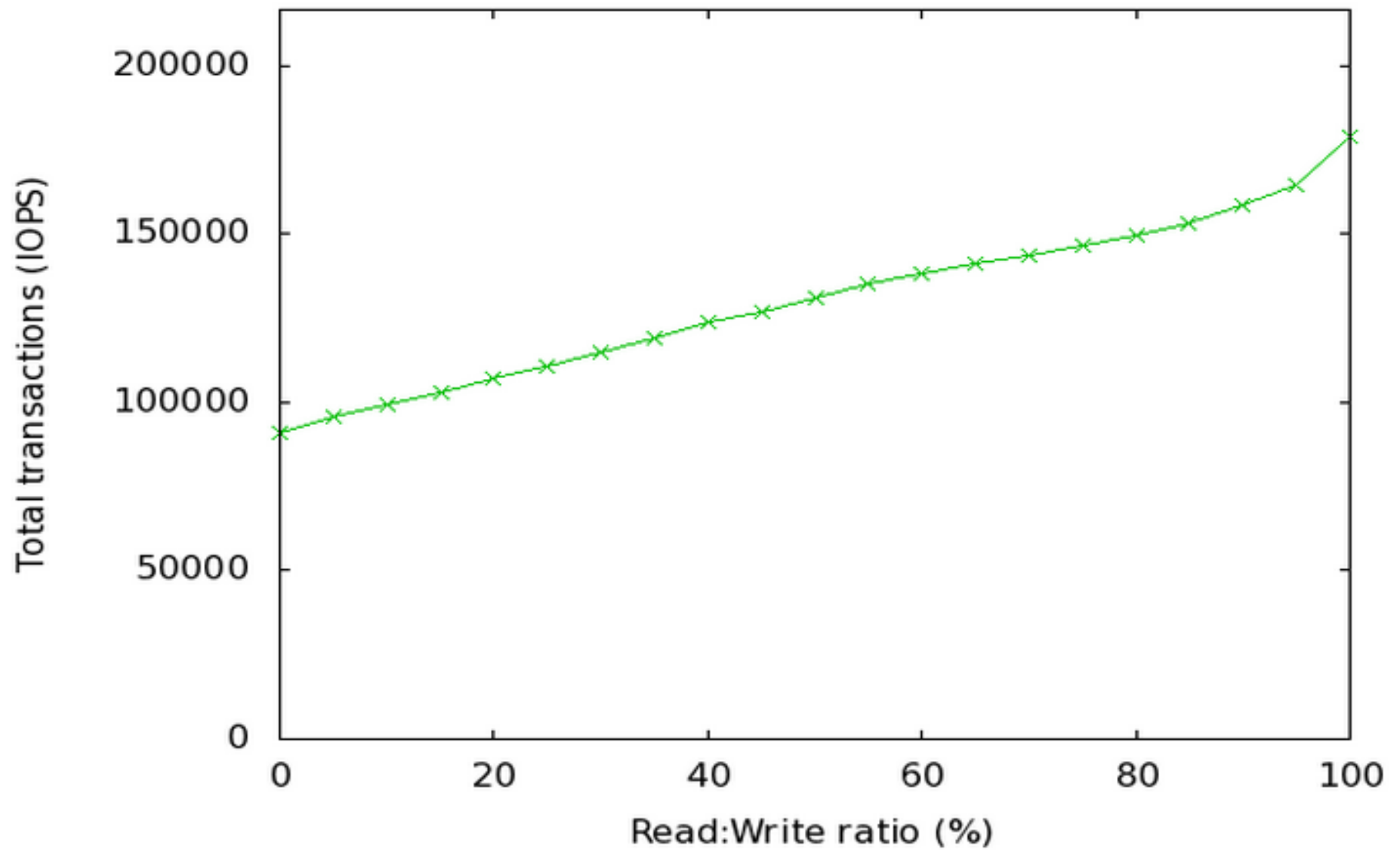
•PCIe card

- Software-based GC
- No RAID across modules
- ×10K Sustained Write IOPS



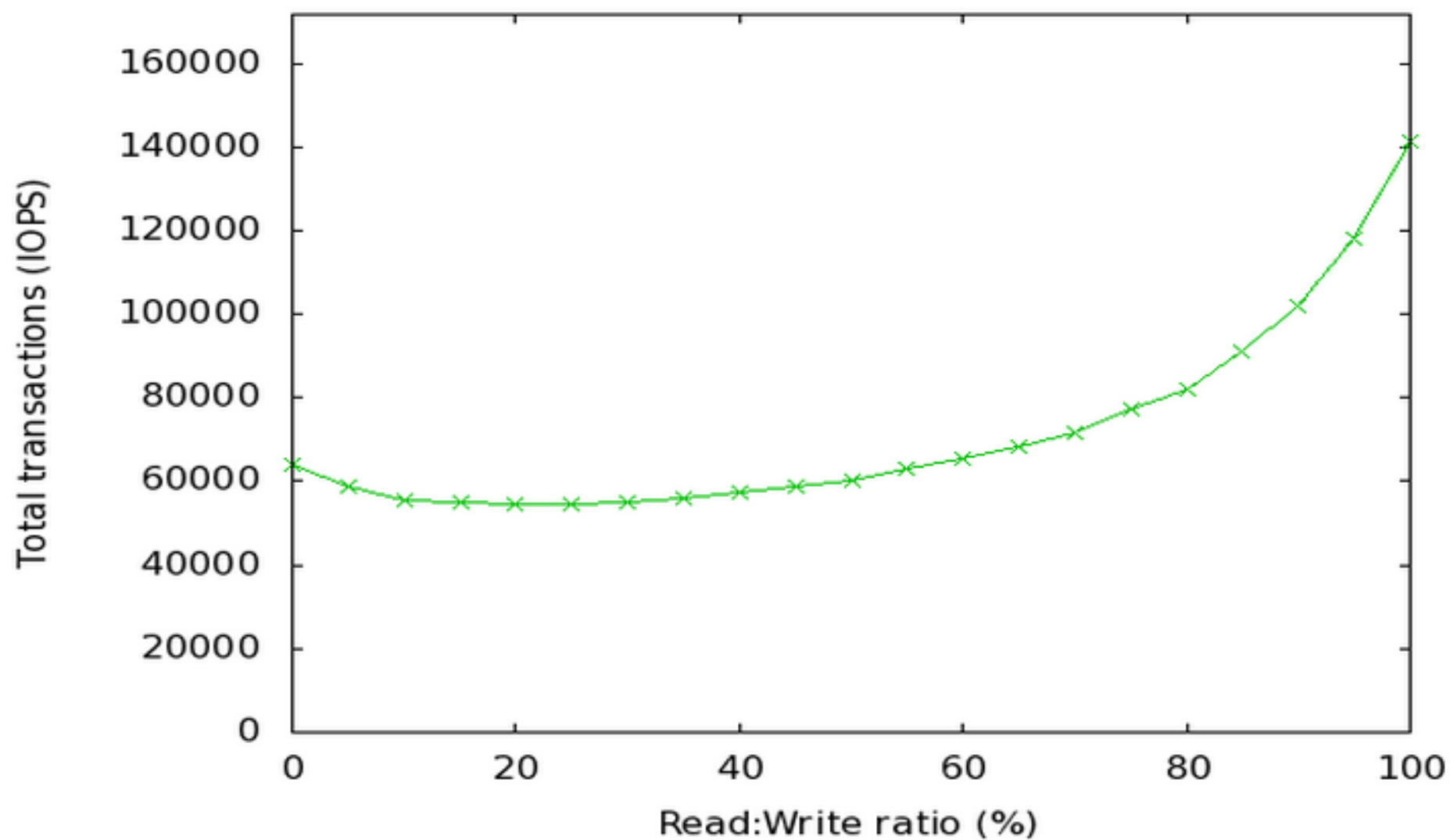
SLC

violin
MEMORY



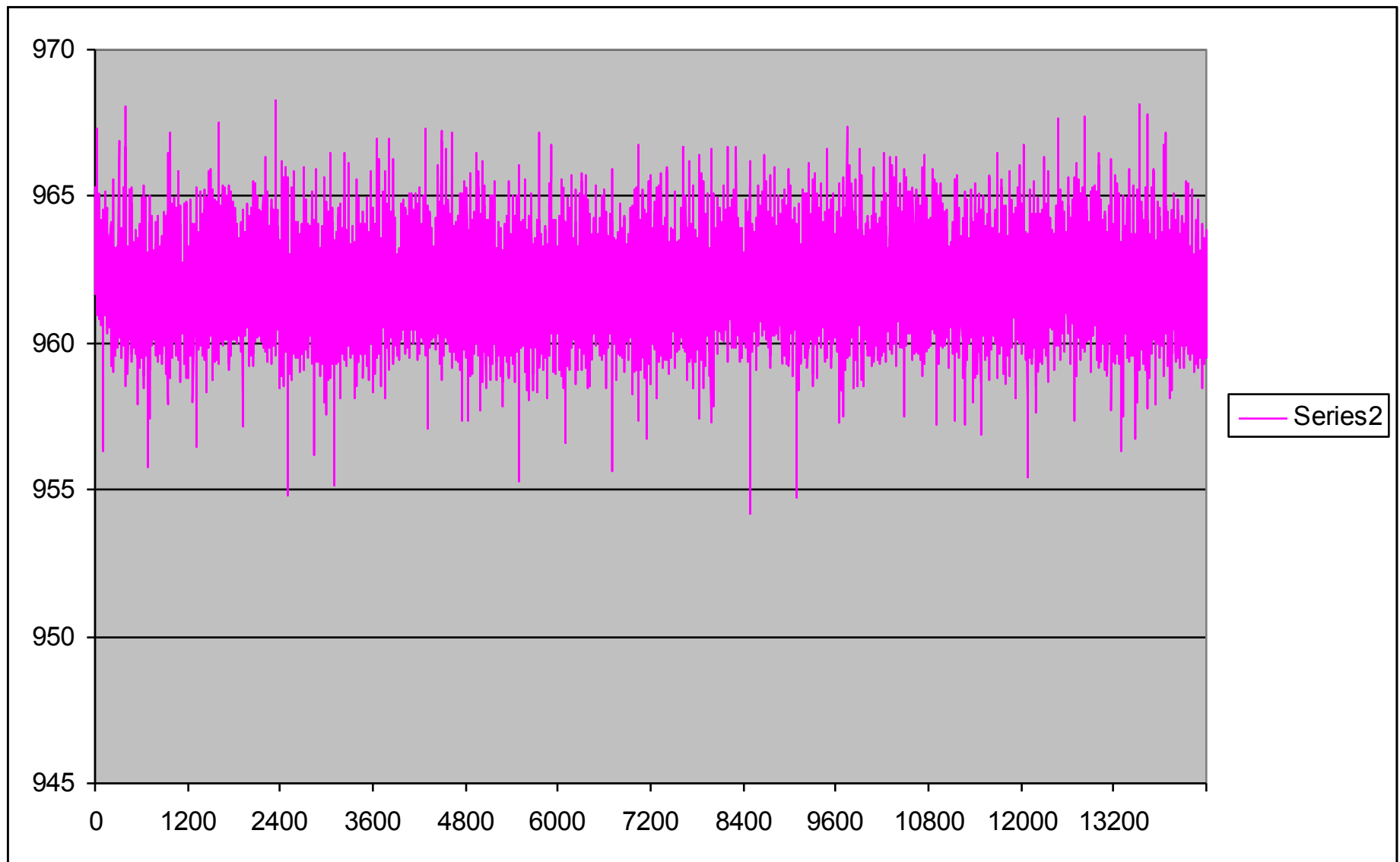
MLC

violin
MEMORY



Sustained random write 4 hour test (962MBps +/- 2%)

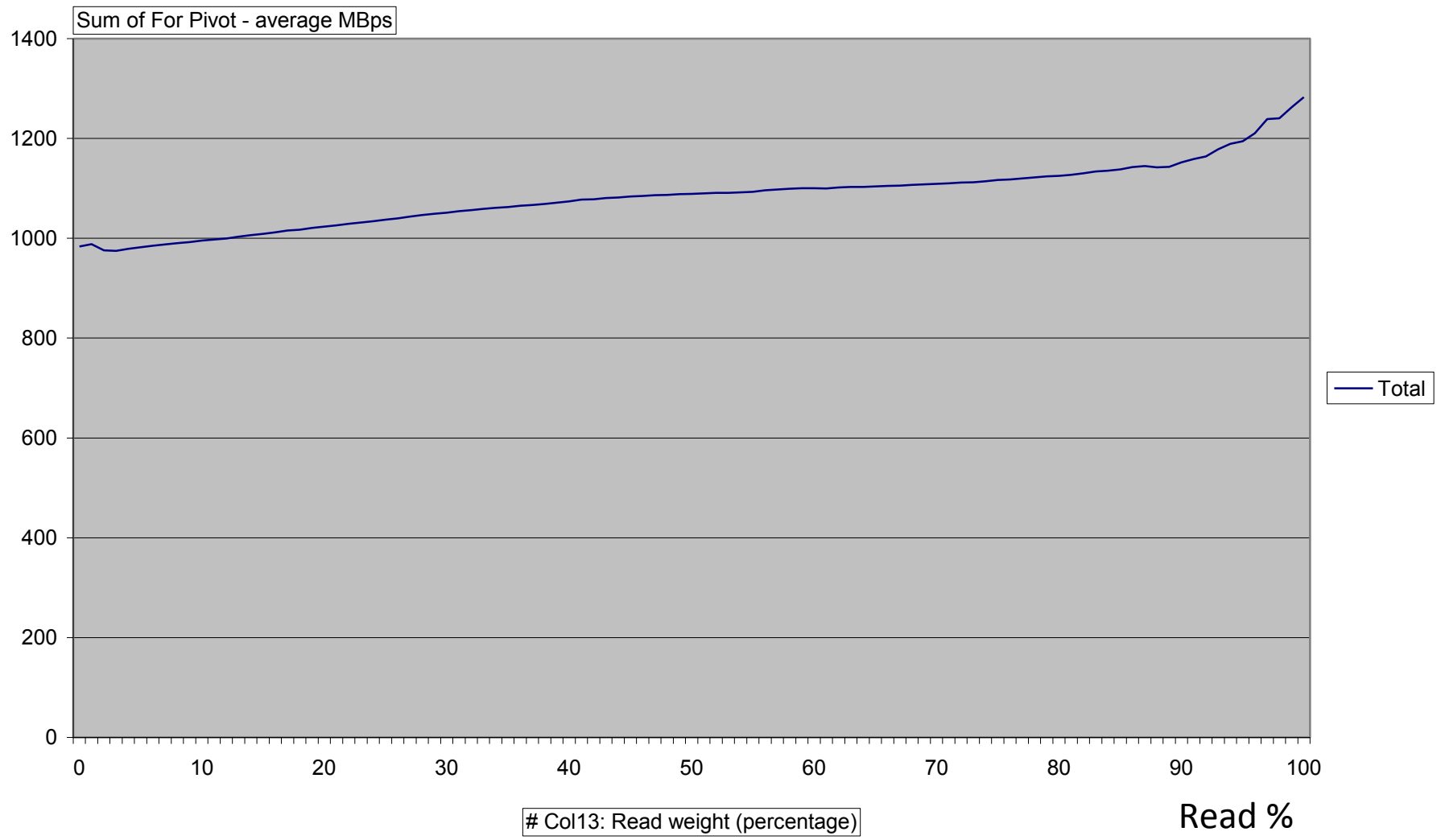
violin
MEMORY



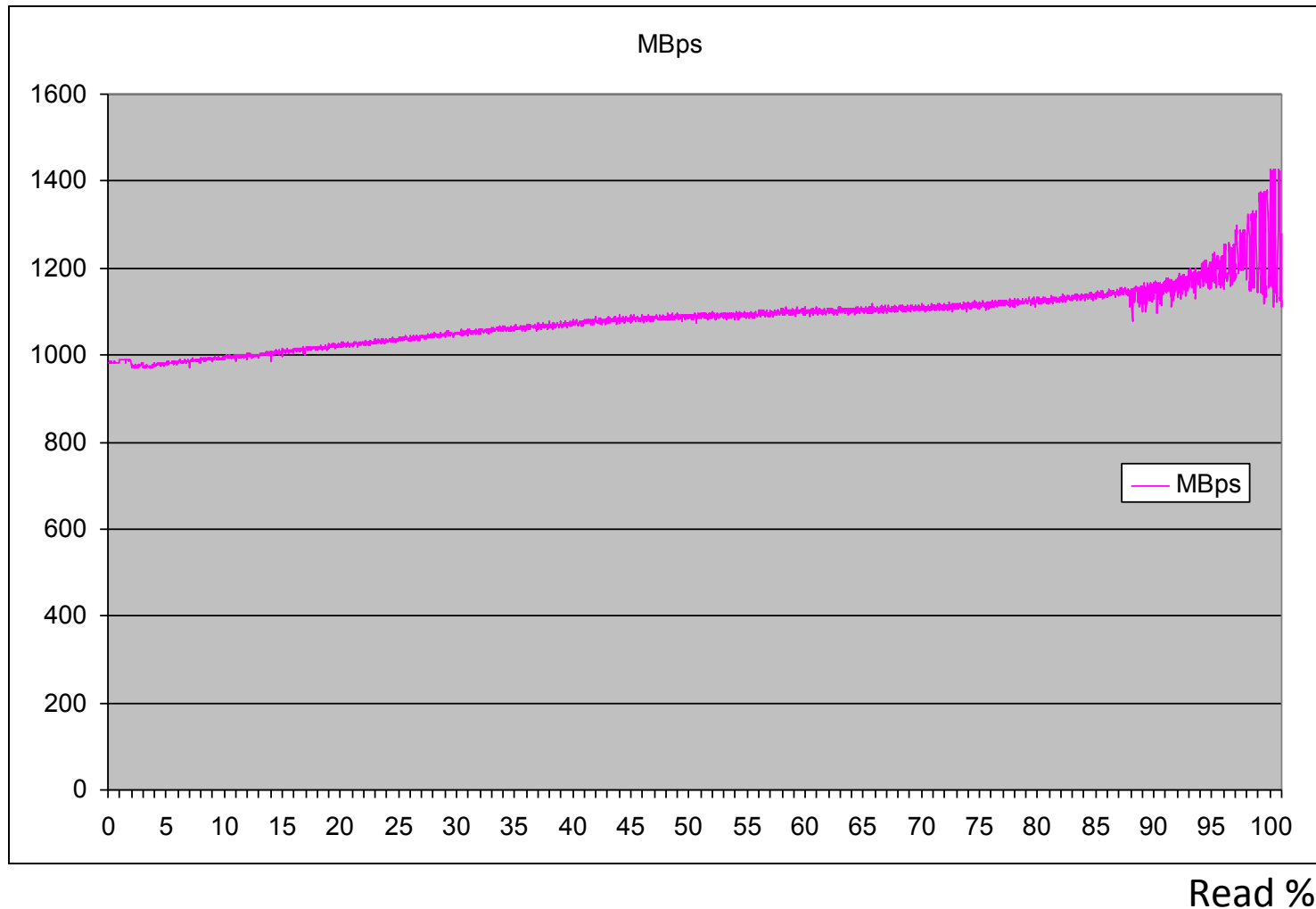
Ave Bandwidth vs R/W mix



Total



Bandwidth vs. R/W mix

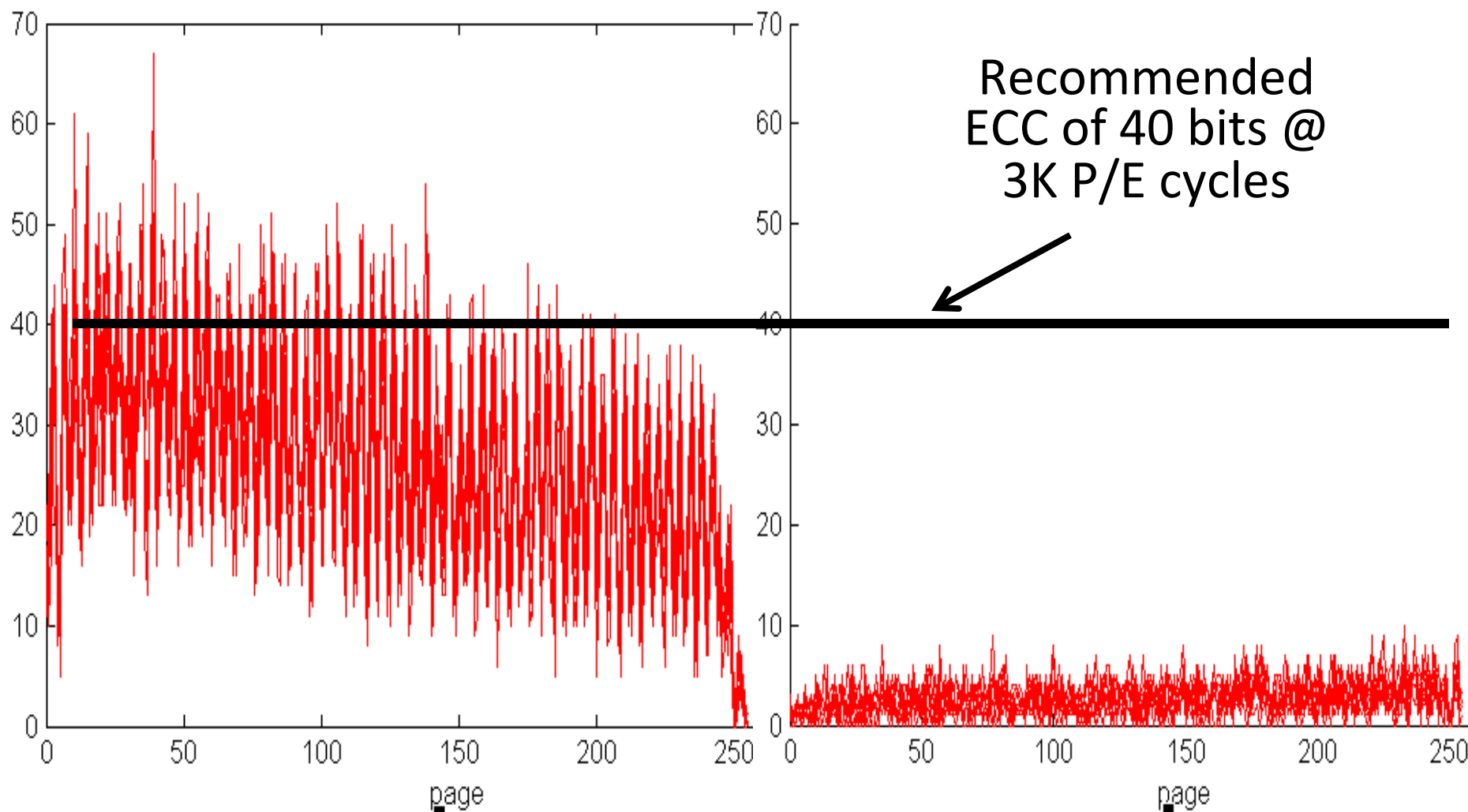


- Random workload, 256 threads, 4KB
- 1 hour, 40 minute (60 secs per %)
- Performance
 - Over 1GBps for most workload mixes
 - Variability @ high read % is due to aggressive preventative read-disturb data movement

Flash wears out



24nm MLC @10K accelerated P/E cycles



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.

- 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

Some people do get it....



From: ????????????????????????????????? 1/21/11

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.

Tell us how you really feel Linus.



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