You make **possible**

Cisco HyperFlex

Architecture Deep Dive, System Design and Performance Analysis

Brian Everitt, Technical Marketing Engineer
@CiscoTMEguy

BRKINI-2016

Barcelona | January 27-31, 2020

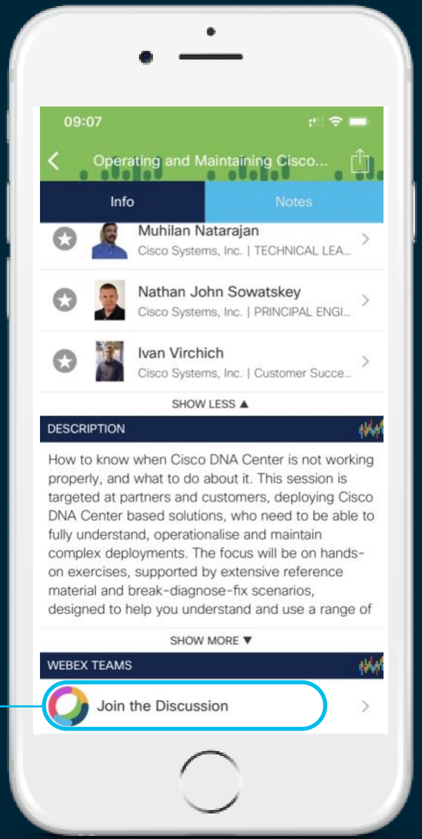
Cisco Webex Teams

Questions?

Use Cisco Webex Teams to chat with the speaker after the session

How

1. Find this session in the Cisco Events Mobile App
2. Click “Join the Discussion”
3. Install Webex Teams or go directly to the team space
4. Enter messages/questions in the team space



Agenda

- Introduction
- What’s New
- HyperFlex All-NVMe
- HyperFlex Acceleration Engine
- Benchmarking Tools
- Performance Testing
- Best Practices
- Conclusion

About Me

Brian Everitt

- 22 years experience as an infrastructure and storage engineer
- 4 years as a Technical Marketing Engineer with the CSPG BU
  - Focus on HyperFlex ESXi CVDs, performance, benchmarking and quality
- 5 years as a Cisco Advanced Services Solutions Architect
  - Focus on SAP Hana appliances and TDI, UCS and storage

What’s New?

What's New With Cisco HyperFlex 4.0?

HyperFlex Edge

2 node (and 4 node) deployments

Cisco Intersight cloud based witness service

Redundant 10GbE option

Next Gen Hardware

All-NVMe nodes

HyperFlex Acceleration Engine hardware offload card

4th Generation Fabric Interconnects and VIC

Cascade Lake CPUs and New drive options (4.0.2a)

Management

One click, full stack software and firmware upgrades

Storage Capacity Forecasts and Trends in Cisco Intersight

GUI license management in HX Connect (4.0.2a)

Technical and Performance Improvements in Cisco HyperFlex 4.0(2a)

- HyperFlex Boost Mode (more details later in this session)
- Dynamic cleaner improvements smooth out performance and reduce garbage
- Native replication scale grows from 200 to 1000 VMs per cluster
- Support for replication across low bandwidth connections down to 10Mbps
- New vCenter HTML5 plugin for cross launch of HX Connect
- Cisco Intersight based upgrades of ESXi for HX Edge clusters
- Health monitoring and alerting for controller VM health (memory use and capacity) and network services such as DNS, NTP and vCenter connectivity
- Pre-upgrade health and compatibility checks that can be run manually to check for issues ahead of time

Scaling Options in HXDP 4.0.2a

	VMware			HyperV	
	SFF (excluding All-NVMe)	LFF	All-NVMe	SFF	LFF
Converged Nodes	3-32	3-16	3-16	3-16	3-16
Compute-Only Nodes	0-32	0-32	0-16	0-16	0-16
	✓ Expansion Supported	✓ Expansion Supported	✓ Expansion Supported	✓ Expansion Supported	✓ Expansion Supported
Max ratio Compute to Converged Nodes*	2:1	2:1	1:1	1:1	1:1
Max Cluster Size	64	48	32	32	32
		! Hybrid Only	! Requires Enterprise License		! Hybrid Only

* 2:1 – Enterprise license (HXDP-P) if # of Compute > # of Converged Nodes.
* 1:1 – Standard license (HXDP-S) if # of Compute <= # of Converged Nodes.

cisco Live!

Scaling Options in HXDP 4.0.2a (cont.)

Edge

SFF
(excluding All-NVMe)

Converged Nodes
2-4

Expansion Not Currently Supported

1GbE and 10GbE Network Options Available

Single and Dual Switch Support

Deploy and Manage via Cisco Intersight

! HX*220 Nodes Only

Stretched Cluster**

SFF

Converged Nodes
2-16 / Site (4-32 / Cluster)

+

Compute-Only Nodes
0-21 / Site (0-42 / Cluster)
[Limited by Max Cluster size]

✓ Expansion Supported***

2:1
Max ratio Compute to Converged Nodes*

Max Cluster Size
32/Site | 64/Cluster

LFF

Converged Nodes
2-8 / Site (4-16 / Cluster)

+

Compute-Only Nodes
0-16 / Site (0-32 / Cluster)

✓ Expansion Supported***

2:1
Max ratio Compute to Converged Nodes*

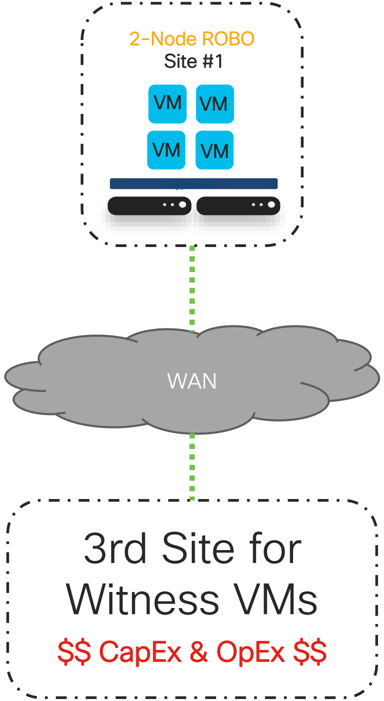
Max Cluster Size
24/Site | 48/Cluster

! Hybrid Only

* 2:1 – Enterprise license (HXDP-P) if # of Compute > # of Converged Nodes, 1:1 – Standard license (HXDP-S) if # of Compute <= # of Converged Nodes.
** Stretched cluster requires Enterprise license (HXDP-P) *** Requires uniform expansion of converged nodes across both sites

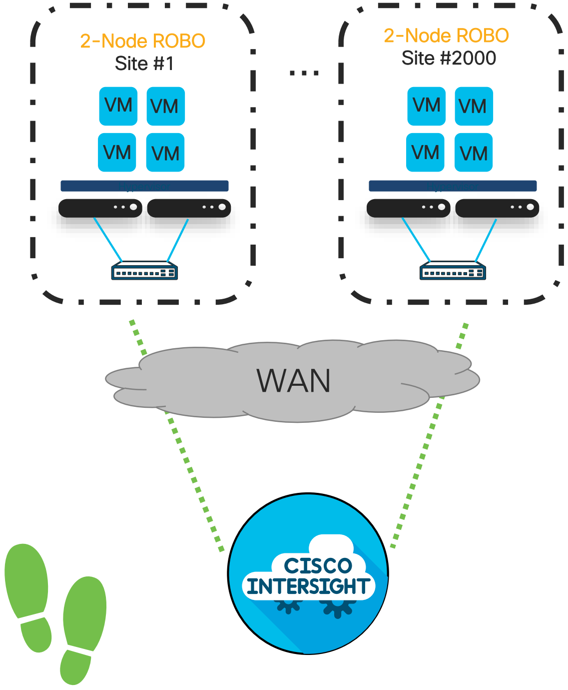
Invisible Cloud Witness, Why?

Traditional 2-Node Deployment



1. Consistency of the file system is essential.
2. To guarantee consistency of a file system a quorum is needed.
3. In a HyperFlex cluster a quorum requires a majority, or 2 of 3 nodes available and in agreement in a cluster.
4. Traditional 2-Node ROBO solutions require a 3rd site for witness VMs.
5. Cisco's innovation of the Intersight Invisible Cloud Witness virtually eliminates the costs with deploying 2-node HCI clusters/

Benefits of Invisible Cloud Witness Architecture

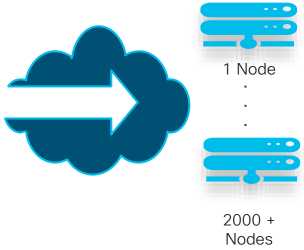


- No additional license cost; Included in an HX Edge Subscription
- No requirement for 3rd site or existing infrastructure
- Cloud-like operations - Nothing to manage
  - No user interface to monitor
  - No user driven software updates
  - No setup, configuration or backup required
  - No scale limitations
- Security
  - Realtime updates with the latest security patches
  - All communications encrypted via TLS 1.2
  - Uses standard HTTPS port 443; no firewall configuration required
- Built on efficient protocol stack
  - No periodic heartbeats
  - No cluster metadata or user data transferred to Intersight
  - Tolerates high latency and lossy WAN connections

Cisco HyperFlex Edge

with Full Lifecycle Management

Multi-Cluster Install



Invisible Cloud Witness Service

On-Going Management Including Full Stack Upgrades

Connected TAC Experience

HyperFlex All-NVMe

Cisco HyperFlex All-NVMe

- All-NVMe UCS C220 M5 Servers, optimized for All-NVMe
- Integrated Fabric Networking
- Ongoing I/O optimizations

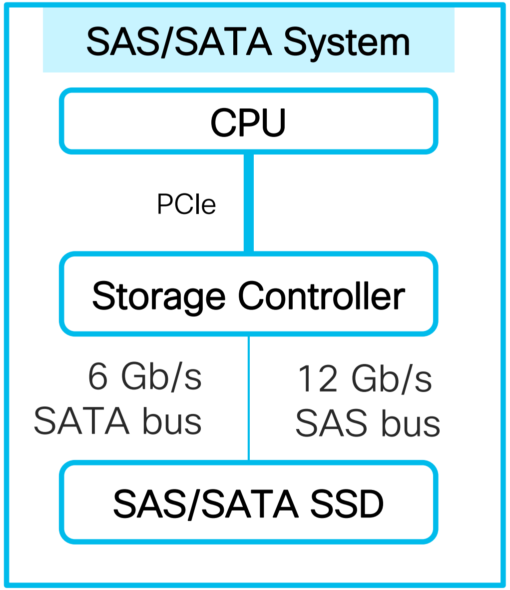
- Co-Engineered with Intel VMD for Hot-Plug & surprise removal cases
- Reliability Availability Serviceability (RAS) Assurance

- NVMe Optane Cache, NVMe Capacity
- Up to 32TB/Node, 40GB Networking
- Fully UCSM Managed for FW, LED etc.

- Higher IOPS
- Lower Latency
- Interop with HW acceleration to deliver the highest performance

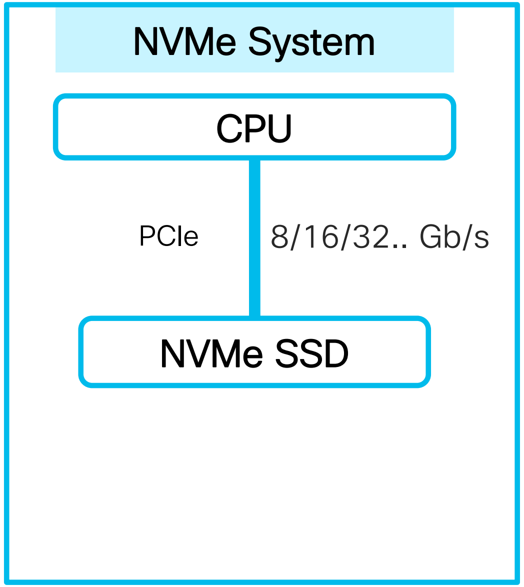
What’s Special About NVMe Drives

Unlocking the drive bottleneck



We speak ATA or SCSI

VS



We speak all new NVMe

Promise of Higher IOPs, Lower Latency & Less CPU cycles per IO

- NVMe is a new protocol and command set for PCIe connected storage
- Drives are still NAND or new 3D XPoint
- 65,535 queues vs.1

What delayed All-NVMe HCI from happening?

Many technical challenges needed to be overcome first…

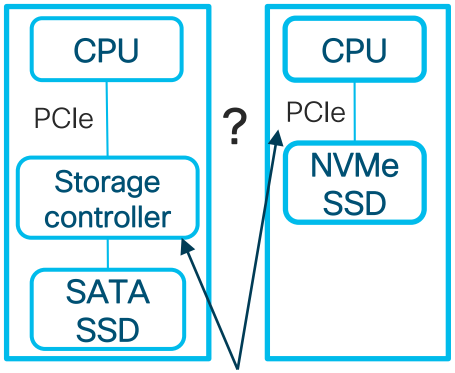
All-NVMe Hardware
Cisco HXAF220c-M5N

Specific all-NVMe design with PCIe M-switch

Overcome limited PCIe lanes for drives and peripherals

Intel Optane high endurance caching disks

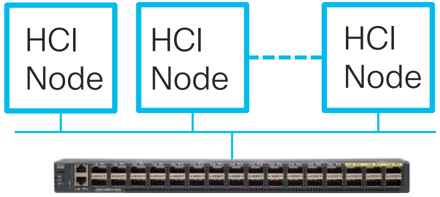
RAS for NVMe Drives
(Reliability, Availability, Serviceability)



RAS Delivery Point?

Hot plug, Surprise drive Removal, Firmware, LED, etc.

Capable Network
(To match drive capabilities)

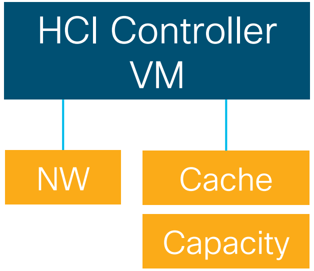


N/W is central to HCI, all writes traverse the network

Network bandwidth and latency both matter

Cisco UCS 10/25/40 GbE is the solution

HCI SW Optimization
(To leverage the Benefits)



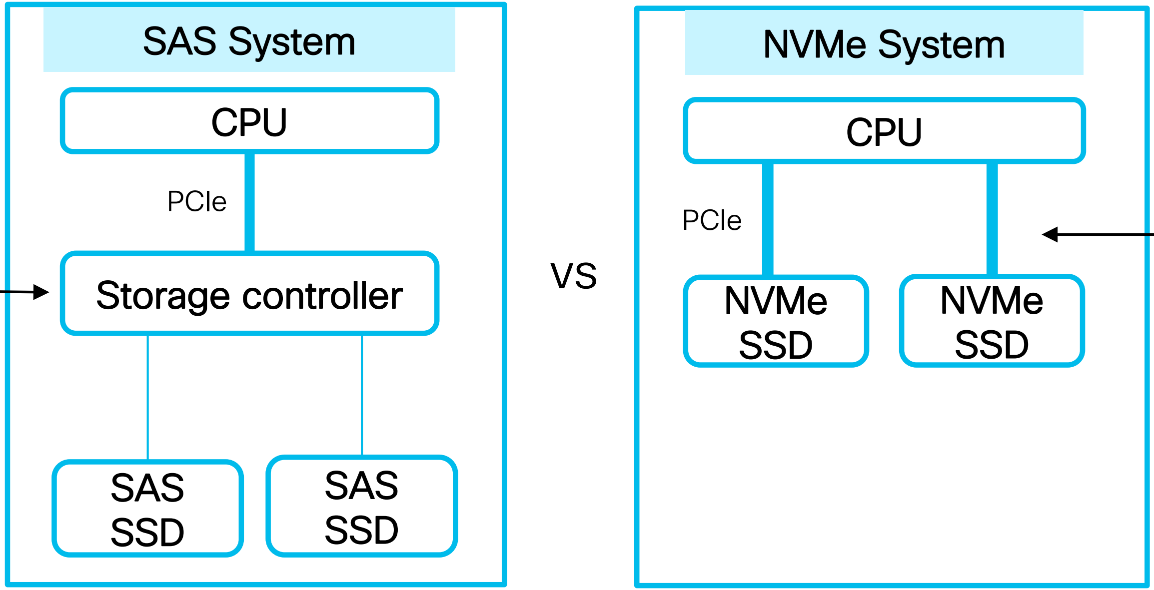
Is it architected to leverage the benefits?

Software Optimization in HX to both NW IO path and local IO path

RAS in SATA/SAS vs NVMe System

Hot Plug, Surprise Removal, Firmware Management, LED etc.

RAS Delivery point



No Controller

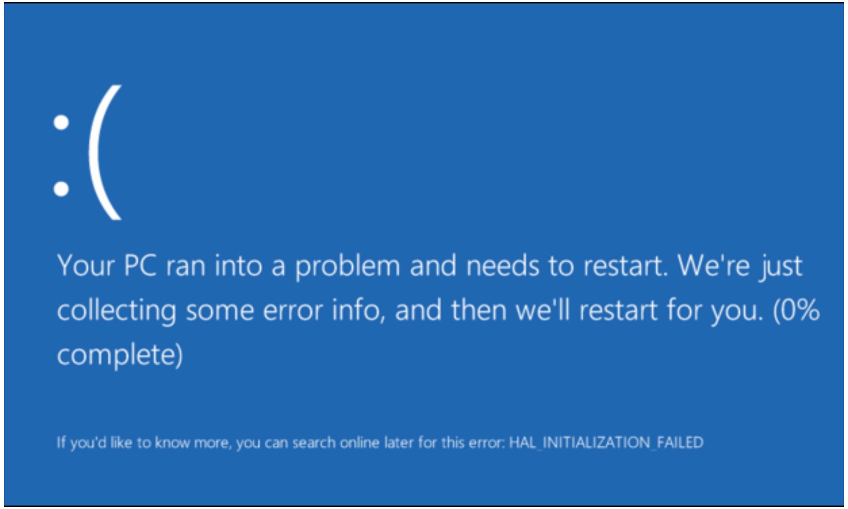
Surprise removal of NVMe drive may crash system

Hot Plug specification for NVMe not there yet

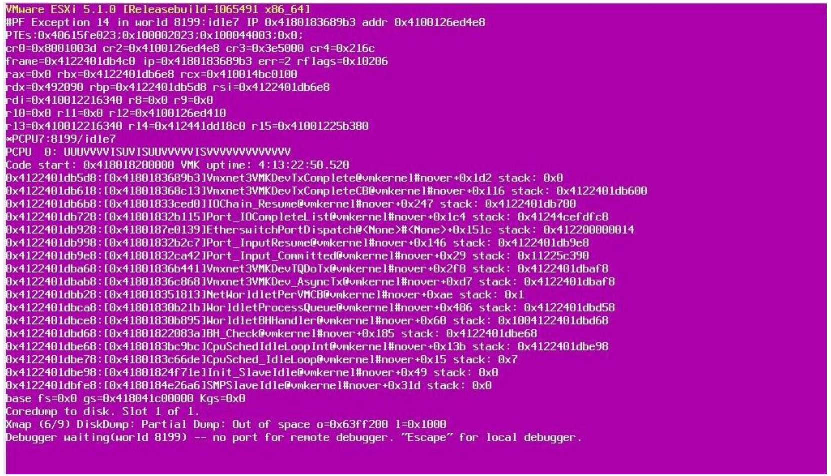
Absence of RAS Means..

BSOD? PSOD?

Hyper-V – Blue Screen due to surprise PCIe SSD removal



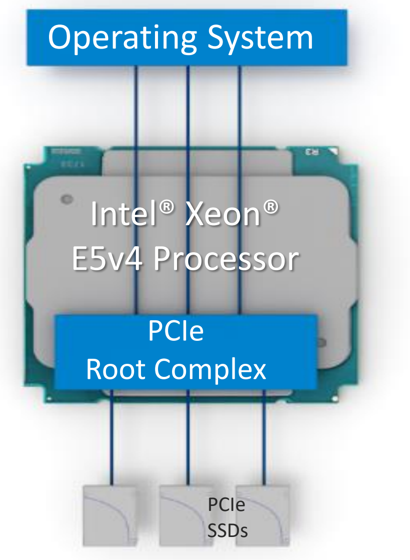
ESXi – Purple Screen due to surprise PCIe SSD removal



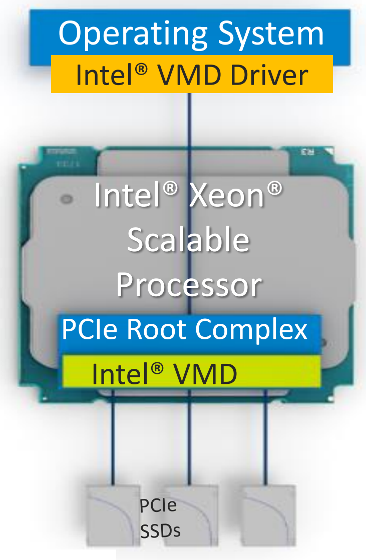
Intel Volume Management Device (VMD) To Rescue

Provides Means To Address HotPlug, Surprise Removal and LED

Before Intel® VMD



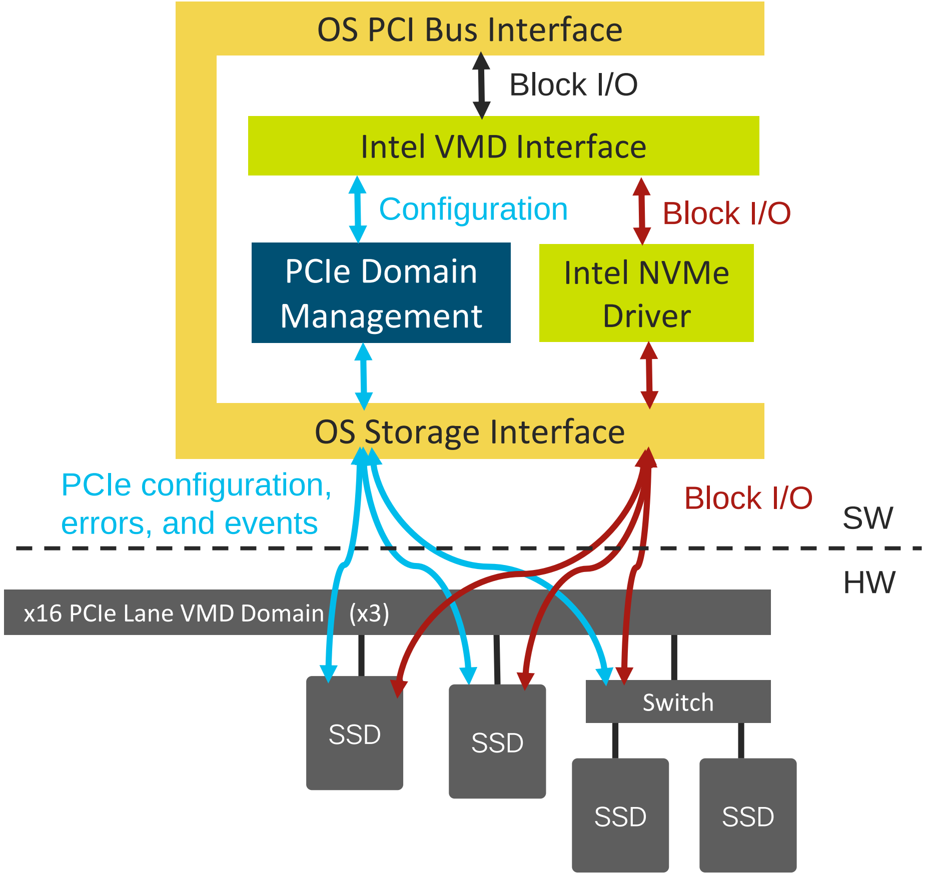
After Intel® VMD



- Intel® VMD is a new technology to enhance solutions with PCIe storage
- Supported for Windows, Linux, and ESXi
- Multi-SSD vendor support
- Intel® VMD enables:
  - Isolating fault domains for device surprise hot-plug and error handling
  - Provide consistent framework for managing LEDs

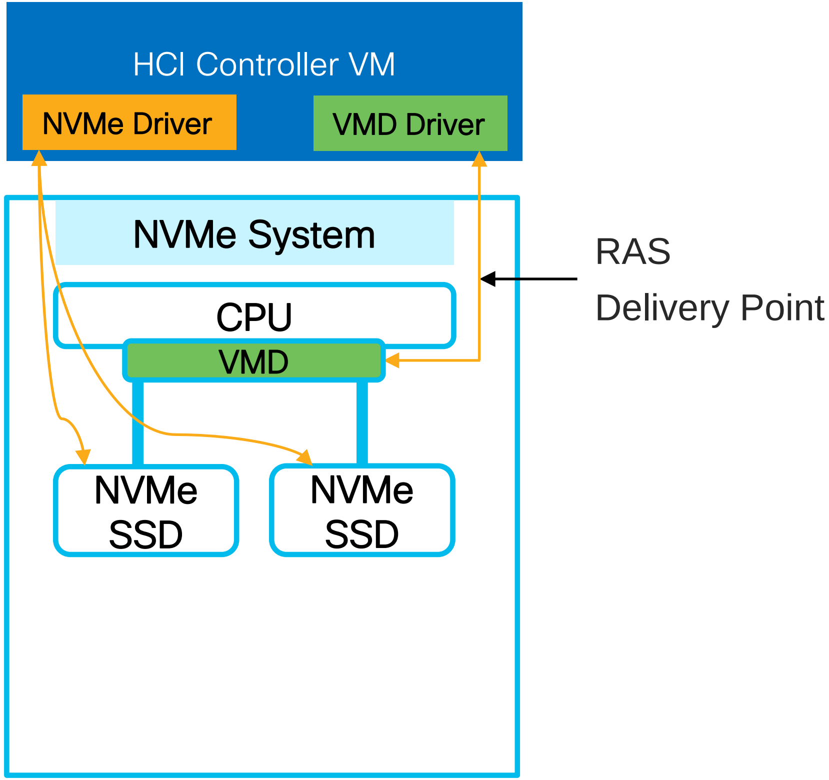
A Closer Look at Intel® VMD

- Intel VMD is an “integrated end point” in Intel® Xeon® Scalable Processor PCIe root complex.
- Maps entire sub PCIe trees (child devices) into VMD address space.
- VMD provides standard SW interface to set up/manage the domain (enumerate, event/error handling, LEDs), mostly out of the IO path.
- 3 VMD domains per Intel® Xeon® Processor, each managing 16 PCIe lanes



RAS in NVMe System With VMD

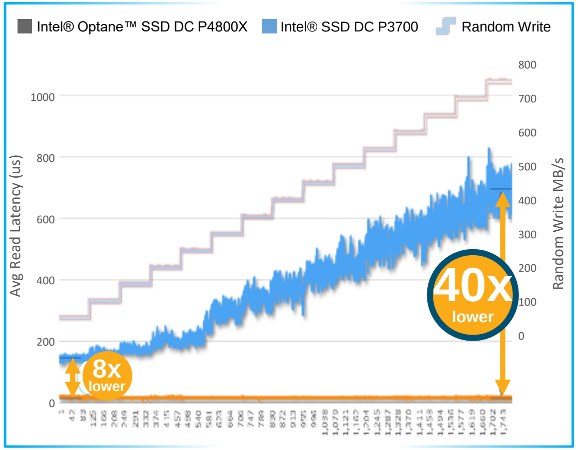
Hot Plug, Surprise Removal, Firmware Management, LED etc.



Intel® Optane™ SSD DC P4800X The Ideal Caching Solution

Lower and more consistent latency

Average Read Latency under Random Write Workload[1]



Higher endurance

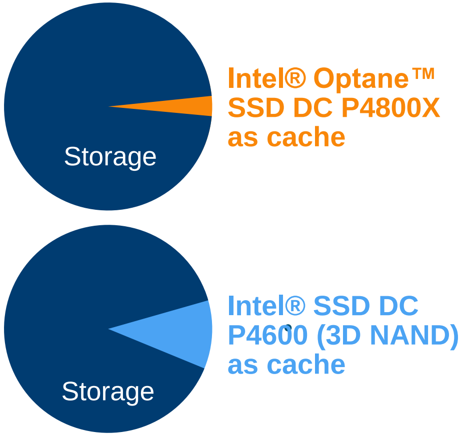
Drive Writes Per Day (DWPD)[2]

Intel® Optane™ SSD DC P4800X — **60 DWPD**

Intel® SSD DC P4600 (3D NAND) — **3.0 DWPD**

More efficient

Cache as a % of Storage Capacity[3]



Lower latency + higher endurance = greater SDS system efficiency

1. Responsiveness defined as average read latency measured at queue depth 1 during 4k random write workload. Measured using FIO 2.15. Common Configuration - Intel 2U Server System, OS CentOS 7.2, kernel 3.10.0-327.el7.x86_64, CPU 2 x Intel® Xeon® E5-2699 v4 @ 2.20GHz (22 cores), RAM 396GB DDR @ 2133MHz. Configuration – Intel® Optane™ SSD DC P4800X 375GB and Intel® SSD DC P3700 1600GB. Latency – Average read latency measured at QD1 during 4K Random Write operations using fio-2.15. Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system.
2. Source – Intel Data Sheets
3. Source – Intel: General proportions shown for illustrative purposes.

HyperFlex All NVMe Delivers Higher Performance

Engineered and Optimized All NVMe

General Workloads	Complex Mixed Workload	Databases
28-57% Higher IOPS	66-90% Higher IOPS	57-71% Higher IOPS
22-40% Lower Total Latency	31-39% Lower Total Latency	34-37% Lower Total Latency

Note: All-NVMe numbers relative to HyperFlex All-Flash

ESG Report: https://research.esg-global.com/reportaction/ciscohyperflexallnvme/Toc?
Note: Results are not to be used as an application or cluster performance sizer. Leverage the HX Sizer

HyperFlex I/O Data Paths

Learn on your own?

- 10 Slides of additional content hidden in this deck
- Slides are viewable in the downloaded PDF version
- Covers the HX read and write I/O data pathways and caching functions
- Content was shown in sessions with the same ID at Cisco Live Las Vegas 2017, Barcelona 2018 and Orlando 2018

Cisco live!

On-Demand Library | Online Events | In-Person Events | About Cisco Live

On-Demand Library

FILTERS

+ EVENT
+ FEATURED
+ TECHNOLOGY
+ PROGRAM
+ FOCUS
+ SOLUTIONS
+ TECHNICAL LEVEL

Search by Keyword

Show Favorites Only

Clear

1 to 3 of 3 results found

Cisco HyperFlex Architecture Deep Dive, System Design and Performance Analysis - BRKINI-2016

Event: 2018 Orlando

Brian Everitt, ENGINEER.TECHNICAL MARKETING

Cisco HyperFlex Design, Performance and Benchmarking for All-Flash and NVMe Clusters - BRKINI-2016

Event: 2018 Barcelona

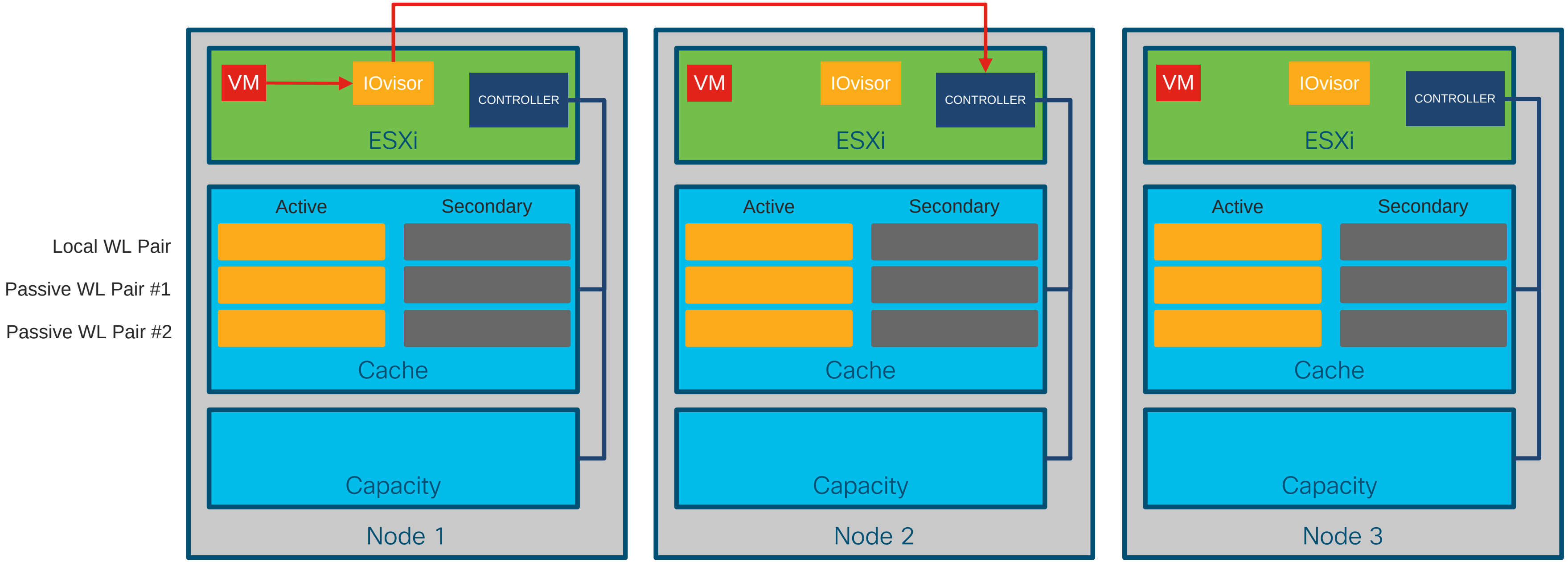
Brian Everitt, ENGINEER.TECHNICAL MARKETING

Cisco HyperFlex Design, Performance and Benchmarking for All-Flash and NVMe Clusters - BRKINI-2016

Event: 2017 Las Vegas

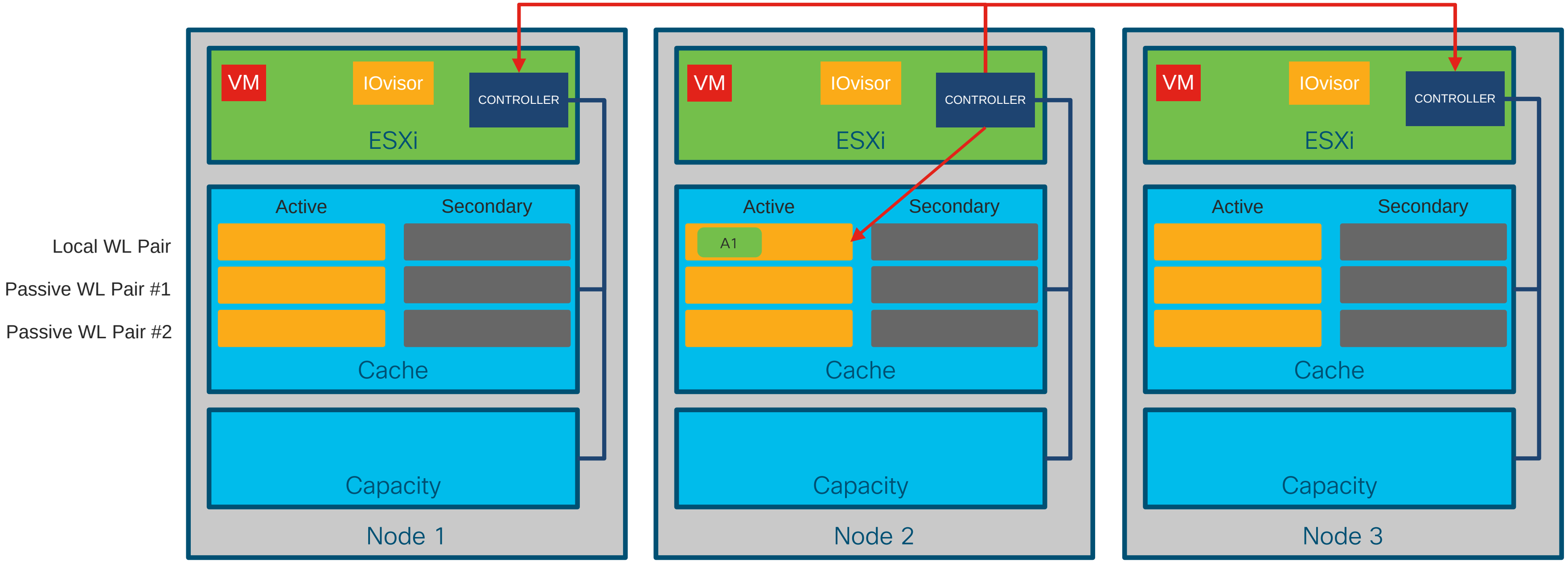
Brian Everitt, ENGINEER.TECHNICAL MARKETING

Input/Output Lifecycles: Writes



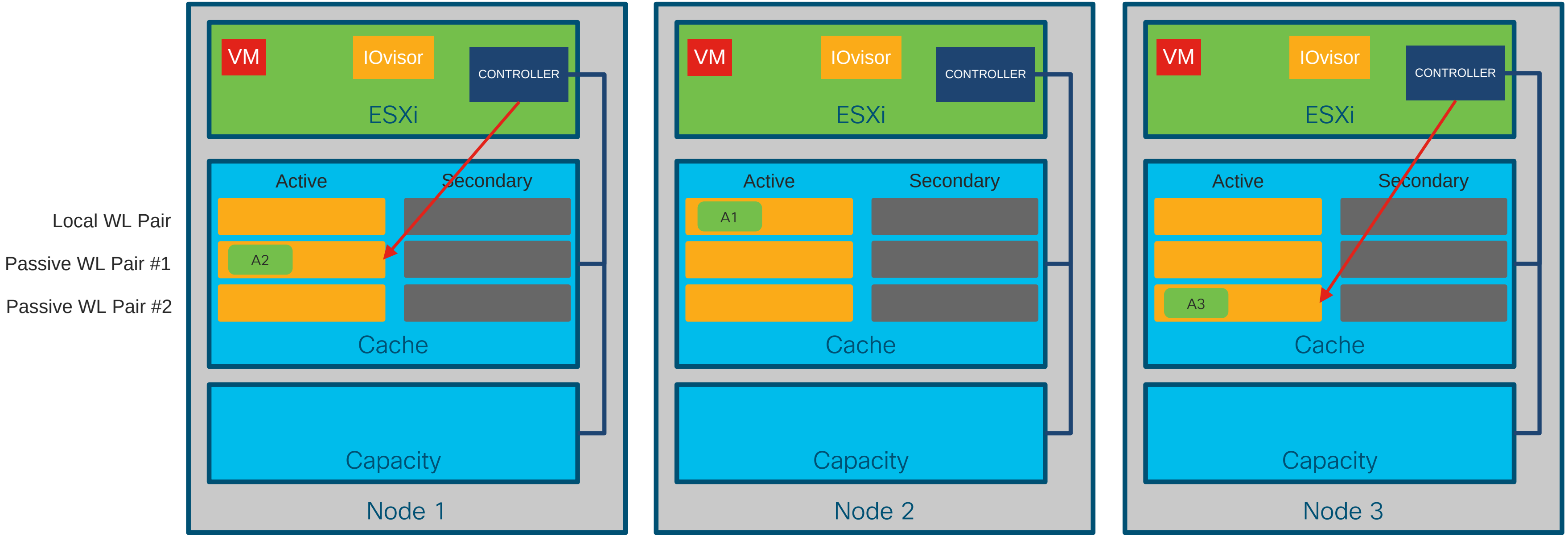
1. VM writes data to HX datastore, received by the IOvisor VIB, hashed to determine the primary node for that block, sent via network to primary node's controller VM.

Input/Output Lifecycles: Writes



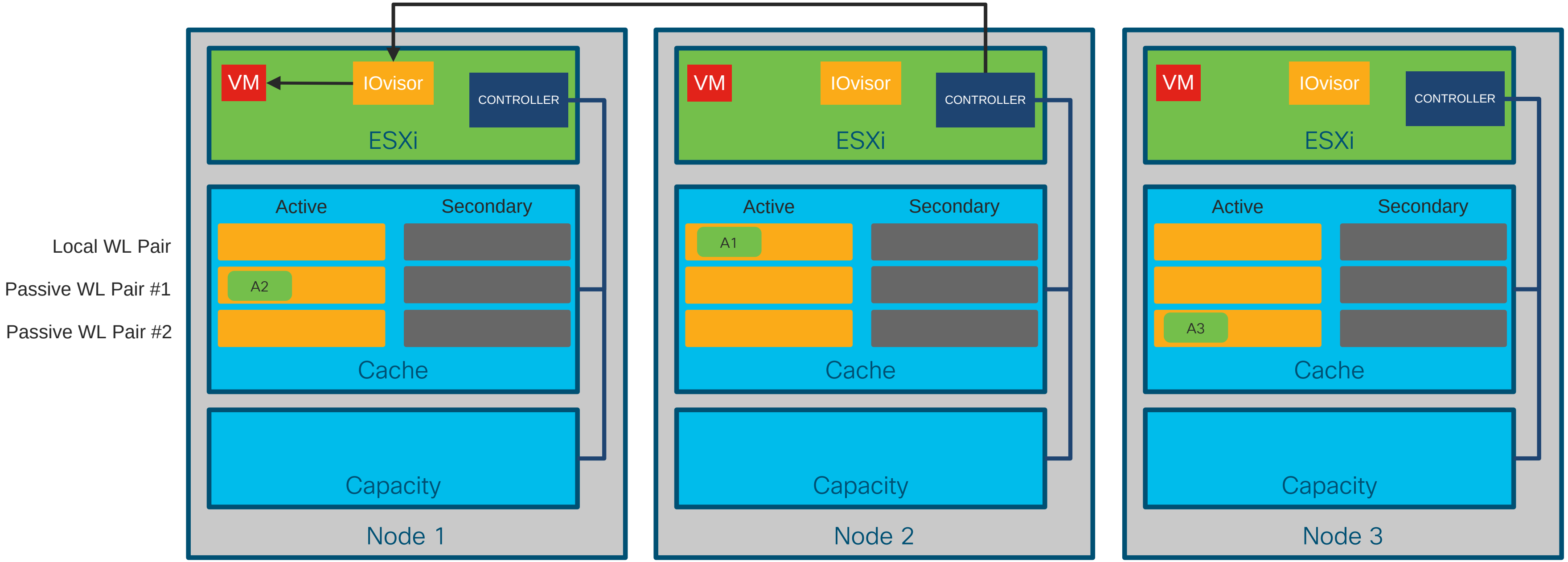
2. Controller VM compresses the data and commits the write to the active local write log on the caching SSD, and sends duplicate copies to the other nodes.

Input/Output Lifecycles: Writes



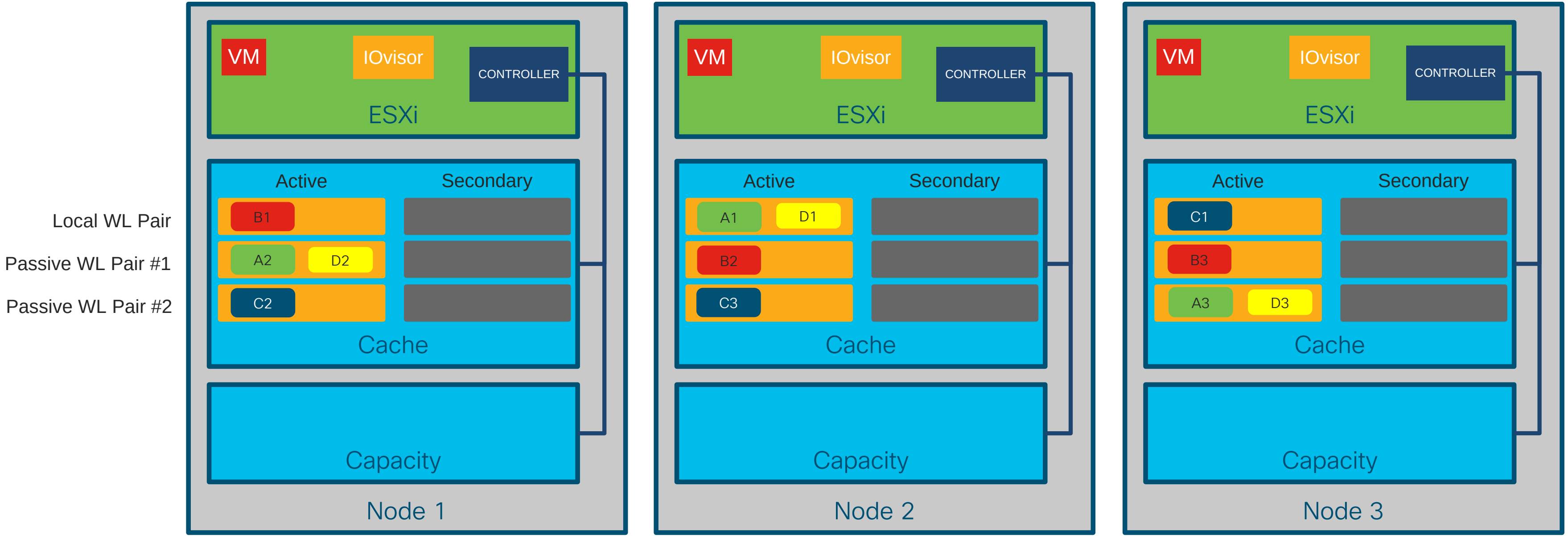
3. Additional nodes commit the write to their passive write logs on the caching SSDs.

Input/Output Lifecycles: Writes



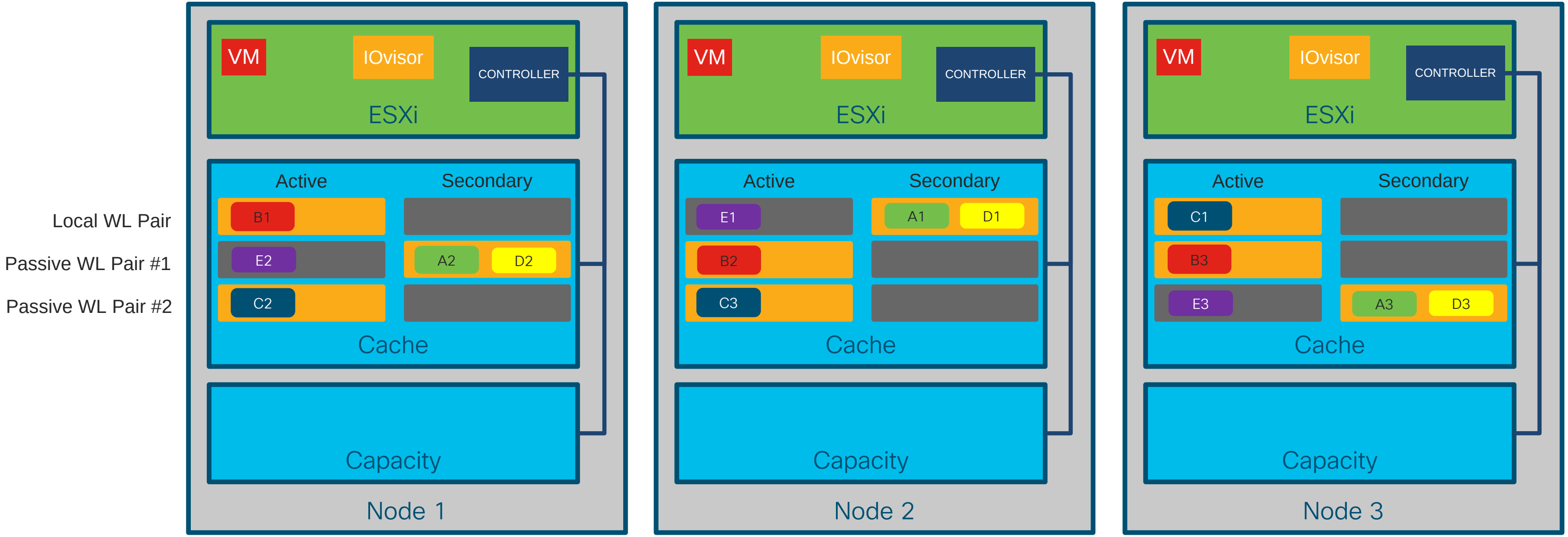
4. After all 3 copies are committed, the write is acknowledged back to the VM via the IOvisor and the datastore like a normal I/O.

Input/Output Lifecycles: Writes



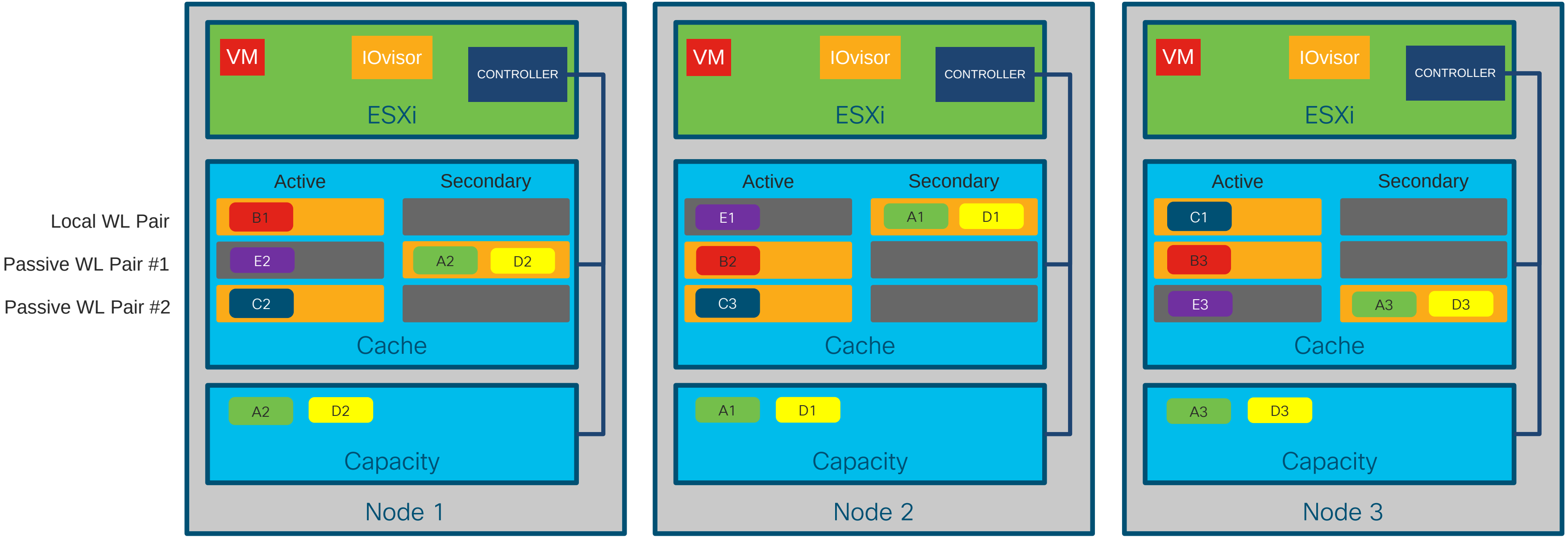
5. Additional write I/O is handled the same way, until the active local write log segment of a node becomes full, then a destage operation is started.

Input/Output Lifecycles: Destage



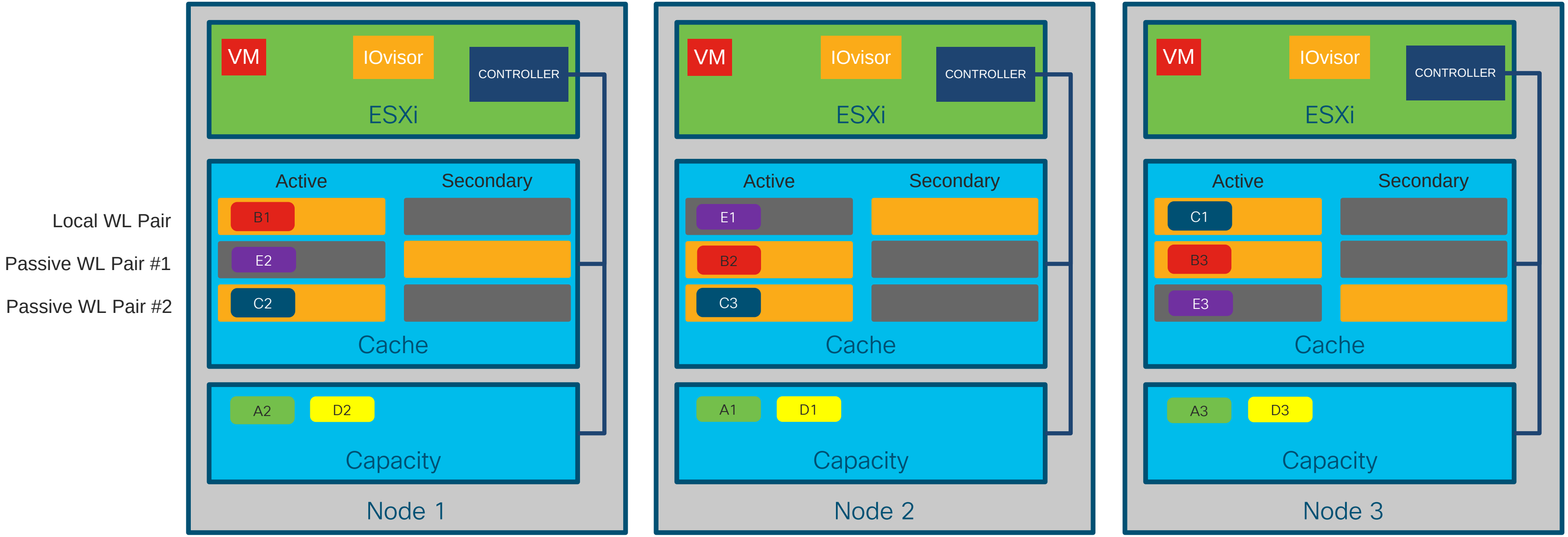
6. The active and secondary log segments are flipped, the full segments are ready to destage meanwhile the empty segments receive new incoming writes.

Input/Output Lifecycles: Destage



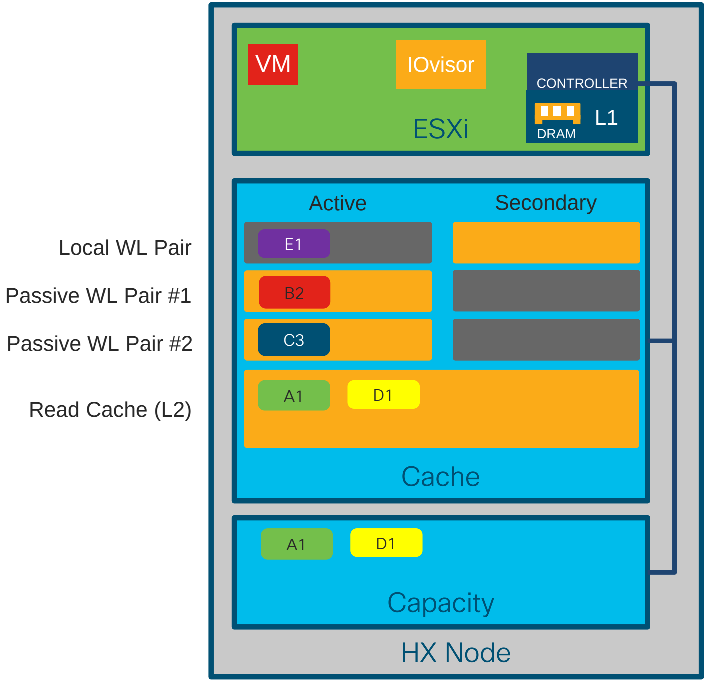
7. The data is deduplicated, and new blocks are committed to the capacity layer disks. Duplicates generate only a metadata update.

Input/Output Lifecycles: Destage



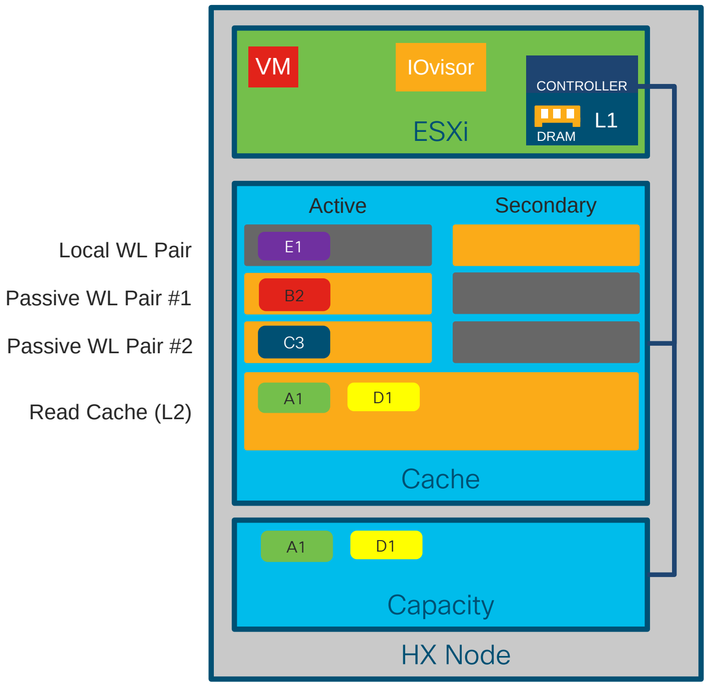
8. Once all data has been destaged, the secondary log segments are emptied, and ready to flip once the active segments become full again.

Input/Output Lifecycles: Read Caches



- All systems have a small amount of RAM in the controller VMs set aside for a read cache (L1).
- Hybrid converged nodes have a large dedicated read cache segment on their caching SSD (L2).
- During destage, a copy of the most recently written data is also copied to the L2 read cache.

Input/Output Lifecycles: Reads



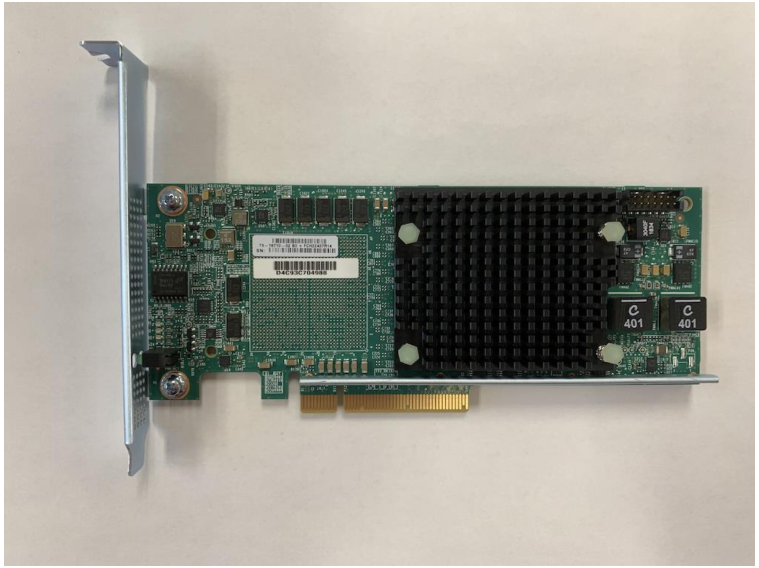
Requests to read data are also received by the IOvisor and serviced in the following order:

1. Active write log of primary node
2. Passive write logs
3. L1 (DRAM) cache
4. L2 read cache (hybrid only)
5. Capacity SSDs/HDDs

Reads are decompressed and cached in L1/L2 as appropriate, and returned to the requesting VM.

HyperFlex Acceleration Engine

HyperFlex Acceleration Engine

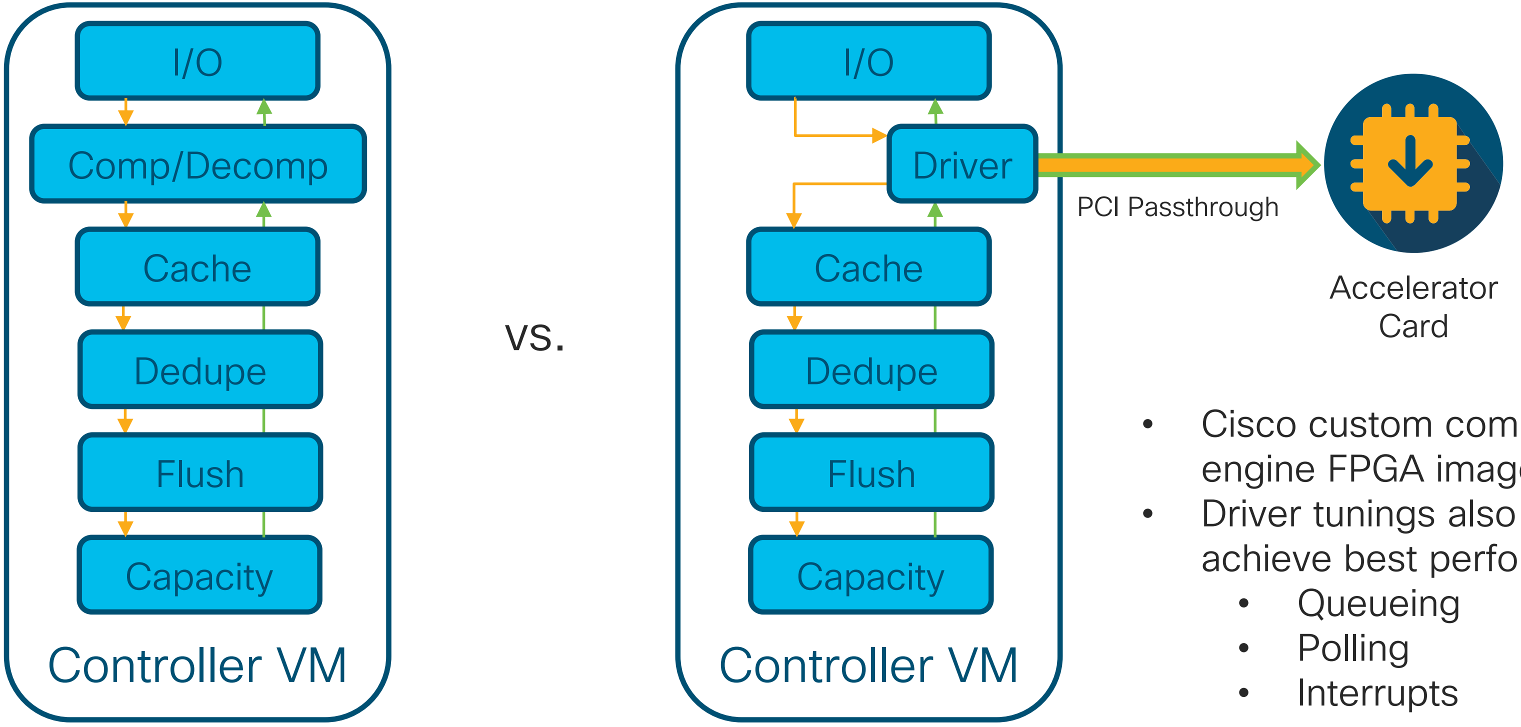


- PCIe FPGA offloading HX filesystem compression
- Higher Compression, Lower $/GB
- Better VM Density, Lower $/VM

- Performance Boost + lower read/write latency
- CPU Utilization & Working Capacity Improvements
- *Other service-offloads in-future: EC, Crypto, Hash etc.*

- Purpose Built & Co-Engineered Software & Hardware
- Ongoing improvements via new images and driver optimize
- Improved TCO

HyperFlex Data Path with Accelerator



- Cisco custom compression engine FPGA image
- Driver tunings also critical to achieve best performance
    - Queueing
    - Polling
    - Interrupts

HyperFlex Acceleration Engine

PID: HX-PCIE-OFFLOAD-1 HyperFlex Acceleration Engine

Requires:

- Enterprise License HXDP-P(1-5YR)
- Homogeneous cluster – all nodes must contain Hercules
- Greenfield – activation @ cluster install

Supported in specific models of nodes:

- HXAF240C-M5SX, All-Flash
- HX240C-M5SX, Hybrid SFF
- HX240C-M5L, Hybrid LFF
- HXAF220-M5N, All-NVMe

Not supported with HX 4.0(1a):

- Stretch clusters
- Edge clusters
- Hyper-V Hypervisor
- SED drives

HX Acceleration Engine - HW Compression

HXDP performs SW Compression using Snappy format – efficient, high speed and reasonable compression ratio.

Acceleration Engine utilizes GZIP in hardware which has a higher compression ratio, and is more computationally intensive than Snappy.

- Compression results will vary depending upon type, content and compressibility of the data.
  - Previously compressed file formats like JPEG video are not well suited for additional compression.
  - Internal tests show roughly 10% additional compression savings

	Original Total Size	Compressed Size Without Card	Savings Without Card	Compressed Size With Card	Savings With Card	Additional Compression Savings With Card
Calgary corpus	3141622	1814576	42.24%	1759282	44.00%	3.00%
Silesia corpus	211938580	114586735	45.93%	107982280	49.05%	5.80%
Canterbury corpus	2810784	1393861	50.41%	1182089	57.94%	15.20%

Benchmarking Tools

Vdbench

- Written by Henk Vandenbergh at Oracle
- Java application that can run on almost all platforms with a JRE
- Configurations to test are written in text based job files
- Highly customizable
- Easy server/client job execution via SSH or RSH
- Easy to read output in HTML files
- Has become the standard tool for the Cisco HX BU TME team
- Issues:
  - Thread and JVM count can make calculating the OIO per disk difficult to calculate

Vdbench Example Job File

```
dedupunit=8k
dedupratio=1.25
compratio=1.4
hd=default,user=root,shell=ssh,jvms=1
hd=vdbench1,system=192.168.100.11
hd=vdbench2,system=192.168.100.12
hd=vdbench3,system=192.168.100.13
hd=vdbench4,system=192.168.100.14

sd=default,openflags=o_direct,threads=16
sd=sd1,host=vdbench*,lun=/dev/sdb

wd=wd1,sd=sd*,xfersize=8k,rdpct=70,seekpct=100

rd=rd1,wd=wd1,iorate=max,elapsed=1h,interval=5,sleep=10m
rd=rd2,wd=wd1,iorate=max,elapsed=1h,interval=5,sleep=10m
rd=rd3,wd=wd1,iorate=max,elapsed=1h,interval=5
```

Vdbench Tips

Calculate the minimum number of total threads required:

- (# of load generators) X (# of disks per load generator) X (# of JVMs per load generator)
- Total # of threads defined depends on where in the job file it is specified:
  - If specified in the sd=default, and only one disk is defined, the value is the total # of threads for the entire test
  - If specified in the sd=default, and multiple disks are defined, it applies to all disks and the total is the sum of all values
  - If specified per disk, the total # of threads is the sum of all values

Calculate the OIO per disk:

- (total threads defined) / ((# of load generators) X (# of disks per load generator))
- Each JVM can service ~100K IOPs, so specifying jvms=1 makes thread calculation easier

Build a load generator template VM with the SSH keys of the master for no password logins

Multi-homed master systems must resolve the load generator facing IP to its hostname via /etc/hosts

Redirect output to a working web server folder, and use .tod for time of day stamps in the folder name

Vdbench Example Job Execution

Ensure each load generator VM has the vdbench java executable in the same location, and copy SSH keys so SSH logins happen without a password.

Since the vdbench job file specifies the clients as well, just run the test:

```
vdbench -f 7030_4k.vdb -o /var/www/html/7030_4k.tod
```

- Easy to loop multiple runs in sequence:

```
for i in {1..3} ; do vdbench -f 7030_4k.vdb -o /var/www/html/ && sleep 3600 ; done
```

- Simple scheduling of runs with Linux at:

```
# cd <location of vdbench job files>
# at 1am tomorrow
at> vdbench -f 7030_4k.vdb -o /var/www/html/output/7030_4k.tod
at> <EOT> (Press CTRL+d to end the job creation)
# atq (view at queued jobs)
# tail -f /var/spool/at/spool/<jobid> (monitor output of running at job)
```

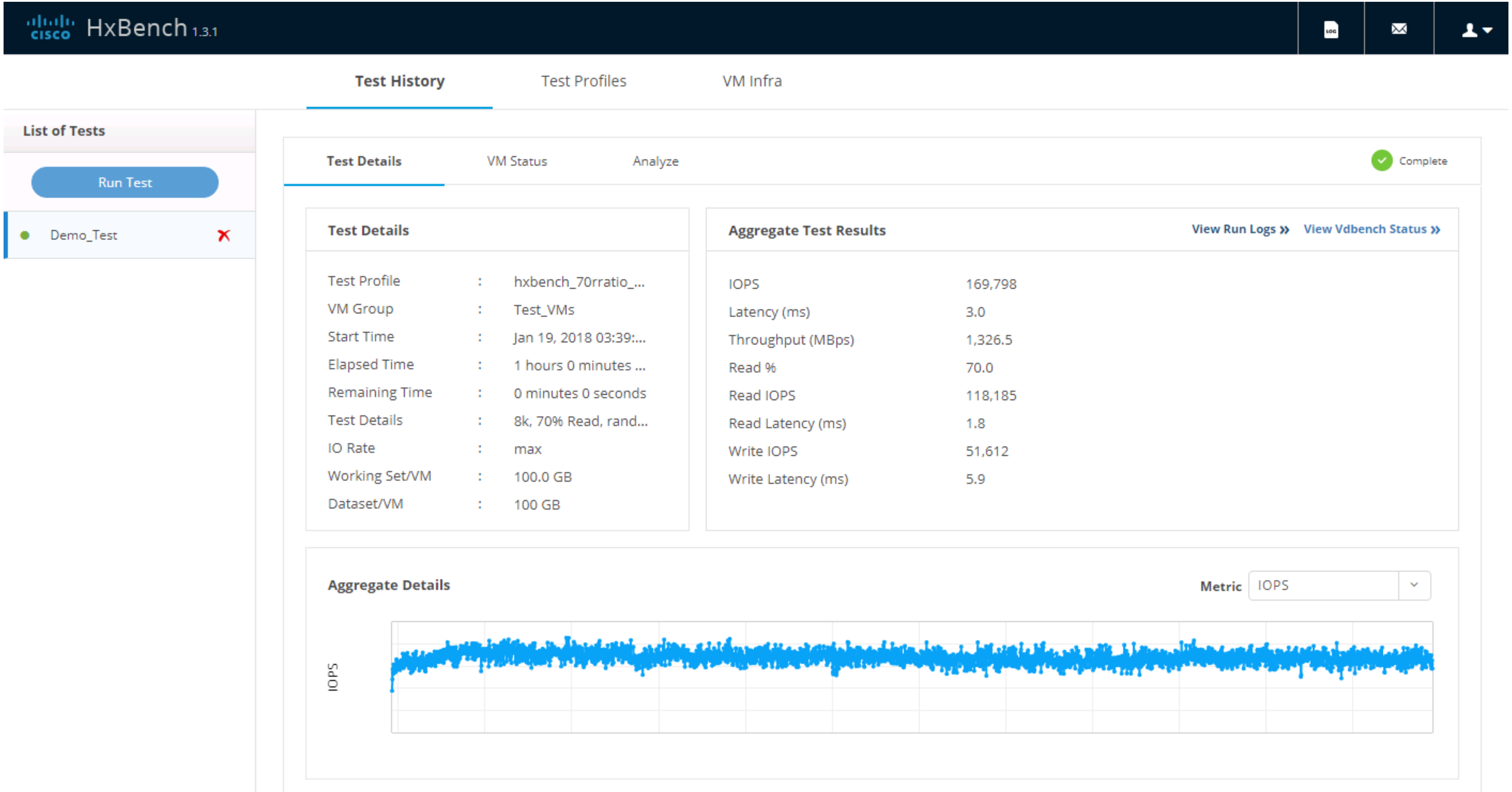
Vdbench Output

```
May 07, 2019  interval         i/o   MB/sec   bytes   read     resp     read    write     resp     resp queue  cpu%  cpu%
                              rate  1024**2     i/o    pct     time     resp     resp      max   stddev depth sys+u   sys
02:47:18.062       151   114282.55  1427.19   13094  70.04    2.607    1.857    4.361   86.760    4.160 298.0   1.0   0.6
02:47:38.051       152   113730.40  1420.30   13094  70.05    2.286    1.573    3.954   59.754    3.731 260.0   1.0   0.6
02:47:58.057       153   114261.05  1427.73   13102  70.02    2.216    1.514    3.857   89.243    3.757 253.3   1.0   0.6
02:48:18.051       154   113950.30  1421.69   13082  70.03    2.098    1.423    3.674   50.198    3.414 239.0   1.0   0.6
02:48:38.052       155   114154.35  1429.40   13129  70.04    2.220    1.479    3.955   54.554    3.577 253.4   1.0   0.6
02:48:58.052       156   114303.10  1426.62   13087  70.05    2.619    1.767    4.611   80.494    4.107 299.4   1.0   0.6
02:49:18.050       157   113833.25  1417.99   13061  70.03    2.992    2.136    4.992  172.096    4.821 340.7   1.0   0.6
02:49:38.056       158   114321.05  1424.61   13066  70.05    2.697    1.994    4.341   58.998    4.018 308.3   1.0   0.6
02:49:58.050       159   113850.65  1420.21   13080  69.99    2.488    1.727    4.262   83.548    3.958 283.3   1.0   0.6
02:50:18.055       160   114379.90  1430.40   13113  69.97    2.432    1.639    4.280   82.884    3.908 278.1   1.0   0.6
02:50:38.052       161   113691.00  1419.88   13095  69.95    2.862    1.982    4.911  125.986    4.287 325.4   0.9   0.6
02:50:58.054       162   114335.40  1430.06   13115  70.02    3.311    2.377    5.494   91.731    5.043 378.6   0.9   0.6
02:51:18.054       163   113751.65  1426.11   13146  70.05    3.271    2.338    5.453  122.152    4.699 372.1   0.9   0.6
02:51:38.051       164   114232.40  1432.91   13153  70.00    3.020    2.187    4.963   60.759    4.425 344.9   0.9   0.6
02:51:58.053       165   114410.00  1433.34   13136  69.96    2.834    2.071    4.612   74.866    4.378 324.3   1.0   0.6
02:52:18.051       166   113927.40  1427.97   13142  69.97    2.742    1.990    4.496   89.488    4.261 312.5   1.0   0.6
02:52:38.056       167   114312.85  1428.52   13103  69.94    3.216    2.412    5.087  376.409    8.220 367.6   0.9   0.6
02:52:58.054       168   113859.30  1422.18   13097  69.98    3.414    2.482    5.587  113.014    5.305 388.8   0.9   0.6
02:53:18.054       169   114330.55  1429.20   13107  69.97    3.084    2.293    4.926   90.576    4.731 352.6   1.0   0.6
02:53:38.051       170   113938.45  1423.24   13098  69.94    3.268    2.432    5.213  154.814    4.903 372.2   0.9   0.6
02:53:58.053       171   114175.30  1428.62   13120  69.98    2.951    2.069    5.008   71.467    4.506 337.0   1.0   0.6
02:54:18.051       172   113697.10  1419.52   13091  69.96    2.584    1.775    4.468  126.972    4.149 293.9   1.0   0.6
02:54:38.051       173   114215.55  1427.92   13109  70.08    2.596    1.775    4.518   64.882    4.005 296.4   1.0   0.6
02:54:58.054       174   114309.35  1427.15   13091  70.05    2.479    1.686    4.333   63.748    3.957 283.4   1.0   0.6
02:55:18.057       175   113745.40  1424.90   13135  70.13    2.860    2.121    4.592  248.715    5.904 325.3   0.9   0.6
02:55:38.053       176   114276.75  1428.09   13103  70.07    2.693    1.903    4.543   70.478    4.179 307.9   1.0   0.6
02:55:58.050       177   113898.90  1421.68   13088  70.10    3.332    2.470    5.351  179.942    5.721 379.3   0.9   0.6
02:56:18.054       178   114223.90  1425.38   13084  70.10    3.505    2.671    5.457  206.941    5.661 400.4   0.9   0.6
02:56:38.051       179   113895.70  1420.08   13073  70.12    3.427    2.595    5.378  224.246    5.812 390.2   0.9   0.6
02:56:58.057       180   114381.55  1425.01   13063  70.09    3.197    2.401    5.061  311.615    5.921 365.6   1.0   0.6
02:56:58.121 avg_2-180   114087.51  1425.81   13104  70.01    2.196    1.517    3.783  376.409    4.051 250.6   0.8   0.5
02:56:59.100 Vdbench execution completed successfully. Output directory: /var/www/html/output/CLUS2019/iops_curve_blender_AF_8node.190506.093637
```


Cisco HX Bench

- Cisco developed tool to automate many facets of benchmark testing
- Uses Vdbench for the testing
- Download at https://hyperflexsizer.cloudapps.cisco.com (CCO login required)
- Deploys a master test controller VM from an OVA file on your HX cluster
- HTML GUI makes test configuration easy
- Automatically deploys lightweight Ubuntu VMs as the test load generators
- Use built in test scenarios or build your own, queue multiple runs and compare the results.
- Test results and charts are customizable and even easier to view than the default HTML output from vdbench

Cisco HX Bench User Interface



Other Benchmark Tools

- Iometer
  - Defunct open source tool, no active development in 4 years
  - Data is highly dedupable, results cannot be trusted

- FIO
  - Similar to Vdbench, highly customizable, output harder to consume

- HCIBench
  - VMware written wrapper for Vdbench

Performance Test Results

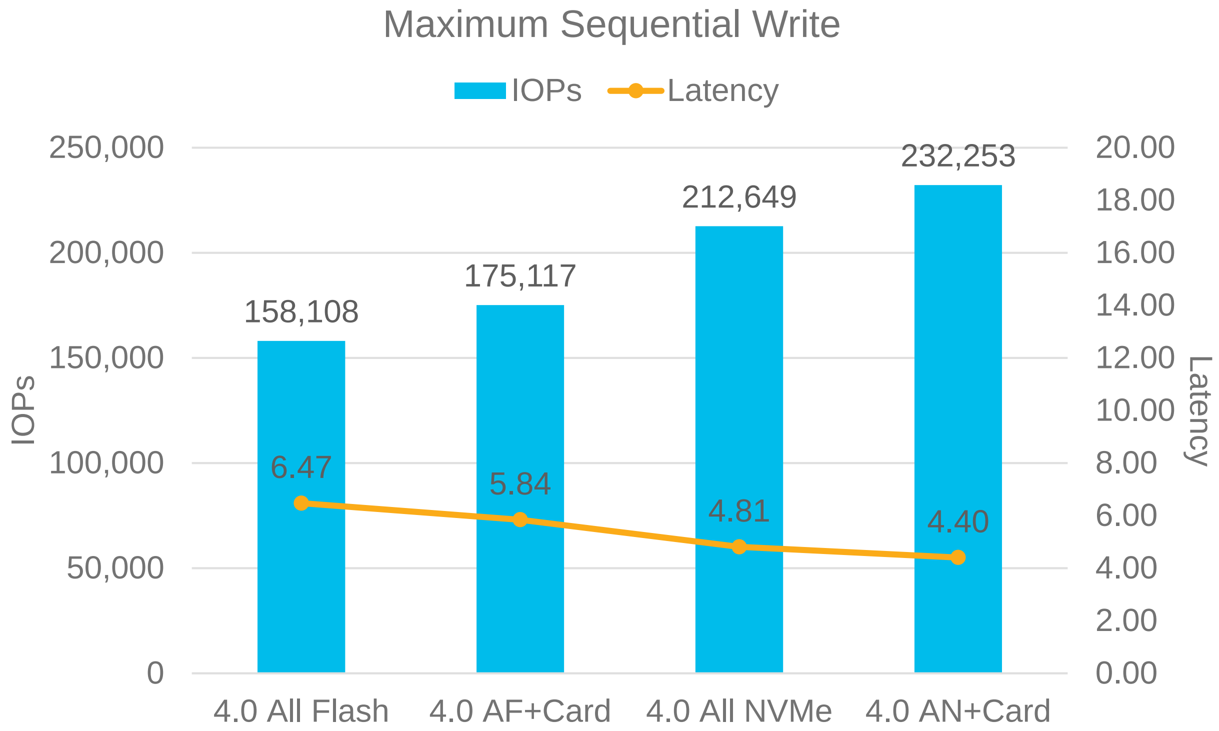
Base Configurations for Vdbench Tests

	HX All-Flash	HX All-Flash + Accelerator	HX All-NVMe	HX All-NVMe + Accelerator
Version	HX 4.0.1b	HX 4.0.1b	HX 4.0.1b	HX 4.0.1b
Hardware Config	8 x HXAF240 M5S, each with: • WL: 375GB Optane • Data: 10 x 960G • VIC 1387 • Xeon 6248 • 384 GB RAM	8 x HXAF240 M5S, each with: • WL: 375GB Optane • Data: 10 x 960G • VIC 1387 • Xeon 6248 • 384 GB RAM	8 x HXAF220 M5N, each with: • WL: 375GB Optane • Data: 6 x 1TB • VIC 1387 • Xeon 6254 • 384 GB RAM	8 x HXAF220 M5N, each with: • WL: 375GB Optane • Data: 6 x 1TB • VIC 1387 • Xeon 6254 • 384 GB RAM
Benchmark Setup	• Cisco UCS Firmware 4.0(4f) • Vdbench 5.04.06 • 8 Linux VMs per host (64 total) • Each VM has 1 x 200GB virtual disk on HXAF240, 125GB on HXAF220, HX filesystem ~50% full before dedupe and compression • Priming: 64K writes 100% sequential until load generator disks are full • Tests run 3 times minimum for 1 hour each, and then averaged			

Benchmark 1

Performance Results Using Vdbench Tests

8K 100% Sequential Write Workload

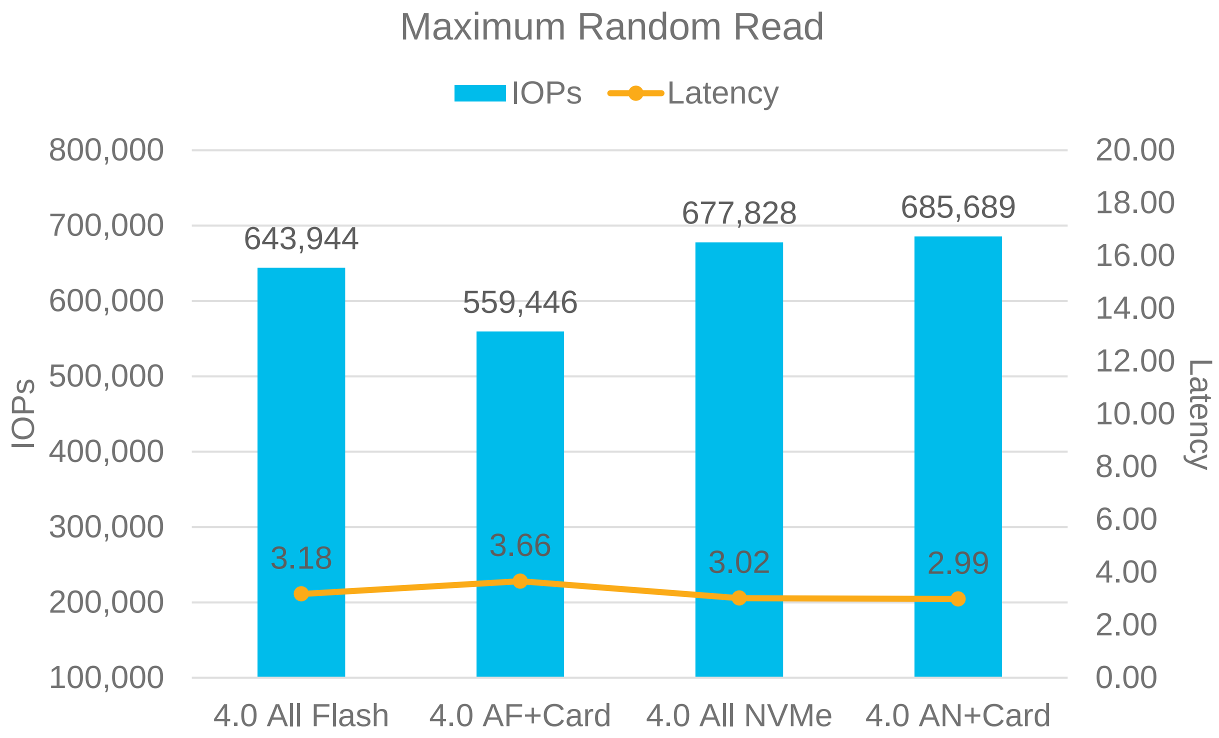


- 16 OIO Per Disk
- ~34% IOPs gain and ~27% latency reduction for All-NVMe compared to All-Flash
- ~9-10% IOPs gain and ~9-10% latency reduction when adding the Accelerator Card

Benchmark 2

Performance Results Using Vdbench Tests

8K 100% Random Read Workload



- 32 OIO Per Disk
- ~13% IOPs loss for All-Flash with Accelerator Card
- Loss only seen in 100% read test, not seen in mixed R/W
- Differences in queues and block interaction being worked to improve result
- No IOPs loss for All-NVMe with Accelerator Card
- All-NVMe results could improve with all 8 capacity disks versus only 6

Benchmark 3

Performance Results Using Vdbench Tests

100% Random 70/30% R/W with Varying Block Sizes

	4K			8K			16K			64K			256K		
	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency
All-Flash	147,872	2.68	5.39	172,536	1.85	5.57	133,665	2.63	6.71	56,199	8.53	10.58	17,821	31.75	21.58
All-Flash + Accelerator	152,912	2.69	4.96	189,514	1.82	4.76	149,745	2.43	5.74	57,414	9.31	8.15	18,947	31.67	16.12
All-NVMe	226,507	1.81	3.35	261,906	1.42	3.20	212,935	1.62	4.22	96,642	3.59	9.26	33,865	10.34	26.18
All-NVMe + Accelerator	243,190	1.73	2.99	277,477	1.41	2.86	225,668	1.63	3.74	101,039	3.75	8.11	35,123	10.93	23.01

- 8 OIO Per Disk
- 52-90% IOPs Gains with HyperFlex All-NVMe compared to All-Flash
- All-NVMe cluster without HyperFlex Accelerator Cards already significantly outperforms All-Flash with the cards
- 4-7% additional IOPs Gains with HyperFlex Accelerator Cards
- 10-12% Write Latency reduction with HyperFlex Accelerator Cards

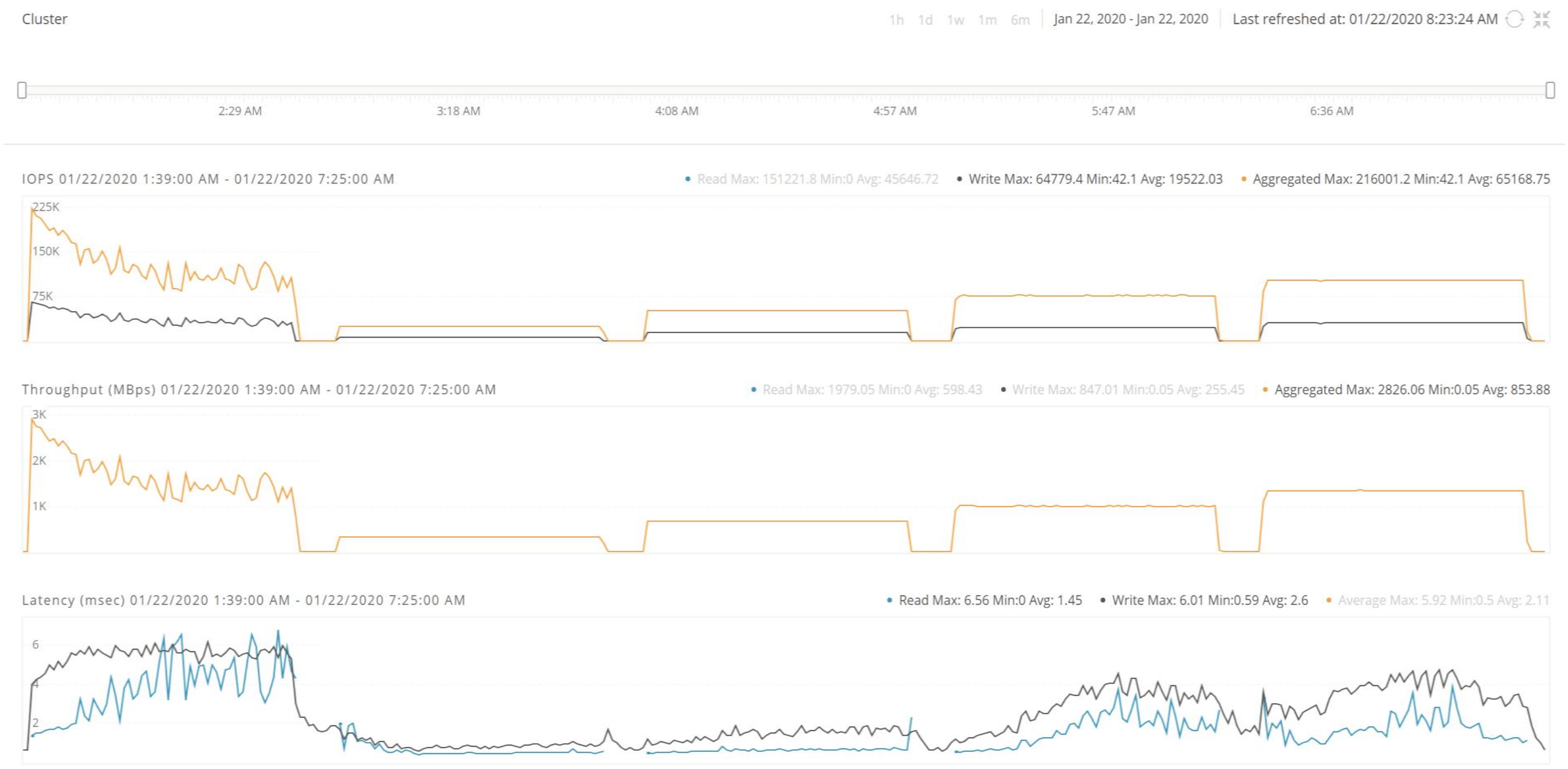
Benchmark 4

Performance Results Using Vdbench Tests

8K 100% Random 70/30% R/W IOPs Curve

	20%			40%			60%			80%			100%		
	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency
All-Flash	34,064	0.3%	0.70	68,091	0.3%	1.07	102,123	0.2%	1.68	135,230	2.1%	2.44	170,103	12.3%	3.02
All-Flash + Accelerator	36,233	0.3%	0.61	72,429	0.2%	1.10	108,653	0.2%	1.68	144,796	0.6%	2.00	181,031	12.5%	2.83
All-NVMe	52,796	0.3%	0.71	105,485	0.3%	1.02	158,209	0.3%	1.27	210,889	0.8%	1.58	263,634	11.1%	1.94
All-NVMe + Accelerator	61,663	0.3%	0.82	123,288	0.3%	0.93	184,880	0.4%	1.09	246,490	1.2%	1.51	308,109	8.4%	1.66

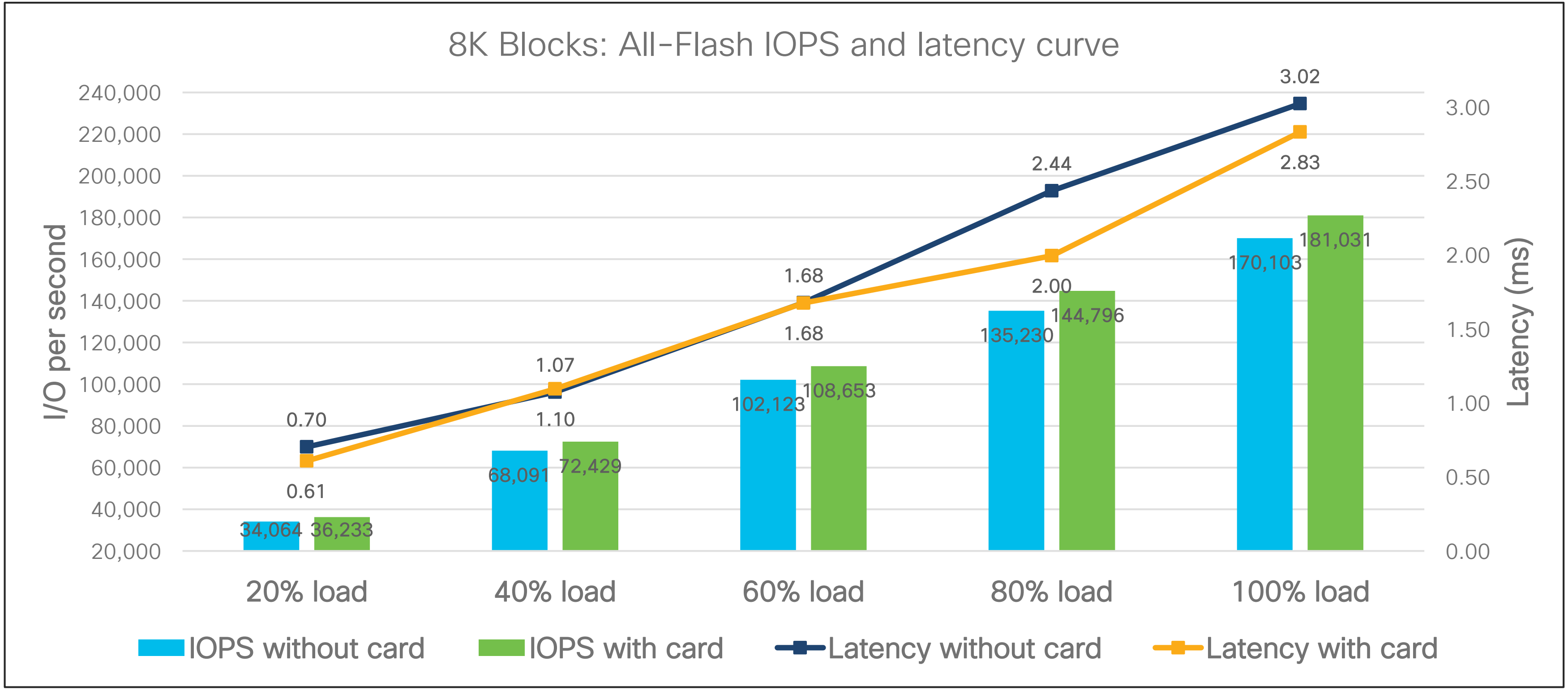
- Curve test measures performance at various performance levels, starting with a 100% unthrottled test, then running the workload again at a different percentages of that maximum result.
- We calculate stability by measuring the standard deviation of IOPs as a percentage of the final average result
- Maximum speed, unthrottled tests will always have more variability due to more aggressive background tasks
- 20% through 80% values are typically very low because the system is not being pushed to the limit


Cluster
1h 1d 1w 1m 6m
Jan 22, 2020 - Jan 22, 2020
Last refreshed at: 01/22/2020 8:23:24 AM
2:29 AM
3:18 AM
4:08 AM
4:57 AM
5:47 AM
6:36 AM
IOPS 01/22/2020 1:39:00 AM - 01/22/2020 7:25:00 AM
Read Max: 151221.8 Min:0 Avg: 45646.72
Write Max: 64779.4 Min:42.1 Avg: 19522.03
Aggregated Max: 216001.2 Min:42.1 Avg: 65168.75
225K
150K
75K
Throughput (MBps) 01/22/2020 1:39:00 AM - 01/22/2020 7:25:00 AM
Read Max: 1979.05 Min:0 Avg: 598.43
Write Max: 847.01 Min:0.05 Avg: 255.45
Aggregated Max: 2826.06 Min:0.05 Avg: 853.88
3K
2K
1K
Latency (msec) 01/22/2020 1:39:00 AM - 01/22/2020 7:25:00 AM
Read Max: 6.56 Min:0 Avg: 1.45
Write Max: 6.01 Min:0.59 Avg: 2.6
Average Max: 5.92 Min:0.5 Avg: 2.11
6
4
2

Benchmark 4

Performance Results Using Vdbench Tests

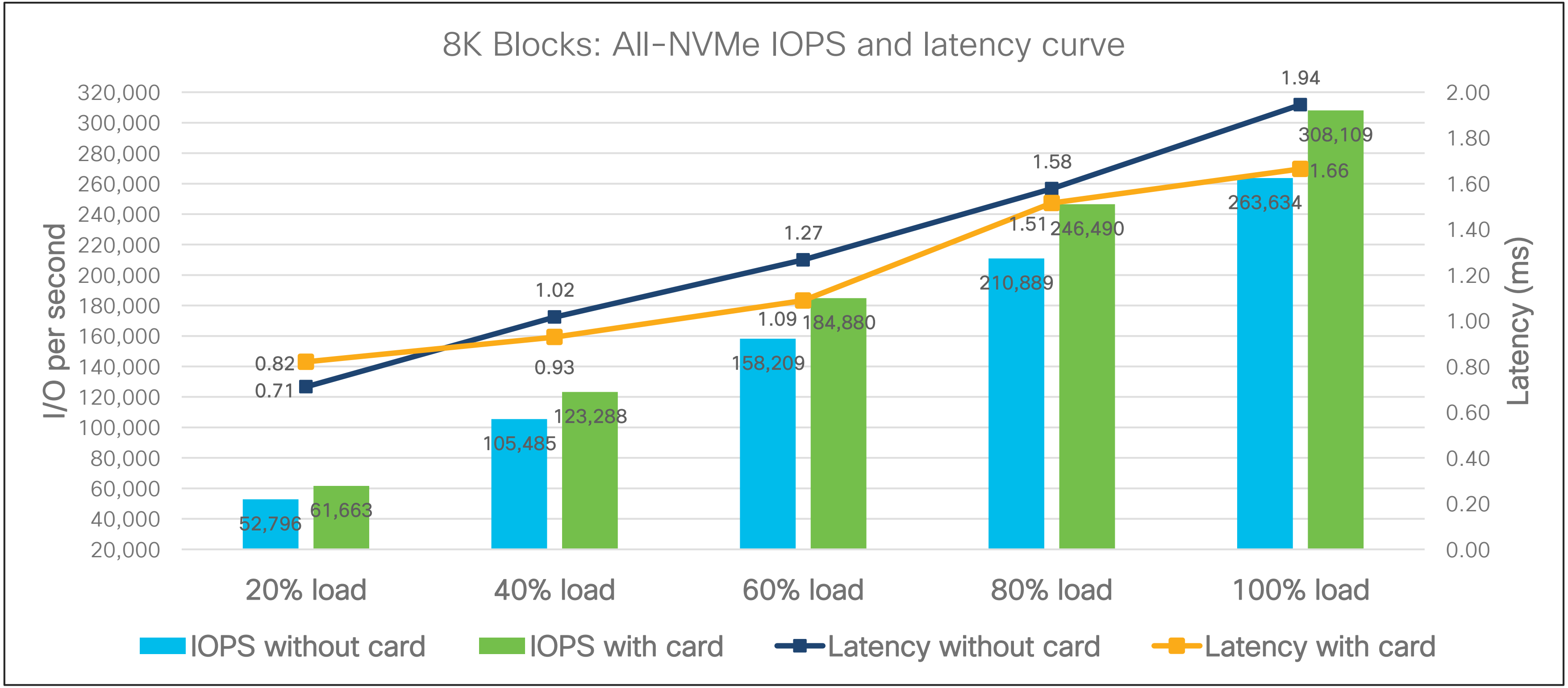
8K 100% Random 70/30% R/W IOPs Curve



Benchmark 4

Performance Results Using Vdbench Tests

8K 100% Random 70/30% R/W IOPs Curve



Benchmark 5

Performance Results Using Vdbench Tests

I/O Blender 100% Random 70/30% R/W IOPs Curve

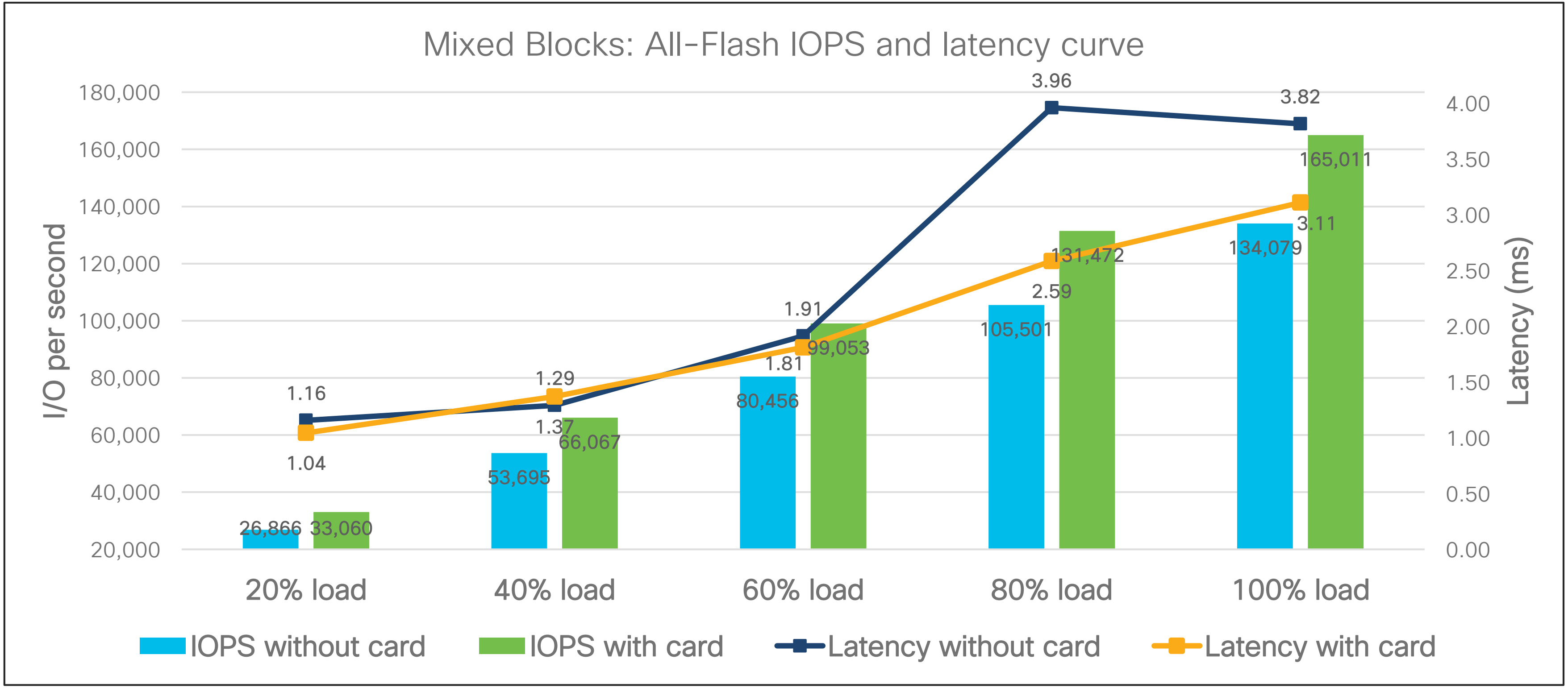
	20%			40%			60%			80%			100%		
	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency	IOPs	IOPs STDV	Avg Latency
All-Flash	26,866	0.3%	1.16	53,695	0.2%	1.29	80,456	0.2%	1.91	105,501	5.2%	3.96	134,079	14.7%	3.82
All-Flash + Accelerator	33,060	0.3%	1.04	66,067	0.2%	1.37	99,053	0.2%	1.81	131,472	2.6%	2.59	165,011	11.7%	3.11
All-NVMe	39,391	0.3%	0.68	78,764	0.3%	1.07	118,088	0.4%	1.45	157,445	0.7%	2.02	196,772	15.2%	2.60
All-NVMe + Accelerator	42,829	0.3%	0.68	85,590	0.2%	0.95	128,380	0.3%	1.42	171,104	0.8%	1.95	213,885	14.6%	2.39

- Curve test measures performance at various performance levels, starting with a 100% unthrottled test, then running the workload again at a different percentages of that maximum result.
- We calculate stability by measuring the standard deviation of IOPs as a percentage of the final average result
- Block Size Mix: 4K 40%, 8K 40%,16K 10%, 64K 10%
- 20% through 80% values are typically very low because the system is not being pushed to the limit
- Every configuration shows higher IOPs at lower latency than the previous result, up to 60% improvement

Benchmark 5

Performance Results Using Vdbench Tests

I/O Blender 100% Random 70/30% R/W IOPs Curve

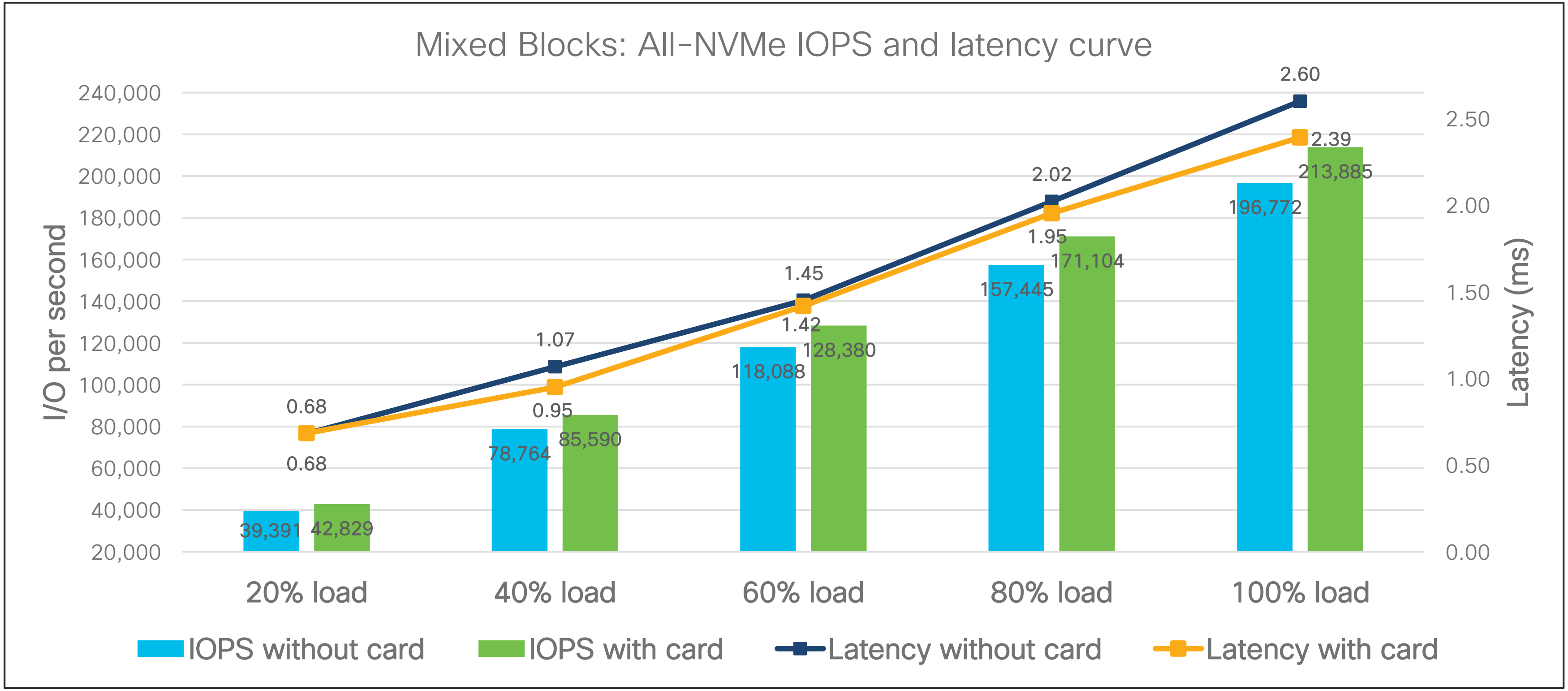


cisco Live!

Benchmark 5

Performance Results Using Vdbench Tests

I/O Blender 100% Random 70/30% R/W IOPs Curve



cisco Live!

Benchmark 6

Performance Results Using Vdbench Tests

8K 100% Random 70/30% R/W IOPs Curve Match

	20%			40%			60%			80%			100%		
	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency
All-Flash	34,064	0.58	0.99	68,091	0.78	1.76	102,123	1.14	2.94	135,230	1.63	4.31	170,103	1.98	5.45
All-NVMe + Accelerator	33,995	0.53	0.64	67,996	0.94	0.86	101,993	0.95	1.22	135,981	0.88	1.34	169,968	0.83	1.56
Improvement		8.7%	35.5%		-21.1%	58.8%		16.9%	58.6%		46.1%	69%		58%	71.4%

- Curve test run as before, but All-NVMe plus Accelerator Card runs throttled to match the results from a standard All-Flash HX 4.0 cluster
- Standardize on IOPs to see the relative difference in latencies when workloads don't demand higher IOPs
- Write latencies improve in all tests, by as much as 71%
- Read latencies also improve by up to 58%
- 40% test read latencies affected by cleaner process getting more aggressive on the smaller All-NVMe cluster earlier

Benchmark 7

Performance Results Using Vdbench Tests

I/O Blender 100% Random 70/30% R/W IOPs Curve Match

	20%			40%			60%			80%			100%		
	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency
All-Flash	26,866	1.03	1.45	53,695	0.93	2.14	80,456	1.25	3.47	105,501	3.11	5.94	134,079	2.57	6.74
All-NVMe + Accelerator	26,999	0.53	0.76	53,995	0.66	1.17	79,991	0.75	1.57	104,986	0.93	2.15	133,985	1.04	2.51
Improvement		49.1%	47.2%		29.3%	45.1%		39.8%	54.8%		70%	63.8%		59.5%	62.8%

- Curve test run as before, but All-NVMe plus Accelerator Card runs throttled to match the results from a standard All-Flash HX 4.0 cluster
- Standardize on IOPs to see the relative difference in latencies when workloads don't demand higher IOPs
- Block size mix: 4K 40%, 8K 40%,16K 10%, 64K 10%
- Write latencies improve in all tests, by as much as 64%
- Read latencies also improve by up to 70%
- 40% test read latencies issue not seen when testing with mixed blocks

HyperFlex Boost Mode

- Increases the number of vCPUs for the HX Storage Controller VMs
  - Only available with HX 4.0(2a) and onward with ESXi
  - All-flash and All-NVMe only, except for HX Edge
  - Increases all-flash from 8 vCPUs to 12
  - Increases all-NVMe from 12 vCPUs to 16
  - The server physical CPUs must have at least the requisite number of physical cores per socket

- Upgrade
  - Change can be made in a rolling fashion via a public process to be made available shortly after release
  - No changes to the CPU reservation settings, i.e. it is still a “soft” reservation leaving CPU cores available to guest VMs when load is low

Benchmark 8

Performance Results Using Vdbench Tests

Boost Mode 100% Read and 100% Write for All-Flash and All-NVMe

	All-Flash 100% Write			All-Flash 100% Read			All-NVMe 100% Write			All-NVMe 100% Read		
	IOPs	W Latency	CPU %	IOPs	R Latency	CPU %	IOPs	W Latency	CPU %	IOPs	R Latency	CPU %
Normal	172,989	5.91	16.8	559,262	3.66	38.2	235,451	4.34	24.3	715,054	2.85	45.6
Boost Mode	218,042	4.69	21.3	668,236	3.06	46.0	266,168	3.84	27.8	711,324	2.88	49.5
Change	+26.0%	-20.7%	+26.5%	+19.5%	-16.3%	+20.5%	+13.0%	-11.5%	+14.6%	-0.5%	+0.8%	+8.4%

- Up to a 26% gain in performance
- Latencies decrease at a similar rate compared to the performance improvement
- ESXi host CPU utilization increases at a rate similar to the overall performance gain
- All-NVMe doesn't gain performance in the 100% read test, suggesting it is constrained elsewhere
- Standard deviation of latency during the 100% read tests fell by 88% and 86% respectively using Boost Mode

Benchmark 9

Performance Results Using Vdbench Tests

Boost Mode All-Flash Block Sweep 100% Random 70/30% R/W

	4K				8K				16K				64K				256K			
	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %
All-Flash + Card	153,197	2.68	4.98	20.3	191,448	1.79	4.73	21.9	148,739	2.48	5.75	22.9	58,571	8.99	8.28	20.3	18,880	31.78	16.16	20.9
All-Flash + Card + Boost	196,274	2.57	2.95	27.6	250,445	1.68	2.91	30.0	188,196	2.42	3.44	28.6	60,979	10.21	4.19	20.2	18,954	35.21	7.86	21.7
Change	+28.1%	-4.0%	-40.8%	+35.4%	+30.8%	-6.2%	-38.5%	+37.1%	+26.5%	-2.3%	-40.1%	+24.5%	+4.1%	+13.6%	-49.3%	-0.1%	+0.4%	+10.8%	-51.4%	+3.8%

- Block sweep runs comparing an All-Flash cluster with Acceleration cards to the same with Boost Mode enabled
- Up to ~31% gain in IOPs by increasing the number of vCPUs of the HX Storage Controller VMs
- Write latencies improve in all tests, by as much as ~51%
- Read latencies also improve by up to ~6%
- CPU utilization increase is nearly linear with the improvement in IOPs, except for at larger block sizes
- Large block sizes see significant decreases in write latency without improvements in IOPs, suggesting a benefit from more parallelism, but bottlenecks exist elsewhere and may be interfering with read latencies

Benchmark 9

Performance Results Using Vdbench Tests

Boost Mode All-NVMe Block Sweep 100% Random 70/30% R/W

	4K				8K				16K				64K				256K			
	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %	IOPs	R Lat	W Lat	CPU %
All-NVMe + Card	236,613	1.79	3.06	31.1	278,087	1.40	2.87	33.8	224,001	1.64	3.77	33.6	100,071	3.78	8.20	32.9	35,319	10.71	23.25	36.9
All-NVMe + Card + Boost	196,274	2.57	2.95	37.8	250,445	1.68	2.91	42.0	270,936	1.50	2.80	40.9	118,025	3.64	5.95	38.1	39,351	11.46	16.59	40.5
Change	+20.6%	-11.8%	-23.8%	+21.6%	+26.8%	-16.7%	-26.4%	+24.3%	+21.0%	-8.9%	-25.9%	+21.7%	+17.9%	-3.9%	-27.4%	+15.7%	+11.4%	+7.0%	-28.6%	+9.8%

- Block sweep runs comparing an All-NVMe cluster with Acceleration cards to the same with Boost Mode enabled
- Up to ~27% gain in IOPs by increasing the number of vCPUs of the HX Storage Controller VMs
- Write latencies improve in all tests, by as much as ~29%
- Read latencies also improve by up to ~17%
- CPU utilization increase is nearly linear with the improvement in IOPs for all tests
- All-NVMe system sees increased performance for all block sizes, showing the benefit of more processing parallelism, alongside the greater parallelism of NVMe at the storage layer

How Far We Have Come

Compare HX 3.0 All-Flash to HX 4.0 All-NVMe

	64K 100% W		8K 100% R		8K 70/30 R/W			64K 70/30 R/W			256K 70/30 R/W		
	IOPs	W Latency	IOPs	R Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency	IOPs	R Latency	W Latency
3.0 All-Flash	34,493	29.42	571,145	3.58	168,258	2.11	5.20	67,484	5.76	11.98	21,374	12.06	51.61
4.0 All-NVMe + Accelerator	70,469	14.11	685,689	2.99	277,477	1.41	2.86	101,039	3.75	8.11	35,123	10.93	23.01
Improvement	104.3%	-52.0%	20.1%	-16.7%	64.9%	-33.1%	-45.1%	49.7%	-34.8%	-32.3%	64.3%	-9.4%	-55.4%

- Compare HX 3.0 All-Flash cluster results with Intel Optane caching drive to HX 4.0 All-NVMe cluster with HyperFlex Accelerator Cards
- Results shown between two HX systems that were released only 1 year apart from each other
- More than doubled throughput and IOPs during 64K block priming tests with latency cut in half
- Mixed R/W testing IOPs improves by 50-65%, nearly 9GB/sec for 256K blocks
- Mixed R/W write latency improves by up to 55%
- Mixed R/W read latencies improve in all tests by up to 35%

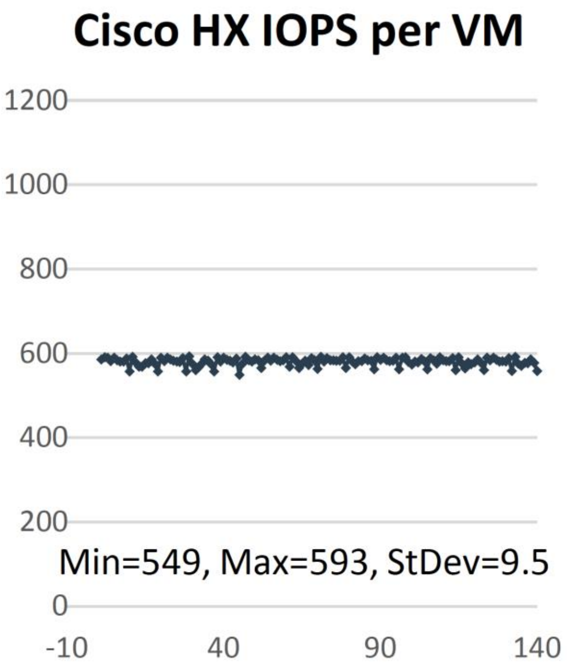
Benchmarking Best Practices

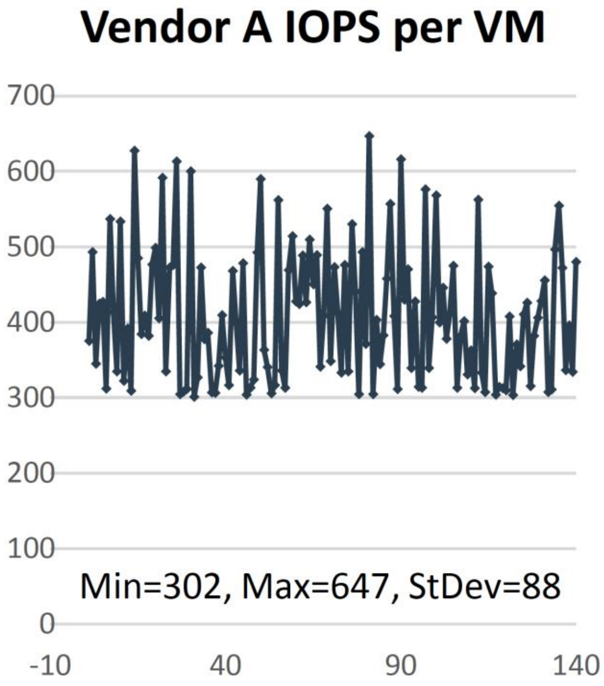
1. Initialize the Disks
   - More realistic: applications don't read data that hasn't been already written
2. Longer Test Duration (1 hour+)
   - Allows background processes to run, performance will stabilize over time
3. Test Appropriate Working Set Sizes
   - Larger working set size is a key benefit for all flash, don't exceed read cache size on hybrid nodes
4. Enable Dedupe and Compression on Competition (On by Default for HX)
   - Most enterprise workloads see significant storage efficiency savings
5. Workload Read / Write Mix
   - Reasonable write percentage (20-35% at least) in the workload

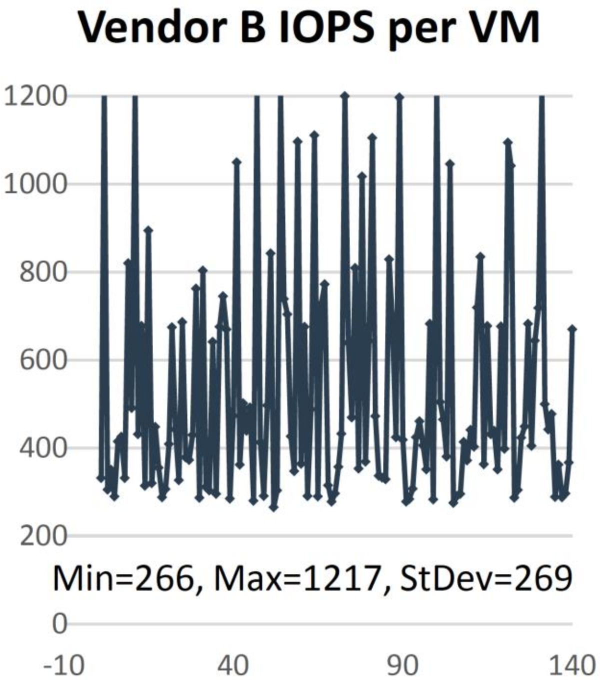
Benchmarking Best Practices

6. Similar hardware configuration
   - Pay careful attention to the BOM – CPU, Memory and SSD (type and count for cache/WL and persistent data SSDs)
7. Similar software configuration
   - Pay careful attention to the resiliency setting – that needs to be identical
   - Comparing Replication Factor = 2 performance with Replication Factor = 3 performance is not comparing apples to apples
8. What to look for in performance results:
   - IOPS / Throughput
   - Latency – not all workloads show sub millisecond latencies (even on all flash!)
   - Consistent Latency / standard deviation of latency
   - Variances across the VMs

Per VM Test Results Variation







Source: https://research.esg-global.com/reportaction/ciscohyperflexcomplexworkloads/Toc

HyperFlex Design and Performance Summary

- Cisco HyperFlex All-NVMe cluster combine best-of-breed technology and partnerships
- HyperFlex All-NVMe clusters take HCI performance to new heights
- HyperFlex Accelerator Cards improve performance, reduce latency and increase compression savings
- Year-over-year performance gains are transformational and enable HX use for all workloads

Complete your online session survey

- Please complete your session survey after each session. Your feedback is very important.
- Complete a minimum of 4 session surveys and the Overall Conference survey (starting on Thursday) to receive your Cisco Live t-shirt.
- All surveys can be taken in the Cisco Events Mobile App or by logging in to the Content Catalog on ciscolive.com/emea.

Cisco Live sessions will be available for viewing on demand after the event at ciscolive.com.

Continue your education

Demos in the
Cisco campus

Walk-in
self-paced labs

Meet the engineer
1:1 meetings

Related sessions

Thank you

You make **possible**