Review of Terabit/sec SDN demonstrations at Supercomputing 2015 and plans for SC16

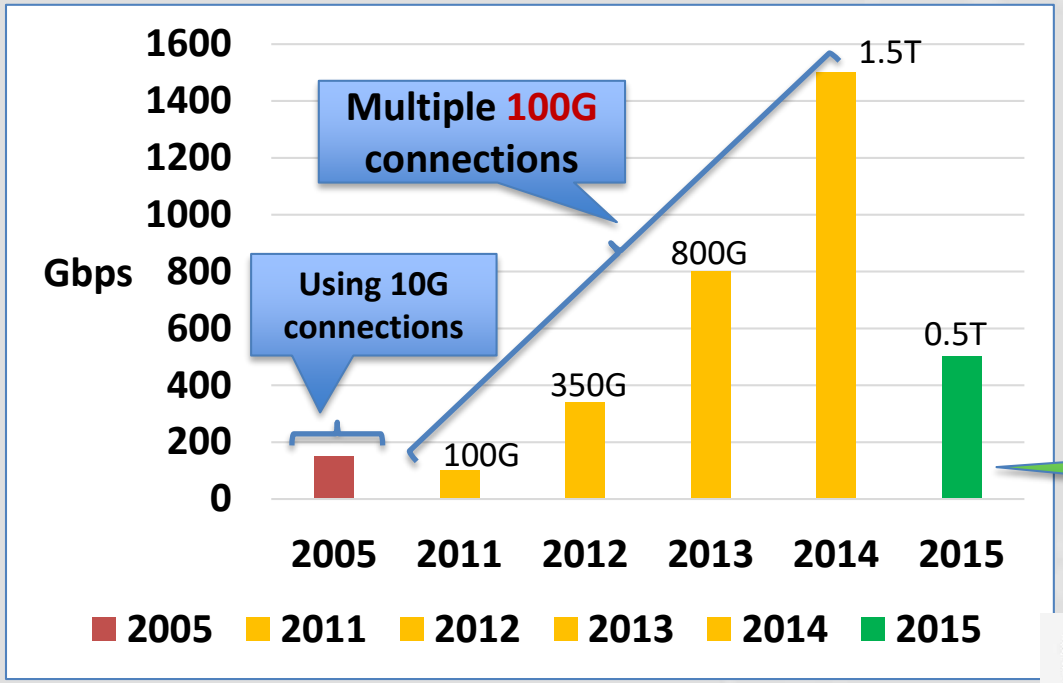
CHEP 2016 Conference, San Francisco,
October 8-14, 2016

Azher Mughal

Caltech

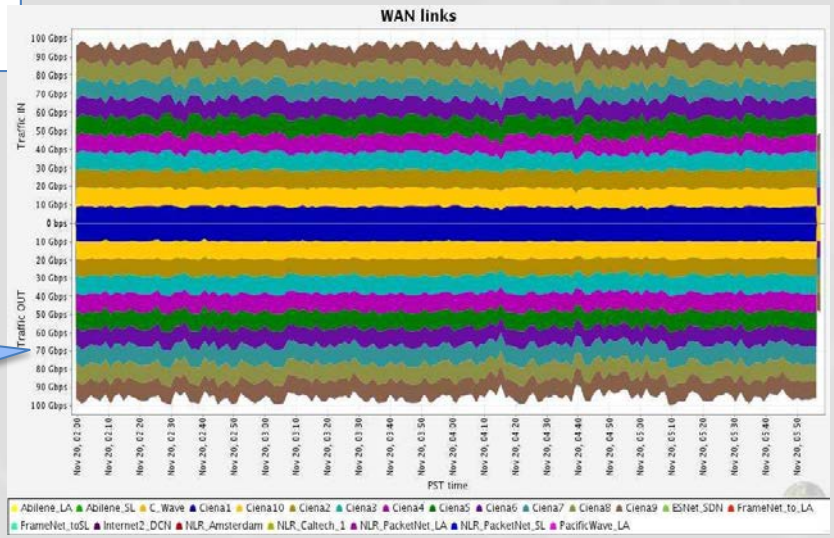
http://supercomputing.caltech.edu/

Bandwidth explosions by Caltech at SC



SC05 (Seattle): 155Gbps
SC11 (Seattle): 100Gbps
SC12 (Salt Lake): 350Gbps
SC13 (Denver): 800Gbps
SC14 (Louisiana): 1.5Tbps
SC15 (Austin): ~ 500Gbps
SC15 (Salt Lake): ~ 2.5Tbps

Fully SDN enabled



First ever 100G OTU-4 trials using Ciena laid over multiple Using 10GE connections in 2008

SC15 Demonstration Goals

SDN Traffic Flows

- Network should solely be controlled by the SDN application (*done partially*)
- Relying mostly on the North Bound interface
  - Install flows among a pair of DTN nodes
  - Re-engineer flows crossing alternate routes across the ring (shortest or with more bandwidth) (*Stability issues: ODL / OF Agent in HW, NICs / Cables*)

NSI WAN Paths

- Connect Caltech booth with the remote sites: FIU, RNP, UMich.

High Speed DTN Transfers

- 100G to 100G (Network to Network)
- 100G to 100G (Disk to Disk)
- 400G to 400G (Network to Network) (*limited to 280Gbps*)

NDN

- Provide/Announce another set of Caltech LHC data (on NDN server) from Austin. Thus clients may find shorter paths and retrieve from show floor.
- Display the overlay traffic flow map designed by CSU. (*not deployed, resource issues*)

SC16 Demonstration Goals

CALTECH HEP NETWORKING

SDN Traffic Flows

- Network should solely be controlled by the SDN application.
- Relying mostly on the North Bound interface
  - Install flows among a pair of DTN nodes
  - Re-engineer flows crossing alternate routes across the ring (shortest or with more bandwidth)

High Speed DTN Transfers

- 100G to 100G (Network to Network)
- 100G to 100G (Disk to Disk)
- 1 Tbps to 1Tbps (Network to Network using RoCE and TCP)

ODL (PhEDEx + ALTO (RSA + SPCE)

- Substantially extended OpenDaylight controller using a unified multilevel control plane programming framework to drive the new network paradigm
- Advanced integration functions with the data management applications of the CMS experiment

OpenDaylight & Caltech/YALE SDN Initiatives

CALTECH HEP NETWORKING

Supporting:

- Northbound and South bound interfaces
- Starting with Lithium, Intelligent services likes ALTO, SPCE, RSA
- OVSDB for OpenVSwitch Configuration, including the northbound interface

MAPLE:

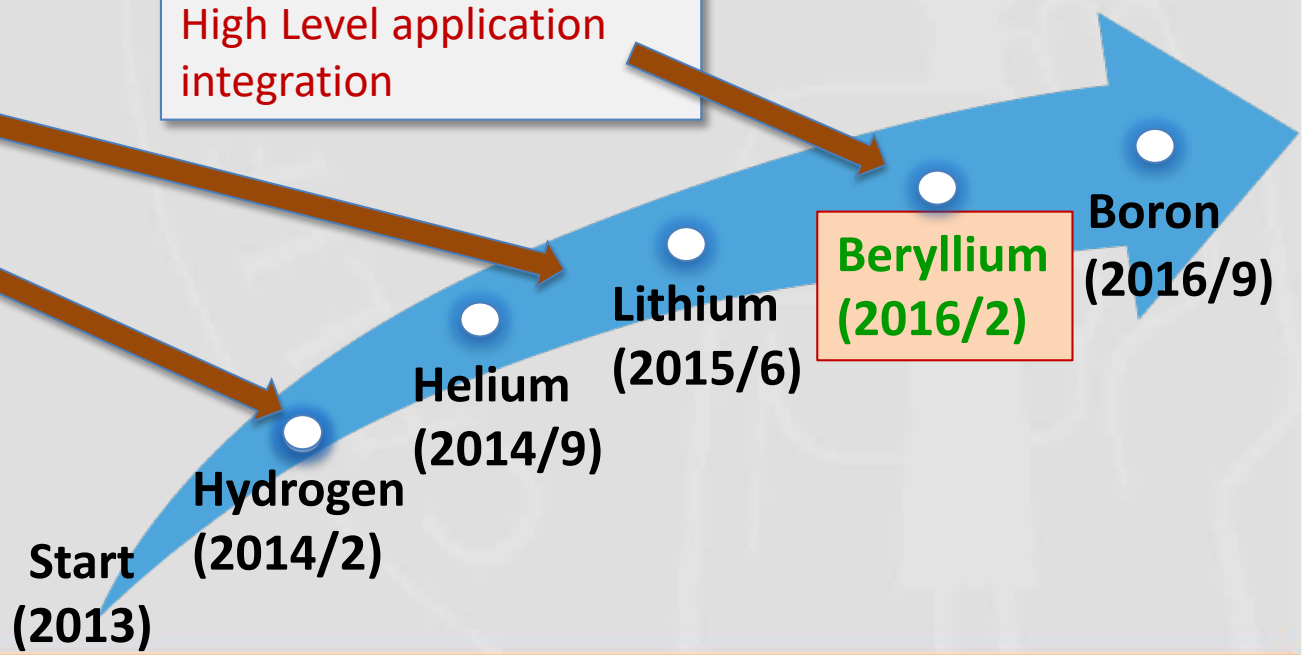
- Rapid application development platform for OpenDaylight, provide an abstraction from Java/environment build complexities

OFNG – ODL:
NB libraries for Helium/Lithium

OLiMPs – ODL:
Migrated to Hydrogen

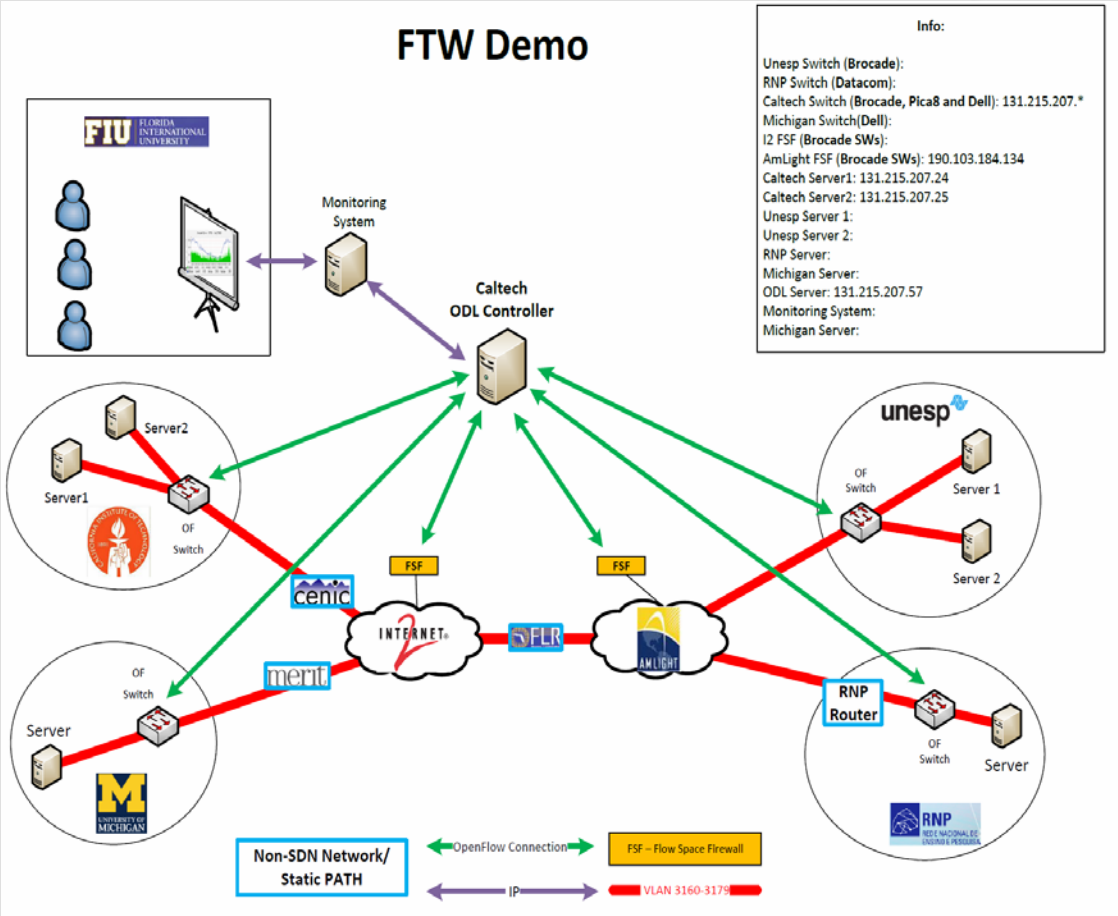
OLiMPs – FloodLight:
Link-layer MultiPath Switching

OFNG - ALTO
High Level application integration



Intelligent SDN-Driven Multipath Circuits

OpenDaylight/OpenFlow Controller Int'l Demo

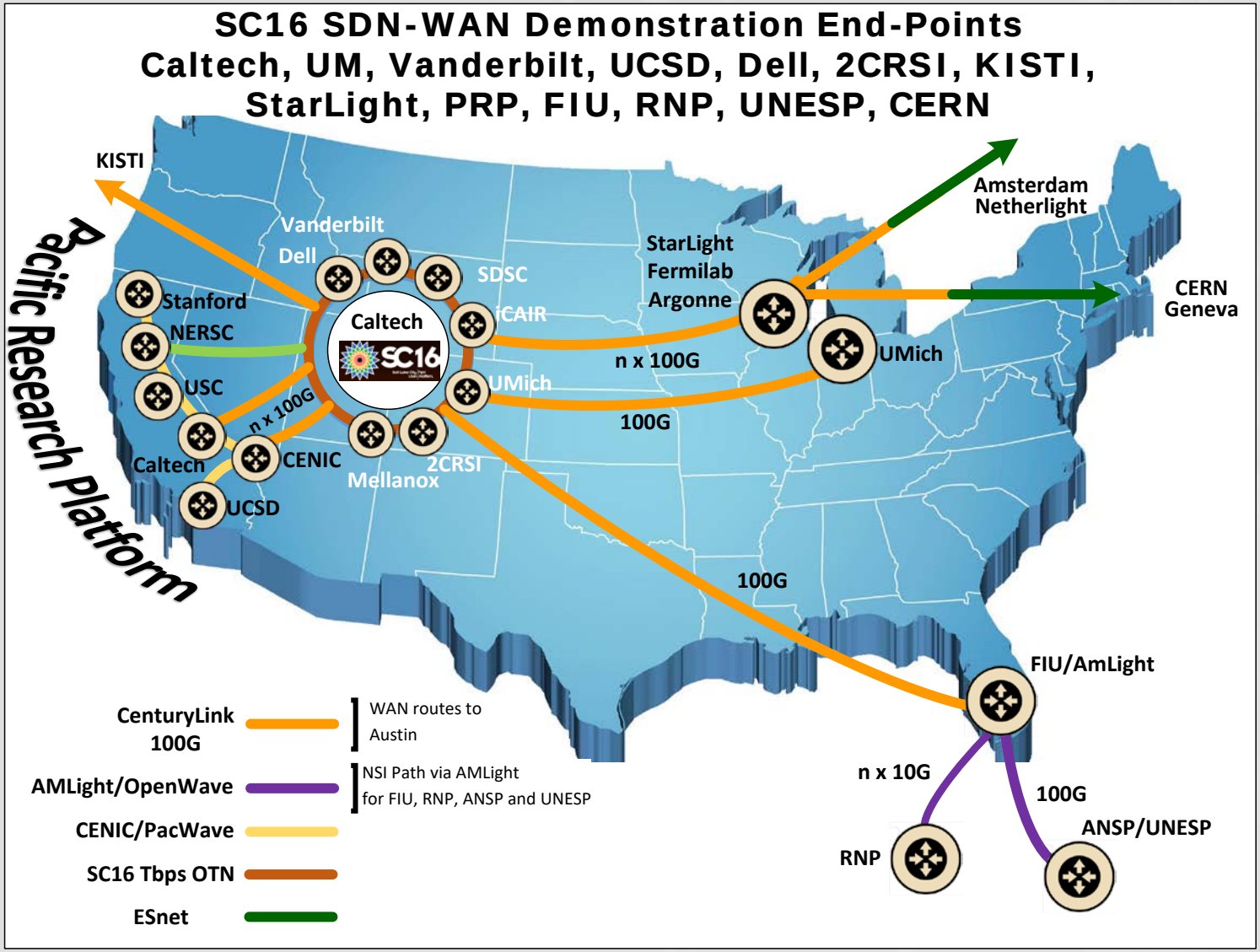


- At the March 31 – April 2 AmLight/Internet2 Focused SDN Workshop
- Dynamic path control under SDN flow rules installed in the switches
- Prelude to ANSE architecture of load-balanced, moderated flows across complex networks for LHC and other data intensive sciences

Caltech, Michigan, FIU, Rio and Sao Paulo, with Network Partners: Internet2, CENIC, Merit, FLR, AmLight, RNP and ANSP in Brazil

SC16 Architecture and Planning

SC16 – Remote End Sites

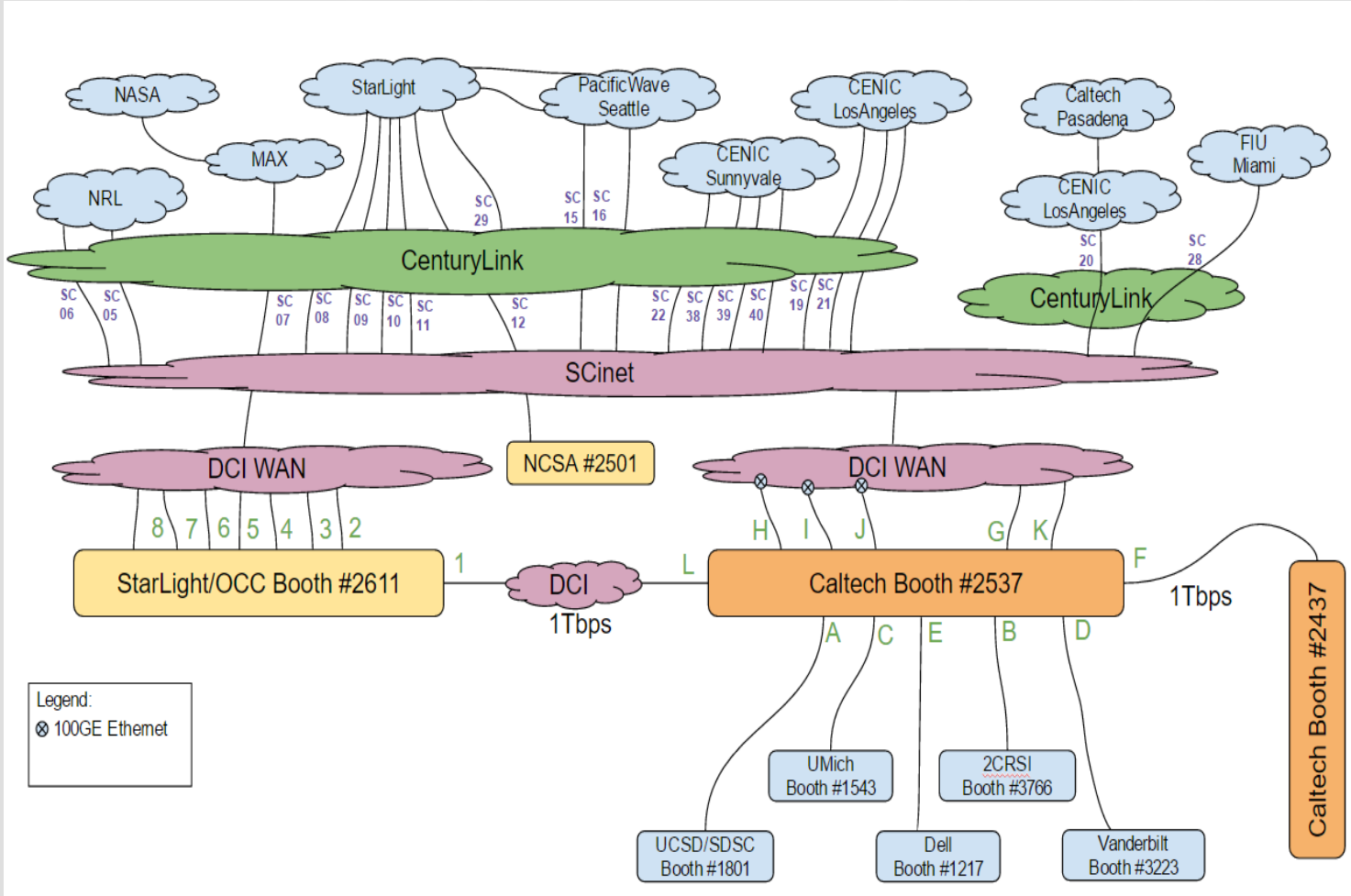


SC16 – Network Connection Providers

SCinet OTN Connection

- All the connections to booths are through the OTN Metro DCI Connections
- 1Tbps Booths
  - Caltech
  - StarLight
  - SCinet
- 100GE Booths
  - 2CRSi
  - Dell
  - UCSD
  - UMich
  - Vanderbilt

Connections
  - 5 x WAN
  - 5 x Dark Fiber
  - 2 x 1Tbps

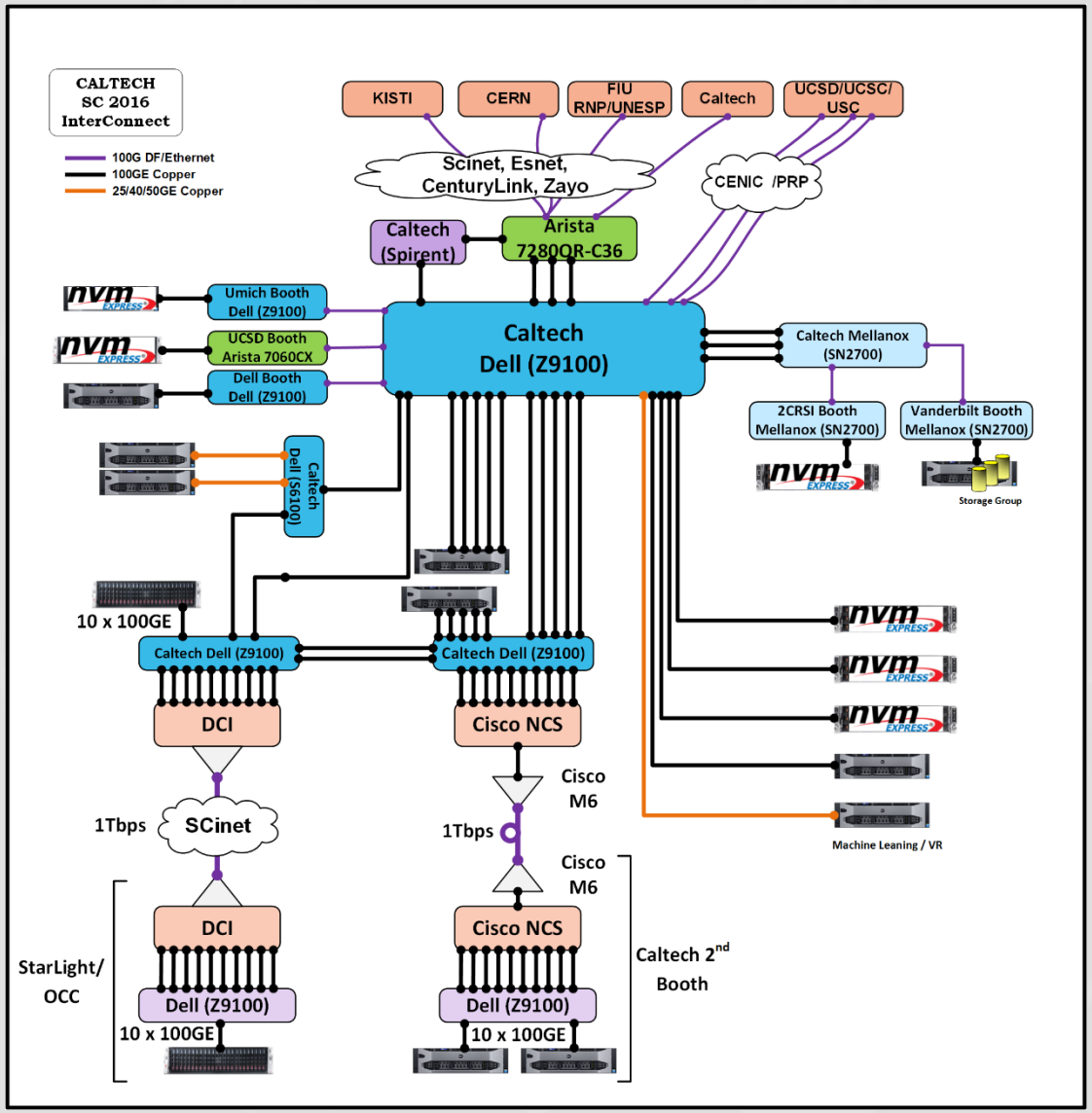


SC16 Network Layout

- All the switches are OF Compliant
- DTN Servers with 25/40/100GE
- DTN Servers are using NVME and connected in NVME fabric structure

There are two 1Tbps links:

- RDMA based: Caltech – StarLight
- TCP based: Between two Caltech Booths



Beyond 100GE -> 200/400GE, Component readiness ?

Server Readiness:

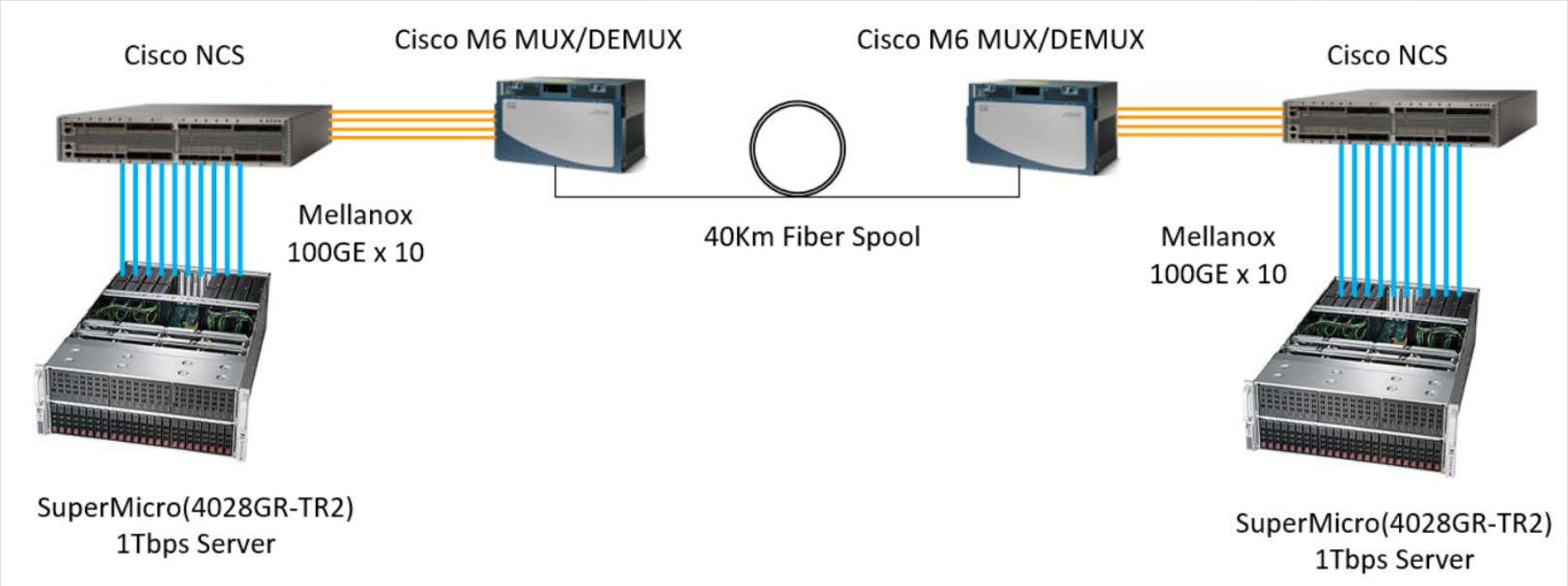
1) Current PCIe Bus limitations
   - PCIe Gen 3.0 (x16 can reach 128Gbs Full Duplex)
   - PCIe Gen 4.0 (x16 can reach double the capacity, i.e. 256Gbps
   - PCIe Gen 4.0 (x32 can reach double the capacity, i.e. 512Gbps
2) Increased number of PCIe lanes within processor

   Haswell/Broadwell (2015/2016)
   - PCIe lanes per processor = 40
   - Supports PCIe Gen 3.0 (8GT/sec)
   - Up to DDR4 2400MHz memory

   Skylake (2017)
   - PCIe lanes per processor = 48
   - Supports PCIe Gen 4.0 (16GT/sec)
3) Faster core rates, or Over clocking (what's best for production systems)
4) Increased memory controllers at higher clock rate reaching 3000MHz
5) TCP / UDP / RDMA over Ethernet

SC16 – 1 Tbps network Layout

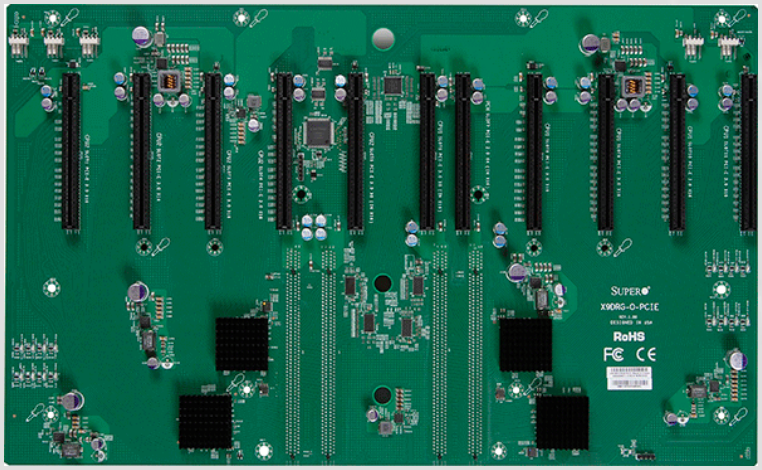
- To demonstrate the maximum throughput from each of the server.
- But to also highlight the current peak limitations in the server hardware, PCIe lanes and PCIe fabric design.
- Still there are challenges on the user memory to network RDMA due to limited PCIe slots

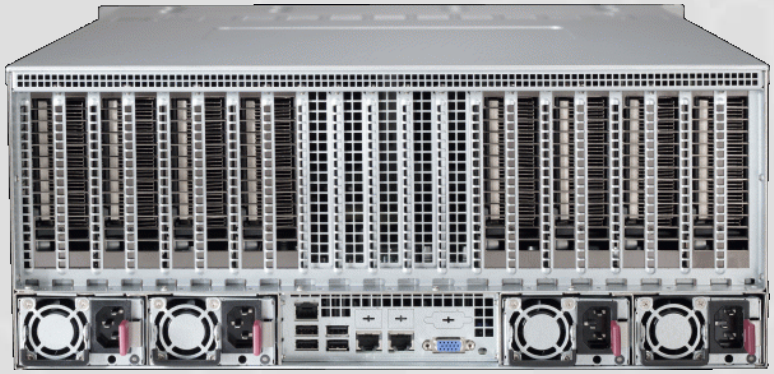


Server Design

SYS-4028GR-TR2

Expansion Board X9DRG-O-PCI-E
(PCIe Gen3 x16 version)





1Tbps Network Server

Slot Layout

GPU Slots:

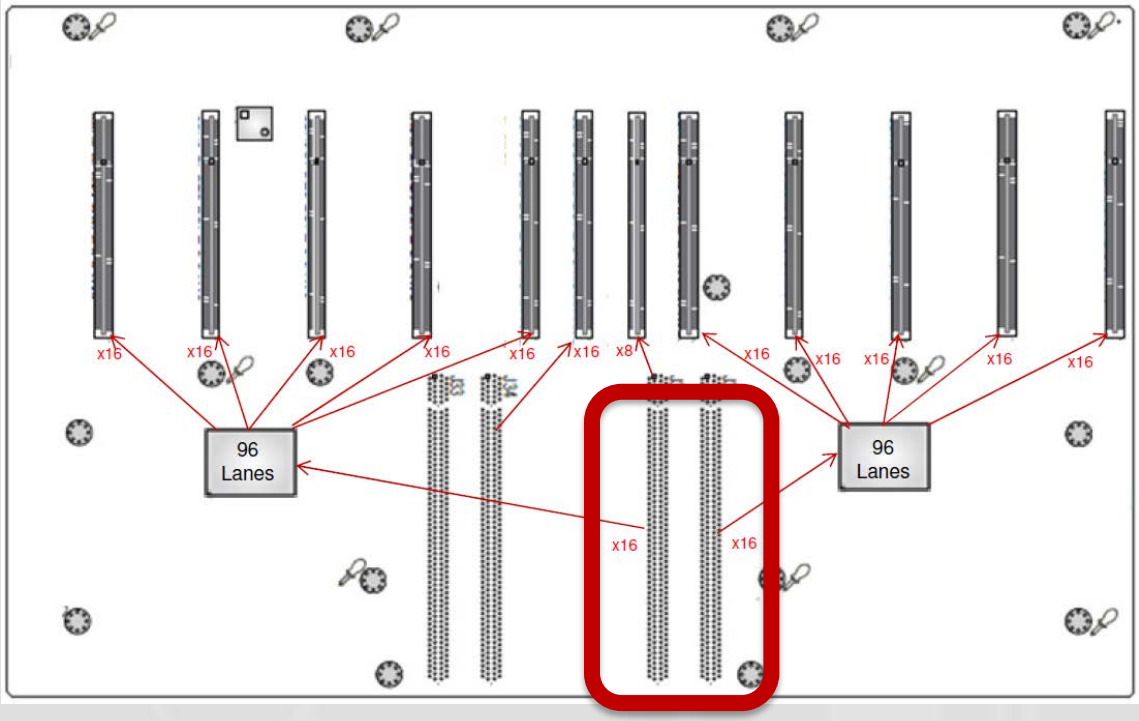
1) CPU-0 = 10 Slots

- Two PCIe Switching chips, each using one x16 Gen3 from CPU-0
- Each switching chip provides 5 x16 slots

Additional Slots:

1) CPU-0 = one x8 slot

2) CPU-1 = one x16 slot



CPU-0

1Tbps Network Server

SM - Different slot and 10GE options

8 Slot GPU

4028GR-TR

Key Features

1. Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 **v4**†/ v3 family; QPI up to 9.6GT/s
2. Up to **3TB**† ECC 3DS LRDIMM , up to DDR4- **2400**†MHz ; 24x DIMM slots
3. Expansion slots:
   8 PCI-E 3.0 x16 (double-width) slots
   2 PCI-E 3.0 x8 (in x16) slots
   1 PCI-E 2.0 x4 (in x16) slot
4. Dual GbE LAN with Intel® i350
5. 24x 2.5" Hot-swap drive bays
6. 8x 92mm RPM Hot-Swap cooling fans
7. 1600W Redundant (2+2) Power Supplies; **Platinum Level (94%+)**

8 Slot GPU with 10GE Ports

4028GR-TRT

Key Features

1. Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 **v4**†/ v3 family; QPI up to 9.6GT/s
2. Up to **3TB**† ECC 3DS LRDIMM , up to DDR4- **2400**†MHz ; 24x DIMM slots
3. Expansion slots:
   8 PCI-E 3.0 x16 (double-width) slots
   2 PCI-E 3.0 x8 (in x16) slots
   1 PCI-E 2.0 x4 (in x16) slot
4. Dual 10GBase-T LAN with Intel® X540
5. 24x 2.5" Hot-swap drive bays
6. 8x 92mm RPM Hot-Swap cooling fans
7. 1600W Redundant (2+2) Power Supplies; **Platinum Level (94%+)**

Imp: Each chassis has other PCIe slots apart from the GPU slots.

10 Slot GPU

4028GR-TR2

Coming Soon

Key Features

- Artificial Intelligence
- Big Data Analytics
- High-performance Computing
- Research lab/National Lab
- Astrophysics
- Business Intelligence

1. Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 **v4**†/ v3 family; QPI up to 9.6GT/s
2. Up to **3TB**† ECC 3DS LRDIMM , up to DDR4- **2400**†MHz ; 24x DIMM slots
3. Expansion slots, **Single Root Complex**,
   11 PCI-E 3.0 x16 (FH, FL) slots
   1 PCI-E 3.0 x8 (in x16) slots
4. 2 GbE LAN with Intel® i350
5. 24x 2.5" Hot-swap drive bays
6. 8x 92mm RPM Hot-Swap cooling fans
7. 2000W Redundant (2+2) Power Supplies; **Titanium Level (96%+)**

10 Slot GPU with 10GE Ports

4028GR-TRT2

Coming Soon

Key Features

- Artificial Intelligence
- Big Data Analytics
- High-performance Computing
- Research lab/National Lab
- Astrophysics
- Business Intelligence

1. Dual socket R3 (LGA 2011) supports Intel® Xeon® processor E5-2600 **v4**†/ v3 family; QPI up to 9.6GT/s
2. Up to **3TB**† ECC 3DS LRDIMM , up to DDR4- **2400**†MHz ; 24x DIMM slots
3. Expansion slots, **Single Root Complex**,
   11 PCI-E 3.0 x16 (FH, FL) slots
   1 PCI-E 3.0 x8 (in x16) slots
4. 2 10GBase-T LAN with Intel® X540
5. 24x 2.5" Hot-swap drive bays
6. 8x 92mm RPM Hot-Swap cooling fans
7. 2000W Redundant (2+2) Power Supplies; **Titanium Level (96%+)**

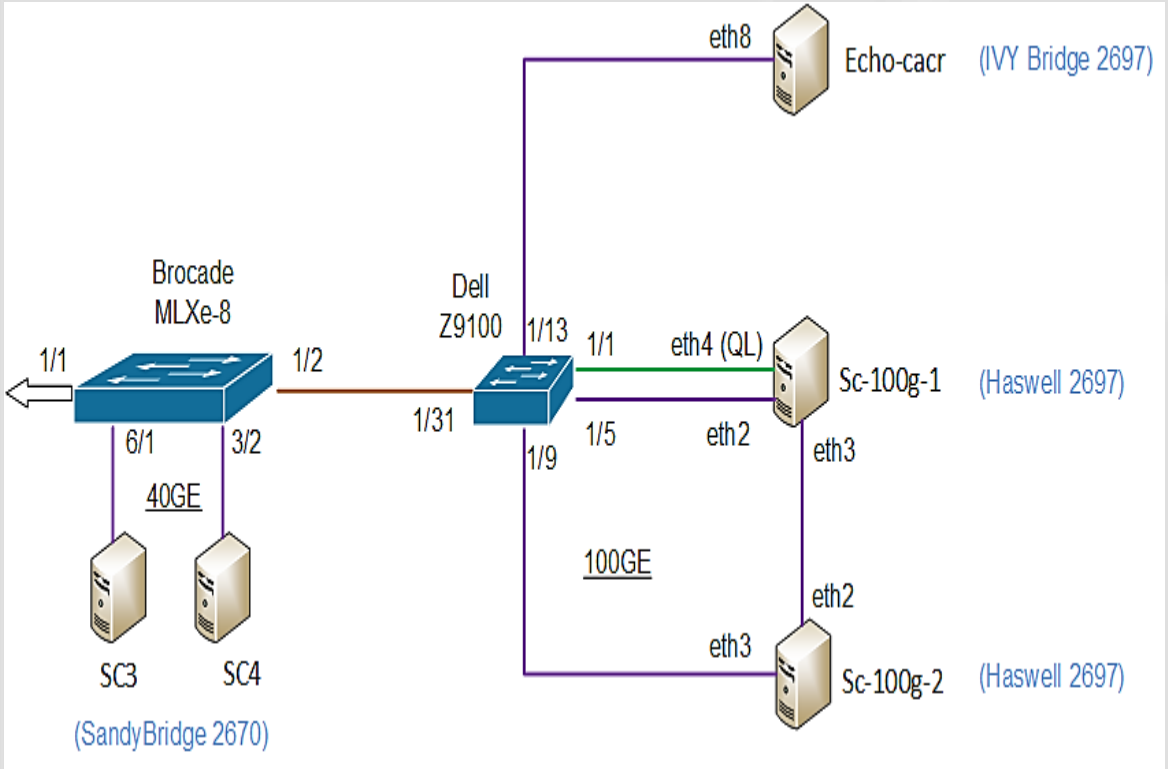
DTN Design Considerations

Different Choices and Opinions …

- How many rack units are needed / available.
- Single socket vs dual socket systems
- Many cores vs Less cores at high clock rates
- SATA 3 RAID Controllers vs HBA Adapters vs NVME
- White box servers vs servers from traditional vendors (design flexibility ?)
- How many PCIe slots are needed (I/O + network). What should be the slot width (x16 for 100GE)
- Onboard Networks cards vs add-on cards
- Airflow for heat load inside the chassis for different work loads (enough fans ?)
- Processor upgradeable motherboard
- Remote BMC / ILOM / IPMI connectivity
- BIOS Tweaking
- Choice of Operating system, kernel flavors
- Redundant power supply

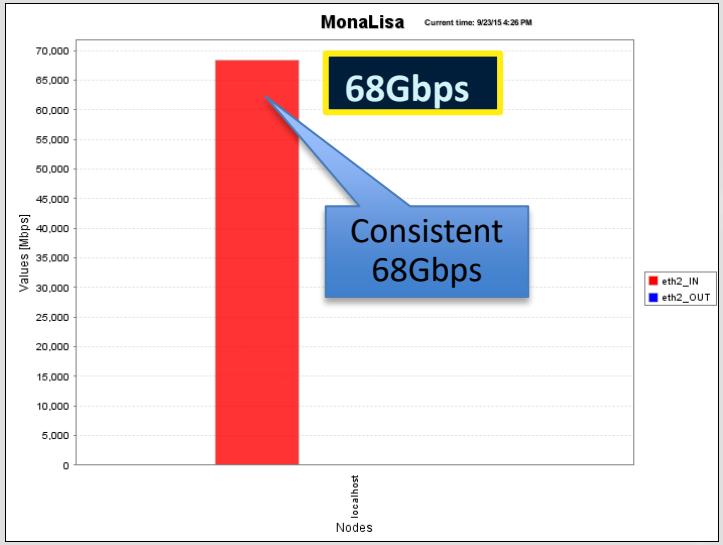
Pre-SC15 Network Performance Tests

Single and multiple TCP Stream
Testing across a variety of Intel processors



- Two Identical Haswell Servers:
  - E5-2697 v3
  - X10DRi Motherboard
  - 128GB DDR4 RAM
  - Mellanox VPI NICs
  - QLogic NICs
  - CentOS 7.1
- Dell Z9100 100GE Switch
- 100G CR4 Copper Cables from Elpeus
- 100G CR4 Cables from Mellanox for back to back connections

Single TCP Stream

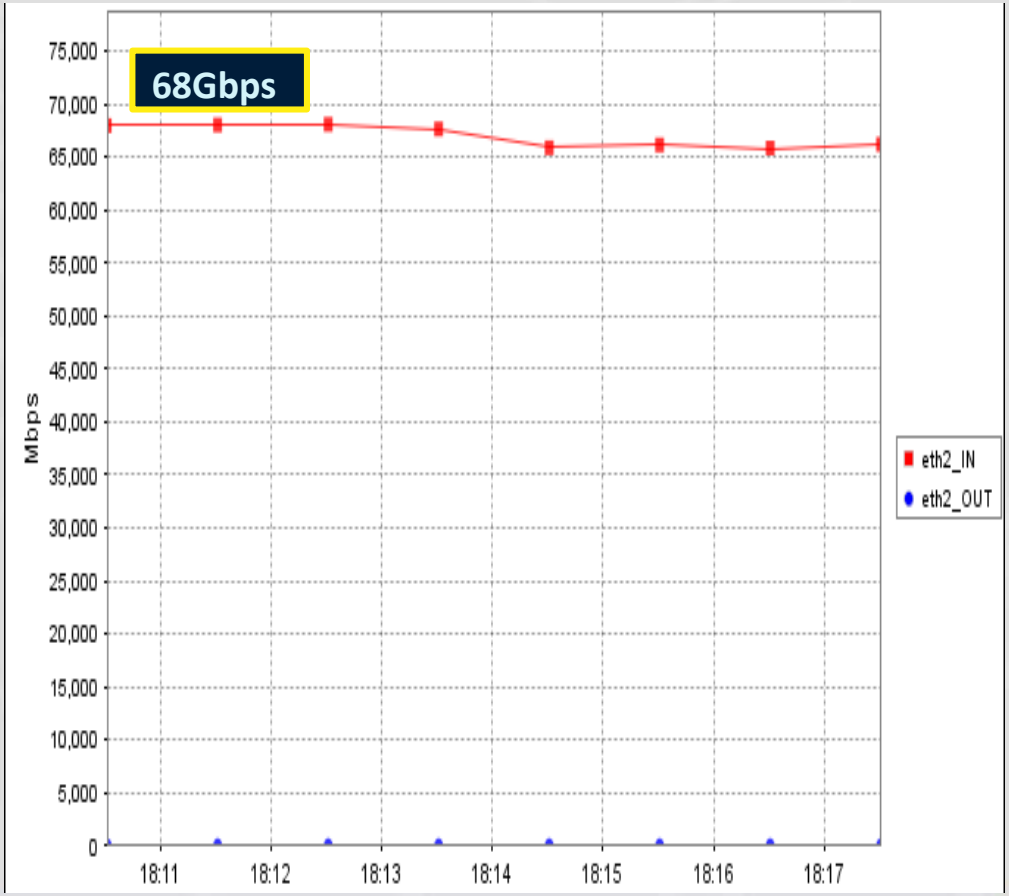


Client

[root@sc100G-1 ~]# numactl --physcpubind=20 --localalloc java -jar fdt.jar -c 1.1.1.2 -netteP 1 -p 7000

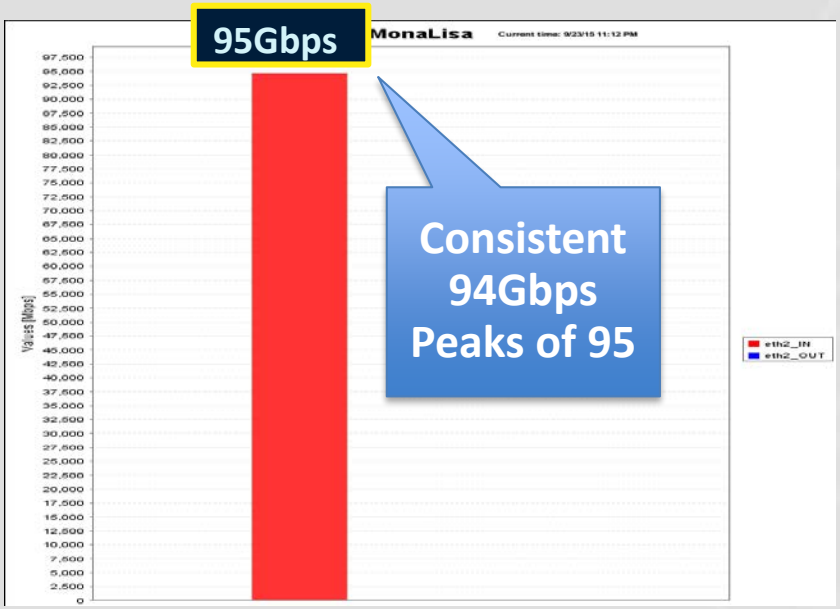
Server

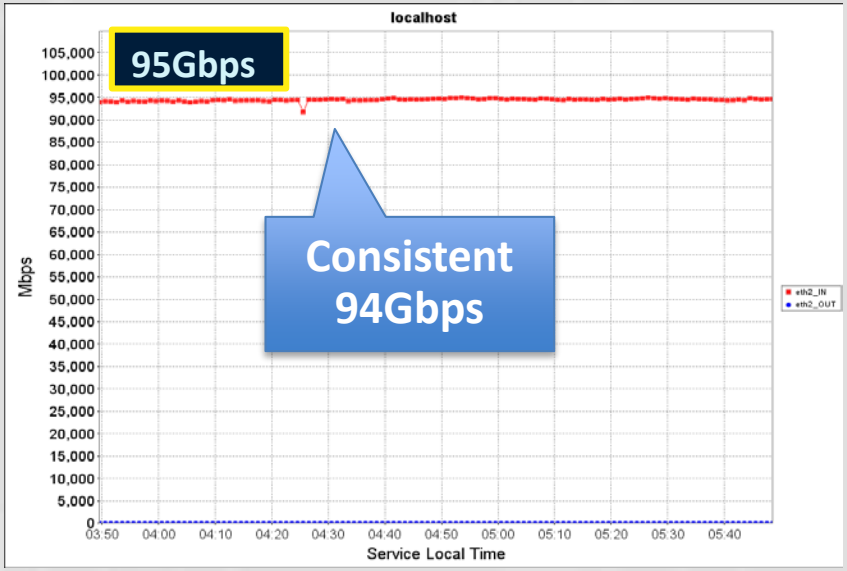
[root@sc100G-2 ~]# numactl --nphyscpubind=20 --localalloc java -jar fdt.jar -p 7000



16:22:04	*Net Out: 67.822 Gb/s*	*Avg: 67.822 Gb/s*
16:22:09	*Net Out: 68.126 Gb/s*	*Avg: 67.974 Gb/s*
16:22:14	*Net Out: 68.159 Gb/s*	*Avg: 68.036 Gb/s*
16:22:19	*Net Out: 68.057 Gb/s*	*Avg: 68.038 Gb/s*
16:22:24	*Net Out: 68.133 Gb/s*	*Avg: 68.057 Gb/s*
16:22:29	*Net Out: 68.349 Gb/s*	*Avg: 68.103 Gb/s*
16:22:34	*Net Out: 68.161 Gb/s*	*Avg: 68.111 Gb/s*
16:22:39	*Net Out: 68.027 Gb/s*	*Avg: 68.101 Gb/s*

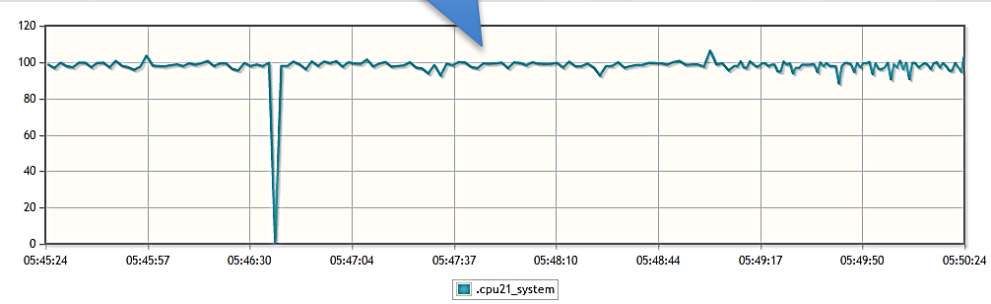
Two TCP Streams in Parallel



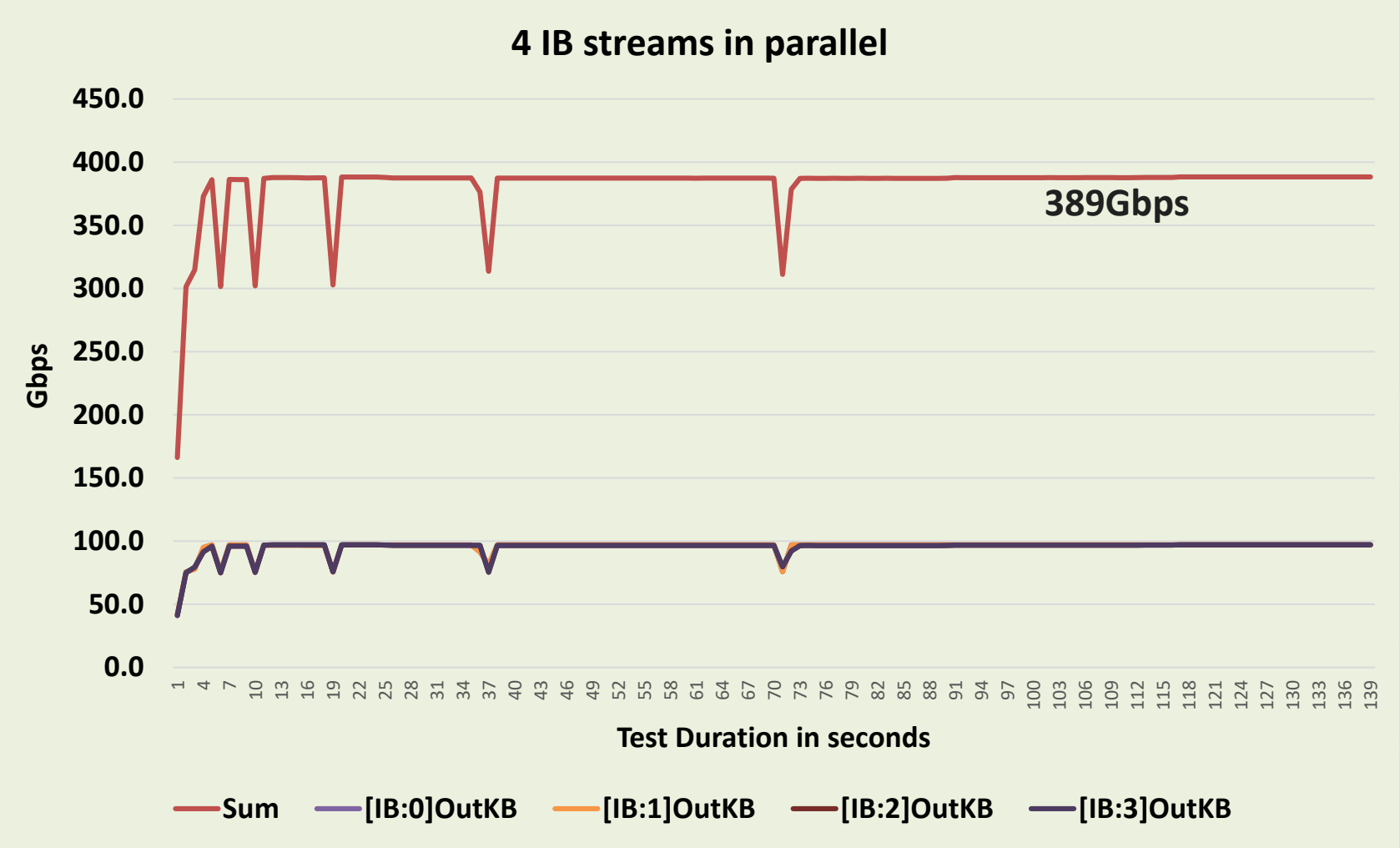


24/09 04:53:29	*Net Out: 94.629 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:34	*Net Out: 94.660 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:39	*Net Out: 94.249 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:44	*Net Out: 94.652 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:49	*Net Out: 94.521 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:54	*Net Out: 94.325 Gb/s*	*Avg: 93.744 Gb/s*
24/09 04:53:59	*Net Out: 94.033 Gb/s*	*Avg: 93.744 Gb/s*
24/09 04:54:04	*Net Out: 94.464 Gb/s*	*Avg: 93.744 Gb/s*
24/09 04:54:09	*Net Out: 94.682 Gb/s*	*Avg: 93.744 Gb/s*

Single Core at 100% (two cores used)



400GE Network Testing



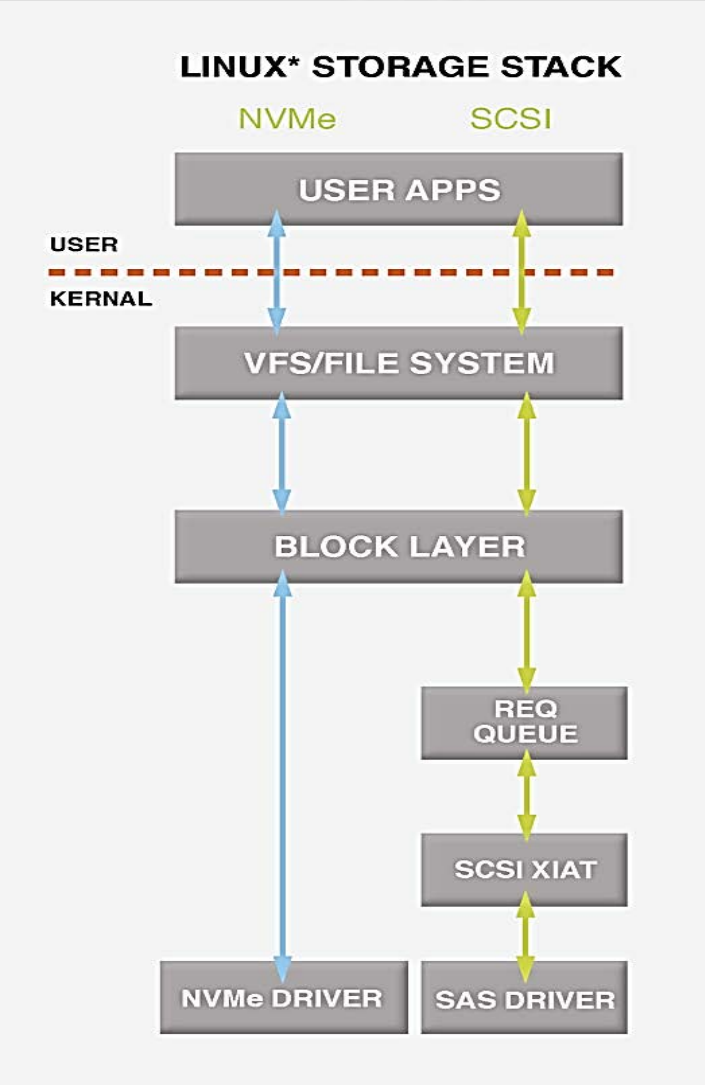
Transmission across 4 Mellnox VPI NICs.

Only 4 CPU cores are used out of 24 cores.

Why to choose NVME based storage devices

NVME Advantages

- Bypasses AHCI / SCSI layers
- Extremely fast (dedicated FPGA) (Seagate recently announced 10GB/sec drive
- Low latency
- Supported by large number of vendors
- Generally available in both 2.5” or PCIe cards form factor (PCIe Gen3 x4/x8/x16)
- Prices are getting low:
  - Sata3 SSDs are about 24 .. 40 cents per GB
  - NVME are about $2 per GB (expensive)



Drives vs Bandwidth Requirement

				Write Performance
10GE			=	13 Gbps
25GE	2 x		=	2 x 14.4 = 28.8 Gbps
40GE	3 x		=	3 x 14.4 = 43.2 Gbps
50GE	4 x		=	4 x 14.4 = 57.6 Gbps
80GE	3 x		=	3 x 28 = 84 Gbps
100GE	4 x		=	4 x 28 = 112 Gbps

NVMe Drive Options

Pros:

Many times faster than standard SSD drives

Cons:

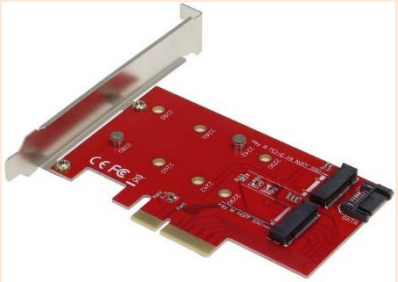
Expensive

2.4 GB/s write

MX 6300 (x8)

Mangstor

Samsung M.2 (PCIe x.2 width



PCIe x4 Adapter (supports two M.2 cards)

2.8 GB/s write

DC P3608 (x8)

intel

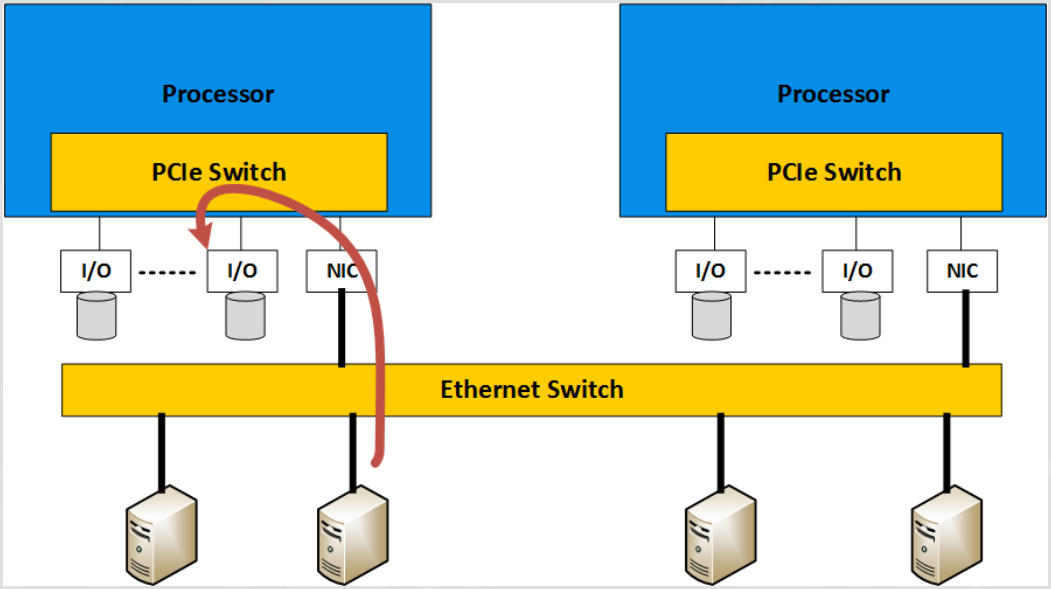
1.75 GB/s write

DC P3700 (x4)

DC P3700 (x4)

NVME Over Fabrics

- Specifications were published in 2016
- Initial code is available starting with Linux kernel 4.7
- Current NVME transports:
  - Fiber Channel
  - NVME fabric is built over exiting



Goals of NVMe over Fabrics is to extend the low-latency efficient NVMe block storage protocol with no more additional 10uSec.

NVMe over Fabrics maintains the architecture and software consistency of the NVMe protocol across different fabric types, providing the benefits of NVMe regardless of the fabric type or the type of non-volatile memory used in the storage target.

http://www.nvmexpress.org/wp-content/uploads/NVMe_Over_Fabrics.pdf

2CRSI / SuperMicro 2U - NVME Servers

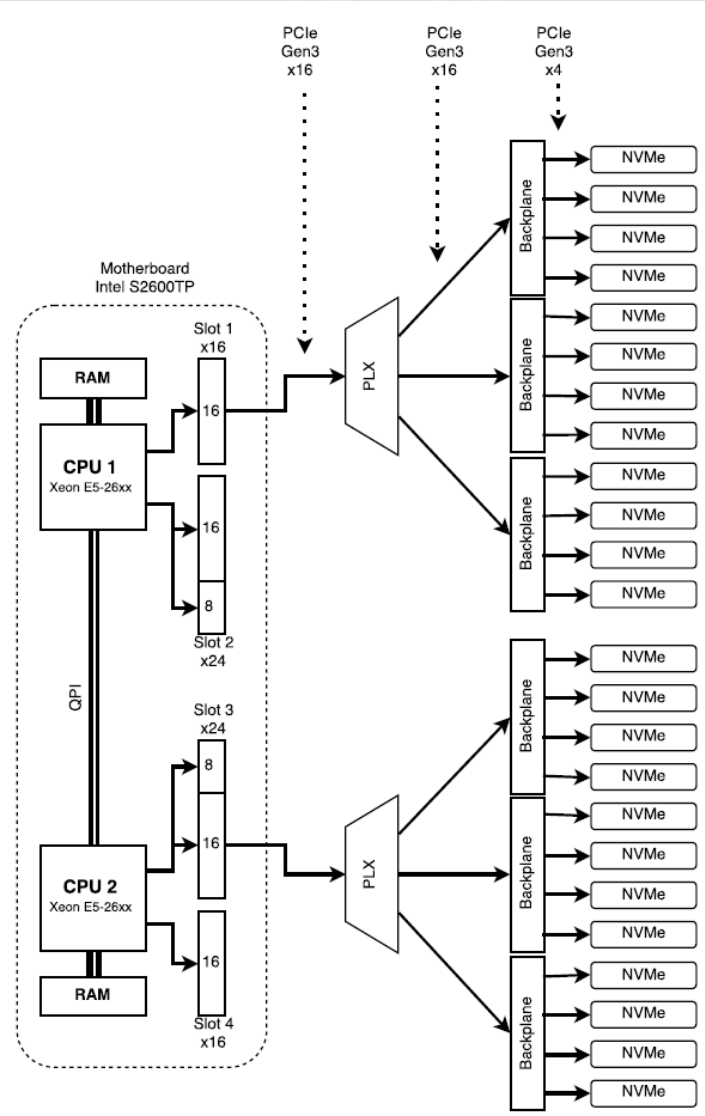
- Both servers are capable to drive 24 NVME drives. SuperMicro also have a 48 drive version.
- Standard 2.5” NVME drive option

2CRSI

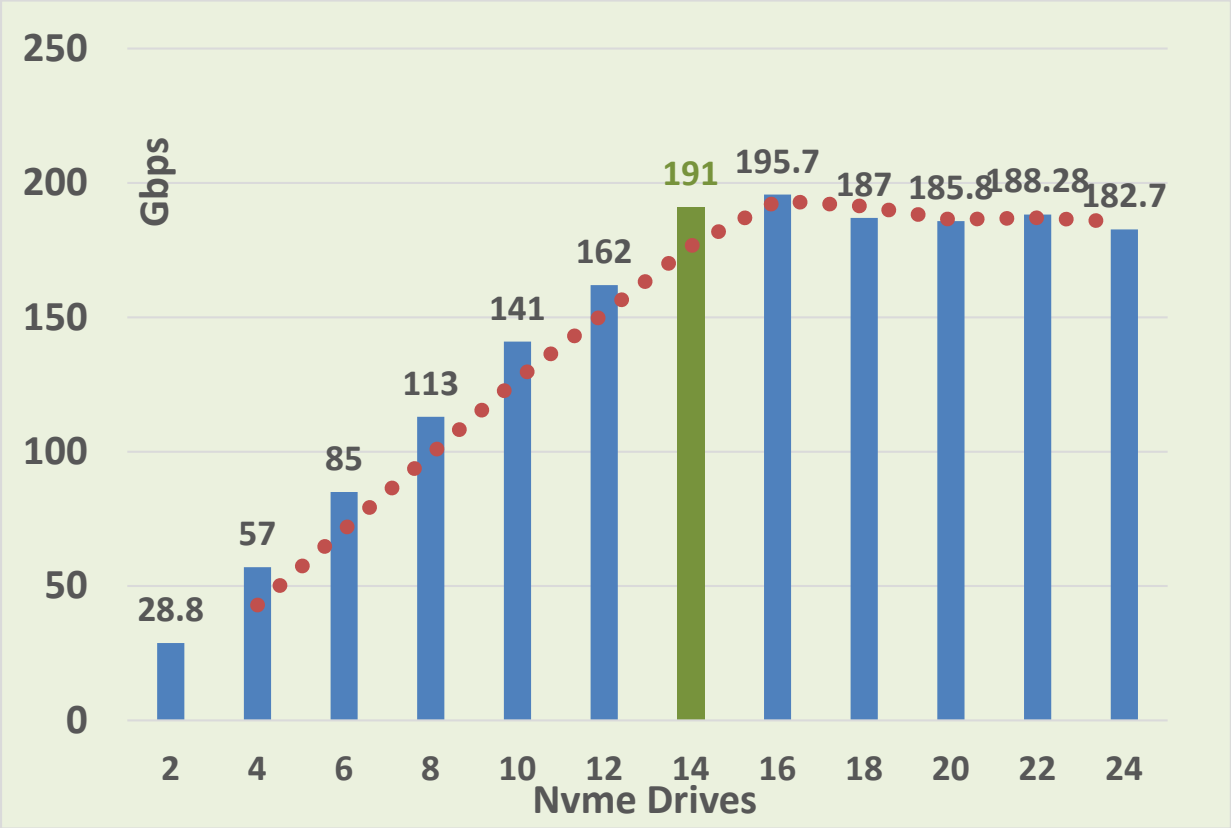
SuperMicro

2.5” NVME Drive

PCIe Switching Chipset for NVME



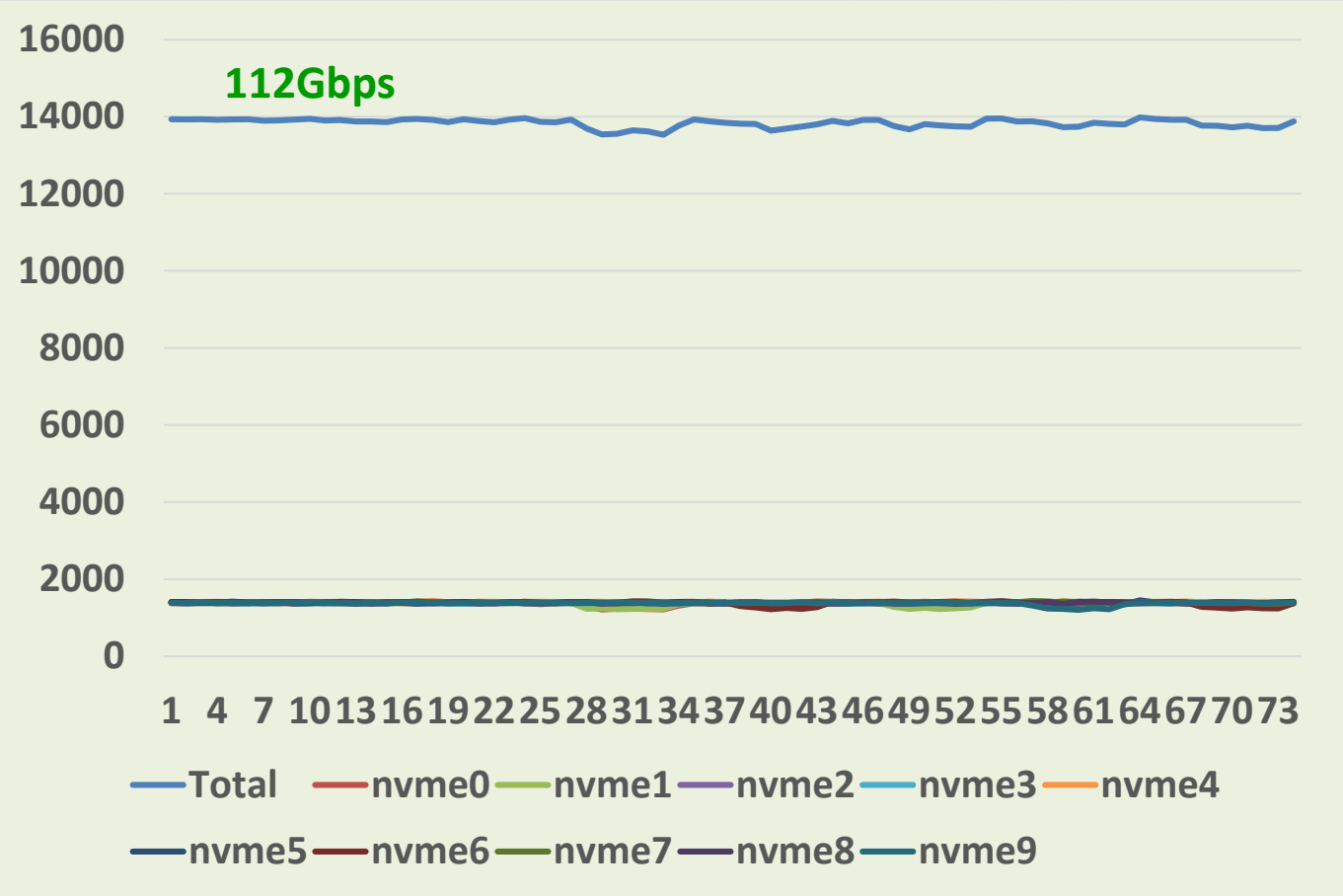
2CRSI Server with 24 NVMe drives

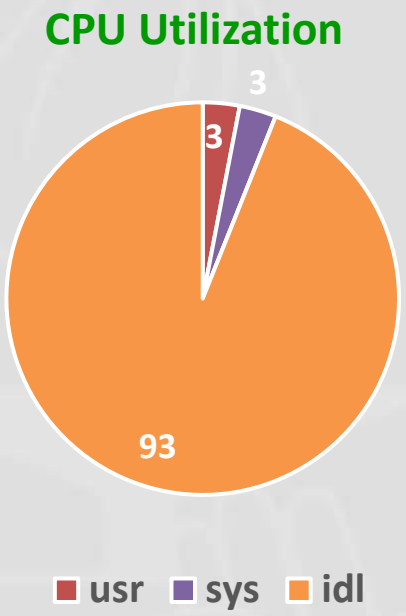


Max throughput reached at 14 drives (7 drives per processor)

A limitation due to combination of single PCIe x16 bus (128Gbps), processor utilization and application overheads.

Intel NVME Raw Disk Benchmarks

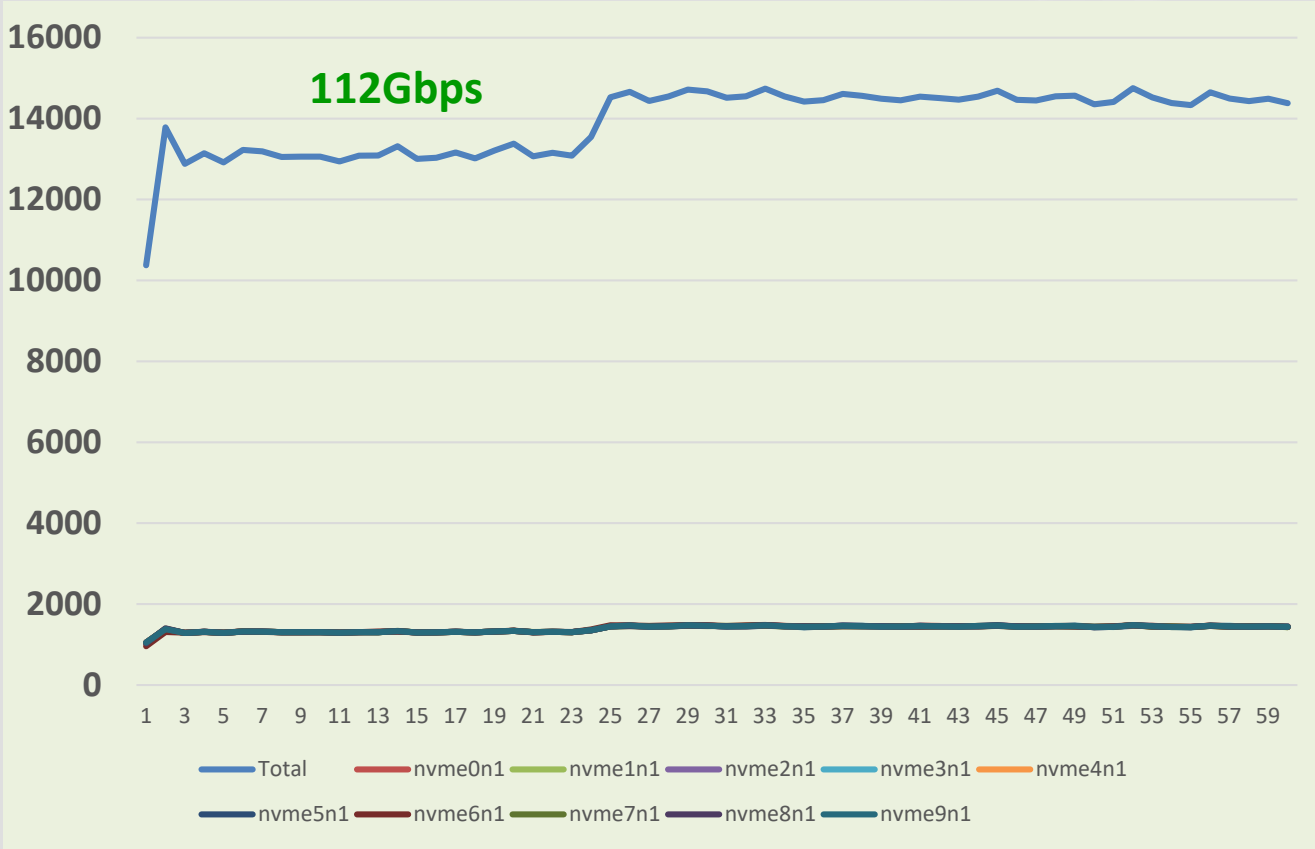


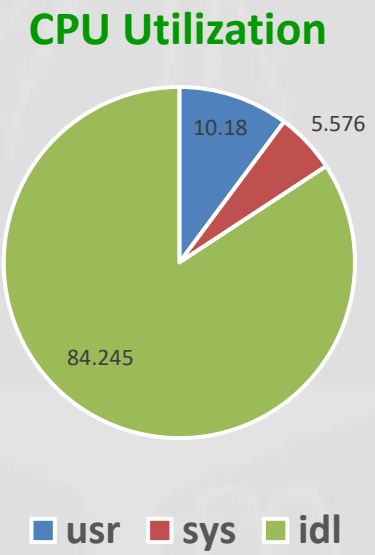


5 x Intel PC 3608 Drives. 5 x PCIe Gen3 x8 slots used on the motherboard.

Each drive is exported as two block devices to Operating system

Intel NVME (XFS, Container) Disk Benchmarks

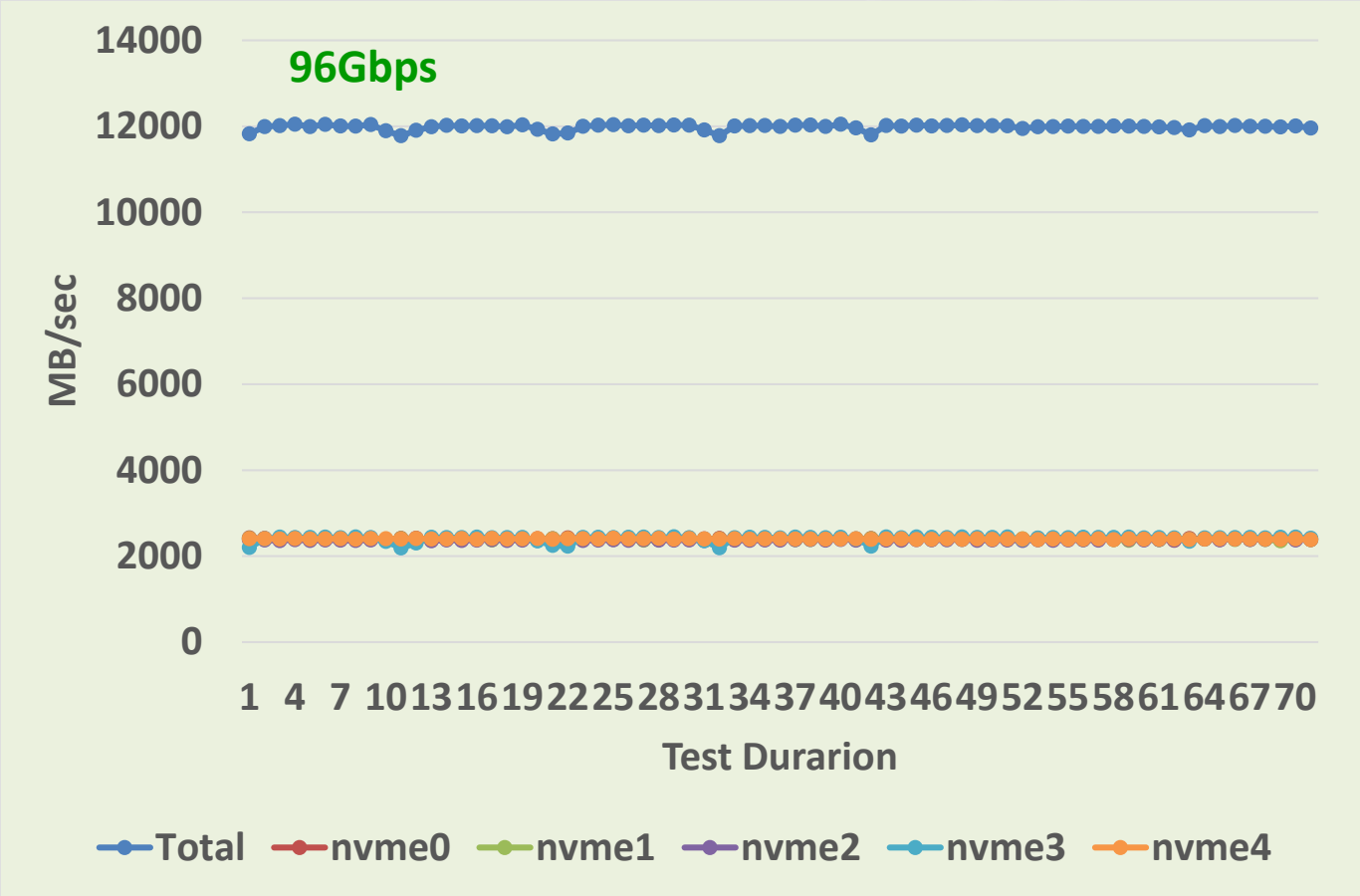


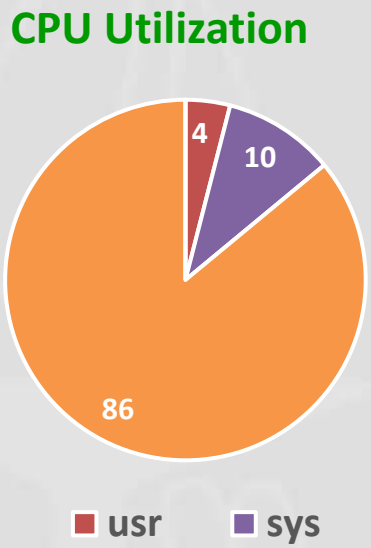


10 x NVME block devices.

5 x PCIe Gen3 x8 slots on the motherboard.

Mangstor NVME Raw Disk Benchmarks

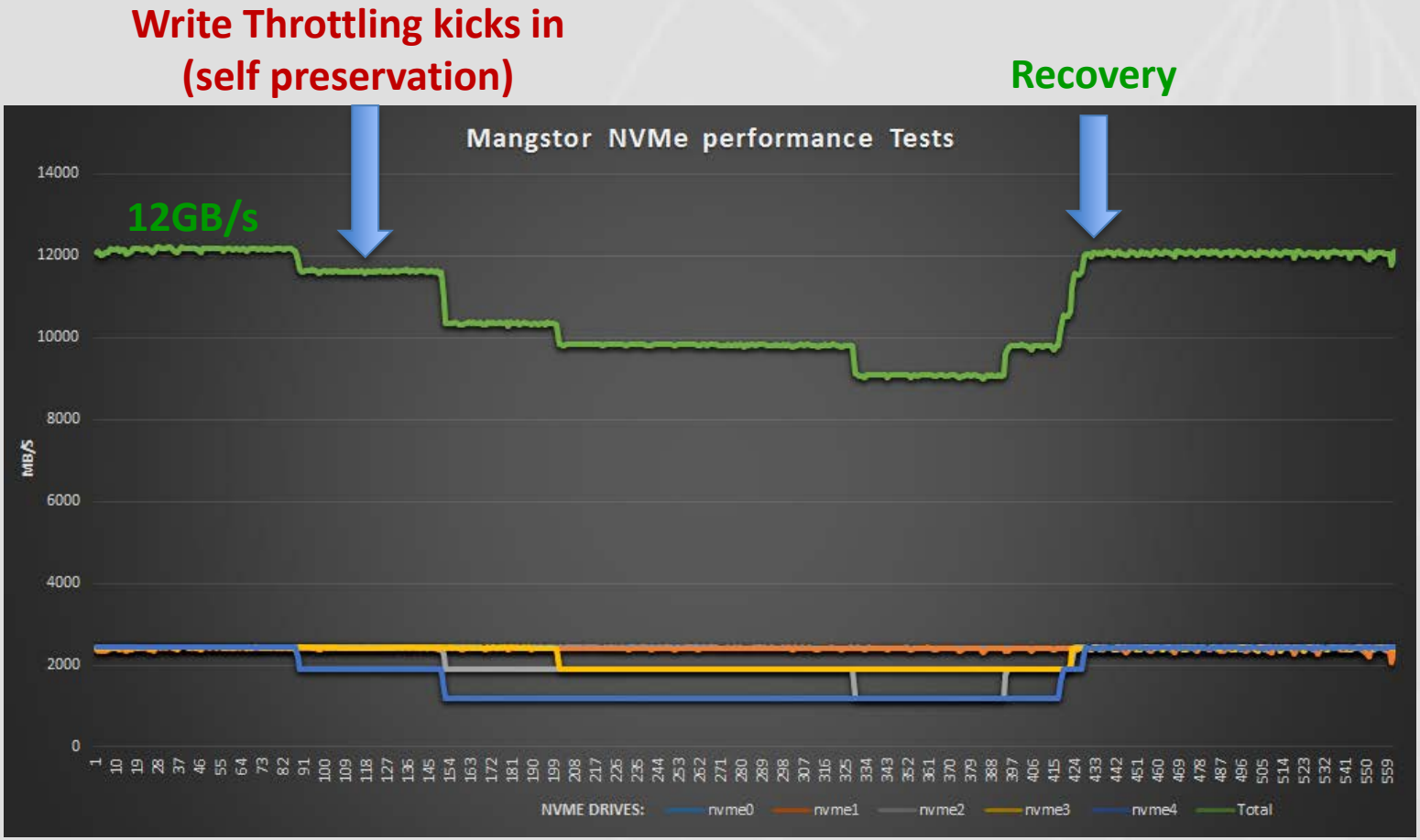




5 x NVME block devices.

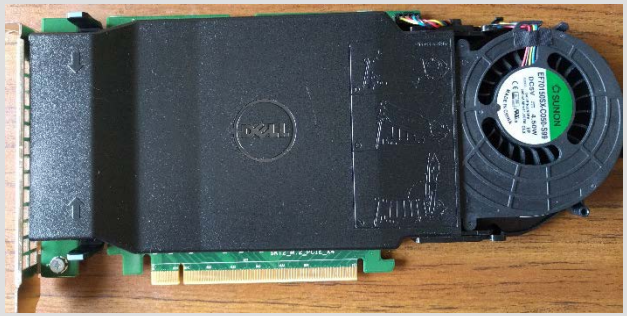
5 x PCIe Gen3 x8 slots on the motherboard.

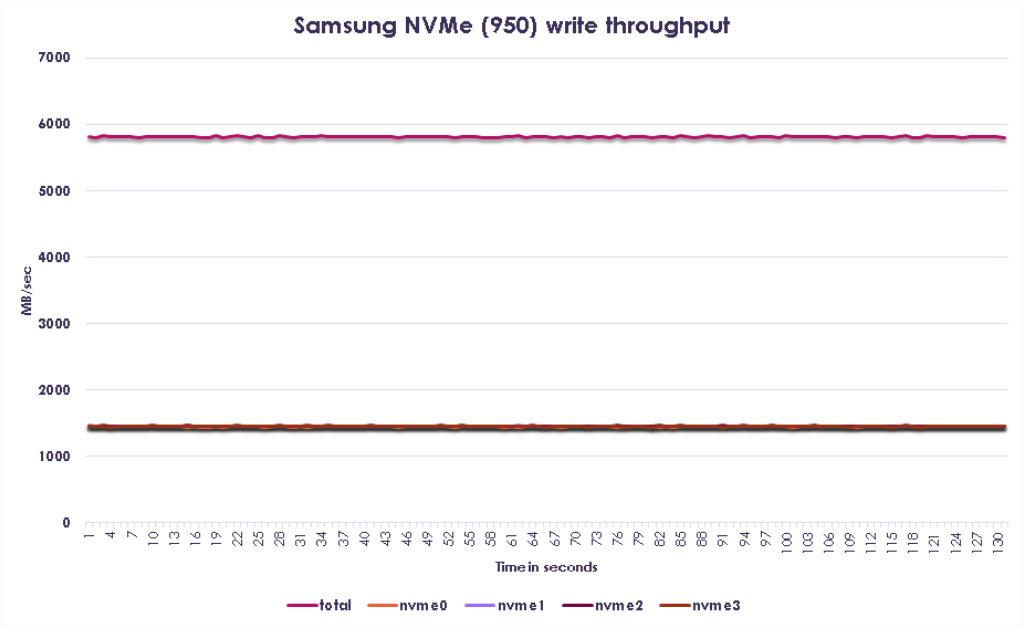
Temperature effects on SSD drives



Follow the manufacturer's guidelines for the minimum airflow requirement.

Let's build a low cost NVME storage





Ingredients:

- Dell quad M.2 adapter card
- Four Samsung 950 PRO M.2 drives

2TB NVMe Storage

~ $1496 or ~$1.3 per GB

~45 Gbps disk I/O

100GE Switches (compact form factor)

- 32 x 100GE Ports
- All ports are configurable
  - 10 / 25 / 40 / 50 / 100GE
- Arista, Dell, and Inventec are based on Broadcom Tomahawk (TH) chip, while Mellanox is using their own spectrum chipset
  - TH: Common 16MB packet buffer memory among 4 quadrants
  - 3.2 Tbps Full Duplex switching capacity
  - support ONIE Boot loader
- Dell /Arista supports two additional 10GE ports
- OpenFlow 1.3+, with multi-tenant support

Next Wave of switches using StrateDNX (Kumron, Jericho) are available in Arista 7280-X

100G Switches, Optics, Cables (confusions …)

- **100GE Optics (QSFP28 iLR4/LR4)**
  - Reliable vendors: InnoLight, Source Photonics
  - Not supported by Mellanox NICs
  - Supported by QLogic (using special knobs)
- **Link Auto Negotiation**
  - QLogic NIC (B0)
  - Mellanox does support
- **Forward Error Correction (FEC) (RS, Firecode)**
  - QLogic NIC (B0)
  - Mellanox does support FEC

Collaboration Partners

Special thanks to …

Research Partners

- Univ of Michigan
- UCSD
- iCAIR / StarLight
- Stanford
- Venderbilt
- UNESP / ANSP
- RNP
- Internet2
- ESnet
- CENIC
- FLR / FIU
- PacWave

Industry Partners

- Brocade (OpenFlow capable Switches)
- Dell (OpenFlow capable Switches)
- Dell (Server systems)
- Echostreams (Server systems)
- Intel (NVME SSD Drives)
- Mangstor (SSD storage)
- Mellanox (NICs and Cables)
- Spirent (100GE Tester)
- 2CRSI (NVME Storage)
- HGST Storage (NVME Storage)

Thank you !

Questions ?

azher@hep.caltech.edu