High Speed Scientific Data Transfers Using Software Driven Dynamic Networks

LJCONE Conference Taipei 2016

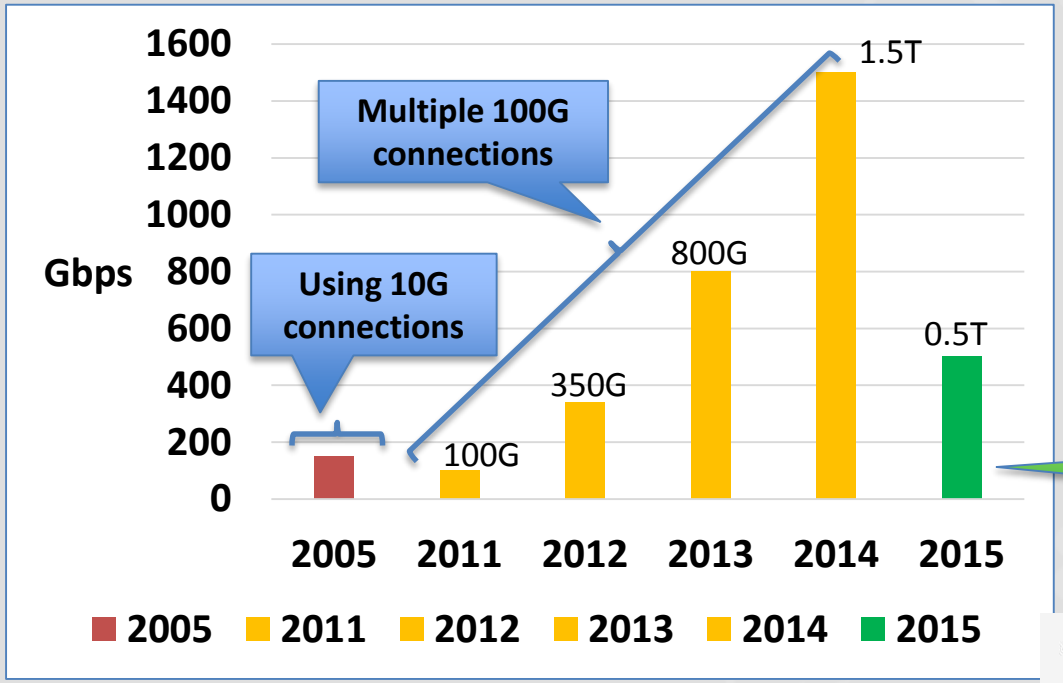
Azher Mughal
Caltech

http://supercomputing.caltech.edu/

Caltech LHCONE Activities

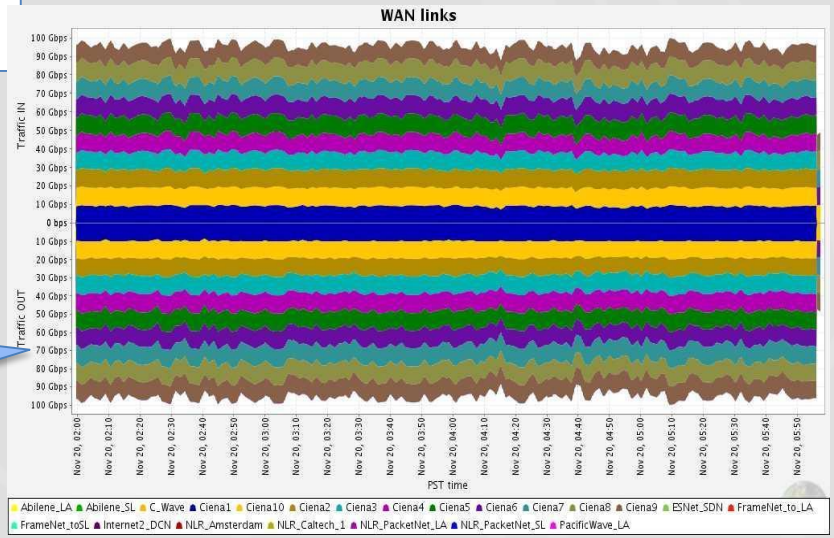
- Caltech HEP group has been participating in PTP experiments for the last many years.
- Several large scale demonstrations during SuperComputing conferences and Internet2 focused workshops, have proved that software has matured well enough to be part of the production services.
- Dynes and ANSE projects took this initiative and delivered applications and metrics which can create end to end dynamic paths and move TeraBytes of data at high transfer rates.
- Moving forward, within the scope of the SENSE project in collaboration with ESnet, the group aims to provide SDN services across multiple domains to various projects using ESnet SENOS to meet the distributed resource environment.
- An active testbed that connects Caltech, UCSD, Univ of Michigan and StarLight from 10G to 100G bandwidth using a single SDN controller.

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

Fully SDN enabled



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

Azher Mughal

SC15 Demonstration Plans

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)

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)

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.

WAN Demonstration Plans

Remote NSI Sites:

1. Caltech
2. StarLight
3. Univ of Michigan
4. Florida International Univ
5. RNP (Brazil)
6. UNESP (Brazil)

Software used to Provision paths to remote end points:

1. OSCARS
2. OESS
3. NSI
4. OpenFlow

Total of 9 x 100GE links (7 links in Caltech Booth)

OpenFlow Lab Topology (ODL dlux View)

CALTECH HEP NETWORKING

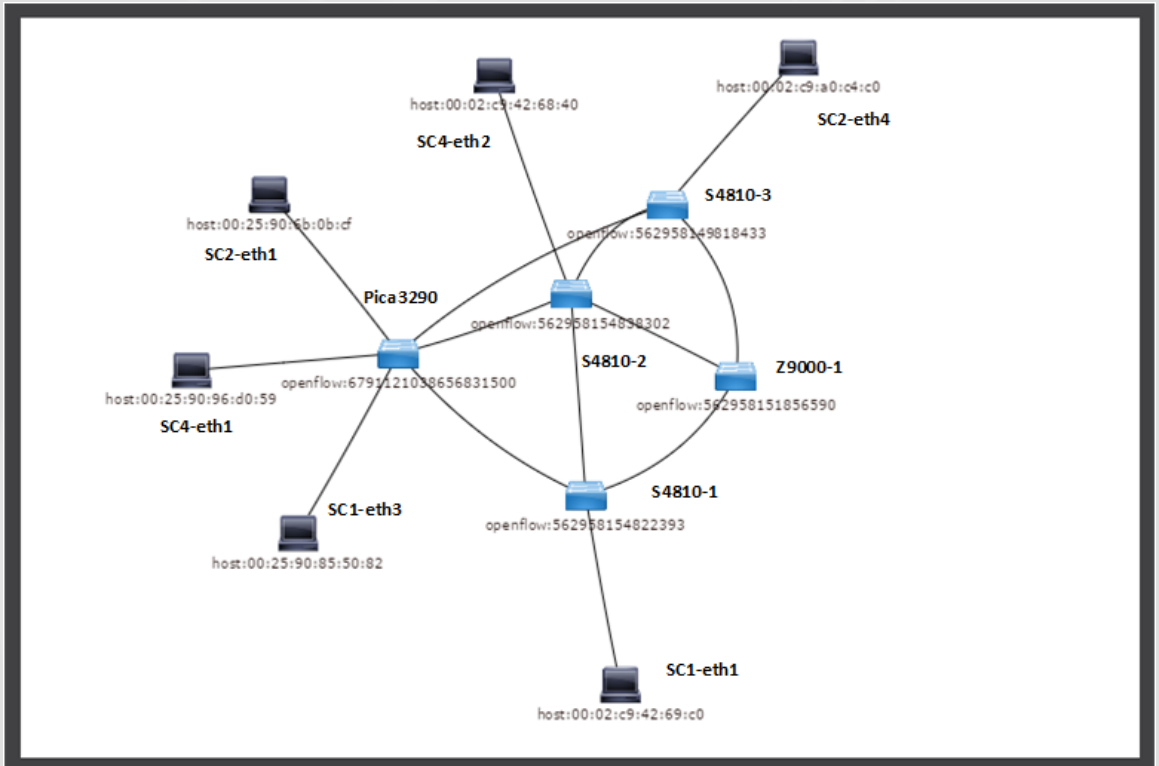
Rich Multi Vendor environment

10/40GE

- Dell s4810
- Dell Z9000
- Dell S4048
- Pica8 3290
- Pica8 3920
- Mellanox 1036/6036

100GE

- Dell Z9100
- Inventec (PicaOS)



Dell:

- OF version 1.0/1.3
- 8 Parallel OF-instances (multi-tenant)

Pica8:

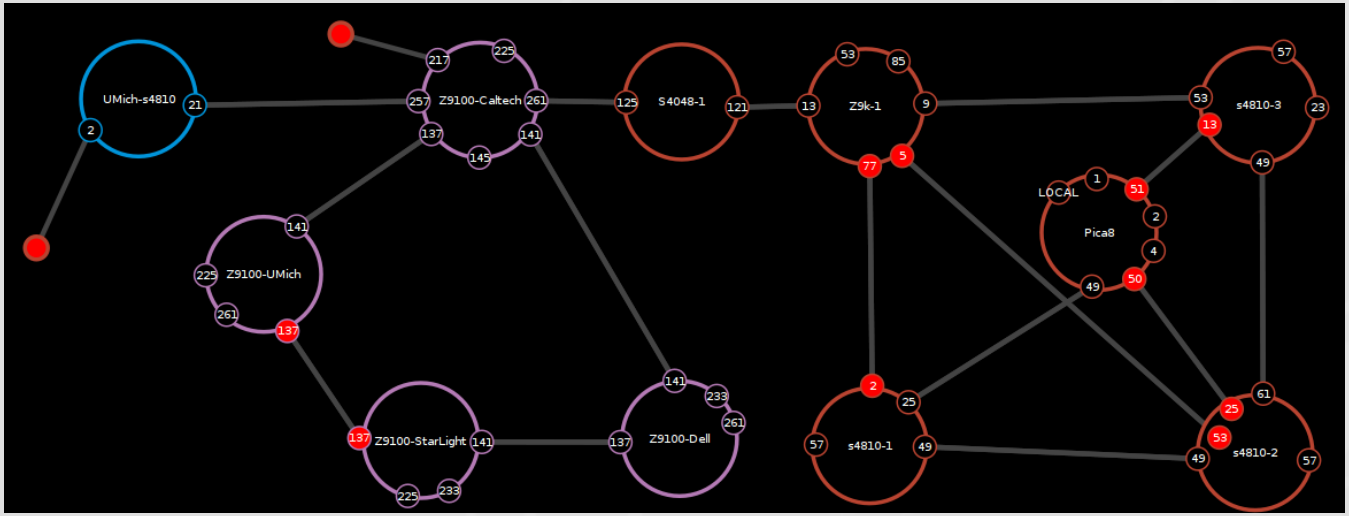
- OF version 1.0/1.3/1.4+
- No limit on OF instances

Mellanox:

- OF 1.0
- Single OF Instance

A versatile Multi-Vendor SDN lab environment for state of the art software development

OpenFlow Lab Topology (OF-NG View)

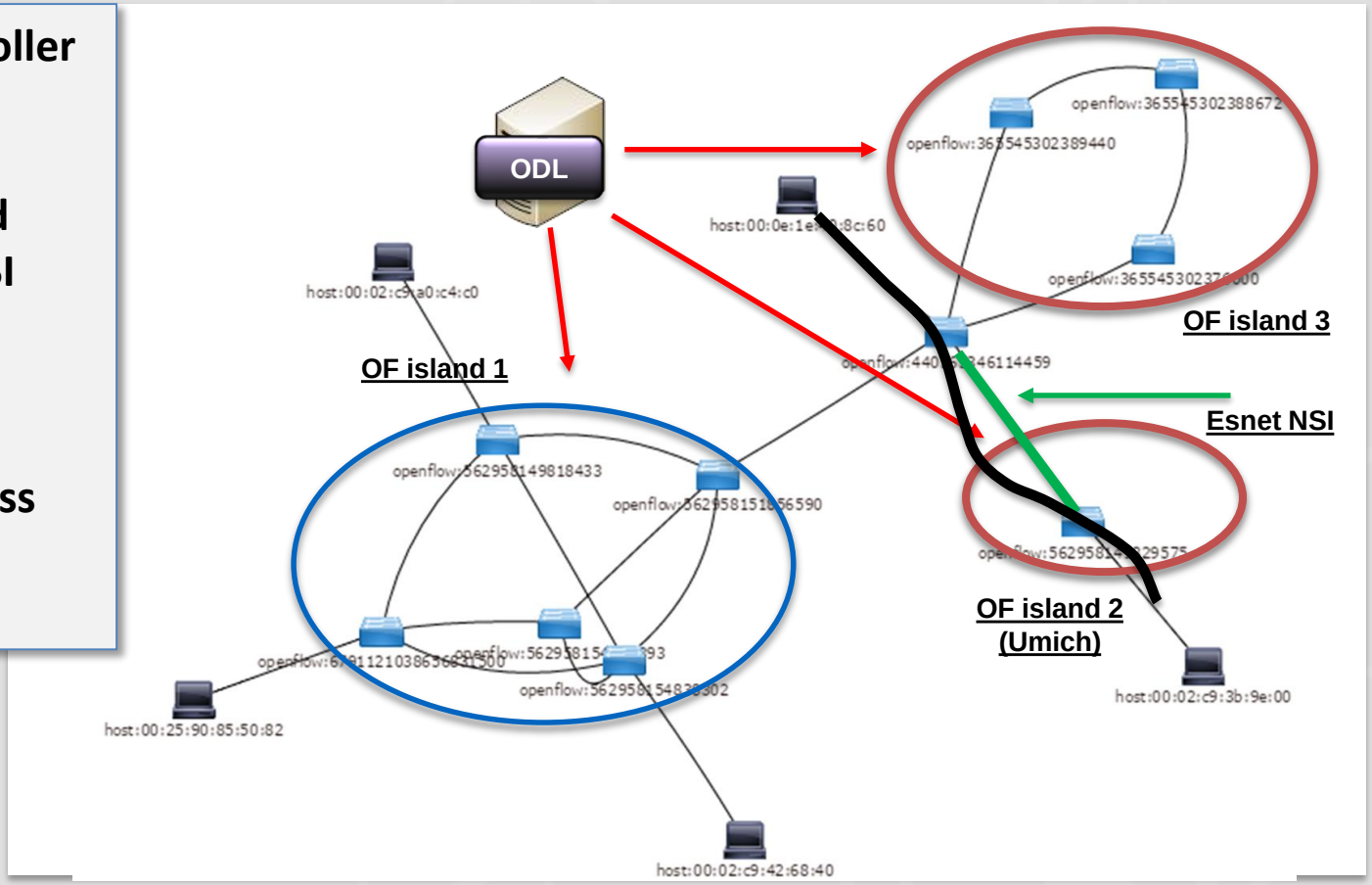


Connectors | Operational Flows | Config Flows

Id	name	port	configuration	hardware_address	status
openflow:440565346114459:217	Hu1/21	217		90:B1:1C:F4:9B:9D	
openflow:440565346114459:145	Hu1/3	145		90:B1:1C:F4:9B:9D	
openflow:440565346114459:225	Hu1/23	225		90:B1:1C:F4:9B:9D	
openflow:440565346114459:141	Hu1/2	141		90:B1:1C:F4:9B:9D	forwarding
openflow:440565346114459:257	Fo1/31/1	257		90:B1:1C:F4:9B:9D	forwarding
openflow:440565346114459:261	Fo1/32/1	261		90:B1:1C:F4:9B:9D	forwarding
openflow:440565346114459:137	Hu1/1	137		90:B1:1C:F4:9B:9D	forwarding

OpenFlow Connecting Remote Islands

- **Single ODL Controller**
- **Global OpenFlow Islands connected using OSCARS/NSI WAN Paths**
- **Layer2/Layer3 configurable across the whole fabric**



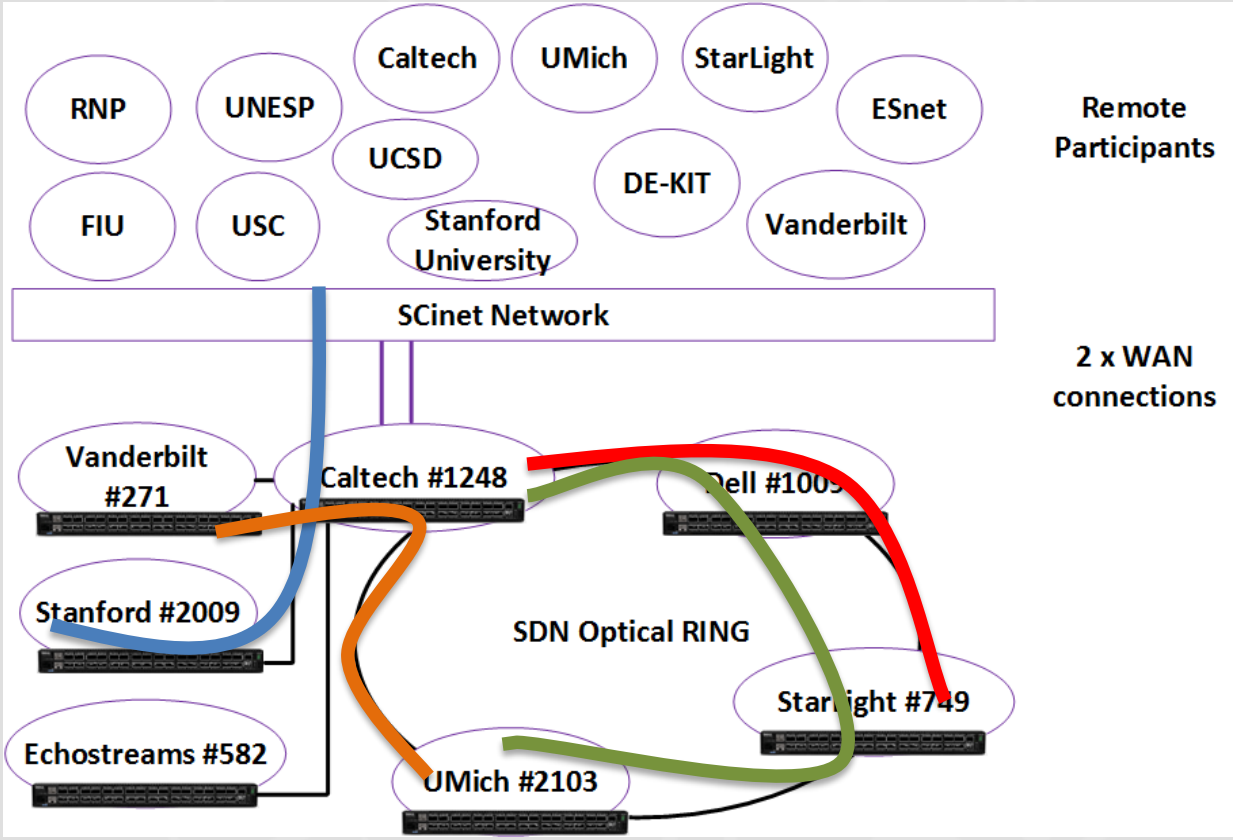
OpenFlow Discovery Protocol (OFDP) using Non-standard LLDP MAC for link discovery across Islands

OpenFlow Traffic Engineering

1. **Install** Flows based on interface statistics or shortest path

2. **Move** Flows dynamically by looking at the interface statistics in the environment and if there is a high bandwidth route available then re-config the flow

3. **Manual** Flow insertion/removal to adapt any requirement



OpenDaylight & Caltech SDN Initiatives

Supports:

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

NetIDE:

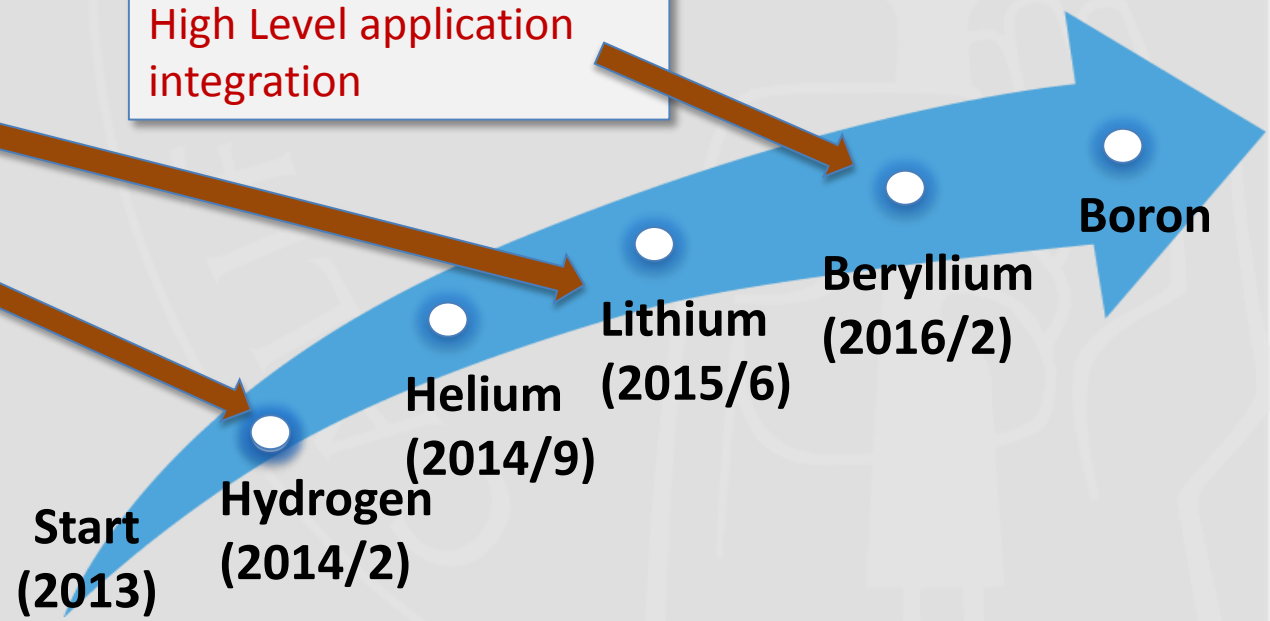
- Rapid application development platform for OpenDaylight and also to convert modules written for other projects to OpenDaylight

OFNG – ODL:
NB libraries for Helium/Lithium

OLiMPs – ODL:
Migrated to Hydrogen

OLiMPs – FloodLight:
Link-layer MultiPath Switching

OFNG - ALTO
High Level application integration



Caltech – UNESP OF-NG Highlights

CALTECH HEP NETWORKING

As part of SC15 conference, we released the OpenFlow Next Generation software (OF-NG), available at GitHub: https://github.com/of-ng

- **Provides a usable, easy to understand and navigate network topology for operators and application developers**
- **Core functionality is written in Python while frontend GUI is extensively using JavaScript for dynamic functionality**
- **Easy drag and drop layout featuring save and load topology**
- **New OpenFlow network elements or hosts appear automatically as they are connected or discovered by the OpenDaylight**
- **OpenFlow Paths:**
  - Route selection and Flow creation/deletion, all paths are available for selection
  - Layer2 and Layer3 path creation at each node or end to end
  - Search and field ordering in the GUI flow tables
  - Plot per flow traffic at any node in the topology
- **Apache mod WSGI interfacing for NB calls. *With this approach, new applications can take advantage of backend library***

OpenDaylight – ALTO Integration

CALTECH HEP NETWORKING

Application Layer Traffic Optimization (ALTO)
Designed by Dr. Richard Yang from YALE University

A high level services approach with intent methodology (paths from A .. Z based on shortest path or bandwidth availability)

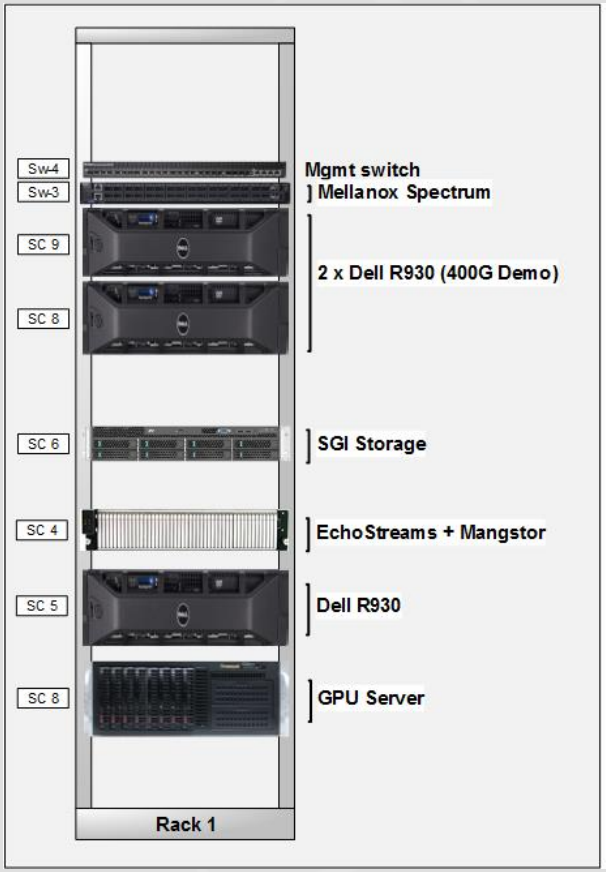
Paths are created using Simple Path Computation Engine (SPCE)

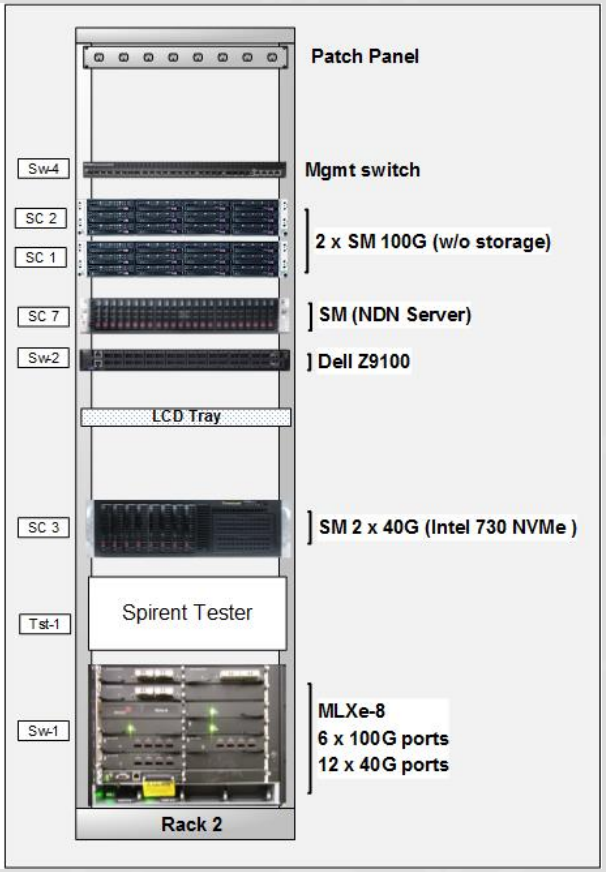
Currently, It offers file transfer scheduler to move large files across different hosts using network and time based metrics

Caltech is working to understand the integration of ALTO with CMS high level services like PhEDEx and ASO.

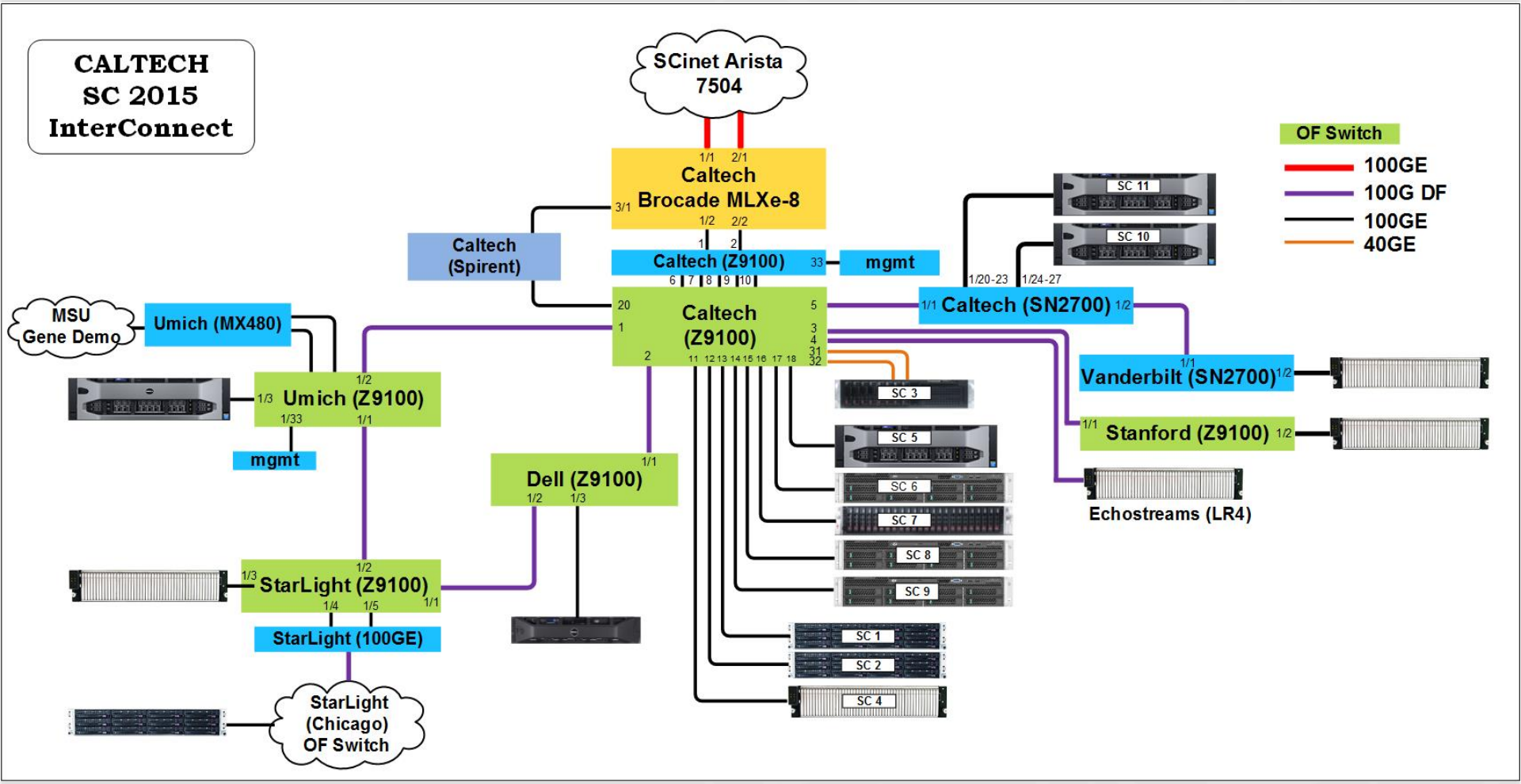
SC15 - Rack Layouts

- **Reduced hardware foot-print with more focus on SDN and 100GE**
- **Disk I/O revolved around NVME**
- **Demonstrate a separate pair of 400GE capable systems**





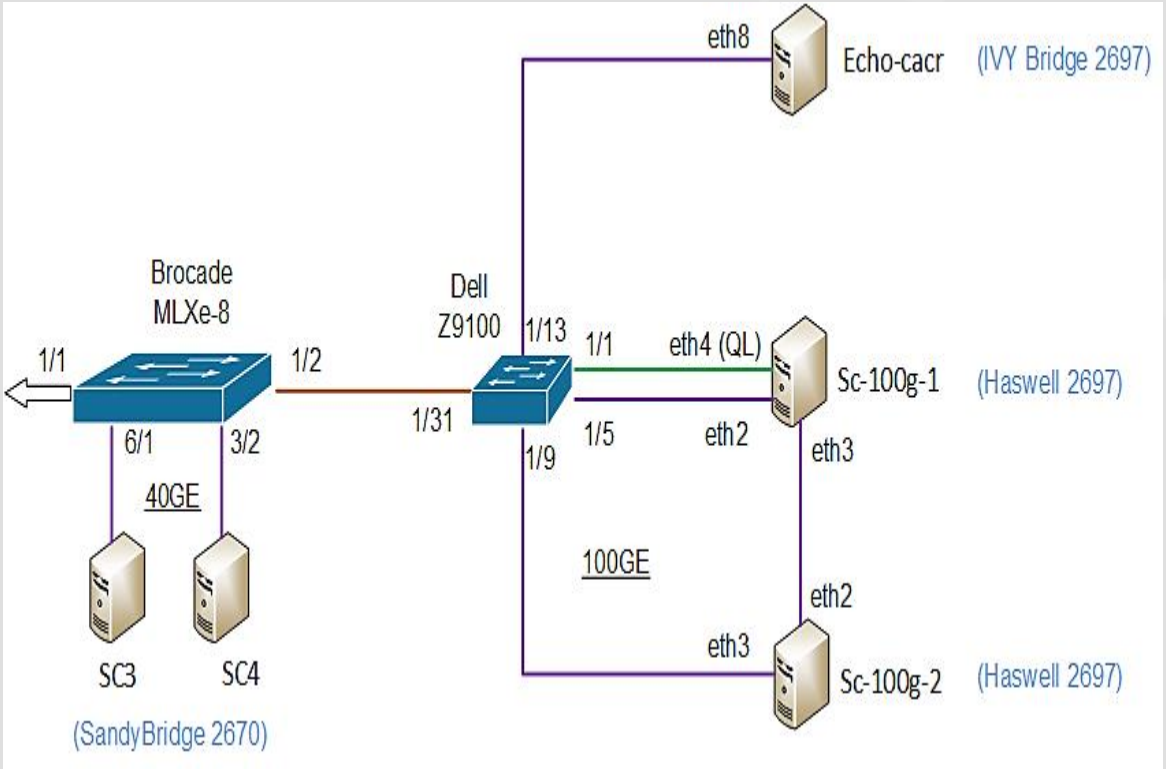
SC15 - Network Layout



Ring of 4 x Dell Z9100 100GE Switches
Above 30, 100GE active ports
Fully SDN controlled using OpenFlow 1.3

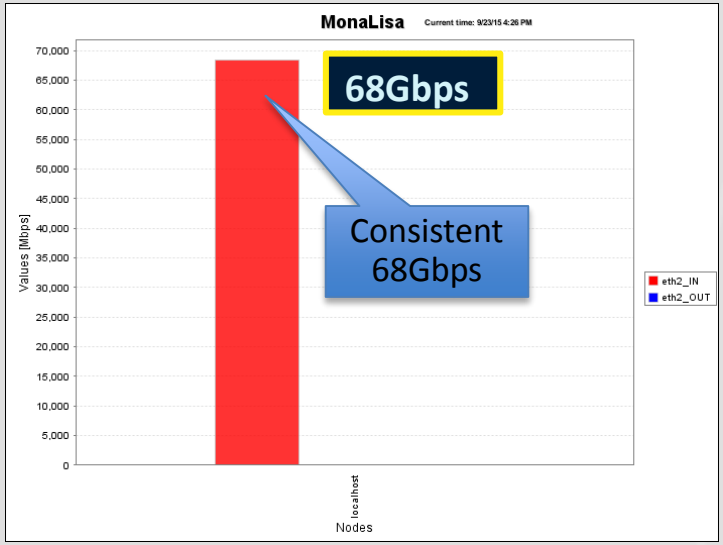
Pre-SC15 Network Performance Tests

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

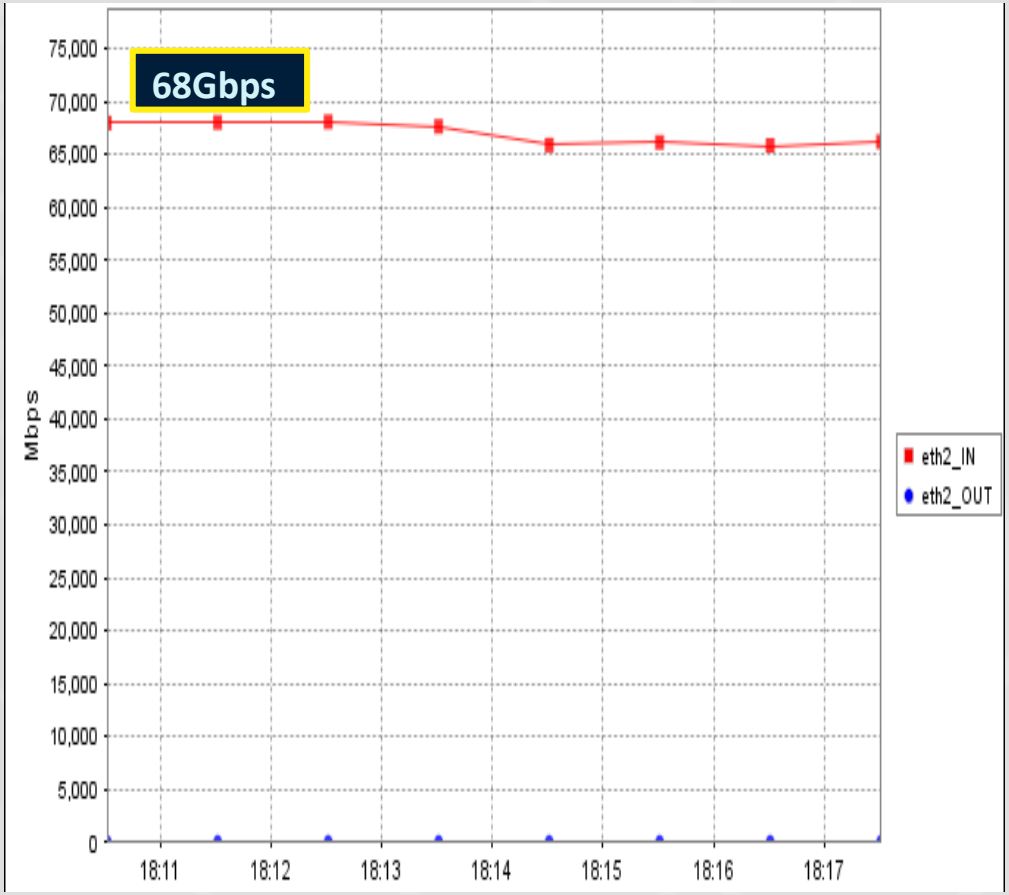


Client

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

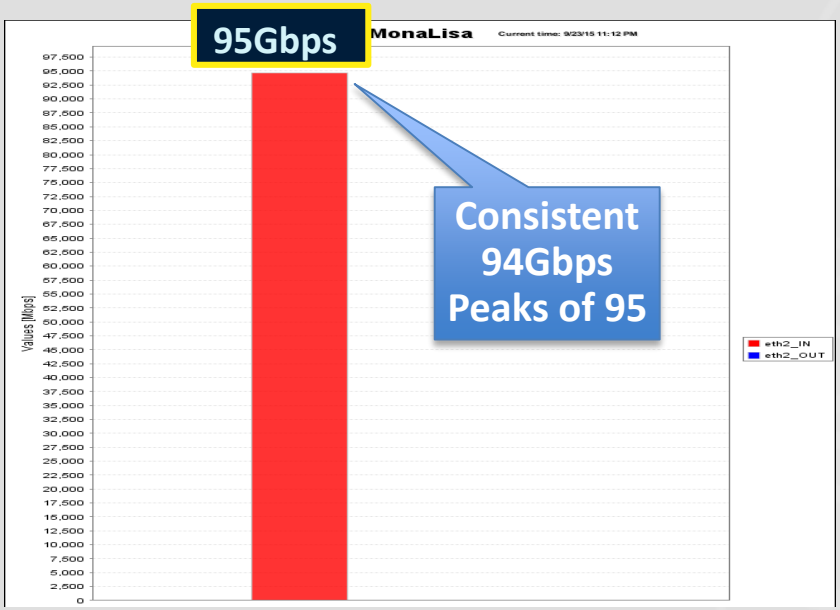
Server

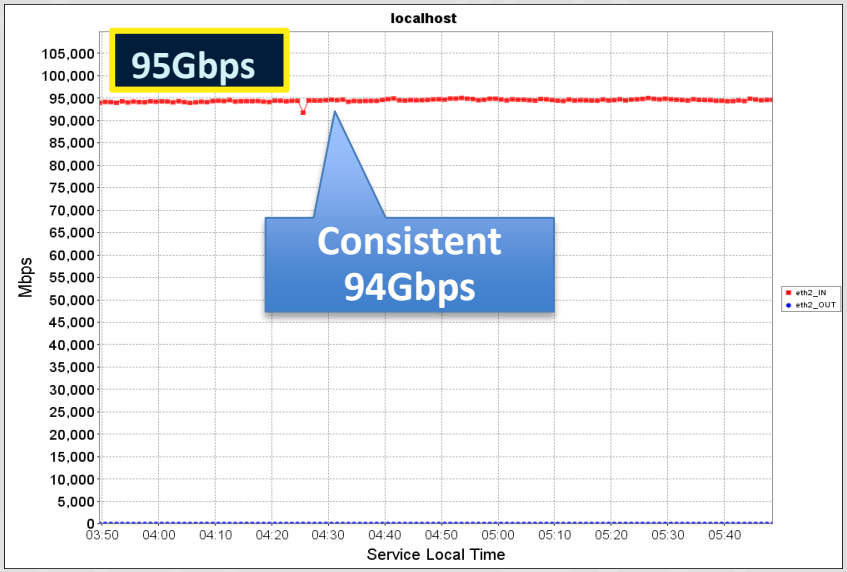
[root@sc100G-2 ~]# numactl --nphyscpubind=20 --localalloc java -jar fdt.jar -p 7000



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*

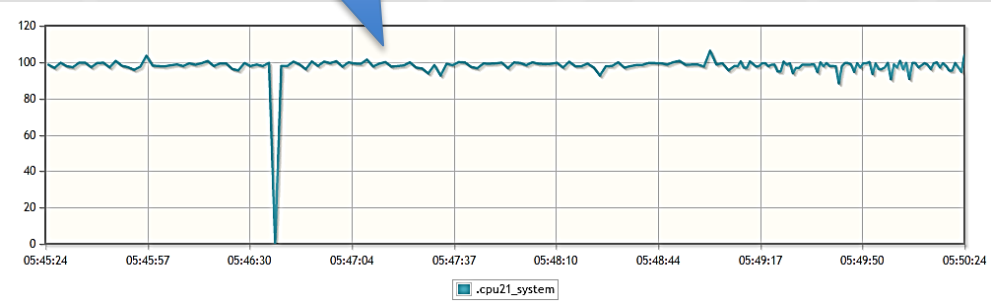
Two TCP Streams in Parallel



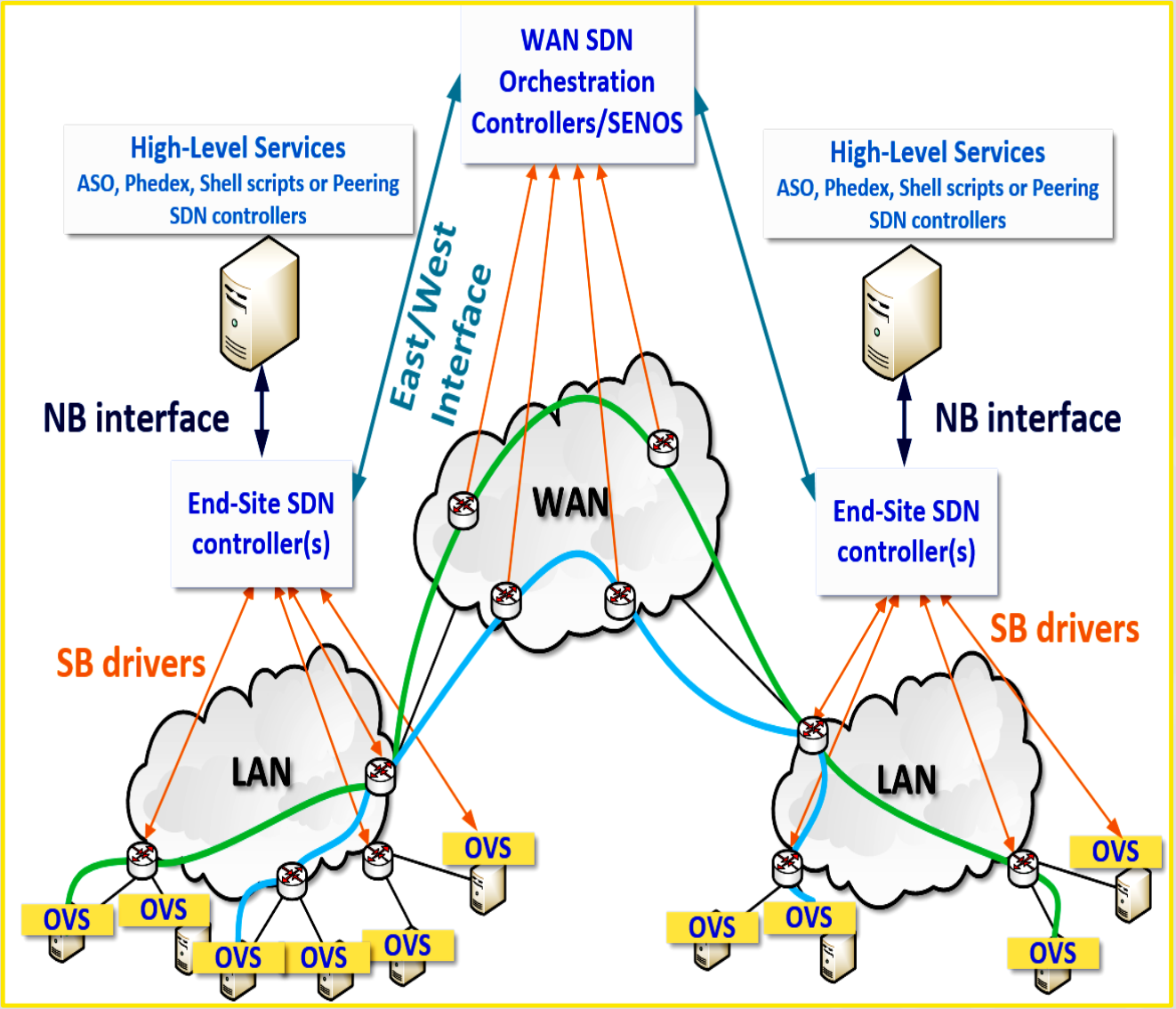


24/09 04:53:29	*Net Out: 94.629 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:34	*Net Out: 94.660 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:39	*Net Out: 94.249 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:44	*Net Out: 94.652 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:49	*Net Out: 94.521 Gb/s*	*Avg: 93.743 Gb/s*
24/09 04:53:54	*Net 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)



End-Site Orchestration



Distributed SDN Controllers

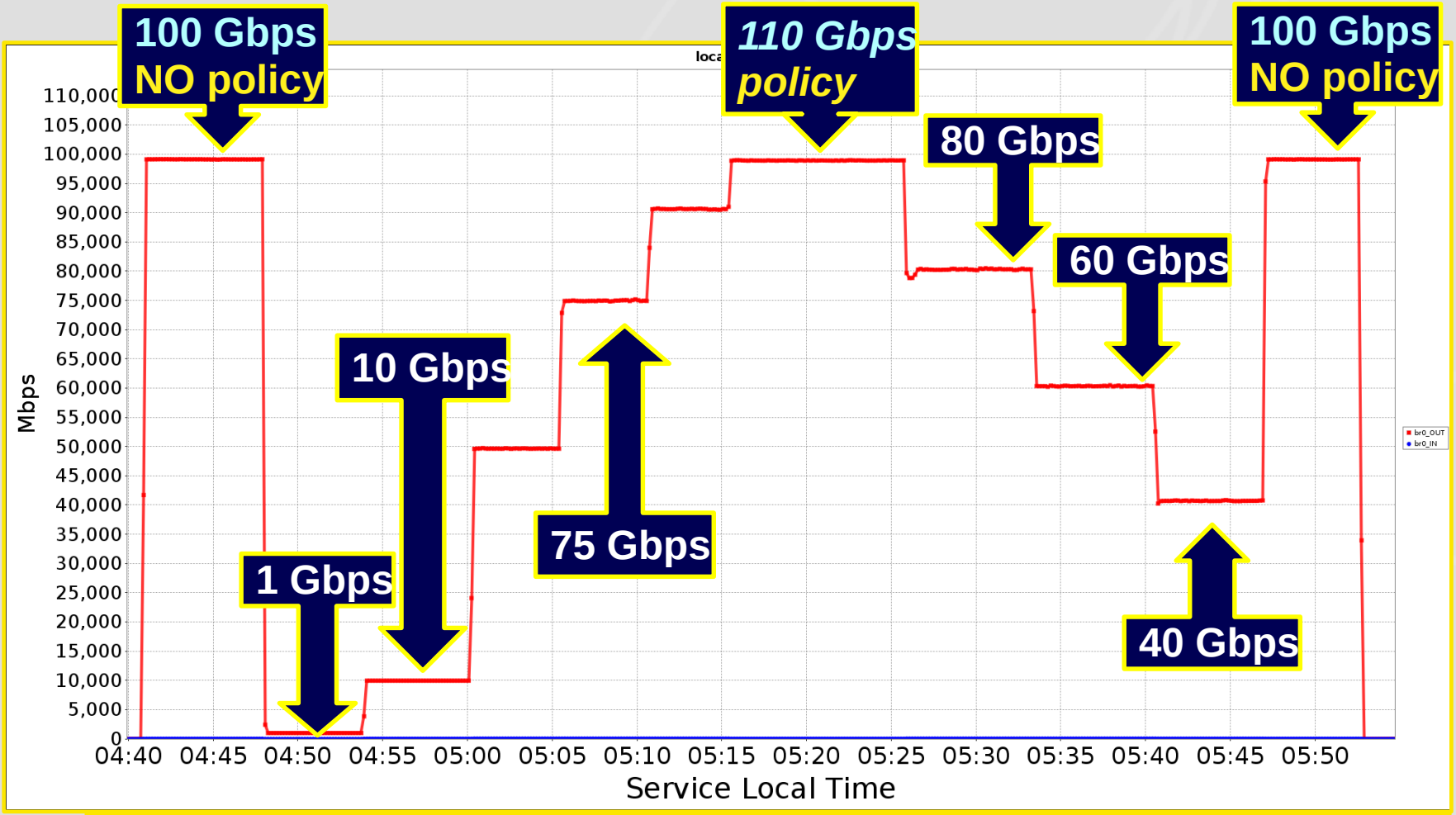
High level services from experiments generating the demand

OVS on end hosts to prioritize or rate-limit the traffic across any pair of traffic flow

Individual site to advertise its resources

Inter-Site Example Multiple Host Groups, Paths, Policies

OVS - Egress Dynamic bandwidth adjustment

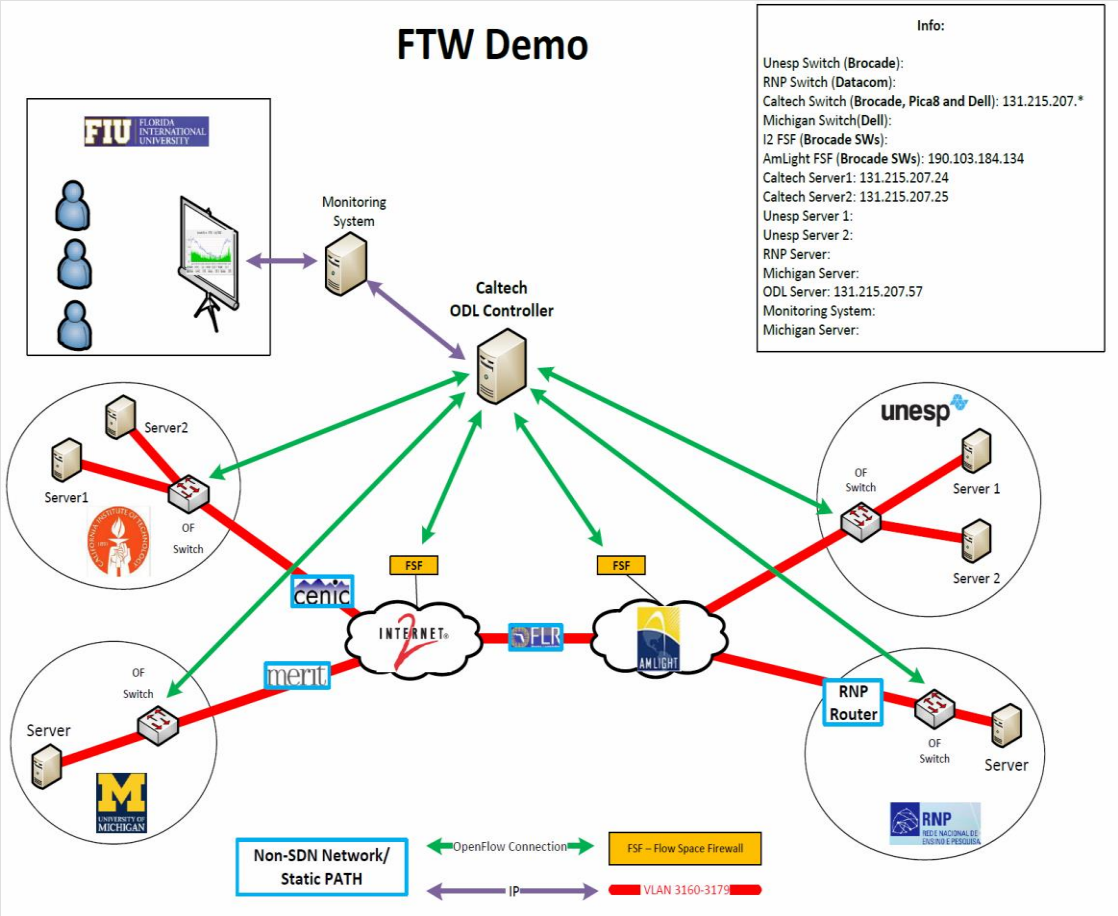


Smooth Stable Flows at any rate up to line rate.
Low CPU penalty on policy usage.

Azher Mughal

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

SDN Demonstration for the FTW Workshop

CALTECH HEP NETWORKING

Dynamic Path creation:

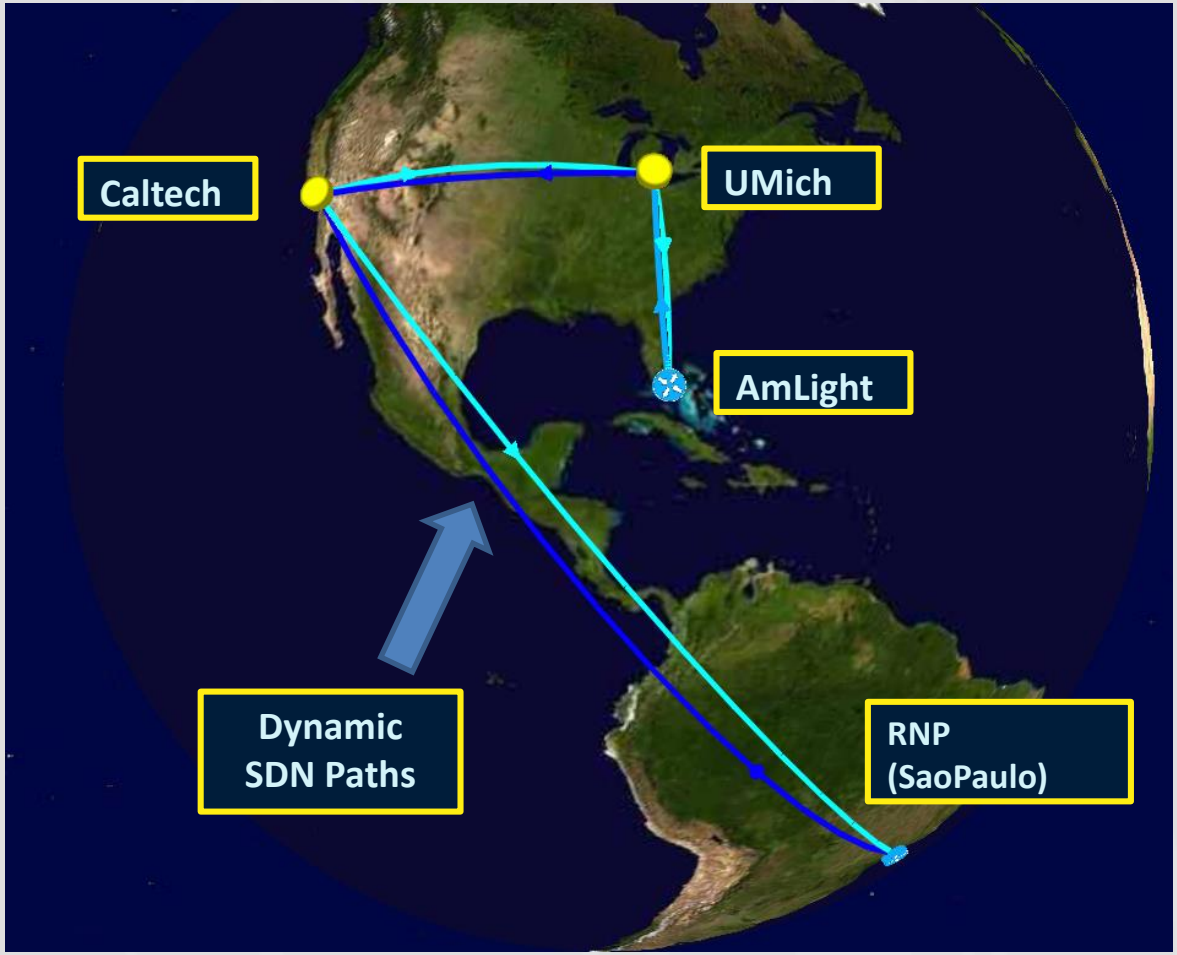
Caltech – UMich

Caltech – RNP

Umich – AmLight

Path initiation by the DYNES FDTAgent using OSCARS API Calls, OESS for OpenFlow data plane provisioning

MonALISA agents at the end-sites providing detailed monitoring information



Successful Multi-domain Dynamic Paths using OSCARS, NSI with SDN as south bound.

Moving forward to integrate with multi-domain Software Defined Exchanges (SDX)

High Performance 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

File System and Drive aggregation

Popular File System Options:

- XFS
- BTRFS
- ceph

Device Aggregation (OS):

- ZFS on Linux
- Software RAID (mdadm)
- Software RAID with Intel additions (imsm)

Device Aggregation (HW RAID):

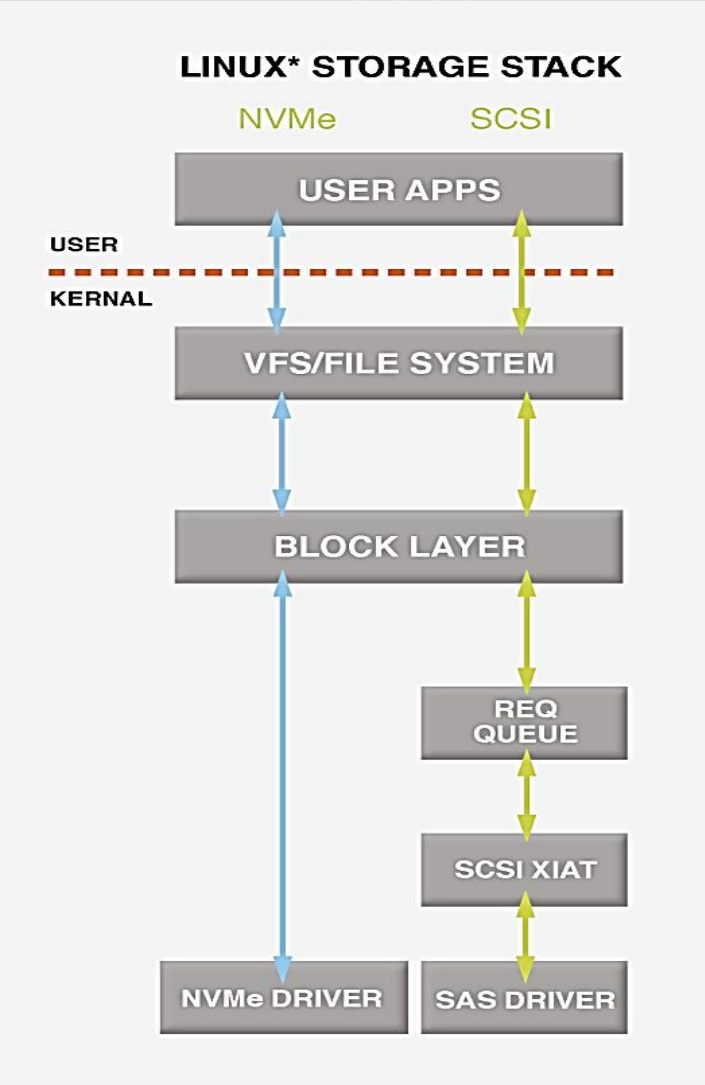
- Hardware RAID Controllers (LSI/PMC)

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 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)



NVMe Drive Options

Pros:

Many times faster than standard SSD drives

Cons:

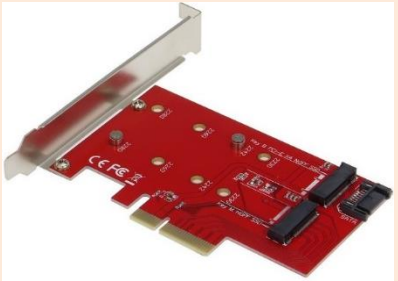
Expensive

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)

2CRSI / SuperMicro 2U - NVME Servers

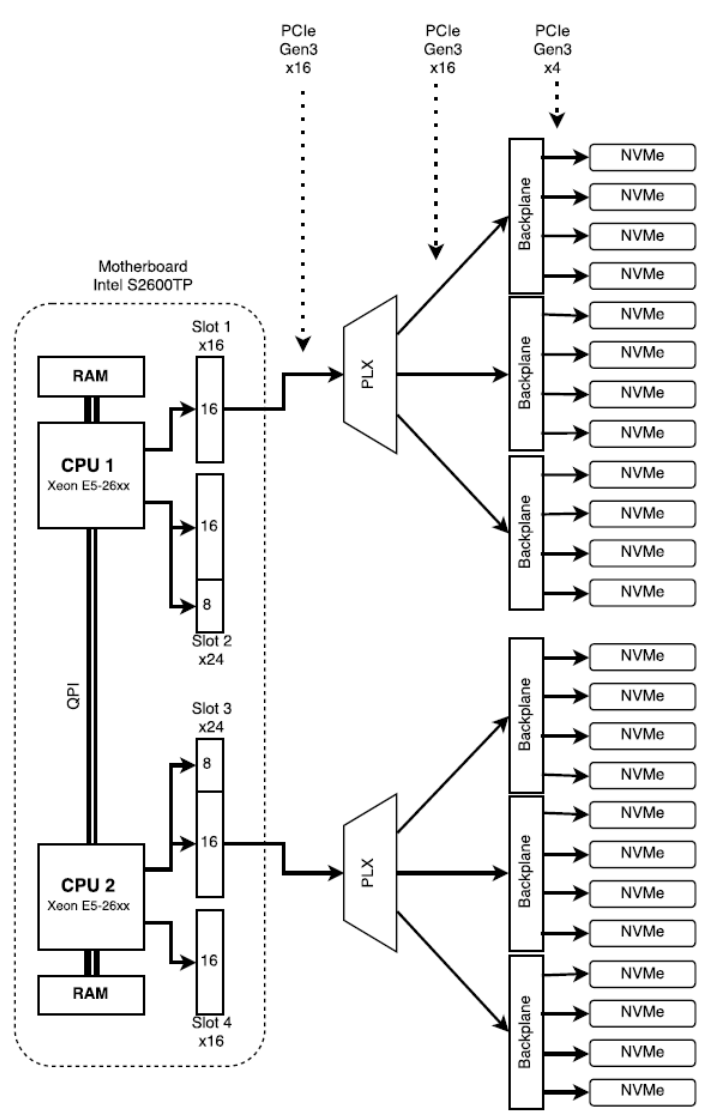
- Both servers are capable to drive 24 NVME drives. SuperMicro also have a 48 drive version.
- Standard 2.5” NVME drive option

2CRSI

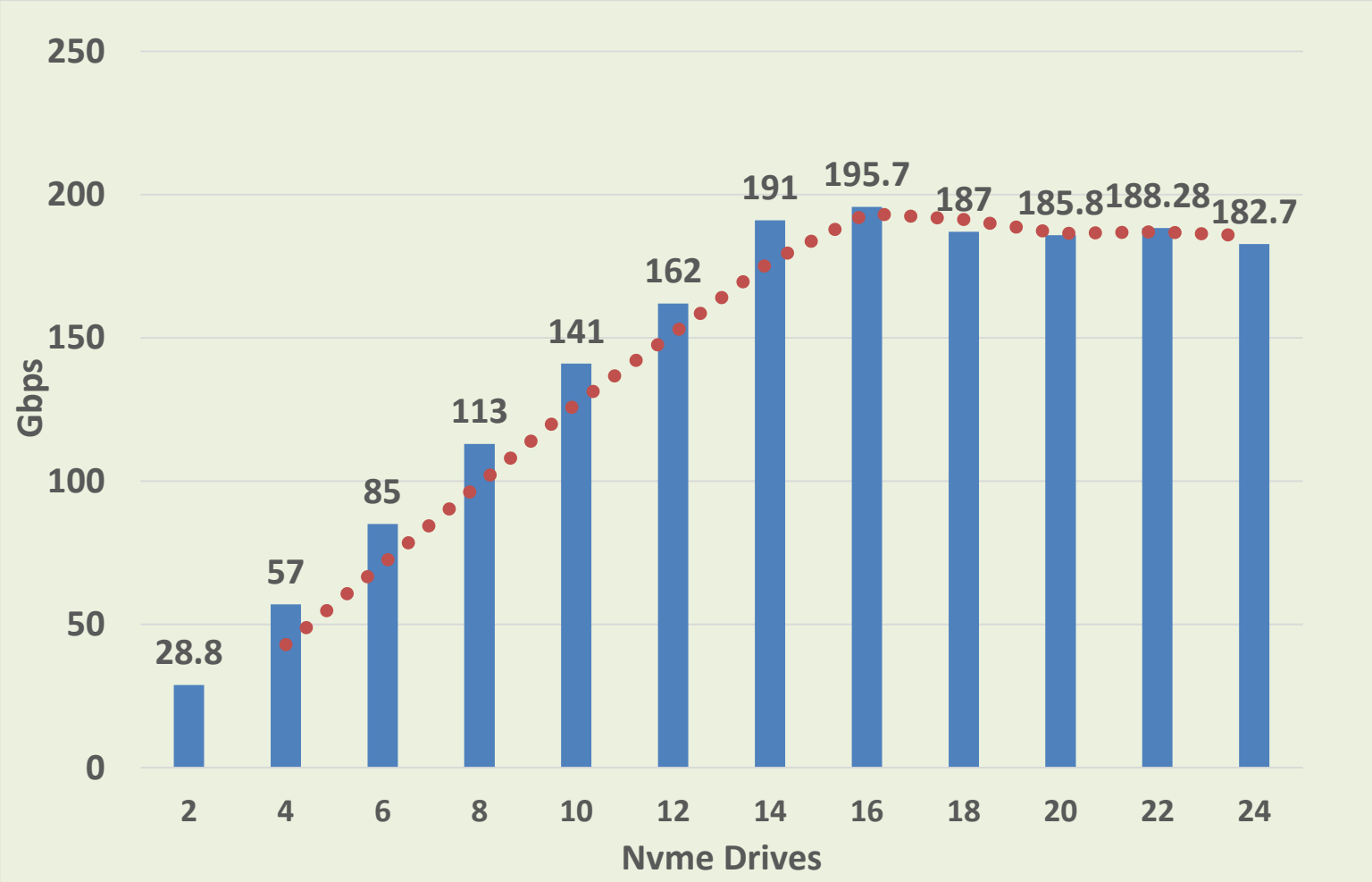
SuperMicro

2.5” NVME Drive

PCIe Switching Chipset for NVME



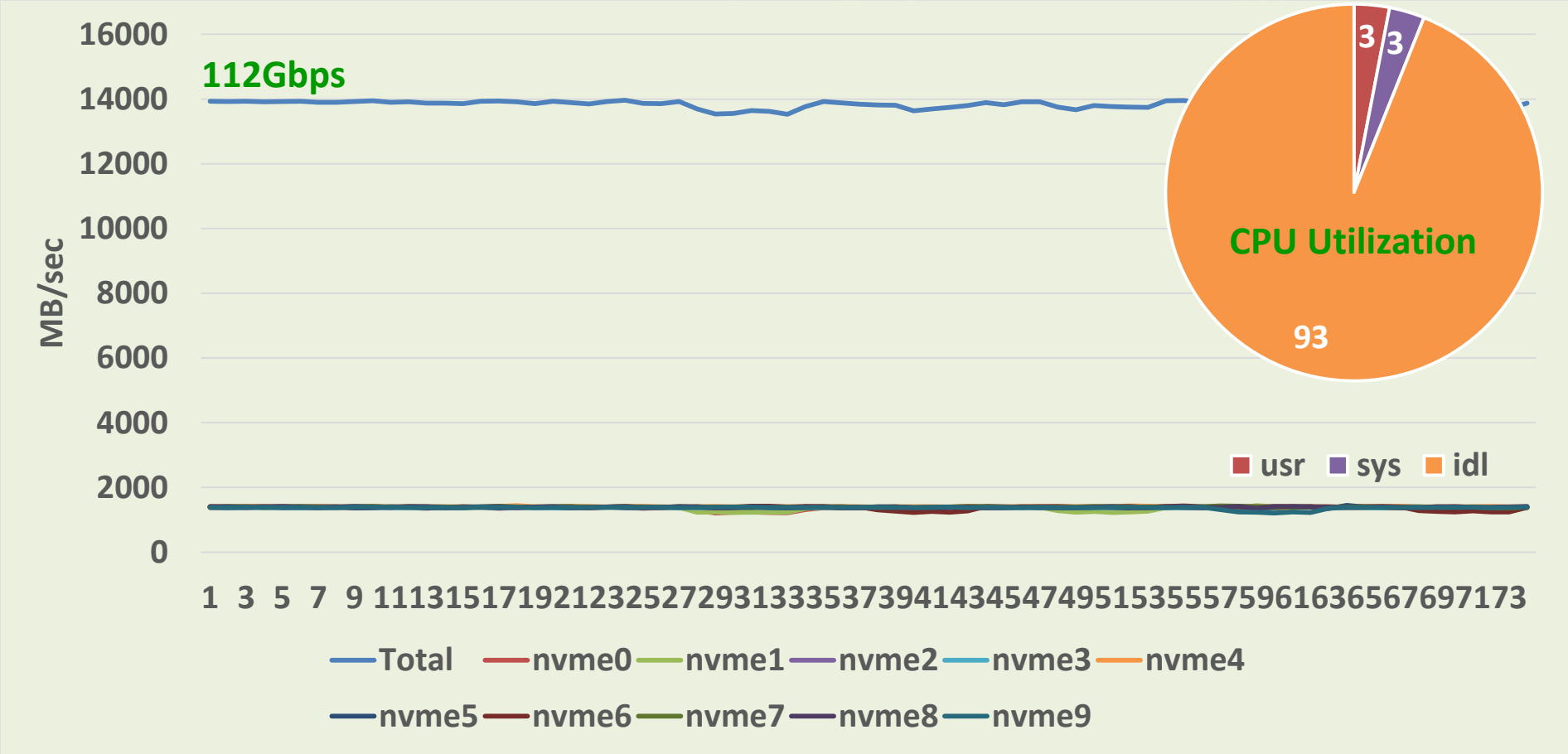
2CRSI Server with 24 NVMe drives



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

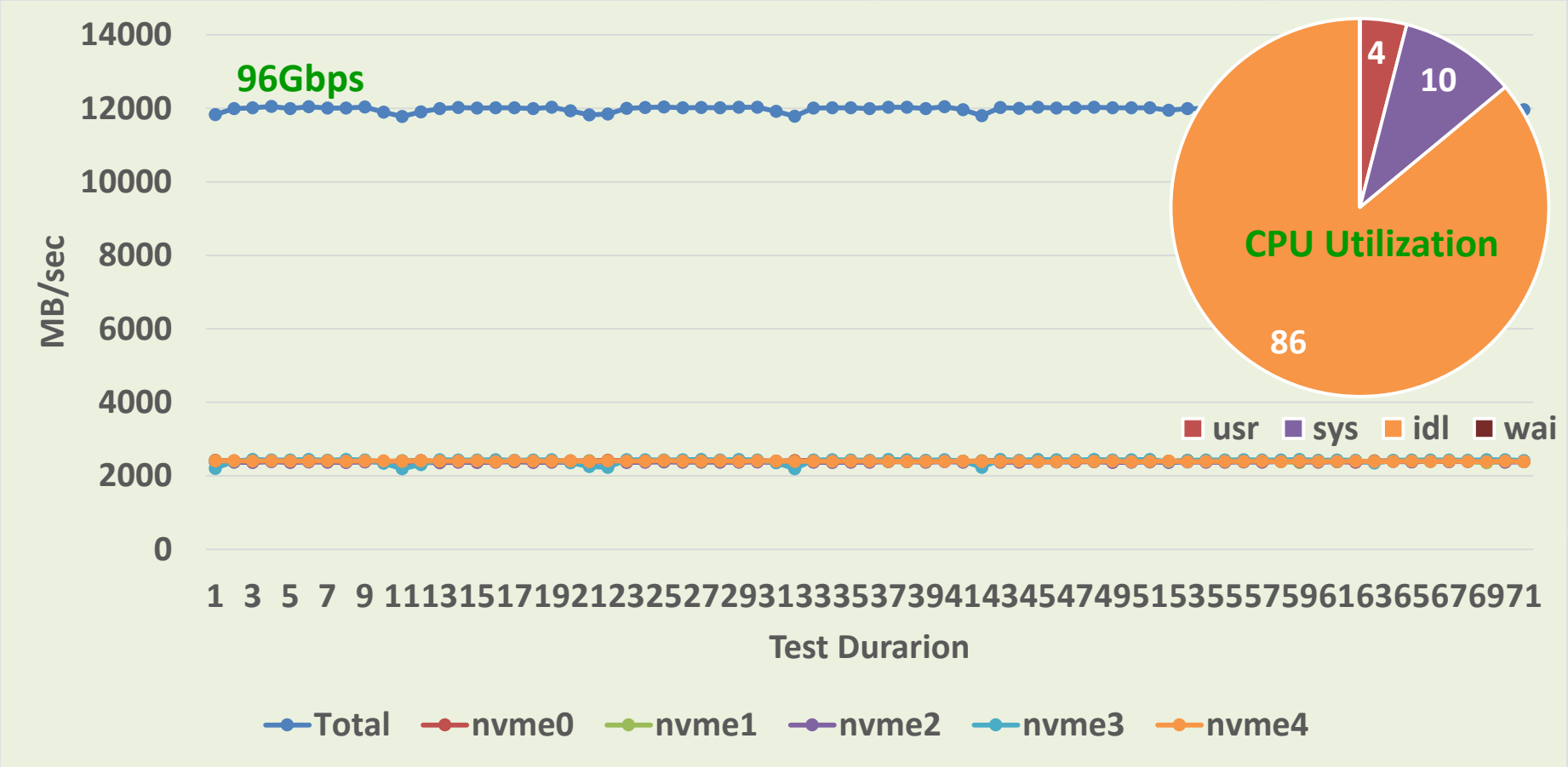
A limitation due to combination of single PCIe x16 bus (128Gbps) and processor utilization.

Intel NVME 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

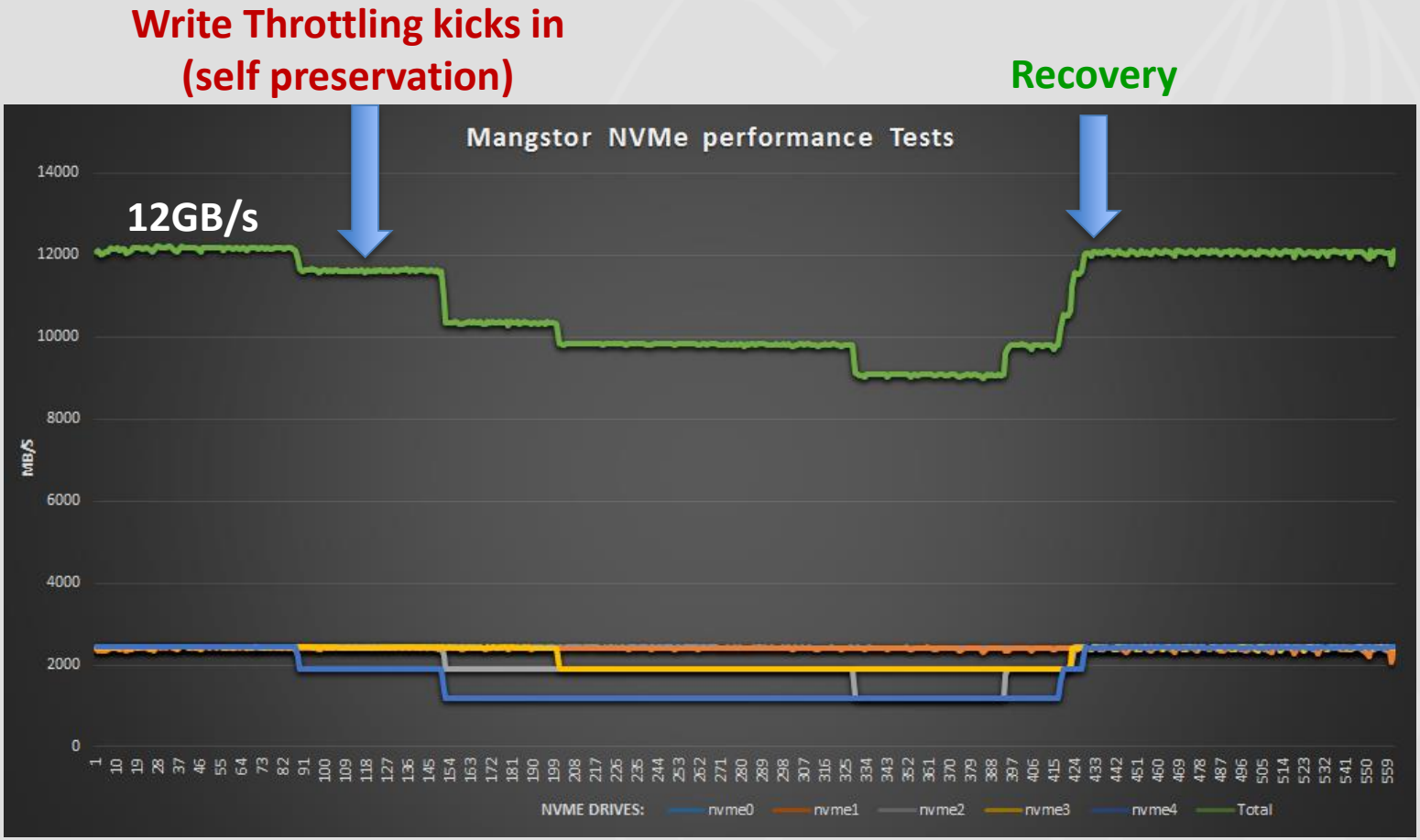
Mangstor NVME Disk Benchmarks



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 minimum airflow requirement.

100GE Switches (compact form factor)

- 32 x 100GE Ports
- All ports are configurable
  - 10 / 25 / 40 / 50 / 100GE
- Dell and Inventec are based on Broadcom Tomahawk chip, While Mellanox is using their own spectrum chipset
  - Common 16MB packet buffer memory
  - 3.2 Tbps Full Duplex switching capacity
- All three switches support ONIE Boot loader
  - Dell = FTOS
  - Inventec = PicaOS
  - Mellanox = Mellanox or Cumulus Linux
- Dell Z9100 supports additional 2 x 10GE SFP+ ports
- OpenFlow 1.3+, with multi-tenant support

Next Wave of switches using StrateDNX (Jericho) should be appearing in 1H 2016.
Promising deep buffers and large routing tables

100G Switches, Optics, Cables (confusions …)

- Only two vendors are shipping LR4 QSFP28 Optics (InnoLight and Source Photonics)
- Source Photonics iLR4/LR4 did not work with juniper (juniper either locks ports or not enough power)
- InnoLight specs says 4W power consumption per QSFP28 port
- Mellanox VPI NICs needs to be converted to *Ethernet*
- QLogic (rev A0) does not support link Auto Negotiation, neither FEC
  - Cannot work with Brocade MLXe Routers
- Mellanox supports FEC but Auto Negotiation takes long time (at times switch ports do not come up)
- Buffers not deep enough to take several bursts at 100Gbps

System Design Considerations for 200GE / 400GE and beyond

Beyond 100GE … 400GE ?

Server Readiness:

1) Current PCIe Bus limitations

- PCIe Gen 3.0 (x16 can reach 128Gbs Full Duplex)

  Still processor has to do PCIe bus arbitration (adding latency)

- PCIe Gen 4.0 (x16 can reach double the capacity, i.e. 256Gbps

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

- rive single stream

4) Increased memory controllers and higher clock rate reaching 3000MHz

Collaboration Partners

CALTECH HEP NETWORKING

Special thanks to …

Research Partners

- Univ of Michigan
- iCAIR / NITRD
- UNESP
- RNP
- Internet2
- ESnet
- CENIC
- FLR
- PacWave

Industry Partners

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

Thank you !

Questions ?