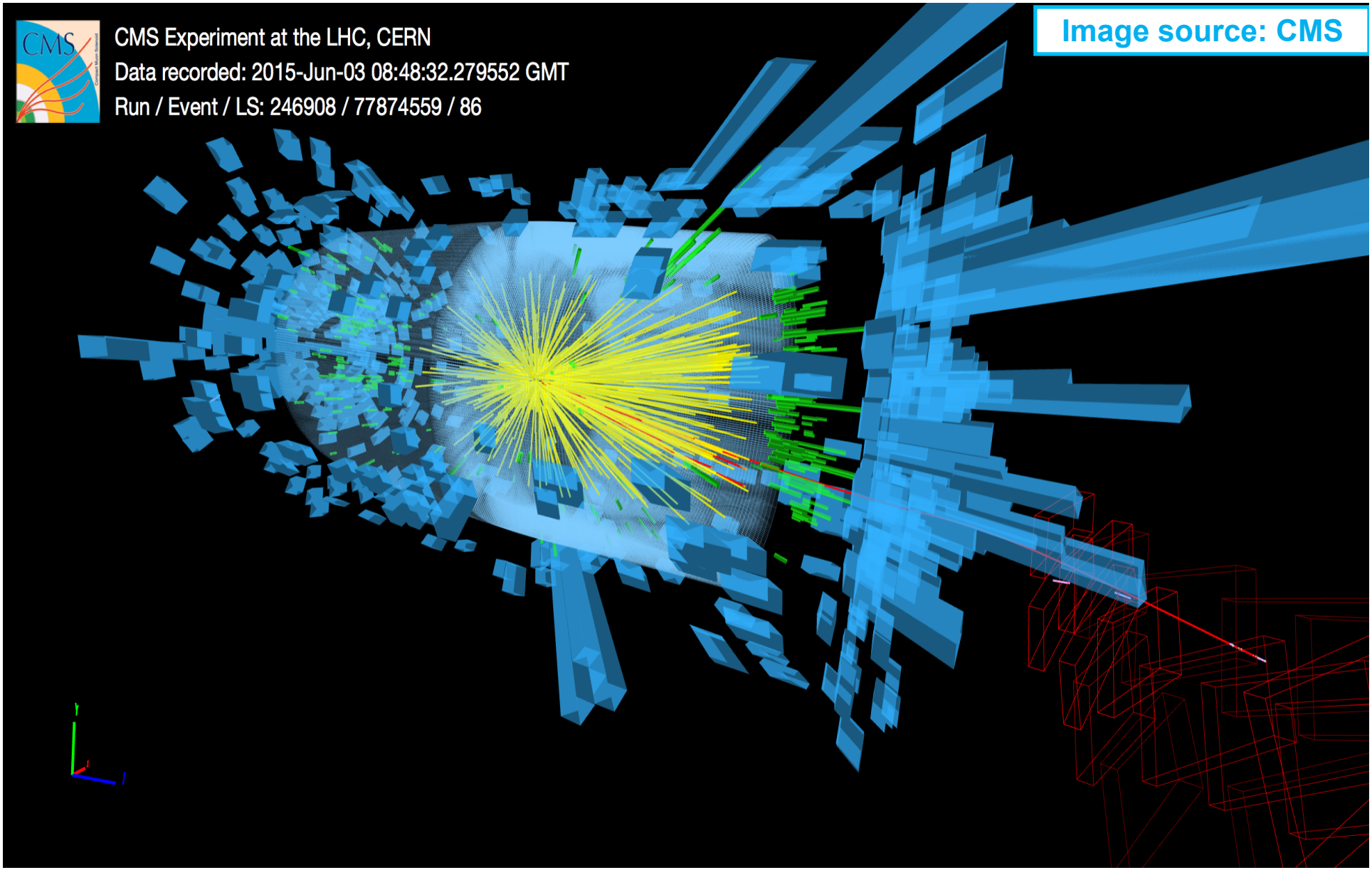
NDN Testbed

Harvey Newman, Azher Mughal, Ramiro Voicu

California Institute of Technology

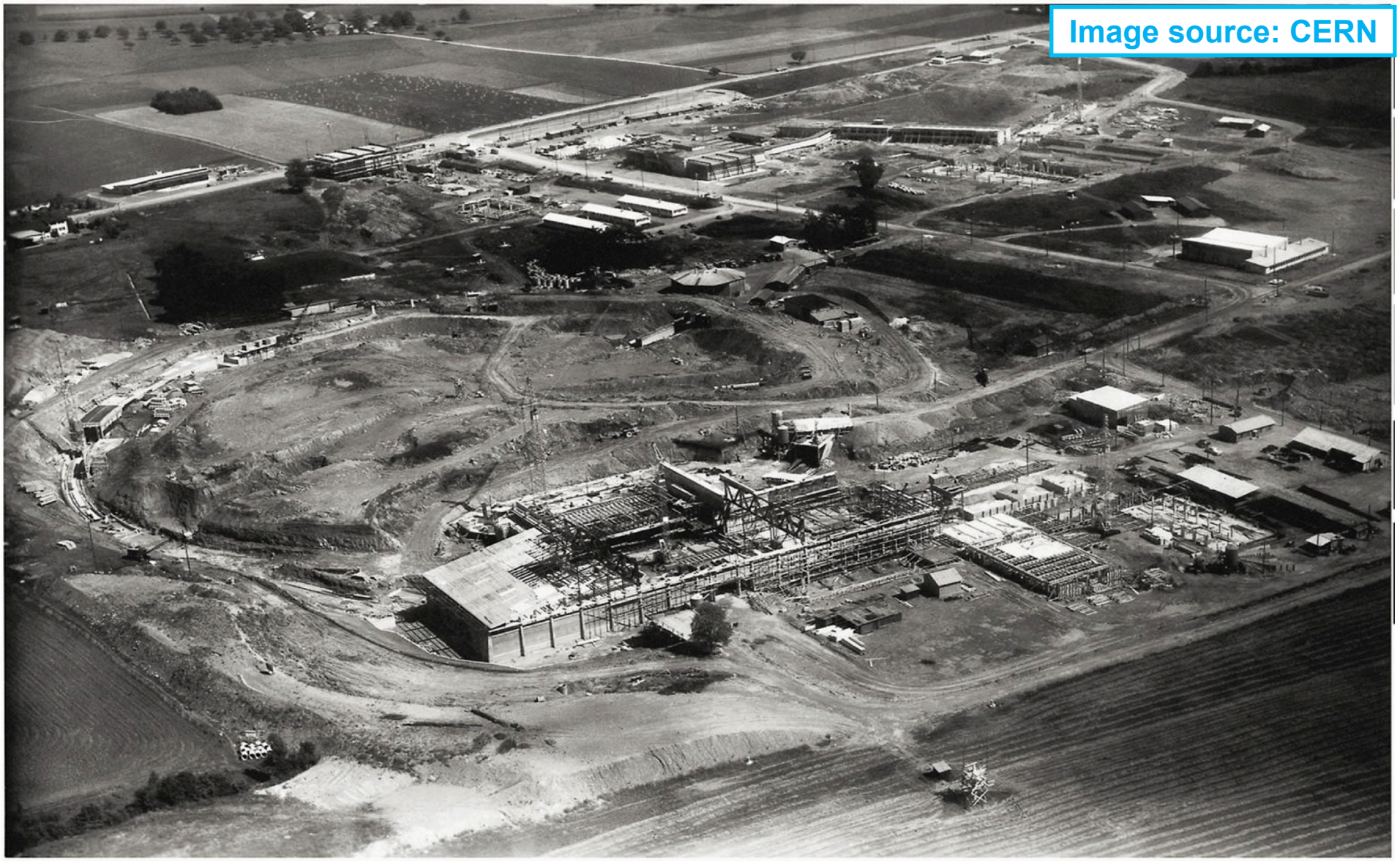
15th Annual Global LambdaGrid Workshop
Prague, Czech Republic, September 28-30, 2015

LHC Collisions at 13 TeV



61 Years ago… CERN was born

29 Sep 1954 – 29 Sep 2015

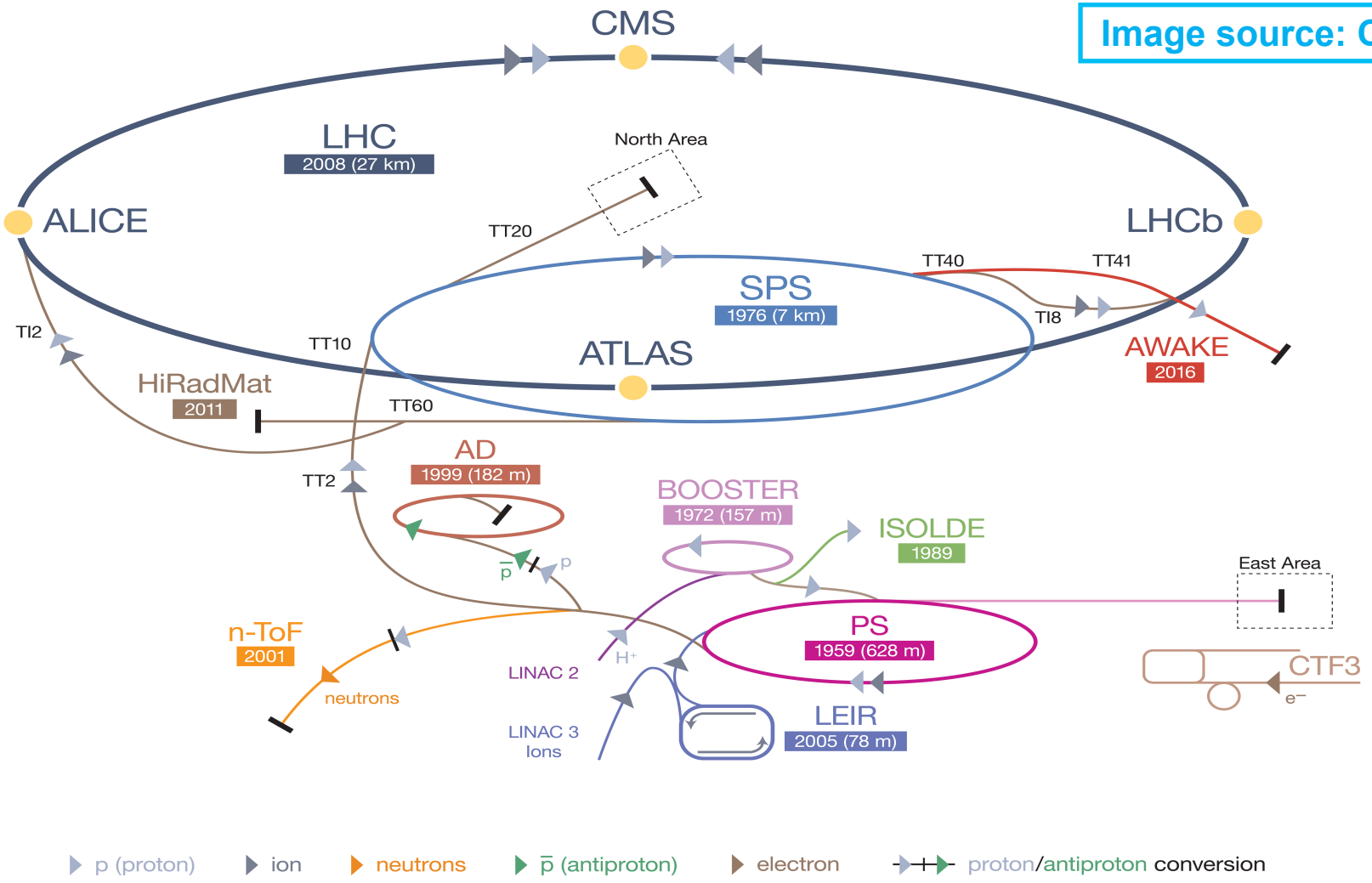


CERN LHC: Aerial view

Image source: CERN

CERN Accelerators

Image source: CERN

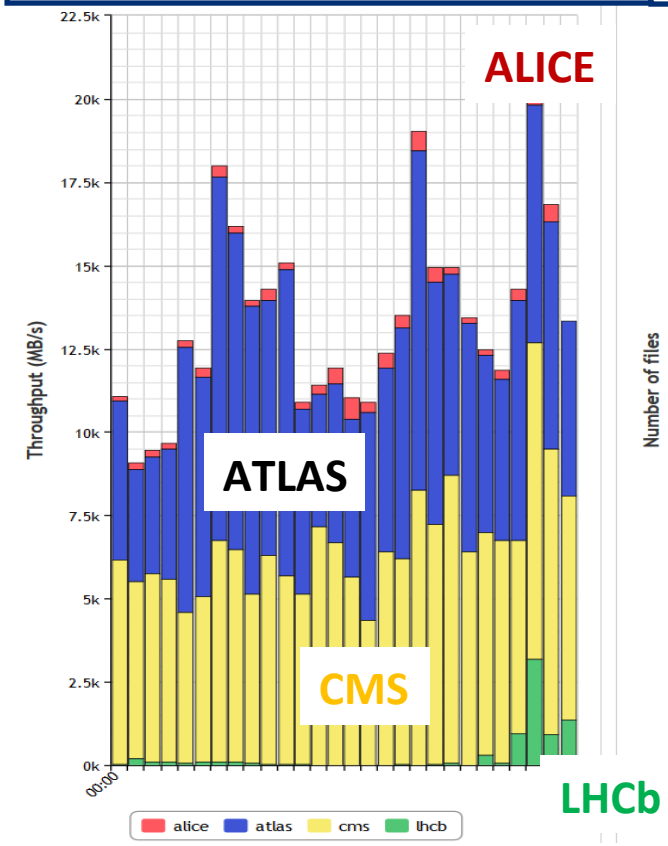


LHC Large Hadron Collider SPS Super Proton Synchrotron PS Proton Synchrotron

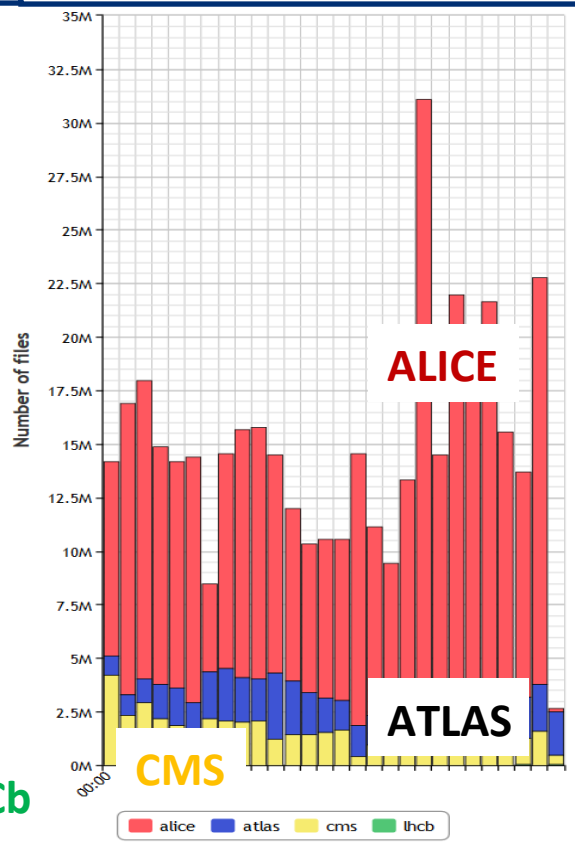
Complex Workflow: the Flow Patterns Have Increased in Scale and Complexity, even at the start of LHC Run2

WLCG: 170 Centers in 40 Countries. 2 Million Jobs Per Day

Transfer Throughput



Transfers Done/Day



13 GBytes/s Typical

25 GBytes/s Peak Transfer Rates

Complex Workflow

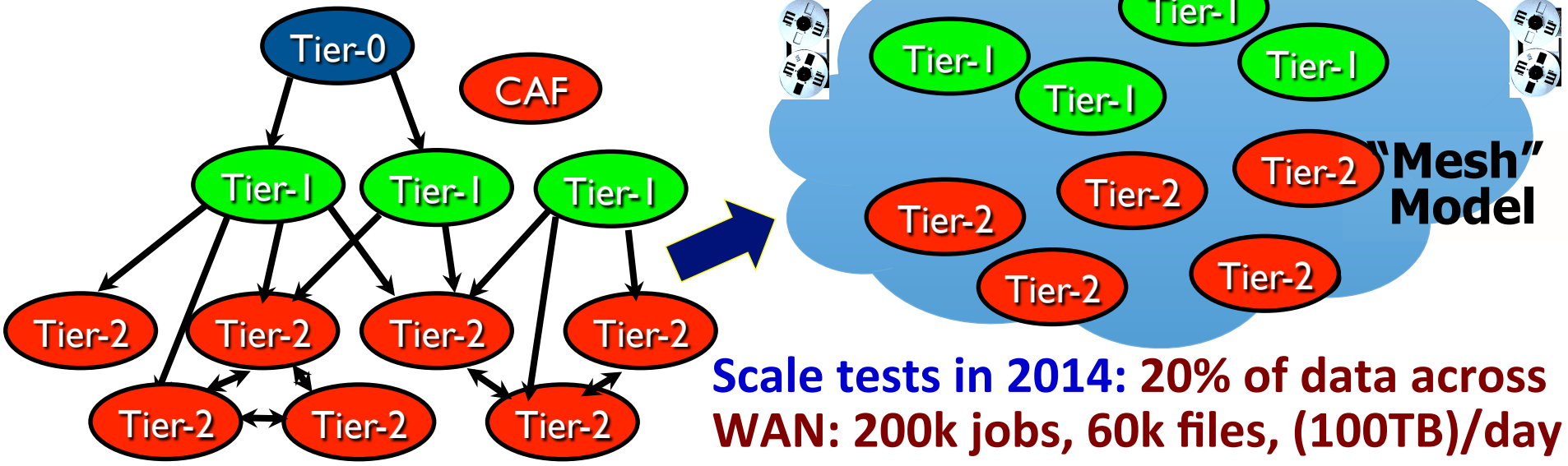
- Multi-TByte Dataset Transfers
- Transfers of Millions of Output Files Daily
- Access to Tens of Millions of Object Collections/Day
- >100k of remote connections (e.g. AAA) simultaneously

WLCG Dashboard Snapshot Aug-Sep 2015

Location Independent Access: Blurring the Boundaries Among Sites + Analysis vs Computing

Maria Girone
CMS Computing

- Once the archival functions are separated from the Tier-1 sites, the functional difference between Tier-1 and Tier-2 sites becomes small [and the analysis/computing-ops boundary blurs]
- Connections and functions of sites are defined by their capability, including the network!

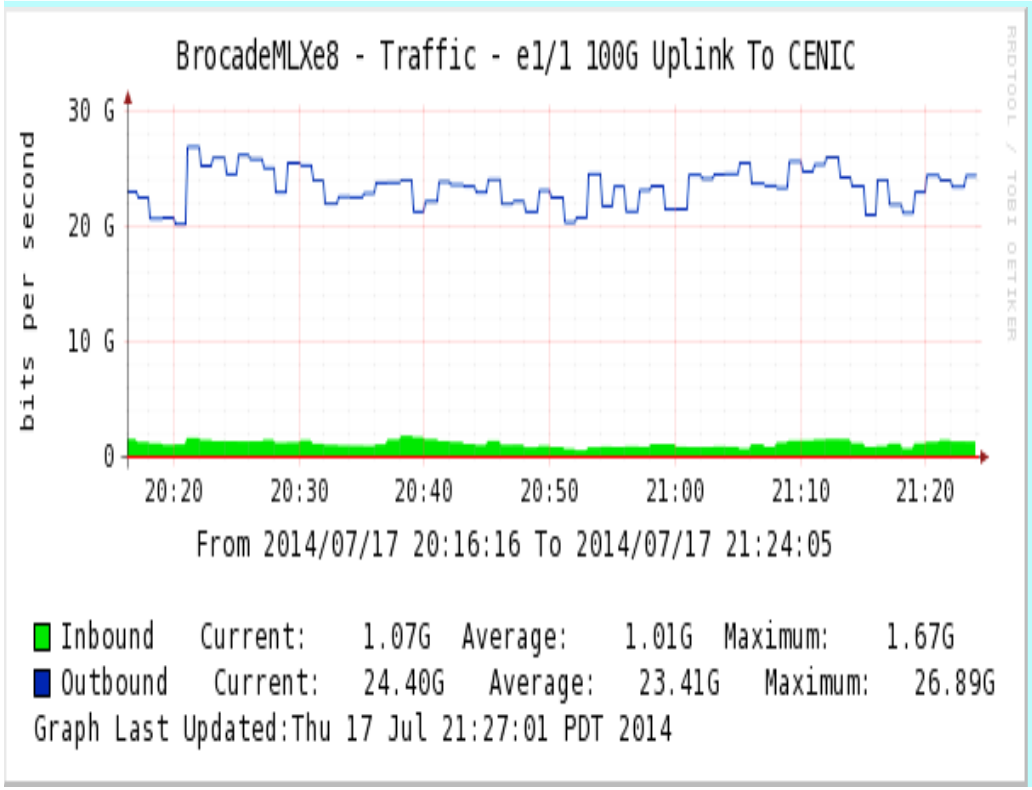


Scale tests in 2014: 20% of data across WAN: 200k jobs, 60k files, (100TB)/day

+ Elastic access of from some Tier2/Tier3 sites

Transfer Rates: Caltech Tier2 to Europe July 2014

- Upload rate: 27 Gbps; 20Gbps to CNAF (Italy) Alone
- By Spring 2015: 12 – 40 Gbps downloads were routine to US CMS Tier2 Sites with 100G Links
- Downloading Terabyte Datasets to Tier3s and Tier4s (to the desktop/laptop) is being explored



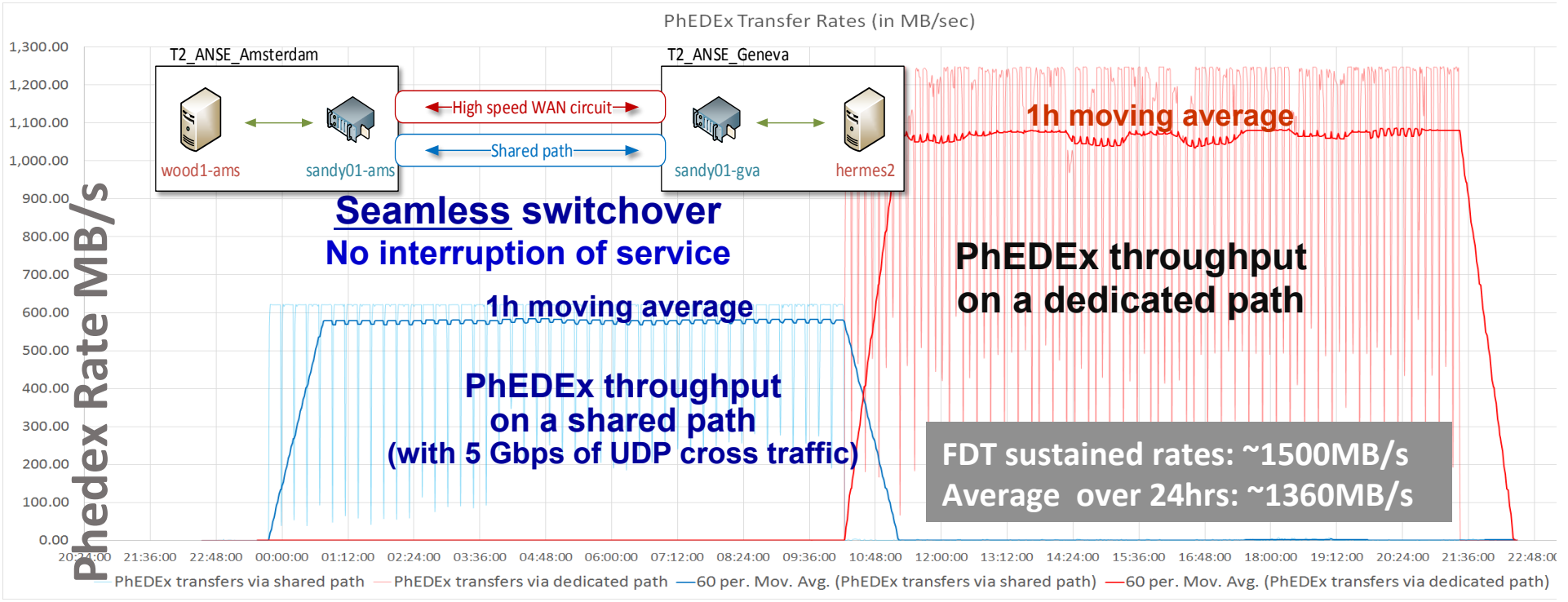
US CMS university based Tier2s have moved to ~100G	
Caltech	100 Gbps
Florida	100 Gbps
MIT	100 Gbps
Nebraska	100 Gbps
Purdue	100 Gbps
UCSD	80 Gbps
Wisconsin	100 Gbps

PhEDEx and Dynamic Circuits

Using dynamic circuits in PhEDEx allows for more deterministic workflows, useful for co-scheduling CPU with data movement

Integrating circuit awareness into the FileDownload agent:

- **Application is backend agnostic; No modifications to PhEDEx DB**
- **All control logic is in the FileDownload agent**
- **Transparent for all other PhEDEx instances**

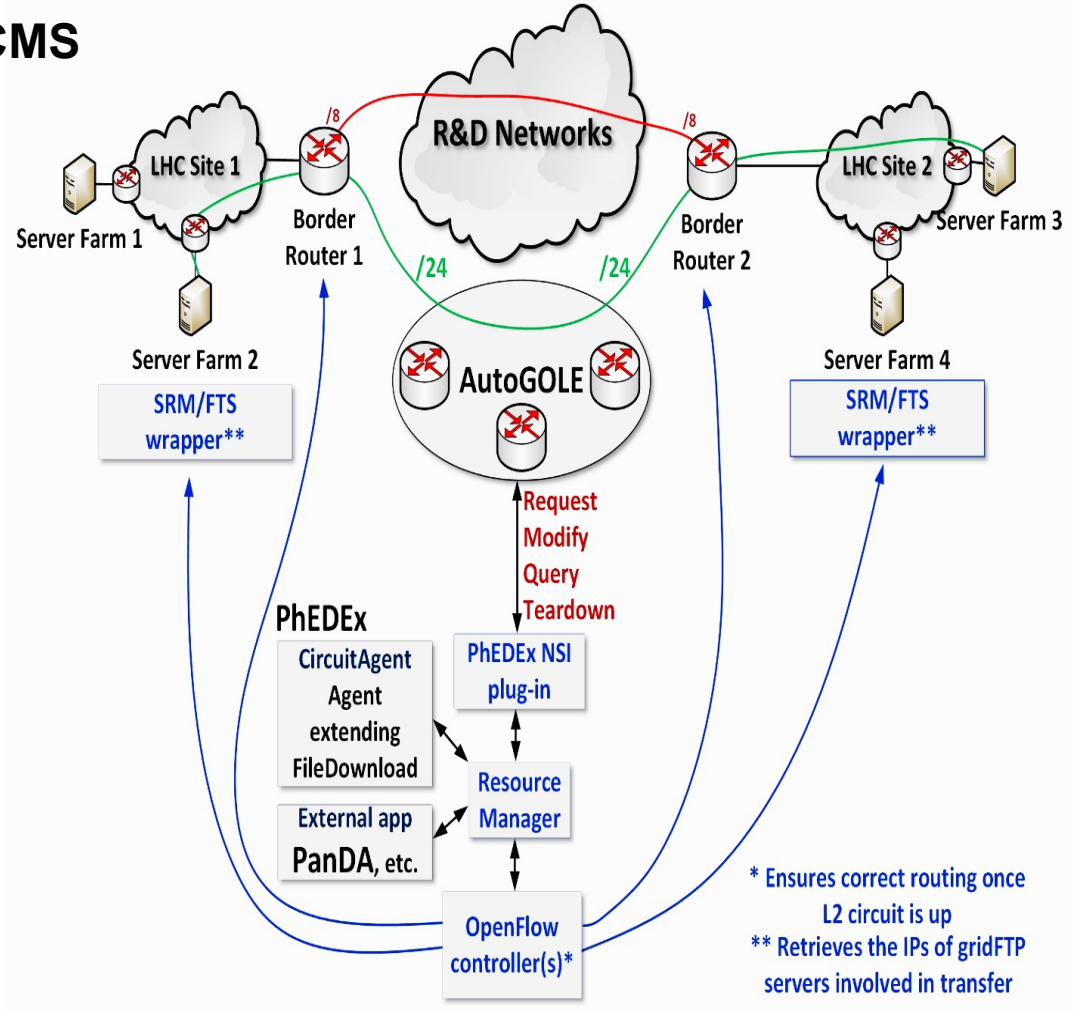


Advanced Network Services for Experiments (ANSE)
Integrating PhEDEx with Dynamic Circuits for CMS

Lapatadescu, Wildish, Mughal, Bunn, Legrand, Newman

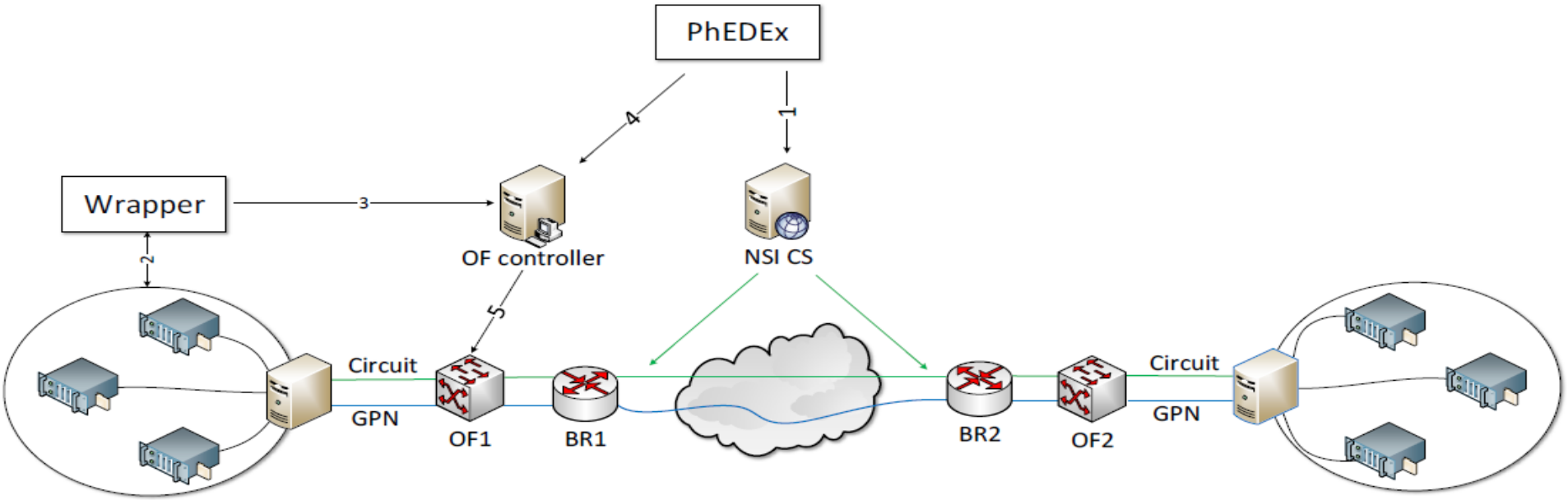
- Building on the AutoGOLE Fabric of Open Exchange Points and NSI emerging standard virtual circuits OSCARS + NSI circuits are used to create WAN paths with reserved bandwidth across the AutoGOLE fabric
- OF flow-matching is done on specific subnets to route only the desired traffic
- Openflow also can be used to select paths outside the fabric

ANSE: 1st real life application integration of NSI and the AutoGOLE fabric with PhEDEx for CMS



Integrating NSI Circuits with PhEDEx: Control Logic

1. PhEDEx requests a circuit between sites A and B; waits for confirmation
2. Wrapper gets a vector of source and destination IPs of all servers involved in the transfer, via an SRM plugin
3. Wrapper passes this information to the OF controller
4. PhEDEx receives the confirmation of the circuit, informs the OF controller that a circuit has been established between the two sites
5. OF controller adds routing information in the OF switches that direct all traffic on the subnet to the circuit

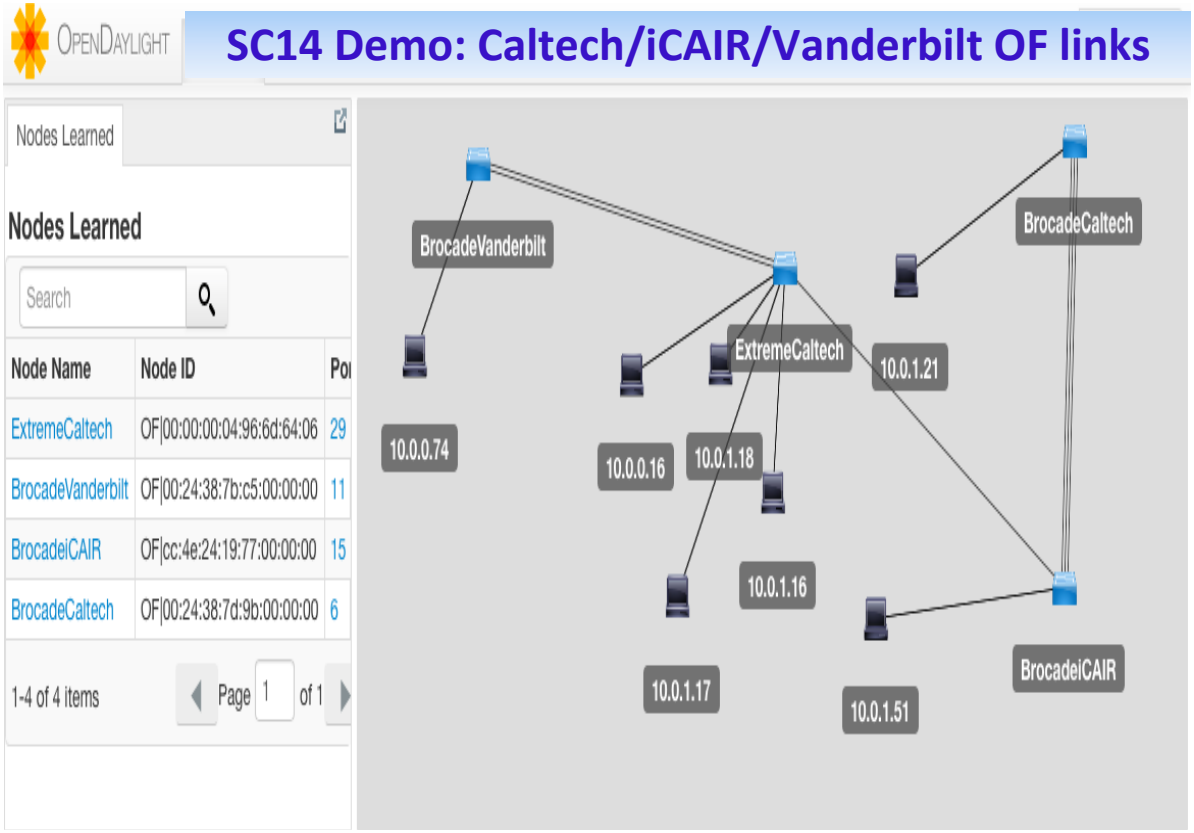


SDN Multipath OpenDaylight Demonstrations at Supercomputing 2014

- 100 Gbit links between Brocade and Extreme switches at Caltech, iCAIR and Vanderbilt booths
- 40 Gbit links from many booth hosts to switches
- Single ODL/Multipath Controller operating in "reactive" mode
  - For matching packets: Controller writes flow rules into switches, with a variety of path selection strategies
  - Unmatched packets "punted" to Controller by switch

Demonstrated:

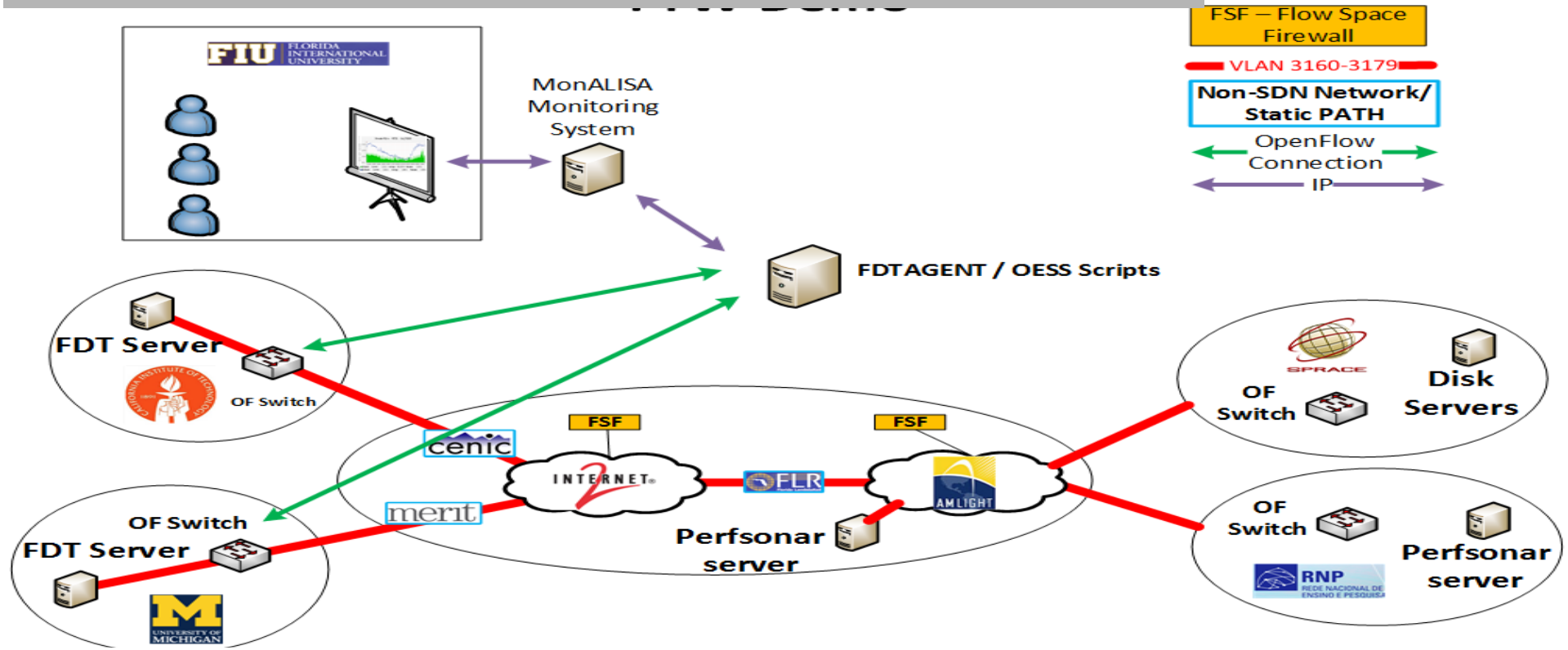
- Successful, high speed, flow path calculation, selection and writing
- OF switch support from vendors
- Resilience against changing net topologies [At layer 1 or 2]
- Monitoring and Control



Focused Technical Workshop Demo 2015:
SDN-Driven Multipath Circuits

A. Mughal
J. Bezerra

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

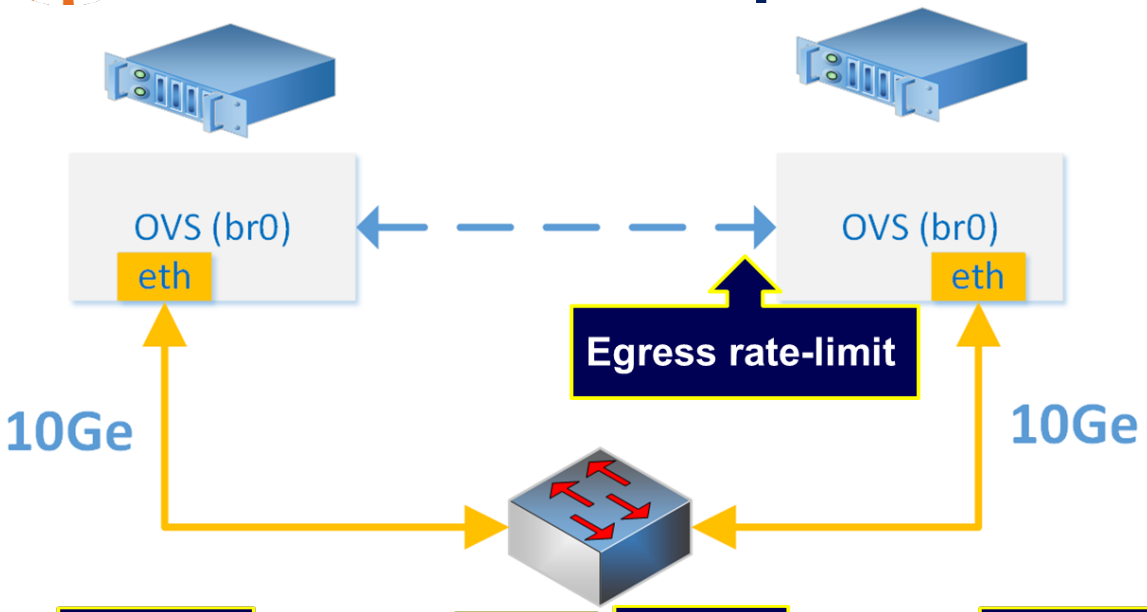


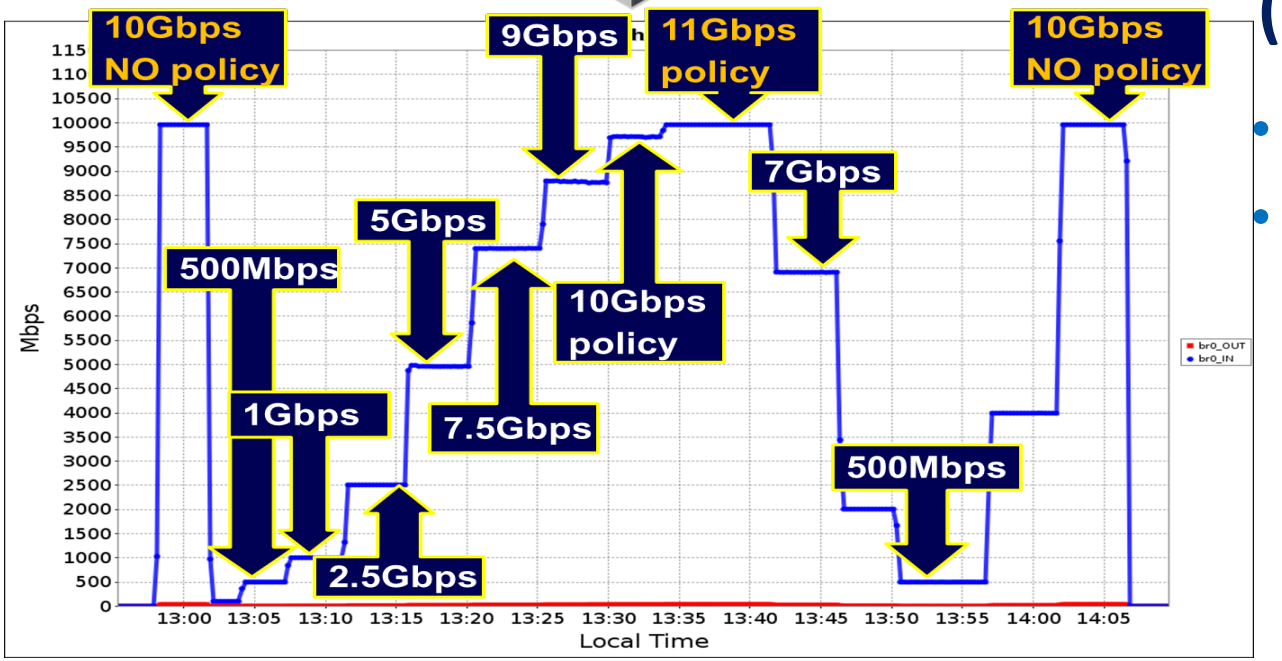
- **Hardened OESS and OSCARS installations at Caltech, Umich, AmLight**
- **Updated Dell switch firmware to operate stably with OpenFlow**
- **Dynamic circuit paths under SDN control**
- **Prelude to the ANSE architecture: SDN load-balanced, moderated flows**

SDN Demonstration at the FTW Workshop. Partners: Caltech, UMich, Amlight/FIU, Internet2, ESnet, ANSP, RNP

- **Dynamic Path creation:**
  - Caltech – Umich
  - Caltech/Umich - SPRACE
  - Caltech – RNP
  - Umich – AmLight
- **Path initiation by the FDTAgent**
- **OESS for OpenFlow data plane provisioning over Internet2/AL2S**
- **MonALISA agents at the end-sites provide detailed monitoring**

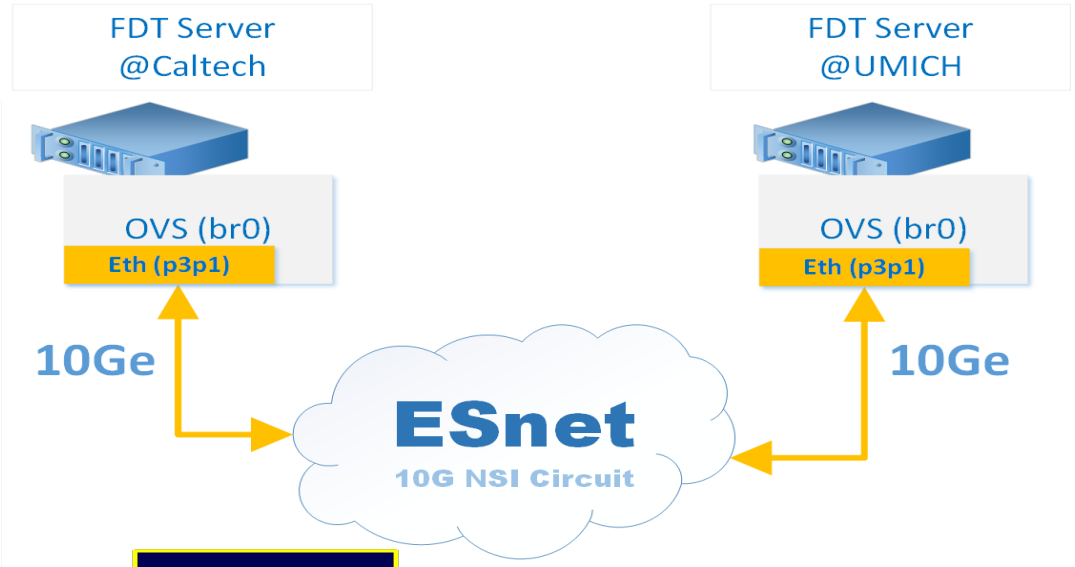
SDN QoS: Traffic Shaping with Open vSwitch (OVS)





- OVS 2.3.1 with stock RH 6.x kernel
- OVS bridged interface achieved same performance as hardware (10Gbps)
- egress rate-limit
- Based on Linux kernel:
  - HTB (Hierarchical Token Bucket)
  - HFSC (Hierarchical Fair-Service Curve)

Use Case: Traffic Shaping with Open vSwitch (OVS) WAN tests over NSI

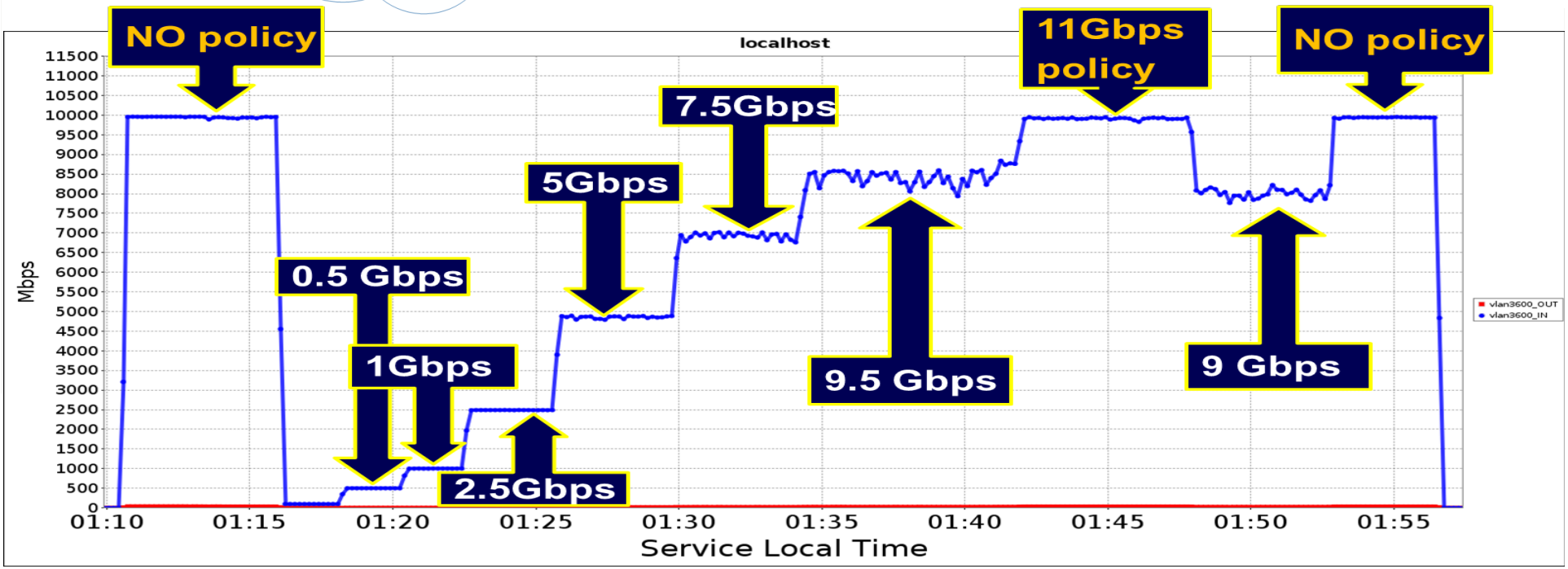


OVS 2.4 with stock kernel

NSI circuit Caltech -> UMICH (~60ms)

Very stable up to 7.5Gbps

Fairly good shaping above 8Gbps (small instabilities)



OVS benefits

- Standard OpenFlow (or OVSDB) end-host orchestration
- QoS SDN orchestration in non-OpenFlow clusters
- OVS works with stock SL/CentOS/RH 6.x kernel used in HEP
- OVS bridged interface achieved the same performance as the hardware (10Gbps)
- No CPU overhead when OVS does traffic shaping on the physical port
- Traffic shaping (egress) of outgoing flows may help performance in such cases when the upstream switch (or ToR) has smaller buffers

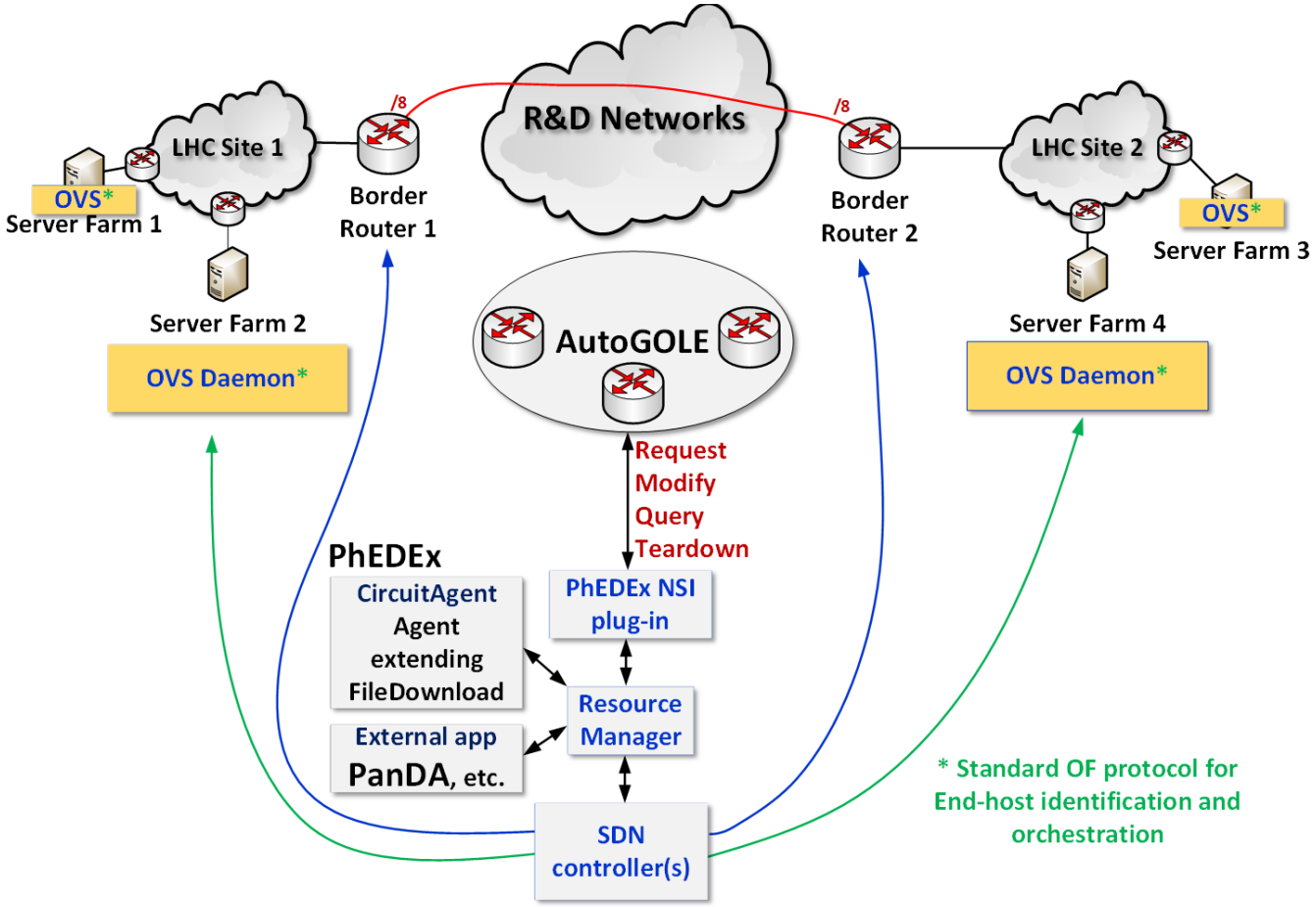
https://indico.cern.ch/event/376098/contribution/24/material/slides/1.pdf

Using OVS for end-host orchestration
Integrating PhEDEx with Dynamic Circuits for CMS

Standard OpenFlow (or OVSDB) protocol for end-host network orchestration (no need for custom SB protocol) Simple procedure to migrate to OVS on the end-host. SDN controller not required in the initial deployment phase

Host type (storage, compute) dynamically discovered using OF identification string

Use SDN controller to create an overlay network from circuit end-point (Border Router) to the storage



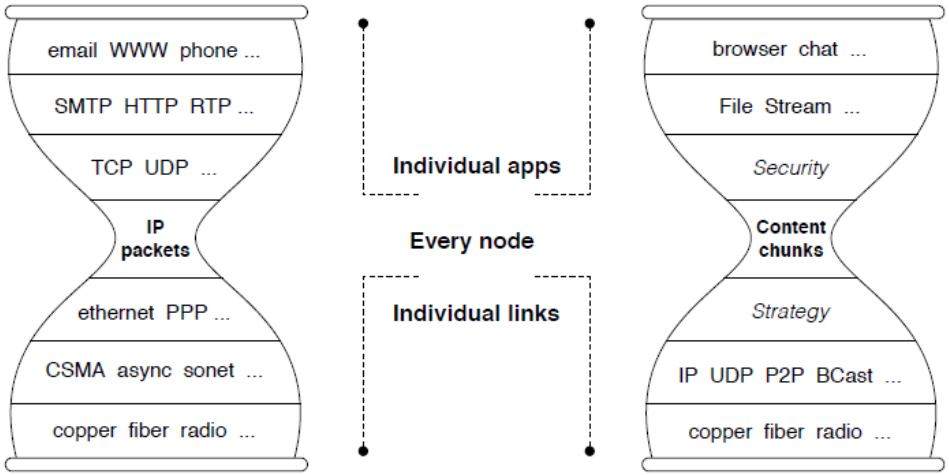
NAMED DATA NETWORKING

Caltech

Named data networking (NDN)

- One of five projects funded by NSF under Future Internet Architecture Program
- Content-Centric networking (CCN) precursor of NDN (2006)
- Novel architecture: evolution from host-centric (IP) to data-centric (NDN)
- "NDN changes the semantics of network service from delivering the packet to a given destination address to fetching data identified by a given name"

- No restrictions imposed on the naming scheme
- The names can identify a data chunck in a file, an end-point, etc



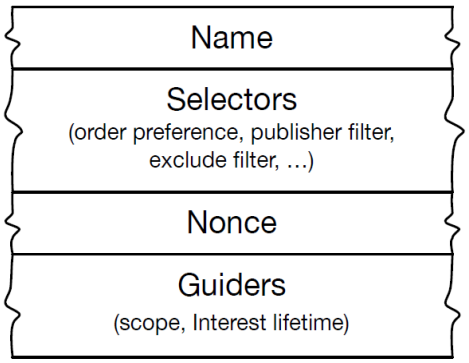
Named data networking (NDN)

- The communication is driven by the receivers
- Two types of packets:
  - Interest – consumer puts the name and sends it
  - Data – contains both the name and the content
- Digital signatures for packets
- NDN routers maintain three tables and a Forwarding Strategy(FS):
  - Pending Interest Table(PIT) – forwarded requests
  - Forwarding information base (FIB) – routing table
  - Content Store (CS) – local router cache
  - Forwarding Strategy (FS) – policies and rules about forwarding

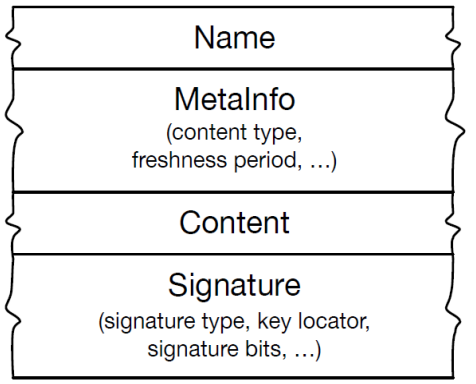
Software and more resources:
NDN collaboration (http://named-data.net):

- NFD (Networking Forwarding Daemon)
- NDN libraries for the client software

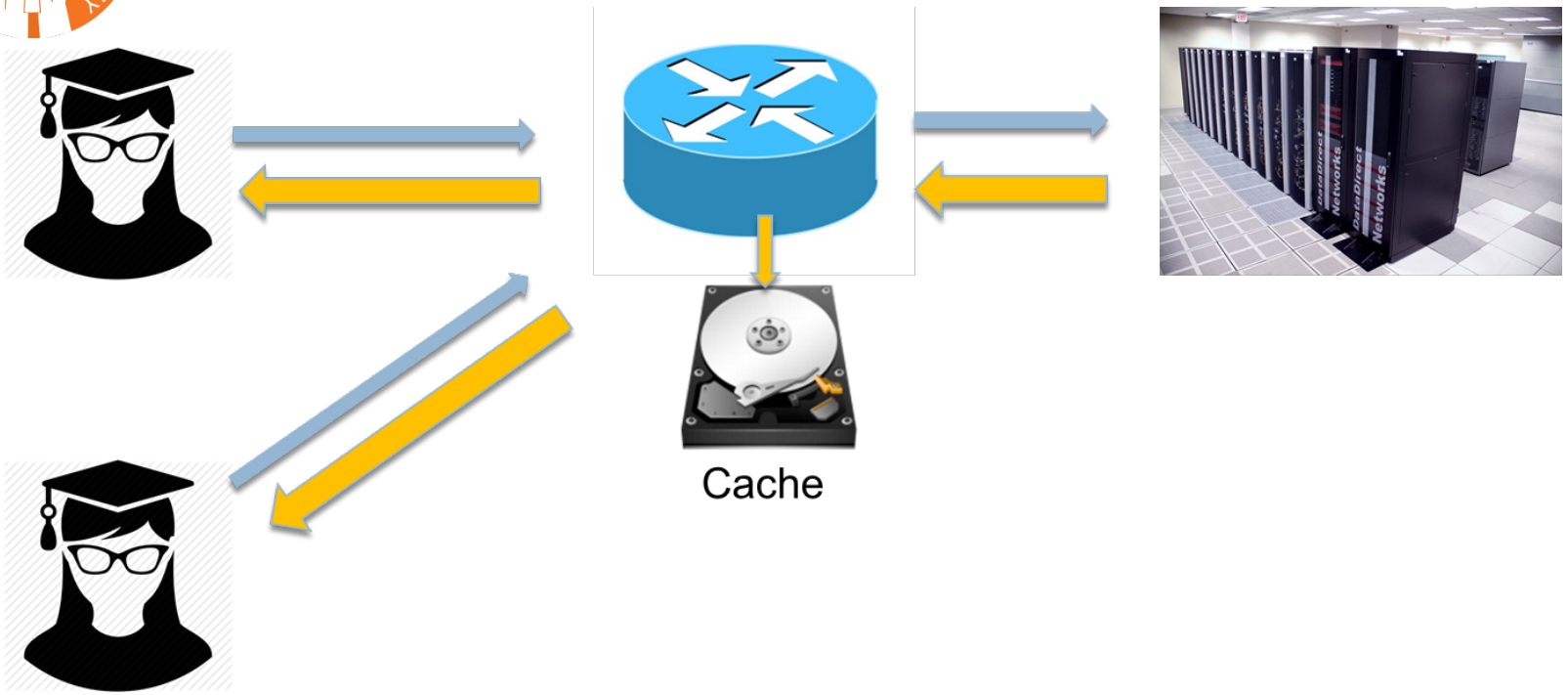
Interest Packet



Data Packet



Content Store(CS) for NDN



Possible CS/Repository server for NDN

- 100G NIC
- High-Performance Persistent Storage (SSD, NVME, …)
- Big Memory FS cache
- DTN looks like a good candidate

NDN & Virtual Interest Packets (VIP)

Edmund Yeh
Northeastern University

- NDN allows combined Forwarding Strategy and Caching to optimally utilize bandwidth and storage

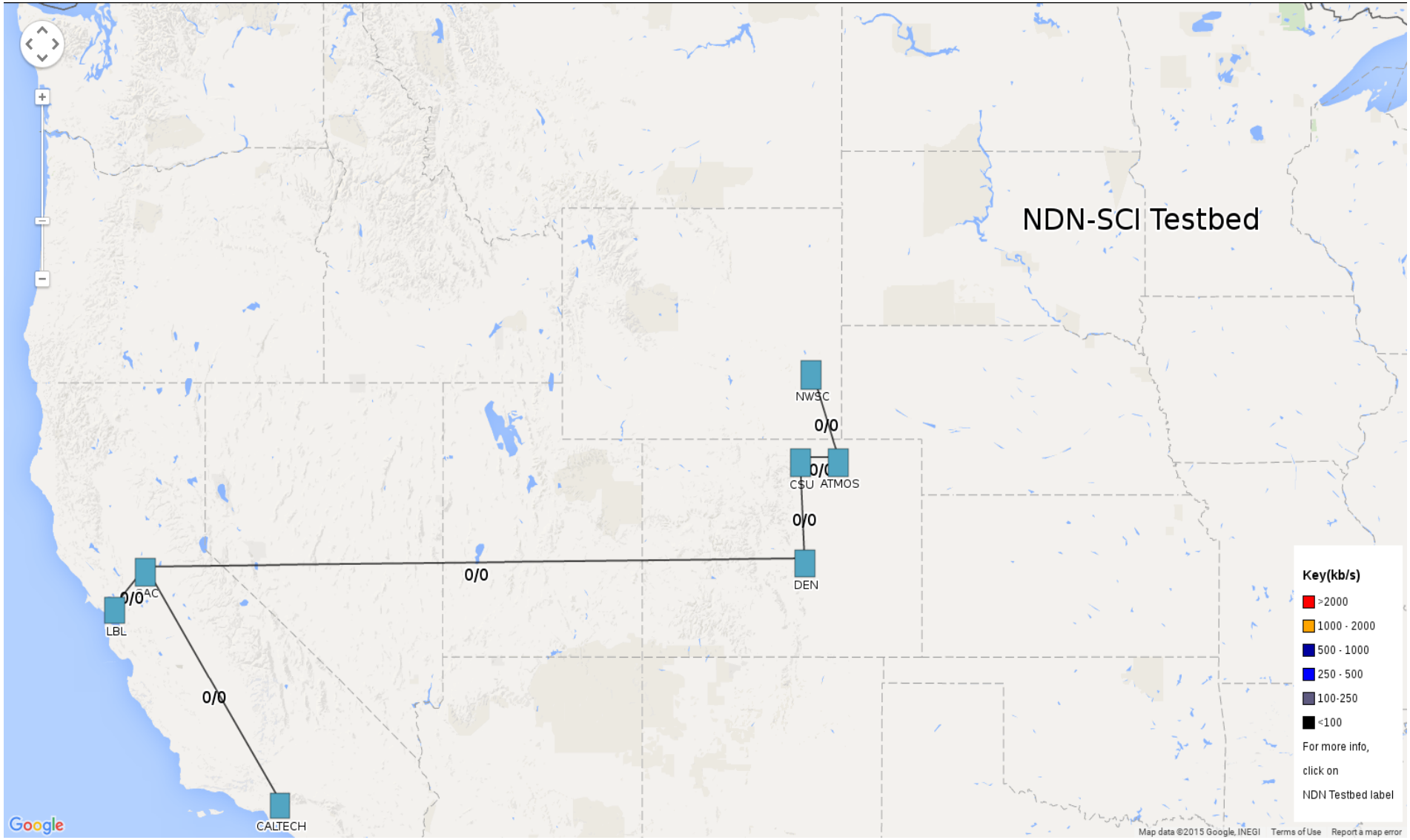
- VIP Algorithm (ACM ICN 2014):
  - On top of an existing NDN infrastructure
  - New metric which capture the measured demand for the respective data objects in the network
  - Maximizes data delivered by network
  - Load balance via multipath forwarding and dynamic caching
  - Superior delay performance
  - Optimally trades off link bandwidth with storage I/O bandwidth
  - Allows data chunks of one object to use same path
  - Can incorporate congestion control at request nodes

NDN - Current activities

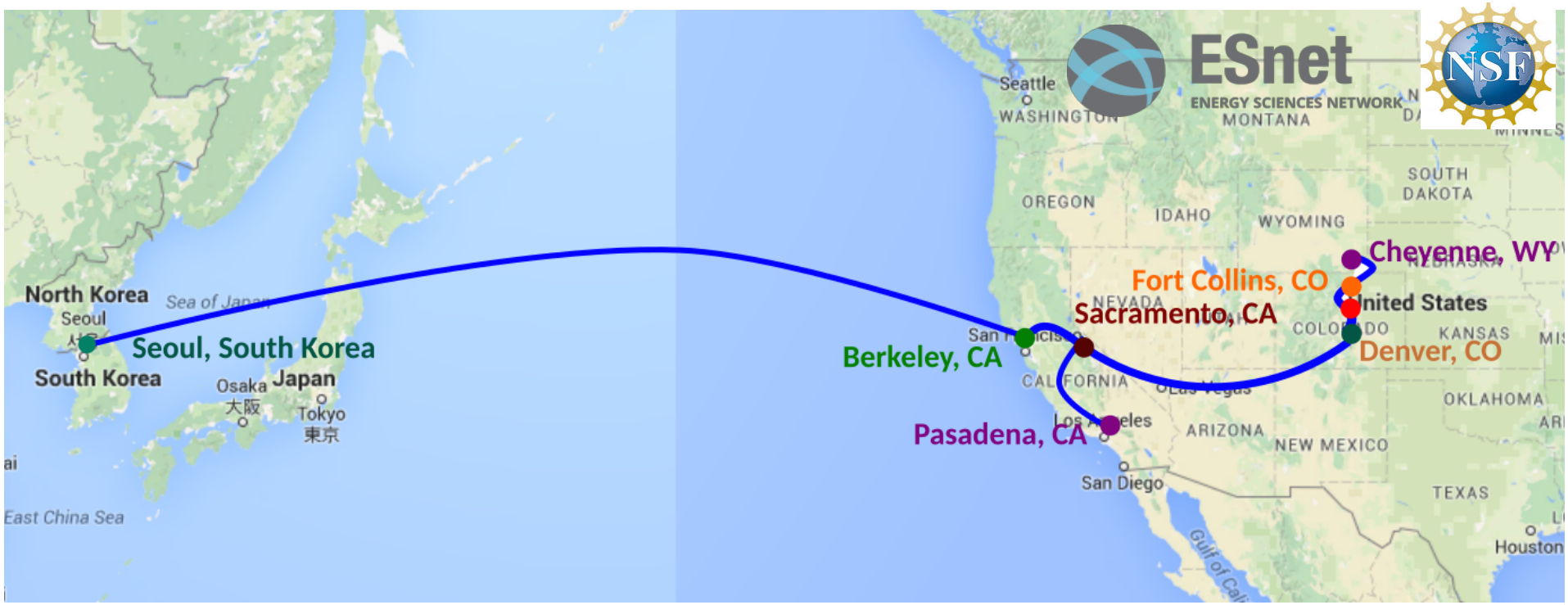
- Build a testbed including core, regional and campus networks
- Deploy a small number of in-network as well as edge caching switch-routers
- Simulation model for feedback + scaling studies (Northeastern)
- Investigate the use of DTN as NDN CS (caching)
- ROOT plugin (NDN client library) – developed by Imperial College London – [CHEP2015]
- Mapping CMS dataset names to an NDN structure (Vlimant, Caltech)

NDN – Science Testbed

Christos Papadopoulos, Susmit Shannigrahi
Colorado State University



NDN – Science Testbed



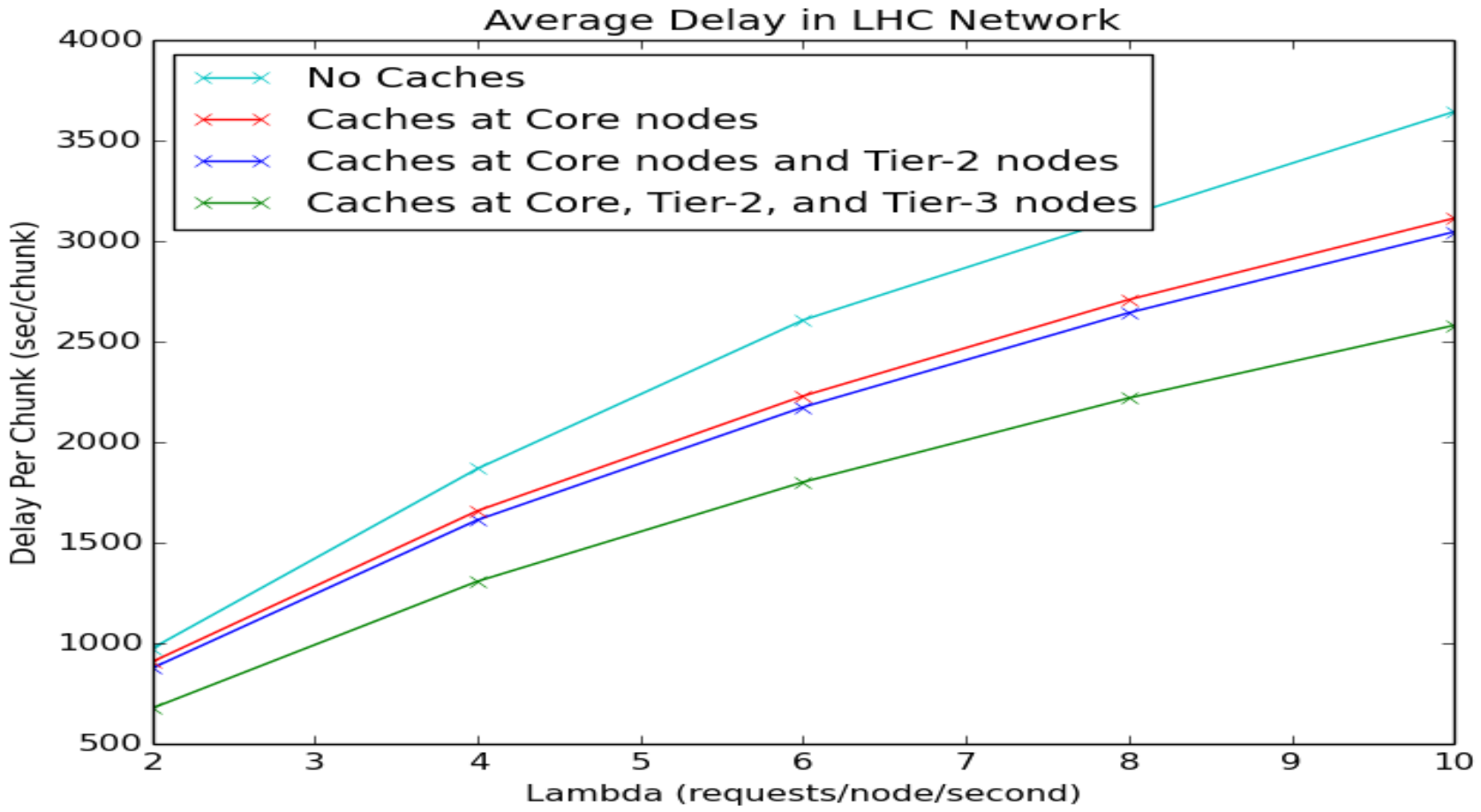
- NSF CC-NIE campus infrastructure award
  - 10G testbed (courtesy of ESnet, UCAR, and CSU Research LAN)
- Currently ~50TB of CMIP5 (Climate), ~70TB of HEP data

LHC data distribution simulation

Edmund Yeh

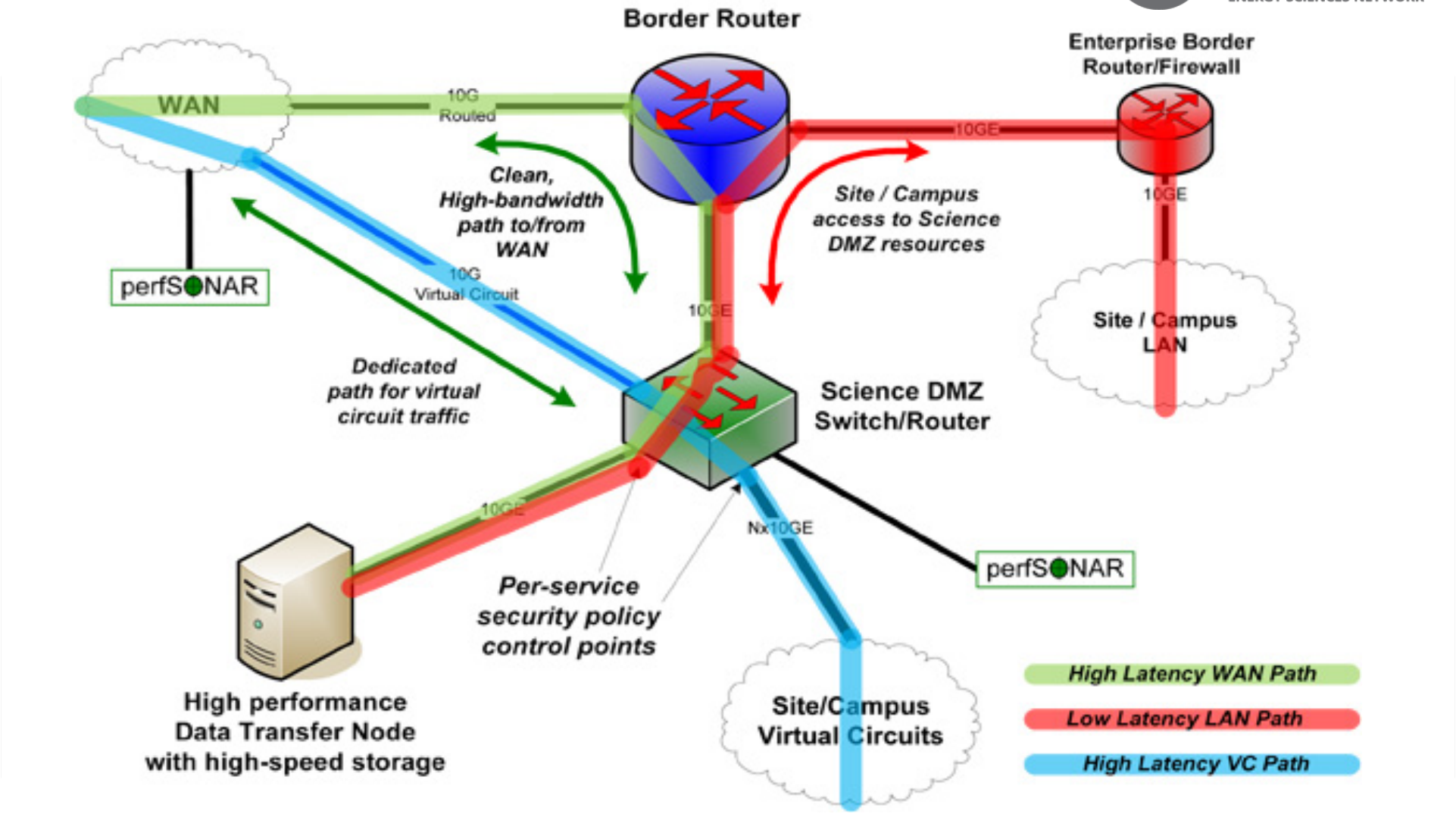
- Simulation using 83 sites: 17 Tier2 and 66 Tier3
- Interest Packet size: 125 Bytes, Data Packet (chunk) size: 200 MBytes, Data Object size: 2 Gbytes(10 chunks/object)
- 20000 objects in total
- Core LHC network using Internet2 topology
- Core link capacities: 100 Gb/sec; Link capacity from exchange points to most edge nodes: 10 Gb/sec; Other link capacities set according to topology
- Requests arrive at nodes following Poisson process with overall rate λ
- Popularity distribution: Zipf distribution with parameter 0.8
- Multipath forwarding/dynamic caching using VIP algorithm

Simulation results



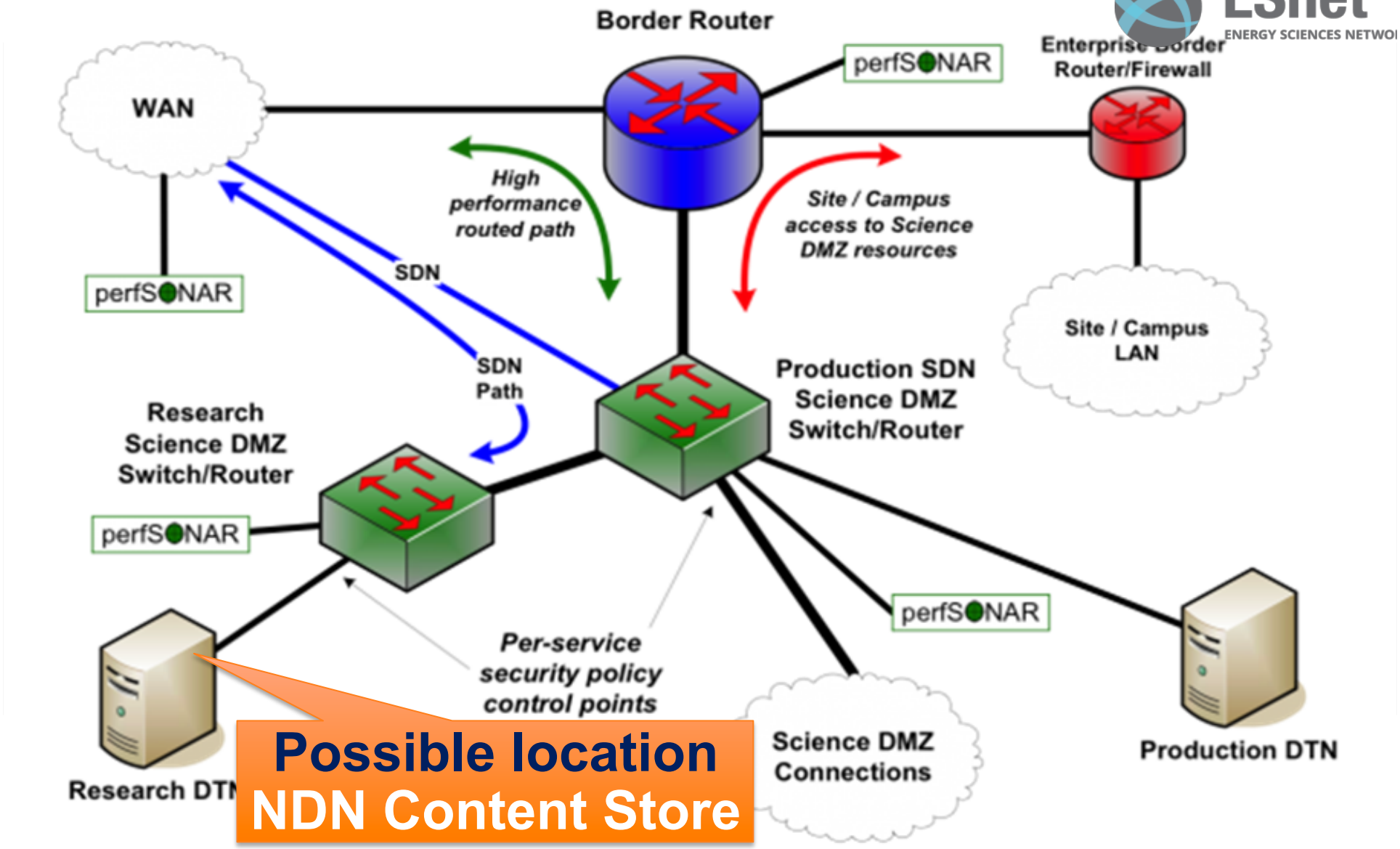
- No caches in network
- 16 TByte caches at 4 core nodes
- 16 TByte caches at 4 core nodes, 8 TByte caches at Tier 2 nodes
- 16 TByte caches at 4 core nodes, 8 TByte caches at Tier 2 and Tier 3 nodes

Science DMZ

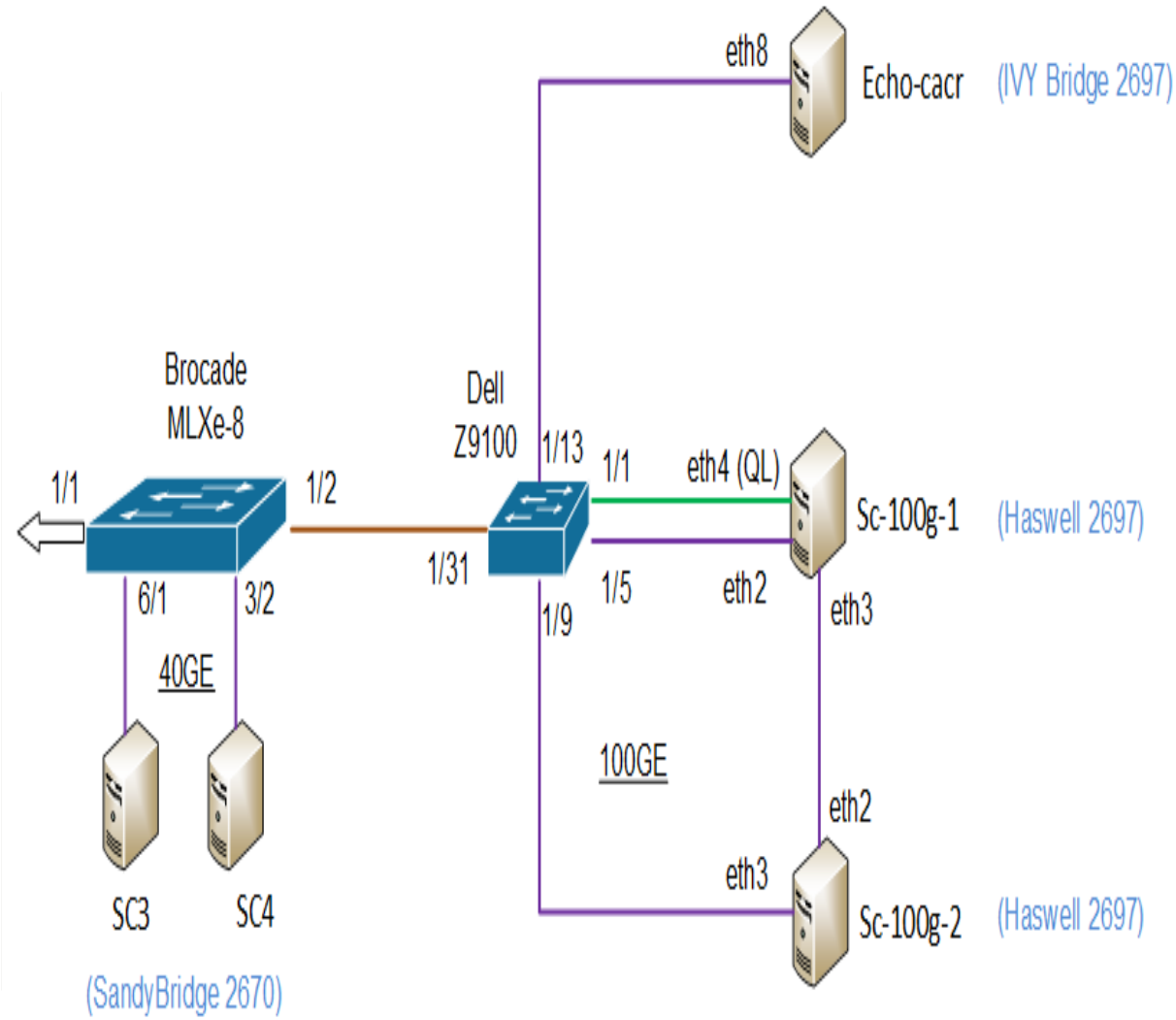


Recommended Science DMZ - Production Deployment:
http://fasterdata.es.net

Research DTN:
Possible intersection - SDN and NDN



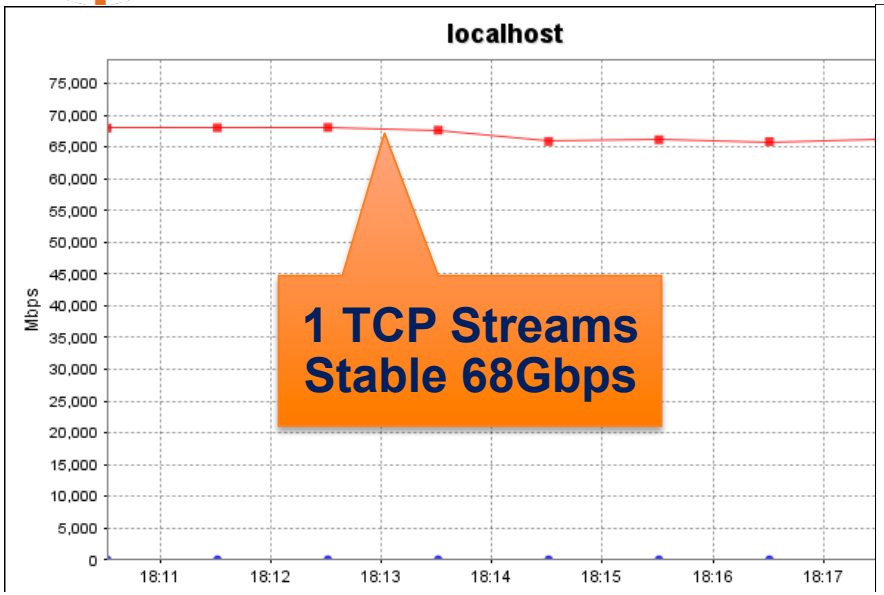
DTN: 100G network tests

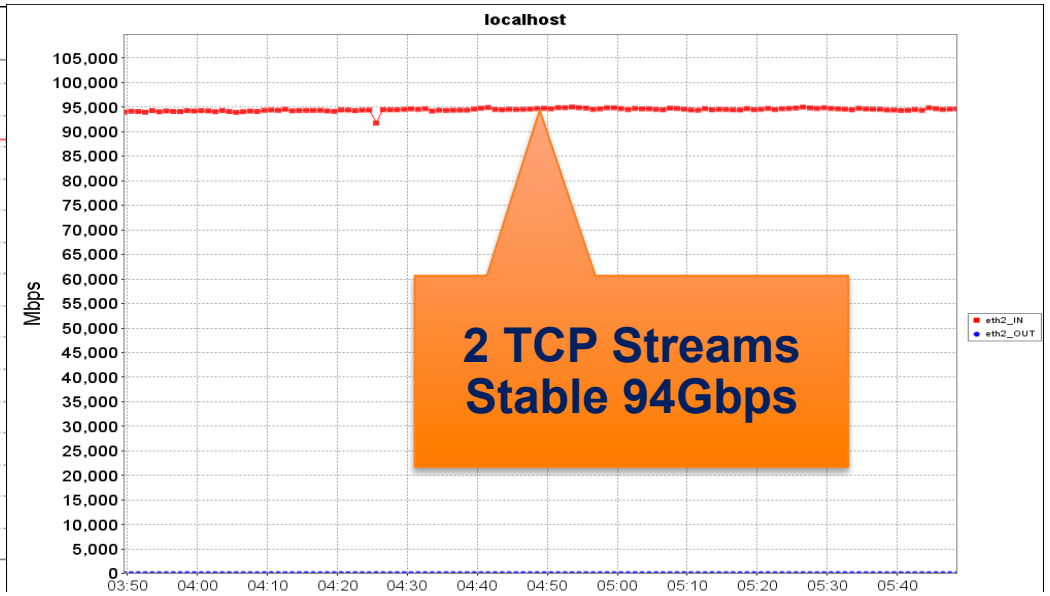


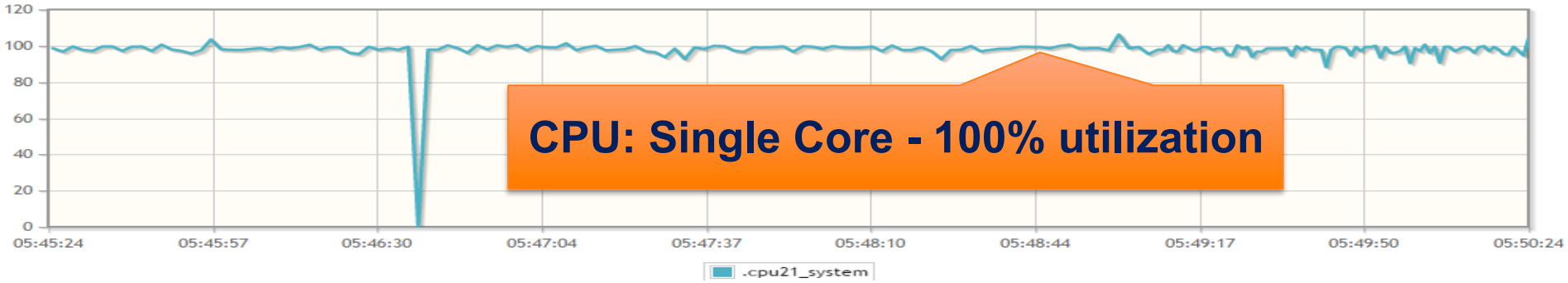
Network Topology:
Servers used in tests: Sc-100G-1 and Sc100G-2

- 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 Cables from Elpeus for switch connections
- 100G CR4 Cables from Mellanox for back to back connections

100G TCP tests







Client #numactl --physcpubind=20 --localalloc java -jar fdt.jar -c 1.1.1.2 –nettest -P 1 -p 7000

Server #numactl --physcpubind=20 --localalloc java -jar fdt.jar -p 7000

Line rate (100G) with 4+ TCP Streams

Pacific Research Platform (DTN Test)

Disk to Disk Transfers (actual files on disk)

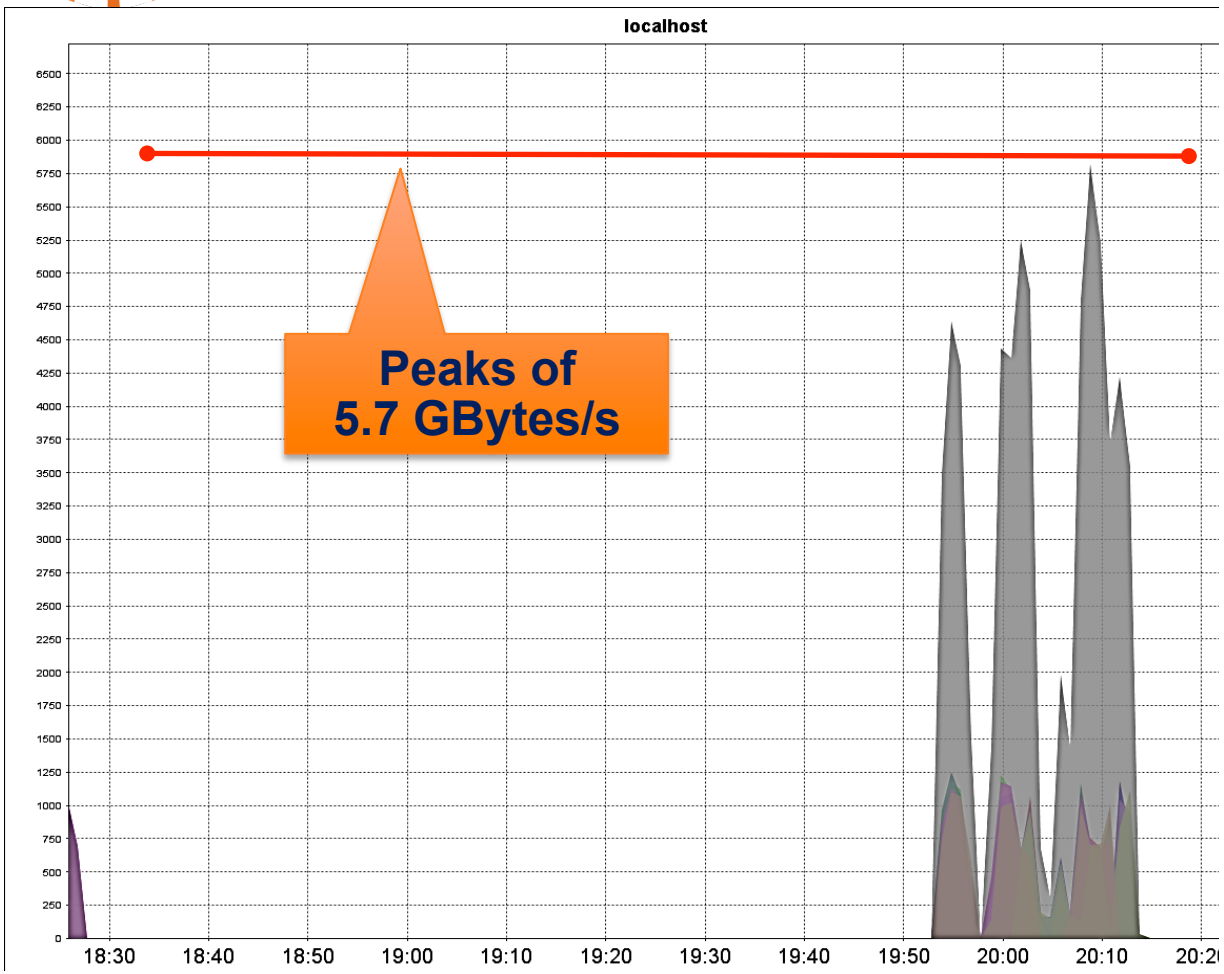
UCSD → Caltech = Avg 36Gbps, Peaks of 40Gbps
UCLA → Caltech = Avg 35Gbps, Peaks of 36Gbps
UCSC → Caltech = Avg 10Gbps

Caltech – FDT Server = *4 x PCIe Gen3 NVME Drives*
UCSD – FIONA = Zpool *(16 x SSD Drives, Intel 530, 120G)*
UCLA – FIONA = Zpool *(16 x SSD Drives, Intel 530, 120G)*

Several issues (*mostly Disk I/O related*):
NVME Drives:

- **Zpool performance Tuning, poor performance compared to individual drives**
- **Software RAID, sync hangs, system crashes**

Dual DTN Test – Disk performance



Reading from disk at both UCSD and UCLA
The disk pools at each of the - Zpool of 16 Intel 530 SSD drives
Receiver side at Caltech 8 Intel D3700 NVME drives were used (4 each for UCSD and UCLA)
Traffic peaked at 5.7GB/sec

At this rate with TCP, CPU cores were fully utilized.

Several issues (*mostly Disk I/O related*):
NVME Drives:

- **Zpool performance Tuning, poor performance compared to individual drives**
- **Software RAID, sync hangs, system occasional hangs**

THANK YOU!

QUESTIONS?

Caltech