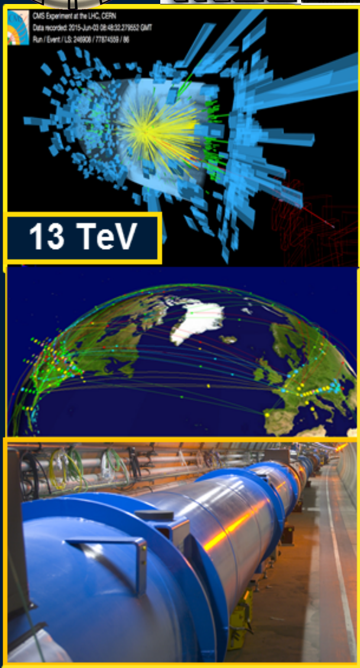
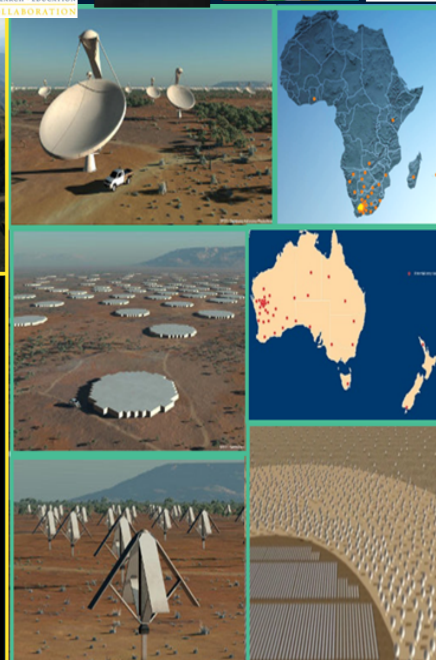
Next Generation Integrated Terabit/sec SDN Architecture for High Energy Physics and Exascale Science

+ INDUSTRY

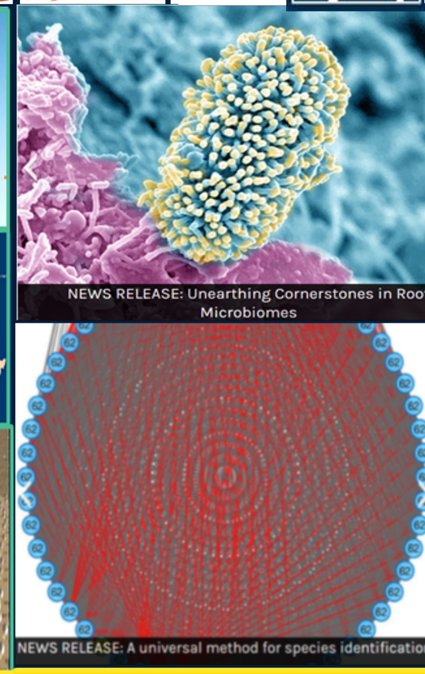


LHC

LSST



SKA



Joint Genome Institute

- *LHC Run2+: Beyond the Standard Model*
- *Data Intensive Exascale LCFs and SDN EcoSystems*

Gateway to a New Era

New Windows On the Universe

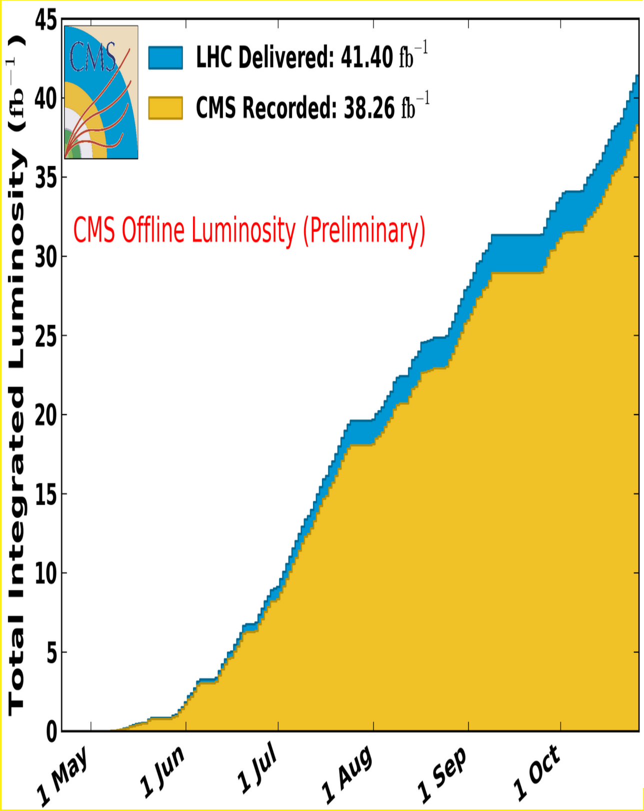
Harvey Newman, Caltech

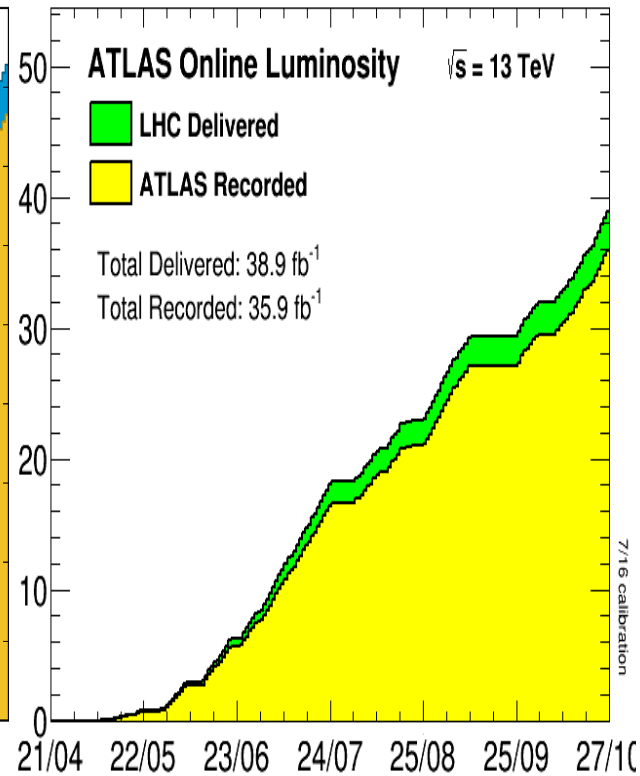
INDIS Workshop

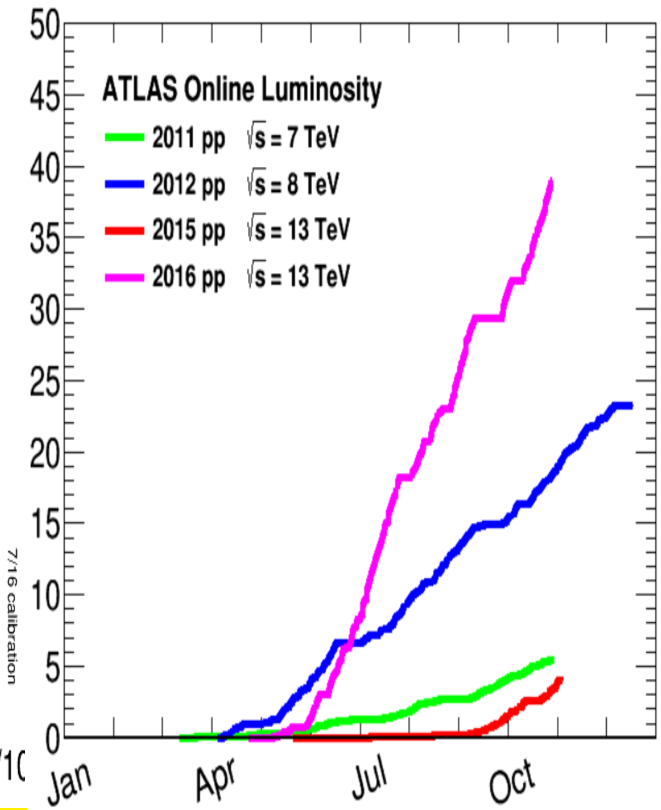
Salt Lake City, November 13 2016

On behalf of The Team and Partners

2016 LHC pp Luminosity to 50% Above Design Higher (to 90% Above ?) in 2017







40 Inverse Femtobarns Delivered ! 92+% Recorded

1.4-1.5 X 10^{34}/cm^2/sec Peak; μ to ~50 ! (Test to 95)

2017 Outlook: to 1.9 X 10^{34}/cm^2/sec, 56/fb with β* = 33 cm ?

Accelerated Challenges: Data Volumes Vs. Available Storage, CPU and Networks starting in 2017-2018

NGenIA
New SDN Paradigm
ExaO LHC rchestrator
Tbps Complex Flows
Machine Learning
LHC Data Traversal
Immersive VR

Thanks to Ecostreams, Orange Labs Silcon Valley

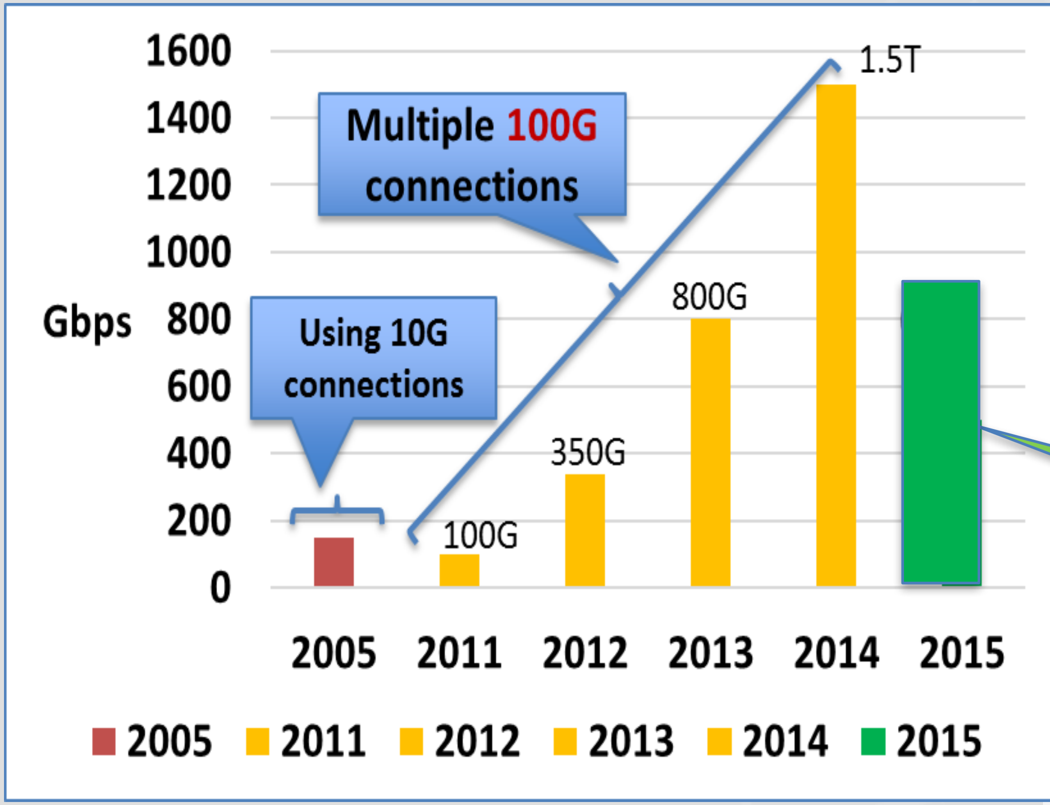
Visit Us at the Caltech Booths 2437, 2537

The Background: High Throughput and SDN

- The Caltech CMS group along with many R&E and industry partners has been participating in Bandwidth Challenges since 2000,and the LHCONE PointToPoint WG experiments for the last many years.
- The NSF/CC* funded projects Dynes and ANSE and the DOE/ASCR OliMPS and SDN NGeniA projects took the initiative to further strengthen these concepts and deliver applications and metrics for creating end to end dynamic paths across multi-domain networks and move TeraBytes of data at high transfer rates.
- Several large scale demonstrations during the SuperComputing conferences and at Internet2 focused workshops, have proved that such SDN-driven path building software is now out of its infancy and can be integrated into production services.

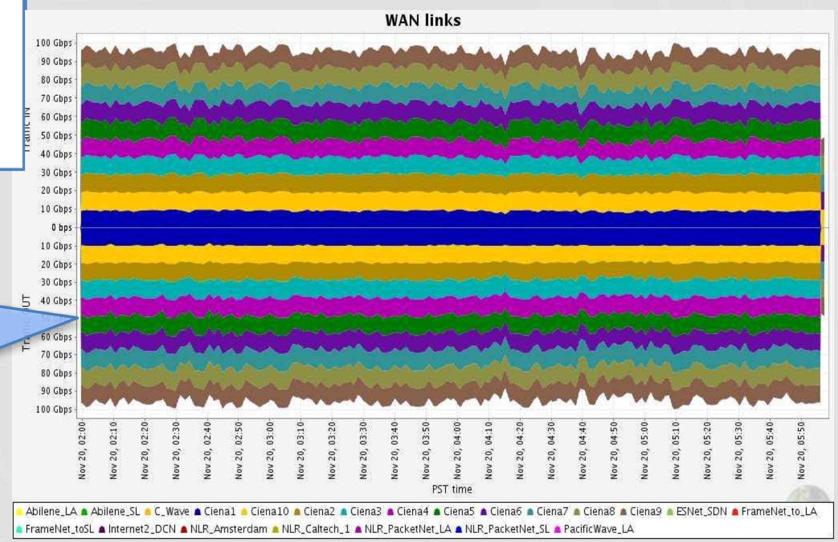
Bandwidth "explosions" by Caltech et al at SC

CALTECH HEP NETWORKING



SC02: FAST
SC05 (Seattle): 155Gbps (15 racks)
SC06: FDT
SC11 (Seattle): 100Gbps
SC12 (Salt Lake): 350Gbps
SC13 (Denver): 800Gbps
SC14 (Louisiana): 1.5Tbps
SC15 (Austin): ~ 750 – 900 Gbps
SC16 (Salt Lake): ~ *2.5Tbps (est.)*

Fully SDN enabled

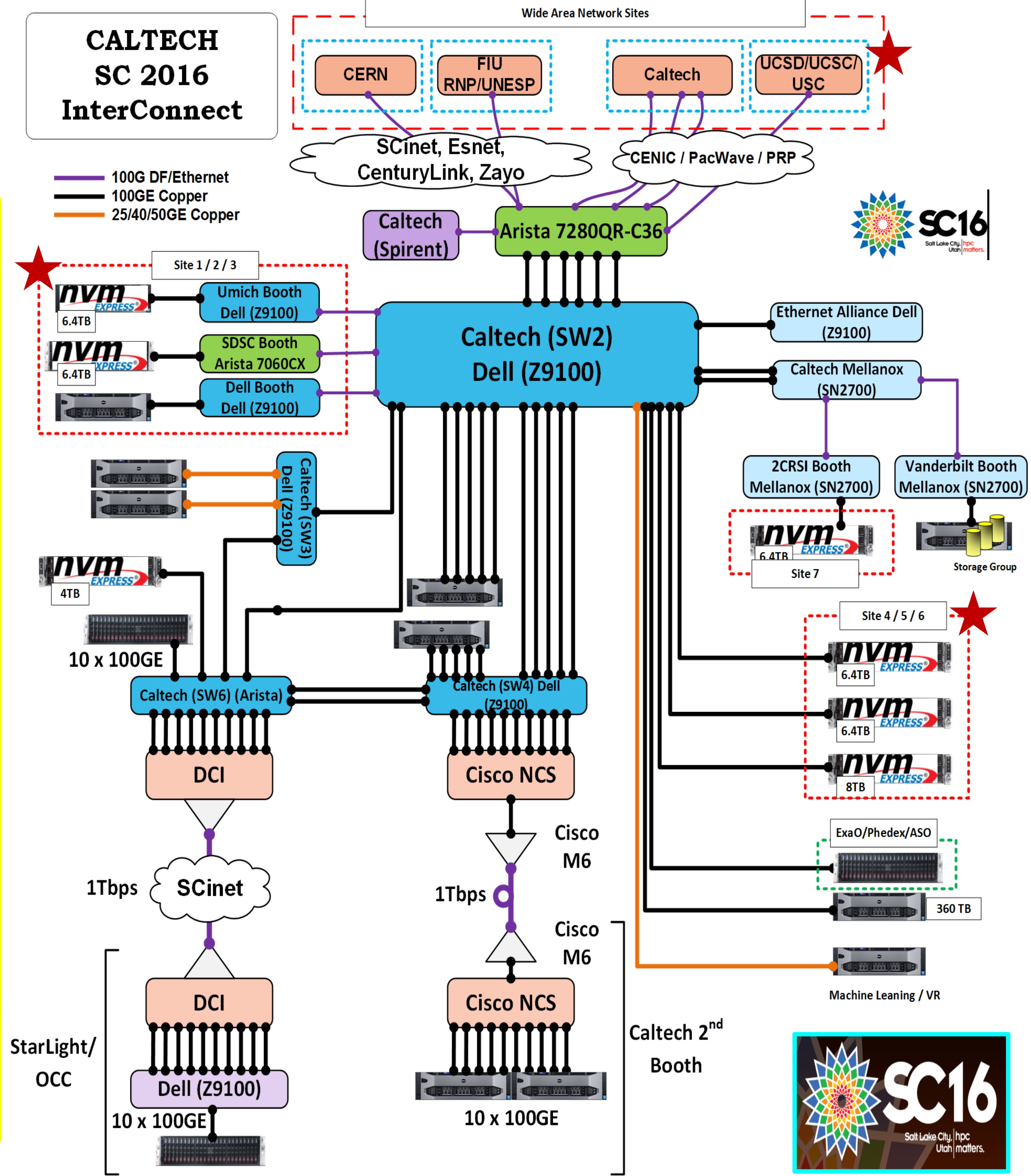


2008: First ever 100G OTU-4 trials using Ciena laid over multiple 10GE connections on the SC08 floor
191 Gbps bidirectional average:
1 Petabyte in 12 hours

Caltech at SC16

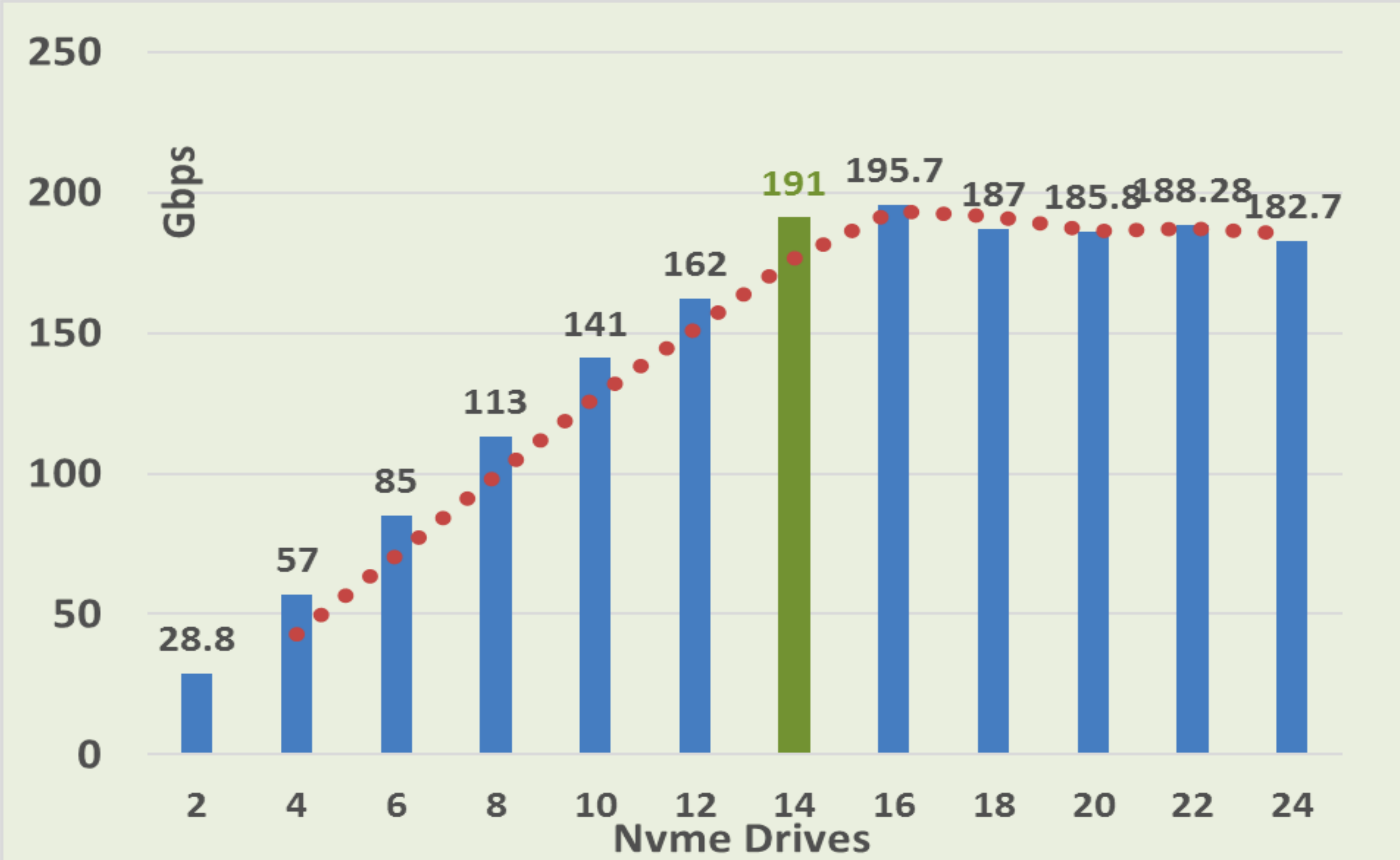
- Terabit/sec ring topology: Caltech – Starlight – SCInet; > 100 Active 100G Ports
- Interconnecting 9 Booths: Caltech 1 to 1 Tbps in booth, and to Starlight 1 Tbps; UCSD, UMich, Vanderbilt, Dell, Mellanox, HGST @100G
- WAN: Caltech, FIU+UNESP, PRP (UCSD, UCSC, USC), CERN, KISTI, etc.

★ ExaO + PhEDEx/ASO CMS Sites



2CRSI + Supermicro Servers with 24 NVMe drives

CALTECH HEP NETWORKING

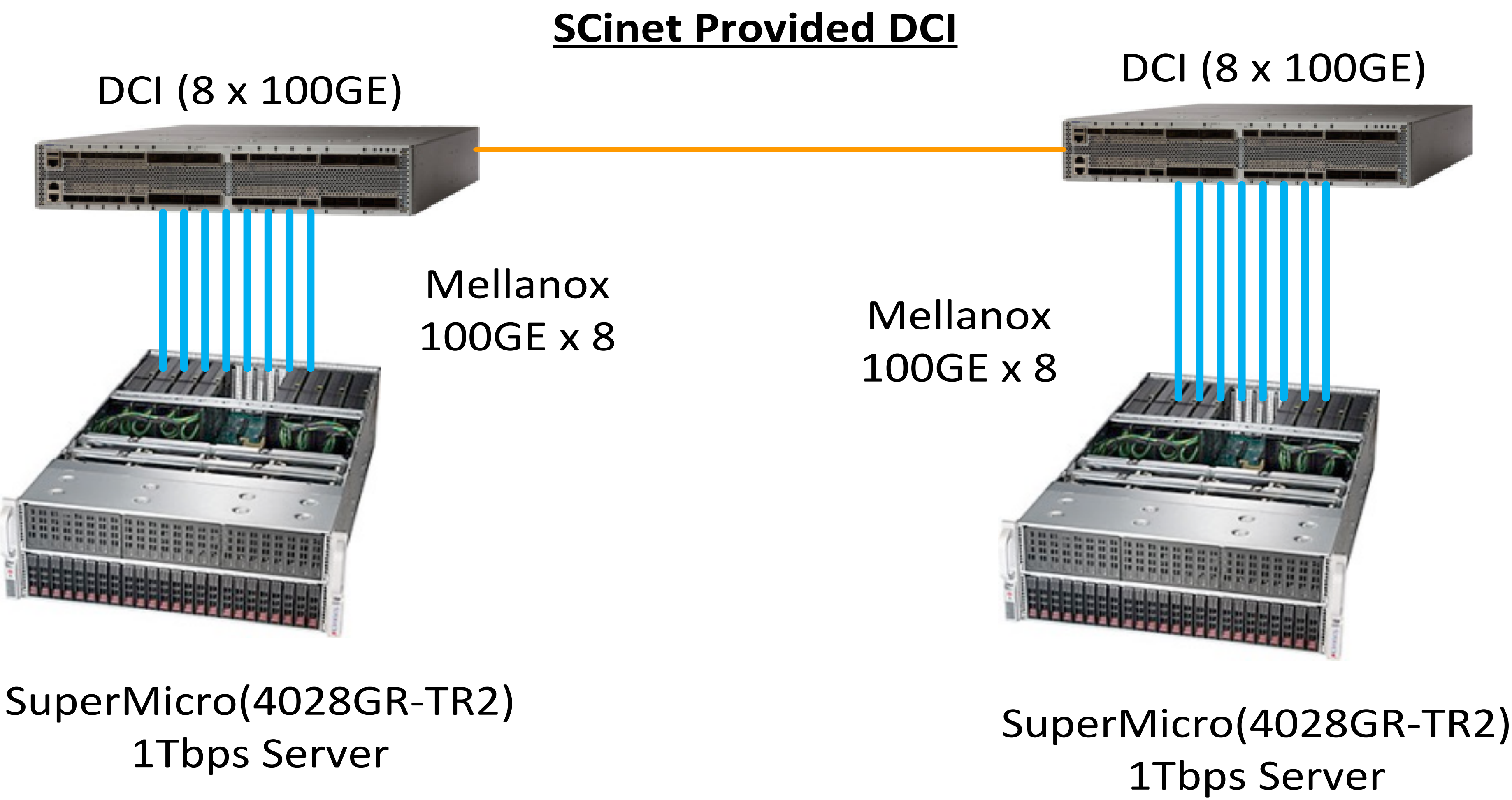


Max throughput reached at 14 drives (7 drives per processor)

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

1Tbps Caltech-StarLight (Between two booths)

- **Scinet + Ciena + Infinera have provided DCI inter-booth connection with a total of 8 (now to 10) 100G links**
- **RoCE based data transfers**
- **A proof of concept to demonstrate the current system capability and explore any limits**



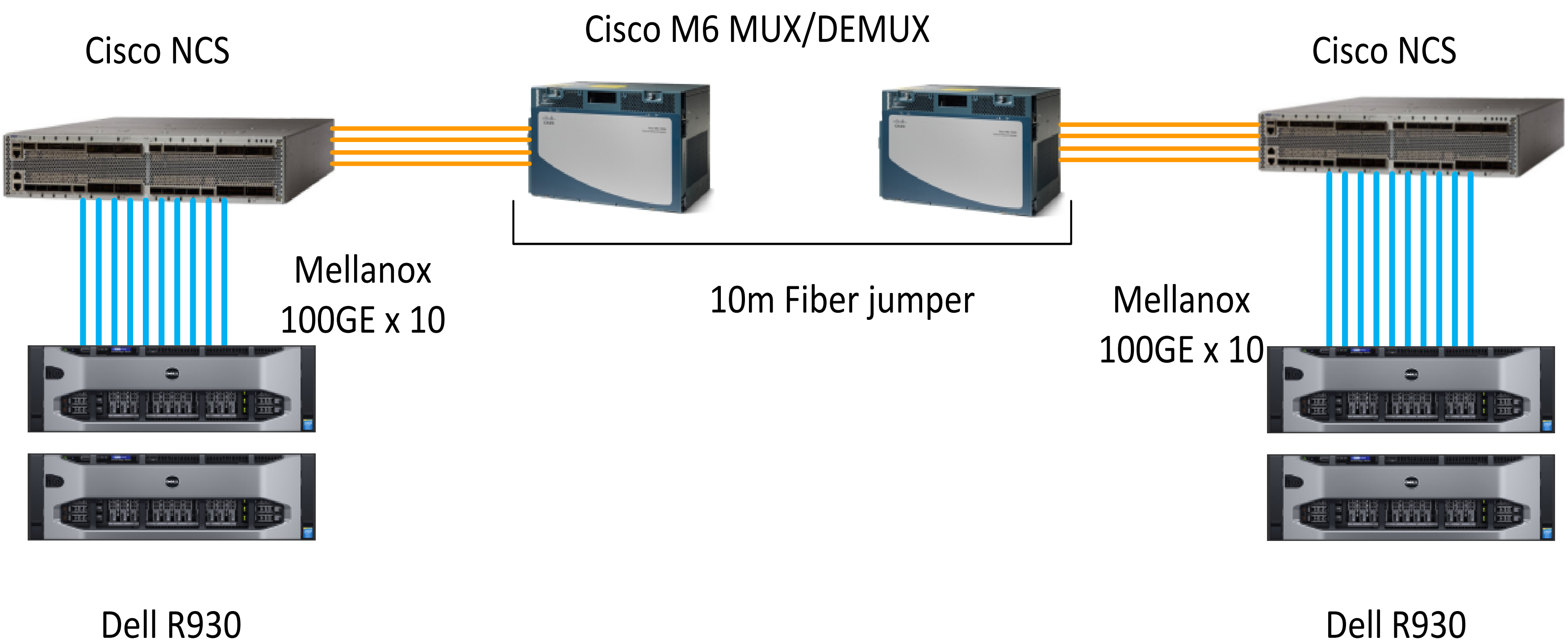
1Tbps Caltech - Caltech

Cisco has provided DCI inter-booth connection with a total of 10 x 100G links

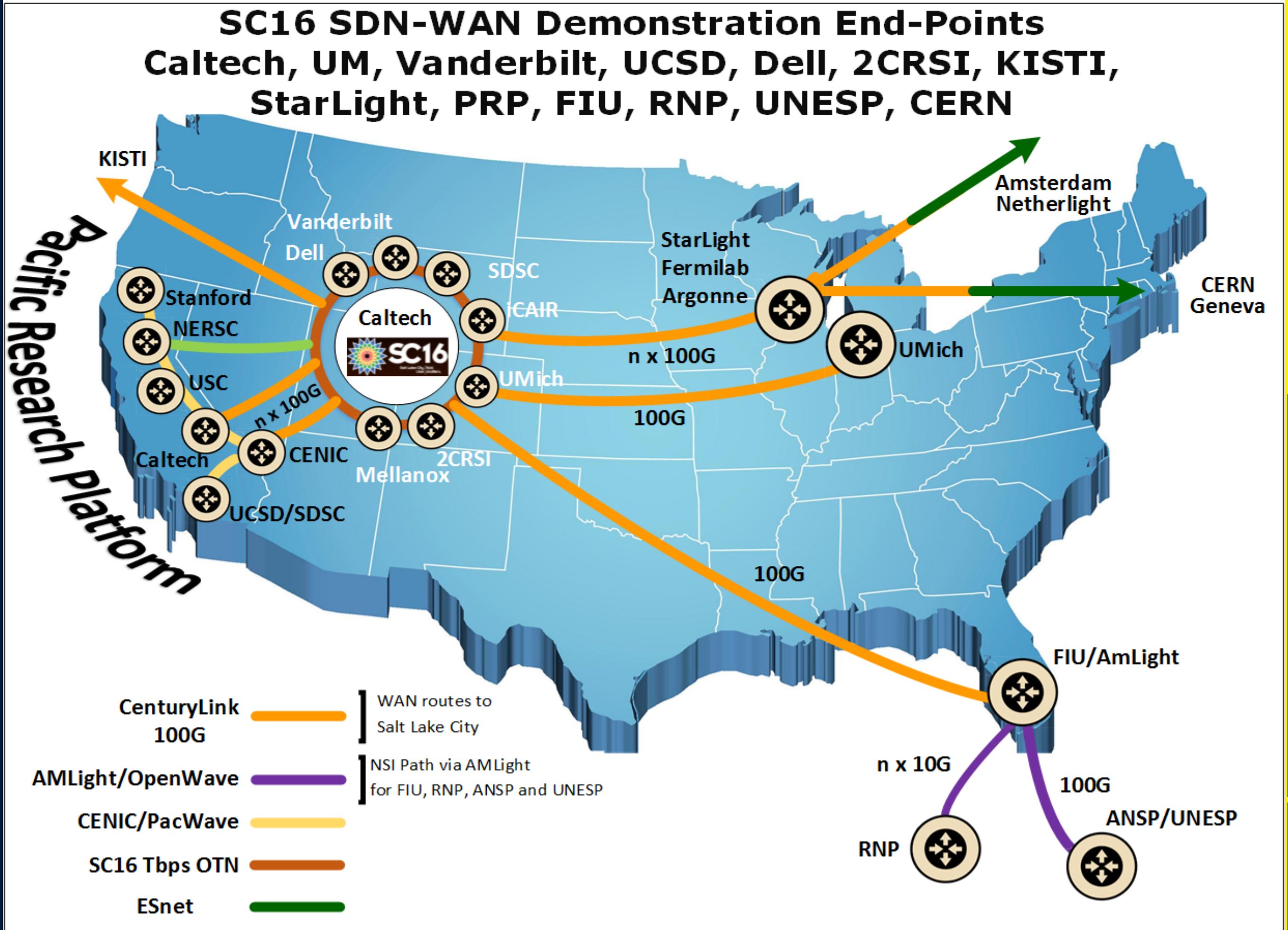
Links Between two pairs of Dell 930 4-socket servers

Transfer Application:

FDT: TCP/IP based application for data transfers



SC16: SDN Next Generation Terabit/sec Integrated Network for Exascale Science



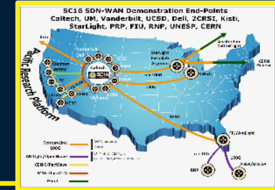
SDN-driven load balanced flow steering and site orchestration Over Terabit/sec Global Networks

Consistent Operations Edge & Core Limits With Agile Feedback: Major Science Flow Classes Up to High Water Marks

Preview PetaByte Transfers to/from Site Edges of Exascale Facilities With 200G+ DTNs

Caltech, Yale, UNESP & Partners: Open Daylight Controller, OVS and ALTO higher level services, New SDN Programming Framework

Caltech and Partner "NGenIA" Demonstrations: Booths 2437, 2537



- Towards "Consistent Operations": New paradigm + programming environment for complex networks linking major facilities & science teams
- A new Terabit/sec network complex interconnecting 9 booths with many 100GE local and wide area connections to remote sites on 4 continents, along with the latest generation of DTNs driving 100-1000 Gbps flows
- A new generation SDN Framework with unified multilevel control plane programming functionality + substantially extended ODL Controller
- ExaO: More advanced, high level integration functions with the data management applications (PhEDEx, ASO) of the CMS experiment
- Protocol-agnostic edge-control and core-control services that cooperate with the science program's data management systems to allocate high bandwidth, load balanced, high throughput flows over selected paths
- Novel deep learning and database architectures and methods for rapid training on, and traversal of LHC data, driving high throughput event classification and characterization, using multi-GPU systems backed by high throughput SSD data stores
- A new immersive VR experience: a virtual tour of the CMS experiment at the LHC, including an inside-out exploration of LHC collision data

Caltech Machine Learning Projects for HEP

- **3D Imaging with LCD datasets :** energy regression and particle identification with 2D/3D convolutional neural nets for the future generation calorimeter.
- **Event classification using particle-level information :** use recurrent neural nets and long short term memory to learn the long range correlations in LHC collision events.
- **Charged particle tracking acceleration :** explore deep neural net methods for new ways of connecting the dots of the HL-LHC trackers and beyond.
- **Distributed learning :** accelerate training of deep neural net models over large datasets using Spark or MPI frameworks.
- **Neuromorphic Hardware :** exploit existing neuromorphic systems for online data processing and event selection. Develop new hardware tailored to the characteristics of LHC data

The knowledge gained also will be applied to Network and Global System optimization and problem resolution

OpenDaylight & Caltech/YALE + UNESP 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 (Yale) in 2016:

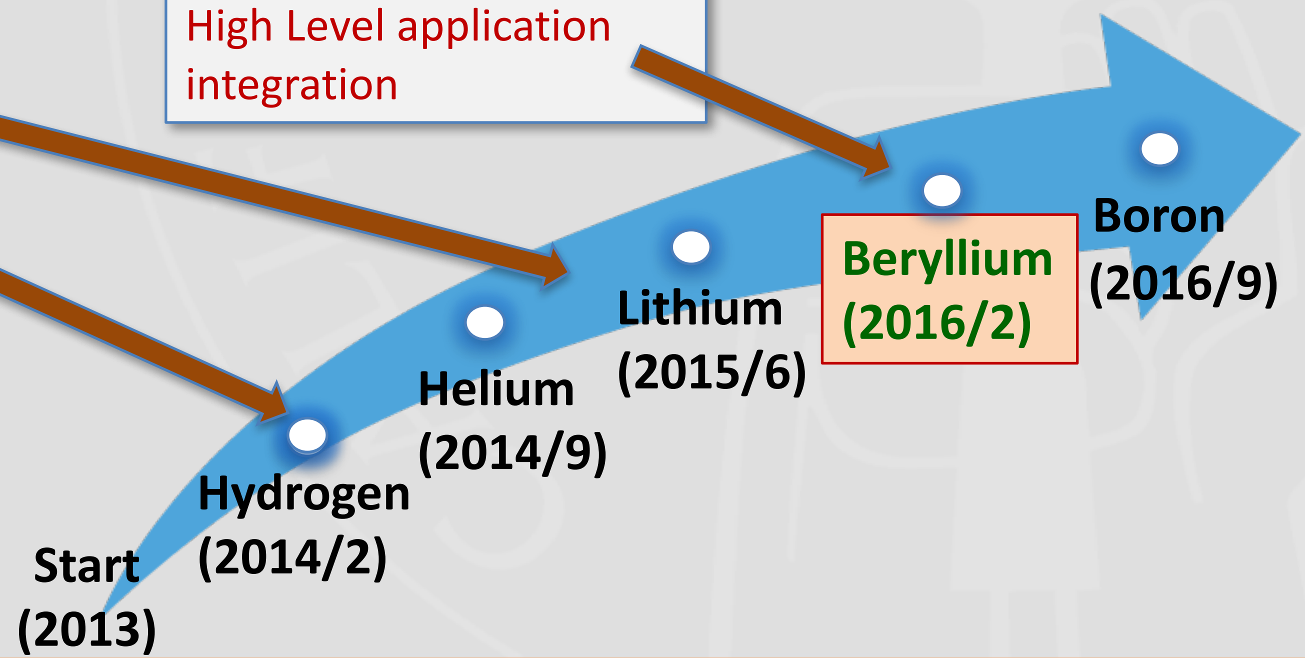
- **Rapid application development platform for OpenDaylight, providing an easy abstraction shielding users/operators from Java/environment build complexities**

OFNG – ODL (w/UNESP):
NB libraries for Helium/Lithium

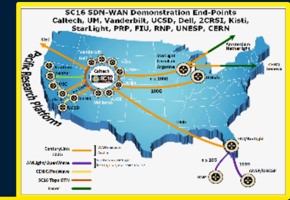
OLiMPs – ODL:
Migrated to Hydrogen

OLiMPs – FloodLight:
Link-layer MultiPath Switching

OFNG - ALTO (w/Yale)
High Level application integration



Yale and Caltech at SC16



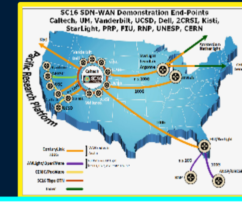
State of the Art SDN Controller + Framework

Driving large load balanced smooth flows over optimally selected paths

See “Traffic Optimization for ExaScale Science Applications”, Q. Xiang et al. IETF Internet Draft https://tools.ietf.org/pdf/draft-xiang-alto-exascale-network-optimization-00.pdf

- We are demonstrating and conducting tutorials at Booths 2437+2537 on our (evolving) state of the art OpenDaylight controller
- Based on a unified control plane programming framework, and novel components and developments, that include:
  - The Application Level Traffic Optimization (ALTO) Protocol
  - A Max-Min fair resource allocation algorithm-set providing flow control and load balancing in the network core
  - A data-driven function store for high-level, change-oblivious SDN programming
  - A data-path oblivious high-level programming framework.
- Smart middleware to interface to SDN-orchestrated data flows over network paths with guaranteed (flow-controlled) bandwidth to a set of DTNs
- Coupled to protocol agnostic (Open vSwitch-based) traffic shaping services at the site edges
- Will be used with Machine Learning to identify key variables controlling the system’s throughput and stability, and for overall system optimization

Yale and Caltech at SC16: State of the Art SDN Framework + ODL Controller



New SDN Framework and Tools : Yale Team

Powerful state of the art, generic tools to substantially simplify SDN programming

Before (manual programming)
- Complex, manual maven programming

→

Web IDE
- Web-based automatic generation of projects
- Programmer focuses only on key aspects

Before (low level programming)
- Low-level, complex OpenFlow rule programming
- Programmer can define only at flow level
- Specific access control allowing only hosts partition

→

Maple programming (high-level programming)
- High-level, completely south-bound agnostic, cross-layer programming
- Programmer sees (logically) each and every packet
- Integrated access control supporting per-user or role based programming

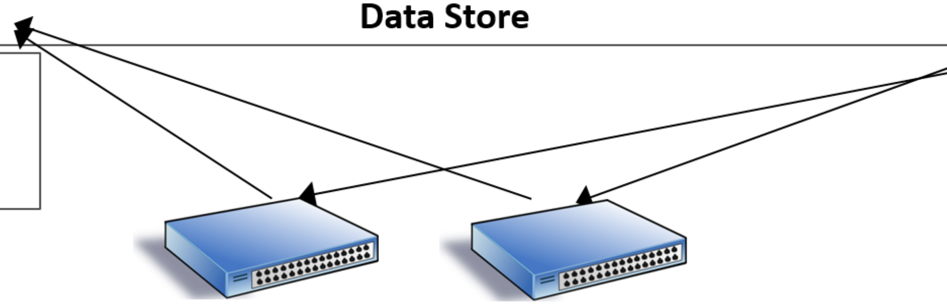
Before (raw data store)
- Complex, manual tracking of execution dependency
- Manual cleanup, re-execute
- Designed directly on raw data store

→

FAST (automated function store)
- Automatic execution dependency tracking
- Automatic cleanup, re-execution (intent ++)
- Can host generic network functions

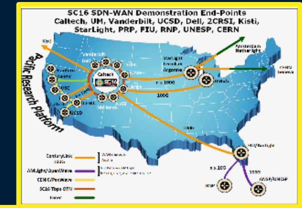
Data Store

Before
- Ad hoc flow rule installation

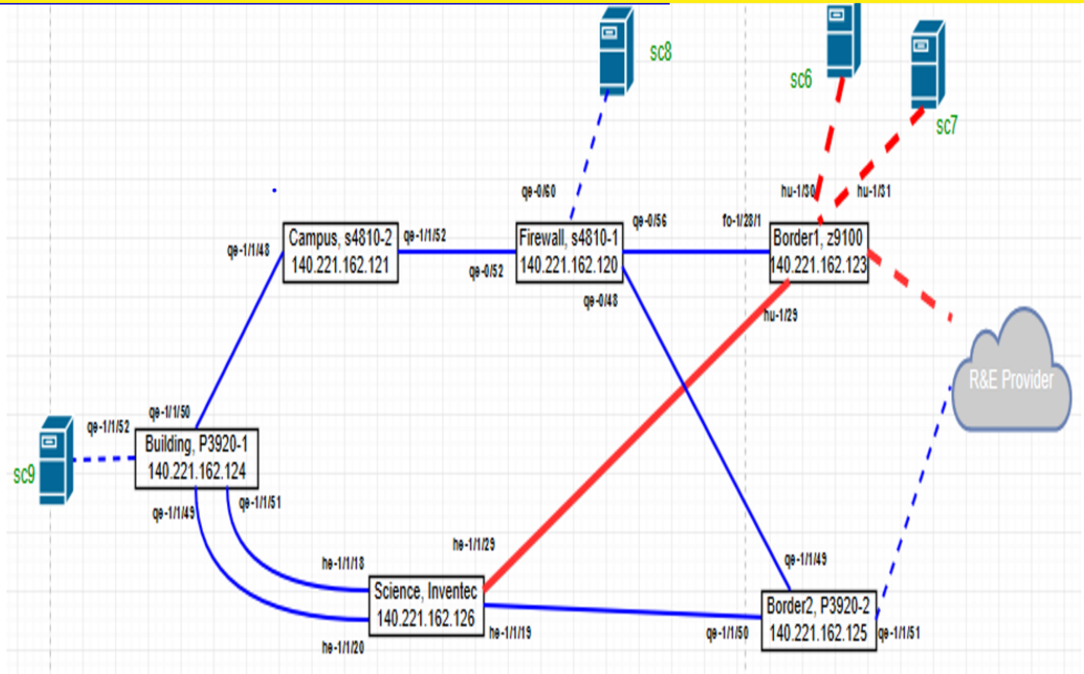


FAST Schedule
- Consistent, optimized flow-mod scheduling

New Tools in Action: Programmable Science DMZ with Maple + FAST



- **Flexible, stateful firewall programmed using Bro**
  - Up to Layer 7 detection
- Generic FW state update to SDN controller using RESTCONF
- **SDN control programming using Maple programming,** executed in FAST function store
- **Achievements**: only 10s of lines code, for a fully adaptive, highly extensive ScienceDMZ



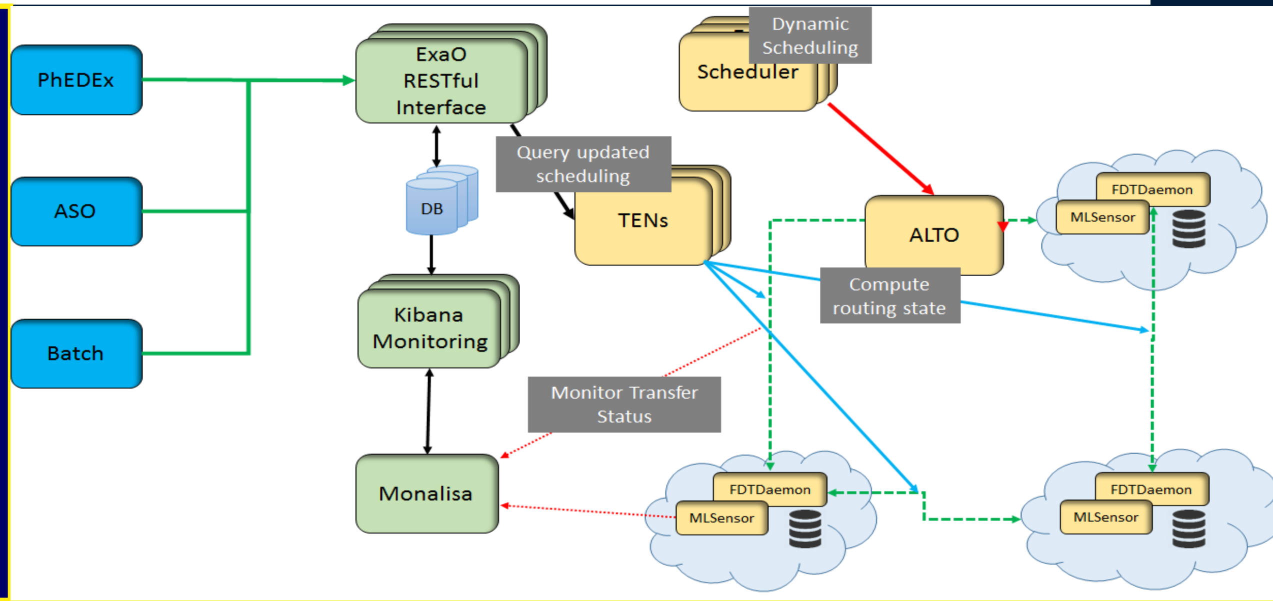
Future:

- **Extensible** to complex, highly stateful, and/or policy driven decisions
- **Enabling** new levels of functionality, handling new levels of complexity

Demonstrations and Tutorials by the Yale Team at the Caltech Booths: 2437 and 2537

CMS at SC16: *ExaO* - Software Defined Data Transfer Orchestrator with Phedex and ASO

Leverage emerging SDN techniques to realize end-to-end orchestration of data flows involving multiple host groups in different domains



- ❐ Maximal link utilization with ExaO:
  - PhEDEx: CMS data placement tool for datasets
  - ASO: Stageout of output files from CMS Analysis Jobs
- ❐ Tests across the SC16 Floor: Caltech, UMich, Dell booths and Out Over the Wide Area: FIU, Caltech, CERN, UMich
- ❐ Dynamic scheduling of PetaByte transfers to multiple destinations

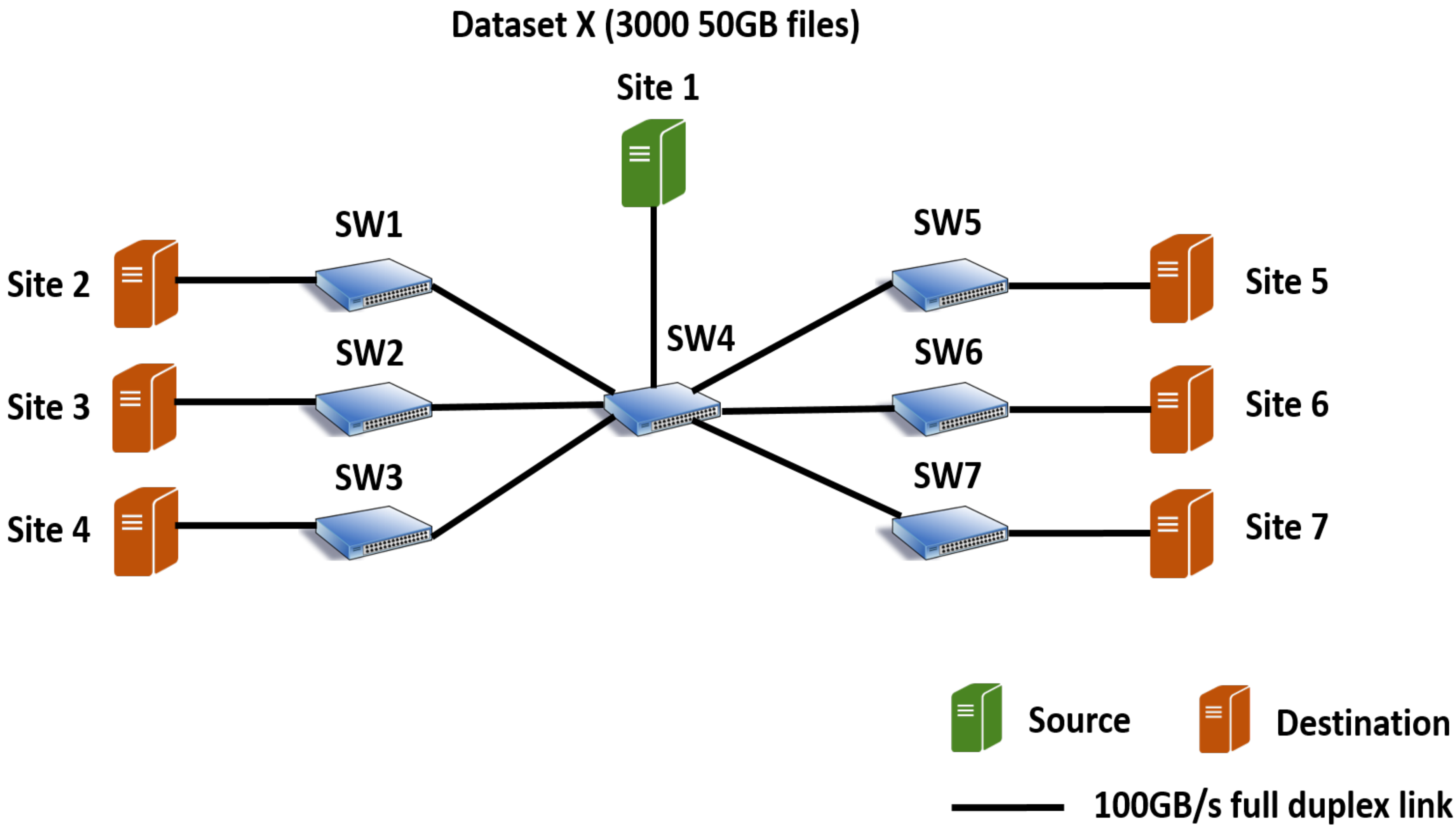
Partners: UMich, StarLight, PRP, UNESP, Vanderbilt, NERSC/LBL, Stanford, CERN; ESnet, Internet2, CENIC, MiLR, AmLight, RNP, ANSP

ExaO: Software Defined Data Transfer Orchestrator

PhEDEx	ExaO
• No real-time, global network view	**Application-Layer Traffic Optimization (ALTO)** • Collect real-time routing information at different domains (ALTO-SPCE) • Compute minimal, equivalent abstract routing state (ATLO-RSA)
• Dataset level scheduling • Destination sites cannot become candidate sources until receiving the whole dataset • Low concurrency	**Scheduler** • Centralized file level scheduling • Destination sites become candidate sources after receiving files • High concurrency
• No network resource allocation scheme • Low utilization	**Scheduler and Transfer Execution Nodes (TEN)** • Global, dynamic rate allocation among transfers (Scheduler) • End host rate limiting to enforce allocation (TEN)

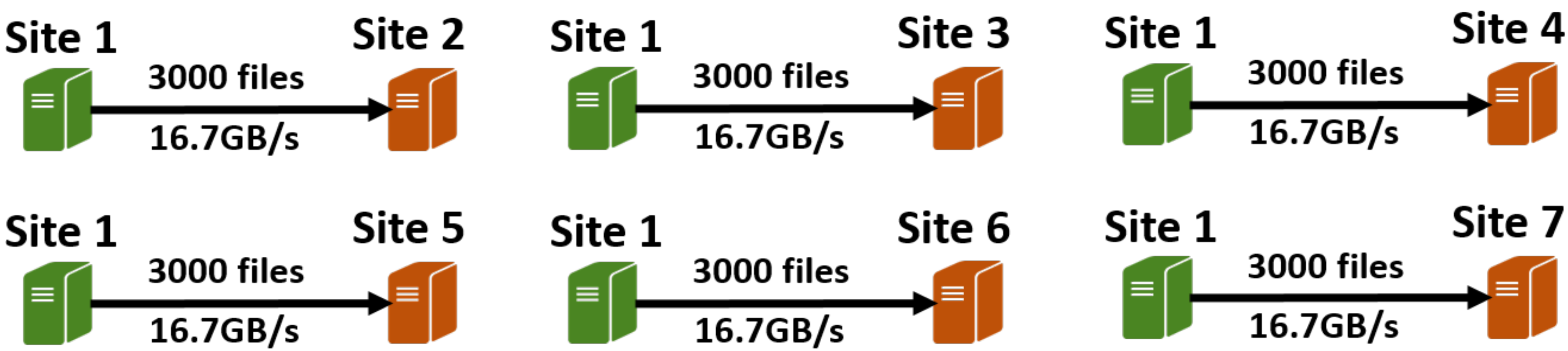
A Major Application of the New SDN Maple+Fast Framework By the Yale Team and Caltech to CMS Data Operations

Case Study: Distribution Dataset X to All the Sites



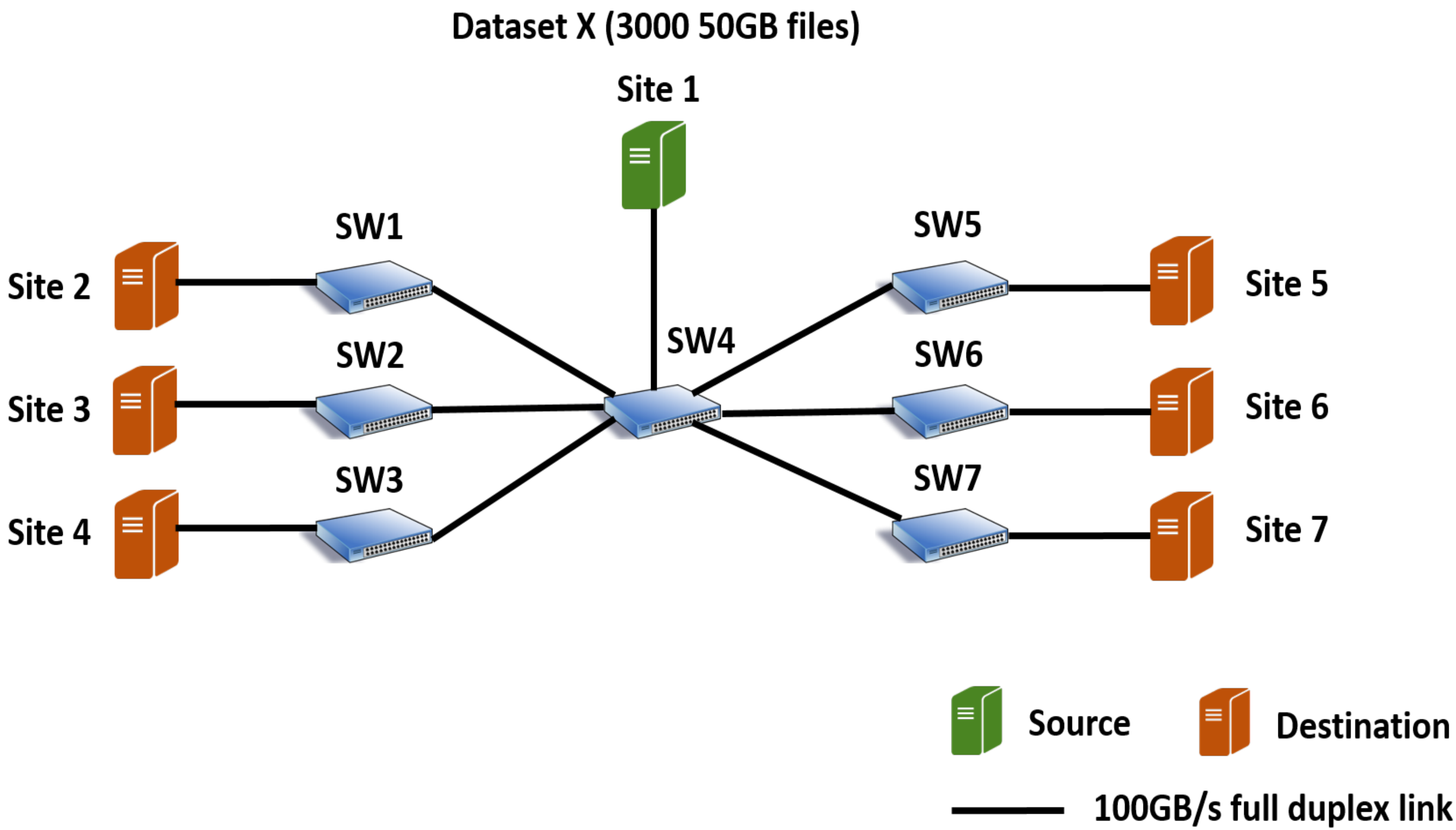
Scheduling Policy Computed by PhEDEx

- Only Site 1 can be the source
- Site 1 sends all 3000 files to each destination site
  - Scheduling decision: (File K; site 1 to X), where K=1..3000, X=2..7
- Leaves the bandwidth allocation to TCP
  - Fair share of each site-to-site flow converges at 100/6=16.7Gbps



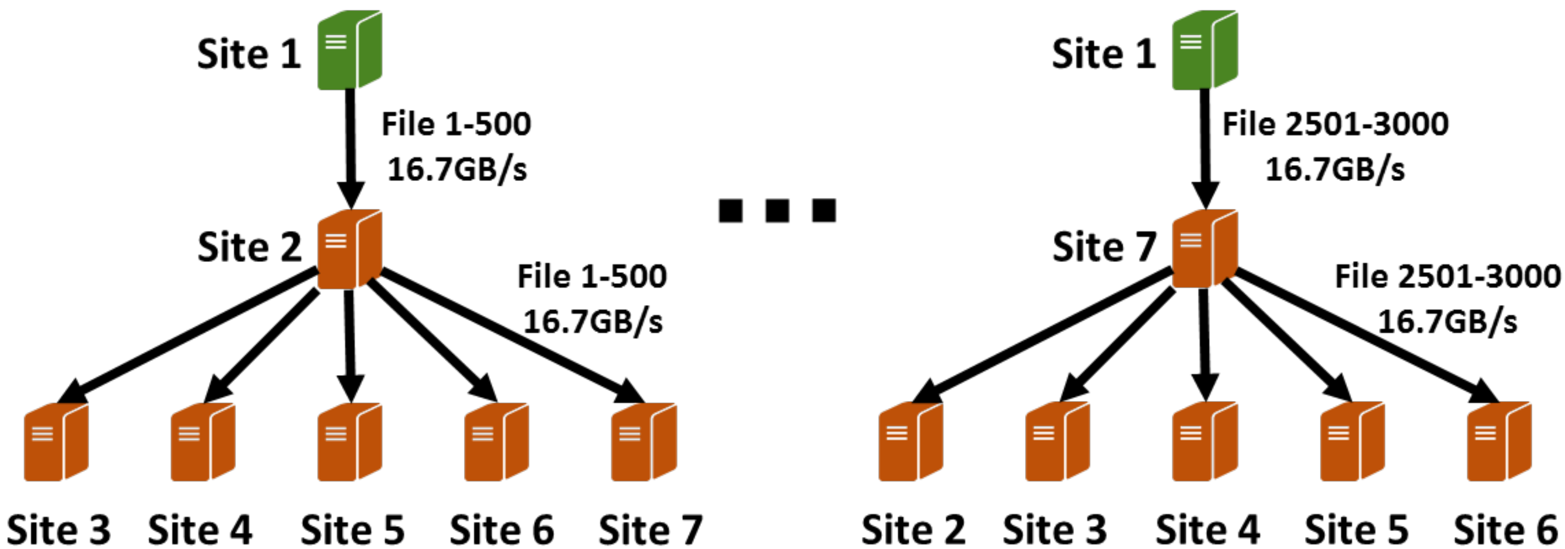
Link Utilization: $\frac{1}{7} \simeq 14\ \%$

Case Revisited: Distribute Dataset X to All Sites



Example: Scheduling Policy Made by ExaO

- Site 1 is the only source at the beginning
- Each site can become a source once receiving certain files
- Site 1 sends 3000/6=500 unique files to each destination site
  - Fair share of each (site 1, site X) flow is 100/6=16.7GB/s
  - Remaining uplink bandwidth of site 1 is 0GB/s
- After receiving a unique file from site 1, site X becomes a source to the other six destination sites
- Site X sends the received file to other destination sites at (100-16.7)/5=16.7GB/s



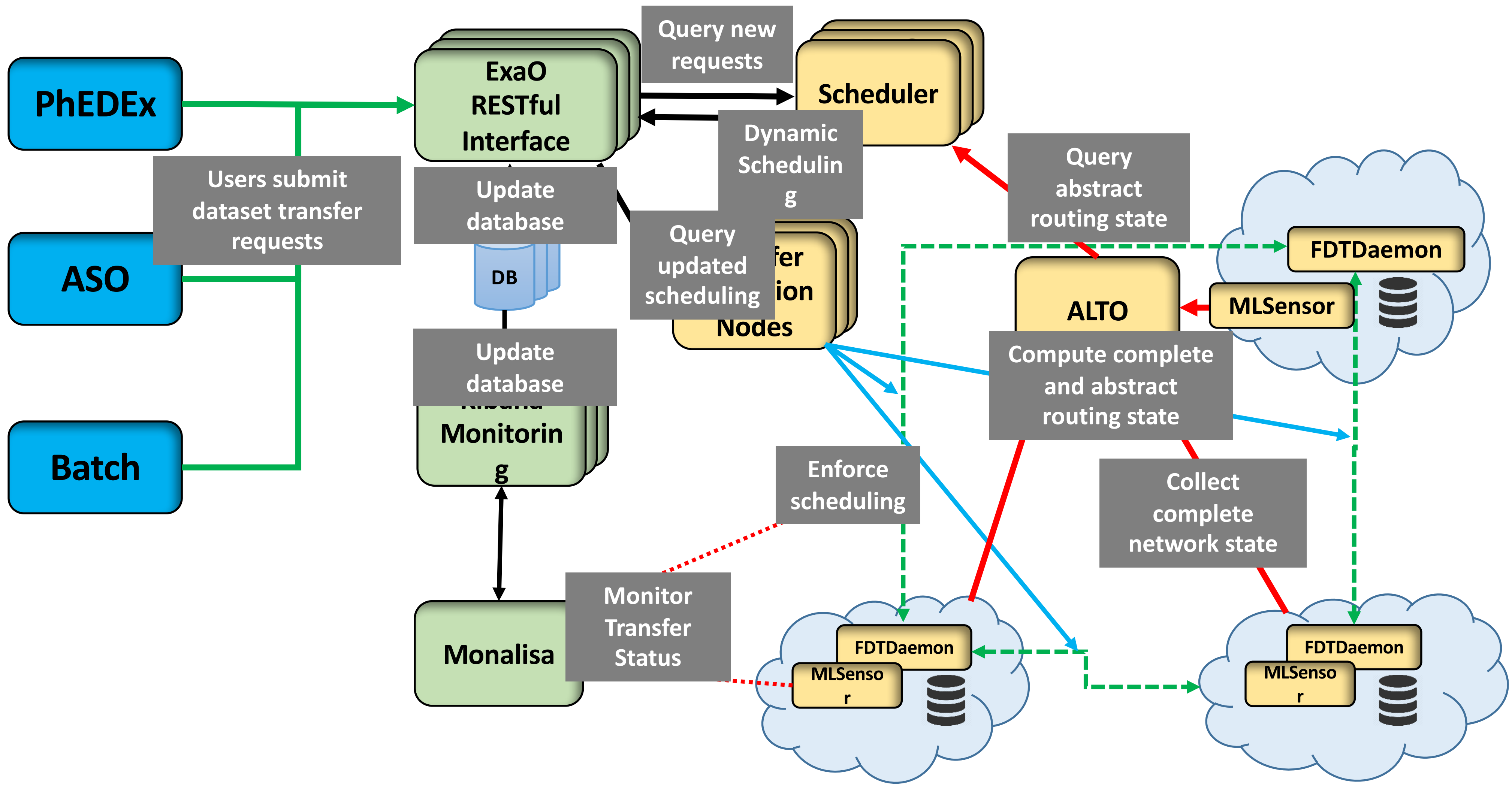
Link Utilization:
$$\frac{6}{7} \simeq 86\ \%$$
(Maximum)

Generalizable to M Source Sites, N Destination Sites, in P Stages With Strategic, Network State and Policy-Sensitive Decisions

Components of ExaO

- **RESTful-API:** allow users submit and manage transfer request through different interfaces
- **ALTO:** collect on-demand, real-time, minimal abstract routing information from different domains
- **ExaO Scheduler:** centralized, efficient file-level scheduling and network resource allocation
- **FDT:** efficient data transfer tools on the end hosts
- **Monalisa:** Distributed monitoring infrastructure for real time monitoring of each flow, transfer

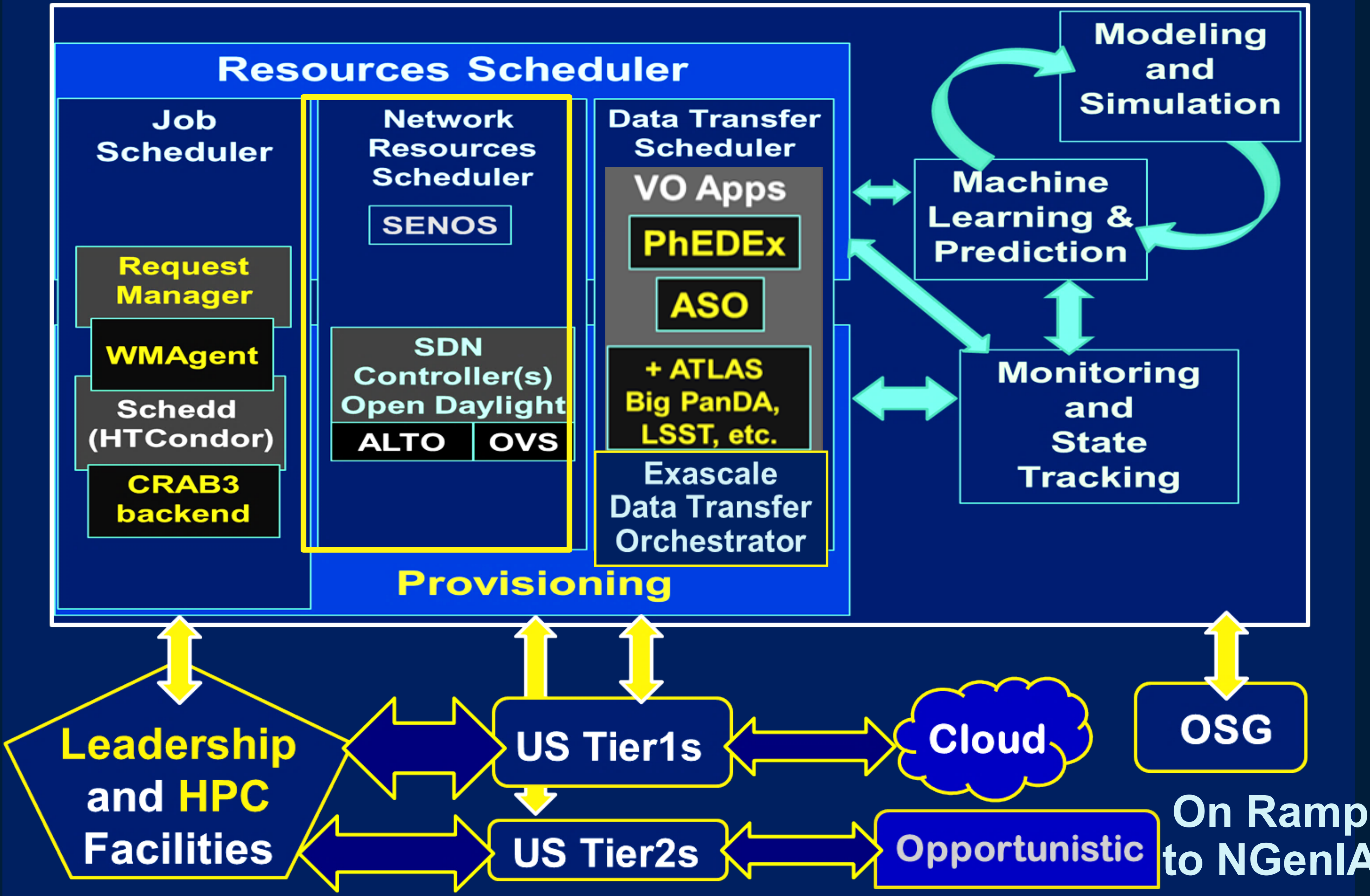
ExaO: Software Defined Data Transfer Orchestrator



Design: Addressing Practical Concerns

- Minimally invasive change on end host groups
- Real-time, dynamic resource allocation under the existence of other network traffic
- Not CMS or HEP specific, hence support any data intensive sciences.
- Dataset distribution to N destination:
  - Maximal link utilization in the testbed
  - N times faster than dataset level scheduling

NGenIA-ES Services and Data Flow Diagram



NGenIA
New SDN Paradigm
ExaO LHC rchestrator
Tbps Complex Flows
Machine Learning
LHC Data Traversal
Immersive VR

Thanks to Ecostreams,
Orange Labs Silcon Valley

Visit Us at the Caltech Booths 2437, 2537
+ the Starlight Booth 2611

Collaboration Partners

CALTECH HEP NETWORKING

Special thanks to …

Research Partners

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

Carrier and R&E Net Partners

- Century Link
- Zayo
- CENIC
- PacWave
- FLR
- MiLR
- Wilcon

Industry Partners

- 2CRSI (NVME Storage, Servers)
- Arista (OpenFlow Switches)
- Ciena (Tbps DCIs, Optics)
- Coriant (Optics/Mux)
- Color Chip (Optics)
- Arista (OpenFlow Switches)
- Dell (OpenFlow Switches; Server systems)
- Echostreams (Server systems)
- Inventec (OpenFlow Switch)
- HGST (NVME SSDs and SAS Disk Arrays)
- Infinera (Optical Interconnects)
- Intel (NVME SSD Drives)
- LIQID (NVME SSD Systems)
- Mellanox (NICs and Cables)
- Orange Labs Silicon Valley (GPUs and Servers)
- Qlogic (NICs)
- Chelsio (NICs)
- Samsung (NVME SSDs)
- Spirent (100GE Tester)
- Supermicro (Servers for GPUs)

200G Dual Socket 2CRSI / SuperMicro 2U NVMe Servers

CALTECH HEP NETWORKING

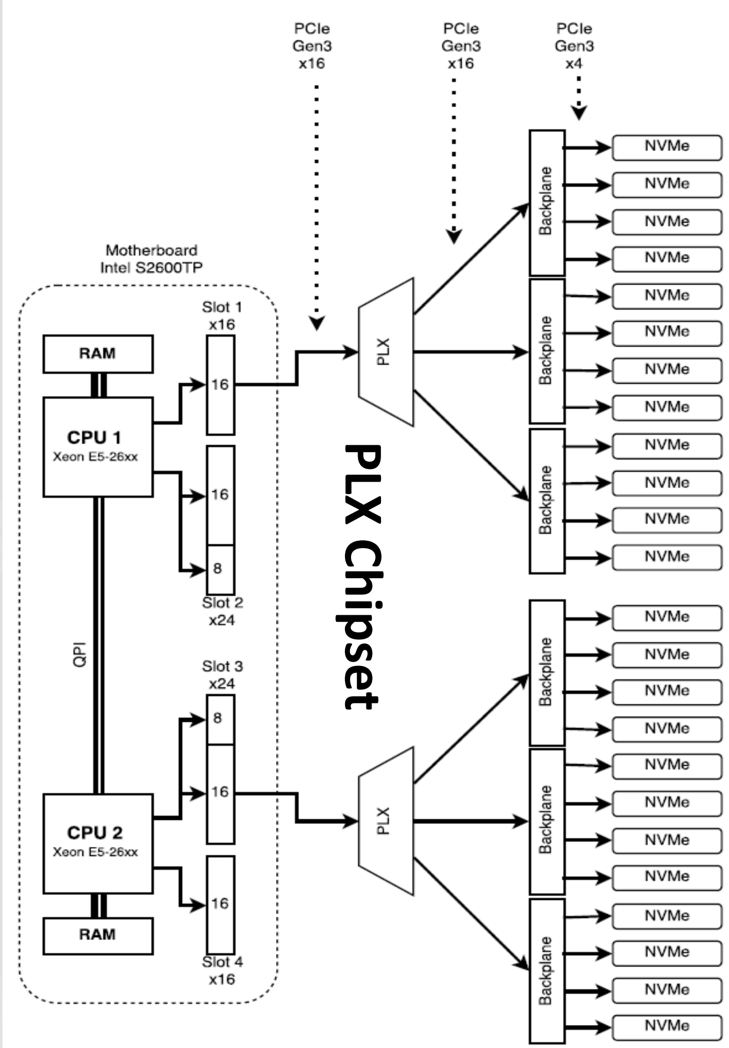
- Both servers are capable to drive 24 x 2.5” NVMe drives. SuperMicro also have a 48 drive version.
- M.2 to U.2 adaptors can be used to host M.2 NVME drives

2CRSI

SuperMicro

2.5” NVMe Drive

PCIe Switching Chipset for NVMe



PCIe Lanes on CPUs are a Major Constraint

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

CALTECH HEP NETWORKING

Server Readiness:

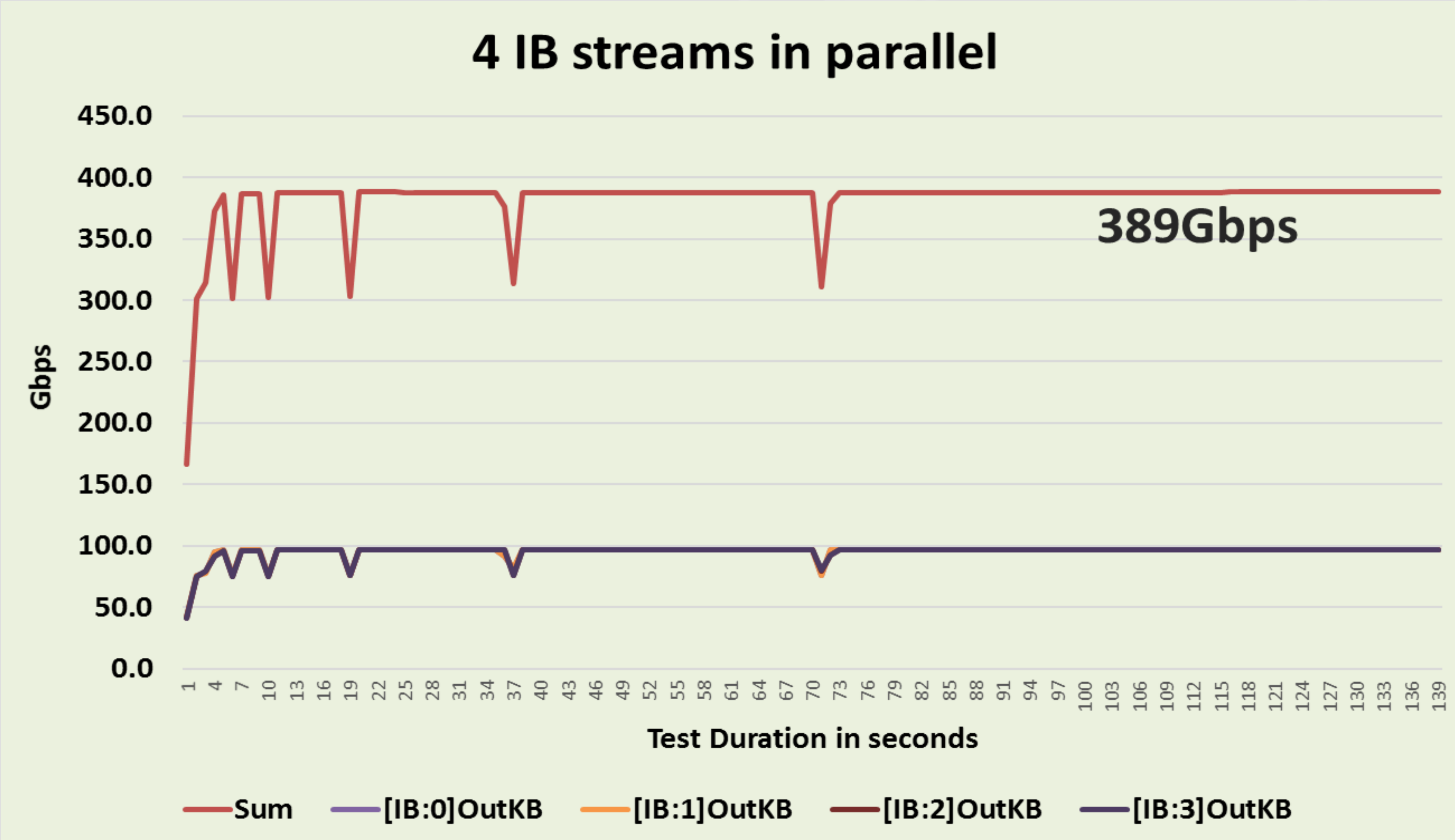
1) Current PCIe Bus limitations
   - PCIe Gen 3.0 (x16 can reach 128Gbs Full Duplex)
   - PCIe Gen 4.0 (x16 can reach double the capacity, i.e. 256Gbps
   - PCIe Gen 4.0 (x32 can reach double the capacity, i.e. 512Gbps
2) Increased number of PCIe lanes within processor

   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

400GE Network Testing: IB Reduces Overheads

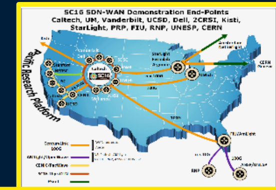
CALTECH HEP NETWORKING



Transmission across 4 Mellnox VPI NICs.

Only 4 CPU cores are used out of 24 cores.

Caltech at SC16 Booths 2437+2537



Acknowledgements

A key factor in the progress and success of the work presented here has been the support and engagement of the DOE Offices of Advanced Scientific Computing and High Energy Physics, and the NSF Directorate for Computer & Information Science and Engineering (CISE).

We thank the agencies for the following grants which supported much of this work:

Caltech Team:

SDN NGenIA, DOE/ASCR Project ID 000219898
SENSE, FNAL PO #626507 under DOE award # DE-AC02-07CH11359
ANSE – NSF award # 1246133
CHOPIN – NSF award # 1341024
CISCO – Award # 2014-128271
Tier2 – NSF award # 1120138
OLIMPS - DOE award # DE-SC0007346 (through 2014)
US LHCNet - DOE # DE-AC02-07CH11359 (through 2015)

Sao Paulo UNESP Team:

São Paulo Research Foundation (FAPESP) under Grant # 2013/01907-0

Huawei do Brasil Telecomunicações Ltda. under Grant Fundunesp # 2481-2015

Yale Team:

NSF: under award# CC*IIE-1440745, award# CNS-1218457

Google: under Google Faculty Research Award (2015)

Huawei: under Huawei Research Award (2014-2015)

A Next Generation Terabit/sec SDN Architecture and Data Intensive Applications for High Energy Physics and Exascale Science

*H. Newman, M. Spiropulu, J. Balcas, T. Hendricks, D. Kcira, I. Legrand, A. Mughal, J. R. Vlimant, R. Voicu***, High Energy Physics, California Institute of Technology
1200 East California Blvd, Pasadena, CA 91125

*S. Novaes, A. Baruchi, R. Iope, B. Leal***, UNESP Center for Scientific Computing
271 R. Dr. Bento Teobaldo Ferraz, São Paulo, Brazil, CEP 01140-070

*K. Gao, M. Wang, Q. Xiang, Y.R. Yang, J. Zhang***, Computer Science, Tongji-Yale Systems Networking Center, Yale University/Tongji University
51 Prospect Street, New Haven, CT 06612

Visit Us at the Caltech Booths 2437, 2537 + the Starlight Booth 2611

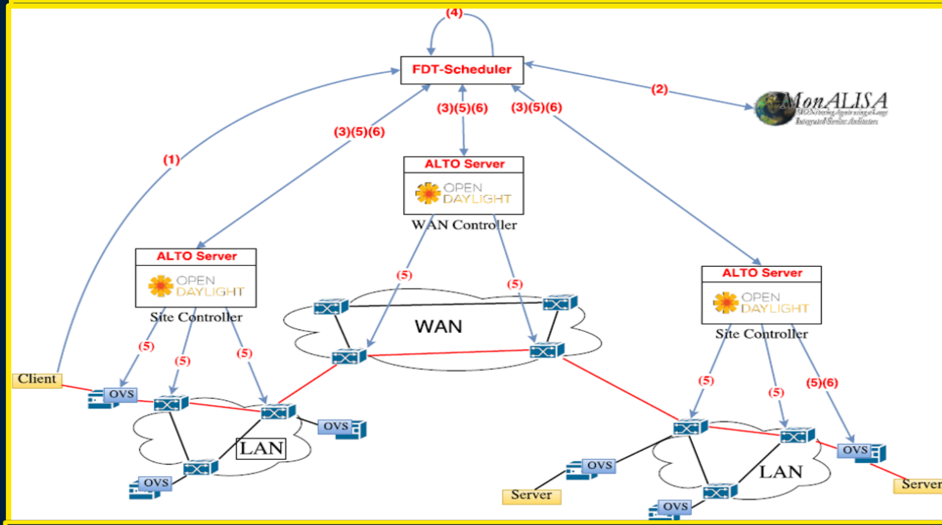
Next Generation “Consistent Operations”

Site-Core Interactions for Efficient, Predictable Workflow

- Key Components: (1) Open vSwitch (OVS) at edges to stably limit flows, (2) Application Level Traffic Optimization (ALTO) in Open Daylight, Maple and Fast for end-to-end optimal path creation/selection + flow metering and high watermarks set in the core
- Flow metering in network fed back to OVS edge instances: to ensure smooth progress of end-to-end flows
- Real-time flow adjustments triggered as below
- Optimization using “Min-Max Fair Resource Allocation” (MFRA) algorithms on prioritized flows

Demos: Internet2 Global Summit in May and at SC16 Booths 2437, 2537 this week

Consistent Ops with ALTO, OVS and MonALISA FDT Schedulers

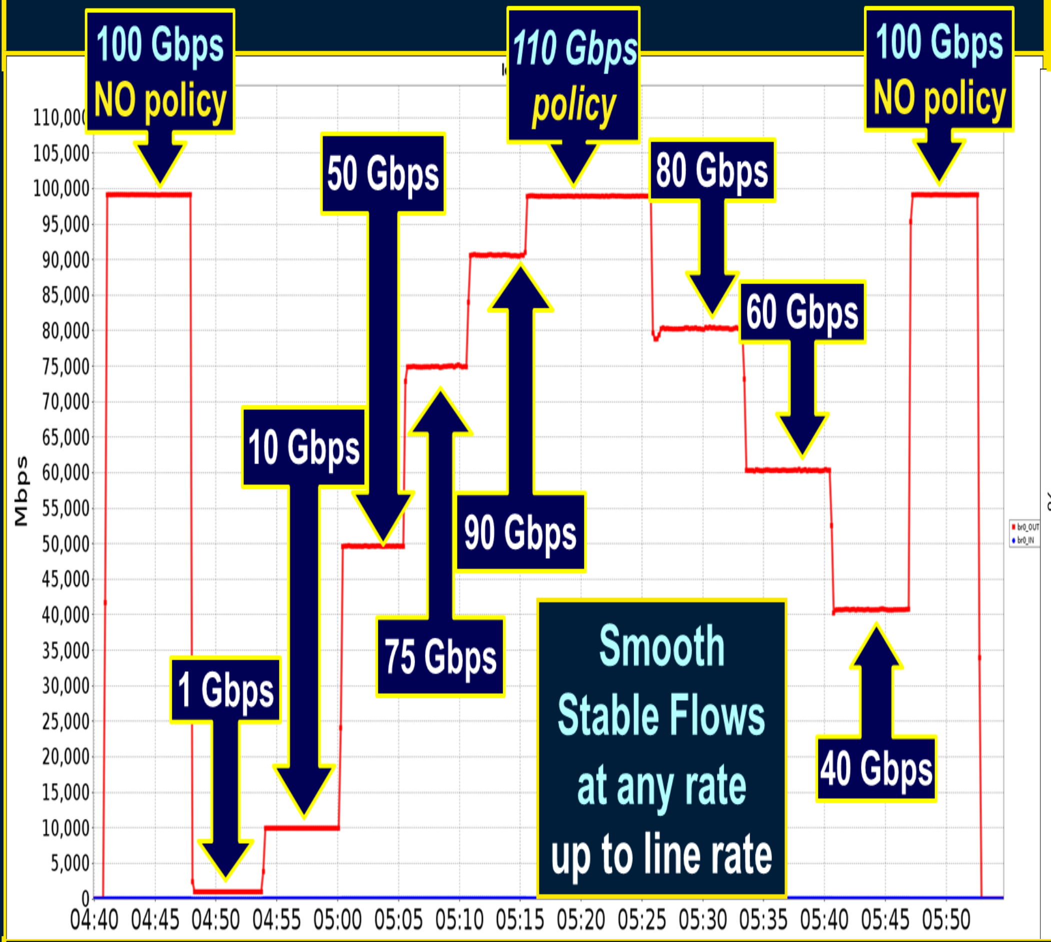


- Real-time adjustment of allocations triggered by: (1) new requests, (2) real-time feedback on progress of transfers, (3) network state changes or error conditions, (4) proactive load-balancing operations, or (5) rate-limiting operations imposed by controllers or emerging network operating systems (e.g. SENOS)

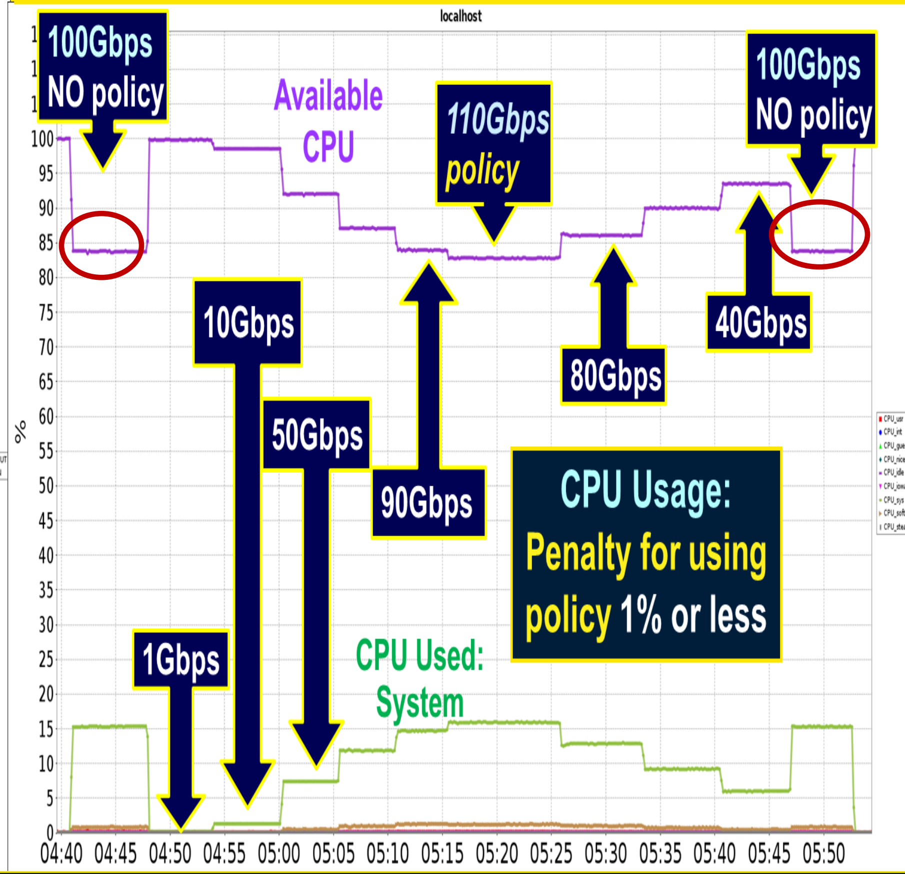
Yale CS Team: Y. Yang, Q. Xiang et al.

OVS Dynamic Bandwidth
100G Rate Limit Tests

RATES



CPU Utilization:
1 Core 16% at full 100G



CPU Usage: Penalty for exerting policy: 1% or less

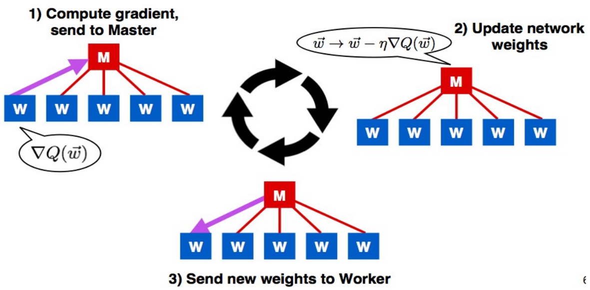
Machine Learning for the LHC Physics Program: Mission Statement

- LHC Data Processing may **use deep learning methods in many aspects** (attend other relevant talks at the Caltech booth)
- **Large volume of collision data** and simulated data to analyze
- Several classes of **LHC Data analysis make use of classifiers** for signal versus background discrimination.
  - ✓ Use of BDT on high level features
  - ✓ Increasing use of MLP-like deep neural net
- Deep learning has delivered **super-human performance** at certain class of tasks (computer vision, speech recognition, ...)
  - ✓ Use of convolutional neural net, recurrent topologies, long-short-term-memory cells, ...
- Deep learning has the advantage of **training on "raw" data**
  - ➢ Several levels of data distillation in LHC data processing

→ Going beyond fully connected networks with advanced deep neural net topologies

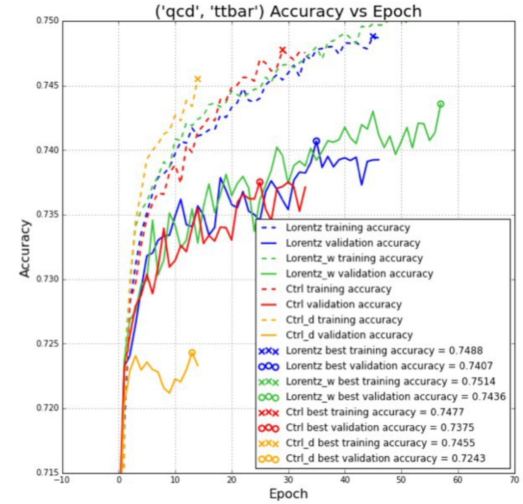
- ➢ **Multi-classification of LHC events from particle-level information**
- ➢ **Charged particle tracking with recurrent and convolutional topologies**
- ➢ **Particle identification and energy regression in the highly granular future CMS calorimeter**

Exploit Supercomputer
Piz Daint @ CSCS



Distributed Training
with MPI or Spark

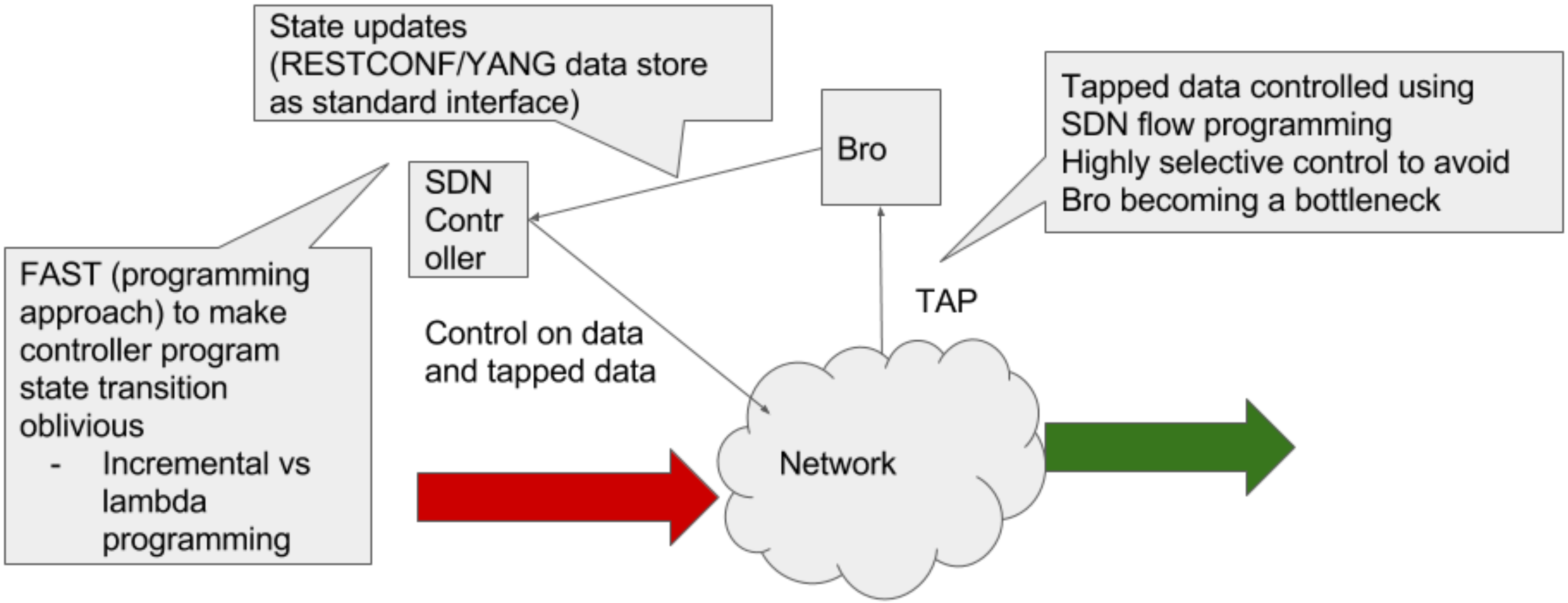
Exploit Local Server
Supermicro @ Caltech



Explore Models Topologies
and Performance

Exploit Supercomputer
Cooley @ ANL

Programmable DMZ using FAST Maple

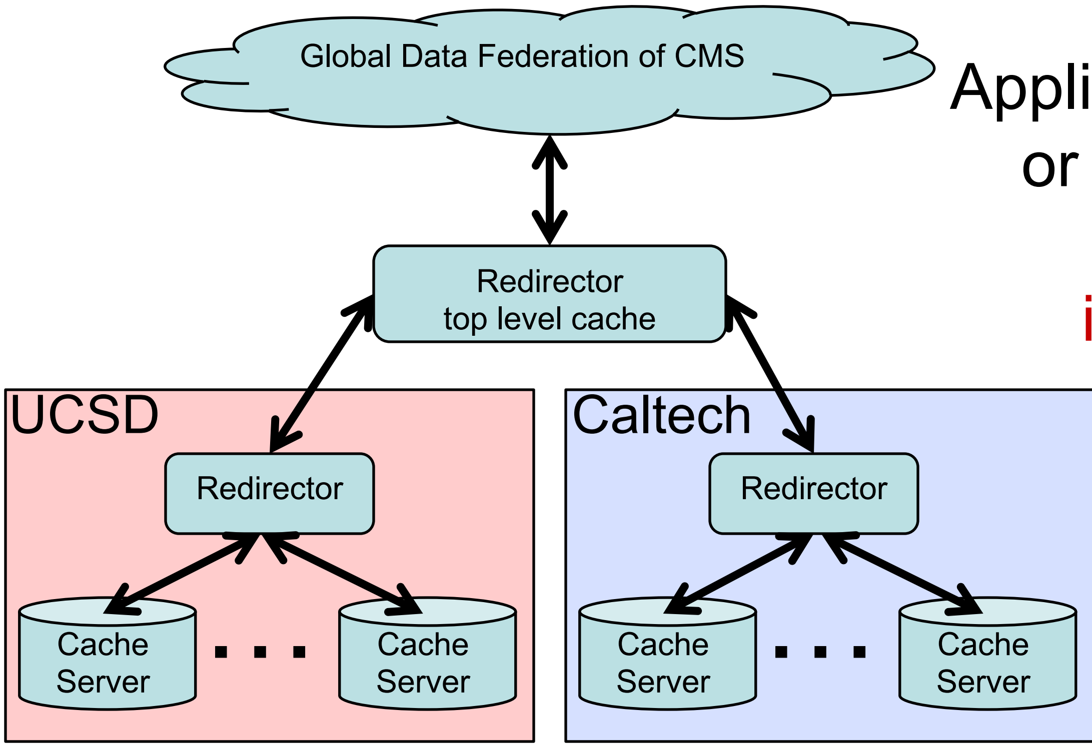


Flexible, generic mechanisms to program the complete system

Three control points

1. What to tap and how to control it ? ***SDN programming model***
2. How to notify from Middlebox to the SDN controller? (RESTCONF to modify SDN data store)
3. How to adaptively control the tapping and the original data?
4. How to program the Middlebox to achieve flexible programing (Bro's programming model)

A Distributed XRootd Cache



Applications can connect at local or top level cache redirector.

⇒ Test the system as individual or joint cache.

Provisioned test systems:

UCSD: 10 x 12 SATA disk of 2TB
@ 10Gbps for each system.

Caltech: 30 SATA disk of 6TB
14 SSD of 512GB
@ 2x40Gbps per system

Production Goal:
Distributed cache that sustains 10k clients reading simultaneously from cache at up to 1MB/s/client without loss of ops robustness.

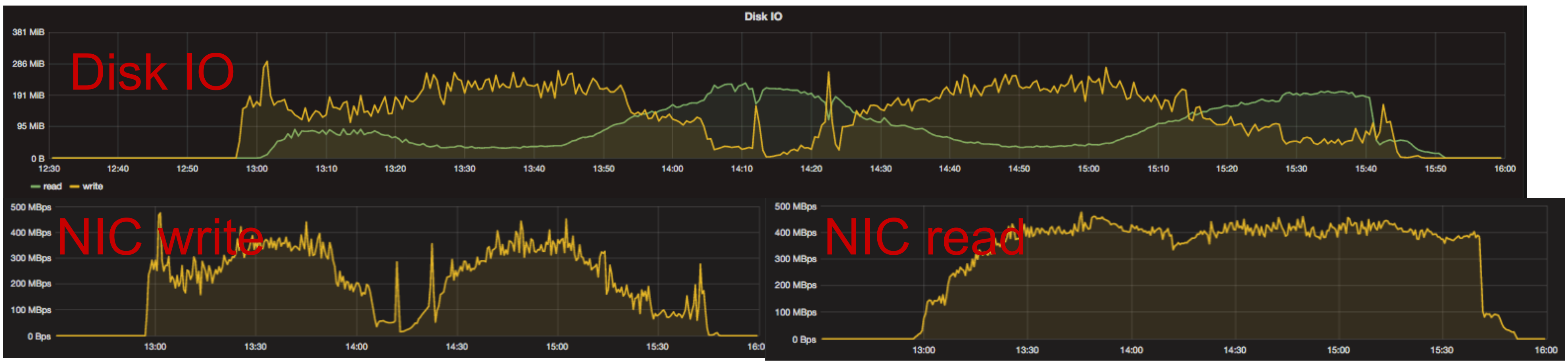
Open Science Grid

Initial Performance Tests

Write as measured at NIC | Read as measured at NIC

Up to 5000 clients reading from 108 SATA disks across 9 servers

Focusing on just one of the servers:



NIC write/read >> Disk IO **=>** by design cache does not always involve disk when load gets high

Robust serving of clients
more important than cache hits

Open Science Grid