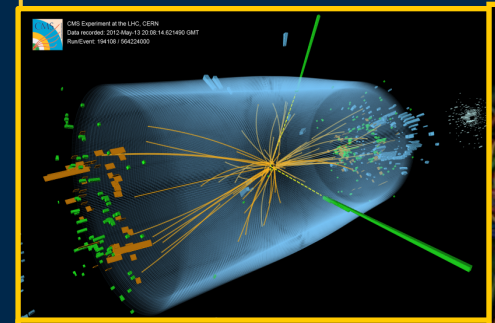
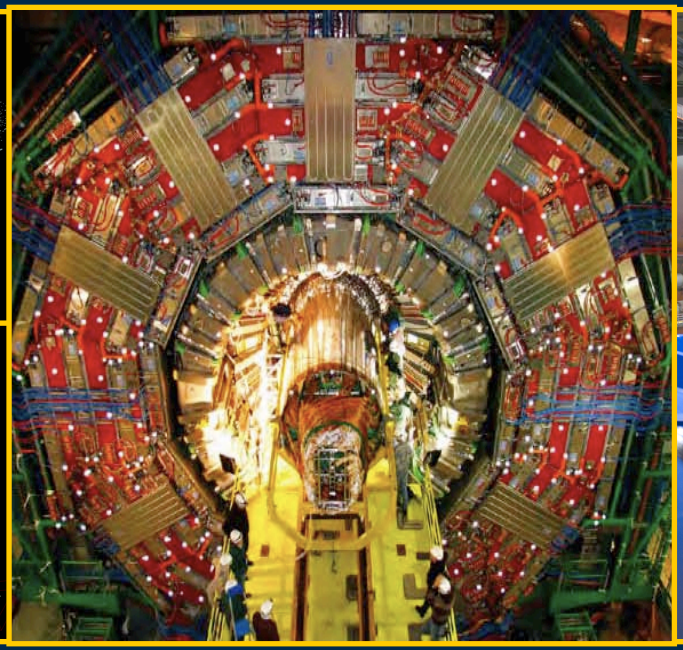
SDN Driven Terabit/sec Workflows for HEP and Data Intensive Science





- ***LHC Run1:*** *Discovery of a New Boson*
- ***LHC Run2:*** *Beyond the Standard Model*

Gateway to a New Era

H. Newman on Behalf of the Caltech and Partner HEP and Network Teams

INDIS Workshop at SC15, Austin

November 16, 2015

Partners: *UMich, Dell, StarLight, Vanderbilt, Stanford, FIU, SPRACE, Echostreams, PRP, ESnet, Internet2, …*

Discovery of a Higgs Boson
July 4, 2012; Nobel Prize 2013-

Physicists Find Elusive Particle Seen as Key to Universe

The New York Times

Englert | Higgs

2013

Theory : 1964

LHC + Experiments Concept: 1984

Construction: 2001

Operation and Discovery: 2009-12

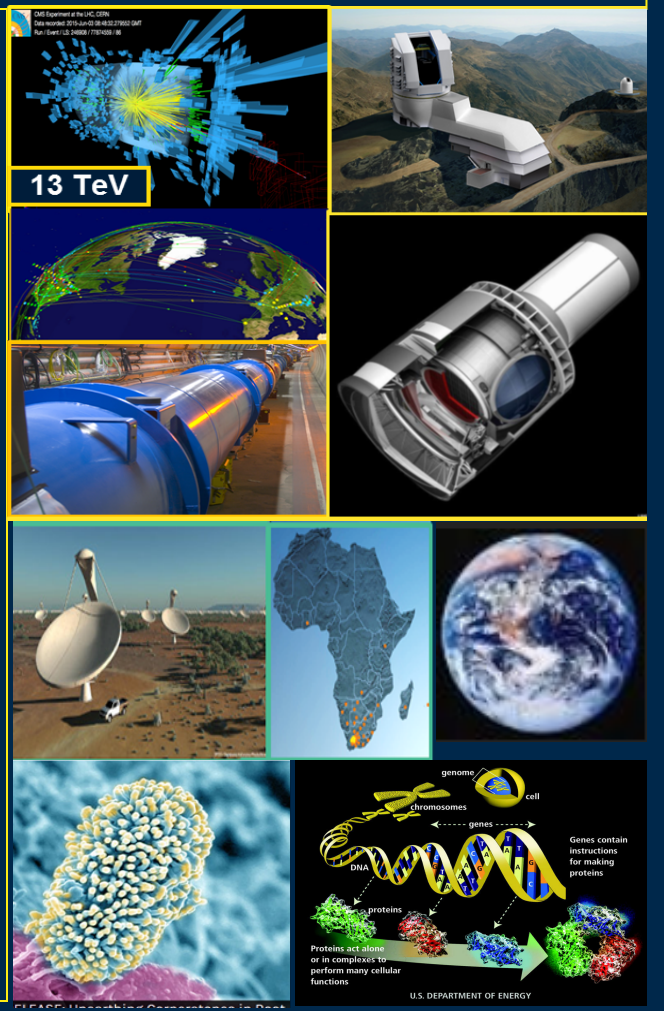
Highly Reliable Advanced Networks Were Essential to the Higgs Discovery and Every Ph.D Thesis of the last 20+ Years

They will be Essential to Future Discoveries, and Every Ph. D Thesis to Come

Entering a new Era of Exploration and Discovery in Data Intensive Sciences

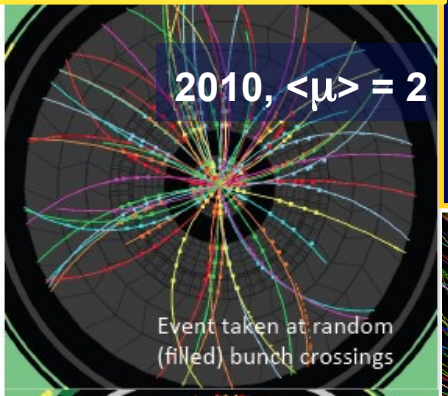
- We are entering a new era of exploration and discovery
  - In many data intensive fields, from HEP and astrophysics to climate science, genomics, seismology and biomedical research

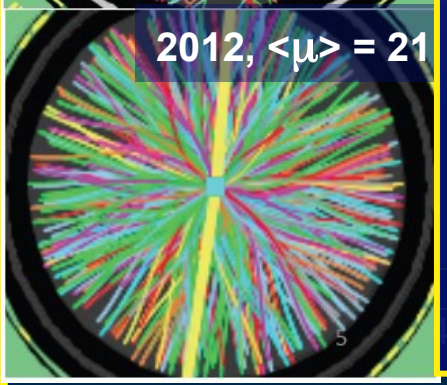
- The largest data- and network-intensive programs from the LHC and HL LHC, to LSST and DESI, LCLS II, the Joint Genome Institute and other emerging areas of growth face unprecedented challenges
  - In global data distribution, processing, access and analysis
  - In the coordinated use of massive but still limited CPU, storage and network resources.
- High-performance networking is a key enabling technology for this research: global science collaborations depend on fast and reliable data transfers and access on regional, national and international scales



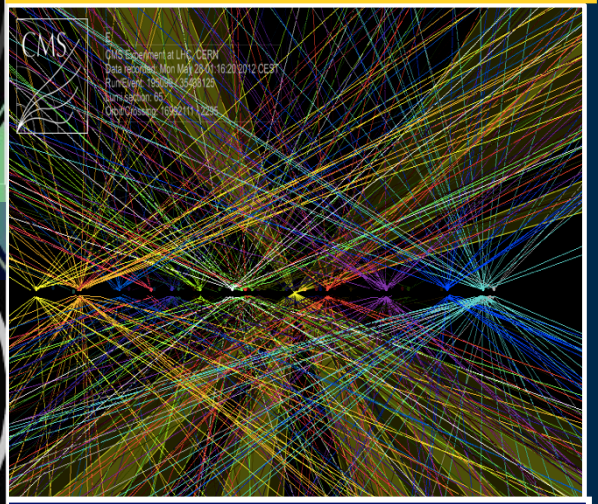
The LHC: Spectacular Performance

Data Complexity: The Challenge of Pileup





~3.5 X 10^{15} pp Collisions
1M Higgs Bosons created in Run 1



~50 Vertices, 14 Jets, 2 TeV

Run2 and Beyond will bring:

- **Higher energy and intensity**
- **Greater science opportunity**
- **Greater data volume & complexity**
- **A new Realm of Challenges**

Average Pileup

Run 1 21
Run 2 42
Run 3 53
HL LHC 140-200

ATLAS Data Flow by Region: 2009-2014

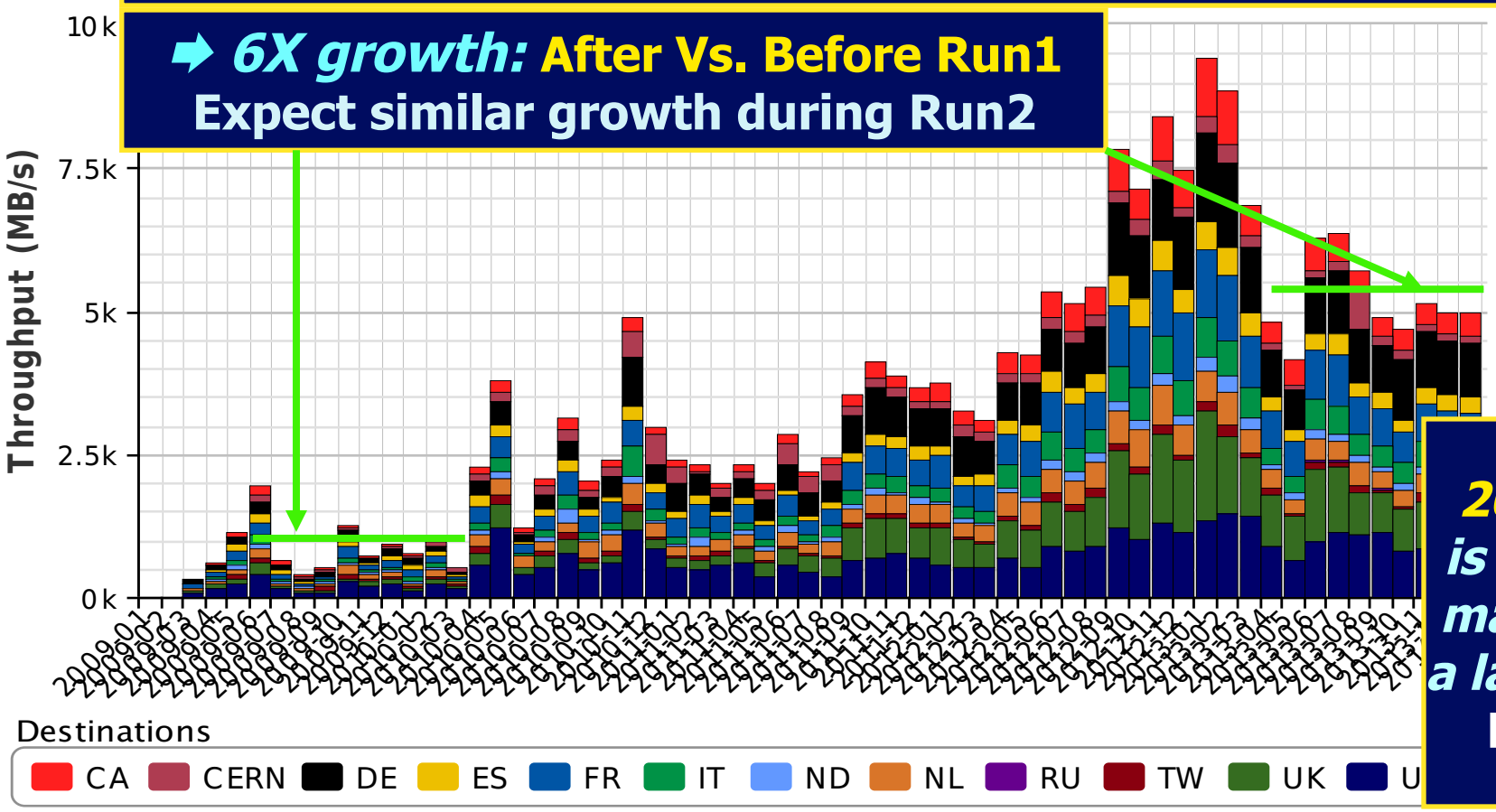
Excellent Grid and Network performance was crucial for the Higgs boson discovery.

171 Petabytes Transferred During 2012

~6 GBytes/sec Average, 14 GBytes/sec Peak

➡ ***6X growth:*** **After Vs. Before Run1**

Expect similar growth during Run2



2012 Vs 2011: +70% Avg; +180% Peak

2014:"10G is becoming marginal for a large Tier2"
R. Mount

Scale of LHC Network Requirements for LHC Run 2: Challenges Ahead

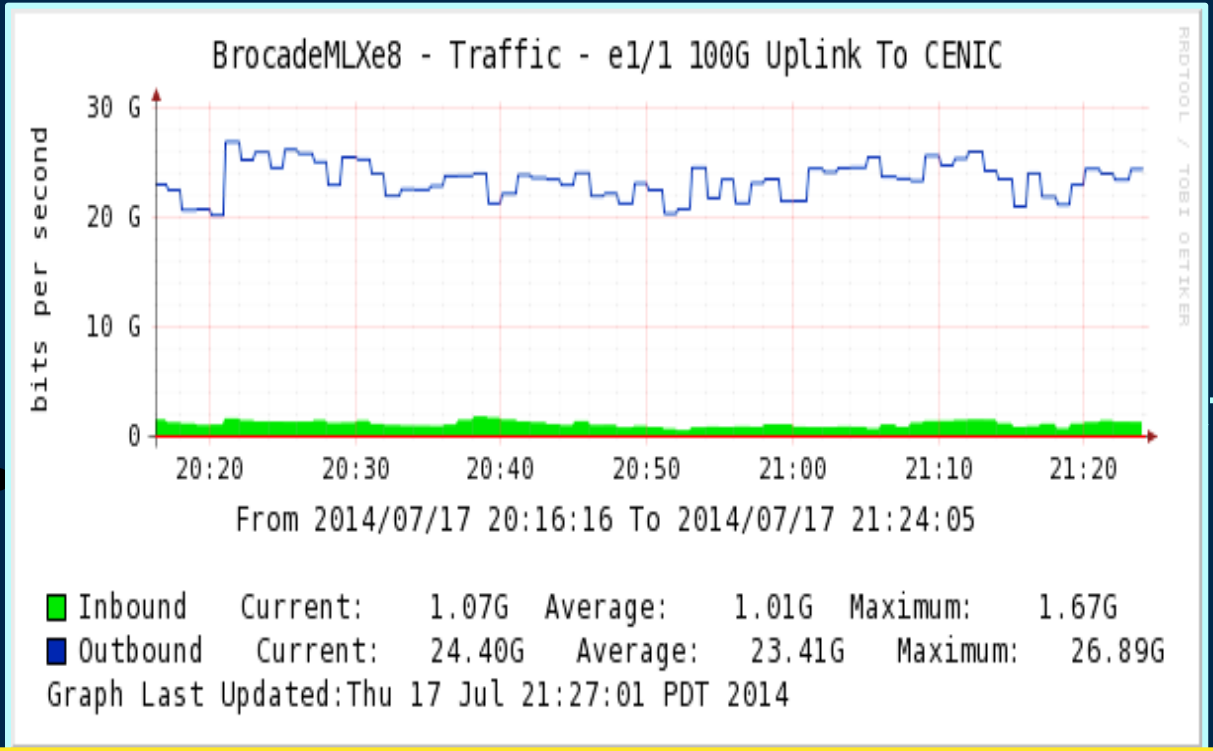
- CMS at 2013 ESnet requirements workshop:
  "Conservative estimates are an increase by a factor of 2 to 4" for 2 to 5 years in the future (2015-2018)
- The ESnet exponential traffic trend is larger, and remarkably steady: 10X every 4.25 Years (since 1992):
  100 PB/Month by 2016
- Case Study of CMS Physics Analysis Needs using location independent "cloud style" data access (AAA) showed:
  A factor of 5-10 within next 5 yrs ➡100G Target for each Tier2
  - Already realized at many US Tier2 sites
- Longer Term Trends: Fisk and Shank at Snowmass:
  100X growth in storage and network needs by LHC Run3 is possible

Transfer Rates: Caltech Tier2 to Europe July 2014

One Day after commissioning the 1st 100G TA research link

Upload rate: 27 Gbps; **20Gbps to CNAF (Italy) Alone**

- **By 2015:** 12 – 50 Gbps Downloads were Routine to US CMS Tier2 Sites; 92 Gbps to Wisconsin Seen in the Fall



US CMS university based Tier2s have moved to ~100G now

Caltech	100 Gbps
Florida	100 Gbps
MIT	100 Gbps
Nebraska	100 Gbps
Purdue	100 Gbps
UCSD	80 Gbps
Wisconsin	100 Gbps

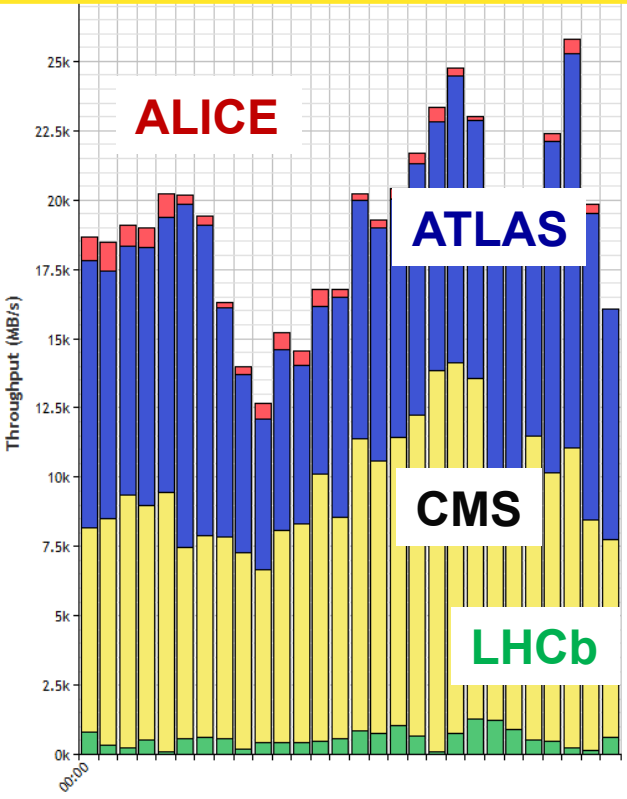
The move to 100G is timely and matches current needs, also at Tier2s. Backbones should continue to advance to meet the needs during Run2.

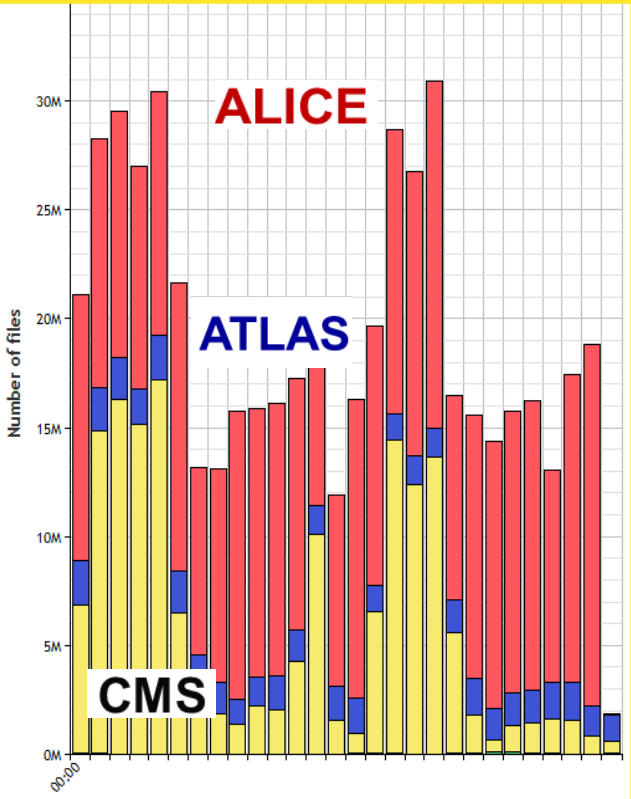
Complex Workflow: the Flow Patterns Have Increased in Scale and Complexity, even at the start of LHC Run2

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

Transfer Throughput | **Transfers Done/Day**

3X Growth from March to October 2015





20 GBytes/s Typical

To 35 GBytes/s Peak Transfer Rates

Complex Workflow

- **Multi-TByte Dataset Transfers**
- **Transfers of 12-35 Million Job Output Files Daily**
- **Access to Tens of Millions of Object Collections/Day**
- **>100k of remote connections (e.g. AAA) simultaneously**

WLCG Dashboard Snapshot Sept-Oct. Patterns Vary by Experiment

Entering a New Era of Technical Challenges
as we Move to Exascale Data and Computing

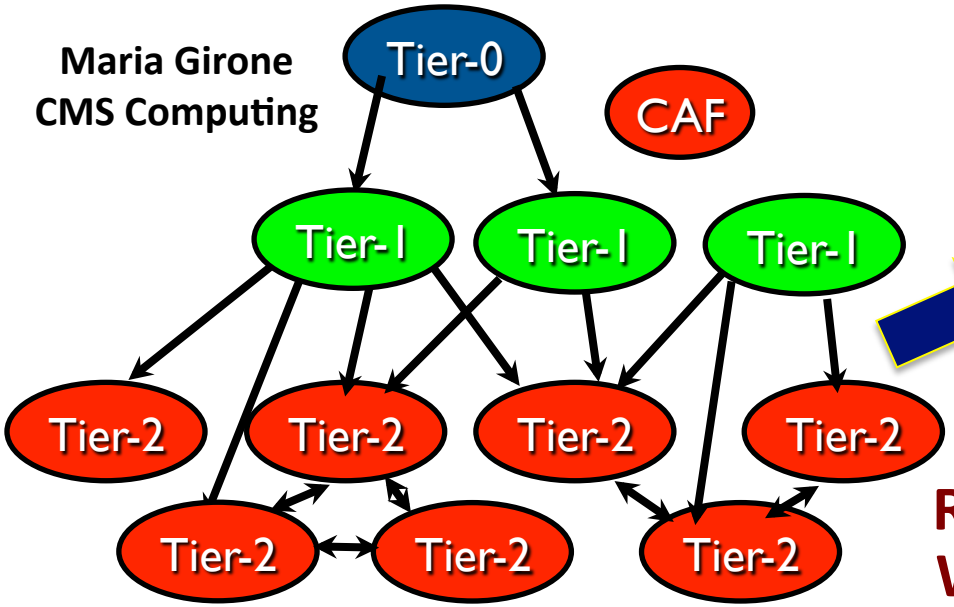
- The largest science datasets today, from LHC Run1, are 300 petabytes
  - Exabyte datasets are on the horizon, by the end of Run2 in 2018
  - These datasets are foreseen to grow by another 100X, to the ~50-100 Exabyte range, during the HL LHC era from 2025
- The reliance on high performance networks will thus continue to grow as many Exabytes of data are distributed, processed and analyzed at hundreds of sites around the world.
- As the needs of other fields continue to grow, HEP will face increasingly stiff competition for the use of large but limited network resources.

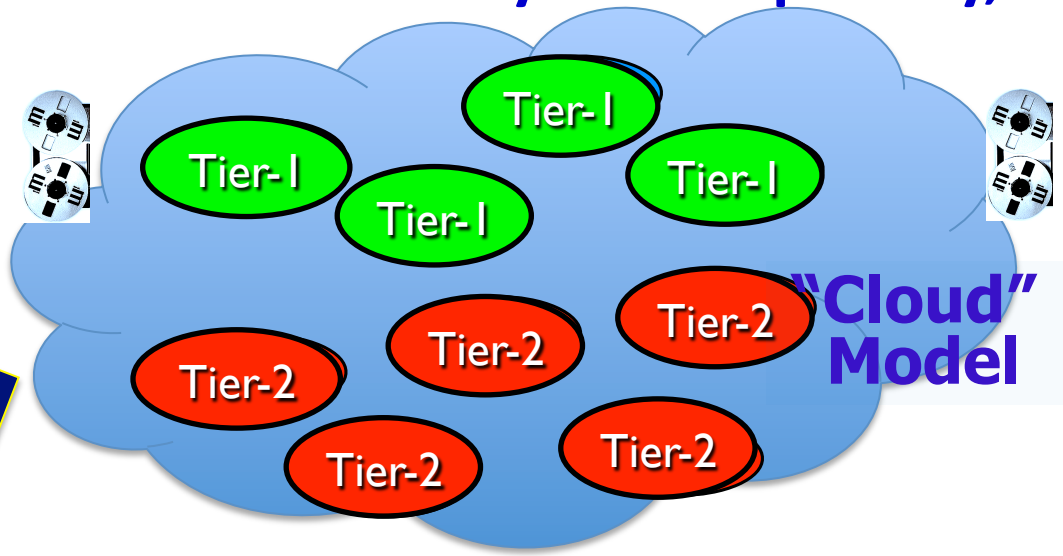
1 EB = 2 milligrams of DNA


Earth Observation

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



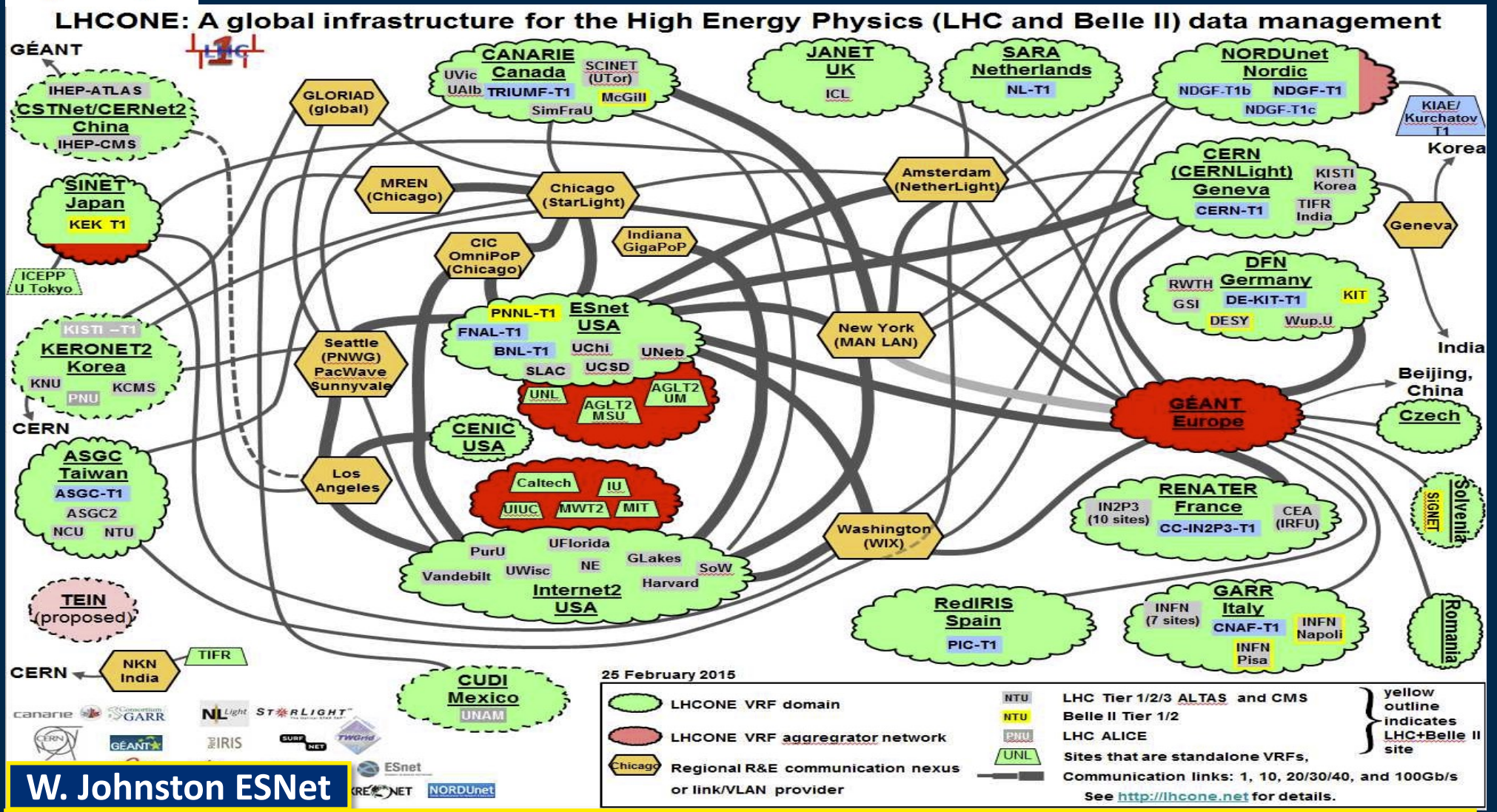


Run2: Scaling to 20% of data across the WAN: 200k jobs, 60k files, (100TB)/day

✸ + Elastic Cloud-like access from some Tier1/2/3 sites

LHCONE: a Virtual Routing and Forwarding (VRF) Fabric

A global infrastructure for HEP (LHC and Belle II) data management



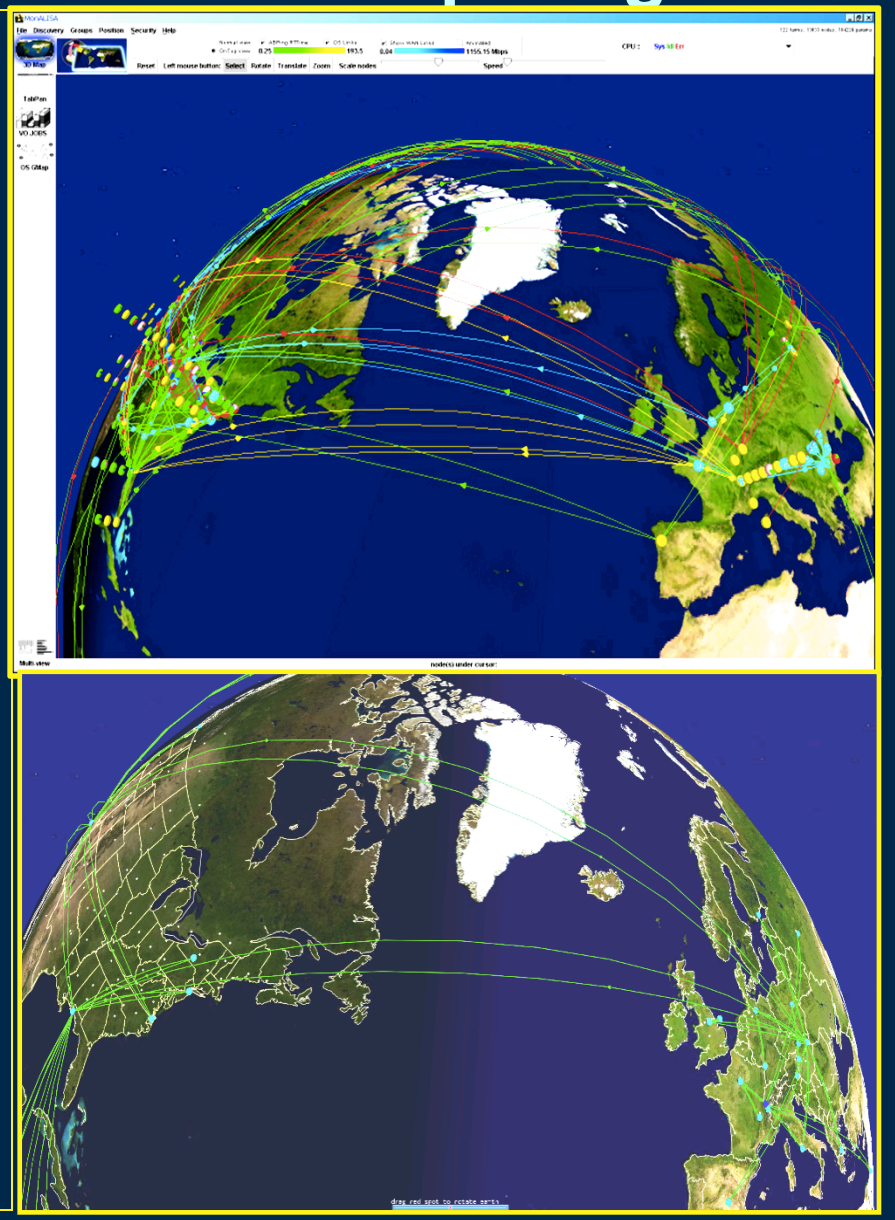
W. Johnston ESNet

The Major R&E Networks Have Mobilized on behalf of HEP

A stepping stone to our joint future: for data intensive sciences

Entering a New Era of Technical Challenges
as we Move to Exascale Data and Computing

- **Beyond network capacity and reliability alone,** the keys to future success are next generation systems **able to:**
  - Respond agilely **to peak and shifting workloads**
  - Accommodate a more diverse set of computing systems **from the Grid to the Cloud** to HPC
  - Coordinate the use of globally distributed computing and storage, and networks that interlink them
    - **In** a manner compatible across fields **sharing common networks**
- **The complexity of the data, and hence the needs for CPU power, will grow disproportionately:** by a factor of several hundred during the same period

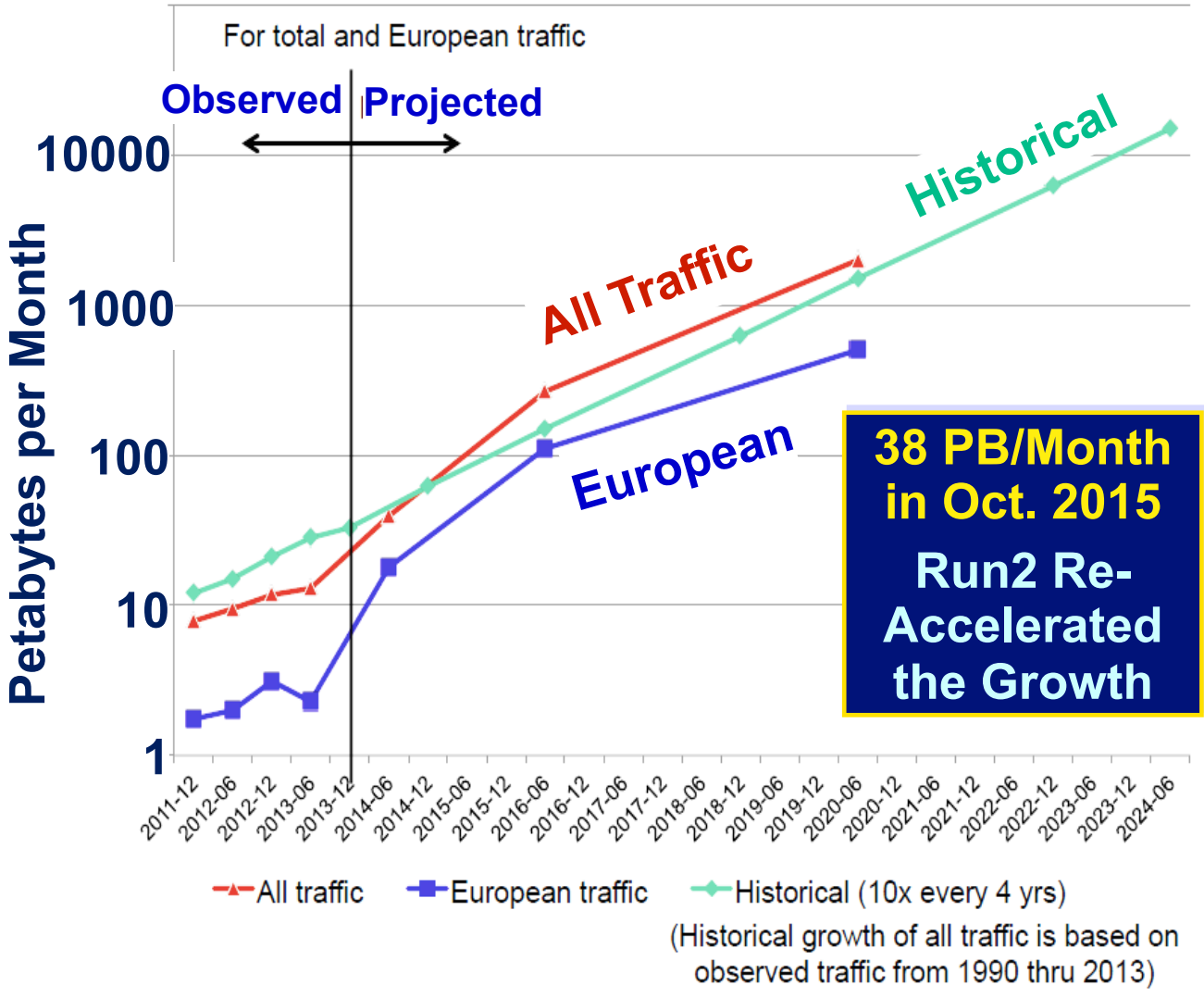


ESnet Science projection to 2024
Compared to historical traffic

E. Dart
W. Johnston

ESnet

Total traffic handled in Petabytes per Month



38 PB/Month in Oct. 2015

Run2 Re-Accelerated the Growth

Projected Traffic Reaches

1 Exabyte Per Month. by ~2020

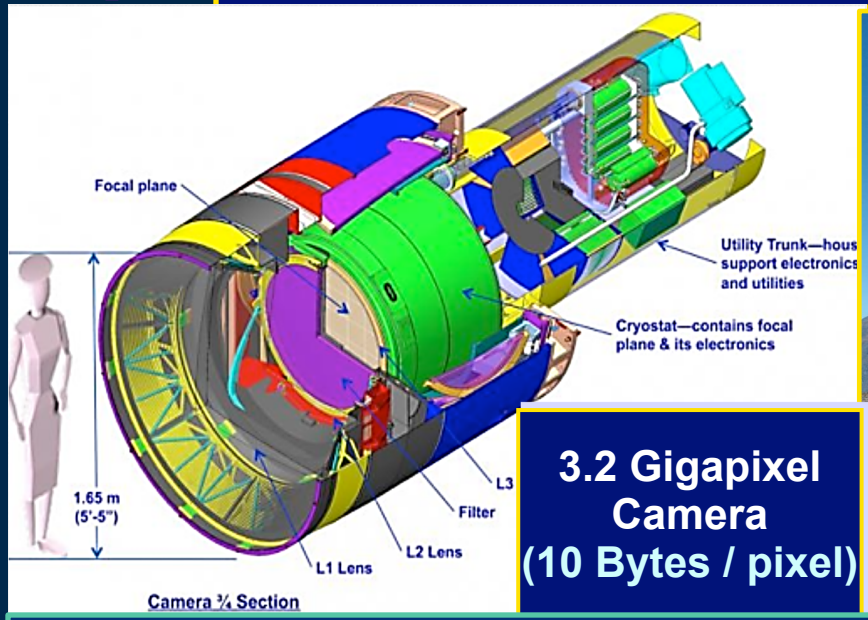
10 EB/Mo. by ~2024

Rate of increase follows or exceeds Historical trend of 10X per 4 Years

HEP traffic will compete with BES, BER and ASCR

LSST + SKA Data Movement

Upcoming *Real-time* Challenges for Astronomy

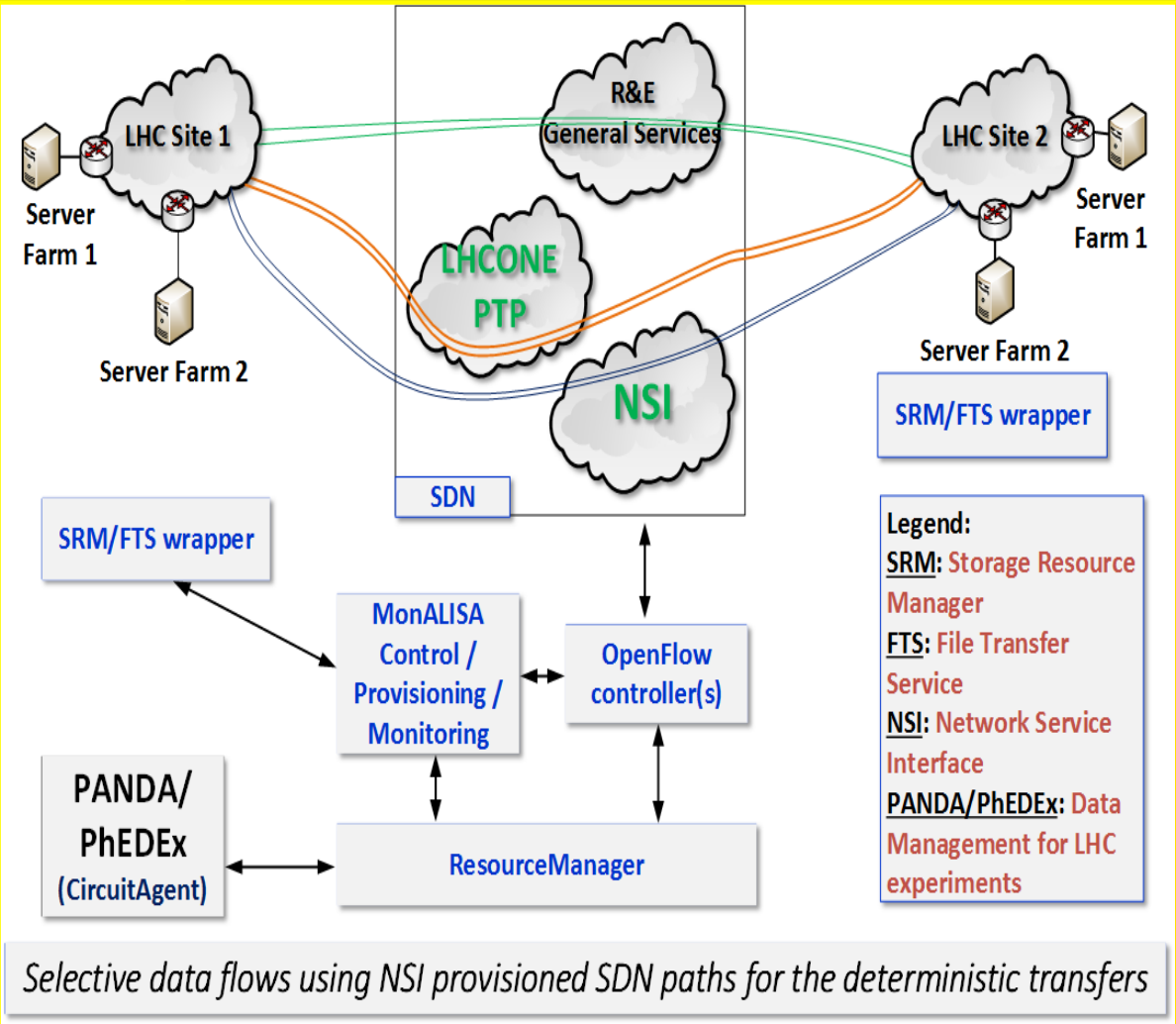


3.2 Gigapixel Camera (10 Bytes / pixel)

SKA

- **Planned Networks**: Dedicated 100G for image data, Second 100G for other traffic, and 40G for a diverse path
- Lossless compressed Image size = 2.7GB (~5 images transferred in parallel over a 100 Gbps link)
  - Custom transfer protocols for images (UDP Based)
- Real-time Challenge: delivery in seconds *to catch cosmic "events"*
- + SKA in Future: 3000 Antennae covering > 1 Million km2; *15,000 Terabits/sec to the correlators ➡ 1.5 Exabytes/yr Stored*

Caltech from ANSE to SDN-NGenIA Dynamic Circuits with Software-defined path building and load balancing



Selective data flows using NSI provisioned SDN paths for the deterministic transfers

Dynamic circuits used to create network paths with reserved bandwidth

OF Flow-matching is done on specific subnets to route only the desired data packets to the circuits

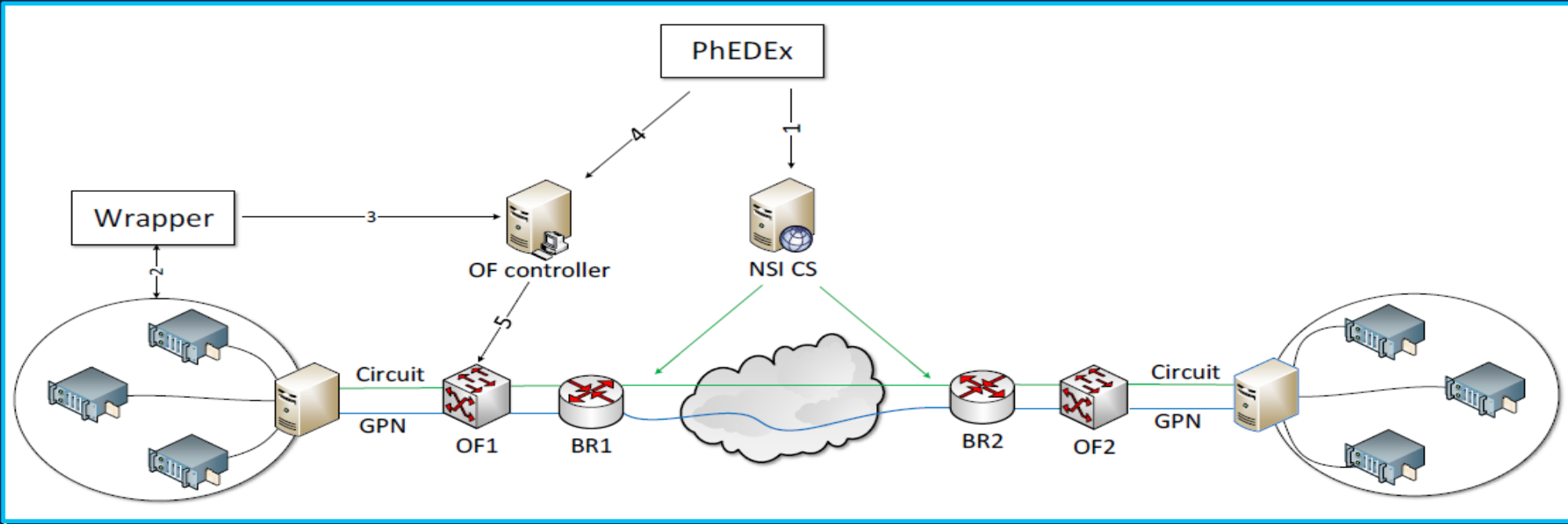
Caltech's controller is used to select paths for the circuits, based on available capacity, load-balancing, etc.

Controller also can be used to load-balance and/or moderate flows over multiple non-circuit paths

Applications: PhEDEx, PanDA File Transfers. ASO Job Output Stageout

Lapatadescu, Bunn, Voicu, Wildish, Mughal, Legrand, Kcira, Balcas, HN

Integrating OSCARS/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 ontroller 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

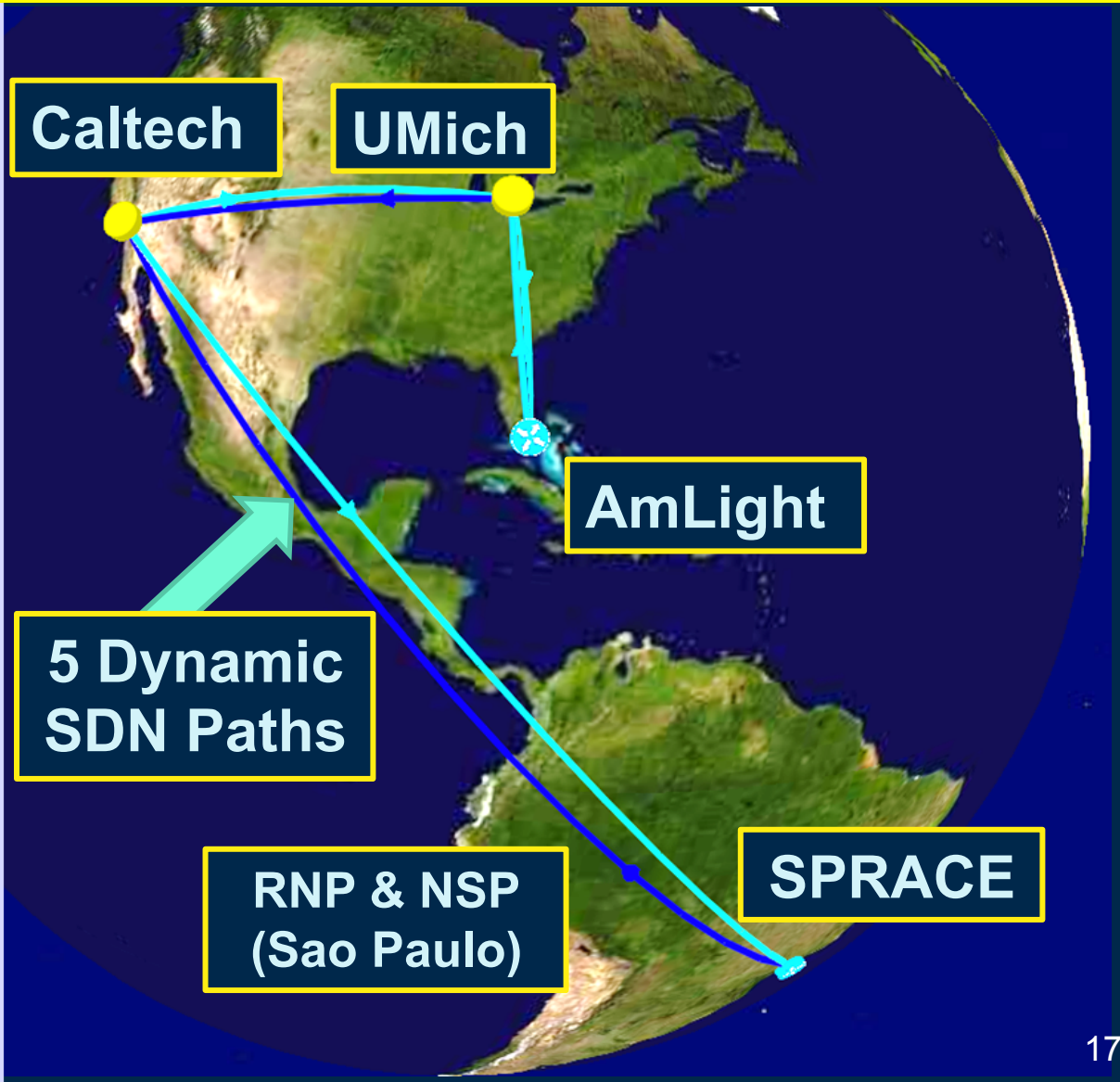
SDN Demonstration at FTW Workshop. Partners: Caltech, UMich, Amlight/FIU, Internet2, ESnet, ANSP+RNP

Dynamic Path creation:
Caltech – UMich
Caltech – RNP
Umich – AmLight

Path initiation by the FDT Agent using OSCARS API Calls

OESS for OpenFlow data plane provisioning over Internet2/AL2S

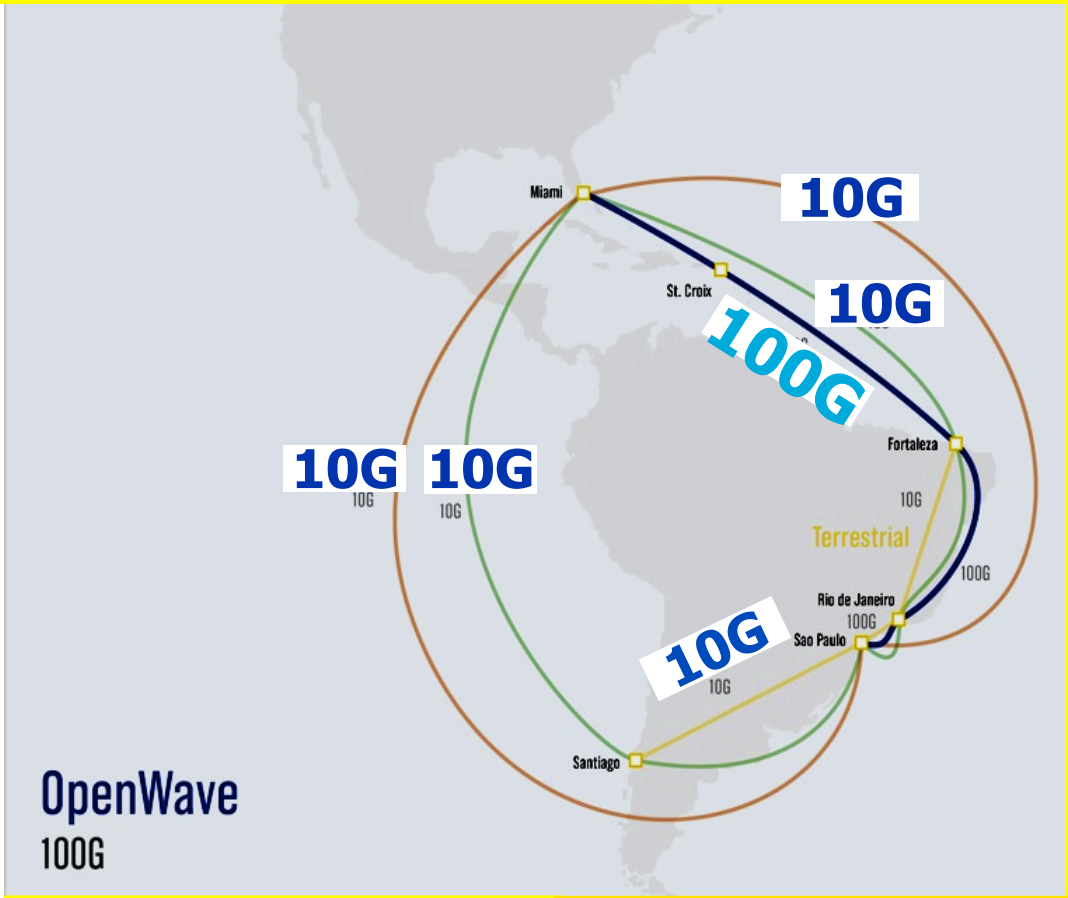
MonALISA agents at the end-sites provide detailed monitoring information



OpenWave: First 100G Link to Latin America in 2015. Connecting LSST

AmLight (US NSF) with RNP, ANSP

Total Capacity for Next Two Years: 140G



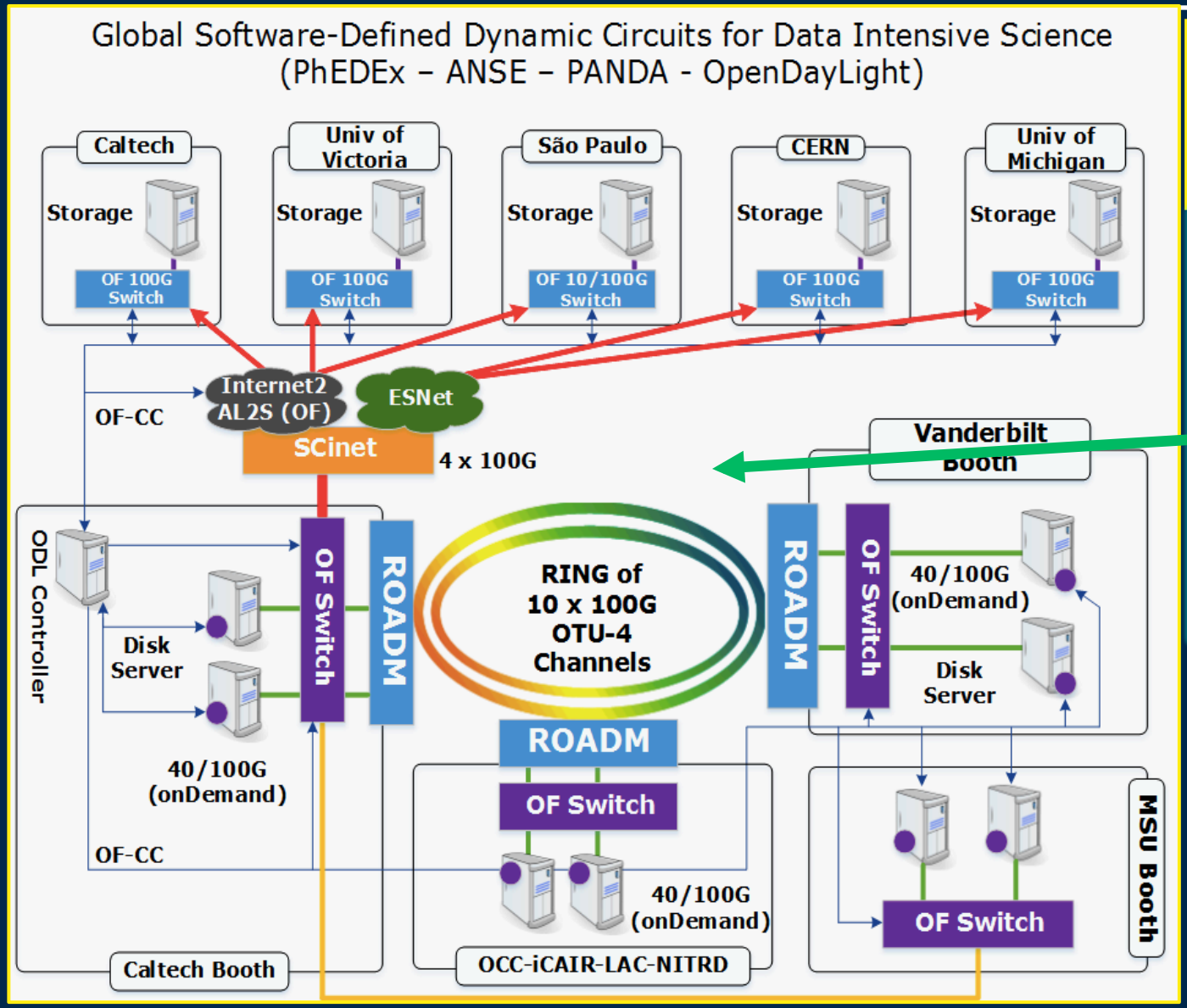
- An "Alien Wave" at 100G on the Undersea Cable
  - ➡ Precedent-setting access to the frequency spectrum by the academic community
  - ➡ Sao Paulo-Rio-Fortaleza -St. Croix-Miami backbone
  - ➡ Scheduled to start soon
- 100G extensions by RNP in Rio and ANSP in Sao Paulo
- Will connect to future 100G cable: Fortaleza-Lisbon
- Will be heavily used by LSST in the future

February 2015

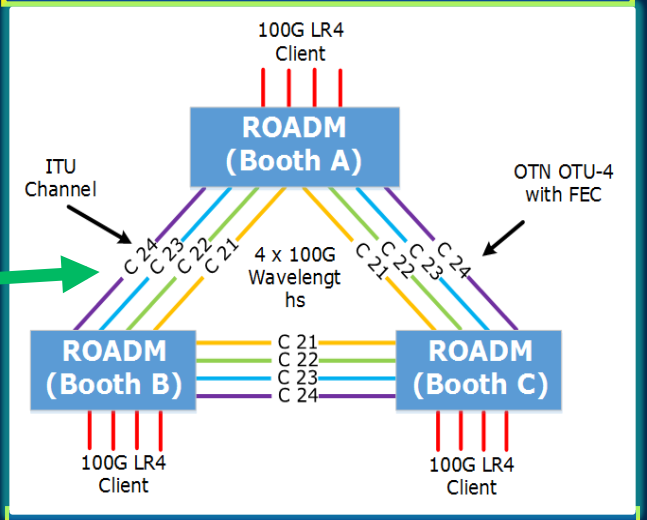
J. Ibarra, AmLight

➡ Using Padtec (BR) 100G equipment. Demonstrations with the HEP team (Caltech et al) at SC2013 and 2014

SC14: Global MultiLayer Software-Defined Dynamic Circuits for Data Intensive Science



Terabit/sec Scale Long Range Networking



30 100G Waves

SDN (ODL) Control of Optical and Switching Systems

Caltech HEP and Partners

1 Tbps Scale Demonstration: Caltech, UVic, Vanderbilt, CERN Sao Paolo, UMich, ESnet

Padtec 10 x 100G DWDM System:

Connected to Vanderbilt Booth and iCAIR

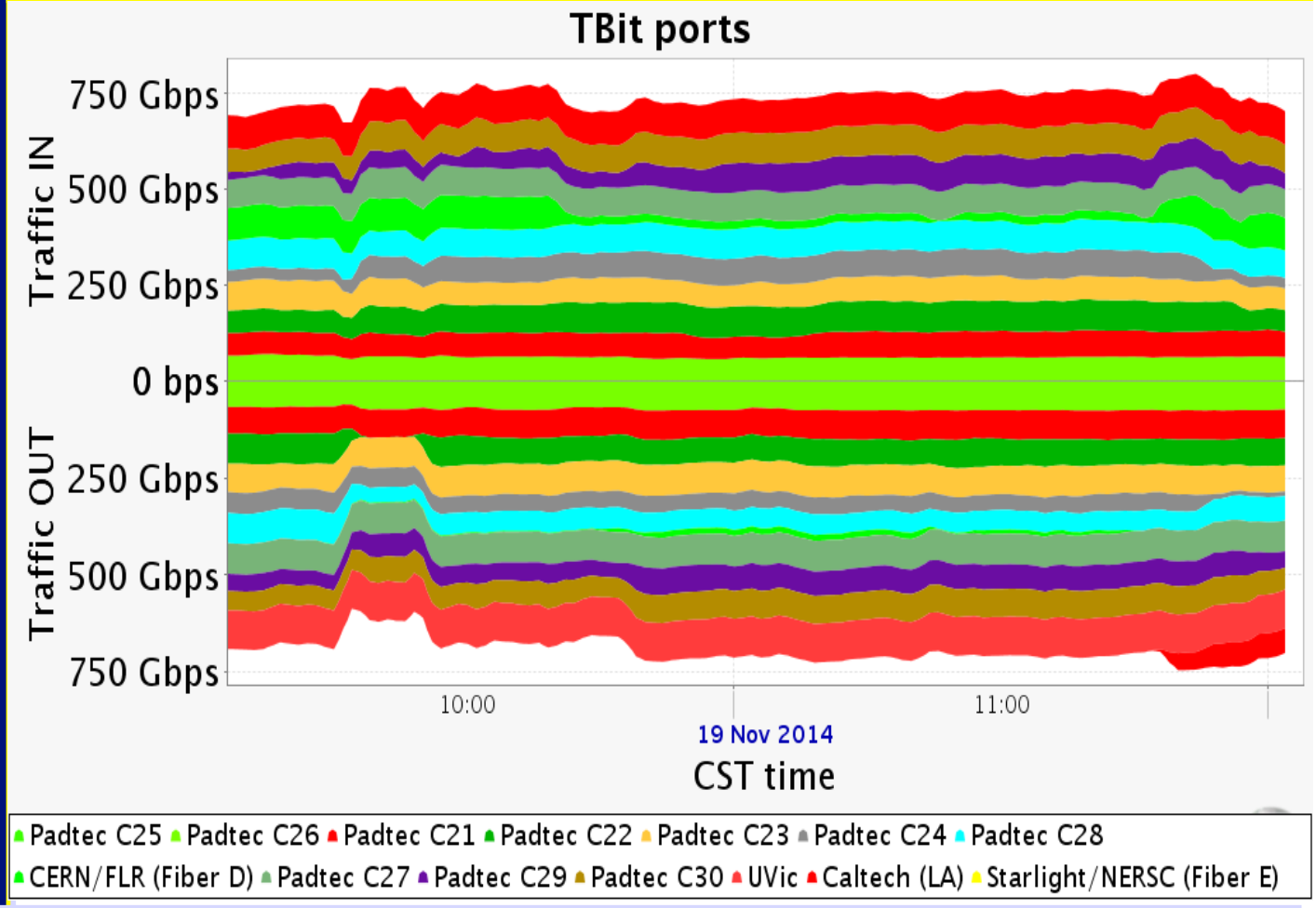
Echostreams 2U Servers: 36 SSD ea.

N x 100G Brocade and Extreme Networks

OpenFlow Switches

~ 900 SSDs

1.35 Tbps Sustained; 400G over WANs

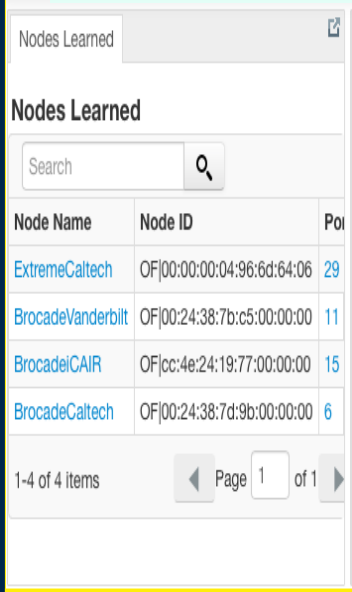


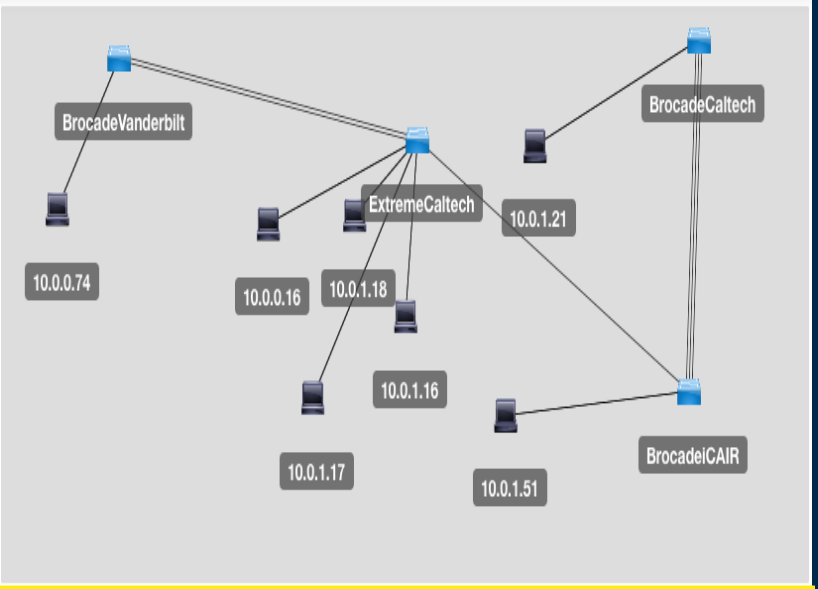
Network partners: SCinet, ESnet, Internet2, ANA-100, CENIC, Wilcon, PacWave, Starlight, MANLAN, MiLR, SURFNet, FLR, RNP, ANSP, AmLight

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 are "punted" to Controller by switch

SC14 Demo: Caltech/iCAIR/Vanderbilt OF links

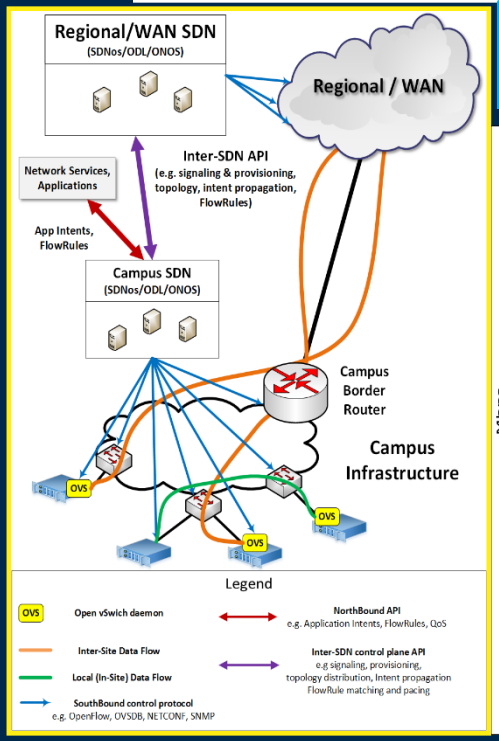


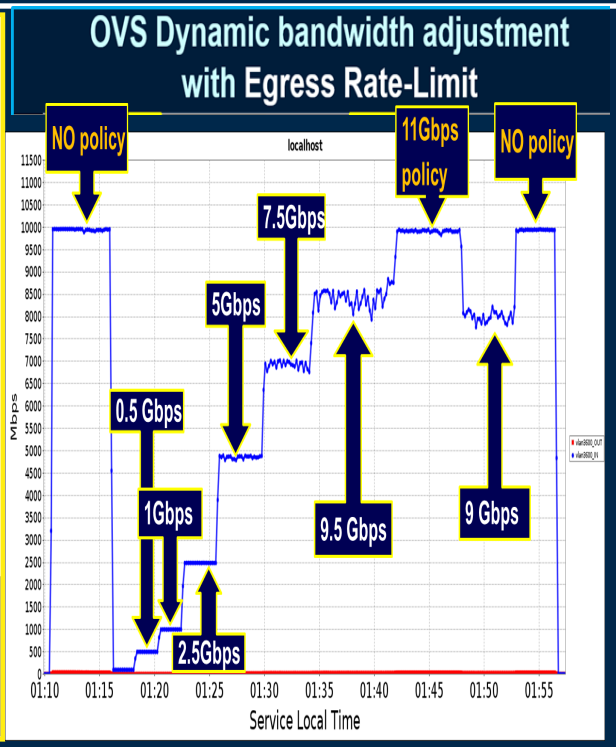


Demonstrated:

- 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

OpenvSwitch: Managing Site Interactions
Locally and with Regional and Wide Area Networks

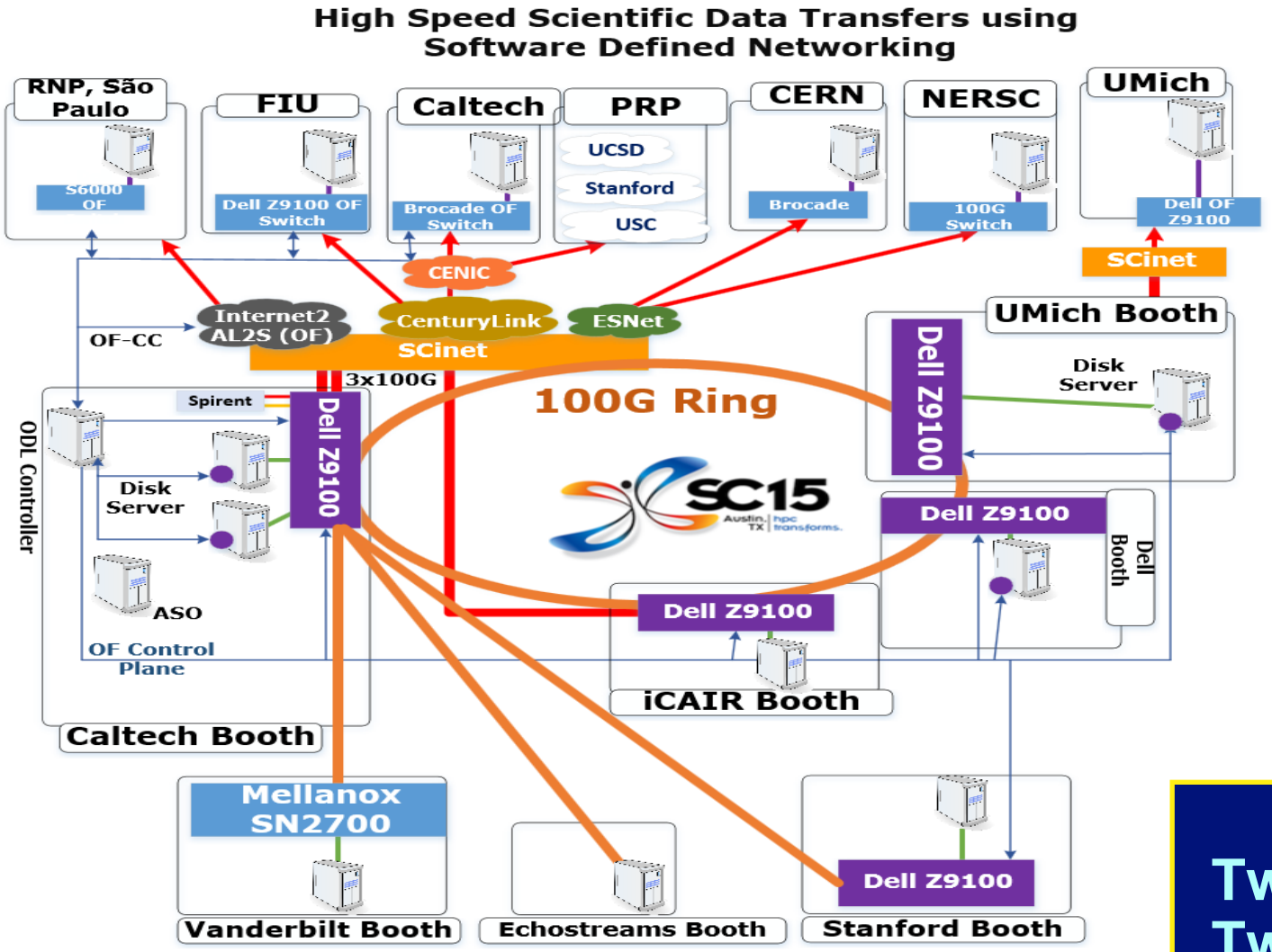




- Caltech OVS Tests: Seamless control of flows at any level, to 10G wire speed
- Very low CPU load: 0 to 5%
- Protocol agnostic and well integrated in the Linux kernel
- 40G/100G Tests Underway

- Standard Northbound API, exposed by the SDN controllers at both Regional/WAN and Site level
- NB API used by external Network Services (or Applications)
  - To request/allocate network resources
  - For automatic failover among subclusters and other intra-site functions
  - To dynamically tune the bandwidth for aggregated flows (e.g. from multiple storage nodes, or even an entire rack, cluster, etc.)
  - To Protect (or Prioritize) different types of individual data flows at the end site, where needed

SC15: SDN Driven Next Generation Terabit/sec Integrated Network for Exascale Science



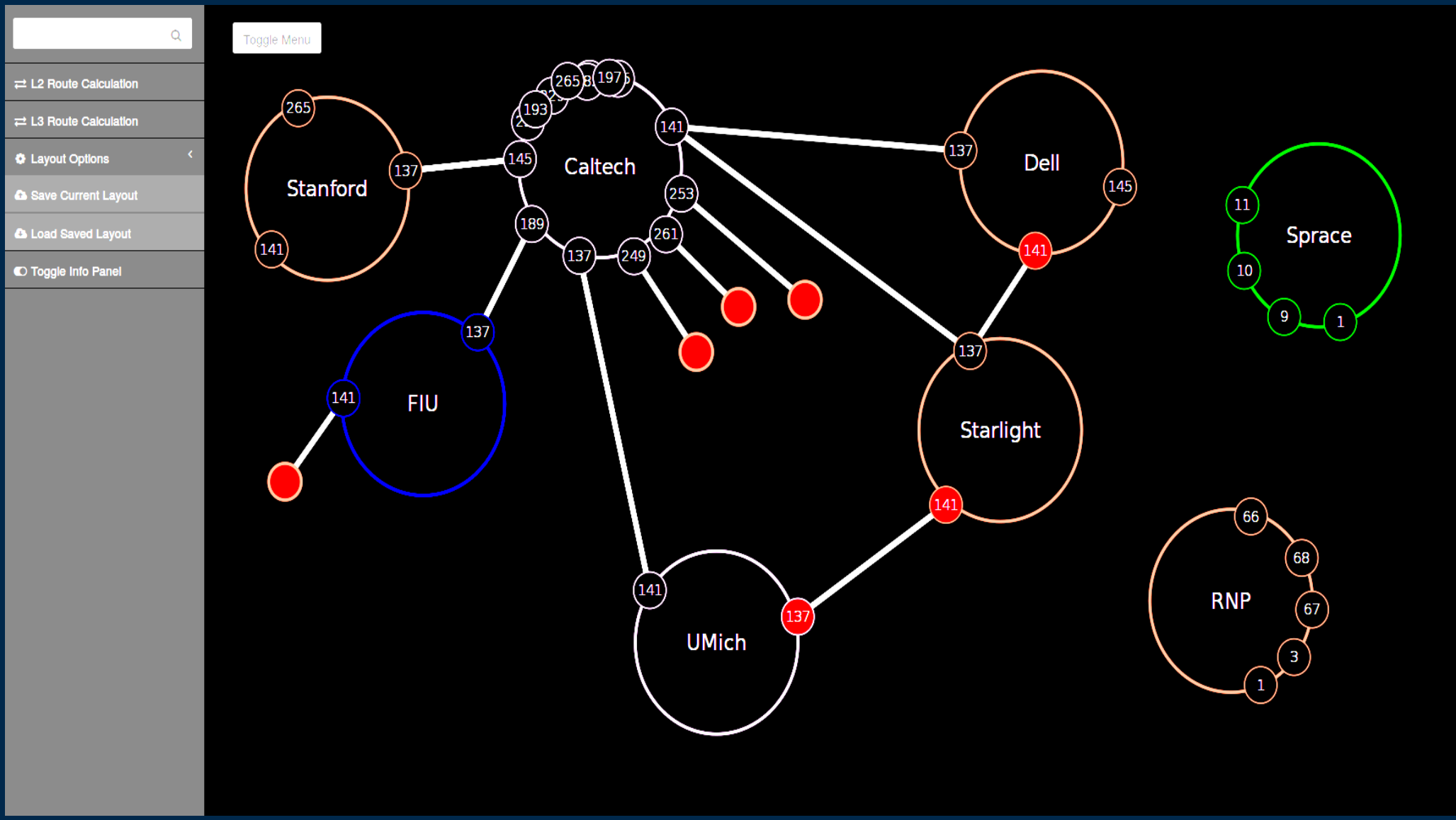
SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

Consistent Operations with Agile Feedback: Supporting Major Science Flows Compatible with other Traffic

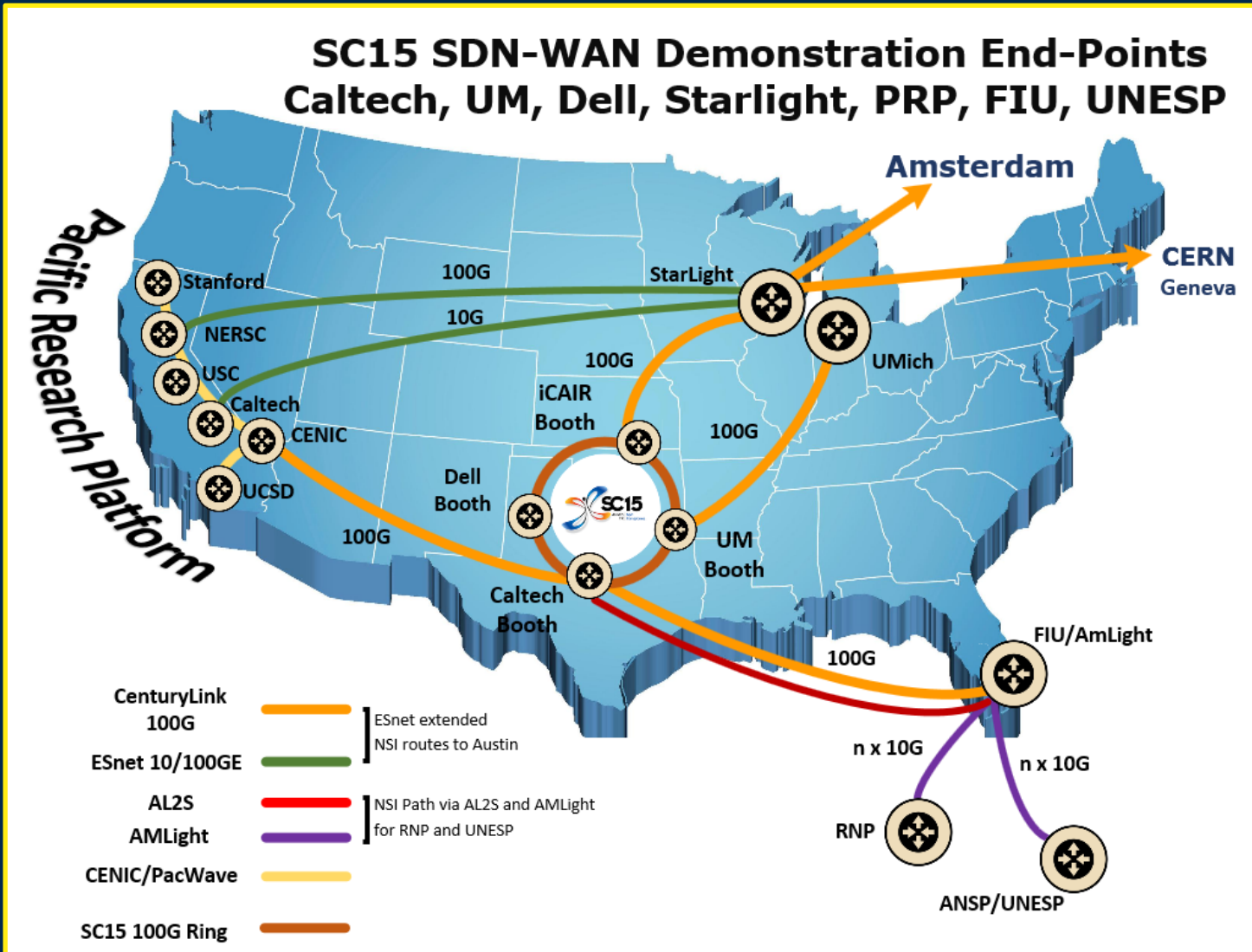
29 100G NICs
Two 4 X 100G DTNs
Two 3 X 100G DTNs
9 32 X100G Switches

Caltech HEP & Partners. Open Daylight Controller

SC15: SDN Driven Terabit/sec Network Topology



SC15: SDN Driven Next Generation Terabit/sec Integrated Network for Exascale Science



Caltech HEP & Partners. Open Daylight Controller

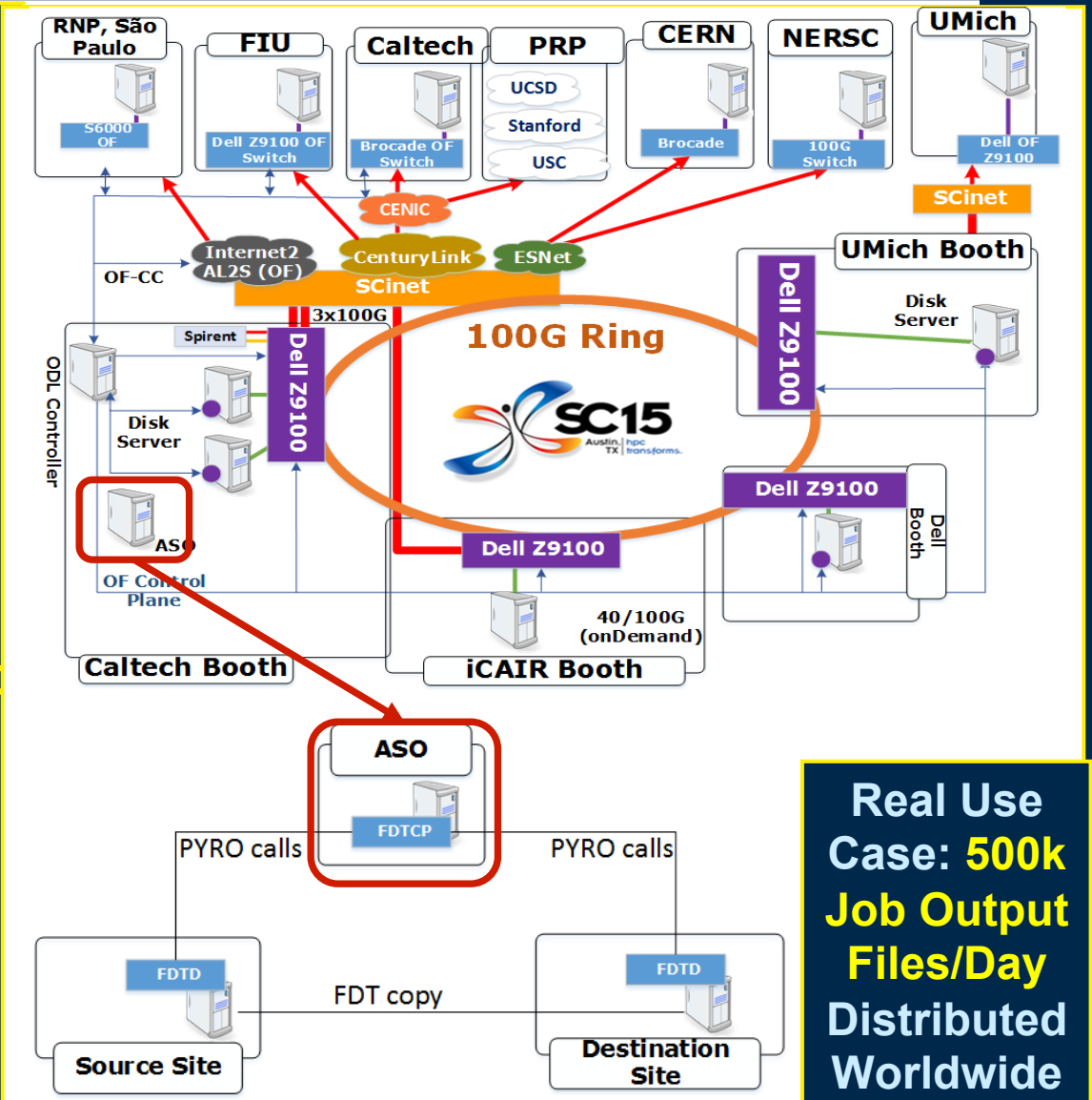
SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

Consistent Operations with Agile Feedback: Major Science Flow Classes Up to High Water Marks

PetaByte Transfers to and From the Site Edges of Exascale Facilities With 400G DTNs

CMS at SC15: Asynchronous Stage Out 3rd Party Copy Demonstration

- All control logic in **ASO**:
  - Group multiple file transfers per link
  - Controls number of parallel transfers
  - Transparent for **ASO** integration
- Only **FDT daemon** has to be installed on storage site
- Tests between end-hosts at **Caltech, Umich, Dell booths** and outside: **FIU, Caltech, CERN, Umich**
- **PetaByte transfers** from multiple sites to multiple locations



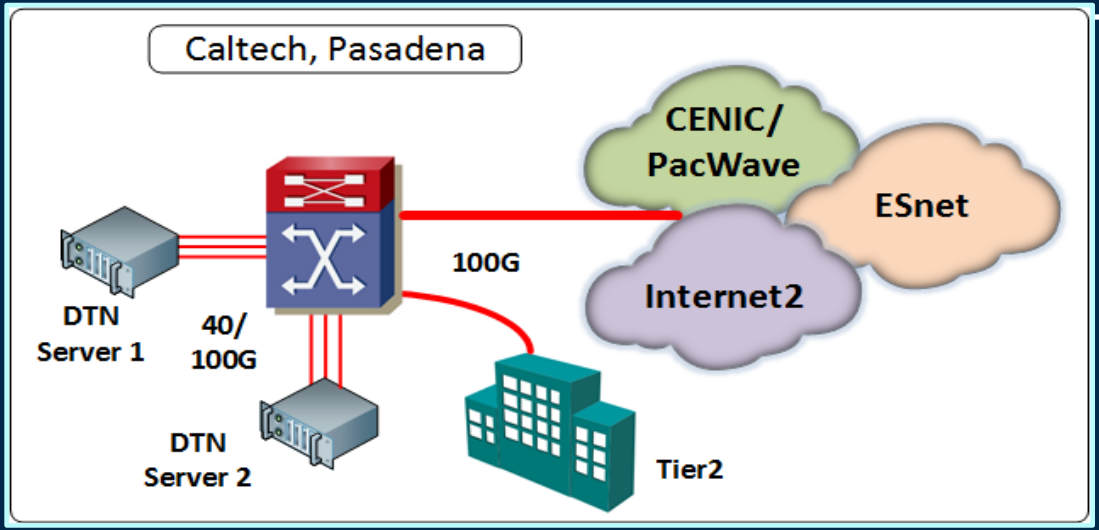
Real Use Case: 500k Job Output Files/Day Distributed Worldwide

Servers at the Caltech Booth
Multi-100G DTNs, SDN, Machine Learning

- 2 400G DTNs: Dell Model R930 4U servers with E7 four socket CPUs; each with 4 100G NICs
- Third R930 server with 5 Model MX6300 Mangstor cards capable of 18+/12 R/W GBytes/sec
- Fourth R930 server with 4 Intel Model DC P3700 SSDs and 1 100G NIC
- 1 Supermicro server with 8 Intel Model DC P3700 SSDs, 2 40G Mellanox NICs (Connects to ESnet)
- 2 Supermicro 4U dual E5-2697 servers each with 3 100G NICs
- 3 SuperMicro (2U dual E5-2670) w/24 OCZ Vertex4 and Intel SSDs each
- 2 Echostreams server (4U and 2U, processors dual E5 2.2 GHz) each with 100G Mellanox NIC
- Echostreams/Orange Labs Server with 16 Tesla K80 GPUs: 100 Teraflops in 4U

Caltech: Data Transfer Node Design

4 X 100G in One DTN



- The Caltech DTNs will be connected to various national backbones: ESnet, CENIC/PacWave, PRP, I2, FLR, AmLight, ANSP, RNP
- Planned use in SDN-NGenIA testbed and pre-production developments
- Production Objective for 2016: Send and receive Petabyte datasets to/from the site perimeter of a Leadership Computing Facility

2015 Milestone: Commission a "400G" (4 X 100G Ethernet) Data Transfer Node

4 X 100G Data Transfer Node

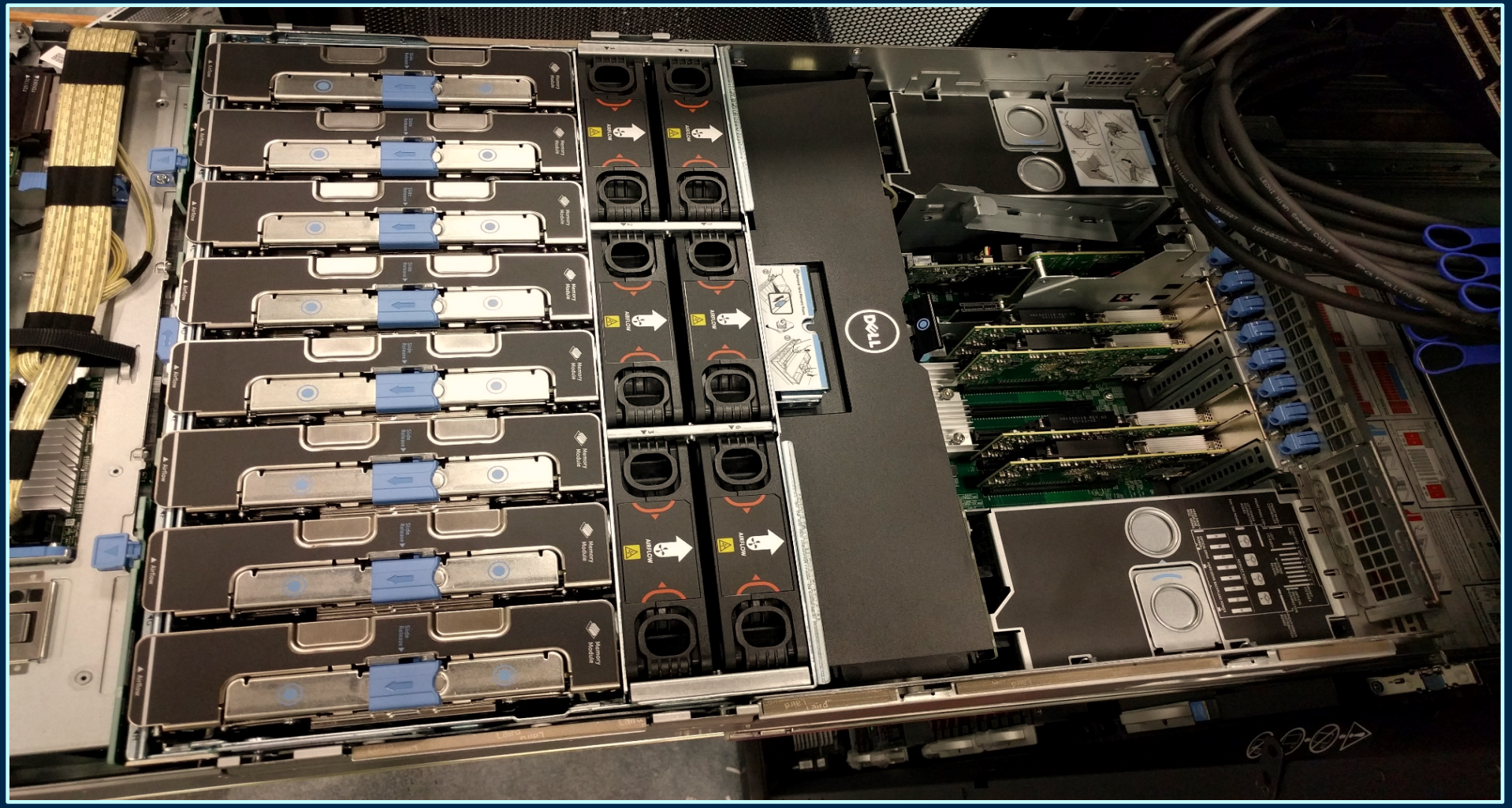
- Dell R930 4U Chassis
- Haswell E7 8891 Processors
- DDR4, 128GB Memory
- Mellanox 100GE NICs

Software

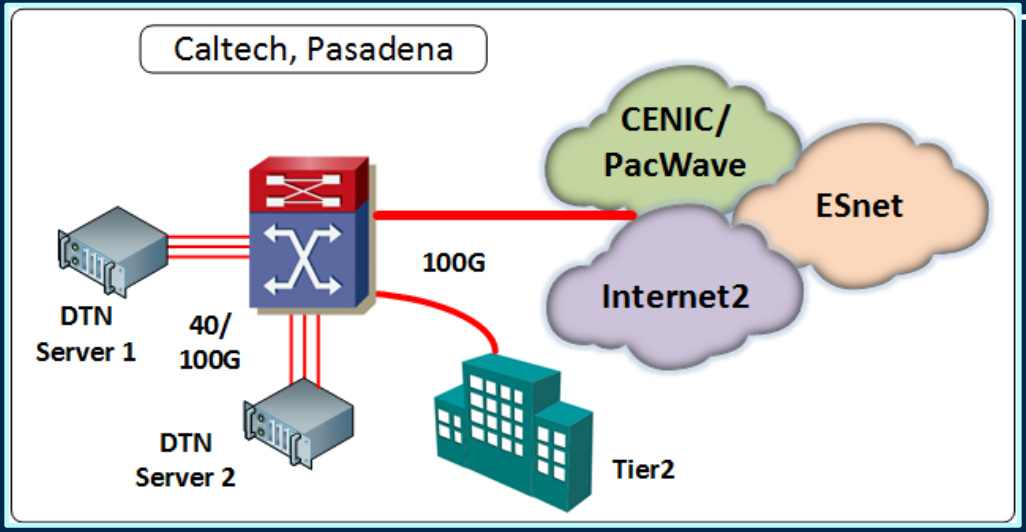
- CentOS 7.1
- Mellanox OFED 3.1
- FDT: Caltech's data transfer application

Servers at the Caltech Booth
Multi-100G DTNs, SDN, Machine Learning

- 2 400G DTNs: Dell Model R930 4U servers with E7 four socket CPUs; each with 4 100G NICs



Caltech Booth: Next Gen Data Transfer Node
3 X 100G in One DTN



- The Caltech DTNs are connected to various national backbones: ESnet, Internet2, CENIC/PacWave
- Planned use in SDN-NGenIA testbed and pre-production developments
- Production Objective for 2016: Send and receive Petabyte datasets to/from the site perimeter of an LCF

2015 Milestone: Commission a "300G" (3 X 100G Ethernet) Data Transfer Node

3 X 100G DTN Design Exists

- SuperMicro 3U chassis
- Haswell X10DRI Motherboard
- Dual Intel Haswell Processors (E5-2697v3)
- DDR4, 128GB Memory
- Mellanox and QLogic 100GE NICs

Software

- CentOS 7.1
- Mellanox OFED 3.1
- FDT: Caltech's data transfer application

Convergence and Collaboration: Tackling the Big Issues for LHC Run2

- Short-term Milestone: Enabling more efficient, manageable workflow by integrating advance networking into the LHC Experiments' mainstream software and data systems, along with CPU and Storage
  - Integrating Network Awareness in CMS (PhEDEx) and ATLAS (PanDA); map priority flows to Dynamic Circuits: *ANSE (NSF)➡ SENSE + SDN NGenIA (DOE/ASCR) Projects*
- Collaborating in Developing Key Technologies
  - Production use of Terabyte to Petabyte Transfers with State of the Art High Throughput
  - Dynamic Circuits to Campuses Across the Country
    - Extending the Science DMZ Concept
  - Software Defined Networking: Using Emerging Standards
  - Named Data Networking: A possible new network paradigm

Convergence and Collaboration
Tackling the Larger Mission

- Empowering Data Intensive Science across multiple fields through efficient, manageable use of national & global infrastructures up to high occupancy levels, including multi-pathing
- Using SDN-driven coordinated use of computing, storage and Network resources for efficient workflow
- Enabled by Pervasive End-to-end Monitoring
- Consistent Operations: Networks ⇔ Science Programs; with feedback
- Key Concepts and Technologies for Success:
  - Dynamic circuits for priority tasks, with Transfer Queuing, Deadline scheduling, Efficient worldwide distribution and sharing
  - Classes of Service by flow characteristics, residency time
  - Load balancing, hotspot resolution, strategic redirection
  - State-based error propagation, localization, resolution
  - SDN driven Intent-based deep site-network orchestration functions
- System Level Optimization Using Machine Learning

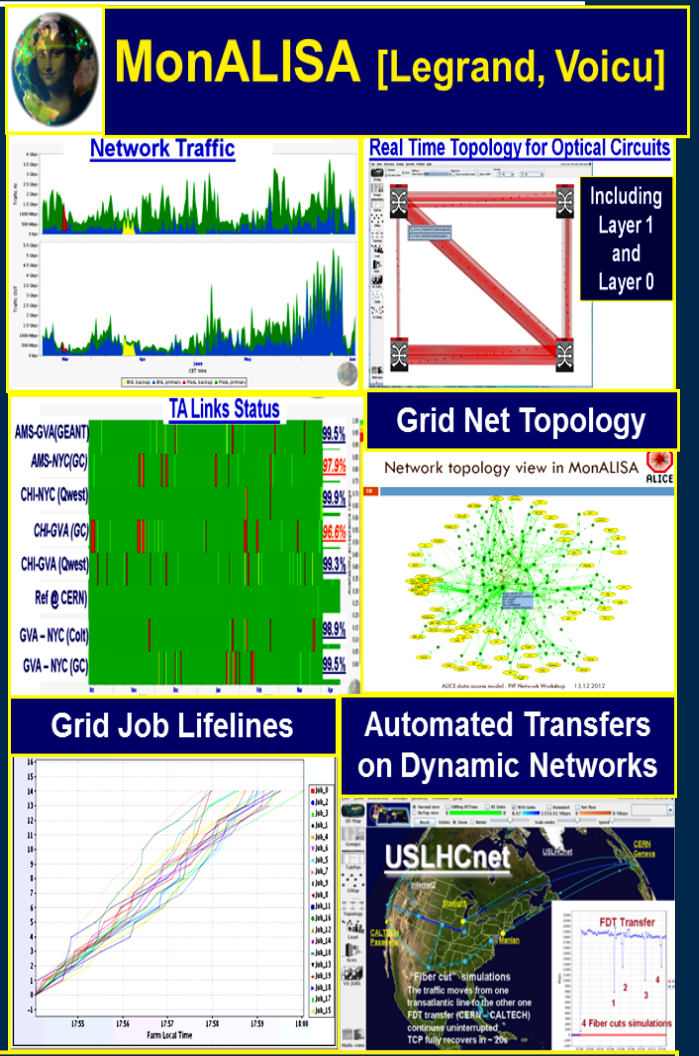
Networks for HEP and Global Science

Our Journey to Discovery

- Run 1 brought us a centennial discovery: the Higgs Boson
- Run 2 will bring us (at least) greater knowledge, and perhaps greater discoveries: *Physics beyond the Standard Model.*
- *Advanced networks will continue to be a key* to the discoveries in HEP and other data intensive fields of science and engineering
- Technology evolution might fulfill the short term needs
- *Near Term Challenges:* A new net paradigm including the global use of circuits will need to emerge during LHC Run2 (in 2015-18)
- *New approaches + a new class of global networked systems to handle Exabyte-scale data are needed [LHCONE, DYNES, ANSE, OliMPS; SENSE+SDNNGenIA]*
- *Worldwide deployment of such systems in ~2020-24 will be:*
  - Essential for the High Luminosity LHC HL-LHC
  - A game-changer, with global impact, shaping both research and daily life

SDN NGenIA: A Winning Strategy
A Next Generation Architecture for Exascale Science

- *Create:* **An agile worldwide-distributed system driving the efficient interplay of software** with an elastic and diverse set of resources: **Computing, Storage, Network**
- Pathway to a Solution
- **Co-design of the methods make consistent, coordinated use of the resources:**
  - Built on a foundation of pervasive monitoring
  - Reacting/adapting to changing conditions, work in progress/work scheduled
  - New modes of steering, use (and reuse) of data products **produced and consumed at many locations**
  - New modes of propagating information on data product availability and the cost of delivery **versus re-computation in real time**



Coherent Interactions **among the experiments' workflow management systems, the end sites,** the network and the user groups *as a System*

Vision: Towards the Next Generation LHC Computing Models

- The experiments are gravitating towards more "location independent", transparent access to data and results, as seen by users
  - Drives the need for the agile, intelligent system mentioned above
- Timescale in steps:

1. Immediate: continue to bring network awareness to the data operations of CMS and ATLAS through PhEDEx and PanDA respectively.
2. During Run2 (2015-18): through developments already underway, deploy first implementations of strategic data distribution and management supported by dynamic network provisioning, and coordinated use of network and other resources.
3. By end of Run2 (2018): Following prototyping, deploy 1st pre-production system underpinning the next generation Computing Model. Exercise the system with the full Run 2 dataset
4. LS2 and Run3: Develop and scale 1st production version of the next generation network-aware Computing Model
5. LS3 and Run 4: Scale to a production-ready Computing Model system able to meet the needs of the experiments at the start of the HL LHC

Backup
Slides Follow

SDN NGenIA: Work Items

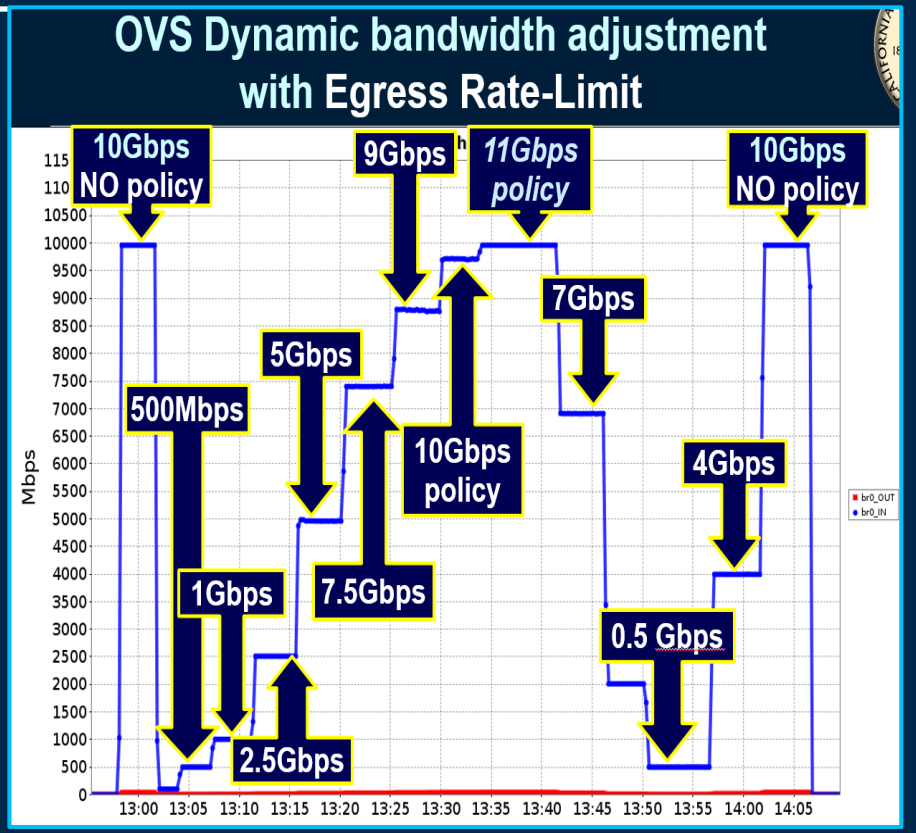
1. Deep site orchestration among virtualized clusters, storage subsystems and subnets to successfully co-schedule CPU, storage and networks
2. Science-program designed site architectures, operational modes, and policy and resource usage priorities, adjudicated across multiple network domains and among multiple virtual organizations
3. Seamlessly extending end-to-end operation across both extra-site and intra-site boundaries through the use of next generation Science DMZs
4. Novel methods of file system integration that enable granular control of extreme scale long distance transfers through flow matching of scattered source-destination address pairs to multi-domain dynamic circuits
5. Funneling massive sets of streams to DTNs at the site edge hosting petascale buffer pools configured for flows of 100 Gbps and up, exploiting state of the art data transfers where possible
6. Adaptive scheduling based on pervasive end-to-end monitoring, including DTN or compute-node resident agents providing comprehensive end-system profiling
7. Unsupervised and supervised machine learning and modeling methods to optimize the workflow involving terabyte to multi-petabyte datasets.

SDN NGenIA Deliverables
Site Orchestration

- System design and first prototype (3 months)
  - ODL orchestration module to control remote OVS instances
- Monitoring Integration and initial testbed installation (3 months)
  - ODL orchestration for end-host and OVS monitoring data gathering; first performance tests over WAN (ESnet)
- Integration of Monitoring and ML optimization feedback (4 months)
  - Monitoring integration and ODL application extension to support QoS orchestration for OVS with feedback from ML
- Integration into a prototypical pre-production system (3 months)
- Migration and/or integration to/with Network OS (3 months)
  - Migrate the orchestration and monitoring modules into the emerging SENOS network operating system

End Site Orchestration: Seamless Operation Across Site Boundaries with Open vSwitch (OVS)

- Seamless operations across site-boundaries (with QoS): a key part of the end-to-end vision
- Open vSwitch (OVS), designed to enable network automation via virtualized programmatic interfaces can do this.
- Using standard and well established protocols for Security, Monitoring, QoS, Automated Control, failover, etc.
- Caltech's tests showed that OVS can support stable flows at specified levels, to 10G wire speed, independent of the transfer protocol.
  - With very low CPU overhead: 0-5%
  - No penalty for applying "policy" (QoS) to individual flows
  - Tests at 40G to 100G are underway



- OVS thus promises to support many long range flows in a diverse production environment
- Gradual migration of part of a cluster is possible and seamless

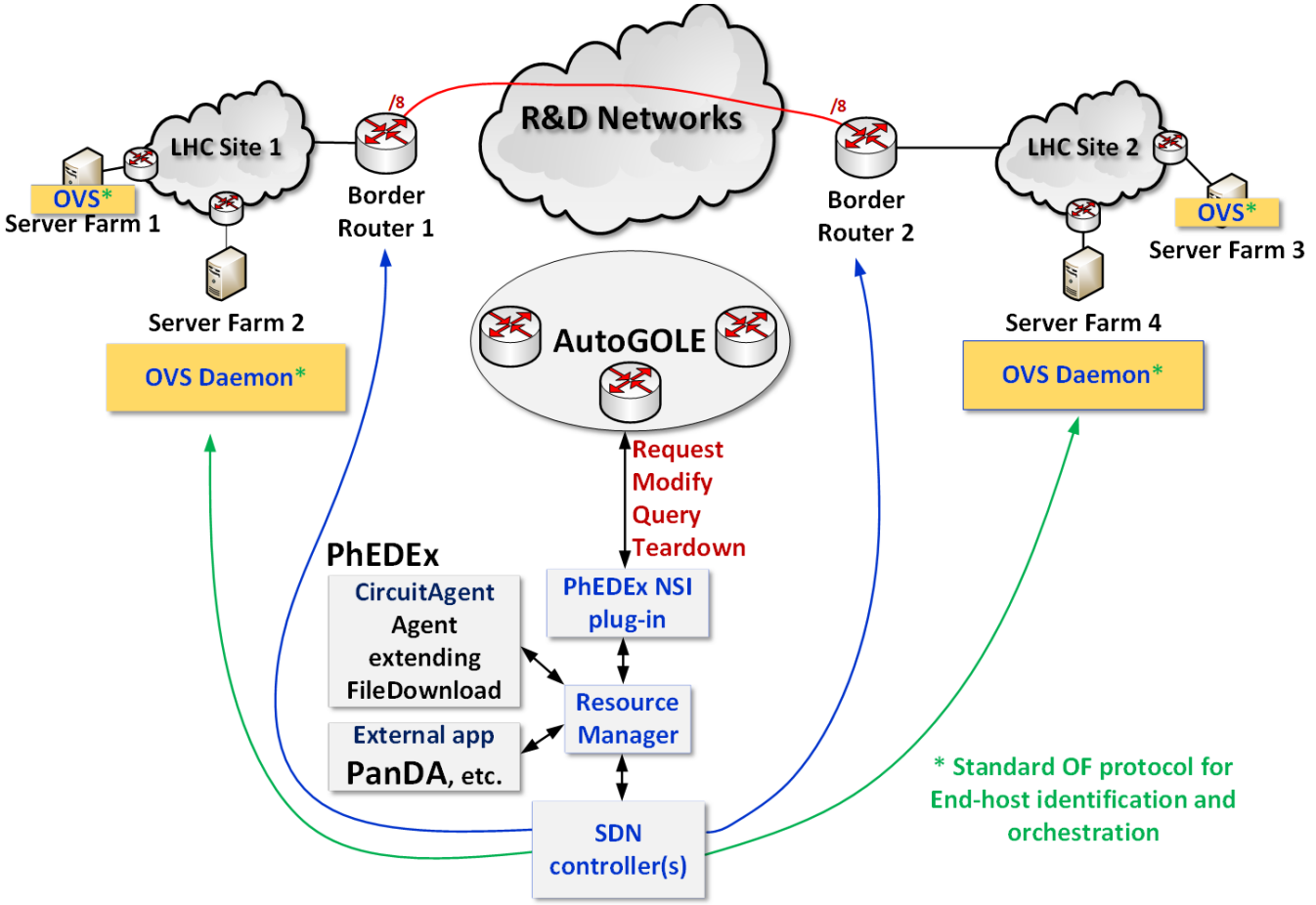
www.openvswitch.org

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



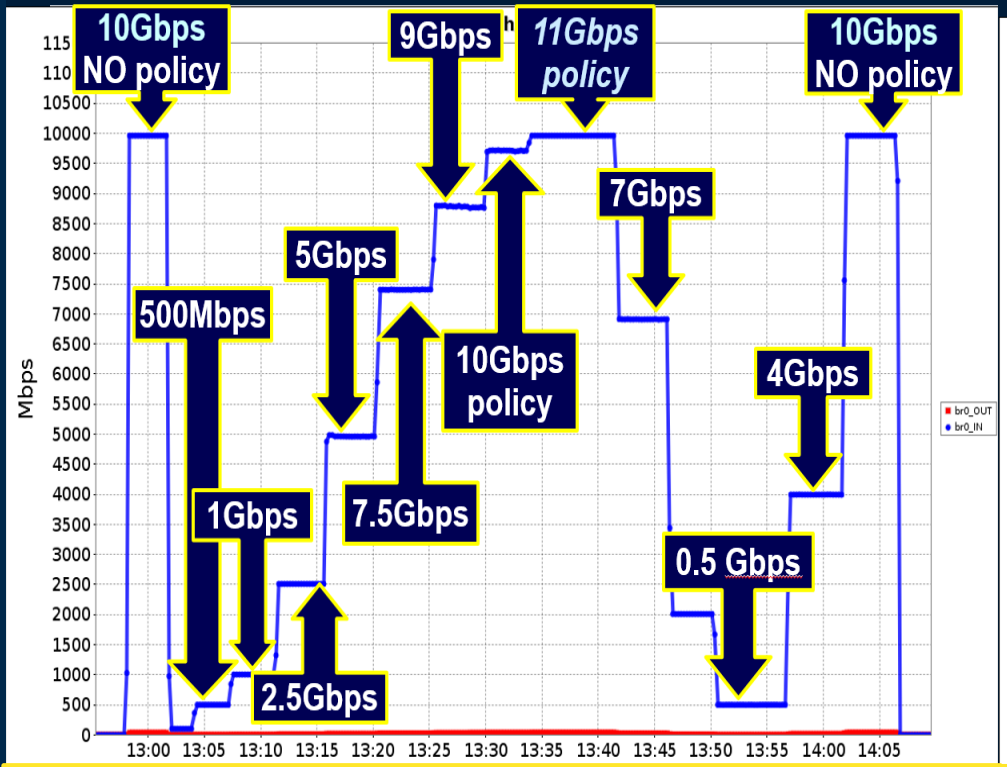
SC14: A Major "Big Data" Event
An Opportunity with the Attention of OSTP

- An opportunity to show our fields' capabilities in advanced networking, computing and frontier science
- Plus concepts and products that could benefit other fields, with a broad impact within and beyond research
- HEP examples:
  - Global scale grid systems
  - State of the art throughput: N X 100G in the wide area, and possibly 1 Tbps between the Caltech and JET booths with DWDM
    - FDT – already used "throughout the Big Data Community"
    - RFTP (with RDMA) – developed with the LBNL/ESnet team
  - Autonomous global monitoring: MonALISA
  - Multi-Layer Dynamic circuits: ANSE: Multipath PhEDEx Transfers
  - Software Defined Networking: Open Daylight Controller with application to Multipath TCP. Plus an ML-Based Orchestrator
- Previewing SC15 when we will have real data coming in

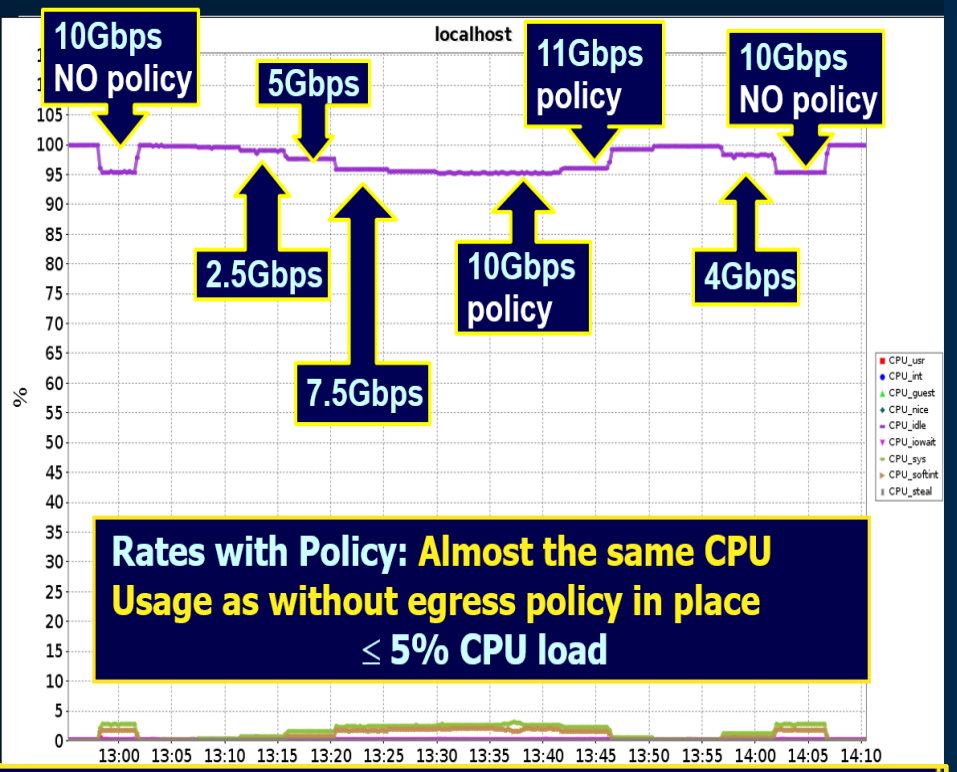
Site Orchestration: OVS Open vSwitch

Protocol independent, SDN managed smooth high-rate flows

OVS Dynamic bandwidth adjustment with Egress Rate-Limit



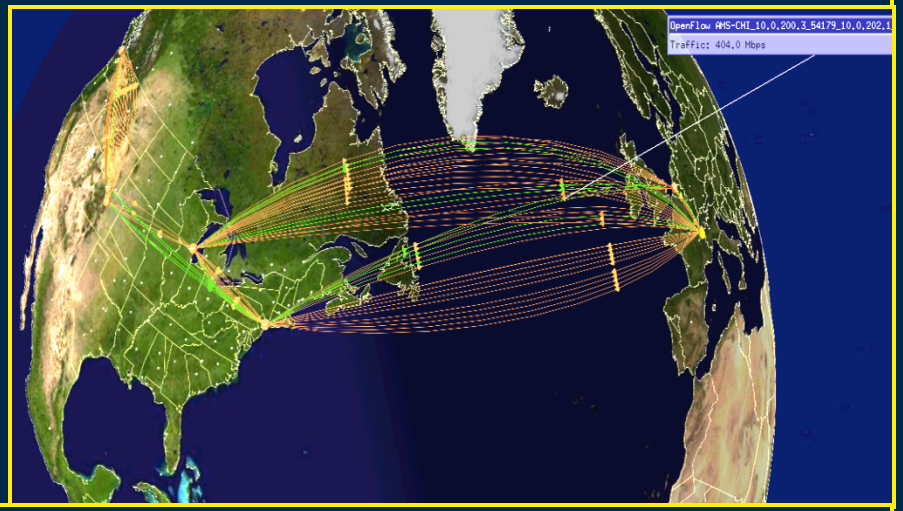
OVS Dynamic bandwidth adjustment Egress rate-limit– Sender CPU



- Smooth egress traffic shaping up to 10Gbps, and up to 7Gbps for ingress
- The CPU overhead is negligible when enforcing QoS

- More testing underway:
  - Ingress over 7 Gbps over long RTT paths
  - 40GE / 100 GE
  - Multiple QoS queues for Egress

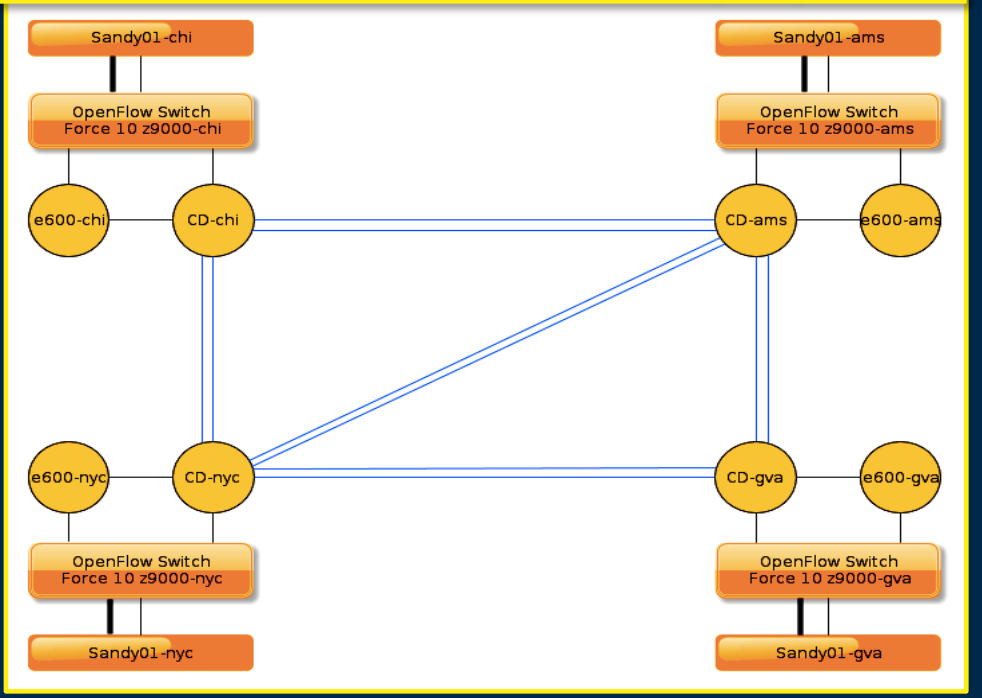
Caltech + Partners: OpenFlow Testbed Demo with MonALISA at SC13



- For SC13, US LHCNet's persistent OpenFlow testbed **was extended to U. Victoria in Canada and USP in Brazil**
- Showed efficient in-network load balancing managing big data transfers among multiple partners
- on three continents **using a single OpenFlow controller**
- Moving to OpenDaylight controller, supported by many vendors

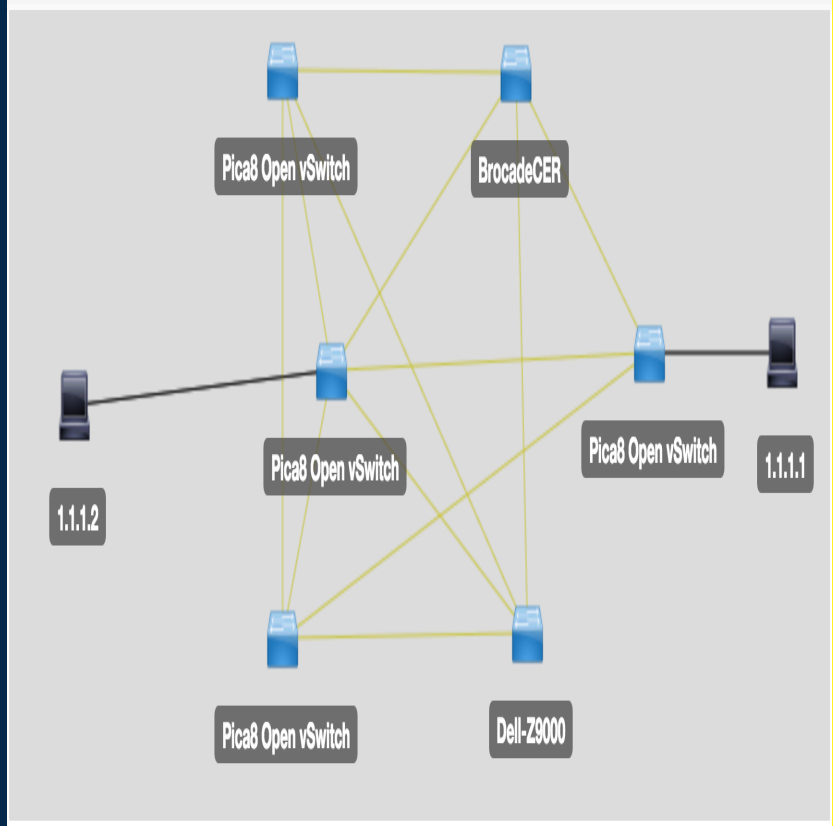
- **Bringing** Software Defined Networking Into *Production Across the Atlantic*

TA Testbed ➡ Production Deployment



- **Leading to** powerful intelligent interfaces between the LHC experiments' data management systems and the network
- **Generally useful:** will be integral to the OpenDaylight Controller

OpenDaylight + OpenFlow Multipath Controller + Testbed: Status and Plans



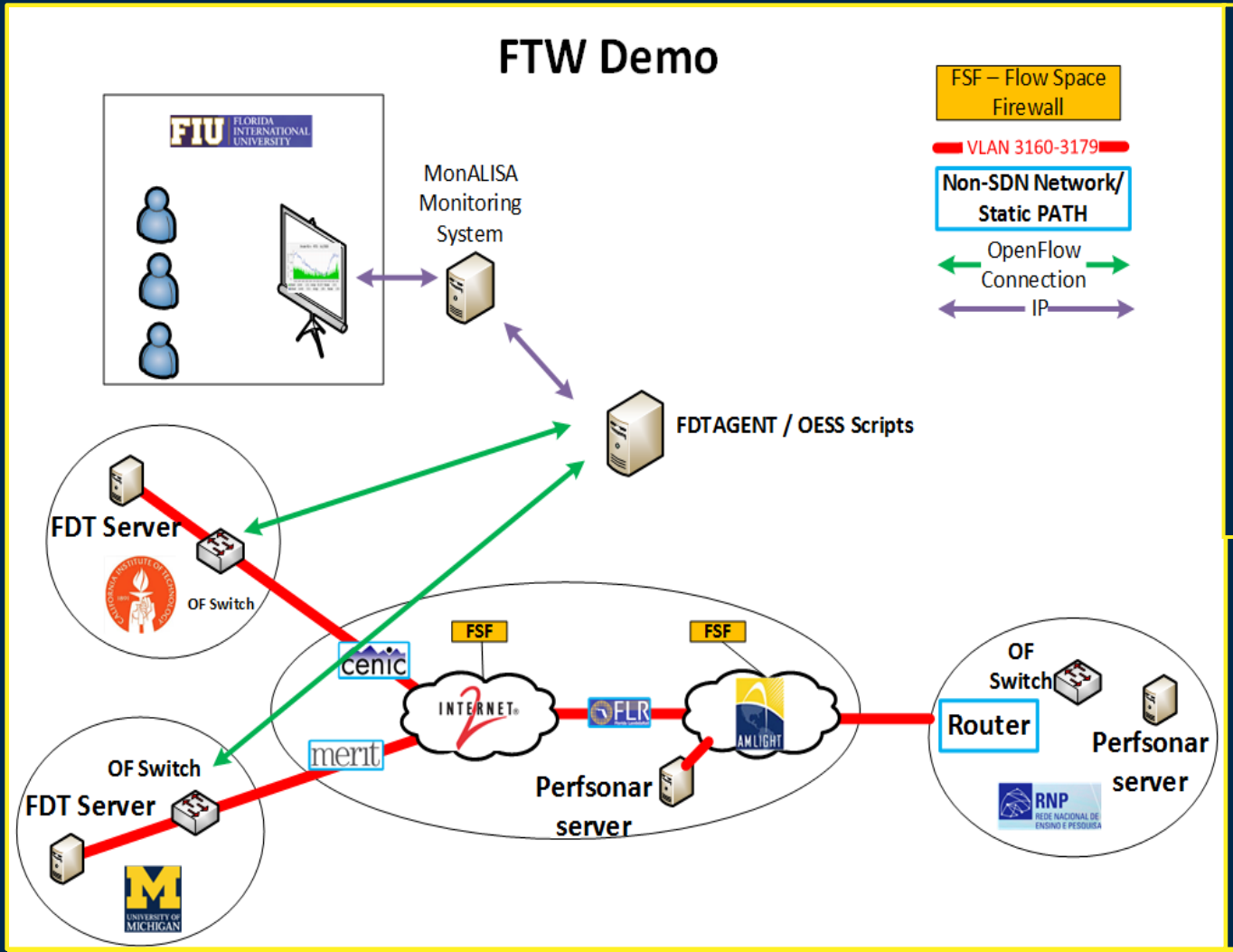
Testbed will be extended to other ANSE sites: **Vanderbilt, Michigan, UT Arlington, FIU, Sao Paulo, Rio, etc.** Then Fermilab, Amsterdam + CERN

Diagram shows current ODL testbed setup, with Six OpenFlow Switches:

Dell Z9000, Brocade CER, 4 Pica8's

- Investigate flow rule write-rates to Pica8, Brocade, Dell, etc. switches
- **Improve HostTracker behavior (auto-discovery of hosts)**
- Add Pro-active/Reactive mode switch to NorthBound interface
- **Continue development of new, enhanced selection algorithms**
- Work on interfaces to SRM/FTS (Find Source/Dest. IP Vectors using OVS)
- **Follow up with Internet2 on use of** our ODL controller on their AL2S footprint
- Moving to the OpenDaylight **Lithium Release**

FTW Demo: SDN-Driven Multipath Circuits
OpenDaylight/OpenFlow Controller Int'l Demo

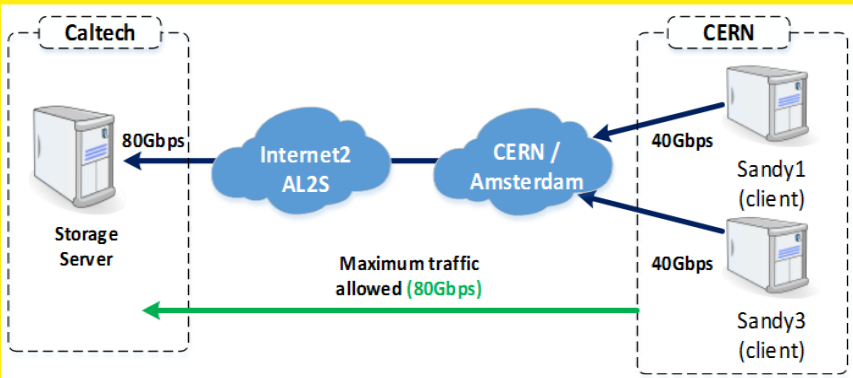


- A. Mughal, J. Bezerra
- *Hardened OESS and OSCARS installations at Caltech, Umich, AmLight*
- *Updated Dell switch firmware to operate stably with OpenFlow*

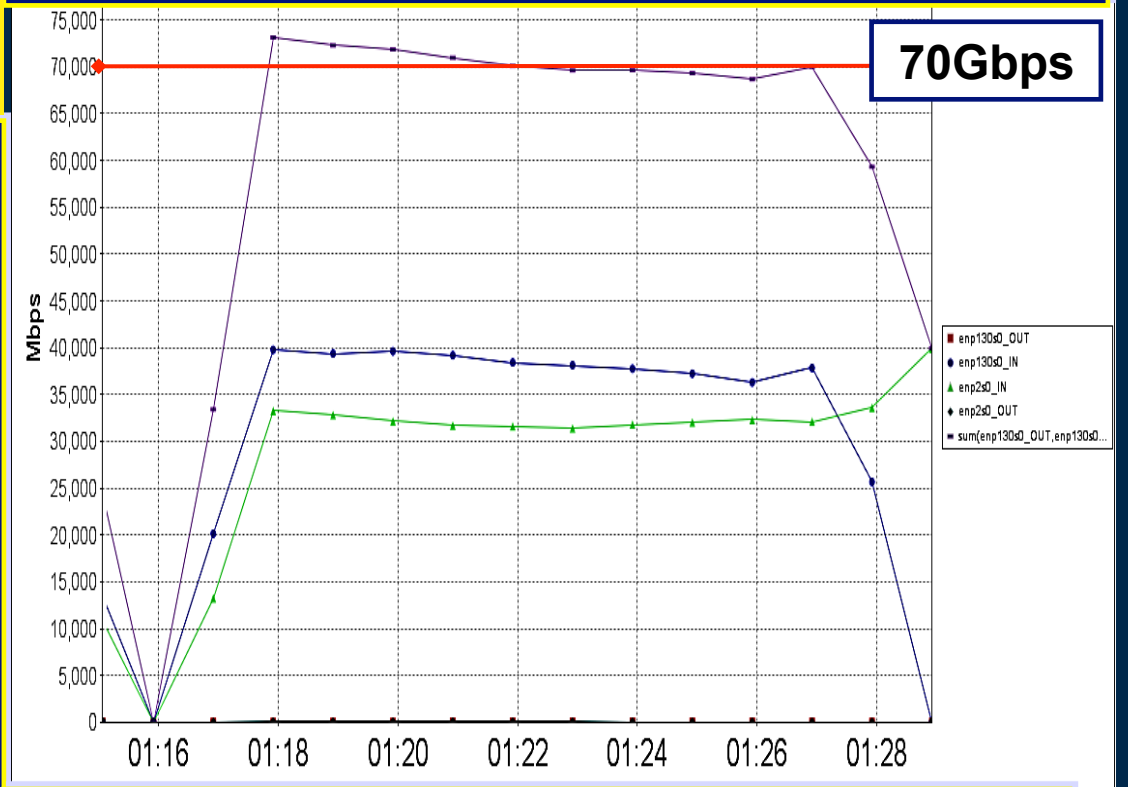
- **Dynamic circuit paths under SDN control**
- **Prelude to the ANSE architecture: SDN load-balanced, moderated flows**

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

Data Transfer Using RFTP (RDMA + FTP): July 16 2014



Two Servers at CERN Writing to One 2U Server at Caltech with 8 Next Generation (nVMe) SSDs and Two 40GE Network Interfaces



Configuration at Caltech

- 4 RFTP daemons listening at unique TCP ports

Configuration at CERN

- 8 RFTP instances running on two servers
- Two client RFTP instances at CERN connect with one RFTPD daemon at Caltech

[*] *RFTP:* by Stony Brook and LBNL

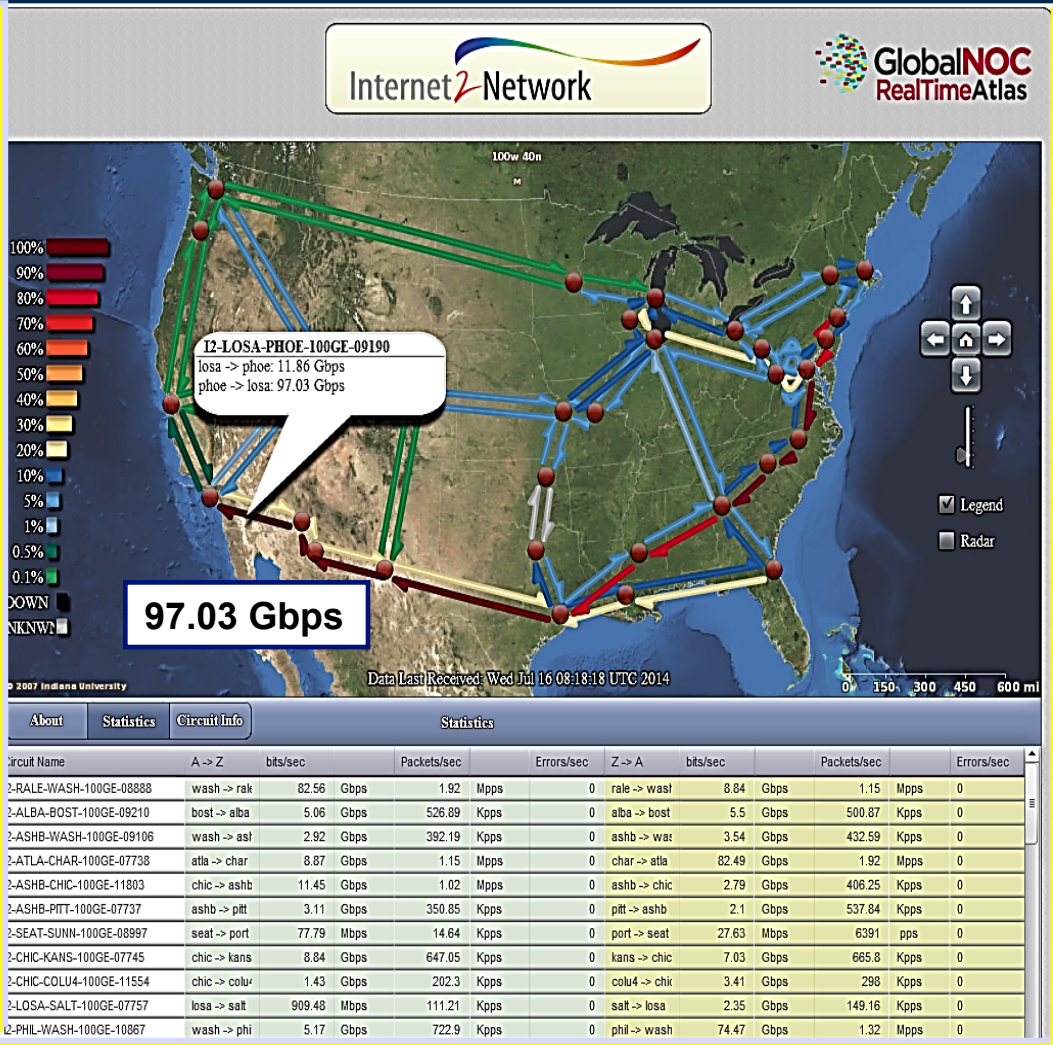
Next Step: Servers with 100GE NICs

Internet2 Network Map
AL2S Traffic Statistics

Traffic peak 97.03 Gbps Phoenix - LA observed during these transfers

A limiting factor on the traffic received at Caltech

Microbursts are often not reported by the monitoring clients



Circuit Name	A -> Z	bits/sec		Packets/sec		Errors/sec	Z -> A	bits/sec		Packets/sec		Errors/sec
I2-RALE-WASH-100GE-08888	wash -> rak	82.56	Gbps	1.92	Mpps	0	rale -> wash	8.84	Gbps	1.15	Mpps	0
I2-ALBA-BOST-100GE-09210	bost -> alba	5.06	Gbps	526.89	Kpps	0	alba -> bost	5.5	Gbps	500.87	Kpps	0
I2-ASHB-WASH-100GE-09106	wash -> ash	2.92	Gbps	392.19	Kpps	0	ashb -> was	3.54	Gbps	432.59	Kpps	0
I2-ATLA-CHAR-100GE-07738	atla -> char	8.87	Gbps	1.15	Mpps	0	char -> atla	82.49	Gbps	1.92	Mpps	0
I2-ASHB-CHIC-100GE-11803	chic -> ashb	11.45	Gbps	1.02	Mpps	0	ashb -> chic	2.79	Gbps	406.25	Kpps	0
I2-ASHB-PITT-100GE-07737	ashb -> pitt	3.11	Gbps	350.85	Kpps	0	pitt -> ashb	2.1	Gbps	537.84	Kpps	0
I2-SEAT-SUNN-100GE-08997	seat -> port	77.79	Mbps	14.64	Kpps	0	port -> seat	27.63	Mbps	6391	pps	0
I2-CHIC-KANS-100GE-07745	chic -> kans	8.84	Gbps	647.05	Kpps	0	kans -> chic	7.03	Gbps	665.8	Kpps	0
I2-CHIC-COLU4-100GE-11554	chic -> colu4	1.43	Gbps	202.3	Kpps	0	colu4 -> chic	3.41	Gbps	298	Kpps	0
I2-LOSA-SALT-100GE-07757	losa -> salt	909.48	Mbps	111.21	Kpps	0	salt -> losa	2.35	Gbps	149.16	Kpps	0
I2-PHIL-WASH-100GE-10867	wash -> phi	5.17	Gbps	722.9	Kpps	0	phil -> wash	74.47	Gbps	1.32	Mpps	0

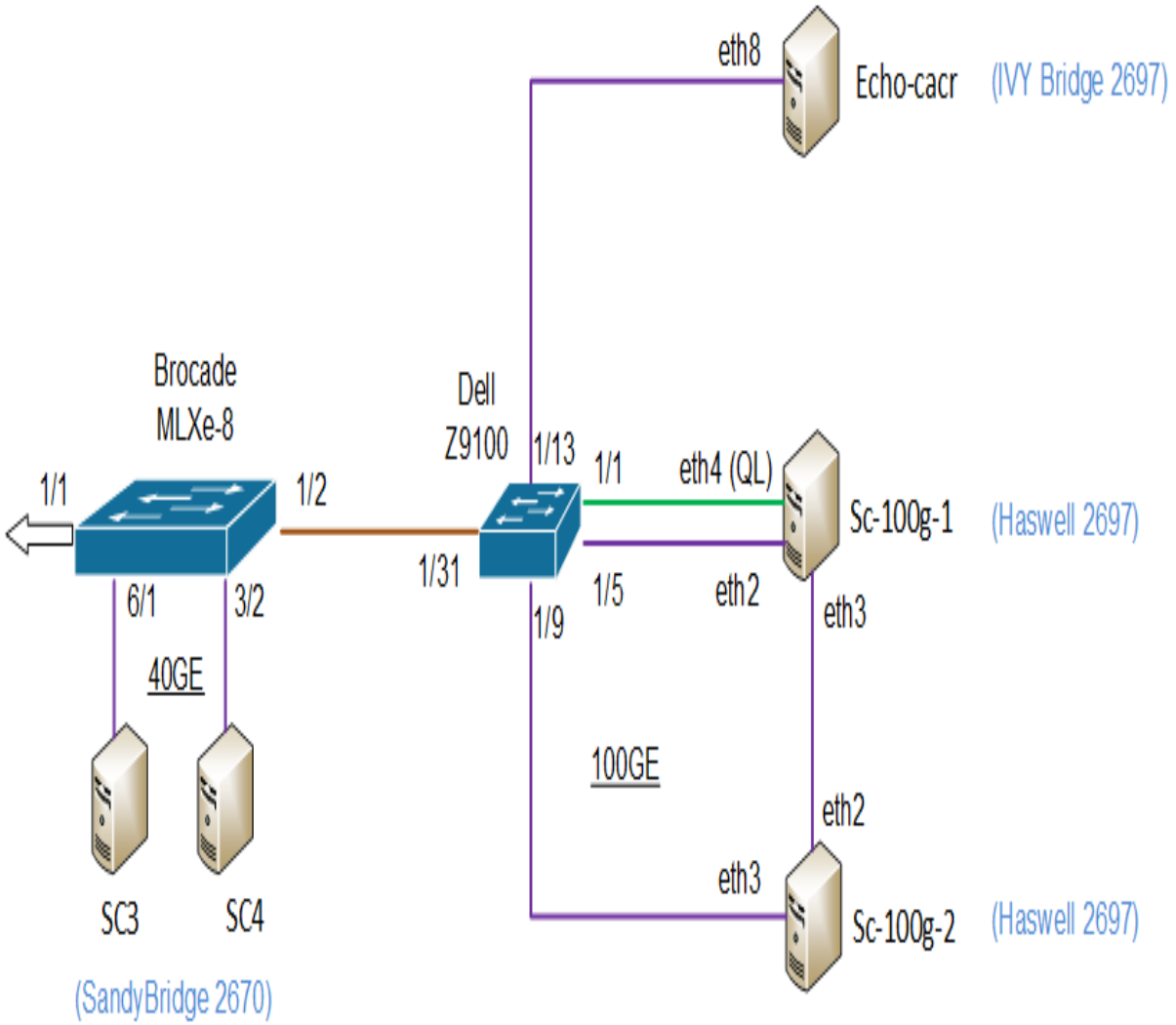
Plan now for Consistent Software Driven Operations

On ESnet and Internet2 covering both universities and Labs

In the US and to CERN via EEX

Message: At this level of capability, we need to control our network use, to prevent saturation as we move into production

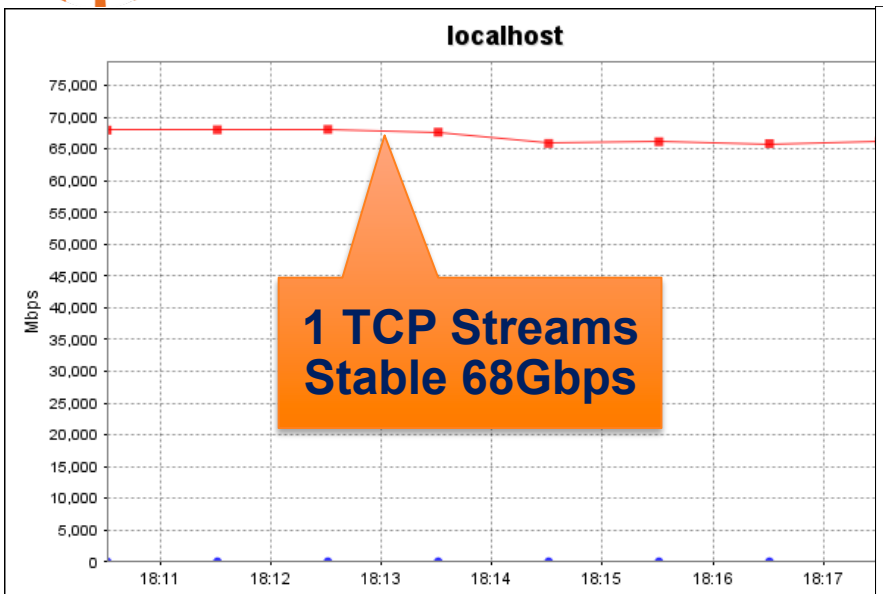
DTN: 100G network tests

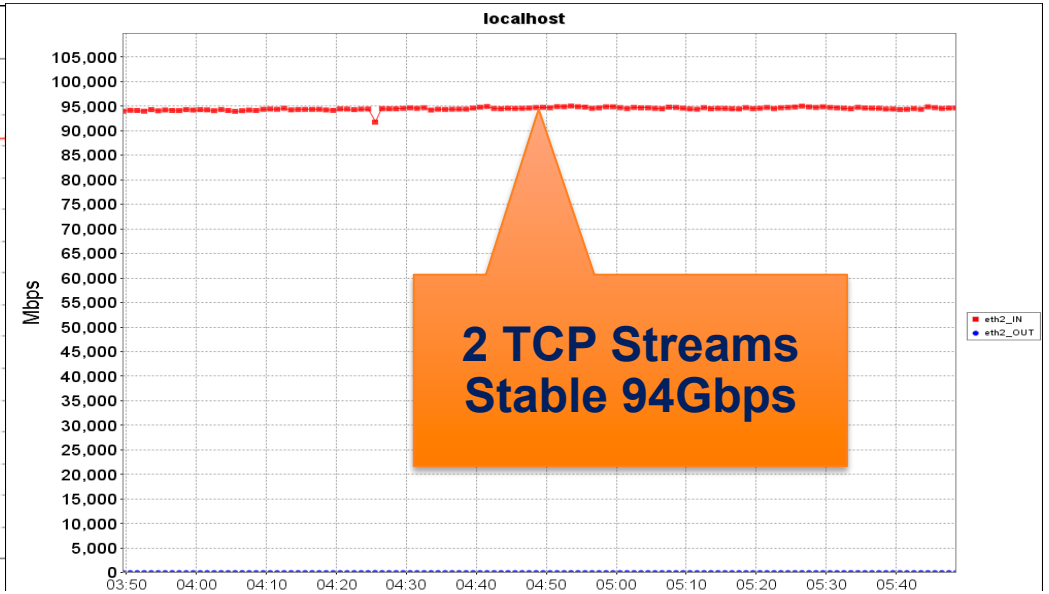


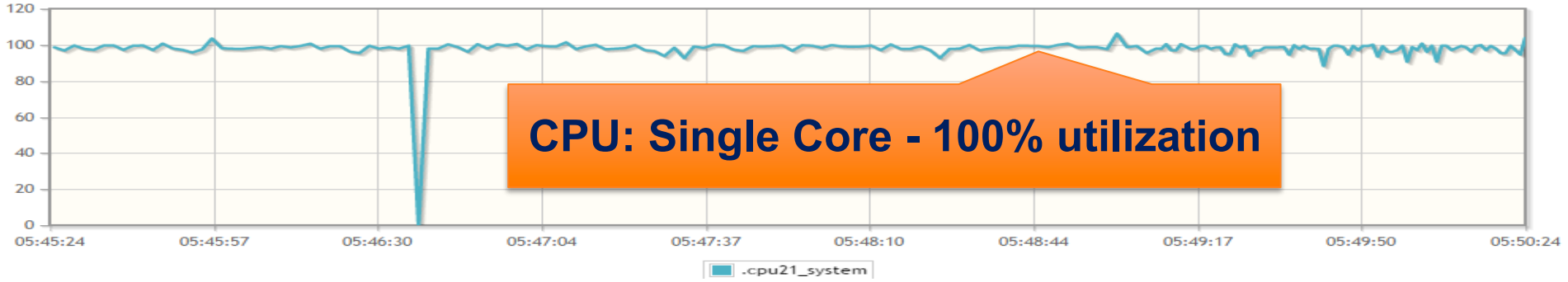
Network Topology:
Servers used in tests: Sc-100G-1 and Sc100G-2

- 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





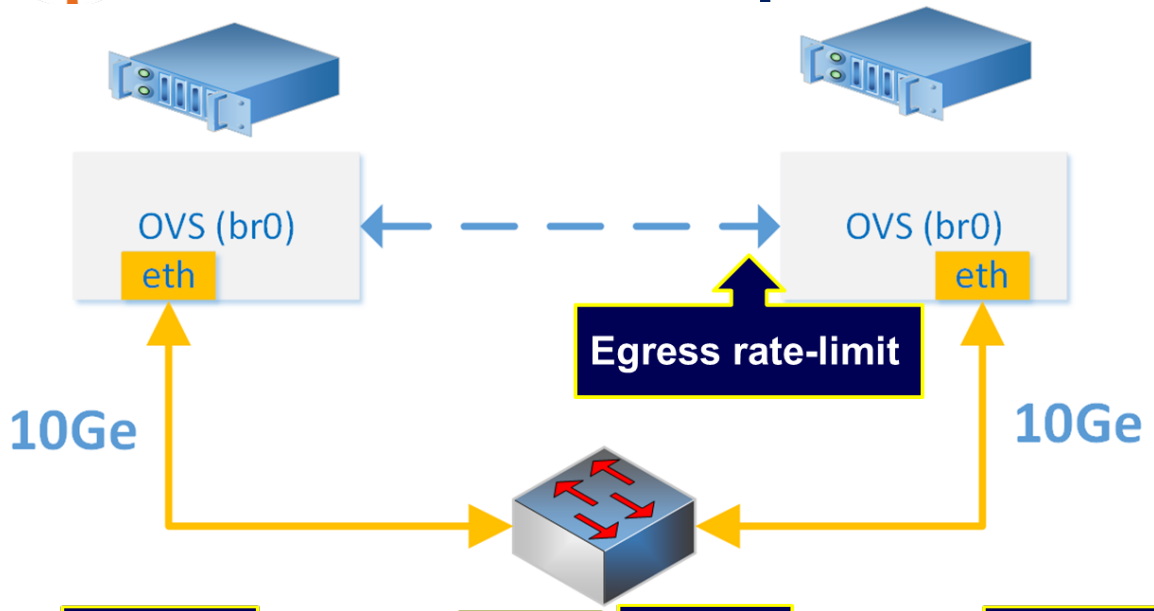


Client #numactl --physcpubind=20 --localalloc java -jar fdt.jar -c 1.1.1.2 –nettest -P 1 -p 7000

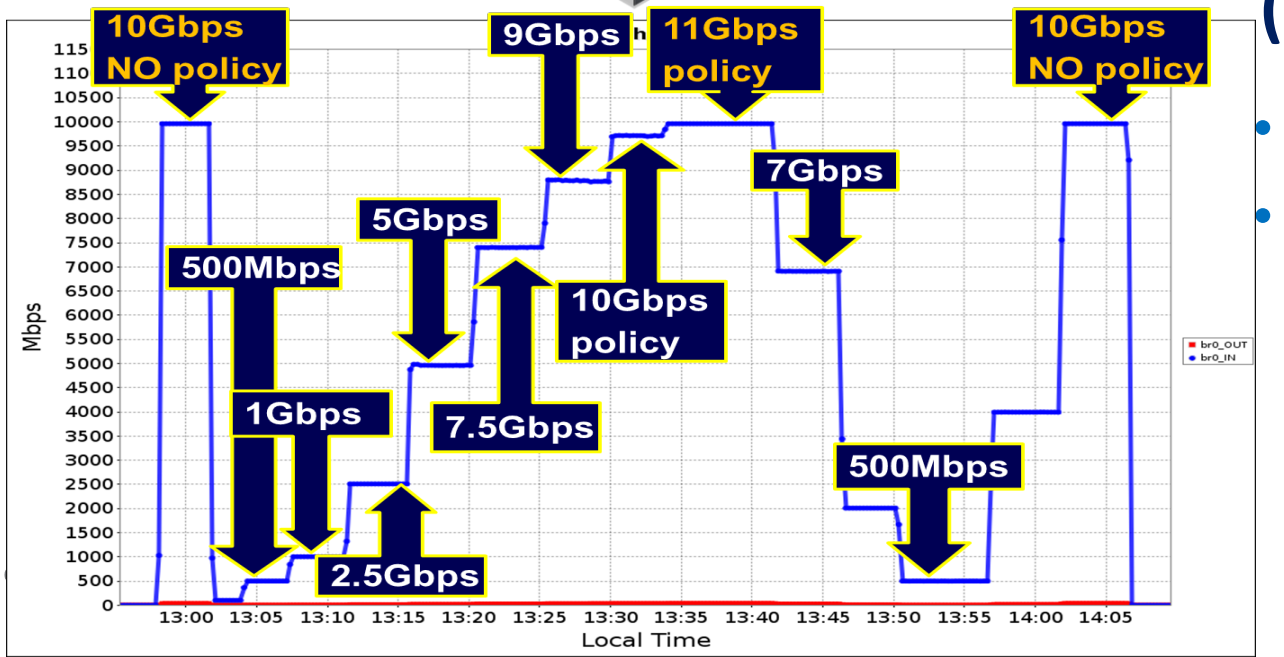
Server #numactl --physcpubind=20 --localalloc java -jar fdt.jar -p 7000

Line rate (100G) with 4+ TCP Streams

SDN QoS: Traffic Shaping with Open vSwitch (OVS)

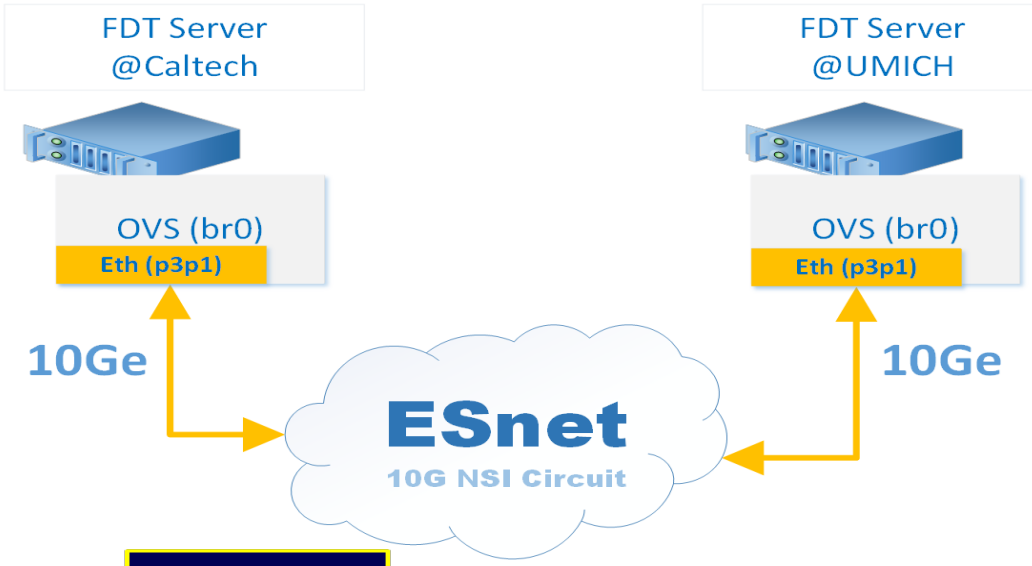


- OVS 2.3.1 with stock RH 6.x kernel
- OVS bridged interface achieved same performance as hardware (10Gbps)



- 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

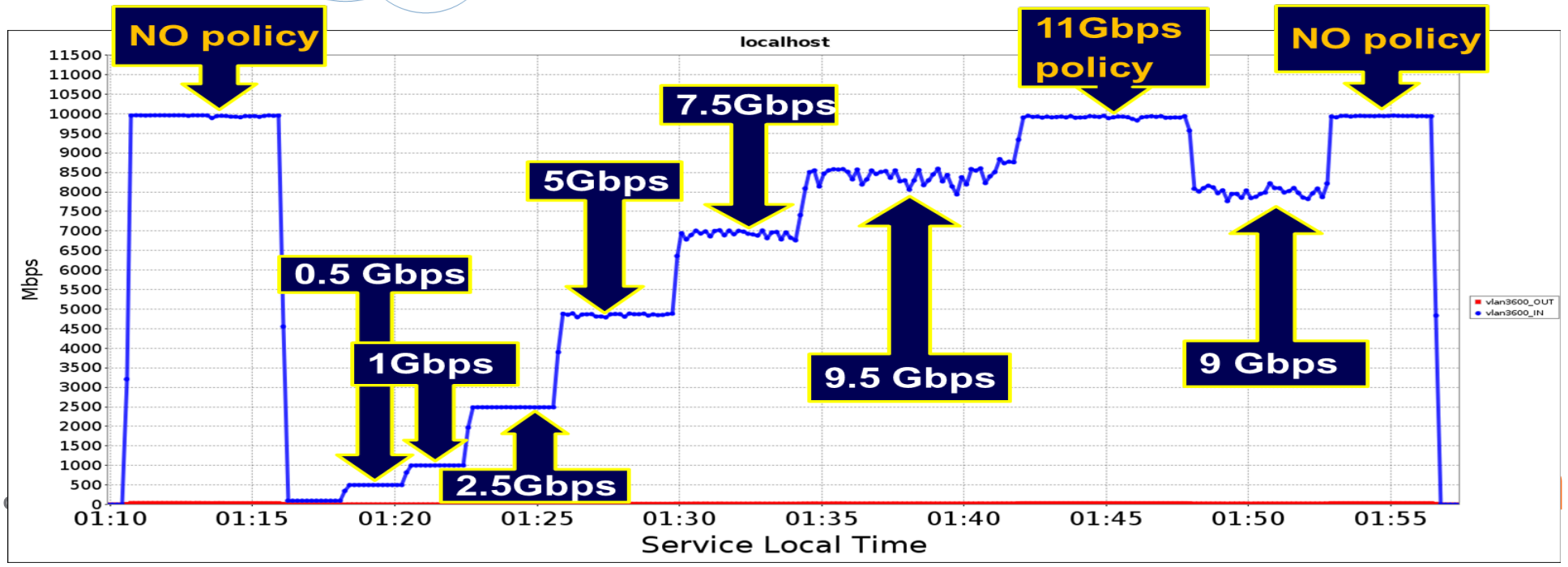


OVS 2.4 with stock kernel

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

Monitoring the Worldwide LHC Grid

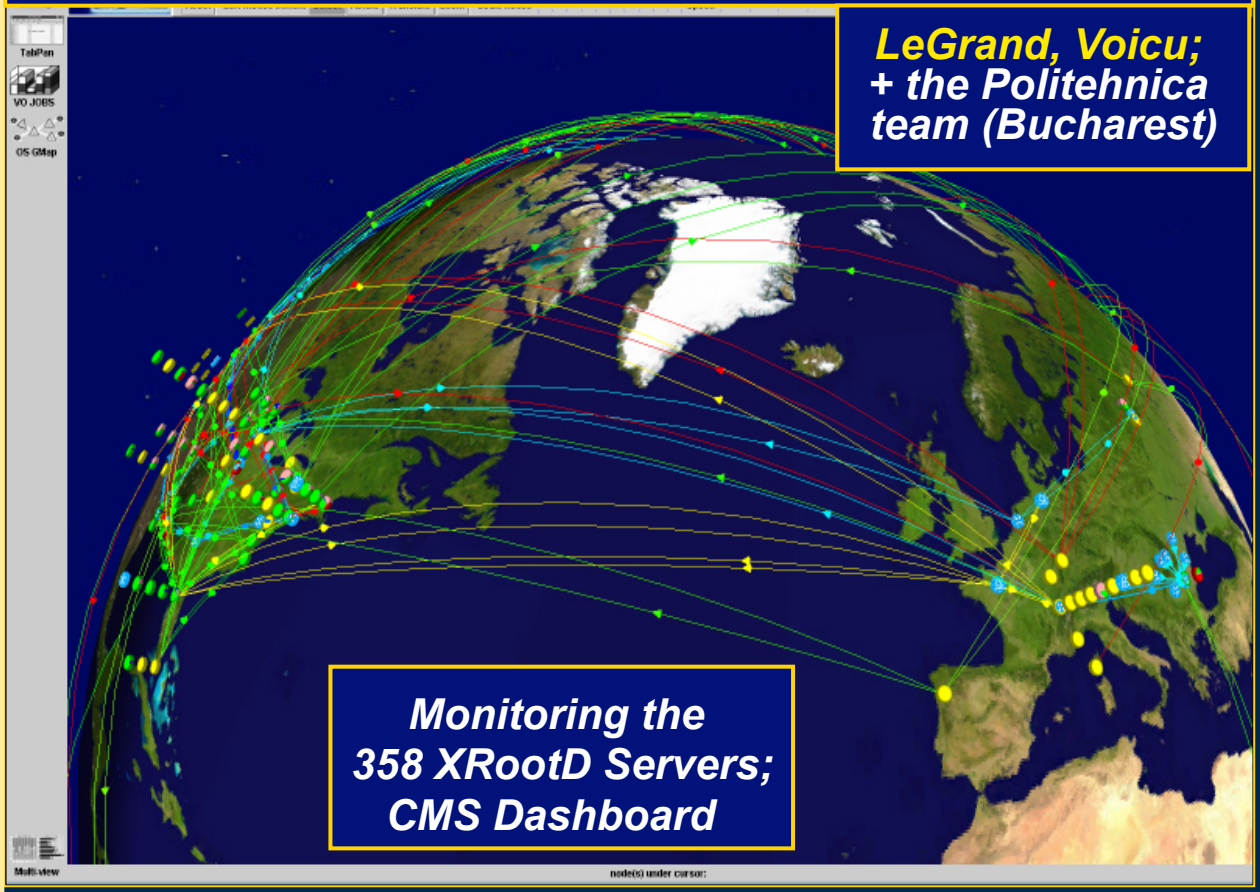
State of the Art Technologies Developed at Caltech

MonALISA Today
Running 24 X 7 at 370 Sites

- Using Intelligent Agents
- Resilient: MTBF >7 Years
- Monitoring
  - 60,000 computers
  - > 100 Links On Major R&E Networks, 14,000 end-to-end paths
- Tens of Thousands of Grid jobs running concurrently
- Collecting 6M persistent and 100M volatile parameters at 35 kHz in real-time
- 10^{12} parameter values served to CMS and ALICE

MonALISA: Monitoring Agents in a Large Integrated Services Architecture

Unique Global Autonomous Realtime System



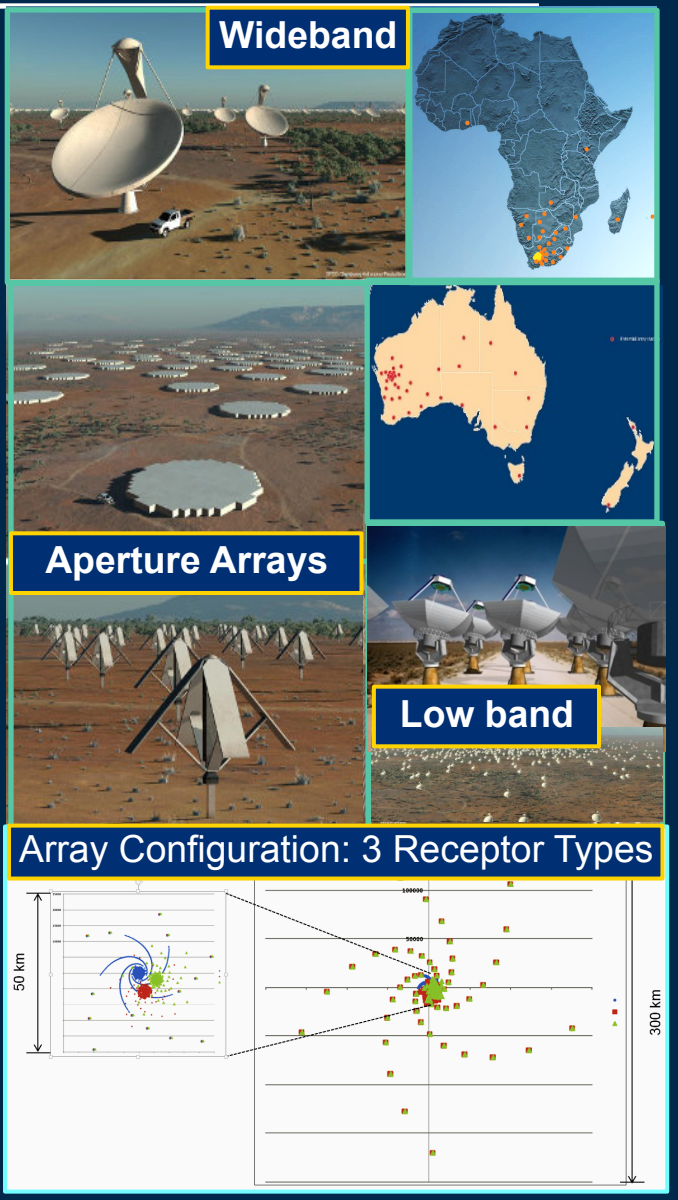
LeGrand, Voicu; + the Politehnica team (Bucharest)

Monitoring the 358 XRootD Servers; CMS Dashboard

SKA: a 1M km2 Radiotelescope

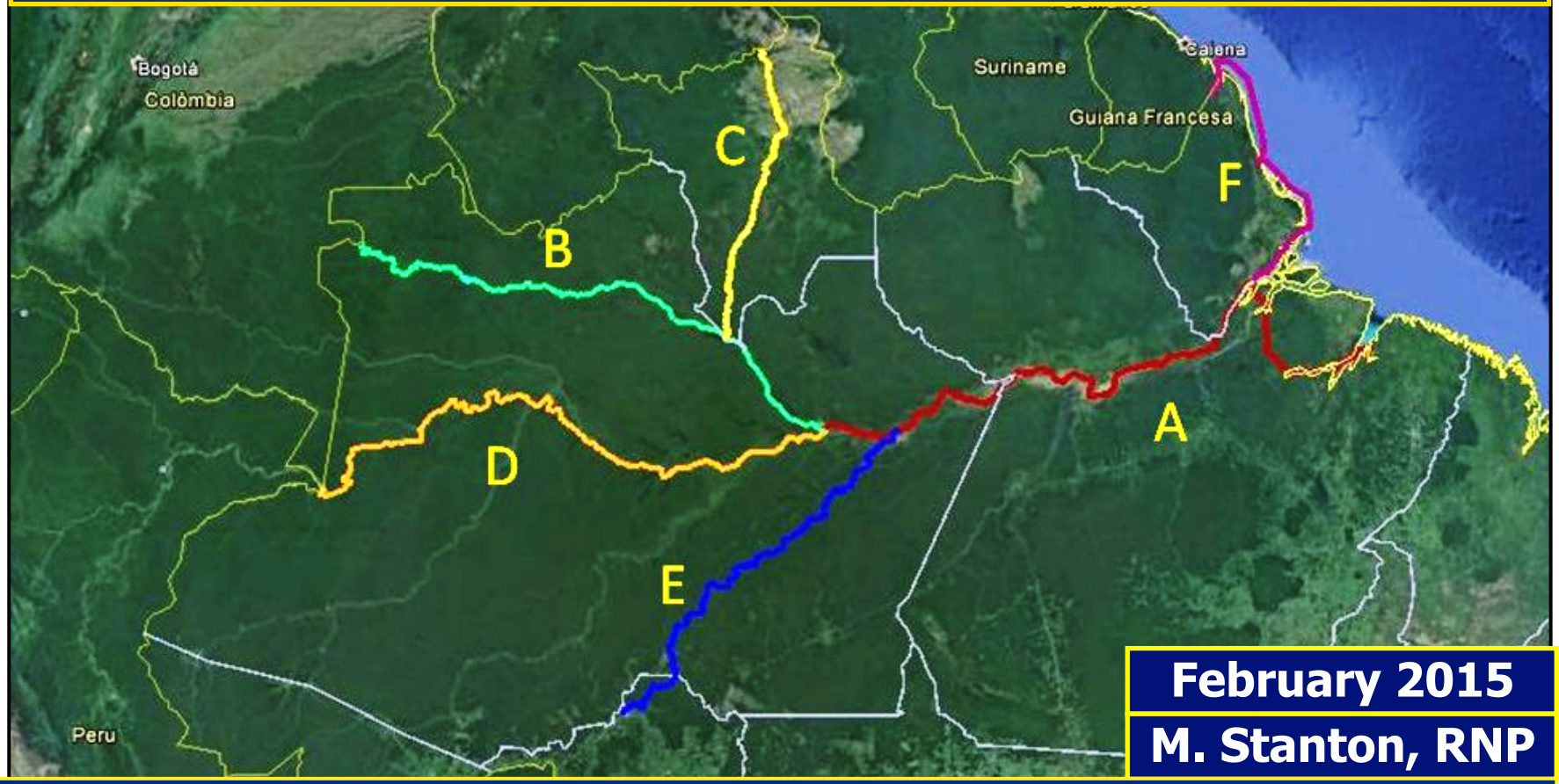
3000 Arrays operating as one instrument

- SKA: ~50 times more sensitive and a million times faster in sky scans than the largest currently operational radiotelescopes
- ~3000 antennae arrays spread over almost a million square kilometers: 80% within 100,000 km^2 of the main South African and Australian sites; Collection area itself is ~1 square kilometer
- An impressive range of Science:
  - Probing the Dark Ages: formation of first structures
  - Galaxy Evolution, Cosmology and Dark Energy
  - Origin and Evolution of Cosmic Magnetism
  - Strong Field Tests of Gravity: Pulsars + Black Holes
  - Cradle of Life: probing the full range of astrobiology
- SKA will generate *~15,000 Terabits/sec* of data that must be moved over networks



Brazil: RNP proposal for cables along major rivers in the north

- Complementing existing fiber infrastructure
- Pilot along Route D may be feasible in 2015



Possible major routes for subfluvial fiber optic cables.
Rivers: A: Amazon; B: Negro; C: Branco; D: Solimões (upper Amazon), E: Madeira; F: Maritime route to French Guiana.