



# SDN for the Large Hadron Collider

**Dorian Kcira**

Caltech CMS Group  
[hep.caltech.edu/cms](http://hep.caltech.edu/cms)

**"Roadmap to Operating SDN-based Networks" Workshop  
Berkeley, California, July 14-16, 2015**



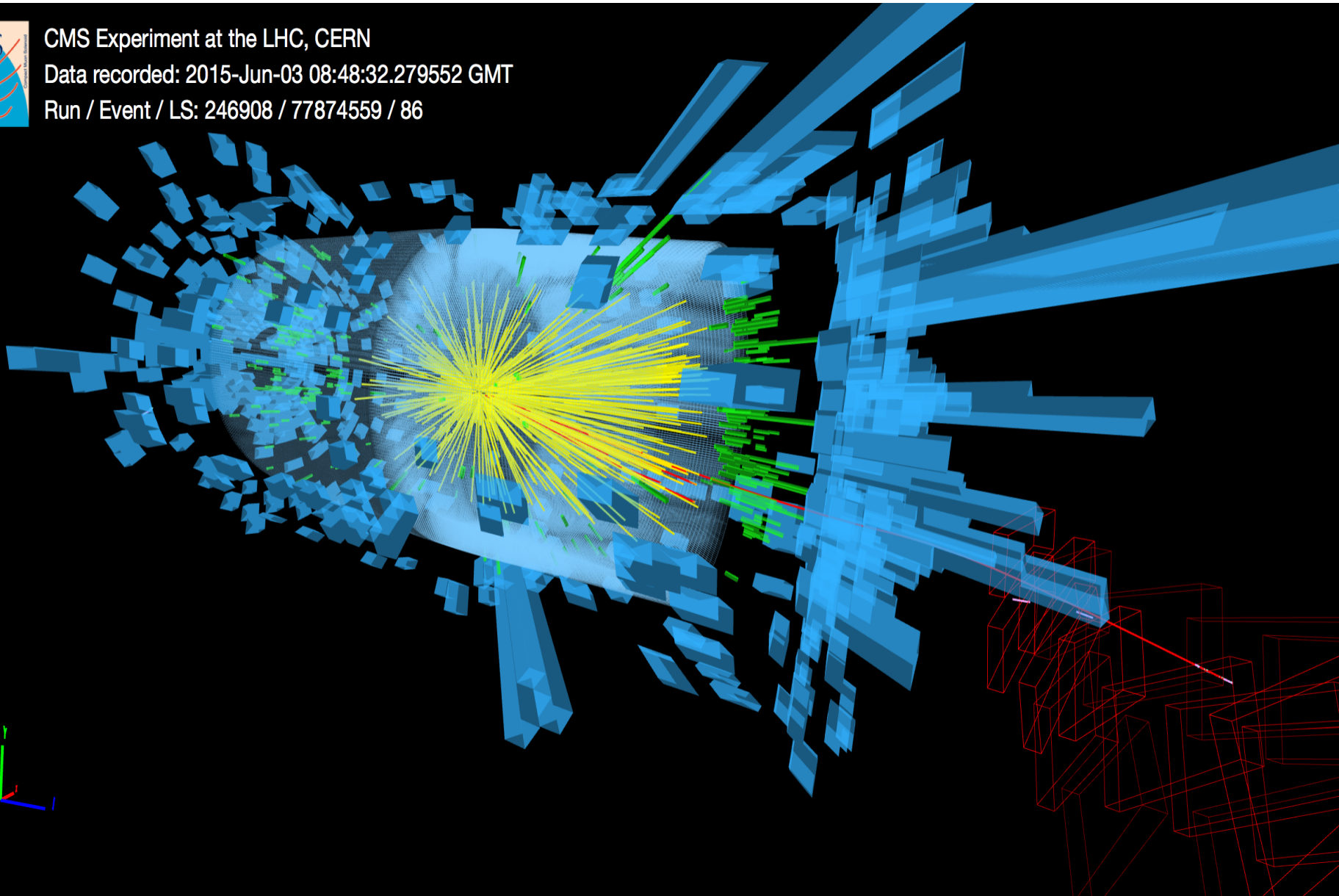
# LHC Collisions at 13 TeV



CMS Experiment at the LHC, CERN

Data recorded: 2015-Jun-03 08:48:32.279552 GMT

Run / Event / LS: 246908 / 77874559 / 86





# Scale of LHC Network Requirements

Proven performance and high reliability are required

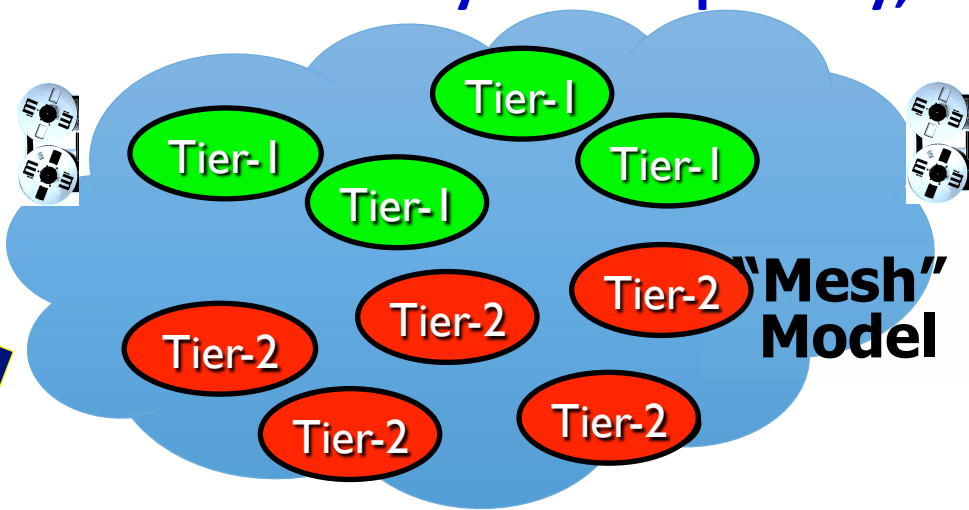
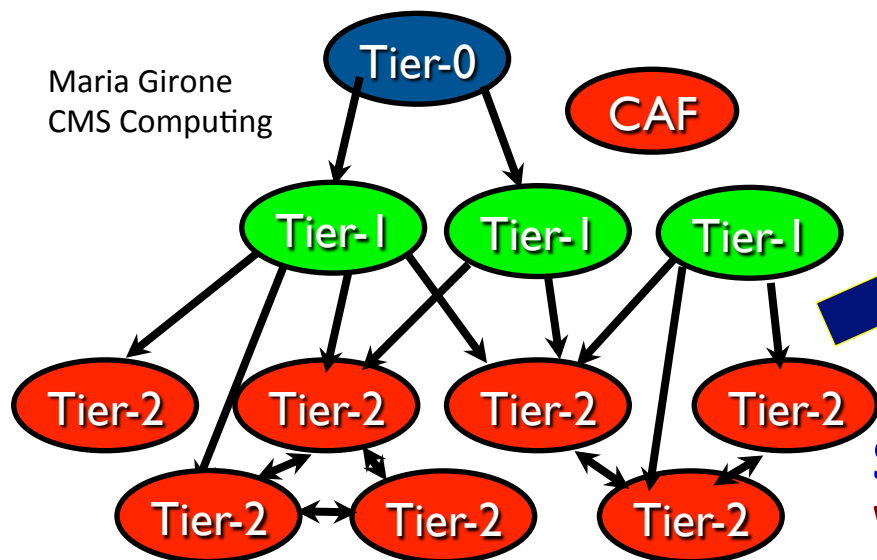
- **A recent conservative baseline estimate given recently is: A factor of ~2 between 2014 and 2017** [This is known to be low]
- **Other bandwidth growth projections and trends are larger, so we need to propose a flexible solution; and better estimates**
- **CMS at recent Esnet requirements workshop states “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)**
- **Case Study of CMS Physics Analysis Needs using location independent “cloud style” data access (AAA) showed: A factor of ~5 within next 5 yrs: 100G Target for each Tier2**
- **Longer Term Trends: Discussion at Snowmass, and long term projections by ESnet showed how 50-100X growth in network needs for multiple fields by ~2020 is**



# 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!!

Maria Girono  
CMS Computing



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

+ Elastic access of from some Tier2/Tier3 sites



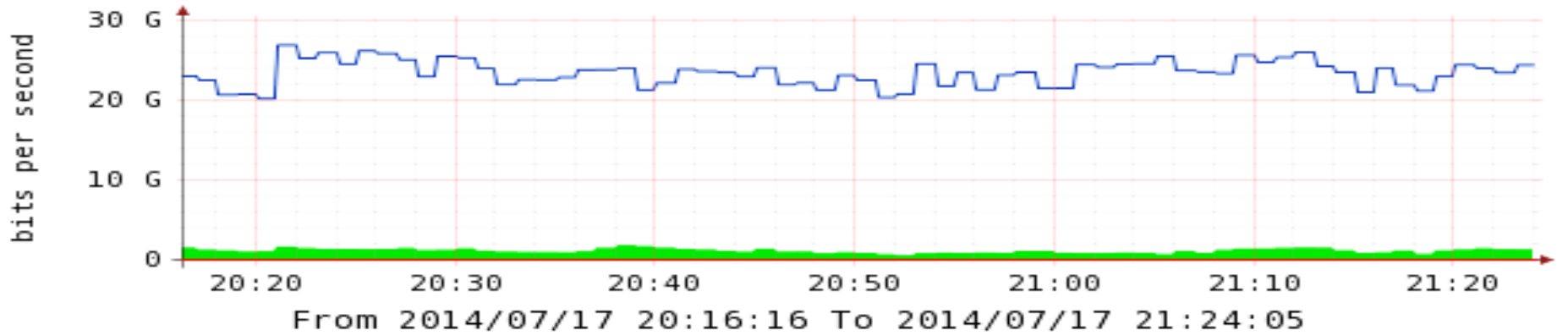
# ANA-100 Link in Service July 16 2014



## Transfer Rates: Caltech Tier2 to Europe July 17

A. Mughal, S. Cury

- Peak upload rate: **26.9 Gbps**
- Average upload rate over 1h of manual transfer requests : 23.4 Gbps
- Average upload rate over 2h (1h manual+ 1h automatic) : 20.2 Gbps
- Peak rate to CNAF alone: **20 Gbps**
- Now: 20 Gbps Milestone for US CMS Tier2 Sites with 100G
- By SC14: Up to ~50 Gbps of production traffic from Caltech Tier2;  
12-40 Gbps Routine for US Tier2s as of Now
- Downloading Terabyte Datasets to Tier3s and Tier4s (to the desktop/laptop) is a use case being explored by CMS (Vlimant)

BrocadeMLXe8 - Traffic - e1/1 100G Uplink To CENIC



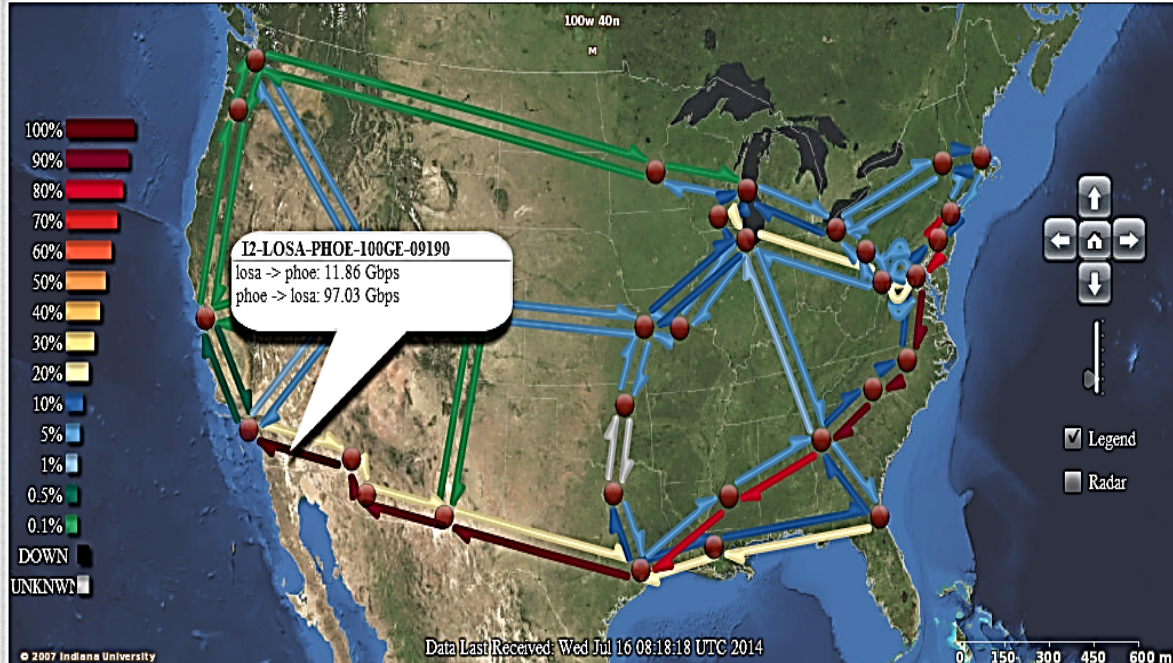
	Inbound	Current:	1.07G	Average:	1.01G	Maximum:	1.67G
	Outbound	Current:	24.40G	Average:	23.41G	Maximum:	26.89G
Graph Last Updated: Thu 17 Jul 21:27:01 PDT 2014							

PROTOCOL / TOR8.0ETIKER



# Internet2 Network Map AL2S Traffic Statistics

A. Mughal



© 2007 Indiana University

About Statistics Circuit Info

Statistics

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 -> wasl	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 -> colu	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-PHI-WASH-100GE-10867	wash -> phi	5.17 Gbps	722.9 Kpps	0	phi -> wash	74.47 Gbps	1.32 Mpps	0

- Traffic peak 97.03 Gbps Phoenix - LA observed during Caltech-CERN transfers
- A limiting factor on the traffic received at Caltech
- Microbursts are often not reported by the monitoring clients
- Plan now for Consistent Software Driven Operations
- On ESnet and Internet2 covering both universities and Labs
- In the US and to CERN via EEX and ANA-200G
- At this level of capability, we need to control our network use, to prevent saturation as we move into production



# SDN Multipath OpenDaylight Demonstrations at Supercomputing 2014

## 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

- 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



## SC14 Demo: Caltech/iCAIR/Vanderbilt OF links

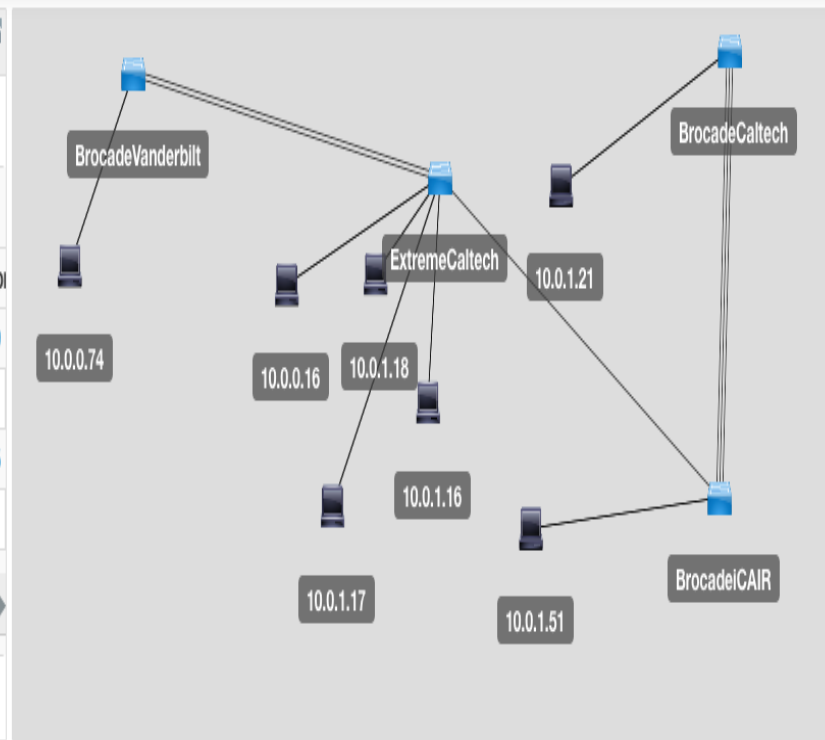
Nodes Learned

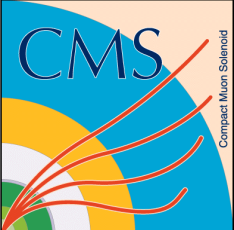
Nodes Learned

Search

Node Name	Node ID	Port
ExtremeCaltech	OF 00:00:00:04:96:6d:64:06	29
BrocadeVanderbilt	OF 00:24:38:7b:c5:00:00:00	11
BrocadeiCAIR	OF cc:4e:24:19:77:00:00:00	15
BrocadeCaltech	OF 00:24:38:7d:9b:00:00:00	6

1-4 of 4 items Page 1 of 1



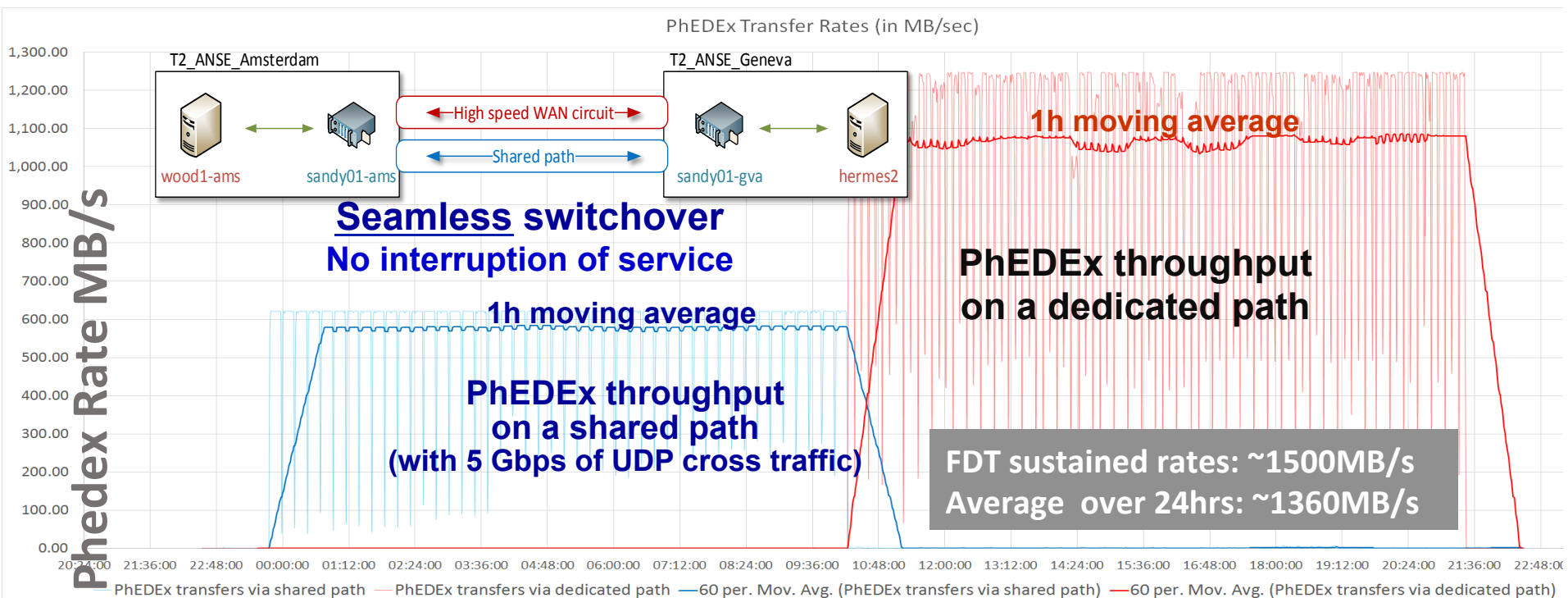


# PhEDEx and Dynamic Circuits

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



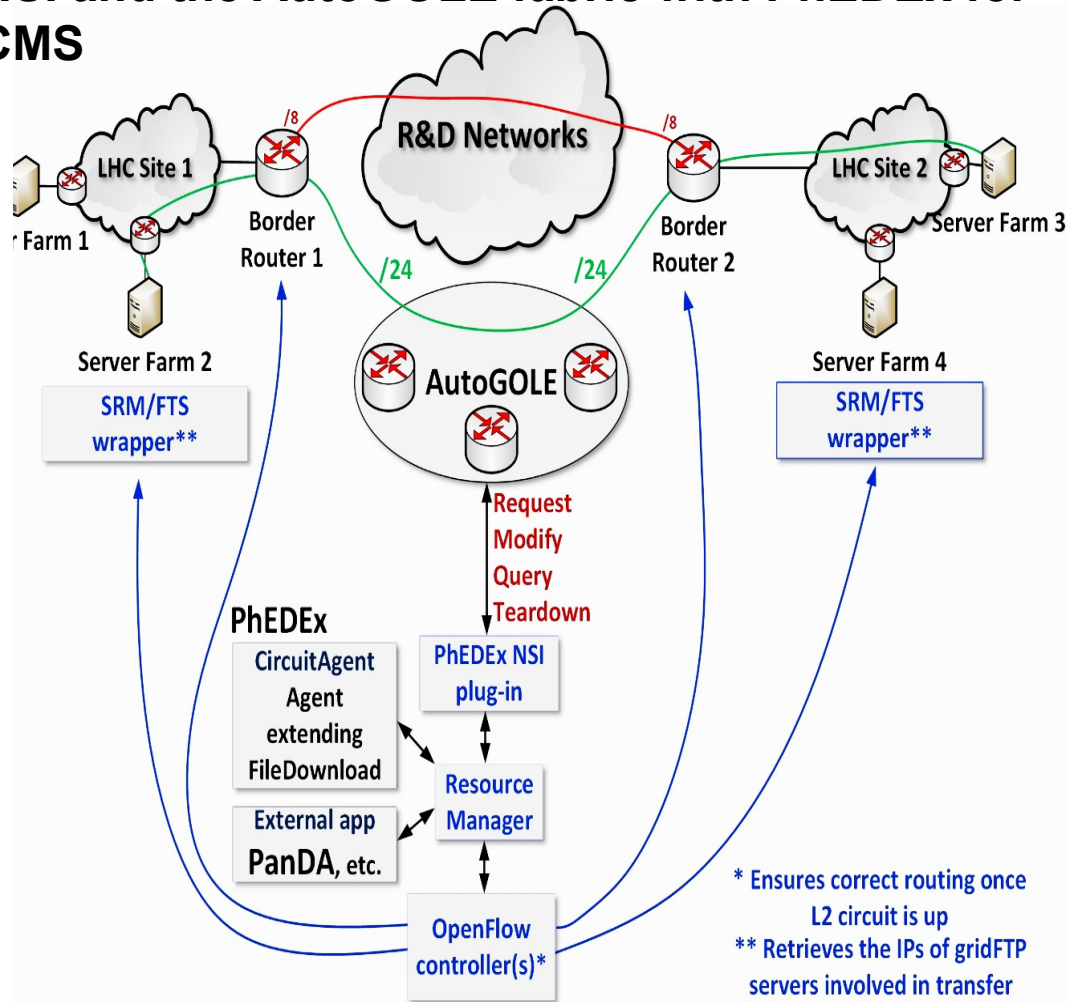




# Advanced Network Services for Experiments (ANSE) Integrating PhEDEx with Dynamic Circuits for CMS

Lapatadescu, Wildish, Mughal, Bunn,  
Voicu, Legrand, Newman

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



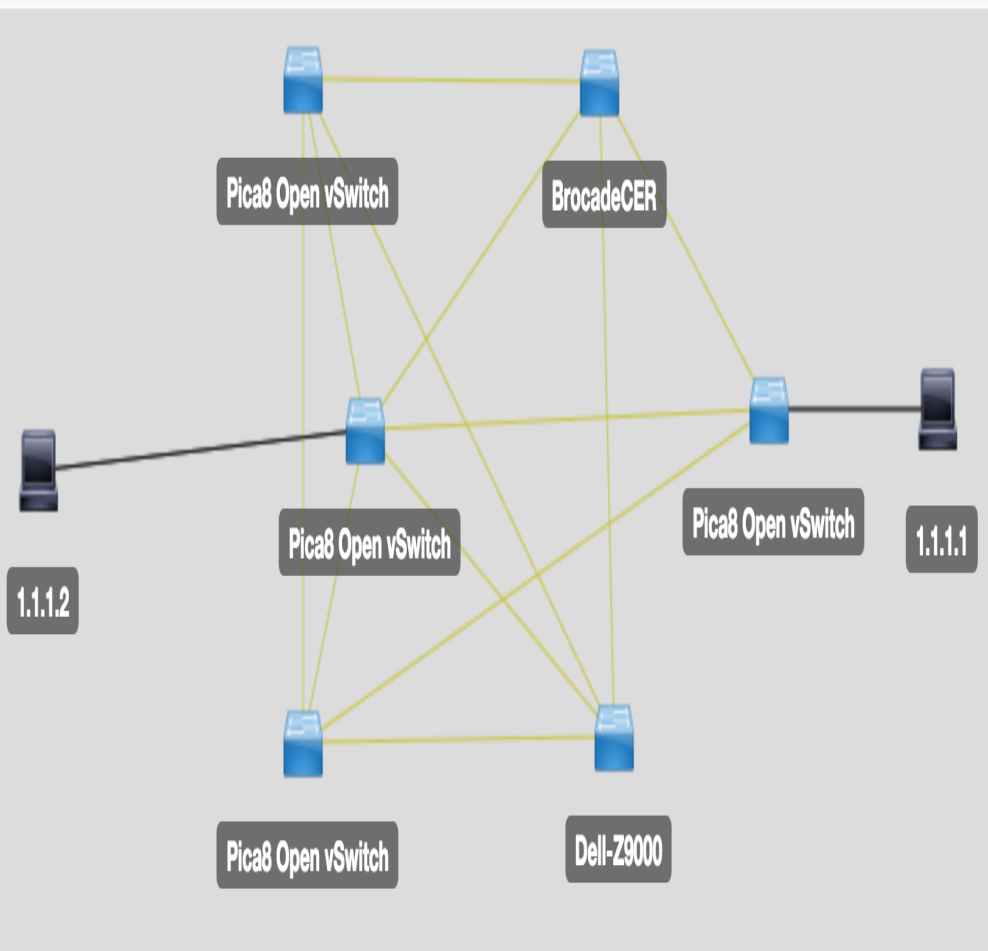
\* Ensures correct routing once  
L2 circuit is up  
\*\* Retrieves the IPs of gridFTP  
servers involved in transfer

- Building on the AutoGOLE Fabric of Open Exchange Points and **NSI emerging standard virtual circuits** (Demoed at SC2012)
- 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**





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



- **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** in context of ANSE (Vectors of Source/Dest. IPs)
- **Move to the OpenDaylight Lithium Release**

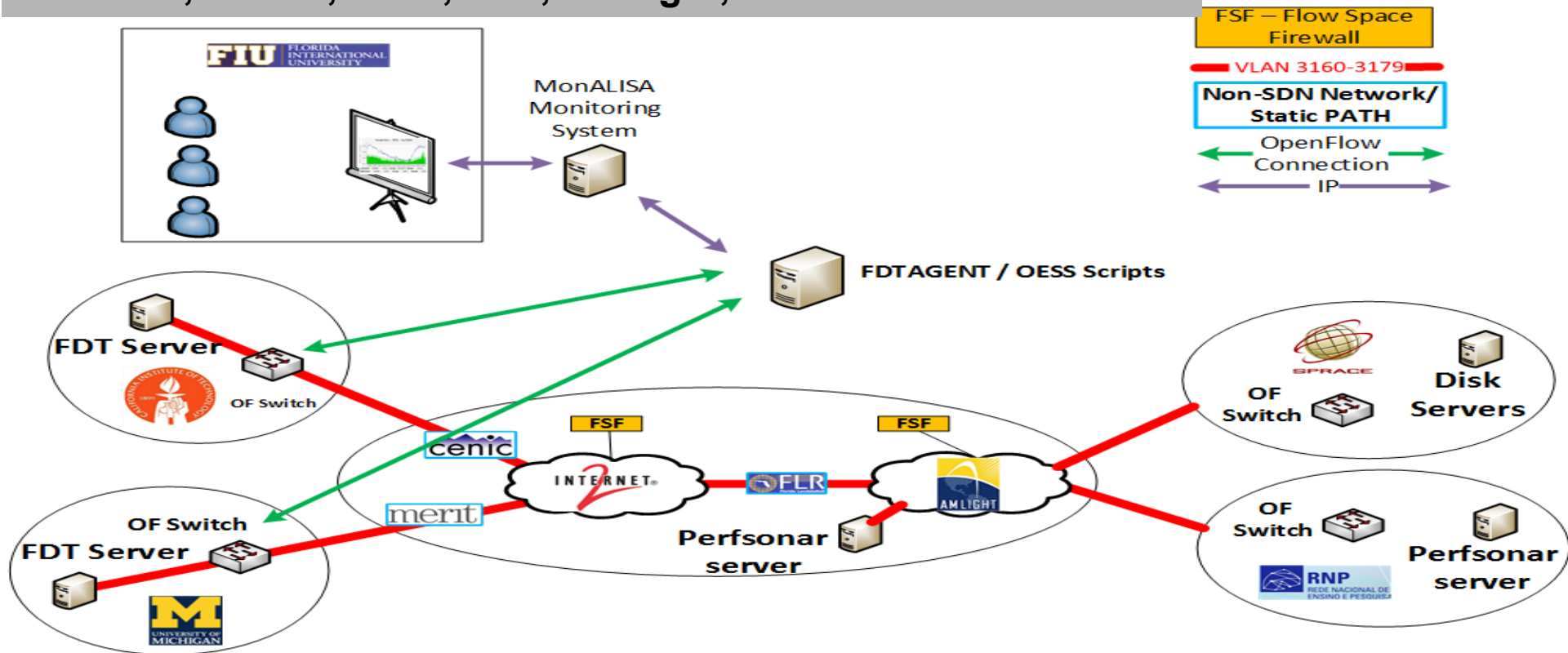
**Testbed will be extended to other ANSE sites: Vanderbilt, Michigan, UT Arlington, FIU, SPRACE, Rio, etc. Then Fermilab, Amsterdam + CERN.**



# Focused Technical Workshop Demo 2015: SDN-Driven Multipath Circuits OpenDaylight/OpenFlow Controller Demo

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

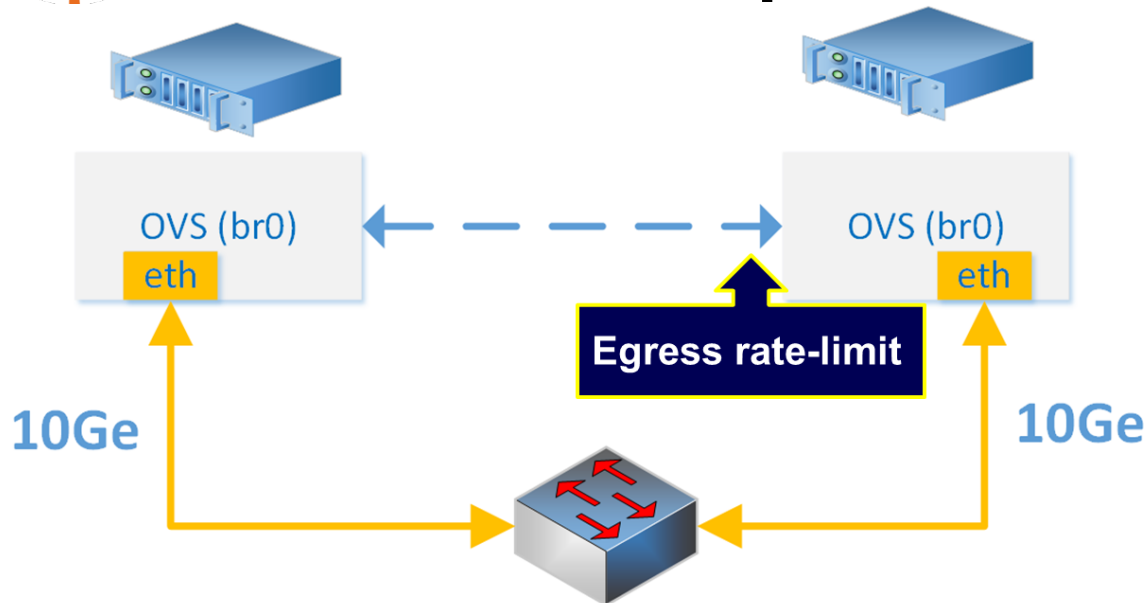
- **Dynamic Path creation:**
  - Caltech – Umich
  - Caltech/Umich - SPRACE
  - Caltech – RNP
  - Umich – AmLight
- **Path initiation by the DYNES FDT Agent using OSCARS API**
- **OESS** for OpenFlow data plane provisioning over Internet2/AL2S
- **MonALISA** agents at the end-sites provide detailed monitoring



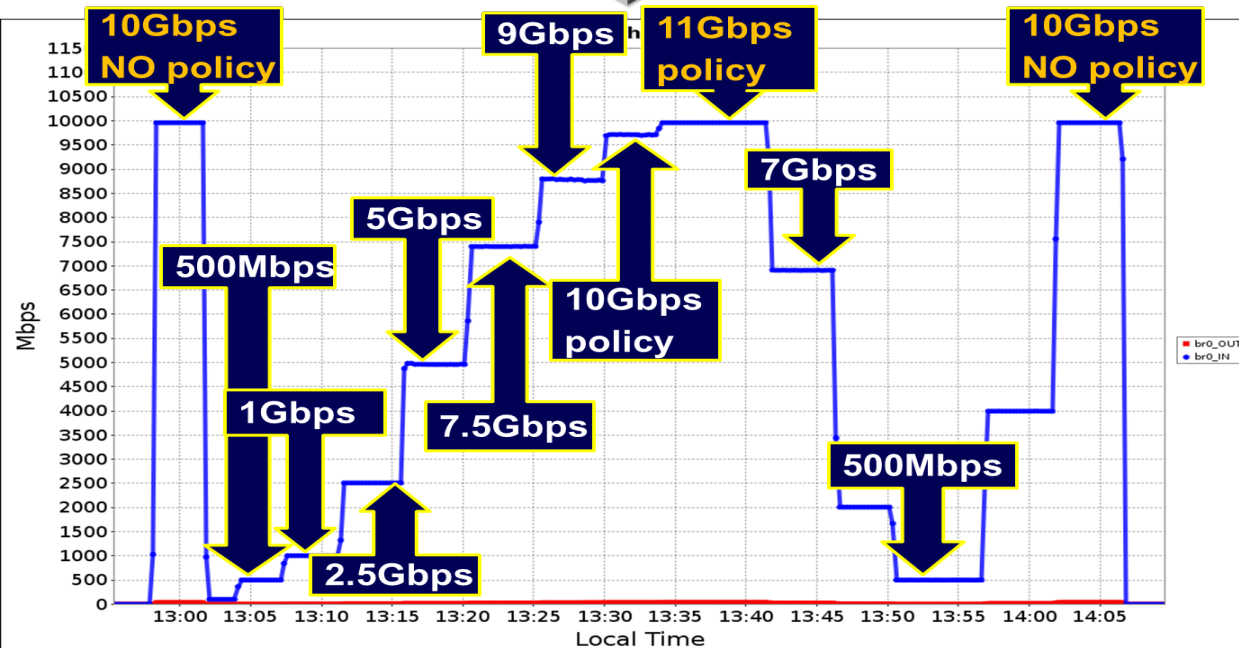


# Use Case: Traffic Shaping with Open vSwitch (OVS)

R. Voicu



- 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)



# Possible OVS benefits

- **The controller gets the “handle” all the way to the end-host**
- **Traffic shaping (egress) of outgoing flows may help performance in cases where upstream switch has smaller buffers**
- **A SDN controller may enforce QoS in non-OpenFlow clusters**
- **OVS 2.3.1 with stock SL/CentOS/RH 6.x kernel**
- **OVS bridged interface achieved the same performance as the hardware (10Gbps)**
- **No CPU overhead for OVS in this scenario**



# Summary

- **LHC Run2 Start: A new Era of Technical Challenges**
  - HEP's **data volumes during Run2 will continue to grow**, but waiting times to complete data transactions cannot grow (much)
  - HEP's **reliance on network performance will continue to grow**
  - Network usage by other research fields will continue to increase: there will be **increased competition for network resources**
- **A new era of physics exploration during Run2, paralleled by the transition to a new era of network technologies**
  - 10G to 100G links: Esnet EEX, Internet2, US CMS Tier2s
  - LHCONE: A Virtual Routing and Forwarding Fabric; and an emerging complementary paradigm of dynamic P2P circuits
  - Need: Emergence of a new generation of Intelligent Software Defined Networks
- **Which may enable us to meet the challenges, with focused work**