OLIMPS: OPENFLOW LINK-LAYER MULTIPATH SWITCHING

Harvey Newman, Artur Barczyk, Michael Bredel

DOE ASCR NGN PIs Meeting

Rockville, Sept. 16/17, 2014

Openflow Link-layer Multipath Switching (OLiMPS)

- **Problem statement:** Improve efficient, manageable use of large networks;

- **Methodology:** Optimization of data flow mapping in complex, multipath topologies

- **Enables per-flow load distribution,** which
  - Increases the robustness
  - Increases efficiency
  - Simplifies management of layer 2 network resources

- **Primary Use Case: LHCONE** Multipath solution

Openflow Link-layer Multipath Switching (OLiMPS)

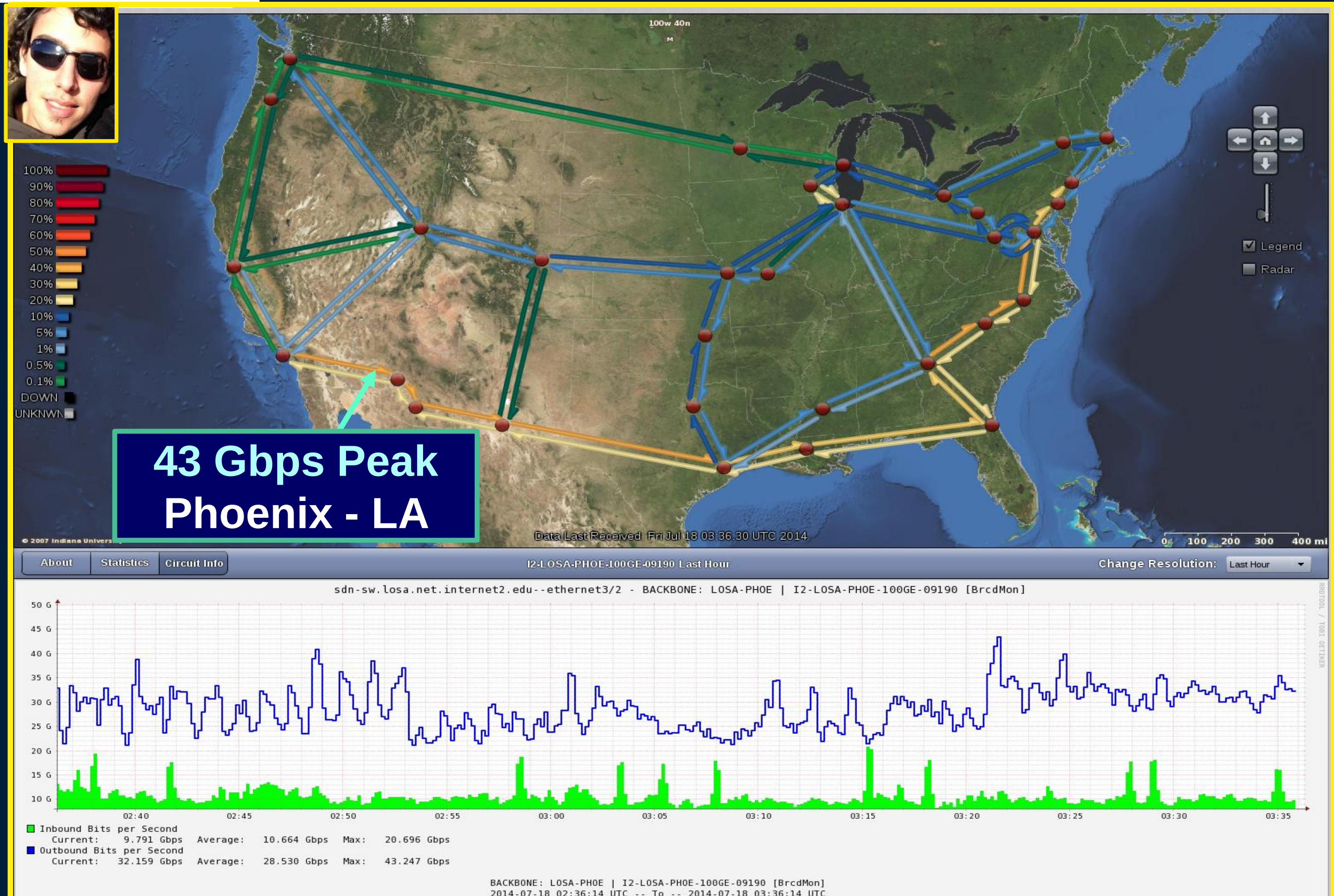
- Our approach used two components:
  - Intelligent path selection in OpenFlow networks
  - Integration with dynamic circuit provisioning systems (OSCARS)

- Synergies with use of efficient data transfer application: FDT, and pervasive robust monitoring: MonALISA

- Side Benefit: Begin work on SDN using OpenFlow; leverage work by our team with a Floodlight controller targeted at multipathing

- Progress:
  - Built and demonstrated a capable Floodlight-based controller
  - Porting code to Open Daylight framework
  - Interfacing to Layer1 dynamic optical paths by SC14 (underway)

Transfer Caltech ➡ Europe elevates usage of Internet2 to > 40% occupancy on some segments



Prelude to Production Use by US Tier2s on ESnet

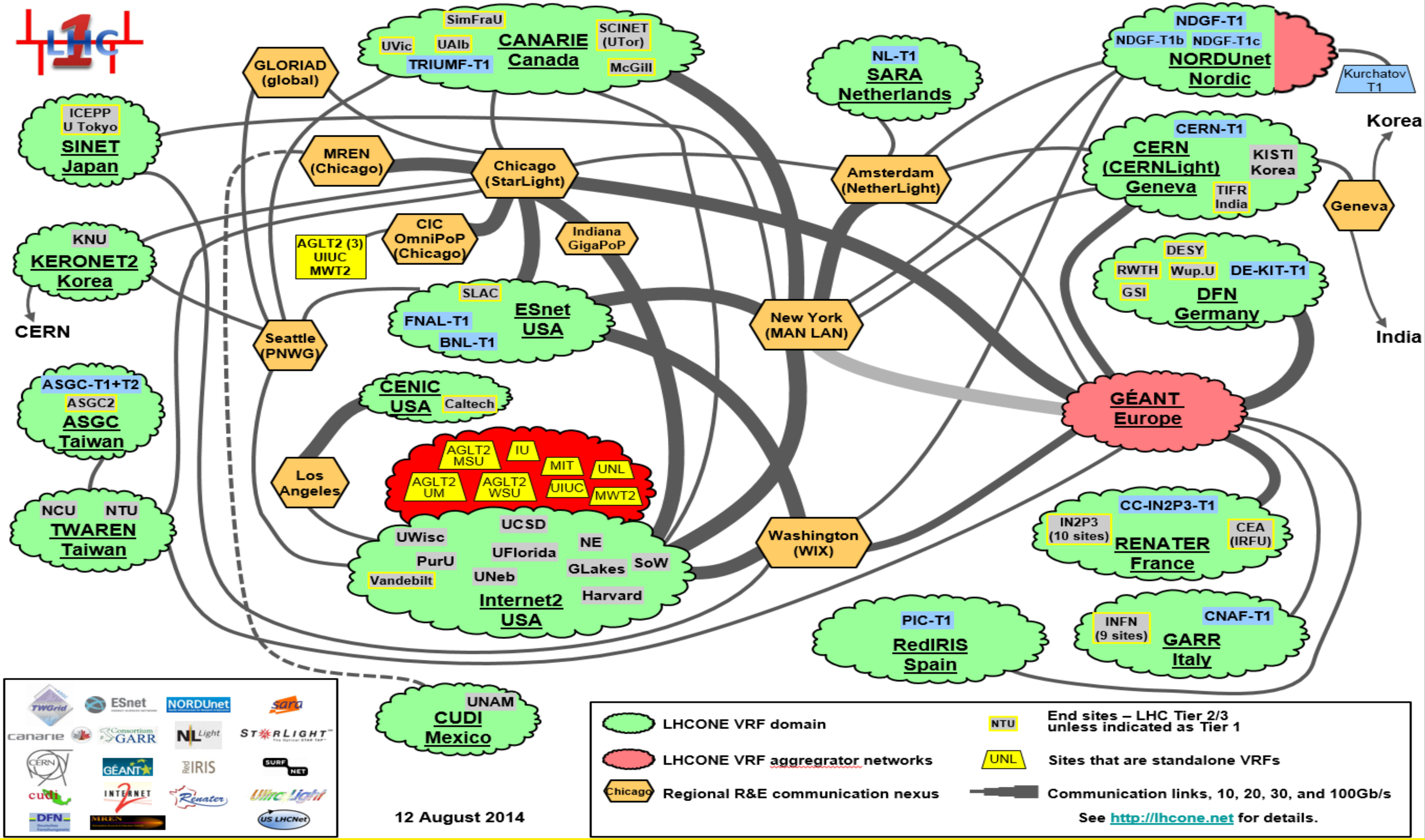
For example: Caltech OliMPS Controller and ESnet OSCARS Circuits this Fall

Can move to OpenDaylight Controller and NSI in 1H 2015

Now: Developing path selection and load balancing methods using the Internet2 Transfer Caltech ➡ Europe elevates usage to > 40% occupancy on some segments

W. Johnston, ESnet: Map of the Global LHCONE Virtual Routing and Forwarding (VRF) Infrastructure Supporting Tier1/2/3 Connectivity

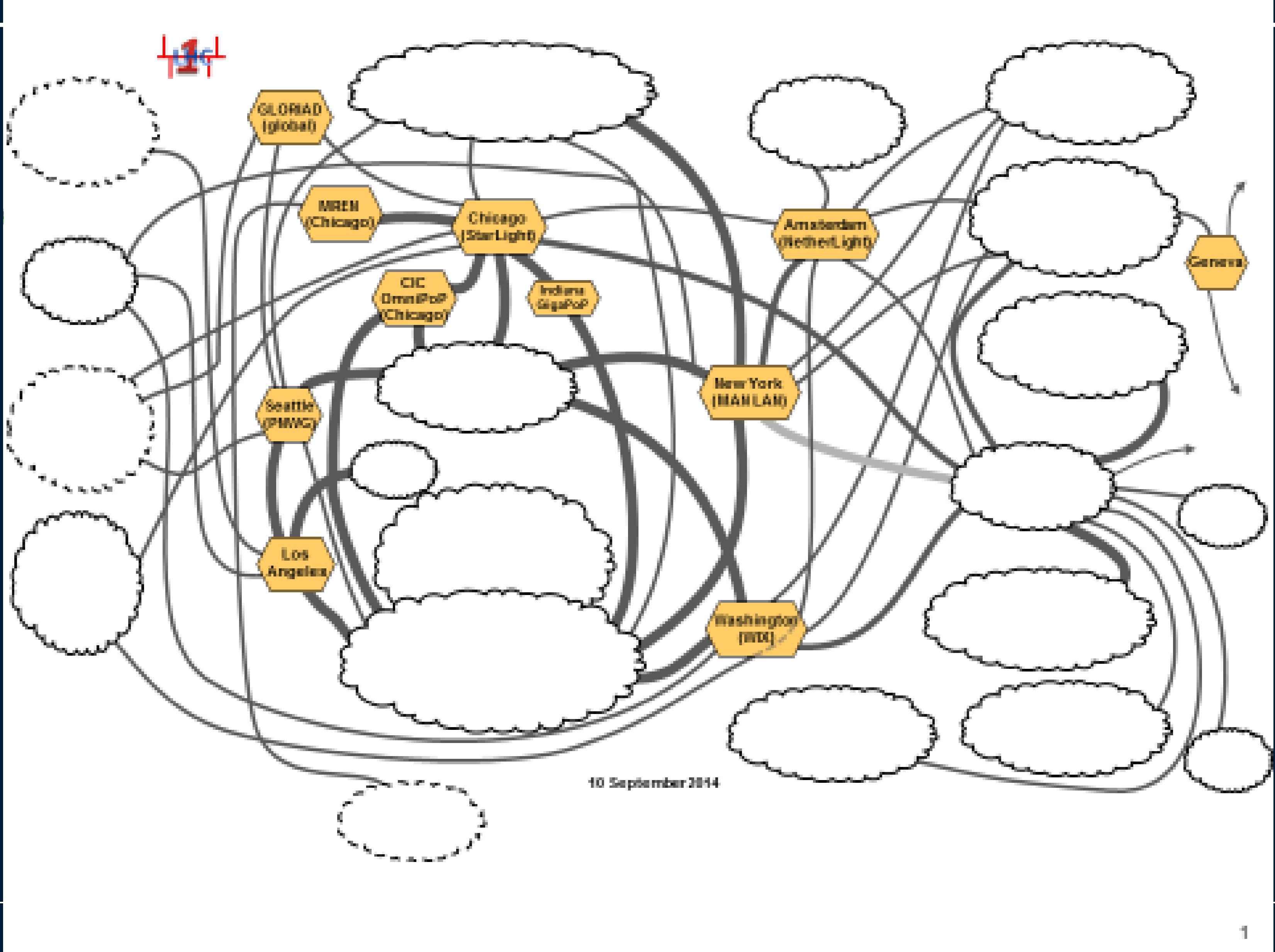
LHCONE: A global infrastructure for the LHC Tier1 data center and Tier 2/3 analysis center connectivity



33

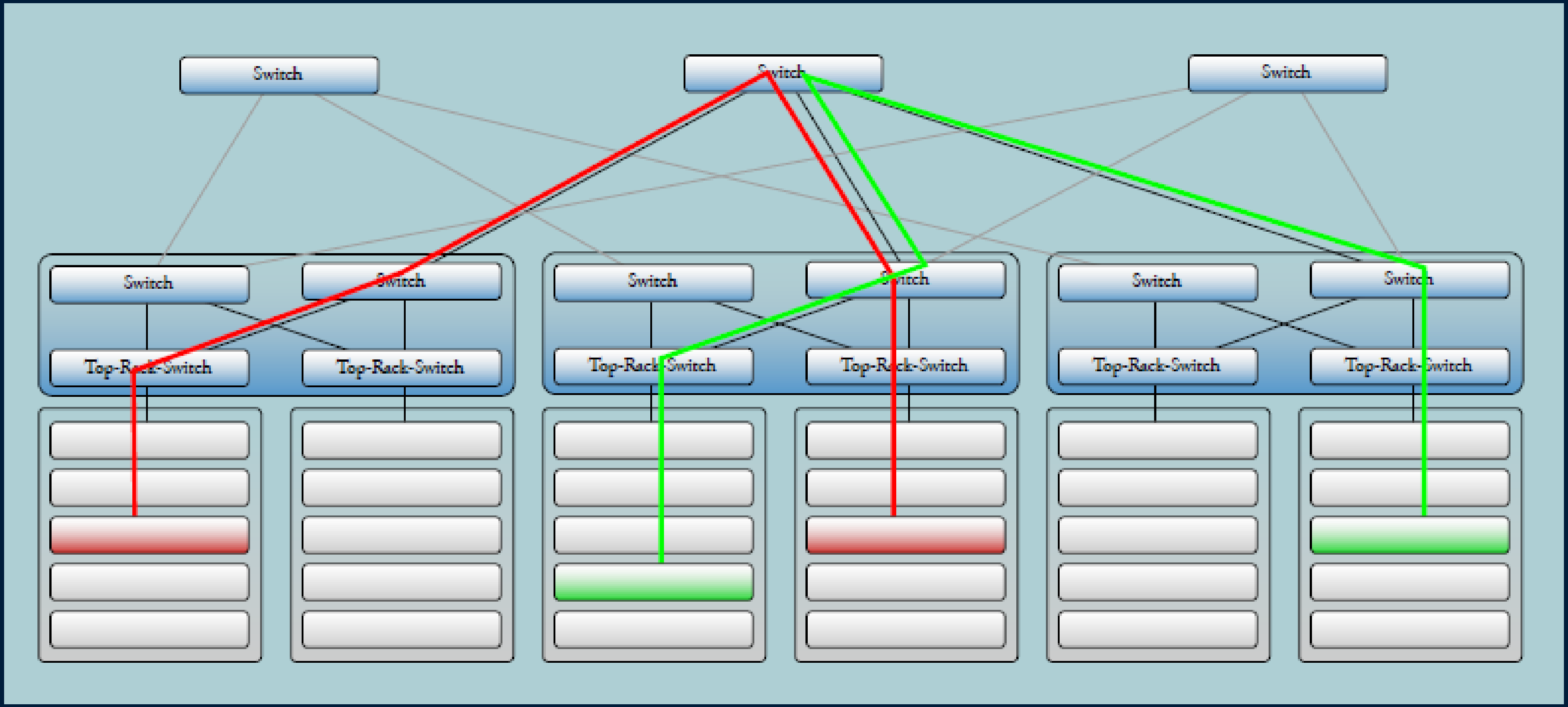
The Major Network R&E Players have Mobilized in Support of the LHC Program

LHONE - Plenty of Paths



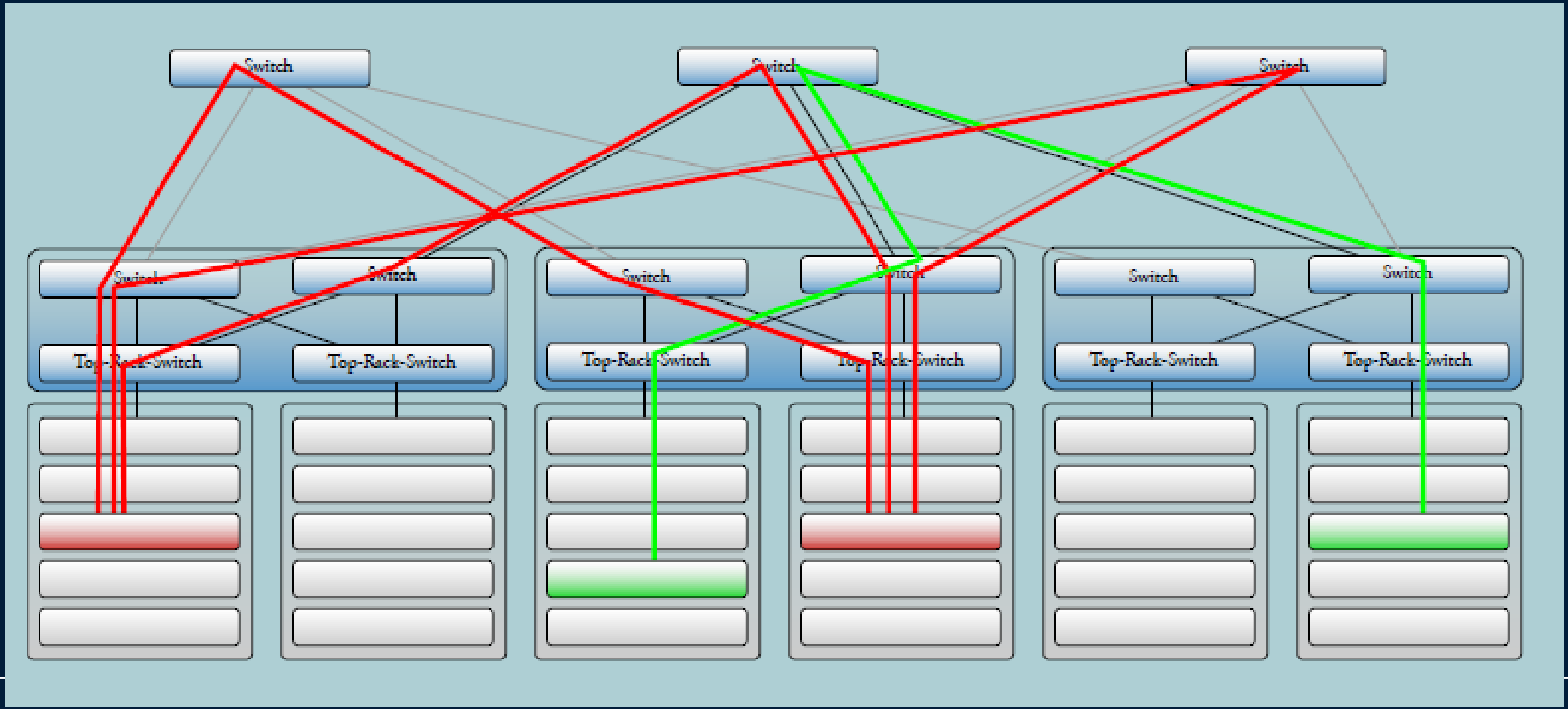
Classical Data Center Topology Limitations

- **Current techniques are limiting performance:**
  - Spanning Tree for loop avoidance
  - LAGs are link-local
  - scaling up involves much configuration work on each involved device



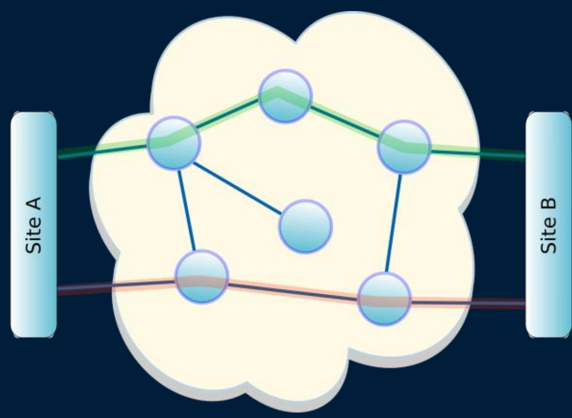
Multipath in Data Center

- Multipath can be achieved in several ways, e.g.
  - Multipath-TCP (IETF RFC 6824)
  - TRILL (IETF RFC 6325)
  - SPB (IEEE 802.1aq)
  - And/Or Load Balancing algorithms in SDN



WAN and LAN considerations

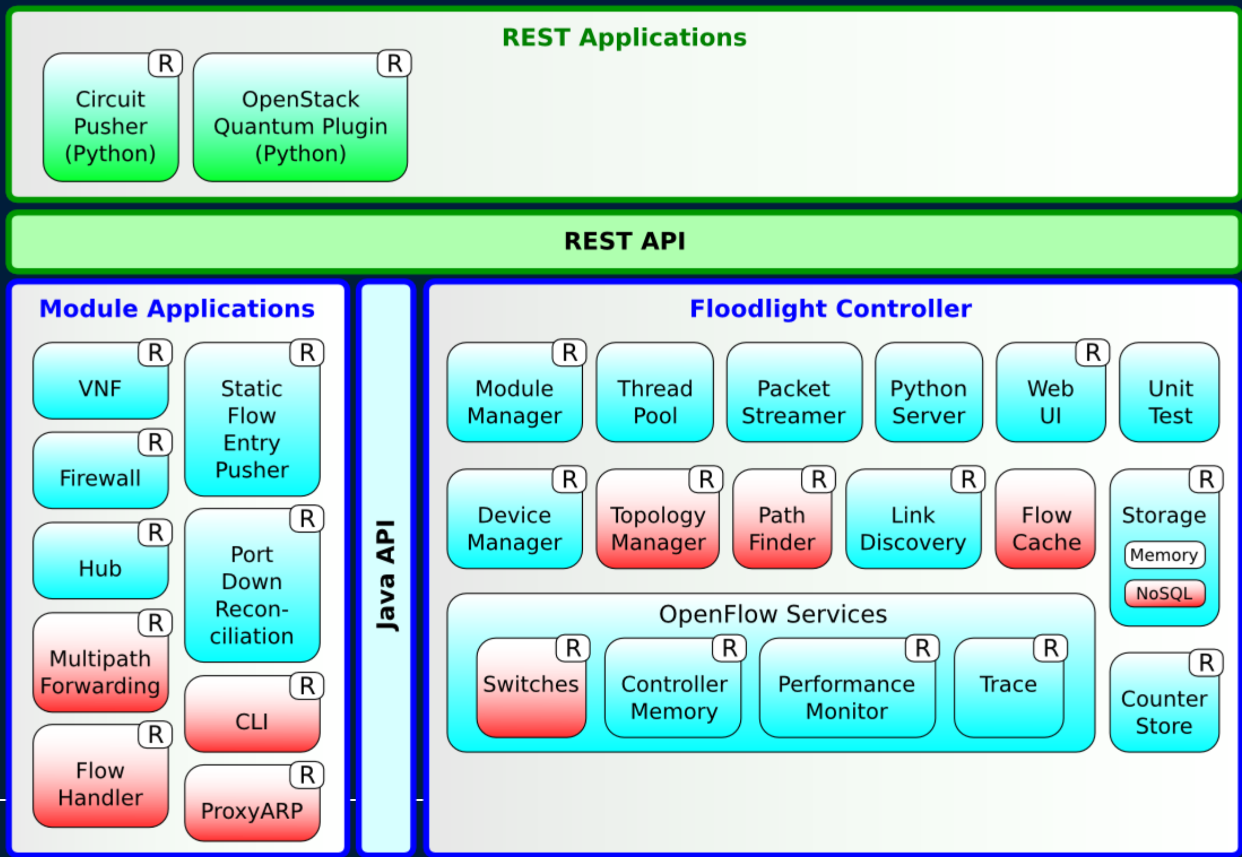
- Need to assume "classical" devices between OpenFlow switches
  - VLANS, STP, broadcast domains



- Latency impacts Controller-device communication
  - Use barrier messages to prevent temporary black holing of traffic

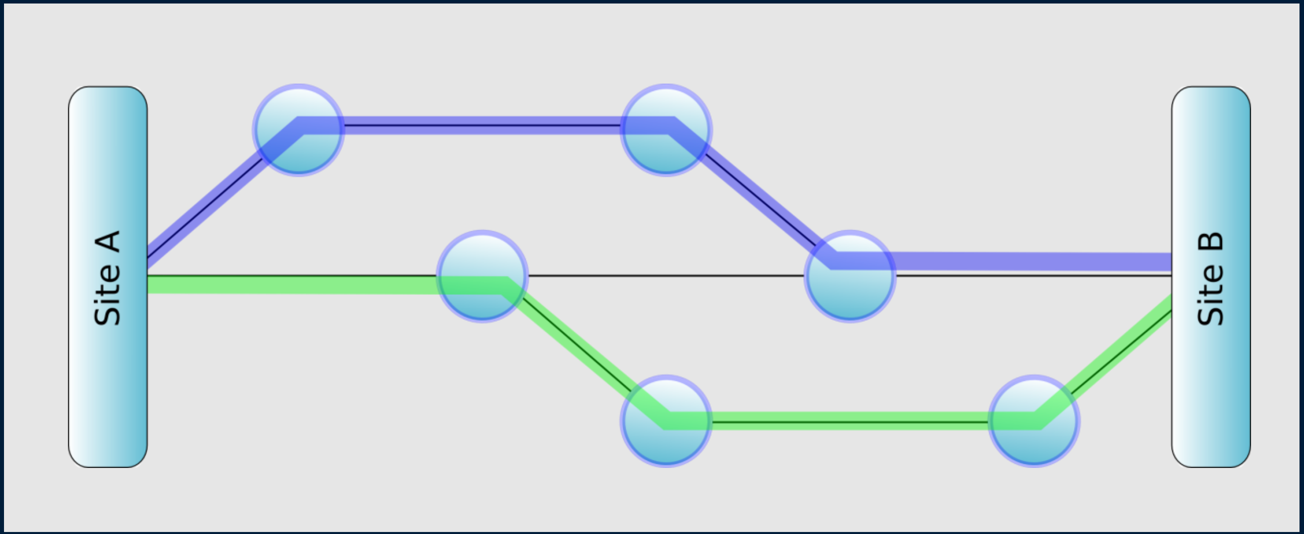
FloodLight extensions

- Curent OLiMPS controller based on FloodLight w/ extensions:
  - CLI tools
  - Topology Management; Loop Prevention
  - Multipath calculation and forwarding



Path computation

- **At stable topology, compute all possible data paths between a source and destination**
  - **Dijkstra's algorithm**
  - **Suurballe's algorithm**



- **In the following, the topology was assumed static, and available paths pre-computed**

Path Selection Algorithms:
Implementation and comparison

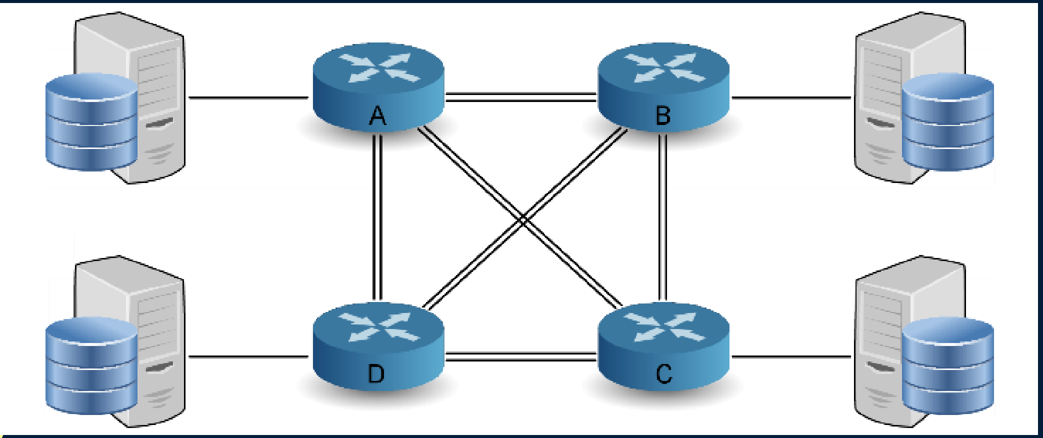
- **Algorithms without considering network state:**
  - **Hash-Based: assign flow to path $\boldsymbol{p_i}$ with** $i = H(flow\ n) mod\ P$
  - **Random: assign flow to a random path**
  - **Round-robin: assign flow to next available path**
- **Algorithms using network information:**
  - **Least-Flows: assign incoming new flow to path $\boldsymbol{p_i}$ with the least number of flows**
  - **Application aware mapping:**
    - Application provides additional data to the controller
      - We used only "amount of data to transfer", $\boldsymbol{S(j)}$
    - Assign new flow to path $\boldsymbol{i}$, minimizing the virtual finishing time $\boldsymbol{T(i)}$

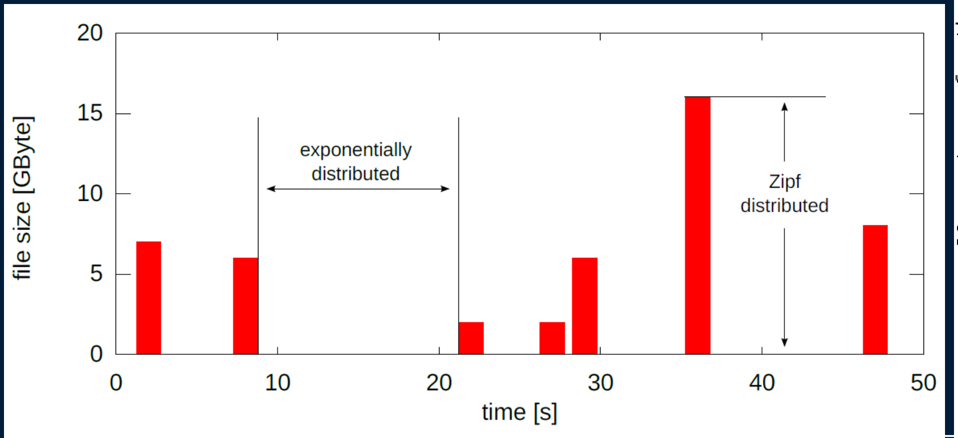
$$T(link_i) = \sum_{j=1}^{J} \frac{\max\{0, S(j) - D(j)\}}{w * C(link_i)}$$

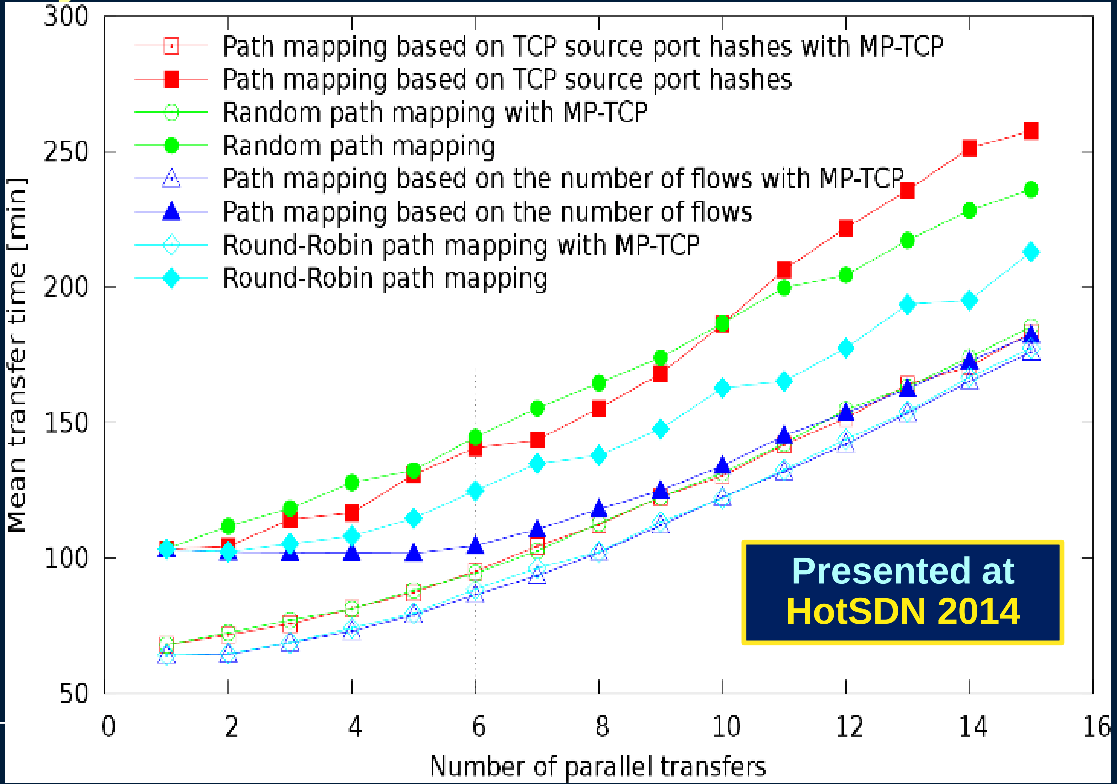
    - $\boldsymbol{D(j)}$ is the amount of data transferred ; $\boldsymbol{C(j)}$ is the capacity of link $\boldsymbol{j}$
- **Also, evaluated performance of each algorithm with and without MP-TCP**

Comparison Results

- For comparative measurements, a test bed was prepared:
  - 10G host interfaces
  - 6 x 1Gbps paths through the network
- N transfers executed simultaneously
- Data set sizes Zipf-distributed
  - 500GB average
- Inter-transfer time exponential

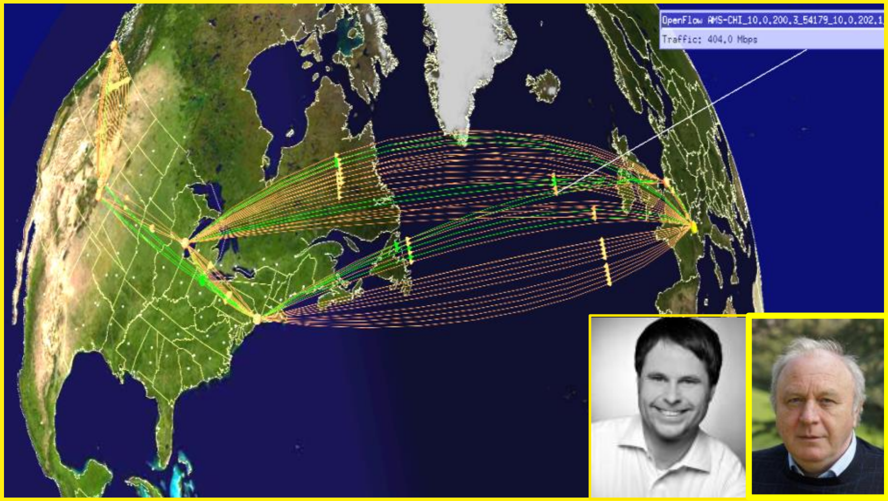






Presented at HotSDN 2014

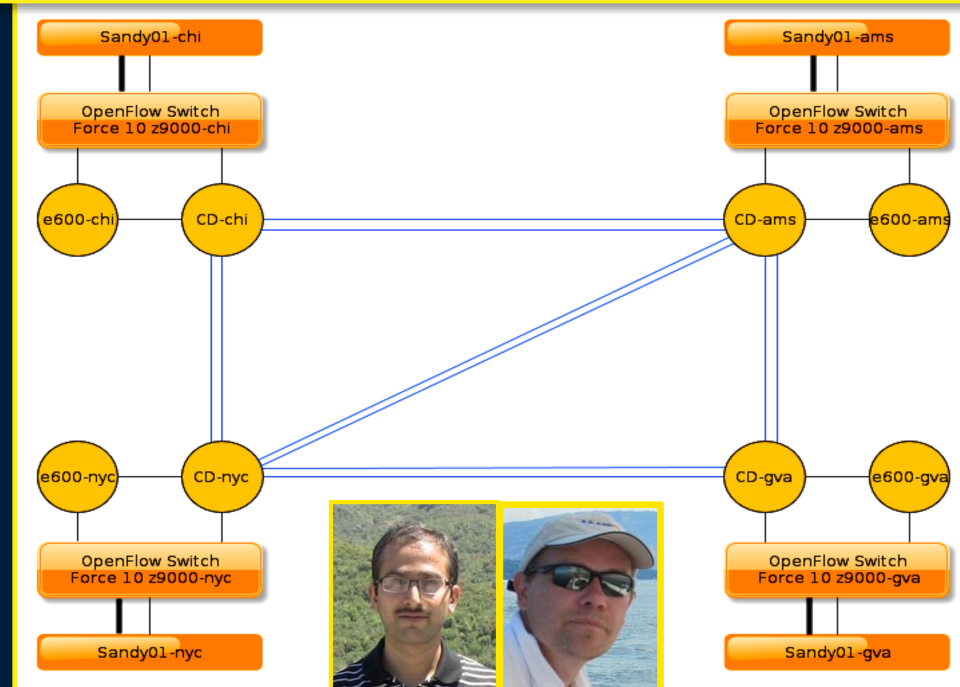
Caltech + Partners: OpenFlow Testbed Demo with MonALISA at SC13



- 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

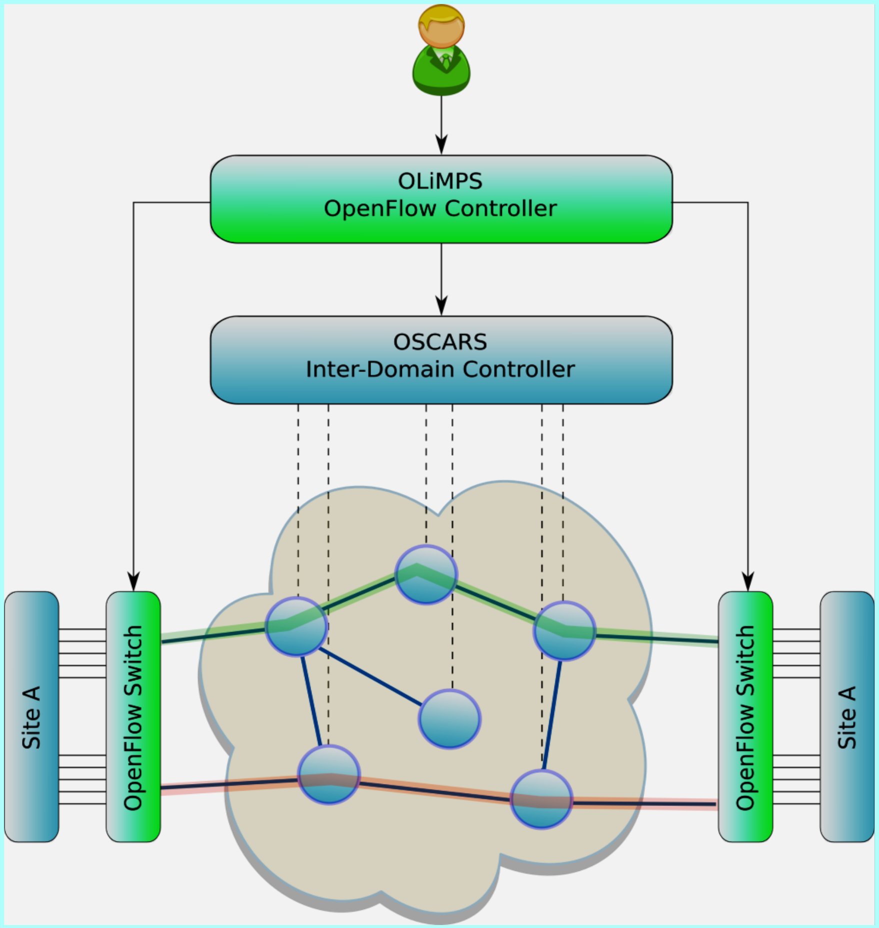


- Leading to powerful intelligent interfaces between the LHC experiments' data management systems and the network
- Generally useful: will be integral to the OpenDaylight Controller

OLiMPS and OSCARS
and Planned Developments using OpenDaylight

OLiMPS/OSCARS Interface

- User traffic analyzed by the OLiMPS controller
  - In active mode: Application provides additional parameters to the controller
- OLiMPS requests setup of multiple paths from OSCARS-IDC
- OSCARS connects OLiMPS-controlled OpenFlow switches through virtual circuits
- OLiMPS transparently maps the traffic onto OSCARS circuits

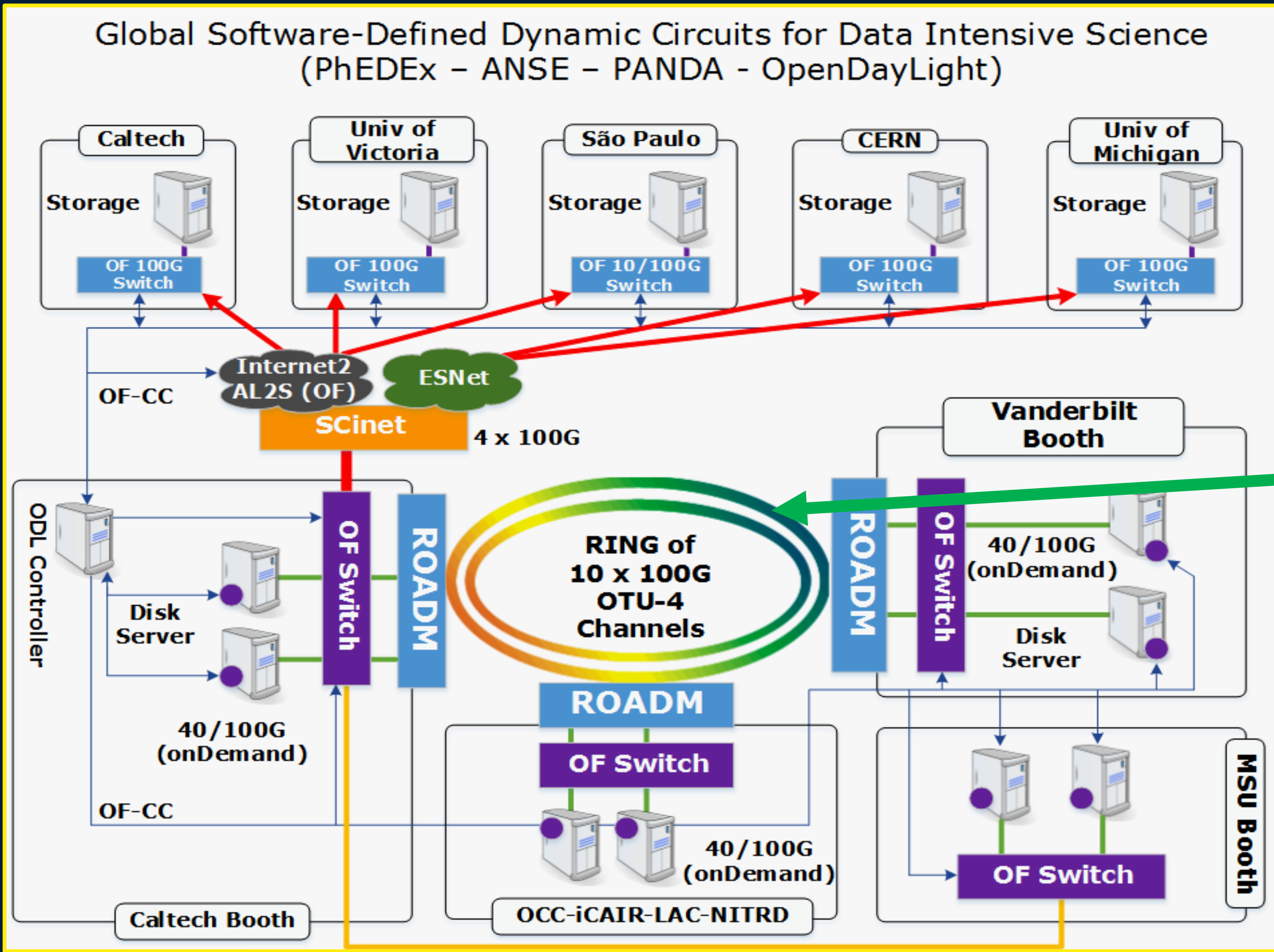


Porting the OLiMPS multipath functionality to OpenDaylight controller framework

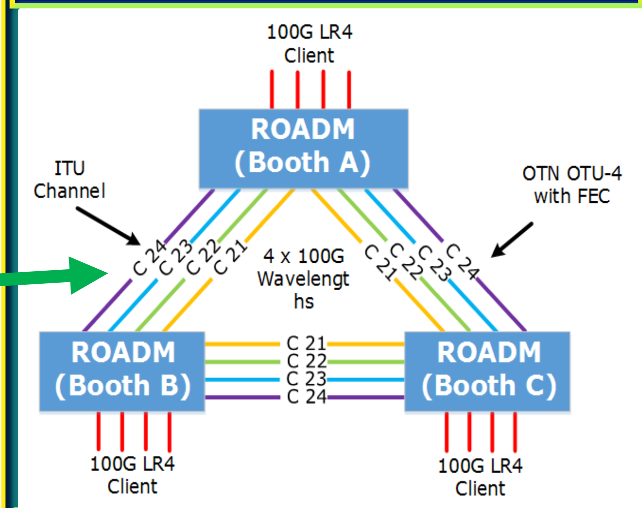
- In collaboration with and sponsored by Cisco Research

To be highlighted including Terabit-scale multi-layer dynamic circuits at SC'14

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

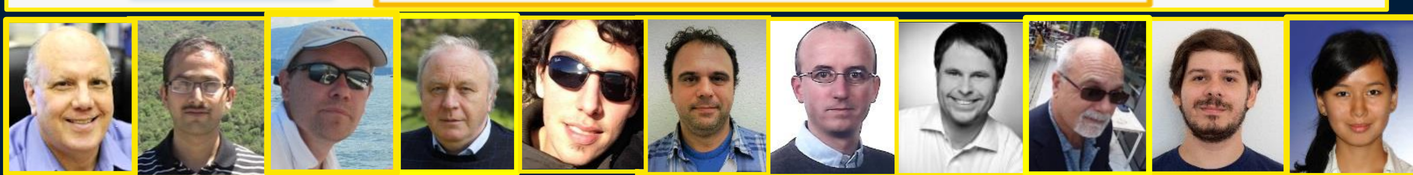


Terabit/sec Scale Long Range Networking



10 100G Waves

SDN Control of Optical and Switching Systems



Caltech HEP and Partners

SC'14 Tbps demo components

- **Collaboration Partners:**
  Caltech, Victoria, Michigan, Vanderbilt, UNESP (Sao Paulo), CERN
- **Software components:**
  - Terabit/sec scale software defined networking
  - Intelligent multilayer dynamic circuits over multiple network paths
- **Connectivity:**
  - Multiwavelength 10 X 100G ROADM network on the conference floor among three booths via dark fibers Caltech, Vanderbilt, and iCAIR/NITRD booth
  - ~5 X 100G external connections across the US and to Europe
- **State of the art system components:**
  - Many 40GE NICS, first 100GE NIC (TBC), SSD, NVMe, CPUs
- **Many network partners:**
  SCInet, ESnet, Internet2, CERN, CENIC, MiLR, CANARIE
- **Strong vendor support:** notably Brocade, Intel, Mellanox, Echostreams, Padtec (Brazil)

Project Website

- **More information at** **http://www.uslhcnet.org/projects/olimps**

- **Source code available on GitHub:** **https://github.com/mbredel/floodlight-olimps**

- **Currently we are porting the OLiMPS multipath functionality to the OpenDaylight controller framework**
  - In collaboration with and sponsored by Cisco Research
  - Will be part of the ODL controller

"What question does your research motivate you to now ask?"

- **Network as a Dynamic System**
  - **Including feedback through monitoring information, e.g.**
    - dynamical behaviour of algorithms with feedback data from the network
    - responsiveness to variations in throughput and network events
  - **What equations govern the flow behaviour? How general can this be formulated? What's the impact of data access patterns?**

- **Deployment in production networks such as the LHCONE**

- **A system for large scale science applications needs**
  - **coherent architecture, pervasive precise monitoring, real-time responsiveness, scalability and heuristic optimization**

THAK YOU!

artur.barczyk@cern.ch