Networking for the HEP Community: LHCONE and More

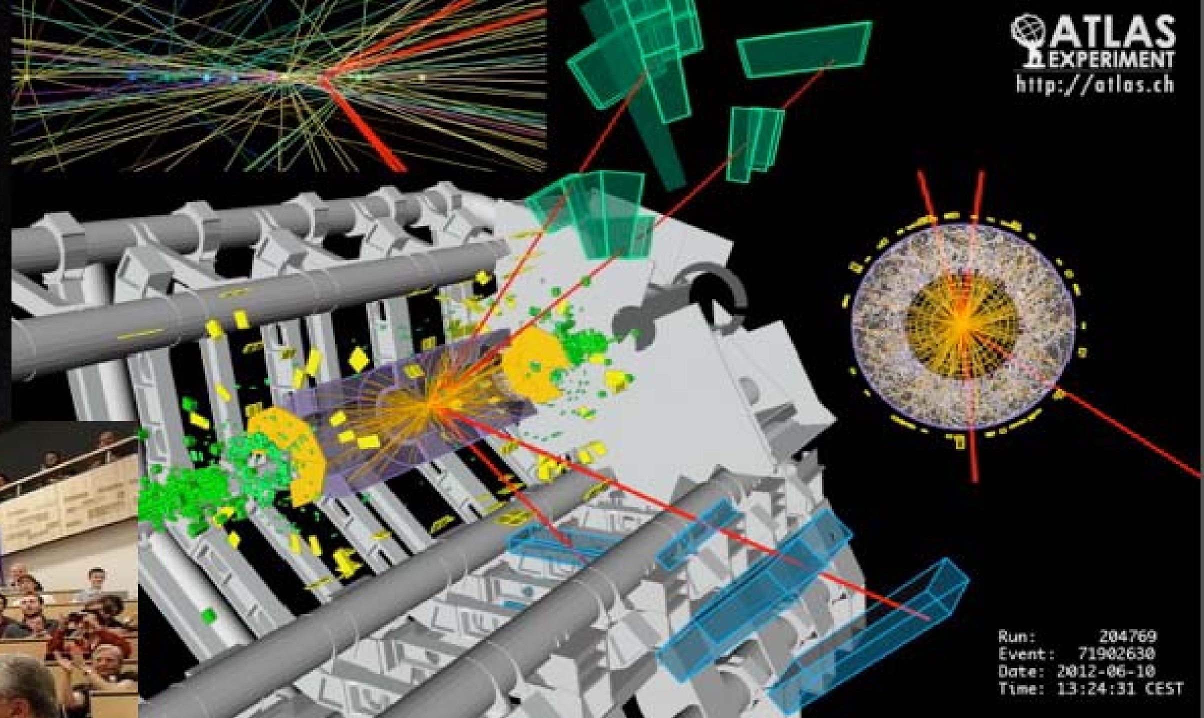
Harvey B Newman
Artur Barczyk

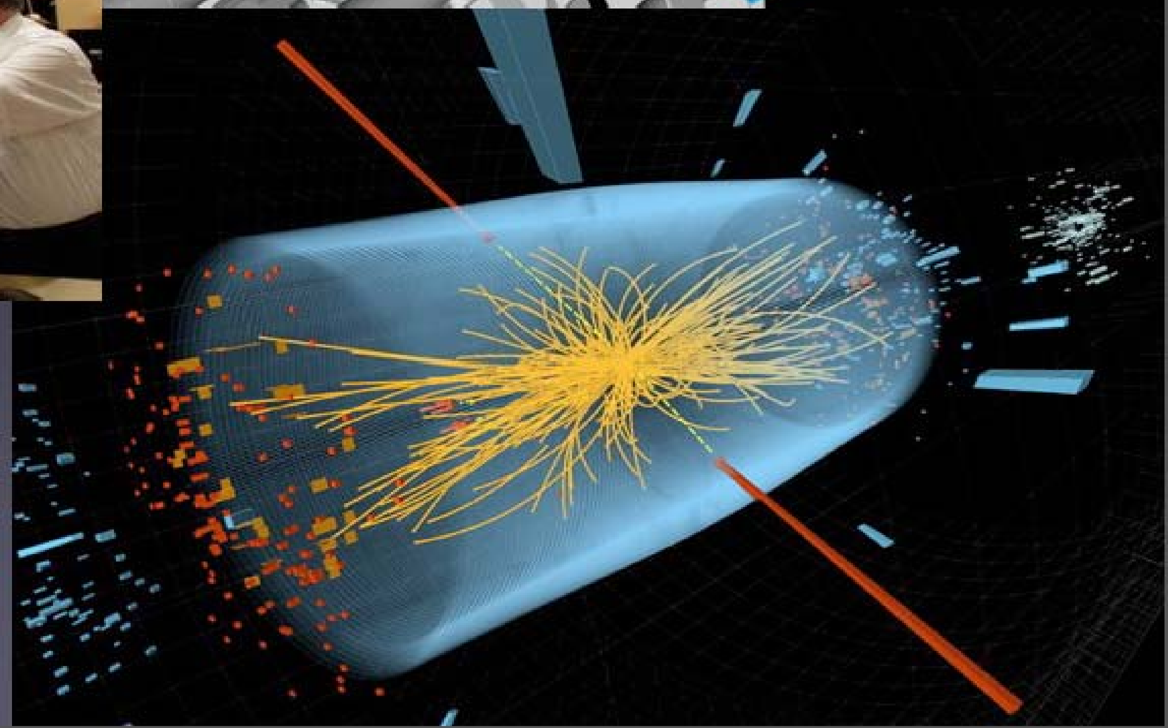
California Institute of Technology

12th Annual Global LambdaGrid Workshop
Chicago, October 11-12, 2012

2012.7.4

discovery of
Higgs-like boson





theory：1964

concept：1984

construction：2001

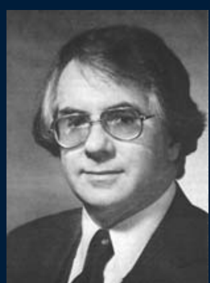
The Standard Model

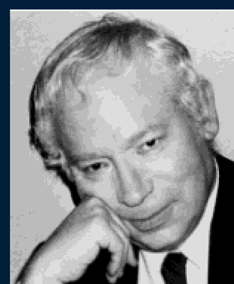
The Origins of Electroweak Symmetry Breaking

A great achievement of the second half of the 20th + 21st Century

Based on relativistic quantum field theories (QFT).

- The first was QED ➡
- The 2nd - Unified Electroweak

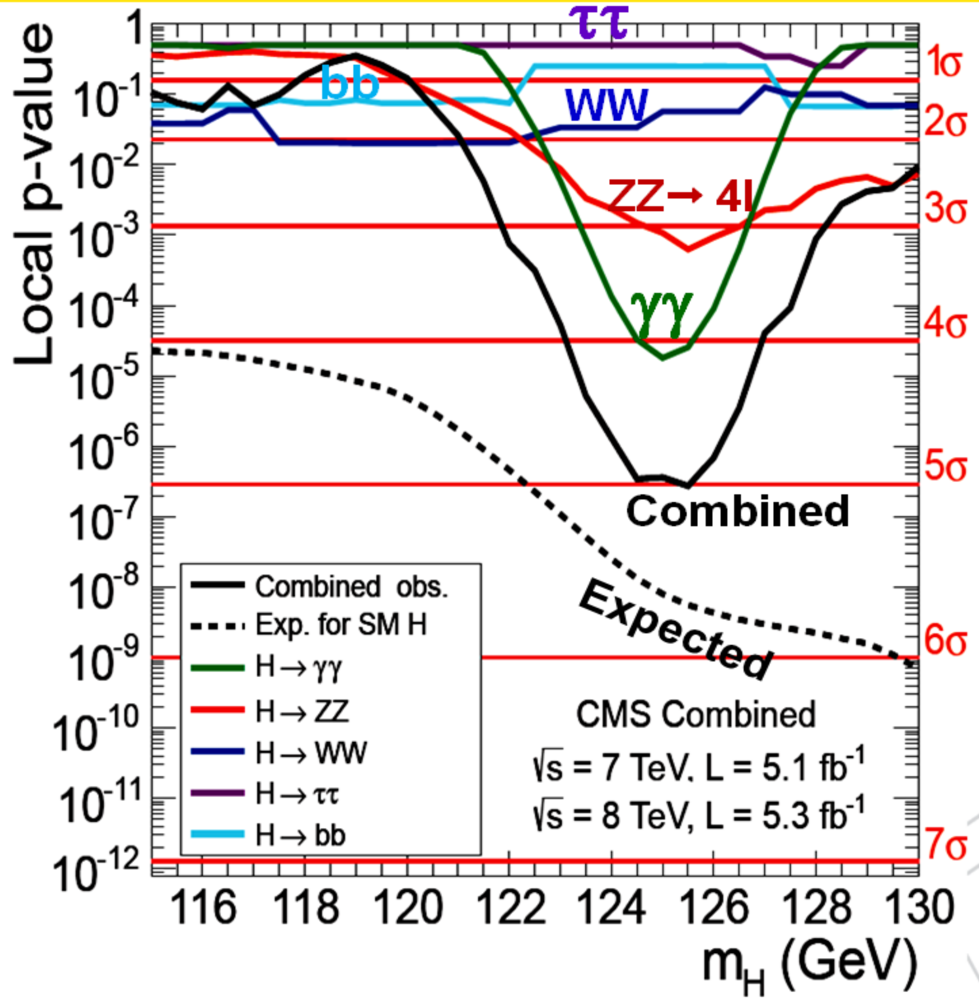




- 3rd: QCD for the Strong Interaction; *Asymptotic Freedom (Politzer et al.)*
- **'The Higgs' boson is the Candidate to explain Electroweak Symmetry Breaking**

Observation of a New Boson Near 125 GeV

p-values and Significance by Channel



arXiv:1207.7235 ; CMS-HIG-12-028
CERN-PH-EP-2012-220

Excess at ~125 GeV seen

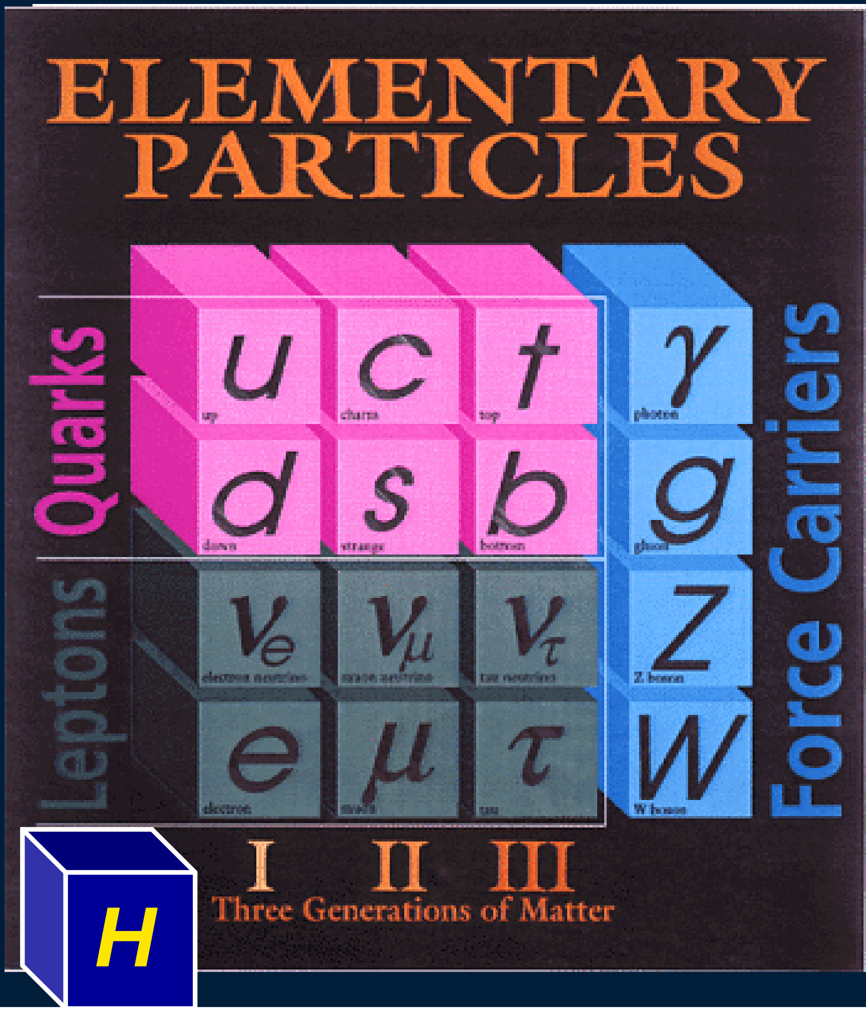
in both 7 TeV data: 3.0 σ
and 8 TeV data: 3.8 σ

High sensitivity, high mass resolution channels: γγ + 4l

- γγ 4.1 σ Excess
- ZZ → 4l: 3.2 σ Excess

	Expected σ	Observed σ
H→γγ	2.8	4.1
H→ZZ	3.6	3.1
H→ ττ + bb	2.4	0.4
H→ γγ + ZZ	4.7	5.0
H → γγ+ZZ+WW	5.2	5.1
H → γγ+ZZ+WW + ττ + bb	5.8	5.0

The Standard Model of Particle Physics: 3 Quark, 3 Lepton Families, 4 Forces



31 particle physicists have won Nobel prizes for making the experimental discoveries and theoretical breakthroughs
[Higgs Generates Masses]

The SM describes the known forces and particles, with one important exception:

Gravity

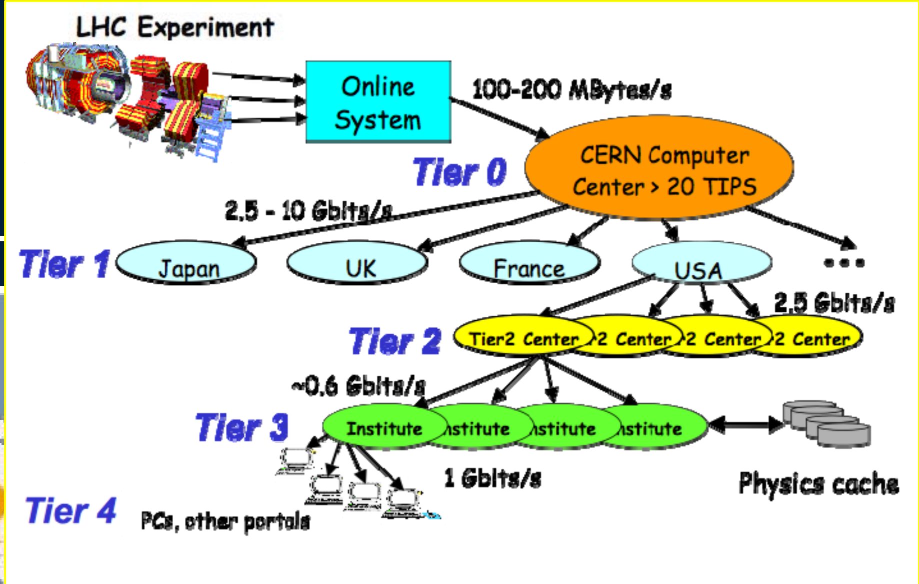
And does not explain:

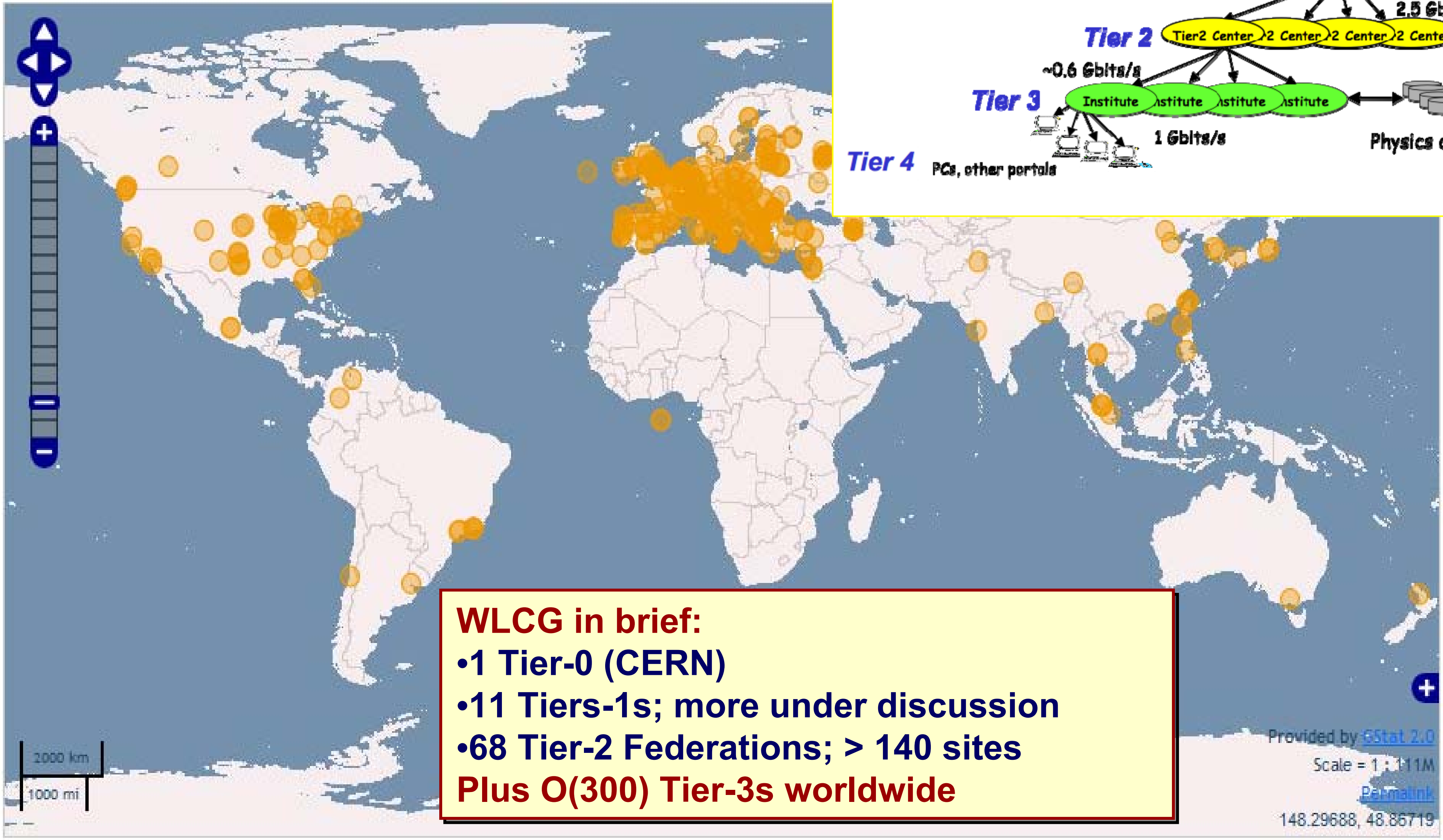
- *The existence of Dark Matter*
- *The unification of forces*
- *Dark energy*

The SM does not work in the early universe

A beautifully simple but *Incomplete* picture; a triumph of 20th *and 21st* century physics Leaving many questions unanswered

LHC Computing Infrastructure





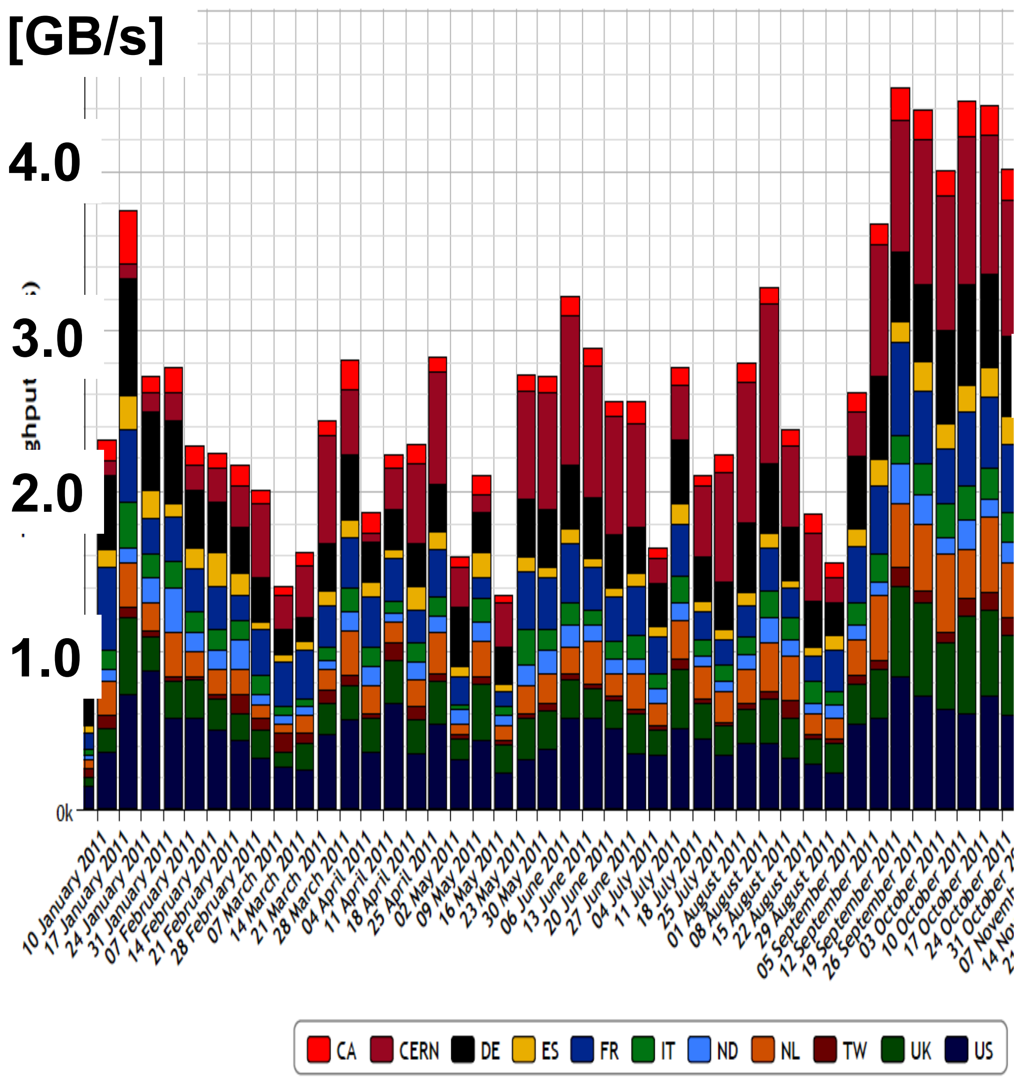
WLCG in brief:

- 1 Tier-0 (CERN)
- 11 Tiers-1s; more under discussion
- 68 Tier-2 Federations; > 140 sites

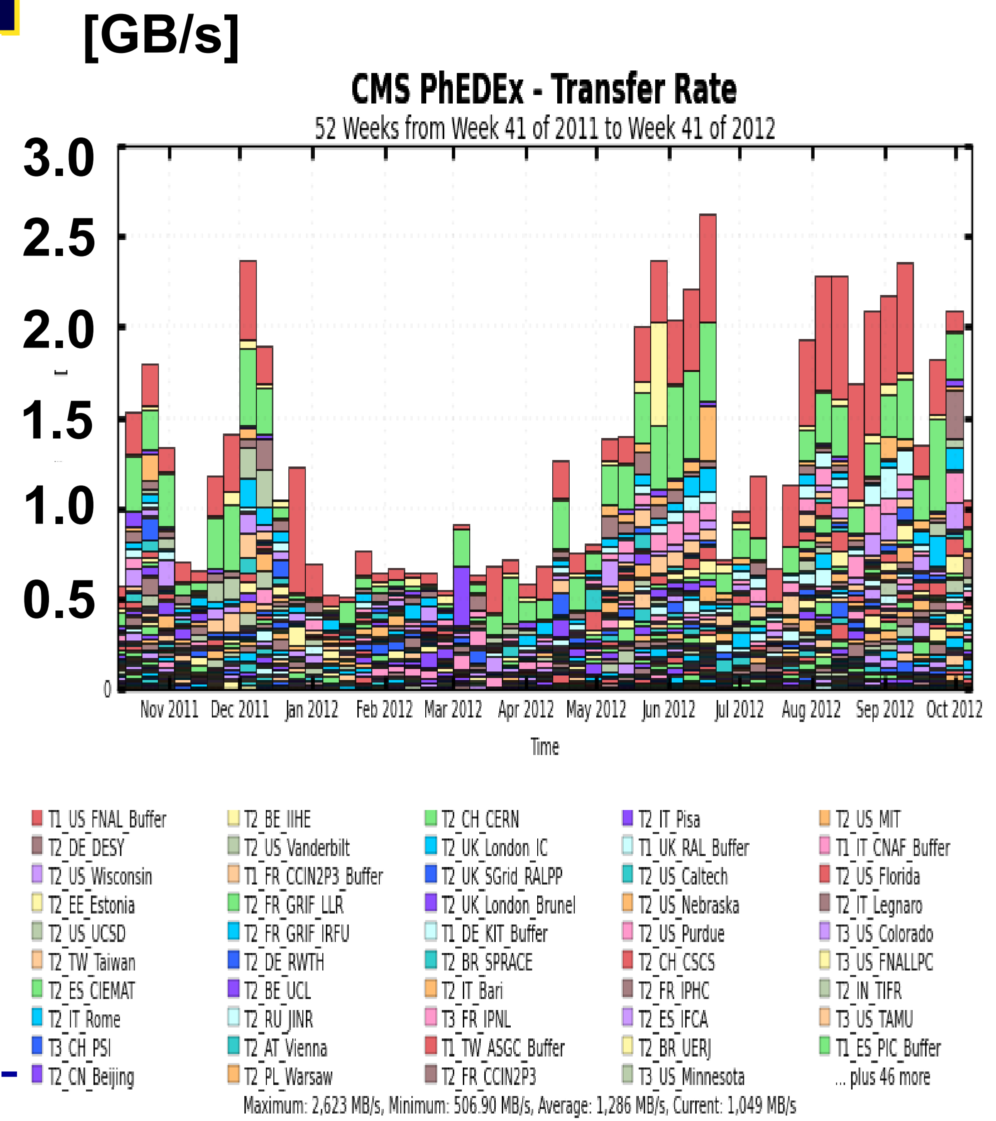
Plus O(300) Tier-3s worldwide

ATLAS Data Flow by Region: Jan. – Nov. 2011

~2.8 Gbytes/sec Average, 4.5 GBytes/sec Peak

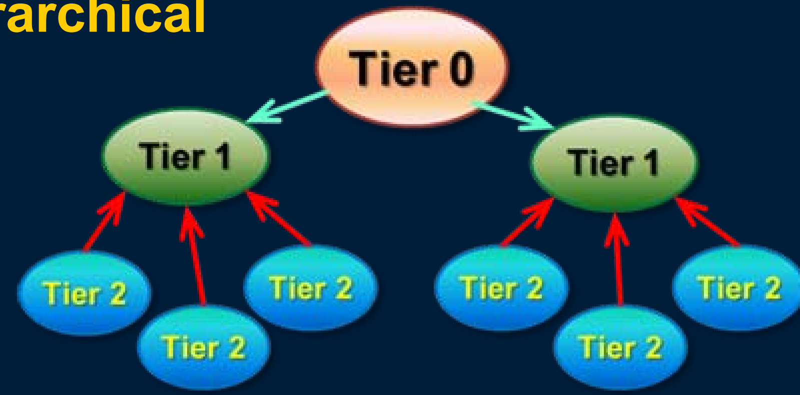


CMS Data Flow by Site: Oct 2011. – Oct. 2012

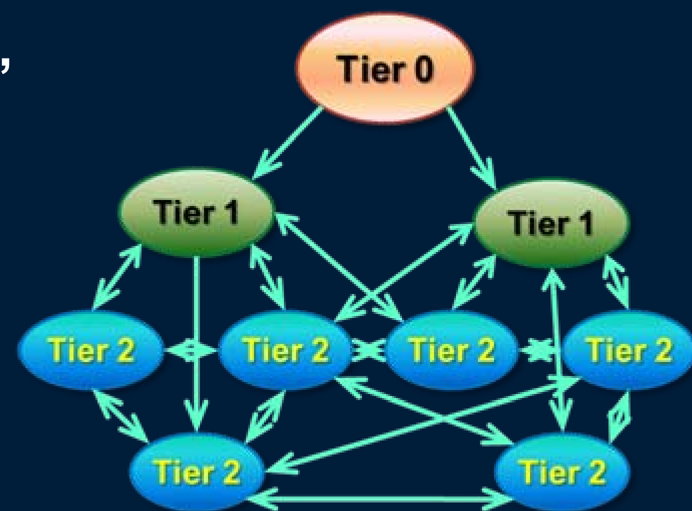


Computing Models Evolution

- **The original MONARC model was strictly hierarchical**
- **Changes introduced gradually since 2010**
- **Main evolutions:**
  - **Meshed data flows: Any site can use any other site as source of data**



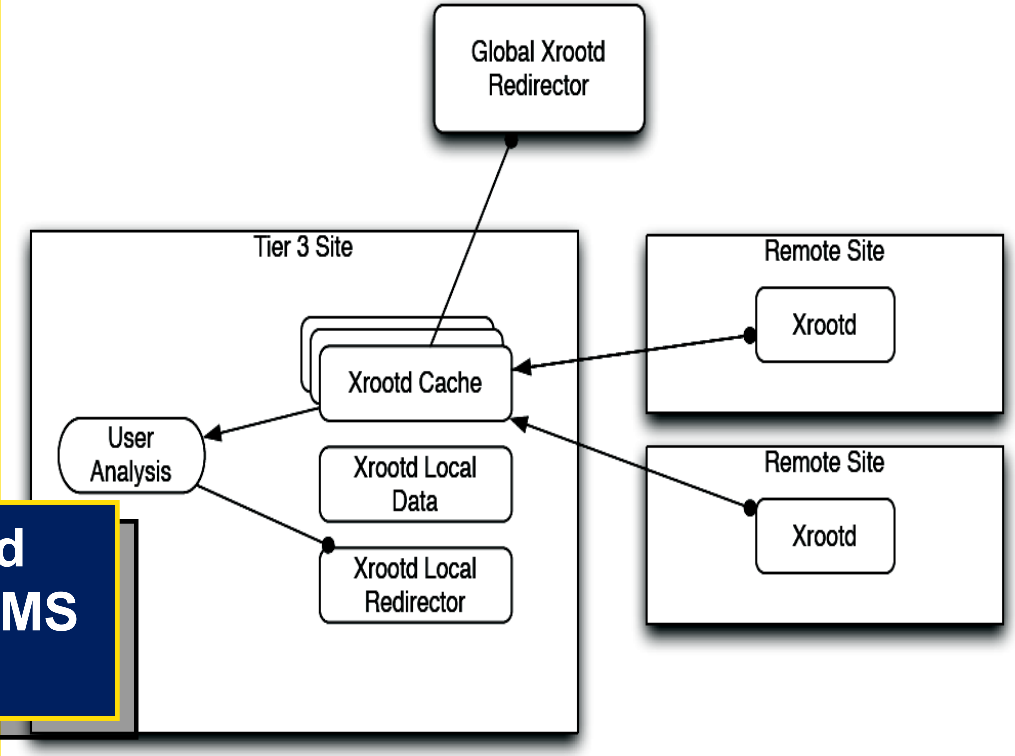
  - **Dynamic data caching: Analysis sites pull datasets from other sites "on demand", including from Tier2s in other regions**
    - **In combination with strategic pre-placement of data sets**
  - **Remote data access: jobs executing locally, using data cached at a remote site in quasi-real time**
    - **Possibly in combination with local caching**

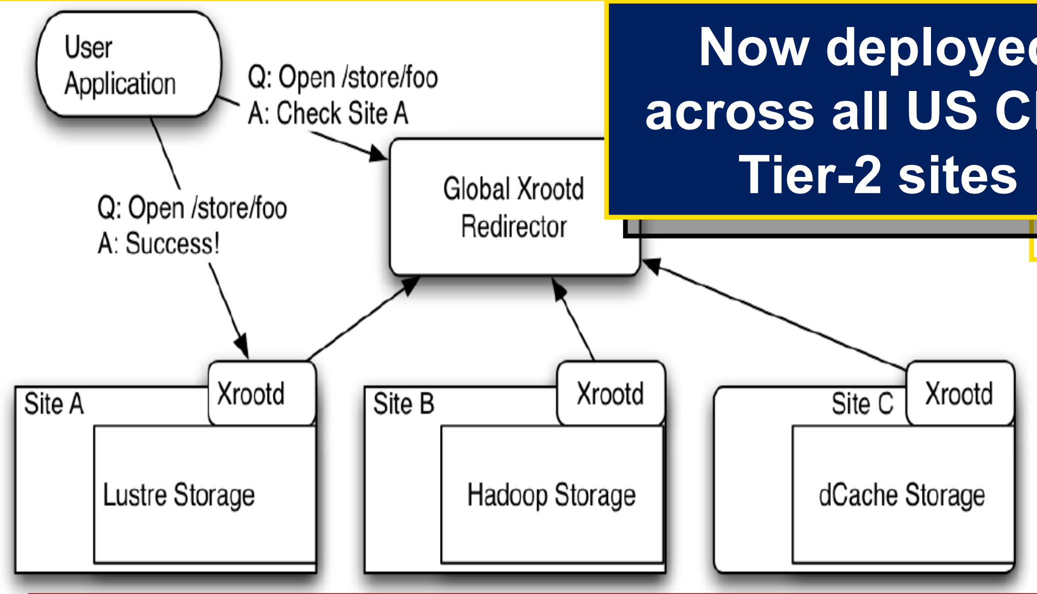


- **Variations by experiment**
- **Increased reliance on network performance !**

Remote Data Access and Processing with Xrootd (CMS)

- Data read through redirector, source hidden from user
- Only selected objects are read (with object read-ahead). No transfer of entire data sets
- Use cases include fallback for read errors, "diskless Tier-3"





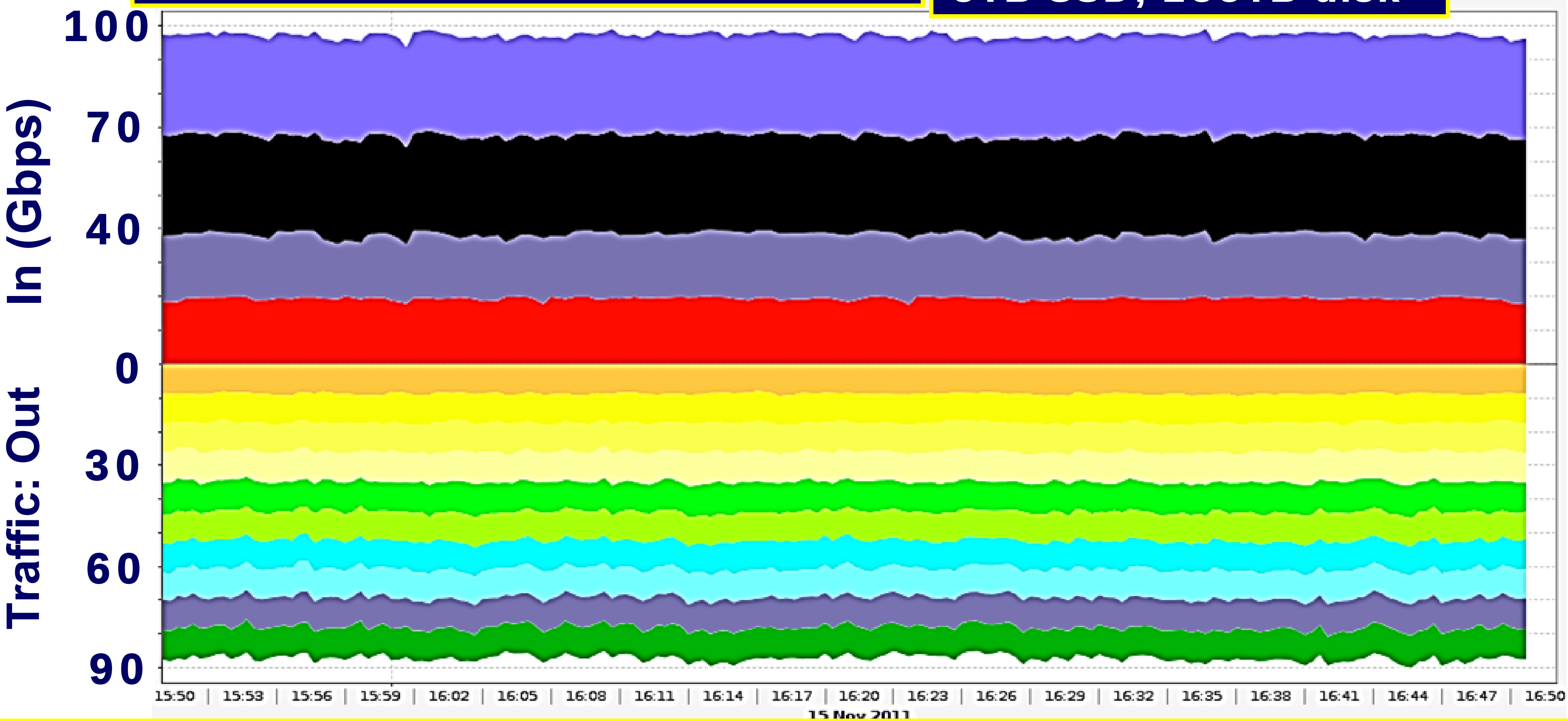
Now deployed across all US CMS Tier-2 sites

Similar operations in ALICE for years

Brian Bockelman, January 2012

Research Partners: UVic, Florida, BNL, FNAL, Michigan, Brazil, Korea, ESnet, NLR, FLR, Internet2, BNL, ESNet, CWave, AWave, IRNC, KREONet

~140CPU Cores, 8 Gen2/3 NICs in 1 Rack of Servers
1 100GE port, 32 40GE Switch Ports;
8TB SSD, 288TB disk

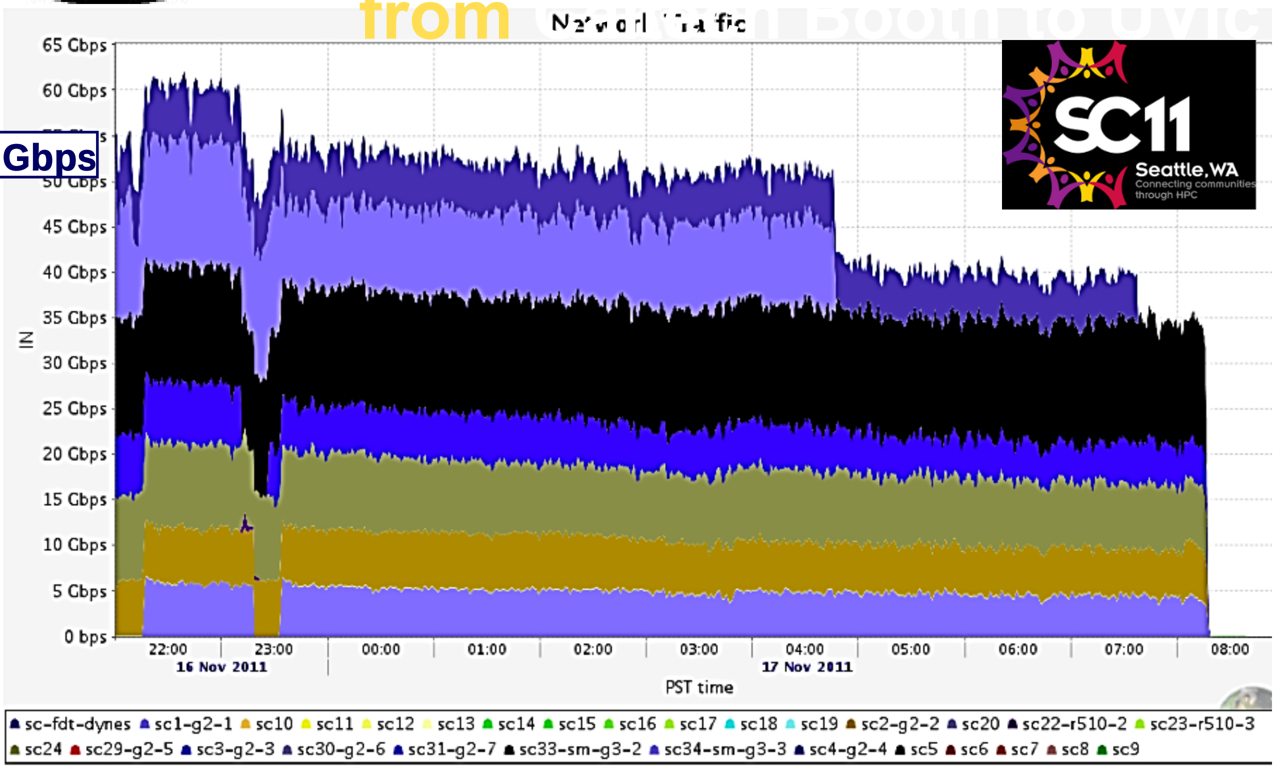


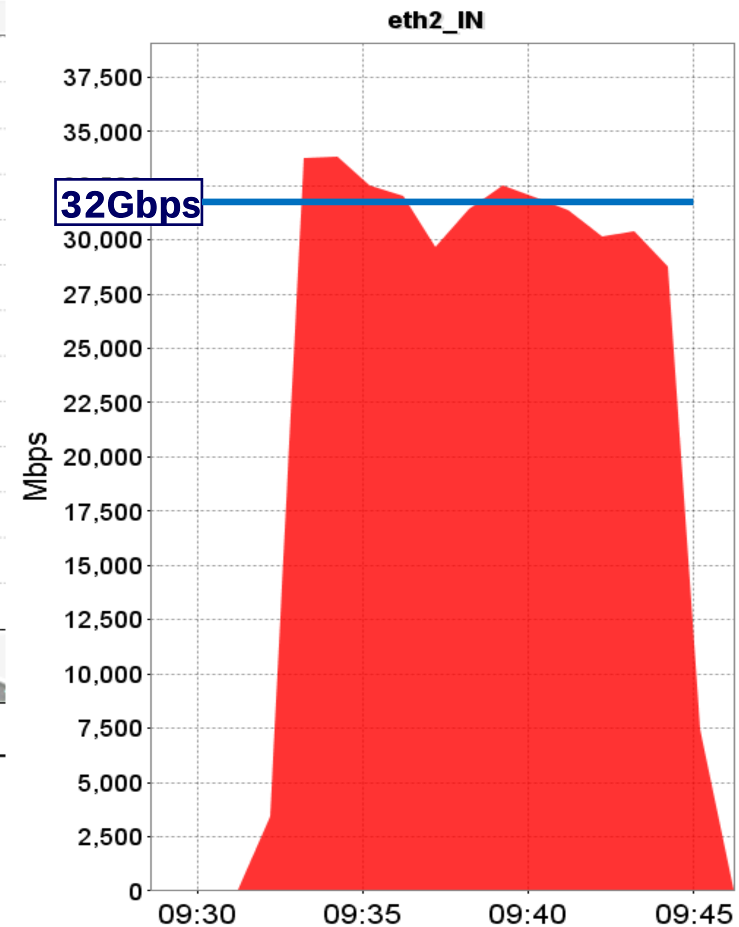
Sustained 186 Gbps; Enough to transfer 100,000 Blu-rays per day

SC12 (Salt Lake): 3 X 100G Demonstration

Disk to Disk Results: 100G Wave from

Latest 40G Server Results

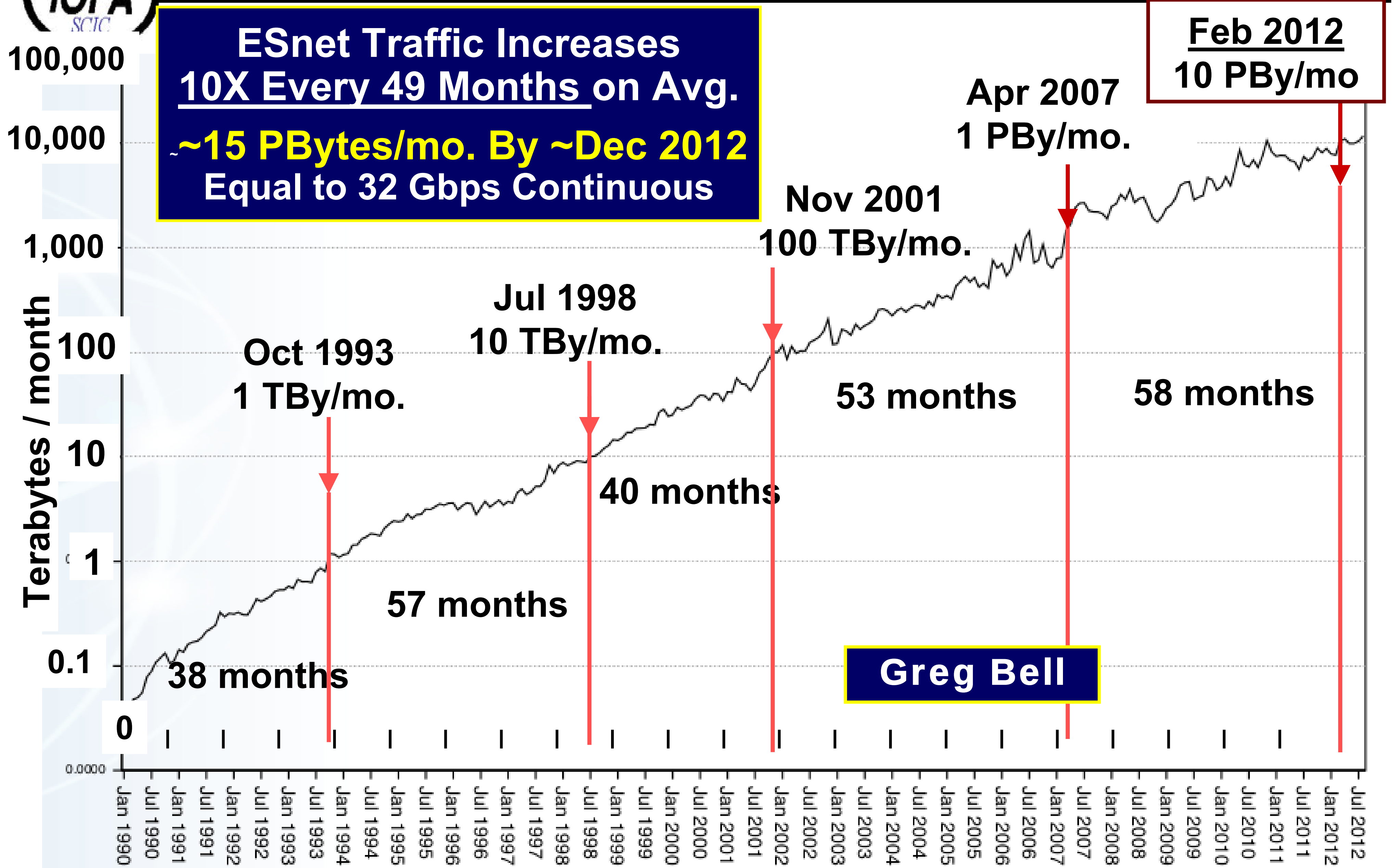




Peaks of 60Gbps disk write on 7 Supermicro and Dell servers with PCI Express Gen 3 buses and 40G Ethernet interfaces

Single Server Gen3 performance: to 36.8Gbps inbound

Remarkable Historical ESnet Traffic Trend Cont'd in 2012



Log Plot of ESnet Monthly Accepted Traffic, January 1990 – July 2012

R&E Network Trends in 2011-12

- **Increased multiplicity of 10G links in the Major R&E networks:** Internet2, Esnet, GEANT, and some European NRENs
- **100G next-generation networks: Backbone in place;** Transition now underway in Internet2 and Esnet !
- **GEANT transition to 100G not far behind;** underway by Fall
- **100G already appearing in Europe and Asia**: e.g. SURFnet – CERN; Romania (Bucharest – Iasi); Korea (Seoul – Daejon)
  - *CERN – Budapest 2 X 100G for LHC Remote Tier0 Center in 2012*
- **Proliferation of 100G network switches and high density 40G data center switches. 40G servers (Dell, Supermicro) with PCIe 3.0 bus**
  - *First int'l 186 Gbps throughput demo: SC11 – U. Victoria*
- **OpenFlow** (Software-defined switching and routing) **taken up by much of the network industry, R&E nets and GLIF**

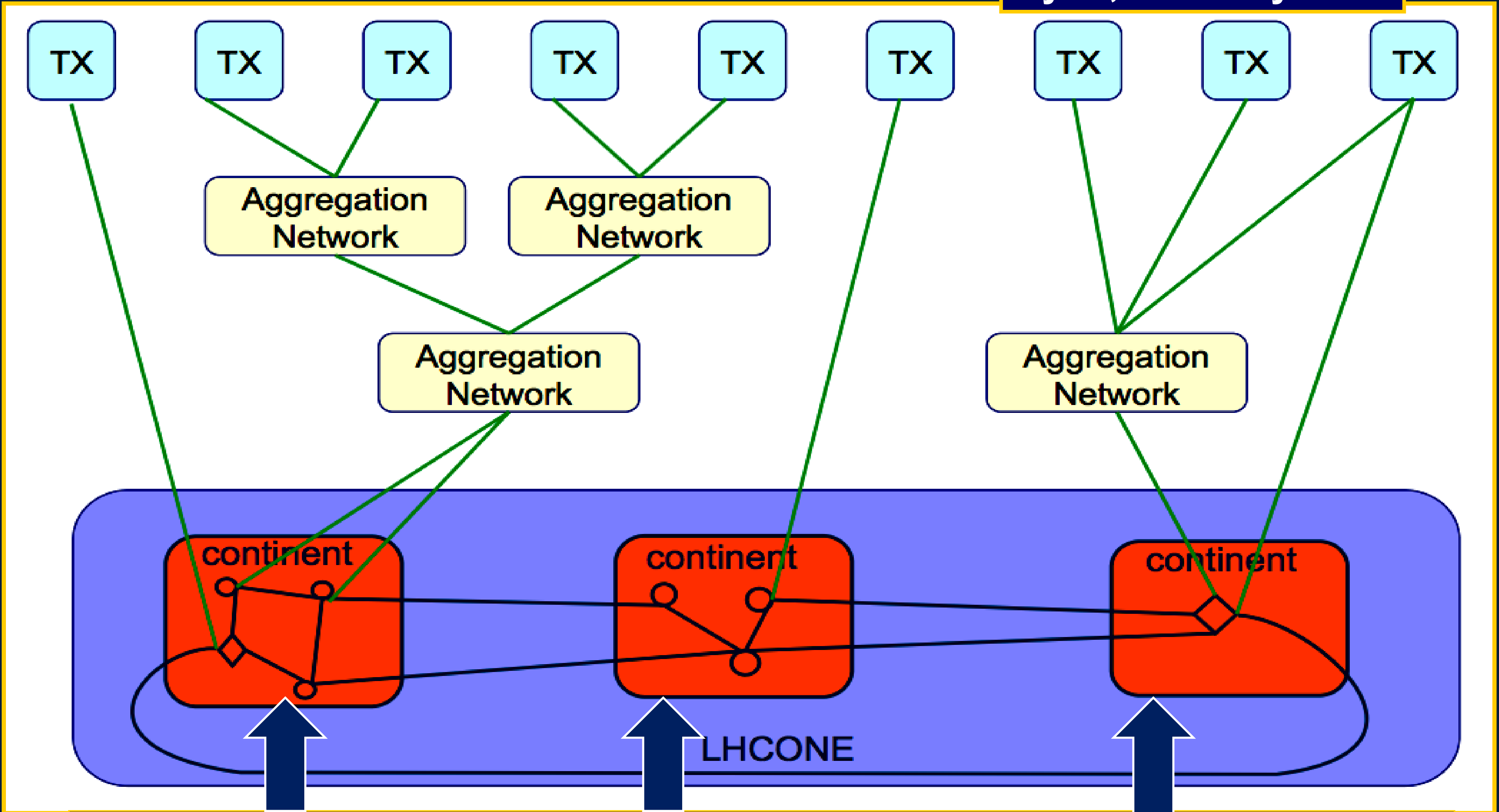
The move to the next generation of 40G and 100G networks is underway and will accelerate as 2012 progresses

LHCONE: 1 Slide Refresher

- In a nutshell, LHCONE was born (out the 2010 transatlantic workshop at CERN) to address two main issues:
  - To ensure that the services to the science community maintain their quality and reliability
  - To protect existing R&E infrastructures against potential "threats" of very large data flows that look like 'denial of service' attacks
- LHCONE is expected to
  - Provide some guarantees of performance
    - Large data flows across managed bandwidth that would provide better determinism than shared IP networks
    - Segregation from competing traffic flows
    - Manage capacity as # sites x Max flow/site x # Flows increases
  - Provide ways for better utilization of resources
    - Use all available resources, especially transatlantic
    - Provide Traffic Engineering and flow management capability
  - Leverage investments being made in advanced networking

LHCONE Initial Architecture, The 30'000 ft View

LHCOPN Meeting
Lyon, February 2011



Sets of Open Exchange Points

Timescales

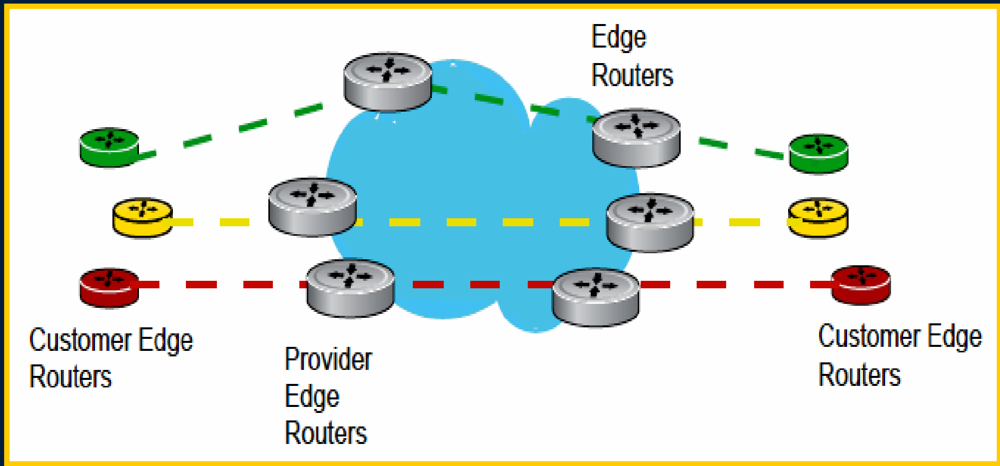
- **In the meantime, we've seen significant increase in backbone as well as GPN transatlantic capacity [as well as HEP traffic]**
  - **True in particular in US and Europe, but this should not lead us to forget that LHCONE is a global framework**
- **WLCG has encouraged us to look a at longer-term perspective rather than rush to implementation**
- **This timescale fits with the LHC Short-term Schedule:**
  - **2012: LHC run will continue through Feb. 2013**
  - **2013-2014: LHC shutdown (Feb. 2013), restart late 2014/beginning 2015**
  - ➡ ***2015: LHC data taking at ~nominal energy (13-14 TeV)***
- **The large experiment data flows will continue to grow:** ***developing effective means to manage such flows is needed***

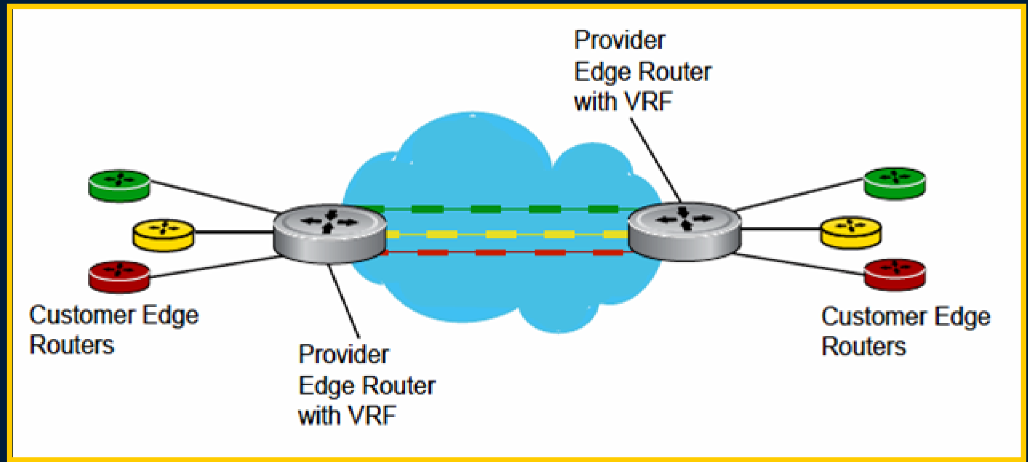
LHCONE Activities

- With the above in mind, LHCONE has defined the following activities:
  1. VRF-based multipoint service: a "quick-fix" to provide multipoint LHCONE connectivity, with logical separation from R&E GPN
  2. Layer 2 multipath: evaluate use of emerging standards such as TRILL (IETF) or Shortest Path Bridging (SPB, IEEE 802.1aq) in WAN environment
  3. Openflow: There was wide agreement at the workshop that SDN is the probable candidate technology for LHCONE in the long-term, however needs more investigations
  4. Point-to-point dynamic circuits pilots
  5. Diagnostic Infrastructure: each site to have the ability to perform E2E performance tests with all other LHCONE sites
- Plus, 6. Overarching: *Investigate impact of LHCONE dynamic circuits on LHC software stacks + computing site infrastructure*

VRF: Virtual Routing and Forwarding

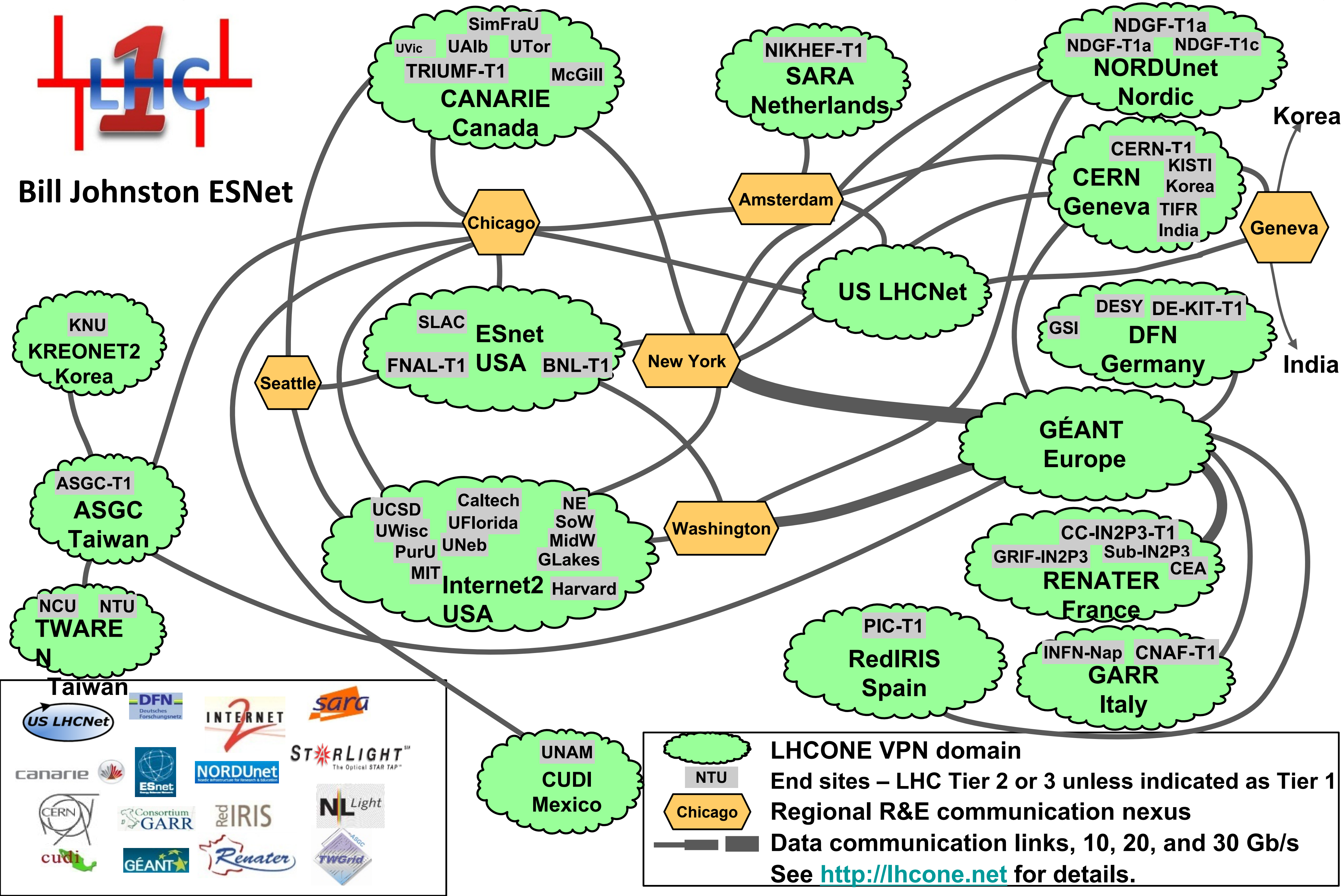
- VRF: in basic form, concerns the implementation of multiple logical router instances inside a physical device
- Logical control plane separation between multiple clients/tenants





- VRF approach in LHCONE: regional networks implement VRF domains to logically separate LHCONE from other flows
- BGP peerings used inter-domain and to the end-sites
- Some potential for Traffic Engineering
  - although scalability is a concern
- BGP communities defined for tuning path preferences

LHCONE: A global infrastructure for the LHC Tier1 Data Center – Tier 2 Analysis Center Connectivity



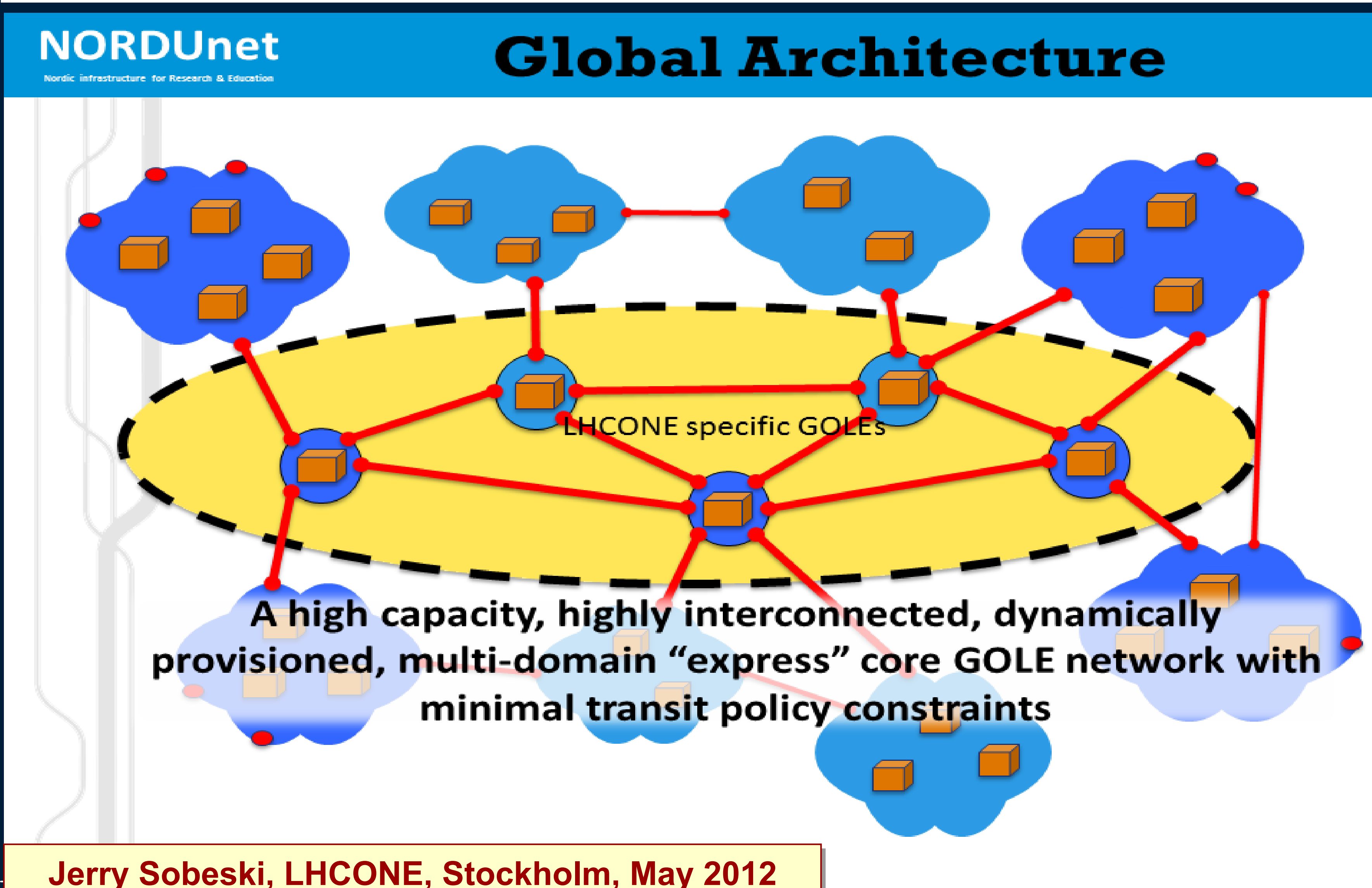
The Case for Dynamic Provisioning in LHC Data Processing

- Data models do not require full-mesh @ full-rate connectivity @ all times
- On-demand data movement will augment and partially replace static pre-placement → Network utilization will be more dynamic and less predictable, if not managed
- Need to move large data sets fast between computing sites; expected performance levels and time to complete operations will not decrease !
  - On-demand: caching
  - Scheduled: pre-placement
  - *Transfer* low-latency + predictability important for efficient workflow
- As data volumes grow, and experiments rely increasingly on the network performance; what will be needed in the future is
  - More efficient use of network resources
  - Systems approach including end-site resources and software stacks
- The solution for the LHC community needs to provide global reach

Point-to-Point Connection Service in LHCONE

- Service definition agreed on in LHCONE
- NSI definition is progressing well
  - See Plugfest NSI V 2.0 demo at this GLIF conference
- AutoGOLEs: Automatic lightpath stitching; could provide the dynamic inter-exchange-point fabric
  - All major R&E networks connect to GOLEs
- Build on nat'l & regional projects for the basic DC technology
  - OSCARS (ESnet, RNP), ION (Internet2), DRAC(SURFNet), AutoBAHN (some EU NRENs)
- Extending into campus:
  - DYNES (Switch and Control Server Equipment)
- Interfacing with LHC experiments/sites
  - DYNES (Software: FDT)
  - ANSE; new NSF funded project aiming at integration of Advanced Network Services with Experiments' data management/workflow SW
    - Caltech, UMICH, Vanderbilt, UTA

P2PCS: Point-to-Point Connection Service



Jerry Sobeski, LHCONE, Stockholm, May 2012

US: DYNES Project supporting LHC data movement

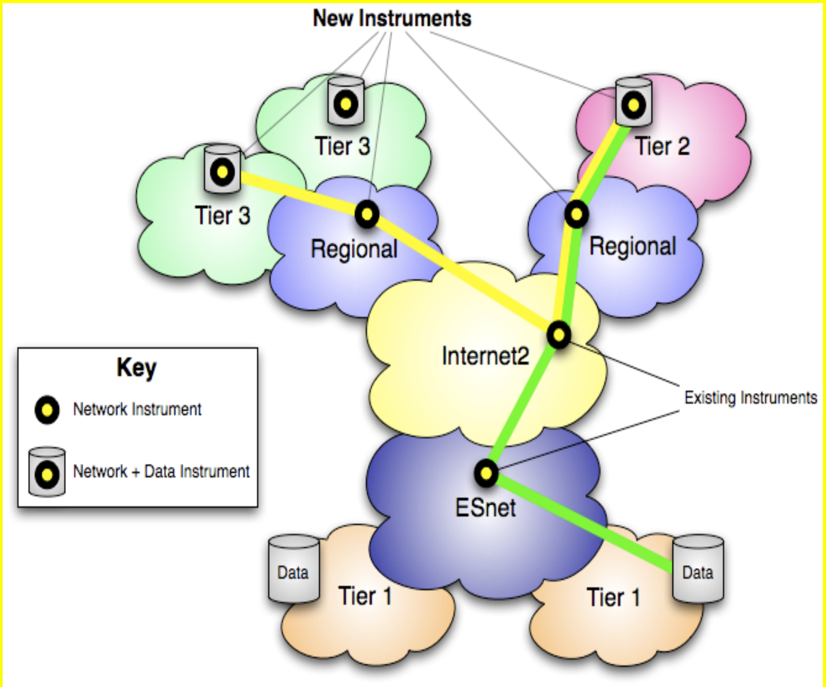
http://internet2.edu/dynes

- NSF funded: Internet2, Caltech, U. Michigan, Vanderbilt
- Nation-wide Cyber-instrument extending hybrid & dynamic capabilities (in production in advanced R&E nets such as Internet2 and ESnet) to campuses & regional networks
- Provides 2 basic capabilities at campuses and regional networks:

1. Network resource allocation such as bandwidth to ensure transfer performance
2. Monitoring of the network and data transfer performance

- Tier2 and Tier3 end-sites need in addition

3. Hardware sites capable of optimal use of the available network resources: IDC controller, switch, data server with FDT

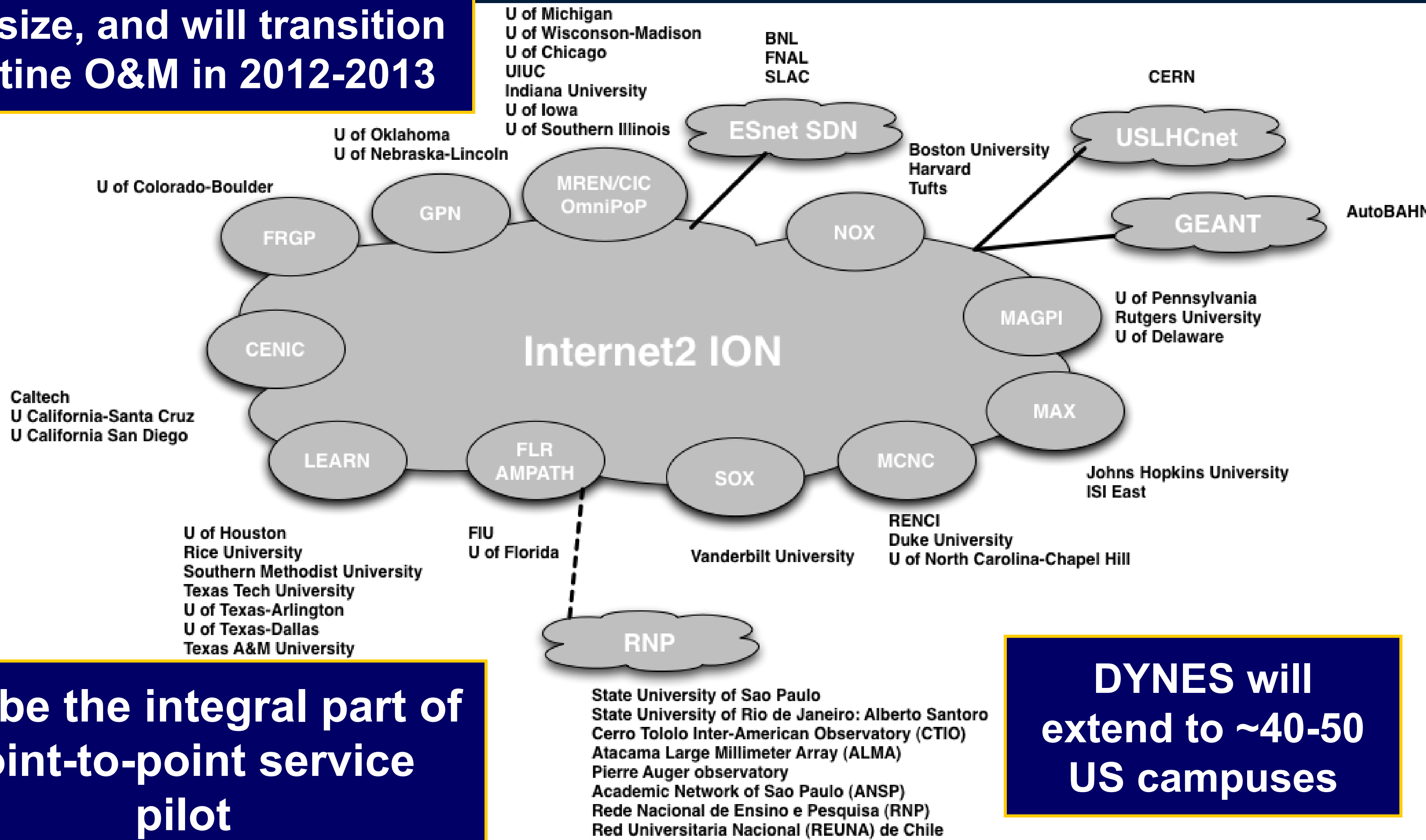


Two typical transfers that DYNES supports: one Tier2 - Tier3 and another Tier1-Tier2.

The Clouds represent the network domains involved in such a transfer.

DYNES Current Logical Topology

DYNES is currently scaling up to full size, and will transition to routine O&M in 2012-2013



Will be the integral part of point-to-point service pilot

DYNES will extend to ~40-50 US campuses

DYNES current status

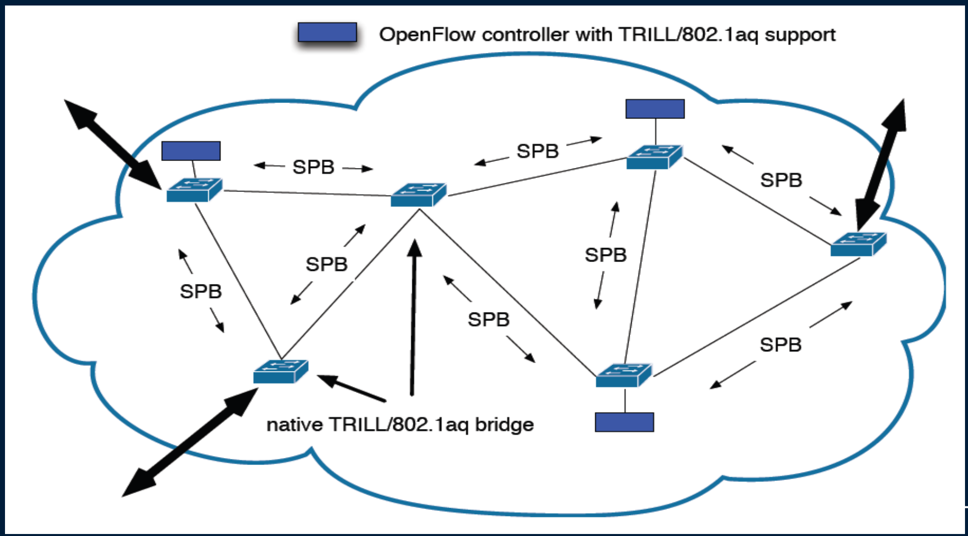
- Deploying at 49 sites (11 regional networks, 38 campuses)
  - completed: 33% (16 sites)
  - in progress: 43% (21 sites)
  - yet to be deployed: 24% (12)
- Beyond installation:
  - Deployment of performance test nodes at all sites
  - Exploring SDN capabilities of the Dell S4810 Switch, and its ability to run the OESS software
  - Exploring RoCE (RDMA over IB/Ethernet) network cards for use with the XSP library, developed by Indiana University

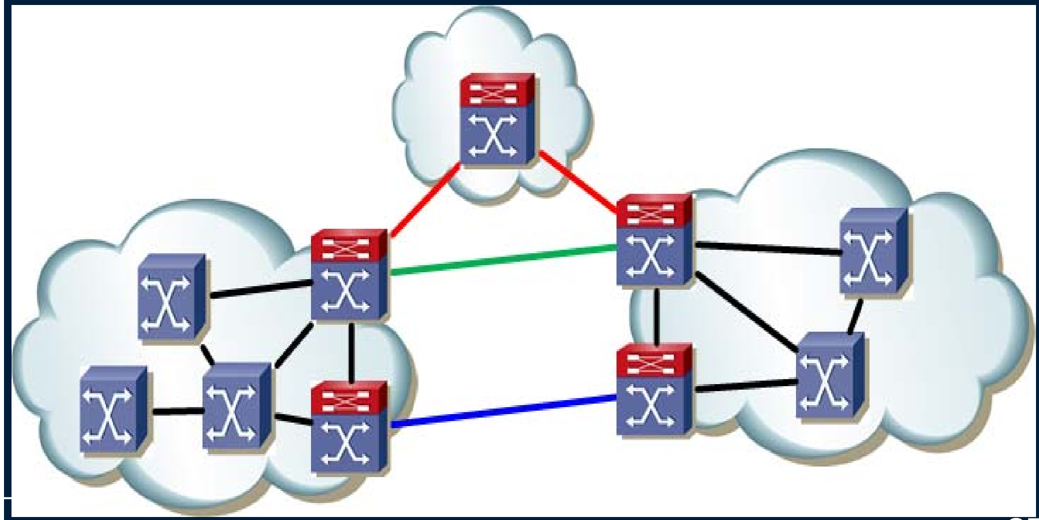
R&D: Solving the Multipath Challenge

- Practical issue in LHCONE:
- There are many possible paths between R&E domains
  - E.g. Trans-Atlantic: USLHCNet (6x10G), ACE/GEANT , NORDUnet, SURFNet
- How to most efficiently distribute the traffic over all these resources?
- No issue for Point-to-Point service
  - delegate to NSI to find available path
- But solution for multipoint-services is not obvious
  - Both at Layer 2 and Layer 3
- Layer 3 (VRF) can use some techniques known from BGP
  - MEDs, AS padding, local preferences, restricted announcements
  - They work in a reasonably small configuration
  - Traffic “control” is complex
  - Not clear if it will scale up to O(100) end-sites (AS’s)
- Layer 2 Multipoint (if considered for LHCONE) must be constrained to tree topology

Multipath in LHCONE

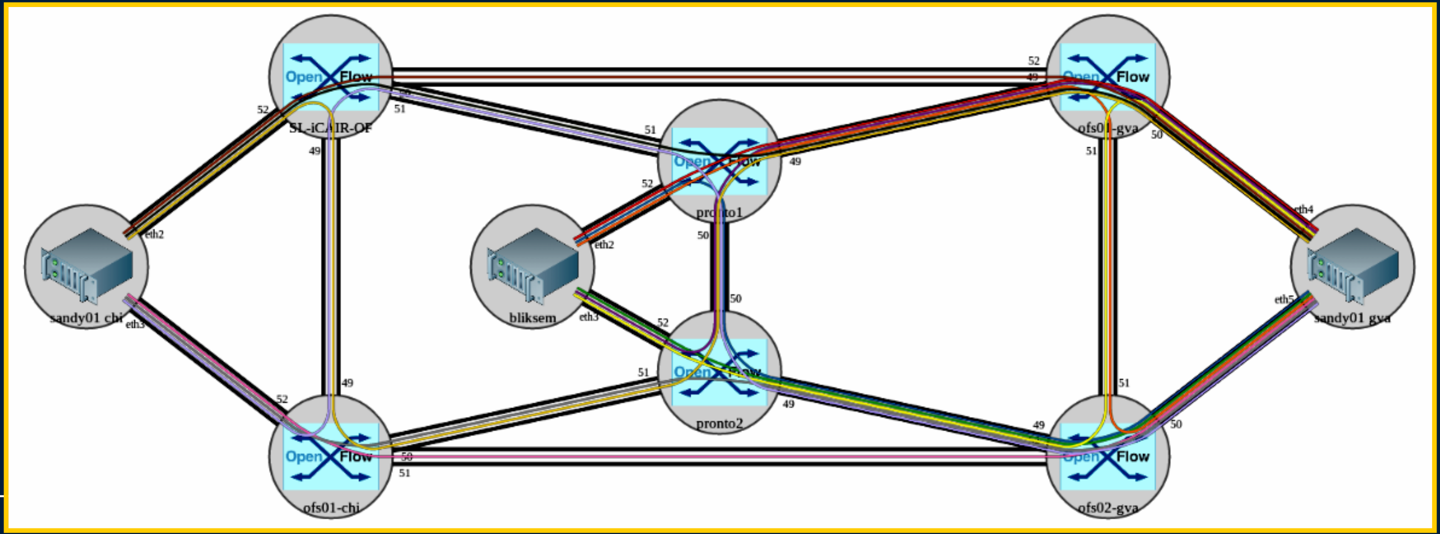
- **For LHCONE, in practical terms:**
  - How to use the many transatlantic paths at Layer 2 among the Many partners: USLHCNet, ACE, GEANT, SURFnet, NORDUnet, …
- **Technologies - Some approaches to Layer 2 multipath:**
  - IETF: TRILL (TRansparent Interconnect of Lots of Links)
  - IEEE: 802.1aq (Shortest Path Bridging)
- **None of those designed for WAN!**
  - Some R&D needed – OpenFlow is the chosen direction





Multipath with Openflow

- Started by Caltech and SARA
  - Caltech: OLiMPS project (DOE OASCR)
    - Implement multipath control functionality using Openflow
  - SARA: investigations in use of MPTCP
- Basic idea:
  - Flow-based load balancing over multiple paths
  - Initially: use static topology, &/or bandwidth allocation (e.g. NSI)
  - Later: real-time information from the network (utilization, topology changes)
  - MPTCP
- Demo NE02
- done yesterday
- at this GLIF Workshop



LHCONE, Networks and Users

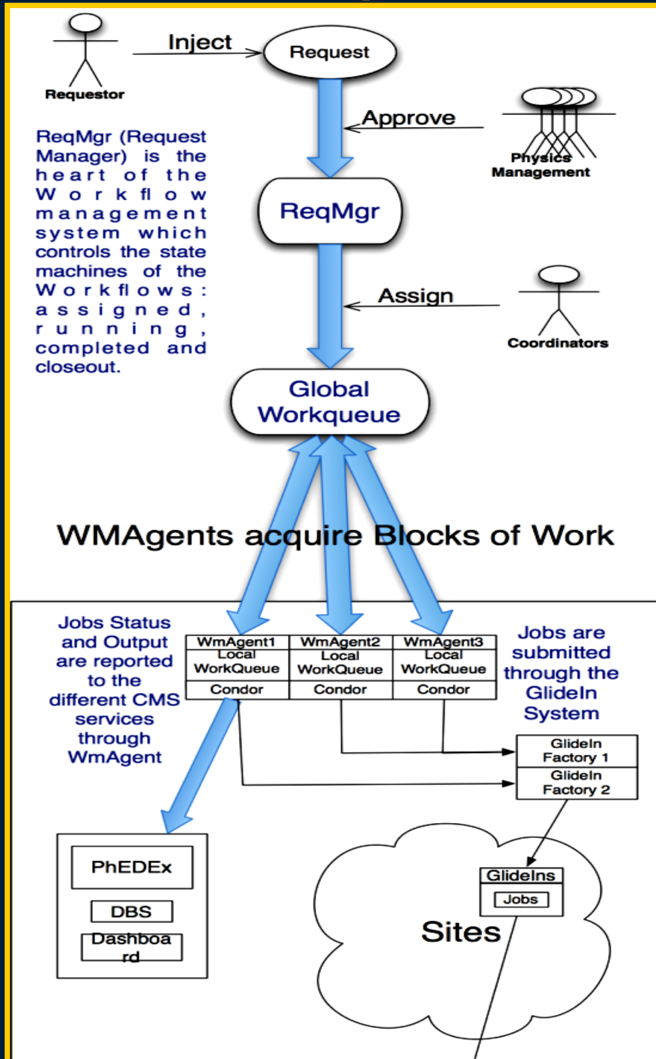
- **Point-to-point pilot implementation requires direct user involvement**

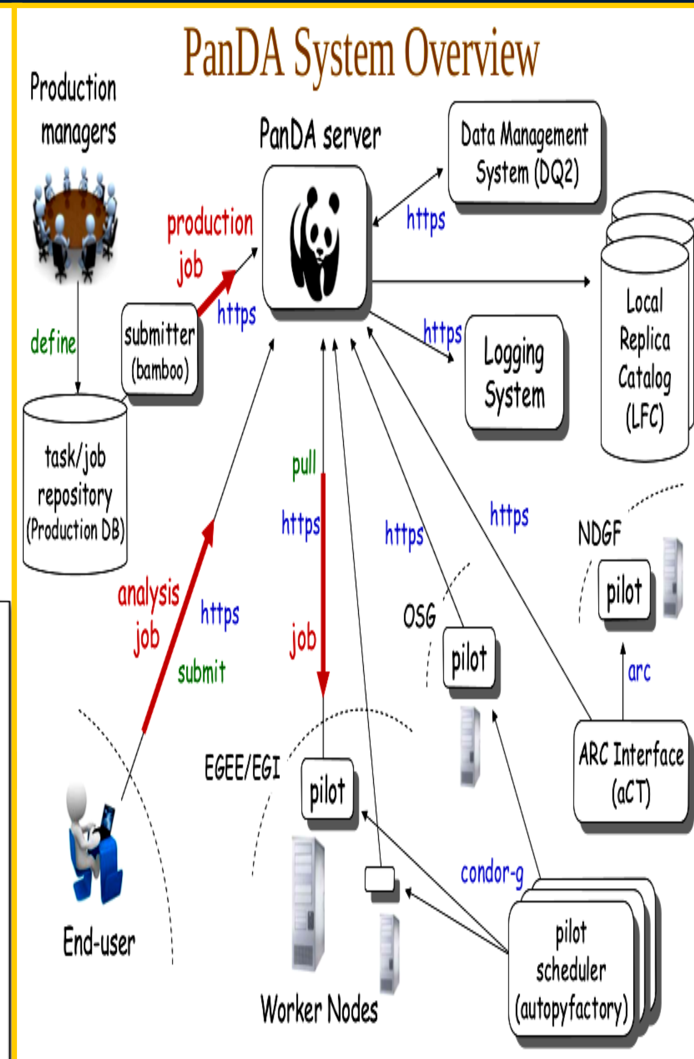
– LHCONE Activity 2

- **For effective use, need integration in LHC experiments' software, workflows & data management structures**

– (Could be) LHCONE Activity 6

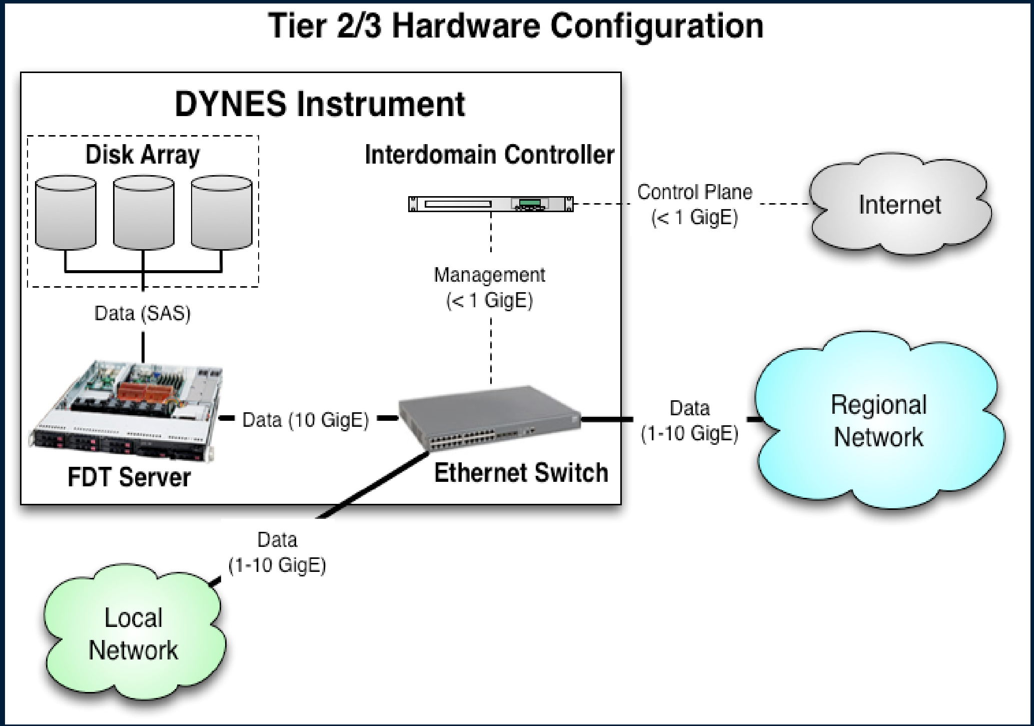
- **CMS: Distributed Workflow Mgmnt (DMWM) with PhEDEx for transfer management**
- **Atlas: …. Distributed Analysis (PaNDA)**





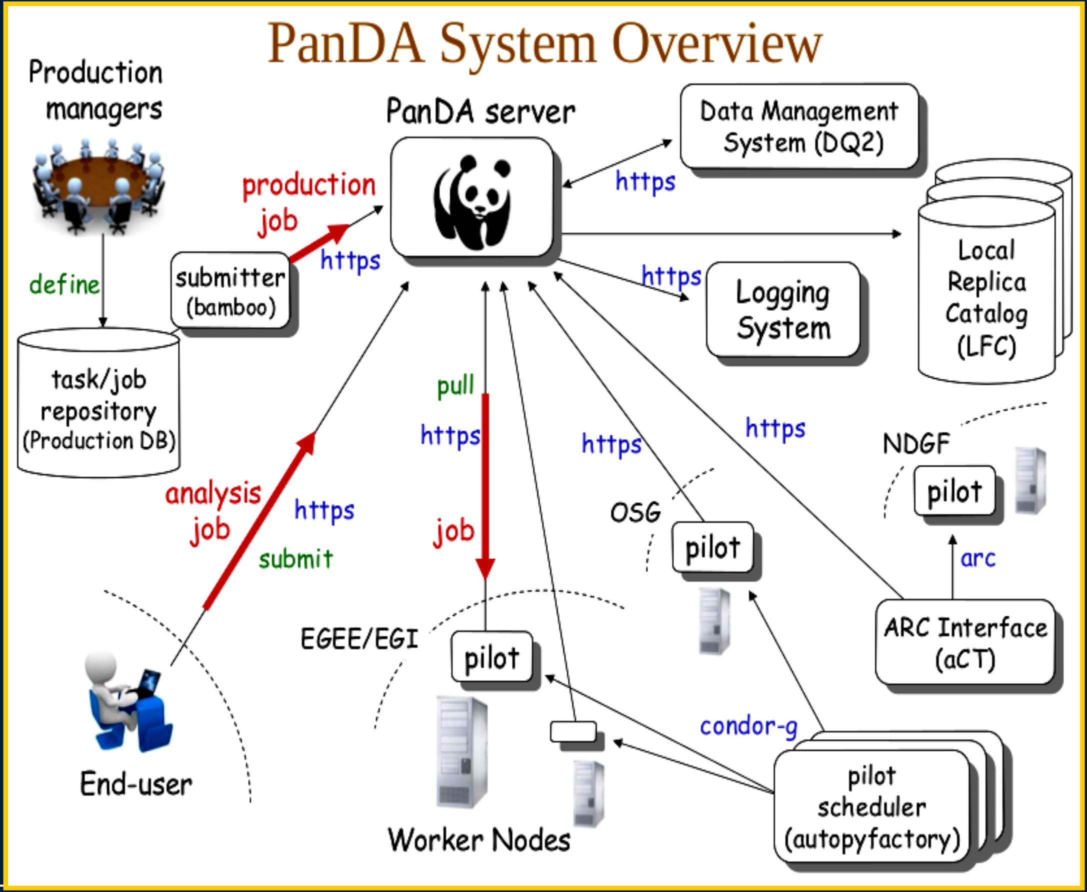
DYNES FDT deployment

- DYNES deployment includes data transfer application: FDT
- FDT uses the IDC API
  - Migration to NSI considered straight forward
- FDT has also been integrated with PhEDEx (in CMS)
- In theory (and soon in practice), US CMS sites could use “Bandwidth on demand”
  - Caveats: (1) No user-side capacity management (FDT calls API, gets resources if available, else use routed path)
  - (2) No advance reservation (other than through Web-GUI and manual operation)
- Could do more with *ANSE:* “Advanced Network Services for Experiments” ➡



ANSE: Advanced Management of LHC data flows

- *Advanced use of dynamic circuits requires higher-levels in software stack to interact with the network*
- **Earlier projects in this area: see Terapaths and StorNet (US ATLAS)**
- **ANSE: NSF funded project**
- **US CMS and US ATLAS collaboration**
  - **Caltech, Vanderbilt, Univ. of Michigan, UT Arlington**
- **Interface advanced network services with LHC data management middleware**
  - *PanDA in (US) Atlas*
  - *Phedex in (US) CMS*



Conclusions

- The LHC computing and data models continue to evolve towards more dynamic, less structured, on-demand data movement
  - large data transfers (requiring high throughput) are complemented by remote data access (latency sensitive)
- LHCONE is on a dual-track:
  - Multipoint VRF implementation: now transitioning to operations
  - Work on innovative technologies, centered around dynamic circuits is advancing in the architecture group
    - Point-to-point services, Openflow, Multipath, Exp. Interface
    - OGF-NSI is a key element
- Synergistic projects such as DYNES are complementing LHCONE activities
- *We are engaging the LHC experiments to implement increased network-awareness and interaction in their data management software stacks: Targeted at LHC restart at full energy in 2014-15*

THANK YOU!

newman@hep.caltech.edu