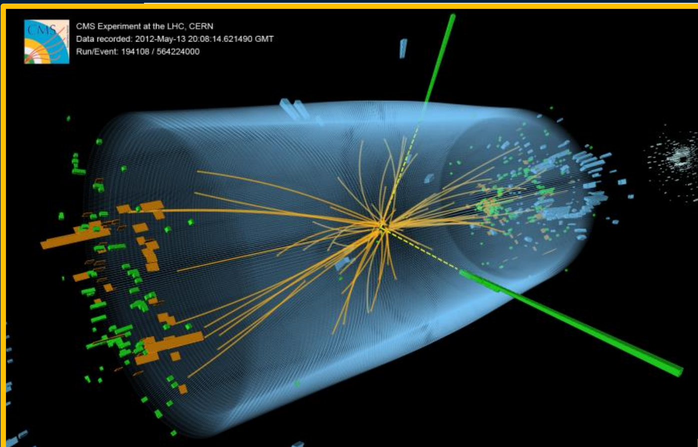
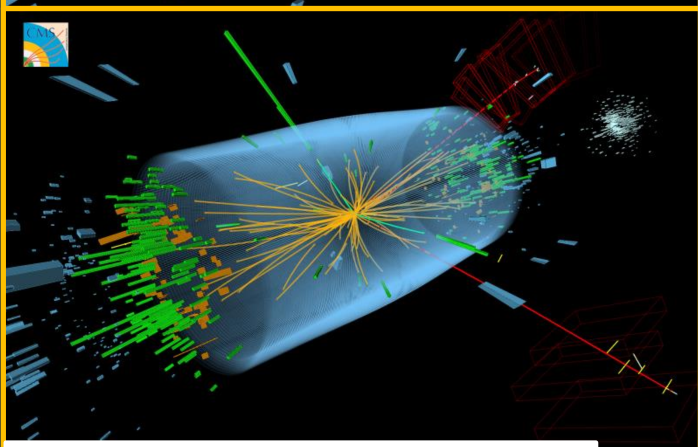
Networking for HEP in the LHC Era:

Global-Scale Developments for Data Intensive Science





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

Gateway to a New Era

50 Vertices, 14 Jets, 2 TeV

Harvey B Newman, Caltech

International School of Physics

"Enrico Fermi": Lecture 1

Networking for HEP in the LHC Era: Global-Scale Developments for Data Intensive Science

- Introduction: Physics Discovery and the role of networks: Historical retrospective
- Network Evolution and Revolution: A new scale during Run 2 (2015-18)
- The LHC Computing Models continue to evolve rapidly
- LHCONE: responding to the changing needs
- Moving Forward – Innovation examples: DYNES, ANSE, OliMPS; SDN
- High Speed Data Transfers: The State of the Art
- The Long View: Challenges and Approaches for the next decade
- Internet World Trends: Usage, Penetration, Traffic Growth & Quality
- ICFA SCIC: A World View of Networks, Trends and Developments; Working to Close the Digital Divide
- SCIC Monitoring WG: Quantifying the Digital Divide
- Closing the Divide Dark with Fiber Networks
- Digital Divide: Model Cases and Problem Areas
- Conclusions

Discovery of a Higgs Like Boson
July 4, 2012

2013

Englert Higgs

Theory : 1964

LHC + Experiments Concept: 1984

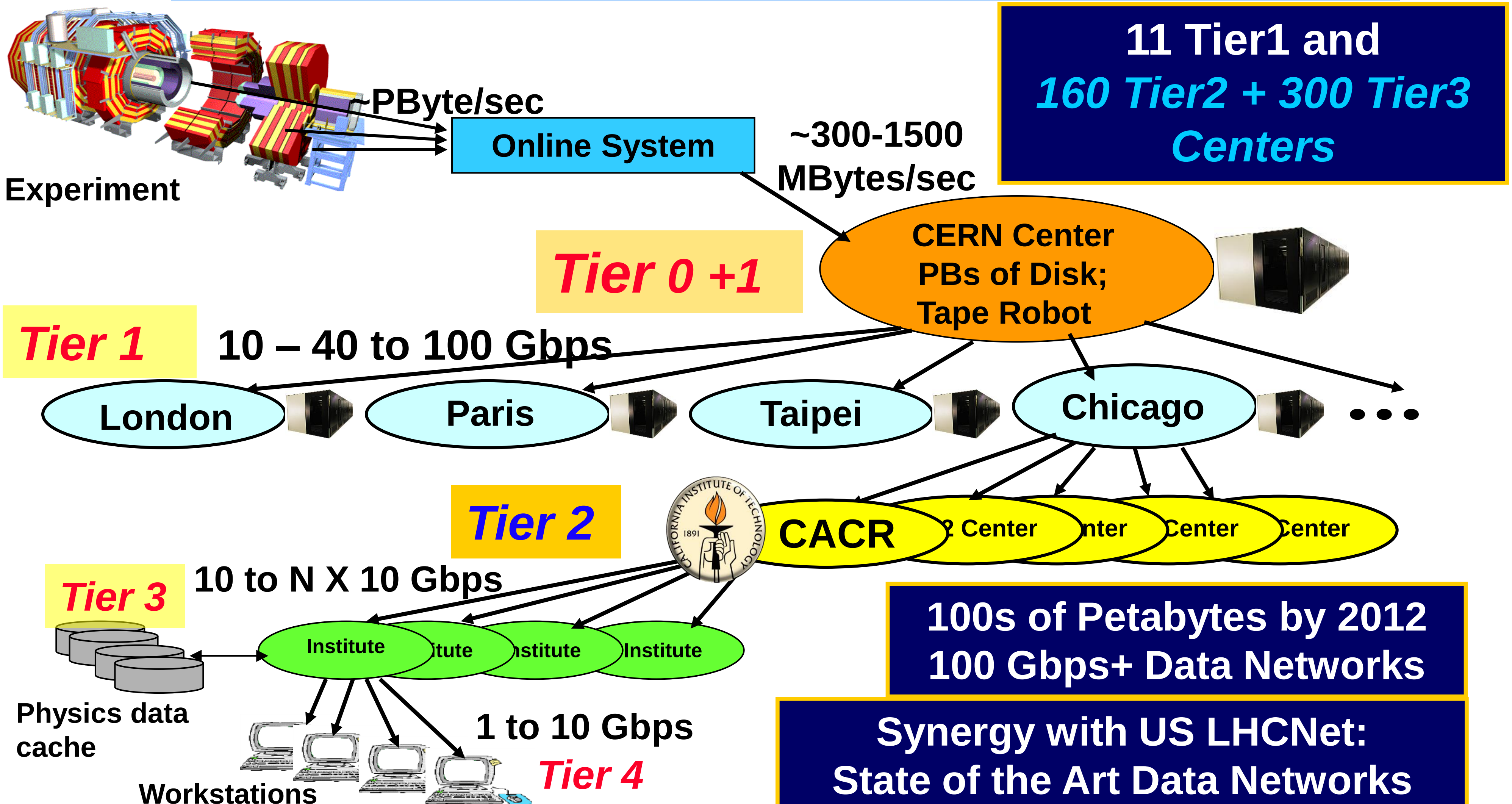
Construction: 2001

Operation and Discovery: 2009-12

A billion people watched

Highly Reliable High Capacity Networks Had an Essential Role in the Higgs Discovery…
And will in Future Discoveries

LHC Data Grid Hierarchy: A Worldwide System Invented and Developed at Caltech (1999)

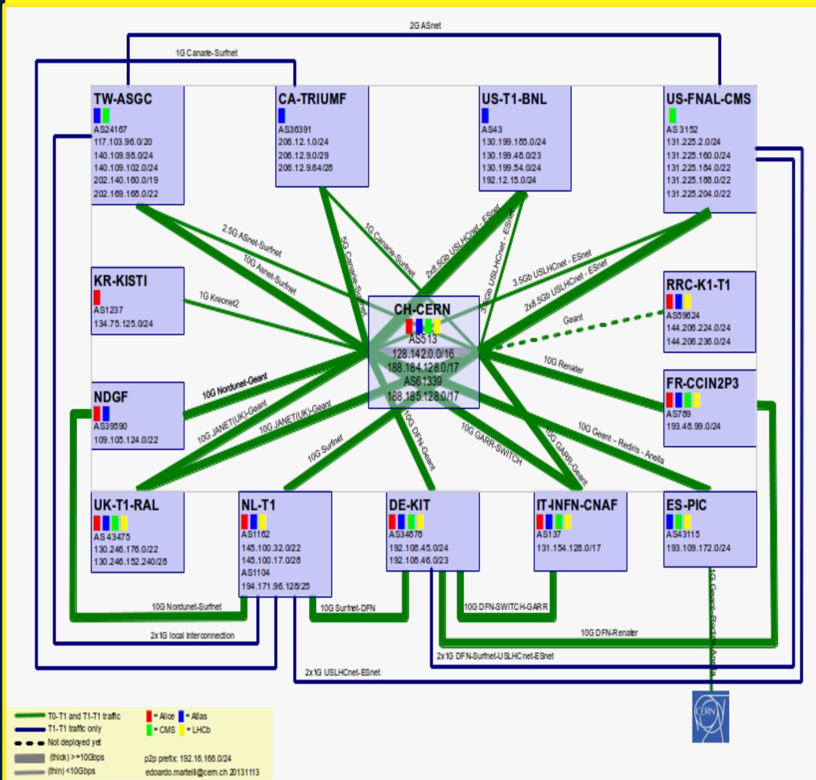


A Global Dynamic System

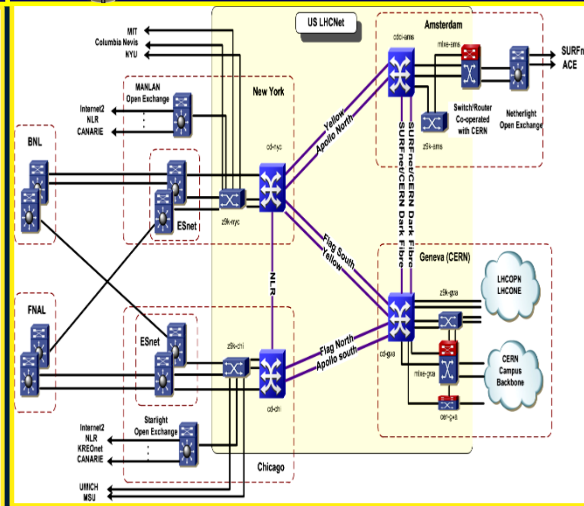
A New Generation of Networks: LHCONE, ANSE

The Core of LHC Networking: LHCOPN and Partners

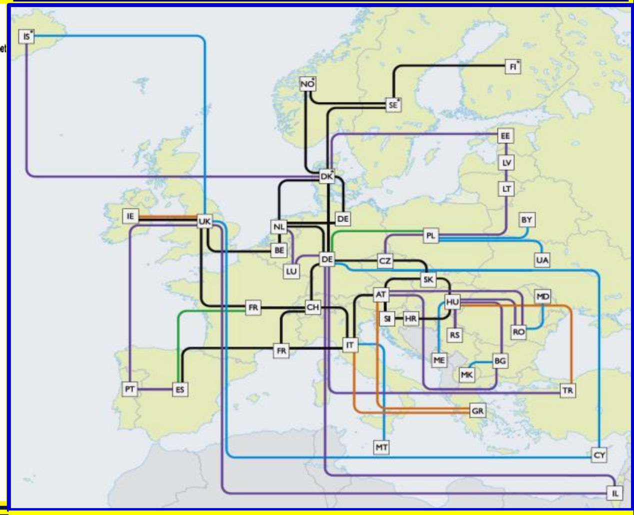
LHCOPN



US LHCNet

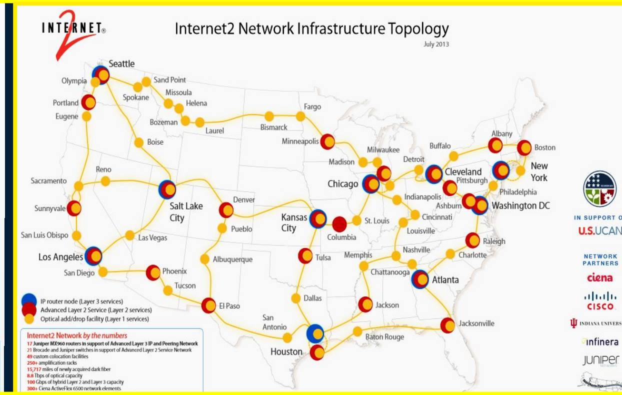


GEANT: 100G Core from 2013



Esnet: 100G from 2012

Internet2: 100G from 2013



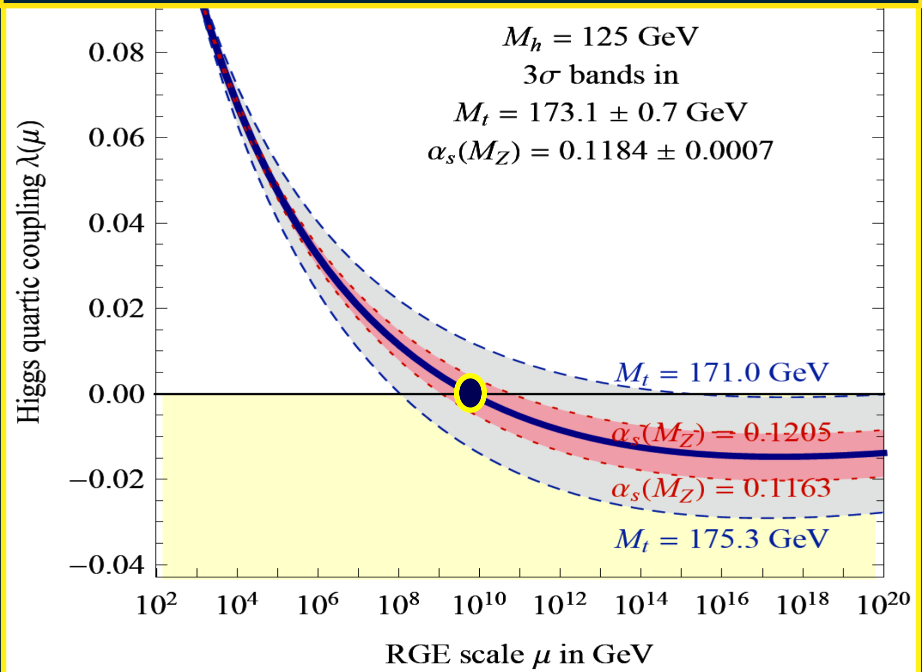
Simple and Highly Reliable, for Tier0 and Tier1 Operations

+ NRENs in Europe, Asia, Latin America, Au/NZ; US State Nets

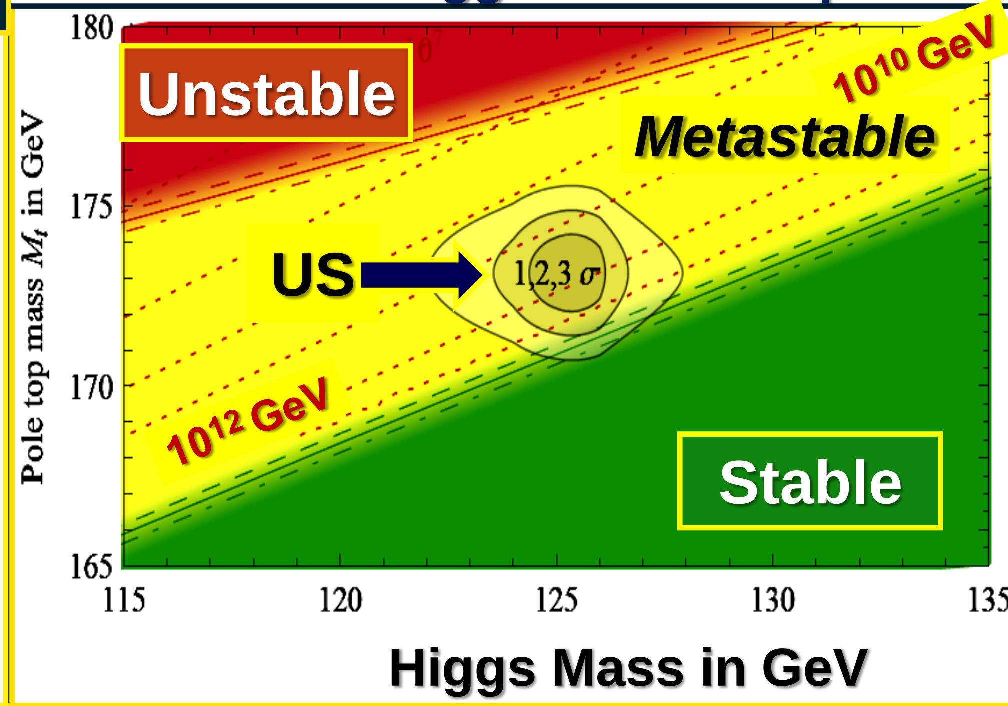
50 100G Connections by 2015

The 125-6 GeV Higgs Mass
Are we just on the wrong side of the Vacuum Stability Bound ?

NNLO Evolution of the Higgs Self-coupling λ(μ)

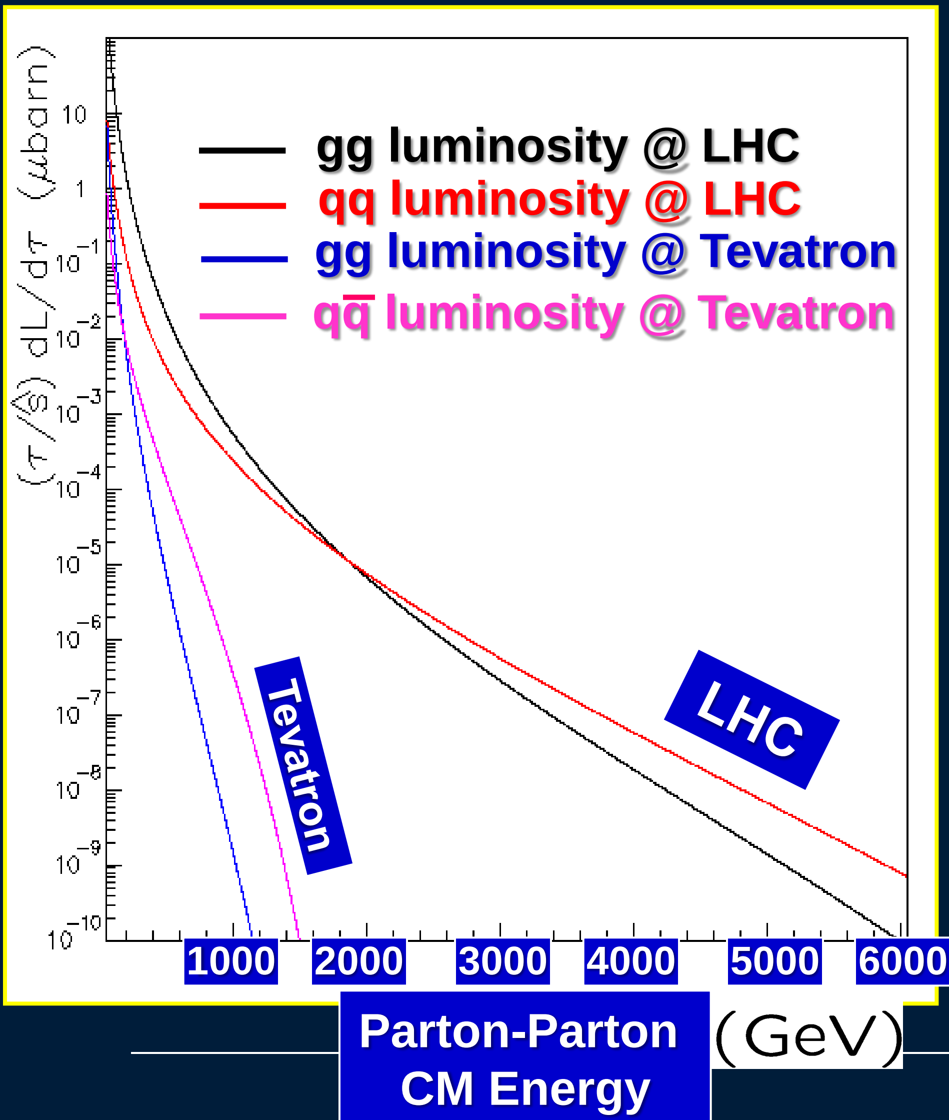


Precise Knowledge of the Top Mass as well as the Higgs Mass is Important



- For a Higgs mass of ~125 GeV
- ➡ λ goes negative ➡ Vacuum we are in is *metastable… ??*
- ➡ *OR:* New physics at an intermediate energy scale ~$10^{10\text{-}12}$ GeV
- *What lies between us and the Big Bang ?*

Opening a Realm of High Energies and a New Era of Discovery



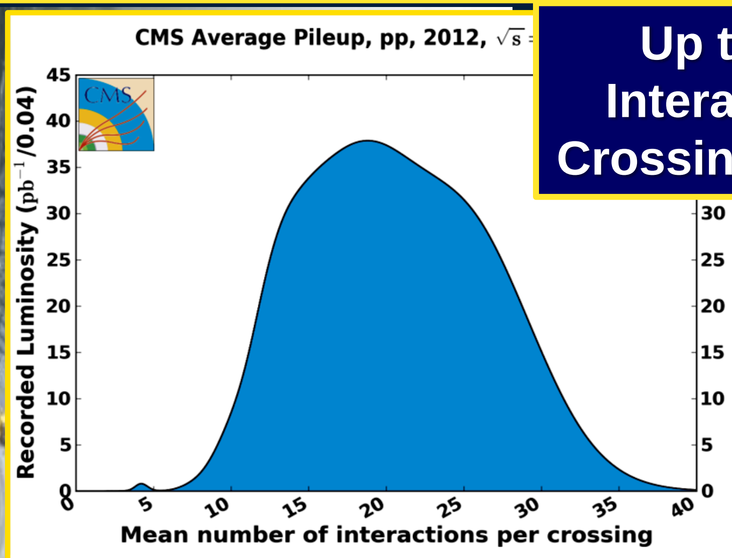
- The LHC is a Discovery Machine
- The first accelerator to probe deep into the Multi-TeV scale
- Its mission is *Beyond the SM*
- There are many reasons to expect new physics

SUSY, Substructures, *Graviton Resonances, Black Holes, Low Mass Strings, … the Unexpected*

We do not know what we will find

The LHC: Spectacular Performance

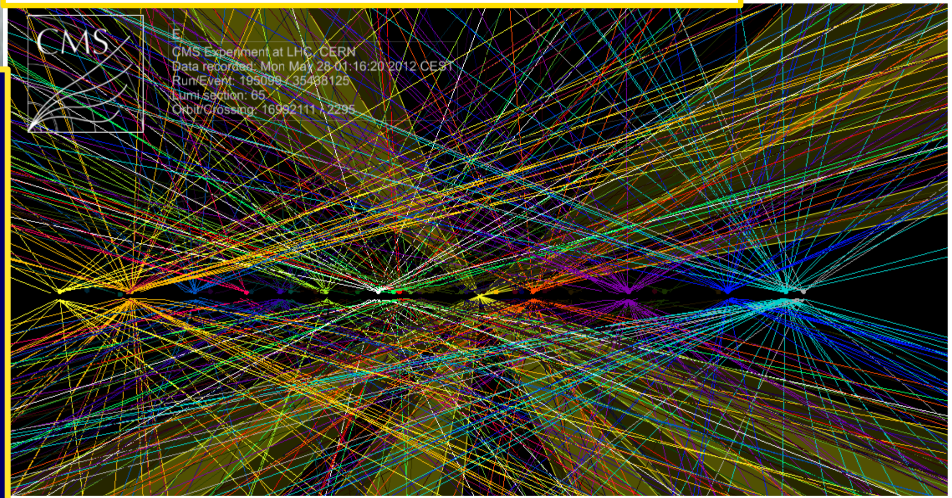
> Design Luminosity: The Challenge of Pileup



Up to ~50 Interactions/ Crossing in 2012

The Next Run will bring:

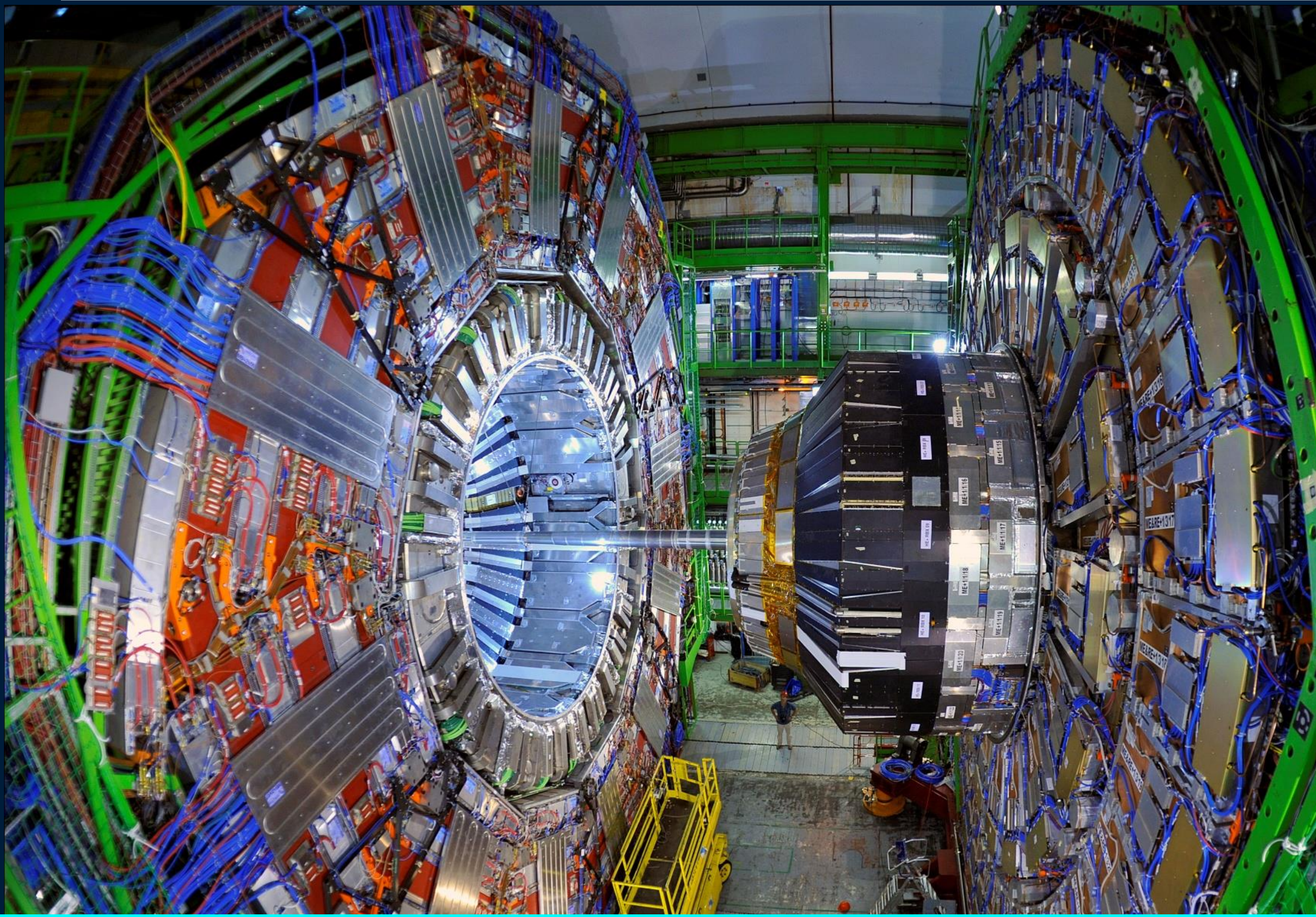
- Higher energy and intensity
- Greater science opportunity
- Greater data volume & complexity
- A new realm of challenges



~50 Vertices, 14 Jets, 2 TeV

CMS: Preparing for a New Era of Physics

What will Nature reveal at 13-14 TeV ?



A Time of Opportunity: for the Next Round of Discoveries

A Time of Challenge Requiring New Technology Advances

Foundations: Caltech Network Team Milestones + 28 Yrs Working with CERN

- 1982: Caltech initiated Transatlantic networking for HEP in 1982, 1982-5: First HEP experience with packet networks (US-DESY)
- 1985-6 Networks for Science: NSFNET, IETF, National Academy Panel
- 1986 Assigned by DOE to operate LEP3Net, the first US-CERN leased line, multiprotocol network for HEP (9.6 – 64 kbps)
- 1987-8: Hosted IBM: they provided the first T1 TA US-CERN link ($3M/Yr)
- 1989-1995: Upgrades to LEP3Net (X.25, DECNet, TCP/IP): 64 – 512 kbps
- 1996 - 2005: USLIC Consortium (Caltech – CERN – IN2P3 – WHO – UNICC). Based on 2 Mbps Links, then ATM, then IP optical links
- 1997: Hosted Internet2 CEO ; CERN Internet2's first Int'l member
- 1996-2000: Created LHC Computing Model (MONARC), & Tier2 Concept
- 2002 – Present: HN serves on ICFA as Chair of the Standing Committee on Inter-regional connectivity: Network Issues, Roadmaps, Digital Divide
- 2006 – Present: US LHCNet, co-managed by CERN and Caltech; Links at 2.5G; then 10G; then 2, 4, 6 10G links. Resilient service.
- Spring 2006 – Present: Caltech took over the primary operation and management responsibility, including the roadmaps and periodic RFPs

Bandwidth Growth of Int'l HENP Networks (US-CERN Example)

- *Rate of Progress >> Moore's Law in 1995-2005 (US-CERN Example)*

9.6 kbps Analog	(1985)	
64-256 kbps Digital	(1989 - 1994)	[X 7 – 27]
1.5 Mbps Shared	(1990-3; IBM)	[X 160]
2 -4 Mbps	(1996-1998)	[X 200-400]
12-20 Mbps	(1999-2000)	[X 1.2k-2k]
155-310 Mbps	(2001-2)	[X 16k – 32k]
622 Mbps	(2002-3)	[X 65k]
2.5 Gbps λ	(2003-4)	[X 250k]
10 Gbps λ	(2005)	[X 1M]

- *A factor of ~1M over a period of 1985-2005 (a factor of ~5k during 1995-2005)*
- *HEP has become a leading applications driver, and also a co-developer of global networks*

Originating the Global Computing and System Concepts for the LHC Experiments

- **Our team has originated and provided many of the key network-related computing concepts and global system deployments underpinning the LHC program, as well as the preceding program (LEP: 1984-2000)**
- **Created the LEP Computing Model in 1984**
  [Unix workstations, Special processors on VME channels, Networks]
- **Developed the first web-based global collaborative software systems:**
  VRVS (1996) ➡ EVO (2006) ➡ Seevogh (2012 to Present)
- **Originated the Computing plans (TDRs) for US CMS and CMS based on "Regional Centers" (1996-1998); SCB Chair through 2001**
- **Led the MONARC project: *Models Of Networked Analysis at Regional Centres* that defined the Computing Model for the LHC Experiments**
  - **Developed the MONARC Simulation System:**
    Leading to the MonALISA system: Monitoring Agents in a Large Integrated Services Architecture
    to monitor/control real global-scale distributed systems
- **Created the Globally Distributed LHC Computing Model: 1999-2000**

Network Evolution and Revolution

A New Scale by LHC Run2 in 2015 and Beyond

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
- Other bandwidth growth projections and trends are larger, so we need 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 5-10 within next 5 yrs ➡100G Target for each Tier2
- ➡ Longer Term Trends: Fisk and Shank at Snowmass showed how 100X growth in storage and network needs by LHC Run3 is possible

ATLAS Data Flow by Region: 2009-2014

Excellent Grid performance was crucial for fast discovery of the Higgs boson.

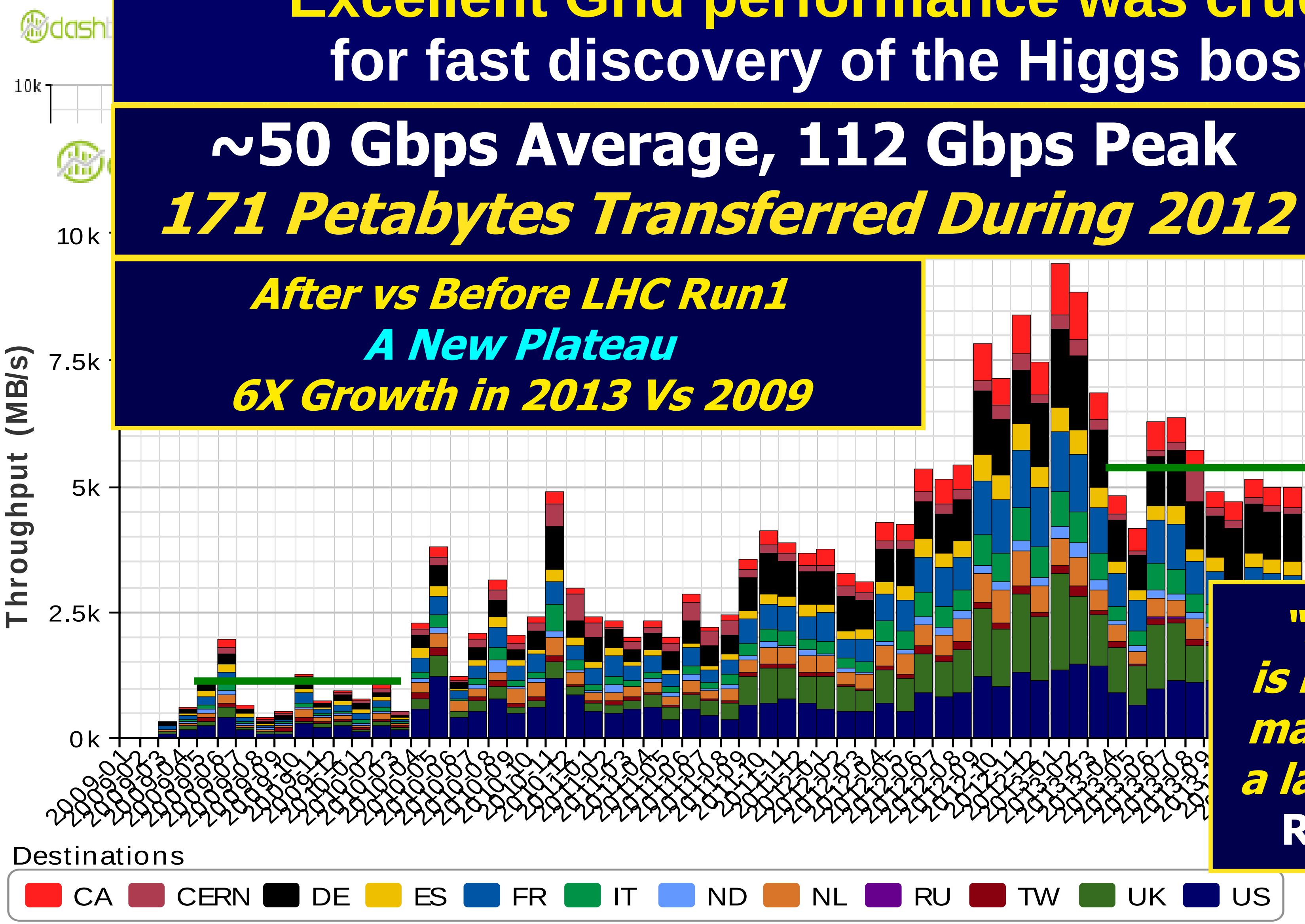
~50 Gbps Average, 112 Gbps Peak

171 Petabytes Transferred During 2012

After vs Before LHC Run1
A New Plateau
6X Growth in 2013 Vs 2009

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

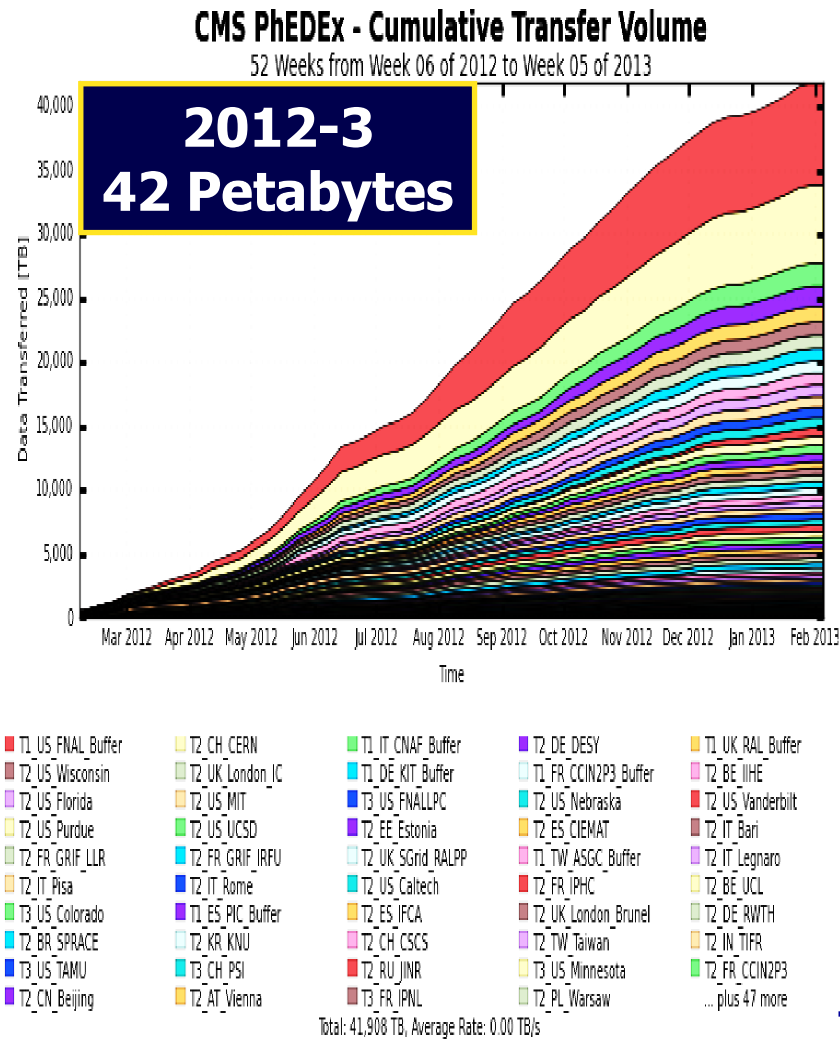
"10 Gbps is becoming marginal for a largeTier2"
R. Mount

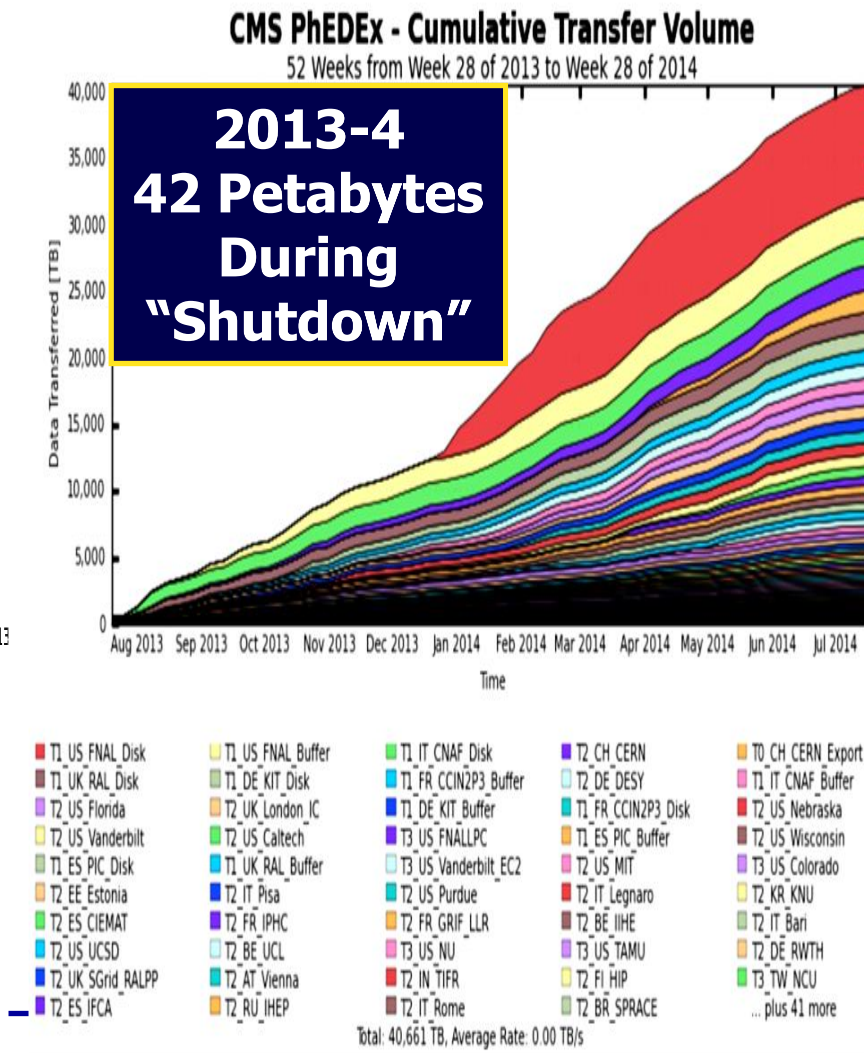


CMS Data Transfer Volume (2012– 2014)

> 80 PetaBytes Transferred Over 24 Months
= 10 Gbps Average (>20 Gbps Peak)

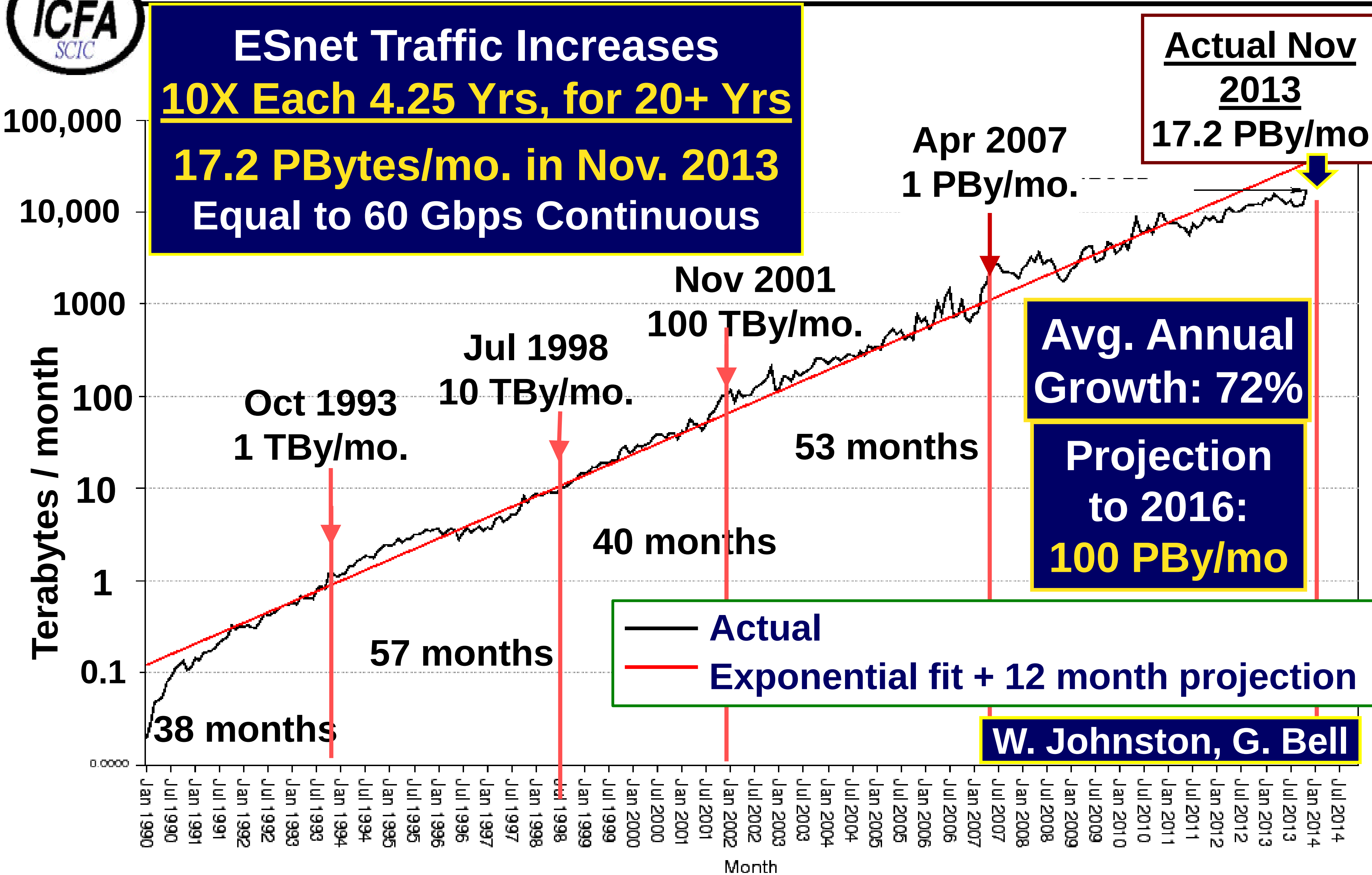
2012 Versus 2011 +45%





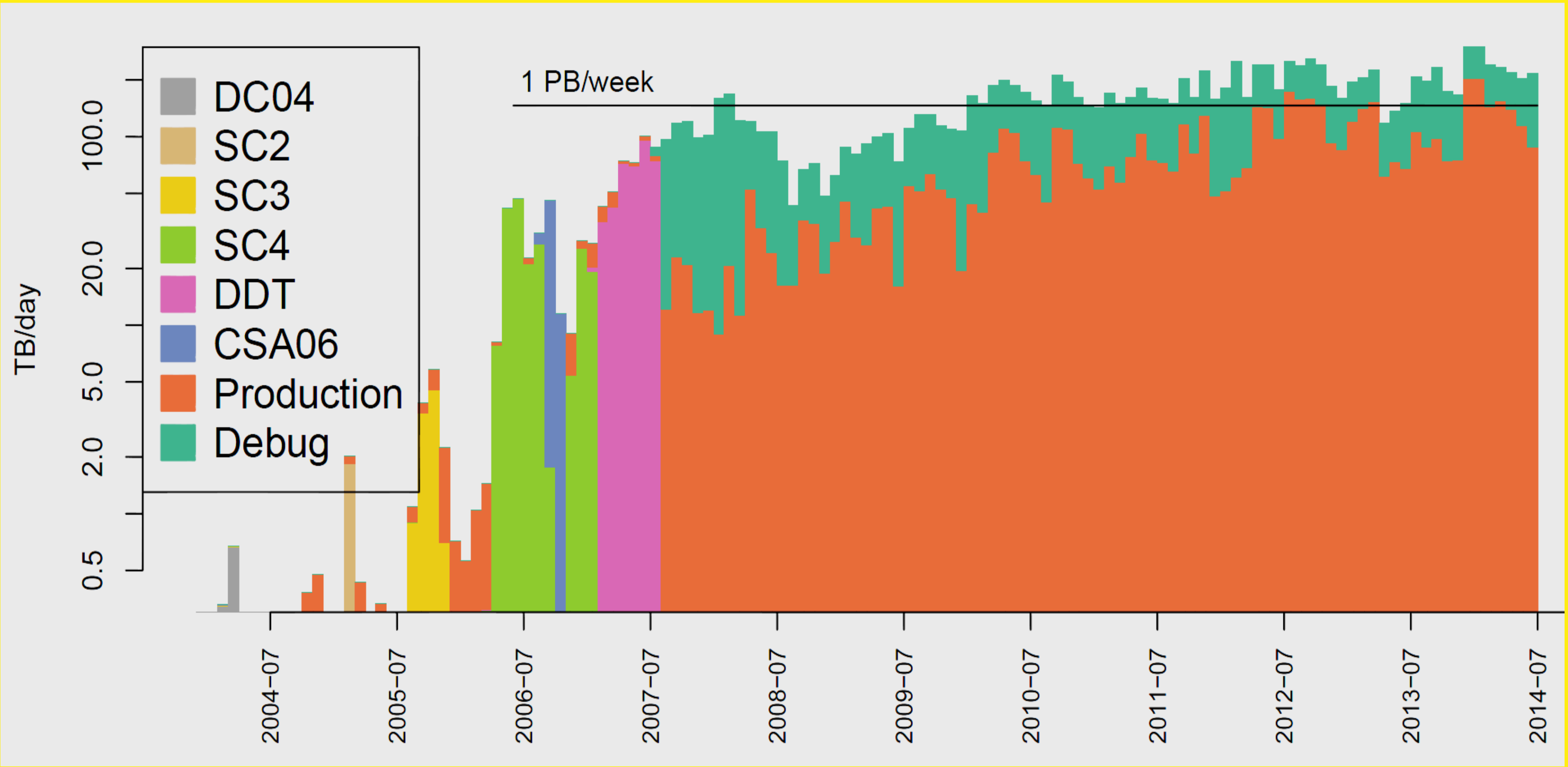
Higher Trigger Rates + Larger Events: Greater Transfer Volume in 2015

Remarkable Historical ESnet Traffic Trend Cont'd in 2013



Log Plot of ESnet Monthly Accepted Traffic, January 1990 – November 2013

PhEDEx: 10+ Years of Data Transfers in CMS



In Addition: "Location Independent Access" (AAA)

To 3 PB/Week LHC Data Taking is Not the Only Driver
Larger Data Flows are Ahead, During LHC Run 2

Network Trends in 2013-14

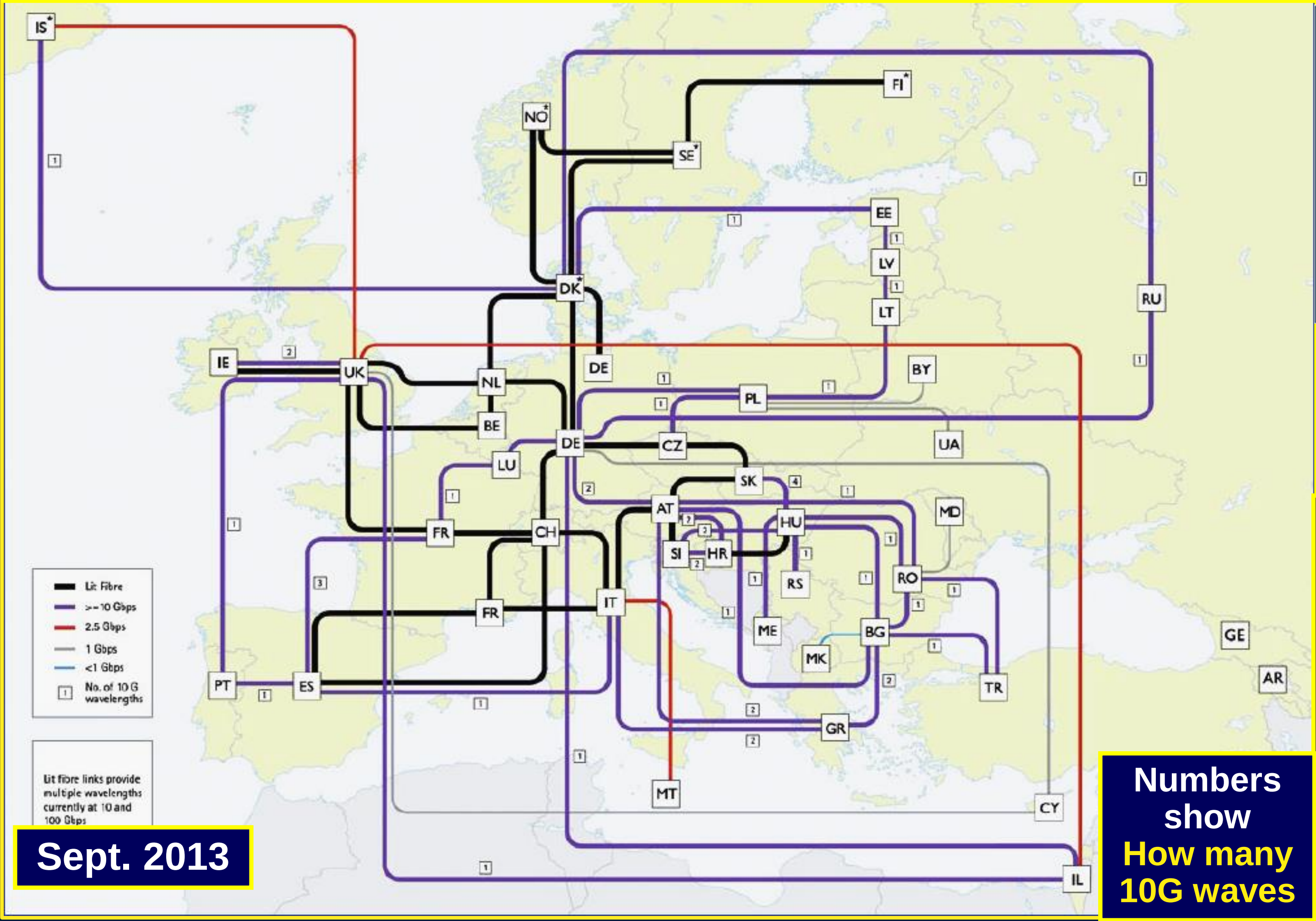
100G Evolution; Optical Transmission Revolution

- Transition to 100G next-generation core backbones: Completed in Internet2 and Esnet in 2012; 100G endsites are proliferating !
- GEANT transition to 100G: Phase 1 already completed in 2013
- Increased multiplicity of 10G links in Many other R&E networks: Internet2, ESnet, GEANT, and leading European NRENs
- 100G already appeared and spreading in Europe and Asia: e.g. SURFnet – CERN; Romania, Czech Republic, Hungary, Poland, China, Korea
- 100G Transatlantic Research Link ANA-100 in use from Fall 2013
- Proliferation of 100G network switches and high density 40G data center switches: 40G servers with PCIe 3.0 bus. *Now awaiting 100G*
- Higher Throughput: 340 G at SC12 and 13 - Caltech, UVic, et al.
- Software Defined Networks (Openflow; OpenDaylight): *A Paradigm Shift* taken up by much of industry and the R&E network community
- Advances in optical network technology even faster: denser phase modulation; *400G production trial (RENATER); 1 Petabit/sec on fiber*

The move to 100G networks is advancing, and accelerating; 400G networks are not so far away

GÉANT Pan-European Backbone
50M Users at 10k Institutions

50 kkm backbone fully migrated to 100G in 2013
5 NRENS Connect to the backbone at 70 Gbps or above
CERN – Wigner Data Center (HU) 100G Already in Service



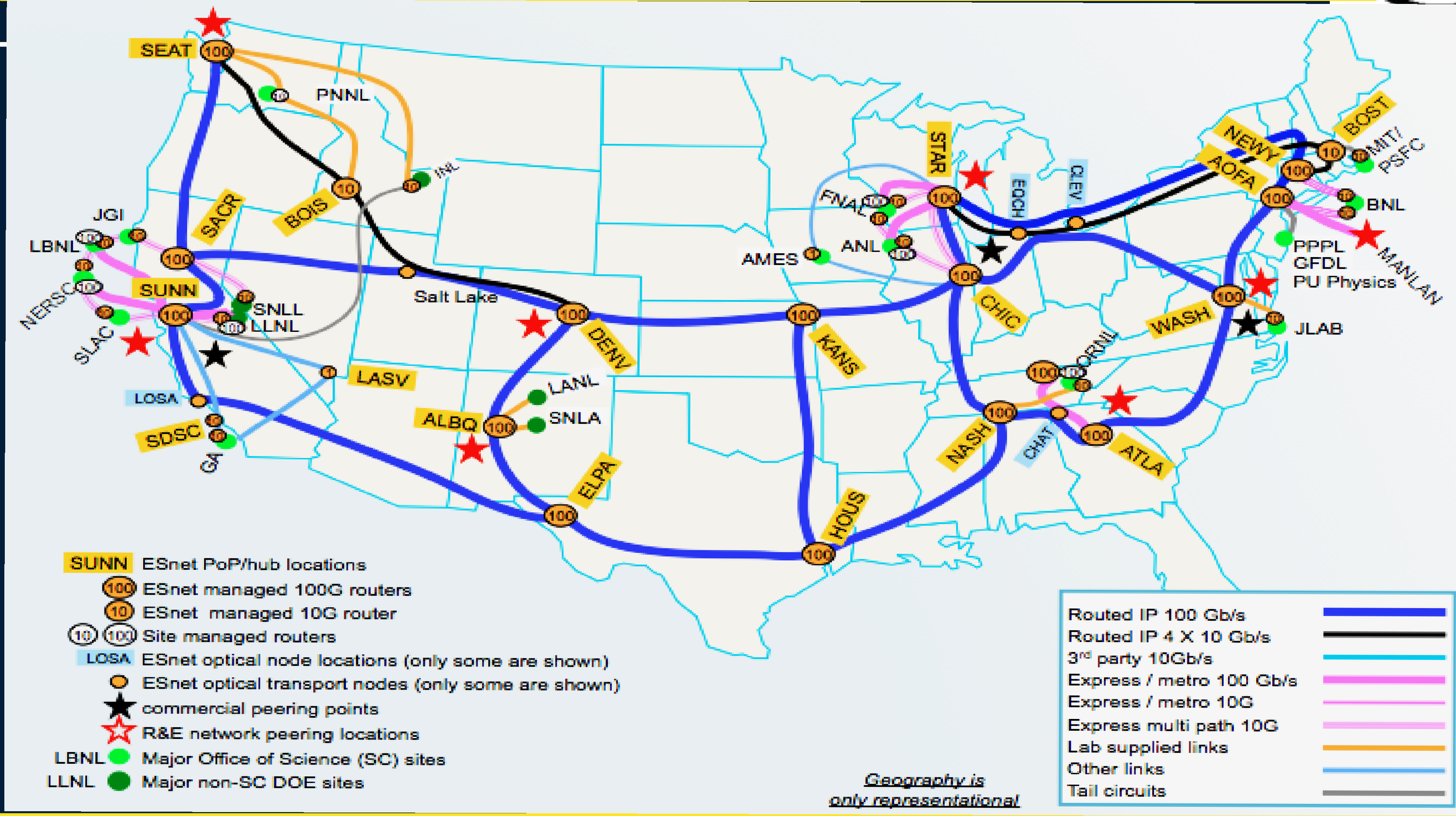
Sept. 2013

Numbers show How many 10G waves

37 NREN Partners

12kkm Dark Fiber Core in 16 Countries
Carrying 26 100G Links So Far

Energy Sciences Network: ESnet5
100G Backbone Completed in Nov. 2012



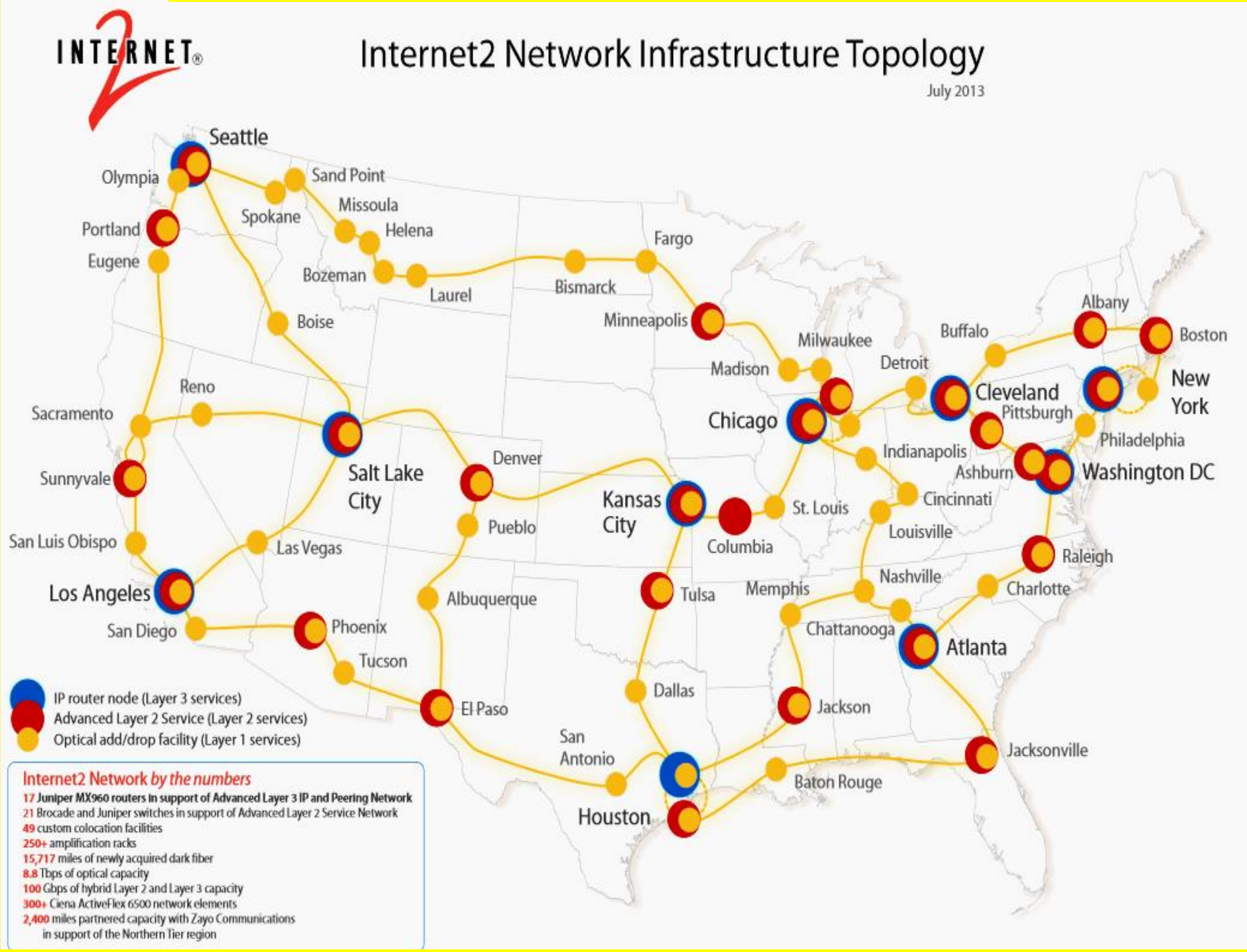
2 X 100G to BNL and 100G to Fermilab; 17 Hubs with N X 100G

Now Only 40G and 100G waves on the backbone

Metro Area Nets in NYC, Chicago, Sunnyvale, Atlanta

100G Dark Fiber Testbed; Share of 100G ANA-100 Transatlantic Link

Internet2 100G Network: Completed in 2012; Innovation Campus Program



Advanced Optical, Switched and Routed Services

Emphasis on 100G: 22 Connectors Plan 50+ 100GE Access Links by 2015

Software defined networking (SDN)

Led DYNES with Caltech

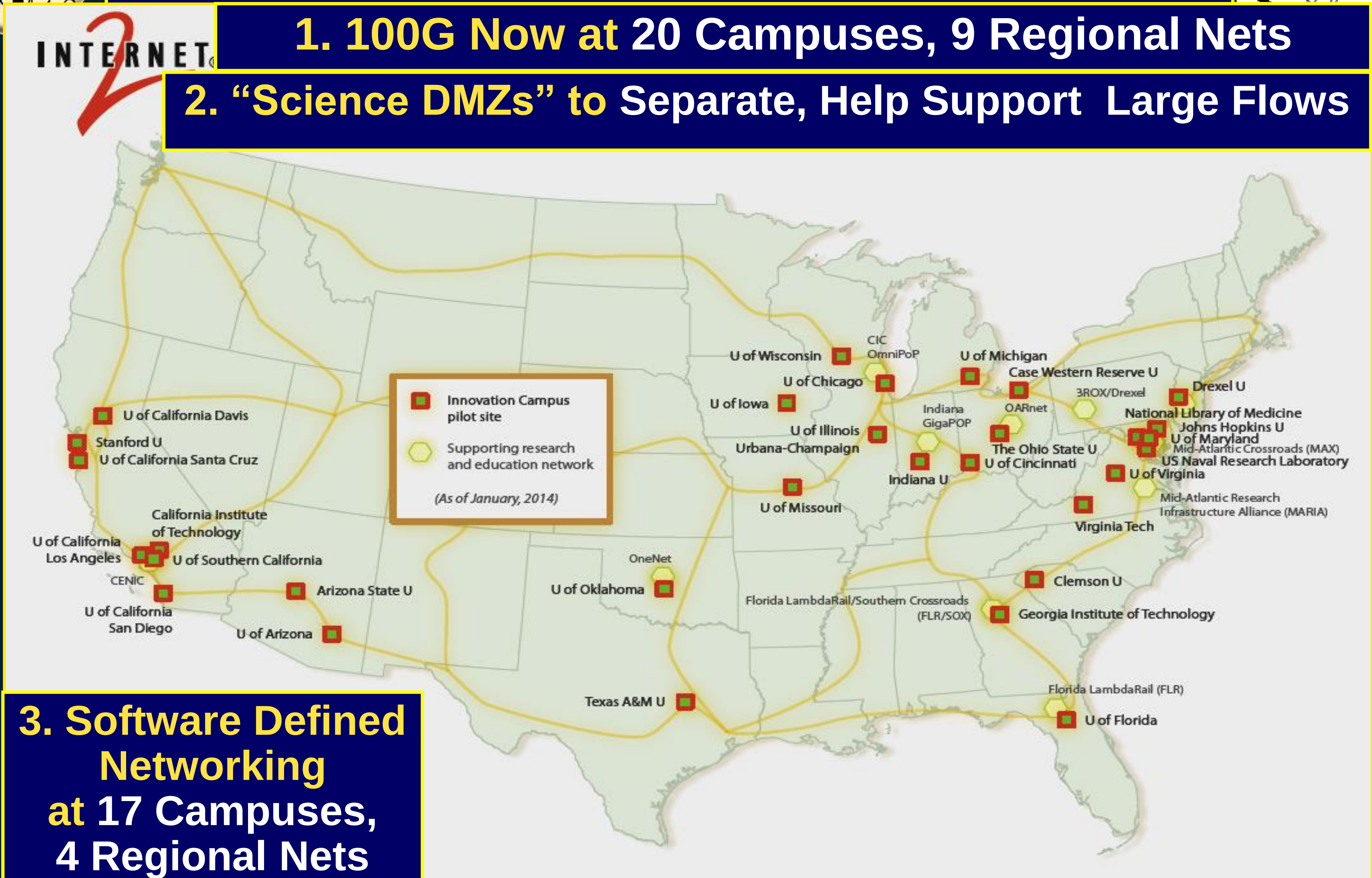
Heavily involved in LHCONE

18k Fiber Miles. Connects to 88 NRENs in Europe, Asia, Latin America, Africa, Middle East

Innovation Campus Pilot Program

1. 100G Now at 20 Campuses, 9 Regional Nets

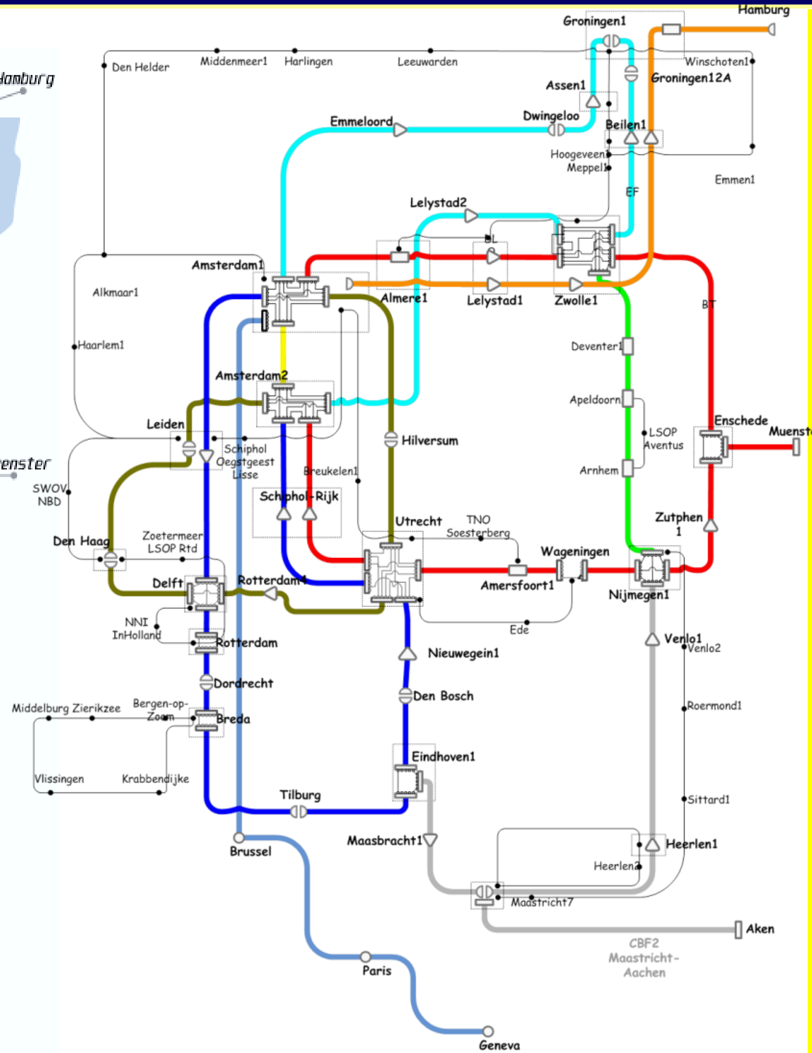
2. "Science DMZs" to Separate, Help Support Large Flows



3. Software Defined Networking at 17 Campuses, 4 Regional Nets

ICFA SCIC

SURFNet and NetherLight: 11000 Km Dark Fiber Flexible Photonic Infrastructure



5 Photonic Subnets

λ Switching at 10G, 100G

3 x 40GE + 1 x 100G links to CERN

158 Fixed or Dynamic Lightpaths

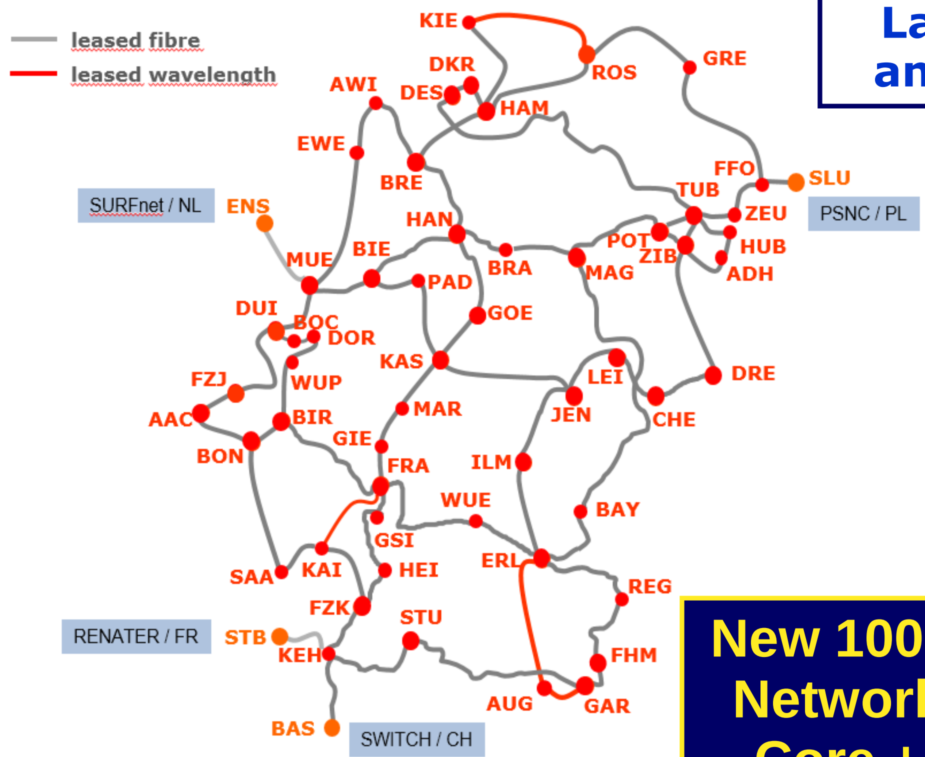
WLCG, EXPRES DEISA,CineGrid

Five Cross Border Fibers: to Belgium, on to CERN (1600km) to Germany: X-Win, On to NORDUnet

Bram Peeters

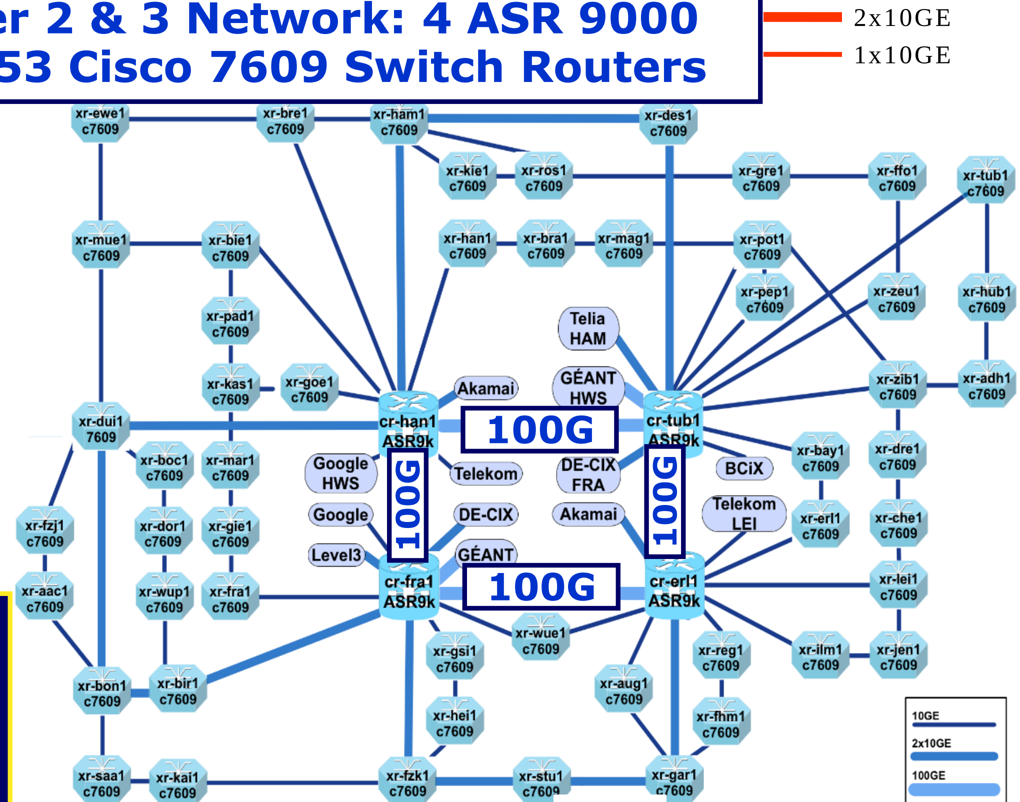
Germany DFN X-WiN: Dark Fiber Network

All New Optical Equipment Supporting 100G Waves in 2014



11,000 km Dark Fiber

Layer 2 & 3 Network: 4 ASR 9000 and 53 Cisco 7609 Switch Routers



New 100G Network Core + 2 X 100G to GEANT

N X 10G LHCOPN Links KIT (Tier1)-CERN. LHCONE 10G T1-T2 Links from KIT to DESY, Aachen, Wuppertal, GSI.

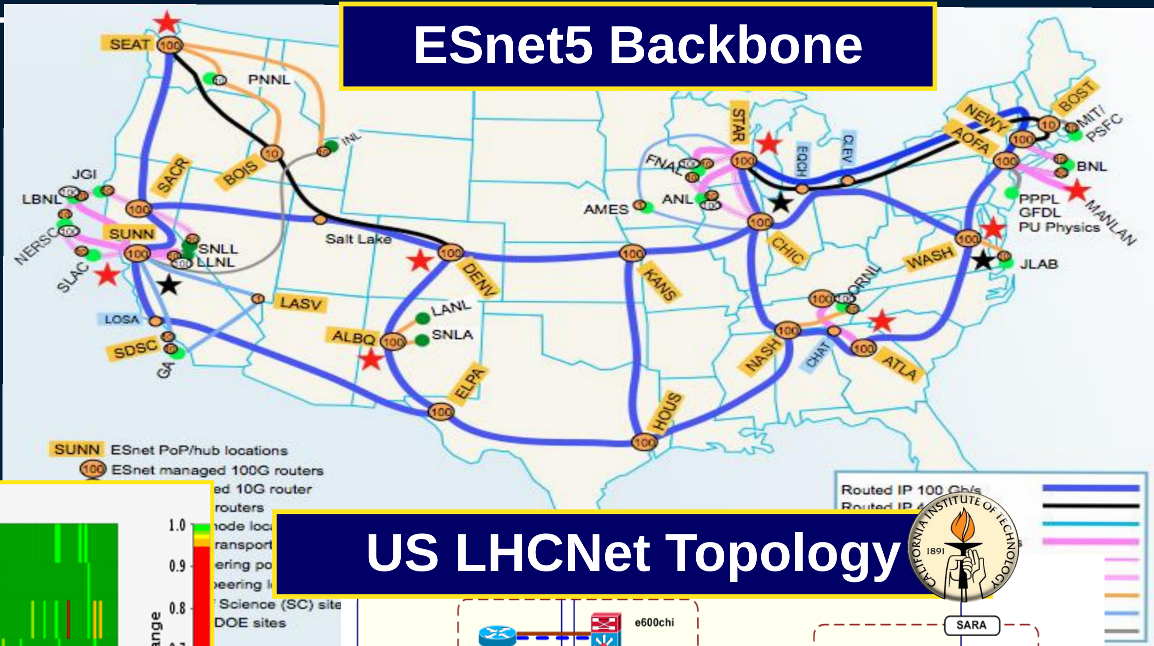
Cross Border Fibers to NL, FR, CH, PL

V. Guelzow

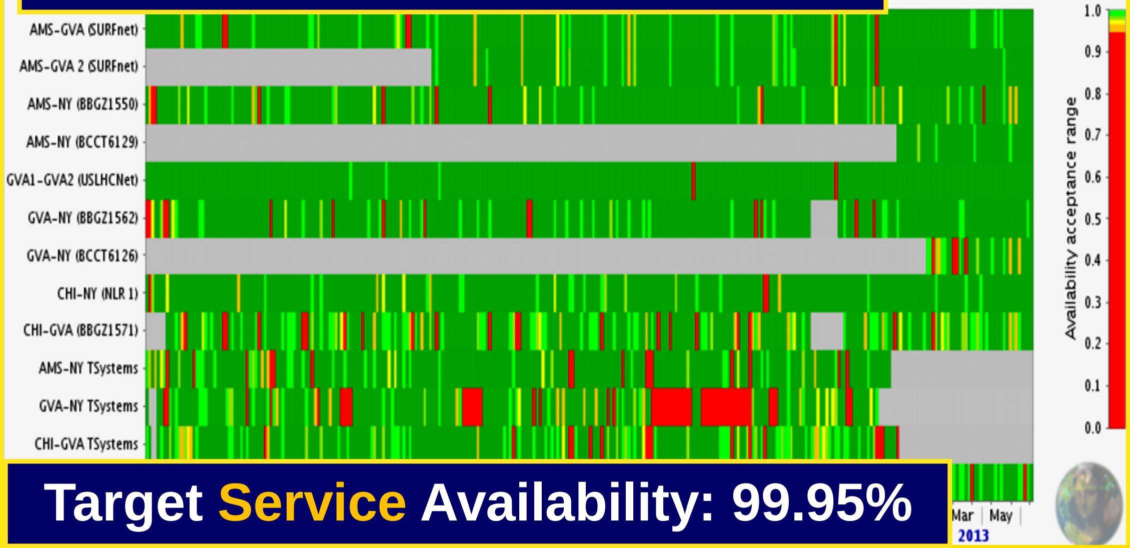
High Performance in Challenging Environments

- Intercontinental links are more complex than terrestrial ones
  - More fiber spans, more equipment; Multiple owners
- Hostile submarine environment
  - A week to Months to repair

ESnet5 Backbone

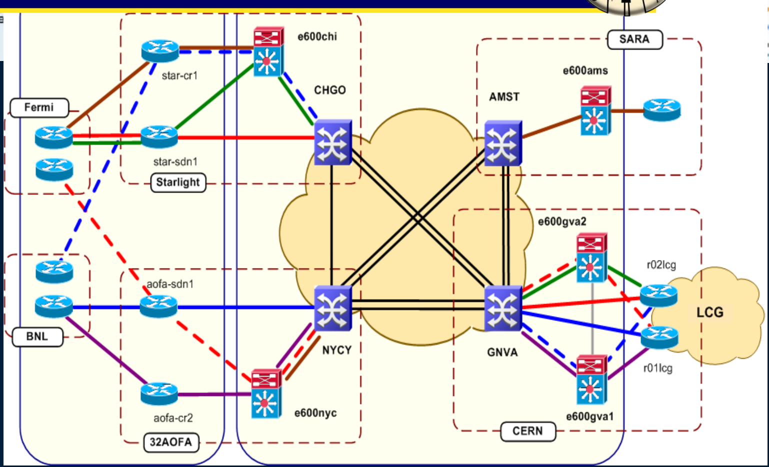


US LHCNet Link Availability



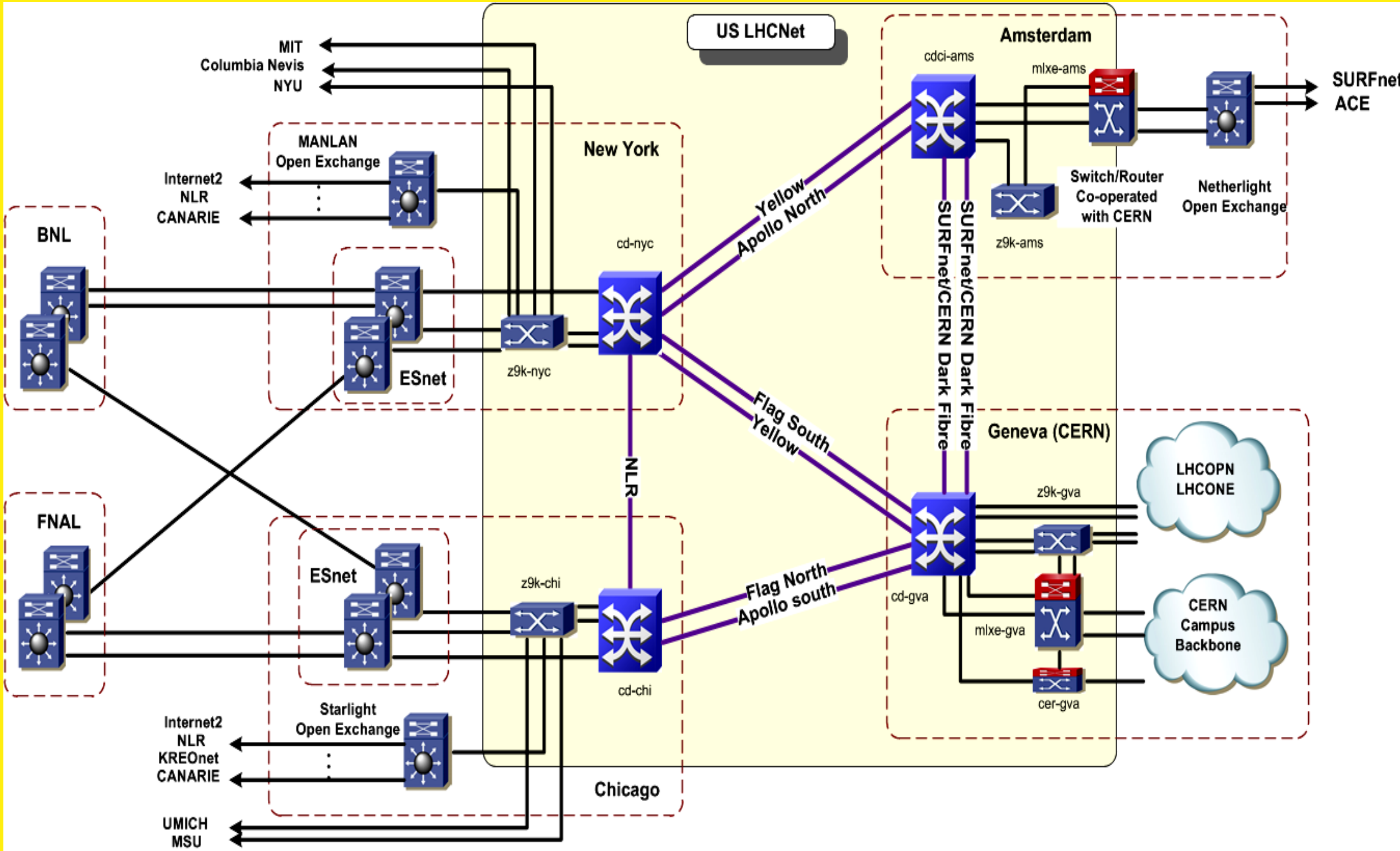
Target Service Availability: 99.95%

US LHCNet Topology



High-Availability Transoceanic solutions require multiple links with carefully planned path redundancy

US LHCNet in 2014

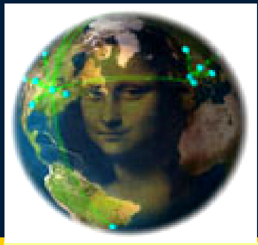


6 X 10GE Links Across the Atlantic +Continental Cross Links

Integrated Into the LHCONE VRF

Transition to ESnet "EEX" (100GE and 40GE Links) in 2015

***Dynamic* circuit-oriented Carrier services with BW guarantees, with robust seamless fallback at Layer 1: Hybrid optical network**



Monitoring the Worldwide LHC Grid

State of the Art Technologies Developed at Caltech

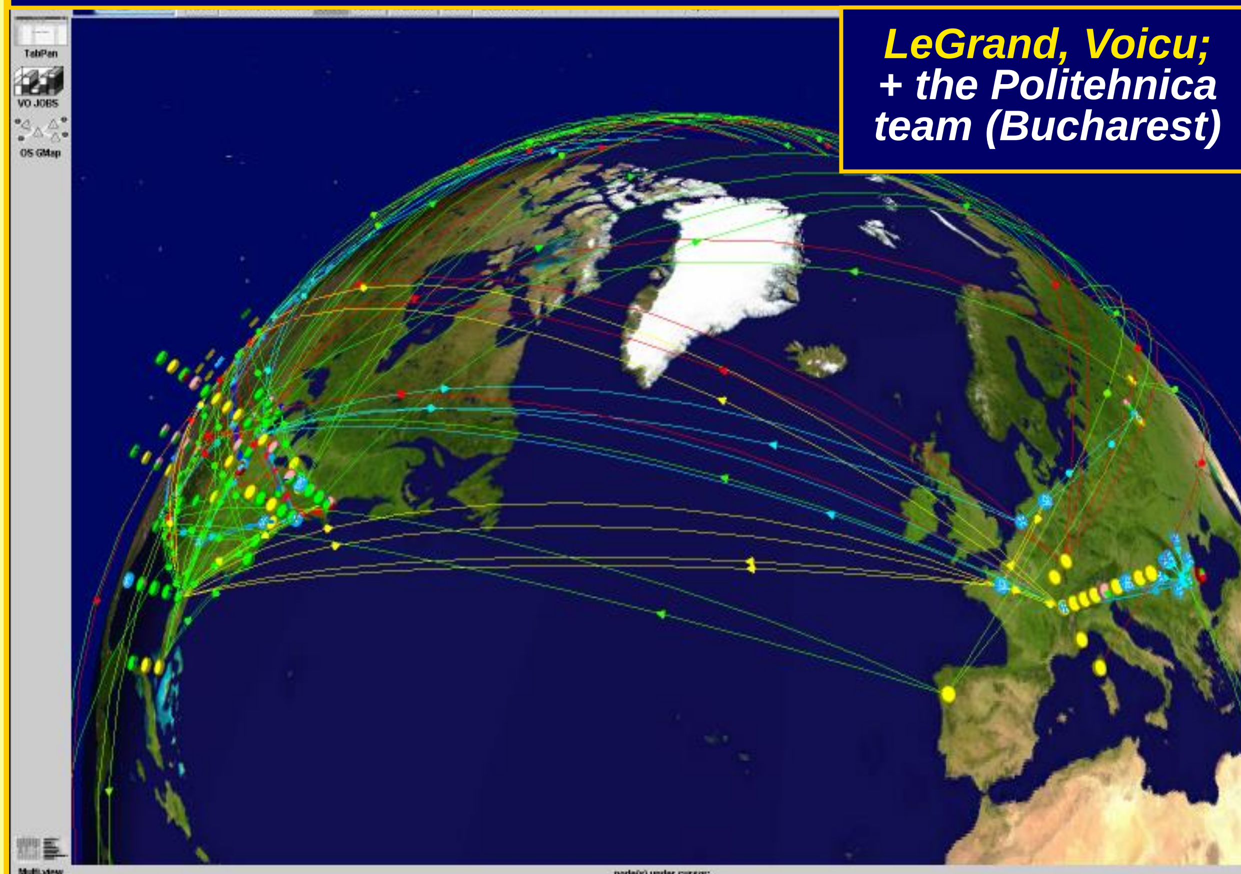
<u>MonALISA Today</u>

Running 24 X 7 at 370 Sites

- **Monitoring**
  - **60,000 computers**
  - **> 100 Links On Major R&E Networks, 14,000 end-to-end paths**
- **Using Intelligent Agents**
- **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**
- **Resilient: MTBF >7 Years**

MonALISA: Monitoring Agents in a Large Integrated Services Architecture

Unique Global Autonomous Realtime System



LeGrand, Voicu; + the Politehnica team (Bucharest)

Caltech Network Team
Synergistic Working Methodology

Production Network

Develop and build next generation networks

High performance
High bandwidth
Reliable network

Pre-Production

TA Testbed: to *N X 10G + 100G*

Lightpath technologies with ESnet and Internet2: *DYNES, OESS/PSS*, OSCARS, DRAC, AutoBAHN

New transport protocols; Software Defined Networking

Ultralight / λStation / Terapaths /PlaNetS/OliMPS/ANSE/Cisco; Vendor Partnerships; SC02-13

HEP & DOE Roadmaps

Testbed for Network Services Development

Networks + Grids for HEP & Data Intensive Research

LHC + Other Experiments: LHCOPN; *LHCONE*

Grid Projects: e.g. PPDG, GriPhyN, iVDGL DISUN; OSG, WLCG

ANSE

R&D efforts tailored for the HEP community and other data intensive science, with direct feedback into high performance production networks.

US CMS Tier2 WAN upgrade plans

Site	Upgrade plan	LHCONE
Caltech	100 Gbit by March 2014	Yes
Florida	100 Gbit available	Planning to
MIT		
Nebraska	100 Gbit in March 2014	Yes
Purdue	100 Gbit available	No plan
UCSD	100 Gbit in August 2014	"Depends"
Wisconsin	40 Gbit by Summer 2014	No plan

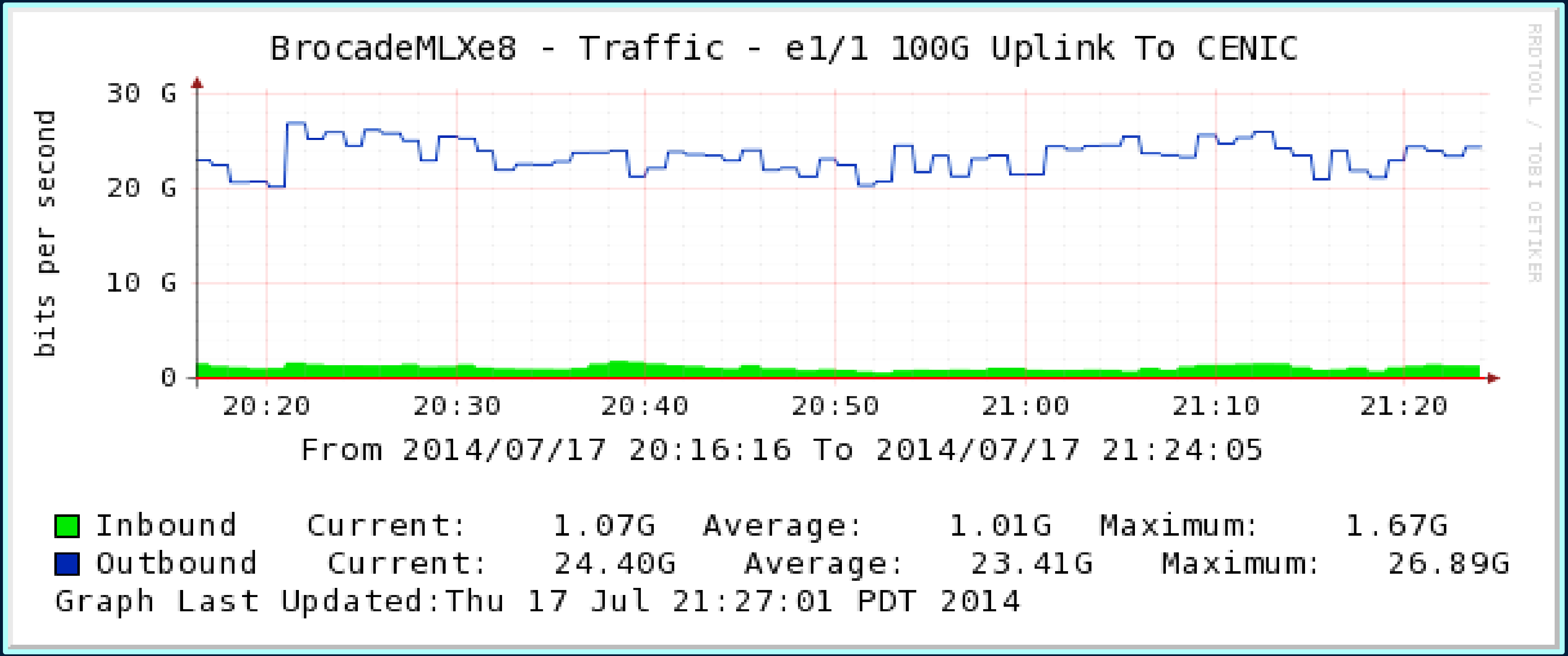
- Note: 1000 T2 batch slots can analyze 2.4 Gbit/s of CMS data
- Needless to say, given the effort and expense needed to upgrade the campus network infrastructure, we want to make the best use of it for scientific productivity

Most US Tier2 Sites at 100G in 2014

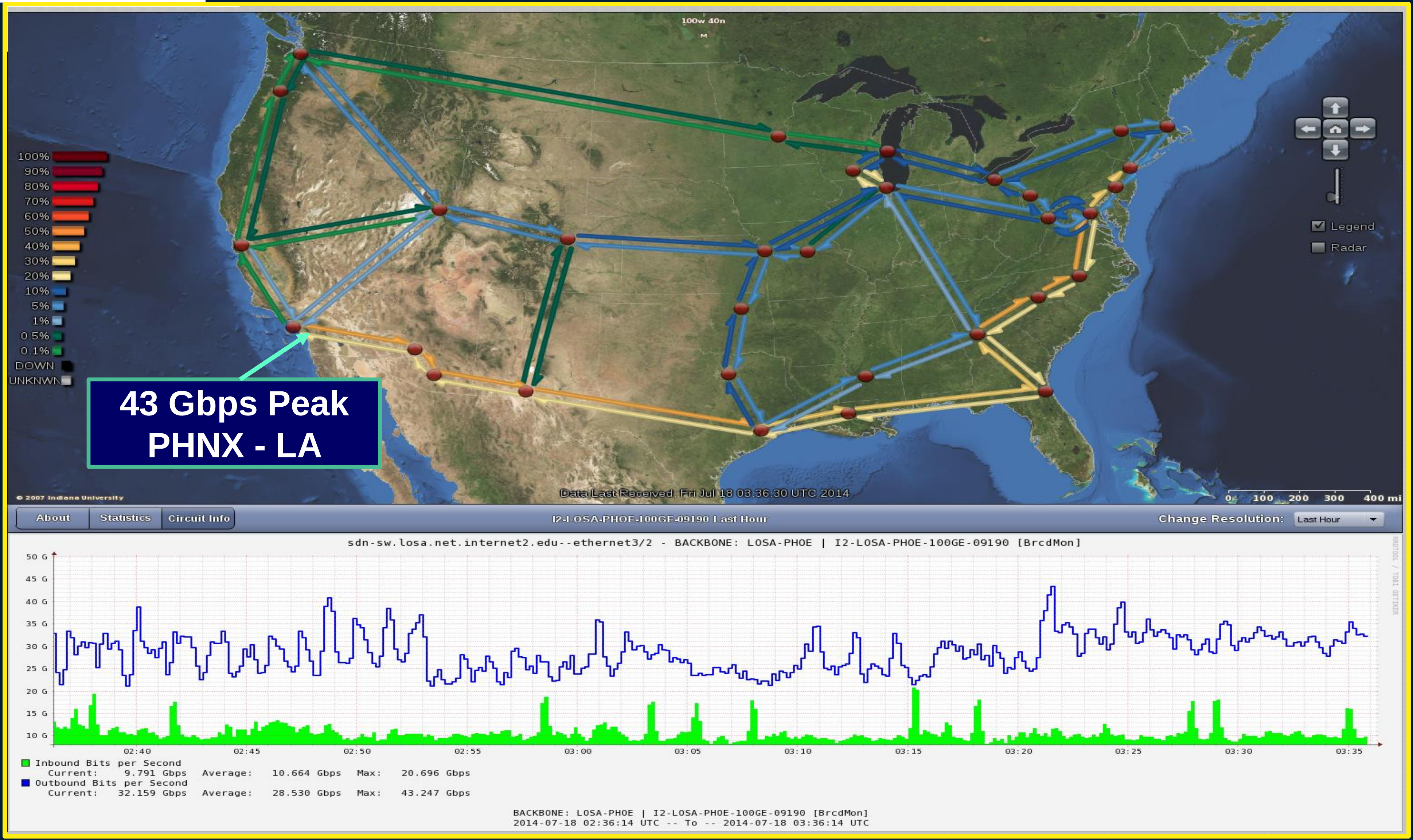
ANA-100 Link in Service July 16

Transfer Rates: Caltech Tier2 to Europe July 17

- 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



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



Just Ahead: LHC Run2

A Time of Opportunity; a Time of Challenge

"If I had asked people what they wanted, they would have said faster horses…"

–Henry Ford

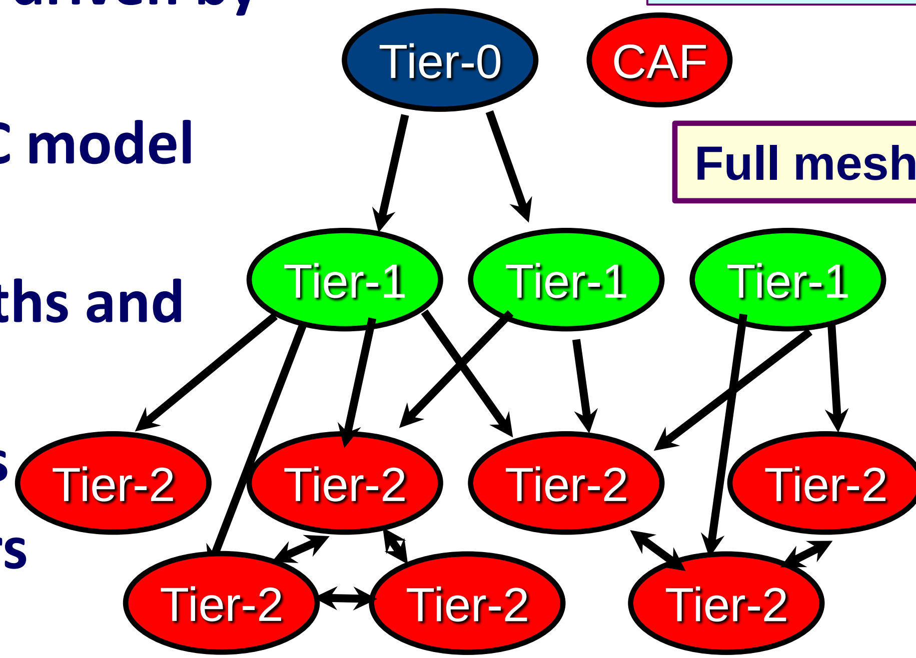
The LHC Computing Models

Continue to Evolve Rapidly

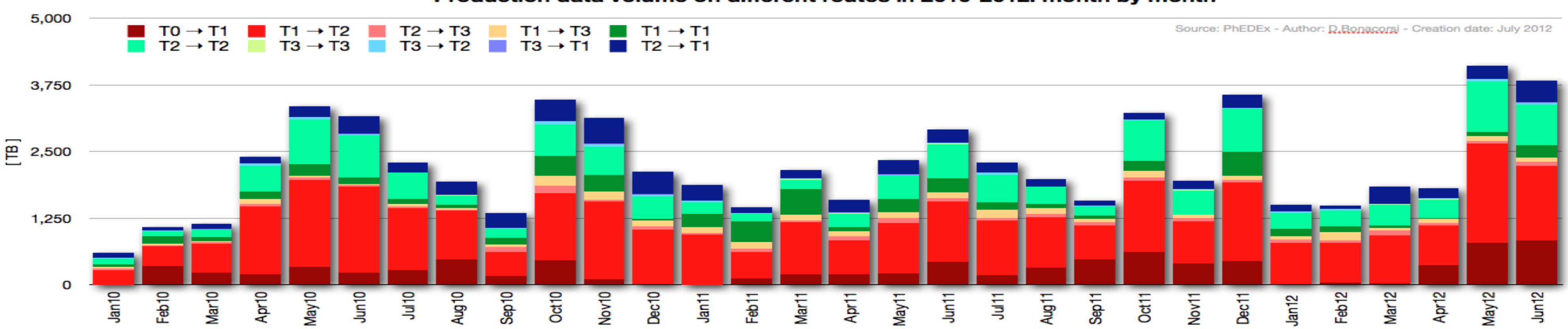
Data Distribution Model in Run1

Maria Girone

- In Run1, CMS network needs were driven by the data distribution model
  - an evolution from the MONARC model but still structured
- Network went through defined paths and large volumes of data were moved
  - Dominated by analysis requests
  - and by Tier-1s to Tier-2 transfers

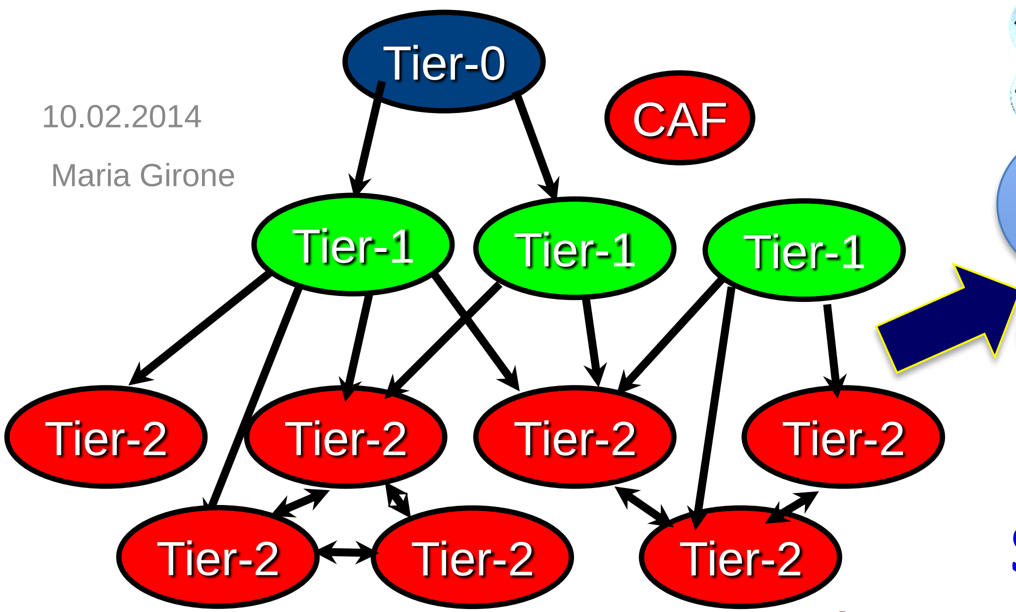


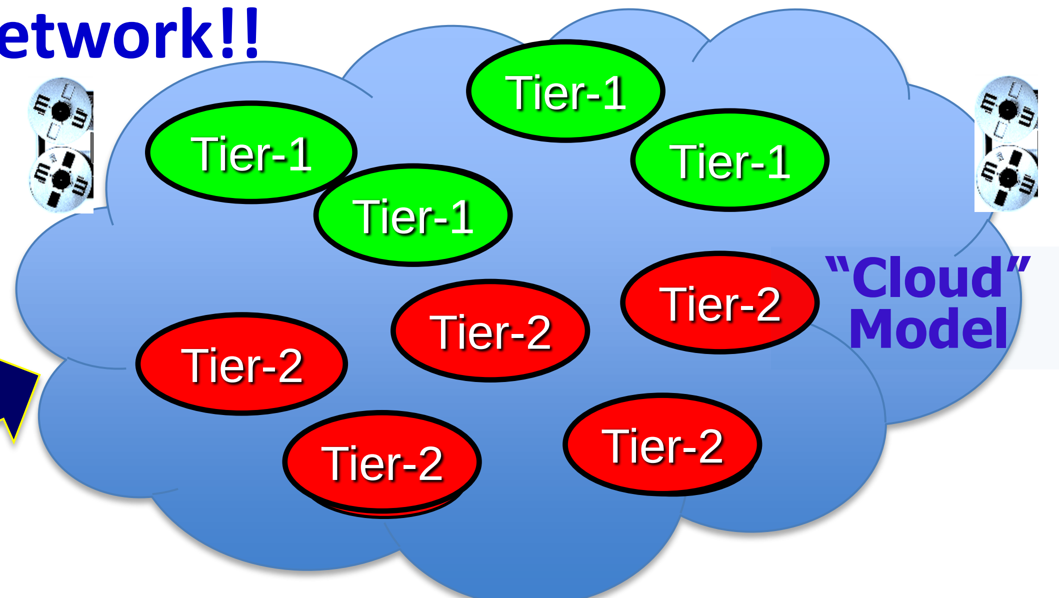
Production data volume on different routes in 2010-2012: month by month



CMS: Location Independent Access: Blurring the Boundaries Among Sites

- Once the archival functions are separated from the Tier-1 sites, the **functional difference between the Tier-1 and Tier-2 sites becomes small**
  - **Connections and functions of sites are defined by their capability, including the network!!**



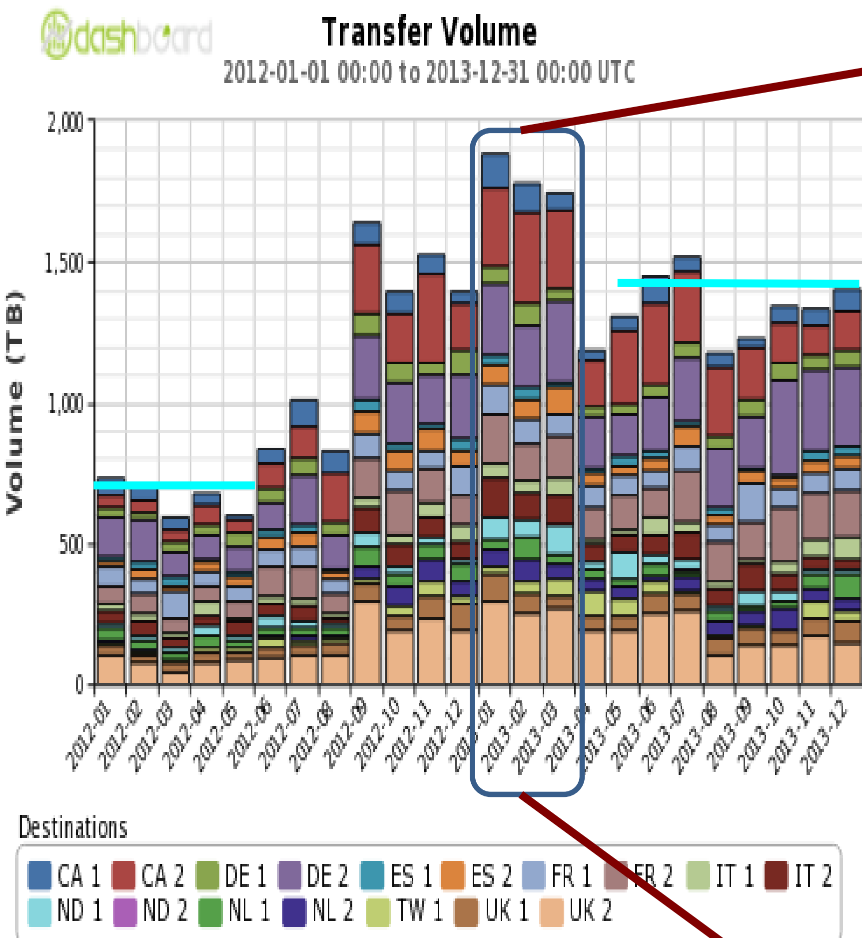


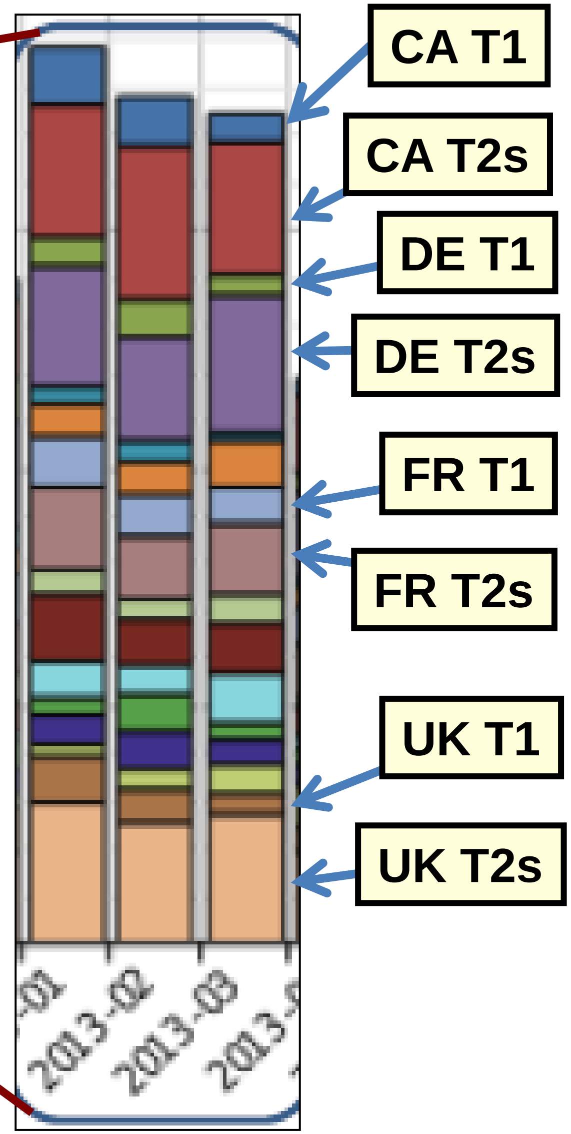
Scale tests ongoing:

Goal: 20% of data across wide area;

200k jobs/day, 60k files/day, O(100TB)/day

ATLAS: T1s vs. T2s from BNL (2013 Winter Conference Preparations)





T2s in several regions are getting ~an order of magnitude more data from BNL than the associated T1s

2H 2013 Volume was ~twice that of 1H 2012, even without data taking.

Exponential growth in data transfers continues, driven by Tier2 data usage.

Expect new peaks by and during LHC Run 2

LHCONE: Responding to Changes in the LHC Computing Models

Qualitative Changes in the Network Landscape During Run2

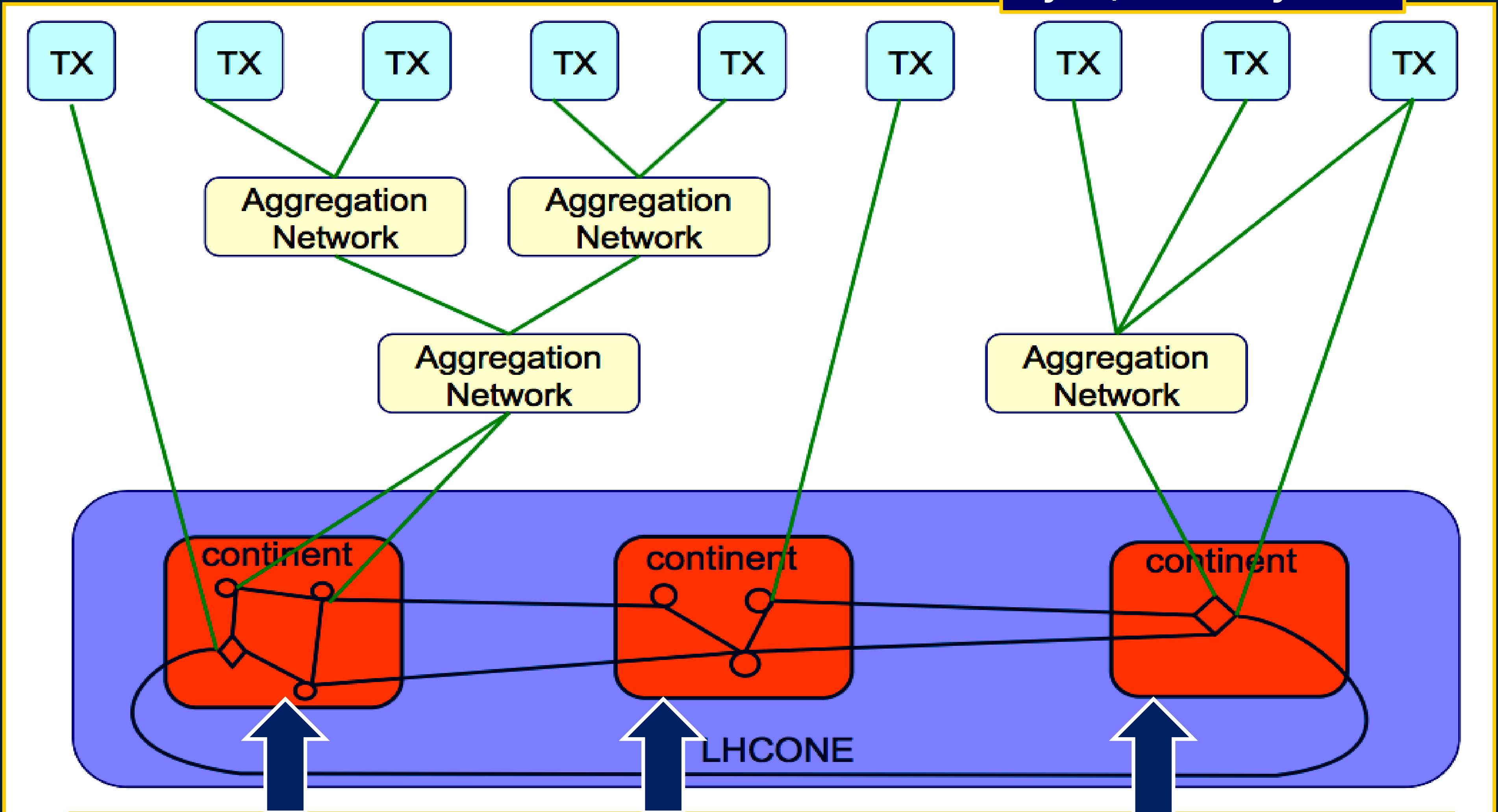
LHCONE: A Global Ensemble of Interconnected Open Exchange Points

- In a nutshell, LHCONE was born out of a 2010 transatlantic workshop at CERN, to address two main issues:
  - To ensure that the services to the science community maintain their quality and reliability; *With a Focus on Tier2/3 operations*
  - To protect existing R&E infrastructures against potential “threats” of very large data flows
- Concepts originated by Caltech
- LHCONE is expected to
  - Provide some guarantees of performance
    - Large data flows sent across managed bandwidth: to provide better determinism than shared IP networks
    - Segregate these from competing traffic flows
    - Manage capacity as # sites x Max flow/site x # Flows increases
  - Provide ways to better utilize network resources
    - Use all available resources, especially transatlantic
    - Provide Traffic Engineering and flow management capability
  - Leverage investments being made in advanced networking

LHCONE Initial Architecture

Basic Idea at 30'000 ft

LHCOPN Meeting
Lyon, February 2011

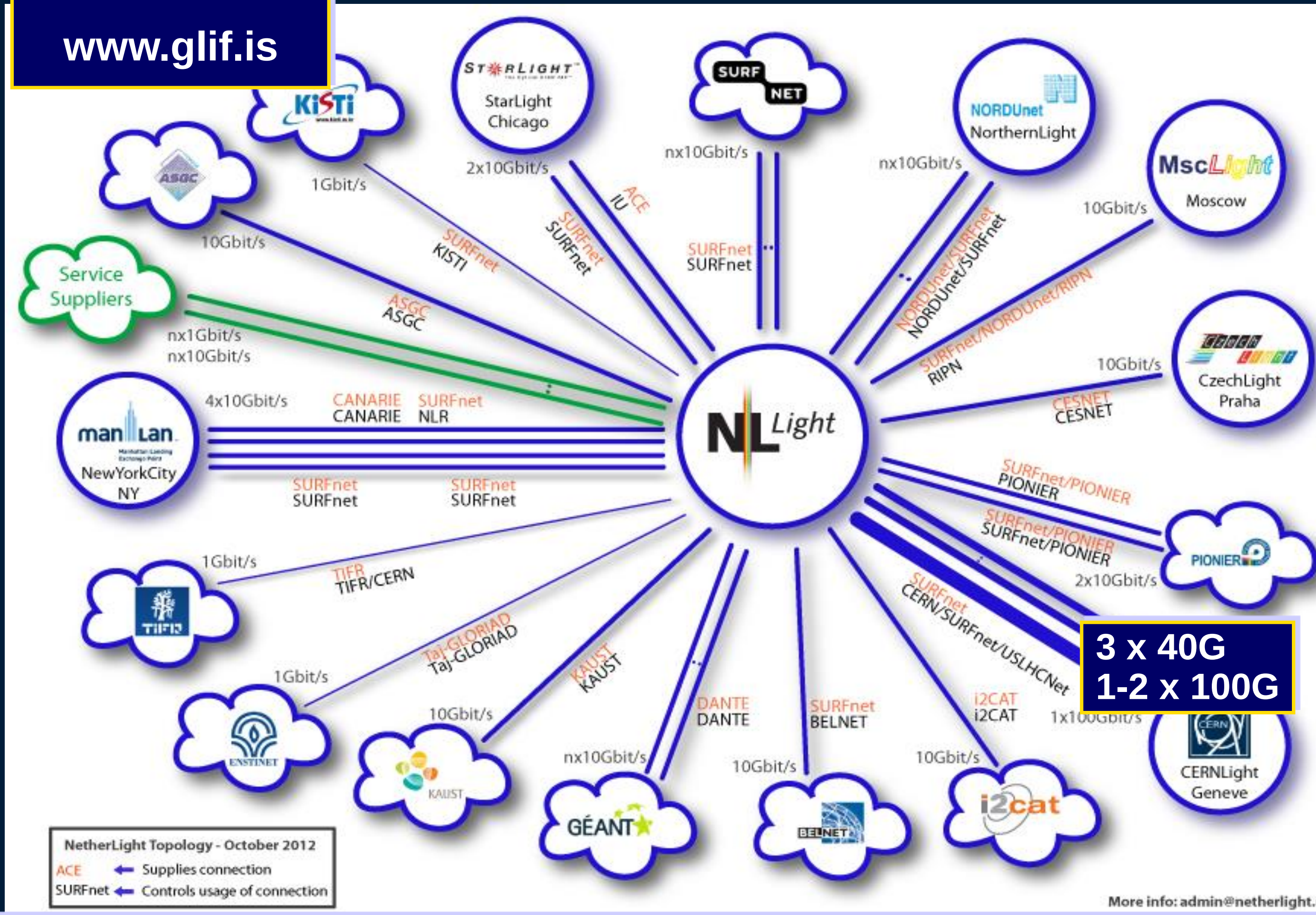


Sets of Open Exchange Points

Open Exchange Points: NetherLight Example

1-2 X 100G, 3 x 40G, 30+ 10G Lambdas, Use of Dark Fiber

www.glif.is



3 x 40G
1-2 x 100G

Inspired Other Open Lightpath **Exchanges**
- Daejeon (Kr)
- Hong Kong (Cn)
- Tokyo (Jp)
- Praha (Cz)
- Seattle
- Chicago
- Miami
- New York

2015-18: Dynamic Lightpaths + IP Services Above 10G

Convergence of Many Partners on Common Lightpath Concepts

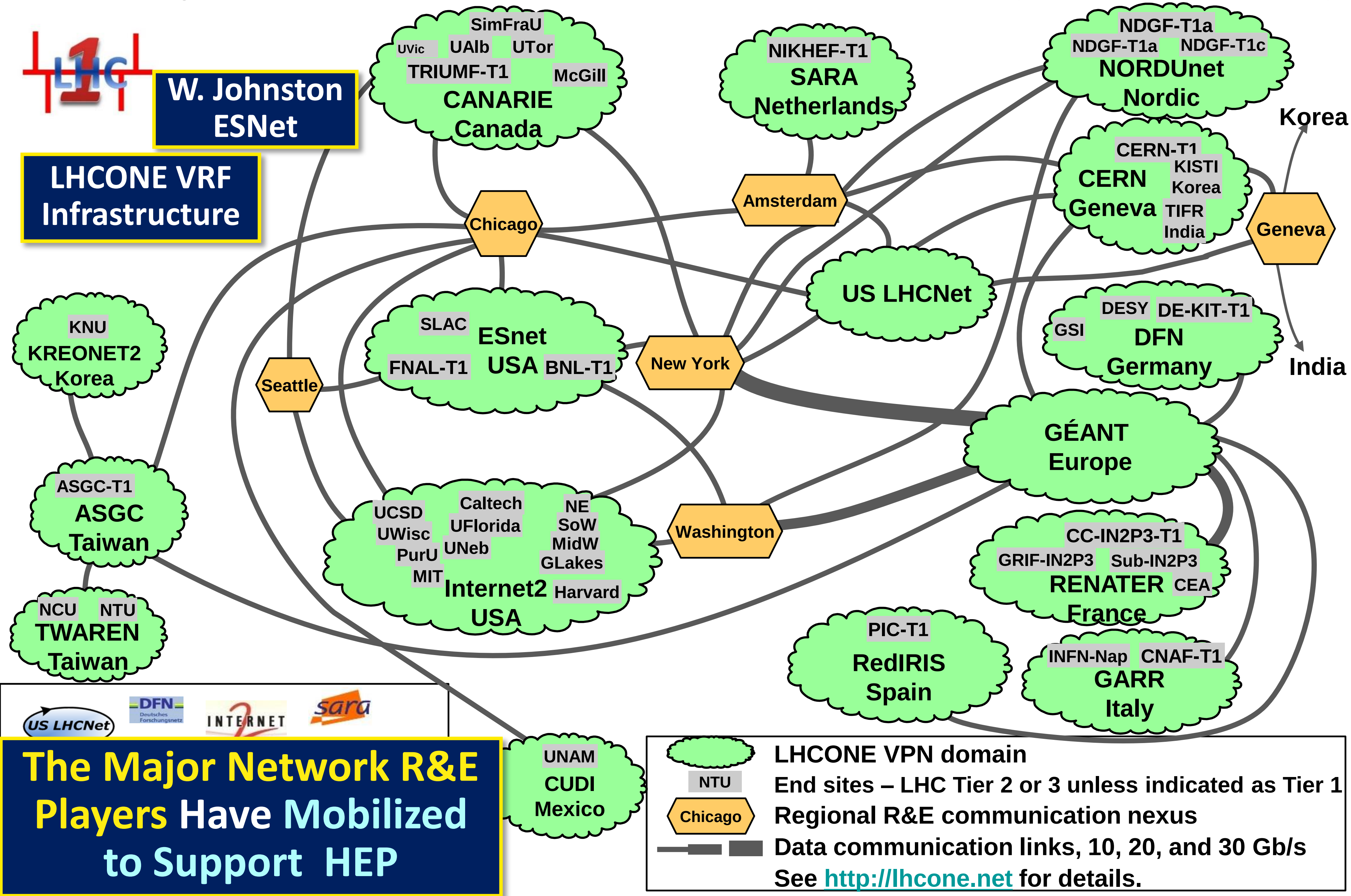
Internet2, ESnet, GEANT, USLHCNet; nl, cz, ru, be, pl, es, tw, kr, hk, in, nordic

LHCONE Activities

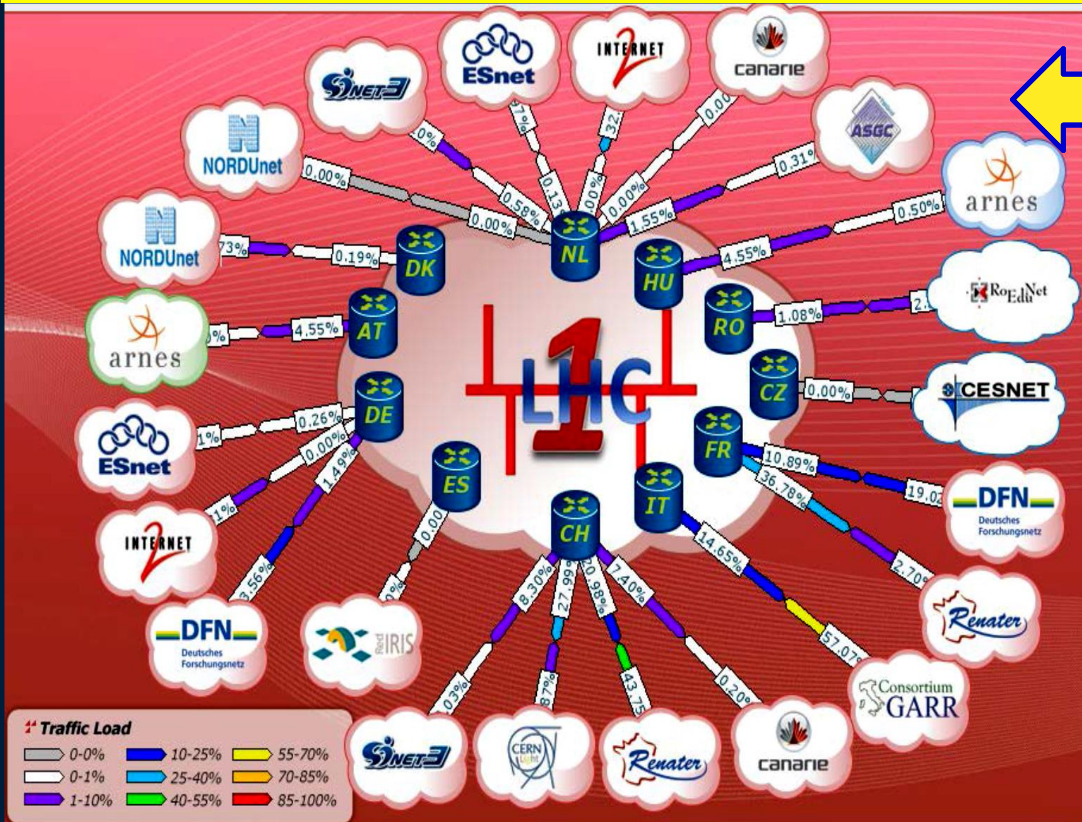
- **Virtual Routing and Forwarding (VRF)-based IP service: a "quick-fix" to provide multipoint LHCONE connectivity, with logical separation of LHC from general purpose R&E traffic**
  - **Successful first phase: in Europe and Canada**
  - **Issue: Policy & technique of restricting to LHC-related clusters**
- **Point to point dynamic virtual circuits service: multi-domain**
  - **Using OSCARS and other existing technologies now**
  - **Migrate to NSI, an emerging worldwide standard**
- **Software Defined Networking: Wide agreement that this is the probable technology of choice for LHCONE in the long-term, with Openflow the leading candidate protocol.**
  - **Promising early results. It needs more development and investigation, to fulfill its (considerable) promise**

Overarching Goals: *Benefit from improved capacity where possible. Investigate the impact of the LHCONE VRF, dynamic circuits (and eventually OpenFlow) on LHC data analysis workflow*

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

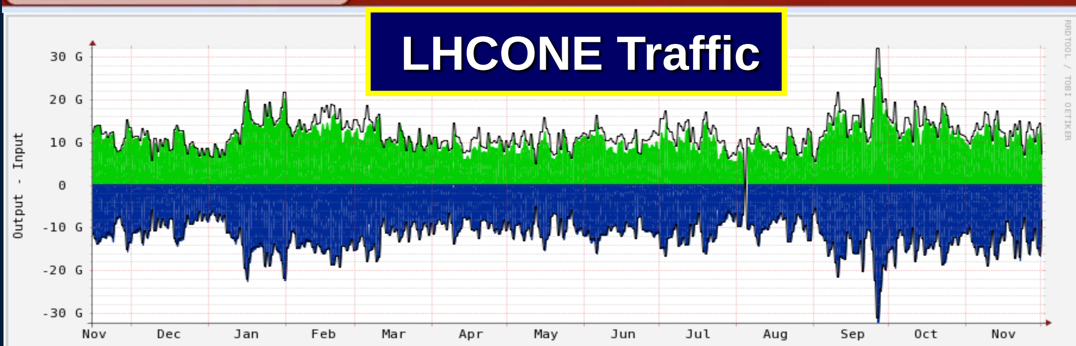


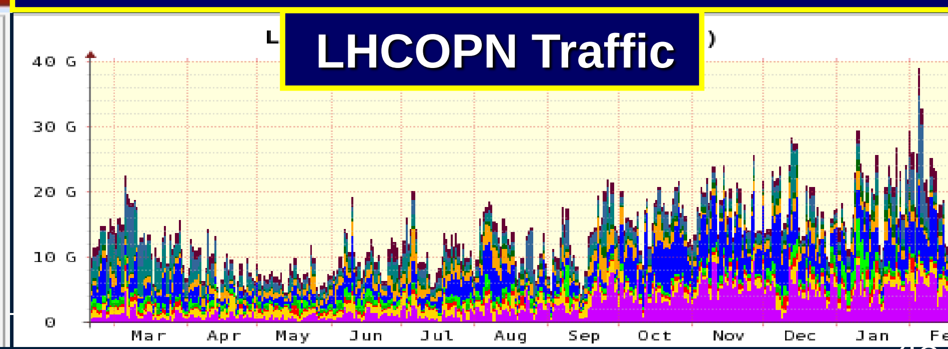
LHCONE Phase1: A "Virtual Routing and Forwarding Fabric" Connecting 8 Tier1s, 40 Tier2s



LHCONE View from Europe

- An important complement to the LHCOPN. Focus on Tier2 and Tier3 operations; Restrict Access to LHC Sites
- Traffic: Steady use above 10 Gbps; peaks of 30 Gbps observed in 2013
- Versus LHCOPN: to 50 Gbps





Canadian Tier1 and Tier2 Sites
Happy So Far with LHCONE

Ian Gable

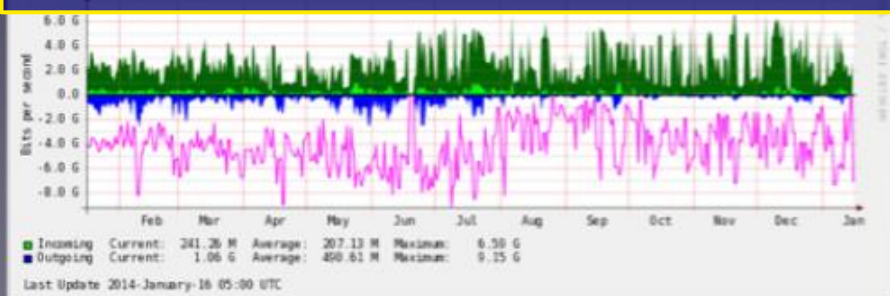
- All sites feel well served by the R&E Networking community.
- In 2012 we moved from using point-to-point circuits to connect TRIUMF - Canadian Tier2s to using the LHCONE within Canada.
  - immediately boosted path utilization and increased performance
  - prevented East coast T2s from communicating with each other via TRIUMF 4000 km away.
  - 2013 CANARIE provisioned a second, dedicated 10G circuit for LHCONE in Canada
  - Additional 10G to TRIUMF LHCONE being added now.

Canada: ATLAS Tier1s and Tier2s
and LHCONE

ATLAS Tier 1 at TRIUMF with 10% of ATLAS Data 8400 km from CERN in a straight line, 175 ms RTT

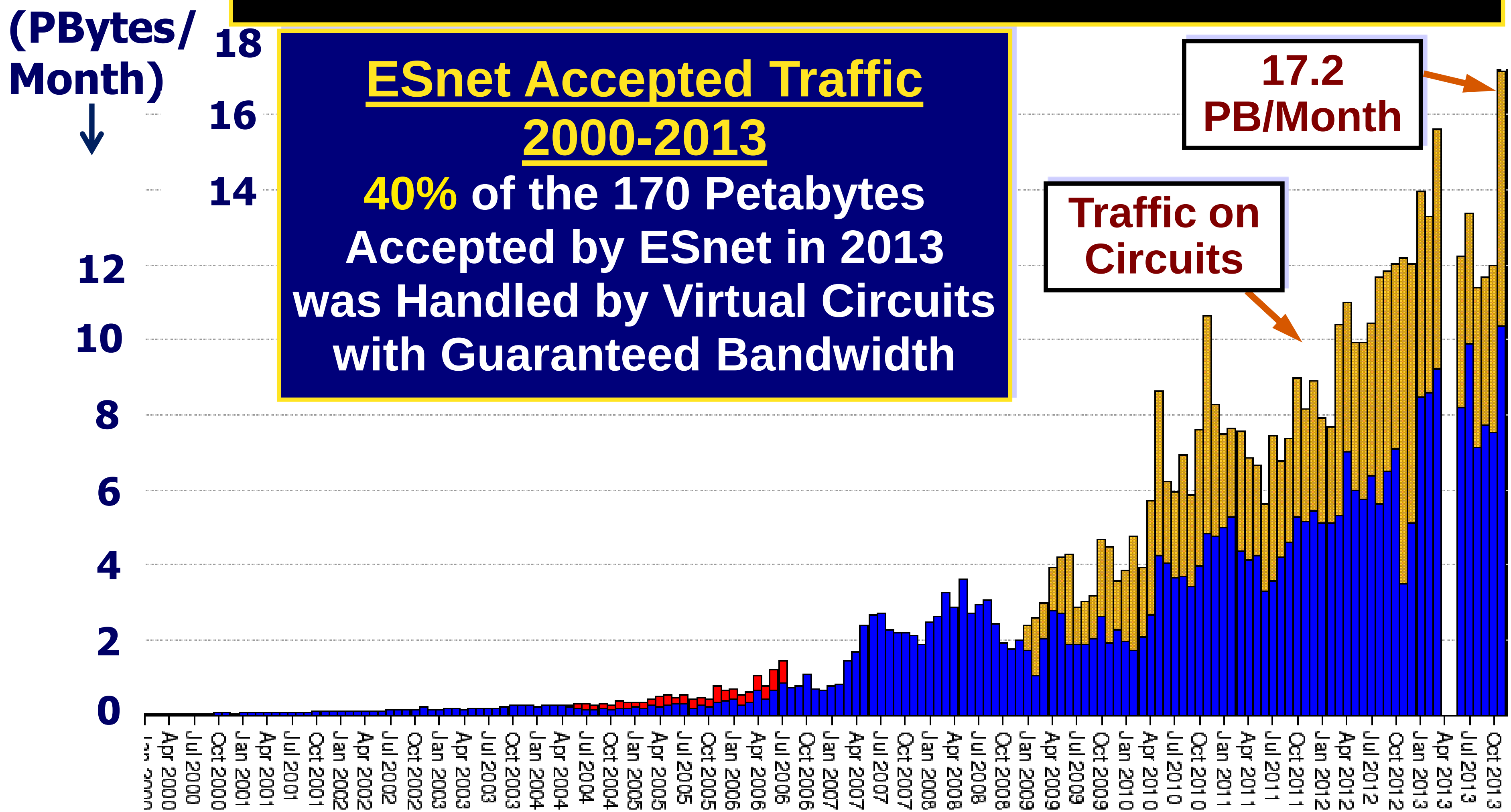
4 ATLAS Tier 2s at University of Victoria, Simon Fraser University, University of Toronto, and McGill University

UVic Tier2 Traffic



Ian Gable

Dynamic Circuits with Bandwidth Guarantees
Next Phase of LHCONE



Large Scale Flows are Handled by Circuits
Using "OSCARs" Software by ESnet and collaborators

What Networks Need to Do

W. Johnston, ESnet Manager (2008)
On Circuit-Oriented Network Services

- For this essential approach to be successful in the long-term it must be routinely accessible to discipline scientists - without the continuous attention of computing and networking experts

- In order to
  - facilitate operation of multi-domain distributed systems
  - accommodate the projected growth in the use of the network
  - facilitate the changes in the types of traffic

➡ the architecture and services of the network must change

- **The general requirements for the new architecture are that it provide:**
  - **1) Support the high bandwidth data flows of large-scale science including scalable, reliable, and very high-speed network connectivity to end sites**
  - ✸ **2) Dynamically provision virtual circuits with guaranteed quality of service (e.g. for dedicated bandwidth and for traffic isolation)**
  - ✸ **3) provide users and applications with meaningful monitoring end-to-end (across multiple domains)**

Traffic Isolation; Security; Deadline Scheduling; High Utilization; Fairness

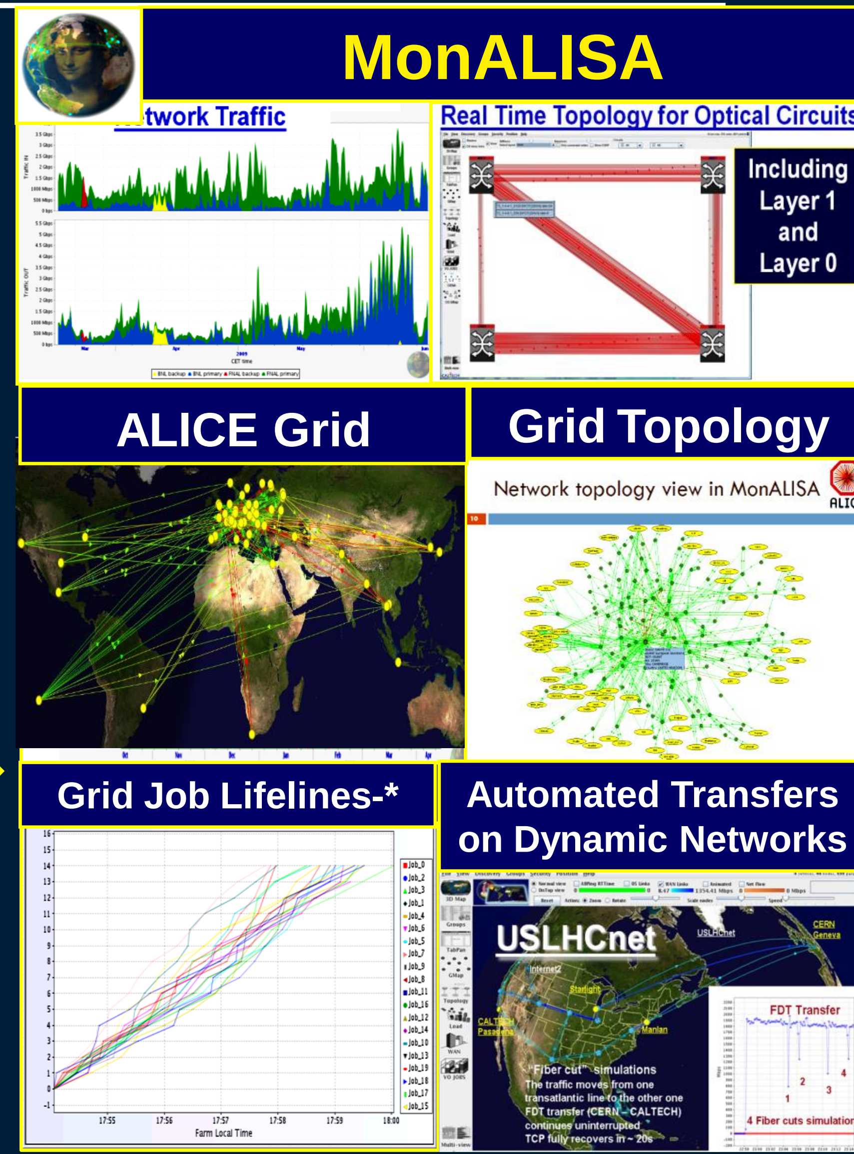
Key Issue and Approach to a Solution:

Next Generation System for Data Intensive Research

- Present Solutions will not scale
- We need: an agile architecture exploiting globally distributed grid, cloud, specialized (e.g. GPU) & opportunistic computing resources
- A Services System that moves the data flexibly and dynamically, and behaves coherently
- Examples do exist, with smaller but still very large scope

MonALISA →

- A pervasive, agile autonomous agent architecture that deals with complexity
- By talented system developers with a deep appreciation of networks

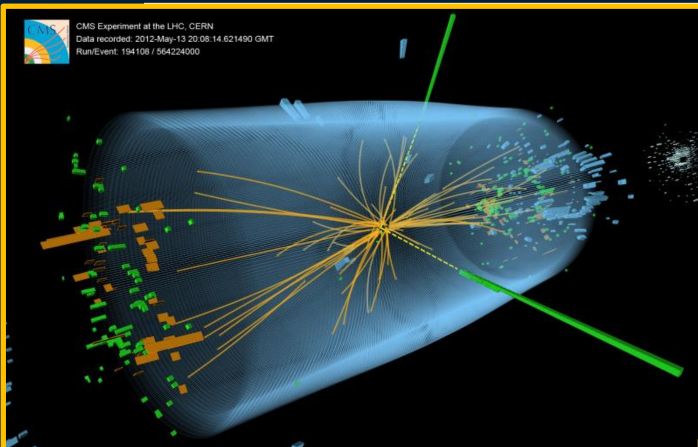


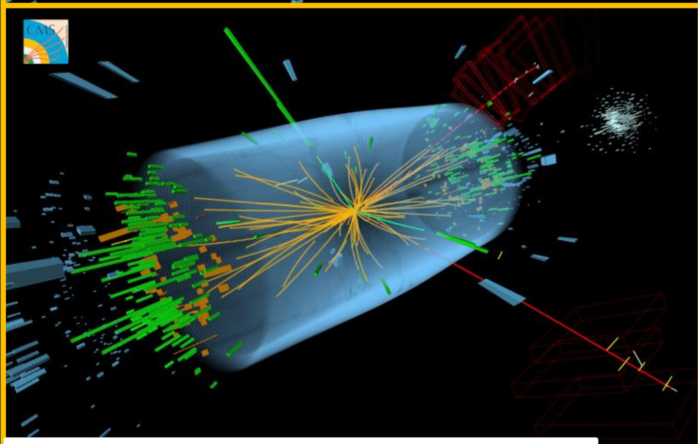
Networks for HEP
Journey to Discovery

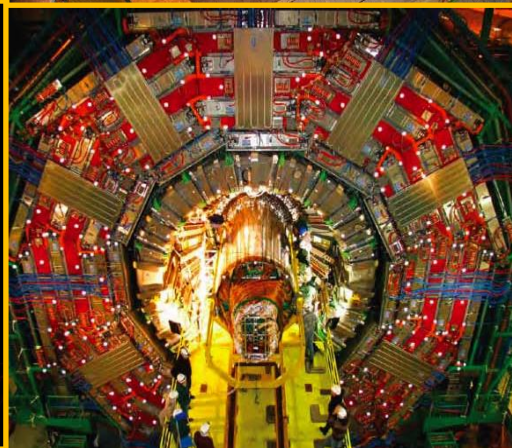
- **Run 1 brought us a centennial discovery: the Higgs Boson**
- **Run 2 will bring us (at least) greater knowledge, and perhaps greater discoveries: Physics beyond the Standard Model.**
- ***Advanced networks will continue to be a key* to the discoveries in HEP and other fields of data intensive science and engineering**
- **Technology evolution *might* fulfill the short term needs**
- **A new paradigm of global circuit based networks 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 [building on LHCONE, DYNES, ANSE, OliMPS]***
- **Worldwide deployment of such systems by 2023 will be:**
  - **Essential for the High Luminosity LHC HL-LHC**
  - **A game-changer that could shape both research and daily life**

Networking for HEP in the LHC Era:

Building on the Caltech Team's Experience and Global-Scale Developments for Data Intensive Science







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

Gateway to a New Era

50 Vertices, 14 Jets, 2 TeV

Harvey B Newman, Caltech

International School of Physics "Enrico Fermi": Lecture 2

Moving Forward: Innovation Examples

DYNES: Dynamic Network System

ANSE: Advanced Network Services for Experiments

OliMPS + Cisco Research: Software Defined Networking with OpenFlow, Open Daylight

CHOPIN: State of the Art US and TA Networks at Caltech

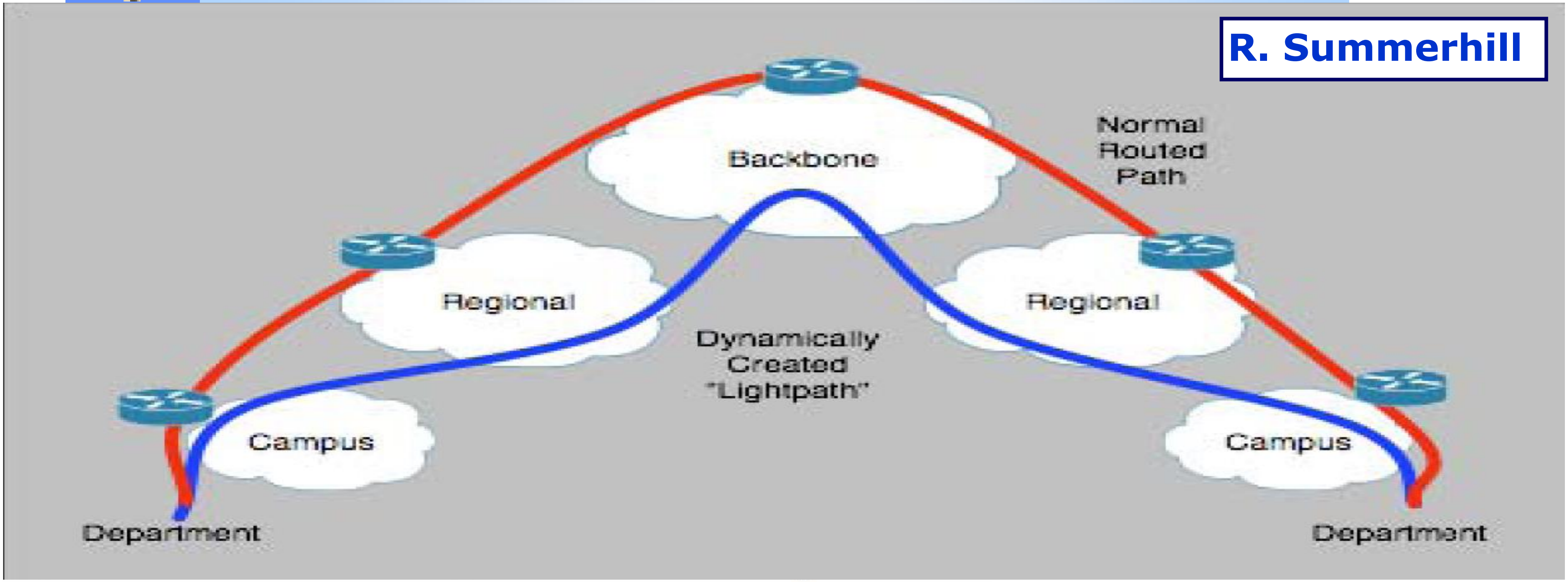
Networking for HEP

Ongoing Innovations by the Caltech Team

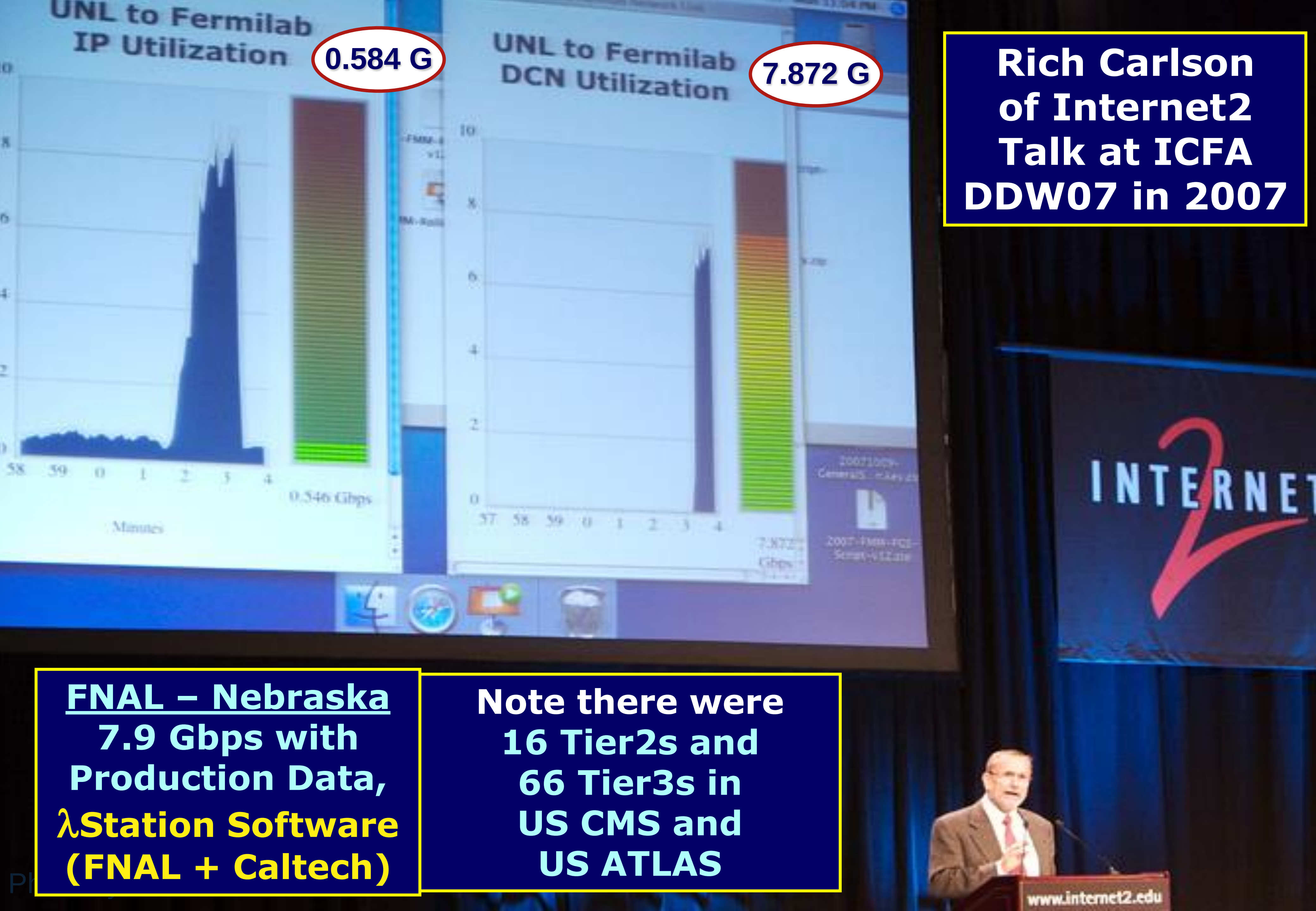
- We are active in several developmental lines important to the Computing Model evolution of large-scale computing for HEP. To name a few:
  - Software Defined Networking (SDN): an application interface to the network
  - Named Data Networking (NDN): a future Internet paradigm
  - Dynamic Circuits and managing the network as a resource (with CPU + storage)
  - Techniques to use 40G and 100G servers efficiently for 100G long distance flows
- With these ongoing developments, future link generations (400G, 1 Tbps) can be accommodated naturally (as we have done in the past) with affordable equipment
  - This also builds on ongoing joint work, such as 100Gbps data transfers during the SC conferences, and ongoing 100G-ANA transatlantic tests.
  - With HEP, network and corporate partners

Building on Ideas from 2006-7
Internet2's DCN Backbone

R. Summerhill



- Initial deployment was 10 x 10 Gbps wavelengths over the footprint
- First round maximum capacity – 80 x 10 Gbps wavelengths; expandable
- Scalability – potential migration to 40 Gbps or 100 Gbps capability
- Reliability – carrier-class standard assurances for wavelengths
- Transition to NewNet: 2006-7

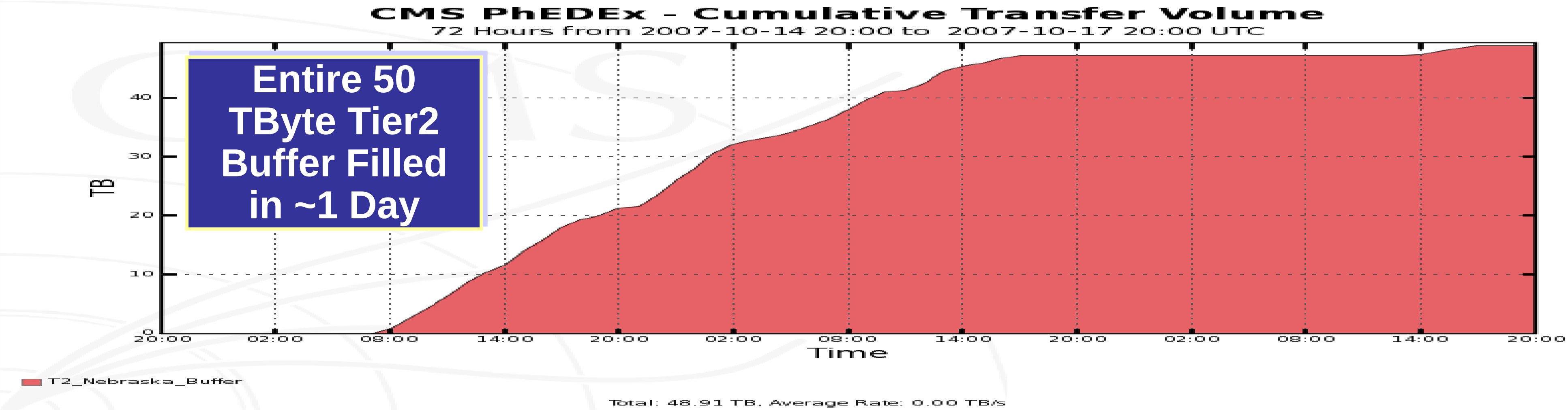


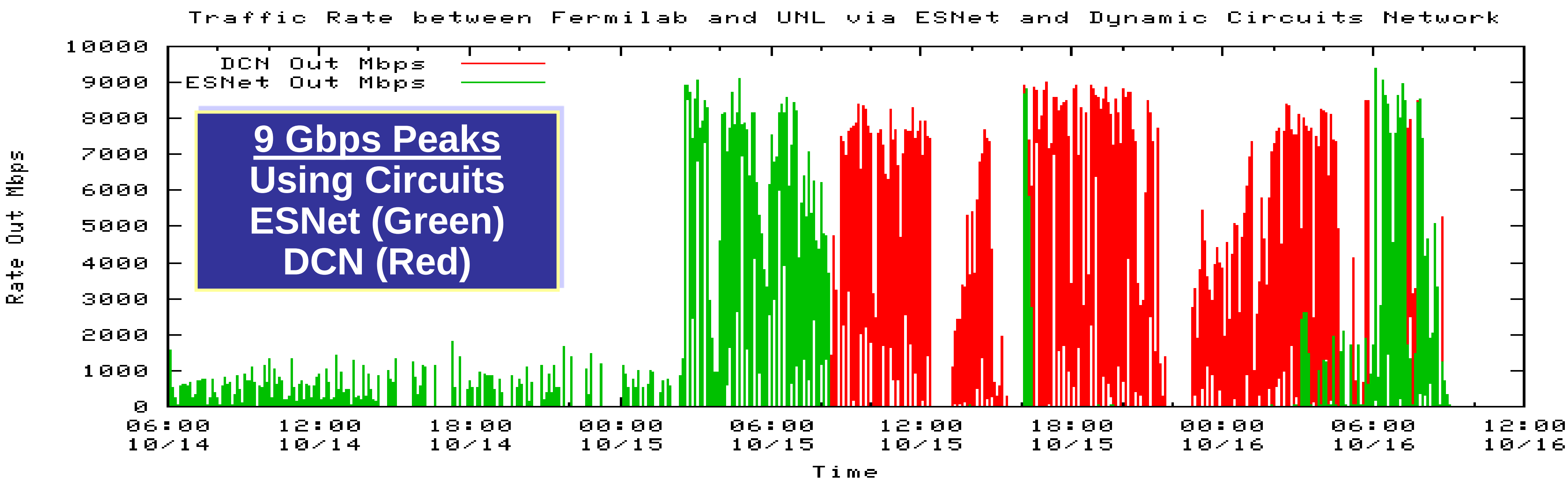
Rich Carlson of Internet2 Talk at ICFA DDW07 in 2007

FNAL – Nebraska
7.9 Gbps with Production Data, λStation Software (FNAL + Caltech)

Note there were 16 Tier2s and 66 Tier3s in US CMS and US ATLAS

CMS data transfer between FNAL and UNL using Internet2's DCN and LambdaStation Software (FNAL + Caltech) Cumulative transfer volume (top) and data rates (bottom)





Findings 5

Don't forget Tier3 Needs

Rich Carlson Internet2 Tier3 Talk

- Tier 3 computing and usage models are ill-defined
  - Bursty traffic demands
    - 1 – 2 TBytes of storage per person
    - 4 hours to move dataset
    - New dataset every 10 – 14 days
  - Some combination of local and remote resources will be used to solve problems
  - Chaotic usage patterns will dominate taking into account think time, data hot spots, and article preparation

Scenario

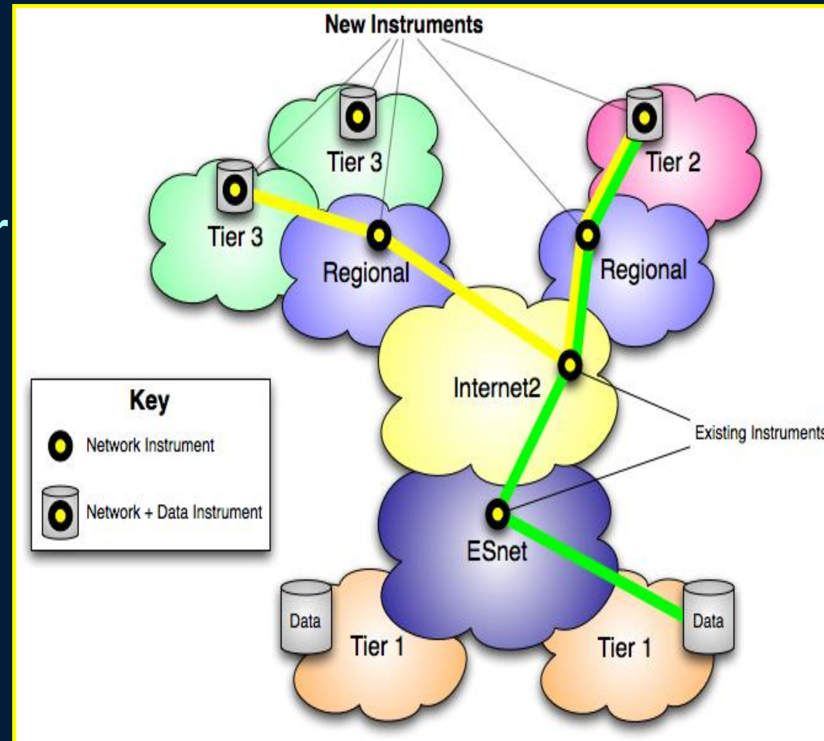
- **0.6 to 1.2 Gbps per flow, each 4 hrs long**
- **~1000 flows/10-14 days on Average; mainly 2 shifts**
- **Implies ~20 flows (total 12 to 24 Gbps) at once, on Average**
- **~1-2 flows per US Tier2 on Average with *Peaks + Spikes***
- **Potential for a lot of inter-regional T1-T1 and/or T1/T2 traffic, to fulfill the needs of the Tier2/Tier3 community**

DYNES: Dynamic Network System

Internet2, Caltech, Michigan, Vanderbilt

- **AIM: extend hybrid & dynamic capabilities to campus & regional networks.**
  - **DYNES cyberinstrument was designed to provide two basic capabilities to the Tier 2S, Tier3s and regional networks:**

1. **Network resource allocation to ensure transfer performance**
2. **Monitoring of the network and data transfer performance for reliability; resilience**



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

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

- **All networks in the path require the ability to allocate network resources and monitor the transfer. This capability currently exists on backbone networks such as ESnet, and in US LHCNet, but is not widespread at the campus and regional level.**
  - In addition Tier 2 & 3 sites require:

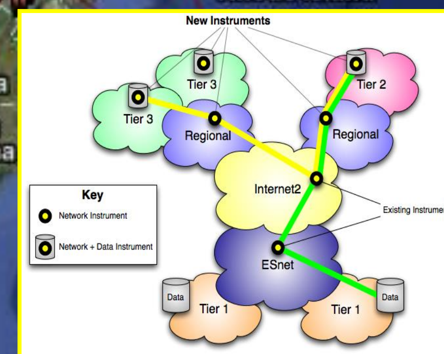
3. **Hardware at the end sites capable of making optimal use of the available network resources:**

DYNES: Dynamic Circuits Nationwide System. Created by Caltech, Led by Internet2

DYNES goal is to extend circuit capabilities to ~50 US campuses
Turns out to be nontrivial ➡

Partners: I2, Caltech, Michigan, Vanderbilt. Working with ESnet on dynamic circuit software

Extending the OSCARS scope; Transition: DRAGON to PSS, OESS



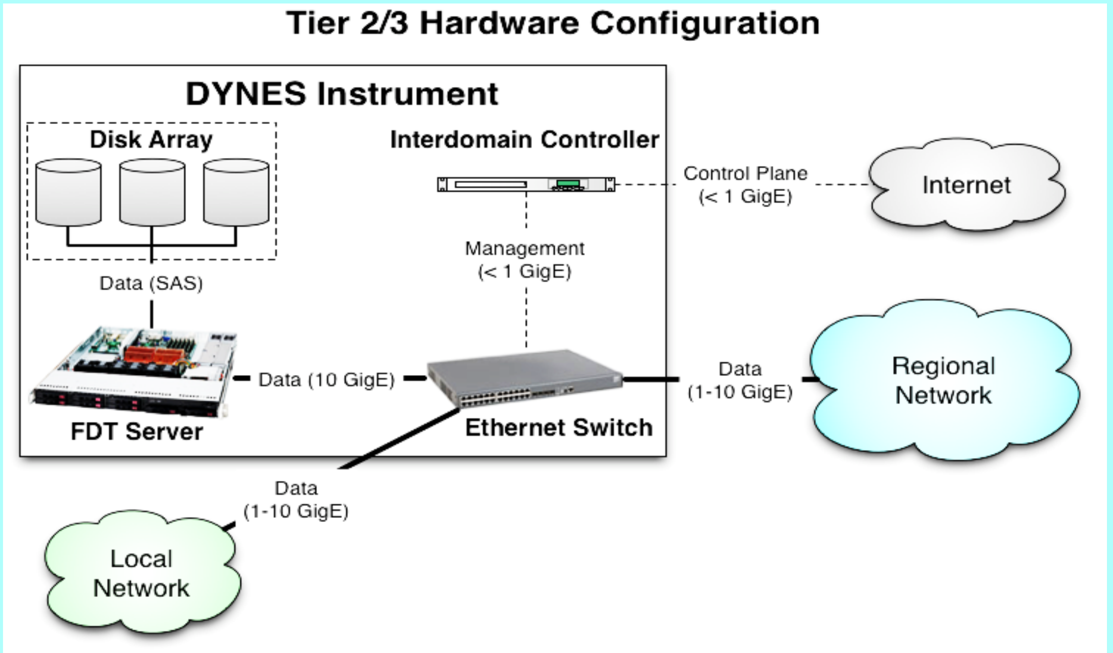
http://internet2.edu/dynes

Functionality will be an integral part of LHCONE point-to-point service: An Opportunity - Via SDN (OpenFlow and OpenDaylight)

DYNES: Tier2 and Tier3 Cyberinstrument Design

- Each DYNES (sub-)instrument at a Tier2 or Tier3 site consists of the following hardware, combining low cost & high performance:

1. An Inter-domain Controller (IDC)
2. An Ethernet switch
3. A Fast Data Transfer (FDT) server. Sites with 10GE throughput capability have a dual-port 10GE network interface in the server.
4. An optional attached disk array capable of several hundred MBytes/sec to local storage.

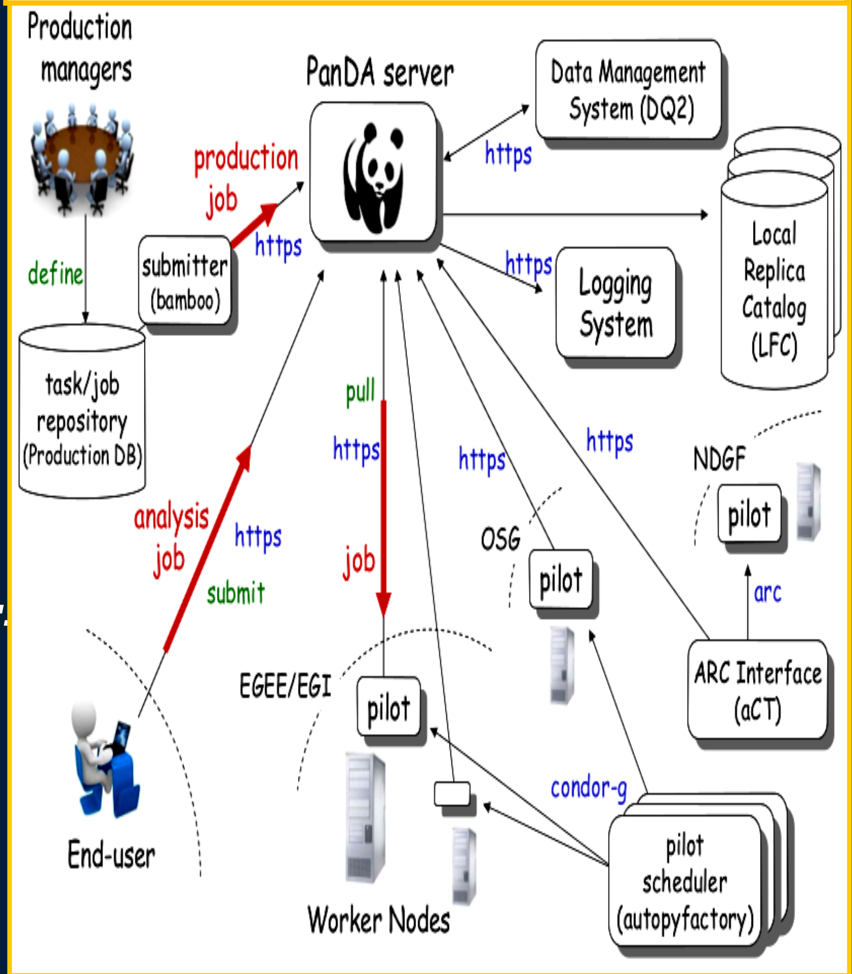


- Fast Data Transfer (FDT) server connects to the disk array and runs FDT software developed by Caltech.
- The disk array stores datasets to be transferred among the sites.
- The FDT server serves as an aggregator/ throughput optimizer in this case, feeding smooth flows over the networks directly to the Tier2 or Tier3 clusters.
- The IDC server handles allocation of network resources on the switch, interactions with other DYNES instruments related to network pro-visioning, and network performance monitoring.
  The IDC creates virtual LANs (VLANs) as needed.

ANSE: Advanced Network Services for Experiments: Manage LHC data flows

- US NSF funded project by Caltech, Vanderbilt, U. Michigan, UT Arlington
- Includes both US CMS and US ATLAS
- *Directly benefit the throughput and productivity of the LHC experiments*
- *Advanced use of dynamic circuits for optimized deterministic workflow*
- Interface advanced network services with LHC data management systems
  - *PanDA in (US) Atlas [De et al.]*
  - *PhEDEx in (US) CMS [Wildish et al.]*
- *Requires that the higher-levels in the experiments' software stacks interact directly with the network*
- *A fertile field for OpenFlow and other SDN Developments*

PanDA Workflow Management System



ANSE Tool Categories

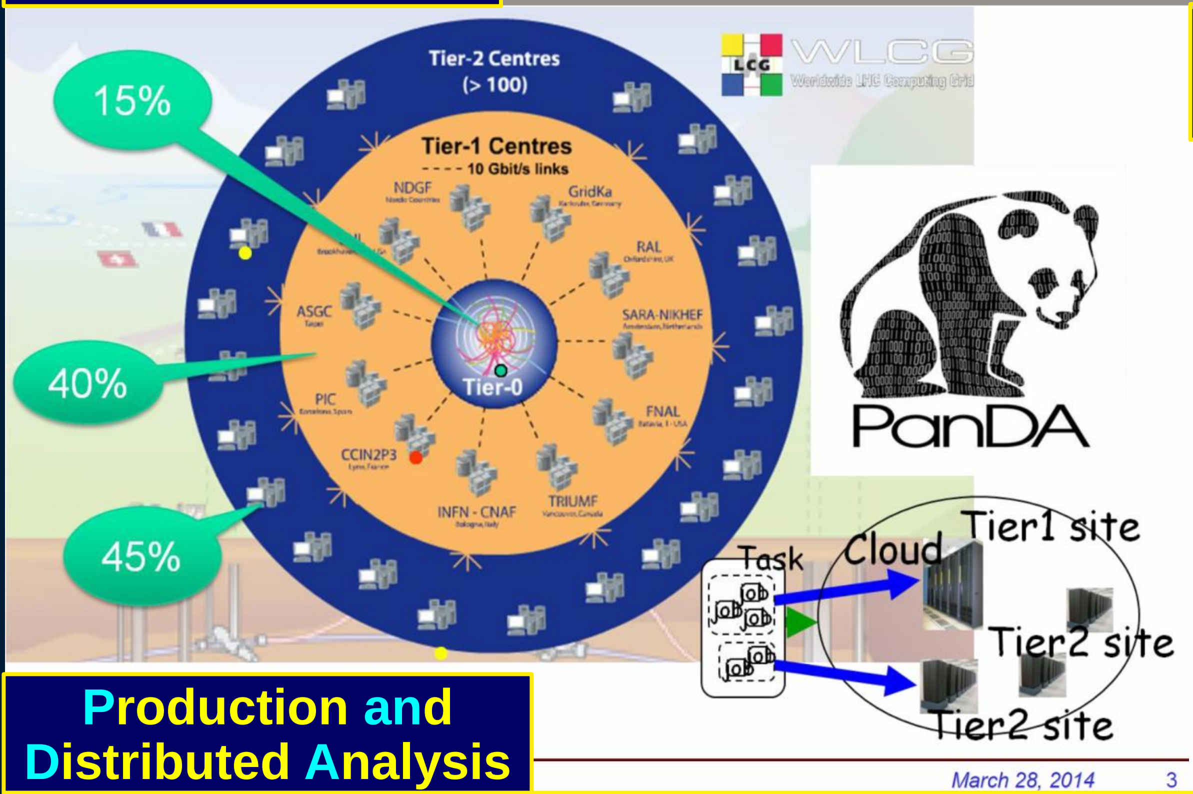
- ***Monitoring (Alone):***
  - Allows *Reactive* Use: React to “events” (State Changes) or Situations in the network
    - Throughput Measurements ➡ Possible Actions:
      (1) Raise Alarm and continue (2) Abort/restart transfers
      (3) Choose different source
    - Topology (+ Site & Path performance) Monitoring ➡ possible actions:
      (1) Influence source selection
      (2) Raise alarm (e.g. extreme cases such as site isolation)
- ***Network Control: Allows Pro-active Use***
  - Reserve Bandwidth Dynamically: prioritize transfers, remote access flows, etc.
  - Co-scheduling of CPU, Storage and Network resources
  - Create Custom Topologies ➡ optimize infrastructure to match operational conditions: deadlines, workprofiles
    - e.g. during LHC running and/or re-reconstruction/re-distribution

ANSE Activities

- Initial sites: UMich, UTA, Caltech, Vanderbilt, CERN, UVIC
- Monitoring information for workflow and transfer management
  - Define path characteristics to be provided to FAX and PhEDEx
    - using perfSONAR info to predict loading for each pair
  - On a NxN mesh of source/destination pairs
  - Could also use LISA agents to gather end-system information
- Dynamic Circuit Systems
  - Working with DYNES at the outset
    - monitoring dashboard, full-mesh connection setup and BW test
  - Deployed a prototype PhEDEx instance for development and evaluation
    - Integration with network services
  - Potentially use LISA agents for pro-active end-system configuration

Integrating Network Awareness in ATLAS Distributed Computing

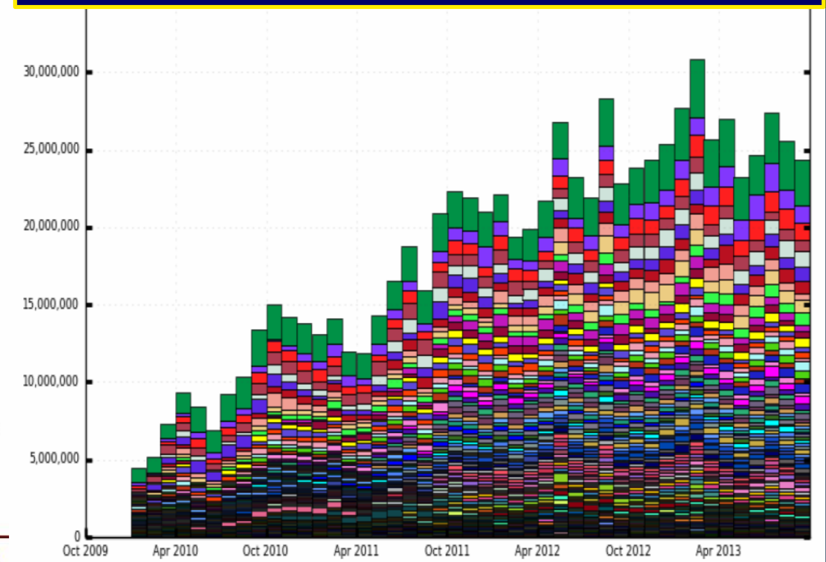
ANSE: Kaushik De



Production and Distributed Analysis

25M Jobs at > 100 Sites Now Completed Each Month

6X Growth in 3 Years (2010-13)

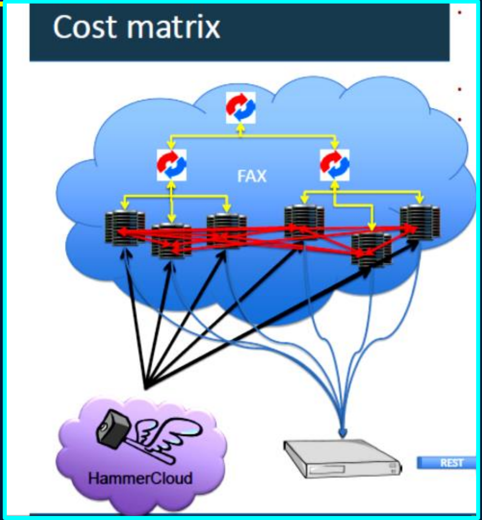


- STEP1: Import network information into PanDA
- STEP2: Use network information directly to optimize workflow for data transfer/access; at a higher level than individual transfers alone
  - Start with simple use cases leading to measureable improvements in workflow/user experience

Integrating Network Awareness in ATLAS Distributed Computing

USE CASES

Kaushik De



1. Faster User Analysis
   - Analysis jobs normally go to sites with local data: sometimes leads to long wait times due to queuing
   - Could use network information to assign work to 'nearby' sites with idle CPUs and good connectivity
2. Cloud Selection
   - Tier2s are connected to Tier1 "Clouds", manually by the ops team (may be attached to multiple Tier1s)
   - To be automated using network info: Algorithm under test
3. PD2P = PanDA Dynamic Data Placement: Asynchronous usage-based
   - Repeated use of data or Backlog in Processing → Make add'l copies
   - Rebrokerage of queues → New data locations
   - PD2P is perfect for network integration
     - Use network for site selection – to be tested soon
     - Try SDN provisioning since this usually involves large datasets; requires some dedicated network capacity

ANSE: Advanced Network Services for Experiments. CMS Developments

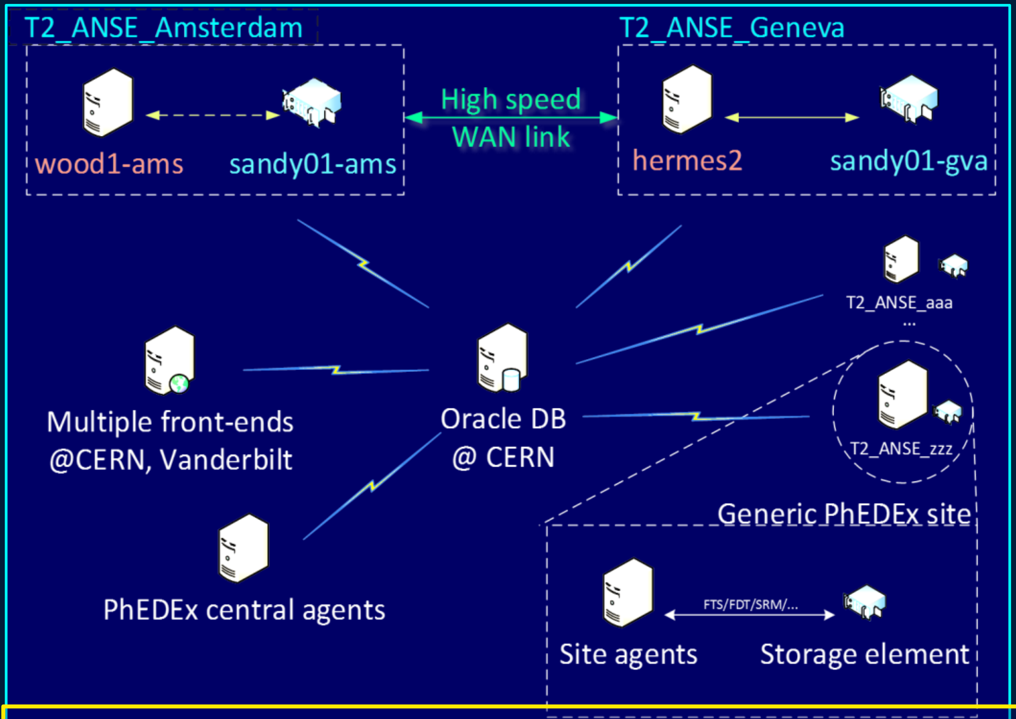
- Implemented circuit interface in PhEDEx
  - Developed a site circuit agent
    - receives creation requests from download agents
    - checks the database and the lookup server to see if circuits are actually allowed on the current link
    - Handles the creation (and tear-down) of the circuits
- Testbed: Switched to using dynamic circuits between Geneva and Amsterdam
  - Over US LHCNet, using OSCARS
  - First results very promising
- Plans: include other DYNES sites; move to pre-production then production use

ANSE: Performance measurements (AMS-GVA) with PhEDEx and FDT for CMS

T2_ANSE_Geneva & T2_ANSE_Amsterdam

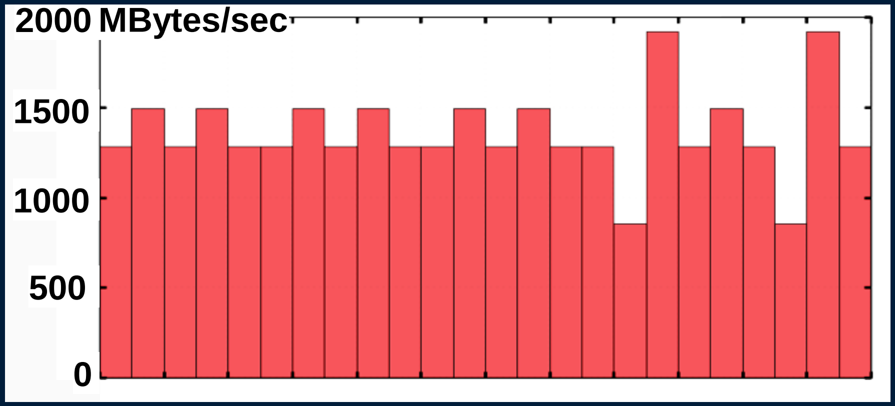
- High Capacity links with dynamic circuit creation between storage nodes
- PhEDEx and storage nodes separate
- 4x4 SSD RAID 0 arrays,
  16 physical CPU cores / machine

PhEDEx testbed in ANSE

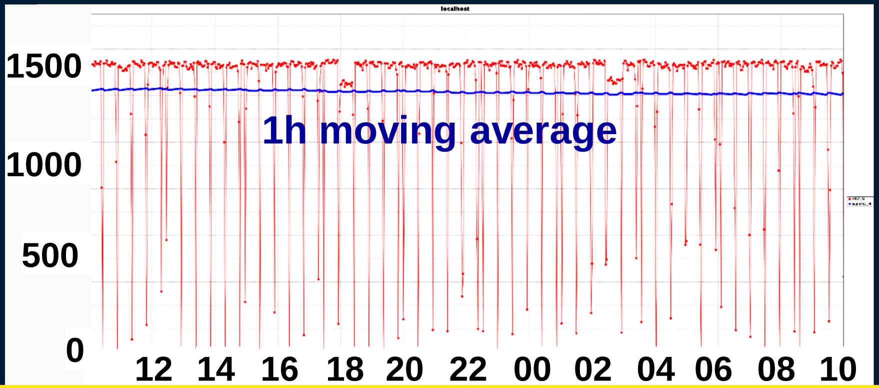


- FDT sustained rates: ~1500 MB/sec
- Average over 24hrs: ~ 1360 MB/sec
- Difference due to delay in starting jobs
- Bumpy plot due to binning + 2 Gbyte file blocks

24 hrs, as Reported by PhEDEx



Throughput as reported by MonALISA



Next Step: Deploy in Production. Development ongoing Now.

OLiMPS

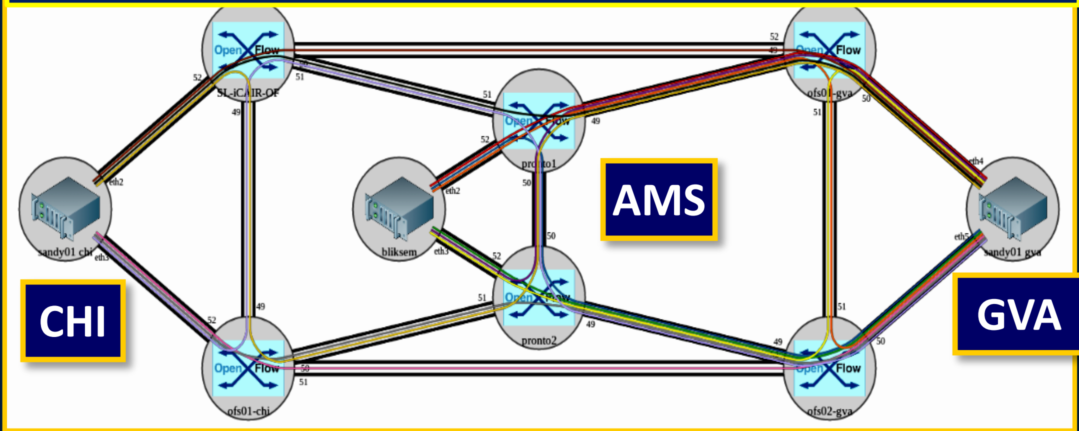
Openflow Link-layer Multipath Switching

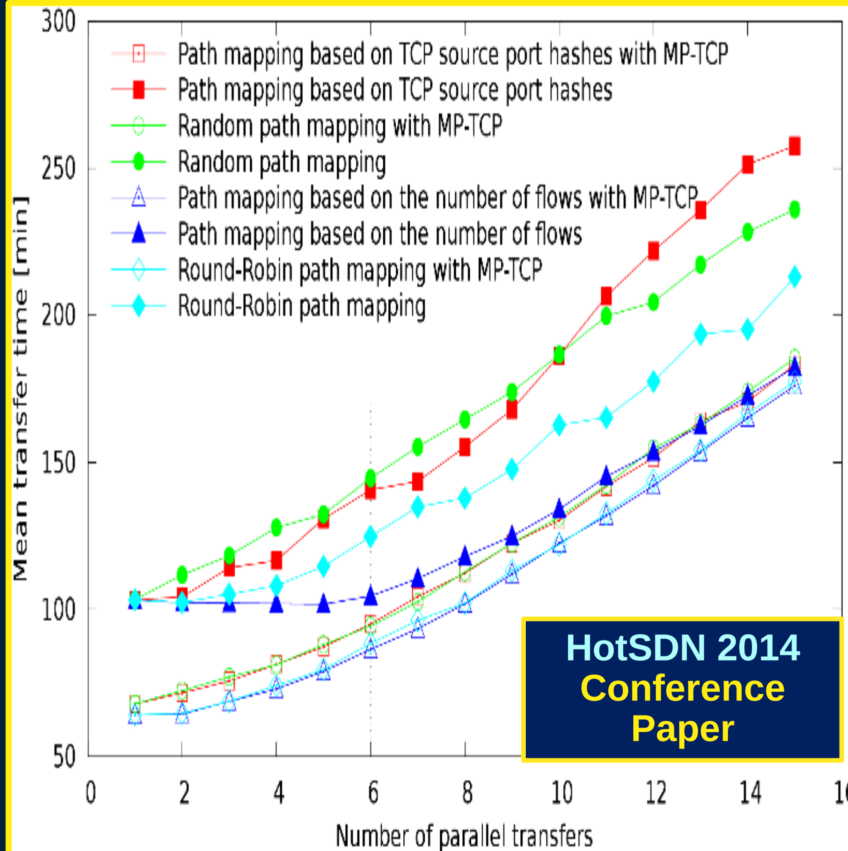
- Project funded by DOE OASCR in 2012-2014
- Research Focus: Efficient data intensive workflow over complex networks
- First Use Case: LHCONE Multipath problem solution
  - Allows for per-flow multipath switching, which
    - Increases the robustness
    - Increases efficiency
    - Simplifies management of layer 2 network resources
  - Construct a robust multi-path system without modifications to the Layer 2 frame structure, using central out-of-band software control
      - A Big Plus: using Openflow, there is no need for new hardware or feature support (other than Openflow)
- Caveat: coding is required, not for the faint-hearted
  - (No, we cannot just buy a controller)

OLiMPS: SDN (OpenFlow) use case in LHCONE: Solving the Multipath problem

- Address the problem of topology limitations in large scale networks

Geneva-Amsterdam-Chicago testbed at SC12

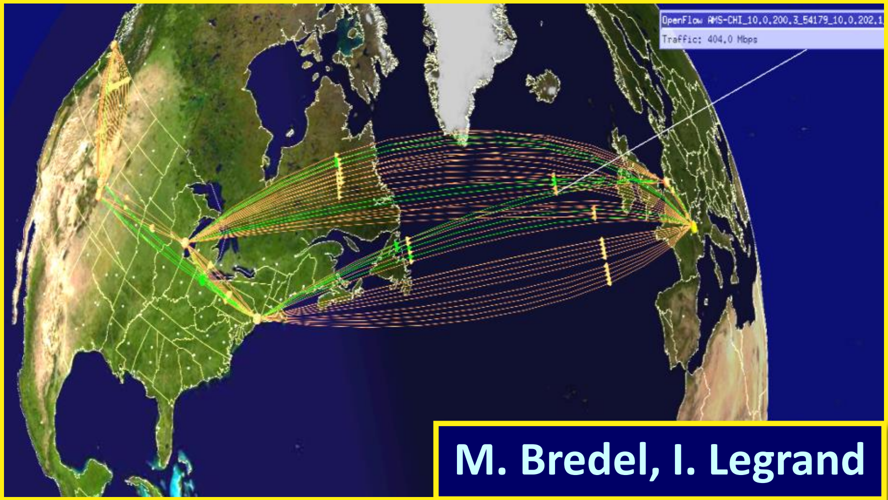




- Idea: Flow-based load balancing over multiple paths ➡ throughput optimization
  - Leverage global network view of the OpenFlow controller
  - Initially: used static topology
  - Next Step (Cisco grant to Caltech): comprehensive real-time info. from the network (utilization, capacity, topology) as well as a full interface to applications

Results: showed a large throughput improvement when using an application interface and load-aware flow assignments

Caltech + Partners: OpenFlow Testbed Demo with MonALISA at SC13

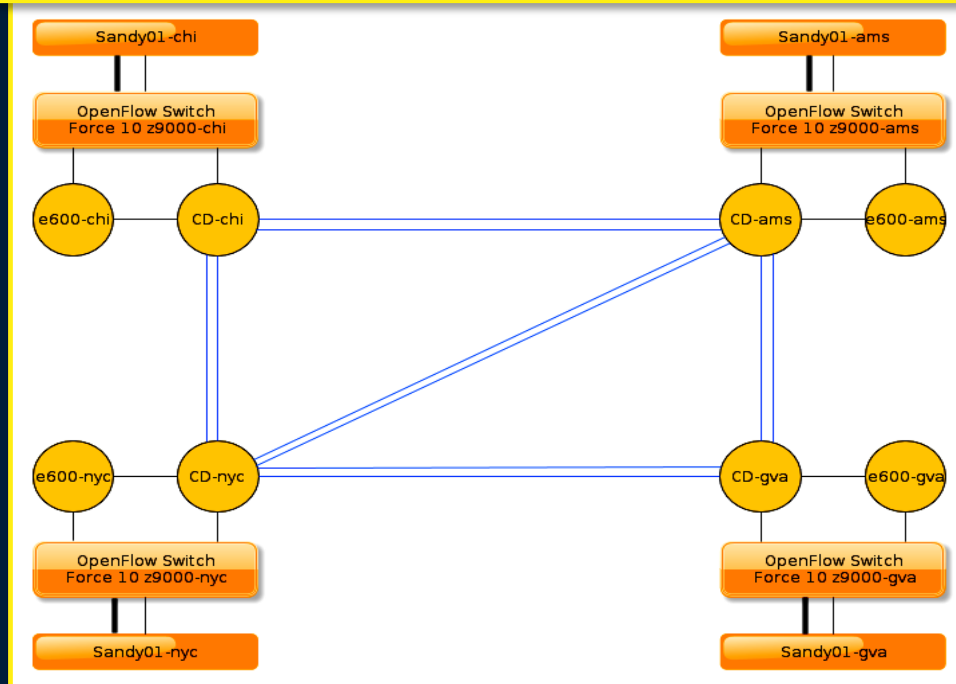


M. Bredel, I. Legrand

- 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

High Speed Data Transfers for HEP

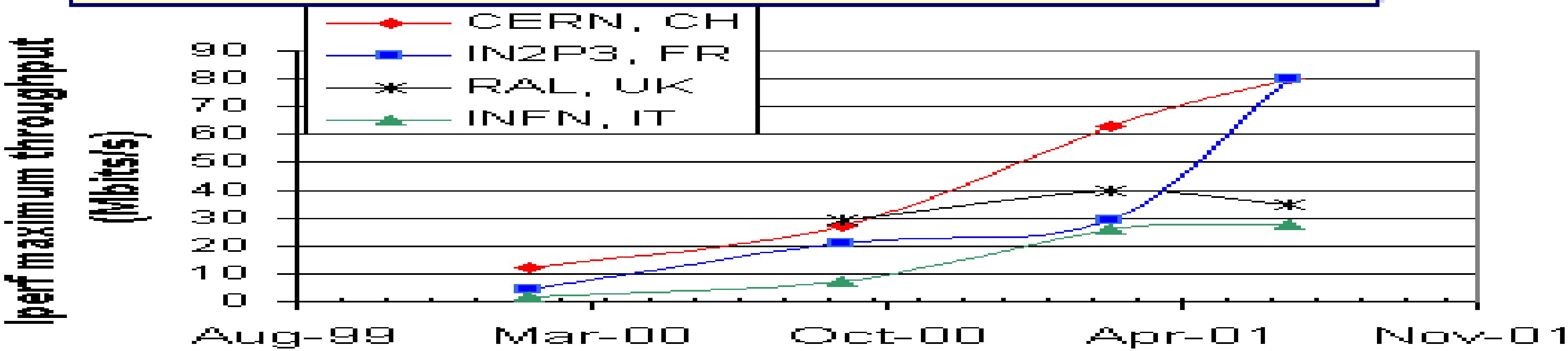
The State of the Art

Major Advances in Data Transfer Applications

Led by HEP with Computer Scientists and Network Engineers

- **2000-2014:** HEP with computer scientists and network engineers developed the knowledge to use long distance networks efficiently, at high occupancy, for the first time
  - "Demystification" of large long range data flows with TCP: From 0.1 to ~1 Gbps streams by 2002
    - 2004-2005: Up to 10 Gbps per flow;
    - One to a few server-pairs matches a 10G link
    - Aggregate from 23 Gbps (SC03) to 151 Gbps (SC05) to 339 Gbps with 175 Gbps storage to storage (SC12)
    - Flows to 40 Gbps starting in 2011; Moving towards ~100G flows from 2012. Waiting for 100G NICs
- Major advances in the TCP stack (FastTCP; Cubic), kernel, end system architecture, network interfaces (10GE, 40GE), drivers and applications, ~since 2002.
- From 2006: Moved to mature storage-to-storage transfer applications; working on transfers among *storage-systems*

1999-2003: HEP Learned to Use 1-10G Networks Fully: Factor of ~50 Gain in Max. Sustained TCP Thruput in 2 Years, On Some US+Transoceanic Routes



- **9/01 105 Mbps 30 Streams: SLAC-IN2P3; 102 Mbps 1 Stream CIT-CERN**
- **5/20/02 450-600 Mbps in 100 Streams SLAC-Manchester on 622 Mbps Link**
- **6/1/02 290 Mbps Chicago-CERN One Stream on 622 Mbps Link**
- **9/02 850, 1350, 1900 Mbps Chicago-CERN 1,2,3 GbE Streams, 2.5G Link**
- **11/02 [LSR] 930 Mbps in 1 Stream California-CERN, and California-AMS FAST TCP 9.4 Gbps in 10 Flows California-Chicago at SC02**
- **2/03 [LSR] 2.38 Gbps in 1 Stream California-Geneva (99% Link Utilization)**
- **5/03 [LSR] 0.94 Gbps IPv6 in 1 Stream Chicago- Geneva**
- **TW & SC2003: 5.65 Gbps (IPv4), 4.0 Gbps (IPv6) in 1 Stream Over 11,000 km**

FAST TCP: Baltimore/Sunnyvale 2002

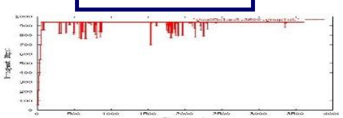
- RTT estimation: fine-grain timer
- Fast convergence to equilibrium
- Delay monitoring in equilibrium
- Pacing: reducing burstiness

Measurements 11/02

- Std Packet Size
- Utilization averaged over > 1hr
- 4000 km Path

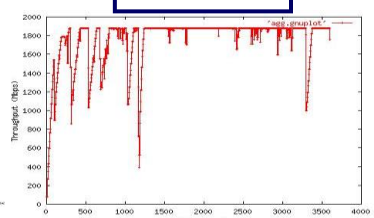
Average utilization

95%



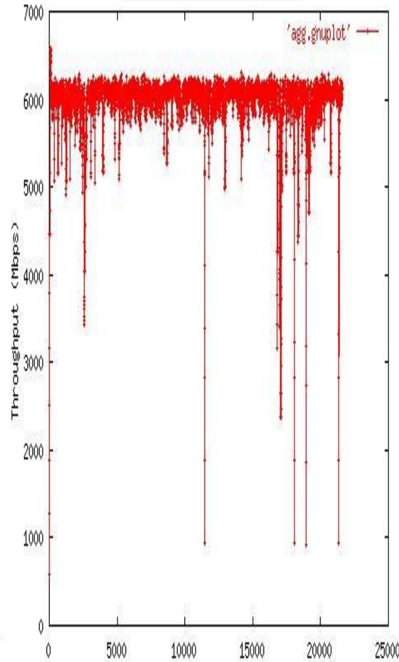
1 flow

92%



2 flows

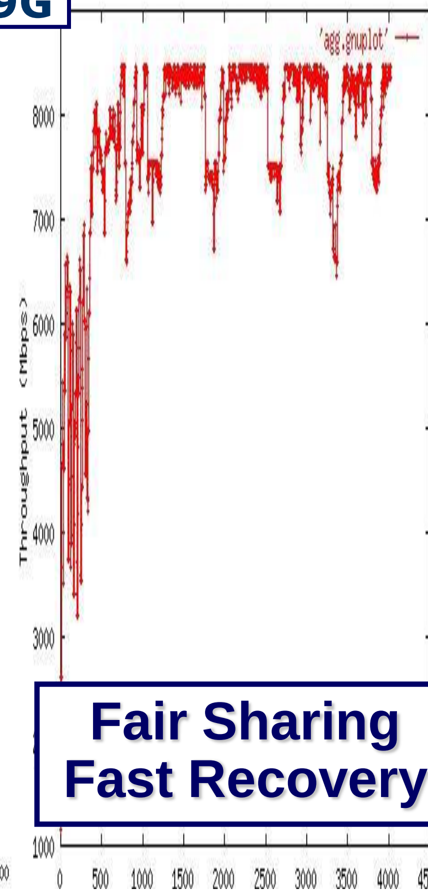
90%



7 flows

9G

90%

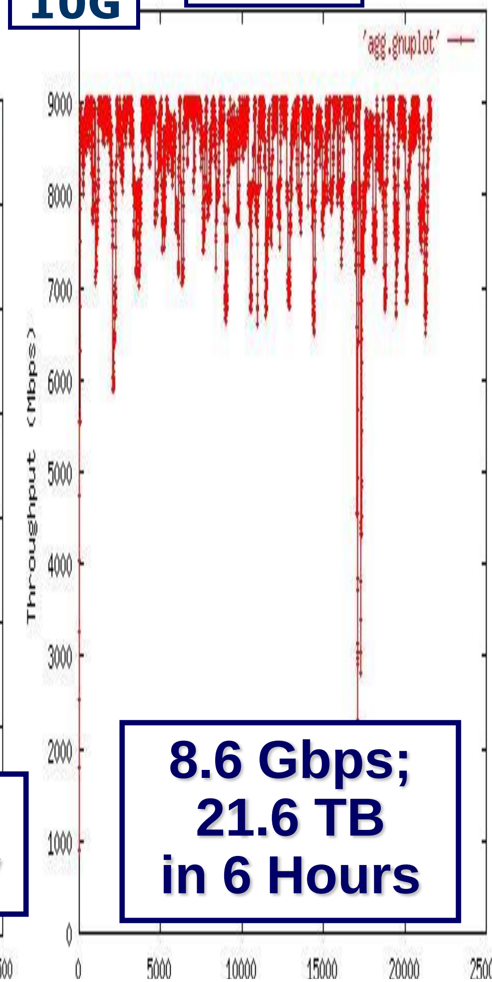


Fair Sharing
Fast Recovery

9 flows

10G

88%

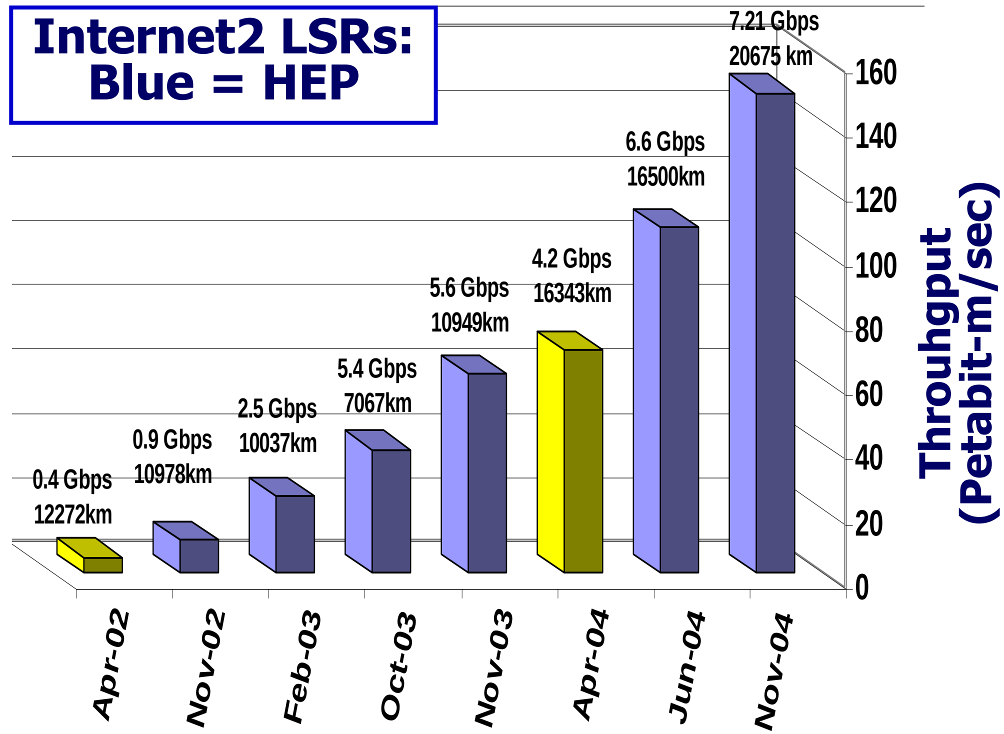


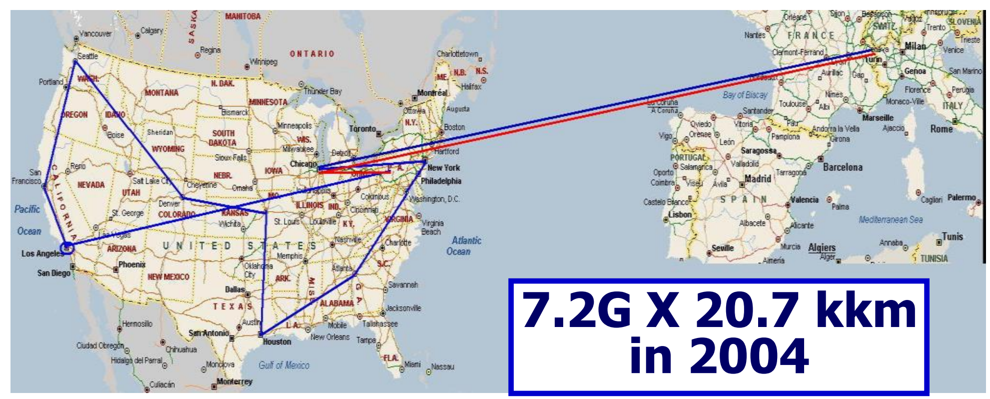
8.6 Gbps;
21.6 TB
in 6 Hours

10 flows

Internet2 Land Speed Records & SC2003-2005 Records

- IPv4 Multi-stream record
  6.86 Gbps X 27kkm: Nov 2004
- PCI-X 2.0: 9.3 Gbps Caltech-StarLight: Dec 2005
- PCI Express 1.0: 9.8 Gbps Caltech – Sunnyvale, July 2006
- *Concentrate now on reliable Terabyte-scale file transfers*
  - *Disk-to-disk Marks: 536 Mbytes/sec (Windows); 500 Mbytes/sec (Linux)*
  - *System Issues: PCI Bus, Network Interfaces, Disk I/O Controllers, Linux kernel,CPU*
- *SC2003-5: 23, 101, 151 Gbps*
- *SC2006: FDT app.: Stable disk-to-disk at 16+ Gbps on one 10G link*

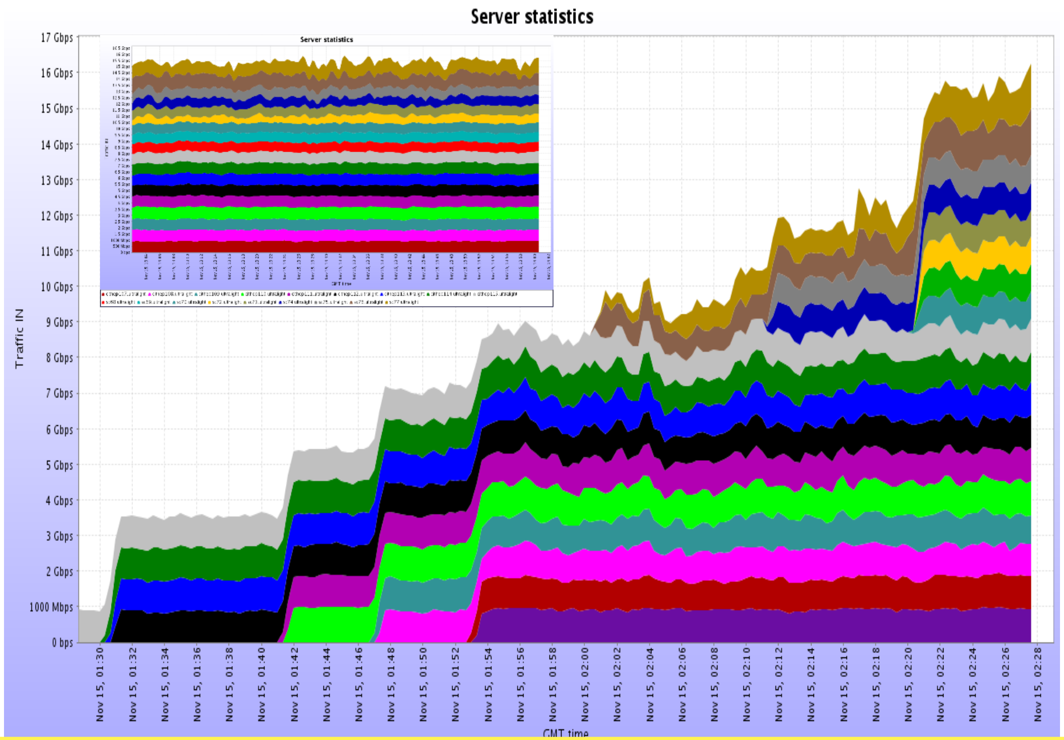




FDT: Fast Data Transport
Results 11/14 – 11/15/06

I. Legrand

- **Stable disk-to-disk flows Tampa-Caltech: Stepping up to 10-to-10 and 8-to-8 1U Server-pairs 9 + 7 = 16 Gbps; then Solid overnight. Using One 10G link**



Capability Level circa 2007: 40-70 Gbps per rack of low cost servers

Efficient Data Transfers

- **Reading and writing at disk speed over WANs (with TCP) for the first time**
- **Highly portable: runs on all major platforms.**
- **Based on an asynchronous, multithreaded system, using Java NIO libraries**
- **Streams a *dataset* (list of files) continuously, from a managed pool of buffers in kernel space, through an open TCP socket**
  - **Smooth data flow from each disk to/from the network**
  - **No protocol start-phase between files**

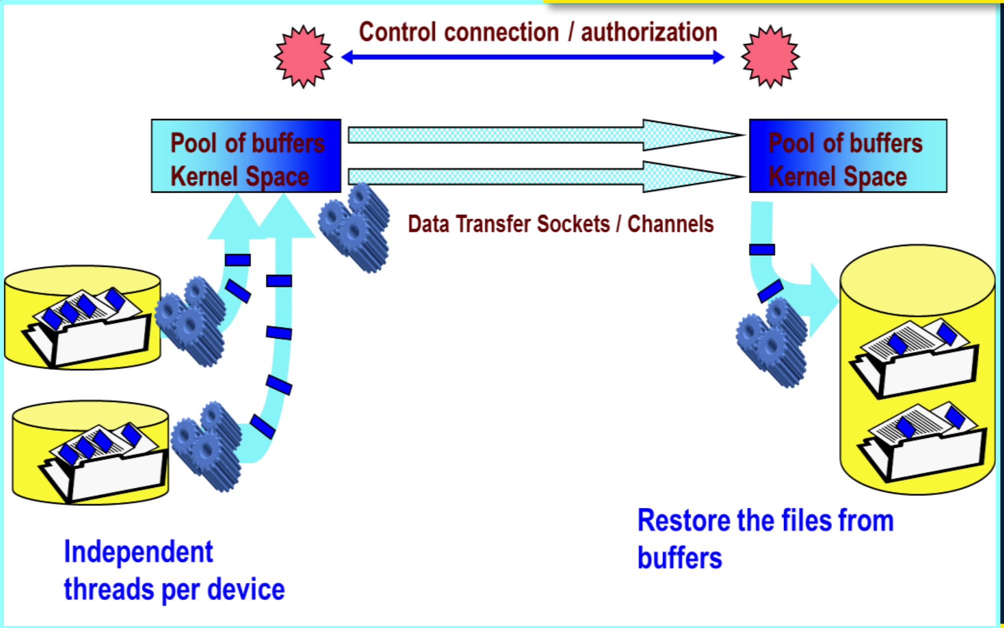
Fast Data Transfer (FDT) 2006

http://monalisa.caltech.edu/FDT

The state of the art in data transfers ever 2006

- FDT: an open source Java application **for WAN efficient data transfers**
- Streams data over long distances at disk speeds through an open TCP socket: **no session starts/stops**
- **Based on an asynchronous, multithreaded system:** schedules many logical threads on a few OS threads
  - **Decomposes** any list of files into a pool of buffers in kernel space
  - Read and write on each physical device **with independent threads**
  - **Appropriately size buffers** to match the end systems' disk IO
  - **Moderate rate of sending buffers** to match the measured net path capacity in real time
  - **Uses parallel streams if needed**

FDT uses IDC API to request dynamic circuit connections

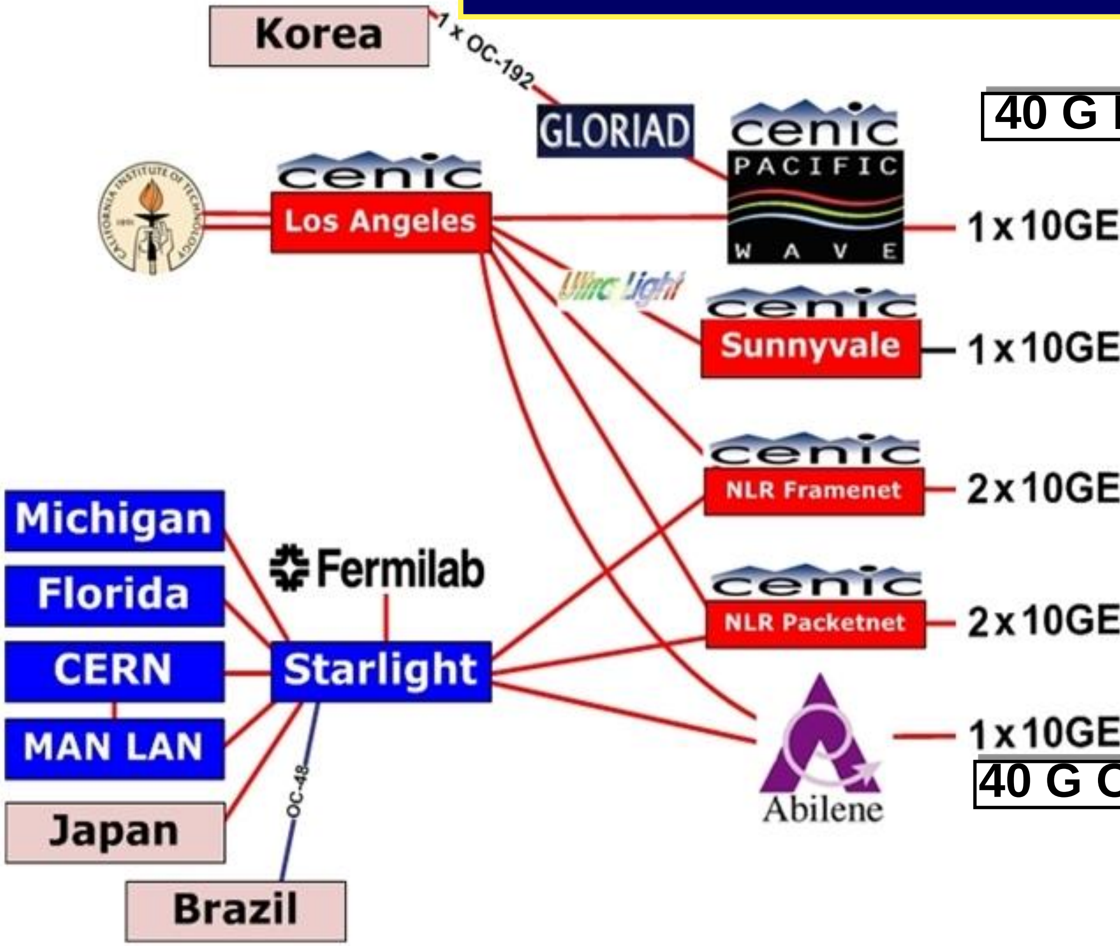


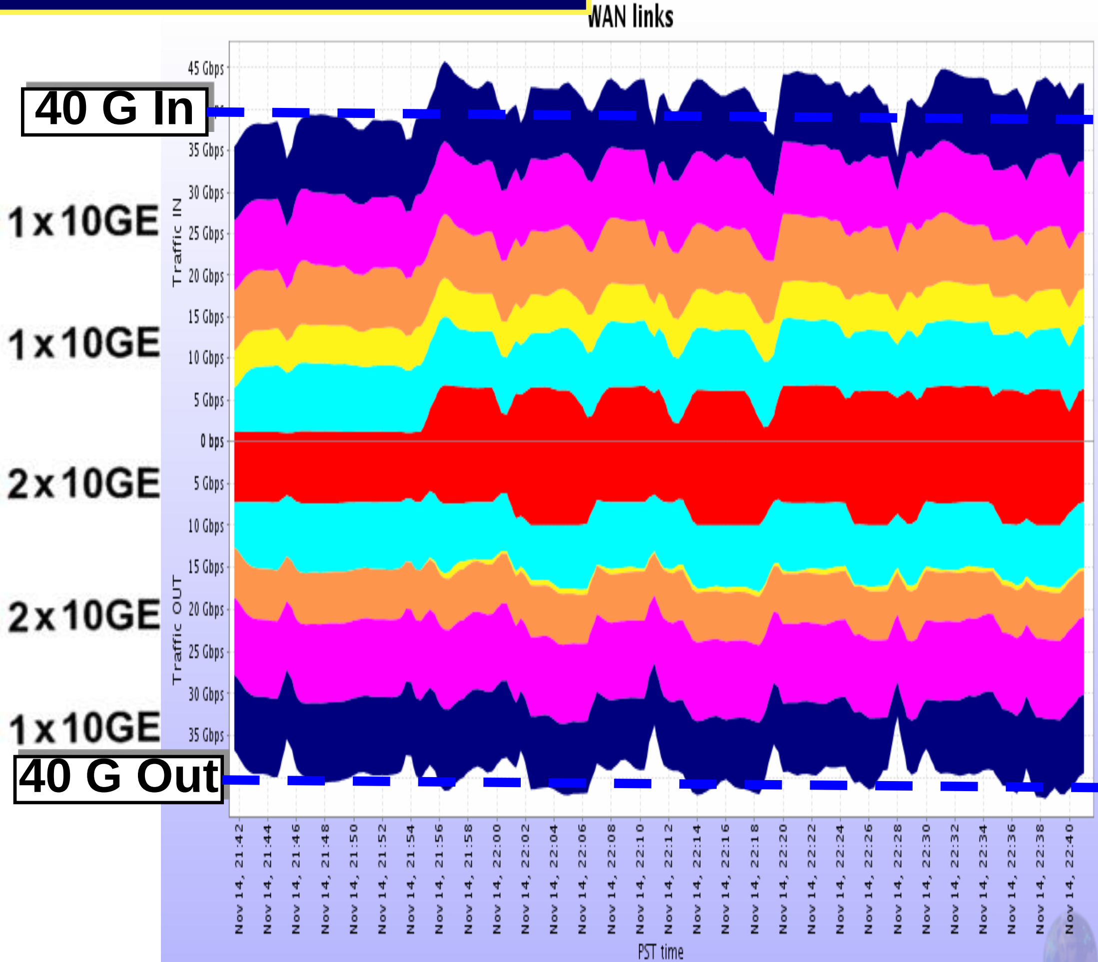
Integration with the main storage systems used by the LHC experiments: dCache, Hadoop, xrootd, Lustre; also PhEDEx and FAX (in progress)

HEP SC2007

One rack of servers: 80+ Gbps Sustained for Hours, also with Non-Zero Packet Loss

A Four Continental Collaborative Effort





Inherent throughput capability of Tier1 & Tier2 servers:
2007 View: Could exceed the affordable transoceanic bandwidth by an order of magnitude or more

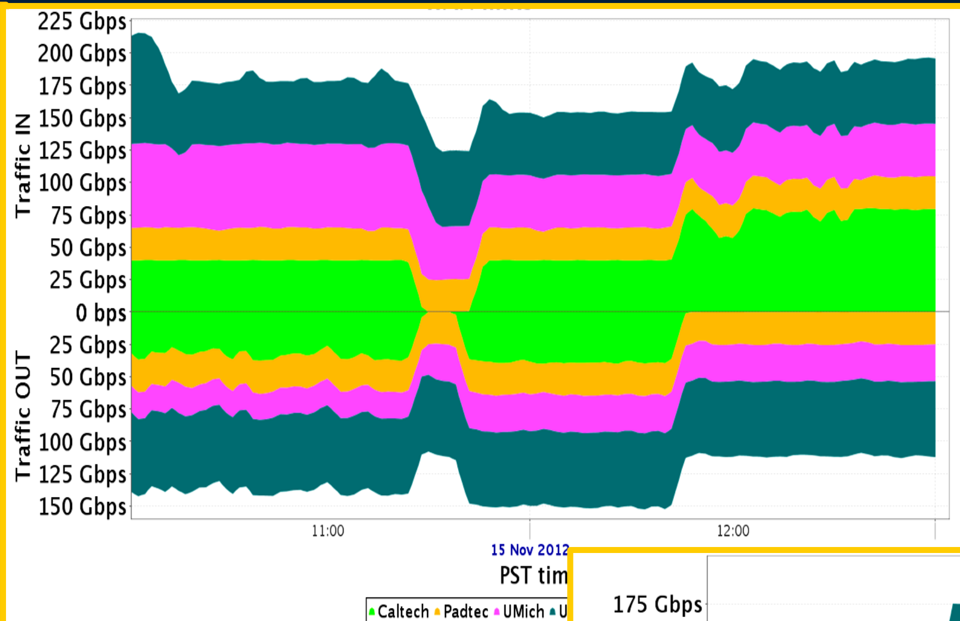
SC12 November 14-15 2012 (4 100G Links)

Caltech-Victoria-Michigan-Vanderbilt; BNL

***FDT* Memory to Memory**

300+ Gbps In+Out Sustained from Caltech, Victoria, UMich

To 3 Pbytes Per Day

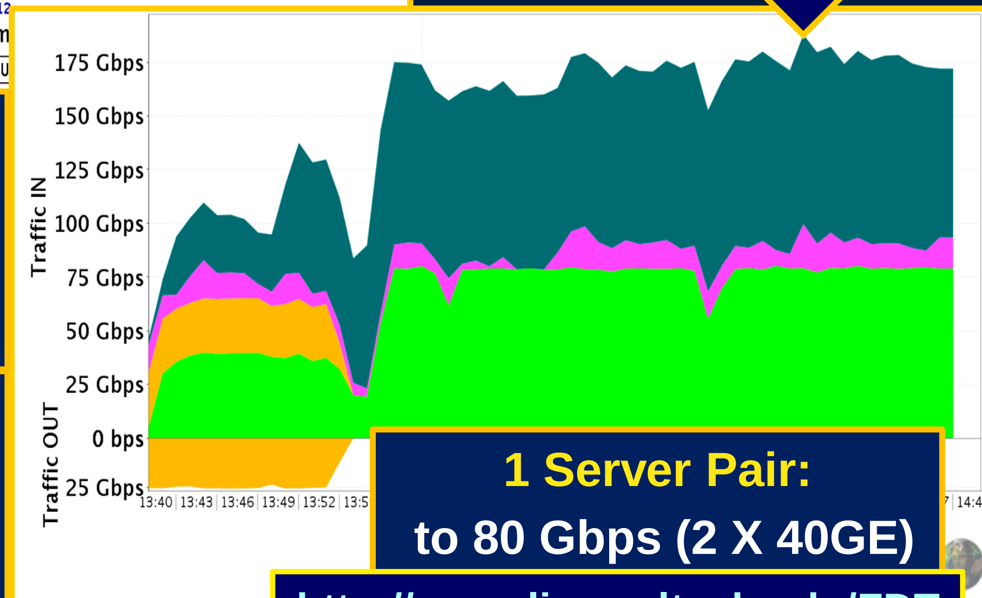


FDT Storage to Storage

http://monalisa.caltech.edu/FDT/

175 Gbps (186 Gbps Peak)

Extensive use of FDT, Servers with 40G Interfaces. + RDMA/Ethernet



1 Server Pair: to 80 Gbps (2 X 40GE)

http://monalisa.caltech.edu/FDT

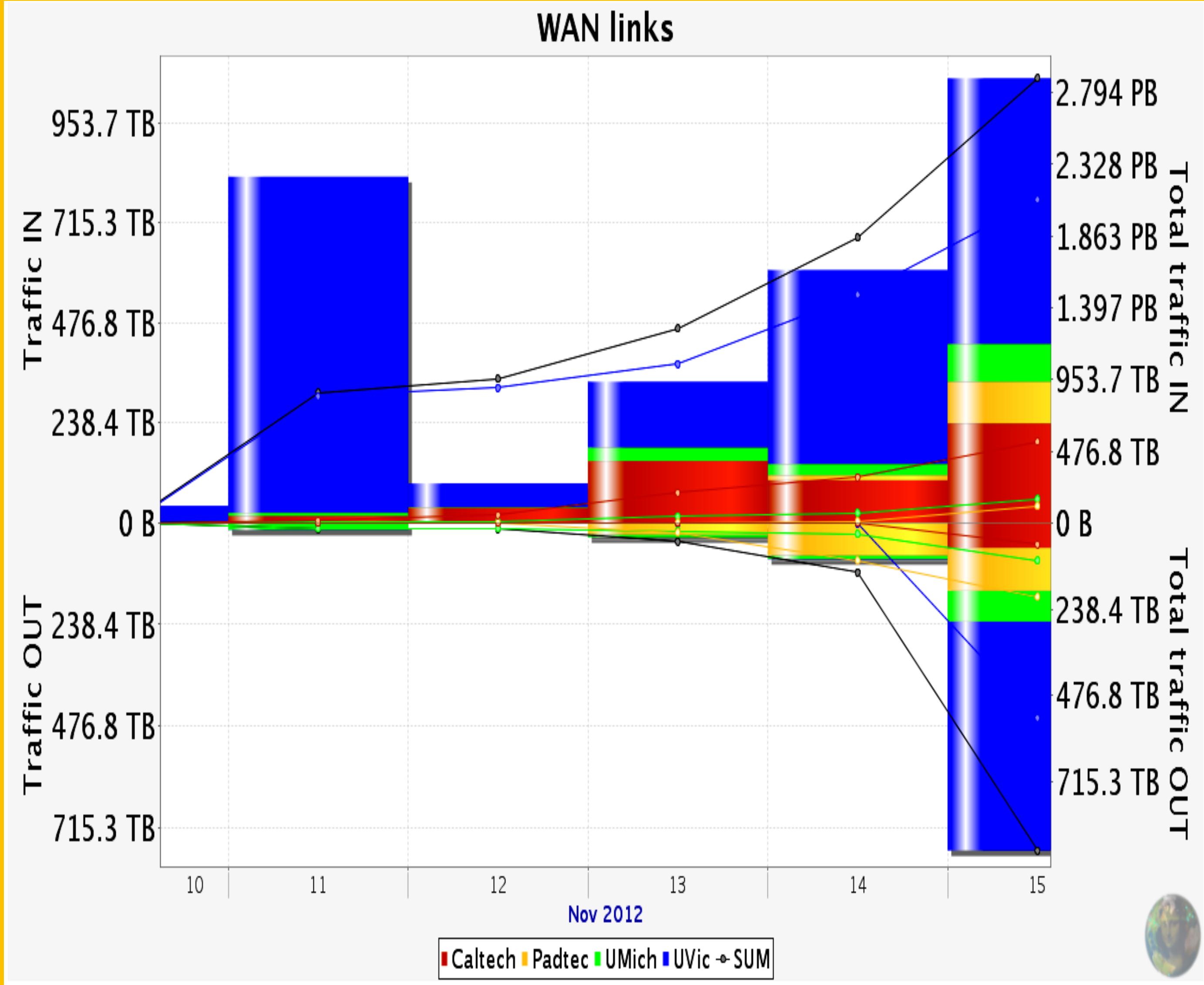
HEP Team and Partners Have defined the state of the art in high throughput long range transfers since 2002

Transferring Petabytes at SC12

FDT and RDMA over Ethernet

3.8 PBytes to and From the Caltech Booth

Including 2 PBytes on Nov. 15

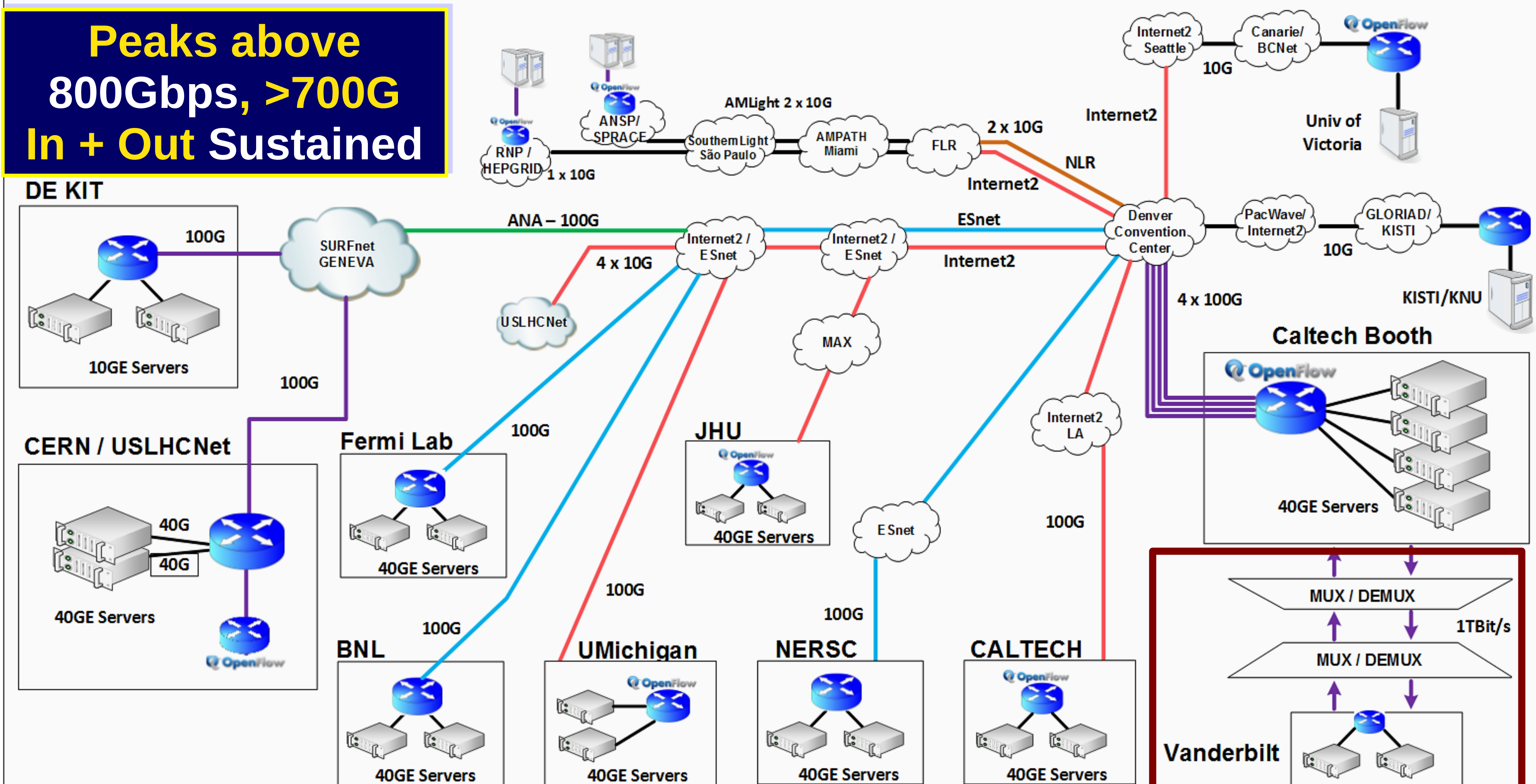


Caltech Booth at SC13 (Denver)
Terabit per second trials

Caltech, UVic, Vanderbilt, Sao Paolo, Karlsruhe, Michigan, JHU, Fermilab, BNL, ESnet

HEP Terabit Wide Area and Local Area Network: Caltech Booth at SC13

Peaks above 800Gbps, >700G In + Out Sustained



Supporting Vendors: Mellanox, Brocade, Echostreams, Intel, Cisco, Dell, Padtec, Ciena, SGI, Seagate, FusionIO, iWnetworks

1 Tbps Scale Demonstration: Caltech, Uvic, Vanderbilt, CERN Sao Paolo, Karlsruhe, Michigan, JHU, Fermilab, BNL, ESnet

Padtec 1 Tbps [*]
DWDM System:
7 X 100G and
8 X 40G Waves

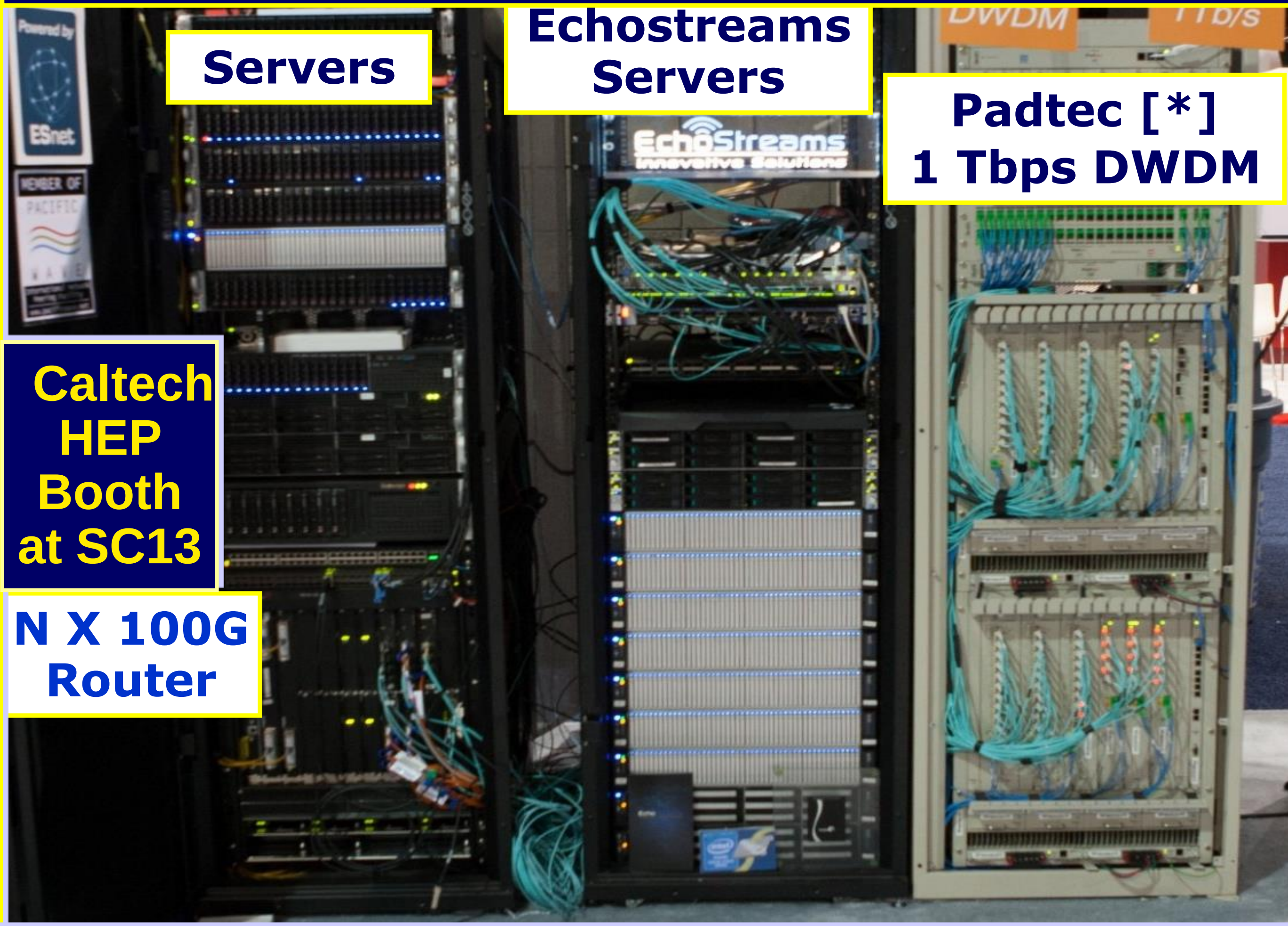
Connected to
Vanderbilt Booth
with similar setup

Echostreams 2U
Servers: 48 SSD ea.

N x 100G Brocade
Switch-Router

700 Seagate
and Intel SSDs

Peaks above 800 Gbps, 750 Gbps Sustained



Network partners: SciNet, ESnet, Internet2, ANA-100, CENIC, Starlight, MANLAN, MiLR, SURFNet, RNP, ANSP, AmLight

1 Tbps Scale Demonstration: Caltech, Uvic, Vanderbilt, CERN Sao Paolo, Karlsruhe, Michigan, JHU, Fermilab, BNL, ESnet

Padtec 1 Tbps [*]

DWDM System:
7 X 100G and
8 X 40G Waves

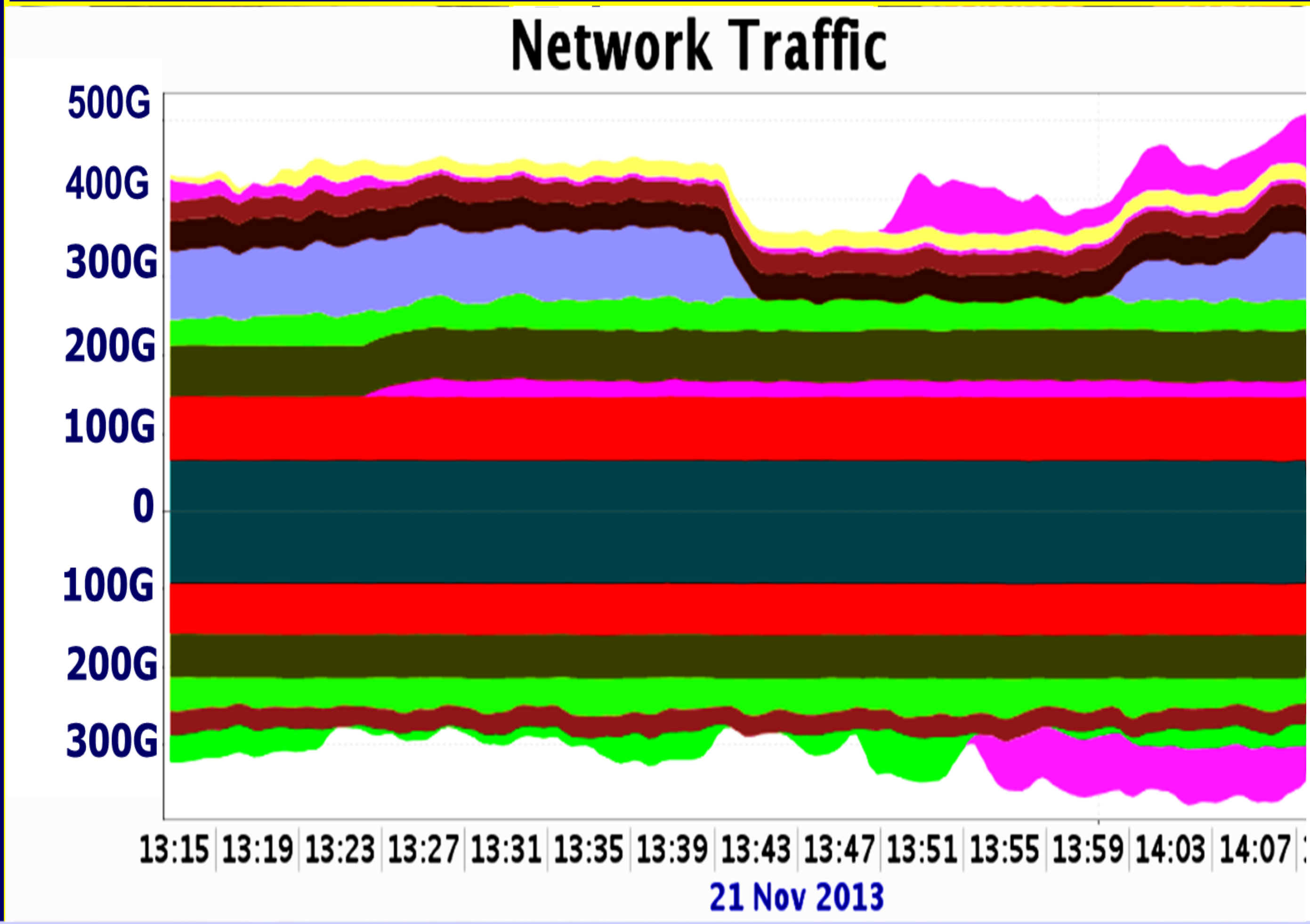
Connected to
Vanderbilt Booth
with similar setup

Echostreams 2U
Servers: 48 SSD ea.

N x 100G Brocade
Switch-Router

700 Seagate
and Intel SSDs

Peaks above 800 Gbps, 750 Gbps Sustained

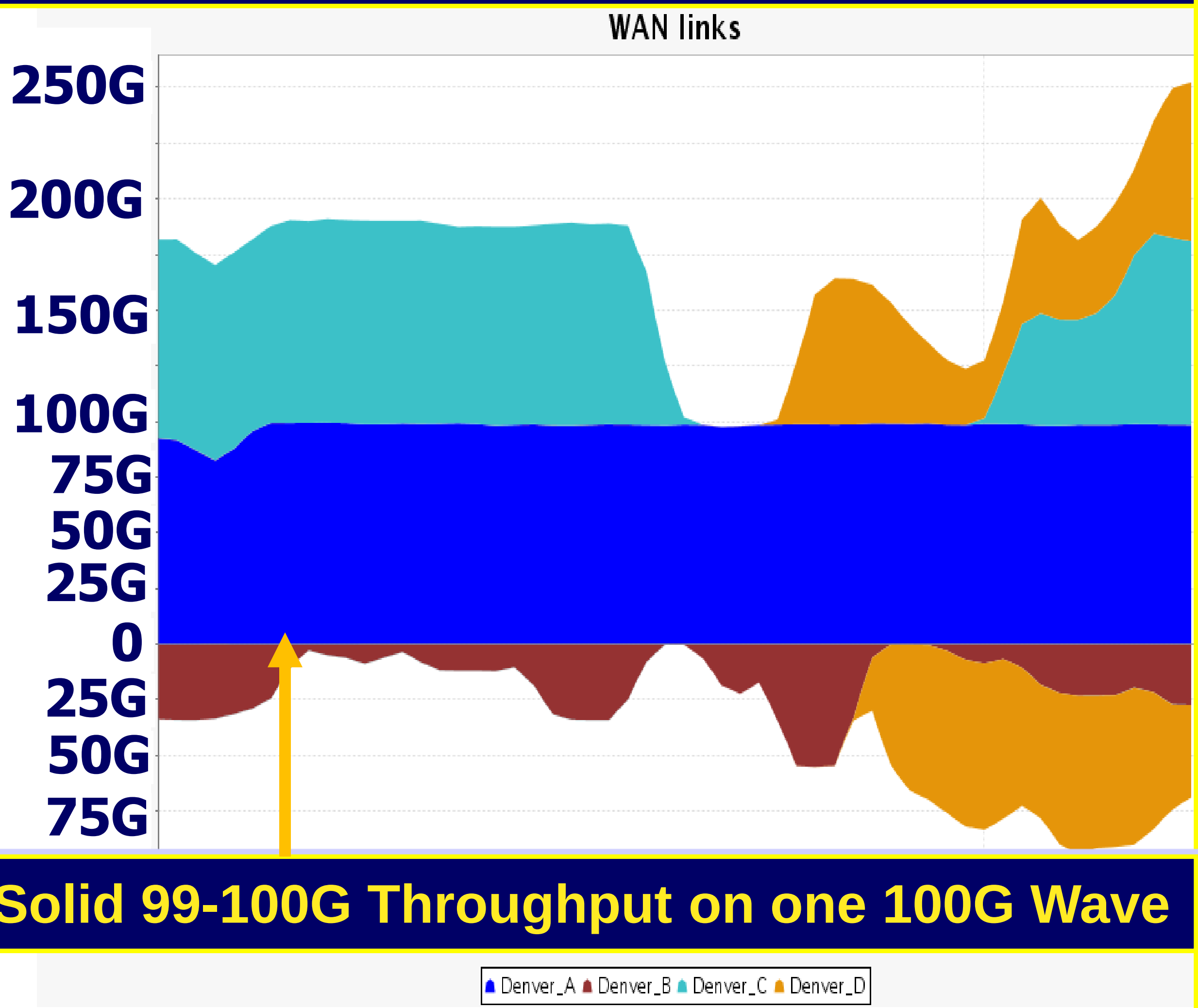


Network partners: SciNet, ESnet, Internet2, ANA-100 CENIC, Starlight, MANLAN, MiLR, SURFNet, RNP, ANSP, AmLight

Caltech Booth at SC13 (Denver)

Wide Area Network Trials Over 4 100G Links

Up to 325G of Wide Area Network Traffic



Solid 99-100G Throughput on one 100G Wave

Including

SC13 to DE-KIT Tier1 on ANA-100
75G Disk-Disk

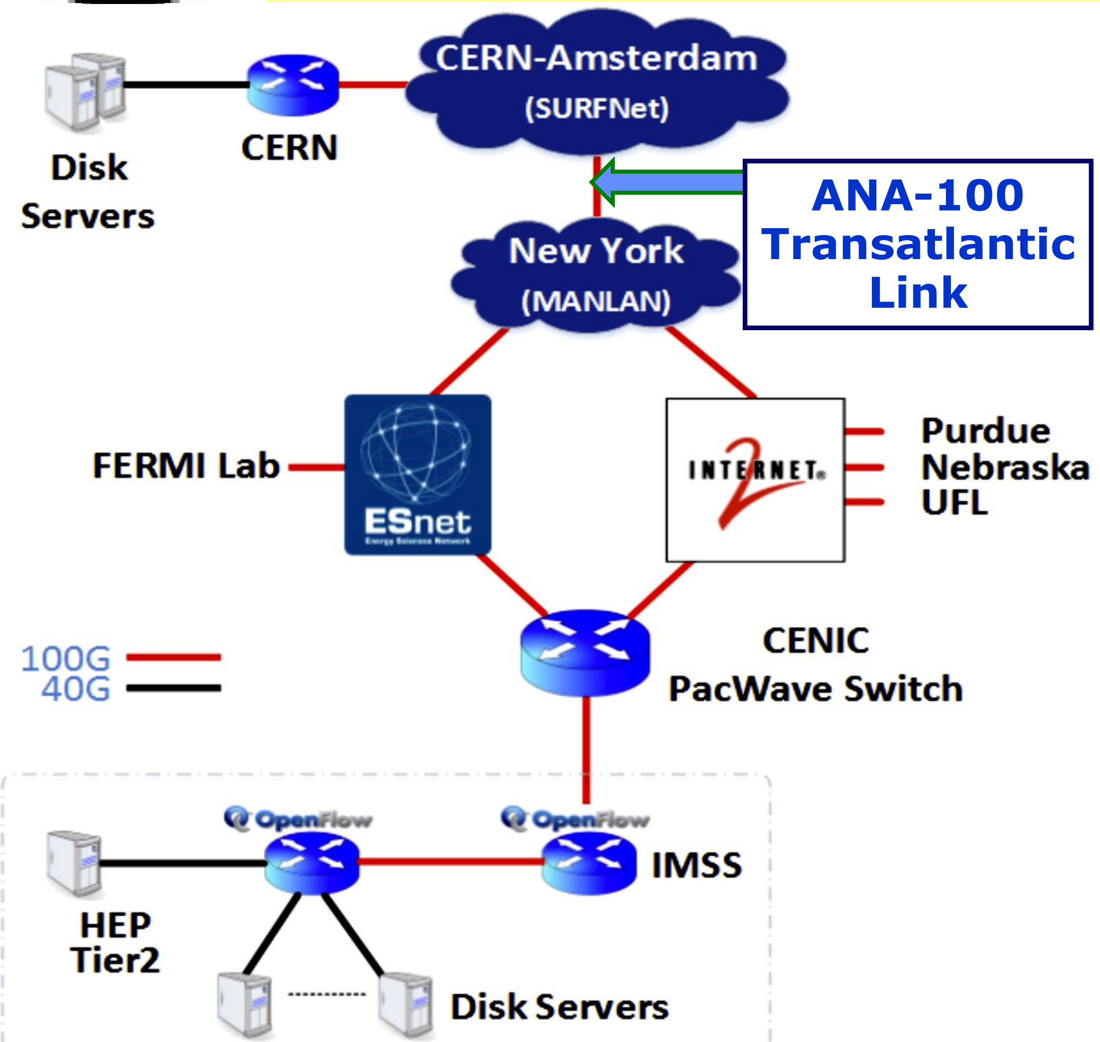
NERSC to SC13 (on ESnet): **90G Disk-Disk**

SC13 to Caltech (on Internet2)
80G Disk to Memory

SC13 to CERN (ESnet)
40G Disk-Disk
75G Memory-Memory

SC13 to BNL (ESnet)
80G Memory-Memory

Feb. 2014: Caltech Connected at 100G to Esnet, Internet2 and CERN via CA Regional Network



Caltech to CERN Sustained Data Transfers at 68Gbps Over 100G TA research link

First 100G TA Trial Direct from a University across the US (ESnet) + the Atlantic

100G Routers and Caltech Link to CENIC funded by NSF CC-NIE campus infrastructure program

Data Transfer Using RFTP (RDMA and FTP): July 2014

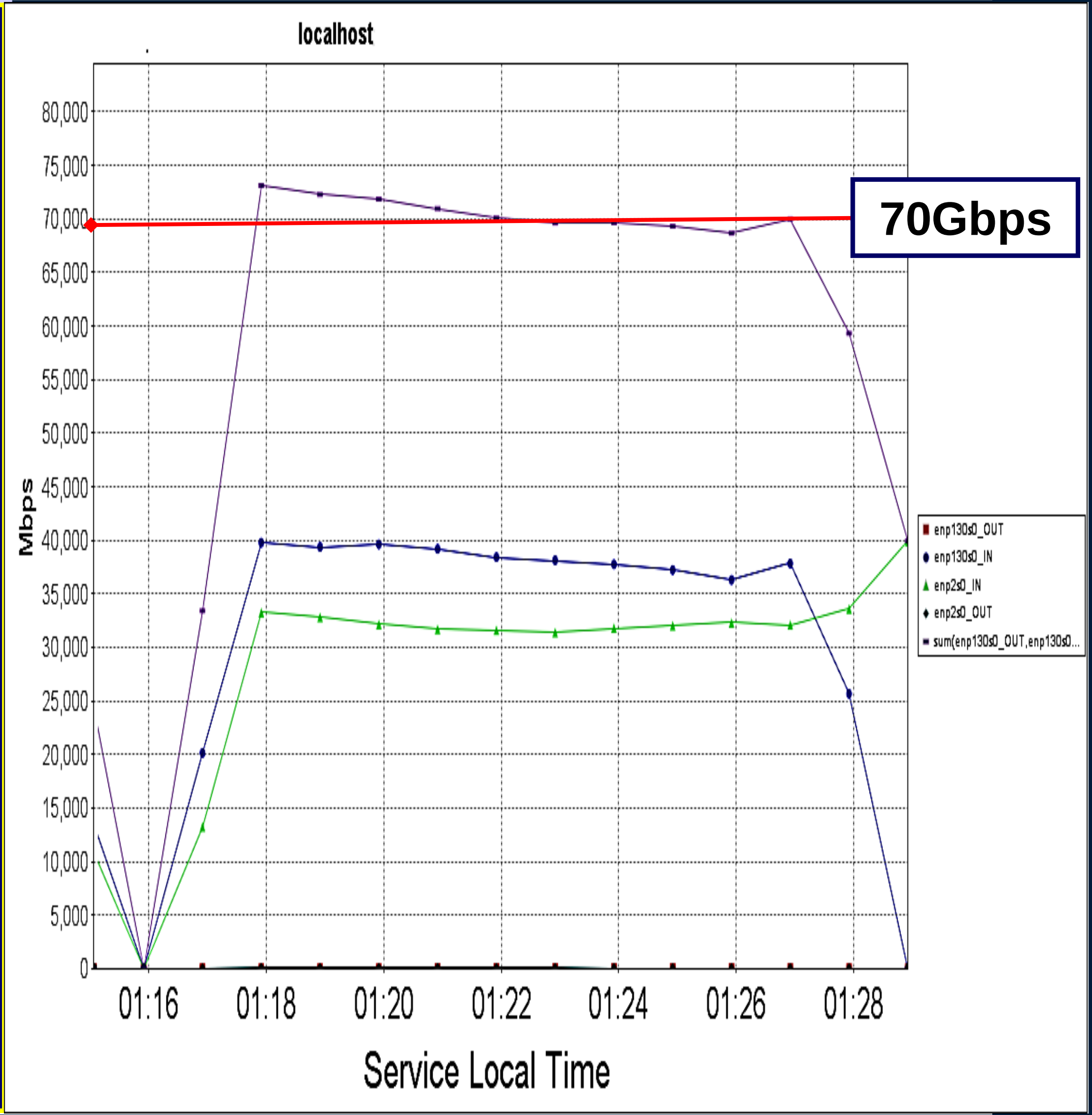
RFTP software in TCP mode transfers multiple source files in parallel

Test Configuration (Server)

- **4 RFTP daemons listening at unique TCP ports**
- **Each RFTP server handles 2 SSD drive mount points (total 8 system mount points)**

Test Configuration (Clients)

- **Total of 8 RFTP clients on two client servers**
- **Two client RFTP processes connect with one RFTPD daemon at destination**



Internet2 Network Map
AL2S Traffic Statistics

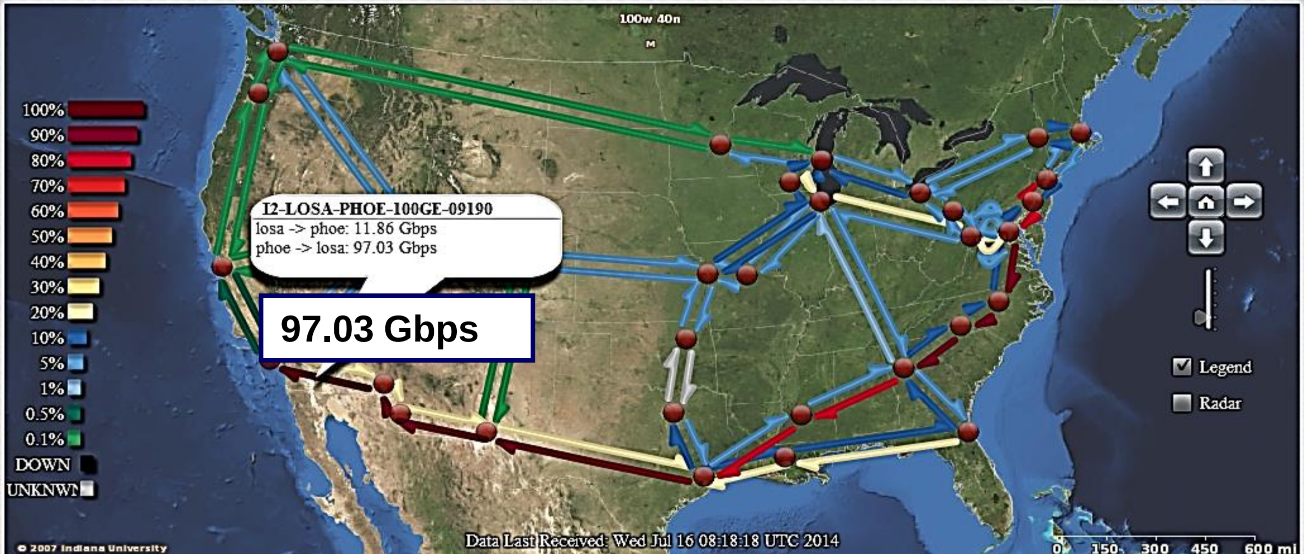
Traffic peak 97.03 Gbps Phoenix - LA observed during these transfers

This is a possible limiting factor on the traffic received at Caltech

Microbursts are often not reported by the monitoring clients

Internet2 Network

GlobalNOC RealTimeAtlas



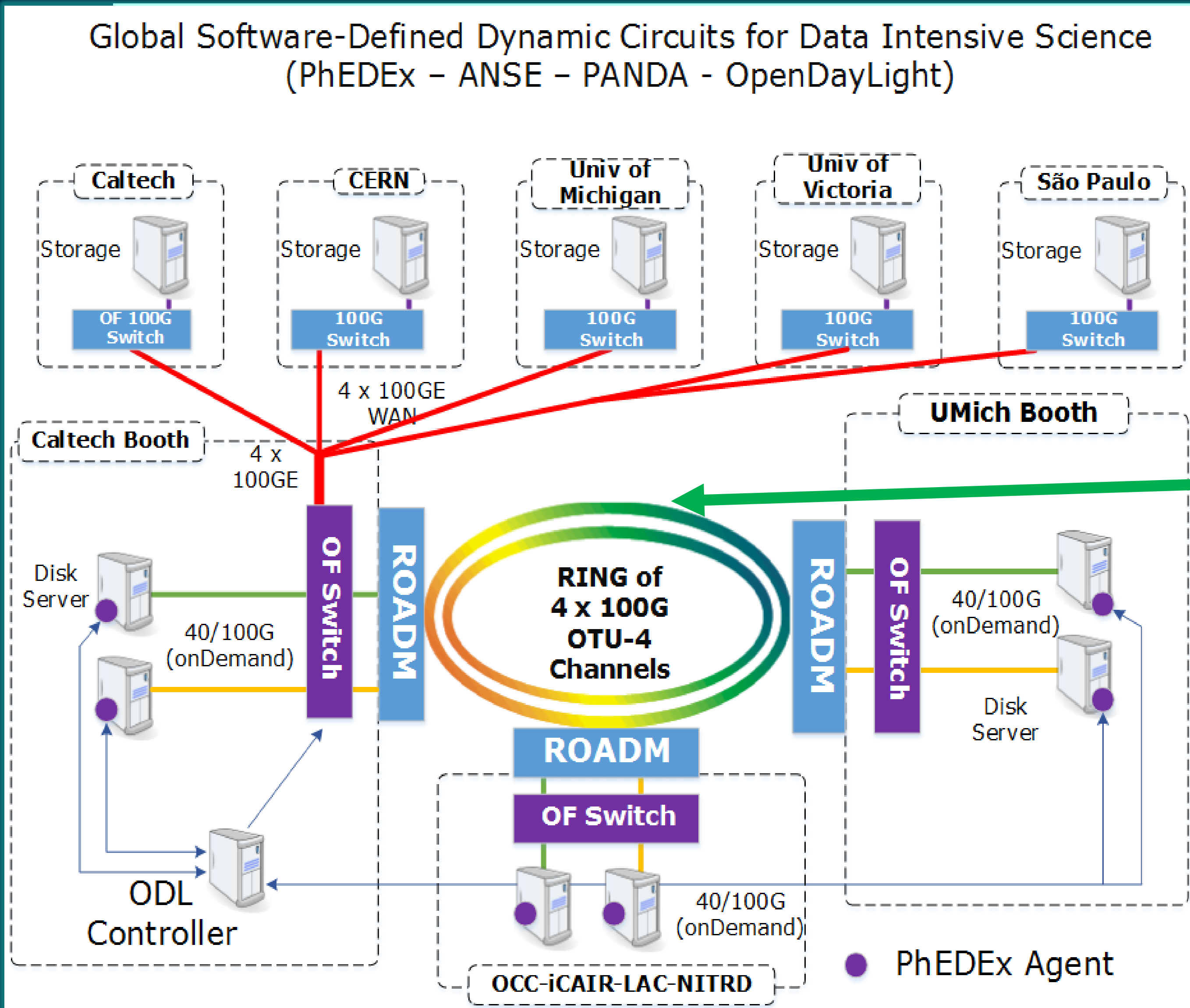
About | Statistics | Circuit Info

Statistics

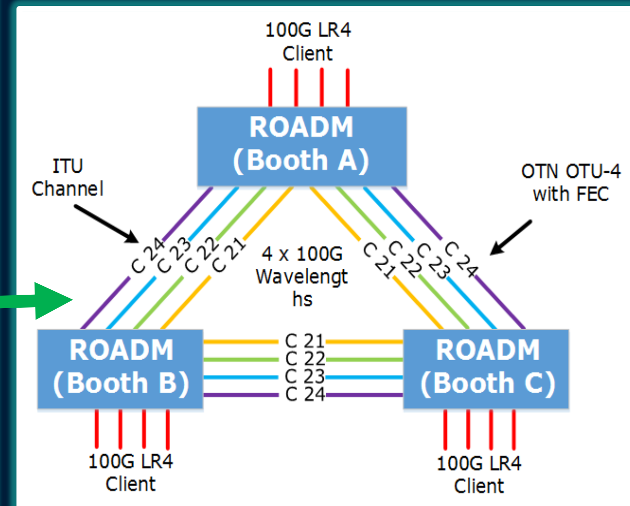
Circuit Name	A -> Z	bits/sec		Packets/sec		Errors/sec	Z -> A	bits/sec		Packets/sec		Errors/sec
I2-RALE-WASH-100GE-08888	wash -> rale	82.56	Gbps	1.92	Mpps	0	rale -> wash	8.84	Gbps	1.15	Mpps	0
I2-ALBA-BOST-100GE-09210	bost -> alba	5.06	Gbps	526.89	Kpps	0	alba -> bost	5.5	Gbps	500.87	Kpps	0
I2-ASHB-WASH-100GE-09106	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

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

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



Terabit/sec Scale Long Range Networking

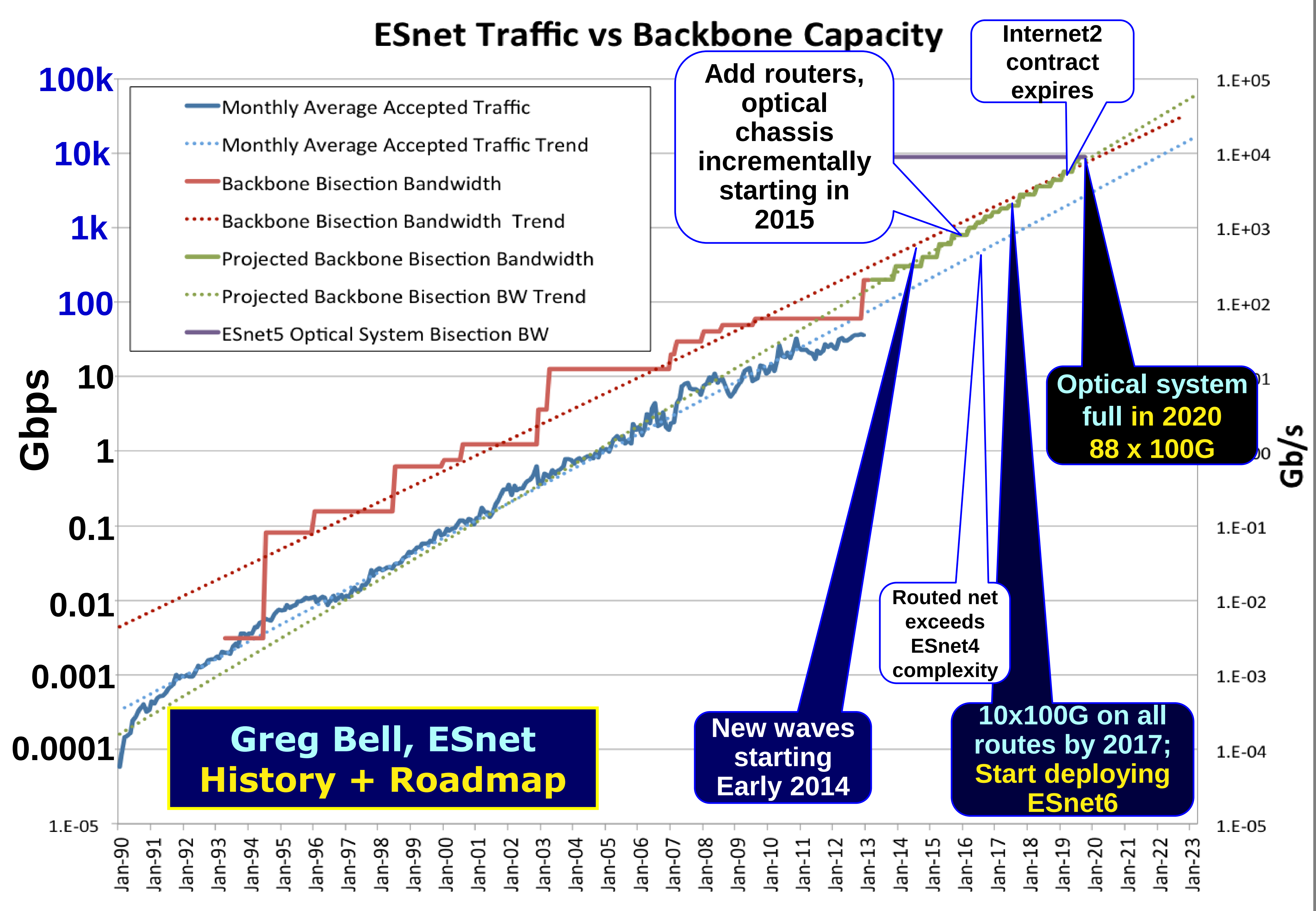


SDN Control of Optical Systems

Caltech HEP and Partners

The Long View: Challenges Ahead

Changes in the Scale and Quality of Data Intensive Networks



ESnet Traffic vs Backbone Capacity
Monthly Average Accepted Traffic
Monthly Average Accepted Traffic Trend
Backbone Bisection Bandwidth
Backbone Bisection Bandwidth Trend
Projected Backbone Bisection Bandwidth
Projected Backbone Bisection BW Trend
ESnet5 Optical System Bisection BW
Gbps
100k
10k
1k
100
10
1
0.1
0.01
0.001
0.0001
1.E-05
Gb/s
1.E+05
1.E+04
1.E+03
1.E+02
1.E-01
1.E-02
1.E-03
1.E-04
1.E-05
Jan-90
Jan-91
Jan-92
Jan-93
Jan-94
Jan-95
Jan-96
Jan-97
Jan-98
Jan-99
Jan-00
Jan-01
Jan-02
Jan-03
Jan-04
Jan-05
Jan-06
Jan-07
Jan-08
Jan-09
Jan-10
Jan-11
Jan-12
Jan-13
Jan-14
Jan-15
Jan-16
Jan-17
Jan-18
Jan-19
Jan-20
Jan-21
Jan-22
Jan-23
Add routers, optical chassis incrementally starting in 2015
Internet2 contract expires
Optical system full in 2020 88 x 100G
Routed net exceeds ESnet4 complexity
New waves starting Early 2014
10x100G on all routes by 2017; Start deploying ESnet6
Greg Bell, ESnet
History + Roadmap

US CMS Tier2 WAN upgrade plans

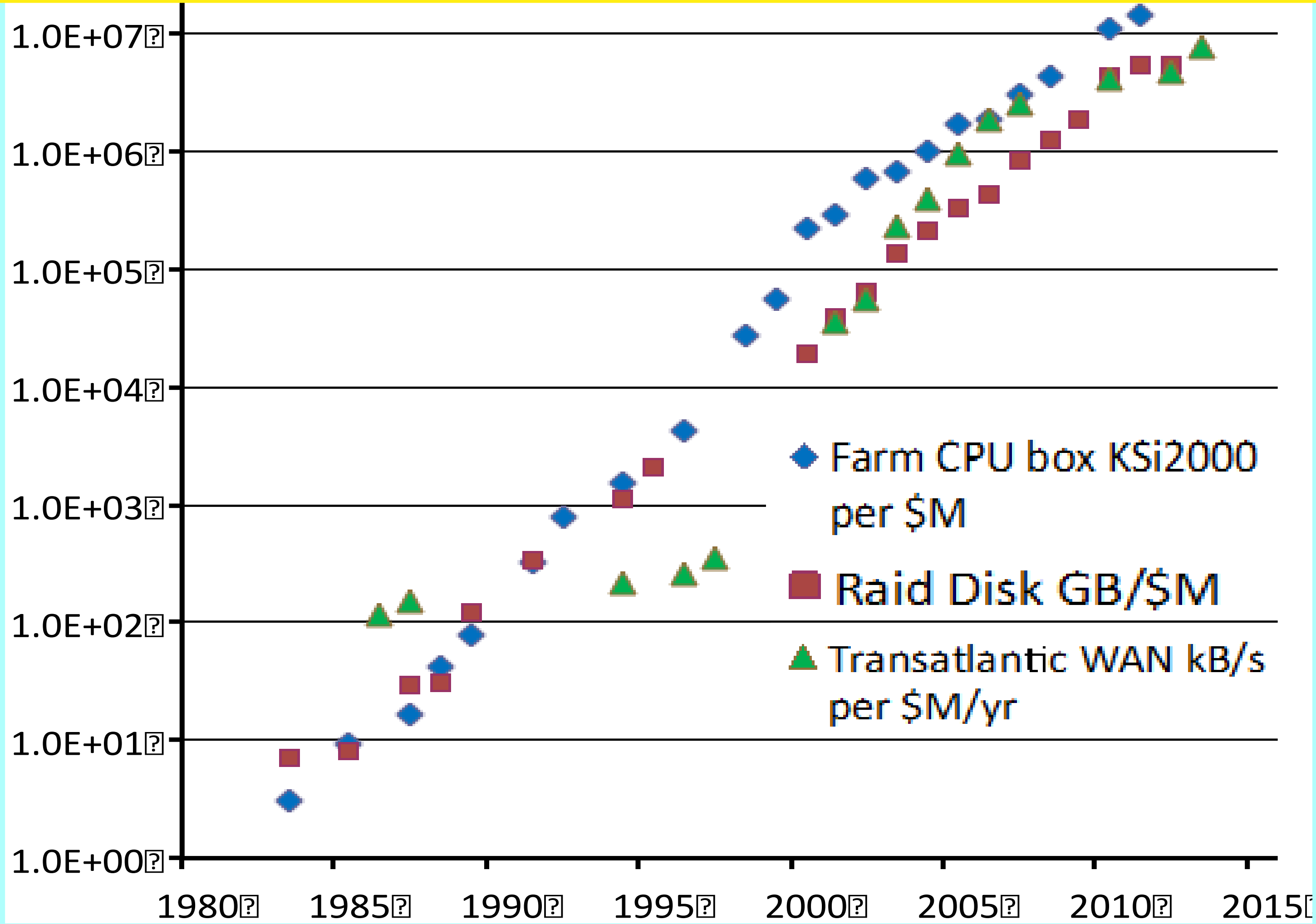
Site	Upgrade plan	LHCONE
Caltech	100 Gbit by March 2014	Yes
Florida	100 Gbit available	Planning to
MIT		
Nebraska	100 Gbit in March 2014	Yes
Purdue	100 Gbit available	No plan
UCSD	100 Gbit in August 2014	"Depends"
Wisconsin	40 Gbit by Summer 2014	No plan

- Note: 1000 T2 batch slots can analyze 2.4 Gbit/s of CMS data
- Needless to say, given the effort and expense needed to upgrade the campus network infrastructure, we want to make the best use of it for scientific productivity

Most US Tier2 Sites at 100G in 2014

Brief Technology History 1983-2014
CPU, Disk, and WAN Bandwidth

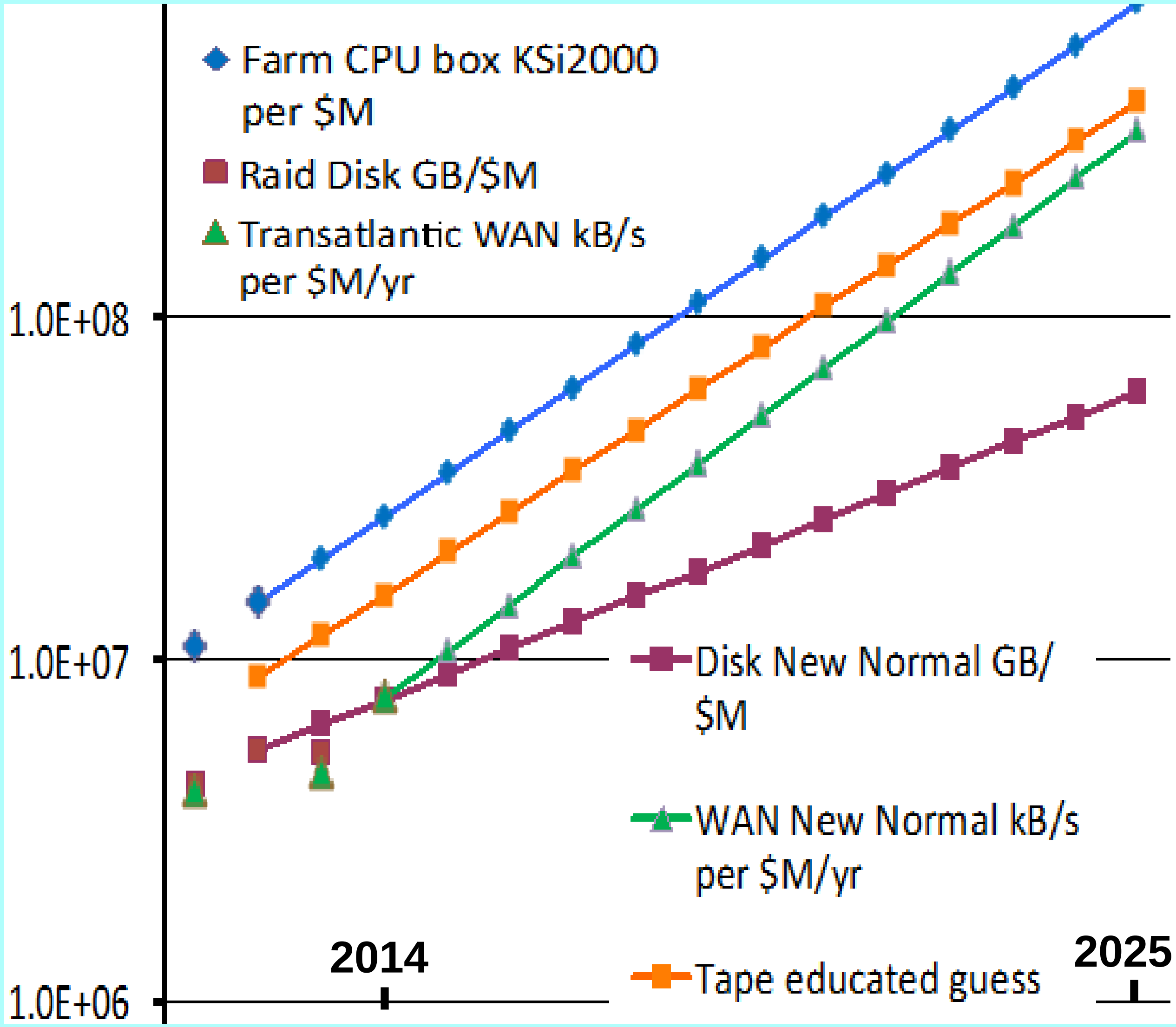
Richard P Mount: Computing in HEP. ICHEP July 9, 2014



“Stuff that Harvey and I bought” – R.P. Mount

Technology Projections to 2025

Performance/Cost Evolution



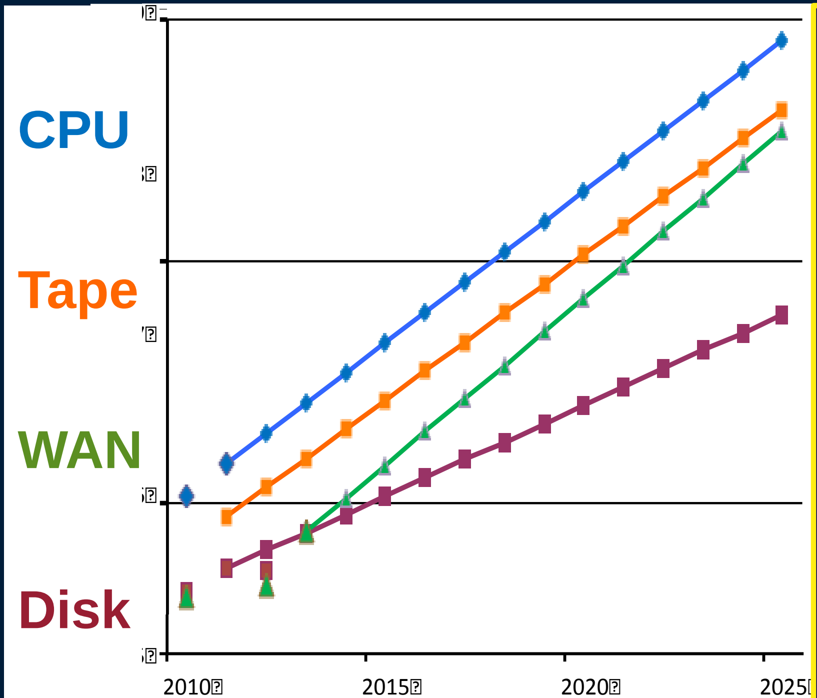
Relative improvement In Performance/Cost Expected in next 10 Years

Technology per unit Cost	Factor
CPU Transistors	10 to 32
Disk Capacity	4 to 8
Tape capacity	8 to 32
WAN bandwidth	10 to 30

Will need to make better use of our resources by HL LHC

Disk Storage might be the biggest issue

Computing Model Outlook for the Next Decade: Minimizing the Storage Needs



Minimize Disk storage needs – options:

- Store less frequently needed data on tape
- Recompute less frequently needed derived data
- Move data rapidly when needed
- Access data remotely (with caching)

Could we automate *ALL* these decisions?

Specify:

- **Lifetime (when can all copies be deleted)**
- **Integrity (tolerable loss/damage probability)**

Leave everything else to "the system" to manage
based on observed and predicted access patterns

R. Mount

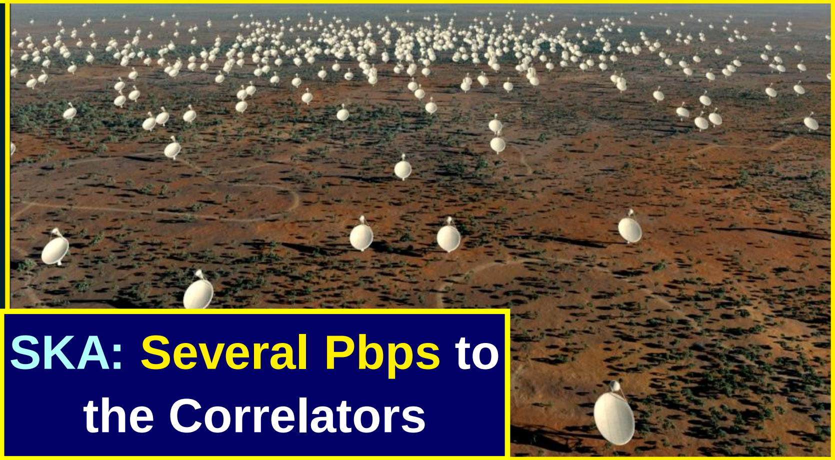
HEP Energy Frontier Computing

Decadal Retrospective and Outlook for 2020(+)

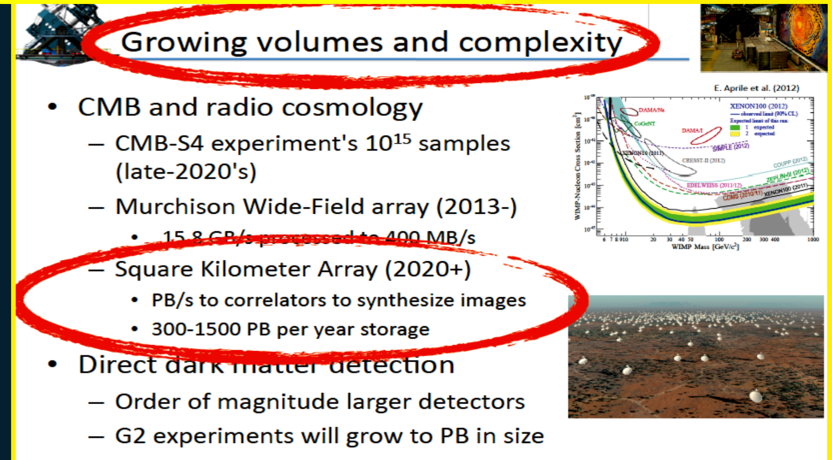
- **Resources & Challenges Grow at Different Rates Compare Tevatron Vs LHC (2003-12)**
  - **Computing capacity/experiment: 30+ X**
  - **Storage capacity: 100-200 X**
  - **Data served per day: 400 X**
  - **WAN Capacity to Host Lab *100 X***
  - **TA Network Transfers Per Day *100 X***
- **Challenge: 100+ X the storage (tens of EB) unlikely to be affordable**
- **Need to better use the technology**
  - **An agile architecture exploiting globally distributed clouds, grids, specialized (e.g. GPU) & opportunistic resources**
  - **A Services System that provisions all of it, moves the data more flexibly and dynamically, and behaves coherently;**
- **Co-scheduling network, CPU and storage**

Snowmass Computing Frontier Sessions

Challenges Shared by Sky Survey, Dark Matter and CMB Experiments.
SKA: 300 – 1500 Petabyes per Year



SKA: Several Pbps to the Correlators



Research and Innovation Agenda

Core Question and a Promising Approach

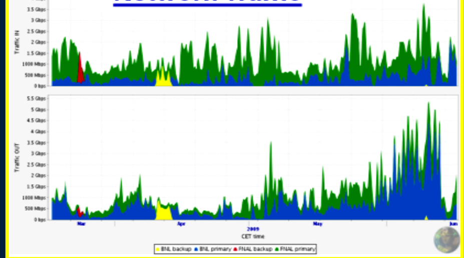
- A Core question: **Can global research networks evolve: into adaptive systems** that respond rapidly to the needs: of HEP and other data intensive sciences ?
- **Examples do exist,** with smaller (but still very large) scope

MonALISA →

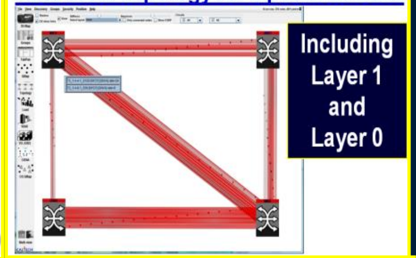
- Pervasive, autonomous agents architecture: deals with, reduces complexity
- **Software Defined Networking is a promising direction: Open services**
  - **Enabling great innovation through** virtualization, deep programmability, and integration
- **Requires talented system architects** with a deep appreciation of networks and their potential

MonALISA [Legrand, Voicu]

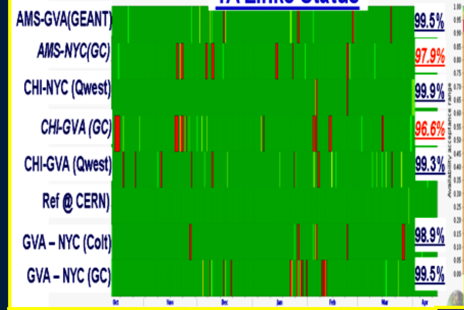
Network Traffic



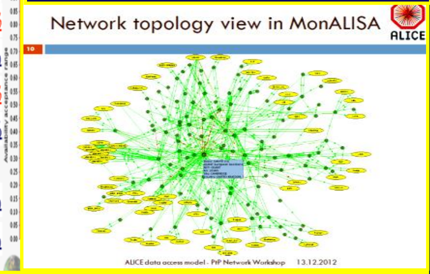
Real Time Topology for Optical Circuits



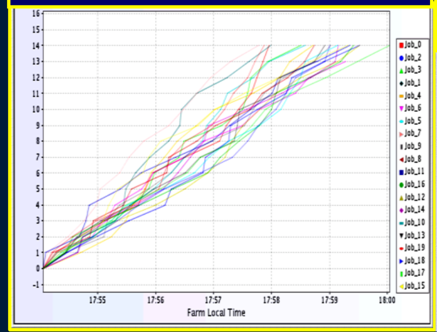
TA Links Status



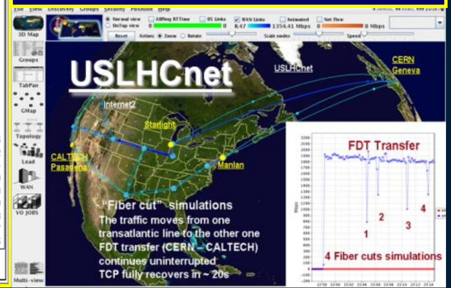
Grid Net Topology



Grid Job Lifelines



Automated Transfers on Dynamic Networks



Raw Bandwidth Projections

- Data from ESnet requirements reviews: http://www.es.net/requirements
- Rolled up by DOE program office
- Units are Gigabits per second (Gbps)

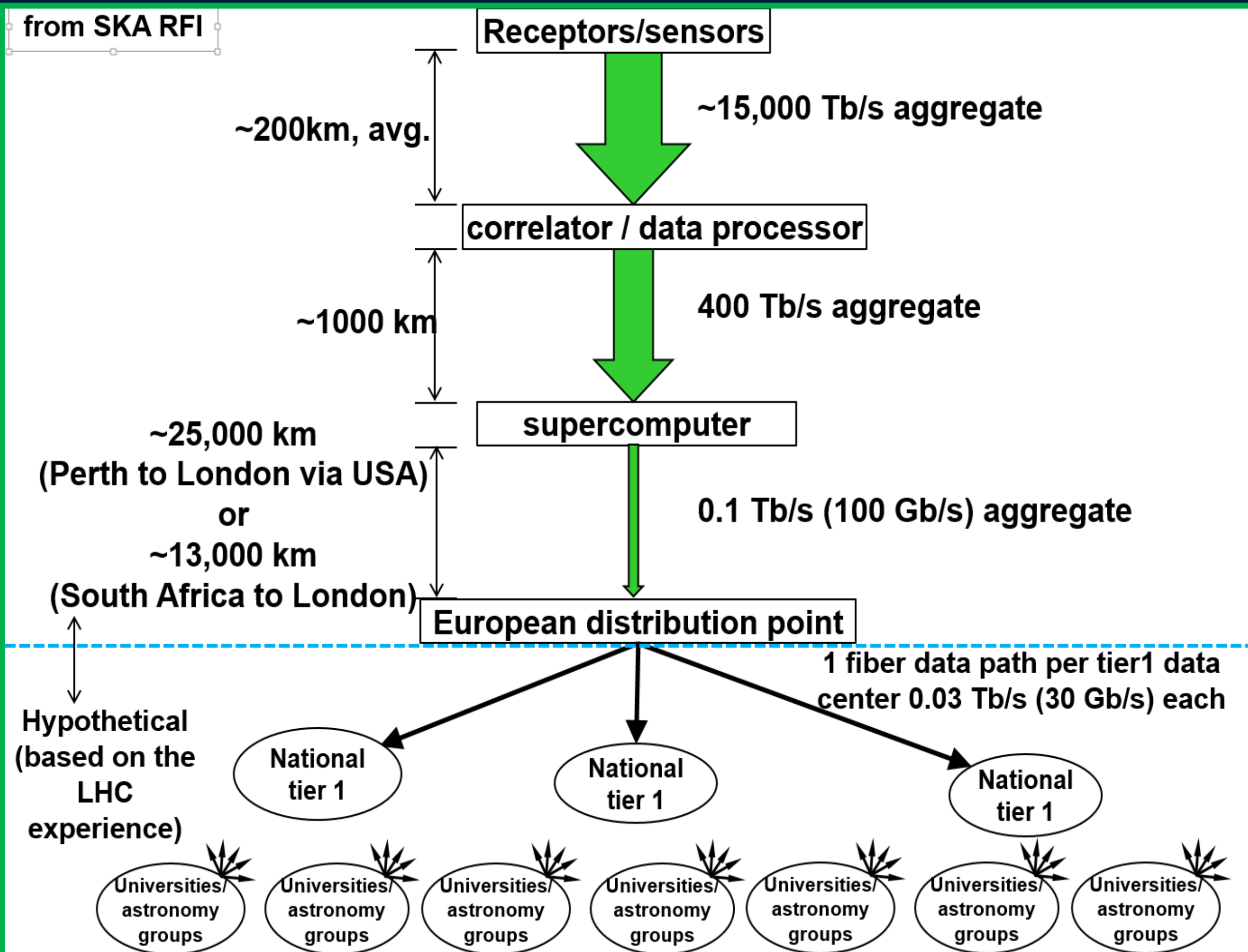
Courtesy Eli Dart

	Basic Energy Sciences (2010)	Fusion Energy Sciences (2011)	Nuclear Physics (2011)	Biological and Environmental Research (2012)	Advanced Scientific Computing Research (2012)	High Energy Physics (2013 est.)	Totals
0-2 yrs.	31	3.2	11	7	12	69	133
2-5 yrs.	178	29	27	106	222	314	875
5+ yrs.	2740	55	66	1430	2300	760	7340

- Timelines mean different things
  - 0-2 years – this is in current budget projections
  - 2-5 years – this is the current technological paradigm or within currently-planned change envelope
  - 5+ years – big events on the horizon (new facilities, facility upgrades, anticipated disruptive technology)
- Many different workflows and classes of workflows present

A Network Centric View of SKA

Bill Johnston



Sensor	Gbps per sensor	Sensors	Petabits/s (10^6 Gbps) Total
Phased Array Feeds	930		
Wideband Single Pixel Feeds	216	130	0.03
SPF with PAFs	1,146	2270	2.6
AA-low (Aperture Arrays)	33,440	250	8.4
AA-mid	16,800	250	4.2
Total		2900	15.2

Stored Data Product Estimates: 300 – 1500 Petabyes/Yr

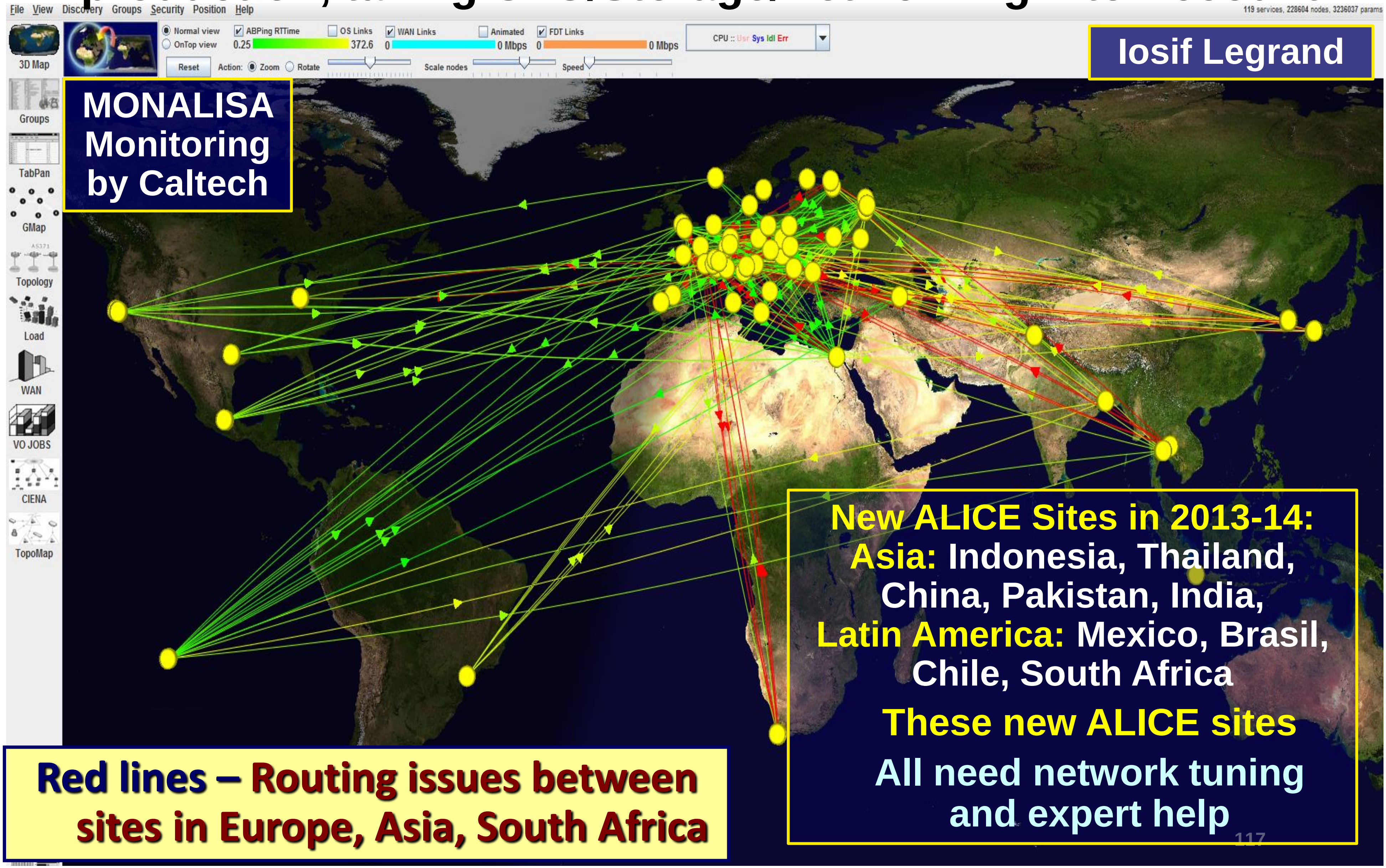
Massive Online and Offline Flows: Analogous to ALICE Triggerless "Flow Through" DAQ System

SKA: A Massive Online System

Once past the Supercomputer, Data flow Might be of same order as the LHC

Need to re-evaluate lower part of the diagram with guesses at 2025 technology

ALICE: MonALISA drives worldwide offline production, taking CPU/Storage/Networking into Account



Iosif Legrand

MONALISA Monitoring by Caltech

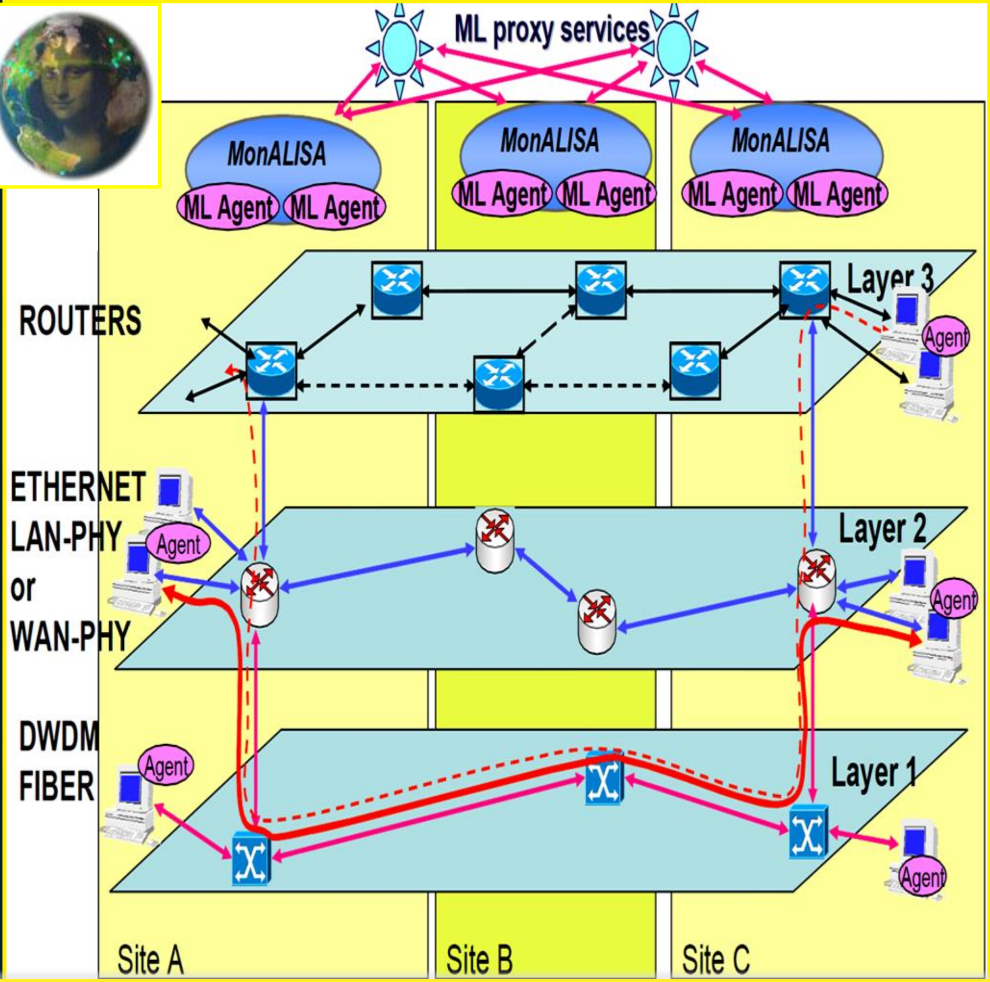
New ALICE Sites in 2013-14:
Asia: Indonesia, Thailand, China, Pakistan, India,
Latin America: Mexico, Brasil, Chile, South Africa

These new ALICE sites

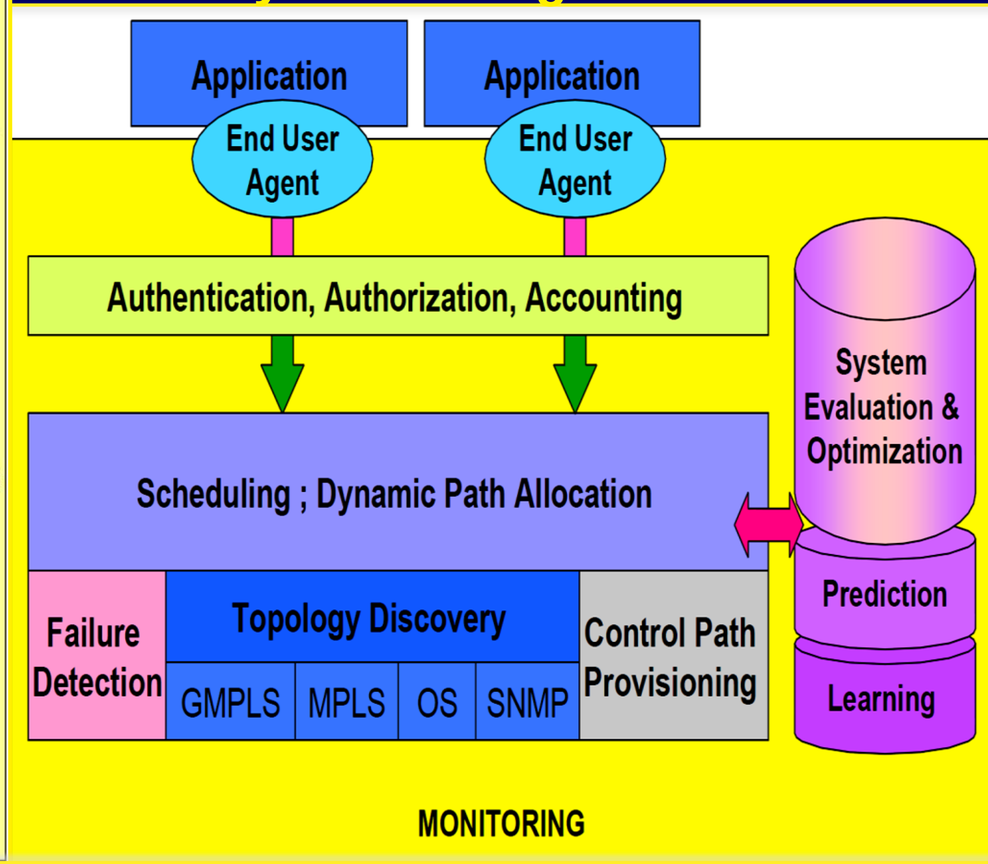
All need network tuning and expert help

Red lines – Routing issues between sites in Europe, Asia, South Africa

VINCI: Virtual Intelligent Networks for Computing Infrastructures



Core Concepts and Real Time System Design: 2005-8



http://monalisa.caltech.edu

VINCI (CHEP06, Mumbai)

http://indico.cern.ch/contributionDisplay.py?sessionId=6&contribId=350&confId=048

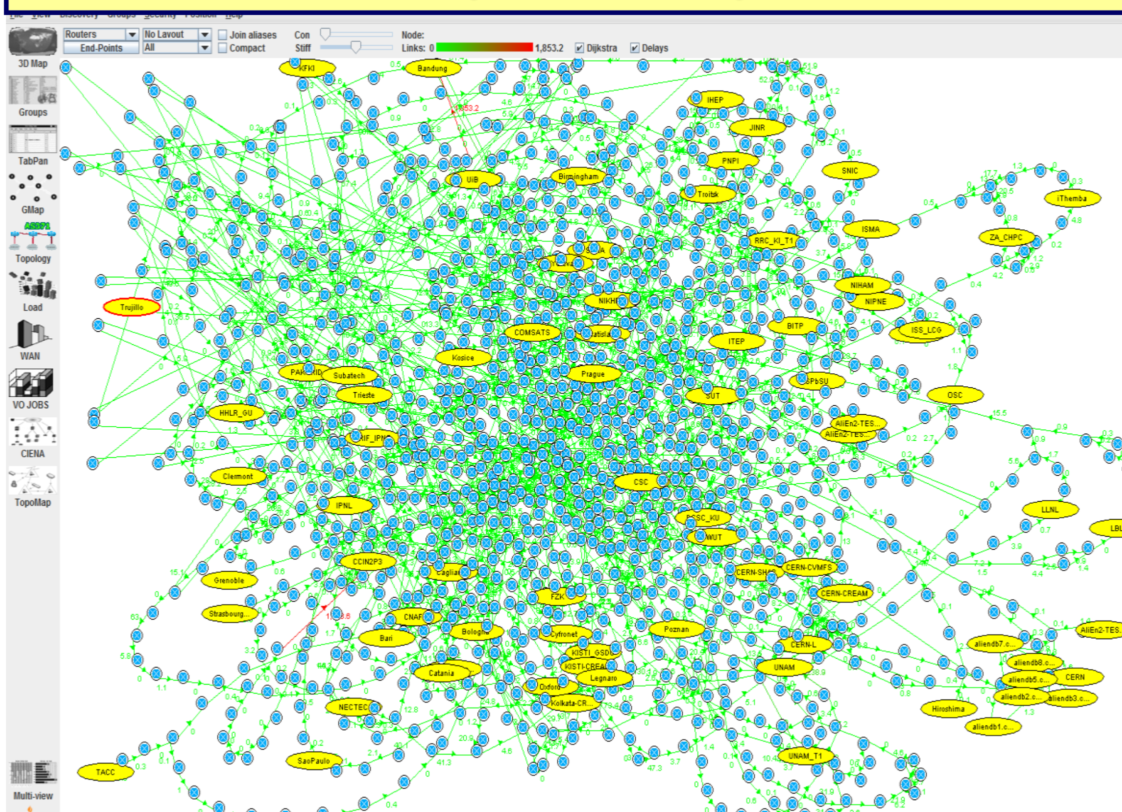
ML Monitoring Network Topology, Latency and Routers in ALICE

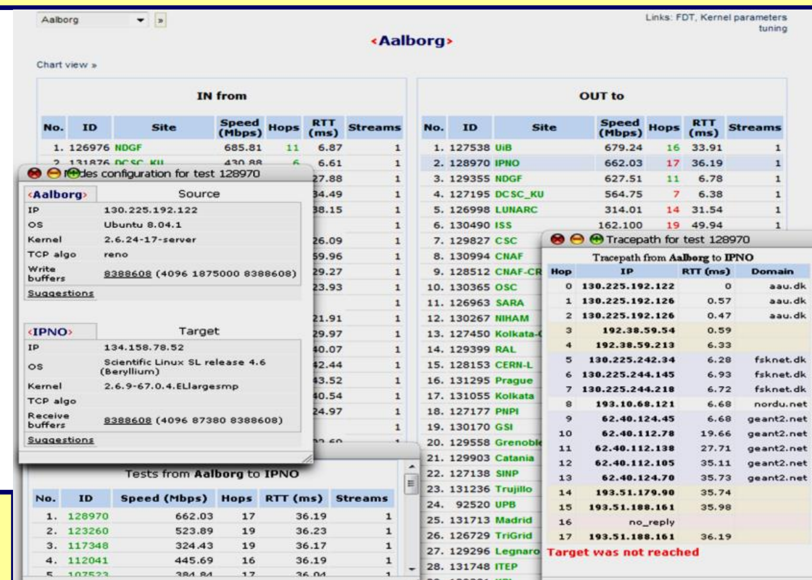
85 x 85 Real Time Site-to-Site Matrix

Proposed: move to all xrootd servers (700 x 700)

Plus:

- **Path monitoring, analysis and identification of routing loops or problem hops**
- **End host monitoring and changes of kernel parameters to improve throughput where needed**



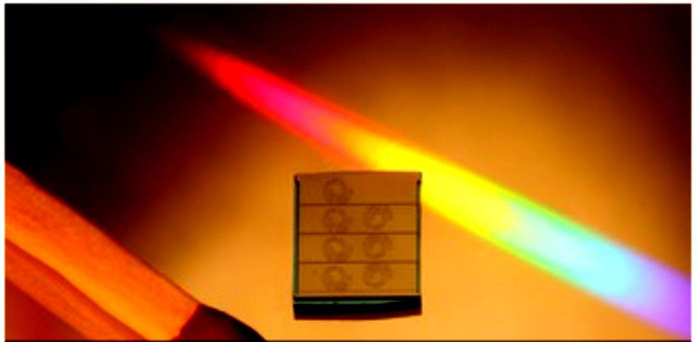


Real Time Topology Discovery & Display

Beyond (or At) Five Years

Physics will find a way

Using light for faster data transmissio

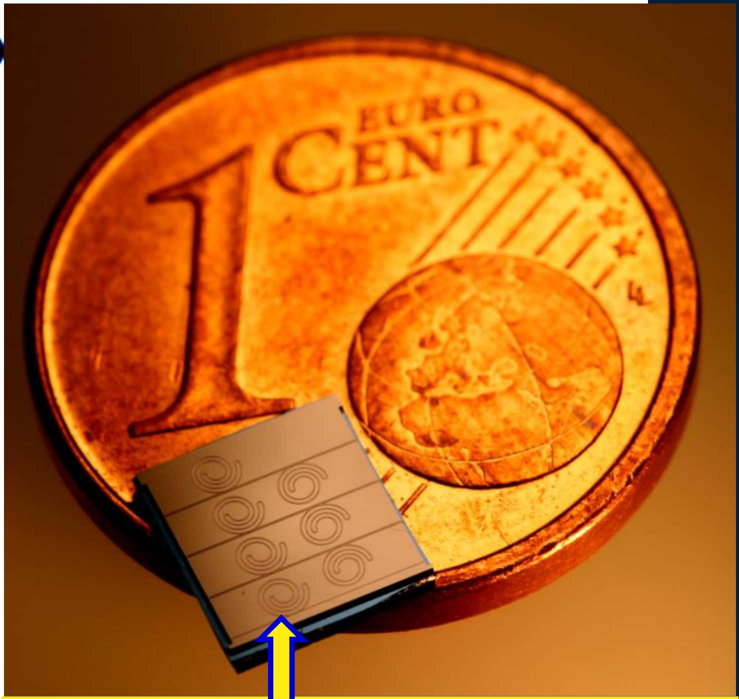


An optical microresonator © 2014 EPFL/Tobias Kippenberg

20.04.14 **- Scientists from EPFL and KIT have achieved data transmissions on a terabit scale with a single laser light frequency using miniaturized optical frequency combs. The findings open the way for using this system in future high-speed communication systems.**

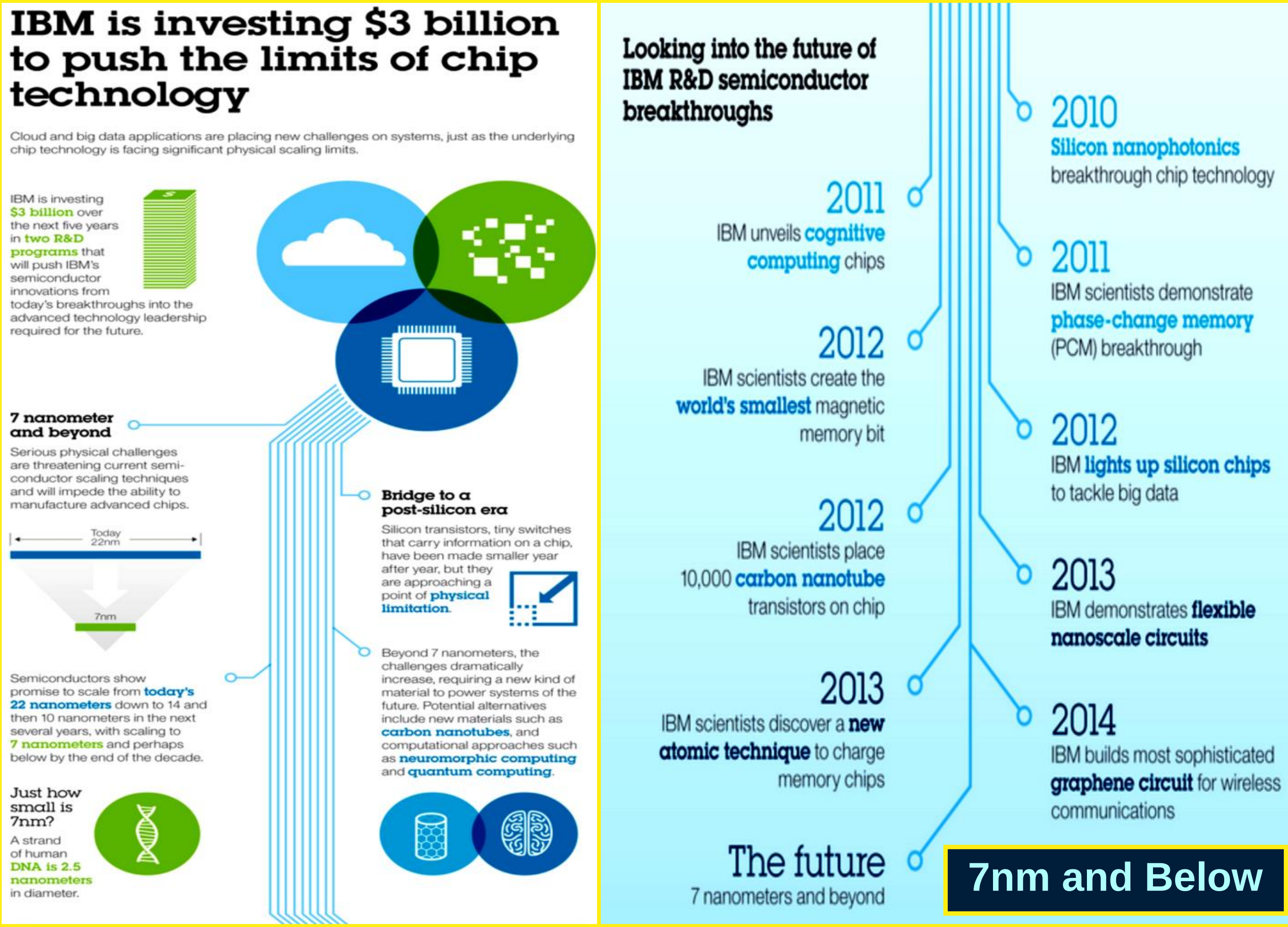
A continuous laser light is made of a single frequency, i.e. a single color. But that single frequency can be divided into separate lines of equal distance, which is referred to as an "optical frequency comb". Practically speaking, that could allow the simultaneous flow of data in optical cables, which could dramatically increase today's speed of data transmission. Optical frequency combs can transmit data on hundreds of separate wavelength channels, meaning that they can overcome transmission bottlenecks in data centers and communication networks. Publishing in *Nature Photonics*, scientists from EPFL and the Karlsruhe Institute of Technology (KIT) have shown that optical frequency combs can achieve a 1.44 Terabit/sec data transmission across a distance of up to 300 Km.

When one light frequency from a laser is fed into a device called an optical microresonator, it is possible to convert it into an "optical frequency comb": a series of densely-spaced spectral lines whose in-between distances are identical and known. These frequencies represent the original light frequency fed into the microresonator, along with hundreds of new frequencies.



- **Microresonators**
- **1 λ → Optical Frequency Comb**
- **1.44 Tbps over 300 km in 20 Comb lines**

Where Do We Go from Here ?
7nm and Below



As new technologies take hold in 2018-25

- Nanophotonics
- Plasmonics
- Silicon Photonics
- Graphene and other 2D materials

With higher density much higher speeds and less energy

The outlook for ICT capabilities will fundamentally change

We should continue to envisage and realize the systems of the future for the next round of science discoveries, and for society

See Richard Feynman's Nantechnology Lecture: "There's Plenty of Room at the Bottom"
https://www.youtube.com/watch?v=4eRCygdW--c

Astrophysics and Other Fields

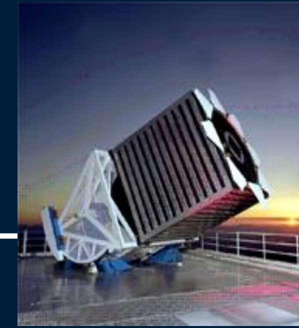
Growing to Massive Data Flows

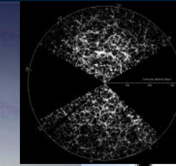
HEP is Not Alone

Collaborations Since 1998

Astro: Sloan *Digital* Sky Survey

"The Cosmic Genome Project" 1992-2008



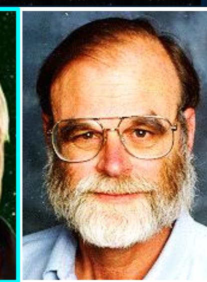


- ❒ Data is public: 5 Terapixels of sky.
  - ❒ 10 TB of raw data → 400TB processed
  - ❒ Originally 0.5 TB catalogs → *> 35TB in the end*
  - ❒ *Now SDSS-3* Served from Johns Hopkins

✸ Skyserver: *Prototype* of 21st Century Data Access

- ❒ 1.4B web hits in 12 years: 4,000,000 distinct users vs. 15k Astronomers
- ❒ Emergence of the "Internet Astronomer"
- ❒ Collaborative server-side analysis by 7K astronomers

Courtesy Alex Szalay JHU



Jim Gray

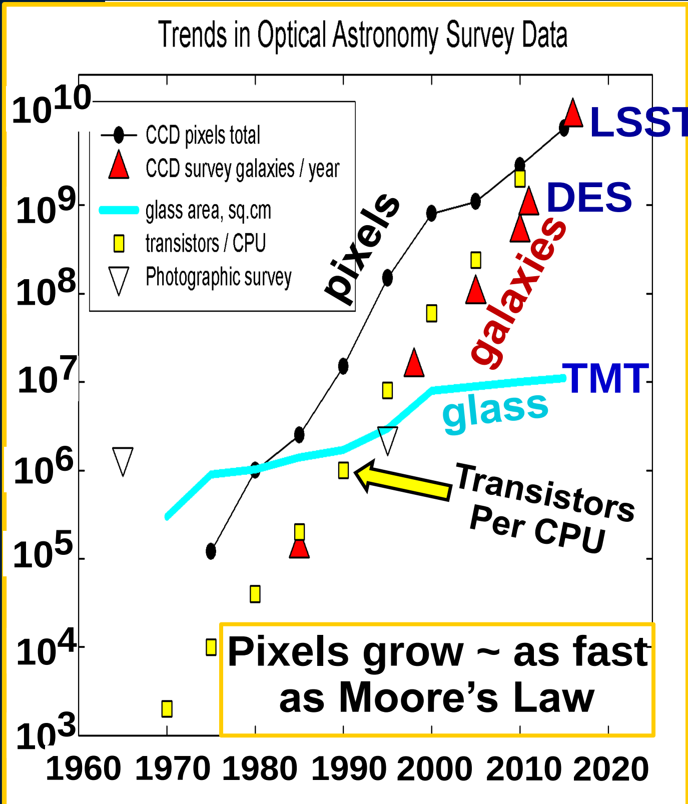
Galaxy Zoo: Crowdsourcing Science (Since 2007)

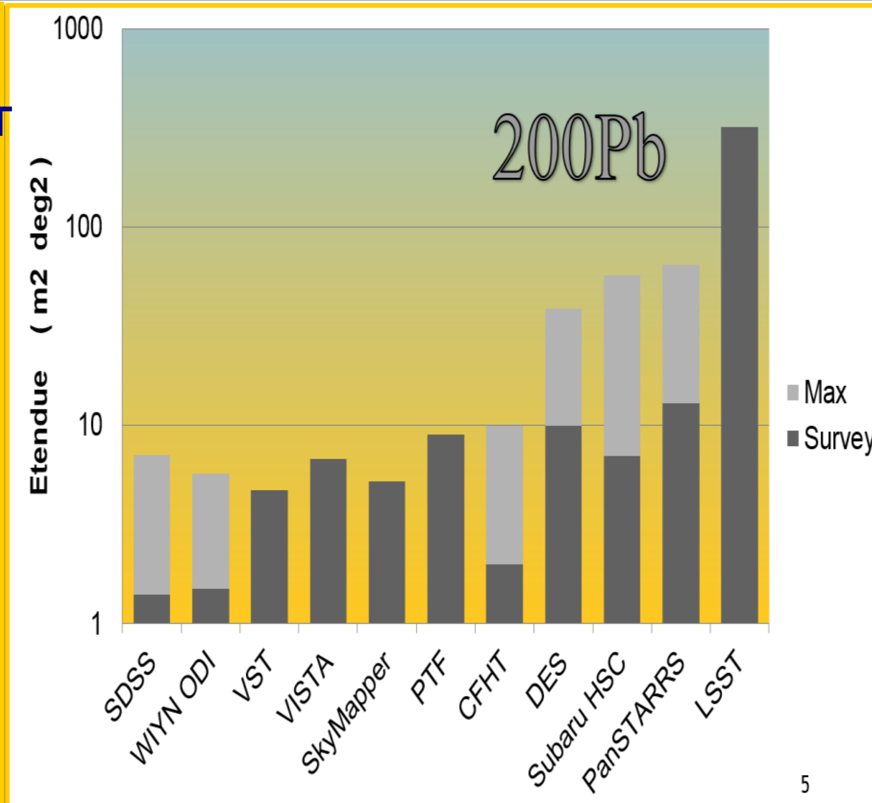
- *It all started back in July 2007, with a data set made up of a million galaxies imaged by SDSS. With so many galaxies, we'd assumed it would take years for visitors to the site to work through them all, but within 24 hours of launch we were stunned to be receiving almost 70,000 classifications an hour. In the end, more than 50 million classifications were received during its first year, contributed by >150,000 people. Now in its 4th Generation: SDSS, Hubble, CANDELS…*

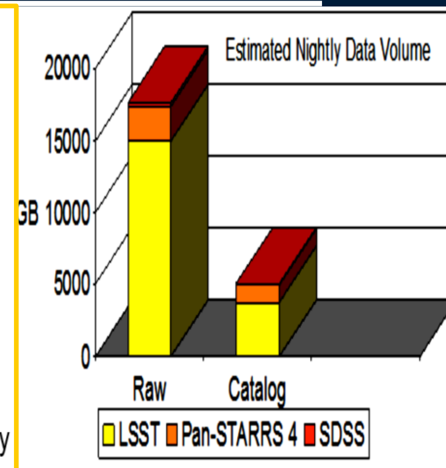
Hanny's Voorwerp

Unprecented Data Volumes

as Sky Surveys Evolve







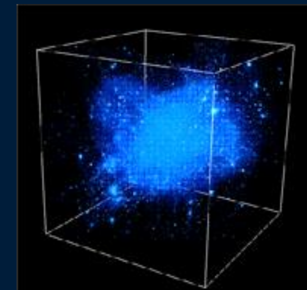
Pixels Vs. Telescopes
(Glass Tops Out)
CCD pixels and Survey
Galaxies/Year:
to 10^{10} by 2020

SDSS was delayed to 2000:
Data Volume grew 70X:
0.0005 to 0.035 Pbytes (all Good)
LSST volume is expected
to be *6000X greater [200 PB]*

Courtesy
Alex Szalay

"Sociology": Structural, Non-Incremental Changes in Experimental Science

- **Multi-faceted challenges:**
  - New computational tools and strategies
  - … Not just statistics, not just computer science, Not just astronomy, not just genomics…



- Science is moving increasingly from hypothesis-driven to (also) data-driven discoveries

Courtesy
Alex Szalay

- **Broad sociological changes:**
  Convergence of Physical and Life Sciences
  - Data collection in ever larger collaborations
  - Virtual Observatories: CERN, VAO, NCBI, NEON, OOI,…
  - Decoupled Analysis using archived data: by smaller groups throughout the world
  - Emergence of the citizen/internet scientist
- **Need to start training the next generations**
  - Π-shaped vs I-shaped people: Early involvement in "Computational (and network) thinking" as well as discipline science

Broad Workflow Classes (Examples)

Large instruments, large collaborations (e.g. LHC)

- Well-organized, large number of institutions
- Broad data distribution to many locations
- Able to adopt common practices and tools
- Need specialized infrastructure to enhance productivity (e.g. LHCONE)
- Always-on use of high-speed data services (all major sites rapidly moving to 100G)

HPC-Centric

- Simulations are a primary driver
- Data movement to secondary analysis
- Data movement between centers (data follows user allocations)
- Support for instruments (e.g. cosmology, fusion)
- Routine movement of data sets 10TB to 1PB in size
- Smaller groups need easy to use tools

Tightly-coupled multi-facility

- Experiment → analysis → decision → experiment
- Data set transaction time is more important than data rate
  - ~10GB in 2 minutes (Fusion); ~6GB in 2 seconds (LSST)

Courtesy Eli Dart