Advanced Networking for HEP, Research and Education in the LHC Era

Harvey B Newman
California Institute of Technology

CHEP2013, Amsterdam
October 17, 2013

Discovery of a Higgs Like Boson
July 4, 2012

Theory : 1964

LHC + Experiments Concept: 1984

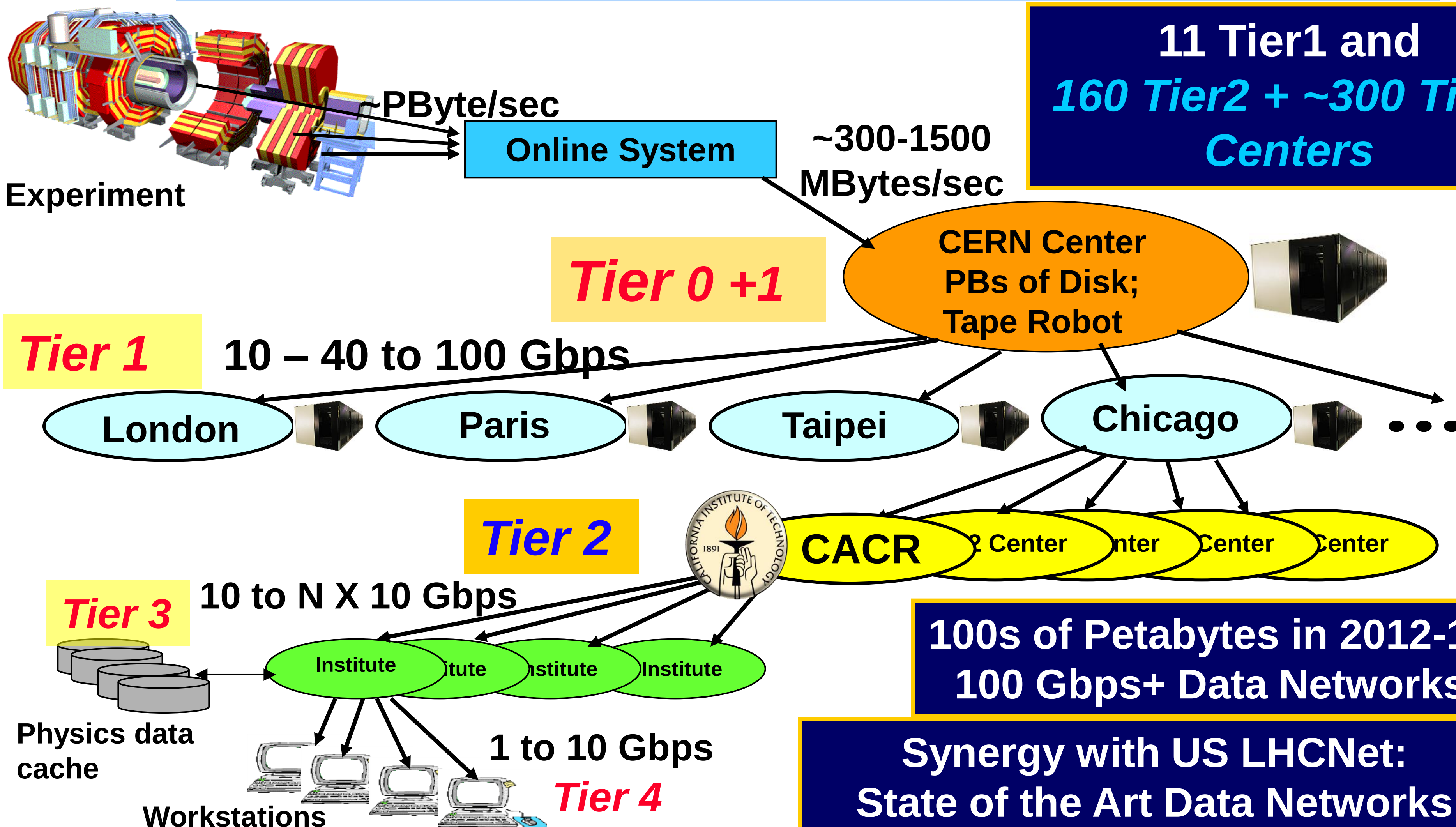
Construction: 2001

Operation: 2009-12

A billion people watched

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

LHC Data Grid Hierarchy
A Worldwide System



11 Tier1 and
160 Tier2 + ~300 Tier3 Centers

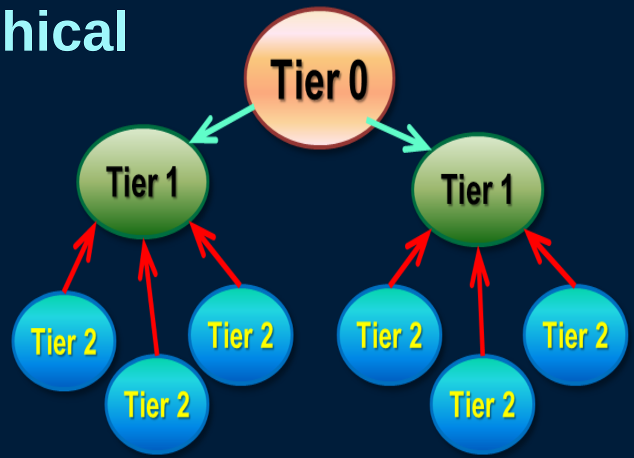
100s of Petabytes in 2012-13
100 Gbps+ Data Networks

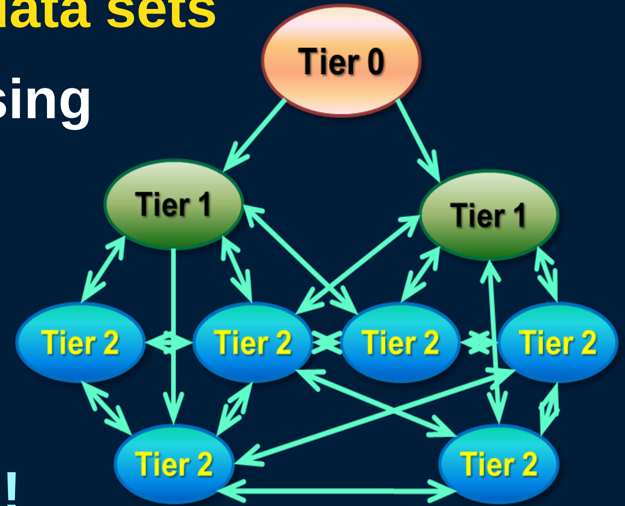
Synergy with US LHCNet:
State of the Art Data Networks

A Global Dynamic System
A New Generation of Networks: LHCONE, ANSE

LHC Computing Model Evolution

- The original MONARC model was (largely) hierarchical
- Main evolutions introduced since 2010:
  - 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 Tier1s and Tier2s in other regions
    - Combined 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
- Federated Data Systems: FAX, PhEDEx, Alien
- Variations by experiment; but a common element is: Increased reliance on network performance !

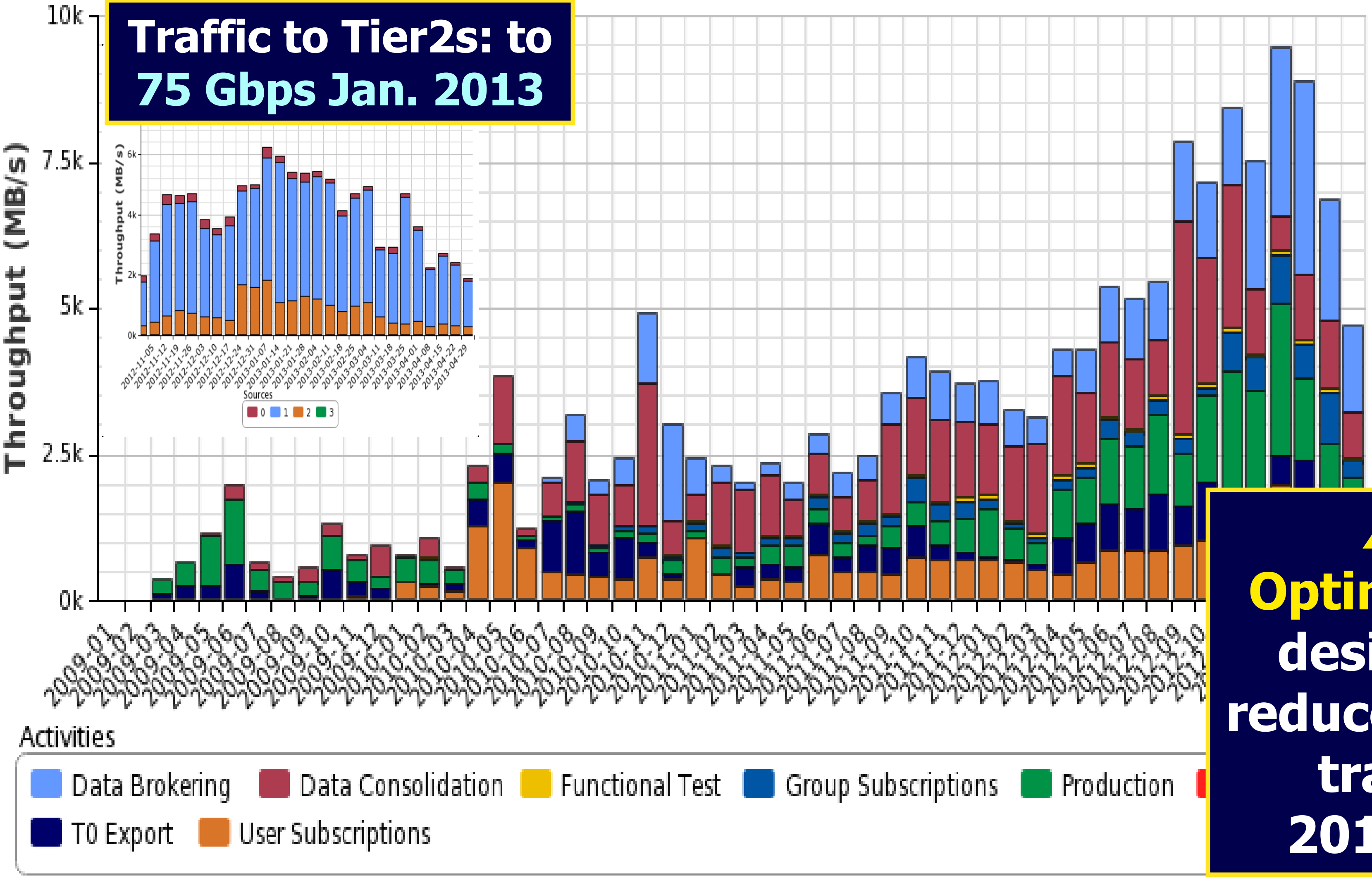




ATLAS Data Flow: 2009- April 2013

2012-13: >50 Gbps Average, 112 Gbps Peak
171 Petabytes Transferred During 2012

Traffic to Tier2s: to 75 Gbps Jan. 2013



2012 Versus 2011: +70% Average +180% Peak

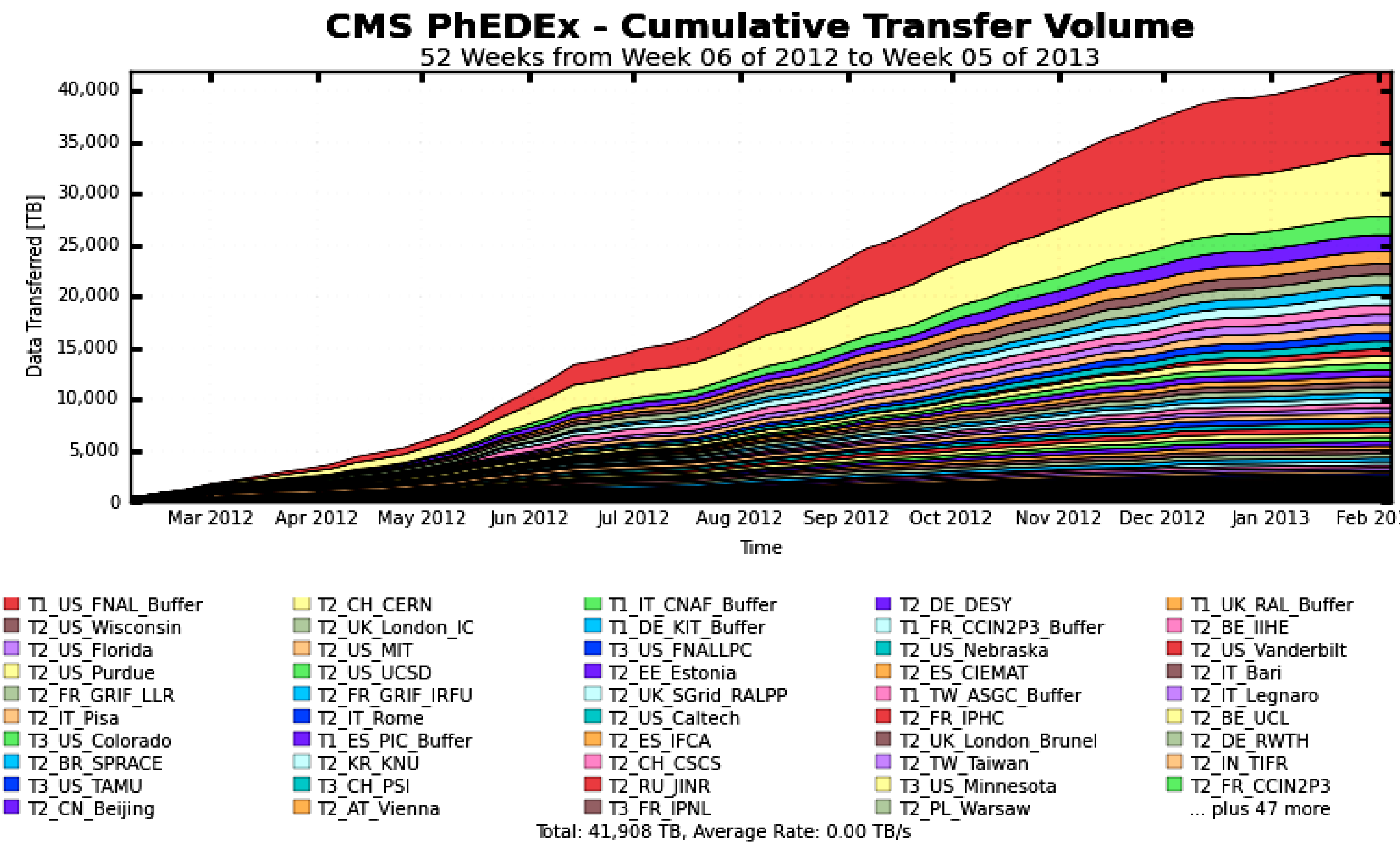
***After* Optimizations designed to reduce network traffic in 2010-2012**

CMS Data Transfer Volume (Feb. 2012– Jan. 2013)

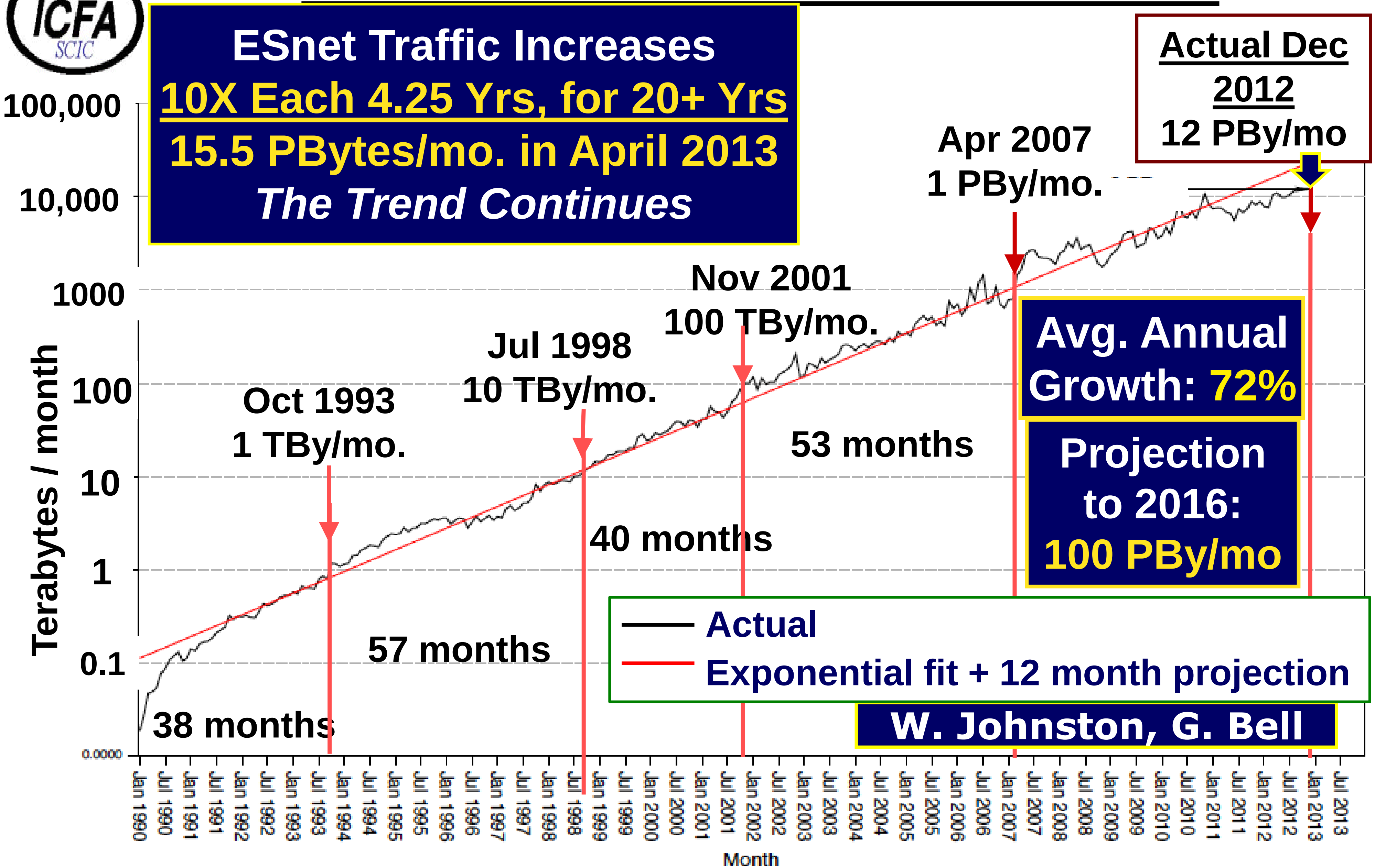
42 PetaBytes Transferred Over 12 Months
= 10.6 Gbps Avg. (>20 Gbps Peak)

2012 Versus 2011: +45%

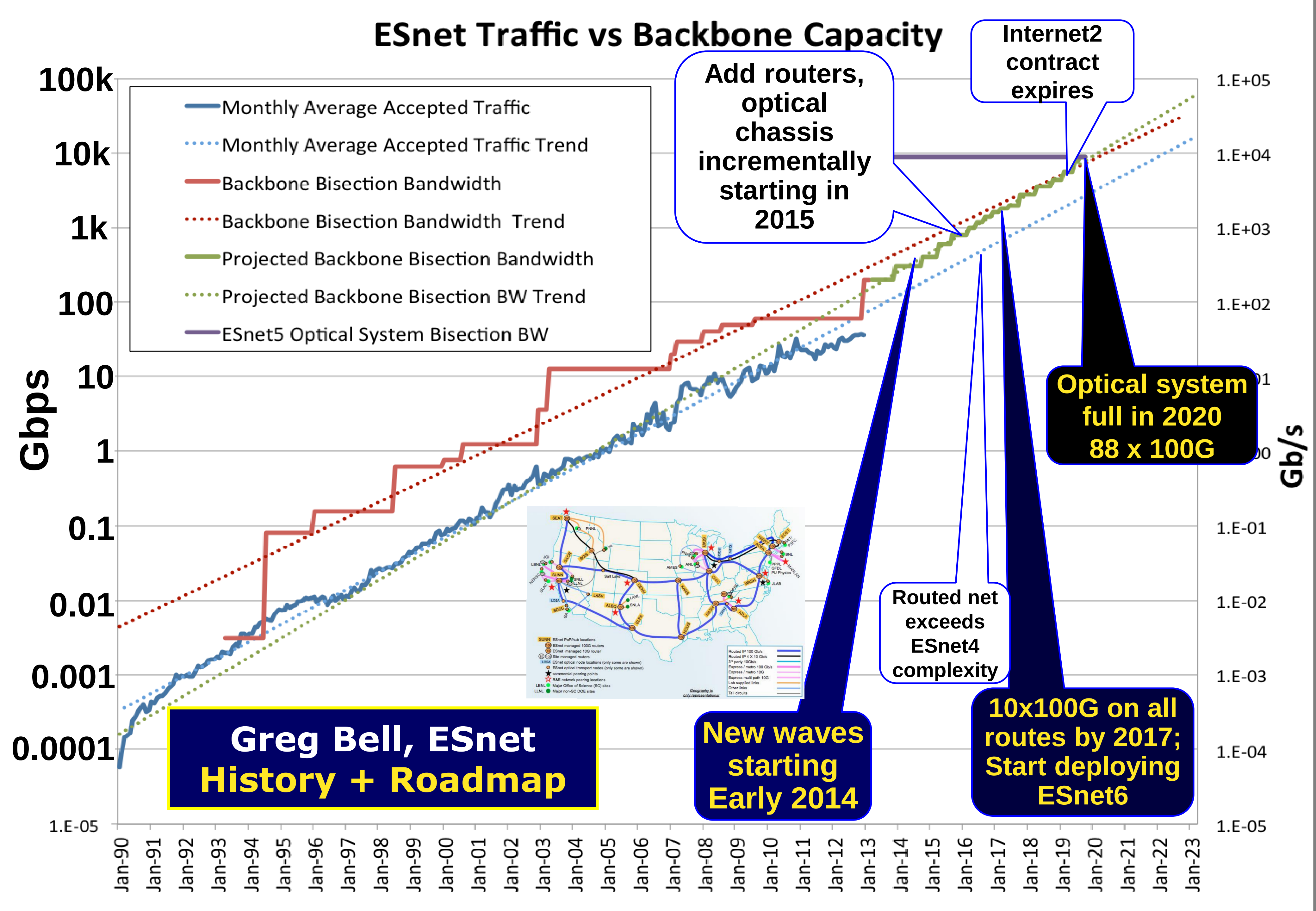
Higher Trigger Rates and Larger Events in 2015



Remarkable Historical ESnet Traffic Trend



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


ESnet Traffic vs Backbone Capacity
Internet2 contract expires
Add routers, optical chassis incrementally starting in 2015
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
Optical system full in 2020 88 x 100G
Routed net exceeds ESnet4 complexity
10x100G on all routes by 2017; Start deploying ESnet6
New waves starting Early 2014
Greg Bell, ESnet
History + Roadmap
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

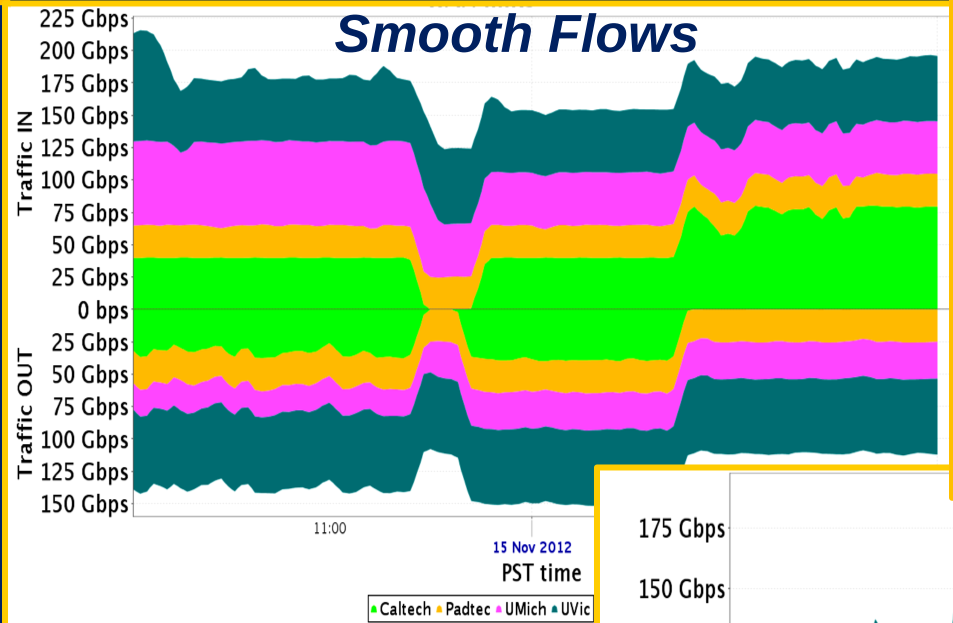
SC12 November 14-15 2012

Caltech-Victoria-Michigan-Vanderbilt; BNL

FDT Memory to Memory

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

To 3 Pbytes Per Day

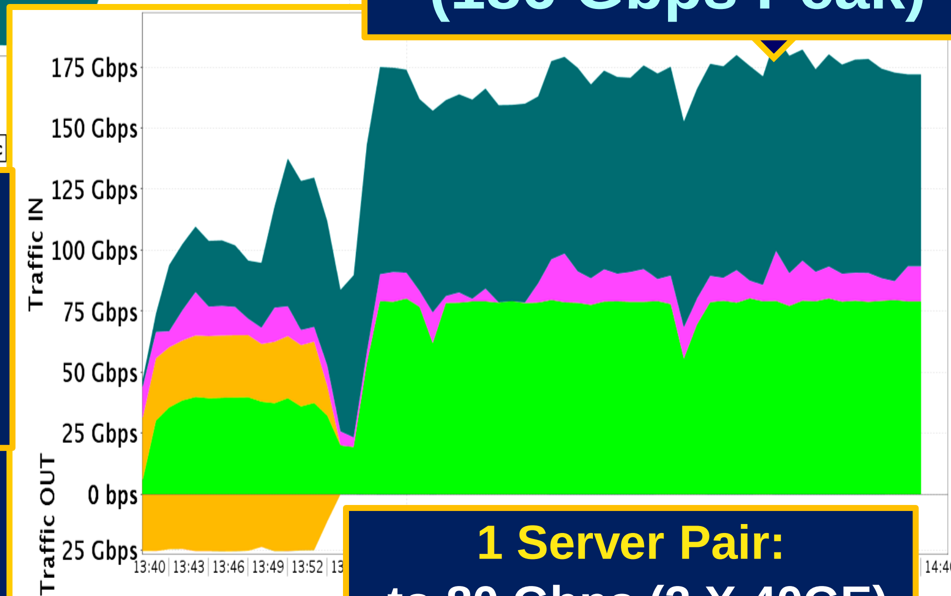


Extensive use of FDT, Servers with 40G Interfaces. + RDMA/Ethernet

FDT Storage to Storage

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

175 Gbps (186 Gbps Peak)



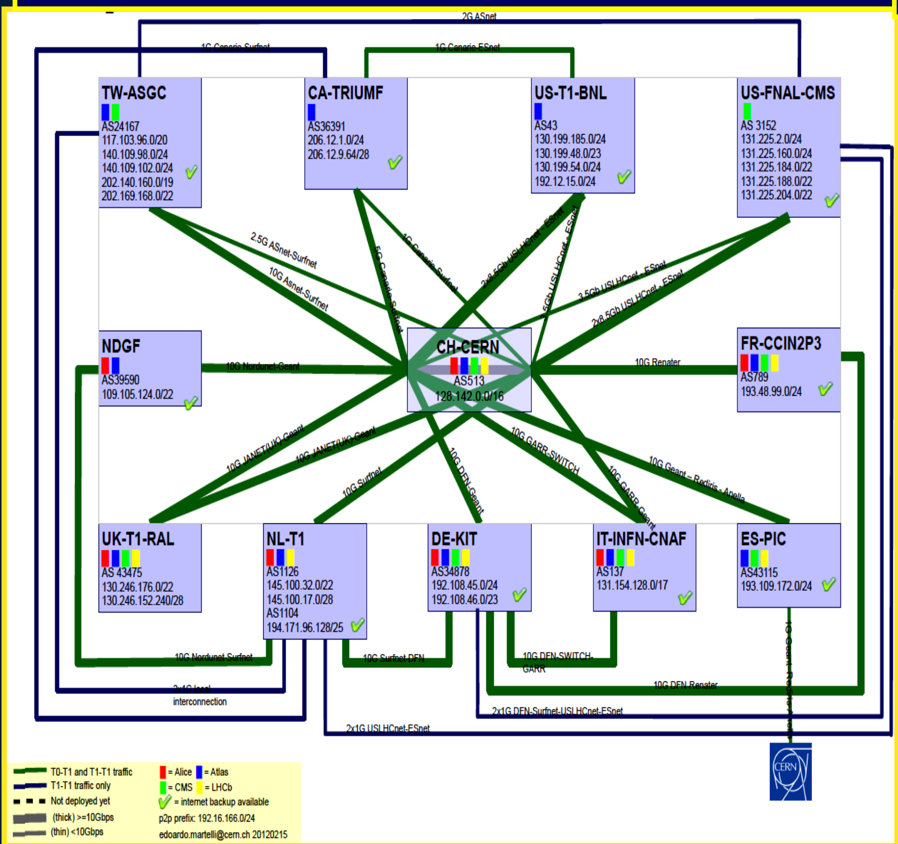
1 Server Pair: to 80 Gbps (2 X 40GE)

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

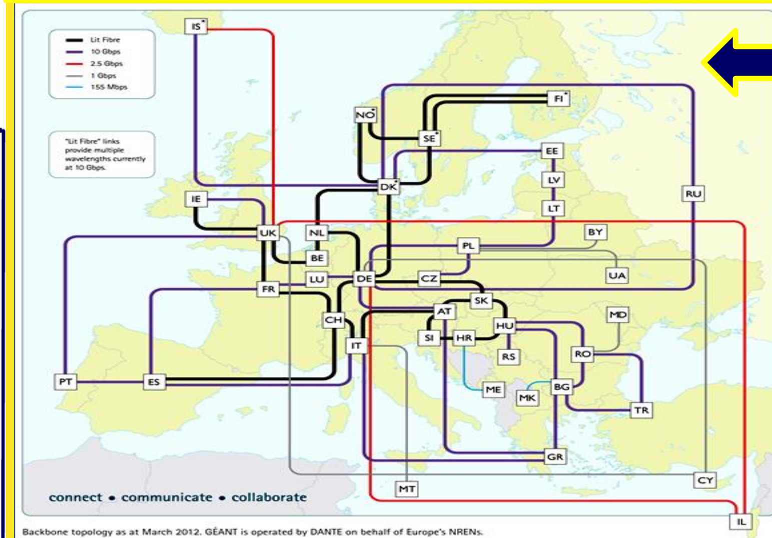
SC13: 1 Terabit/sec Trials

The Core of LHC Networking: LHCOPN and Partners

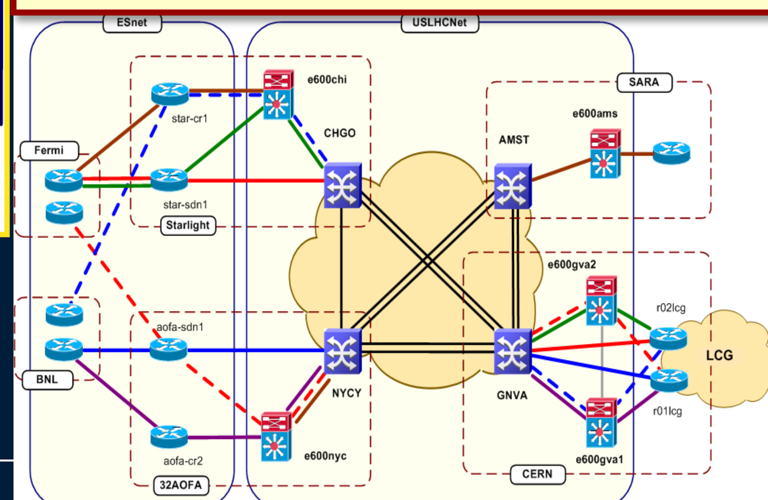
LHCOPN



GEANT



US LHCNet and ESnet



Dark Fiber Core Among 19 Countries:

- Austria
- Belgium
- Croatia
- Czech Rep.
- Denmark
- Finland
- France
- Germany
- Hungary
- Ireland
- Italy
- Netherlands
- Norway
- Slovakia
- Slovenia
- Spain
- Sweden
- Switzerland
- UK

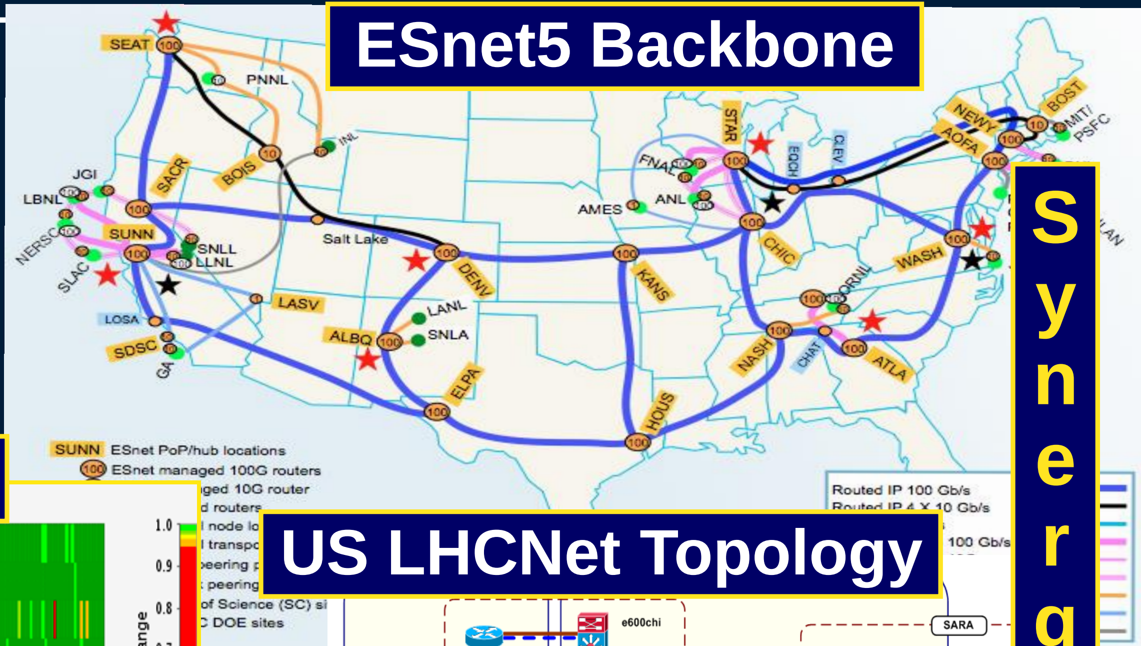
+ ESnet, NRENs in Europe, Asia; Internet2, NLR, Latin Am.,Au/NZ

Simple and Reliable, for Tier0 and Tier1 Operations

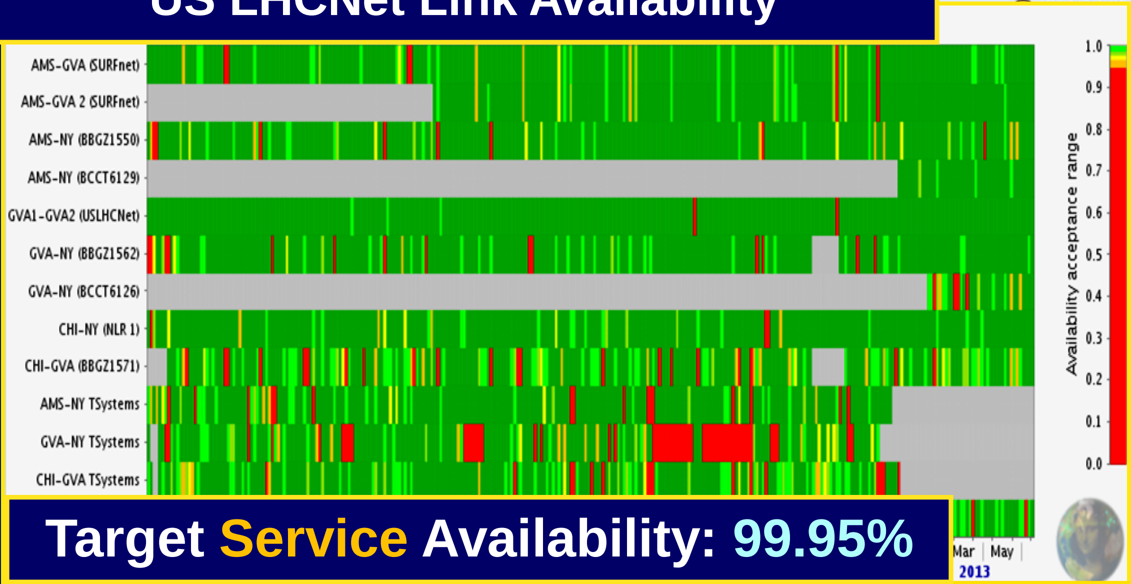
Guaranteeing High Performance in Challenging Environments

- 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

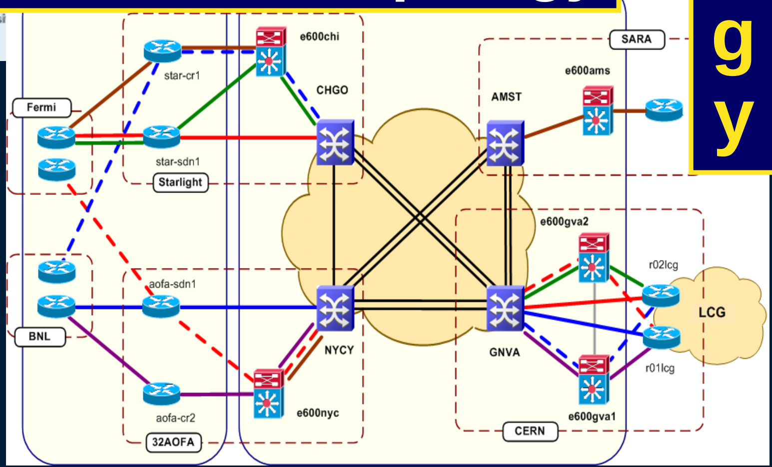


US LHCNet Link Availability



Target Service Availability: 99.95%

US LHCNet Topology



Synergy

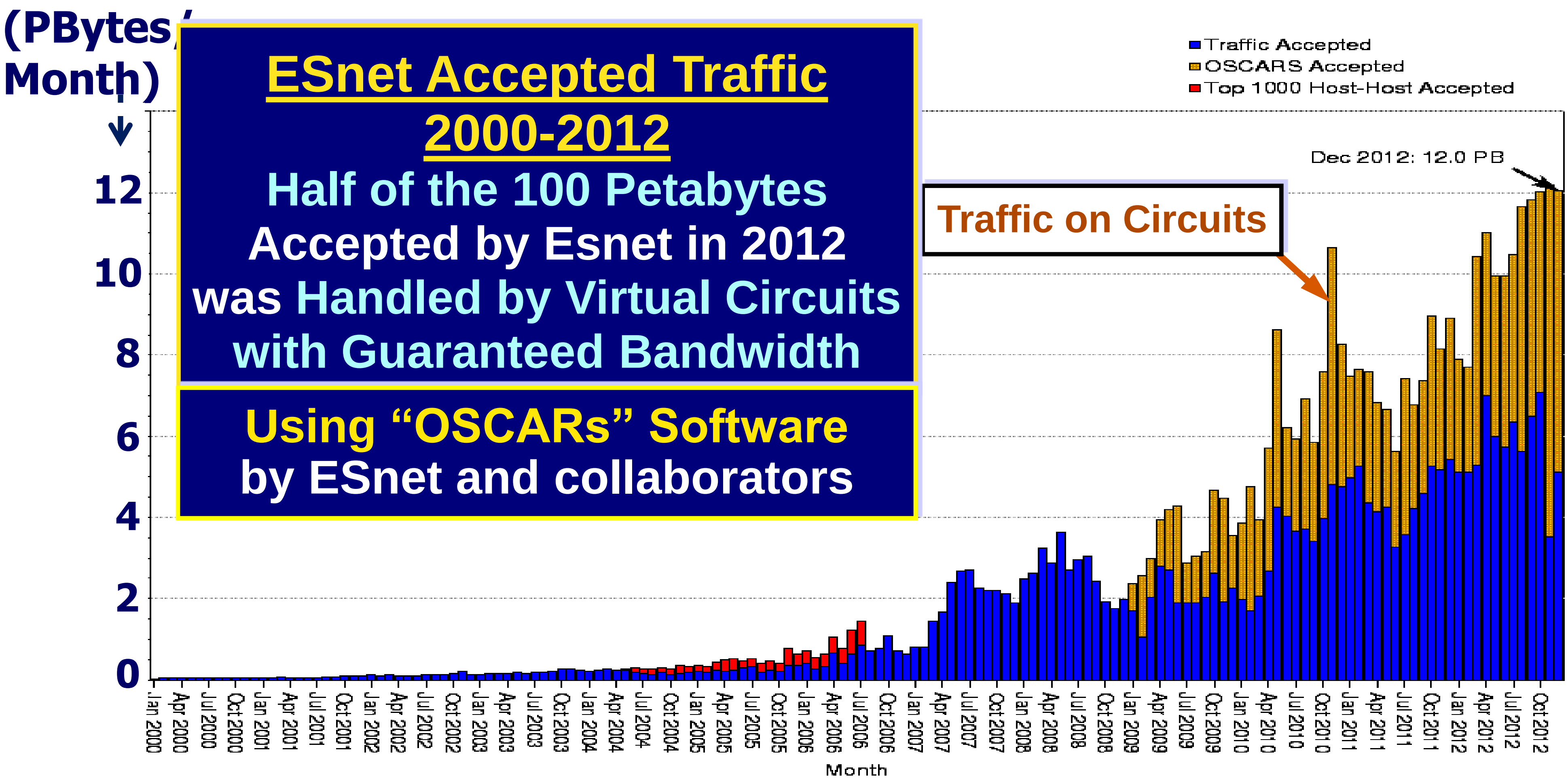
High-Availability Transoceanic solutions require multiple links with carefully planned path redundancy

Hybrid Networks: Dynamic Circuits with Bandwidth Guarantees

ESnet Accepted Traffic 2000-2012

Half of the 100 Petabytes Accepted by Esnet in 2012 was Handled by Virtual Circuits with Guaranteed Bandwidth

Using "OSCARs" Software by ESnet and collaborators



Large Scale Flows are Handled by (Dynamic) Circuits: Traffic separation, performance, fair-sharing, management

LHCONE: A Global Fabric of Interconnected Open Exchange Points

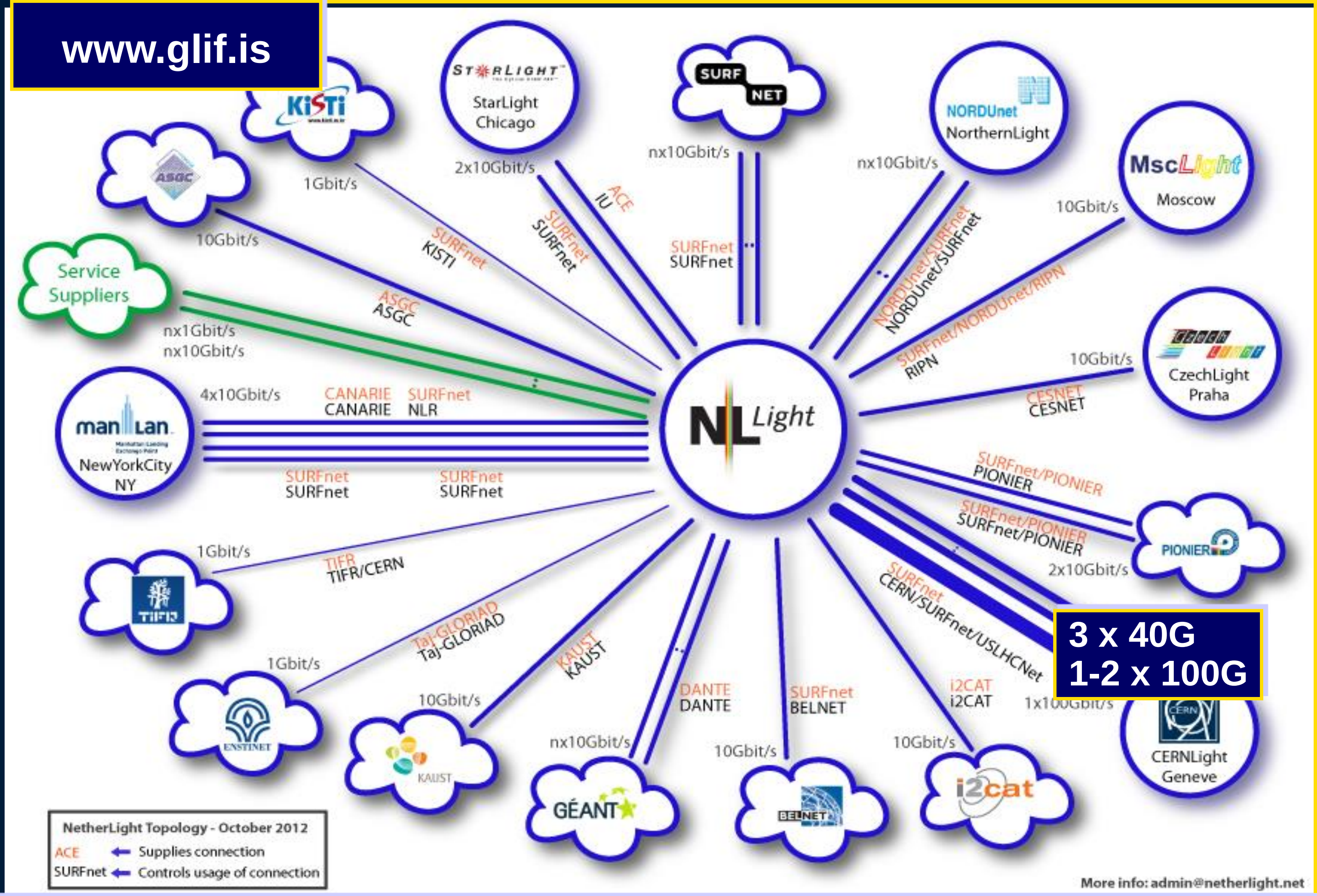
- In a nutshell, LHCONE was born (out the 2010 transatlantic workshop at CERN) to address two main issues:
  - To ensure the services to the science community maintain their quality and reliability; *Focus on Tier2/3 operations*
  - To protect existing R&E infrastructures against potential “impacts” of very large data flows
- 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 TA and other network resources
    - Through traffic Engineering and flow management capability
  - Leverage investments being made in advanced networking

Open Exchange Points: NetherLight Example

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

US LHCNet

www.glif.is



Inspired Other Open Lightpath Exchanges

- Daejeon (Kr)
- Hong Kong
- Tokyo
- Praha (Cz)
- Seattle
- Chicago
- Miami
- New York

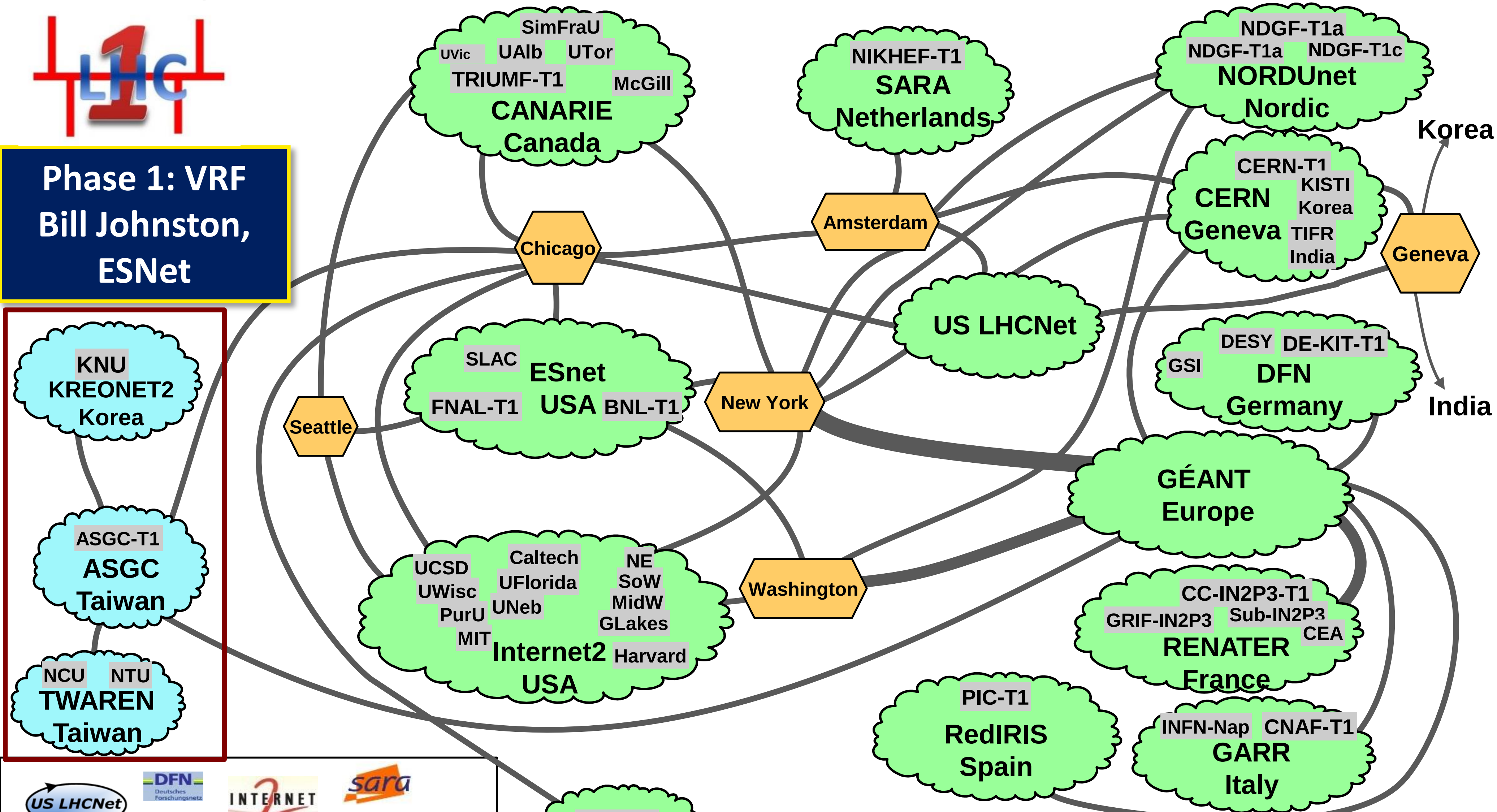
Convergence of Many Partners on Common Lightpath Concepts

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

LHCONE Overview and Activities

- Current activities fall in three areas:
  - Multipoint connectivity through a L3VPN (with Virtual Routing and Forwarding); should be restricted to the LHC Community
  - Routed IP, virtualized service
  - Point-to-point dynamic circuits
- R&D, targeting demonstration this year
  - Develop a Point-to-point service prototype, aka "experiment"
- Common to both is logical separation of LHC traffic from the General Purpose Network (GPN)
  - Avoids interference effects
  - Allows trusted connections and firewall bypass
  - Matches guaranteed bandwidth to priority class of work
- More R&D in SDN/OpenFlow for LHC traffic
  - For tasks which cannot be done with traditional methods

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



LHCONE VPN domain

NTU End sites – LHC Tier 2 or 3 unless indicated as Tier 1

Chicago Regional R&E communication nexus

Data communication links, 10, 20, and 30 Gb/s

See http://lhcone.net for details.

DYNES Map

DYNES is extending circuit capabilities to ~50 US campuses

DYNES is ramping up to full scale, and will transition to routine Operations in 2013-14

An excellent example of NREN/Science partnership

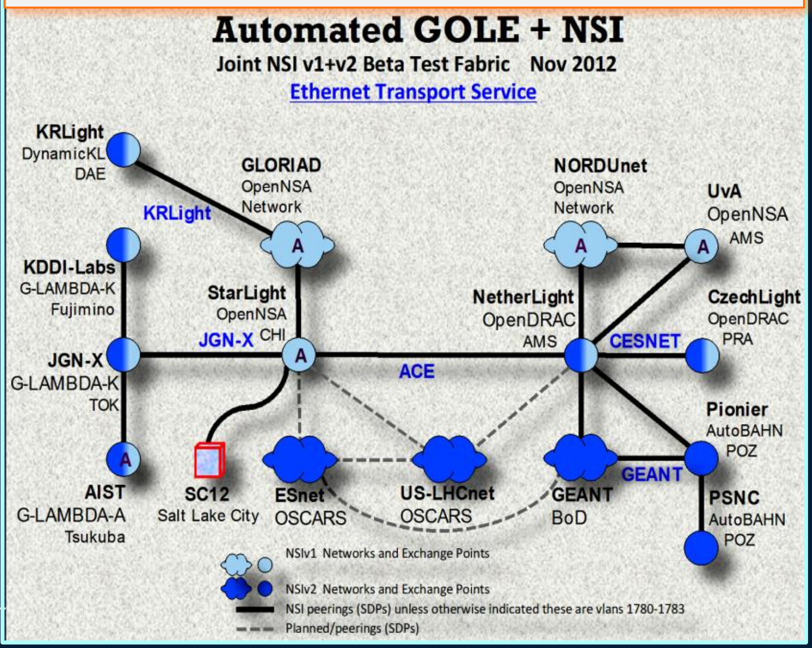
Will be an integral part of the point-to-point service in LHCONE

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

Path to LHCONE Dynamic Point-to-Point Circuit Service

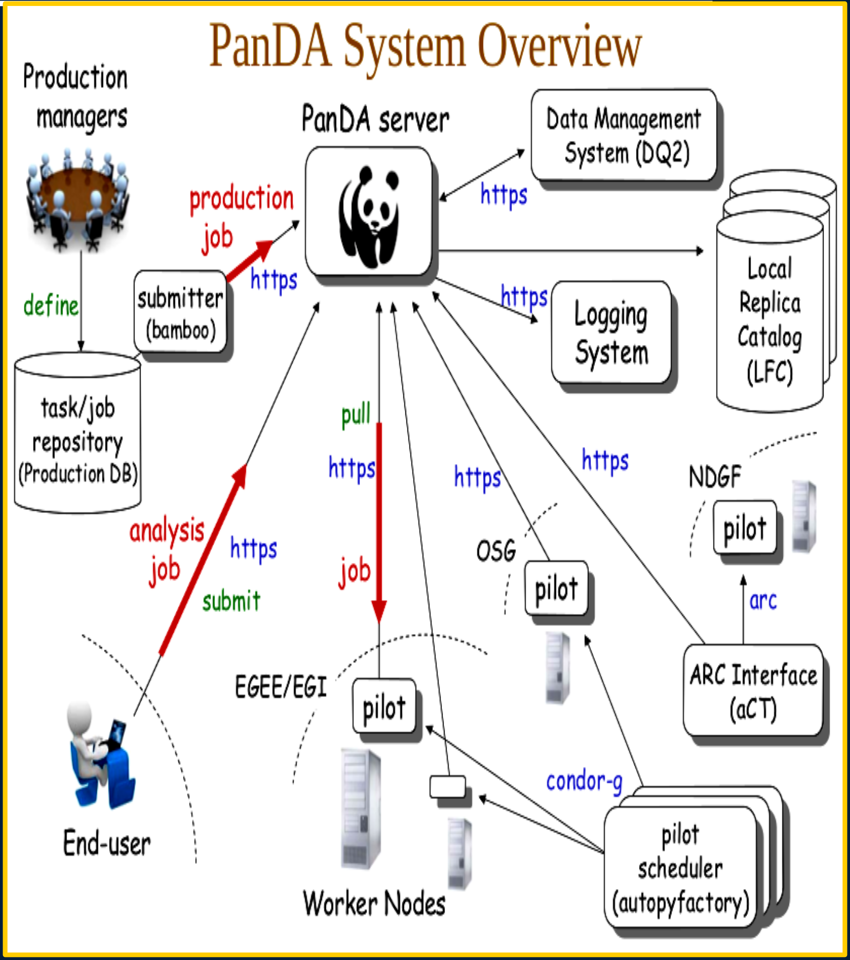
- The Goal: Provide reserved bandwidth between a pair of end-points.
- Several provisioning systems developed by R&E community: OSCARS (ESnet), OpenDRAC (SURFnet), AutoBAHN (GEANT), G-Lambda-A (AIST), G-Lambda-K (KDDI)
- For Inter-domain: moving towards an emerging standard(s)
  - OGF NSI: The Network Services Interface
  - Connection Service (NSI CS):
    - v1 'done' and demonstrated e.g. at GLIF 2012 and SC'12
- GLIF: testbed for NSI-based systems
  - E.g. Automated-GOLE Working Group is actively developing the notion of exchange points automated through NSI
  - GOLE = GLIF Open Lightpath Exchange www.glif.is

GLIF R&D and demo infrastructure at SC12



ANSE: Advanced Network Services for Experiments. Management of LHC data flows

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



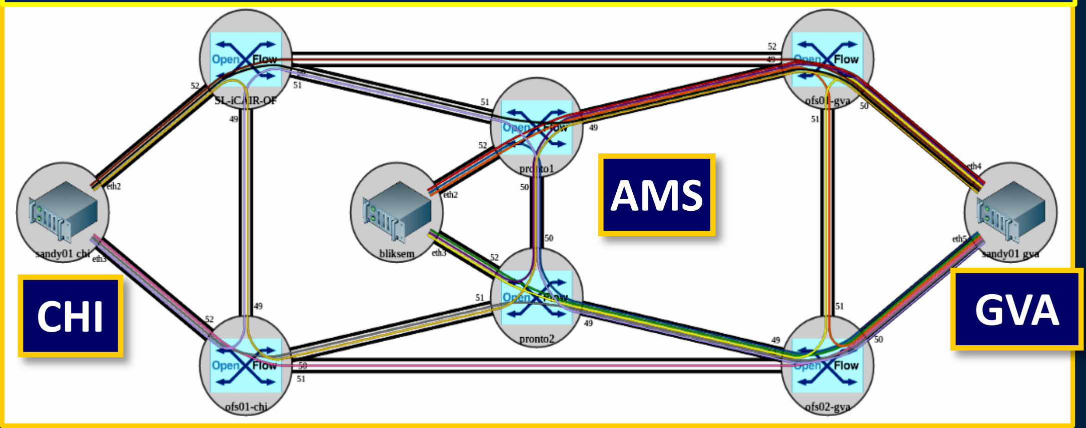
ANSE: Advanced Network Services for [LHC] Experiments

- Goals: Improve overall throughput and task times to completion
- Enable strategic workflow planning including network capacity as well as CPU and storage as a co-scheduled resource
  - Use network resource allocation along with storage and CPU resource allocation in planning data and job placement
- Path: Integrate advanced network-aware tools in the mainstream production workflows of ATLAS and CMS
  - Use accurate (as much as possible) monitoring information about the network (capacity, load, topology) to optimize workflows
  - Use existing tools and installations where they exist [FAX, PhEDEx, AAA] ; extend functionality of the tools to match experiments' needs
  - Identify and develop tools and interfaces where they are missing
- Exploit state of the art in high throughput long distance data transport, network monitoring and control

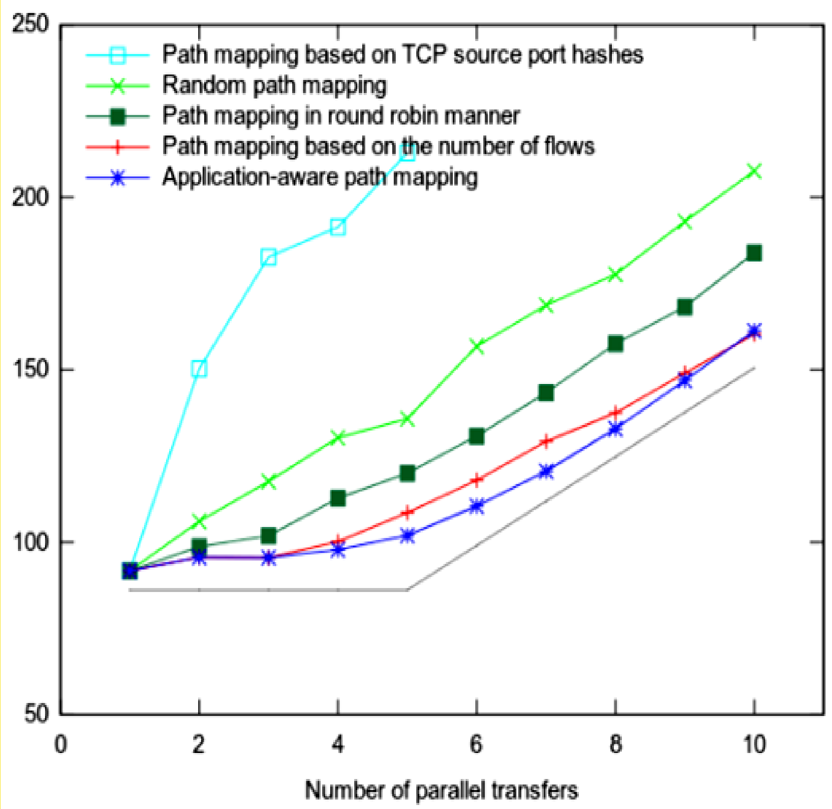
Openflow Link Level Multipath Switching: SDN use case in LHCONE

- Address problem of topology limitations in large scale networks

Geneva-Amsterdam-Chicago testbed



- Basic idea: Flow-based load balancing over multiple paths
  - Leverage global network view of the OpenFlow controller
  - Initially: use static topology
  - Later: comprehensive real-time information from the network (utilization, topology changes) as well as interface to applications



Early results: show a large throughput improvement when using an application interface and load-aware flow assignments

What about the world network scene ? The ICFA SCIC in 2012-13

http://cern.ch/icfa-scic

2013 Reports: A Banner Year

LHC Data Rampup and Discovery; but Deepening Digital Divide

- *Main Report: "Networking for HEP"* [HN, A. Mughal, A. Barczyk]
  - ➔ Updates on the Digital Divide, World Network Status
- *37 New Annexes + A World Network Overview*
  Status and Plans of Nat'l & Regional Networks, HEP Labs, & Optical Network Initiatives
- *Monitoring Working Group Report* [R. Cottrell, A. Satar, S. McKee]
- *LHCONE (www.lhcone.net):* A New Global Architecture of Open Exchange Points

Also See:

- TERENA *2012 Compendium* (www.terena.org):
  R&E Networks in Europe
- http://internetworldstats.com: Worldwide Internet Use
- OECD Broadband Portal http://www.oecd.org/sti/ict/broadband

Network Trends in 2012-13

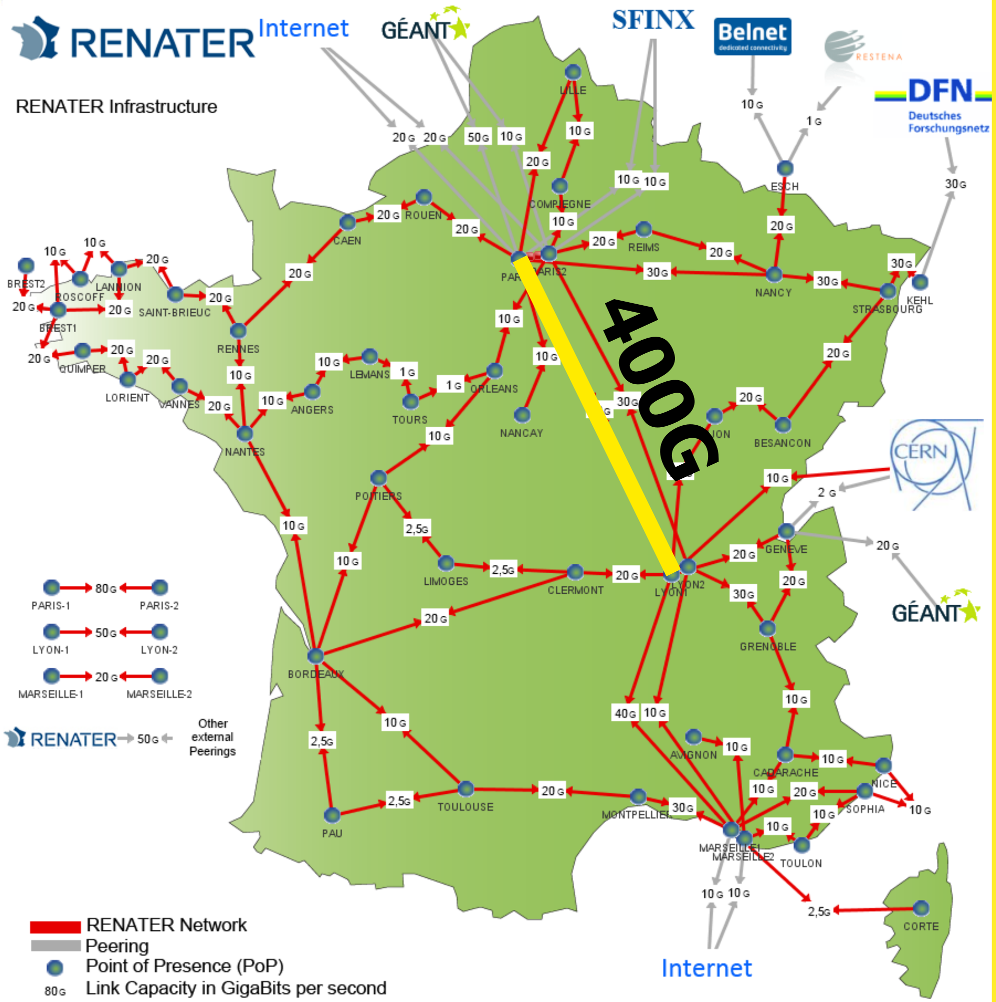
100G Evolution; Optical Transmission Revolution

- Increased multiplicity of 10G links in Major R&E networks: Internet2, ESnet, GEANT, and leading European NRENs
- Transition to 100G next-generation core backbones: Completed in Internet2 and Esnet in 2012; US 100G endsites proliferating !
- GEANT transition to 100G: Phase 1 Completed by Mid-2013
- NREN 100G already appeared and spreading in Europe and Asia: e.g. SURFnet & Budapest - CERN; Romania, Czech Rep., Hungary, China, Korea
- 100G Transatlantic (Initial trials) in 2013
- Proliferation of 100G network switches and high density 40G data center switches. 40G servers (Dell, Supermicro) with PCIe 3.0 bus
- Higher Throughput: 300G+ at SC12 – UVic, Caltech, Mich., Vanderbilt
- Trend towards SDN (Openflow, etc.): a Major Focus taken up by much of the global R&E network community and industry
- Advances in optical network technology even faster: denser QAM modulation; *400G in production (RENATER); 1 Petabit/sec on a fiber*

The move to the next generation 100G networks is well underway and accelerating; 200G, 400G production networks not far away

France: RENATER5 Dark Fiber Infrastructure

Laurent Gyde, and E.Camisard, CEF Workshop Prague Sept. 2012



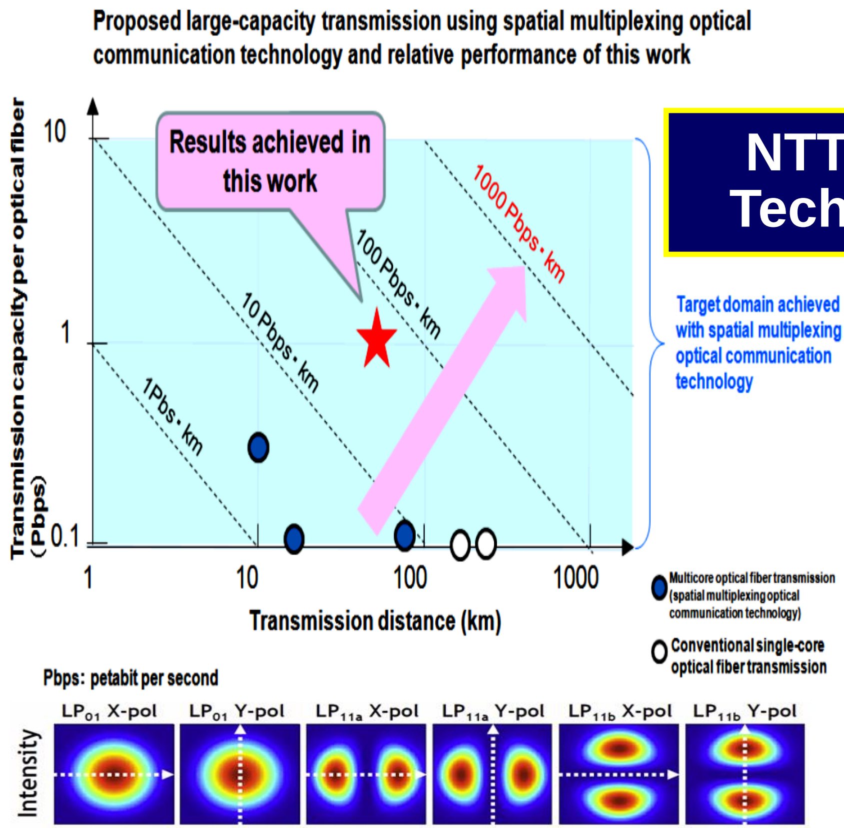
- **11,900 km dark fiber**
  120 links, 72 PoPs, 84 huts
- **665 institutions connected over 1346 sites**, in French cities and overseas
- **125 10GbE wavelengths** and 200 DWDM chassis on backbone
- **External connections: ~100 Gbps in Total**
- **Traffic: More than 100 Pbytes** internationally in 2012
- ➡ **100G to GEANT and then Major RENATER Sites** Started this year

First 400G live link in the world (Paris-Lyon) Jan. 2013

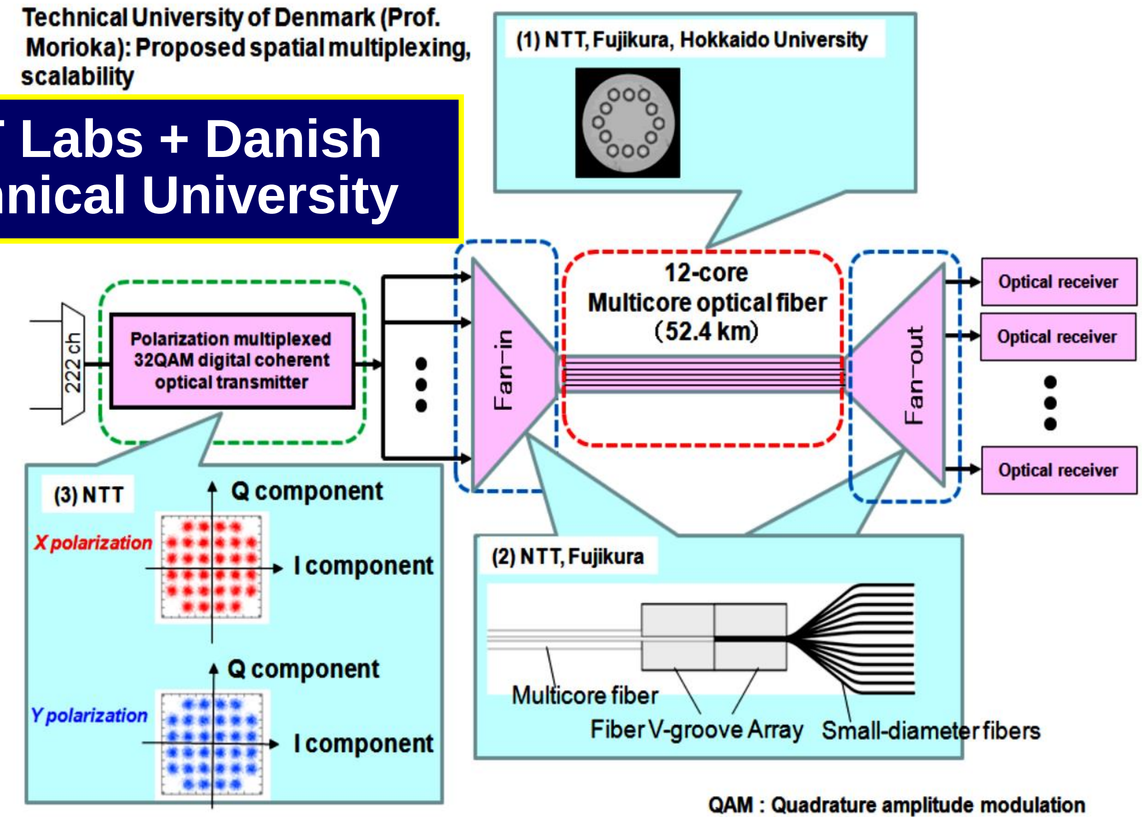
+ Other Vendor Trials: 200G wave NYC – Boston; 2 Tbps (3.3 kkm)

Optical Data Transmission: State of the Art

1 *Petabit/sec* On a 12-Core Fiber over 52 km
Spatial Mode + 32 QAM + Polarization + WDM Multiplexing

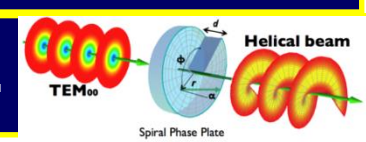


NTT Labs + Danish Technical University

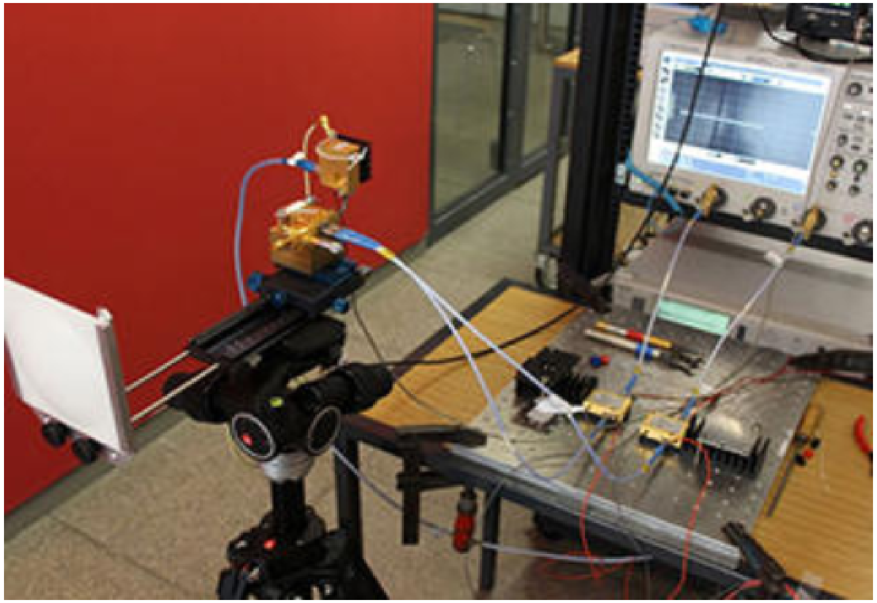


1.01 Pbps Throughput: 12 Cores X 222 Channels/Core X 380G/Channel
96 bps/Hz across 11 THz http://www.ntt.co.jp/news2012/1209e/120920a.html

Other developments: Using *Orbital ang. momentum;* Willner et al.



10/15: Wireless Data Transmission: State of the Art
Karlsruhe Institute of Technology
100 Gbps over 20m; 40 Gbps over 1 km



Setup for the world record of wireless data transmission at 100 gigabits per second: The receiver unit (left) receives the radio signal that is recorded by the oscilloscope (right). Courtesy of KIT

Extension of cable-based telecommunication networks requires high investments in both conurbations and rural areas. Broadband data transmission via radio relay links might help to cross rivers, motorways or nature protection areas at strategic node points, and to make network extension economically feasible. In the current issue of *Nature Photonics*, researchers present a method for wireless data transmission at a world-record rate of 100 gigabits per second (Gb/s).

In their record experiment, 100 gigabits of data per second were transmitted at a frequency of 237.5 GHz over a distance of 20 m in the laboratory. In previous field experiments under the "Millilink" project funded by the BMBF, rates of 40 gigabits per second and transmission distances of more than 1 km were reached. For their latest world record, the scientists applied a photonic method to generate the radio signals at the transmitter. After radio transmission, fully integrated electronic circuits were

"Wireless sub-THz communication system with high data rate"
Nature Photonics, doi: 10.1038/nphoton.2013.275, http://www.nature.com/nphoton/index.html.

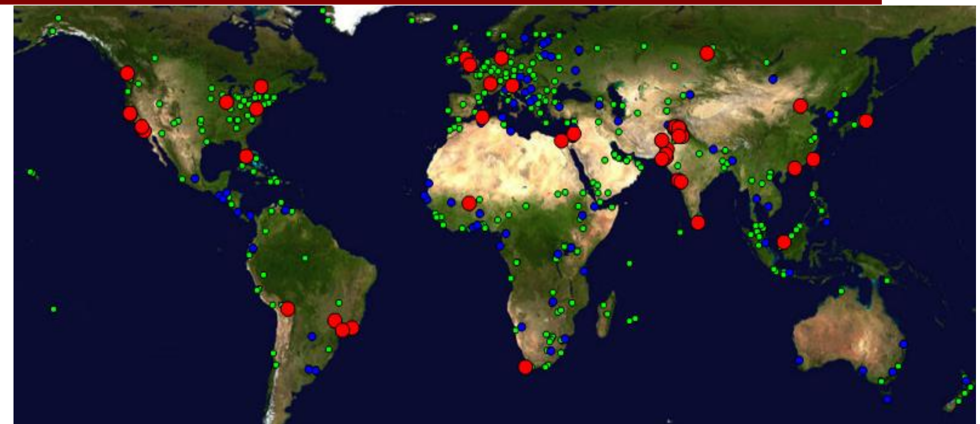
SCIC Monitoring WG
PingER (Also IEPM-BW)

IEPM
Internet End-to-end Performance Monitoring

ICFA SCIC

R. Cottrell

Monitoring & Remote Nodes Dec2012)



- **Measurements from 1995 On**

  Reports link reliability & quality

- **Countries monitored**
  - ➔ **Contain 99% of world pop.**
  - ➔ **99.5% of World's Internet Users**

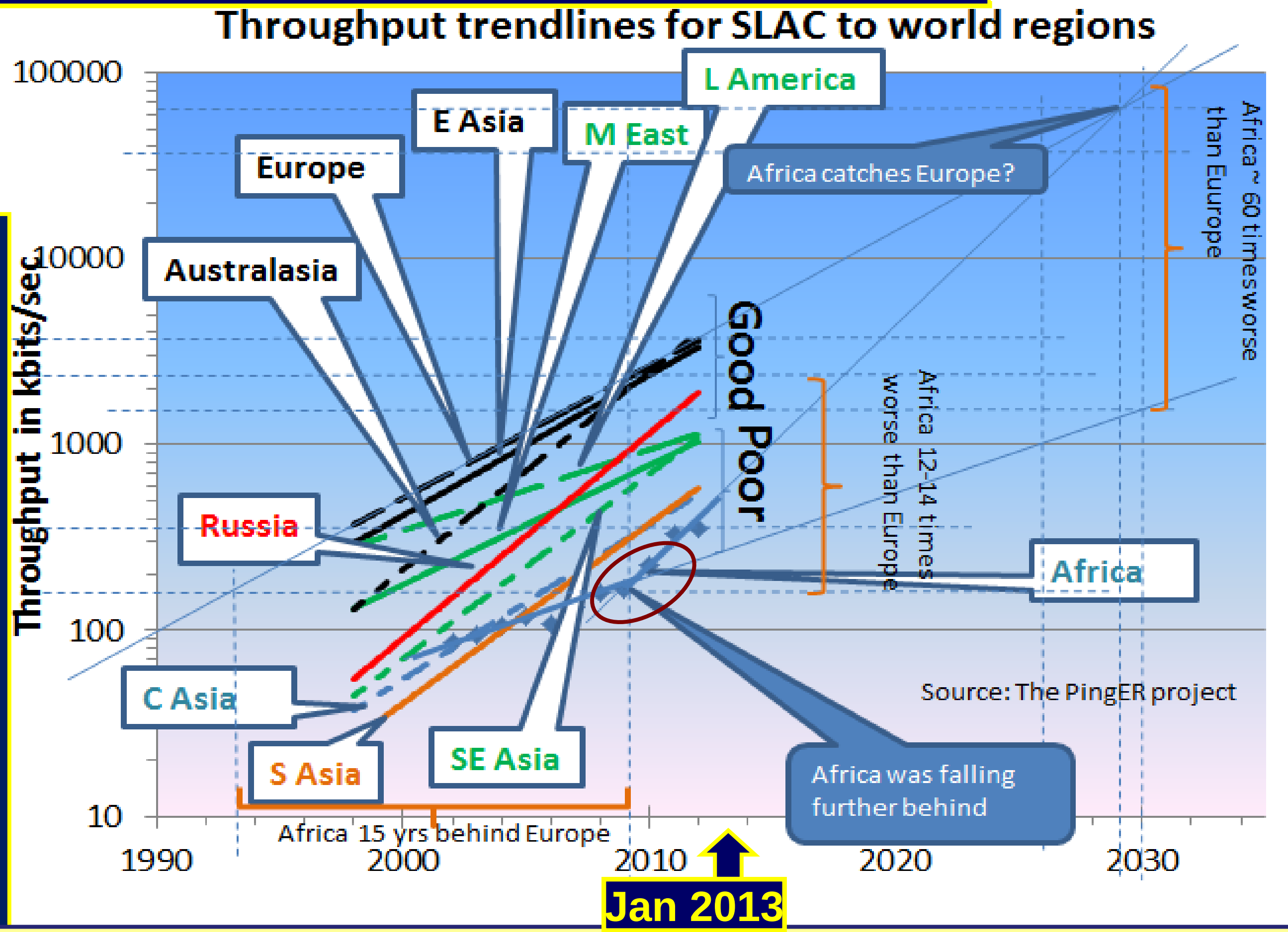
- **810 remote nodes at 775 sites in 165 nations; 86 monitoring nodes;**

- **Strong Collaboration with ICTP Trieste, NUST(Pk), U. Malaysia**

- **Excellent, Vital Work: A Volunteer Effort**

Countries: N. America (3), Latin America (*22*), Europe (31), Balkans (*10*), Africa (48), Middle East (*15*), Central Asia (9), South Asia (*8*), East Asia (4),SE Asia (9), Russia (1), Oceania (4)

Throughput Trendlines from SLAC 1998 - 2013



Top 4

Europe, N. America, East Asia & Australasia

Behind Europe

5 Yrs: Russia, Latin America, Middle East

9 Yrs: Southeast Asia

12-14 Yrs: So+Central Asia

15 Years: Africa

Derived TCP Throughput = 1460 Bytes*8bits/Byte/ (RTT * Sqrt(loss)); Matthis et al.

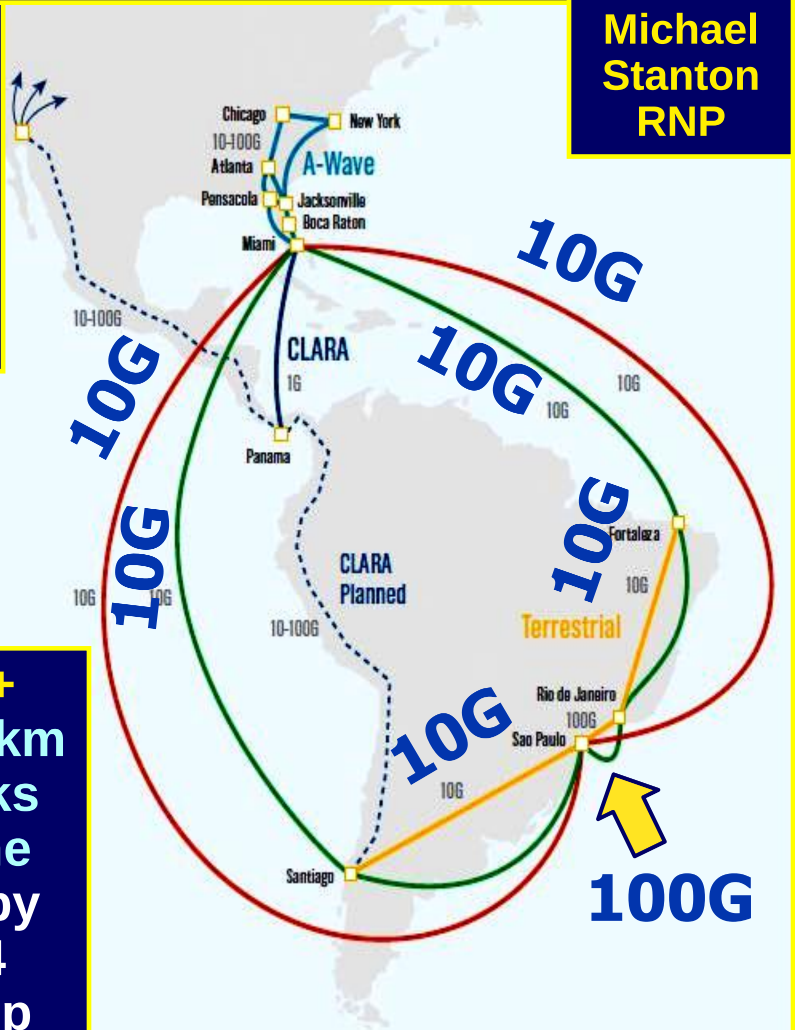
In 10 years: Russia & Latin America should catch up. Africa was falling farther behind; *new cables to Africa are making a difference since 2011*

Closing the Digital Divide: R&E Networks in/to Latin America in 2011 - 2013

RNP, ANSP, AmLight (US NSF) RedCLARA (EU)

AmLight Connects to Atlantic Wave at 10G in Miami

Michael Stanton RNP



Huawei + Vivo: 2100km WDM links across the Amazon by the 2014 World Cup

- Subsea Links Upgraded to Four 10G links on two cables: Sao Paulo, Rio, Santiago to Miami (RNP + AmLight)
- Supports Rio and Sao Paulo Tier2s, and GridUNESP Regional Tier1
- Dark Fiber metro nets in 24 of 27 State capitals; last 3 this year
- Terrestrial 10G backbone: Santiago – Sao Paulo – Rio – Fortaleza
- 100G link Rio – Sao Paulo
- AmLight Andes: Link to Chile, shared with the US Astronomy community
- Advanced net projects: *GIGA, CIPO, Future Internet Testbed*

Networks in (LHC) DAQ Systems
Evolution and Challenges

- Data Center Links: 1GE is now a "commodity"; 10GE widely available from ~2007; 40GE from ~2011; 100GE from 2015?; 400GE from 2022 ?
- Data Center Switches: 24 to 48 x 10GE nearly "mass market"; Small Switches: 32 X 40GE (Dell Z9000), and Large Switches: up to 648 X 40GE or 56G IB (Mellanox SX6536) Exist today
  - One Switch could handle all the LHC Experiments' DAQ at Run2
- Server ports, NICs: 1 GE is now "commodity"; 10 GE from ~2008; 40GE/56 IB (Mellanox) from 2012; 100GE expected in 2014-15; 400 GE by 2022 ?
- HL LHC: 100GE Links and NICs should be common by Start of Run3
- But the Total Rate HLT input rate (Projections ~ 6-32 Terabits/s each) would still require hundreds of these links:

Neufeld (TDOC), ECFA Workshop

	Event-size [kB]	Rate [kHz]	Bandwidth [Gb/s]	Year
ALICE	20000	50	8000	2019
ATLAS	4000	200	6400	2022
CMS	4000	1000	32000	2022
LHCb	100	40000	32000	2019

ATLAS and CMS: Triggered vs. Triggerless Architectures

- 1 MHz (Triggered): CMS Example
  - Network Throughput: 1 MHz with ~4 MB: Aggregate ~32 Tbps
    - Links: Event Builder-cDAQ: ~400 links of 100 Gbps
    - Switch: almost possible today; “No problem” by 2022
- 40 MHz (Triggerless): Is this feasible ?
  - Network Throughput 40 MHz with ~5 MB: Aggregate ~2000 Tbps (2 Petabits/sec)
    - Event Builder Links: ~2,500 links of 400 Gbps
    - Switch: has to grow by factor ~25 in 10 years; with Backplane capacities of 100s of Tbps: Difficult !
  - Front End Electronics
    - Tracker Readout Cables: Copper !
      ➡ Energy, Heat and Material: Show Stoppers !
- NOTE: LHCb (40 MHz) & ALICE (50 kHz Pb-Pb) will run Triggerless Post-LS2
- Nota Bene: Nanophotonics, and/or plasmonics, graphene transistors could bring major changes by the 2030’s [Think HE LHC]

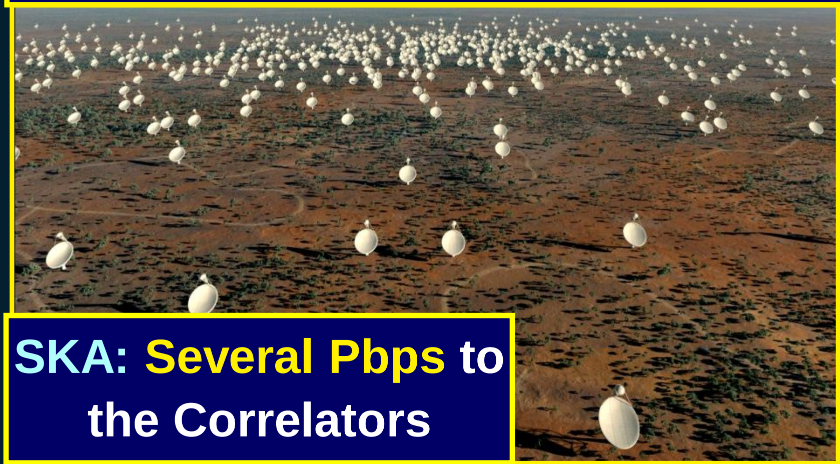
W. Smith (TDOC) at ECFA Workshop

HEP Energy Frontier Computing

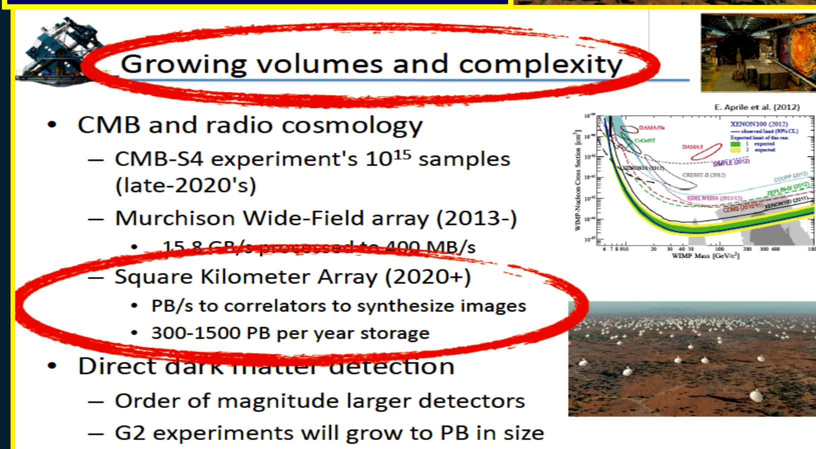
Decadal Retrospective and Outlook for 2020 [Fisk]

- 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
- We need to learn How to make better use of 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

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



SKA: Several Pbps to the Correlators

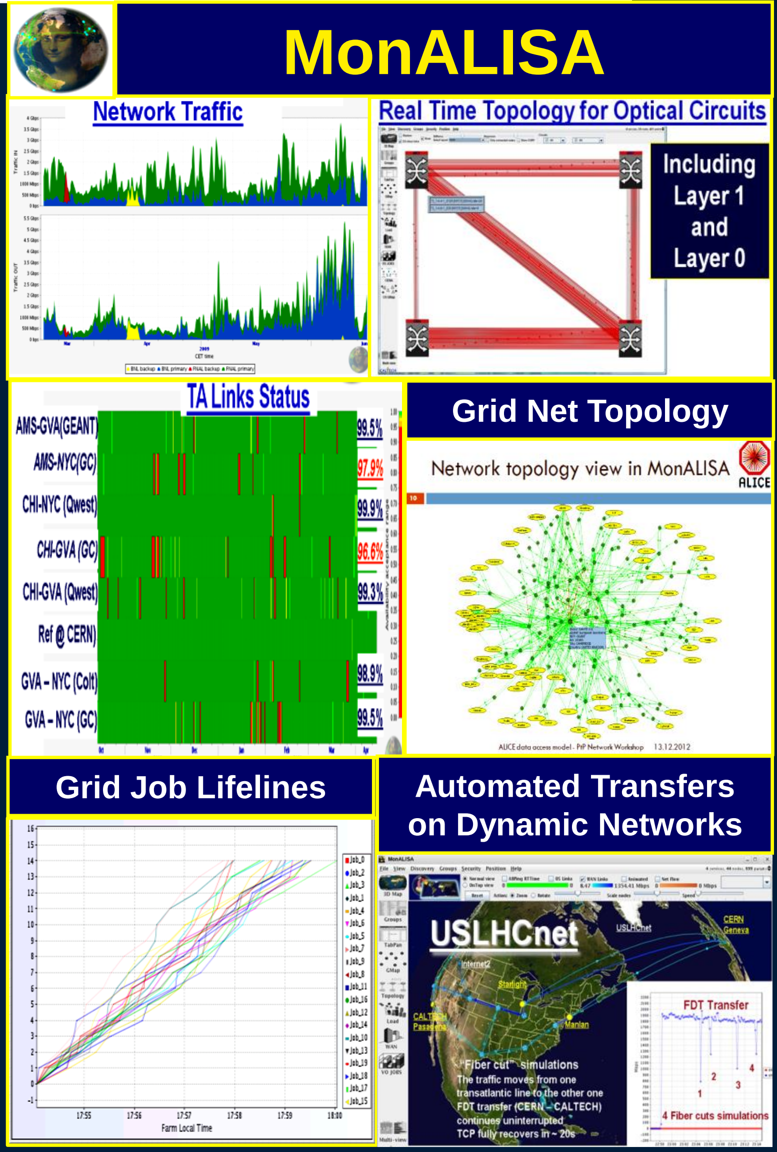


Snowmass Computing Frontier Sessions

Research and Innovation Agenda

A Core Question and a Promising Approach

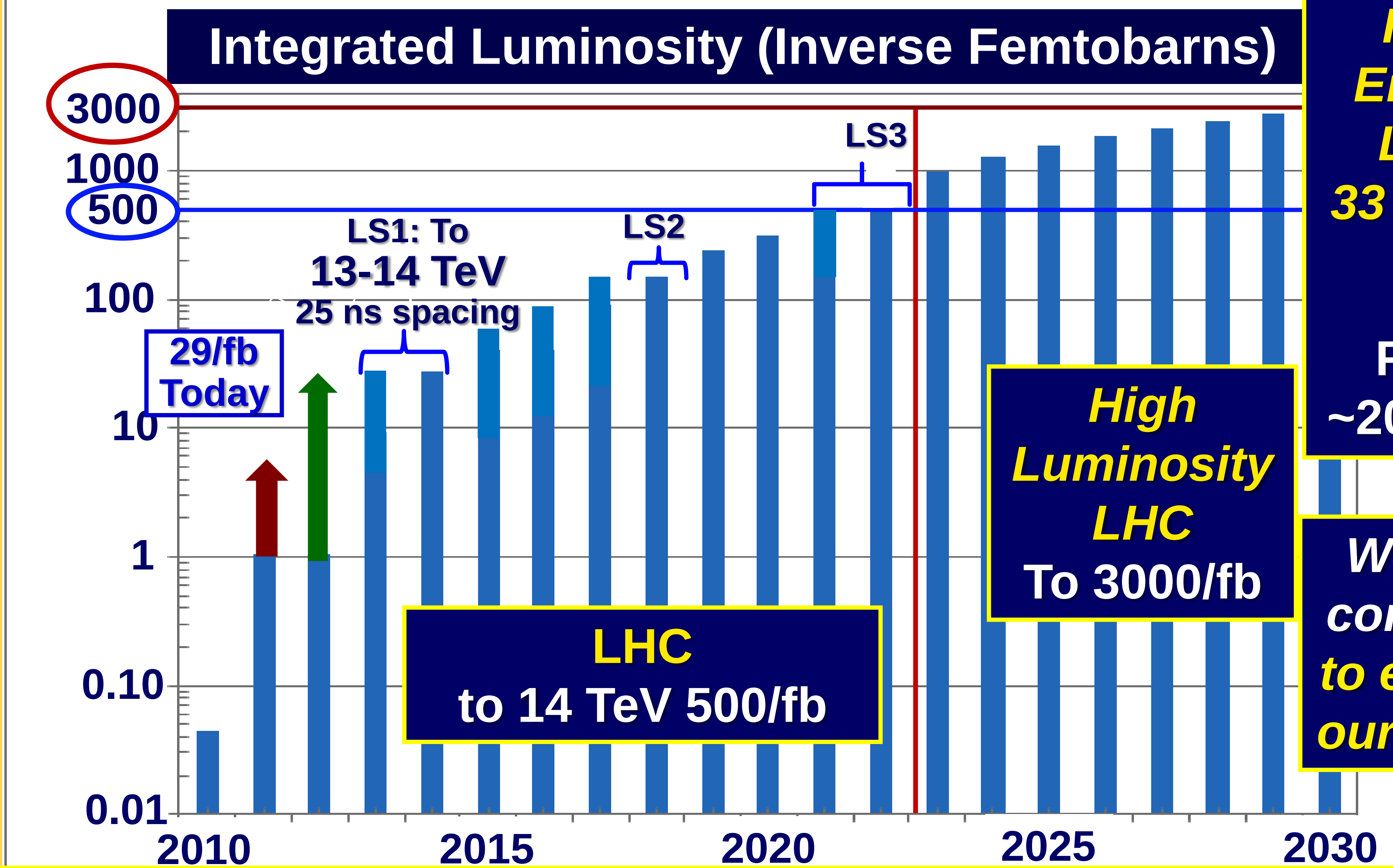
- A Core question: Can global research networks evolve: into adaptive, self-organizing systems that respond quickly to meet the needs of HEP for Petabyte -scale operations ?
- Software Defined Networking is a very promising research direction
  - Reimagine today's inflexible, proprietary HW/SW systems as open, deeply programmable components
  - Have the potential to enable innovation by facilitating virtualization, programmability, integration.
- Achieving the goal will require talented real-time system development, and code
  - Examples do exist, with smaller (but still large) scope



Off to a great start in 2010-12

LHC Outlook: to 2040+

The Road to Higher Luminosity and Energy



High Energy LHC: 33 or 100 TeV

From ~2035 On

We will continue to extend our reach

We have just begun

Networks for HEP
Conclusions and Outlook

- Run 1 brought us a centennial discovery.
- Run 2 will bring us (at least) greater knowledge,
  and perhaps great(er) discoveries.
- *Advanced networks have been, and will continue to be a key* to the discoveries. In HEP and in other fields of data intensive science
- Technology evolution may allow us to meet the short term network needs, but more attention and sufficient budgets will be needed
- In the medium term, by LHC Run2:
  A new paradigm of circuit based networks will need to emerge
- In the longer term, by Run3: *Evolution alone will not suffice*
- *A new class of global networked system are needed, building on:*
  - *The experiments data federations: FAX, AAA, Alien*
  - *New dynamic networks and methods: LHCONE, DYNES, ANSE*
- Successful development of such a system, in cooperation with expert network teams both within HEP, and beyond our community:
  - Will be essential for HL-LHC
  - Would be a game-changer with global impact

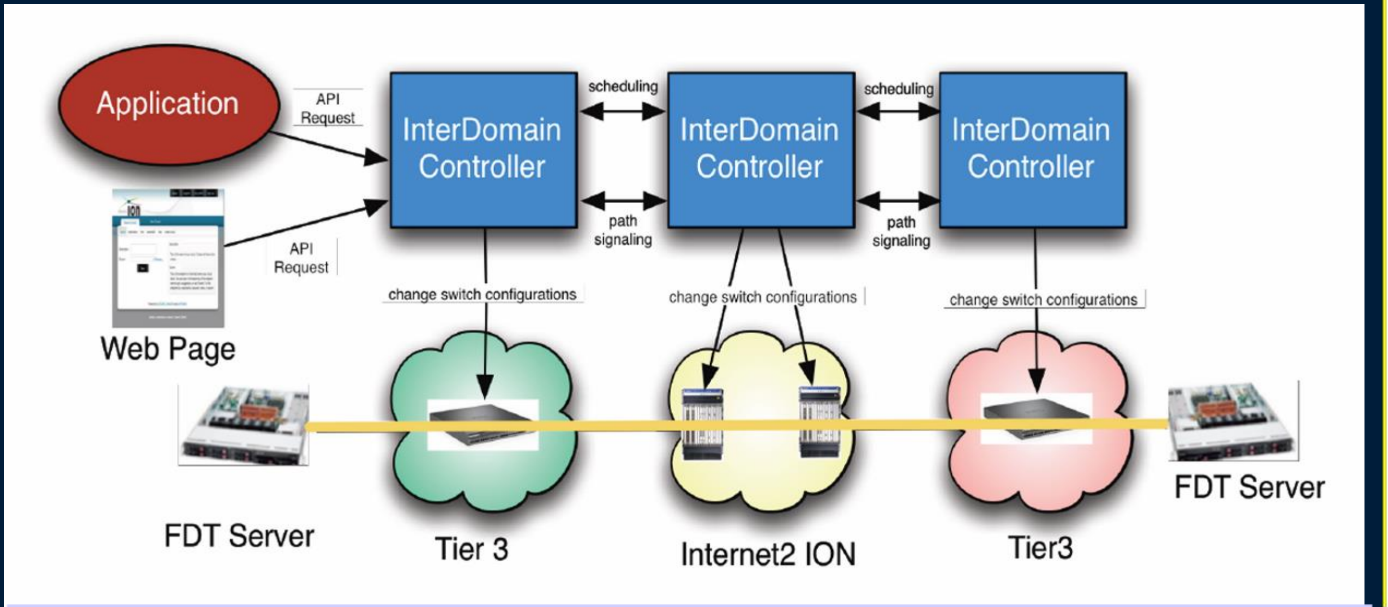
THANK YOU!

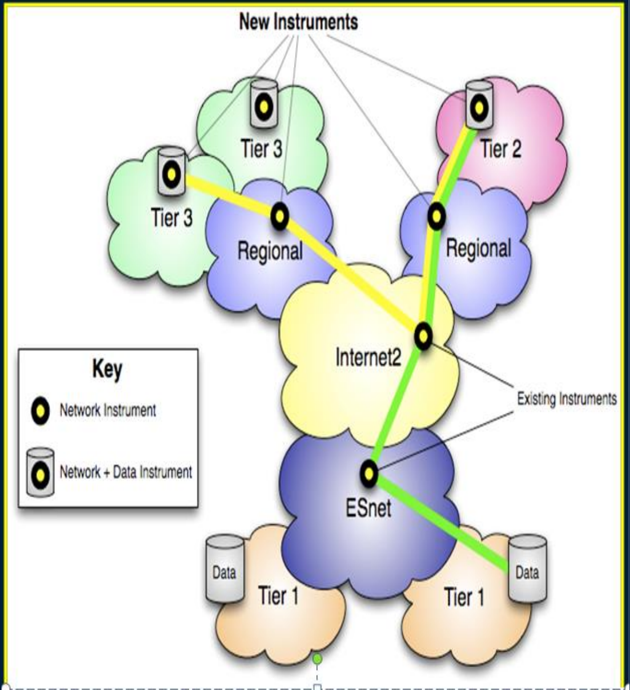
Harvey Newman
newman@hep.caltech.edu

SLIDES FOR LONGER VERSION OF THE TALK

DYNES: Dynamic Network System

- A distributed virtual cyber-instrument spanning about 50 US universities and 11 Internet2 connectors which interoperates with ESnet, GEANT, APAN, US LHCNet, and many others.
- Synergetic projects include OLiMPS and ANSE





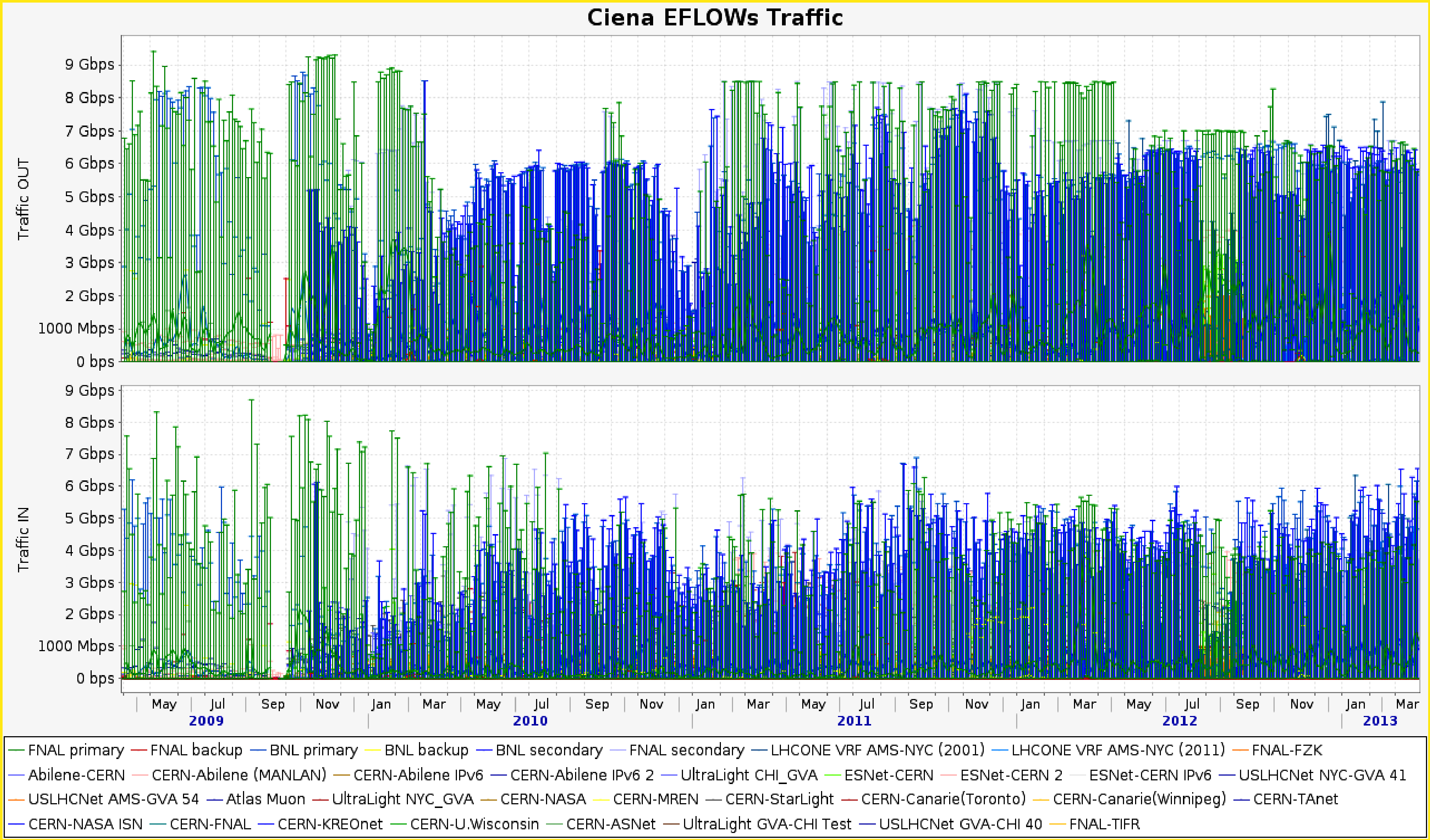
Additional work: Ensuring traffic protection, while adapting to campus and regional configurations and policies. New methods such as SDN.

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

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

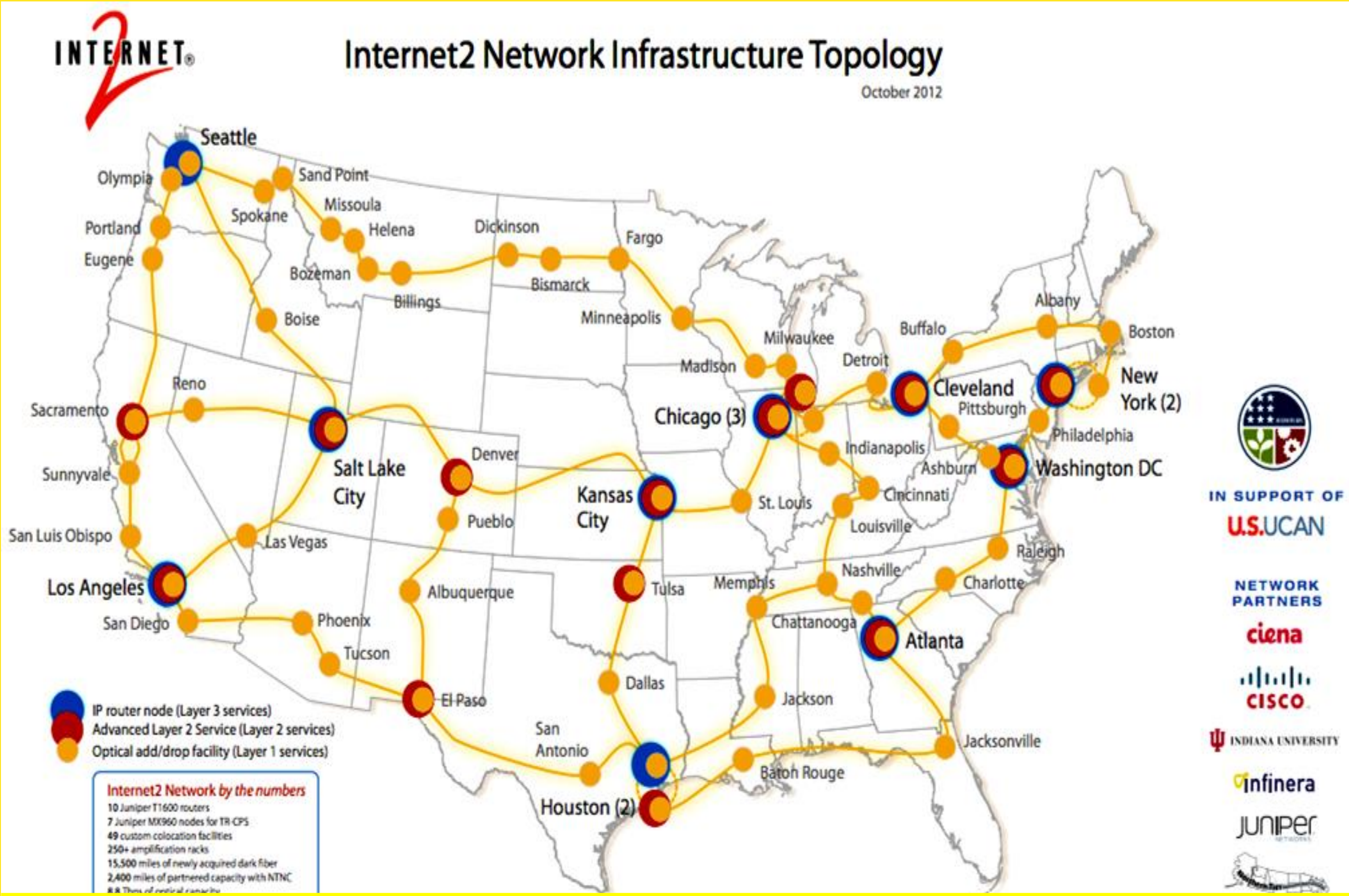
BACKUP SLIDES

US LHCNet Peak Utilization: Many Peaks of 5 to 9+ Gbps (for hours) on each link



Internet2 100G Network

Completed Fall 2012



Up to 8.8 Tbps Optical, 49 PoPs, 100G IP Service. 15.5k + 2.4k Fiber Miles

Innovation Campuses with 100G Connections: Science DMZs, Enabled by SDN by 2014

Advanced Optical, Switched and Routed Services

22 Connectors Plan

50+ X 100GE Access Links by 2015

Advanced Layer 2 Services (AL2S), including dynamic circuits

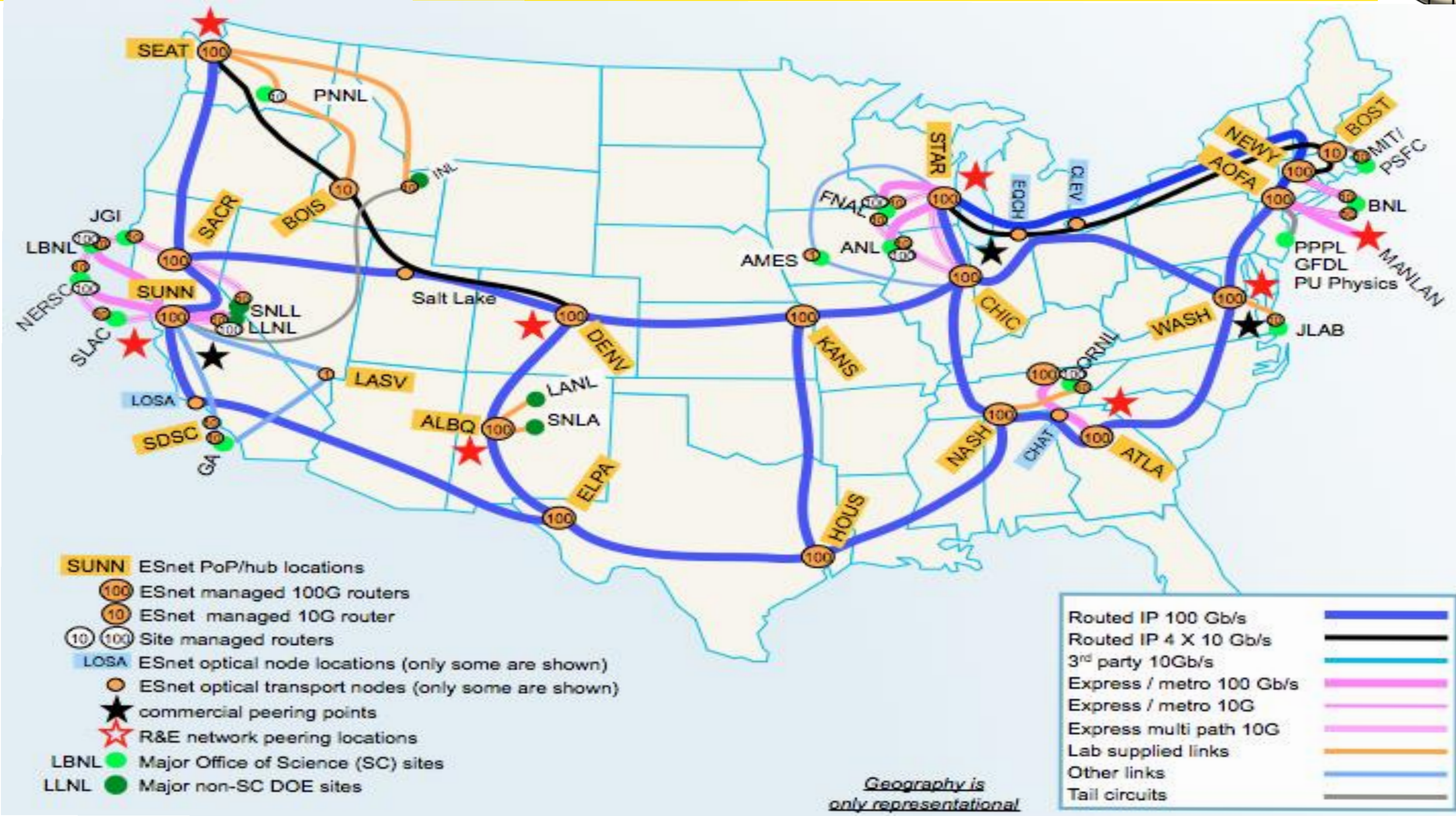
Heavily involved in LHCONE

Leading DYNES with Caltech

Moving towards **software defined networking (SDN)**

Energy Sciences Network: ESnet5

100G Backbone Completed in Nov. 2012



2 to 6 X 10G in ESnet4 in 2011-12; 100G ESnet5 from Nov. 2012

Now 15 100G Hubs; MANs (LI, CHI, SNV); Advanced 100G Testbed

Scaling up to 40 X 100G; Dark Fiber to Carry 10/40/100G Waves

2 X 100G to BNL and 100G to Fermilab: installation 1H 2012

GEANT 2012-13 Highlights
Transitions to N X 10G, now *100G*

- Monthly traffic volumes doubled in 2012, from 6 PBytes in January to 12 PBytes in December
- Many GÉANT backbone links were upgraded to 2-4 X 10G in 2011-12 due to increased use
  - 17 Links NL-UK, CH-FR, FR-UK; CH-IT, NL-US, CZ-DE; AT-CZ, AT-DE, AT-HU, AT-IT, CH-ES, DE-DK, ES-FR, FR-CH, HU-BG, HU-CZ, NL-DK
  - 5 NRENs set to access GEANT at 50G or More this year
- Transition to 100G links across Europe is now underway; Eventually plan to have 2 Tbps across the backbone.
  - CERN-Budapest link 100G for distributed LHC Tier0 In Service
  - Phase 1 500G optical ring completed; with 11 100G links.
  - Regular status updates on GEANT 100G transition at http://www.geant.net/Network/Terabit_network/Pages/home.aspx
  - Accesses at 100G and potentially N X 100G planned
- Advanced Developments: *LHCONE*

400G Production-Ready Waves Demonstrated
400GE Link in Production (RENATER)

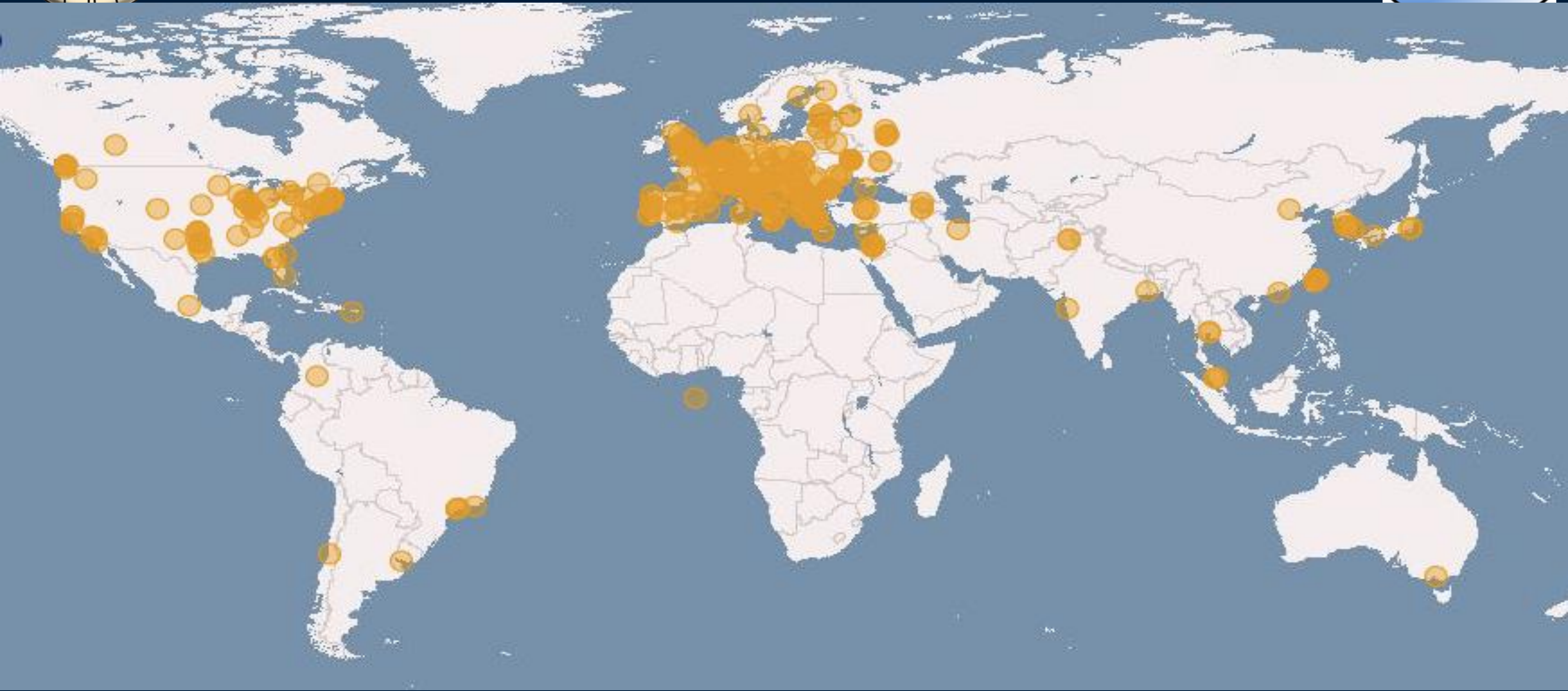
Chinese telecoms equipment vendor Huawei successfully completed a field trial using new optical fiber transmission technologies on Vodafone's live network, reaching 2 Terabit/s transmission over 3,325 km, or 2066 miles. This capacity is ~20 times higher than current commercially deployed 100G systems.
http://www.huawei.com/en/about-huawei/newsroom/press-release/hw-202114-vodafone.htm

February 6: Orange, Alcatel-Lucent provide a live 400G link to RENATER (Paris – Lyon)

France Telecom-Orange and Alcatel-Lucent have deployed the world's first optical link with a capacity of 400 Gbps per wavelength in a live network. Following a successful field trial, the 400-Gbps-per-wavelength fiber-optic link is now operational between Paris and Lyon (289 miles).
[System capacity: 17.6 Tbps on 44 400G waves.]

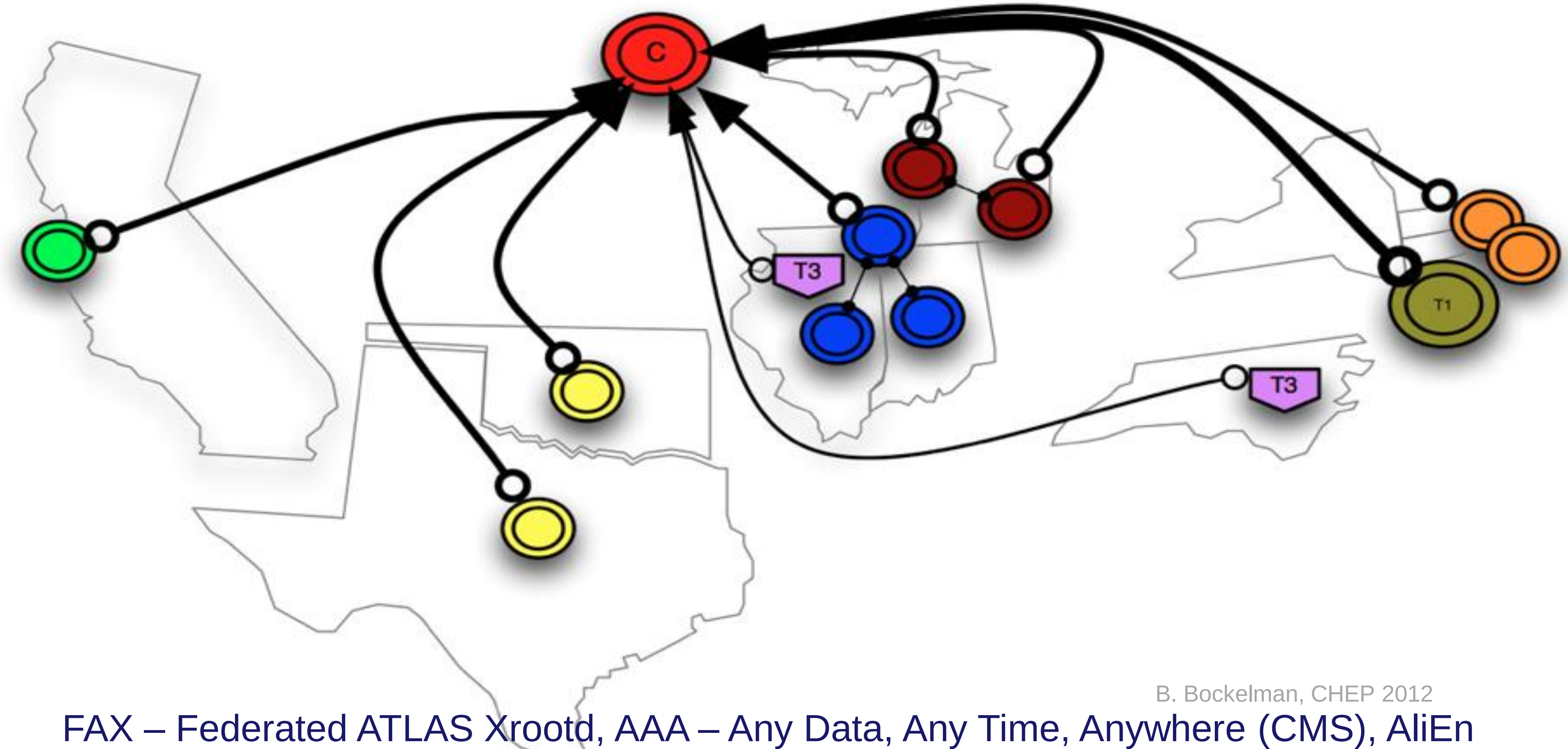
http://www.lightwaveonline.com/articles/2013/02/orange--alcatel-lucent-provide-live-400g-link-to-renater.html

WLCG Collaboration



- Distributed infrastructure of 150 computing centers in 40 countries
- 300+ k CPU cores (~ 2M HEP-SPEC-06)
- The biggest site with ~50k CPU cores, 12 T2 with 2-30k CPU cores
- Distributed data, services and operation infrastructure

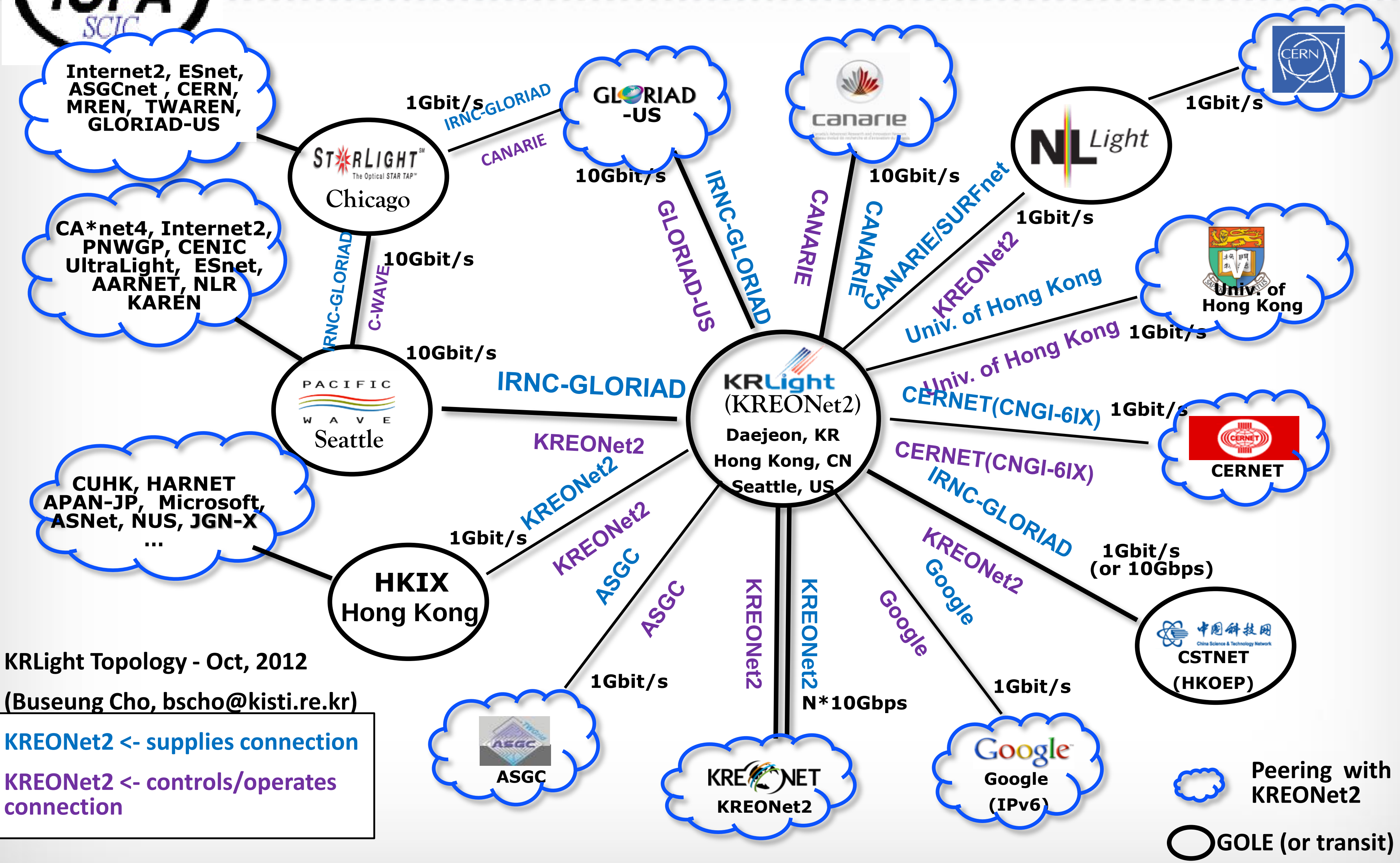
Global Data Federation



B. Bockelman, CHEP 2012

FAX – Federated ATLAS Xrootd, AAA – Any Data, Any Time, Anywhere (CMS), AliEn (ALICE)

KOREA: KRLight



KRLight Topology - Oct, 2012

(Buseung Cho, bscho@kisti.re.kr)

KREONet2 <- supplies connection

KREONet2 <- controls/operates connection

SCIC Presentation to ICFA

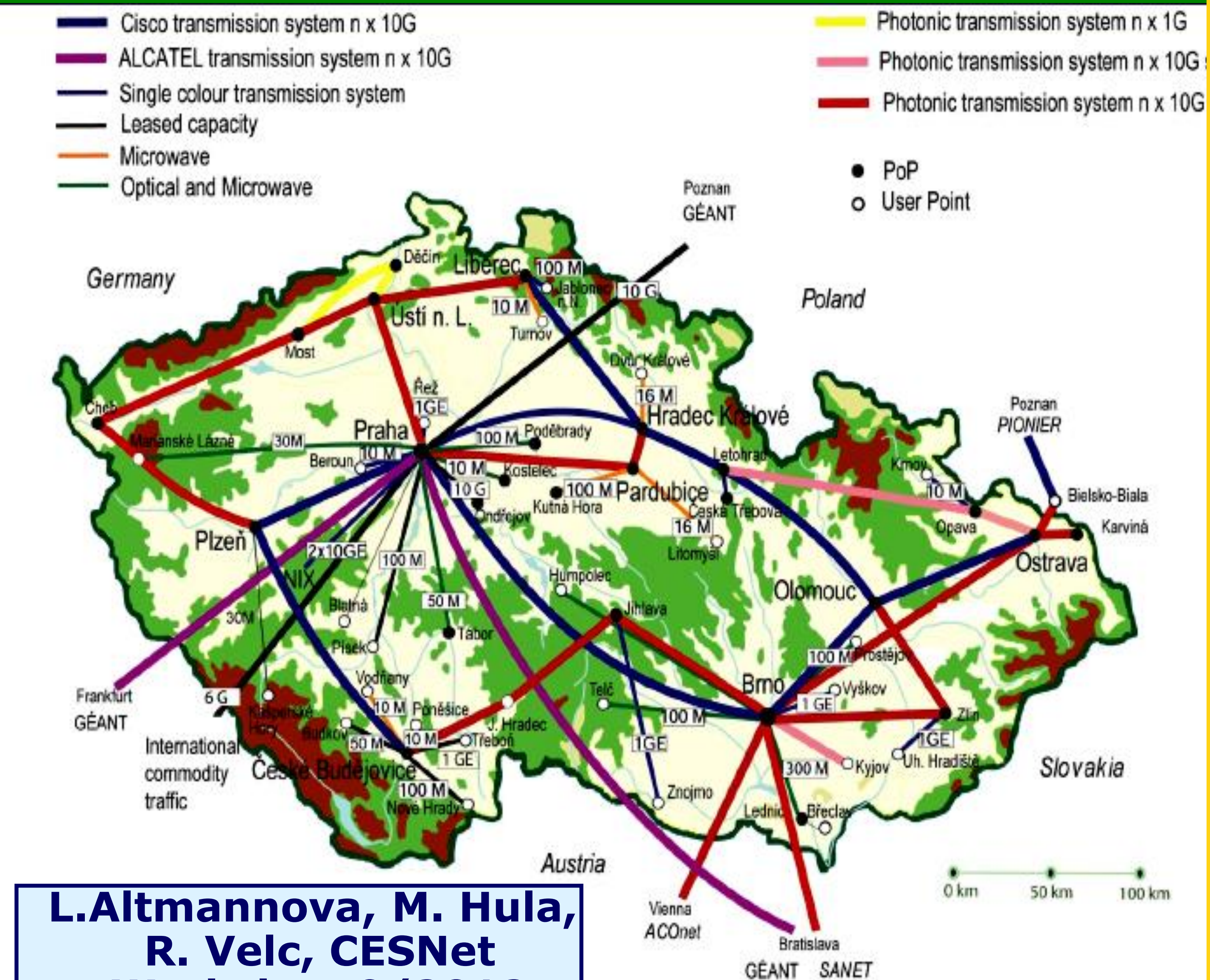
Highlights: http://cern.ch/icfa-scic

- Internet World Trends: Users, Penetration, Traffic Growth
- ICFA SCIC Reports, Work in 2012 and Conclusions
- Networking for HEP in the LHC Era; Evolution and Revolution in 2012-13
- The Move to New LHC Computing Models: LHCONE Ramps Up
- SCIC Monitoring Group: Mapping the Digital Divide
- *Closing the Digital Divide: Model Examples and Problem Areas*
- Advances in High Speed Data Transfers
- Optical Data Transmission: the State of the Art
- SCIC Monitoring WG: Updates, Key Observations, Funding

SCIC Reports, Annexes, and Full Presentation: Comprehensive Information

- Networking for HEP in the LHC Era: Part2
- The New Computing Models and LHCONE: Part2
- *SCIC Monitoring WG: Mapping the Digital Divide, Part2*
- *Closing the Digital Divide: Model & Problem Areas, Part2*
- Internet World Trends, Part2
- Advances in High Speed Data Transfers, Part2
- The Rise of Broadband: 2nd Digital Divide
- The Rise of Dark Fiber Networks; Dark Fiber Networks Closing the Digital Divide
- Nat'l, Continental and Transoceanic Network Infrastructures: Transition to 40G and 100G Cores
- Dynamic Circuits for Large Flows: OSCARS; the DYNES Project
- Global Subsea Cable Status; Capacity Growth and Price Trends
- Optical Data Transmission: the State of the Art, Part 2

CESNet2 and CESNet EF: Advanced Digital and All Photonic Networks



5340 km
Leased Fiber
420 Km Dark Fiber
738 km Exp. Net Facility

All-Photonic Service

Fixed Bandwidth with Fixed Delays

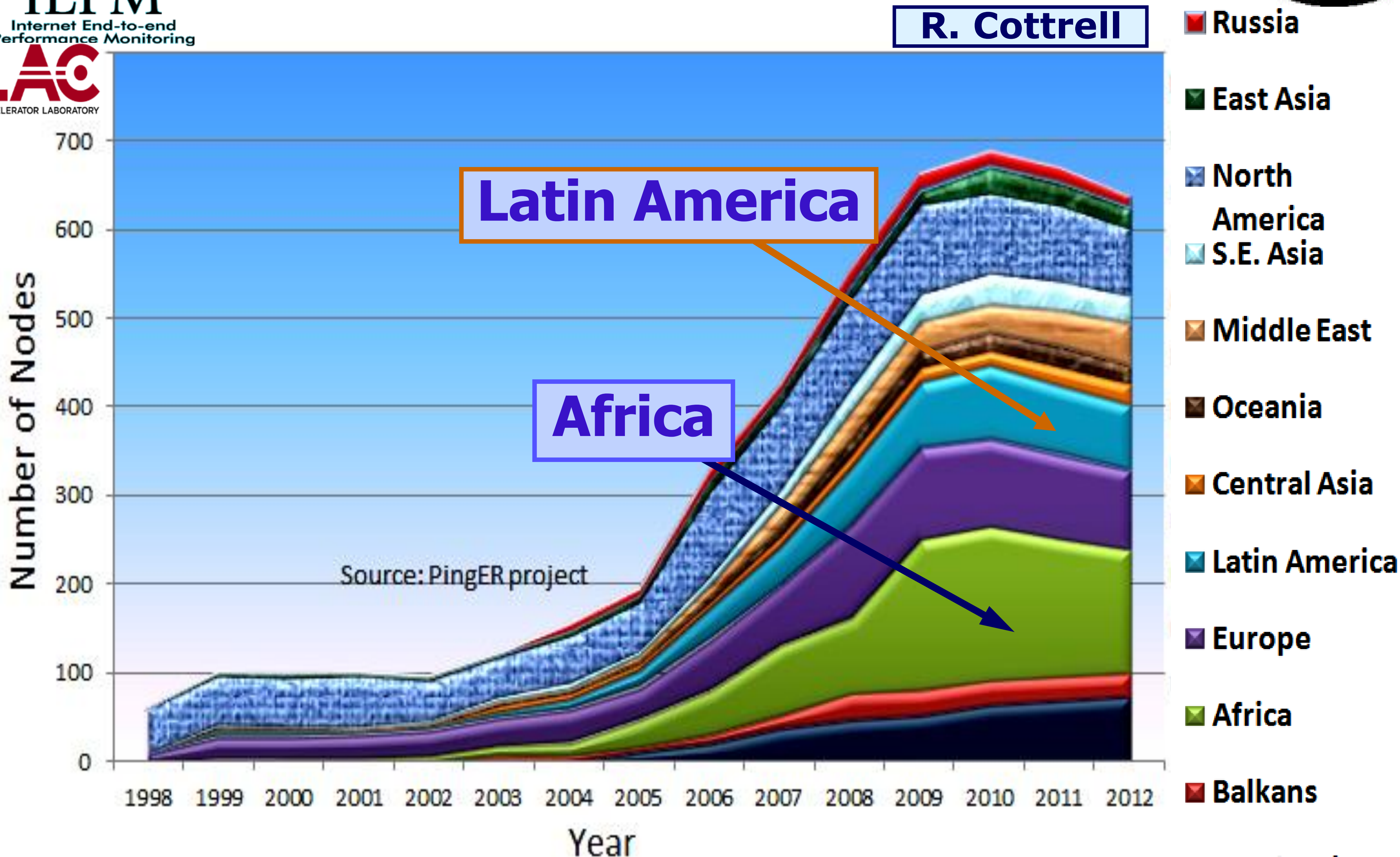
Useful for New Applications

Precise Timing, Real Time, Interaction with External Processes

L.Altmannova, M. Hula, R. Velc, CESNet Workshop 9/2012

Number of Hosts Monitored By Region: 1998 - 2012

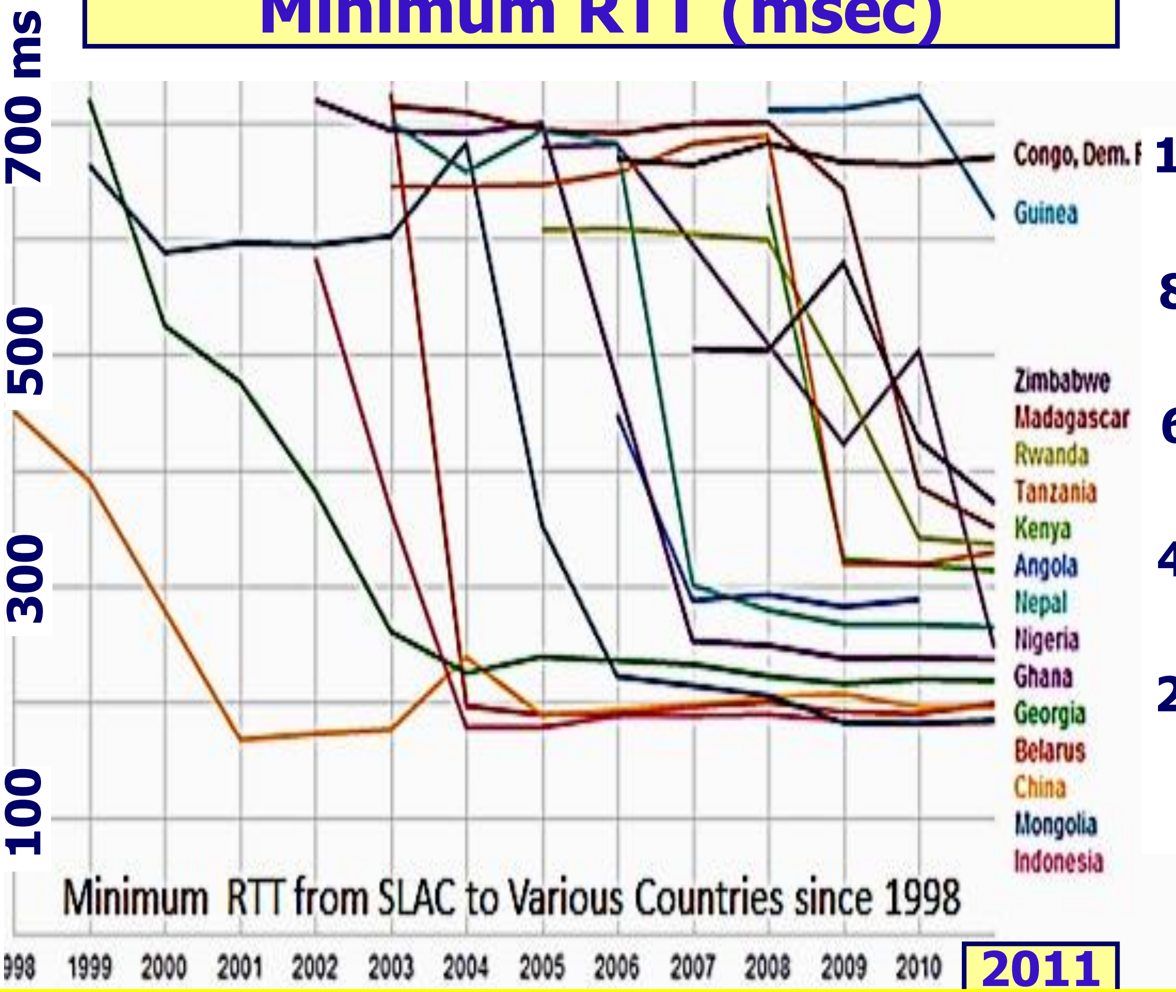
R. Cottrell



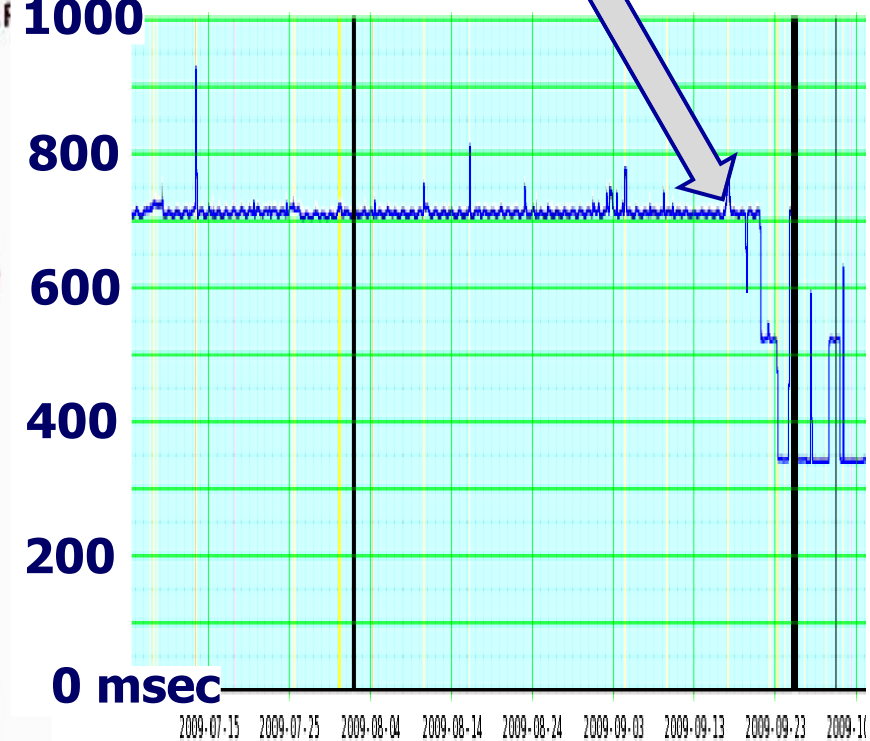
Maintaining access is manpower intensive

Round Trip Time (from SLAC) Drops
as African Nations Move from Geostationary Satellites to the New Undersea Cables

Minimum RTT (msec)



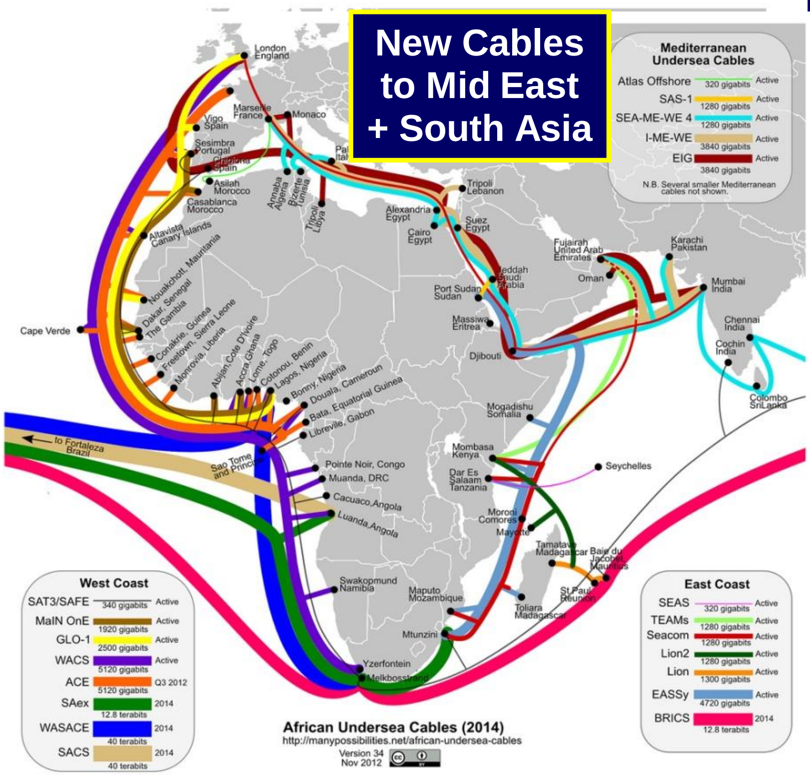
Rwanda: RTT shift from GEOS to Fiber



Lower RTT Tends *to Increase Performance*

R. Cottrell

New African Undersea Cables Following the 2010 World Cup: to Europe, India, Middle East



- Undersea cables continue to arrive at both African coasts (since 2009); 1000X Potential capacity
- To multi-Terabits/sec; 10X more by 2014
- Seacom, EASSy, TEAMS, Lion, Lion2, MainOne, GLO1, WACS *in production*
  - *+ ACE, BRICS, SAex, WASACE, SACS by 2014*
- Triggered by the 2010 World Cup.
- Connections to the African interior *spreading*
- Plus new Mediterranean Cables to Mideast+Gulf

http://manypossibilities.net/african-undersea-cables

More comprehensive map (with terrestrial fiber):
http://www.ubuntunet.net/sites/ubuntunet.net/files/Intra-Africa_Fibre_Map_v6.pdf

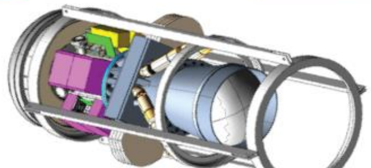
Seacom	EASSy	TEAMs	MainOne	WACS	GLO1	ACE	SAex	WASACE	BRICS
$ 650M	$ 265M	$ 130M	$ 240 M	$ 600M	$ 800 M	$ 700M			
13.7 kkm	10 kkm	4.5 kkm	7 kkm	14 kkm	9.5 kkm	14 kkm	9 kkm	9 kkm	34 kkm
1.28 Tbps	4.72 Tbps	1.28 Tbps	1.92 Tbps	3.84 Tbps	2.5 Tbps	5.12 Tbps	12.8 Tbps	40 Tbps	12.8 Tbps
Active 2009	Active 2010	Active 2009	Active 2010	Active 2012	Active 2010	2013	Q2 2013	2014	2014

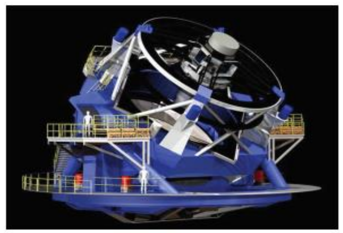
Big Data at the Cosmic Frontier
of Astrophysics and HEP

L. Bauerdick, Snowmass

A decade of data: DES to LSST

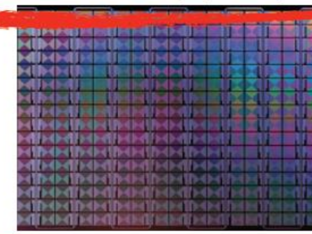
- Wide field and deep
  - DES: 5,000 sq degrees
  - LSST: 20,000 sq degrees
- Broad range of science
  - Dark energy, dark matter
  - Transient universe
- Timeline and data
  - 2012-16 (DES)
  - 2020 – 2030 (LSST)
  - 100TB - 1PB (DES)
  - 10PB - 100 PB (LSST)





Technology developments

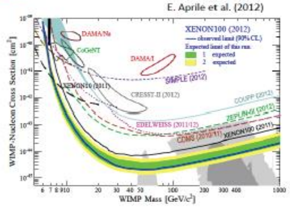
- Microwave Kinetic Inductance Detectors (MKIDs)
  - Energy resolving detectors (extended to optical and UV)
  - Resolving power: 30 < R < 150 (~5 nm resolution)
  - Coverage: 350nm – 1.3 microns
  - Count rate: few thousand counts/s
  - 32 spectral elements for uv/optical/ir photons



Growing volumes and complexity

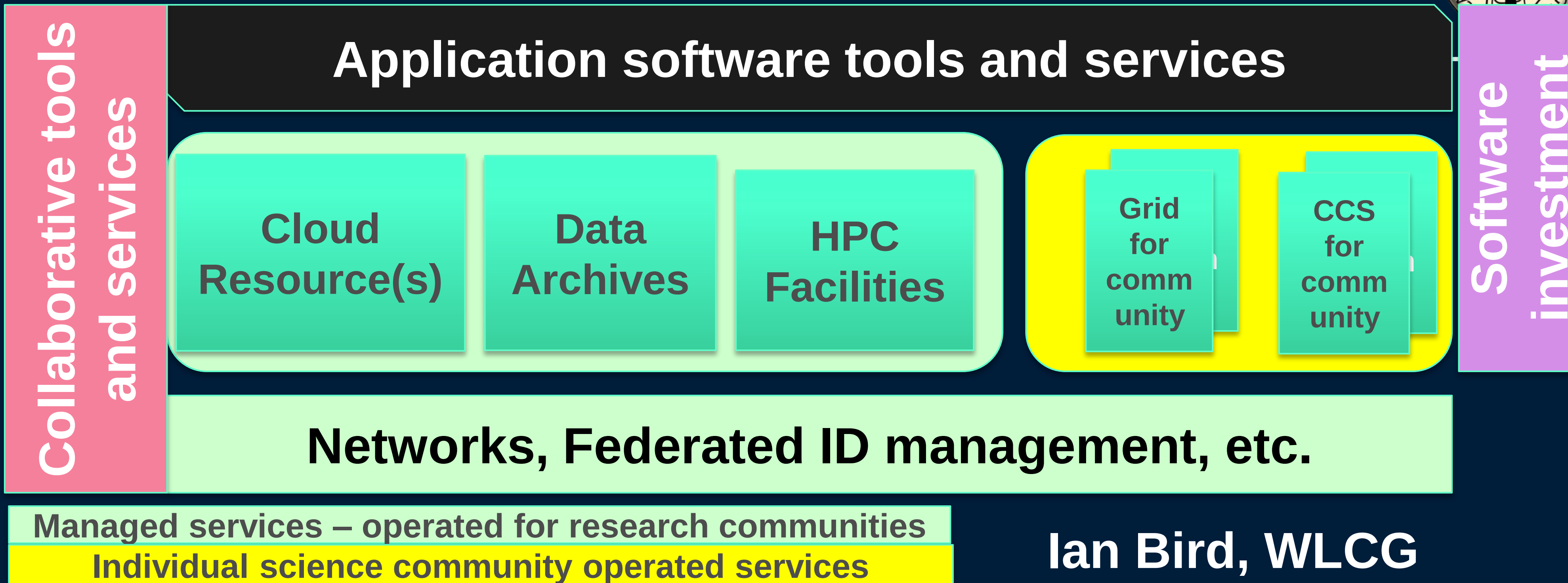
- CMB and radio cosmology
  - CMB-S4 experiment's 10^{15} samples (late-2020's)
  - Murchison Wide-Field array (2013-)
    - 15.8 GB/s processed to 400 MB/s
  - Square Kilometer Array (2020+)
    - PB/s to correlators to synthesize images
    - 300-1500 PB per year storage
- Direct dark matter detection
  - Order of magnitude larger detectors
  - G2 experiments will grow to PB in size

E. Aprile et al. (2012)



- From tabletop to cosmological surveys
  - Huge image data and catalogs
    - DES 2012-2016
      - 1PB images
      - 100TB catalog
    - LSST 2020-2030
      - 6PB images/yr, 100 PB total
      - 1PB catalogs, 20 PB total
  - large simulations

Future e-Infrastructure System for Science?



Ian Bird, WLCG

Key principles:

- Governed & driven by science/research communities
- **Business model: Operations should be self-sustaining:**
  - Managed services are paid by use (e.g. Cloud services, data archive services, …)
  - Community services operated by the community at their own cost using their own resources (e.g. grids, citizen cyberscience)
- **Software support:** open source, funded by collaborating developer institues

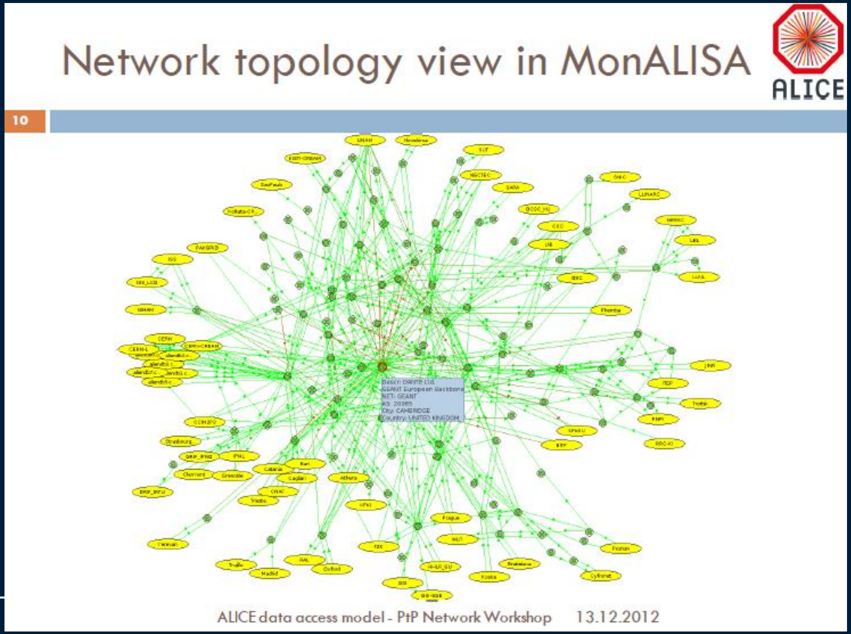
HEP Computing Circa 2020:
Possible Vision: A Global Data Intensive CDN

Ian Fisk, Snowmass

- A Global Content Delivery Network
  - Data management resources that deliver data on demand
  - Cached & replicated; intelligent about data placement + mobility
  - Large independent local storage systems connected to clusters is probably not the most efficient scheme
  - The data federations already being deployed are a first step, but more work – and system development – is needed
- Dynamic data delivery systems of this kind give a lot of flexibility in how to make use of diverse computing systems
  - But put strong requirements on network capacity + capability
- While a 10k core cluster typical for 2020 will require 10 Gbps (or more) of networking for organized processing
- End users doing Analysis would require ~ 100 Gbps for rapid delivery of multi-Terabyte "Small" datasets
- Hundreds of such end users will present a challenge: *also to the next-generation networks of 2020*

Components for a working system (III)

- Monitoring: PerfSONAR and MonALISA
- All LHCOPN and many LHCONE sites have PerfSONAR deployed
  - Goal is to have all LHCONE instrumented for PerfSONAR measurement
- Regularly scheduled tests between configured pairs of end-points:
  - Latency (one way)
  - Bandwidth
- Currently used to construct a dashboard
- Could provide input to algorithms developed in ANSE for PhEDEx and PanDA
- ALICE and CMS experiments are using MonALISA monitoring framework
  - accurate bandwidth availability
  - complete topology view



Monitoring the Worldwide LHC Grid

State of the Art Technologies Developed at Caltech

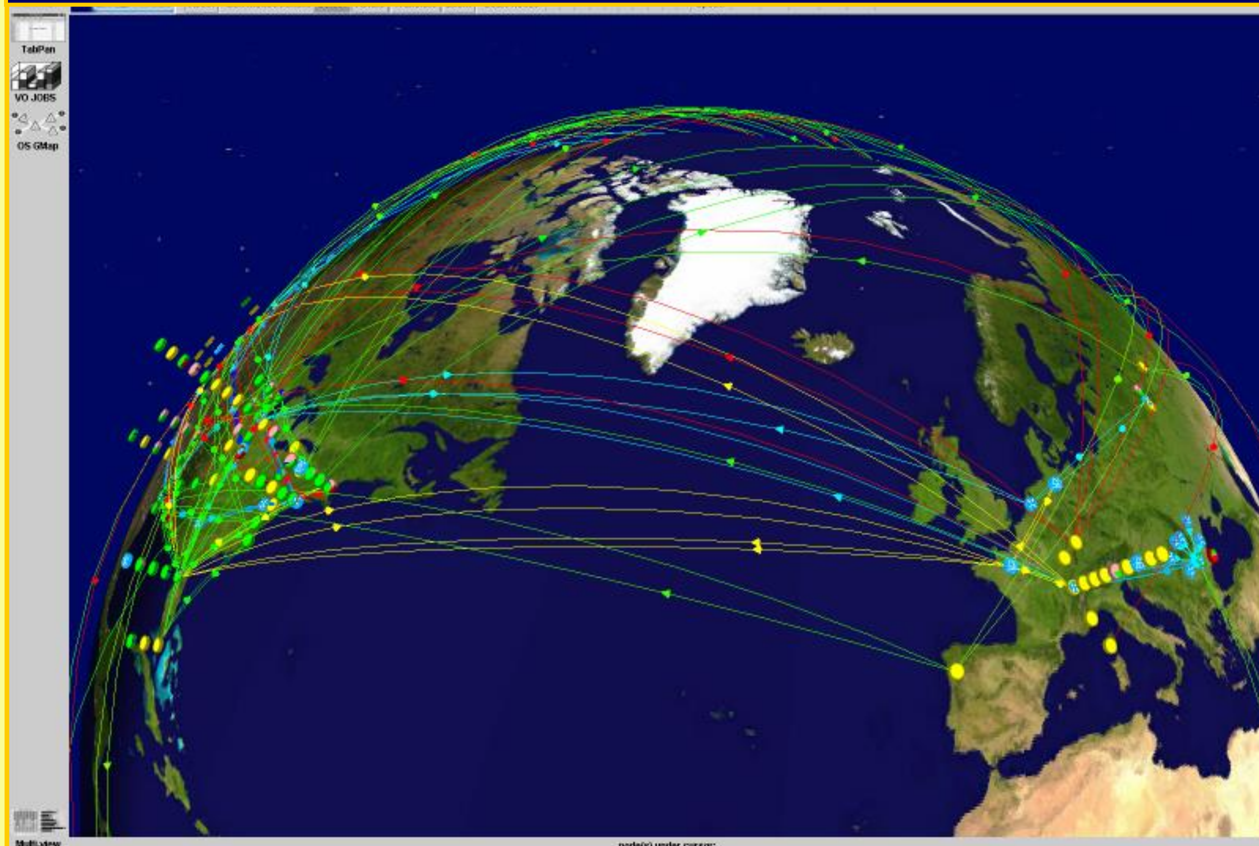
MonALISA Today

Running 24 X 7 at 380 Sites

- Monitoring
  - 40,000 computers
  - \> 100 Links On Major Research and Education Networks
- Using Intelligent Agents
- Tens of Thousands of Grid jobs running concurrently
- Collecting > 4M parameters in real-time

MonALISA: Monitoring Agents in a Large Integrated Services Architecture

A Global Autonomous Realtime System



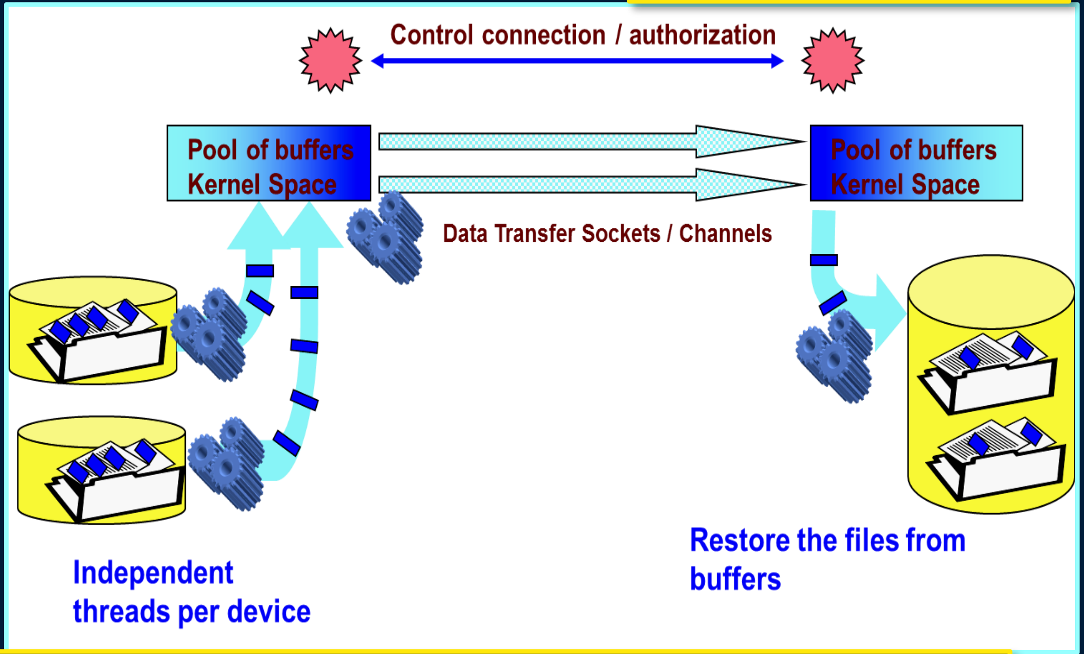
World expertise in high data throughput over long range networks

Fast Data Transfer (FDT)

http://monalisa.caltech.edu/FDT

- FDT is an open source Java application for efficient data transfers
- Easy to use: similar syntax with SCP, iperf/netperf
- Based on an asynchronous, multithreaded system
- Uses the New I/O (NIO) interface and is able to:
  - Decompose/Stream/Restore any list of files
  - Use independent threads to read and write on each physical device
  - Transfer data in parallel on multiple TCP streams, when necessary
  - Use appropriate size of buffers for disk IO and networking
  - Resume a file transfer

FDT uses IDC API to request dynamic circuit connections



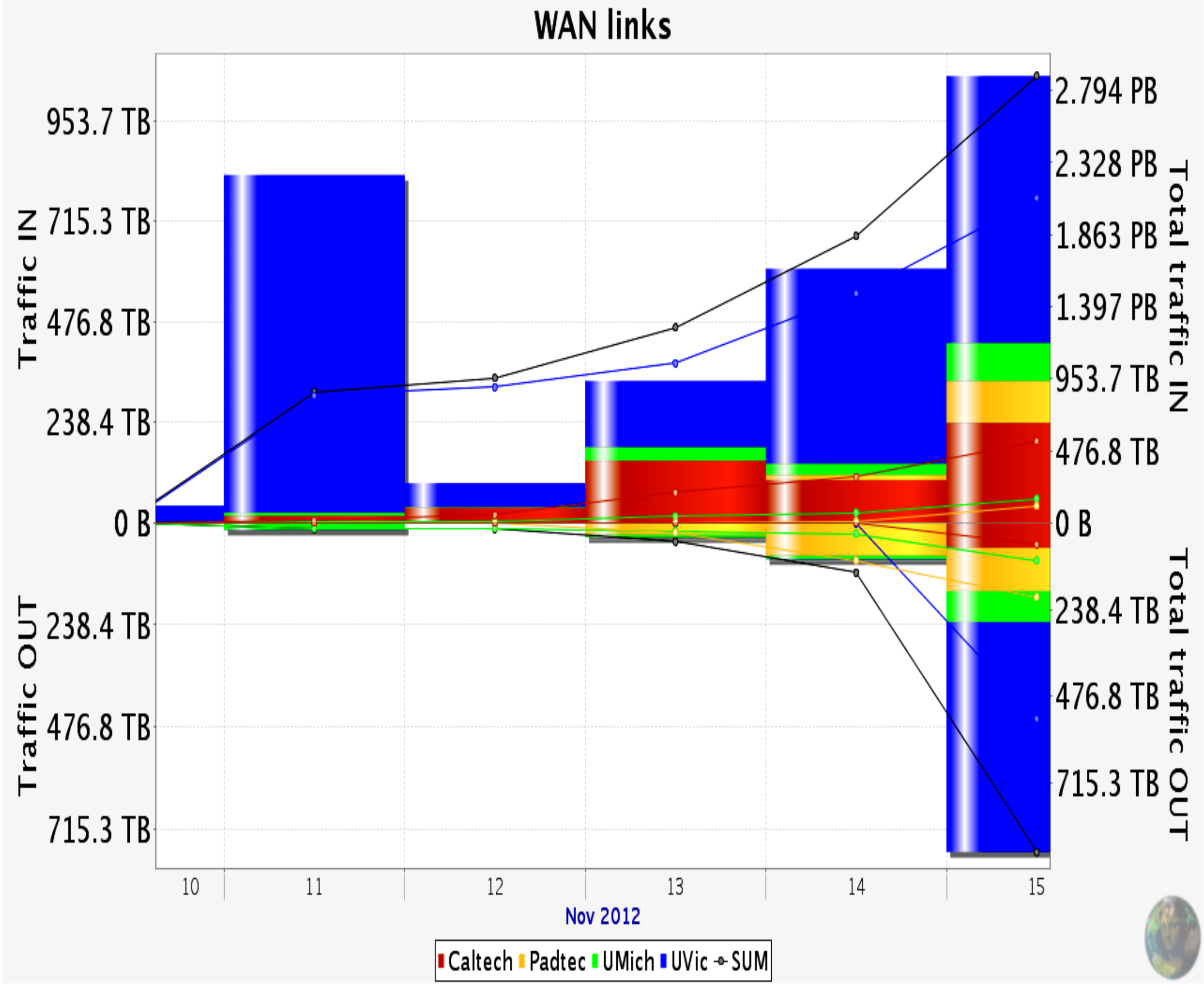
Open source TCP-based Java application; the state of the art since 2006

Transferring Petabytes at SC12

FDT and RDMA over Ethernet

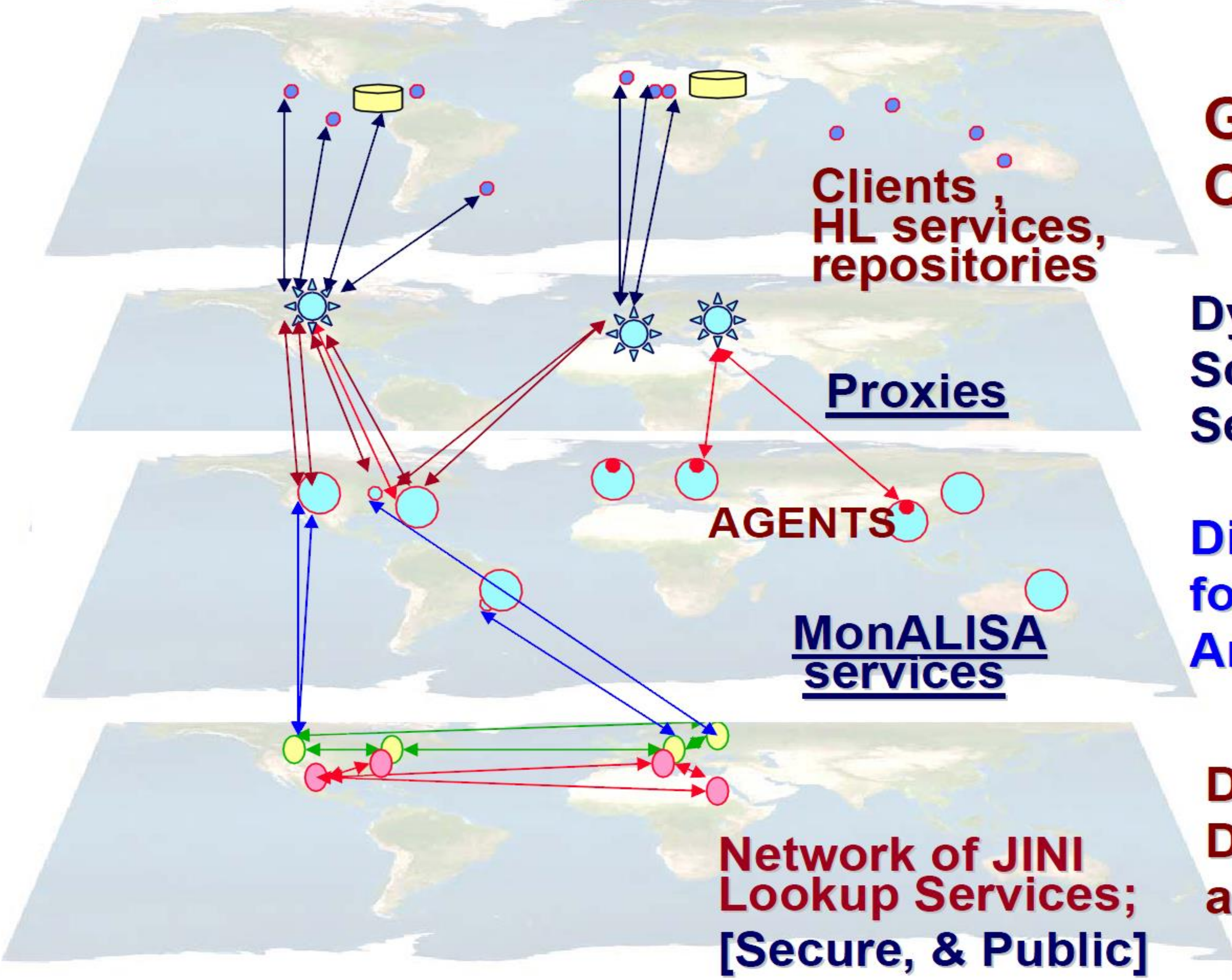
3.8 PBytes to and From the Caltech Booth

Including 2 PBytes on 11/15



MonALISA : An Agent-based System of Distributed Services

Fully Distributed System with no Single Point of Failure



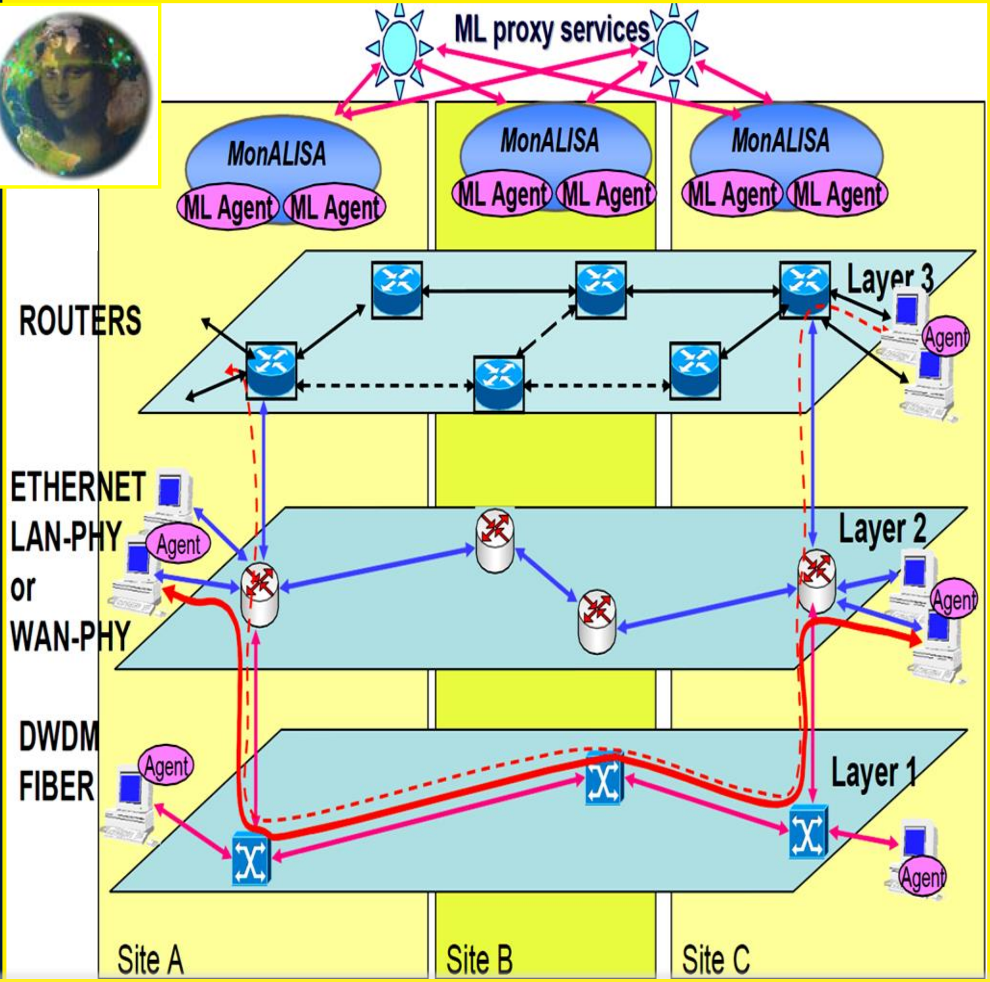
Global Services or Clients

Dynamic load balancing
Scalability & Replication
Security AAA for Clients

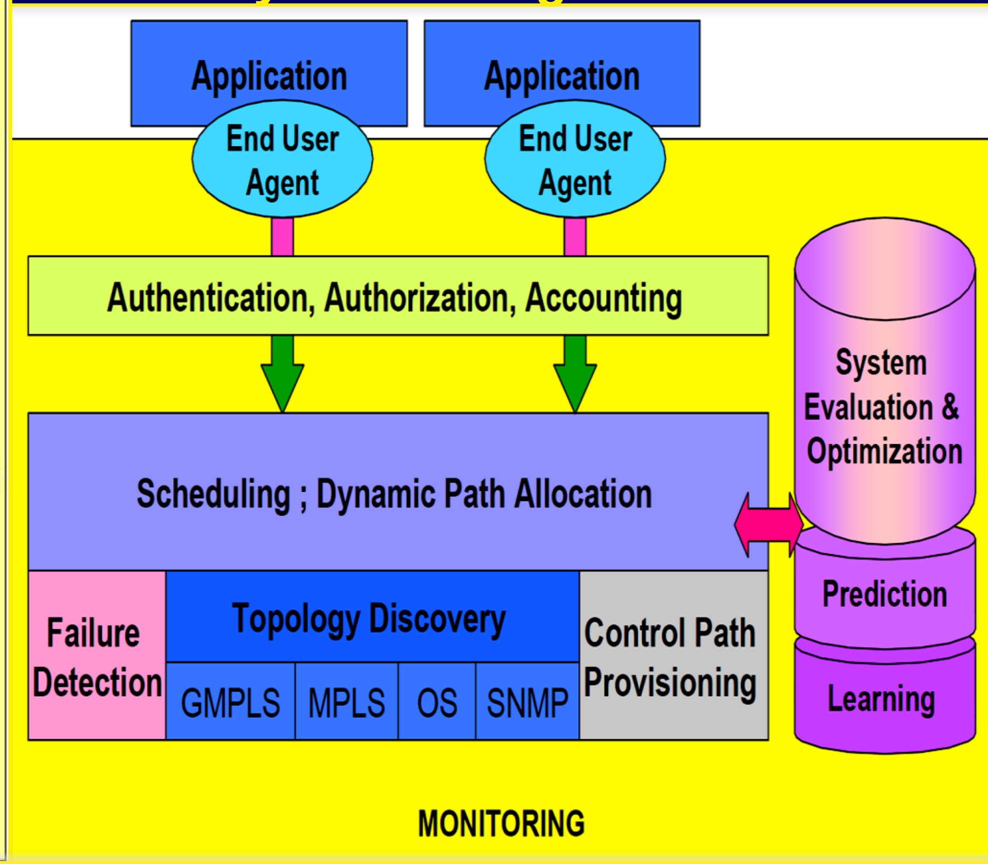
Distributed System for gathering and Analyzing Information.

Distributed Dynamic Discovery- based on a lease Mechanism

VINCI: Virtual Intelligent Networks for Computing Infrastructures



Core Concepts and Real Time System Design: 2005-6



http://monalisa.caltech.edu

VINCI (CHEP06, Mumbai)

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

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