A New Generation of Networks and Grid Computing Models for HEP in the LHC Era

Harvey B Newman

California Institute of Technology

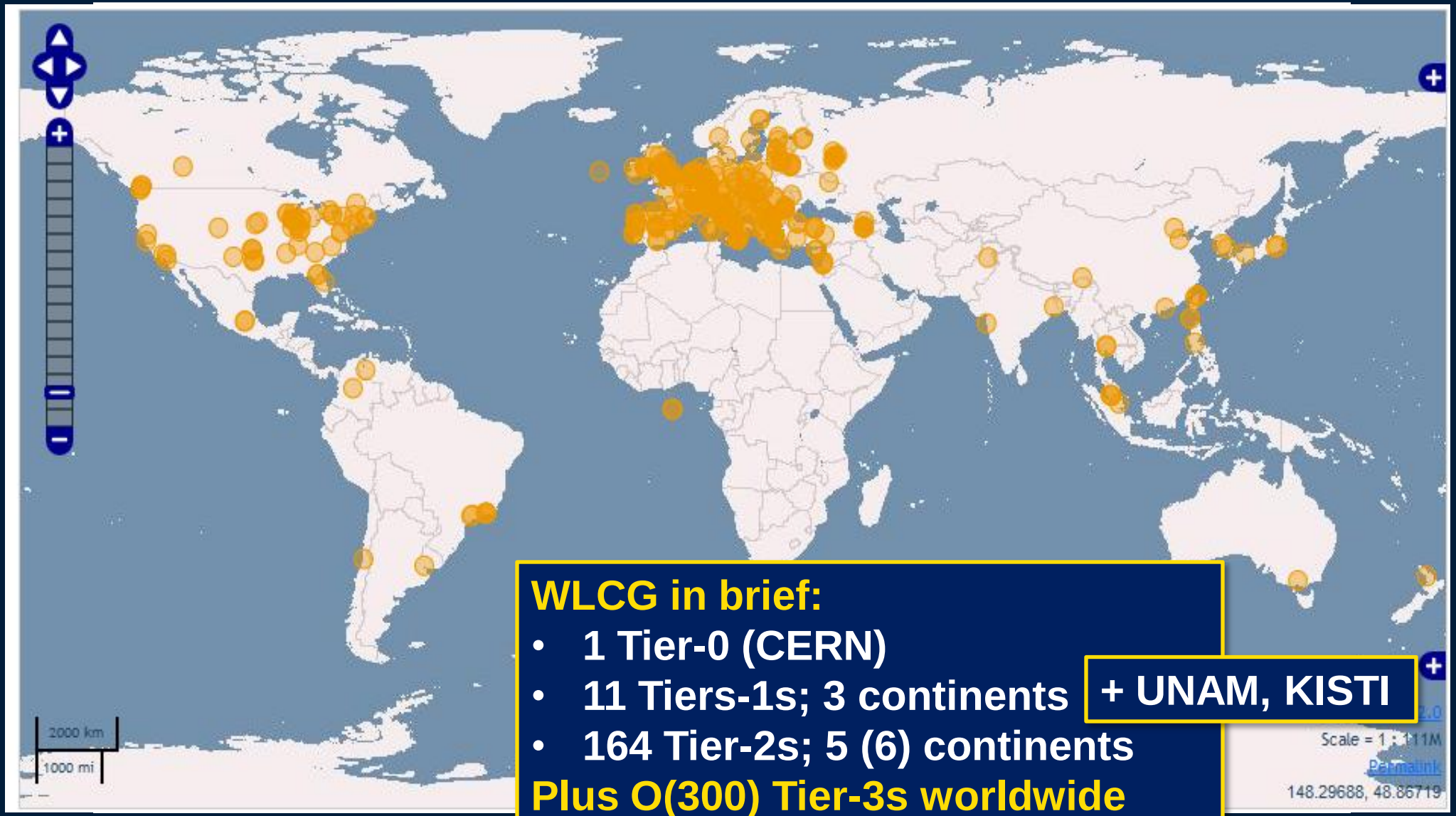
Grid of the Americas Workshop
ICN-UNAM, February 9, 2011

OUTLINE

- Introduction
- The Evolving Network Requirements
- Network Requirements WG
- Emergence of New LHC Computing Models
- LHCONE: A New Architecture of Open Exchange Points
- UNAM and CUDI in LHCONE

Bandwidth Evolution and Capacity Planning
Transition to 40G and 100G Next Generation Technologies
Advances in Data Transfer Applications
Closing the Digital Divide

LHC Computing Infrastructure



WLCG in brief:

- 1 Tier-0 (CERN)
- 11 Tiers-1s; 3 continents
- 164 Tier-2s; 5 (6) continents

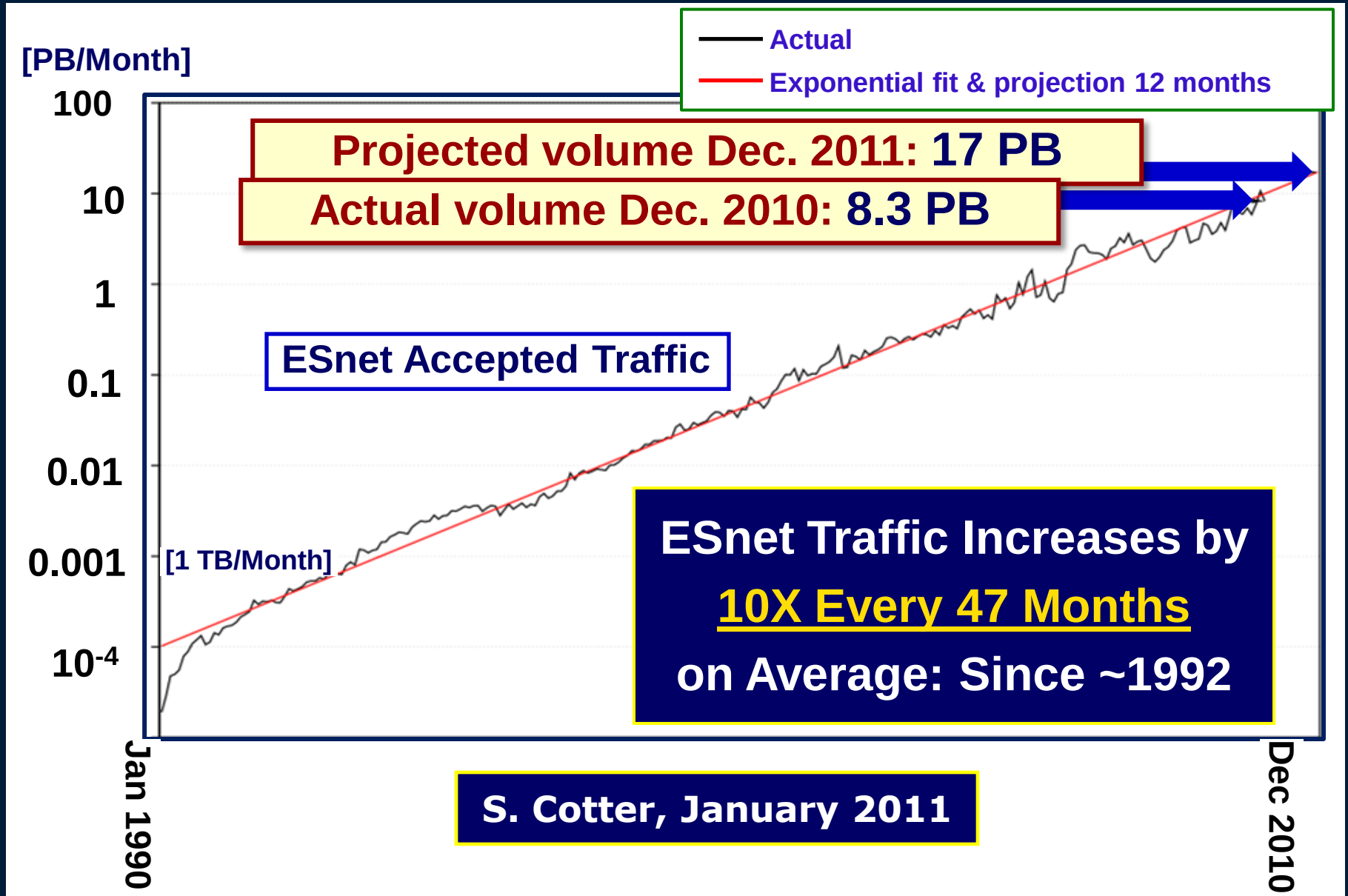
Plus O(300) Tier-3s worldwide

+ UNAM, KISTI

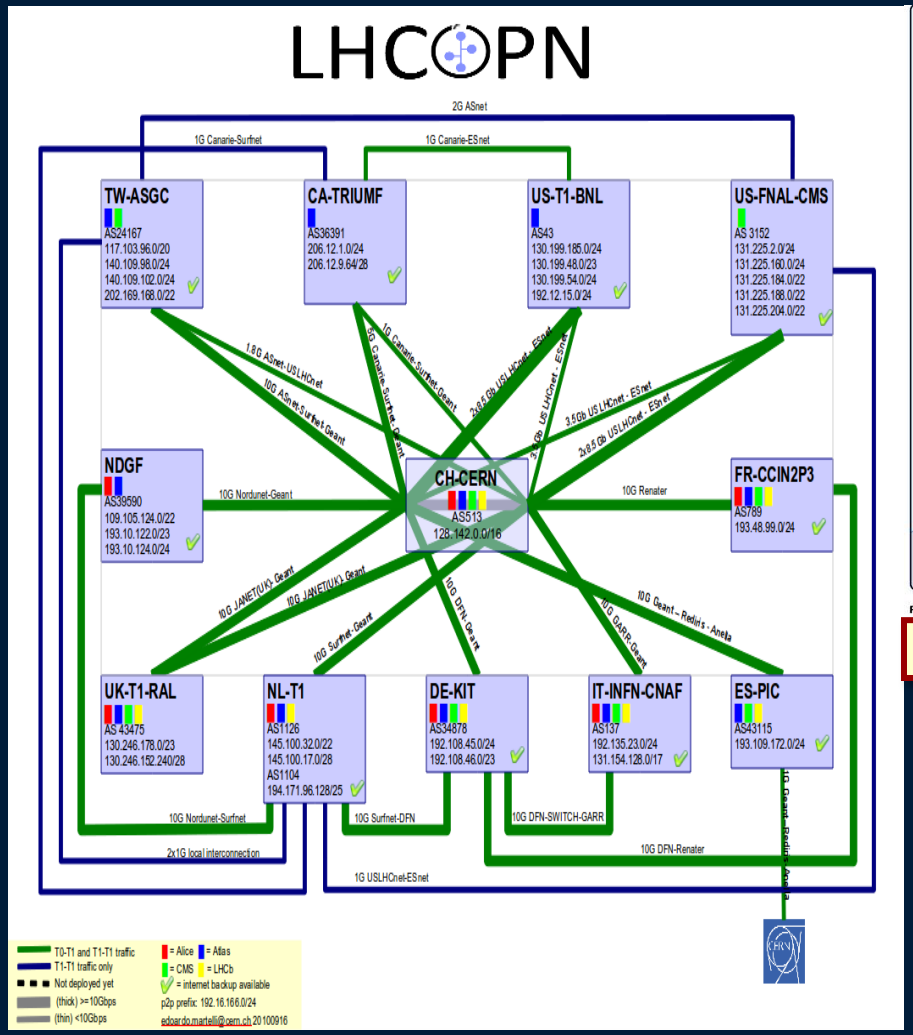
Networking in the LHC Era

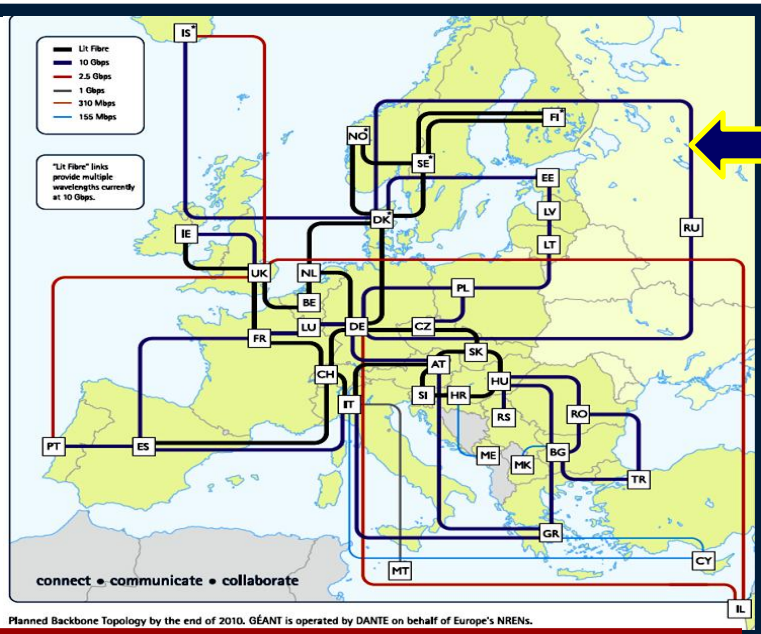
- HEP's reliance on long range networks continues its 30 year trajectory, marked by:
- An exponential growth in capacity
  - 10X in usage every 47 months in Esnet, over 18 years
  - 6 X 10^6 times capacity growth over 25 years across the Atlantic (LEP3Net in 1985 to US LHCNet in 2010)
- New technology generations & standards each few years
  - The transition from 10G to 40G (2011-12) and 100G (2011-14) are the next steps
  - Along with the first set of standards integrating optical transport (ITU OTN hierarchy) and Ethernet (IEEE 802.3)
- A sustained ability to use ever-larger continental and transoceanic networks effectively: high throughput transfers
- HEP as a driver of R&E and mission-oriented networks

ESnet: Continued Exponential Growth

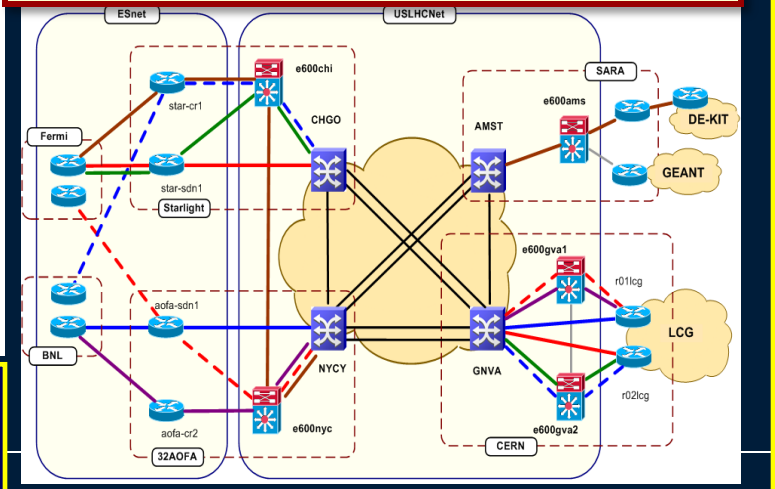


The Core of LHC Networking: LHCOPN and Partners





US LHCNet and ESnet



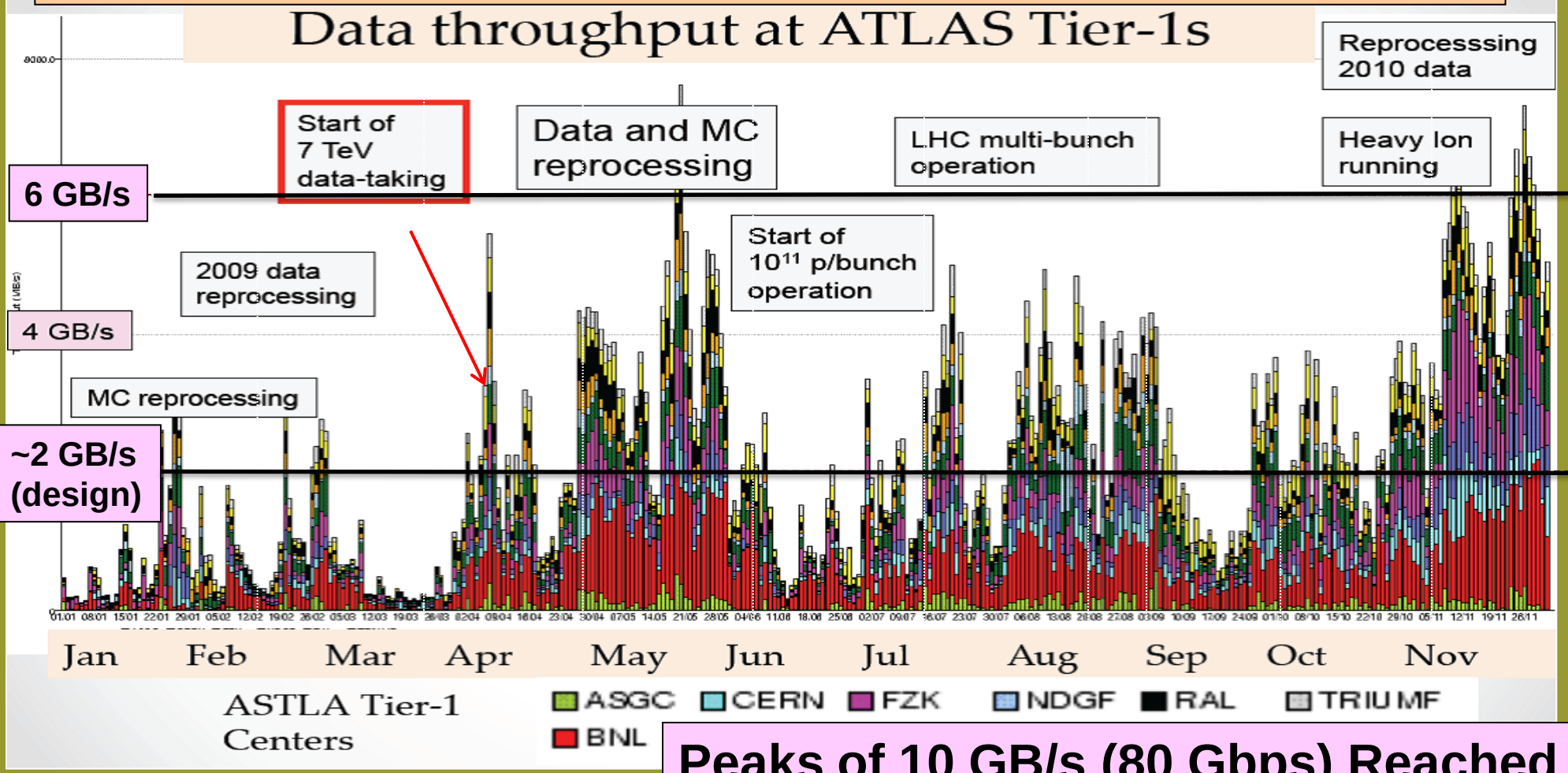
Dark Fiber Core Among 19 Countries:

- 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

Worldwide data distribution and analysis (F.Gianotti)

Total throughput of ATLAS data through the Grid: 1st January → November.



Peaks of 10 GB/s (80 Gbps) Reached

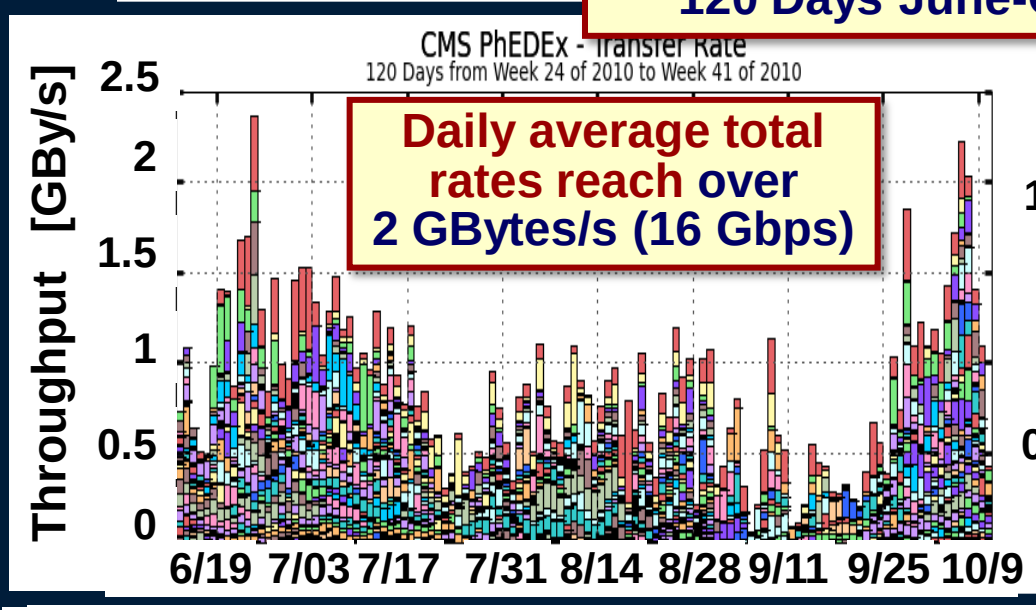
Grid-based analysis in Summer 2010: >1000 different users; >15M analysis jobs

The excellent Grid performance has been crucial for fast release of physics results.
E.g. ICHEP: Full data sample taken until Monday was shown at conference Friday

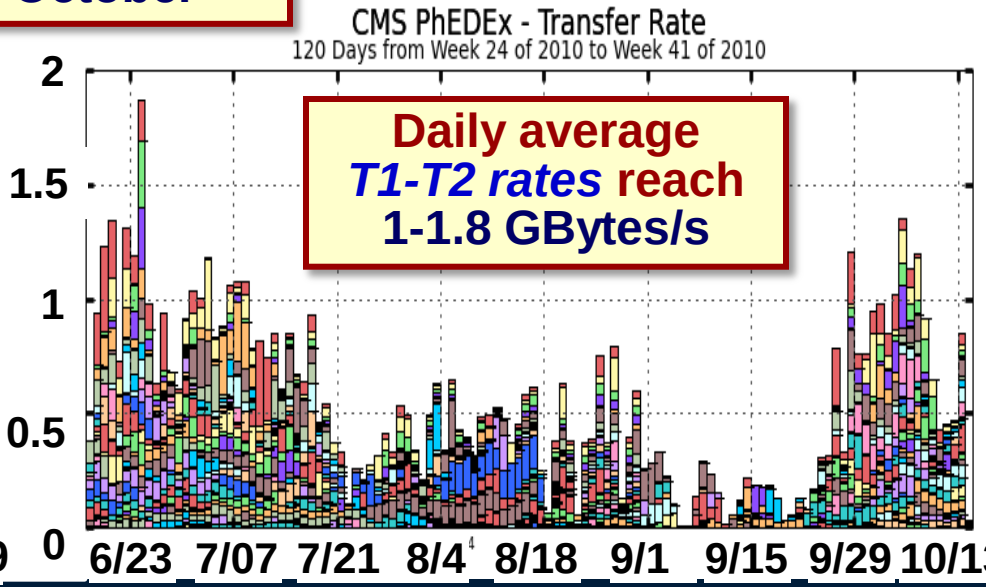
CMS Data Movements

(All Sites and Tier1-Tier2)

120 Days June-October

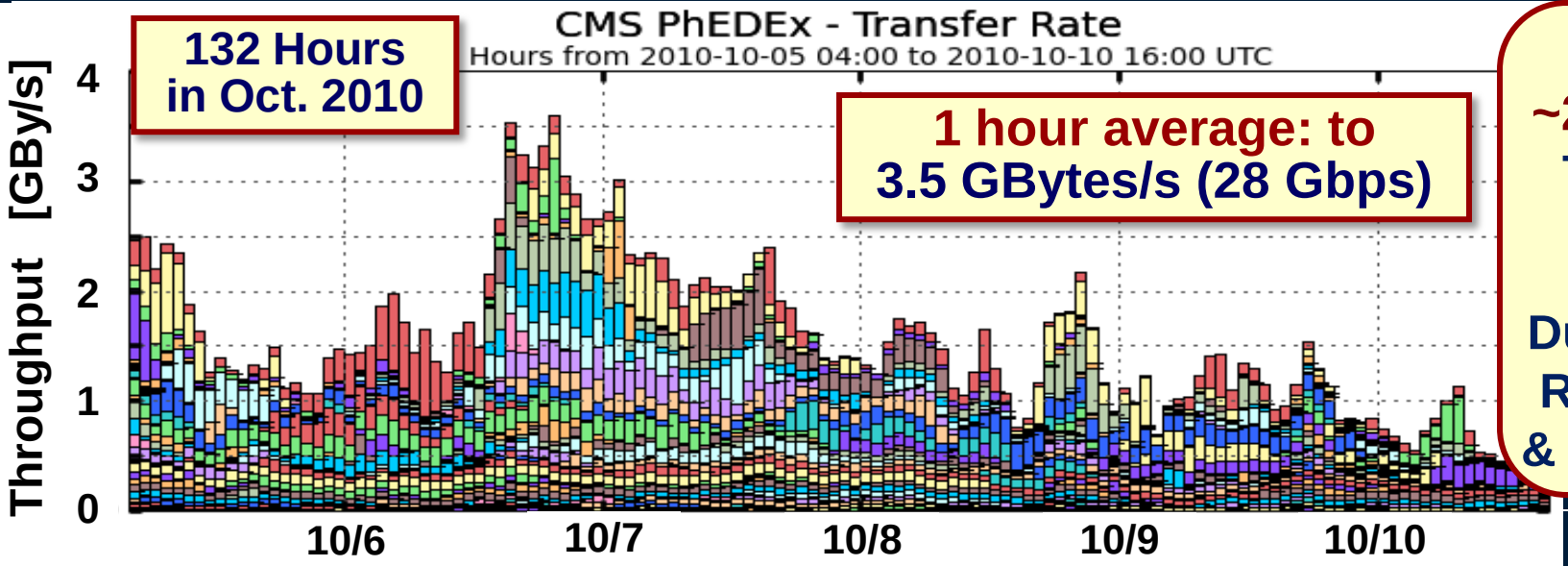


Daily average total rates reach over 2 GBytes/s (16 Gbps)



Daily average *T1-T2 rates* reach 1-1.8 GBytes/s

132 Hours in Oct. 2010



1 hour average: to 3.5 GBytes/s (28 Gbps)

Tier2-Tier2 ~25% of Tier1-Tier2 Traffic

To ~50% During Dataset Reprocessing & Repopulation

Networking in the LHC Era

- The LHC experiments, with their distributed Computing Models and worldwide involvement in LHC physics, have brought a renewed focus on networks
- The prospect of discoveries in the 2011-12 run, has brought a renewed emphasis on both network bandwidth and "reliability"
  - The service uptime goal of 99.95% has been set, and achieved: through the implementation of resilient network infrastructures
- Reliability of the networks in 1st months of running at 7 TeV has been highlighted at ICHEP as a major element in the LHC program's success
- This has given the experiment the confidence to seek more agile and effective Models of data distribution and/or remote access
  - To harness the efforts of physicists worldwide in pursuit of discoveries at the LHC, and to increase their competitiveness
  - Bringing new physics opportunities, and also *new challenges to the worldwide network infrastructures* supporting the LHC program

This also means we must continue to address the Digital Divide in many world regions, as the rate of progress in the developed world accelerates

Workshop on Transatlantic Connectivity CERN 2010

- A workshop on transatlantic connectivity was held 10-11 June 2010 at CERN
  - http://indico.cern.ch/conferenceDisplay.py?confId=88883
  - ~50 Participants representing the major stakeholders in R&E networking
    - ESnet, Internet2, GEANT, NRENs, NSF, DOE, Industry, Major Labs etc
- And Revealed the following:
  - Flows are already larger than foreseen in the LHC program, even at the lower luminosities seen in Spring 2010
  - Some Tier2's are very large (not new), and growing larger
    - All US ATLAS and US CMS T2's have 10G capability; some larger.
  - Some Tier1-Tier2 flows are quite large (several to 10Gbps)
  - Tier2-Tier2 data flows are also starting to be quite significant.
  - The vision progressively moves away from all-hierarchical models towards peer-to-peer
    - True for both CMS and ATLAS
    - For reasons of reduced latency, increased working efficiency, agility
- Expectations of network capability are reaching unrealistic proportions without forward planning.

David Foster, CERN-IT ; Harvey Newman, Caltech

LHC Experiments' Future Networking Requirements Working Group

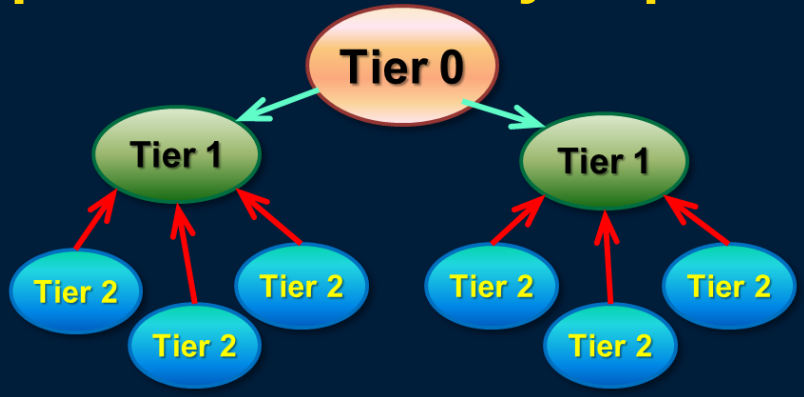
- A Requirements Working Group was formed in June 2010 among the experiments and network providers, to investigate and then respond to future network requirements
  - Following the Workshop on Transatlantic Networking for the LHC Experiments
- Harvey Newman (US LHCNet) Bill Johnston (ESnet)
  Jerry Sobieski (NORDunet) Klaus Ullmann (DFN, DANTE)
  David Foster (CERN) Ian Fisk (CMS)
  Kors Bos (ATLAS, Chair) Artur Barczyk (US LHCNet)
  Eric Boyd (Internet2)
- The new data and computing models incorporate greater reliance on network performance
  - Will rely more on network infrastructure bandwidth & robustness
- Requirements need to be based on a complete operational model that includes *all* significant network flows

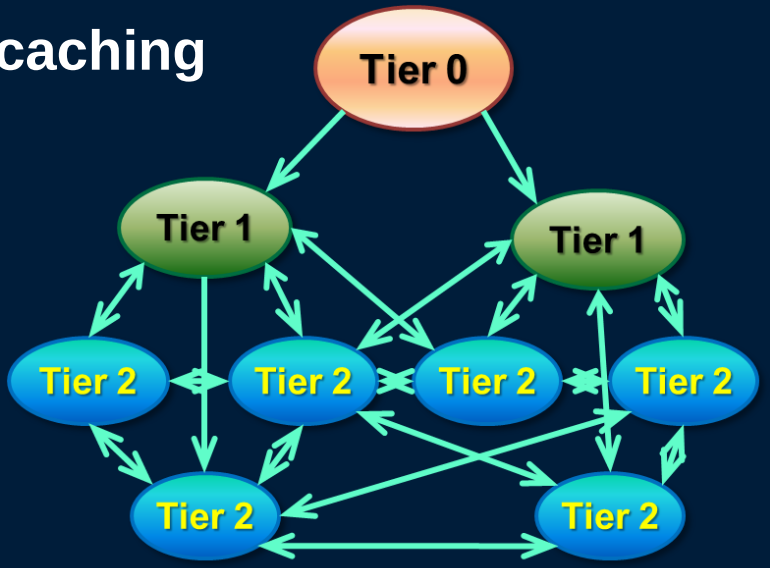
Findings: Three Levels of Tier2 Throughput Requirements

- Minimal Tier2: 1 Gbps Throughput ➡ 2 Gbps Provisioned
  - Mainly MC production; functions but not flexible; no "QoS"
- Nominal Tier2: 5 Gbps Throughput ➡ 10 Gbps Provisioned
  - Samples updated in reasonable time; Tier2 storage updated regularly
- Leadership Tier2: 10+ Gbps Throughput ➡ to ~20 Gbps Provisioned
  - Substantial analysis facilities supporting large numbers of users
  - Large local storage updated frequently; datasets provided to other Tier2s
- It is expected that Tier2s will move from minimal to nominal and from nominal to leadership over time
- All categories on an increasing scale: ~2X / 2Yrs for Nominal & Leadership; ~2X per year for Minimal
- Costs have to be understood; including future evolution
  - Networking requirements need to be included in budgets
- Major LHCOPN players have been tasked with developing the architectural design and operational plan to meet these needs

Changing LHC Data Models

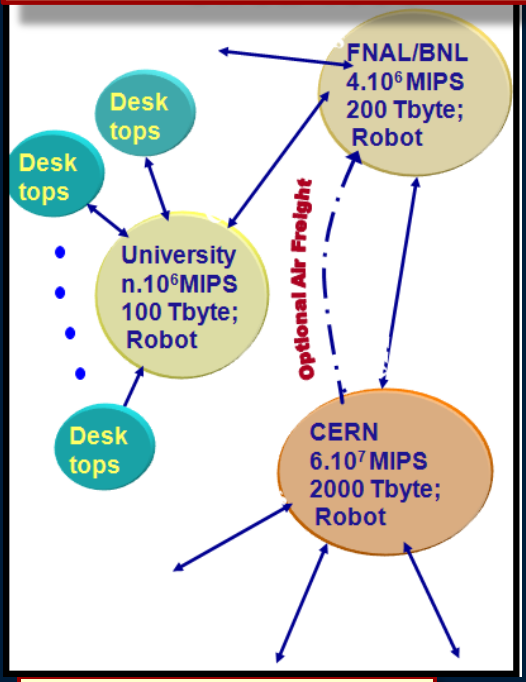
- 3 Recurring Themes:
  - **Flat(ter) hierarchy:** Any site might in the future pull data from any other site hosting it.
  - Data caching: Analysis sites will **pull datasets** from other sites "on demand", including from Tier2s in other regions
    - Possibly in combination with strategic pre-placement of datasets
  - **Remote data access:** jobs executing locally, using data cached at a remote site in quasi-real time
    - Possibly in combination with local caching
- Expect variations by experiment



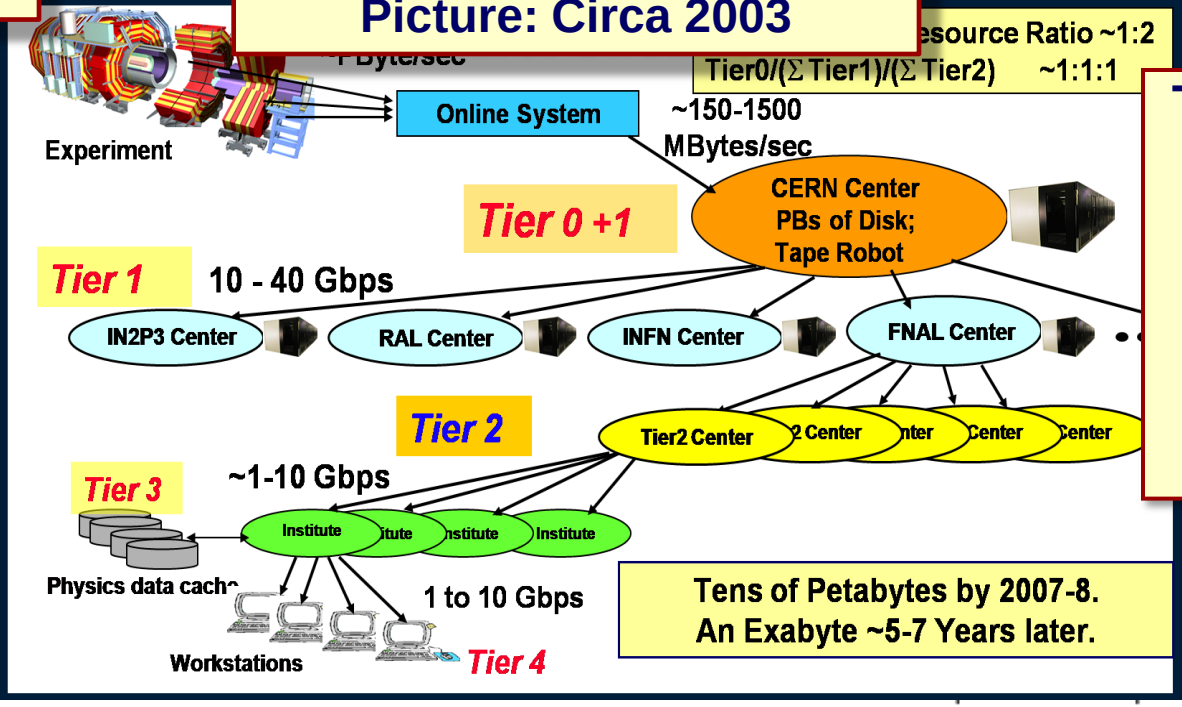


Experiments' Data Models

Circa 1996



The Evolving MONARC Picture: Circa 2003

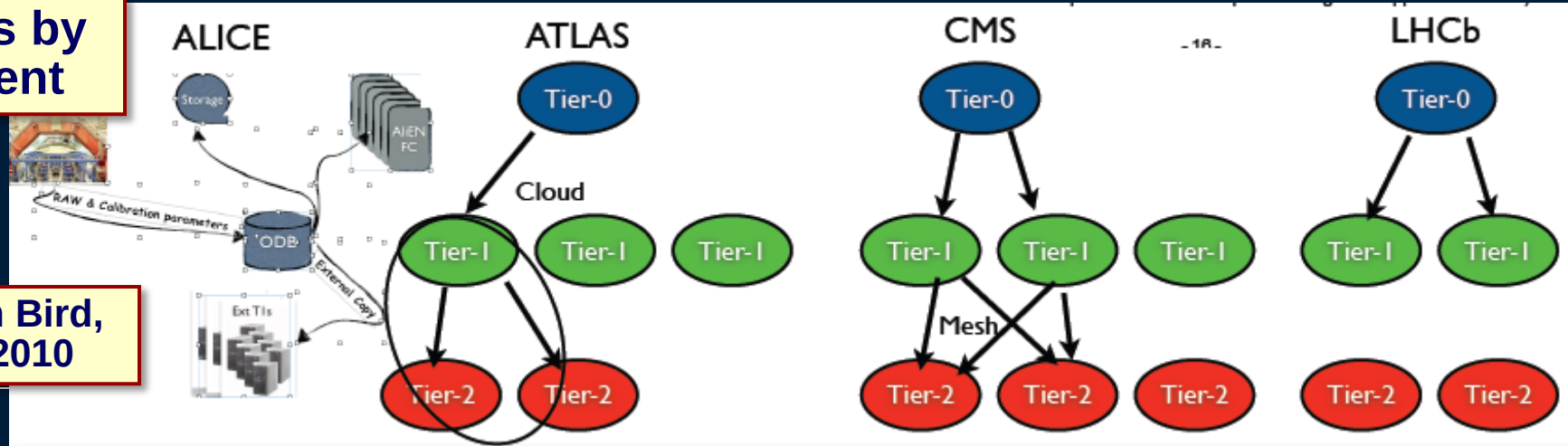


The models are based on the MONARC model

Now 10+ years old

Tens of Petabytes by 2007-8.
An Exabyte ~5-7 Years later.

Variations by experiment



From Ian Bird, CHEP 2010

The Changing LHC Computing Models

Ian Bird, CHEP conference, Oct 2010

Evolution of data placement

Move towards caching of data rather than strict planned placement

- Download the data when required
  - Selects popular datasets automatically
  - When datasets no longer used will be replaced in the caches
- Data sources can be any (Tier 0, 1, 2)
- Can still do some level of intelligent pre-placement
- Understanding a distributed system built on unreliable and asynchronous components means
  - Accepting that catalogues may be not fully updated
  - Data may not be where you thought it was
  - Thus must allow remote access to data (either by caching on demand and/or by remote file access)

The Changing LHC Computing Models

Ian Bird, CHEP conference, Oct 2010

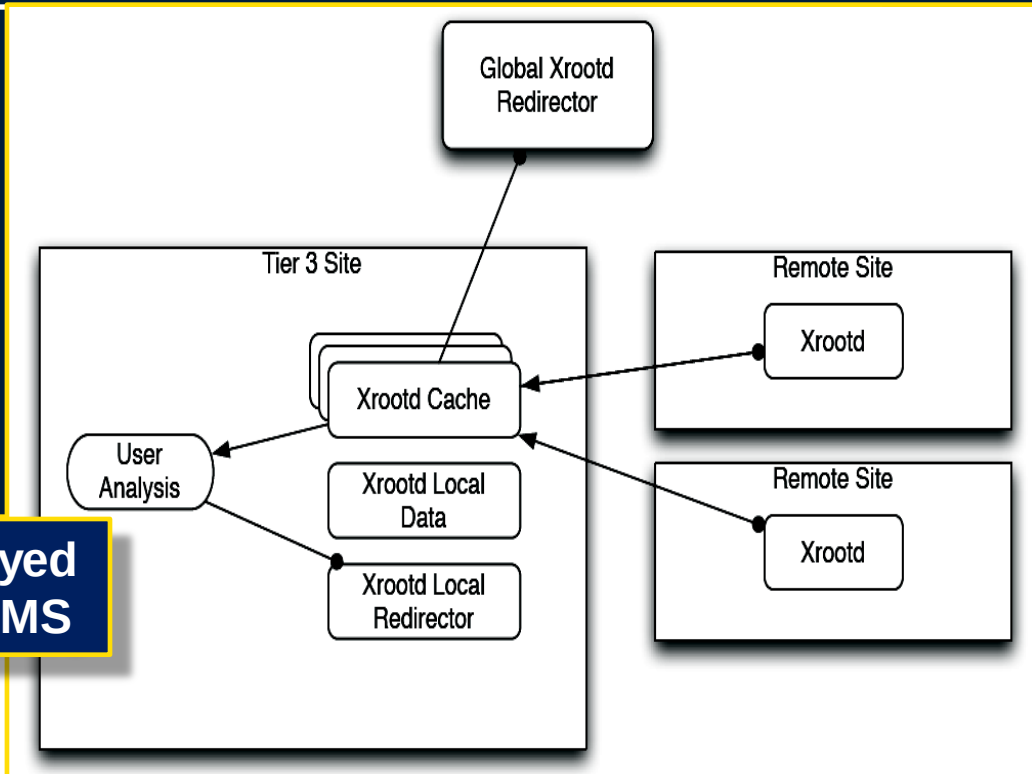
Implications for networks

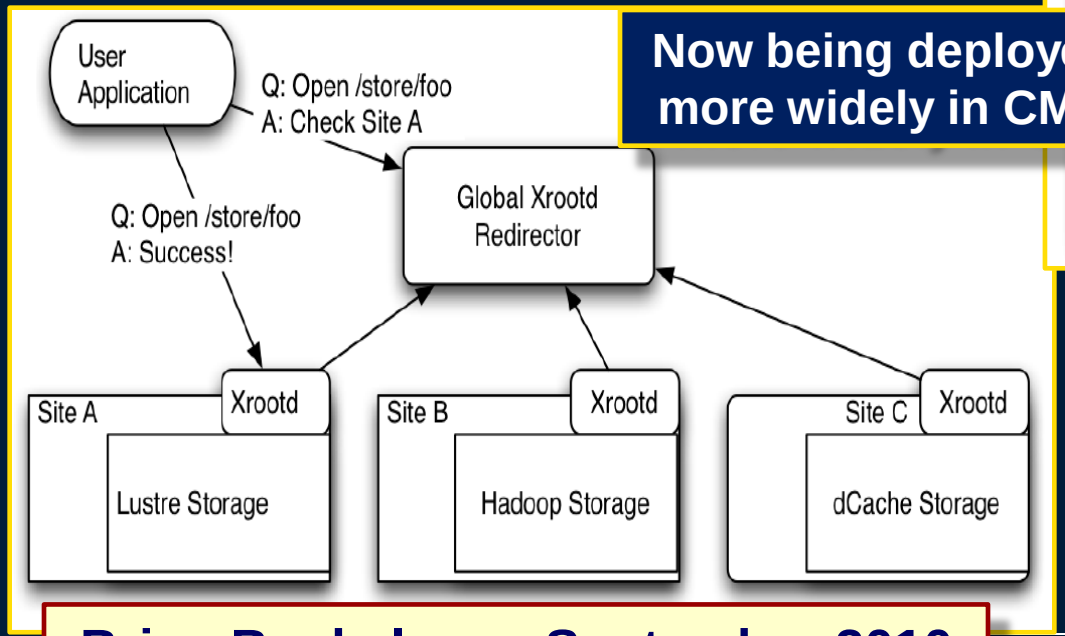
- Hierarchy of Tier 0, 1, 2 no longer so important
- Tier 1 and Tier 2 may become more equivalent for the network
- Traffic could flow more between countries as well as within (already the case for CMS)
- Network bandwidth (rather than disk) will need to scale more with users and data volumes
- Data placement will be driven by demand for analysis and not pre-placement

Ian Bird, CHEP conference, Oct 2010

Remote Data Access with Local Caching and Processing with Xrootd (CMS)

- Useful for smaller sites with less data storage
- Only selected objects are read (with object read-ahead).
  No transfer of entire data sets
- CMS demonstrator: Omaha diskless Tier3, served data from Caltech and Nebraska (Xrootd)





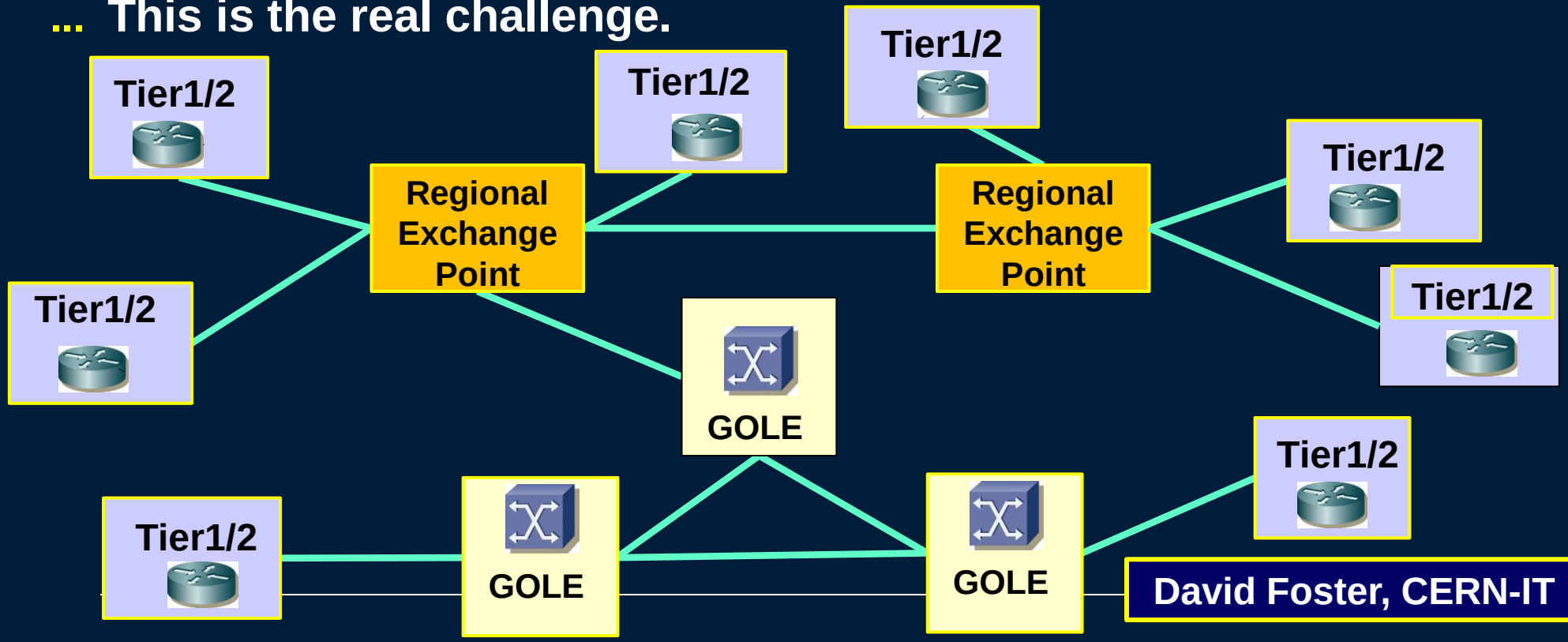
Now being deployed more widely in CMS

Similar operations in ALICE for years

Brian Bockelman, September 2010

A Possible Future:
An Infrastructure of Infrastructures

- Many players: LHCOPN, R&E Networks, CBDF Links + Commercial Links; Domains Interconnected through "Open Light Path Exchanges" (GOLE)
- No Central funding: So organic growth based on need and capability is essential.
- Federated + Open, engaging all parties & using all opportunities to connect
- The devil is in the details: Funding, Interoperability, Coord. Operations, etc. ... This is the real challenge.



SOLUTION PROPOSAL
LHCONE
NOW UNDER DEVELOPMENT

Design Inputs

- **Given the scale, geographical distribution and diversity of the sites as well as funding, only a federated solution is feasible**
- **The current LHC OPN is not modified**
  - OPN will become part of a larger whole
  - Some purely Tier2/Tier3 operations
- **Architecture has to be Open and Scalable**
  - Scalability in bandwidth, extent and scope
- **A resilient core is required; allow resilient edge-connections**
- **Bandwidth guarantees are required → determinism**
  - End-to-end systems approach
  - Reward effective use
- **Operation at Layer 2 (Switching) and below (Optical)**
  - Advantage in performance, costs, power consumption

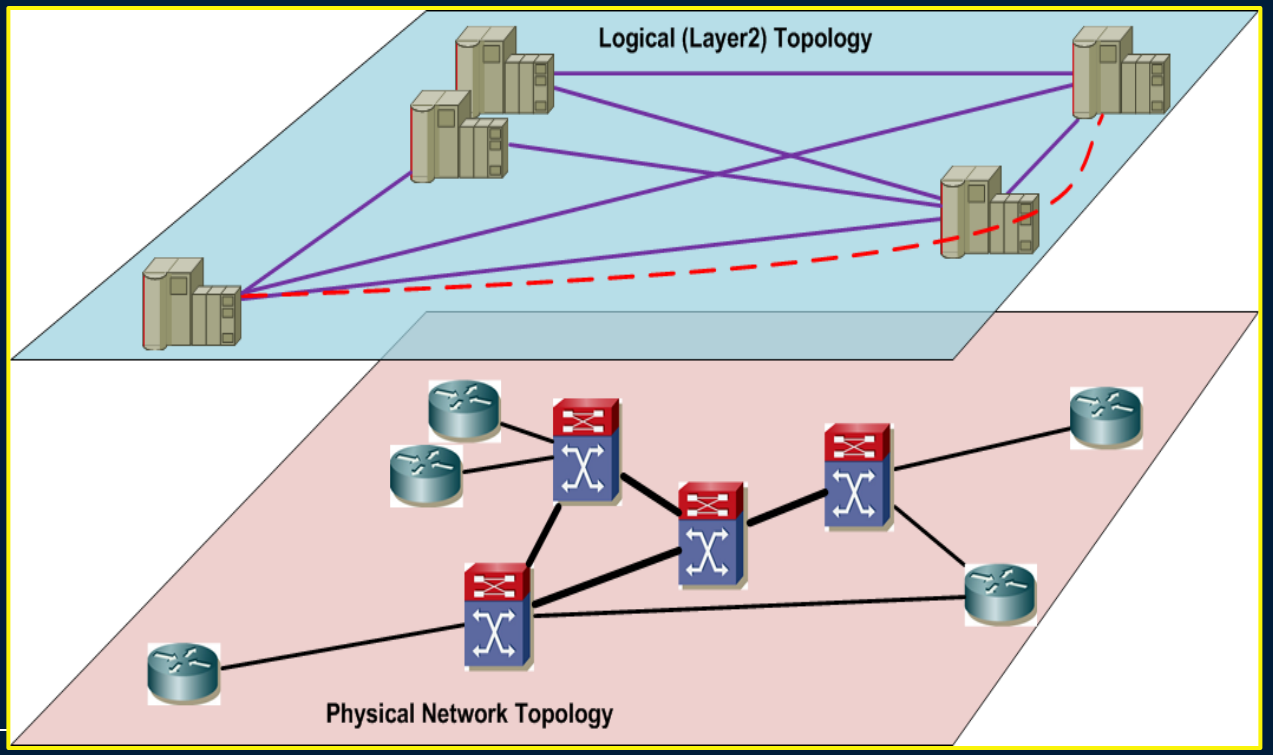
Lessons learned

- **The LHC OPN has proven itself; We shall learn from it**
- **Simple architecture**
  - Point-to-point Layer 2 circuits
  - Flexible and scalable topology
- **Grew Organically**
  - From star to partial mesh
  - Open to several technology choices
    - each of which satisfies requirements
- **Federated Governance Model**
  - Coordination between stakeholders
  - No single administrative body required
  - Made extensions and funding straight-forward
- **Remaining Challenge: monitoring and reporting**
  - More of a systems approach

➡ ***NB: Solved in ALICE and US LHCNet by MonALISA***

LHCONE: To Better Serve Tier1s, 2s and 3s in the LHC Era

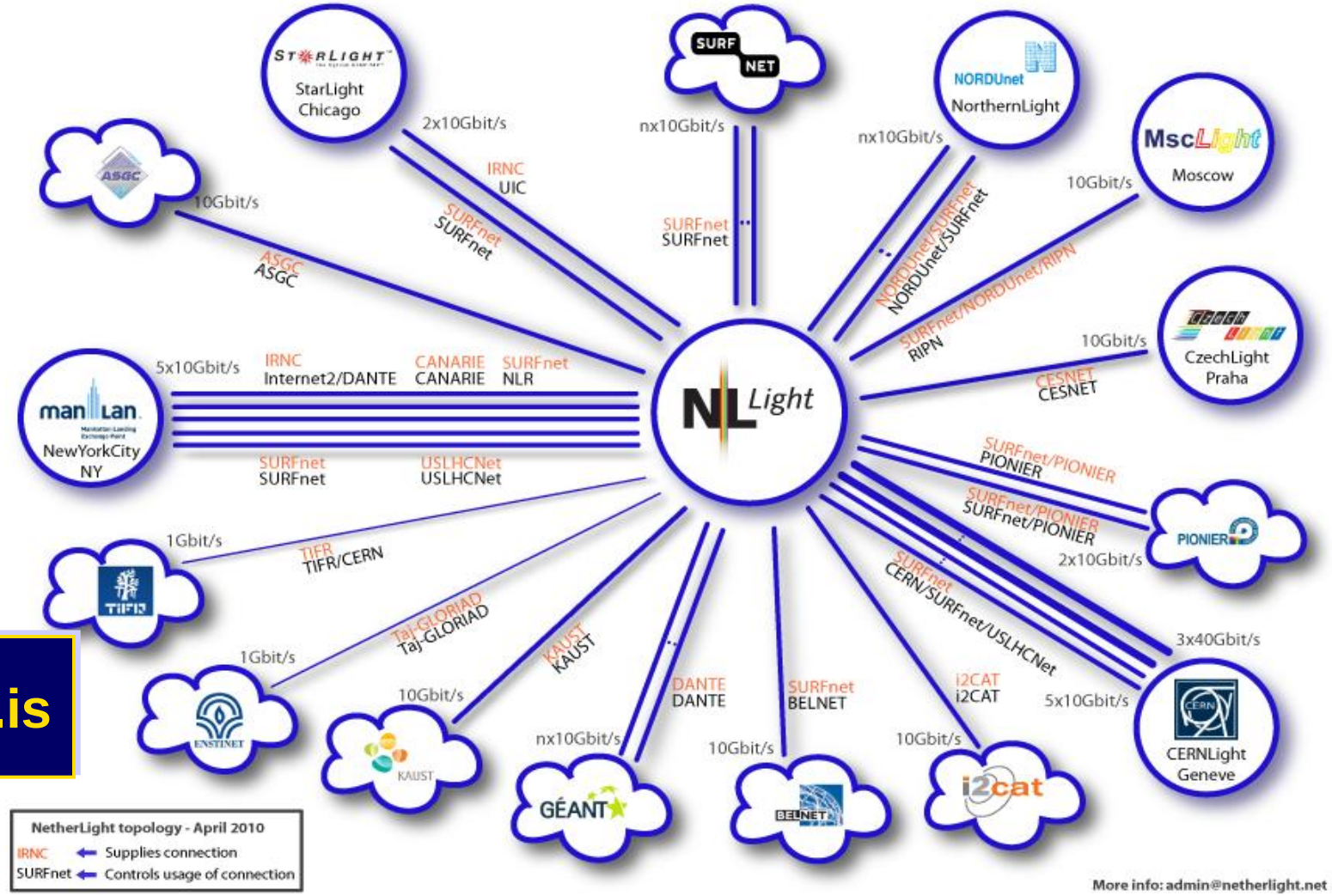
- A design satisfying all the requirements:
  Switched Core with Routed Edge
- **Sites interconnected through Lightpaths**
  - Site-to-site Layer 2 connections, static or dynamic
- **Switching is far more robust and cost-effective for high-capacity interconnects**
- **Routing at the end-sites also often is necessary**



LHCONE: Switched Core of Open Exchange Points

- **Strategically placed core exchange points**
  - E.g. start with 2-3 in Europe, Starlight and ManLAN in NA, Southern Light and AmLight in Latin America, 1-2 in Asia
  - E.g. existing devices at Tier1s, GOLEs, GEANT nodes, …
- **Interconnected through high capacity trunks**
  - 10-40 Gbps today, soon 100Gbps
- **Trunk links can be CBF, multi-domain Layer 1/ Layer 2 links, …**
  - E.g. Layer 1 circuits with virtualized sub-rate channels, sub-dividing 100G links in early stages
- **Resiliency, where needed, provided at Layer 1/ Layer 2**
  - E.g. SONET/SDH Automated Protection Switching, Virtual Concatenation
- **At later stage, automated Lightpath exchanges will enable a flexible "stitching" of dynamic circuits**
  - Demonstration (proof of principle) done at last GLIF meeting & SC10

Open Exchange Points: NetherLight Example
3 x 40G, 30+ 10G Lambdas, Use of Dark Fiber



www.glif.is

Convergence of Many Partners on Common Lightpath Concepts

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

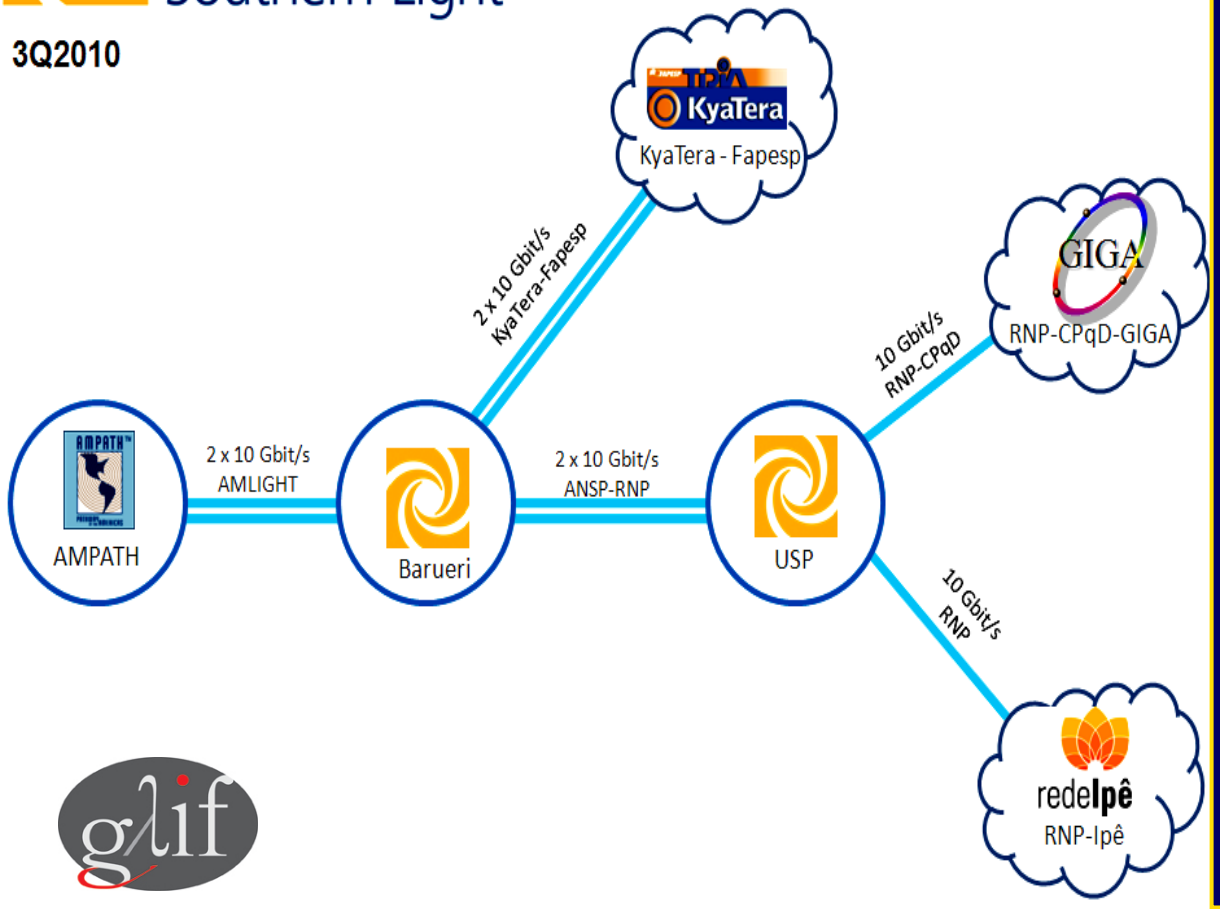
SouthernLight

Latin American Open Exchange Point

M. Stanton, RNP

SOL Southern Light

3Q2010



Additional GLIF resources include the multigigabit core of the Ipê network, to be greatly extended in early 2011, the experimental GIGA network, operated jointly by RNP and CPqD, and the KyaTera network in São Paulo state. Figure shows the current configuration of the SouthernLight GOLE.

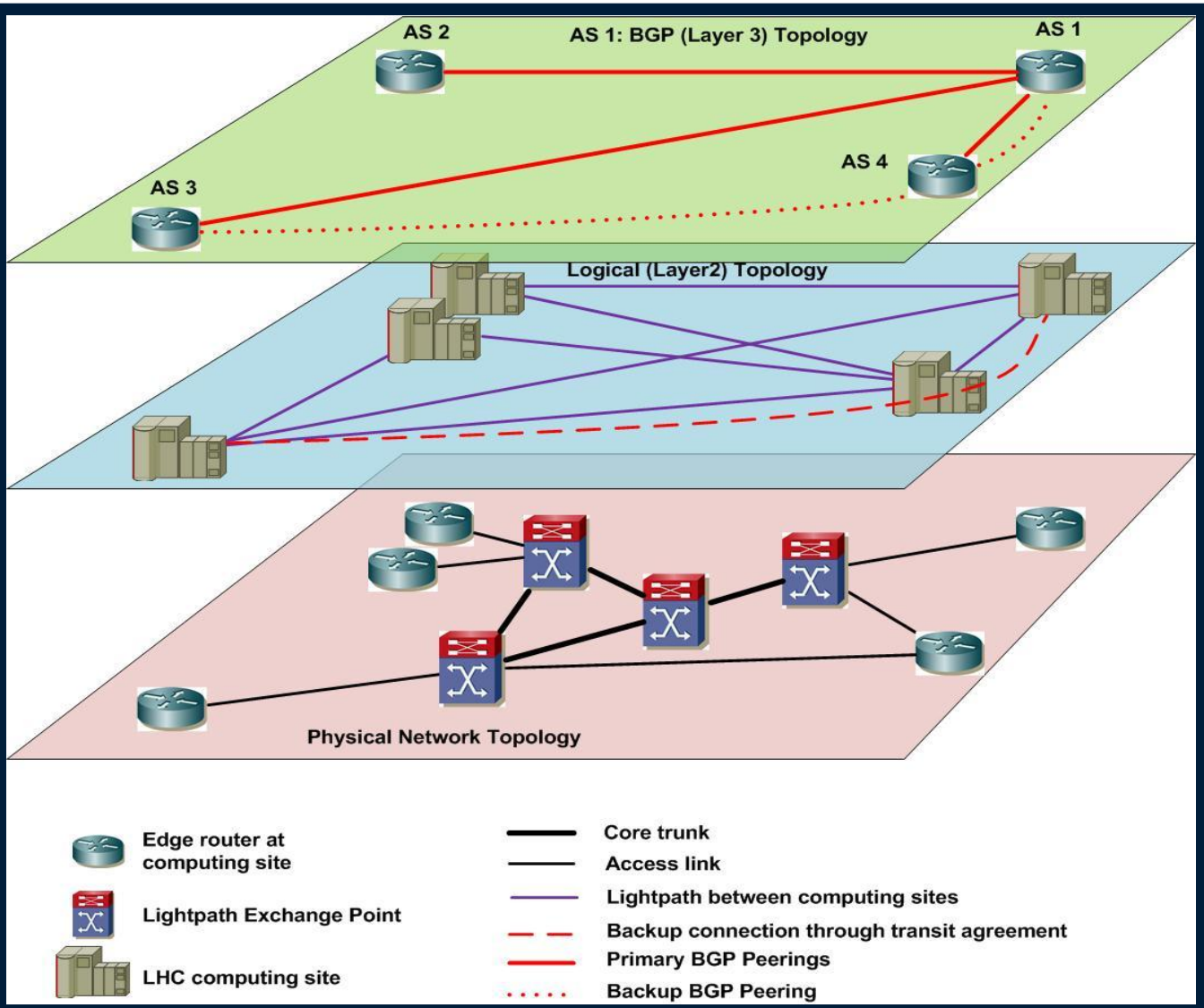
RNP has committed itself to demonstrate an interoperable dynamic circuit service, and it is planned to deploy an experimental service in the next upgrade of the Ipê network in 2011. Such a facility will greatly enhance RNP's capability to manage the widespread use of end to end circuits. RNP was able to carry out experimental studies and is on track to transfer this technology to the future Ipê network.

Southern Light is Recognized as a GOLE by

LHCONE's Routed Edge

- **End sites (might) require Layer 3 connectivity in the LAN**
  - Otherwise a true Layer 2 solution might be adequate
- **Lightpaths terminate on a site's router**
  - Site's border router, or, preferably
  - The Router closest to the storage elements
- **All IP peerings are peer-to-peer, site-to-site**
  - Reduces convergence time, avoids issues with flapping links
- **Each site decides and negotiates with which remote site it desires to peer**
  - (e.g. based on experiment's connectivity design)
- **Router (BGP) advertises *only* the SE subnet(s) through the configured Lightpath**

LHCONE Layer1 through Layer 3



How do End-Sites Connect?
A Simple Example

- **The Tier1 at UNAM needs 1 Gbps (soon a few Gbps) connectivity (each) to 2 sites in Europe, 2 in US and the KISTI Tier1 in Korea**
- **5 x 1G intercontinental circuits are cost-prohibitive**
- **The Tier2 could however afford the needed circuit to the next GOLE (e.g. Starlight via San Antonio + NLR or Southern Light + AmLight**
  - ➡ Through CUDI, AmLight, NLR and Internet2
- **The GOLE connects to other GOLEs via trunks (StarLight, NetherLight, KRLight, etc.) (trunks) and then onward to the sites**
- **Static bandwidth allocation (first stage)**
  - ➡ The end-site has a 1Gbps link, with 5 VLANS, each one terminating at one of the desired remote sites
  - ➡ Bandwidth is allocated by the exchange points to fit the needs
- **Dynamic allocation (later this year); with BW guarantees**
  - ➡ The end-site has a 1Gbps link, with configurable remote end-points and bandwidth allocation

LHCONE Summary

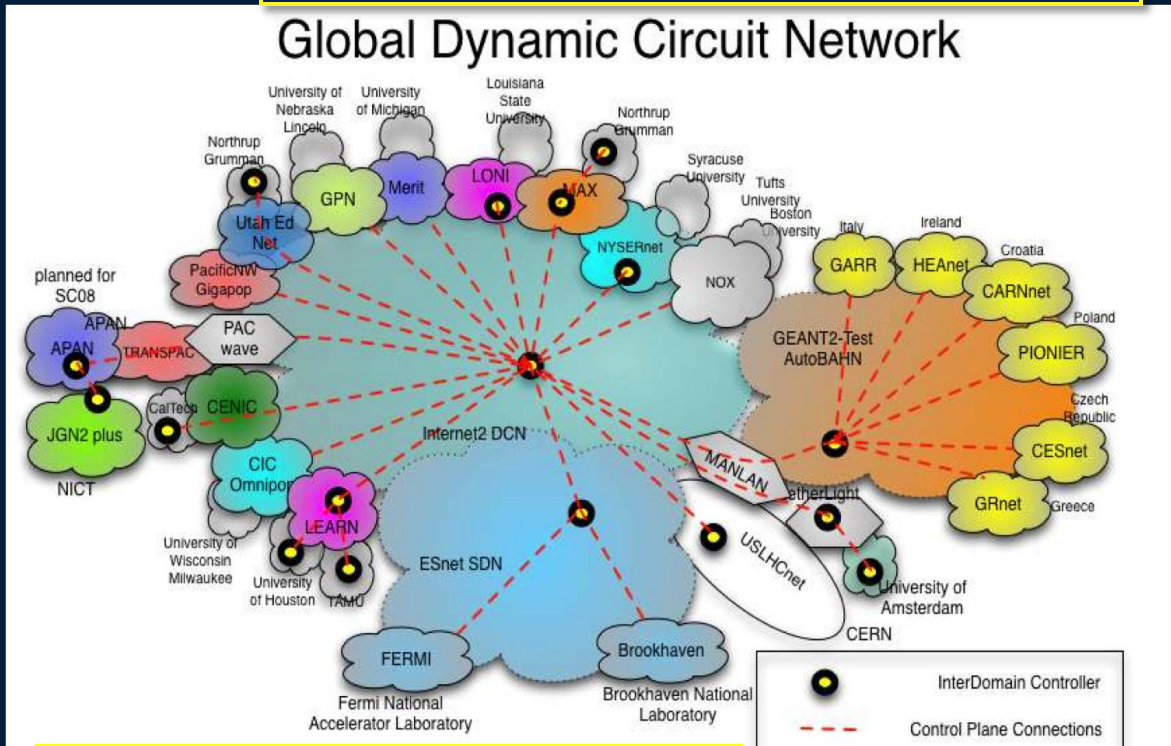
- **LHCONE:** A robust, scalable & comparatively low-cost solution based on a switched core with routed edge architecture
- Core consists of sufficient number of strategically placed Open Exchange Points interconnected by properly sized trunk circuits
  - Scaling rapidly with time as in the requirements document
- IP routing is implemented at the end-sites
- Initial deployment: predominantly static Lightpaths, later predominantly using dynamic circuits + resource allocation
- A federated governance model has to be used due to the global geographical extent and diversity of funding sources
  - Organic growth; Key Role of NRENs (most notably CUDI)

Dynamic Circuit Networks

- Separate high impact data flows from general network traffic
- Provide Quality of Service Guarantees
  - Bandwidth, availability; latency, jitter
- Network resource reservation in advance, or "on demand"
- Manage, schedule resources
- Hybrid: support various network technologies
  - Optical (λ-switched)
  - Packet switched
  - Routed (IP/MPLS, GMPLS)
- DYNES: A recent example relevant to the UNAM Tier1 Plan

Pioneering implementations in production:
ESnet OSCARS; Internet2 ION, SURFnet DRAC

Together with other developments and prototypes (AutoBAHN))

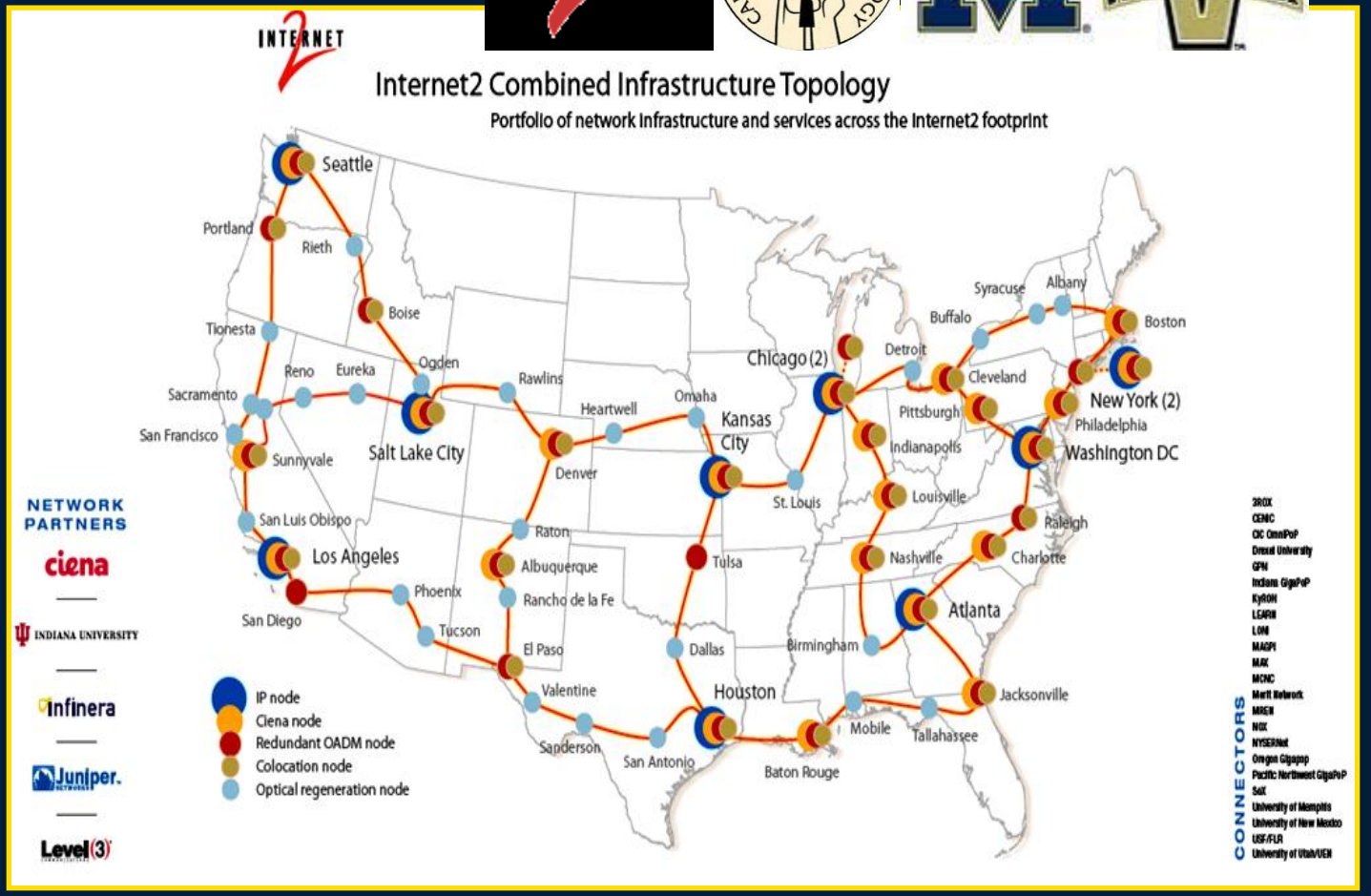


Int'l Demonstration Setup

DYNES: http://www.internet2.edu/dynes
Dynamic Network System Project

- **Internet2, Caltech, Vanderbilt, Univ. of Michigan**

- **PI:** Eric Boyd (Internet2)
- **Co-PIs:**
  - Harvey Newman (Caltech)
  - Paul Sheldon (Vanderbilt)
  - Shawn McKee (Michigan)

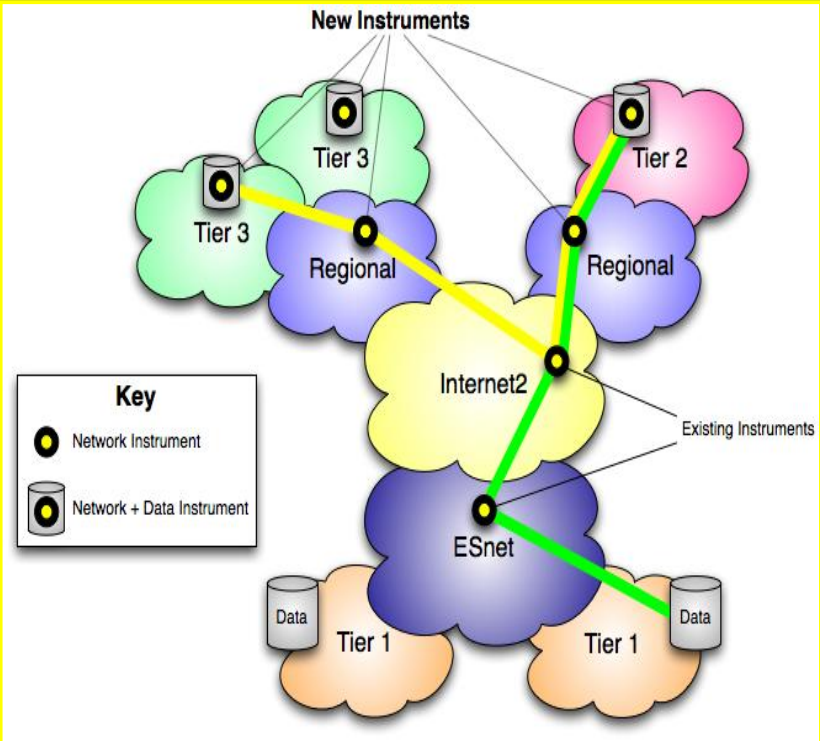


DYNES Overview

- **What is DYNES?**
  - A U.S-wide dynamic network "cyber-instrument" spanning ~40 US universities and ~14 Internet2 connectors
  - Extends Internet2's dynamic network service "ION" into U.S. regional networks and campuses; Aims to support LHC traffic (also internationally)
  - Based on the implementation of the Inter-Domain Circuit protocol developed by ESnet and Internet2; Cooperative development also with GEANT, GLIF
- **Who is it?**
  - The project team: Internet2, Caltech, Univ. of Michigan, Vanderbilt
  - The LHC experiments, astrophysics community, WLCG, OSG, other VOs
  - The community of US regional networks and campuses
  - International Partners
- **What are the goals?**
  - Support large, long-distance scientific data flows in the LHC, other programs (e.g. LIGO, Virtual Observatory), & the broader scientific community
  - ➡ ***Build a distributed virtual instrument at sites of interest to the LHC, but available to R&E community generally***

DYNES System Description

- **AIM: extend hybrid & dynamic capabilities to campus & regional networks.**
  - A DYNES instrument must provide two basic capabilities at the Tier 2S, Tier3s and regional networks:
  1. Network resource allocation such as bandwidth to ensure transfer performance
  2. Monitoring of the network and data transfer performance
- **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 Internet2 and ESnet, 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

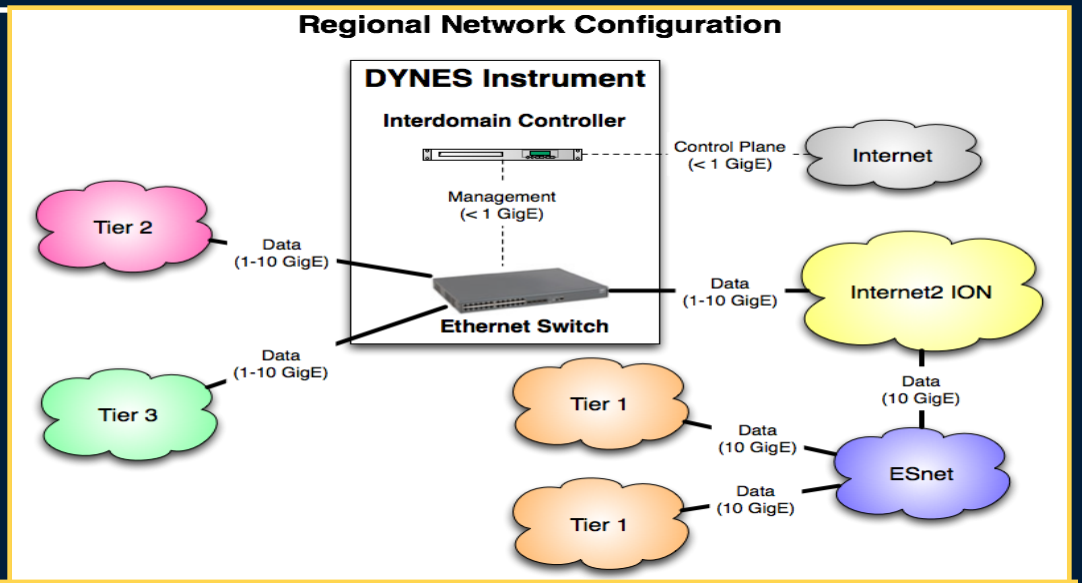


Two typical transfers that DYNES supports: one Tier2 - Tier3 and another Tier1-Tier2.

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

DYNES: Regional Network - Instrument Design

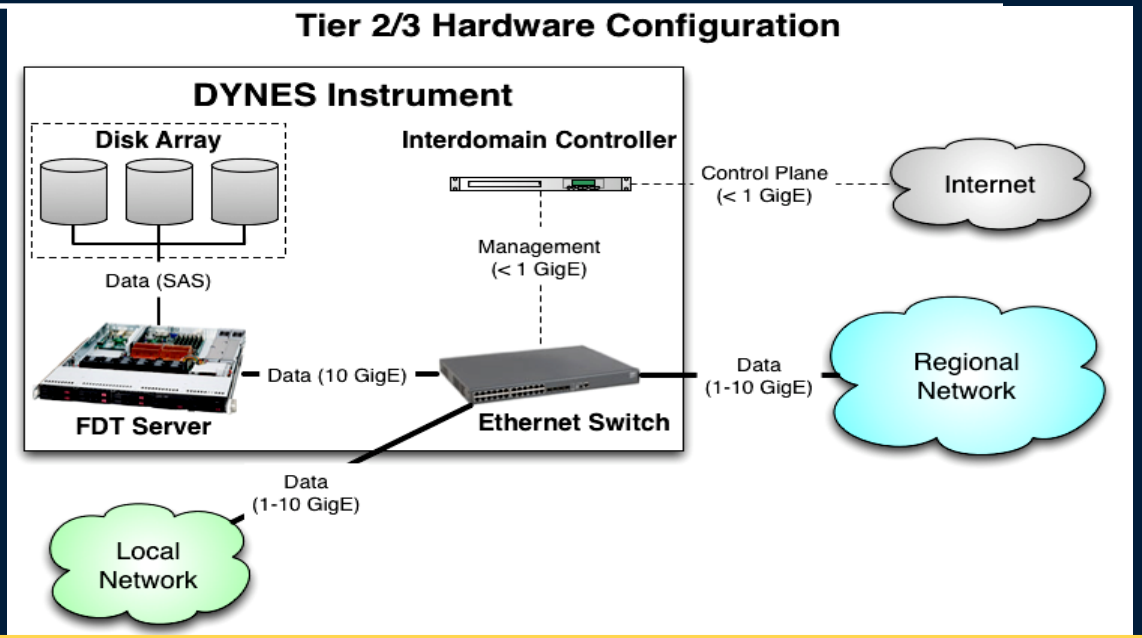
- **Regional networks require**
  1. An Ethernet switch
  2. An Inter-domain Controller (IDC)
- **The configuration of the IDC consists of OSCARS, DRAGON, and perfSONAR. This allows the regional network to provision resources on-demand through interaction with the other instruments**
- **A regional network does not require a disk array or FDT server because they are providing transport for the Tier 2 and Tier 3 data transfers, not initiating them.**



At the network level, each regional connects the incoming campus connection to the Ethernet switch provided. Optionally, if a regional network already has a qualified switch compatible with the dynamic software that they prefer, they may use that instead, or in addition to the provided equipment. The Ethernet switch provides a VLAN dynamically allocated by OSCARS & DRAGON. The VLAN has quality of service (QoS) parameters set to guarantee the bandwidth requirements of the connection as defined in the VLAN. These parameters are determined by the original circuit request from the researcher / application. Through this VLAN, the regional network provides transit between the campus IDCs connected in the same region or to the global IDC infrastructure.

DYNES: Tier2 and Tier3 Instrument Design

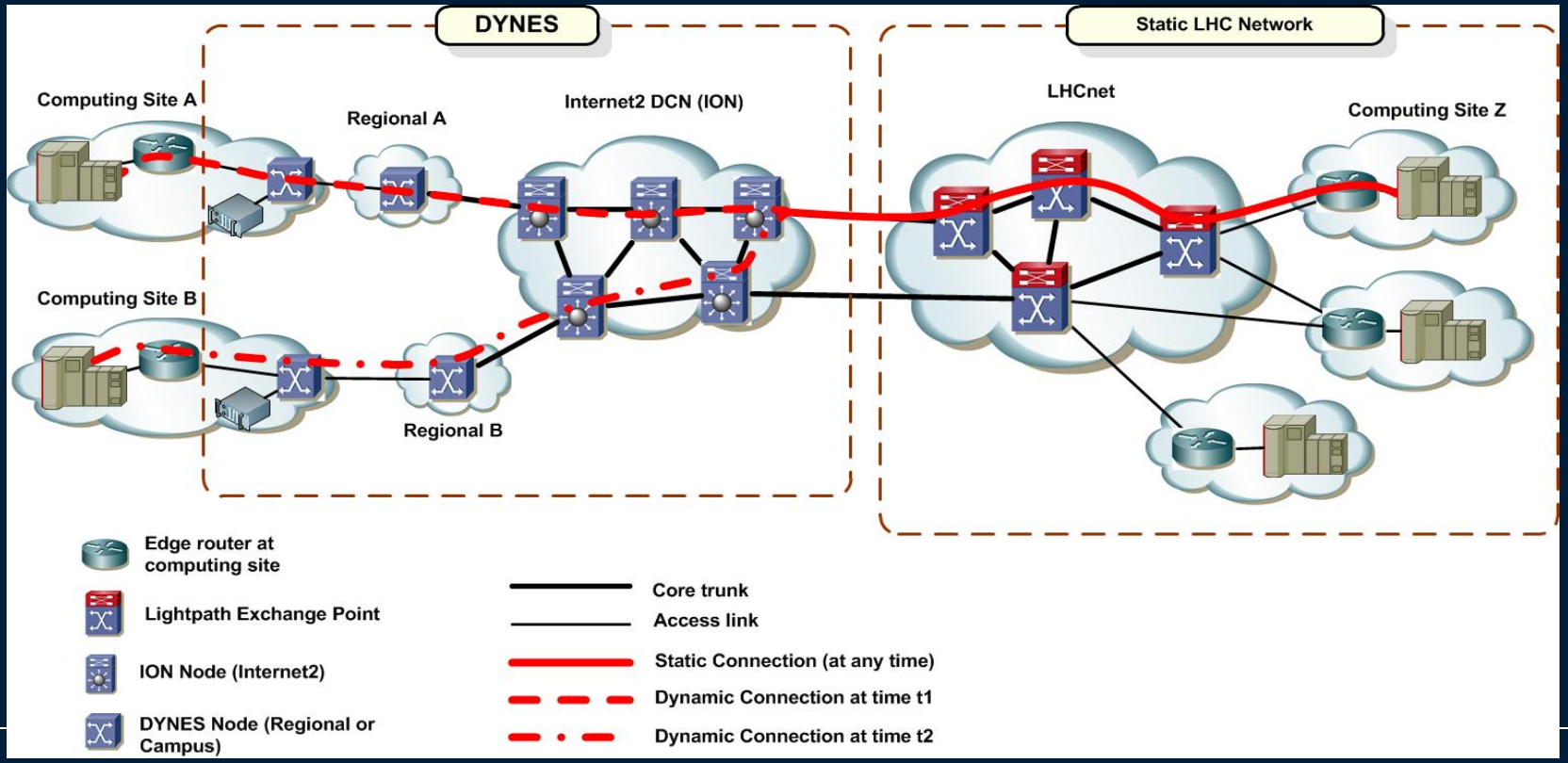
- **Each DYNES (sub-)instrument at a Tier2 or Tier3 site consists of the following hardware, combining low cost & high performance:**
  1. An Inter-domain Controller (IDC)
  2. An Ethernet switch
  3. A Fast Data Transfer (FDT) server. Sites with 10GE throughput capability will have a dual-port Myricom 10GE network interface in the server.
  4. An optional attached disk array with a Serial Attached SCSI (SAS) controller capable of several hundred MBytes/sec to local storage.



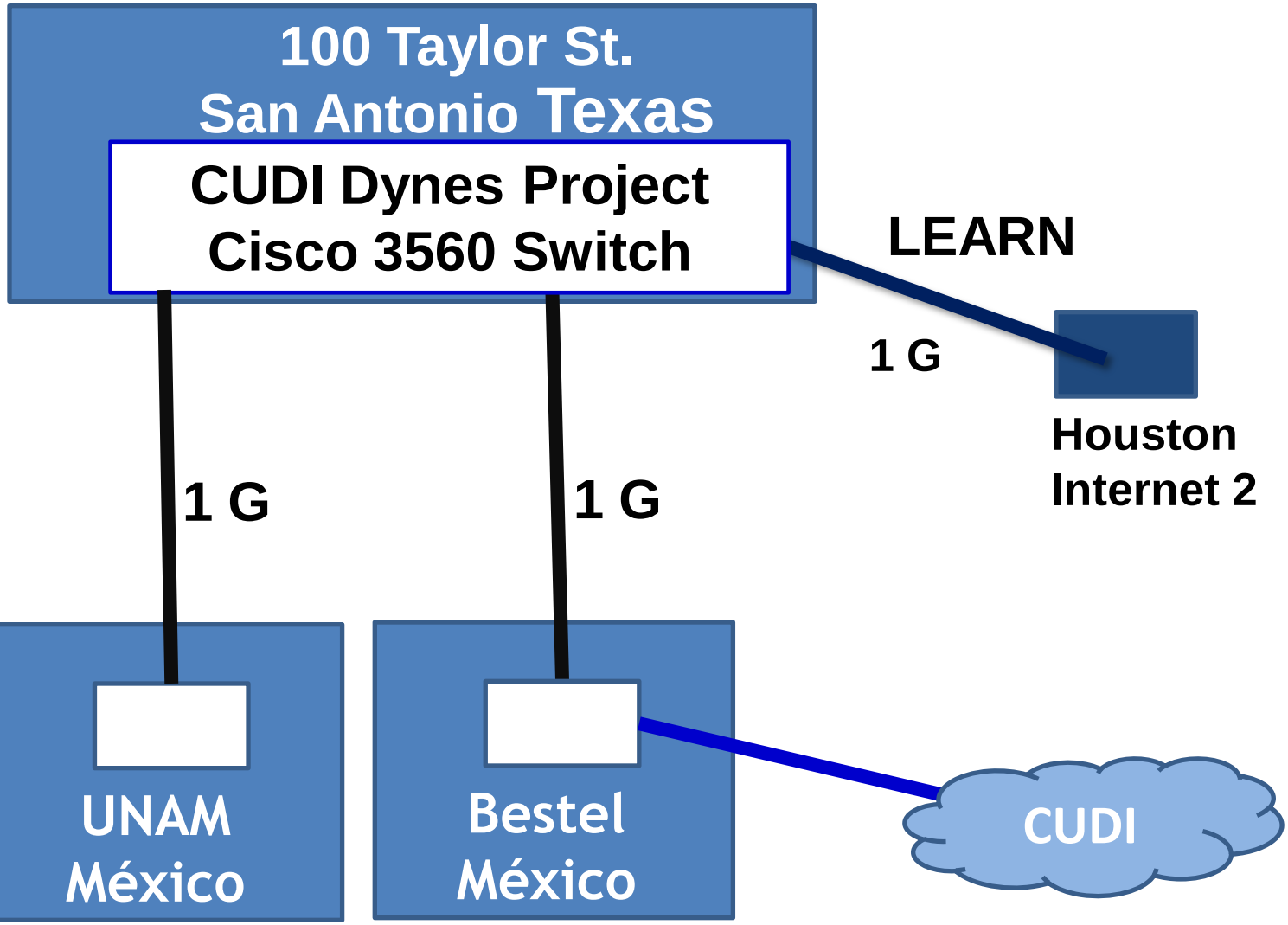
Fast Data Transfer (FDT) server connects to the disk array via the SAS controller and runs FDT software developed by Caltech, an asynchronous multithreaded system that automatically adjusts I/O and network buffers to achieve maximum network utilization. The disk array stores datasets to be transferred among the sites in some cases. 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.

How Can DYNES be Leveraged in LHCONE ?

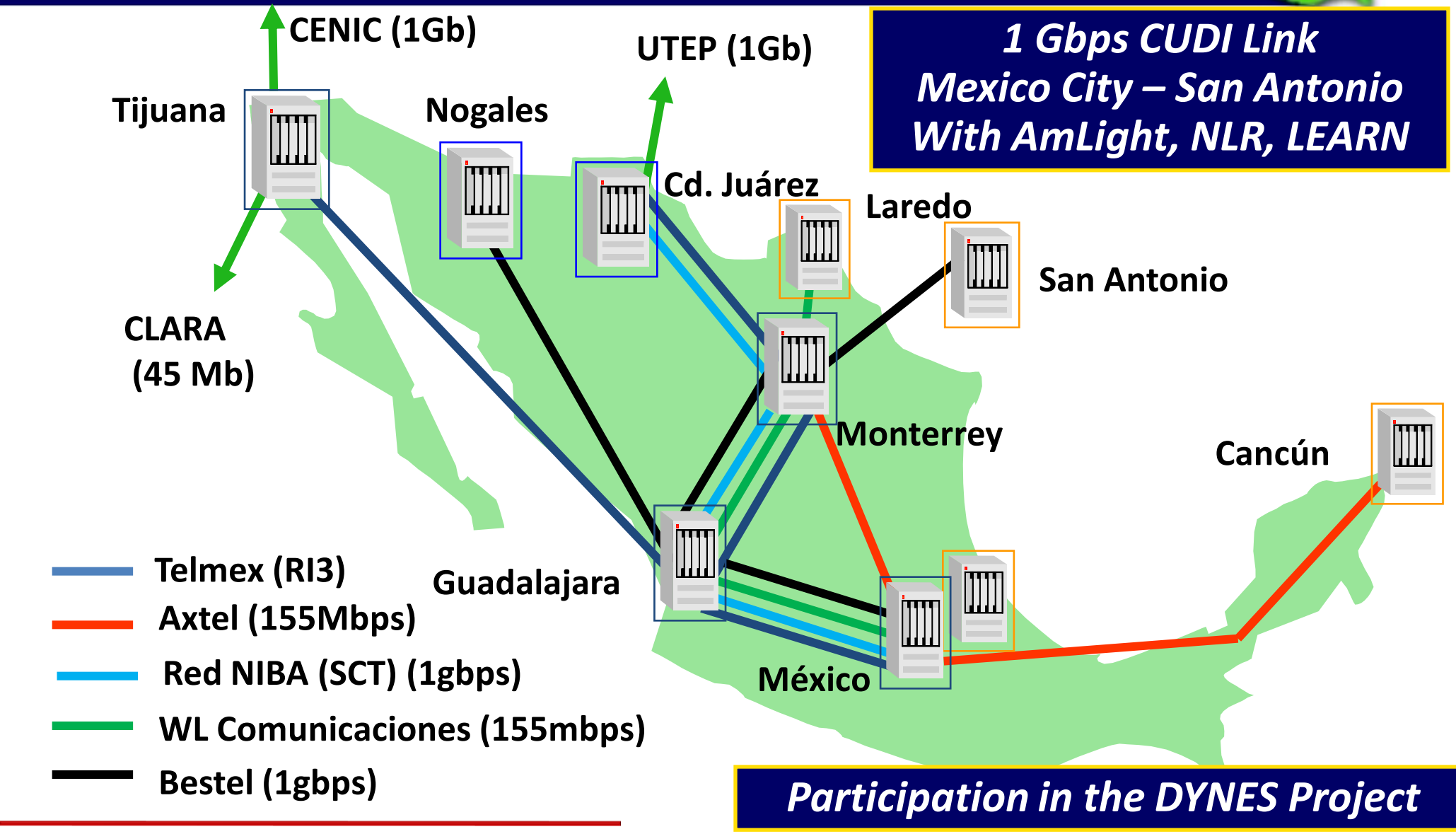
- The Internet2 ION service currently has end-points at two GOLEs in the US: MANLAN & StarLight; + I2 "Distributed OEP" Proposed
- A static Lightpath from any end-site to one of these GOLE sites can be extended through ION to any of the DYNES sites (LHC Tier2 or Tier3)



The UNAM 1 Gbps Link into San Antonio can also be connected to the CUDI switch



Backbone of the CUDI Network

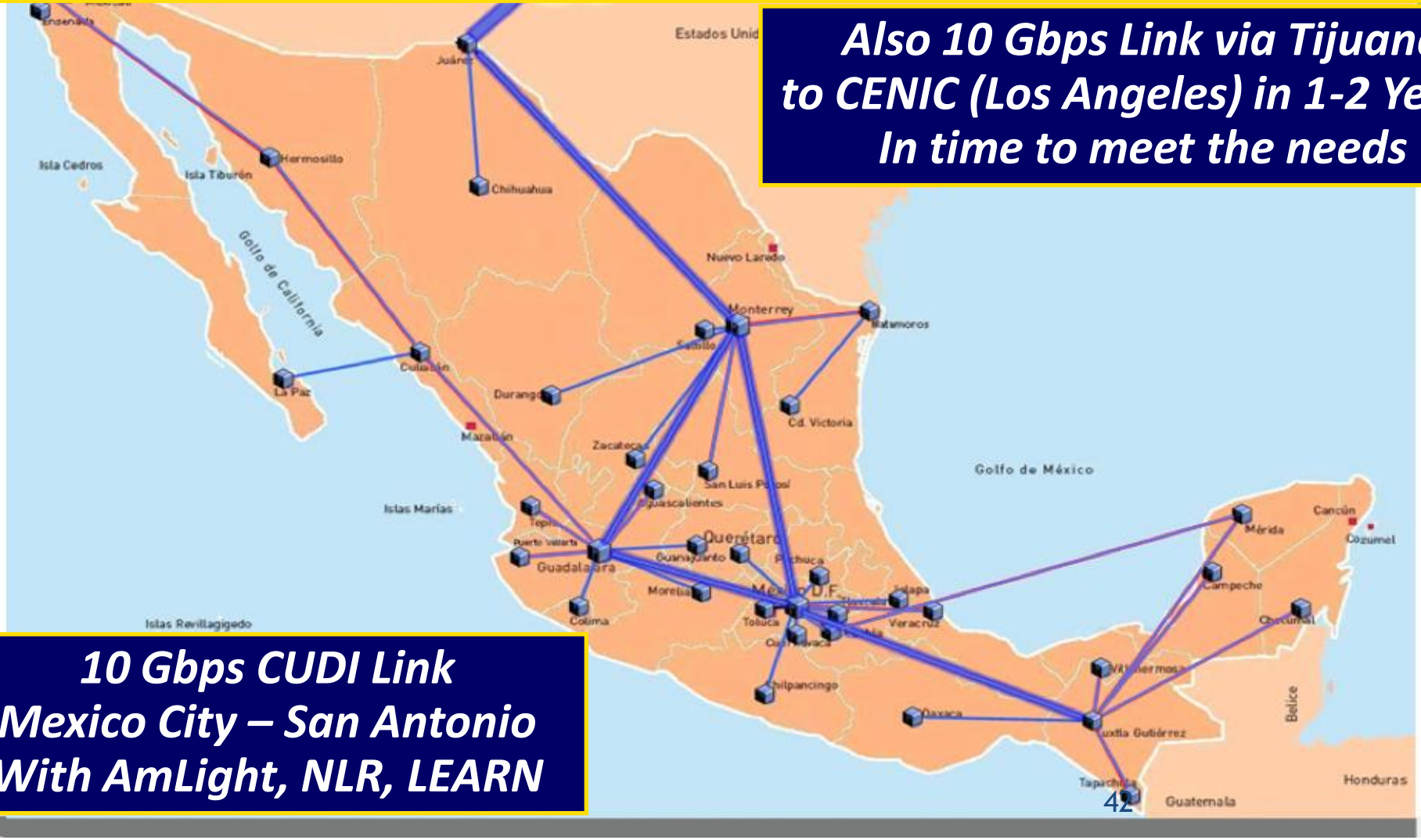


1 Gbps CUDI Link
Mexico City – San Antonio
With AmLight, NLR, LEARN

Participation in the DYNES Project

Backbone of the CUDI Network

Expansion of the Backbone to 10 Gbps (SCT, CUDI, Conacyt in June)



Also 10 Gbps Link via Tijuana to CENIC (Los Angeles) in 1-2 Years; In time to meet the needs

10 Gbps CUDI Link Mexico City – San Antonio With AmLight, NLR, LEARN

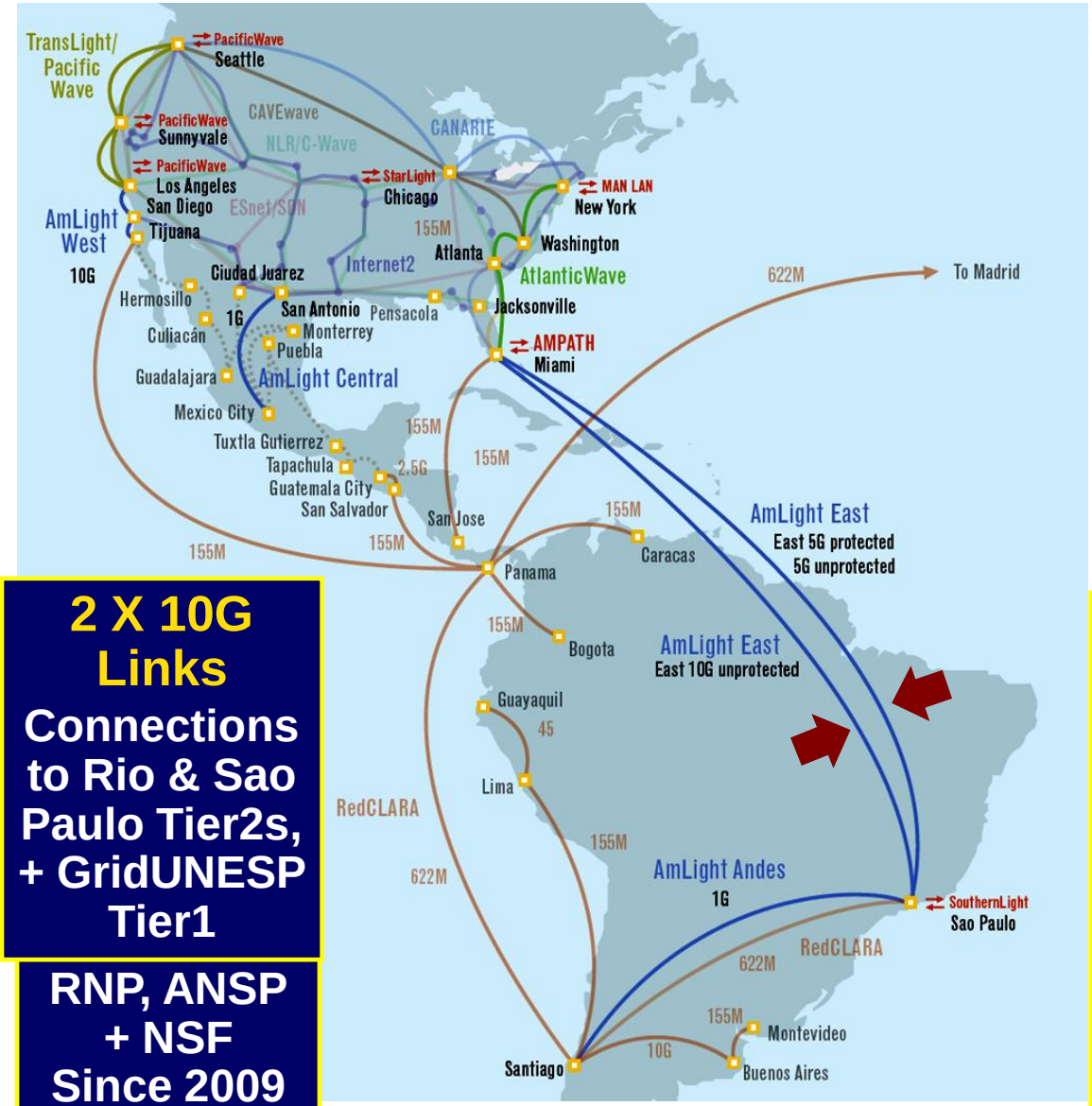
Americas Lightpaths (AMLIGHT)

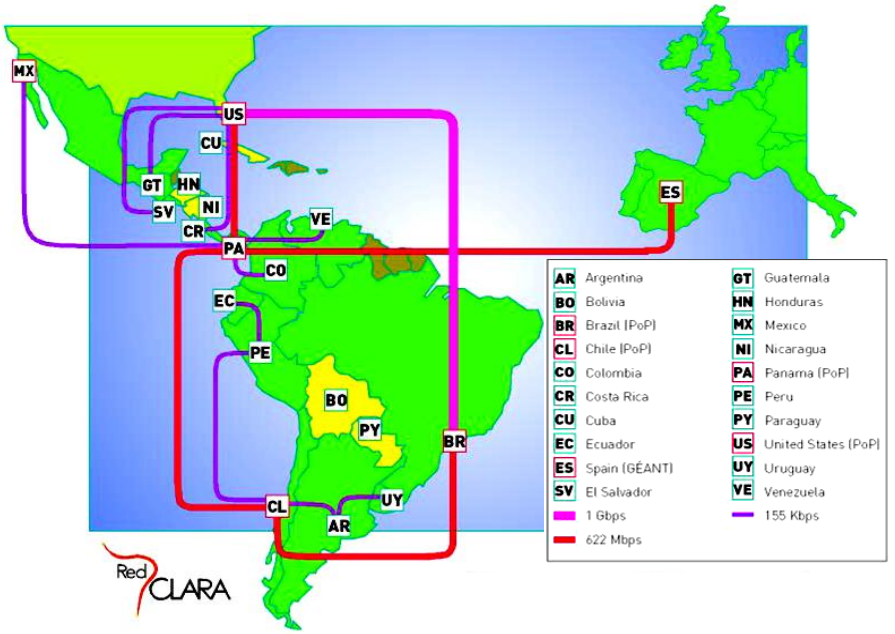
- AmLight is a project with support from U.S. Nat'l Science Foundation & its collaborators
- AmLight aims to enhance science research and education in the Americas by
  - Providing operation of production infrastructure
  - Engaging U.S. and Latin American science and engineering research and education communities
  - *Creating an open instrument for collaboration*

USA-Mexico Links Supported by AMLIGHT

- Connects USA and Mexico Research & Education (R&E) Communities via California and Texas
- Project collaborators are FIU, CUDI, CENIC and LEARN
- One 1 Gb wave between Mexico City & San Antonio
  - Connections to Internet2 and NLR operated by LEARN and CUDI
- Includes two 1 Gb waves between Tijuana & Los Angeles, *increasing to 10 Gb in 2011*
  - Connections to PacificWave and international networks operated by CENIC and CUDI

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





2 X 10G Links

Connections to Rio & Sao Paulo Tier2s, + GridUNESP Tier1

RNP, ANSP + NSF Since 2009

- **RedCLARA (EU-Funded)**
  155/622M Connections Among 18 Latin Am. NRENs; 622 M to GEANT
- **EEC 2 / 3 of cost; 2nd round funding 18 M€ (2009-2012)**
- **Using these resources to acquire fiber assets to connect to most countries**
- **10G Link Santiago ↔ Buenos Aires for Auger**

GLIF 2010 Map DRAFT: Brazil

RNP-Ipe

RNP Giga

Kyatera

(Sao Paulo)

CLARA/RNP

Innova Red

(br, ar, cl)

REUNA-ESO

AmLight East

AmLight Andes

➡ **Cross Border Dark Fiber Initiatives** underway with Argentina, Chile, Uruguay, Paraguay

Summary and Conclusions: Networks in the LHC Era and the UNAM Tier1

- The capacity and capability of HEP's networks continues to advance; we will soon be taking the next step to 40G/100G on major routes
- The experiments are building a new round of Computing Models, with greater reliance on networks
  - More intensive use of Tier2s & Tier3s; more complex flows
  - More agile, and more effective for discoveries
- The LHCOPN team has designed and is developing a new architecture based on a global set of Open Exchange Points to meet the needs: *LHCONE*
  - Experiments and network providers need to work together now, to complete the Phase 1 plan and begin operations in 2011
- Working with CUDI, AmLight, RNP, NLR, Internet2, US LHCNet and other partners the UNAM Tier1 will poised to secure the necessary network resources, and participate in these important inter-regional developments
- We must continue to work on the Digital Divide in Latin America
  - Starting with all the universities and projects in Mexico, with CUDI

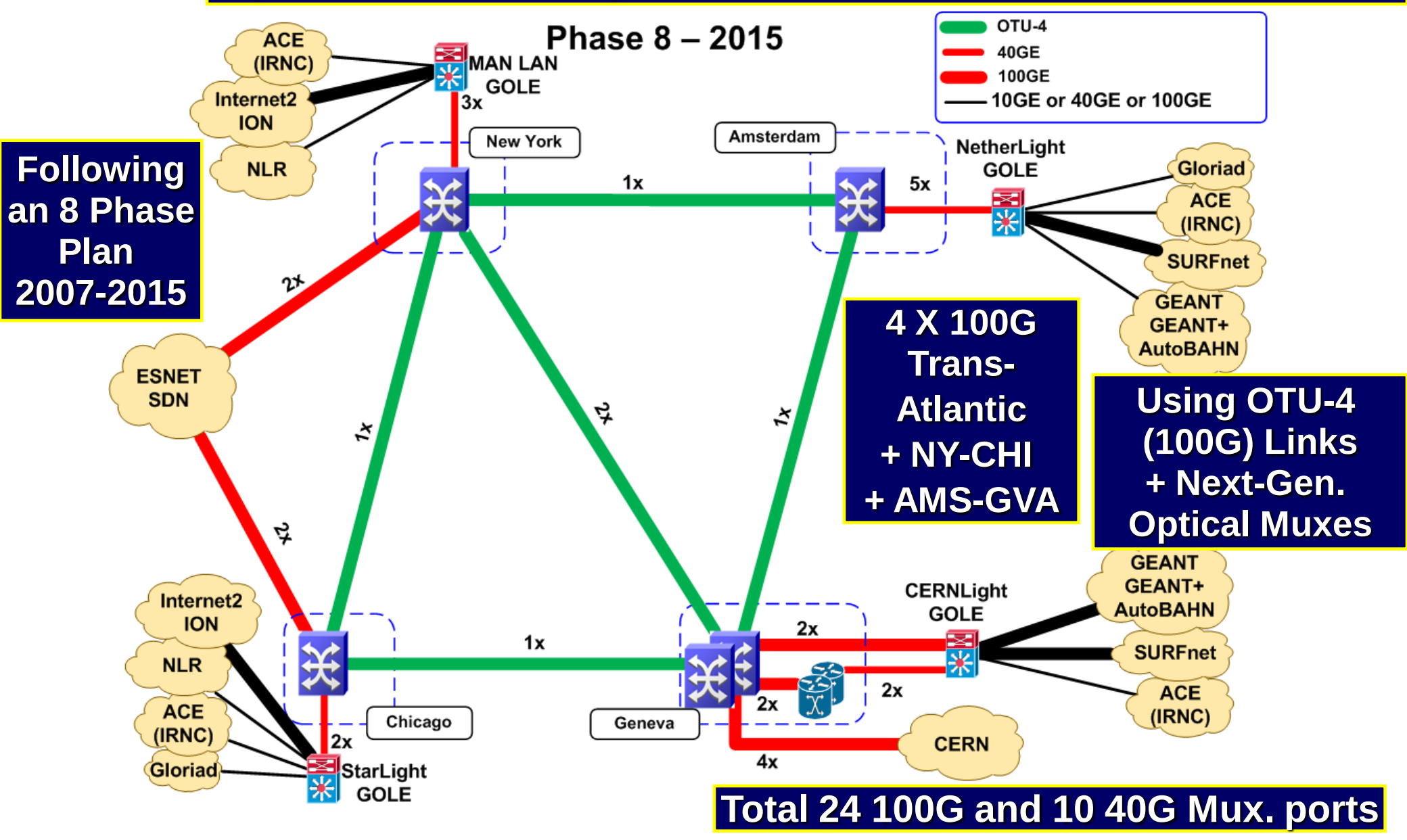
THANK YOU!

newman@hep.caltech.edu

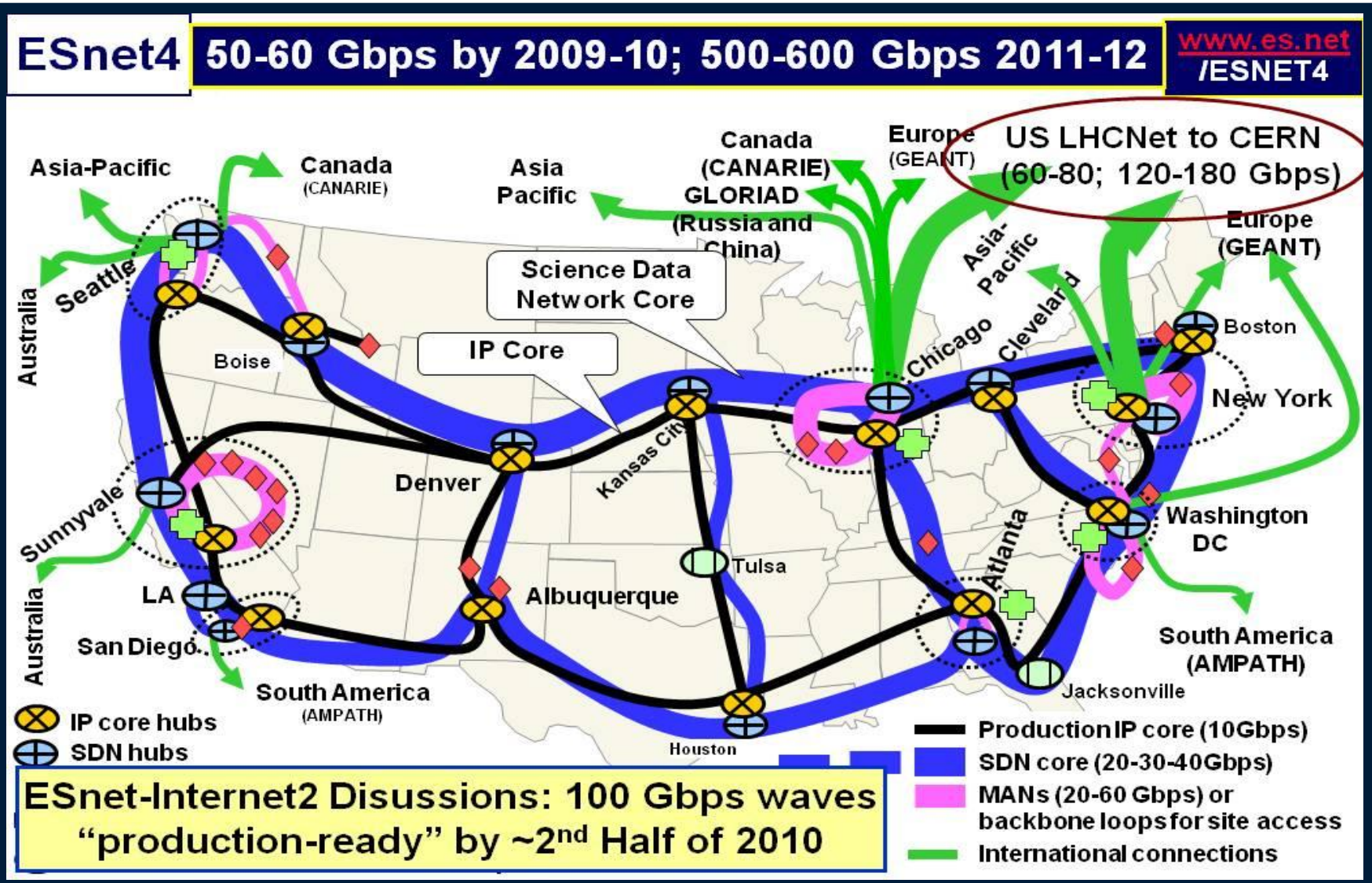
EXTRA SLIDES

TOWARDS THE NEXT GENERATION 40G & 100G NETWORK TECHNOLOGIES

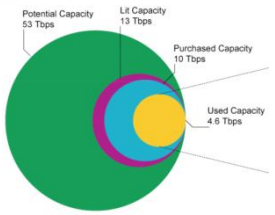
Implementation: USLHCNet Scenario Phase 8 (2015 or 2014 ?): Transition to Full Use of 100G



ESnet Future N X 100G Network



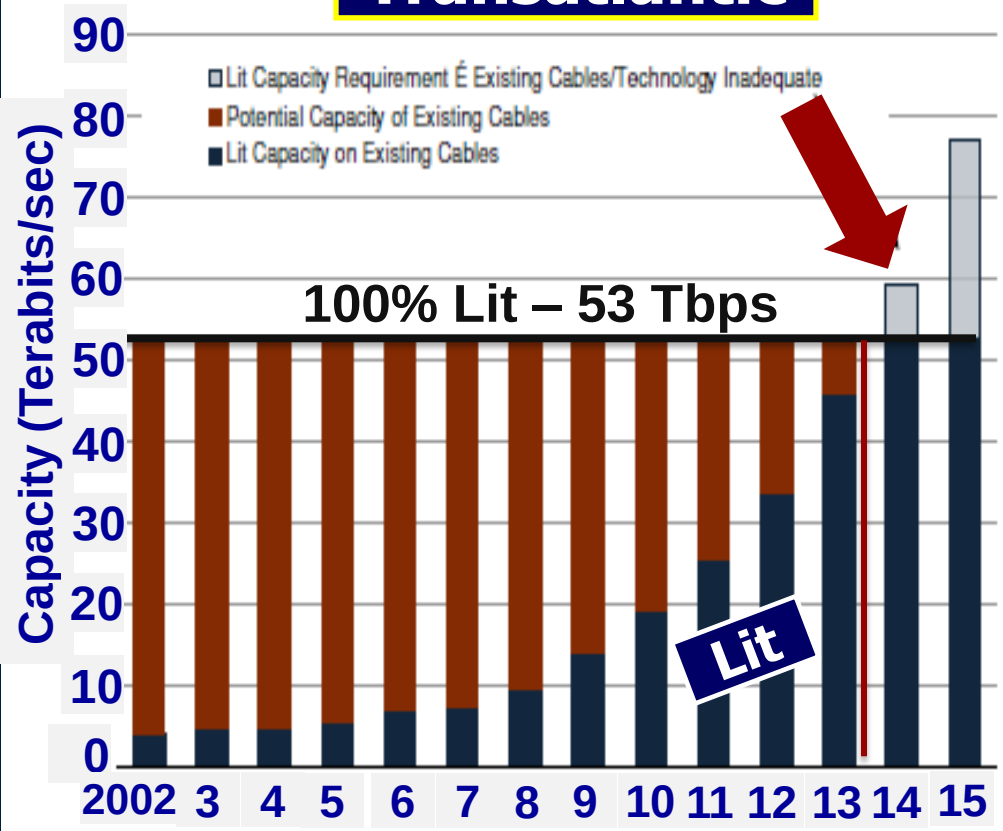
Internet Traffic Growth 65% Per Year; 100+% in Developing Regions



When will We Run Out of Capacity?

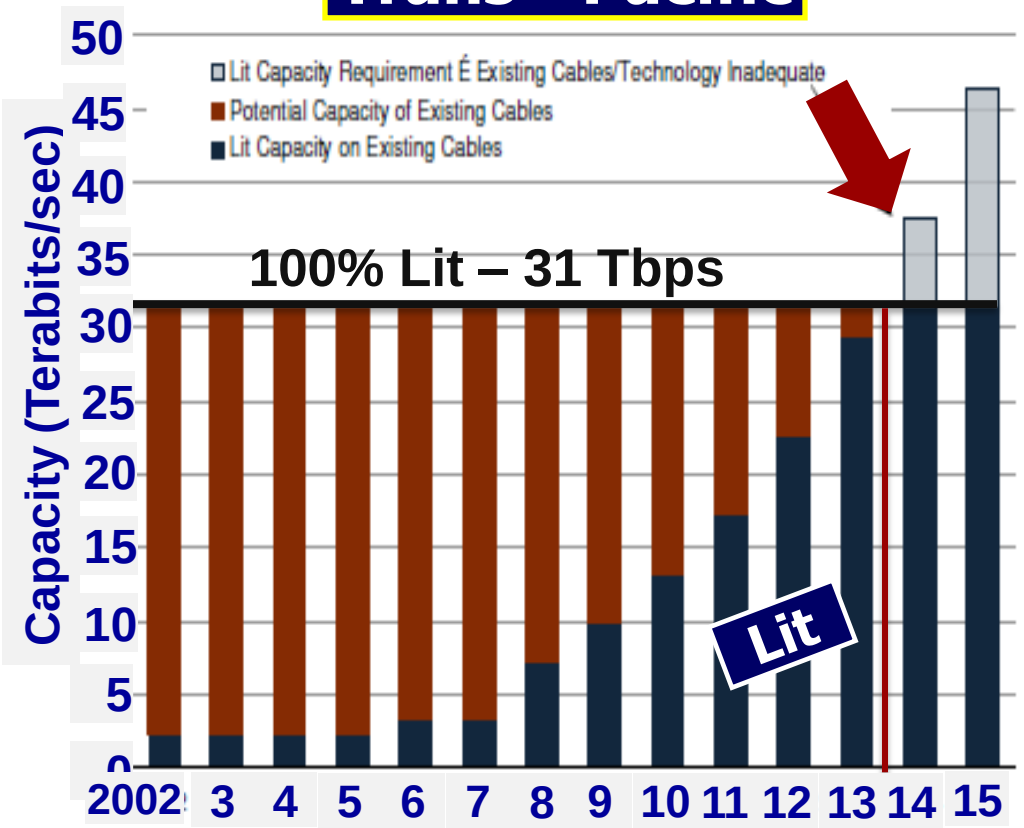
Trans-Atlantic

Transatlantic



Trans-Pacific

Trans - Pacific

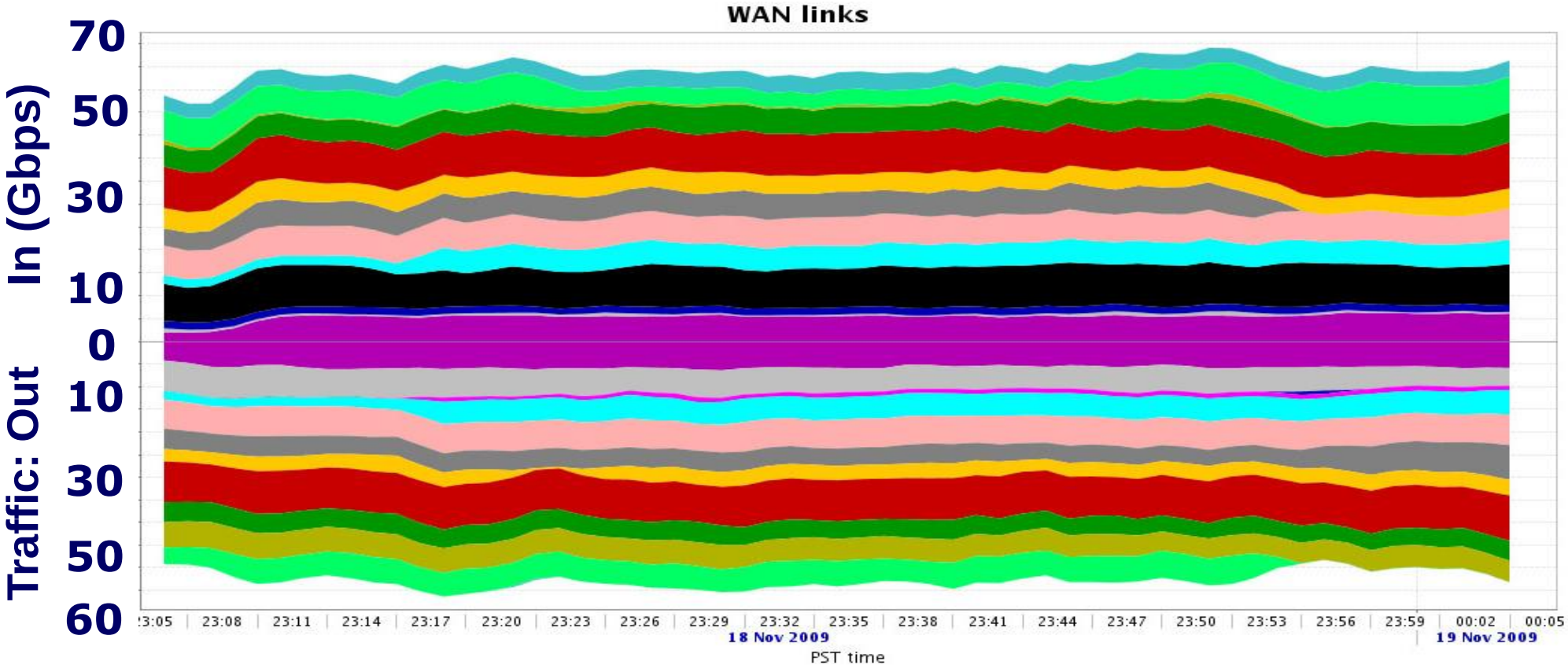


Transatlantic and Trans-Pacific BW Will Become Very Scarce in 2013 ➡ Drive Transition to 40G or 100G Waves by ~2012. Examples: Pacific, SE-ME-WE4

SC09

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

~616 CPU Cores and 38 10GE NICs in 1 Rack of Servers
53 10GE Switch Ports; ~100 TB Disk



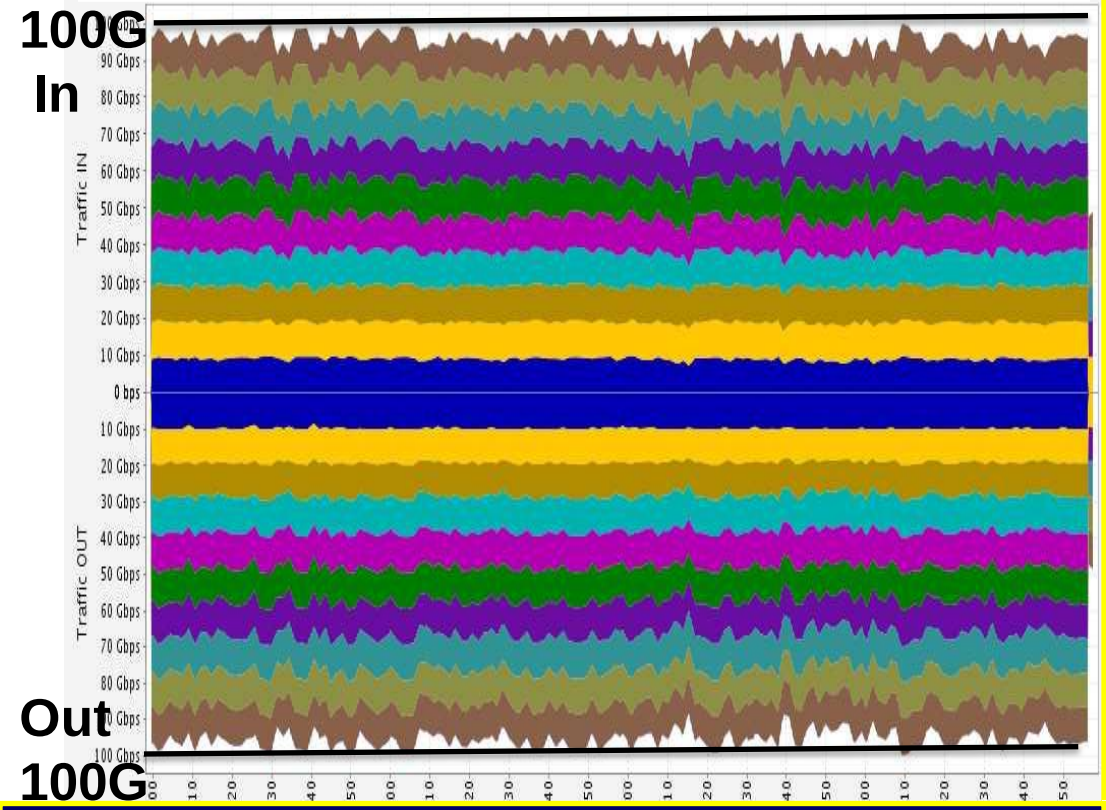
Max. 119 Gbps; 110 Gbps Sustained; 65 Gbps Outbound

Using FDT and FDT/Hadoop Storage/Storage; Now FDT/PhEDEx

SC08

Caltech and CIENA: 191 Gbps Avg., 199.90 Gbps Max on An OTU4 (Standard 100G) Wave at SC2008: 80km

1.02 Petabytes Overnight



10 X 10G Waves at the Caltech HEP Booth

Used Fully, in Both Directions with Caltech's *FDT:*
TCP-Based Java
Open Source Application

Previewing the USLHCNet Transition to 4 X 100G by ~2015

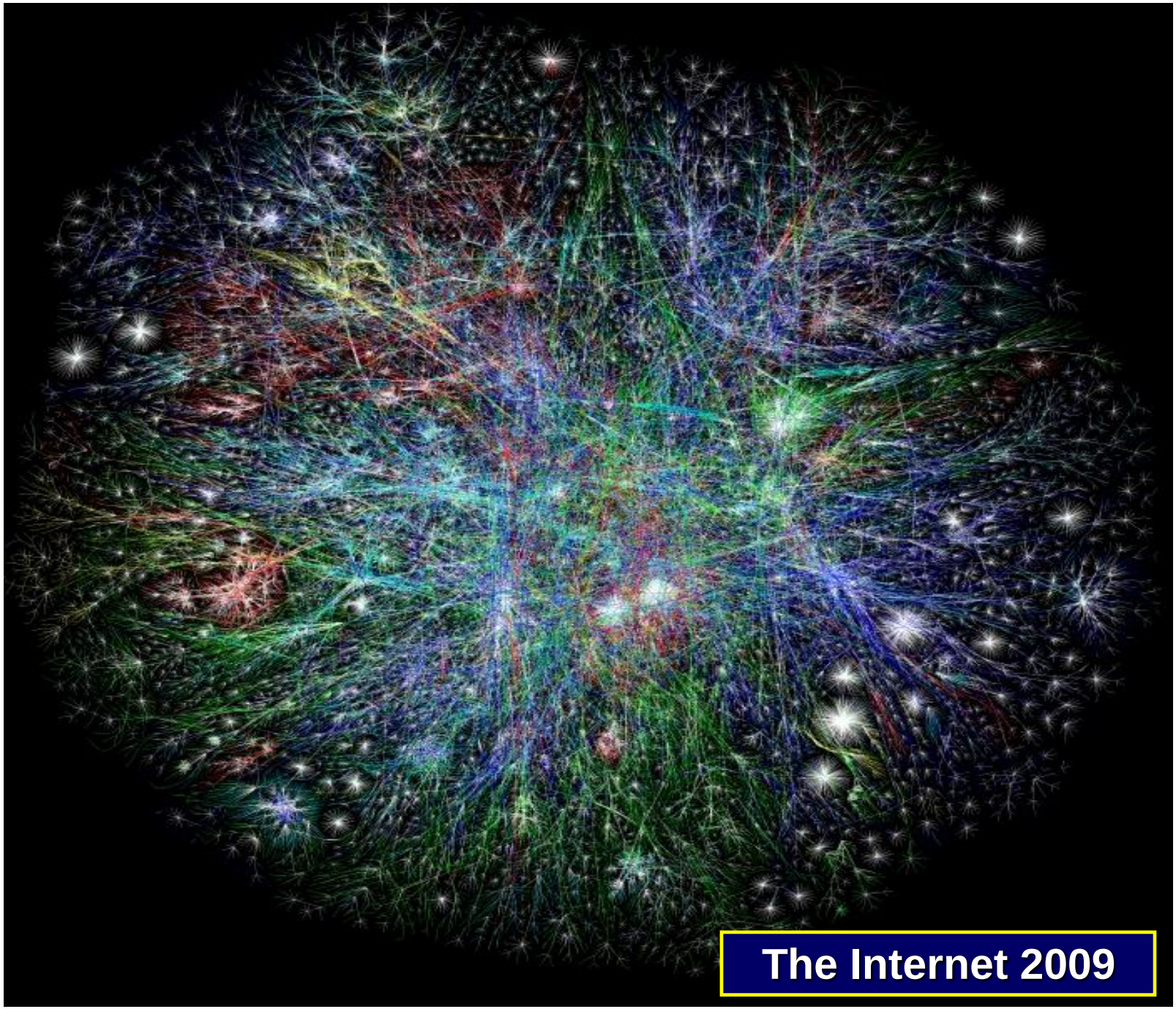
***GLIF:* 40 GE Xfer + Streaming from SSDs Amsterdam-CERN 10/13/10**

Parallel Sessions:
FDT/Hadoop & PhEdeX

EXTRA SLIDES

THE DIGITAL DIVIDE CONTINUES

ICFA
SCIC



SCIC Main Conclusions for 2010

- *It is more urgent than ever, as we enter the LHC era in earnest that we act to Close the Digital Divide*
  - ➔ *To make physicists from all world regions full partners in the upcoming scientific discoveries*
- *We are learning to help do this effectively, in some cases in partnership with many agencies and HEP groups:*
  - ➔ *Brazil (RNP), Mexico (CUDI)*
  - ➔ *AmLight (FIU)*
  - ➔ *"Taj" Extension of GLORIAD to Middle East and India*
- *But we are indeed beginning to leave other countries and regions behind, for example: the Rest of Latin America; Most of the Middle East, South Asia; Africa*
- *A great deal of work remains: Support for the IEPM Monitoring Effort at SLAC is vital for this work*

ICFA Report 2010 - Main Trends Accelerate:
Dark Fiber Nets, Dynamic Circuits, 40-100G
http://cern.ch/icfa-scic

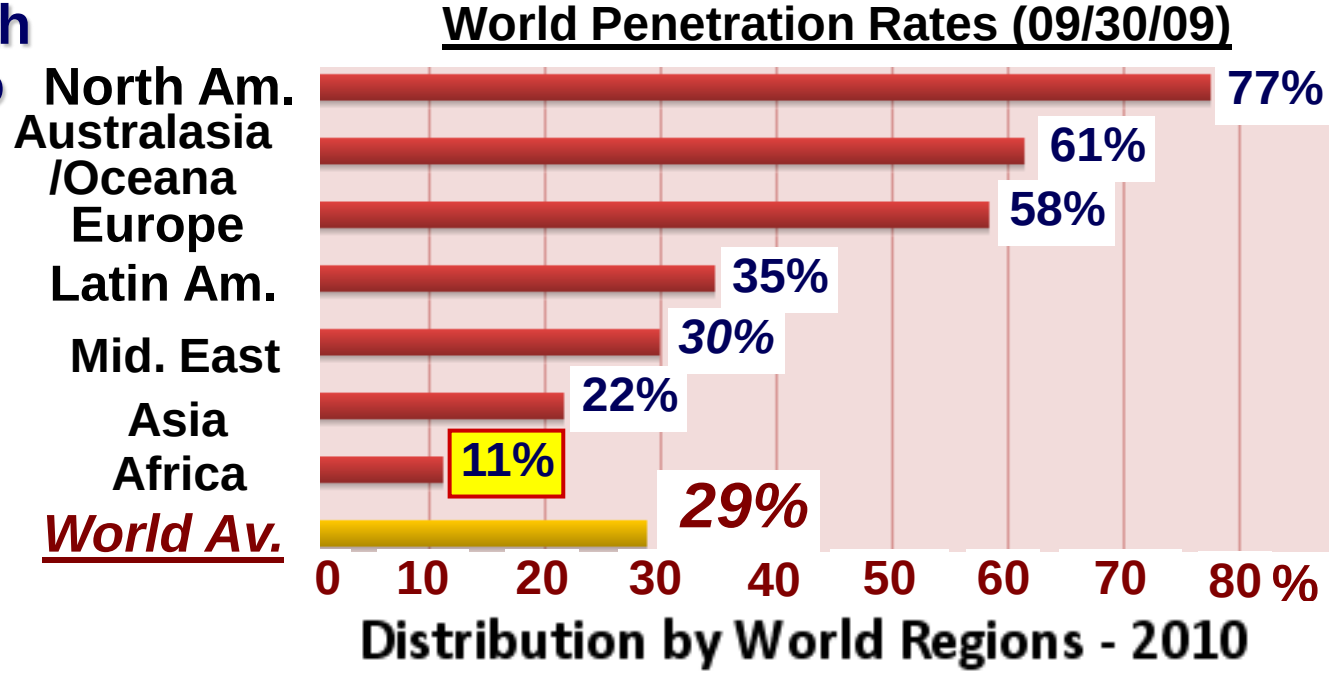
- Current generation of 10 Gbps network backbones and major Int'l links arrived in 2002-8 in US, Europe, Japan, Korea; Now *China, Brazil*
  - ➔ Bandwidth Growth: from 16 to >10,000X in 7 Yrs. >> Moore's Law
- Proliferation of 10G links across the Atlantic & Pacific since 2005
  - ➔ Installed Bandwidth for LHC well above 200 Gbps in aggregate
- Rapid Spread of "Dark Fiber" and DWDM: Emergence of Continental, Nat'l, State & Metro *N X 10G "Hybrid" Networks in Many Nations*
- ➔ Point-to-point "Light-paths" for HEP and "Data Intensive Science"
  - ➔ *Now Dynamic Circuits; Managed Bandwidth Channels*
- Technology continues to drive Performance Higher, Costs Lower
  - ➔ Commoditization of GE now 10 GE ports on servers; 40 GE starting
  - ➔ Cheaper and faster storage (< $100/Tbyte); 100+ Mbyte/sec disks
  - ➔ Multicore processors with Multi-Gbyte/sec interconnects
- *Appearance of terrestrial 40G and 100G MANs/WANs:*
  - ➔*40G optical backbones in commercial and R&E networks*
  - ➔*100G pilots/tests in 2009-10, first service deployments in 2011*
- Transition to 40G, 100G links: by 2011-12 (on land), ~2012-13 (undersea)
- Outlook: *Continued growth in bandwidth deployment & use*

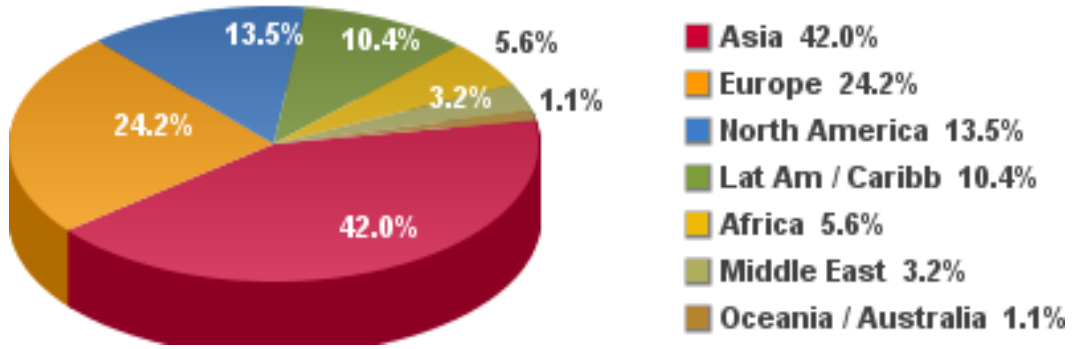
"Long Dawn" of the Information Age

1.97B Internet Users; 550M with Broadband (6/30/10)

http://internetworldstats.com

- Explosion of bandwidth use: ~6,000 PBytes/mo
- Rise of broadband
- Rise of Video + Mobile Traffic: ~20 Exabytes Per mo. (64%) by 2013
- Web 2.0: Billions of Web Pages, embedded apps.
  - Facebook, Twitter, Skype; 4G Mobile
- Beginnings of Web 3.0: Social, streaming, SOA; ubiquitous information
- Broadband as a driver of modern life: from e-banking to e-training to e-health





Broadband: 100M+ in China, 84M in US

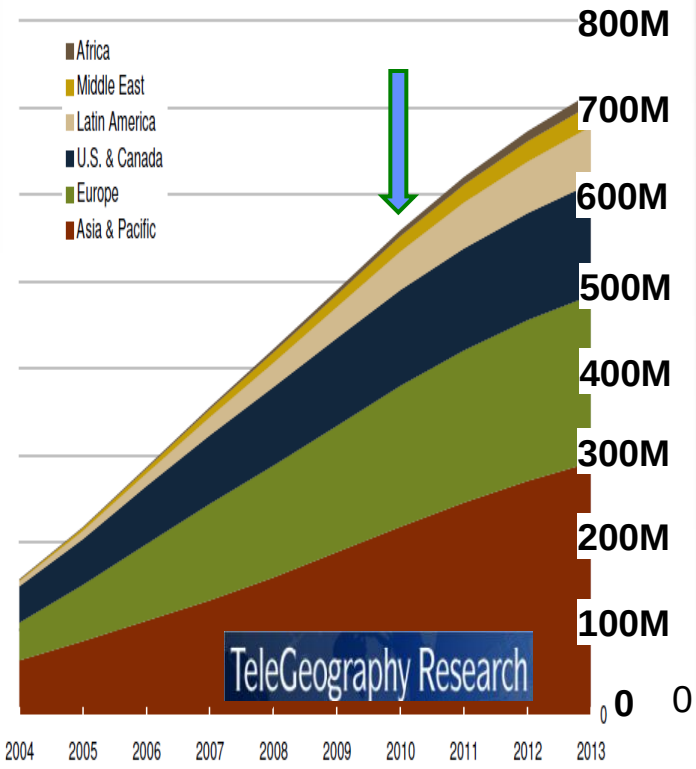
ITU: Announces A World Broadband Plan 9/2010
Closing the New Digital Divide

http://www.broadbandcommission.org

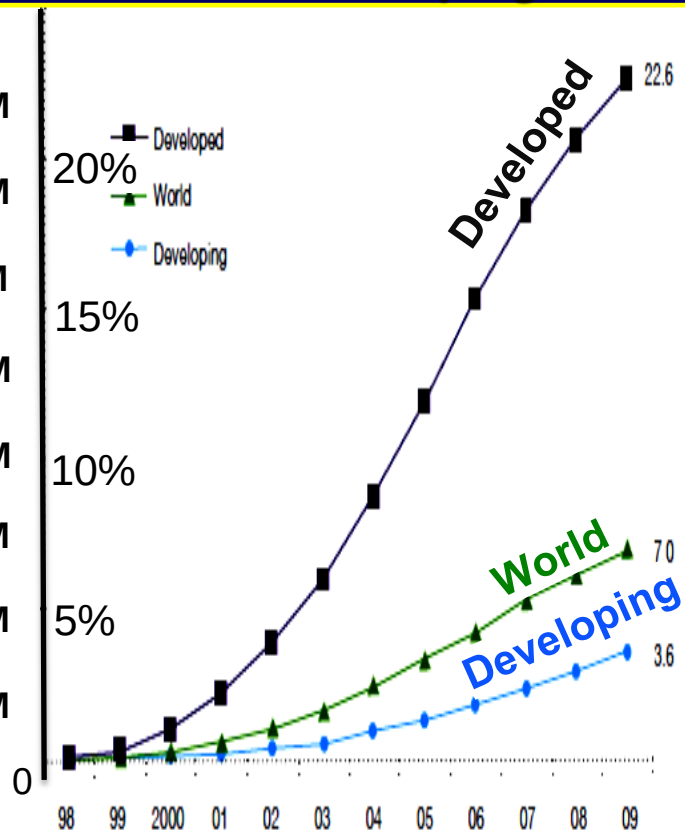
Goal: 50% of World Population with Broadband by 2015

Reason #1 for Growth:
Broadband Subscriber Increases

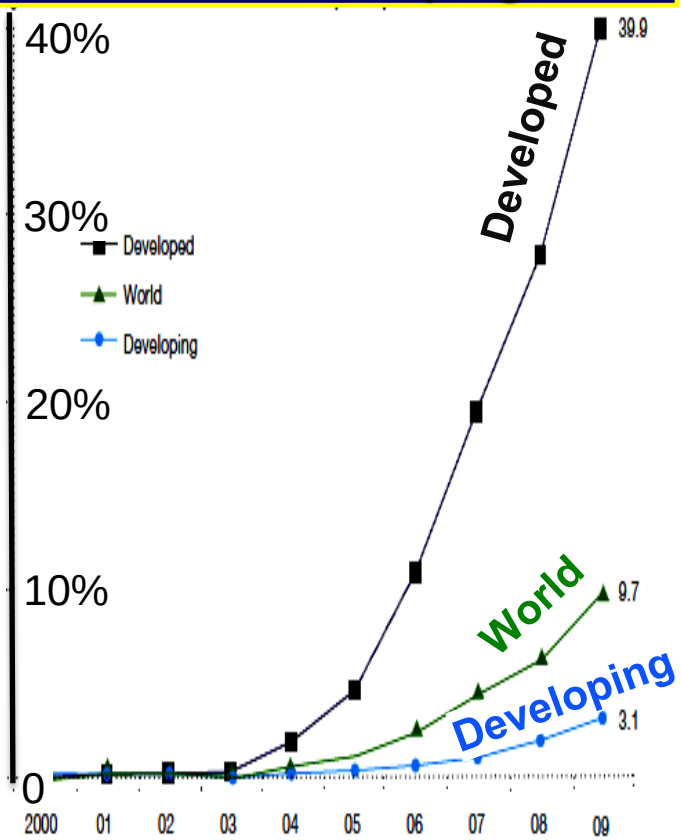
Broadband Subscribers by Region



Fixed Broadband per 100 Inhabitants
23% in the developed world
3.6% in the developing world



Mobile Broadband per 100 Inhabitants
40% in the developed world
3.1% in the developing world



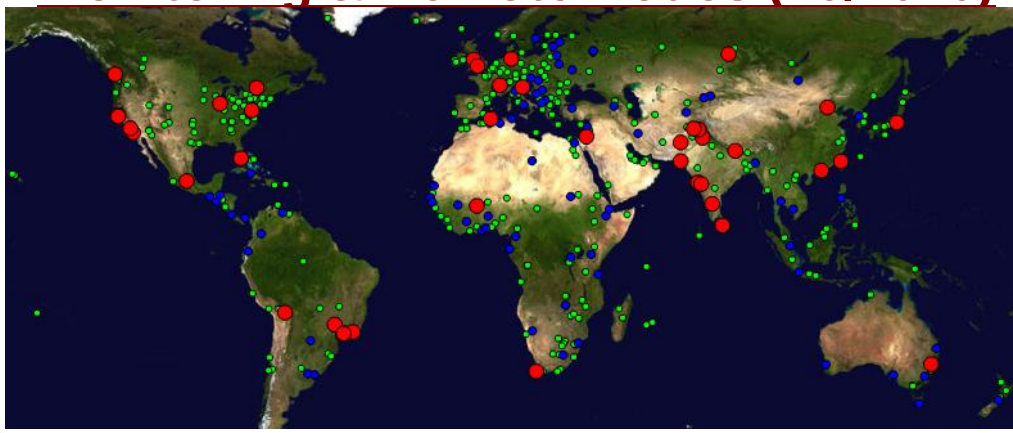
SCIC Monitoring WG
PingER (Also IEPM-BW)

R. Cottrell

IEPM
Internet End-to-end Performance Monitoring

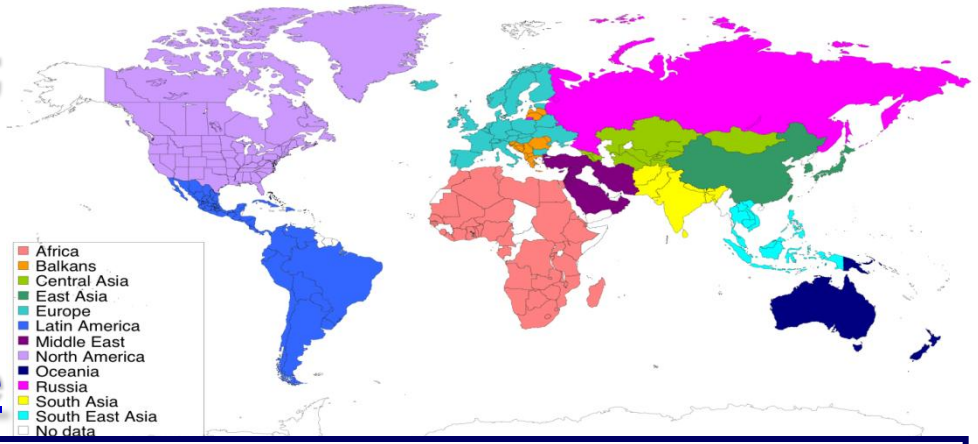
ICFA SCIC

Monitoring & Remote Nodes (10/2010)



- **Measurements from 1995 On**
  - *Reports link reliability & quality*
- **Countries monitored**
  - ➔ Contain 98% of world pop.
  - ➔ 99% of World's Internet Users
- 930 remote nodes at 786 sites in 164 nations; 55 monitoring nodes; 169 nodes in 50 African countries
- Strong Collaboration with ICTP Trieste and NUST/SEECS (Pakistan)
- Excellent, Vital Work; Funding issue

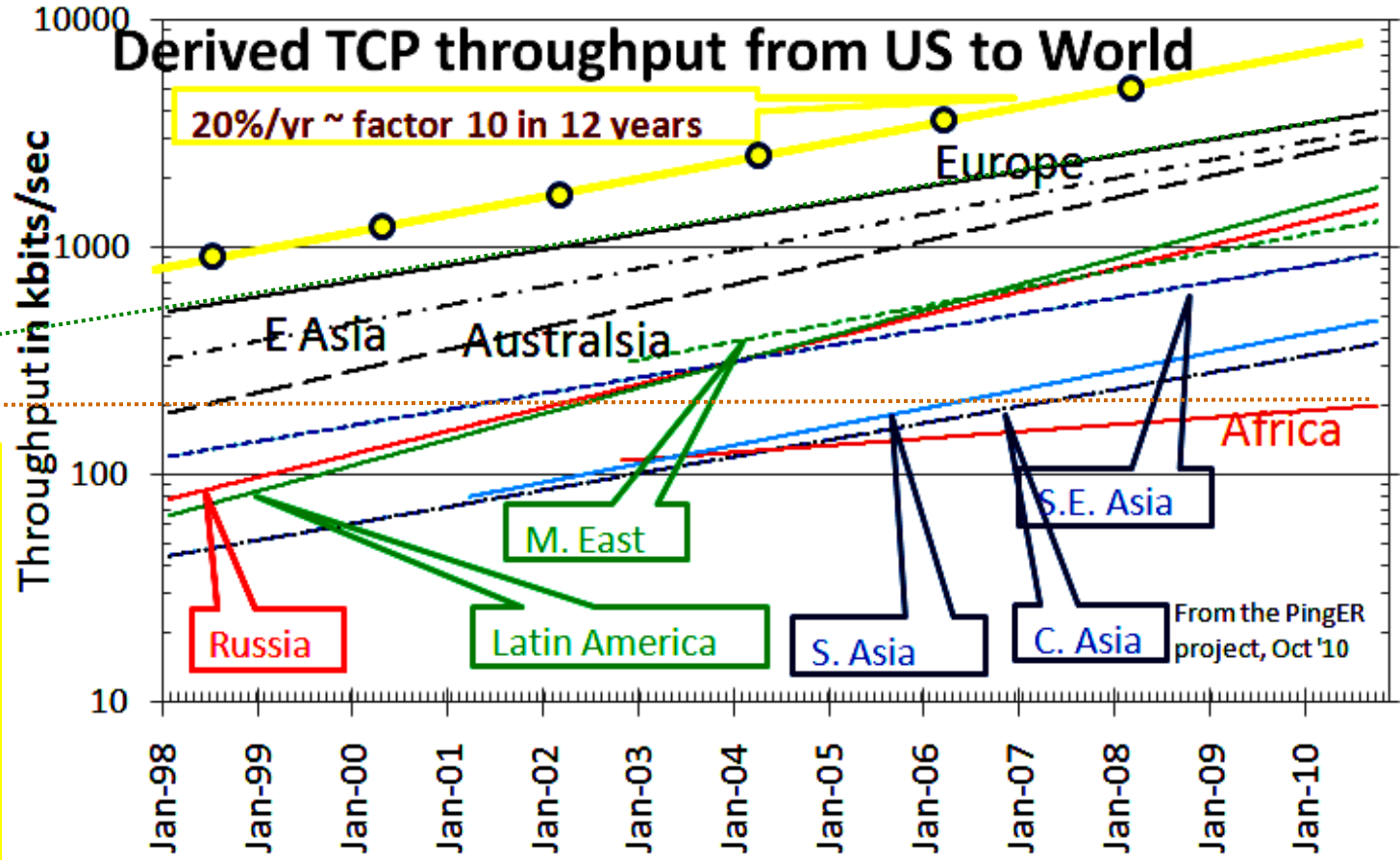
PingER Regions



Countries: N. America (3), Latin America (*21*), Europe (30), Balkans (*10*), Africa (*50*), Middle East (*13*), Central Asia (9), South Asia (*8*), East Asia (4), SE Asia (10), Russia (1), China (1) and Oceania (4)

SCIC Monitoring WG: Throughput improvements 1998-2010

IEPM
Internet End-to-end Performance Monitoring



Top 4
Europe, N. America, East Asia & Australasia

Behind Europe

5 Yrs: Russia, Latin America, Middle East

8 Yrs: SE Asia

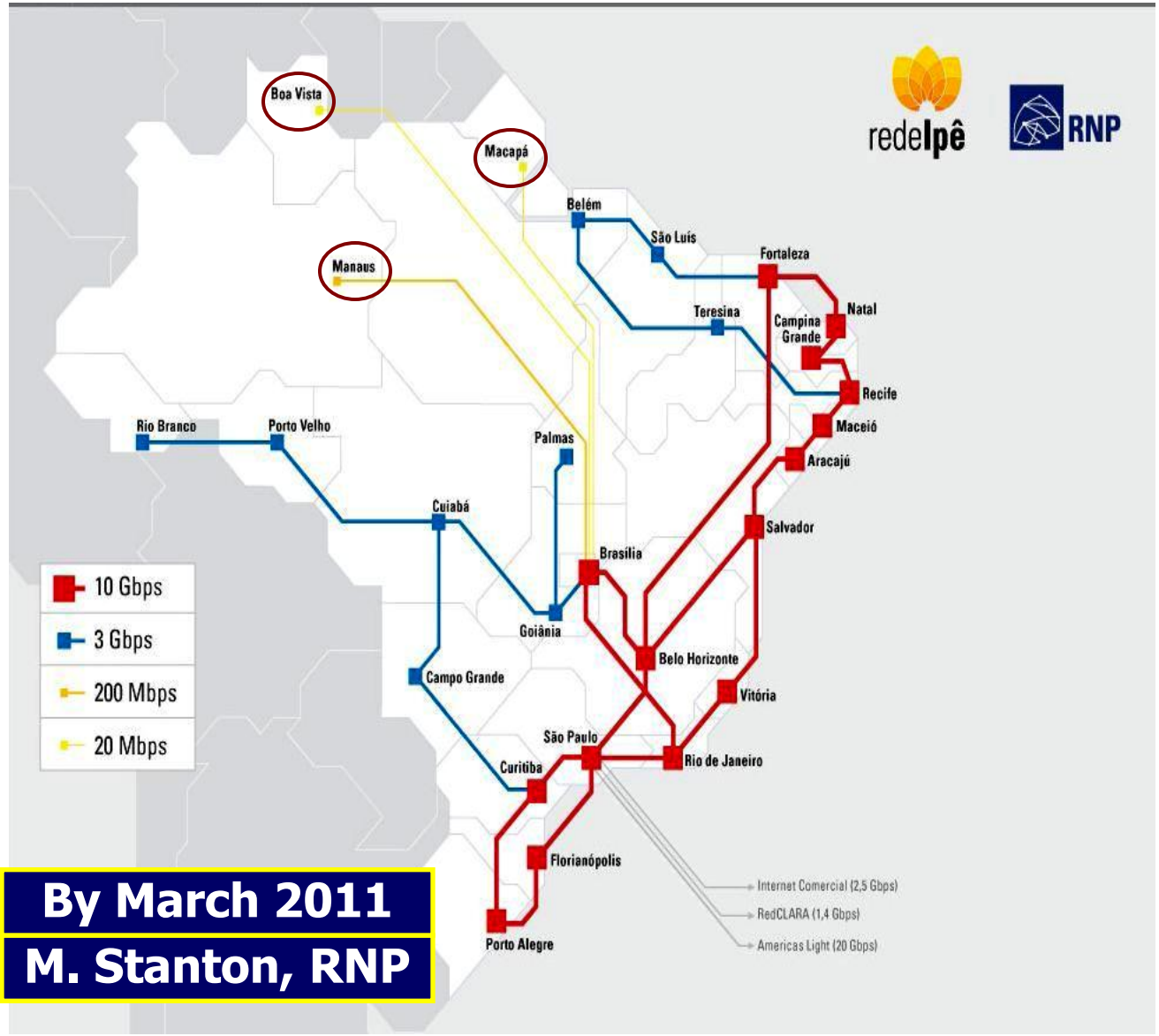
13 Yrs: S. Asia, C. Asia

18 Yrs: Africa

Derived TCP Throughput = 1460 Bytes*8bits/Byte/ (RTT * sqrt(loss))
Mathis et. Al.

In 10 years:
Russia and Latin America should catch up with top 4.
Africa falling further behind, factor 60 behind East Asia

Brazil in 2011: Next-Generation "Ipê" 10G Core Network



- Oi Telco Providing 29,000 km of fiber to RNP; +Free OPEX
- 29 10G or 3G Waves
- Will connect 24 of 27 state capitals by 2011
- Hydroelectric power lines, and optical fibers will reach the 3 northern capitals by 2013

2nd Continental-Scale Transformation Since 2005

By March 2011

M. Stanton, RNP

POLAND: PIONIER 6000 km Dark Fiber Network in 2010

LCG/EGEE

POLTIER2

Distributed Tier2

(Poznan, Warsaw, Cracow) Connects to Karlsruhe Tier1

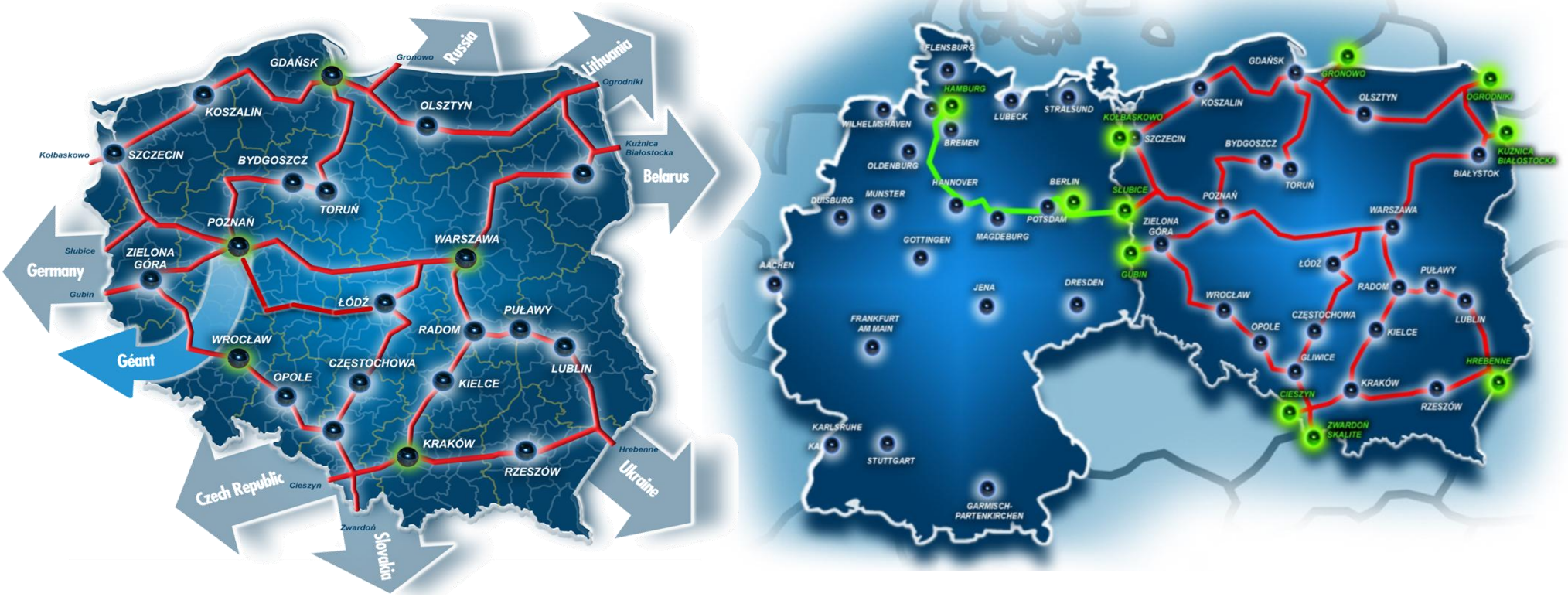
Cross Border Dark Fiber Links

to Russia, Ukraine, Lithuania, Belarus, Czech Republic, and Slovakia

2,4, or 6 X 10G Among 24 Major Univ. Centers

R. Lichwala

PIONIER: Direct Connections to DFN, SURFnet, NORDunet



Cross Border Dark Fiber Links At 4 X 10G

R. Lichwala

SLOVAK Academic Network
All 10 GbE Switched Ethernet

sanet.
Slovak Academic Network

~10,000x Increase 2002-2010

SANET to Schools 1GE to 500 Schools In 54 Cities By 2012 (92 schools connected in 2009)

http://www.sanet.sk/en/index.shtm

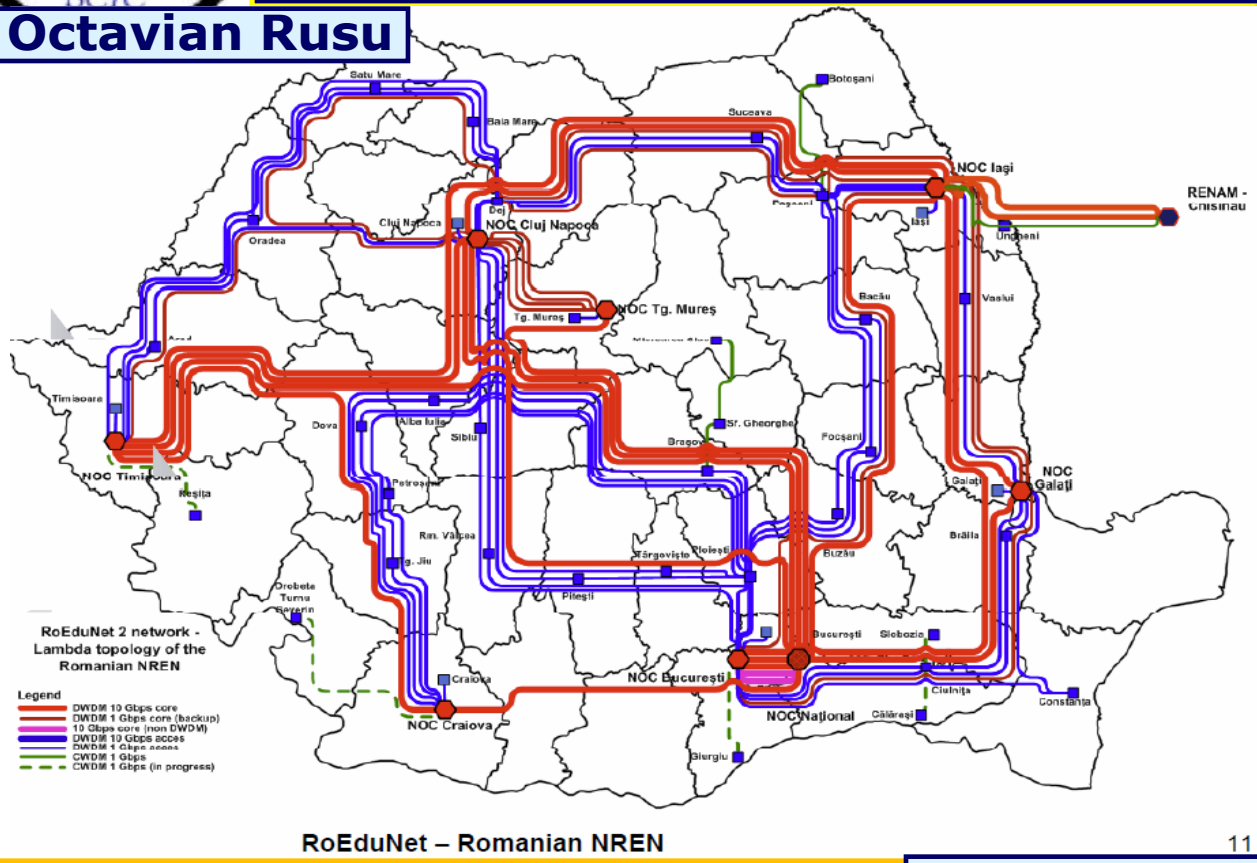
Weis Horvath

- ❒ 2002 - 2004: Dark Fiber Links to Austria, Czech Republic, Poland
- ❒ 2005-6: Complete 1 GbE links to all main sites
- ❒ 2006: 10 GbE Cross-Border Dark Fiber to Austria & Czech Republic; 8 X 10G over 224 km with Nothing In-Line Demonstrated
- ❒ 2007-10: Transition Backbone to 10G Done; All CB Dark Fibers to 10G

RoEduNet2 (ROMANIA)

> 10,000X Since 2002: Pan-European "Role of Science in the Information Society" Ministerial Meeting with HEP Bucharest

Octavian Rusu



RoEduNet – Romanian NREN

11

4240 Km Dark Fiber

600 Km with WDM

38 10G + 41 1G Waves

56 Sites

Separate Optical Control Plane (Nortel);

No regeneration: Up to 1000 km spans

Cross Border Dark Fiber to Modova

www.ces.net/events/2010/cef/p/rusu.pdf

2001 – RoEduNet joined GEANT as partner
2006 – RoEduNet2 project approved
2007 – New modern data centers in Bucharest: National NOC and Bucharest NOC
2007 – More than 40 new routers installed in network, layer 3 of network completely upgraded
2008 August – GEANT POP installed in Bucharest: 10 Gbps to GEANT, 2.5 Gbps committed
2008 December – RoEduNet2 network in production
2010 – 1st CBF from Romania installed: Iasi – Chisinau (Moldava) DWDM segment operational

USA-RUSSIA-CHINA-KOREA-NETHERLANDS-CANADA-DENMARK-FINLAND-ICELAND-NORWAY-SWEDEN

INDIA-EGYPT-SINGAPORE

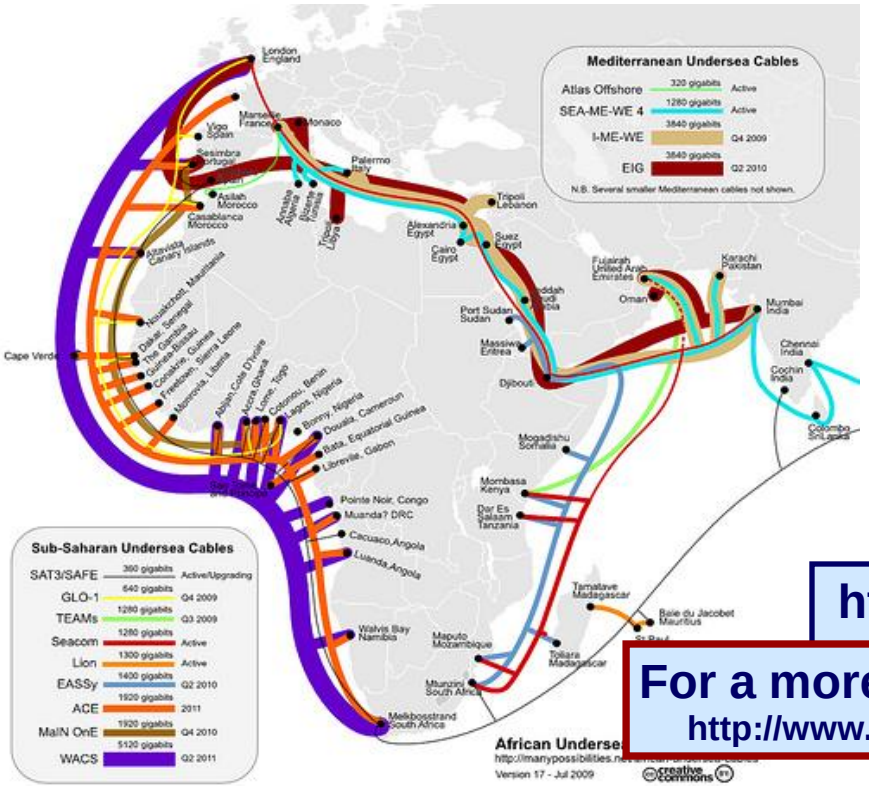
GLORIAD-Taj Expansion

October 14, 2009

The National Science Foundation (NSF)-funded Taj network has expanded to the Global Ring Network for Advanced Application Development (GLORIAD), wrapping another ring of light around the northern hemisphere for science and education.

Taj now connects India, Singapore, Vietnam and Egypt to the GLORIAD global infrastructure and dramatically improves existing U.S. network links with China and the Nordic region.

New African Undersea Cables to Europe, India, Middle East



- Ambitious plans again underway to better-connect African continent, both East & West.
- Potential increase in capacity 1000X: to multi-Terabit/s range.
- Seacom, EASSy, TEAMS, MainOne already in production
- Spurred by the World Cup: Outlook is some of these will succeed

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

For a more comprehensive map (with terrestrial fiber):
http://www.ubuntunet.net/sites/ubuntunet.net/files/Intra-Arica_Fibre_Map_v6.pdf

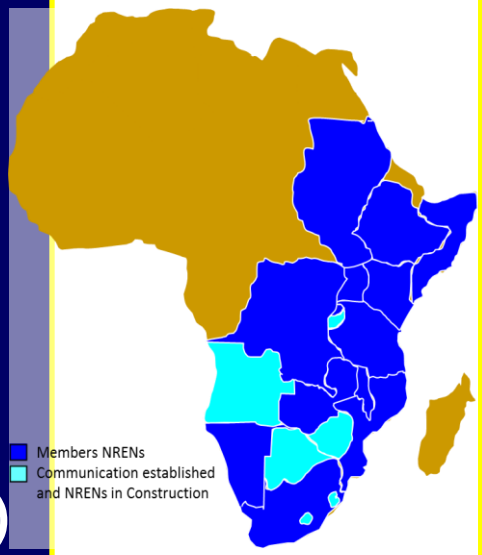
Seacom	EASSy	TEAMs	WACS	MainOne	GLO1	ACE
$ 650M	$ 265M	$ 130M	$ 600M	$ 240 M	$ 800 M	$ 700M
13.7 kkm	10 kkm	4.5 kkm	13 kkm	14 kkm	9.5 kkm	12 kkm
1.28 Tbps	3.84 Tbps	1.28 Tbps	3.84 Tbps	1.92 Tbps	2.5 Tbps	5.12 Tbps
July 2009	July 2010	Sept. 2009	Q3 2011	Q2 2010	Q3 2010	Q2 2012

The UbuntuNet Alliance www.ubuntunet.net

UbuntuNet Alliance

13 Eastern and Southern Africa NRENs

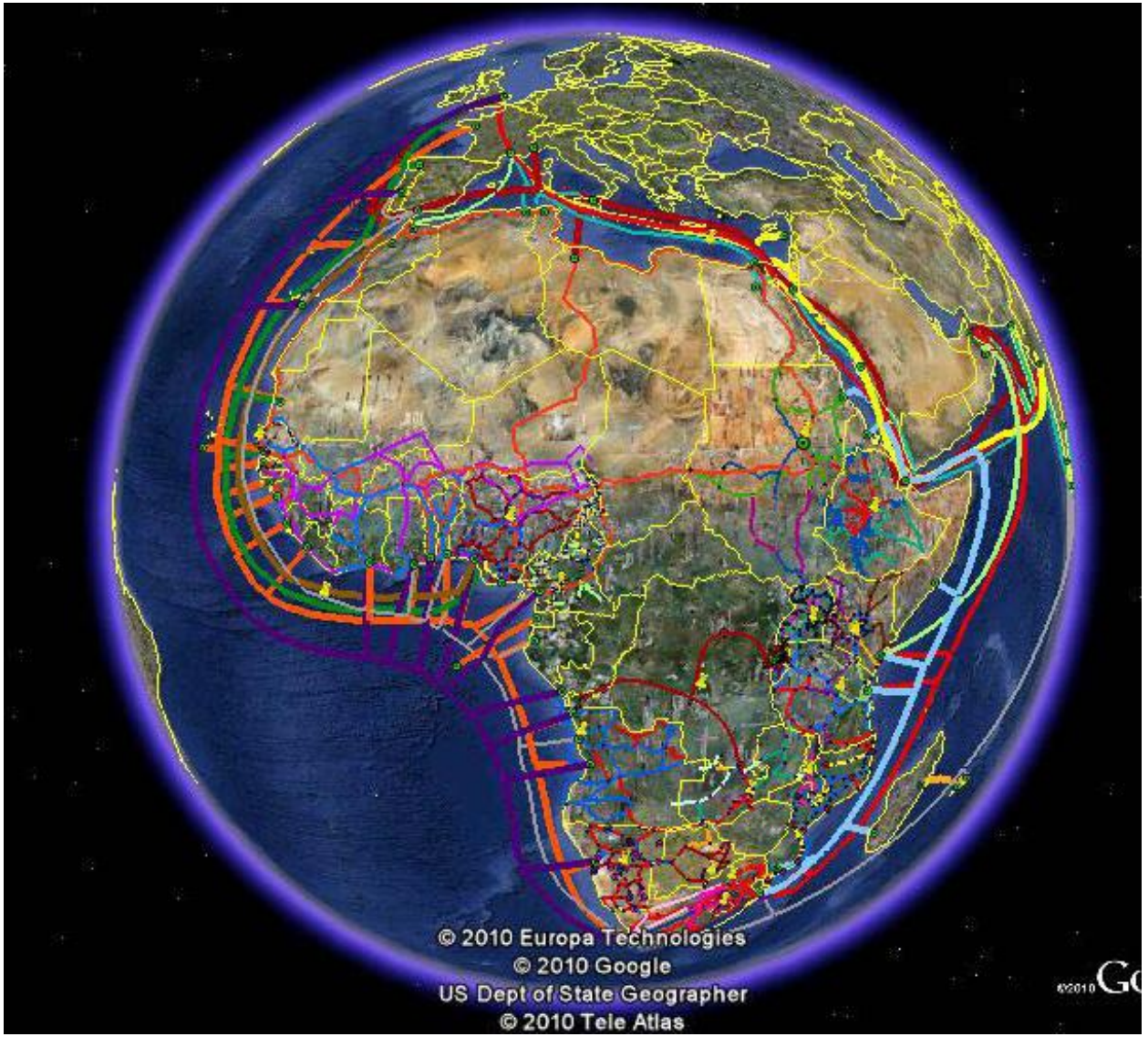
Eb@le (Rep. of Congo)
EthERNet (Ethiopia)
KENET (Kenya)
MAREN (Malawi)
MoRENet (Mozambique)
RENU (Uganda)
RwEdNet (Rwanda)
SomaliREN (Somalia)
SUIN (Sudan)
TENET (South Africa)
TERNET (Tanzania)
Xnet (Namibia)
ZAMREN (Zambia)



Key developments during Jan 2010 – Jan 2011 include

- Growth of membership to thirteen members, the latest being Xnet, the Namibian NREN;
- Securing euro 15million in support from the European Union Commission, through the African Union Commission, for rolling out the regional network. 20% of this will be provided by the member NRENs of the Alliance. Implementation will start during the first quarter of 2011.
- Increase in the connections from member NRENs to the Alliance router in London from 64 STM-1s to 69 STM-1s.
- Increasing the interconnection between the Alliance and GÉANT from 1 Gbps to 20 Gbps to cope with the growing traffic. This includes a 10Gbps point to point link to enable high capacity high volume data transfers.

Intra-African Fiber Map

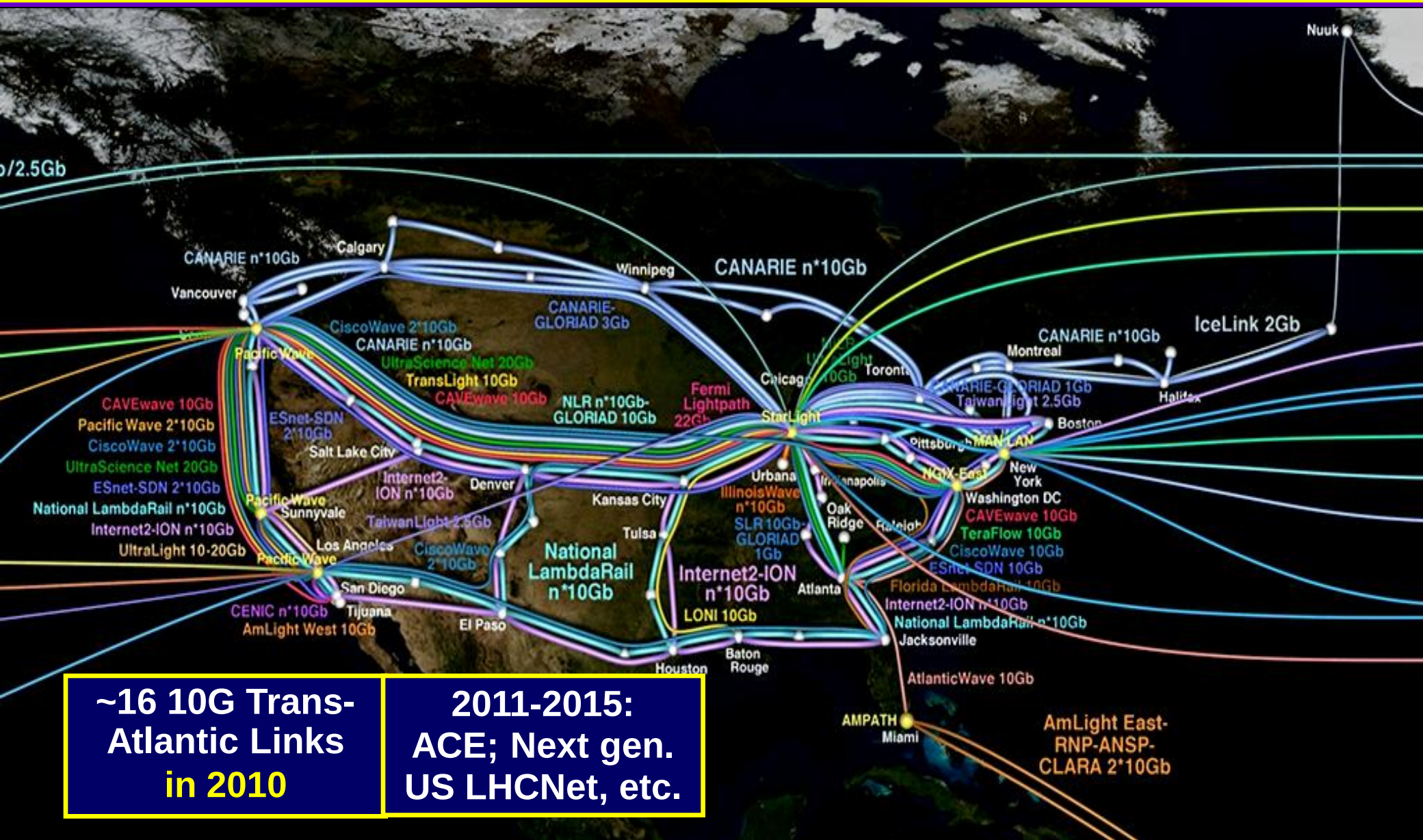


UbuntuNet Alliance

13 Eastern and Southern Africa NRENs

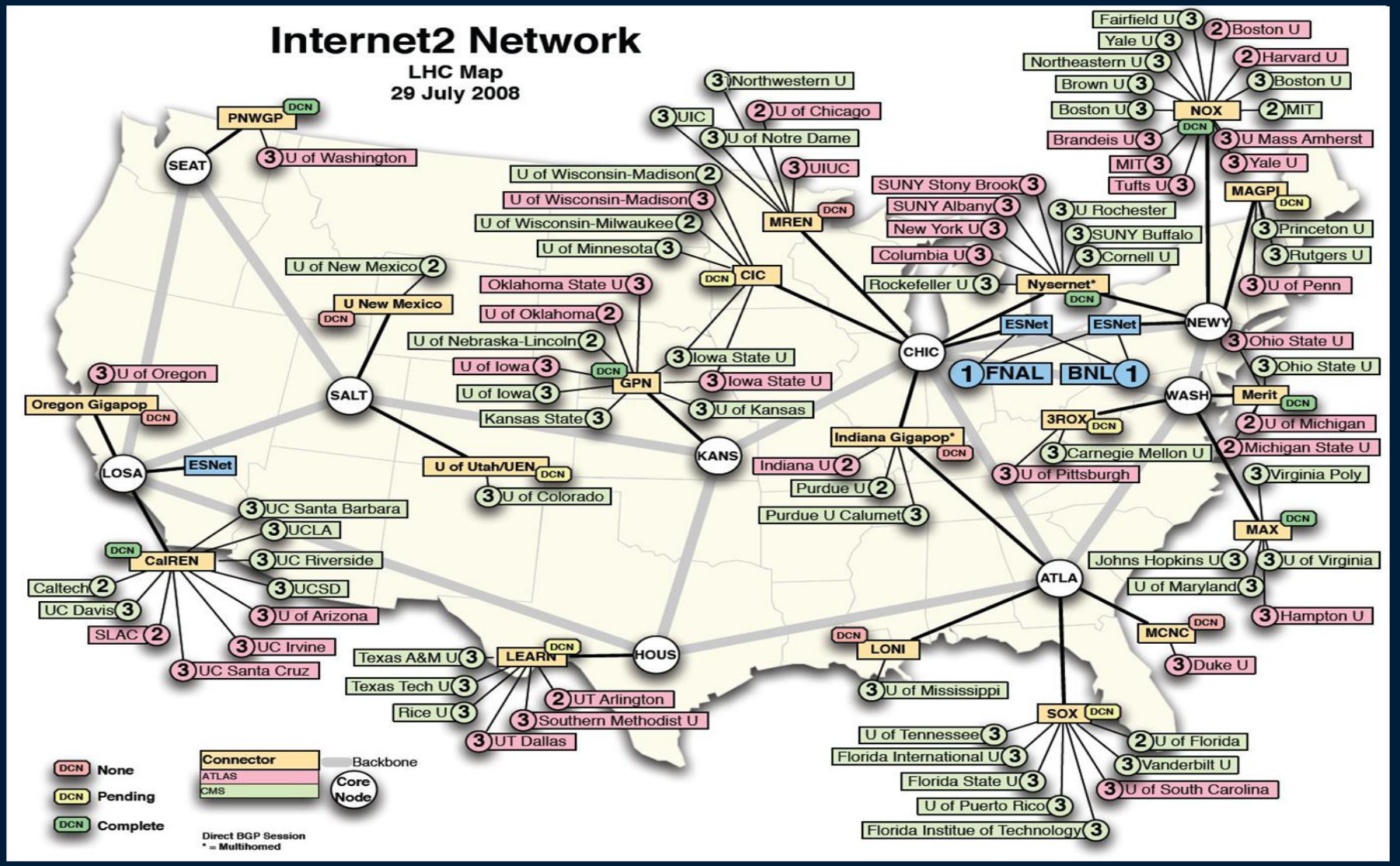
- Eb@le (Rep. of Congo)
- EthERNet (Ethiopia)
- KENET (Kenya)
- MAREN (Malawi)
- MoRENet (Mozambique)
- RENU (Uganda)
- RwEdNet (Rwanda)
- SomaliREN (Somalia)
- SUIN (Sudan)
- TENET (South Africa)
- TERNET (Tanzania)
- Xnet (Namibia)
- ZAMREN (Zambia)

A Global Partnership of R&E Networks and Advanced R&D Projects Supporting HEP



~16 10G Trans-Atlantic Links in 2010	2011-2015: ACE; Next gen. US LHCNet, etc.

Internet2 Dynamic Circuit Network



Path to a Solution for LHC Computing and Networking

- **Define what you need**
  - Excessive use of general purpose networks will cause "defensive action"
  - Implementing int'l network architectures with sufficient reliability and capacity to cope with the traffic growth & flow patterns is not trivial; it needs planning & time.
  - Experiments must work with the network community to create an infrastructure to support T1-T2-T3 matrix flows. [And "Any Data Anywhere ?"]
- **Pay for it**
  - Given an agreed architectural plan with capacity and other objectives (e.g. resilience and adaptability to shifting flows)
    - Cost-optimal solutions can be found
    - The limits of what could be afforded can be understood
    - The funding bodies can plan to support it within feasible cost bounds.
    - The sites can budget to connect.
  - This requires conviction & excellent justification of the costs
- **Integrate it into a System with real end-to-end awareness**
  - From the end-systems to the interfaces to the networks

David Foster, CERN-IT ; Harvey Newman, Caltech