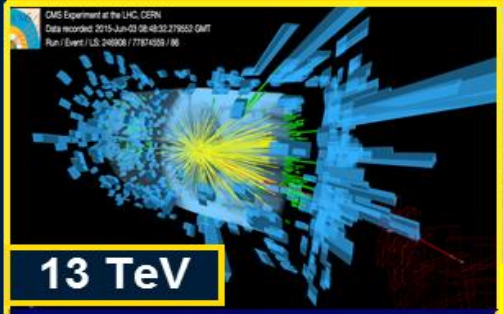
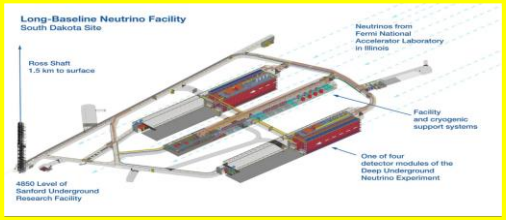
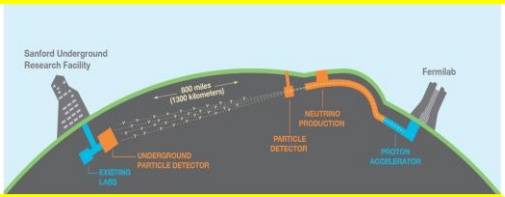
ICFA SCIC and the Global Network Advancement Group:
Next Generation System for LHC, HL LHC & Data-Intensive Sciences



LHC

LSST





LBNF/DUNE

SKA

LHC Run3 and HL-LHC

DUNE

VRO SKA

BioInformatics

Earth Observation

Gateways to a New Era

Harvey Newman, Caltech

Data Lifecycle Panel Meeting

July 16 2024

ICFA Standing Committee on Interregional Connectivity (SCIC)

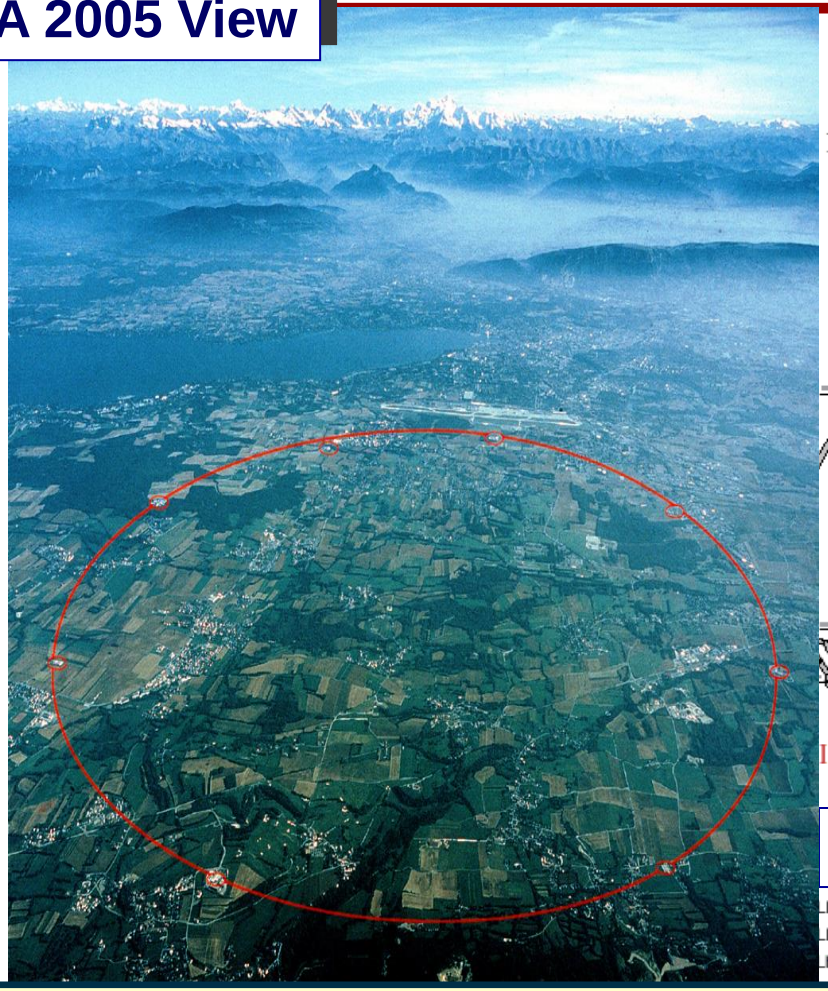
- ICFA Visionary Statement of 1996; ICFA Network Task Force
- SCIC Created in July 1998 in Vancouver [HN Chair since 2002]

CHARGE:

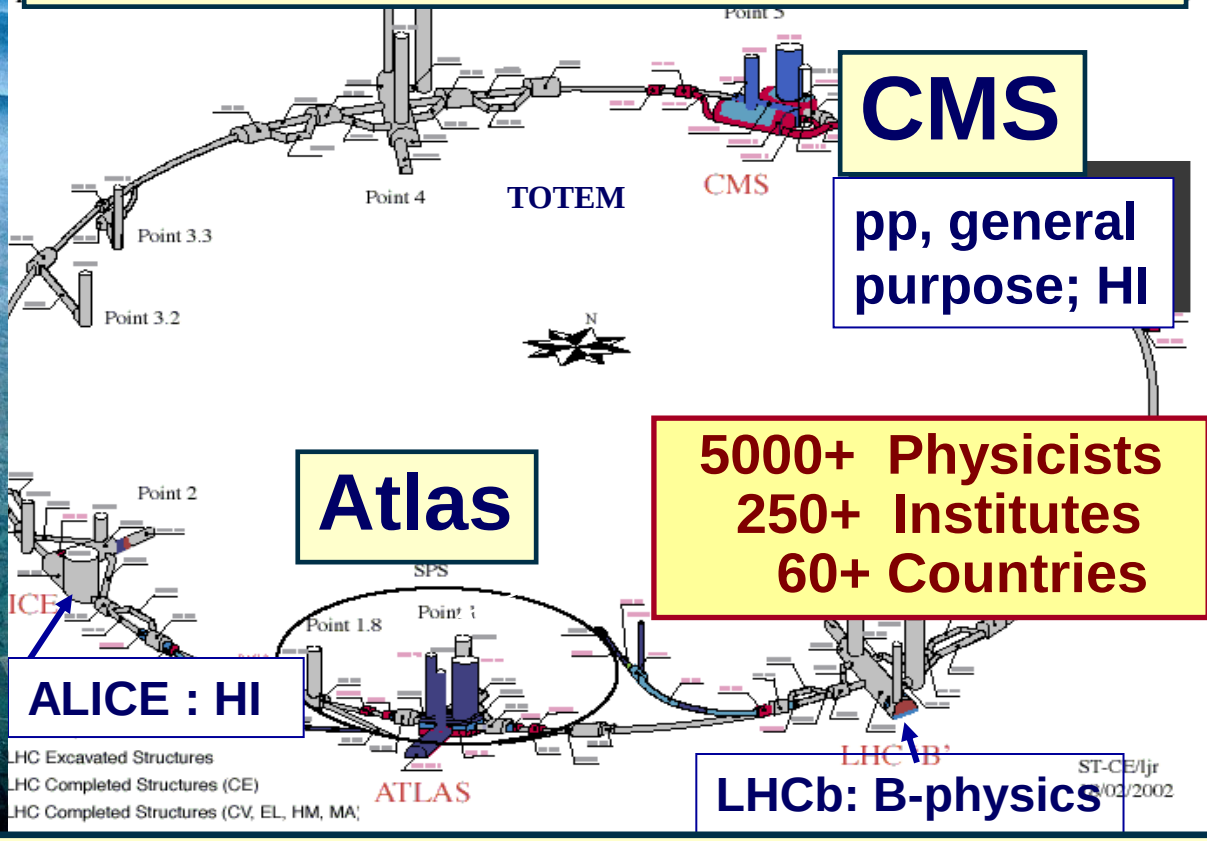
- Make recommendations to ICFA concerning the connectivity between the Americas, Asia and Europe
- As part of the process of developing these recommendations, the committee should
  - Monitor traffic on the world's networks
  - Keep track of technology developments
  - Periodically review forecasts of future bandwidth needs, and
  - Provide early warning of potential problems
- Representatives: Major labs, ECFA, ACFA, North and Latin American Users, Russia, China
- Digital Divide Focus and Many Workshops in 2002-17

Large Hadron Collider at CERN,
Geneva: Plans from 1989; 2009 Start

A 2005 View



- pp $\sqrt{s}$ =14 TeV L=10^{34} cm^{-2} s^{-1}
- 27 km Tunnel in Switzerland & France



CMS

pp, general purpose; HI

Atlas

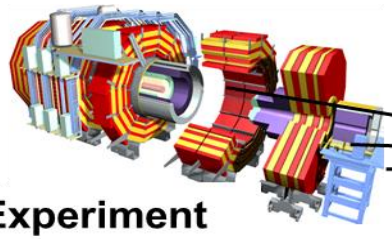
5000+ Physicists
250+ Institutes
60+ Countries

ALICE : HI

LHCb: B-physics

Challenges: Analyze petabytes of complex data cooperatively
Harness global computing, data & network resources

Global Data Flow: LHC Grid Hierarchy
A Worldwide System Invented by Caltech (1999)



Experiment

~PBytes/sec

Online System

~1000 MBytes/sec

13 Tier1, *170 Tier2*
+ 300 Tier3 Centers:
1k to N X 10k Cores

Tier 0

CERN Center
100 PBs of Disk;
Tape Robot

Tier 1

N X100 Gbps

London | Paris | Bologna | Chicago | ...

10 to 100 Gbps

Tier 2

CACR | ? Center | nter | Center | Center

Tier 3

Physics data cache

Institute | tute | stitute | Institute

10 to 100 Gbps

Tier 4

Workstations

- **Tier0:** Real Time Data Processing
  Tier1 and Tier2: Data and Simulation Production Processing
  Tier2 and Tier3: Data Analysis

Increased Use as a Cloud Resource (Any Job Anywhere)
Increased "Elastic" Use of Additional HPC and Cloud Resources
A Global Dynamic System: Fertile Ground for Control with ML

LHC: Discovery of the Higgs Boson and Beyond; 75+ Years of Exploration !

Physicists Find Elusive Particle Seen as Key to Universe

2013 Nobel Prize

Englert	Higgs

	Energy Frontier	Intensity Frontier	Cosmic Frontier
Higgs Boson	●		
Neutrino Mass		●	●
Dark Matter	●	●	●
Cosmic Acceleration			●
✸ Explore the Unknown	●	●	●

48 Year Search; 75 Year Exploration

Theory (1964): 1950s – 1970s;

LHC + Experiments Concept: 1984

Construction: 2001; Operation: 2009

Run1: Higgs Boson Discovery 2012

Run2, Run3 and Going Forward: Precision Measurements and Beyond SM Exploration: 2013 - 2042

Advanced Networks Were Essential to Higgs Discovery and Every Ph.D Thesis; They will be Essential to All Future Discoveries

- NOTE: ~90% of Data Still to be Taken
- HL LHC in 2029-40: 3-5X Greater Intensity: Upgraded detectors for more complex events
- *To 25X Data Rate to Storage*

SCIC Work Areas

2015 Snapshot

- *Closing the Digital Divide (Workshops 2002-17)*
  - Monitoring the world's networks, with a focus on the Divide; work towards greater equality of scientific opportunity
  - Encouraging the development of national advanced R&E network infrastructures: *through knowledge sharing, and joint work*
  - Work on throughput improvements; problem solutions
- Advanced network technologies and systems
  - Creation and Development of new network concepts and architectures: with many network partners
    - *LHCOPN, LHCONE*
    - Software defined networking: and OpenFlow; OpenDaylight; ONOS
  - High throughput methods; + community engagement to apply the methods in many countries, for the LHC and other major programs (HEP, LIGO, AMS, et al.)
  - Integration of advanced network methods with experiments' mainstream data distribution and management systems

ICFA Standing Committee on Inter-regional Connectivity

- **Missions in 2024**
  - Bring Issues to the attention of ICFA
  - Track advanced computing, storage, network and associated software technologies; highlight opportunities and coming issues
  - With a focus on major programs: LHC to HL-LHC, LIGO, LSST, SKA, DUNE et al
  - Track and help understand and set requirements via both community meetings (e.g. LHCONE/ LHCOPN) and reviews (e.g. ESnet 2020-21+Updates)
  - Inform and enable the global community to use networks effectively in support of its science goals
    - Enabling communities in all parts of the world to work effectively as full partners in the science
- **R&E Network Development**
  - Work with R&E network partners to help develop the continental, transoceanic and regional network infrastructures
- **Founded the Global Network Advancement Group (GNA-G) and its working groups in 2019-20** as a major venue uniting the R&E networks and science collaborations worldwide behind the mission
  - Chair Data Intensive Sciences WG

Beyond the Basic Infrastructures: Meeting the Challenges

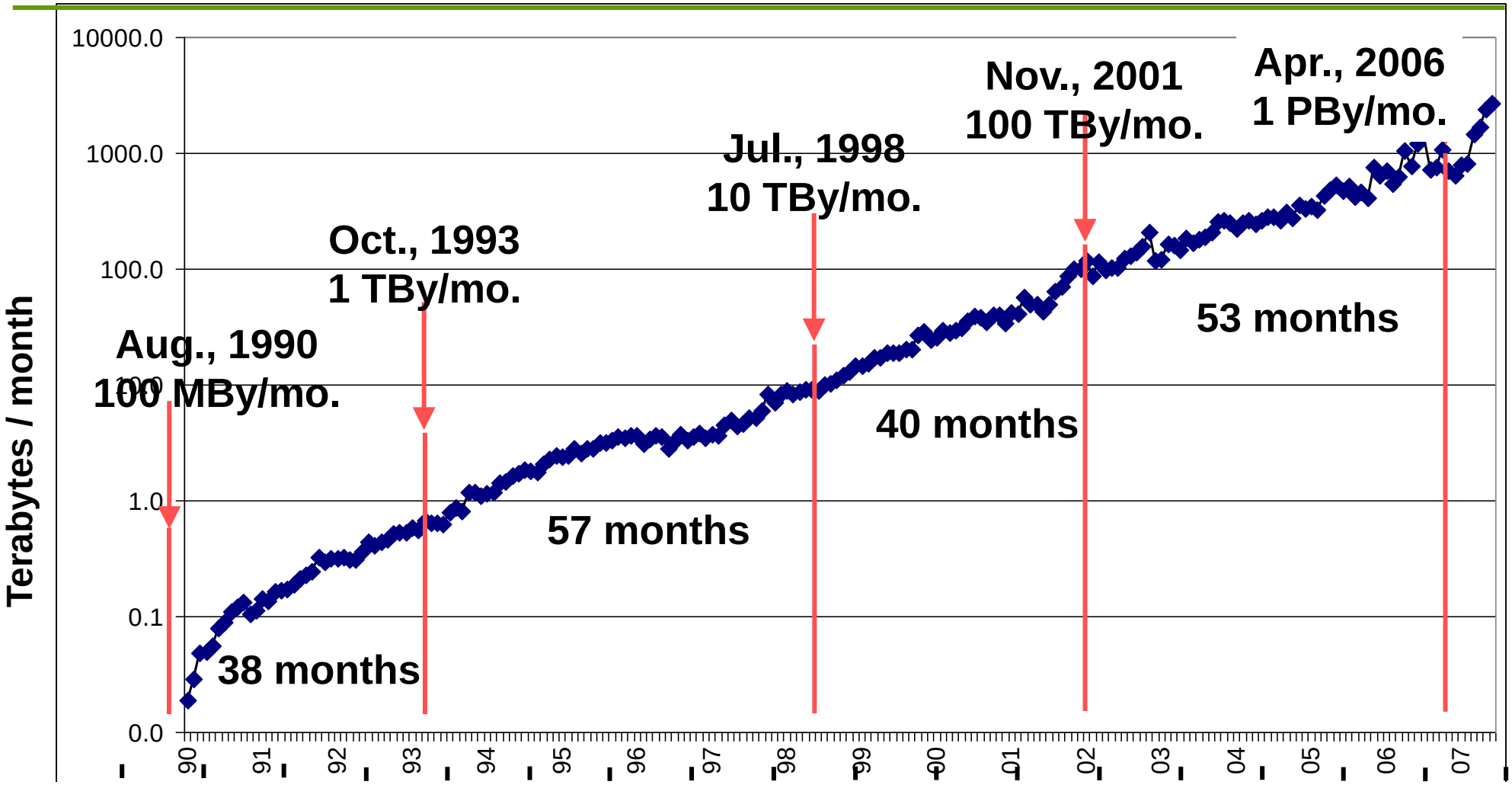
- Formation of a global programmable fabric supporting data intensive research Learning from and going beyond the LHCOPN/LHCONE + DC experience
- Developing integrated systems including networks as a first class resource, across the US, Europe, Asia Pacific and Latin America. Learn from
- Deploying Global Testbeds with Advanced Services: SENSE/AutoGOLE, Global P4 Lab, NSF FABRIC
- Enabling the science programs to meet their needs while accommodating other traffic serving academia and research

Engagement

- With all of the experiments' computing managements, the major R&E network organizations, and advanced network R&D projects supporting data intensive science
- Develop, test and prove the new needed methods and tools; migrate them into production use to meet the requirements
- Engage in persistent proof of concept, prototype, pre-production exercises and demonstrations

ESnet Traffic has Increased by 10X Every 47 Months, on Average, Since 1990

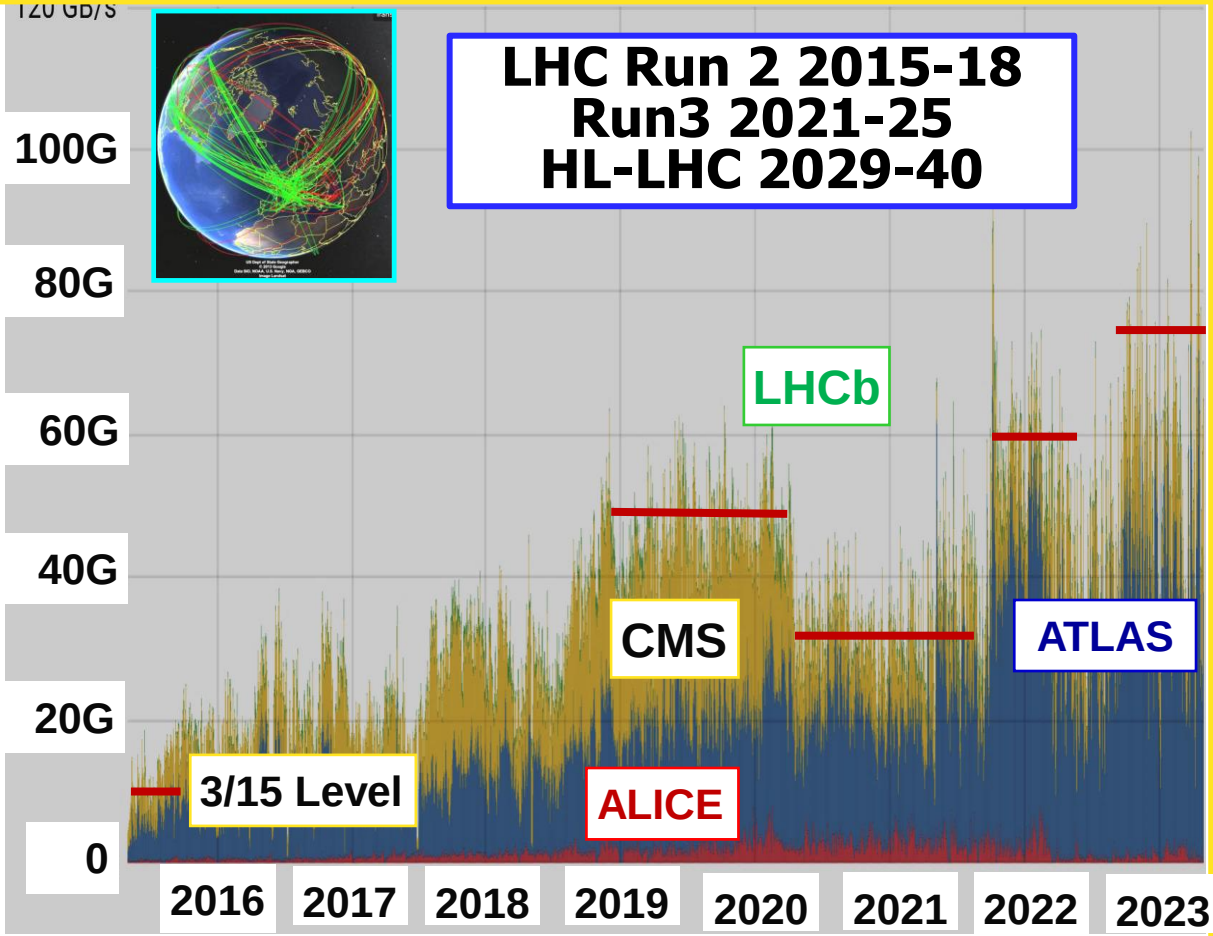
J. Metzger
ICFA DDW07



Log Plot of ESnet Monthly Accepted Traffic, January, 1990 – June, 2007

LHC Data Flows Have *Increased* in Scale and Complexity since the start of LHC Run2 in 2015

WLCG Transfers Dashboard: Throughput April 2015 – April 2023



30-85 GBytes/s Weekly Avg

To 100+ GBytes/s Daily Avg

Complex Workflow

- To ~1.3 M jobs (threads) simultaneously
- Multi-TByte to Petabyte Transfers
- To ~25 M File Transfers/Day
- 100ks of remote connections

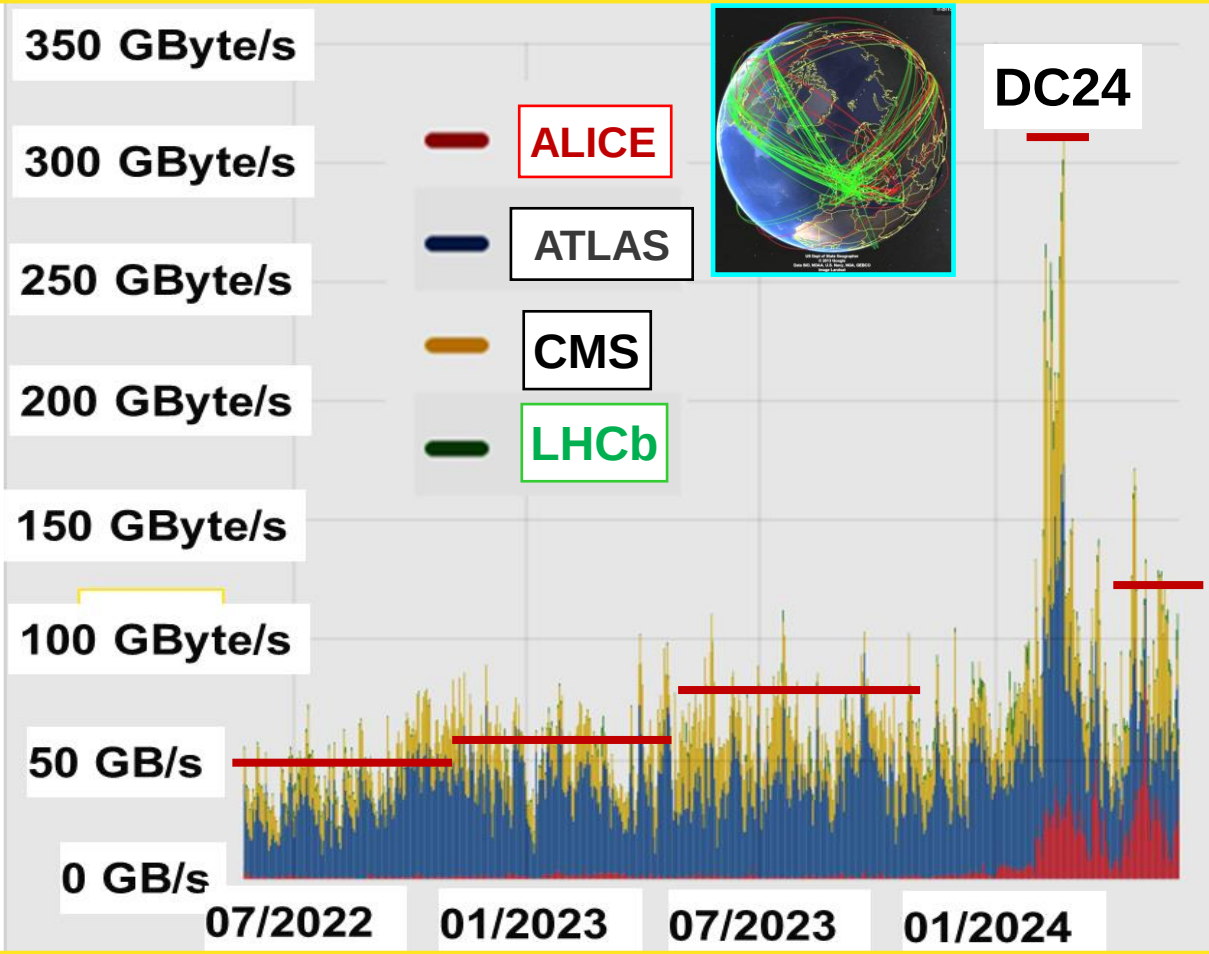
- Effects of Covid from Spring 2020 are evident
- Fast recovery also evident

~12X Growth in Throughput 2015-2023: *+40%/Yr;* *+ Much Faster Growth Bursts*

https://monit-grafana.cern.ch/d/AfdonIvGk/wlcg-transfers?orgId=20&from=now-8y&to=now

LHC Data Flows Increase in Scale and Complexity: Another Burst Upward in 2023-4

WLCG Transfers Dashboard: Throughput June 2022 – May 2024



70-150 GBytes/s Weekly Avg

To 170+ GBytes/s Daily Avg

Complex Workflow

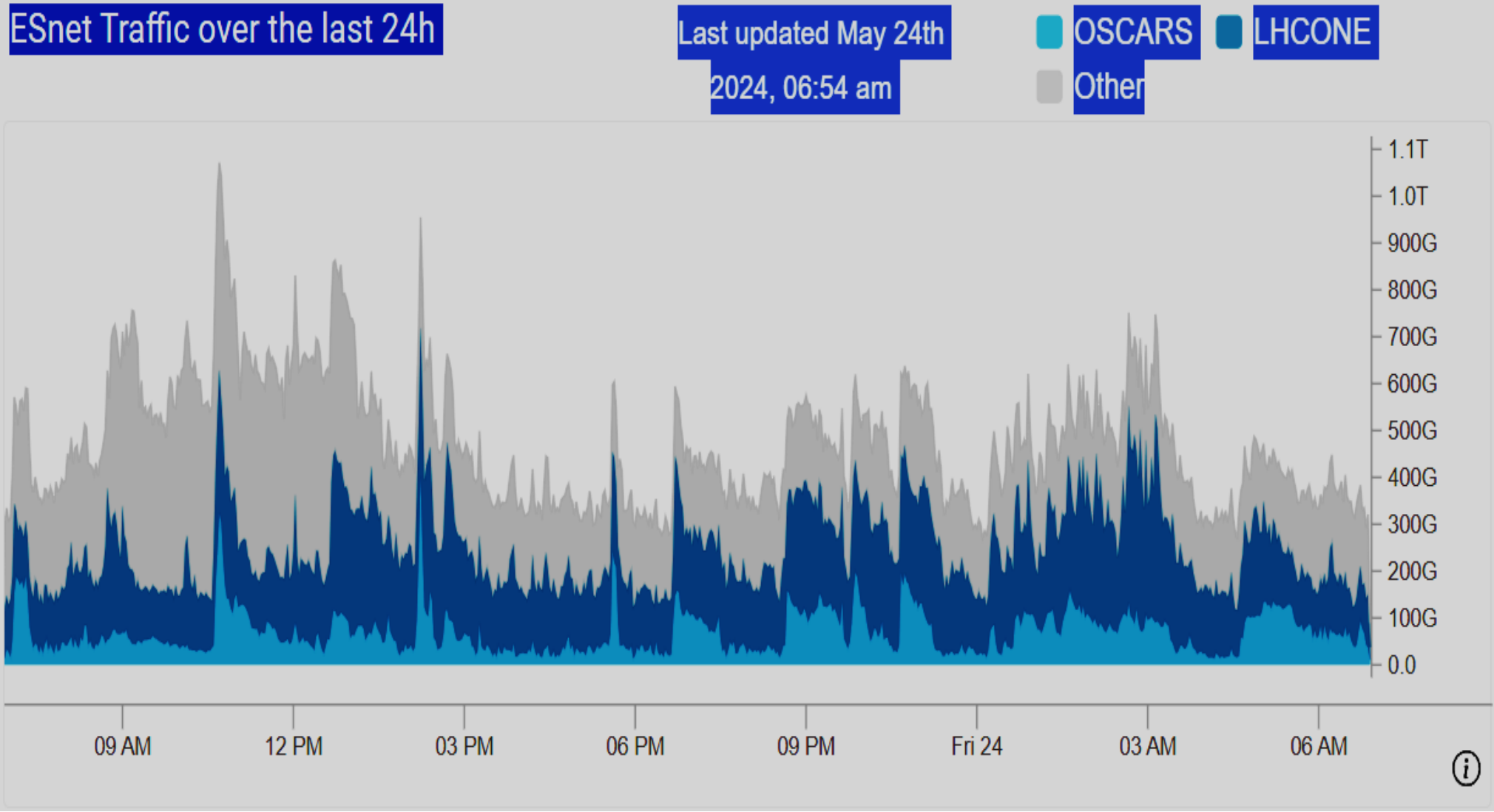
- To ~2 M jobs (threads) simultaneously
- Multi-TByte to Petabyte Transfers
- To ~75 M File Transfers/Day
- Millions of remote connections

- *Another Sea Change in 2023-4*
- 2X in Transfer Rates and Files Transferred
- DC24 (25% HL LHC): 300+ GB/s

~1.8 to 2X Growth in 12 Months: 350-1000X Per Decade Equivalent (?)

https://monit-grafana.cern.ch/d/AfdonIvGk/wlcg-transfers?orgId=20&from=now-2y&to=now

ESnet Today: Traffic 0.3 to 1.1 Terabit/sec
with LHCONE and OSCARs 1/3 to 2/3



Estimates at the time of DC21: Data Rate Table

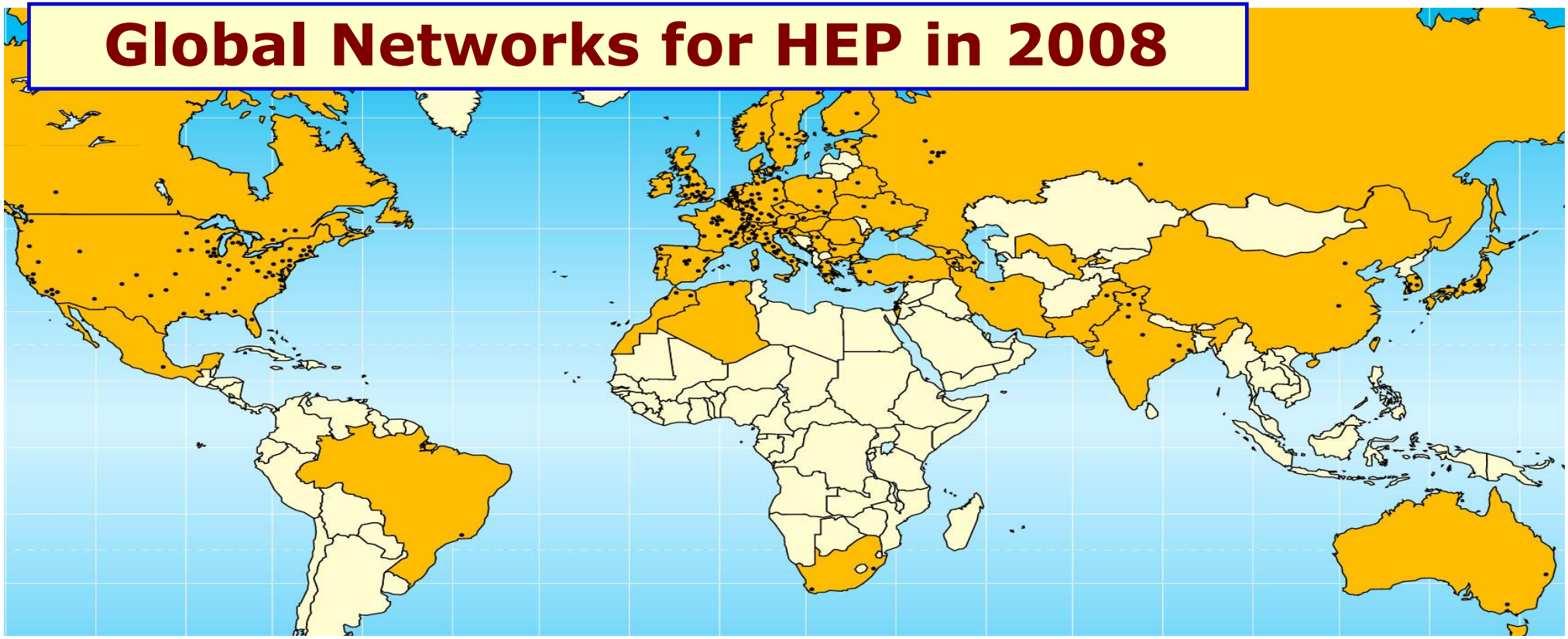
M. Lassnig at WLCG GDB July 12, 2023

- ATLAS & CMS T0 to T1 per experiment
  - 350 PB RAW annually, taken and distributed during typical LHC uptime of 7M seconds / 3 months (50GB/s aka. 400Gbps)
  - Another 100Gb/s estimated for prompt reconstruction data (AOD, other derived output)
  - In total approximately 1Tbps for CMS and ATLAS together
- ALICE & LHCb
  - 100 Gbps per experiment estimated from Run-3 rates
- **Minimal model":** ∑ (ATLAS,ALICE,CMS,LHCb)
  *2 (for bursts) *2 (overprovisioning) = ***4.8Tbps***
- **Flexible model**
- Assumes reading of data from above for reprocessing/reconstruction within 3 months
- Means doubling the Minimal Model: ***9.6Tbps***

- But: Only data flows from the T1s to T2s and T1s accounted for !
  - ***Nota Bene:*** *No MC production flows nor re-creation of derived data included in the 2021 modelling!*
- *ESnet: Requirements Review Update Sessions: July 17 and 26 2023*

ICFA Standing Committee on Interregional Connectivity (SCIC)

Global Networks for HEP in 2008



Harvey B. Newman

California Institute of Technology

ICFA Meeting, DESY
February 11, 2008

SCIC in 2007-8

http://cern.ch/icfa-scic

Three 2008 Reports: An Intensive Year

Rapid Progress, Deepening Digital Divide

- ***Main Report: "Networking for HENP"*** **[HN, D. Nae et al.]**
  - **Includes Updates on the Digital Divide, World Network Status; Brief updates on Monitoring and Advanced Technologies**
- ***34* Appendices: A World Network Overview**
  Status and Plans for the Next Few Years of Nat'l & Regional Networks, HEP Labs, & Optical Net Initiatives
- ***Monitoring Working Group Report*** **[R. Cottrell]**

Also See:

- **TERENA (www.terena.nl) *2007 Compendium:***
  In-depth Annual Survey on R&E Networks in Europe
- **http://internetworldstats.com: Worldwide Internet Use**
- **http://websiteoptimization.com/bw: Broadband (the new Digital Divide ?)**
- ***SCIC 2003 Digital Divide Report*** **[A. Santoro et al.]**

Work on the Digital Divide (2002-17) from Several Perspectives

2008 Snapshot

- Share Information: *Monitoring*, Tracking *BW Progress*; Dark Fiber Projects & Pricing
  - Track Planning (focus on LHC) and Leading Edge Progress
  - Model Cases: Poland, Slovakia, Czech Rep., Brazil, China …
  - Encourage Access to Dark Fiber; Modern technology choices
- Raise Awareness: Locally, Regionally & Globally
  - Many Digital Divide Workshops
  - Diplomatic Events: WSIS, RSIS, Bilateral: e.g. US-India
- Technical Help with Modernizing the Infrastructure:
  - Provide Tools for Effective Use: Data Transport, Monitoring, Collaboration
  - Design, Commissioning, Development
  - *India "Knowledge Network", and Mumbai – CERN Link*
- Encourage, and Work on Inter-Regional Projects
  - GLORIAD, Russia-China-Korea-US-Europe Optical Ring
  - Latin America: CHEPREO/WHREN (US-Brazil); RedCLARA
  - Mediterranean: EUMEDConnect; Asia-Pacific: TEIN2; Asia: ORIENT

HEPGRID and Digital Divide Workshop
UERJ, Rio de Janeiro, Feb. 16-20 2004

NEWS:

Bulletin: ONE TWO
WELCOME BULLETIN
General Information
Registration
Travel Information
Hotel Registration

Tutorials

- C++
- Grid Technologies
- Grid-Enabled Analysis
- Networks
- Collaborative Systems

Theme: Global Collaborations, Grids and Their Relationship to the Digital Divide

For the past three years the SCIC has focused on understanding and seeking the means of reducing or eliminating the Digital Divide, and proposed to ICFA that these issues, as they affect our field of High Energy Physics, be brought to our community for discussion. This led to ICFA's approval, in July 2003, of the 1st Digital Divide and HEP Grid Workshop.

More Information:

http://www.lishep.uerj.br

SPONSORS

CLAF CNPQ FAPERJ UERJ

Sessions & Tutorials Available (w/Video) on the Web

International ICFA Workshop on HEP Networking, Grids, and Digital Divide Issues for Global e-Science

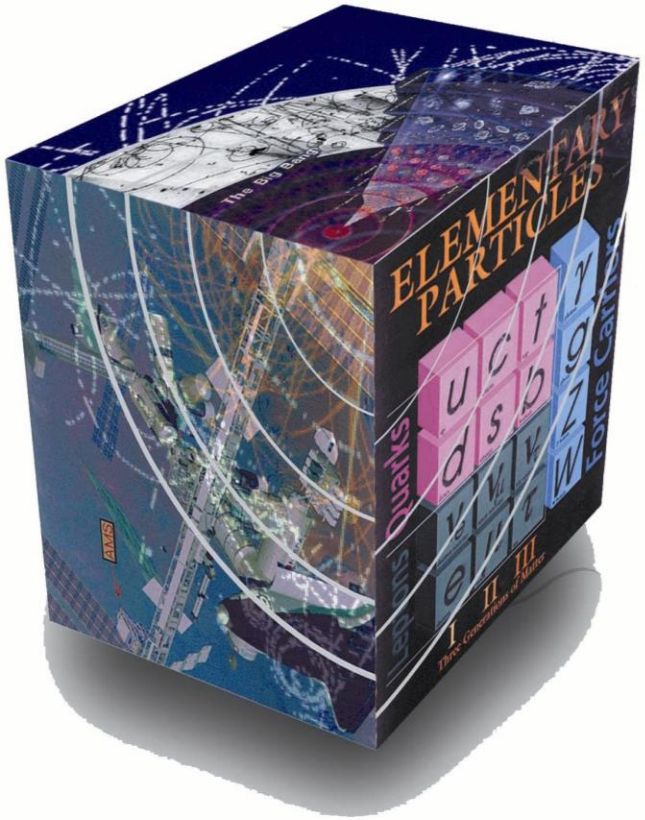
http://chep.knu.ac.kr/HEPDG2005 (Daegu)

Workshop Missions

- Review the status and outlook, and focus on issues in data-intensive Grid computing, inter-regional connectivity and Grid enabled analysis for high energy physics
- Relate these to the key problem of

the Digital Divide

- Promote awareness of these issues in various regions, focusing on the Asia Pacific, Latin America, Russia, and Africa
- Develop approaches to eliminate the Divide and
- Help ensure that the basic requirements for global collaboration are met, related to all of these aspects



SCIC Workshops and Meetings
(12 in 2006-7 Alone)

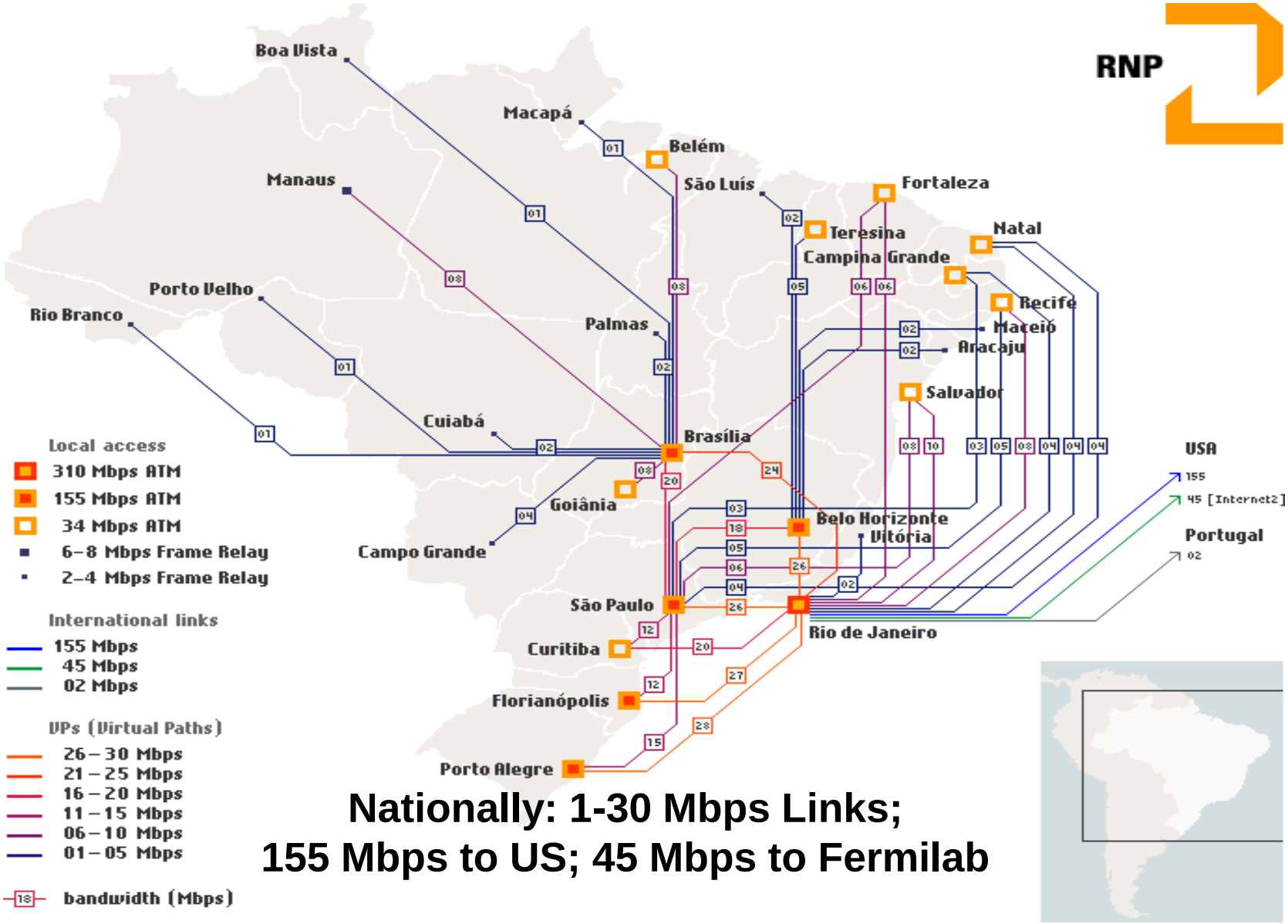
- *October 2006: 3rd ICFA Digital Divide Workshop in Cracow & Sinaia*
- *March 2007: Nat'l Center for Physics Inauguration, Islamabad Pakistan*
- *March – April: Tier3 Network and Data Workshops (5 Workshops)*
- *April 14-17 at APS (Jacksonville): "Bridging the Digital Divide" Sessions, Sponsored by Forum for International Physics*
- *April 26: Internet2 Meeting: Enhancing Research & Education Connectivity to and Within South Asia, Arlington Virginia*
- *June 1: US Liaison Committee to IUPAP, the Nat'l Academies, Wash. DC*
- *Sept. 2-7: Computing for HEP (CHEP07), Victoria Canada*
- *October 8: Internet2 Fall Members Meeting, San Diego*
- *October 24-27: 4th ICFA Digital Divide Workshop, Universidad Iberoamericana, Mexico City*
- *November 10-16: Supercomputing 2007, Reno Nevada*
- *November 11-16: Int'l Heliophysical Year Africa Workshop, Addis Ababa*
- *December 10-12: Internet and Grids in Africa: An Asset for African Scientists for the Benefit of African Society, Montpelier, France*

ICFA Report 2008 Update: Main Trends Accelerate; + *Dynamic Circuits*

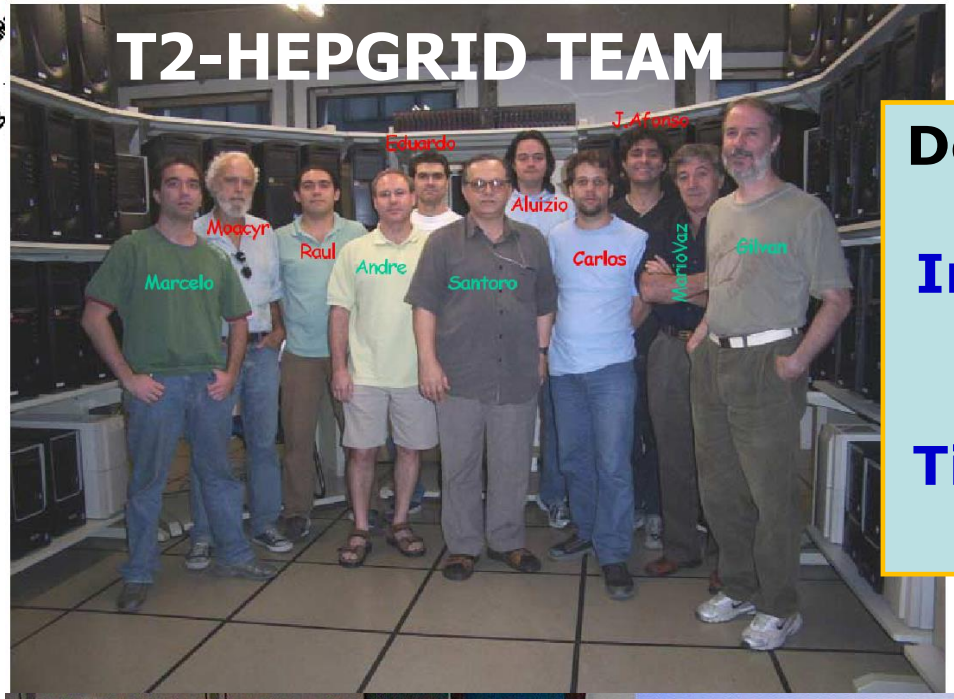
- Current generation of 10 Gbps network backbones and major Int'l links arrived in 2001-7 in US, Europe, Japan, Korea; Now *China*
  - Bandwidth Growth: from 4 to 2500 Times in 5 Years; >> Moore's Law
- Rapid Spread of "Dark Fiber" and DWDM: the emergence of Continental, Nat'l, State & Metro "Hybrid" Networks in Many Nations
  - *Now Dynamic Circuits, and Managed Bandwidth Channels*
- Moving to N X 10G Backbones, complemented by Point-to-point "Light-paths" for "Data Intensive Science", notably HEP
- Proliferation of 10G links across the Atlantic & Pacific; Use of multiple 10G Links (e.g. US-CERN) along major paths began in Fall 2005
  - On track for >10 X 10G networking for LHC, in production by 2007-8
- Technology evolution continues to drive performance higher, equipment costs Lower
  - Commoditization of Gigabit and now 10-Gigabit Ethernet on servers
  - Use of new busses (PCI Express) in PC's and network interfaces in 2006
  - Improved Linux kernel for high speed data transport; multi-CPUs

➡ Outlook: *Continued growth in bandwidth deployment & use*

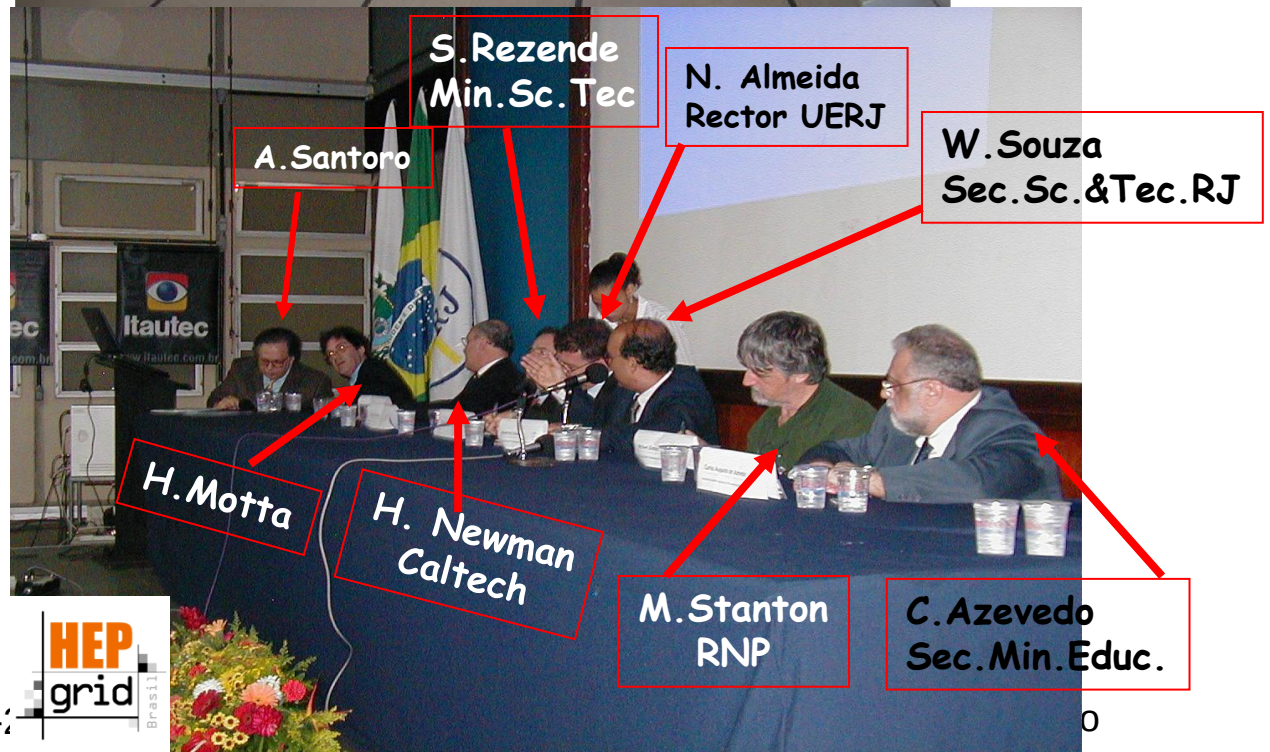
Brazil: RNP in Early 2004



Nationally: 1-30 Mbps Links;
155 Mbps to US; 45 Mbps to Fermilab



December 20, 2004
Inauguration of the HEPGrid Tier2 at UERJ In Rio



SLOW? YES! But we continue our main project:

The purpose of HEPGRID-CMS/BRAZIL is to become
- ➡ **At Regional Level, Federate with CBPF, UFRJ, UFRGS, UFBA, UERJ & UNESP**
- ➡ **At International Level, Federate with Caltech, T1-FNAL, GRID3/OSG...**
- ➡ **Strong cooperation with CALTECH**

On line systems

Brazilian HEPGRID

T0 +T1

CERN

Soon not more a Dream!

2.5 - 10 Gbps

T1

France **Germany** **Italy** **BRAZIL** **622 Mbps** **USA**

UNESP/USP SPRACE-Working

UERJ Regional Tier2 Ctr

T2 T1

T3 T2

UFRGS

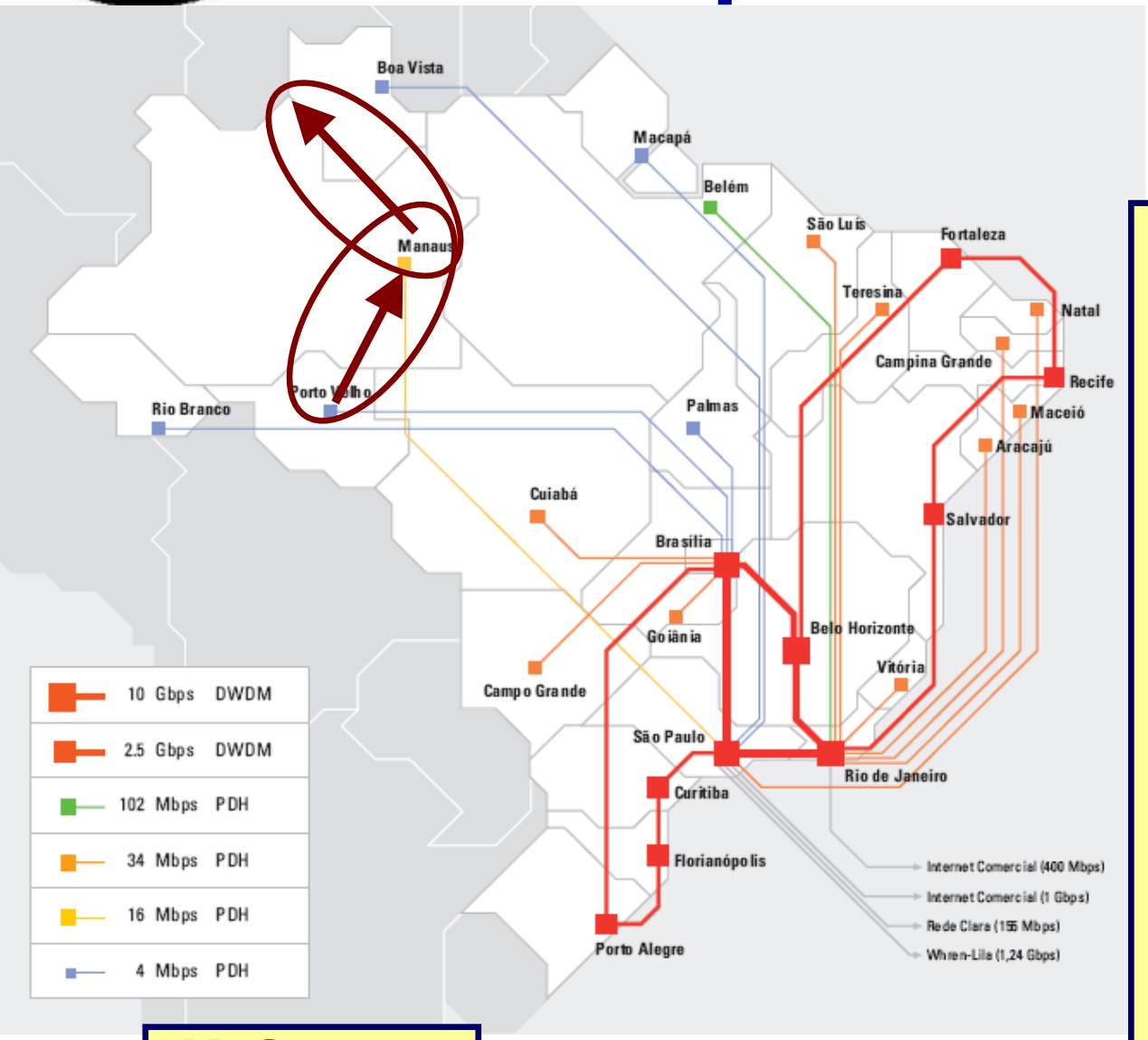
Gigabit

UERJ **UFBA** **UFRJ** **CBPF**

UERJ: T2T1, 100500 Nodes; Plus T2s to 100 Nodes

T4 Individual Machines

Brazil: RNP2 Next-Generation "Ipê" Backbone



New vs. Old
A factor of
70 to 300 in Bandwidth

2006:

- **Buildout of dark fiber nets in 27 cities with RNP PoPs underway: Manaus, Belem**
- **2.5G (to 10G) WHREN (NSF/ANSP) Link to US; 622M Link to GEANT**
- **Prospect of Upgrading Sao Paulo Link to 10G (RNP, ANSP + NSF)**
- **Dark Fiber across the Amazon to Manaus**
- **102 Mbps Rio-Belem**

M. Stanton

CHEP06: Pres. of India Collaborating with US, CERN, Slovakia via EVO

Coincident with Data Transfers of ~500 Mbps

15 TBytes to/from India in 2 Days

+ MonALISA

Mumbai-Japan-US Links

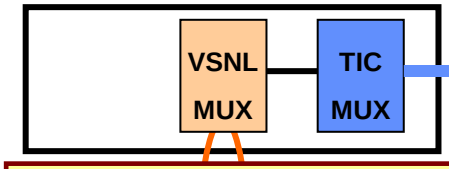
INDIA

Chennai POP VSNL
LANDING STATIONS

JAPAN

JAPAN LAND
STANDING

SINGAPORE LANDING STATION



TIC

Cable

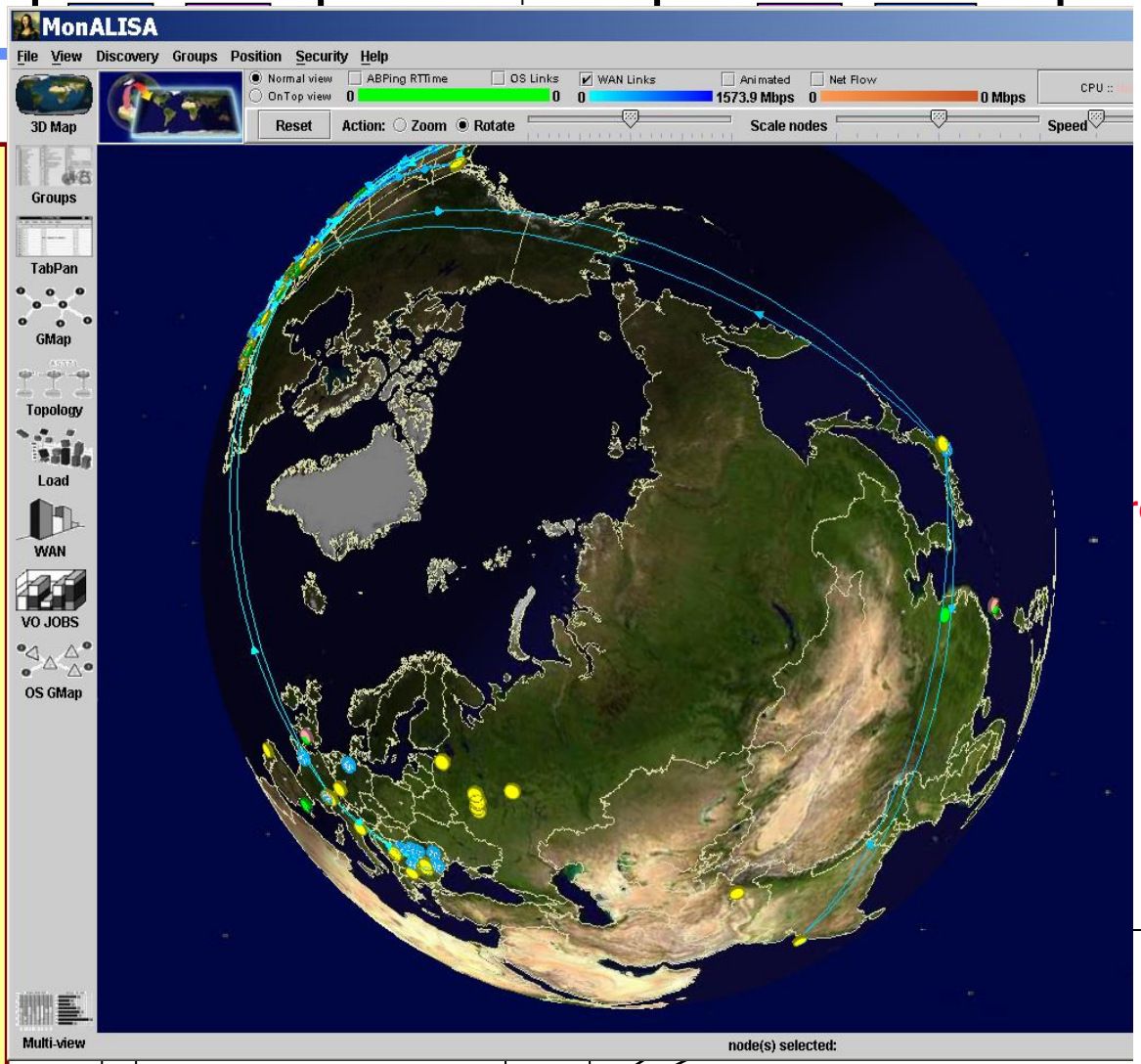
TIFR Link to Japan
+ Onward to US & Europe

Loaned Link from VSNL at CHEP06

End to End Bandwidth
4 X 155 Mbps
on SeMeWe3 Cable

Goal is to Move to
10 Gbps on SeMeWe4

Helped spark planning for Next Generation R&E "Knowledge Network" in India



INTERFACE TYPES

STM 4

INTERFACE TYPES

OC-12

TIFR Mumbai, INDIA

NTT Otemachi Bldg, JAPAN

+ Onward to US, Europe ➡

India: Knowledge Commission Recommendation to Create a National "Knowledge Network": Approved by Prime Minister January 25, 2007

- *"Build a National Knowledge Network with gigabit capabilities to connect all universities, libraries, laboratories, hospitals and agricultural institutions to share data and resources across the country." [~ $2B]*
- *5000 Institutions; 500-1000 in Phase 1*
- *Minimum connectivity at end nodes, 100 Mbps (to gigabit)*
- *Phase 1: Start with existing commercial networks*
  - *"Slide" into hybrid network with inner core owned by the stakeholders*
- *Migrate core to N X 10 Gbps, providing gigabit connectivity*

Ongoing issue: Pricing of 10 G int'l Links to the Region [also Pakistan]

Sinaia, Romania
October 13-18, 2006
http://niham.nipne.ro/events2006/

Mihai Petrovici

SLOVAK Academic Network January 2008: 1 GbE to 10 GbE Switched Ethernet

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

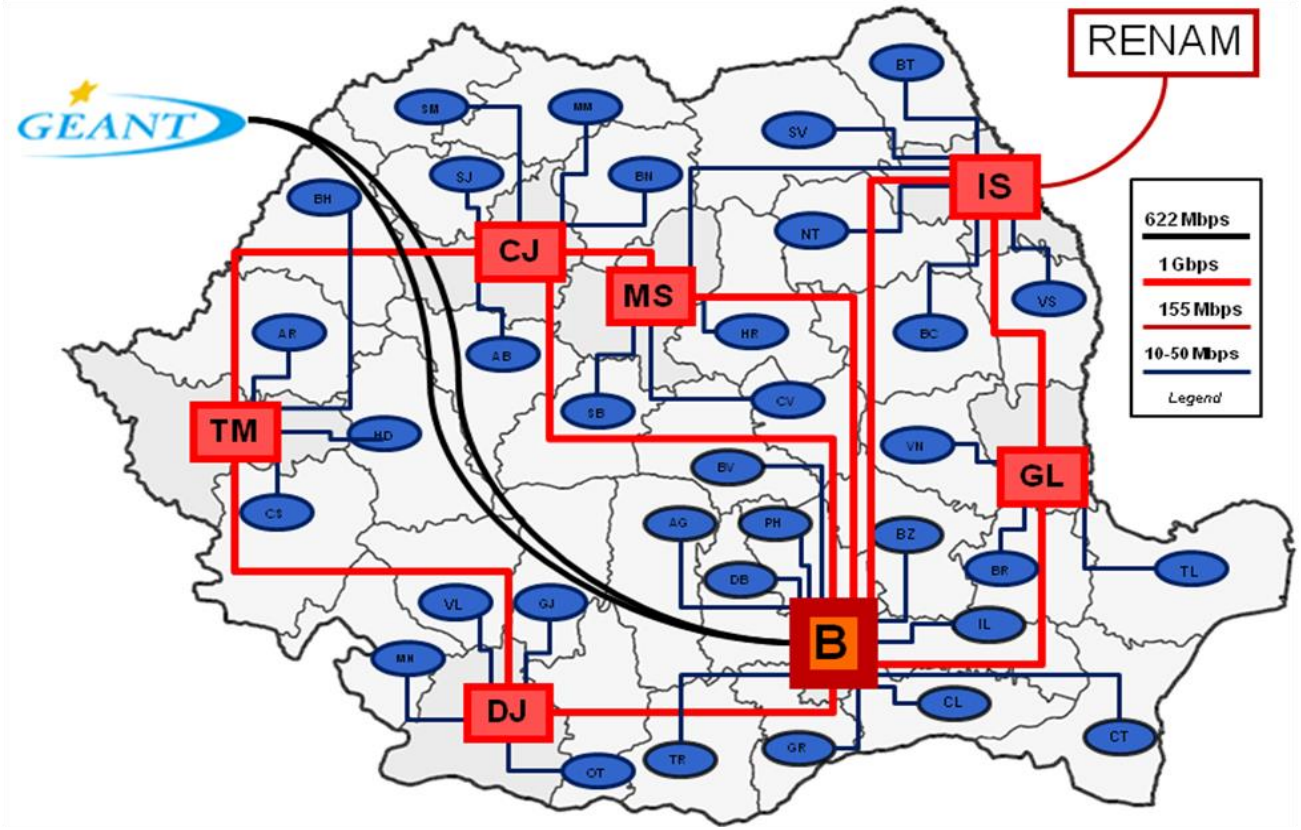
2500x Increase Since 2002

Weis Horvath

- 1660 km of Dark Fiber CWDM Links
- 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 (11/2006); 8 X 10G over 224 km with Nothing In-Line
- 2007: All 4 CBDF links to 10 Gbps; Transition backbone to 10G begun

Romania: Roedunet (2/08)

N. Tapus
O. Rusu



1G Backbone + Accesses

10G Connections to Grid Centers in Bucharest, Iasi

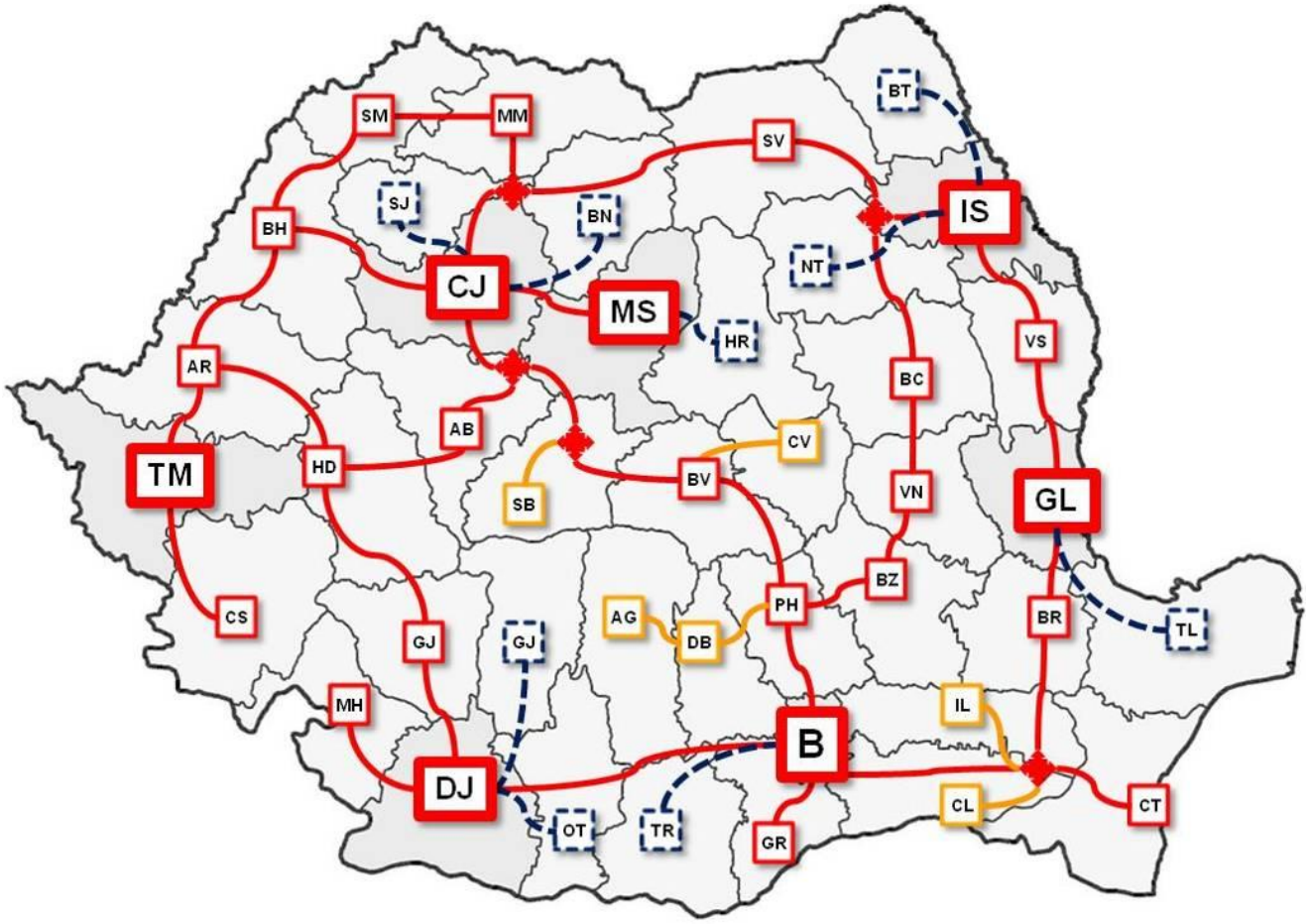
2 X 622 Mbps to GEANT2

10G Connection to GEANT due in March or April 2008

5+ Years Since 2002 SCIC Grid Workshop and WSIS Pan-European Ministerial Meeting in Bucharest

Romania: Roedunet2 Planned Dark Fiber Topology

N. Tapus
O. Rusu



97 Links Planned

DWDM (Red):
37 10 Gbps
+ 60 8x1 Gbps

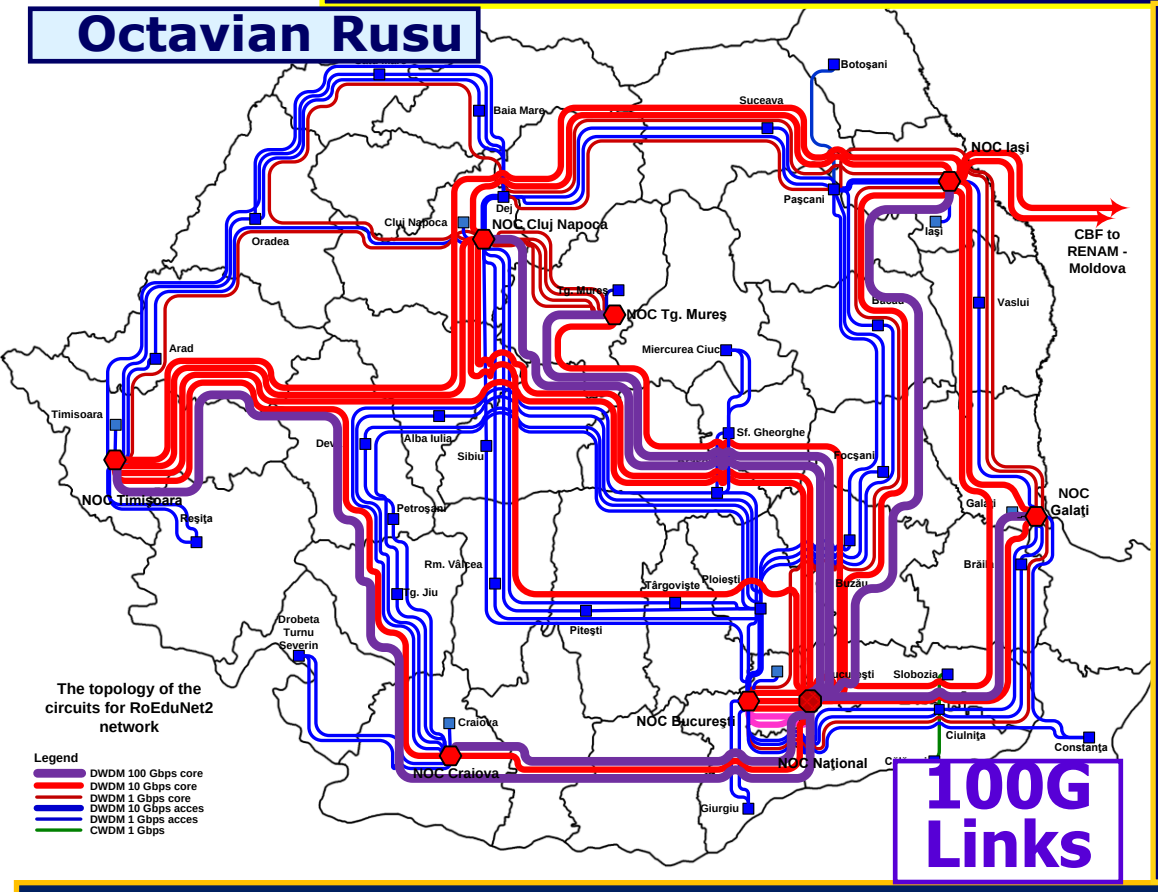
CWDM (Tan)
6 10Gbps links

70% of Network Planned to be Completed by End 2008

RoEduNet2 (ROMANIA): New 100G Core in 2013

100,000X Since 2002: Pan-European "Role of Science in the Information Society" Ministerial Meeting with HEP Bucharest

Octavian Rusu



100G Links

5636 Km Dark Fiber

41 Segments

56 Optical Sites

18 Dynamically Configurable

35 10GE + 54 10 X 1GE Links

2012: 1st 100GE Link

Bucharest – Iasi

2013: Deploying

100G Core Infrastructure

2006 – RoEduNet2 project approved

2008 – August – GEANT POP installed in Bucharest: 10 Gbps to GEANT, 2.5 Gbps committed

2010 – 1st CBF from Romania installed: Iasi – Chisinau (Moldava) DWDM segment operational

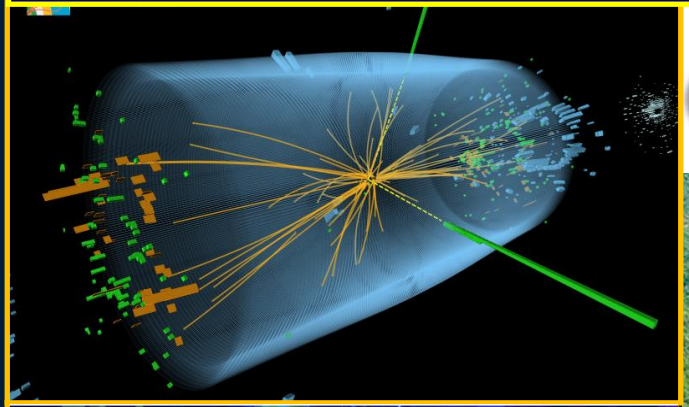
2011 – Total fiber length 5636 km; Interconnects 40 of the 41 county capitals

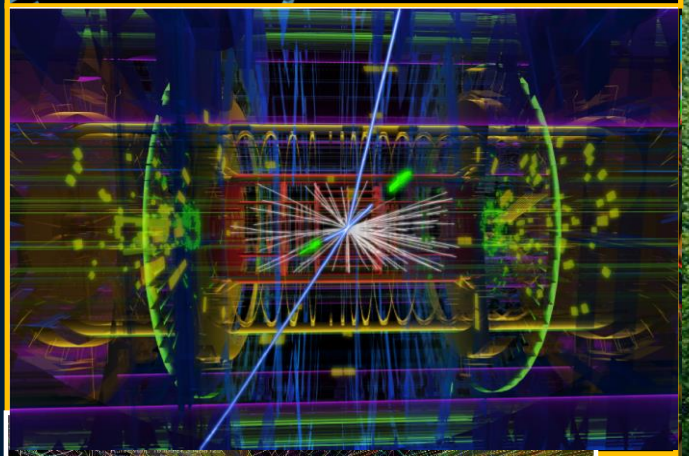
2012 – 1st 100 Gbps Ethernet link installed from Bucharest to Iasi;

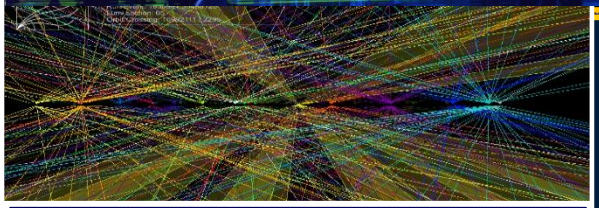
2013 – Upgrade to 6 100G Waves Bucharest to Iasi; wider 100G deployments

Crossing the Digital Divide
in Brazil and the Amazon Region

ICFA SCIC:http://icfa-scic.web.cern.ch/







50 Vertices, 14 Jets, 2 TeV

LISHEP 2015

Harvey B Newman

LISHEP 2015 Panel

August 5, 2015

Embarking on a
River of Discovery

SCIC in 2002-15

http://cern.ch/icfa-scic

2015 Reports: A Worldview of *Networks for and from HEP*

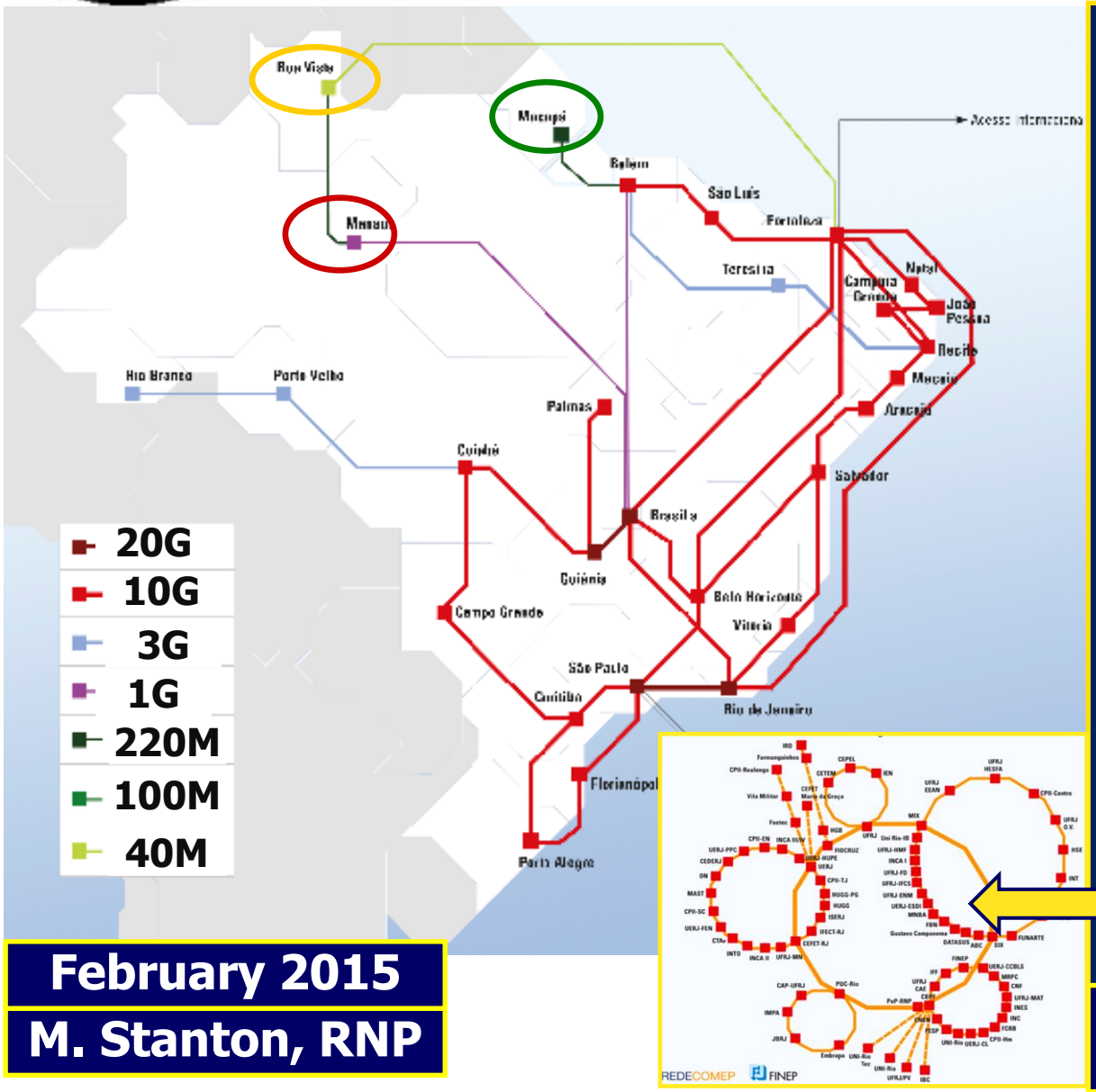
Focus on the LHC Program during Run2 and Beyond

- ***Executive Summary and Main Report*: "Networking for HEP"** [HN, A. Mughal et al.]: Updates on the Digital Divide, World Network Status, Transformative Trends in the Global Internet
- ***31* New Annexes + A World Network Overview**
  Status and Plans of International, Nat'l & Regional Networks, HEP Labs, and Advanced Network Projects
- ***Monitoring Working Group Report*** [R. Cottrell, S. McKee]: Quantifying the Digital Divide: PingER Data from a worldwide set of monitors
- ***LHCONE (www.lhcone.net):*** A Global Architecture of Open Exchange Points and a Routing "Fabric" supporting the LHC Computing Models: ***Focus on Tier1/2/3 Operations; Building on a Successful 1st Phase***

Also See:

- TERENA *2015 Compendium* (www.terena.org): R&E Networks in Europe
- http://internetlivestats.com: Worldwide Internet Use
- Telegeography.com; Interactive Submarine Cable Map: http://submarinecablemap.com

Brazil in 2015: 6th Phase "Ipê" 10G Core Network



- 4000 km 10G Footprint (East+South) Completed
- New 2nd fiber across Amazon reaches the northern capitals Macapa and Manaus
  - RNP expects to get multi-Gbps to these cities soon
- Completion of the optical fiber footprint: Manaus – Boa Vista
- Brasilia – Manaus Link to 1G, via existing terrestrial link
- Metro R&E dark fiber nets in 26 of 27 state capitals in operation 19 at 2X10G; 2 at 10G+3G; only Porto Velho left to be built.
- Completion of long-awaited 10G 300km long metro ring in Rio used by 60 campi, including the HEPGrid Tier2

Impact of First ICFA Digital Divide Wkshp in Rio in 2004

February 2015

M. Stanton, RNP

Optical Fiber Through the Amazon: Porto Velho-Manaus

935 Km Optical Cables
8,000 Utility Poles
7 Optical Repeaters (622 Mbps)
60m Towers
Microwave across the Amazon

Aerial Crossing of the Amazon at Jurupari: 2100m span between 300m towers

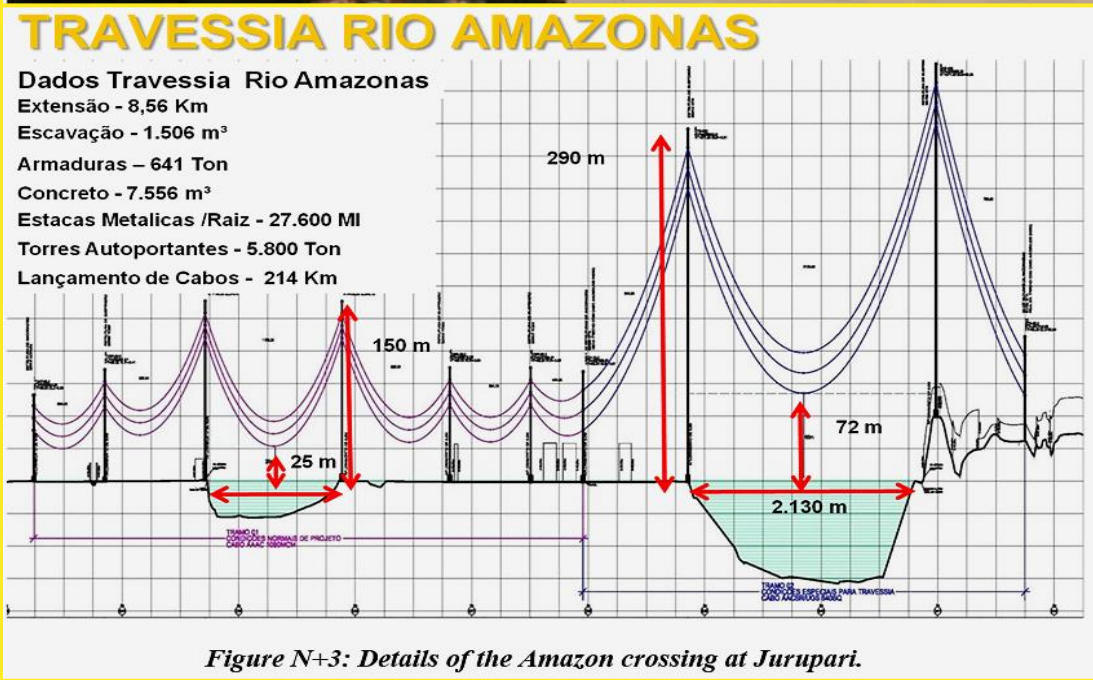
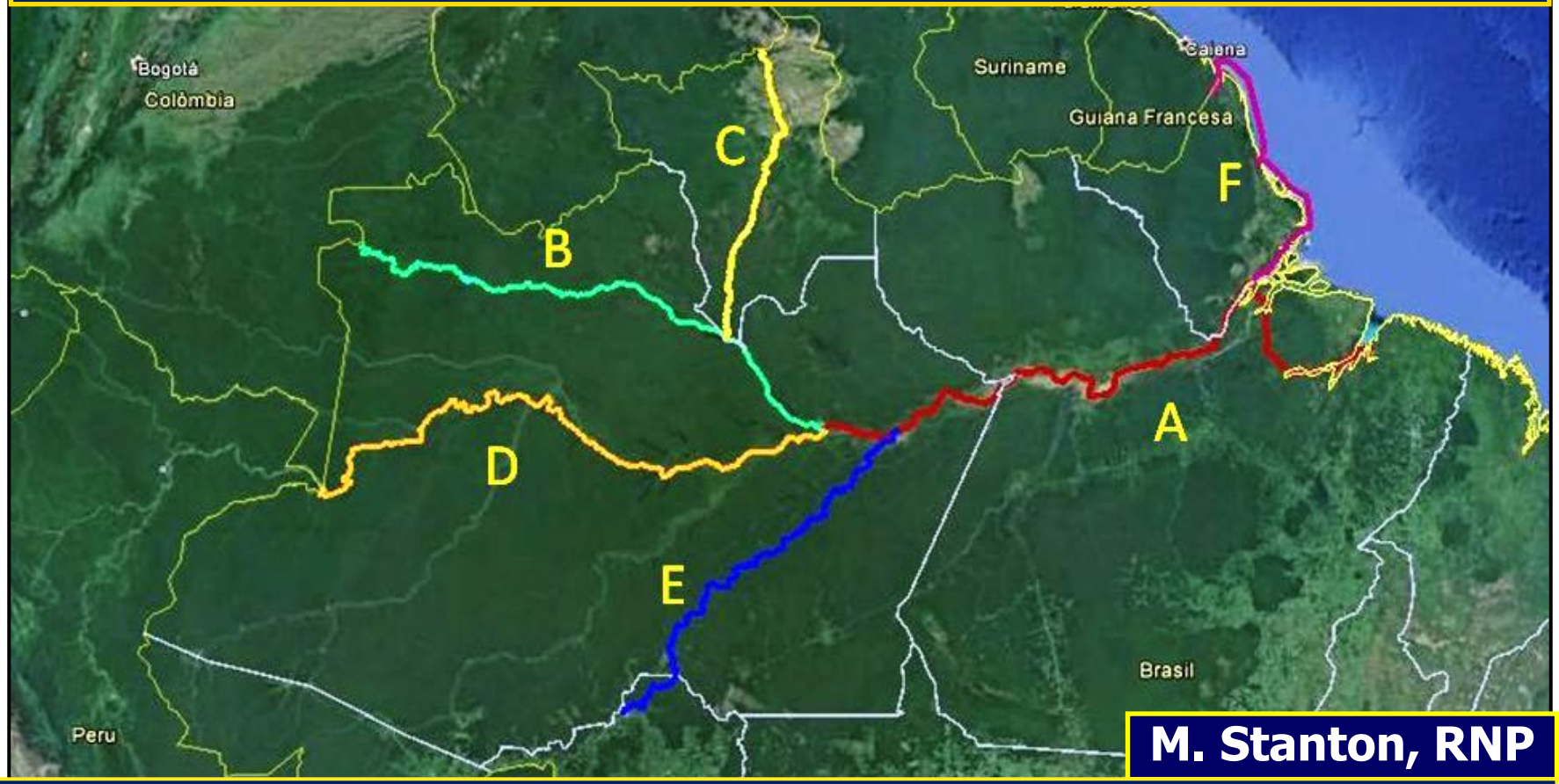


Figure N+3: Details of the Amazon crossing at Jurupari.

- 2nd fiber across the Amazon reaches the northern capital cities **Macapa and Manaus**
  - **Brings competition** to the 1st terrestrial link to Manaus
  - **RNP expects to get multi-Gbps access to these cities soon**

Brazil: RNP proposal for cables along major rivers in the north

- Complementing existing fiber infrastructure
- Pilot along Route D may be feasible in 2015



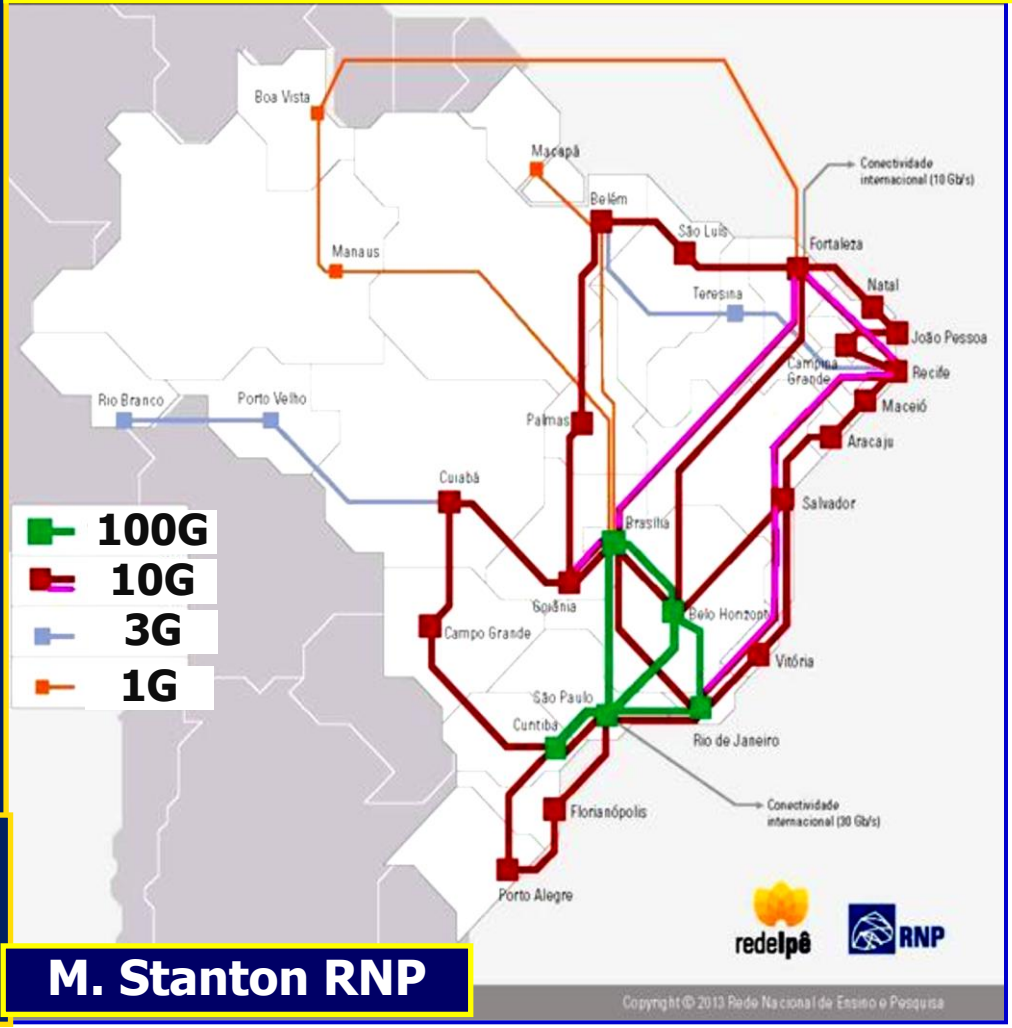
Possible major routes for subfluvial fiber optic cables. Rivers: A: Amazon; B: Negro; C: Branco; D: Solimões (upper Amazon), E: Madeira; F: Maritime route to French Guiana.

Brazil: Future Internet, SDN and 100G in the Backbone

- Layer 2 Circuit technology is used to construct large scale testbeds for R&D in Future Internet architectures
- SDN (OpenFlow) has been used to automate some L2 dynamic circuit operations in the AmLight infrastructure
- RNP intends to deploy several 100G links at the core of its network already in 2015
- ANSP and RNP will deploy SDN on their networks

- This will serve as a pilot for its SDN-based future network with appropriately abundant capacity

Concept for a 100G Core in the RNP Network

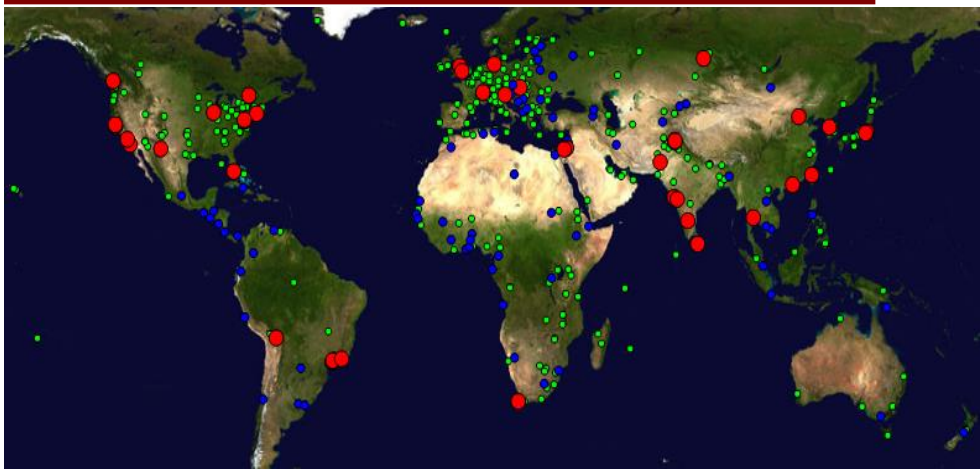


M. Stanton RNP

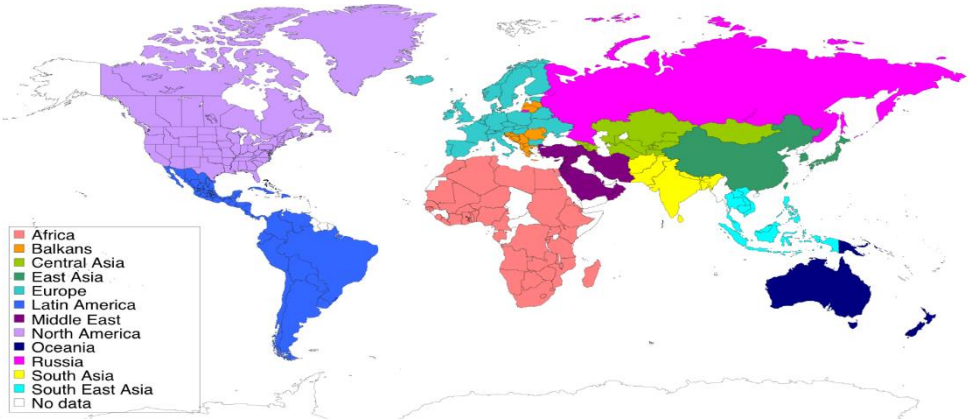
SCIC Monitoring WG PingER (Also IEPM-BW)

R. Cottrell

Monitoring & Remote Nodes (2/08)



PingER Regions

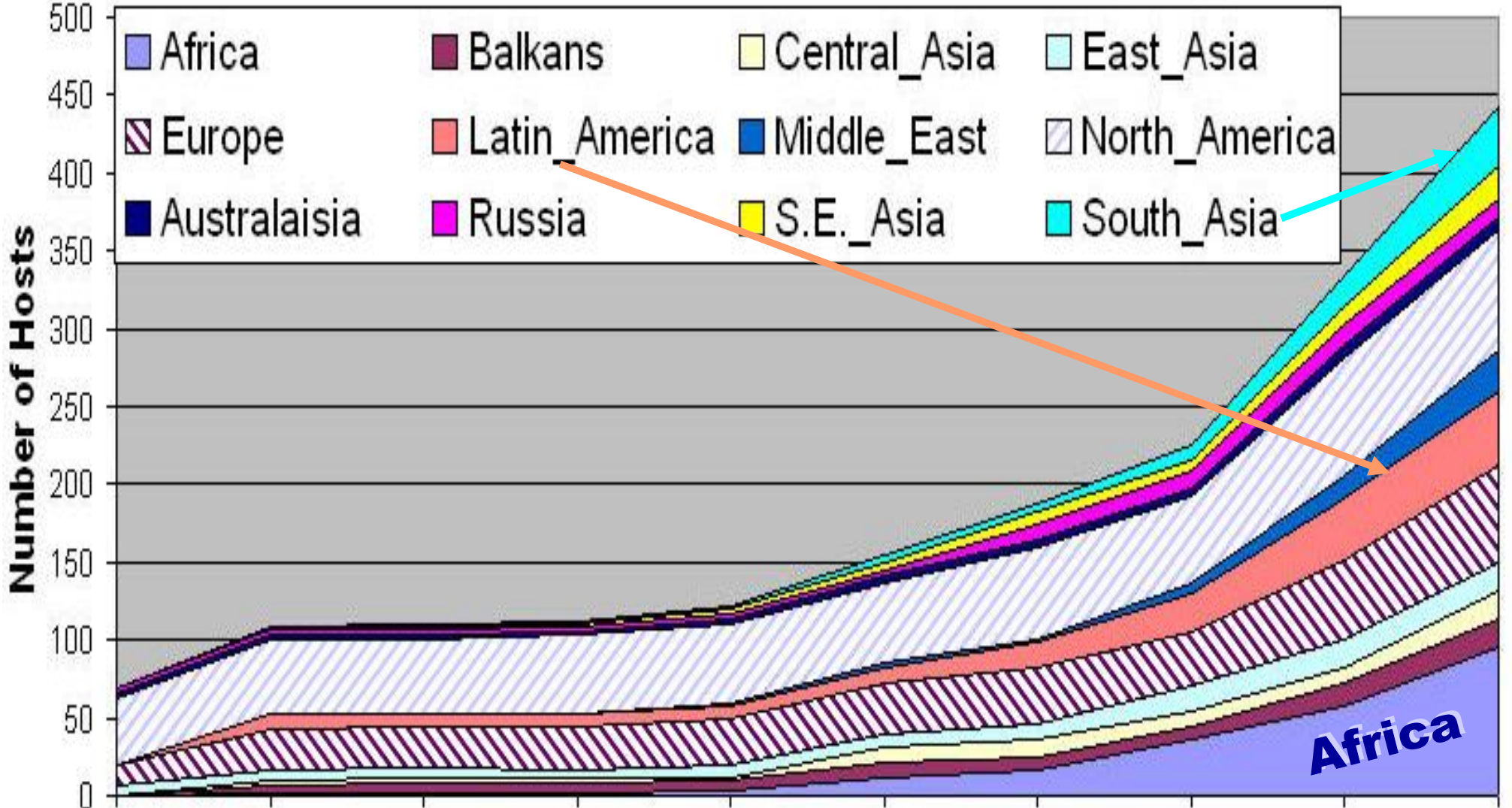


- **Measurements from 1995 On**
  - *Reports link reliability & quality*
- **Countries monitored**
  - Contain 95.5% of world pop.
  - 99% of world's Internet users
- 800+ remote nodes at 675 sites in 155 nations; 40 monitoring nodes 87 Sites in 45 African countries
- Strong Collaboration with ICTP Trieste and NIIT (Pakistan)
- Excellent, Vital Work; funding issue

Countries: N. America (2), Latin America (20), Europe (40), Africa (45), Middle East (12), Central Asia (9), South Asia (7), East Asia (4), Sotheast Asia (9), Russia (1), China (1) and Oceania (4)

Number of Hosts Monitored By Region: 1998 - 2007

R. Cottrell



Throughput Trendlines from SLAC 1998 - 2013

IEPM Internet End-to-end Performance Monitoring

SLAC NATIONAL ACCELERATOR LABORATORY

Top 4

Europe, N. America, East Asia & Australasia

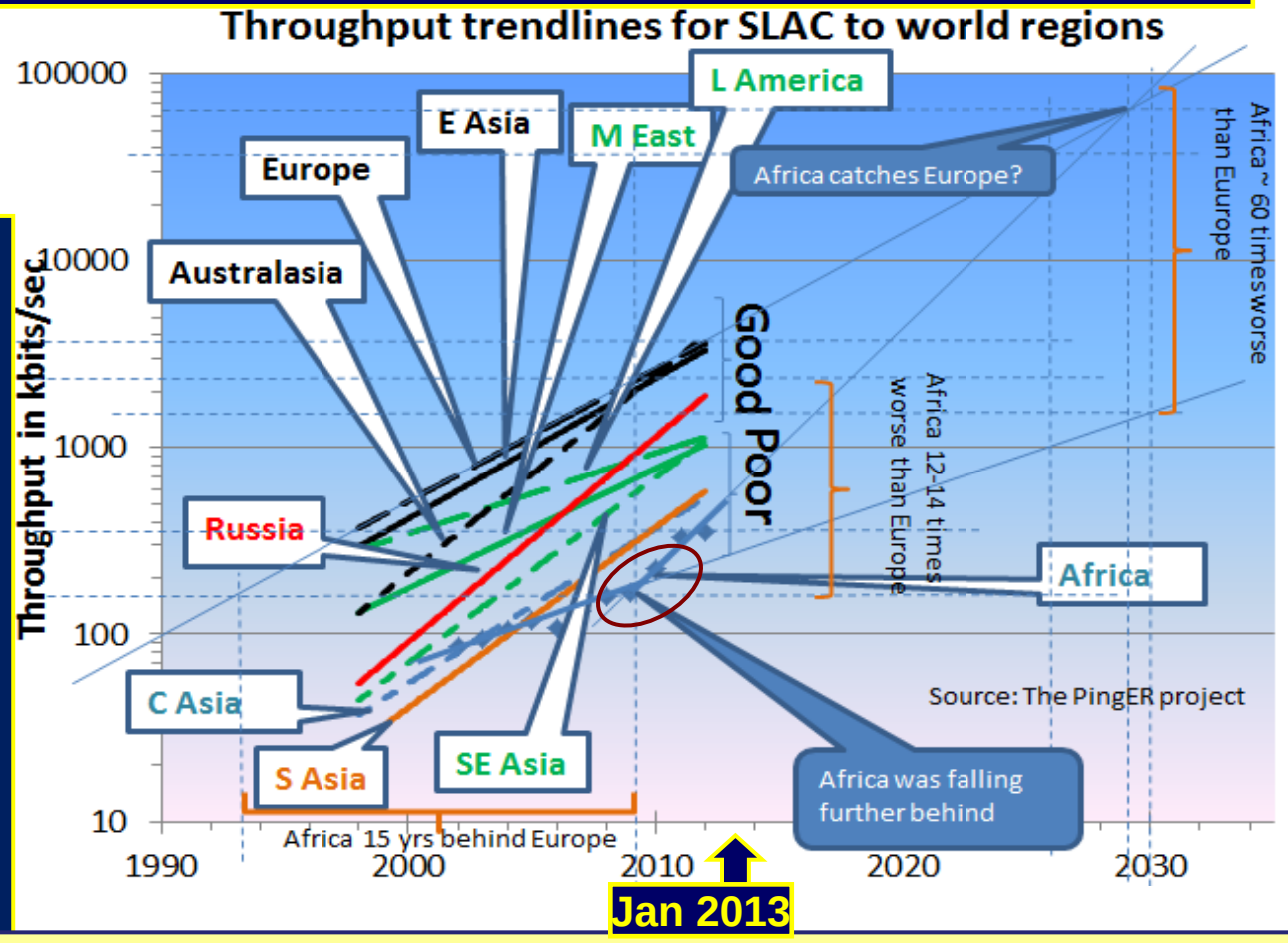
Behind Europe

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.

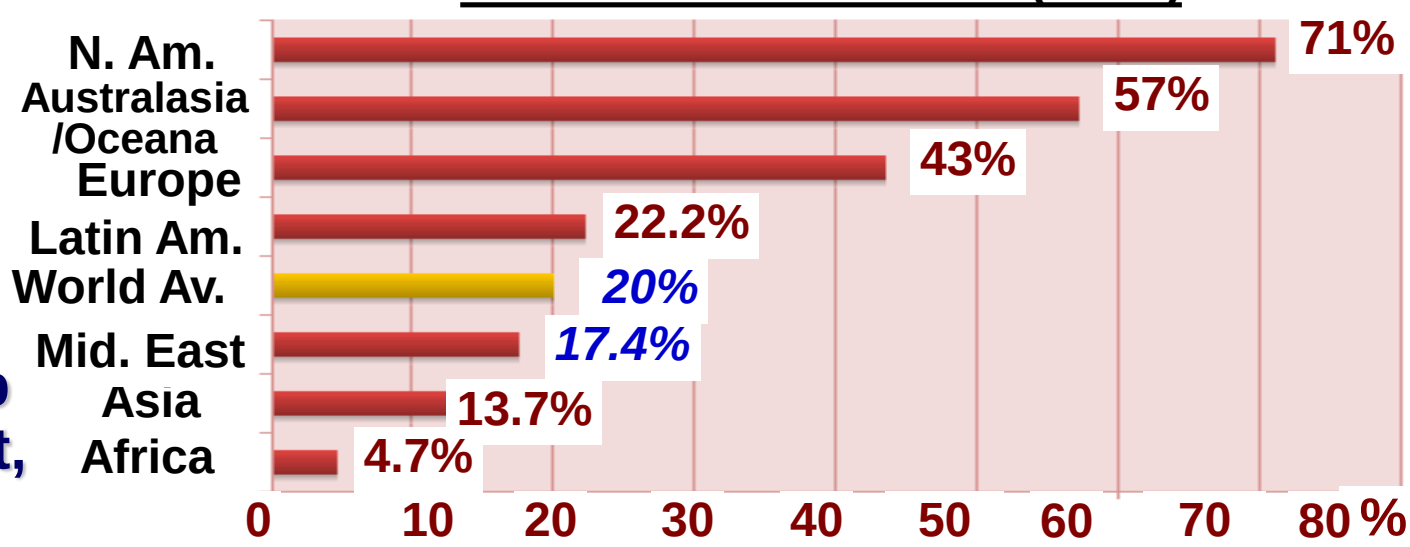
In 10 years: Russia & Latin America should catch up. Africa was falling farther behind; ***new subsea cables making a difference since 2011***

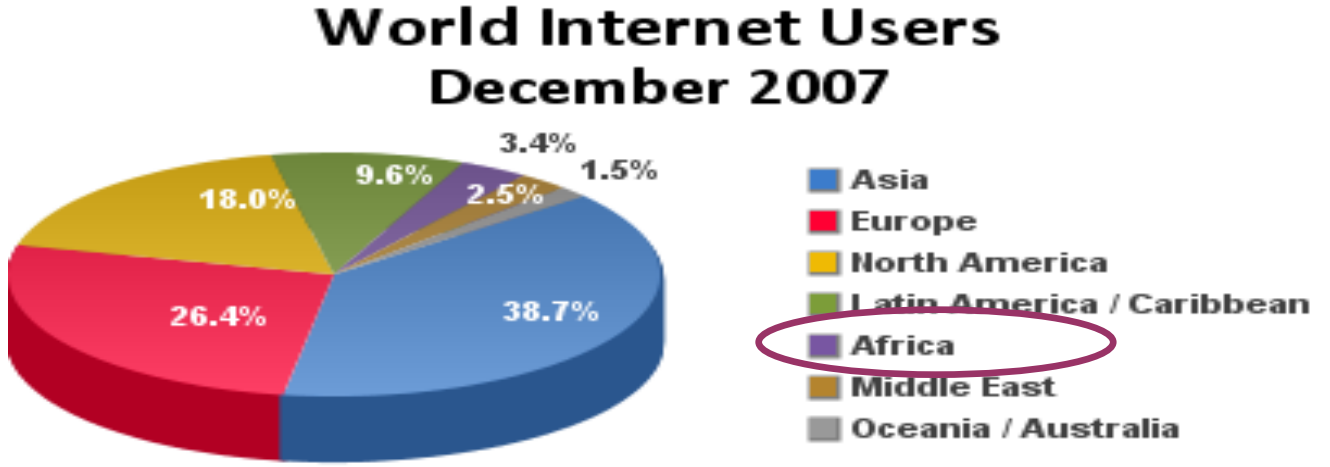
1st Revolution: "Long Dawn" of the Information Age

1.32B Internet Users; 300+M with Broadband (12/07)

http://internetworldstats.com

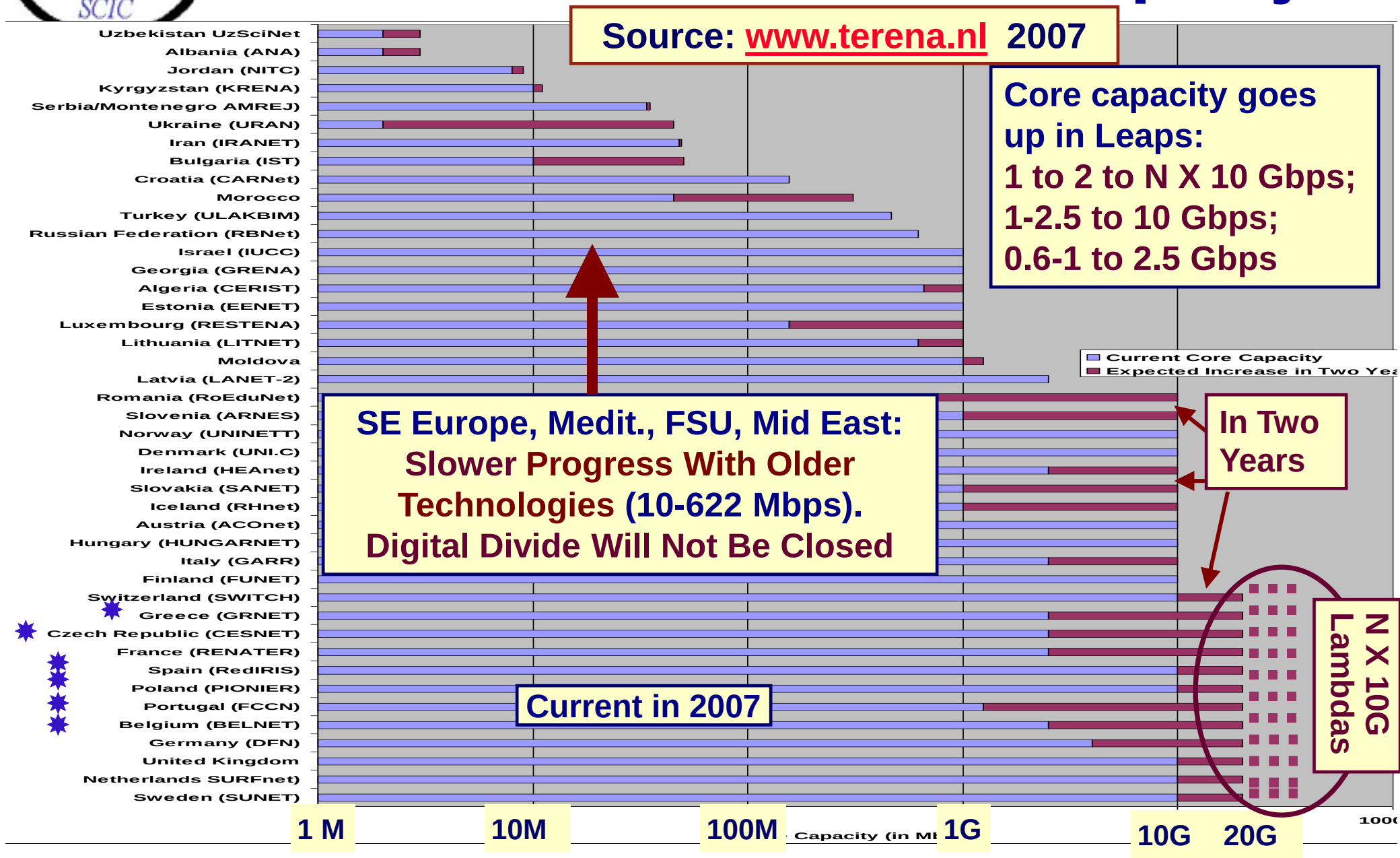
- Explosion of bandwidth use: ~1 TByte/sec
- Raw capacity still largely unused
- Emergence of Web 2.0: Billions of Web Pages, rich content, embedded apps.
  - Wikipedia
  - MySpace
  - Second Life
- Signs of Web 3.0: Rich, persistent streaming content – ubiquitous information





Broadband: 65-70M Each in the US and China

Digital Divide Illustrated by Network Infrastructures: TERENA Core Capacity



Global Network Advancement Group (GNA-G) Leadership Team: Since September 2019

leadershipteam@lists.gna-g.net

Buseung Cho
KISTI (Korea)

Marco Teixera
RedCLARA
(Latin America)

Ivana Golub
PSNC, GEANT
(Europe)

Harvey Newman
Caltech (US)

David Wilde,
Chair
Aarnet (Australia)

Alex Moura
KAUST
(Saudi Arabia)

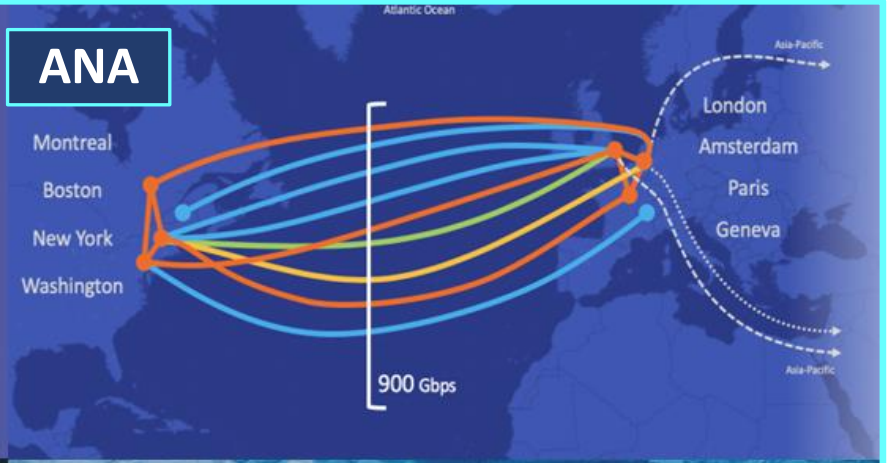
- **An open volunteer group devoted to developing the blueprint** to make using the Global R&E networks both simpler and more effective
- **Its primary mission is to support global research and education** using the technology, infrastructures and investments of its participants.
- **The GNA-G is a data intensive research & science engager** that facilitates and accelerates global-scale projects by **(1) enabling high-performance data transfer,** and **(2) acting as a partner in the development of next generation intelligent network systems** that support the workflow of data intensive programs

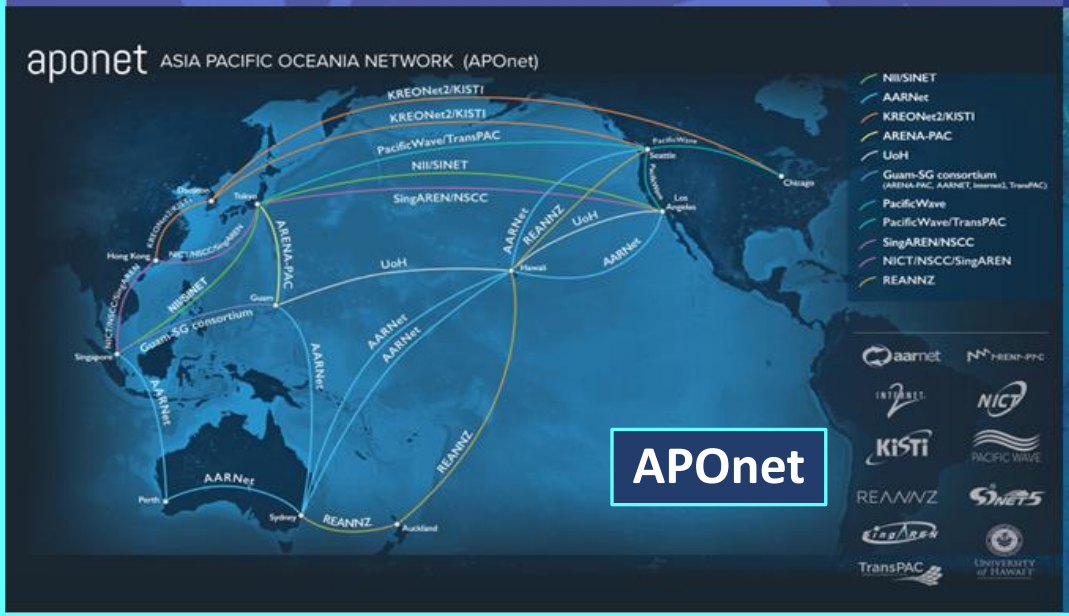
See https://www.dropbox.com/s/qsh2vn00f6n247a/GNA-G%20Meeting%20slides%20-%20TechEX19%20v0.8.pptx?dl=0

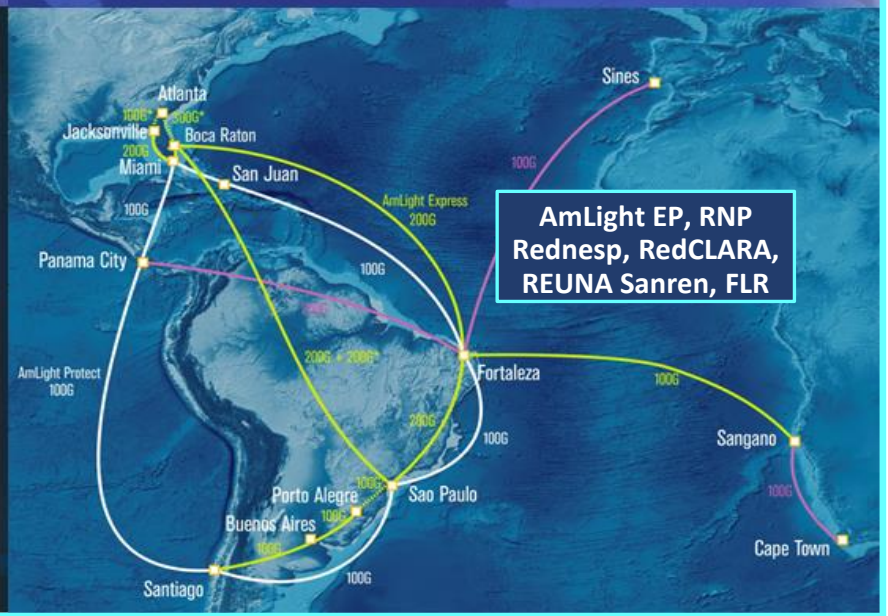
GNA-G
Global Network Advancement Group

https://www.gna-g.net/

Mission: Support global research and education using the technology, infrastructures and investments of its participants

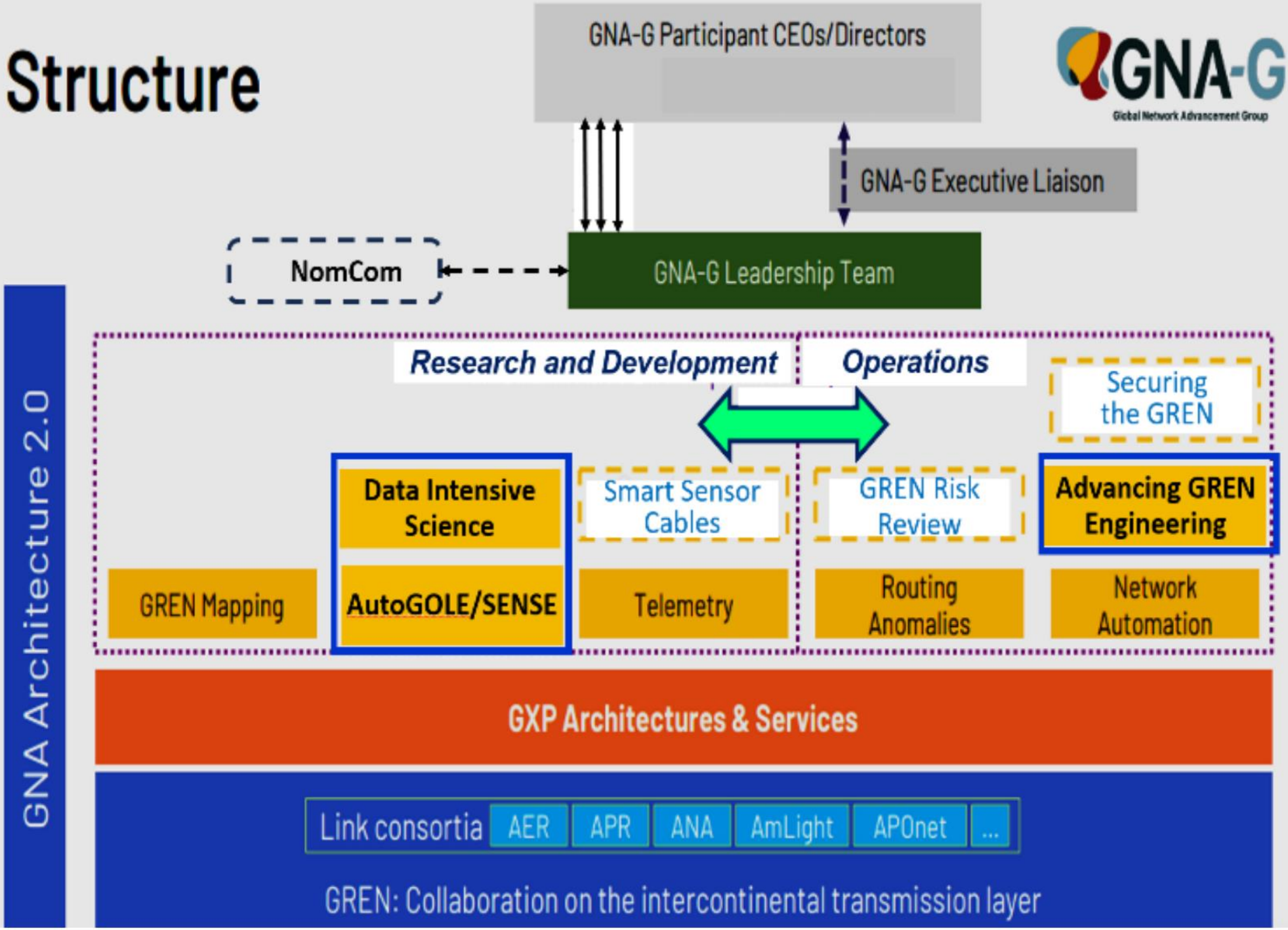






The GNA-G exists to bring together researchers, National Research and Education Networks (NRENs), Global eXchange Point (GXP) operators, regionals and other R&E providers, in developing a common global infrastructure to support the needs

Structure


GNA-G Participant CEOs/Directors
GNA-G
Global Network Advancement Group
GNA-G Executive Liaison
NomCom
GNA-G Leadership Team
GNA Architecture 2.0
Research and Development
Operations
Securing the GREN
Data Intensive Science
Smart Sensor Cables
GREN Risk Review
Advancing GREN Engineering
GREN Mapping
AutoGOLE/SENSE
Telemetry
Routing Anomalies
Network Automation
GXP Architectures & Services
Link consortia
AER
APR
ANA
AmLight
APOnet
...
GREN: Collaboration on the intercontinental transmission layer

AutoGOLE / SENSE Working Group

- **Worldwide collaboration of open exchange points and R&E networks interconnected to deliver network services end-to-end in a fully automated way.** NSI for network connections, SENSE for integration of End Systems and Domain Science Workflow facing APIs.
- **Key Objective:**
  - The AutoGOLE Infrastructure should be persistent and reliable, to allow most of the time to be spent on experiments and research.
- **Key Work areas:**
  - Control Plane Monitoring: Prometheus based, Deployments underway
  - Data Plane Verification and Troubleshooting Service: Study and design group formed
  - AutoGOLE related software: Ongoing enhancements to facilitate deployment and maintenance (Kubernetes, Docker based systems)
  - Experiment, Research, Multiple Activities, Use Case support: Including Fabric, NOTED, Gradient Graph, P4 Topologies, Named Data Networking (NDN), Data Transfer Systems integration & testing.
- **WG information**
  https://www.gna-g.net/join-working-group/autogole-sense

The GNA-G Data Intensive Sciences WG

Charter: https://www.dropbox.com/s/4my5mjl8xd8a3y9/GNA-G_DataIntensiveSciencesWGCharter.docx?dl=0

- Principal aims of the GNA-G DIS WG:

(1) To meet the needs and address the challenges faced by major data intensive science programs

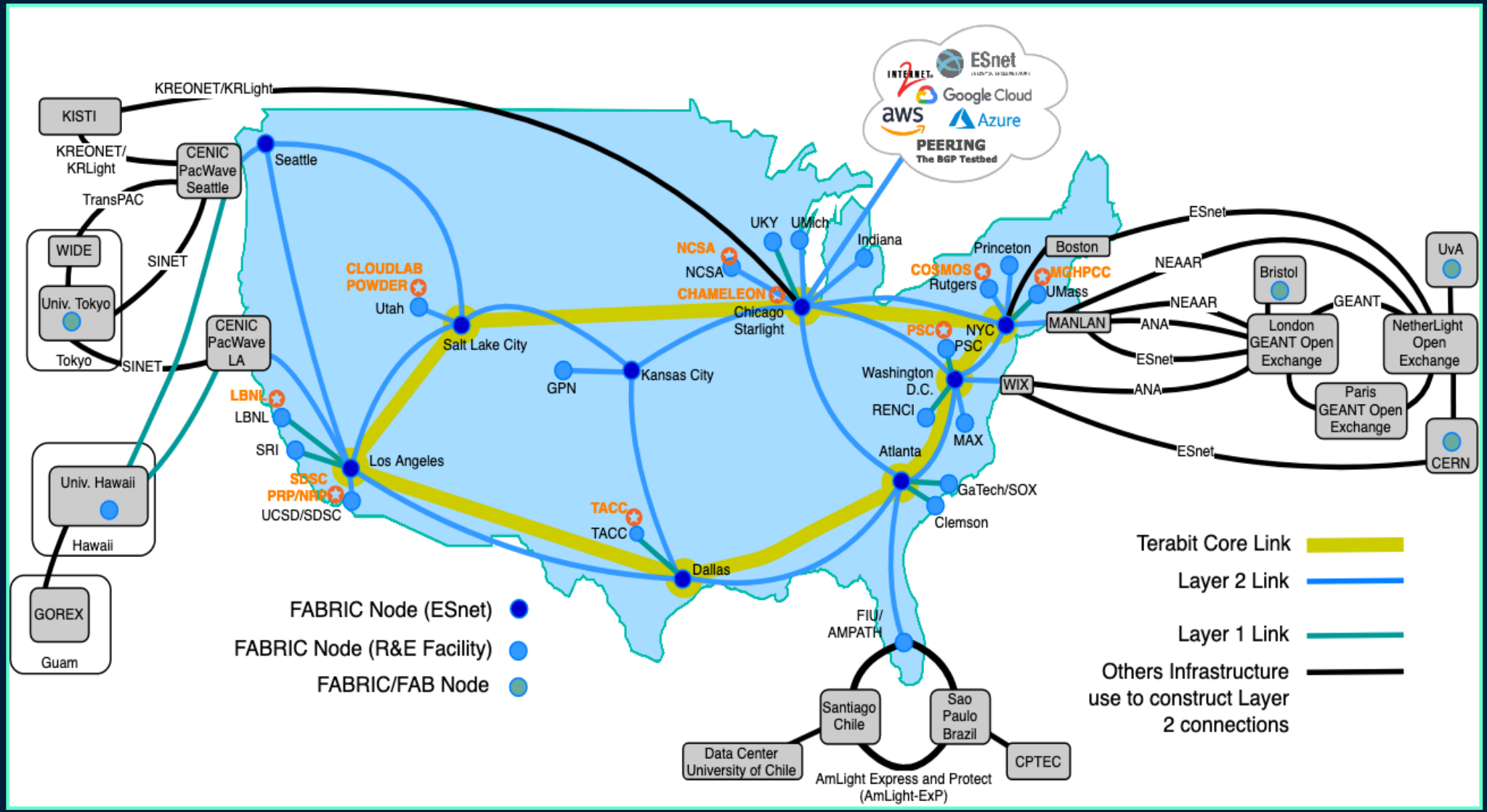
- In a manner consistent and compatible with support for the needs of individuals and smaller groups in the at large A&R communities

(2) To provide a forum for discussion, a framework and shared tools for short and longer term developments meeting the program and group needs

- To develop a persistent global testbed as a platform, to foster ongoing developments among the science and network communities

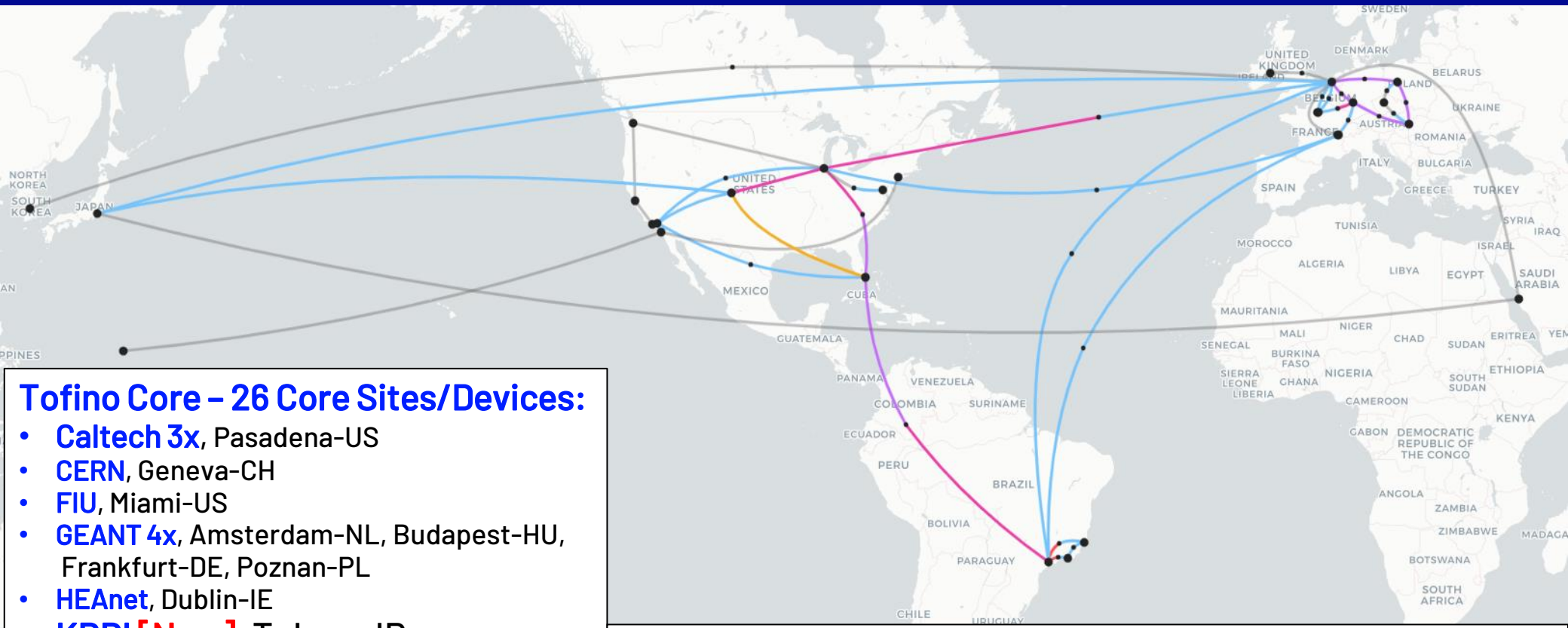
- While sharing and advancing the (new) concepts, tools & systems needed
- Members of the WG partner in joint deployments and/or developments of generally useful tools and systems that help operate and manage R&E networks with limited resources across national and regional boundaries
- A special focus of the group is to address the growing demand for
  - Network-integrated workflows
  - Comprehensive cross-institution data management
  - Automation, and
  - Federated infrastructures encompassing networking, compute, and storage
- Working Closely with the AutoGOLE/SENSE WG

FABRIC and FAB: Terabit/sec R&D Backbone Across the US. Transoceanic Links and Intercontinental Partnerships



US, Europe, Asia Pacific and Latin America

Global P4 Lab (GP4L)



Tofino Core – 26 Core Sites/Devices:

- **Caltech 3x**, Pasadena-US
- **CERN**, Geneva-CH
- **FIU**, Miami-US
- **GEANT 4x**, Amsterdam-NL, Budapest-HU, Frankfurt-DE, Poznan-PL
- **HEAnet**, Dublin-IE
- **KDDI [New]**, Tokyo-JP
- **KISTI**, Daejeon-KR
- **RENATER**, Paris-FR
- **RNP**, Rio de Janeiro-BR
- **SC23 [New]**, Denver-US
- **SouthernLight**, São Paulo-BR
- **StarLight**, Chicago-US
- **SWITCH 6x [New]**, Geneva-CH
- **Tennessee Tech**, Cookeville-US
- **UFES**, Vitória-BR
- **UMd/MAX**, College Park-US

BlueField-2/DPDK Islands – 7 Sites/Devices [New]:

- **Pacific Wave/UCSD**, Chicago-US, GUAM-GU, Los Angeles-US, New York-US, San Diego-US, Seattle-US, Sunnyvale-US

x86/DPDK Islands – 4 Sites/Devices:

- **FABRIC [New]**, Miami-US
- **2x GEANT**, Paris-FR, Prague-CZ
- **KAUST [New]**, Saudi Arabia-SA

Major Advances in Data Transfer Applications
Led by HEP with Computer Scientists and Network Engineers

- 2000-2007: HEP with computer scientists and network engineers developed the knowledge to use long distance networks efficiently, at high occupancy, for the first time
  - "Demystification" of large long range data flows with TCP:
    - 2004-2005: Up to 10 Gbps per flow;
    - One to a few server-pairs matches a 10 Gbps link
    - Aggregate from 23 Gbps (SC03) to 151 Gbps (SC05)
- Major advances in the TCP stack (e.g. FAST TCP), Linux Kernel (2.6.19-21), end system architecture, network interfaces and drivers
- 2006-2007: Moved to mature storage-to-storage transfer applications; working on transfers among *storage-systems*

Internet2 Land Speed Records 2002-4

Nov. 2004 Record Network

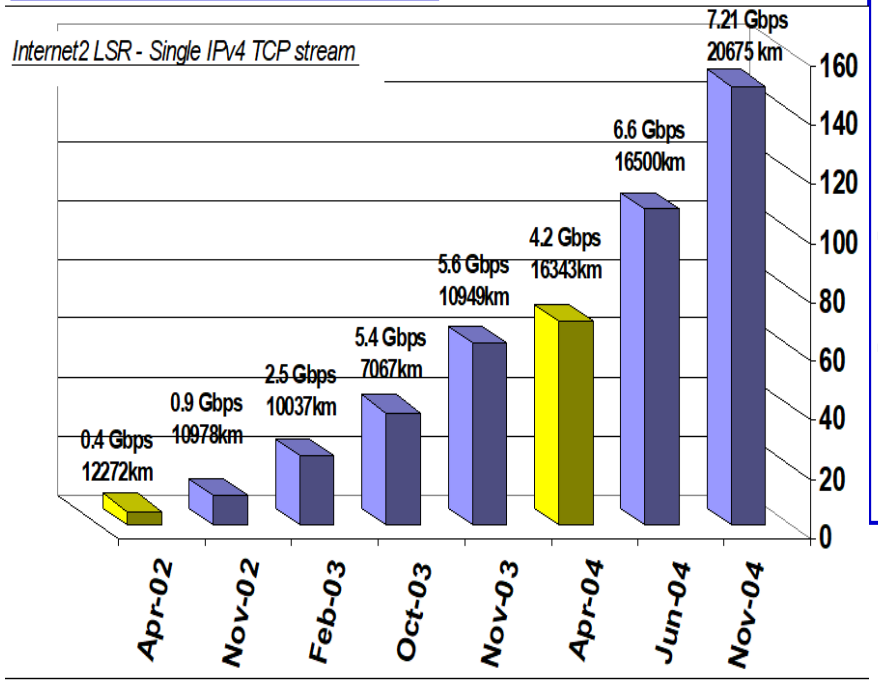
- IPv4 Multi-stream record with FAST TCP: 6.86 Gbps X 27kkm: Nov 2004
- IPv6 record: 5.11 Gbps between Geneva and Starlight: Jan. 2005
- *Disk-to-disk Marks:*
  536 Mbytes/sec (Windows);
  500 Mbytes/sec (Linux)
- *End System Issues: PCI-X Bus, Linux Kernel, NIC Drivers, CPU*

NB: Manufacturers' Roadmaps for 2006: One Server Pair to One 10G Link

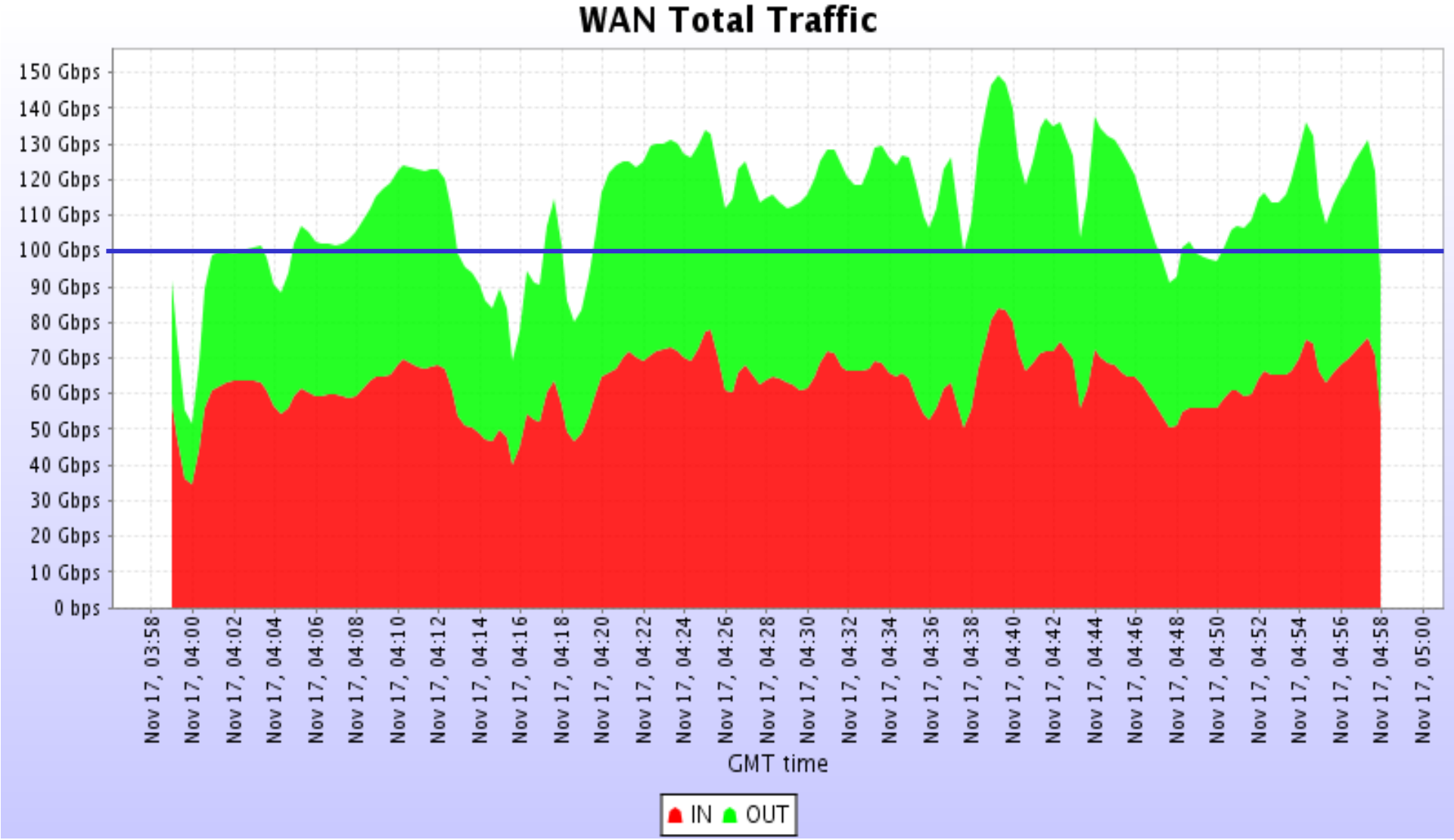
[Compare to 2024: to ~800G per server with Caltech's FDT]

Internet2 LSRs: Blue = HEP

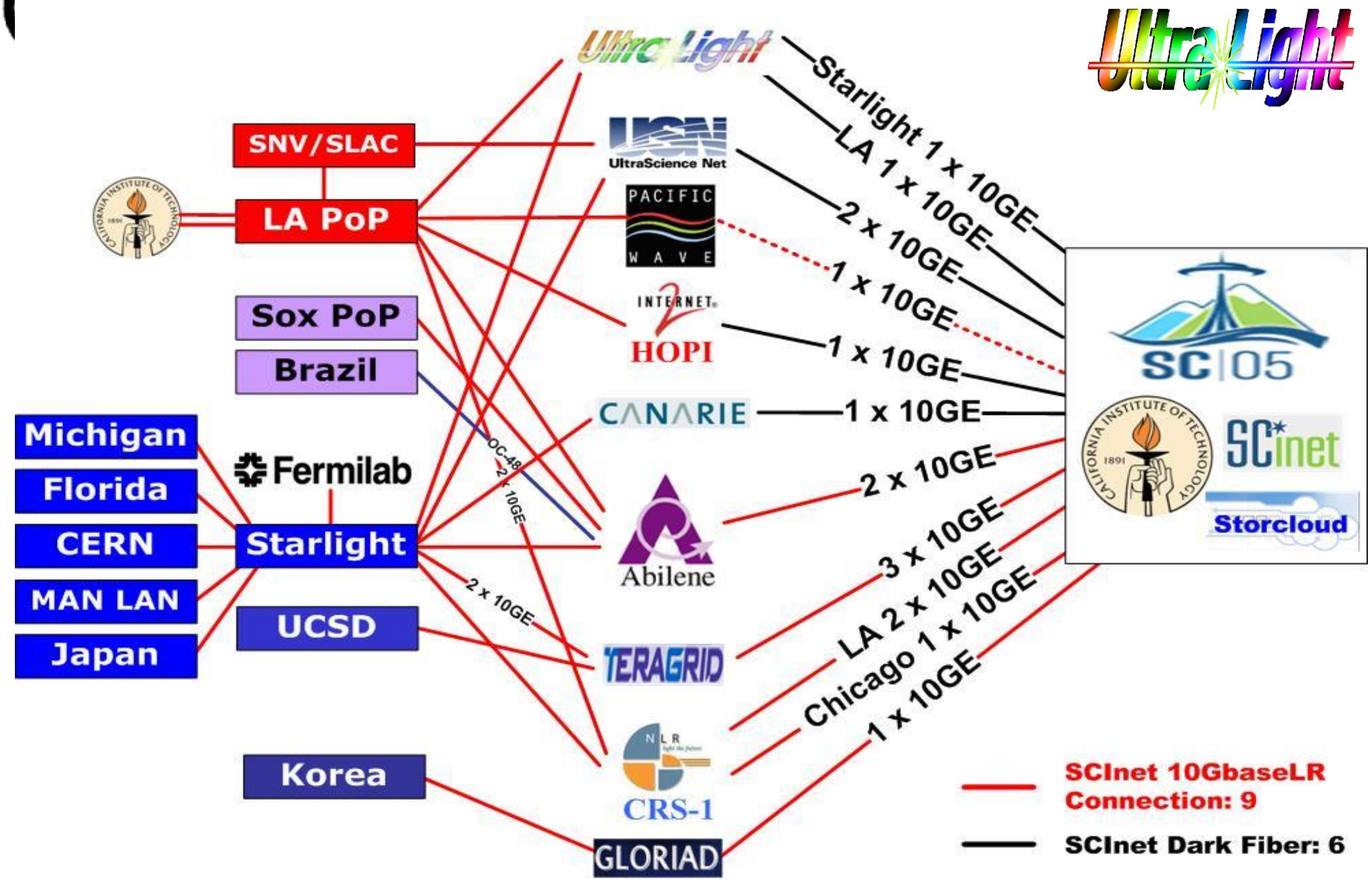
7.2G X 20.7 kkm



151 Gbps Peak; > 100 Gbps for Hours

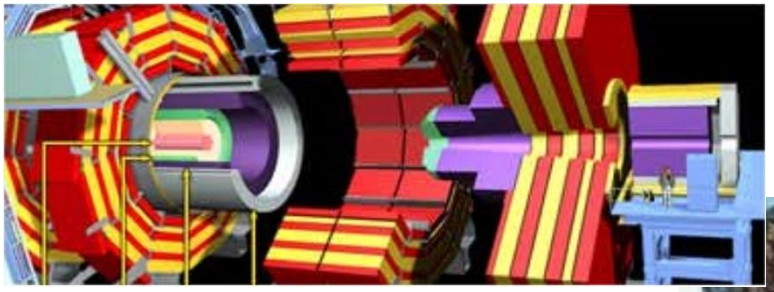


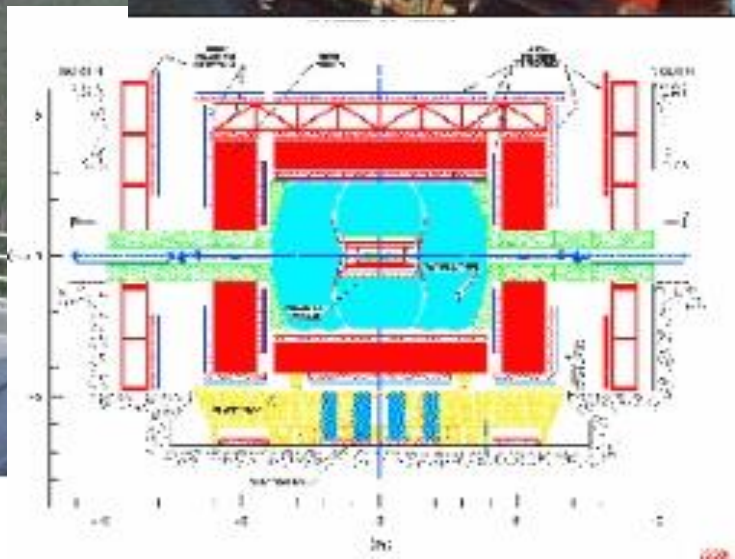
SC2005 BWC Data Flows to Caltech Booth



Fermilab

- BWC data sources were the *Production Storage Systems and File Servers* used by:
  - CDF
  - DØ
  - US CMS Tier 1
  - Sloan Digital Sky Survey
- ~600 gridftp servers (of 1000s) were directly involved
- Each of these produces, stores and moves Multi-TB to PB-scale data: Tens of TB per day



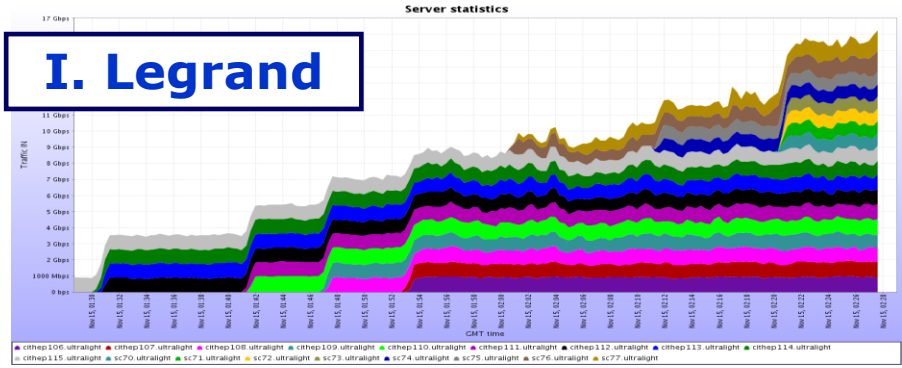


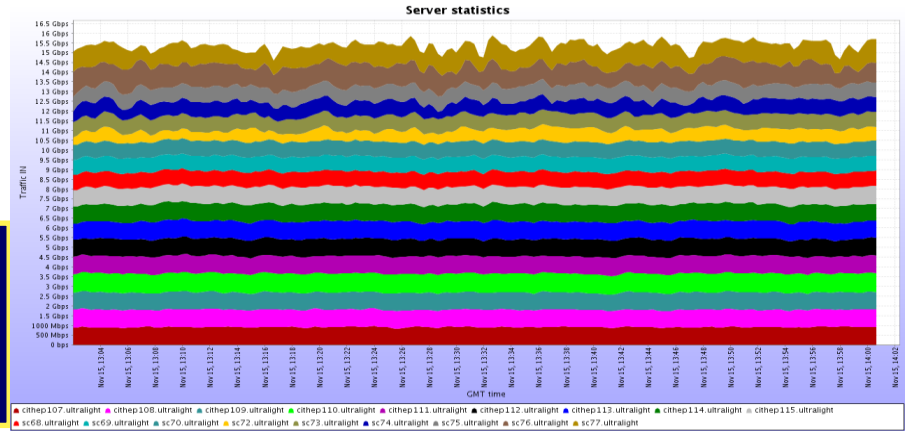
SC06-SC08 BWC: Fast Data Transfer
http://monalisa.cern.ch/FDT

- An easy to use open source Java application that r*uns on all major platforms*
- Uses asynch. multithreaded system to achieve smooth, linear data flow:
  - Streams a dataset (list of files) contin-uously through an open TCP socket
  - ➡ No protocol Start/stops between files
  - Sends buffers at rate matched to the monitored capability of end to end path
  - Use independent threads to read & write on each physical device
- Secure: Can "plug-in" external AAA APIs from major projects

- SC06 BWC: Stable disk-to-disk flows Tampa-Caltech: 10-to-10 and 8-to-8 1U Server-pairs for 9 + 7 = 16 Gbps; then Solid overnight. Using One 10G link
- 17.77 Gbps BWC peak; + 8.6 Gbps to and from Korea

I. Legrand

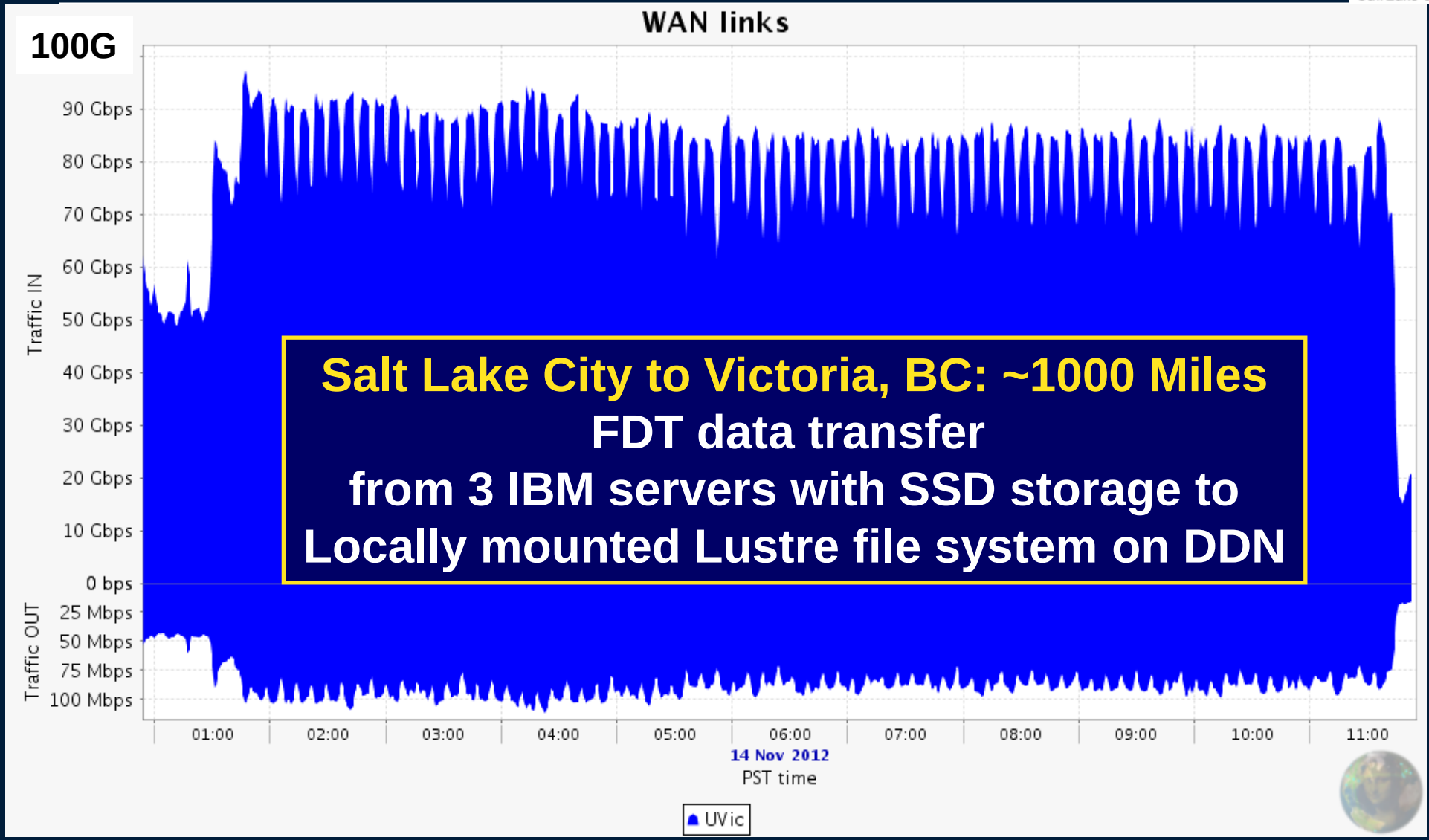




SC07: ~70-100 Gbps per rack of low cost 1U servers

SC12: The Real Thing

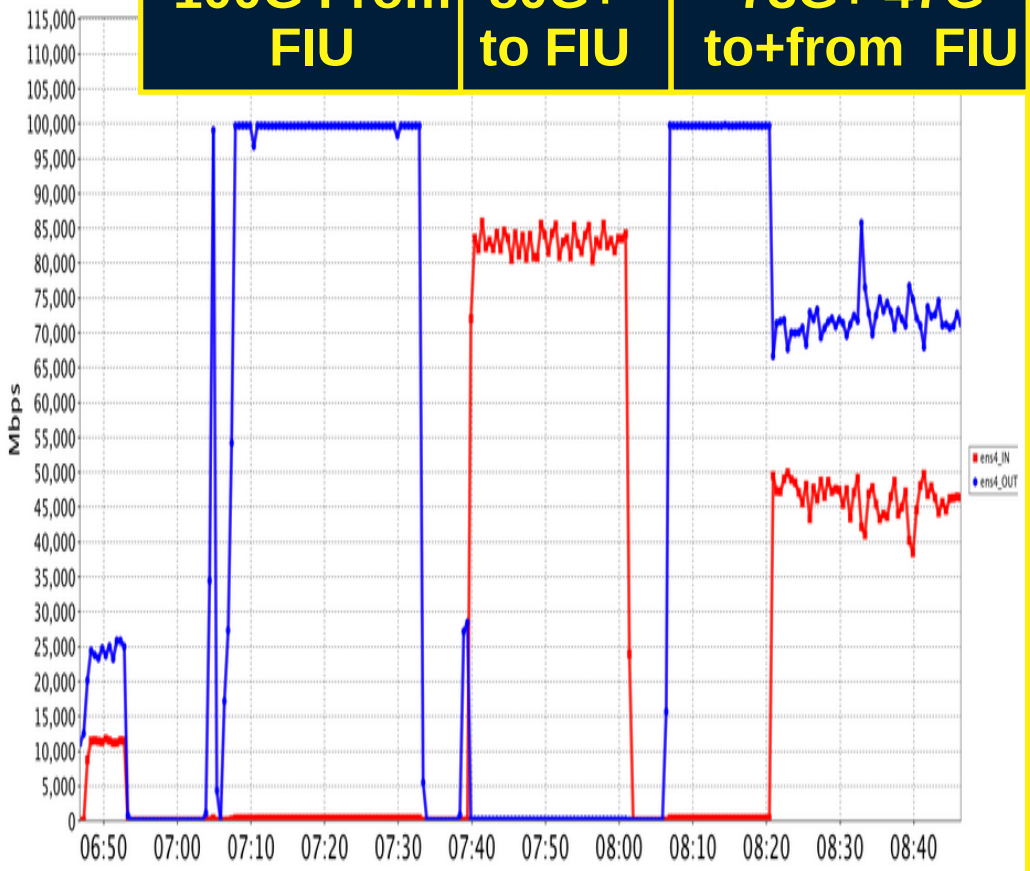
Storage-to-Storage over 100Gbps with FDT



Mellanox and Qlogic 100G and Mellanox N X 100G NIC Results at SC15

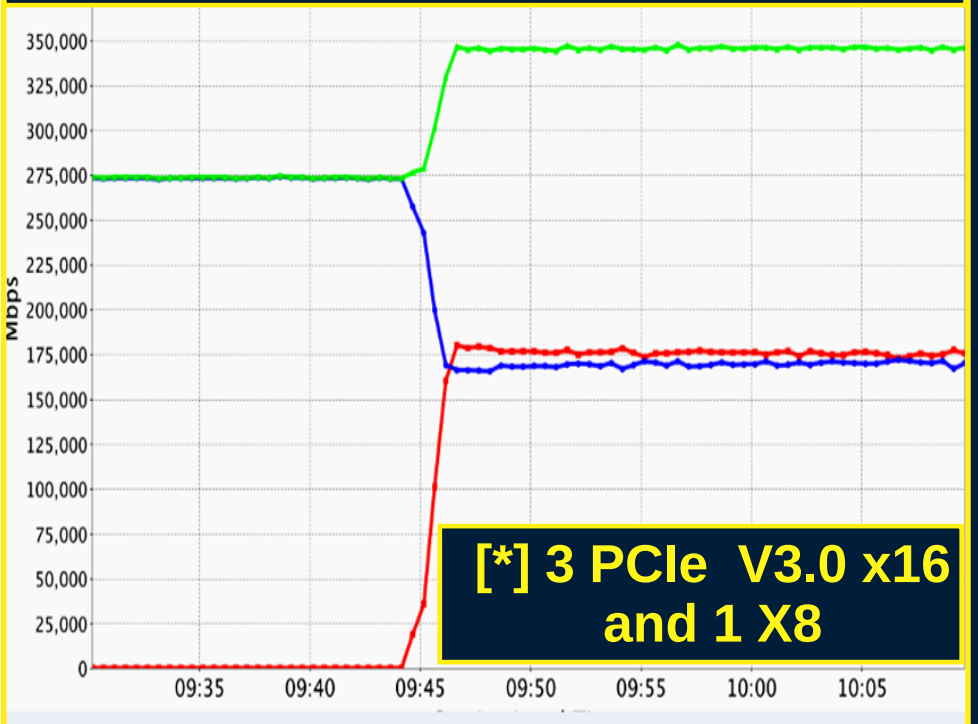
FIU – Caltech Booth – Dell Booth

100G From FIU	80G+ to FIU	73G+ 47G to+from FIU



4 X 100G Server Pair in the Caltech Booth

275G out; 350G in+out [*] Stable Throughput



[*] 3 PCIe V3.0 x16 and 1 X8

Using Caltech's FDT Open Source TCP Application
http://monalisa.caltech.edu/FDT

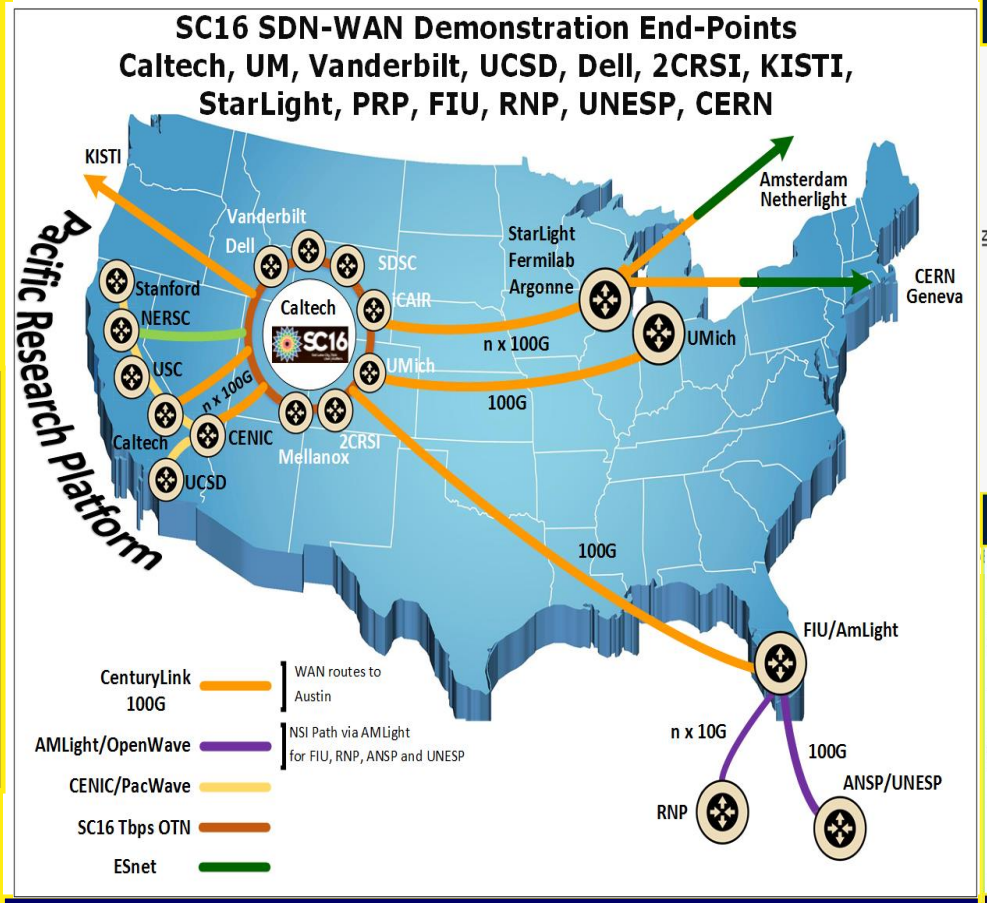
SC15-23: SDN Next Generation Terabit/sec Ecosystem for Exascale Science

supercomputing.caltech.edu

SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

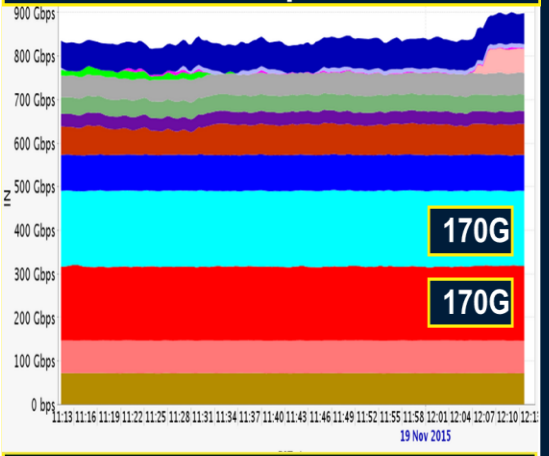
SC16+: Consistent Operations with Agile Feedback Major Science Flow Classes Up to High Water Marks

Preview PetaByte Transfers to/ from Sites With 100G - 1000G DTNs

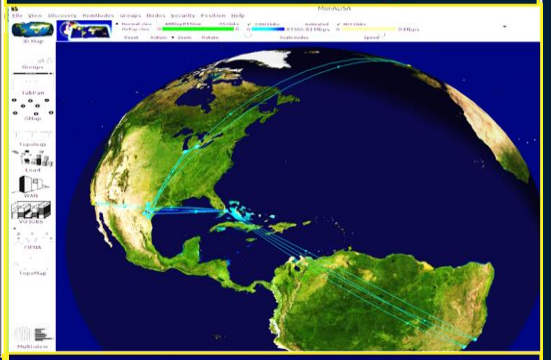


LHC at SC15: Asynchronous Stageout (ASO) with Caltech's SDN Controller

900 Gbps Total
Peak of 360 Gbps in the WAN



Global Topology



29 100G NICs; Two 4 X 100G and Two 3 X 100G DTNs; 1.5 Tbps Capability in one Rack; 9 32 X100G Switches

Tbps Rings for SC18-23: Caltech, Ciena, Scinet, StarLight + Many HEP, Network, Vendor Partners

DTN: ASUS RS520A-E12-RS12U

PCIe 5.0 Ports: Two x16, Two x8, 1 OCP 3.0 x16

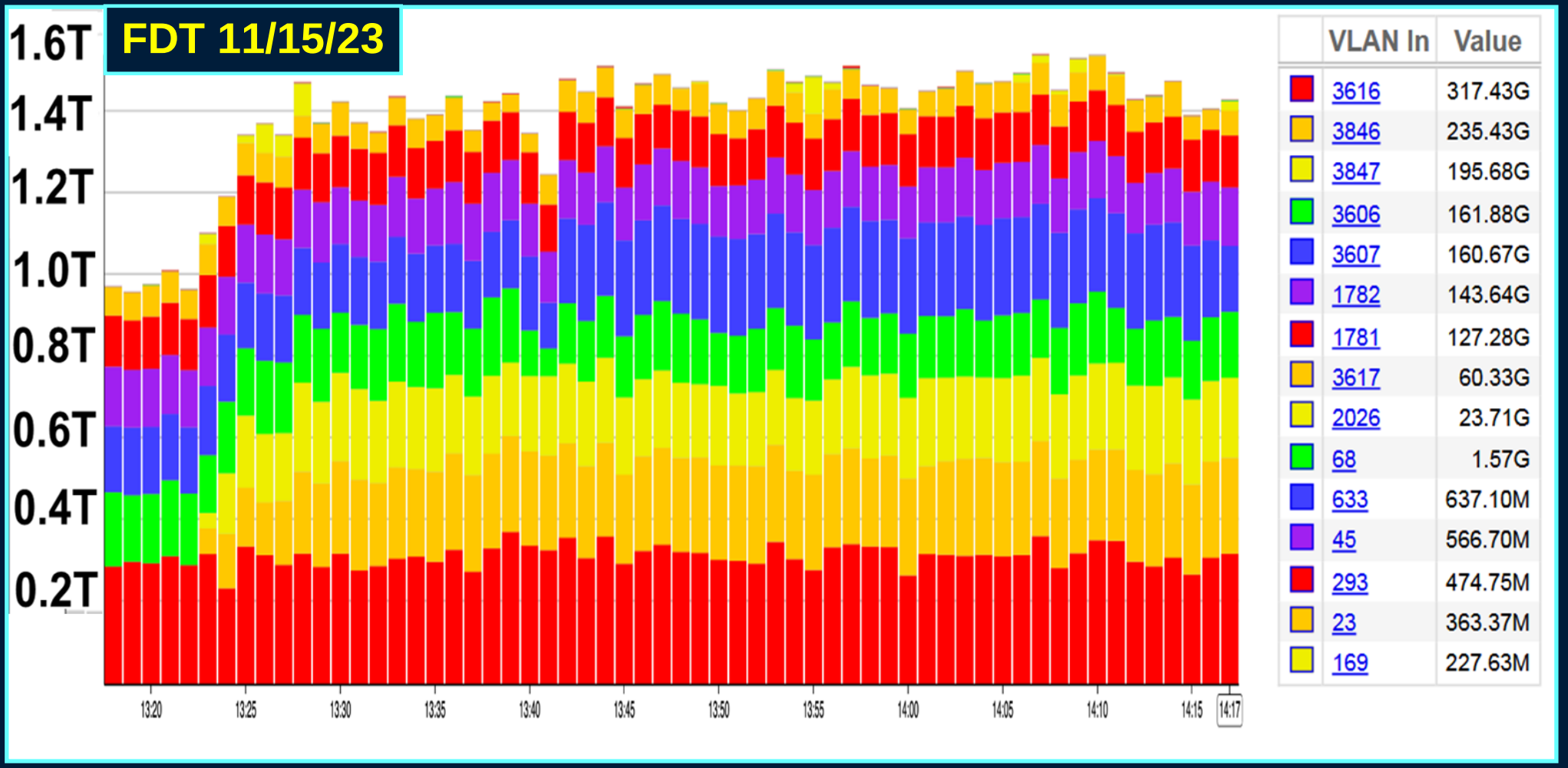
US CMS DTN: CPU EPYC 9374F
3.85 GHz, to 4.3 GHz 32 Core

NIC Setup at SC23 (x2)

ConnectX-7 400GE (200GE)
Two ConnectX-6 200GE
One ConnectX-6 x8 100GE
One 100GE OCP3.0

Tofino1 TNTech	42
Tofino1 BUR001	41
Tofino1 BUR002	40
	39
Dell Z9432F 32 X 400G Switch	38
Arista 7060DX4 32 X 400G Switch	37
	36
ASUS Gen5 DTN1 400G + 3 X 200G + 100G	35 34
ASUS Gen5 DTN2 400G + 3 X 200G + 100G	33 32
Dell Z9664F-ON 64 X 400G Switch	31 30
	29
Dell 730XD DTN 2 X 100G UCSD 1 (2U)	28 27
Dell 730XD DTN 2 X 100G UCSD 2 (2U)	26 25
	24
Dell Z9100 32 X 100G Switch	23
Console	22
Dell S60 Switch	21
Dell 730XD DTN 2 X 100G UCSD3 (2U)	20 19
Dell 730XD DTN 2 X 100G UCSD4 (2U)	18 17
	16
Dell 730XD DTN 2 X 100G UCSD5 (2U)	15 14
Dell 730XD DTN 2 X 100G UCSD6 (2U)	13 12
	11
Dell 730XD DTN 2 X 100G NEU 1 (2U)	10 9
Dell 730XD DTN 2 X 100G SANDIE 9 (2U)	8 7
To ~3 Tbps in a single rack	6 5 4 3 2 1

NRE-13 Top Sources: To 1.5+ Tbps
on 4 X 400G Circuits with Dynamic Transfer Limit

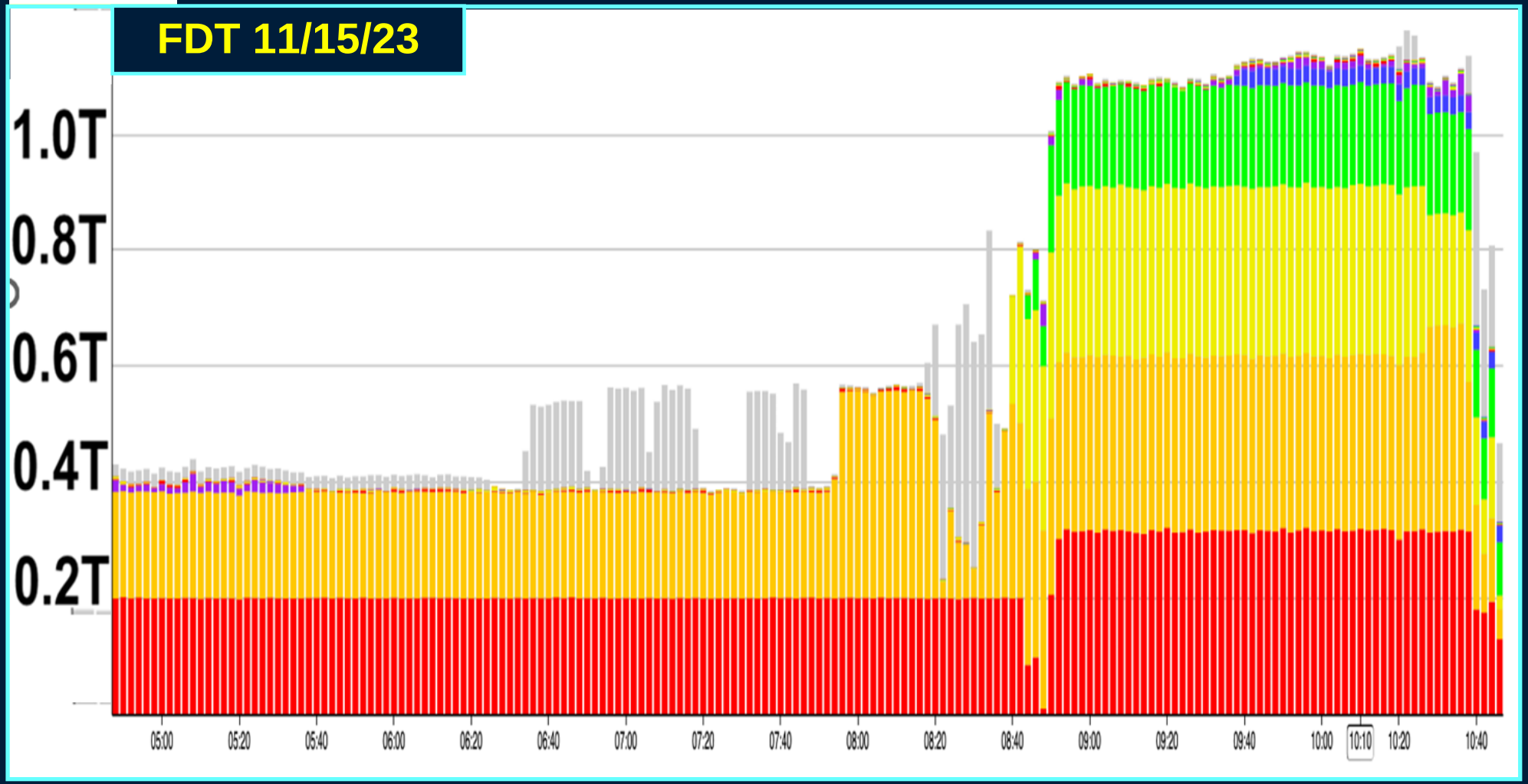


	VLAN In	Value
	3616	317.43G
	3846	235.43G
	3847	195.68G
	3606	161.88G
	3607	160.67G
	1782	143.64G
	1781	127.28G
	3617	60.33G
	2026	23.71G
	68	1.57G
	633	637.10M
	45	566.70M
	293	474.75M
	23	363.37M
	169	227.63M

With Just 2 Gen5 + 2 (of 6) Gen3 Servers at SC23 and 3 Gen5 Servers at Caltech

NRE-13: 1.1 Tbps on 2 X 400G Circuits

Stabilized with Dynamic Thread Management



With Just 2 Gen5 Servers at SC23 and 2 at Caltech

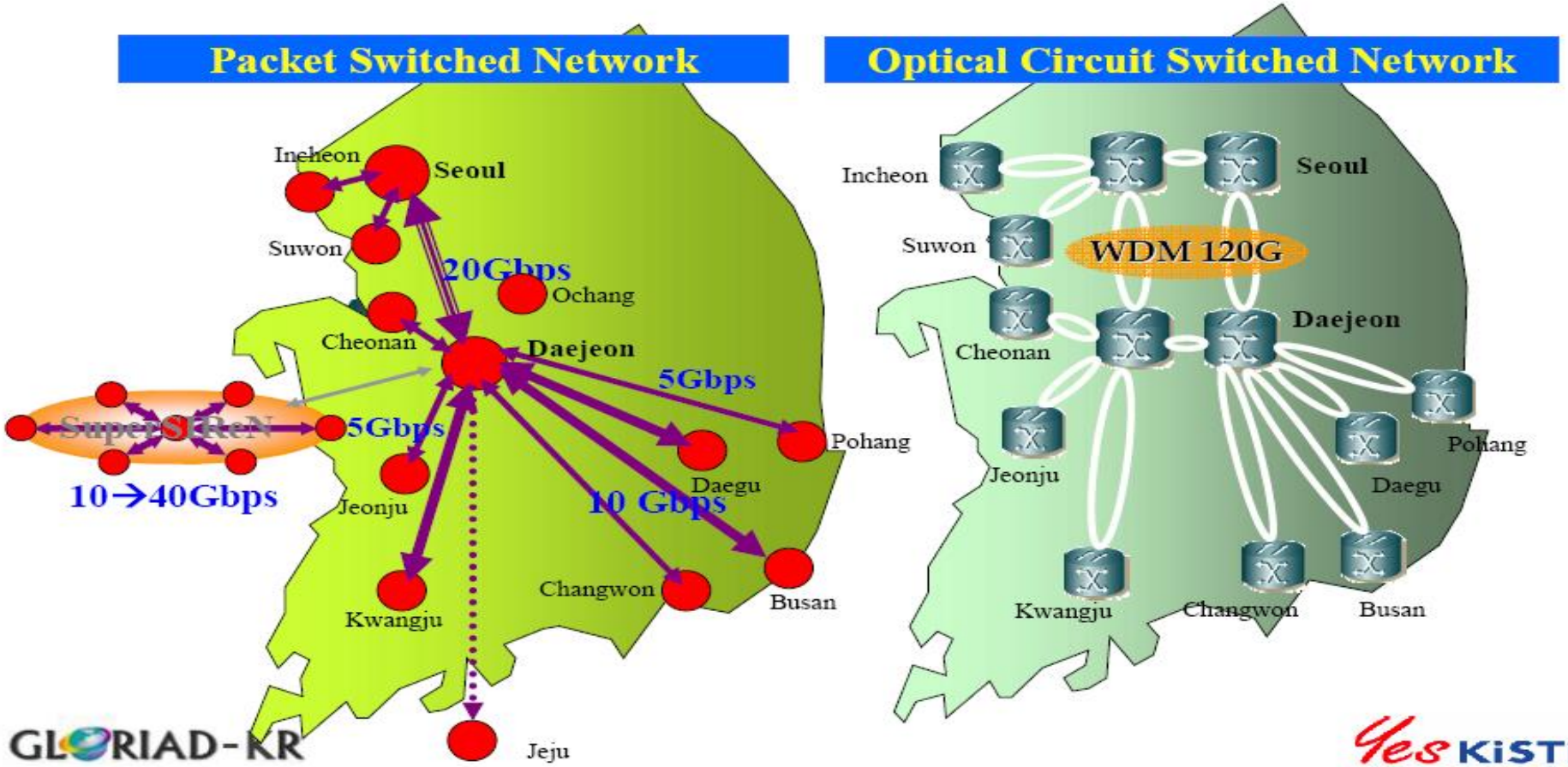
Next Generation Network-Integrated System for HL-LHC and Data Intensive Sciences

- **Top Line Message:** To realize the discovery potential and meet the challenges of the HL-LHC and data intensive science programs, we require a system which:
  - Coordinates worldwide networks as a first class resource along with computing, storage and Analysis Facility Services across world regions
- A global fabric that flexibly allocates, balances and makes best use of the available network resources; AF Services add a real-time dimension
  - Network services negotiating with site services to accelerate workflow
- Builds on ongoing R&D projects: from regional caches/data lakes to intelligent control and data planes to ML-based optimization
- Leverages the worldwide move towards a fully programmable ecosystem of networks and end-systems (P4, SONIC; PolKA, SRv6), and operations platforms (OSG, NRP; GRP, global SENSE Testbed, Global P4 Lab)
- Simultaneously supports the LHC experiments and other data intensive programs while accommodating the traffic that supports the at-large worldwide academic and research community
- The LHC experiments, in particular the CMS data production and AF teams, along with the GNA-G and its Working Groups, and the worldwide R&E network community are key players

Extra Slides Follow

The KREONET Infrastructure Showing the upgraded core network interconnecting 10 regional centers

Hybrid Backbone on KREONet2



Data Lake Concept: Site and In-Network Caches in California; US and Europe

Roughly 30,000 cores across Caltech & UCSD … half typically used for analysis
A 2+ Pbyte Working Example in Production

Site Caches: Plan to include other Tier3s; Tier2s across the US; and at CERN
In-Network Caches: ESnet and Internet2 across US; AMS and London exchange points

Scaling to HL LHC: ~30 Pbytes Per Tier2, ~10 Pbyte Caches, 1 Pbyte Refresh in a Shift Requires 400G Link. **Still relies on use of compact event forms, efficiently managed data transport**

GNA-G - strategic pillars