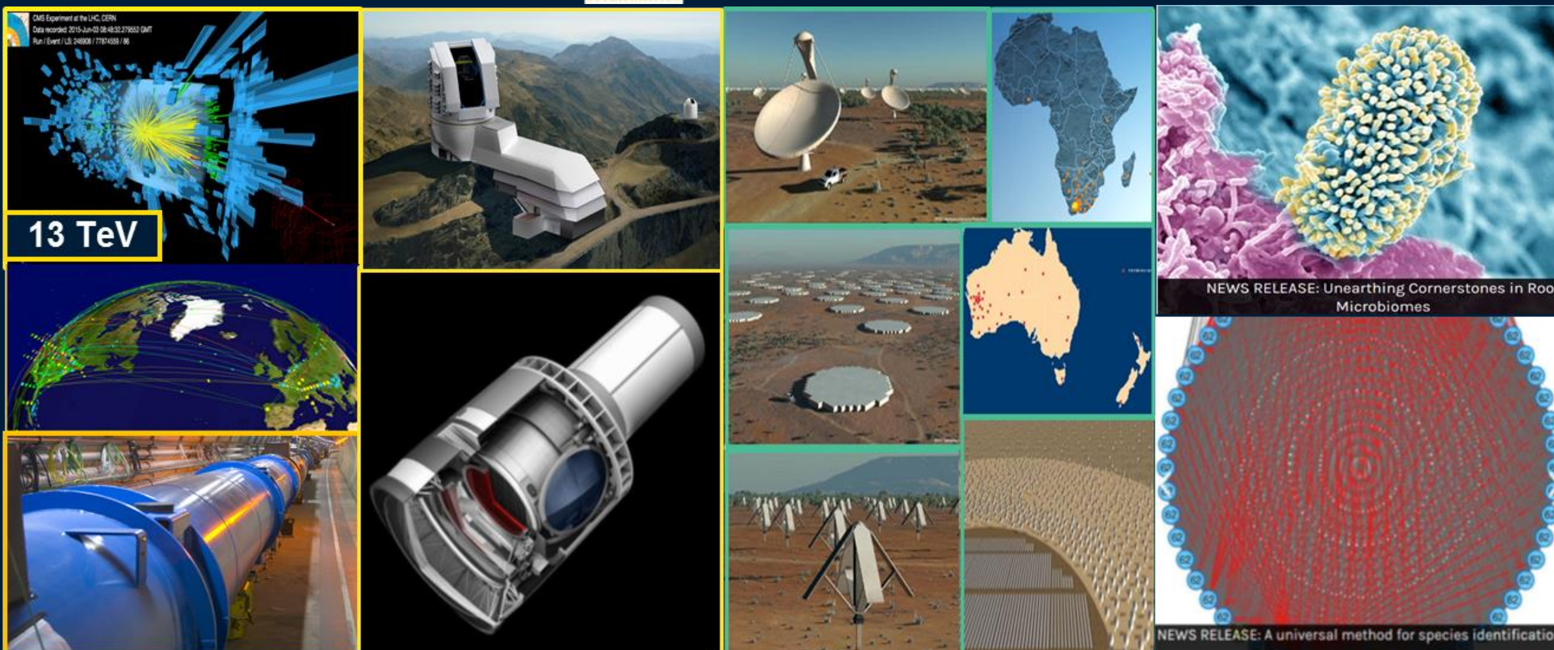
LISHEP 2018: UFBA Salvador, Bahia

Future Computing and Networking for HEP:
Next Generation Systems; Work on the Digital Divide

Future Computing and Networking for HEP:
Next Generation Systems; Work on the Digital Divide

+ INDUSTRY



LHC | LSST | SKA | Joint Genome Institute

LHC Beyond the Higgs Boson

LSST SKA

BioInformatics

Earth Observation

Gateways to a New Era

Harvey Newman, Caltech

LISHEP 2018, UFBA Salvador

September 12, 2018

Discovery of the Higgs Boson
July 4, 2012

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	●	●	●

A 48 Year Search

Theory : 1964

LHC + Experiments
Concept: 1984

Construction: 2001

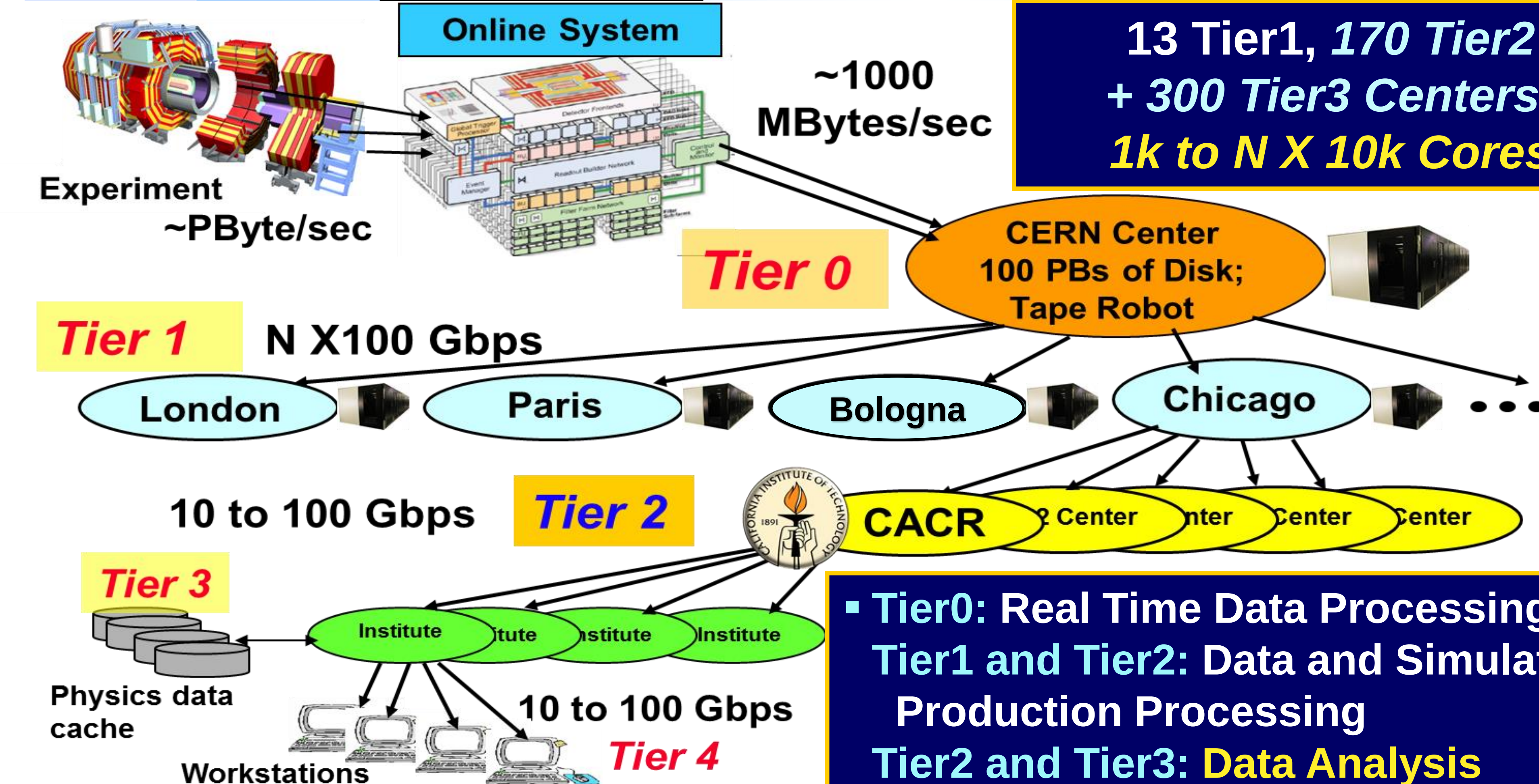
Operation: 2009
Discovery: 2012

Advanced Networks
Were Essential to the Higgs Discovery and Every Ph.D Thesis of the last 20+ Years

They will be Essential to
Future Discoveries,
and Every Ph. D Thesis to Come

Global Data Flow: LHC Grid Hierarchy
A Worldwide System Originated at Caltech (1999)

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



- **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)
Increasing Use of Additional HPC and Cloud Resources
A Global Dynamic System: Fertile Ground for Control with ML

Entering a new Era of Exploration and Discovery in Data Intensive Sciences

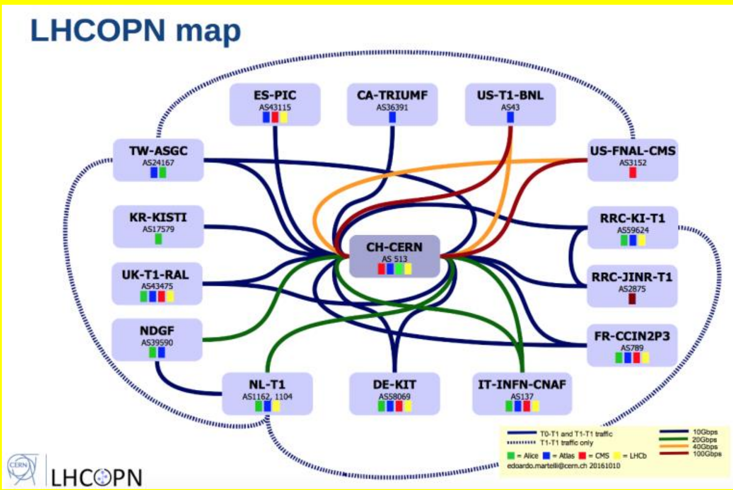
- We have entered a new era of exploration and discovery
  - In many data intensive fields, from HEP and astrophysics to climate science, genomics, seismology and biomedical research

- The largest data- and network-intensive programs from the LHC and HL LHC, to LSST, LCLS II, the Joint Genome Institute and other emerging areas of growth will face a new round of unprecedented challenges
  - In global data distribution, processing, access and analysis
  - In the coordinated use of massive but still limited CPU, storage and network resources.

- High-performance networking is a key enabling technology for this research: global science collaborations depend on fast and reliable data transfers and access on regional, national and international scales

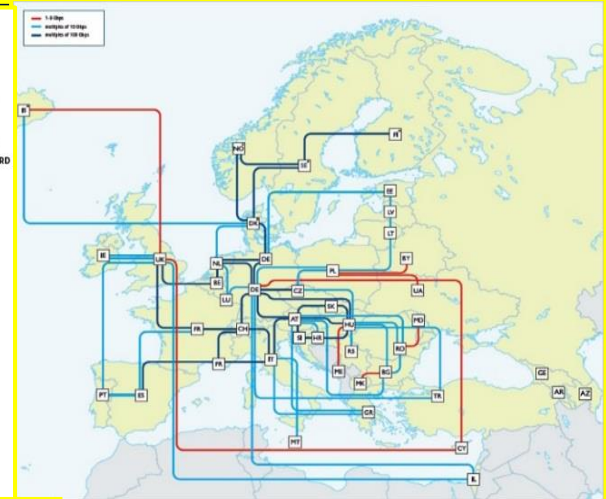
Core of LHC Networking LHCOPN, LHCONE, GEANT, ESnet, Internet2, CENIC

LHCOPN: Simple & Reliable Tier0+1 Ops

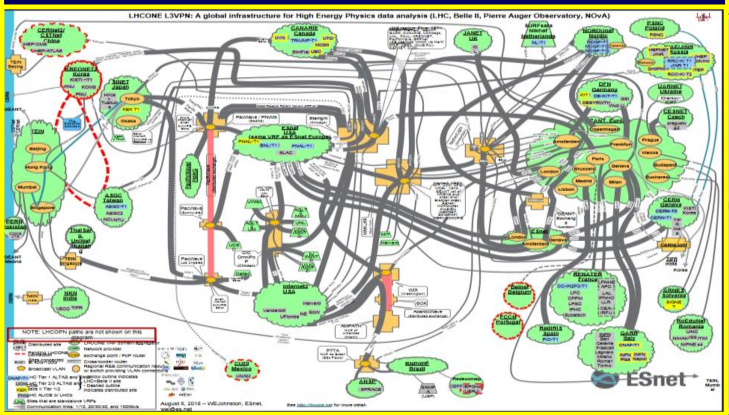


Internet2

GEANT

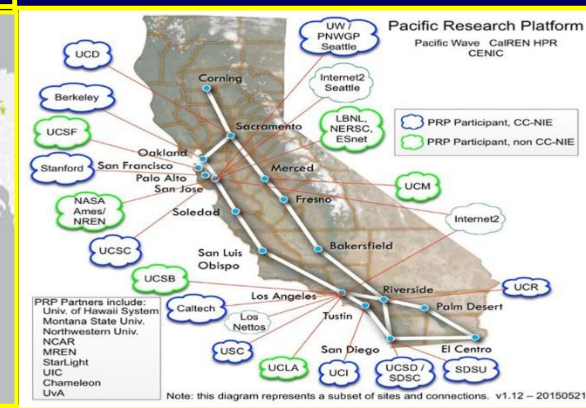


LHCONE VRF: 170 Tier2s



ESnet (with EEX)

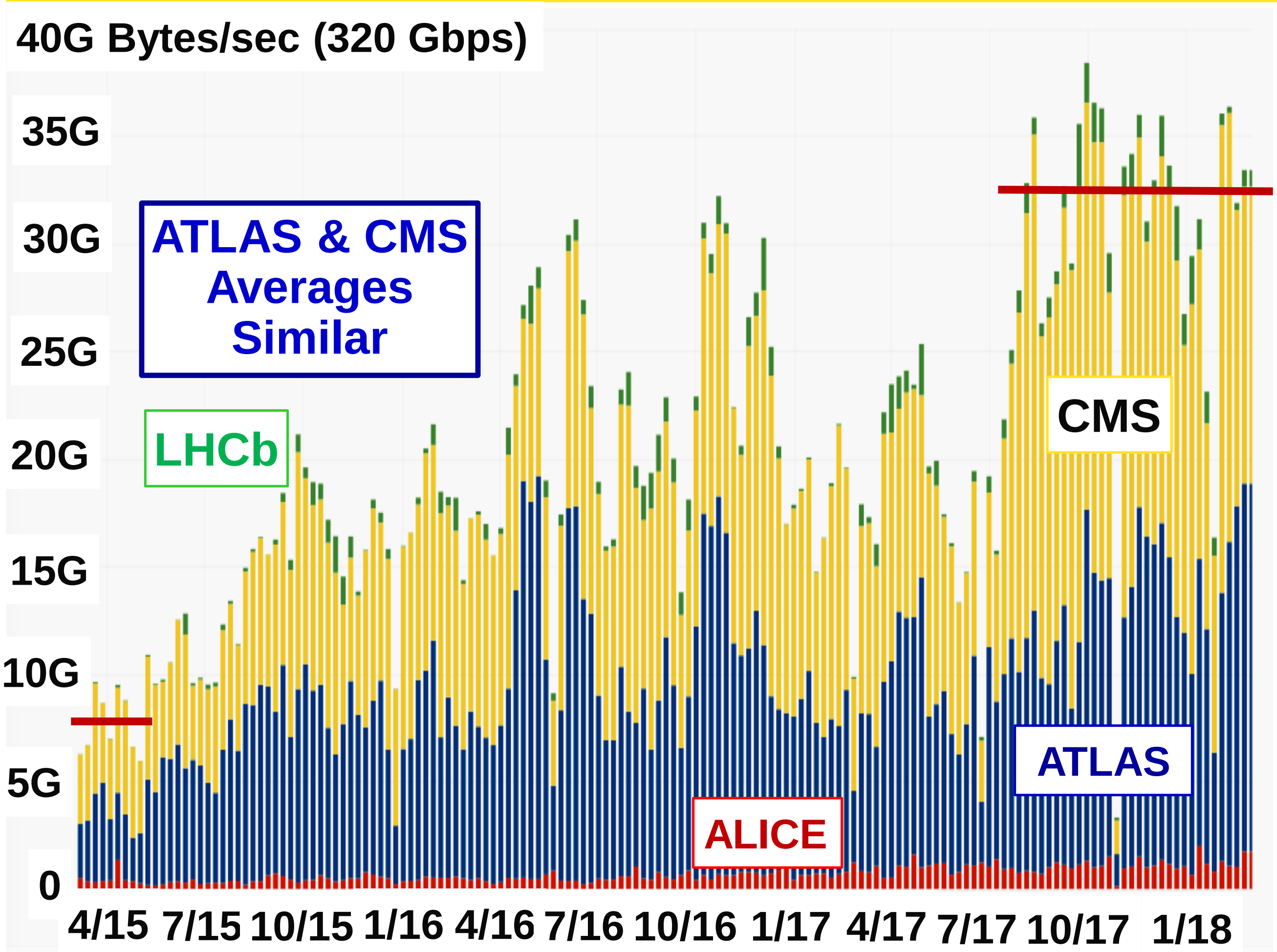
CENIC and PRP



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

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

WLCG Transfers Dashboard: Throughput March 2015 – March 2018



32 GBytes/s Sustained

45+ GBytes/s Peaks

Complex Workflow

- Multi-TByte Dataset Transfers;
- 6-12 M Transfers/Day
- >100k of remote connections (e.g. AAA) simultaneously

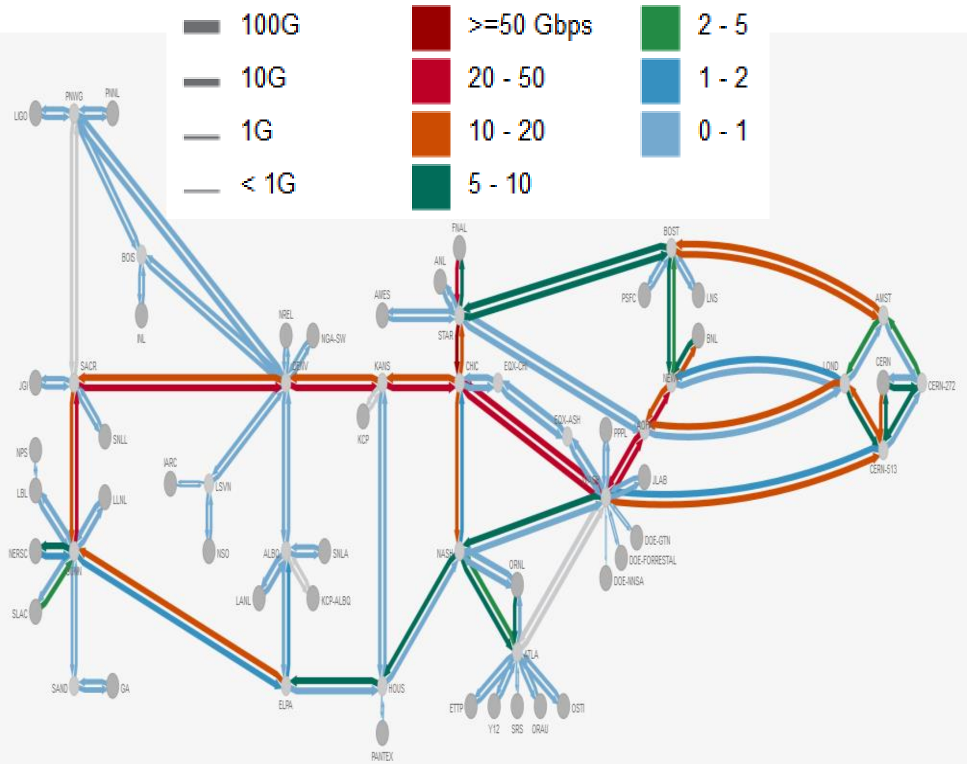
4X Growth in Sustained Throughput in 3 Years: +60%/Yr; 100X per Decade

Energy Sciences Network
2018 Updates and Outlook

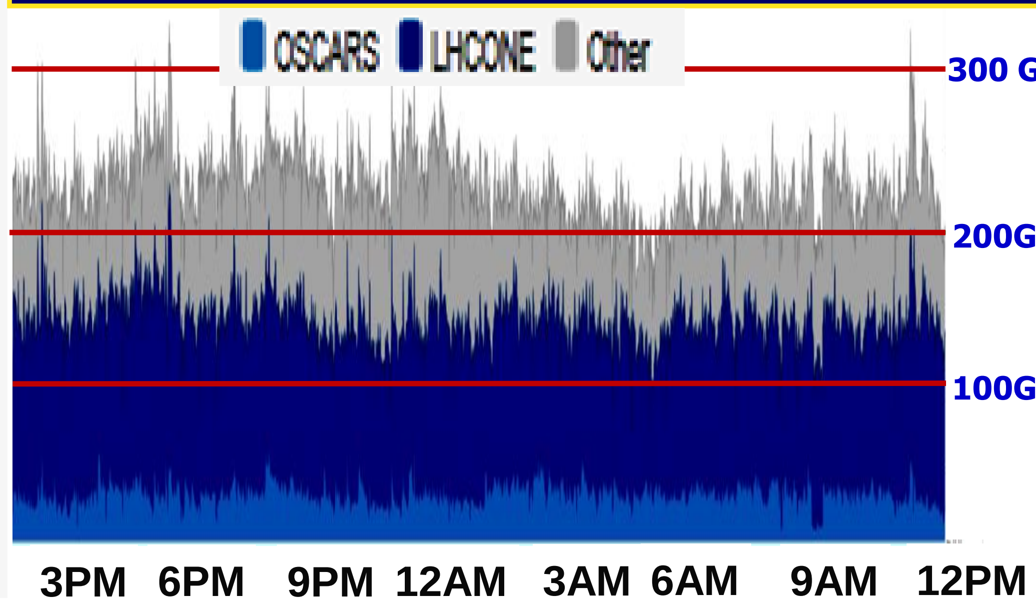
ESnet

my.es.net Portal

- 230-260 Gbps Typical: Peaks to 300+ Gbps
- Long term traffic growth of 72%/year (10X per 4 Years)
- Incoming Traffic 78 PB/mo in 12/17
- LHCONE *Rapid* Growth: 1.7 to 37 PB/Mo from 1/2015 to 12/2017



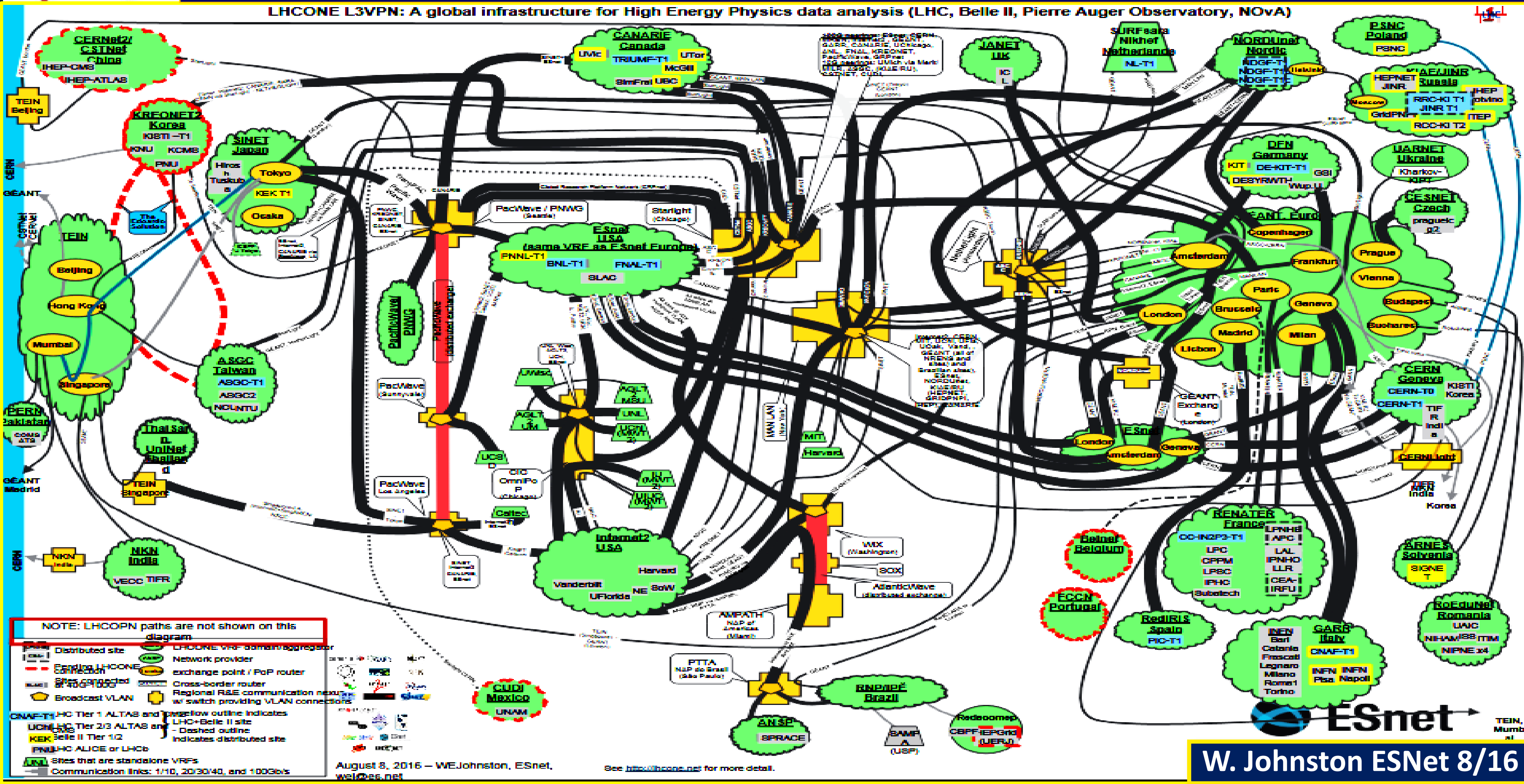
- LHCONE: *Now the largest class of ESnet traffic*



✷ *ESnet6:* next SDN-enabled generation for 2019-2025: *Tbps* on major routes

LHCONE: a Virtual Routing and Forwarding (VRF) Fabric

A global infrastructure for *HEP (LHC, Belle II, NOvA, Auger, Xeon)* data flows



W. Johnston ESNet 8/16

Good News: The Major R&E Networks Have Mobilized on behalf of HEP

Issue: A complex system with limited scaling properties.

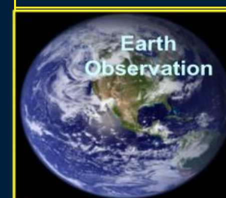
LHCONE traffic grew by ~25X in 24 months: a challenge during Run2

A New Era of Challenges: Global Exabyte Data Distribution, Processing, Access and Analysis

- Exascale Data for the LHC Experiments
  - 0.5 EB in 2017, 1 EB by end of Run2
  - To ~50 EB during HL LHC Era
- Network Total Flow of >1 EB this Year
  - 900 Pbytes flowed over WLCG in 2017
- Projected (2016-26) HL LHC Shortfalls
  - CPU ~3X, Storage ~3X, *Networks*
- Network Dilemma: Per technology generation (~10 years)
  - Capacity at same unit cost: 4X
  - Bandwidth growth: 30X (Internet2); 50X (GEANT), 70X (ESnet)
  - Similarly for NICs: 400G by ~2026
- *During LHC Run3 (~2022) we will likely reach a network limit*
- This is unlike the past
  - Optical, switch advances are evolutionary; *physics barriers ahead*

New Levels of Challenge

- Global data distribution, processing, access and analysis
- Coordinated use of massive but still limited *diverse* compute, storage and network resources
- Coordinated operation and collaboration *within and among* scientific enterprises

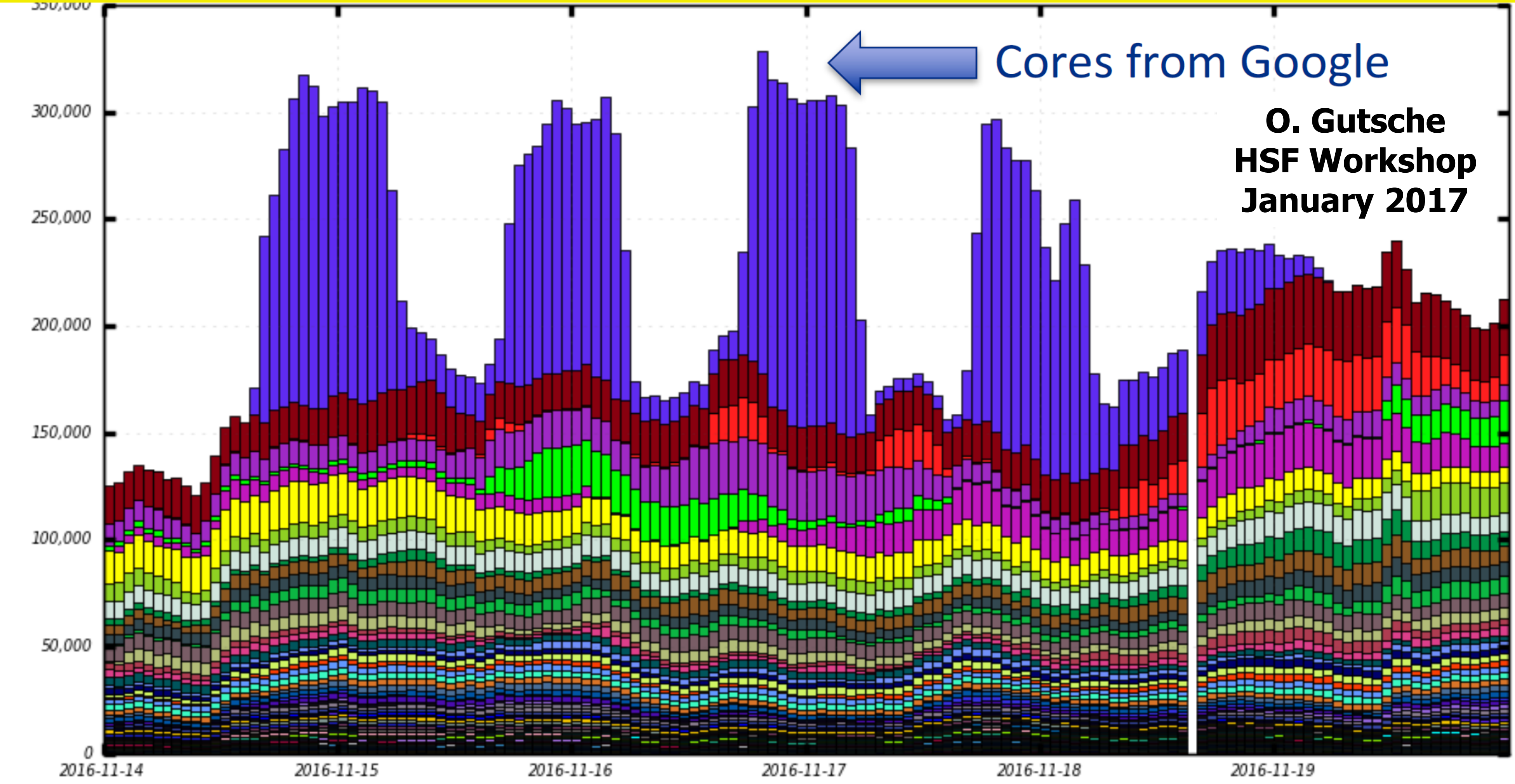



- HEP will experience increasing Competition from other data intensive programs
  - Sky Surveys: LSST, SKA
  - Next Gen Light Sources
  - Earth Observation
  - Genomics

HEPCloud Facility: Fermilab Bursting to the Cloud.

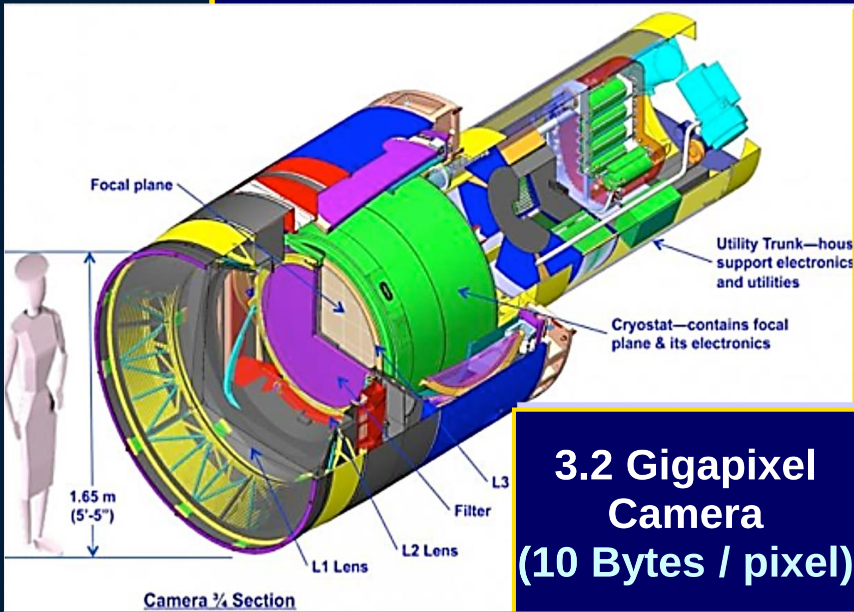
Doubling CMS Peak Compute Capacity

Meeting the Challenge of Bringing Diverse Resources to Bear



Issue: Cloud Business Model for Data + Network Intensive Use Cases

LSST + SKA Data Movement
Upcoming *Real-time* Challenges for Astronomy



3.2 Gigapixel Camera (10 Bytes / pixel)

Dark Matter
Dark Energy
Changing Sky
Solar System
Milky Way

SKA

- **Planned Networks**: Dedicated 100Gs (Spectrum !) for image data, +100G for other traffic, and ~40G and up for diverse paths
- **Lossless compressed Image size = 2.7GB (~5 images transferred in parallel over a 100 Gbps link)**
  - **Custom transfer protocols for images (UDP Based)**
- **Real-time Challenge: delivery in seconds to catch cosmic “events”**
- **+ SKA in Future: 3000 Antennae covering > 1 Million km2;** ***15,000 Terabits/sec to the correlators ➡ 0.3-1.5 Exabytes/yr Stored***

Facts of Life: Towards the Next Generation Computing Model

- The projected needs and shortfalls in computing, storage and networks for the HL LHC era remain daunting (20 X CPU, 50 X Disk and Tape)
  - This remains true with our best code improvements, and foreseeable technology evolution in industry through 2026
  - We must take steps to mitigate this mismatch: efficient use of all resources, including exascale computing systems, vectorization, massive parallelization (GPUs, FPGAs) etc.
- Industrial technology visions are towards “heterogeneous architectures”
  - Adapted to “AI” at all levels, from HPC systems to desktops, laptops, cellphones, and IOT devices and systems (e.g. smart cities)
  - General Direction: more towards enabling new capabilities (and market opportunity) rather than lowering costs
- Beyond 2026: “physics barriers” will be felt as limits are reached
  - The transition to nanoscale post-silicon electronics, 2D materials and new interfaces will require large global investments; not yet evident
- The network dimension: traffic growth rate far outstrips cost evolution
  - Constraints may already be felt during LHC Run3; definitely by HL LHC
- Pointing towards a new deeply network-aware next generation system

Perspective on Improvements
Aimed at Reducing the Challenge (I)

- *Parallelization to reduce the memory footprint:*
  Already in place in CMSSW; how much farther can we go ?
- *Use of vectorization features in processors:*
  How much marginal benefit can be achieved in practice (e.g. GEANT V experience)
- *Use of GPUs, FPGAs and coming processors with "built in AI": Training phase (on GPUs) remains CPU, power and cost-intensive:*
  - *Do we build special clusters or run on national HPC (eventually exascale) facilities ?*
  - *If the former: Are the new architectures affordable ?*
  - *If the latter: how much can this scale before we must pay for support manpower, front end large caches, edge networks, etc.) ?*
  - *NOTE: Inference phase more favorable (including in our triggers)*

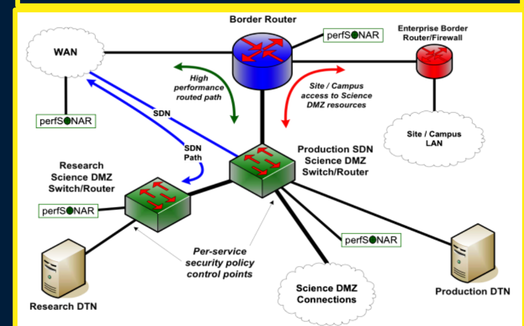
Operational Pilot for Exascale and other HPC Facilities with Petabyte Transactions

➡ Targeting the CPU Needs at LHC Run3 and HL LHC

- Develop system architectures in HW + software for petabyte transactions (to/from exabyte stores)
  - Edge clusters with petabyte caches
    - Input + output pools: ~10 to 100 Pbytes
  - A handful of proxies at the edge
    - To manage and focus security efforts
  - Extending Science DMZ concepts
    - Enabling 100G to Tbps SDNs with Edge/WAN Coordination
  - Identifying + matching HEP units of work to specific sub-facilities adapted to the task
  - Site-Network End-to-End Orchestration
    - Efficient, smooth petabyte flows over 100G then 400G (2020) then ~1 Tbps (2025) networks

Leadership

Next Gen Science DMZ



Major Changes for 2026 and Beyond:

Several Technology/Physics Barriers are Being Approached

- *Processor designs: 10 nm feature size in processors is difficult/delayed; 7 nm or perhaps 5 nm feature size appears to be a practical limit*
- *Disk Storage: Below ~10nm: "superparamagnetic limit", large investments will be needed for new techniques.*
- *Network price/performance improvement has similarly slowed.*
  - *Time for 400GE to be at same price point as the present 100GE generation has slowed; likely to be 10 years (Price/perf. ~ -13%/year)*
  - *Network usage is increasing much faster than this over the medium and long term: ~4X since 2015 (50-60% per year)*
  - *By ~2026 we will likely hit an energy density limit in switches*
- *Conclude: By around 2026-2028 the need for nanoscale processes, 2D materials (and huge investments) will likely be felt throughout industry*
- *Bottom Line: We need to carefully follow these technology developments; work with prototype and 1st-generation systems when relevant*

Vision: Next Gen Integrated Systems for Exascale Science: a Major Opportunity

Opportunity: Exploit the Synergy among

1. **Global operations data and workflow management systems developed by HEP programs, *to respond to both steady state and peak demands***
   - **Evolving to work with *increasingly diverse (HPC) and elastic (Cloud) resources***
2. **Deeply programmable, agile software-defined networks (SDN), emerging as multidomain network operating systems (e.g. SENSE & SDN NGenIA; Multidomain multicontroller SDN [*])**
3. **Machine Learning, modeling and game theory: Extract key variables; move to real-time self-optimizing workflows with Reinforcement Learning.**

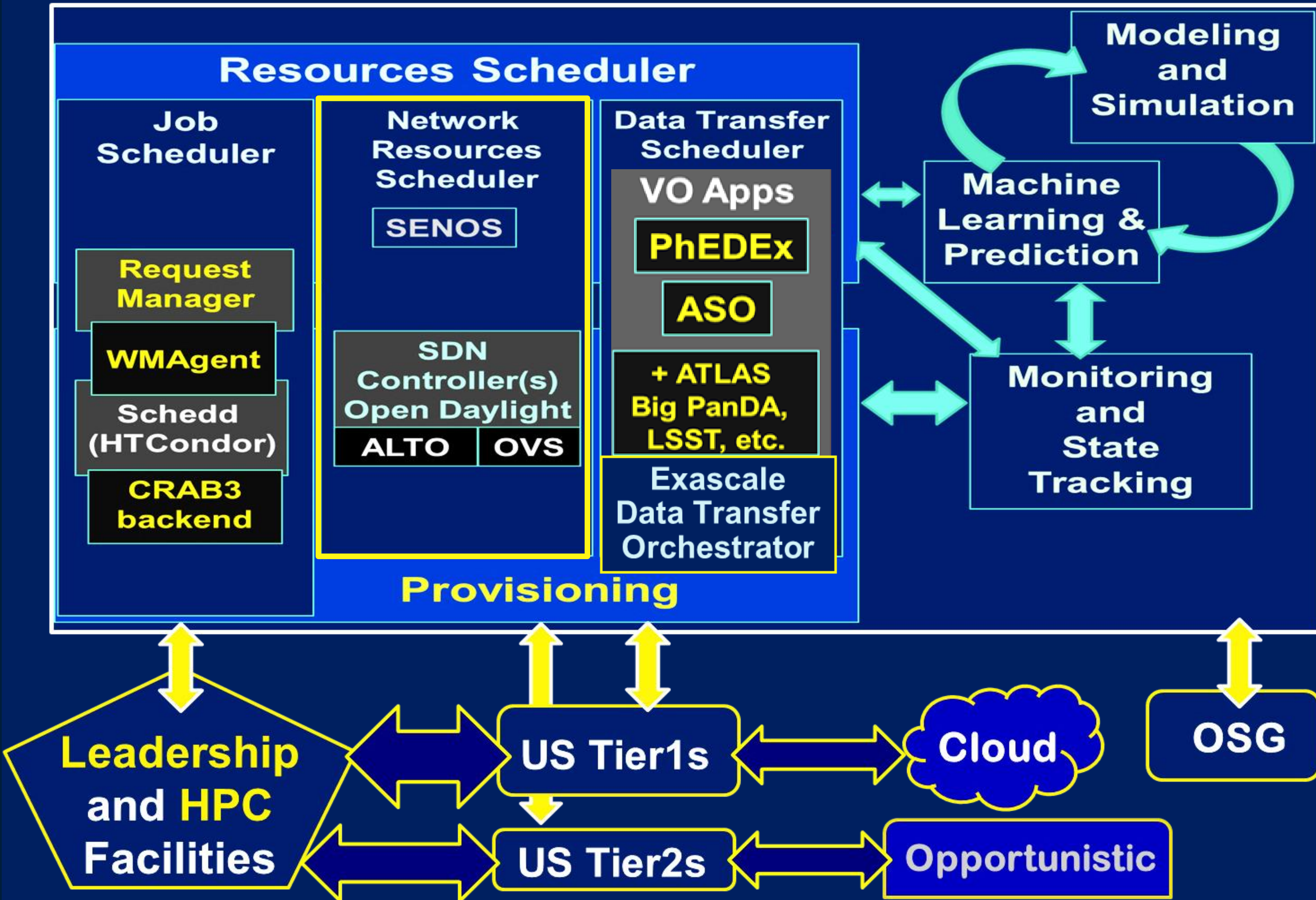
[*] *See https://www.ietf.org/id/draft-xiang-alto-multidomain-analytics-01.txt*

Responding to the Challenges

New Overarching "Consistent Operations" Paradigm

- VO Workflow Orchestration systems that are real-time and reactive ☑;
    - New: Making then *Deeply network aware, and proactive*
- *Network Orchestrators with similar, real-time character*
- Together responding moment-to-moment to:
  - State changes in the networks and end systems
  - Actual versus estimated transfer progress, access IO speed
- Prerequisites:
  - End systems, data transfer applications and access methods capable of high of throughput [➡ FDT]
  - Realtime end-to-end monitoring systems [End sites + networks]
- Modern Elements for efficient operations within the limits: SDN-driven bandwidth allocation, load balancing, flow moderation at the network edges and in the core
- Goal: Best use of available network, compute and storage resources while avoiding saturation and blocking other network traffic

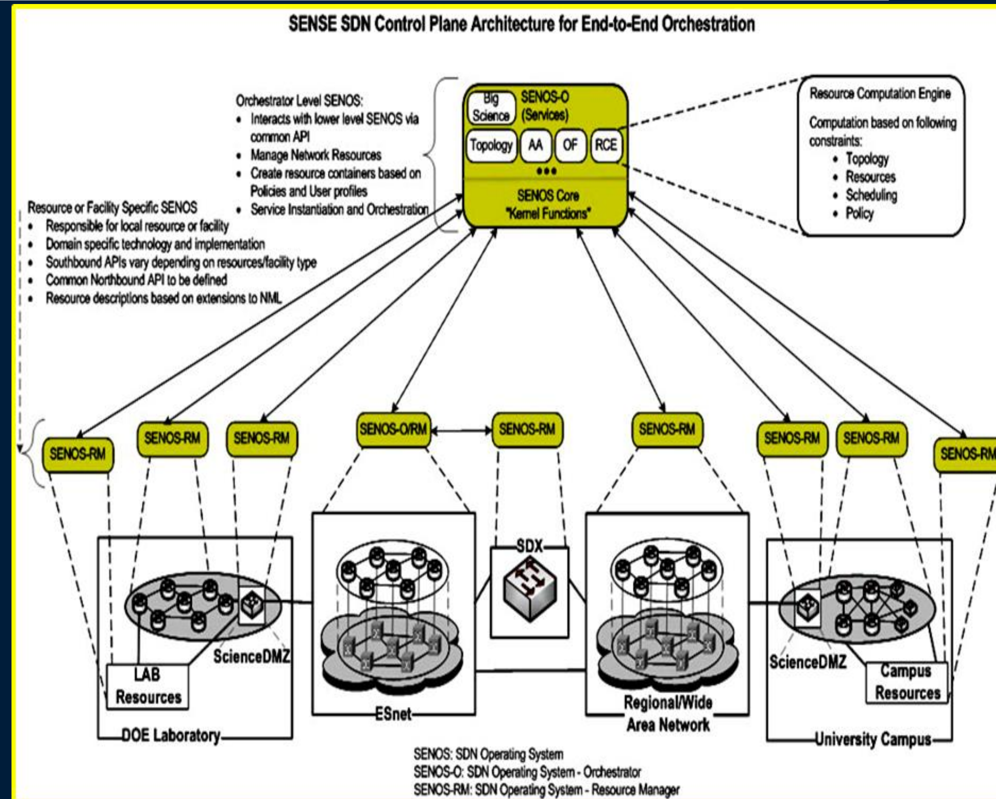
NGenIA-ES Services and Data Flow Diagram



SENSE: SDN for End-to-end Networked Science at the Exascale

ESnet Caltech Fermilab LBNL/NERSC Argonne Maryland

- **Mission Goals:**
  - Improve end-to-end performance of science workflows
  - Enabling new paradigms: e.g. creating dynamic distributed 'Superfacilities'.
- **Comprehensive Approach:** An end-to-end SDN Operating System, with:



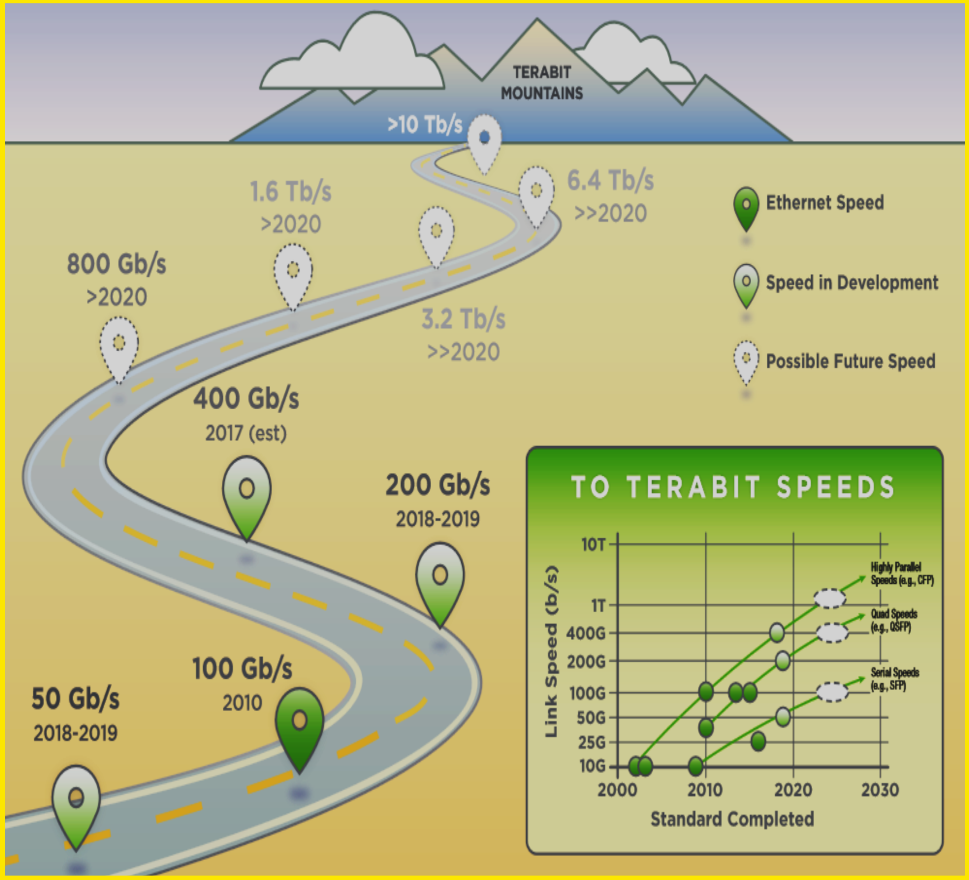
  - Intent-based interfaces, providing intuitive access to intelligent SDN services
  - Policy-guided E2E orchestration of resources
  - Auto-provisioning of network devices and Data Transfer Nodes
  - Network measurement, analytics and feedback to build resilience

Riding the Ethernet Wave: Petabyte Transactions
To Create the NG and NNG Ecosystems: 2018 – 2026+

- A petabyte transfer would occupy a 100G link for 24 hrs at wire speed now
- Circa 2021, a PB transfer could take 6 hours on a 400G link
- In the exascale era, circa 2023-5 a PB would take 90 minutes on a 1.6 Tbps link
  - Providing some agility
  - Beginning to allow Multiple transactions
- Through the HL LHC era we can foresee Next-to-Next Generation Systems with
  - Increasing agility
  - Larger and multiple transactions

Ethernet Alliance Roadmap

http://www.ethernetalliance.org/wp-content/uploads/2015/03/Ethernet-Roadmap-2sides-29Feb.pdf



Beyond the Terabit Mountains: Physics - the Nanoscale Era

SC15-18: SDN Next Generation Terabit/sec Ecosystem for Exascale Science

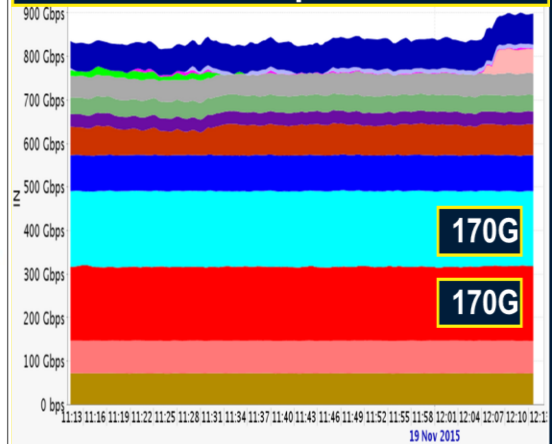
supercomputing.caltech.edu

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 Site Edges of Exascale Facilities With 100G -1000G DTNs

900 Gbps Total
Peak of 360 Gbps in the WAN



Global Topology

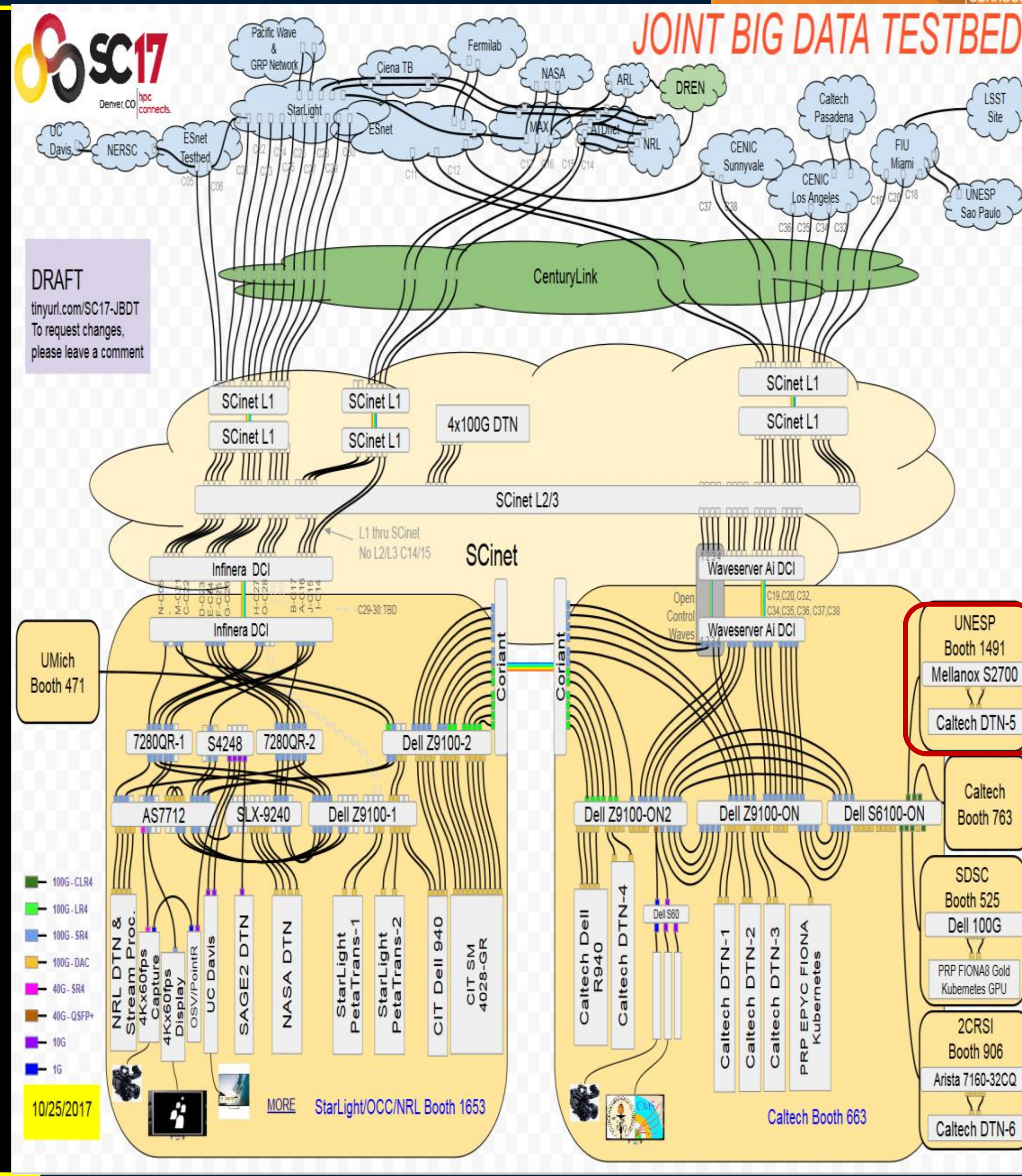
LHC at SC15: Asynchronous Stageout (ASO) with Caltech's SDN Controller

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 Ring for SC17: Caltech, Ciena, Scinet, OCC/ StarLight + Many HEP, Network, Vendor Partners at SC16

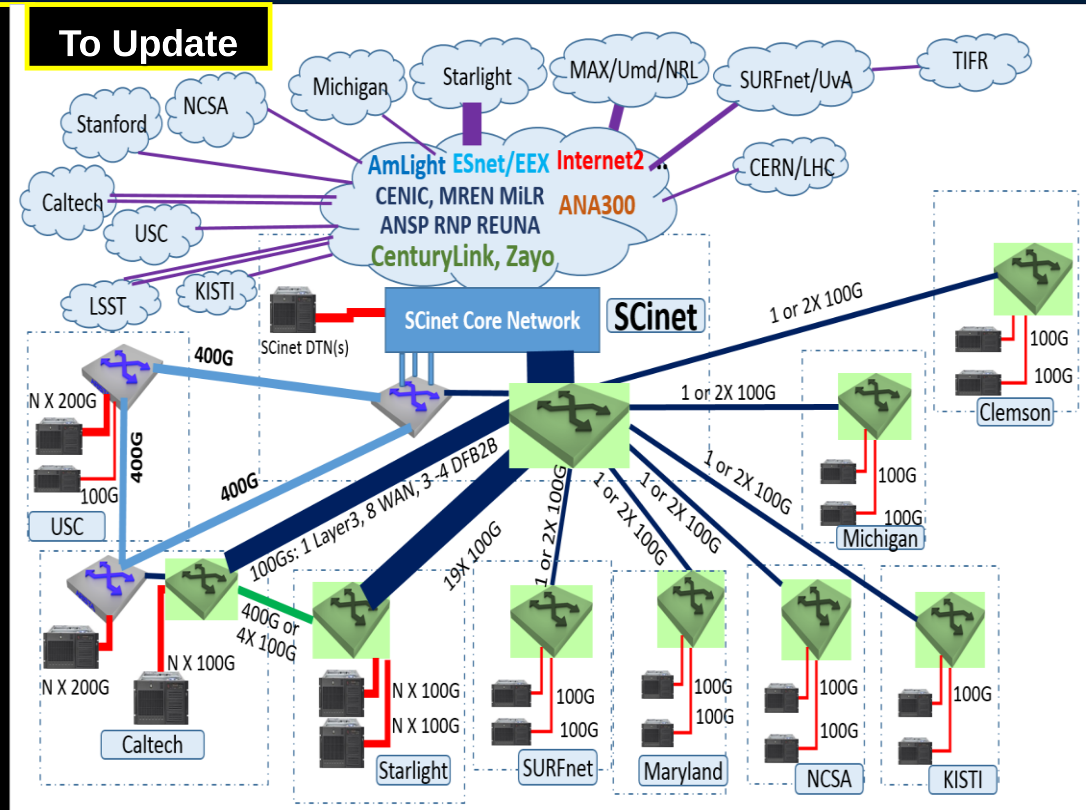
Caltech and Partners at SC17

- ~3 Tbps capacity each at the Caltech and Starlight/OCC booths; 3 Tbps External
- 1+ Tbps between the Booths and ~3Tbps to the WAN
- Caltech Booth: 200G dedicated to Caltech campus; 300G to PRP (UCSD, Stanford, UCSC, et al); 300G to Brazil+Chile via Miami; 200G to ESnet
- Waveserver Ai + Coriant DCIs in the booths: *N X 100GE to 400G, 200G waves*
- N X 100G Servers: *Supermicro and Dell; Mellanox 100G ConnectX-4 NICs*
- Intel P4500, 4608, Optane SSDs



Caltech and Partners at SC18

- ~2-3 Tbps Capacity each at the Caltech and Starlight Booths
- Science + Network Booth Partners: SCinet + USC, Starlight (4), SURFnet (2), Maryland, KISTI, Michigan, Clemson, (NCSA)
- WAN Sites: Caltech, Maryland, Starlight, PRP [UCSD, Stanford, LBNL], CENIC, NCSA, Fermilab, AmLight, LSST, UNESP, SURFnet, KISTI/KASI, CERN, TIFR
- Caltech Booth: 200G dedicated to Caltech campus; 300G to PRP via CENIC (UCSD, Stanford, LBNL, UCSC, et al); 200G to Brazil+Chile via Miami; 200G to Esnet via Sunnyvale
- Ciena 8180 Platform in the Caltech booth: *400G waves, NX100G clients*
- Four 400GE Arista Switches: “CUS” Triangle + Starlight Booth Leg; OSFP Optics, 400G to 4 X 100G Splitter Cables



LHC, LSST Programs. SENSE, SANDIE, Mercator, Kytos, etc

- Servers: 26 to 38 X 100G NICs
  2 or 3 4-Node Supermicro with 16-24 100G NICs and Intel SSDs; 2 Rackspace Barreleye G2 (Power9) PCIe Gen4: 2 X 200G in each booth; 4 Echostreams 1U with 8-12 X 100G NICs

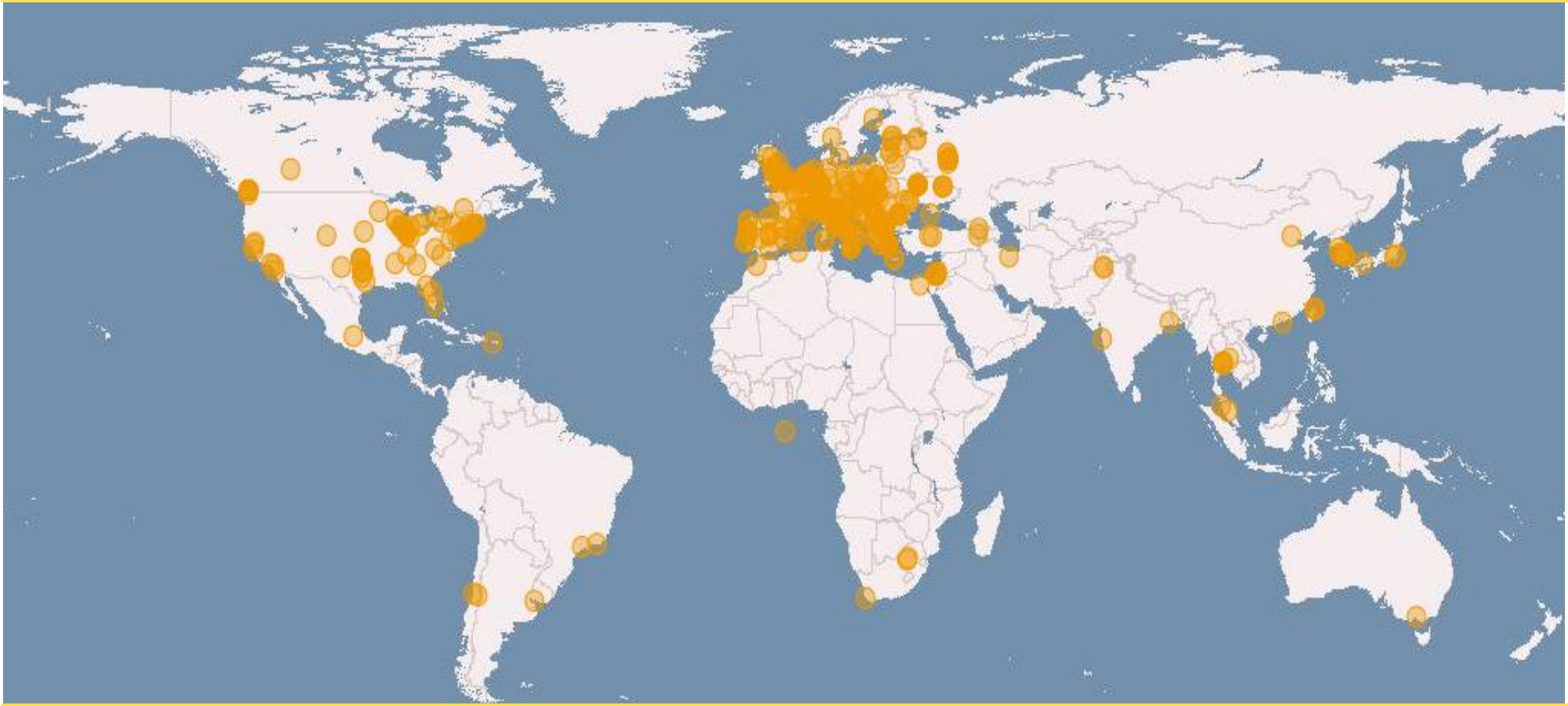
★ **Microcosm: Creating the Future Scinet and of Networks for Science**

Next Generation Computing and Networking for LHC and HEP Summary

- *To meet the challenges of globally distributed Exascale data and computation faced by the LHC and other major science programs*
- *New approaches are required: Deeply programmable software driven networked systems. We have just scratched the surface*
  - *A new "Consistent Operations" paradigm:* SDN-driven goal-oriented end-to-end operations, founded on
    - *Stable, resilient high throughput flows (e.g. FDT)*
    - *Controls at the network edges, and in the core*
    - *Real-time dynamic, adaptive operations among the sites and networks*
    - *Increasing negotiation, adaptation, with built-in intelligence*
    - *Coordination among VO and Network Orchestrators*
- Bringing Exascale HPC and Web-scale cloud facilities, into the data intensive ecosystems of global science programs
  - Petabyte transactions a driver of future network and server technology generations; Tbit/sec system demonstrators along the way
- We require a systematic approach, forward looking R&D program

ICFA Standing Committee on Interregional Connectivity (SCIC)

Global Networks for HEP *http://icfa-scic.web.cern.ch*



Closing the Digital Divide

Work with Alberto Santoro

SCIC Work Areas

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

SCIC in 2002-18:
Focus on the Digital Divide

http://cern.ch/icfa-scic

A Worldview of *Networks for and from HEP*

2017 Presentation: "Networking for HEP"

Updates on the Digital Divide, World Network Status, Transformative Trends in the Global Internet

- **32 Annexes for 2016-17 [22 New]: A World Network Overview**
  Status and Plans of International, Nat'l & Regional Networks, HEP Labs, and Advanced Network Projects
- ***2016 Monitoring Working Group Report*** **[S. McKee, R. Cottrell, et al]: Quantifying the Digital Divide: PingER Data from worldwide monitor set PerSONAR and WLCG Monitoring Efforts**
- **Also See: http://internetlivestats.com: Worldwide Internet Use**
- ***GEANT (formerly TERENA) Compendia*** **(https://compendium.geant.org/compendia): R&E Networks in Europe**
- **Telegeography.com; Interactive Submarine Cable Map: http://submarinecablemap.com**

ICFA SCIC

- We are continuing (since 2002) our work in many countries <u>to Close the Digital Divide</u>
  - Both in the physics community and in general
- To make physicists from all world regions full partners in the scientific discoveries
- We have worked in partnership with advanced networks, many agencies and HEP groups to help:
  - Brazil (RNP, ANSP), Asia Pacific (APAN), Mexico (CUDI)
  - AmLight (FIU): US – Latin America
  - GLORIAD Ring Around the Earth, Including to Russia, China, Middle East and India
- But we are leaving other countries and regions behind, for example: Africa, Most of the Middle East, South and SE Asia
  - *Work to improve the situation in Brazil and the rest of Latin America is progressing: Since 2004*
  - *Professor Santoro in particular has had a key role*
  - A great deal of work remains

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

<u>Theme:</u> 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

NEWS:

Bulletin: ONE TWO

WELCOME BULLETIN

General Information

Registration

Travel Information

Hotel Registration

Tutorials

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

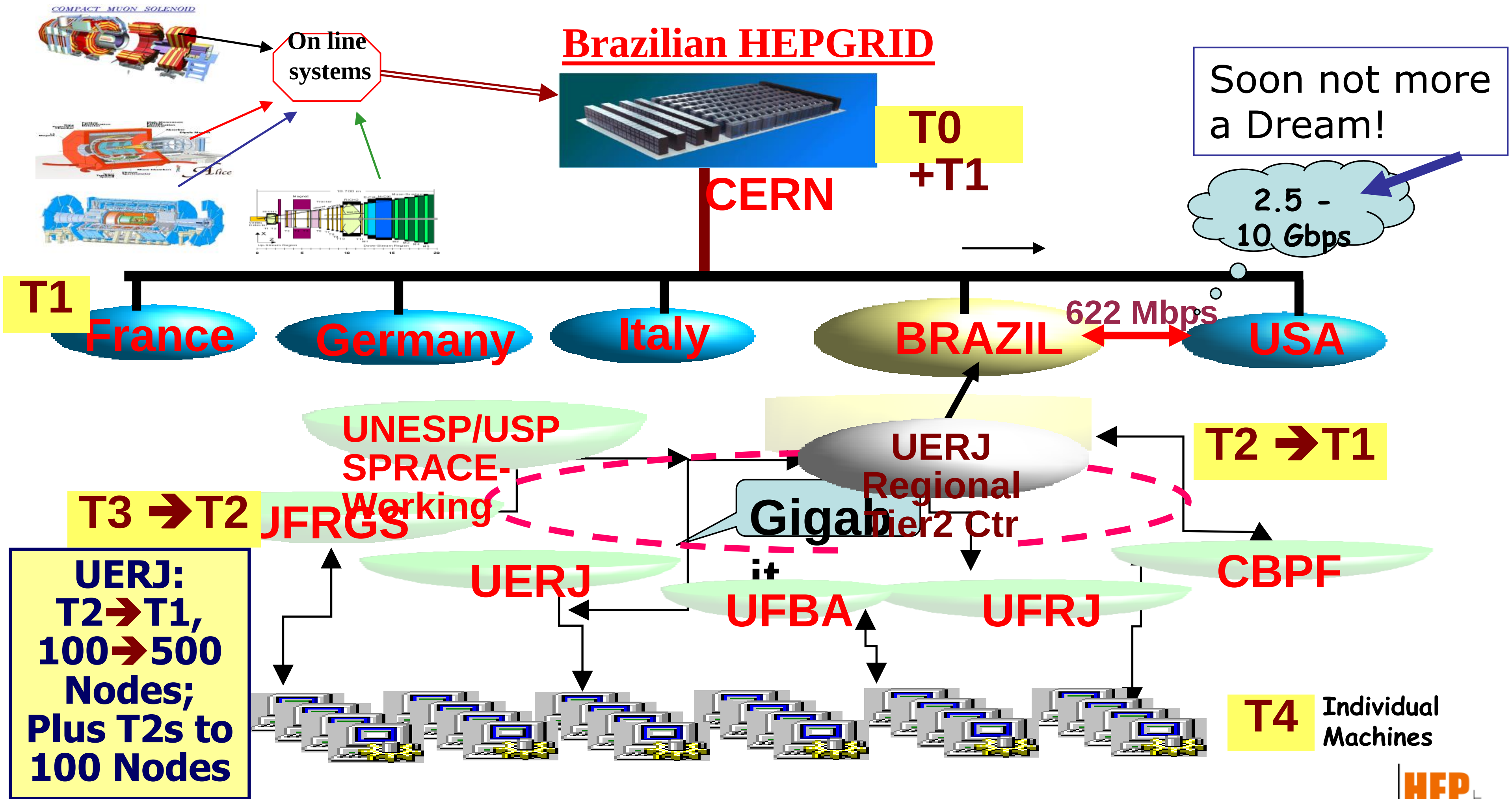
SPONSORS

CLAF CNPQ FAPERJ UERJ

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

The purpose of HEPGRID-CMS/BRAZIL is to become

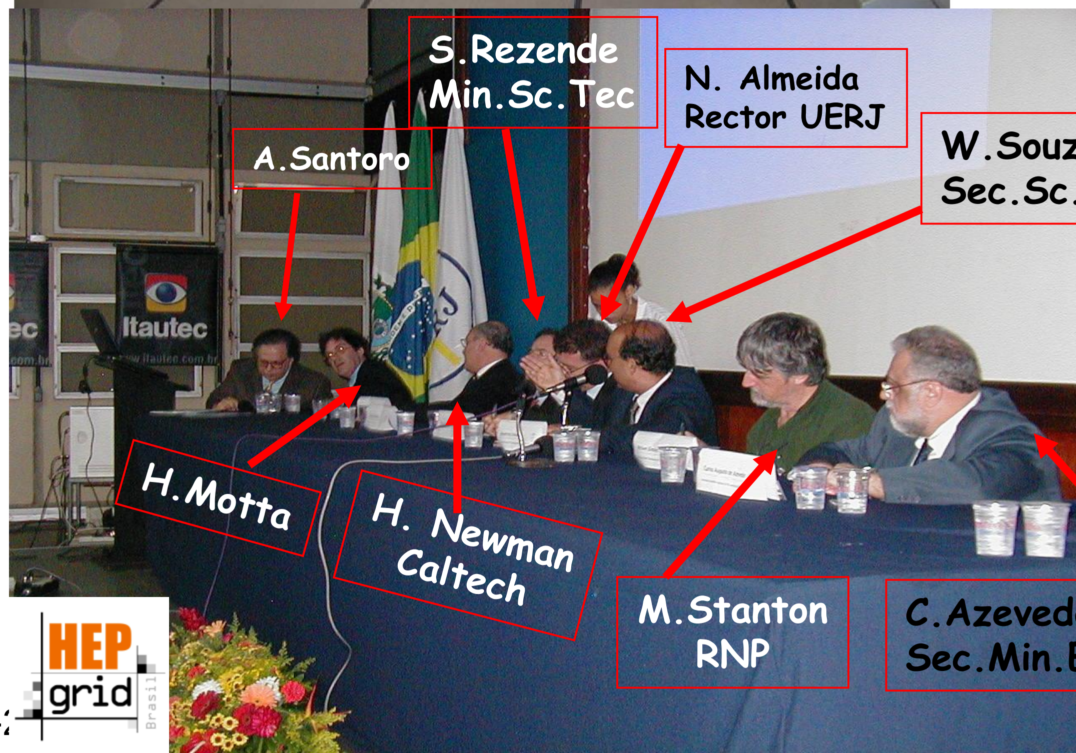
- At Regional Level, Federate with CBPF, UFRJ, UFRGS, UFBA, UERJ & UNESP
- At International Level, Federate with Caltech, T1-FNAL, GRID3/OSG...
- Strong cooperation with CALTECH



T2-HEPGRID TEAM

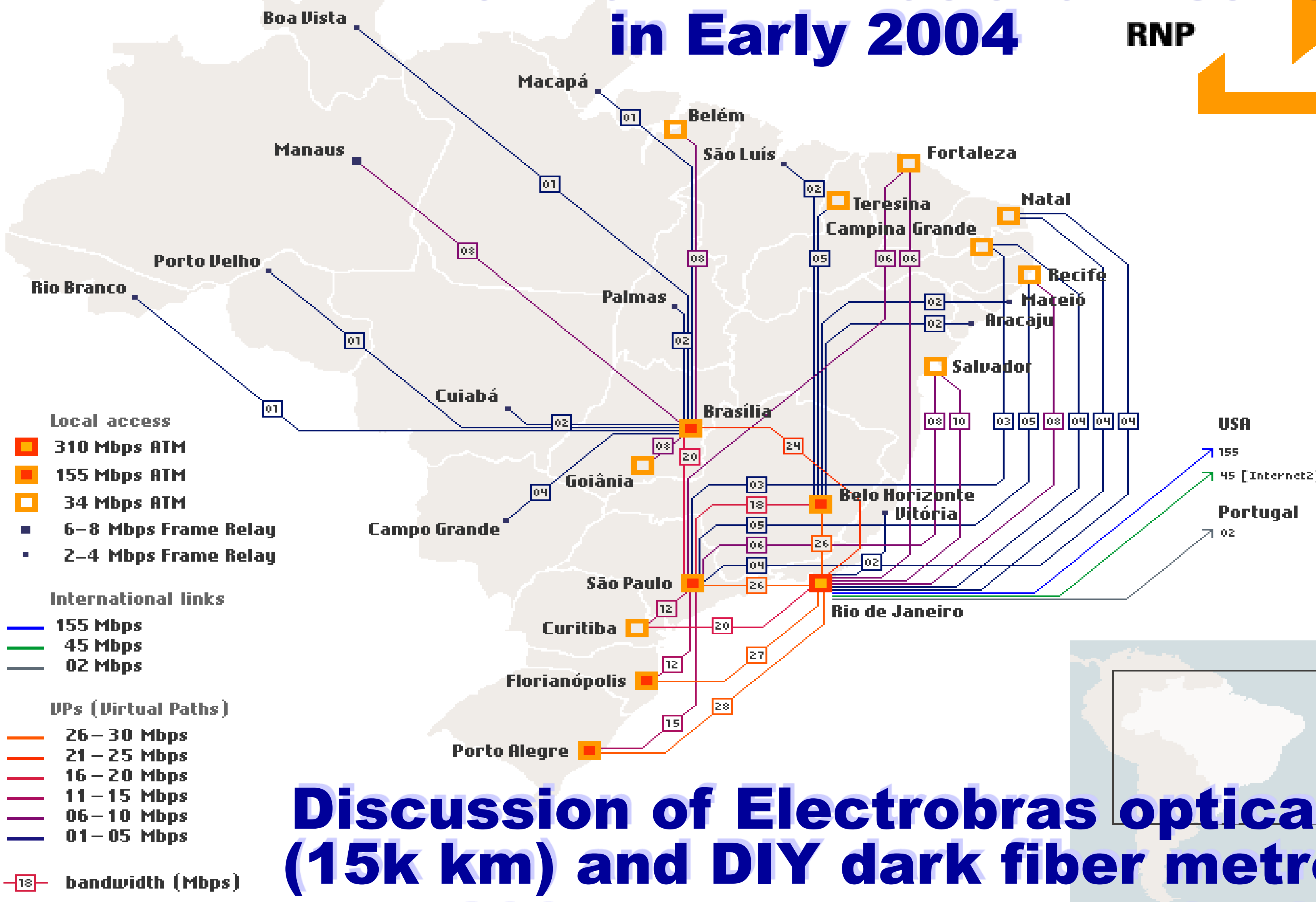
December 20, 2004
Inauguration of the HEPGrid Tier2 at UERJ In Rio


Harvey Newman




Alberto Santoro

RNP: Brazilian RNP National Network in Early 2004

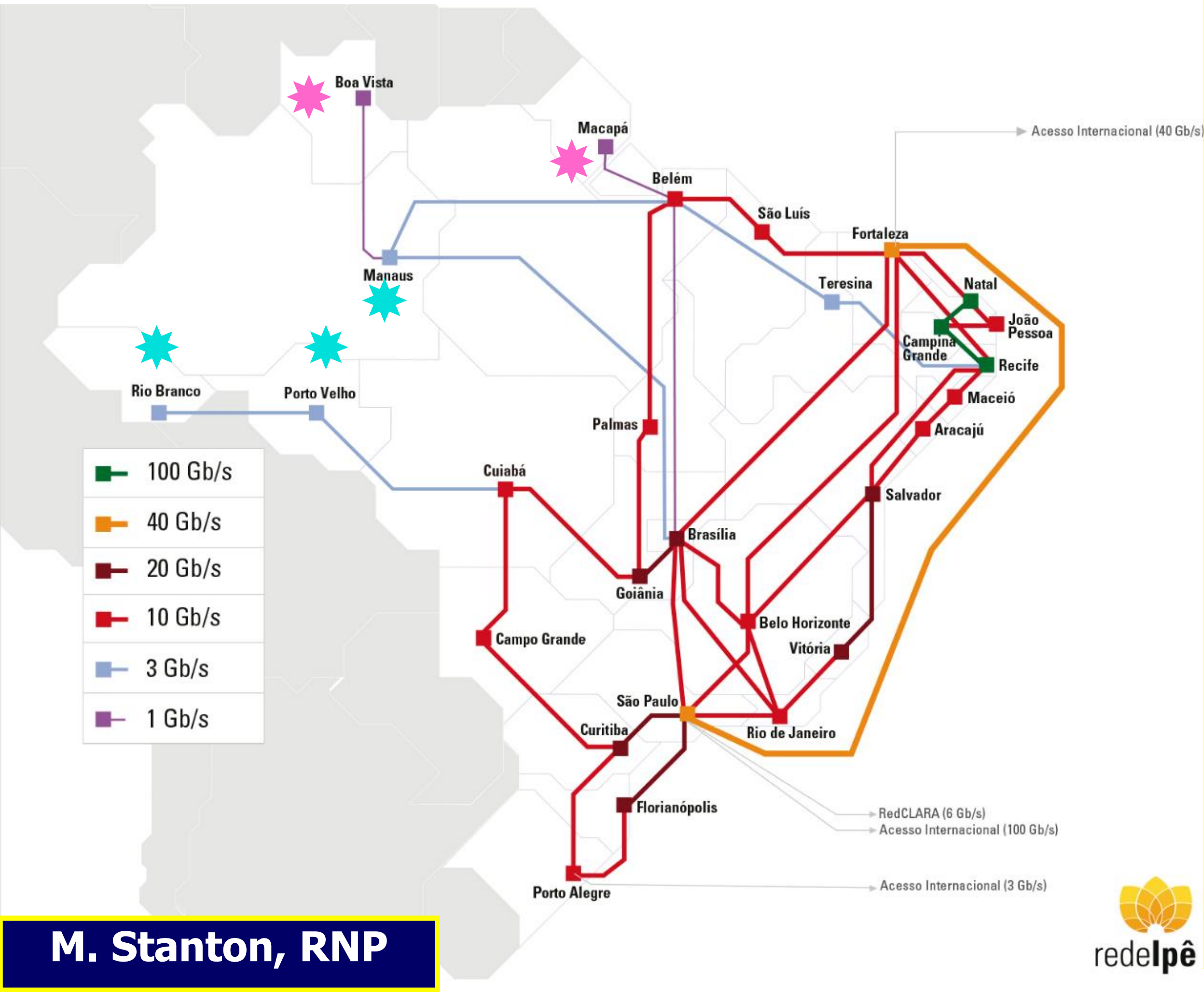


Discussion of Electrobras optical fiber (15k km) and DIY dark fiber metro nets

Then 2005: Reverse Auction for 2.5G + 10G over 4000 km Footprint

Brazil: RNP Phase 6 Backbone: 601 Gbps Aggregate Capacity; 149 Gbps Int'l

2018

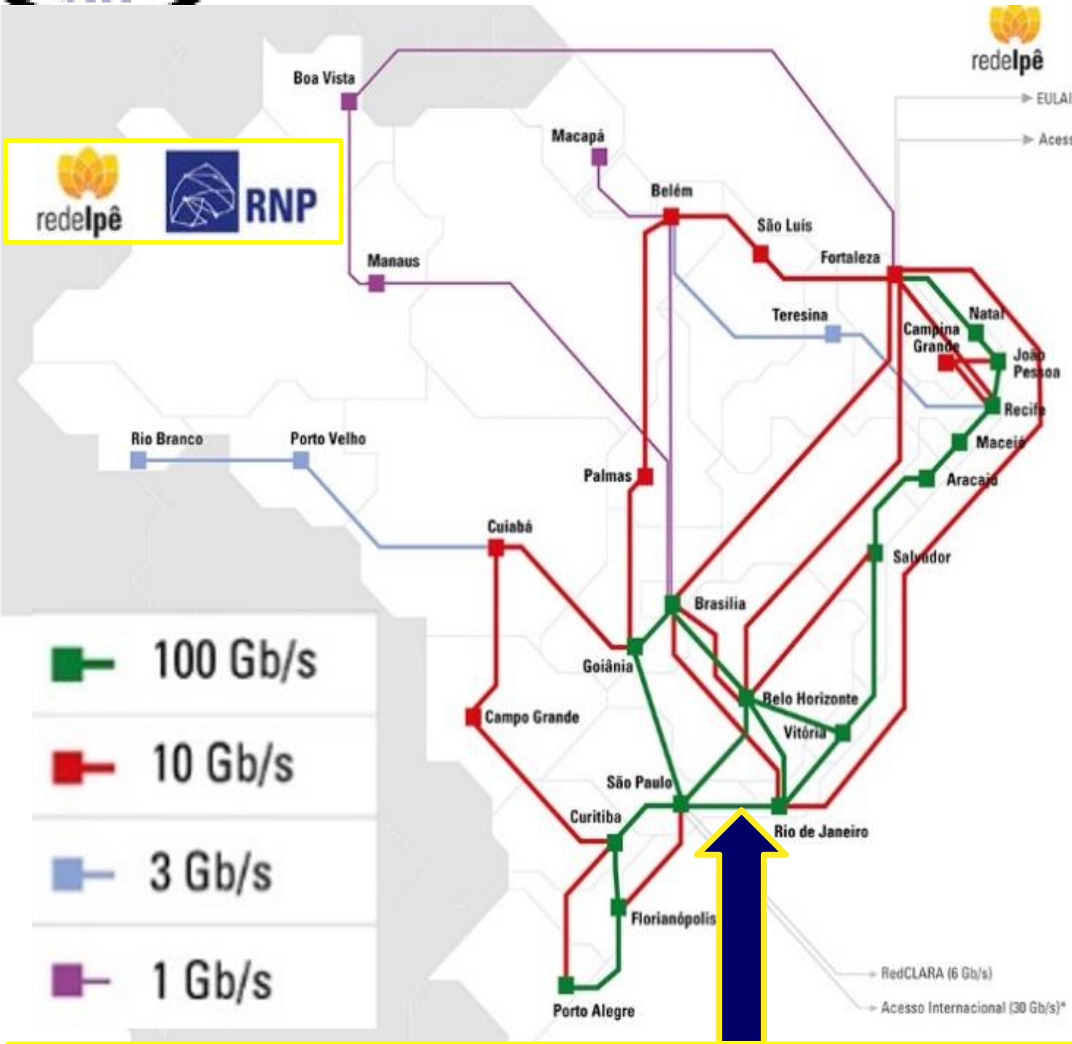


- First 100G connections in the Core in 2018
- 10G + some 2 X 10G links in the Core
- Connecting all the State Capitals, each with DIY dark fiber metro networks
- 1 to 3G Link across the Amazon to Manuas +1G to the NW capitals [✸]
- 3G Links to the West capitals [✸]
- First 100G Int'l link arrived already in 2016

M. Stanton, RNP

ICFA SCIC

RNP Phase 7 Backbone with major 100G Core Planned by 2019



Phase 7 RNP Backbone with 100G Core planned by 2019

- Requirement to support 100G waves started in 2017
- By 2019 100G central rings and a 4000 km 100G backbone planned along the eastern coast
- RNP is acquiring long-term rights to an extensive optical fiber infrastructure for the 100G transition

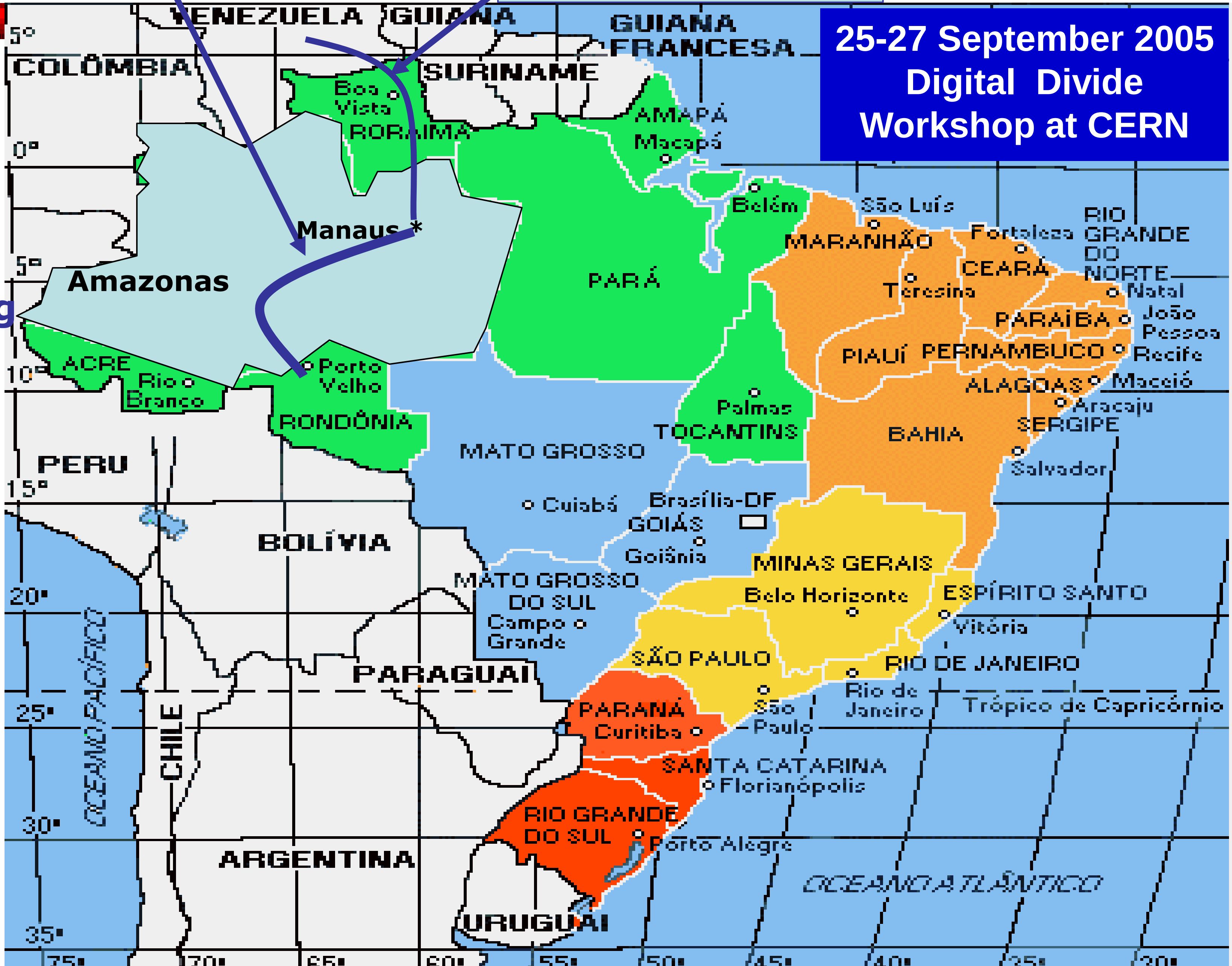
Suggestion#2 for Fiber Installation

Suggestion#1 for Fiber Installation

25-27 September 2005 Digital Divide Workshop at CERN

RNP + LOCAL Network are paying Attention to this Region Also

In 2005



Optical Fiber Through the Amazon: Porto Velho-Manaus

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

+ Aerial Cable Across the Amazon

CD

Presented at LISHEP 2011

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

M. Stanton
RNP

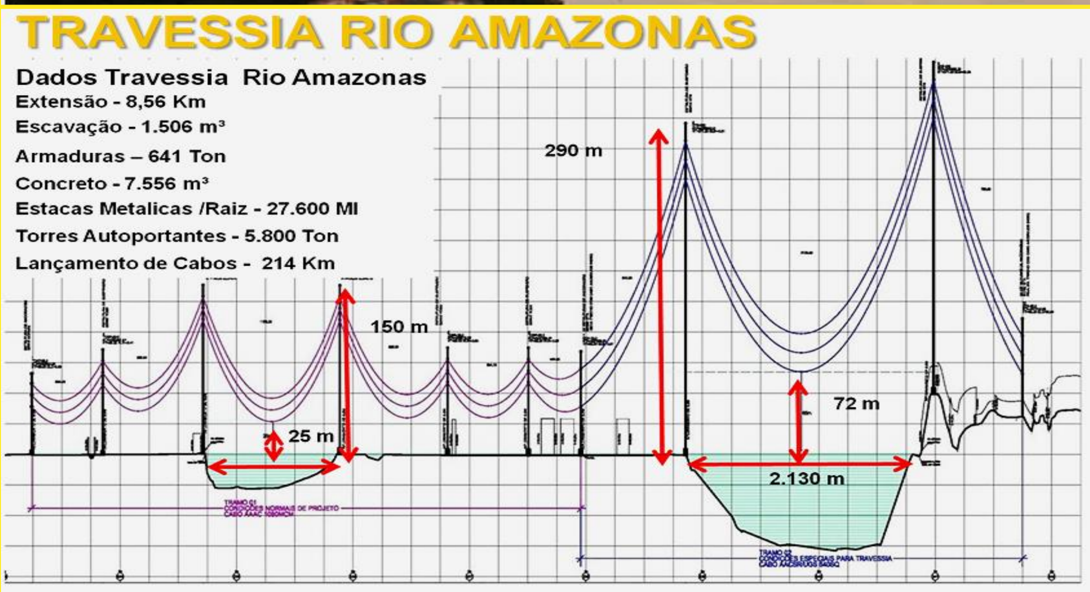


Figure N+3: Details of the Amazon crossing at Jurupari.

- 2nd fiber across the Amazon reaches the northern capital cities **Macapa and Manaus**
  - **Brings competition** to the 1st terrestrial link to Manaus
  - RNP achieved **multi-Gbps access to Manaus and the West Capitals soon after**

RNP and the Brazilian Army: Amazonia Conectada Project

http://www.amazoniaconectada.eb.mil.br/eng/

7000 km of Data Highways (Infovias) planned along the Negro, Solimoes, Jurua, Purus and Madeira Rivers

M. Stanton RNP

Introduced at LISHEP 2015 in Manaus

Purple lines are proposed subfluvial fiber

2016-17

700 km Solimões R: Manaus – Tefé

100 km Negro River: Manaus – Novo Airão

Next goal: Tefé – Tabatinga

Peru and the FITEL plan to connect Yurimaguas to Iquitos

400+ km subfluvial cable connection to be tendered along the Huallaga and Marañon Rivers between Yurimagauas and Nauta

Terrestrial FO already along Nauta-Iquitos highway

Future possibility: interconnect Brazilian and Peruvian projects between Iquitos and Tabatinga to provide a new broadband link from the Atlantic to the Pacific

M. Stanton RNP

Purple lines are proposed subfluvial fiber

2018-19

Yurimaguas

is on the Peruvian terrestrial National Backbone FO Network (RDNFO)

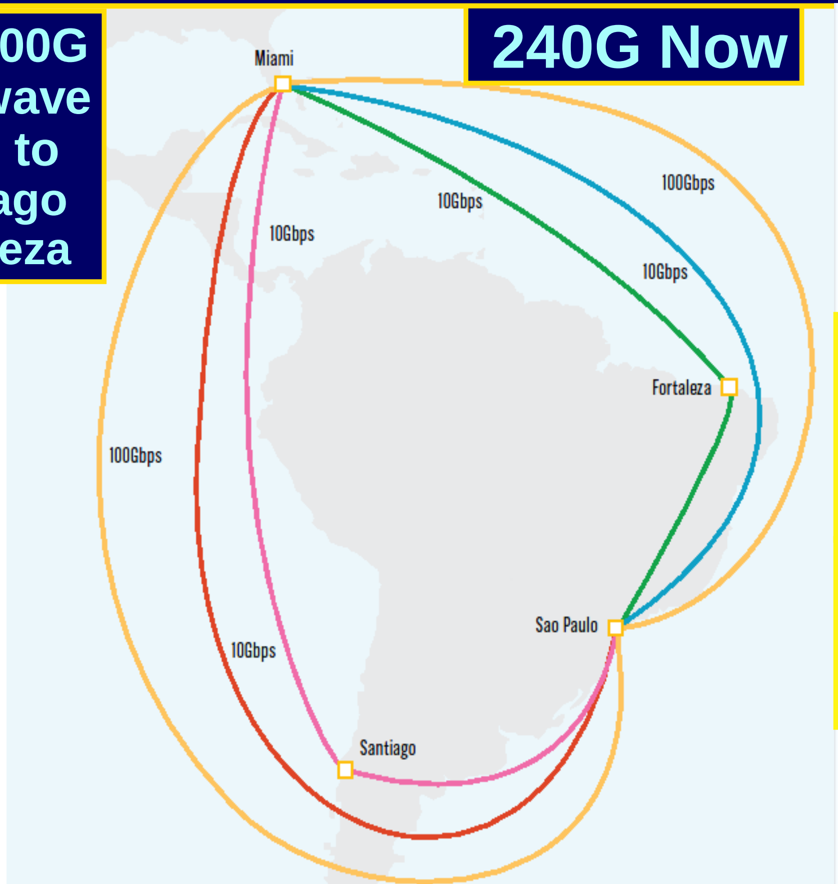
being built by FITEL to support digital inclusion

AmLight Express and Protect

US-Latin America Amlight Backbone Plan: 240G to 680G+

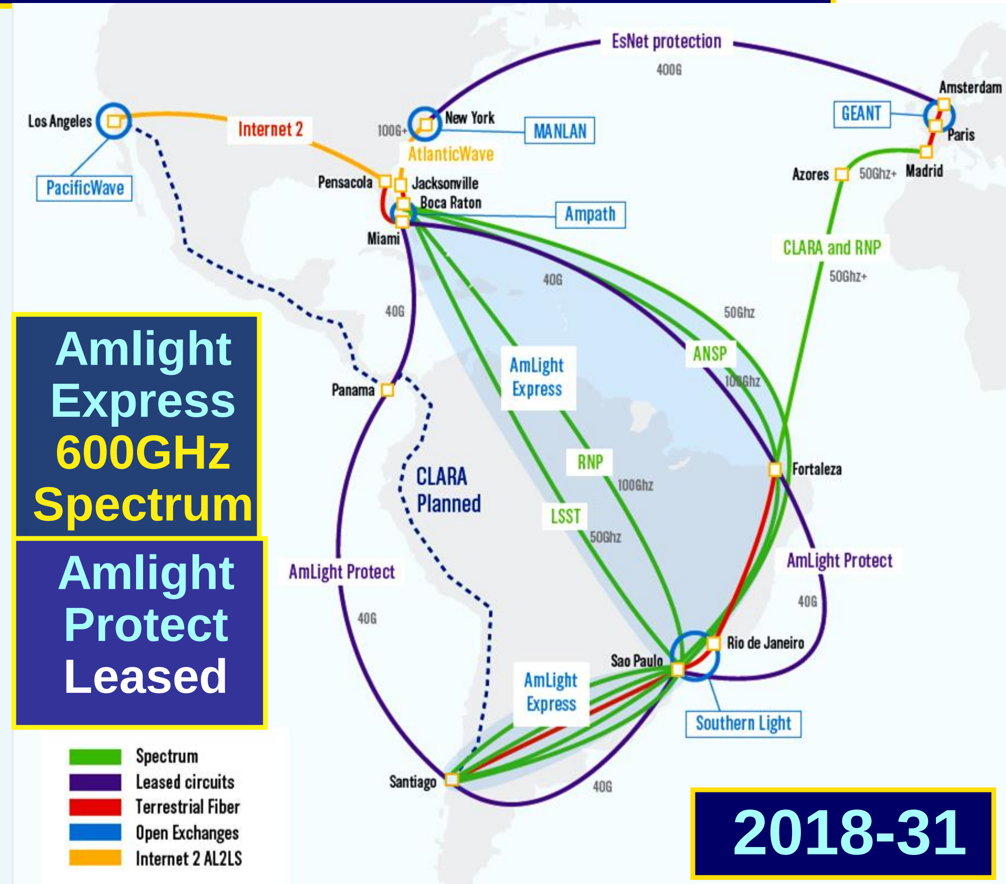
Next 100G Openwave 100G to Santiago Fortaleza

240G Now



Amlight Express 600GHz Spectrum

Amlight Protect Leased



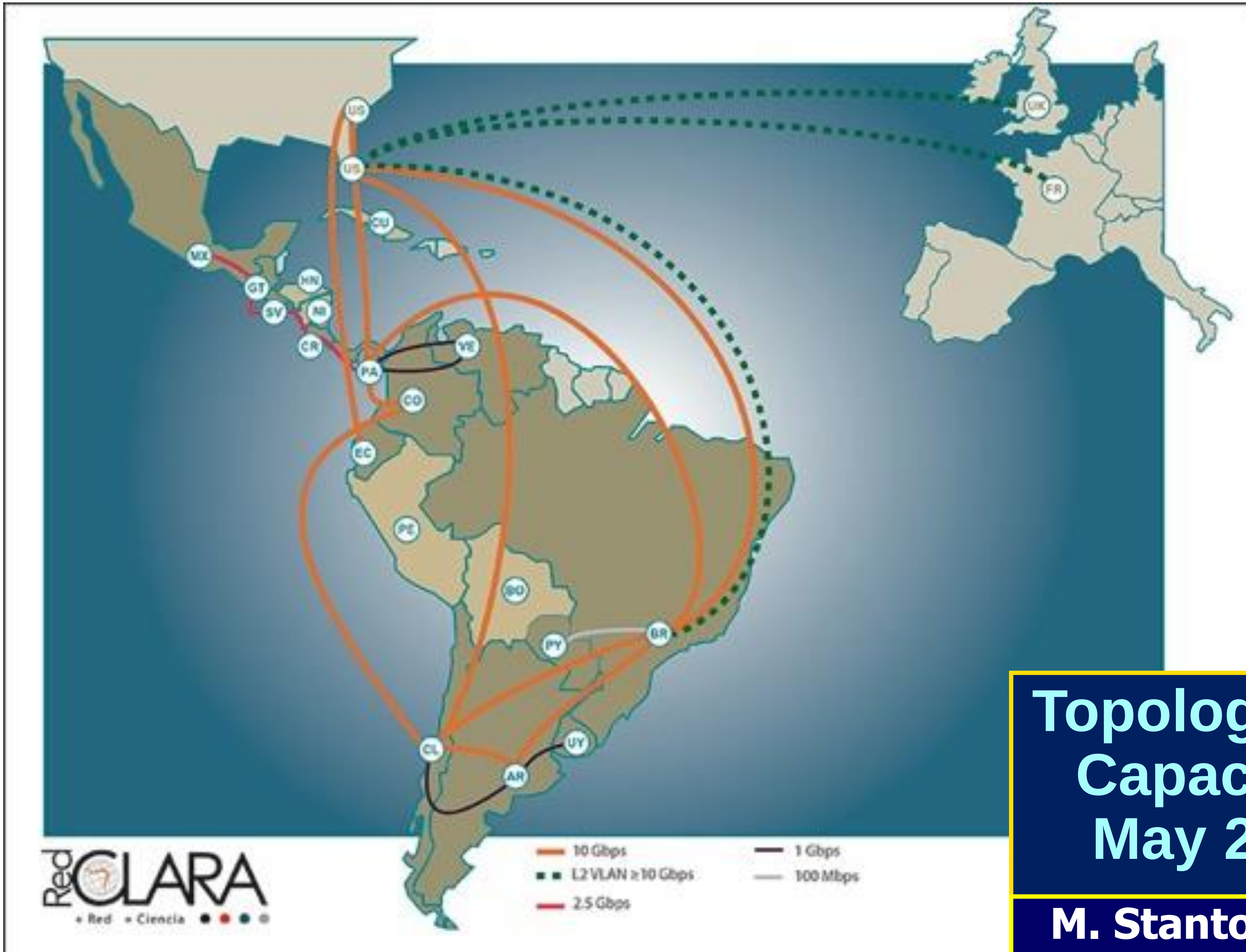
2018-31

NSF support for *AmLight* is part of a scalable rational architecture, designed to support the U.S.-WH R&E community and the evolving nature of discovery and scholarship.

Supporting LSST, LHC, WH R&E Networking

NSF Award # ACI-0963053 , 2010-2017

RedCLARA: Interconnecting Latin American NRENs



Topology and Capacities May 2017

M. Stanton, RNP

RedCLARA: Extra-regional Connectivity to Participating Latin American Networks

Marco Teixeira (RNP)

Country	NREN	Link Access Bandwidth	External Bandwidth
Argentina	InnovaRed	10 Gbps	500 Mbps
Brasil	RNP	10 Gbps	4 Gbps
Chile	REUNA	10 Gbps	1.5 Gbps
Colômbia	RENATA	10 Gbps	2.0 Gbps
Costa Rica	CRNET	2 Gbps	500 Mbps
Ecuador	CEDIA	10 Gbps	300 Mbps
El Salvador	RAICES	1 Gbps	125 Mbps
Guatemala	RAGIE	125 Mbps	125 Mbps
México	CUDI	300 Mbps	200 Mbps
Uruguay	RAU	500 Mbps	155 Mbps
Venezuela	REACCIUN	1 Gbps	100 Mbps

RedCLARA: Low External Bandwidth Issue

A New Generation of Cables with 100G Channels to Brazil in 2016-20

M. Stanton, RNP

September 2018

BELLA: RedCLARA and GEANT Program

Linking Latin American NRENs to a BR-EU Cable

Country	Brazil	Argentina	Chile	Peru	Ecuador	Colombia	TOTAL
Route length (km)	6223	2500	2000	2594	1330	1803	16450

M.Stanton, RNP

EllaLink Projected Landing points in Fortaleza, São Paulo

Planned Access Network to South America

Bella-S: acquire 25 yr IRU for 45 spectrum slots on new cable (EllaLink) between BR and EU Contract signed 2018.

Bella-T: acquire IRU for FO spectrum in each participating country **to support elastic optical network shared between NRENs and RedClara.**

2018: partially built in **Brazil, Chile, Colombia, Ecuador.** Tenders in course.

Bella program supported by EC via Géant, and by S. American NRENS via RedClara

III – A Visionary Conclusion

[by *Alberto Santoro in 2005*]

- Networks: From RNP: We will go to 10 Gbps.
- Got Certifications
- Other Brazilian regions, Very far of the big Brazilian cities: Amazonas State
- Infrastructure of the University (UERJ) is being upgraded with Optical Fibers to expand good links in the future.
- All that demonstrate that RNP and Regional networks with exceptions, are taken seriously Digital Divide.
- Please, we have to understand, that, there are a lot of local effort! But,

IT IS ALSO A CONSEQUENCE OF OUR MEETINGS IN RIO, WITH MANY AUTHORITIES WITH RNP!!! THANKS!

Closing the Digital Divide

The pioneering work of Alberto Santoro

Brazil as a Global Model Example

Backup Slides Follow

The Future of Big Data Circa 2025:
Astronomical or Genomical ? By the Numbers

PLoS Biol 13(7): e1002195. doi:10.1371/journal.pbio.1002195

Domains of Big Data in 2025. In each, the projected annual CPU and storage needs are presented, across the data lifecycle

Basis: 0.1 to 2B Human Genomes, replicated 30Xs;
+ Representative Samples of 2.5M Other Species' Genomes

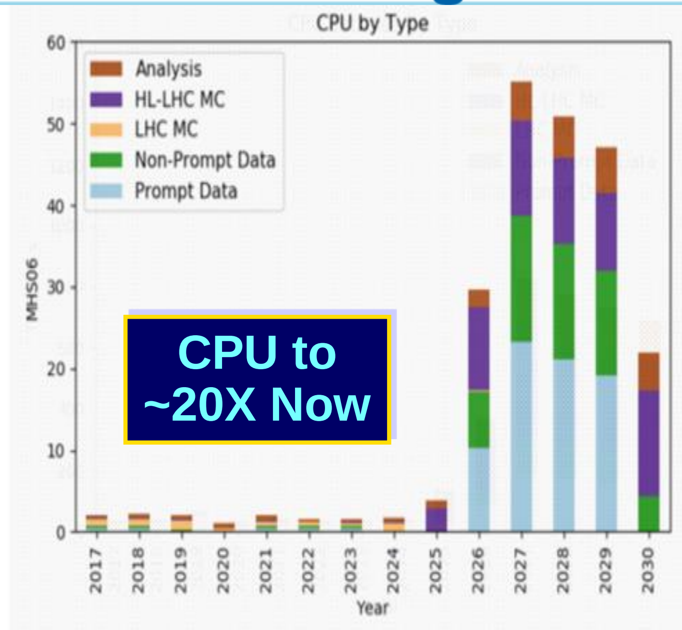
Data Phase	SKA	Twitter	YOU TUBE	GENOMICS	HL LHC
Acquisition	25 ZB/Yr	0.5–15 billion tweets/year	500–900 million hours/year	1 Zetta-bases/Yr	
Storage	1.5 EB/Yr	1–17 PB/year	1–2 EB/year	*2-40 EB/Yr*	2-10 EB/Yr
Analysis	In situ data Reduction	Topic and sentiment mining	Limited requirements		
	Real-time processing	Metadata analysis		Variant Calling 2×10^{12} CPU-h	
	Massive Volumes			All-pairs genome alignment 10^{16} CPU-h	0.065 to 0.2 X 10^{12} CPU Hrs
Distribution	DAQ 600 TB/s	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many at 10 MBps Fewer at 10 TB/sec	DAQ to 10 TB/s Offline ~0.1 TB/s

(1) Genomics Needs Realtime Filtering/Compression Before a Meaningful Comparison Can Be Made (2) New Knowledge is Transforming the Problem

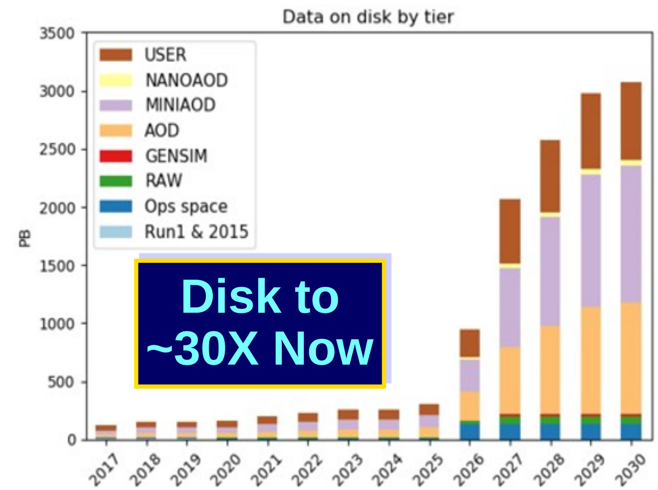
Facts of Life: Towards the Next Generation Computing Model (CMS Example)

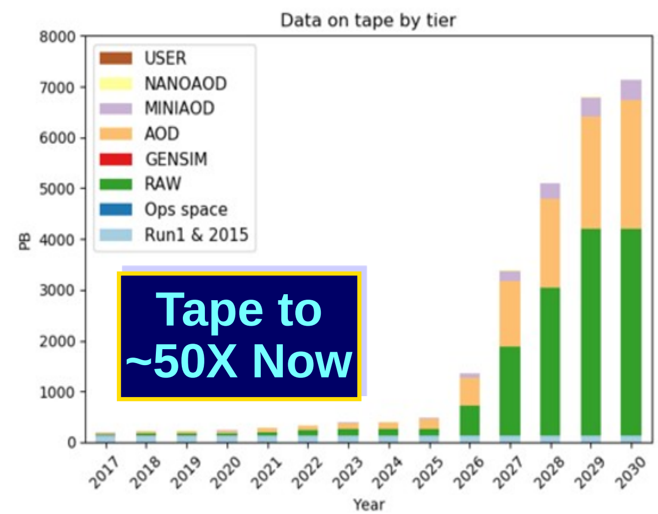
"Naïve" Extrapolations: Daunting!

- HL-LHC scales for CMS computing
  - Exa-byte scale disk and tape storage (x50 w/r to now)



  - CPU needs 5M cores (x20 w/r to now)





  - transfer of exa-byte-sized data samples across the Atlantic at 250-500 Gbps (ESnet now has allocated 40Gbps transatlantic for the LHC)
- These estimates got DOE's attention…

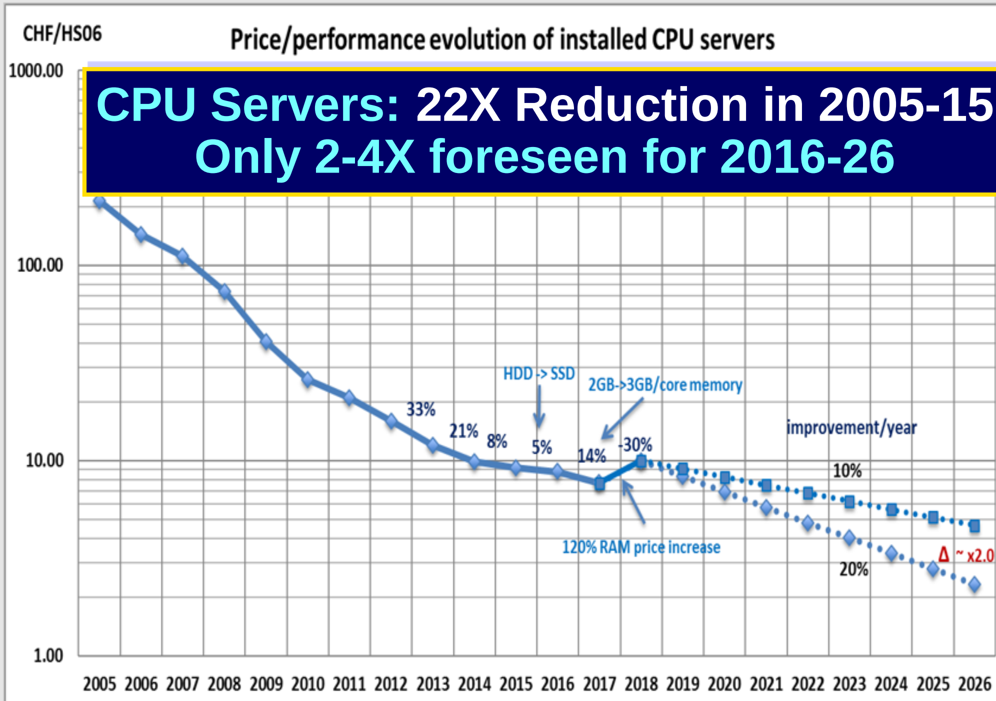
Towards the Next LHC Computing Model:
Techology eveolution alone is not the solution

- **Market trends indicate slow price/performance improvements in both computing and data storage servers: (see B. Panzer, https://twiki.cern.ch/twiki/bin/view/Main/TechMarketPresentations)**
  - **~ -15% per year for compute; ~ -22% per year for storage servers**
- **Projected shortfalls assuming constant budgets remain high: E.g. If we get a 2X speedup from code improvements, and 1.5X reduction in storage from use of the NanoAOD, shortfalls would be 2.3 to 4X in CPU and ~3X for the storage by 2026**
- **There are no technology breakthroughs or competitive market forces on the horizon that will markedly change this**
  - **Disk storage evolution are accounted for**
  - **Intel dominance of the processor market continues: little competitive pressure. AMD response is not compelling**
- **Note: Tape price evolution (-30% per year) may continue, but increased reliance on tape has its own issues of "inertia"**
- ***http://www.npd.no/Global/Norsk/3-Publikasjoner/Presentasjoner/24-august-2017/IBM_Ed%20Childers_IBM%20Tape%20roadmap%20and%20strategy.pdf***

Technology Projections – CPU and Data Servers

See https://twiki.cern.ch/twiki/bin/view/Main/TechMarketPresentations

CPU Servers: 22X Reduction in 2005-15
Only 2-4X foreseen for 2016-26



Based on CERN procurements during the last years

Current assumption:

Future CPU server price/performance improvement: 15%

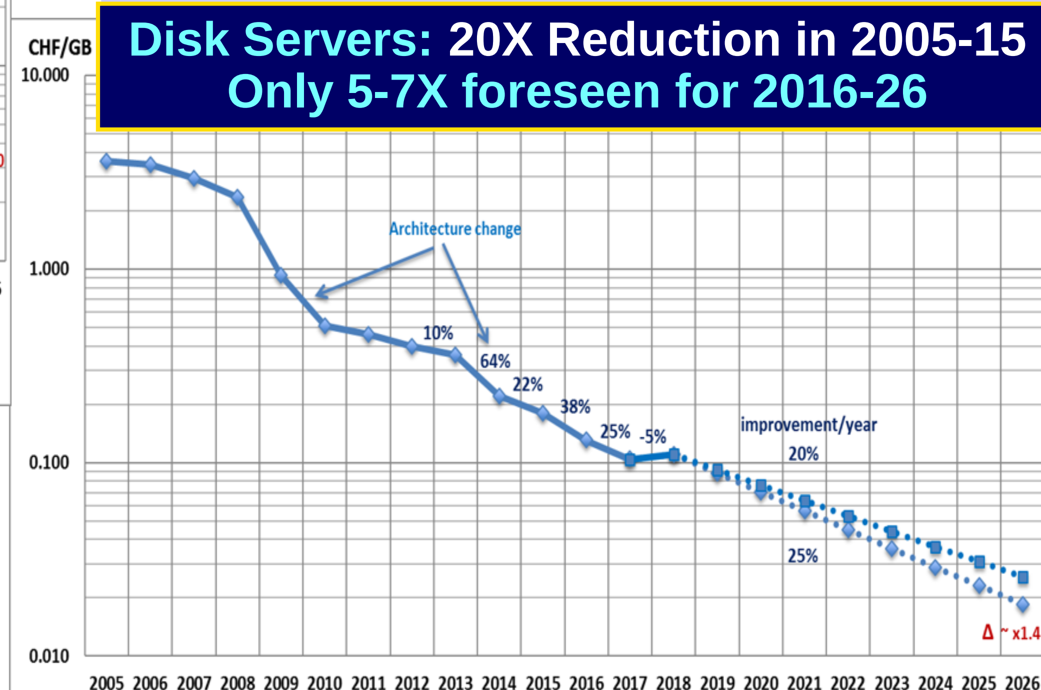
Future Disk server price/space improvement: 20%

26. March 2018 Bernd Panzer-St

Server Cost Evolution

Disk server: very hard to estimate the REAL costs of HDDs e.g. there are 70 different 6TB HDD models in the market with a price difference of a factor 2.5. At CERN we saw price difference of a factor >2 between low street prices and purchase prices. more variations between 6 TB and 8TB disks

Disk Servers: 20X Reduction in 2005-15
Only 5-7X foreseen for 2016-26



ESnet

Towards ESnet6

C12 J. OPT. COMMUN. NETW./VOL. 9, NO. 4/APRIL 2017

Beyond 100 Gb/s: Capacity, Flexibility, and Network Optimization

Kim Roberts, Qunbi Zhuge, Inder Monga, Sebastien Gareau, and Charles Laperle

Labs. Significant growth in scientific data, the number of data sets, and the evolution of scientific workflows has caused traffic to grow in excess of 70% a year. It is estimated that by 2025, this long-haul core network will have over 40% of its light paths carrying traffic in excess of 10 Tb/s and 10% with traffic in excess of 20 Tb/s. Therefore, to fulfill its mandate, ESnet must deploy state-of-the-art transmission technology that will enable the minimization of the transport cost and, if possible, allow the use of the installed capacity with the utmost flexibility. To satisfy these requirements, intensive research on optical communications has been conducted to reduce the cost per bit and increase the spectral efficiency of backbone networks [2].

Bottom Line: National and continental R&E networks ***Might*** keep pace with demand, for the next several years.

The outlook for intercontinental links is more uncertain.

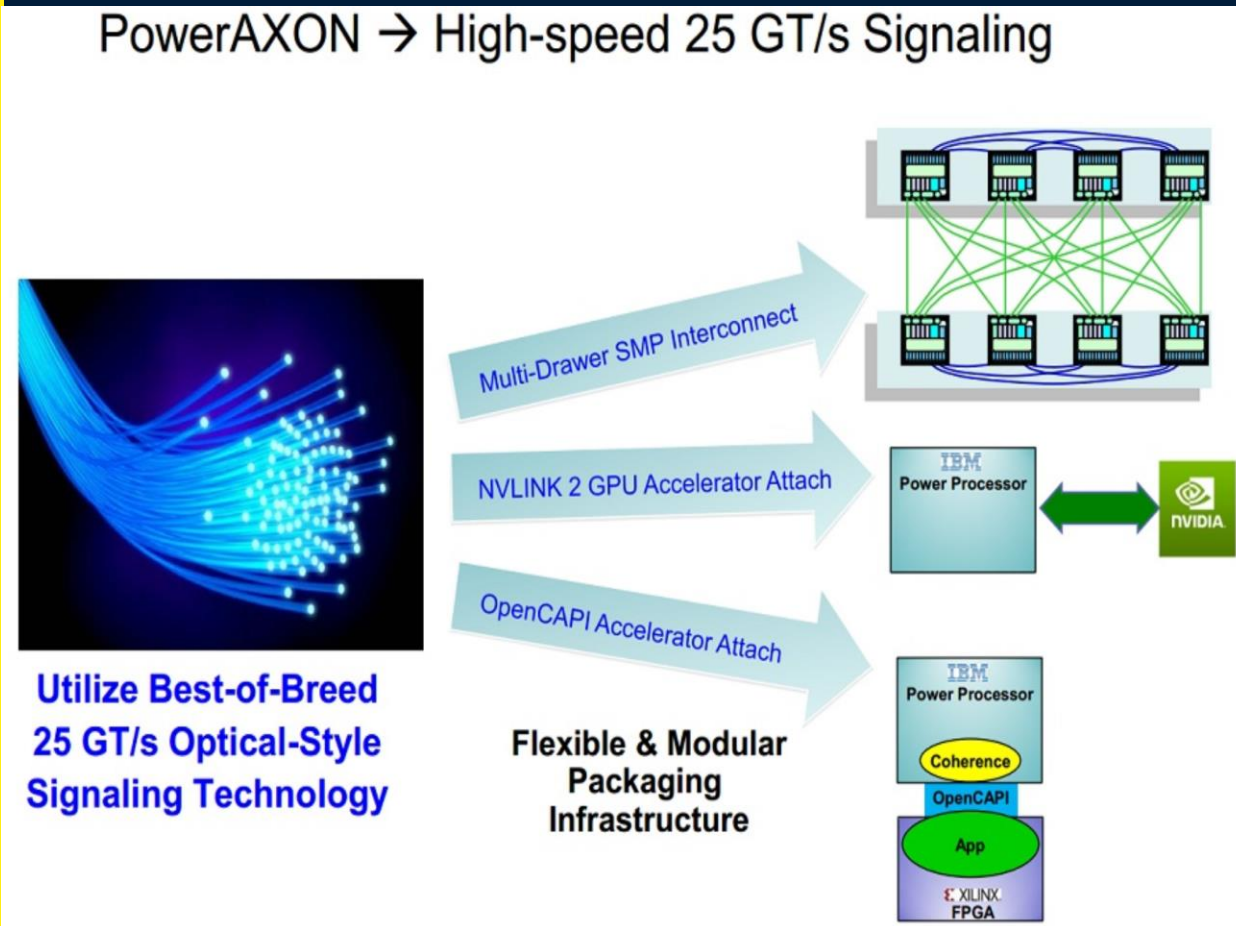
Upcoming Computing Architecture Trends
and How They May Affect Us (I)

- *The failure to keep up with Moore's Law and the shift towards AI-oriented applications has some major vendors (e.g. IBM, HP, Xilinx, Kingston) seeking a better way:*
  - *Disaggregated or "Heterogeneous" Computing: Interconnect CPU, storage, GPUs and FPGAs to form a unified "fabric" via higher speed buses and links.*
  - *Overcome basic limits of current architectures: Loading the main CPU; PCIe Gen3 BW and Limited Lanes to the CPUs*
  - *Enable higher throughput parallel processing: e.g. with multiple GPUs or FPGAs in compact ~mass-market systems*
- **Example Vision: IBM Power 9 and 10 Futures Presented at HotChips 30**
- **https://www.hpcwire.com/2018/08/23/ibm-at-hot-chips-whats-next-for-power/**
- **System Throughputs: 210 Gbytes/sec in 2018 to 435+ Gbytes/sec in 2020+**

IBM at Hot Chips 30:
Heterogeneous Computing Concept

https://www.hpcwire.com/2018/08/23/ibm-at-hot-chips-whats-next-for-power/

"With processor, memory and networking technologies all racing to fill in for an ailing Moore's law, the era of the heterogeneous datacenter is well underway, and IBM is positioning its chips to be the air traffic controller at the center of it all."



The CPU
as a Big Switch
For Flexible SMP, GPU- and FPGA-attached architectures

"AXON": A and X:
On and off-chip SMP buses;
O: OpenCapi
N: NVLink2 (Open)
25.6 GT/sec: ~5X 2400 MHz DRAM; Many IO channels

IBM Heterogeneous Computing Concept
Power 9 to Power 10 Roadmap

Power Systems

Proposed POWER Processor Technology and I/O Roadmap

IBM

Todays talk

	POWER7 Architecture		POWER8 Architecture		POWER9 Architecture			POWER10
	2010 POWER7 8 cores 45nm New Micro-Architecture New Process Technology	2012 POWER7+ 8 cores 32nm Enhanced Micro-Architecture New Process Technology	2014 POWER8 12 cores 22nm New Micro-Architecture New Process Technology	2016 POWER8 w/ NVLink 12 cores 22nm Enhanced Micro-Architecture With NVLink	2017 P9 SO 12/24 cores 14nm New Micro-Architecture Direct attach memory New Process Technology	2018 P9 SU 12/24 cores 14nm Enhanced Micro-Architecture Buffered Memory	2019 P9 w/ Adv. I/O 12/24 cores 14nm Enhanced Micro-Architecture New Memory Subsystem	2020+ P10 TBA cores New Micro-Architecture New Technology
Sustained Memory Bandwidth	65 GB/s	65 GB/s	210 GB/s	210 GB/s	150 GB/s	210 GB/s	350+ GB/s	435+ GB/s
Standard I/O Interconnect	PCIe Gen2	PCIe Gen2	PCIe Gen3	PCIe Gen3	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen4 x48	PCIe Gen5
Advanced I/O Signaling	N/A	N/A	N/A	20 GT/s 160GB/s	25 GT/s 300GB/s	25 GT/s 300GB/s	25 GT/s 300GB/s	32 & 50 GT/s
Advanced I/O Architecture	N/A	N/A	CAPI 1.0	CAPI 1.0 , NVLink 1.0	CAPI 2.0, OpenCAPI3.0, NVLink2.0	CAPI 2.0, OpenCAPI3.0, NVLink2.0	CAPI 2.0, OpenCAPI4.0, NVLink3.0	TBA

Statement of Direction, Subject to Change

Upcoming Computing Architecture Trends and How They May Affect Us (II)

- **Another Vision: Intel Acquires Altera ($ 16.7B) December 2015: Intel + Stratix 10 FPGA "Hyperflex" Architecture** https://www.intel.com/content/dam/www/programmable/us/en/pdfs/literature/wp/wp-01251-enabling-nextgen-with-3d-system-in-package.pdf

3D SiP-based transceiver tiles: maximizing scalability and flexibility

Heterogeneous 3D SiP technology decouples the transceiver tile or die from the core fabric die: the transceiver is separate and sits next to the core fabric die. Therefore, the transceiver

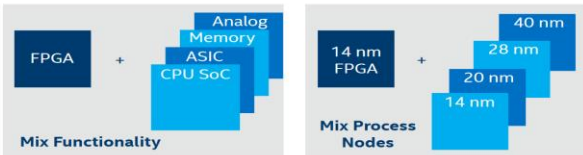
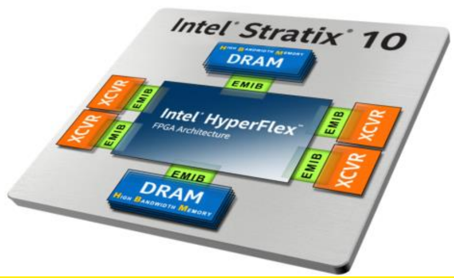
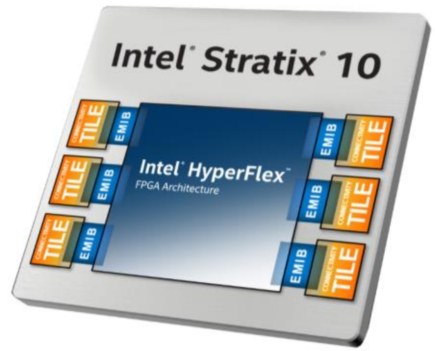


Figure 6. Heterogeneous In Package Integration with 3D SiP Technology





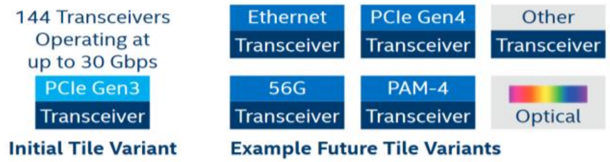


Figure 8. Enhanced Flexibility and Scalability with Separate Transceiver Tiles

The EMIB advantage

The EMIB technology offers a simpler manufacturing flow, higher performance, enhanced signal integrity, and reduced complexity. Figure 9 depicts the physical package construction. The construction heterogeneously integrates the FPGA fabric die (1) and two transceiver die (2). (The non-FPGA die can be a transceiver die, memory die, CPU die, or any other functionality.) The three die rest on a standard flip-chip ball grid array (FCBGA) package substrate (3), which connects to the underlying PCB. The routing between the die and the package balls uses standard FCBGA routing (4). This assembly is encapsulated with a standard package lid (5) to create a single package solution. The package substrate utilizes several EMIB connections (6). The EMIB enables the heterogeneous in-package integration by connecting the die using an ultra-high-density interconnect.

- **Other Interesting Developments (August 2018): Hot Chips and Flash Memory Storage Conferences. Examples: (1) NRAM (Carbon nanotubes) poised to replace DRAM ?** https://www.anandtech.com/show/13252/hot-chips-2018-nanotubes-as-dram-by-nantero-live-blog **; (2) MRAM Developer Day at FMS**

Perspective on Improvements
Aimed at Reducing the Challenge (II)

- *New machine learning methods for reconstruction and analysis*
  - *How much can we gain in scaling certain algorithms (e.g. KF tracking, shower GAN) algorithms ?*
- *Resource savings vs scope expansion: What will be balance between*
  - *Resource use reduction through more efficient algorithms versus*
  - *Enablement of more sophisticated methods and greater physics reach, but with equal or greater resource use ?*
- *What will we gain from new processor + disaggregated system architectures ?*
  - *Cost effectiveness for us compared to today's conventional mass-market computer architectures (e.g. Supermicro, Dell) is TBD*

✸ *In order to understand the impact and potential savings*

✸ *We must engage in a multipronged R&D program; along several lines*

✸ *We require a forward looking, open view in designing the Next Computing Model for the Run3 + LHC Era*

Facts of Life Beyond Data Centers, Processing and Analysis: Cultural Changes

- There are already major changes underway that will have a great impact on our designs and plans
- Between now and 2026, and then beyond
- Emergence of machine-to-machine (MTM) communication
  - To 27B devices by ~2020; pervasive before HL LHC
  - Digital Assistants, AI smartphones are the tip of the iceberg
- IOT: Expanding intelligence, autonomy, and coordination
  - Emergence of smart apartments/homes, buildings, neighborhoods and cities
- Real-time *low power* AI devices: responsive on microsecond to millsecond timescale: already on the radar for the Phase II Trigger
  - Can we use this for real-time event cleanup ? Pre-processing ?
  - [Joel Butler]: Can we reconstruct all < 2 GeV tracks in real-time ?
- Autonomous tracking of workflows: real-time remediation if workflow is delayed; managing user/system interactions ?

ICFA SCIC

770M Internet Users in China (7/18)
Rise of Mobile Social Software Services

Over 770m internet users in China

By Yu Xiaoming | chinadaily.com.cn | Updated: 2018-07-13 13:27

People use their mobile phones outside an office building in Beijing on May 24, 2018. [Photo/VCG]

China had 772 million netizens by the end of 2017, with a penetration rate of 55.8 percent, and 40.74 million more than the number in the previous year, Beijing News reported citing a report released by the Internet Society of China.

Among them 753 million surfed the internet via mobile phones, 57.34 million more than the previous year. The users averaged 27 hours per week surfing the internet, the report said.

A total of 398 million Chinese used internet livestreaming last year, and the report predicted the figure will top 500 million by 2019. However, the industry saw a slower user growth of 28.4 percent last year, and it's expected to continue to fall to 10.2 percent by 2019.

Mobile social software has become a absolute necessity for mobile phone uses, with nearly half of them using these software more than three times each day. The top three most used social apps are WeChat Moments, Qzone and Sina Weibo, with usage rate of 87.3 percent, 64.4 percent and 40.9 percent respectively.

China's online shopping market was worth 7.18 trillion yuan last year, up 32.2 percent year-on-year, according to the report.

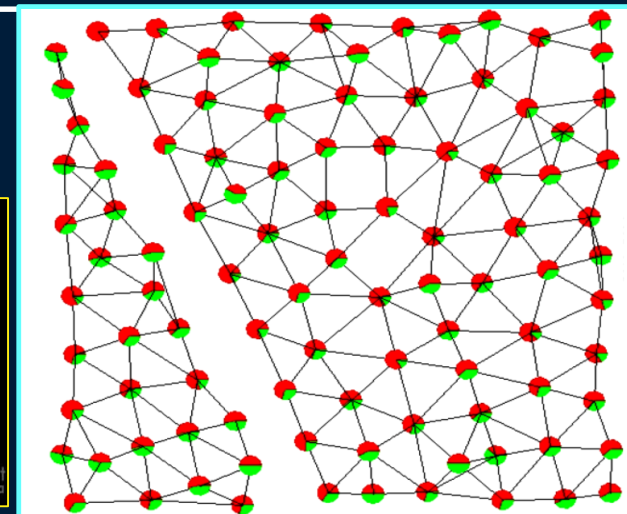
China had 531 million online payment users at the end of last year, 11.9 percent or 56.61 million more than the previous year. Among them, 527 million people use mobile online payment, 12.3 percent or 57.83 million more than the previous year.

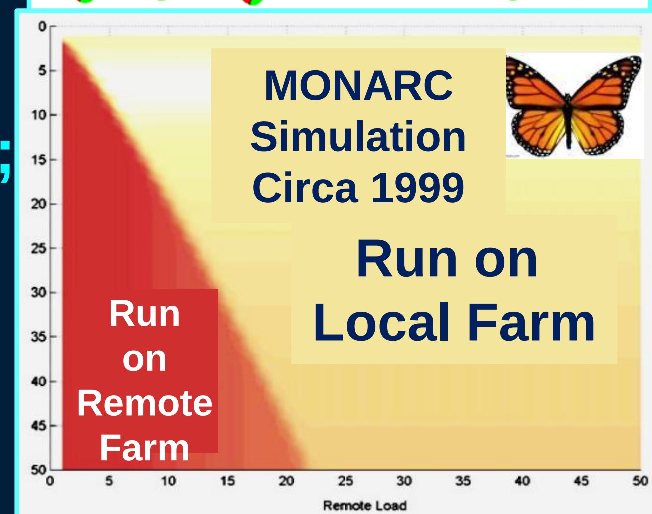
By the end of 2017, a total of 435 million Chinese used ride-hailing services, a rise of 19.2 percent compared with a year earlier. Among them, 287 million use ride-hailing services to book taxis, up 27.5 percent from a year earlier, while 236 million people use these services to book private vehicles, up 40.6 percent compared to the end of 2016.

Last year, 253 million Chinese used internet medical services, up 29.7 percent year-on-year. The report said over the next few years the compound annual growth rate is expected to reach 55.01 percent. By 2010, the market of China's mobile medical care will be worth 69.74 billion yuan, the report added.

Key Developments from the HEP Side: Machine Learning, Modeling, Game Theory

- Applying Deep Learning + Self-Organizing systems methods to optimize LHC workflow
  - Unsupervised: to extract the key variables and functions
  - Supervised: to derive optima
  - Iterative and model based: to find effective metrics and stable solutions [*]
- Complemented by modeling and simulation; game theory methods [*]
- Progressing to real-time agent-based dynamic systems
- With application to LHC Workflow





Self-organizing neural network for job scheduling in distributed systems

[*] T. Roughgarden (2005). *Selfish routing and the price of anarchy*

ECOM2X: Technology choices and speeds today

Storage Type	Random IOPS	IOPS Ratio to HDD	Serial IO	Serial IO Ratio to HDD	$/TB	$/TB Ratio
Tape	~0	~0	350 MB/sec	1.8 – 3.5	10-20	0.2 – 0.5
Hard Disk	~100-200	1	100 – 200 MB/sec	1	40-50	1
SATA SSD NVMe SSD	25k – 100k 200k-500k	125 – 1k 1k – 5k	500 MB/sec 1.5 – 3.5 GB/sec	1.5 8-35	200-300 300-600	4-6 7-12
Memory			~20 GB/sec		11k	2.2-3k

New Directions to Meet the Challenge (III)
Exascale HPC Facility Use

- *Use of Exascale facilities for our mainstream computing: Issues*
  - Recasting our codes: (a) reducing the memory footprint and (b) redoing the architecture to run highly parallel (e.g. 512 threads and up)
  - Only "efficient" use of some of these facilities (e.g. in the US) would be allowed on a large scale (e.g. make full use of GPUs)
  - Manpower to support highly data-intensive multi-user operations is very limited at these facilities
    - We need to work out an actual workflow with their staff, and design how it might work at scale
  - Network impact of data ingest and export needs to be understood, and solutions need to be implemented

LHC Computing Model
Outlook for the HL LHC

- ***Additional Perspectives:***
- ***Lothar Bauerdick,***
  "CMS Computing for HL-LHC & Exascale Initiatives"

  https://www.dropbox.com/s/71levo3qg63hjys/CMSComputingHLLHCExascaleInitves_LATBauerdickPAC071818.pdf?dl=0

- ***Liz Sexton-Kennedy, Plenary Talk at ICHEP 2018:***
  "The Future of Software and Computing for HEP: Pushing the Boundaries of the Possible"

  https://www.dropbox.com/s/ybwyb5w7q72tu24/FutureSoftwareandComputingforHEP_LizSextonKennedyICHEP2018Plenary.pdf?dl=0

- ***J. Flix, M. Schulz, A. Sciaba, HEPIX Fall 2017 (Tsukuba)***
  "Optimizing resource needs for LHC Computing: Ideas for a Common Approach"
  https://www.dropbox.com/s/g3rfsdk46wx87j5/OptimisingResourceNeedsforLHCComputing_FlixetalOctober2017.pdf?dl=0

A New Generation of Cables to Brazil with 100G Channels in 2017-20

Cable	Owners	Ready for service	Capacity	Length (km)	Landing points in Brazil	Other countries served
Monet	Google, Antel, Angola Cables, Algar Telecom	2017	64 Tb/s	10,556	Fortaleza (branch) Santos	USA (Boca Ratón, FL)
South Atlantic Cable System (SACS)	Angola Cables	2018	40 Tb/s	6,165	Fortaleza	Angola (Luanda)
Ellalink	Telebras, IslaLink	2020	48 Tb/s	9,501	Fortaleza (branch) Santos	Portugal (Sines)
Tannat	Google, Antel	2018	90 Tb/s	2,000	Santos	Uruguay (Maldonado)
Seabras-1	Seaborn Networks	2017	72 Tb/s	10,500	Fortaleza (branch) Santos	USA (New York)
South Atlantic Interlink (SAIL)	Camtel, China Unicom	2018	32 Tb/s	5,900	Fortaleza	Cameroon (Kribi)
BRUSA	Telefonica	2018		11,000	Fortaleza Rio de Janeiro	USA (Virginia Beach)

M. Stanton, RNP

Expected future international Nx 100G routes from Brazil

- Monet to Florida (2018)
- SACS to Angola (2018)
- SACS+WACS to South Africa (2019)
- New RedClara (2019)
- Ellalink to Europa (2020/1)
- probable Ellalink extention to French Guiana (and maybe Colombia)

M. Stanton RNP

Potential great use for data intensive science programs, including the ALMA (Chile) and SKA (South Africa to the Americas) telescope arrays