A Next Generation Global Network System
for the LHC and Data-Intensive Sciences

LHC | LBNF/DUNE | SKA

LHC Run3 and HL-LHC

DUNE

VRO SKA

BioInformatics

Earth Observation

Gateways to a New Era

Harvey Newman, Caltech

LISHEP2023

March 9, 2023

GNA-G Global Network Advancement Group

ICFA SCIC

California Institute of Technology 1891

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

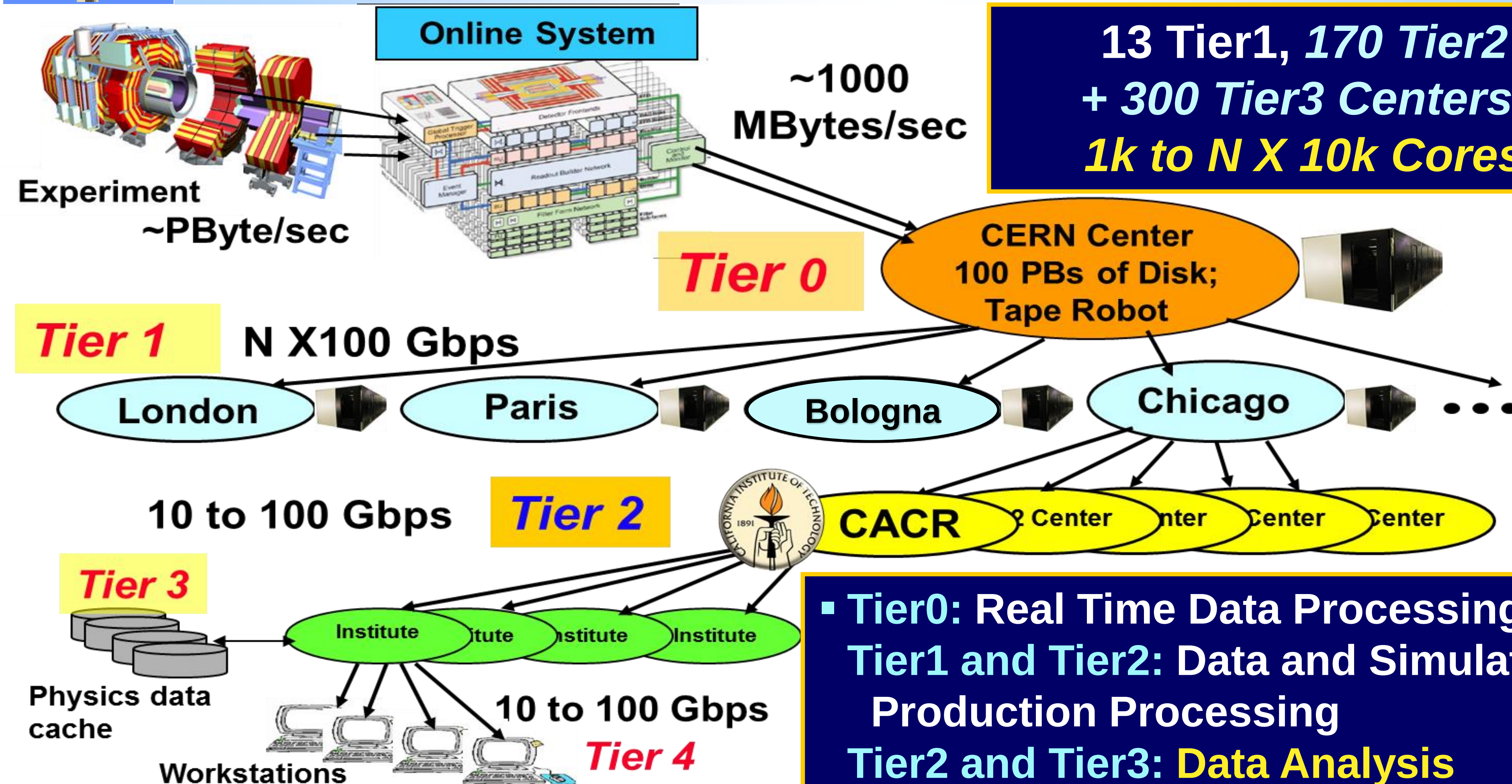
Run1: Higgs Boson Discovery 2012

Run2 and Going Forward: Precision Measurements and BSM Exploration: 2013 - 2042

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

- NOTE: ~95% of Data Still to be Taken
- Greater Intensity: Upgraded detectors for more complex events
- *To 25X Data Taking Rate in 2029-40*

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



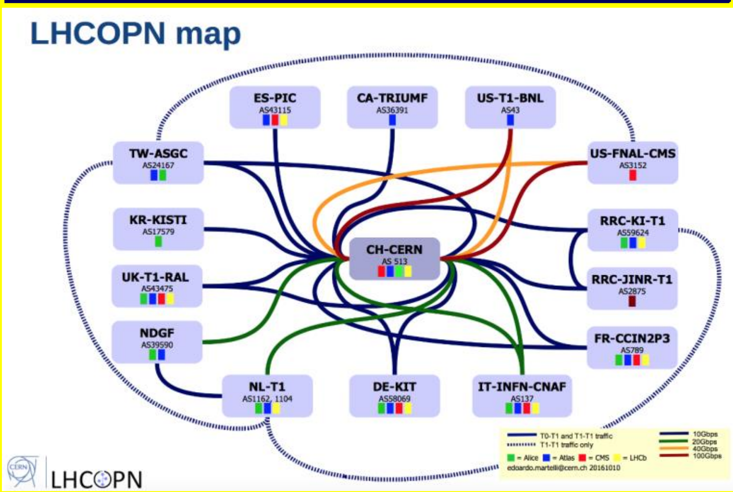
13 Tier1, *170 Tier2*
+ 300 Tier3 Centers:
1k to N X 10k Cores

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

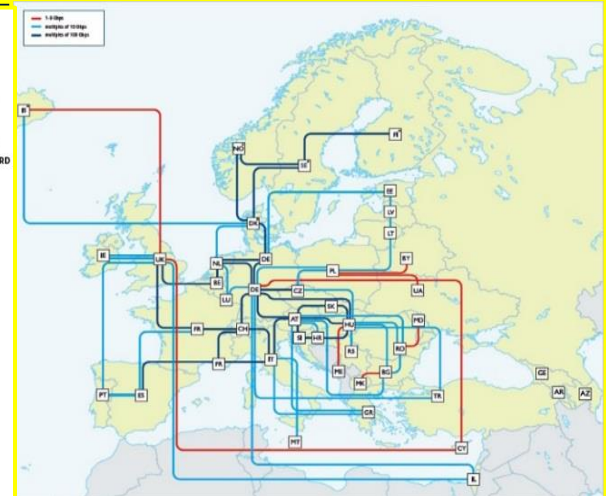
Core of LHC Networking LHCOPN, LHCONE, GEANT, ESnet, Internet2, CENIC…

LHCOPN: Simple & Reliable Tier0+1 Ops

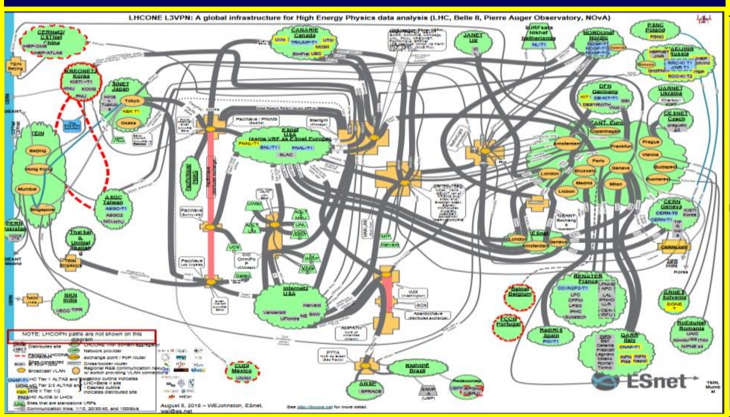


Internet2

GEANT



LHCONE VRF: 170 Tier2s ++



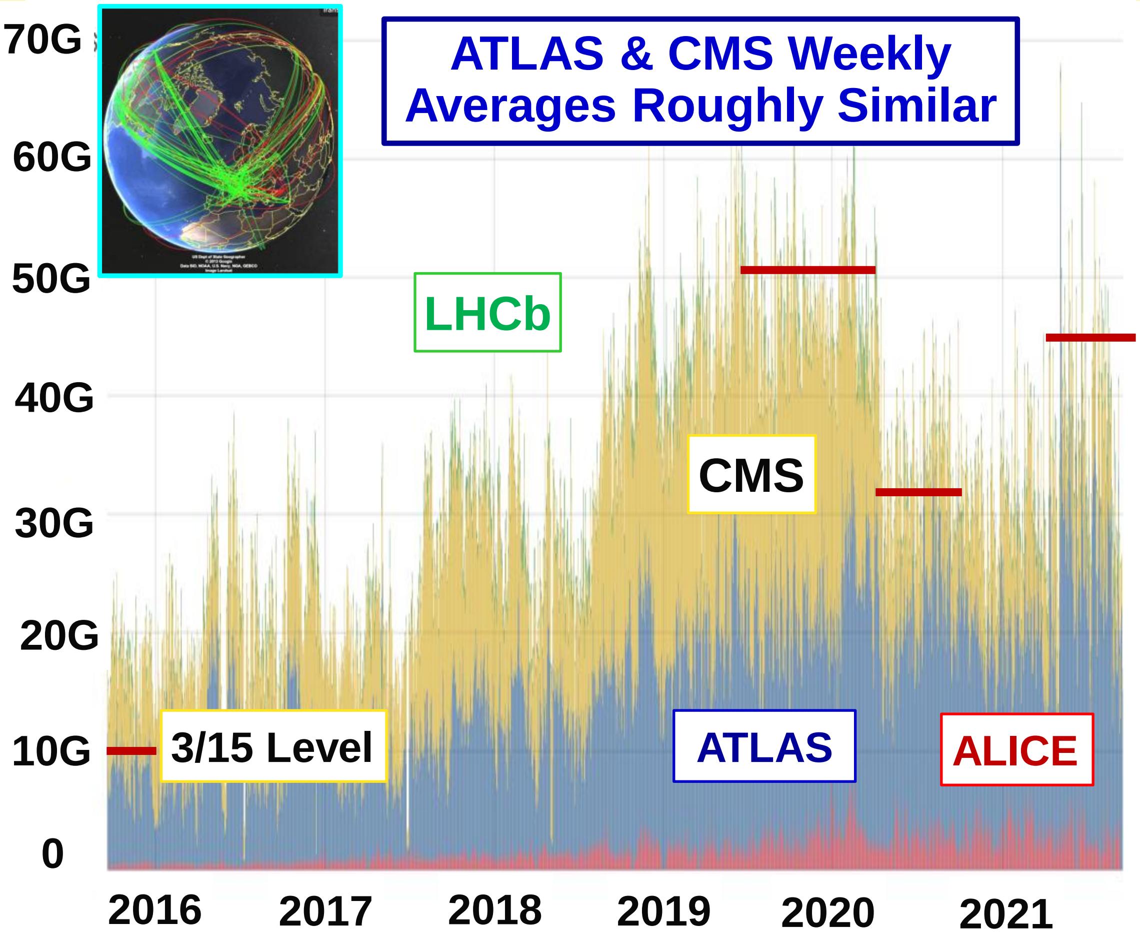
ESnet (with EEX)

CENIC and NRP

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

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

WLCG Transfers Dashboard: Throughput Sept 2015 – Sept 2021



15 to 58 GBytes/s Week Avg
To 70+ GBytes/s Daily Avg

Complex Workflow

- ~ 1M jobs (threads) simultaneously
- Multi-TByte to Petabyte Transfers;
- To ~10 M File Transfers/Day
- 100ks of remote connections

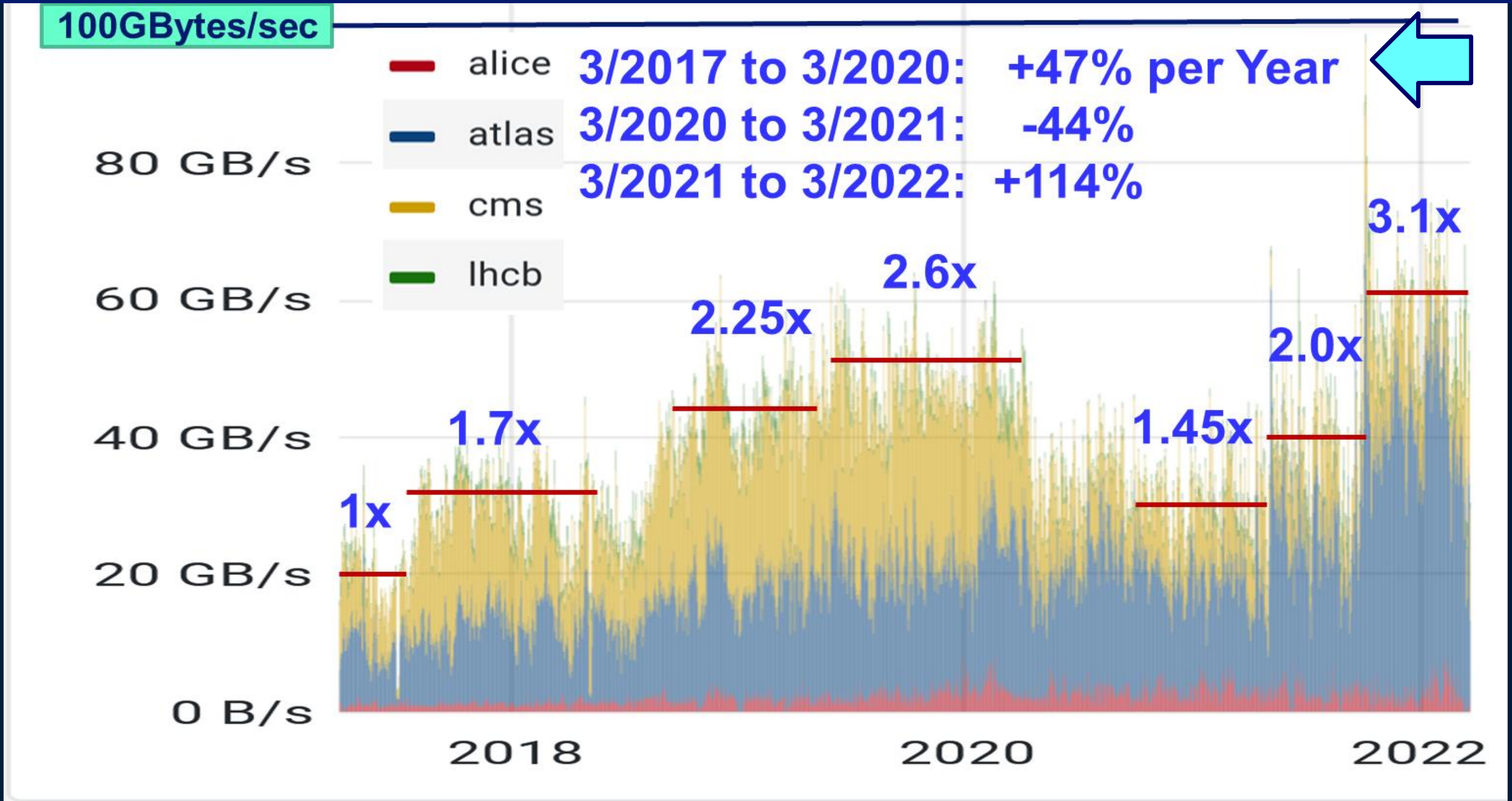
- The effects of Covid from Spring 2020 are evident
- The recovery is also evident

5X Growth in Throughput in 2015-2019: +50%/Yr; ~60X per Decade

https://monit-grafana.cern.ch/d/AfdonlvGk/wlcg-transfers?orgId=20&from=now-6y&to=now

WLCG Transfers Dashboard: 3/2017 to 3/2022

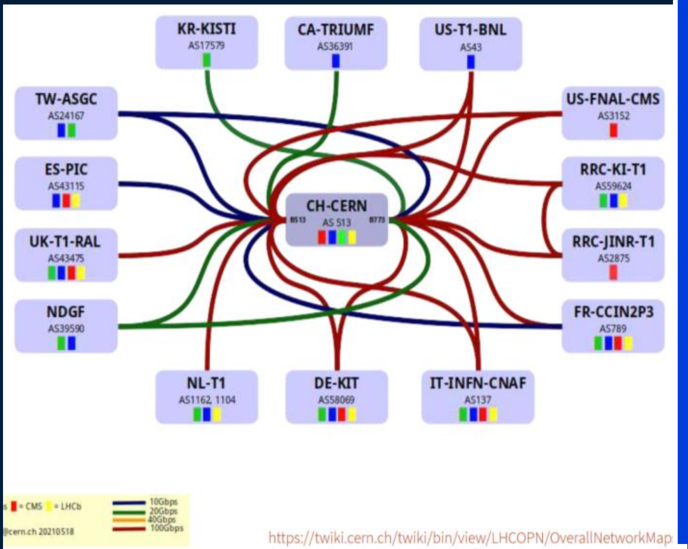
https://monit-grafana.cern.ch/d/AfdonIvGk/wlcg-transfers?orgId=20



HL LHC "10%" Data Challenge in October 2021 exceeded 100 Gbytes/sec for >1 Day
The "30%" Challenge Scheduled for 2024 will be a real challenge
General WLCG Traffic is again at or above the pre-pandemic level

Network Traffic Evolution 2021-22 to and from CERN

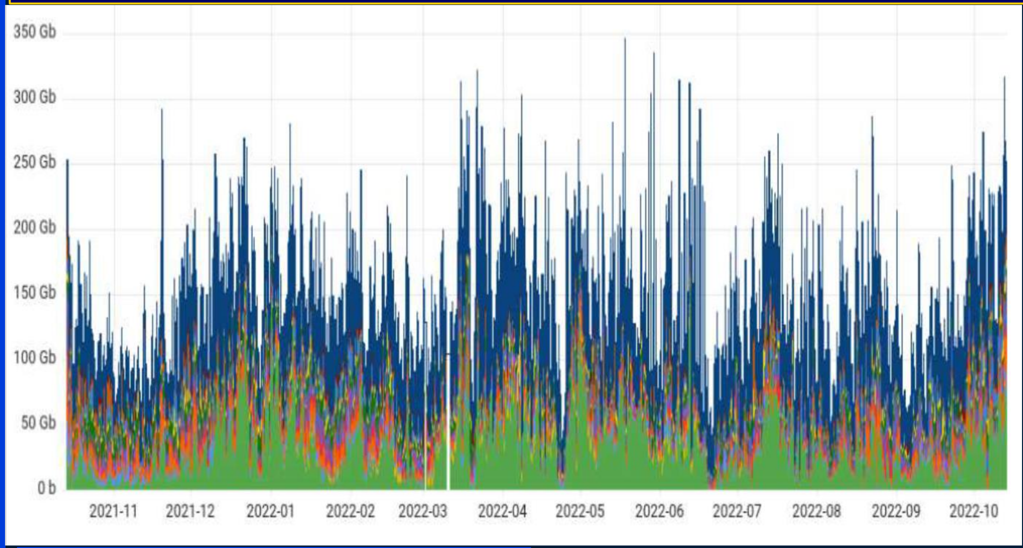
LHCOPN



Numbers

- 14 Tier1s + 1 Tier0
- 12 countries in 3 continents
- Dual stack IPv4-IPv6
- 1.3 Tbps to the Tier0

LHCOPN Traffic CERN to Tier1s



Overall CERN Traffic: LHCOPN, LHCONE, Internet

E. Martelli, CERN/IT

Numbers:

Sent out ~536 PB in the last 12 months

+34% compared to previous year (398PB)

Numbers:

Moved ~457 PB in the last 12 months

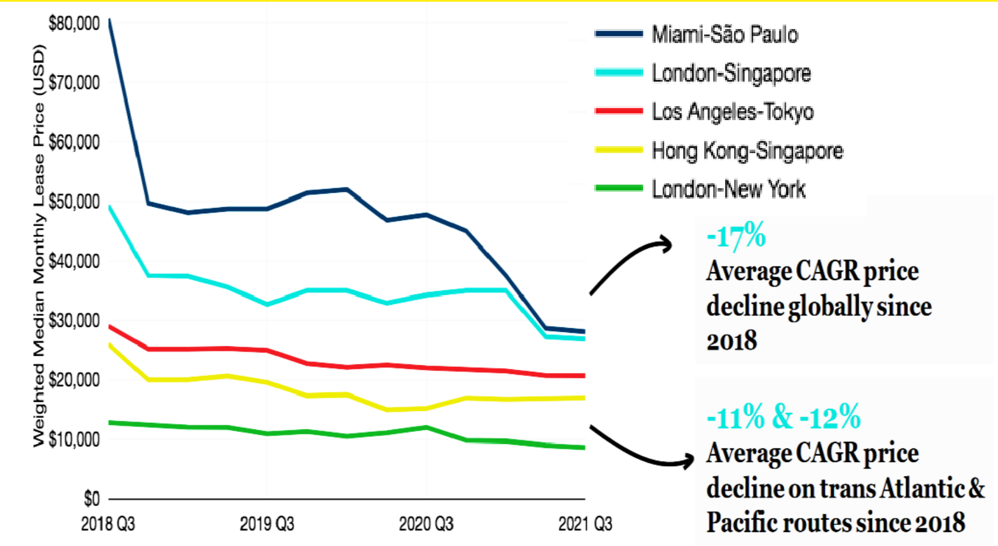
+34% compared to previous year (341PB)

Peak at ~391Gbps

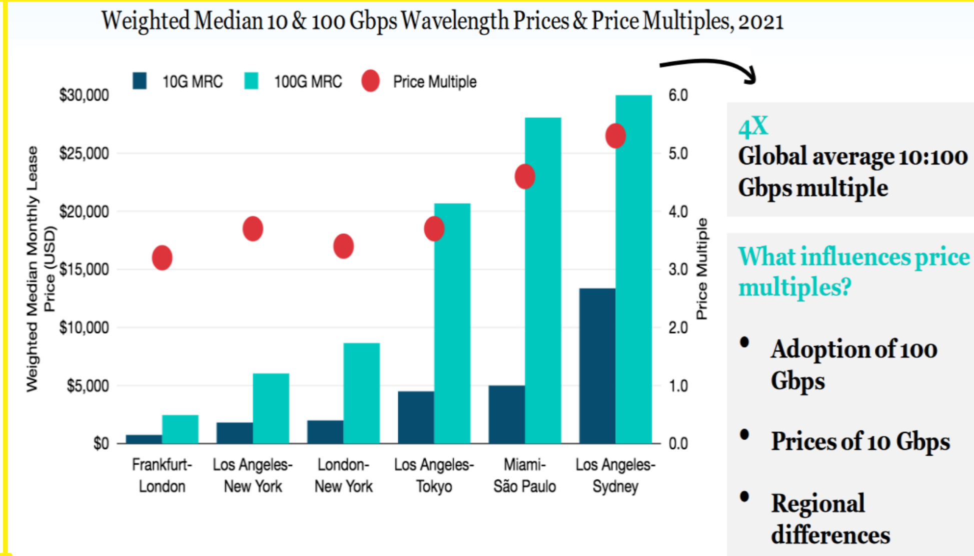
International Bandwidth Pricing Trends

https://www2.telegeography.com/hubfs/2022/Presentations/2022%20PTC%20Workshop.pdf

Weighted Median 100G Monthly Lease Prices 2018-21

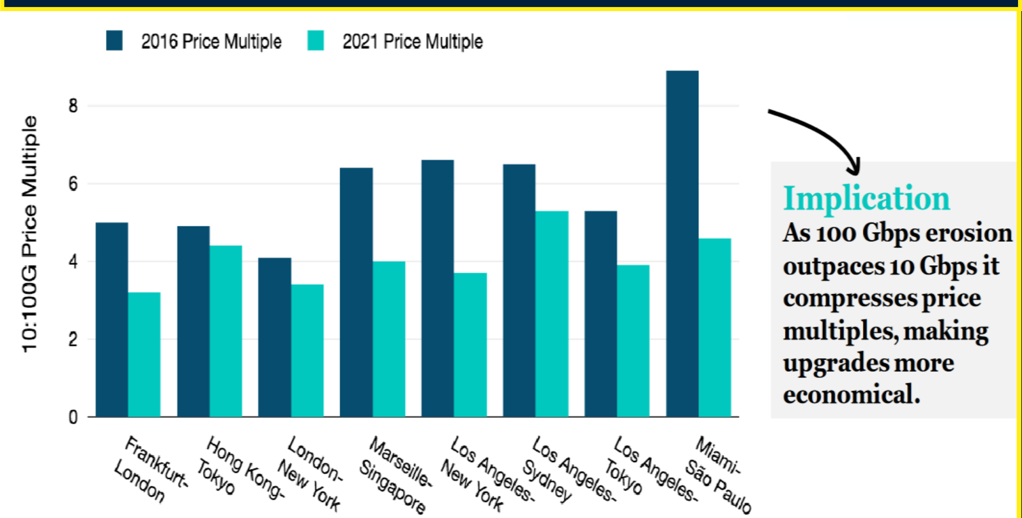


Weighted Median 100G/10G Price Multiples 2018-21



Compressing Price Multiples Over Time

Price Multiples 100G/10G by Region in 2016 and 2021



- **Price Evolution 2018-21 + Future**
  - -17% Price CAGR Average
  - Only -11 to -12 % CAGR NYC – London and LA-Tokyo
  - To -6% 2019-20 due to COVID
  - 100G/10G Price Multiple: Still ~4X, from 6.4X in 2015
  - 400G/100G Price Multiple Evolution Expected to be Similar: When: from 2025-28 ?

(Southern) California ((So)Cal) Cache

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

CPU in both places can access storage in both places.

How much disk space is enough?

Cache MINI and measure working set accessed:
0.45 Petabytes in October 2019

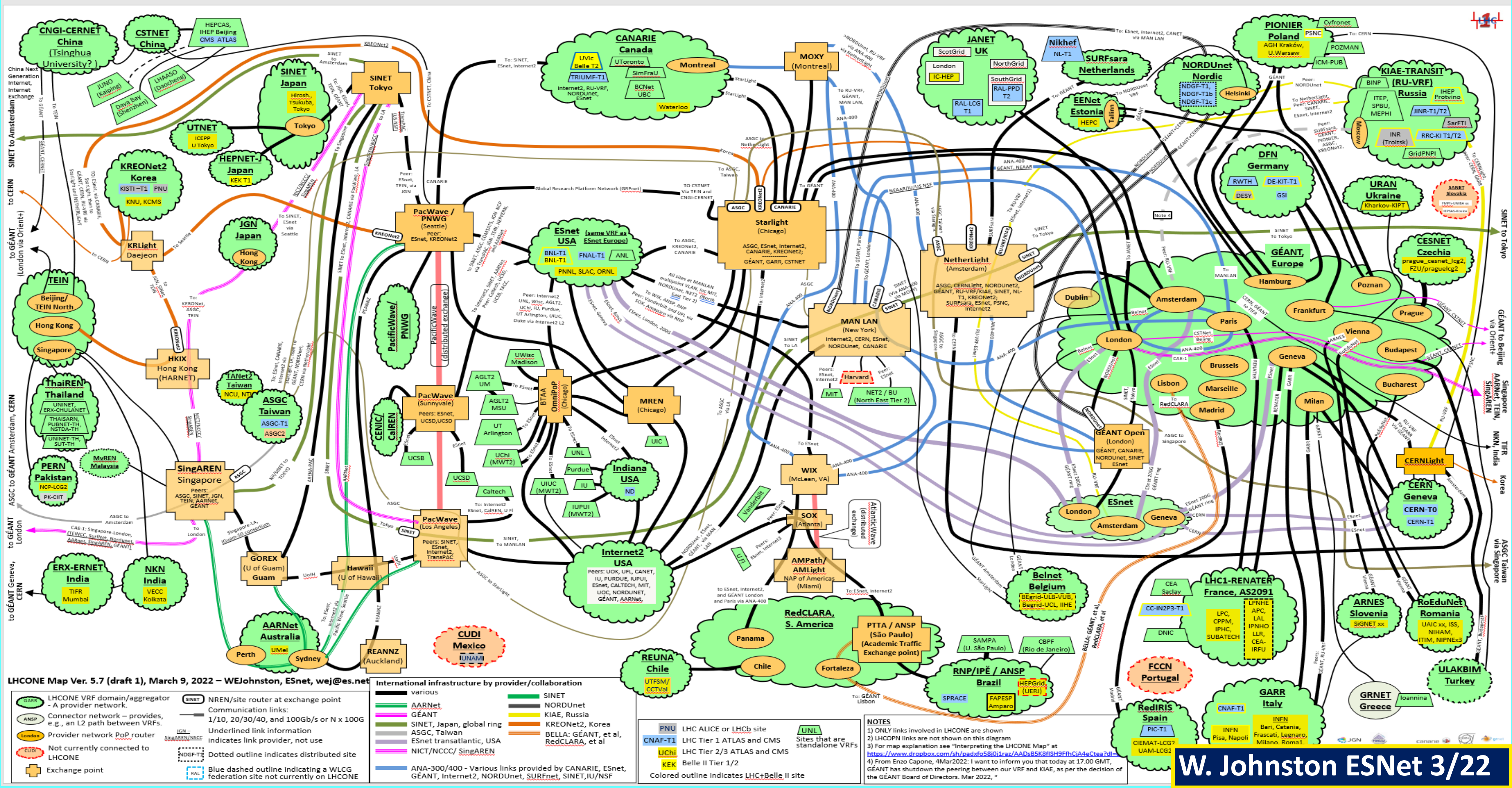
In early May, we added a cache at the ESNet POP in Sunnyvale to the SoCal cache.

Plan to include Riverside and other SoCal Tier3s;
ESnet and Internet2: additional in-network caches

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

LHCONE: a Virtual Routing and Forwarding (VRF) Fabric

Global infrastructure for *HEP* (*LHC, DUNE, Belle II, NOvA, Auger, Xenon, JUNO…*)



W. Johnston ESNet 3/22

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

A complex system with limited scaling properties. So: Multi-ONE ? New Mode of Sharing ?

LHCONE traffic growing by 60-70%/Yr: a challenge already in LHC Run3 (2022-5)

Annual CMS Data Volume

	# of collisions	# of events simulated	RAW event size [MB]	AOD event size [MB]	Total per year [PB]
Today	9 Billion	22 Billion	0.9	0.35	~20
HL-LHC	56 Billion	64 Billion	6.5	2	~600

The beams get "brighter" by x6
Data taking rate goes up by x6
Simulations go up by x3 (++)

Primary Data volume per year goes up by x30

Conclusion: CMS Data ~Exabyte/Year by ~2029 at HL-LHC

Towards a Computing Model for the HL LHC Era
Challenges: Capacity in the Core and at the Edges

- Programs such as the LHC have experienced rapid exponential traffic growth, at the level of 40-60% per year
  - At the January 2020 LHCONE/LHCOPN meeting at CERN, CMS and ATLAS expressed the need for Terabit/sec links on major routes, by the start of the HL-LHC in ~2029
    - **This is projected to outstrip the affordable capacity**
  - The needs are further specified in "blueprint" Requirements documents by US CMS and US ATLAS, submitted to the ESnet Requirements Review in August 2020, and captured in a comprehensive DOE Requirements Review report for HEP: *https://escholarship.org/uc/item/78j3c9v4*
  - Three areas of particular capacity-concern by 2028-9 were identified:

    (1) Exceeding the capacity across oceans, notably the Atlantic, served by the Advanced North Atlantic (ANA) network consortium

    (2) Tier2 centers at universities requiring 100G 24 X 7 X 365 average throughput with sustained 400G bursts (a petabyte in a shift), and

    (3) Terabit/sec links to labs and HPC centers (and edge systems) to support multi-petabyte transactions in hours rather than days

HL-LHC Network Needs and Data Challenges Requirements Study 2021-2

- **Export of Raw Data from CERN to the Tier1s (350 Pbytes/Year):**
  - **400 Gbps *Flat* each for ATLAS and CMS Tier1s; +100G each for other data formats; +100 G each for ALICE, LHCb**
- **"Minimal" Scenario [*]: Network Infrastructure from CERN to Tier1s Required**
  - **4.8 Tbps Aggregate: Includes 1.2 Tbps Flat (24 X 7 X 365) from the above, x2 to Accommodate Bursts, and x2 for overprovisioning, for operational headroom: including both non-LHC use, and other LHC use.**
  - ✸ **This includes *1.4 Tbps Across the Atlantic for ATLAS and CMS alone***
- **Note that the above Minimal scenario is where the network is treated as a scarce resource, unlike LHC Run1 and Run2 experience in 2009-18.**
- **In a "Flexible Scenario" [**]: *9.6 Tbps, including 2.7 Tbps Across the Atlantic* Leveraging the Network to obtain more flexibility in workload scheduling, increase efficiency, improve turnaround time for production & analysis**
  - **In this scenario: Links to Larger Tier1s in the US and Europe: ~ 1 Tbps (some more); Links to Other Tier1s: ~500 Gbps**
- **Tier2 provisioning: 400Gbps bursts, 100G Yearly Avg: ~Petabyte Import in a shift**
  - **Need to work with campuses to accommodate this: it may take years**

[*] NOTE: Matches numbers presented at ESnet Requirements Review (Summer 2020)

[**] NOTE: Matches numbers presented at the January 2020 LHCONE/LHCOPN Meeting

HL-LHC Online System Needs: 2023 Versus 2029
CMS Only (Phase 2 TDR)

- Event Network Throughput: 32x
- Computing Power: 53x
- Storage Throughput: 26x
- Storage Capacity Per Day: 16x (3.3 Pbytes)

CMS detector	LHC Phase-1	HL-LHC Phase-2	
Peak ⟨PU⟩	60	140	200
L1 accept rate (maximum)	100 kHz	500 kHz	750 kHz
Event Size at HLT input	2.0 MB [a]	6.1 MB	8.4 MB
Event Network throughput	1.6 Tb/s	24 Tb/s	51 Tb/s
Event Network buffer (60 s)	12 TB	182 TB	379 TB
HLT accept rate	1 kHz	5 kHz	7.5 kHz
HLT computing power [b]	0.7 MHS06	17 MHS06	37 MHS06
Event Size at HLT output [c]	1.4 MB	4.3 MB	5.9 MB
Storage throughput [d]	2 GB/s	24 GB/s	51 GB/s
Storage throughput (Heavy-Ion)	12 GB/s	51 GB/s	51 GB/s
Storage capacity needed (1 day [e])	0.2 PB	1.6 PB	3.3 PB

- Not Just a future issue:
  - Greater luminosity and pileup in 2023 compared to 2022: More CPU, more storage, greater network required
  - 50% greater CPU from external HPCs needed in 2023

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

leadershipteam@lists.gna-g.net





Buseung Cho
KISTI (Korea)

Marco Teixera
RedCLARA

Ivana Golub
PSNC, GEANT

Harvey Newman
Caltech (US)

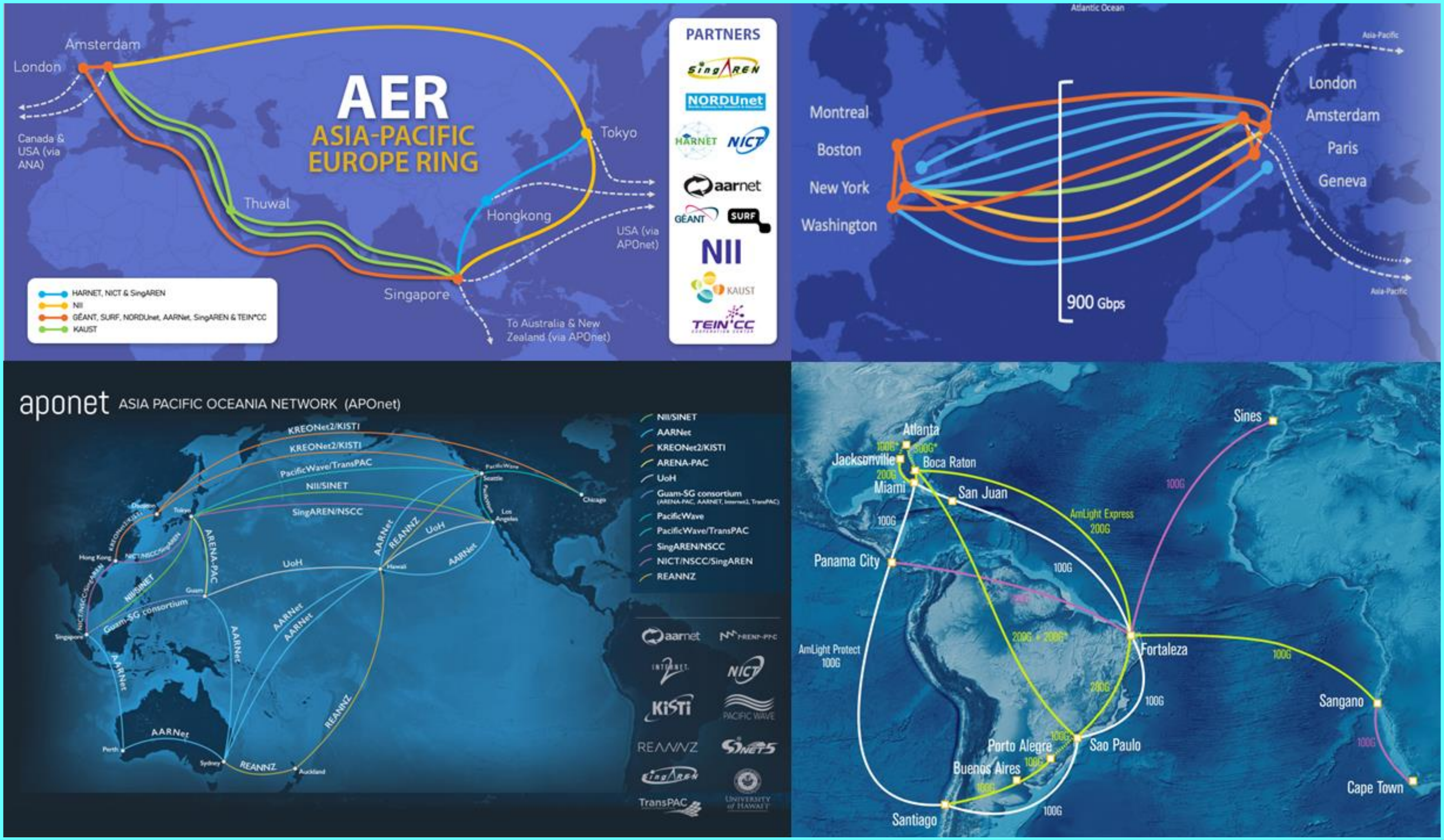
David Wilde, Chair
Aarnet (Australia)

- The GNA-G is an open volunteer group devoted to developing the blueprint to make using the Global R&E networks both simpler and more effective, operating under GNA-G.
- Its primary mission is to support global research and education using the technology, infrastructures and investments of its participants.
- The GNA-G needs to be 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 ata intensive programs

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

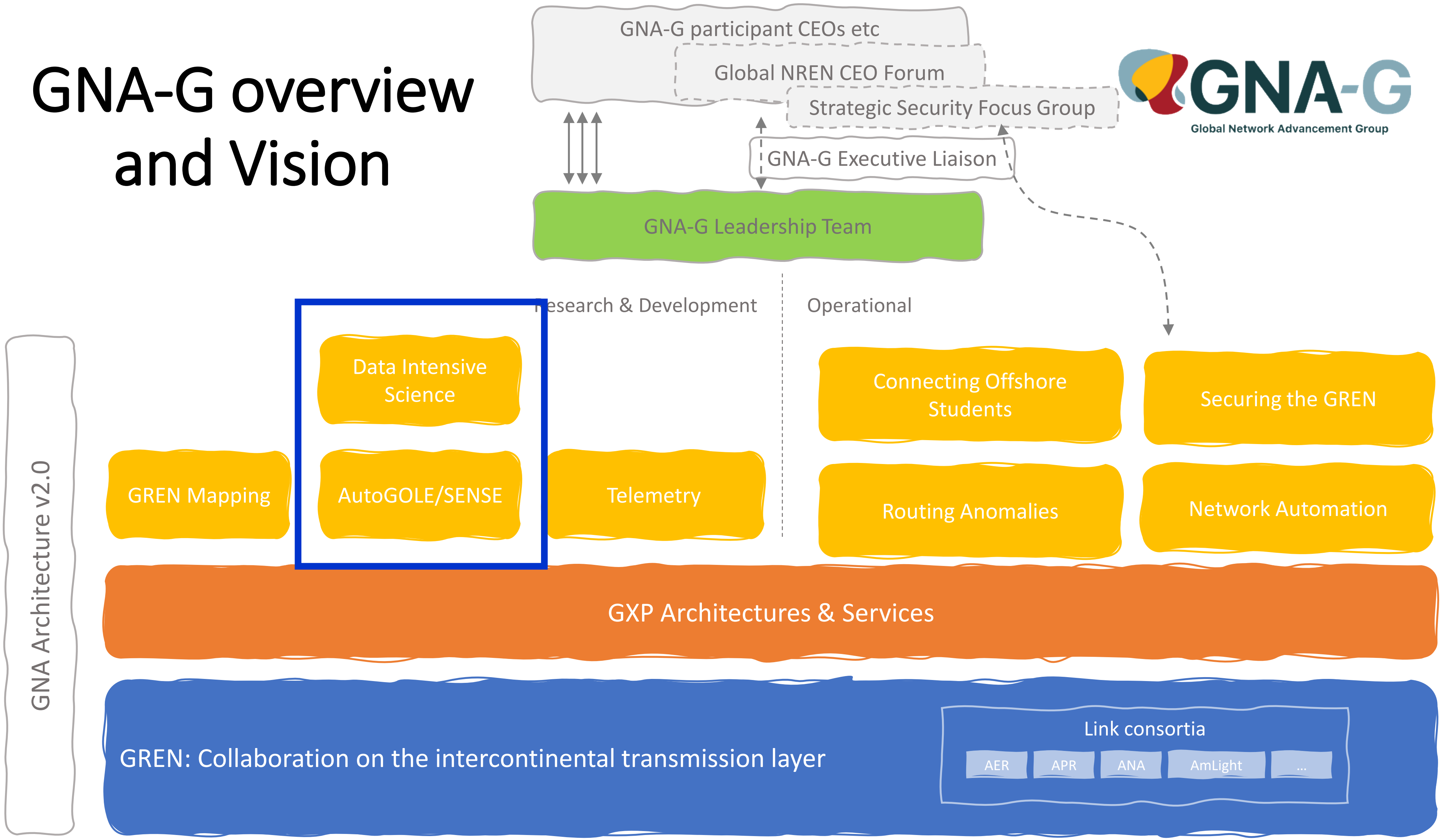
GNA-G
Global Network Advancement Group
https://www.gna-g.net/

Mission: To 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.

GNA-G overview and Vision



The GNA-G Data Intensive Sciences WG

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

- *Mission: Meet the challenges of globally distributed data and computation faced by the major science programs*
- *Coordinate provisioning the feasible capacity across a global footprint, and enable best use of the infrastructure:*
  - *While meeting the needs of the participating groups, large and small*
  - *In a manner Compatible and Consistent with use by the at-large A&R communities*
- *Members:*

  Alberto Santoro, Azher Mughal, Bijan Jabbari, Brian Yang, Buseung Cho, Caio Costa, Carlos Antonio Ruggiero, Carlyn Ann-Lee, Chin Guok, Ciprian Popoviciu, Dale Carder, David Lange, David Wilde, Dima Mishin, Edoardo Martelli, Eduardo Revoredo, Eli Dart, Eoin Kenney, Frank Wuerthwein, Frederic Loui, Harvey Newman, Heidi Morgan, Iara Machado, Inder Monga, Jeferson Souza, Jensen Zhang, Jeonghoon Moon, Jeronimo Bezerra, Jerry Sobieski, Joao Eduardo Ferreira, Joe Mambretti, John Graham, John Hess, John Macauley, Julio Ibarra, Justas Balcas, Kai Gao, Karl Newell, Kaushik De, Kevin Sale, Lars Fischer, Liang Zhang, Mahdi Solemani, Maria Del Carmen Misa Moreira, Marcos Schwarz, Mariam Kiran, Matt Zekauskas, Michael Stanton, Mike Hildreth, Mike Simpson, Ney Lemke, Phil Demar, Raimondas Sirvinskas, Richard Hughes-Jones, Rogerio Iope, Sergio Novaes, Shawn McKee, Siju Mammen, Susanne Naegele-Jackson, Tom de Fanti, Tom Hutton, Tom Lehman, William Johnston, Xi Yang, Y. Richard Yang
- *Participating Organizations/Projects:*
- *ESnet, Nordunet, SURFnet, AARNet, AmLight, KISTI, SANReN, GEANT, RNP, CERN, Internet2, CENIC/Pacific Wave, StarLight, NetherLight, Southern Light, Pacific Research Platform, FABRIC, RENATER, ATLAS, CMS, VRO, SKAO, OSG, Caltech, UCSD, Yale, FIU, UERJ, GridUNESP, Fermilab, Michigan, UT Arlington, George Mason, East Carolina, KAUST*
- *Working Closely with the AutoGOLE/SENSE WG*

✸ *Meets Weekly or Bi-weekly; All are welcome to join.*

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

Towards a Next Generation Network-Integrated System For the LHC Program and Data Intensive Sciences

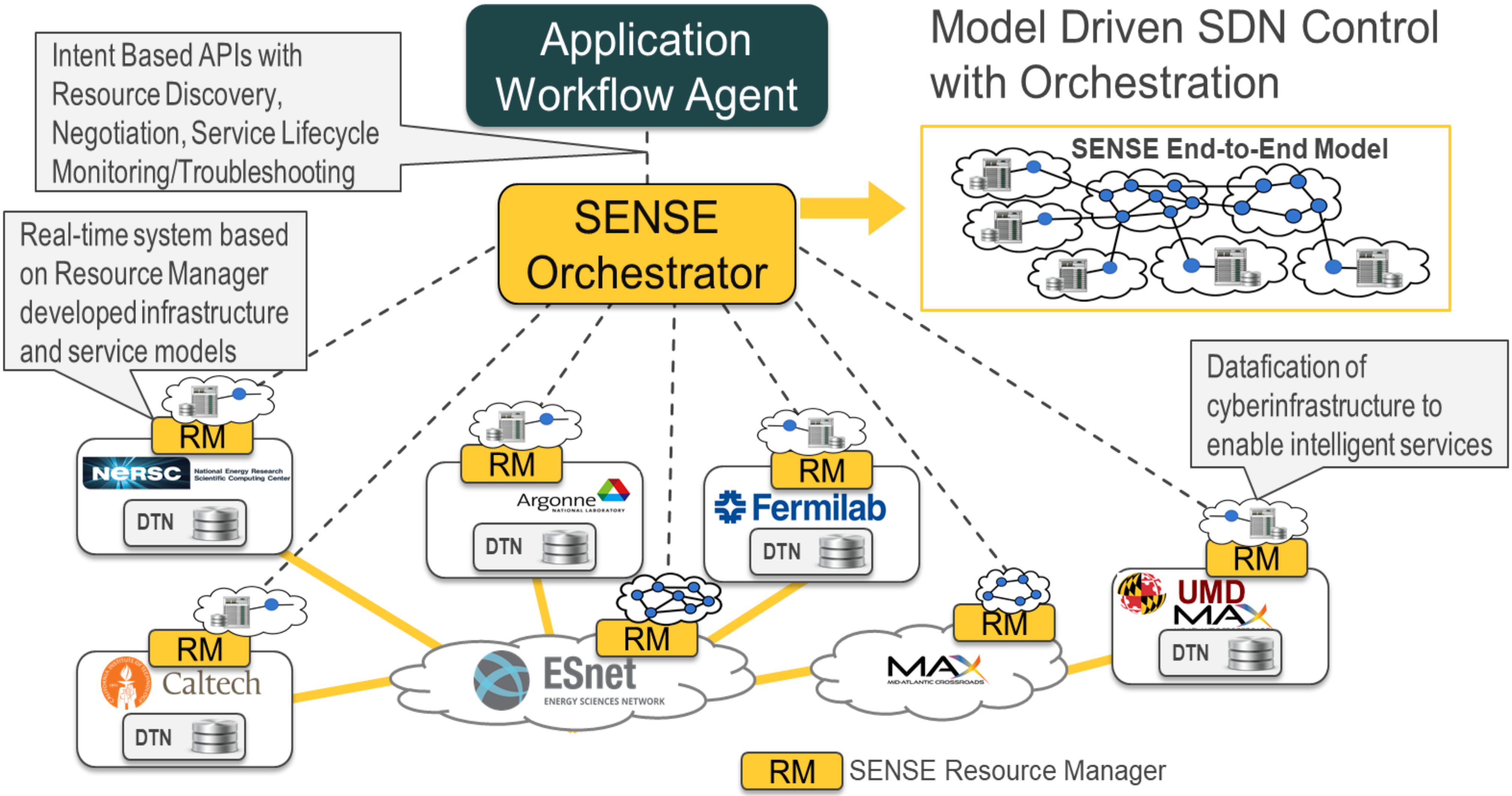
Next Generation Network-Integrated System

- Top Line Message: To meet the needs, we need to transition to a new dynamic and adaptive software-driven system, which
  - Coordinates worldwide networks as a first class resource along with computing and storage, across multiple domains
  - Simultaneously supports the LHC experiments, other major DIS programs and the larger worldwide academic and research community
- Systems design approach: A global dynamic fabric that flexibly allocates, balances and conserves the available network resources
  - Negotiating with site systems that aim to accelerate workflow
- Builds on ongoing R&D projects: from regional caches/data lakes to intelligent control and data planes to ML-based optimization [E.g. SENSE/AutoGOLE, NOTED, ESNet HT, GEANT/RARE, AmLight, Fabric, Bridges; NetPredict, DeepRoute, ALTO, PolKA ...]
- We are also leveraging the worldwide move towards a fully programmable ecosystem of networks and end-systems (P4, PINS; SRv6), plus operations platforms (OSG, NRP; global SENSE Testbed)
- The LHC experiments together with the WLCG, the GNA-G and its Working Groups, and the worldwide R&E network community, are the key players
  - Directions being taken up by other programs: LBNF/DUNE, VRO, SKA

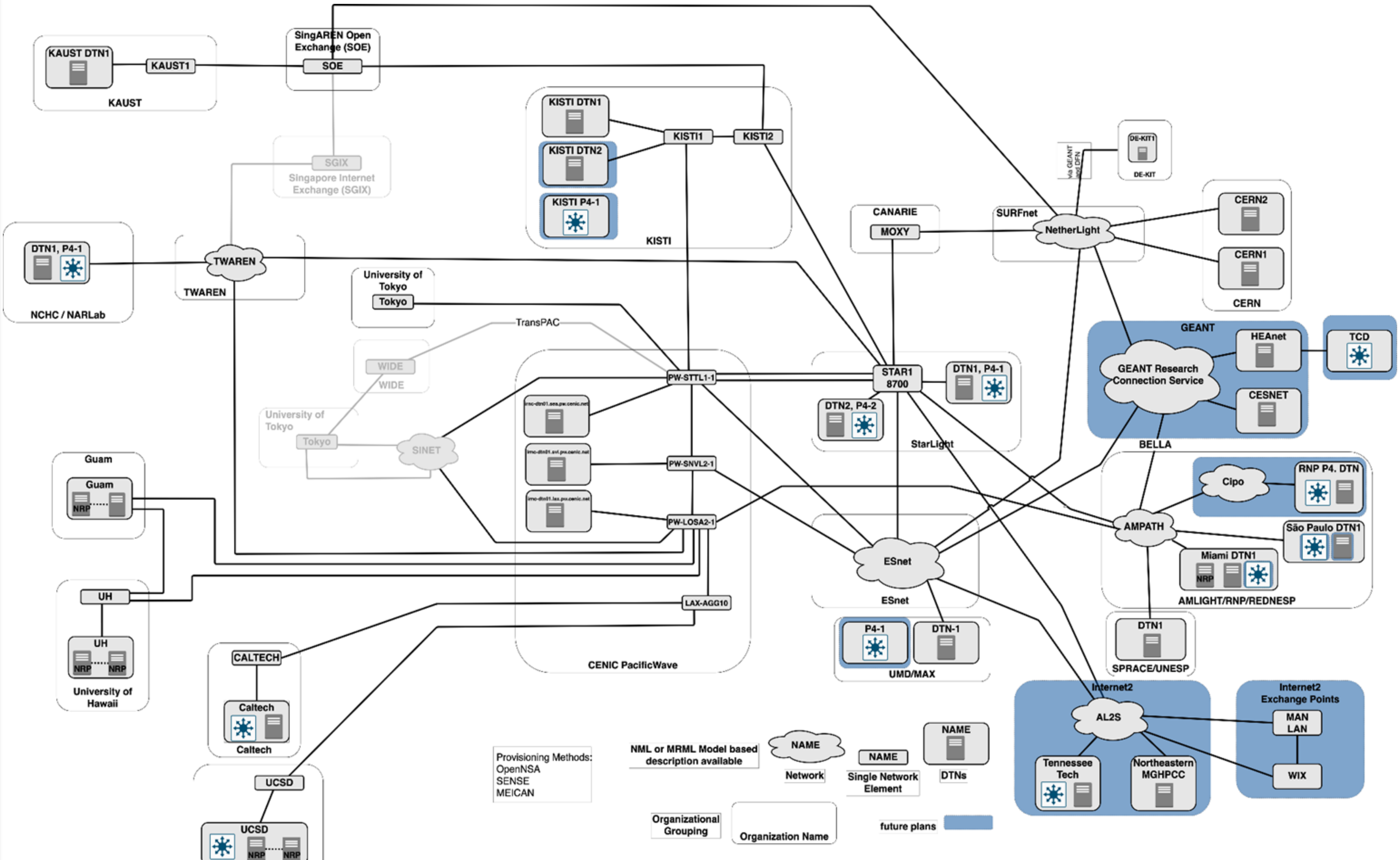
SENSE: SDN Enabled Networks for Science at the Exascale

ESnet, Caltech, Fermilab, LBNL https://arxiv.org/abs/2004.05953

Creates Virtual Circuit Overlays. Site and Network RMs, Orchestrator



GNA-G AutoGOLE/SENSE WG Global Persistent Testbed



Software Driven Network OSes; Programmability at the Edges and in the Core

The Rising Transpacific and Asia Pacific Network Community In a Global Context

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:
    Promtheus based, deployments now 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 Activity, Use Case support:
    Including NOTED, Gradient Graph, P4 Topologies, Named Data Networking (NDN), Data Transfer Systems integration and testing.
- WG information
  https://www.gna-g.net/join-working-group/autogole-sense

SENSE Services for LHC and Other Open Science Grid (OSG) Workflows: Rucio, FTS, XRootD Integration

- Goals center around SENSE providing key network-related capabilities to LHC + 30 other science programs
- Primary Motivations:
  - Developing mechanisms for an application workflow to obtain information regarding the network services, capabilities, and options; to a degree similar to what is possible for compute resources
  - Giving applications the ability to interact with the network: to exchange information, negotiate performance parameters, discover expected performance metrics, and receive status/ troubleshooting information in real time, during long transactions.
- Key pathways:
  (1) Interfacing and interaction with RUCIO, FTS and XRootD data federation and storage systems, through Engagement: with CMS, ATLAS and those development teams
  (2) Enabled by the ESnet6 plan: with up to ~75% of link capacity on key routes available for policy-driven dynamic provisioning and management by ~2027-28
- 4 Phase 1 Year Schedule, Completing in 2023: from system evaluation to design, to prototype deployments and tests, to a plan for the transition to operations and the additional R&D needed

Interfacing to Multiple VOs With FTS/Rucio/XRootD

LHC, Dark Matter, ν, Heavy Ions, VRO, SKAO, LIGO/Virgo/Kagra; Bioinformatics

CMS

ATLAS EXPERIMENT

LHCb

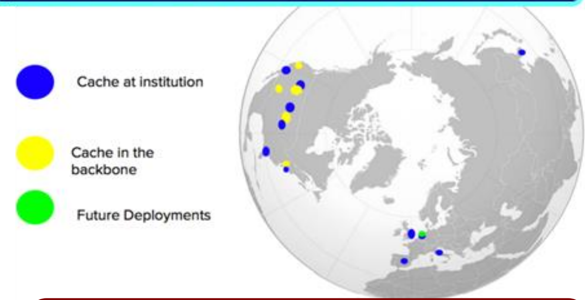
ALICE

ICECUBE SOUTH POLE NEUTRINO OBSERVATORY

Xe XENON Dark Matter Project

ESCAPE European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

OSG Data Federation



More than a dozen caches deployed across 3 continents

Collaboration	Working Set	Data Read	Reread Multiplier
DUNE	25GB	131TB	5.4k
LIGO (private)	41.4TB	3.8PB	95
LIGO (public)	4.3TB	1.5PB	318
MINERVA	351GB	116TB	340
DES	268GB	17TB	66
NOVA	268GB	308TB	1.2k
RPI_Brown	67GB	541TB	8.3k

7 most popular data areas

FAIR Facility for Antiproton and Ion Research

cta cherenkov telescope array — the observatory for ground-based gamma-ray astronomy

EST european solar telescope

ESO

SKA SQUARE KILOMETRE ARRAY

AENEAS Advanced European Network of E-infrastructures for Astronomy with the SKA

European Science Data Center

LSC LIGO Scientific Collaboration

VIRGO

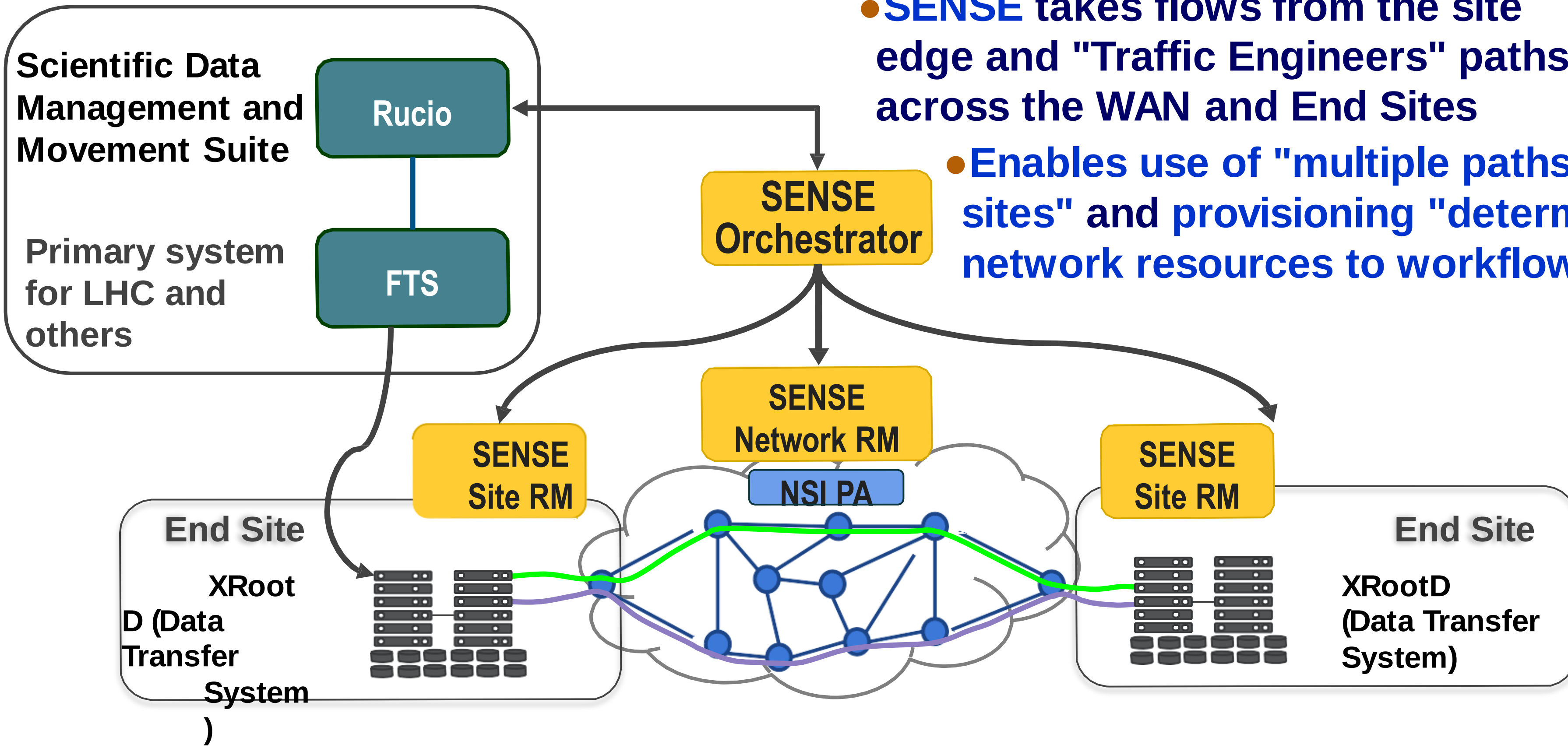
KAGRA

Vera Rubin Observatory

SENSE and Rucio/FTS/XRootD Interoperation

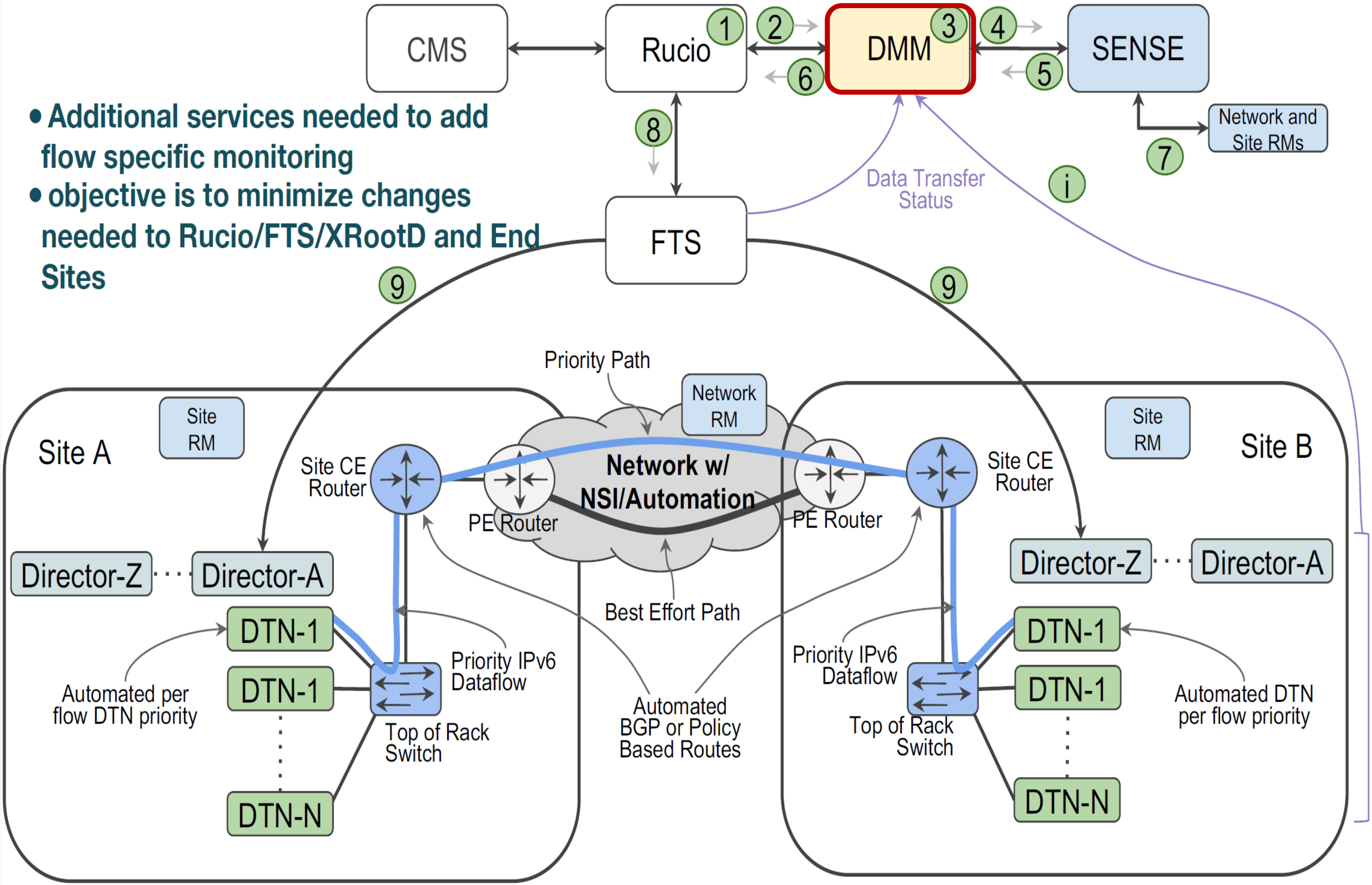
- **Rucio identifies groups of data flows (IPv6 subnets) which are "high priority"**

- **SENSE takes flows from the site edge and "Traffic Engineers" paths across the WAN and End Sites**
  - **Enables use of "multiple paths between sites" and provisioning "deterministic" network resources to workflows**



SENSE and Rucio/FTS/XRootD Interoperation

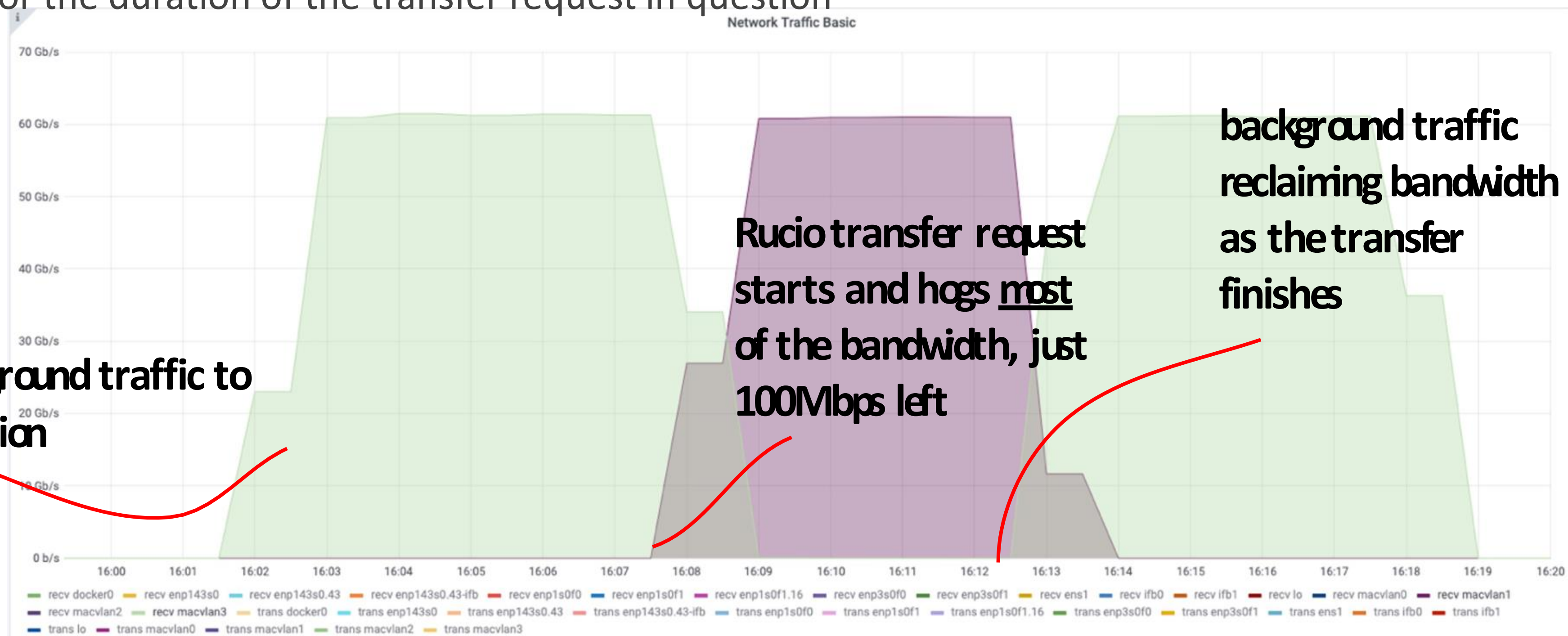
- Additional services needed to add flow specific monitoring
- objective is to minimize changes needed to Rucio/FTS/XRootD and End Sites



Development and Testing Ongoing

Creating a priority service between 2 sites:

- On demand i.e. triggered solely by the creation of a rule in Rucio
- On a congested network path (to show QoS)
- Just for the duration of the transfer request in question



60 Gbps

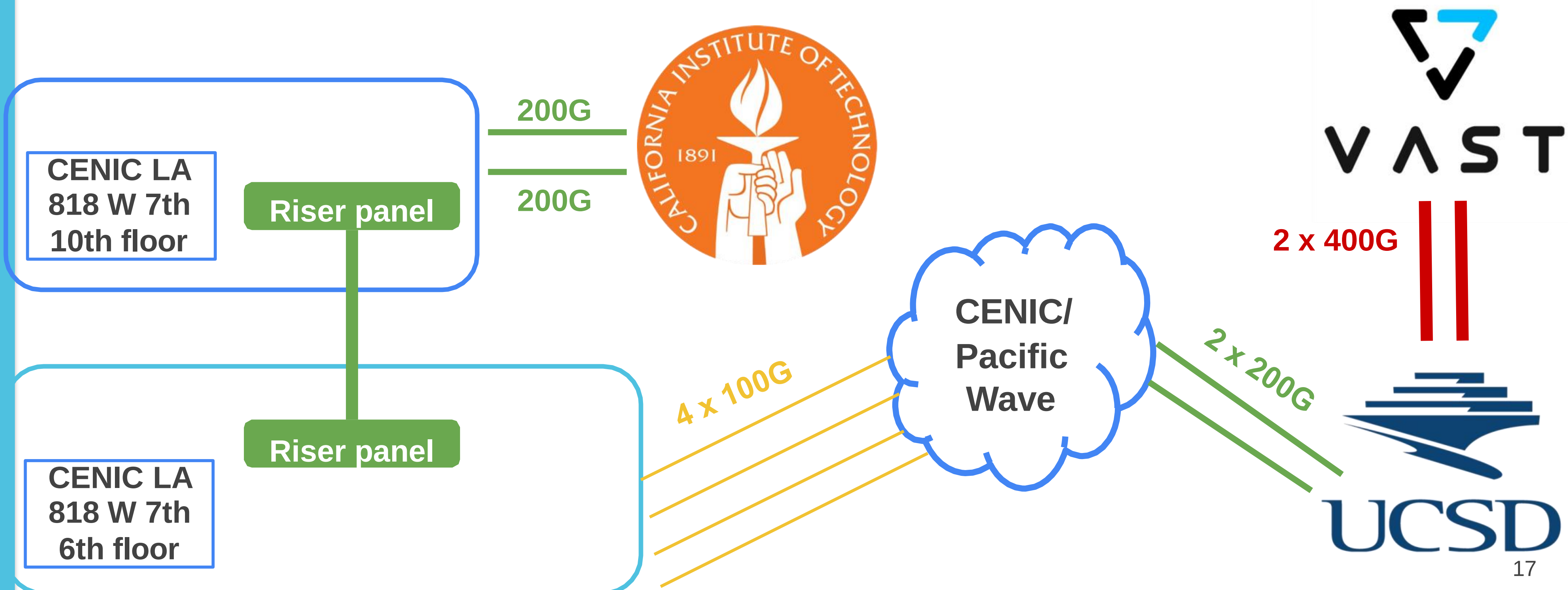
Artificial background traffic to produce congestion

Rucio transfer request starts and hogs <u>most</u> of the bandwidth, just 100Mbps left

background traffic reclaiming bandwidth as the transfer finishes

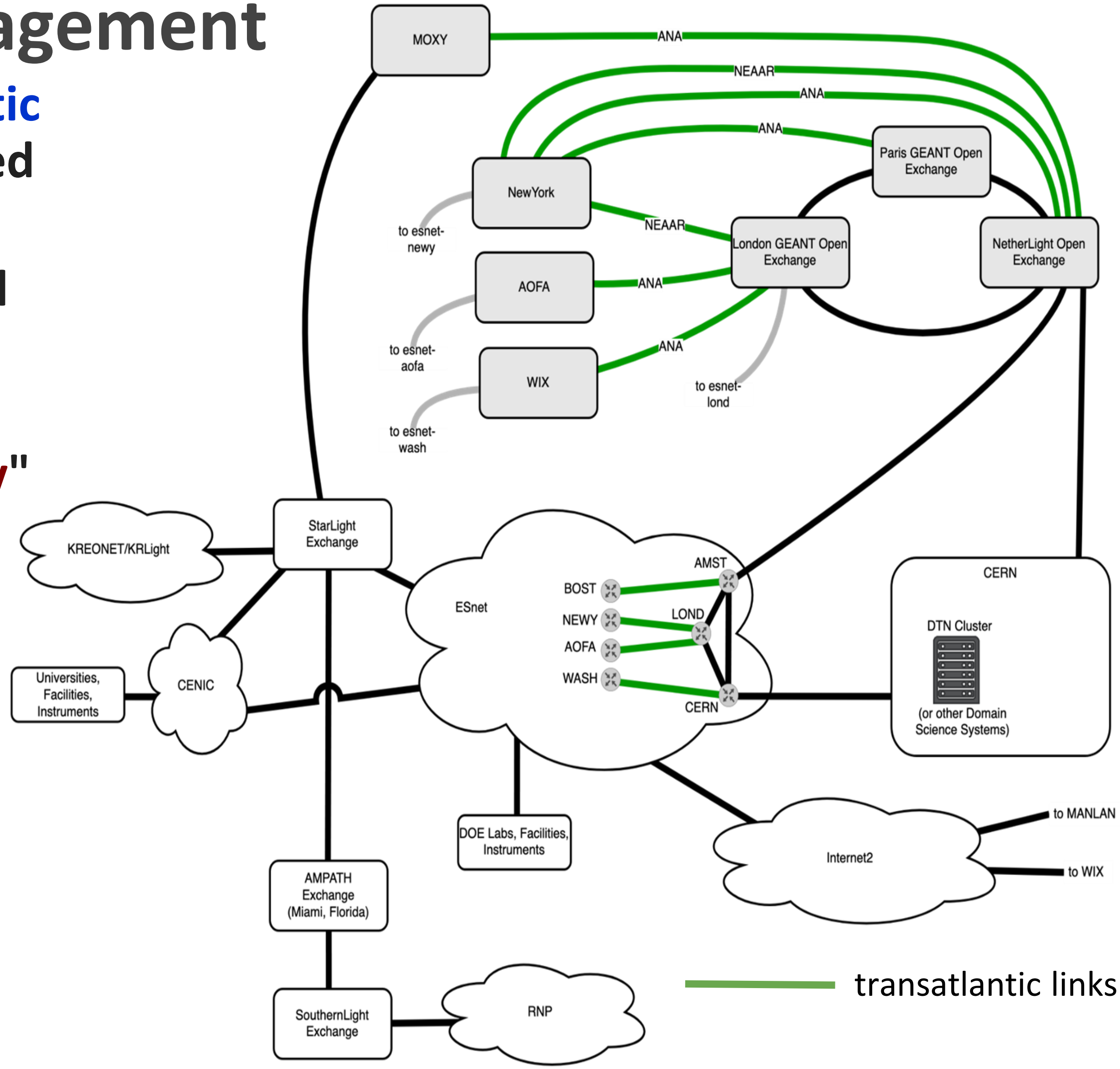
Working toward operation at 400Gbps between sites

UCSD to Caltech XRootD Transfers at 300 Gbs, working thru several issues: XRootD push vs pull, host level QoS configuration



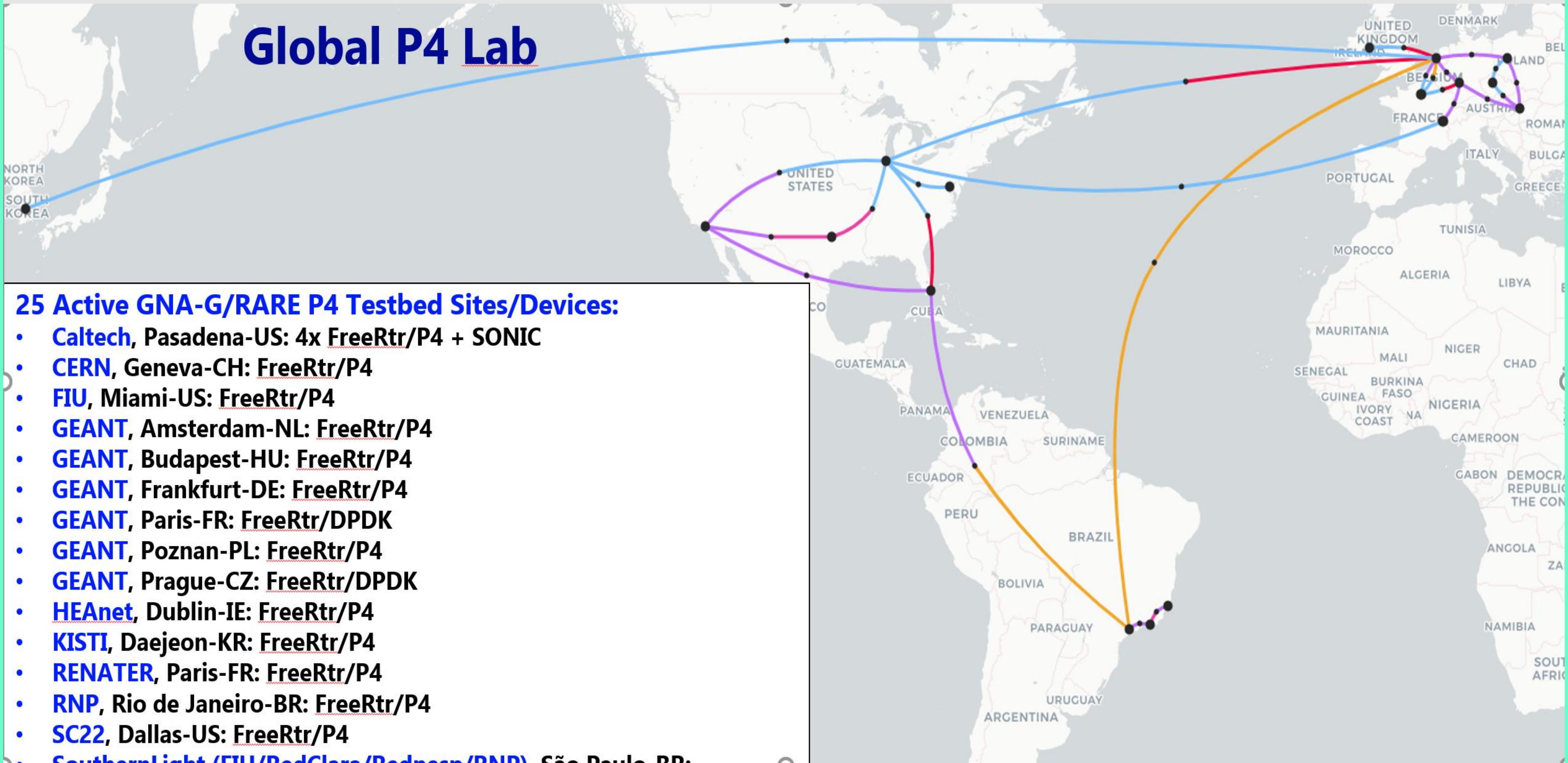
Important Link Management

- There are **multiple transatlantic and transpacific links,** operated by multiple organizations
- Goal is to more flexibly control how these are utilized on a per flow, group, or use basis
- Do not want to manage "every" flow in the network, however should be able to manage "any" flow in the network



- An equally important goal is to understand the load and leave room for other traffic
- Compatible with other network operations
- Result is to make better use of the available resources overall
- Two timescales: **SENSE overlay network of virtual circuits with BW guarantees is relatively stable; IPv6 subnets and Directors provide more dynamic flow handling**

P4 + SONIC Programmable Global PersistentTestbed



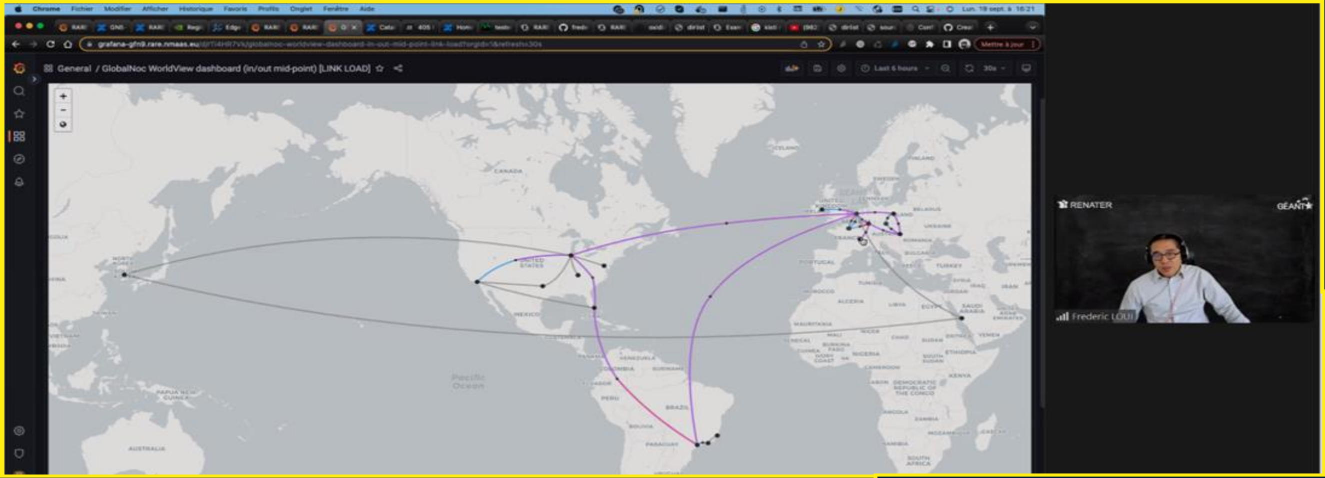
25 Active GNA-G/RARE P4 Testbed Sites/Devices:

- Caltech, Pasadena-US: 4x FreeRtr/P4 + SONIC
- CERN, Geneva-CH: FreeRtr/P4
- FIU, Miami-US: FreeRtr/P4
- GEANT, Amsterdam-NL: FreeRtr/P4
- GEANT, Budapest-HU: FreeRtr/P4
- GEANT, Frankfurt-DE: FreeRtr/P4
- GEANT, Paris-FR: FreeRtr/DPDK
- GEANT, Poznan-PL: FreeRtr/P4
- GEANT, Prague-CZ: FreeRtr/DPDK
- HEAnet, Dublin-IE: FreeRtr/P4
- KISTI, Daejeon-KR: FreeRtr/P4
- RENATER, Paris-FR: FreeRtr/P4
- RNP, Rio de Janeiro-BR: FreeRtr/P4
- SC22, Dallas-US: FreeRtr/P4
- SouthernLight (FIU/RedClara/Rednesp/RNP), São Paulo-BR: FreeRtr/P4
- StarLight, Chicago-US: FreeRtr/P4
- SWITCH, Geneva-CH: FreeRtr/P4
- TCD, Dublin-IE: FreeRtr/P4
- Tennessee Tech, Cookeville-US: FreeRtr/P4
- UFES, Vitória-BR: FreeRtr/P4
- UMd/MAX, College Park-MD: FreeRtr/P4

+5 Expected Sites/Devices (dec/2022):

- JISC, London-UK: FreeRtr/P4
- KAUST, Saudi Arabia-SA: FreeRtr/DPDK
- RNP, Rio de Janeiro-BR: +1 SONiC/P4
- UCSD, San Diego-US: SONiC/P4
- UFES, Vitória-BR: +1 FreeRtr/P4

GEANT (Now Global) P4 Lab
and Dataplane of P4 Programmable Switches



GEANT + RNP
Frederic Loui
Narcos Scwarz
et al.

A new worldwide platform for
- New agile and flexible feature development
- New use case development corresponding to research programs’ requirements
- Multiple research network overlay slices

A global playing field for
- Next generation network monitoring tools and systems
- Development of fully automated network deployment
- New network operations paradigm development

VISION: Federate and integrate multiple testbeds and toolsets such as AutoGOLE / SENSE

LATEST: Rapidly Deployable Digital Twins
with high fidelity (Container Lab) for complex network topologies
Plan: Smooth Transition from Simulation to Real-World Global Operations

Global P4 Lab: Motivation – Marcos Schwarz, RNP
Frederic Loui, GEANT

Can we increase the rate of evolution without interfering with production networks?

To develop and operate end-to-end / multi-domain orchestration services:

- Resource reservation (guaranteed bandwidth)
- Resource provisioning (Circuits, VRFs)
- Underlay observability
- Dynamic traffic steering/engineering
- Dynamic creation of L3 VPNs
- Closed loop multi-domain visibility/intelligence/controllability

How can we create/sustain an integration initiative/platform to propose and validate next generation protocol and services?

- Proposition: Use programmable P4 devices to experiment on pre-production networks leveraging industry/R&E open ecosystems

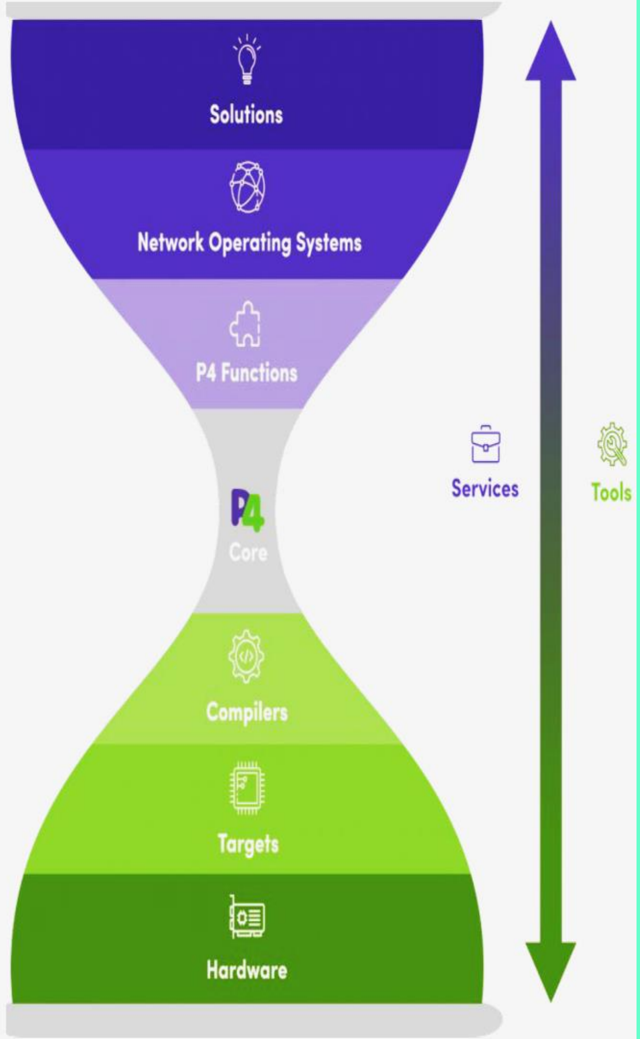
Why P4 ?

Open Networking Ecosystem for Hardware and Software

- Programmable ASICs, FPGAs or generic CPUs
- Open implementation of industry standards
- Ease the development of experimental/research features

Extensible Production Grade Network Operating Systems

- SONiC/PINS (LF/ONF)
- RARE/FreeRtr (GEANT)



Global P4 Lab Goals

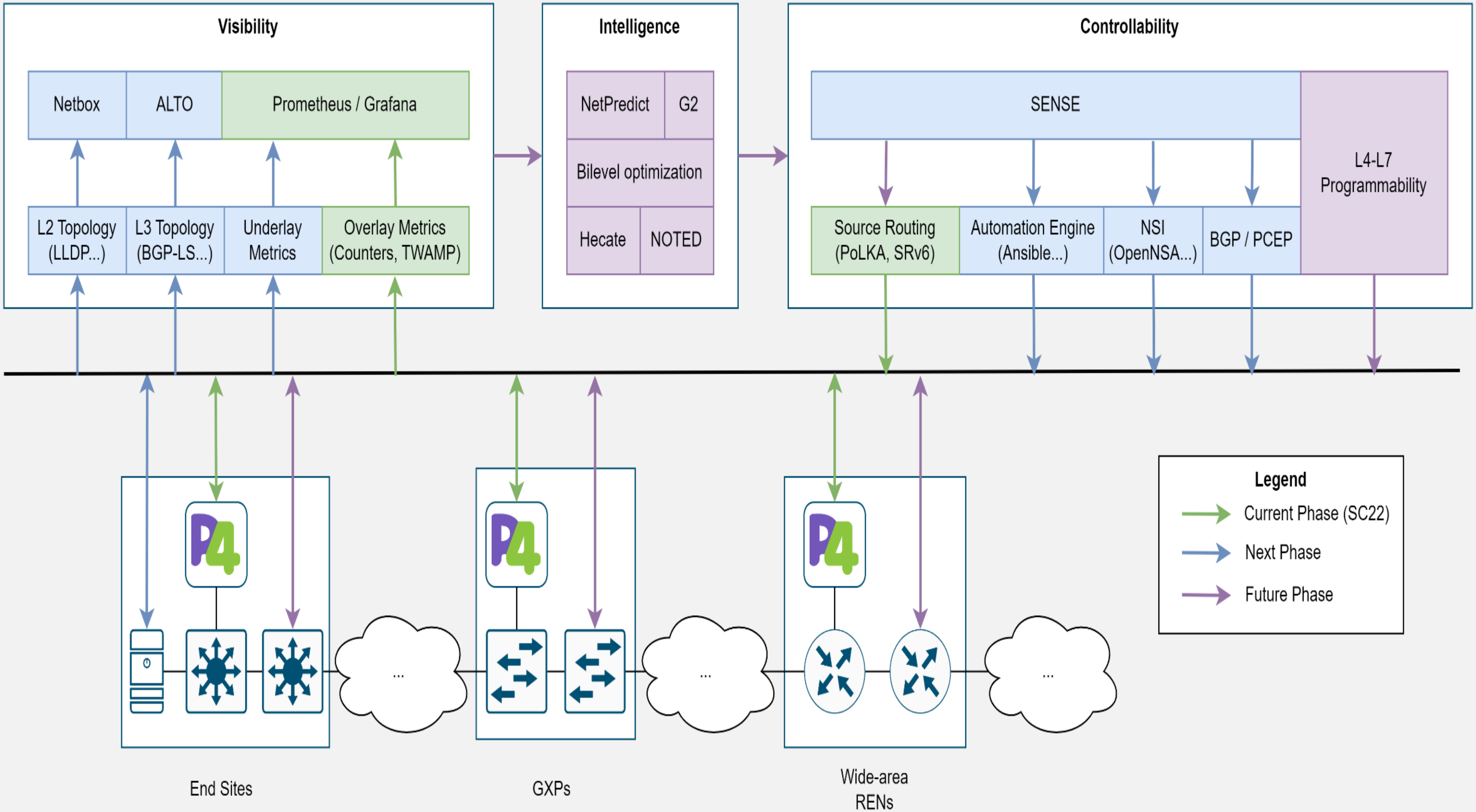
Current (Q1 2023)

- Persistent global L3 overlay network based on P4 switches
- Intercontinental high capacity transfers (100G and over) exploring multiple source routing solutions (Segment Routing and PolKA)
- Management infrastructure and tools used to operate this global network
- Core network based on RARE/FreeRtr and edge networks based also on SONiC

Future (2023-4)

- Capability to support multiple virtual networks that implement different choices of routing stacks, traditional and SDN based:
  on the same devices
- Integration with initiatives for visibility, controllability & intelligence
- Work as a reference state of the art / next generation R&E network

Integration Path Among Initiatives



Self Driving Network

- Adaptive Routing (e.g. Real time data for routing decisions)
- Learns to Avoid Congestion
- Congestion Free ➜ Loss Free
- Towards 100% Utilization
- Proactive Fault Repair

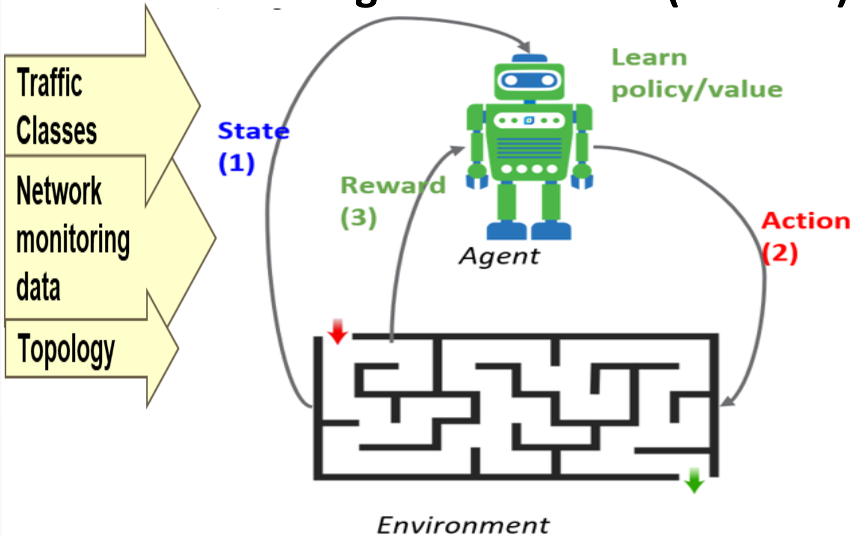
M. Kiran, C, Guok et al (ESnet)

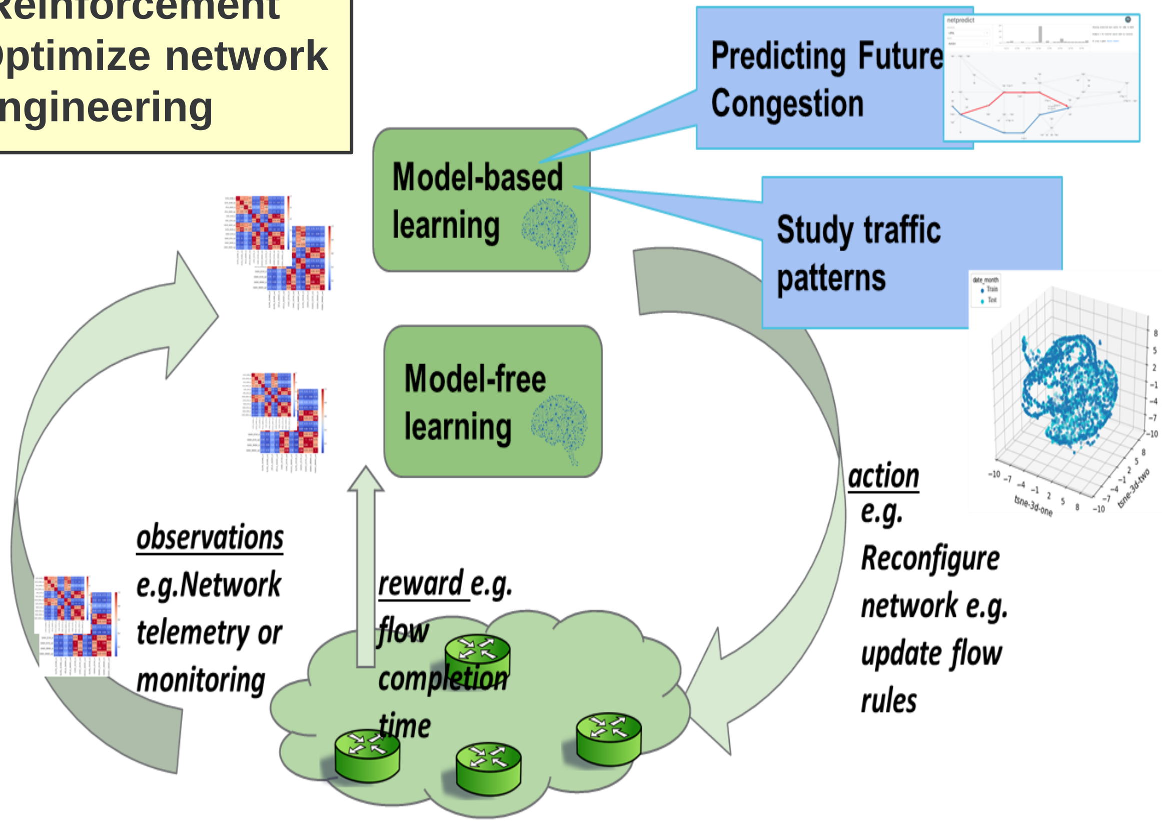
Mariam Kiran (ESnet) et al. Intelligent Networks DOE Project

Self-Driving Network for Science

Use Deep Reinforcement Learning to Optimize network traffic engineering

Intelligent Controller (HECATE)



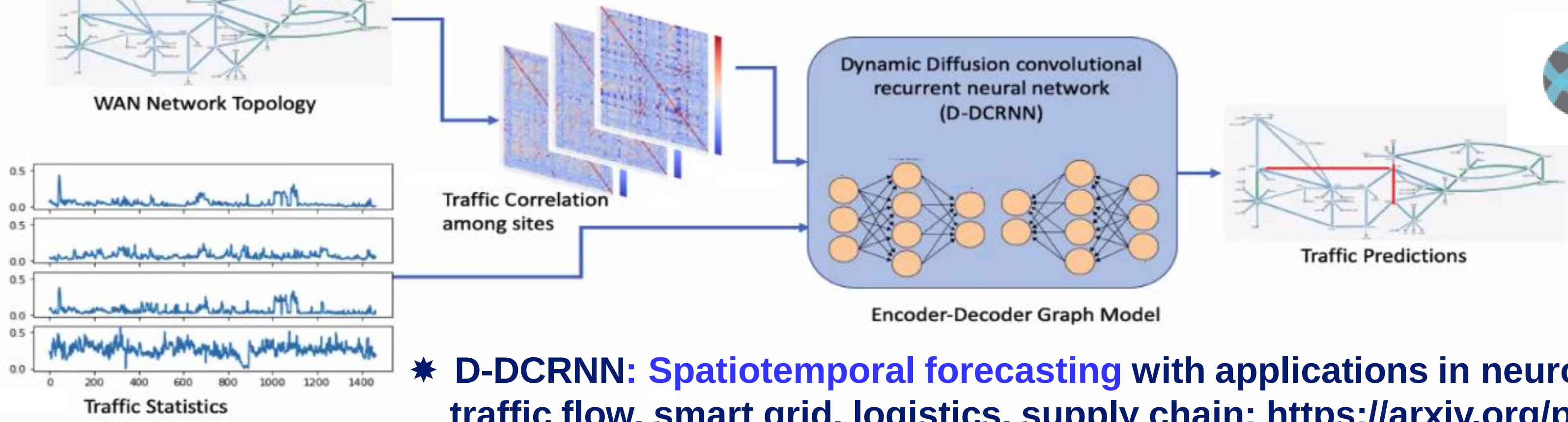


Case Studies:

1. **Model free:** Path selection for large data transfers: better load balancing
2. **Model Free:** Forwarding decisions for complex network topologies: Deep RL to learn optimal packet delivery policies vs. network load level
3. **Model Based:** Predicting network patterns with Netpredict

DAPHNE[2] - Deep and Autonomous High Performance Networks

NetPredict[3]: Deep Learning Model for Network Congestion

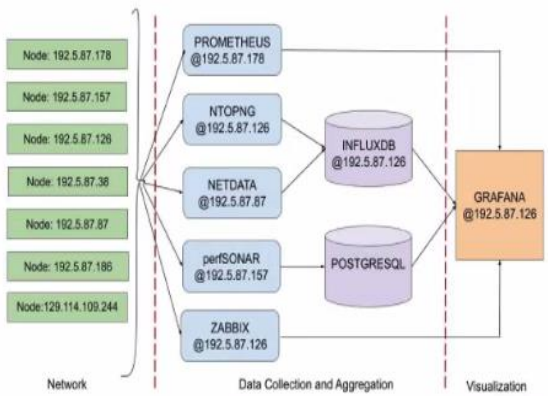


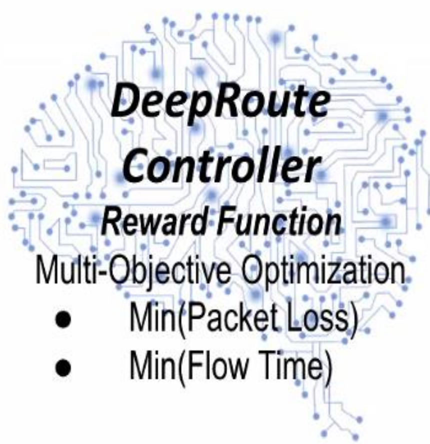
ESnet

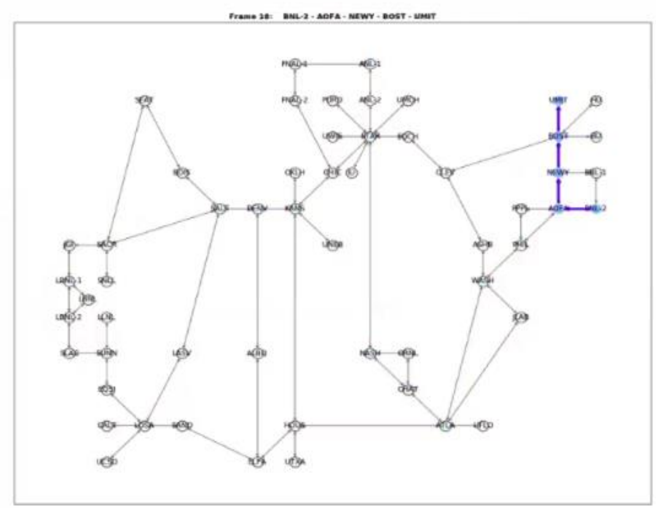
- **D-DCRNN: Spatiotemporal forecasting with applications in neuroscience, climate, traffic flow, smart grid, logistics, supply chain: https://arxiv.org/pdf/1707.01926.pdf**

DeepRoute[4]: Intelligent Traffic Engineering

Network Dashboard collects data and feeds to controller







- **Future: Hooks for a Richer, More Stateful Objective Function (Policy, Priority, Deadlines, Path Quality, Co-Flows...); Event Response; Applications to 5G and Quantum Networks**

Mariam Kiran (ESnet)
mkiran@lbl.gov et al

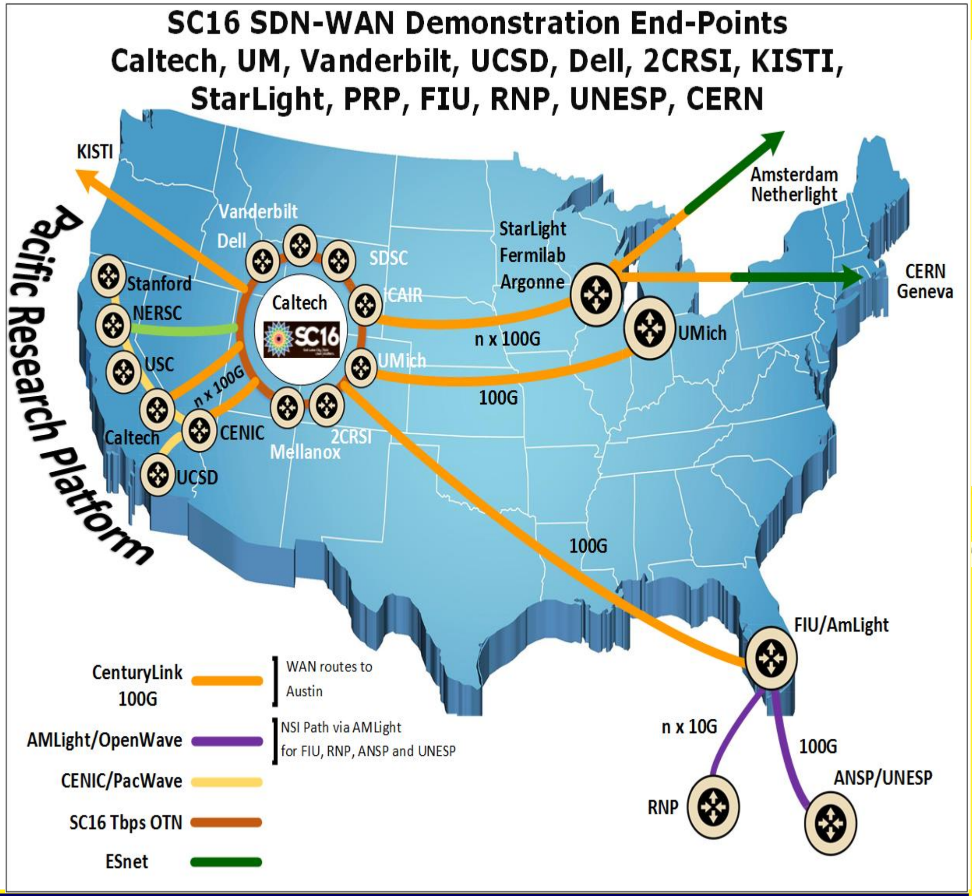
SC15-22: SDN Next Generation Terabit/sec Ecosystem for Exascale Science

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

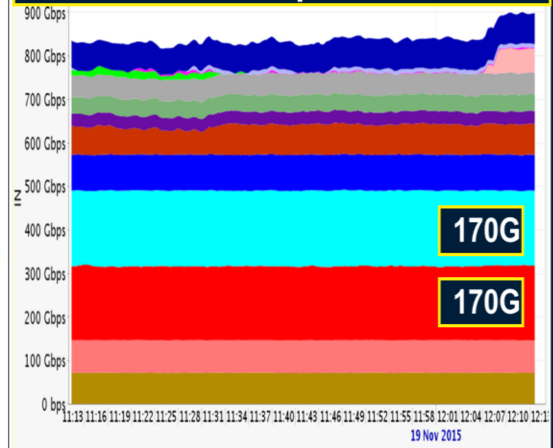
Preview PetaByte Transfers to/ from Site Edges of Exascale Facilities With 100G -1000G DTNs



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 SC18-22: Caltech, Ciena, Scinet, StarLight + Many HEP, Network, Vendor Partners

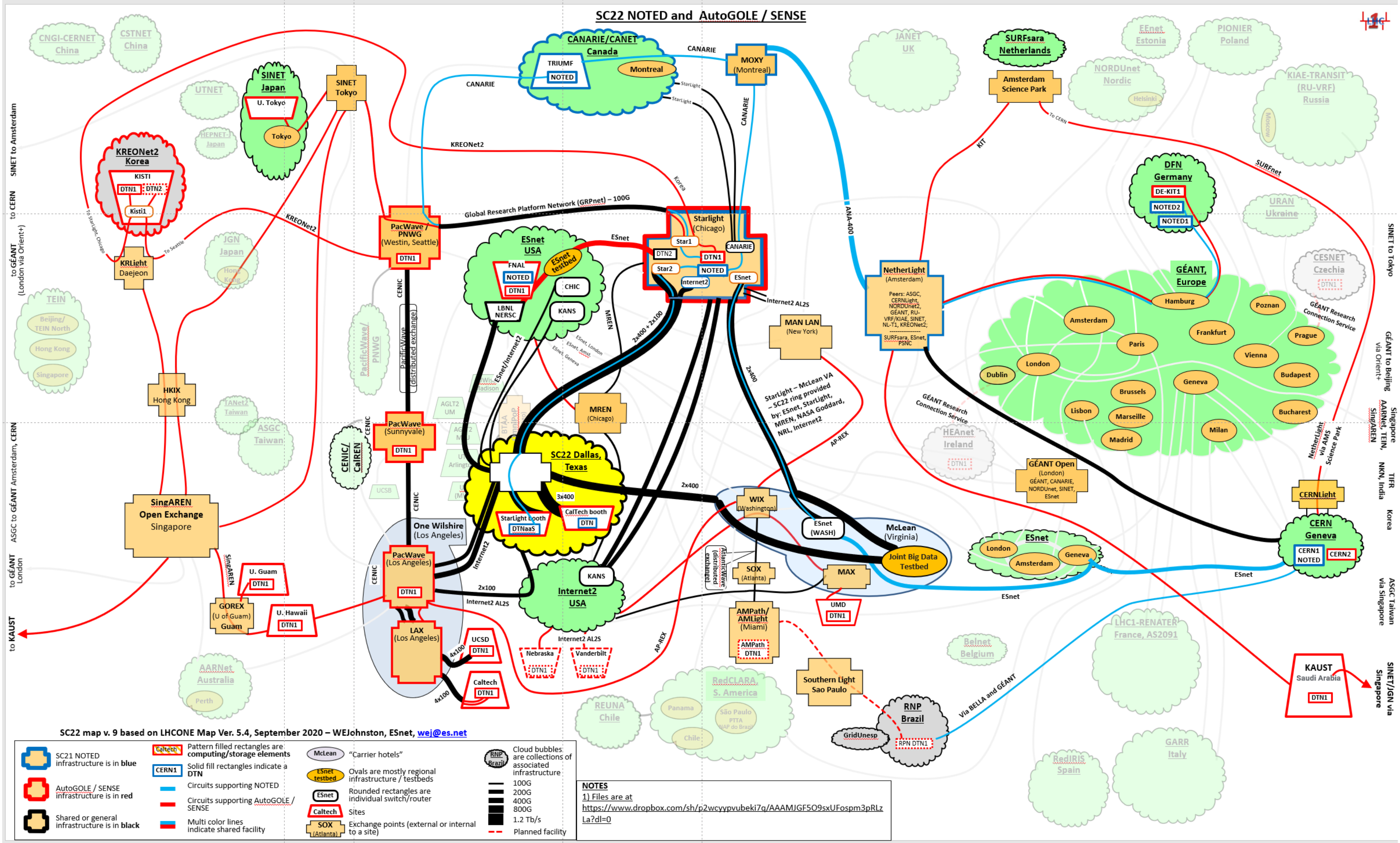
GNA-G DIS WG: Worldwide Partnerships at SC22

Partners: Group Leads and Participants, by Team

- **Caltech HEP:** Harvey Newman (newman@hep.caltech.edu), Justas Balcas (jbalcas@caltech.edu), Raimondas Sirvinskas (raimis.sirvis@gmail.com), Catalin Iordache, Preeti Bhat, Andres Moya, Sravya Uppalapati
- **Caltech IMSS:** Jin Chang (jin.chang@caltech.edu), Azher Mughal (azher@caltech.edu), Dawn Boyd, Larry Watanabe, Don S. Williams
- **UCSD/SDSC/NRP:** Frank Wuerthwein (fkw888@gmail.com), Tom deFanti (tdefanti@eng.ucsd.edu), Larry Smarr, John Graham, Tom Hutton (hutton@ucsd.edu), Dima Mishin, Jonathan Guiang, Diego Davila, Igor Sfiligoi, Aashay Arora,
- **Yale:** Richard Yang (yry@cs.yale.edu), Jensen Zhang
- **Northeastern University:** Edmund Yeh (eyeh@ece.neu.edu), Yuanhao Wu, Volkan Mutlu, Yuezhou Liu
- **Tennessee Tech:** Susmit Shannigrahi (sshannigrahi@tntech.edu), Sankalpa Timilsina
- **UCLA:** Lixia Zhang (lixia@cs.ucla.edu), Jason Cong (cong@cs.ucla.edu), Michael Lo, Sichen Song
- **Fermilab:** Oliver Gutsche (gutsche@fnal.gov), Phil Demar (demar@fnal.gov)
- **ESnet:** Inder Monga (imonga@es.net), Chin Guok (chin@es.net), Tom Lehman (tlehman@es.net), John MacAuley, Xi Yang, Justas Balcas, Mariam Kiran
- **LBNL/NERSC:** Alex Sim (asim@lbl.gov)
- **Nebraska/UNL:** Garhan Attenbury (garhan.attebury@unl.edu)
- **Vanderbilt:** Andrew Melo, andrew.m.melo@accre.vanderbilt.edu
- **CERN:** Edoardo Martelli (edoardo.martelli@cern.ch), Carmen Misa (carmen.misa@cern.ch)
- **Qualcomm Gradient Graph:** Jordi Ros-Giralt (jros@qti.qualcomm.com), Sruthi Yellamraju
- **UFES:** Magnos Martinello, Moises R.N. Ribeiro (moises@ele.ufes.br), Christina Dominicini (cristina.dominicini@ifes.edu.br), Everson Borges (everson@ifes.edu.br), Rafael Guimaraes
- **RNP:** Marcos Schwarz (marcos.schwarz@rnp.br), Leandro Ciuffo (leandro.ciuffo@rnp.br)
- **RENATER/GEANT/RARE:** Frédéric LOUI (frederic.loui@renater.fr)
- **UNESP (SPRACE NCC UNESP):** Sergio Novaes (Sergio.Novaes@cern.ch), Rogerio Iope (rogerio.iope@unesp.br)
- **Rednesp:** Antonio J F Francisco, Ney Lemke **(UNESP)** (ney.lemke@unesp.br), Carlos Antonio Ruggiero **(USP)** (toto@ifsc.usp.br), Jorge Marcos de Almeida (USP) (jorge@usp.br)
- **UERJ:** Alberto Santoro (Alberto.Santoro@cern.ch)
- **George Mason/BRIDGES:** Bijan Jabbari (bjabbari@gmu.edu), Jerry Sobieski, Liang Zhang
- **Xiamen:** Qiao Xiang (xiangq27@gmail.com), Chenyang Huang, Ridi Wen, Yuxin Wang, Jiwu shu
- **Colorado State:** Chengyu Fan (chengy.fan@gmail.com)

- **CENIC:** Louis Fox (lfox@cenic.org), Sana Bellamine (sbellamine@cenic.org), Tony Nguyen
- **Pacific Wave/USC:** Celeste Anderson (celestea@usc.edu)
- **Starlight/MREN/iCAIR:** Joe Mambretti (j-mambretti@northwestern.edu), Jim Chen, Fei Yeh
- **Internet2:** Christian Todorov, (ctodorov@internet2.edu), Rob Vietzke (rvietzke@internet2.edu)
- **AmLight/FIU:** Julio Ibarra (Julio@fiu.edu), Jeronimo Bezerra, Vasilka Chergarova
- **Amlight/ISI:** Heidi Morgan (hlmorgan@isi.edu)
- **Ciena:** Scott Kohlert (skohlert@ciena.com), Rod Wilson
- **KISTI/KREONET:** Buseung Cho (bscho@kisti.re.kr), Mazahir Hussain, Tergel Munkhbat
- **CANARIE:** Thomas Tam (Thomas.Tam@canarie.ca)
- **KAUST:** Alex Moura (alex.moura@kaust.edu.sa), Kevin Sale
- **DE-KIT:** Bruno Hoeft (bruno.hoeft@kit.edu)
- **JPL:** Lee, Carlyn-Ann (Carlyn-Ann.Lee@jpl.nasa.gov)
- **NIST:** Davide Pasavento (davide.pasavento@nist.gov)
- **Hawaii:** Chris Zane (czane@hawaii.edu)
- **SURFNet:** Hans Trompert (hans.trompert@surfnet.nl)
- **CESNET:** Michal Hažlinský, (hazlinsky@cesnet.cz)
- **Clemson:** Cole McKnight (cbmckni@g.clemson.edu)
- **NCHC/TAWREN:** Li-Chi Ku, lku@narlabs.org.tw
- **GNA-G/AArNet:** David Wilde (David.Wilde@aarnet.edu.au)
- **GNA-G AutoGOLE / SENSE WG Members**: https://www.gna-g.net/join-working-group/autogole-sense
- **GNA-G Data Intense Science WG Members:** https://www.gna-g.net/join-working-group/data-intensive-science/
- **STORDIS:** Waldemar.Scheck@stordis.com

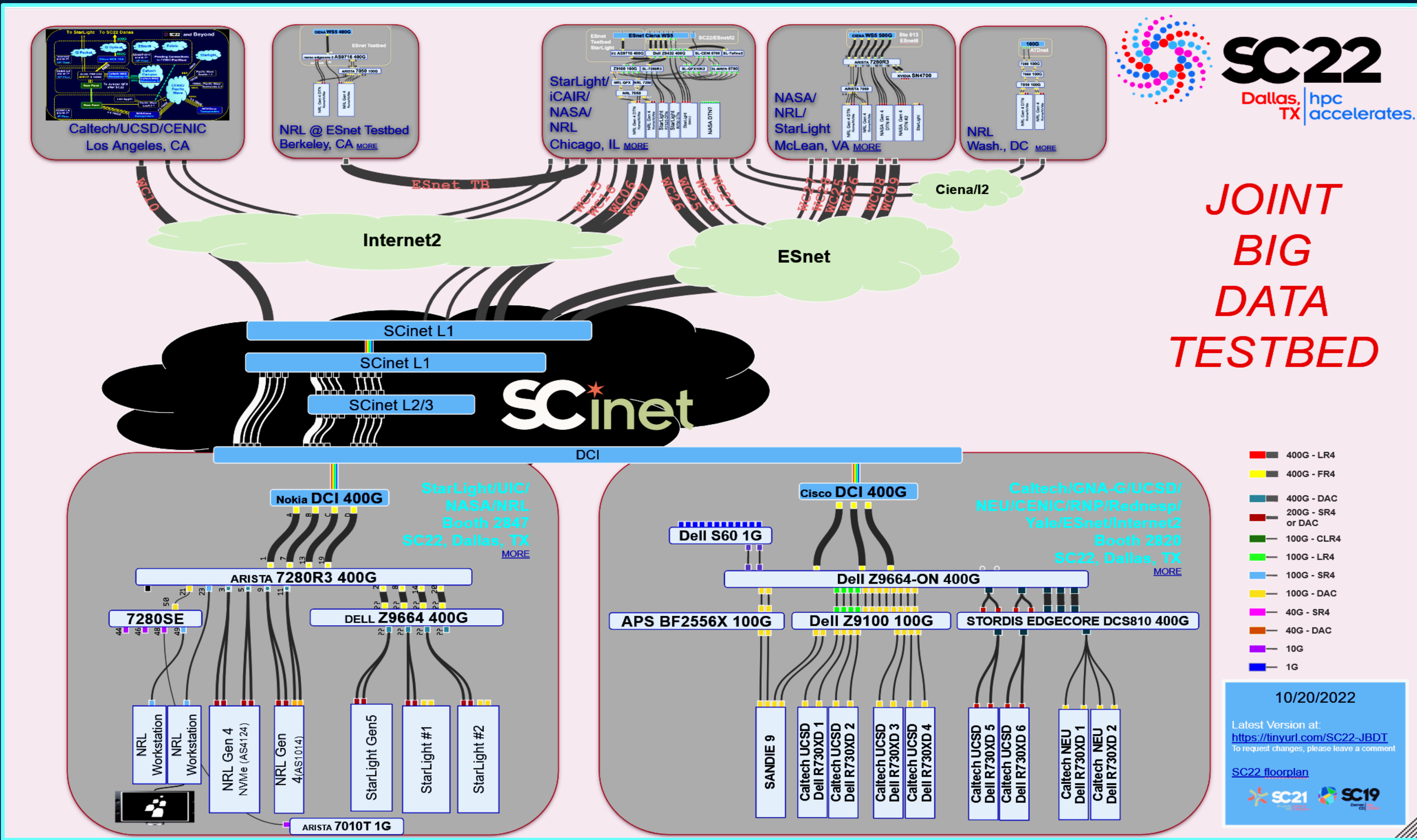
SC22 NREs: AutoGOLE / SENSE and NOTED (Draft)



SC21: Global footprint. Multiple 400G Optical Wide Area Links Bill Johnston 9/23/22

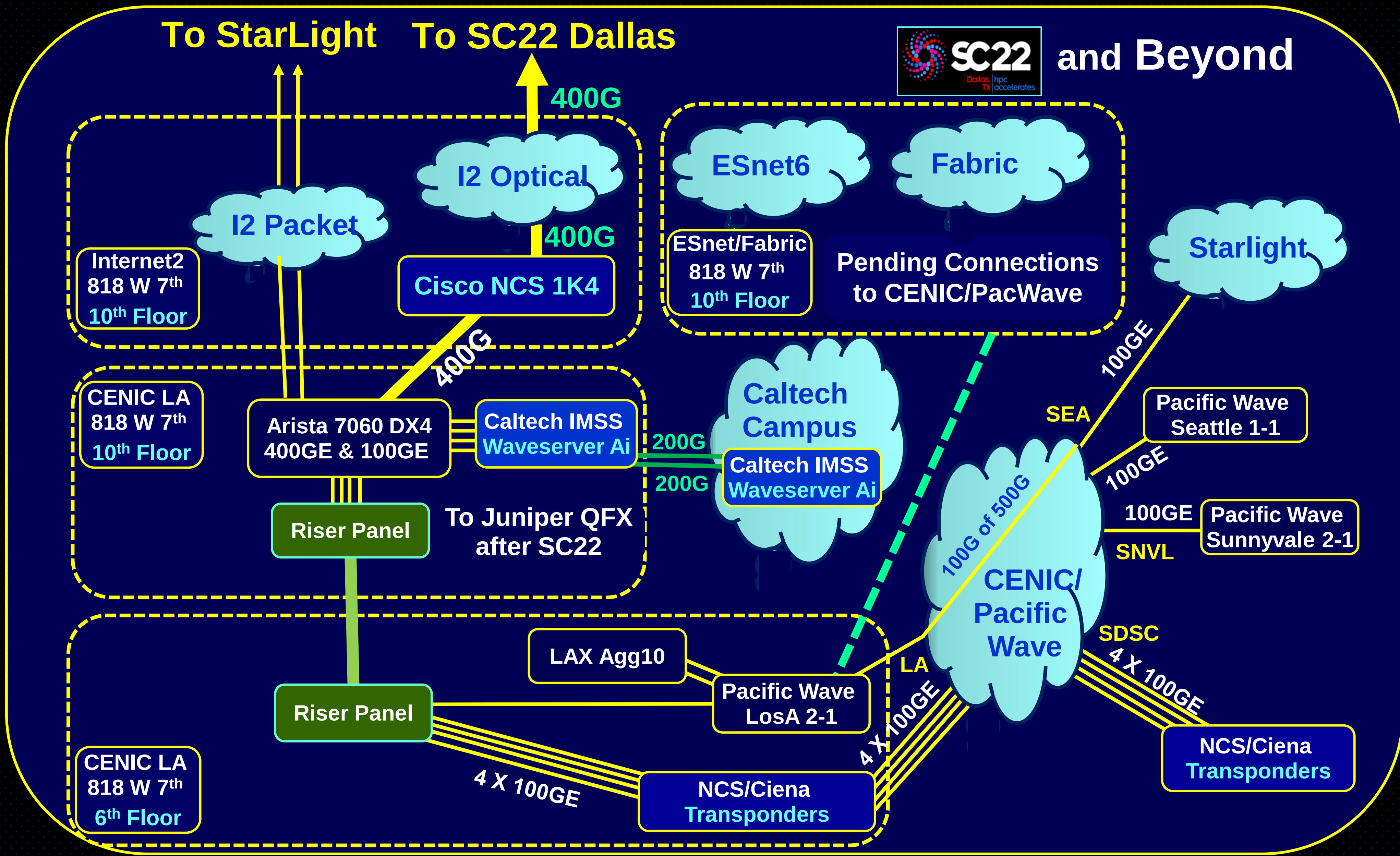
SC22: Global footprint. Terabit/sec Triangle Starlight – McLean – Dallas; 400G to LA; 4 X 100G to Caltech and UCSD/SDSC; 3 X 400G to the Caltech Booth

Caltech, StarLight, NRL and Partners at SC22



Microcosm: Creating the Future of SCinet and of Networks for Science

A New Generation Persistent 400G/100G Super-DMZ: CENIC, Pacific Wave, ESnet, Internet2, Caltech, UCSD, StarLight ++



Global Petascale to Exascale Workflows for Data Intensive Sciences Accelerated by Next Generation Programmable Network Architectures and Machine Learning Applications

Network Research Exhibition NRE-19. Abstract:
https://www.dropbox.com/s/qcm41g7f7etjxvy/SC22_NRE_GlobalPetascaleWorkflows_V8072022.docx?dl=0

- A Vast Partnership of Science and Computer Science Teams, R&E Networks and R&D Projects
  - Convened by the GNA-G Data Intensive Sciences WG
- **Mission**
  - Meeting the challenges faced by leading edge data intensive experimental programs in high energy physics, astrophysics, genomics and other fields of data intensive science
  - Clearing the path to the next round of discoveries
- **Demonstrating a wide range of the latest advances in:**
  - Software defined and Terabit/sec networks
  - Intelligent global operations and monitoring systems
  - Workflow optimization methodologies with real time analytics
  - State of the art long distance data transfer methods and tools, local and metro optical networks and server designs
  - Emerging technologies and concepts in programmable networks and global-scale distributed system

Global Petascale to Exascale Workflows for Data Intensive Sciences

- Advances Embedded and Interoperate within a 'composable' architecture of subsystems, components and interfaces, organized into several areas:
  - **Visibility:** Monitoring and information tracking and management including IETF ALTO/OpenALTO, BGP-LS, sFlow/NetFlow, Perfsonar, Traceroute, Qualcomm Gradient Graph congestion information, Kubernetes statistics, LibreNMS, P4/Inband telemetry
  - **Intelligence:** Stateful decisions using composable metrics (policy, priority, network- and site-state, SLA constraints, responses to 'events' at sites and in the networks, ...), using NetPredict, Hecate, RL-G2, Yale Bilevel optimization, Coral, Elastiflow/Elastic Stack
  - **Controllability:** SENSE/OpenNSA/AutoGOLE, P4/PINS, segment routing with SRv6 and/or PolKA, BGP/PCEP
  - **Network OSes and Tools:** GEANT RARE/freeRtr, SONIC, Calico VPP, Bstruct-Mininet environment, ...
  - **Orchestration:** SENSE, Kubernetes (+k8s namespace), dedicated code and APIs for interoperation and progressive integration

Next Generation System for Data Intensive Sciences

- Overarching Concept: Consistent Network Operations:
  - Stable load balanced high throughput workflows cross optimally chosen network paths
  - Provided by autonomous site-resident services dynamically interacting with network-resident services
  - Responding to (or negotiating with) site demands from the science programs' principal data distribution and management systems
  - Up to preset or flexible *high water marks* to accommodate other traffic
- Architecture: Data Center Analogue
  - Classes of "Work" (work = transfers, or overall workflow), defined by task parameters and/or priority and policy
  - Adjusts rate of progress in each class to respond to network or site state changes, and "events"
  - Moderates/balances the rates among the classes to optimize a multivariate objective function with constraints

Next Generation System for Data Intensive Sciences

- The development of effective optimization methods, and multidimensional, real-world metrics
  - Are themselves challenging, groundbreaking activities
- Strategic Aim: Compatible coexistence of programmable goal-oriented networks, and production networks
  - Simultaneously meeting the needs of the leading edge science programs and the at-large A&R communities

Closing the Digital Divide

Brazil as a Global Model Example

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

Alberto Santoro and Harvey B Newman with RNP

from 2002 Onward

ICFA SCIC Perspective and Outlook

- ***Missions***
  - *Inform and enable the global community to use networks effectively in support of the communities' science goals*
  - *Track advanced computing, storage, network and associated software technologies; highlight opportunities and coming issues*
  - *With a mission focus on major programs LHC to HL-LHC, LSST, SKA, DUNE et al*
  - *Track and help understand and set requirements via both community meetings (e.g. LHCONE/LHCOPN) and agency reviews (e.g. ESnet in July)*
  - *Bring Issues to the attention of ICFA*
- ***Activities***
- *Work with R&E network partners to help develop the continental, transoceanic and regional network infrastructures*
- *Beyond the basic infrastructures: Formation of a global fabric supporting data intensive research Learning from and going beyond the LHCONE experience*
- *Developing integrated systems including networks as a first class resource, across a global footprint*
- ***Engagement***
- *With all of the experiments' computing managements, the major R&E network organizations, key network projects supporting major science programs Also leading edge development projects: SENSE, P4 Global Tesbed, NRP, GRP, etc.*
- *Engage in proof of concept, prototype, pre-production exercises and demonstrations to test and prove requirements*

SCIC Work Areas

- *Closing the Digital Divide (A. Santoro, WG Chair)*
  - 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 networking: Multiple Network OS's
  - 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, SKA, VRO, et al.)

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

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

NEWS:

Bulletin: ONE TWO

WELCOME BULLETIN

General Information

Registration

Travel Information

Hotel Registration

Tutorials

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

SPONSORS

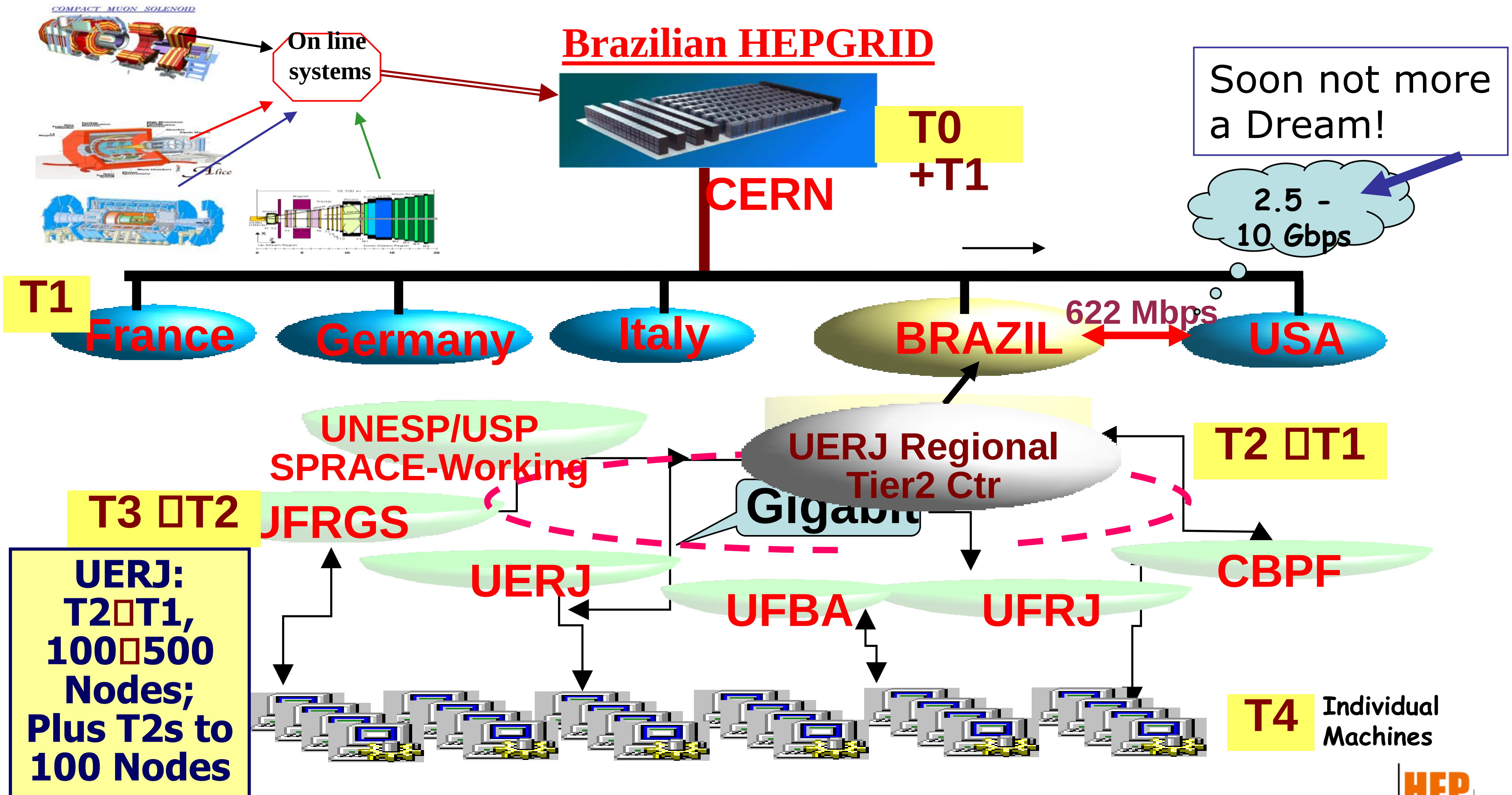
CLAF CNPQ FAPERJ UERJ

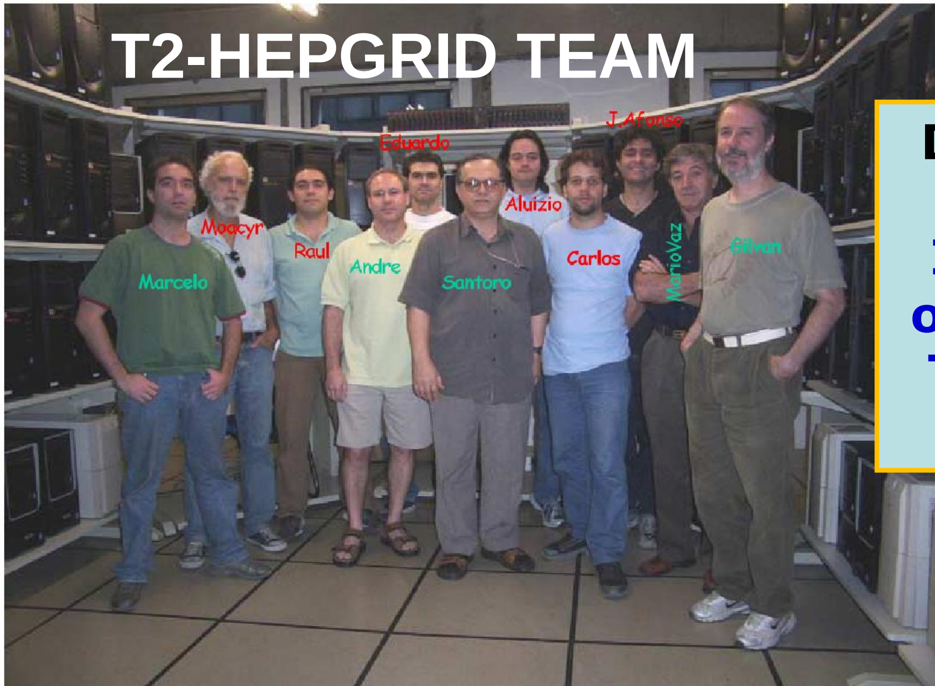
Sessions & Tutorials Available (w/Video) on the Web

SLOW? YES! But we continue our main project:

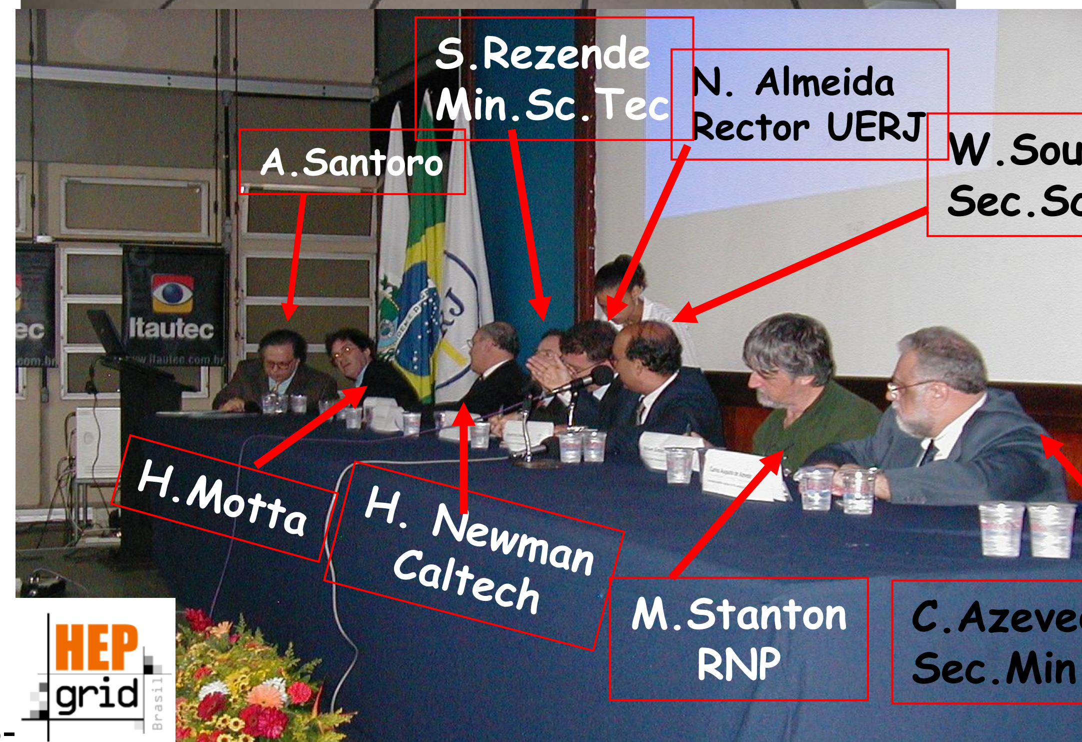
The purpose of HEPGRID-CMS/BRAZIL is to become

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

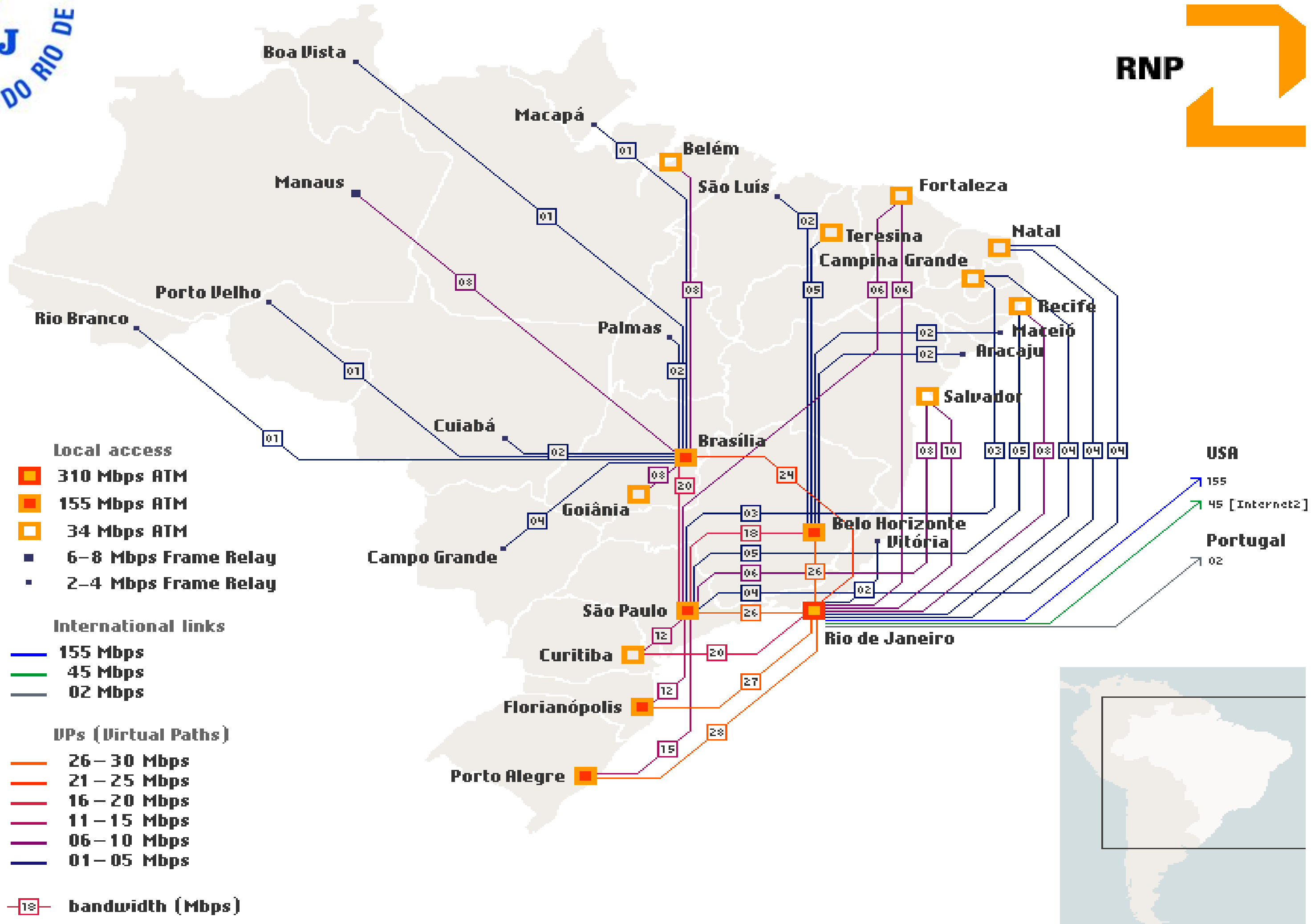




December 20, 2004
Inauguration of the HEPGrid Tier2 at UERJ In Rio



Brazil: RNP in Early 2004



Beginning a Process of Innovation: Reverse Auction for 2.5 and 10 Gbps in 2005

HEPGRID BRAZIL -GRID and Digital Divide

HEPGRID BRAZIL -GRID and Digital Divide

International Workshop on African Research & Education Networking

Alberto Santoro
UERJ - Brazil

September 25-27 2005
CERN

Outline

I – Introduction

II - T2–HEPGRID BRASIL – Other News

III - Conclusion

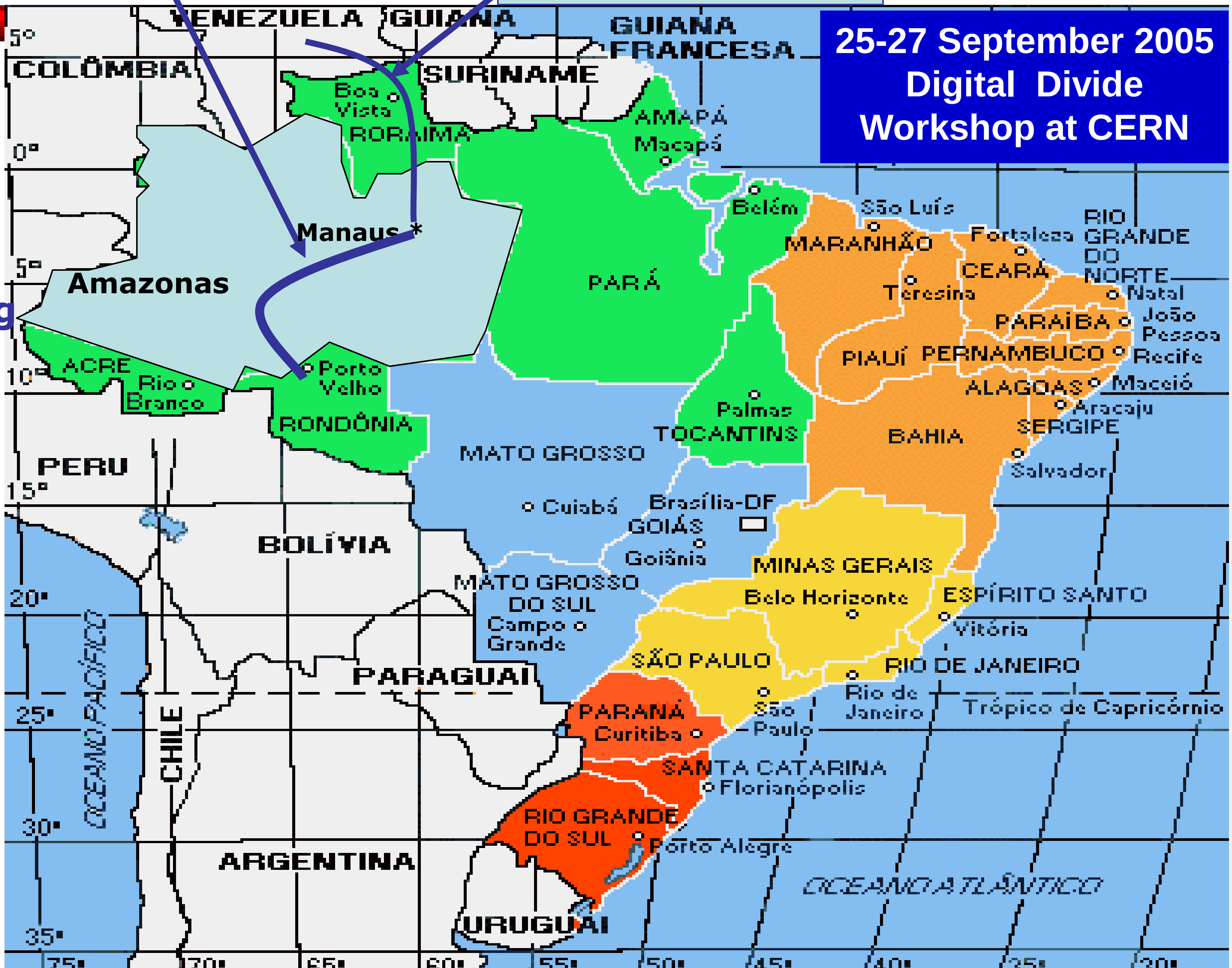
Suggestion#2 for Fiber Installation

Suggestion#1 for Fiber Installation

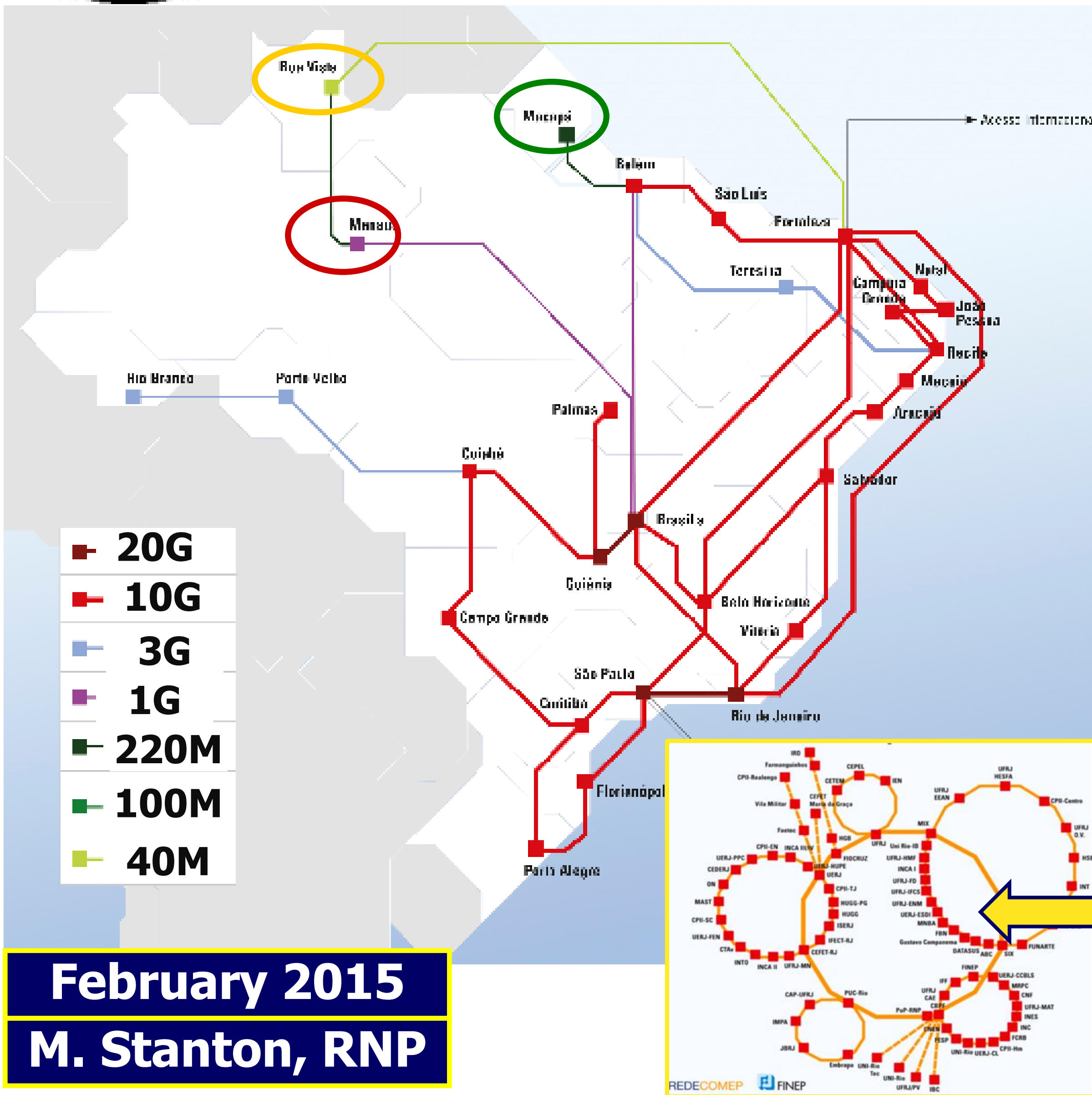
25-27 September 2005 Digital Divide Workshop at CERN

RNP + LOCAL Network are paying Attention to this Region Also

In 2005



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

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

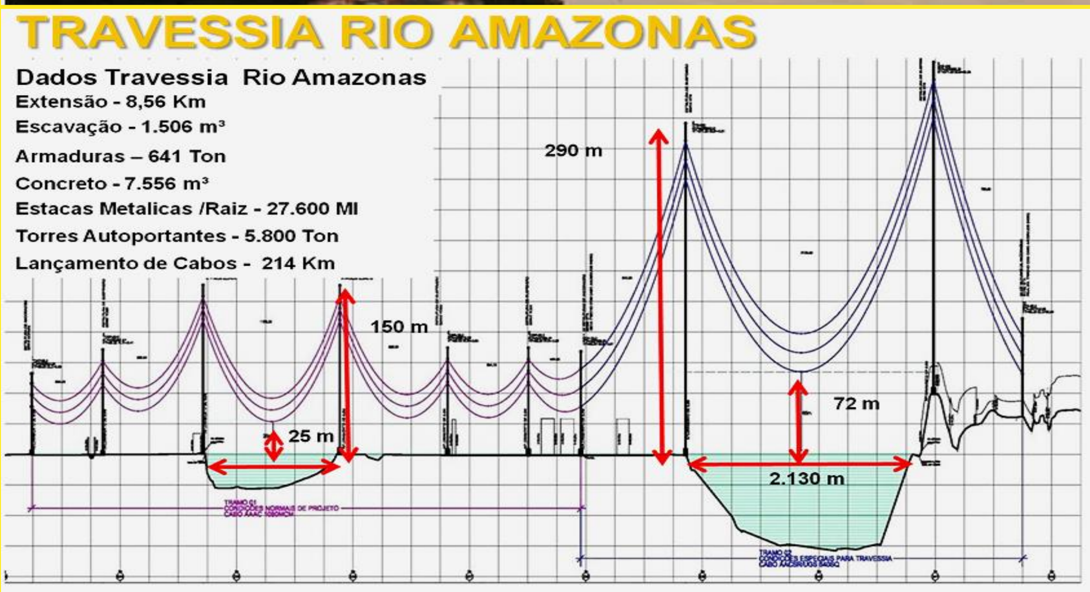
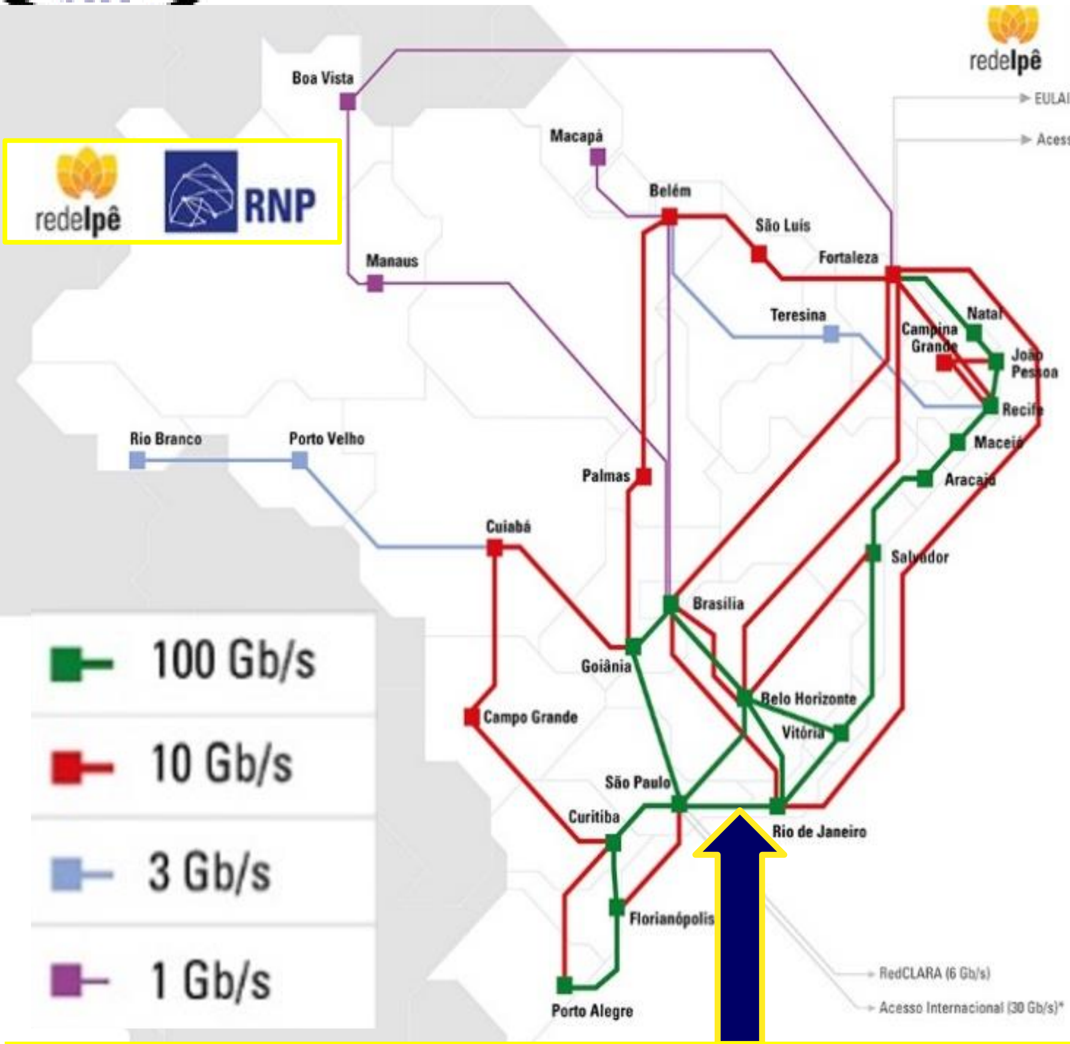


Figure N+3: Details of the Amazon crossing at Jurupari.

- 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**

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



Phase 7 RNP Backbone with 100G Core from 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 acquired long-term rights to an extensive optical fiber infrastructure for the 100G transition

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

2016-19

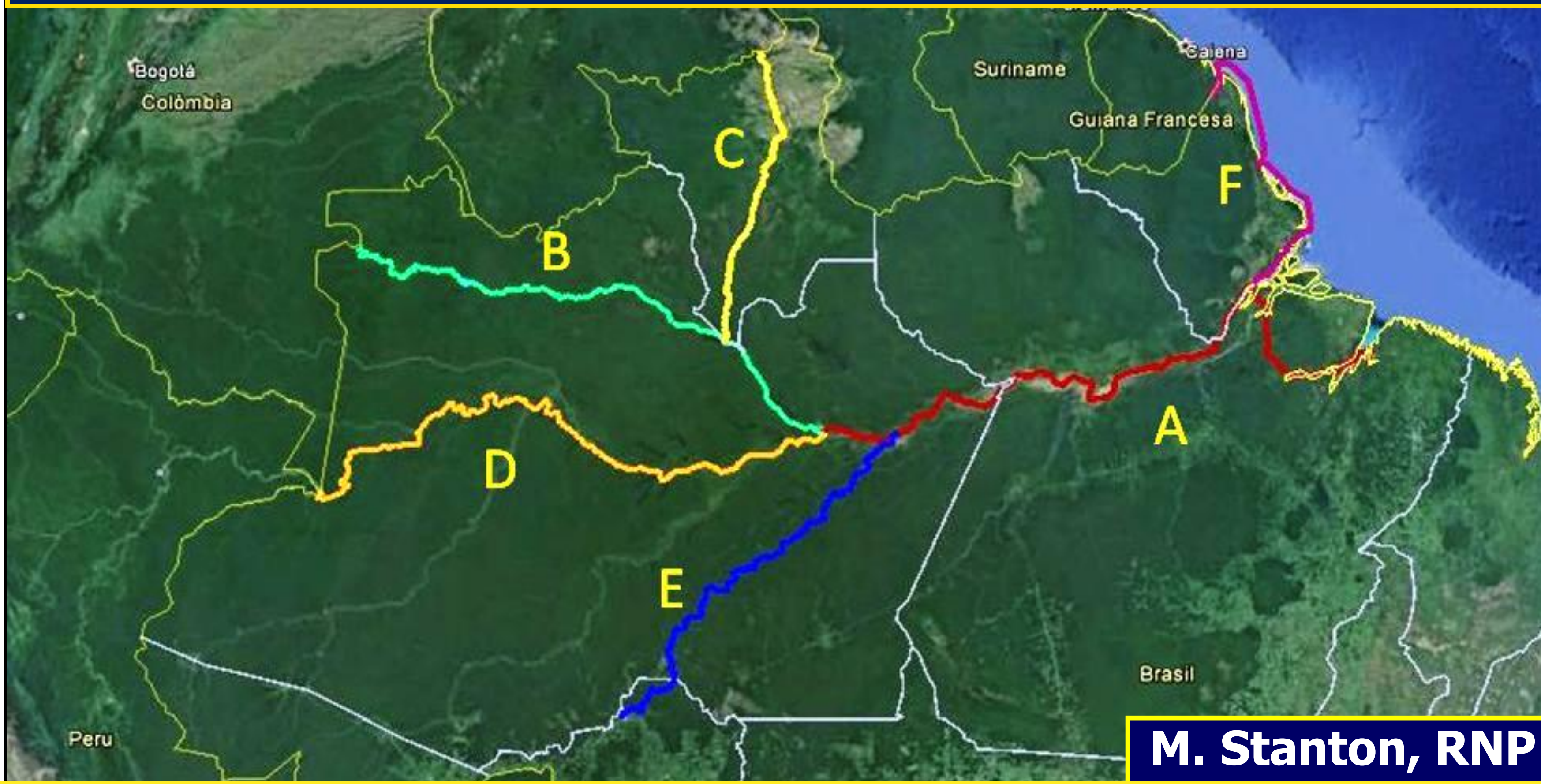
700 km Solimões R: Manaus – Tefé

100 km Negro River: Manaus – Novo Airão

Next: Tefé – Tabatinga Interconnect Brazil and Peru

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

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



M. Stanton, RNP

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.

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

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

2018-19

Yurimaguas is on the Peruvian terrestrial National Backbone FO Network (RDNFO) being built by FITEL to support digital inclusion

M. Stanton RNP

Purple lines are proposed subfluvial fiber

Closing the Digital Divide
Brazil as a Global Model Example

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

LISHEP 2023
International School on High Energy Physics

LISHEP 2021
International School on High Energy Physics

Rio +Worldwide

UFBA
LISHEP 2018
International School on High Energy Physics

Salvador

LISHEP 2015

Manaus

On a *River of Discovery*

Alberto Santoro and Harvey B Newman
from 2002 Onward