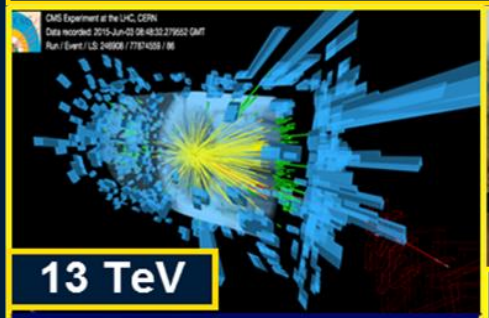
Global Network Advancement Group

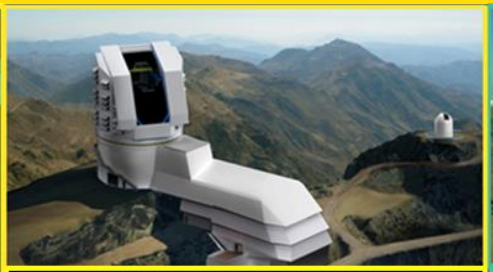
Next Generation Network-Integrated System for Data Intensive Sciences

SC24 ATLANTA NOV 17–22

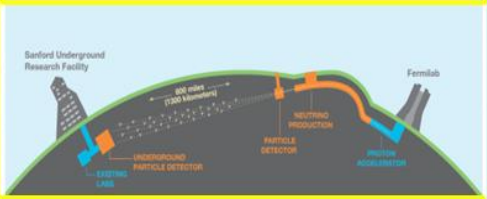
Booth 845



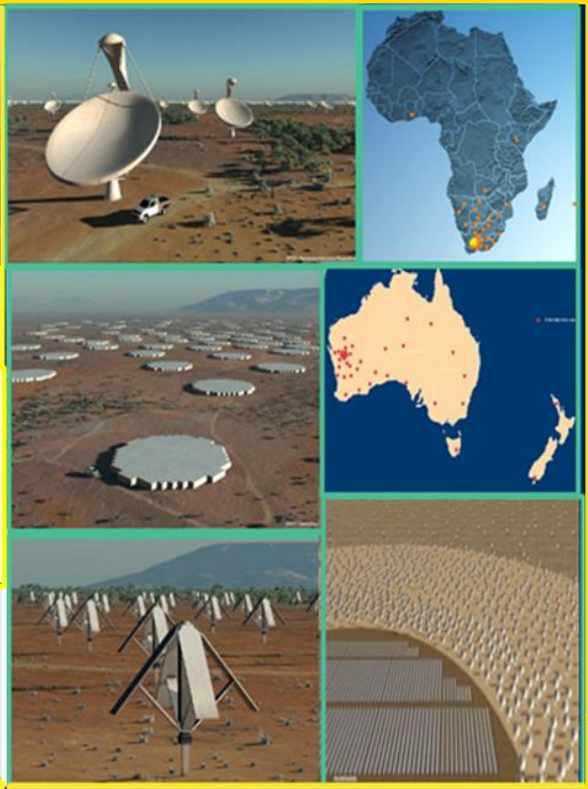
LHC



Rubin Observatory



LBNF/DUNE



SKA

LHC Run 3 and HL-LHC

Rubin Observatory

SKA

BioInformatics

Earth Observation

Gateways to a New Era

GNA-G Global Network Advancement Group

SC24 Network Research Exhibition

NRE-13 and 9 Partner NREs

INDIS Workshop November 18, 2024

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: ~85% of Data Still to be Taken**
- **Greater Intensity: Upgraded detectors for more complex events**
- ***To 5X Data Taking Rate in 2029-40***

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

Experiment

~PByte/sec

Online System

~1000 MBytes/sec

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

Tier 0

CERN Center
100 PBs of Disk;
Tape Robot

Tier 1

N X100 Gbps

London

Paris

Bologna

Chicago

10 to 100 Gbps

Tier 2

CACR

Center

Center

Center

Tier 3

Institute

Physics data cache

10 to 100 Gbps

Tier 4

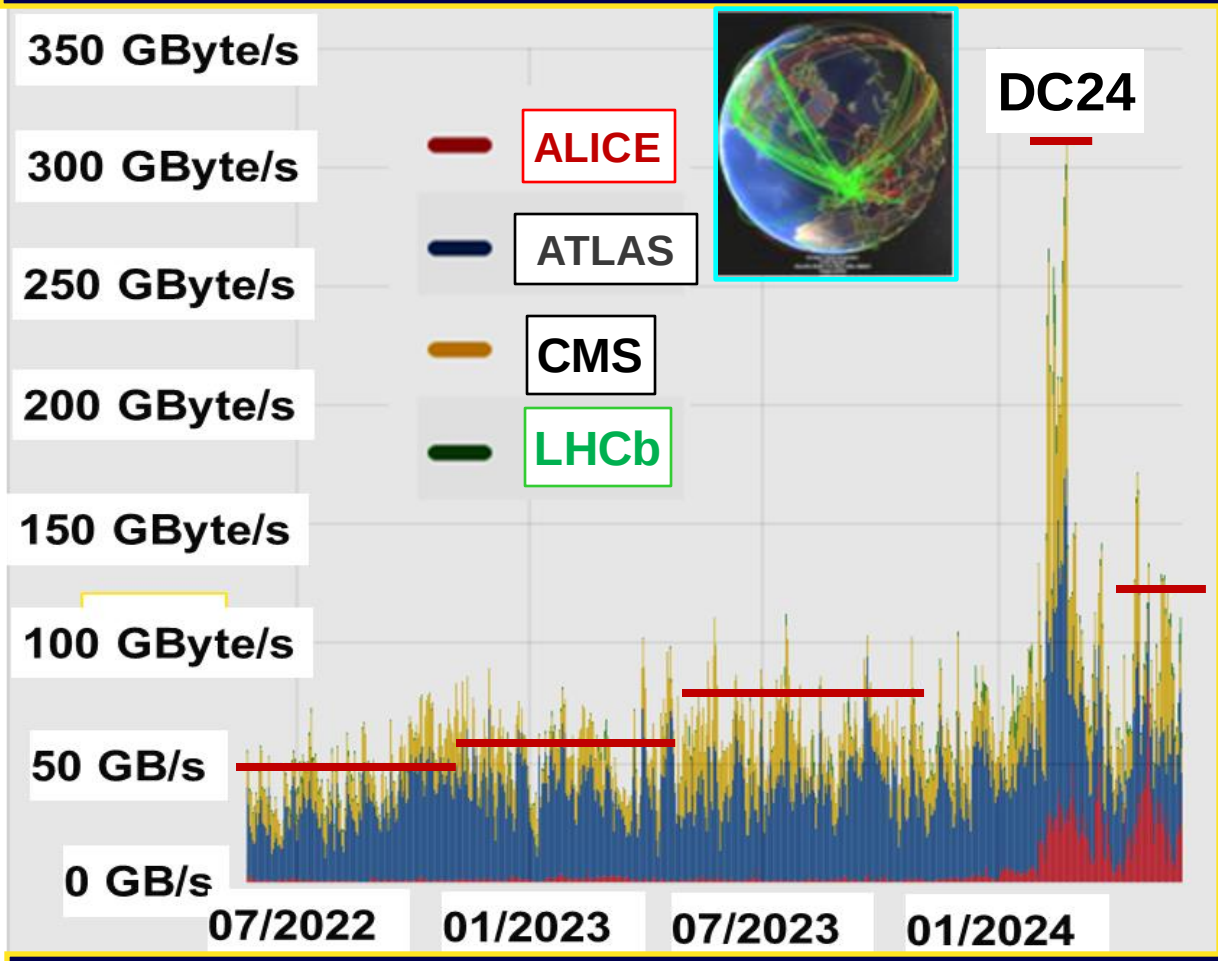
Workstations

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

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

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

WLCG Transfers Dashboard: Throughput June 2022 – May 2024



70-150 GBytes/s Weekly Avg
To 170+ GBytes/s Daily Avg

Complex Workflow

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

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

~1.8 to 2X Growth in 12 Months: 100 to 1000X Per Decade Equivalent (?)

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

Next Generation Network-Integrated System

- **Top Line Message:** To realize the physics discovery potential and meet the challenges of data intensive sciences, we need a new dynamic system which:
  - Coordinates worldwide networks as a first class resource along with computing and storage, across and among world regions
- Follows a systems design approach: A global fabric that flexibly allocates, balances and makes best use of 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
- Leverages the worldwide move towards a fully programmable ecosystem of networks and end-systems (P4, SONIC; PolKA, SRv6), plus operations platforms (OSG, NRP; global SENSE Testbed, Global P4 Lab, FABRIC)
- Simultaneously supports the LHC experiments, VRO other data intensive programs and the larger worldwide academic and research community
- The LHC experiments together with the GNA-G and its Working Groups, the WLCG and the worldwide R&E network community are key players
  - Together with the major programs: LHC, LBNF/DUNE, VRO, SKA
- **SC24 is a Major Milestone, and a Leap Forward Towards this Goal**

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

leadershipteam@lists.gna-g.net

Buseung Cho
KISTI (Korea)

Ivana Golub
PSNC, GEANT (Europe)

Harvey Newman
Caltech (US)

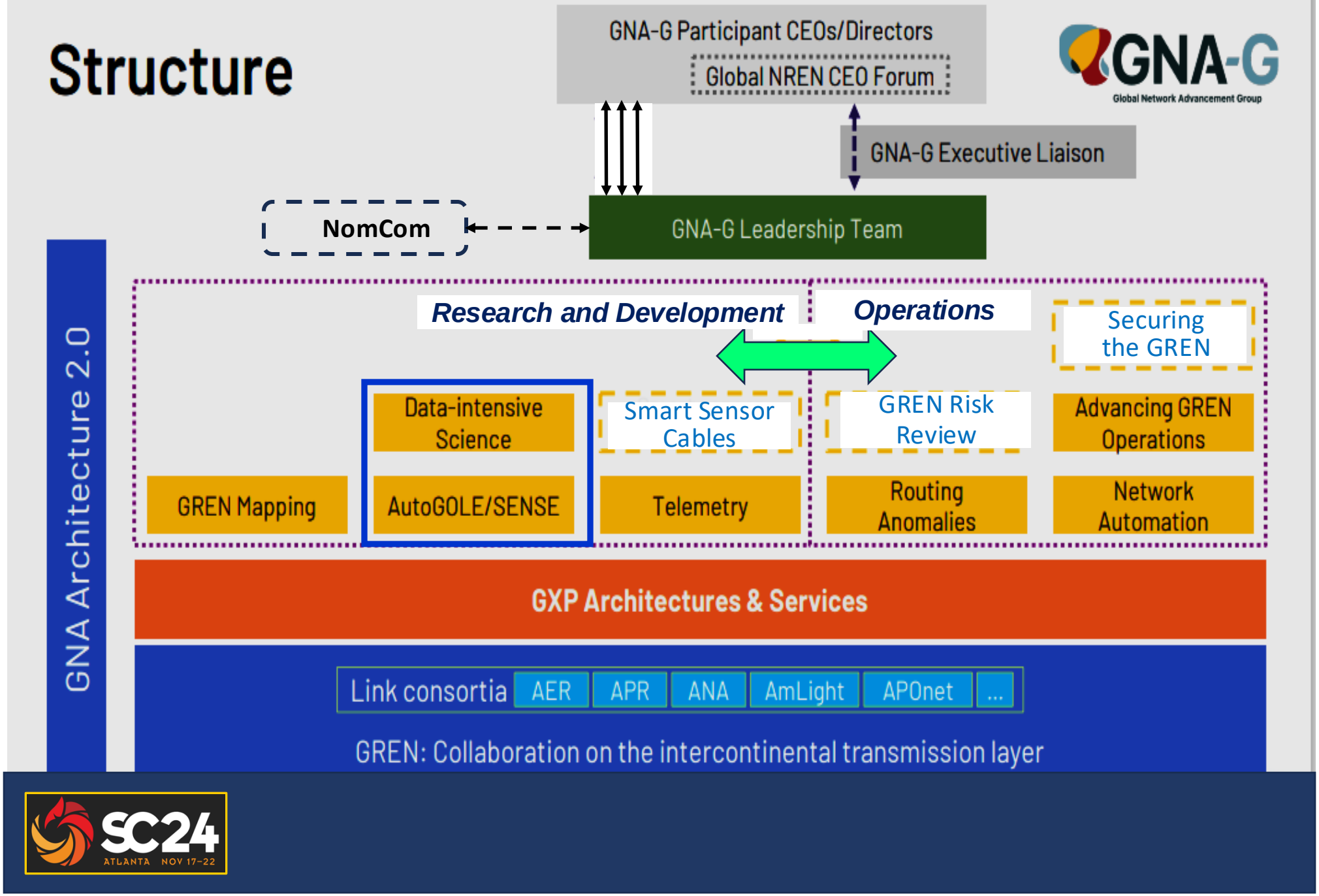
David Wilde, Chair
Aarnet (Australia)

Alex Moura
KAUST (Saudi Arabia)

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

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

Structure


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

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

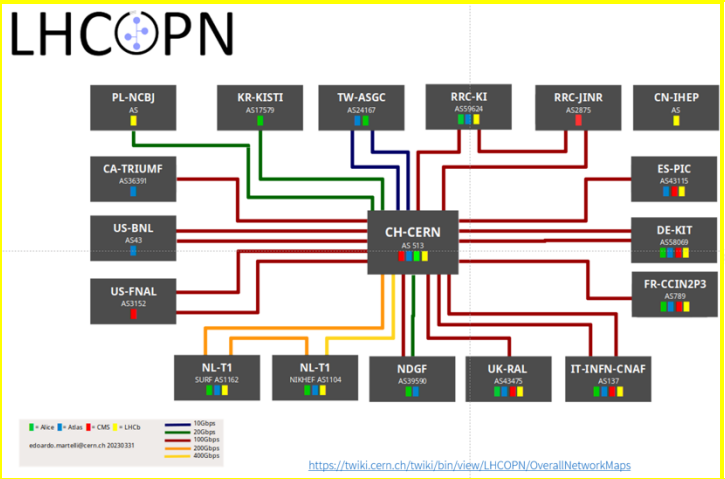
- **To develop a persistent global testbed as a platform, to foster ongoing developments among the science and network communities**

- **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**

The Core of LHC Networking

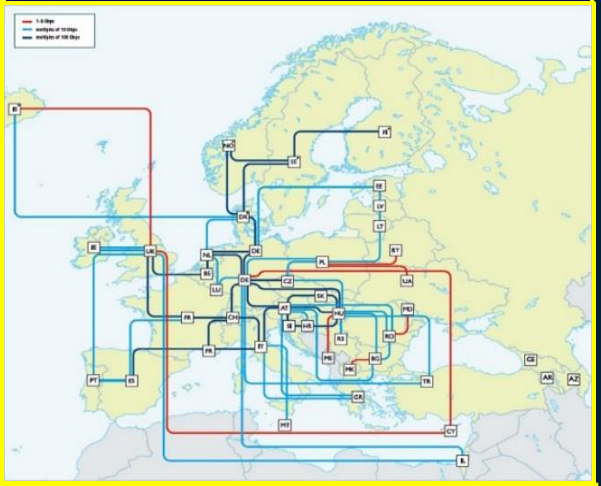
LHCOPN, LHCONE, GEANT, ESnet, Internet2, CENIC…

LHCOPN: Simple & Reliable
Tier0+1 Ops

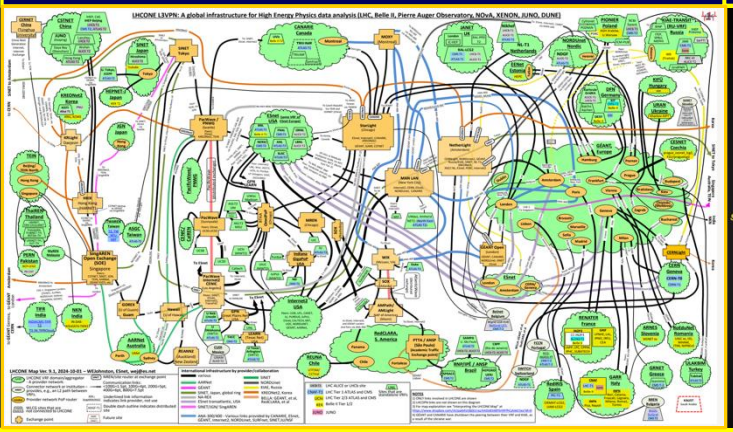


Internet2n

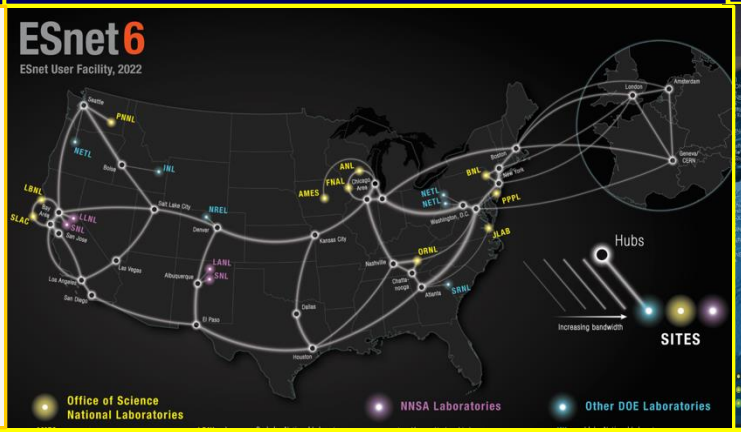
GEANT



LHCONE VRF: 170 Tier2s ++



ESnet6



CENIC and NRP

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

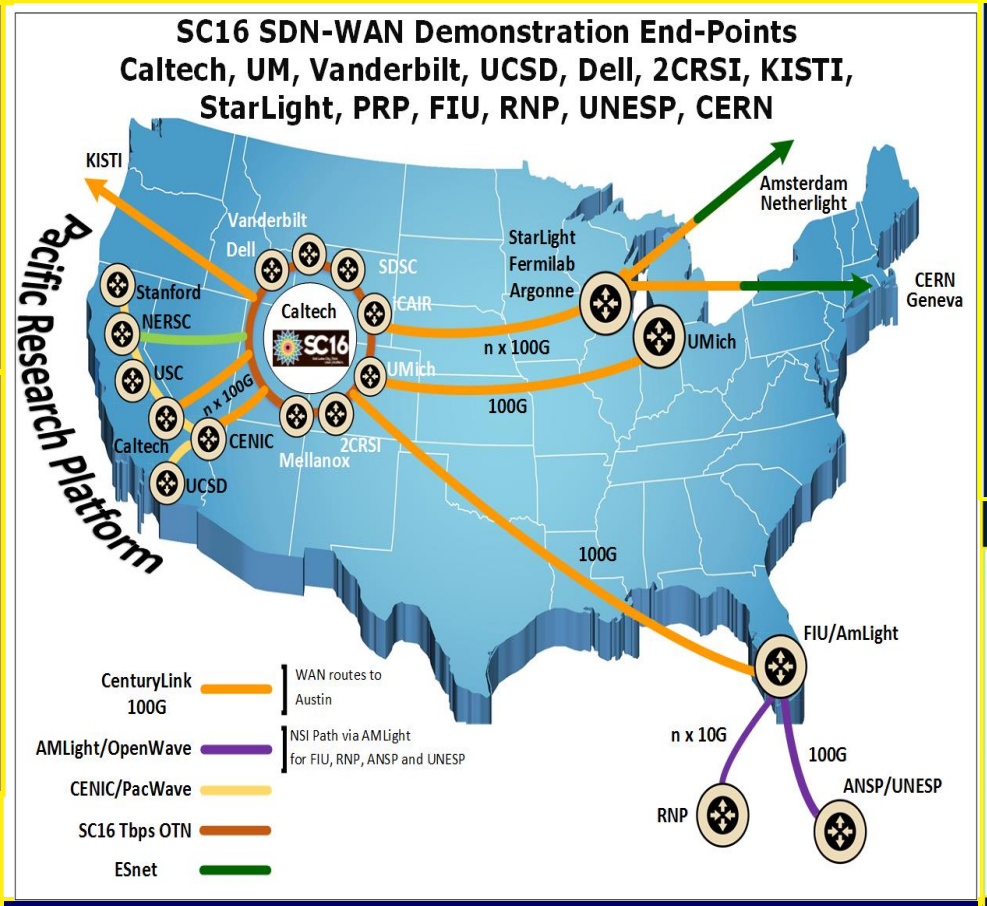
SC15-24: SDN Next Generation Terabit/sec Ecosystem for Exascale Science

supercomputing.caltech.edu

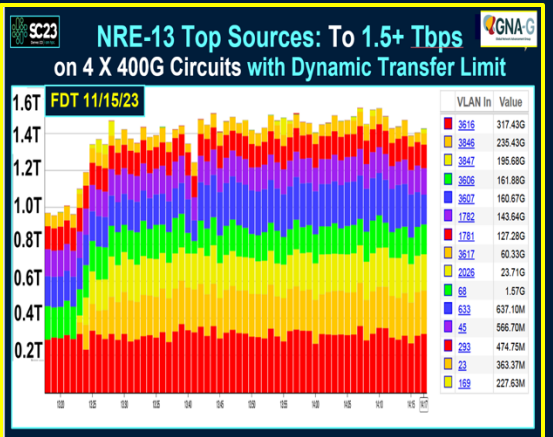
SDN-driven flow steering, load balancing, site orchestration Over Terabit/sec Global Networks

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

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



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



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 SC18-24: Caltech, Ciena, SCInet, StarLight + Many HEP, Network, Vendor Partners

Global Network Advancement Group: Next Generation Network-Integrated System for Data Intensive Sciences

Network Research Exhibition NRE-13

- A Vast *Partnership* of Science and Computer Science Teams, R&E Networks and R&D Projects; Convened by the GNA-G DIS WG; with GRP, AmRP, NRP
- Mission: Demonstrate the road ahead
  - Meet the challenges faced by leading-edge data intensive programs in HEP, astrophysics, genomics and other fields of data intensive science;
    ✸ *Compatible with other use*
  - Clearing the path to the next round of discoveries
- Demonstrating a wide range of 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 systems
- Hallmarks: Progressive multidomain integration; compatibility internal + external; *A comprehensive systems-level approach*

SC24 ATLANTA NOV 17–22

GNA-G Global Network Advancement Group

Worldwide Partnership at SC24 and Beyond

GNA-G
Global Network Advancement Group

SC24
ATLANTA NOV 17–22

Global Petascale to Exascale Workflows for Data Intensive Sciences

CALIFORNIA INSTITUTE OF TECHNOLOGY
1891
HIGH ENERGY PHYSICS + TERABIT/SEC NETWORKS – THE FRONTIERS OF EXASCALE SCIENCE – SUPERCOMPUTING.CALTECH.EDU

Accelerated by Next Generation Programmable SDN Architectures and Machine Learning Applications

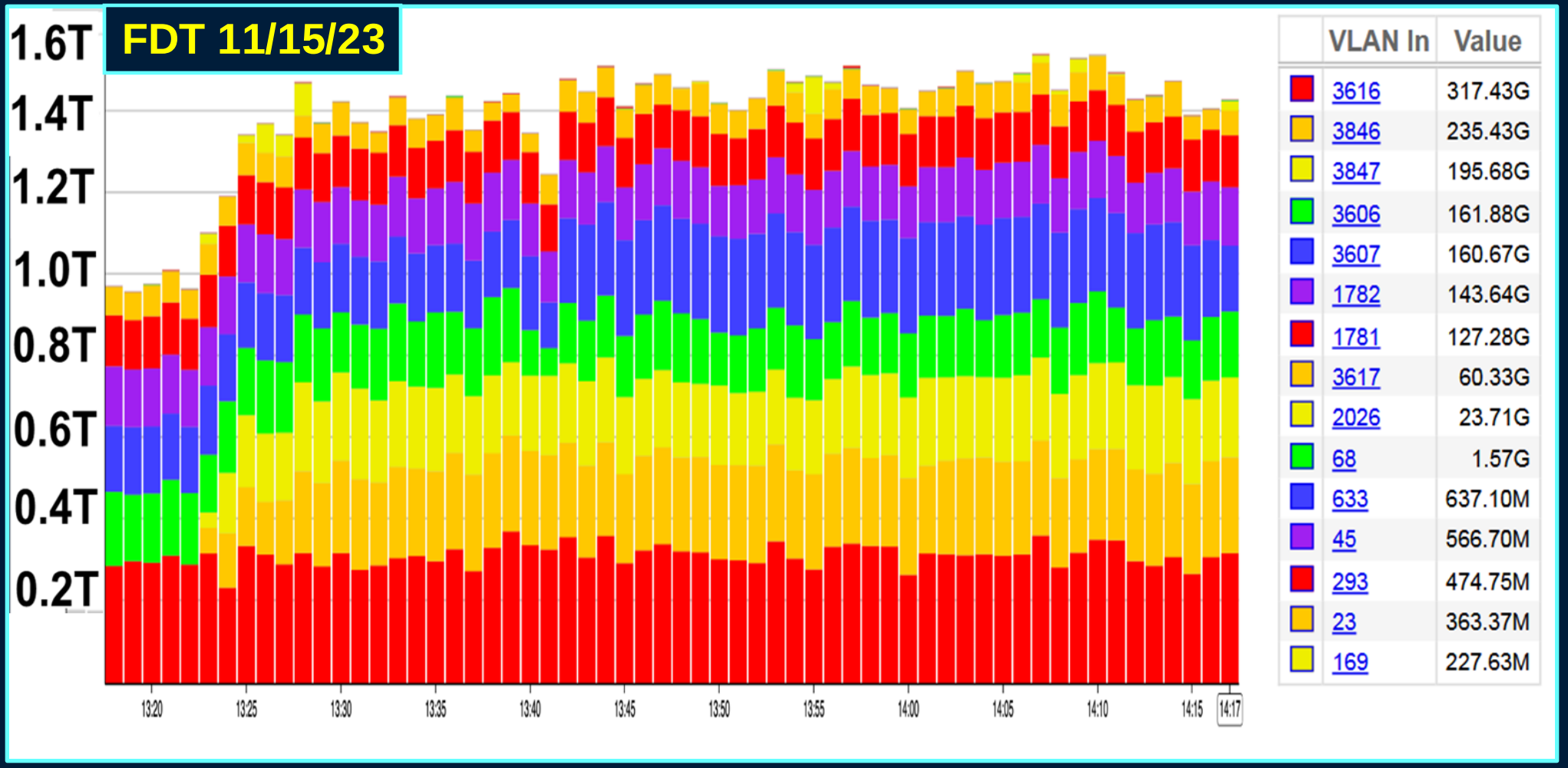
Global Petascale to Exascale Workflows for Data Intensive Sciences

- **Advances Embedded and Interoperate within a 'composable' architecture of subsystems, components and interfaces, organized into several areas; coupled to rising Automation**
  - **Visibility: Monitoring and information tracking and management** including IETF ALTO/OpenALTO, BGP-LS, sFlow/NetFlow, Perfsonar, Traceroute, Qualcomm Gradient Graph congestion information, Kubernetes statistics, Prometheus, P4/Inband telemetry, ***InMon***
  - **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, GradientGraph, Yale Bilevel optimization, Coral, Elastiflow/Elastic Stack
  - **Controllability: SENSE/AutoGOLE/SUPA,** P4, segment routing with SRv6, SR/MPLS 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

SC24 Network Research Exhibition NRE-113 and Partners
NREs Hosted at or Partnering with Caltech Booth 845

NRE-04	Joe Mambretti (Northwestern) et al.	1.2 Tbps Services WAN Services: Architecture, Technology and Control Systems
NRE-05	Joe Mambretti (Northwestern) et al.	Global Research Platform
NRE-16	Chris Wilkinson (Internet2) et al.	NA-REX Prototype Demonstration
NRE-13	Harvey Newman (Caltech) et al.	The Global Network Advancement Group: A Next Generation System for Data Intensive Sciences
NRE-13a	Alex Moura (KAUST) et al.	Exploring FDT, QUIC, BBRv2 and HTTP/3 in High Latency WAN Paths
NRE-13b	Edmund Yeh (Northeastern) et al.	N-DISE: NDN for Data Intensive Science Experiments
NRE-15	Kasandra Pillay (SANReN/CSIR) et al.	High Speed Data Transfers from South Africa to USA !
NRE-19	Carlyn Ann-Lee (JPL) et al.	A Federated Learning Guide to Cosmic Ray Events
NRE-22	Tom Lehman (ESnet) et al.	AutoGOLE/SENSE: End-Site Resource Integration with Network Services
NRE-23	Tom Lehman (ESnet) et al.	SENSE and Rucio/FTS/XRootD/dCache Interoperation
NRE-25	Mariam Kiran (ORNL), M. Martinello (UFES) et al.	HECATE merges with PolKA AI-enabled traffic engineering For data-intensive science
NRE-29	John Graham (UCSD/SDSC), Marcos Schwarz (RNP) et al.	Multi Domain experiments using ESnet SENSE on the National Research Platform / PacWave / FABRIC
NRE-30	Marcos Schwarz (RNP), Carlos Ruggiero (USP/Rednesp) et al.	Global P4 Lab: Programmable Networking with P4, GEANT RARE/freeRtr and SONIC; Digital Twin
NRE-31	Carlos Ruggiero (USP/Rednesp) et al.	High Performance Networking with the Sao Paulo Backbone SP Linking 8 Universities and the Bella Link
NRE-32	Everson Borges (IFES), Magnos Martinello (UFES) et al.	PolKA Routing Approach to Support Traffic Engineering for Data-intensive Sciences
NRE-33	Y. Richard Yang (Yale) et al.	XTS: Scaling and Optimizing Efficiency and Control of Data Transport for Data-intensive Networks

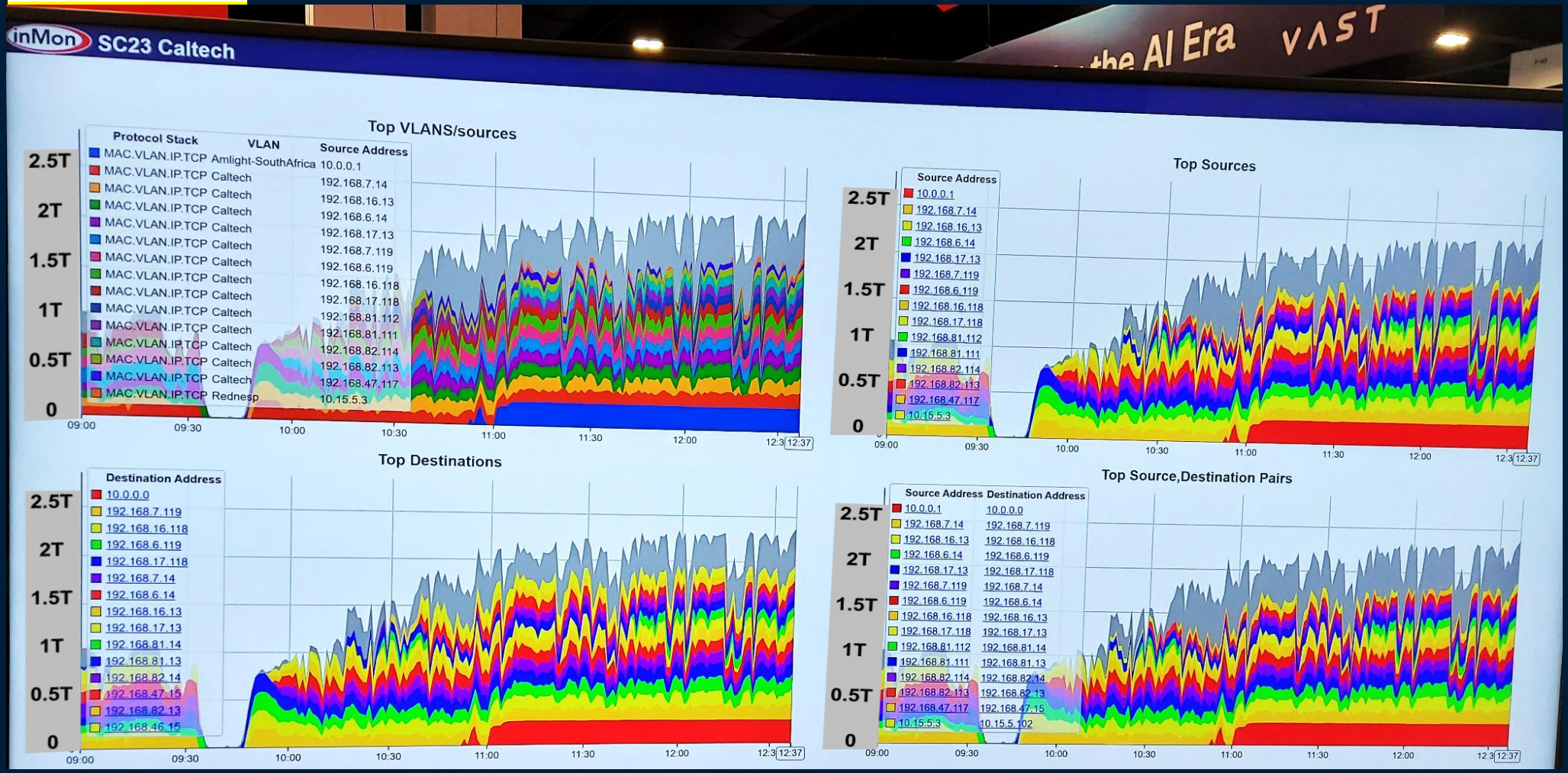
NRE-13 Top Sources: To 1.5+ Tbps
on 4 X 400G Circuits with Dynamic Transfer Limit



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

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

SC23 Stress Test 11/16/23
Caltech Results: Up to 2.4 Tbps

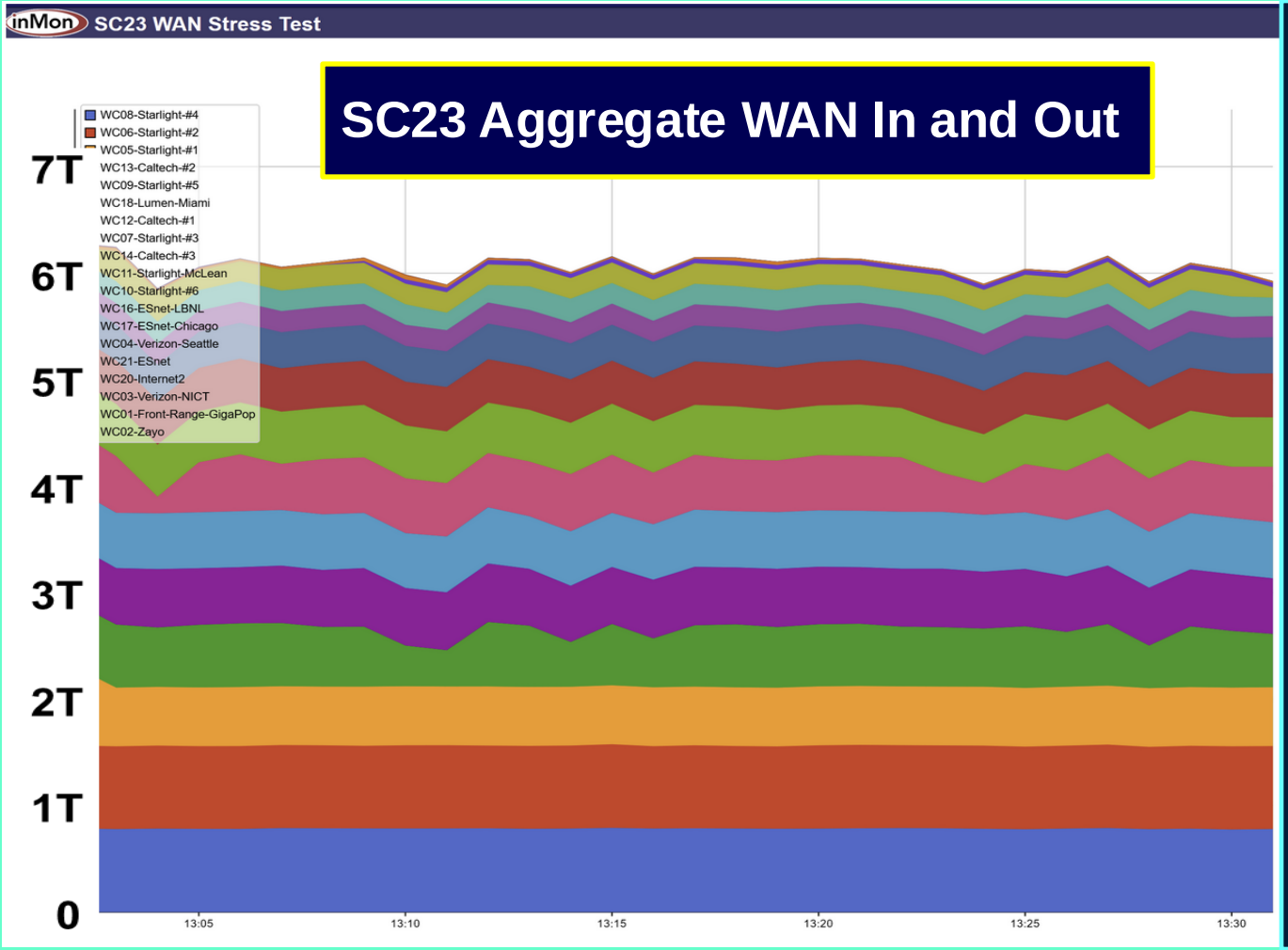


With 2 Gen5 + Gen3 Servers at SC23 and 3 Gen5 Servers at Caltech

FDT 11/16/23

SC23 Stress Test 11/16/23

Caltech Booth providing 2.3 Tbps of 6.2 Tbps



Going Forward

- **Latest kernels:** full use of all PCIe slots
- **400GE with CX7 NICs** and DR4 Transceivers
- **Multi-User: Scheduled stable N X 100G flows** with FDT & SENSE
- **SENSE 400G paths:** ESnet production, NA-REX via StarLight; Links to CERN
- **NVMe SSD Front End** Operations + HSM
- **PCIe 6.0 and CXL** DTN tests by ~SC24

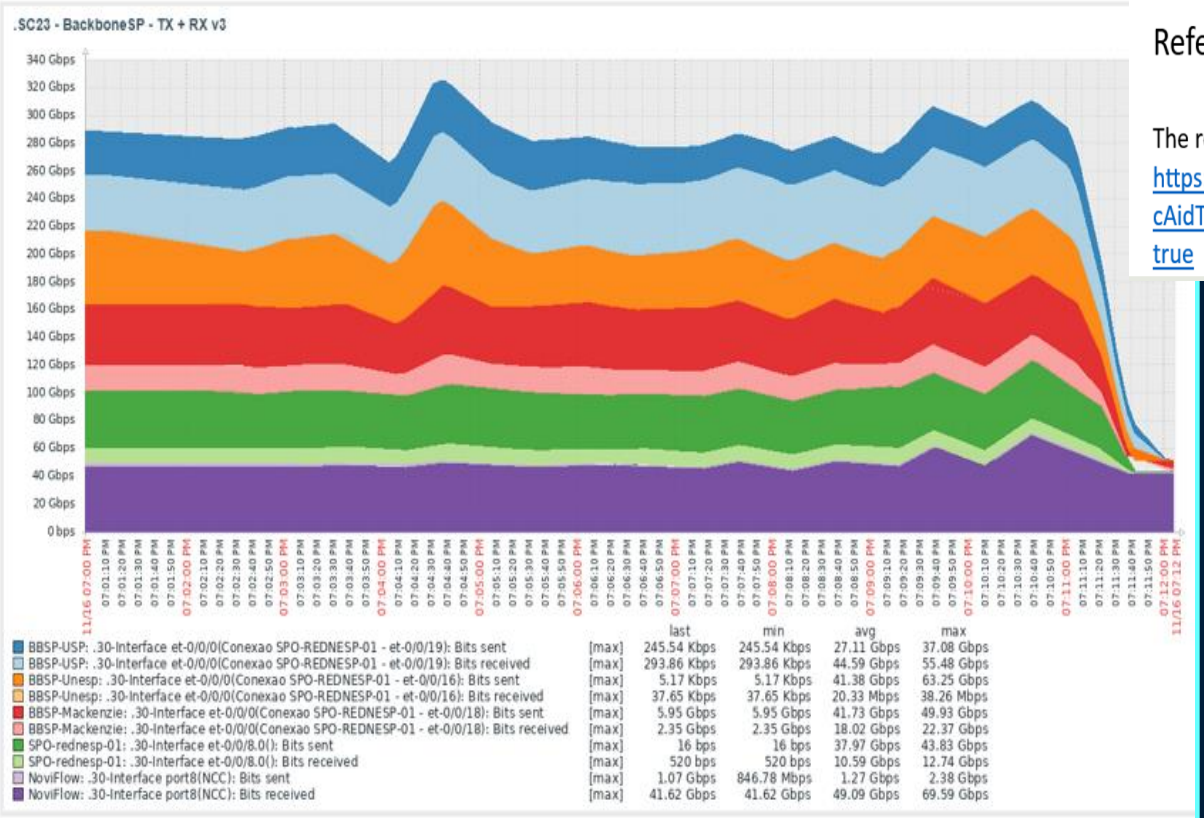
With 2 Gen5 + Gen3 Servers at SC23 and 3 Gen5 Servers at Caltech

FDT 11/16/23

This Just In (11/23) Rednesp Backbone: Record US ⇔ Brazil Results

Two networking tools were used to generate traffic: **iperf3** and **fdt**.

During SC23 data tsunami, on November the 16th, a peak of 330 gbps (considering data from Brazil to the USA and vice-versa) was achieved and can be seen in the next figure.

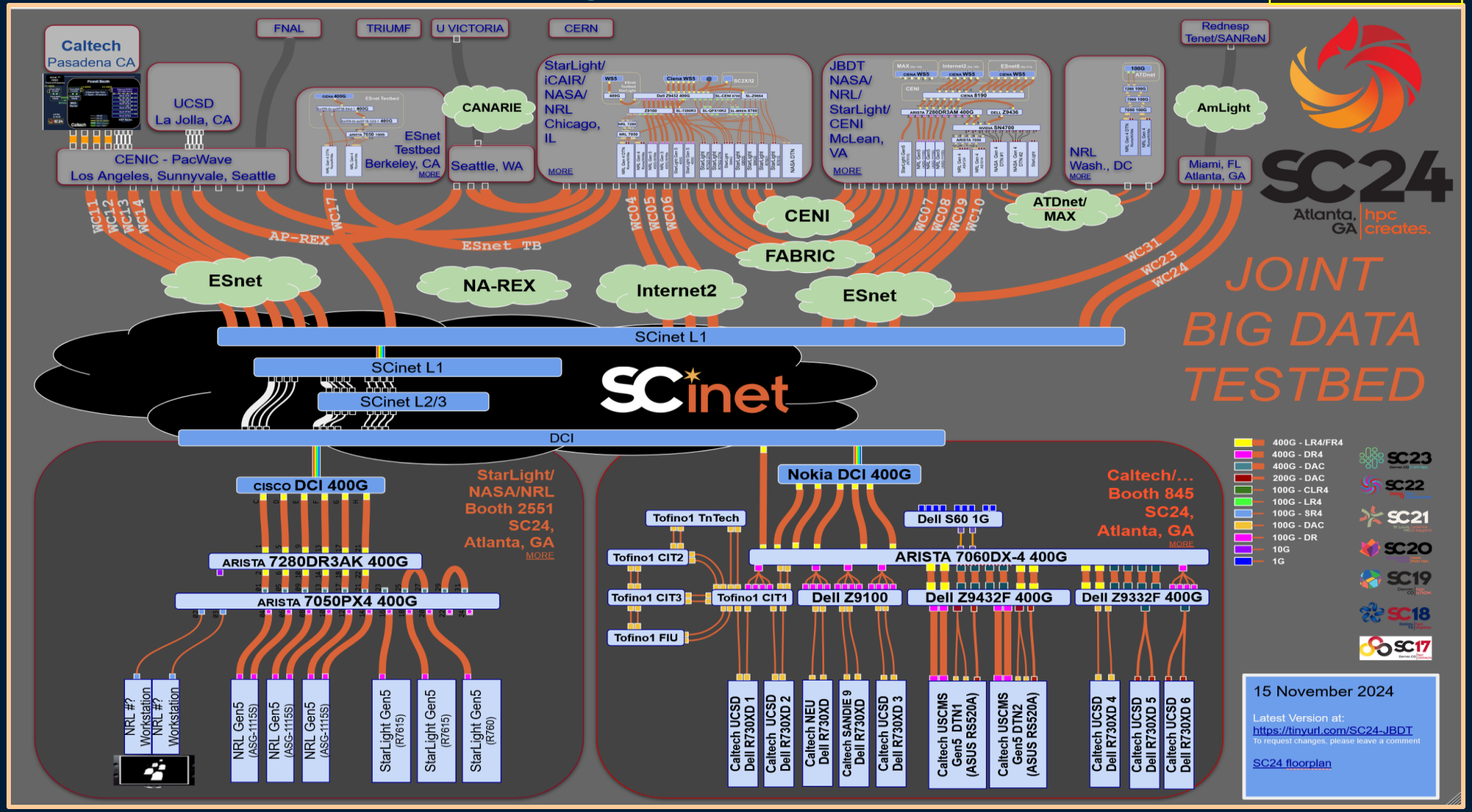


These results are very good, considering that the 100 gbps links also carry production traffic. However, it is certainly possible to achieve higher bandwidths with more tuning and with a more controlled bandwidth allocation in the links. Rednesp is now trying to optimize its infrastructure to achieve a more efficient use of the intercontinental links connecting Sao Paulo, Brazil, to the USA, to Europe and to other countries in South America.

References

The rednesp presentation slides can be seen at https://docs.google.com/presentation/d/1qUX1mvP3Ohb5zMuP18-cAidTvCcFEWF6/edit?usp=drive_link&ouid=114820778254083128813&rtpof=true&sd=true

Caltech and StarLight/NRL Booths at SC24



SC24: Global footprint. Terabit/sec Triangle Starlight – McLean – Atlanta;
6 X 400G to the Caltech Booth: 2X 400G LA-ATL; 4 X 400G to the Caltech Campus,
2 X 400G + 100G to Latin America and South Africa; 400G to Fermilab; 400G to CERN
with CENIC, Ciena, Internet2, ESnet, StarLight, US CMS and Network Partners

DTN: ASUS RS520A-E12-RS12U

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

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

NIC Setup at SC24: 2 X 1.1 Tbps

Two ConnectX-7 2 X 400GE;
One ConnectX-6 200GE (x8)
One Broadcom OCP3.0 2 X 100G

Tofino1 CIT3
Tofino1 CIT2
Tofino1 TnTech
Tofino1 CIT1
Tofino1 FIU
Dell Z9432F 32 X 400G Switch
ASUS Gen5 DTN1 2X400G + 200G + 2X100G
ASUS Gen5 DTN1 2X400G + 200G + 2X100G

Arista 7060DX4 32 X 400G Switch
Dell 730XD DTN 2 X 100G GWU (2U)
Dell 730XD DTN 2 X 100G UCSD 1 (2U)
Dell 730XD DTN 2 X 100G UCSD 2 (2U)

Dell S60 1G Management Switch

Console
Dell Z9100 32 X 100G Switch

Dell 730XD DTN 2 X 100G UCSD 3 (2U)
Dell 730XD DTN 2 X 100G UCSD 4 (2U)

Dell 730XD DTN 2 X 100G UCSD5 (2U)
Dell 730XD DTN 2 X 100G UCSD6 (2U)

Dell 730XD DTN 2 X 100G NEU 1 (2U)
Dell 730XD DTN 2 X 100G SANDIE 9 (2U)

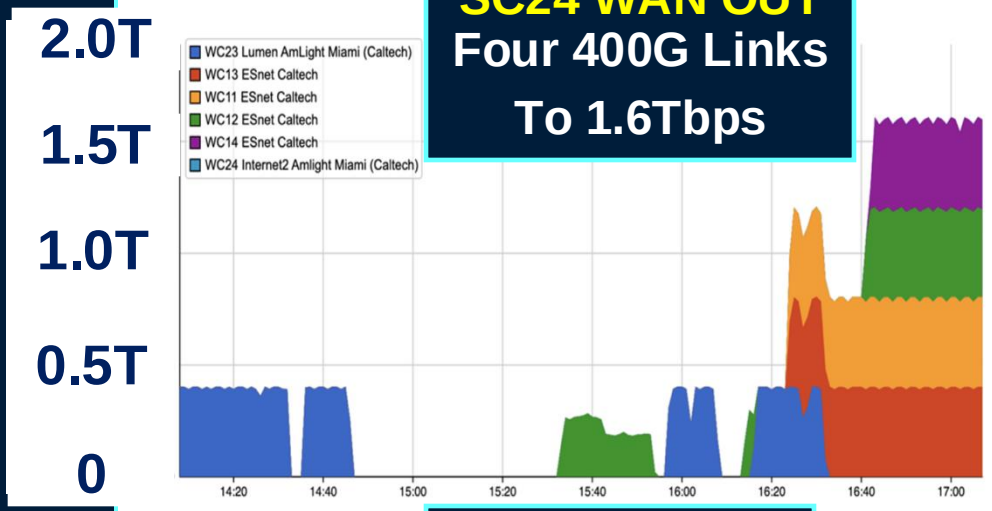
To ~4 Tbps
in a single rack

NRE-13 First Results: To 3.2 Tbps

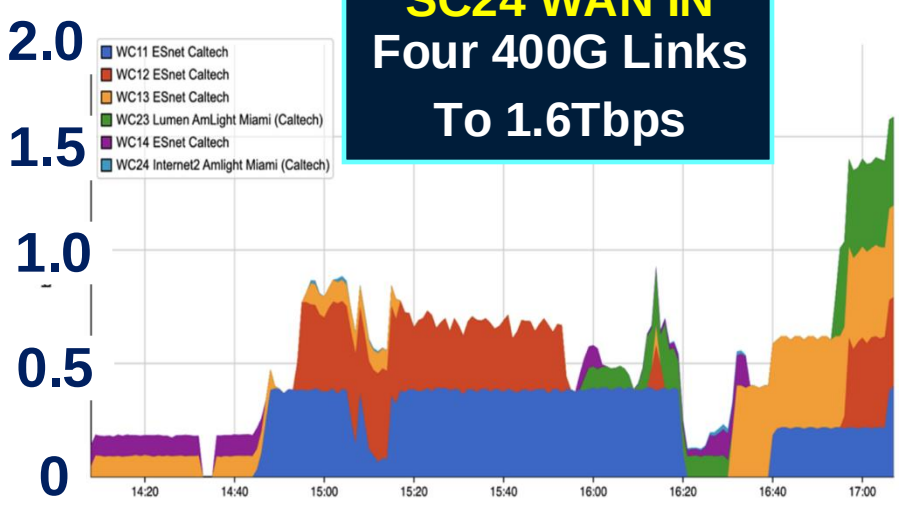
on 4 of 6 X 400G Circuits: 4 LA-ATL, 2 Miami-ATL

FDT++ 11/19/24

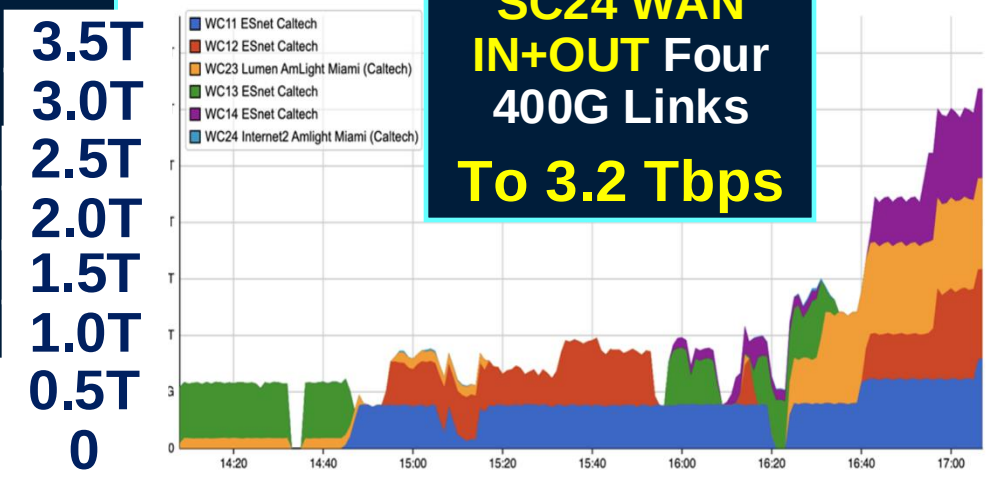
SC24 WAN OUT
Four 400G Links
To 1.6Tbps



SC24 WAN IN
Four 400G Links
To 1.6Tbps



SC24 WAN IN+OUT Four 400G Links
To 3.2 Tbps

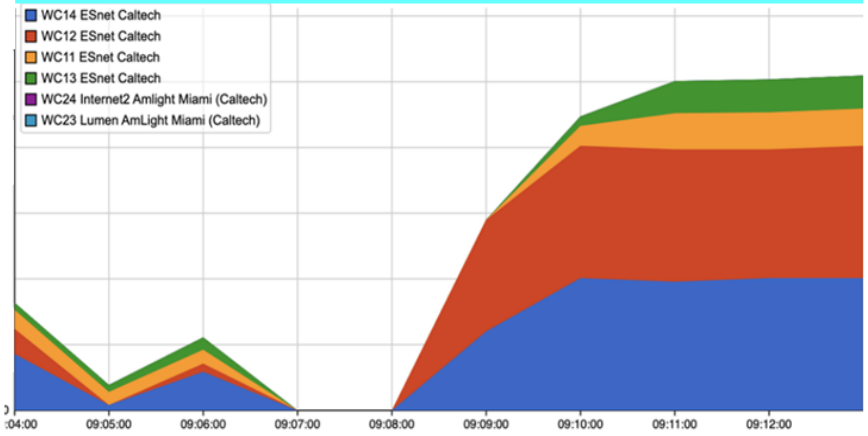


Just Getting Started (on 4 of 6 Links)

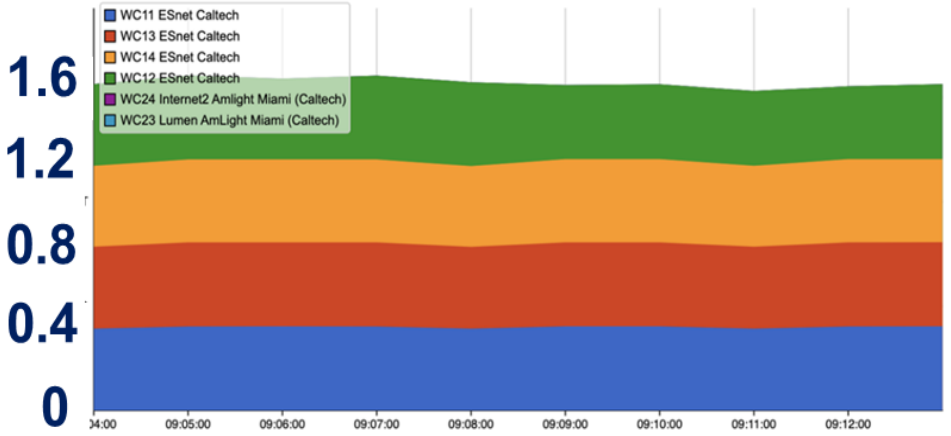
+ Studying the limits in depth of a single 32 core CPU and server with 1.1 Tbps of NIC ports

NRE-13 Server Limits

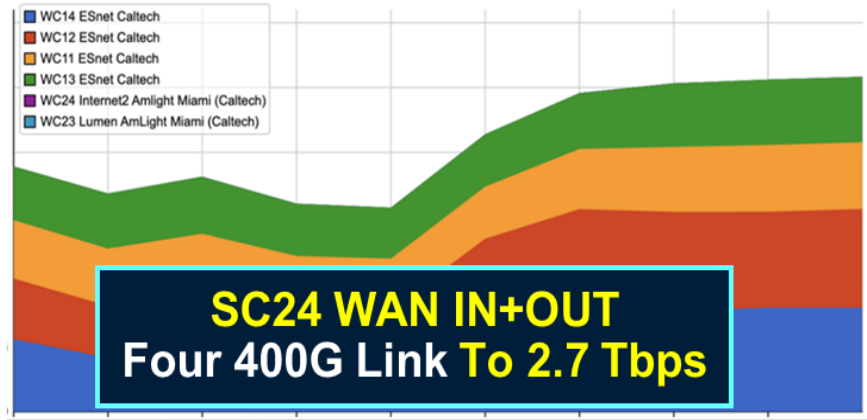
SC24 WAN OUT from 1 Server to Four 400G LinksTo 1.1 Tbps



SC24 WAN IN from 2 Servers with Four 400G Links To 1.6 Tbps



SC Aggregate WAN IN+OUT



SC24 WAN IN+OUT Four 400G Link To 2.7 Tbps

Just Getting Started

Wire Speed: One server can do 800G in + 800G Out with iperfUsing all 32 Cores

Acknowledgements

This ongoing work is partially supported by the US National Science Foundation (NSF) Grants OAC-2030508, OAC1841530, OAC-1836650, MPS-1148698, and PHY-1624356, along with research grants from many international funding agencies and direct support from the many regional, national, and continental network and industry partners mentioned. The development of SENSE is supported by the US Department of Energy (DOE) Grants DE-SC0015527, DESC0015528, DE-SC0016585, and FP-00002494.

Finally, this work would not be possible without the significant contributions and the collaboration of the many HEP, network and computer and research teams partnering in the Global Network Advancement Group, in particular the GNA-G Data Intensive Sciences and AutoGOLE/SENSE Working Groups and the Global P4 Lab led by GEANT and the RNP Brazilian National Network, together with many industry partners, most notably Ciena, Dell and Arista