Global Network Advancement Group

Next Generation Network-Integrated System for Data Intensive Sciences

LHC Run 3 and HL-LHC

Rubin Observatory

SKA

BioInformatics

Earth Observation

Gateways to a New Era

SC23 Network Research Exhibition

NRE-13 and Partner NREs

Worldwide Partnership at SC23 and Beyond

GNA-G

SC23

Global Petascale to Exascale Workflows for Data Intensive Sciences

Accelerated by Next Generation Programmable SDN Architectures and Machine Learning Applications

Foreseeing, Comprehending and Meeting the Challenges

- LHC to HL LHC Challenges: Scale, Complexity and Global Extent;
  Complex workflows not fully captured in requirements documents so far
- GNA-G: since 2019 towards a comprehensive next generation global system in the Global Network Advancement Group and its DIS and SENSE WGs:
  Bringing together computing, storage and networks, all as first class subsystems;
  Meeting the challenges while accommodating traffic of the at-large A&R community
- Parallel lines of development, and progressive integration; two global testbeds
  Scaling from 100G to 400G links, nationally, transoceanic, on campuses
- Moving towards a flexible architecture that will accommodate regional and VO specific developments; multiple open source network OSes
  - Leveraging multiple open source developments in academia and industry
    Programmable core services of increasing power and sophistication
- Model Concepts: Data center analogue; multidomain collaborative services; metric of success takes into account multiple dimensions: priority, policy, network and site state and deadlines. Digital Twin to go from prototypes into production.
- PolKA: Polynomial Key-based architecture for traffic engineering and management of competing Terabit/sec flows across intercontinental networks
- Machine Learning for system optimization; developing relevant metrics

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
  - Needs are further specified in "blueprint" Requirements documents by US CMS and US ATLAS, submitted to the ESnet Requirements Review, and captured in a comprehensive 2021 DOE Requirements 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

[*] Another Update of the Requirements Report is coming in 2024

Estimates at the time of DC21: Data Rate Table

M. Lassnig at WLCG GDB July 12, 2023

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

  2 (for bursts) *2 (overprovisioning) = *4.8Tbps
- **Flexible model: Assumes reading of data from above for reprocessing/reconstruction within 3 months**
- **Means doubling the Minimal Model: *9.6Tbps;***

  Including 2.7 Gbps Transatlantic for ATLAS and CMS Alone

- **But: Only data flows from the T1s to T2s and T1s accounted for.**
  - ***Nota Bene: No MC production flows nor re-creation of derived data included in the 2021 modelling !***

400G Transatlantic Plans: ESnet, Internet2 and Canarie

400G Transatlantic Capacity Additions/Upgrades on Amitié cable

- 1 x 400G for Internet2/CANARIE
- 2 x 400G for ESnet
- Mid 2023 (wet-plant in Q2 and terrestrial in Q3)
- Add Boston as open exchange point
- Early effort to acquire spectrum services from commercial providers – lessons learned

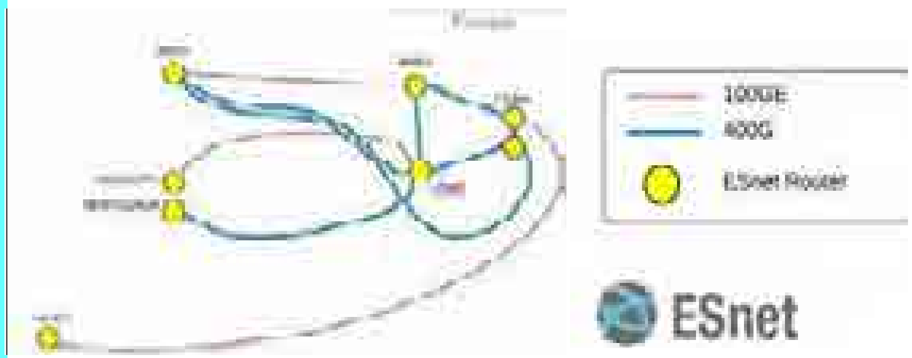
Exploring second 400G link into New York, Washington or other east-coast city TBD

Christian Todorov, Internet2
LHCOPN/LHCONE Meeting April 2023

ESnet Transatlantic and EU Ring Upgrades

Eli Dart, ESnet
LHCOPN/LHCONE Meeting
April 2023

- In Production:
  400G NYC-London; peering with GEANT
- Underway (by the fall):
  - 400G Boston – London
  - 400G Boston – CERN
  - 400G European Ring
- Transatlantic Capacity Targets
  - 1.5T in advance of DC24
  - ... To 3.2T well in advance of Run4
    (by 2028; if funding is as expected)

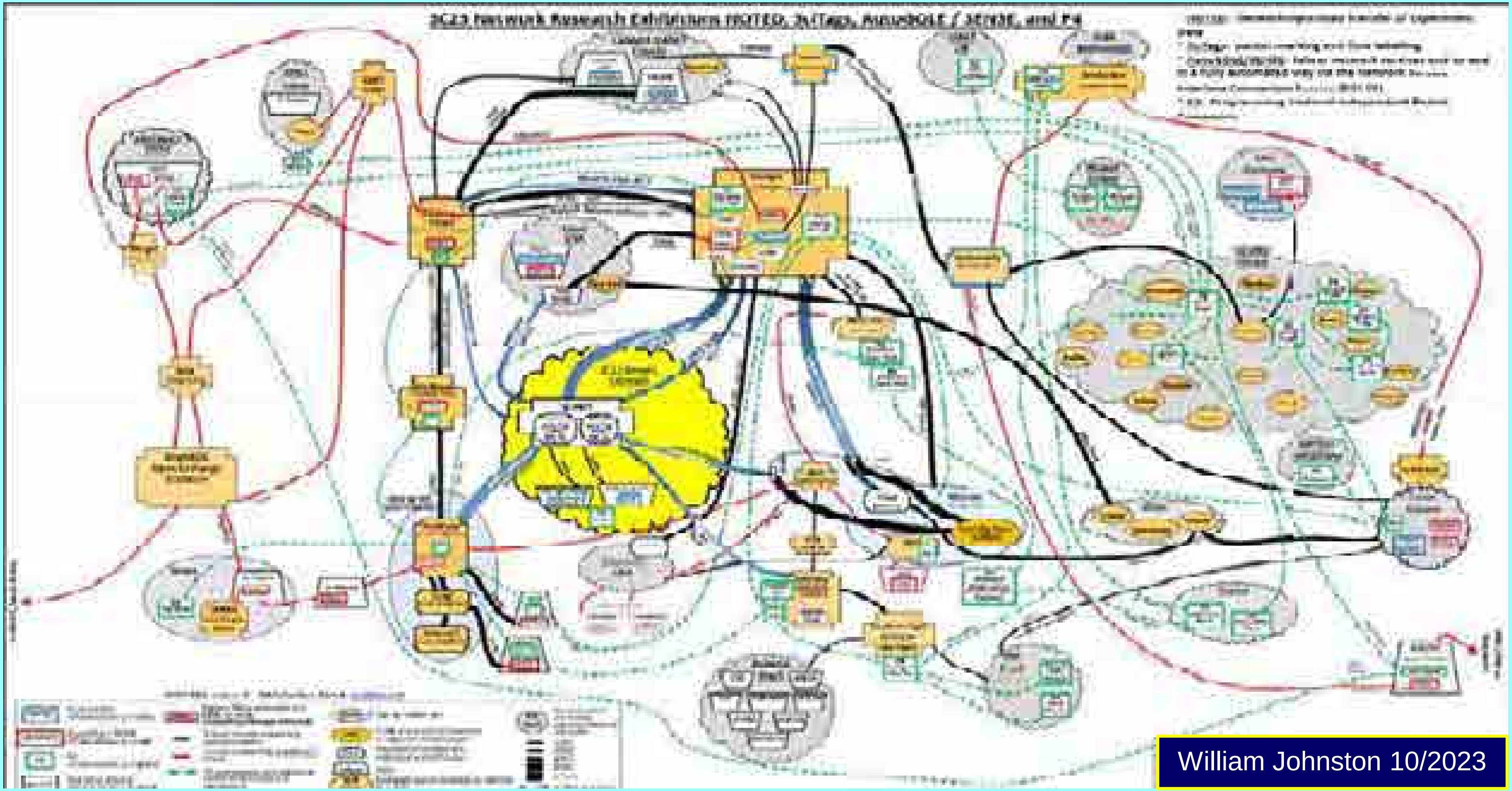


Exchange Points with Automation in Boston, Washington, NYC

Provide support to the community for 400G diversity among Internet2, ESnet, ANA, FIU/AmLight, APONet et al.

GNA-G

LHCONE Map Highlighting 400G to 1.6T Links for SC23



SC23: Global footprint. Terabit/sec Triangle Starlight – McLean – Denver;
3 X 400G to LA; 4 X 400G to the Caltech Campus, and 3 X 400G to the Caltech Booth
with CENIC, Ciena, Internet2, ESnet, StarLight, US CMS and Network Partners

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

leadershipteam@lists.gna-g.net

Buseung Cho
KISTI (Korea)

Marco Teixera
RedCLARA
(Latin America)

Ivana Golub
PSNC, GEANT
(Europe)

Harvey Newman
Caltech (US)

David Wilde,
Chair
Aarnet (Australia)

Alex Moura
KAUST
(Saudi Arabia)

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

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

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

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

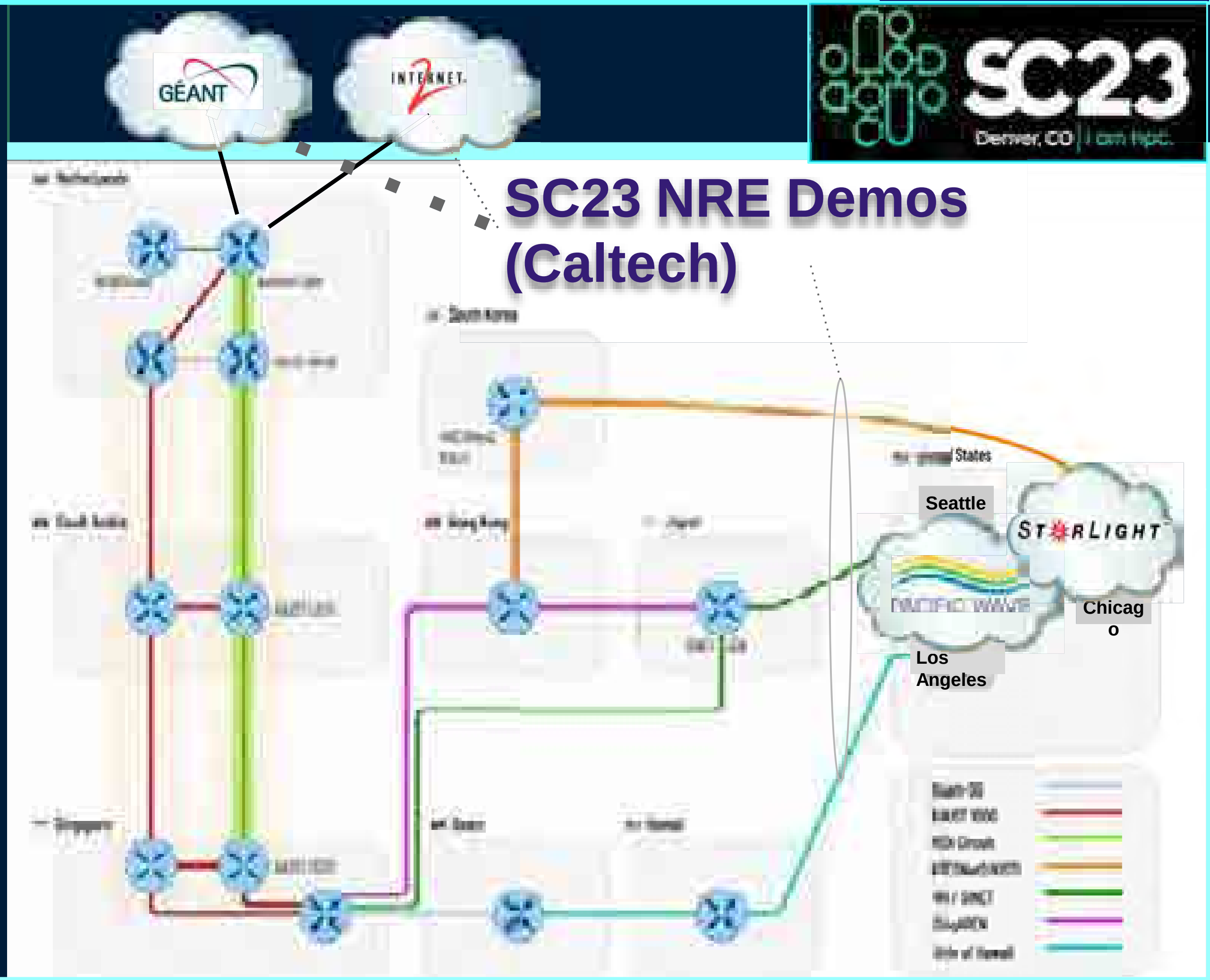
Rednesp and RNP: Expanding Capacity Among Latin America, US, Europe and Africa

- Total 600 Gbps capacity between Brazil and the USA:
- Rednesp (formerly ANSP) has 3 links to the USA, connecting Sao Paulo to Florida.
  - Atlantic 100G link: direct São Paulo- Miami
  - Pacific 100G link: São Paulo to Santiago (Chile), from there to Panama, San Juan and Miami
  - 200 Gbps link: São Paulo to Florida through Fortaleza, on the Monet cable (Angola cables)
- RNP also a 200 Gbps link using the Monet cable

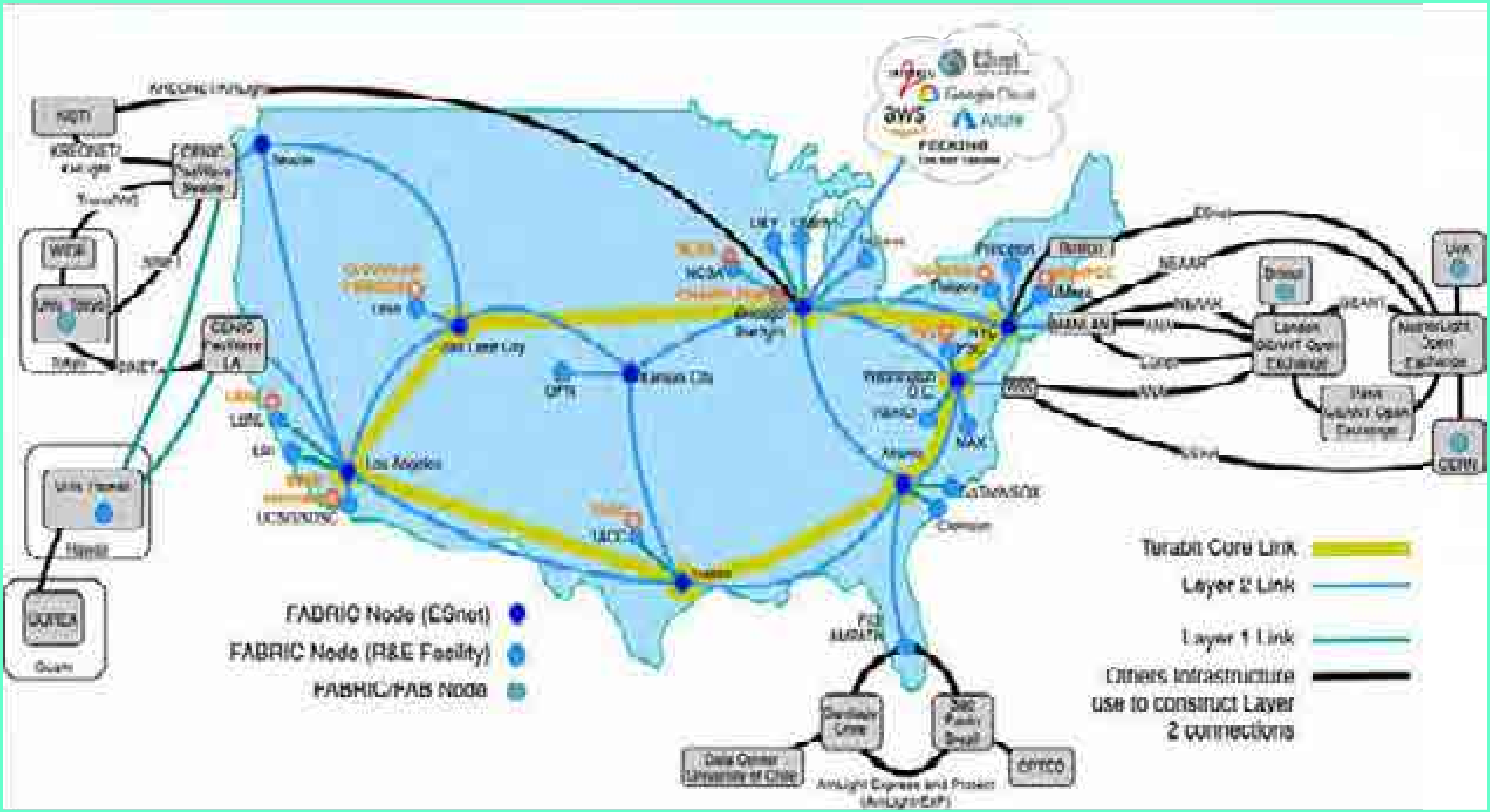
- Transatlantic Links:
  The Bella link between São Paulo and Sines in Portugal, and a link connecting São Paulo, Angola and South Africa.
- Deployment of the "Backbone SP" interconnecting most universities in the Sao Paulo (state) with 100G links has been completed by May 2023
  - UNICAMP (University of Campinas),
  - UNIFESP (Federal University at São Paulo)
  - State University of São Paulo)
  - USP (University of São Paulo)
  - UFSCAR (Federal University of São Carlos)
  - Mackenzie University
  - ITA (Aeronautics Institute of Technology)
    UFABC (Federal University of ABC) to SP4

KAUST and the AsiaPacific Oceania Network (APONet): Closing the Global Ring East and West

- Since the AER MoU, KAUST is coordinating with REN partners on shared deployment of spare capacity
- KAUST is supporting the following partners by offering point-to-point circuits for submarine cable backup paths:
  - AARnet
  - GÉANT
  - NetherLight
  - NII/SINET
  - SingAREN
- The SC23 NRE Demonstrations are also supported by KAUST closing the ring from Amsterdam to Singapore and back to the US

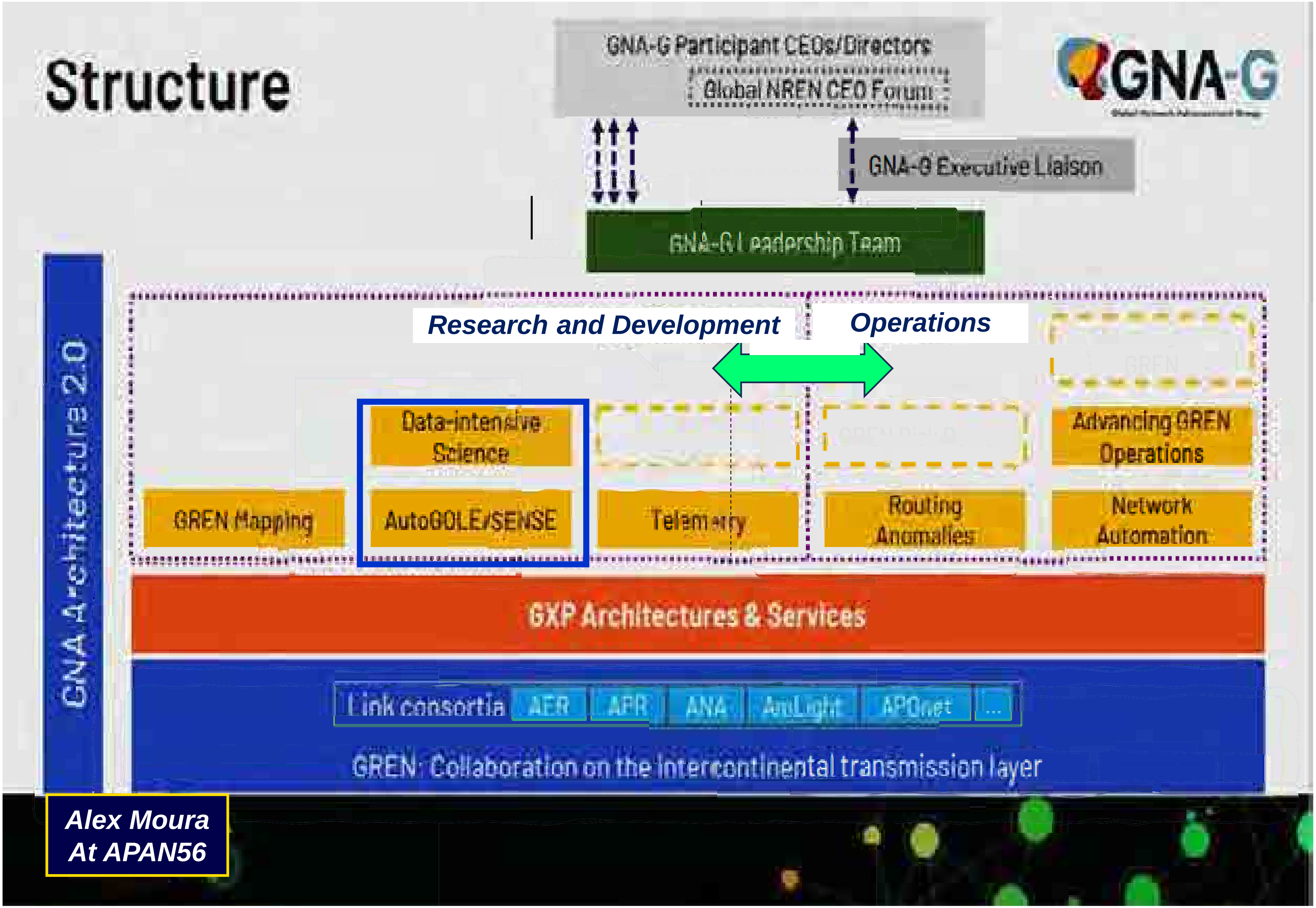


FABRIC and FAB: Terabit/sec Across the US. Transoceanic Links and Intercontinental Partnerships



US, Europe, Asia Pacific and Latin America

CENIC and Pacific Wave: A “Regional” Network for California with Global Reach and Vision East and West



Structure
GNA-G Participant CEOs/Directors
Global NREN CEO Forum
GNA-G Executive Liaison
GNA-G Leadership Team
GNA Architecture 2.0
Research and Development
Operations
Data-intensive Science
Advancing GREN Operations
GREN Mapping
AutoGOLE/SENSE
Telemetry
Routing Anomalies
Network Automation
GXP Architectures & Services
Link consortia
AER
AFR
ANA
AmLight
APOnet
...
GREN: Collaboration on the intercontinental transmission layer
Alex Moura
At APAN56

The GNA-G Data Intensive Sciences WG

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

- ***A Vast Worldwide Partnership of* R&E networks, physics programs, advanced network R&D projects, scientists and engineers in multiple disciplines**
- ***Members:***

  Alberto Santoro, Alex Moura, Azher Mughal, Bijan Jabbari, Buseung Cho, Caio Costa, Carlos Antonio Ruggiero, Carlyn Ann-Lee, Chin Guok, Chris Bruton, Chris Wilkinson, Ciprian Popoviciu, Cristina Domenicini, Dale Carder, David Lange, David Wilde, Dima Mishin, Edoardo Martelli, Eduardo Revoredo, Eli Dart, Eoin Kenney, Everson Borges, 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, Kevin Sale, Lars Fischer, Liang Zhang, Mahdi Solemani, Carmen Misa Moreira, Magnos Martinello, Marcos Schwarz, Mariam Kiran, Matt Zekauskas, Michael Stanton, Mike Hildreth, Mike Simpson, Moises Ribeiro, Ney Lemke, Oliver Gutsche, Phil Demar, Preeti Bhat, Rafael Guimaraes, Raimondas Sirvinskas, Richard Hughes-Jones, Rogerio Iope, Rogerio Motitsuki, Sergio Novaes, Shawn McKee, Susanne Naegele-Jackson, Tim Chown, Tom de Fanti, Tom Hutton, Tom Lehman, William Johnston, Xi Yang, Y. Richard Yang, Ryan Yang
- ***Participating Organizations/Projects/Supporters:***
- *ESnet, AARNet, AmLight, Rednesp, KAUST, KISTI, SANReN, GEANT, RNP, CERN, Internet2, CENIC/Pacific Wave, StarLight, NetherLight, SURFnet, Nordunet, Southern Light, National Research Platform, FABRIC, RENATER, ATLAS, CMS, VRO, SKAO, OSG, Caltech, UCSD, Yale, FIU, UFES, UERJ, GridUNESP, Fermilab, Nebraska, Vanderbilt, Michigan, UT Arlington, George Mason, East Carolina; Ciena, Arista, Dell*

✸ ***Meets Weekly or Bi-weekly***

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

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/Supa for network connections, SENSE for integration of End Systems and Domain Science Workflow facing APIs.
- Key Objective:
  - The AutoGOLE Infrastructure should be persistent and reliable, to allow most of the time to be spent on experiments and research.
- Key Work areas:
  - Control Plane Monitoring: Prometheus based, Deployments underway
  - Data Plane Verification and Troubleshooting Service: Study and design group formed
  - AutoGOLE related software: Ongoing enhancements to facilitate deployment and maintenance (Kubernetes, Docker based systems)
  - Experiment, Research, Multiple Activities, Use Case support: Including XRootD/Rucio Integration, Fabric, NOTED, Qualcomm GradientGraph, P4 Topologies, Named Data Networking (NDN), Data Transfer Systems... integration & testing.
- WG information https://www.gna-g.net/join-working-group/autogole-sense

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**

Next Generation System for Data Intensive Sciences

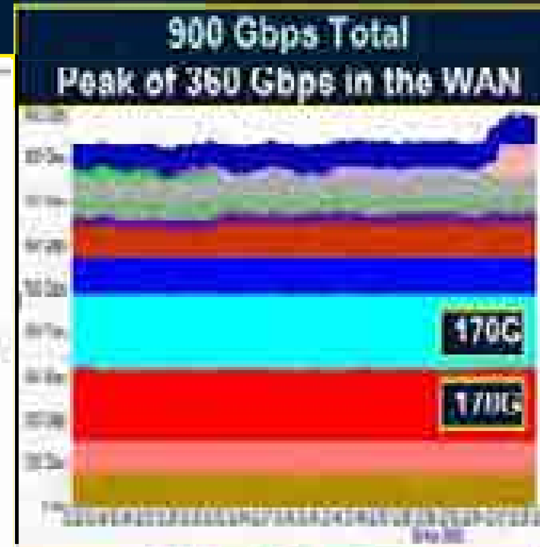
- Architectural Model: Data Center Analogue
  - Classes of "Work" (work = transfers, or overall workflow), defined by VO, task parameters and/or priority and policy
  - Adjusts rate of progress in each class to respond to network or site state changes, and "events"
  - Moderates/balances the rates among the classes
  - Optimizes a multivariate objective function with constraints
- Overarching Concept: Consistent Network Operations:
  - Stable load balanced high throughput workflows crossing 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 serving the at-large academic and research community
- Developing a new operational paradigm, enabling the community; protecting the world's R&E networks as site knowledge/capability rise

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

supercomputing.caltech.edu

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

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



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

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

Tbps Rings for SC18-23: 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*

SC23 Network Research Exhibition NRE-13 and Partners
NREs Hosted at or Partnering with Caltech Booth 1255

NRE-001	Joe Mambretti (Northwestern) et al.	1.2 Tbps Services WAN Services: Architecture, Technology and Control Systems
NRE-002	Joe Mambretti (Northwestern) et al.	400 Gbps E2E WAN Services
NRE-003	Joe Mambretti (Northwestern) et al.	NA-REX Prototype Demonstration
NRE-004	Joe Mambretti (Northwestern) et al.	Global Research Platform
NRE-005	Edoardo Martelli (CERN) et al.	LHC Networking and NOTED
NRE-009	Qiao Xiang (Xiamen) et al.	Fully Automated Network Configuration for Large Scale Networks
NRE-013	Harvey Newman (Caltech) et al.	The Global Network Advancement Group: A Next Generation System for Data Intensive Sciences
NRE-014	Tom Lehman (ESnet) et al.	AutoGOLE/SENSE: End-to-End Network Services and Workflow Integration
NRE-015	Tom Lehman (ESnet) et al.	SENSE and Rucio/FTS/XRootD Interoperation
NRE-016	Tom Lehman (ESnet) et al.	FABRIC
NRE-019	Jeronimo Bezerra (FIU) et al.	AmLight 2.0: Flexible Control, Deep Visibility and Programmability @ Tbps
NRE-020	Marcos Schwarz (RNP) et al.	Global P4 Lab: Programmable Networking with P4, GEANT RARE/freeRtr and SONIC; Digital Twin
NRE-021	Y. Richard Yang (Yale) et al.	ALTO-TCN: Application-Defined Network Control for Data Intensive Sciences Though Deep Network Visibility
NRE-022	Mariam Kiran (ORNL) et al.	5G on the Showfloor
NRE-024	Alex Moura (KAUST) et al.	Exploring FDT, QUIC, BBRv2 and HTTP/3 in High Latency WAN Paths
NRE-025	Edmund Yeh (Northeastern) et al.	N-DISE: NDN for Data Intensive Science Experiments
NRE-032	Magnos Martinello (UFES) et al.	PolKA Routing Approach to Support Traffic Engineering for Data-intensive Sciences

GNA-G: Next Generation Network-Integrated System
NRE-13: Elements of the Demonstrations

- 400 G (Switch, Server) to 1.6 T (4 X 400GE, 2 X 800G Coherent) Next Generation Networks Transformation of the LA CENIC/Pacific Wave PoP
- National Research Platform
- Global Research Platform (GRP); Software Defined Int'l Open Exchanges (SDXs)
- SENSE: Automated virtual circuit and flow control services for data intensive science programs; FTS and Rucio integration for LHC workflows
- Rednesp High performance networking with the Bella Link & Sao Paulo Backbone
- AmLight Express & Protect (AmLight-EXP) With SANREN, TENET and CSIR: US-Latin America (Rednesp); US-South Africa
- N-DISE: Named Data Networking for Data Intensive Experiments
- PolKA: Polynomial Key-Based Architecture: Creation of an overlay network with Source Routed tunnels forming virtual circuits
- Towards Fully-Automated Network Configuration Management for Large-Scale Science Networks with Scalable Distributed Data Plane Verification
- KAUST: Exploring Efficient Data Transfer Protocols Across High Latency Networks
- KISTI-SCION: Scalability, Control and Isolation on Next Gen (Round the World) Networks
- 5G/Edge Computing Application Performance Optimization; High-Performance Routing of Science Network Traffic
- Network traffic prediction and engineering optimizations with graph neural network and other emerging deep learning methods, developed by ESnet's Hecate /DeepRoute project
- ALTO/TCN: Application-Level Traffic Optimization and Transport Control +Integration of OpenALTO and Qualcomm GradientGraph

NRE-13: An Exciting Agenda
at the Caltech Booth 1255

SC23 - Caltech/GNA-G Booth 1255 | Agenda

Tuesday, November 14		
10:30 AM	The Global Network Advancement Group by The Data Intensive Sciences and AutoGOLE/SENSE Working Groups	Link
10:40 AM	Global Research Platform and SC23 by Joe Mambretti (in-person)	Link
11:20 AM	Title: AmLight 2.0: Flexible control, deep visibility, and programmability @ Tbps! Presenter: Jeronimo Bezerra	Link
11:40 AM	Compute Naming with NDN and Kubernetes by Sankalpa Timilsina	Link
1:55 PM	Title: AutoGOLE/SENSE: End-to-End Network Services and Workflow Integration Presenter: by Tom Lehman	Link
2:20 PM	NOTED by Bruno Hoeft	Link
2:30 PM	Moving towards IPv6 only in the German Tier 1 Data Center of the CERN Large Hadron Collider by Bruno Hoeft	Link

Wednesday, November 15		
10:30 AM	Global P4 Lab by Marcos Schwarz	Link
10:56 AM	SENSE/Rucio by Preeti P Bhat and Justas Balcas	Link
11:20 AM	N-DISE: NDN for Data-Intensive Science Experiments by Edmund Yeh	Link
11:45 AM	High performance networking with São Paulo Backbone SP connecting 8 universities by Rogerio Motitsuki	Link
1:30 PM	Beyond a Centralized Verifier: Scaling Data Plane Checking via Distributed, On-Device Verification by Zhenpei Huang (in person)	Link
1:55 PM	Toward Fully-Automated Network Configuration Management for Large-Scale Science Networks by Huiban Xu (in person)	Link
2:20 PM	KISTI-SCION: Scalability, Control and Isolation on Next Generation Networks by François Wirz, Chanjin Park (in-person)	Link
2:45 PM	PolKA routing approach to support traffic engineering for data-intensive science by Rafael Guimarães (in-person)	Link
3:35 PM	White Rabbit sync and High Energy Physics by Francisco Girela Lopez	Link
4:00 PM	GNA-G Community Meeting by GNA-G Leadership Team	Link
4:25 PM		
4:50 PM		
5:15 PM		
5:40 PM		

- Presentation Agenda Complementary to the SCinet Theater Nearby
- Intensive Activities at the Booth 11/9 – 11/16
- Rapid Startup: Excellent Infrastructure: SCinet, VLANs, Power and Crate Delivery

https://docs.google.com/spreadsheets/d/1unE5efWM5FRg3pvbLBSY1Y5JcM1VGBo6b3JRLy8VBv4/edit#gid=213419659

Global P4 Lab (GP4L)

Tofino Core – 26 Core Sites/Devices:

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

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

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

x86/DPDK Islands – 4 Sites/Devices:

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

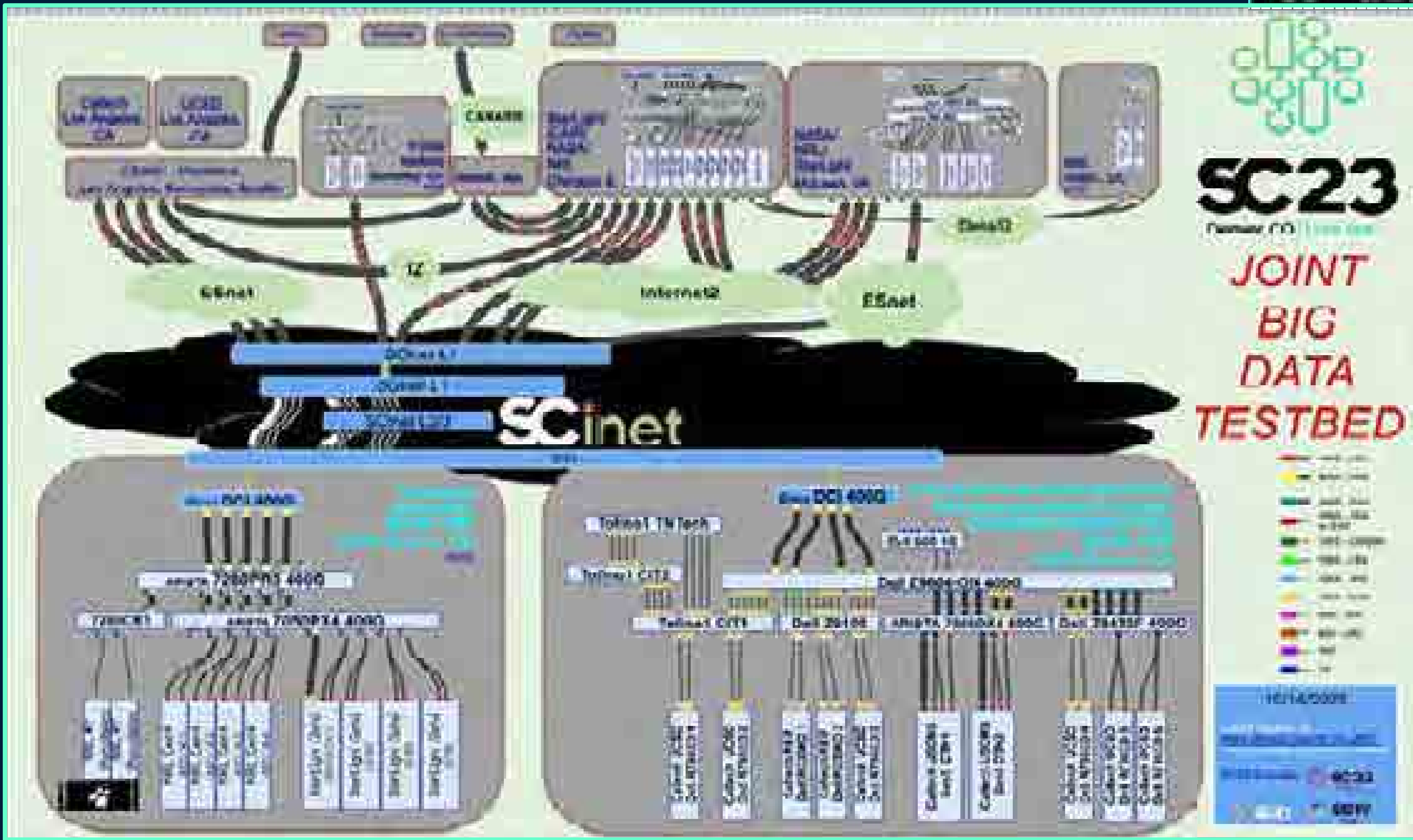
PolKA: An Efficient Source Routing Approach to Meet the Requirements of Data Intensive Sciences

No tables in the core	Fixed length header	Topology agnostic multipath routing	Support in prog. switches	Open source/ Interoperable

Talk by Rafael Guimaraes (UFES)

- **PolKA: Polynomial Key-based Architecture for Source Routing Implementation**
  - **Stateless Core:** A single user-defined encoded/decoded label defines the path: identifying each switch and port along the way
  - Polynomial Residue Number System (RNS)
  - Chinese Remainder Theorem (CRT)
  - Packet forwarding based on mod operation:
    using switch CRC hardware for speed (> 100 Gbps achieved)
  - Packets traverse fixed function switches in the path as needed
- Easy Setup of paths/tunnels using a standard CLI
- Open Source Implementation in RARE/freeRtr
- Many powerful network applications: Proof of transit, PBR, multipath, multicast, failure protection, telemetry, ...

Caltech and StarLight/NRL Booths at SC23



SC23: Global footprint. Terabit/sec Triangle Starlight – McLean – Denver;
3 X 400G Denver-LA; 4 X 400G to the Caltech Campus, and 4 X 400G to the Caltech Booth
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

SC23

NIC Setup at SC23 (x2)

NVIDIA ConnectX-7 400GE (200GE)

Two ConnectX-6 200GE

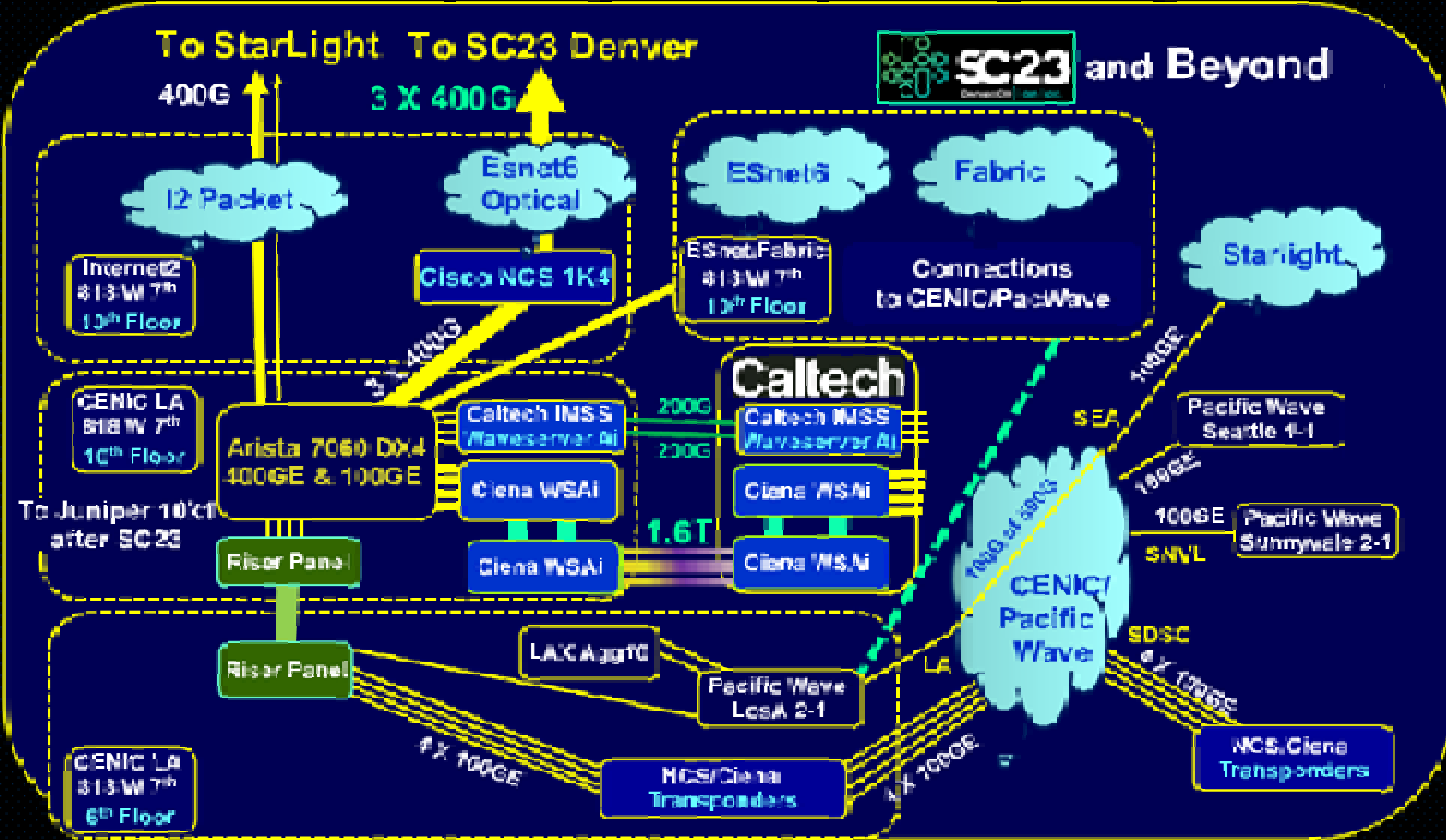
One ConnectX-6 x8 100GE

One 100GE OCP3.0

Tofino1 TNTech
Tofino1 BUR001
Tofino1 BUR002
Dell Z9432F 32 X 400G Switch
Arista 7060DX4 32 X 400G Switch
ASUS Gen5 DTN1 400G + 3 X 200G + 100G
ASUS Gen5 DTN2 400G + 3 X 200G + 100G
Dell Z9664F-ON 64 X 400G Switch
Dell 730XD DTN 2 X 100G UCSD 1 (2U)
Dell 730XD DTN 2 X 100G UCSD 2 (2U)
Dell Z9100 32 X 100G Switch
Console
Dell S60 Switch
Dell 730XD DTN 2 X 100G UCSD3 (2U)
Dell 730XD DTN 2 X 100G UCSD4 (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)

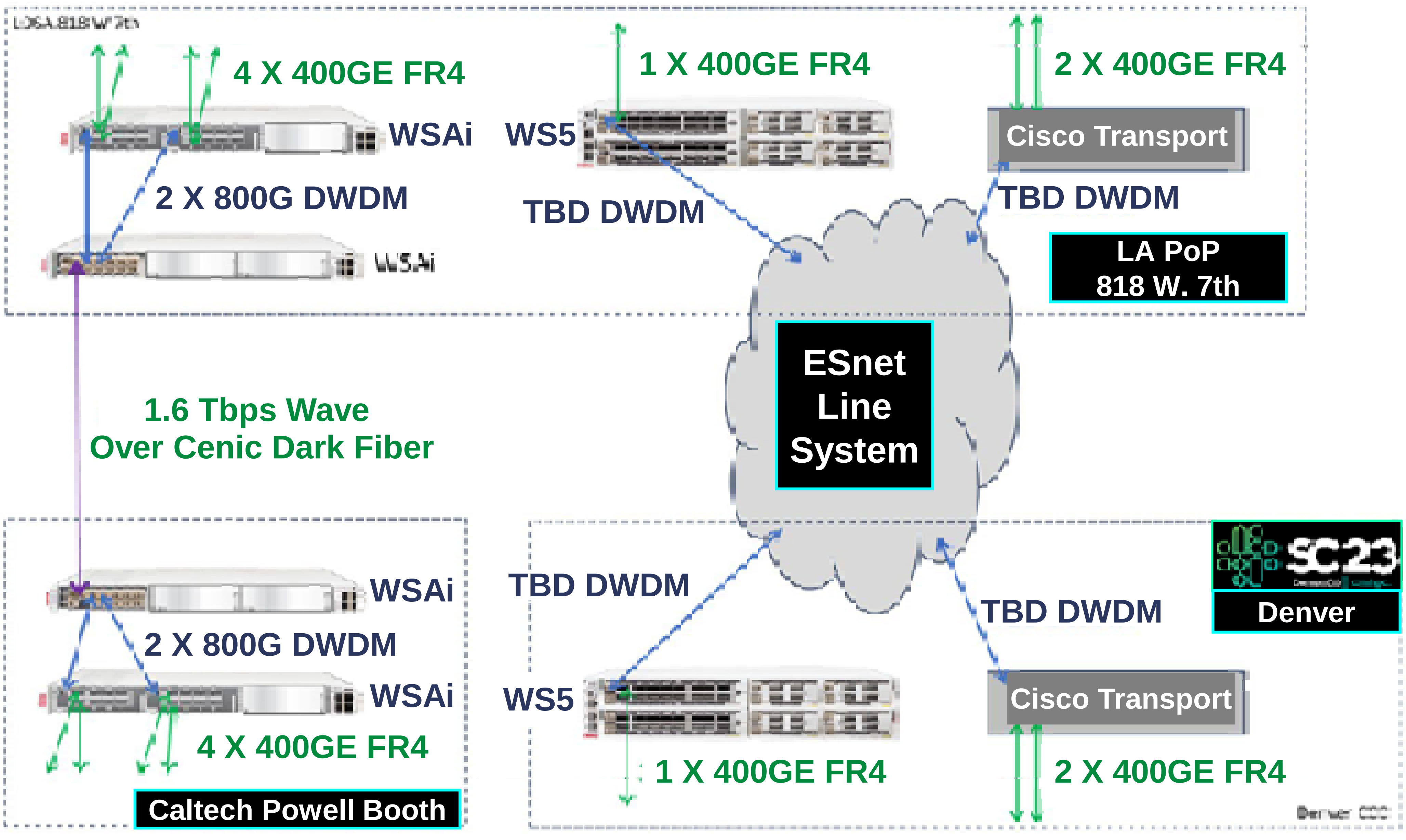
To ~3 Tbps in a single rack

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



SC23: 3 X 400G on ESnet Denver - LA: Ciena, Caltech and CENIC using WSAis and a dark fiber pair. Bringing 4 X 400GE via 2 800G Waves direct to the campus

Ciena WaveServer Ais and Waveserver 5s: Site Connections at the SC23 (Denver), the CENIC PoP (LA), and Caltech (Pasadena)c



CENIC, ESnet and Internet2 at the LA PoP

400G + 4 X 100G to Caltech via WS Ais

3 X 400G LA-Denver via ESnet

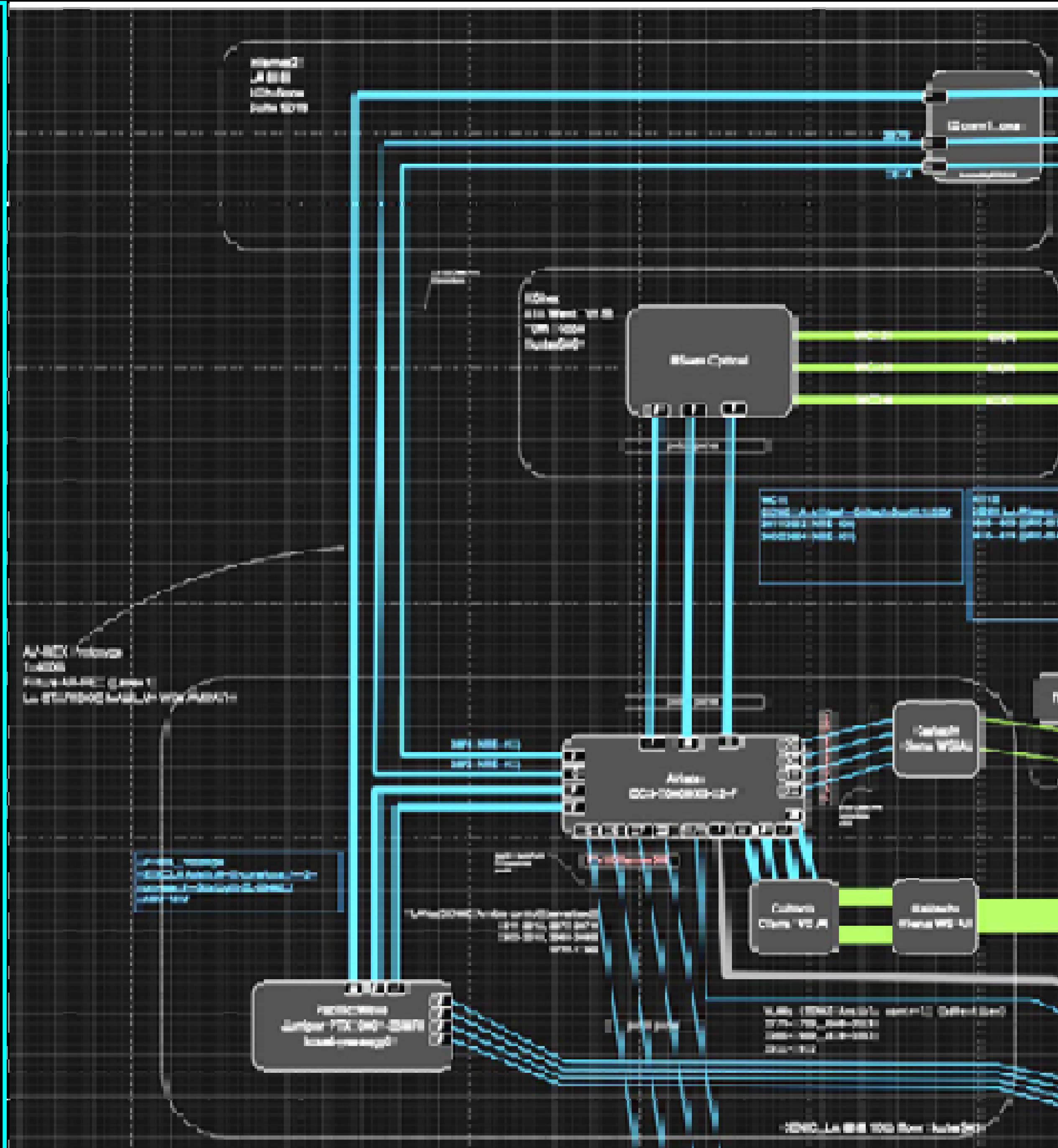
4 x 100G to UCSD/SDSC

2 X 400G to Pacific Wave via CENIC

Permanent:

400G NA-REX Prototype

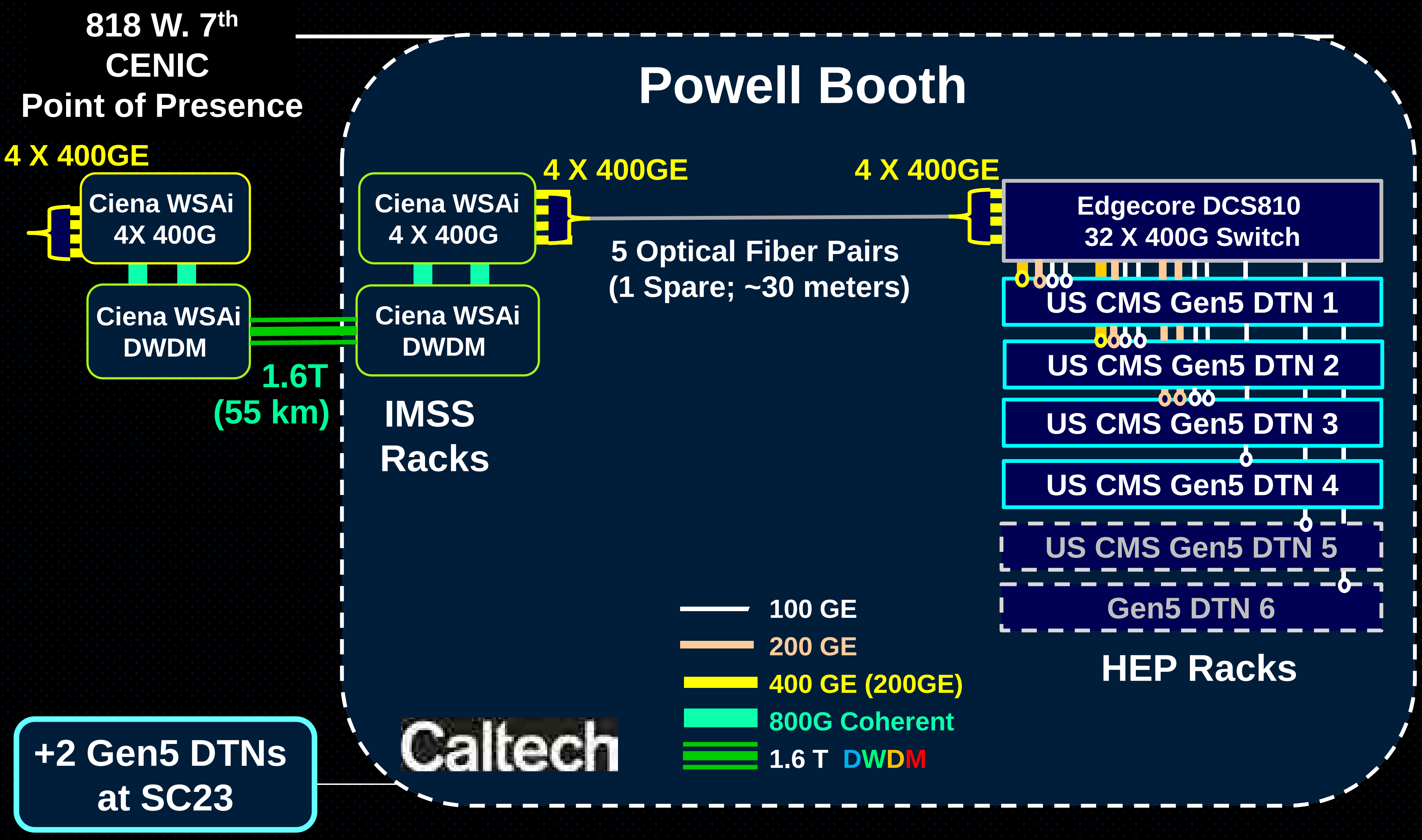
400G to ESnet Production



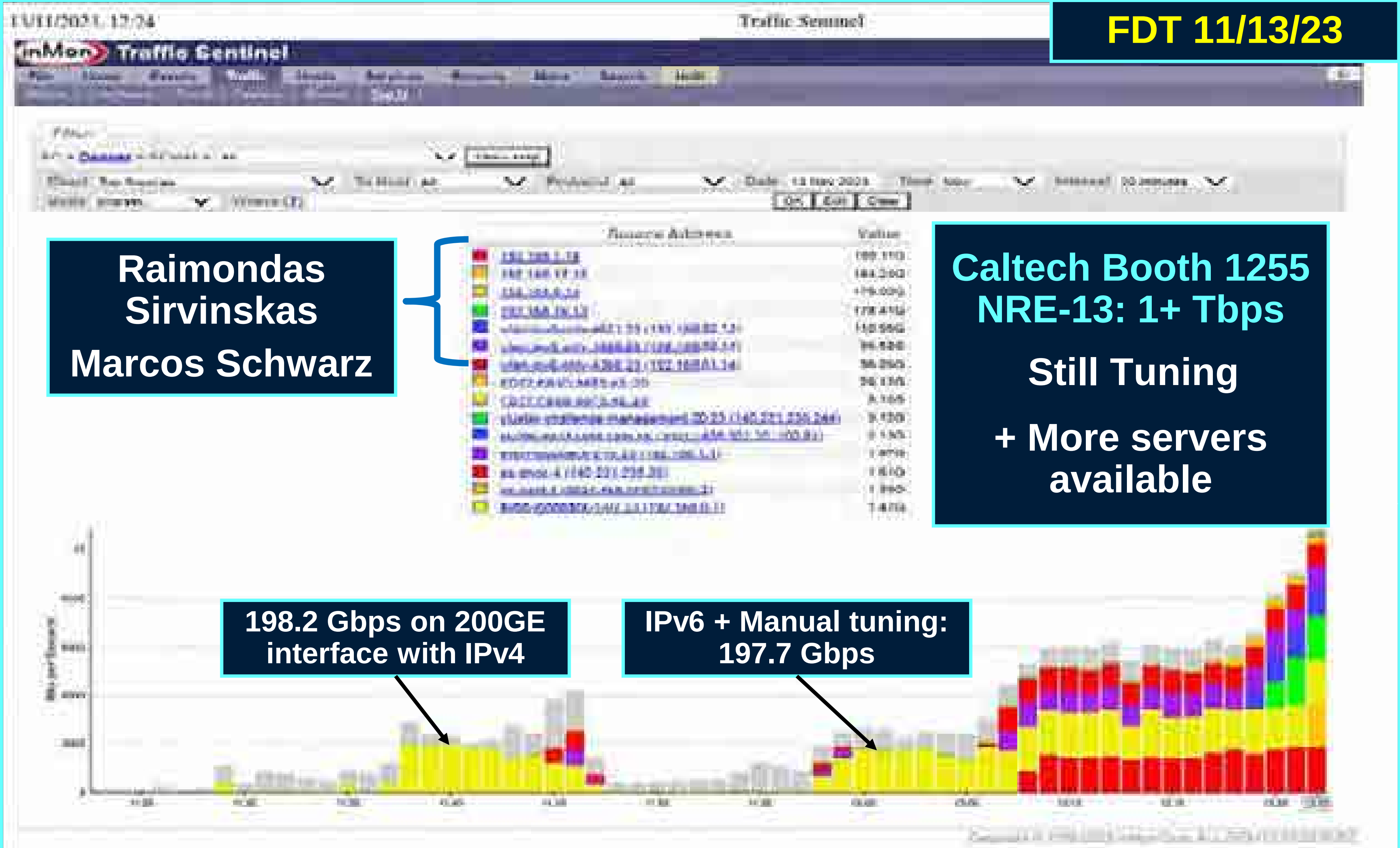
NA-REX: North America Research and Education Exchange

October 2023

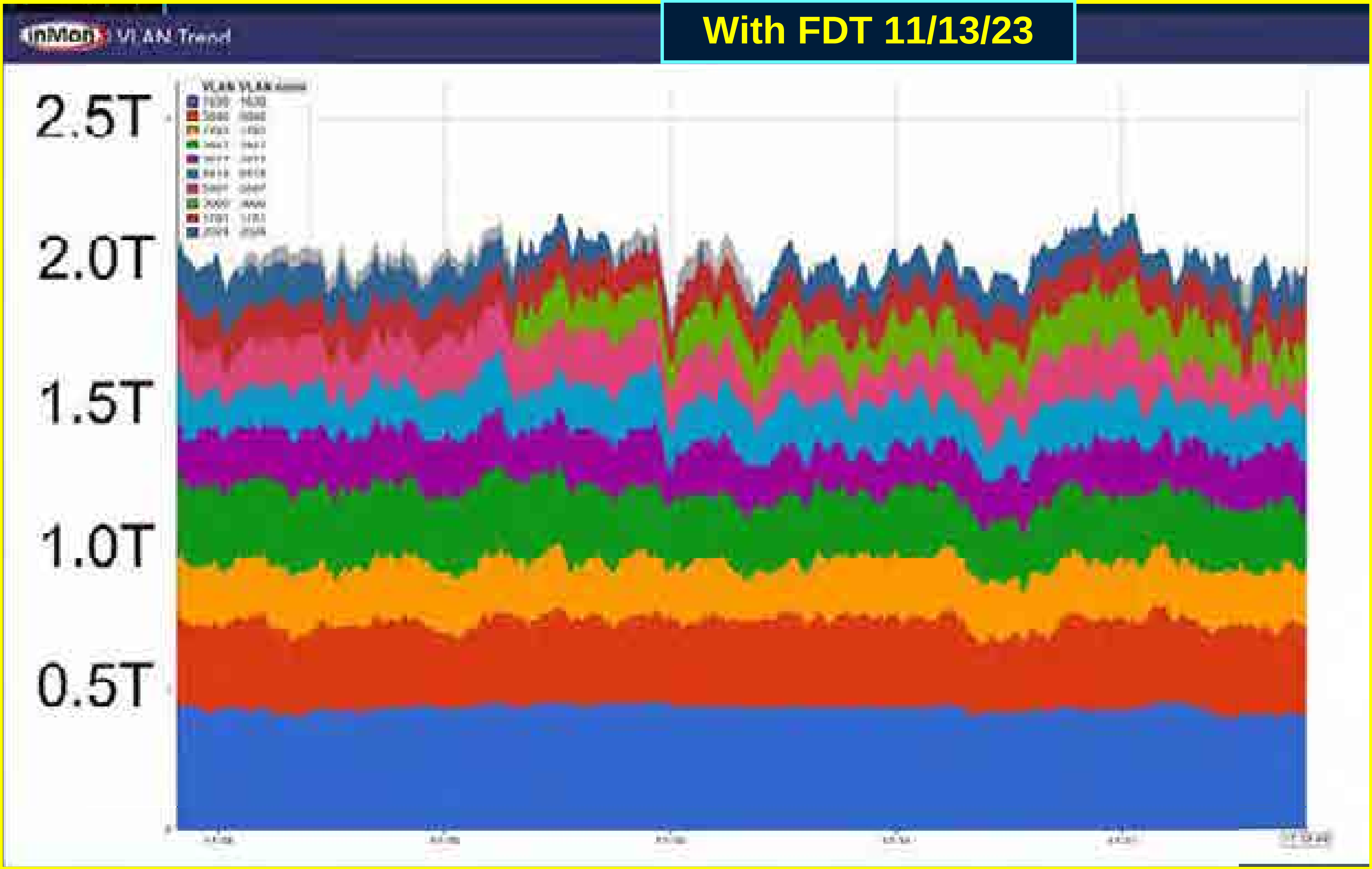
Simplified Caltech – LA Layout for SC23



NRE-13: 1+ Tbps with FDT: From One Pair of Gen5 Servers at Caltech To One Pair at SC23

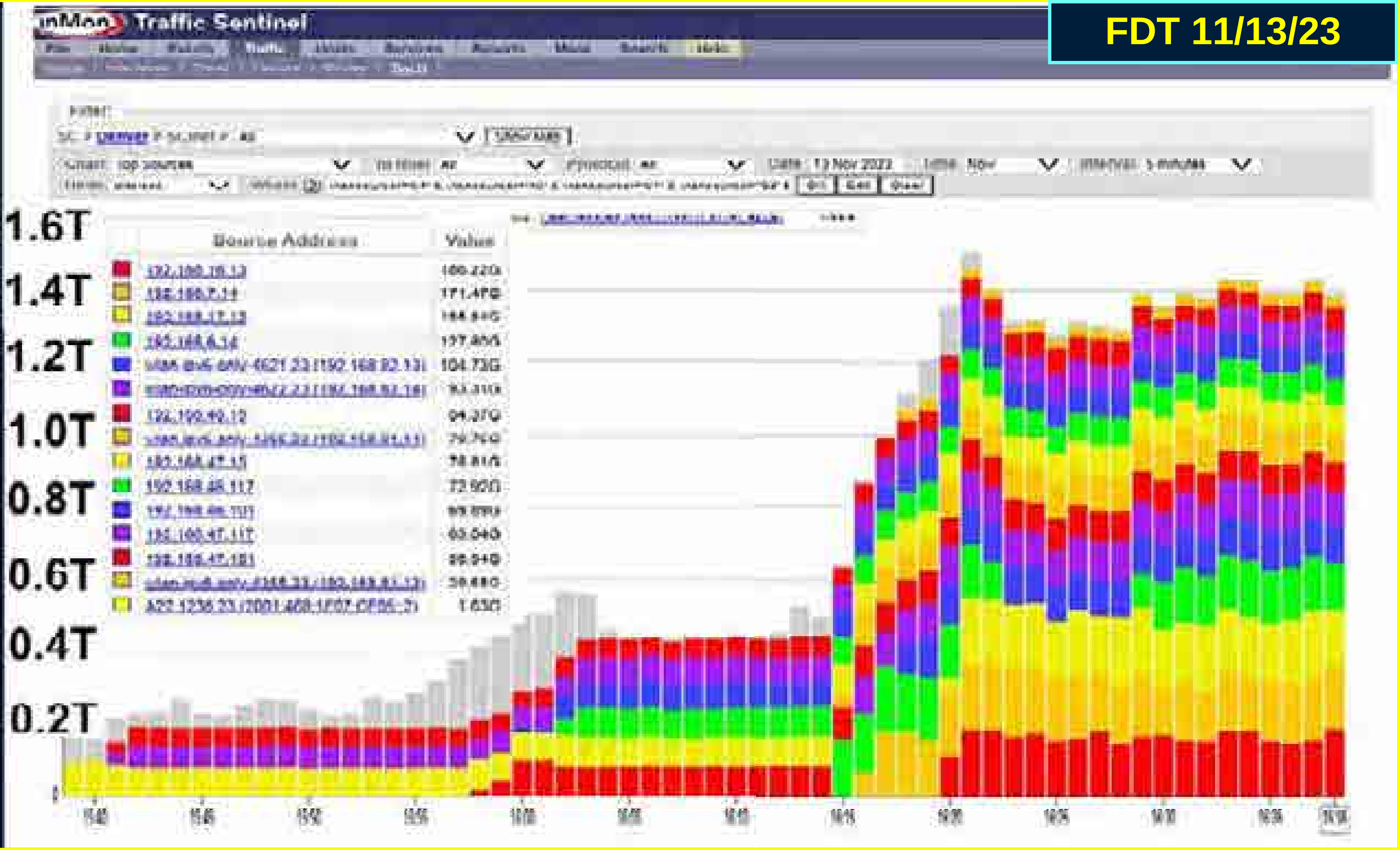


NRE-13 VLANs: To 1.4 Tbps of 2 Tbps

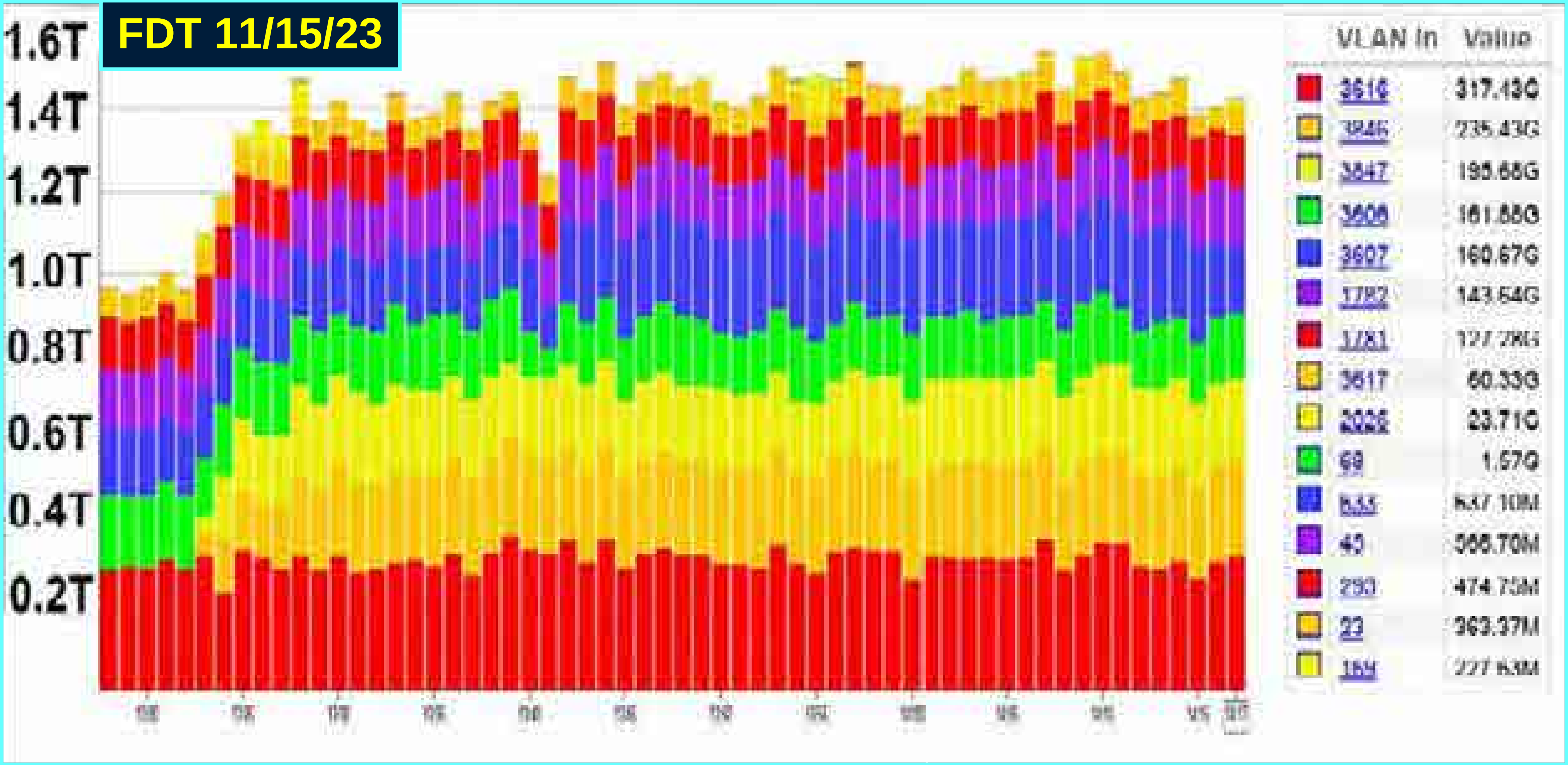


NRE-13 Top Sources: To 1.4 Tbps

FDT 11/13/23



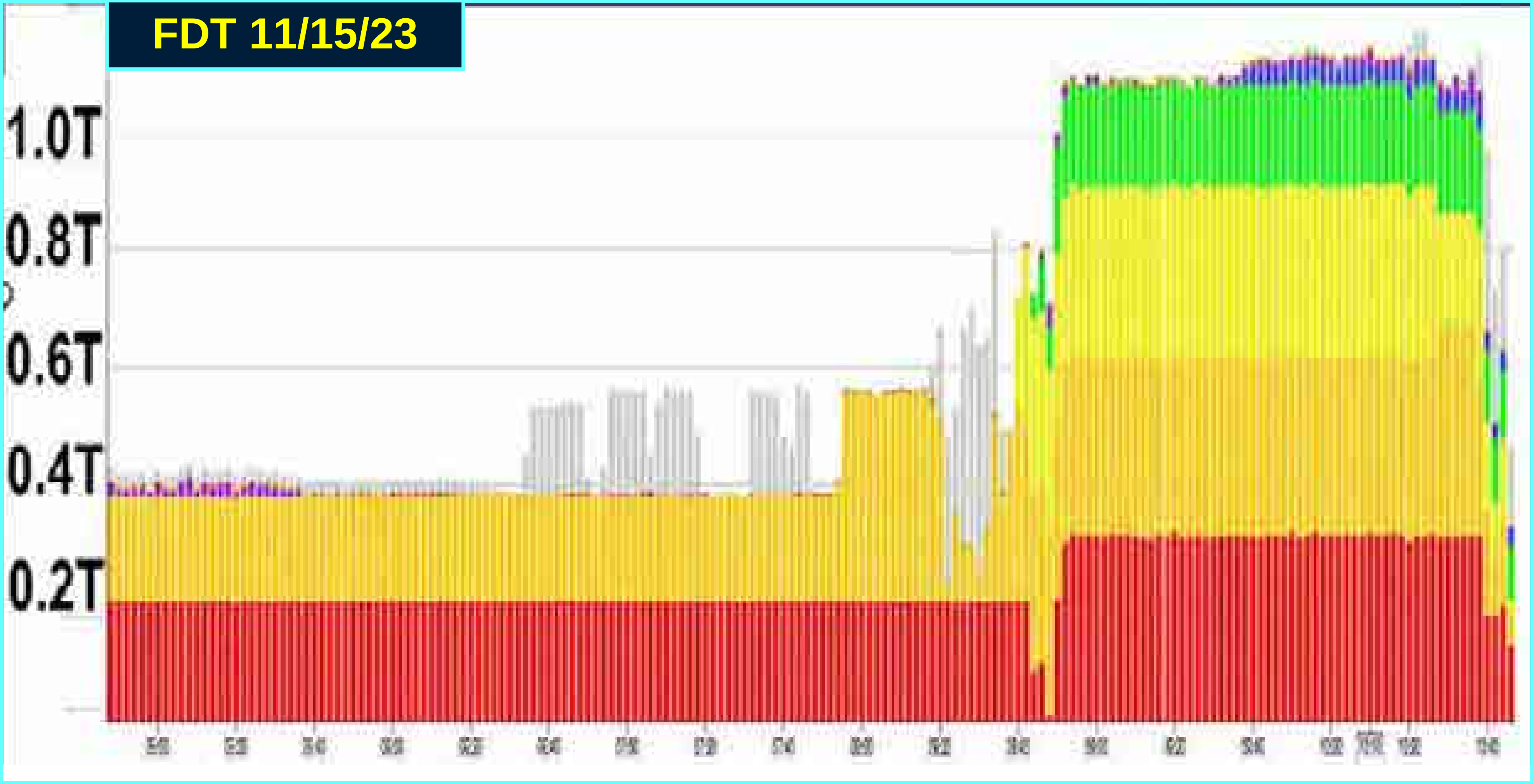
NRE-13 Top Sources: To 1.5+ Tbps on 4 X 400G Circuits with Dynamic Transfer Limit



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

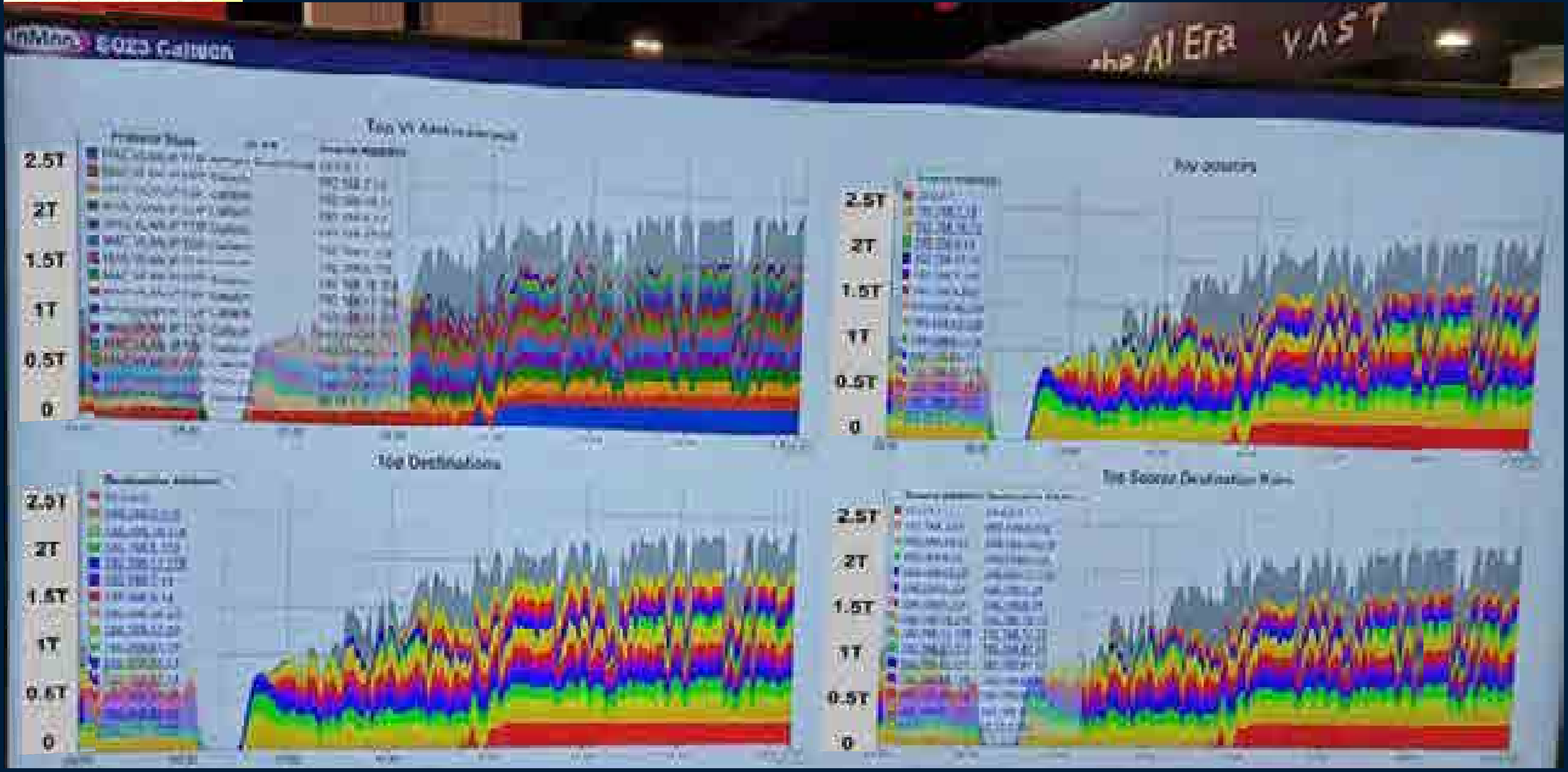
NRE-13: 1.1 Tbps on 2 X 400G Circuits

***Stabilized* with Dynamic Thread Management**



With Just 2 Gen5 Servers at SC23 and 2 at Caltech

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

General rules for better throughput [*]

Raimondas Sirvinskas and Marcos Schwarz

Lessons learned from previous Supercomputing conferences:

- We should not trust any kernel version, except the one(s) we have tested and confirmed to work well.

  For example kernel 4.18.0 comes with AlmaLinux 8.8 by default:
  It tested well on two 200Gbit links using single direction transfers
  but it failed when we started a third transfer on a third interface
  at the same time, or 2x 200Gbit bidirectional transfers

- Recommended kernel version was 6.5.10
- Ensure Jumbo frames is set on each interface and VLAN
- Ensure the CPU governor is set to performance
- Turn Adaptive RX off
- Set txqueuelen to 10000
- Set the network interface RX and TX buffers to the maximum supported values
- Set the interface to use BIG TCP (newer kernel feature available since 6.3)

[*] Also see https://www.dropbox.com/scl/fi/kgxjynwpbrqv5fkd2u6hc/SC23-path-to-high-throughput-1.pptx?rlkey=zlv3rmzqhzt08fmin651dso6z&dl=0

Narrative: The Road to High Throughput

- We started with an initial tuning parameter set from previous work during SC conferences. Initial tests showed that we could achieve 198 Gbps on a 200GE interface, and ~197 Gbps using IPv6 so we continued launching transfers on other interfaces.
- We then noticed that after reaching close to 400Gbps, we had lots of packet drops. This was unexpected, because we prepared the same servers at Caltech and got ~600 Gbps from each server doing bidirectional transfers.
- After quick investigations we saw that we were using default kernels at the SC23 booth versus the 6.5.10 kernel at Caltech.
  - So we updated the kernel on the booth servers and continued testing.
- The initial goal was to get better results than the previous year (at SC22) which was ~850Gbps. That was achieved during the first day of preparations.
  - As this was achieved quite easily, we continued work to get better throughput, and we got better results each day than the previous day.
- Common issues were that some tuning was missing on the server, and overloading of servers by running multiple transfers on each of multiple interfaces.
  - So we reduced the number of threads to the minimum needed for our transfers.
  - For example: we needed 4 threads to get 197.9 Gbps, and using 5 threads we got 198.2 Gbps (only +0.15%), so we decided to use 4 threads to lower the system load when running multiple transfers on multiple interfaces.

Raimondas Sirvinskas and Marcos Schwarz

Kernel Parameters

Most kernel tunings were from previous years but we still study all available parameters and possible effects on interface throughput or system load.

- Set the socket receive and send buffers in bytes to maximum: 256 Mbytes
- Turn on window scaling which can enlarge the transfer window
- Tell TCP to make decisions that would prefer lower latency
- Enable select acknowledgments (SACK)
- Maximize the amount of memory that any TCP receive buffer can allocate
- Maximize size of the receive queue
- Dynamically adjust the receive buffer size of a TCP connection
- Set the default queuing discipline to use for network devices - fq
- Set the time and number of packets softirqd can process in a polling cycle
- Turn timestamps off to reduce performance spikes related to timestamp generation
- Do not cache metrics on closing connections
- Set the congestion control algorithm (Cubic, BBRv3) that gives best results

Testing configuration using multiple threads

Run FDT on Server A:

java -jar fdt.jar

And then run FDT on Server B:

java -jar fdt.jar -c <Server_A_IP_ADDRESS> -nettest -P 2 # two threads

Two thread FDT output:

```
03/12 09:17:26  Net Out: 139.139 Gb/s   Avg: 139.139 Gb/s
03/12 09:17:31  Net Out: 139.985 Gb/s   Avg: 139.562 Gb/s
03/12 09:17:36  Net Out: 144.703 Gb/s   Avg: 141.266 Gb/s
03/12 09:17:41  Net Out: 132.307 Gb/s   Avg: 139.027 Gb/s
03/12 09:17:46  Net Out: 133.943 Gb/s   Avg: 138.010 Gb/s
03/12 09:17:51  Net Out: 140.426 Gb/s   Avg: 138.408 Gb/s
```

Add more threads to get maximum throughput and find the spot when performance stops increasing. Larger RTT may require more threads.

Kernel parameters: Comparing Congestion Control Algorithms

Congestion control:
net.ipv4.tcp_congestion_control=cubic   # Depends on the system and situation

Cubic versus bbr:

Cubic (better for Gen5 200G)

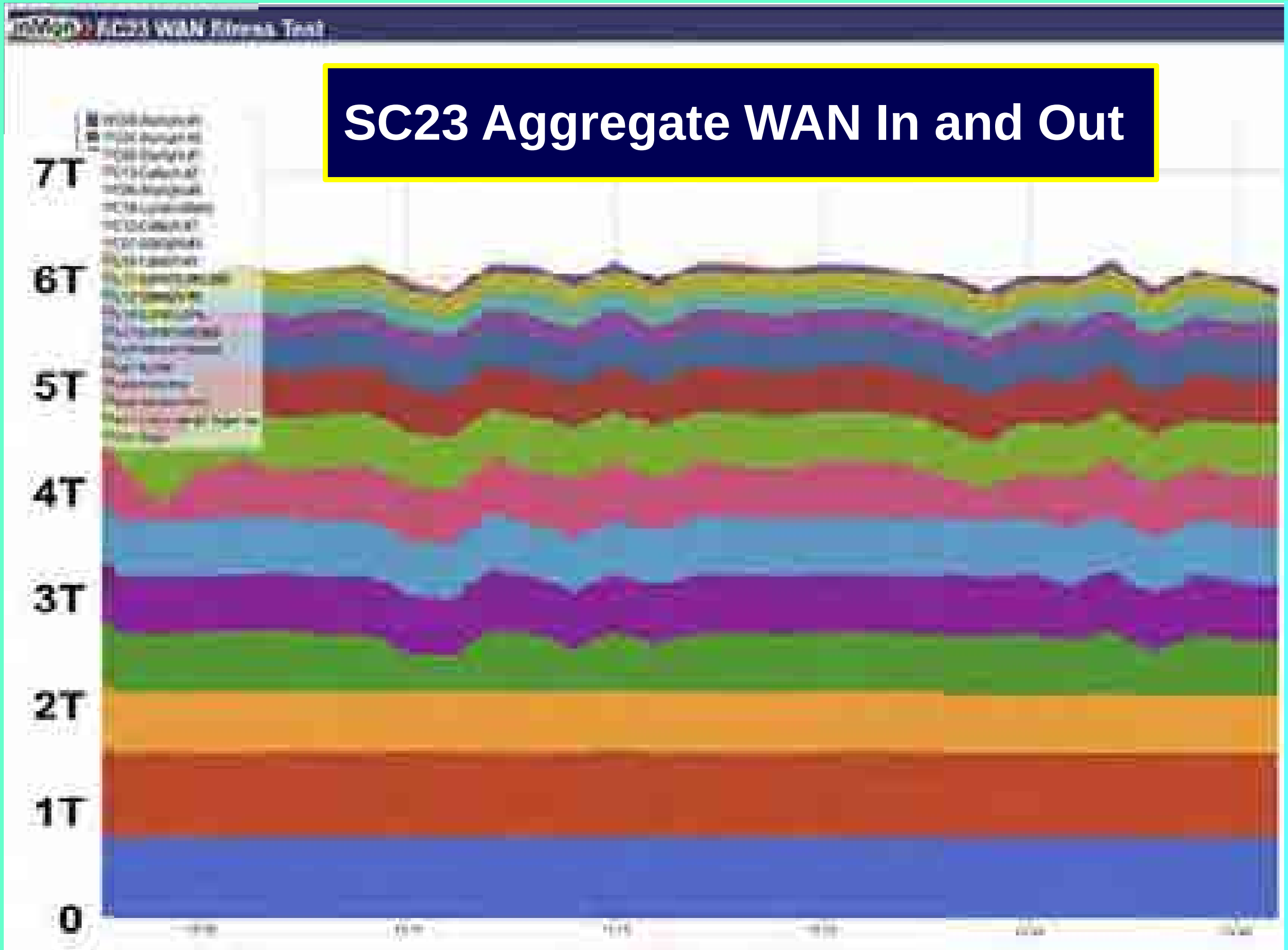
03/12 07:49:46	Net Out: 197.963 Gb/s	Avg: 197.963 Gb/s
03/12 07:49:51	Net Out: 197.941 Gb/s	Avg: 197.952 Gb/s
03/12 07:49:56	Net Out: 197.804 Gb/s	Avg: 197.890 Gb/s
03/12 07:50:01	Net Out: 197.762 Gb/s	Avg: 197.858 Gb/s
03/12 07:50:06	Net Out: 197.985 Gb/s	Avg: 197.875 Gb/s
03/12 07:50:11	Net Out: 197.706 Gb/s	Avg: 197.847 Gb/s
03/12 07:50:16	Net Out: 197.874 Gb/s	Avg: 197.851 Gb/s
03/12 07:50:21	Net Out: 198.116 Gb/s	Avg: 197.879 Gb/s
03/12 07:50:26	Net Out: 197.920 Gb/s	Avg: 197.884 Gb/s

BBR (but better for Gen3 100G)

03/12 08:09:38	Net Out: 172.095 Gb/s	Avg: 172.095 Gb/s
03/12 08:09:43	Net Out: 153.968 Gb/s	Avg: 163.031 Gb/s
03/12 08:09:48	Net Out: 163.870 Gb/s	Avg: 163.308 Gb/s
03/12 08:09:53	Net Out: 162.587 Gb/s	Avg: 163.122 Gb/s
03/12 08:09:58	Net Out: 161.984 Gb/s	Avg: 162.894 Gb/s
03/12 08:10:03	Net Out: 156.793 Gb/s	Avg: 161.872 Gb/s
03/12 08:10:08	Net Out: 163.431 Gb/s	Avg: 162.095 Gb/s
03/12 08:10:13	Net Out: 167.796 Gb/s	Avg: 162.807 Gb/s
03/12 08:10:18	Net Out: 166.956 Gb/s	Avg: 163.265 Gb/s

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
- **NVMe SSD Front End** Operations + HSM
- **PCIe 6.0 and CXL** DTN tests by ~SC24
- **SENSE 400G paths:** ESnet production, NA-REX via StarLight; Links to CERN

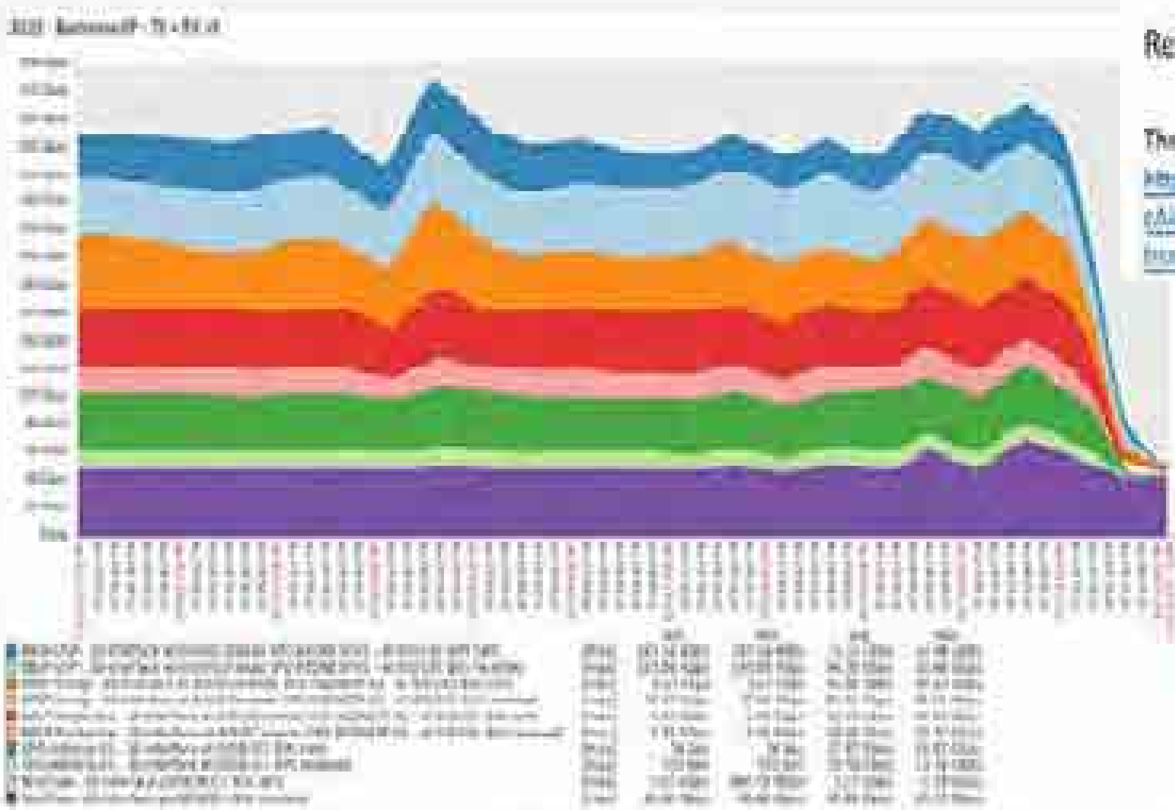
FDT 11/16/23

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

This Just In: 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/[illegible]

SC23 NRE-13: Take Away Messages

A Major Milestone pushing us forward on many fronts

- We had great results, reflected in the demos and presentations before and during SC23.
  - This will be fleshed out through the reports/feedback from each partner or hosted NRE
- We now have two global testbeds with expanding capabilities.
  - Beyond virtual circuits alone, we can do traffic engineering at the edge and in the core.
  - Applications such as FDT also can limit the sending or receiving rate stably, so these capabilities can be impedance matched, for precise scheduling of large flows.
- There are many other important emerging capabilities: Including the programmable Global P4 Lab including Bluefield2 and other smart edge devices, the Container-Lab based digital twin, ESnet High Touch, NOTED among Tier1s, PolKA and SRv6, among others
- Both the GNA-G Leadership Team and our DIS working group are seeking a system-level path to the next generation advanced network, and the architectural structure(s) and operations that go with it.
- There is an increasing gulf between current capabilities and the requirements as pre-conceived in 2020-22. Actual requirements will be in the middle, also exploiting then-current technology.
- Forward looking exercises using/stressing current capabilities as they emerge are needed: to properly gauge future requirements, and to feed into and craft effective system designs.
- We have important permanent elements left behind after SC23: Including The 400G link to the ESnet production network in LA which is useful for DC24 and beyond, and the 400G link between the CENIC Juniper and StarLight with 2 X 400G to the SENSE-controlled Arista in LA.
- We are also discussing the possibility of keeping the additional fiber pair between the Caltech campus and LA with CENIC, which would have multiple uses.
- With Mariam Kiran (now at ORNL) we will resume the effort on using machine learning/AI to optimize network operations: tactically; and with the emerging system-level picture – strategically

Next Generation Network-Integrated System

- Top Line Message: To realize the physics discovery potential and meet the challenges of the HL LHC era, 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)
- Simultaneously supports the LHC experiments, 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
- **SC23 is a Major Milestone, and a Leap Forward Towards this Goal**

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

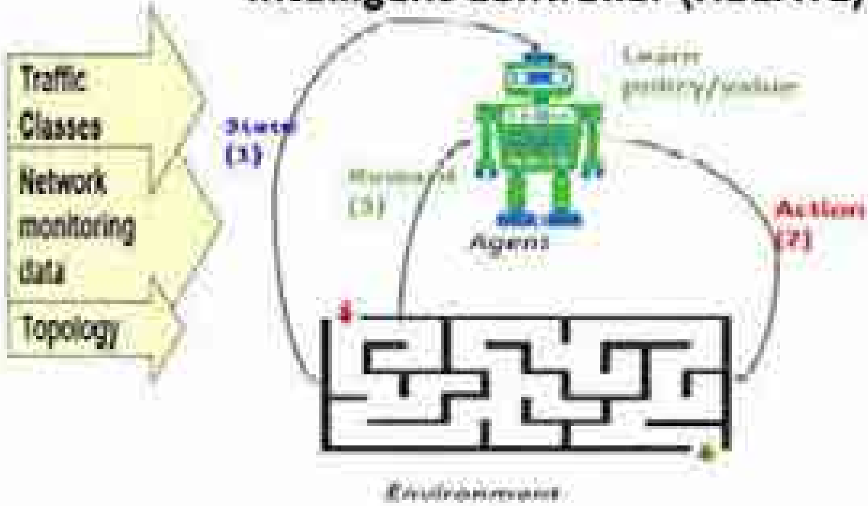
Mariam Kiran (ESnet) et al, Intelligent Networks DOE Project

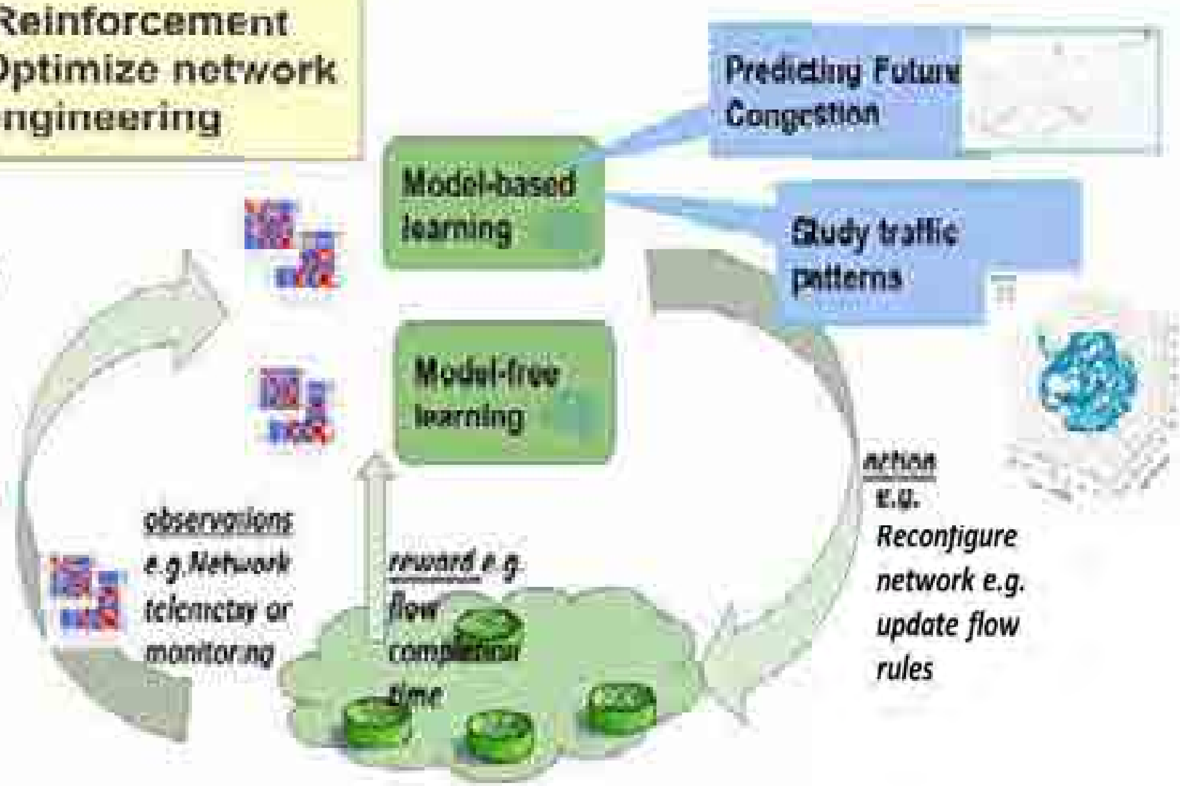
Self-Driving Network for Science

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

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

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

Related Work: CHEP2023 Talks and Poster

https://indico.jlab.org/event/459/ May 2023

- **58. J. Balcas (Caltech) et al., Track 1, Monday 2 PM:**
  "Automated Network Services for Exascale Data Movement"
- **125. S. Shannigrahi (Tennessee Tech) et al., Track 1, Monday 2:15 PM: "A Named Data Networking Based Fast Open Storage System Plugin for XRootD"**
- **584. Y. Richard Yang (Yale) et al., Track 1, Monday 2:30 PM:**
  "ALTO/TCN: Toward an Architecture of Efficient and Flexible Data Transport Control for Data-Intensive Sciences using Deep Infrastructure Visibility"
- **32. A. Sim (LBL) et al., Track 4, Tues. 11:15 AM: "Predicting Resource Usage Trends with the Southern California Petabyte Scale Cache"**
- **450. P. Bhat (Caltech) et al., Poster Session Tues. 3:30 PM:**
  "Scientific Community Transfer Protocols, Tools & Their Performance Based on Network Capabilities"
- **60. C. Guok (ESnet) et al., Track 7, Tues. 5:15 PM: "Complete End-to-End Network Path Control for Scientific Communities with QoS Capabilities"**
- **59. J. Balcas (Caltech) et al., Track 1, Thurs. 11:45 AM:**
  "Job CPU Performance Comparison Based on MiniAOD Reading Options: Local vs. Remote"
- **614. A. Arora (UCSD) et al., Track 1, Thurs. 2:45 PM:**
  "400 Gbps Benchmark of XRootD HTTP-TPC"

PolKA References and Presentations

- References
- [1] C. Dominicini et al., "PolKA: Polynomial Key-based Architecture for Source Routing in Network Fabrics," 2020 6th IEEE Conference on Network Softwarization (NetSoft), 2020, pp. 326-334, doi: 10.1109/NetSoft48620.2020.9165501.
  [2] R. S. Guimarães et al., "M-PolKA: Multipath Polynomial Key-Based Source Routing for Reliable Communications," in IEEE Transactions on Network and Service Management, pp. 2639-2651, Sept. 2022, doi: 10.1109/TNSM.2022.3160875.
- [3] Jordi Ros-45/33667Giralt et al., "On the Bottleneck Structure of Congestion-Controlled Networks". Proc. ACM Meas. Anal. Comput. Syst. (2019). https://doi.org/10.1107

- PolKA Applications or future demonstrations (with further development):
  [1] In-situ Proof-of-Transit for Path-Aware Programmable Networks https://ieeexplore.ieee.org/document/10175482

  [2] In-Band Telemetry: https://link.springer.com/chapter/10.1007/978-3-031-28451-9_45

  [3] Initiatives on using QTBS integrated with PoLKA for TE Optimization

  - https://docs.google.com/presentation/d/1oatze9SEOCtf8rIpsry7ocWXwBwL08xEQptq4GpLRRg/edit?usp=sharing
  - https://docs.google.com/presentation/d/1qnZyEnF6by41ERwBo9Uv3BB_BrbG3YhrP1ZbKZyTEQM/edit?usp=sharing
  - https://sol.sbc.org.br/index.php/wpeif/article/view/24659 n

Next Generation System for Data Intensive Sciences

- The development of effective ML optimization methods, and multidimensional, real-world metrics
  - Are themselves challenging, groundbreaking activities
  - A new area of application in multidomain distributed systems
- 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

Global P4 Lab

Dataplane of P4 and Other Programmable Switches

GEANT + RNP
Frederic Loui
Marcos Scwarz
et al.

2023: From 26 to 38 Sites

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

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