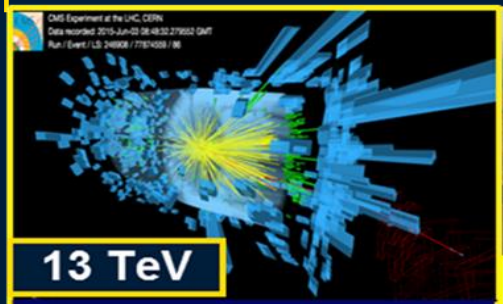
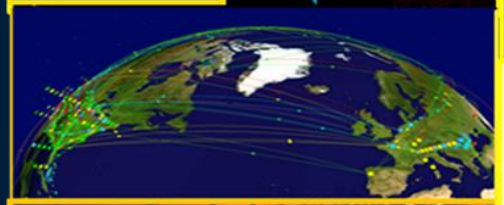
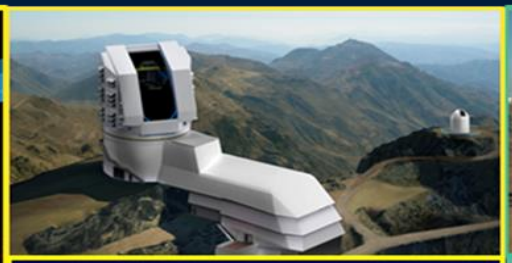
Global Networks for HL LHC and Data Intensive Sciences

1.5 Tbps+ Trials and Lessons Learned: from SC23 and Beyond

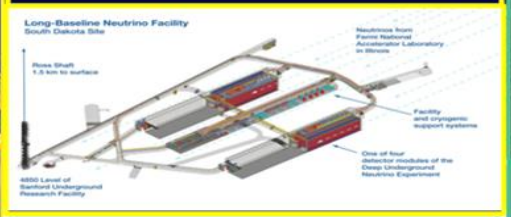


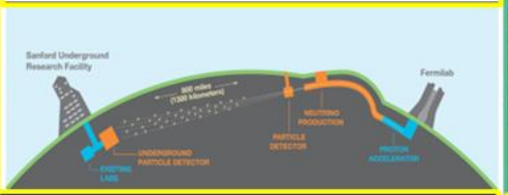


LHC



Rubin Observatory





LBNF/DUNE

SKA

LHC Run 3
and HL-LHC

Rubin
Observatory

SKA

BioInformatics

Earth
Observation

Gateways
to a New Era

Harvey Newman, Caltech

DOMA/WLCG General Meeting

June 26, 2024

GNA-G
Global Network Advancement Group

CALIFORNIA INSTITUTE OF TECHNOLOGY 1891

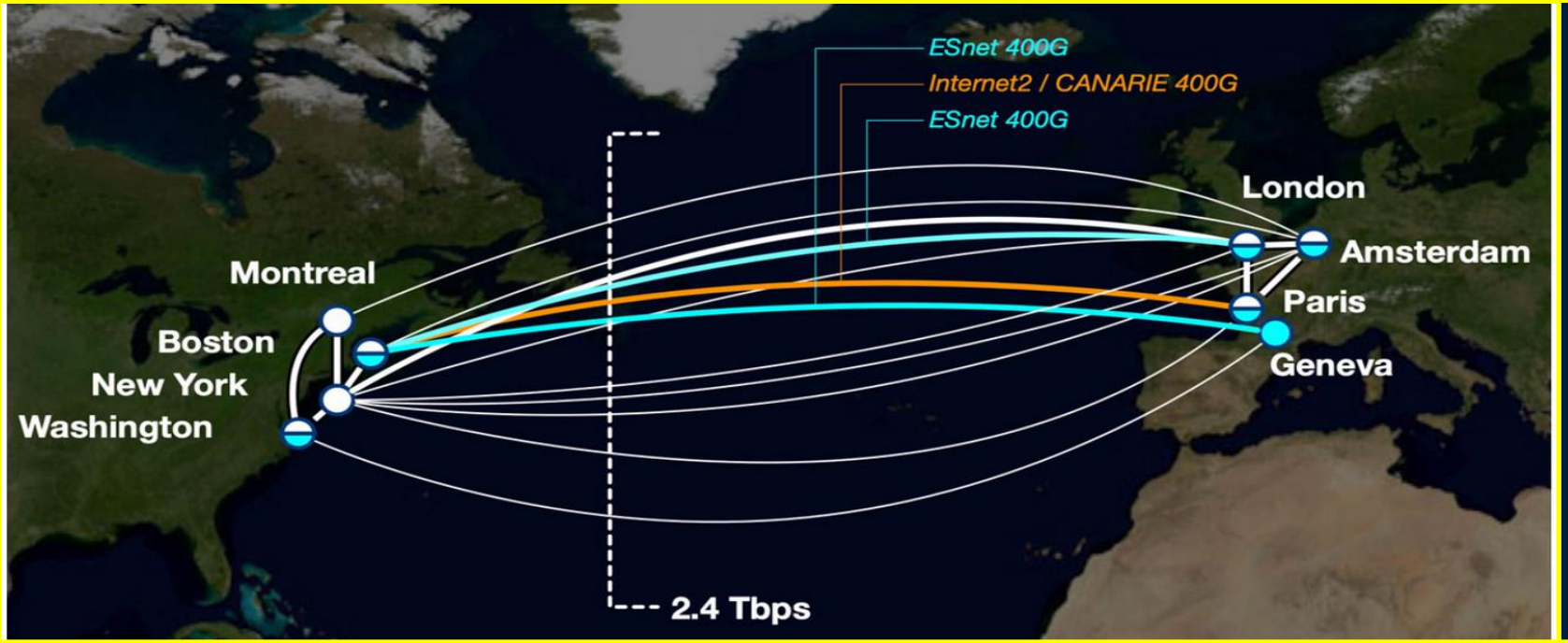
GNAG-G Context and Points for Today

Goal: Appropriately Frame Requirements and Challenges

- **LHC to HL LHC Challenges: Scale, Complexity and Global Extent**
  - **Workflow scale & complexity not fully captured in requirements documents so far**
  - **Scaling from N X 100G to N X 400G links: nationally, transoceanic, on campuses**
  - **Data flow growth now, and projections, point towards network-constrained infrastructures, and hence Stateful, potentially complex decisions on resource use**
  - **No explicit consideration of Analysis Use Case aggregate flow requirements so far**
  - **Full Throughput Tools (even today) are a central element; for setting requirements**
- **Work towards a comprehensive next generation system to respond to the challenges has progressed since 2019:**
  In the Global Network Advancement Group, its DIS and SENSE WGs, ICFA SCIC [*]
  Partnering with many R&E network and project partners (OSG, NRP, DOMA, etc.)
- **System Characteristics**
  - **Computing, storage, networks: all as first class, jointly managed resources**
  - **Designed to carefully allocate network resources Within Limits; to meet the challenges while accommodating the traffic supporting the at-large A&R community**
  - **Integrated with LHC workflow: Rucio/FTS/XRootD; also applicable to DUNE, others**
  - **Parallel lines of development, and progressive integration; Using the SENSE and GP4L worldwide testbeds + StarLight, FABRIC; Transoceanic Link Consortia**
- **[*] Ongoing work with the new ICFA Data Lifecycle Panel from Spring 2024**

The Upgraded ANA (to 2.4 Terabits/sec): ESnet, Internet2, GEANT, CANARIE Transoceanic Links and Intercontinental Partnerships

ANA's network expansion supports multinational, data-intensive science collaborations, including the Large Hadron Collider (LHC), the world's largest and most powerful particle accelerator, and the Square Kilometer Array (SKA), the ongoing effort to build the world's largest radio astronomy observatory. It adds much-needed capacity for transmitting instrument findings to researchers globally, enabling ground-breaking discoveries.

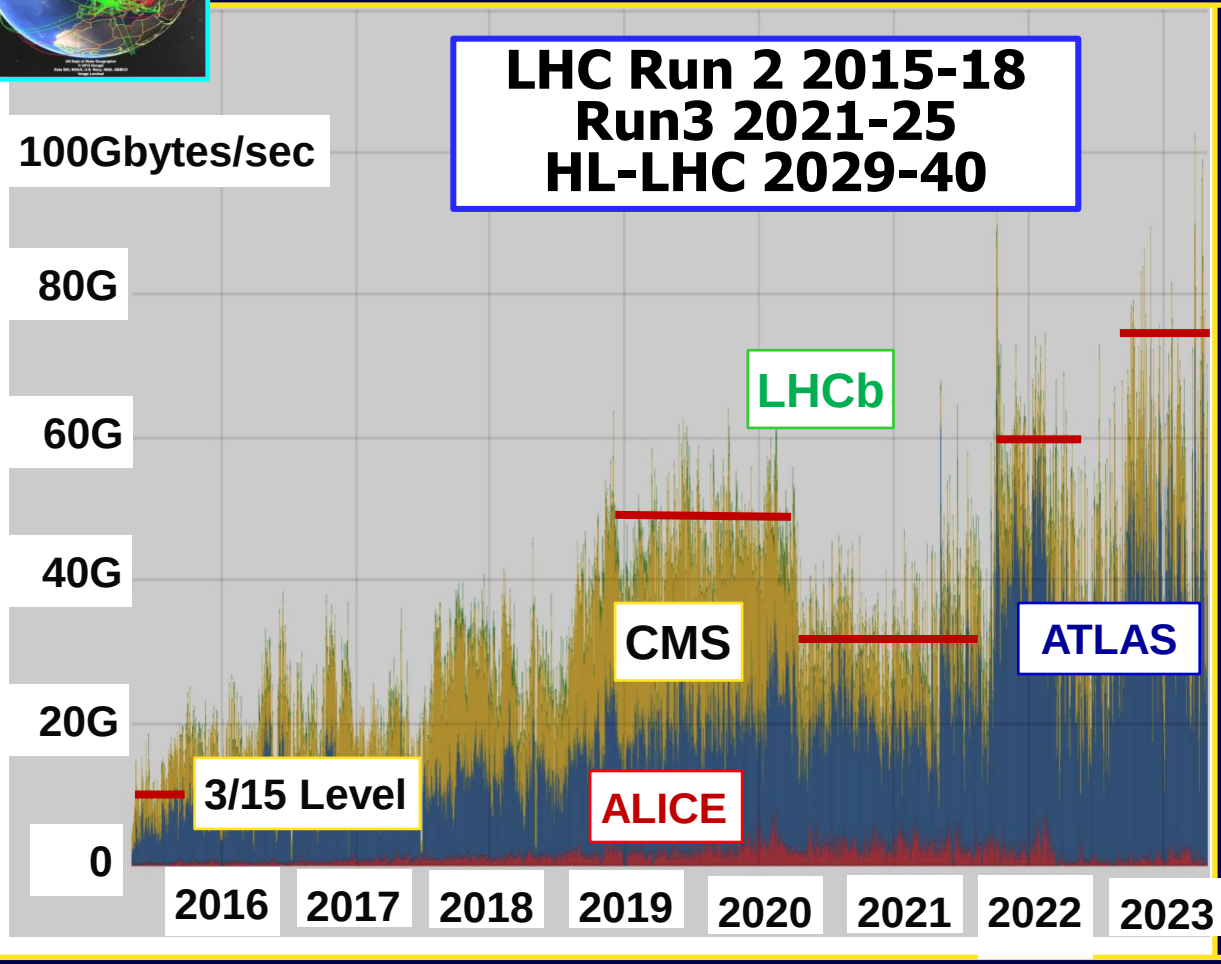


The joint effort adds three new 400 Gbps spectrum circuits between exchange points in the U.S., U.K., and France. The connections utilize the record-breaking 400 terabits per second (Tbps) transAtlantic Amitié subsea cable system spanning 6,783 kilometers. The flexibility and scalability of these spectrum circuits enable significant capacity growth, through future upgrades at exchange points.

https://internet2.edu/canarie-esnet-geant-and-internet2-unveil-highest-capacity-transoceanic-connectivity/

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

WLCG Transfers Dashboard: Throughput April 2015 – April 2023



30-85 GBytes/s Weekly Avg

To 100+ GBytes/s Daily Avg

Complex Workflow

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

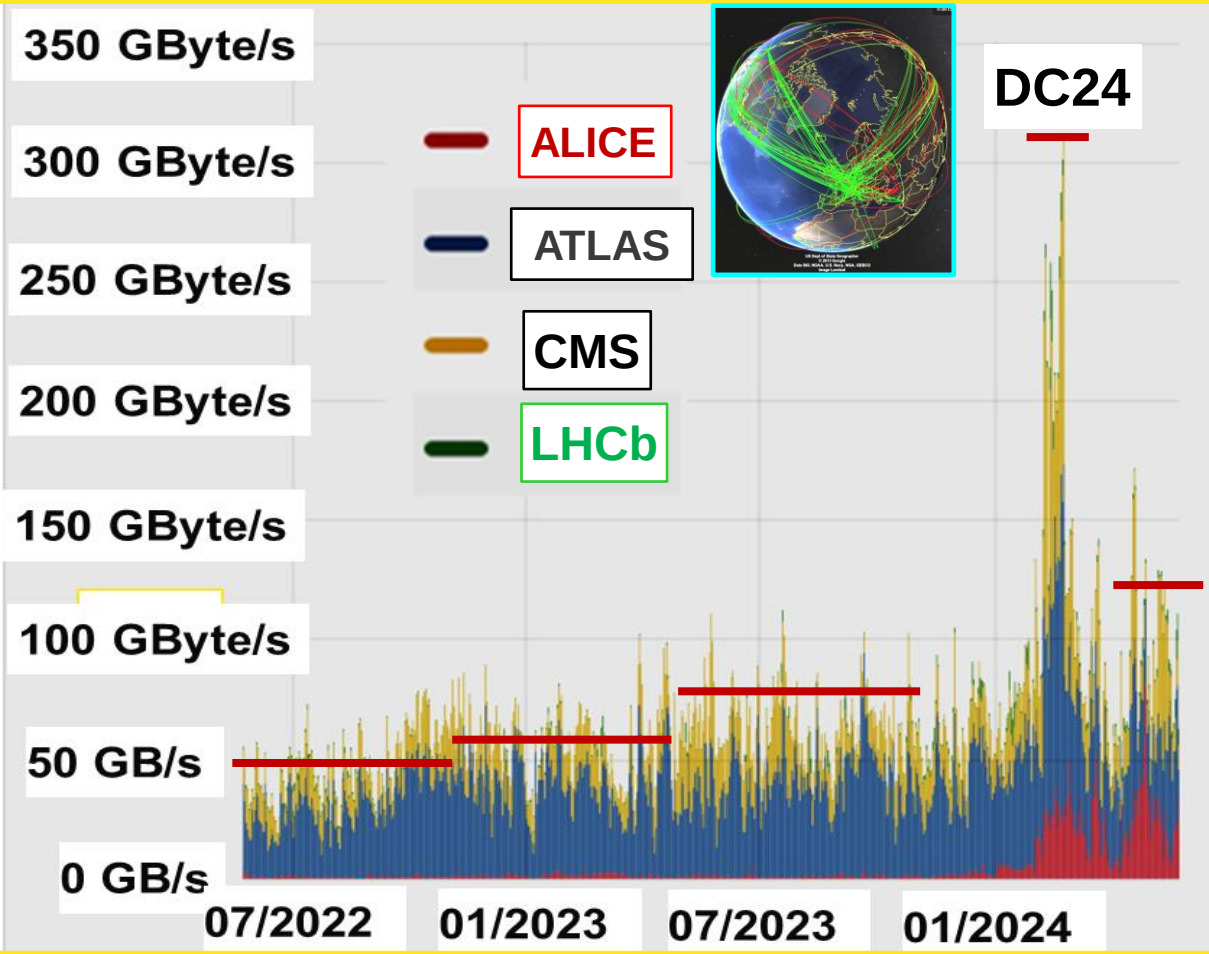
- Effects of Covid from Spring 2020 are evident
- Fast recovery also evident

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

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

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

WLCG Transfers Dashboard: Throughput June 2022 – May 2024



70-150 GBytes/s Weekly Avg

To 170+ GBytes/s Daily Avg

Complex Workflow

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

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

~1.7 to 2X Growth in 12 Months: 200-1000X Per Decade Equivalent (?)

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

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 (50GBytes/s, i.e. 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":** $\sum$ (ATLAS,ALICE,CMS,LHCb)
  *2 (for bursts) *2 (overprovisioning) = ***4.8Tbps***
- **Flexible model**
- Assumes reading of data from above for reprocessing/reconstruction within 3 months
- Means doubling the Minimal Model: ***9.6Tbps***

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

Supercomputing 2002 BWC : Baltimore Nov. 16-22

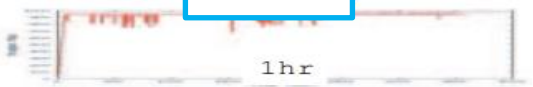
Breakthrough: Stable High Throughput for Hours over Transcontinental and Transatlantic Distances

Aggregate Throughput Traces from 1 to 10 Flows: to 9.1 Gbps in 2002

FAST TCP

1 Flow for 1 Hour

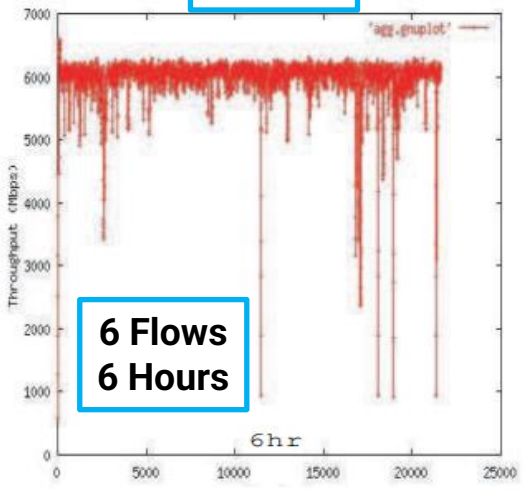
95%



1 flow

90%

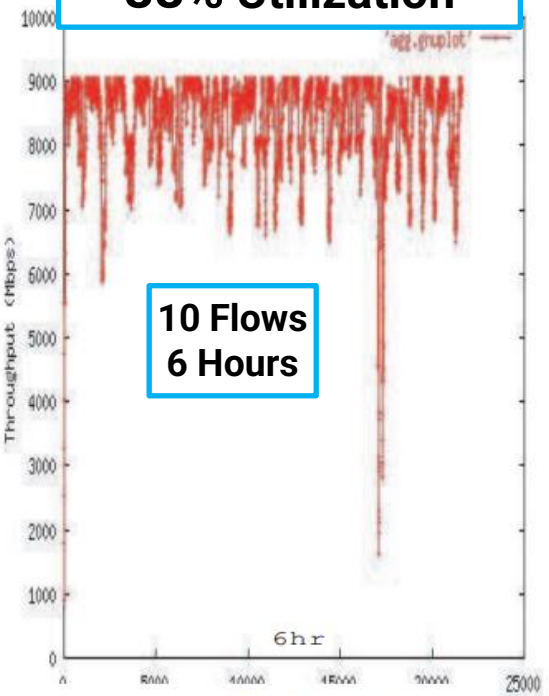
6 Flows 6 Hours



7 flows

88% Utilization

10 Flows 6 Hours



10 flows

10G Links:
US LHCNet (Caltech)
Internet2
TeraGrid
CERN

10G Links
Among Sites:
Caltech, SLAC, Sunnyvale, StarLight, Baltimore (3948 km)
CERN (10,037 km)

#flow	throughput Mbps	utilization	delay ms	distance km	duration s	bmps 10^{15}
1	925 (266)	95% (27%)	180	10,037	3,600	9.28 (2.67)
2	1,797 (931)	92% (48%)	180	10,037	3,600	18.03 (9.35)
7	6,123	90%	85	3,948	21,600	24.17
9	7,940	90%	85	3,948	4,030	31.35
10	8,609	88%	85	3,948	21,600	33.99

Stability:
Gentle interactions in the presence of other aggressive protocols competing for bandwidth

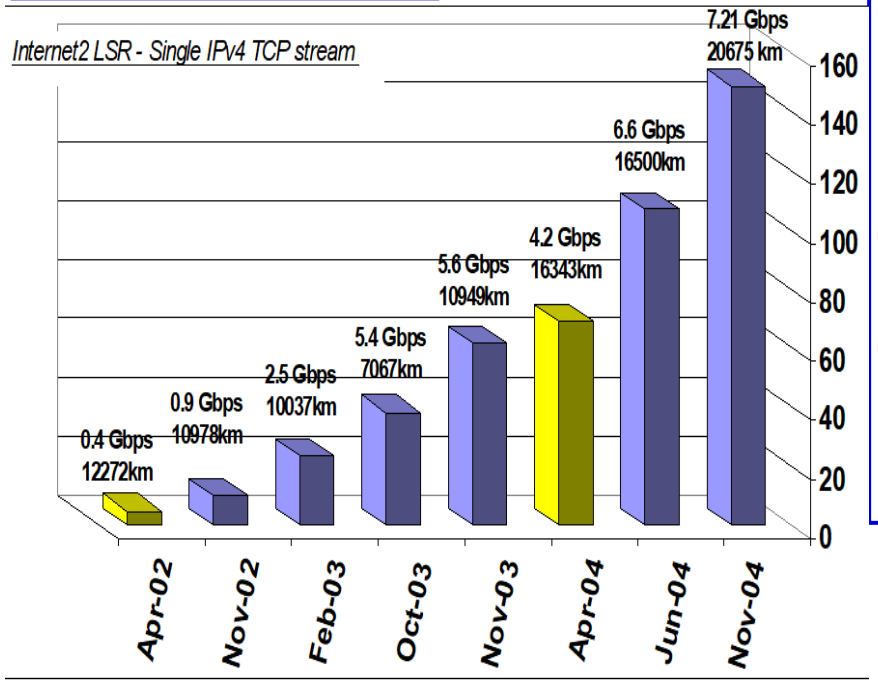
Internet2 Land Speed Records 2002-4

Nov. 2004 Record Network

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

Internet2 LSRs:
Blue = HEP

7.2G X 20.7 kkm



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

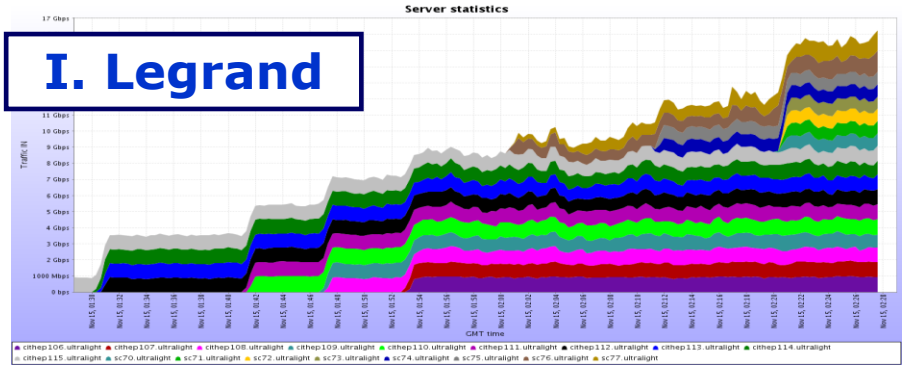
[Compare to 2024: to ~800G per server with Caltech's FDT]

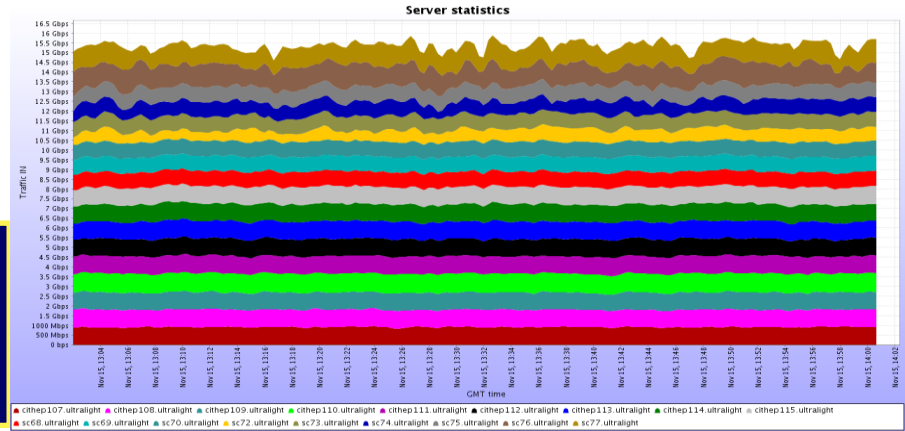
SC06-SC08 BWC: Fast Data Transfer
http://monalisa.cern.ch/FDT

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

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

I. Legrand





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

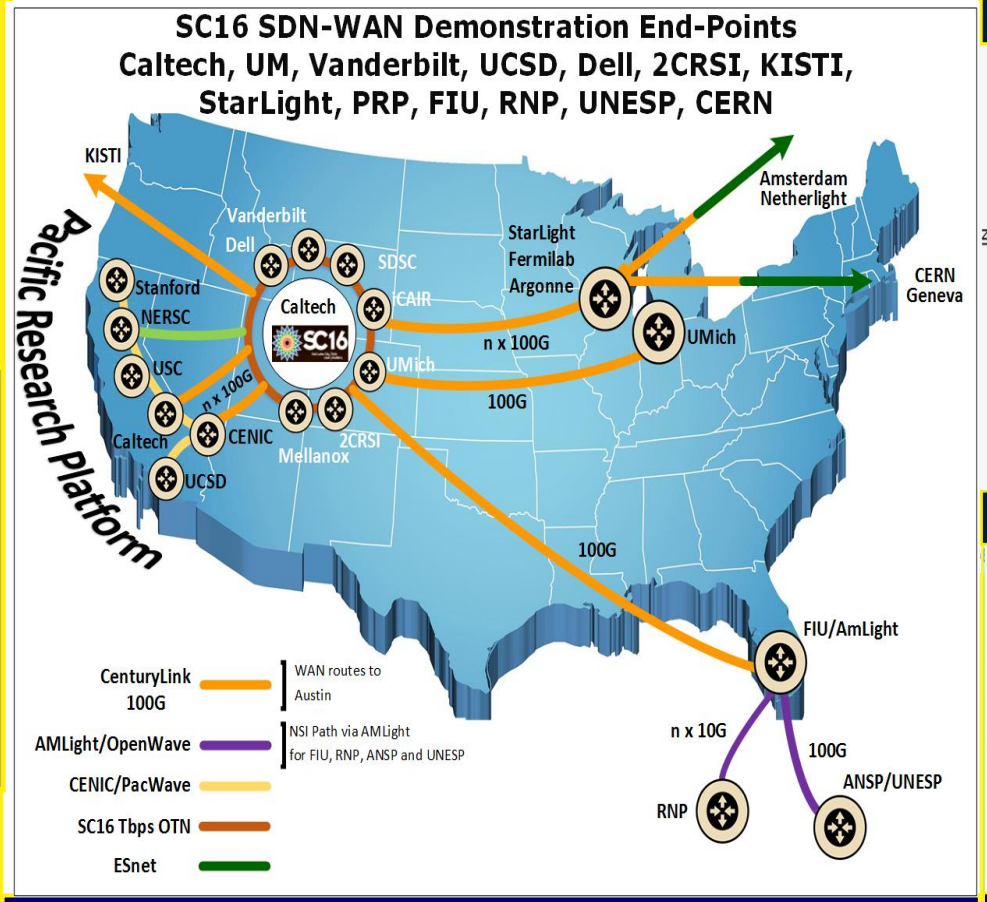
SC15-23: SDN Next Generation Terabit/sec Ecosystem for Exascale Science

supercomputing.caltech.edu

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

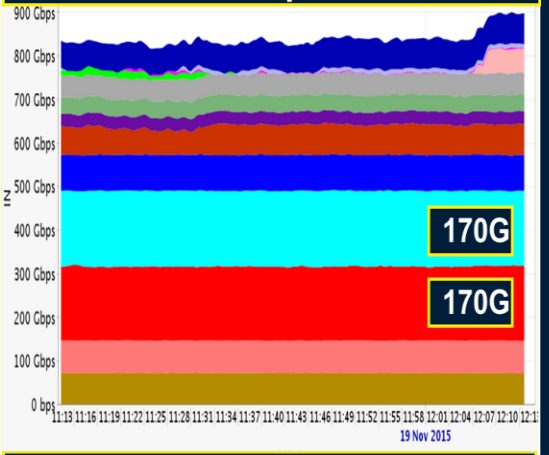
SC16+: Consistent Operations with Agile Feedback Major Science Flow Classes Up to High Water Marks

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

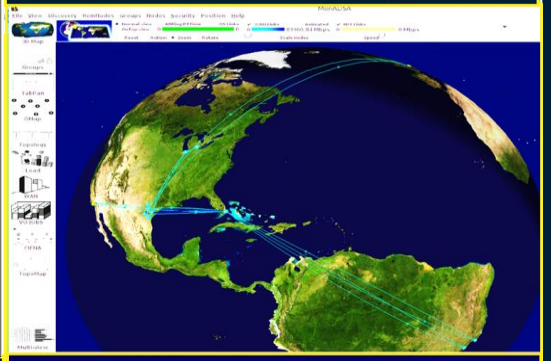


LHC at SC15: Asynchronous Stageout (ASO) with Caltech's SDN Controller

900 Gbps Total
Peak of 360 Gbps in the WAN



Global Topology



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

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

Worldwide Partnership at SC23 and Beyond

GNA-G
Global Network Advancement Group

SC23
Denver, CO | i am hpc.

Global Petascale to Exascale Workflows for Data Intensive Sciences

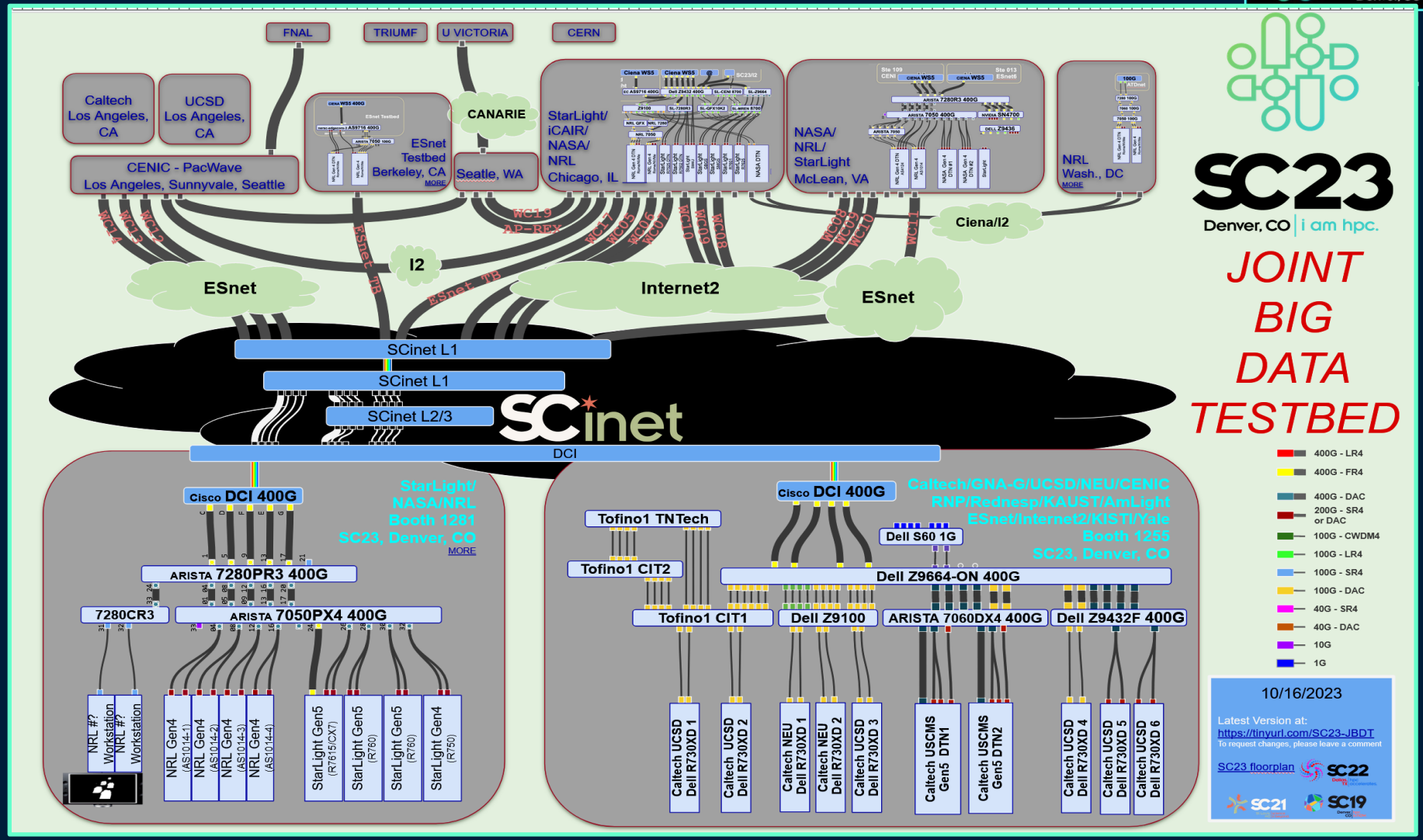
CALTECH.EDU HIGH ENERGY PHYSICS + TERABIT/SEC NETWORKS – THE FRONTIERS OF EXASCALE SCIENCE – SUPERCOMPUTING

CALIFORNIA INSTITUTE OF TECHNOLOGY

1891

Accelerated by Next Generation Programmable SDN Architectures and Machine Learning Applications

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 Caltech Campus; 4 X 400G to Caltech Booth with CENIC, Ciena, Internet2, ESnet, StarLight, US CMS and Network Partners

Technology Push: Actual Requirements will be a "hybrid" between preconceived notions and then-current capabilities .

Example: One 2023 Tbps Tier2 DTN + FDT

DTN: ASUS RS520A-E12-RS12U

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

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

NIC Setup at SC23 (x2)

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

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

To ~3 Tbps in a single rack

CENIC, ESnet and Internet2 at the LA PoP

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

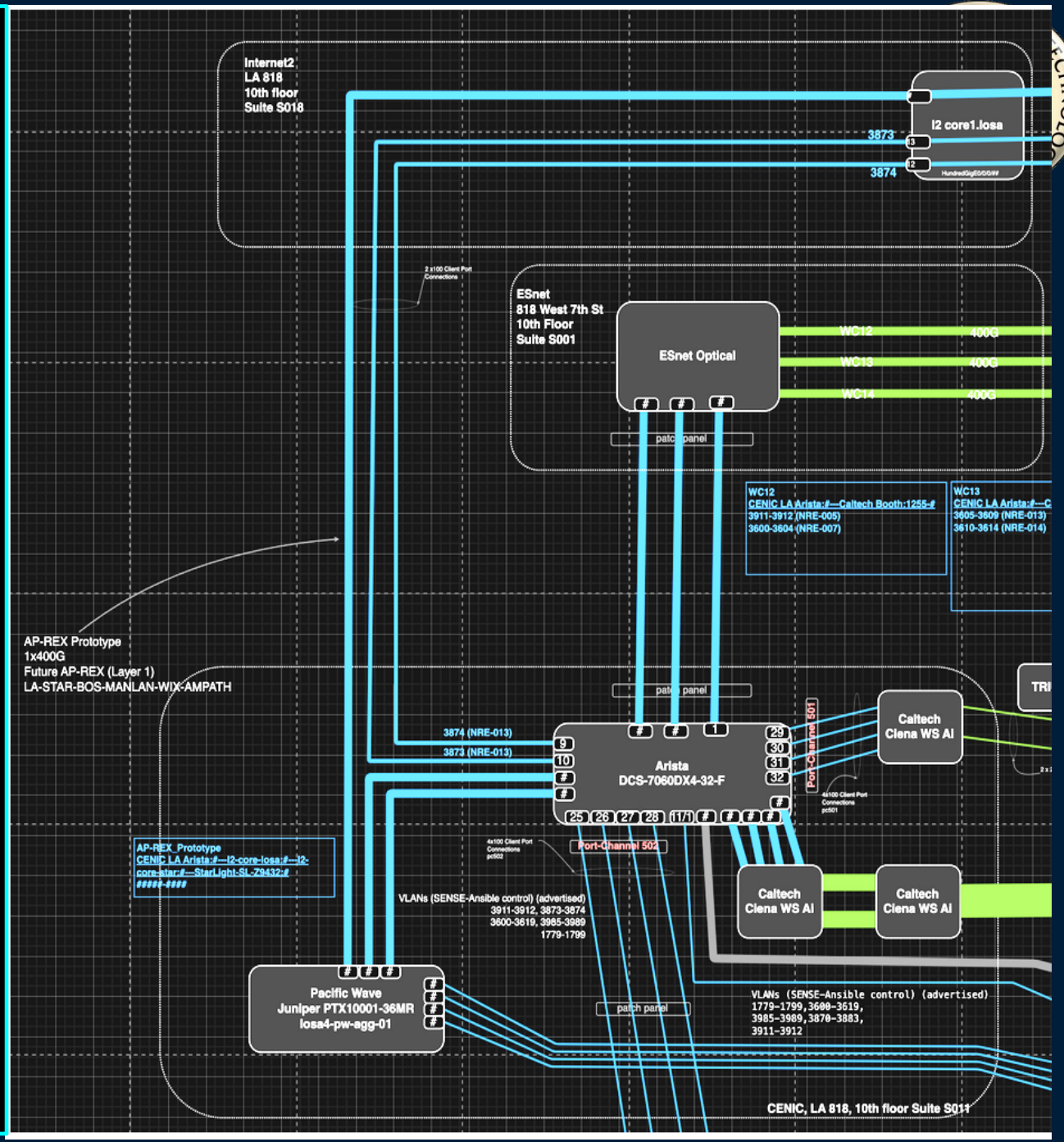
3 X 400G LA-Denver via ESnet

4 x 100G to UCSD/SDSC

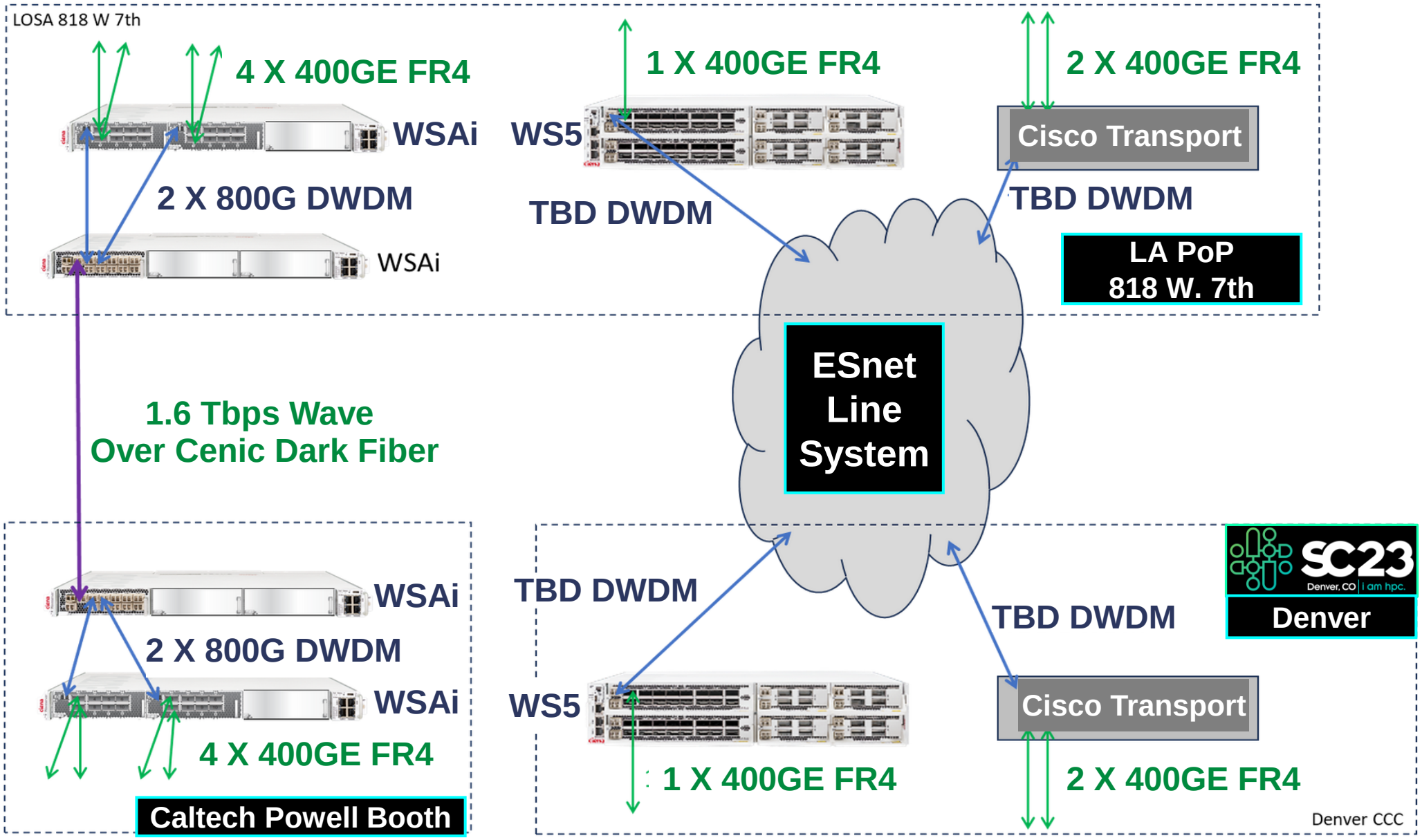
2 X 400G to Pacific Wave via CENIC

Permanent:

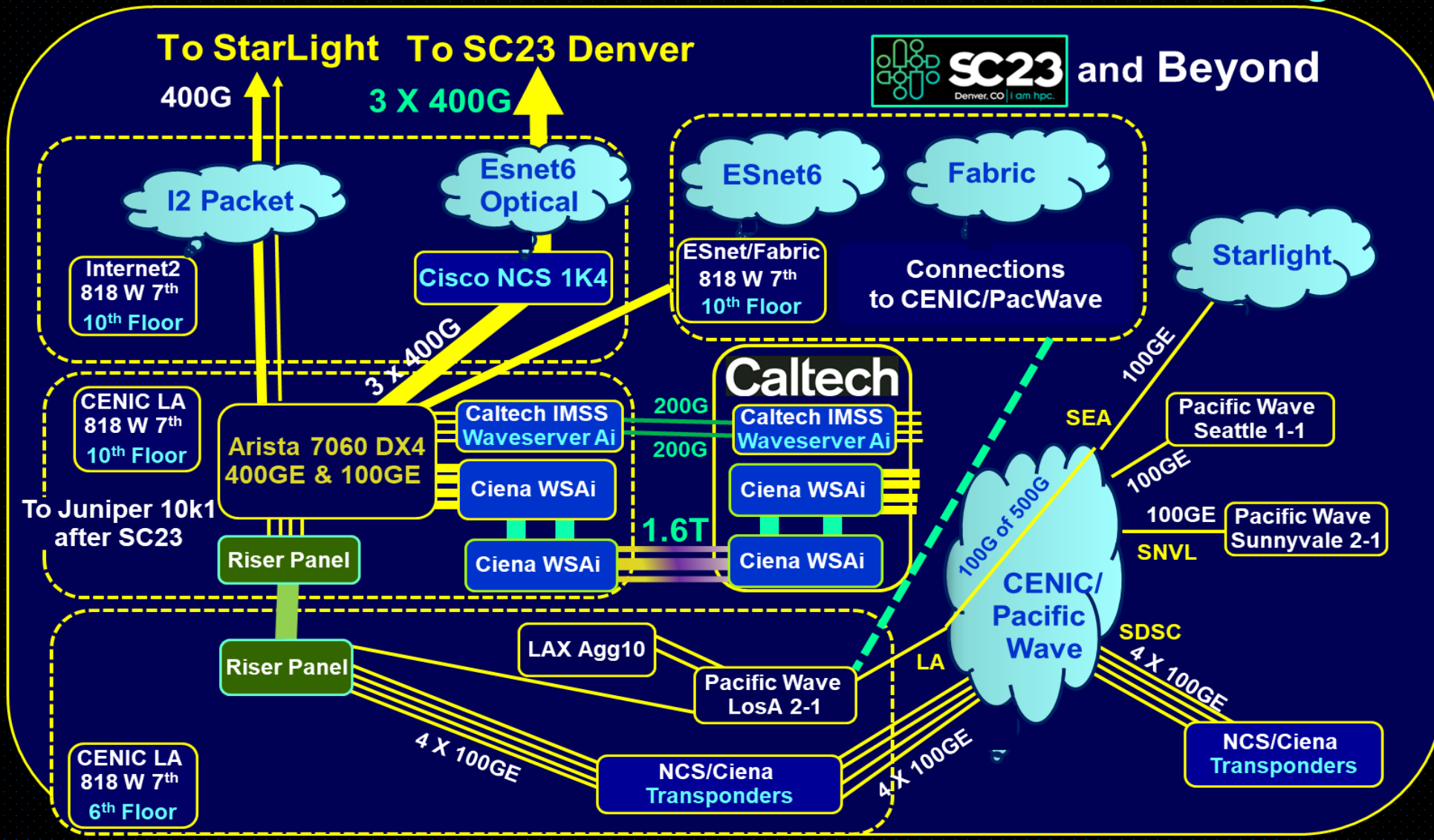
400G NA-REX Prototype
400G to ESnet Production



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

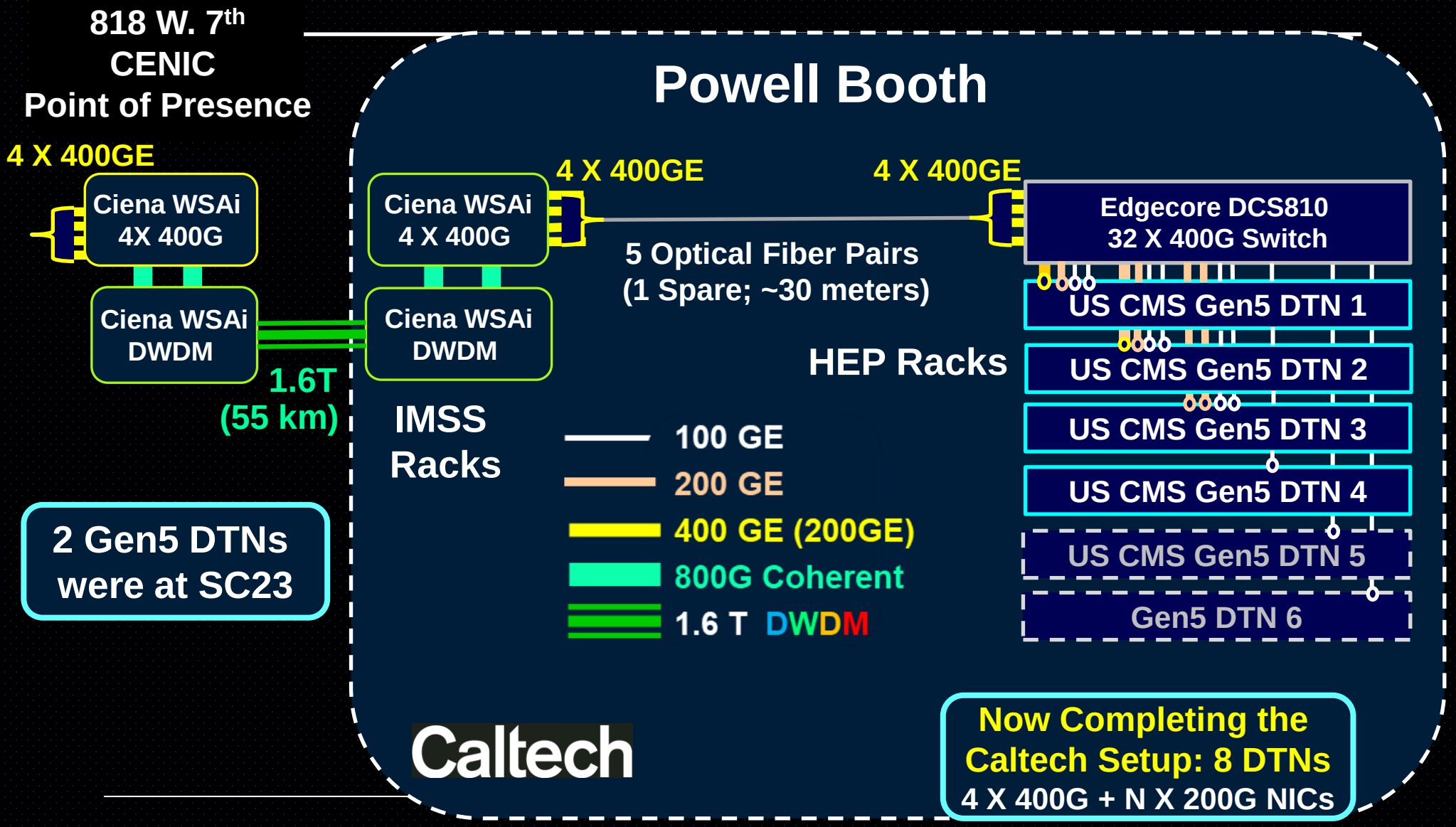


A New Generation Persistent 400G Super-DMZ: Ciena, Arista, CENIC, Pacific Wave, ESnet, Internet2, Caltech, UCSD, StarLight++



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

Simplified Caltech – LA Layout for SC23



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

Fast Start: 1+ Tbps with FDT within ~1 hour:
From One Pair of Gen5 Servers at Caltech To One Pair at SC23

13/11/2023, 12:24 Traffic Sentinel

FDT 11/13/23

inMon Traffic Sentinel

File | Home | Events | Traffic | Hosts | Services | Reports | Maps | Search | Help

Status | Interfaces | Trend | Factors | Circles | Top N

Filter:

SC > Denver > SCinet > All Show Map

Chart Top Sources To Host All Protocol All Date 13 Nov 2023 Time Now Interval 20 minutes

Units Bits/sec. Where (?) OK Edit Clear

Raimondas Sirvinskas

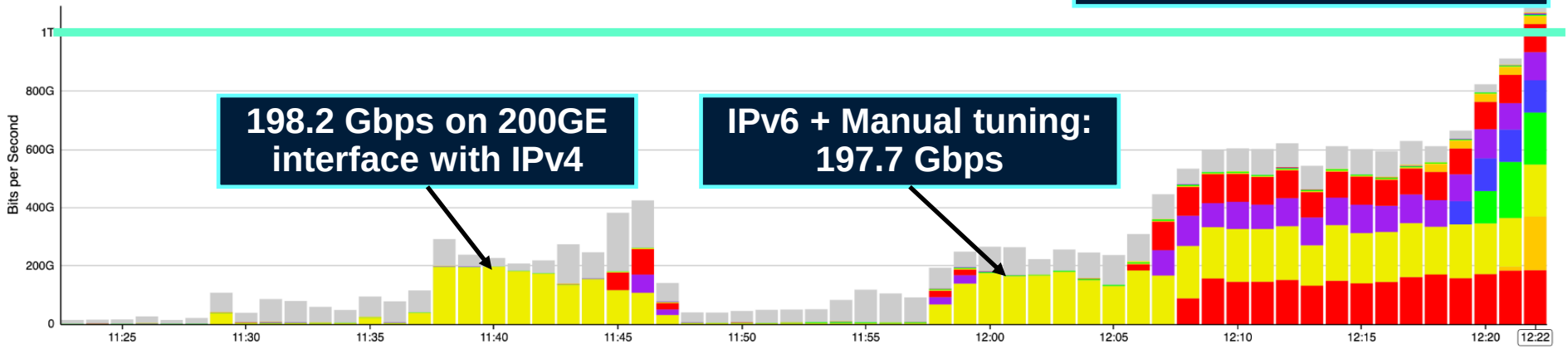
Marcos Schwarz

	Source Address	Value
	192.168.7.14	185.11G
	192.168.17.13	184.24G
	192.168.6.14	179.02G
	192.168.16.13	178.41G
	wlan-ipv6-only-4621.23 (192.168.82.13)	110.95G
	wlan-ipv6-only-4622.23 (192.168.82.14)	96.52G
	wlan-ipv6-only-4366.23 (192.168.81.14)	96.26G
	FD77:FB3D:88F5:43::20	26.12G
	FD77:FB3D:88F5:42::20	3.16G
	cluster-challenge-management-20.23 (140.221.235.244)	3.12G
	st-096-hh151458.cern.ch (2001:1458:301:10::100:81)	2.13G
	evpn-loopback-2-rtr.23 (192.168.1.1)	1.67G
	ps-dnoc-4 (140.221.235.26)	1.61G
	ps-conf-1 (2001:468:1F07:CF00::2)	1.59G
	evpn-loopback-1-07.23 (192.168.0.7)	1.47G

Caltech Booth 1255
NRE-13: 1+ Tbps

Still Tuning

+ More servers available



Copyright © 1999-2023 InMon Corp. ALL RIGHTS RESERVED

NRE-13 Top Sources: To 1.4 Tbps

FDT 11/13/23

inMon Traffic Sentinel

File | Home | Events | Traffic | Hosts | Services | Reports | Maps | Search | Help

Status | Interfaces | Trend | Factors | Circles | Top N

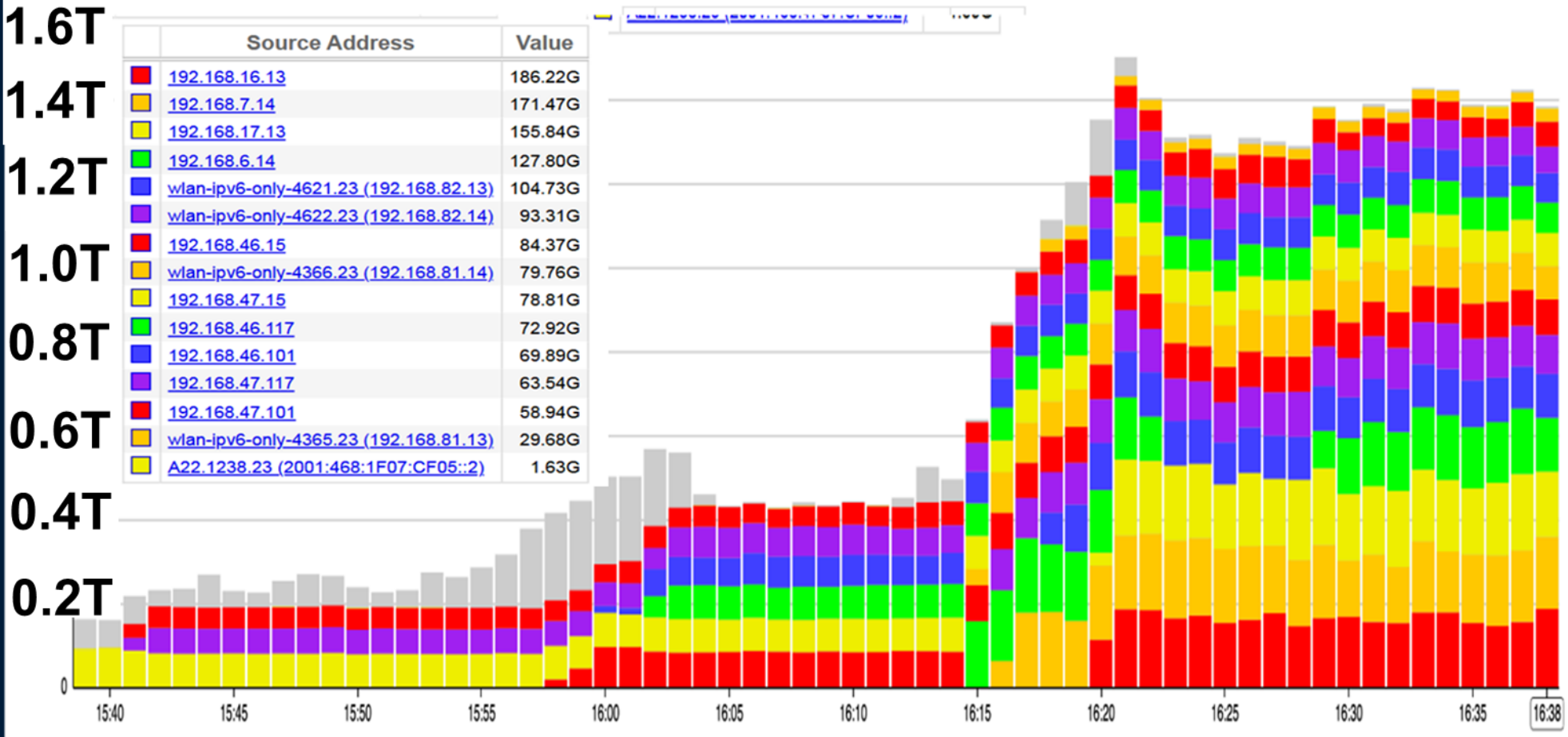
Filter:

SC > Denver > SCinet > All [Show Map]

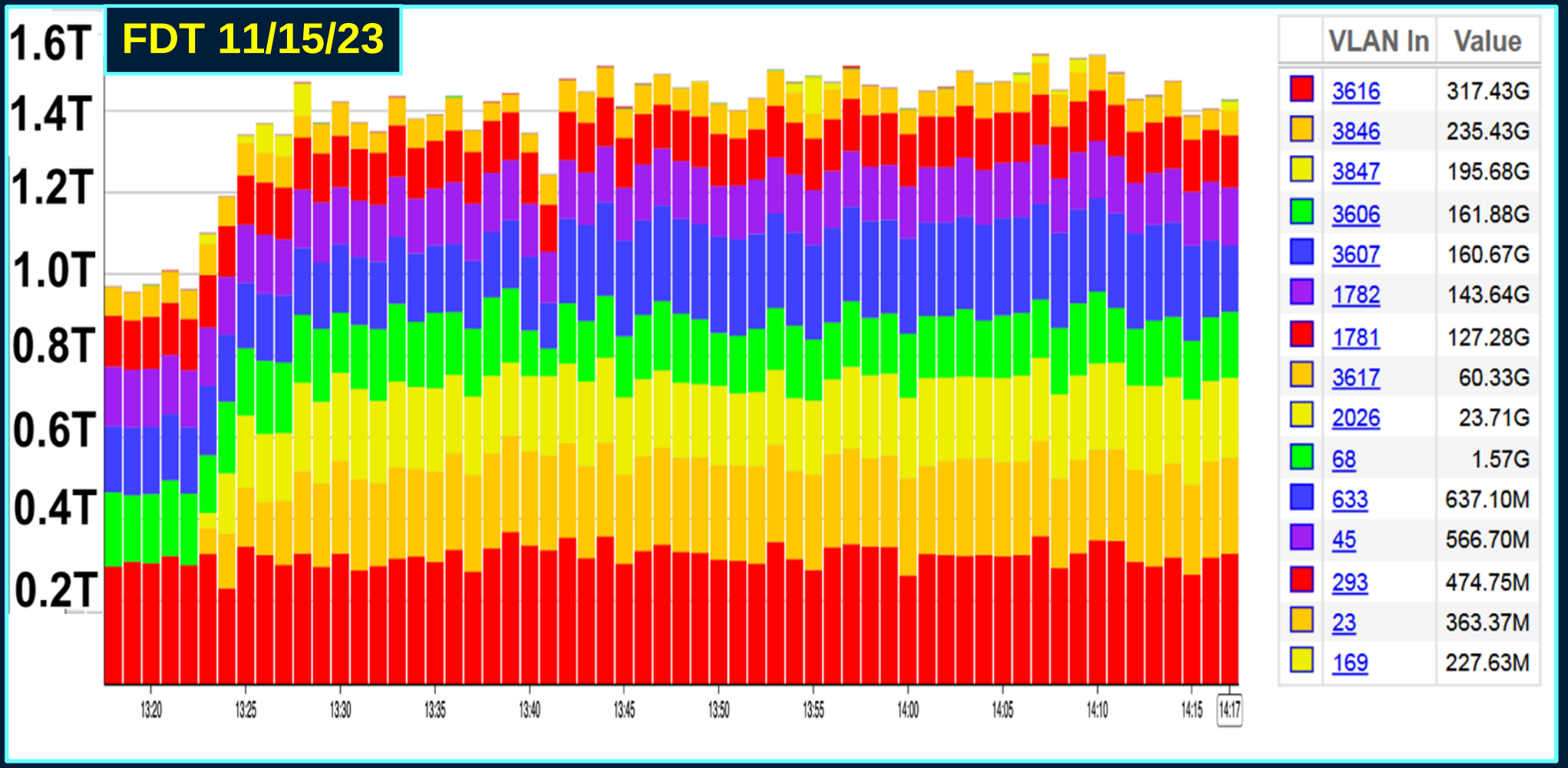
Chart: Top Sources | To Host: All | Protocol: All | Date: 13 Nov 2023 | Time: Now | Interval: 5 minutes

Units: Bits/sec. | Where (?): vlansource!='64' & vlansource!='46' & vlansource!='61' & vlansource!='63' & | OK | Edit | Clear

	Source Address	Value
	192.168.16.13	186.22G
	192.168.7.14	171.47G
	192.168.17.13	155.84G
	192.168.6.14	127.80G
	wlan-ipv6-only-4621.23 (192.168.82.13)	104.73G
	wlan-ipv6-only-4622.23 (192.168.82.14)	93.31G
	192.168.46.15	84.37G
	wlan-ipv6-only-4366.23 (192.168.81.14)	79.76G
	192.168.47.15	78.81G
	192.168.46.117	72.92G
	192.168.46.101	69.89G
	192.168.47.117	63.54G
	192.168.47.101	58.94G
	wlan-ipv6-only-4365.23 (192.168.81.13)	29.68G
	A22.1238.23 (2001:468:1F07:CF05::2)	1.63G



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

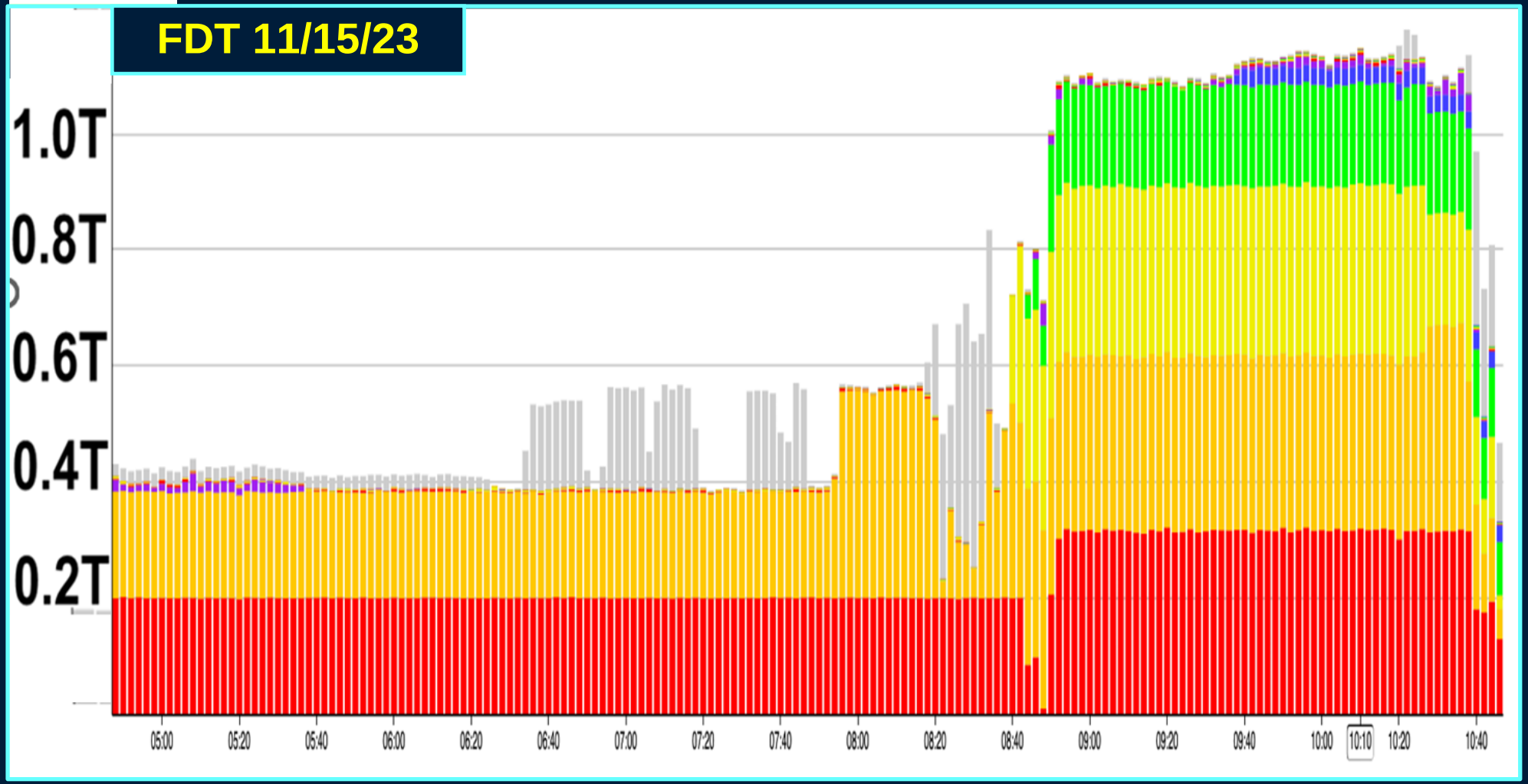


	VLAN In	Value
	3616	317.43G
	3846	235.43G
	3847	195.68G
	3606	161.88G
	3607	160.67G
	1782	143.64G
	1781	127.28G
	3617	60.33G
	2026	23.71G
	68	1.57G
	633	637.10M
	45	566.70M
	293	474.75M
	23	363.37M
	169	227.63M

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

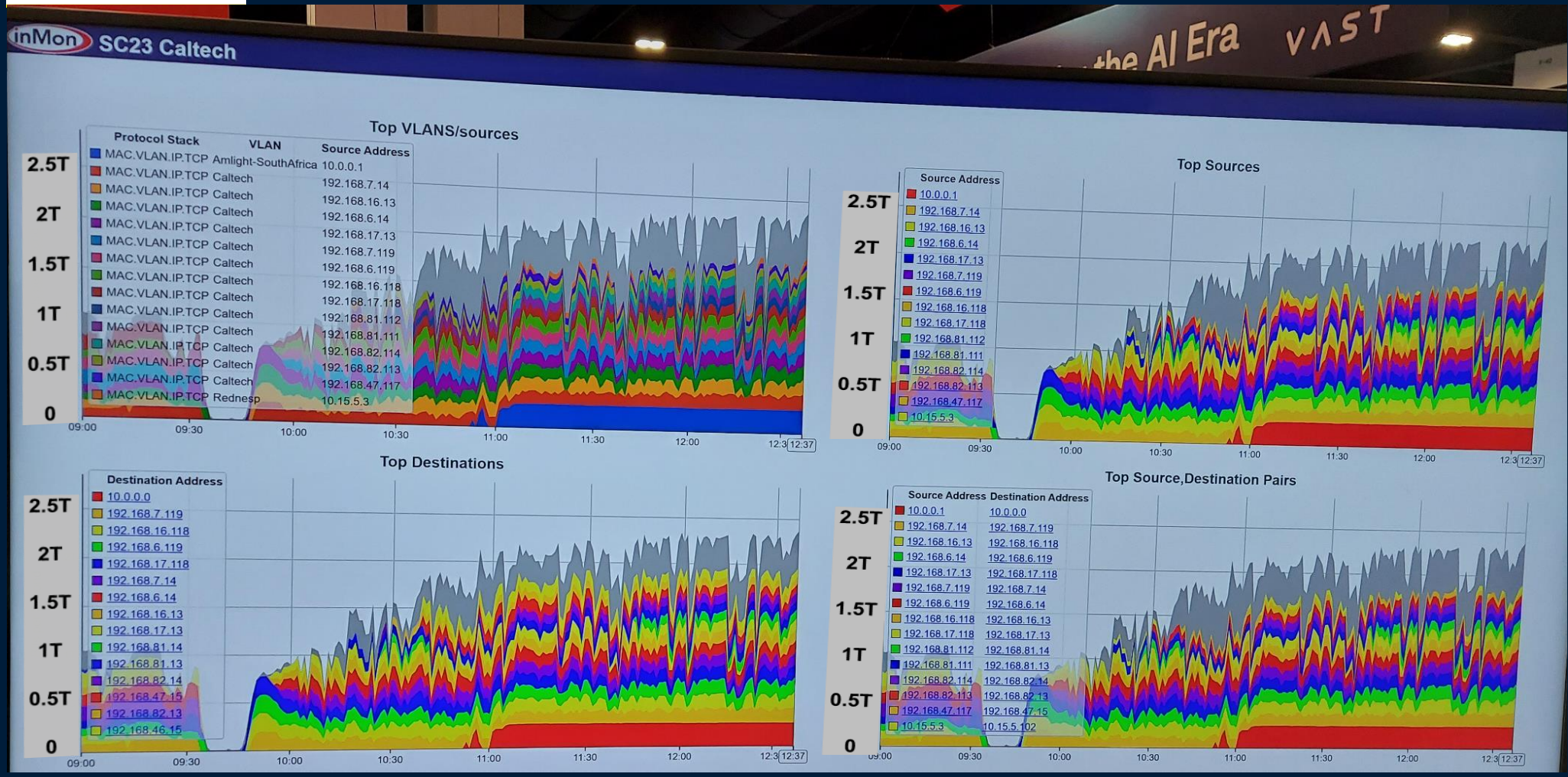
NRE-13: 1.1 Tbps on 2 X 400G Circuits

Stabilized with Dynamic Thread Management



With Just 2 Gen5 Servers at SC23 and 2 at Caltech

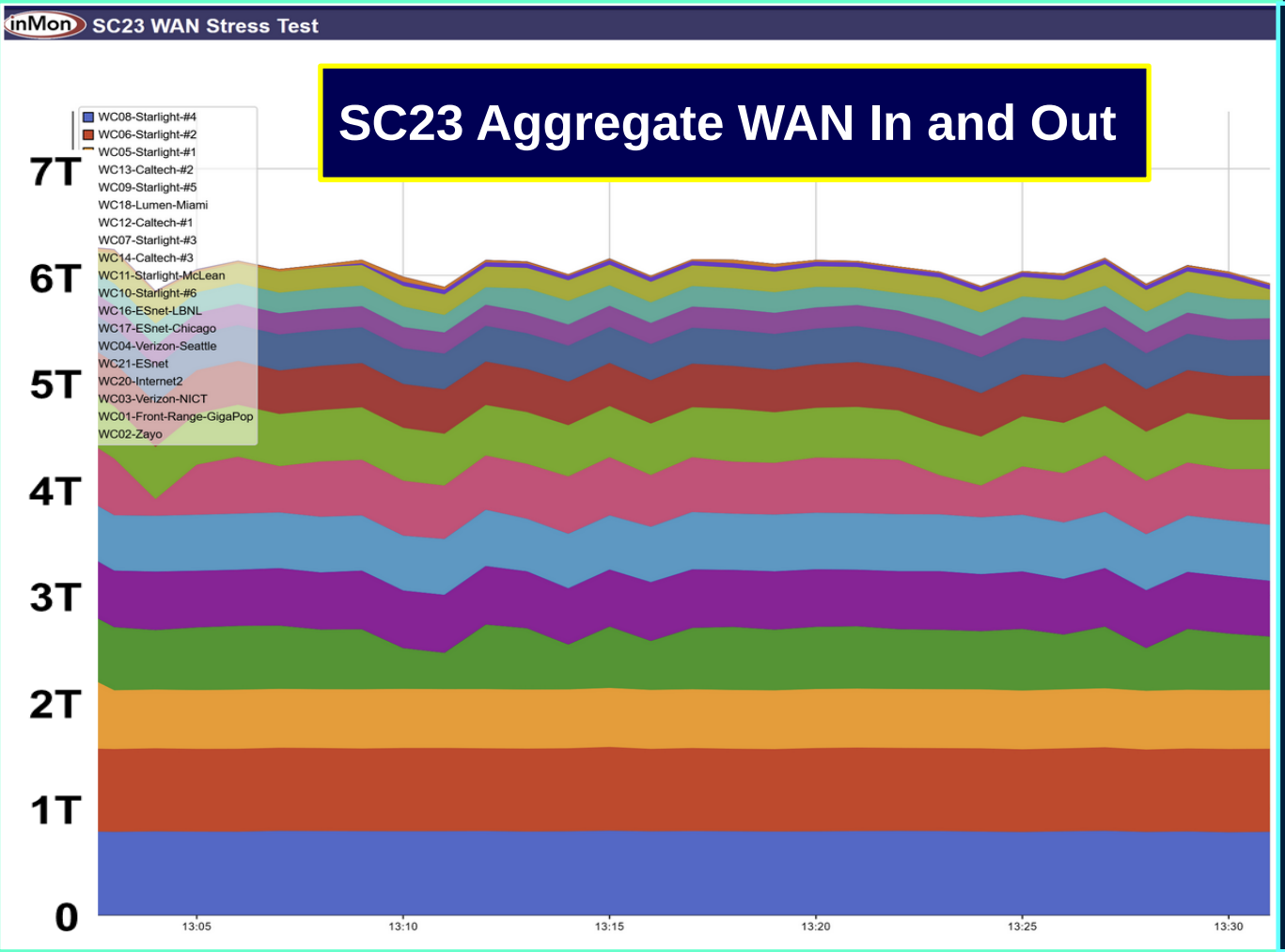
SC23 Stress Test 11/16/23
Caltech Results: Up to 2.4 Tbps



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 [Hit Target]



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 or 2025
- **SENSE 400G paths:** ESnet production, NA-REX via StarLight; Links to CERN

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

Caltech

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 then was 6.5.10 [Now 6.8.x]
- 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

Added Explanations: Rules for Higher Throughput

- Use large packets ➔ MTU 9000 (as above)
- Do not allow CPU to enter power save mode: set to stay at high frequency all the time
- Adaptive RX is an algorithm to improve rx latency at low packet-receiving rates and to improve throughput at high packet-receiving rates.
  Some NIC drivers do not support this feature.
- Changing txqueuelen allows you to set the length of the data queue for network interfaces. When the queue reaches the specified value of txqueuelen, then the data is transmitted.
- Interface has the capability of using transmit and receive buffer description ring into the main memory. They use direct memory access (DMA) to transfer packets from the main memory to carry packets independently from the CPU.
- BIG TCP was implemented mainly for IPV6, however it has some IPv4 implementation too as it gives benefits for IPv4 also. Setting BIG TCP for the interface and testing single thread transfers we got a nice improvement:

  sudo ip link set dev eth1 gro_max_size 185000 gso_max_size 185000

  - The single thead transfer performance increased from 65Gbps to 74Gbps on a local transfer using a 200GE link [M. Schwarz working toward 100G]

Explanation: About Big TCP and Higher Throughput

- Enabling Big TCP across both send/receive systems can lead to huge throughput improvements and lower latencies for high speed networking environments.
- The idea behind Big TCP is to use a header in the packet that is bigger than 16 bits, which is the maximum value that can be specified in an IP header.
- Big TCP is a technology that allows going past the current 64KB TSO/GRO packet limit size for IPv6 traffic by way of the IPv6 Jumbogram extension header.
- Big TCP makes the packets processed in the kernel bigger, so that the overall number of packets to be processed will be further reduced to improve performance.

Kernel Parameters

Caltech

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
- 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

Caltech

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 (4): Comparing Congestion Control Algorithms

Caltech

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

Cubic versus bbr:

Cubic

```
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: 197.705 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: 197.928 Gb/s   Avg: 197.884 Gb/s
```

BBR

```
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: 163.300 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
```

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

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**
- **There is an increasing gulf between current capabilities and the requirements as pre-conceived in 2020-23.** Actual requirements will be in the middle, also exploiting then-current technology.
- **We now have three global testbeds with expanding capabilities, to close the gap.**
  - **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 3 and other smart edge devices, the Container-Lab based digital twin, ESnet High Touch, NOTED among Tier1s, PolKA and SRv6+MicroSIDS, and 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.
- **With Mariam Kiran (now at ORNL) we are resuming the effort on using machine learning/AI**
  to optimize network operations: tactically; and with the emerging system-level picture – strategically
- **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 after SC23: Including**
  - **400G link to the ESnet production network in LA which is useful for DC24 and beyond**
  - **400G link between the CENIC Juniper &nd StarLight with 2 X 400G to the SENSE-controlled Arista in LA**
  - **The additional fiber pair between the Caltech campus and LA with CENIC,**
    which has multiple uses, and currently supports 4 X 400G links connecting to 4 X 400G + N X 200G DTNs

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

Extra Slides Follow

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)

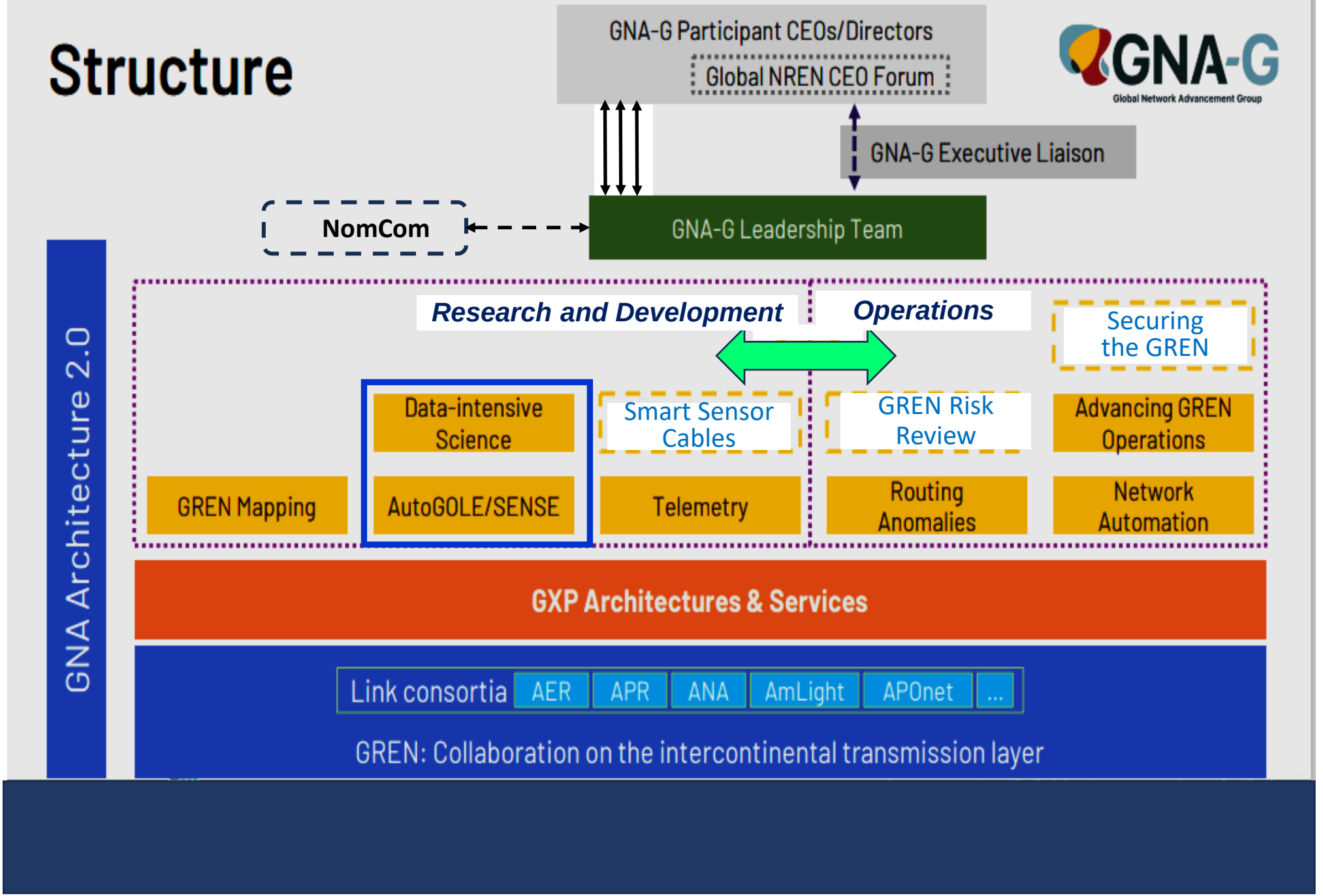
David Wilde,
Chair
Aarnet (Australia)

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

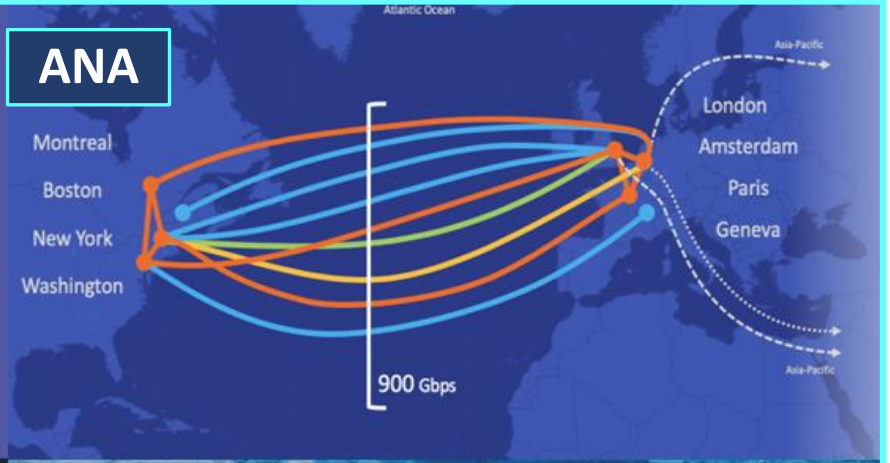
Structure

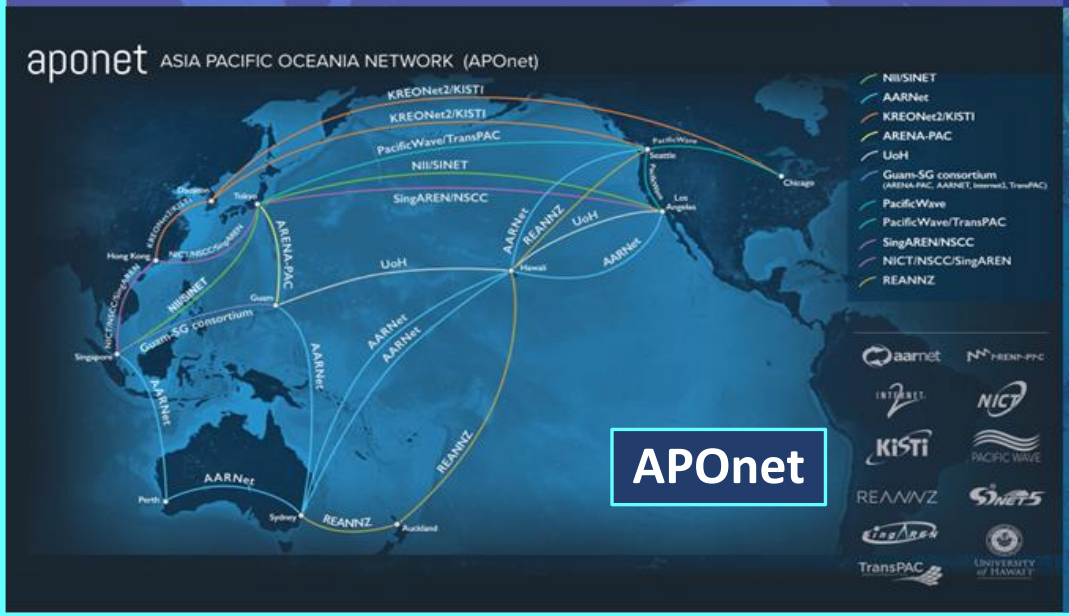

GNA-G Participant CEOs/Directors
Global NREN CEO Forum
GNA-G
Global Network Advancement Group
GNA-G Executive Liaison
NomCom
GNA-G Leadership Team
GNA Architecture 2.0
Research and Development
Operations
Securing the GREN
Data-intensive Science
Smart Sensor Cables
GREN Risk Review
Advancing GREN Operations
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

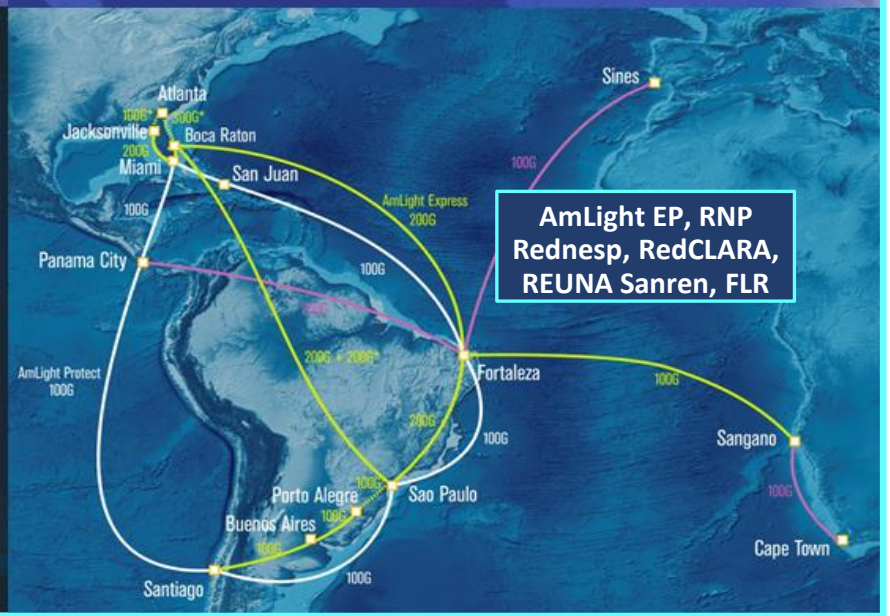
GNA-G
Global Network Advancement Group

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

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







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

NA-REX: North America Research and Education Exchange

October 2023

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

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

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 Denver, CO | i am hpc.

GNA-G Global Network Advancement Group

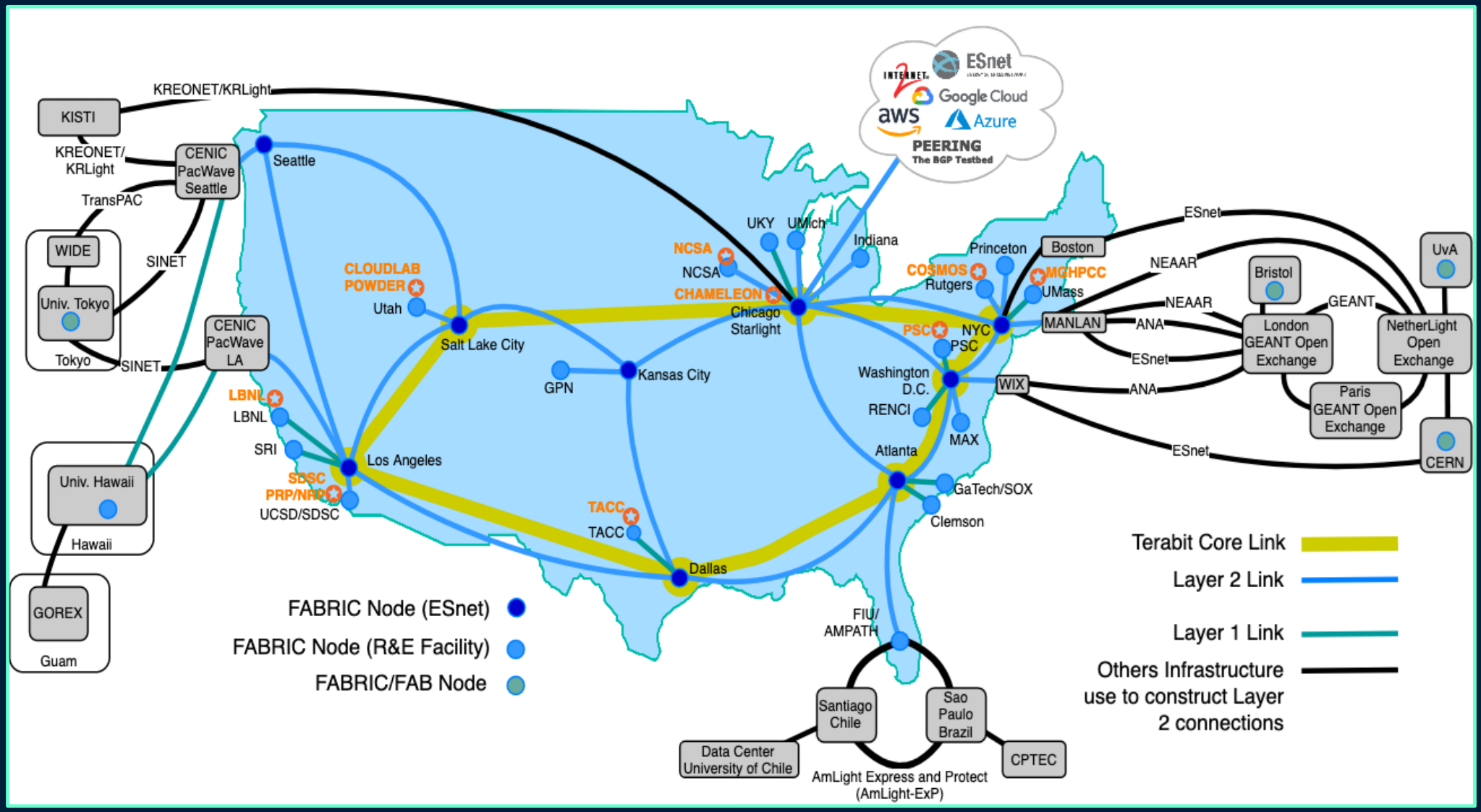
Next Generation System for Data Intensive Sciences

- 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

AutoGOLE / SENSE Working Group

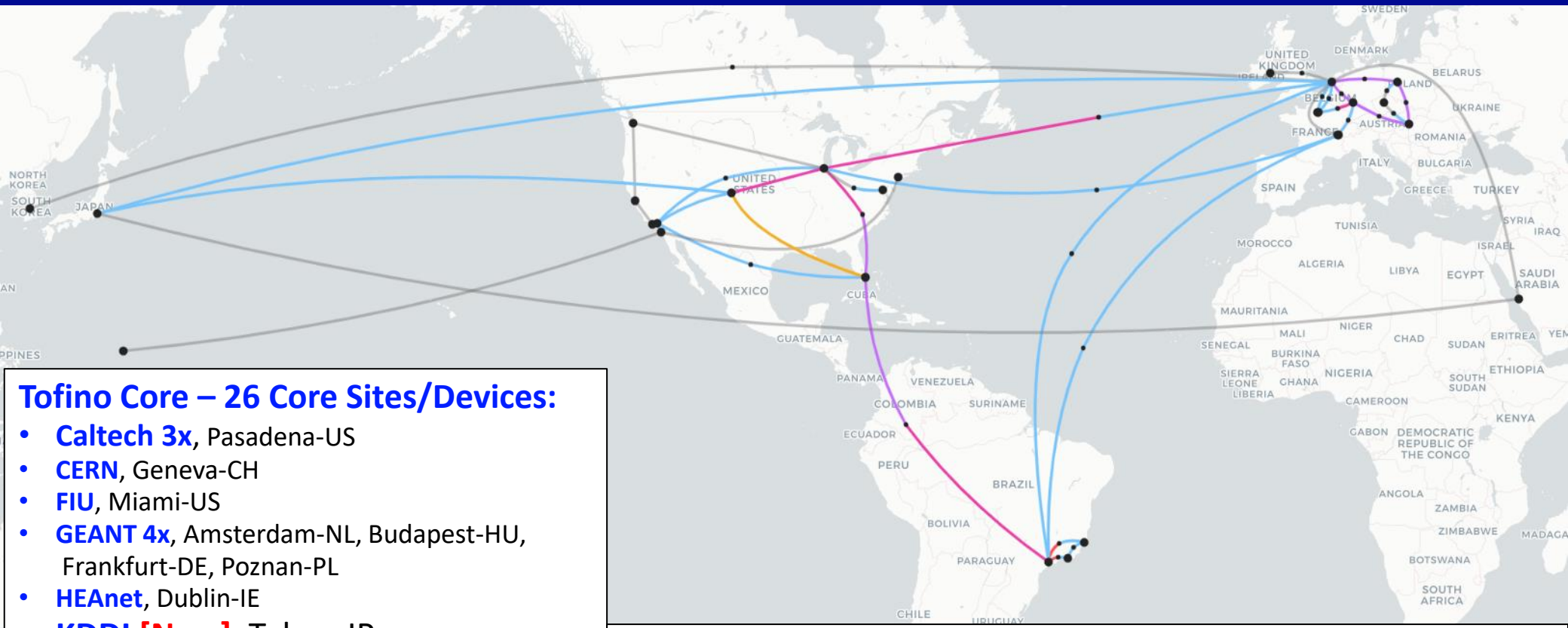
- **Worldwide collaboration of open exchange points and R&E networks interconnected to deliver network services end-to-end in a fully automated way.** NSI/Supa for network connections, SENSE for integration of End Systems and Domain Science Workflow facing APIs.
- **Key Objective:**
  - The AutoGOLE 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

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



US, Europe, Asia Pacific and Latin America

Global P4 Lab (GP4L)



Tofino Core – 26 Core Sites/Devices:

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

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

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

x86/DPDK Islands – 4 Sites/Devices:

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

Global P4 Lab (GP4L)
Status and Ongoing Progress

SC23 Denver, CO | i am hpc.

Talk by
Marcos Schwarz RNP

Achieved by SC23

- Persistent global L3 overlay network based on P4 switches
- Core network based on RARE/FreeRtr and edge networks based also on SONiC
- Management infrastructure and tools used to operate this global network
- Support for new device types: Bluefield2 Smart NIC Islands (using emulated P4)
- Intercontinental high capacity transfers (100G and up) exploring multiple source routing solutions and next generation protocols (e.g. PolKA)
- Creation of an on-demand digital twin of the GP4L and production networks
- Interconnections with other testbeds (FABRIC)

In Progress

- Automated generation of a real time world map + dashboards of GP4L
- Capability to support multiple virtual networks that implement different choices of routing stacks on the same devices: traditional and SDN based
- Integration with initiatives for visibility, controllability and intelligence
- ***Working as a reference state of the art / next generation R&E network***

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, ...

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

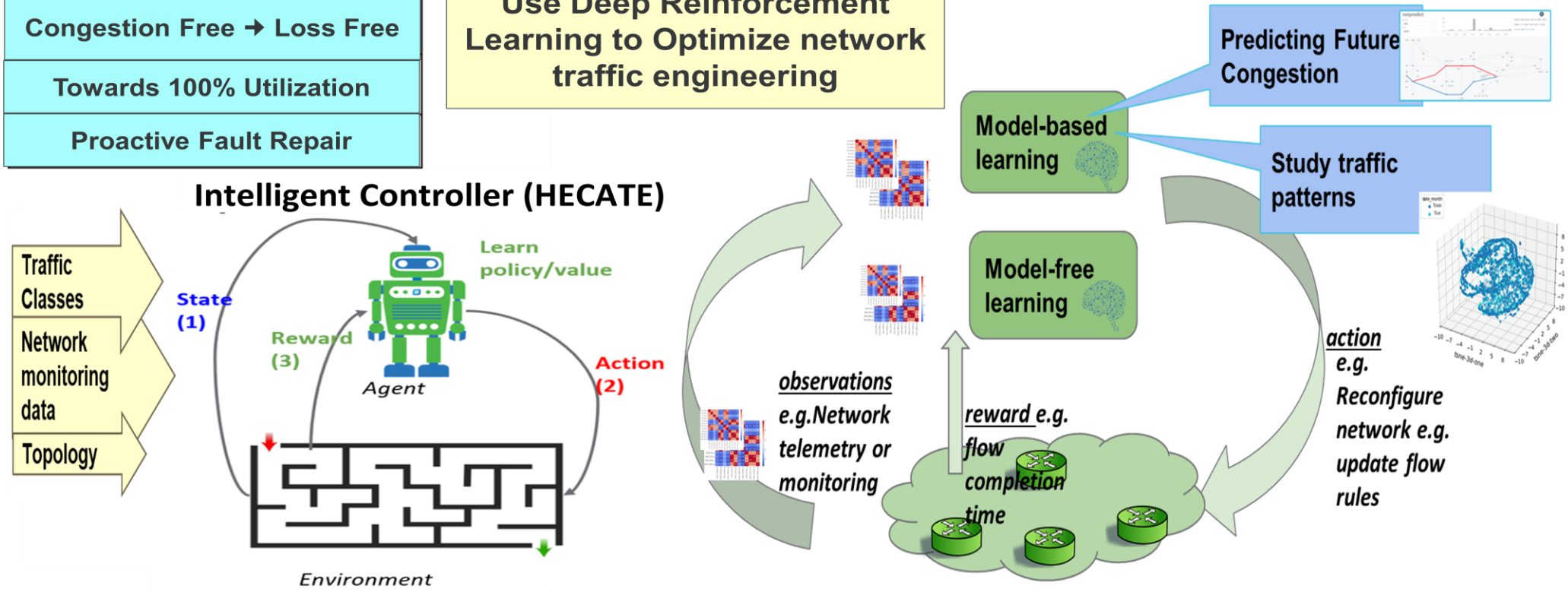
Mariam Kiran (ESnet) et al. Intelligent Networks DOE Project

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



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**

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*

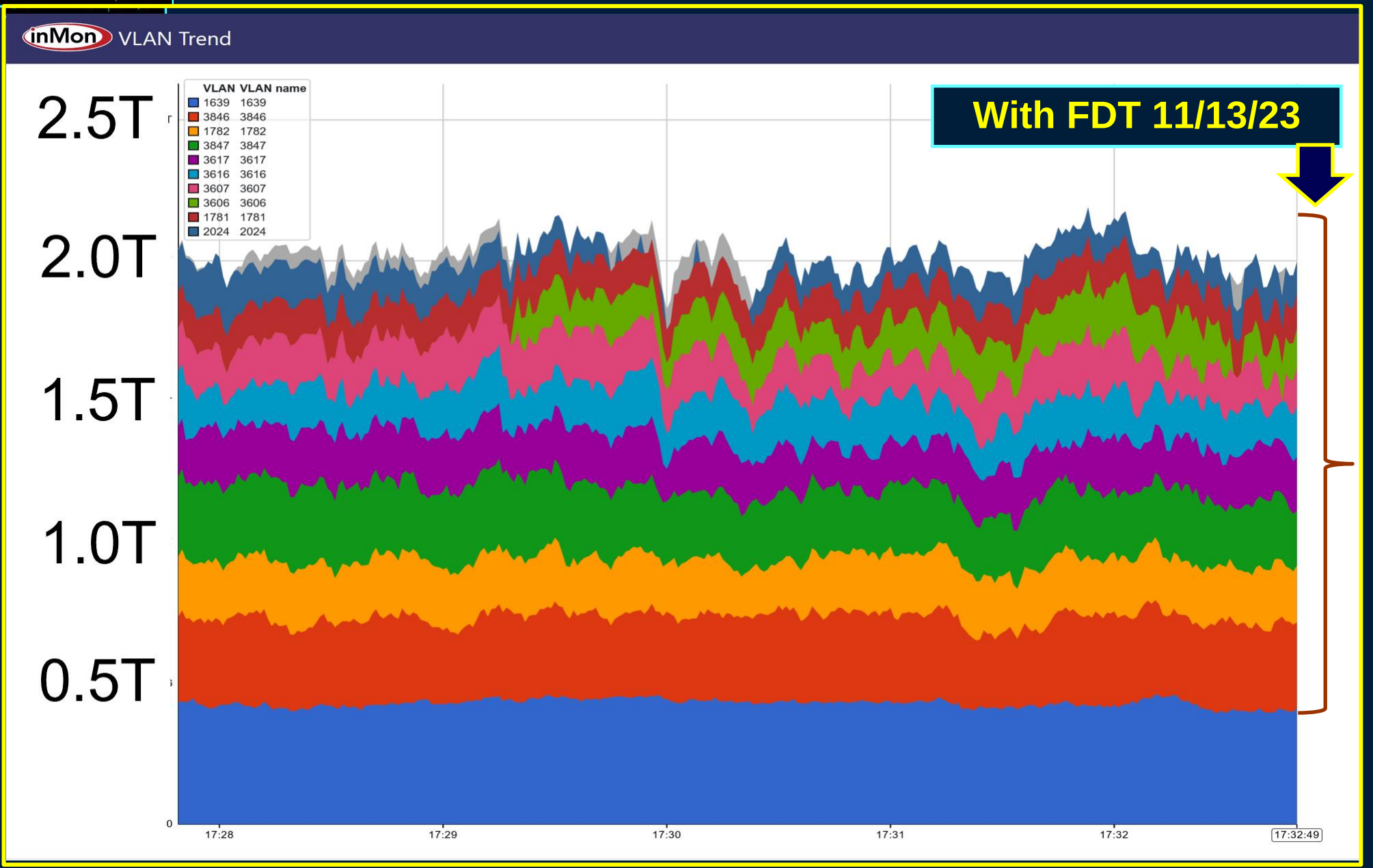
SC23 Denver, CO | i am hpc.

GNA-G Global Network Advancement Group

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**

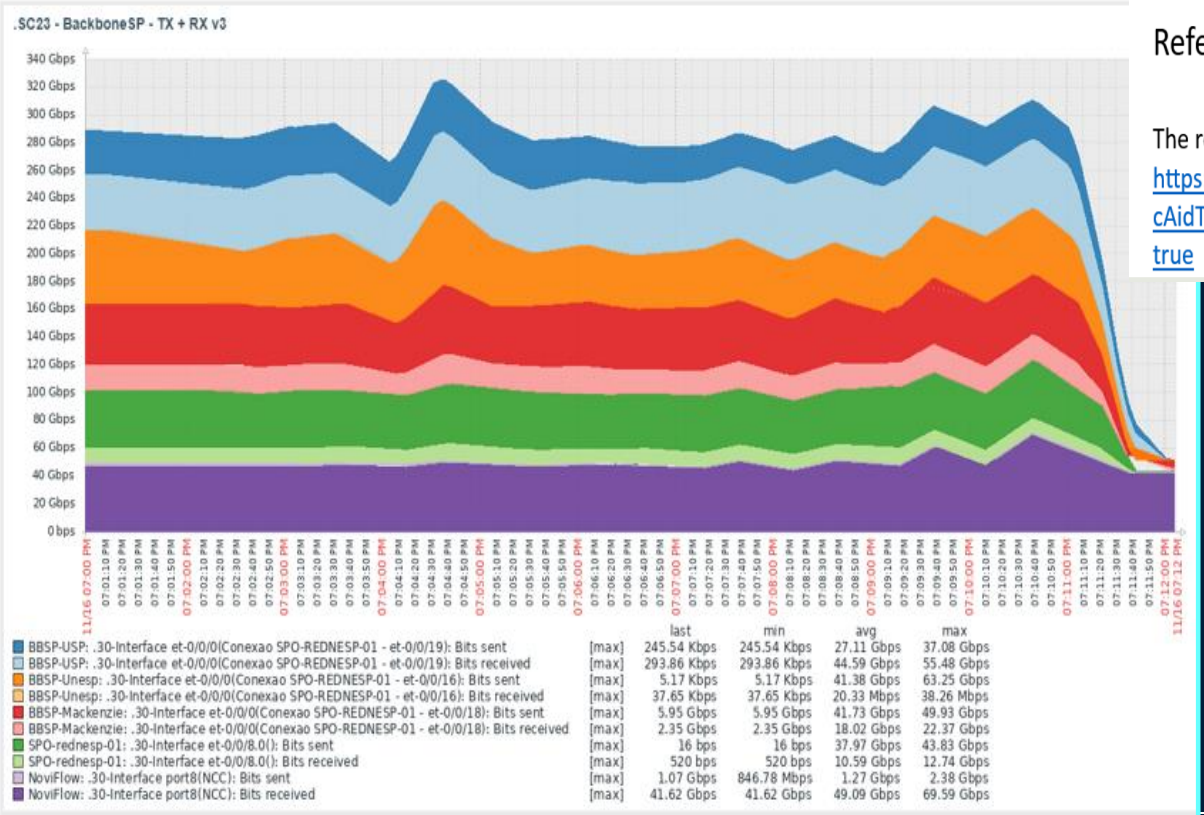
NRE-13 VLANs: To 1.4 Tbps of 2 Tbps



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.



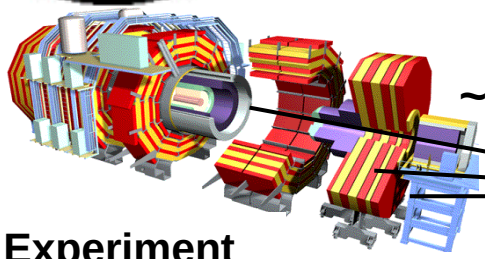
		last	min	avg	max
BBSP-USP: .30-Interface et-0/0/0(Conexao SPO-REDNESP-01 - et-0/0/19): Bits sent	[max]	245.54 Kbps	245.54 Kbps	27.11 Gbps	37.08 Gbps
BBSP-USP: .30-Interface et-0/0/0(Conexao SPO-REDNESP-01 - et-0/0/19): Bits received	[max]	293.86 Kbps	293.86 Kbps	44.59 Gbps	55.48 Gbps
BBSP-Unesp: .30-Interface et-0/0/0(Conexao SPO-REDNESP-01 - et-0/0/16): Bits sent	[max]	5.17 Kbps	5.17 Kbps	41.38 Gbps	63.25 Gbps
BBSP-Unesp: .30-Interface et-0/0/0(Conexao SPO-REDNESP-01 - et-0/0/16): Bits received	[max]	37.65 Kbps	37.65 Kbps	20.33 Mbps	38.26 Mbps
BBSP-Mackenzie: .30-Interface et-0/0/0(Conexao SPO-REDNESP-01 - et-0/0/18): Bits sent	[max]	5.95 Gbps	5.95 Gbps	41.73 Gbps	49.93 Gbps
BBSP-Mackenzie: .30-Interface et-0/0/0(Conexao SPO-REDNESP-01 - et-0/0/18): Bits received	[max]	2.35 Gbps	2.35 Gbps	18.02 Gbps	22.37 Gbps
SPO-rednesp-01: .30-Interface et-0/0/8.0(): Bits sent	[max]	16 bps	16 bps	37.97 Gbps	43.83 Gbps
SPO-rednesp-01: .30-Interface et-0/0/8.0(): Bits received	[max]	520 bps	520 bps	10.59 Gbps	12.74 Gbps
NoviFlow: .30-Interface port8(NCC): Bits sent	[max]	1.07 Gbps	846.78 Mbps	1.27 Gbps	2.38 Gbps
NoviFlow: .30-Interface port8(NCC): Bits received	[max]	41.62 Gbps	41.62 Gbps	49.09 Gbps	69.59 Gbps

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

LHC Data Grid Hierarchy
(Invented by Caltech: from 1999)
A 2005 View



Experiment

~PByte/sec

Online System

~150-1500 MBytes/sec

CERN/Outside Resource Ratio ~1:2
Tier0/(Σ Tier1)/(Σ Tier2) ~1:1:1

Tier 0 +1

CERN Center
PBs of Disk;
Tape Robot

Tier 1

10 - 40 Gbps

IN2P3 Center

RAL Center

INFN Center

FNAL Center

~10 Gbps

Tier 2

Tier2 Center

Institute

Institute

Institute

Institute

Tier 3

~1-10 Gbps

Physics data cache

1 to 10 Gbps

Workstations

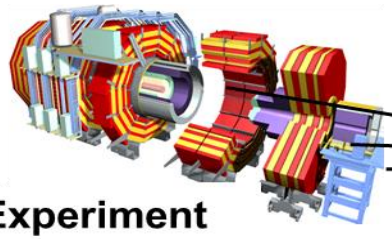
Tier 4

Tens of Petabytes by 2007-8.
An Exabyte ~5-7 Years later.

+1st Tier2 Center at Caltech
CACR 1999-2000

Emerging Vision: A Richly Structured, Global Dynamic System

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



Experiment

~PBytes/sec

Online System

~1000 MBytes/sec

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

Tier 0

CERN Center
100 PBs of Disk;
Tape Robot

Tier 1

N X100 Gbps

London | Paris | Bologna | Chicago | ...

10 to 100 Gbps

Tier 2

CACR | ? Center | nter | Center | Center

Tier 3

Physics data cache

Institute | tute | stitute | Institute

10 to 100 Gbps

Tier 4

Workstations

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

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

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

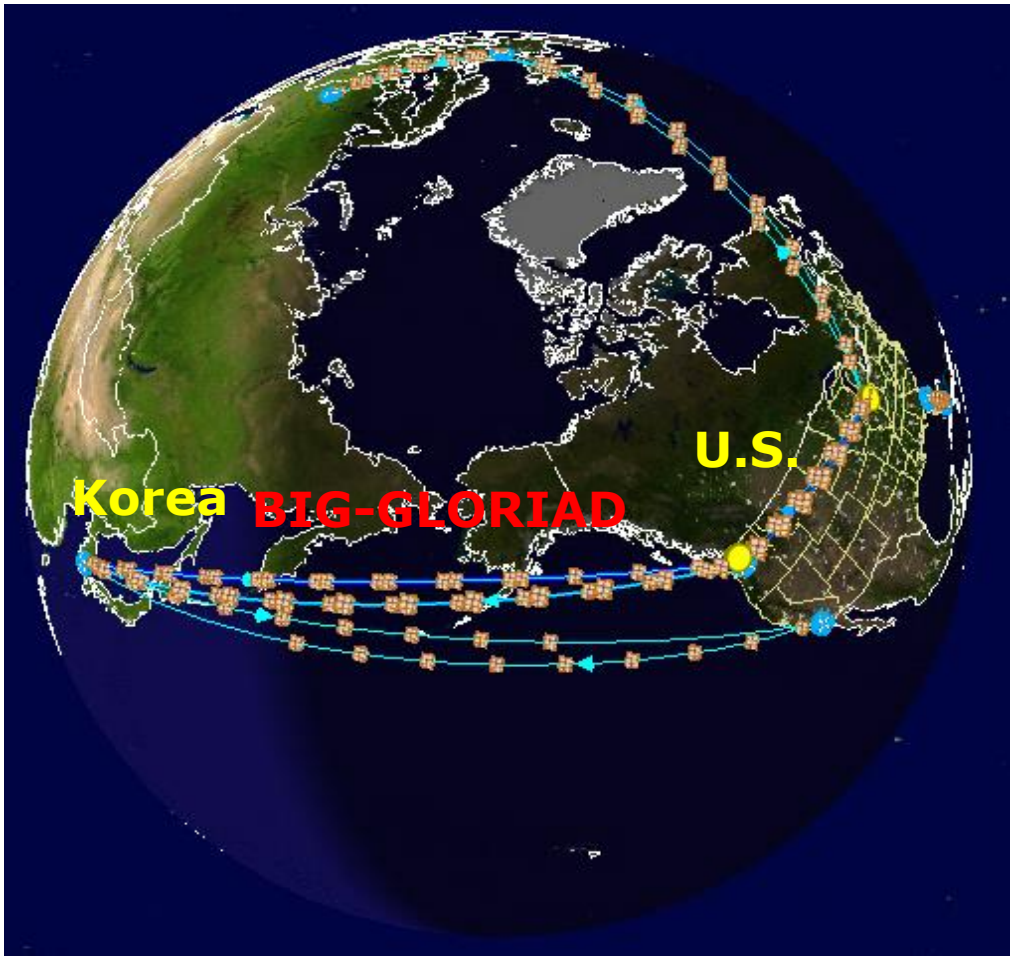
Acknowledgements

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.

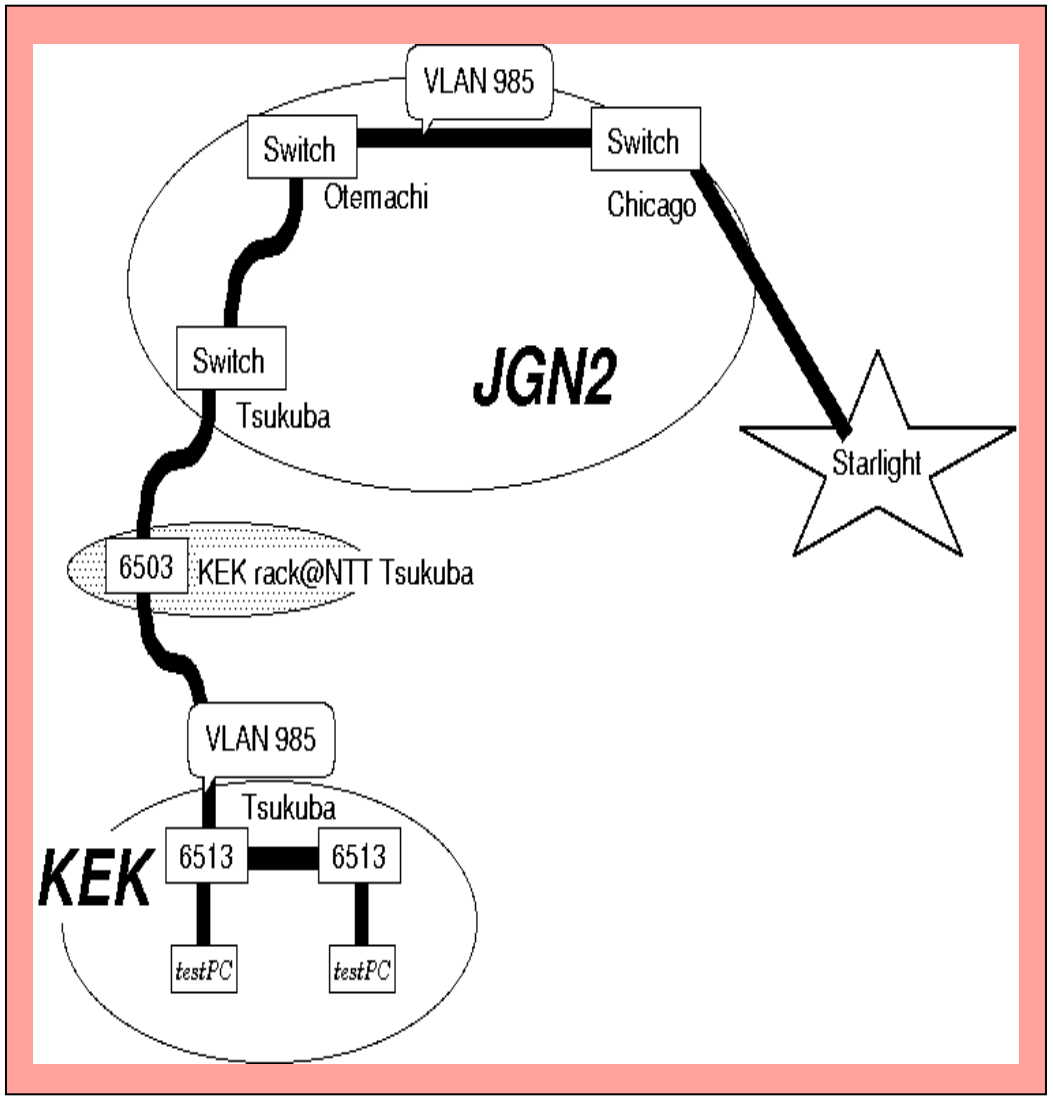
Finally, this work would not be possible without the significant contributions and the collaboration of the many HEP, network and computer and research teams partnering in the Global Network Advancement Group, in particular the GNA-G Data Intensive Sciences and AutoGOLE/SENSE Working Groups and the Global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

KNU (Korea) Main Goals

- Uses 10Gbps GLORIAD link from Korea to US, which is called BIG-GLORIAD, also part of UltraLight
- Try to saturate this BIG-GLORIAD link with servers and cluster storages connected with 10Gbps
- Korea is planning to be a Tier-1 site for LHC experiments



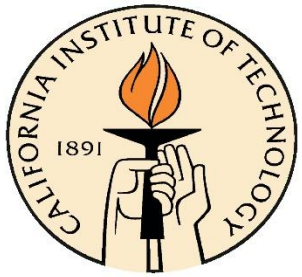
KEK (Japan) at SC05
10GE Switches on the
KEK-JGN2-StarLight Path



JGN2: 10G Network Research Testbed

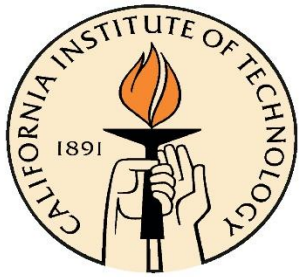
- Operational since 4/04
- 10Gbps L2 between Tsukuba and Tokyo Otemachi
- 10Gbps IP to Starlight since August 2004
- 10Gbps L2 to Starlight since September 2005

Otemachi–Chicago OC192 link replaced by 10GE WANPHY in September 2005



"Global Lambdas for Particle Physics"
A Worldwide Network & Grid Experiment

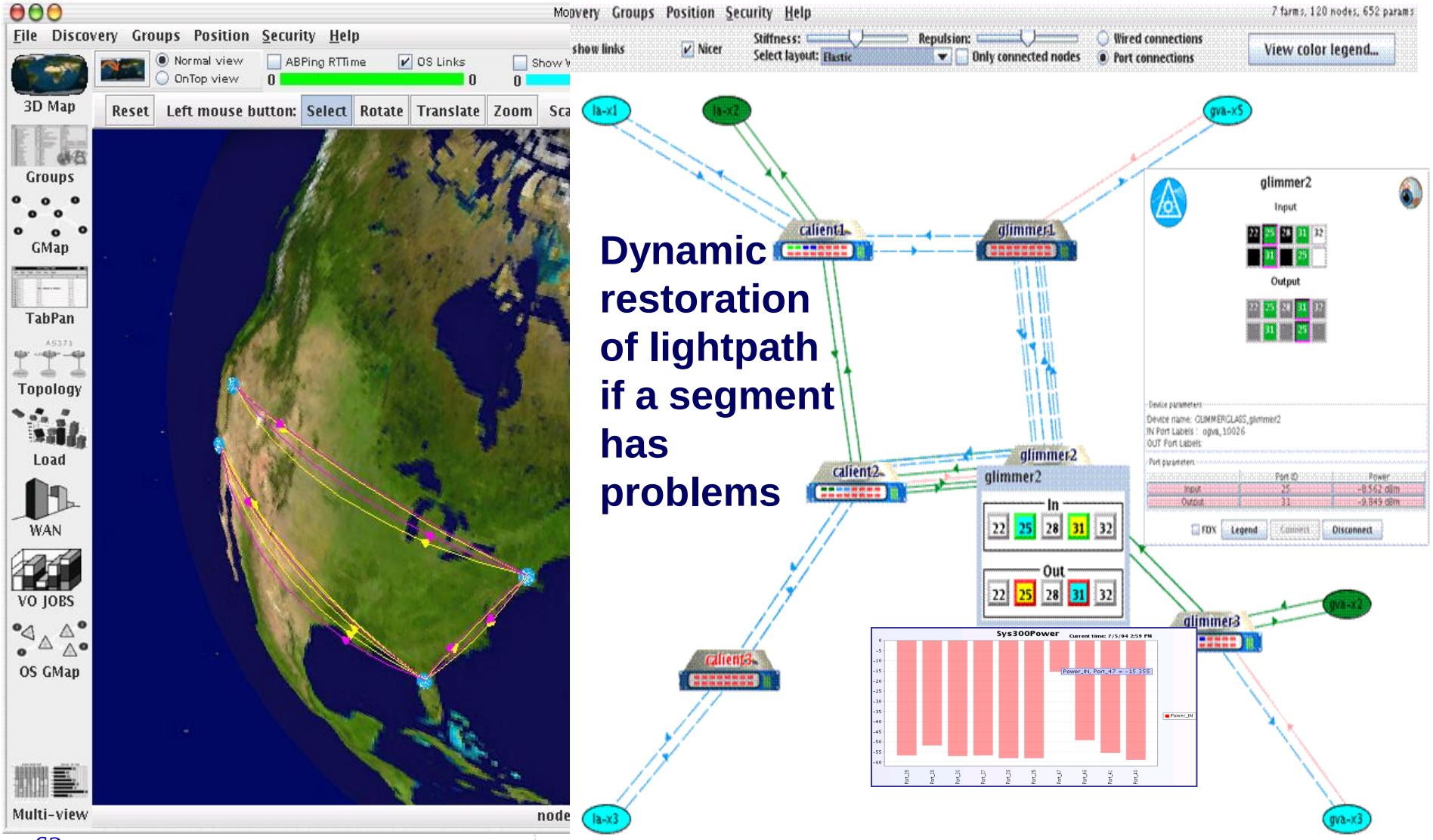
- **We have Previewed the IT Challenges of Next Generation Science at the High Energy Frontier (for the LHC and other major programs)**
  - **Petabyte-scale datasets**
  - **Tens of national and transoceanic links at 10 Gbps (and up)**
  - **100+ Gbps aggregate data transport sustained for hours; We reached a Petabyte/day transport rate for real physics data**
- **We set the scale and learned to gauge the difficulty of the global networks and transport systems required for the LHC mission**
  - **But we set up, shook down and successfully ran the system in <1 week**
- **We have substantive take-aways from this marathon exercise**
  - **An optimized Linux (2.6.12 + FAST + NFSv4) kernel for data transport; after 7 full kernel-build cycles in 4 days**
  - **A newly optimized application-level copy program, bbcp, that matches the performance of iperf under some conditions**
  - **Extension of Xrootd, an optimized low-latency file access application for clusters, across the wide area**
  - **Understanding of the limits of 10 Gbps-capable systems under stress**



"Global Lambdas for Particle Physics"
A Worldwide Network & Grid Experiment

- **We are grateful to our many network partners: SCInet, LHCNet, Starlight, NLR, Internet2's Abilene and HOPI, ESnet, UltraScience Net, MiLR, FLR, CENIC, Pacific Wave, UKLight, TeraGrid, Gloriad, AMPATH, RNP, ANSP, CANARIE and JGN2.**
- **And to our partner projects: US CMS, US ATLAS, D0, CDF, BaBar, US LHCNet, UltraLight, LambdaStation, Terapaths, PPDG, GriPhyN/iVDGL, LHCNet, StorCloud, SLAC IEPM, ICFA/SCIC and Open Science Grid**
- **Our Supporting Agencies: DOE and NSF**
- **And for the generosity of our vendor supporters, especially Cisco Systems, Neterion, HP, IBM, and many others, who have made this possible**
- **And the Hudson Bay Fan Company…**

A Real-World Working Example: Agents Create an Optical Path on Demand

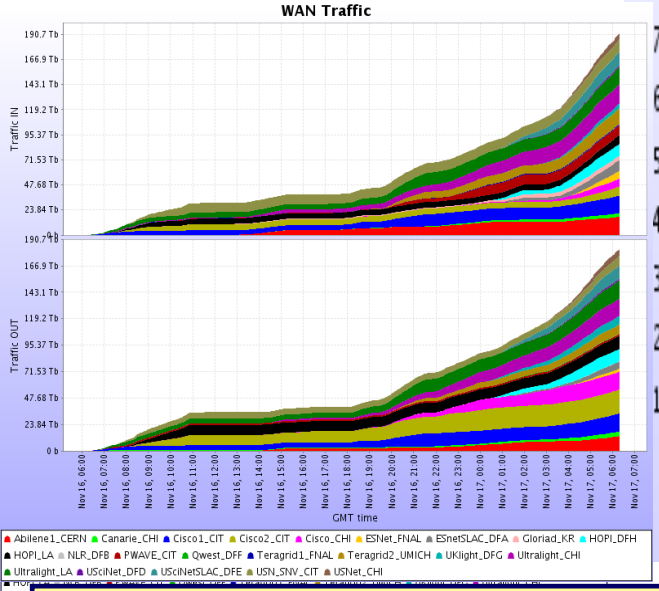



UltraLight
gridunesp
Franca
S.J. Rio Preto
Jaboticabal
Dracena
Tupã
Araraquara
Pres. Prudente
Marília
Rosana
Assis
Rio Claro
Ourinhos
S. J. dos
Guaratinguetá
Sorocaba /
Iperó
São Paulo
Itapeva
Registro
SPRACE
Brazil HEPGrid:
Rio de Janeiro (UERJ)
and Sao Paulo (UNESP)

2nd Quiet Revolution in Science and R&E Networks Continues

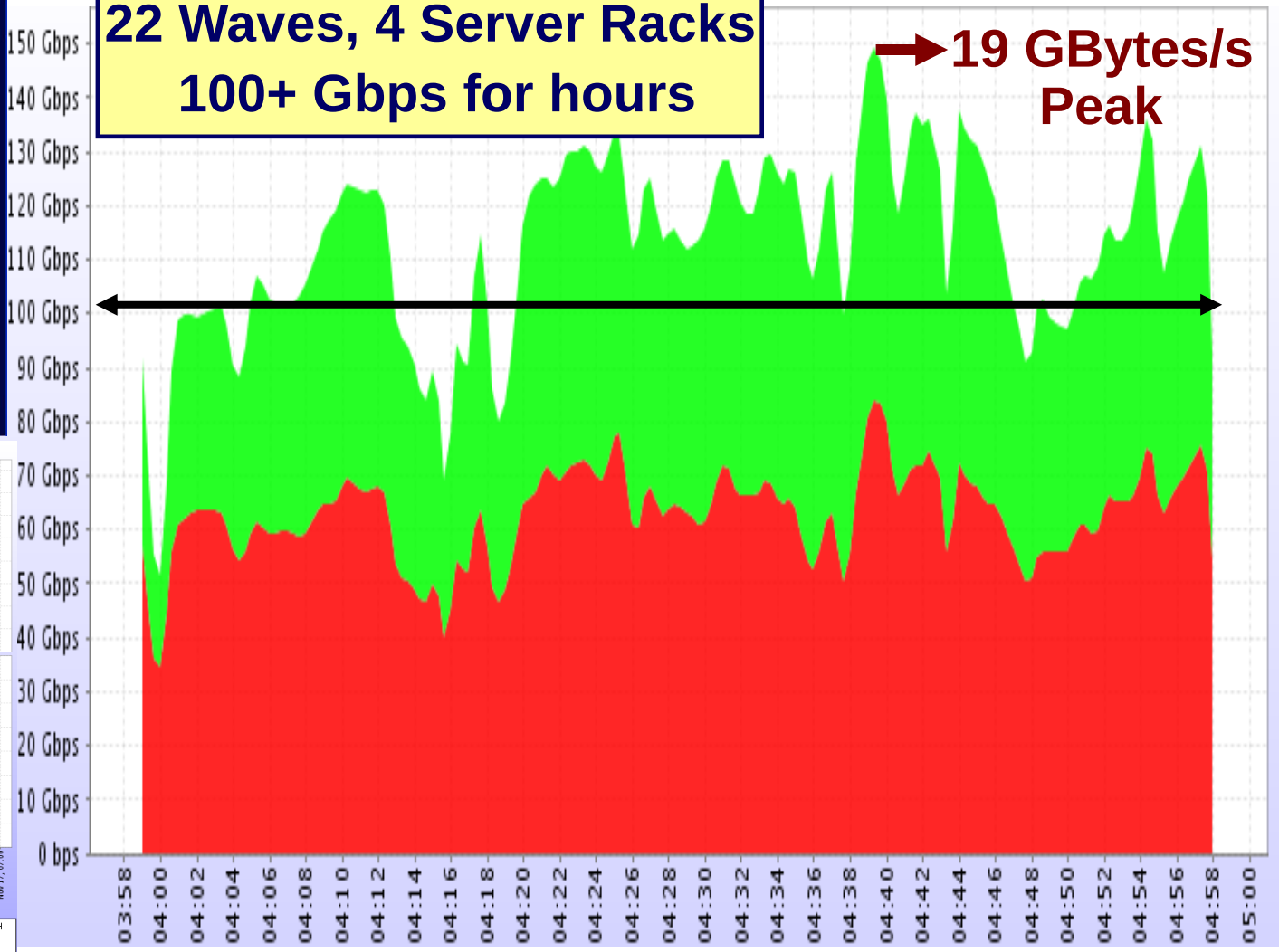
- ***2000-2007: HEP, working with computer scientists and network engineers, has developed the knowledge to use long distance networks efficiently, at high occupancy***
- **"Demystification" of large long range data flows with TCP**
  - **Exploit advances in the TCP stack (e.g. FAST TCP), Linux Kernel (2.6.20), end system architecture, network interfaces and drivers**
  - **Matching the capacity of the end-to-end path**
- **Just one to a few server-pairs with 10 GbE interfaces match a 10 Gbps Link**
- **Making the advances *widely* accessible**

SC2005 Bandwidth Challenge:
Caltech, CERN, FNAL, BNL, SLAC, UM, UF, ESNet, I2…



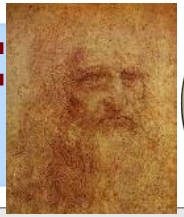
22 Waves, 4 Server Racks
100+ Gbps for hours

19 GBytes/s Peak



475 TB Total in < 24h; Sustained Rate of 1.1 Petabyte Per Day

FDT Automatic Path Recovery: Fiber Cut Simulations



MonALISA

File View Discovery Groups Security Position Help

4 services, 44 nodes, 899 params

3D Map | Groups | TabPan | GMap | Topology | Load | WAN | VO JOBS | Multi-view

Normal view / OnTop view | ABPing RTTime | OS Links | WAN Links | Animated | Net Flow

0 — 0 | 8.47 — 1354.41 Mbps | 0 — 0 Mbps

Reset | Action: Zoom / Rotate | Scale nodes | Speed

CERN Geneva

USLHCnet

Internet2

Starlight

CALTECH Pasadena

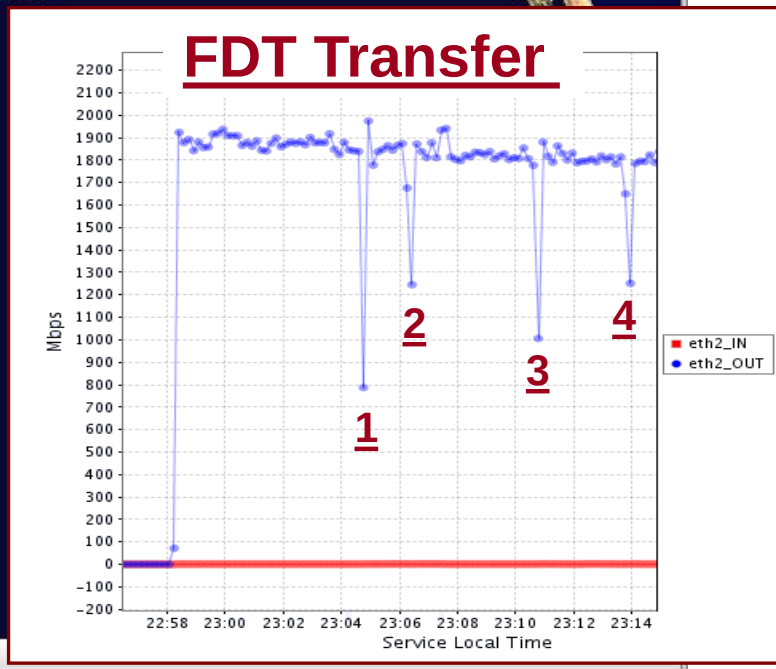
Manlan

"Fiber cut" simulations

The traffic moves from one transatlantic link to the other one

FDT transfer (CERN – CALTECH) continues uninterrupted

TCP fully recovers in ~ 20s



node(s) under cursor:

CALTECH

Mumbai-Japan-US Li

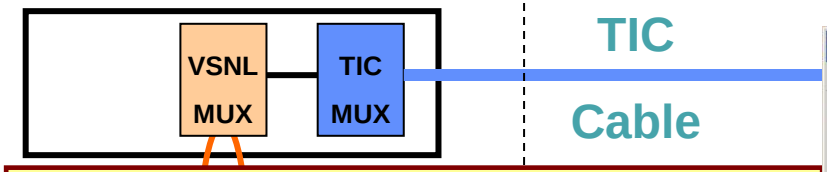
CHEP06; "Moving India Into the Global Community Through Advanced Networking"

INDIA

Chennai POP VSNL

LANDING STATIONS

SINGAPORE LANDING STATION



TIFR Link to Japan
+ Onward to US & Europe

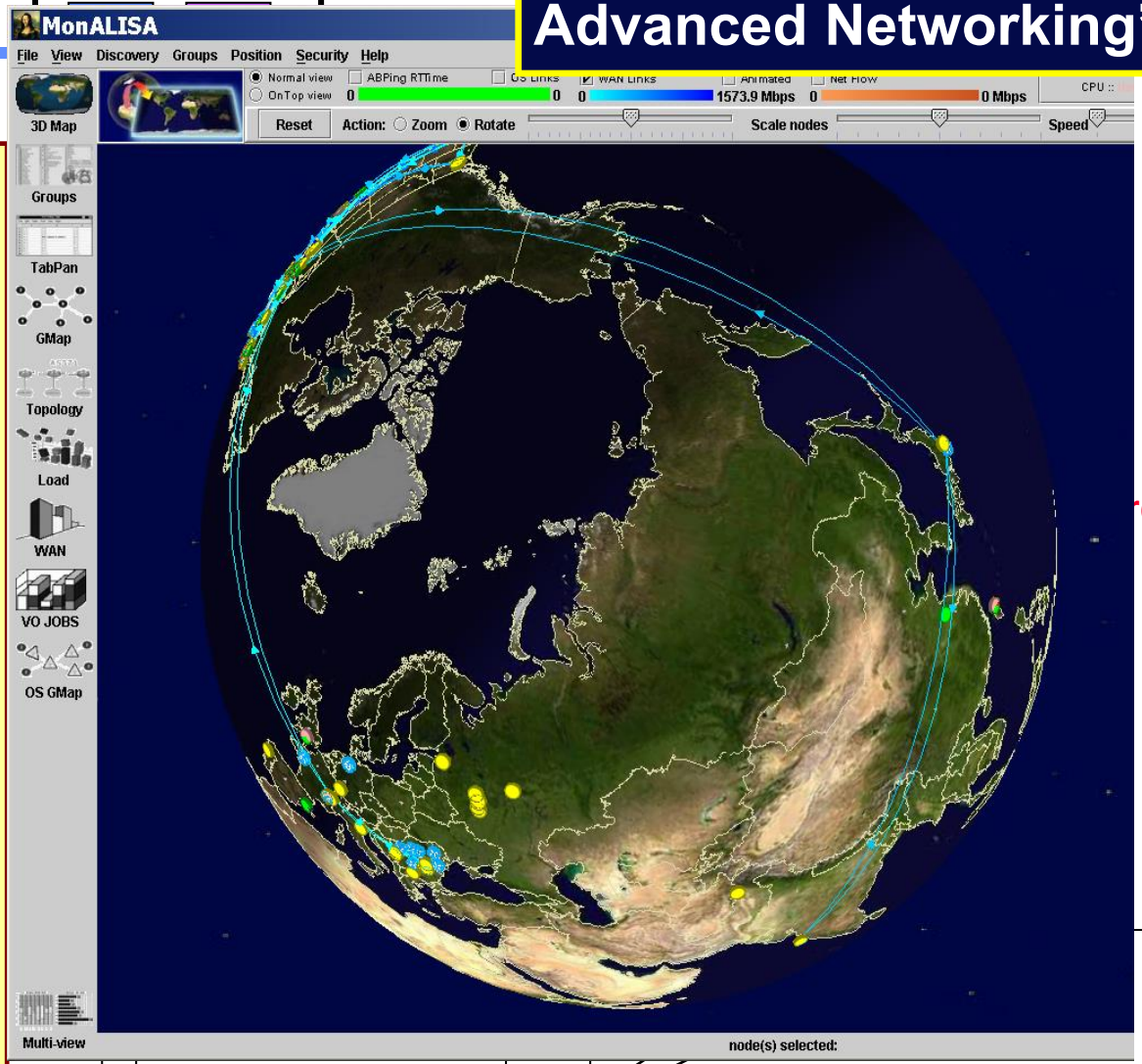
Loaned Link from VSNL at CHEP06

End to End Bandwidth 4 X 155 Mbps on SeMeWe3 Cable

Advanced applications: Data transfers (15 Tbytes in 2 days), LCG, EVO

Goal: Move to 10 Gbps

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



INTERFACE TYPES

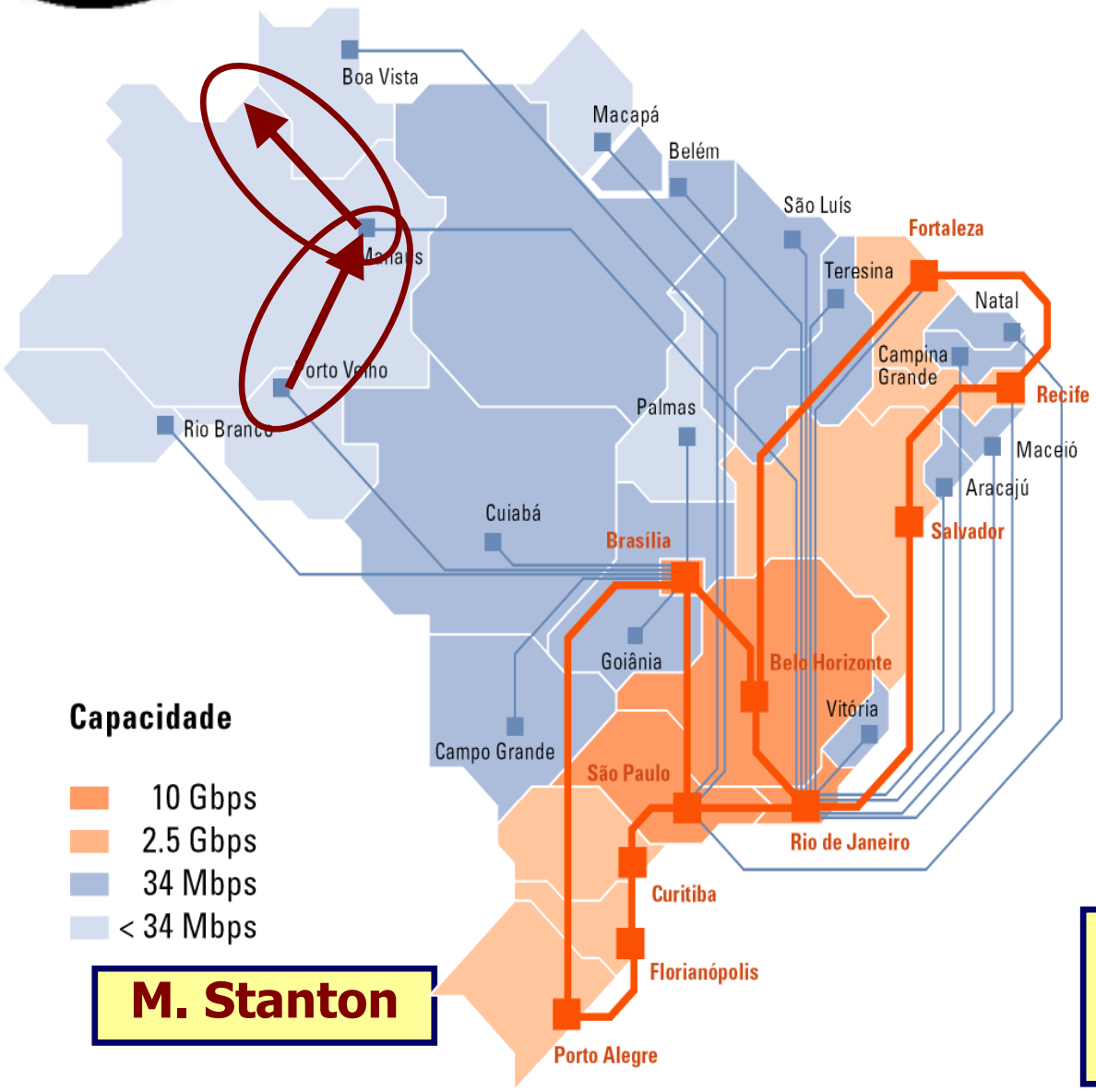
OC-12

TIFR Mumbai, INDIA

NTT Otemachi Bldg, JAPAN

+ Onward to US, Europe ➡

Brazil: RNP2 Next-Generation Backbone



M. Stanton

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

2006-7:

- **Buildout of dark fiber nets in 27 cities with RNP PoPs underway**
- **200 Institutions Connected at 1 GbE by End-2008 (Well-advanced)**
- **2.5G (to 10G) WHREN (NSF/RNP) Link to US; 622M Link to GEANT**

Now extending to the Northwest; Dark fiber across the Amazon to Manaus

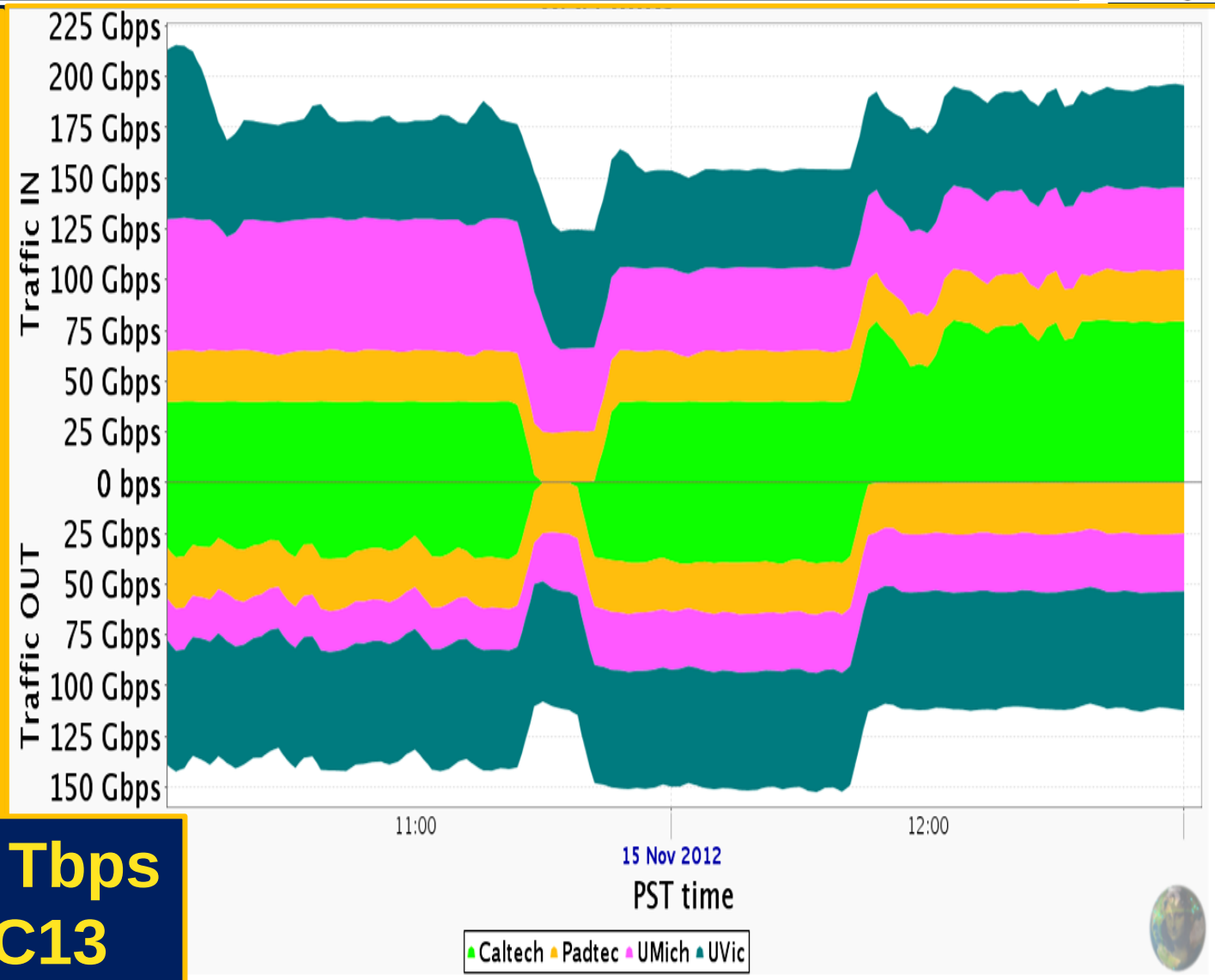
Memory to Memory at SC12
11/14-15/12 Using Caltech's FDT

FDT Memory to Memory

339 Gbps Peak

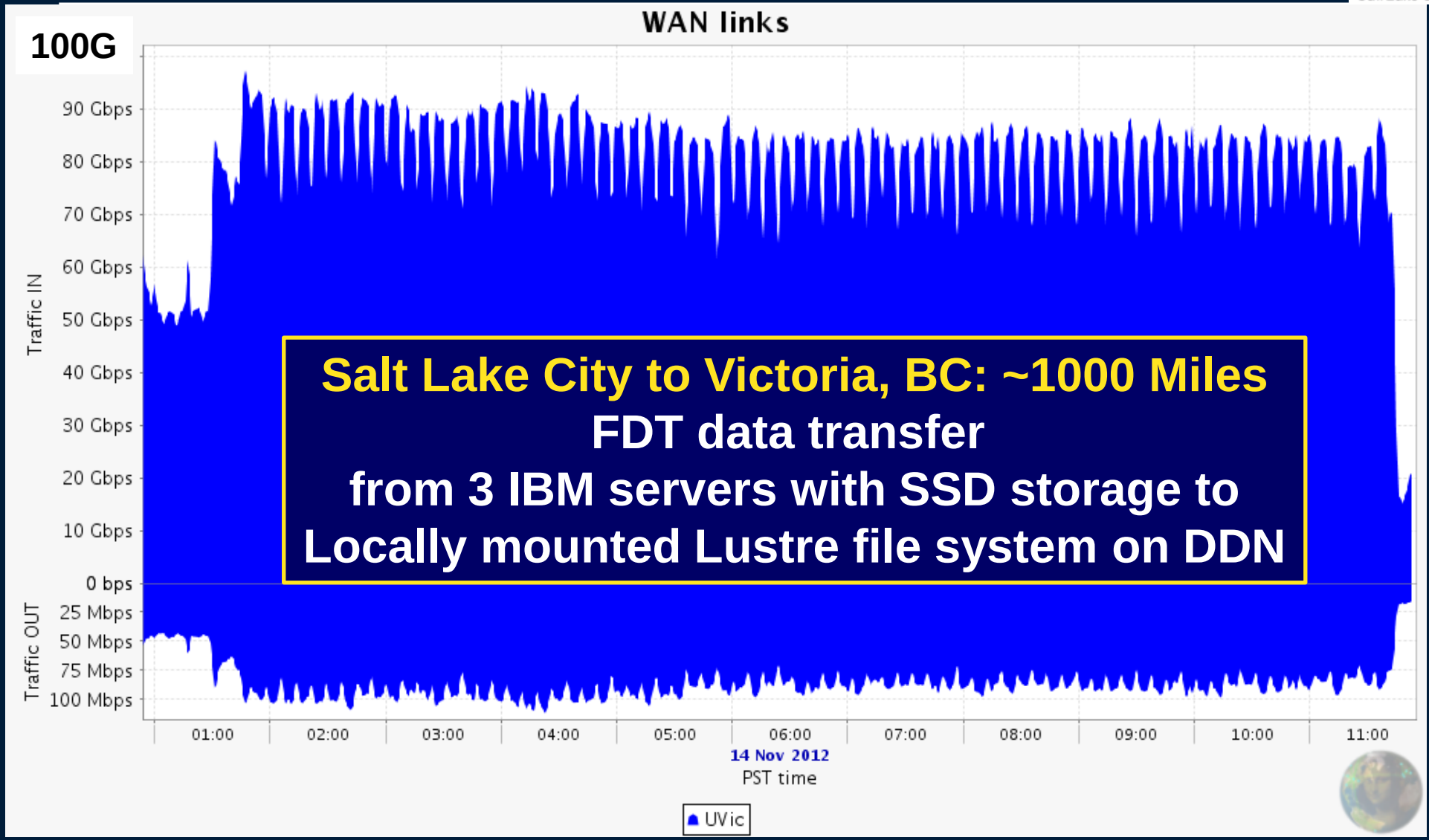
300+ Gbps Sustained from Caltech, Victoria, Umich

To 3 Pbytes Per Day



We are ready for Tbps Transfers at SC13

SC12: The Real Thing: Storage-to-Storage over 100G with FDT

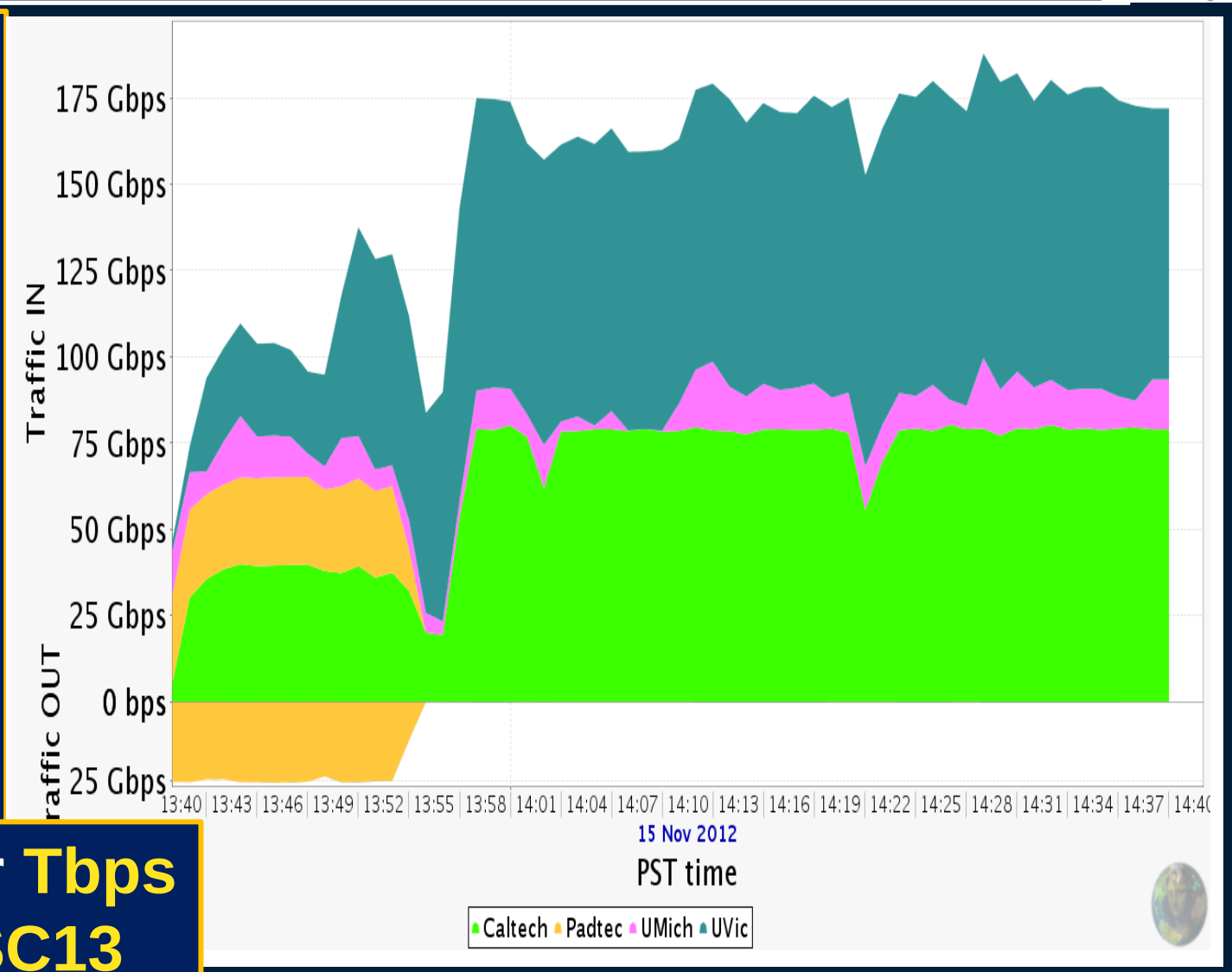


FDT Storage to Storage at SC12 Nov. 15 2012

FDT Storage to Storage

175 Gbps (186 Peak) from Caltech, Victoria and Michigan to SLC

We are ready for Tbps Transfers at SC13

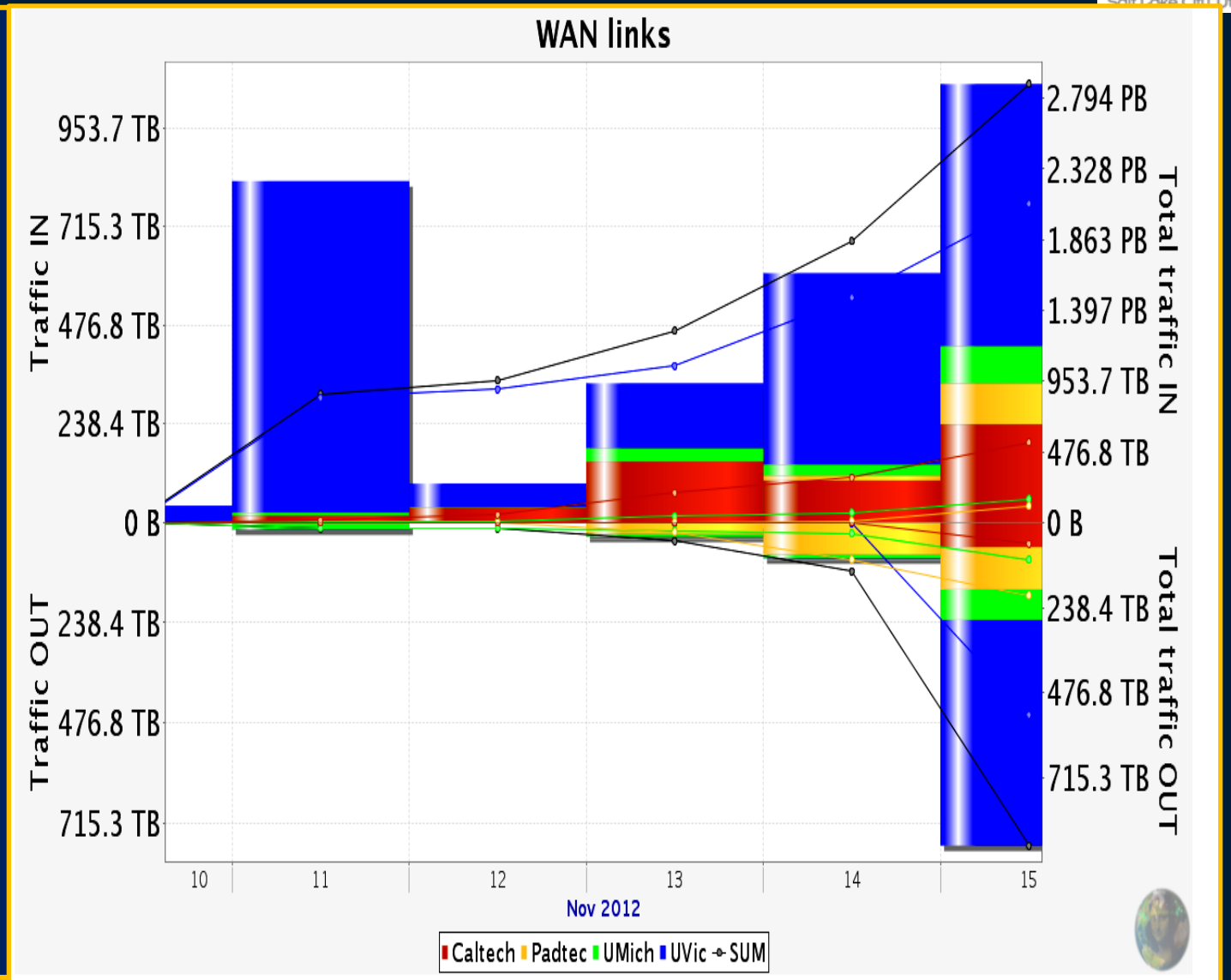


Transferring Petabytes at SC12

FDT and RDMA over Ethernet

3.8 PBytes to and From the Caltech Booth

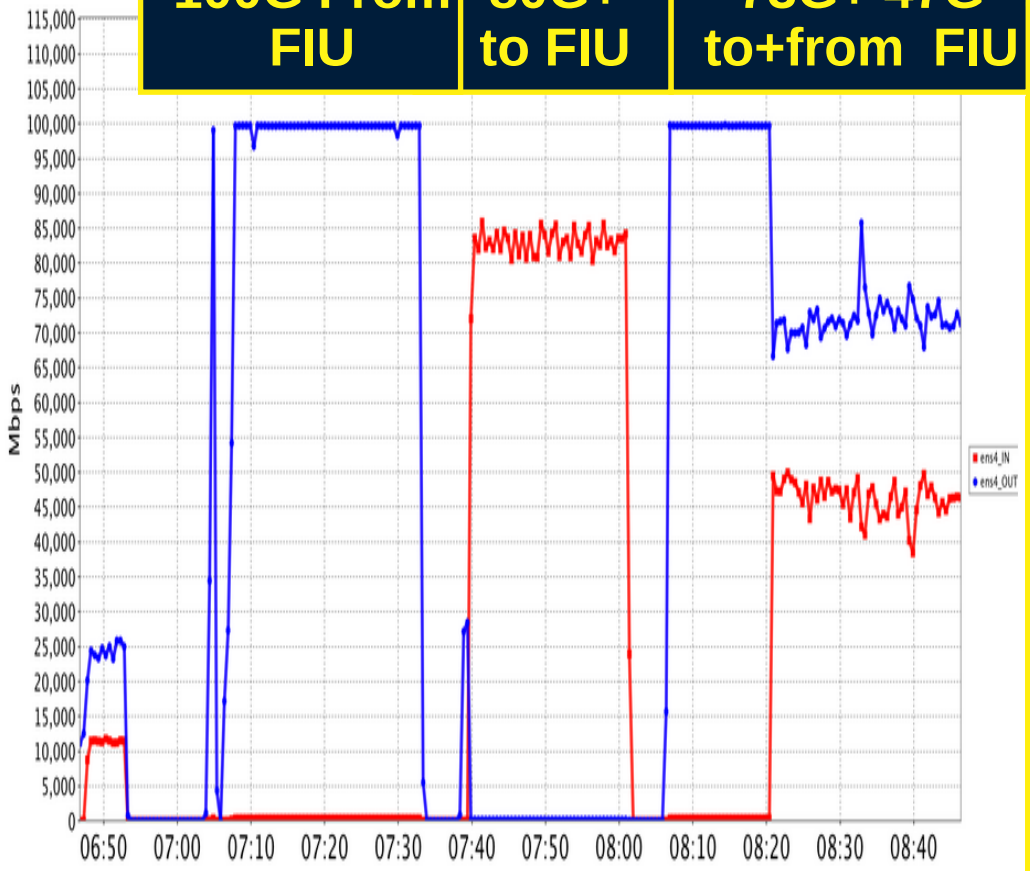
Including 2 PBytes on Nov. 15



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

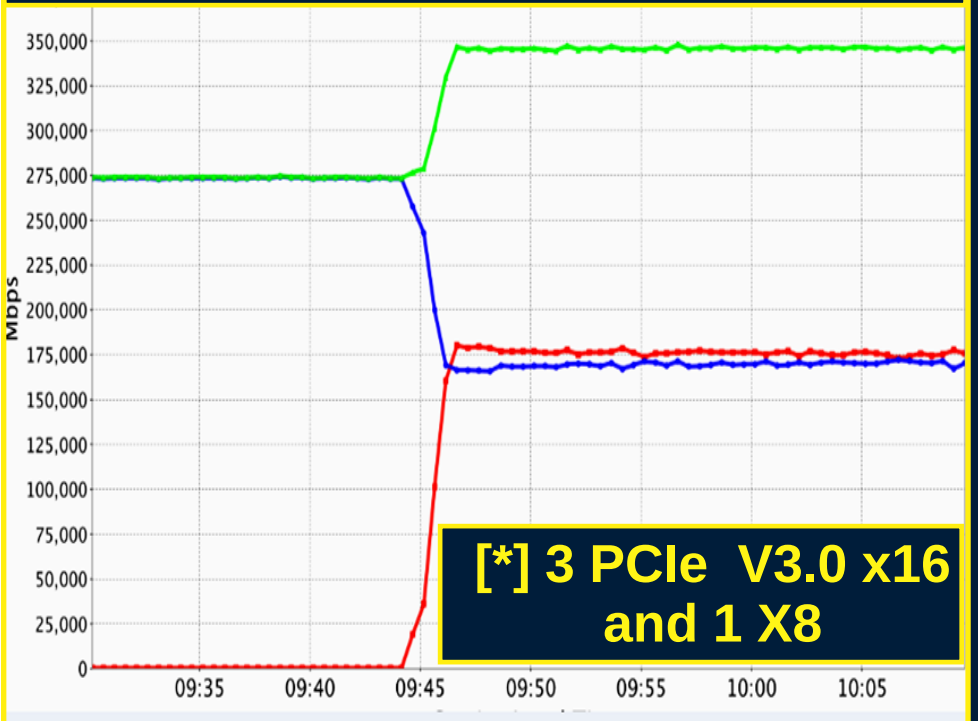
FIU – Caltech Booth – Dell Booth

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



4 X 100G Server Pair in the Caltech Booth

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



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

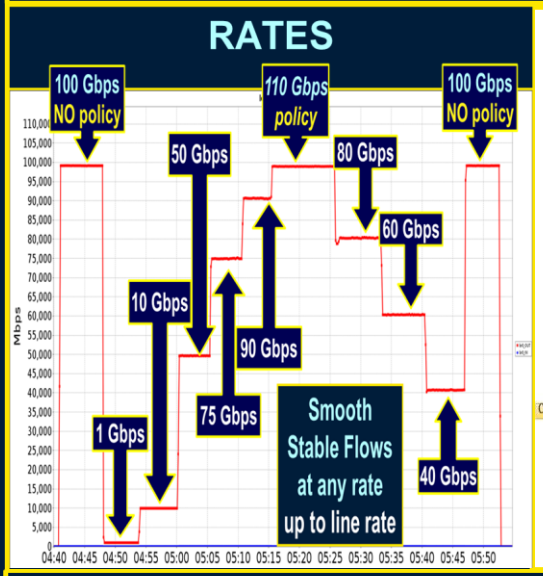
Using Caltech's FDT Open Source TCP Application
http://monalisa.caltech.edu/FDT

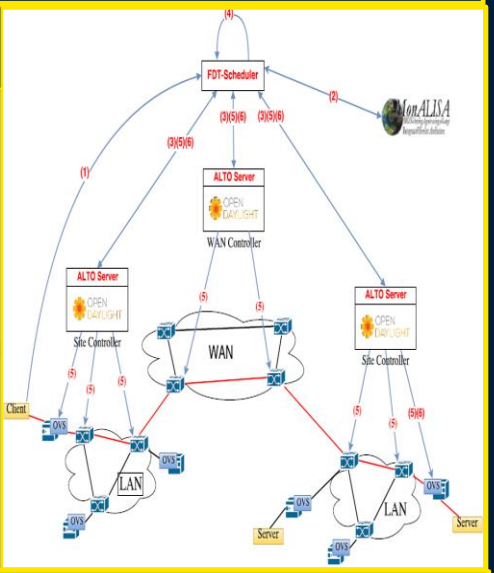
Next Generation "Consistent Operations": Site-Core Interactions for Efficient, Predictable Workflow

- Key Components: (1) Open vSwitch (OVS) at edges to stably limit flows, (2) Application Level Traffic Optimization (ALTO) in Open Daylight for end-to-end optimal path creation, flow metering and high watermarks set in the core
- Real-time flow adjustments triggered as above
- Optimization using "Min-Max Fair Resource Allocation" (MFRA) algorithms on prioritized flows
- Flow metering in the network fed back to OVS edge instances; to ensure smooth progress of flows end-to-end

Demos: Internet2 Global Summit in May
SC16 in November

Consistent Ops with ALTO, OVS and MonALISA FDT Schedulers





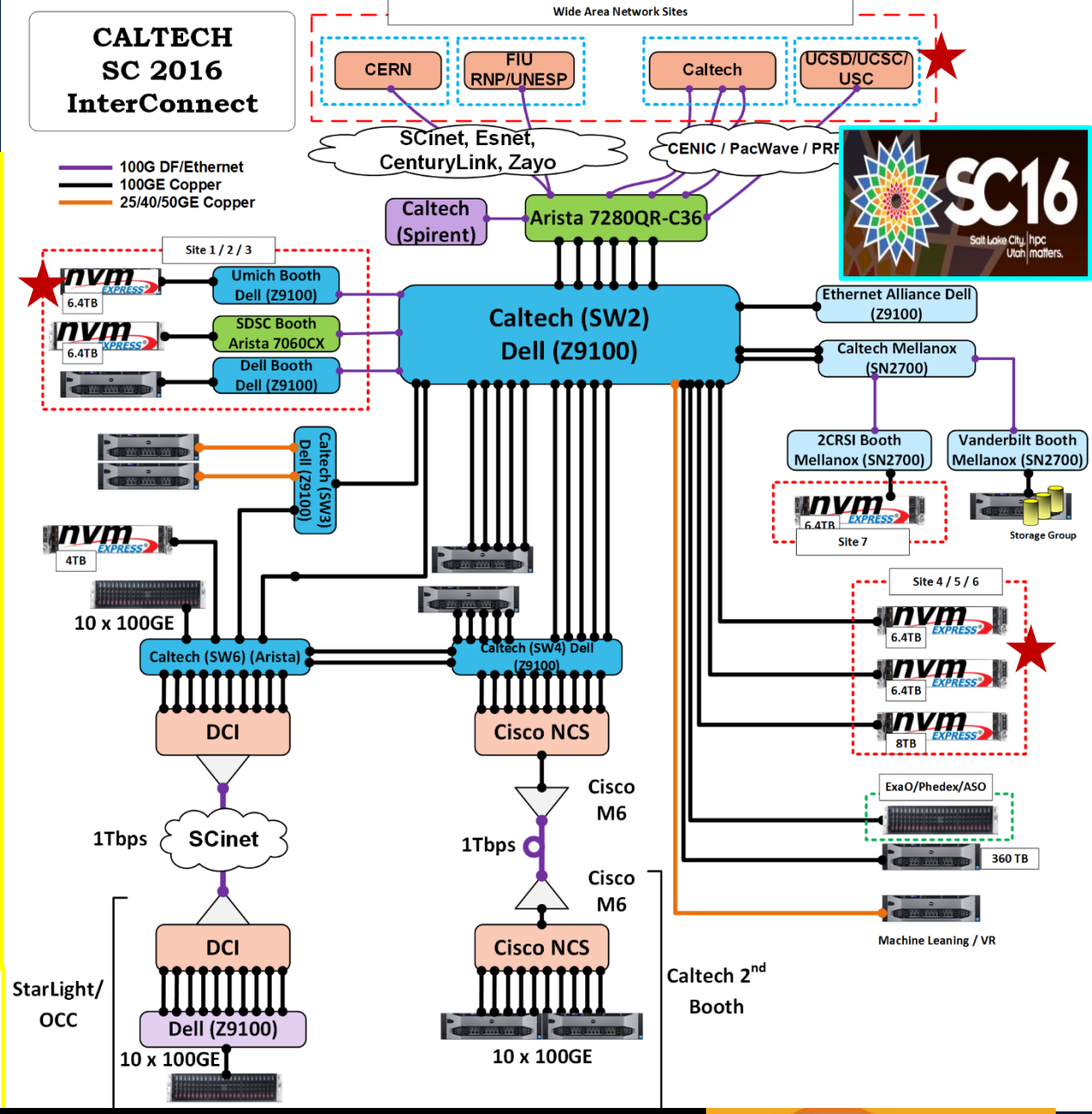
- Real-time adjustment of allocations triggered by: (1) new requests, (2) real-time feedback on progress of transfers, (3) network state changes or error conditions, (4) proactive load-balancing operations, or (5) rate-limiting operations imposed by controllers or emerging network operating systems (e.g. SENOS)

With Yale CS Team: Y. Yang, Q. Xiang et al.

Caltech at SC16

- Terabit/sec ring topology: Caltech – Starlight – SCInet; > 100 Active 100G Ports
- Interconnecting 9 Booths: Caltech 1 to 1 Tbps in booth, and to Starlight 1 Tbps; UCSD, UMich, Vanderbilt, Dell, Mellanox, HGST @100G
- WAN: Caltech, FIU +UNESP (Sao Paulo), PRP (UCSD, UCSC, USC), CERN, KISTI, etc.

★ ExaO + PhEDEx/ASO CMS Sites

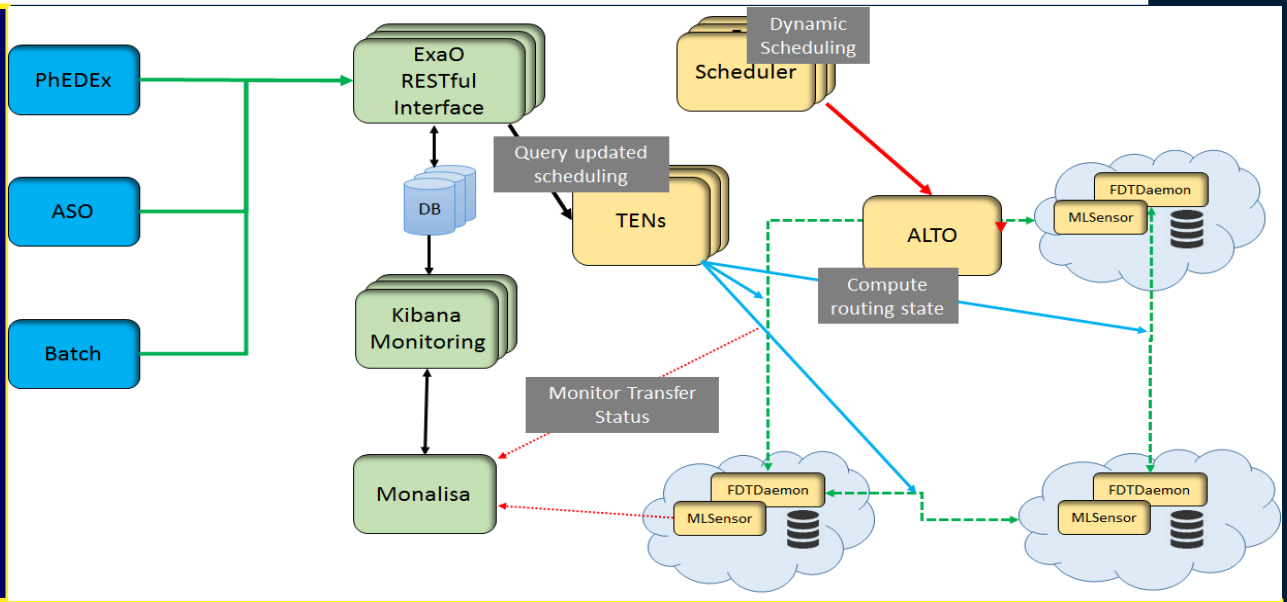


Looking Forward: **We will start work on SC17 and will be looking for network and research site partners Soon**

SC17 Denver, CO hpc connects.

CMS at SC16: *ExaO* - Software Defined Data Transfer Orchestrator with Phedex and ASO

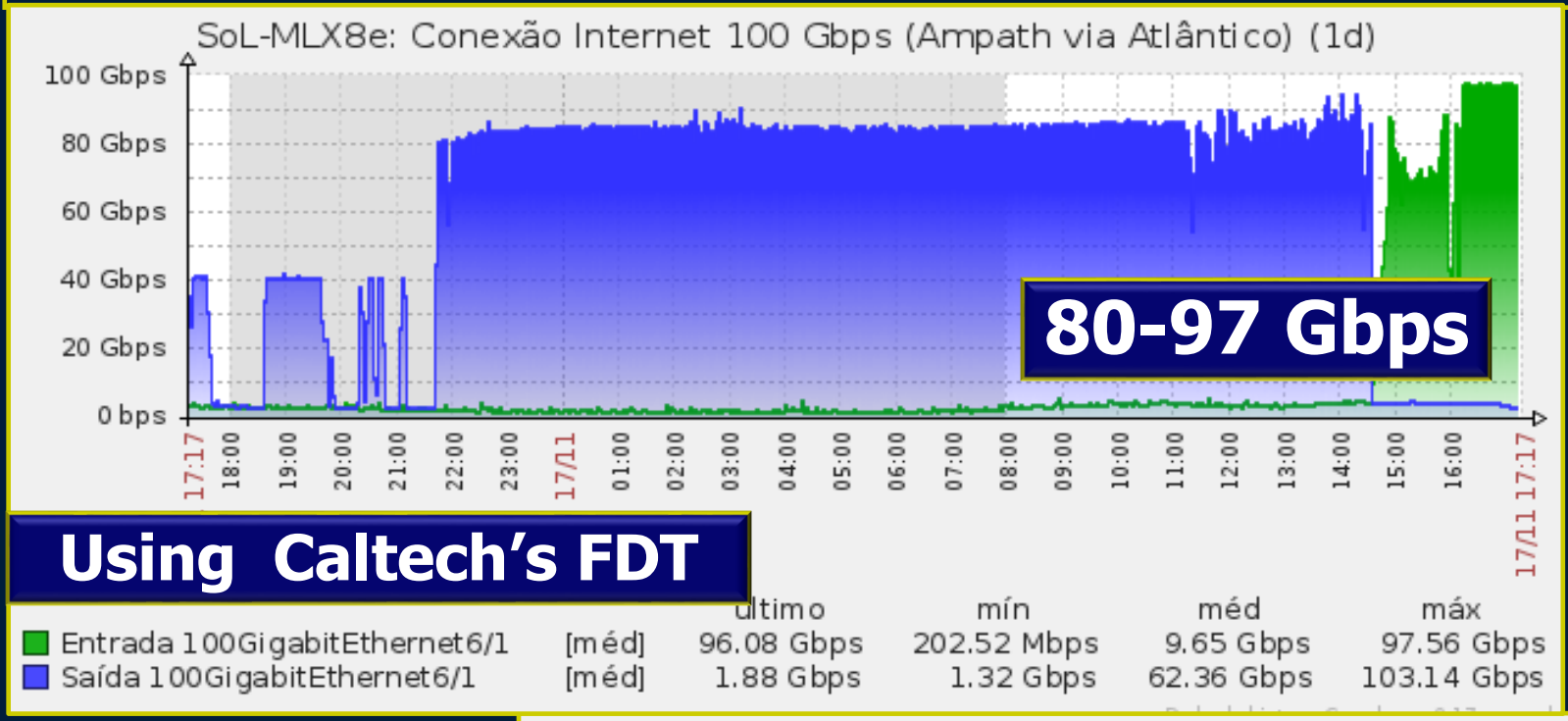
Leverage emerging SDN techniques to realize end-to-end orchestration of data flows involving multiple host groups in different domains



- Maximal link utilization with ExaO:
  - PhEDEx: CMS data placement tool for datasets
  - ASO: Stageout of output files from CMS Analysis Jobs
- Tests across the SC16 Floor: Caltech, UMich, Dell booths and Out Over the Wide Area: FIU, Caltech, CERN, UMich
- Dynamic scheduling of PetaByte transfers to multiple destinations

Partners: UMich, StarLight, PRP, UNESP, Vanderbilt, NERSC/LBL, Stanford, CERN; ESnet, Internet2, CENIC, MiLR, AmLight, RNP, ANSP

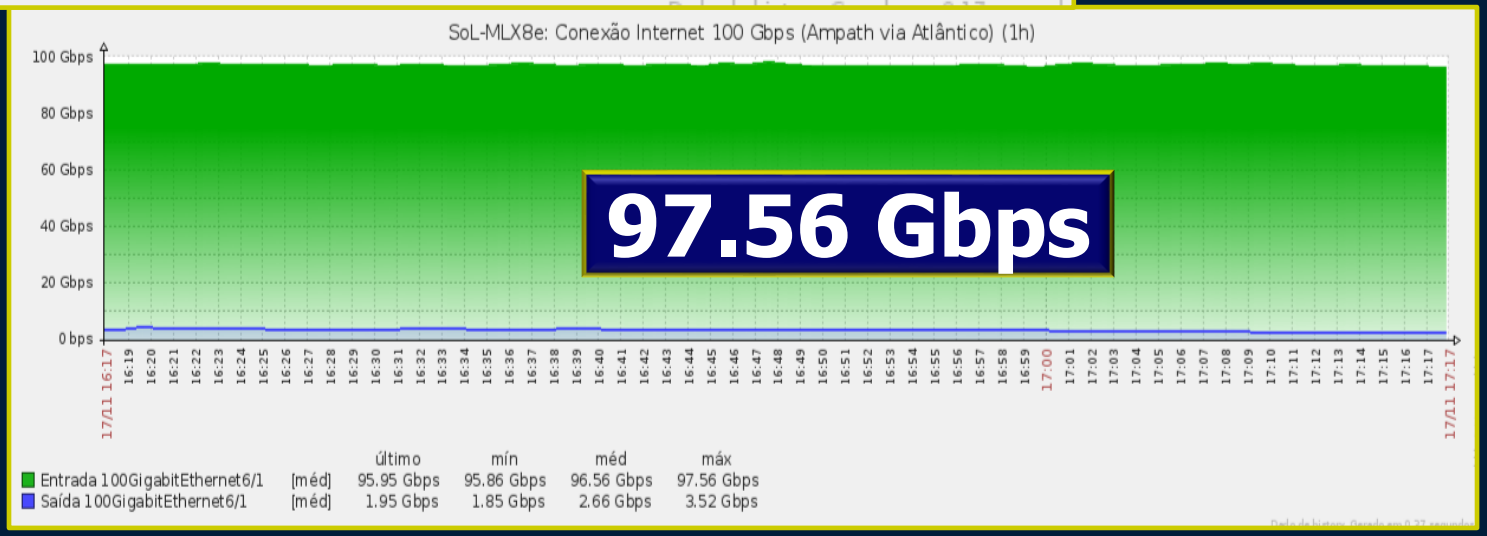
GridUnesp: Transfer Demo at SC16



Using Caltech's FDT

17 Hour transfer overnight on Miami-Sao Paulo Atlantic link

1 Hour transfer on Miami-Sao Paulo Atlantic link



Game Theory and the Future of Networking

http://blog.eai.eu/game-theory-and-the-future-of-networking/

- **Game theory:** Mathematical models of conflict and cooperation among intelligent rational decision-makers
  - Studies participants' behavior in strategic situations.
- Motive and the need for Increased Reach induce selfish entities to cooperate in pursuit of a common goal
- Application Pull: the Internet calls for analysis and design of systems that span multiple entities with diverging information and interests
- Technology Push: math and science mindset of GT is similar to that of (many) scientists
- Fields of Use: economics, political science, psychology, logic, computer science, biology, poker… and now HEP exascale data

- Emergence of the internet has motivated development of GT algorithms for finding equilibrium in games, markets, auctions, peer-to-peer systems, security and information markets

- GT is now applied to a wide range of behaviors
- It has become an umbrella term for the science of logical decision making
- In and among humans and computers

- *Coherent Interactions* among the experiments' workflow management systems, the end sites, the network and the user groups *as a System*

openlab.web.cern.ch