ESnet
ENERGY SCIENCES NETWORK

DC24 from the ESnet perspective

Kate Robinson
Eli Dart
Dale Carder

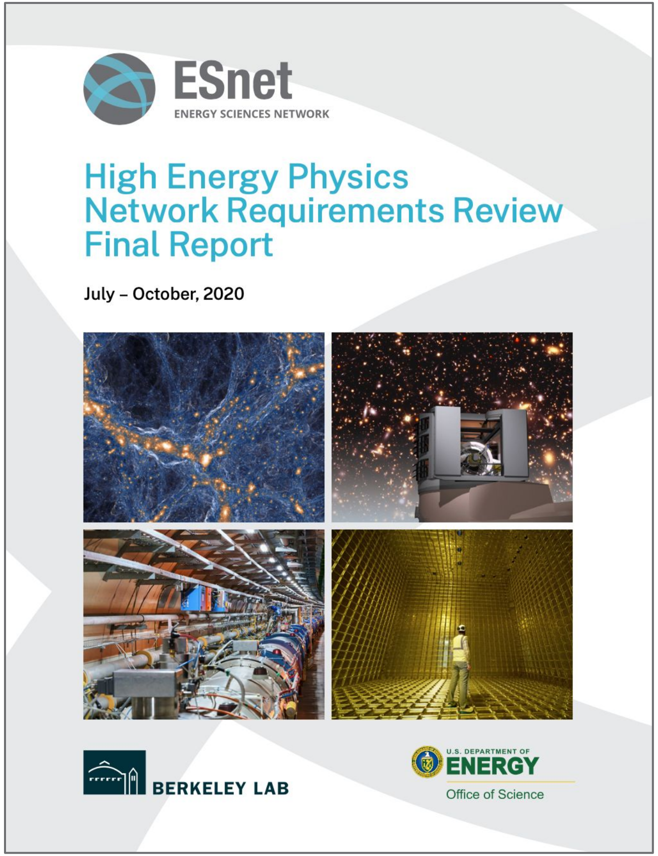
LHCOPN/LHCONE #52
University of Catania
10-12 April 2024

U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

Process started > 4yrs ago!

- Report identified connectivity needs for US-ATLAS and US-CMS
- Projections codified into data challenges
- Data Challenges (full software stack)
  - 1: 10% of the target 2021
  - 2: 25% in 2024
  - 3: 50%?



ESnet
ENERGY SCIENCES NETWORK

High Energy Physics Network Requirements Review Final Report

July – October, 2020

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

https://doi.org/10.2172/1804717

Campaign to engage w/ each LHC site on ESnet

- Multi-year effort to meet with *everyone* in preparation for DC24
  - University Physics Dept PI's
  - Tier-2 site staff
  - University central IT network groups
  - Regional networks
  - R&E Exchange points
  - National Labs
  - CERN
- Over 70 individual issues tracked for follow up
  - Upgrades
  - performance issues
  - Many efforts required multi-party coordination!

Regional Connectivity Upgrades

- ESnet focused extensively on upgrading connections to regional exchange points for US Tier-2 sites in preparation for DC24
  - 2x100G CENIC in Los Angeles
    - CalTech, UCSD
  - 400G GPN
    - Nebraska, SWT2
  - 100G NOX 100G in Boston
    - Supporting MIT & New NET2
  - 400G SOX Nashville
    - Vanderbilt, U-Florida
  - 4x100G + 4x100G OmniPoP
    - Michigan
    - Michigan State
    - U-Chicago
    - IU-Bloomington
    - UIUC
    - Purdue
    - UW-Madison

Other DC24 preparation work

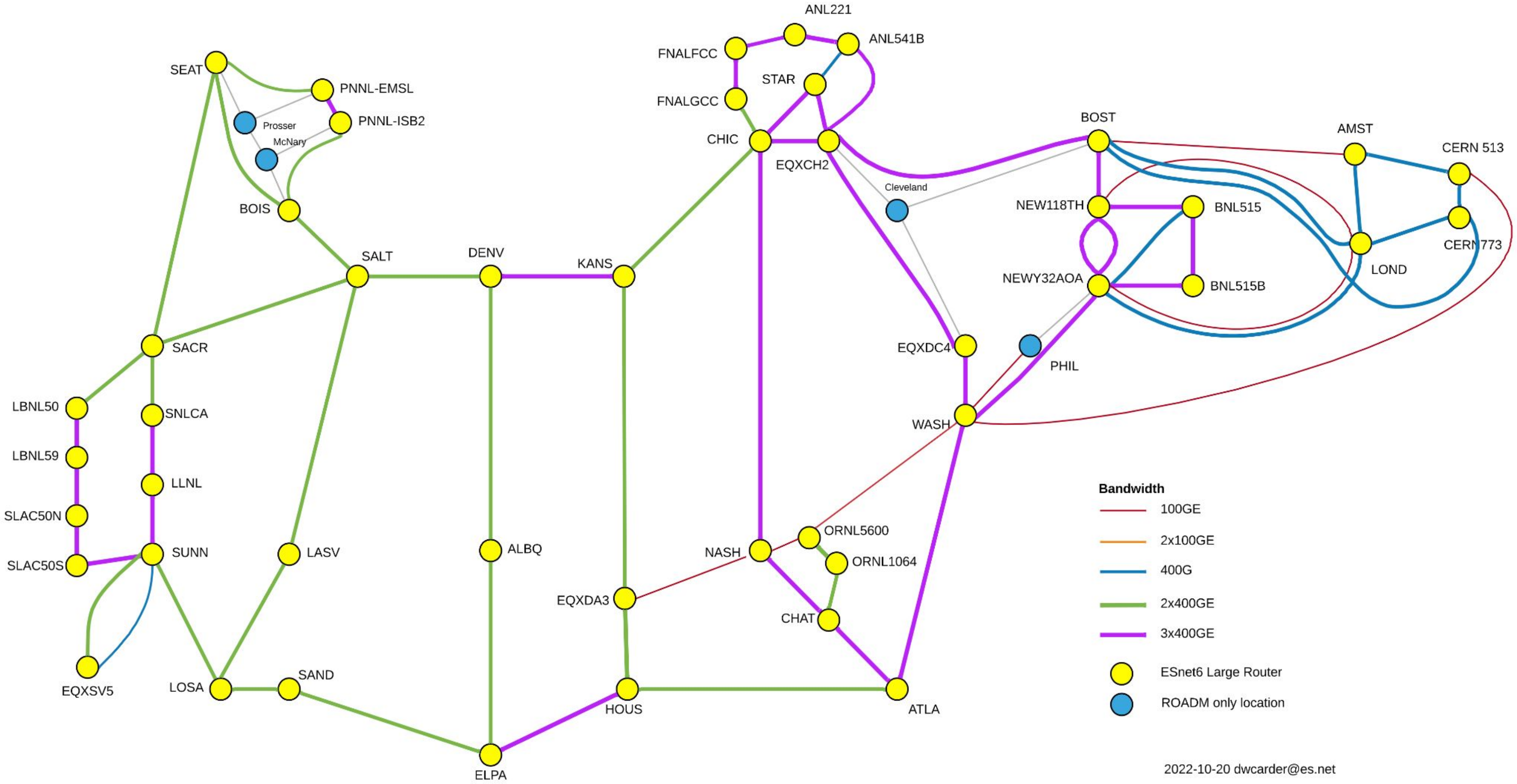
- Developed traffic-engineering solution (discussed in TA talk)
- 100G ⟶ 400G peering upgrades to GEANT
- PerfSonar deployments on LHCONE
- Stardust dashboards
- Portal updates, portal updates, portal updates (adding TA links..)

ESnet

US Tier 1 Connectivity during DC24

- ESnet6 installed routers collocated at our sites
- Most are connected to our optical system at 1.2Tbit + redundancy
- We are now ready to accommodate upgrades as sites are able
  - BNL - US ATLAS Tier 1
    - 2 x 400G + 2 x 400G
  - FNAL - US CMS Tier 1
    - Current: 400G (4 x 100G + 2 x 100G)
    - Near (very near) Future: 800G (1 x 400G + 1 x 400G)

ESnet Topology during DC24

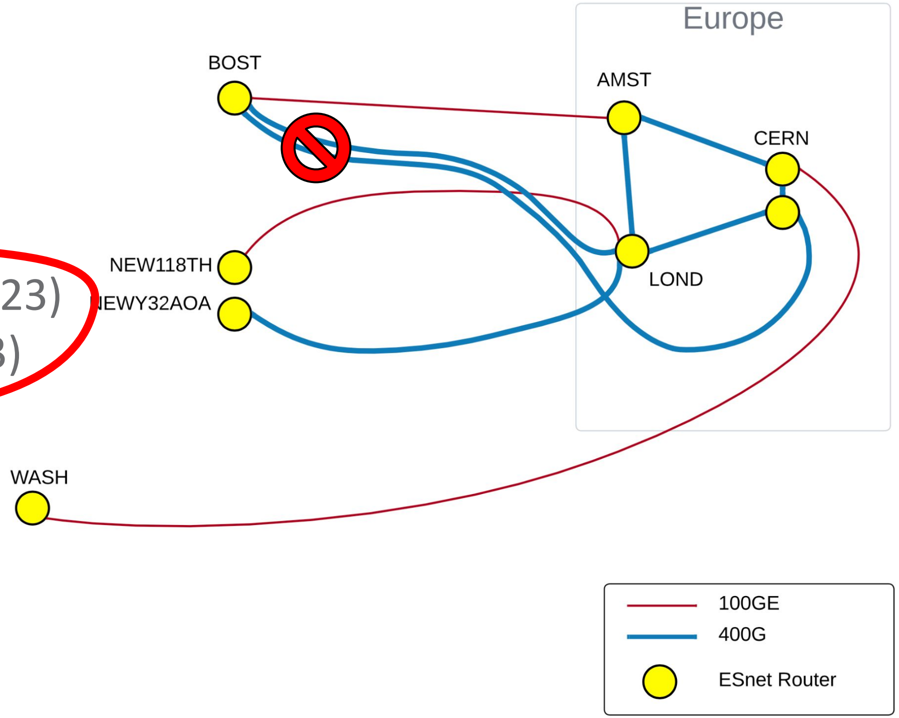


2022-10-20 dwcarder@es.net

Transatlantic 400G upgrades were delayed by the cable vendor

- **In Production:**
  - 400G New York - London
- Currently underway:
  - 400G Boston - London (Estimated 12/23)
  - 400G Boston - CERN (Estimated 12/23)

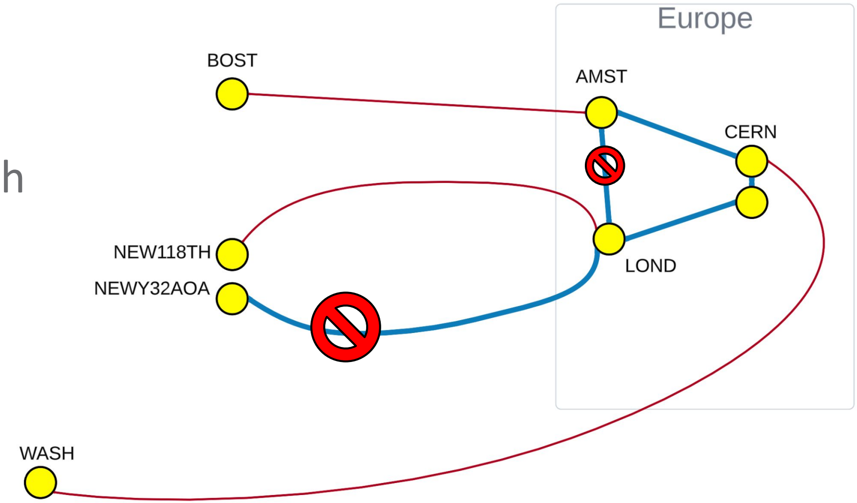
- Trans-Atlantic capacity targets
  - 1.5T in advance of DC24 **:-(**
    - **reality was closer to ~800**
    - **and then we had an outage ...**



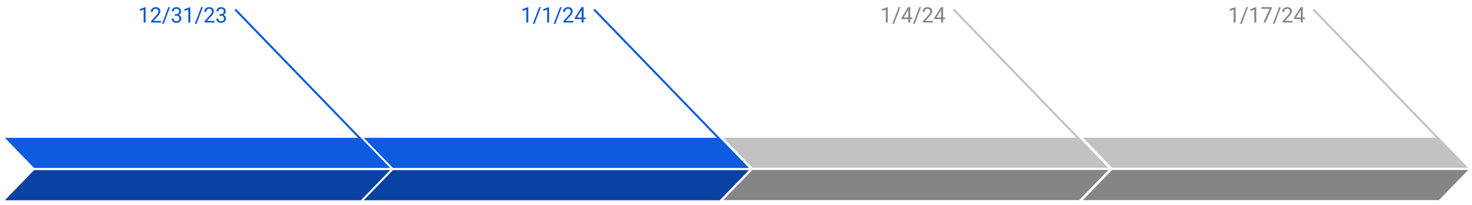
ESnet

Outages, as per data-challenge tradition!

- From 800G to 300G
- New York - London
  - 400G + 100G from NEAAR on same path
  - Cut
  - Restored
  - Planned maintenance saga
  - Mitigation Plan
- Amsterdam - London
  - Cut
  - Restored
  - Impacted Primary LHCOPN path to FNAL (Also CERN to RAL)



AEC-1 Cable outage



12/31/23

Outage began

Outage has been identified as a subsea wet plant issue on AEC1 between Ireland and New York. The fault location is believed to be between repeaters 10 and 11 from the Killala side and engineers continue to investigate further

1/1/24

Ship sent for repairs

Aqua Comms sent ship to repair cable. However, Aqua Comms has also identified some preventative maintenance needed on AEC-1 so cable will not be back in production until AFTER DC24.

1/4/24

ESnet coordinated with Aqua Comms

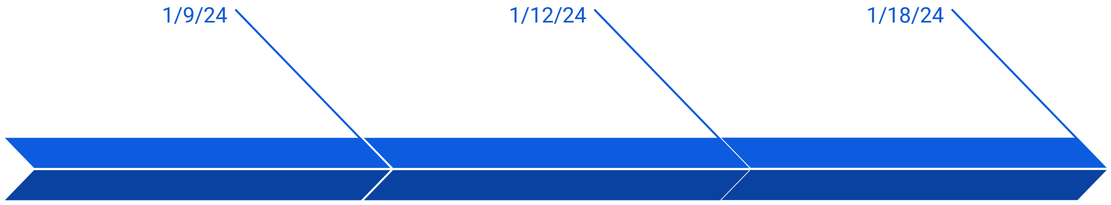
ESnet successfully coordinated with Aqua Comms at PTC and requested they move the preventative maintenance to after DC24

1/17/24

Back in Service

Cabled spliced and service restored

Added AEC-2 100G TA link



1/9/24

Emailed IU

IU International networks had reached out previously to offer us a backup on AEC-2 cable

1/12/24

IU Provisioned

Provisioned end to end

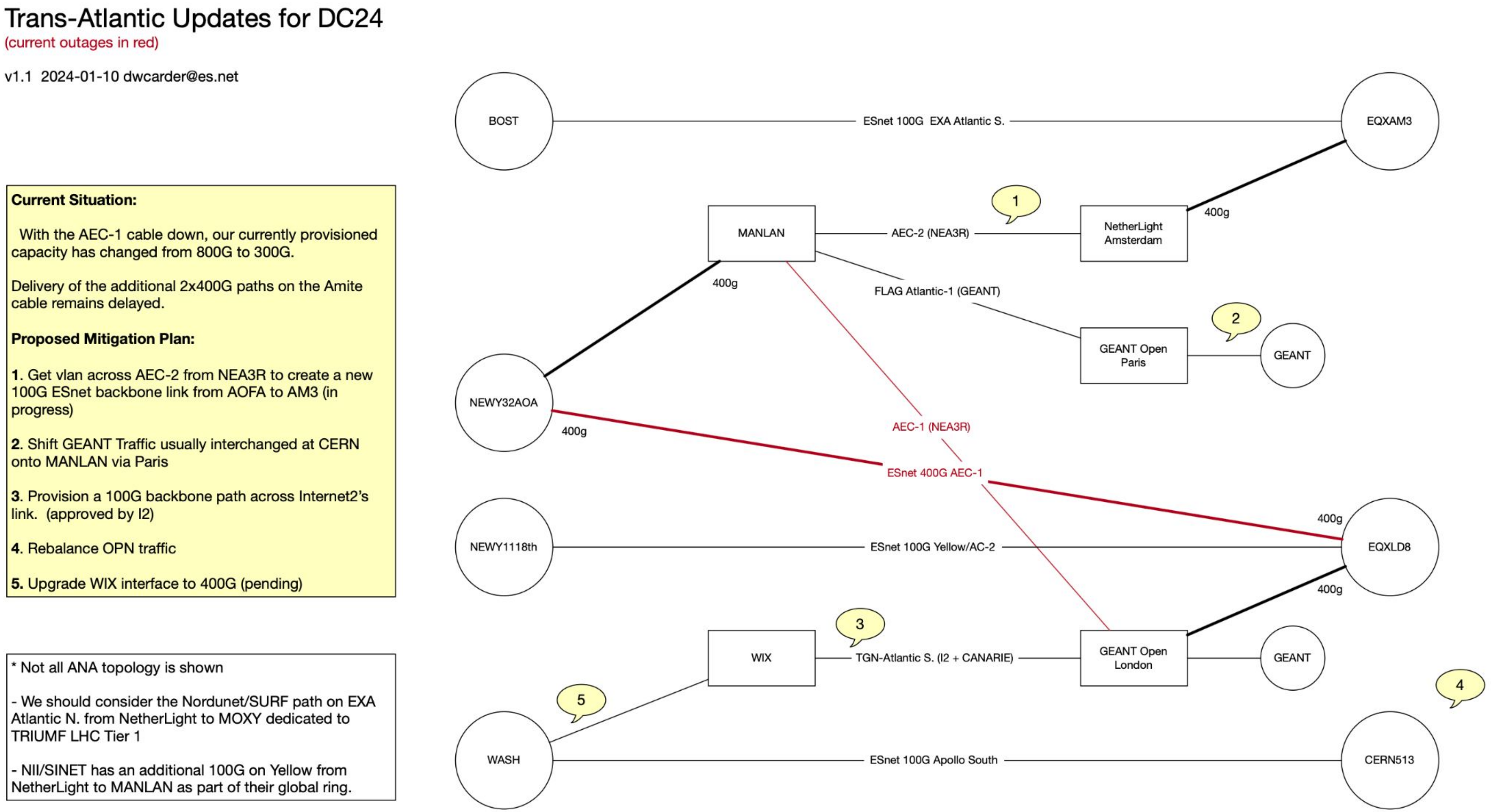
1/18/24

Tested, In-service

ESnet tested and moved circuit to production

Shoutout to ANA, IU International networks / Brenna Meade for all coordination and fast provisioning from MANLAN/I2 and NetherLight/SURF

Mitigation: use all R&E 100G TA links



Trans-Atlantic Updates for DC24

(current outages in red)

v1.1 2024-01-10 dwcarder@es.net

Current Situation:

With the AEC-1 cable down, our currently provisioned capacity has changed from 800G to 300G.

Delivery of the additional 2x400G paths on the Amite cable remains delayed.

Proposed Mitigation Plan:

1. Get vlan across AEC-2 from NEA3R to create a new 100G ESnet backbone link from AOFA to AM3 (in progress)

2. Shift GEANT Traffic usually interchanged at CERN onto MANLAN via Paris

3. Provision a 100G backbone path across Internet2's link. (approved by I2)

4. Rebalance OPN traffic

5. Upgrade WIX interface to 400G (pending)

* Not all ANA topology is shown

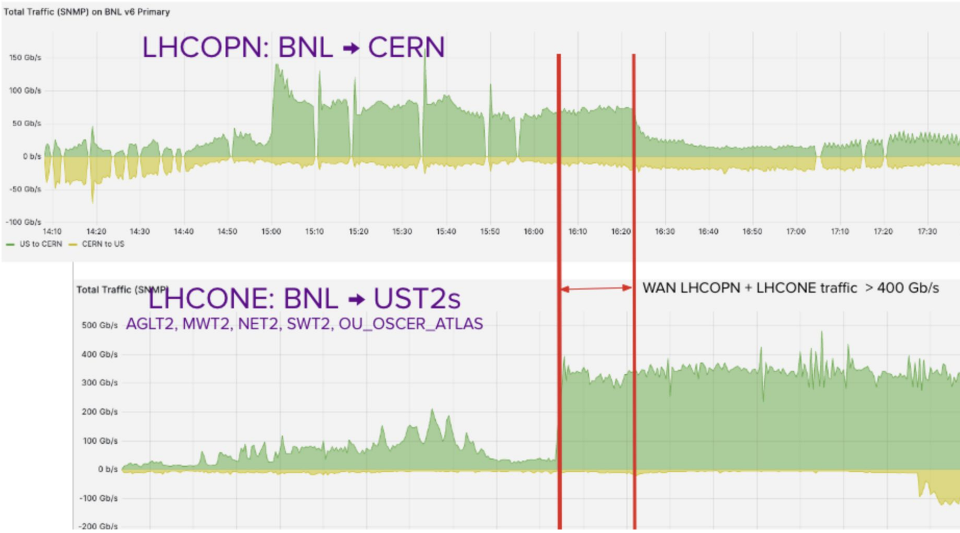
- We should consider the Nordunet/SURF path on EXA Atlantic N. from NetherLight to MOXY dedicated to TRIUMF LHC Tier 1

- NII/SINET has an additional 100G on Yellow from NetherLight to MANLAN as part of their global ring.

Other Issues Encountered

- Topology changes due to outages affected LHCOPN paths
  - Needed to move around outages & overlap
- Changing an OSCARS (LHCOPN) L2 circuit induces outages
  - Current process is delete old, create new
  - (post DC24) implemented phase 1 of improvements to change LSP's
    - adjust bandwidth
- Currently no easy way to visualize multiple LSPs are for LHCOPN users other than the ESnet portal.

Pre-DC24 Efforts



WAN LHCOPN + LHCONE traffic > 400 Gb/s

- US-ATLAS testing
  - BNL hit levels **exceeding** targets
- Nearly Last-minute issues
  - Congestion to U-Florida
    - Moved to new 400G path
  - Packet loss to UT-Arlington
    - David Nichols @ LEARN found a degrading, but not yet failed optic
- Mitigation plans ready for certain regional congestion scenarios, but did not need to exercise them
- US-CMS testing
  - found strange 200G limit facing Fermilab (more later)

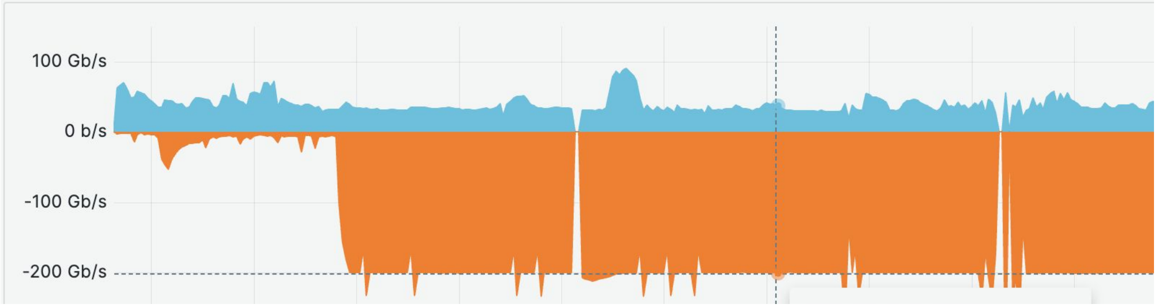
perfSONAR

- ESnet has a robust perfSonar deployment, but previously there were only a few LHCONE hosts in ESnet (down to just 1)
- We recently deployed interfaces on all current ESnet perfSonar servers in LHCONE as we wanted to avoid additional hardware deployments for LHCONE. We also automated the deployment.
- IPv6 only perfSonar interfaces in LHCONE
- Selected perfSonar hosts nearest to US Tier 1 and 2 sites
- LHCONE dashboard
- ATLAS mesh dashboard
- CMS mesh dashboard
- What other hosts (especially non-ESnet hosts) should be added to this?
- The pS deployment was helpful in testing leading up to DC24 (e.g. UT Arlington)

Fermilab issue

4x100G LAG for science traffic

(due to supply-chain woes)



- During US-CMS testing in December & retest in late January we saw a 200G bottleneck
  - on the LHCONE vlan *only.* 300G+ worked fine on LHCOPN
- ESnet led a 2-day intense test-a-thon to identify & fix.
  - repurposed Chicago area perfsonar nodes for bit blasting using multi-threaded iperf3
- Issue related to QoS differences between OSCARS (LHCOPN) and Best-Effort (LHCONE) traffic with respect to Nokia egress queue scheduling across multiple linecard LAG
- ***Temporarily Mitigated*** the worst of the impact by loosening the OPN bandwidth guarantee on egress to FNAL.
- Much thanks to:
  - Andrey & the FNAL network group
  - MULTI-THREADED iperf3!!!!

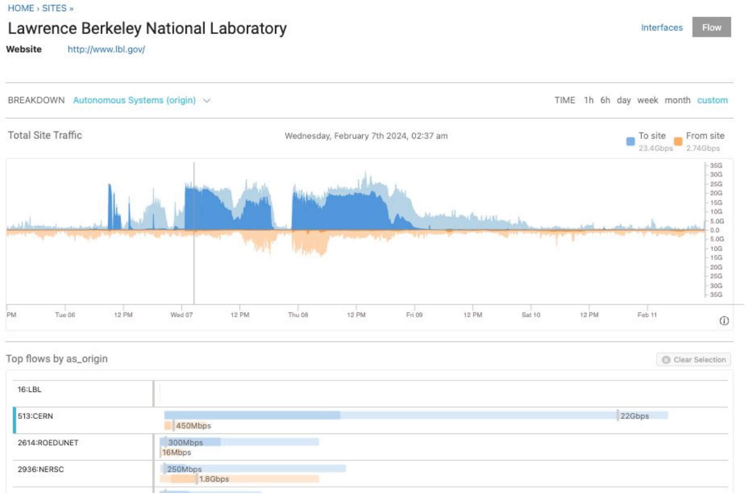
ESnet

Early surprise, BOST-AMST 100G loaded with ALICE traffic for LBNL

ESnet's LHC traffic tends to be dominated by ATLAS and CMS

ALICE usually doesn’t use the network in this way. Is this is permanent change?

Applied load-balancing to mitigate this shortest-path preference



Lawrence Berkeley National Laboratory

Website http://www.lbl.gov/

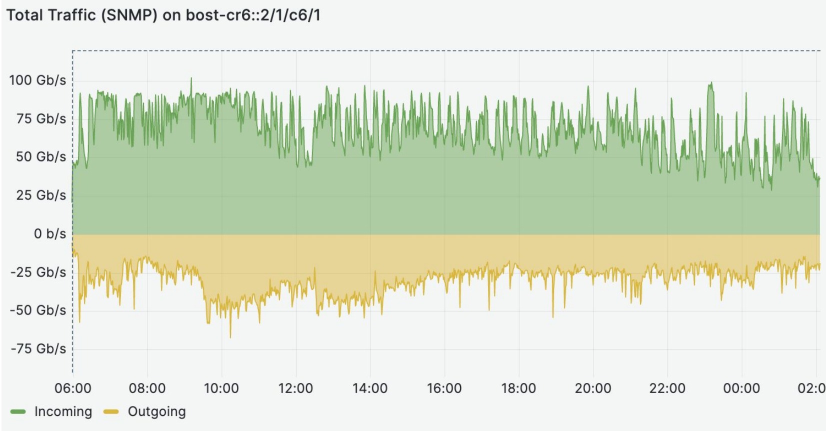
BREAKDOWN Autonomous Systems (origin)

Total Site Traffic

Wednesday, February 7th 2024, 02:37 am

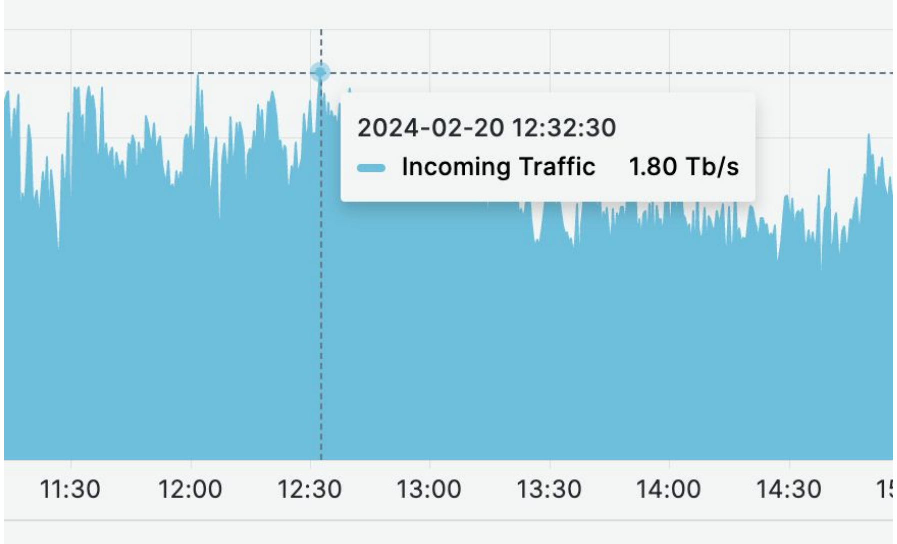
Top flows by as_origin

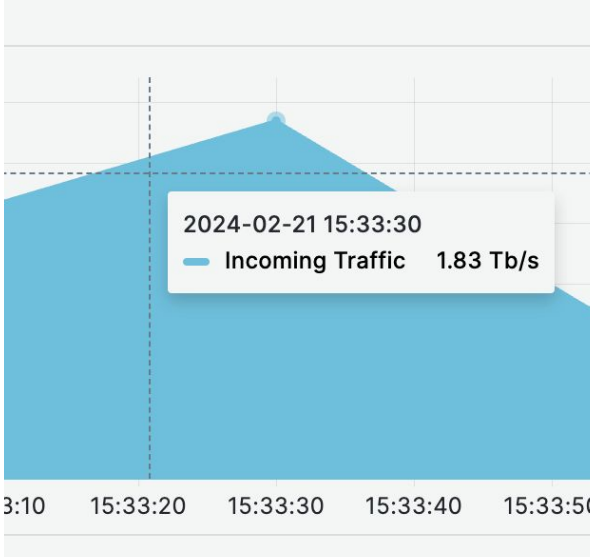
16:LBL		
513:CERN	22Gbps	450Mbps
2614:ROEDUNET	300Mbps	16Mbps
2936:NERSC	250Mbps	1.8Gbps



Total Traffic (SNMP) on bost-cr6::2/1/c6/1

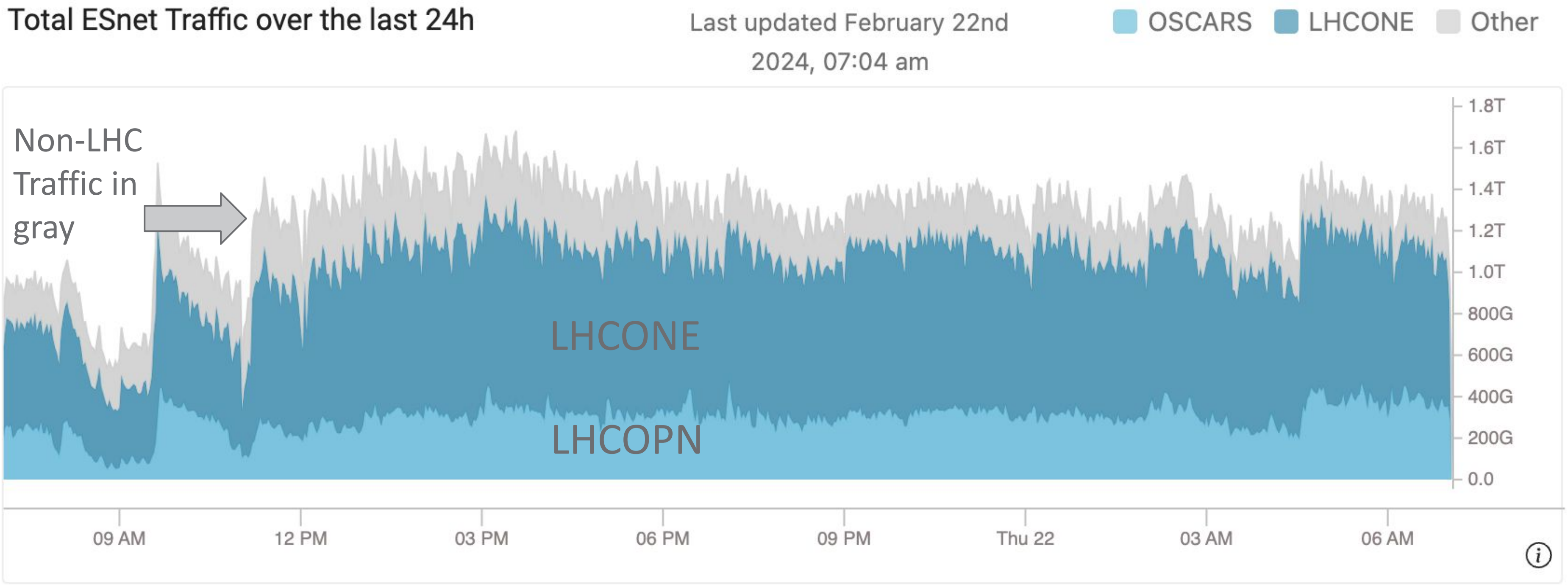
ESnet peak of 1.83 Tbit/sec offered load during DC24





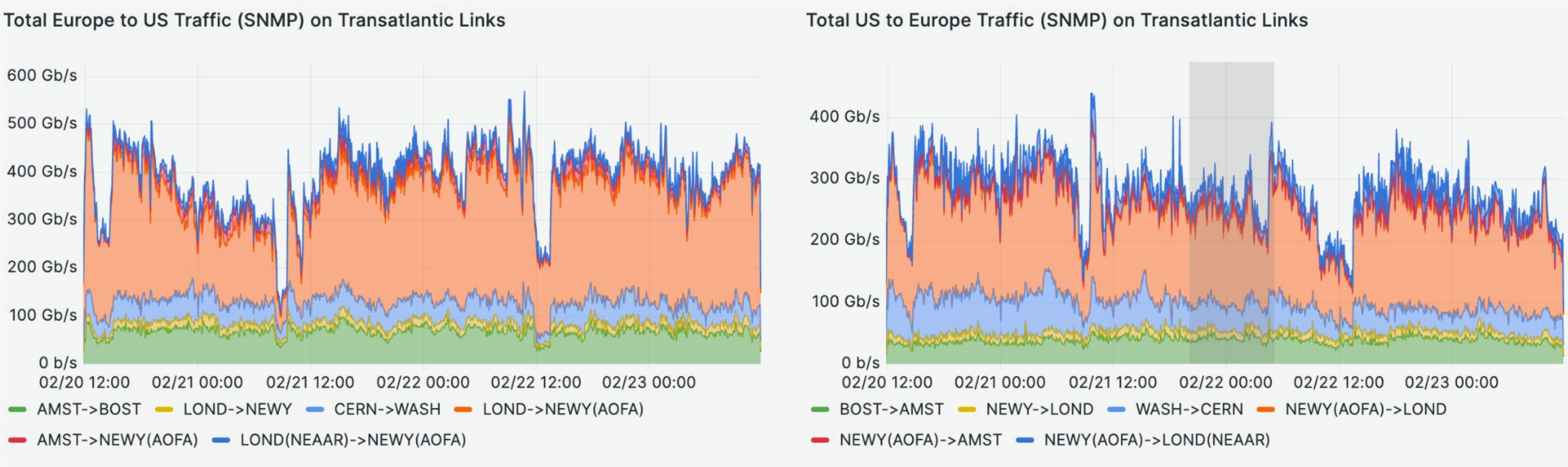
https://dashboard.stardust.es.net/d/Bi0-rzg4z/welcome?orgId=1&from=1708437425634&to=1708467355046

LHCOPN (OSCARS) & LHCONE DC24 Traffic vs Everything Else on ESnet


Total ESnet Traffic over the last 24h

Last updated February 22nd 2024, 07:04 am

ESnet Transatlantic Usage



Total Europe to US Traffic (SNMP) on Transatlantic Links

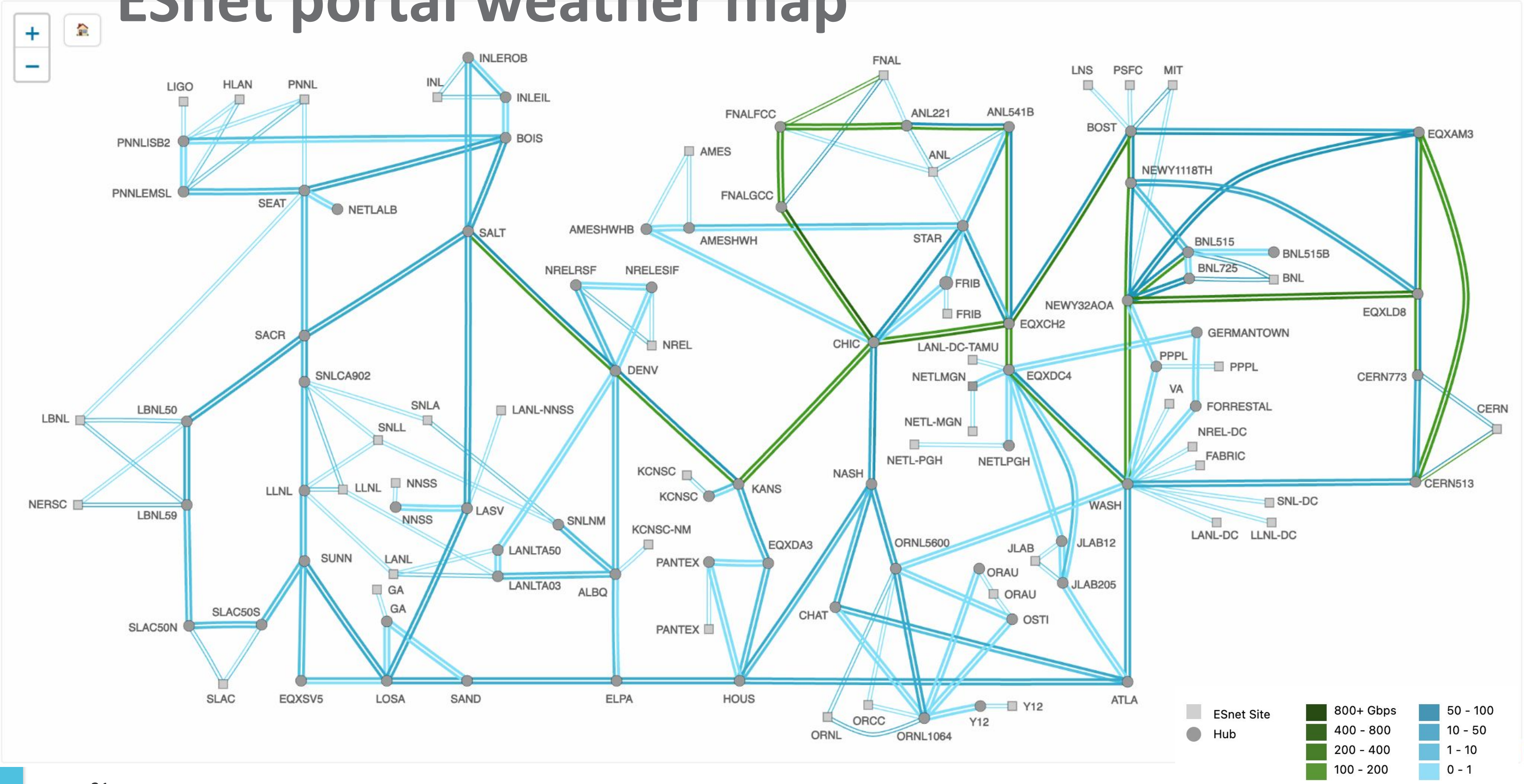
Total US to Europe Traffic (SNMP) on Transatlantic Links

LHCOPN Traffic placement + LHCONE weighted load-balancing

https://public.stardust.es.net/d/IkFCB5Hnk/lhc-data-challenge-overview?orgId=1&from=1708451852305&to=1708711052305

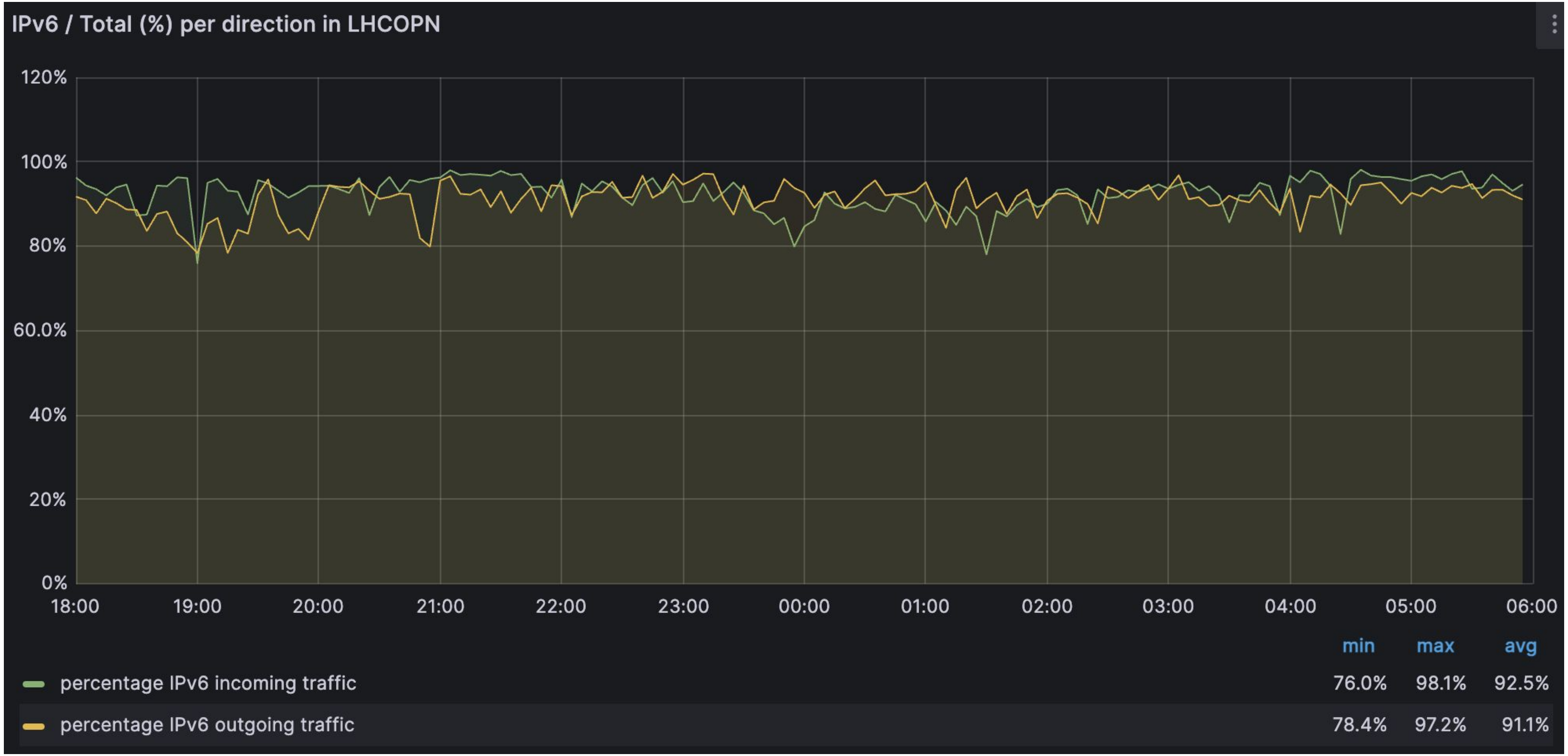
ESnet

ESnet portal weather map



■ ESnet Site	■ 800+ Gbps	■ 50 - 100
● Hub	■ 400 - 800	■ 10 - 50
	■ 200 - 400	■ 1 - 10
	■ 100 - 200	■ 0 - 1

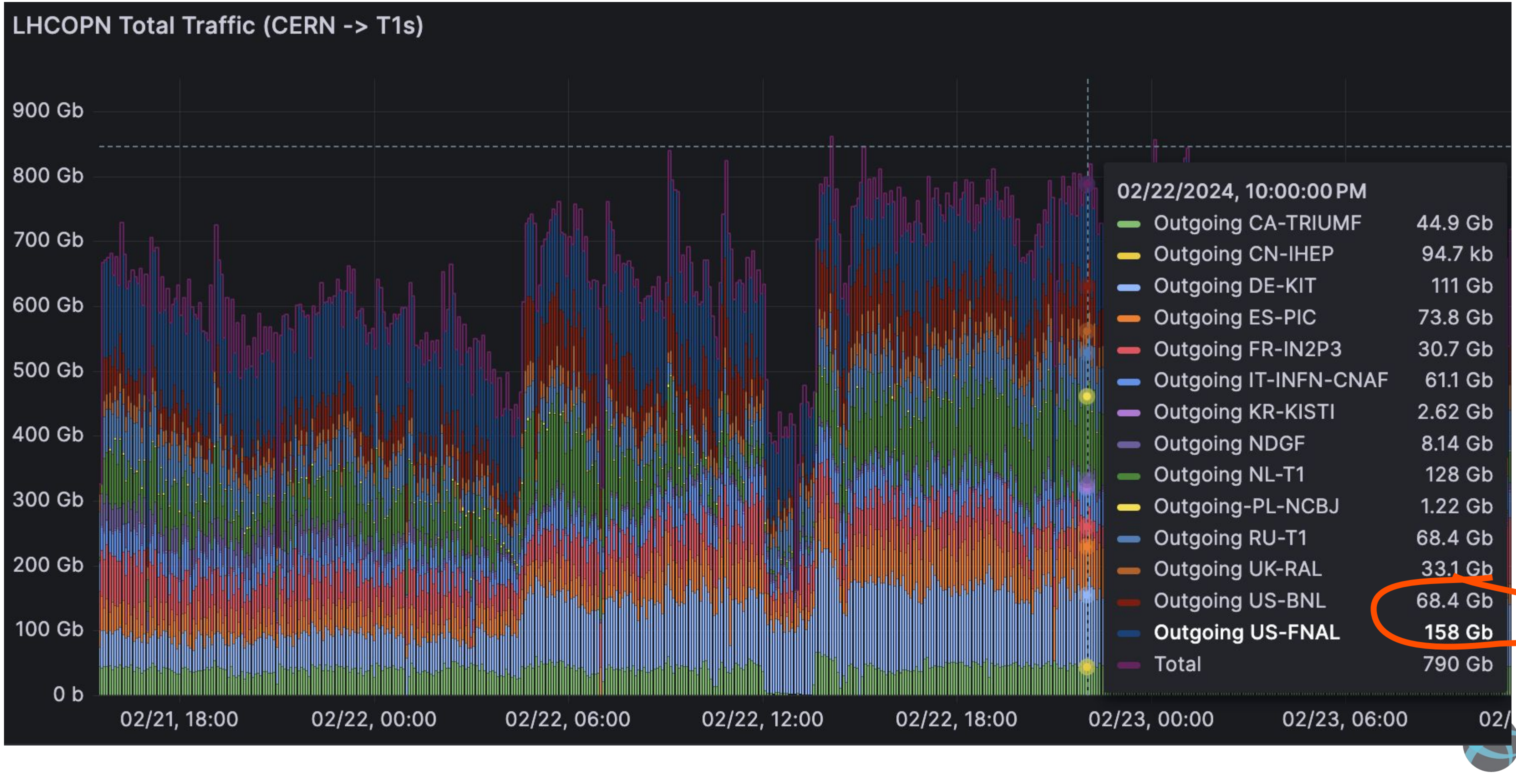
IPv4? Never heard of it.



	min	max	avg
🟩 percentage IPv6 incoming traffic	76.0%	98.1%	92.5%
🟨 percentage IPv6 outgoing traffic	78.4%	97.2%	91.1%

https://monit-grafana-open.cern.ch/d/cumEJJb4z/lhcopn-one-ipv6-vs-ipv4?orgId=16&from=1707436800000&to=1707480000000&var-source=raw&var-bin=5m&var-lhcopn_interfaces_ipv6=All&var-lhcopn_interfaces_ipv4=All

FNAL and BNL as major contributors to LHCOPN

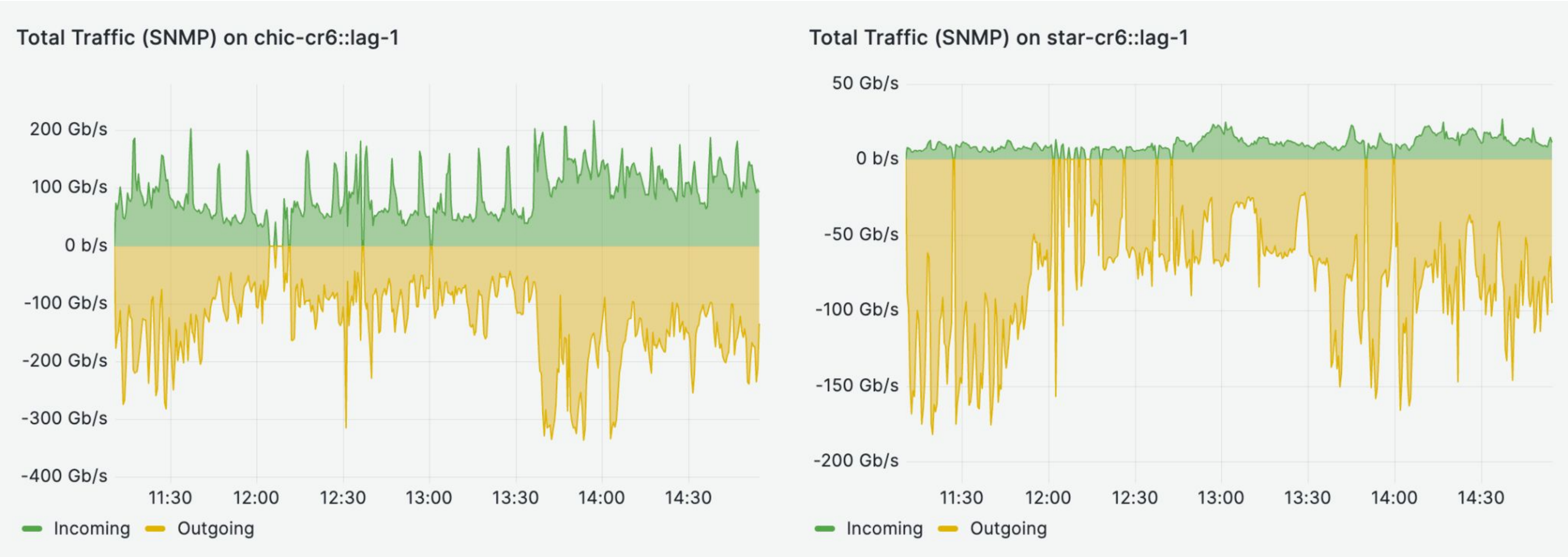


LHCOPN Total Traffic (CERN -> T1s)

02/22/2024, 10:00:00 PM

Outgoing CA-TRIUMF	44.9 Gb
Outgoing CN-IHEP	94.7 kb
Outgoing DE-KIT	111 Gb
Outgoing ES-PIC	73.8 Gb
Outgoing FR-IN2P3	30.7 Gb
Outgoing IT-INFN-CNAF	61.1 Gb
Outgoing KR-KISTI	2.62 Gb
Outgoing NDGF	8.14 Gb
Outgoing NL-T1	128 Gb
Outgoing-PL-NCBJ	1.22 Gb
Outgoing RU-T1	68.4 Gb
Outgoing UK-RAL	33.1 Gb
Outgoing US-BNL	68.4 Gb
Outgoing US-FNAL	**158 Gb**
Total	790 Gb

OmniPoP 2 x 4 x 100G

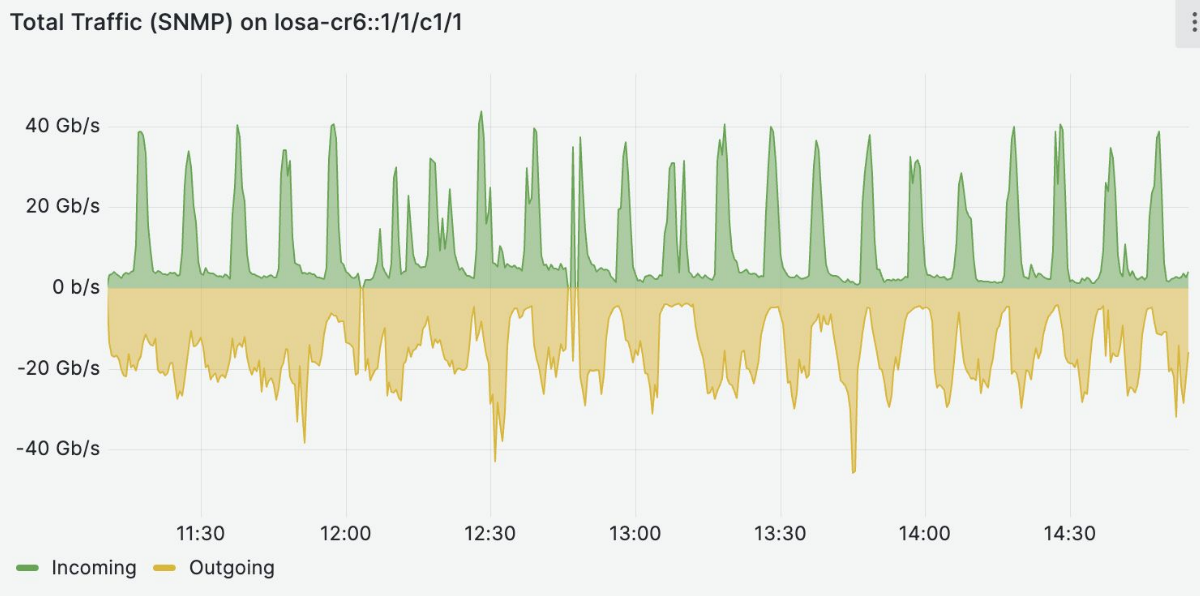

Total Traffic (SNMP) on chic-cr6::lag-1

Total Traffic (SNMP) on star-cr6::lag-1

https://public.stardust.es.net/d/b2c3a9c5-42e5-4b92-91eb-1e6f14a57fa8/lhc-data-challenge-regionals?orgId=1&from=1708621847254&to=1708635274282

Characteristic Tier-2 data rates during DC24

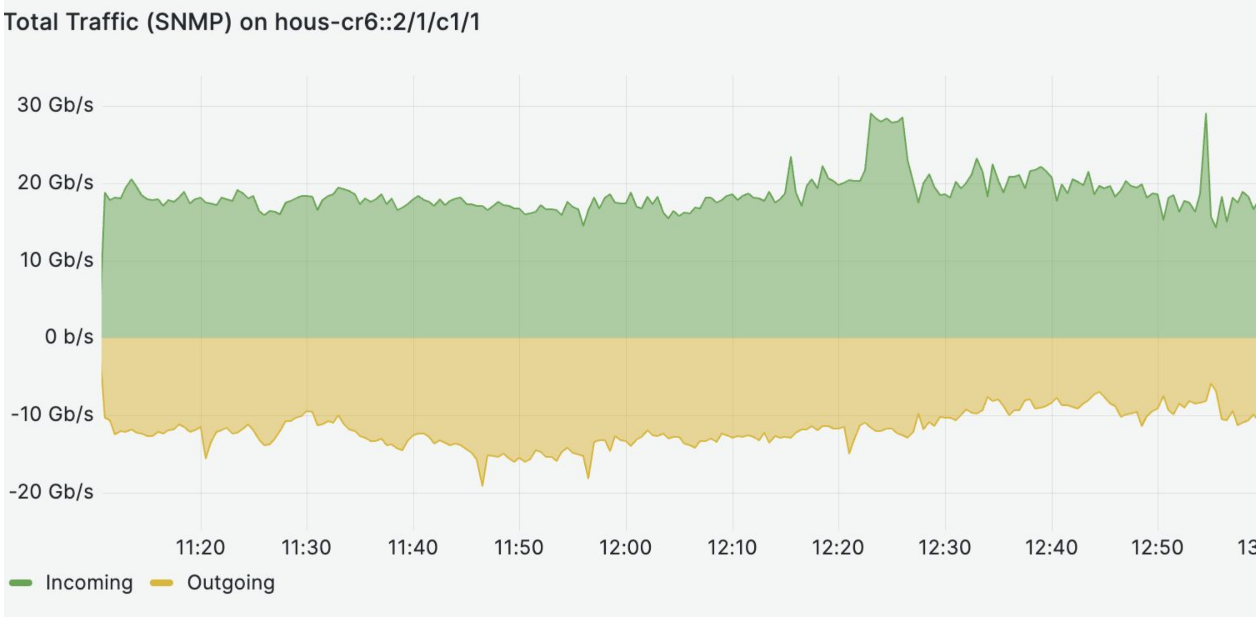
Bursts believed to be an artifact of storage system outpacing job submission



Caltech

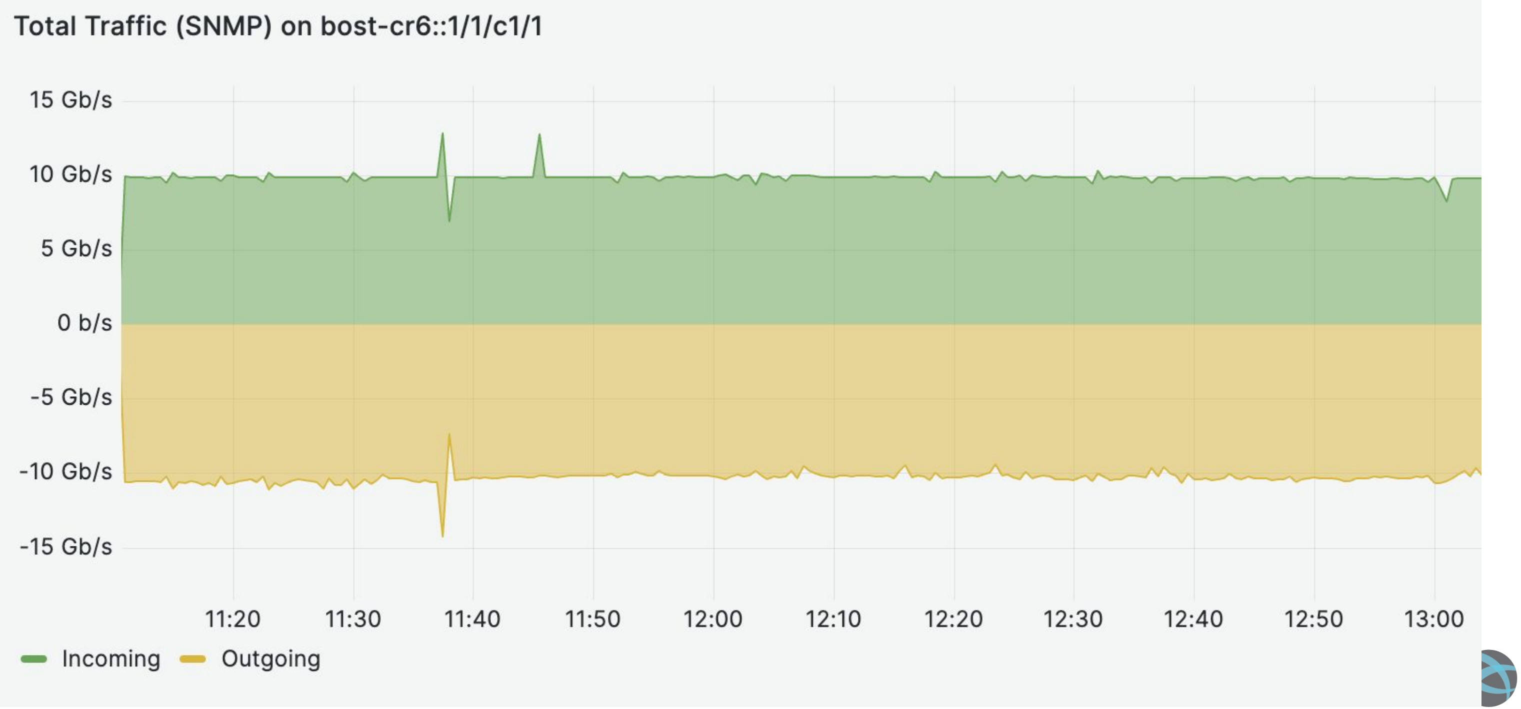
Characteristic Tier-2 data rates

Site known to have an internal storage limitation, no bursts



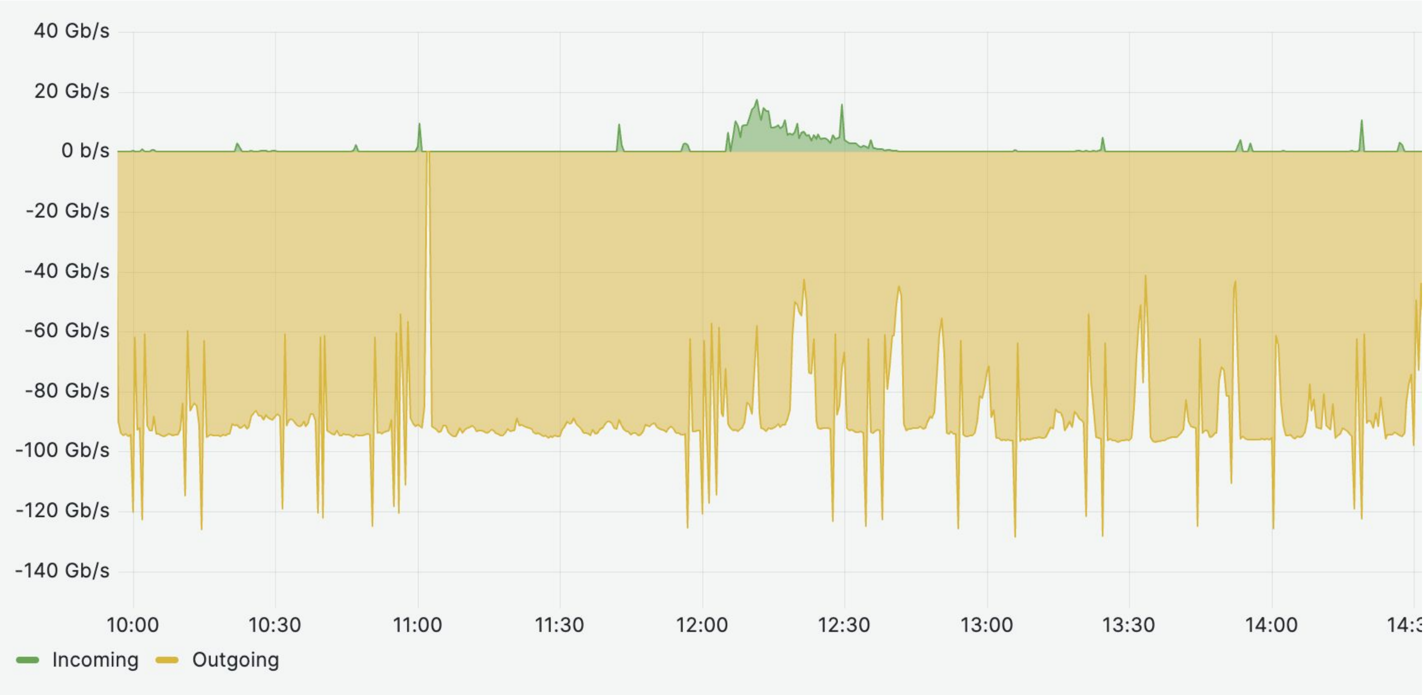
Characteristic Tier-2 data rates

Site known to have a 10G network bottleneck



Characteristic Tier-2 data rates

University of Florida, 100G network bottleneck

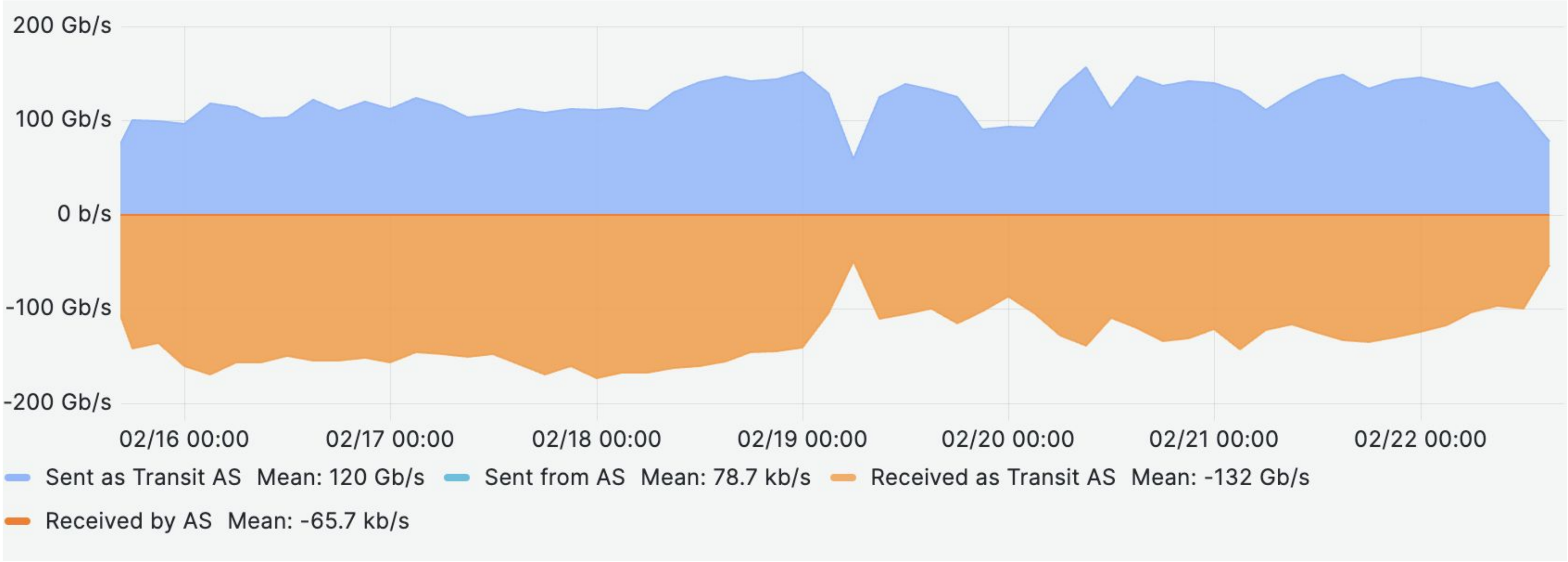


visible snmp polling errors under investigation

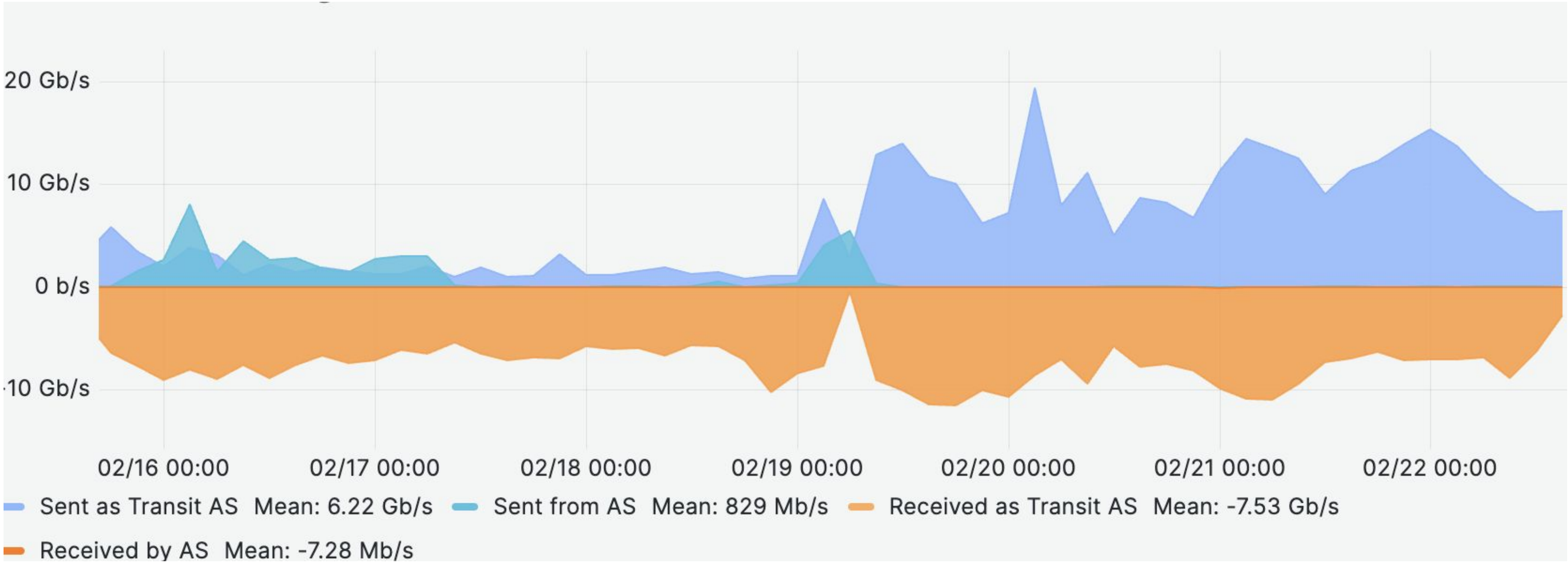
Notable LHCONE Peers

- Highlights from selected peerings follow

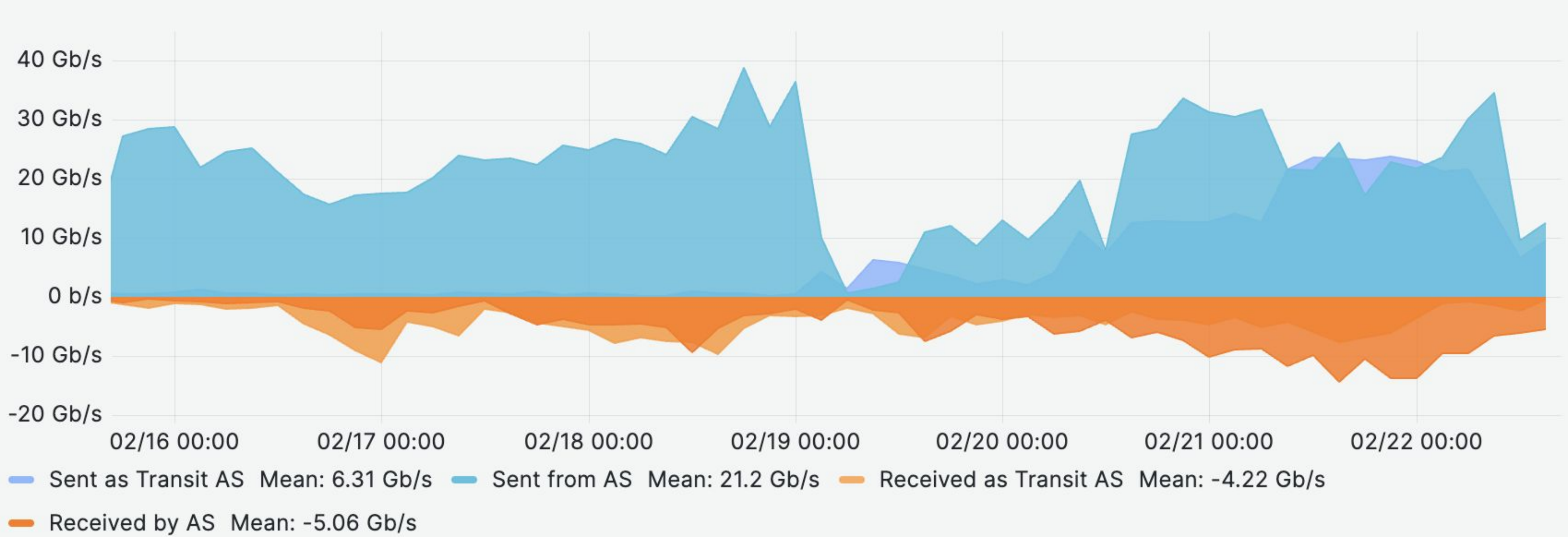
ESnet--GEANT Traffic



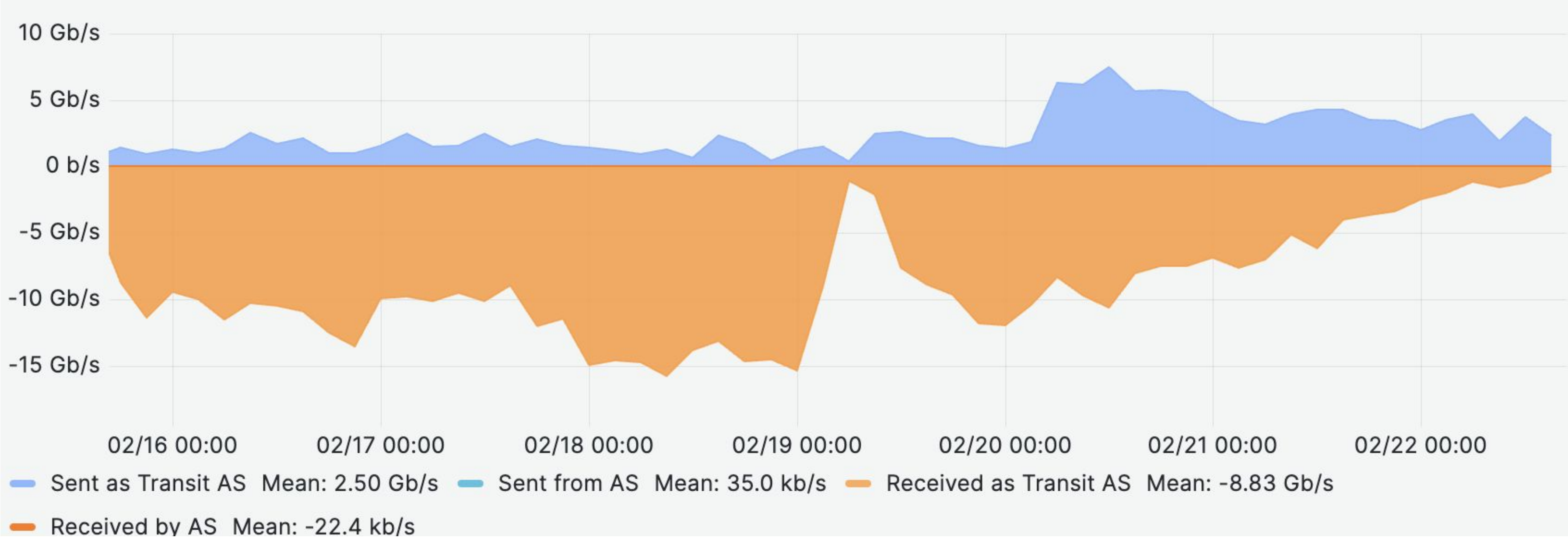
ESnet--SINET Traffic



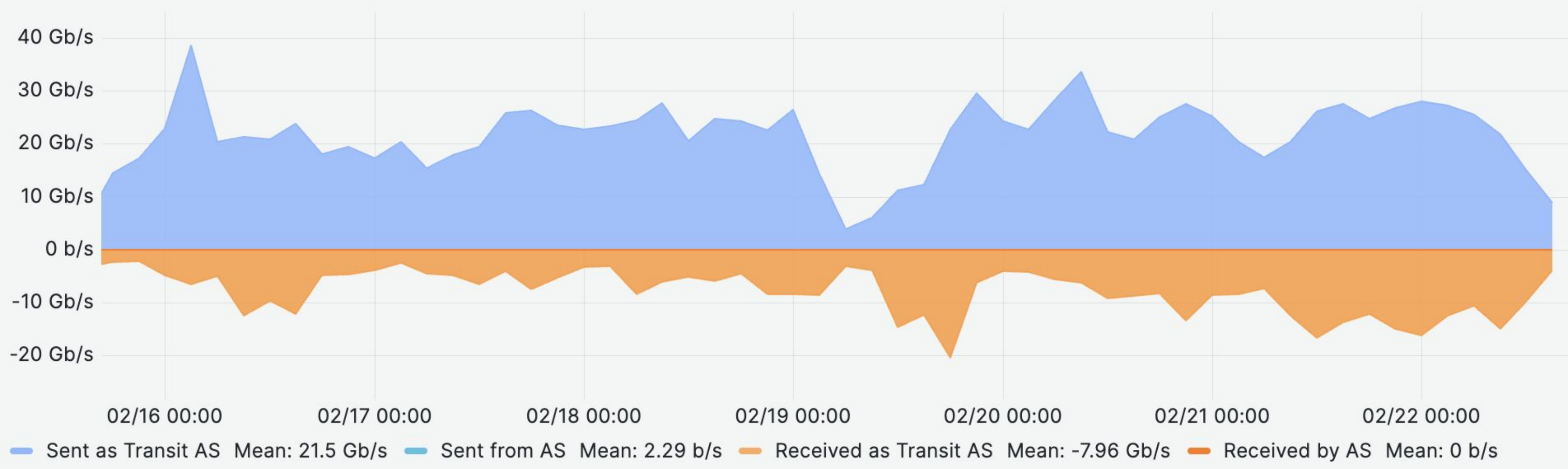
ESnet--SURFNET Traffic



ESnet--NORDUNET Traffic



ESnet--CANARIE Traffic



Data caching current use cases

- **US CMS: Southern California Petabyte Scale Cache**
  - Regional storage cache for US CMS user analysis at Caltech and UCSD
  - 23 federated XCache nodes: Approximately 2PB of total storage capacity
    - 12 nodes at UCSD: each with 24 TB, 10 Gbps
    - 9 nodes at Caltech: each with storage sizes ranging from 96TB to 388TB, 40 Gbps
    - 1 node at LBNL59 (ESnet): 44 TB storage, 40 Gbps
- **US CMS: Chicago**
  - Regional cache for U. Wisc Madison, Purdue and Notre Dame
  - 6 federated XCache nodes: Approximately 345 TB in total
    - 5 nodes at U. Wisc Madison: each with 35TB, 10Gbps
    - 1 node at CHIC (ESnet): 184TB on 100Gbps (LHCONE)
- **US CMS: Boston**
  - Regional cache for MIT
  - 1 node at BOST (ESnet): 300TB on 100Gbps (LHCONE)
- **OSG/OSDF: London and Amsterdam**
  - Mainly for DUNE and LIGO for transatlantic traffic from the US
  - Each 300TB, 100Gbps

Data caching future plans

- **Immediate (2024-2025)**
  - Pilot project with US ATLAS
    - Testing multi-service platform on BOST node with DTNasS containers, targeting a BNL's VP
  - Continue discussion and possible pilot work with DUNE
  - Comparing characteristics from different regional caching nodes
  - Additional pilot node deployment
    - CMS and OSDF, multi-service case, at Atlanta
    - Upgrade storage capacity at CHIC and BOST
- **Near-term (2025-2026)**
  - Possible caching service expansion for LHC/OSDF experiments
  - Longer term traffic projection
  - Making conclusions on the caching pilot
- **Beyond (2026-)**
  - Monitoring testbed
  - Prediction on data distribution to connect to the traffic engineering
    - Possible connection to the high touch data
  - Storage-integrated networking testbed
    - Edge storage, longer term storage for pre-staged data with some (simple) computing power
    - Metadata management testbed

Conclusions

- No ESnet congestion during DC24 despite multiple unexpected issues
- Make sure to get your voice heard for the High Energy Physics Requirements Review in 2025
  - Link to HEP 2020 RR https://doi.org/10.2172/1804717
- Starting US CMS/ATLAS Tier 1 & 2 outreach to prepare for DC26
- Measurement was so useful in spotting and addressing issues.
  - ESnet portal
  - CERN
  - WLCG

Thanks!

katerobinson@es.net

Energy Sciences Network (ESnet)

Lawrence Berkeley National Laboratory

http://my.es.net/

http://www.es.net/

http://fasterdata.es.net/

ESnet