U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

ESnet
ENERGY SCIENCES NETWORK

ESnet Provider Update

Kate Robinson, Paul Wefel, Eli Dart
Network Engineers, ESnet

2025 SA3CC Meeting
La Serena, Chile
7 May 2025

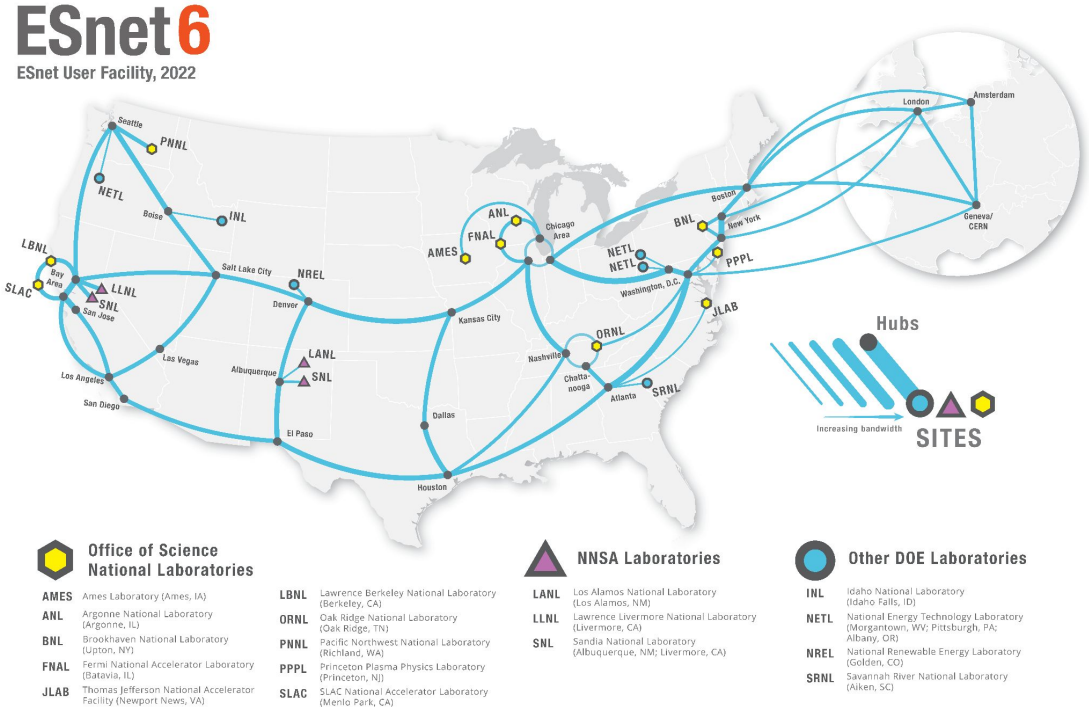
Agenda

- ESnet overview
- Requirements Review update
- Network Upgrades
- TransAtlantic upgrades and strategy
- Measurement tools
- Troubleshooting SLAC to GCP

ESnet

ESnet is DOE'S data circulatory system…

- ESnet supports the DOE scientific research ecosystem.
- Interconnects all National Labs and User Facilities
- Provides reliable, high-performance connectivity to the global research collaborations, Cloud, and the larger Internet

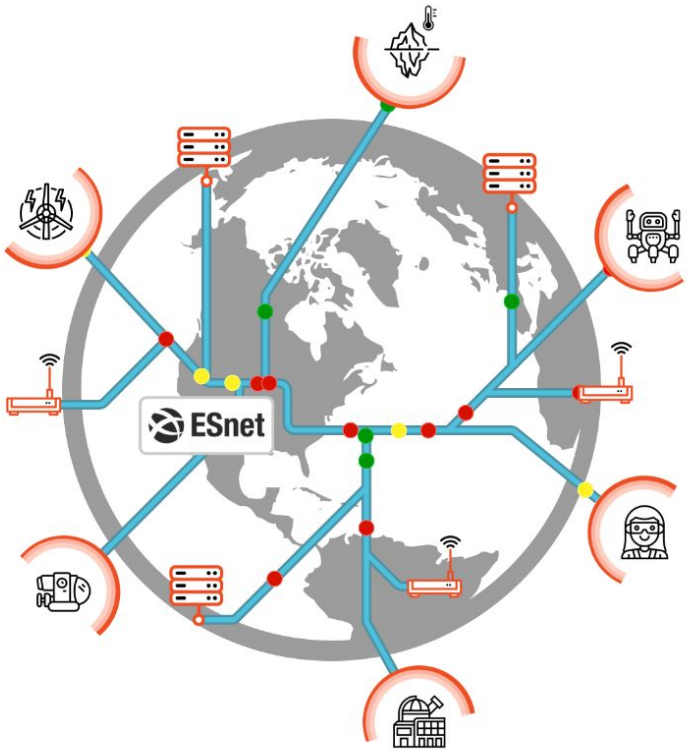


ESnet

…and the stage for a global science laboratory.

ESnet's Vision: Scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data.

ESnet's Mission: Networking that accelerates science.



Science requirements drive design and upgrades

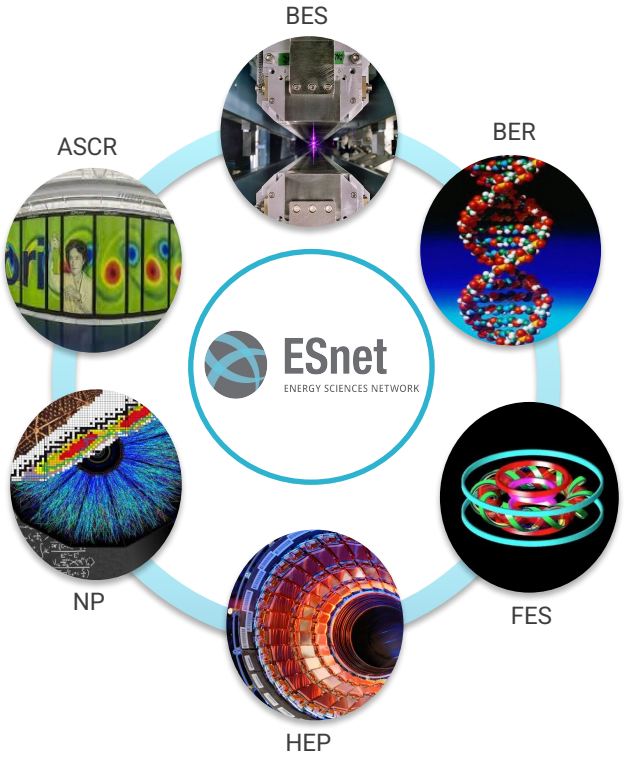
ESnet's Requirements Review program builds knowledge, understanding, and collaboration with DOE/SC programs and facilities allowing the larger ecosystem to learn from each other.

We capture:

- Major science experiments and facilities, both in operation and planned.
- The process of science used for knowledge discovery
- The volume of data produced now, and anticipated in the future, and where the data must be shared, computed and/or stored.
- The current technology capabilities (network, computation, storage, and software stack) plus planned upgrades, additions or improvements.

This dialogue provides the foundation for ESnet services and strategy.

ESnet's Requirements Review reports have become a strategically essential set of documents for the Office of Science.



ESnet

Requirements Review Overview

ESnet's core partnership program created to comprehensively evaluate:

- ***Major science experiments and facilities,*** both in operation and planned.
- ***The process of science*** used for knowledge discovery, and including scientists' interactions with the instruments and facilities.
- ***The volume of data*** produced now, and anticipated in the future, with an emphasis on geographical location of where the data must be shared, computed and/or stored.
- ***The current technology capabilities*** (network, computation, storage, and software stack) used by each science collaboration/facility as well as any planned upgrades, additions or improvements.

Scientific Case Studies (14 Total, about 350pg)

- Cosmological Simulation Research
- Dark Energy Science Collaboration (DESC)
- Dark Energy Spectroscopic Instrument (DESI)
- The Vera C. Rubin Observatory (Rubin Observatory) & the Legacy Survey of Space and Time (LSST)
- Cosmic Microwave Background - Stage 4 (CMB-S4)
- Alpha Magnetic Spectrometer (AMS) Experiment
- Muon Experimentation at Fermilab
  - Muon G minus two (g-2)
  - Muon-to-electron-conversion experiment (Mu2e)
- Belle II Experiment
- Neutrino Experiments at Fermilab
  - Short-Baseline Neutrino Program (SBN)
  - The Deep Underground Neutrino Experiment (DUNE)
- Super Cryogenic Dark Matter Search (Super CDMS)
- Large Hadron Collider (LHC) Experimentation and Operation
  - ATLAS (A Toroidal LHC ApparatuS) Experiment
  - Compact Muon Solenoid (CMS) Experiment
  - LHC Operations
  - High Luminosity (HL) Era of the LHC

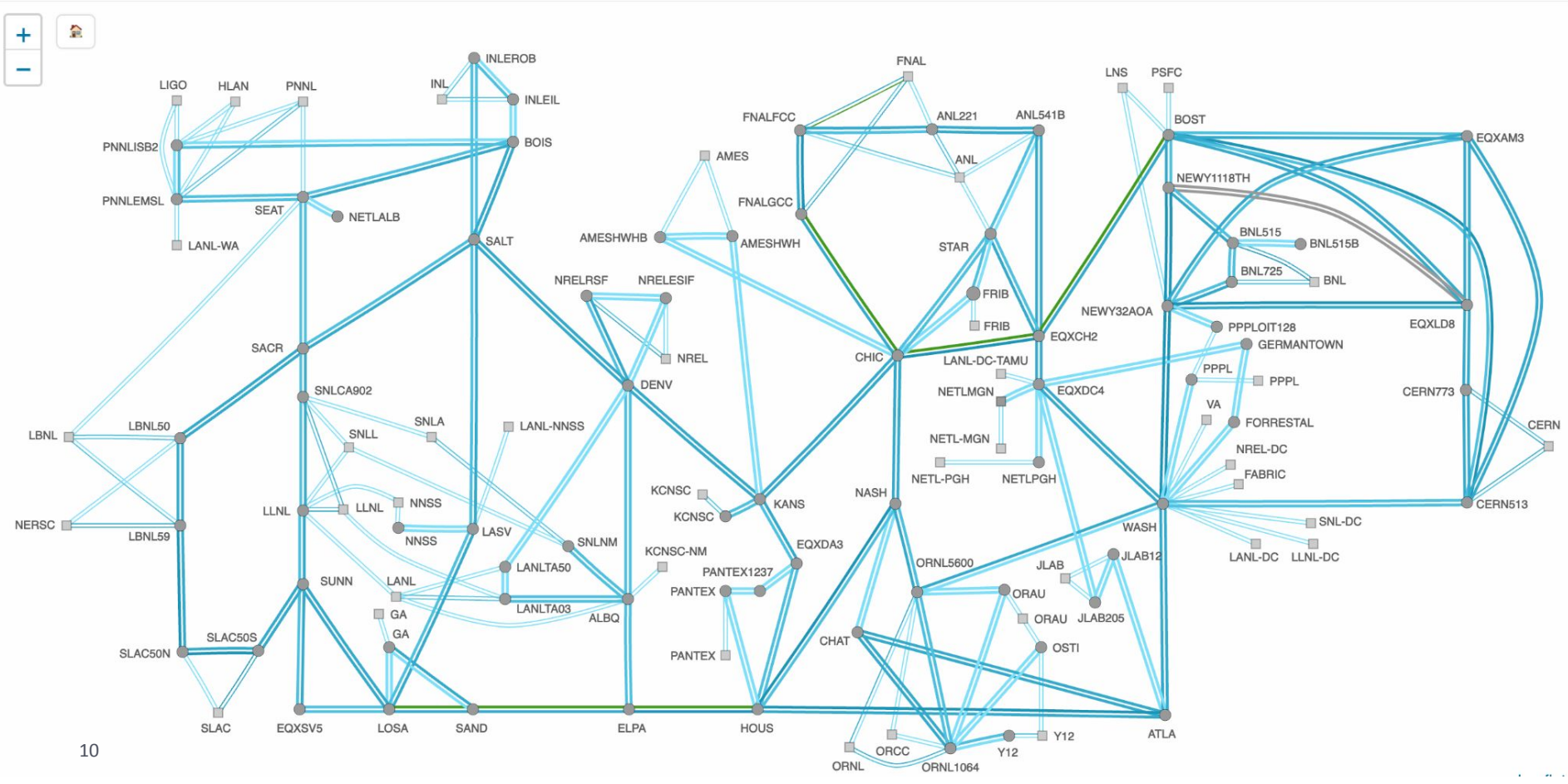
Summit / LSSTCam	Yes – primary initial copy, retained for ~30 days	Quick-look capabilities for observing staff	30 TB / night	Continuous transfer during the night to SLAC, after shutter closes on each exposure	No	Per data security requirements, data must be transferred over IPSec encrypted tunnel to SLAC
USDF (located at SLAC)	Yes – primary archival copy	Cloud-based portal for all users backed by on-prem data holdings; real-time broadcast of alert packets to selected brokers	5 PB/yr. With ancillary and other image data products, up to 500 PB by end of survey in 2035	Nightly (images and metadata northbound) and Alerts to brokers each minutes	Processed data sent to data center in Chile; calibration and QA data sent to Summit	Large number of small files may impact network transfer performance
Chile	Yes	Data Portal for Chilean users	5 PB/yr	Nightly raw images stored. Weekly/ monthly calibration data. Annual data-release data products	No	No
Brazil	No	None, data only transits	5 PB/yr	Same as Chile (same data transits Brazil)	Same as USDF (same data transits Brazil)	No
CC-IN2P3 / FrDF	Yes – secondary copy	Portal (possibly), default is USDF	5 PB/yr	Nightly raw images sent at low or modest rate after embargo period, plus additional data as part of distributed annual data-release processing	40% of annual data-release products generated here and returned to USDF	No
IRIS / UKDF	No	Full IDAC user interface	few PB/yr	Data transferred from USDF to UKDF to enable (a) 25% of annual data-release processing and (b) full IDAC for annual data-release	25% of annual data-release products generated here and returned to USDF	UK serves as both processing partner and full IDAC

ESnet

Breaking news: (rough) 2024 Findings

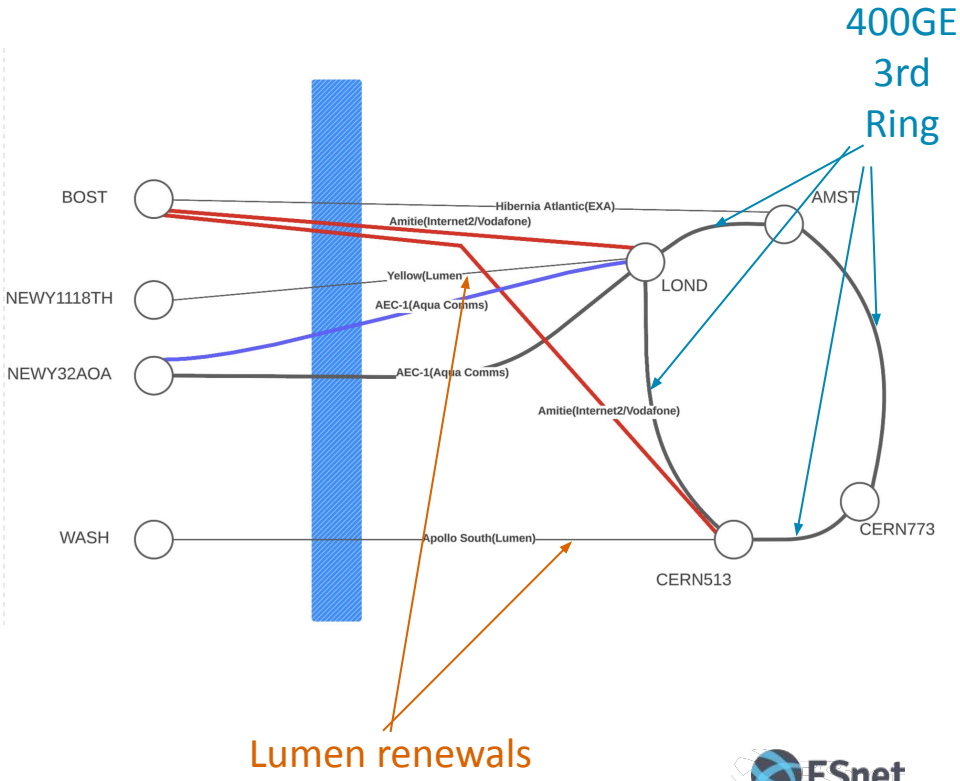
- ***Data volume Increases***:
  - Across the board - PB era is now common for the large experiments. Small have graduated to TB scale
  - Frequency of production increasing, along with fidelity
- ***Ability to support multi-facility workflows***:
  - Sensor, Computation, Storage, and People may all be in different locations - regularly.
  - Focus in DOE space: Integrated Research Infrastructure (https://iri.science)
  - Ability to execute 'near real-time' workflows is now routine (thanks R&E Networking Community!)
  - This is the only way we will scale - but requires significant coordination and cooperation
- ***Cloud use***:
  - Would still consider this the 'dabbling' stage for the majority. A significant number of workflows are now 'cloud-ready' and can burst there if the costs allow it (hint: they still do not)
  - Some experiments have built successful hybrid model (Rubin): cloud 'front end' to support the users, backed against DOE-based storage and production/analysis computation.
- ***Data Challenges = Good Thing***

ESnet Backbone



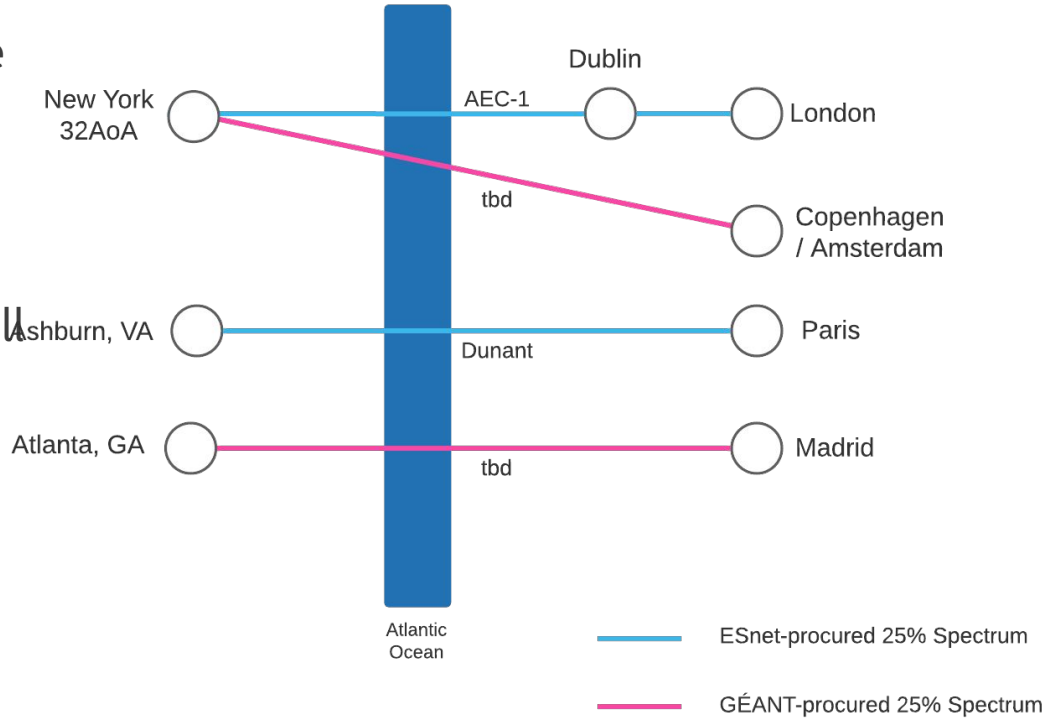
Transatlantic Updates

- Currently in Production
  - 400G + 100G New York to London
  - 100G Boston to Amsterdam
  - 100G Washington to CERN
  - 400G Boston to CERN
  - 400G Europe Ring
    - working with GEANT
    - Next phase will be 3x400G
  - 400G Boston to London
- AEC-1 Spectrum WL6 Upgrade
- Lumen 2x100GE
  - Term ends 10/2024
  - Migrate Yellow to 32AOFA
  - Quoting 400G WASH-CERN
- EU 3rd 400GE Ring
  - GÉANT preparing for service delivery



Future US - Europe Connectivity Plans

- Collaborating with GEANT to share spectrum on subsea
- Plan: collectively acquire optical spectrum across 4+ diverse cables
- ESnet-targeted cables are fixed; still some variability in GEANT plans
- Depending on GEANT's spectrum procurement timelines, we may investigate additional 400GE lit services for interim diversity



Monitoring and Analysis

Stardust

Network Measurement and Analysis for ESnet

Extensible / Open Architecture

NSF NetSAGE project derived

Leverage Open Source components where we can, and innovate where it makes a difference.

Multiple access methods

Dashboards, Indexed APIs and “Raw”

Today, we are focused on users creating and sharing visual dashboards.

In the future, we expect direct programmatic access will become increasingly common for ML work and external collaboration.

Multi Datasource

Extensible and Open

Traffic Accounting, Link and Resource Use, Performance Testing Results, Others not yet invented.

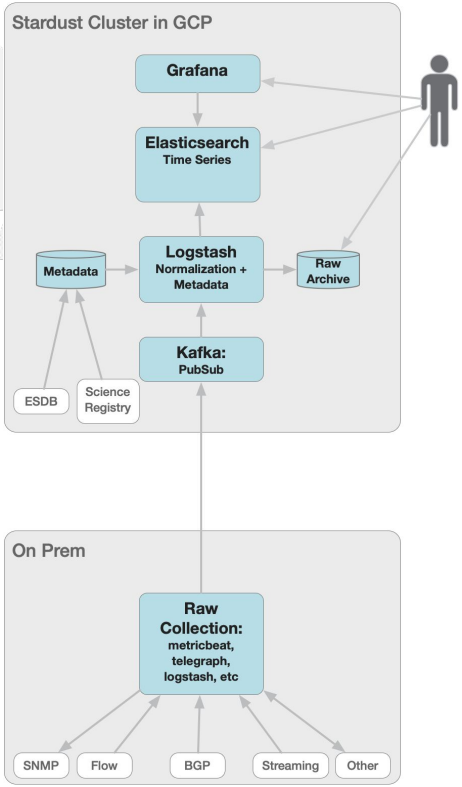
Each has its own set of measurements to which we add a common core set of metadata.

Flexible aggregation

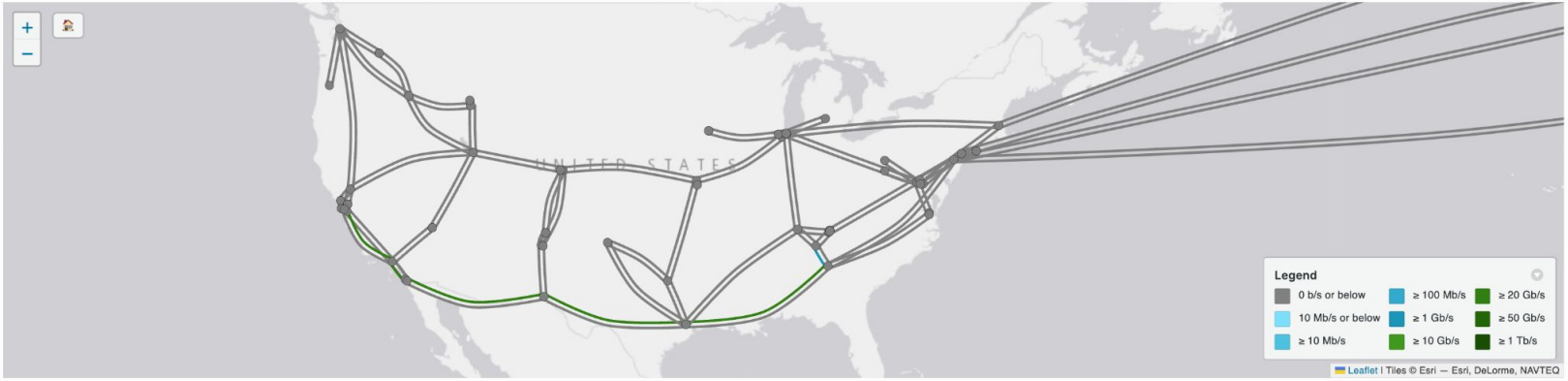
Time Frame and Relationships

The metadata we add to measurements is used to summarize data to tell stories, and having common metadata allows us to use multiple data sets in a story.

- How are researchers moving science data and how has that changed over the last 3 years?
- What just caused that huge spike in traffic on the links to europe in the last 15 minutes and is that likely impacting data transfers?



ESnet



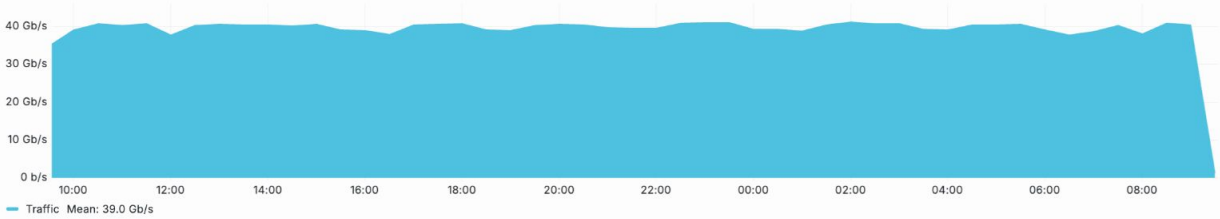
Bytes Transferred in 1d

429 TB

# of Flows in 1d

97.9 K

Traffic Rate Over Time



Ingress Interface Receiving By AS (Flow)

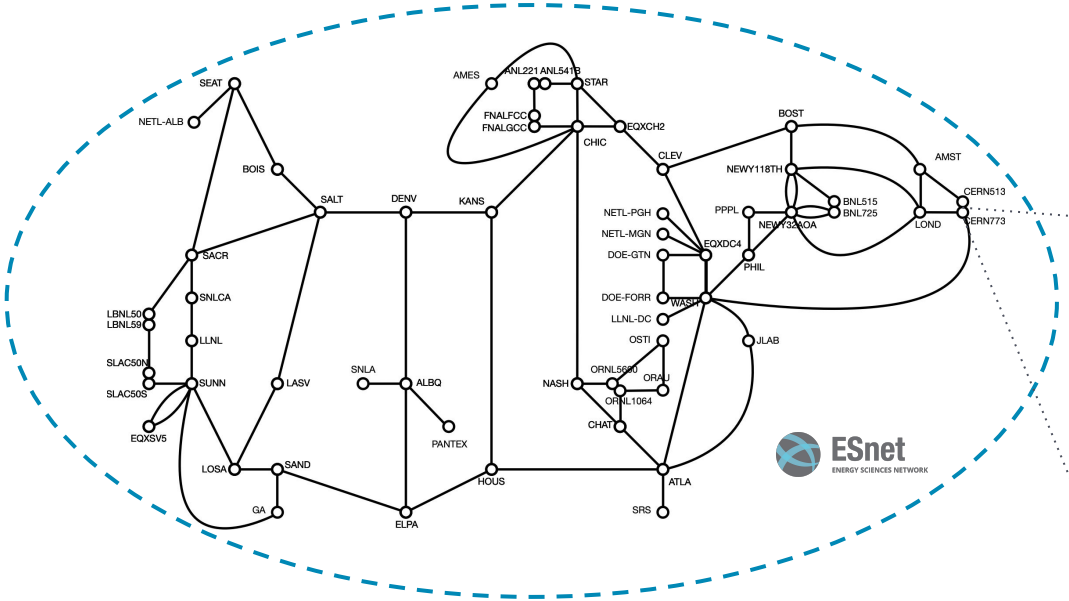
Ingress Interface	Rate	Volume
atla-cr6::lsst_se-1625	39.2 Gb/s	423.7 TB
atla-cr6::atla-ps-hc-lsst-thru	97.7 Mb/s	1.1 TB
slac50n-cr6::slac_se-849	88.6 Mb/s	956.9 GB
chat-cr6::chat-ps-hc-lsst-t...	85.7 Mb/s	926.0 GB
slac50n-cr6::slac50n-ps-h...	83.7 Mb/s	904.0 GB
slac50s-cr6::slac50s-ps-h...	72.8 Mb/s	786.3 GB

Top Interfaces by Incoming Traffic (SNMP)

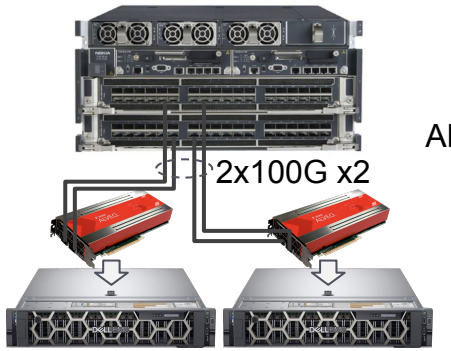
Interface	Volume
atla-cr6::lsst_se-1625	434.0 TB
atla-cr6::atla-ps-hc-lsst-thru	1.1 TB
slac50n-cr6::slac_se-849	970.4 GB
chat-cr6::chat-ps-hc-lsst-thru	929.7 GB
slac50n-cr6::slac50n-ps-hc-lsst-thru	906.2 GB
slac50s-cr6::slac50s-ps-hc-lsst-thru	789.6 GB

ESnet

High-Touch: High Precision Network Measurements



High-Touch Server Hardware Deployment



ESnet6 Core Router

AMD Xilinx Alveo U280 FPGA (1 per server)

FPGA Servers (2 per hub site)

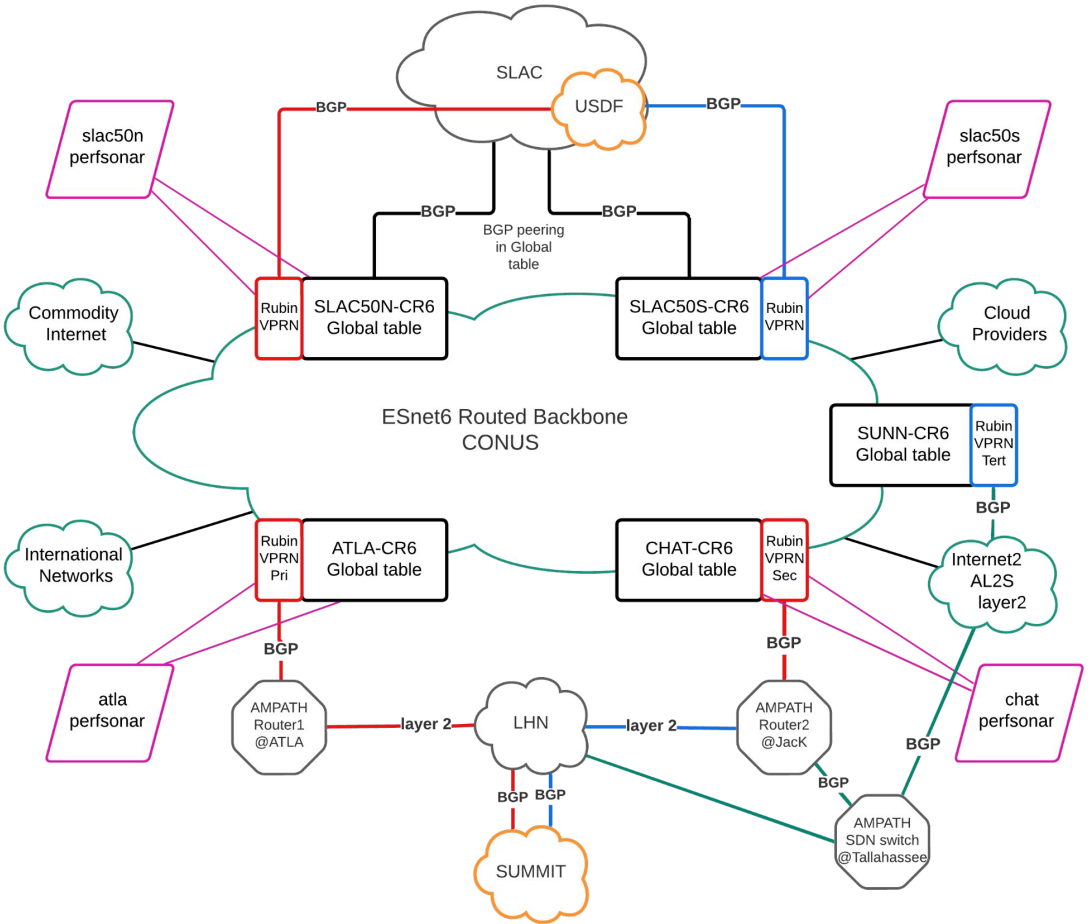
- 42 deployment locations
- Near 100% perimeter coverage*
- Router packet mirroring allows 100% packet inspection

*NB: Certain ports are omitted due to security sensitivities.

vlan_id	any(sap_routing_instance)	exporting_node	ip_src	ip_dst	tot_packets	tot_retrans	tot_loss	direction
0	IPT	eqxsv5-ht2	106.8.139.229	134.79.138.132	4912	1	3	in
1134	IPT	slac50s-ht1	106.8.139.229	134.79.138.132	4919	1	3	out
316	LSST	atla-ht1	139.229.137.20	134.79.23.9	250538	17575	17377	in
314	LSST	slac50s-ht1	139.229.137.20	134.79.23.9	250520	17575	17377	out
316	LSST	atla-ht1	139.229.137.21	134.79.23.9	829065	76840	76188	in
314	LSST	slac50s-ht1	139.229.137.21	134.79.23.9	830253	77295	76639	out
316	LSST	atla-ht1	139.229.137.22	134.79.23.9	237474	16524	16394	in
314	LSST	slac50s-ht1	139.229.137.22	134.79.23.9	237447	16524	16394	out
316	LSST	atla-ht1	139.229.137.23	134.79.23.9	1509943	41887	38068	in
314	LSST	slac50s-ht1	139.229.137.23	134.79.23.9	1509661	41887	38068	out
316	LSST	atla-ht1	139.229.137.24	134.79.23.9	10416	305	305	in
314	LSST	slac50s-ht1	139.229.137.24	134.79.23.9	10389	305	305	out
316	LSST	atla-ht1	139.229.137.25	134.79.23.9	123669	931	827	in
314	LSST	slac50s-ht1	139.229.137.25	134.79.23.9	123653	931	827	out
316	LSST	atla-ht1	139.229.140.13	134.79.235.226	1508663	30	8	in
313	LSST	slac50n-ht1	139.229.140.13	134.79.235.226	1508666	25	8	out
316	LSST	atla-ht1	139.229.140.135	134.79.23.9	2575	0	0	in
314	LSST	slac50s-ht1	139.229.140.135	134.79.23.9	2571	0	0	out
316	LSST	atla-ht1	139.229.140.135	134.79.235.226	1364	51	9	in
313	LSST	slac50n-ht1	139.229.140.135	134.79.235.226	1393	53	9	out
316	LSST	atla-ht1	139.229.140.137	134.79.235.226	1999432	64	1	in
313	LSST	slac50n-ht1	139.229.140.137	134.79.235.226	1999445	52	0	out
316	LSST	atla-ht1	139.229.153.204	134.79.23.240	3912742	4685	4144	in
314	LSST	slac50s-ht1	139.229.153.204	134.79.23.240	3912742	4685	4144	out
316	LSST	atla-ht1	139.229.153.205	134.79.23.240	3932814	5140	4540	in
314	LSST	slac50s-ht1	139.229.153.205	134.79.23.240	3932814	5140	4540	out
316	LSST	atla-ht1	139.229.153.239	134.79.23.240	3919270	5065	4477	in
314	LSST	slac50s-ht1	139.229.153.239	134.79.23.240	3919270	5065	4477	out
316	LSST	atla-ht1	139.229.153.240	134.79.23.240	3904322	5323	4708	in
314	LSST	slac50s-ht1	139.229.153.240	134.79.23.240	3904322	5323	4708	out
316	LSST	atla-ht1	139.229.153.242	134.79.23.240	3919965	5510	4859	in
314	LSST	slac50s-ht1	139.229.153.242	134.79.23.240	3919965	5510	4859	out
316	LSST	atla-ht1	139.229.153.247	134.79.23.240	3889732	4697	4156	in
314	LSST	slac50s-ht1	139.229.153.247	134.79.23.240	3889732	4697	4156	out
316	LSST	atla-ht1	139.229.153.248	134.79.23.240	3976364	5187	4552	in
314	LSST	slac50s-ht1	139.229.153.248	134.79.23.240	3976364	5187	4552	out
316	LSST	atla-ht1	139.229.153.249	134.79.23.240	599070	535	440	in
314	LSST	slac50s-ht1	139.229.153.249	134.79.23.240	599070	535	440	out
316	LSST	atla-ht1	139.229.175.76	134.79.23.9	555090	0	0	in
314	LSST	slac50s-ht1	139.229.175.76	134.79.23.9	554740	0	0	out
316	LSST	atla-ht1	139.229.180.90	134.79.23.9	12772802	575	412	in
314	LSST	slac50s-ht1	139.229.180.90	134.79.23.9	12771664	573	413	out
316	LSST	atla-ht1	139.229.180.91	134.79.23.9	11842	1	1	in
314	LSST	slac50s-ht1	139.229.180.91	134.79.23.9	11839	1	1	out
316	LSST	atla-ht1	139.229.180.92	134.79.23.9	101697296	1113498	1075621	in
314	LSST	slac50s-ht1	139.229.180.92	134.79.23.9	101657118	1113068	1075198	out
316	LSST	atla-ht1	139.229.180.93	134.79.23.9	88930566	980998	950049	in
314	LSST	slac50s-ht1	139.229.180.93	134.79.23.9	88892320	980657	949679	out
316	LSST	atla-ht1	139.229.180.94	134.79.23.9	106729818	1309422	1273145	in
314	LSST	slac50s-ht1	139.229.180.94	134.79.23.9	106689674	1308971	1272697	out
316	LSST	atla-ht1	139.229.181.20	134.79.23.9	22957348	24590	18274	in
314	LSST	slac50s-ht1	139.229.181.20	134.79.23.9	22682593	24457	18177	out
99	IPT	eqxsv5-ht1	69.192.139.229	134.79.229.218	1713	0	0	in
1134	IPT	slac50s-ht1	69.192.139.229	134.79.229.218	1736	0	0	out
1134	IPT	slac50s-ht1	69.192.139.229	134.79.81.30	1105	5	0	out
1134	IPT	slac50s-ht1	69.192.139.229	134.79.93.126	1187	4	0	out
vlan_id	any(sap_routing_instance)	exporting_node	ip_src	ip_dst	tot_packets	tot_retrans	tot_loss	direction

ESnet

High Level Diagram: ESnet Portion Of LHN



SLAC to GCP for Rubin Science Platform

- Problem statement: transfer of data objects from S3DF at SLAC (Rubin US Data Facility) to GCP is too slow
- Next level of detail:
  - SLAC DTNs have already been tuned
    - Including VERY large TCP buffers
  - Data objects are approximately 100MB each
  - Data transfers via HTTP/REST
  - Combination of data rate and data object size means transfer times are short (~2 seconds), but needs to be much faster

ESnet

SLAC to GCP for Rubin Science Platform

- Additional details
  - Nginx load balancer in front of DTNs
  - Download performance from SLAC to ESnet DTNs shows that performance inversely proportional to distance
    - Red flag - issue is latency sensitive
  - Transfers of the same file get different performance
    - First transfer is 10x slower than second and subsequent
      - Transfer from storage is much slower than transfer from cache
- Put all this together, and what do you get?

ESnet

SLAC to GCP for Rubin Science Platform

- Three separate components involved
- Storage:
  - Storage itself is much slower than storage system cache
  - Limitation of the storage itself - can’t fix that with networking
- Architecture:
  - DTNs are behind nginx load balancer
  - Load balancer host TCP config is what matters
- TCP:
  - One TCP connection per file
  - Data transfers aren’t getting out of slow start (done in 2 seconds)
  - Big buffers alone won’t help

ESnet

SLAC to GCP for Rubin Science Platform

- Solution: increase cwnd past RFC6928 values
  - Increase slow start flight size from 10 segments to 50
- But, do it on the load balancer
  - DTNs don't make TCP connections with GCP hosts
  - DTN → nginx → GCP
    - WAN path is between nginx and GCP
- Transfers now take 0.7 seconds
  - But only if they come from storage system cache
  - Raw storage is still ~8x slower → separate issue
    - Network can't fix raw storage performance
- Important notes:
  - **DO NOT** do this as part of a normal DTN tuning exercise
  - Very-short-duration TCP transactions are difficult!

ESnet


U.S. DEPARTMENT OF
ENERGY
BERKELEY LAB
ESnet
ENERGY SCIENCES NETWORK
Thanks!
https://my.es.net/
https://www.es.net/
https://fasterdata.es.net/