ESnet
ENERGY SCIENCES NETWORK

Evolution of ESnet - A Changing Landscape in Scientific Networking

Chin Guok
CTO
Energy Sciences Network
Lawrence Berkeley National Laboratory

Mini-GRP Workshop @ SCA23
Singapore
Mar 1, 2023

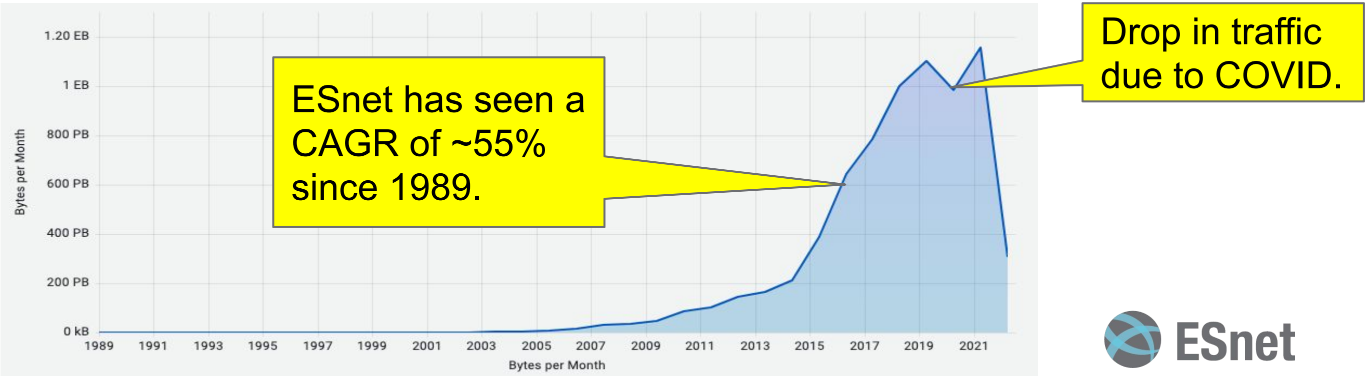
U.S. DEPARTMENT OF ENERGY
Office of Science

BERKELEY LAB

ESnet is a Science Mission Network

ESnet provides the high-bandwidth, reliable connections that **link scientists** at national laboratories, universities, and other research institutions, **enabling them to collaborate** on some of the world's most important scientific challenges including **energy**, **climate science**, and the **origins of the universe**. **Funded by the DOE Office of Science**, ESnet is managed and operated by the Scientific Networking Division at Lawrence Berkeley National Laboratory. As a nationwide infrastructure and DOE User Facility, ESnet provides scientists with **access to unique DOE research facilities and computing resources**.

ESnet's Mission is to enable and accelerate scientific discovery by delivering unparalleled network infrastructure, capabilities, and tools.



DOE Office of Science - Largest supporter of basic research in the physical sciences in the US

The mission of the **Advanced Scientific Computing Research (ASCR)** program is to discover, develop, and deploy computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to the Department of Energy (DOE).

Basic Energy Sciences (BES) supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security.

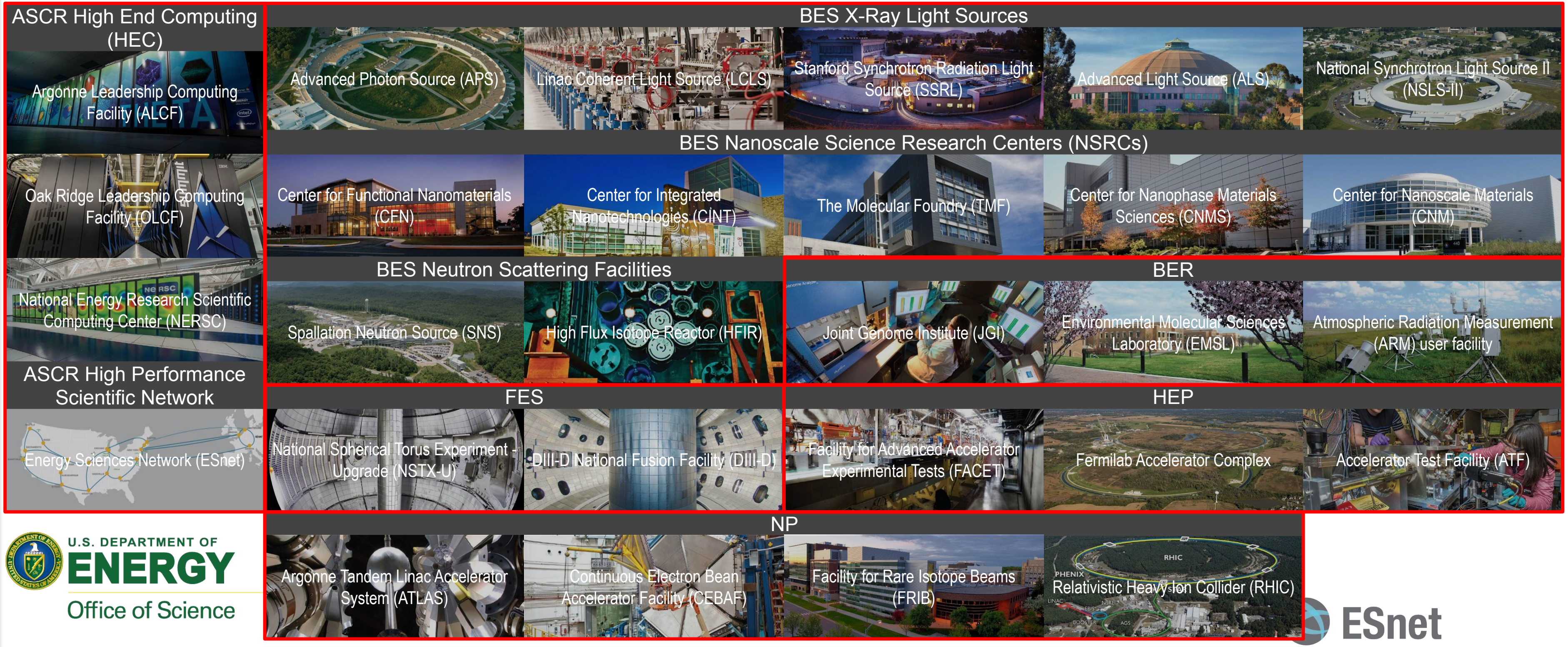
The mission of the **Biological and Environmental Research (BER)** program is to support transformative science and scientific user facilities to achieve a predictive understanding of complex biological, earth, and environmental systems for energy and infrastructure security, independence, and prosperity.

The **Fusion Energy Sciences (FES)** program mission is to expand the fundamental understanding of matter at very high temperatures and densities and to build the scientific foundation needed to develop a fusion energy source.

The mission of the **High Energy Physics (HEP)** program is to understand how our universe works at its most fundamental level.

The mission of the **Nuclear Physics (NP)** program is to discover, explore, and understand all forms of nuclear matter.

DOE Office of Science - Uniquely positioned for large scale collaborative science*

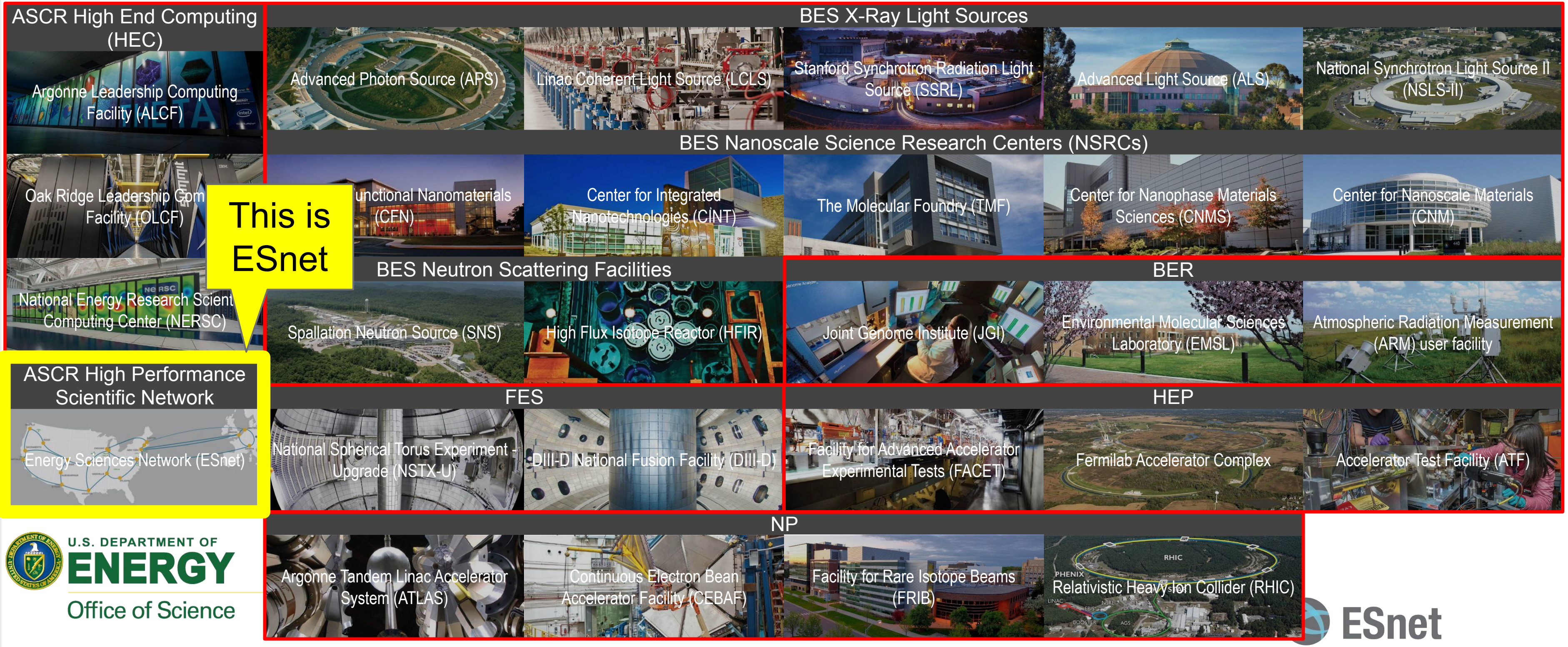


U.S. DEPARTMENT OF ENERGY
Office of Science

ESnet

DOE Office of Science facilities also support other collaborations, e.g., LHC, LSST, etc

DOE Office of Science - Uniquely positioned for large scale collaborative science*

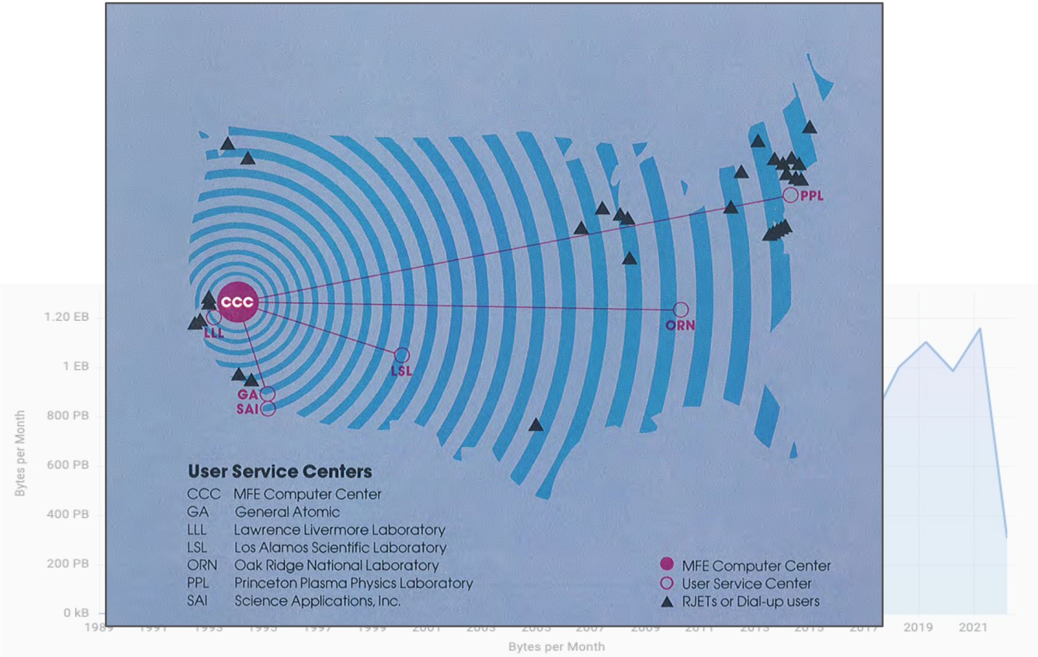


U.S. DEPARTMENT OF ENERGY
Office of Science

ESnet

DOE Office of Science facilities also support other collaborations, e.g., LHC, LSST, etc

Evolution of the Energy Sciences Network (ESnet)



Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

1970s | 1980s | 1990s | 2000s | 2010s | Present

ESnet

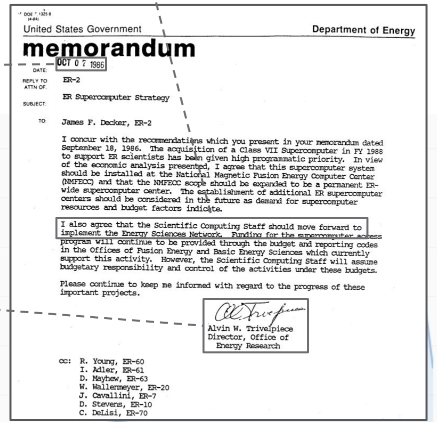
Evolution of the Energy Sciences Network (ESnet)

I also agree that the Scientific Computing Staff should move forward to implement the Energy Sciences Network. Funding for the supercomputer ac

ESnet(1)
[1986 - 1994]
Building an open standards network
- *IPv4*
- *BGP*

Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

OCT 0 7 1986



DOE F 1325.8
(4-84)

United States Government Department of Energy

memorandum

DATE: OCT 0 7 1986

REPLY TO ATTN OF: ER-2

SUBJECT: ER Supercomputer Strategy

TO: James F. Decker, ER-2

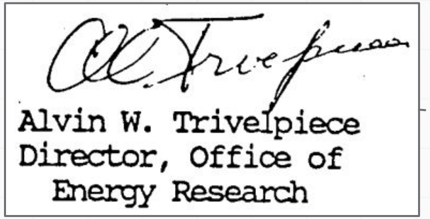
I concur with the recommendations which you present in your memorandum dated September 18, 1986. The acquisition of a Class VII Supercomputer in FY 1988 to support ER scientists has been given high programmatic priority. In view of the economic analysis presented, I agree that this supercomputer system should be installed at the National Magnetic Fusion Energy Computer Center (NMFECC) and that the NMFECC scope should be expanded to be a permanent ER-wide supercomputer center. The establishment of additional ER supercomputer centers should be considered in the future as demand for supercomputer resources and budget factors indicate.

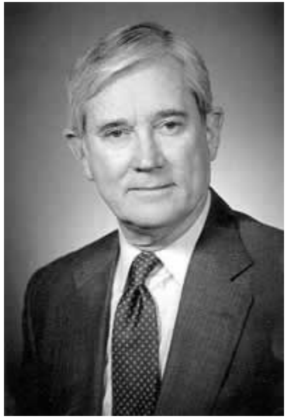
I also agree that the Scientific Computing Staff should move forward to implement the Energy Sciences Network. Funding for the supercomputer access program will continue to be provided through the budget and reporting codes in the Offices of Fusion Energy and Basic Energy Sciences which currently support this activity. However, the Scientific Computing Staff will assume budgetary responsibility and control of the activities under these budgets.

Please continue to keep me informed with regard to the progress of these important projects.

Alvin W. Trivelpiece
Director, Office of
Energy Research

cc: R. Young, ER-60
I. Adler, ER-61
D. Mayhew, ER-63
W. Wallenmeyer, ER-20
J. Cavallini, ER-7
D. Stevens, ER-10
C. DeLisi, ER-70





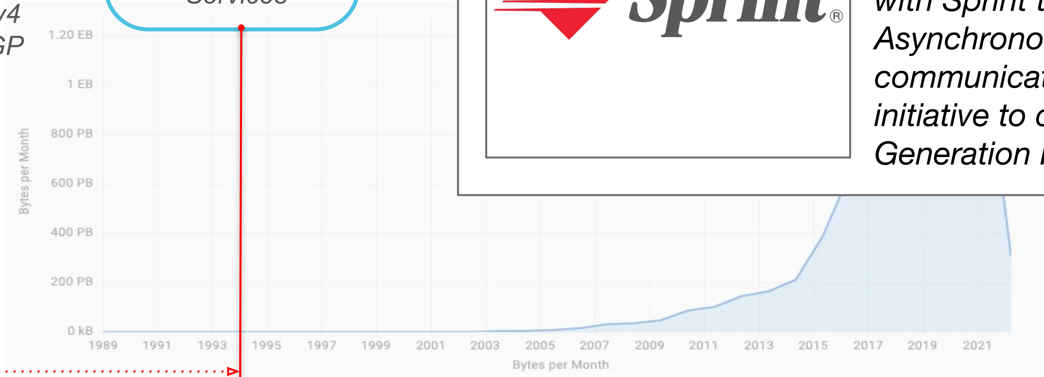
1970s | 1980s | 1990s | 2000s | 2010s | Present

Evolution of the Energy Sciences Network (ESnet)

Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

ESnet(1)
[1986 - 1994]
Building an open standards network
- *IPv4*
- *BGP*

ESnet2
[1994 - 2000]
Adopting emerging technologies
- *ATM*
- *IPv6*
- *Video Conf Services*



SPRINT AGREEMENT FOR "NEXT-GEN INTERNET"

ESnet signs a multi-year agreement with Sprint to provide high-speed Asynchronous Transfer Mode (ATM) communications as part of a federal initiative to develop the "Next Generation Internet."

1970s | 1980s | 1990s | 2000s | 2010s | Present

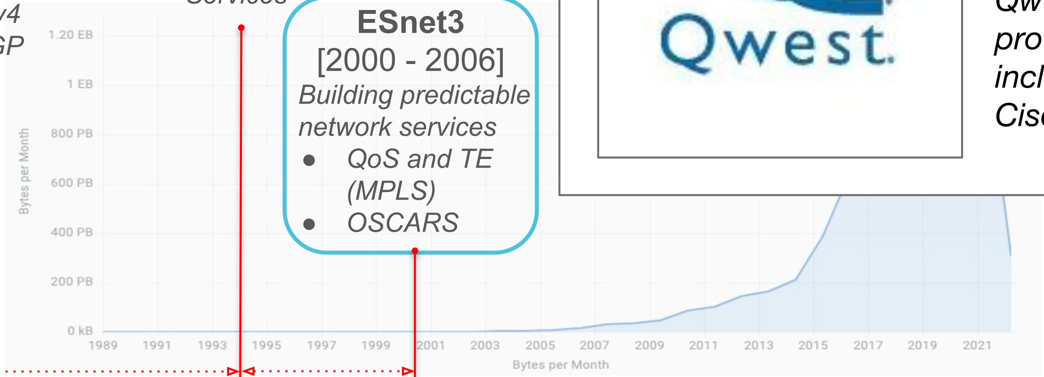
Evolution of the Energy Sciences Network (ESnet)

Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

ESnet(1)
[1986 - 1994]
Building an open standards network
- *IPv4*
- *BGP*

ESnet2
[1994 - 2000]
Adopting emerging technologies
- *ATM*
- *IPv6*
- *Video Conf Services*

ESnet3
[2000 - 2006]
Building predictable network services
- *QoS and TE (MPLS)*
- *OSCARS*



PARTNERSHIP SHIFT TO QWEST

Transition from Sprint to Qwest as ESnet's network provider. Equipment includes Juniper M20 and Cisco 8540 & 6500 routers

1970s | 1980s | 1990s | 2000s | 2010s | Present

Evolution of the Energy Sciences Network (ESnet)

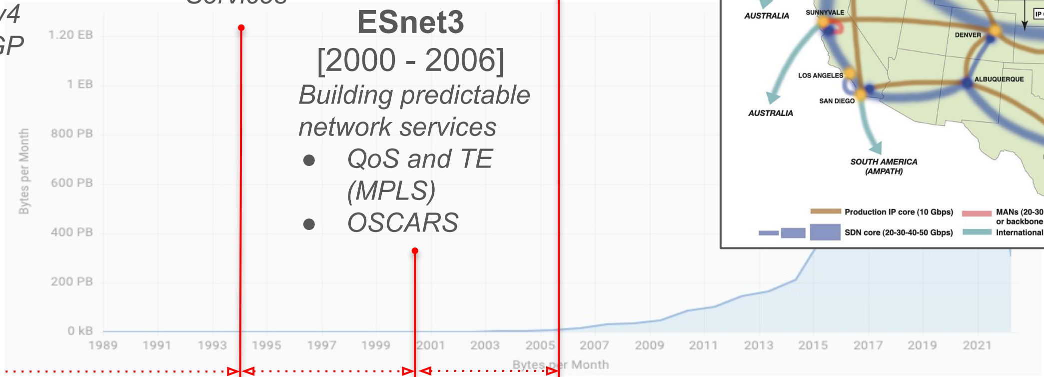
Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

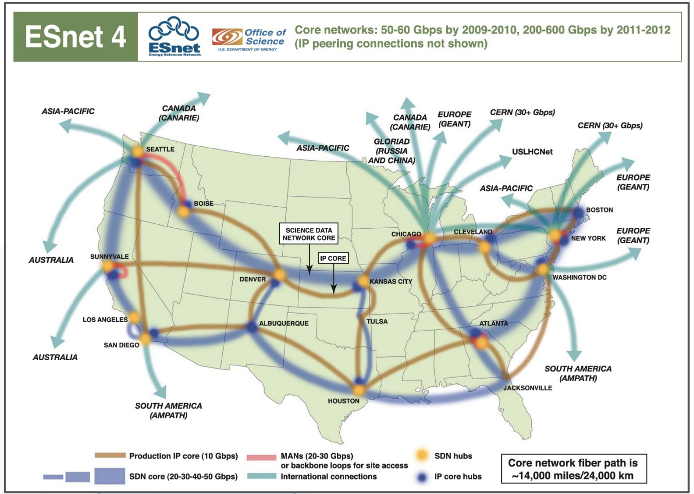
ESnet(1)
[1986 - 1994]
Building an open standards network
- *IPv4*
- *BGP*

ESnet2
[1994 - 2000]
Adopting emerging technologies
- *ATM*
- *IPv6*
- *Video Conf Services*

ESnet3
[2000 - 2006]
Building predictable network services
- *QoS and TE (MPLS)*
- *OSCARS*

ESnet4
[2006 - 2011]
Deploying purpose build architectures
- *Science Data Network (SDN)*
- *ScienceDMZ*
- *perfSONAR*





1970s → 1980s → 1990s → 2000s → 2010s → Present

ESnet Energy Sciences Network

ESnet

Evolution of the Energy Sciences Network (ESnet)

Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

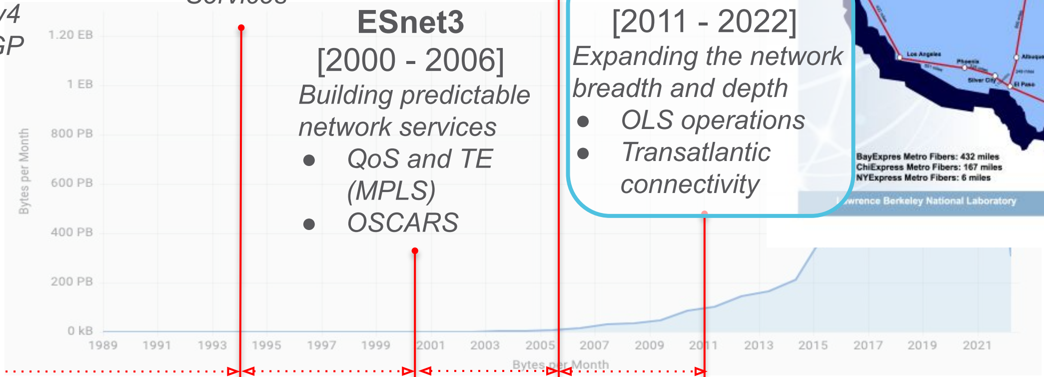
ESnet(1)
[1986 - 1994]
Building an open standards network
- *IPv4*
- *BGP*

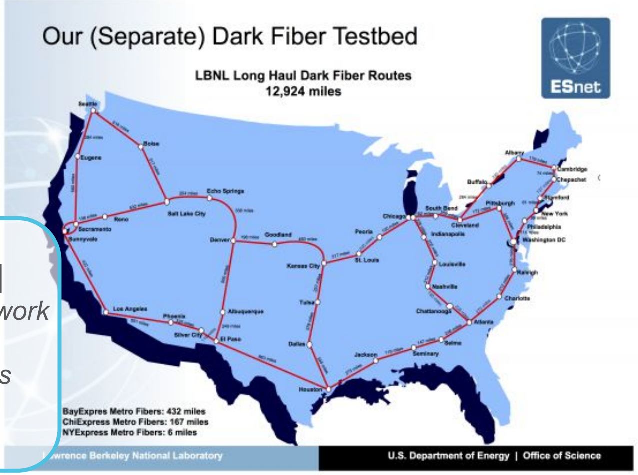
ESnet2
[1994 - 2000]
Adopting emerging technologies
- *ATM*
- *IPv6*
- *Video Conf Services*

ESnet3
[2000 - 2006]
Building predictable network services
- *QoS and TE (MPLS)*
- *OSCARS*

ESnet4
[2006 - 2011]
Deploying purpose build architectures
- *Science Data Network (SDN)*
- *ScienceDMZ*
- *perfSONAR*

ESnet5
[2011 - 2022]
Expanding the network breadth and depth
- *OLS operations*
- *Transatlantic connectivity*





1970s | 1980s | 1990s | 2000s | 2010s | Present

Evolution of the Energy Sciences Network (ESnet)

Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

ESnet(1)
[1986 - 1994]
Building an open standards network
- *IPv4*
- *BGP*

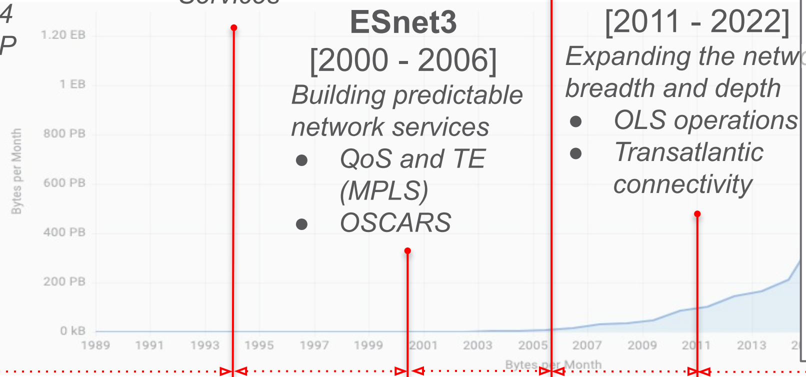
ESnet2
[1994 - 2000]
Adopting emerging technologies
- *ATM*
- *IPv6*
- *Video Conf Services*

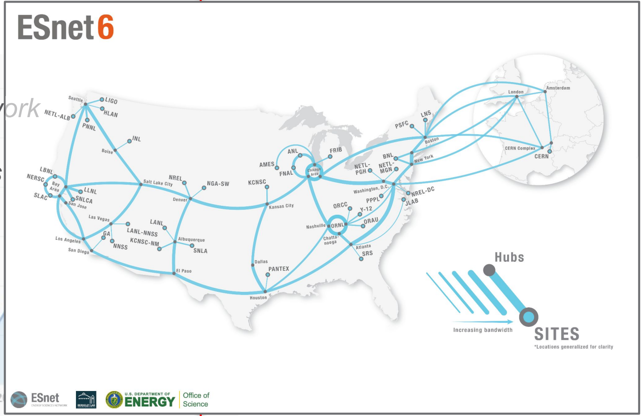
ESnet3
[2000 - 2006]
Building predictable network services
- *QoS and TE (MPLS)*
- *OSCARS*

ESnet4
[2006 - 2011]
Deploying purpose build architectures
- *Science Data Network (SDN)*
- *ScienceDMZ*
- *perfSONAR*

ESnet5
[2011 - 2022]
Expanding the network breadth and depth
- *OLS operations*
- *Transatlantic connectivity*

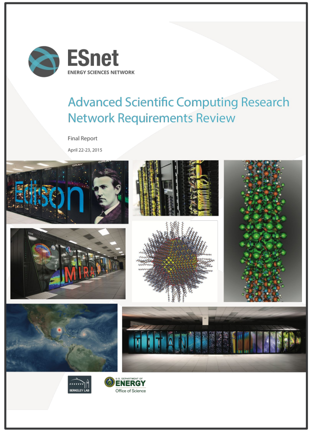
ESnet6
[2022 - TBD]
Developing comprehensive network automation and visibility capabilities
- *ESnet6 Orchestration/Automation*
- *ESnet6 High-Touch*

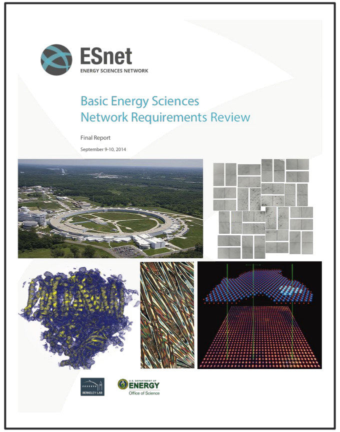


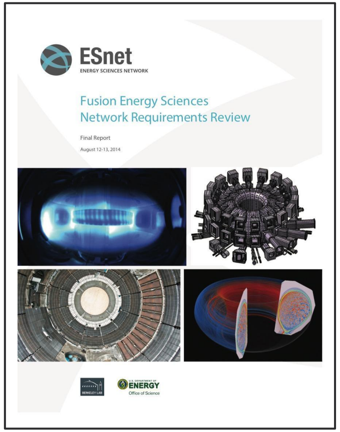


1970s | 1980s | 1990s | 2000s | 2010s | Present

ESnet Services Definition Process







- Requirements gathered from (6) DOE Office of Science program office requirements workshops.
  - Advanced Scientific Computing Research (ASCR)
  - Basic Energy Sciences (BES)
  - High Energy Physics (HEP)
  - Biological and Environmental Research (BER)
  - Fusion Energy Sciences (FES)
  - Nuclear Physics (NP)
- Input on requirements are documented as workflows, which are then formalized as services, driving the design and architecture.

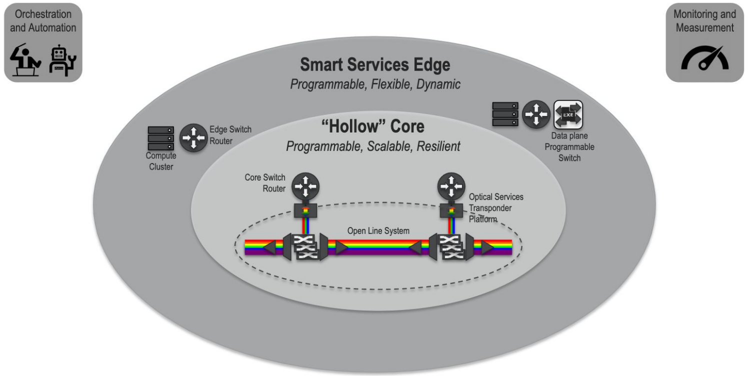
ESnet

ESnet6 (Relevant) Capabilities

- **Bandwidth:** 4x+ of bandwidth ensures that there is no loss and data can be transferred faster and easier than hard drive shipping
- **Capacity:** Dedicated optical system (OLS) allows for cost-effective bandwidth to support growing needs
- **Resiliency:** Redesign of site connectivity and adoption of new network protocols to ensure dependability and predictability for data transfer
- **Orchestration/Automation:** Automation allows us to create custom services with rich and highly integrated set of capabilities quickly, and consistently, with high reliability
- **Connectivity & Scale:** Rich set of connections allows flexibility and broad access to needed resources (other collaborators in DOE, R&E, and cloud) anywhere in the US and into Europe
- **Telemetry:** Visibility allows scientists to understand and optimize their performance
- **Consulting:** Engagement team helps scientists to make best use of resources without networking expertise
- **Security:** Secure services support dependability and data integrity and privacy

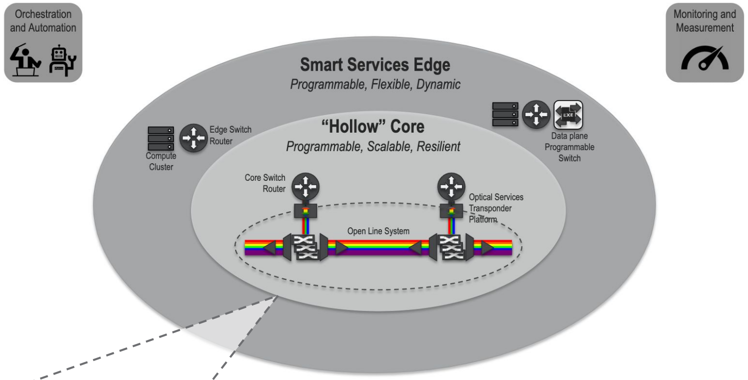
ESnet6 Design and Build (in a nutshell)

ESnet6 "Hollow" Core Architecture

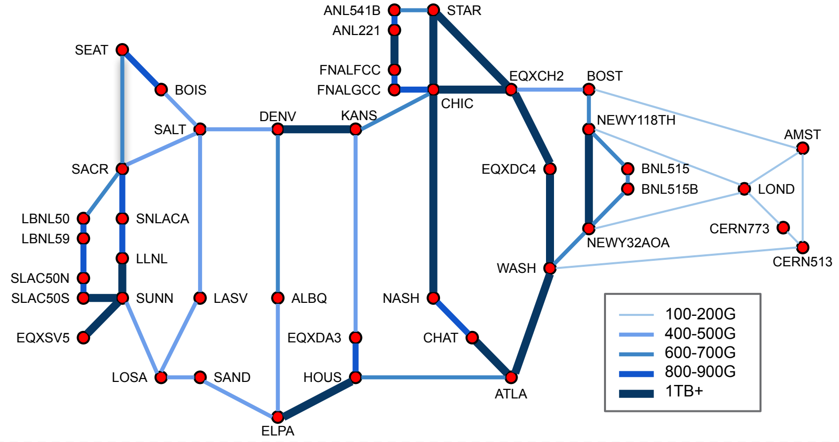


ESnet6 Design and Build (in a nutshell)

ESnet6 “Hollow” Core Architecture

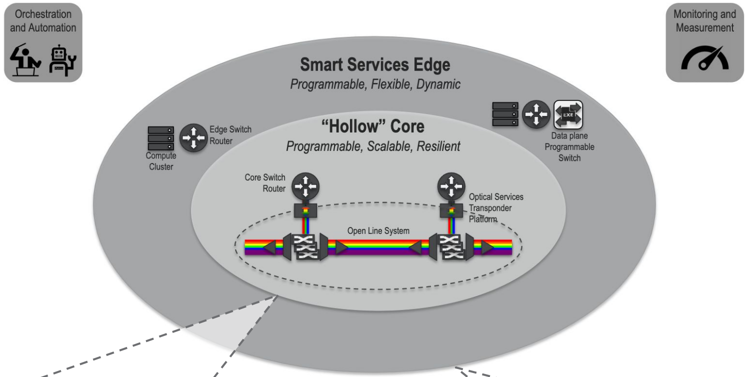


ESnet6 Network Capacity (as of May 2022)

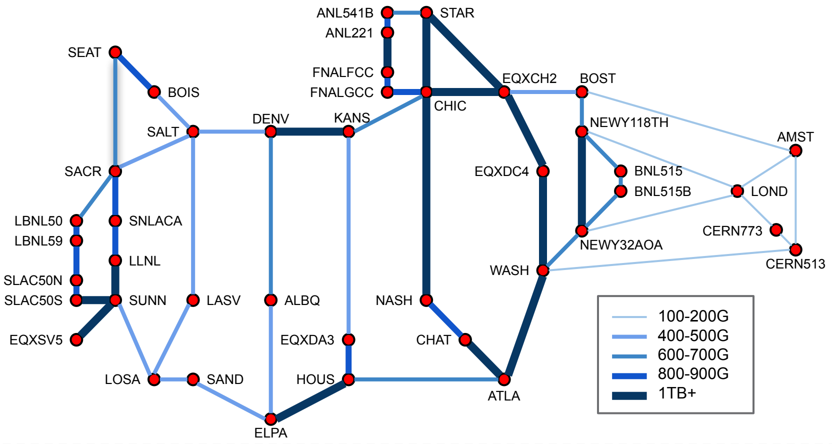


ESnet6 Design and Build (in a nutshell)

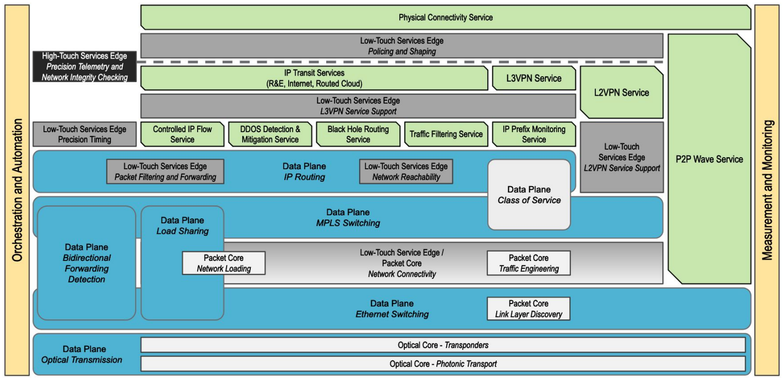
ESnet6 “Hollow” Core Architecture



ESnet6 Network Capacity (as of May 2022)

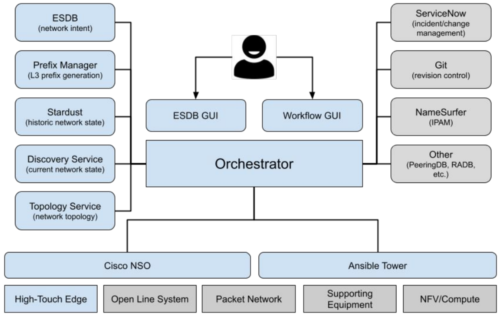


ESnet6 Services and Capabilities Structure

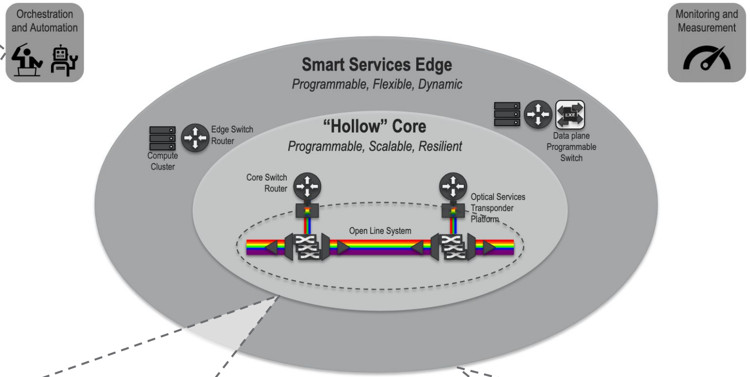


ESnet6 Design and Build (in a nutshell)

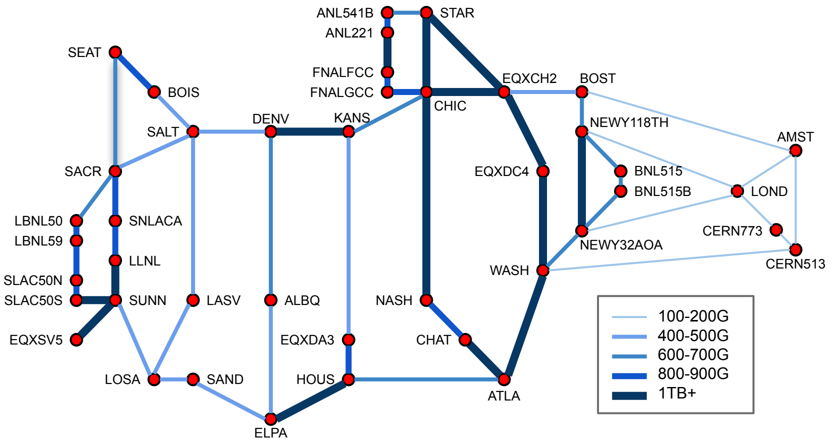
ESnet6 Orchestration & Automation Framework



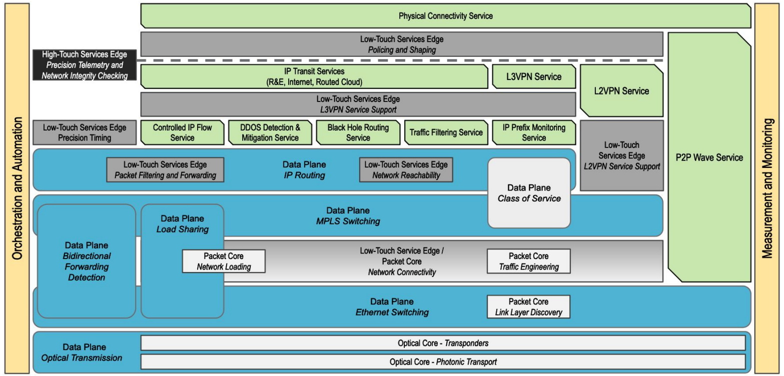
ESnet6 "Hollow" Core Architecture



ESnet6 Network Capacity (as of May 2022)

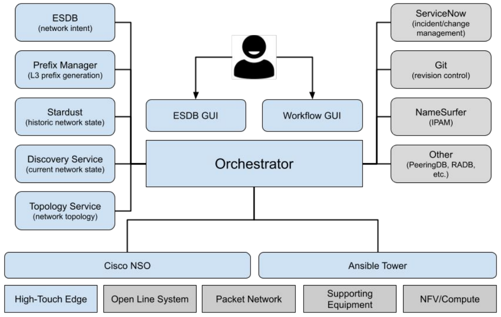


ESnet6 Services and Capabilities Structure

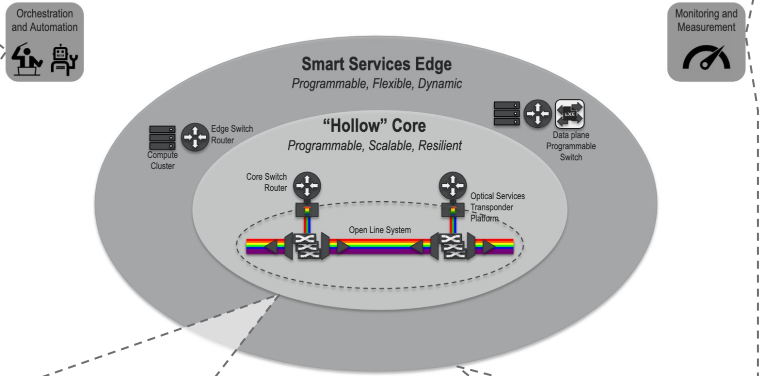


ESnet6 Design and Build (in a nutshell)

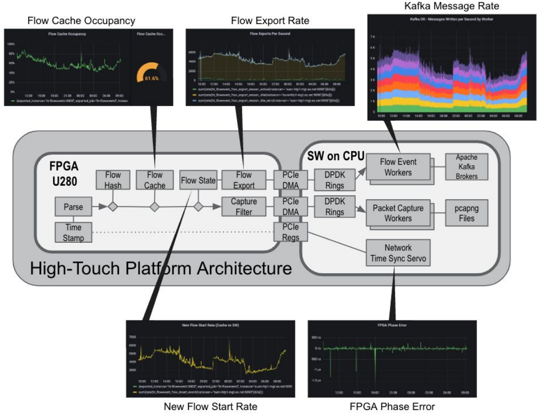
ESnet6 Orchestration & Automation Framework



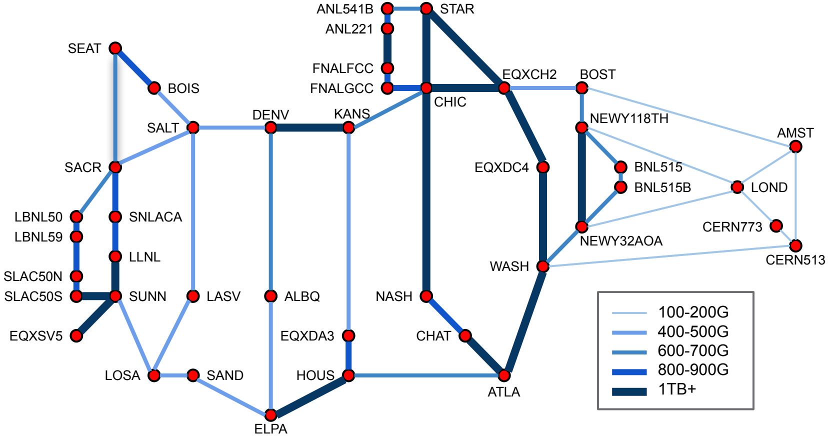
ESnet6 "Hollow" Core Architecture



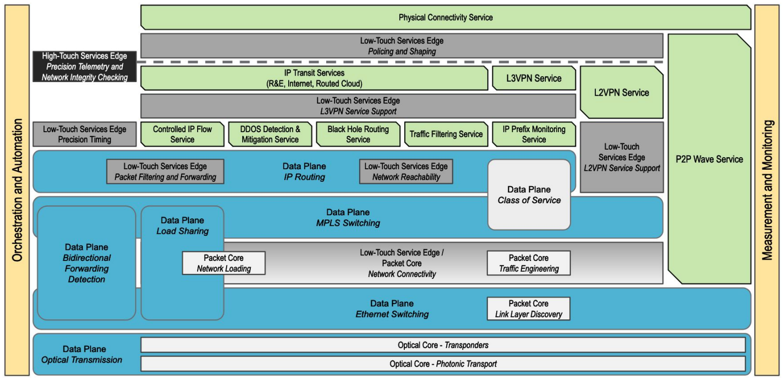
ESnet6 High-Touch Precision Network Telemetry Platform



ESnet6 Network Capacity (as of May 2022)

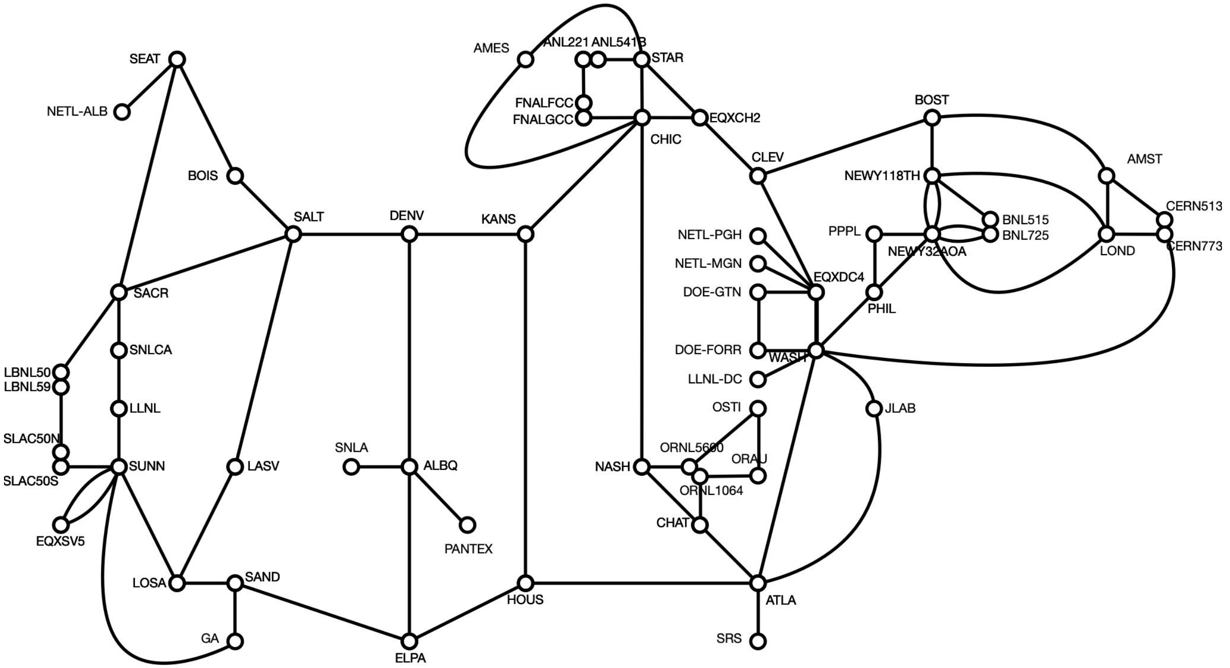


ESnet6 Services and Capabilities Structure



ESnet6 High-Touch System Deployment

42 deployment locations, each location will have 2 High-Touch servers

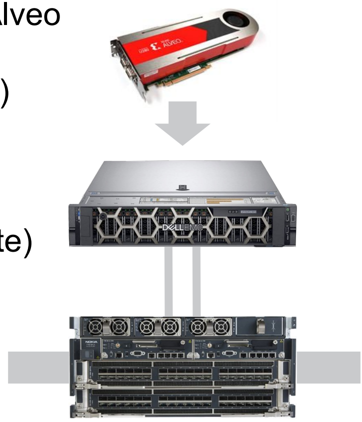


High-Touch Server Hardware Deployment

AMD Xilinx Alveo U280 FPGA (1 per server)

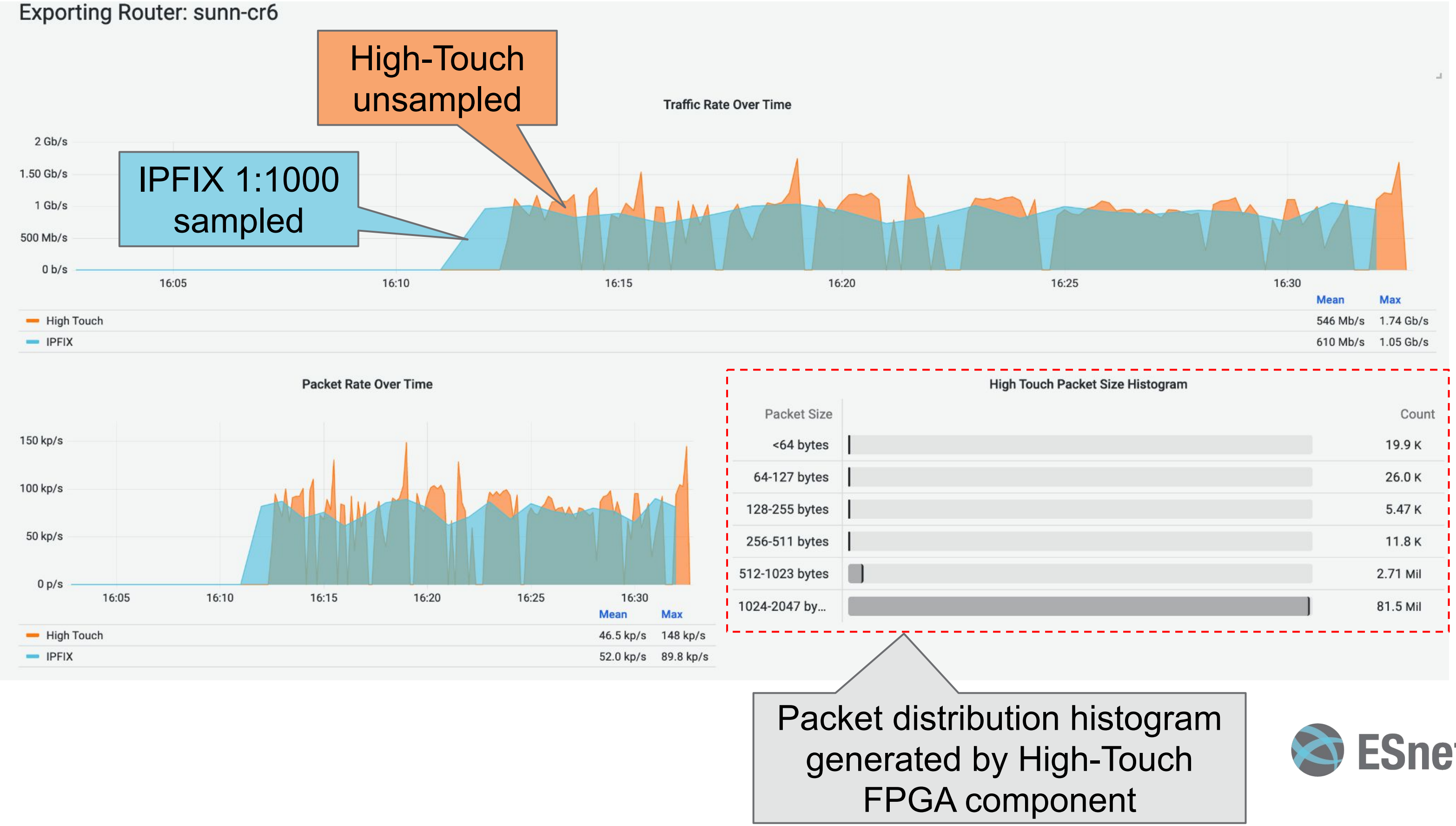
High-Touch Server (2 per hub site)

ESnet6 Core Router



ESnet

Visibility into Network Flow Performance



ESnet

Visibility into Network Flow Details

HighTouch
programmable, high-resolution network telemetry

Billions of flows served daily

- Column-oriented (SQL) database (ClickHouse)
- Built to handle trillions of rows, petabytes of data
- Unsampled High-Touch flow exports (5 tuple)
- Centralized collection
- 2.5 billions rows inserted daily (11 HT nodes) (as of 02.09.2023)
- queries process millions rows/second
- Goal: retain 30 days of exports
  - ~ 1.15 trillion rows
  - ~ 100 TB

ESnet

SQL to access All Flows across the Entire Network

Columnar SQL Database

Clickhouse *

```sql
SELECT
    round(log2(packets), 0) AS bin,
    exp2(bin)               AS bin_low,
    exp2(bin + 1) - 1       AS bin_high,
    max(packets)            AS max_pkts_in_export,
    count(packets)          AS cnt,
    round(cnt / tot, 2)     AS pct,
    bar(cnt, 0, tot, 30)    AS packet_distribution_tcp
FROM ht.all_flows
WHERE (end_time_ns >= toDateTime('2023-02-09 00:00:00'))
  AND (end_time_ns <  toDateTime('2023-02-09 12:00:00'))
  AND (ip_proto_num = 6)
GROUP BY bin
ORDER BY bin ASC
```

bin	bin_low	bin_high	max_pkts_in_export	cnt	pct	packet_distribution_tcp
0	1	1	1	632870729	67	
1	2	3	2	80841816	9	
2	4	7	5	104701309	11	
3	8	15	11	46543951	5	
4	16	31	22	40139184	4	
5	32	63	45	10509970	1	
6	64	127	90	4720006	1	
7	128	255	181	4044598	0	
8	256	511	362	3052280	0	
9	512	1023	724	3037506	0	
10	1024	2047	1448	2313619	0	
11	2048	4095	2896	2995997	0	
12	4096	8191	5792	2218766	0	
13	8192	16383	11585	1213550	0	
14	16384	32767	23170	590475	0	
15	32768	65535	46340	276320	0	
16	65536	131071	92681	211259	0	
17	131072	262143	185363	102852	0	
18	262144	524287	370726	84040	0	
19	524288	1048575	741370	71900	0	
20	1048576	2097151	1482801	16990	0	
21	2097152	4194303	2956901	3057	0	
22	4194304	8388607	5893181	207	0	
23	8388608	16777215	8031073	19	0	

24 rows in set. Elapsed: 9.778 sec. Processed 2.41 billion rows, 40.98 GB (246.50 million rows/s.

NB: Use of Clickhouse is part of prototyping effort

ESnet

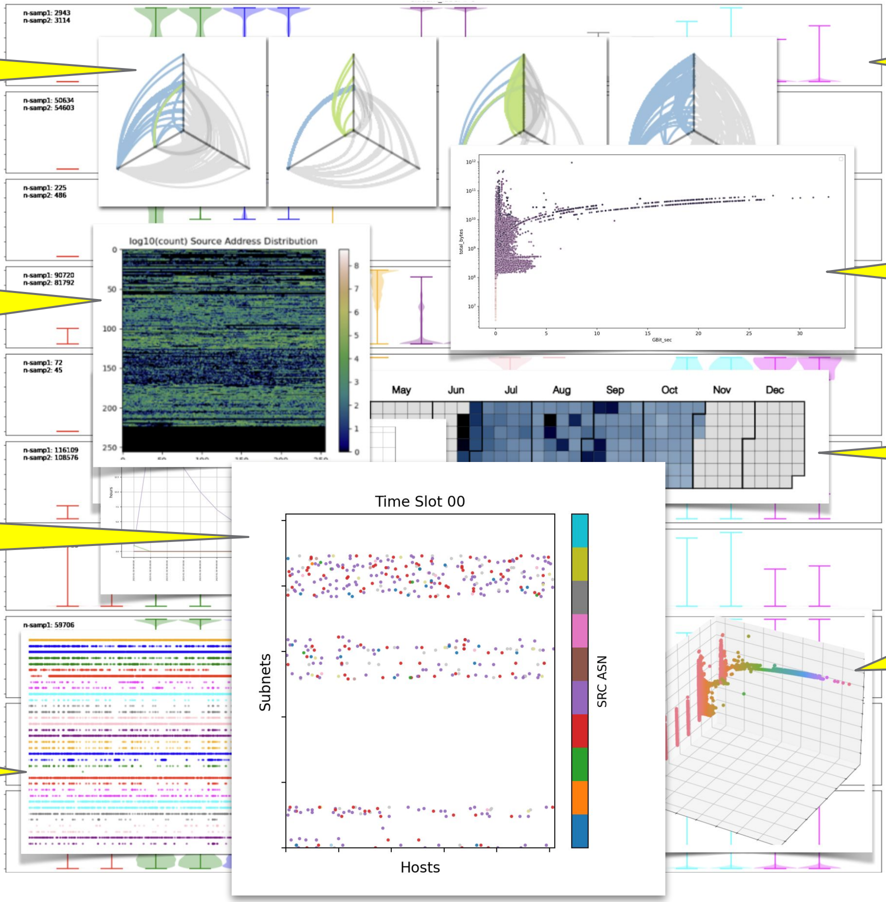
Easy Integration with Data Science and ML Libraries

Host fingerprinting to detect anomalous connections

Background radiation monitoring for security breaches

Data import lag analysis for understanding ingest issues

Port and subnet scanning to identify malicious activities



Cluster attribute analysis for identifying common traffic characteristics

Throughput analysis to determine data movement performance

Calendar view of activity for network planning

Packet size analysis to understand flow characteristics and behaviors

ESnet

Haberman et al

Evolution of the Energy Sciences Network (ESnet)

Magnetic Fusion Energy Network (MFENET)
[1976 - 1986]

ESnet(1)
[1986 - 1994]
Building an open standards network
- *IPv4*
- *BGP*

ESnet2
[1994 - 2000]
Adopting emerging technologies
- *ATM*
- *IPv6*
- *Video Conf Services*

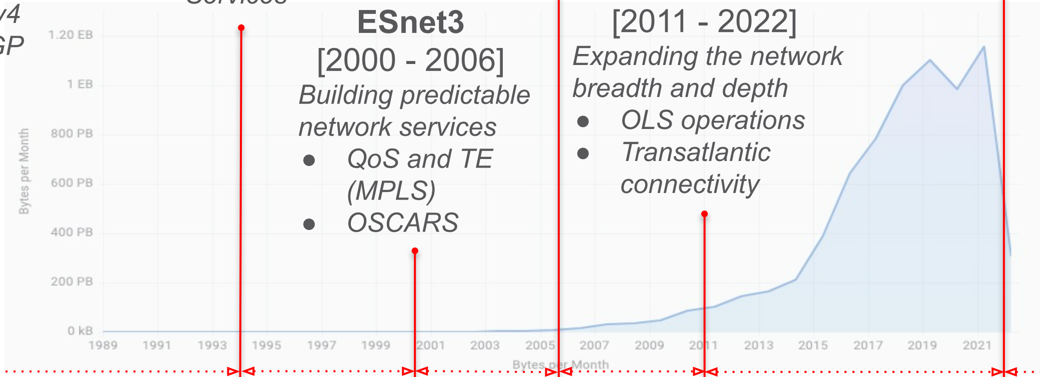
ESnet3
[2000 - 2006]
Building predictable network services
- *QoS and TE (MPLS)*
- *OSCARS*

ESnet4
[2006 - 2011]
Deploying purpose build architectures
- *Science Data Network (SDN)*
- *ScienceDMZ*
- *perfSONAR*

ESnet5
[2011 - 2022]
Expanding the network breadth and depth
- *OLS operations*
- *Transatlantic connectivity*

ESnet6
[2022 - TBD]
Developing comprehensive network automation and visibility capabilities
- *ESnet6 Orchestration/Automation*
- *ESnet6 High-Touch*

ESnet7
[TBD]
Developing intuitive and intelligent network functions and data centric services
- *AI/ML driven Networking*
- *Computational Storage*



1970s | 1980s | 1990s | 2000s | 2010s | Present

Scheduling Workflows based on Network Traffic Predictions (using Graph neural networks)

Scientific Achievement

Highly accurate 48 hour predictions for ESnet reliable predictions. These results will inform how scientific workflows will be scheduled for optimum network performance.

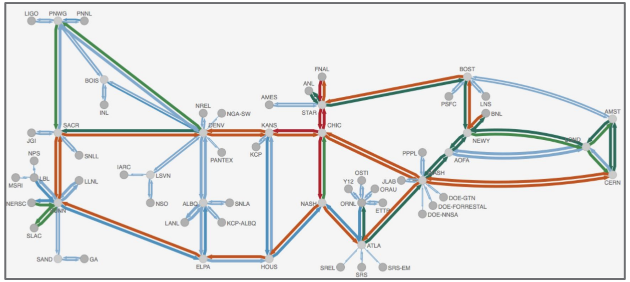
Significance and Impact

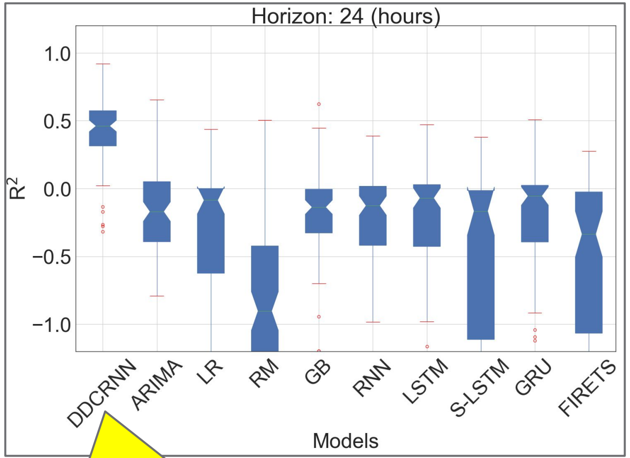
A novel Dynamic Diffusion based neural network that takes temporal and spatial information to produce accurate predictions

- Compared to existing approaches, this is highly improved results
- Adding capability to add multiple workflow legs so that end-to-end schedules

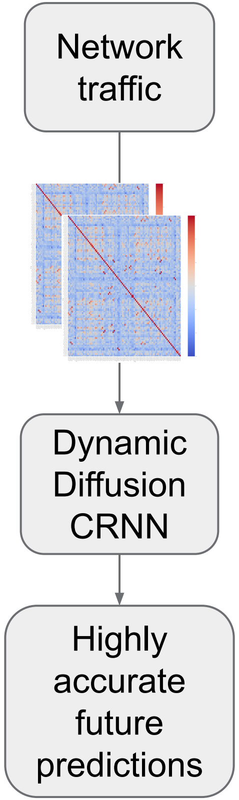
Research Details

- Developing a mathematical framework for hyperparameter for Dynamic Diffusion Convolutional Recurrent Neural Network (DDCRNN) model
- Testing with various network data sets
- Deploying as an easy to use GUI and open source project





Our approach shows significant improvement compared to other approaches

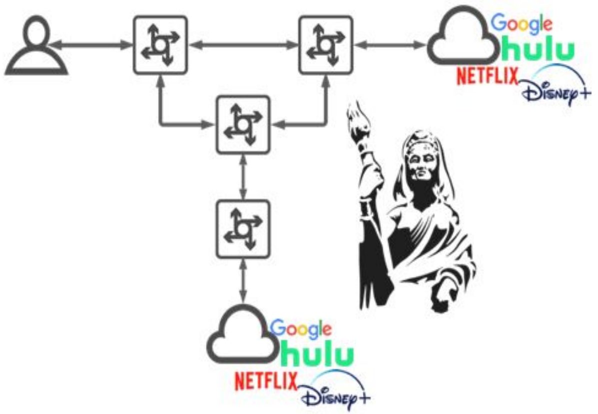


ESnet

BERKELEY LAB

Kiran et al

Optimizing Traffic Flows in Real-time : HECATE*



Scientific Achievement

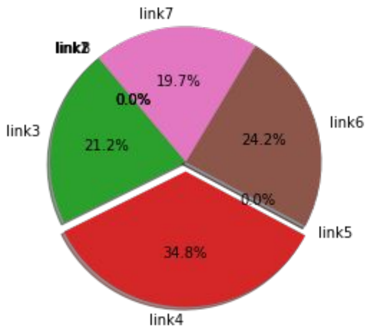
Create “be-spoke networks” for engineers to optimize of multiple criteria like loss, throughput, traffic patterns and site characteristics.

Significance and Impact

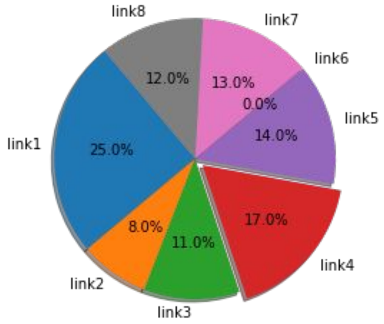
Every network is different and often need constant human attention. We develop a reinforcement learning approach coupled with unsupervised learning to help HECATE learn optimal patterns and then optimize the network when HECATE is turned on.

Research Details

- No-compromise on performance: Hecate monitors network “health” and actively reroutes traffic
- Caters to many applications: Hecate self-learns traffic classes to guarantee service
- Seamlessly integrate multiple network solutions
- Deployable as hardware solution



Before



After

NB: Improvement of load sharing across all links and overall average utilization

ESnet

**Patent filed – Deep Learning informed Traffic Engineering*

Kiran et al

In-Network Data Caching

Scientific Achievement

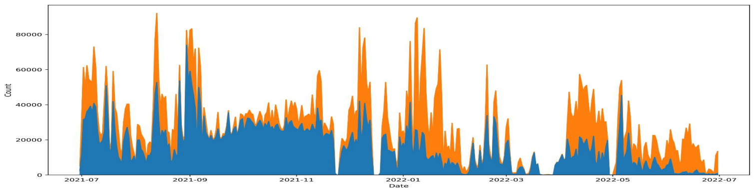
We experimented and demonstrated the capability of a network-based temporary data cache; how in-network caching mechanism helps network traffic performance, how much data can be shared within the network, and how much network traffic volume can be reduced consequently.

Significance and Impact

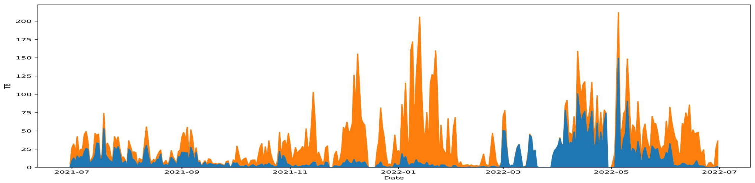
- In-network services such as temporary data caching could potentially have a big impact on traffic engineering and how the remote data is being accessed.
- Data caching mechanism in a region is expected to reduce the redundant data transfers, saved network traffic, and lower data access latency improving overall application performance.
- It also provides the unique capability for a network provider to design data hotspots into the network topology. The appropriate bandwidth resources and traffic engineering techniques can manage traffic movement and congestion.

Research Details

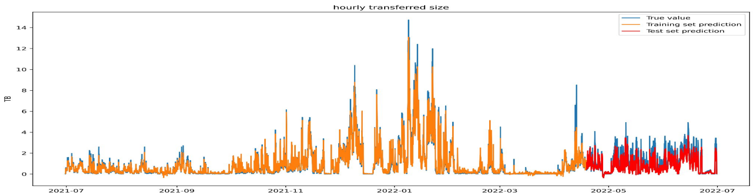
- ESnet cache node as a part of SoCal Petabyte scale cache in collaboration with Caltech, UCSD, and US CMS.
- Studied 1-year's operational from SoCal Repo from Jul 2021 to Jun 2022.
- On average 67.6% of file requests were satisfied by the cache, which translated to 4.5PB (35.4%) of requested bytes (12.7PB) served by the cache.
- Network traffic was reduced by up to 29TB per day due to cached data.
- Sim et al. "Effectiveness and predictability of in-network storage cache for Scientific Workflows", IEEE ICNC, 2023



Daily file requests (count) - cache misses (in orange) and cache hits (in blue)



Daily traffic volume - cache misses (in orange) and cache hits (in blue)



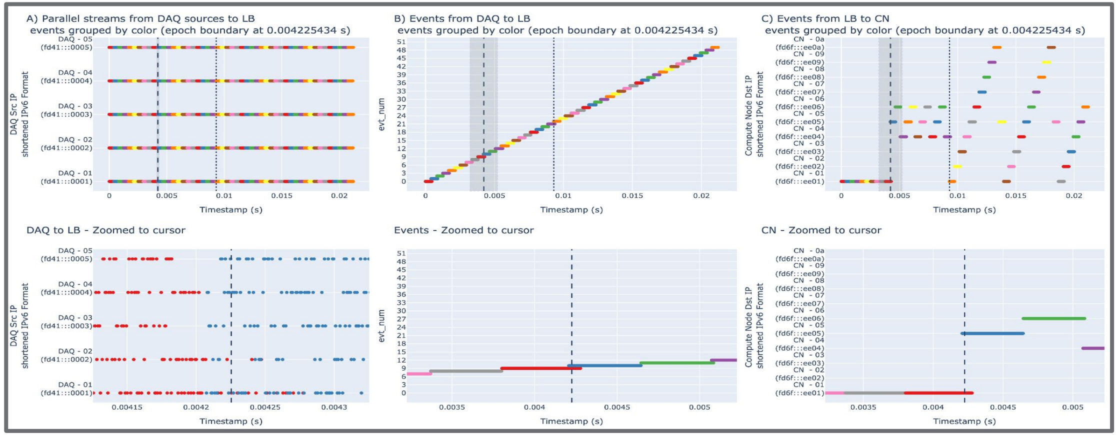
Hourly volume (in bytes) of cache misses from the LSTM model output

Sim et al

EJ-FAT FPGA Accelerated Transport Load Balancer

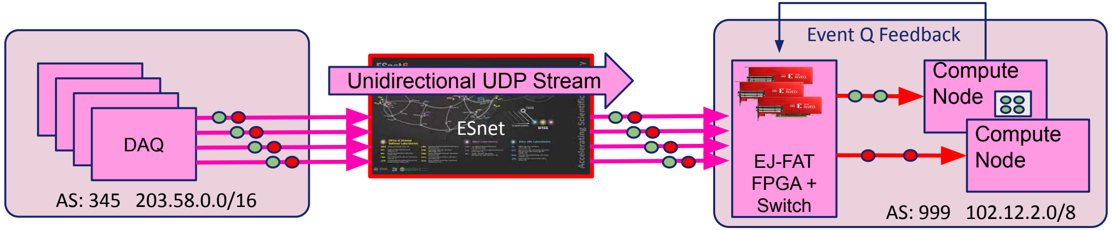
Scientific Achievement

The real-time load balancer is designed to support WAN latencies for geographically distributed accelerator facilities and high performance computing centers, and has been successfully integrated with JLab's ERSAP processing pipeline for end-to-end- event processing.



Significance and Impact

- Horizontal scale. Keep adding parallel FPGAs and switches to achieve Terabits of throughput. All the elements work together to get related pieces of data to each compute node.
- Multi-Domain. DAQ source only need to know 1 dst IP for the load balancer. Compute nodes can register their IP address with the LB and receive work.
- Work is broken into 1uS or shorter packet bursts. Zero packet loss or accidental reshuffling in the load balancer, even in dynamic environments with compute nodes changing on the fly. Overlapping event times do not confuse the load balancer.



Research Details

- Separation of IP Addresses between labs
- In network sorting of Event Data
- Stateless load balancing
- Compute node feedback for dynamic LB
- Hit-less reconfiguration of LB table
- Unidirectional UDP streaming

Kumar et al

Questions…

Chin Guok <chin@es.net>