ESnet6:
Building the infrastructure to support the next-generation of science

Inder Monga

Executive Director, ESnet

Division Director, Scientific Networking

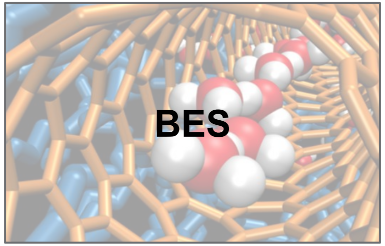
Lawrence Berkeley National Laboratory

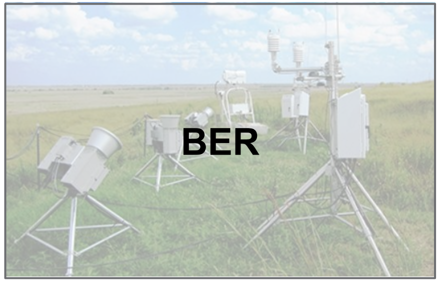
ASCAC

September 2022

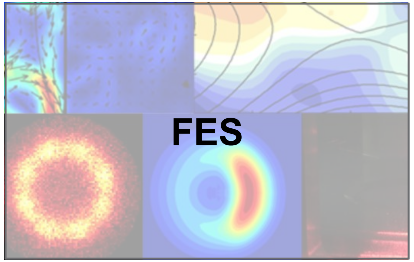
ESnet

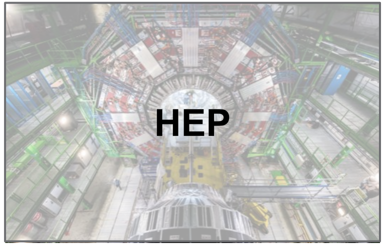
The ESnet user facility:
A complex system tuned for science

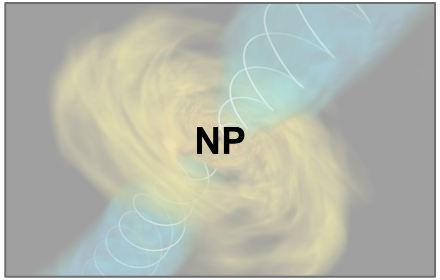




U.S. DEPARTMENT OF
ENERGY
Office of Science







ESnet

Facility upgrade supports the evolution of the scientific process

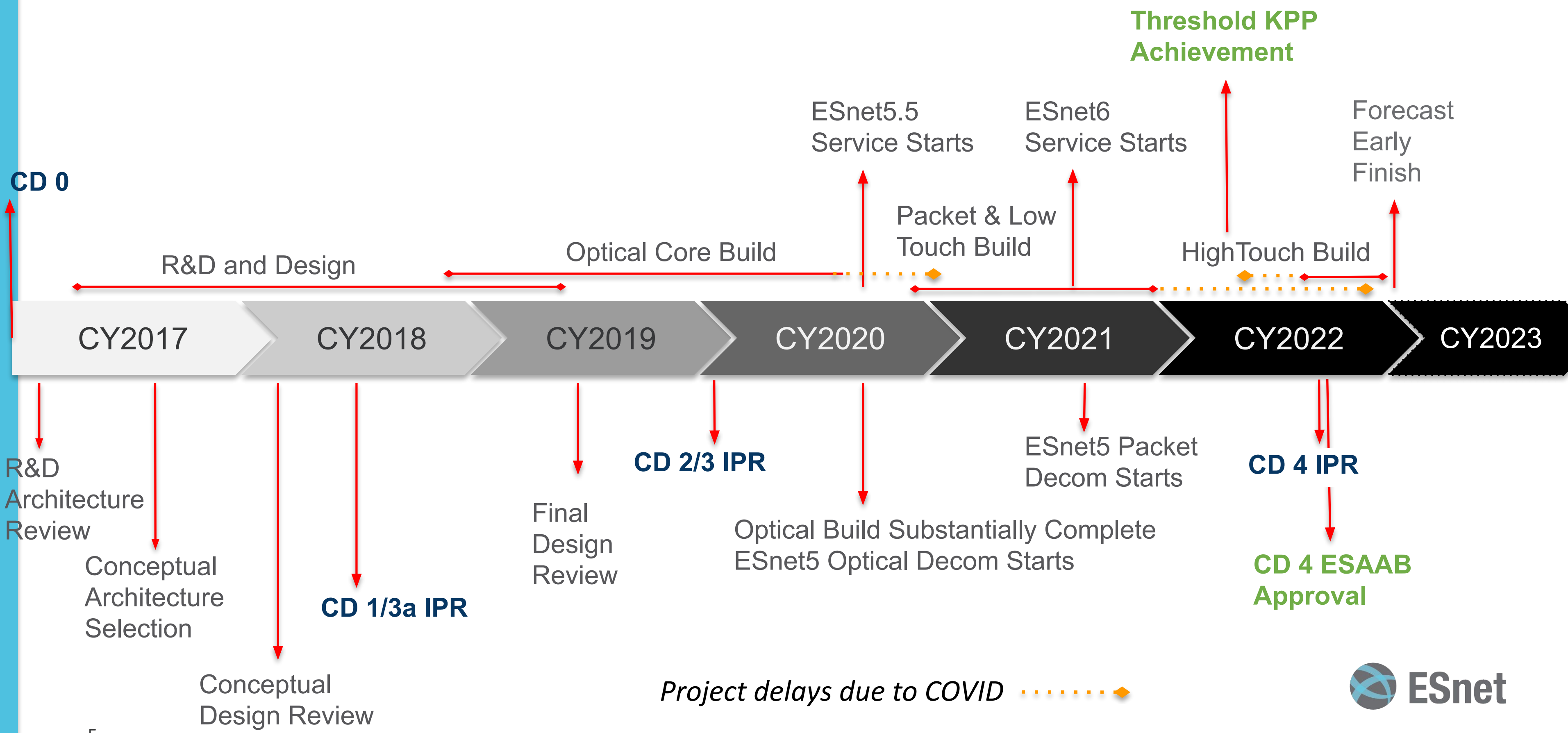
Exponential increases in data	→	**Network capacity** to handle traffic **Just-in-time ability to add capacity**
Productivity of science and national labs depends on the network	→	**Improve resiliency,** including cyber
New scientific workflows require convergence of data sources and facilities	→	**Flexibility** through **automation** and **programmability**, ability to create custom data and network services

Transformative but challenging project

..and we are deconstructing the older plane and transferring the passengers to the new one in parallel

- First DOE 413.3 project for the facility
- First greenfield design and build of the entire network by ESnet team
- First time implementing and operating the optical layer
- >50% of the team hired and onboarded during the project execution.
- ~10x increase in coordination, communication and reporting due to the Pandemic
- ~Zero unplanned downtime, and limited off hours planned downtime
- **Strong support from ASCR/DOE and Congress**

ESnet6 Project: Six years from concept to done



ESnet

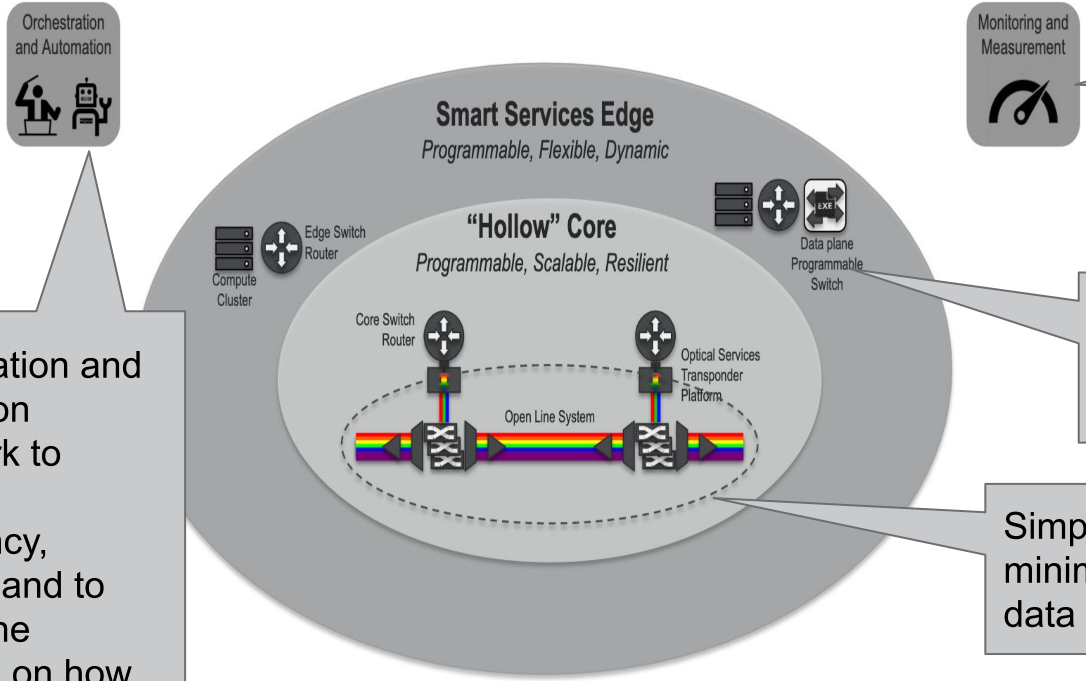
Threshold KPP's were met earlier this year

Description	Threshold KPPs	Objective KPPs
1. **Network Backbone:** Deliver a new Tb-scale ESnet6 networking backbone with at least 2X the capability of ESnet5 that can deliver sufficient data movement capacity for the next 7-10 years	T1a. Installed and commissioned new optical equipment to support wave transmission on 40 fiber segments *Baseline: 7/2020* ***Actual : 3/2021***	O1a. Installed and commissioned new optical equipment to support wave transmission on at least 52 fiber segments *Baseline: 8/2021* ***Actual : 9/2021***
	T1b. Deployed and commissioned 15.5 Tbps of network capacity on the backbone *Baseline: 2/2022* ***Actual : 1/2022***	O1b. Deployed and commissioned at least 20.6 Tbps of network capacity on the backbone *Baseline: 4/2022* ***Actual: 5/2022***
	T1c. Installed and commissioned new routing equipment at the Network Backbone Hub Locations *Baseline: 9/2021* ***Actual: 2/2022***	O1c. Installed and commissioned new routing equipment at the Network Backbone Hub Locations and Connected Sites *Baseline: 11/2021* *Forecast: 12/2022*
2. **Automation:** Using an integrated network orchestration platform, commission automated provisioning and monitoring of network operations and security services	T2a. Deployed automated provisioning of one network service *Baseline: 11/2021* ***Actual: 10/2021***	O2a. Deployed automated provisioning of two or more network services *Baseline: 8/2021* ***Actual: 8/2021***
	T2b. and one security service *Baseline: 04/2022* ***Actual : 3/2022***	O2b. and two or more security services *Baseline: 1/2023* *Forecast: 1/2023*
3. **Programmable Network Flexibility:** Design and implement a highly programmable data plane for development and deployment of innovative science data services	T3. Demonstrated one service using a programmable data plane (i.e., high-touch service), at two sites *Baseline: 4/2022* ***Actual : 3/2022***	O3. Deployed one or more services using a programmable data plane (i.e., high-touch services), among more than two sites *Baseline 09/2022* *Forecast: 12/2022*

What did we accomplish?

ESnet6 Design (in a nutshell)

ESnet6 "Hollow" Core Architecture



Unprecedented visibility into the network through telemetry

Innovative edge, providing standard and custom services with programmability

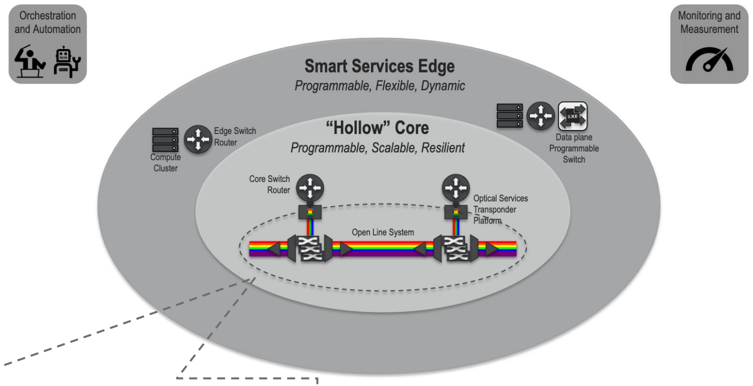
Simple network core focusing in minimal processing and high-speed data movement between the edges

Orchestration and automation framework to provide consistency, reliability and to change the paradigm on how networks are built and run

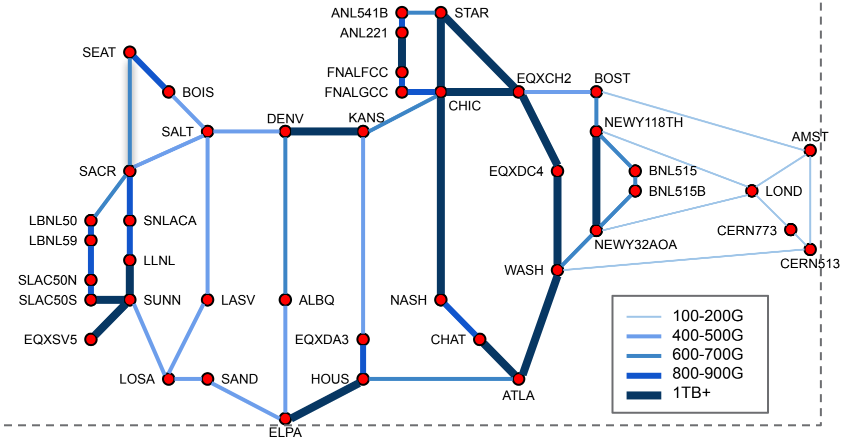
ESnet

ESnet6 Design and Build (in a nutshell)

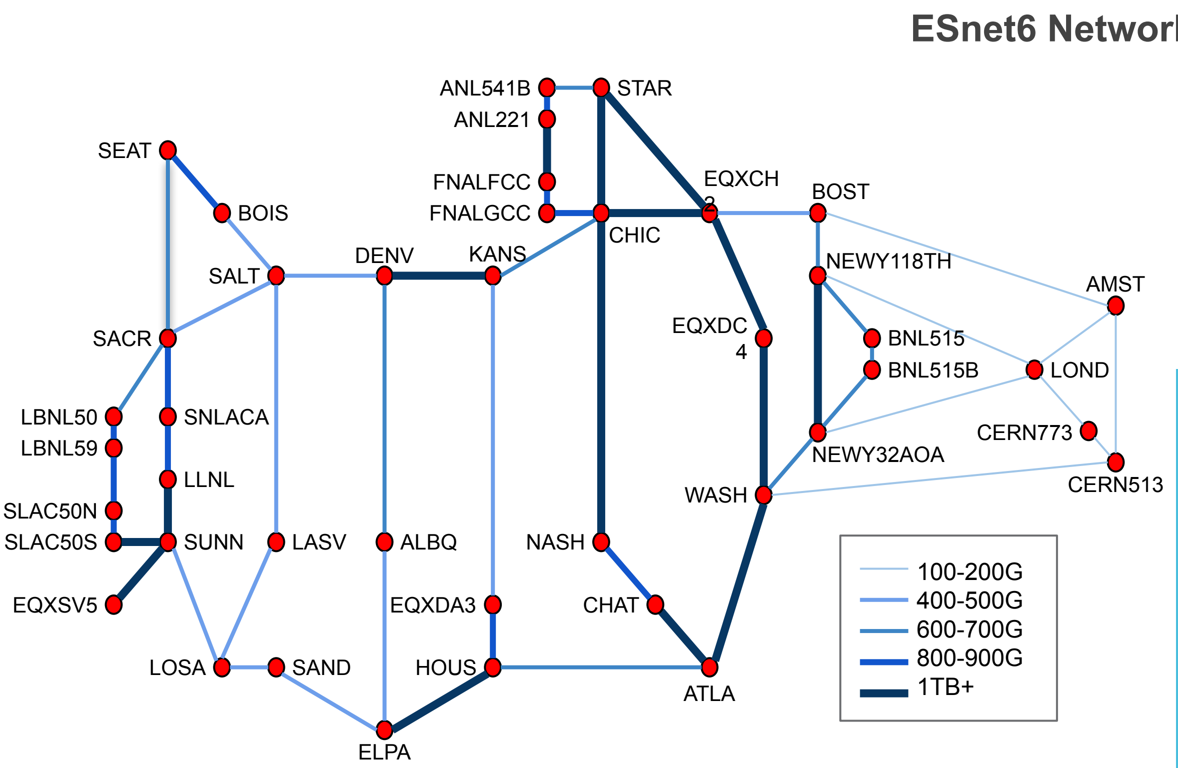
ESnet6 "Hollow" Core Architecture



ESnet6 Network Capacity (as of May 2022)



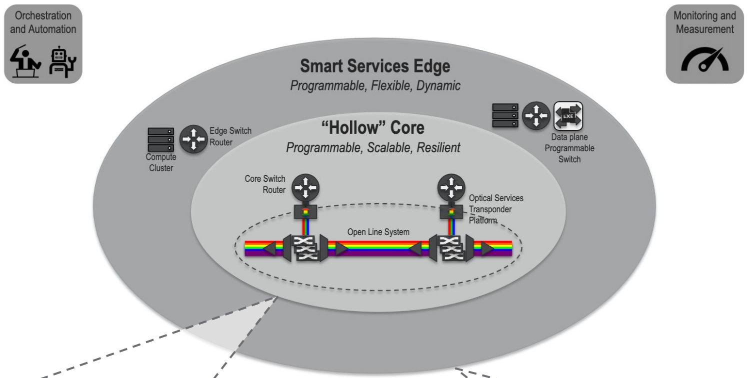
ESnet6 Design and Build (in a nutshell)



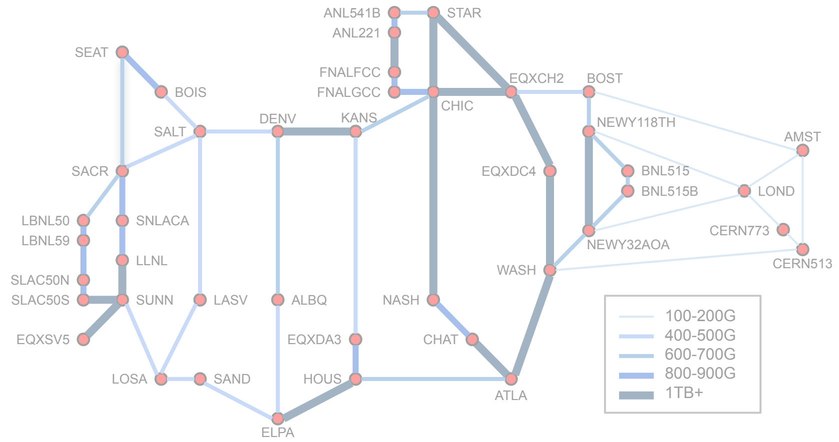
- **15,000 miles** of dark fiber lit up
- **300 leased spaces** installed across the US with ESnet owned equipment
- New fiber spans to **increase reliability and reduce latency**
- **46.1 Tbps** aggregate capacity deployed
- **400Gbps - 1 Tbps** services available

ESnet6 Design and Build (in a nutshell)

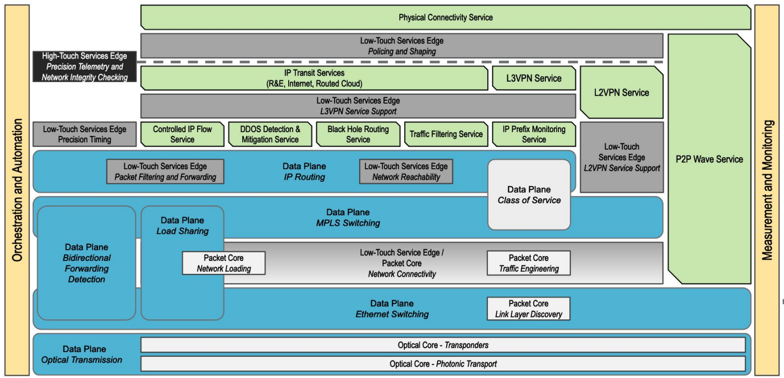
ESnet6 “Hollow” Core Architecture



ESnet6 Network Capacity (as of May 2022)

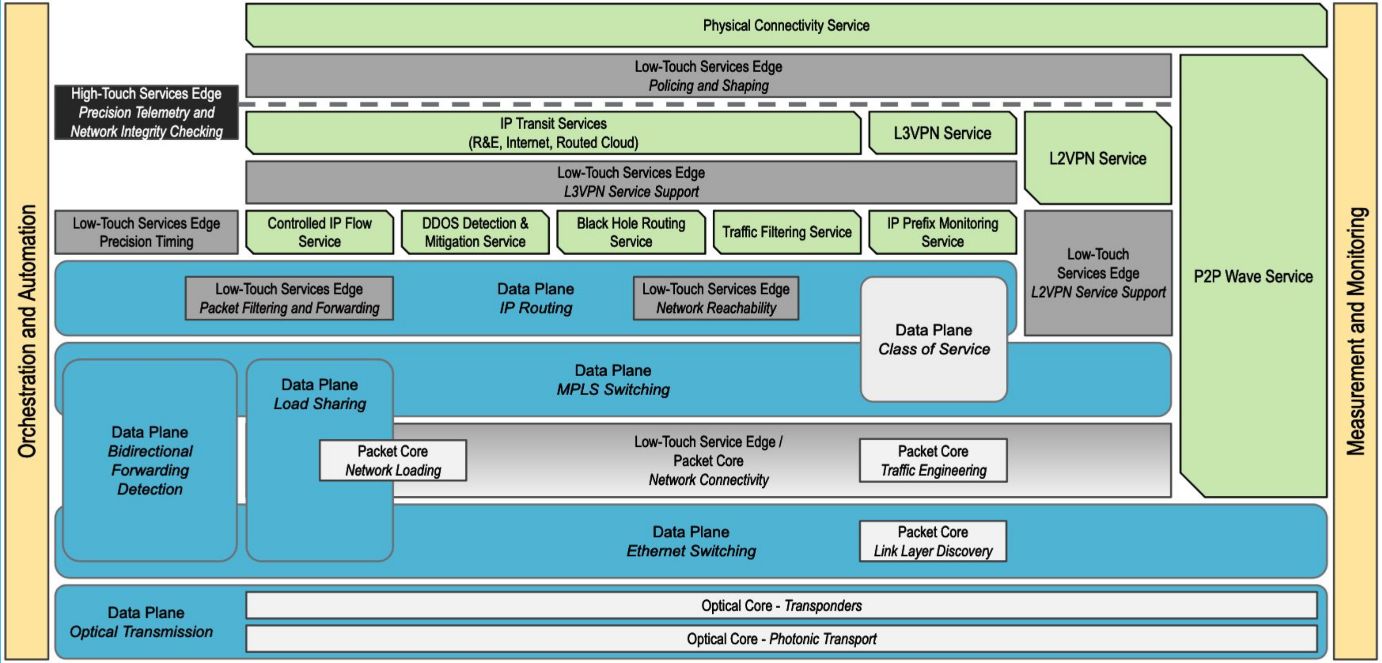


ESnet6 Services and Capabilities Structure



ESnet6 Design and Build (in a nutshell)

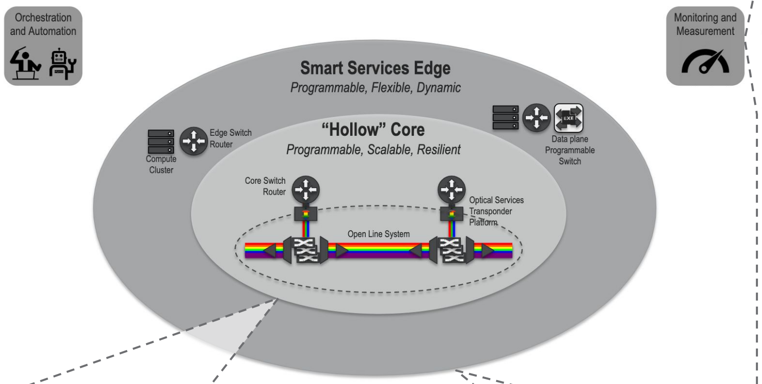
Well defined Services architecture, with a flexible set of services to support most common use-cases. These services were implemented using new packet platforms (routers)



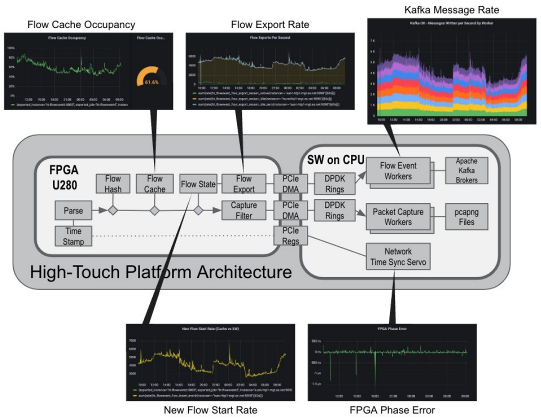
- All of the backbone routers replaced, provisioned, and in-service
- New routers at most of our connected sites
- Older routers decommissioned in parallel (and power / colo released)
- All of the core services have been developed and deployed
- Few aspirational services still under development

ESnet6 Design and Build (in a nutshell)

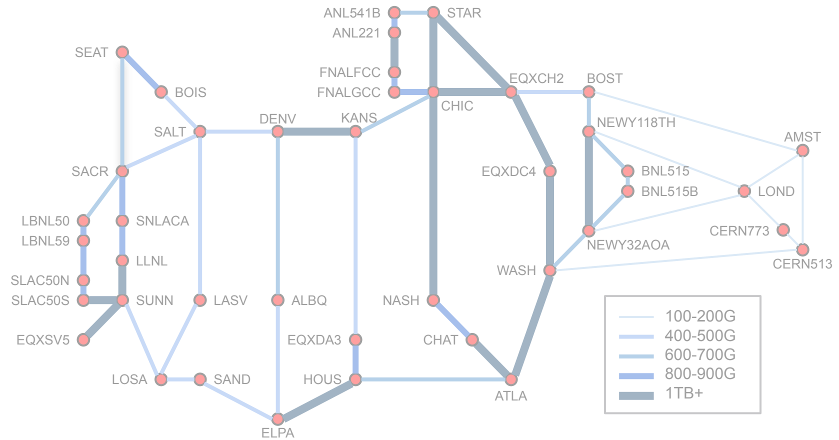
ESnet6 “Hollow” Core Architecture



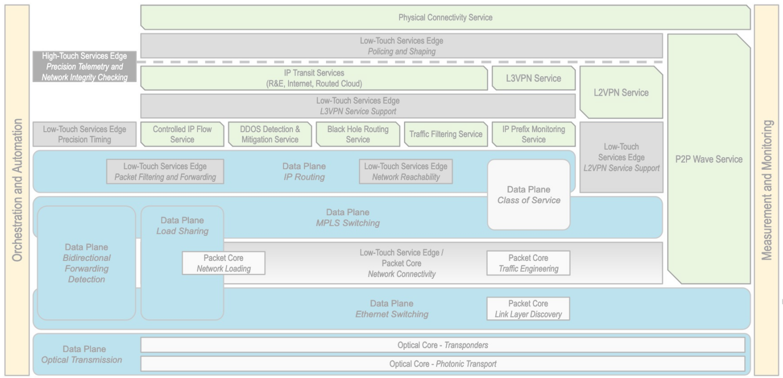
ESnet6 High-Touch Precision Network Telemetry Platform



ESnet6 Network Capacity (as of May 2022)

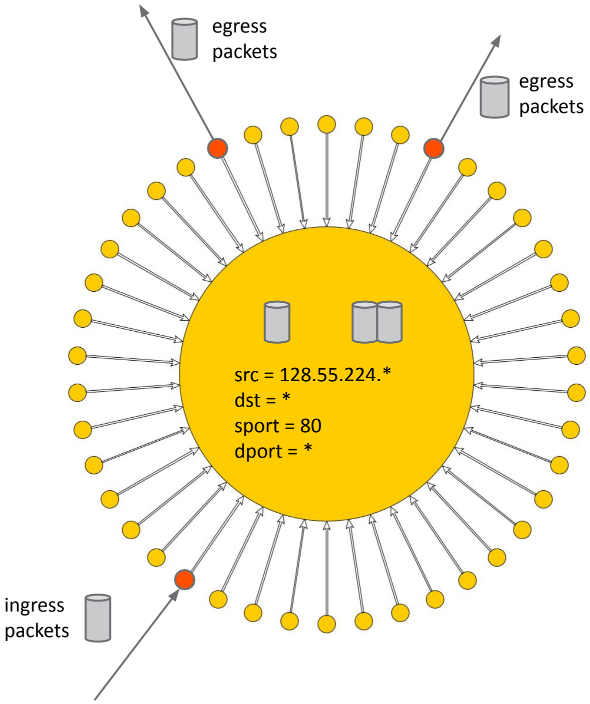


ESnet6 Services and Capabilities Structure



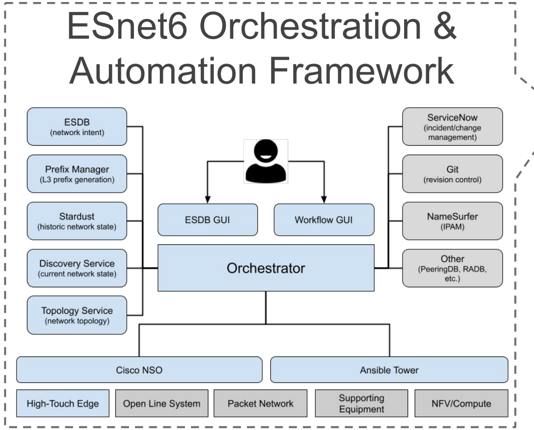
What is ‘High-Touch Precision Telemetry Platform’?

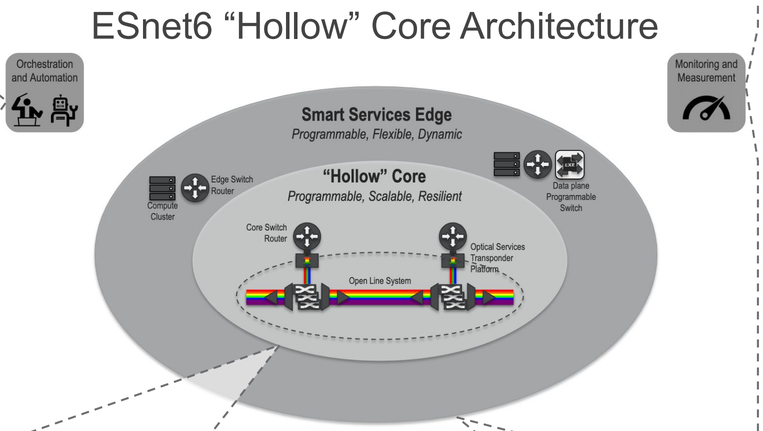
- Objective KPP Service
- Ability to choose flows we want to look at with this ‘packet microscope’
- High Touch HW (~SmartNIC) selects packets of interest out of the unsampled mirror stream and send them to software (no sampling unlike commercial platforms)
- Adds nanosecond precision timestamps on each packet
- Software writes packets into disk for future retrieval and merging with simultaneous captures across the entire footprint
- Users can see individual packets enter and leave the ESnet ASN even in the presence of asymmetric routing
- Many services can be built on this platform, in ESnet’s plan for the future

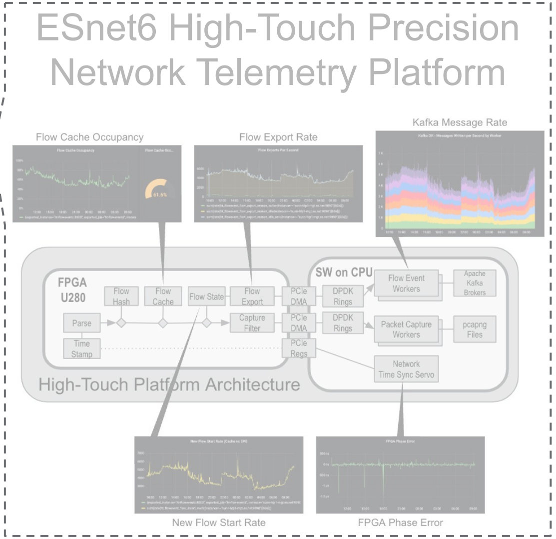


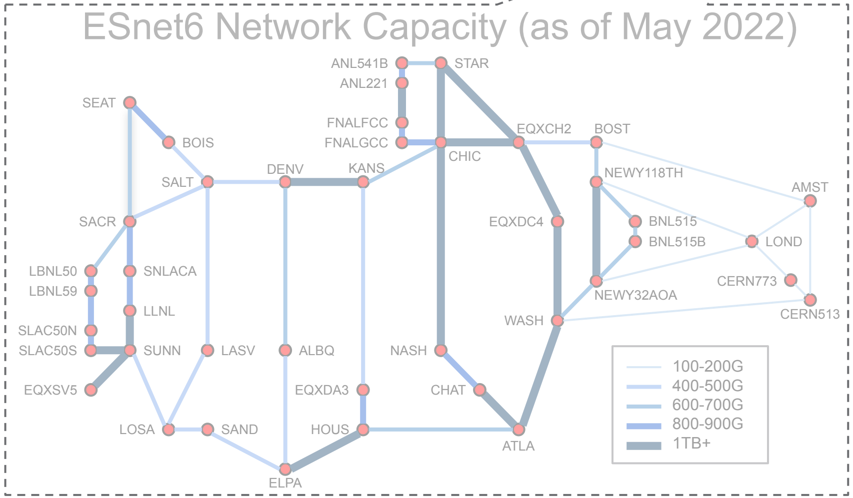
ESnet

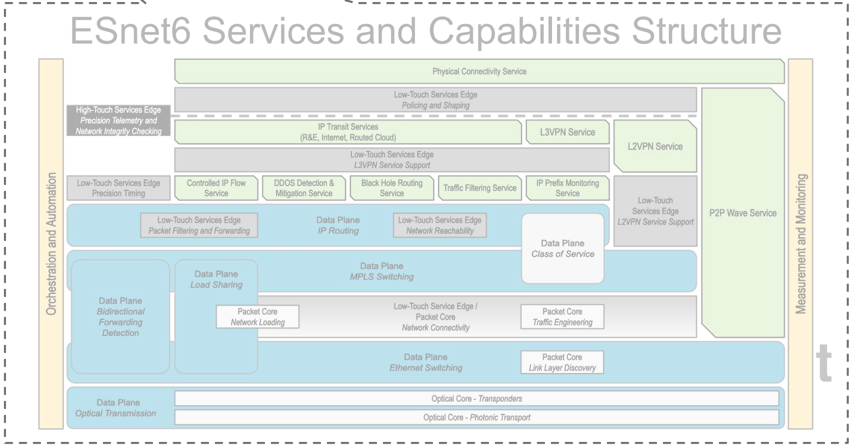
ESnet6 Design and Build (in a nutshell)











Sidebar:
In the before times…

- Network devices were treated as pets, not cattle
  - Engineers assigned to specific routers
- Documentation was completed after the work was done
- Bespoke business processes, dictated by individual circumstances
- CLI* configuration and operations was commonplace and accepted
- Typical process was to "Cut and Paste" configuration
- Hand-tended scripts and basic automation, including use of complex Jinja2 templating system

ESnet

Command Line Interface - common way to configure routers

Orchestration and Automation Goals

- Consistent configurations for complex services
- Consistent method for service deployment and ongoing management
- Reduce probability of human error
- Enhance network reliability
- Enable engineers to focus on more design than deployment (less busy-work)

Note: ***Orchestration is not a replacement for Humans***

Expansive vision for ESnet software stack, with limited scope planned for ESnet6 project

ESnet6 Software Components

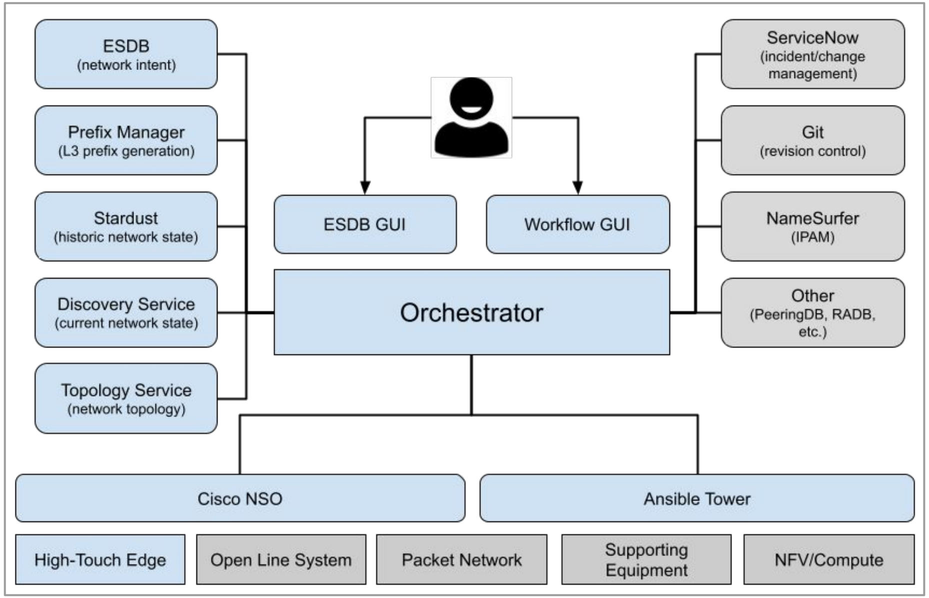
Assurance
- Fault management
- Root cause analysis
- Incident management

Provisioning
- Workflow orchestration
- Automated provisioning

Security
- Audit logging
- Black hole routing
- IP prefix monitoring

Analytics
- Time series data
- Streaming telemetry
- High touch telemetry
- Traffic planning

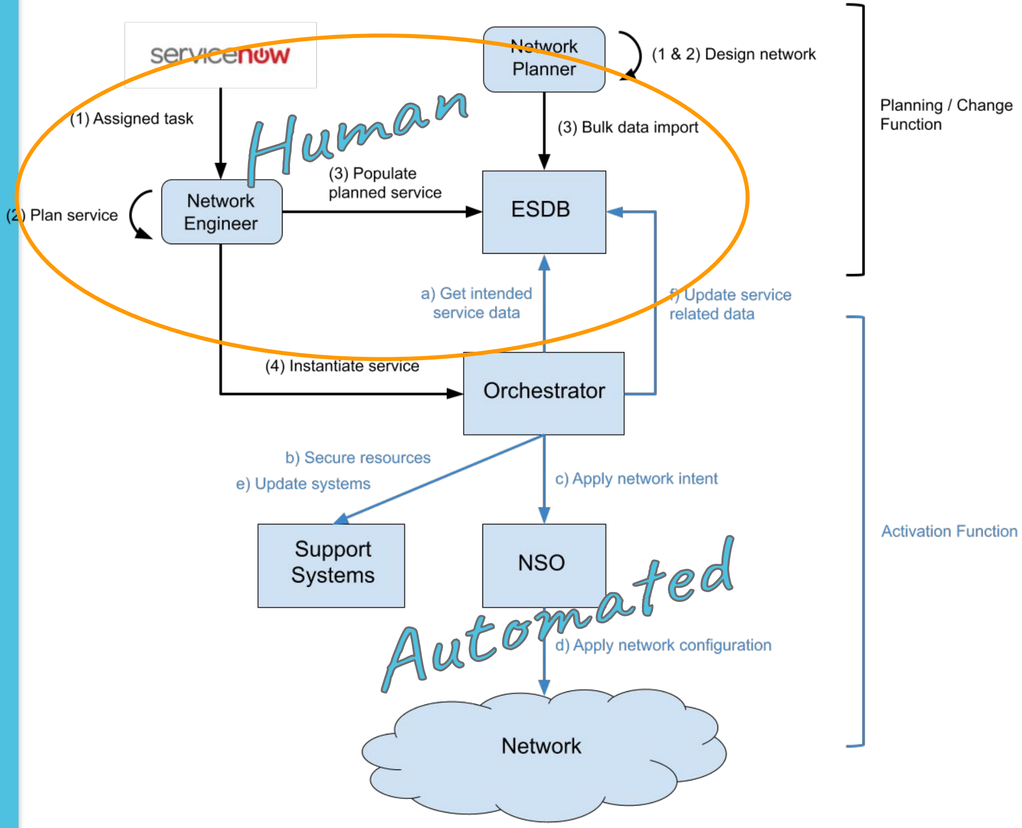
Unified Data Model
- Intent
- Discovery
- Topology

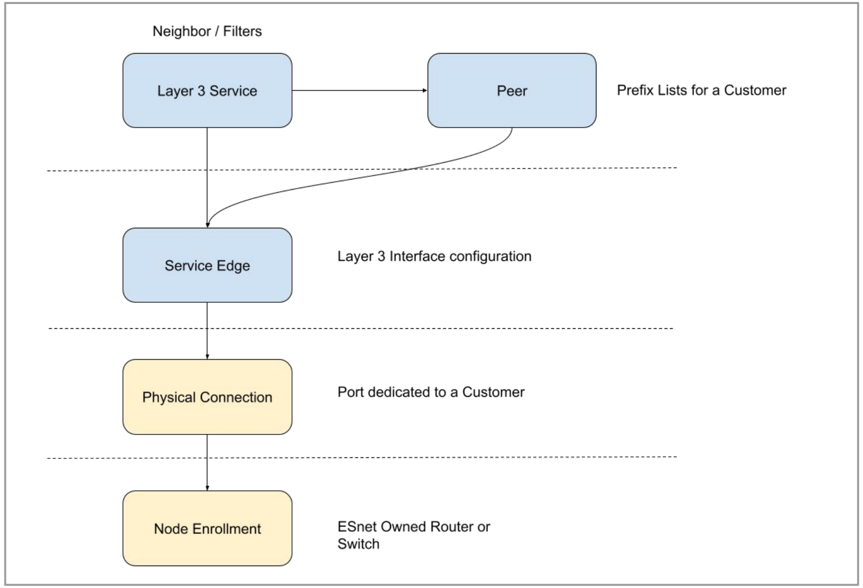


Provisioning Automation Arch.

ESnet

Most of router provisioning activities are using the automation stack to deploy services



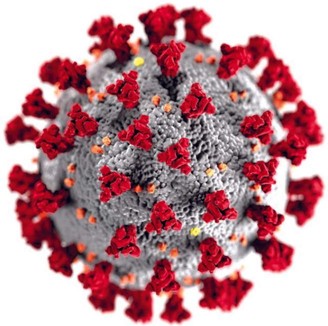


Automation and Services are composable and can stack to tackle complex activities

Lessons Learnt:
Formal Risk Management

Structured planning for problems like this

helped us manage problems like this!



- Expert training of all staff on formal risk management processes.
- Dedicated Risk Manager worked to identify, quantify, and develop mitigation strategies, and ensured we continuously updated and communicated about risks and issues throughout the project.
- All the early efforts paid huge dividends in the end.

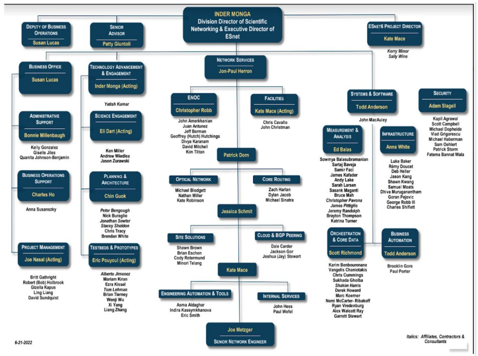
Lessons Learnt:
Team Growth!

Hiring and **onboarding** was critical to executing the project successfully!

We could never have completed it without our highly skilled, diverse, distributed team, with deep experience working and thriving in a virtual environment

2017
45 People

2022
120 People



Lessons Learnt:
DOE 413.3B: A growth opportunity!

DOE Office of Science allows *tailoring* the 413.3b processes.

In hindsight, tailoring the 413.3b process saved the project:

- Used Milestone Execution Index(MEI) metrics to measure and track progress, instead of Earned Value Management.
- Used Project Acceptance Memos (PAMs) to incrementally accept and retire scope from the project, and transition it to Program.

Working in this framework ensured that LBL Management, our Federal Project Director, Federal Program Manager, and other DOE leadership were 100% in sync with us and able to fully support us throughout the project.

Our entire organization now has a much better understanding of project management!

ESnet

Project accomplishments: Summary

- 15,000 miles of dark fiber lit up
- 300 locations around the US where ESnet equipment was installed and turned up
- 46.1 Terabits/second aggregate capacity installed
- ESnet5 decommissioned while ESnet6 was being installed without interruption in service
- > 70 new routers installed while decommissioning 53 of existing ESnet5 routers
- Automation framework built to help deploy and manage network services including security
- Innovation delivered that exceeds any commercially available functionality

How does this help Science?

Scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data.

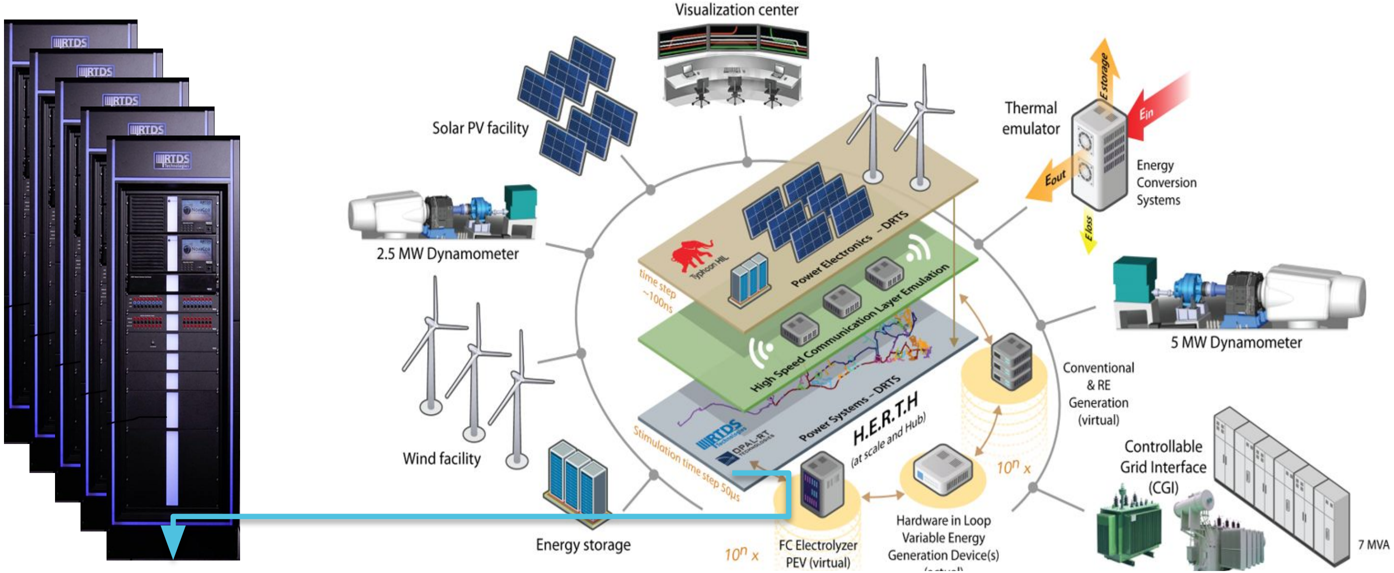
ESnet

NREL ARIES Objectives

1. Increasing variability in the physical size of new energy technologies

2. Controlling large numbers of interconnected devices

3. Integrating diverse technologies that have not previously worked together

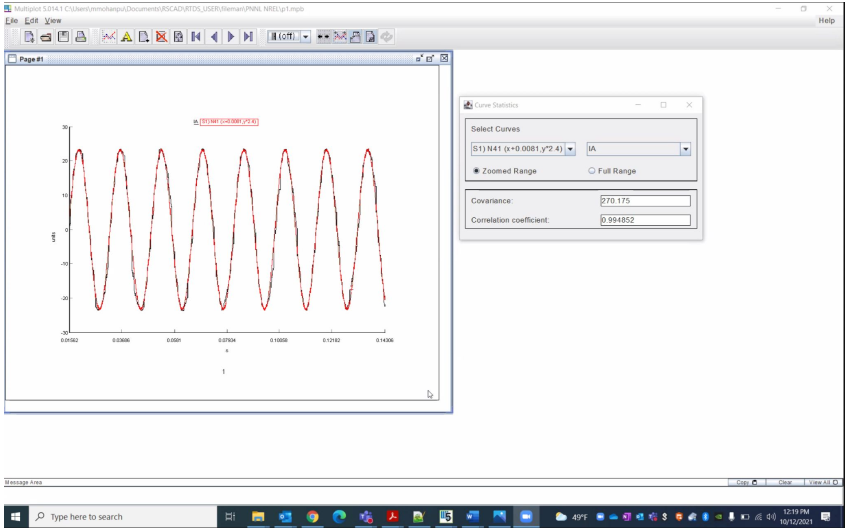


ESnet

ARIES Networking Challenges

- Distributed electrical grid control & simulation requires highly reliable, deterministic, **low latency, low jitter** connectivity.

- Very different network capabilities needed vs. high-throughput for many other ESnet applications (HEP, NP, etc)



Plot shows result from a ARIES test showing very low jitter - a sine wave was transferred from NREL to PNNL and then back to NREL - received data stream (black) shows very little difference from generated signal (red).

ESnet

Network Improvements from ESnet5 OSCARS to ESnet6 OSCARS

- Shorter path on new ESnet6 network
- Improved overall latency of network between instruments by ~30%
- Improved overall variability of network by ~70%

	Between Instruments		ESnet end to end Average Network Measurements			
	Round-Trip Time (RTT)	RTT Diff	Round-Trip Time (RTT)	RTT Max	RTT Min	RTT Diff
Start	37	4.6	32.4	32.5	32.3	0.2
Finish	23.9	1.207	22.693	22.726	22.671	0.055
Network Improvement	35.41%	73.76%	29.96%	30.07%	29.81%	72.50%

ESnet

Next Steps

In FY 23 – **10,000+ devices will be connected with real-time simulators in DOE National Lab Complex** to **provide a platform to emulate large area power systems**

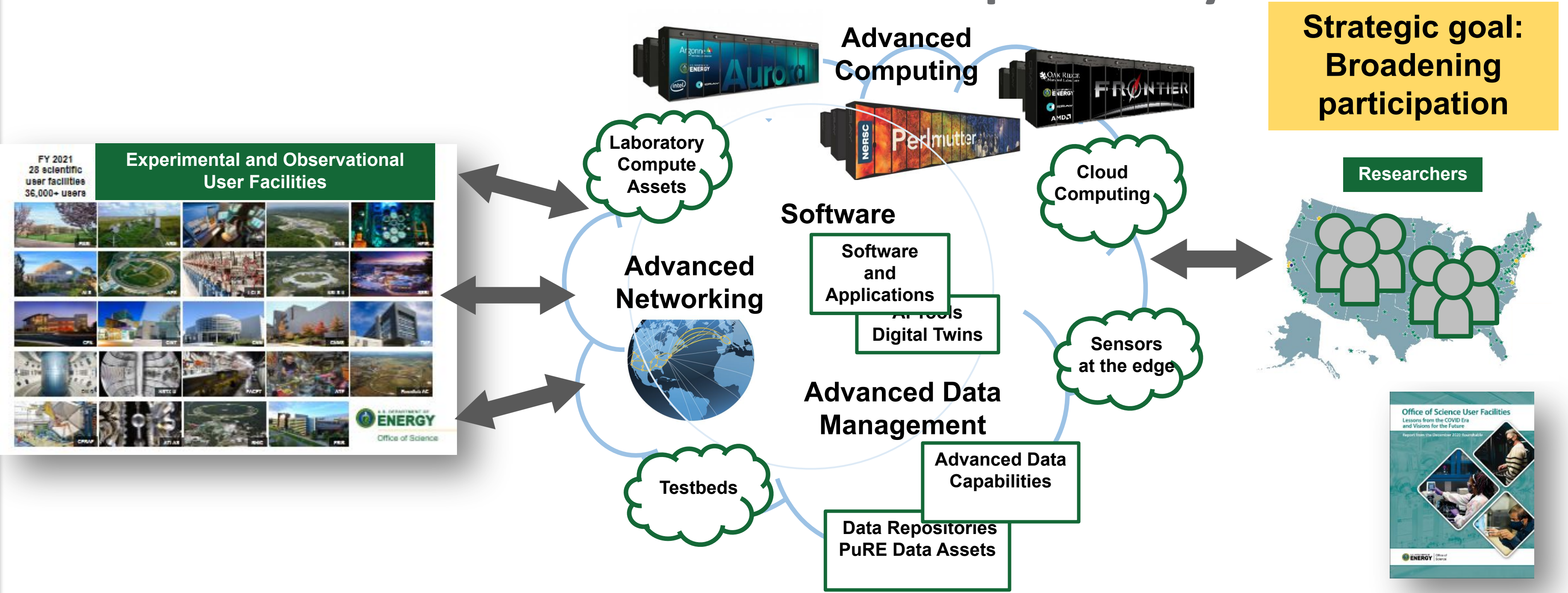
10,000 Devices cluster

Federated Real-Time Simulators with Power Systems Models

OBJECTIVE OF THE PLATFORM

Simulate a power systems extreme event of national consequences [major hurricanes, or cyber-attacks]

The vision: A DOE/SC **integrated research ecosystem** that transforms science via seamless interoperability



New modes of integrated science

Rapid data analysis and steering of experiments	Novel models for multi-facility allocation/utilization	AI-enabled insight from dynamic, vast multi-modal data	Seamless user interconnectivity via federated IDs

What does this mean for networks*?

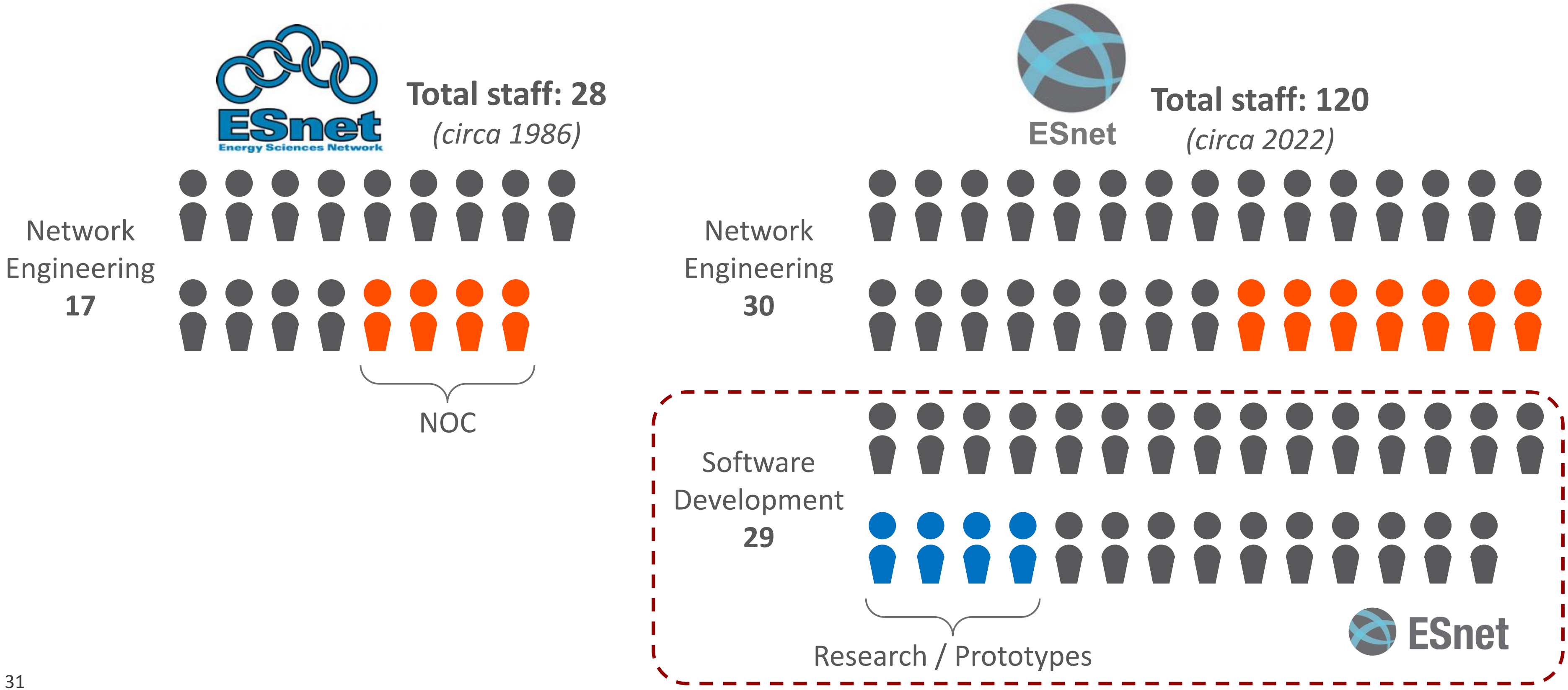
Promoting networks as “first class” resources, similar to instruments, compute and storage, e.g.,

- Accessible
  - Security frameworks for accessing (selected) services
  - APIs to interact with services
- Controllable
  - Resource/service selection/negotiation
  - Service scheduling
- Transparent
  - Resource (general) availability
  - Service (specific) status
- Adaptable
  - Ability to integrating compute and/or storage into the network
  - Rapid prototyping of new services

Networking is an end-to-end service, inter-domain interoperability and service consistency is critical!

ESnet

Due to ESnet6, team well positioned to tackle IRI challenges
(Growing emphasis on software, workflows and orchestration)




Sunset for ESnet6 project, Sunrise for ESnet
ESnet6 is now simply, ESnet.

Join us in October for #ESnet6Week!

Oct 11: ESnet6 - Unveiling Ceremony (free, online)

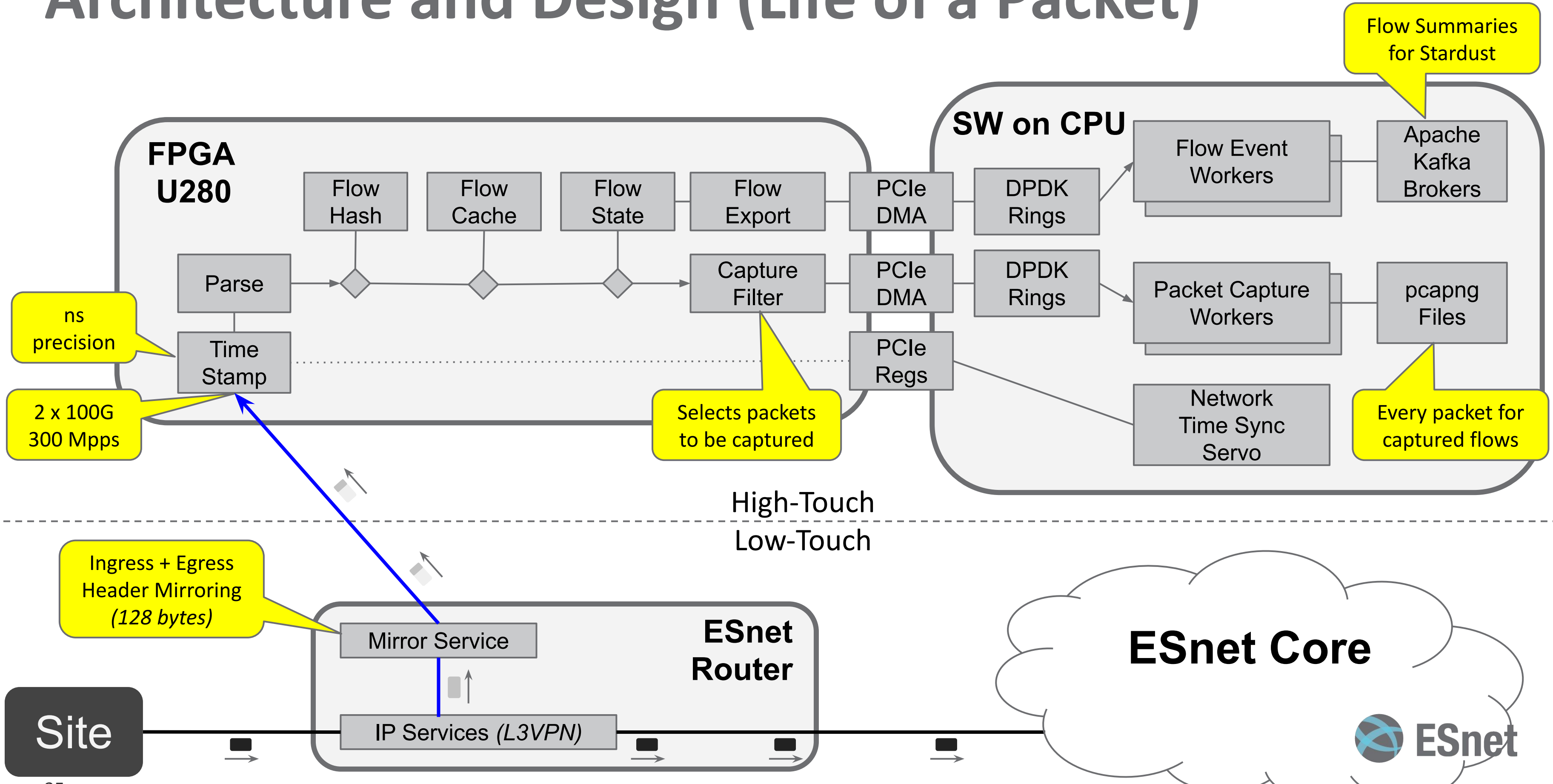
Oct 12-13: Confab22, our first user meeting (online & in-person)

Register here!

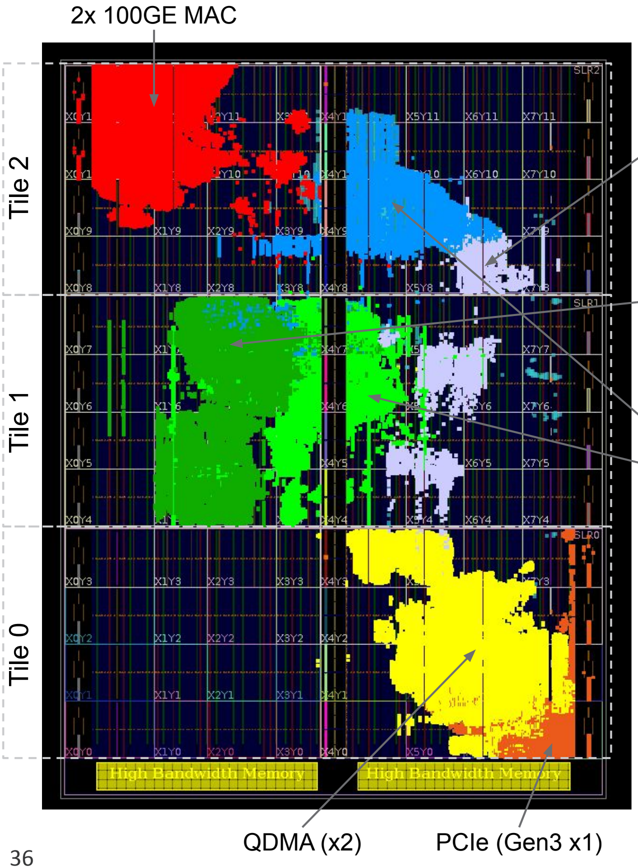
Thank you!

ESnet

Architecture and Design (Life of a Packet)



High-Touch Integrated FPGA Logic Blocks form the SmartNIC



Xilinx Open NIC Shell (open source)

- Provides pin mappings, CMAC + PCIe/DMA interfaces
- ESnet was a pre-release user and provided user feedback

Xilinx SDNet (P4 program -> logic)

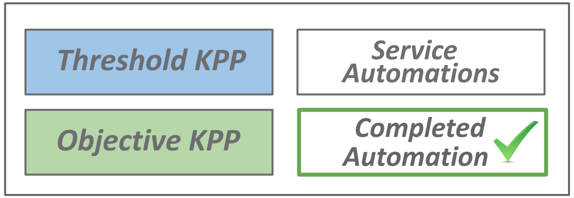
- Packet parsing, table lookups, packet filtering, packet edits
- Compiles a user-provided P4 program into FPGA logic

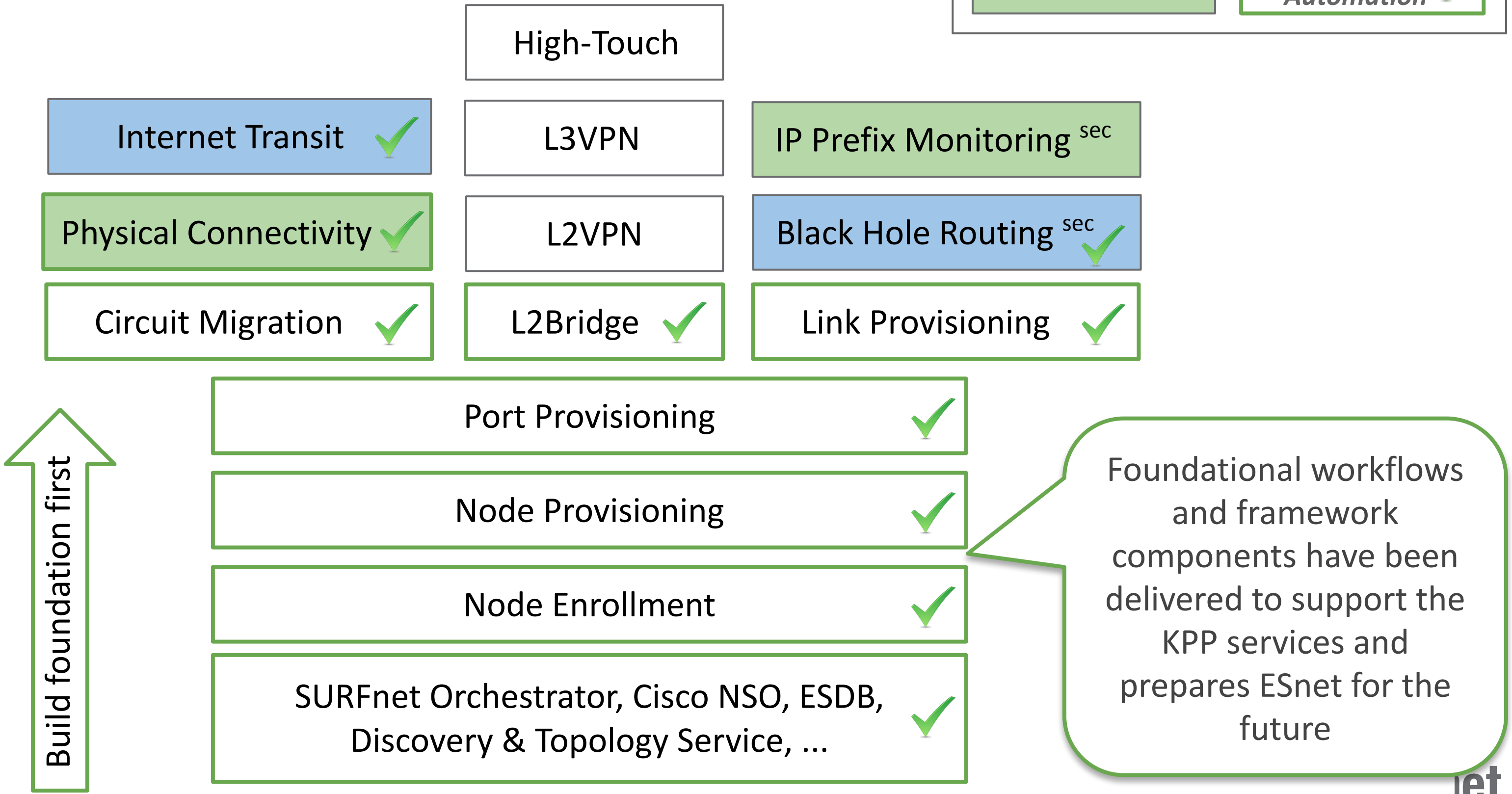
ESnet Custom Logic

- Processes 100% of the packet headers on the wire
  - **2x100G (or 300 million packets per second)**
- Per-Flow state tracking block (new function for P4 program)
  - Unsampled packet/byte counts
  - Packet size histograms
- PCIe register interfaces
- (Room for more stuff!)

ESnet

ESnet6 Design and Build (in a nutshell)

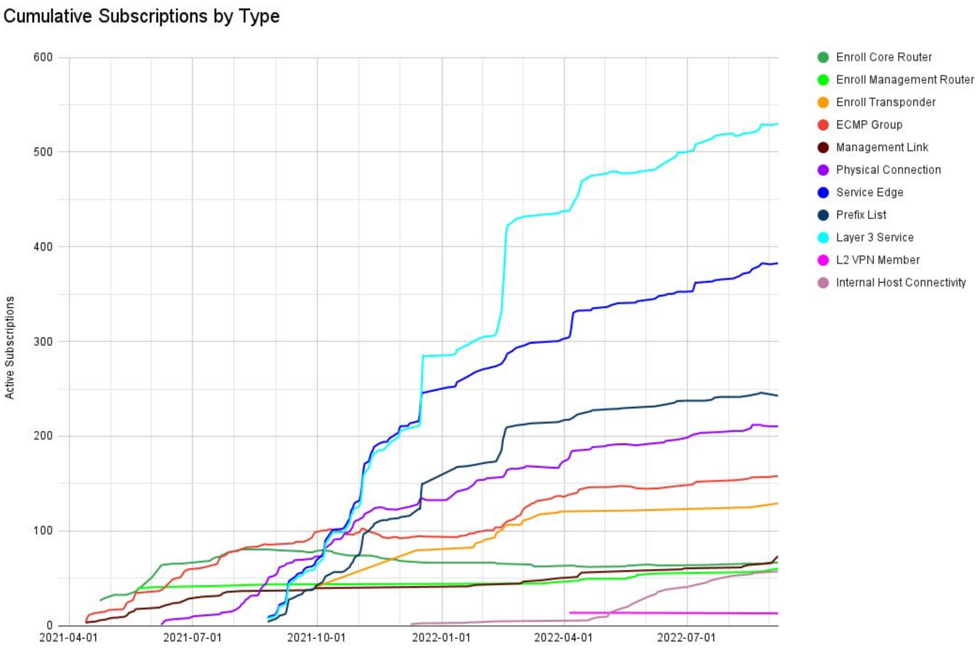




technology evolution needs pairing with culture change

Subscription Counts

ECMP Group	157
Management Link	68
Enroll Core Router	86
Enroll Management Router	72
Enroll Transponder	131
Physical Connection	213
Service Edge	383
Prefix List	245
Layer 3 Service	536
L2 VPN Member	13
Internal Host Connectivity	65



ESnet

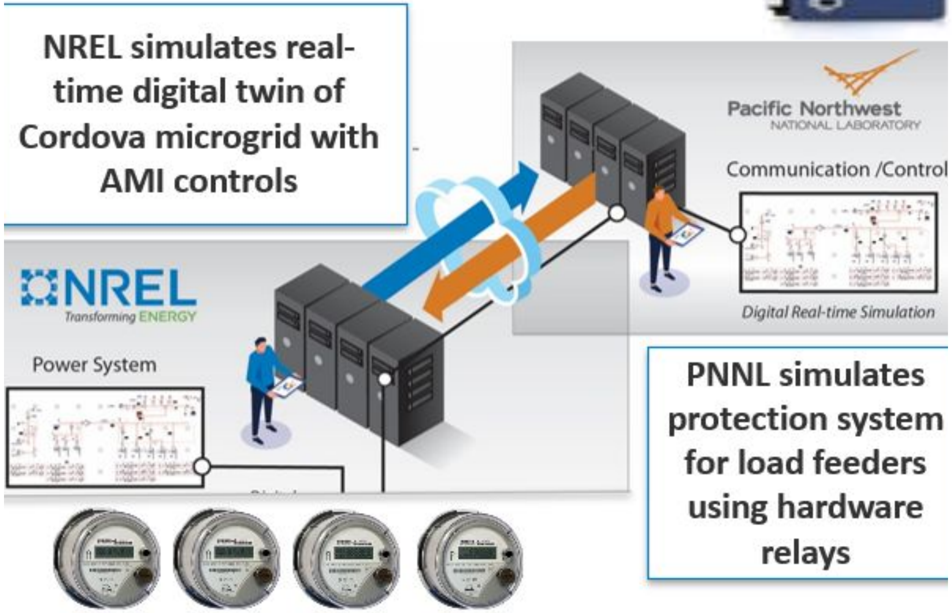
NREL-PNNL successful demo in December

WITH AMI

1. Hydro plant trips □
2. AMI sheds non-critical load □
3. Critical Load Served

WITHOUT AMI

1. Hydro plant trips □
2. Protection relay trips □
3. Critical Load Lost



AMI-based load control prevents power disruption to critical loads (airport, hospital, coast guard, etc.)

ESnet

DOE ASCR IRI Task Force contemplated operational models and guiding principles.

ASCR Integrated Research Infrastructure Task Force — March 8, 2021

Toward a Seamless Integration of Computing, Experimental, and Observational Science Facilities: A Blueprint to Accelerate Discovery

About the ASCR Integrated Research Infrastructure Task Force

There is growing, broad recognition that integration of computational, data management, and experimental research infrastructure holds enormous potential to facilitate research and accelerate discovery.[1] The complexity of data-intensive scientific research—whether modeling/simulation or experimental/observational—poses scientific opportunities and resource challenges to the research community writ large.

Within the Department of Energy's Office of Science (SC), the Office of Advanced Scientific Computing Research (ASCR) will play a major role in defining the SC vision and strategy for integrated computational and data research infrastructure. The ASCR Facilities provide essential high end computing, high performance networking, and data management capabilities to advance the SC mission and broader Departmental and national research objectives. Today the ASCR Facilities are already working with other SC stakeholders to explore novel approaches to complex, data-intensive research workflows, leveraging ASCR-supported research and other investments. In February 2020, ASCR established the Integrated Research Infrastructure Task Force[2] as a forum for discussion and exploration, with specific focus on the operational opportunities, risks, and challenges that integration poses. In light of the global COVID-19 pandemic, the Task Force conducted its work asynchronously from April through December 2020, meeting via televideo for one hour every other week. The Director of the ASCR Facilities Division facilitated the Task Force, in coordination with the ASCR Facility Directors.

The work of the Task Force began with these questions: Can the group arrive at a shared vision for integrated research infrastructure? If so, what are the core principles that would maximize scientific productivity and optimize infrastructure operations? This paper represents the Task Force's initial answers to these questions and their thoughts on a strategy for world-leading integration capabilities that accelerate discovery across a wide range of science use cases.

B. Brown, C. Adams, K. Antypas, D. Bard, S. Canon, E. Dart, C. Guok, E. Kissel, E. Lancon, B. Messer, S. Oral, J. Ramprakash, A. Shankar, T. Uram, <https://doi.org/10.2172/1863562>

"Our vision is to integrate across scientific facilities to accelerate scientific discovery through productive data management and analysis, via the delivery of pervasive, composable, and easily usable computational and data services."

Areas

- AL Allocations
- AC Accounts
- DA Data
- AP Applications
- SC Scheduling
- WF Workflows
- PB Publication
- AR Archiving

Principles

Flexibility. Assembly of resource workflows is facile; complexity is concealed

Performance. Default behavior is performant, without arcane requirements

Scalability. Data capabilities without excessive customizations

Transparency. Security, authentication, authorization should support automation

Interoperability. Services should extend outside the DOE environment

Resiliency. Workloads are sustained across planned and unplanned events

Extensibility. Designed to adapt and grow to meet unknown future needs

Engagement. Promotes co-design, cooperation, partnership

Cybersecurity. Security for facilities and users is essential.

Brown et al