ESnet: DOE’s High Performance Network (HPN) Scientific User Facility

**Mission of DOE Office of Science:**
Delivery of scientific discoveries and major scientific tools to transform our understanding of nature.

**Mission of Energy Sciences Network:**
*Science network user facility* designed to **accelerate** scientific research and discovery.
Optimized for Enabling Big-Data Science

Provides connectivity to all of the DOE labs, experiment sites, & user facilities (> 34417 users)
ESnet is a Science Mission Network

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

Mission: To Enable and Accelerate Scientific Discovery by Delivering Unparalleled Network Infrastructure, Capabilities, and Tools.
Each Major Upgrade Transforms the Facility to Meet Science Needs with Innovative, Cutting Edge Technologies

ESnet (Gen 3) 2005
Asynchronous Transfer Mode technology (new, now obsolete), provided by Carrier Sprint

ESnet (Gen 4) 2011
Traditional IP Routing with ESnet innovation OSCARS (new) to build Science Data Network (SDN)

ESnet (Gen 4) 2011
Transatlantic network

ESnet (Gen 5) 2021
Shared optical substrate (New) with traditional IP routing. Acquisition of Dark Fiber resource

ESnet (Gen 6) 2027
Innovative architecture with capability to meet capacity, reliability, and flexibility mission need

2005: Transatlantic network

2014:

2027:

5
ESnet Services Definition Process

- Requirements gathered from (6) DOE Office of Science program office requirements workshops.
  - Advance Scientific Computing Research (ASCR)
  - Basic Energy Sciences (BES)
  - High Energy Physics (HEP)
  - Biological and Environment 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 Capacity Projections

Log scale

62% Year-on-Year Growth

103 PB – Mar ‘19

ESnet Accepted Traffic: Jan 1990 - Mar 2019 (Log Scale)

Projected volume for Mar 2020: 405.7 PB
Actual volume for Mar 2019: 103.1 PB

Log scale
ESnet6 Design Objectives

1. **No single point of failure within the ESnet network backbone**
   - Fiber topology of interconnected rings.
   - HA hardware and software, e.g., redundant routing engines, switching fabrics, power supplies, automated job migration within a compute cluster.
   - SRLG-like connection ideology.

2. **Agility to add and/or move bandwidth capacity**
   - Dynamic bandwidth management at different network layers, e.g., wave provisioning at L1, TE tunnels at L2/2.5.

3. **Fine-grained traffic engineering**
   - Explicit hop paths per TE tunnel, distinct set of TE tunnels per service instance.

4. **Comprehensive automated management**
   - “Zero” touch provisioning, platform for intelligent decision making (e.g., AI).

5. **Highly programmable and flexible services**
   - Velocity and flexibility to develop new services unconstrained by vendor or ASIC-based limitations.
ESnet6 R&D Phase: Architecture and Technologies Matrix

Orchestration and Automation

Layer 3
- Router and DWDM Ethernet Switch Architecture
- Packet Transport Router Architecture
- Router and OTS Architecture
- Router and PKT/OTN OTS Architecture
- SDN Router and OTS Architecture
- SDN Router and PKT/OTN OTS Architecture
- Router and DWDM Ethernet Switch Architecture

Layer 2
- DWDM Ethernet Switches
- Transport Router DWDM
- SDN Routers
- SDN Switches
- PKT-OTN Optical Transport Systems
- P2P Optical Transport Systems

Layer 1
- Alien Wave Optical Transport
- Traditional Network Layer Separation
- Network Layer Integration
- Software Defined Networking
ESnet6 Hybrid Architecture Selection

1. **Layer 1**
   - Alien Wave Optical Transport
   - Open Line System

2. **Layer 2**
   - DWDM Ethernet Switches
   - Transport Router - DWDM

3. **Layer 3**
   - WAN Router Platforms

**Options**

- **(A)** Router and DWDM Ethernet Switch Architecture
- **(B)** Packet Transport Router Architecture
- **(C)** Router and OTS Architecture
- **(D)** Router and PKT/OTN OTS Architecture
- **(E)** SDN Router and PKT/OTN OTS Architecture
- **(F)** SDN Router and OTS Architecture

**Orchestration and Automation**
- Programmable Dataplane Switches / SmartNICs
- 3rd Party Transponders
- P2P Optical Transport Systems
- SDN Routers
- SDN Switches
ESnet6 ("Hollow-Core") Architecture Overview

**“Hollow” Core**
- **Programmable** – Software driven APIs to allocate core bandwidth as needed, and monitor status and performance.
- **Scalable** – Increased capacity scale and flexibility by leveraging latest technology (e.g. FlexGrid spectral partitioning, tunable wave modulation).
- **Resilient** – Protection and restoration functions using next generation Traffic Engineering (TE) protocols (e.g. Segment Routing (SR)).

**Smart Services Edge**
- **Programmable** – Software driven APIs to manage edge router/switch and retrieve telemetry information.
- **Flexible** - Data plane programmable switches (e.g. FPGA, NPU) in conjunction with compute resources to prototype new services (driven by Software Defined Networks (SDN)).
- **Dynamic** – Dynamic instantiation of services using SDN paradigms (e.g. Network Function Virtualization (NFV), Virtual Network Functions (VNF), service chaining).
Optical Core

Packet Core

Low-Touch Services Edge

High-Touch Services Edge
ESnet6 Optical Core Characteristics

- **High Bandwidth and Scalable Capacity**
  - Variable modulation rates (e.g., 100-400G waves) and Flex-grid for spectral efficiency.
  - Designed for full-fill (in C-band).
  - 100GE handoffs day-1, 400GE supported opportunistically (requires FlexO-ish solution).

- **Hardware and Provisioning Resiliency**
  - Topology of interconnected rings
  - High-availability and redundant hardware, with distribution of connections across redundant components.
  - SRLG-like provisioning of waves across disparate hardware.

- **Flexible**
  - Designed for colorless, directionless, and (limited) contentionless.
  - Open Line System with 3rd party transponder support.

- **Secure**
  - Isolated (walled-garden) management plane with protected access for 3rd party L1 NOC.

- **Programmable**
  - Automated provisioning, will require integrated 3rd party transponder and OLS management.
ESnet6 Dark Fiber Footprint

* Actual fiber distances based on OTDR Analysis Information
** Derived distances based on line-of-sight + 20%

Site Owned/Leased Dark Fiber
ESnet IRU Dark Fiber
Optical Add/Drop Locations
Optical Core

Packet Core

Low-Touch Services Edge

High-Touch Services Edge
ESnet6 Packet Core Characteristics

- **High-Bandwidth with Fine-grain Traffic Engineering**
  - (MPLS) Switching within the Packet Core, eliminating overhead of routing.
  - Distinct set of SR-TE LSPs associated to each service instance.
  - QoS for user traffic to support guaranteed bandwidth services.

- **Packet Layer Resiliency**
  - TI-LFA FRR will facilitate fast recovery for local failures.
  - Unrecoverable SR-TE LSP failures will result in fall back to ISIS-SR LSPs.
  - S-BFD will allow the LSP head-end router to quickly detect non-local failures and react accordingly.

- **Flexible**
  - Use of PCE to manage SR-TE LSPs provides for potential to support non-standard multi-constrained path finding (e.g., predictive usage).

- **Secure**
  - Separate forwarding tables and QoS queues for management and control traffic to prevent interaction with user traffic.

- **Programmable**
  - Automated provisioning and management of SR-TE LSPs per service instance.
ESnet6 Day-1 Planned Backbone Capacity (Jan 2021)

Jun 2021 (Optical) Bandwidth Capacity Plan

(Based on 29-year (adjusted) trend usage prediction analysis which includes Feb 2019 traffic data, and additional site input)
“High-Touch” vs “Low-Touch” Services (Packets PoV)

“High-Touch”
- Flexible packet manipulation.
- Customizable (programmable) forwarding lookup functions.
- Configurable packet pipeline based on programmable switch processor.

“Low-Touch”
- Simple forwarding and filtering functions.
- Constrained (fixed) forwarding lookup functions.
- Static packet pipeline based on fixed Application-Specific Integrated Circuit (ASIC) functions.
“High-Touch” vs “Low-Touch” Services (Packets PoV)

“High-Touch”
- Flexible packet manipulation.
- Customizable (programmable) forwarding lookup functions.
- Configurable packet pipeline based on programmable switch processor.

“Low-Touch”
- Simple forwarding and filtering functions.
- Constrained (fixed) forwarding lookup functions.
- Static packet pipeline based on fixed Application-Specific Integrated Circuit (ASIC) functions.

“No-Touch”
- Opaque forwarding*.

*NB: Specific to P2P Wave Service.
ESnet6 Low-Touch Services Edge Characteristics

• Flexible Service Instantiations
  – All users services will be supported by a distinct L2/L3VPN service instance with associated LSPs to enable per service traffic engineering.
  – Policing and shaping functions to support guaranteed bandwidth services.
  – Support of standards based protocols (e.g., eBGP) for external interaction.

• Secure
  – Distinct VRF instances per service prevents route leaks and contains issues like DDoS to within a service instance.
  – Packet filtering and forwarding functions to support security services such as traffic filtering or DDoS mitigation.

• Programmable
  – Automated provisioning of service instances based on workflow templates. This could include pushing down configurations on the Low-Touch routers in addition to spinning up VMs on the local compute cluster (e.g., for services like Zeek-on-the-WAN).
ESnet6 Network Services

Other Services
- Cloud Site
- Testbed
- Data Cache
- NDN
- Premium
- Flow Telemetry
- Upstream Blackhole
- IP Prefix Monitoring

Packet Routing, Discard and Filtering Services
- Black Hole Routing
- Traffic Filtering
- Traffic Redirection
- DDOS Detection & Mitigation
- Controlled IP Flow
- Rate Limiting
- Filtered Internet

Packet Forwarding Services
- Internet Transit
- Cloud Transit
- R&E Transit
- L3VPN

Frame Forwarding Services
- L2VPN P2P & MP
- Cloud Backhaul
- Transparent Point-to-Point Ethernet
- P2P Wave

Physical Connectivity Service
Optical Core
Packet Core
Low-Touch Services Edge
High-Touch Services Edge
ESnet6 High-Touch Services Edge Characteristics

- Highly Programmable Dataplane
  - Programmable packet pipeline for customized services.
  - Local compute resources available for (real-time) localized processing and analysis.
- Flexible Data Path Configuration
  - Out-of-path packet processing on duplicate packet stream received from the Low-Touch router using the copy-and-forward (e.g., port mirroring) functions.
  - Bump-in-the-wire packet processing can be done by service chaining High-Touch platform into the data path using redirection on Low-Touch router (e.g., L2 cross connects, SR-labels, etc).

ESnet views the High-Touch Services Edge as a strategic capability that will initiate new conversations, facilitate research collaborations, and enable novel services.
“High-Touch” vs “Low-Touch” Hardware

“High-Touch”
Programmable data-plane

Pros:
- Flexible to customize for specialized use cases

Cons:
- Complexity of designing / implementing specialized use cases
- Higher cost

“Low-Touch” (and “No-Touch”) Application-Specific Integrated Circuits (ASCI) based data-plane

Pros:
- Optimized for specific tasks
- Lower cost

Cons:
- Inflexible

- Xilinx Kintex UltraScale (FPGA+)
- Barefoot Torfino (P4)
- BarefExchip NP-4 (NPU)

- Broadcom Jericho
- Juniper Trio
- Nokia FP4
High-Touch *Precision Network Telemetry*

1. A 100 MHz Clock has a period of 10 ns:
   Any contemporary FPGA / ASIC can operate counters at this rate.
   A 100 Gbit serial bit-stream transmits 1,000 bits in 10 ns.
   A small 125 byte packet contains 1,000 bits.

   ![Diagram of 10 ns and 125 Byte Packets]

   With 10 ns resolution timestamps, we can distinguish individual packets.

2. The speed of light is 1 ns / foot.

   ![Diagram of 10', 20', and 30' Intervals]

   With 10 ns resolution timestamps, we can locate a packet within 10 feet.
   The 3,000 mile distance from Berkeley to New York represents 15 million feet, or 1.5 million packets on the optical cable.

   ![Diagram of Time vs. Distance]

   We can locate where our packets are on the wire, or inside a node.
   We can measure buffer fill levels up to the resolution of a single packet.
   We can make these measurements without modifying a single router or switch.
High-Touch Architecture Overview
1 Gbps iPerf flow - 600,000 packets

Per-packet rates reaching 10 Gbit/s (line rate of the sender)

Note: Average rate is calculated using a time-weighted average of per-packet rates.
1 Gbps iPerf flow - 1% packet drop

Per-packet rates maximum: 1 Gb/s

5 Mb/s average flow rate (100 pkt window)

Note: only 23 packets were dropped all together, taking bandwidth down to 5 Mb/s from 1 Gb/s.
Flow throughput drops when a retransmission happens
Orchestration and Automation
The ESnet(5) Operations Support System

Collection of software tools used by ESnet to manage daily network operations.

Grown organically over the life of ESnet.

New tools and processes incorporated as a need identified.

Consists of a mix of purchased, open source, and in-house developed software.
The five key components to a cohesive orchestration solution:

- Workflow management;
- Automated provisioning;
- Network intent;
- Network discovery;
- Network topology.
Automated Provisioning

- The automation of pre-defined procedures to programmatically change configuration within the network
- ESnet6 requires automated service provisioning across network, compute, and application resources
- Automation solutions are typically focused on domain specific provisioning tasks
- Model driven design and provisioning
  - Each network service is defined using a service model that is decomposed into domain specific provisioning operations.

Ansible, NSO, NMS
Questions...
Backup Slides