U.S. DEPARTMENT OF ENERGY

BERKELEY LAB

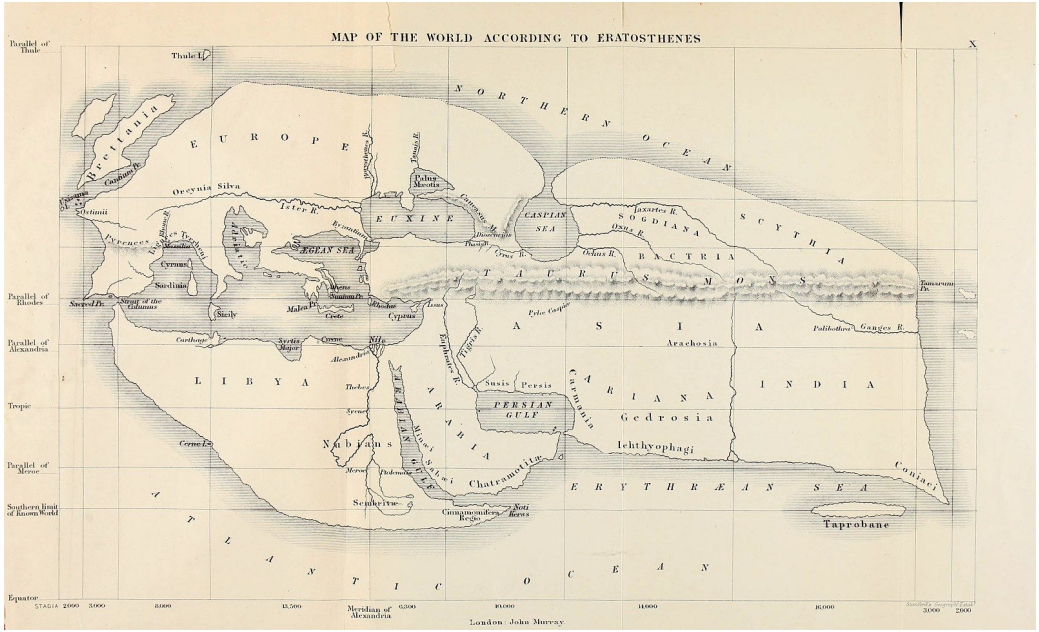
ESnet

Software Defined Network control for LHC Experiments

Justas Balcas
Software Engineer
ESnet

CHEP 2024
Krakow, Poland
2024-10-24

If a {human} was a packet, how did it travel?



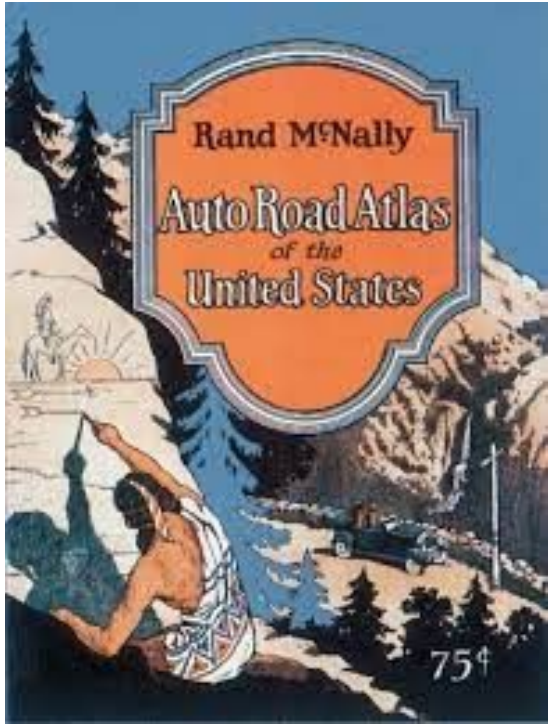
Map by
Eratosthenes of Cyrene
(276 B.C - ~194 B.C)

Father of Geography

Equivalent of hop-by-hop, store-forward routing

- Maps introduced rough guide on directions and location
- Tools helped to align to those directions
- Refinement of directions was based on observing intermediate landmarks or asking

If a {human+automobile} was a packet, how did it travel?



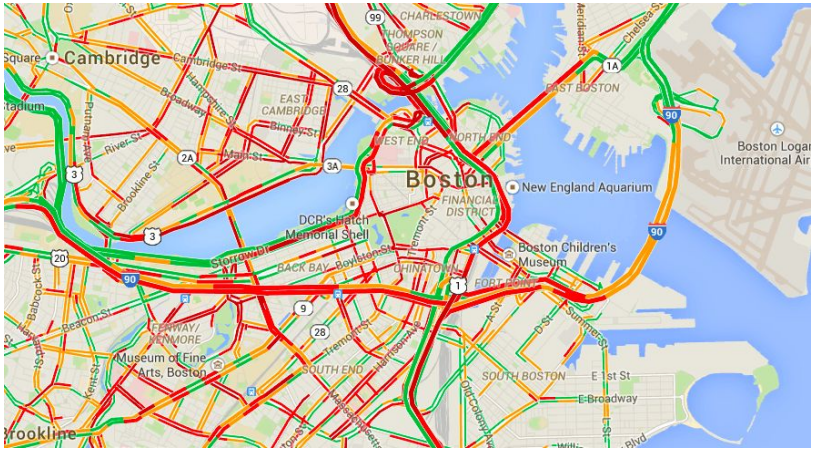
With the advent of automobiles, Rand McNally published its first road atlas called “Auto Chum” in 1924

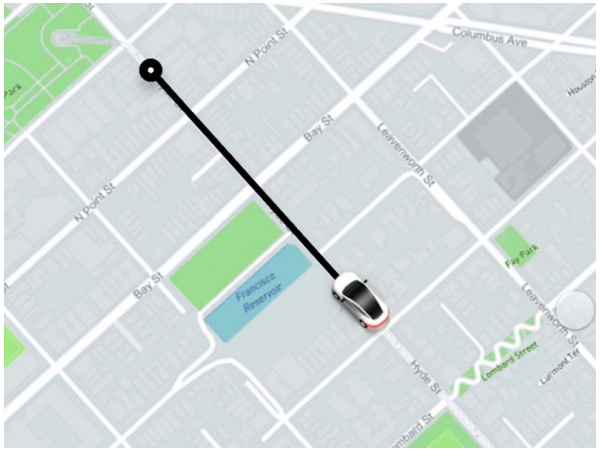
Routes were pre-computed by human brain before getting on the road, re-routing happened on the fly by stopping and manually determining the route again

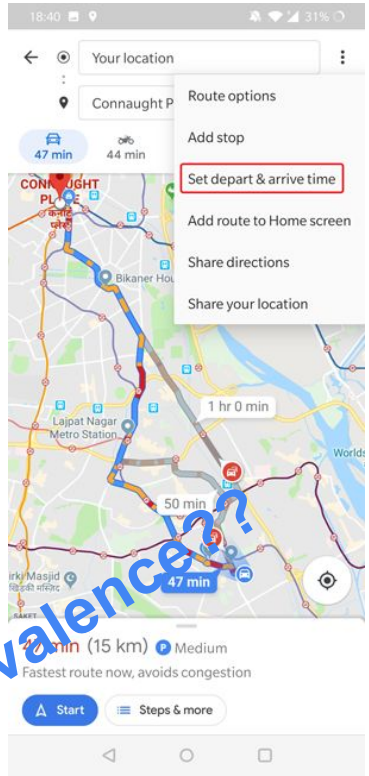
Equivalent of MPLS or Layer 2.5 based routing

Prediction and planning was hard, and depended on personal experience or hearsay

With the advent of digital technology, the {human + vehicle} packets have real-time + historical knowledge







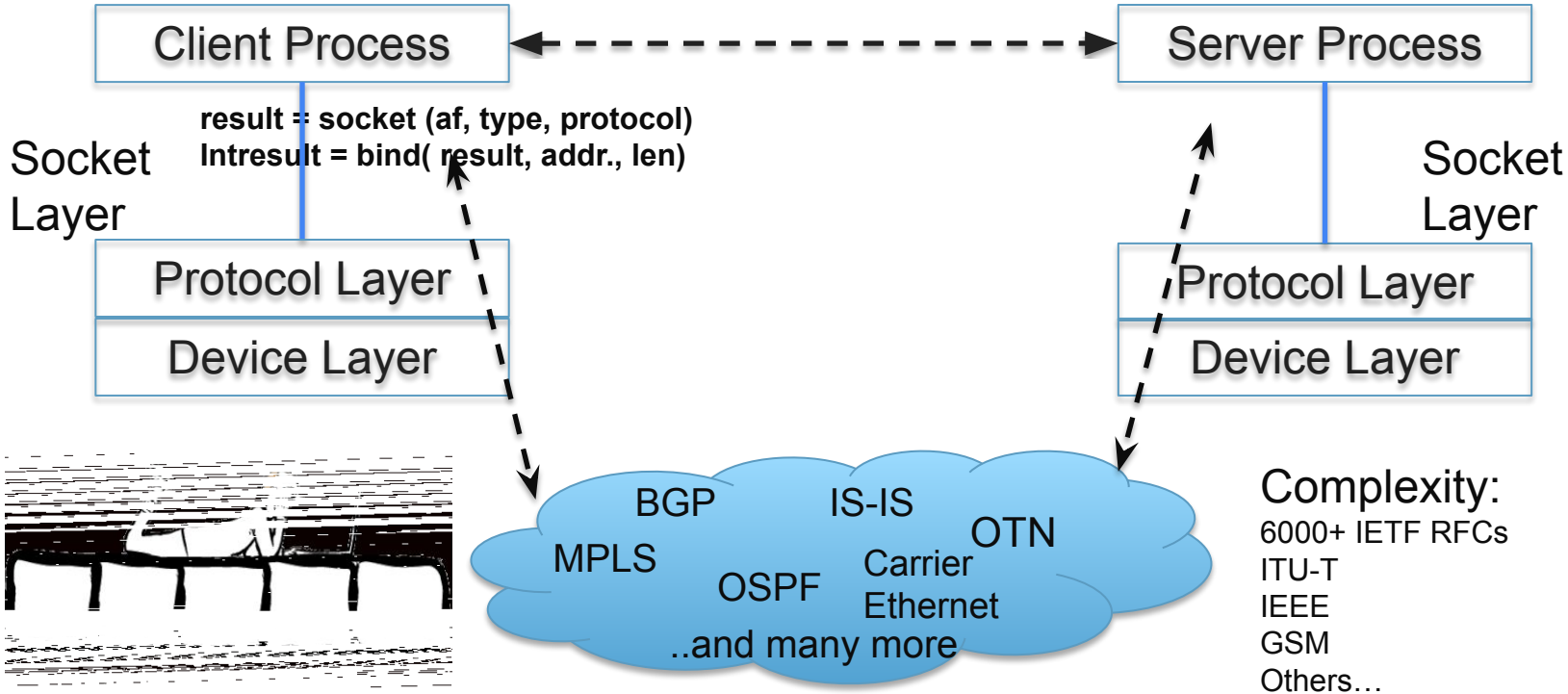
Real-time traffic and traffic prediction helps plan with just in time information, and features such as dynamic rerouting and updated accurate data on when the destination will be reached

Equivalence??

Inder Monga Keynote at 7th Rucio workshop (recording): https://indico.cern.ch/event/1343110/sessions/557886/#20241001

Aspirational Goal: How can we provide predictability and resilience to certain data flows given the huge variability of background traffic?

The Unix Socket Interface:

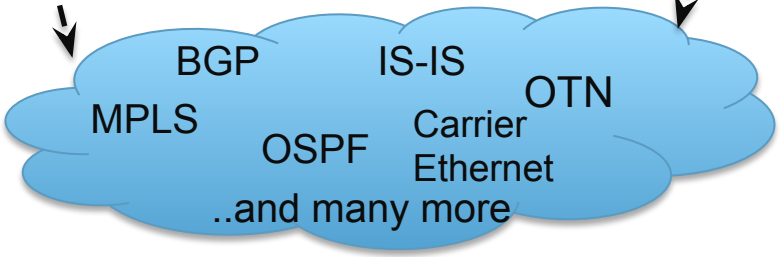


The Unix Socket Interface: Network became a “**black box**”

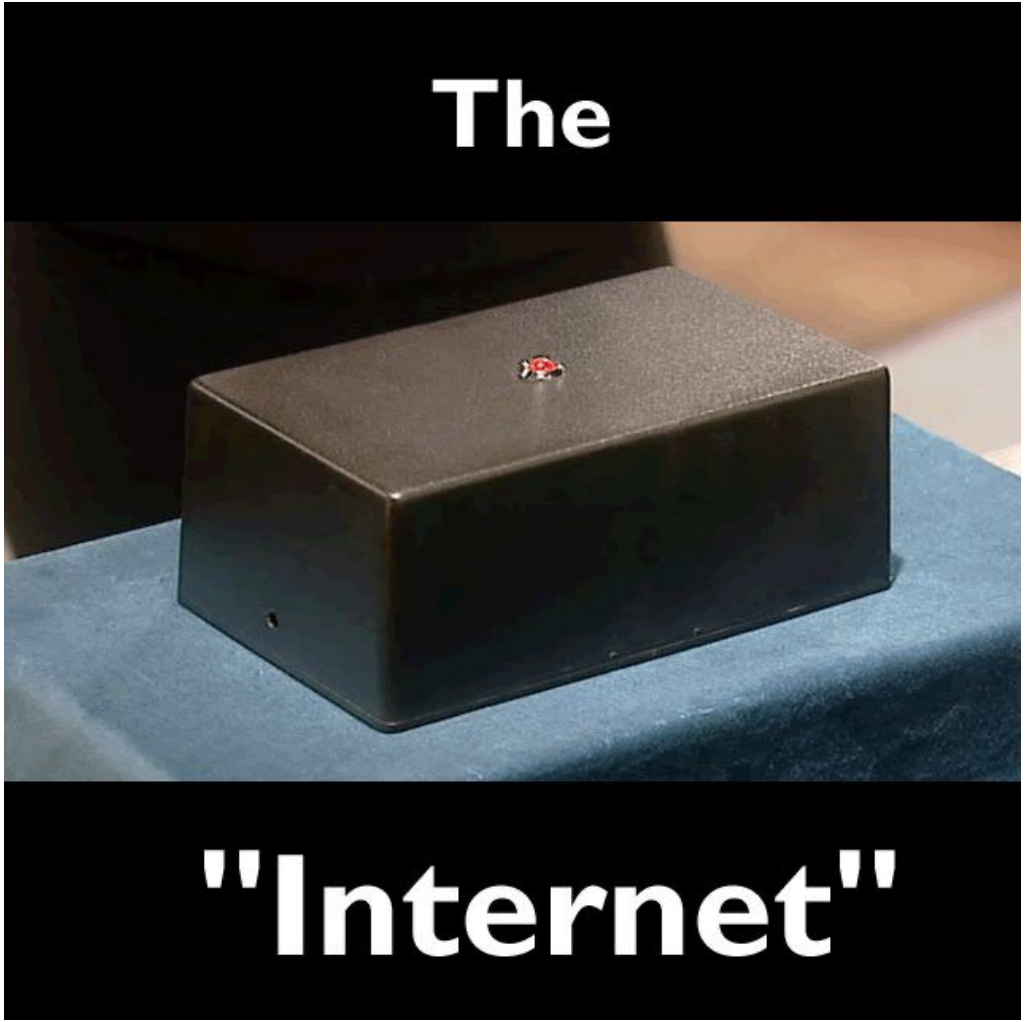
- Gives file system like abstraction to the network
- Hides the complexity of the network and its operations

result = socket (af, type, protocol)

- Application gets no feedback on the progress of the transfer
- There is no reasons given when a transfer fails, the only approach is try again, and again…..
- Network has no responsibility (unlike UPS or Amazon…)

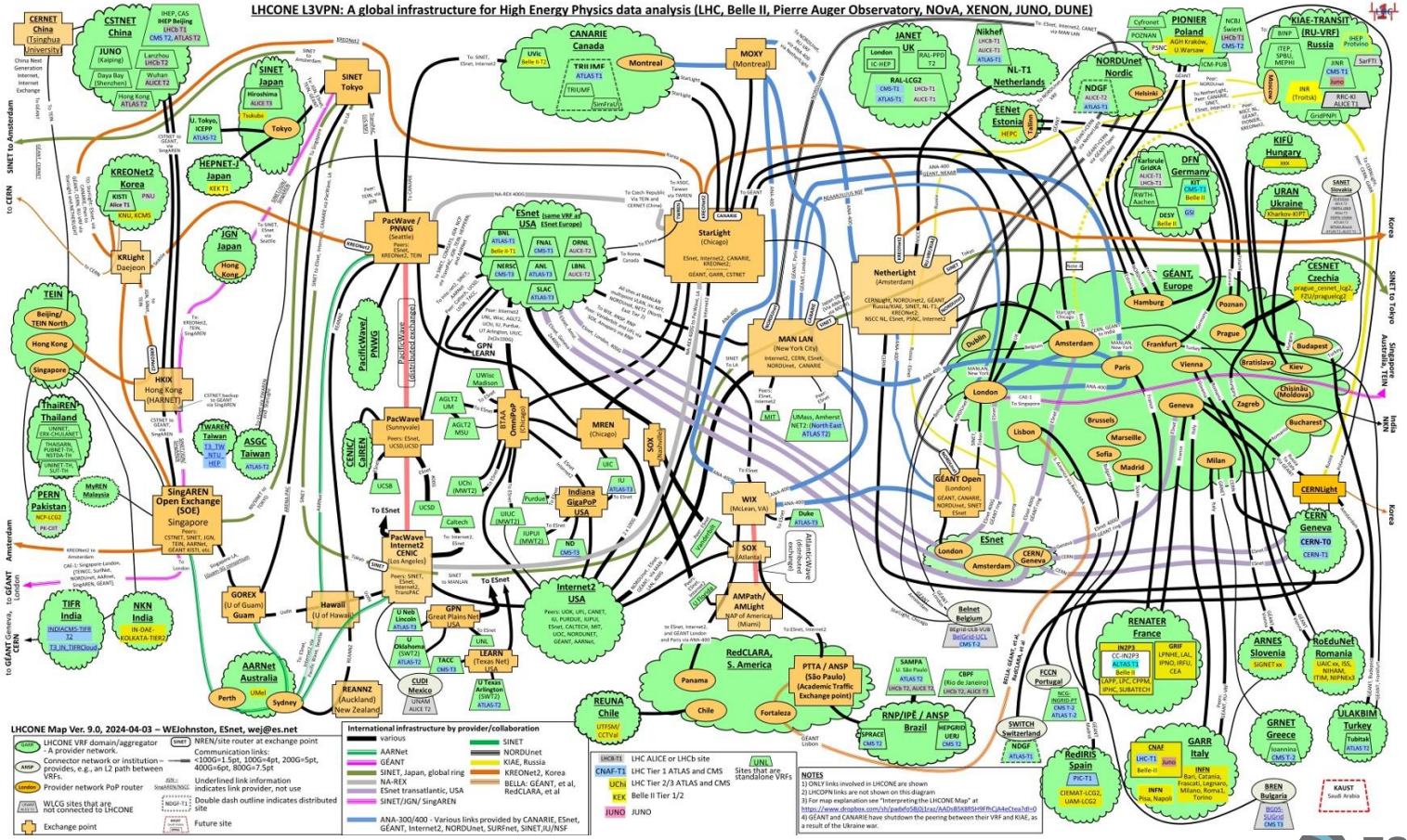


Complexity:
6000+ IETF RFCs
ITU-T
IEEE
GSM
Others…

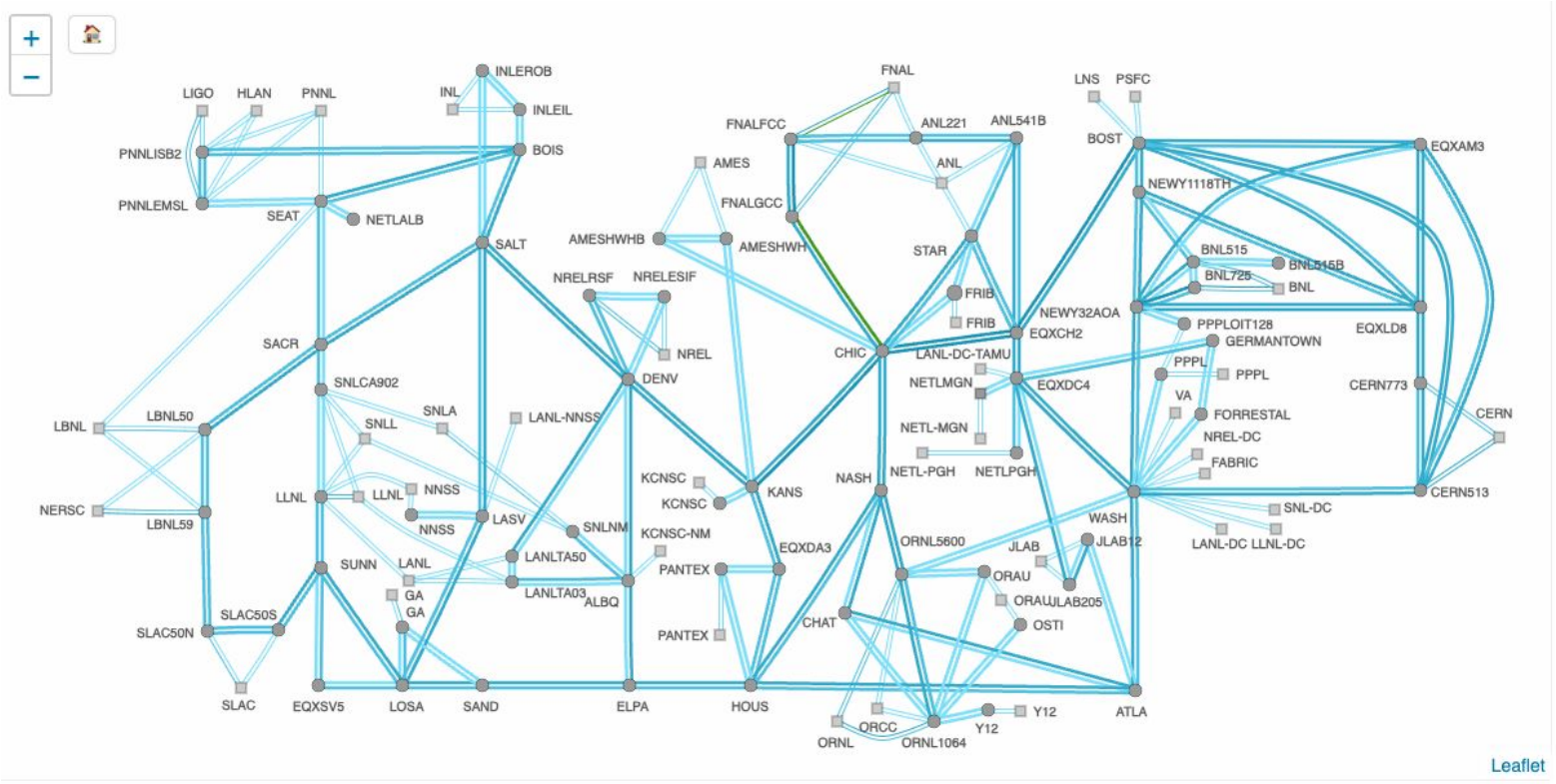


ESnet

LHCONE L3VPN

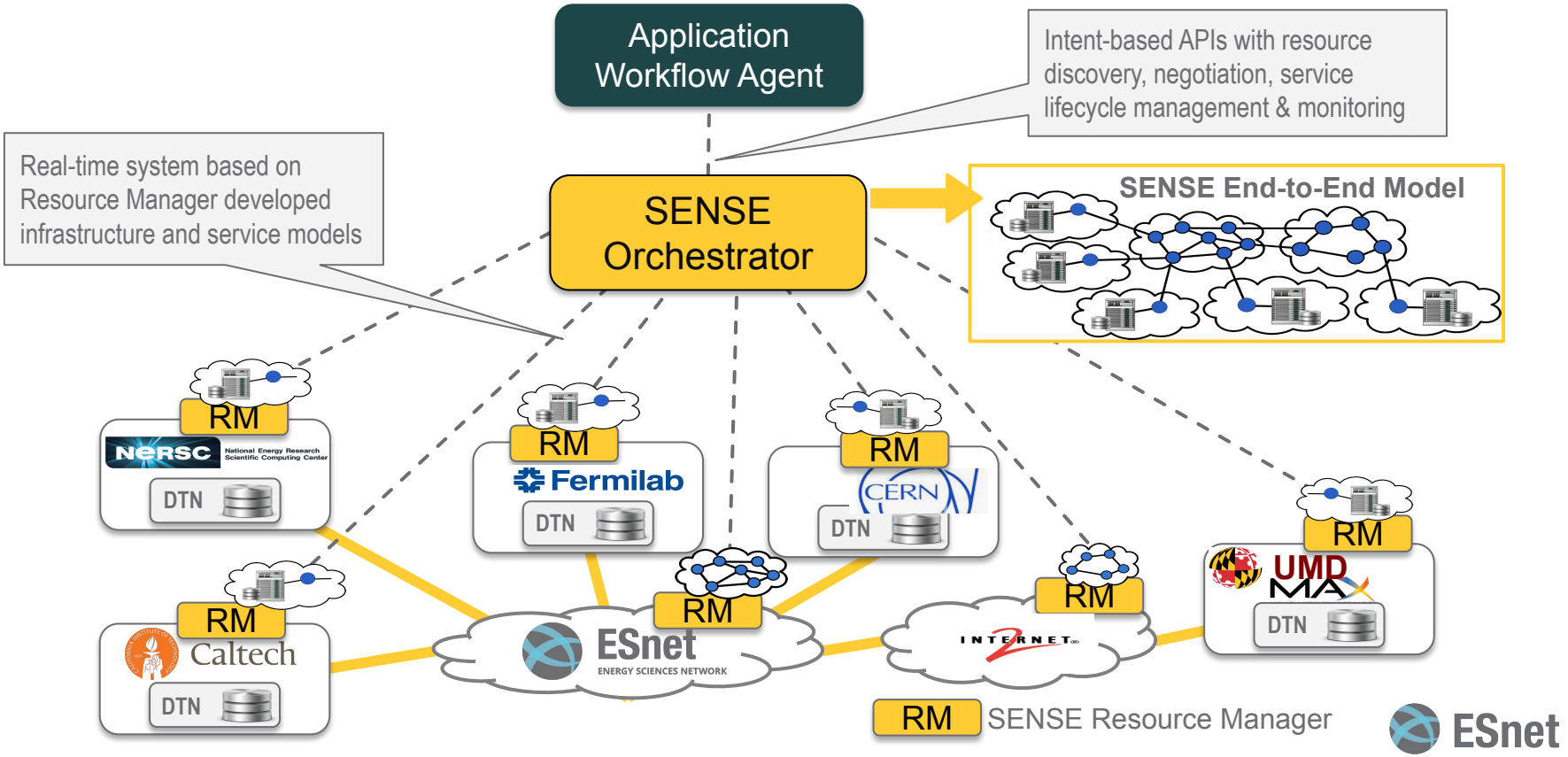


Network topology from my.es.net

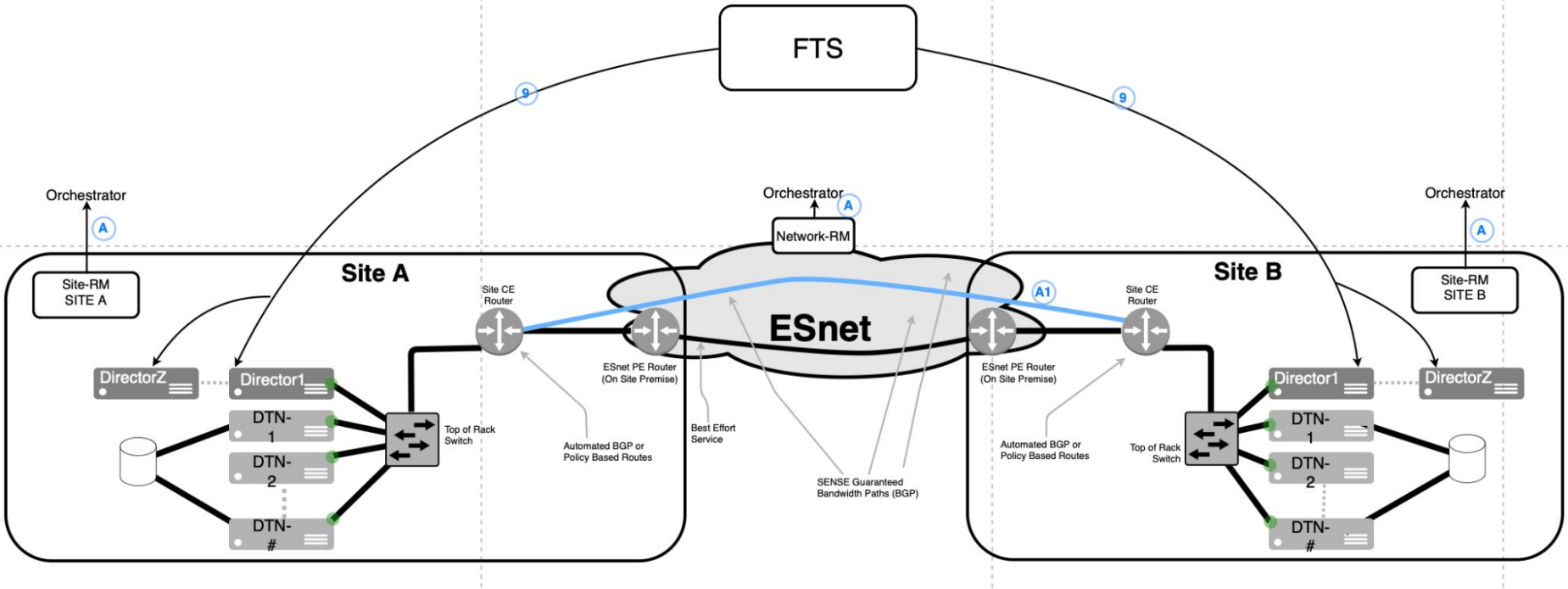


ESnet

The SENSE Architecture



What is possible via SENSE?

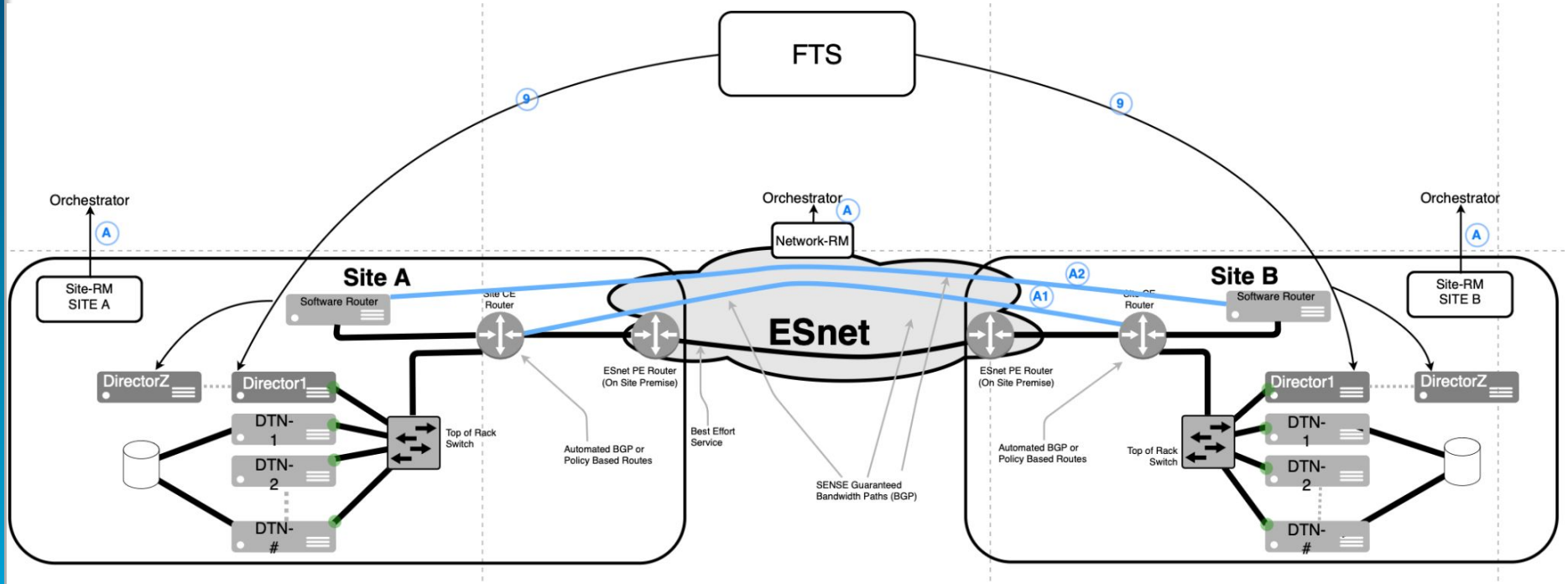


L2/L3/BGP/QoS/Modify/Vlan Translation (Dell, Arista, Cisco, Juniper, SONiC, FreeRTR)

End-to-End (last mile issue solved)

ESnet

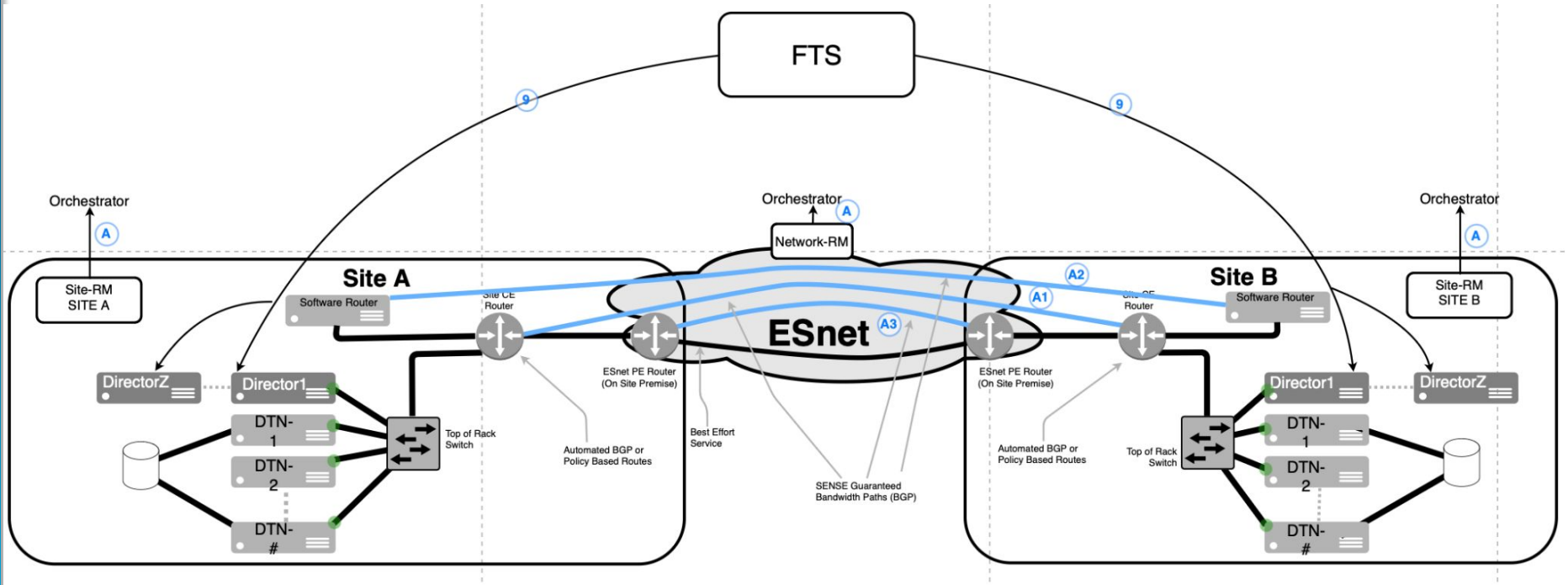
What is possible via SENSE?



No vendor lock, no switch/router access. Support - FRR, FreeRTR, SONiC, OVS DPDK/VPP offload with supported NICs

ESnet

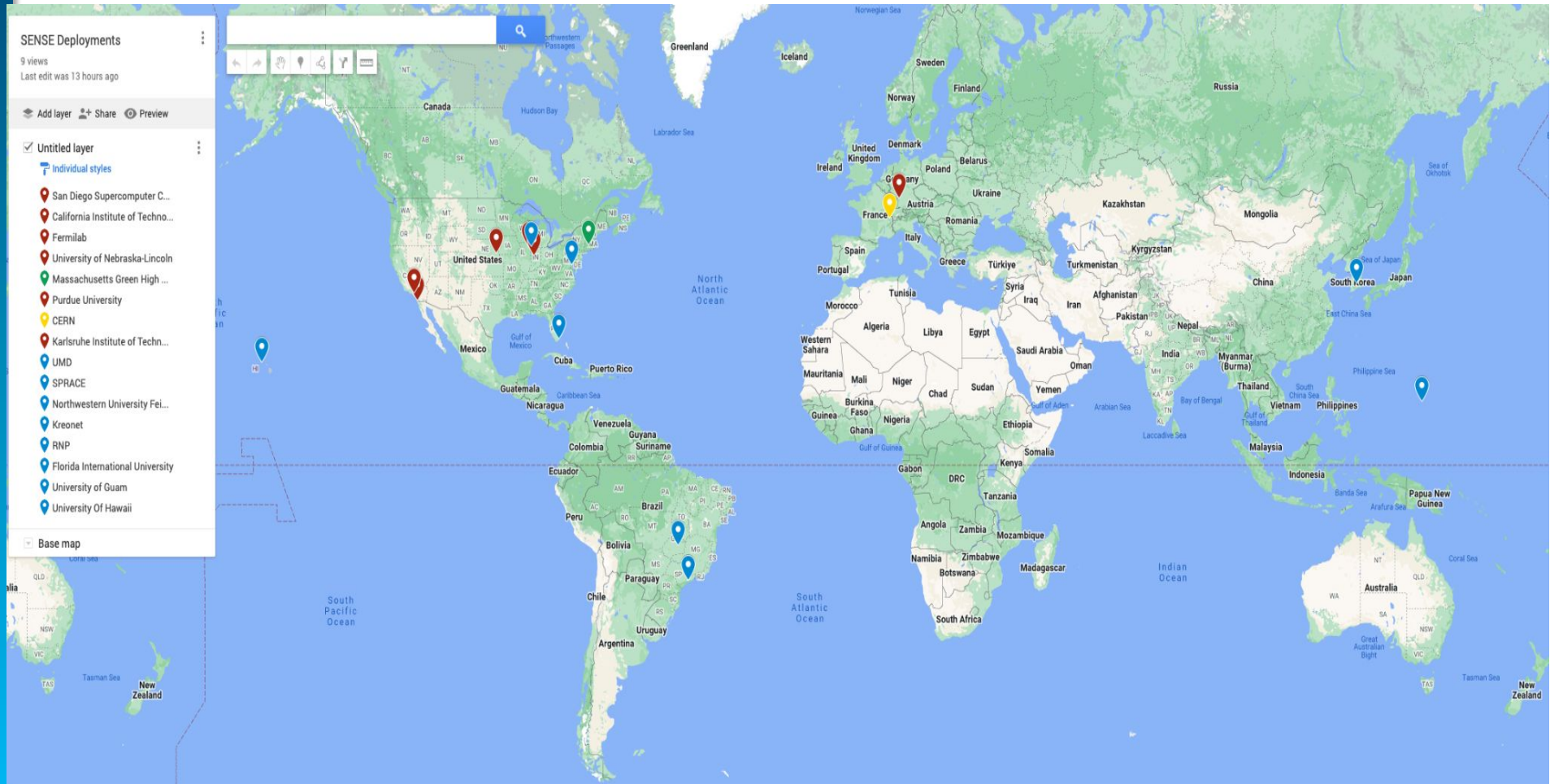
What is possible via SENSE?



No Site changes, all routing at NRE (currently L2/QoS support, L3/BGP/QoS - soon)

ESnet

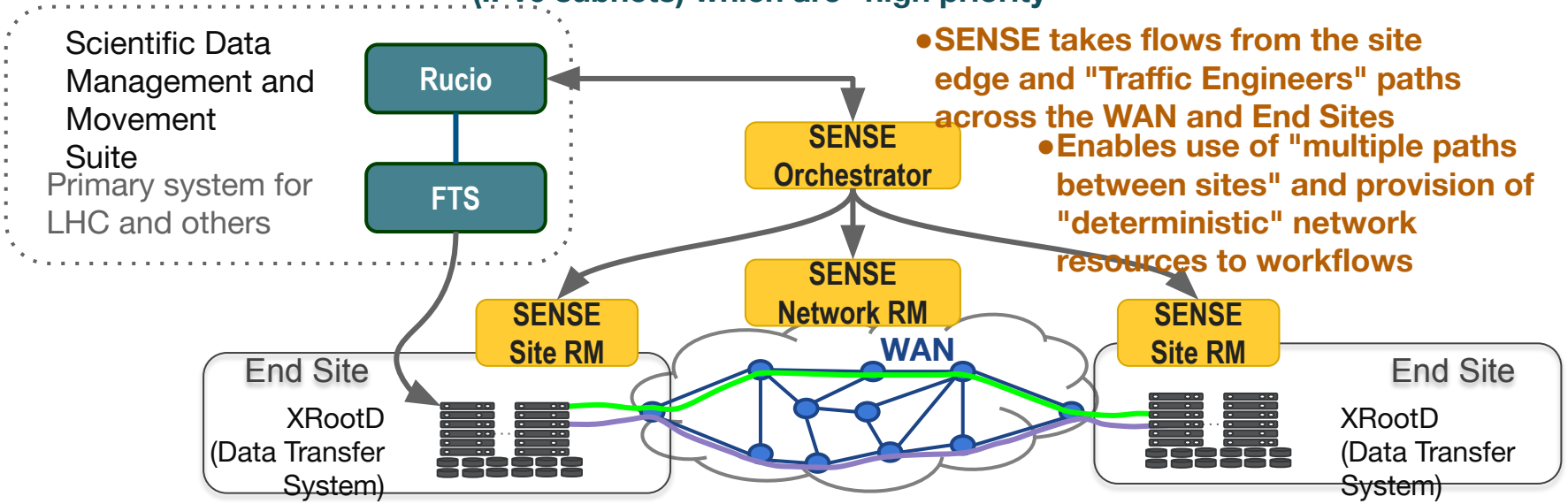
SENSE deployments: 52 Servers, 16 sites, 20 network domains




U.S. DEPARTMENT OF
ENERGY
BERKELEY LAB
ESnet
SENSE
Latest developments and results

SENSE and Rucio/FTS/XRootD Interoperation (DC24 and beyond)

- **Rucio identifies groups of data flows (IPv6 subnets) which are "high priority"**
- **SENSE takes flows from the site edge and "Traffic Engineers" paths across the WAN and End Sites**
- **Enables use of "multiple paths between sites" and provision of "deterministic" network resources to workflows**



Data Movement Manager (DMM) for the SENSE-Rucio Interoperation Prototype

22 Oct 2024, 17:09
18m
Room 1.B (Medium Hall B)

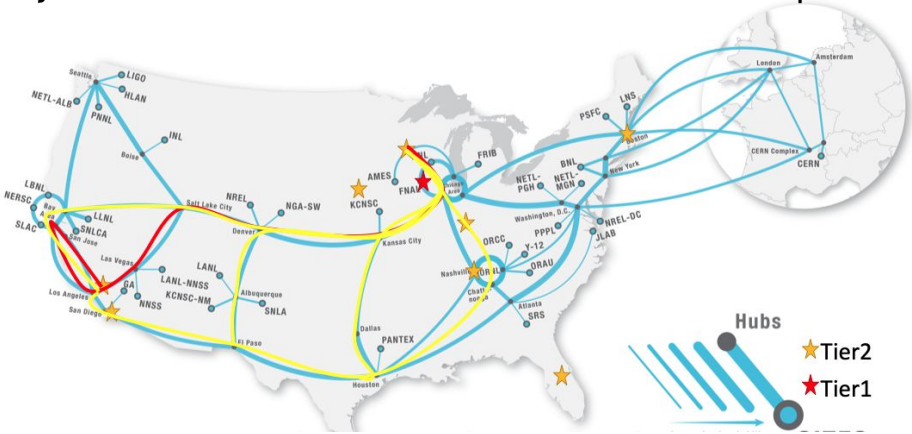
Talk | Track 1 - Data and ... | Parallel (Track 1)

Speaker

Aashay Arora (Univ. of California San Diego (US))

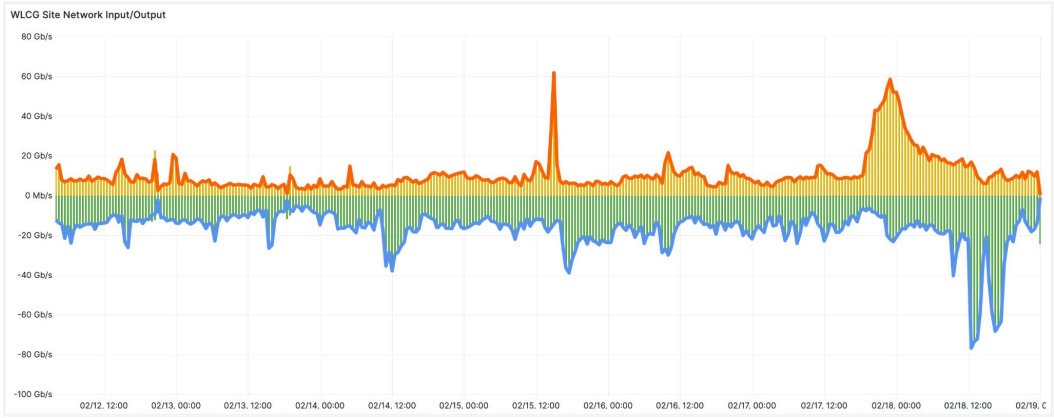
SENSE/Rucio (Network Orchestration)

The objective is to provide Rucio with capabilities to request network services via SENSE in order to: ***a) improve accountability, b) increase predictability, and c) isolate and prioritize transfer requests.*** This project uses a dedicated Rucio as well as XRootD instances so it would not interfere with Production systems. Data was transferred across a mix of production and next generation network paths.

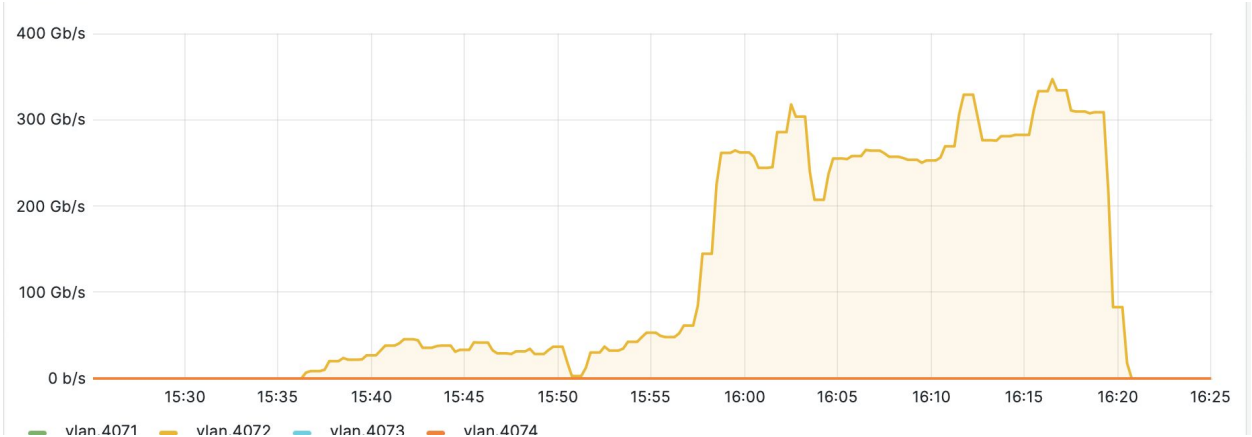


Between Fermilab, Caltech, UCSD Rucio-DMM/SENSE-FTS-XRootD multiple Rucio-triggered data flows were managed between multiple pairs of sites; The modify feature of DMM was used to change bandwidth allocation on the fly in response to Rucio requests. The following Quality of Service policies were demonstrated: Hard QoS / Soft QoS on Server; Hard QoS at the network level. DMM Real time API-driven FTS tuning was used to adjust active/max transfers settings. Additional US-CMS Tier2 sites are evaluated for deployment.

DC24 and after (CMS Caltech Tier2 Production)



Caltech Tier2 During DC24
80gbit/s max

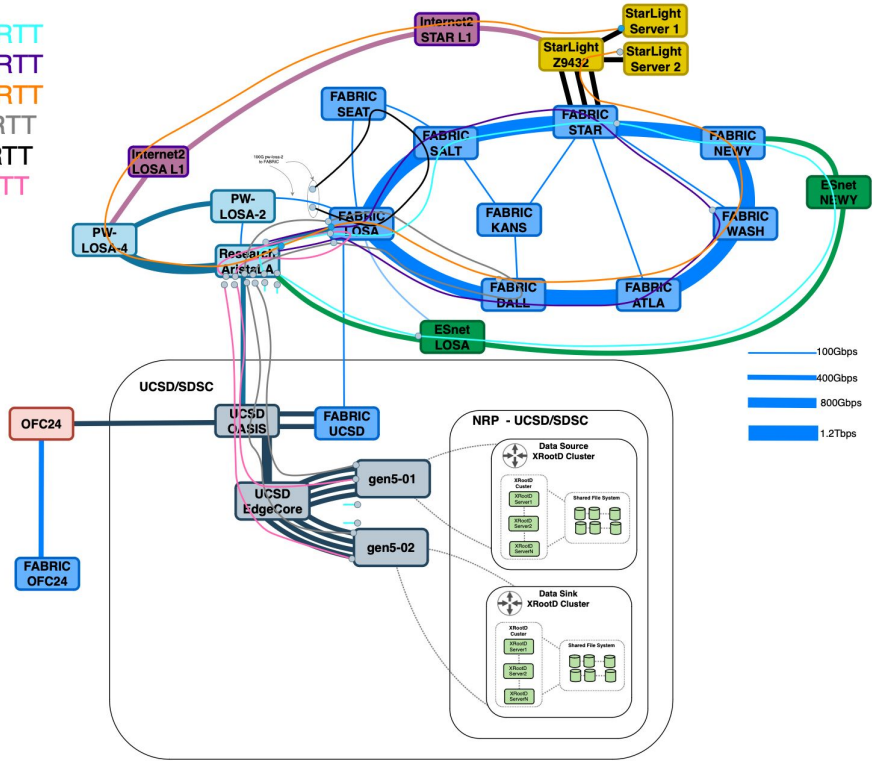


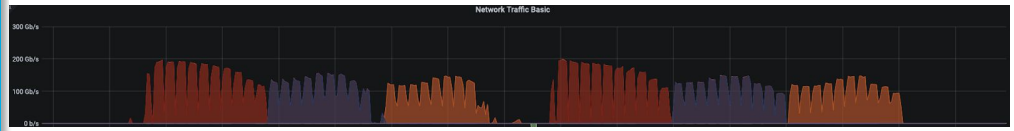
June 2024 via SENSE
Tunings:
New transfer nodes (2×100G)
Network limit removals;
NIC replacements;
JBOD SAS Configuration;
Ceph Object Size Increase 4MB-16MB;

ESnet

SENSE/Fabric/XRootD/NRP/Kubernetes/Multus

131 ms RTT
122 ms RTT
108 ms RTT
80 ms RTT
58ms RTT
6 ms RTT





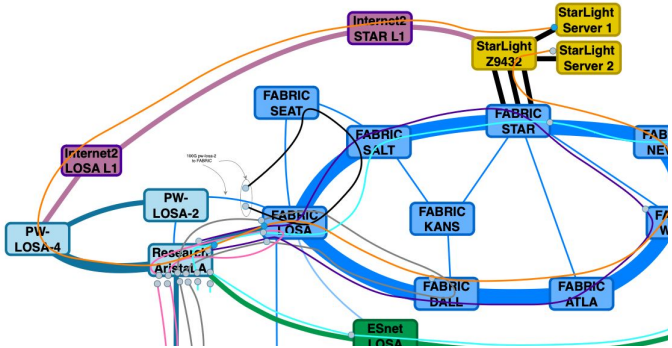
2 Servers:
2U Supermicro (SYS-621H-TN12R)
2× 32 core CPU (Intel Gold 6430)
1TB DDR5 (64GB DDR5-5600)
12x Samsung PM1733A (Raid0, 42TB)
400G NVIDIA CX7

- Can we sustain 400Gbps to/from the site using XRootD HTTPs?
- Where are the Software limitations?
- How does latency affects throughput?
- How do jumbo frames affect throughput?
- Should hyperthreading be ON or OFF for storage endpoints?
- What are CPU and Memory Requirements?
- What is the overhead when adding storage (Memdisk, local NVMe Raid, DFS)?

ESnet

SENSE/Fabric/XRootD/PRP/Kubernetes/Multus

131 ms RTT
122 ms RTT
108 ms RTT
80 ms RTT
58ms RTT
6 ms RTT



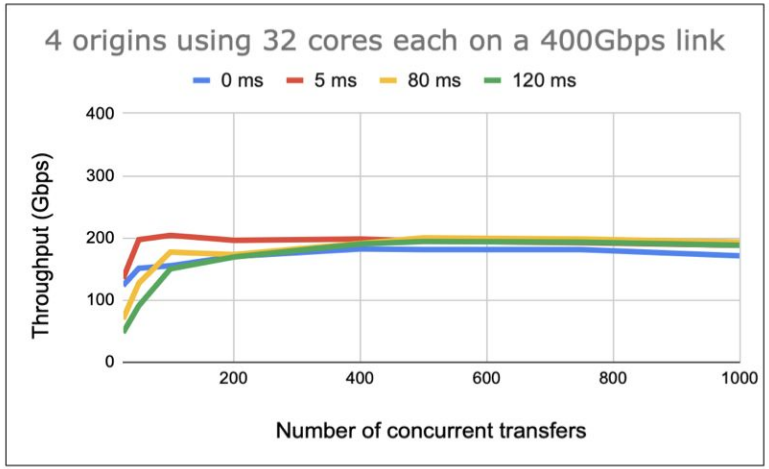
Benchmarking XRootD-HTTPS on 400Gbps Links with Variable Latencies

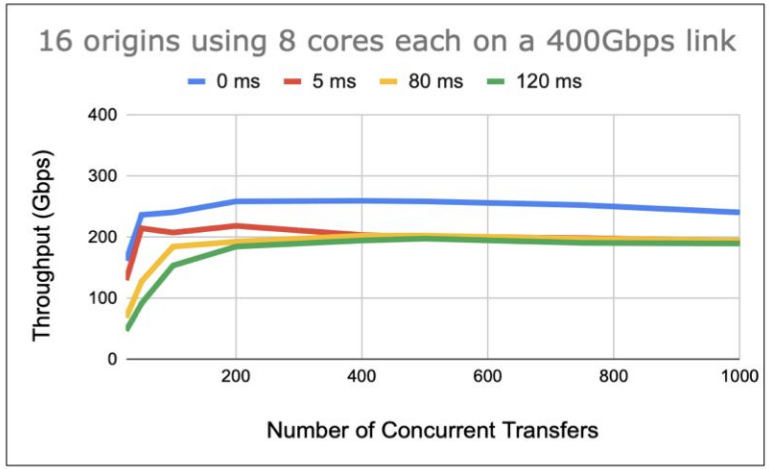
23 Oct 2024, 08:18
57m
Room 4

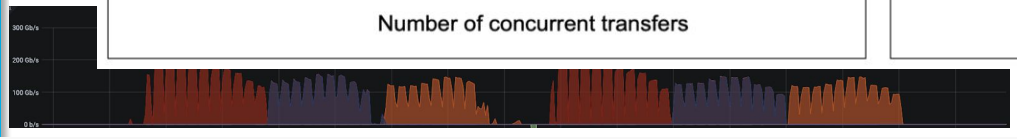
Poster | Track 1 - Data and ... | Poster session

Speaker

Aashay Arora (Univ. of California San Diego (US))





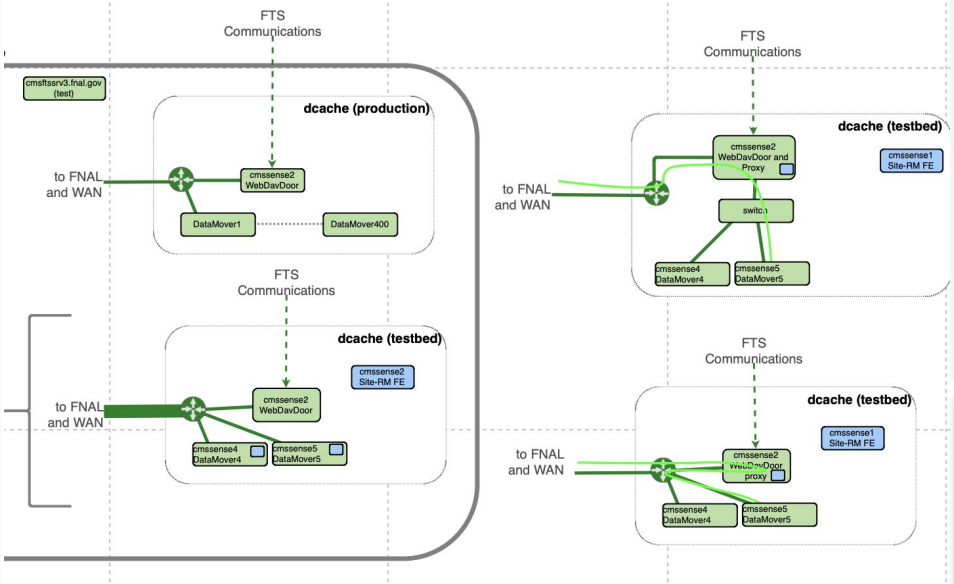


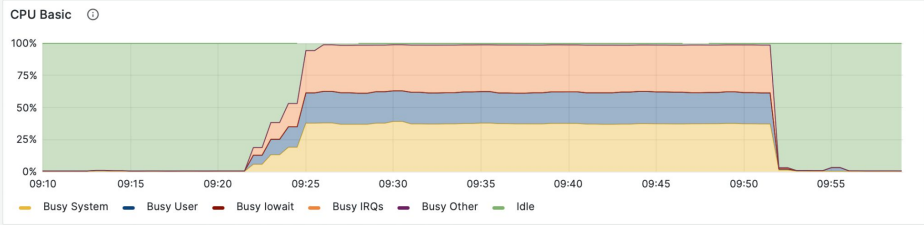
…m the site
…ons?
…ghput?
…roughput?
…r OFF for
…ding
storage (local NVMe Raid, DFS)?

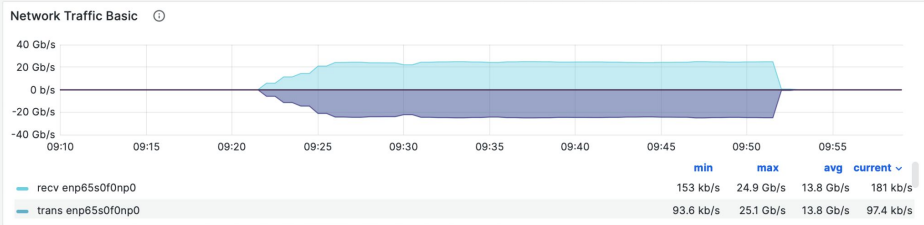
https://indico.cern.ch/event/1338689/contributions/6010884/attachments/2952346/5190108/CHEP_XRootD_poster.pdf

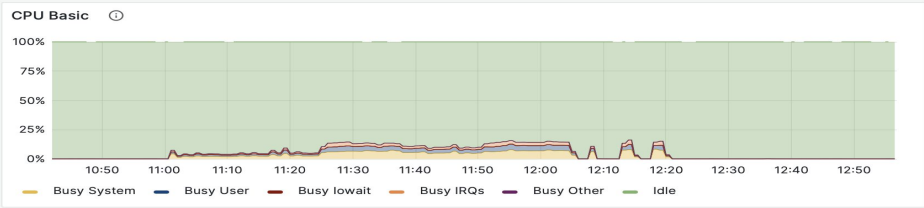
ESnet

Fermilab Dcache (Proxy/NoProxy) to SoCal



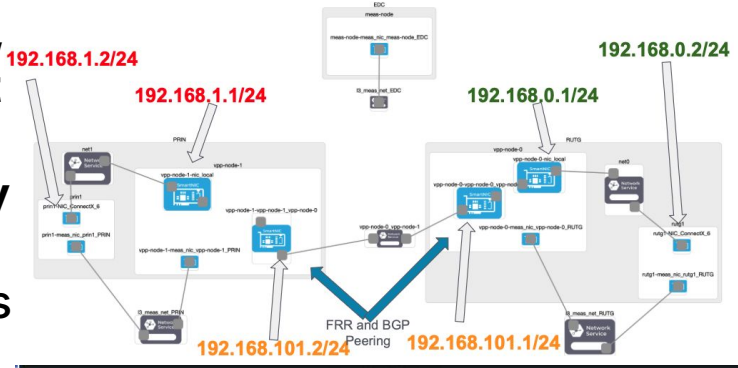






Software Router for SENSE/Rucio on FABRIC

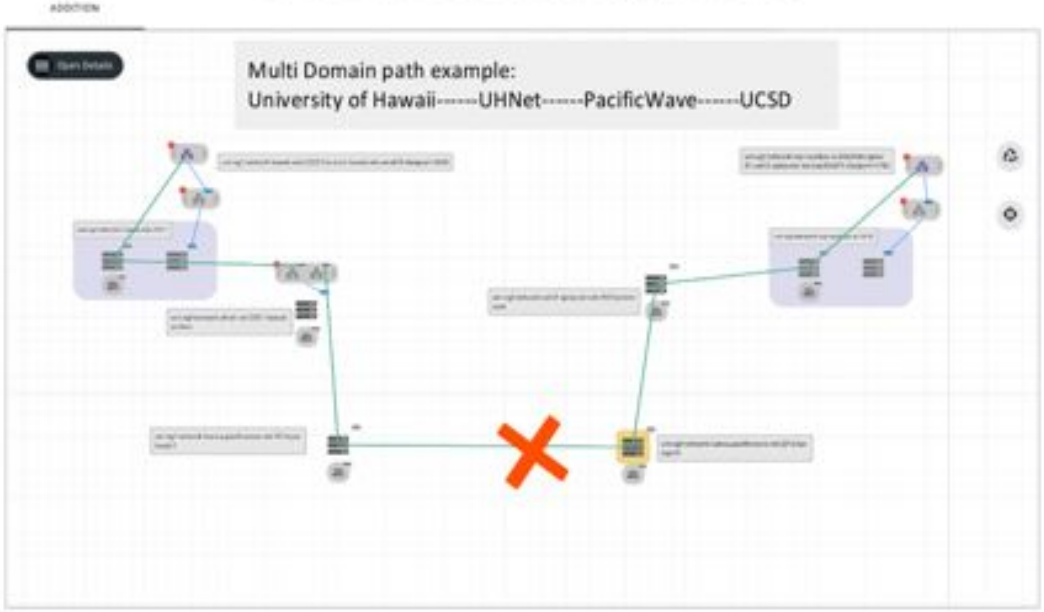
- **FABRIC** - Nation wide programmable network, provides **GPU, FPGAs, NICs, QoS, Interconnect national facilities. Allows to design, test applications, protocols and services at any node in the network**
- SENSE/Rucio need to support control at Sites without network device access.
- Hardware/Software in use:
  - ConnectX-6 (PCI passthrough, 2×100G)
  - VPP with DPDK
  - FRRouting (without/with DPDK via VPP)
  - FreeRTR with DPDK
- Stable 50Gbps with 2 cores/4gb RAM VM (FRRouting only, no DPDK)
- VPP - 60 Gbps (with DPDK)
- FreeRTR - 30 Gbps (no Jumbo frames support)



https://github.com/sdn-sense/vpp-frr

Real Time Debugging

Imagine knowing where the network path is broken at a glance!

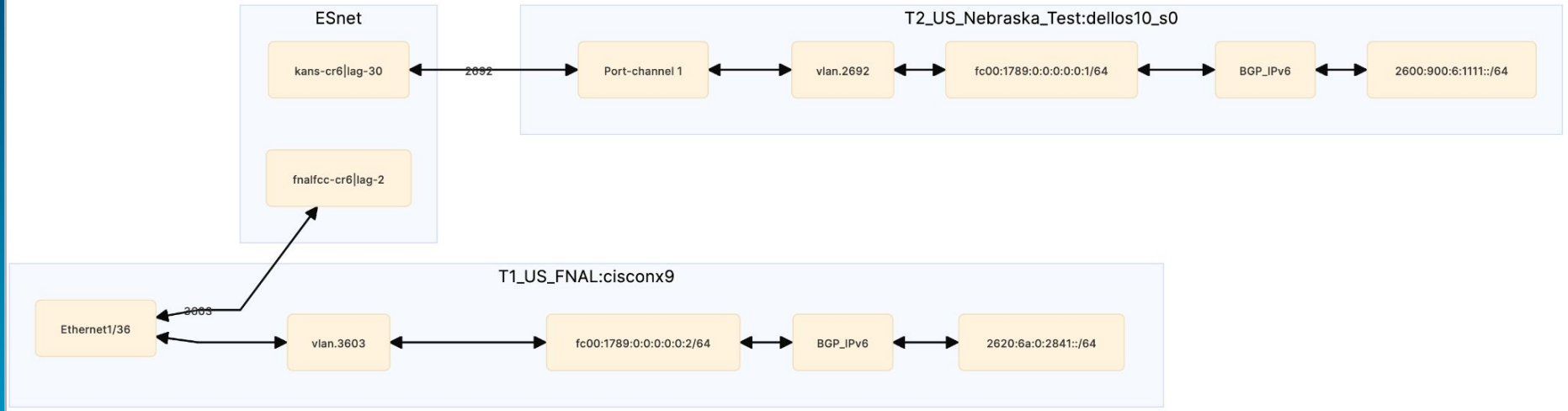


Sharks' attraction to undersea fiber-optic cables has been well-documented over the years.

Screenshot / YouTube

ESnet

L3 BGP peering (end-to-end real time)



ESnet

Special thank you to many colleagues

Frank Würthwein, Jonathan Guiang, Aashay Arora, Diego Davila, John Graham, Dima Mishin, Thomas Hutton, Igor Sfiligoi, Harvey Newman, Maria Spiropulu, Justas Balcas, Raimondas Sirvinskas, Preeti Bhat, Marcos Schwarz, Sravya Uppalapati, Andres Moya, Tom Lehman, Inder Monga, Xi Yang, Chin Guok, John MacAuley, Hans Trompert, Evangelos Chaniotakis, Joe Mambretti, Sana Bellamine, Christopher Bruton, Oliver Gutsche, Asif Shah, Chih-Hao Huang, Dmitry Litvinsev, Phil Demar, Andrew Melo, David A Mason, Garhan Attebury, Hans Trompert, Rafael Coelho, Jessa Westclark, Moya Andres

and many others from NRE communities


U.S. DEPARTMENT OF
ENERGY
BERKELEY LAB
ESnet
SENSE
Backup slides

SENSE and Rucio for USCMS (During DC24)

interface eth1 req-123 input/output rate 100gbit

class flow1 commit 40gbit # Hard QoS

Match 2620:fc00::/64

Class flow2 commit 60gbit # Hard QoS

default

interface eth1 req-123 input/output rate 100gbit

class flow1 commit 40gbit max 100gbit # Soft QoS

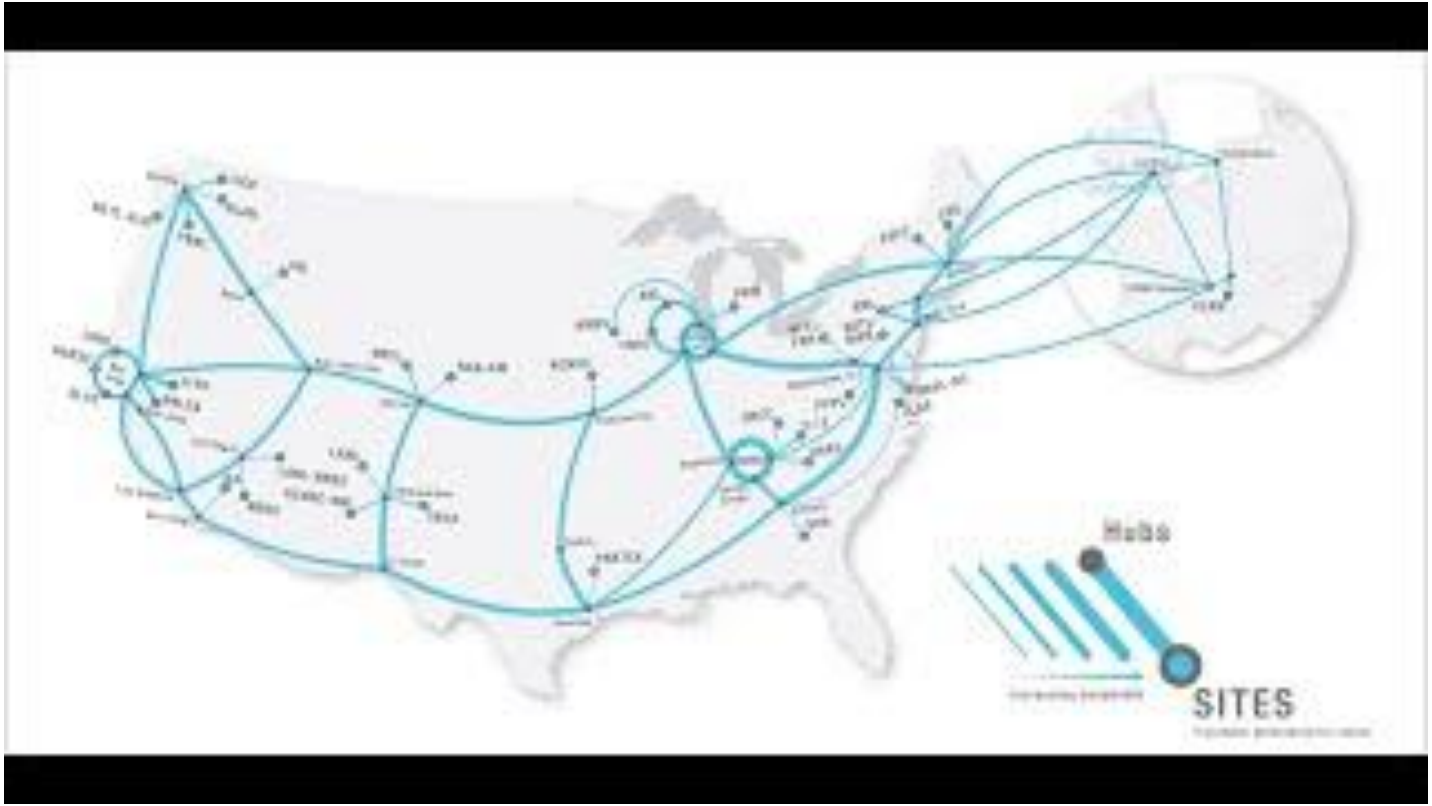
Match 2620:fc00::/64

class flow2 commit 10gbit max 100gbit # Soft QoS

default

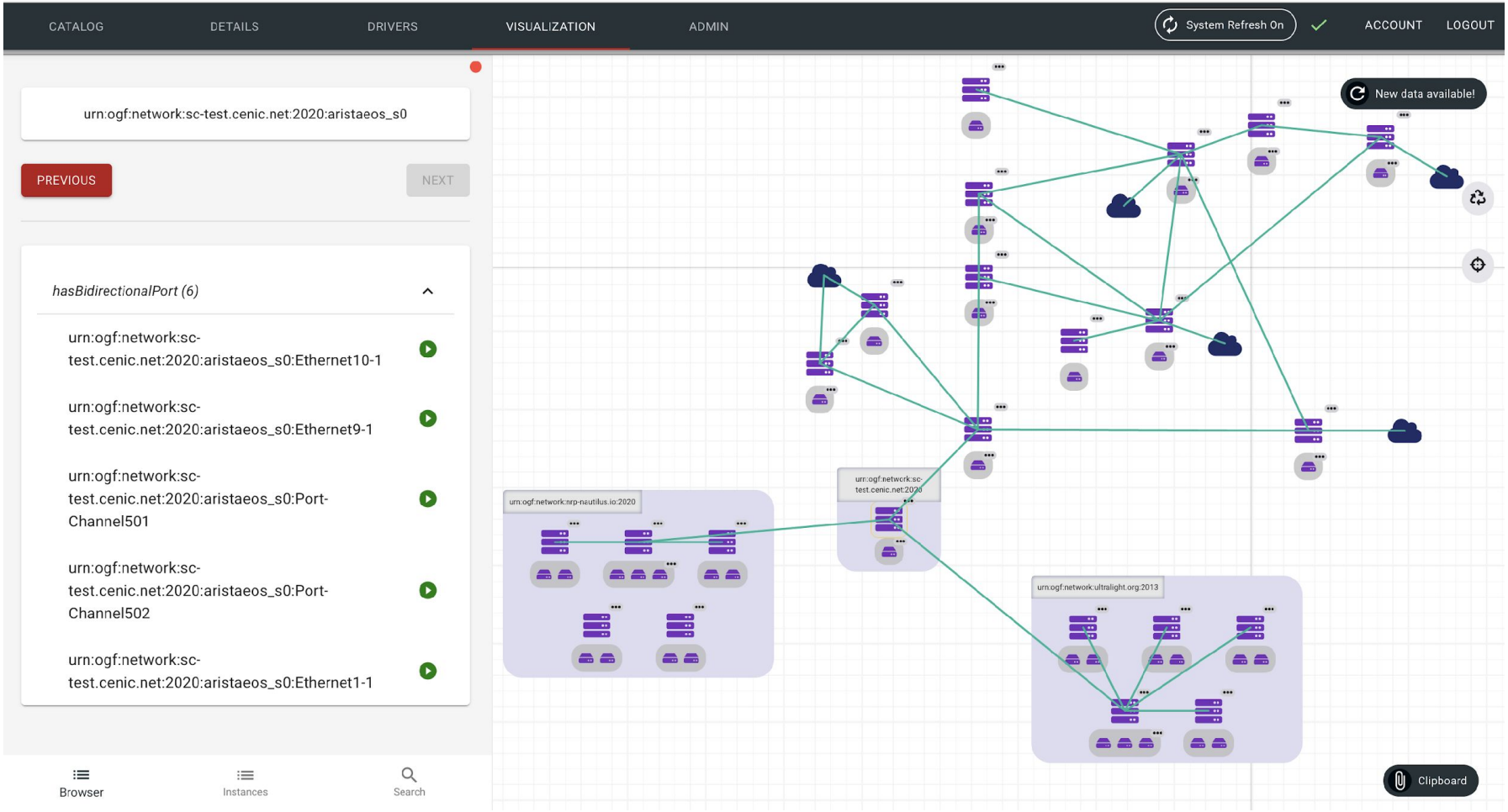
ESnet

Demo time (Recorded during SC23)



ESnet

SENSE - Semantic Modeling of Global Resources in Real Time



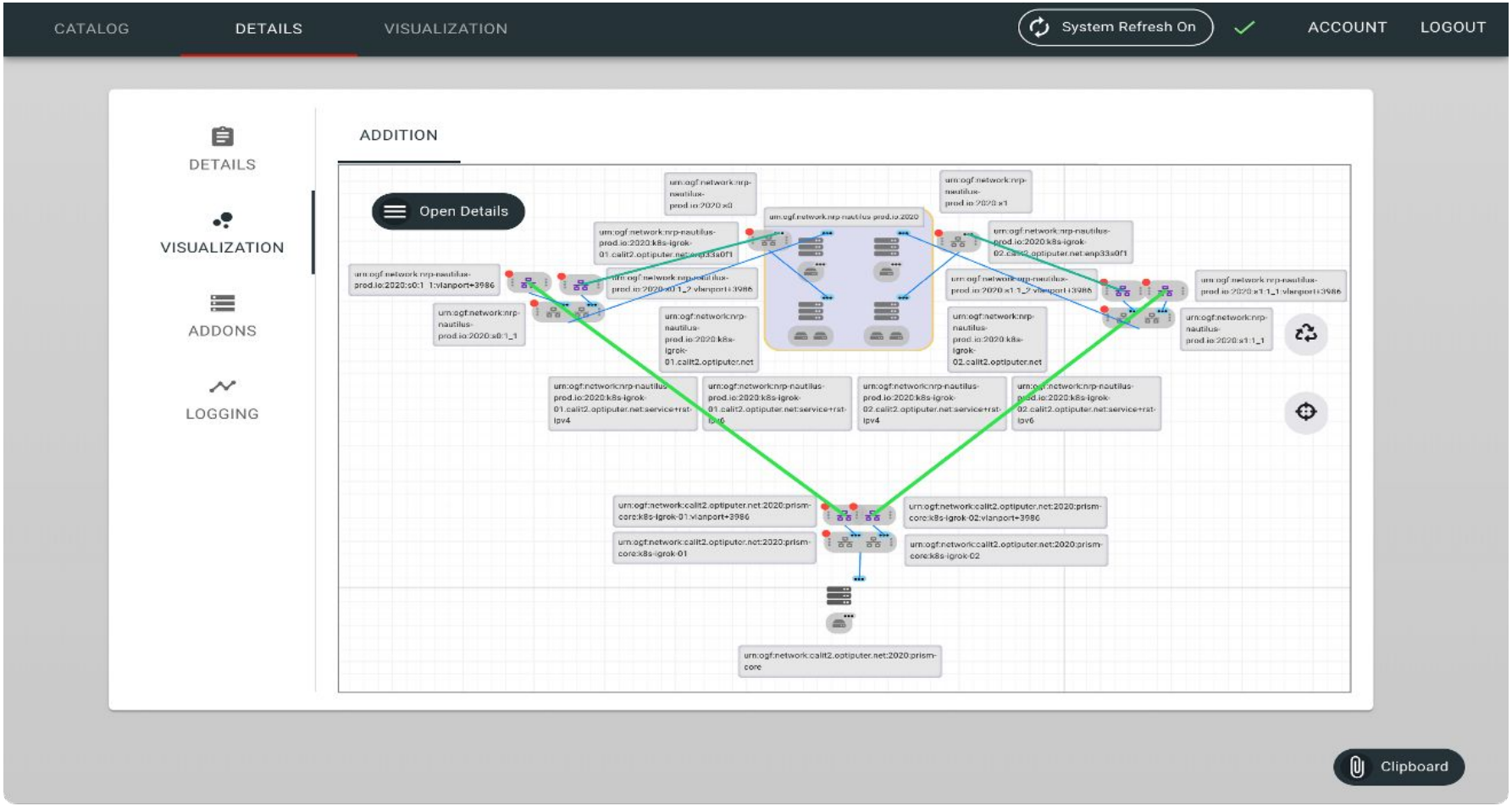
Science Focused Automation and Orchestration with SENSE

- History of the SENSE Orchestrator
  - The development of Multi-Resource Markup Language (MRML) as a SENSE precursor and foundation of “semantic modeling of everything in the cyberinfrastructure”.
  - 2015-2019, “SDN for End-to-end Networked Science at the Exascale” (SENSE) sponsored by DOE with a focus on orchestration and automation of end-to-end SDN networks across WAN domains, end-sites and host servers.
  - SENSE today is specialized in integrating multi-facility, multi-network, multi-cloud infrastructures and presenting as normalized, abstracted, single-point-of-touch services to the workflows.

- A taste of the SENSE orchestration service
  - Allocate a data transfer host in a DOE lab and a VM cluster in Amazon AWS cloud
  - Interconnect them into an overlay of interconnected L2VPN and L3VPN across the lab site, ESnet, Internet2 and cloud provider networks.
  - The end-to-end automation and orchestration is API driven by an application workflow agent with an intent-based service definition that is customized and abstract.
  - Interactive workflow assistance is provided with negotiation, co-scheduling, auditing and full service lifecycle management.

ESnet

SENSE Orchestrated Service Instance as a Resource Model “Delta”



SENSE Service Profile - Workflow Intent for End-to-End CI Needs

ID: 719e9823-d628-4cf0-8c25-ad5c6fd5642b

[SC23] Rucio-DMM-FNAL-UCSD

FOLDER: RUCIO

DEMO

Licenses

aarora@ucsd.edu - 5 slot(s) given.
ALLOCATION

jbalcas@caltech.edu - 3 slot(s) given.
ALLOCATION

xiyang@es.net - 1 slot(s) given.
ALLOCATION

+

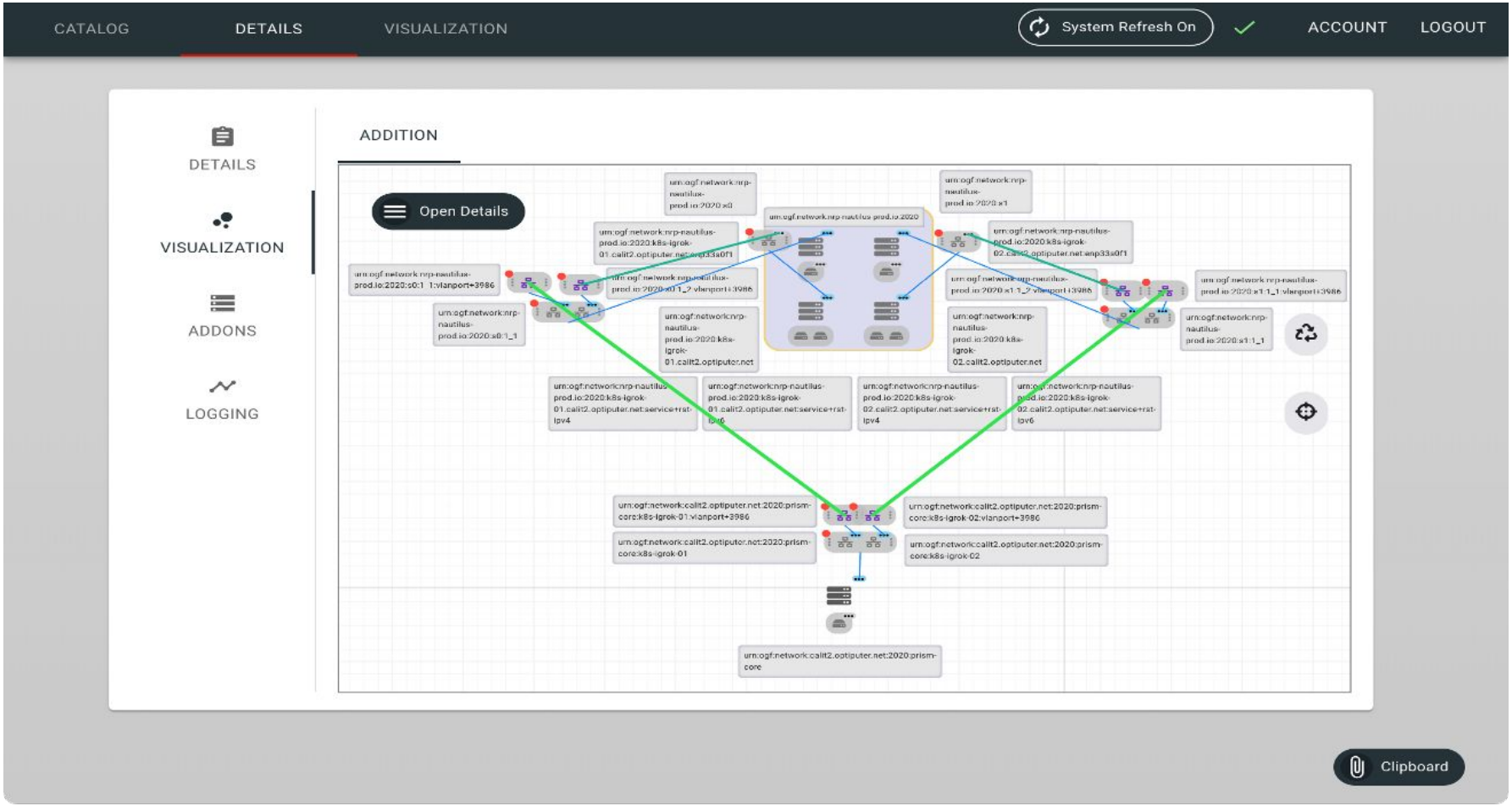
Select a node...


```
▼ DNC root schema {2}
  ▼ data {2}
      type : Site-L3 over P2P VLAN
    ▼ connections [1]
      ▼ 0 {5}
          name : Connection 1
        ▼ terminals [2]
          ▼ 0 {4}
              uri : urn:ogf:network:fnal.gov:2023
              vlan_tag : any
              ipv6_prefix_list : 2620:6a:0:2842::/64
              assign_ip : true
          ▼ 1 {4}
              uri : urn:ogf:network:nrp-nautilus.io:2020
              vlan_tag : any
              ipv6_prefix_list : 2001:48d0:3001:111::/64
              assign_ip : true
        ▼ path_profile {1}
          ▼ exclusion_list [1]
            ▼ 0 {1}
                uri : urn:ogf:network:stack-fabric:2024:topology
```


JSON View | SAVE AS | SAVE | DELETE | CLOSE | Alias | SUBMIT

ESnet

SENSE Orchestrated Service Instance as a Resource Model “Delta”



SENSE Service Instance - API Driven Full Lifecycle Management