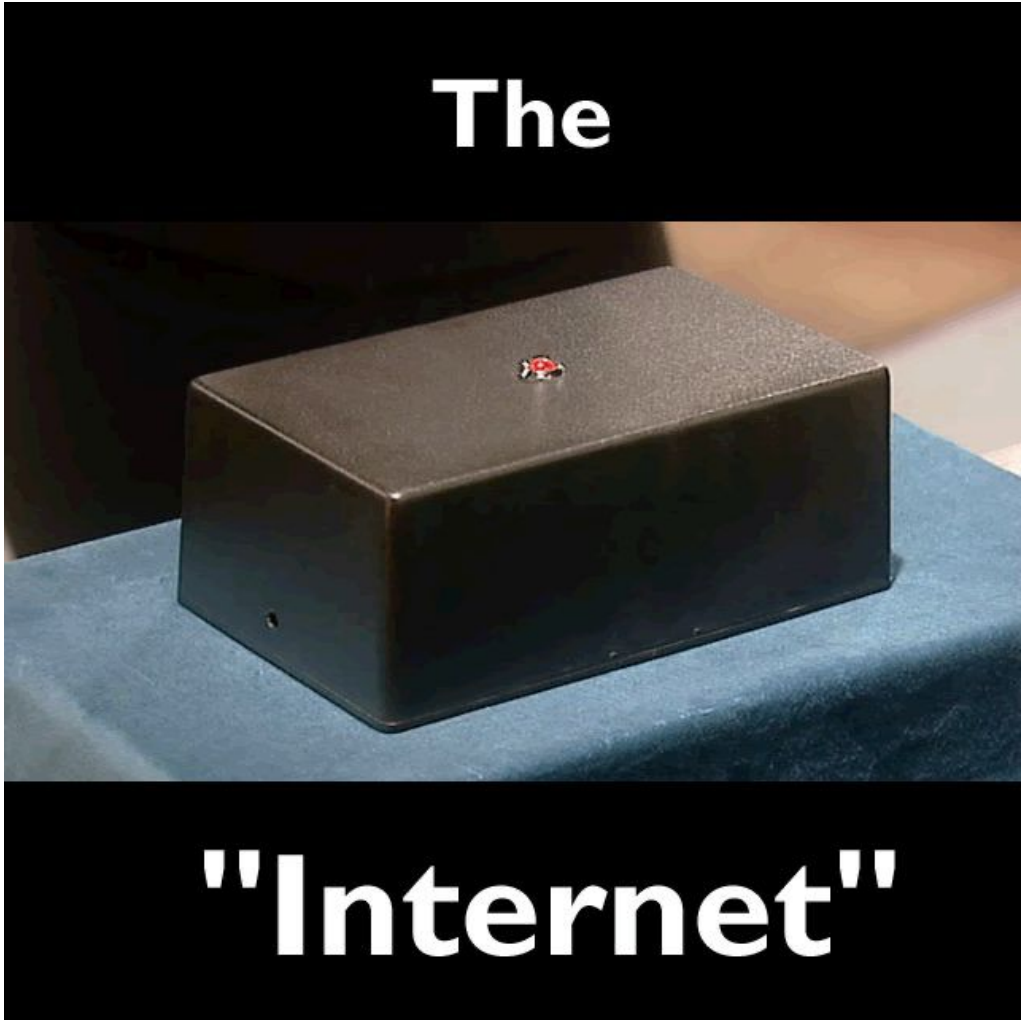
SENSE in ATLAS

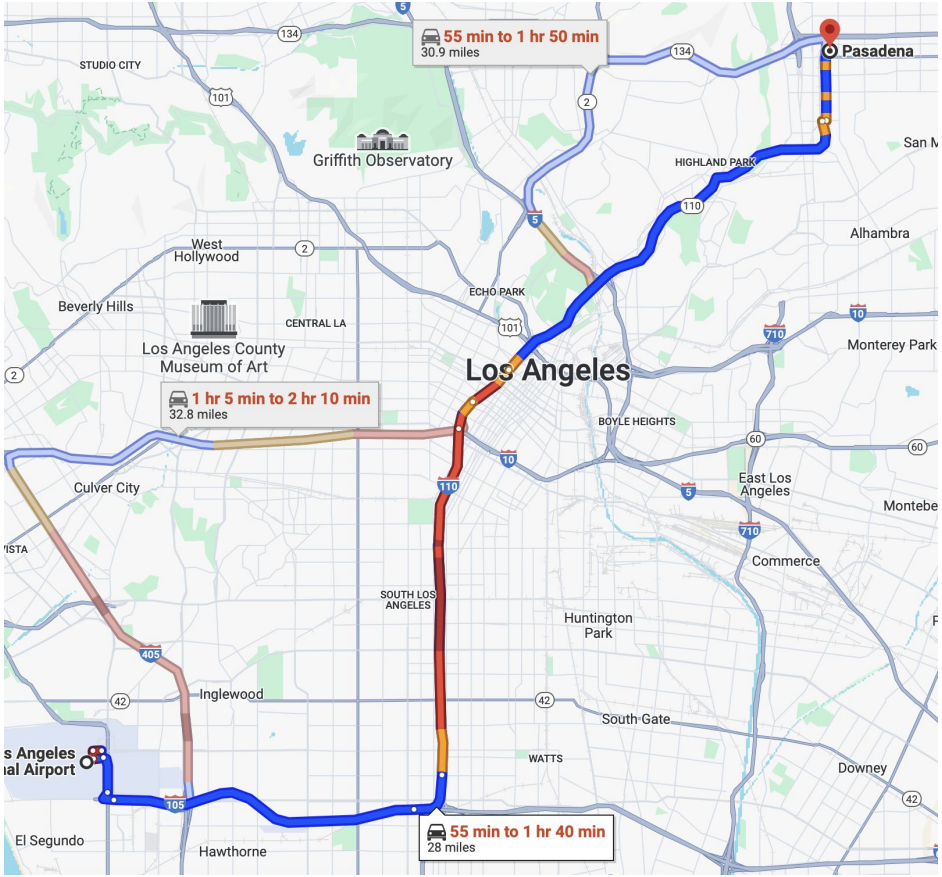
HTC 2026

Rafael Coelho, Justas Balcas on behalf of the Rucio/SENSE team

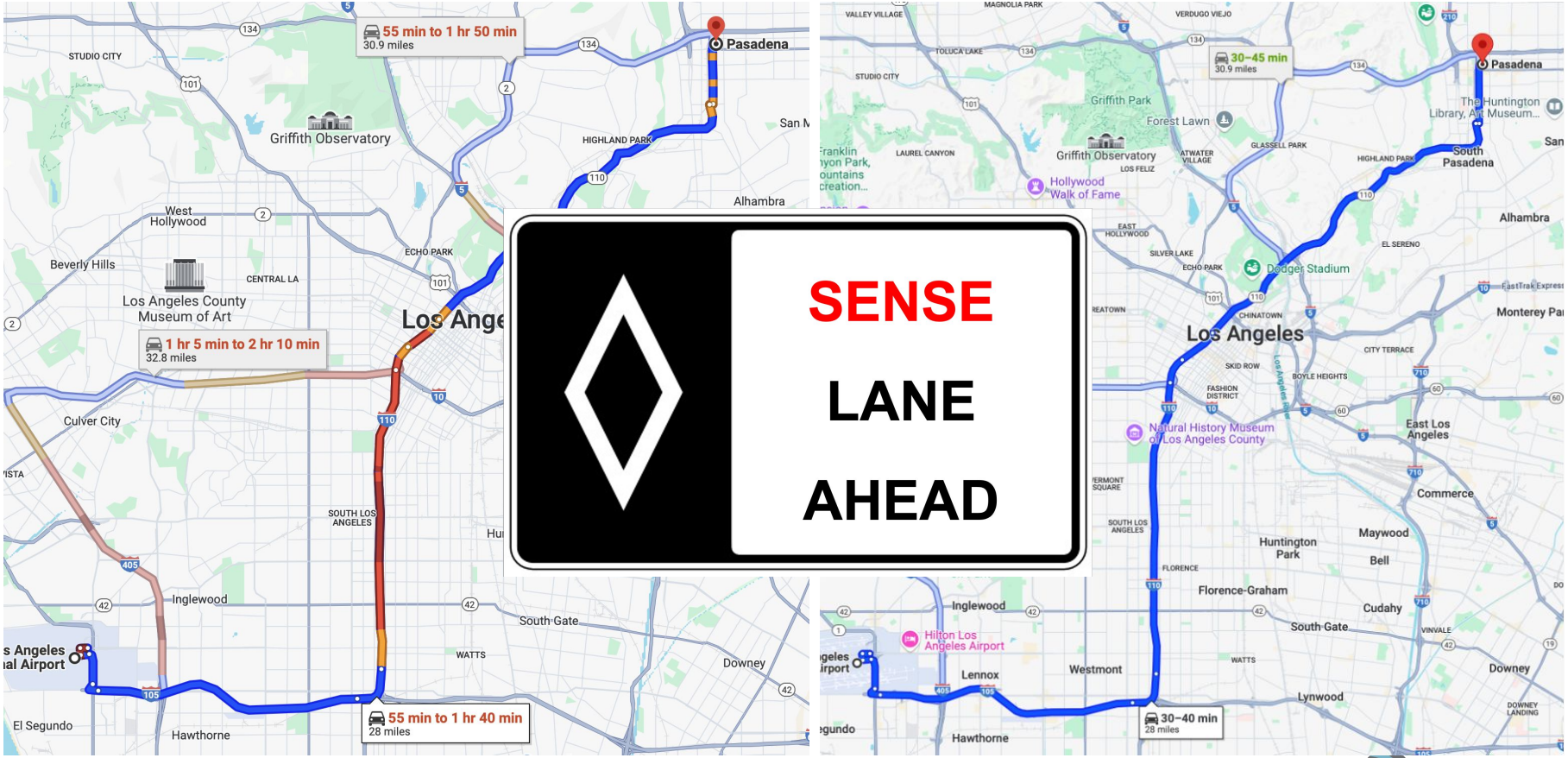
ESnet

When will I get home?



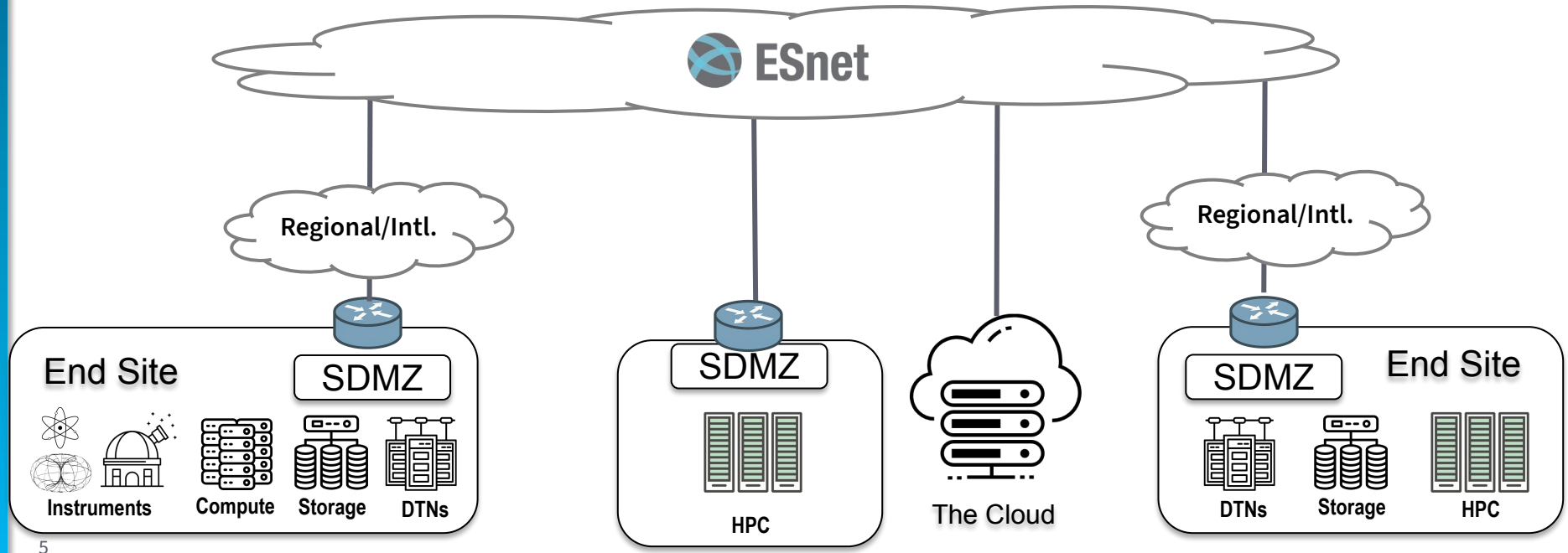
ESnet

SENSE Lanes and Route planning

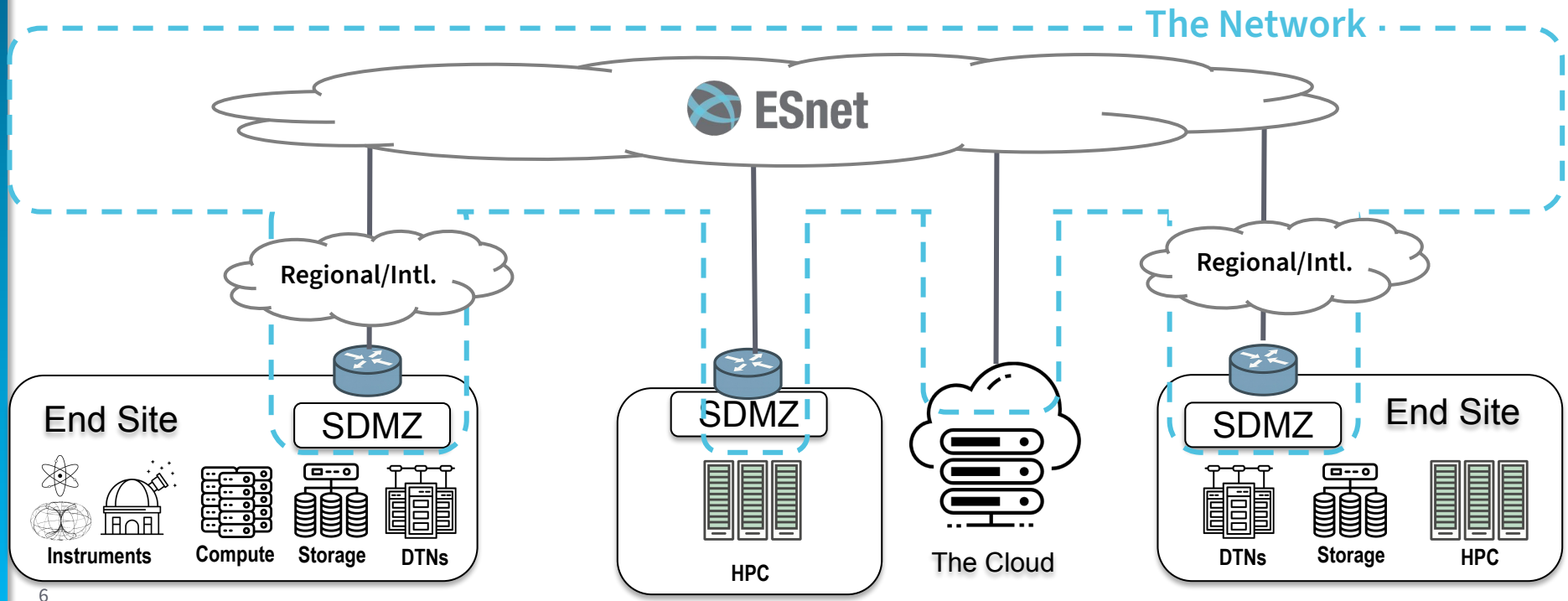


ESnet

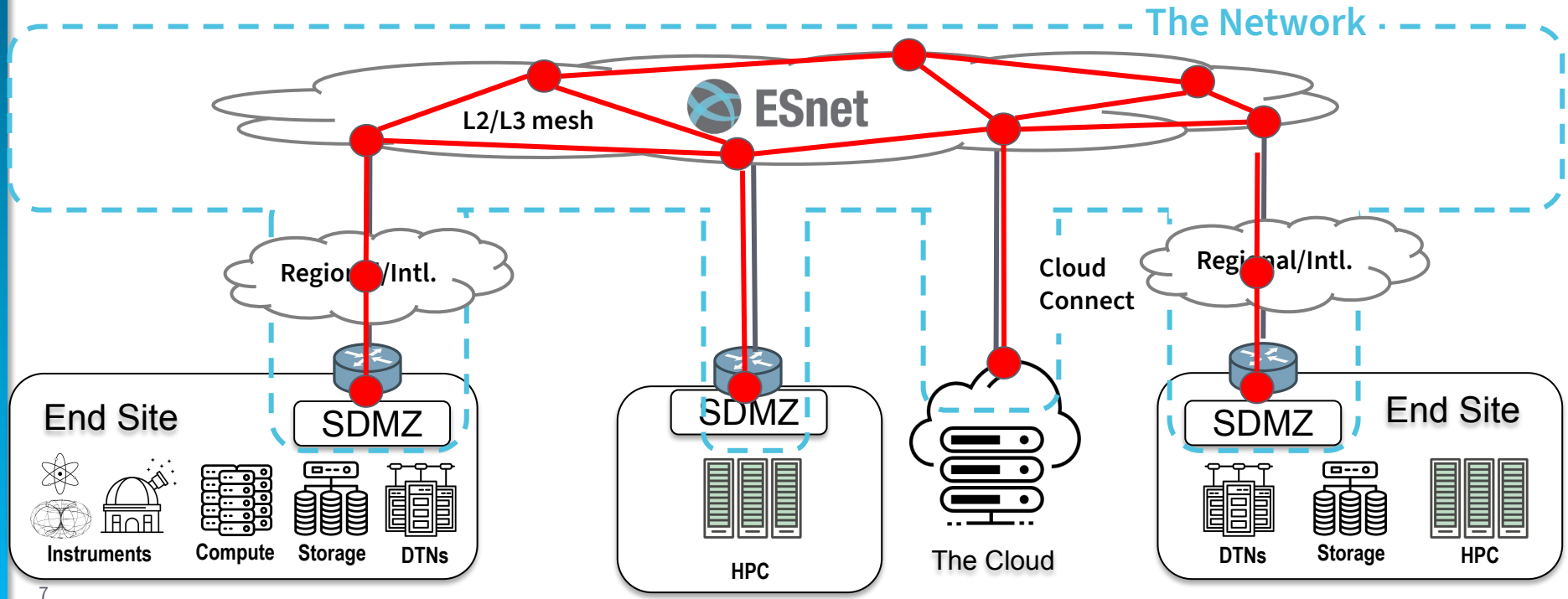
SENSE: Build your own Network with Guarantees



SENSE: Build your own Network with Guarantees



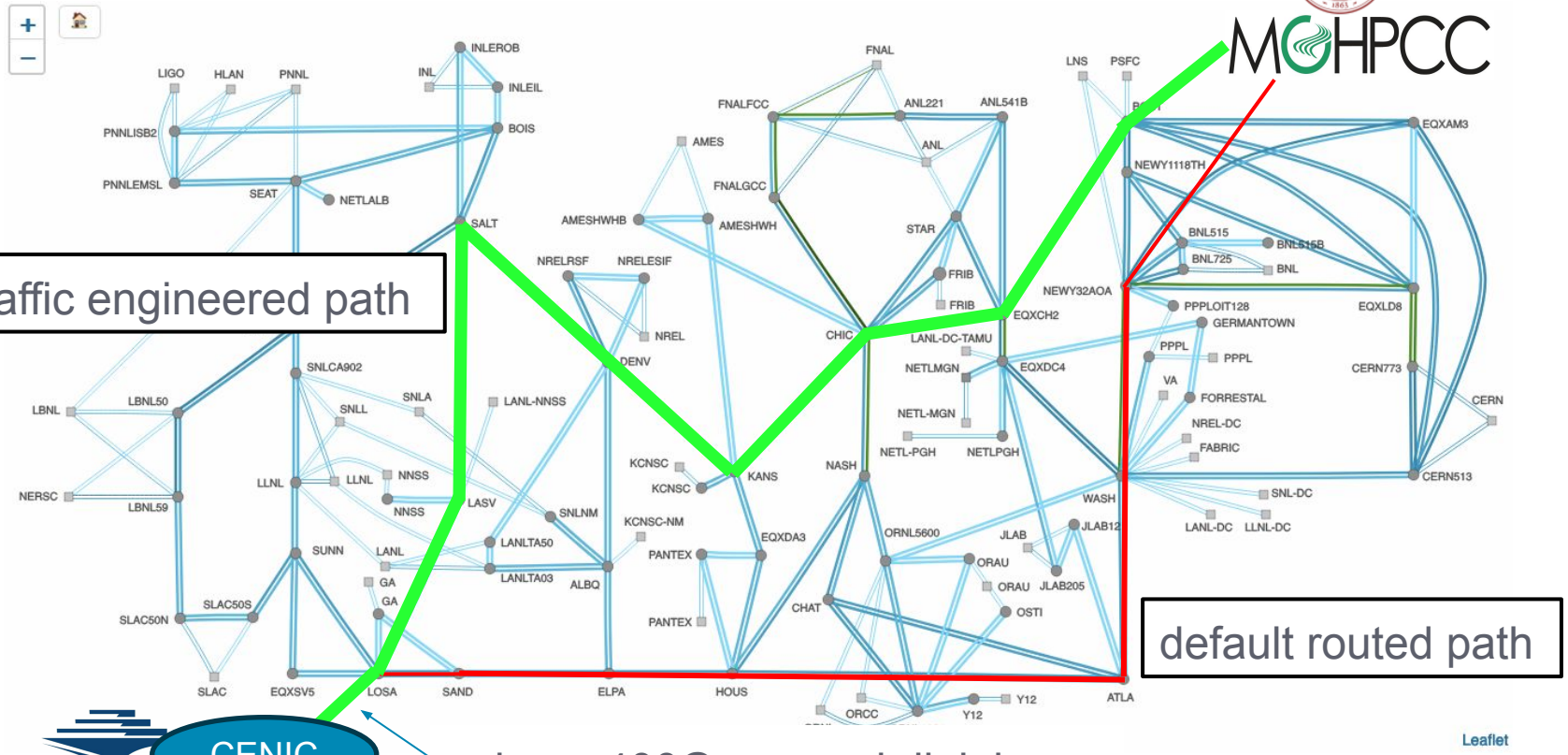
SENSE: Build your own Network with Guarantees



Paths MGHPCC ⇒ UCSD/SDSC

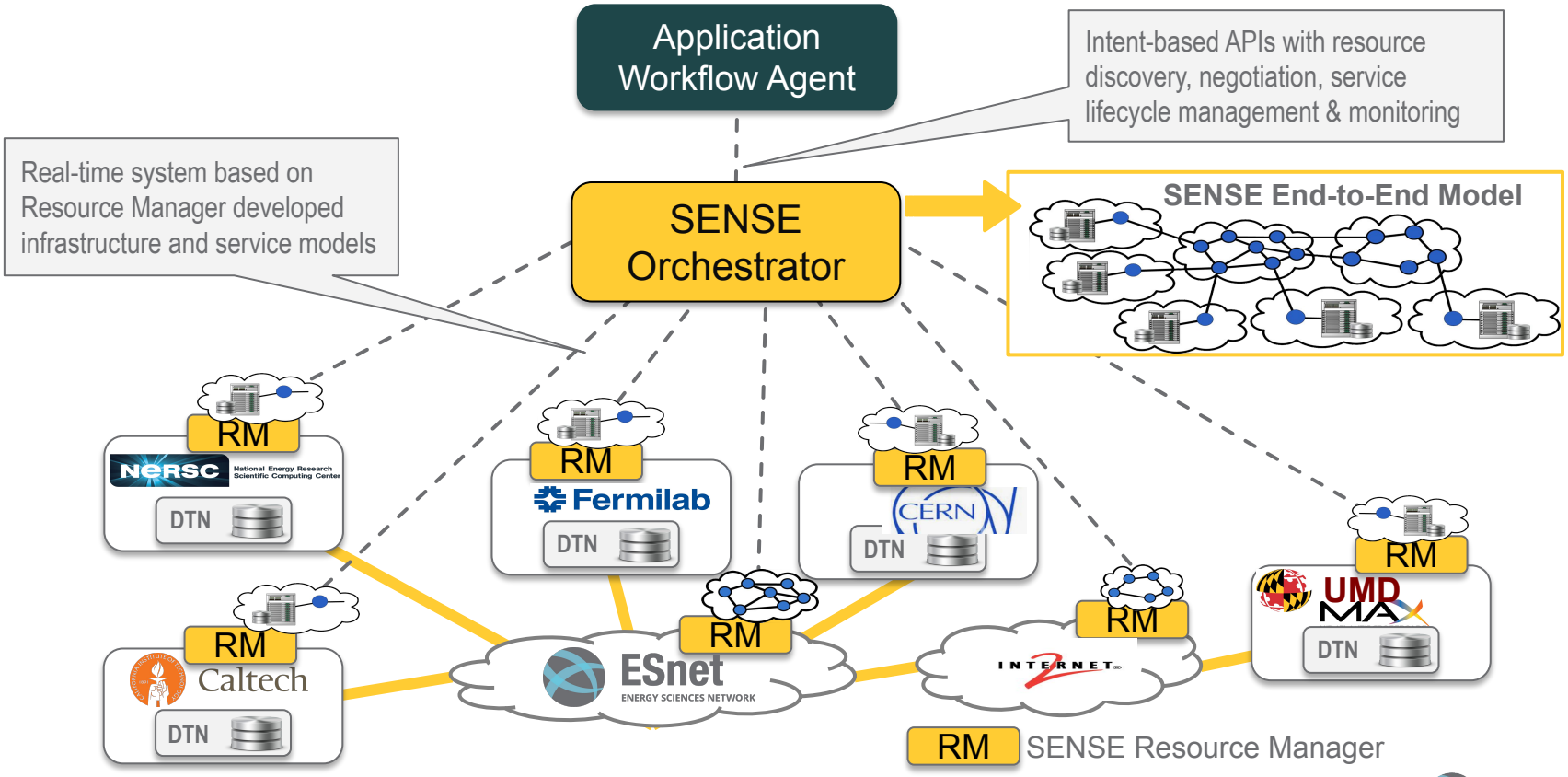
University of Massachusetts Amherst

MGHPCC



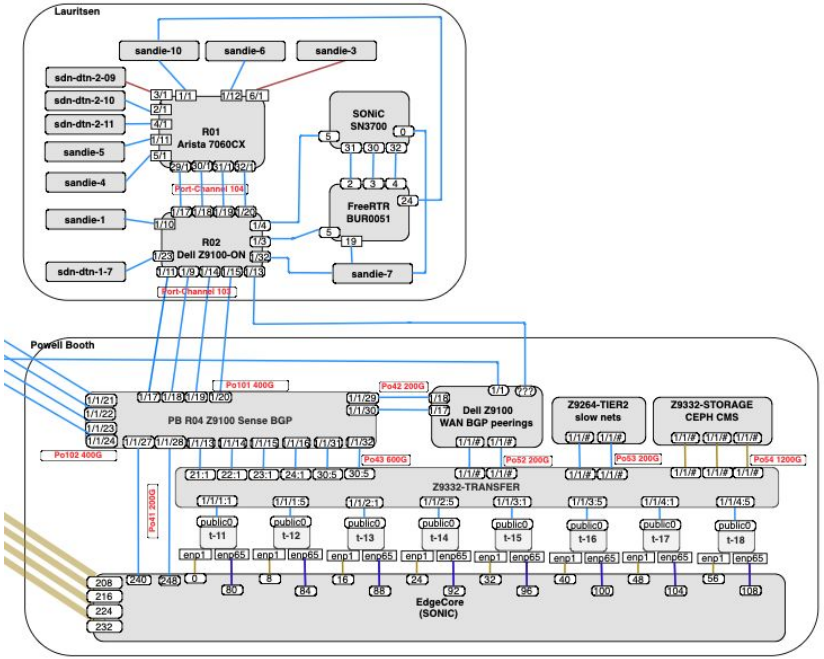
SENSE can be used to find and create these alternative paths.

The SENSE Architecture



ESnet

The devil is in the details



Heterogeneous Device Ecosystem

- Dell OS9 & OS10, Arista, Sonic, FreeRTR

Varied Command-Line Interfaces (CLI)

- Each vendor utilizes distinct CLI syntax and commands

Inconsistent Configuration Management

- Standardizing configurations across different platforms is complex

Limited Unified Ansible Modules

- Need for specialized modules or custom scripts for each device type

Increased Maintenance Effort

- Continuous updates required to handle CLI changes and new device models

Complex Error Handling

- Troubleshooting issues becomes more challenging with diverse device responses

ESnet

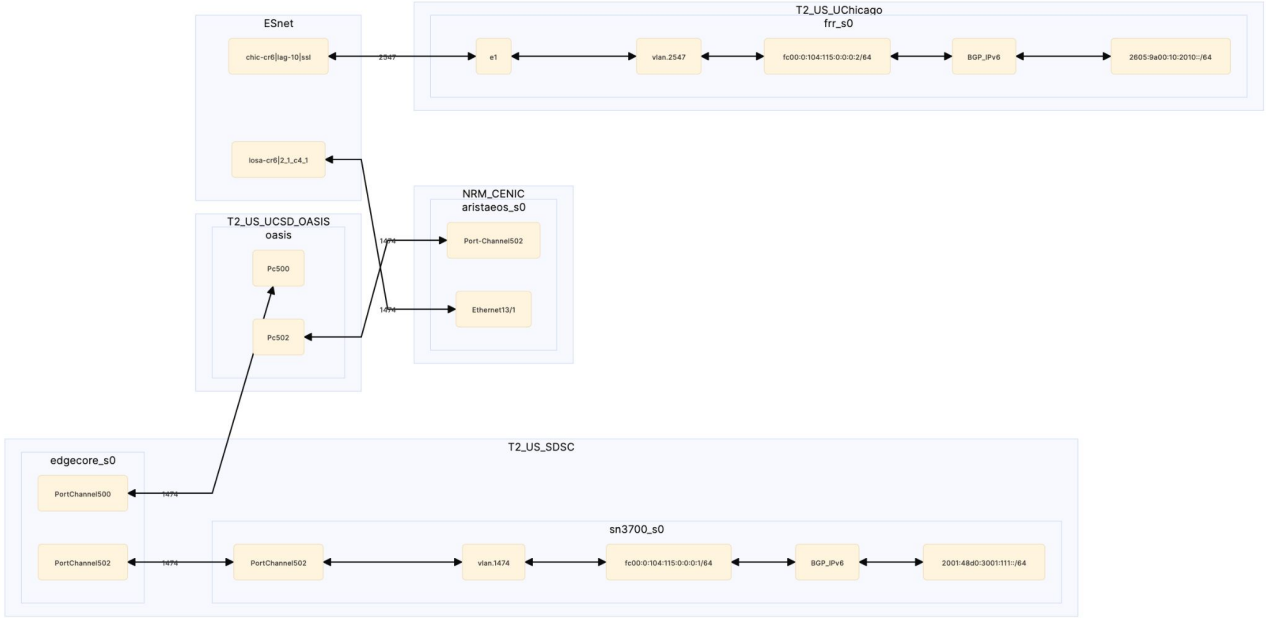
Supported network control devices

Switch OS	Viz in MRML	VLAN Creation	VLAN Translation	Ping/ Traceroute	BGP Control	BGP Multipath	QoS	Comments
RAW	1	0	0	0	0	0	0	RAW Plugin (Fake switch, no control on hardware. use only if instructed by SENSE Team)
Dell OS 9	1	1	0	1	1	0	1 (see note below)	Dell OS9 Ansible Collection
Dell OS 10	1	1	0	1	1	1	0	Dell OS10 Ansible Collection
Azure SONiC	1	1	0	1	1	1	0	Azure SONiC Ansible Collection
Arista EOS	1	1	0	1	0	0	1	Arista EOS Ansible Collection
Juniper Junos	1	1	0	1	1	1	0	Juniper Junos Ansible Collection
FreeRTR	1	0	0	0	0	0	0	FreeRTR Ansible Collection
Cisco Nexus 9/10	1	1	0	1	1	1	0	Cisco Nexus 9 Ansible Collection
FRRouting (FRR)	1	1	0	1	1	1	0	FRRouting Ansible Collection
FRRouting (FRR+VPP)	1	1	0	1	1	1	0	FRRouting Ansible Collection
Mellanox OS	0	0	0	0	0	0	0	Development, expected 2026
Nokia SR OS	0	0	0	0	0	0	0	Nokia SR-OS Collection — Development, expected 2026

https://sdn-sense.github.io

ESnet

End-to-End Monitoring

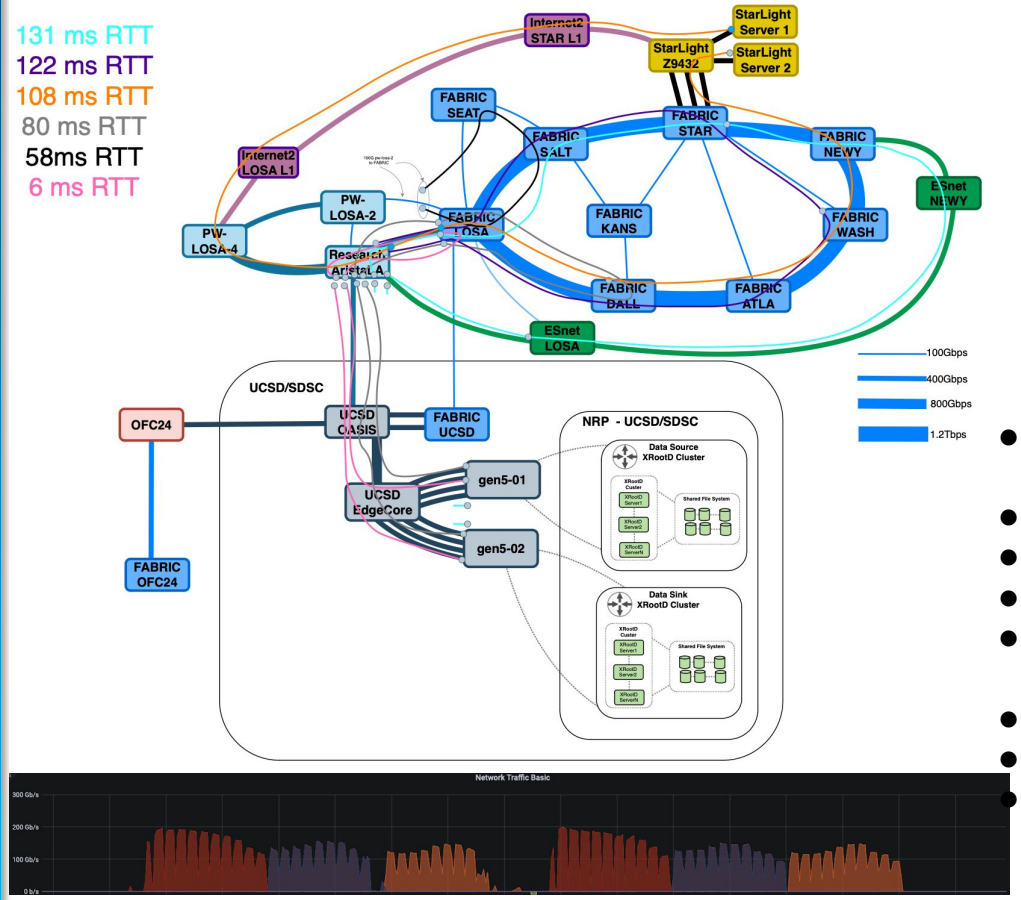


ESnet

Port-by-Port statistics

ESnet

SENSE/Fabric/XRootD/NRP/Kubernetes/Multus



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?
- Optimal sysctl and node tunings?
- What are CPU and Memory Requirements?
- What is the overhead when adding storage (Memdisk, local NVMe Raid, DFS)?

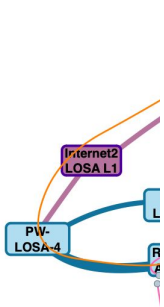
ESnet

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:

21H-TN12R)
old 6430)
-5600)
(Raid0, 42TB)



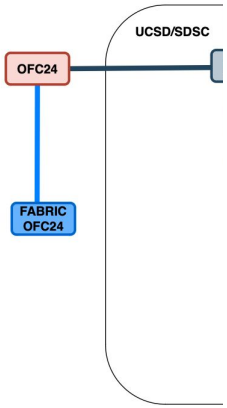
Benchmarking XRootD-HTTPS on 400Gbps Links with Variable Latencies

23 Oct 2024, 08:18
57m
Room 4

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

Speaker

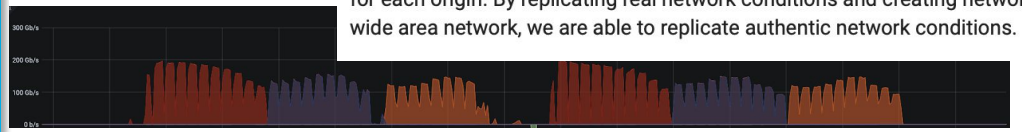
Aashay Arora (Univ. of California San Diego (US))



Description

In anticipation of the High Luminosity-LHC era, there's a critical need to oversee software readiness for upcoming growth in network traffic for production and user data analysis access. This paper looks into software and hardware required improvements in US-CMS Tier-2 sites to be able sustain and meet the projected 400 Gbps bandwidth demands, while tackling the challenge posed by varying latencies between sites. Specifically, our study focuses on identifying the performance of XRootD HTTP third-party copies across multiple 400 Gbps links and exploring different host and transfer configurations.

Our approach involves systematic testing with variations in the number of origins per cluster and CPU, Memory allocations for each origin. By replicating real network conditions and creating network "loops" that traverse multiple switches across the wide area network, we are able to replicate authentic network conditions.



to/from the site

nitations?
throughput?
ect throughput?
ON or OFF for

unings?
y Requirements?
n adding storage
(Memdisk, local NVMe Raid, DFS)?

https://www.epj-conferences.org/articles/epjconf/pdf/2024/05/epjconf_chep2024_01001.pdf

ESnet

Summary on SENSE

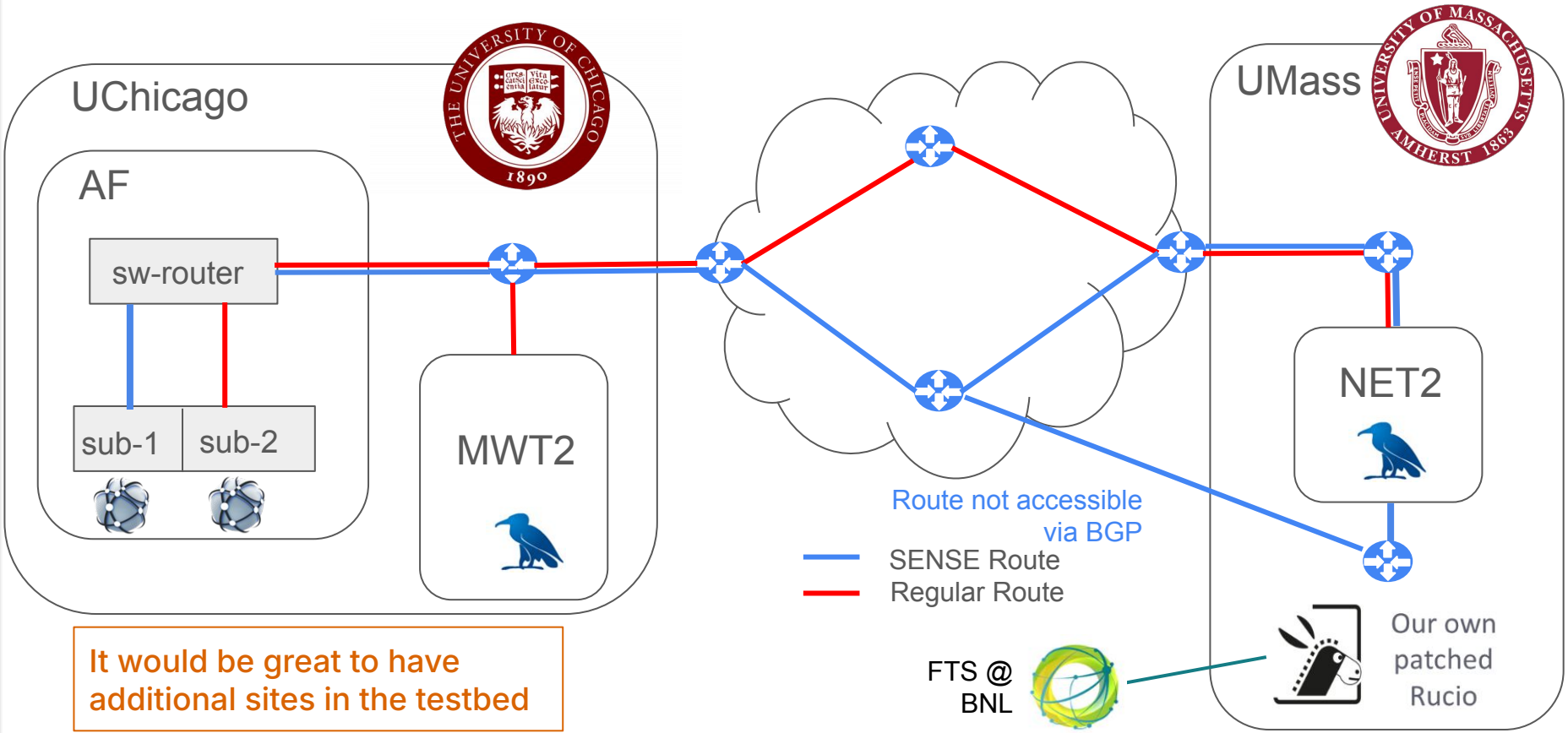
- SENSE allows clients to instruct the network about the priority traffic and SENSE can provide network guarantees end-to-end:
  - bandwidth allocation
  - data flow isolation → fine-grain monitoring
  - temporarily increase network capacity
- CMS plans to show SENSE usage in production by DC27
- ATLAS have two sites with SENSE installation (UMass and UChicago) and Rafael will talk more about this!

https://sdn-sense.github.io

Towards a SENSE testbed in ATLAS

- We deployed a dedicated ATLAS-SENSE Rucio + DMM + SENSE stack in a NRP machine at NET2
  - Rucio server, auth, and all daemons (submitter, finisher, poller, etc.) deployed via Helm
- Integrated the NET2 test RSE and the AF test RSE in the ATLAS-SENSE Rucio instance
  - At UMass NET2, integrated with dCache storage
  - At UChicago AF, no real storage. Just xrootd servers on test machines (it would be great if it could be connected to an actual storage for large test transfers with large files)
- Integrated ATLAS-SENSE Rucio instance with BNL FTS for testing.

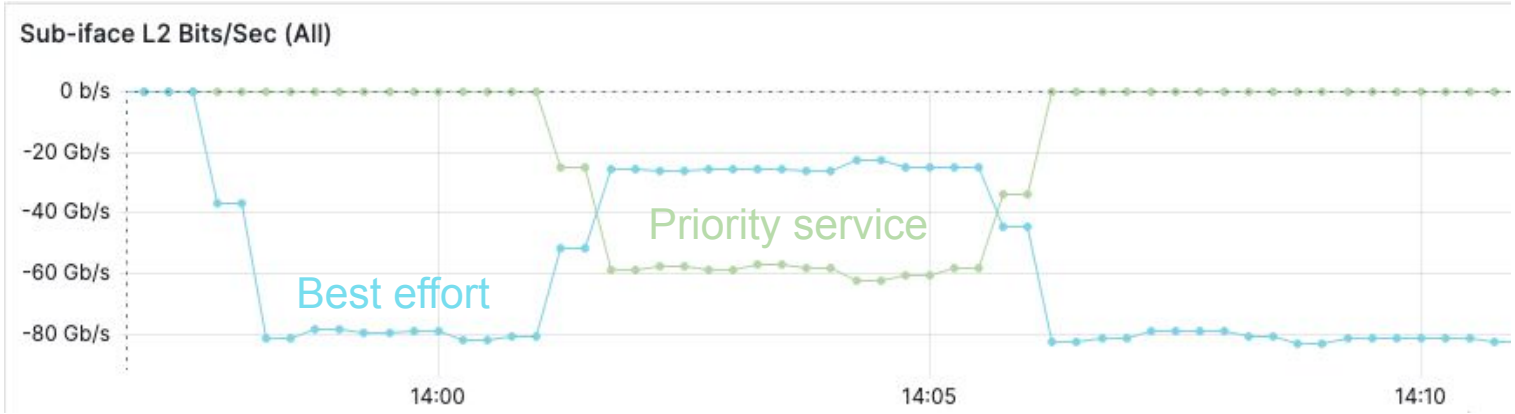
ATLAS-SENSE testbed



It would be great to have additional sites in the testbed

First tests and one outstanding issues

- Initial tests without BNL FTS show expected performance (UMass - UChicago transfer during 2026 mini-data challenge)



- When using the BNL FTS endpoint, the Optimizer throttles transfers (as expected)
  - We are coordinating with the BNL team on the required DN whitelisting so that our Rucio can manage the relevant optimizer settings.

Automating tests between NET2 and AF

- Our goal is to test SENSE transfers between UMass and UChicago under different network utilization conditions:
  - Live bash transfer speed output script to monitor tests
  - We specify 3 pairs of endpoints for 3 sense circuits and creates 3 transfers on these pairs
    - 2 with different priority levels, and a third that changes stochastically to mimic other transfers partially saturating the circuit
    - Stochasticity is achieved by modulating the bandwidth of the sense circuit during a transfer
    - Modulation of a live transfer has been achieved

Towards an evaluation of SENSE in ATLAS

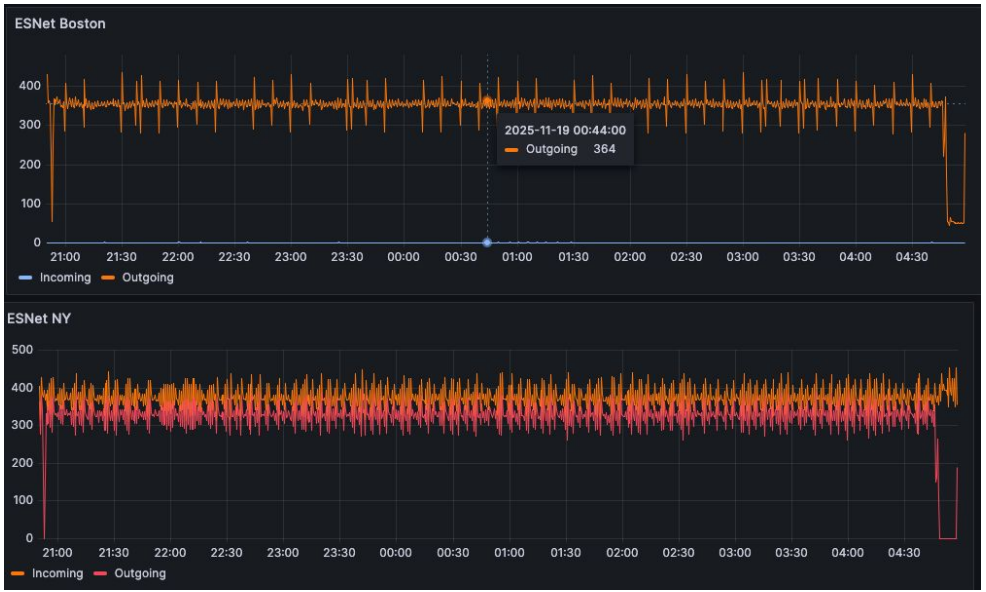
- With our infrastructure we can **quantify** the transfer time with and without SENSE of realistic HL-LHC datasets under different assumptions of network utilization that may happen in the future.
  - SENSE provides uniform, predictable and accountable transfer times.
- The testing transfer times are being fed into ATLAS grid simulation (done in collaboration with Fatih Akman and Verena Martinez as part of the REDWOOD efforts) to evaluate gains in realistic workflows.

Acknowledgments

- I would like to acknowledge that, at UMass, the work presented here has been done by non US-ATLAS facility people who have been excited to collaborate with the NET2 team (Eduardo), the AF team (Fengping), and the SENSE team (Diego, Justas, Tom, Xi, ...) on this project
- A large part of the work has been done by Remy Steele with important contributions from Fatih Akman and Jessa Westclark.

Backup Slides

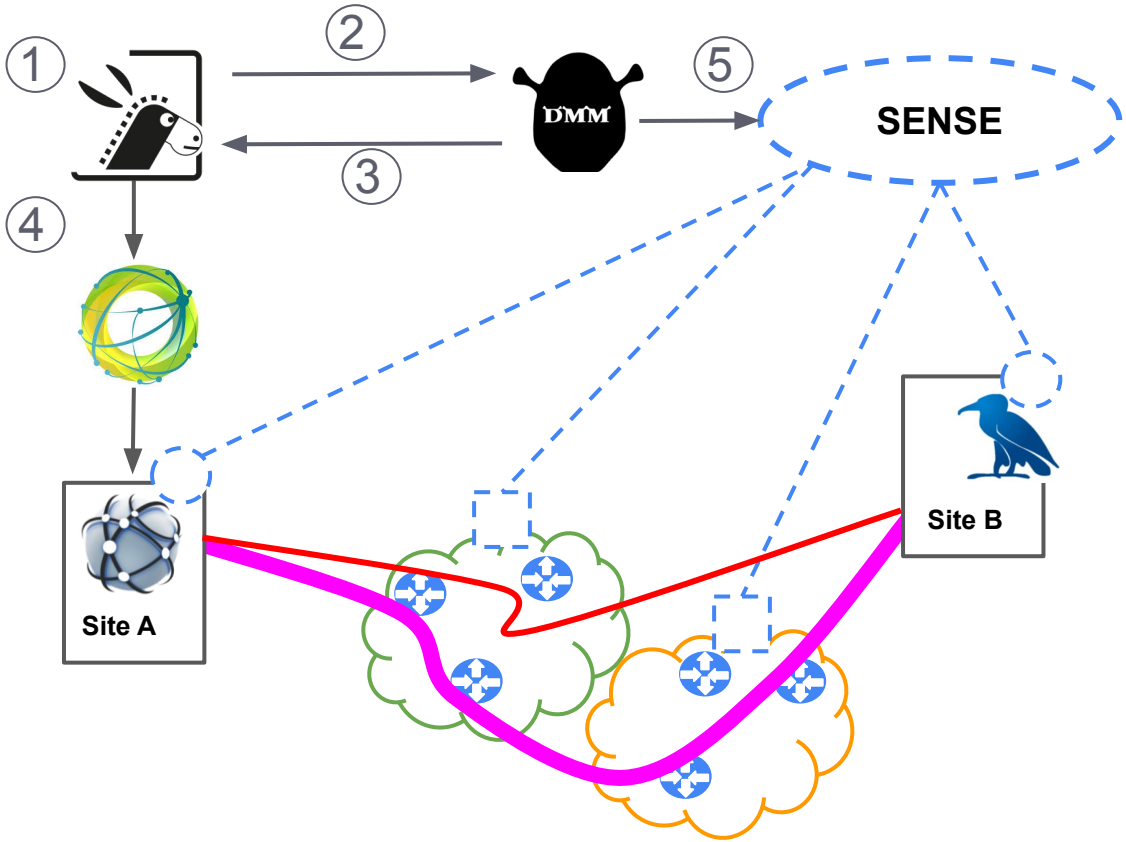
Story of Success: UMass



UMass performing 1.1Tbps at SC25.
Top: UMass/Boston => SC25 400 Gbps
Bottom: UMass/NY <=> SC25 400 + 300 Gbps

Using SENSE we were able to configure UMass to temporarily make use of their backup network path to increase their network capacity from 400 to 800 Gbps (symmetric) for the duration of the demonstration.

What Makes SENSE? End-to-End Orchestration



1. A user creates a Rucio rule with the ***SENSE*** flag (“I want priority”).

2. Rucio contacts *Data Movement Manager* to get new pair of **endpoints**.

3. These are the **SENSE network touchpoints** on the **same sites** but in **special IPv6 subnets**.

4. Rucio-FTS transfers now **bind to** the SENSE priority path by replacing the old endpoints with the new ones.

5. Behind the curtain, *DMM* requests the creation of the SENSE path.

** SENSE error? Transfer remains on or fails back to default path **transparently.***

Network Control - Requirements

From the point of view of a site and in order to achieve the above, SENSE requires access to the following network capabilities:

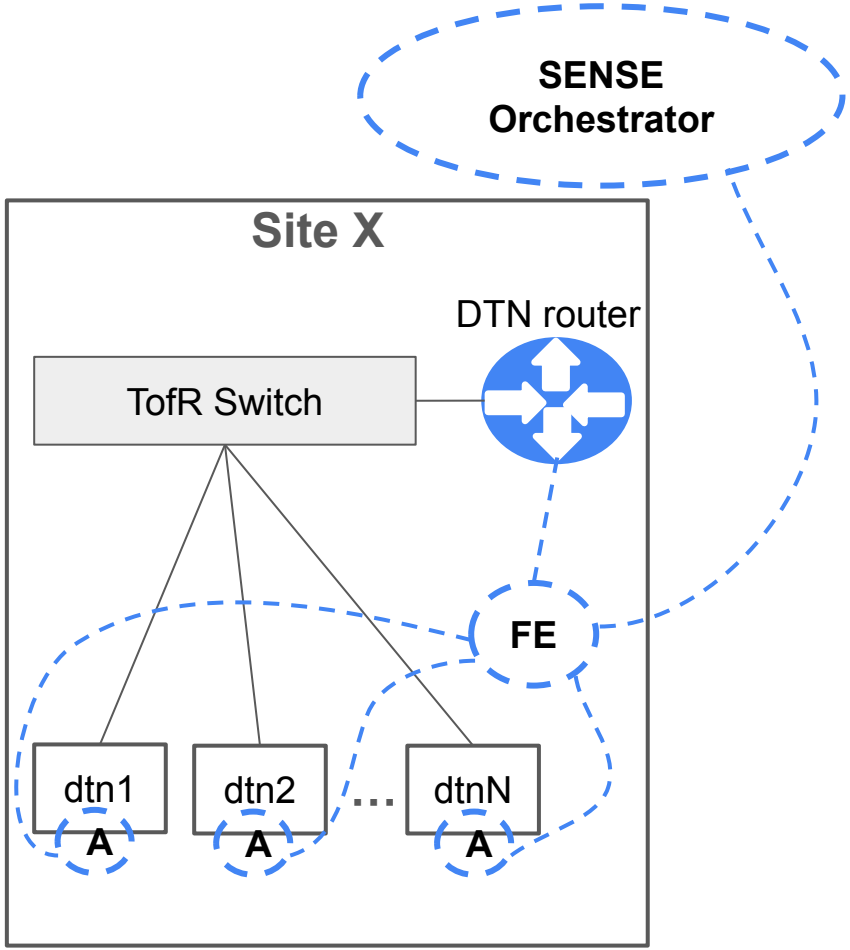
- Modify the routing of the DTNs data flows
  - Manage VLANS to create Point-2-Point L2 network paths
  - Manage BGP announcements to route desired traffic over the above L2
- Configure QoS policies on the DTNs
  - Used to provide bandwidth allocations
  - Done using the Linux Traffic Controller

These requirements are **restricted to a defined set of subnets and VLANS** which are used only for SENSE-managed data flows.

Network Control - SiteRM

At the site level, SENSE requires the deployment of 2 main Site Resource Managers (SiteRM):

1. SiteRM Frontend.
    a. Receives commands from the Orchestrator
    b. Applies changes in the DTN router
    c. Sends commands to the Agents
2. SiteRM Agent.
    a. Runs in the DTNs
    b. Execute commands on the DTN
    c. Enforces QoS policies



More info about SiteRM: https://sdn-sense.github.io/#siterm-resource-manager