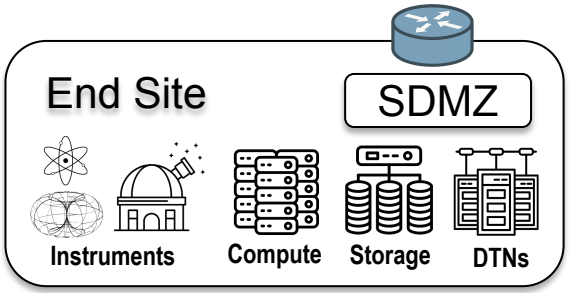
Capability Challenge: SENSE

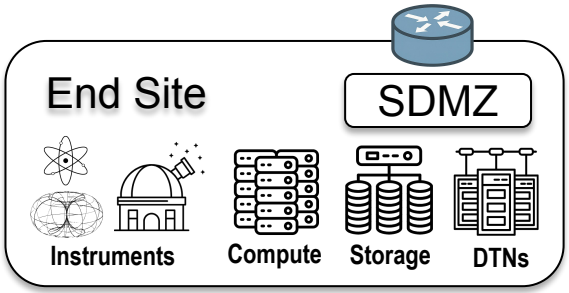
Diego Davila (UCSD), Justas Balcas (ESnet)

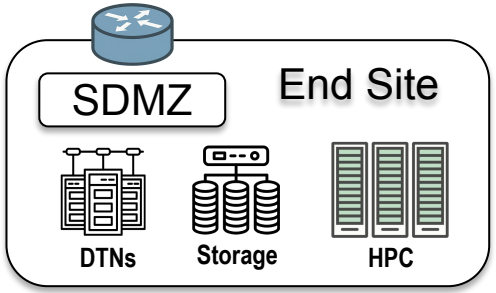
HTC25 - June 5th, 2025


SENSE: Build your own Network with Guarantees

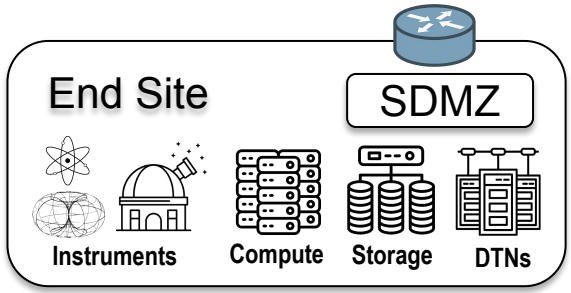


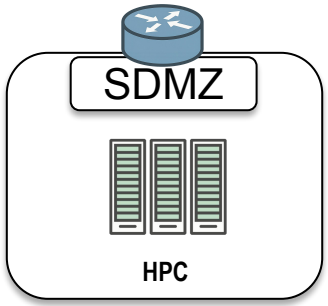
SENSE: Build your own Network with Guarantees

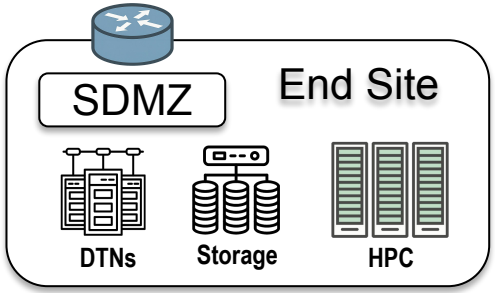




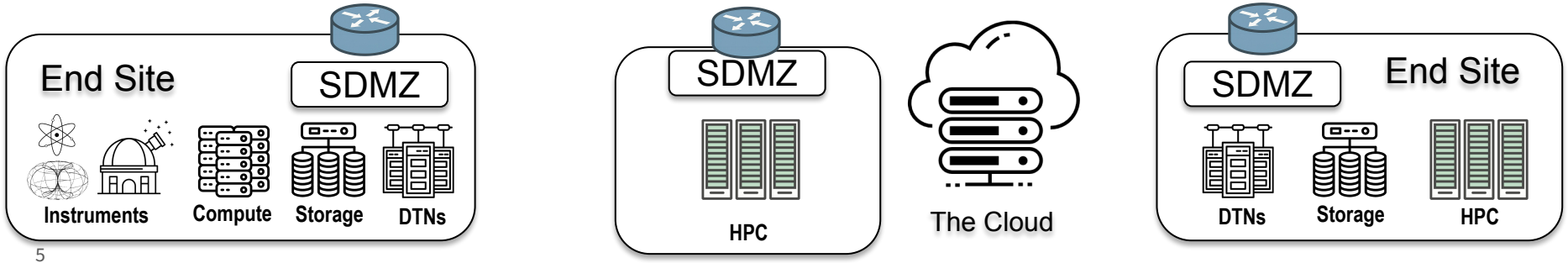
SENSE: Build your own Network with Guarantees



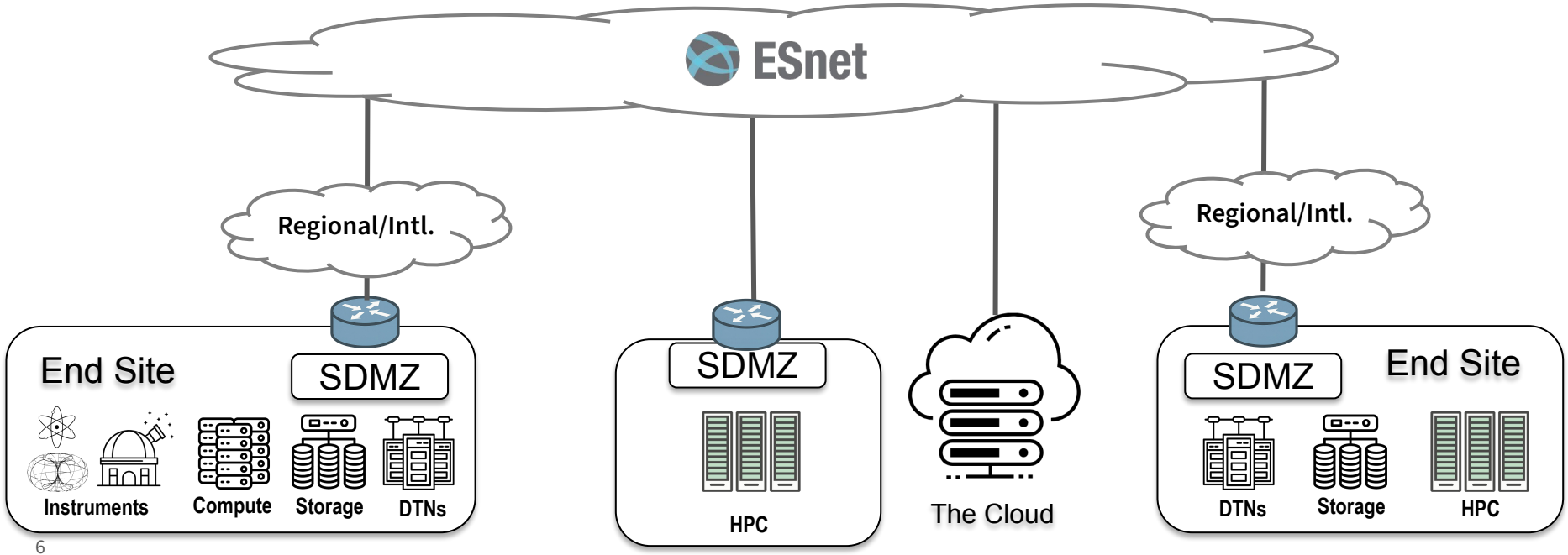




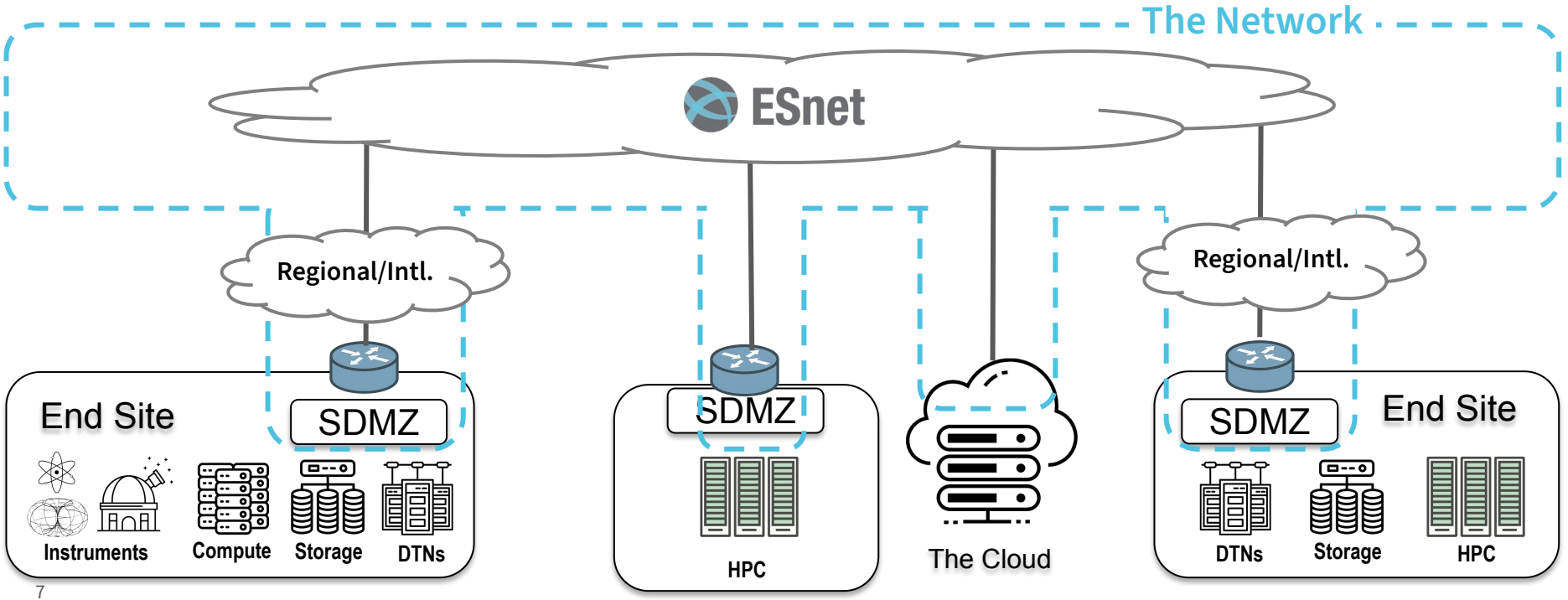
SENSE: Build your own Network with Guarantees



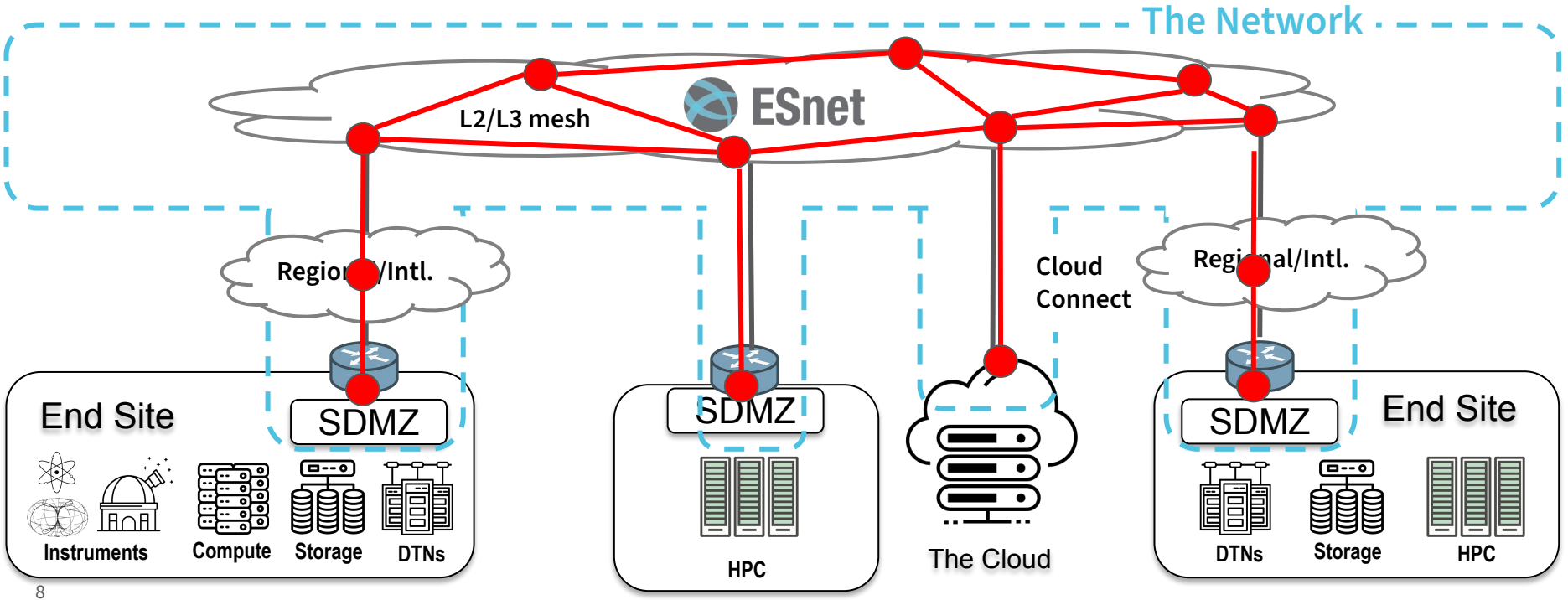
SENSE: Build your own Network with Guarantees



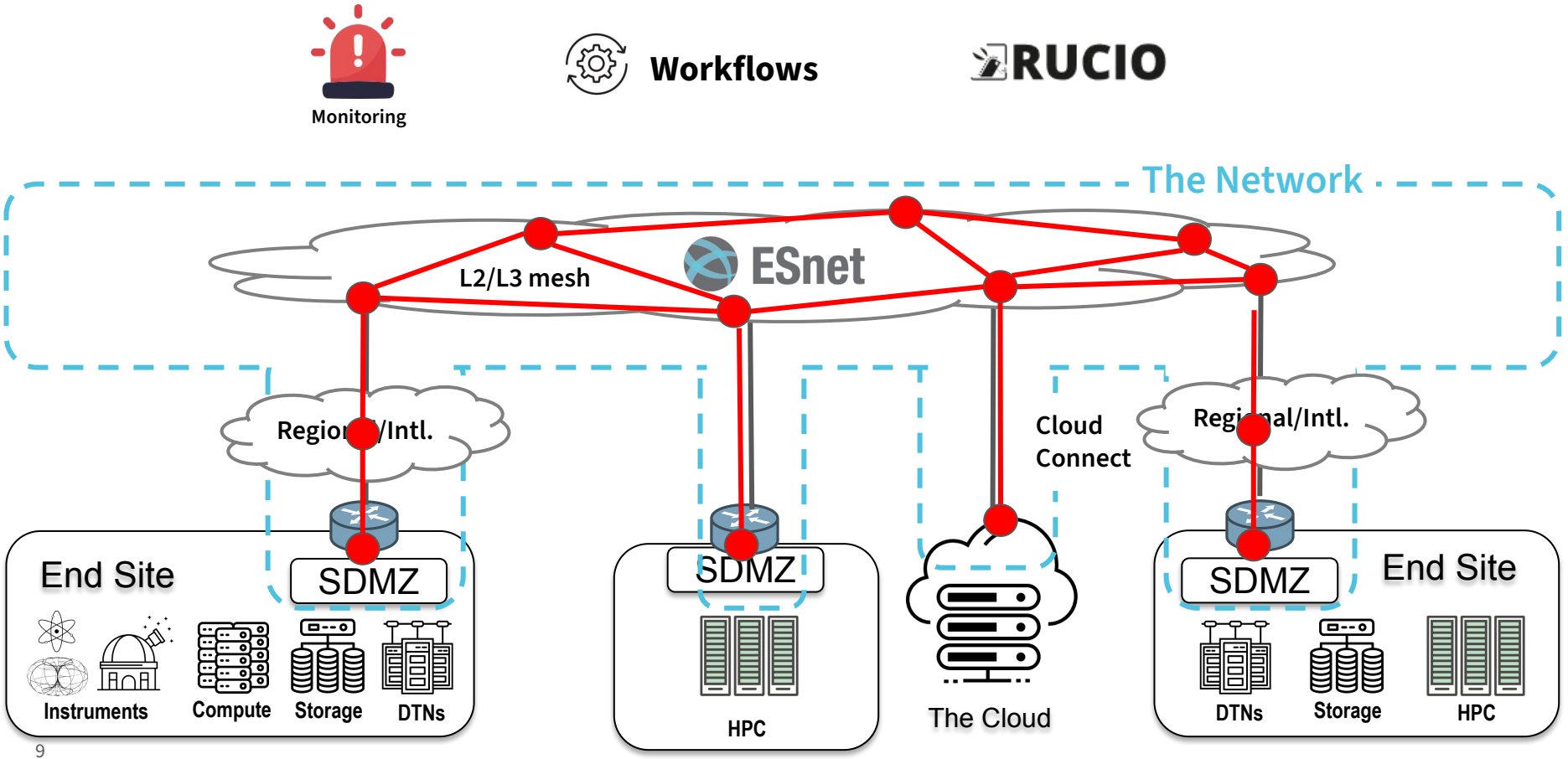
SENSE: Build your own Network with Guarantees



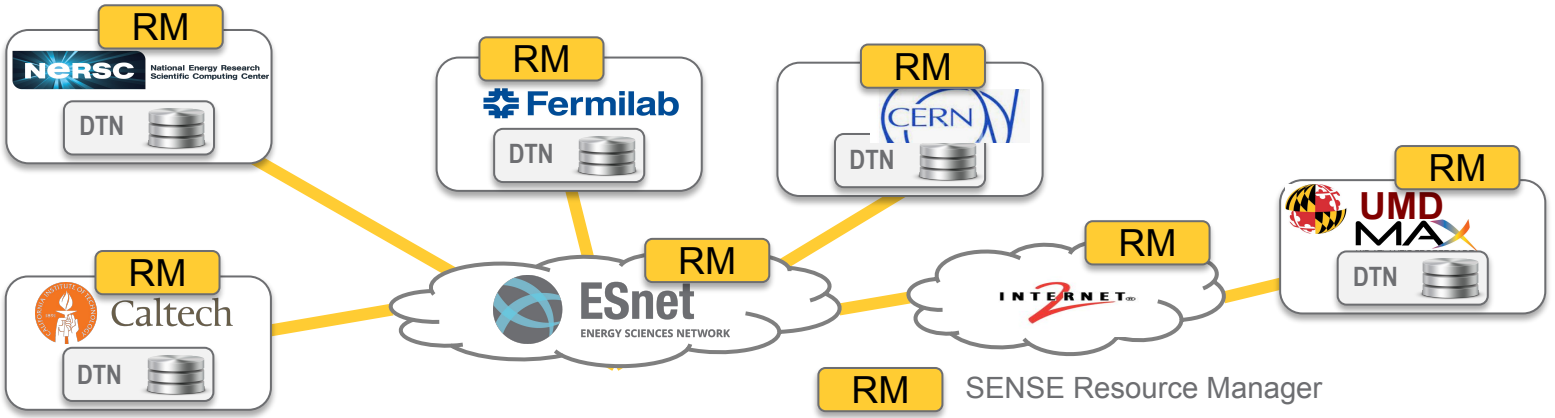
SENSE: Build your own Network with Guarantees



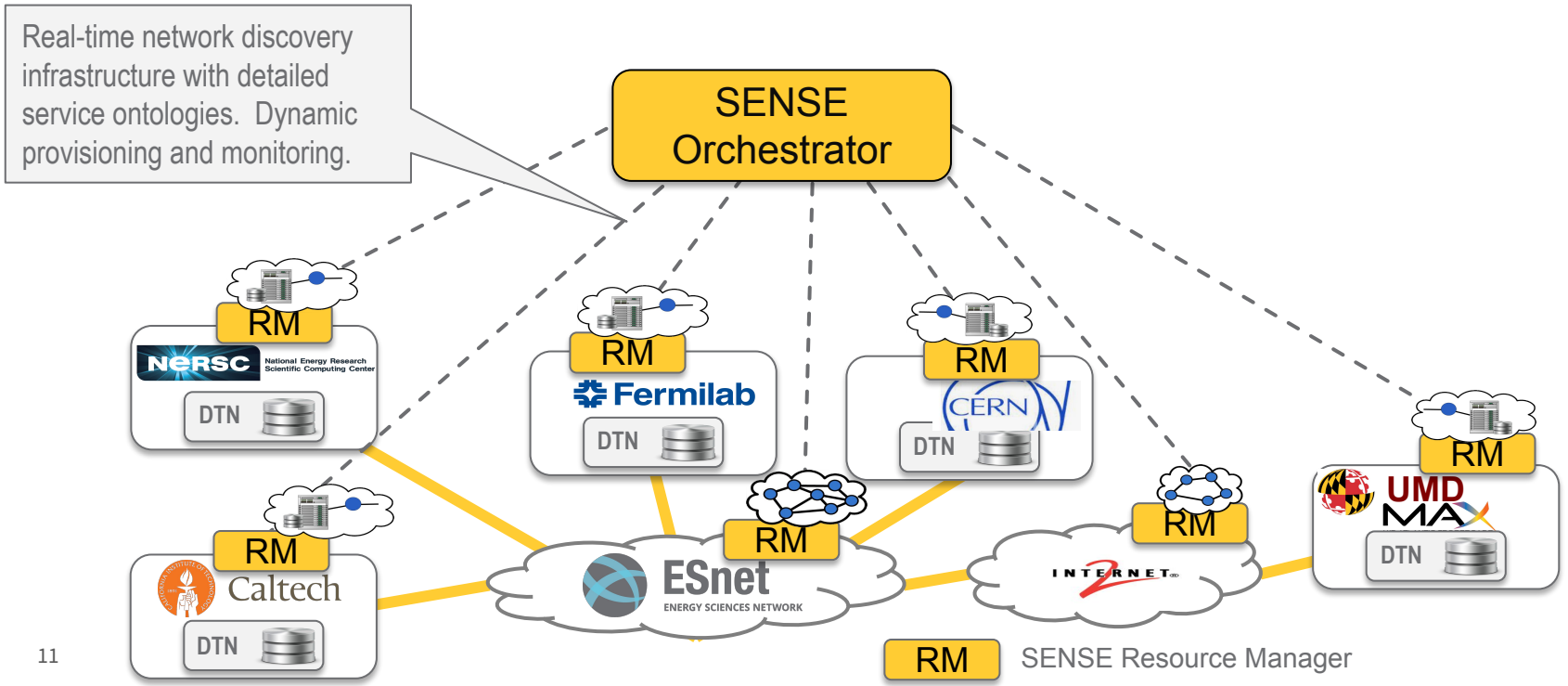
SENSE: Build your own Network with Guarantees



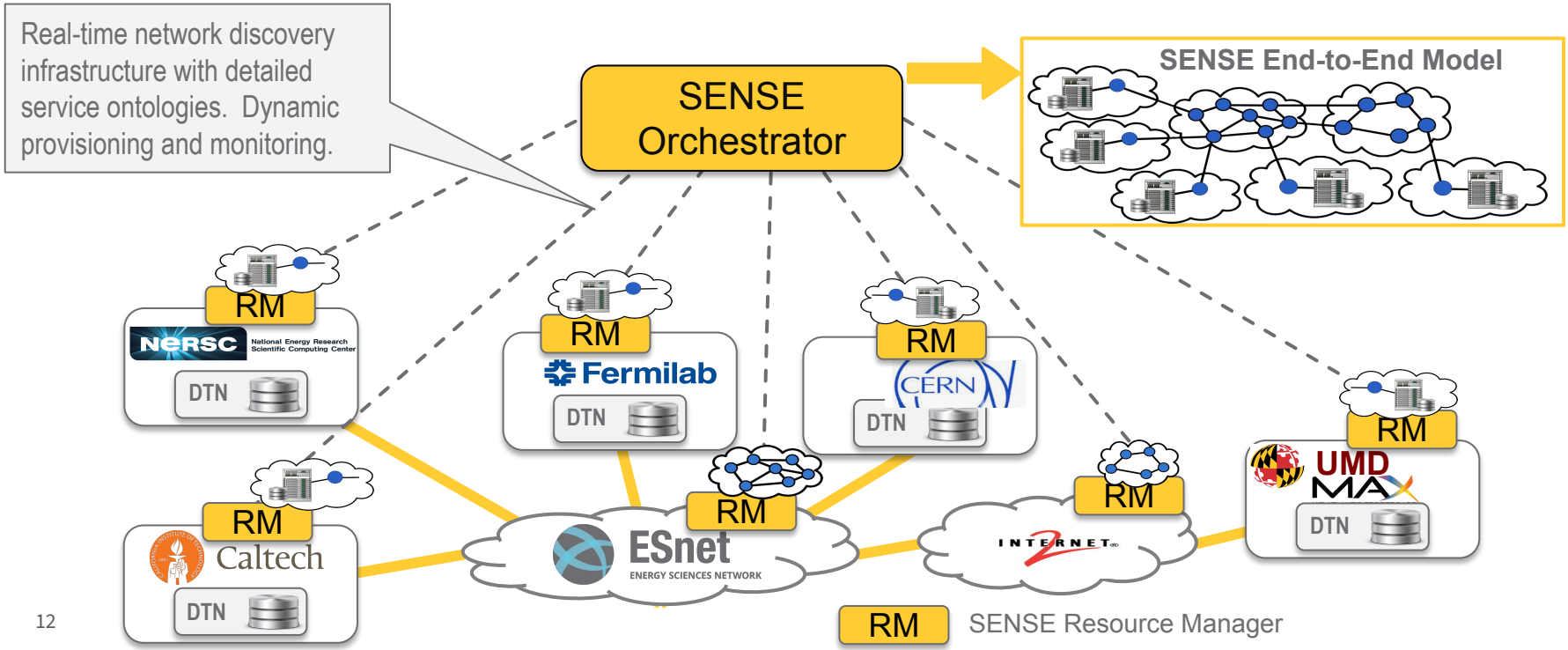
The API-Driven SENSE Architecture



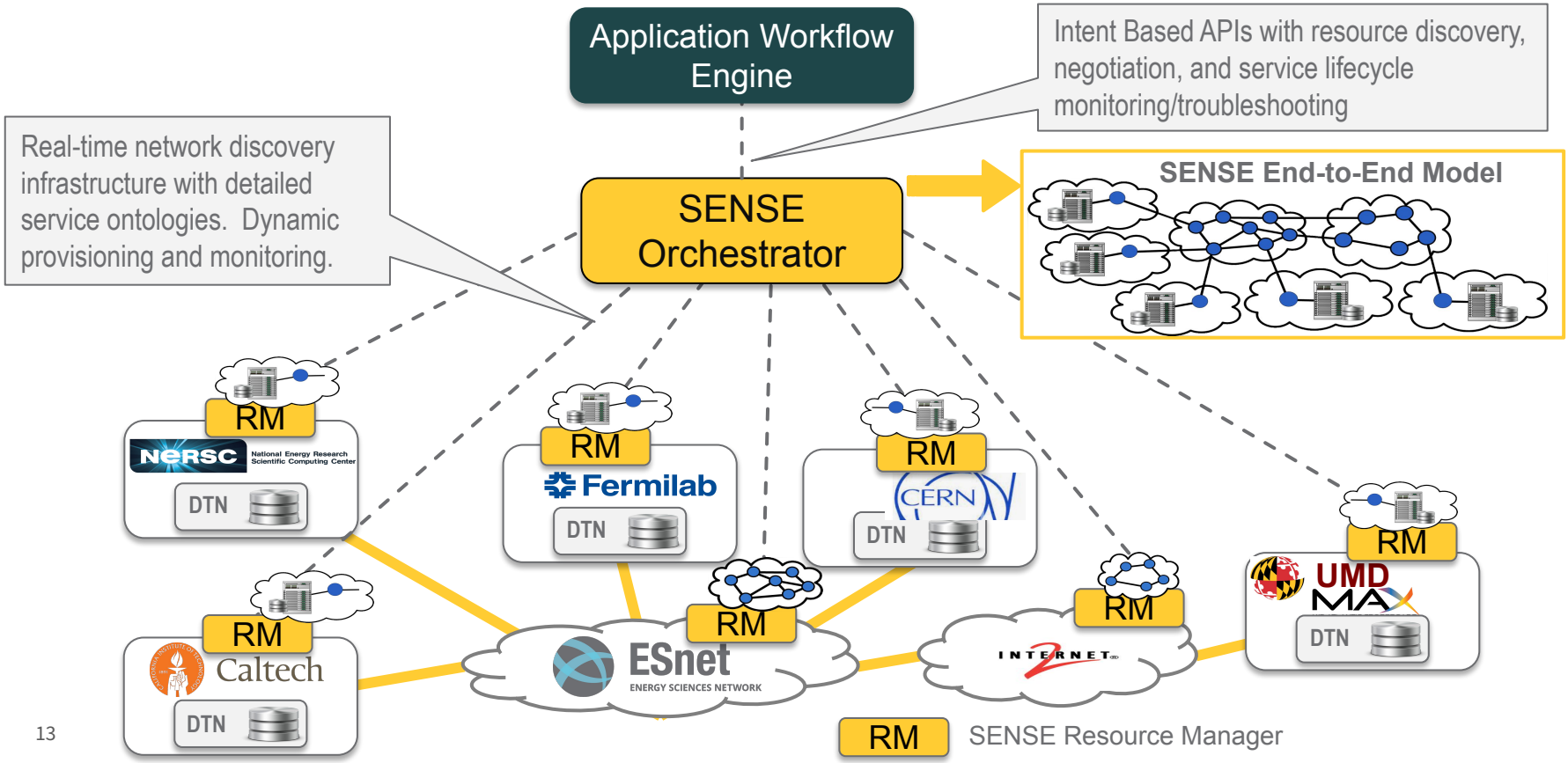
The API-Driven SENSE Architecture



The API-Driven SENSE Architecture

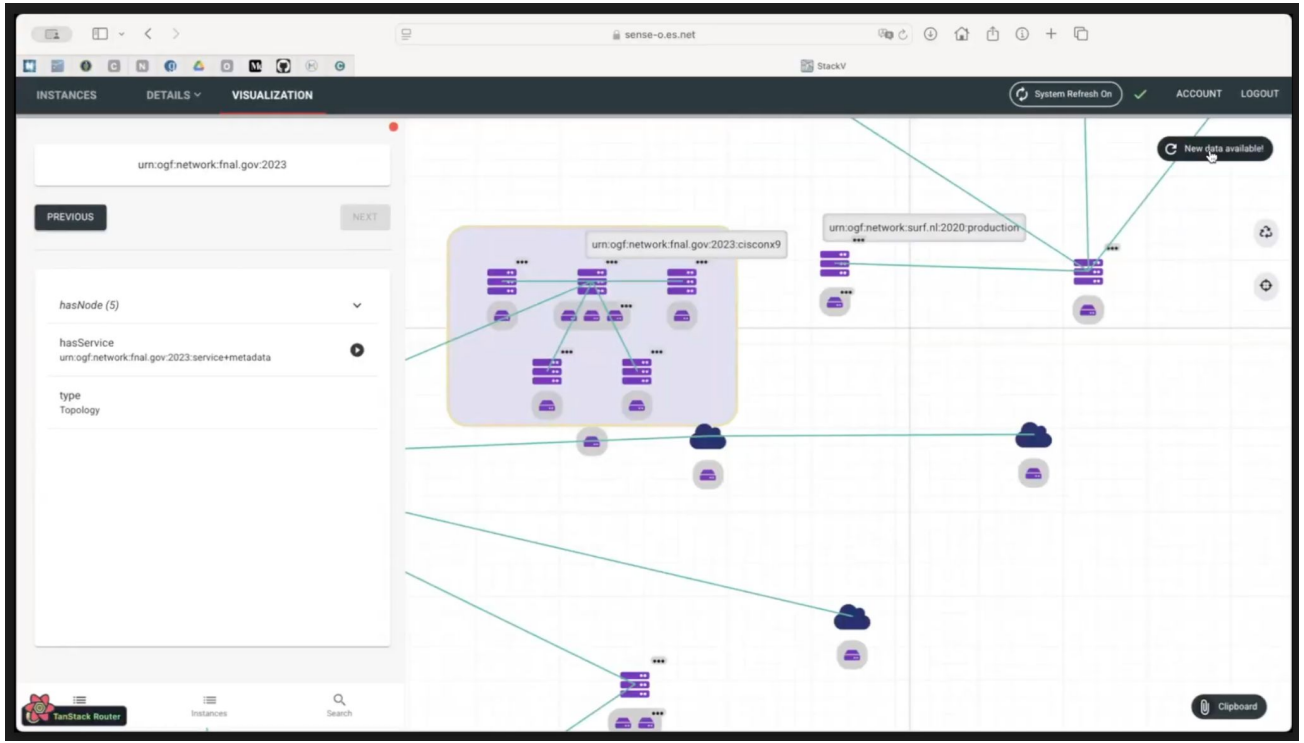


The API-Driven SENSE Architecture

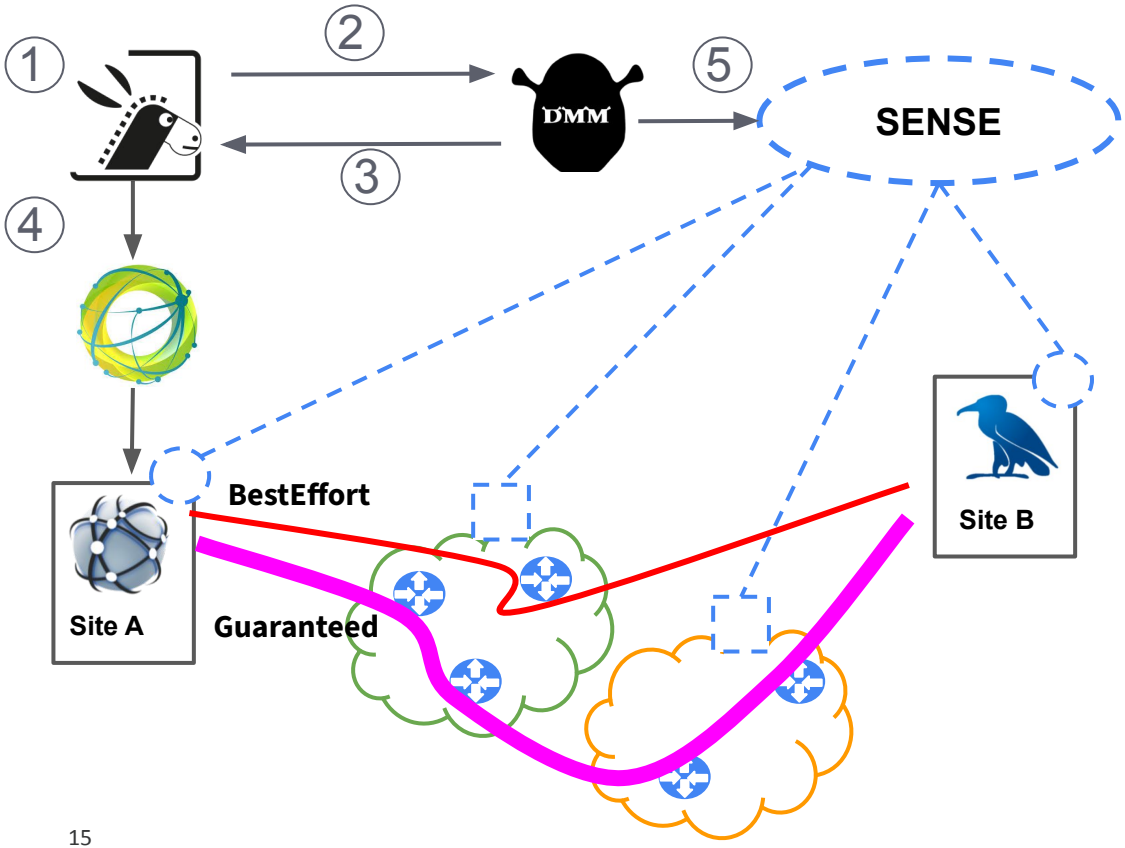


To learn more about SENSE

SENSE Tutorial



Rucio with SENSE



1. A user creates a Rucio rule with the –sense flag.

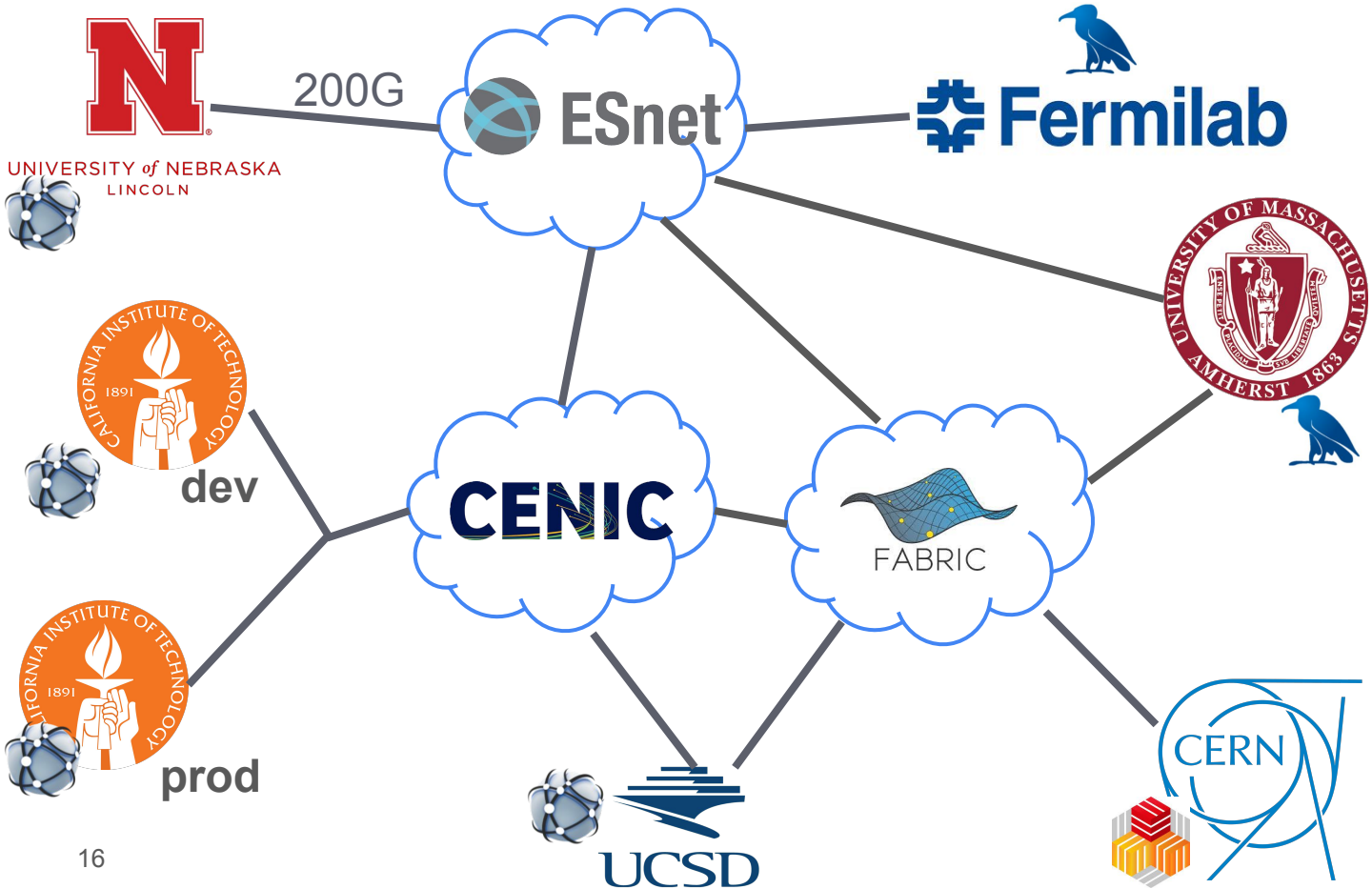
2. Rucio contacts DMM to get a new pair of endpoints

3. DMM pick a free endpoint on each site and gives them back to Rucio. These endpoints point to the **same sites** but on **different subnets**

4. Rucio replaces the old endpoints with the new ones and continues as usual i.e. sends the list of files to transfer to FTS

5. Behind the curtain DMM requests the creation of the special path

Current testbed



400 Gbps Everywhere but Nebraska

Our own patched Rucio

Fts(dev) @FNAL

Capability mini-challenge: Path to production

The main idea behind the “Capability Mini-challenge” was to show **new technology,** previously shown to **improve data transfers** in testbeds, **in production**

Our (SENSE/Rucio team) plan was to show the use of SENSE as close to production as possible

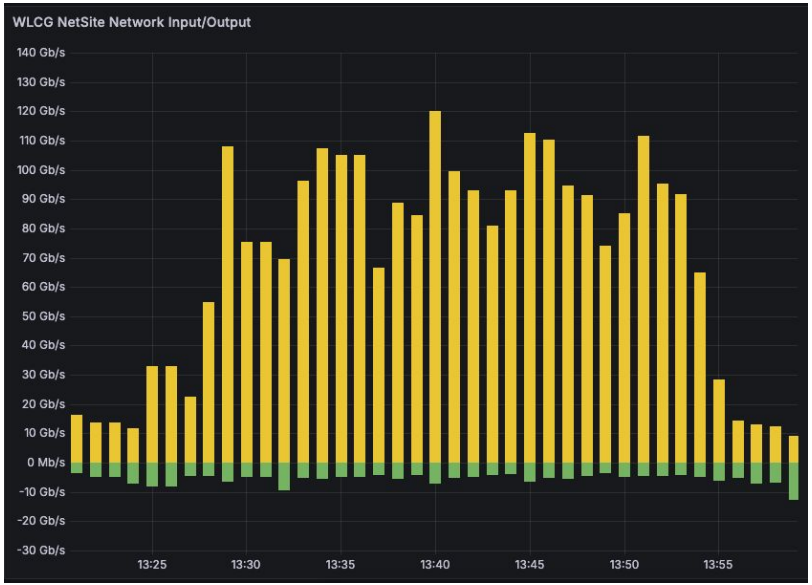
- We picked our 2 more advanced sites: Nebraska and Caltech
- Mounted their production File System
  - We use paths outside the CMS namespace to avoid possible conflicts
- Triggered a Rucio/SENSE 100+ Gbps data flow

SENSE in Production

In **Caltech**, SENSE is fully integrated into their production system.

Nebraska uses separate DTNs to support SENSE but they mount the same Ceph FS used by production

The next step is to patch CMS Rucio to allow SENSE supported transfers between these 2 sites



Nebraska => Caltech Rucio/SENSE data flow
http://monit-grafana-open.cern.ch/goto/UTbaE2fHg?orgId=16

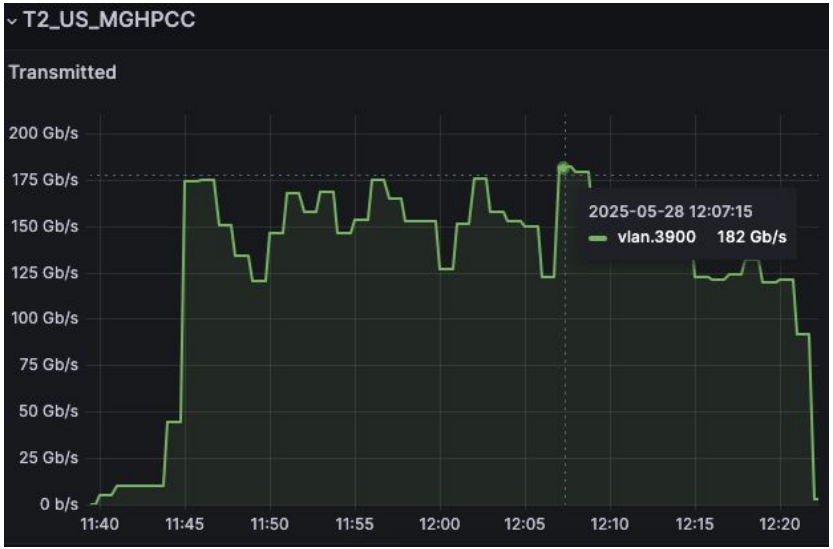
Others sites getting close to production

UMASS

We deployed a **single** 400 Gbps capable node into the SENSE testbed and carried out **initial tests** to SDSC, reaching a max throughput of 180 Gbps (RTT=70ms)

The next steps are:

- Repeat tests aiming to higher throughput
- Deploy a dCache proxy in front of their production storage



UMASS => SDSC, Rucio/SENSE data flow
https://autogole-grafana.nrp-nautilus.io/d/000000048/throughput-by-interface?orgId=1&from=1748457567776&to=1748460135776

Others sites getting close to production

FNAL

Using a dCache setup of 3 data servers + 2 proxies and running hundreds of parallel transfers using gfal-copy we were able to pull **190Gbps** out of FNAL.

The next step is to deploy a set of proxies in front of their production dCache cluster

FNAL => SDSC, Rucio/SENSE data flow
https://autogole-grafana.nrp-nautilus.io/d/000000048/throughput-by-interface?orgId=1&from=1747679751957&to=1747681228407

Bonus of the SENSE testbed

Not doing only-SENSE R&D but also testing new tech, system tunings, etc on a highly controlled Long Fat Network

Our usual tests show that we can do high throughput with FEW powerful servers

- We are able to sink ~200 Gbps into SDSC using only 2 gen5(*) nodes and 5 servers in FNAL as source (RTT ~60ms)
- Caltech has demonstrated 300Gbps + using only 8 transfer nodes

(*)PCIe Gen-5 ConnectX-7

What’s next (R&D)

- Benchmark every site in the testbed and document their maximum data transfer rate capability (with and without SENSE)
- Get more ATLAS sites involved
  - UMass ready to play, UChicago getting there
- CERN
  - Show high bandwidth to US (high RTT)
- Adding all US sites into the testbed
  - Currently working with Purdue and Wisconsin
- We need experience with EOS (hope to get this from Purdue)

What’s next: Path to Production

- Caltech and UNL already in prod
- Focus on getting SENSE supported in production by CMS Rucio
- Getting as many sites using SENSE in prod as possible
  - FNAL and UCSD where we have Network control are obvious candidates
- Sites with Ceph, Hadoop, XRootD and dCache are also good candidates

Acknowledgements

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

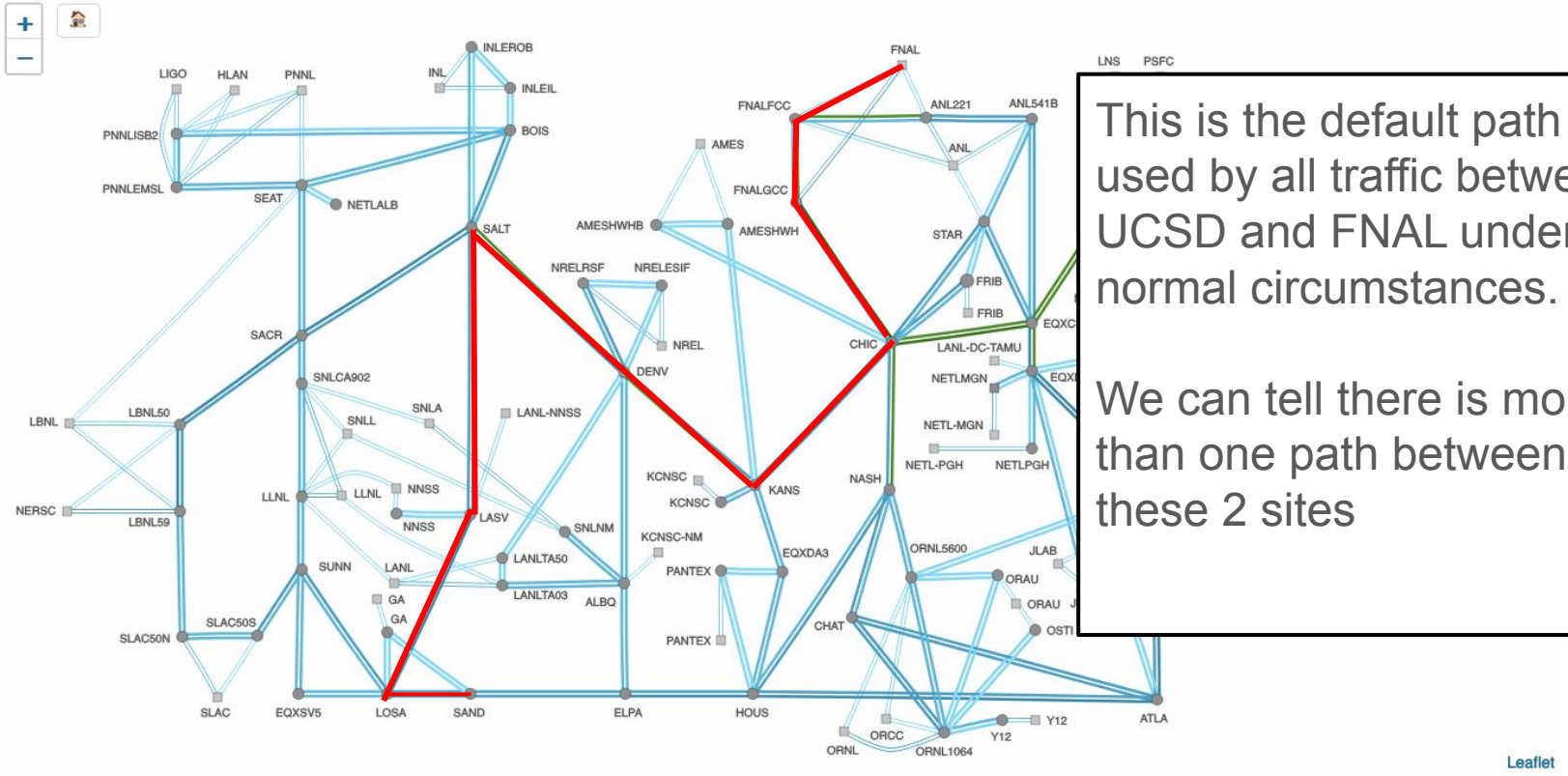
This ongoing work is partially supported by the US National Science Foundation (NSF) Grants OAC-1836650, PHY-2323298, PHY-2121686 and OAC-2112167. In addition, the development of SENSE is supported by the US Department of Energy (DOE) Grants DE-SC0015527, DE-SC0015528 and DE-AC02-07CH11359

Finally, this work would not be possible without the significant contributions of collaborators at ESNet, Caltech, FNAL and SDSC.

Thanks!
Questions?

Background

Traceroute UCSD => FNAL

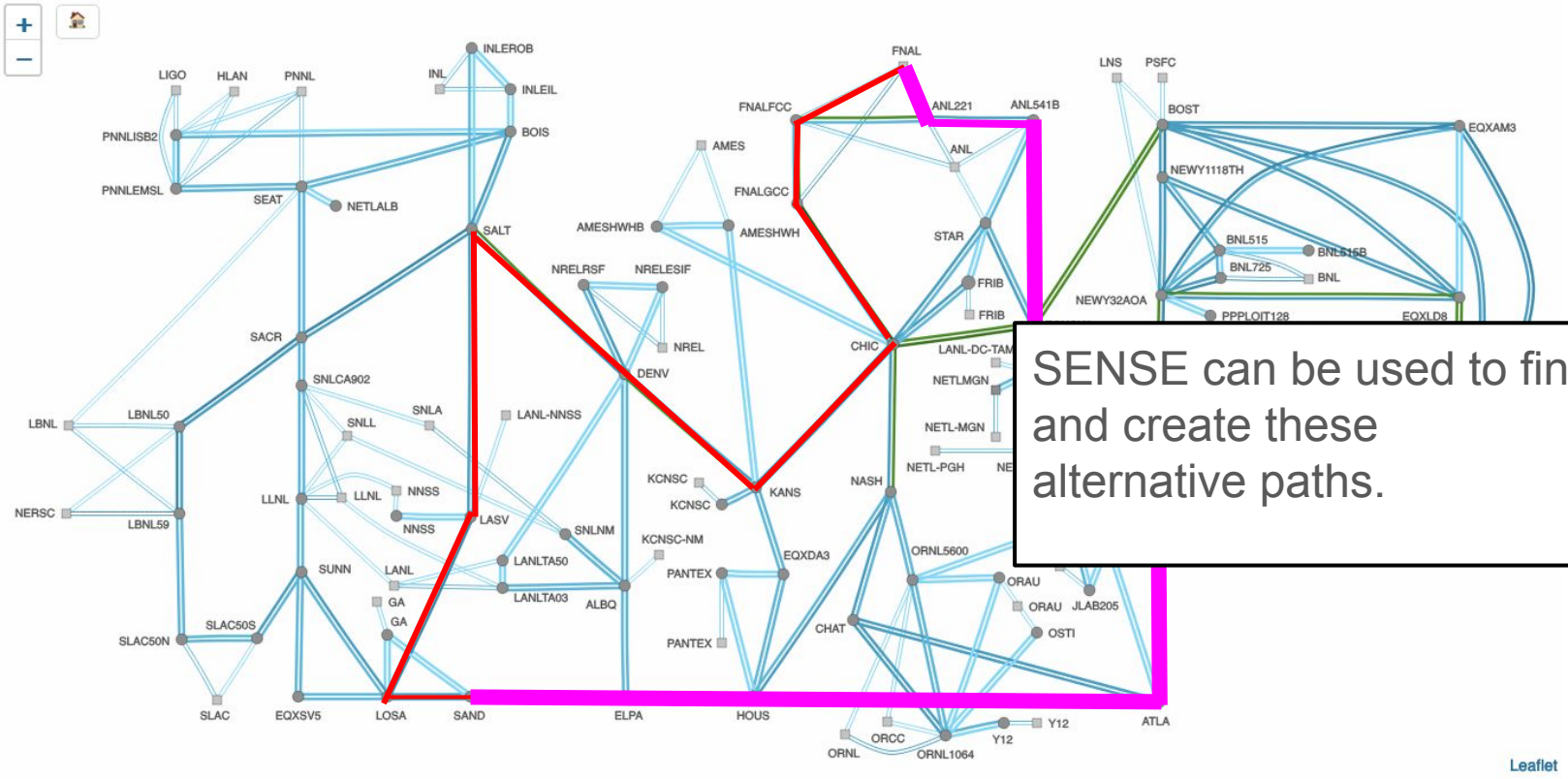


This is the default path used by all traffic between UCSD and FNAL under normal circumstances.

We can tell there is more than one path between these 2 sites

ESnet

Alternative path UCSD => FNAL



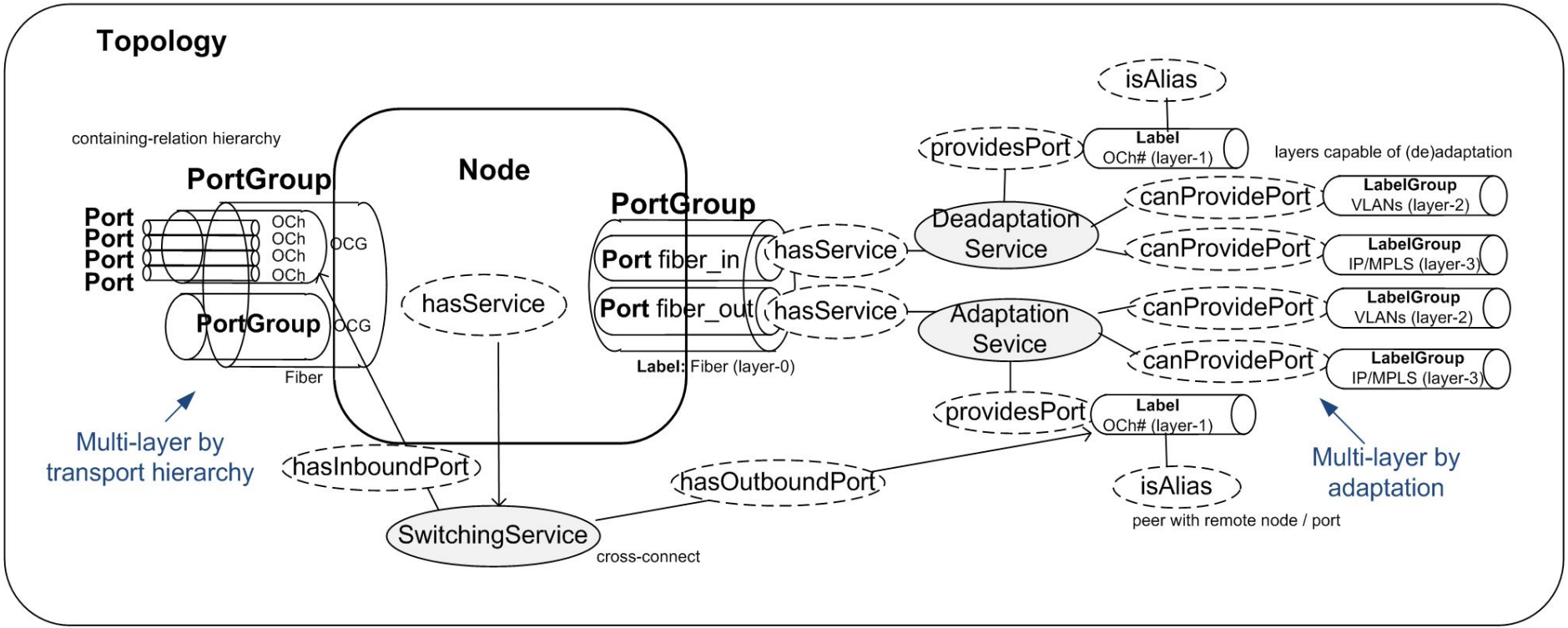
ESnet

Wish list

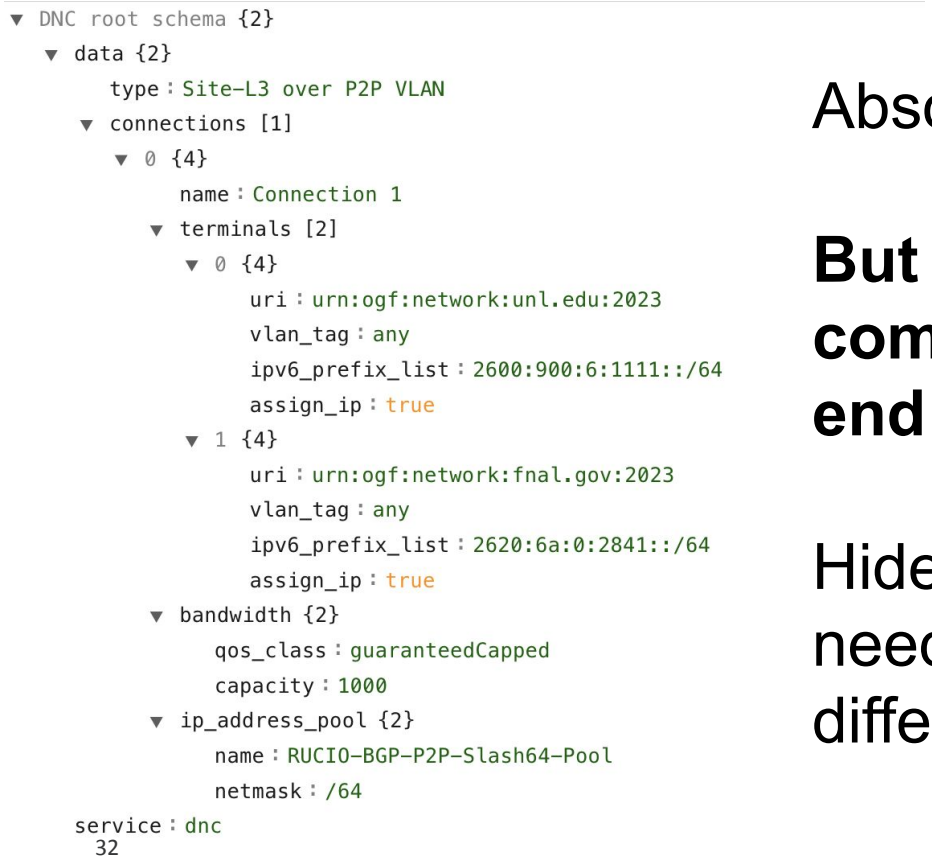
- All US sites in the testbed
  - Even 1 node would do
- Tell us about your switch models

BACKUP SLIDES

How complex it is?!



Hard for us, easy for end-user!

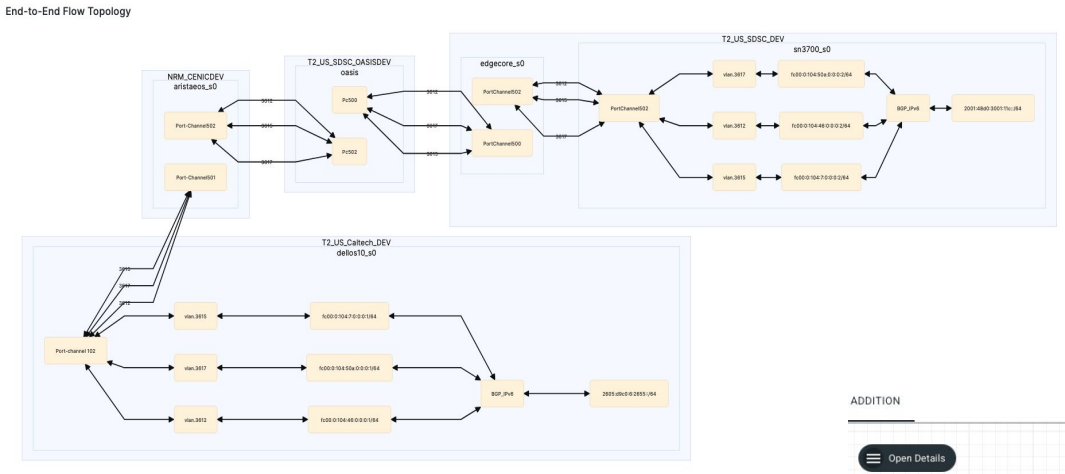


Absolutely, this is complicated stuff!

But our job is to take this complexity and simplify it for the end user.

Hide complexity the user does not need to see and normalize the differences.

Build it as you want!



L2/L3 VPNs
Single Path, Multipath (ECMP/UCMP)
IP assignment on hops;
BGP, ACL, QoS control;
Hop-by-Hop monitoring:

- In/Out/Err/Discards
- Unicast/Multicast