Automated Network Services for Exascale Data Movement

Frank Würthwein, Jonathan Guiang, Aashay Arora, **Diego Davila**, John Graham, Dima Mishin, Thomas Hutton, Igor Sfiligoi, Harvey Newman, Justas Balcas, Preeti Bhat, Tom Lehman, Xi Yang, Chin Guok, Oliver Gutsche, Phil Demar, Marcos Schwarz

5th Rucio Community Workshop, Nov 2022

ESnet

Motivation

Scientific collaborations reaching the Exascale

- LHC experiments doing **Millions of transfers** every day, will increase as we approach the High Luminosity LHC
- Lots of **transfer failures** currently
- Understand failures can be hard e.g. **http 500 error**
- Understanding **poor performance** is even harder

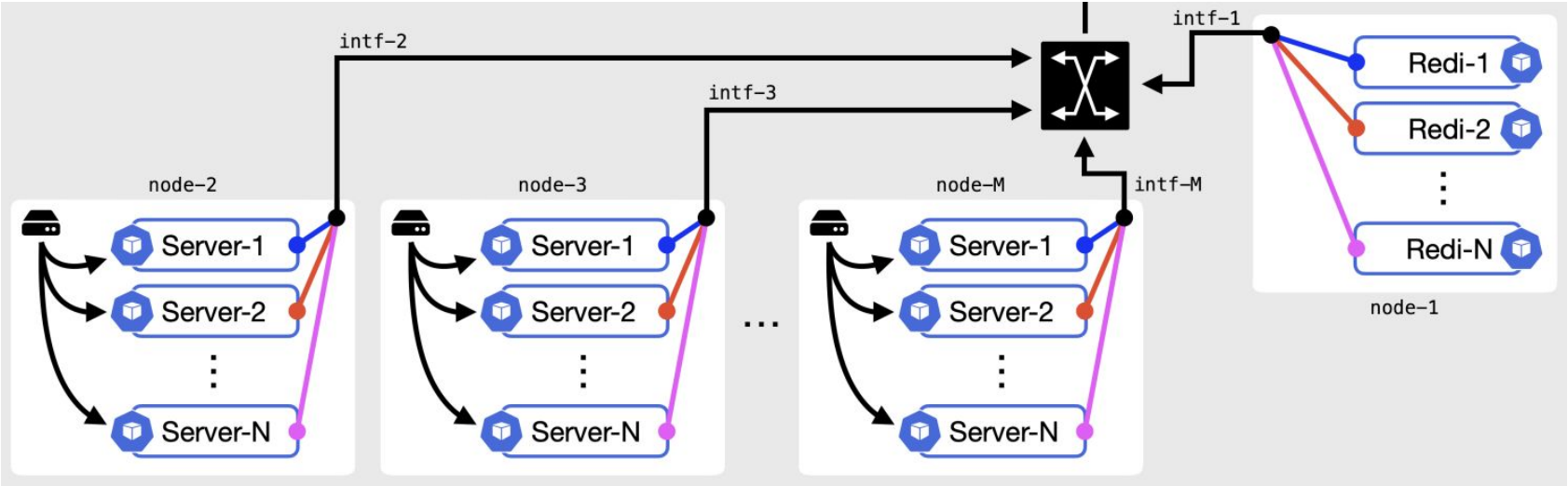
What we want

- A better way to manage our transfers, to **improve accountability**
- Being able to **Isolate large flows** would make them easier to understand
- Once isolated we can
  - Use **Quality of Service (QoS)** to provide a bandwidth guarantee
  - Use **VPNs** to select a fixed path

We can focus on the largest flows (not ALL transfers)

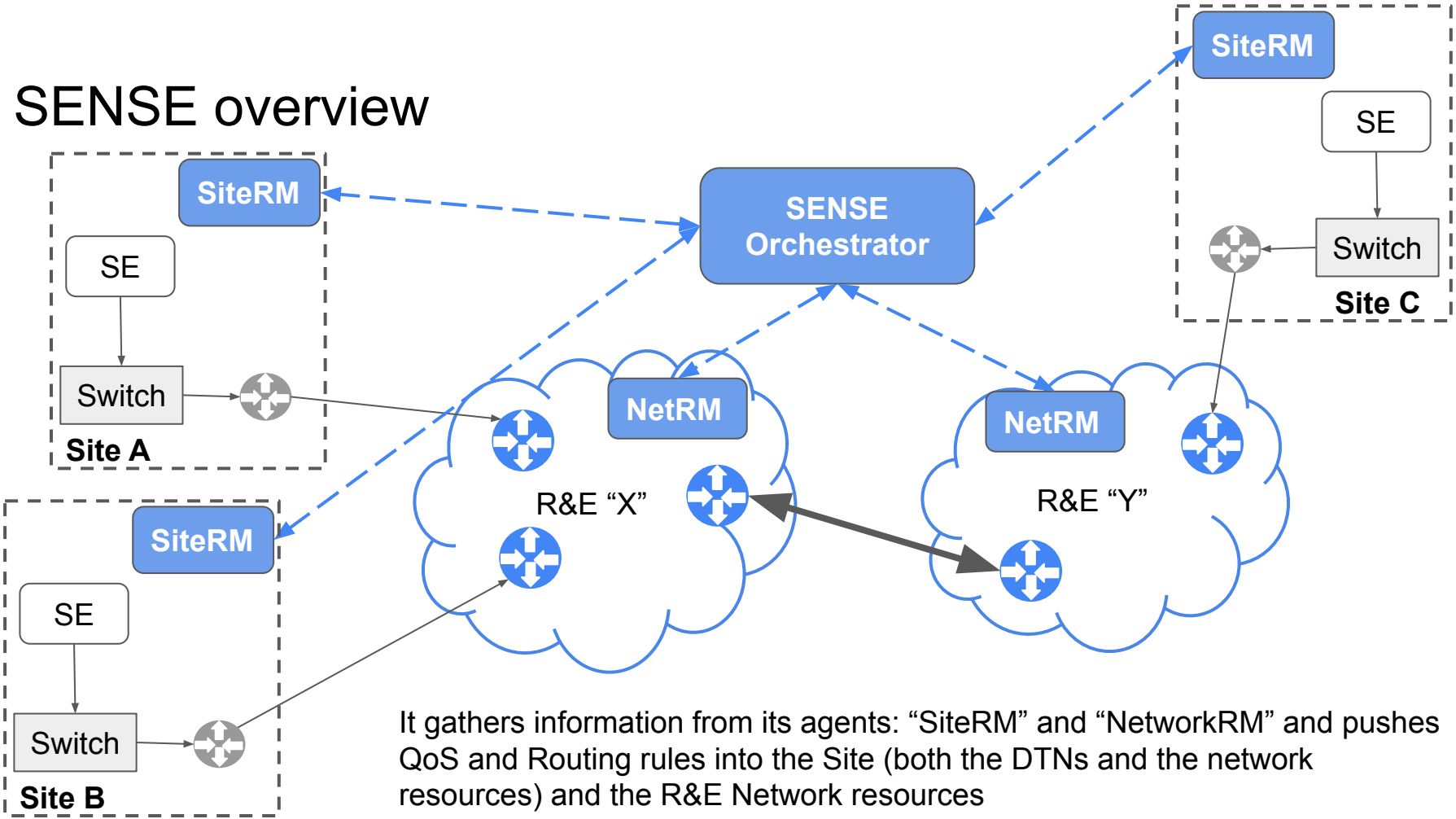
Isolation using XRootD multi-endpoint

- A single data server is configured to listen in N different IPv6 addresses.
- We use IPv6 because we need many IP addresses



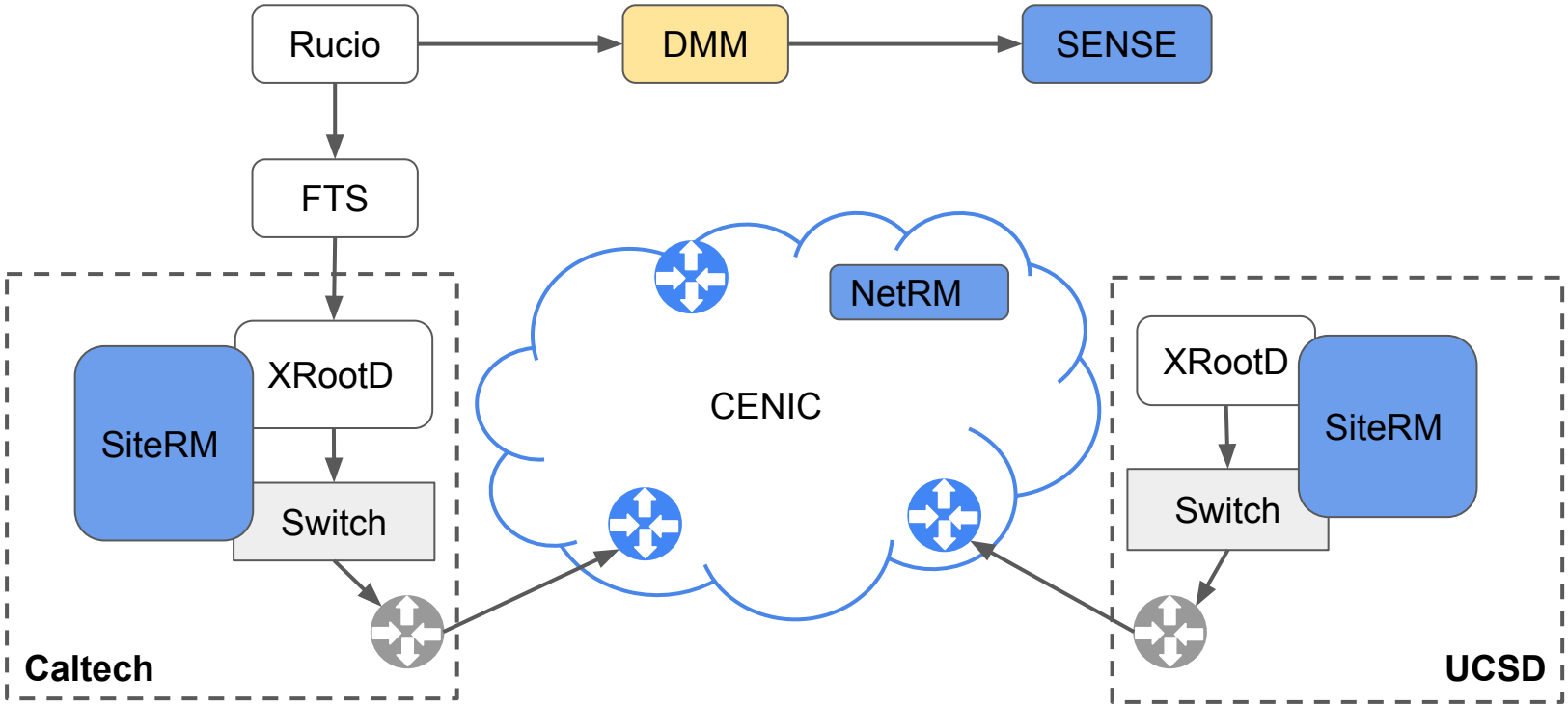
XRootD cluster with M servers and N subnets, Every color represents a different subnet

SENSE overview



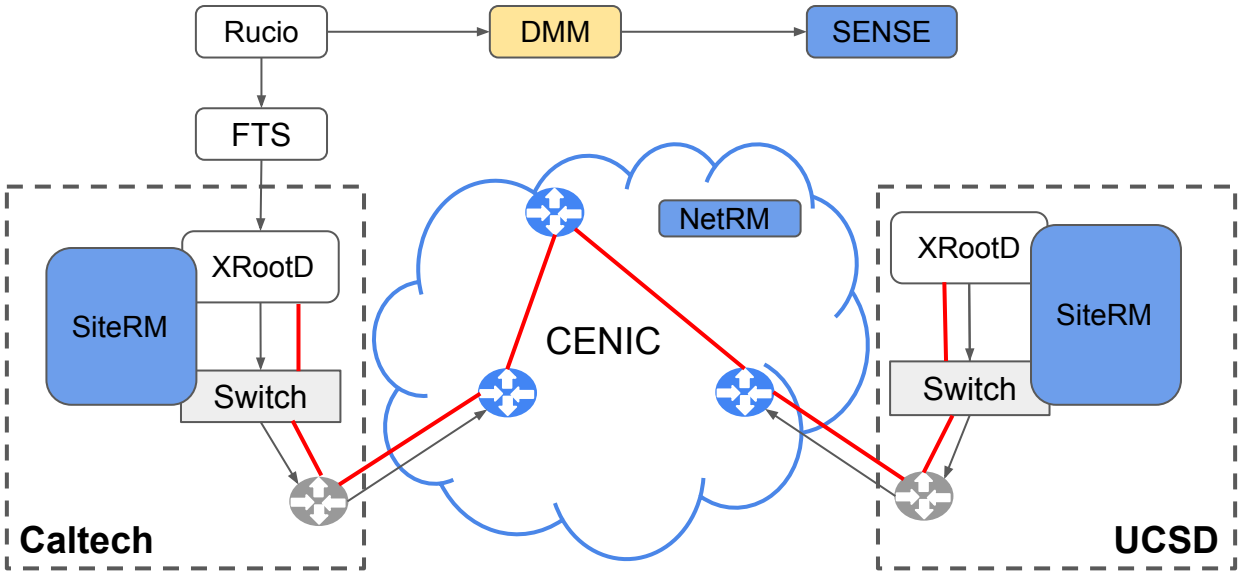
It gathers information from its agents: "SiteRM" and "NetworkRM" and pushes QoS and Routing rules into the Site (both the DTNs and the network resources) and the R&E Network resources

How does Rucio + SENSE looks like



DMM: Data Movement Manager (interface between Rucio and SENSE … and much more)

How it works? For a **non-priority** Rucio request

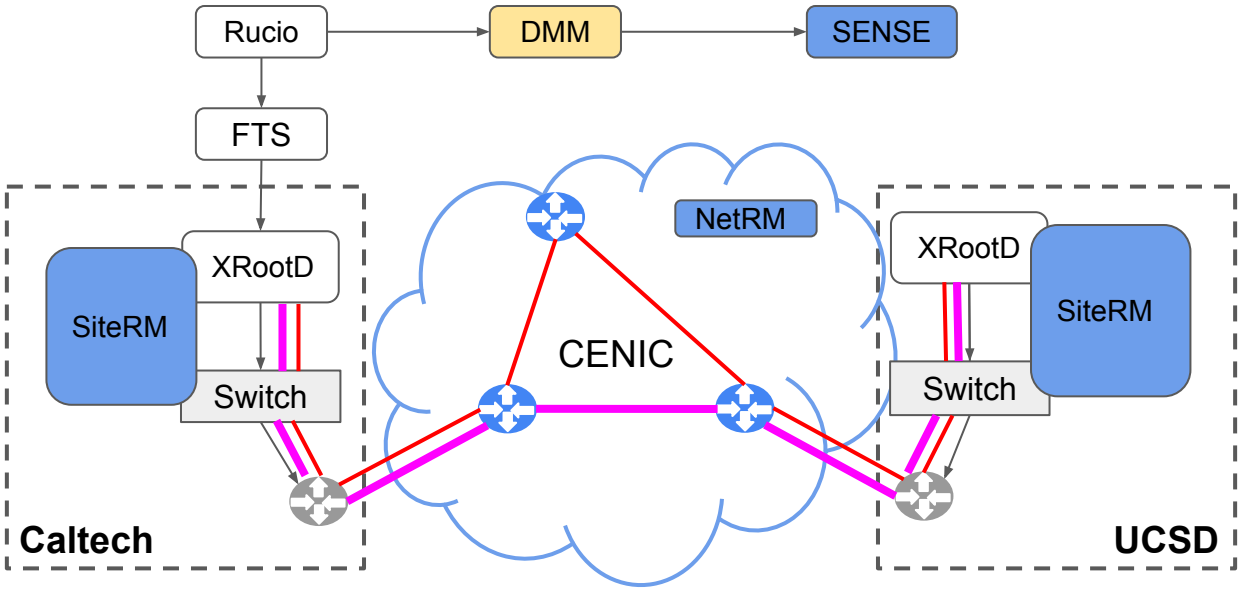


For every Rucio request, Rucio contacts DMM to ask for the endpoints (IP addresses) to use before contacting FTS

For a regular request (red) DMM will return the IPv6 addresses selected for “best effort”

SENSE is only contacted by DMM in order to get the set of IPv6 addresses of the 2 sites involved in the transfer. This information is cached

How it works? For a priority Rucio request



For a priority Rucio request (pink) DMM picks a pair of free IPv6s and requests a bandwidth allocation on them to SENSE

DMM return the selected pair of IPv6s to Rucio

SENSE instructs SiteRM to implement specific routing and QoS on the given IPv6s at the site level

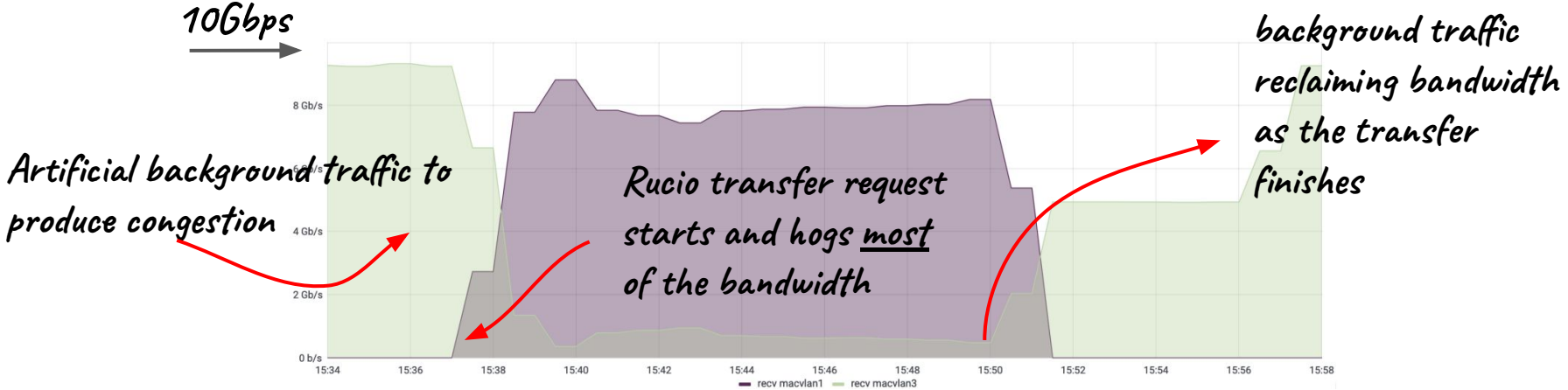
SENSE instructs NetworkRM to implement specific routing and apply QoS in CENIC nodes in between the 2 IPv6 endpoints

When the transfer is finished Rucio signals DMM which request the deallocation of the priority services

Our Proof of Concept

As a PofC we wanted to prove that we could create a priority service between 2 sites:

- On demand i.e. triggered solely by the creation of a rule in Rucio
- On a congested network path (to show QoS)
- Just for the duration of the transfer request in question



Network traffic on 2 different virtual interfaces in the receiving XRootD server

DMM

Designing **effective policies on how bandwidth should be shared** is one of the main tasks of DMM and also a key conceptual challenge for this project in the long term.

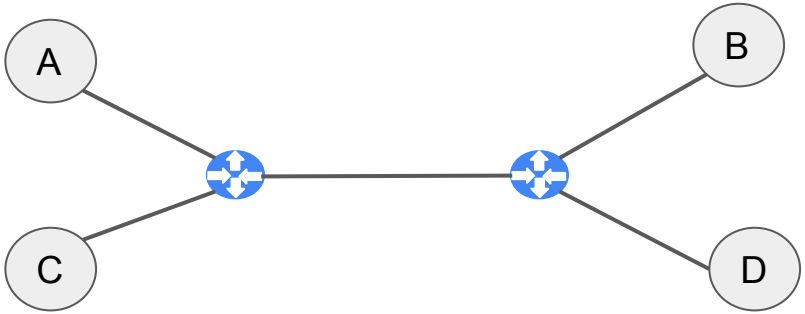
DMM needs to keep track/manage:

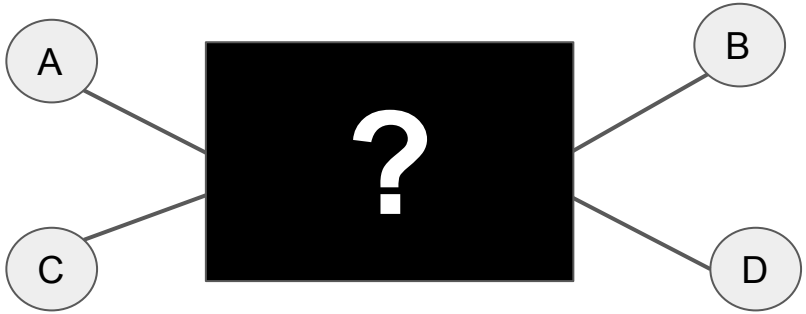
- Available bandwidth on each site and each link
- Number of IPv6 addresses available
- Recompute fair-share every time something changes (new/finish transfer)
- (future) Modify the established network services if conditions change e.g. network or site changes

DMM

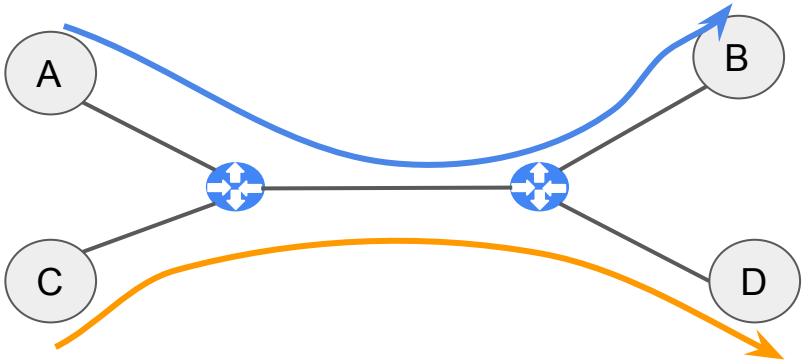
Implementing **effective fair sharing is not a trivial** task due to the possibility of having:

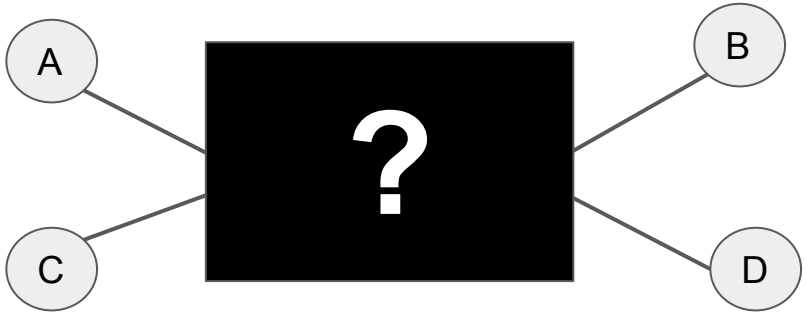
- several independent transfers using **overlapping segments of the network.**
- an incomplete picture of the network topology
- a combination of the above





DMM





Assuming 2 ongoing transfers:
A->B & C->D

Any new transfer from any of these sites (A,B,C,D) implicates recalculating fairshares

Early lower priority transfer A -> B Could hog the bandwidth of a Later higher priority C -> D

Simulation

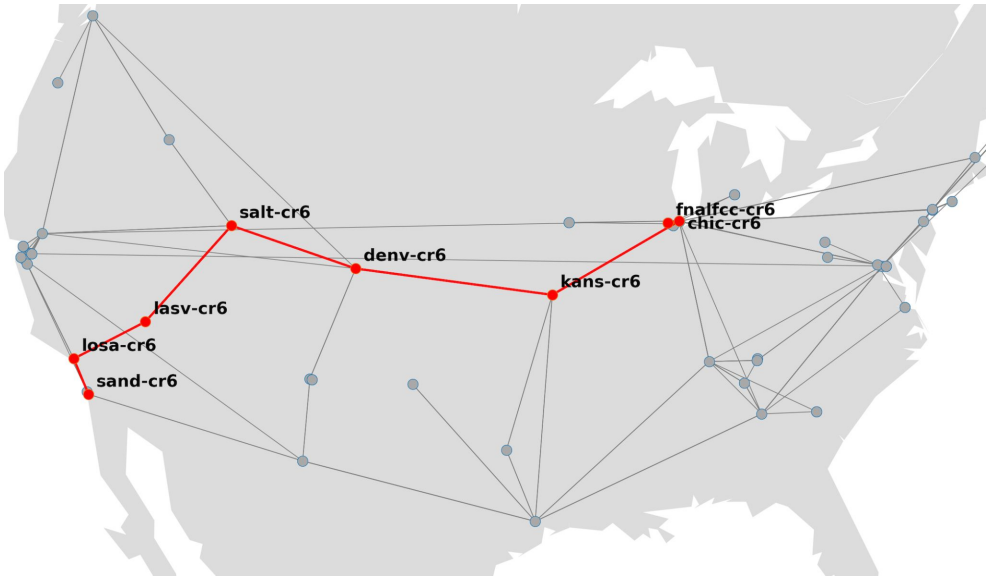
To **facilitate exploration** of this problem we have started developing a simulation of the entire system surrounding DMM including the network topology.

The main objectives of the simulation are:

1. **Validate our observations** of the behavior of the testbed
2. Playback annual sequences of actual transfers to **show SENSE benefits**
    a. We plan to use the monitoring records from Rucio and/or FTS for that
3. Collaborate with CS researchers to **develop policies for network bandwidth allocation**

Simulation example

- 90G from SiteA to SiteB, p=1
- Sleep 5s
- 50G from SiteC to SiteB, p=0
- 60G from SiteC to SiteB, p=5
- …



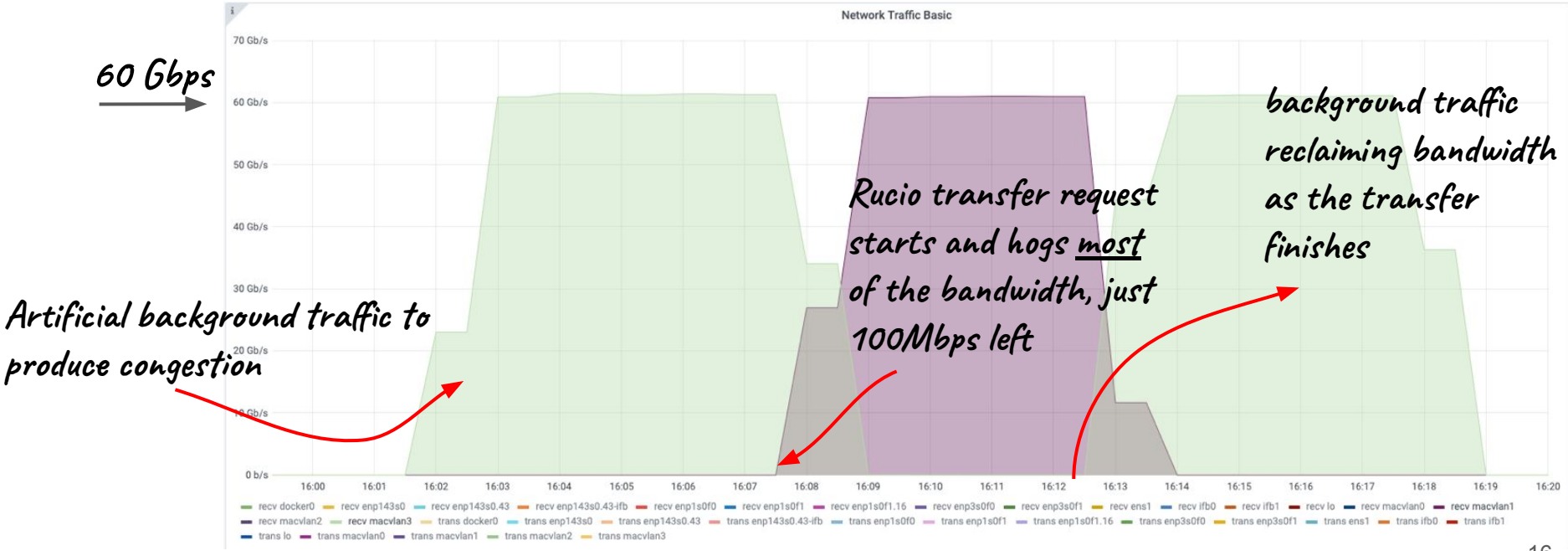
ESnet topology as seen by our simulation algorithm

Simulation status

- All sim-software components developed
- Pulled/cleaned monitoring records from Rucio for 2022
- Got/parsed ESnet topology
- Ran few simple validation tests
- Got a new student to do all the remaining work

Next: 400Gbps test

Currently a max of **60Gbps**, still far from the target…



ACKNOWLEDGMENTS

This ongoing work is partially supported by the US National Science Foundation (NSF) Grants OAC-2030508, OAC- 1841530, OAC-1836650, MPS-1148698, and PHY-1624356. In addition, the development of SENSE is supported by the US Department of Energy (DOE) Grants DE-SC0015527, DE- SC0015528, DE-SC0016585, and FP-00002494. Finally, this work would not be possible without the significant contributions of collaborators at ESNet, Caltech, and SDSC.

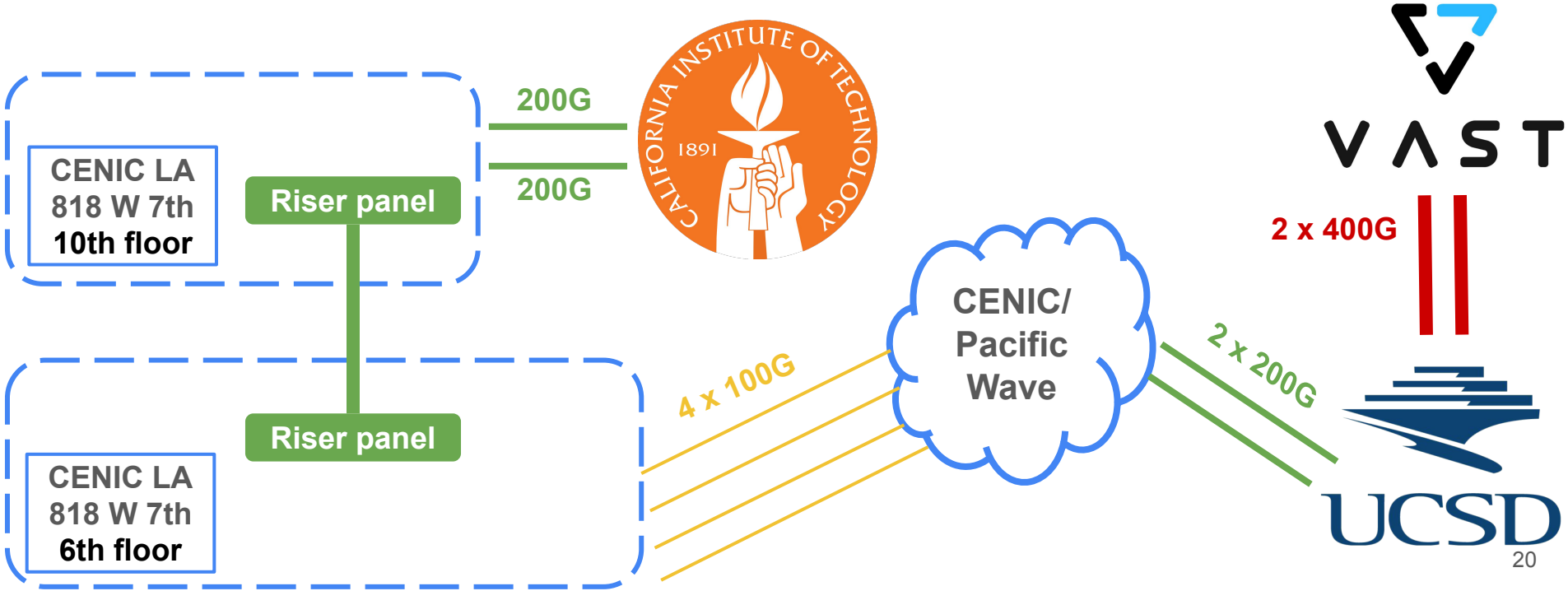
Questions?

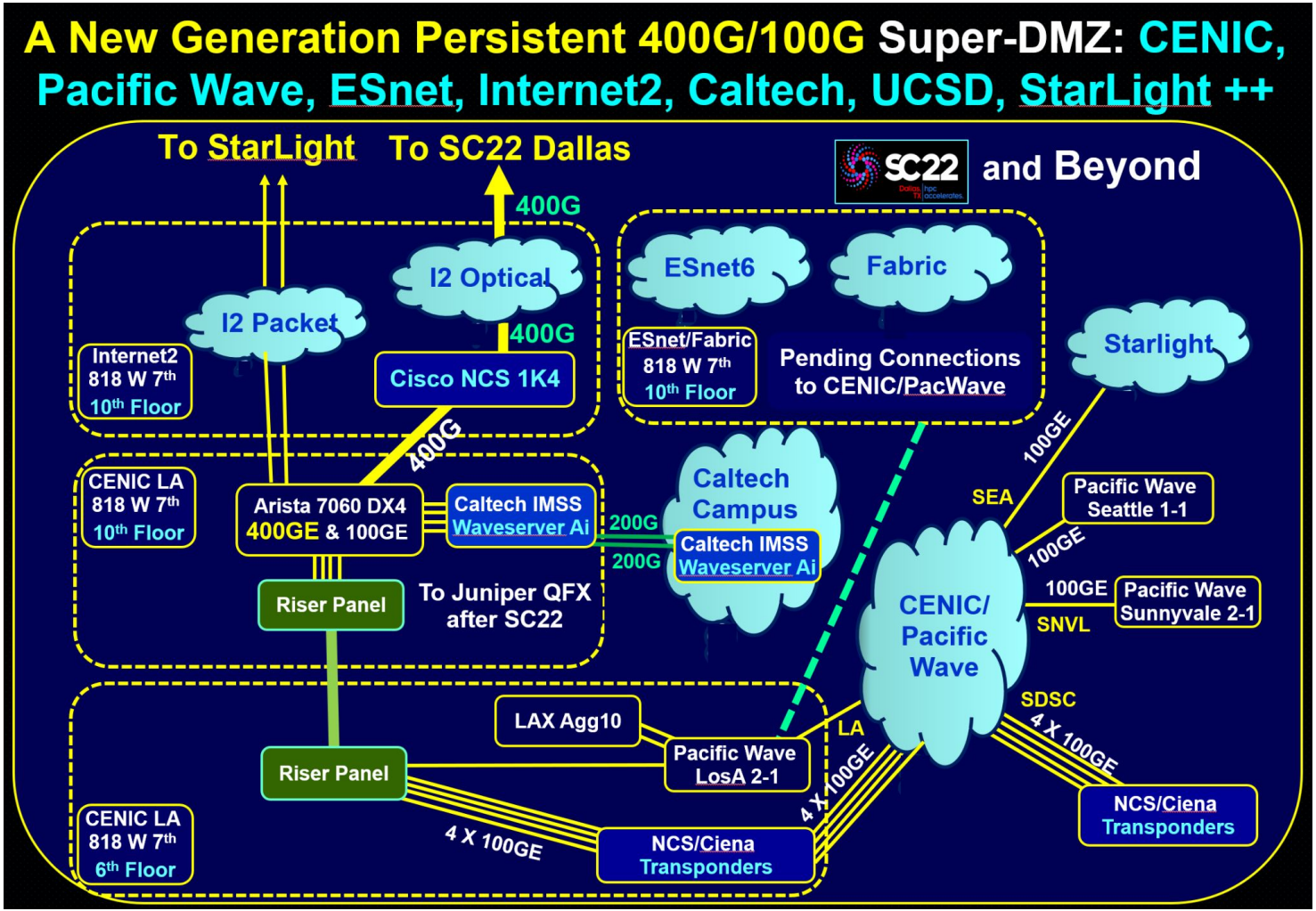
Want to join SENSE Testbed?
Or ask questions?
Drop an email to SENSE Group:
sense-info@es.net

Backup slides

Coming soon: new test at 400Gbps

The PofC was done at 10Gbps. In principle this should work at any scale … but it would be nice to show: *“How the future of transfer requests will look”*




A New Generation Persistent 400G/100G Super-DMZ: CENIC, Pacific Wave, ESnet, Internet2, Caltech, UCSD, StarLight ++
To StarLight
To SC22 Dallas
SC22
and Beyond
400G
I2 Optical
400G
I2 Packet
Internet2
818 W 7th
10th Floor
Cisco NCS 1K4
400G
ESnet6
Fabric
ESnet/Fabric
818 W 7th
10th Floor
Pending Connections
to CENIC/PacWave
Starlight
100GE
CENIC LA
818 W 7th
10th Floor
Arista 7060 DX4
400GE & 100GE
Caltech IMSS
Waveserver Ai
200G
200G
Caltech
Campus
Caltech IMSS
Waveserver Ai
SEA
Pacific Wave
Seattle 1-1
100GE
100GE
Pacific Wave
Sunnyvale 2-1
SNVL
Riser Panel
To Juniper QFX
after SC22
CENIC/
Pacific
Wave
SDSC
4 X 100GE
LAX Agg10
Pacific Wave
LosA 2-1
LA
4 X 100GE
Riser Panel
NCS/Ciena
Transponders
CENIC LA
818 W 7th
6th Floor
4 X 100GE
NCS/Ciena
Transponders