Rucio/SENSE implementation for CMS

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

Throughput Computing
July 12th, 2023

Motivation

- CMS currently performs Millions of transfers per day
- We have a significant failure rate: https://monit-grafana.cern.ch/goto/pG3Zt7C4z?orgId=20
- Hard to understand failures/degradation (see Shawn McKee presentation[*])
- Hard to predict ETAs
- We cannot take advantage of different network paths
- We cannot express our priorities to the Network
- All data accesses are treated equally (remote reads v.s. TPCs)

[*]https://agenda.hep.wisc.edu/event/2014/contributions/28489/attachments/9162/11048/Progress%20and%20Plans%20in%20OSG%20Networking.pdf

Objective

Allow Rucio to use SENSE to **fine-grain manage** its largest data flows

By “manage” we mean being able to create priority paths between sites and use them to transfer large data flows that are time sensitive i.e. need precise ETAs.

Within these “priority paths” data flows will:

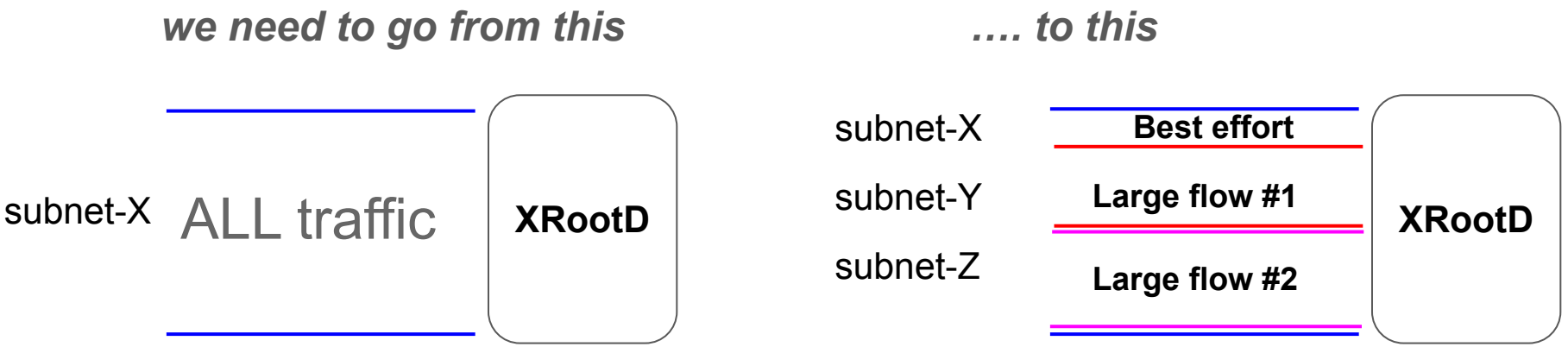
- Travel isolated from the rest of the data transfers
- Have a bandwidth guaranteed (QoS)
- Make use of fixed routes (VPNs)

Only for large time-sensitive data flows, the rest of the transfers will continue to travel on “best effort” mode

Implementation

First things first: SENSE services are configured on a per-subnet basis

So we need our SEs to expand over more than a single subnet



Implementation

The XRootD multi-subnet recipe

1. Assign N IPv6 addresses to 1 server, each IP is in a different subnet
2. Create N virtual interfaces each will have a different IP
3. Run N XRootD instances per server, each server is linked to a different virtual interface

Notes:

- This doesn’t need more hardware, just a bunch of configuration :)
- The setup above can be “easily” extrapolated to a XRootD cluster (see backup slides)
- K8s + multus-plugin makes the setup a lot easier non-k8s is also possible
- IPv6 is not required but highly recommended

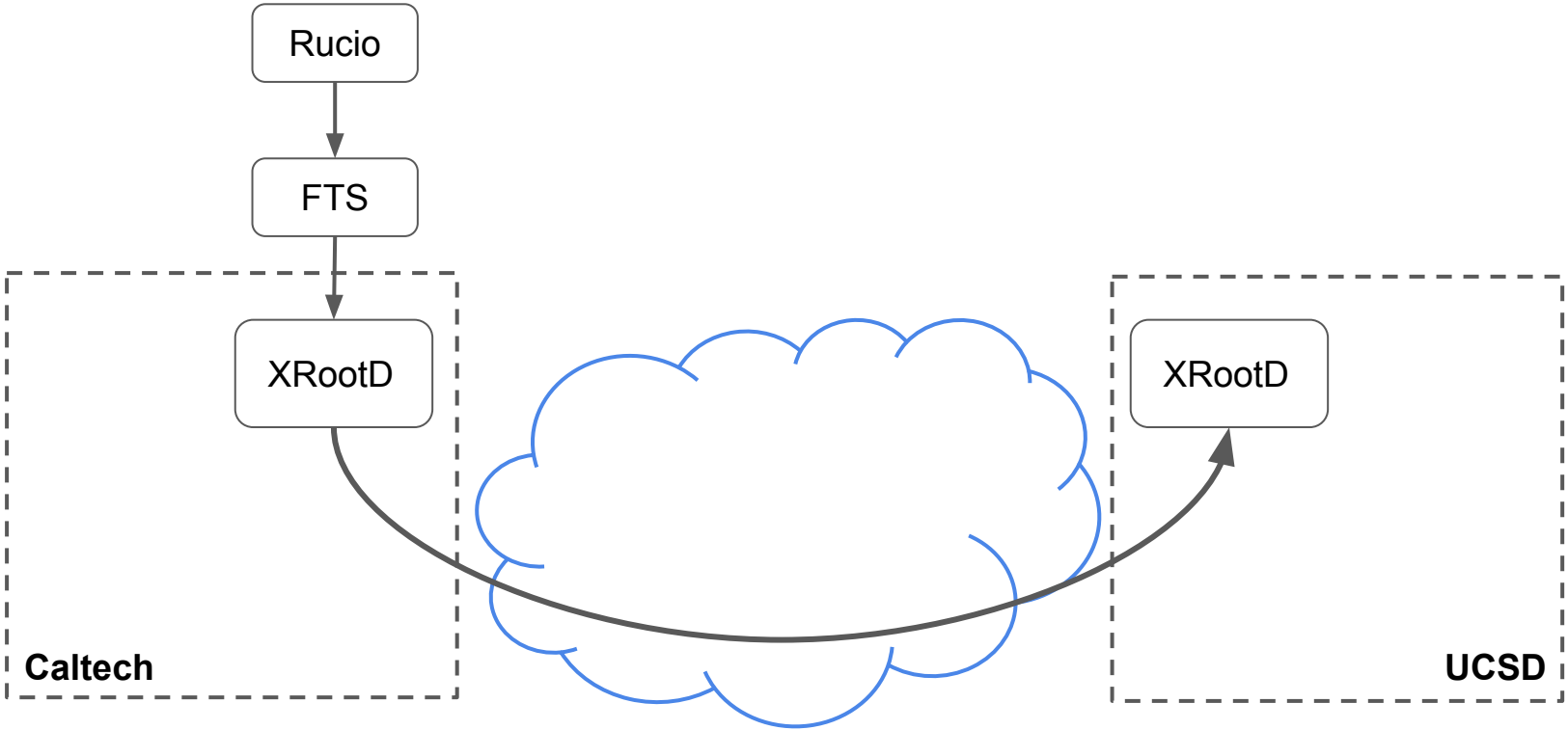
Implementation

So far we have:

- SEs that can talk over multiple different subnets
- SENSE that can create priority paths between these subnets

So where and how Rucio comes into play in all this?

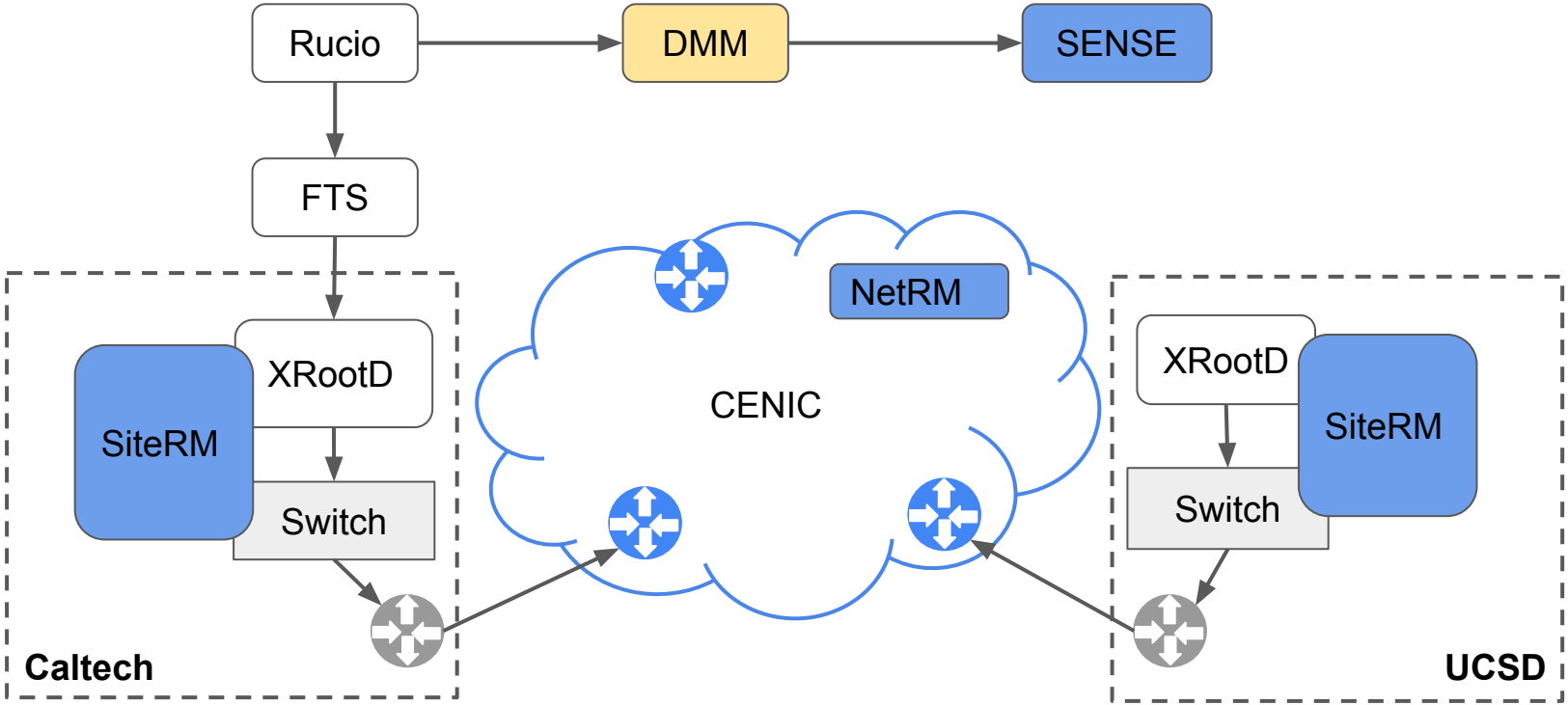
A “normal” TPC transfer looks like this nowadays



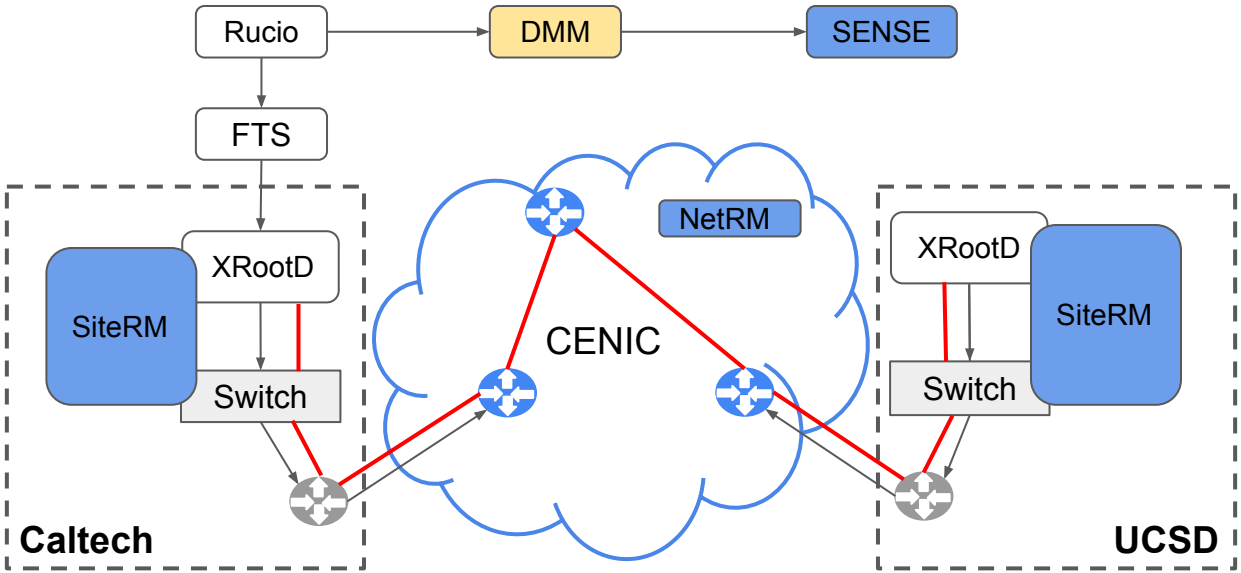
Data Movement Manager (DMM)

- Home-made SW product that sits between Rucio and SENSE
- Acts like a consultant for Rucio when it tries to talk to FTS
- Knows about the different subnets available on each site (via SENSE)
- Request network services to SENSE based on Rucio’s priorities
- Calculates bandwidth request based on relative priorities
- Monitors the usage of priority paths (in the near future)

This is how Rucio + DMM + SENSE looks like



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

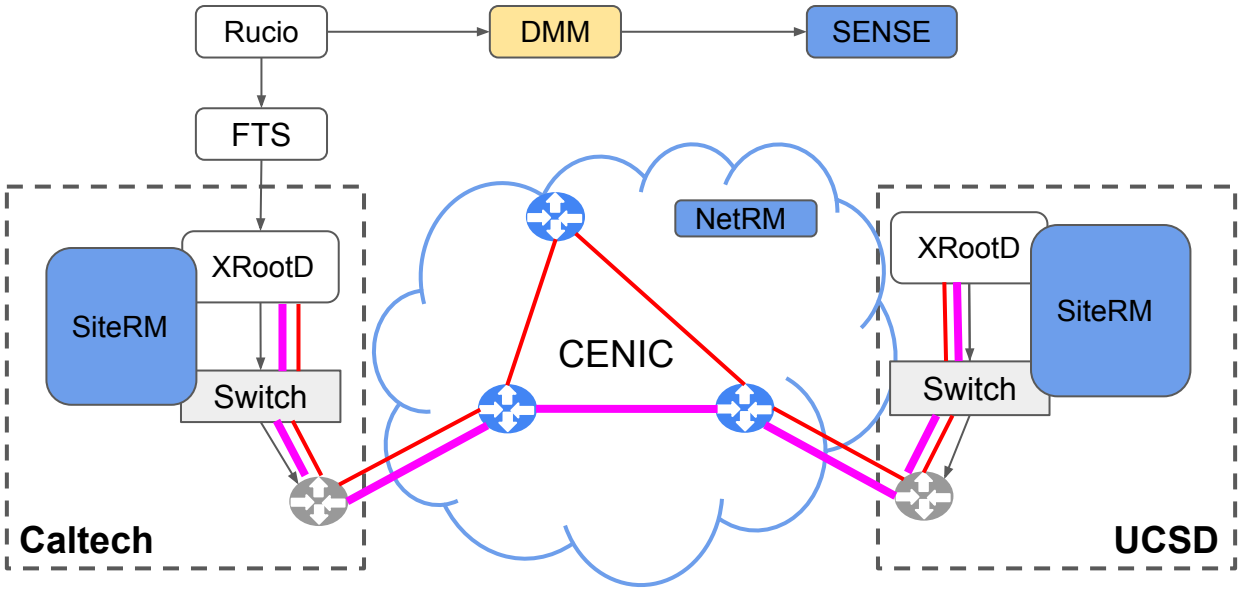


For every Rucio request, Rucio contacts DMM to ask for the endpoints (IP addresses) to use before contacting FTS

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

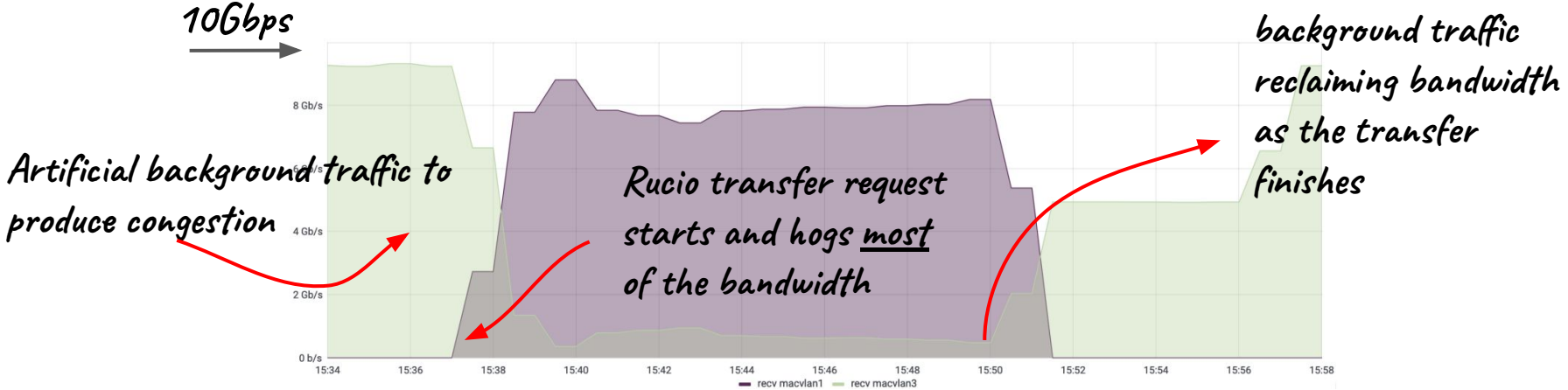
SENSE instructs NetworkRM to implement specific routing and apply QoS in CENIC nodes in between the 2 IPv6 endpoints

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

Status

- **Currently working on a repeat of our PofC at 400Gbps**
  - Already managed to do >300Gbps XRootD to XRootD: https://indico.jlab.org/event/459/contributions/11303/attachments/9681/14120/400gbps_benchmark.pdf
- **Implement monitoring**
  - Compare allocated vs achieved bandwidth using DTN network traffic + FTS records
- **Add more sites to our testbed**
  - Coming soon:
    - Fermilab: site-to-site testing is expected to begin by the end of July.
    - Nebraska: design work underway at the Tier2 Facility
  - Coming not that soon: Vanderbilt and Sprace
  - Started negotiations with CERN

What’s next

- **Implementation**
  - DMM policy implementation and simulation (on pause due to lack of effort)
  - Add support for more NOS (Network Operating Systems) in SiteRM
  - DTN-as-a-Service – Auto Start/Stop Transfer Service on Request
  - How can we include Sites without network control?

- **Demos**
  - Participate as a prototype in the WLCG Data Challenge 2024
  - Mini LHC Data Challenge using UCSD, CAltech, FNAL and Nebraska in Fall 2023
  - SC23 demonstrations in Fall of 2023.

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.

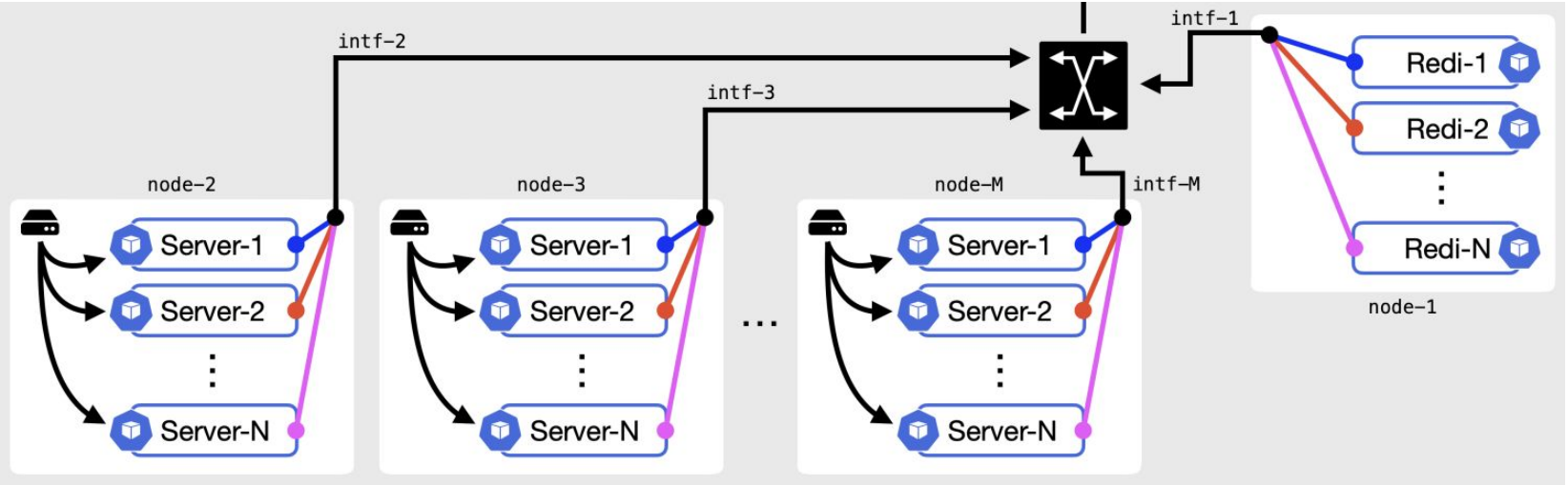
Questions?

didavila@ucsd.edu

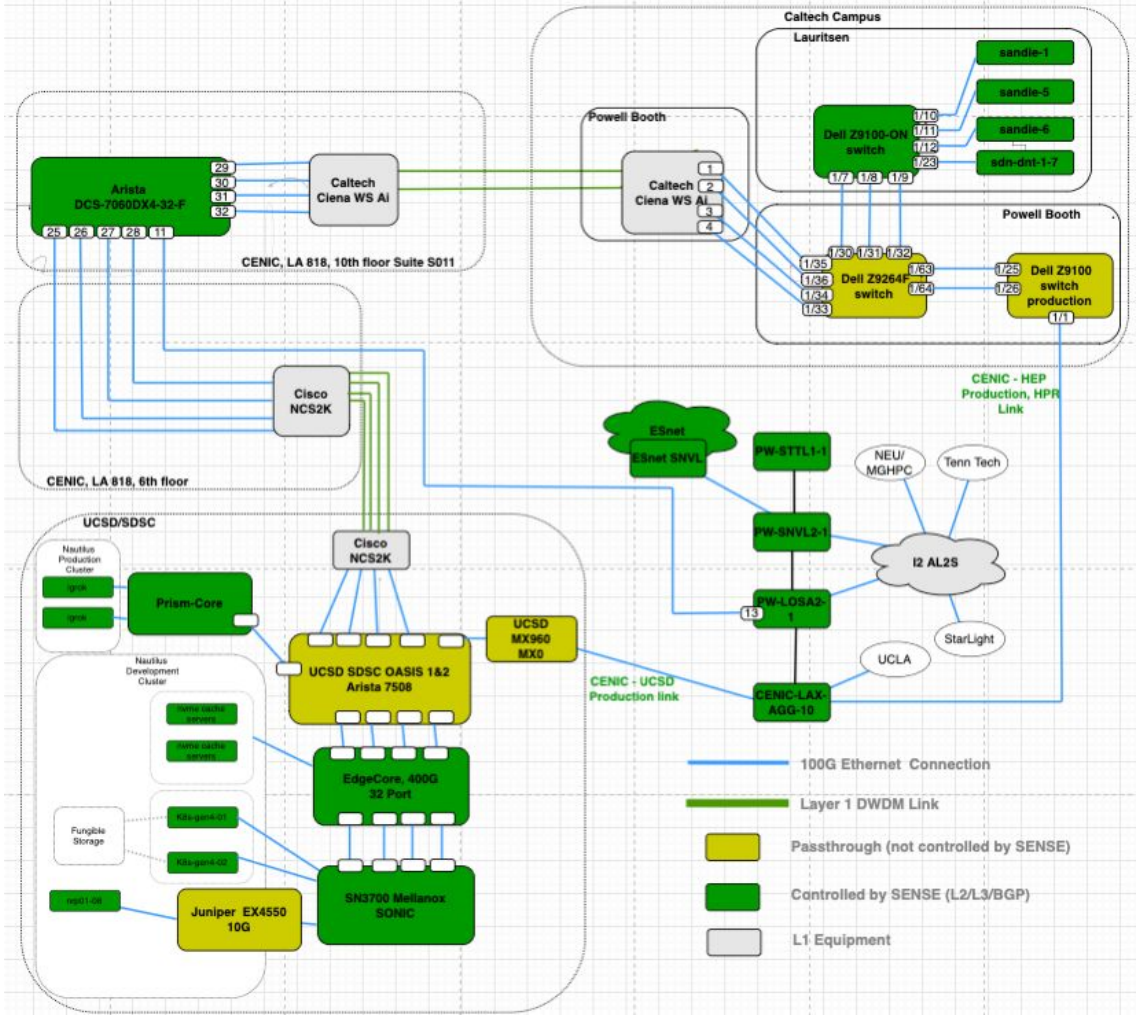
Backup Slides

XRootD cluster multi-subnet

- Priority services (QoS and VPN) are established on a subnet basis
- An XRootD cluster requires N different subnets to participate in N priority services.
- An XRootD cluster with M servers will require M x N IP addresses i.e. every server will have an IP in each subnet



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


Caltech Campus
Lauritsen
sandie-1
sandie-5
sandie-6
sdn-dnt-1-7
Dell Z9100-ON switch
1/10
1/11
1/12
1/23
1/7
1/8
1/9
Powell Booth
Caltech Ciena WS Ai
1
2
3
4
Arista DCS-7060DX4-32-F
29
30
31
32
25
26
27
28
11
Caltech Ciena WS Ai
CENIC, LA 818, 10th floor Suite S011
Powell Booth
1/30
1/31
1/32
1/35
1/36
1/34
1/33
Dell Z9264F switch
1/63
1/64
1/25
1/26
Dell Z9100 switch production
1/1
CENIC - HEP Production, HPR Link
Cisco NCS2K
CENIC, LA 818, 6th floor
ESnet
ESnet SNVL
PW-STTL1-1
NEU/ MGHPC
Tenn Tech
PW-SNVL2-1
I2 AL2S
PW-LOSA2-1
13
StarLight
UCLA
CENIC-LAX-AGG-10
UCSD/SDSC
Nautilus Production Cluster
Prism-Core
Cisco NCS2K
UCSD MX960 MX0
UCSD SDSC OASIS 1&2 Arista 7508
CENIC - UCSD Production link
Nautilus Development Cluster
EdgeCore, 400G 32 Port
Fungible Storage
SN3700 Mellanox SONIC
Juniper EX4550 10G
100G Ethernet Connection
Layer 1 DWDM Link
Passthrough (not controlled by SENSE)
Controlled by SENSE (L2/L3/BGP)
L1 Equipment