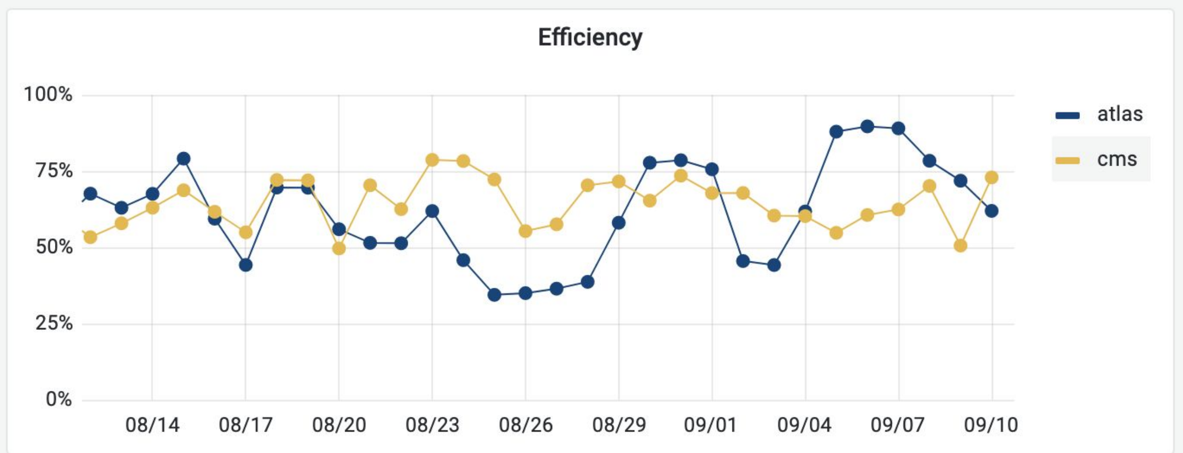
Rucio/SENSE overview and our plans for DC24

IRIS-HEP Retreat - September, 2023

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

Objective: Improve bulk transfers (aka TPCs)

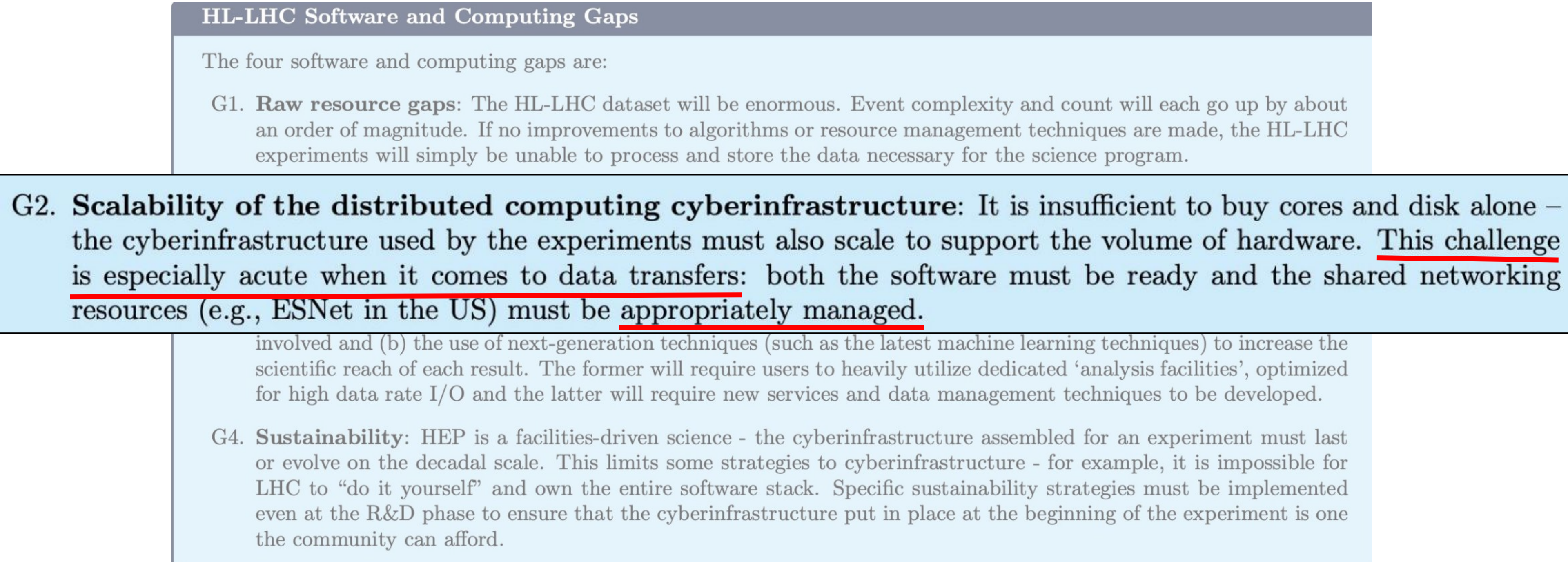
- IRIS-HEP: fill the gap between the resources we have and the ones we need
- On TPC land there is enough room for improvement



CMS and ATLAS TPC efficiency in the last 30 days[1]

- Hard to understand failures/degradation (HTC23 Shawn McKee presentation[1])

Stolen form Peter’s talk on Monday[3]



HL-LHC Software and Computing Gaps

The four software and computing gaps are:

G1. **Raw resource gaps**: The HL-LHC dataset will be enormous. Event complexity and count will each go up by about an order of magnitude. If no improvements to algorithms or resource management techniques are made, the HL-LHC experiments will simply be unable to process and store the data necessary for the science program.

G2. **Scalability of the distributed computing cyberinfrastructure**: It is insufficient to buy cores and disk alone – the cyberinfrastructure used by the experiments must also scale to support the volume of hardware. This challenge is especially acute when it comes to data transfers: both the software must be ready and the shared networking resources (e.g., ESNet in the US) must be appropriately managed.

involved and (b) the use of next-generation techniques (such as the latest machine learning techniques) to increase the scientific reach of each result. The former will require users to heavily utilize dedicated ‘analysis facilities’, optimized for high data rate I/O and the latter will require new services and data management techniques to be developed.

G4. **Sustainability**: HEP is a facilities-driven science - the cyberinfrastructure assembled for an experiment must last or evolve on the decadal scale. This limits some strategies to cyberinfrastructure - for example, it is impossible for LHC to “do it yourself” and own the entire software stack. Specific sustainability strategies must be implemented even at the R&D phase to ensure that the cyberinfrastructure put in place at the beginning of the experiment is one the community can afford.

“appropriately managed”

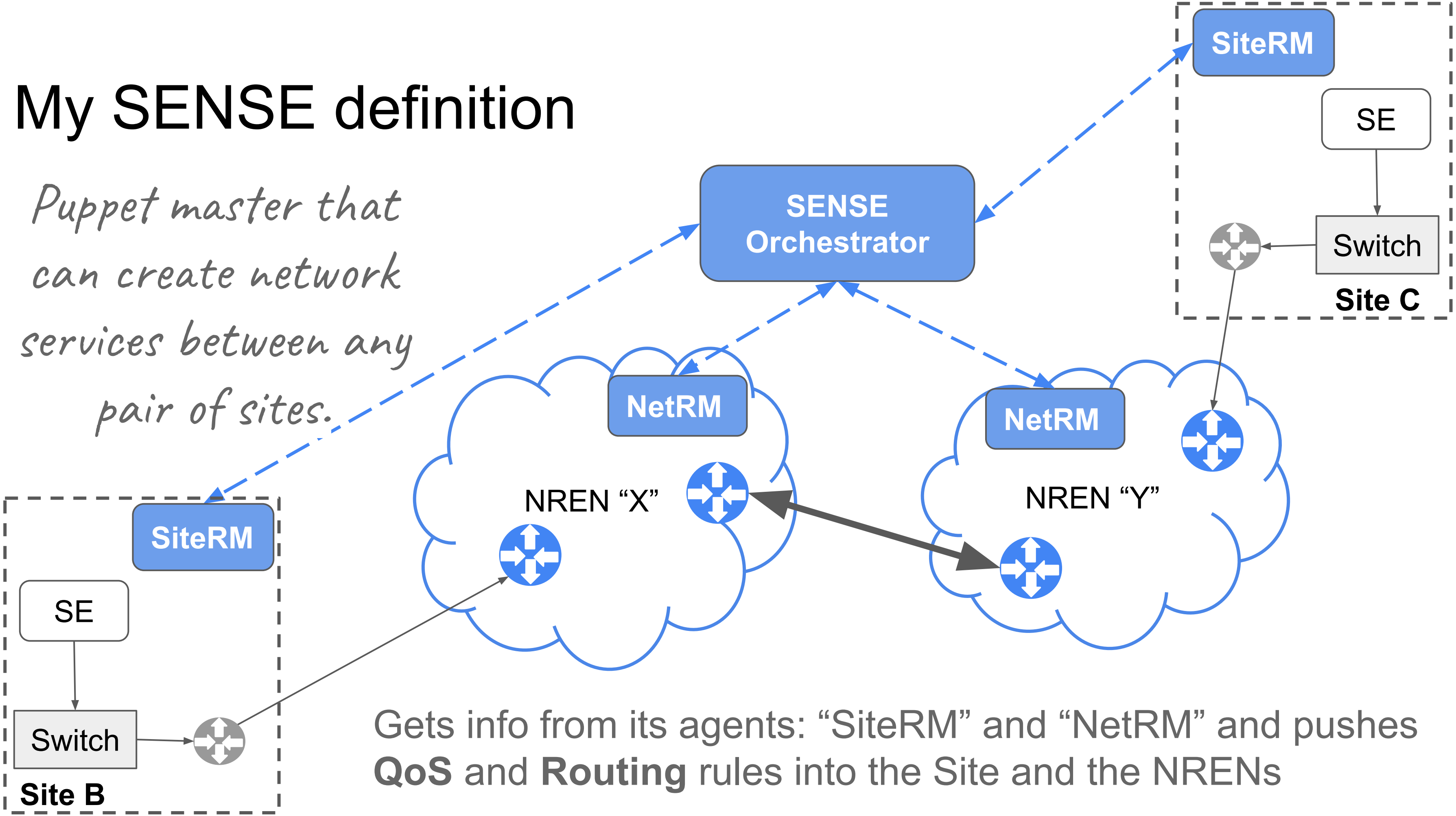
How can we improve TPCs?

By mixing Rucio and SENSE

SENSE: Software Defined Networking for End-to-End Networked Science at the Exascale[4]

My SENSE definition

Puppet master that can create network services between any pair of sites.



Gets info from its agents: "SiteRM" and "NetRM" and pushes **QoS** and **Routing** rules into the Site and the NRENs

Multi subnet Storage System

- Sense services are created based on subnets
- Current Storage Systems live in a single subnet
- We need our Storage Systems to be exposed via multiple subnets for this to work
- We managed to do this in XRootD by adding a bunch of configuration
  - No extra hardware is needed

More details here:
https://indico.cern.ch/event/1185600/contributions/5109192/attachments/2545788/4383989/Automated%20Network%20Services%20for%20Exascale%20Data%20Movement%20(1).pdf

Rucio & SENSE

Rucio: Data Management System used by CMS and ATLAS, it knows the data workflows, how big, where they have to go, how important are.

By joining forces we can let Rucio leverage SENSE capabilities to:

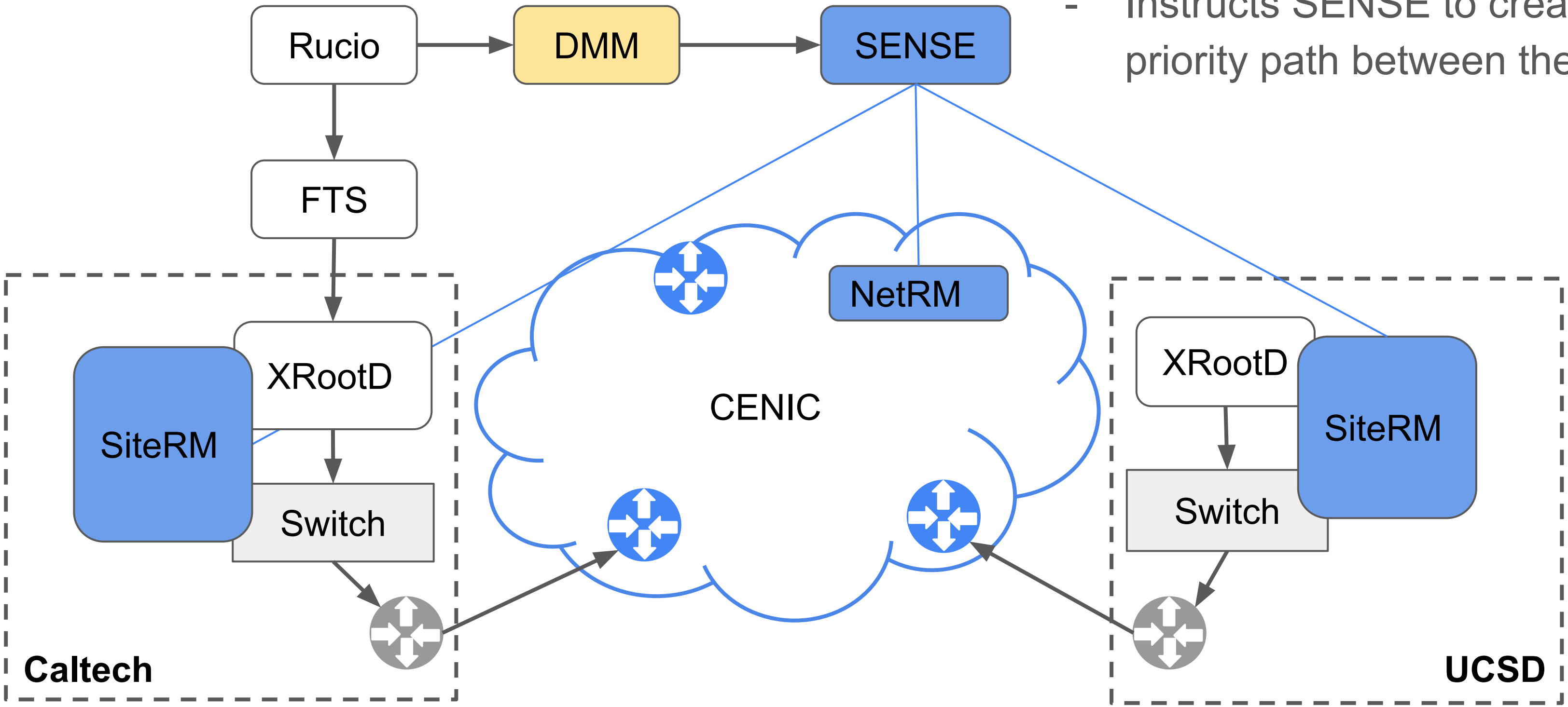
- Isolate => different data workflows travel on different subnets
- QoS => Allocate bandwidth
- Routing => Select the best path

ONLY for the LARGEST and more TIME-SENSITIVE data workflows

How it looks

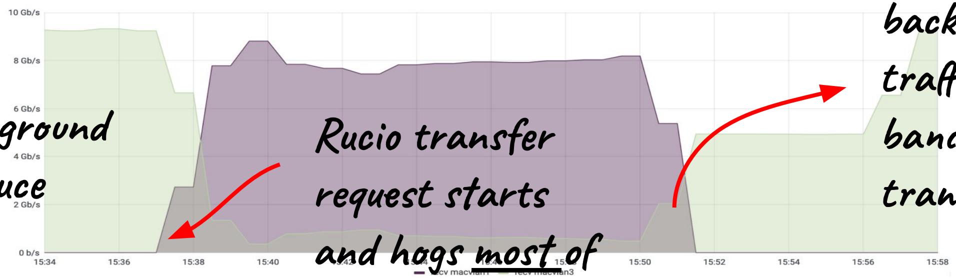
DMM:

- Calculates bandwidth allocation
- Picks a free subnet at each site
- Instructs SENSE to create a priority path between them



Status

We did a PofC @10Gbps last year[2]



Artificial background traffic to produce congestion

Rucio transfer request starts and hogs most of the bandwidth

background traffic reclaiming bandwidth as the transfer finishes

Since then we have focus on:

a) redo at a higher scale (400Gbps)

b) add more sites to the testbed

c) improve stability

c) add support for more Network OS

Plans for DC24

We have proposed **2 mini-challenges**:

1. Sep. High bandwidth demonstration of **multiple** Rucio initiated priority data transfers between UCSD and Caltech
2. Nov. Demonstration of 3 priority paths between 3 different pairs of sites:
    a. FNAL => UCSD
    b. Caltech => UNL
    c. UCSD => Caltech

For the **actual challenge** we can provide additional artificial traffic amongst the sites in our testbed

Questions?

ACKNOWLEDGMENTS

This ongoing work is partially supported by the US National Science Foundation (NSF) Grants OAC-1841530, OAC-1836650, PHY-2323298 and PHY-1624356. In addition, the development of SENSE is supported by the US Department of Energy (DOE) Grants DE-SC0015527, DESC0015528, DE-SC0016585, and FP-00002494. Finally, this work would not be possible without the significant contributions of collaborators at ESnet, Caltech, and SDSC.

References

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

[2]https://indico.cern.ch/event/1185600/contributions/5109192/attachments/2545788/4383989/Automated%20Network%20Services%20for%20Exascale%20Data%20Movement%20(1).pdf

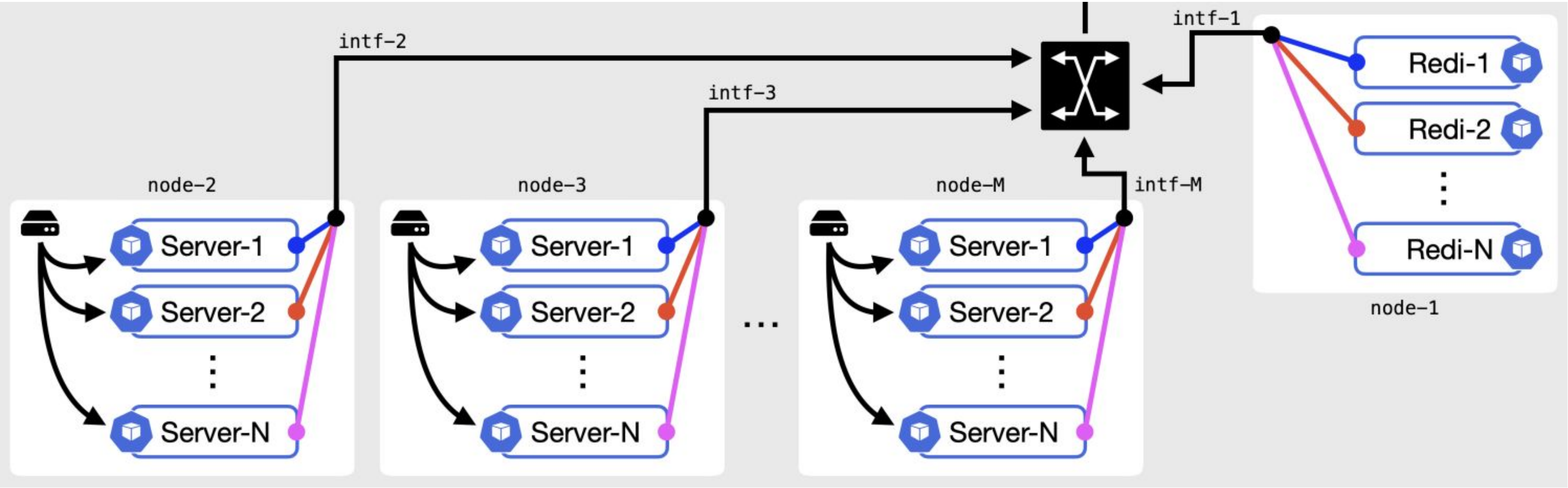
[3]https://indico.cern.ch/event/1288444/contributions/5413959/attachments/2712256/4709900/20230911-IRIS-HEP-Institute-Retreat-Introduction.pdf

[4]https://www.es.net/network-r-and-d/sense/

Backup slides

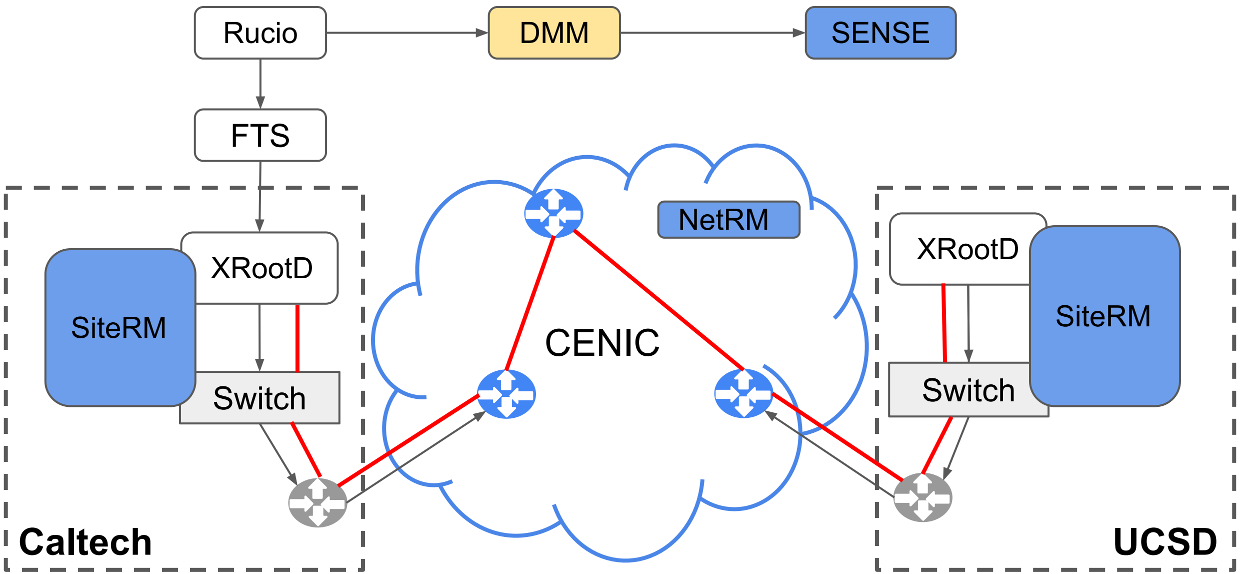
Isolation using XRootD multi-endpoint

- 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

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

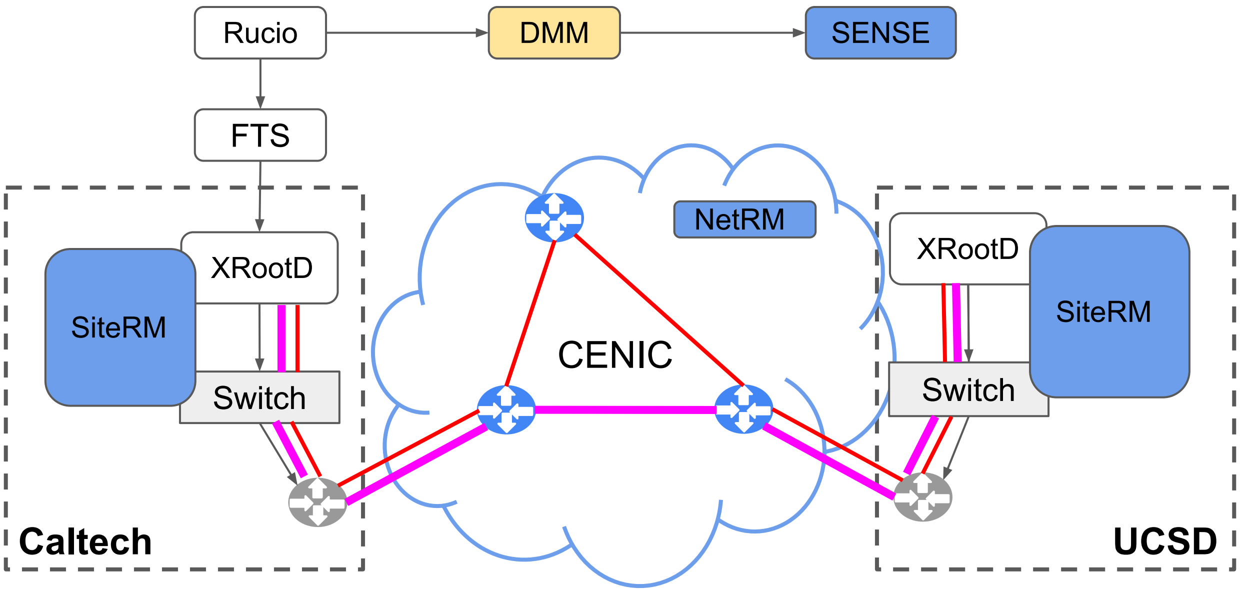


For every Rucio request, Rucio contacts DMM to ask for the endpoints (IP addresses) to use before contacting FTS

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