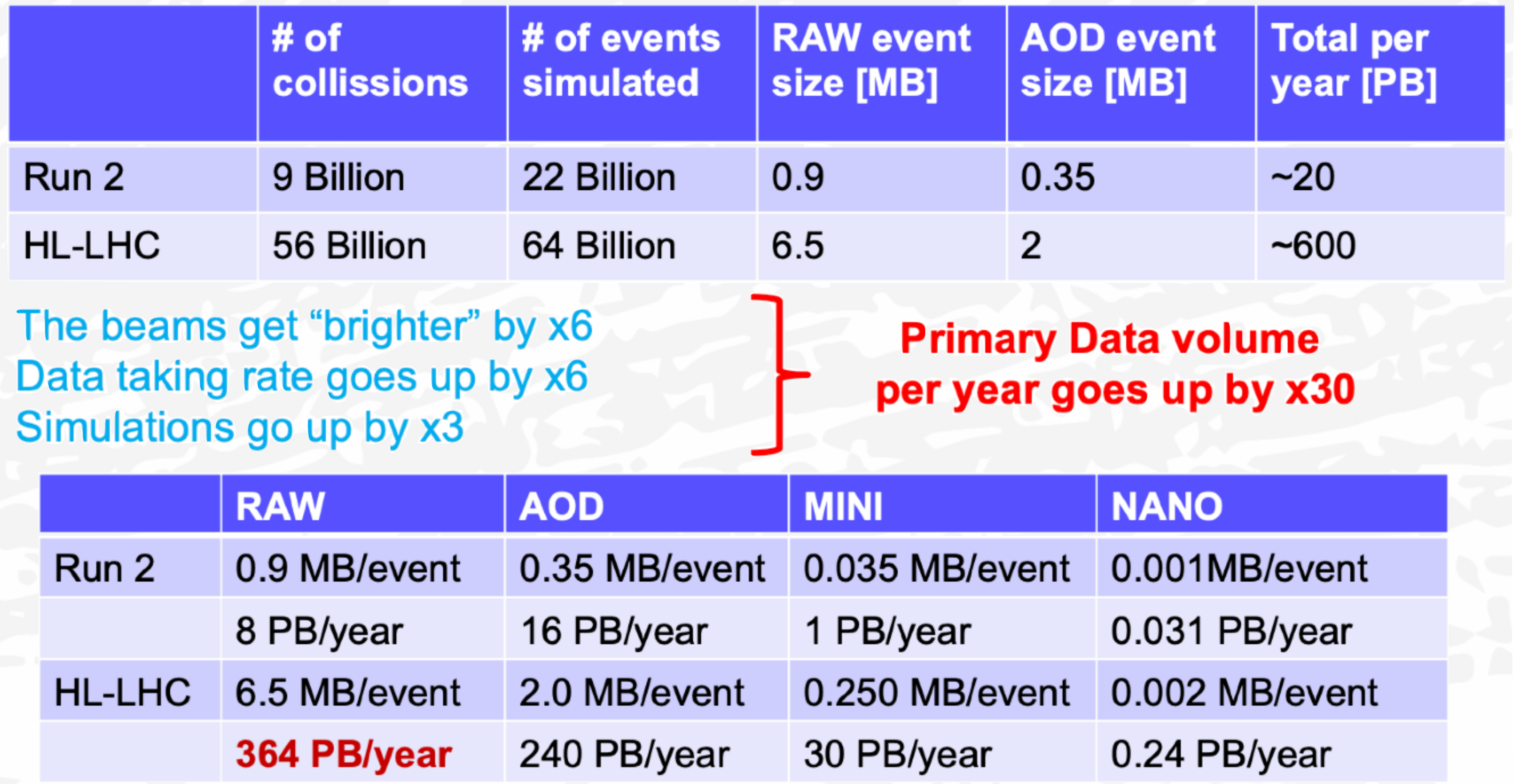
Towards Accountable Network Bandwidth Utilization via SDN

Frank Würthwein[1], Jonathan Guiang[1], **Aashay Arora**[1], Diego Davila[1], John Graham[1], Dima Mishin[1], Thomas Hutton[1], Igor Sfiligoi[1], Harvey Newman[2], Justas Balcas[2], Preeti Bhat[2], Tom Lehman[3], Xi Yang[3], Chin Guok[3], Oliver Gutsche[4], Asif Shah[4], Chih-Hao Huang[4], Dmitry Litvinsev[4], Phil Demar[4], Marcos Schwarz[4]

1. University of California San Diego / San Diego Supercomputer Center
2. California Institute of Technology
3. ESNet, Lawrence Berkeley National Laboratory
4. Fermilab

Overview

- We are approaching the exa-scale computing era for most large collaborative experiments, for e.g. (HL-)LHC



	# of collissions	# of events simulated	RAW event size [MB]	AOD event size [MB]	Total per year [PB]
Run 2	9 Billion	22 Billion	0.9	0.35	~20
HL-LHC	56 Billion	64 Billion	6.5	2	~600

The beams get "brighter" by x6
Data taking rate goes up by x6
Simulations go up by x3

Primary Data volume per year goes up by x30

	RAW	AOD	MINI	NANO
Run 2	0.9 MB/event	0.35 MB/event	0.035 MB/event	0.001MB/event
	8 PB/year	16 PB/year	1 PB/year	0.031 PB/year
HL-LHC	6.5 MB/event	2.0 MB/event	0.250 MB/event	0.002 MB/event
	364 PB/year	240 PB/year	30 PB/year	0.24 PB/year

  - Current model of data transfers, namely summarized as push-now-worry-later will not be feasible in the near future, controlled data-flows with high accountability will become a necessity.

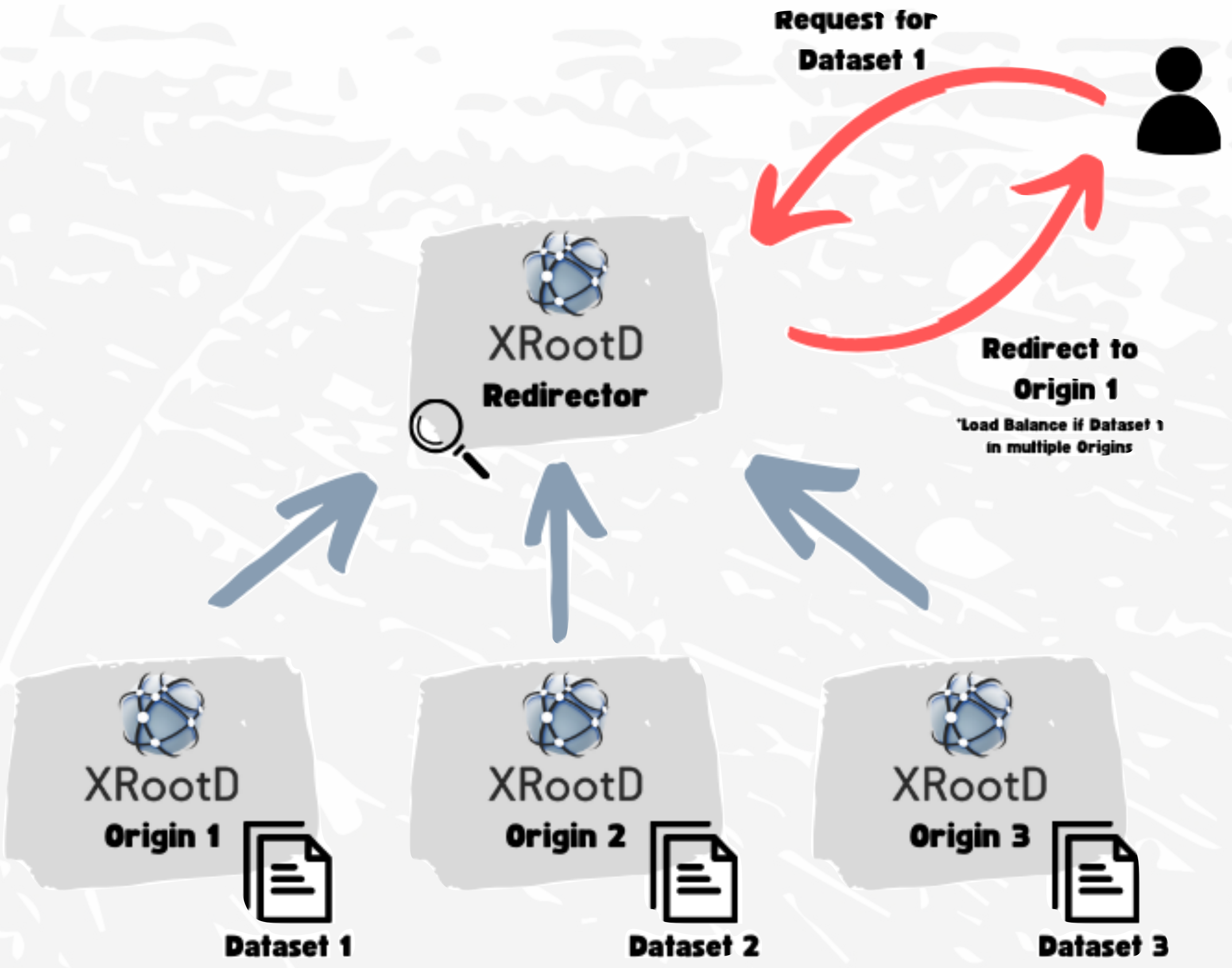
- Software defined networking (SDN) controlled data-flows allow for end-to-end accountability of network utilization, and allows the different stakeholders (e.g. large experiments) to manage their priorities.

- Using HEP (in particular CMS) software stack as the control testing ground, we are integrating the existing tools with SDN.

Pieces of the Testbed

XRootD

- XRootD is an open-source, high-performance data transport protocol and software suite designed for large-scale, distributed data-intensive applications.
- Facilitates seamless distributed data access through its hierarchical deployment system.
- Supports HTTP third-party-copy data-flows
- Can be deployed at scale using Kubernetes.



Storage Element

Rucio

- Rucio is a data management system designed for scientific and HPC environments.
- Manages large-scale, distributed data in scientific and HPC environments.
- Optimizes data placement based on policies, considering factors like popularity and access patterns.
- Catalogs detailed metadata for tracking data lineage, access control, and provenance.
- Enables different user groups or projects to share infrastructure with controlled data access.
- Works with FTS, the transfer orchestrator that submits HTTP copy requests.

Rucio

Data management service

knows the data workflows, the size, where its going and how important it is

SENSE

- **S**oftware-Defined Network for **E**nd-to-end **N**etworked **S**cience at the **E**xascale
- Provides the mechanisms to enable multi-domain orchestration for a wide variety of network and other cyberinfrastructure resources in a highly customized manner.
- These orchestrated services can be customized for individual domain science workflow systems and requirements.
  - Services include Layer 2 Point to Point Network Connections, Layer 2 Multipoint Network Topologies, and Layer 3 Routed/Virtual Private Network (VPN) services.
- Agents: SiteRM and NetRM push QOS and routing rules into the Site and NRENs.

SENSE

SDN orchestration layer

puppet master that can create network services between sites (SEs)

Data Movement Manager

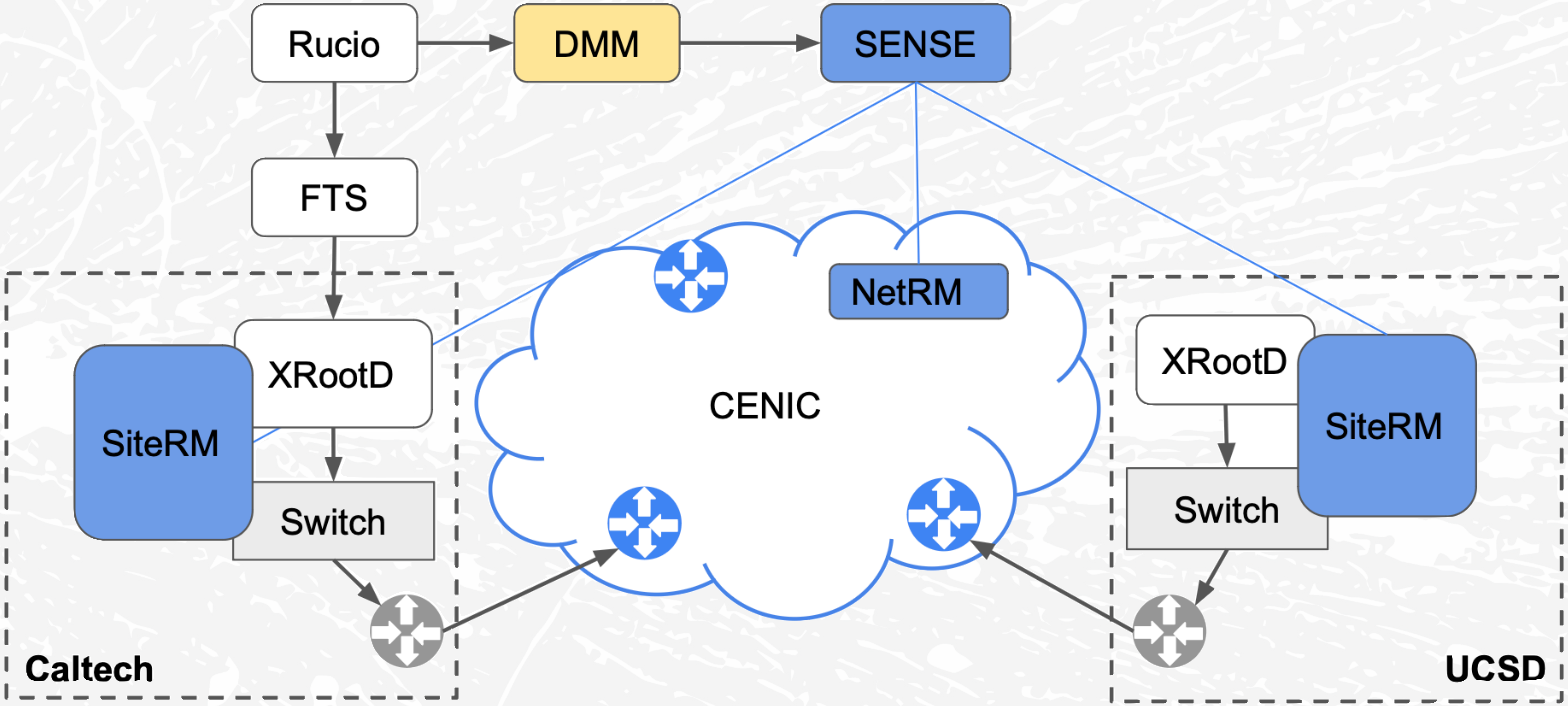
- Interface between Rucio and SENSE, making SDN operated HEP data-flows possible
  - Gets transfer metadata like source, destination, number of bytes and priority from Rucio
  - Gets bandwidth between endpoints from SENSE
  - Based on metadata, makes decision on bandwidth allocation for multiple requests.
  - Keeps state of all the data-flows, monitors performance and creates reports of underperforming flows.
- In long term, we predict that this will be a component of Rucio itself but for now, keeping it separate is the prototypical architecture.

Data Movement Manager

Interface between Rucio and SENSE

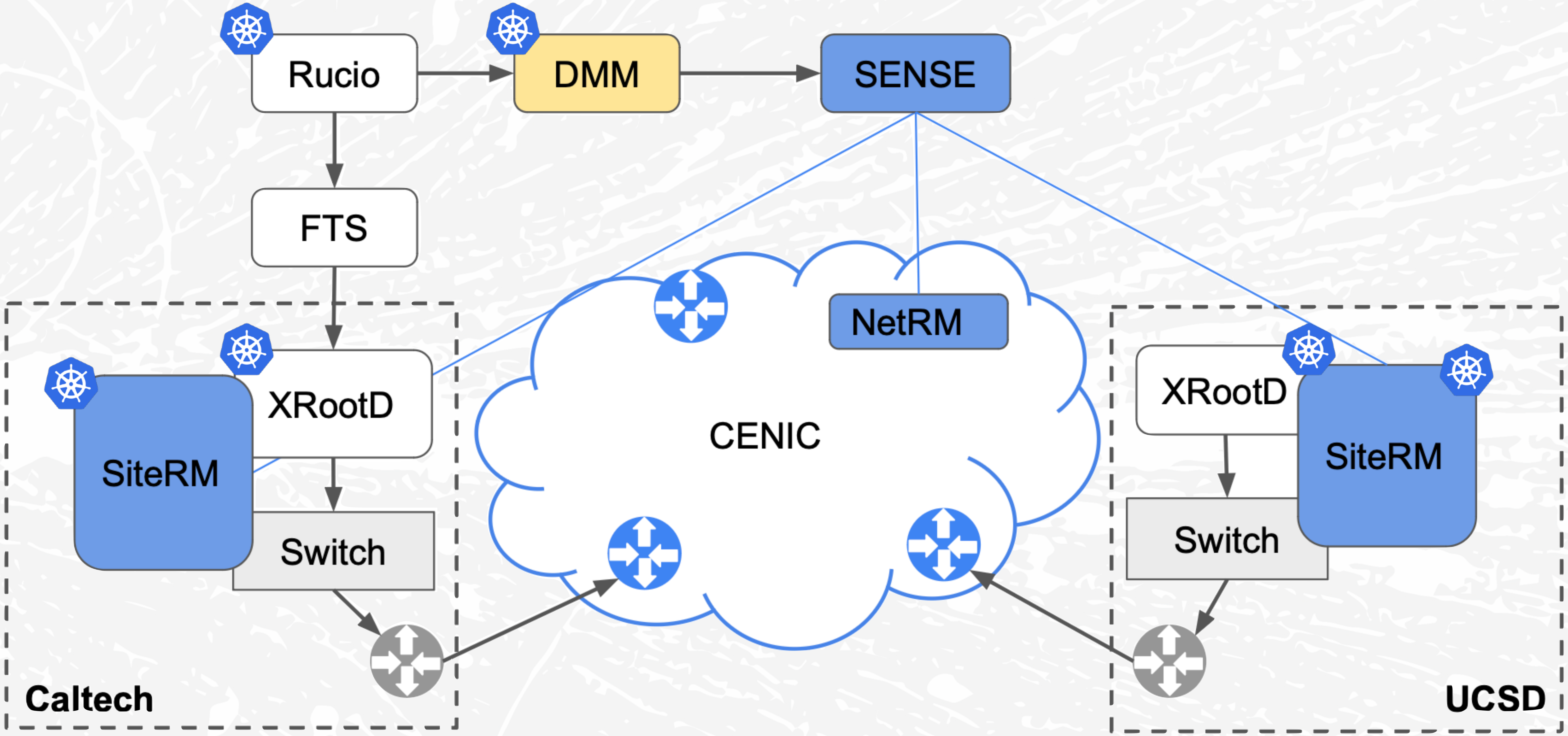
decision making and monitoring

Overall Picture



Rucio → DMM → SENSE → DMM → Rucio → FTS → XRootD

Deployment

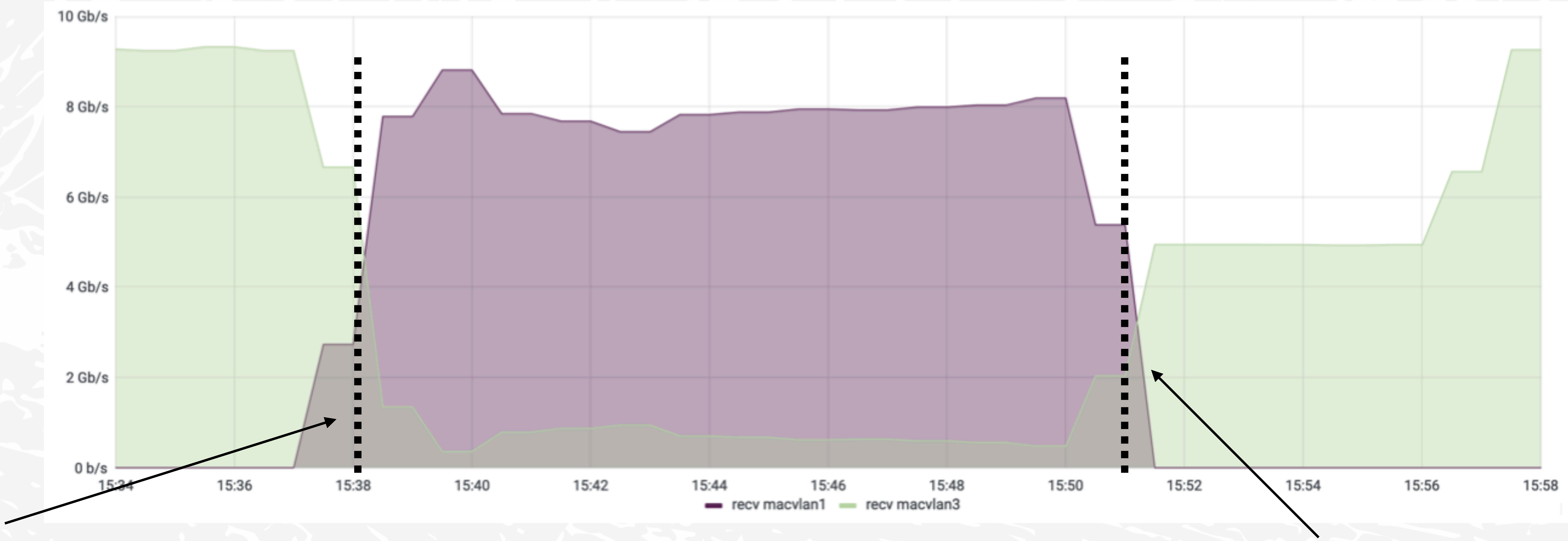


Most pieces are deployed using Kubernetes, others are production services

Current Status

Last Year

- Proof of Concept at 10Gbps with 1 managed allocation and background traffic.



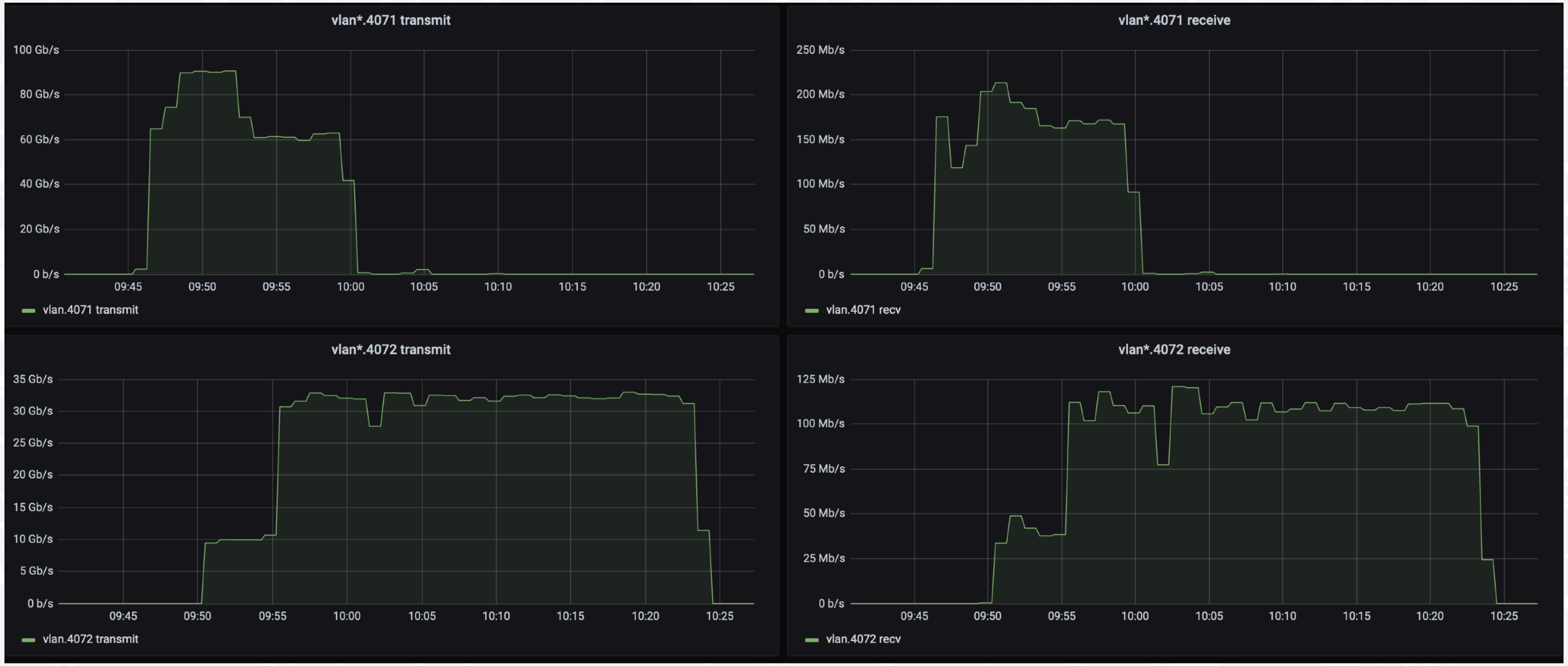
SENSE path is created

SENSE path is deleted

Diagram showing background (green) and priority (purple) traffic through one of the interfaces.

Today

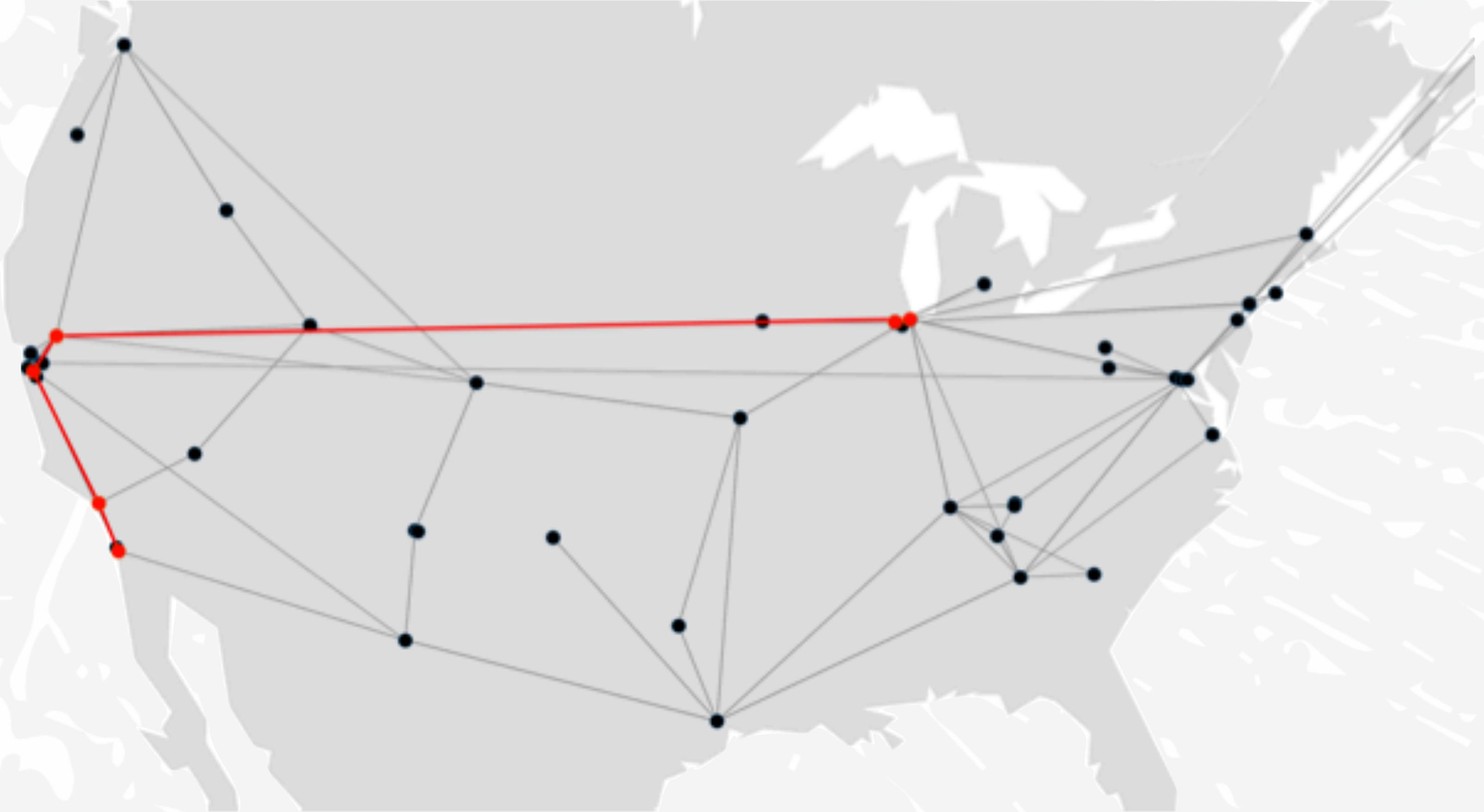
- Two priority flows on a 100Gbps managed link.



100Gbps link between UCSD and Caltech being shared by 2 Priority Paths created by SENSE, each of them using only its allocated share 33/66 Gbps

Future

- Working on new site deployments at University of Nebraska-Lincoln (midwest) and Fermilab.
- Agreed with the Rucio team on initial plan for integration of DMM into Rucio
- New 400Gbps link from LA to ESnet allows us to expand our testbed beyond UCSD and Caltech
- Planning to do high-throughput O(100Gbps) tests at higher latencies



References

- F. Würthwein, J. Guiang, A. Arora, D. Davila, J. Graham, D. Mishin, T. Hutton, I. Sfiligoi, H. Newman, J. Balcas, T. Lehman, X. Yang, & C. Guok. (2022). Managed Network Services for Exascale Data Movement Across Large Global Scientific Collaborations. In 2022 4th Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP). IEEE.

- T. Lehman, X. Yang, C. Guok, F. Wuerthwein, I. Sfiligoi, J. Graham, A. Arora, D. Mishin, D. Davila, J. Guiang, T. Hutton, H. Newman, and J. Balcas, "Data transfer and network services management for domain science workflows," 2022. [Online]. Available: https: //arxiv.org/abs/2203.08280

- J. Zurawski, D. Brown, B. Carder, E. Colby, E. Dart, K. Miller et al., "2020 high energy physics network requirements review final report," Lawrence Berkeley National Laboratory, Tech. Rep. LBNL-2001398, Jun 2021. [Online]. Available: https://escholarship.org/uc/item/78j3c9v4

- I. Monga, C. Guok, J. MacAuley, A. Sim, H. Newman, J. Balcas, P. DeMar, L. Winkler, T. Lehman, and X. Yang, "Software- defined network for end-to-end networked science at the exascale," Future Generation Computer Systems, vol. 110, pp. 181–201, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0167739X19305618

Acknowledgements

- 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 CENIC, ESnet, Caltech, and SDSC.

Thank You!
Questions?

UC San Diego SDSC SAN DIEGO SUPERCOMPUTER CENTER Caltech ESnet ENERGY SCIENCES NETWORK Fermilab

Backup

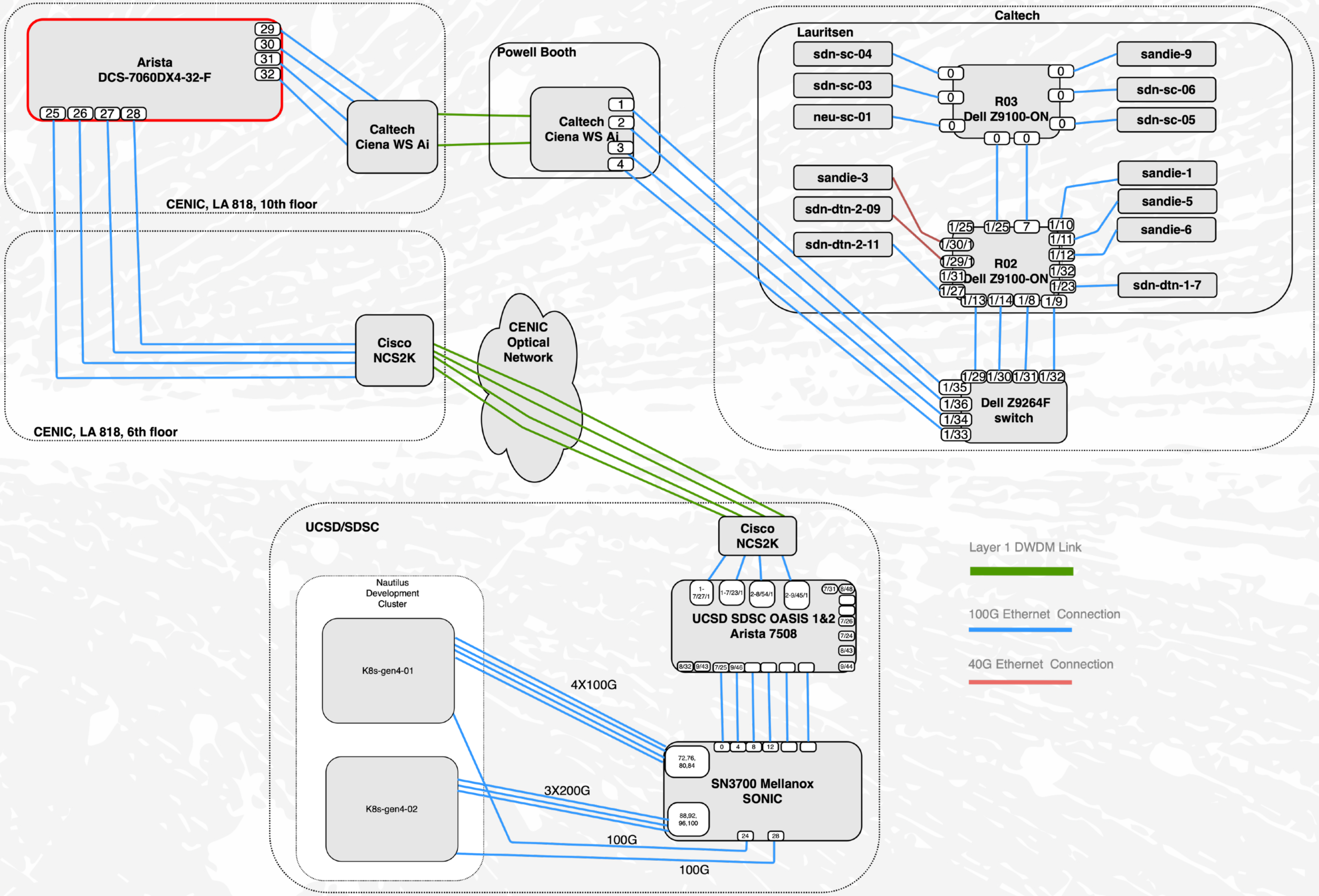
UC San Diego SDSC SAN DIEGO SUPERCOMPUTER CENTER Caltech ESnet ENERGY SCIENCES NETWORK Fermilab

Nuances

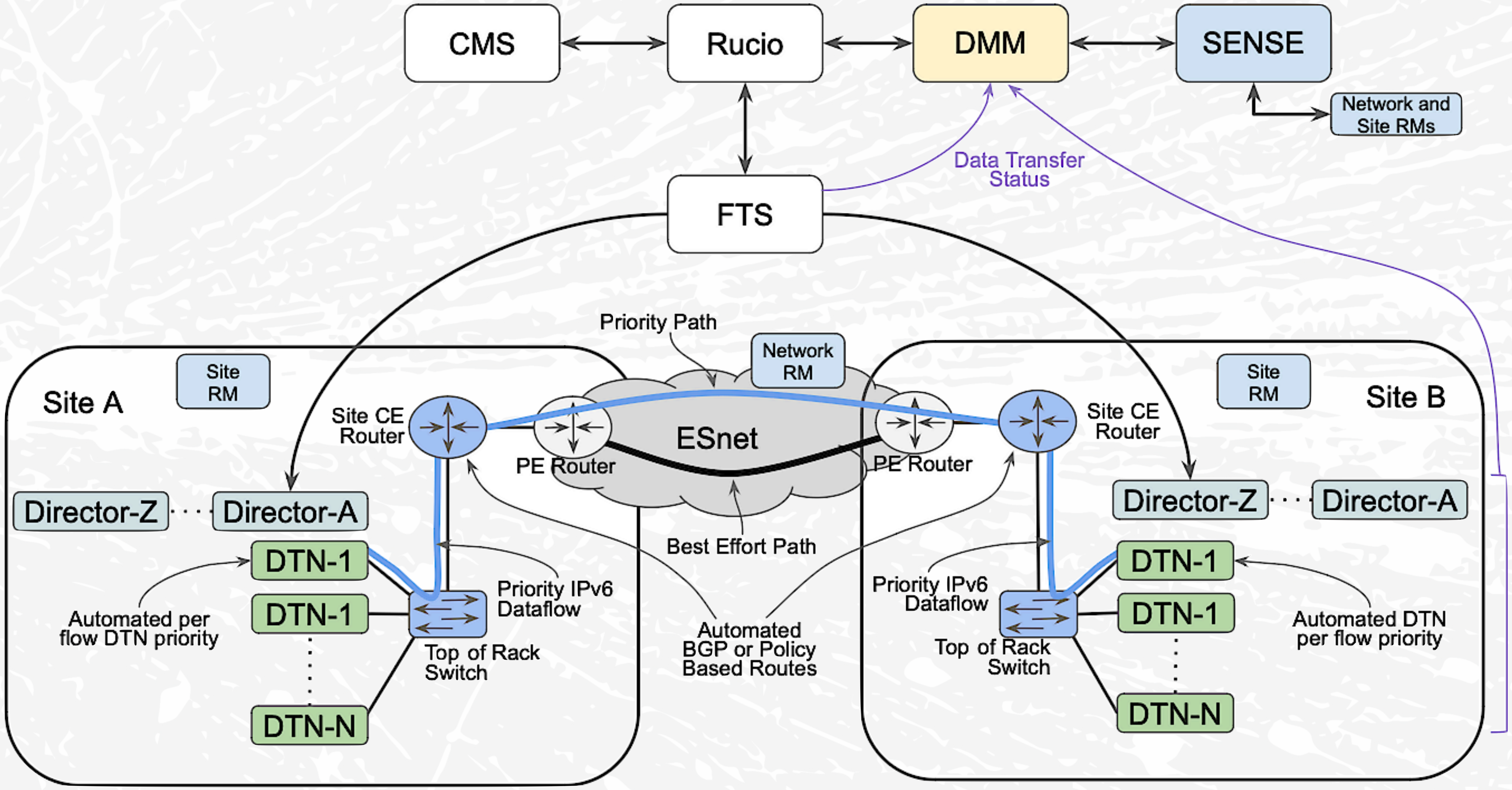
- We need a custom version of Rucio which can pass the transfer metadata to DMM, we deploy this using Helm using the official Rucio charts and apply a patch with our modifications.

- SENSE services are created based on IP subnets but current storage systems listen to a single endpoint.
  - Solution: expose storage systems over multiple subnets
    - Using Kubernetes, deploy many instances of XRootD (clusters), and allocate different IPv6 subnets to each (in our case, using Multus CNI).

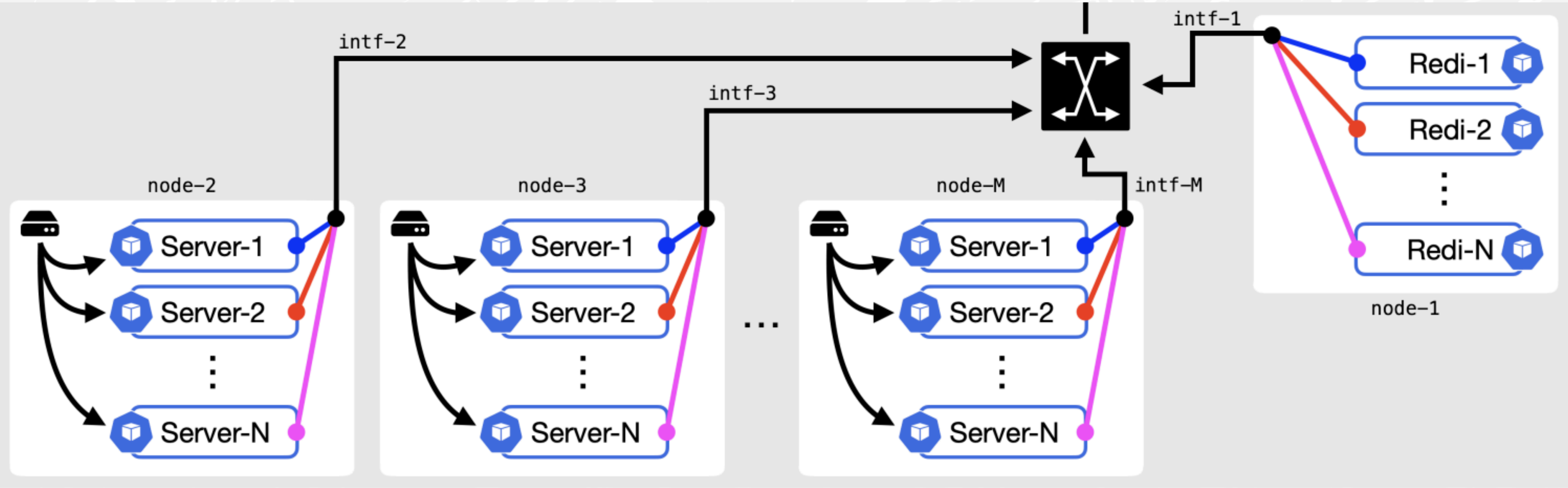
UCSD-Caltech Testbed



Overall Picture (more detailed)



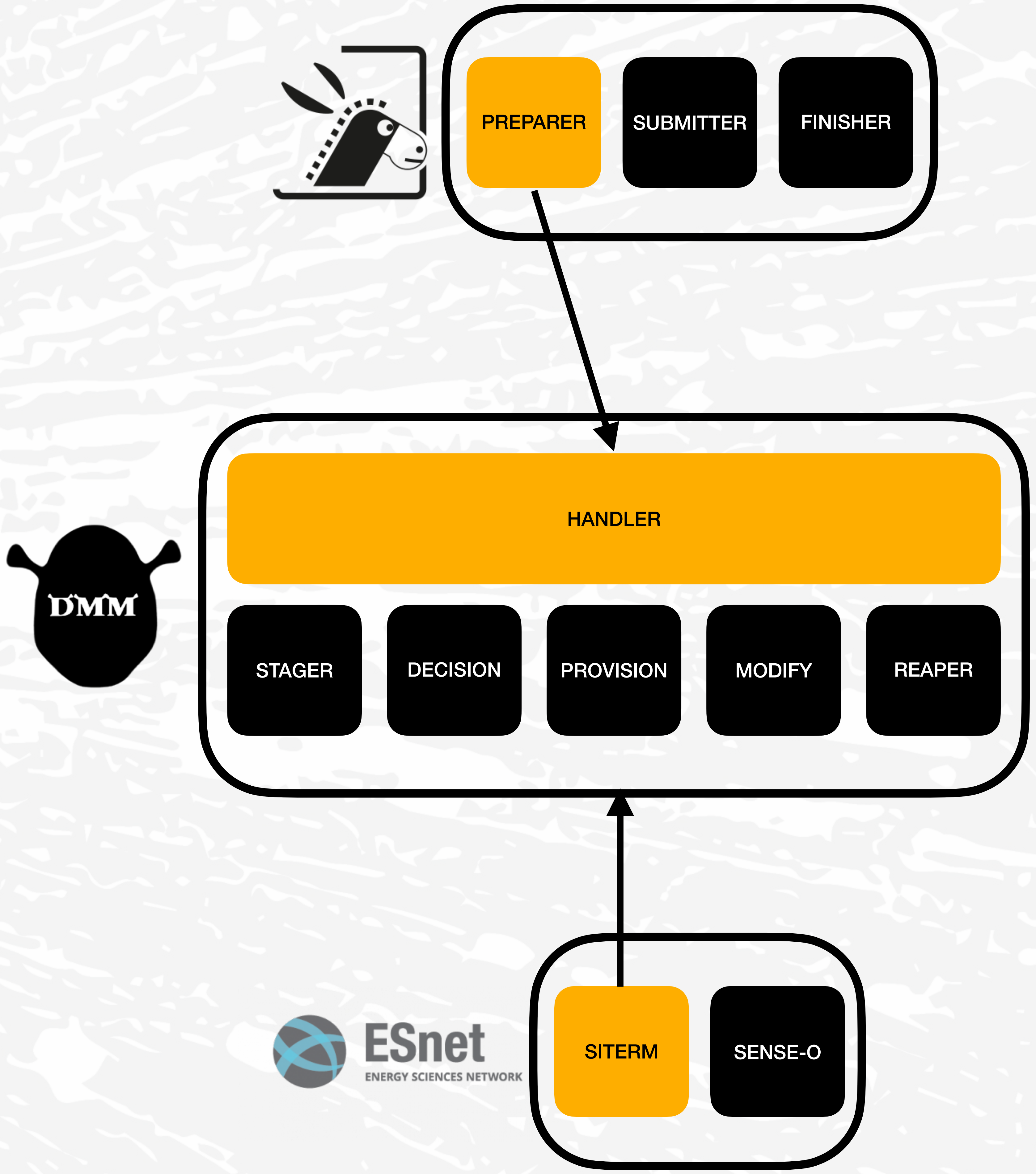
Multi-subnet XRootD deployment



Multiple XRootD clusters deployed over M DTNs. Each color represents a different IPv6 subnet.

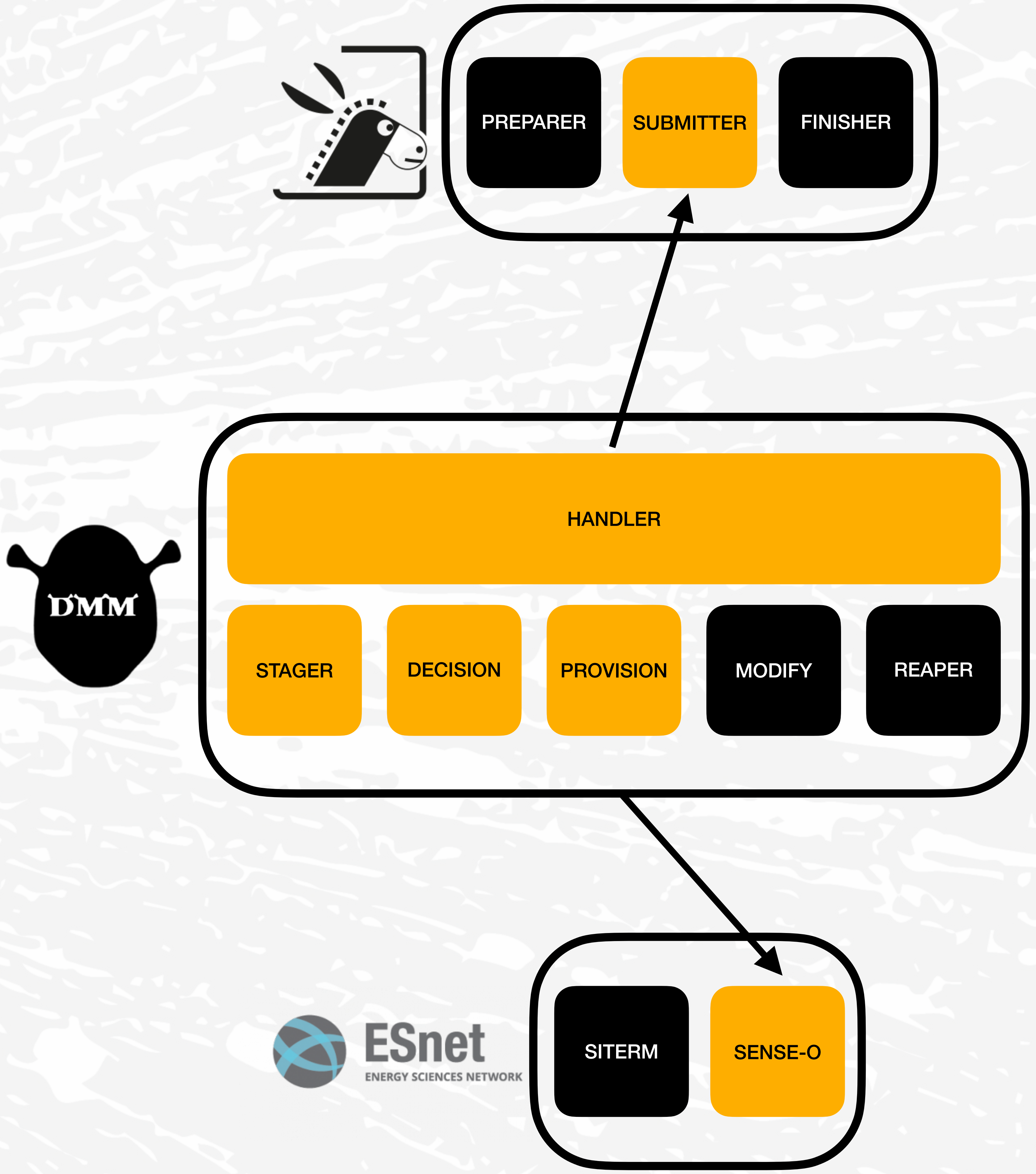
1. add-rule, preparer-handler

- Preparer sends transfer metadata (HTTP)
  - Rule ID, source site, destination site, number of files in the dataset/container, number of bytes to be transferred.
- DMM handler receives the request, adds it to the DB. [INIT]
  - IPv6s are allocated: [ALLOCATED]
    - SiteRM has a list of IPs for a site, DMM queries it, checks if its in use.



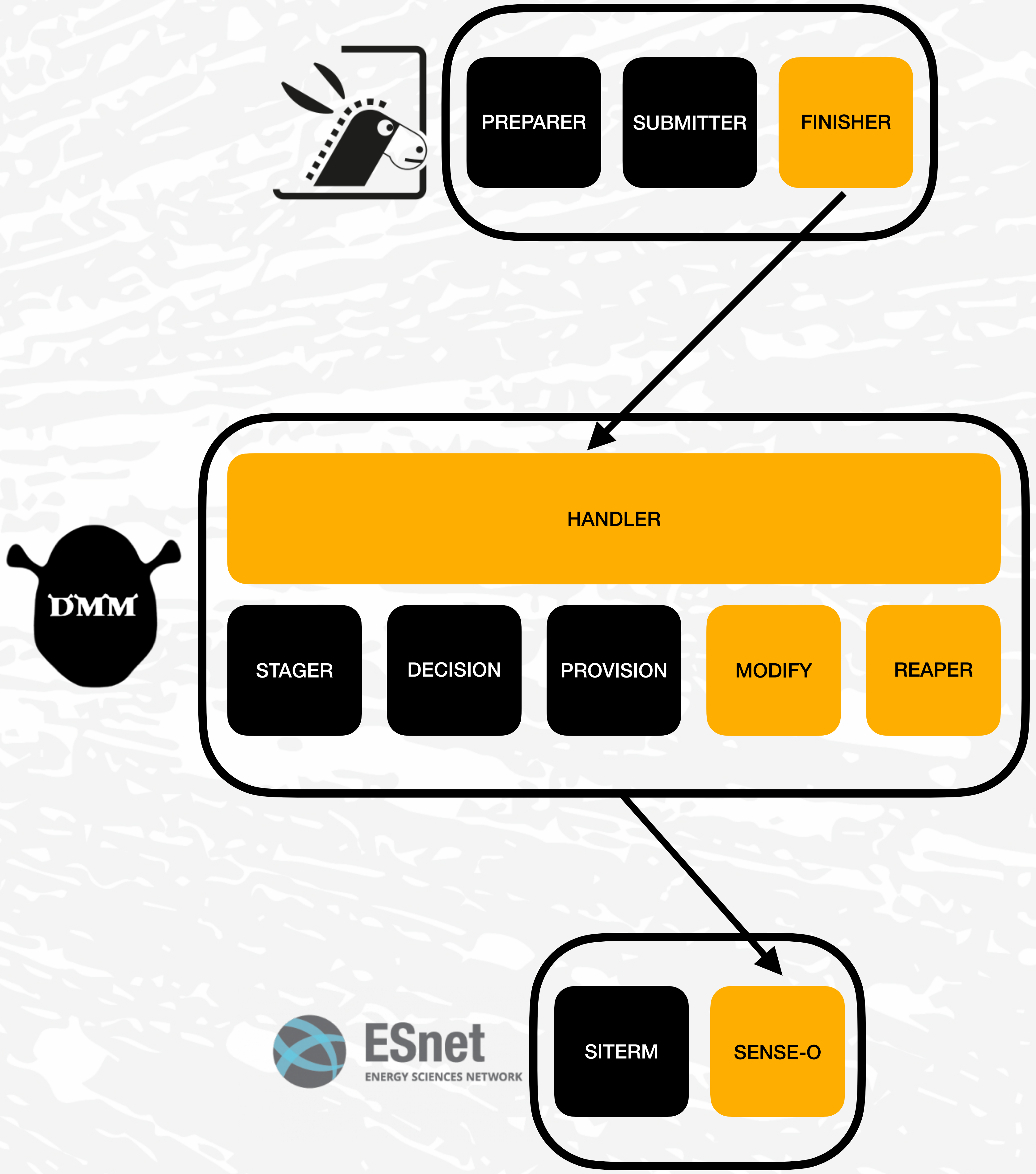
2. stager, decision, submitter-handler, provision

- submitter-handler receives submitter requests and returns the allocated IPs, at this point the transfers start already, although only a few are active (see next point).
  - In Rucio, we replace the IPs right before the FTS submission - not the best way of doing things but it works
- stager daemon stages the SENSE path with the source and destination ips, FTS SE and Link limits are modified, set to a very low number. [STAGED]
- decision daemon (the brain of DMM) creates a network graph for all the requests and traverses through the nodes to allocate bandwidths. [DECIDED]
- provision daemon allocates the bandwidths calculated by the decision daemon to the staged SENSE path. [PROVISIONED]
  - Once the path is provisioned, FTS SE and Link limits are modified, set to very high.
- SENSE path is created and the provisioned bandwidth kicks in.



3. finisher-handler, reaper, modifier

- Finisher handler listens to the finisher-conveyor daemon.
  - One all transfers are finished, the link is free-ed (not deleted) [FINISHED]
  - If another request comes in for the same set of endpoints, we reuse the path by modifying the bandwidth (modify is a lot cheaper than building a path).
  - If no new requests come in within a set window, the path is deleted. [DELETED]


PREPARER
SUBMITTER
FINISHER
HANDLER
DMM
STAGER
DECISION
PROVISION
MODIFY
REAPER
ESnet
ENERGY SCIENCES NETWORK
SITERM
SENSE-O