Towards Accountable Network Bandwidth Utilization via SDN

A.K.A SENSE/Rucio Interoperation

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] and more…

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

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

Motivation

- 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.

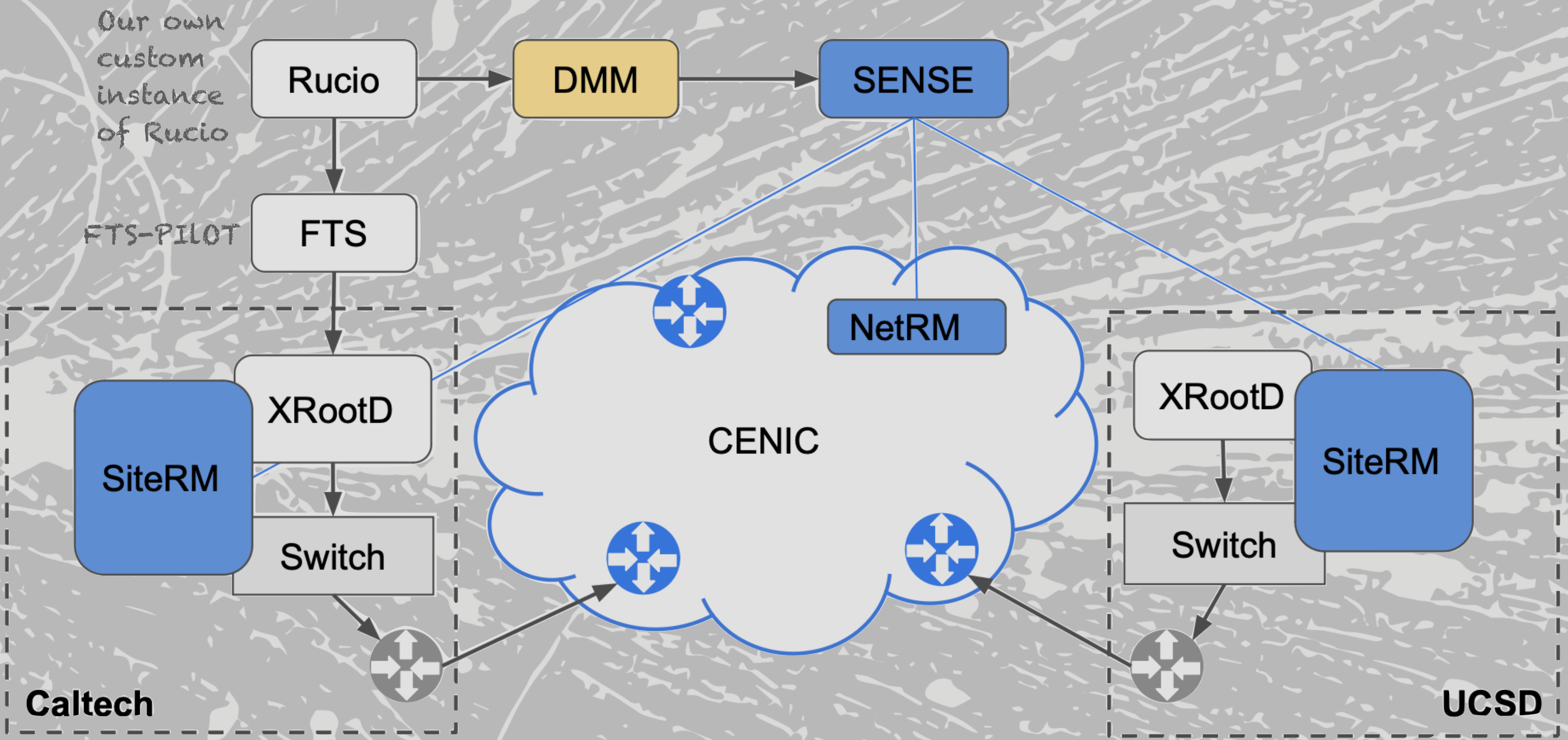
- Software defined networking (SDN) controlled data-flows allow for end-to-end accountability of network utilization, and allows the different VOs to manage their priorities.

- Using the HEP software stack for data movement, i.e. Rucio+FTS+XRootD as the control testing ground, we are integrating these existing tools with SENSE, the SDN service.

SENSE

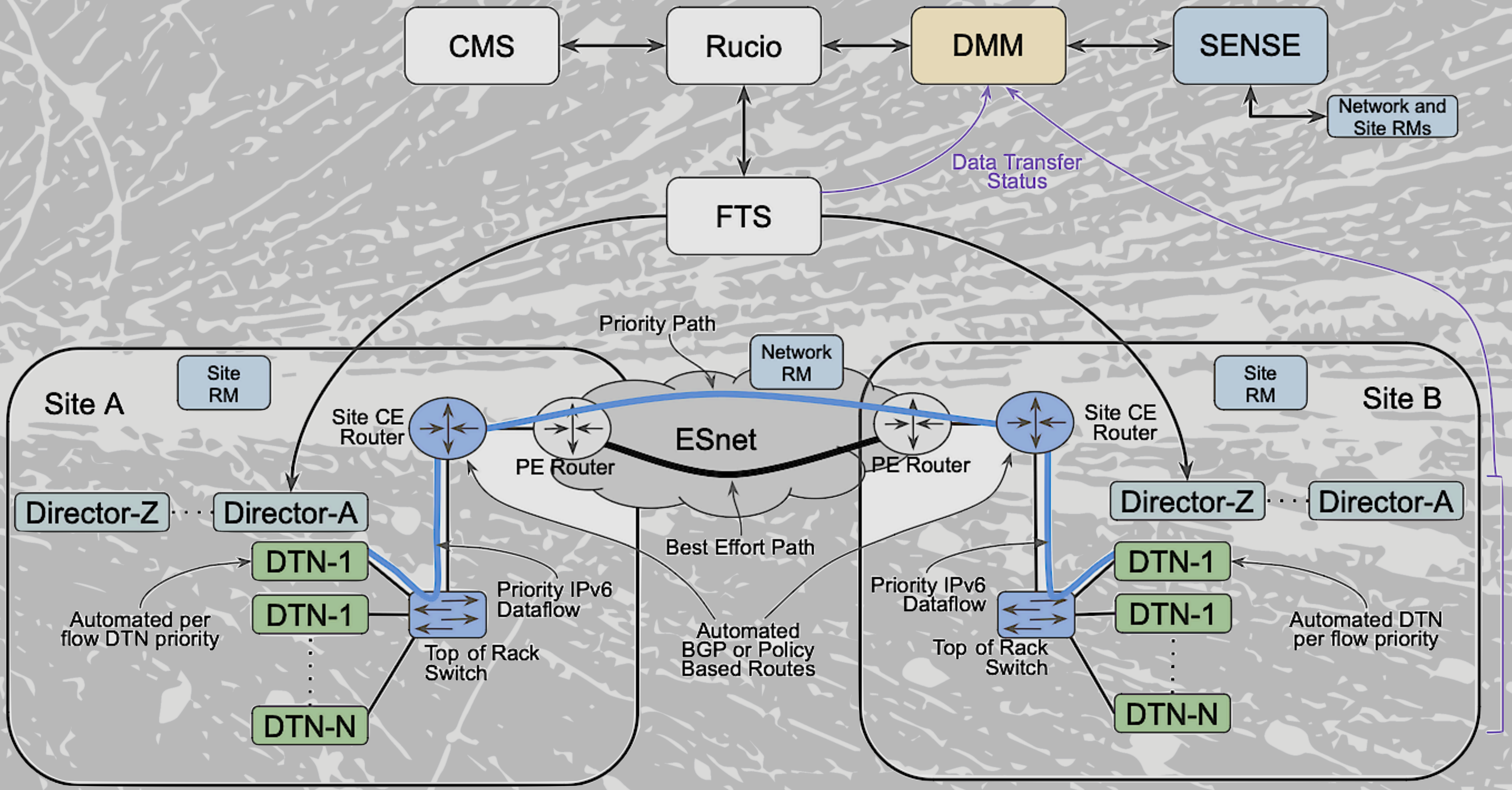
- **S**oftware-Defined Network for **E**nd-to-end **N**etworked **S**cience at the **E**xascale
- Provides the mechanisms to enable multi-domain orchestration for a wide variety of network and other cyberinfrastructure resources in a highly customized manner.
- These orchestrated services can be customized for individual domain science workflow systems and requirements.
  - It can create network services like routing and bandwidth allocation on demand.
- Agents: SiteRM and NetRM push QOS and routing rules.

Overall Picture



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

Overall Picture (more detailed)



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

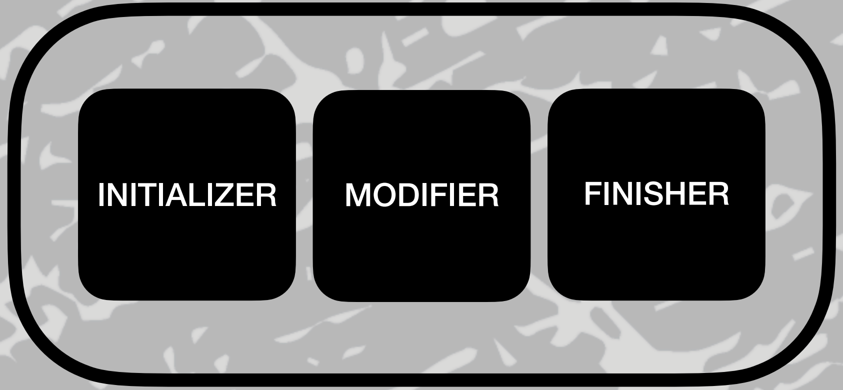
Data Movement Manager (DMM)

- Interface between Rucio and SENSE, making SDN operated HEP data-flows possible
  - **FROM RUCIO**: Gets transfer metadata like source and destination RSE names, size of the rule (number of bytes) and priority of the rule.
  - **FROM SENSE**: Gets SENSE IPv6s corresponding to the RSEs, bandwidth between the endpoints
  - **[TO RUCIO]:** Injects the IPv6s into Rucio at the FTS submission step.
  - Based on rule metadata, makes decision on bandwidth allocation/provision for all rules and tells SENSE to build a dedicated P2P VLAN between the endpoints.
  - Keeps state of all the data-flows, monitors performance and creates reports of underperforming flows.

DMM Technical Details

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

DMM-Rucio Daemons



- DMM, inspired by Rucio, also operates on a daemon based model.
- For Rucio Interactions, there are three daemons:

1. **Initializer**: DMM queries Rucio (using the Rucio python API) for all the rules (would be helpful if rules could be indexed by when they were submitted so DMM could query only for the rules added in the last n hours)
    - Rule ID, source site, destination site, number of files in the dataset/container, number of bytes to be transferred and the priority.
    - DMM sees a new rule, adds it to the DB in [INIT] state
2. **Modifier:** If priority of a rule gets modified, updates the priority in the DB
3. **Finisher**: When rule is marked “OK” or “STUCK” for too long, Mark request as [FINISHED]

DMM-Core Daemons

1. **Allocator**: Picks an unused IPv6 from the SENSE address pool and assigns it to the rule.

2. **Decider**: Brain of DMM, does a graph based calculation for bandwidth allocation: sites represent nodes and rules are edges. The graph gets updated with each rule and based on the relative priority, the bandwidth is allocated based on fair share.

3. **FTS Config**: Set FTS SE and link limits appropriately based on bandwidth allocation and latency.

DMM-SENSE Daemons

PROVISION MODIFIER DELETER

1. **Provision**: submits request for the SENSE circuit to be built

2. **Modifier:** based on priority changes / new rules being added, modifies the allocated bandwidth of the existing circuits.

3. **Deleter:** Building circuits is expensive, so once a rule is marked as finished, DMM keeps the circuit up for 10 minutes (change the allocated bandwidth to 1Gbps), if no new rule between the same pair of endpoints comes in then take down the circuit so the endpoints can be used for other rules.

Other Features

1. **Monitoring Dashboard:** shows status of the request.

2. **Accountability reports:** DMM has access to a lot more information about the request status, namely the host level information through node_exporter + prometheus (runs in site-RM) and also FTS-monit information (CERN Opensearch).

   - Using all of these metrics, we can determine which sites underperform, i.e. are not able to meet the bandwidth that was allocated to them for a rule. And based on that, change the provisions + generate reports for the site admin to look at to see what is the point of failure.

Our Custom Rucio Instance

- Deployed using the official Helm charts, apply a patch for our changes.
- At the FTS submission step, from within the transfertool/fts.py script, query DMM using the rule_id
- DMM returns the new endpoints right away (there is almost no delay, i.e. new endpoints are allocated in < 10 seconds of the rule being added to DMM).
- Change the endpoints in the FTS post request and it proceeds as usual.
- Error handling: If there is an issue, forget DMM and proceed as normal and use RSE's configured endpoints.

Results / Status

- We have showed demos of the end-to-end workflow involving all components at multiple occasions, most notably at SC23, DC24 and OFC over high bandwidth links (100Gbps).
- We were able to expand our testbed from just UCSD + Caltech to also include Fermilab and now UNL. Also trying to go to 400Gbps.
- Our tests until now involved multiple rules between UCSD, Caltech and FNAL with different priorities.

Next Steps

- Add more sites, including CERN and SPRACE.
- Experiment with other storage systems
- Have this in prod for DC26

Thank you!

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

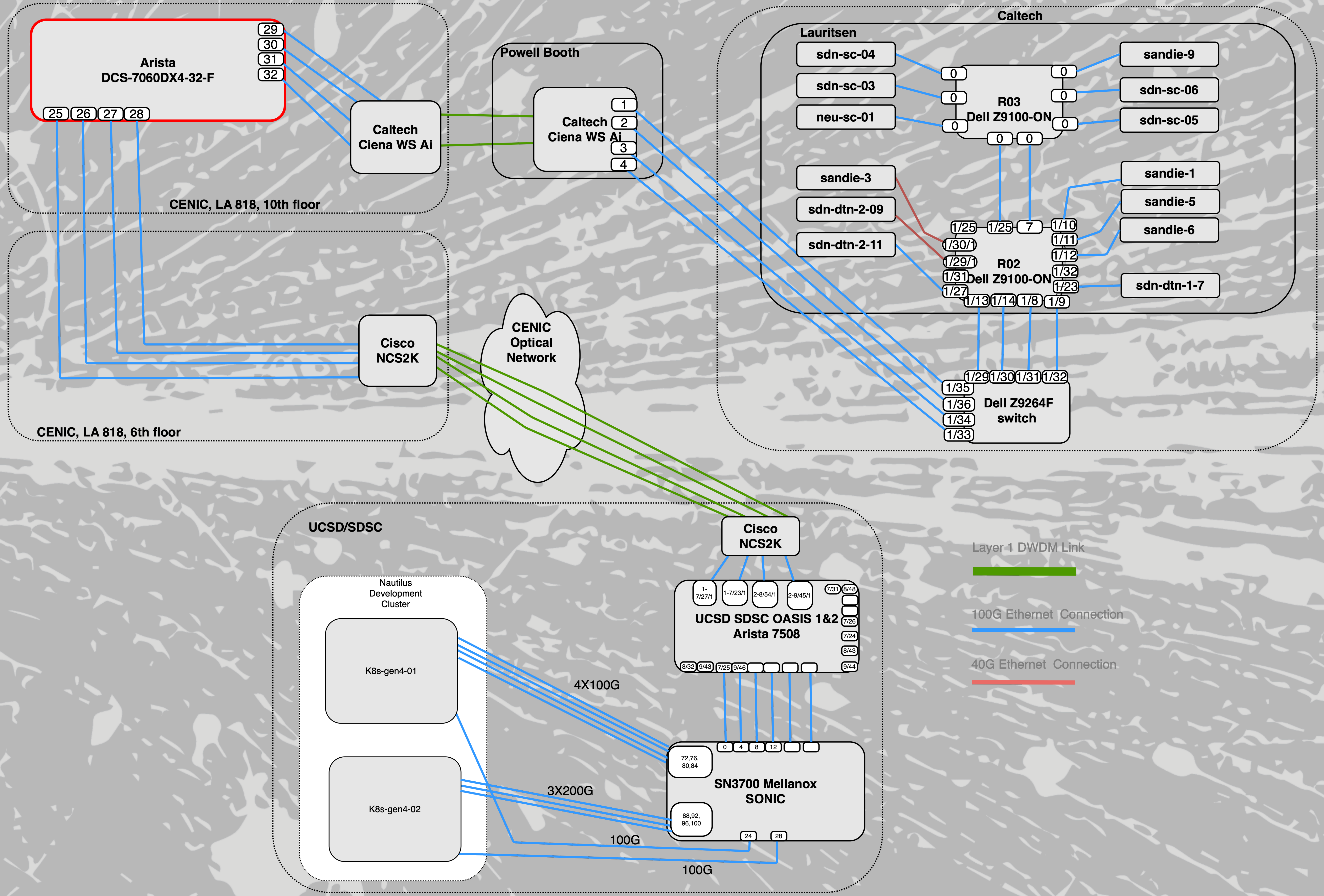
Acknowledgements

- 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.

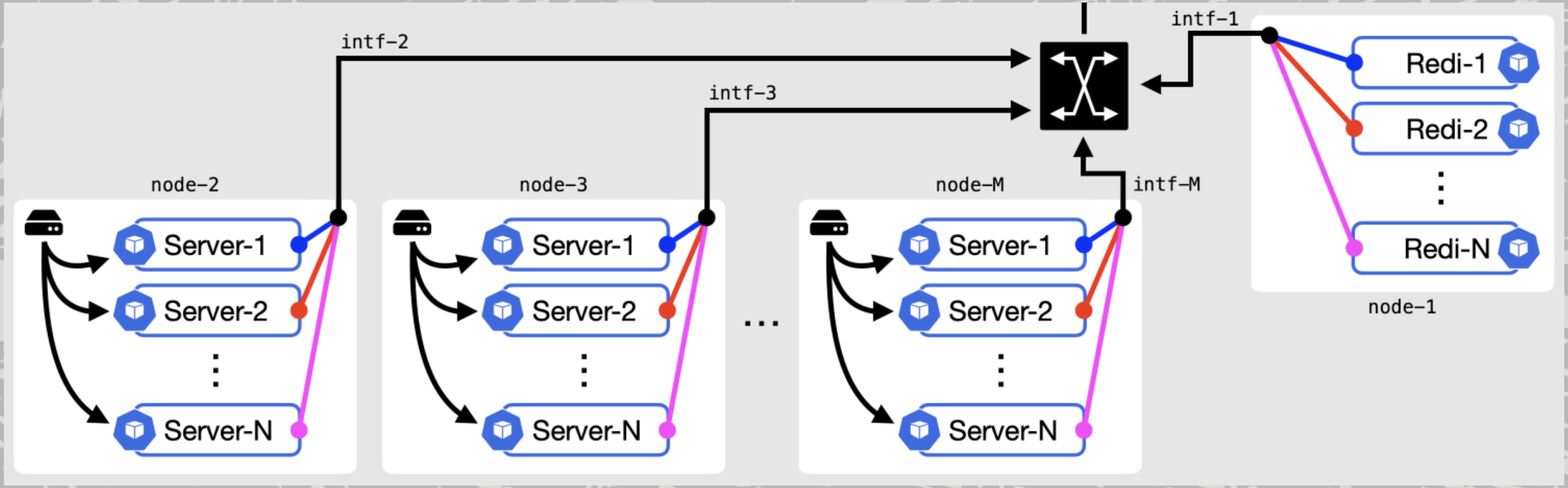
Backup

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

UCSD-Caltech Testbed



Multi-subnet XRootD deployment



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

Multiple interface setup is managed using Multus Kubernetes CNI.