

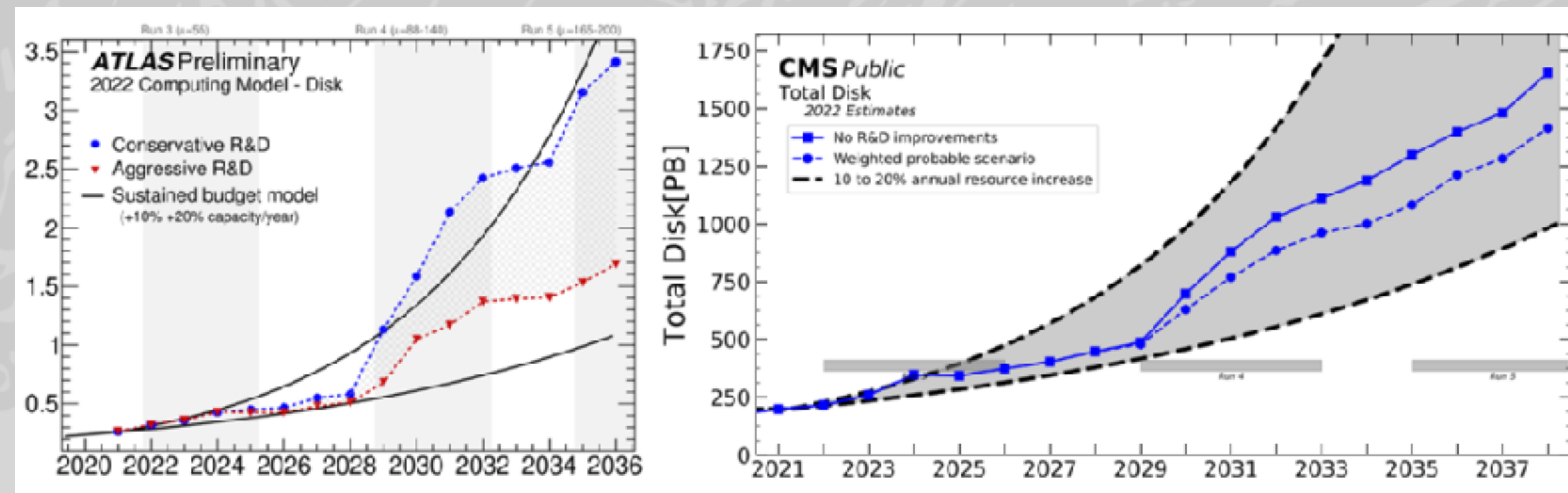
Data Movement Manager (DMM) for the SENSE-Rucio Interoperation Prototype

Frank Würthwein¹, Jonathan Guiang¹, **Aashay Arora**¹, Diego Davila¹, John Graham¹, Dima Mishin¹, Thomas Hutton¹, Igor Sfiligoi¹, Harvey Newman², Justas Balcas³, Preeti Bhat², Tom Lehman³, Xi Yang³, Chin Guok³, Oliver Gutsche⁴, Asif Shah⁴, Chih-Hao Huang⁴, Dmitry Litvinsev⁴, Phil Demar⁴, Marcos Schwarz⁴ and more...

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

Motivation

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



	# of collisions	# 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

- All this additional data is a challenge on the storage side, but what about network?
- Global collaborations depend on efficient data transfer, network is a limited resource.
 - Current model of data transfers, namely summarized as push-now-worry-later will not be feasible in the near future, we need accountability.
- What if we could have controlled data-flows for the largest datasets which dominate network usage?**

Software Defined Networking (SDN)

- Traditional networks are static, with fixed paths and configurations, limiting flexibility.
- SDN decouples the control plane (how data is routed) from the data plane (how data moves), allowing dynamic, programmable networks.
- Allows for guaranteed bandwidth-allocated paths.
- Enables us to make sure that high-priority data flows receive the necessary bandwidth and resources, optimizing the performance of large-scale transfers across multiple sites.



SENSE (SDN for End-to-end Networked Science at the Exascale)

- SENSE is an SDN-based service designed by ESNNet.
 - It enables automated and dynamic provisioning of network paths tailored for large-scale science data flows.
 - Allows for optimizing bandwidth usage and reducing latency.
- Setting up a SENSE path for a data transfer requires knowledge of available network paths and other users' priorities
 - **How can we streamline this through automated data management systems like Rucio that already have information about these priorities?**



SENSE (SDN for End-to-end Networked Science at the Exascale)

- SENSE is an SDN-based service designed by ESNet.



Software defined network control for LHC Experiments



24 Oct 2024, 17:09

18m

Room 2.A (Seminar Room)

Talk

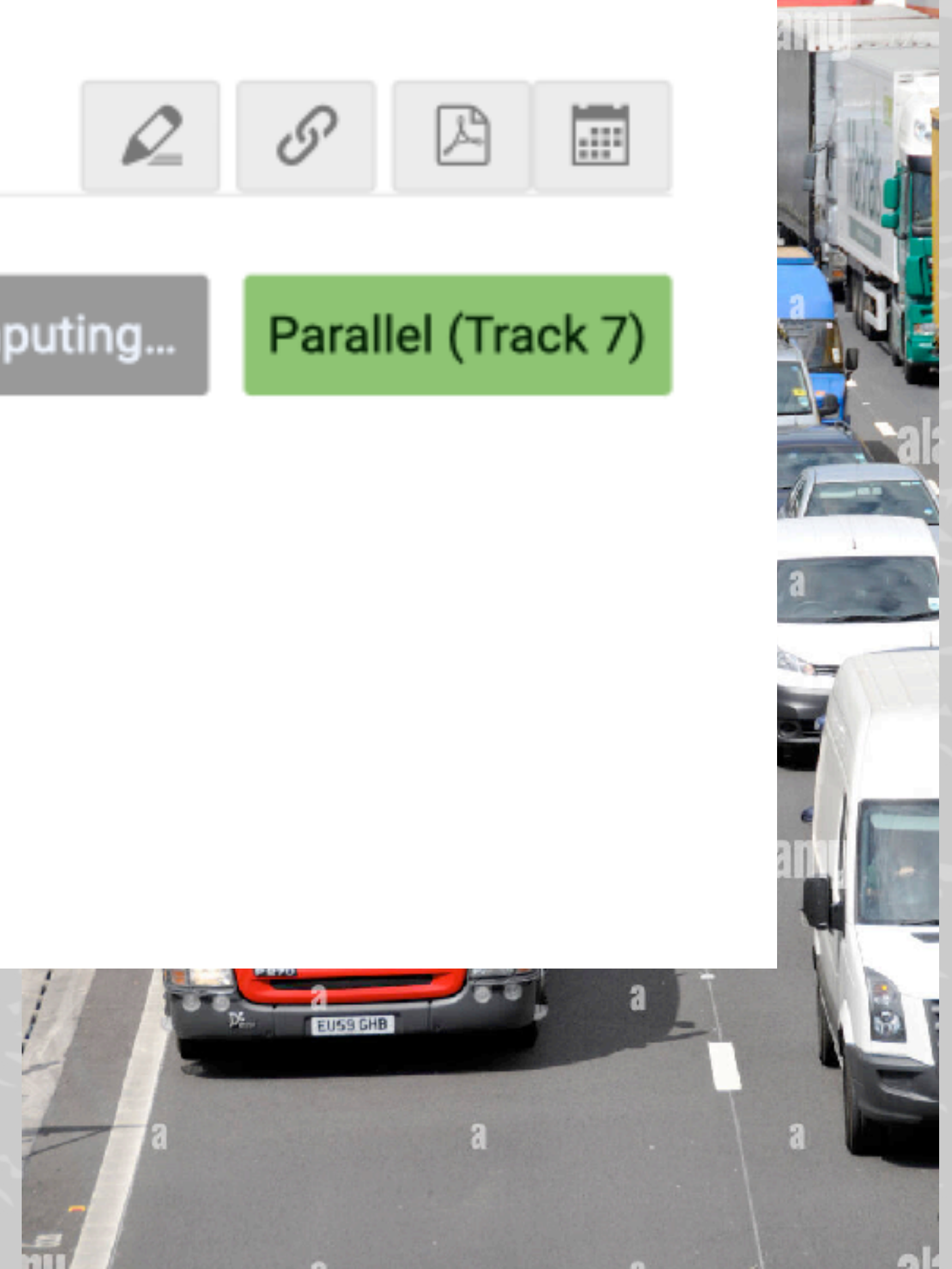
Track 7 - Computing...

Parallel (Track 7)

Speaker

Justas Balcas (ESnet)

management systems like Rucio that already have information about these priorities?



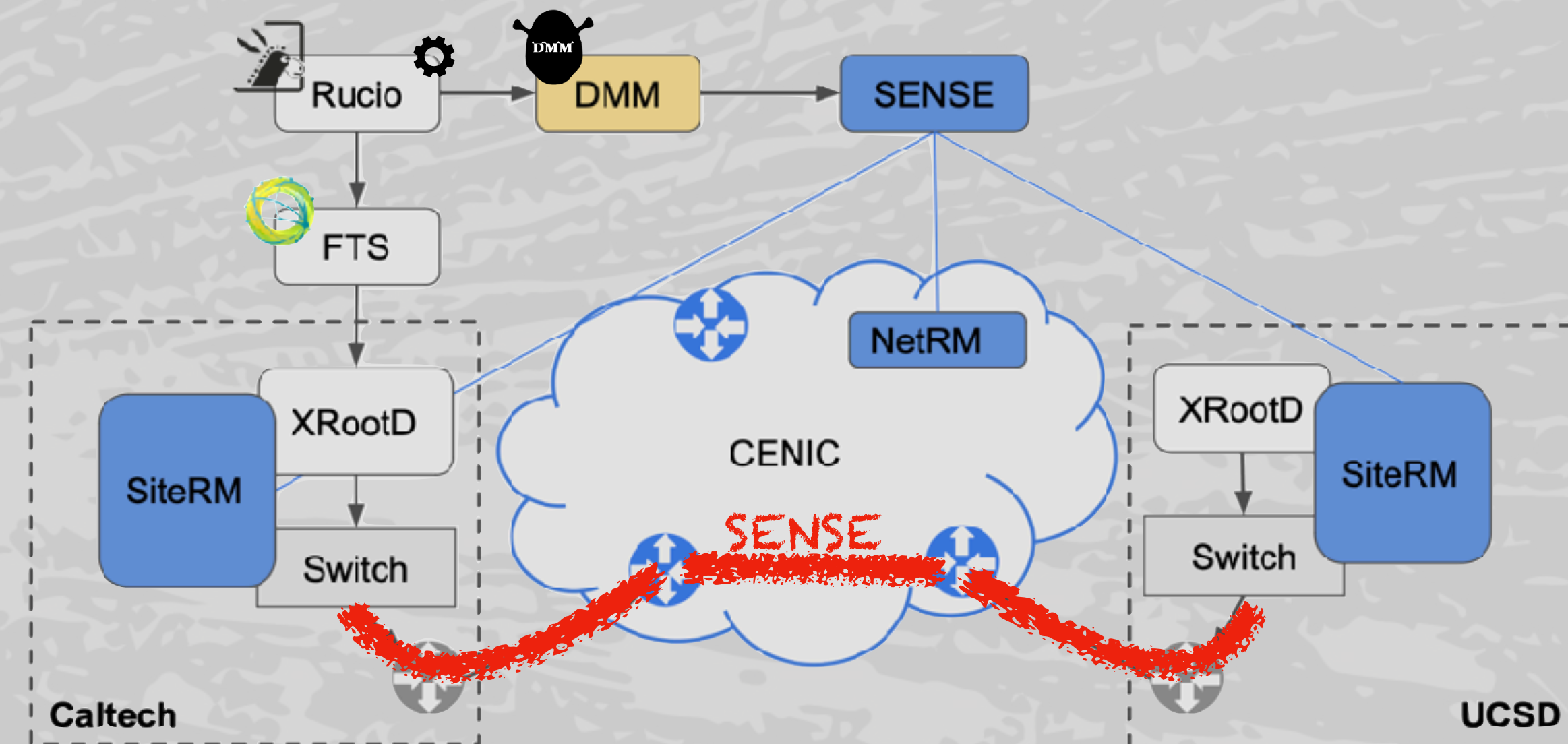
Data Movement Manager (DMM)



- **DMM: Bridging Rucio and SENSE**
- A prototype interface connecting Rucio's data management with SENSE's SDN service.
- Enables SDN-optimized data transfers across the WLCG.
- Allows for prioritization of important data transfers to ensure critical data flows are completed quickly and efficiently.
- Monitors network-level and transfer-level metrics to identify underperforming flows.
- Optimizing the throughput by managing the number of active transfers for a rule.

How does it work? In Summary:

- SENSE can establish connections between endpoints, providing Quality of Service (QoS) at the IP subnet level.
- Projected XRootD deployment exposes multiple IP ranges (info in backup).
- Rucio manages data transfers and needs specific endpoints to pass to the File Transfer Service (FTS) for the transfer jobs.
- DMM maintains an updated list of SENSE-controlled IP ranges along with their corresponding endpoints.
 - When a new transfer rule is added in Rucio, it contacts DMM and gets back this new pair of endpoints.
 - DMM tells SENSE to create a circuit between the specific IP ranges, allocating bandwidth according to the transfer priority set in Rucio.
 - It also changes the number of active jobs on FTS to meet this bandwidth.
 - When the rule is finished, DMM marks the circuit as available for other rules which use the same endpoint, if no such rule is seen within 10 minutes, the circuit is deleted.



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

DMM Integration with Rucio

- Done primarily using the Rucio Python client.
- DMM periodically queries Rucio for new rules.
 - When it sees a new rule, it gets the rule metadata, namely source and destination sites, rule priority, and total transfer size.
- Rucio contacts DMM through DMM's REST API with the rule_id and gets the SENSE endpoints in return (more on this in the next slide)
- DMM monitors each ongoing rule for modifications in priority.
- When a rule is marked as finished in Rucio, DMM frees up the IP ranges to be used for a new rule.



Changes in Rucio

- To replace the transfer endpoints with the ones controlled by SENSE, we add a small patch into Rucio.
- The patch, added at the FTS submission step can be summarized as:

```
if rule_needs_sense():
    for transfer in transfers_to_be_submitted_to_fts:
        new_endpoints = query_dmm(transfer.rule_id)
        transfer.replace(curr_endpoints, new_endpoints)
else:
    continue
.. ..
transfer.submit_to_fts()
```

- These changes cause no disruption to the normal Rucio workflow, DMM returns the new endpoints right away (there is almost no delay, i.e. new endpoints are allocated in $O(\text{seconds})$ of the rule being added to Rucio).

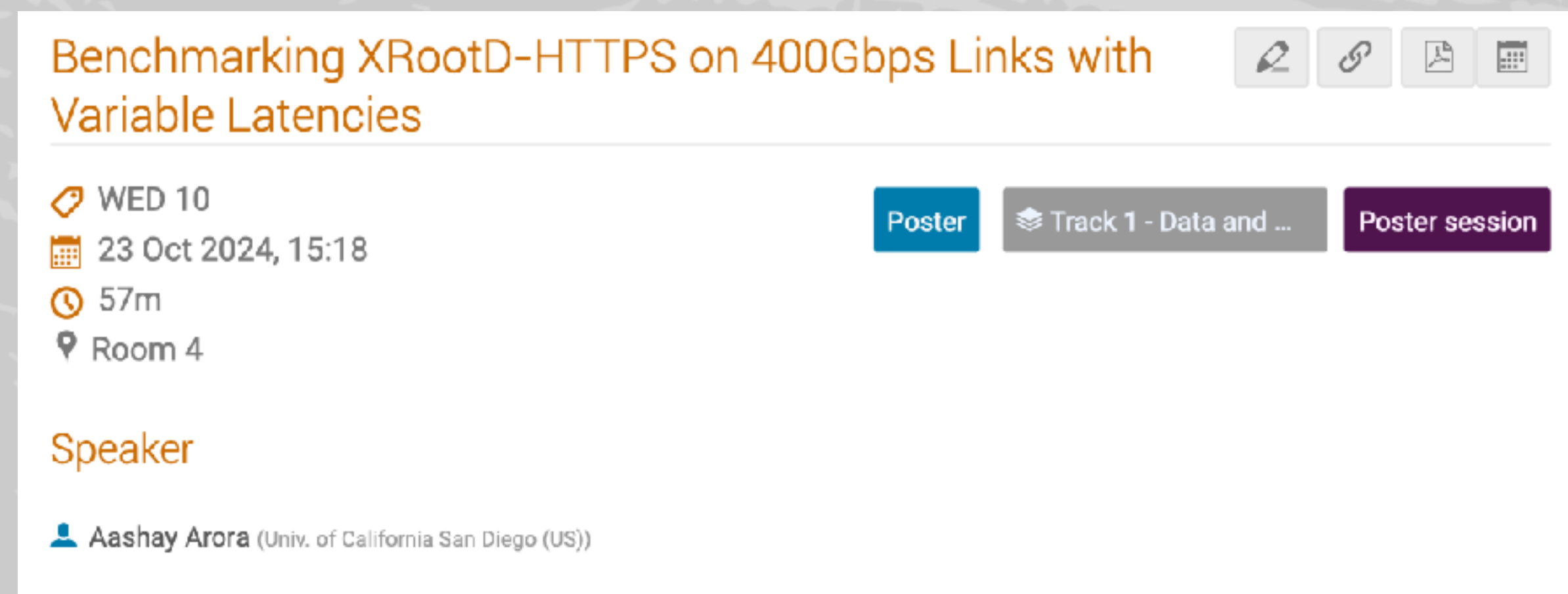
DMM Interaction with SENSE

- DMM interacts with SENSE using the REST API
- Primary communication involves
 - Retrieving the list of endpoints and IP ranges for all the sites in the DMM configuration
 - At the time of requesting a SENSE circuit:
 - Retrieving an IP range which is not in use by any other circuit.
 - Request the provisioning of the circuit with a dedicated bandwidth.
 - Modifying the bandwidth of a circuit in case the priority of an ongoing rule is changed.
- Taking down the circuit when it's not in use.



Interaction with FTS

- Since the maximum attainable throughput is a function of the number of transfers in flight, we need to tune the number of active jobs in FTS. DMM is able to do this using FTS' REST API.
- The number of jobs should be decided based on the allocated bandwidth and the latency between the sites.
- How many active transfers do we need to saturate a given link?



Benchmarking XRootD-HTTPS on 400Gbps Links with Variable Latencies

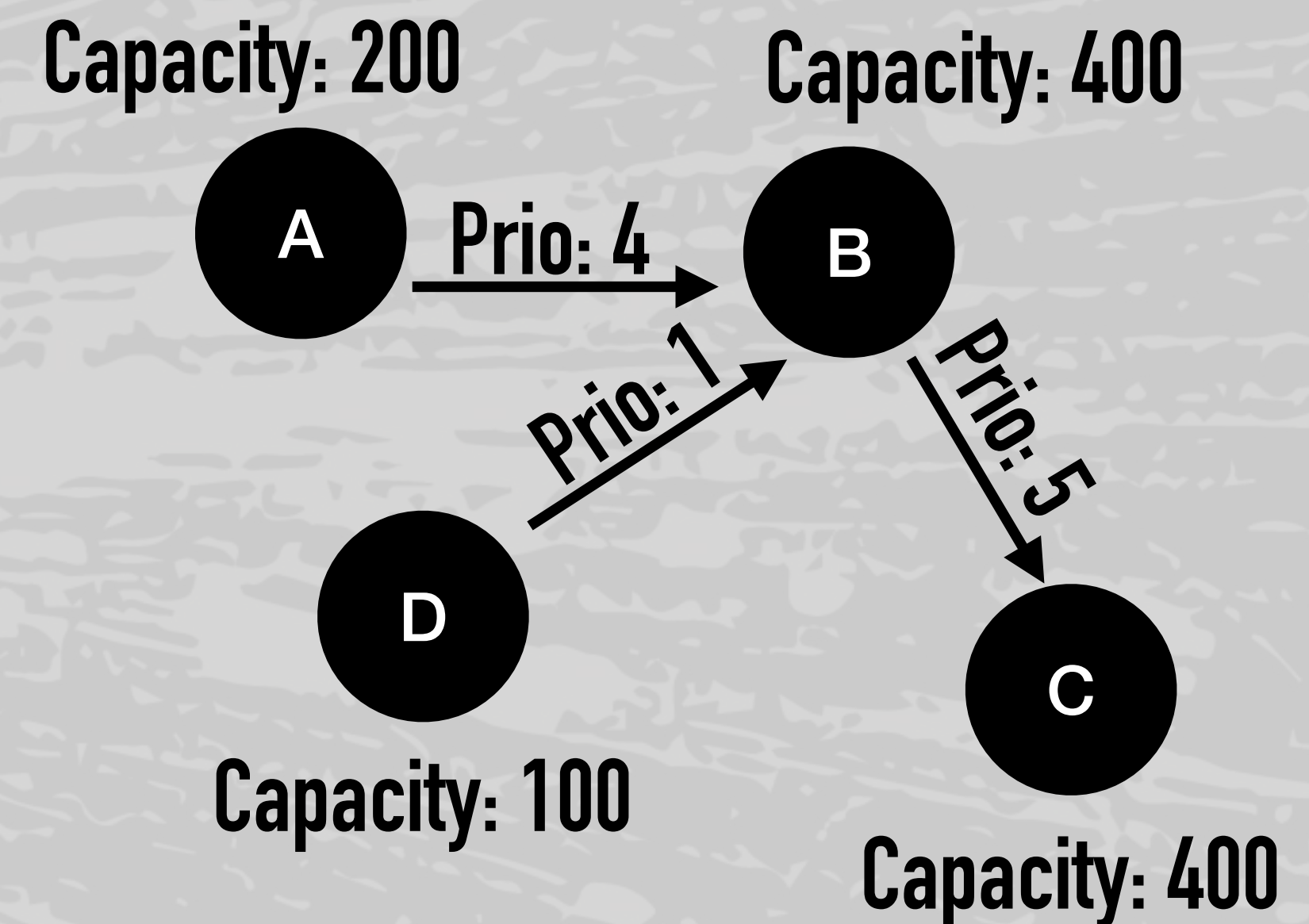
WED 10
23 Oct 2024, 15:18
57m
Room 4

Poster Track 1 - Data and ... Poster session

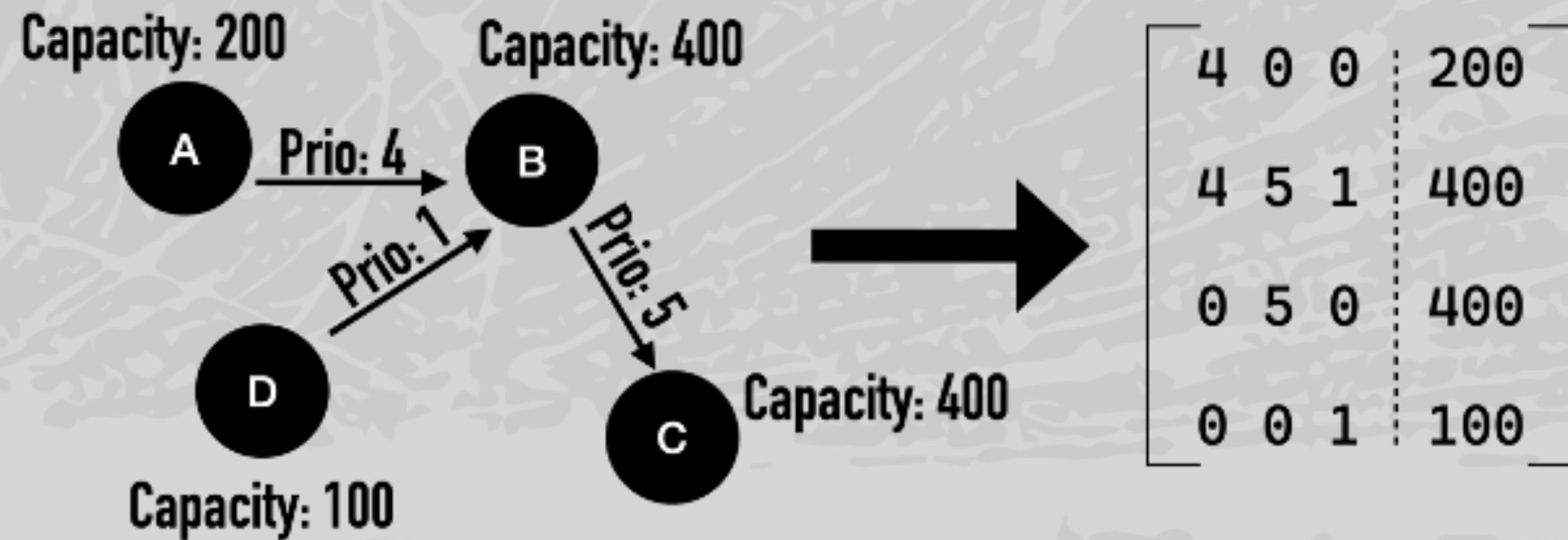
Speaker
Aashay Arora (Univ. of California San Diego (US))

Bandwidth Decisions

- The decision of how much of the bandwidth to allocate to a given Rucio rule is made based on the relative priorities of the rules.
- This is a hard problem due to the possible cycles in the transfer topology, in which case iteratively traversing the graph can be an infinite loop.
- We choose a heuristic based on a linear programming optimization according to the following scheme:
 - A priority-weighted adjacency matrix is constructed from the transfer (multi-)graph where the nodes represent sites and edges represent rules.
 - The constraints of the optimization are the uplink capacities of the sites.
 - The allocated bandwidth is based on the product of the rule's priority and the weight obtained for an edge by solving this ill-defined system.
 - From the set of viable solutions, we pick the one which matches the ratios of the priorities the closest.



Bandwidth Decisions: illustrated



$$\left[\begin{array}{ccc|c} 4 & 0 & 0 & 200 \\ 4 & 5 & 1 & 400 \\ 0 & 5 & 0 & 400 \\ 0 & 0 & 1 & 100 \end{array} \right]$$

Viable solution set of allocated bandwidths:
(in order [A→B, B→C, D→B])

- [20. 375. 5.]
- [40. 350. 10.]
- [60. 325. 15.]
- [80. 300. 20.]
- [100. 275. 25.]
- [120. 250. 30.]
- [190. 175. 35.]
- [160. 200. 40.]

$$\longrightarrow \frac{c_0}{c_1} = 0.8, \frac{c_1}{c_2} = 5, \frac{c_0}{c_2} = 4$$

Objective:

$$\max_x c^T x \text{ for } Ax \leq b$$

b: uplink capacity, c: priorities

Pick the last one because the ratios match that of the priorities the closest (in this case exactly)

Online and Offline Monitoring

SENSE Utilizes tools like Prometheus, Node Exporter, and Grafana and enables real-time monitoring and debugging across network paths hop-by-hop over multiple domains.

Additionally, by accessing FTS metrics from CERN's MonIT system, DMM can gather detailed job-level data, including information on failures, retries, and overall transfer performance.

- **Real-Time Performance Monitoring:**

- DMM periodically queries Prometheus to monitor the throughput of ongoing transfers and can adjust bandwidth allocations dynamically based on real-time performance.

- **Job-Level Insights from FTS:**

- DMM correlates job-level data with host-level metrics to generate comprehensive reports for each transfer, providing deep insights into both system performance and potential issues.



DMM Frontend

Frontend Dashboard: Shows all active requests being handled by DMM, as well as the status of the sites.

Data Movement Manager														
Home Sites														
Rule ID	DMM Status	Source RSE	Source IPv6 Range	Source Hostname	Destination RSE	Destination IPv6 Range	Destination Hostname	Request Priority	Allocated Bandwidth (Gbps)	SENSE Instance UUID	SENSE Circuit Status	Throughput (Gbps)	Health	Details
ba721f59fa854264831b7f...	DELETED	T2_US_Caltech	2605:d9c0:5:2646::/64	redn-c9.i2-sense.ultraigh...	T2_US_SDSC	2001:48c0:3001:112::/64	xrozd-sense-ucsd-redie...	3	325.745	02b5a639-bff5-446d-997...	CANCEL - READY	0.0		See More
4fb556c11c6547999879c...	DIRTYFD	T1_US_PNL	2620:6a0:0:2841::/64	cmsama4-origi-2841-1...	T2_US_SDSC	2001:48c0:3001:111::/64	xrozd-sense-ucsd-redie...	5	100.0	a1e45c54-0125-4a3b-9f...	CANCEL - READY	0.0		See More

Rule ID

Source and Destination IP ranges & Endpoints

Priority

Allocated Bandwidth and current Throughput

Status and next steps

- DMM is currently deployed in a controlled testbed environment, involving a small number of sites and handling $O(10)$ data transfer requests.
- **Next Steps:**
 - **Transition to Pseudo-Production:** We plan to move DMM into a pseudo-production phase, starting with CMS's “integration” instance of Rucio.
 - **Exploring Rucio Integration:** Currently, DMM operates as a standalone service, but there is strong potential for it to become a fully integrated component of Rucio in order to streamline operations and enhance the system’s functionality.

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.

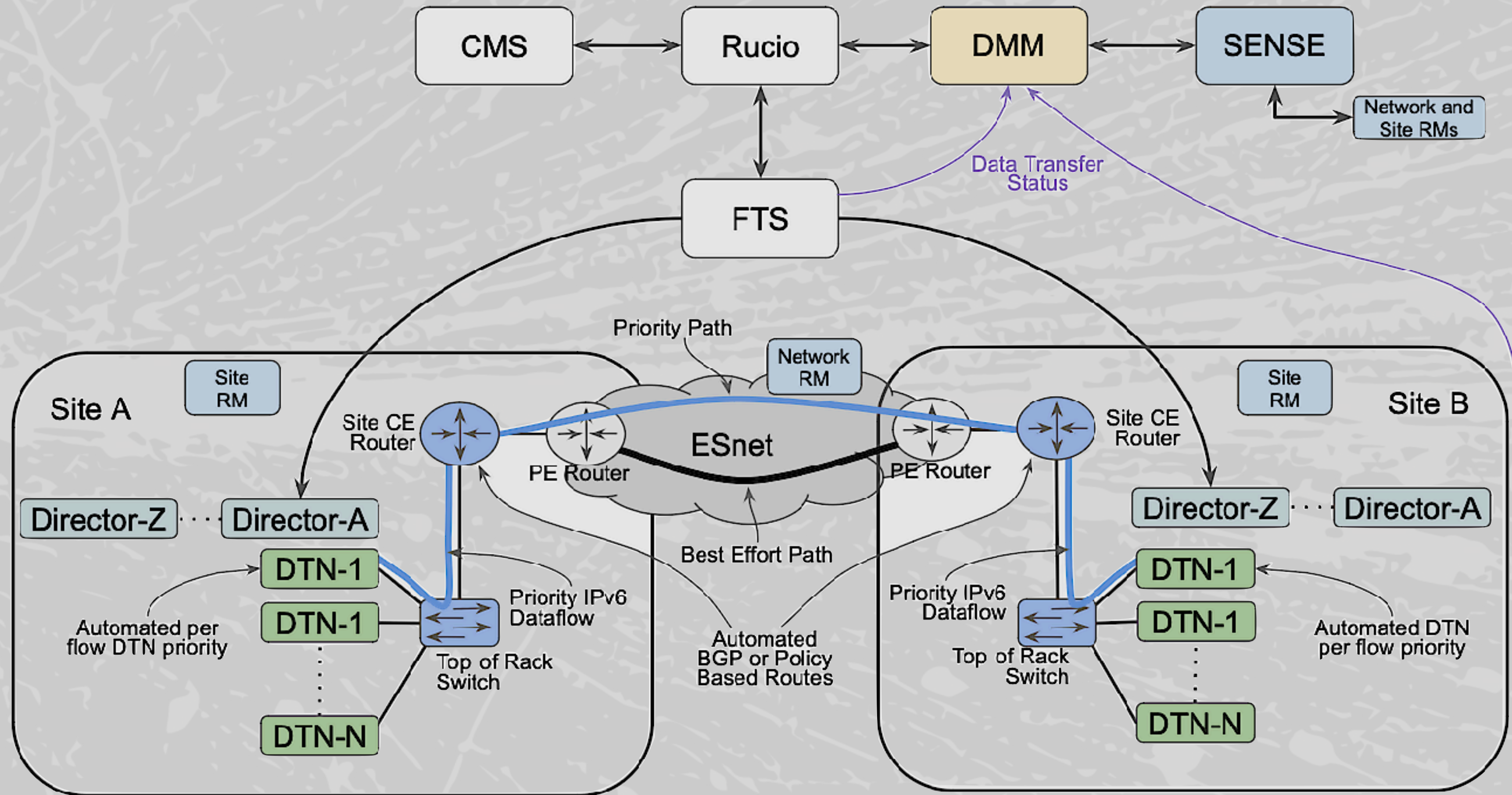
Backup

Rucio



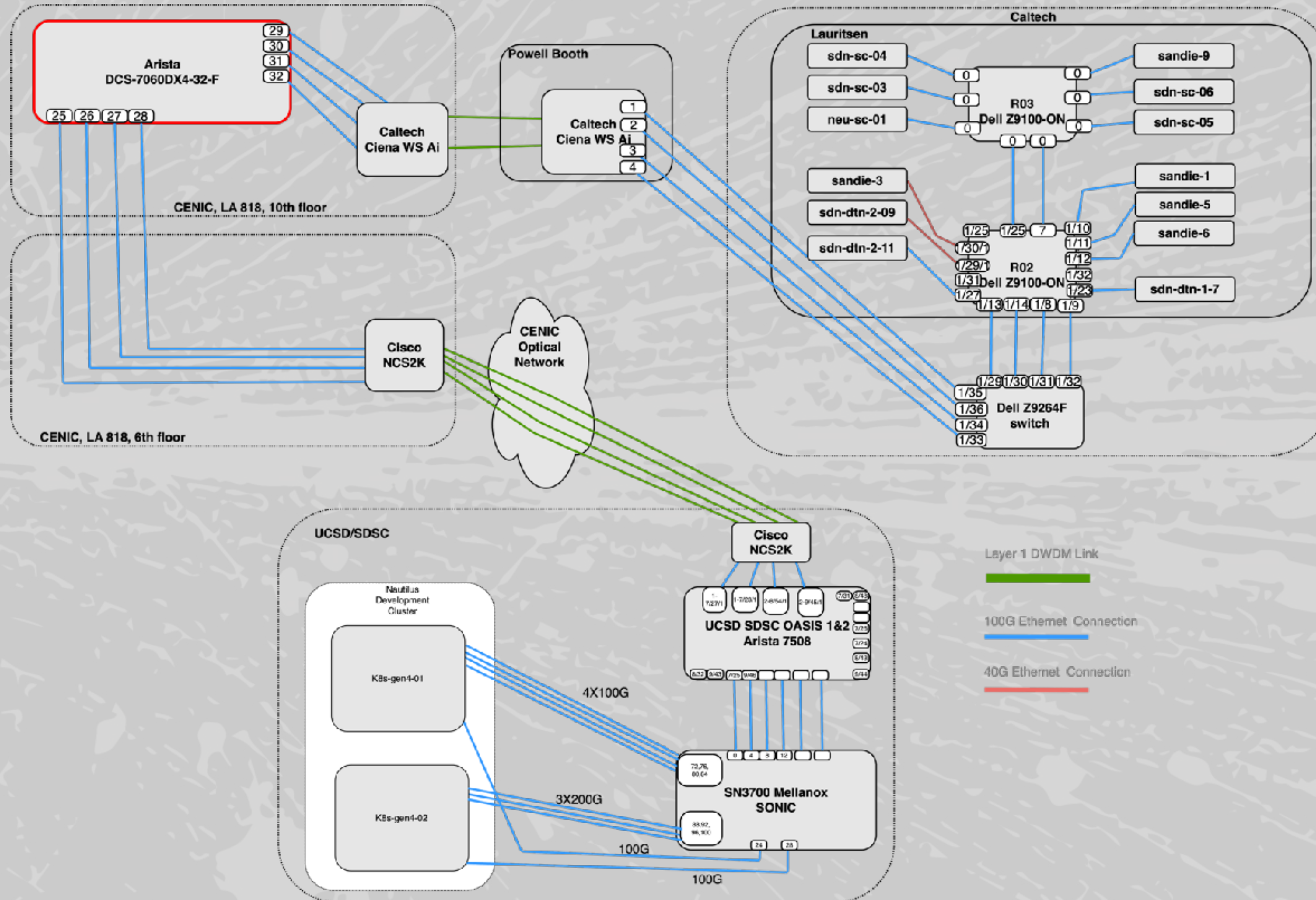
- Rucio is a data management system designed for scientific and HPC environments.
- Designed to handle petabytes of data generated by high-energy physics (HEP) experiments like the Large Hadron Collider (LHC).
- Used for data replication, placement, transfer, and deletion across the WLCG.

Overall Picture (more detailed)

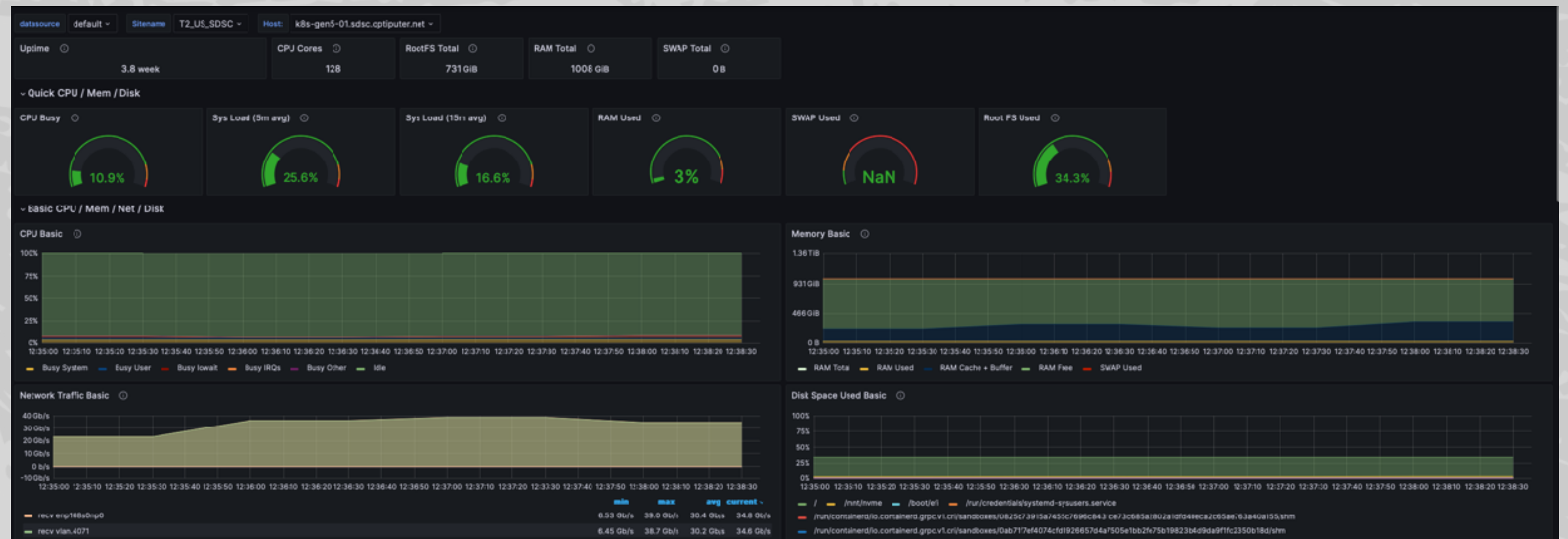
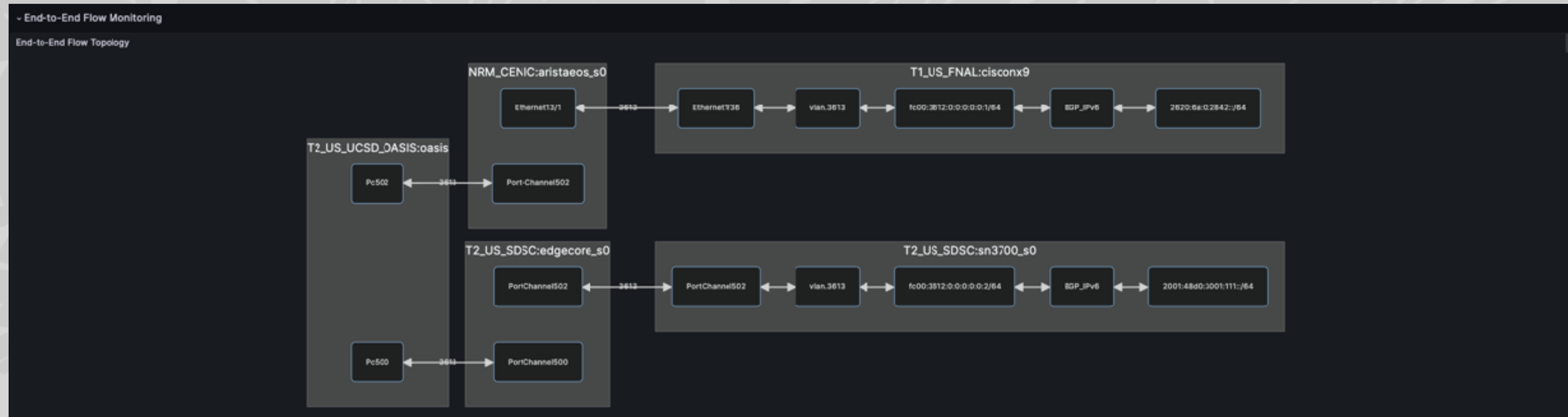


Rucio → DMM → SENSE → DMM → Rucio → FTS → XRootD

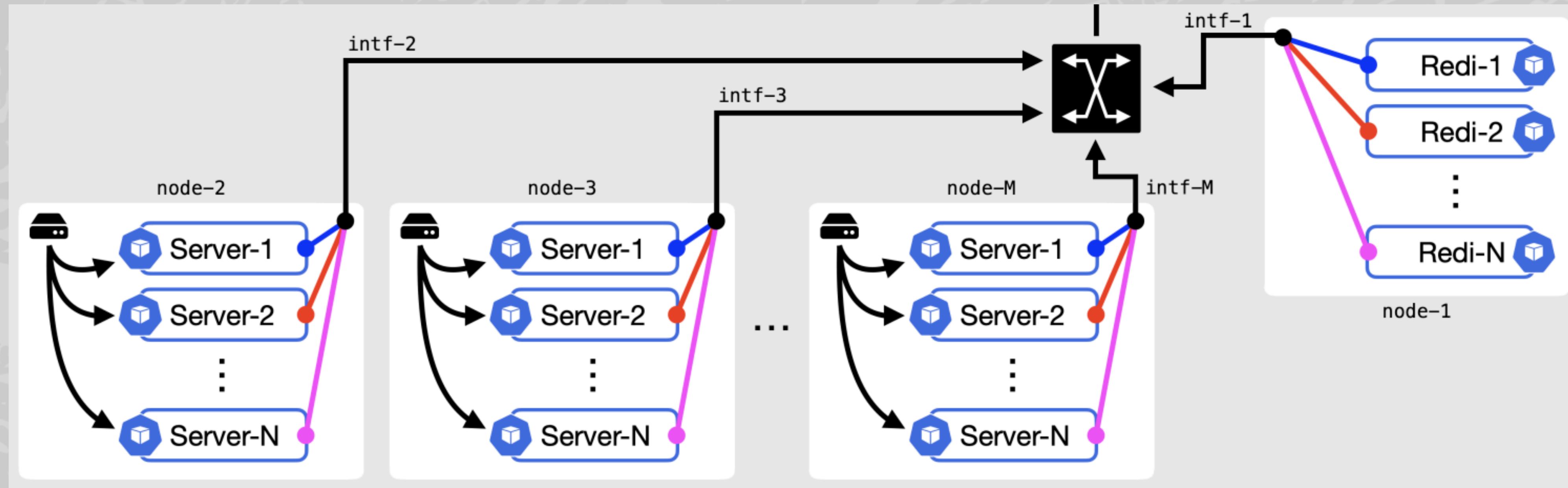
UCSD-Caltech Testbed



Monitoring Dashboards



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.