SENSE and Rucio/FTS/XRootD Interoperation

ESnet
Tom Lehman, Xi Yang, Chin Guok

UCSD
Frank Würthwein, Jonathan Guiang, Aashay Arora, Diego Davila, John Graham, Dima Mishin, Thomas Hutton, Igor Sfiligoi

Caltech
Harvey Newman, Justas Balcas

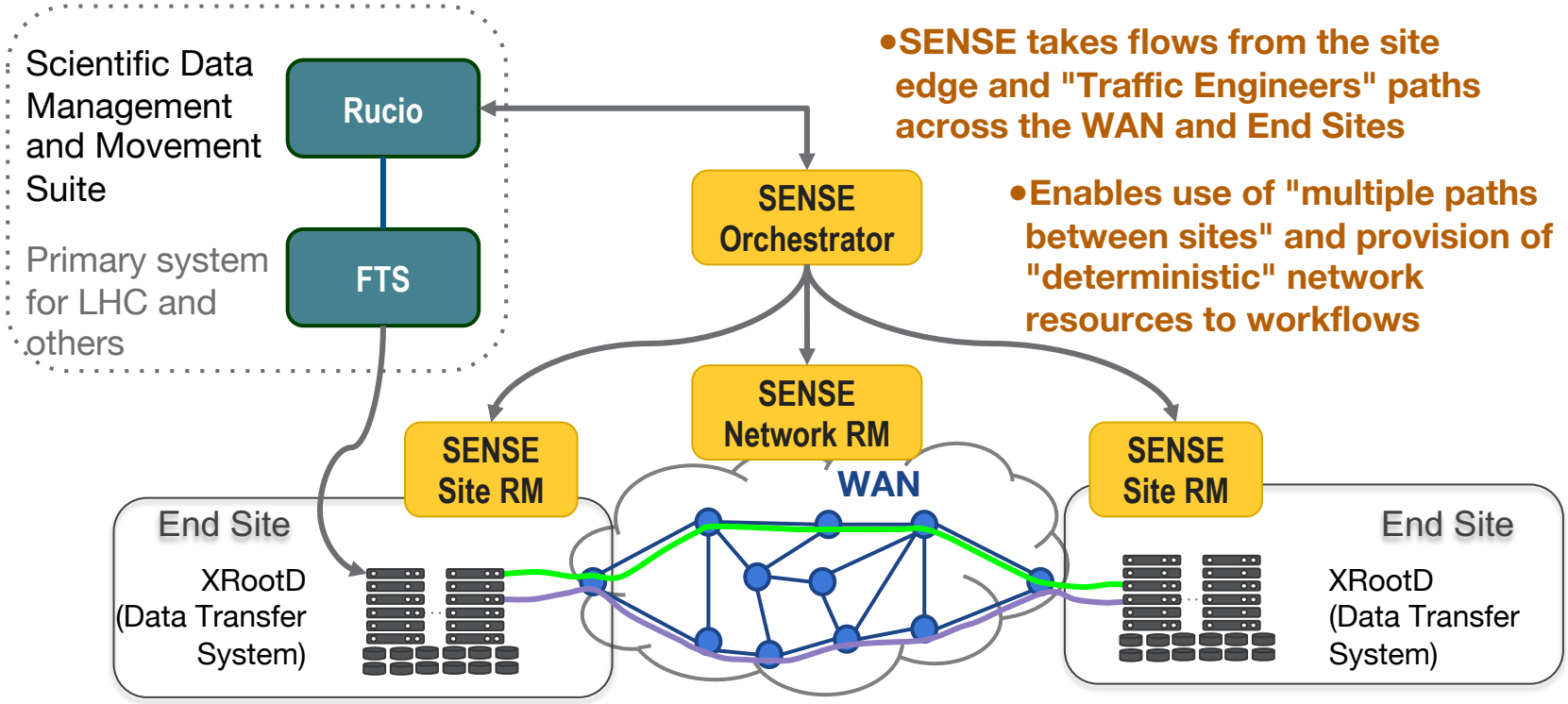
Data Challenge 24 Preparation
LHCOPN-LHCONE meeting #50
FZU, Prague CZ
April 18-19, 2023

ESnet

UCSD

SENSE and Rucio/FTS/XRootD Interoperation

- Rucio identifies groups of data flows (IPv6 subnets) which are "high priority"



- SENSE takes flows from the site edge and "Traffic Engineers" paths across the WAN and End Sites
- Enables use of "multiple paths between sites" and provision of "deterministic" network resources to workflows

Objectives

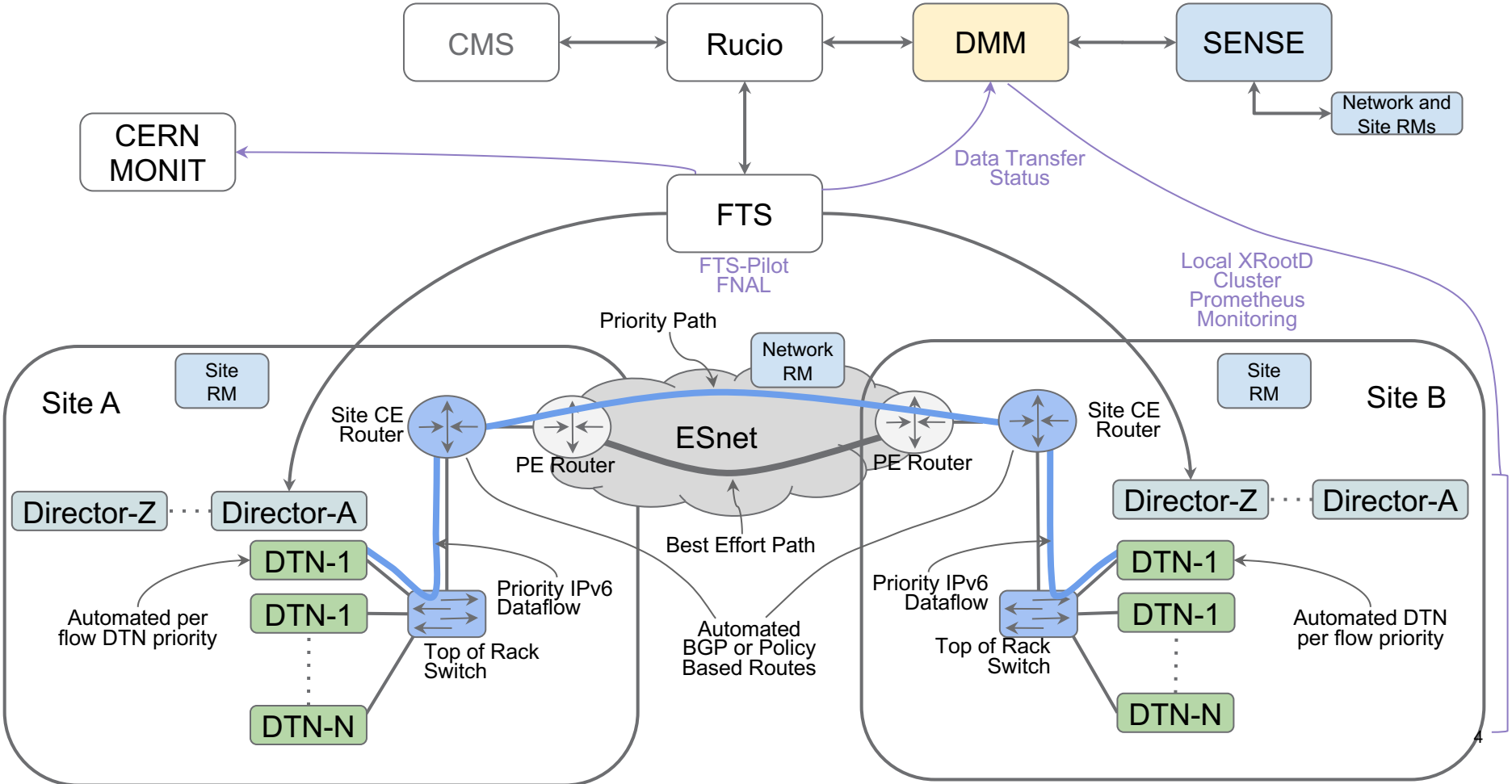
Overall objective is to develop a better way to manage CMS transfers

Accountability: determine where the issues are and develop a process to correct

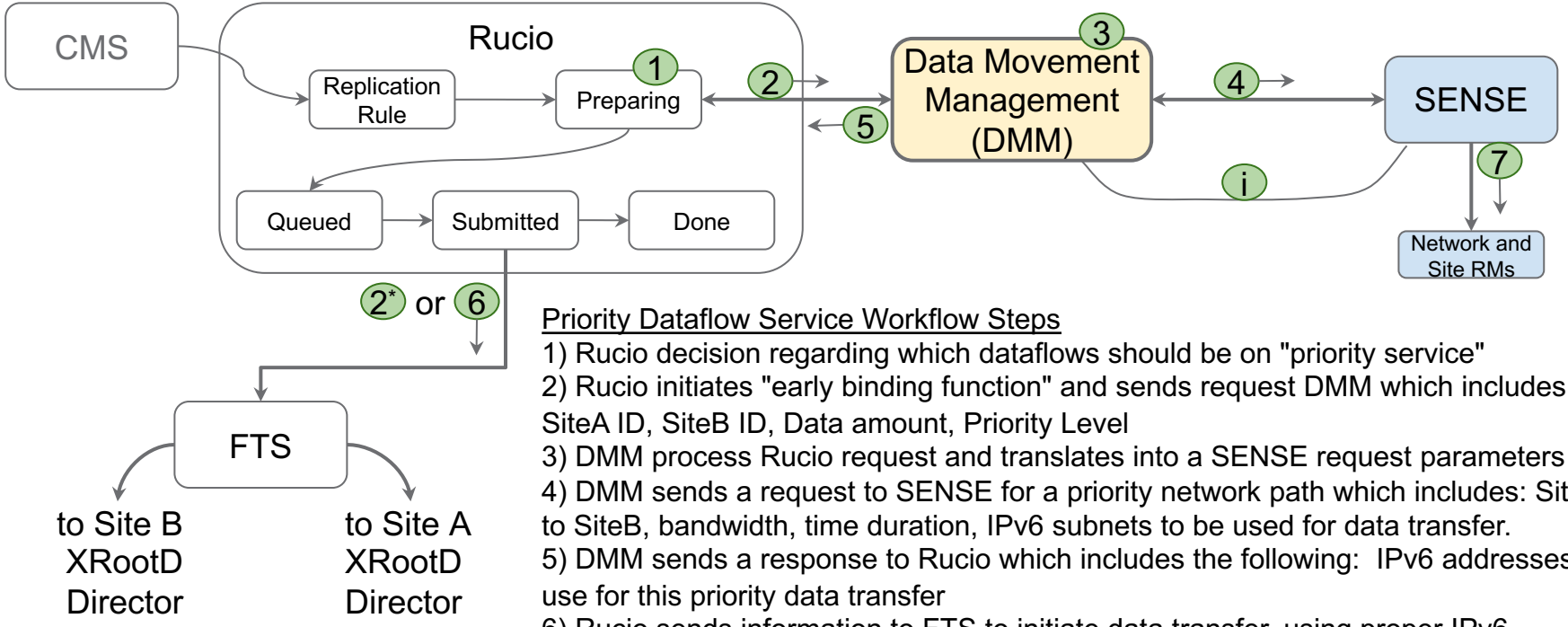
Focus on the largest flows (not ALL transfers)

Plan to use this system as part mini-Data Challenges in 2023 and official Data Challenge in 2024

SENSE Rucio/FTS/XRootD Workflow



Rucio, DMM, SENSE Workflow



Priority Dataflow Service Workflow Steps

1) Rucio decision regarding which dataflows should be on "priority service"
2) Rucio initiates "early binding function" and sends request DMM which includes: SiteA ID, SiteB ID, Data amount, Priority Level
3) DMM process Rucio request and translates into a SENSE request parameters
4) DMM sends a request to SENSE for a priority network path which includes: SiteA to SiteB, bandwidth, time duration, IPv6 subnets to be used for data transfer.
5) DMM sends a response to Rucio which includes the following: IPv6 addresses to use for this priority data transfer
6) Rucio sends information to FTS to initiate data transfer, using proper IPv6 addresses
7) SENSE sends request to Network and Site Resource Managers to instantiate priority network service

*Rucio to FTS and DMM interactions can be asynchronous

i) DMM to SENSE "discovery services" (one time at DMM startup)
This is the mechanism for DMM to discover information about sites which includes: sites available for service, IPv6 subnets available, site network connection speed

DMM - Data Movement Manager

- React to and process Rucio's "priority" data flow request
- Translate that into actionable information
  - Network provisioning (via SENSE)
  - Data Transfer initiation (identify the proper IPv6 subnet for Rucio-FTS-XRootD to use for a data flow)
- Longer term Focus: Designing effective policies for how "priority" should be established, who decides, what is the proper mix between priority services and best effort
  - Eventually DMM functions may be distributed between Rucio, SENSE, and/or other parts of the Domain Science Workflow.

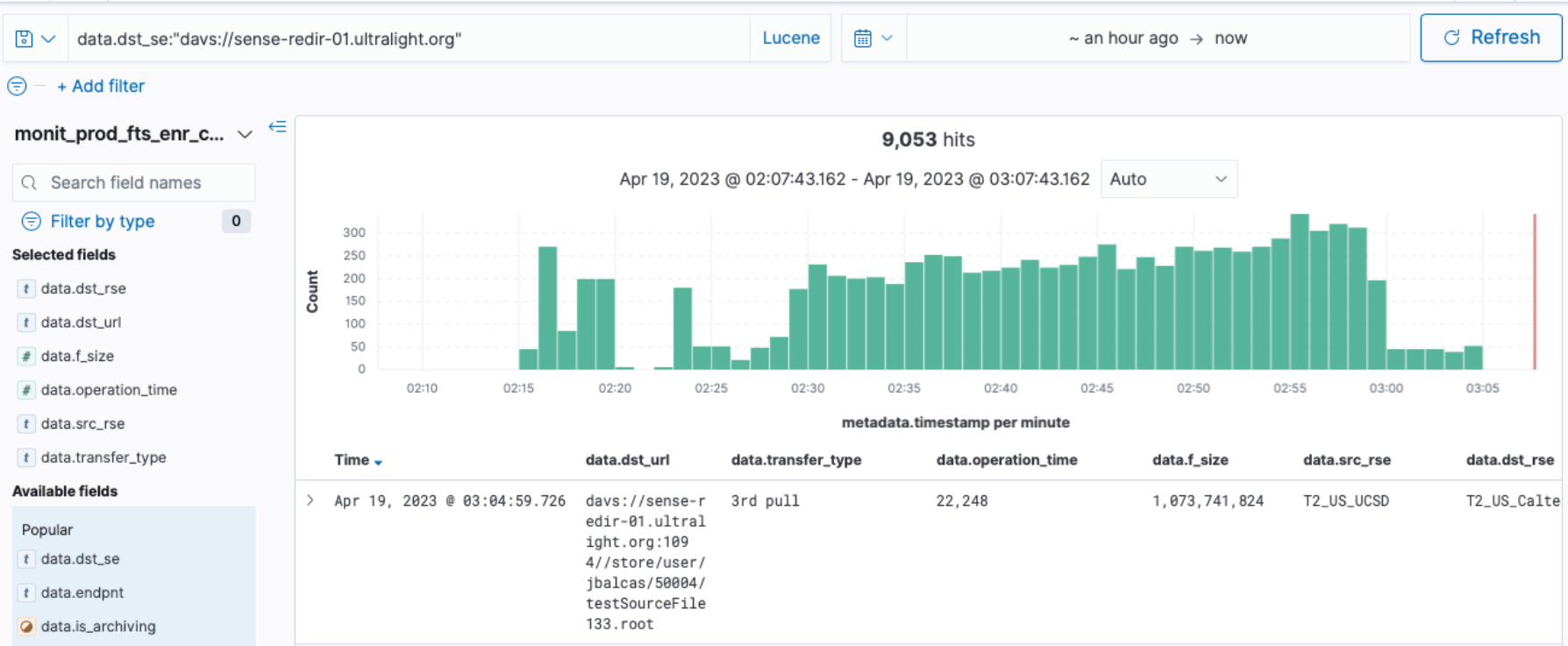
Rucio, DMM, SENSE Workflow

- A “priority” data flow is a flexible concept, and could be:
  - all data between Site A and Site B for a specific time period
  - all data between Site A and Site B on a specific IPv6 subnet
  - almost anything based on Site and IPv6/subnet parameters
- End-to-End Data Transfer monitoring
  - Performance evaluation (was the performance as expected?)
  - If not, analysis of why? (network?, congestion? where? end-system config/tuning? data movement protocols? other?)

End-to-End Performance Monitoring

- From local XRootD cluster Prometheus
  - Allocated vs achieved bandwidth
  - Total data transferred vs total transfer size
  - DMM summarizes when a transfer finishes
- FTS records in monIT
  - Data transfer performance from FTS/XRootD perspective
- Still working on the details of data collection, storage, and correlation/analysis

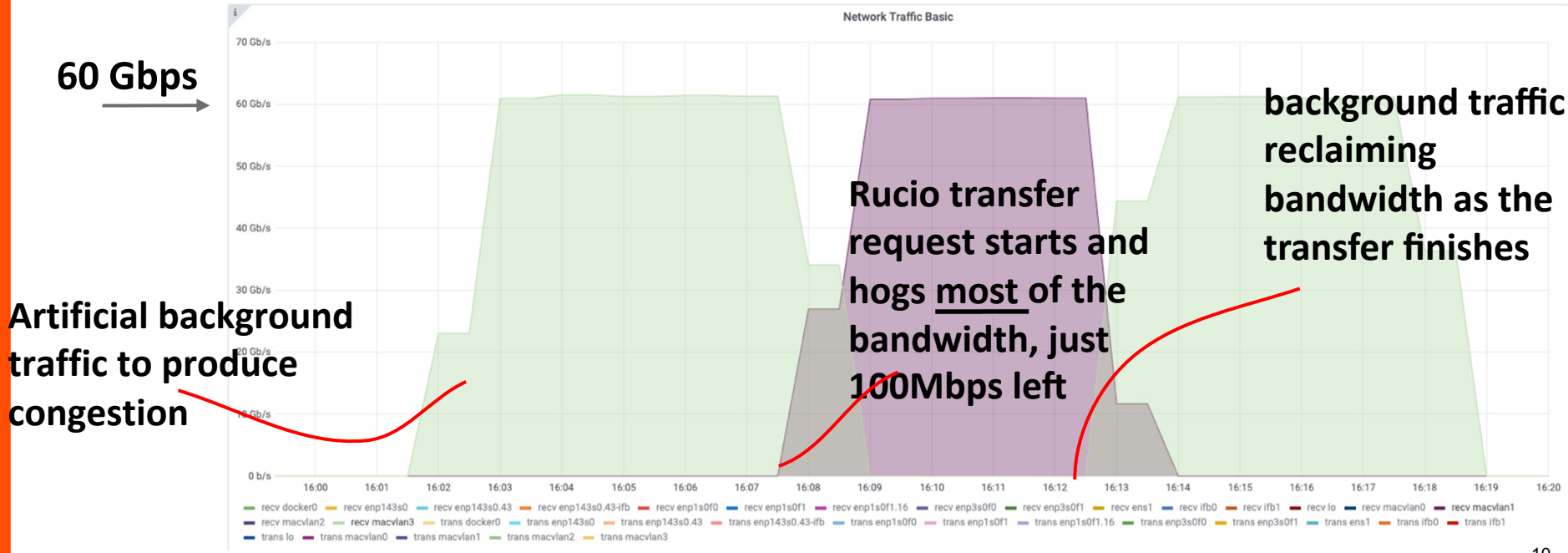
FTS Transfers via SENSE Path logged in MONIT (using CERN FTS3@Pilot Instance)



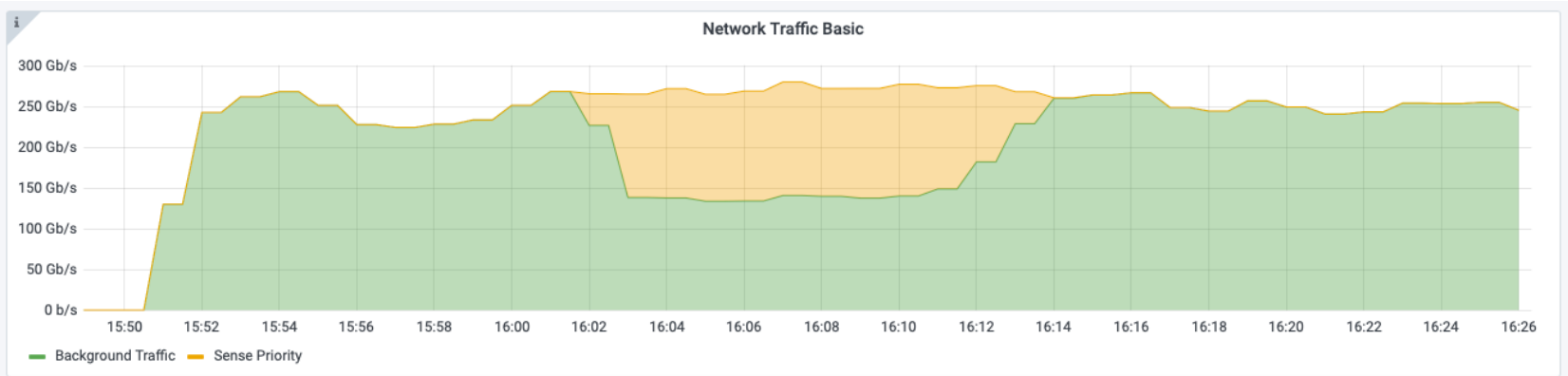
https://monit-opensearch.cern.ch/dashboards/goto/be5243def2962d4f7e222f0c0502d179?security_tenant=global
(CERN Account login needed)

Proof of Concept Testing

Currently working toward ~400 Gbps site-to-site. Only a few hosts needed for these rates. Working thru some technical issues in the areas of End System QoS, FTS/XRootD configurations.



UCSD to Caltech Testing at higher speeds

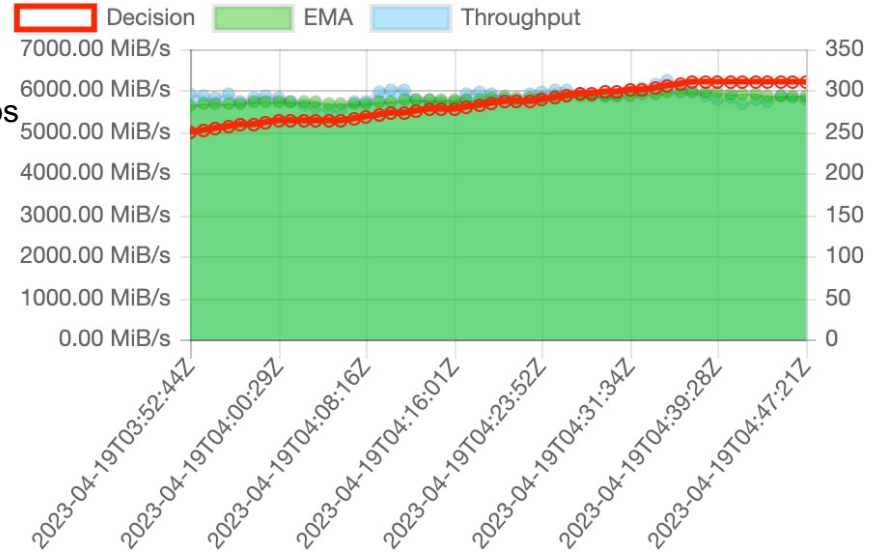


- Using FDT (Not FTS/XRootD)
- Green – background traffic, Yellow – Priority path requested via SENSE
- Total Capacity between UCSD-Caltech (300gbps). Background 200G, Priority 100G.
- Working thru some issues with Linux TC, Kubernetes/Multus Private NS Issues. Also evaluating use of BPF and Smart NICs for end-system options.

Higher Speed transfers using FTS/XRootD

- Once FTS Transfers are submitted, FTS Slowly increase number of active transfers (see red line).
- Due to this, XRootD endpoints do not get enough streams to reach >200gbps.
- Submission to FTS Ongoing, an you can see its progress on the link below
- Working to increase these transfer rates

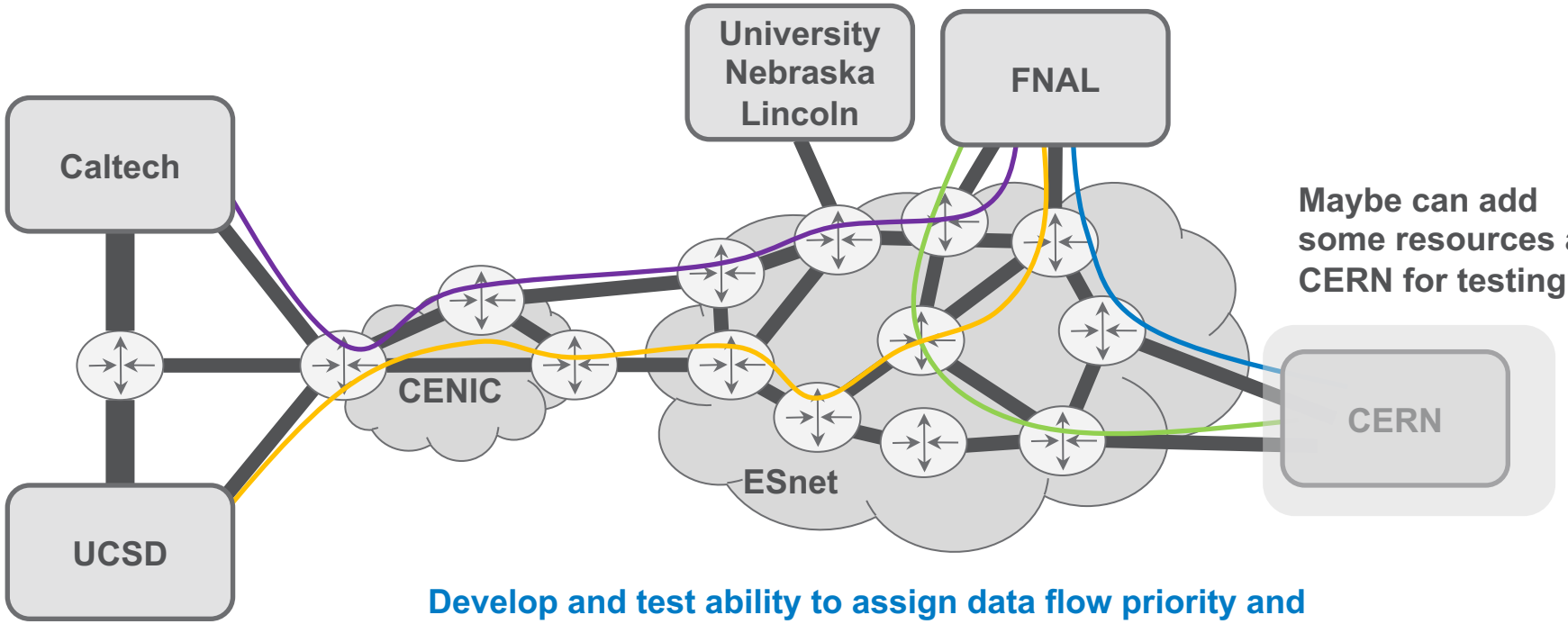
~48 Gbps



Source	Destination	VO	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel	1h)	Thr.
davs://xrootd-sense-ucsd-redirector.sdsc	davs://sense-redir-01.ultralight.or	cms	1284	190	–	–	–	14117	–	–	100.00 %	5223.57 MiB/s
			1284	190	0	0	0	14117	0	0	100.00 %	–

https://fts3-pilot.cern.ch:8449/fts3/ftsmon/#/?vo=&source_se=&dest_se=davs:%2F%2Fsense-redir-01.ultralight.org&time_window=1
(CERN Account login needed)

SENSE Rucio/FTS/XRootD Interoperation System Deployment

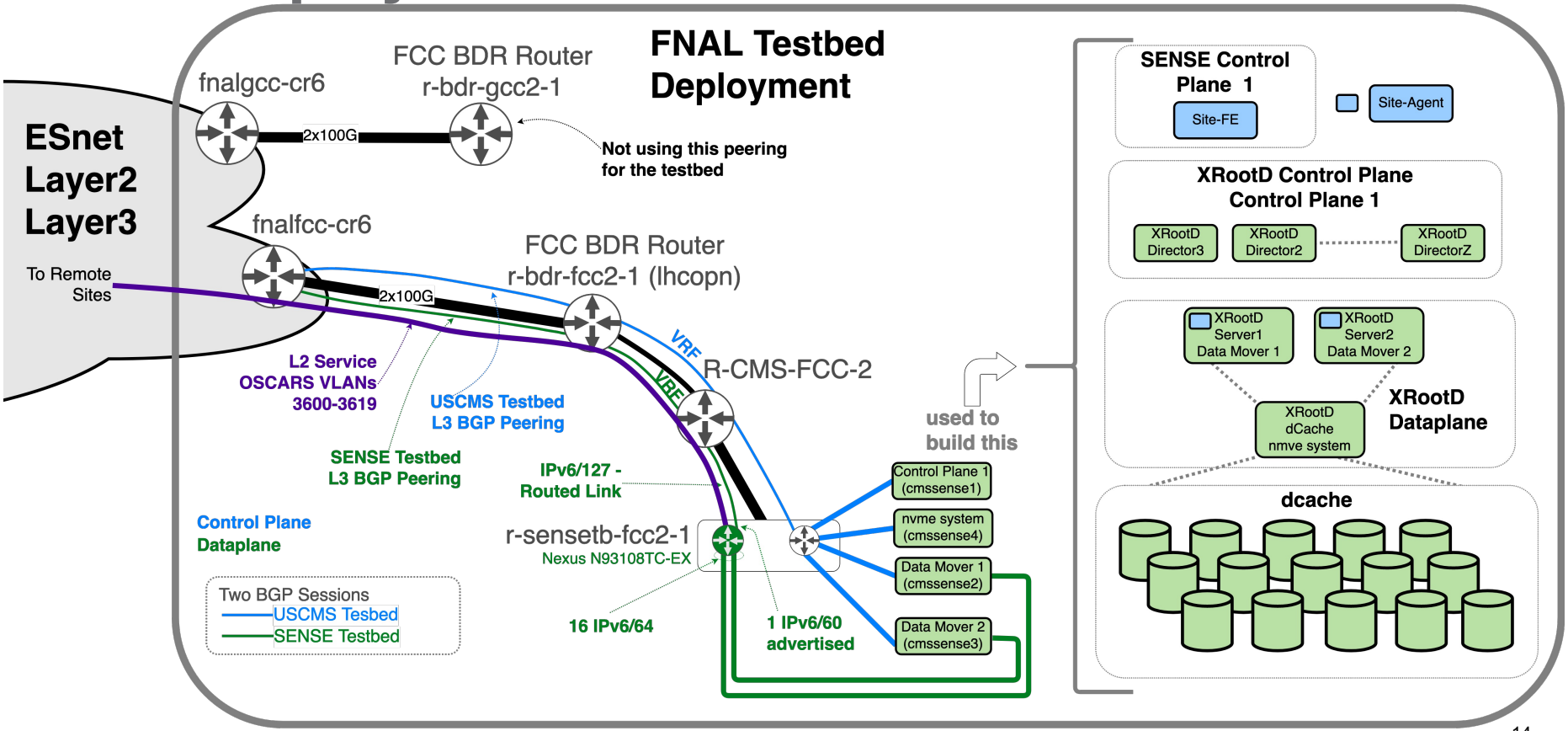


Maybe can add some resources at CERN for testing?

Develop and test ability to assign data flow priority and traffic engineer different end-to-end paths

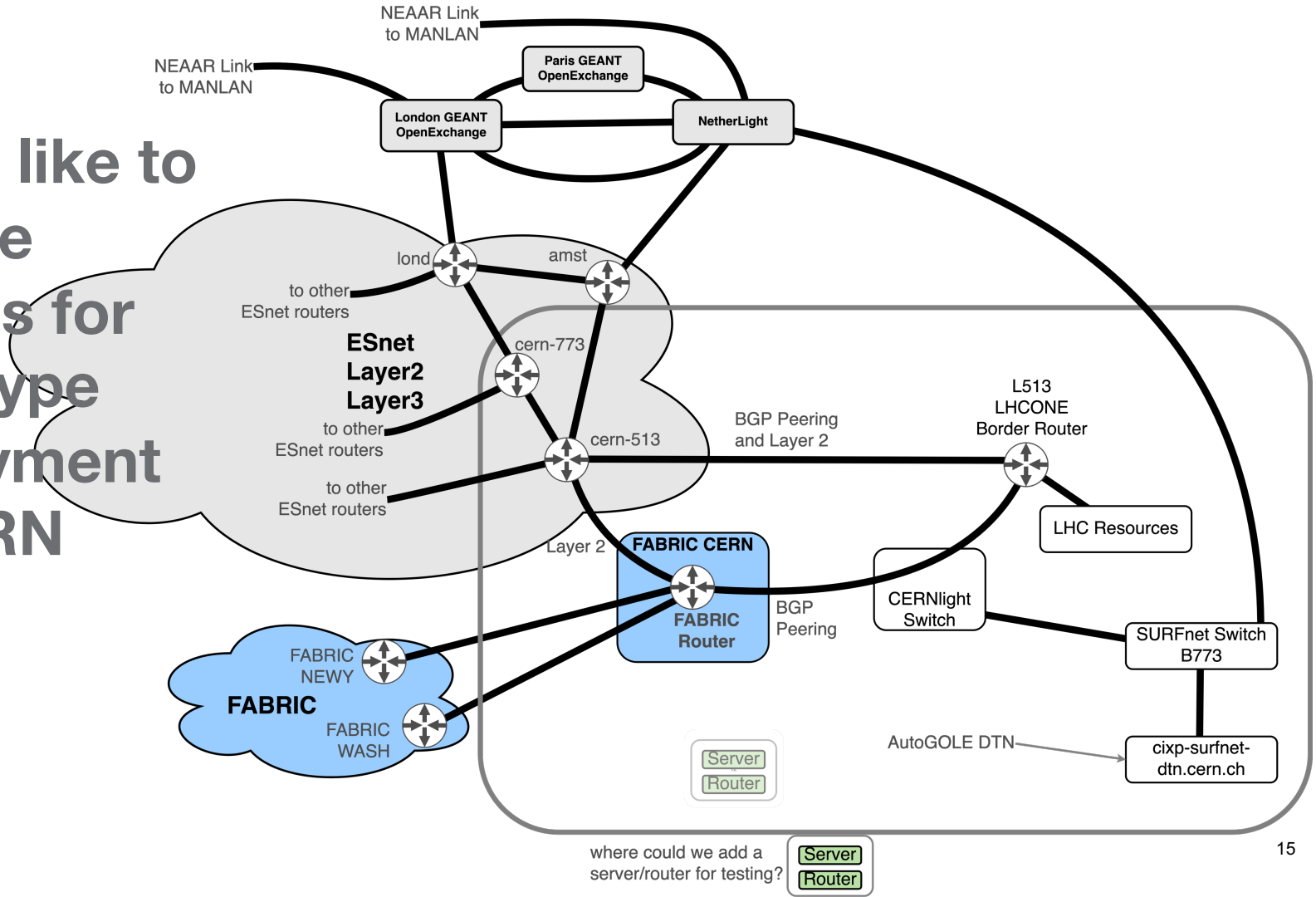
May add other sites: CERN, Vanderbilt, SPRACE

FNAL Deployment



Similar deployments at UCSD, Caltech

Would like to explore options for prototype deployment at CERN

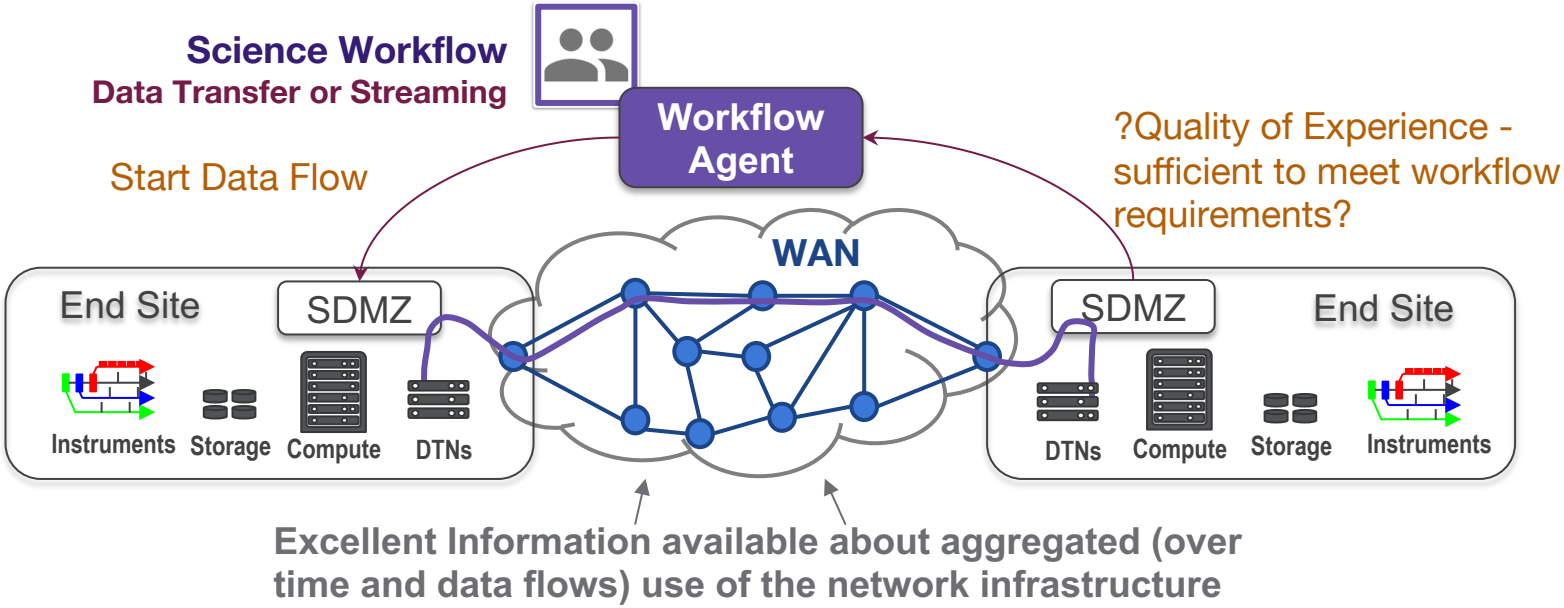


Objectives

- Provide mechanisms for domain science workflows and middleware (Rucio) to identify “priority” data flows
- Realtime integration of site data flows and wide area traffic engineering
  - in response to “priority” request
  - and/or just allow better overall network (link) utilization via traffic distribution/optimization
- Traffic engineering may include paths with QoS, or to traverse lightly loaded links

Enable Science Workflow and Network Interaction with Deterministic "Quality of Experience"

- No realtime per flow data available for planning or monitoring
- No "deterministic" network services available
- Start data flow, and hope for the best



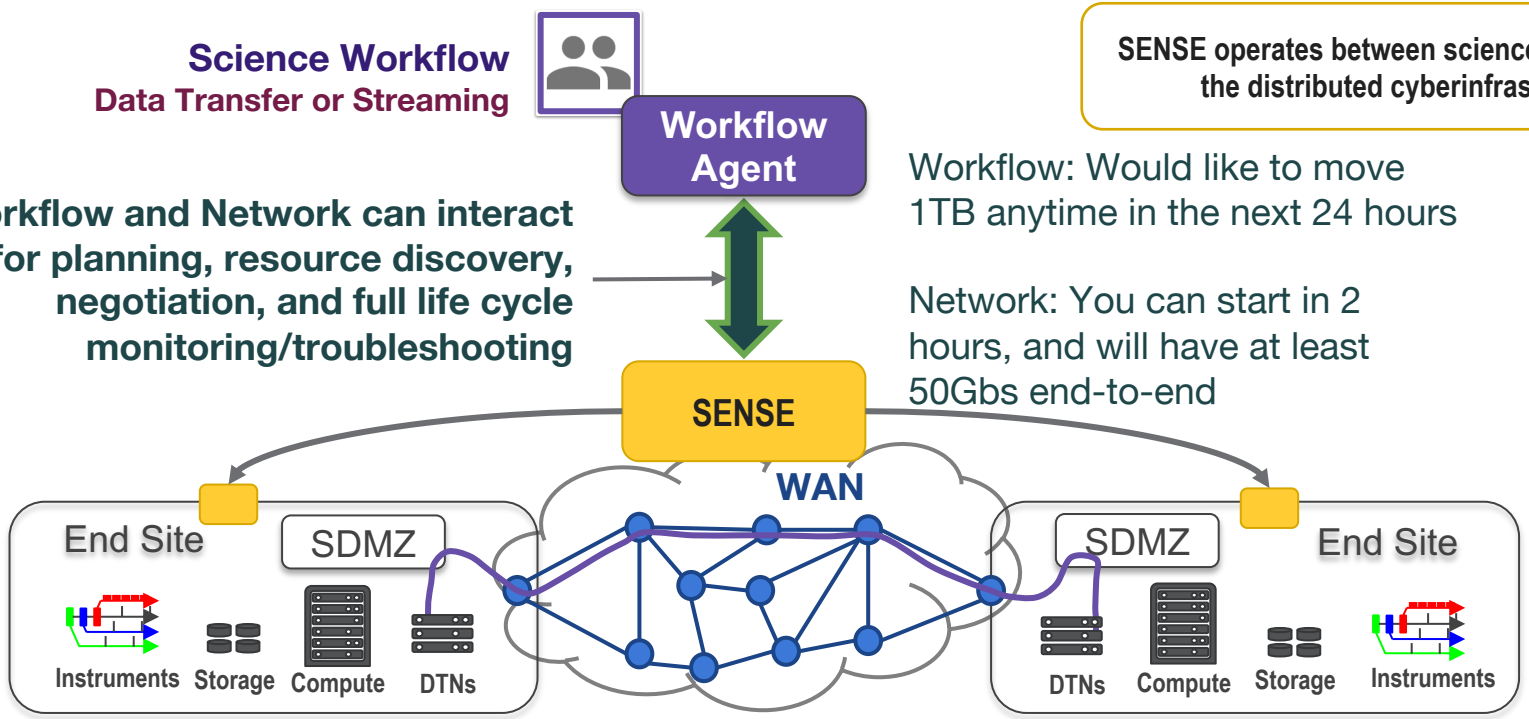
Elevate Network to First Class Resource

API driven Automation and Orchestration

Science Workflow
Data Transfer or Streaming

Workflow and Network can interact for planning, resource discovery, negotiation, and full life cycle monitoring/troubleshooting

SENSE operates between science workflow and the distributed cyberinfrastructure



- Allows workflows to identify data flows which are higher priority
- Allows the network to traffic engineer to fully utilize all network paths

Key Themes

- Today, science workflows view the network as an opaque infrastructure - inject data and hope for an acceptable Quality of Experience
- We should allow workflow agents to interact with the network - ask questions, see what is possible, get flow specific data and resources
- Science workflow planning should be able to include the networks as a first-class resource (alongside compute, storage, instruments)
- This requires collaborative cross-discipline teams for workflow co-design
- The same mechanisms that allow the above can also be used by individual networks to distribute traffic more efficiently across entire infrastructure

Thanks