LHC Workflow R&D/Testing using AutoGOLE, SENSE, and FABRIC

ESnet, UCSD, Caltech, FNAL, FABRIC

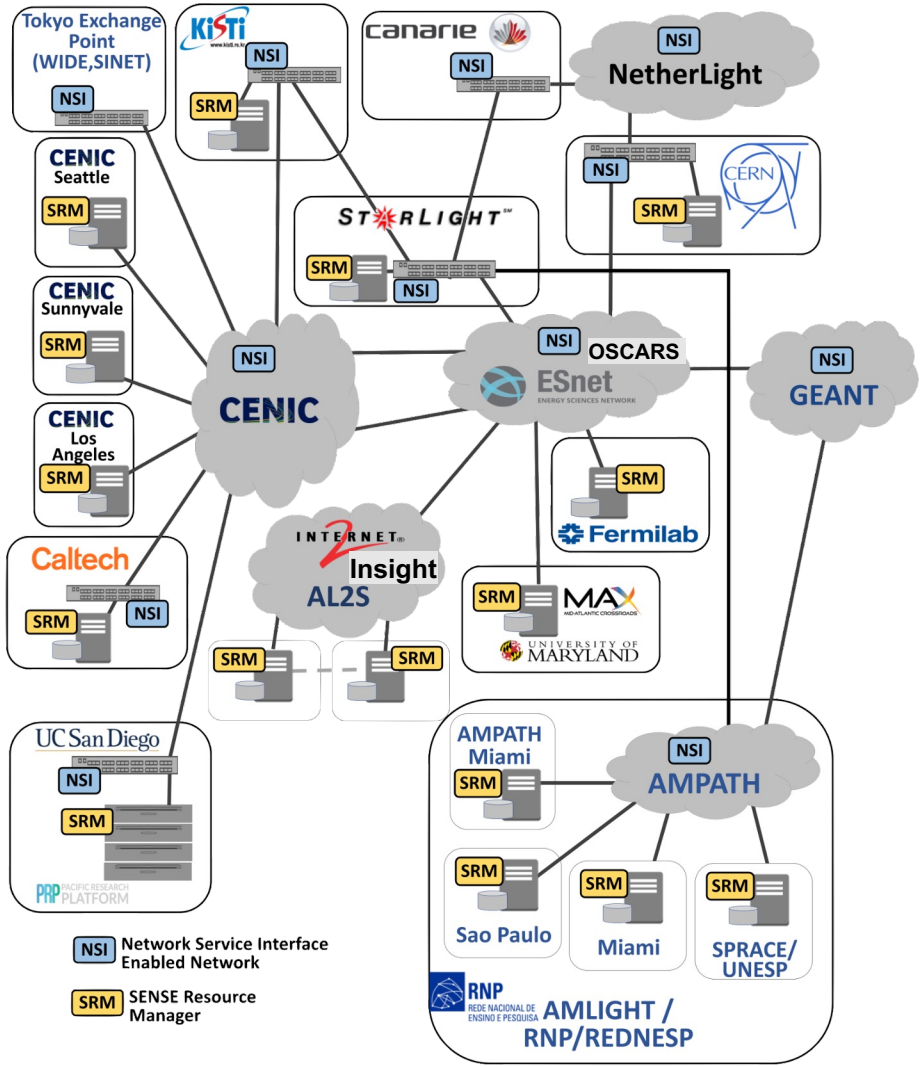
LHCOPN-LHCONE meeting #52
INFN Catania Italy
April 9-11, 2024

Outline

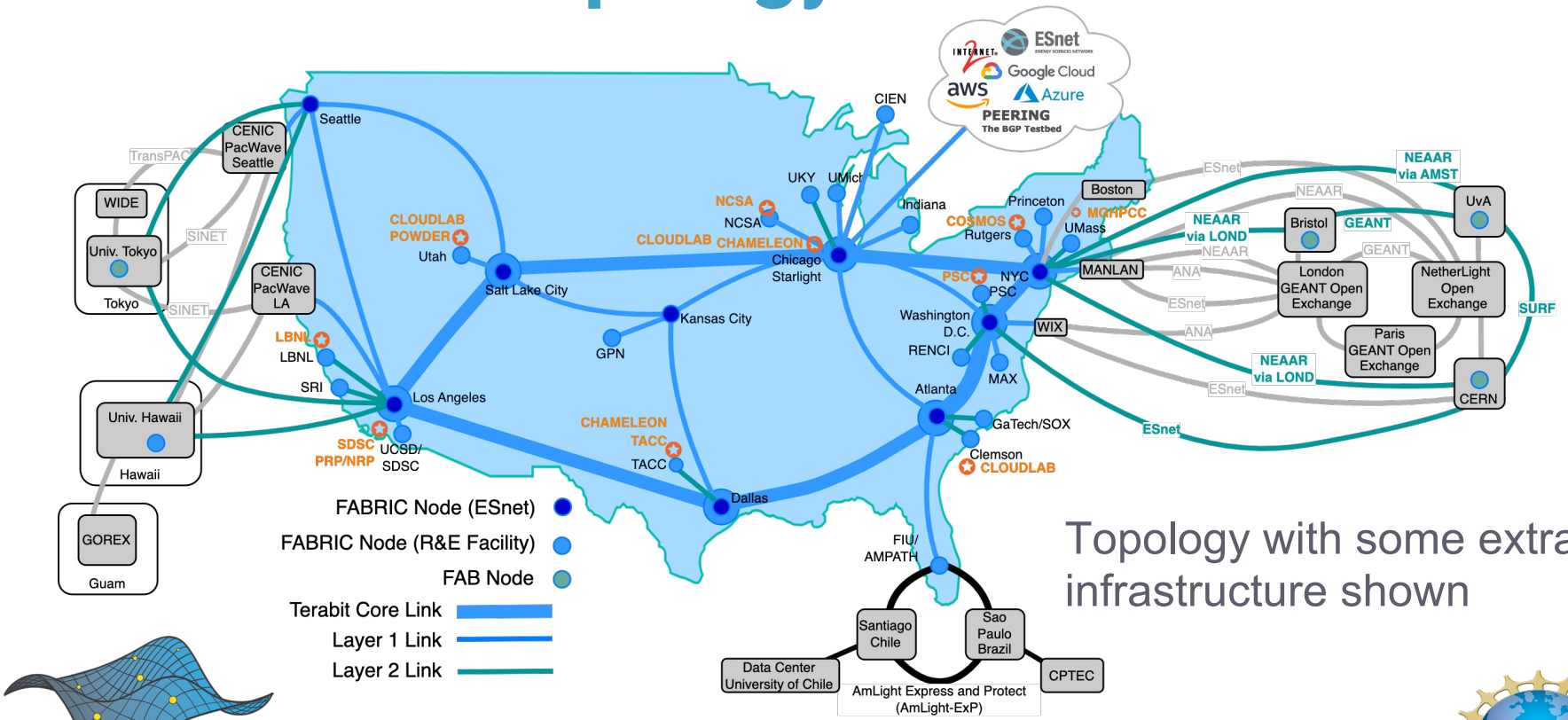
- R&E Infrastructure for Testing, Development, Research
  - AutoGOLE
  - SENSE
  - FABRIC
- Rucio/FTS/XRootD use of SENSE Services for CMS workflows
- XRootD performance Testing
- Future Research/Development/Test Areas

SENSE/AutoGole

- AutoGOLE, NSI, and SENSE working together provide the mechanisms for complete end-to-end services which includes the network and the attached End Systems (DTNs).
- Possible Provisioning Objectives: Layer 2 isolation, Guaranteed QoS, Managing Flows Path/Link usage



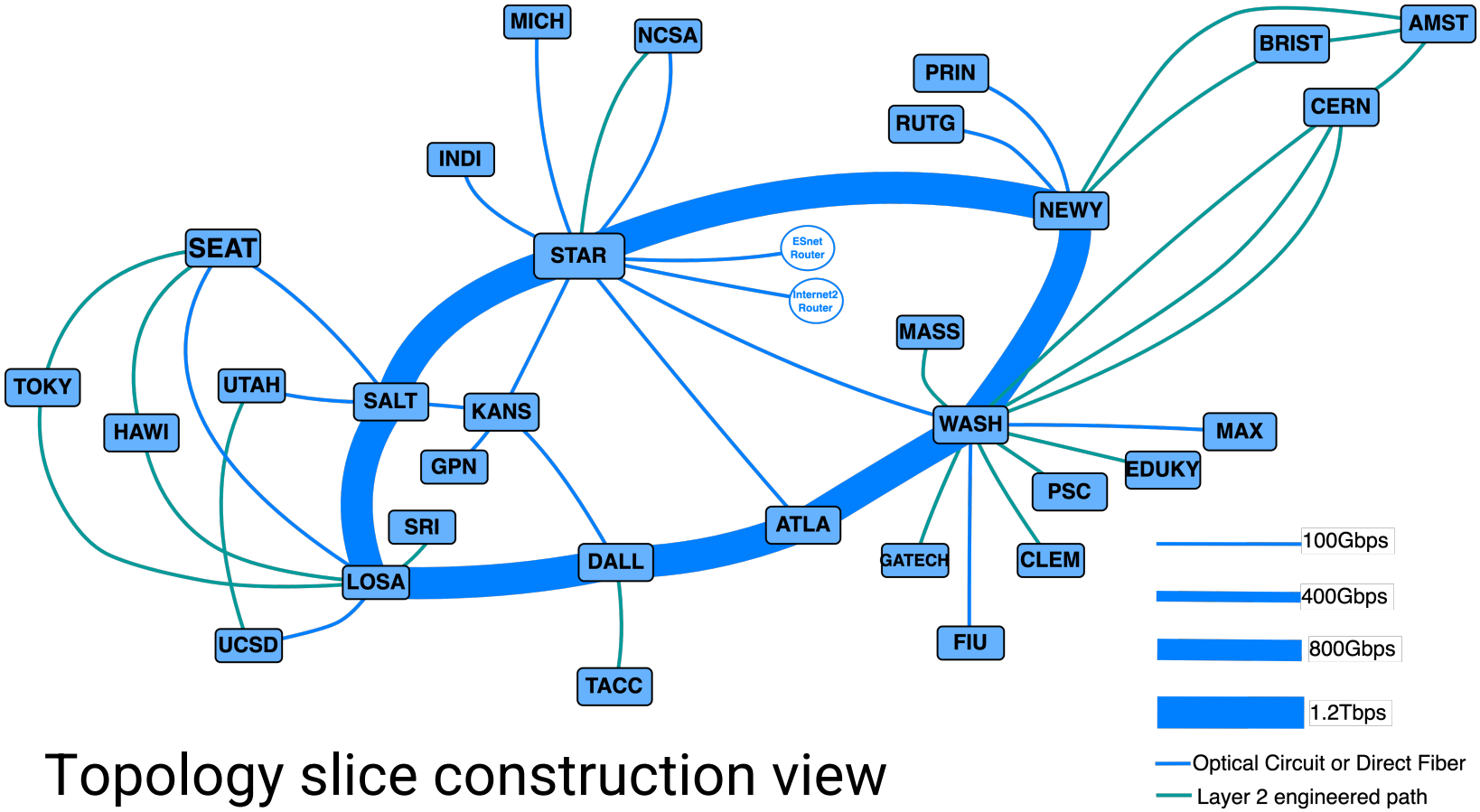
FABRIC Topology



Topology with some extra infrastructure shown

FABRIC

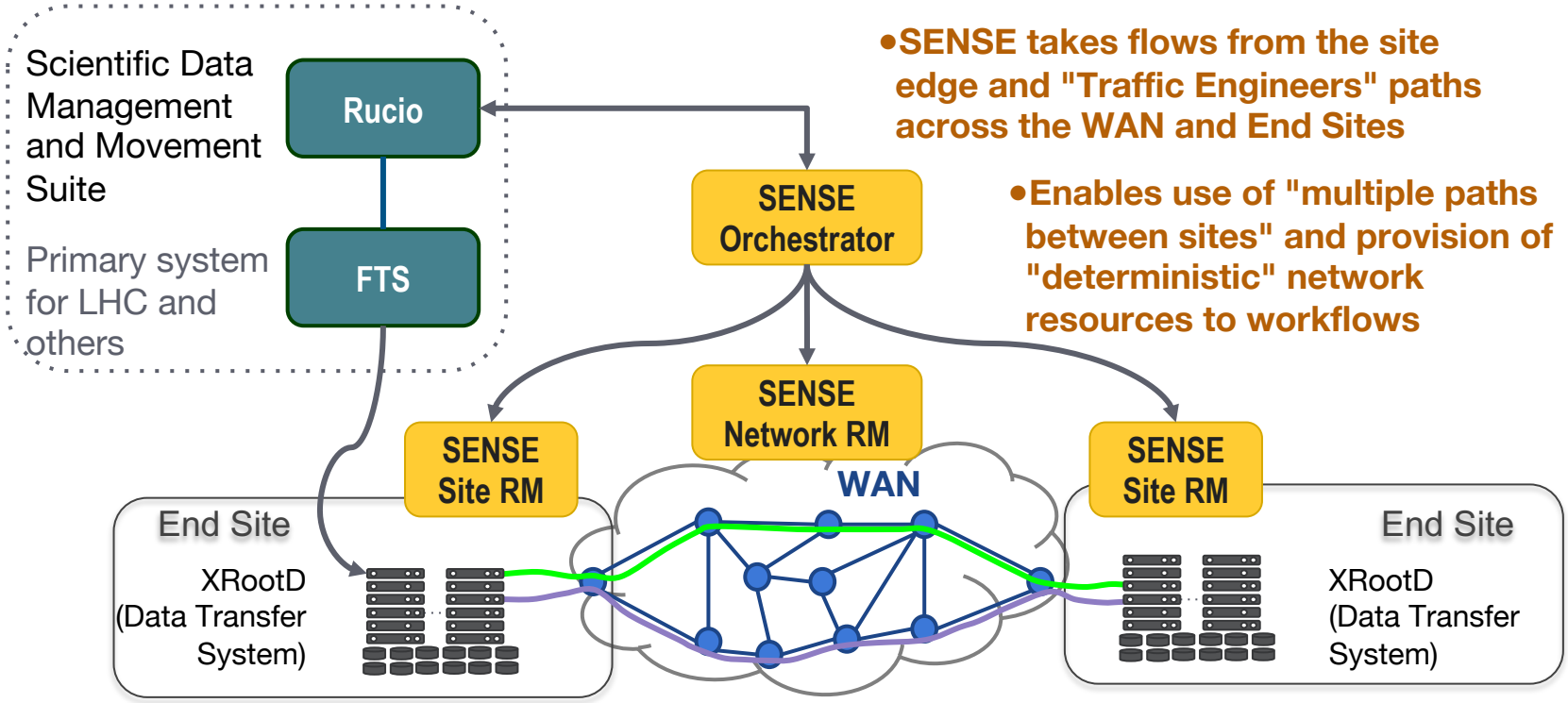
FABRIC Topology



Topology slice construction view

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

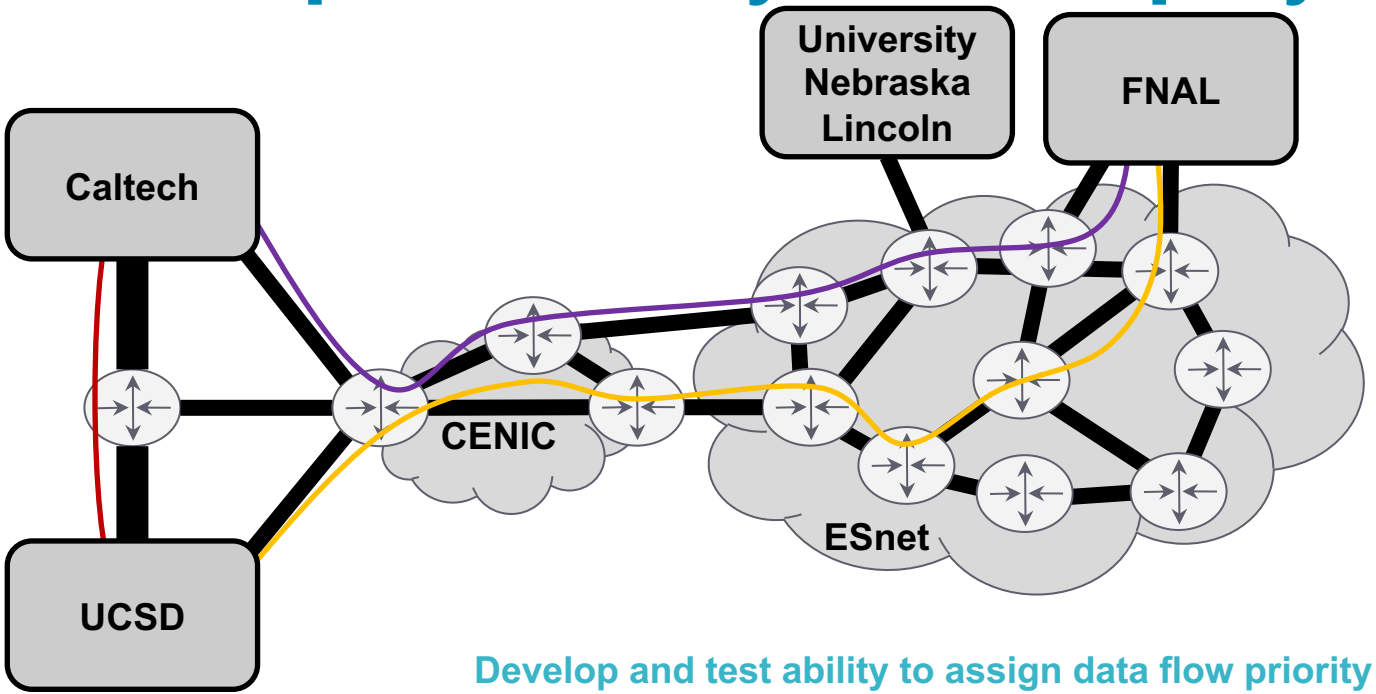
Objectives

Overall objective is to develop an improved way to manage CMS transfers

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

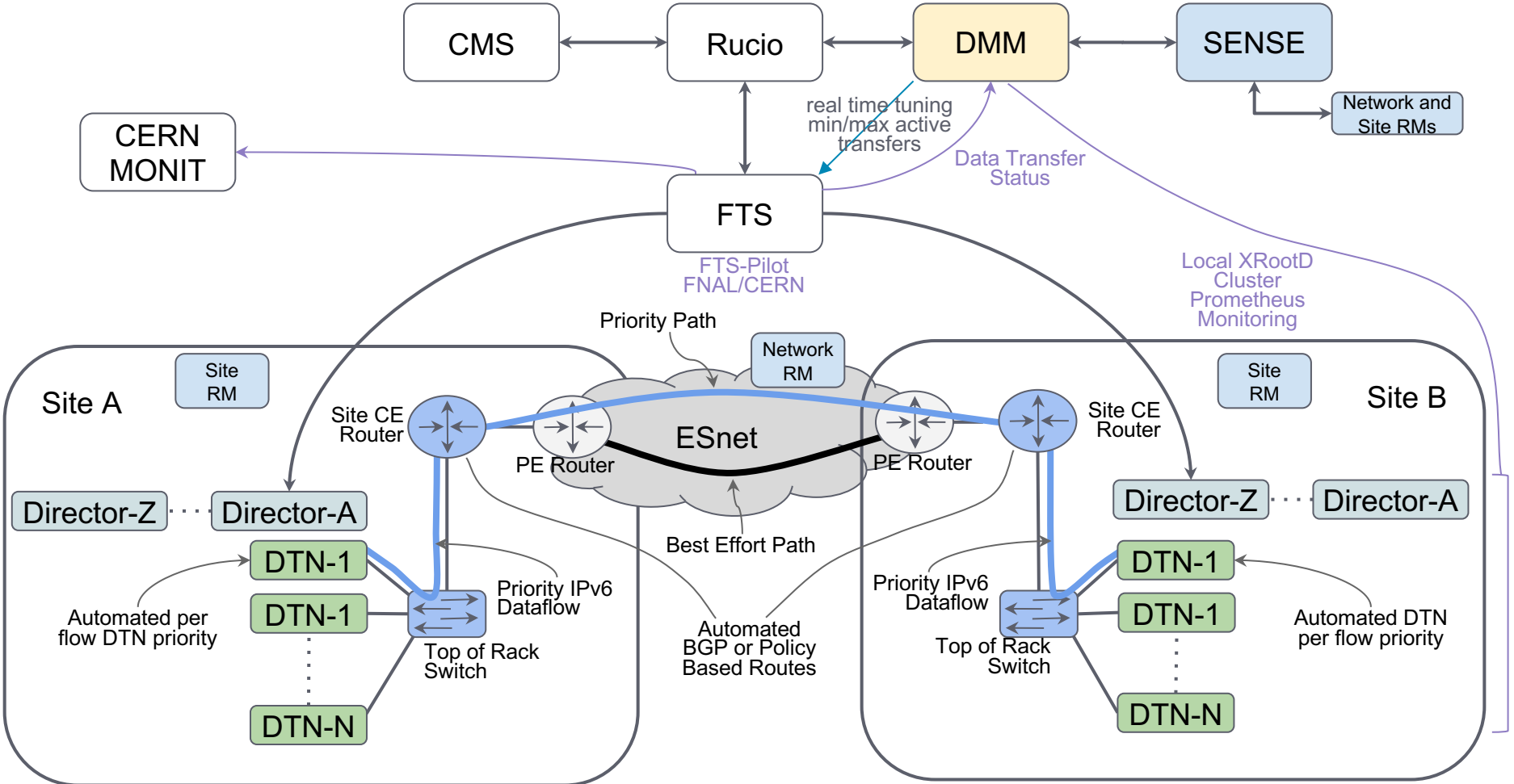
Focus on the largest flows (not ALL transfers)

SENSE Rucio/FTS/XRootD Interoperation System Deployment

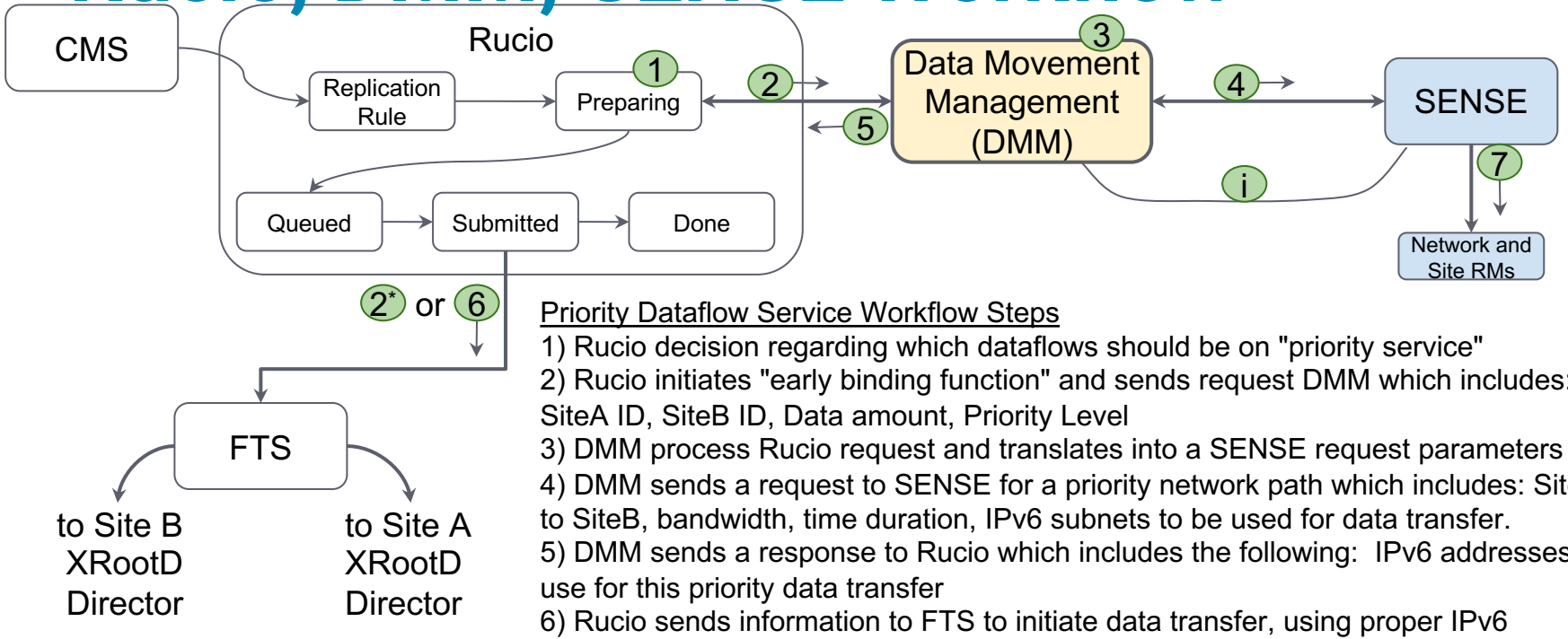


Develop and test ability to assign data flow priority and traffic engineer different end-to-end paths

SENSE Rucio/FTS/XRootD Workflow



Rucio, DMM, SENSE Workflow



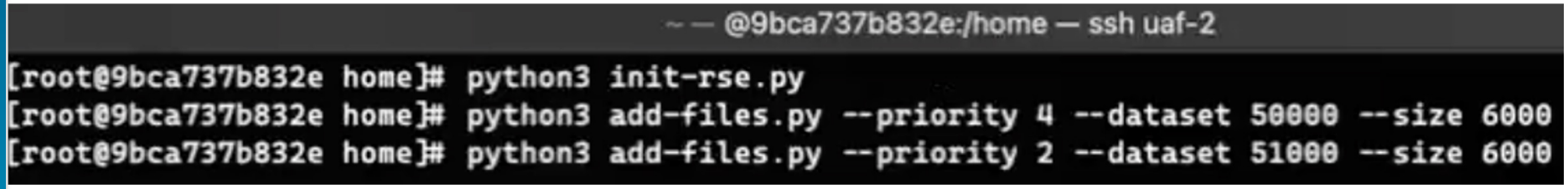
Priority Dataflow Service Workflow Steps

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

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

*Rucio to FTS and DMM interactions can be asynchronous

Rucio Replication Rules with Priorities



- Rucio knows all about file locations, and what data needs to be moved between which sites, and can define a "priority" for the data transfers
- DMM translates this Rucio defined "priority" into specific network service requests to SENSE.
- In this case, for a 100Gbps link, the priority 4 data transfer will get ~67 Gbps, and the priority 2 data transfer will get ~33 Gbps
- These priorities and allocations can be modified as needed during the lifecycle of existing transfers or in response to new transfer requests

Rucio Replication Rules with Priorities

1) Priority 4 data transfer starts first and get ~100 Gbps

3) Priority 4 transfer now gets ~67 Gbps

4) Priority 4 transfer ends and Priority 2 continues at ~33 Gbps

2) Priority 2 transfer starts and gets ~33 Gbps

DMM to FTS Tunning

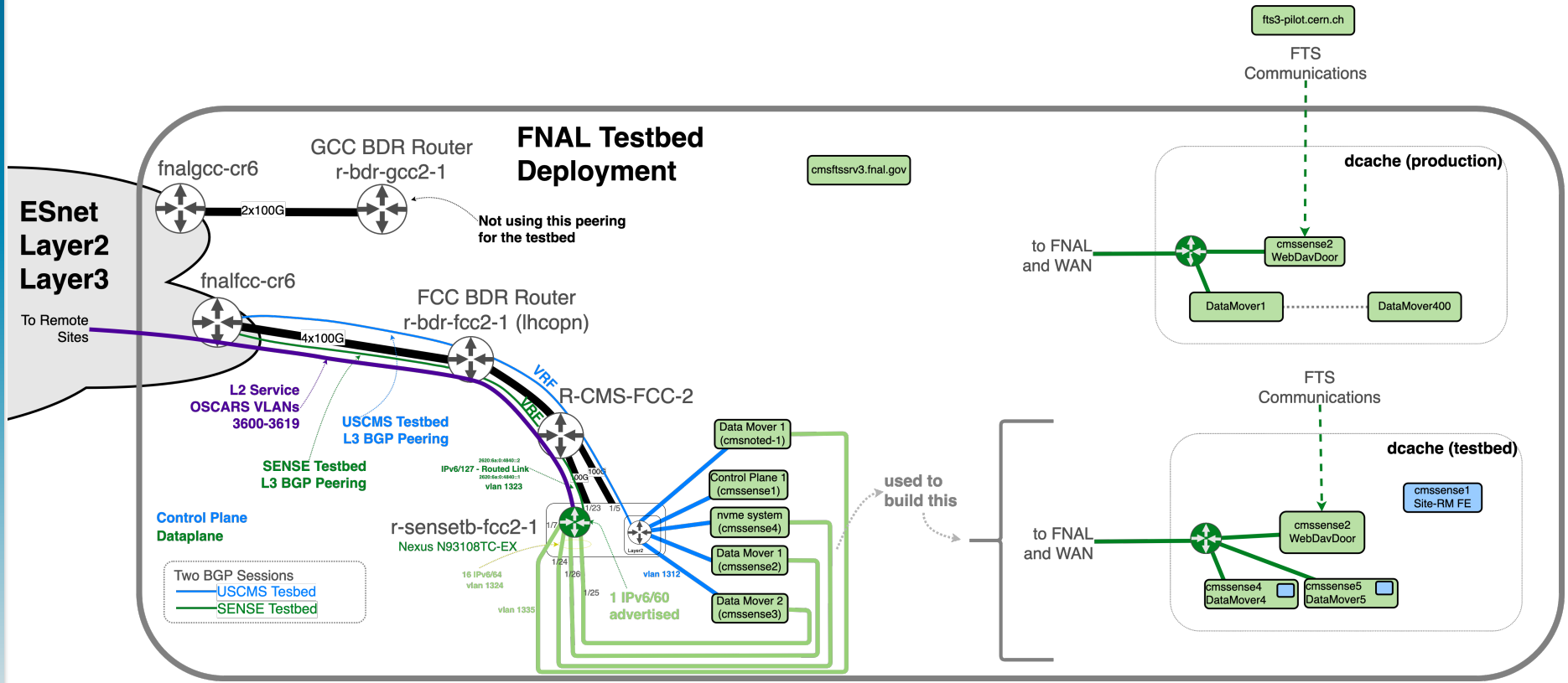
Link configuration

First | Previous | 1 | 2 | Next | Last

Symbolic name	Source	Destination	Min. Active	Max. Active
✚ davs://cmssense4-origin-2841-1.fnal.gov-davs://sense-redir-01.ultralight.org	davs://cmssense4-origin-2841-1.fnal.gov	davs://sense-redir-01.ultralight.org	20	20
✚ davs://cmssense4-origin-2841-1.fnal.gov-davs://sense-redir-02.ultralight.org	davs://cmssense4-origin-2841-1.fnal.gov	davs://sense-redir-02.ultralight.org	1600	1600

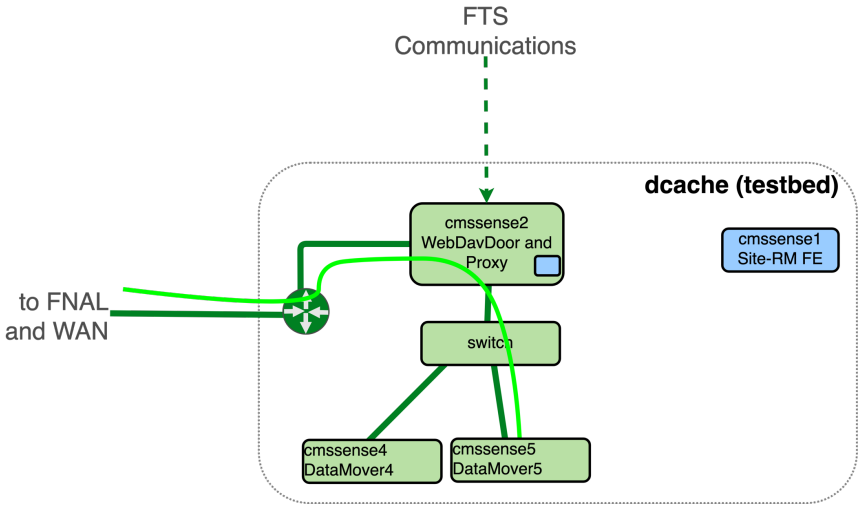
Need to tune FTS so concurrent transfer config matches network service provisioned based on Rucio defined priority

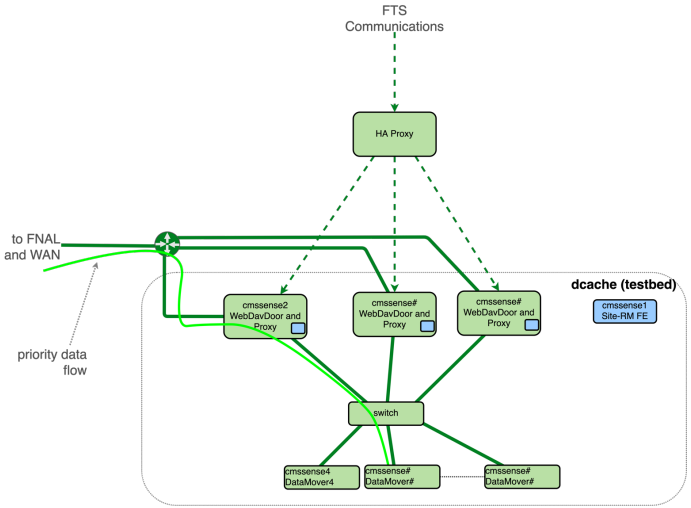
FNAL Testbed



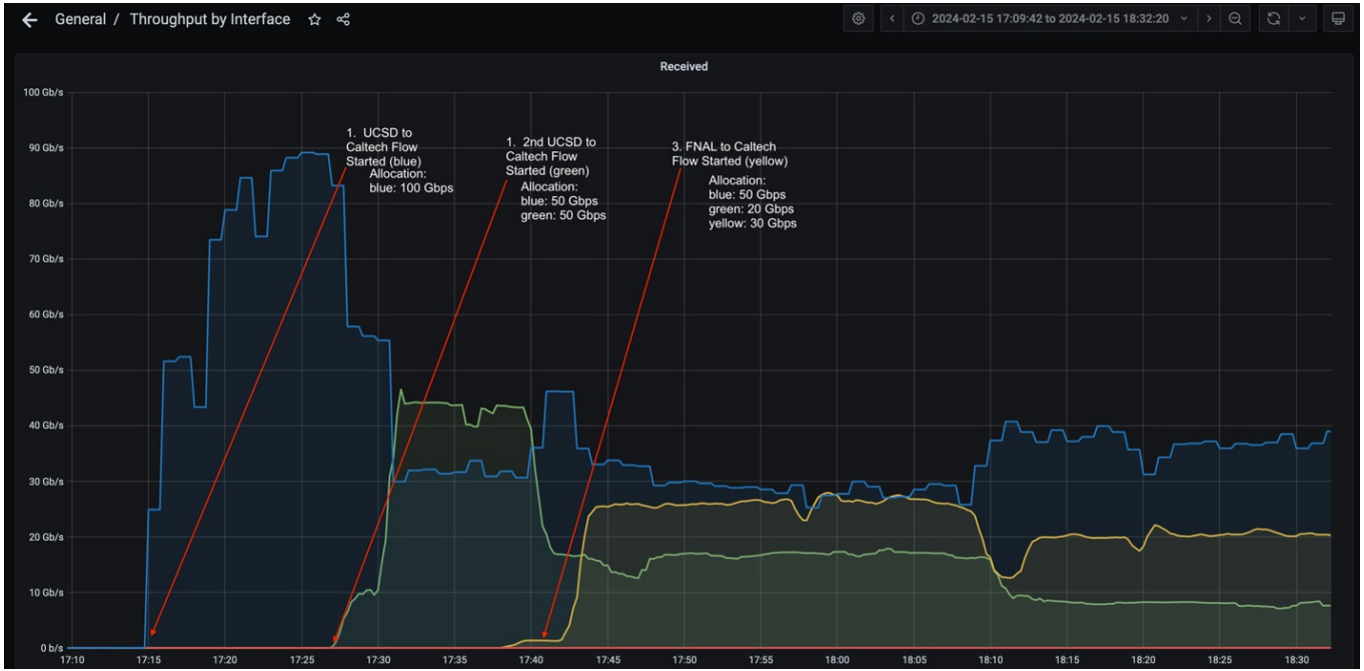
FNAL Testbed

- Working on testing with dCache WebDavDoor Proxy and High Availability features next





LHC DC24 Testing



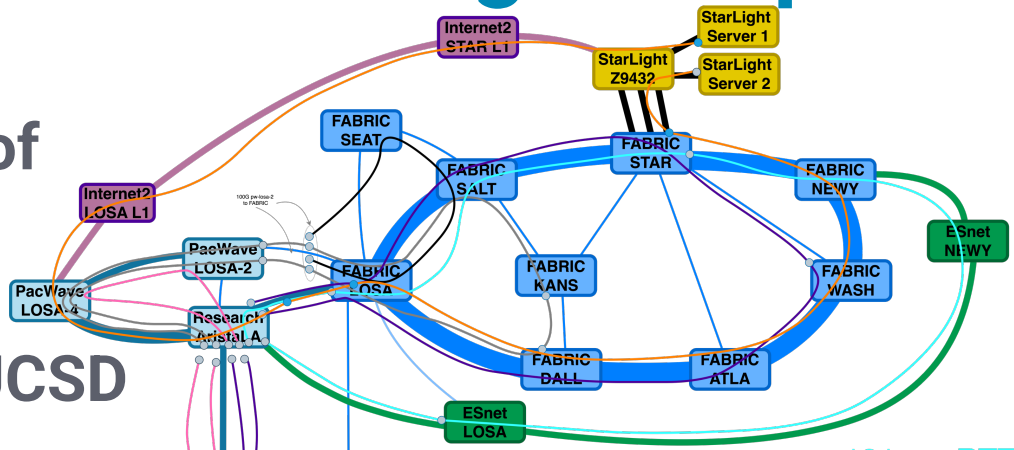
- yellow – FNAL to Caltech
- blue – UCSD to Caltech
- green – UCSD to Caltech

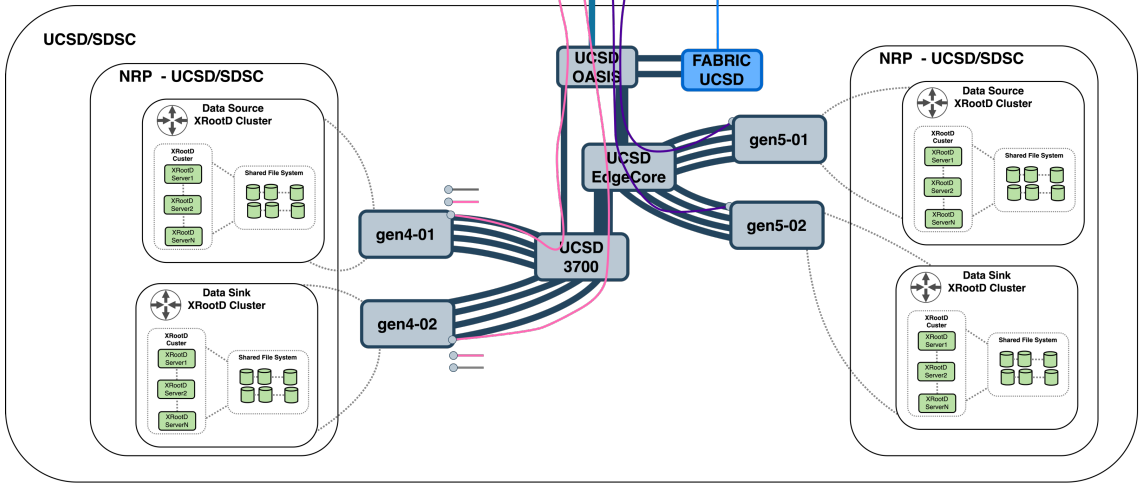
- Demonstrated ability for Rucio to define priorities and the workflow to react (modify)

XRootD Performance Testing - Loops

- R&E AutoGOLE, FABRIC used to setup a topology of variable RTT Loops
- XRootD Testing
- Source and Sink both at UCSD

- FABRIC, Internet2, ESnet, CENIC, StarLight, others





131 ms RTT
122 ms RTT
108 ms RTT
80 ms RTT
58ms RTT
6 ms RTT

100Gbps
400Gbps
800Gbps
1.2Tbps

XRootD Performance Testing - Loops

- two servers, source and sink
- multiple tests and variables
- # of XRootD instances, cores, latency, # concurrent transfers

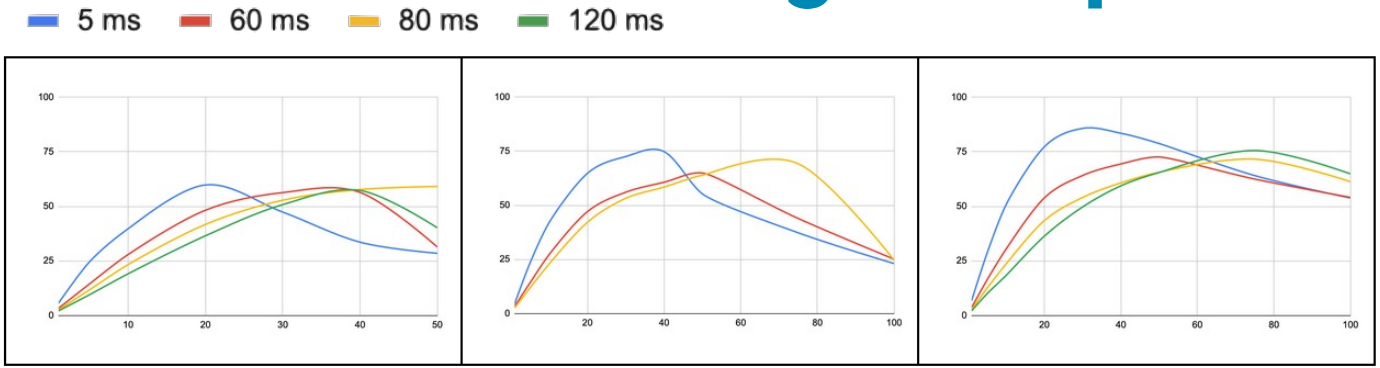


Table 1. XRootD single instance. Total cores per server: Left: 16, Middle: 32, Right: 128

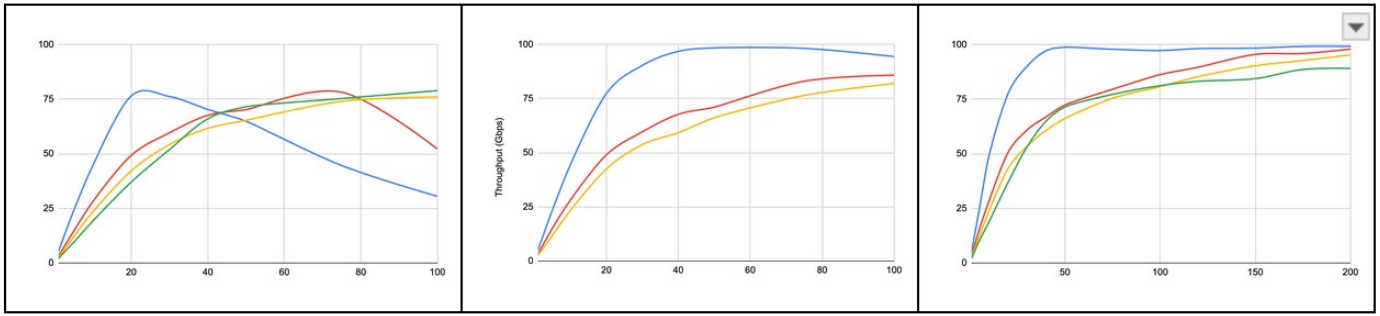


Table 2. XRootD dual instance. Total cores per server: Left: 16, Middle: 32, Right: 128

x-axis: number of concurrent transfers

y-axis: throughput (Gbps)

SENSE Rucio/FTS/XRootD Interoperation Next Steps

- Continue testing at currently deployed sites:
  - UCSD, Caltech, FNAL
- Evaluate options for other US CMS Sites
- Evaluate options for prototype deployment at SPRACE (Brazil) and CERN
- Use of Smart NICs (Nvidia Bluefield-3) as Network Service Termination point and as XRootD DataOrigins and/or WebDavDoor Proxy)

XRootD Performance Testing
Next Steps

- Current XRootD System
  - more testing at 400 Gbps loops end-to-end, multiple RTTs
  - summarize performance and optimal configuration and tuning parameters
- R&D and testing of enhancements in the areas of different data movement protocols, data bundling, and other options

Network Services Co-Design

- Current "push now worry later" data transfer model may not work so well once end sites are upgraded/tuned
- Coordination and co-design across Compute, Storage AND Network Services may be needed
- Will allow for end-to-end accountability of network utilization by end systems, and allow different stakeholders (e.g. large experiments) to express and manage priorities
- Working with Rucio/FTS/XRootD/dCache software stacks to build and test

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!

Questions? Comments?

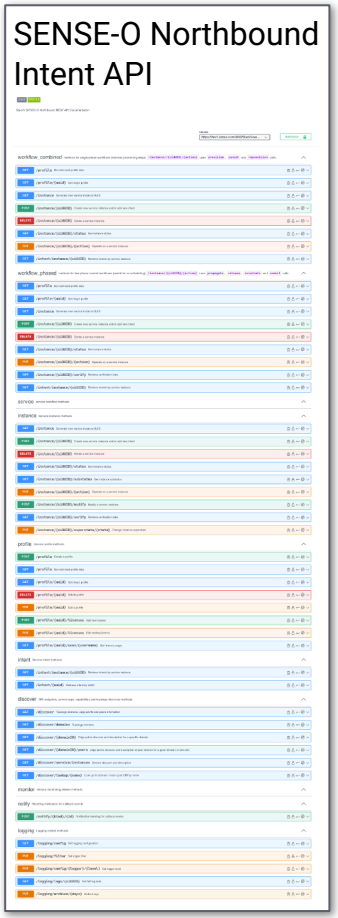
tlehman@es.net

Extra Slides

SENSE-How does an Application interact with the Network?

https://app.swaggerhub.com/apis/xi-yang/SENSE-O-Intent-API/2.0.2

SENSE-O Northbound Intent API



Functional Primitives

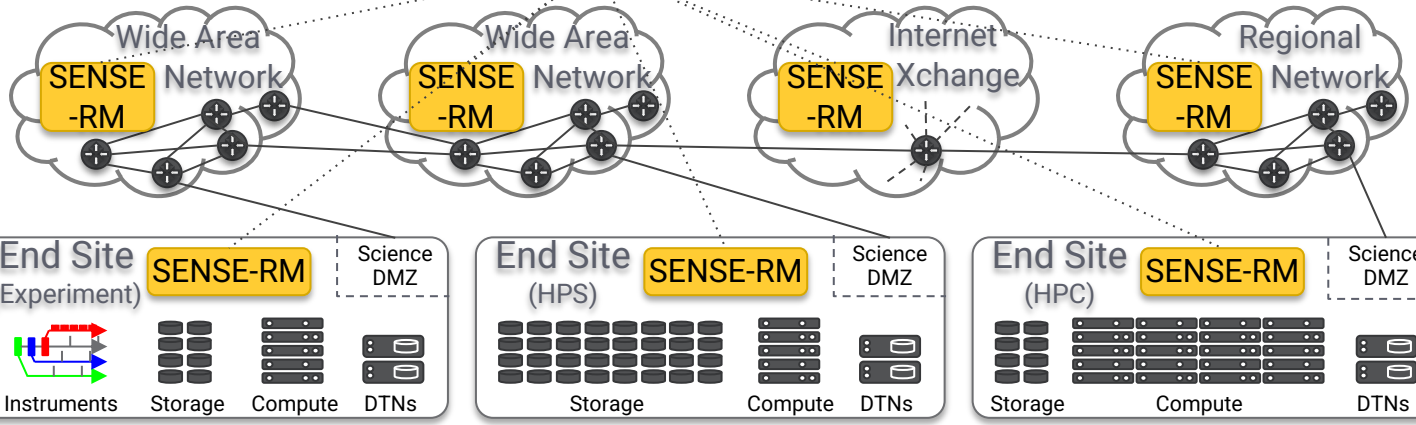
- Resource discovery
- Reserve
- Commit
- Terminate
- Modify
- Status/Audit

Types of Interactions

- What is possible?
- What is recommended?
- Request and/or negotiate for service (guarantees)
- Service status and troubleshooting

Application Workflow Agent

SENSE-Orchestrator(s)



AutoGOLE / SENSE WG

- GNA-G AutoGOLE/SENSE WG homepage
  - https://www.gna-g.net/join-working-group/autogole-sense
- Co-Chairs:
  - Tom Lehman (ESnet)
  - Marcos Felipe Schwarz (RNP)
  - Hans Trompert (SURF)
  - Buseung Cho (KISTI)
- AutoGOLE/SENSE Working Group mailing list
  - autogole@lists.gna-g.net
- Zoom meetings
  - every two weeks on Tuesdays, 10am ET

NSI Software

- OpenNSA
  - https://github.com/BandwidthOnDemand/opennsa
  - https://github.com/NORDUnet/opennsa
  - https://nordunet.github.io/opennsa/
- SuPA (SURF ultimate Provider Agent)
  - https://workfloworchestrator.org/SuPA/
  - https://github.com/workfloworchestrator/SuPA

SENSE Software

- SENSE
  - Orchestrator
    - sense.es.net
  - Site Resource Manager
    - https://github.com/sdn-sense
    - https://sdn-sense.github.io/
  - Network Resource Manager
    - https://github.com/esnet/sense-rm
- SENSE has drivers for NSI, Internet2 Insight Console, ESnet OSCARS, FABRIC, SENSE Site RM, SENSE Network RM, Cloud Providers