Rucio/SENSE in DC24

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

Nov 9th, 2023

Introduction

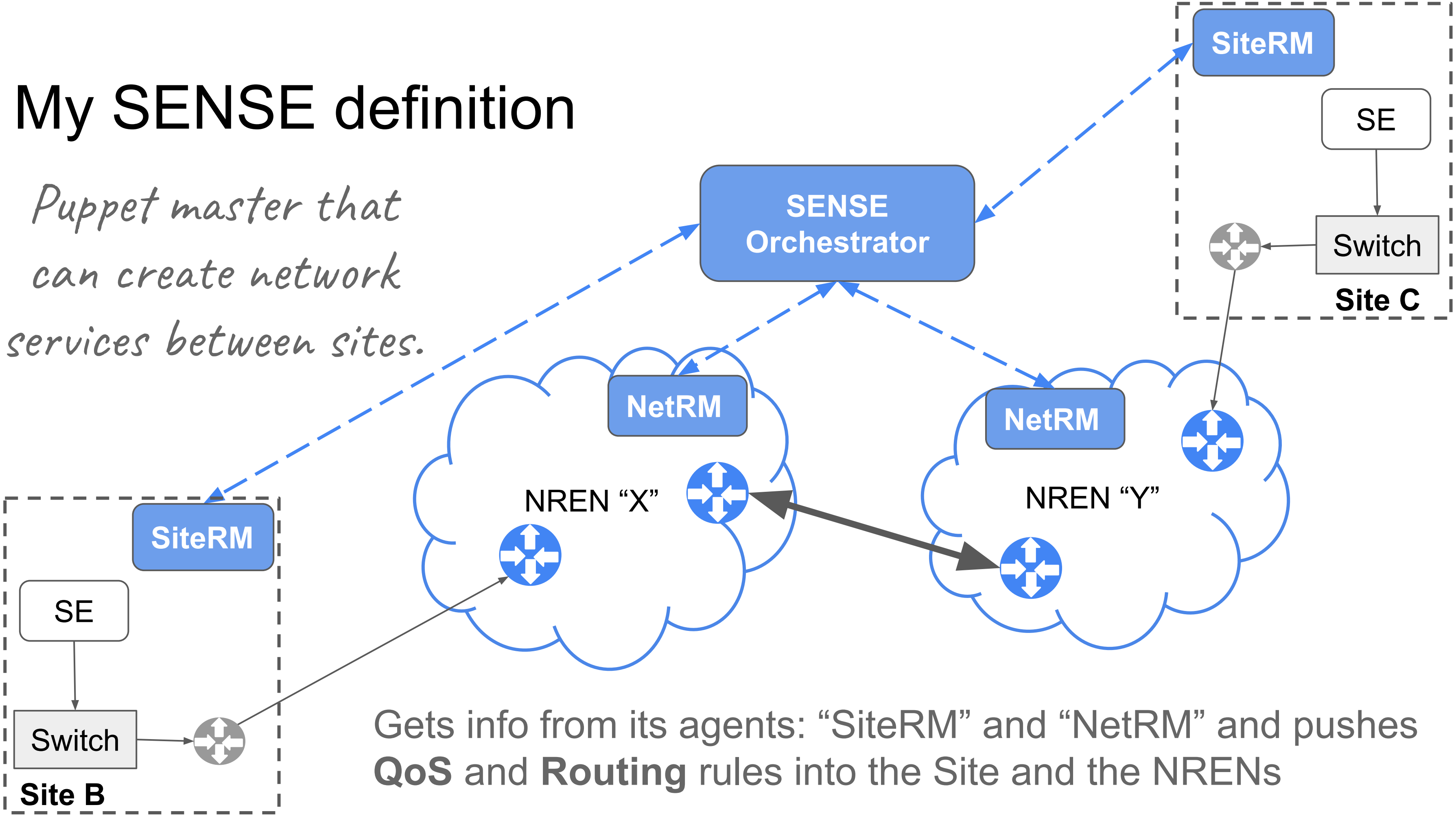
The overall goal is to enable Rucio to negotiate and manage network capabilities with SENSE

This creates end-to-end accountability of the network utilization and allows individual stakeholders (e.g. VOs) to manage their internal priorities.

Prioritization between different stakeholders are managed at the SENSE layer

My SENSE definition

Puppet master that can create network services between sites.



Gets info from its agents: "SiteRM" and "NetRM" and pushes **QoS** and **Routing** rules into the Site and the NRENs

Multi subnet Storage System

- Sense services are created based on subnets
- **Problem:** current Storage Systems live in a single subnet
- **Solution:** expose our Storage Systems over multiple subnets
- We managed to do this in XRootD by adding a bunch of configuration
  - No extra hardware needed

More details here:
https://indico.cern.ch/event/1185600/contributions/5109192/attachments/2545788/4383989/Automated%20Network%20Services%20for%20Exascale%20Data%20Movement%20(1).pdf

Rucio & SENSE

Rucio: Data Management System used by CMS and ATLAS, it knows the data workflows, how big, where they have to go, how important are.

By joining forces we can let Rucio leverage SENSE capabilities to:

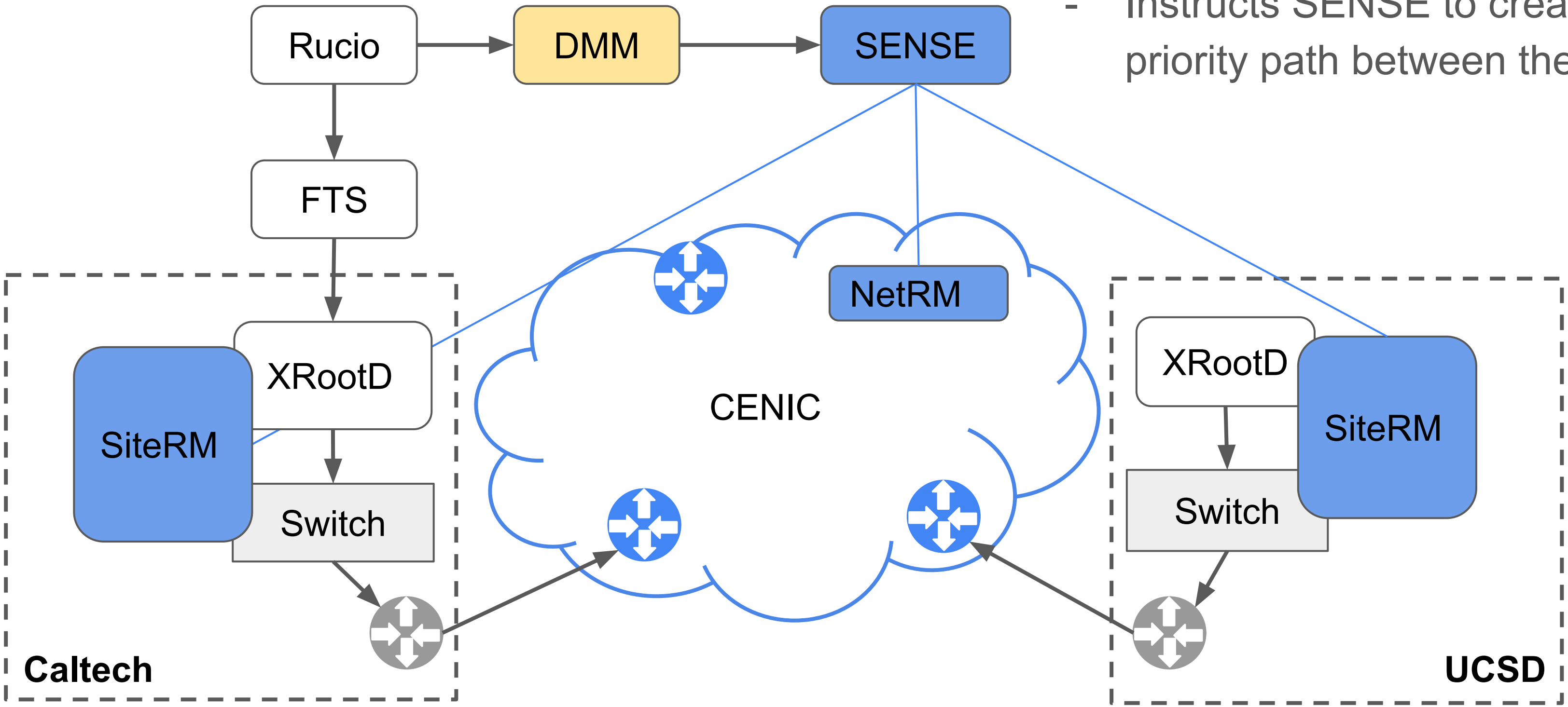
- Isolate => different data workflows travel on different subnets
- QoS => Allocate bandwidth
- Routing => Select different paths

Our target is ONLY the LARGEST and/or TIME-SENSITIVE data workflows

How it looks

DMM - Data Movement Manager:

- Picks a free subnet at each site
- Calculates bandwidth allocation
- Instructs SENSE to create a priority path between them



Project status

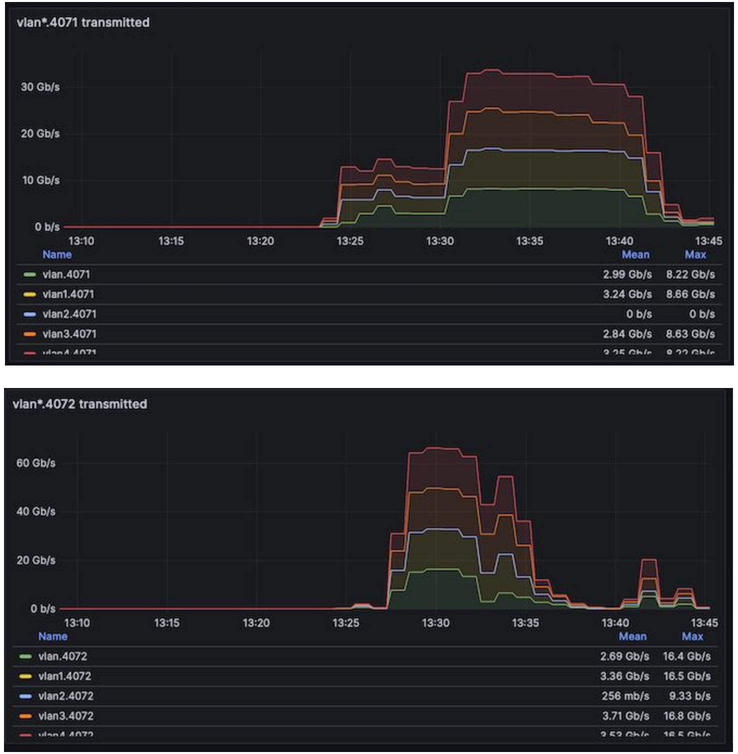
- Last year: PofC at 10Gbps with 1 managed allocation and background traffic.
- Last week: Managed 100Gbps between 2 allocations (more on this later)
- Working on new site deployments at UNL and FNAL
- Agreed with the Rucio team on initial plan for integration of DMM into Rucio
- Demonstrated 350Gbps transfers within our testbed (no SENSE managed)
    - Presented in CHEP23
- 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

Mini-challenges status

#1 (COMPLETED). High bandwidth demonstration of **multiple** Rucio initiated priority data transfers between UCSD and Caltech

In the image on the right we can see a 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 (top/bottom)

#2 (ONGOING). Demonstration of 3 priority paths between 3 different pairs of sites. We are currently testing FNAL’s deployment.



Name	Mean	Max
🟩 vlan.4071	2.99 Gb/s	8.22 Gb/s
🟨 vlan1.4071	3.24 Gb/s	8.66 Gb/s
🟦 vlan2.4071	0 b/s	0 b/s
🟧 vlan3.4071	2.84 Gb/s	8.63 Gb/s
🟥 vlan4.4071	3.25 Gb/s	8.22 Gb/s

Name	Mean	Max
🟩 vlan.4072	2.69 Gb/s	16.4 Gb/s
🟨 vlan1.4072	3.36 Gb/s	16.5 Gb/s
🟦 vlan2.4072	256 mb/s	9.33 b/s
🟧 vlan3.4072	3.71 Gb/s	16.8 Gb/s
🟥 vlan4.4072	3.53 Gb/s	16.5 Gb/s

Mini-challenge #1. Two Priority Paths created between SDSC and Caltech

Proposal for DC24

1. **Any day**: 400Gbps of SENSE managed throughput **SDSC => Caltech**
    1. completely orthogonal: via CENIC i.e. no prod links used
    2. do we want to monitor/account this for as DC?
2. **Day 4** (“reprocessing”): **FNAL => Caltech & SDSC => Caltech** share of 100Gbps of SENSE managed throughput. The share 66/33 gets swapped in intervals of 1 hr
    1. Will use a non-Prod path from FNAL -> ESnet
    2. within ESnet level the link is shared with Prod
3. **Day 11 or 12** (Contingency): **FNAL => Caltech** at max throughput available via SENSE
    1. Same as #2

For 2 and 3: we would ask for **green-light from FNAL** and **synch up with DUNE** activities**.** Also we **need to agree on how to monitor**: “RSE names” and “Activity”

For all 3 cases NO Production Storage would be used

How Rucio/SENSE can improve DC in the future?

Monitoring is a very important part of DC but obtaining precise throughput between a given pair of sites is not trivial.

Rucio/SENSE can create priority paths between a given pair of sites for the “Data Challenge” injected data and provide not only a bandwidth guaranteed path but also instantaneous throughput monitoring for such path.

Questions?

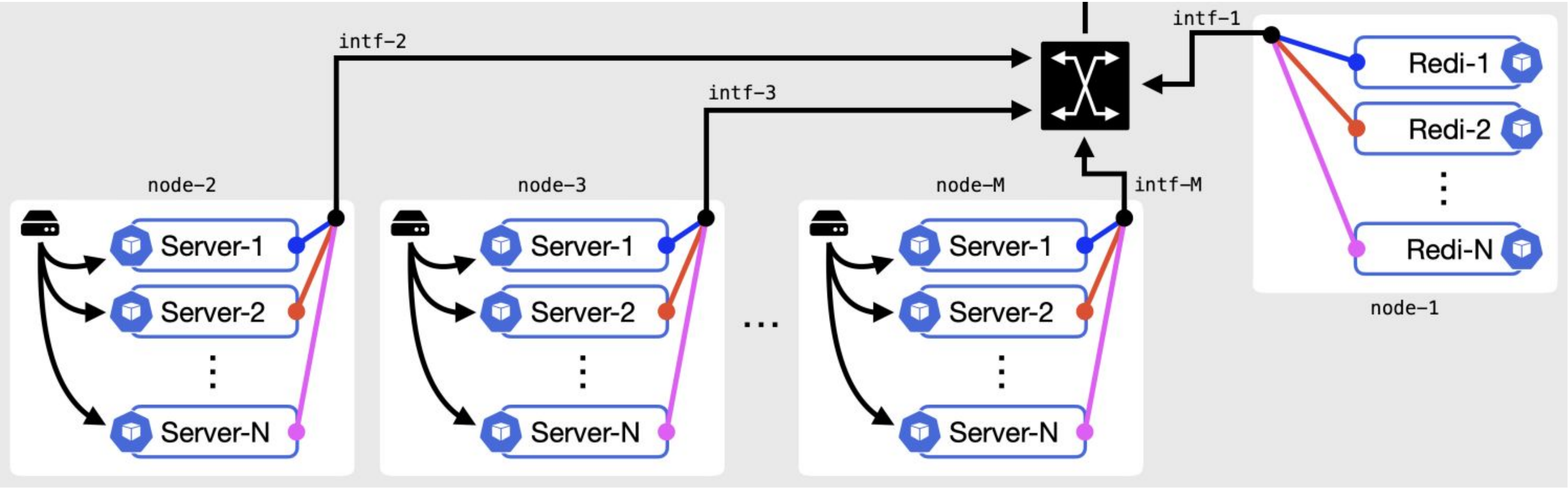
ACKNOWLEDGMENTS

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 slides

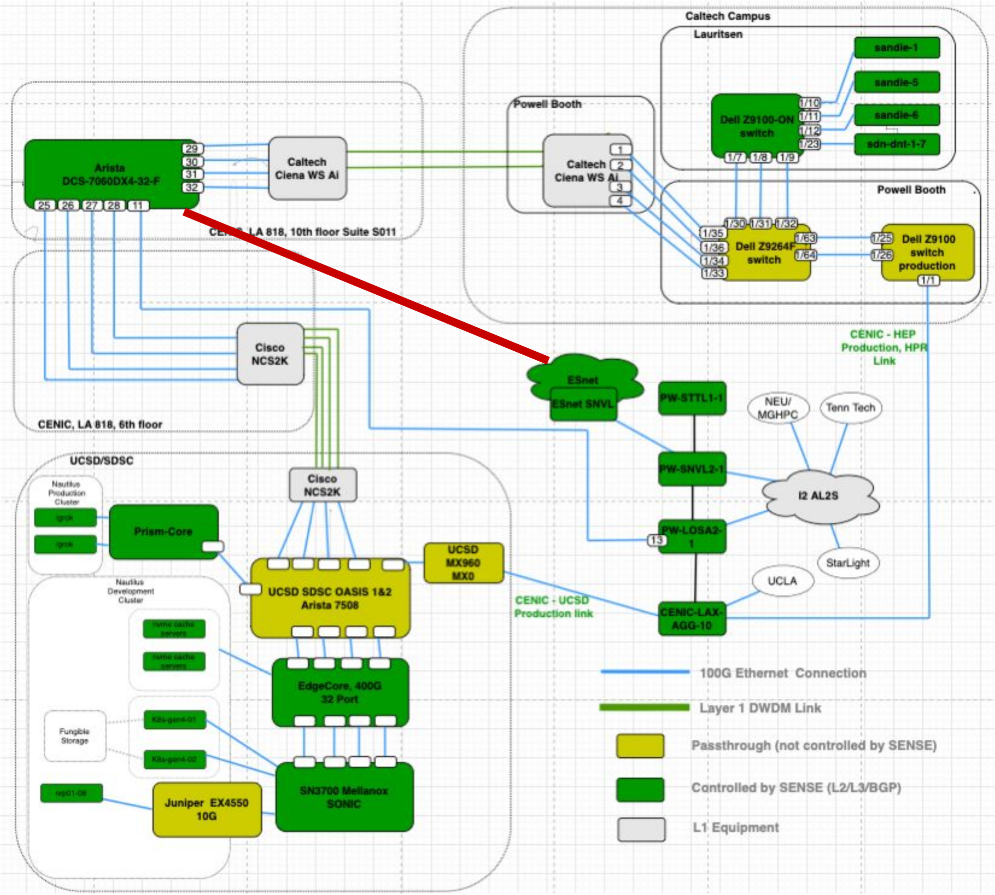
Isolation using XRootD multi-endpoint

- A single data server is configured to listen in N different IPv6 addresses.
- We use IPv6 because we need many IP addresses



XRootD cluster with M servers and N subnets, Every color represents a different subnet

New 400Gbps


Caltech Campus
Lauritsen
sandie-1
sandie-5
sandie-6
sdn-dnt-1-7
Dell Z9100-ON switch
1/10
1/11
1/12
1/23
1/7
1/8
1/9
Powell Booth
Caltech Ciena WS Ai
1
2
3
4
Powell Booth
1/30
1/31
1/32
1/35
1/36
1/34
1/33
Dell Z9264F switch
1/63
1/64
1/25
1/26
Dell Z9100 switch production
1/1
Arista DCS-7060DX4-32-F
29
30
31
32
25
26
27
28
11
Caltech Ciena WS Ai
CENIC, LA 818, 10th floor Suite S011
Cisco NCS2K
CENIC, LA 818, 6th floor
ESnet
ESnet SNVL
PW-STTL1-1
PW-SNVL2-1
PW-LOSA2-1
13
CENIC-LAX-AGG-10
CENIC - HEP Production, HPR Link
NEU/ MGHPC
Tenn Tech
I2 AL2S
StarLight
UCLA
UCSD/SDSC
Nautilus Production Cluster
igrok
igrok
Prism-Core
Cisco NCS2K
UCSD SDSC OASIS 1&2 Arista 7508
UCSD MX960 MX0
CENIC - UCSD Production link
Nautilus Development Cluster
nvme cache servers
nvme cache servers
EdgeCore, 400G 32 Port
Fungible Storage
K8s-gen4-01
K8s-gen4-02
nrp01-08
Juniper EX4550 10G
SN3700 Mellanox SONIC
100G Ethernet Connection
Layer 1 DWDM Link
Passthrough (not controlled by SENSE)
Controlled by SENSE (L2/L3/BGP)
L1 Equipment