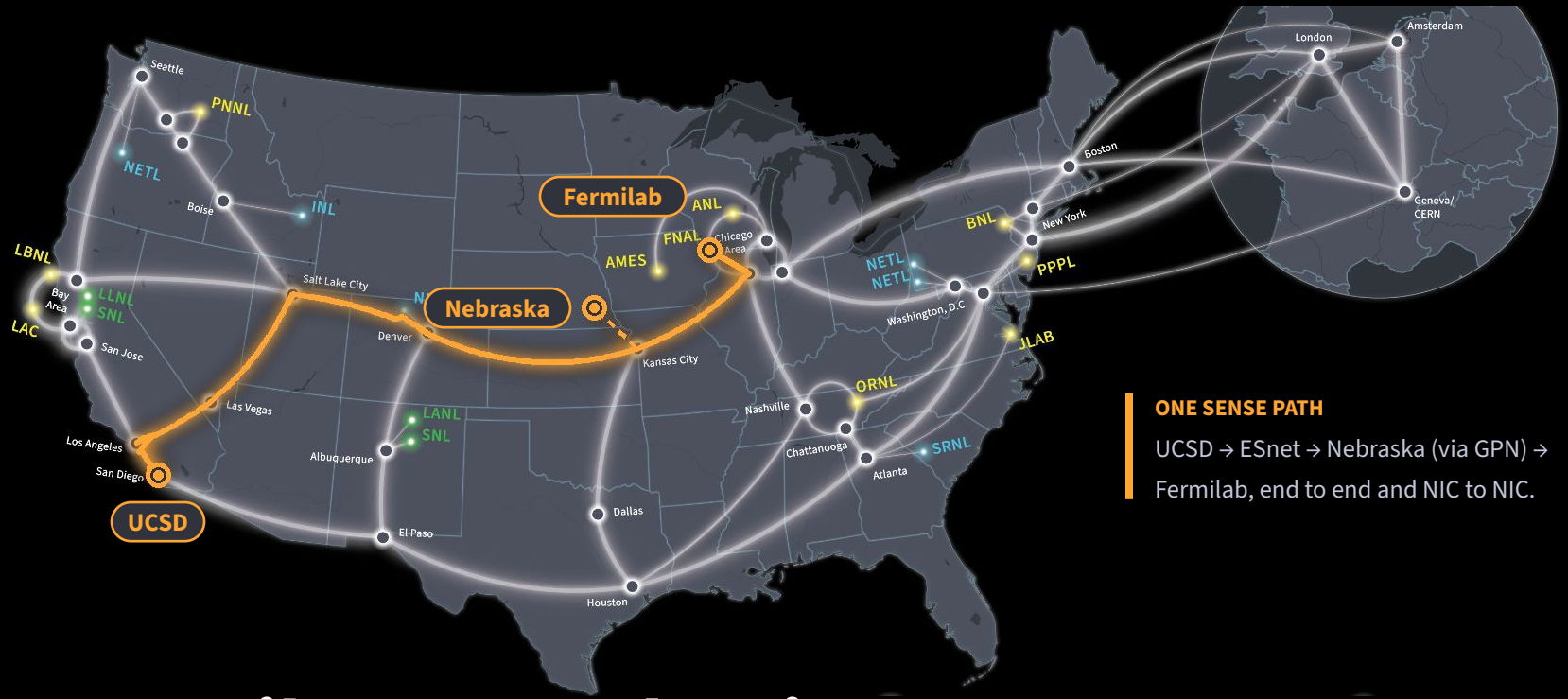
ESnet
ENERGY SCIENCES NETWORK

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY



ONE SENSE PATH

UCSD → ESnet → Nebraska (via GPN) → Fermilab, end to end and NIC to NIC.

From Pilot to Production

Scaling SENSE/SiteRM across sites

Justas Balcas
ESnet - on behalf of the SENSE team
[HEPiX · Lincoln · 22/10/26]

The network as a first-class resource

WORKFLOW “I need to move 1 PB anytime in the next 24 hours.”

NETWORK “Start in 2 hours. You get at least 400 Gbps end-to-end, for 8 hours.”

Today: start the flow and hope

No real-time per-flow view, no deterministic service, best effort on every hop.

SENSE: intent-based APIs

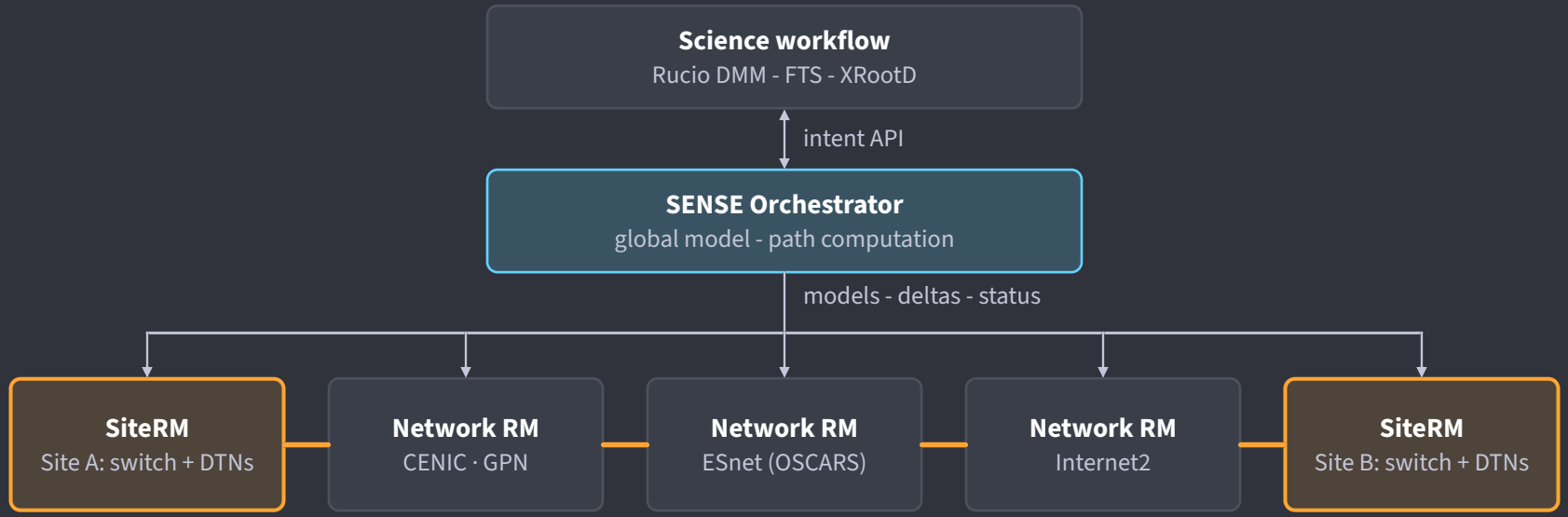
Discovery, negotiation, provisioning and lifecycle monitoring across every domain on the path.

Owners stay in control

Each domain runs its own Resource Manager and decides what SENSE may change.

ESnet

SENSE spans 1 to N domains on the path

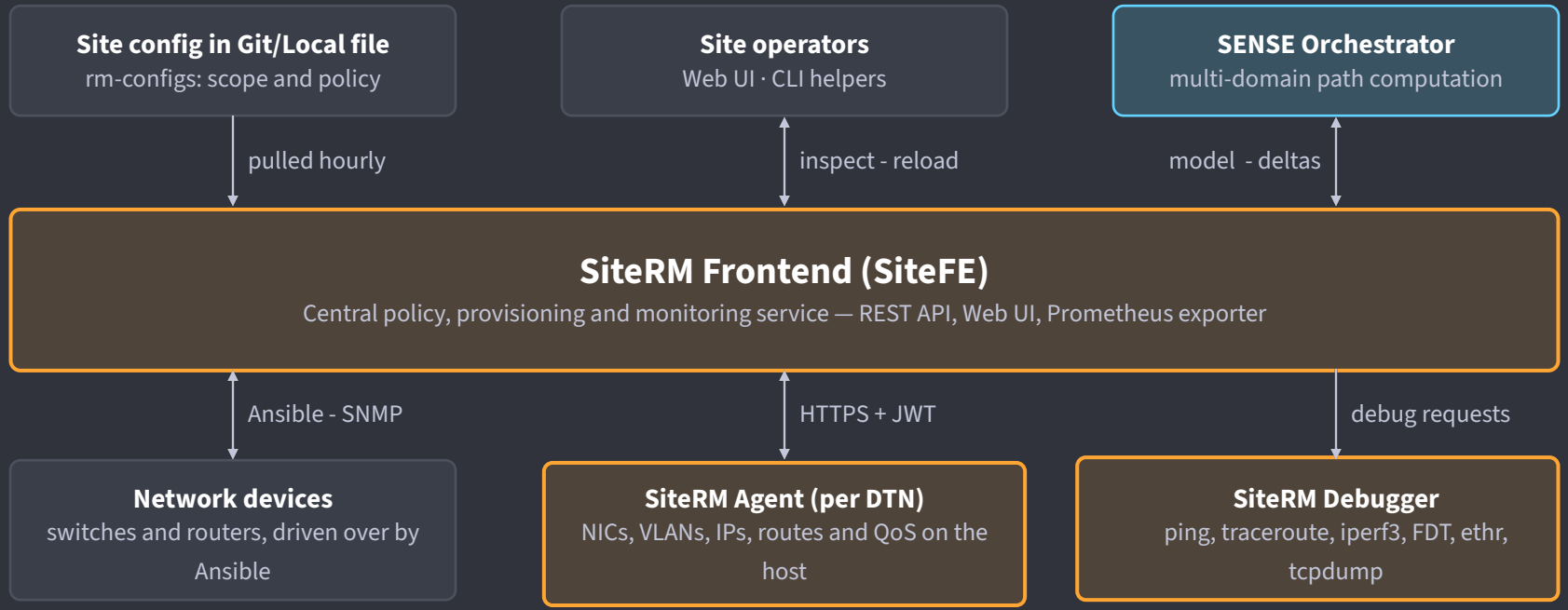


End-to-end data path: DTN - site network - regional - WAN - site network - DTN

Every Resource Manager is run by Site. Site decides what SENSE is allowed to do.

ESnet

One Frontend per site, one Agent/Debugger per DTN



ESnet

Container 1(anywhere at Site) Frontend: focused services

REST API + Web UI
JWT and X.509 auth, Swagger, topology, deltas, service states.

LookUpService
Builds the site model (MRML) from device facts, agent reports and Git/Local config.

PolicyService
Accepts deltas from SENSE-O and checks VLAN, IP, capacity and capability conflicts.

Provisioning
Prepares VLAN, BGP and QoS config and applies only what differs on the device.

BGPMonitoring
Session state and prefix counts for every SENSE-owned BGP peer.

Validator
Cross-checks config against reality: LLDP, device output, BGP health.

SNMPMonitoring
Port and VLAN counters, MAC tables; Prometheus exporter.

ESnet

Container on each DTN: the SiteRM Agent

Observes

- NICs, speeds, MACs, VLANs and IPs
- LLDP neighbors via lldpd
- Routes, ARP/NDP, interface counters
- Kubernetes and Multus pod interfaces
- Host certificate validity

Applies

- VLAN sub-interfaces and IPs
- Policy routing rules for L3 paths
- Linux tc QoS per VLAN or per IP range
- DSCP marking: 46 guaranteed, 18 soft, 0 best effort

Guards

- Touches only interfaces in its config
- Checks new IPs against host addresses
- Refuses QoS that oversubscribes the NIC
- DSCP marking can be disabled per agent

Runs as a container with **hostNetwork** and **NET_ADMIN** on bare metal, Docker/Podman or Kubernetes (Helm). It talks only to its own Frontend.

ESnet

Container on Each DTN: Debugger

Reachability

`rapid-ping`

`rapid-pingnet`

`traceroute`

`traceroute-net`

From a DTN or from the switch itself, inside the SENSE VRF.

Throughput

`iperf3 server/client`

`fdt server/client`

`ethr server/client`

Ports, streams and timeouts within site-defined limits.

Packets and tables

`tcpdump`

`arp-table`

`prometheus-push`

Capture on the SENSE VLAN; push results to Prometheus.

Operators submit from the Web UI or REST API.

RTMon keeps pings running on every active path.

The smoke tester pings after every provision.

ESnet

What a site can offer through SiteRM

Layer 2 point-to-point
A tagged VLAN between any two endpoints: DTN NIC, switch port or router port.

Multipoint Layer 2
One broadcast domain shared by many endpoints across many sites.

Layer 3 with BGP
L2 path between routers plus private BGP that carries only the service /64s.

Multipath BGP
Several VLAN and BGP peer pairs to one destination for load sharing and redundancy.

QOS CLASSES

`guaranteedCapped`
Dedicated bandwidth with a hard upper limit.

`softCapped`
Guaranteed baseline; may burst into spare capacity.

`bestEffort`
Whatever is left, at lower priority.

> **Rucio Priority Network Service** = Layer 3 BGP over a guaranteed L2 path between routers. Only the priority /64 is rerouted; everything else stays on the default routed path.

The site stays in control

Narrow scope
VLAN ranges, listed ports, IPv4/6 pools, private ASN and VRF. SiteRM touches nothing else.

Reviewed changes
Policy lives in Git (rm-configs): pull requests, a config validator, hourly reload
(Or local files without reload)

Scoped access
JWT for people, X.509 challenge for agents, read-only SENSE-O tokens, IP allow-lists.

Announced downtime
FE/downtime.yaml tells SENSE-O and monitoring to pause new requests.

No surprises
Every device command is documented per vendor; cleanup tools, test tools

FE/MAIN.YAML · EXAMPLE

```
T2_US_Example:
  plugin: ansible
  ipv6-subnet-pool:
    - 2001:db8:0:1120::/64
  switch: [dellos10_s0]
dellos10_s0:
  private_asn: 64514
  vlan_range: [3600-3619]
  ports:
    Port-channel 1:
      capacity: 100000
      wanlink: true
  allports: false
```


ESnet

Real-time NIC-to-NIC topology (MRML Model)

Switch LLDP neighbors

Switch MAC tables

Host LLDP (lldpd)

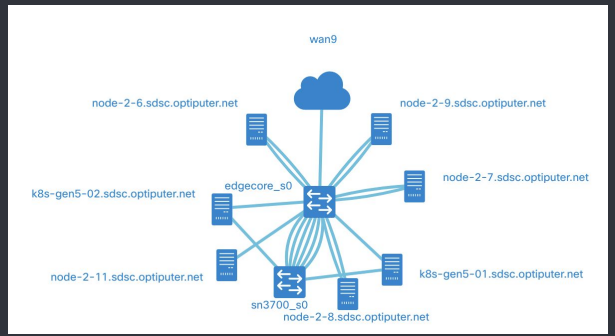
WAN links from Git

Topology builder/LookupService
matches chassis IDs and MACs into topology file

Web UI topology
switches, hosts, WAN links

Validator
host MAC vs. switch LLDP

Model to SENSE-O
paths end at NIC ports



ESnet

What is Delta?

```
<urn:ogf:network:nrp-nautilus.io:2020:sn3700_s0:service+rst-ipv6:table+e8839f27-23b3-42a6-88e2-41dcd3325918:route+wan-1:conn+Connection_1:to-prefix-list-1>
    a     owl:Thing , rdfs:Resource , <http://schemas.ogf.org/mrs/2013/12/topology#NetworkAddress> ;
    <http://schemas.ogf.org/mrs/2013/12/topology#type>
        "ipv6-prefix-list" ;
    <http://schemas.ogf.org/mrs/2013/12/topology#value>
        "2620:6a:0:2842::/64" .
<urn:ogf:network:nrp-nautilus.io:2020:sn3700_s0:service+rst-ipv6:table+e8839f27-23b3-42a6-88e2-41dcd3325918:route+wan-1:conn+Connection_1:via-1>
    a     owl:Thing , rdfs:Resource , <http://schemas.ogf.org/mrs/2013/12/topology#NetworkAddress> ;
    <http://schemas.ogf.org/mrs/2013/12/topology#type>
        "ipv6-address" ;
    <http://schemas.ogf.org/mrs/2013/12/topology#value>
        "fc00:0:204:126:0:0:0:1/64" .
<urn:ogf:network:nrp-nautilus.io:2020:sn3700_s0:PortChannel502:conn+e8839f27-23b3-42a6-88e2-41dcd3325918:vt+l2-policy-Connection_1:vlanport+3116:service+bw>
    a     <http://schemas.ogf.org/mrs/2013/12/topology#BandwidthService> ;
    <http://schemas.ogf.org/mrs/2013/12/topology#maximumCapacity>
        "50000000000"^^xsd:long ;
    <http://schemas.ogf.org/mrs/2013/12/topology#reservableCapacity>
        "50000000000"^^xsd:long ;
    <http://schemas.ogf.org/mrs/2013/12/topology#type>
        "guaranteedCapped" ;
    <http://schemas.ogf.org/mrs/2013/12/topology#unit>
        "bps" .
<urn:ogf:network:nrp-nautilus.io:2020:sn3700_s0:PortChannel502:conn+e8839f27-23b3-42a6-88e2-41dcd3325918:vt+l2-policy-Connection_1:vlanport+3116>
    a     <http://schemas.ogf.org/nml/2013/03/base#BidirectionalPort> ;
    <http://schemas.ogf.org/mrs/2013/12/topology#hasNetworkAddress>

<urn:ogf:network:nrp-nautilus.io:2020:sn3700_s0:PortChannel502:conn+e8839f27-23b3-42a6-88e2-41dcd3325918:vt+l2-policy-Connection_1:vlanport+3116:ipv6-address+fc00_0_204_126_0_0_0_2_64> ;
    <http://schemas.ogf.org/nml/2013/03/base#hasService>
        <urn:ogf:network:nrp-nautilus.io:2020:sn3700_s0:PortChannel502:conn+e8839f27-23b3-42a6-88e2-41dcd3325918:vt+l2-policy-Connection_1:vlanport+3116:service+bw> ;
    <http://schemas.ogf.org/nml/2013/03/base#isAlias>
        <urn:ogf:network:nrp-nautilus.io:2020:edgecore_s0:PortChannel502:conn+e8839f27-23b3-42a6-88e2-41dcd3325918:vt+l2-policy-Connection_1:vlanport+3116> .
```

NDL (Network Description Language) - early semantic network modeling language.

NML (Network Markup Language) - formalized network topology modeling through the Open Grid Forum.

MRML (Multi-Resource Markup Language) — extended NML to represent heterogeneous resources such as DTNs, compute, storage, and cloud infrastructure.

ESnet

One intent, many vendor dialects

DESIRED STATE

```
interface:
  vlan3607:
    vlanid: 3607
    mtu: 9216
    vrf: lhcone
    ipv6_address:
      fc00::16/124: present
    tagged_members:
      <port>: present
    state: present
```



DELL OS10

```
interface vlan 3607
 mtu 9216
 ip vrf forwarding lhcone
 ipv6 address fc00::16/124
interface Port-channel 102
 switchport trunk allowed vlan 3607
```

ARISTA EOS

```
vlan 3607
interface Vlan 3607
 vrf lhcone
 ipv6 address fc00::16/124
interface Ethernet31/1
 switchport trunk allowed vlan add 3607
```

The same structure drives Cisco NX-OS, SONiC, Junos (EX/QFX and MX/PTX), FRR and FreeRTR, and covers BGP, prefix-lists, route-maps and QoS.

ESnet

Monitoring: from every port to a Slack alert

1. Collect
Switch SNMP and Ansible facts, Agent reports, BGP summaries, service states.

→

2. Export
SiteFE Prometheus endpoint: port and VLAN rates, MACs, QoS and BGP state, exc codes.

→

3. Watch
Autogole Prometheus and Grafana on NRP: site dashboards, probes, alert rules.

→

4 .Alert
Slack: channel per SiteRM and NSI domain, on failure and on recovery.

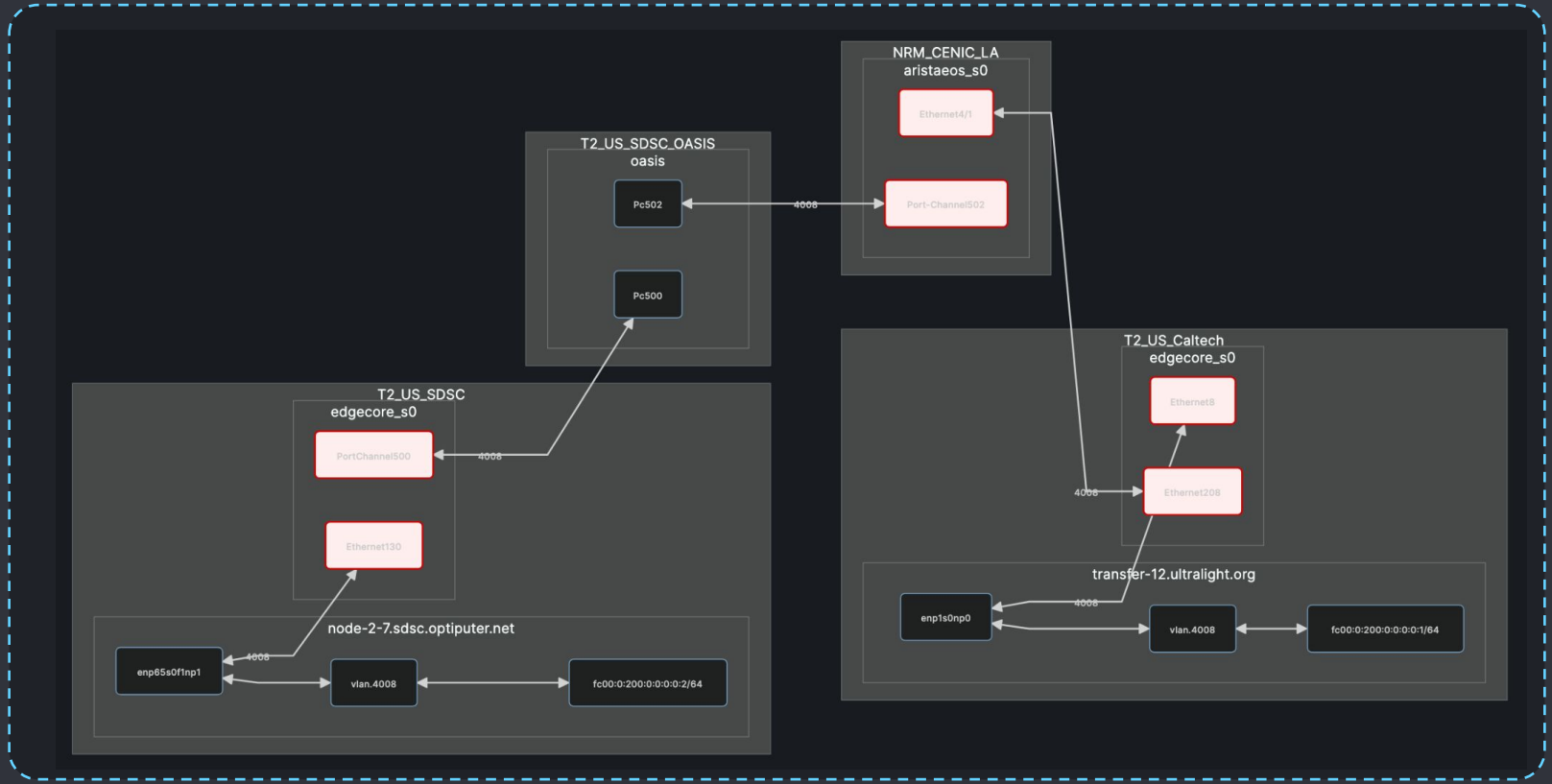


Per-request view: RTMon

Reads each active request's manifest from SENSE-O and generates its own Grafana dashboard: every hop from site switches (SiteRM) through ESnet Stardust, Internet2 TSDS and CENIC SNMP, pings from the endpoints with annotations, and an L2 MAC-learning panel. Removed when the request ends.

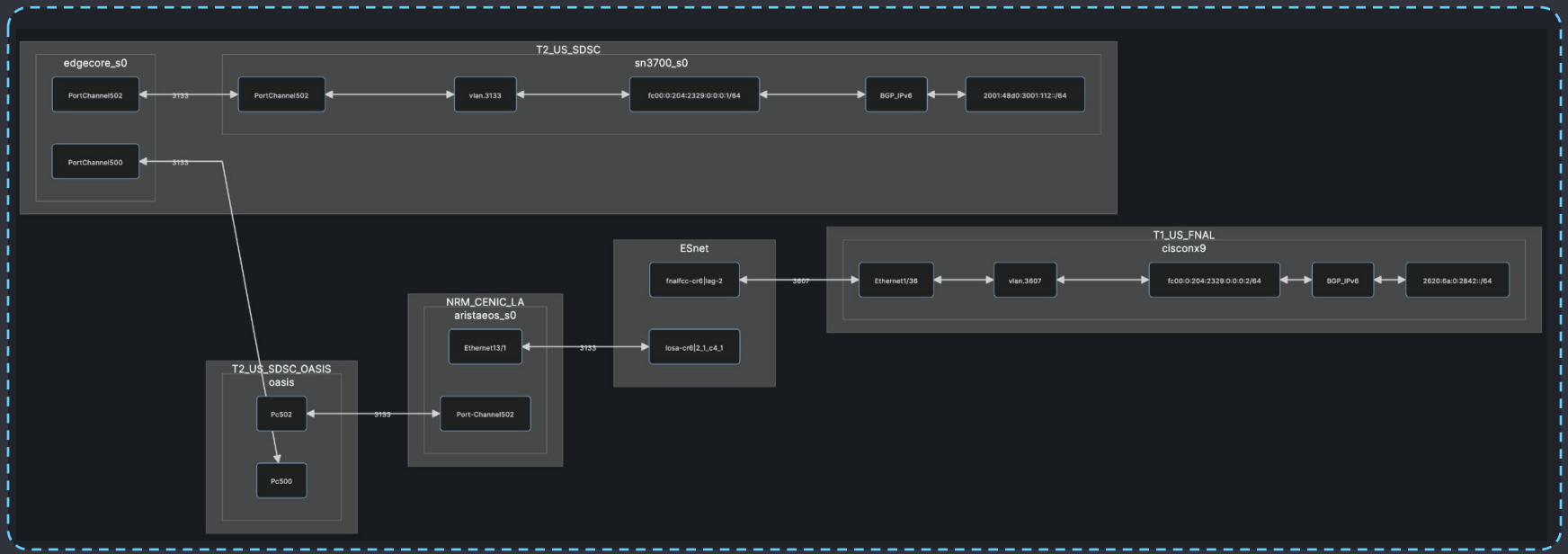
ESnet

RTMon: one dashboard per request



ESnet

RTMon: one dashboard per request



ESnet

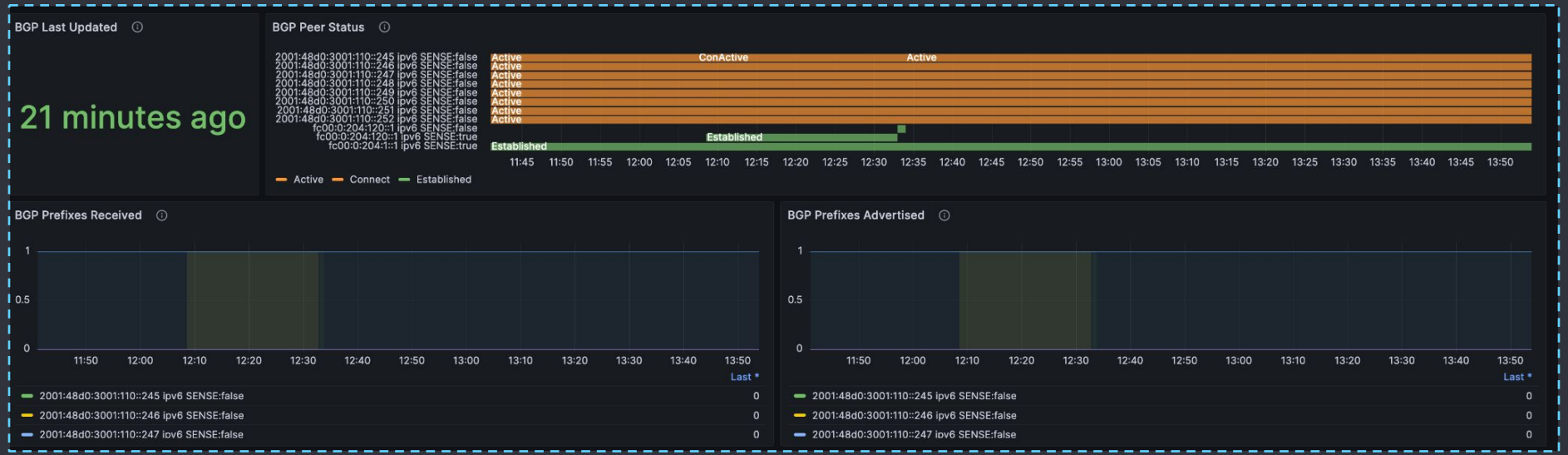
RTMon: one dashboard per request



⚠ Switch T2_US_SDSC:edgecore_s0 has not learned the MAC 94:6d:ae:a6:66:41 of T2_US_SDSC node-2-7.sdsc.optiputer.net on VLAN 4008. It is marked in red in the flow diagram.
⚠ Switch NRM_CENIC_LA:aristaeos_s0 has not learned the MAC 94:6d:ae:a6:66:41 of T2_US_SDSC node-2-7.sdsc.optiputer.net on VLAN 4008. It is marked in red in the flow diagram.
⚠ Switch T2_US_Caltech:edgecore_s0 has not learned the MAC 94:6d:ae:a6:66:41 of T2_US_SDSC node-2-7.sdsc.optiputer.net on VLAN 4008. It is marked in red in the flow diagram.
⚠ No BGP session data in Prometheus for switch T2_US_SDSC:edgecore_s0. Its BGP panel is not shown.
⚠ No SNMP data in Prometheus for switch T2_US_SDSC_OASIS:oasis. Its flow and L2 debugging panels are not shown.
⚠ No BGP session data in Prometheus for switch T2_US_SDSC_OASIS:oasis. Its BGP panel is not shown.
⚠ No BGP session data in Prometheus for switch NRM_CENIC_LA:aristaeos_s0. Its BGP panel is not shown.
⚠ No BGP session data in Prometheus for switch T2_US_Caltech:edgecore_s0. Its BGP panel is not shown.

ESnet

RTMon: one dashboard per request



ESnet

Failure 1: a switch port goes down

1. **Detect**
   Every LookupService cycle reads oper status and line protocol for each configured port and port-channel member.
2. **Decide**
   Anything but up or connected (notconnect, errdisabled, down, lowerLayerDown) counts as unhealthy.
3. **Protect**
   The port is left out of the model sent to SENSE-O, so no new path is computed over a dead link.
4. **Signal**
   WARNING -119 SWITCH_PORT_OPER_DOWN (member down: -116) → Prometheus → Grafana → Slack.
5. **Recover**
   Once the port is up again it returns to the model automatically. No manual step in SiteRM.

Why it matters

Path computation assumes every port in the model is live. Previously, only bundle members were checked, so a port-channel taken down by a loop still looked usable.

SERVICE STATE

```
WARNING · exc code -119
Port <switch><port> is enabled in
SiteRM config but operstatus is
'down' (lineprotocol 'down');
excluded from the model.
```


ESnet

Failure 2: a BGP session goes down

1. **Scope**
   BGPMonitoring checks only devices with an active SENSE BGP delta: FRR, SONiC, Dell OS9/OS10, Cisco NX-OS, Junos.
2. **Collect**
   One cross-vendor BGP summary playbook; a peer is tagged sense=true when it matches the delta's next hop.
3. **Timing**
   Hourly; three back-to-back scans after the peer set changes; on demand via "Force BGP rescan".
4. **Validate**
   -121 not established · **-122** 0 prefixes received · **-123** 0 advertised · **-120** no data yet
5. **Expose**
   Session state, prefixes received and advertised, uptime and last-scan time as Prometheus metrics.

"Established" is not enough

A session can be up while a prefix-list filters everything. Traffic then quietly takes the default routed path: the transfer works, the guarantee does not. Prefix counts catch it.

ESnet

Other failure modes, same pattern

Situation	What SiteRM does	Signal
Port-channel member down	Bundle stays; degraded member flagged	-116
Device returns no data	Device and ports flagged; re-scan after 1 h	-107 / -108
Host MAC ≠ switch LLDP	Miscabling or wrong port mapping flagged	-109
VLAN on device, not in a delta	Never handed out; flagged for cleanup	-114
VLAN range used up	No free VLANs in the model; operators warned	-112 / -113
Agent cannot reach its FE	Keeps applied config; FAILED after 10 min	-34
Model changes every cycle	Update loop suspected after 60 changes	-115
Certificate near expiry	Alert before authentication breaks	probe alert

Debounced: a warning must repeat **5 times** before it is raised, so a single slow poll never pages anyone.

ESnet

Drift happens: continuous reconciliation

On switches and routers

- Every cycle, compare each active delta (VLAN, IPs, BGP, QoS) with live device facts
- Apply only the difference
- After a restart, skip config the device already carries
- Optional retry of failed applies, with back-off
- Force-apply/cancel a single request from the helper CLI

On every DTN

- Every cycle, compare active deltas with what is on the host
- Re-create missing VLAN interfaces, IPs, routes and QoS
- Remove whatever is no longer active
- Honor start and end times locally
- Keep applied config when the Frontend is unreachable

Desired state lives in SiteRM. Reboots, rollbacks and accidental deletions converge on the next cycle; manual additions are flagged, never silently removed.

Seven network OS families, one policy

Network OS	VLAN	BGP	MultipathBGP	QoS	In production at
Cisco NX-OS	yes	yes	yes	yes*	Fermilab, Wisconsin
Dell OS10	yes	yes	yes	yes*	Caltech, Nebraska
Azure SONiC	yes	yes	yes	no	Caltech, UCSD, SDSC
Juniper Junos	yes	yes	yes	yes*	MGHPCC, CENIC
FRR / FRR+VPP	yes	yes	yes	no	UChicago, FABRIC nodes
Arista EOS	yes	no	no	yes	CENIC (Los Angeles)
Dell OS9	yes	yes	yes	yes*	Caltech

All of them are modeled for SENSE.

In development: Nokia SR OS, Mellanox OS.

* - device limit for QoS capabilities.

QoS end to end: host, site network, WAN

DTN · AGENT

Linux tc per VLAN or per IP range, plus DSCP marking. An L3 request is shared across every DTN behind the /64.

SITE SWITCH · FE

Traffic class and policer per VLAN. Guaranteed first; soft and best effort share the rest, soft at higher priority.

WAN · ESNET

OSCARS-managed QoS on the MPLS control plane, requested by SENSE-O for the same path.

Fair share across DTNs

```
10 DTNs × 10 Gbps behind one /64
L3 requests: 10, 20, 50 Gbps
→ per DTN: 1, 2, 5 Gbps
What if 2 DTNs down?
→ per DTN: 1.25, 2.5, 6.25 Gbps
```

Still hard

- Bandwidth guarantee ≠ throughput: DMM/Rucio tunes FTS concurrency
- tc is per interface and per namespace; Multus pods escape host shaping
- No QoS on SONiC or Junos MX/PTX (without license)
- Different node types (CPU/Mem speed), other load, NIC Speeds, Server Up/Down, Reserved Bandwidth, L2/L3 requests complicates calculation

ESnet

MGHPCC Prod to UCSD/SDSC (2 servers)

Best Effort (default) **Red** used full capacity when no competition.
SENSE Priority traffic **Blue** took over the capacity when requested.



ESnet

The devil is in the details

Dell OS9
At most 3 rate limits per port: the QoS budget is a device capability, not a constant.

Dell OS10
Deleting a VLAN also drops it from trunks; other systems need an explicit removal step. QoS same limit as Dell OS9

Juniper
EX uses vlan units, QFX uses irb; MX/PTX bridge through a virtual-switch instance. QoS requires License

SONiC
BGP lives in the embedded FRR, ping needs ip vrf exec, and QoS is not supported.

Arista
Police rates have a ceiling, so max_policy_rate and burst size are set per switch.

Everyone
LLDP IDs, MAC formats, switchport flags, BGP counters, servers, Docker/Kube, network setup all differ; normalize first.

> **What worked:** our own Ansible collection per vendor, deterministic fact subsets, and per-device capability flags (vlanmode, vlanip, qos_policy, allports).

Four sites, four different shapes

	Caltech	SDSC / NRP	UChicago	CERN on FABRIC
Devices	Dell OS9, Arista EOS, Dell OS10 ×2 + SONiC	SONiC ×2 (SN3700, Edgecore)	FRR + VPP router on host	FRR + VPP on a VM with 2x Dual Connect X6 (400Gbps)
WAN	CENIC 4×400G + 280G	OASIS 760G to ESnet, CENIC	ESnet Chicago, 100G	FABRIC facility ports
Endpoints	8 DTNs, dual-homed	16 Kubernetes nodes	6 DTNs, bonded NICs	XRootD VMs in NRP node
Routing	BGP v4+v6 in a VRF	BGP v4+v6	BGP v6	BGP v6
IPv6 /64s	16	12	4	2
Host QoS	private namespace	private namespace/host level	private namespace	host level
Telemetry	SNMP from the FE	external SNMP + Thanos	VPP exporter	VPP exporter

Same SiteRM code and the same API to SENSE-O; **everything site-specific lives in config**.

Caltech: four vendors, dual-homed DTNs

DEVICES

Dell OS 9, Arista EOS, 2× Dell OS10 + 1× SONiC (Edgecore)

WAN

100G Prod, 4×400G + 280G to CENIC in Los Angeles; 400G to Caltech's SDN Testbed

DTNS

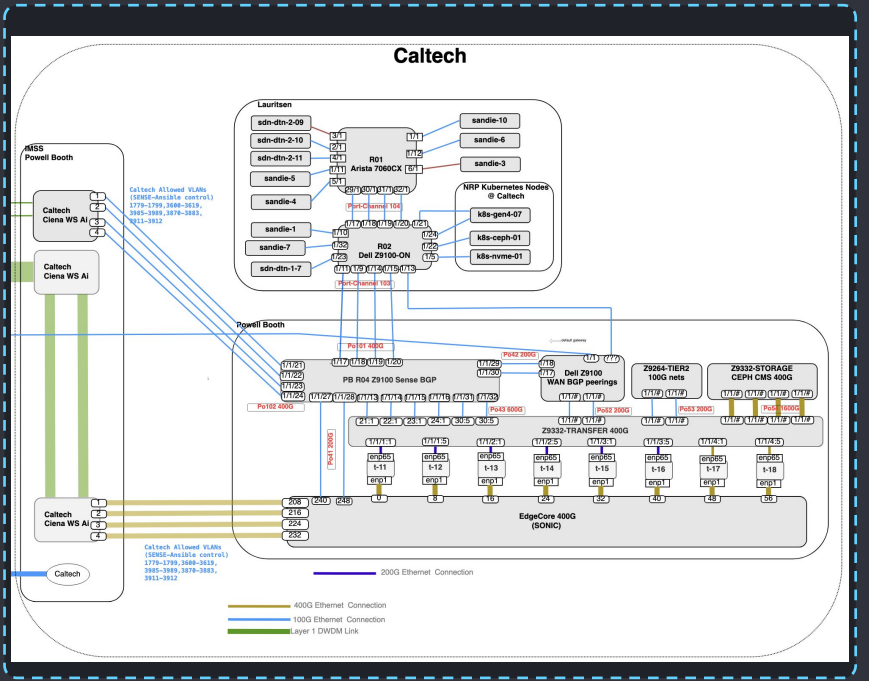
8 agents, dual-homed: a bond to Dell plus a 400G NIC to SONiC (+14 SDN Testbed)

ROUTING

BGP v4+v6 in a VRF; 16 IPv6 /64s, one per XRootD redirector (x2 SDN Separate)

CONFIG SNIPPET

```
dellos10_s0:
  rsts_enabled: ipv4,ipv6
  vrf: Vrf_sense01
  ports:
    Port-channel 102:
      capacity: 280000
      isAlias: …cenic.net…:Port-Channel501
```



ESnet

SDSC / NRP: Kubernetes nodes behind SONiC

DEVICES

2× SONiC (Mellanox SN3700, Edgecore) on a 1.2 Tbps port-channel

WAN

760G to the OASIS aggregation: ESnet, CENIC, FABRIC, PRISM, UCSD

NODES

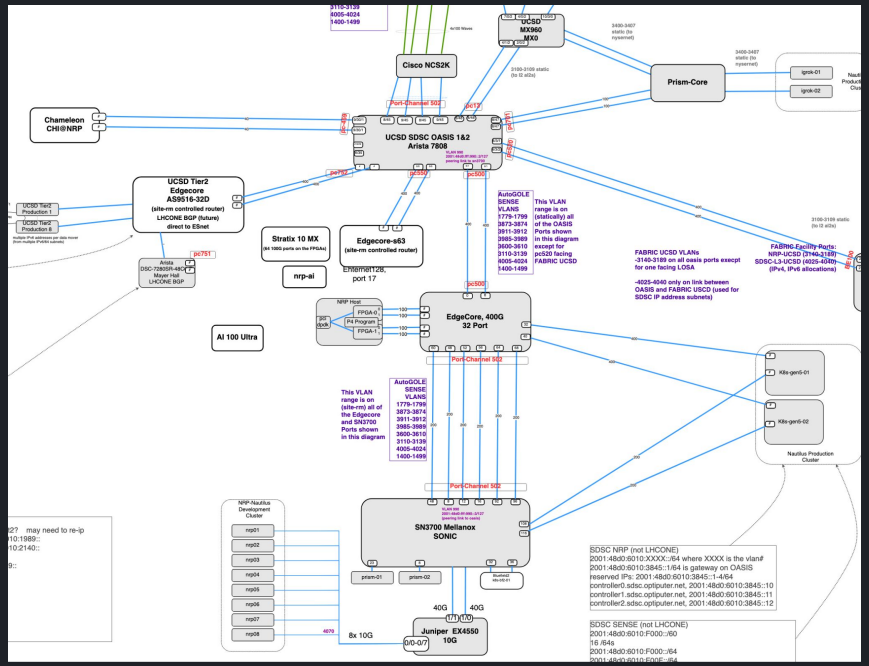
16 NRP Kubernetes nodes with 1 to 5 SENSE NICs each

TELEMETRY

Switch SNMP from a separate exporter; host metrics from NRP's Thanos

CONFIG SNIPPET · ONE DTN

```
general:
  prometheus_federate: http://thanos-federate…
enp65s0f1np1:
  switch: edgecore_s0
  port: Ethernet128
qos:
  policy: privatens
```



ESnet

UChicago: a software router into ESnet

ROUTER

FRR + VPP software router, managed by SiteRM like any switch

WAN

One 100G port straight to ESnet's Chicago router, no regional hop

DTNS

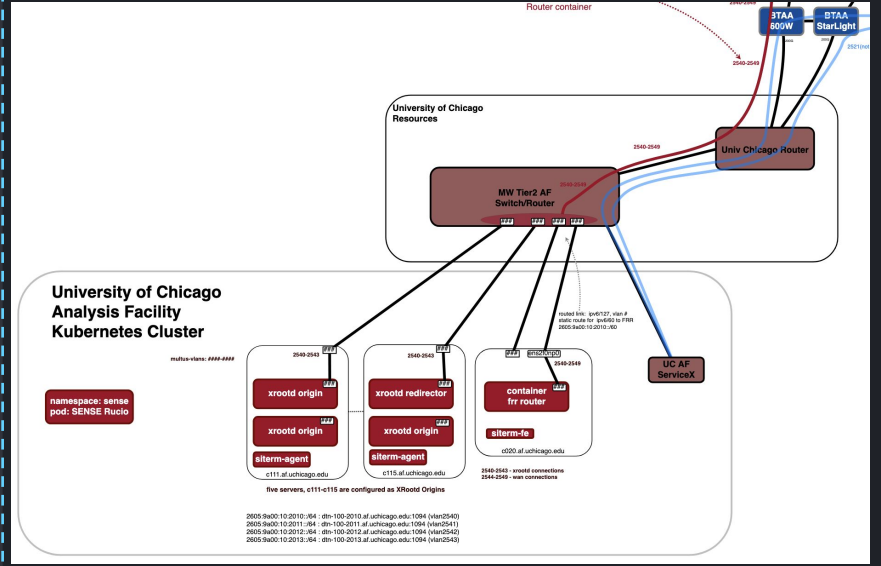
6 agents with bonded NICs behind a switch SiteRM does not manage

QOS

Host tc in private namespaces, one VLAN interface per IPv6 /64

CONFIG SNIPPET · ONE DTN

```
qos:
  policy: privatens
  interfaces:
    vlan.2540:
      ipv6_range: 2001:db8:10:2010::/64
      master_intf: bond0
```



ESnet

CERN via FABRIC/ESnet/NRP:

ROUTER

FRR + VPP on a FABRIC VM with dedicated NICs; no physical switch control

PORTS

4× 100G + 1G FABRIC facility ports, aliased into FABRIC's model

HOSTS

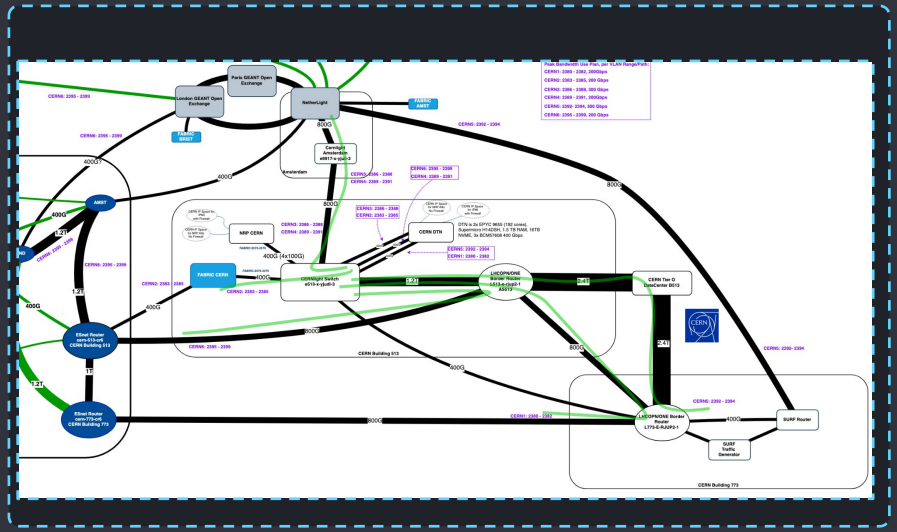
XRootD hosts on sub-interfaces on NRP Node

ROUTING

BGP v6 with a private ASN; one /64 per XRootD host

CONFIG SNIPPET

```
frr_s0:
  rsts_enabled: ipv6
  ports:
    e1:
      capacity: 100000
      isAlias: …stack-fabric…:node+CERN:port+…
```



ESnet

Six lessons/principles from day-to-day operations

01

Don’t trust the model

Check the device every cycle: oper status, LLDP, BGP prefixes.

02

Codes before dashboards

Stable numeric error codes made cross-site alerting possible.

03

Debounce before alerting

Multiple repeats before a warning; stale devices re-scanned after an hour.

04

Up does not mean it is working

Green services and established sessions still need pings and end-to-end tests.

05

Keep the site in full control

Narrow scope, reviewed config and documented commands speed up approvals.

06

Need to tune the app

A bandwidth guarantee is not a throughput guarantee; transfers, dtns, network, storage need tuning.

Where SiteRM runs today

14
sites: 6 LHC,
plus 4 non-LHC

150+
DTN endpoints with a
SiteRM Agent

22
SiteRM instances in the
SENSE-O model

15
R&E networks and
exchanges

LHC SITES	Fermilab (T1) - Caltech - UCSD - MWT2 - NET2 - Nebraska - Wisconsin* - CERN**
NON-LHC SITES	SDSC - National Research Platform at Internet2 PoPs - FABRIC (CERN) - StarLight
NETWORKS	ESnet - Internet2 - CENIC - GÉANT - SURF - CANARIE - Pacific Wave - AmPath - AtlanticWave-SDX - StarLight - MREN - CalIT2 PRISM - Univ. of Hawaii - FABRIC
DEVICES	Seven network OS families on site switches and routers (Dell, Arista, Cisco, Juniper, SONiC, FRR/VPP) + Nokia SR OS and Mellanox soon

* – Wisconsin In progress (install on their own)

** - CERN via NRP/Fabric and HAProxy/Nginx Davs testing. Adding CERN DTN to NRP (Scitags for SC26: Mariam Babik, Edoardo Martelli)

ESnet

Road to DC27: scaling across LHC sites

CMS Rucio + SENSE, proposed path

NOW

DMM instance under CMS-Rucio team ownership

END OF 2026

Patch the CMS Rucio integration instance

PRE-PRODUCTION

Demonstrate SENSE integration end to end

LATE JANUARY 2027

Patch CMS Rucio production before DC27

Timing subject to alignment with the CMS-Rucio team.

Scaling SiteRM

- More Sites, on production networks and servers
- Simplify onboarding: one config PR, Helm or Docker, documented device permissions
- New platforms: Nokia SR OS, Mellanox OS
- Close QoS gaps: SONiC, Junos MX/PTX, container namespaces
- Explore SmartNIC/DPU offload for priority routing and QoS
- Dynamic allocation of bandwidth on a given port based on group or project (multi-vo fairshare)
- L3VPN with-in ESnet (control BGP Routes on ESnet). Last mile issue remains.

ESnet

THANK YOU

Questions?

DOCS AND RELEASES

sdn-sense.github.io

SENSE

sense.es.net · sense-info@es.net

CODE

github.com/sdn-sense/siterm

PAPER

Software-Defined Network for End-to-end Networked Science at the Exascale doi.org/10.1016/j.future.2020.04.018

Thanks to the site, network and experiment teams at Fermilab, Caltech, UCSD, SDSC, NRP, UChicago, Wisconsin, MGHPCC, Nebraska, ESnet, Internet2, CENIC, GPN, FABRIC and StarLight.

ESnet
ENERGY SCIENCES NETWORK

Backup slides



Lifecycle: from intent to running config

1 accepted
SENSE-O submits a delta and PolicyService checks it. Not committed within 5 min → removed.

→

2 committed
Every domain on the path agrees and SENSE-O commits; resources are reserved.

→

3 activated
At start time the FE and the Agents apply config on switches and DTNs and report state.

→

4 removed
On cancel or end time, config removed from devices and hosts; VLANs and bandwidth are freed.

Checked before a delta is accepted

- The target host is alive and its Agent is running
- VLAN is in the allowed range and free for the whole time window
- IPs come from site pools, are unique, and are not already on a device
- IPv6 prefixes are no more specific than /64
- Guaranteed bandwidth fits the port's remaining capacity

⚠ **When it fails**

The delta moves to **failed** with an exc code. SENSE-O, the Web UI and Grafana all see the same code.

ESnet

Smoke tests catch problems before data does



Every 12 hours
All pairs of production endpoints. 1 Gbps guaranteedCapped with IPv6 on both ends; 40 minutes allowed per step.

Failures stay visible
A failed instance stays in SENSE-O for debugging and its pair is skipped until resolved. “No path found” is retried next cycle.

Results in Grafana
Outcome per step, ping loss and RTT per pair; dashboards for ESnet, Internet2, FABRIC, Caltech and AmPath.

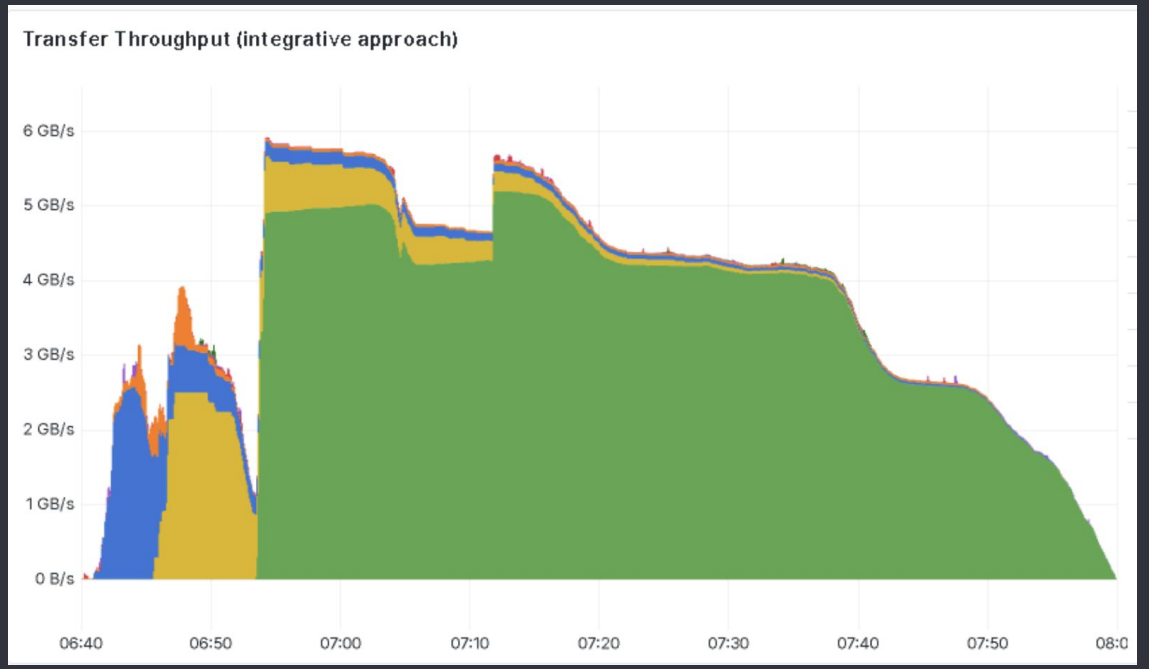
Plus: **Rucio/DMM** injects short-lived SENSE rules every 8 h on the CMS testbed, and **smoke tests** transfers

ESnet

From this month Mini-Data Challenges

- Oct 14 - French T2: IN2P3-CPPM
- Multi VO Test: ATLAS and LHCb
- Results:
  - ATLAS: 0.3 Gbps
  - LHCb: 3.1 Gbps

Seems to have been caused by having very different FTS limits on each instance

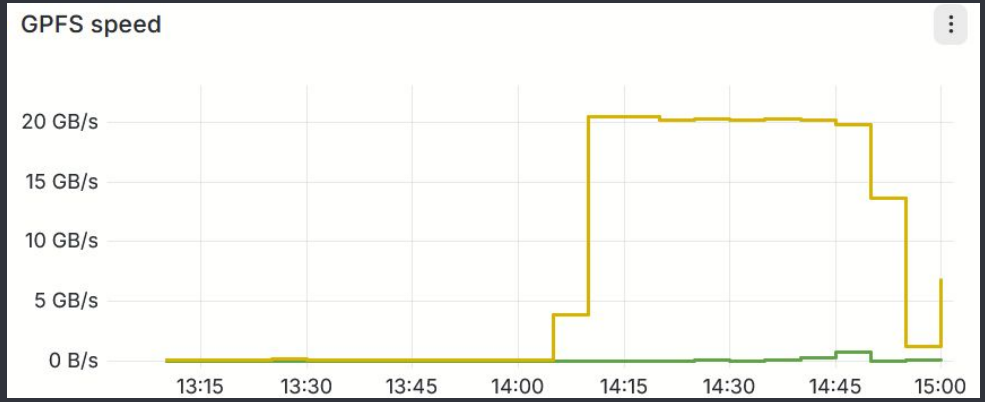


Writes to IN2P3-CPPM: Green - LHCb Yellow -ATLAS

From this month Mini-Data Challenges (cont'd)

- Oct 19 - CERN to CNAF
- Multi VO Test: ATLAS, CMS and LHCb
- Targets:
  - ATLAS: 32 Gbps
  - CMS: 32Gbps
  - LHCb: 16 Gbps

A mistake on the CMS injection caused the whole bandwidth to be used by CMS.

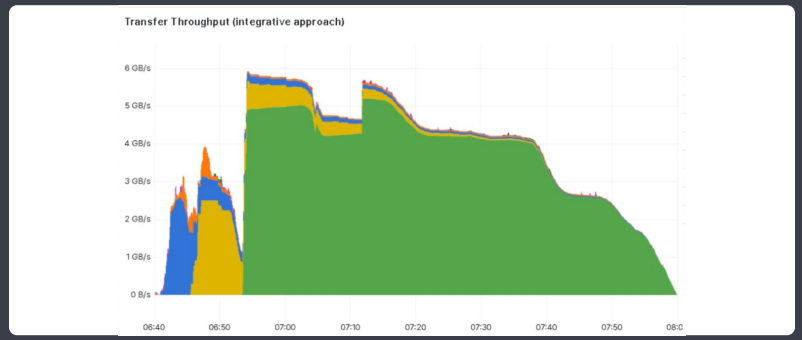


Writes to CNAF: Green - ATLAS , Yellow - CMS

Mini-DC, October: avoiding VO starvation

OCT 14 · WRITES TO IN2P3-CPPM

ATLAS + LHCb



Green: LHCb · Yellow: ATLAS

WHAT HAPPENED

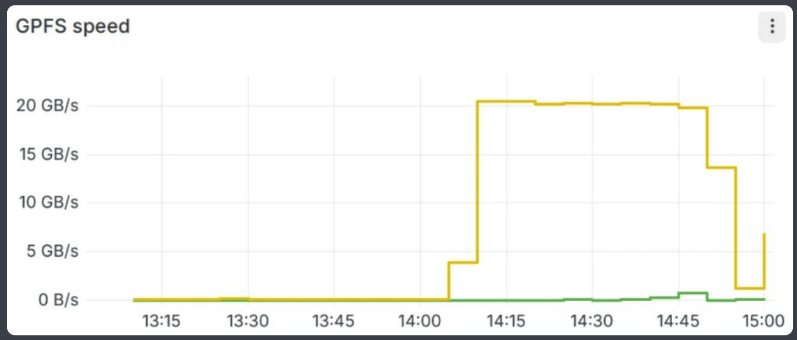
Very different FTS limits per instance: ATLAS got 0.3 Gbps, LHCb 3.1 Gbps.

WITH SITERM QOS

ATLAS's /64 keeps its guaranteed floor; FTS settings no longer decide who gets the link.

OCT 19 · CERN TO CNAF

ATLAS 32 · CMS 32 · LHCb 16 Gbps targets



Green: ATLAS · Yellow: CMS

WHAT HAPPENED

A CMS injection mistake took the whole link, about 20 GB/s; ATLAS was starved.

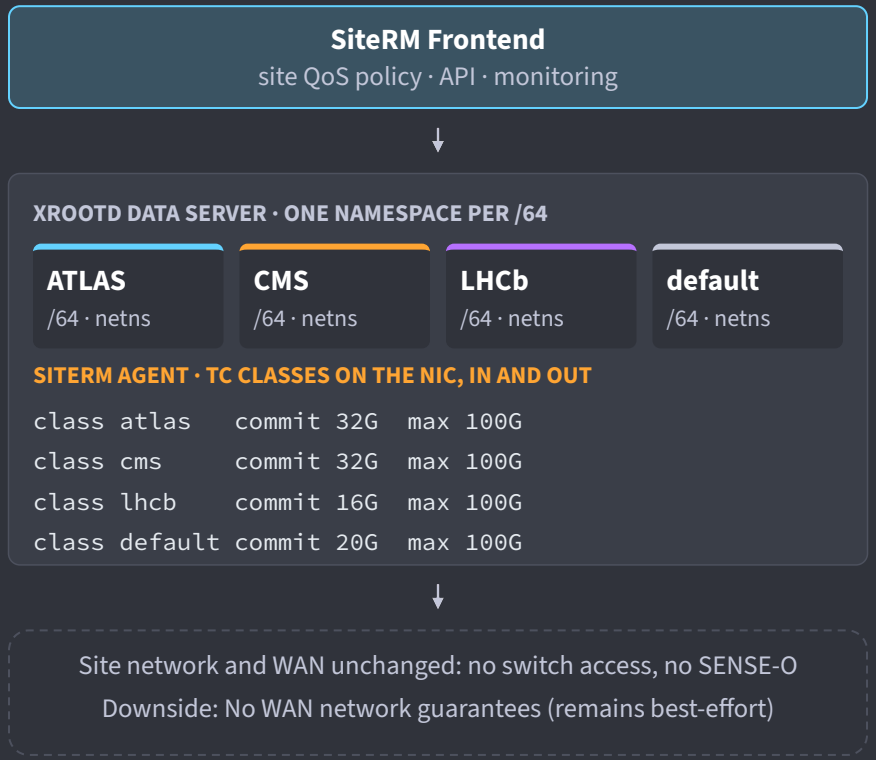
WITH SITERM QOS

CMS is held to its share while ATLAS is active, and still uses whatever is idle.

Charts: Mini-DC reports, October 2026 [add reference]

ESnet

SiteRM alone: QoS without network control



1. **Split by subnet**
   One XRootD/CmsD instance per VO or priority class, each in its own namespace and IPv6 /64 (or Kubernetes + Multus).
2. **Deploy only SiteRM**
   Frontend plus Agents on the data servers. No switch credentials, no Orchestrator, no circuits.
3. **The site sets the policy**
   Per /64: a guaranteed floor (commit) and a ceiling (max). One /64 stays the default class.
4. **The Agent enforces it**
   Linux tc (HTB via FireQOS) on the NIC, both directions. Idle capacity is borrowed, so nothing is wasted.
5. **Change it live**
   SiteAdmin, Operators or DMM shift priorities at runtime. Add SENSE network control later, without redeploying.

ESnet