tnc23

DIGITAL GENERATIONS

TIRANA, ALBANIA | 5-9 JUNE 2023

# Managed Network Services for Large Data Transfers

## SENSE Orchestration

**Tom Lehman (ESnet), Chin Guok (ESnet)**

TNC23

June 6, 2023

GÉANT

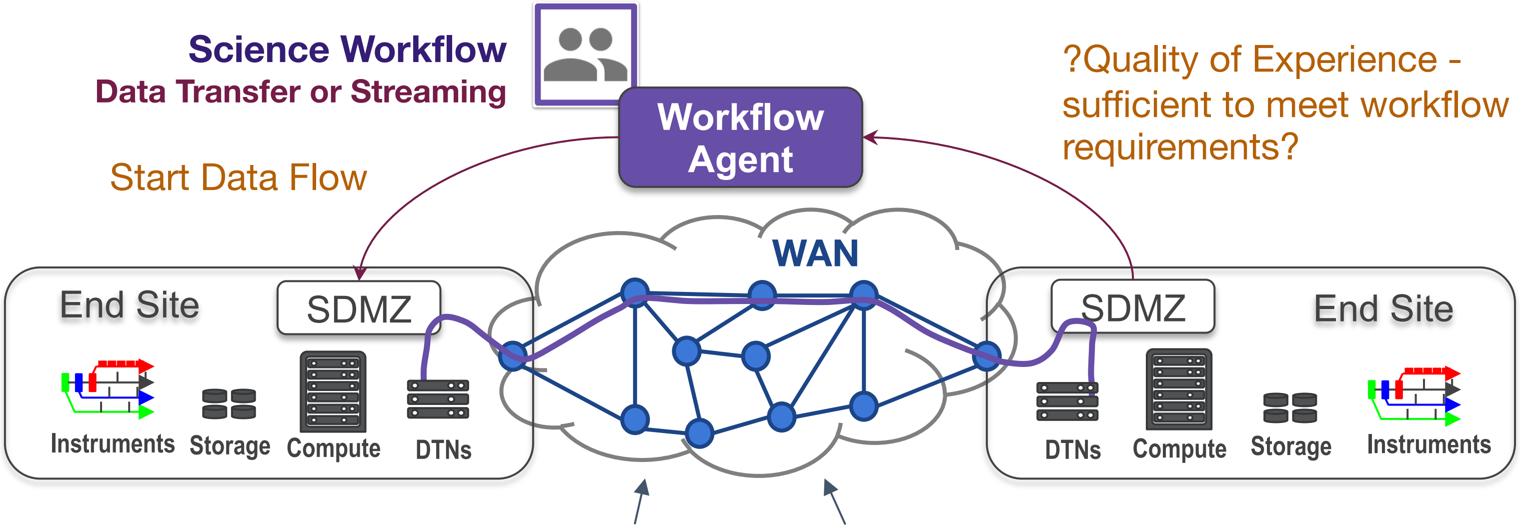
Co-funded by
the European Union

# Presentation Outline

- Multi-Domain, Multi-Resource Service Orchestration
  - objectives, issues, approach
- SENSE Orchestration System, Architecture, Implementation
- SENSE Orchestration services for Rucio/FTS/XRootD Data Movement and Management System
  - with a focus on LHC CMS workflows
- Next Steps

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

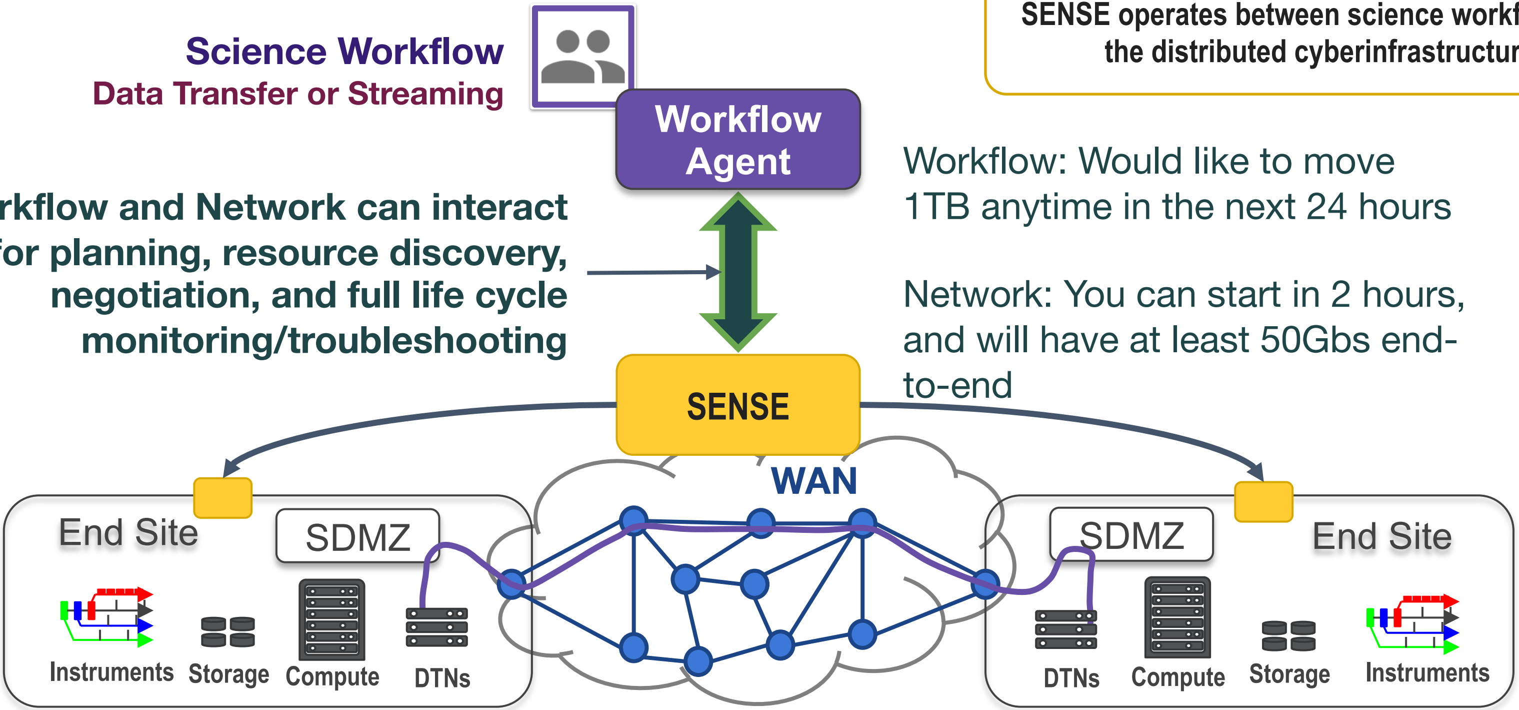
- No realtime per flow data available for planning or monitoring
- No "deterministic" network services available
- Start data flow, and hope for the best



Excellent Information available about aggregated (over time and data flows) use of the network infrastructure

# Elevate Network to First Class Resource
# API driven Automation and Orchestration

SENSE operates between science workflow and the distributed cyberinfrastructure



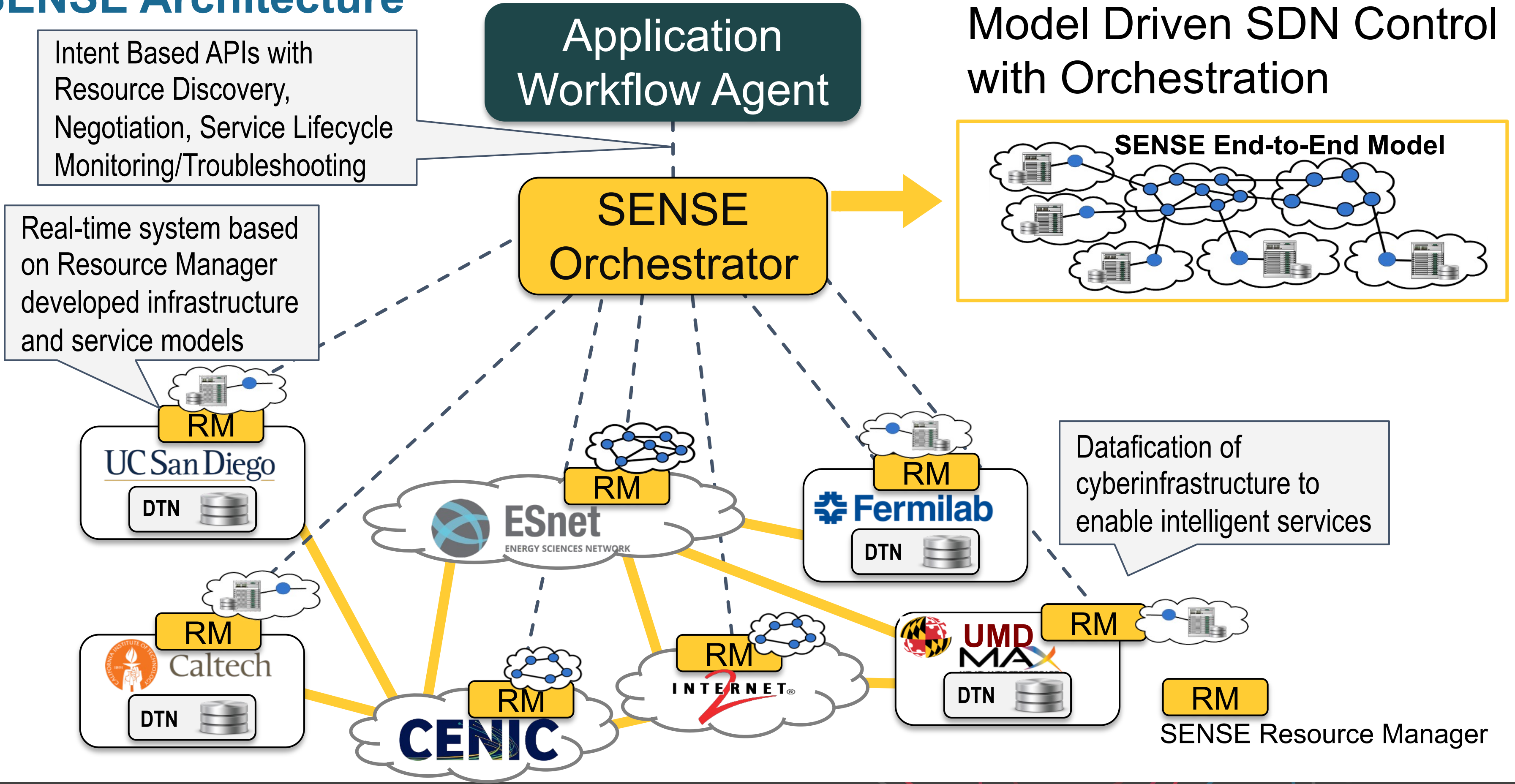
Workflow and Network can interact for planning, resource discovery, negotiation, and full life cycle monitoring/troubleshooting

Workflow: Would like to move 1TB anytime in the next 24 hours

Network: You can start in 2 hours, and will have at least 50Gbs end-to-end

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

# SENSE Architecture



Model Driven SDN Control with Orchestration

Intent Based APIs with Resource Discovery, Negotiation, Service Lifecycle Monitoring/Troubleshooting

Real-time system based on Resource Manager developed infrastructure and service models

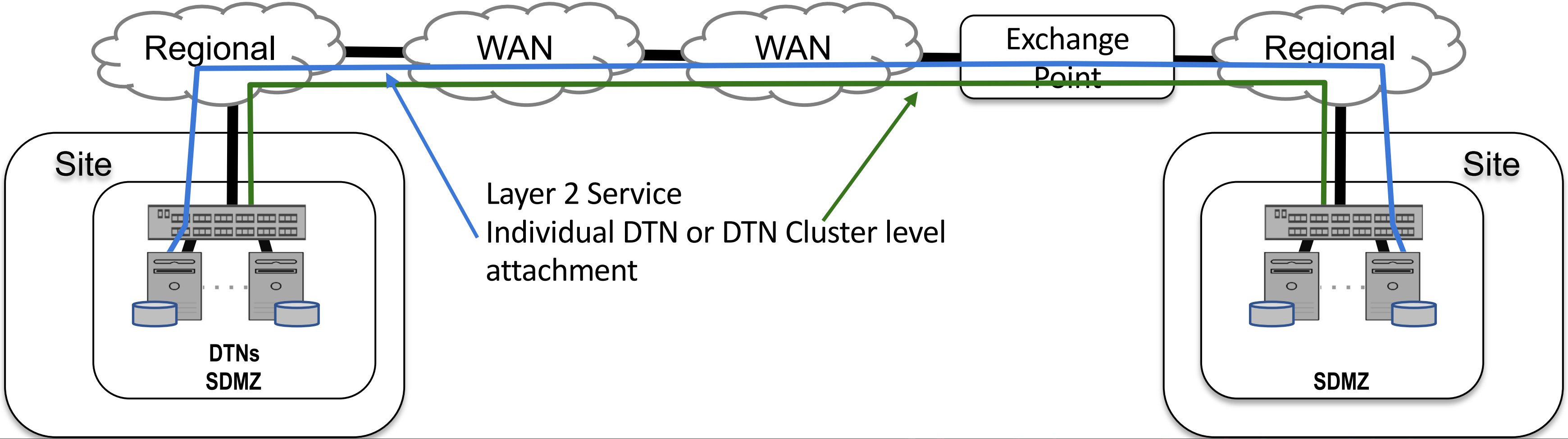
Datafication of cyberinfrastructure to enable intelligent services

# SENSE Solution Approach – Application Interactions

- **Intent Based** – Abstract requests and questions in the context of the application objectives.
- **Interactive** – What is possible? What is recommended? Let’s negotiate.
- **Real-time** – Resource availability, provisioning options, service status, troubleshooting.
- **End-to-End** – Multi-domain networks, end sites, and the network stack inside the end systems.
- **Full Service Lifecycle Interactions** – Continuous conversation between application and network for the service duration.

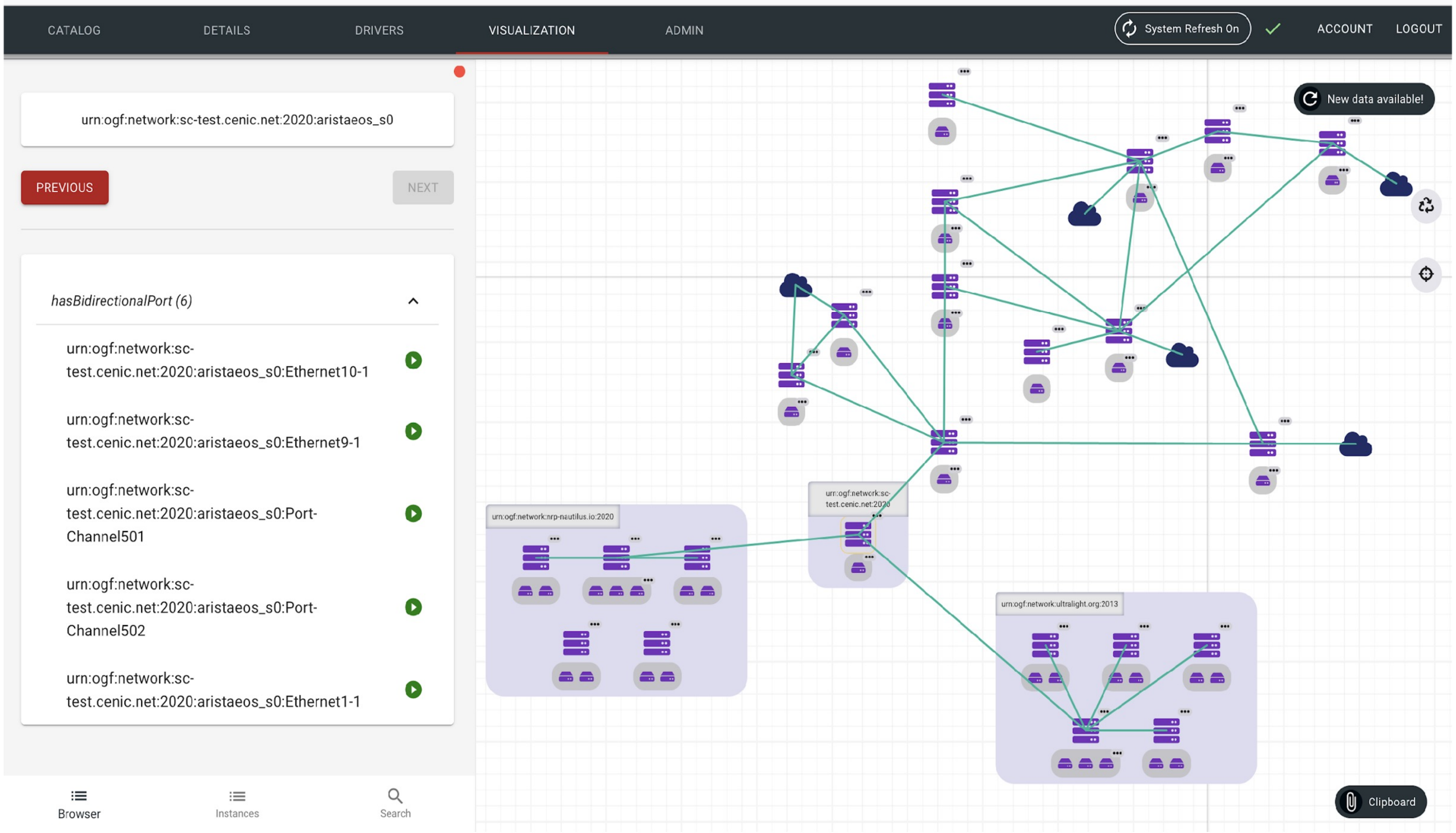
# SENSE Services

- **Orchestration** (of other domain owned systems)
- **Multi-Resource** (networks, end systems, instruments, clouds)
- **Multi-Domain** (Sites, Regionals, WANs, Exchange Points)
- **Multi-Service** (L2 Point-to-Point, L2 MultiPoint, L3VPN, QoS, Traffic engineered paths)
- **Intelligent Services** (realtime interaction, full-lifecycle monitoring)



Layer 2 Service
Individual DTN or DTN Cluster level attachment

# SENSE - Model based Resource Descriptions



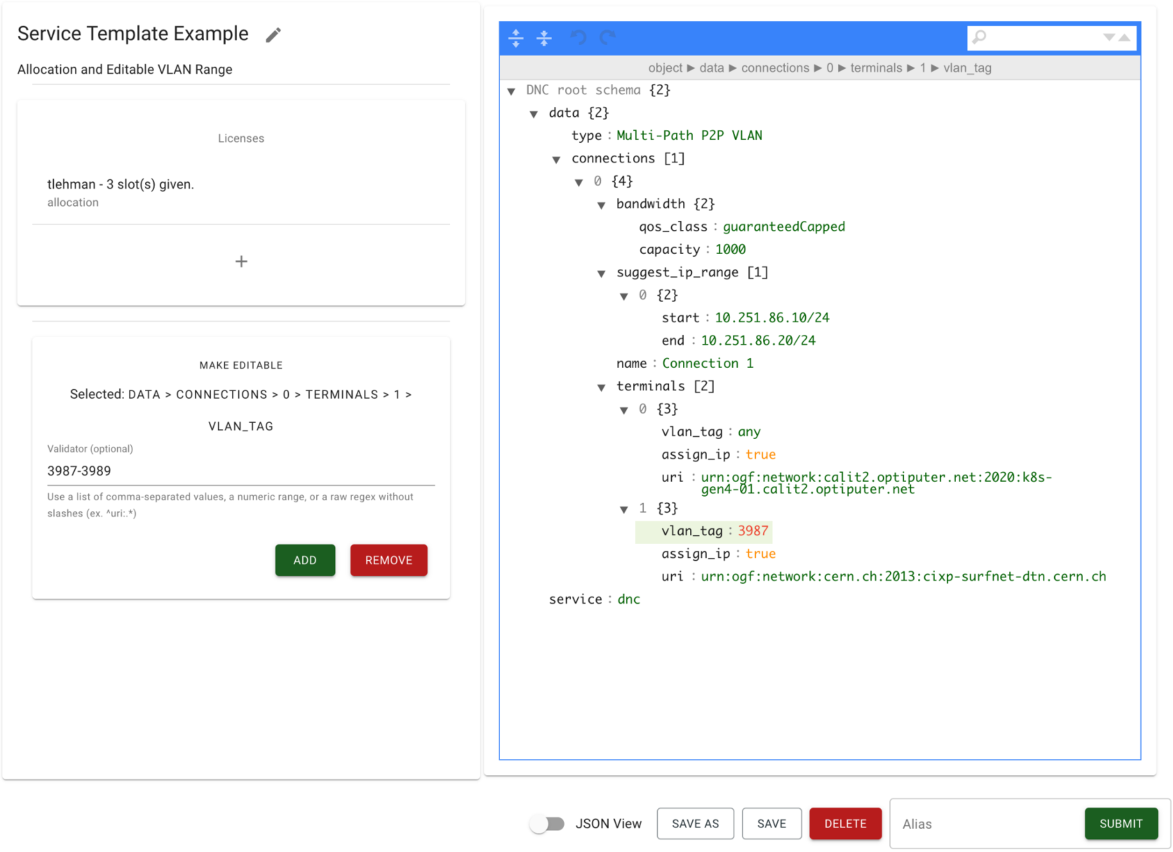
urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0

*hasBidirectionalPort (6)*

- urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0:Ethernet10-1
- urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0:Ethernet9-1
- urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0:Port-Channel501
- urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0:Port-Channel502
- urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0:Ethernet1-1

# SENSE - Model based Resource Descriptions

- Read only and optionally with user editable parameters
- Allows user to run with one time "ticket" or multiple time-use allocations



Service Template Example

Allocation and Editable VLAN Range

Licenses

tlehman - 3 slot(s) given.
allocation

+

MAKE EDITABLE

Selected: DATA > CONNECTIONS > 0 > TERMINALS > 1 > VLAN_TAG

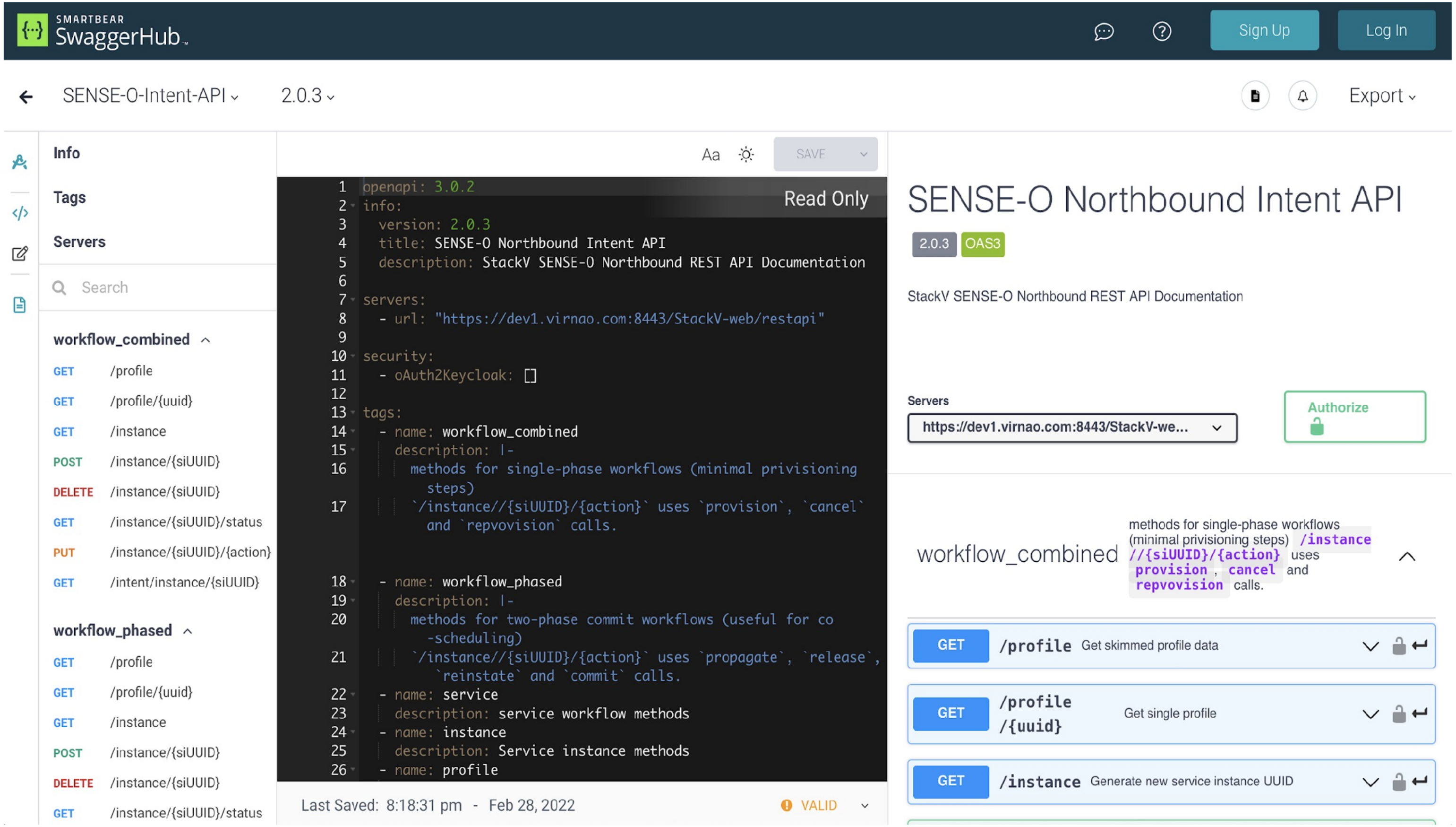
Validator (optional)
3987-3989

Use a list of comma-separated values, a numeric range, or a raw regex without slashes (ex. ^uri:.*)

object ▶ data ▶ connections ▶ 0 ▶ terminals ▶ 1 ▶ vlan_tag

```
▼ DNC root schema {2}
  ▼ data {2}
      type : Multi-Path P2P VLAN
    ▼ connections [1]
      ▼ 0 {4}
        ▼ bandwidth {2}
            qos_class : guaranteedCapped
            capacity : 1000
        ▼ suggest_ip_range [1]
          ▼ 0 {2}
              start : 10.251.86.10/24
              end : 10.251.86.20/24
          name : Connection 1
        ▼ terminals [2]
          ▼ 0 {3}
              vlan_tag : any
              assign_ip : true
              uri : urn:ogf:network:calit2.optiputer.net:2020:k8s-
                    gen4-01.calit2.optiputer.net
          ▼ 1 {3}
              vlan_tag : 3987
              assign_ip : true
              uri : urn:ogf:network:cern.ch:2013:cixp-surfnet-dtn.cern.ch
    service : dnc
```

# SENSE - Northbound API



workflow_combined

| | |
|---|---|
| GET | /profile |
| GET | /profile/{uuid} |
| GET | /instance |
| POST | /instance/{siUUID} |
| DELETE | /instance/{siUUID} |
| GET | /instance/{siUUID}/status |
| PUT | /instance/{siUUID}/{action} |
| GET | /intent/instance/{siUUID} |

workflow_phased

| | |
|---|---|
| GET | /profile |
| GET | /profile/{uuid} |
| GET | /instance |
| POST | /instance/{siUUID} |
| DELETE | /instance/{siUUID} |
| GET | /instance/{siUUID}/status |

```
openapi: 3.0.2
info:
  version: 2.0.3
  title: SENSE-O Northbound Intent API
  description: StackV SENSE-O Northbound REST API Documentation

servers:
  - url: "https://dev1.virnao.com:8443/StackV-web/restapi"

security:
  - oAuth2Keycloak: []

tags:
  - name: workflow_combined
    description: |-
      methods for single-phase workflows (minimal privisioning
        steps)
      `/instance//{siUUID}/{action}` uses `provision`, `cancel`
        and `repvovision` calls.


  - name: workflow_phased
    description: |-
      methods for two-phase commit workflows (useful for co
        -scheduling)
      `/instance//{siUUID}/{action}` uses `propagate`, `release`,
        `reinstate` and `commit` calls.
  - name: service
    description: service workflow methods
  - name: instance
    description: Service instance methods
  - name: profile
```

Last Saved: 8:18:31 pm - Feb 28, 2022 VALID

## SENSE-O Northbound Intent API

2.0.3 OAS3

StackV SENSE-O Northbound REST API Documentation

Servers

https://dev1.virnao.com:8443/StackV-we...

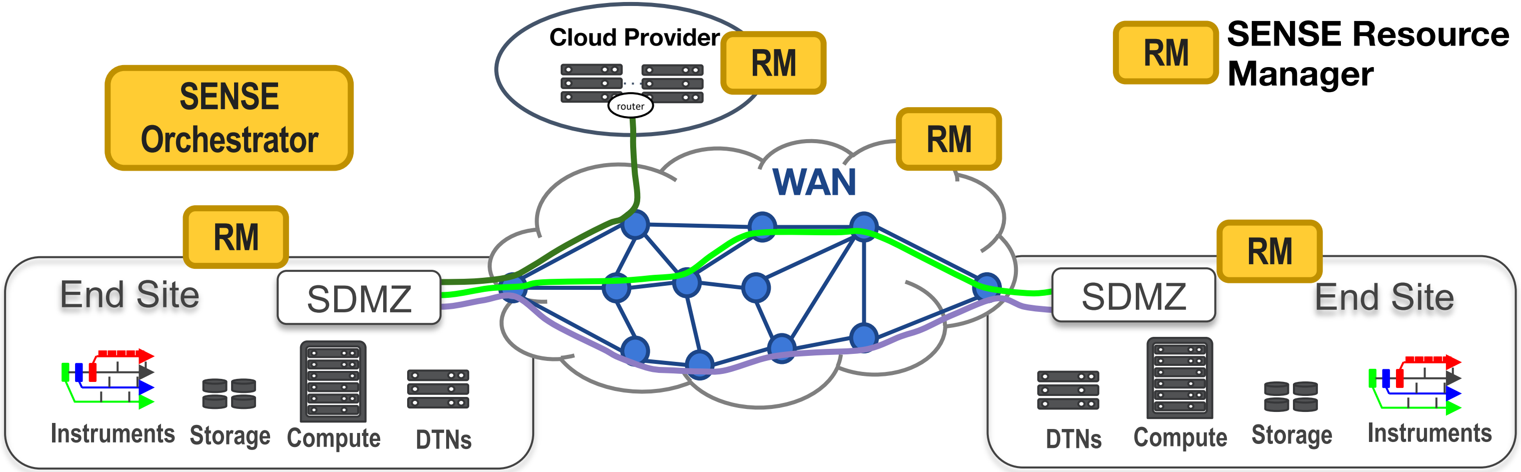
Authorize

workflow_combined — methods for single-phase workflows (minimal privisioning steps) `/instance//{siUUID}/{action}` uses `provision` , `cancel` and `repvovision` calls.

| | | |
|---|---|---|
| GET | /profile | Get skimmed profile data |
| GET | /profile/{uuid} | Get single profile |
| GET | /instance | Generate new service instance UUID |

# Multi-Resource Orchestration

- Networks, End-Systems, Cloud Resources, Instruments
- No need to manage/orchestrate all of the resources end-to-end, just the ones that matter
  - congestion, performance, or policy reasons

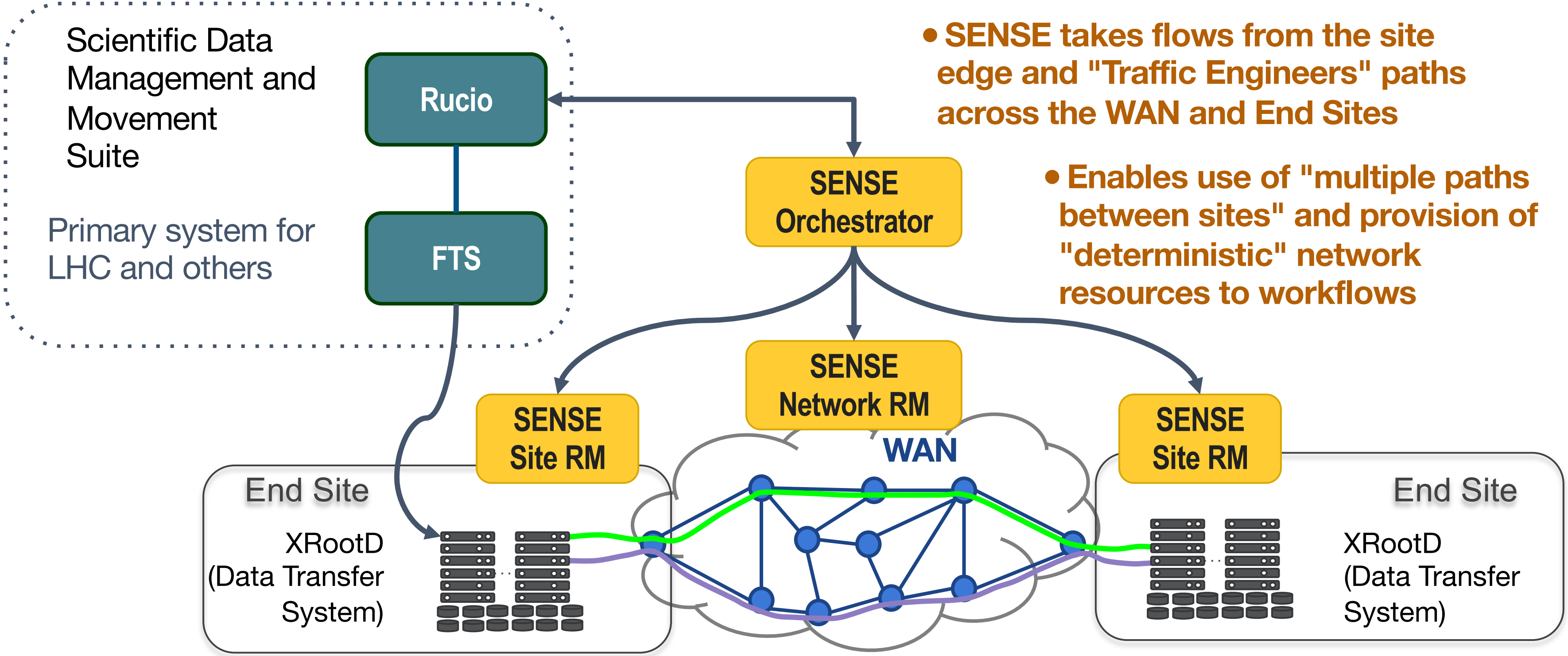


Cloud provider resources and connections

Traffic Engineered End-to-End Paths

# SENSE and Rucio/FTS/XRootD Interoperation

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



Scientific Data Management and Movement Suite
Primary system for LHC and others
Rucio
FTS
SENSE Orchestrator
SENSE Network RM
SENSE Site RM
SENSE Site RM
WAN
End Site
XRootD (Data Transfer System)
End Site
XRootD (Data Transfer System)

- 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

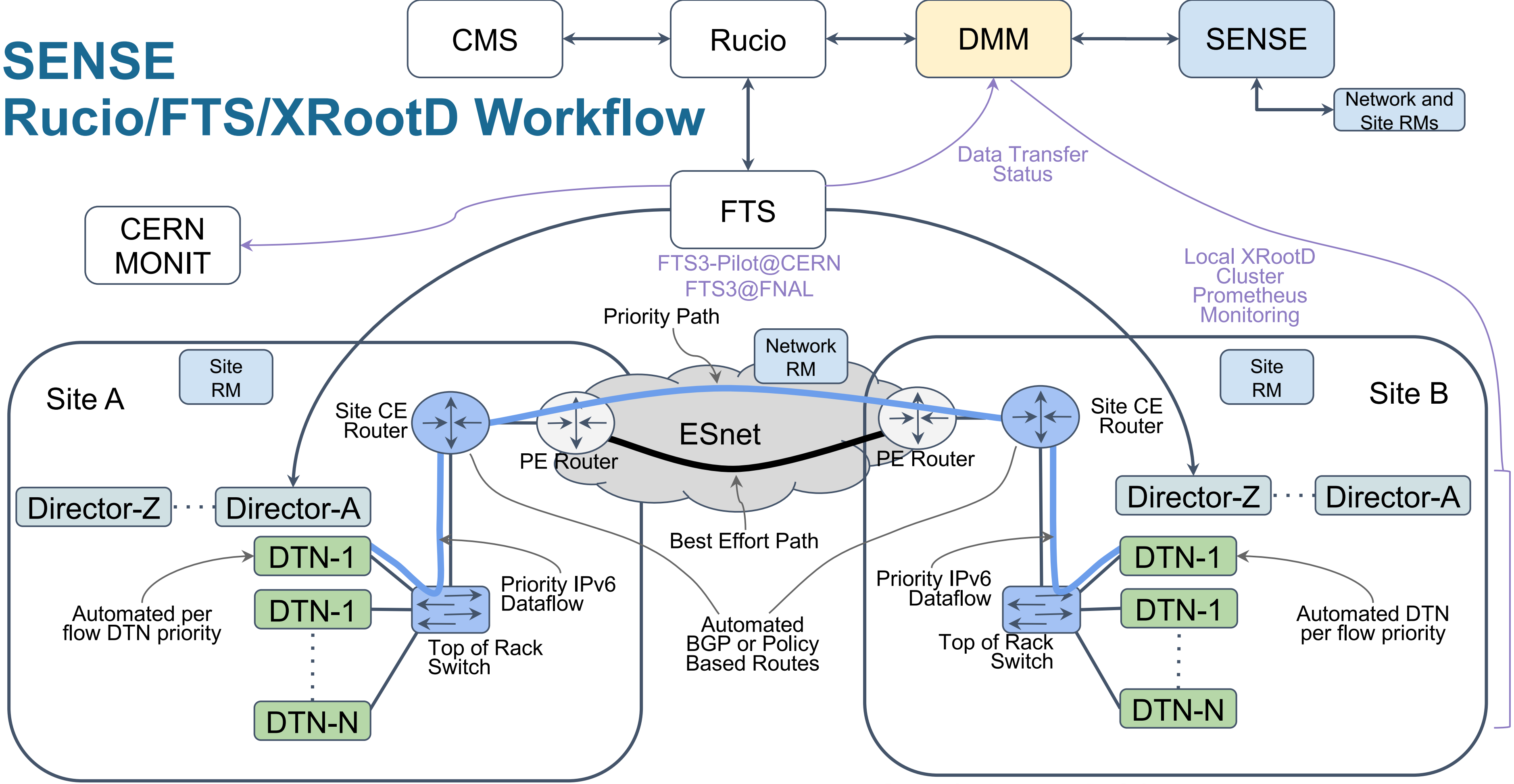
*Overall objective is to develop a better way to manage CMS transfers*

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

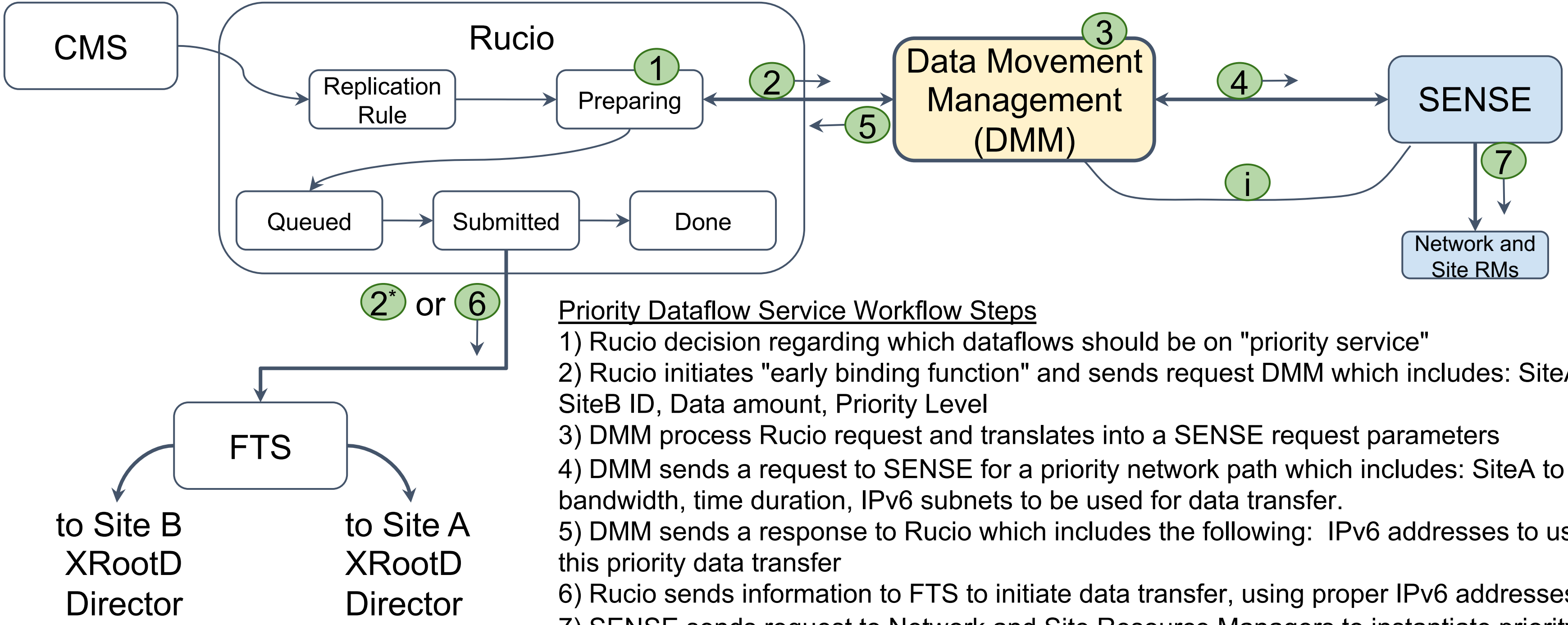
*Focus on the largest flows (not ALL transfers)*

*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

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

# 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

# Rucio, DMM, SENSE Workflow

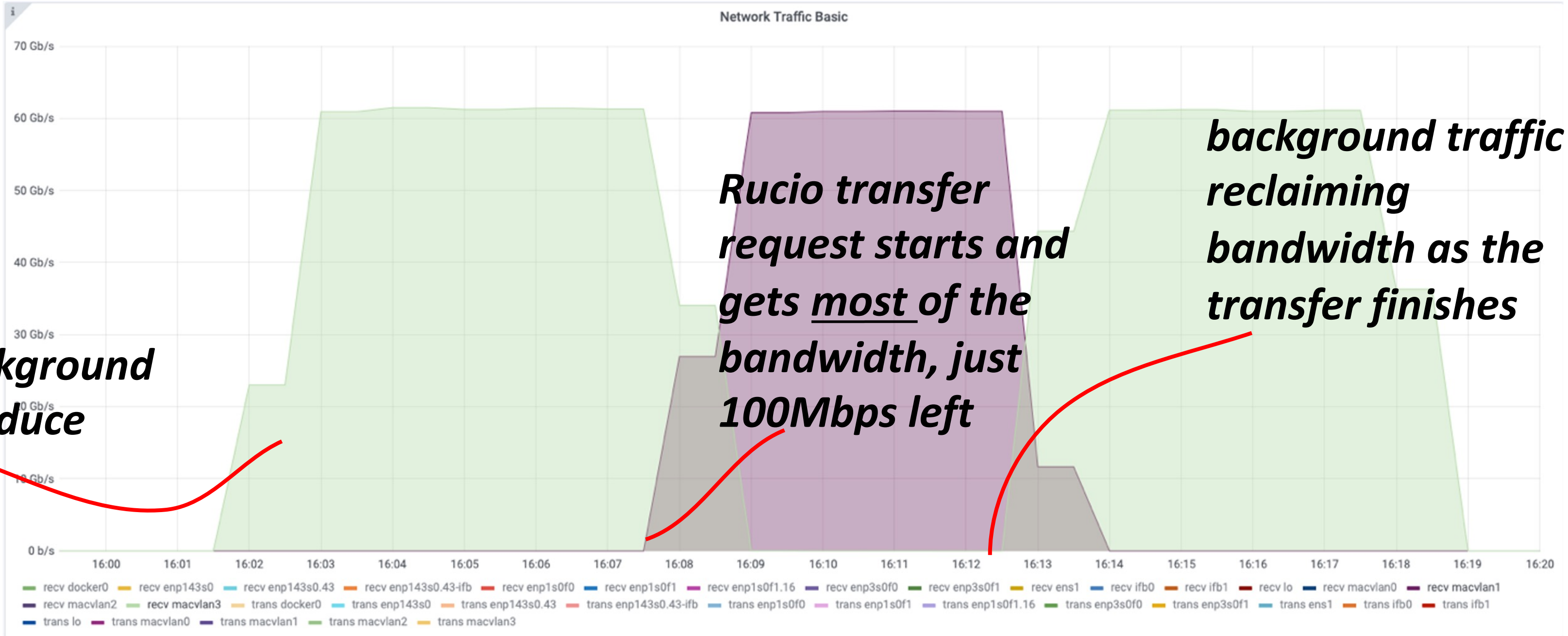
- A “priority” data flow is a flexible concept, and could be:
  - all data between Site A and Site B for a specific time period
  - all data between Site A and Site B on a specific IPv6 subnet
  - almost anything based on Site and IPv6/subnet parameters
- End-to-End Data Transfer monitoring
  - Performance 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
- Correlate data transfer layer throughput with network utilization
  - Still working on the details of data collection, storage, and correlation/analysis

# Proof of Concept Testing

Currently working toward ~400 Gbps site-to-site. Only a few hosts needed for these rates.



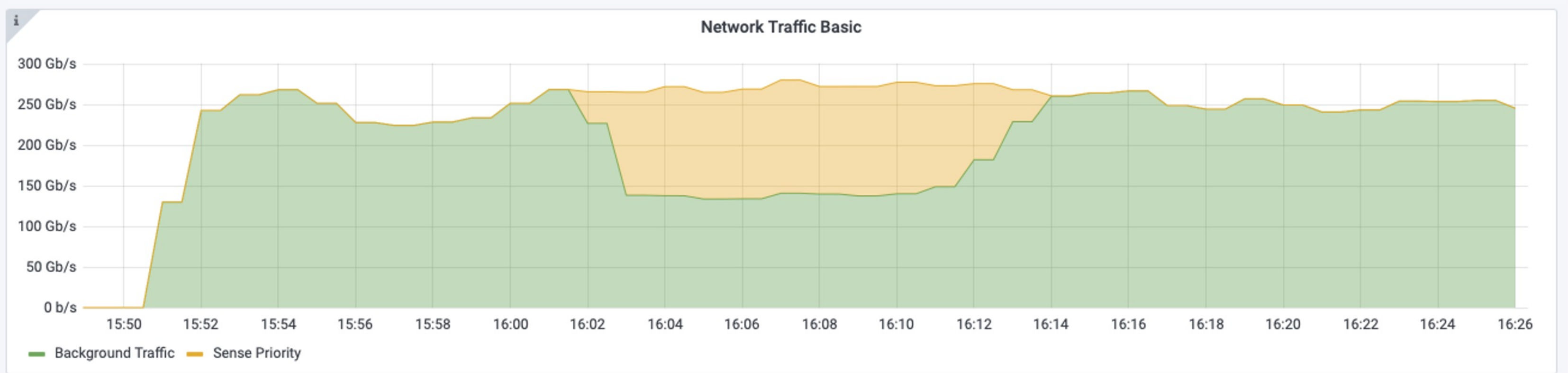
**60 Gbps**

***Artificial background traffic to produce congestion***

***Rucio transfer request starts and gets <u>most</u> of the bandwidth, just 100Mbps left***

***background traffic reclaiming bandwidth as the transfer finishes***

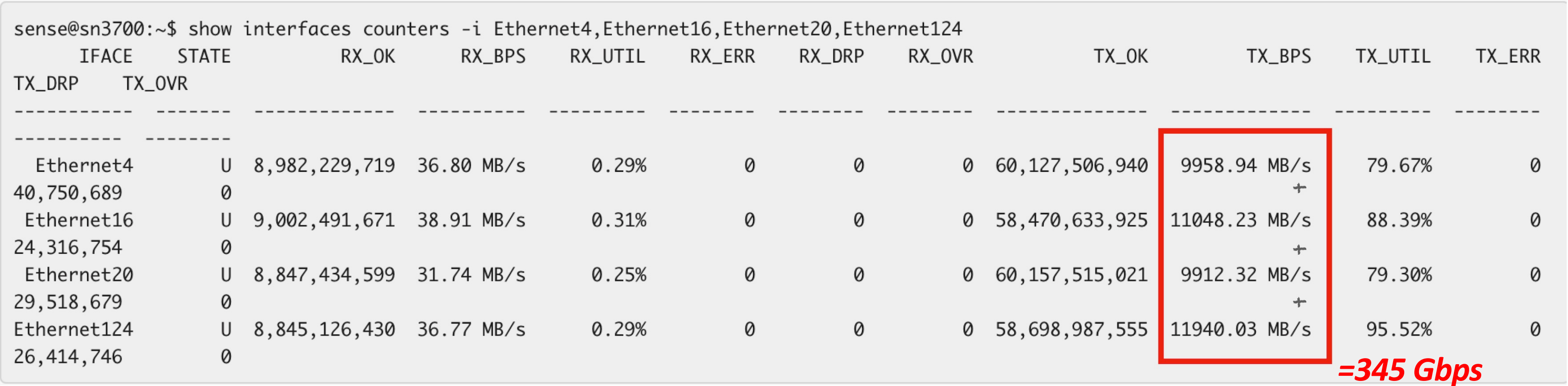
# UCSD to Caltech Testing at higher speeds



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

# 400Gbps Benchmark of XRootD- HTTP third-party-copy Transfers

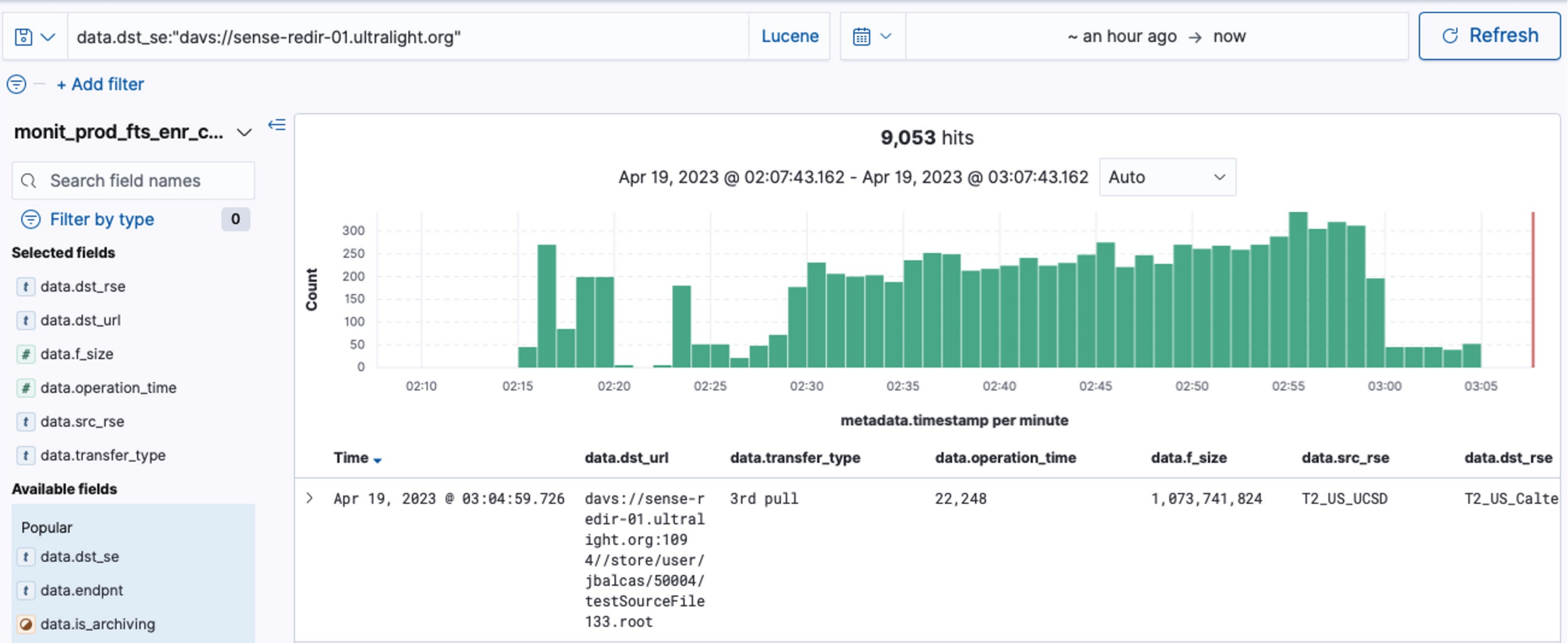
- We can reach ~400* Gbps and sustain it for hours! (345 Gbps over a network path capable of doing 350 Gbps). Using 40 streams of 1 GB files for each of the 13 servers with Caltech as sink, i.e. 520 streams coming out of UCSD
- XRootD-HTTP is capable of supporting the high throughputs required for the HL-LHC era
- Systematically running transfers can enable us to parameterize by number of CPU cores, number of streams, etc. Need at least $\mathcal{O}(10)$ streams per XRootD instance for ideal throughput.
- Use of redirectors does not affect performance. Choice of transfer tool does affect throughput.
- Reference UCSD, Caltech team presentation for more details:
    - CHEP23, https://indico.jlab.org/event/459/contributions/11303/



```
sense@sn3700:~$ show interfaces counters -i Ethernet4,Ethernet16,Ethernet20,Ethernet124
      IFACE    STATE          RX_OK      RX_BPS    RX_UTIL    RX_ERR    RX_DRP    RX_OVR           TX_OK         TX_BPS    TX_UTIL    TX_ERR
TX_DRP    TX_OVR
-----------  -------  -------------  ----------  ---------  --------  --------  --------  --------------  -------------  ---------  --------
----------  --------
  Ethernet4        U  8,982,229,719  36.80 MB/s      0.29%         0         0         0  60,127,506,940   9958.94 MB/s     79.67%         0
40,750,689         0                                                                                                 +
 Ethernet16        U  9,002,491,671  38.91 MB/s      0.31%         0         0         0  58,470,633,925  11048.23 MB/s     88.39%         0
24,316,754         0                                                                                                 +
 Ethernet20        U  8,847,434,599  31.74 MB/s      0.25%         0         0         0  60,157,515,021   9912.32 MB/s     79.30%         0
29,518,679         0                                                                                                 +
Ethernet124        U  8,845,126,430  36.77 MB/s      0.29%         0         0         0  58,698,987,555  11940.03 MB/s     95.52%         0
26,414,746         0
```

=345 Gbps

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



data.dst_se:"davs://sense-redir-01.ultralight.org"

9,053 hits

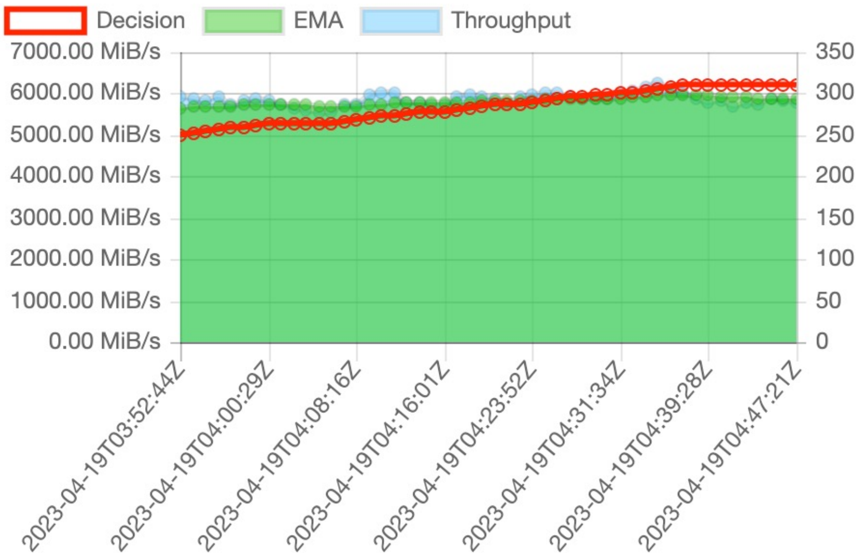
Apr 19, 2023 @ 02:07:43.162 - Apr 19, 2023 @ 03:07:43.162

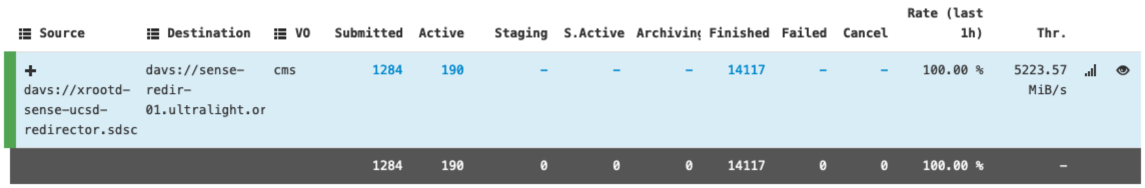
| Time | data.dst_url | data.transfer_type | data.operation_time | data.f_size | data.src_rse | data.dst_rse |
|---|---|---|---|---|---|---|
| Apr 19, 2023 @ 03:04:59.726 | davs://sense-redir-01.ultralight.org:1094//store/user/jbalcas/50004/testSourceFile133.root | 3rd pull | 22,248 | 1,073,741,824 | T2_US_UCSD | T2_US_Calte |

# 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.
- Working to increase transfer rates
- Including a dynamic way to control submission rate (FTS to XRootD)

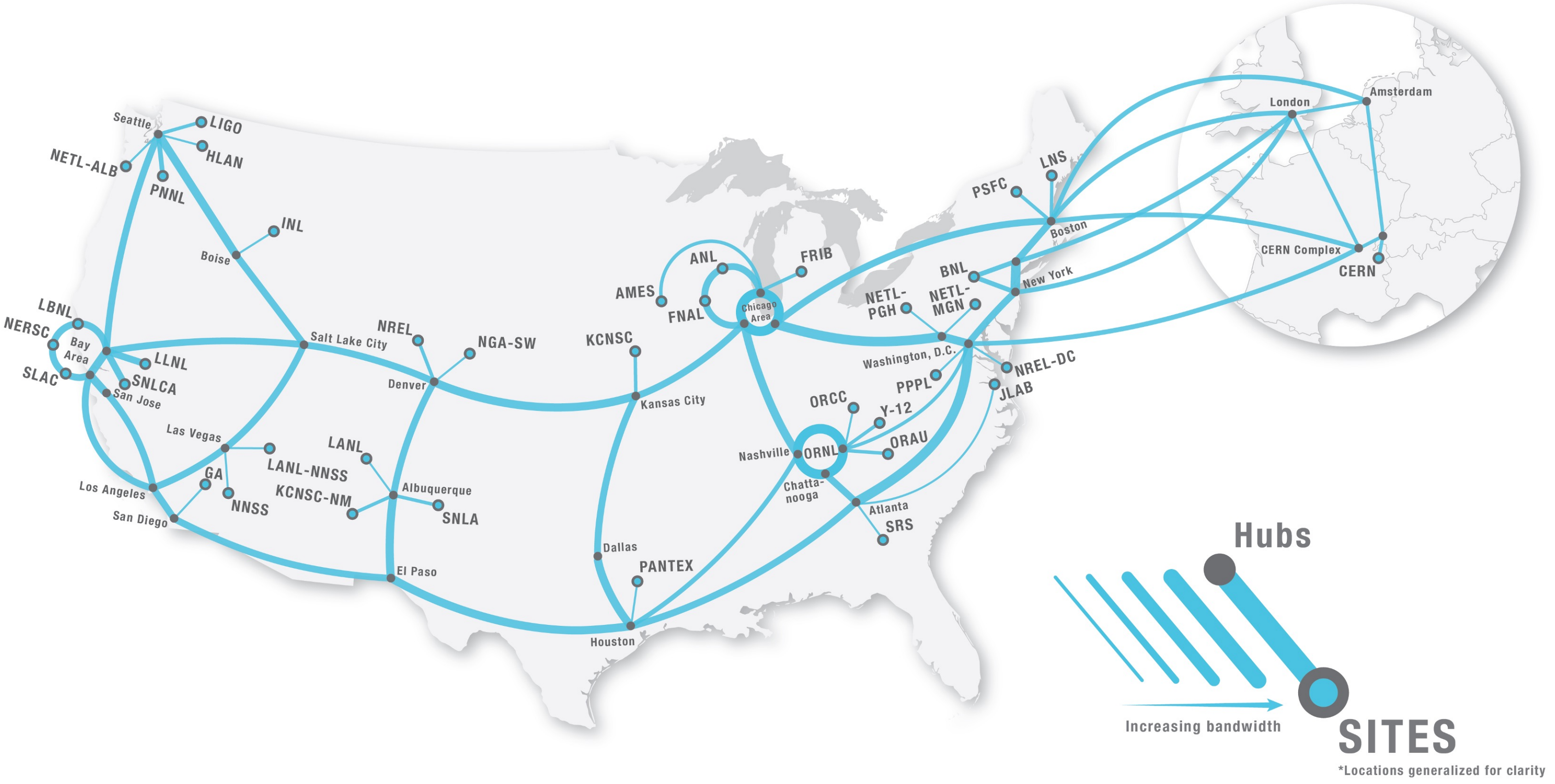
~48 Gbps



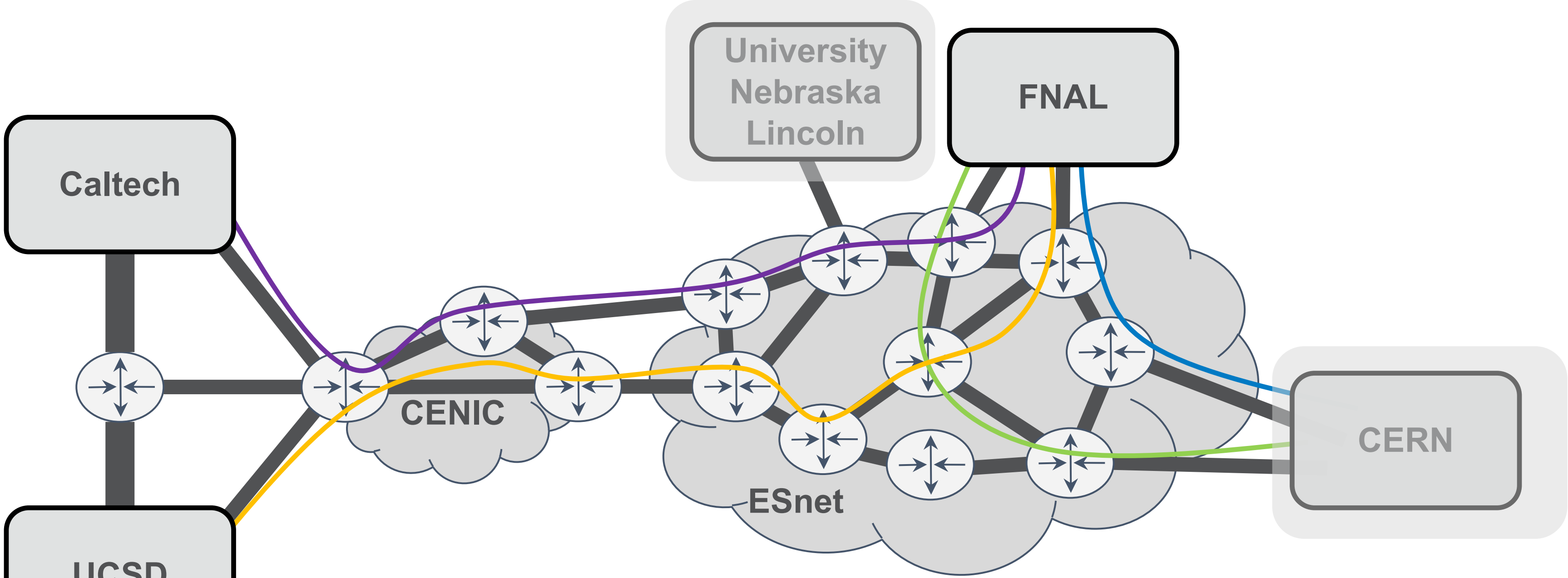


| Source | Destination | VO | Submitted | Active | Staging | S.Active | Archivin | Finished | Failed | Cancel | Rate (last 1h) | Thr. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| + davs://xrootd-sense-ucsd-redirector.sdsc | davs://sense-redir-01.ultralight.or | cms | 1284 | 190 | - | - |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

# ESnet Network Topo + US-CMS Sites



# SENSE Rucio/FTS/XRootD Interoperation System Deployment



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

▨ Deployment underway

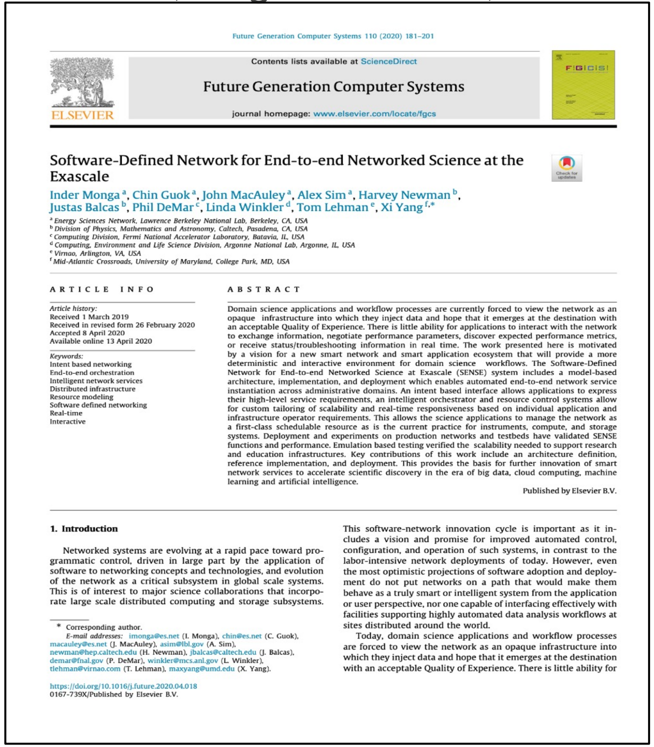
May add other sites: CERN, UNL, Vanderbilt, SPRACE

# Next Steps

- Development Goals:
  - DMM Development and policies. Allow it be adaptable – and define importance of data transfer.
  - Add more sites – US (Fermilab (T1), Nebraska (T2), Vanderbilt (T2)), Brazil - Sprace (T2), CERN (T2). Looking for more European site(s).
  - More NOS (Network Operating Systems) support in SiteRM (Dell OS 10, FreeRTR, Juniper)
  - Quality of Service (Hard QoS, Soft QoS) What to do once underutilized/oversubscribed?
  - Link weights on WAN:
    - Caltech-LasVegas-CERN (130ms, 10gbit max); Caltech-SFO-CERN (163ms, 20gbit max)
  - Policy for fair-share between experiments. Who gets how much and what?
  - Automated End-to-End troubleshooting, monitoring, alarming. (pin-point exact hop failing, alerting)
  - Other experiment use cases and support in SENSE.
- Participate in the WLCG Data Challenge 2024

# SENSE Information and Contacts

- Software-Defined Network for End-to-end Networked Science at the Exascale, Elsevier Future Generation Computer Systems, Volume 110, September 2020, Pages 181-201, https://doi.org/10.1016/j.future.2020.04.018
- SENSE Northbound API Program
  - https://app.swaggerhub.com/apis/xi-yang/SENSE-O-Intent-API
- Contacts
  - Xi Yang, xiyang@es.net
  - Tom Lehman, tlehman@es.net
  - SENSE Information, sense-info@es.net
- SENSE Website: sense.es.net



Future Generation Computer Systems 110 (2020) 181–201

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

## Software-Defined Network for End-to-end Networked Science at the Exascale

Inder Monga[a], Chin Guok[a], John MacAuley[a], Alex Sim[a], Harvey Newman[b], Justas Balcas[b], Phil DeMar[c], Linda Winkler[d], Tom Lehman[e], Xi Yang[f,*]

[a] Energy Sciences Network, Lawrence Berkeley National Lab, Berkeley, CA, USA
[b] Division of Physics, Mathematics and Astronomy, Caltech, Pasadena, CA, USA
[c] Computing Division, Fermi National Accelerator Laboratory, Batavia, IL, USA
[d] Computing, Environment and Life Science Division, Argonne National Lab, Argonne, IL, USA
[e] Virnao, Arlington, VA, USA
[f] Mid-Atlantic Crossroads, University of Maryland, College Park, MD, USA

ARTICLE INFO

*Article history:*
Received 1 March 2019
Received in revised form 26 February 2020
Accepted 8 April 2020
Available online 13 April 2020

*Keywords:*
Intent based networking
End-to-end orchestration
Intelligent network services
Distributed infrastructure
Resource modeling
Software defined networking
Real-time
Interactive

ABSTRACT

Domain science applications and workflow processes are currently forced to view the network as an opaque infrastructure into which they inject data and hope that it emerges at the destination with an acceptable Quality of Experience. There is little ability for applications to interact with the network to exchange information, negotiate performance parameters, discover expected performance metrics, or receive status/troubleshooting information in real time. The work presented here is motivated by a vision for a new smart network and smart application ecosystem that will provide a more deterministic and interactive environment for domain science workflows. The Software-Defined Network for End-to-end Networked Science at Exascale (SENSE) system includes a model-based architecture, implementation, and deployment which enables automated end-to-end network service instantiation across administrative domains. An intent based interface allows applications to express their high-level service requirements, an intelligent orchestrator and resource control systems allow for custom tailoring of scalability and real-time responsiveness based on individual application and infrastructure operator requirements. This allows the science applications to manage the network as a first-class schedulable resource as is the current practice for instruments, compute, and storage systems. Deployment and experiments on production networks and testbeds have validated SENSE functions and performance. Emulation based testing verified the scalability needed to support research and education infrastructures. Key contributions of this work include an architecture definition, reference implementation, and deployment. This provides the basis for further innovation of smart network services to accelerate scientific discovery in the era of big data, cloud computing, machine learning and artificial intelligence.

Published by Elsevier B.V.

### 1. Introduction

Networked systems are evolving at a rapid pace toward programmatic control, driven in large part by the application of software to networking concepts and technologies, and evolution of the network as a critical subsystem in global scale systems. This is of interest to major science collaborations that incorporate large scale distributed computing and storage subsystems. This software-network innovation cycle is important as it includes a vision and promise for improved automated control, configuration, and operation of such systems, in contrast to the labor-intensive network deployments of today. However, even the most optimistic projections of software adoption and deployment do not put networks on a path that would make them behave as a truly smart or intelligent system from the application or user perspective, nor one capable of interfacing effectively with facilities supporting highly automated data analysis workflows at sites distributed around the world.

Today, domain science applications and workflow processes are forced to view the network as an opaque infrastructure into which they inject data and hope that it emerges at the destination with an acceptable Quality of Experience. There is little ability for

* Corresponding author.
*E-mail addresses:* imonga@es.net (I. Monga), chin@es.net (C. Guok), macauley@es.net (J. MacAuley), asim@lbl.gov (A. Sim), newman@hep.caltech.edu (H. Newman), jbalcas@caltech.edu (J. Balcas), demar@fnal.gov (P. DeMar), winkler@mcs.anl.gov (L. Winkler), tlehman@virnao.com (T. Lehman), maxyang@umd.edu (X. Yang).

https://doi.org/10.1016/j.future.2020.04.018
0167-739X/Published by Elsevier B.V.

# Acknowledgements

Chin Guok, Tom Lehman, Inder Monga, Xi Yang

Harvey Newman, Justas Balcas, Preeti Bhat

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

Oliver Gutsche, Phil Demar

tnc23

DIGITAL GENERATIONS

TIRANA, ALBANIA | 5-9 JUNE 2023

# Thank you

Any questions?

Tom Lehman <tlehman@es.net>

GÉANT

Co-funded by
the European Union

# Extra Slides

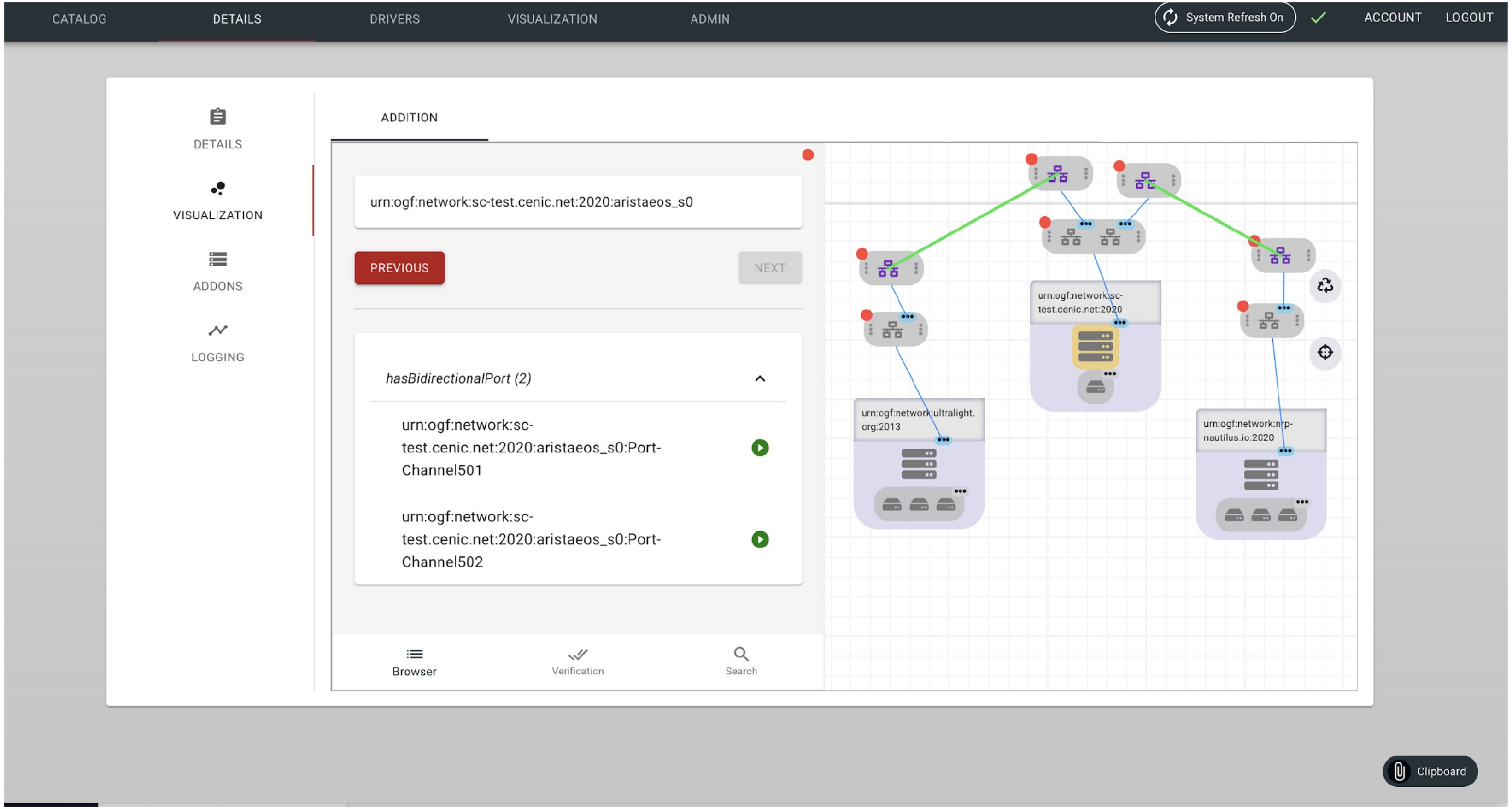
# Key Themes

- 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

# 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

# SENSE - Site Layer 3 Flow to WAN Traffic Engineered Path Service



urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0

*hasBidirectionalPort (2)*

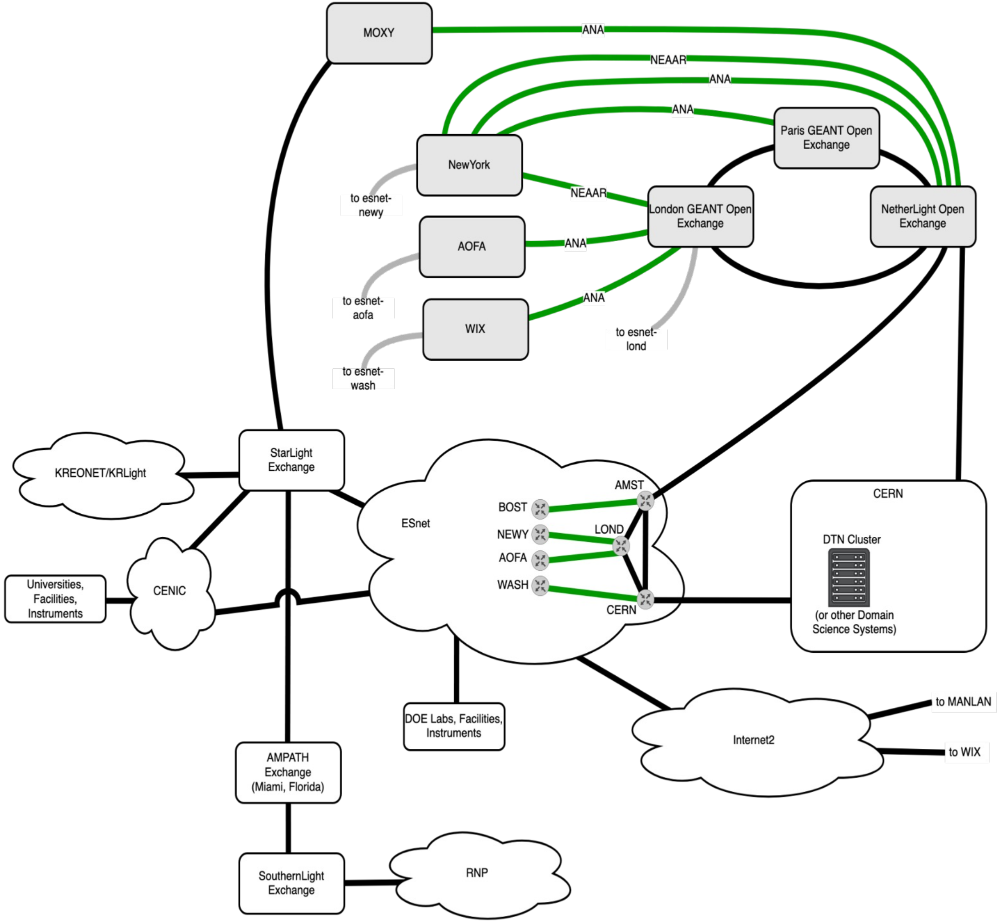
- urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0:Port-Channel501
- urn:ogf:network:sc-test.cenic.net:2020:aristaeos_s0:Port-Channel502

# Objectives

- Make Rucio capable to **schedule transfers on the network.**
- **Improve accountability**.
- **Predetermined transfer speed and quality of service (time to completion).**
- **Fine-grain managed transfers can be also fine-grain monitored** since they travel alone within a well-identified network channel.
- Comparing **Achieved V.S. Allocated bandwidth** will make network & endpoint issues evident.

# Important Link Management

- There are **multiple transatlantic and transpacific links**, operated by multiple organizations
- Goal is to more flexibly control how these are utilized on a per flow, group, or use basis
- Do not want to manage "**every**" flow in the network; but we should be able to manage "**any**" flow in the network
- **An equally important goal** is to understand the load vs capacity and leave room for other traffic
- **Remain compatible** with other network operations
- **Two timescales:** SENSE overlay network of virtual circuits with BW guarantees is relatively stable; IPv6 subnets and Directors provide more dynamic flow mapping to various traffic engineered paths


MOXY
ANA
NEAAR
ANA
ANA
Paris GEANT Open Exchange
NewYork
to esnet-newy
NEAAR
London GEANT Open Exchange
NetherLight Open Exchange
AOFA
ANA
to esnet-aofa
WIX
ANA
to esnet-lond
to esnet-wash
KREONET/KRLight
StarLight Exchange
ESnet
AMST
BOST
NEWY
LOND
AOFA
WASH
CERN
CERN
DTN Cluster
(or other Domain Science Systems)
Universities, Facilities, Instruments
CENIC
DOE Labs, Facilities, Instruments
Internet2
to MANLAN
to WIX
AMPATH Exchange (Miami, Florida)
SouthernLight Exchange
RNP