Automated Network Services for Exascale Data Movement

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, Phil Demar

May 8, 2023

Using SENSE

Automated Network Services for Exascale Data Movement

To move data in Rucio

For CMS sites

But let's keep in mind that this work is extensible to any collaboration that uses Rucio (adaptable to other data management systems too) to move data across sites

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, Phil Demar

May 8, 2023

ESnet

CALIFORNIA INSTITUTE OF TECHNOLOGY 1891

UCSD

CMS

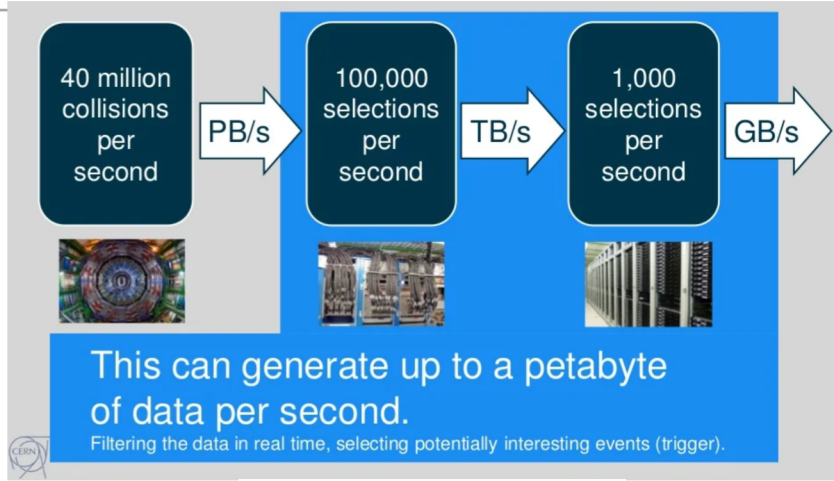
Motivation = HL-LHC

Caltech

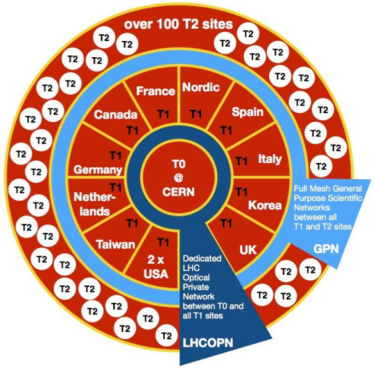
CMS expects more than half **an exabyte of new data for each year of LHC** operations during the High-Luminosity LHC era from about 2029-2040

- One annual processing workflow of a few hundred PBs
- Every 3 years, exabyte scale re-processing workflow

Total aggregate data flows are expected to be **dominated by the largest flows**.



- Given that:
  - WAN Network resources are not unlimited!
  - We don't have infinite $$; We cannot just throw hardware to the problem



We need to be smart and **use our resources efficiently**

Objectives

Caltech

Make Rucio capable to **schedule transfers on the network**.
Improve accountability.

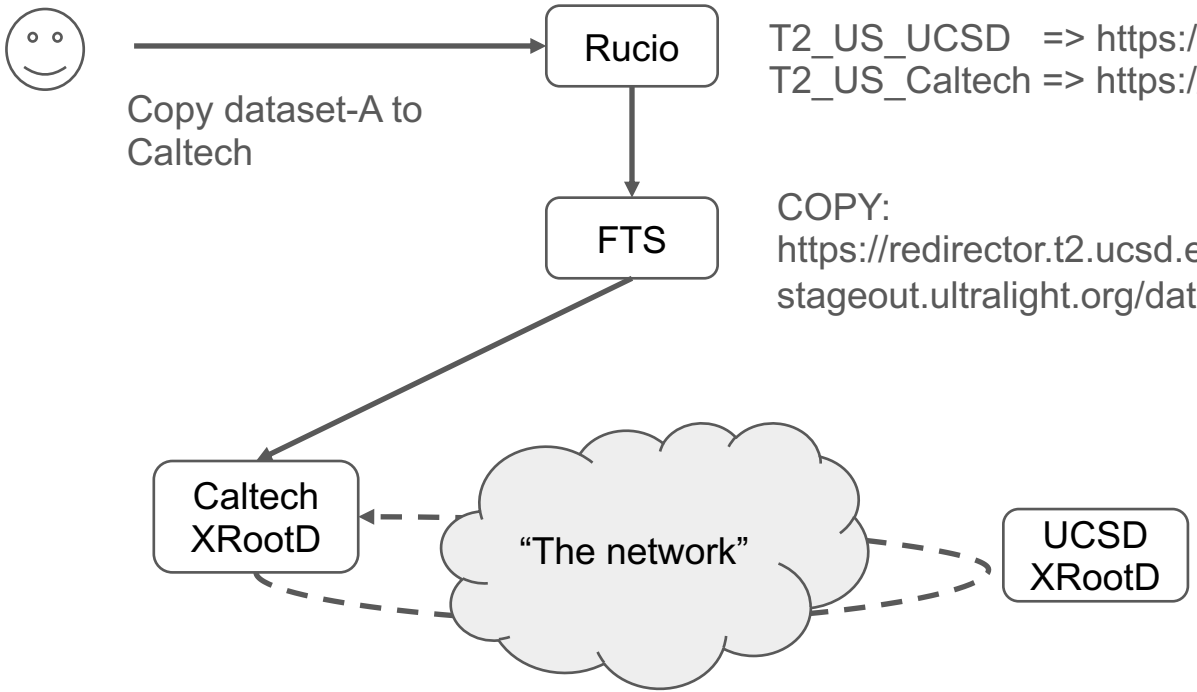
Predetermined transfer speed and quality of service (time to completion).

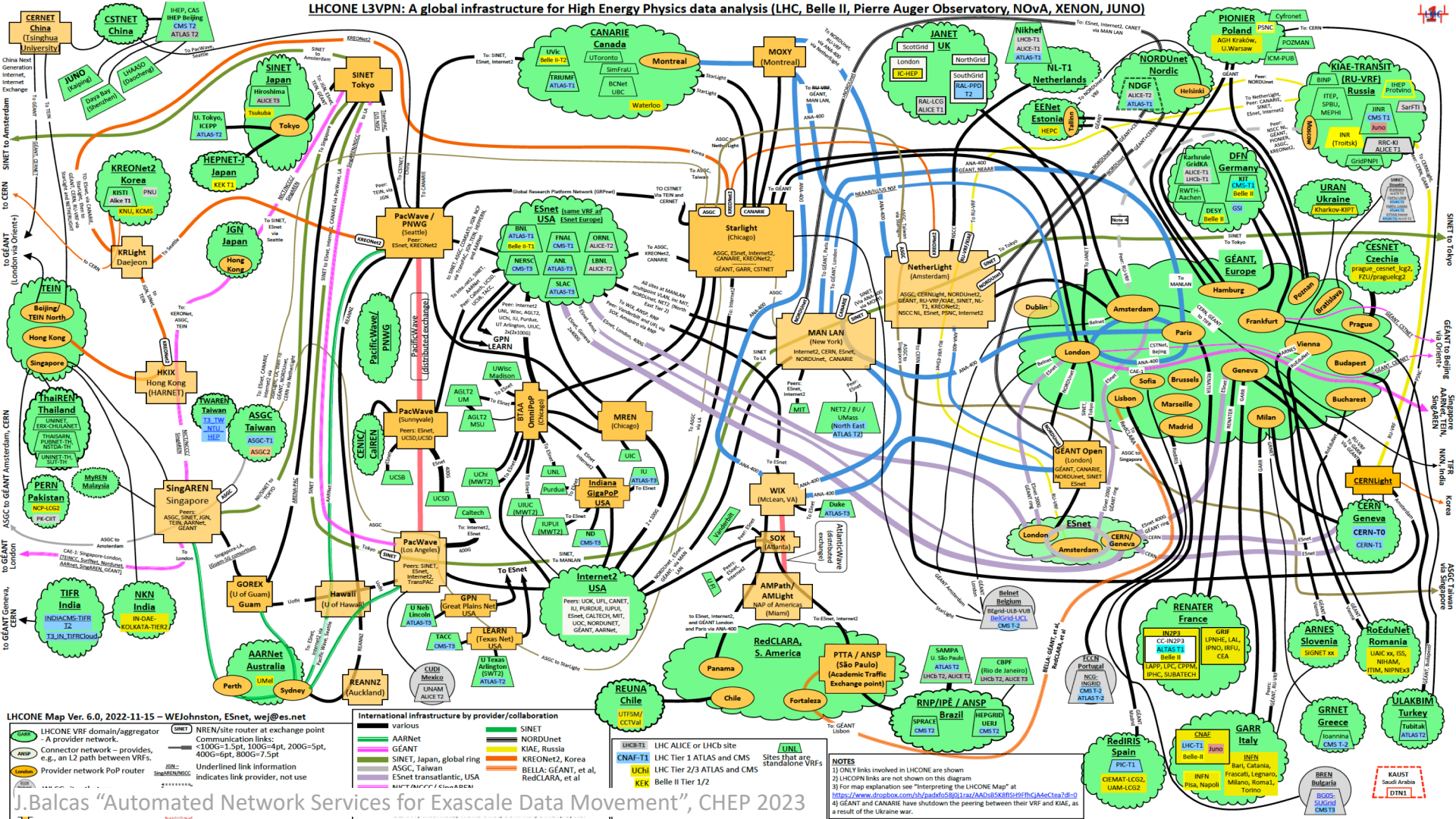
Fine-grain managed transfers can be also fine-grain monitored since they travel alone within a well-identified network channel.

Comparing **Achieved V.S. Allocated bandwidth** will make network & endpoint issues evident.

Quick review: How transfers work nowadays?

Caltech



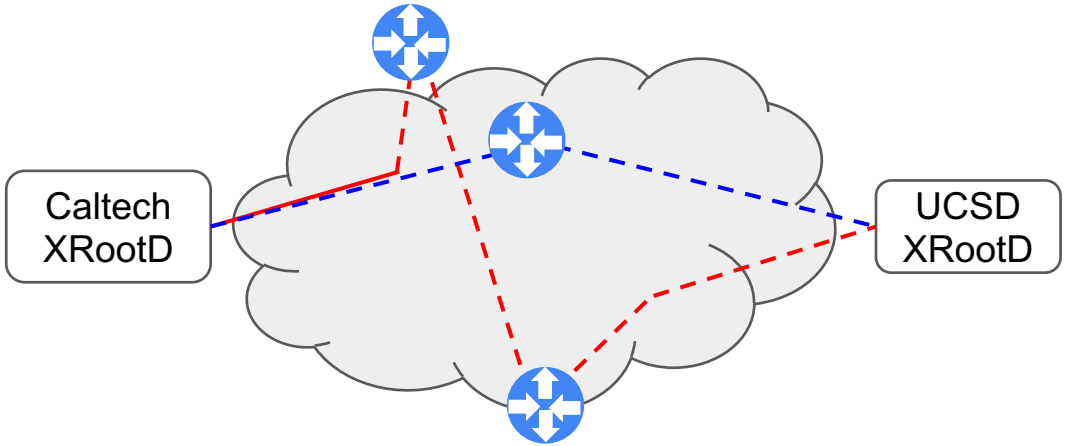

LHCONE L3VPN: A global infrastructure for High Energy Physics data analysis (LHC, Belle II, Pierre Auger Observatory, NOvA, XENON, JUNO)
CERNET China (Tsinghua University)
CSTNET China
IHEP, CAS IHEP Beijing CMS T2 ATLAS T2
JUNO (Kaijing)
Daya Bay (Shenzhen)
LHAASO (Daocheng)
China Next Generation Internet, Internet Exchange
SINET Japan Hiroshima ALICE T3 Tsukuba
Tokyo
SINET Tokyo
U. Tokyo, ICEPP ATLAS-T2
HEPNET-J Japan KEK T1
KREONet2 Korea KISTI Alice T1 PNU KNU, KCMS
KRLight Daejeon
JGN Japan Hong Kong
TEIN Beijing/ TEIN North Hong Kong Singapore
HKIX Hong Kong (HARNET)
ThaiREN Thailand UNINET, ERX-CHULANET THAISARN, PUBNET-TH, NSTDA-TH UNINET-TH, SUT-TH
PERN Pakistan NCP-LCG2 PK-CIIT
MyREN Malaysia
SingAREN Singapore Peers: ASGC, SINET, JGN, TEIN, AARNet, GÉANT
TWAREN Taiwan T3_TW_NTU_HEP
ASGC Taiwan ASGC-T1 ASGC2
TIFR India INDIACMS-TIFR T2 T3_IN_TIFRCloud
NKN India IN-DAE-KOLKATA-TIER2
AARNet Australia UMel Perth Sydney
GOREX (U of Guam) Guam
Hawaii (U of Hawaii)
REANNZ (Auckland)
CUDI Mexico UNAM ALICE T2
REUNA Chile UTFSM/ CCTVal
RedCLARA, S. America Panama Chile Fortaleza
PTTA / ANSP (São Paulo) (Academic Traffic Exchange point)
RNP/IPÊ / ANSP Brazil SPRACE CMS T2 HEPGRID UERJ CMS T2
SAMPA U. São Paulo ATLAS T2
CBPF (Rio de Janeiro) LHCb T2, ALICE T3
RedIRIS Spain PIC-T1 CIEMAT-LCG2, UAM-LCG2
CANARIE Canada UVic Belle II-T2 UToronto SimFraU TRIUMF ATLAS-T1 Montreal BCNet UBC Waterloo
MOXY (Montreal)
Starlight (Chicago) ASGC, ESnet, Internet2, CANARIE, KREONet2; GÉANT, GARR, CSTNET
ESnet (same VRF as ESnet Europe) BNL ATLAS-T3 Belle II-T1 FNAL CMS-T1 ORNL ALICE-T2 NERSC CMS-T3 ANL ATLAS-T3 LBNL ALICE-T2 SLAC ATLAS-T3
PacWave / PNWG (Seattle) Peer: ESnet, KREONet2
PacificWave (distributed exchange)
PacificWave/ PNWG
GPN LEARN
PacWave (Sunnyvale) Peers: ESnet, UCSD,UCSC
CENIC/ CalREN
UCSB
UCSD
Caltech
UWisc Madison
AGLT2 UM
AGLT2 MSU
BTAA OmniPoP (Chicago)
UChi (MWT2)
UNL
Purdue
UIUC (MWT2)
IUPUI (MWT2)
ND CMS-T3
MREN (Chicago)
UIC
IU ATLAS-T3
Indiana GigaPoP USA
PacWave (Los Angeles) Peers: SINET, ESnet, Internet2, TransPAC
GPN Great Plains Net USA
U Neb Lincoln ATLAS-T3
TACC CMS-T3
LEARN (Texas Net) USA
U Texas Arlington (SWT2) ATLAS-T2
Internet2 USA Peers: UOK, UFL, CANET, IU, PURDUE, IUPUI, ESnet, CALTECH, MIT, UIOC, NORDUNET, GÉANT, AARNet,
MAN LAN (New York) Internet2, CERN, ESnet, NORDUnet, CANARIE
MIT
NET2 / BU / UMass (North East ATLAS T2)
WIX (McLean, VA)
Duke ATLAS-T3
SOX (Atlanta)
AMPath/ AMLight NAP of Americas (Miami)
JANET UK ScotGrid London IC-HEP NorthGrid SouthGrid RAL-PPD T2 RAL-LCG ALICE T1
NL-T1 Netherlands Nikhef LHCB-T1 ALICE-T1 ATLAS-T1
EENet Estonia HEPC Tallinn
NDGF NORDUnet Nordic Helsinki
PIONIER Poland AGH Kraków, U.Warsaw PSNC Cyfronet POZMAN ICM-PUB
KIAE-TRANSIT (RU-VRF) Russia BINP IHEP Protvino ITEP, SPBU, MEPHI JINR CMS T1 Juno INR (Troitsk) SarFTI RRC-KI ALICE T1 GridPNPI Moscow
URAN Ukraine Kharkov-KIPT
CESNET Czechia prague_cesnet_lcg2, FZU/praguelcg2
DFN Germany Karlsruhe GridKA ALICE-T1 LHCb-T1 RWTH-Aachen DESY Belle II GSI KIT CMS-T1 Belle II
GÉANT, Europe Hamburg Poznan Bratislava Dublin Amsterdam Frankfurt Paris Prague Vienna Budapest London Geneva Sofia Brussels Lisbon Marseille Madrid Milan Bucharest
NetherLight (Amsterdam) ASGC, CERNLight, NORDUnet2, GÉANT, RU-VRF/KIAE, SINET, NL-T1, KREONet2; NSCC NL, ESnet, PSNC, Internet2
GÉANT Open (London) GÉANT, CANARIE, NORDUnet, SINET ESnet
ESnet London Amsterdam CERN/ Geneva
CERNLight CERN Geneva CERN-T0 CERN-T1
RENATER France IN2P3 CC-IN2P3 ATLAS T1 Belle II LAPP, LPC, CPPM, IPHC, SUBATECH GRIF LPNHE, LAL, IPNO, IRFU, CEA
Belnet Belgium BGrid-ULB-VUB BeLGrid-UCL CMS T-2
ECCN Portugal NCG-INGRID CMS T-2 ATLAS T-2
GARR Italy CNAF LHC-T1 Juno Belle-II INFN Bari, Catania, Frascati, Legnaro, Milano, Roma1, Torino INFN Pisa, Napoli
ARNES Slovenia SIGNET xx
RoEduNet Romania UAIC xx, ISS, NIHAM, ITIM, NIPNEx3
GRNET Greece Ioannina CMS T-2
ULAKBIM Turkey Tubitak ATLAS T2
BREN Bulgaria BG05-SUGrid CMS T3
KAUST Saudi Arabia DTN1
LHCONE Map Ver. 6.0, 2022-11-15 – WEJohnston, ESnet, wej@es.net
LHCONE VRF domain/aggregator - A provider network.
Connector network – provides, e.g., an L2 path between VRFs.
Provider network PoP router
Underlined link information indicates link provider, not use
NREN/site router at exchange point
Communication links: <100G=1.5pt, 100G=4pt, 200G=5pt, 400G=6pt, 800G=7.5pt
International infrastructure by provider/collaboration
various
AARNet
GÉANT
SINET, Japan, global ring
ASGC, Taiwan
ESnet transatlantic, USA
SINET
NORDUnet
KIAE, Russia
KREONet2, Korea
BELLA: GÉANT, et al, RedCLARA, et al
LHCB-T1 LHC ALICE or LHCb site
CNAF-T1 LHC Tier 1 ATLAS and CMS
UChi LHC Tier 2/3 ATLAS and CMS
KEK Belle II Tier 1/2
UNL Sites that are standalone VRFs
NOTES
1) ONLY links involved in LHCONE are shown
2) LHCOPN links are not shown on this diagram
3) For map explanation see "Interpreting the LHCONE Map" at https://www.dropbox.com/sh/pad4fo58l01rzaz/AADsiBSK8fiH9FfhGjA4eCtea?dl=0
4) GÉANT and CANARIE have shutdown the peering between their VRF and KIAE, as a result of the Ukraine war.
to GÉANT Amsterdam, CERN
to GÉANT London
to GÉANT Geneva, CERN
to GÉANT (London via Orient+)
to CERN
SINET to Amsterdam
SINET to Tokyo
GÉANT to Beijing via Orient+
Singapore AARNet, SingAREN
TIFR NKN, India
Korea
ASGC Taiwan via Singapore
J.Balcas "Automated Network Services for Exascale Data Movement", CHEP 2023

What can be improved?

Caltech

Let's STOP using the network as a black box and instrument it!

There can be multiple paths between 2 SEs, but tools don't get to pick which one to use. The routing algorithms know nothing about our priorities



At the SE we put all transfers in the same bucket, no prioritization

*TPC – Third Party Copy, *AAA – Any Data, Any Time, Anywhere

Esnet Network Topo + US-CMS Sites

Caltech

J.Balcas “Automated Network Services for Exascale Data Movement”, CHEP 2023

SENSE Architecture

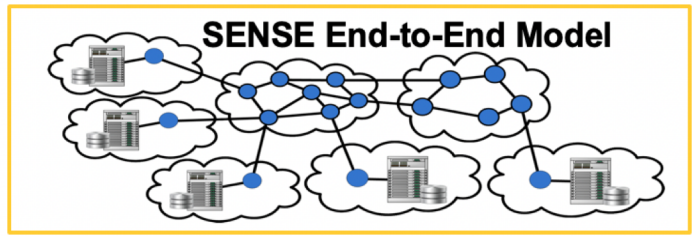
Intent Based APIs with Resource Discovery, Negotiation, Service Lifecycle Monitoring/Troubleshooting

Application Workflow Agent

SENSE Orchestrator

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

Model Driven SDN Control with Orchestration



RM UC San Diego TransferService

RM ESnet ENERGY SCIENCES NETWORK

RM Fermilab TransferService

Datafication of cyberinfrastructure to enable intelligent services

RM Caltech TransferService

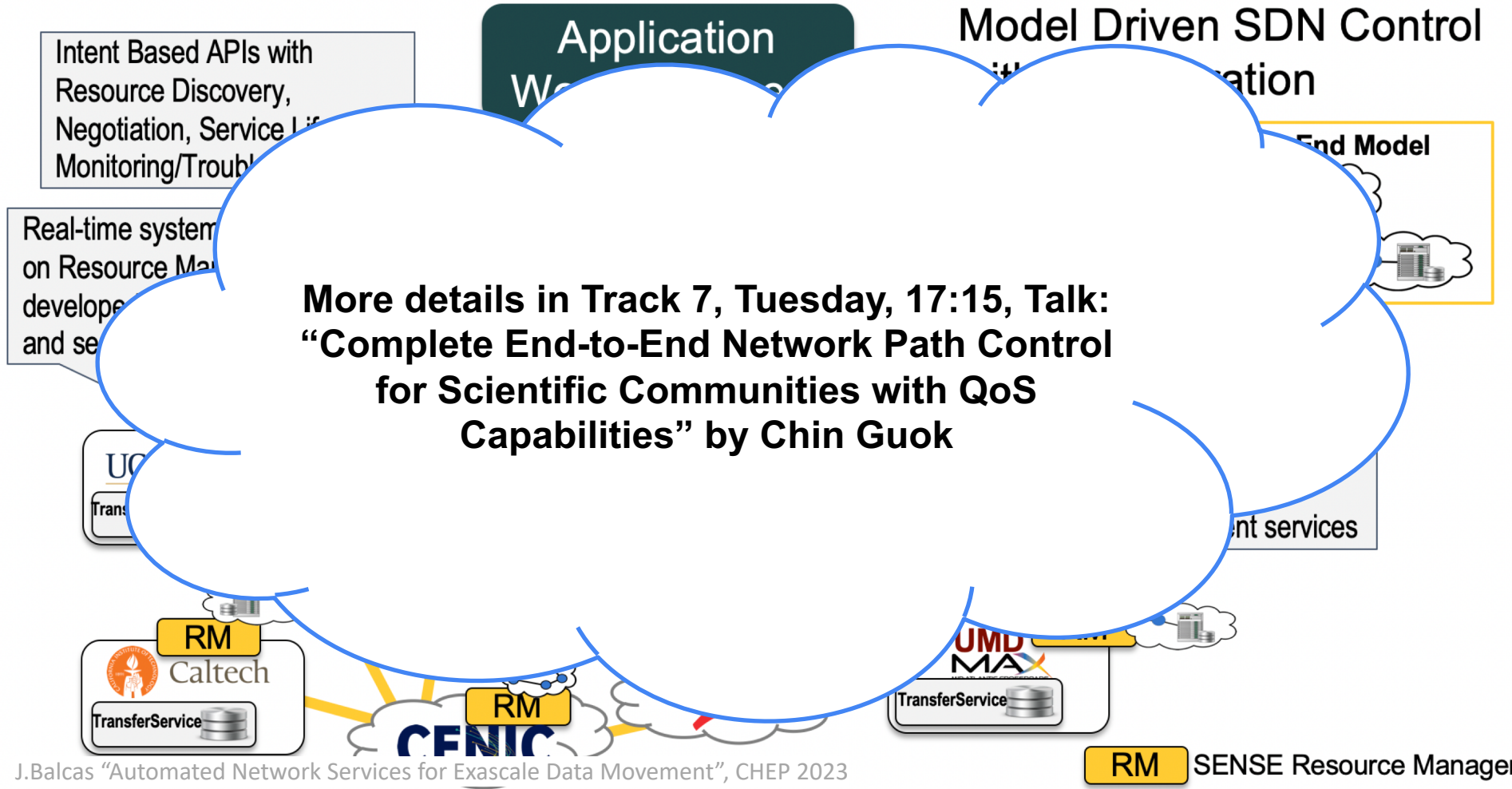
RM CENIC

RM INTERNET2

RM UMD MAX TransferService

RM SENSE Resource Manager

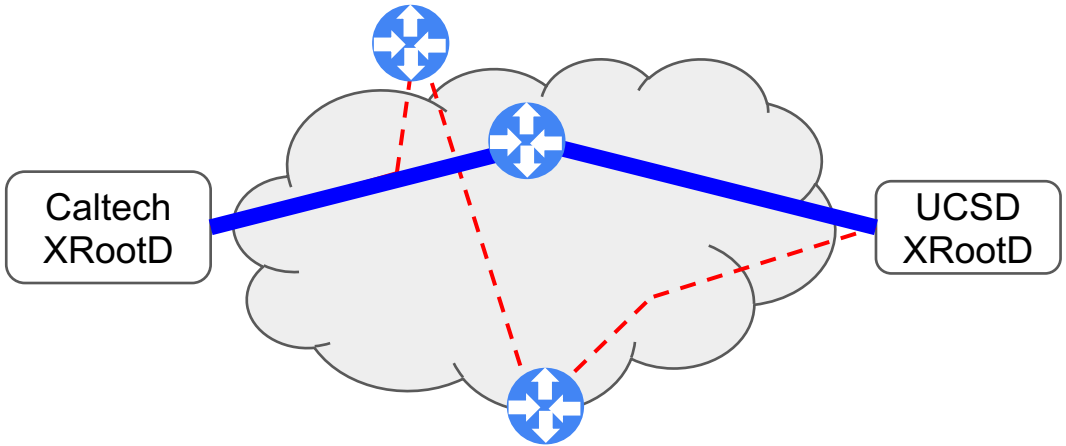
SENSE Architecture



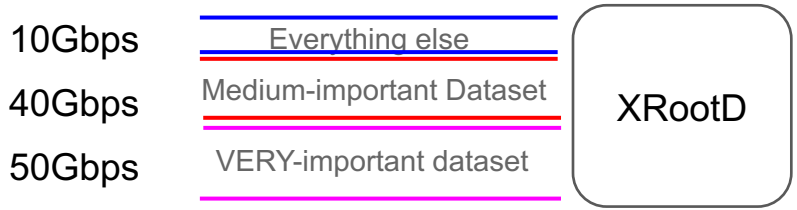
How do we improve things? Using SENSE we can:

Caltech

Configure **VPNs** between Routers so we can enforce a given path to be used for specific set of transfers

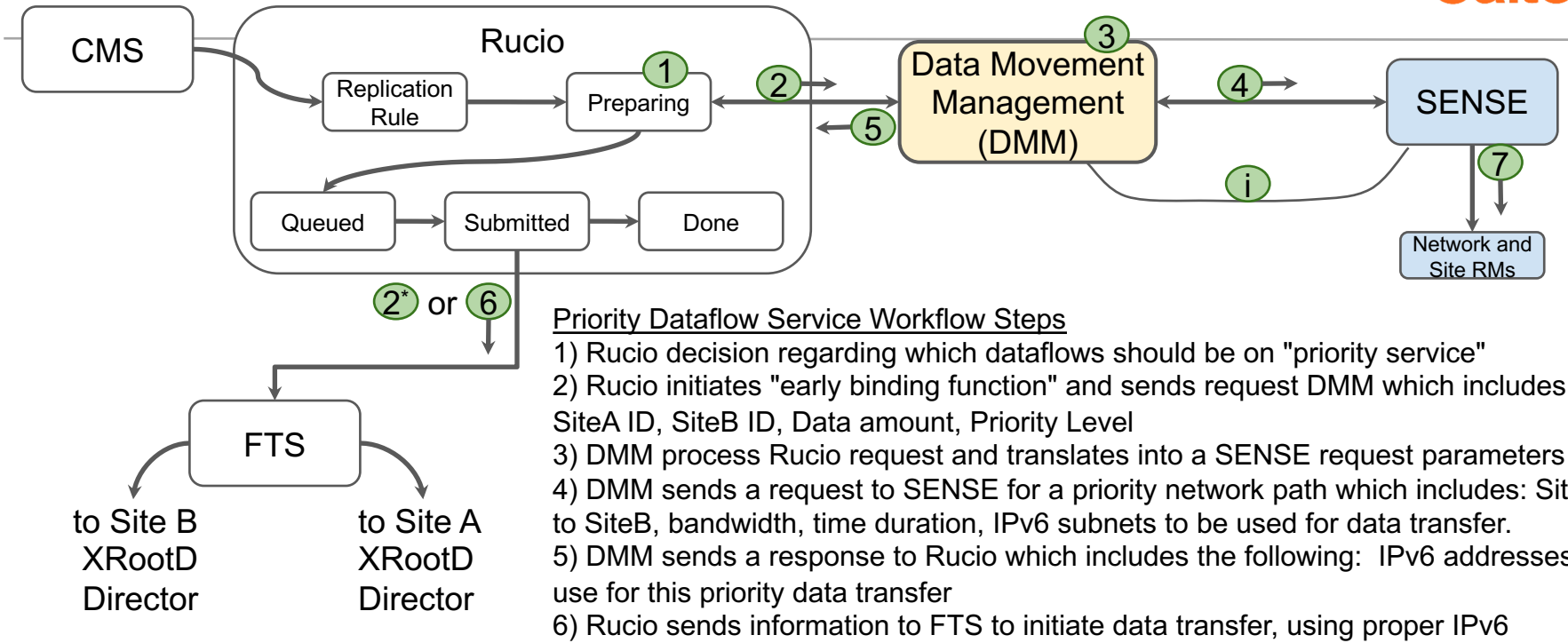


Implement **QoS** so that we can prioritize certain transfers at the DTN level/Router Level



Rucio/DMM/SENSE Workflow

Caltech



*Rucio to FTS and DMM interactions can be asynchronous

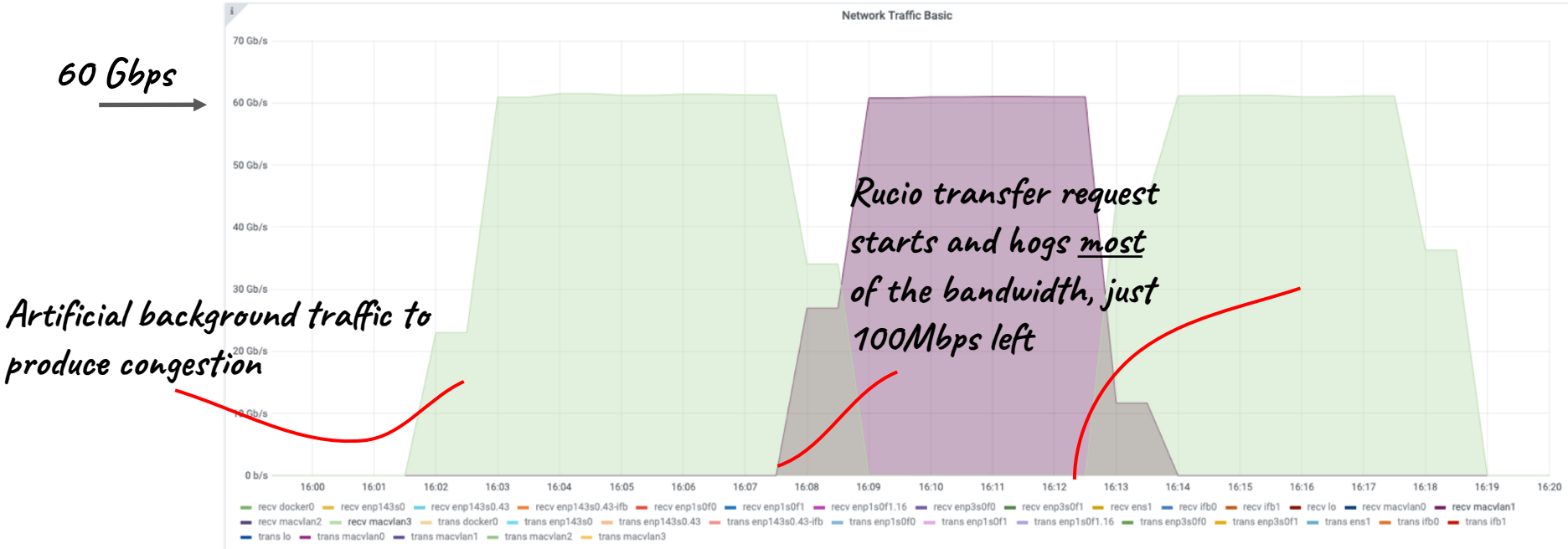
Priority Dataflow Service Workflow Steps

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

Proof of Concept Testing

Caltech

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

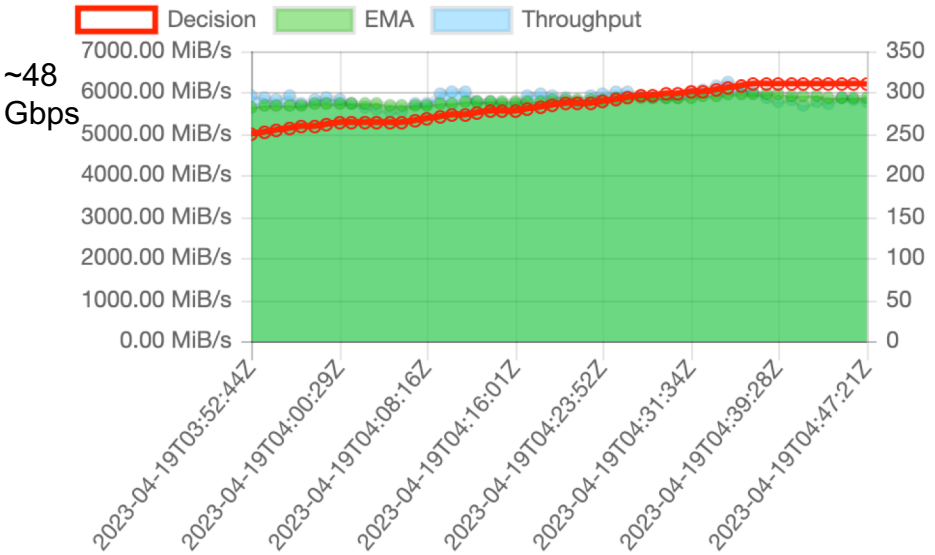


Higher Speed transfers using FTS/XRootD

Caltech

- 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
- Need a dynamic way to control submission rate (to FTS and FTS to XRootD)

~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 %	–

Key takeaways

Caltech

- 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

Thank You!

Want to try out? Join SENSE Testbed? Ask questions?
Drop an email to SENSE Group:
sense-info@es.net

ESnet

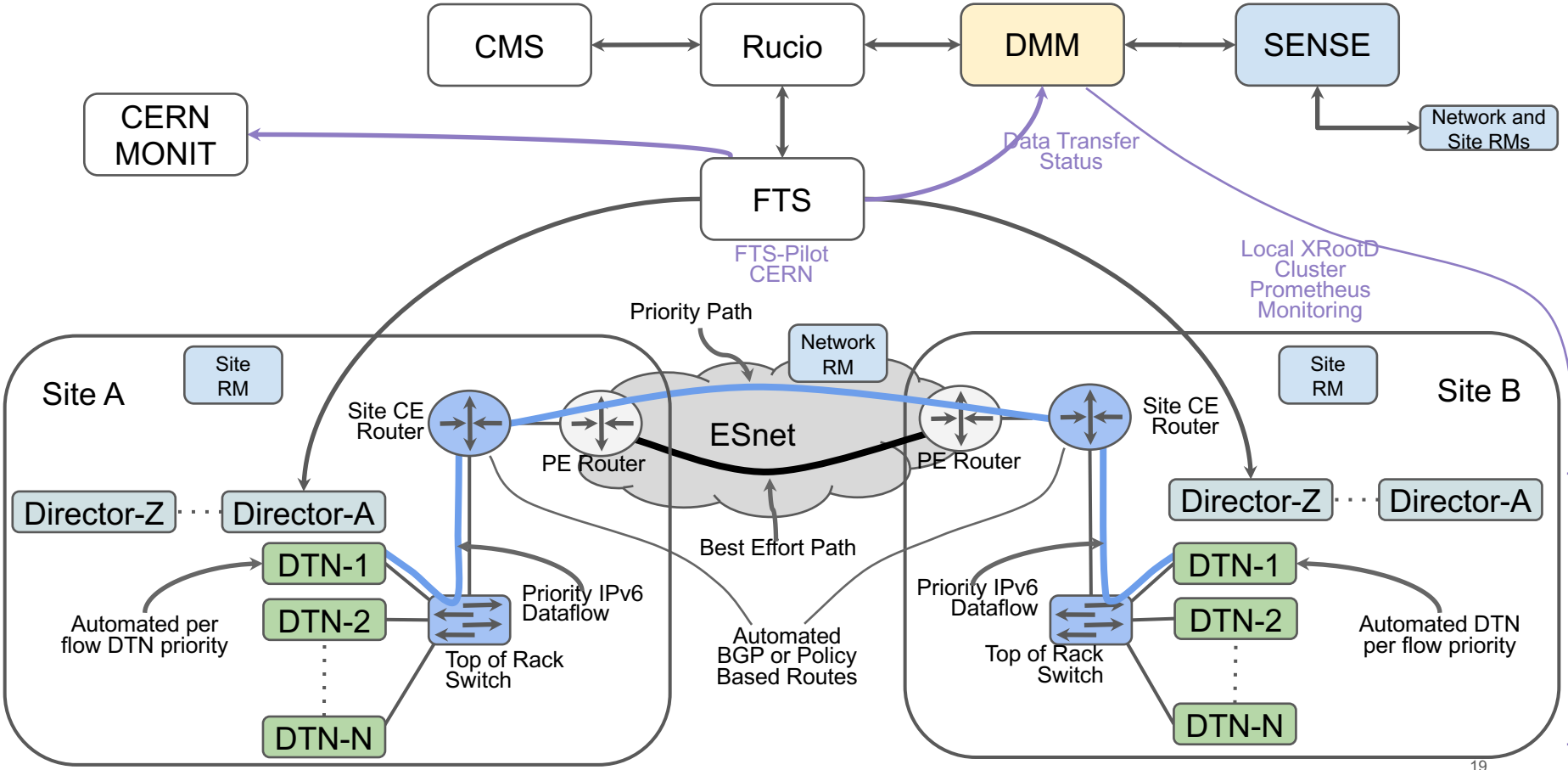
Acknowledgments

This ongoing work is partially supported by the US National Science Foundation (NSF) Grants OAC-2030508, OAC-1841530, OAC-1836650, MPS-1148698, and PHY-1624356. In addition, the development of SENSE is supported by the US Department of Energy (DOE) Grants DE-SC0015527, DE-SC0015528, DE-SC0016585, and FP-00002494. Finally, this work would not be possible without the significant contributions of collaborators at ESNet, Caltech, and SDSC.

Backup slides

SENSE/Rucio/FTS Workflow

Caltech



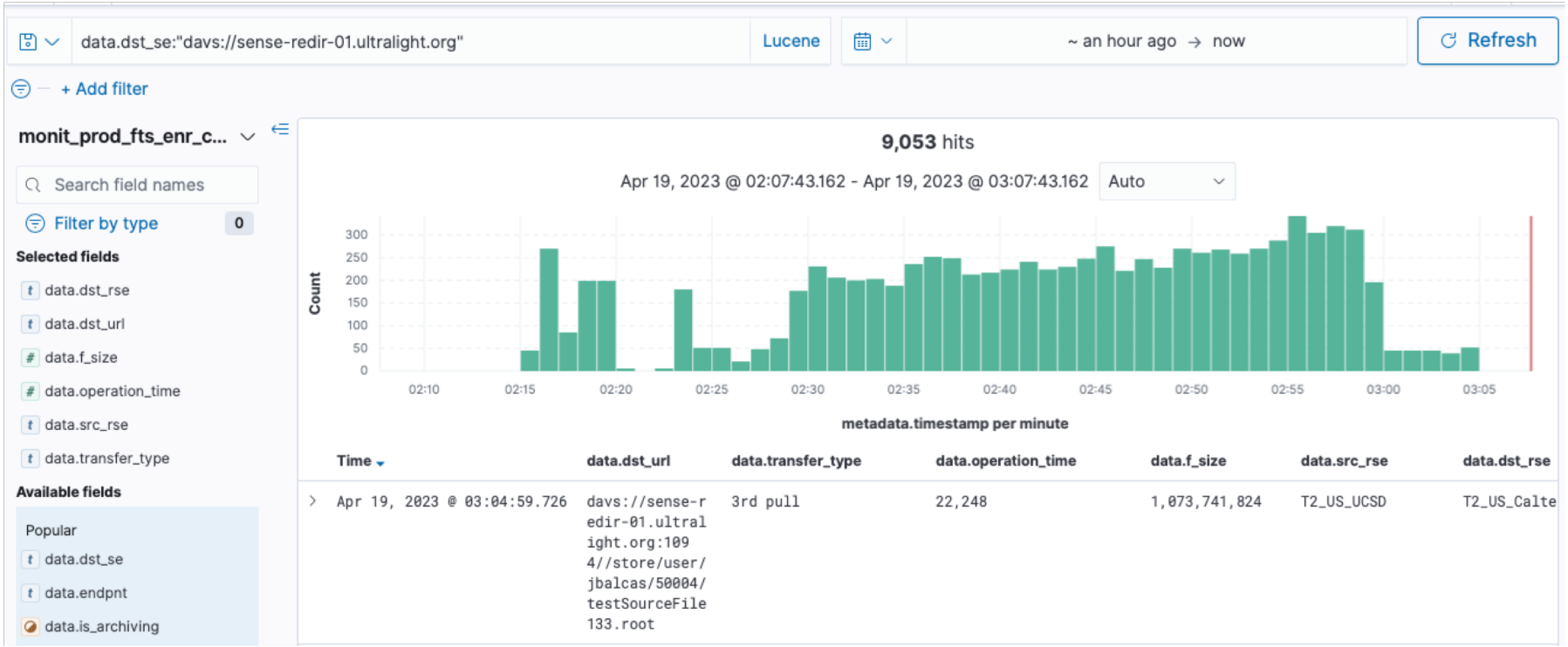
What's next?

Caltech

- 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

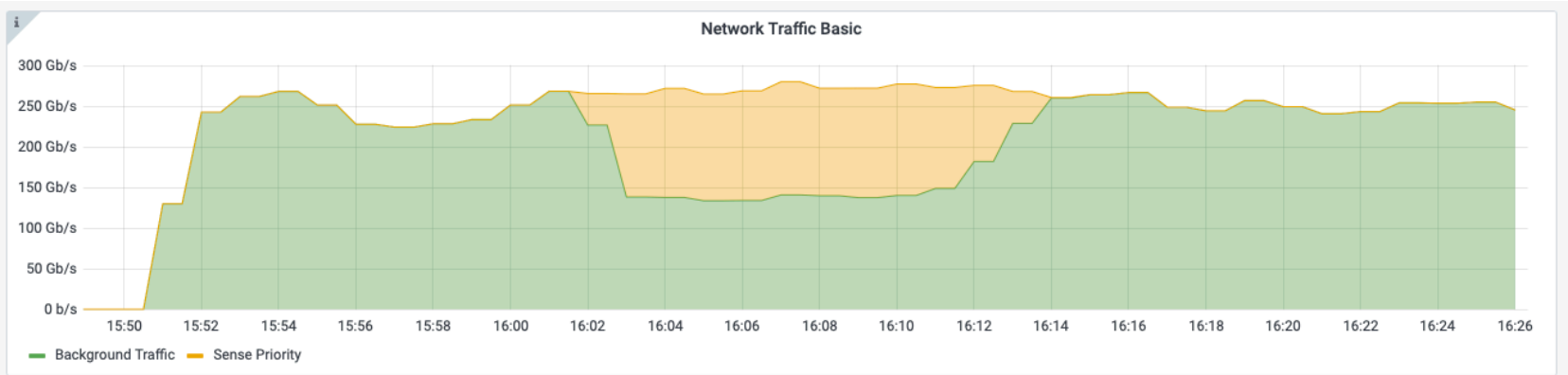
FTS Transfers via SENSE Path logged in MONIT (using CERN FTS3@Pilot Instance)

Caltech



UCSD to Caltech testing at higher speeds

Caltech



- 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 through some issues with Linux TC, Kubernetes/Multus Private NS Issues. Also evaluating use of BPF and Smart NICs for end-system options, P4TC.

TC – Traffic Control, NS – Namespaces, BPF – Berkeley Packet Filter,
P4TC – Protocol-Independent Packet Processor Traffic Control