

# **Making SENSE of Application-Network Interfaces**

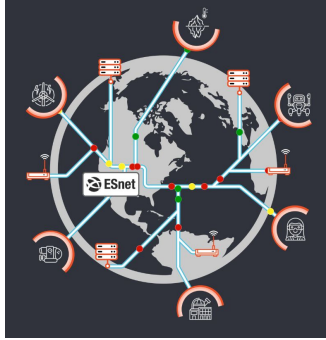
Keynote, Rucio Workshop

Inder Monga

Director, Energy Sciences Network  
Division Director, Scientific Networking Division  
Lawrence Berkeley National Lab



# Talk Flow



## ESnet Introduction: Data Circulatory System



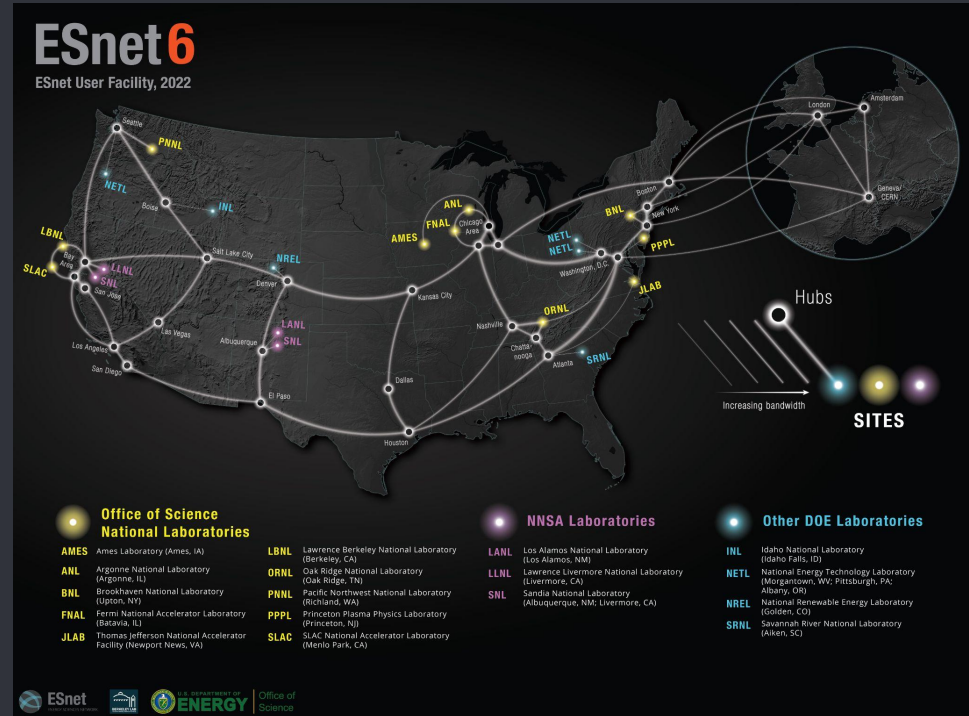
Analogy: Case for a richer Network-Application interface

$f(x)$

Journey towards application-network integration

# ESnet is the DOE'S data circulatory system...

- ESnet supports the DOE scientific research ecosystem.
- Interconnects all national labs and user facilities
- Provides reliable, high-performance connectivity to global research collaborations, the Cloud, and the larger Internet.



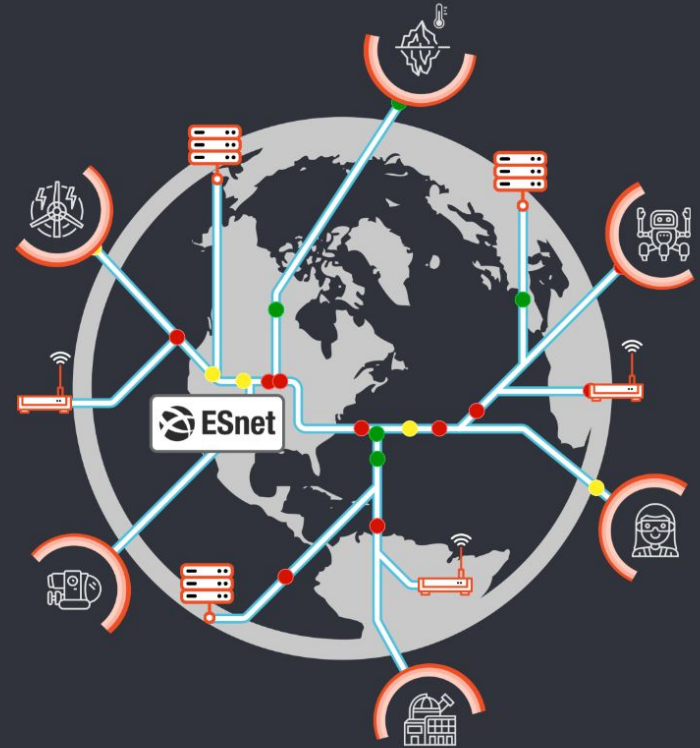
# ...and the stage for a global science laboratory.

## ESnet's Vision

Scientific progress will be completely unconstrained by the physical location of instruments, people, computational resources, or data.

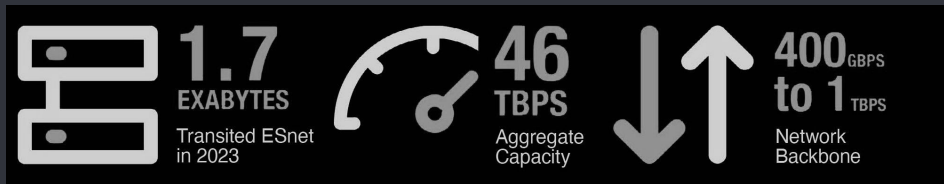
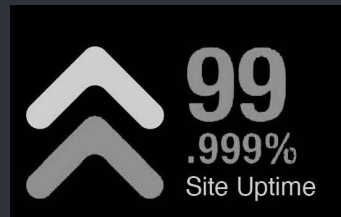
## ESnet's Mission

Networking that accelerates science.



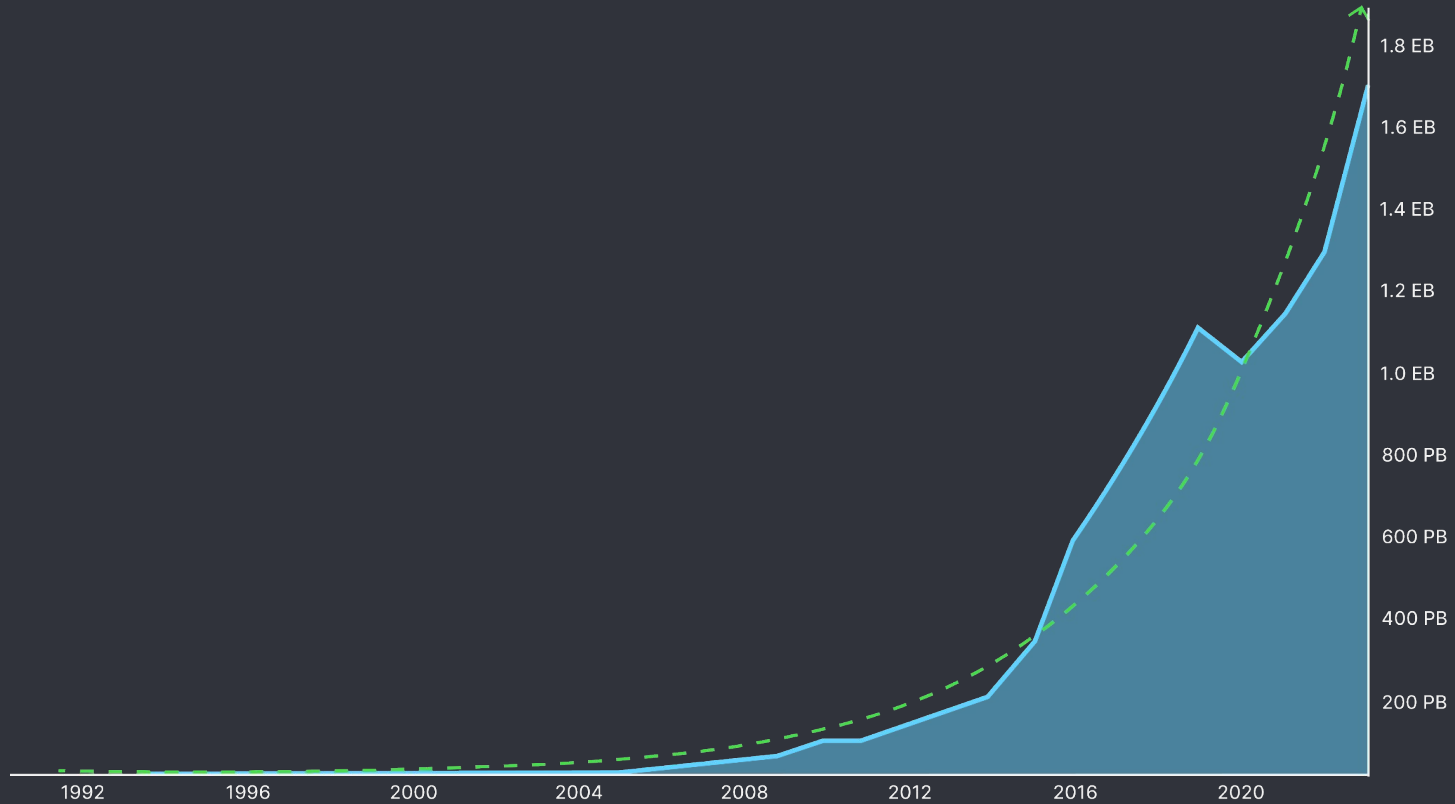


# By the numbers (2023)





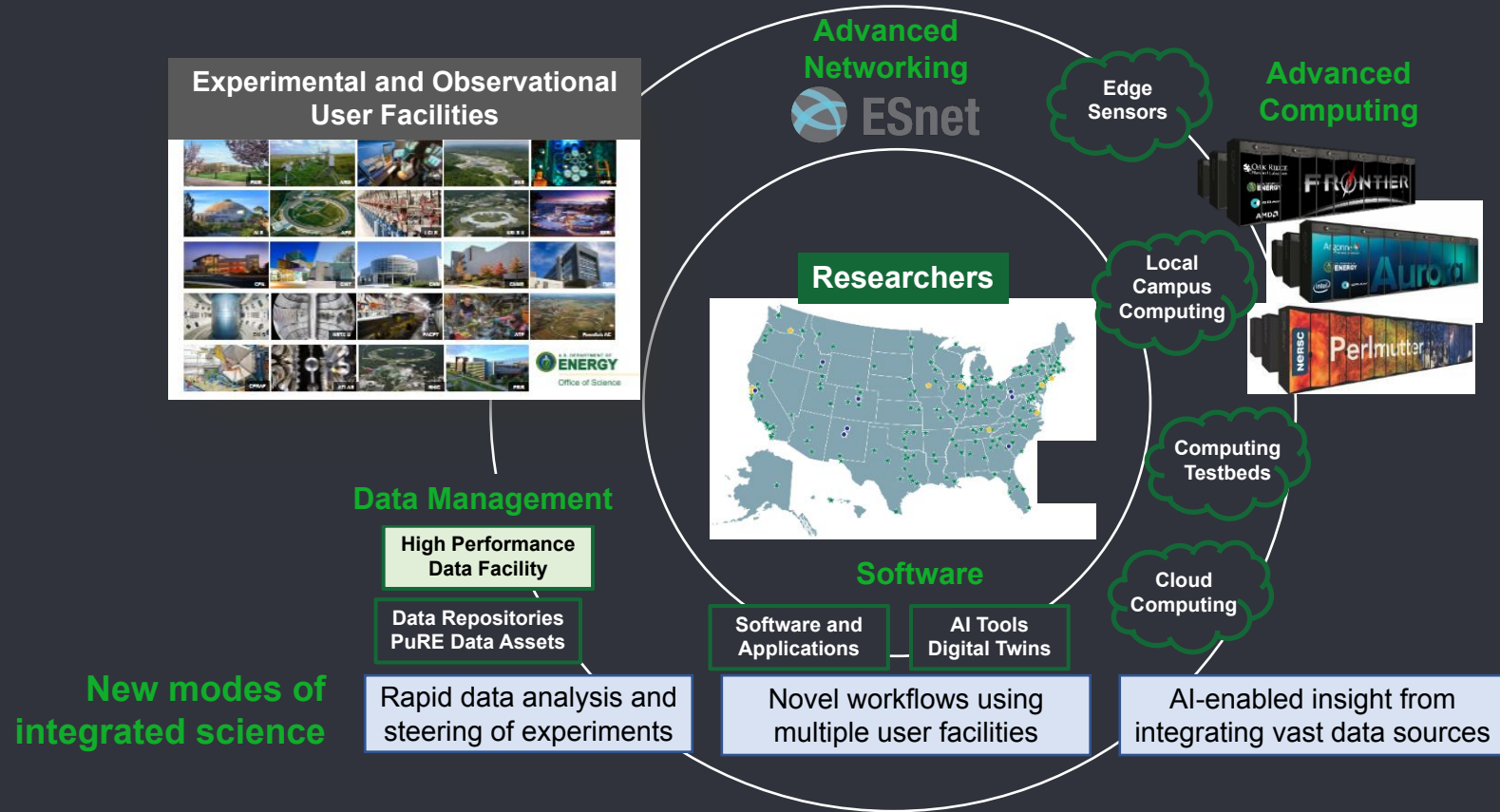
# Exponential traffic growth



# Multi-facility integration and Data intensive flows are drivers for a better application-network integration

- Integrated Research Infrastructure (IRI)
- HPDF Hub and Spoke model
- AI for Science (FASST)

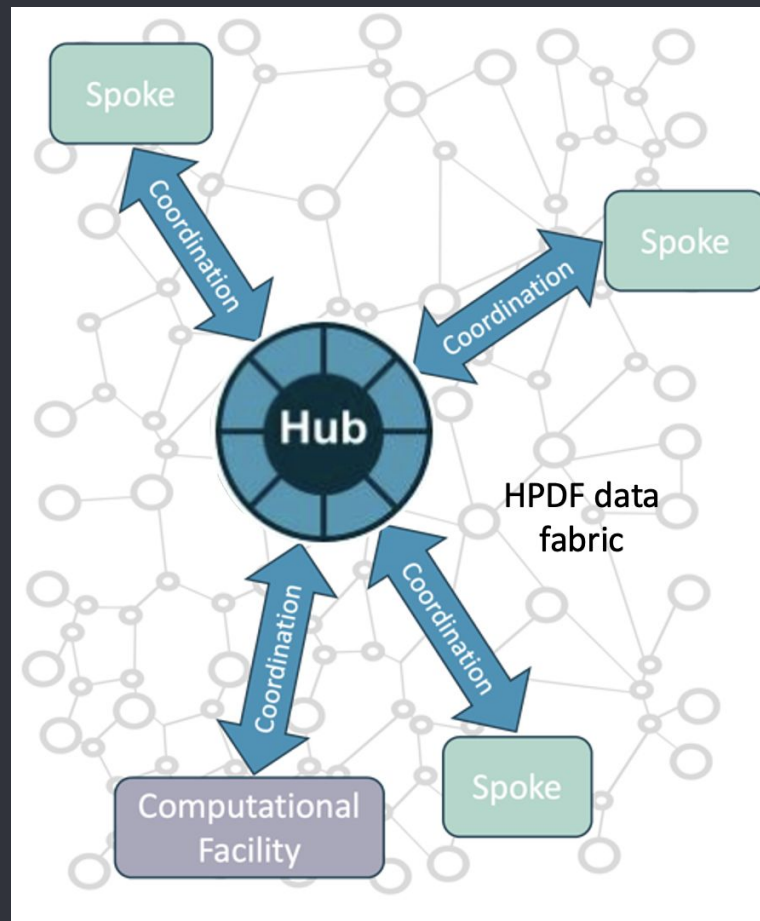
# ESnet's central role in IRI: connecting researchers and user facilities to HPC, HPDF, Cloud, testbeds, and the edge



## ESnet services are integral to HPDF distributed facility

Integration with ESnet will be key for

- Advanced data services to handle science workflows
- Geographically and operationally resilient active-active failover
- Deploying distributed computing or storage resources between Hub(s) and Spokes





# Distributed, large data infrastructure and movement needed for foundational and large-parameter **AI models**

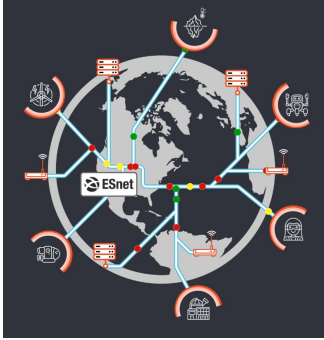
## Chapter 19: Data Infrastructure

- In pursuit of the active collective memory concept introduced above, we may imagine a **malleable, tiered set of AI foundation models with high bandwidth connections**.
- These varying foundation models would also connect and

expected to advance quickly. Ultra-fast and reliable ESnet connectivity, broadly deployed data and computing connections, and extensive task automation [10] will make it trivial to implement and run flows that link experiments and simulations with AI agents, data repositories, and other elements of an AI-enabled and AI-enabling DOE science infrastructure. Continued work on policy will be required to avoid bureaucratic barriers to effective resource sharing and collaborative work.



# Talk Flow



ESnet Introduction: Data Circulatory System



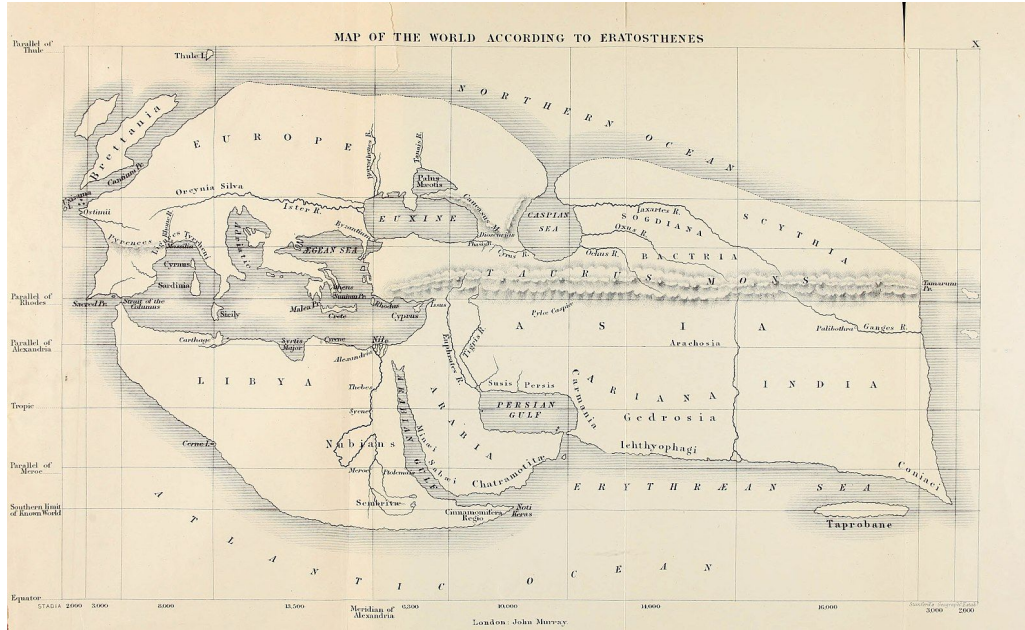
**Analogy: Case for a richer Network-Application interface**

$f(x)$

Journey towards application-network integration



# If a {human} was a packet, how did it travel?



Map by  
Eratosthenes of Cyrene  
(276 B.C - ~194 B.C)

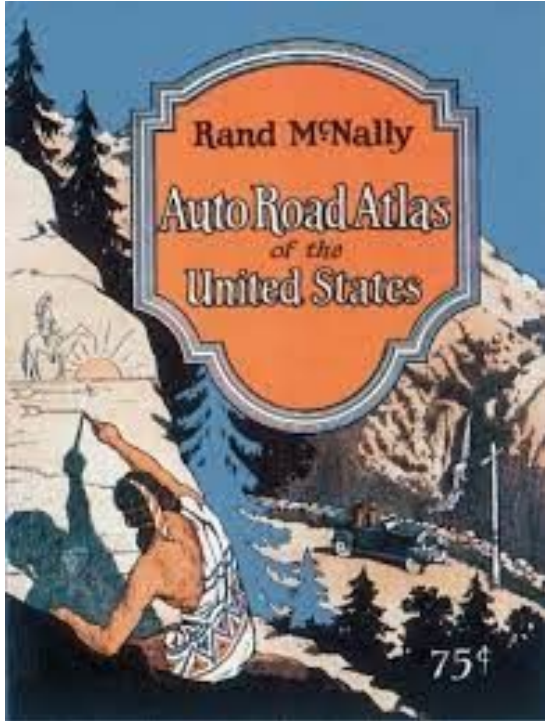
Father of Geography



Equivalent of  
hop-by-hop,  
store-forward  
routing

- Maps introduced rough guide on directions and location
- Tools helped to align to those directions
- Refinement of directions was based on observing intermediate landmarks or asking

# If a {human+automobile} was a packet, how did it travel?



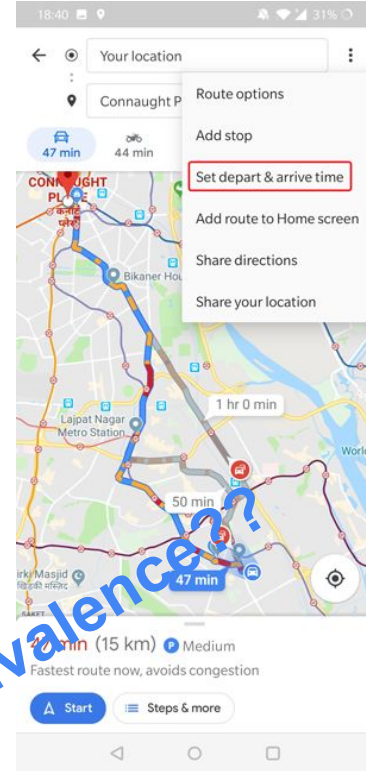
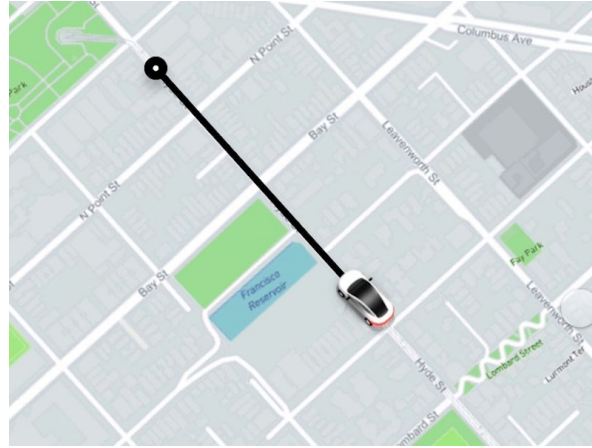
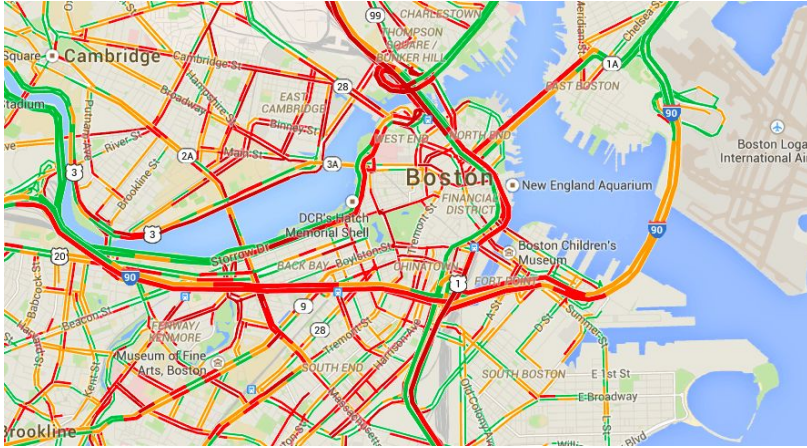
With the advent of automobiles, Rand McNally published its first road atlas called “Auto Chum” in 1924

Routes were pre-computed by human brain before getting on the road, re-routing happened on the fly by stopping and manually determining the route again

*Equivalent of  
MPLS or Layer  
2.5 based  
routing*

Prediction and planning was hard, and depended on personal experience or hearsay

# With the advent of digital technology, the {human + vehicle} packets have real-time + historical knowledge

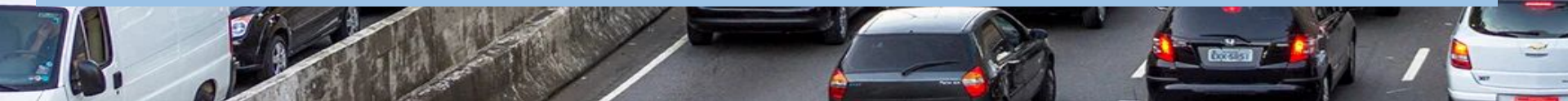


Real-time traffic and traffic prediction helps plan with just in time information, and features such as dynamic rerouting and updated accurate data on when the destination will be reached

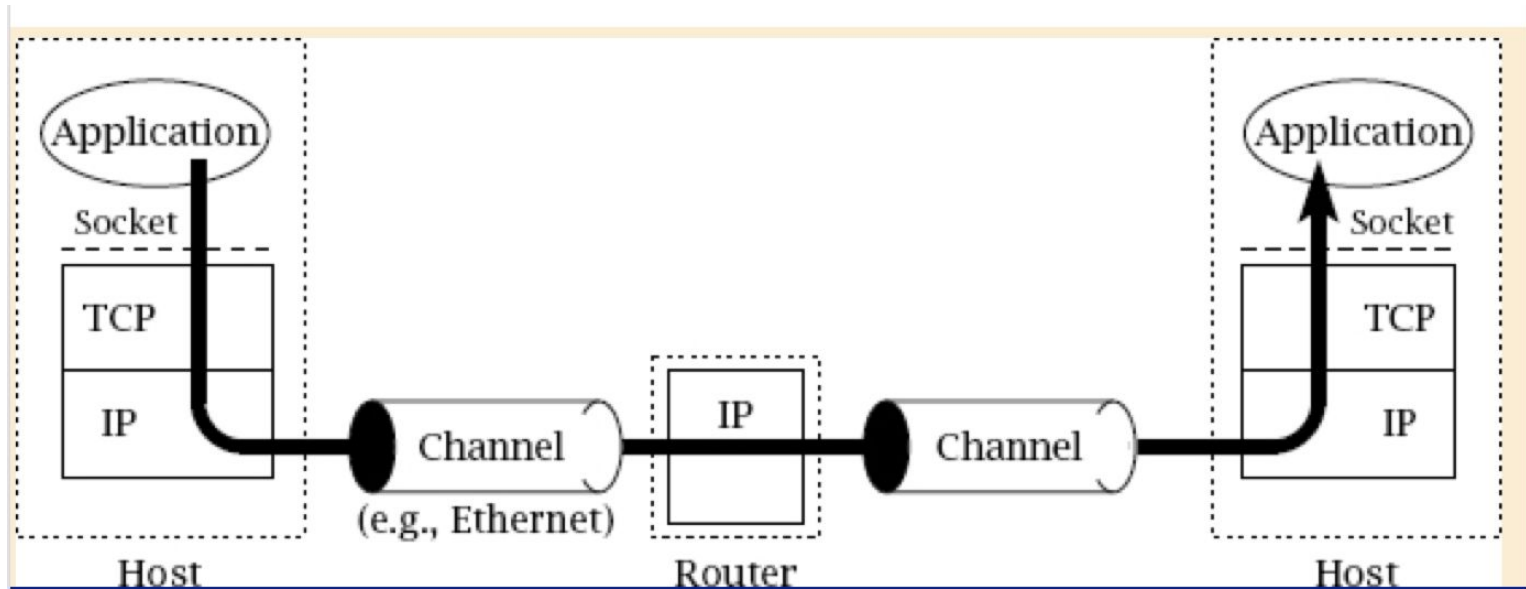




**Aspirational Goal:** How can we provide predictability and resilience to certain data flows given the huge variability of background traffic?



# Looking deeper: the socket interface



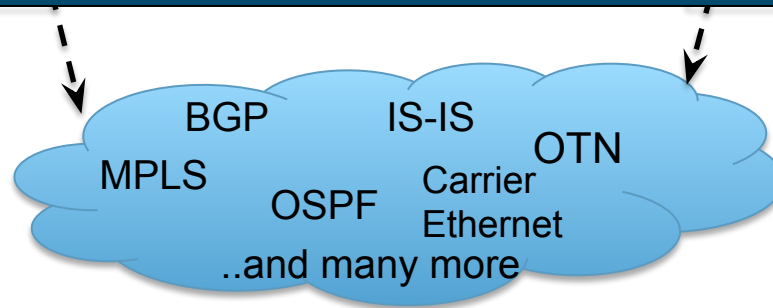
Adapted from: [Donahoo](#), Michael J., and Kenneth L. Calvert. TCP/IP sockets in C: practical guide for programmers. Morgan Kaufmann, 2009.

# The Unix Socket Interface: Most Successful Data Plane Abstraction



result = socket (af, type, protocol)

- Gives file system like abstraction to the network
- Hides the complexity of the network and its operations

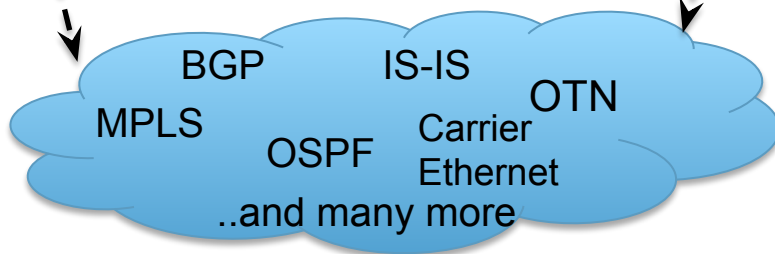


Complexity:  
6000+ IETF RFCs  
ITU-T  
IEEE  
GSM  
Others...

# The Unix Socket Interface: Network became a “black box”

- Application gets no feedback on the progress of the transfer
- There is no reasons given when a transfer fails, the only approach is try again, and again.....
- Network has no responsibility (unlike UPS or Amazon...)

The socket interface was not built for massively shared networks



Complexity:  
6000+ IETF RFCs  
ITU-T  
IEEE  
GSM  
Others...





Globus Transfer requests over last two years; those involving NERSC in red

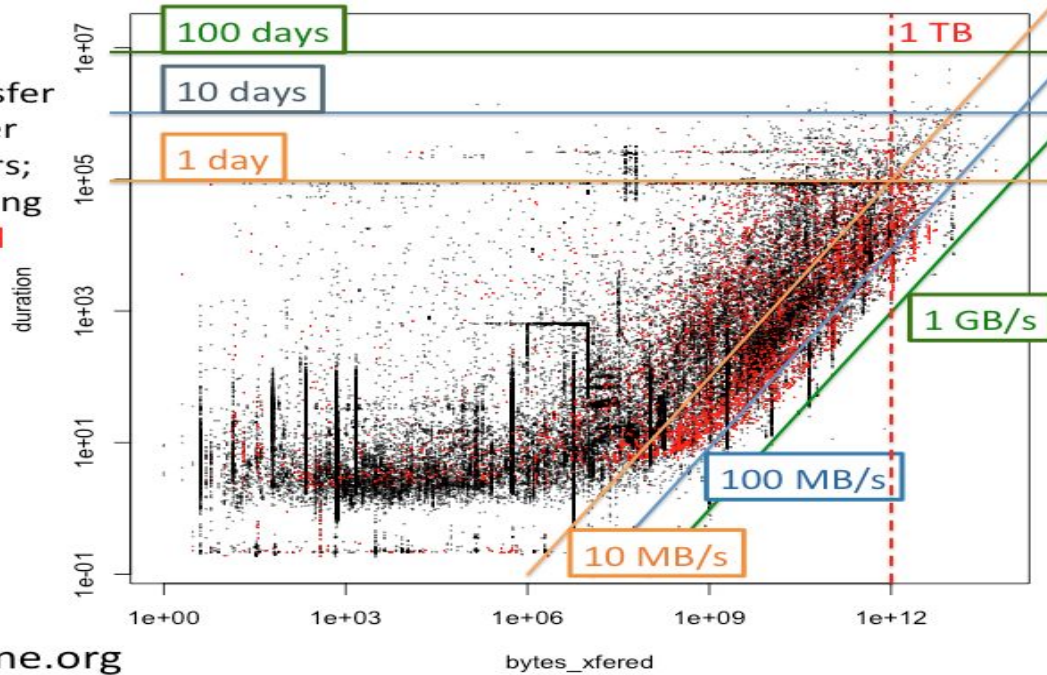


Image from Ian Foster

globusonline.org

Goal: How do we provide workflows a non-onerous but rich interface that allows two way exchange of *relevant* information that help manage the workflows overall objective?



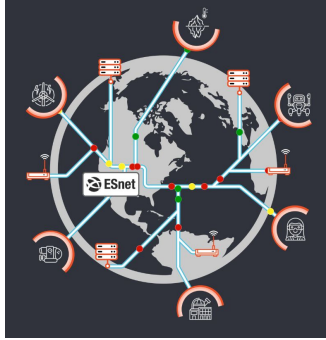
# Network-application interface experience is roughly equivalent to the 1900's maps

- Applications send data (packets) without any knowledge of the state of the end-to-end path, the networks it is transiting etc.
- Their knowledge is limited to rudimentary information like local link state, reachability, and round-trip time
- Real-time monitoring and knowledge does not propagate to the applications, and even when available, not across domains
- Applications have no way to query the network or state their requirements for certain critical flows

While the applications have themselves benefitted from the advent of the Internet, the **art of networking itself is yet to take full advantage of its own feature set and offer advanced capabilities** \*\*

*\*\* there are proprietary advancements in this area and research interest*

# Talk Flow



ESnet Introduction: Data Circulatory System



Analogy: Case for a richer Network-Application interface

$f(x)$

**Journey towards application-network integration**

# Phase 1: Grid Applications drove the first exploration

## Project DRAC: Creating an applications-aware network

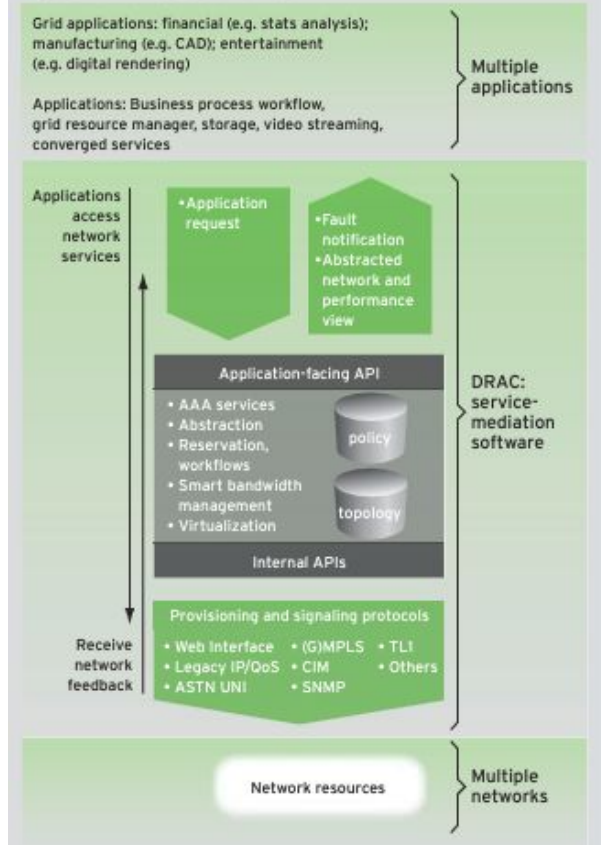
by Franco Travostino, Rob Keates, Tal Lavian, Inder Monga, and Bruce Schofield

Intelligent networking and the ability for applications to more effectively use all of the network's capability, rather than just the transport "pipe," have been elusive. Until now. Nortel has developed a proof-of-concept software capability – service-mediation "middleware" called the Dynamic Resource Allocation Controller (DRAC) – that runs on any Java platform and opens up the network to applications with proper credentials, making available all of the properties of a converged network, including service topology, time-of-day reservations, and interdomain connectivity options. With a more open network, applications can directly provision and invoke services, with no need for operator involvement or point-and-click sessions. In its first real-world demonstrations in large research networks, DRAC is showing it can improve user satisfaction while reducing network operations and investment costs.

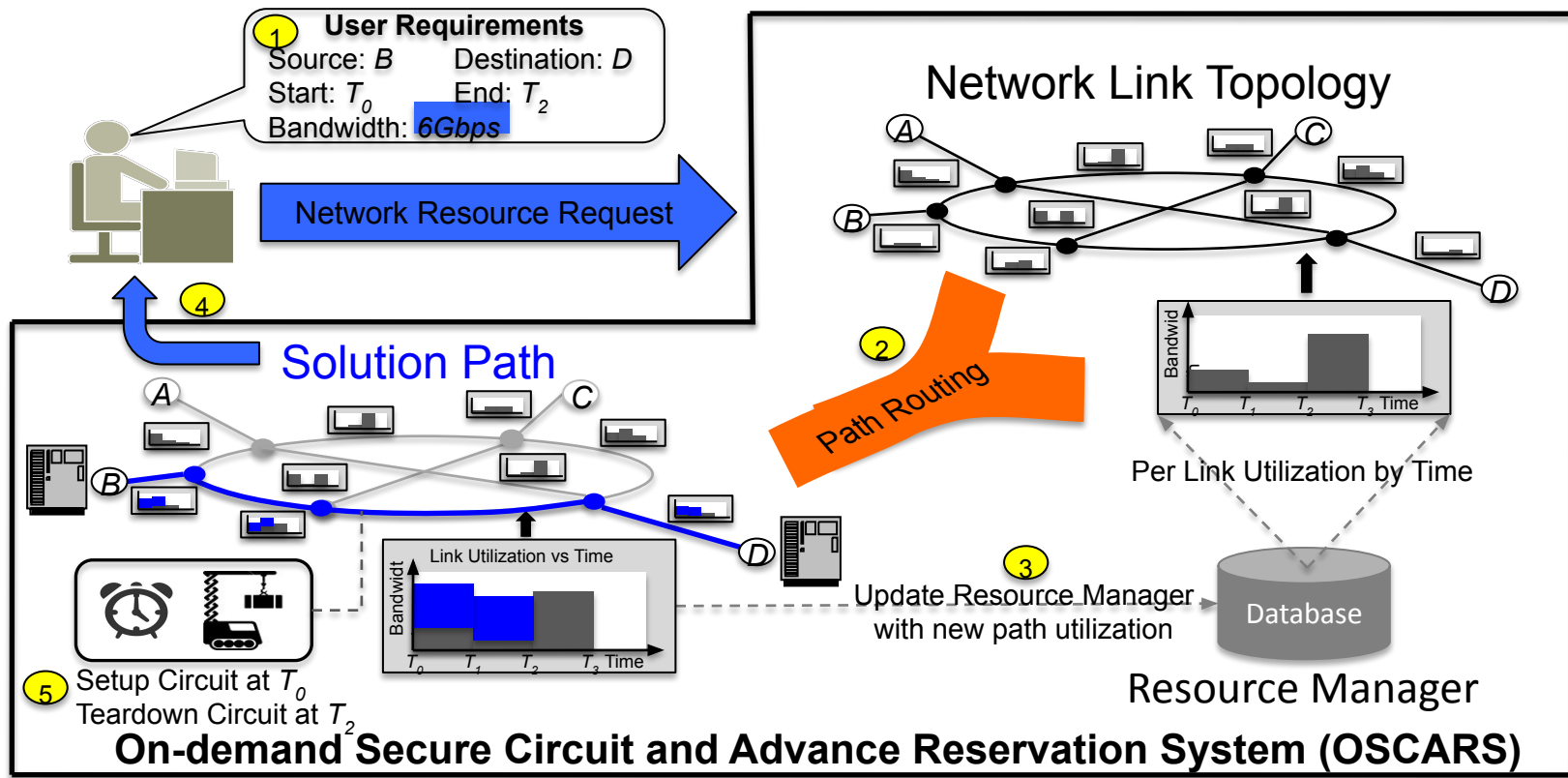
Moreover, as data travels through the end-to-end network and across different networks, it typically encounters different types of network technologies – from packet, circuit, wireless, and wireline to various access environments – each with its own separate topologies, protocols, and features, again leading to missed opportunities or high CapEx/OpEx costs.

### Dynamic Resource Allocation Controller

Figure 1. DRAC core framework

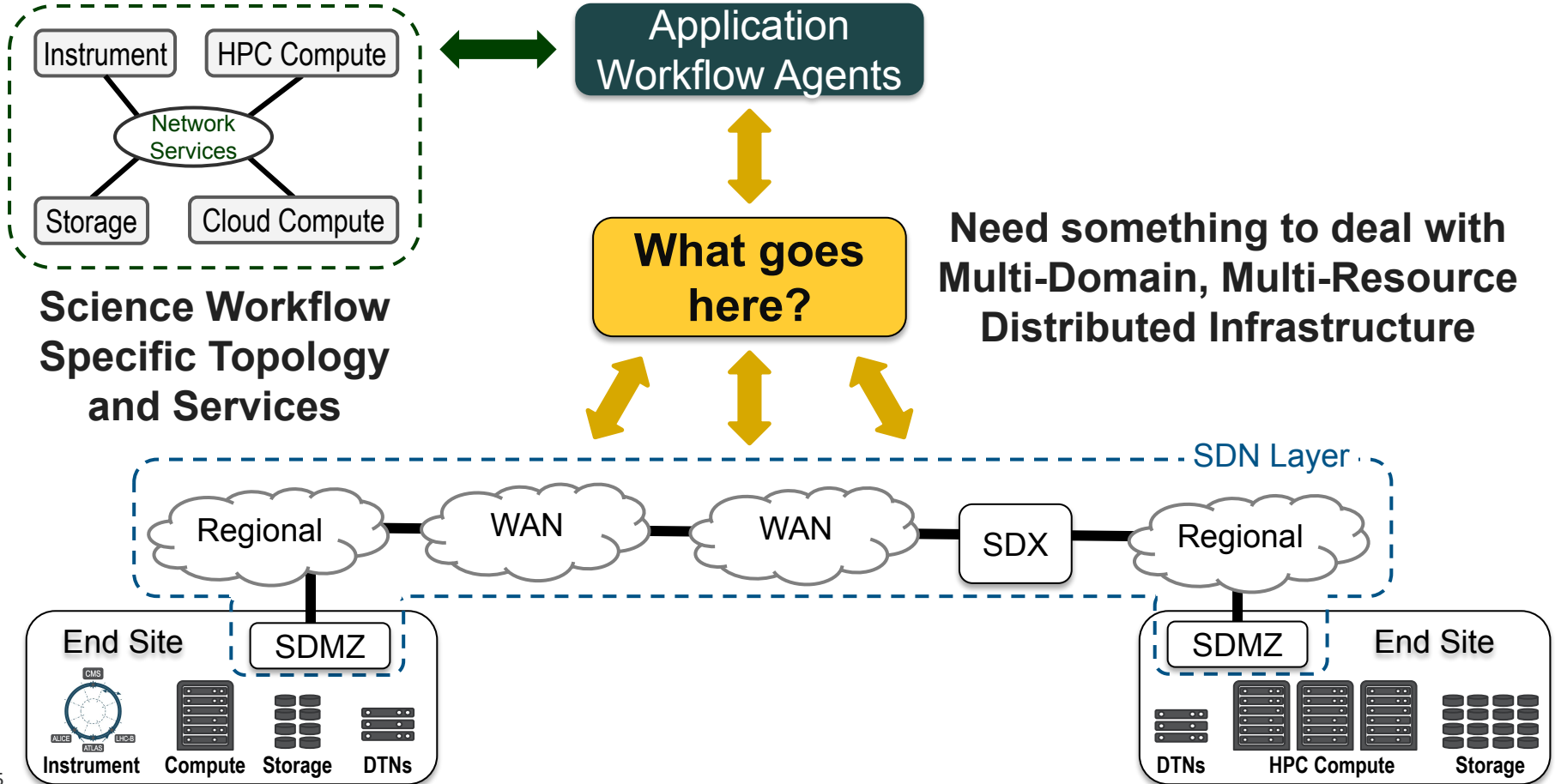


# Phase 2: Manual reservation with guaranteed bandwidth

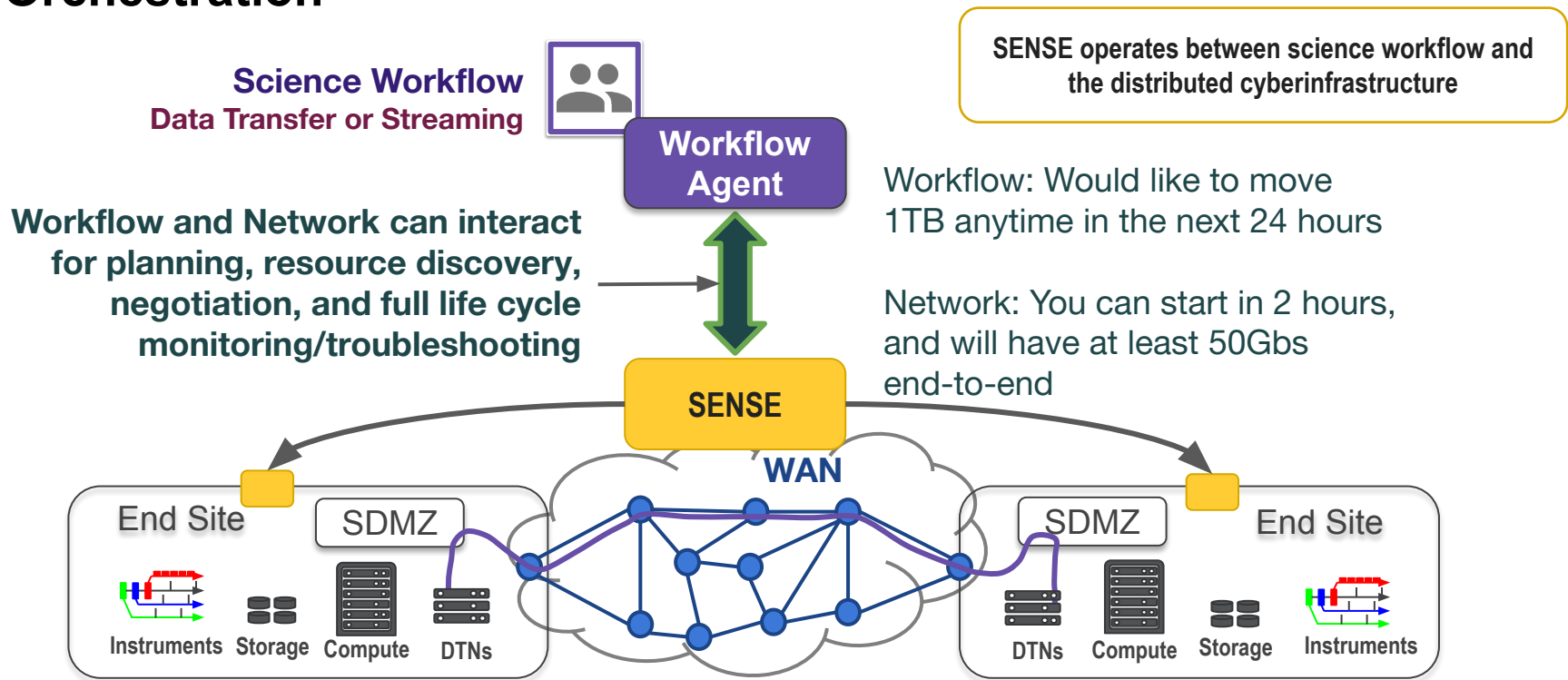


Operator driven, edge to edge automated setup, No end-systems knowledge or application interaction

# Phase 3: Focus on data mover to data mover

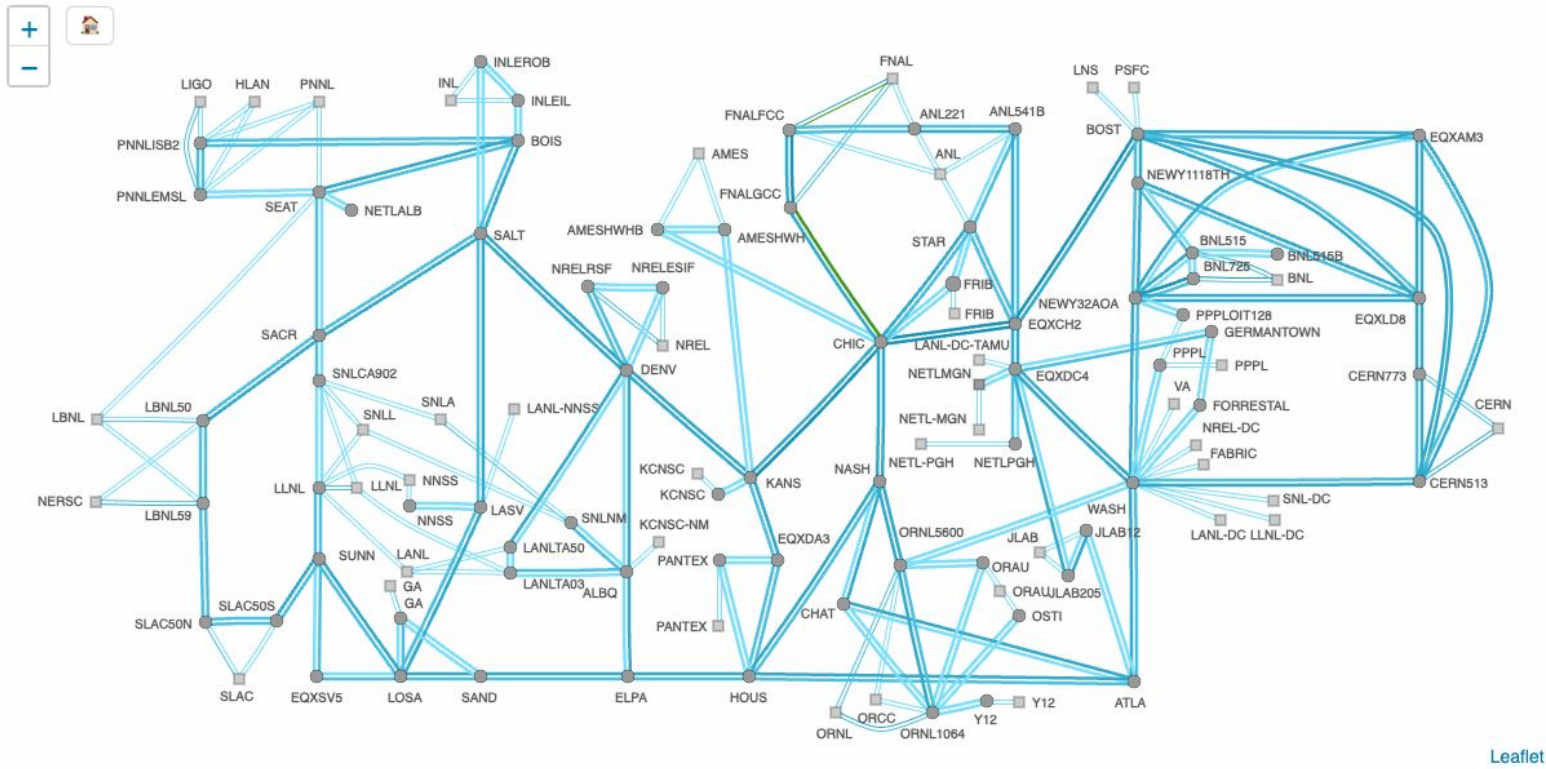


# Elevate Network to **First Class Resource API** driven Automation and Orchestration



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

# Allows the network to traffic engineer to fully utilize all network paths

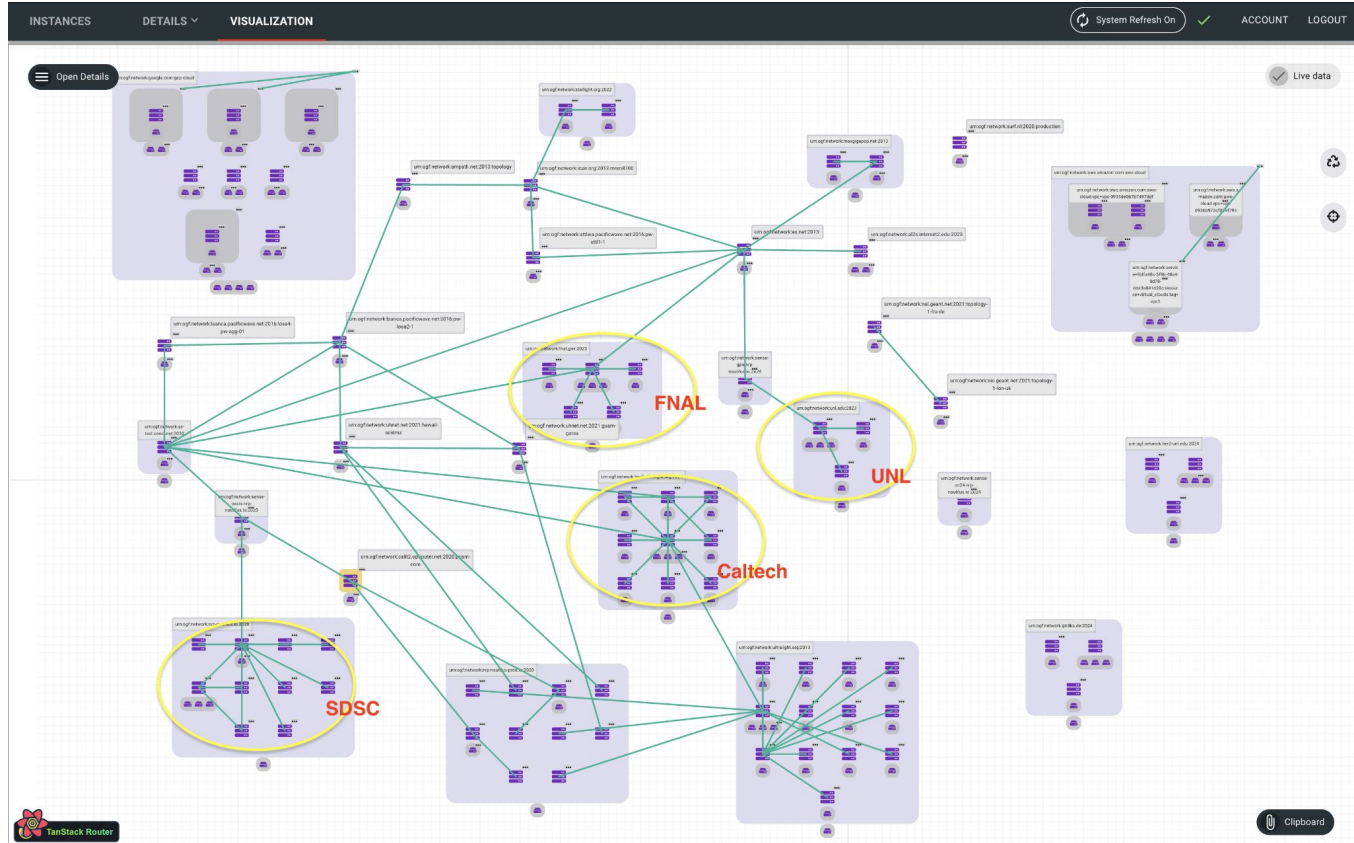


Leaflet

Network topology from my.es.net



# Semantic models help create visual representations





# Ultimate vision: SENSE – 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.

# Workflows can express their intent with SENSE API (also part of IRI Interfaces working group)

SMARTBEAR  
SwaggerHub

SENSE-O-Intent-API 2.0.3

Info

Tags

Servers

Search

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

```
1 openapi: 3.0.2
2 info:
3   version: 2.0.3
4   title: SENSE-O Northbound Intent API
5   description: StackV SENSE-O Northbound REST API Documentation
6
7 servers:
8   - url: "https://dev1.virnao.com:8443/StackV-web/restapi"
9
10 security:
11   - oAuth2Keycloak: []
12
13 tags:
14   - name: workflow_combined
15     description: |-
16       methods for single-phase workflows (minimal provisioning
17         steps)
18         /instance/{siUUID}/{action} uses `provision`, `cancel`
19         and `repvovision` calls.
20
21   - name: workflow_phased
22     description: |-
23       methods for two-phase commit workflows (useful for co
24         -scheduling)
25         /instance/{siUUID}/{action} uses `propagate`, `release`,
26         `reinstate` and `commit` calls.
27
28   - name: service
29     description: service workflow methods
30
31   - name: instance
32     description: Service instance methods
33
34   - name: profile
35     description: Profile methods
```

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

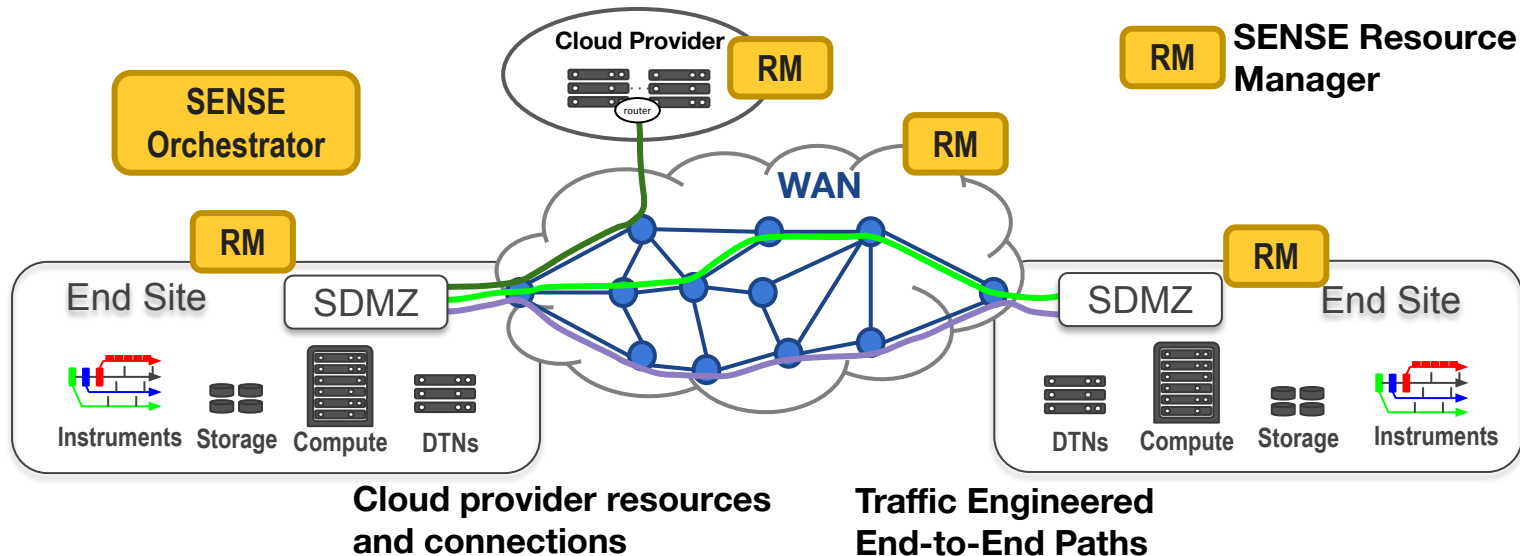
https://dev1.virnao.com:8443/StackV-we... Authorize

workflow\_combined

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

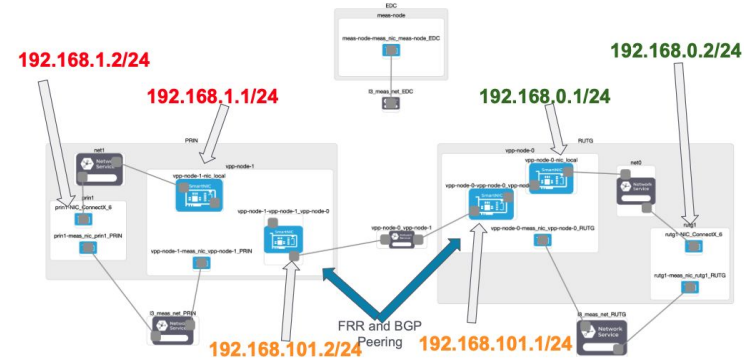
# Multi-Resource Orchestration

- 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



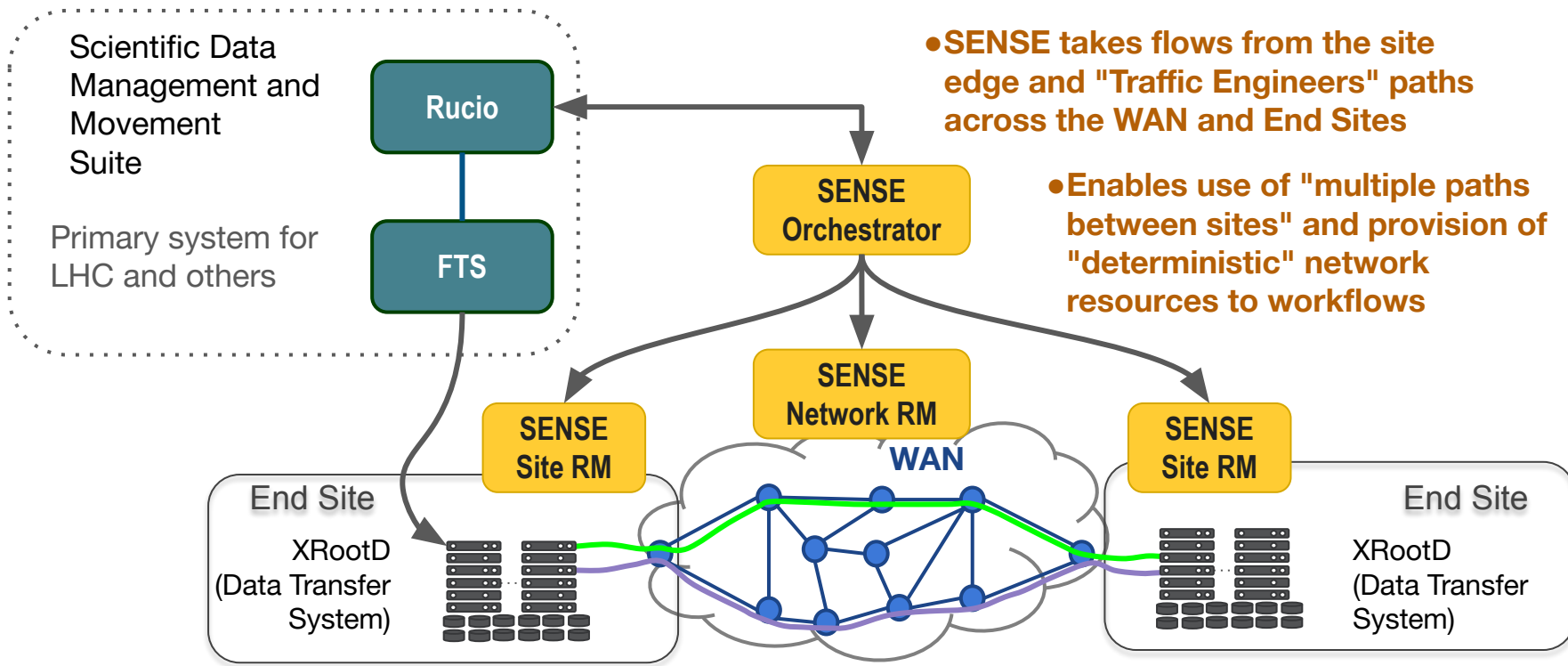
# Software Router for SENSE/Rucio on FABRIC

- **FABRIC** - Nation wide programmable network, provides GPU, FPGAs, NICs, QoS, Interconnect national facilities. Allows to design, test applications, protocols and services at any node in the network
- SENSE/Rucio need to support control at Sites without network device access
- Hardware/Software in use:
  - ConnectX-6 (PCI passthrough, 2x100G)
  - VPP with DPDK
  - FRRouting (without/with DPDK via VPP)
  - FreeRTR with DPDK
- Stable 50Gbps with 2 cores/4gb RAM VM (FRRouting only, no DPDK)
- VPP - 60 Gbps (with DPDK)
- FreeRTR - 30 Gbps (no Jumbo frames support)



# SENSE and Rucio/FTS/XRootD Interoperation (DC24 and beyond)

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



# DC24 SENSE/Rucio (Network Orchestration)

The objective was to provide Rucio with capabilities to request network services via SENSE in order to:

- a) improve accountability,**
- b) increase predictability,**
- and c) isolate and prioritize transfer requests.**

This project used a dedicated Rucio as well as XRootD instances so it would not interfere with Production systems. Data was transferred across a mix of production and next generation network paths.



Between Fermilab, Caltech, UCSD Rucio-DMM/SENSE-FTS-XRootD multiple Rucio-triggered data flows were managed between multiple pairs of sites; The modify feature of DMM was used to change bandwidth allocation on the fly in response to Rucio requests. The following Quality of Service policies were demonstrated: Hard QoS / Soft QoS on Server; Hard QoS at the network level. DMM Real time API-driven FTS tuning was used to adjust active/max transfers settings. Additional US-CMS Tier2 sites will be evaluated for deployment.

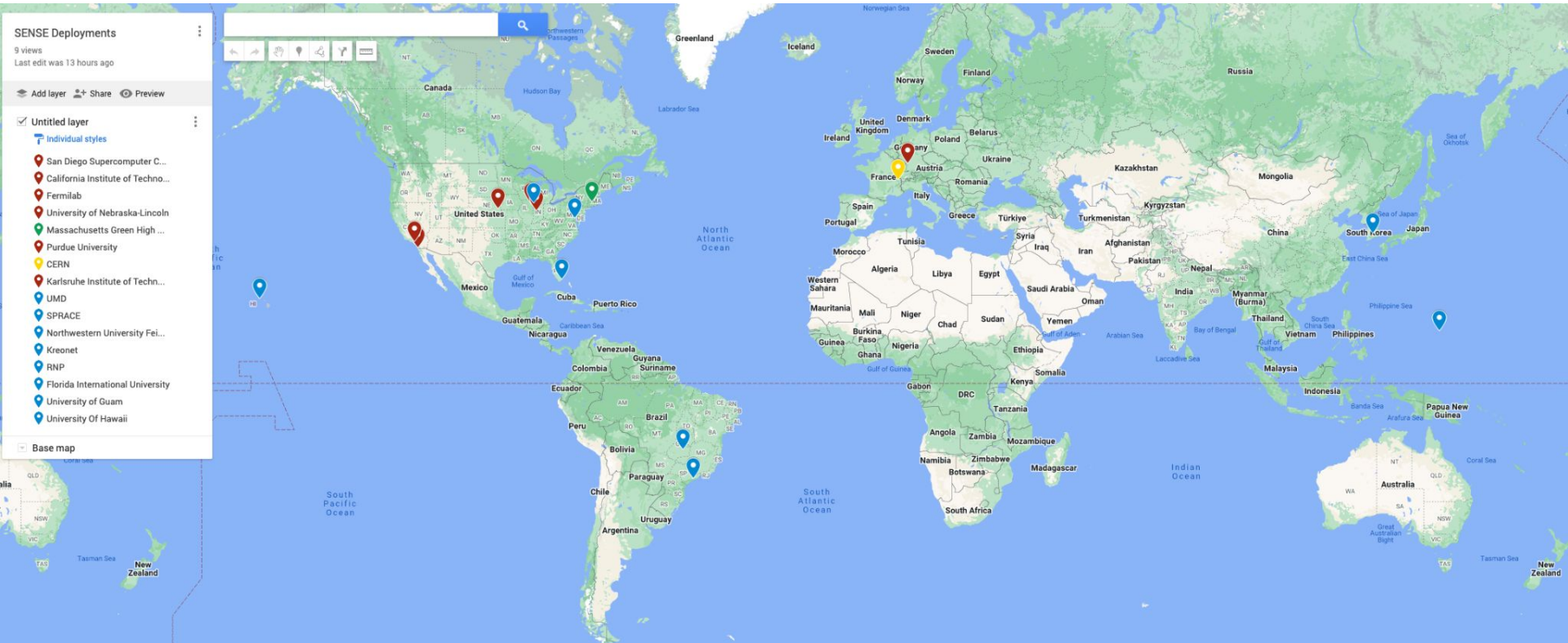
# SENSE and Rucio for USCMS (During DC24)





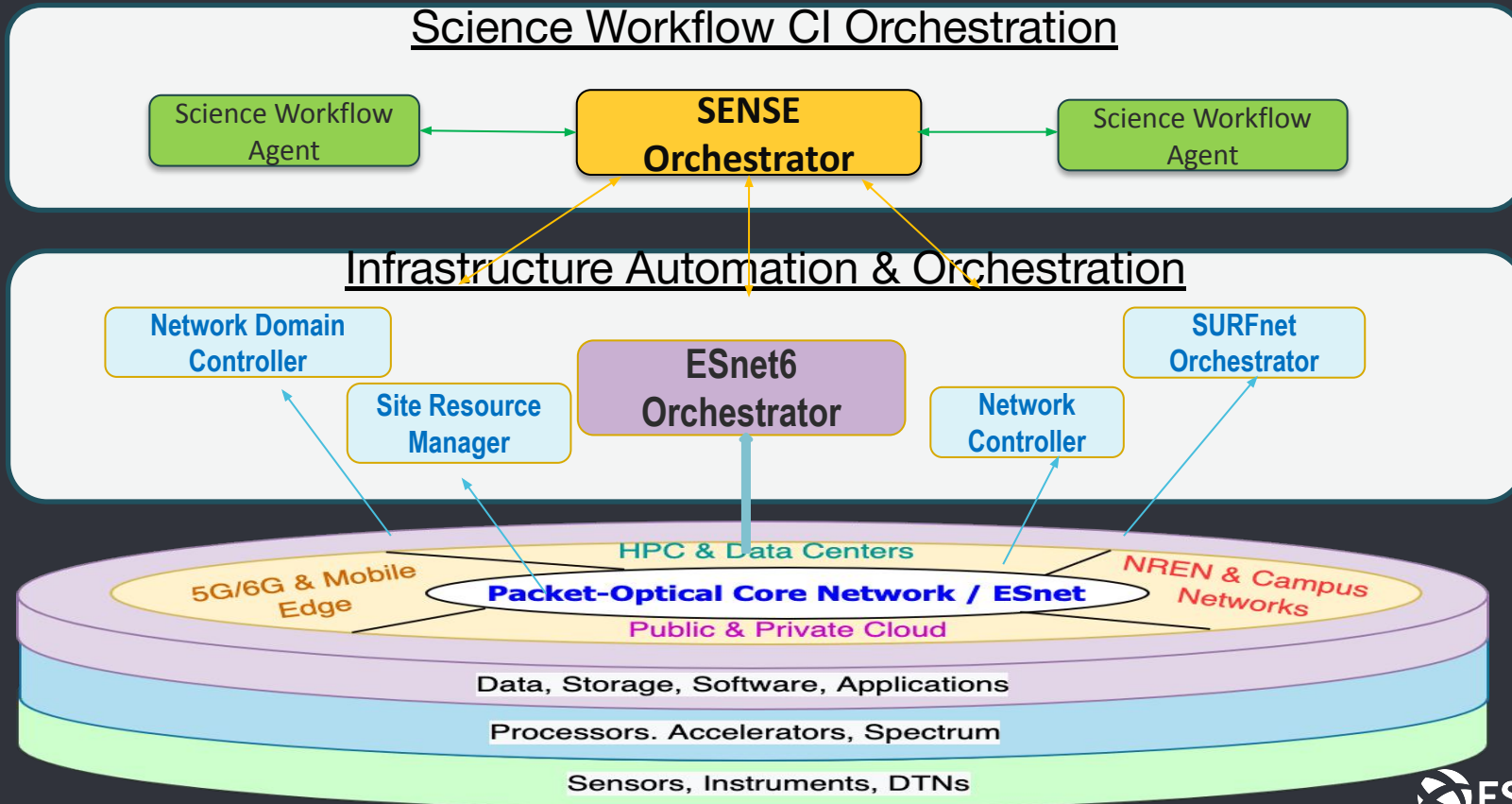
# SENSE deployments: 52 Servers, 16 sites, 20 network domains

## Free to try and use!!!





# An ESnet Vision for Automation & Orchestration



## For more details

1. Justas, Xi: Software-Defined Network for End-to-end Networked Science at Exascale at 11 Pacific on Wednesday  
<https://indico.cern.ch/event/1343110/contributions/6065564/>
2. Diego: Integration between Rucio and SENSE at 11:25 Pacific on Wednesday  
<https://indico.cern.ch/event/1343110/contributions/6119458/>

# Takeaways

- Increasingly data-intensive and complex workflows are motivating new conversations between the science applications and the research infrastructure
- Science application “infrastructure user experience” can be enhanced by ‘API’ interaction between applications, and research infrastructure, including networks
- Many experiments like DC24 will showcase the value of the integration, and highlight new challenges that need to be addressed

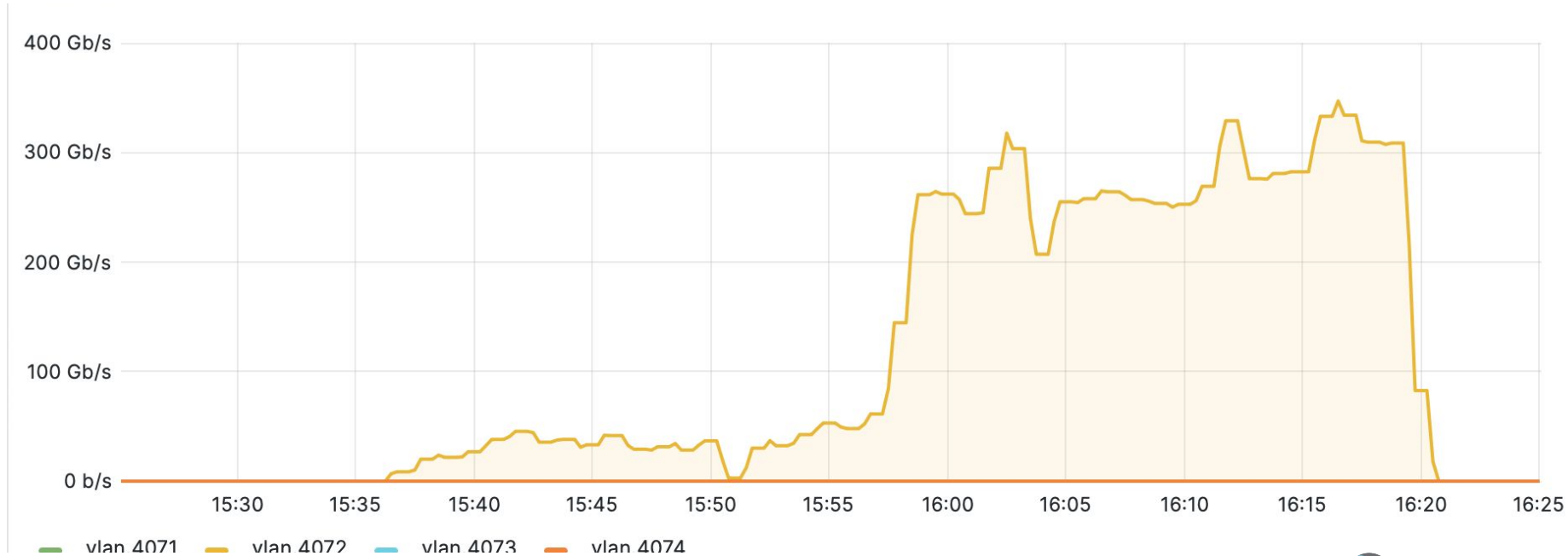
SCIENTIFIC DATA MANAGEMENT



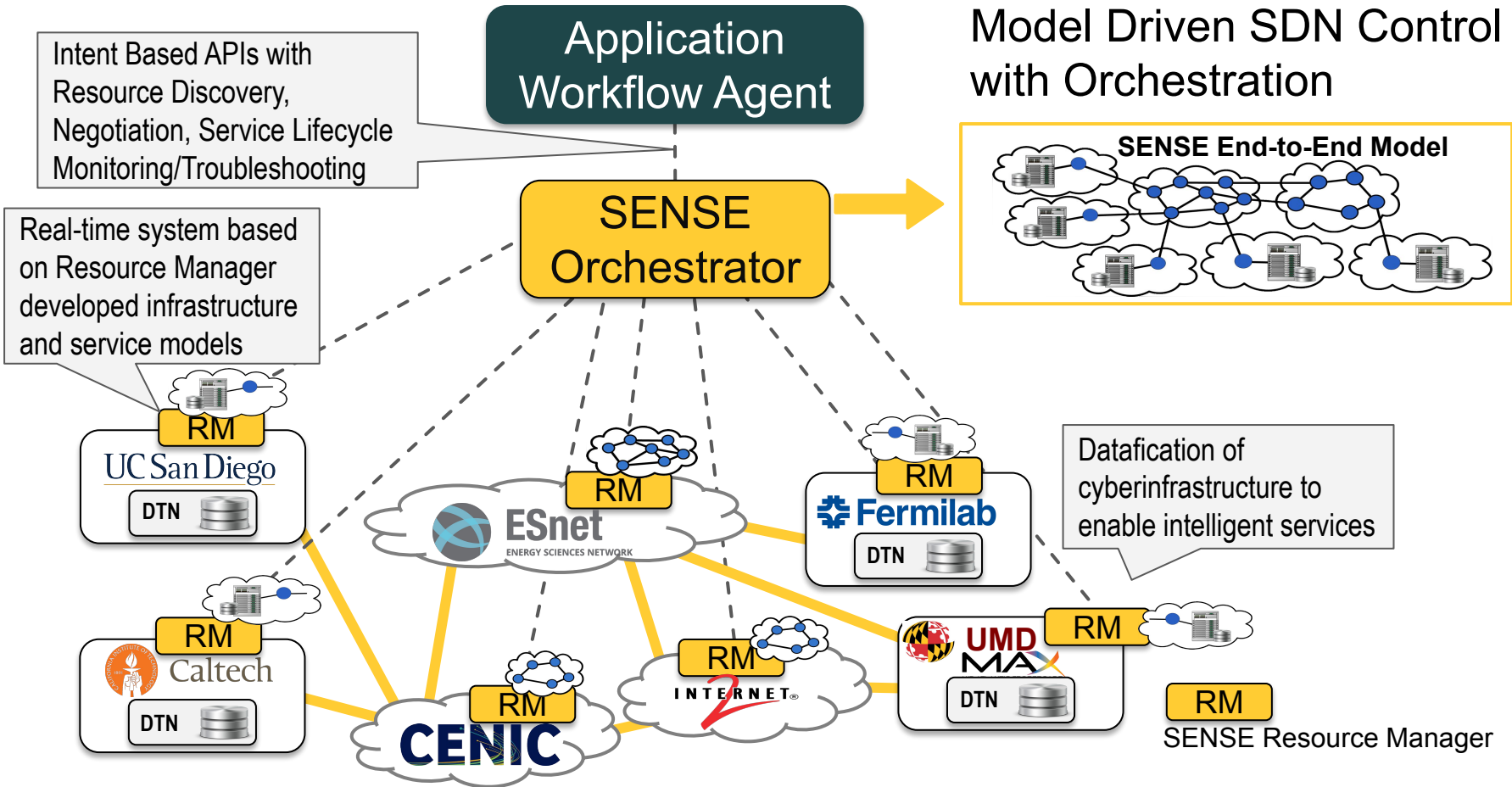
Backup

# SENSE and Rucio (Caltech Production storage)

Source	Destination	V0	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel	Rate (Last 1h)	Thr.
+ davs://redir-11.t2-sense.ultralight.org	davs://xrootd-sense-ucsd-redirector-112.sdsc.optiputer.net	cms	-	529	-	-	-	10472	-	-	100.00 %	32215.50 MiB/s

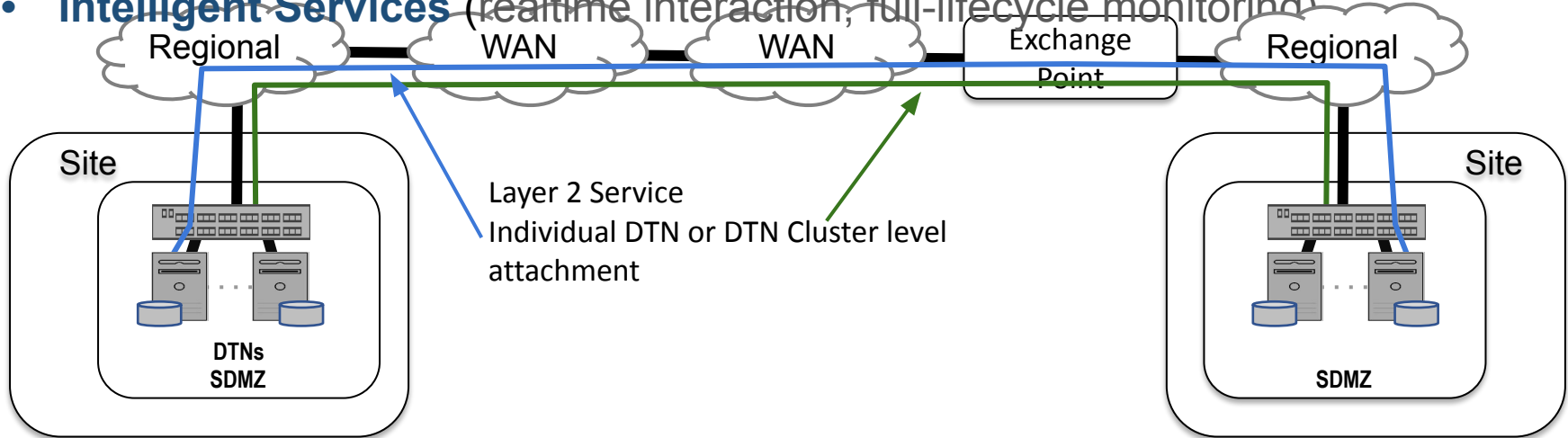






# SENSE Services

- **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)



# 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

The screenshot displays the SENSE interface for a Service Template Example. The left pane shows the user interface for editing the template, and the right pane shows the resulting JSON model.

**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:\*)

ADD REMOVE

**JSON Model**

```
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
```

JSON View SAVE AS SAVE DELETE Alias SUBMIT

# SENSE Papers and Info

- 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>

— Accepted Manuscript:

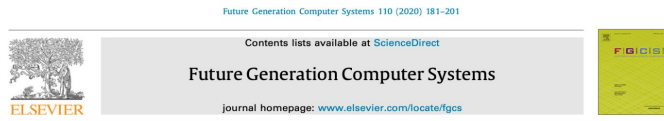
<https://arxiv.org/abs/2004.05953>

- [SENSE Northbound API Program](#)

— <https://app.swaggerhub.com/apis/xi-yang/SENSE-O-Intent-API>

- SENSE Website

— [sense.es.net](https://sense.es.net)



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

Inder Monga<sup>a</sup>, Chin Guok<sup>a</sup>, John MacAuley<sup>a</sup>, Alex Sim<sup>a</sup>, Harvey Newman<sup>b</sup>, Justas Balcas<sup>b</sup>, Phil DeMar<sup>c</sup>, Linda Winkler<sup>c</sup>, Tom Lehman<sup>d</sup>, Xi Yang<sup>e,f</sup>

<sup>a</sup>Energy Sciences Network, Lawrence Berkeley National Lab, Berkeley, CA, USA  
<sup>b</sup>Division of Physics, Mathematics and Astronomy, Caltech, Pasadena, CA, USA  
<sup>c</sup>Computing Division, Fermi National Accelerator Laboratory, Batavia, IL, USA  
<sup>d</sup>Computing, Environment and Life Science Division, Argonne National Lab, Argonne, IL, USA  
<sup>e</sup>Virsoo, Arlington, VA, USA  
<sup>f</sup>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

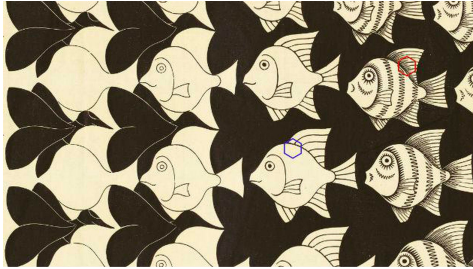
\* Corresponding author.  
E-mail addresses: imonga@es.net (I. Monga), chin@es.net (C. Guok), macauley@es.net (J. MacAuley), asim@es.net (A. Sim), newman@ep.caltech.edu (H. Newman), jbalcas@caltech.edu (J. Balcas), demar@fnl.gov (P. DeMar), winkler@mcs.anl.gov (L. Winkler), tiehman@virsoo.com (T. Lehman),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

<https://doi.org/10.1016/j.future.2020.04.018>  
0167-7390/Published by Elsevier B.V.

# Talk Flow



Analogy: Case for a richer Network-Application interface

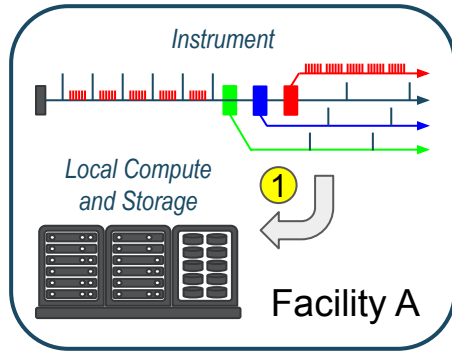


Data-intensive science: motivating the case for integrated research infrastructure

$$f(x)$$

Example of application-network integration

# Example of a Traditional Workflow for Large Scale Distributed Science Experiments

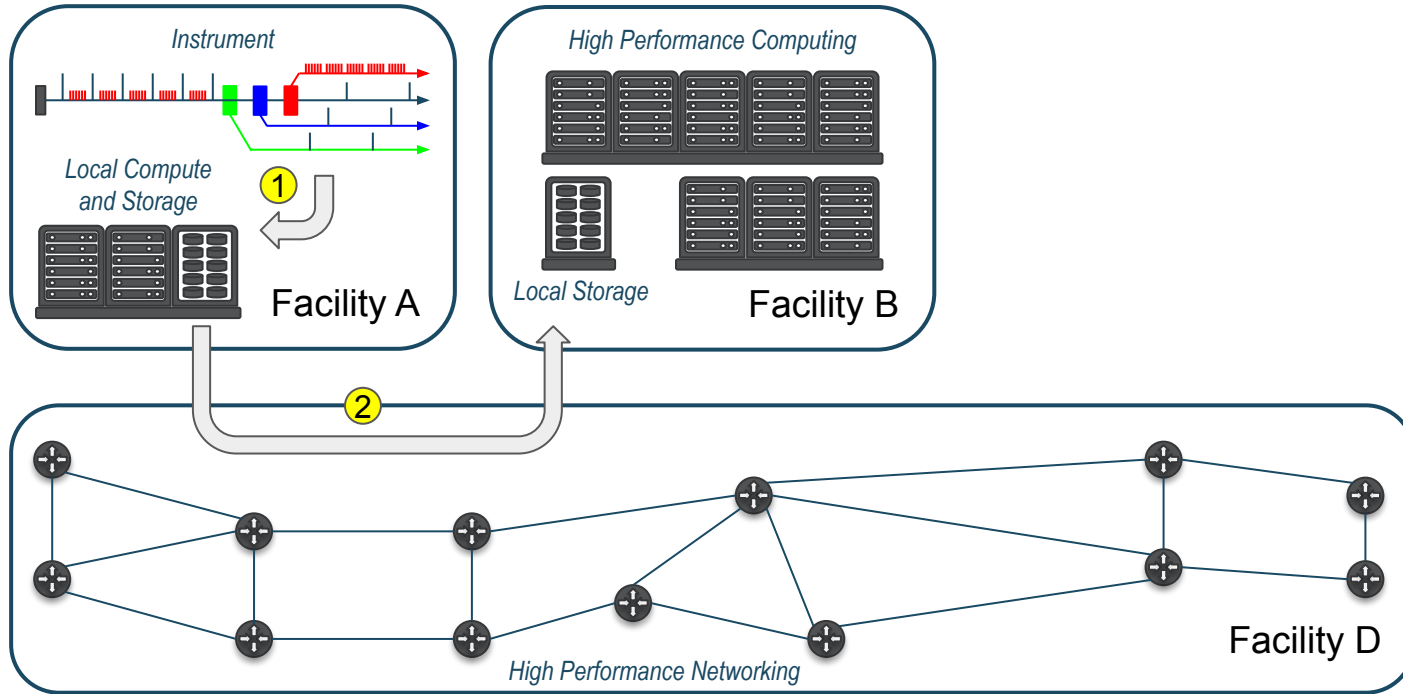


## Discrete Data Flow

1. Initial data processing on local compute and staged in local storage.



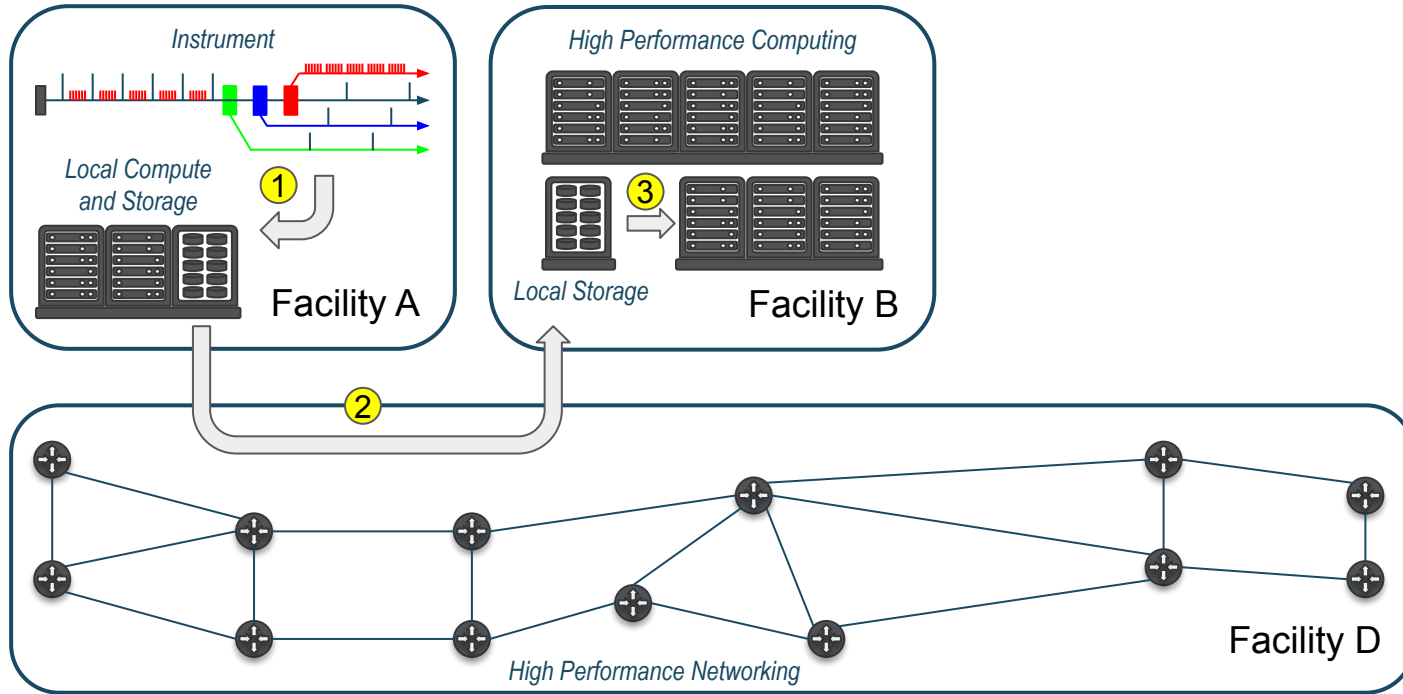
# Example of a Traditional Workflow for Large Scale Distributed Science Experiments



## Discrete Data Flow

1. Initial data processing on local compute and staged in local storage.
2. Data transfer over WAN and staged in HPC local storage.

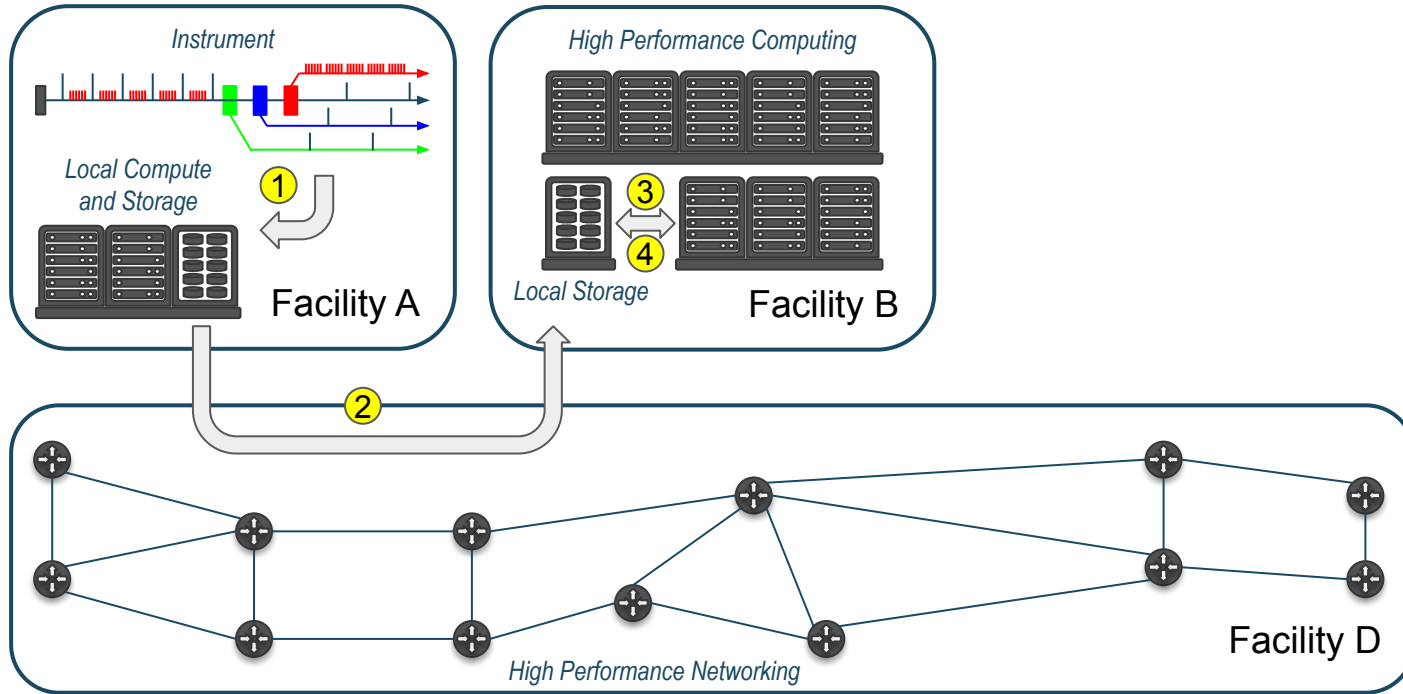
# Example of a Traditional Workflow for Large Scale Distributed Science Experiments



## Discrete Data Flow

1. Initial data processing on local compute and staged in local storage.
2. Data transfer over WAN and staged in HPC local storage.
3. HPC fetches data from local storage for processing.

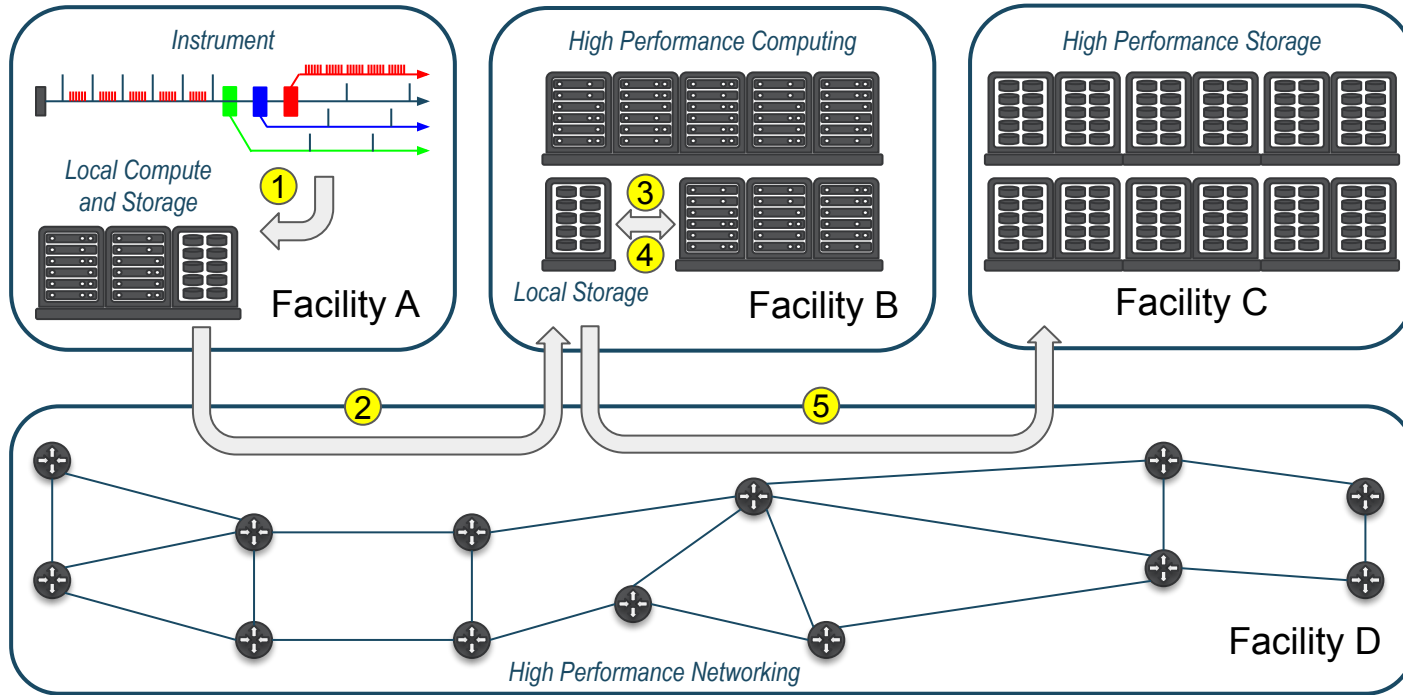
# Example of a Traditional Workflow for Large Scale Distributed Science Experiments



## Discrete Data Flow

1. Initial data processing on local compute and staged in local storage.
2. Data transfer over WAN and staged in HPC local storage.
3. HPC fetches data from local storage for processing.
4. HPC stages processed data in local storage for transfer.

# Example of a Traditional Workflow for Large Scale Distributed Science Experiments



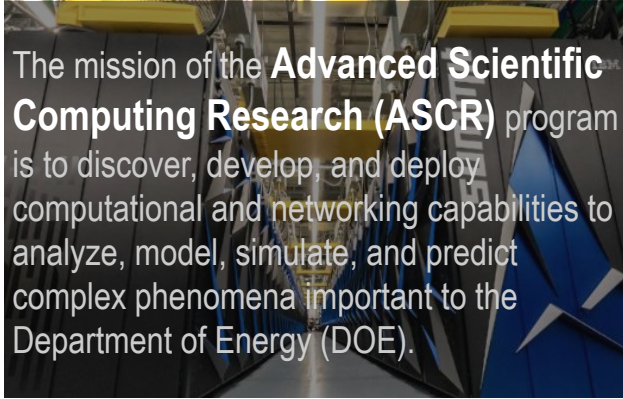
## Discrete Data Flow

1. Initial data processing on local compute and staged in local storage.
2. Data transfer over WAN and staged in HPC local storage.
3. HPC fetches data from local storage for processing.
4. HPC stages processed data in local storage for transfer.
5. Data transfer over WAN into HPS for long term storage.


# Where are the inefficiencies?

- Experiments use local compute to compensate for the lack of *reliably* schedulable (shared) compute resources.
  - Not every job is a good fit for an HPC.
  - Discrepancy between HPC job schedule and need for real-time computing to support the experiment.
- **Networks are treated as “black-box” entities, e.g., unpredictable performance, unknown status.**
  - **(Temporary) storage is used to stage data and compensate for lack of network performance predictability, resulting in multiple data transfers.**
- Access to resources across domains is inconsistent
  - Different security mechanisms and allocation policies.
  - Different APIs and architectures, affecting job portability.
- ○○○

# DOE Office of Science - Largest supporter of basic research in the physical sciences in the US (\$6-\$7 B)




The mission of the **Advanced Scientific Computing Research (ASCR)** program is to discover, develop, and deploy computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to the Department of Energy (DOE).



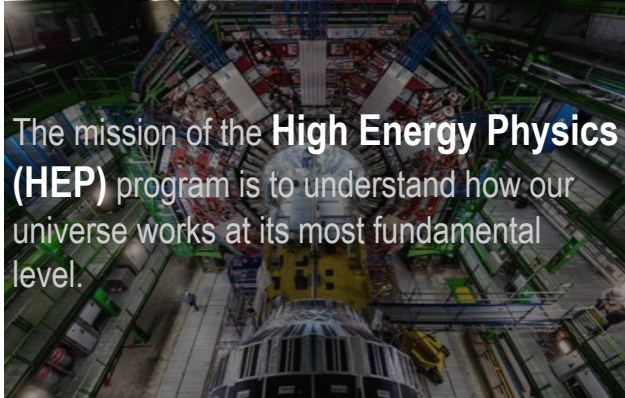
**Basic Energy Sciences (BES)** supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security.



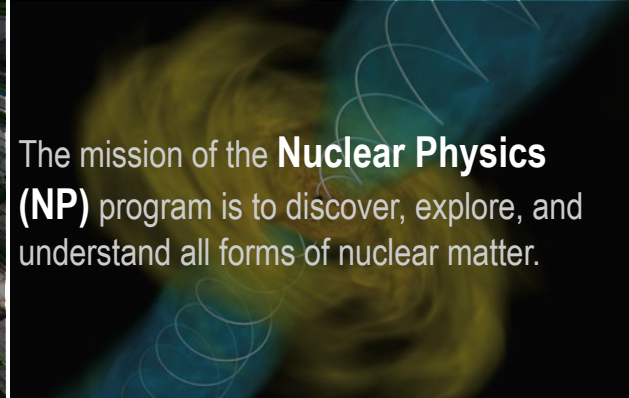
The mission of the **Biological and Environmental Research (BER)** program is to support transformative science and scientific user facilities to achieve a predictive understanding of complex biological, earth, and environmental systems for energy and infrastructure security, independence, and prosperity.



The **Fusion Energy Sciences (FES)** program mission is to expand the fundamental understanding of matter at very high temperatures and densities and to build the scientific foundation needed to develop a fusion energy source.















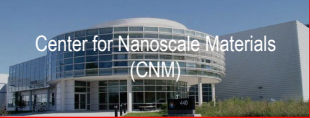














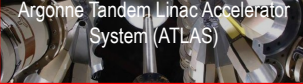




The mission of the **High Energy Physics (HEP)** program is to understand how our universe works at its most fundamental level.



The mission of the **Nuclear Physics (NP)** program is to discover, explore, and understand all forms of nuclear matter.



# DOE Office of Science - Uniquely positioned for large scale collaborative science with 28 large user facilities\*

<p>ASCR High End Computing (HEC)</p> <p>Argonne Leadership Computing Facility (ALCF)</p> 	<p>BES X-Ray Light Sources</p>				
<p>Oak Ridge Leadership Computing Facility (OLCF)</p> 	<p>Advanced Photon Source (APS)</p> 	<p>Linac Coherent Light Source (LCLS)</p> 	<p>Stanford Synchrotron Radiation Light Source (SSRL)</p> 	<p>Advanced Light Source (ALS)</p> 	<p>National Synchrotron Light Source II (NSLS-II)</p> 
<p>BES Nanoscale Science Research Centers (NSRCs)</p>					
<p>Oak Ridge Leadership Computing Facility (OLCF)</p> 	<p>Center for Functional Nanomaterials (CFN)</p> 	<p>Center for Integrated Nanotechnologies (CINT)</p> 	<p>The Molecular Foundry (TMF)</p> 	<p>Center for Nanophase Materials Sciences (CNMS)</p> 	<p>Center for Nanoscale Materials (CNM)</p> 
<p>National Energy Research Scientific Computing Center (NERSC)</p> 	<p>BES Neutron Scattering Facilities</p>				
<p>ASCR High Performance Scientific Network</p> <p>Energy Sciences Network (ESnet)</p> 	<p>Spallation Neutron Source (SNS)</p> 	<p>High Flux Isotope Reactor (HFIR)</p> 	<p>Joint Genome Institute (JGI)</p> 	<p>Environmental Molecular Sciences Laboratory (EMSL)</p> 	<p>Atmospheric Radiation Measurement (ARM) user facility</p> 
<p>FES</p>			<p>BER</p>		
<p>Energy Sciences Network (ESnet)</p> 	<p>National Spherical Torus Experiment - Upgrade (NSTX-U)</p> 	<p>DIII-D National Fusion Facility (DIII-D)</p> 	<p>Facility for Advanced Accelerator Experimental Tests (FACET)</p> 	<p>Fermilab Accelerator Complex</p> 	<p>Accelerator Test Facility (ATF)</p> 
<p>HEP</p>					
<p>NP</p>					
<p>U.S. DEPARTMENT OF ENERGY Office of Science</p> 	<p>Argonne Tandem Linac Accelerator System (ATLAS)</p> 	<p>Continuous Electron Beam Accelerator Facility (CEBAF)</p> 	<p>Facility for Rare Isotope Beams (FRIB)</p> 	<p>RHIC</p> 	<p>PHENIX Relativistic Heavy Ion Collider (RHIC)</p> 



# DOE Office of Science - Uniquely positioned for large scale collaborative science with 28 large user facilities\*

<p>ASCR High End Computing (HEC)</p> <p>Argonne Leadership Computing Facility (ALCF)</p>	<p>BES X-Ray Light Sources</p> <p>Advanced Photon Source (APS)</p> <p>Linac Coherent Light Source (LCLS)</p>		<p>Stanford Synchrotron Radiation Light Source (SSRL)</p> <p>Advanced Light Source (ALS)</p>		<p>National Synchrotron Light Source II (NSLS-II)</p>
<p>Oak Ridge Leadership Com Facility (OLCF)</p>	<p>BES Nanoscale Science Research Centers (NSRCs)</p> <p>Functional Nanomaterials (CFN)</p> <p>Center for Integrated Nanotechnologies (CINT)</p> <p>The Molecular Foundry (TMF)</p> <p>Center for Nanophase Materials Sciences (CNMS)</p> <p>Center for Nanoscale Materials (CNM)</p>				
<p>National Energy Research Scientific Computing Center (NERSC)</p>	<p>BES Neutron Scattering Facilities</p> <p>Spallation Neutron Source (SNS)</p> <p>High Flux Isotope Reactor (HFIR)</p>		<p>BER</p> <p>Joint Genome Institute (JGI)</p> <p>Environmental Molecular Sciences Laboratory (EMSL)</p> <p>Atmospheric Radiation Measurement (ARM) user facility</p>		
<p>ASCR High Performance Scientific Network</p> <p>Energy Sciences Network (ESnet)</p>	<p>FES</p> <p>National Spherical Torus Experiment - Upgrade (NSTX-U)</p> <p>DIII-D National Fusion Facility (DIII-D)</p>		<p>HEP</p> <p>Facility for Advanced Accelerator Experimental Tests (FACET)</p> <p>Fermilab Accelerator Complex</p> <p>Accelerator Test Facility (ATF)</p>		
<p>NP</p> <p>Argonne Tandem Linac Accelerator System (ATLAS)</p> <p>Continuous Electron Beam Accelerator Facility (CEBAF)</p> <p>Facility for Rare Isotope Beams (FRIB)</p> <p>RHIC</p> <p>PHENIX Relativistic Heavy Ion Collider (RHIC)</p>					

This is ESnet

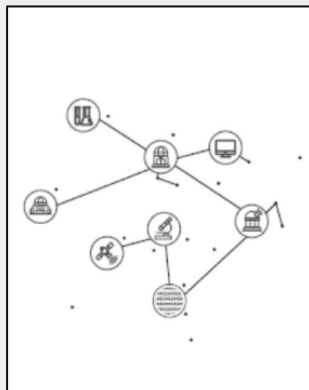
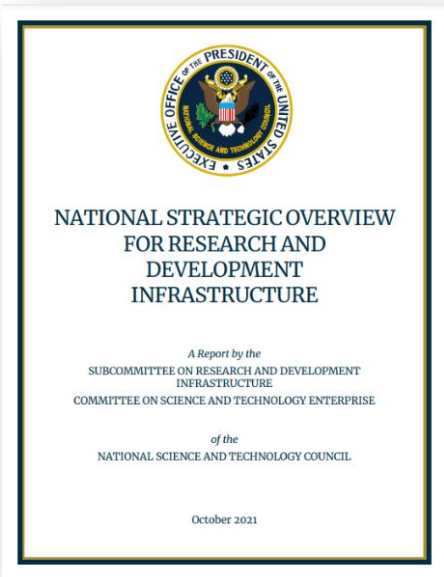


56 \*DOE Office of Science facilities also support other collaborations, e.g., LHC, LSST, etc with other funding agencies like NSF

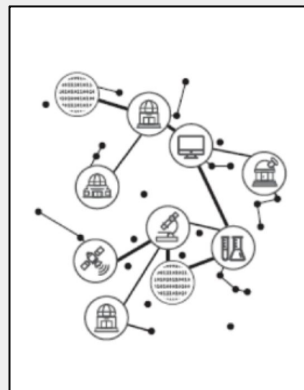
# Interconnectivity and integration of instrumentation, data and computing have been explicitly recognized as strategic requirements for national R&D

The 2021 National Strategic Overview from the Subcommittee on Research and Development Infrastructure formally redefined “federal R&D Infrastructure” to now include computing, data, and networking facilities, resources and services.

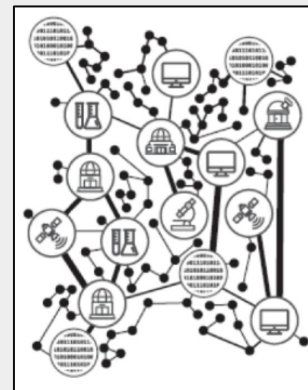
***“R&D continues to shift from smaller to bigger science, driven in large part by advances in computing and other research cyberinfrastructure, which interlink[s] research data, analytics, ... and experimental instrumentation.”***



Past

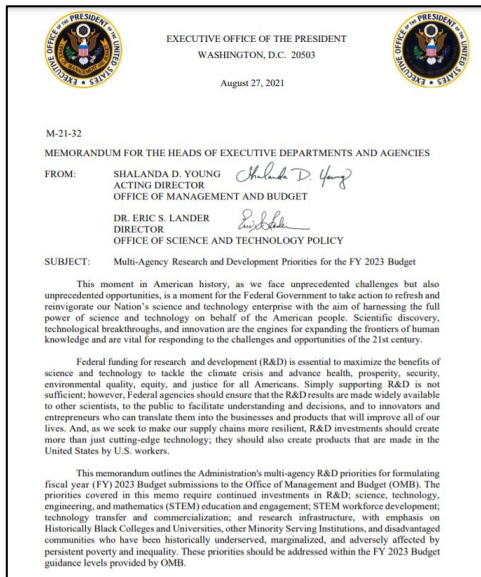


Present



Future

# National imperatives for US leadership in strategic areas all require an interlinked ecosystem of instruments, compute, data



Multi-Agency R&D Priorities for the FY 2023 Budget (August 27, 2021)

**Pandemic readiness and prevention.  
Tackling climate change.**

**Catalyzing research and innovation in critical and emerging technologies.**

*Artificial intelligence (AI)*

*Quantum Information Science (QIS)*

*High-performance computing*

*Advanced communications technologies,  
Microelectronics, Biotechnology, Robotics,  
Space technologies.*

**Innovation for equity.**

**National security and economic resilience.**

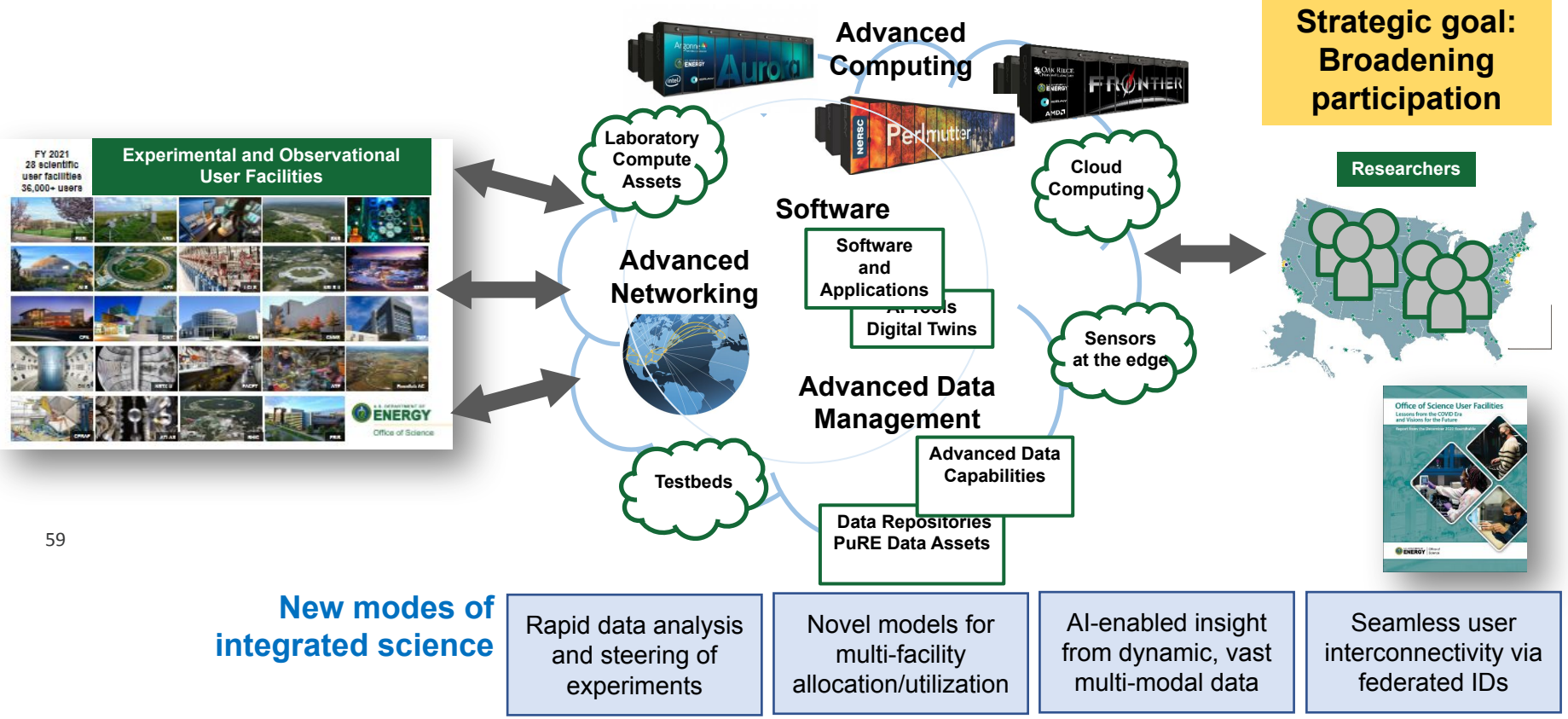


Emerging interagency concepts and initiatives towards a **National Research Ecosystem**

**DOE is uniquely placed to lead and shape the national conversation and initiatives.**



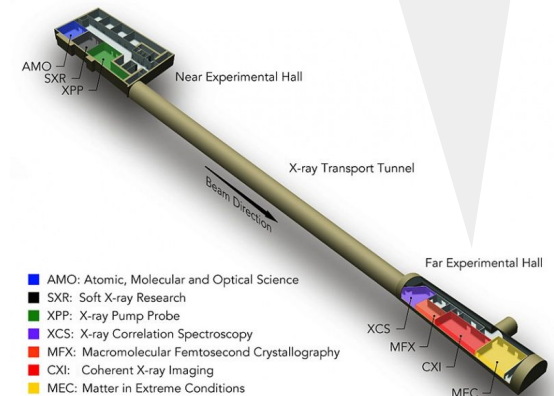
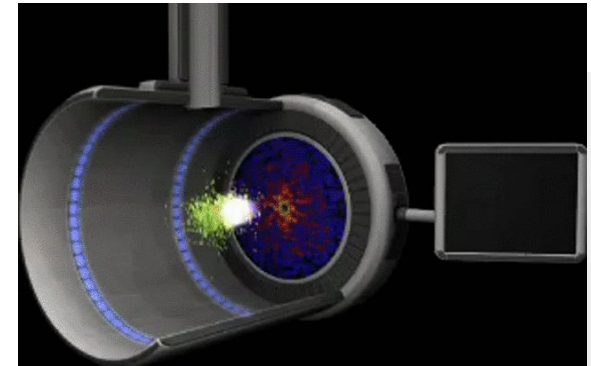
# The vision: A DOE/SC integrated research ecosystem that transforms science via seamless interoperability



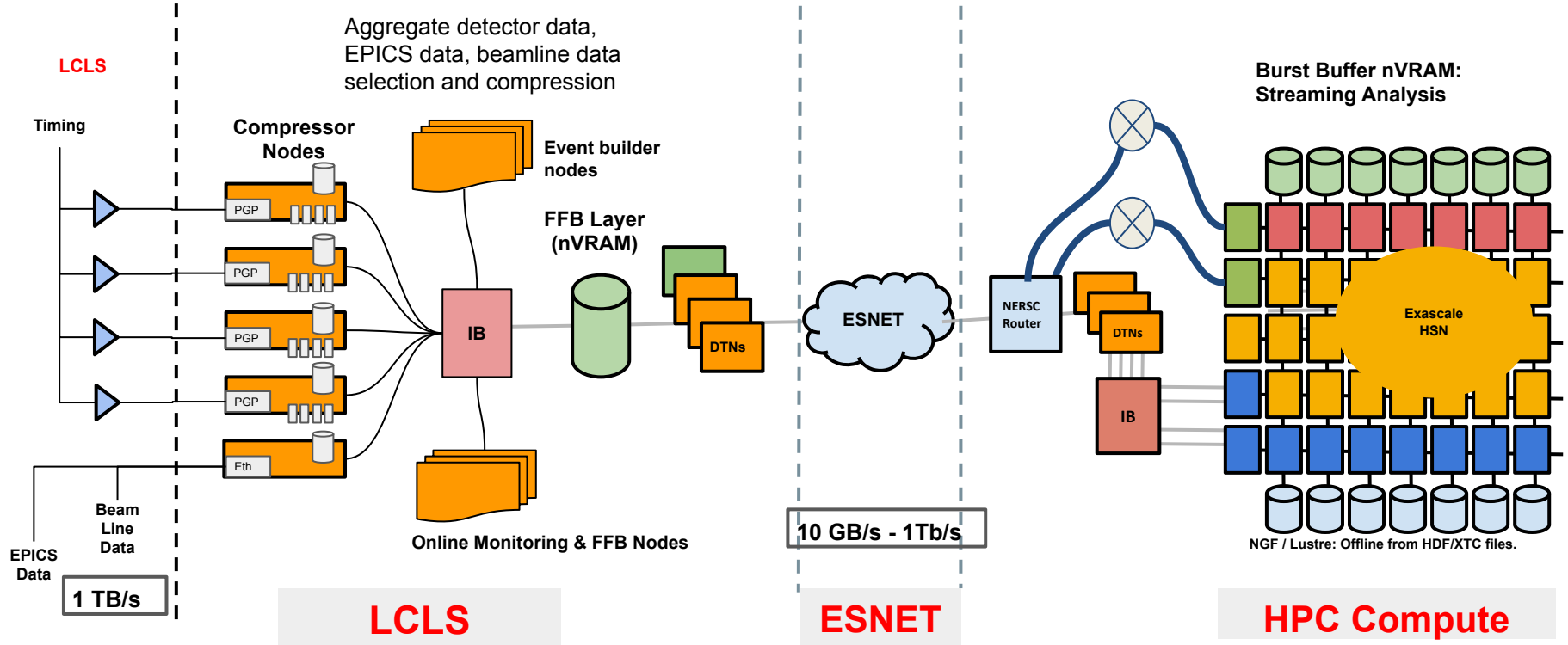
# IRI Use Case 1 - Fast feedback to adjust experiment parameters

## Linac Coherent Light Source (LCLS)

- Ultrafast X-ray pulses from LCLS are used like flashes from a high-speed strobe light, producing stop-action movies of atoms and molecules.
- Both data processing and scientific interpretation demand intensive computational analysis.
- Leverage HPC resources to process initial results to verify proper alignment. Misalignment results in wasted experiment.

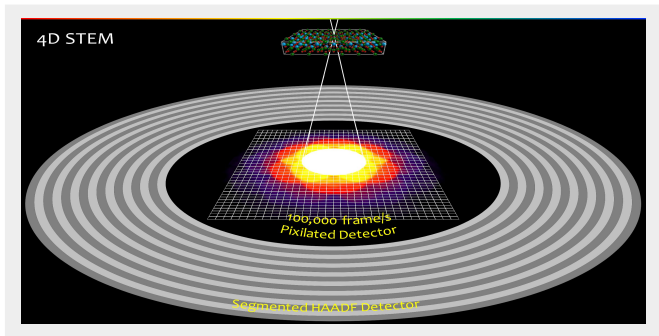


# LCLS ExaFEL Data Transfer Workflow





# IRI Use Case 2 - Reduction or elimination of site local compute and storage

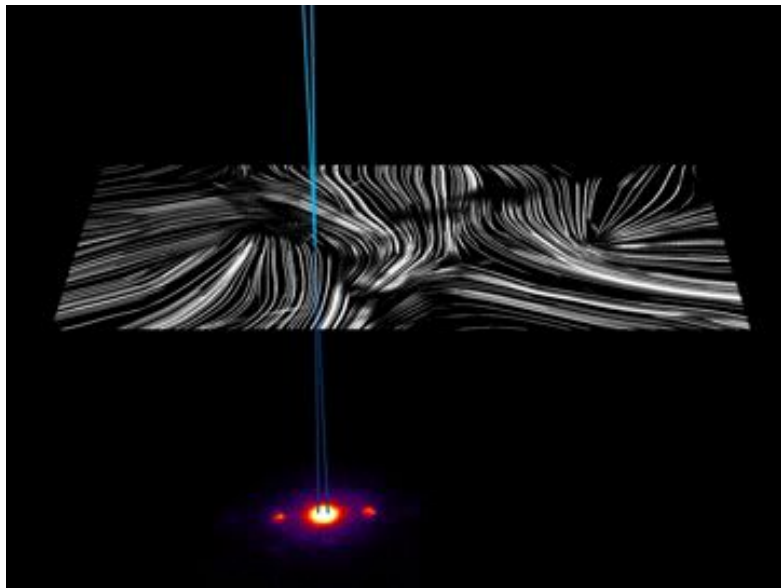


## National Center for Electron Microscopy (NCEM) - 4D STEM

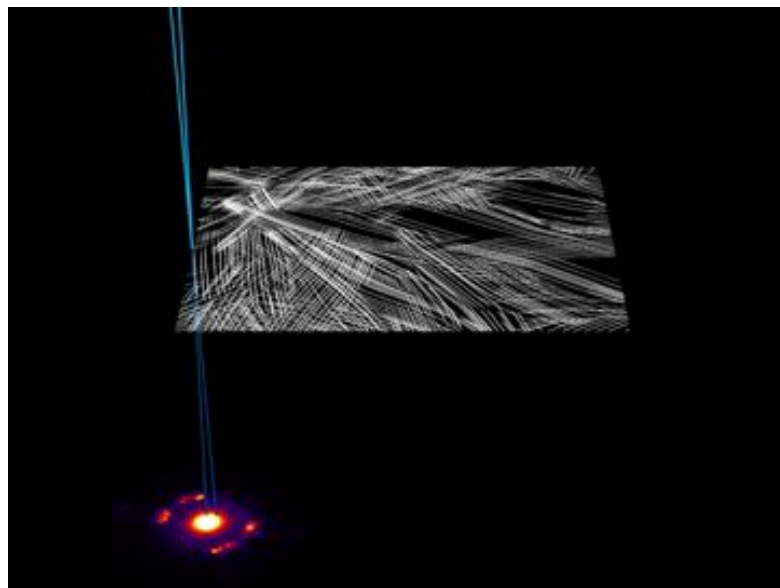
- NCEM is developing a high frame rate (100KHz) 4D detector system to enable fast real-time data analysis of scanning diffraction experiments in scanning transmission electron microscopy (STEM)
- High frame rate development aims to improve scanning diffraction experiments and will be installed on the Transmission Electron Aberration-corrected Microscope (TEAM)
- Direct high speed data transfer of raw image sets from microscope to HPC for online analysis and storage of data.



# High-speed detectors can capture atoms in action at up to 1,600 frames per second



4D-STEM scan of small-molecule organic semiconductor before DIO is added. The diffraction patterns show the orientation of the molecular arrangements in the film. (Credit: Colin Ophus/Berkeley Lab)

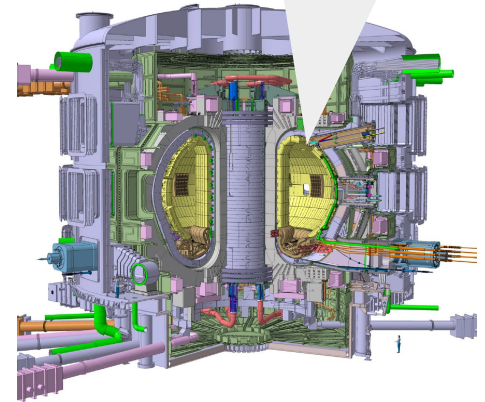
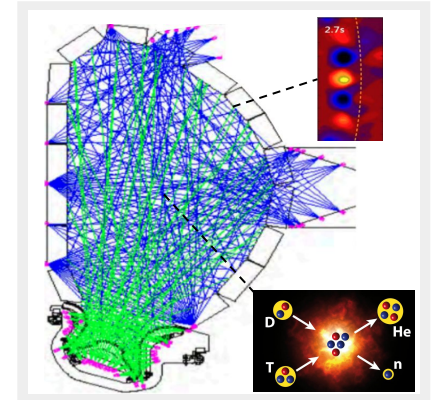


4D-STEM scan of small-molecule organic semiconductor after DIO is added. (Credit: Colin Ophus/Berkeley Lab)

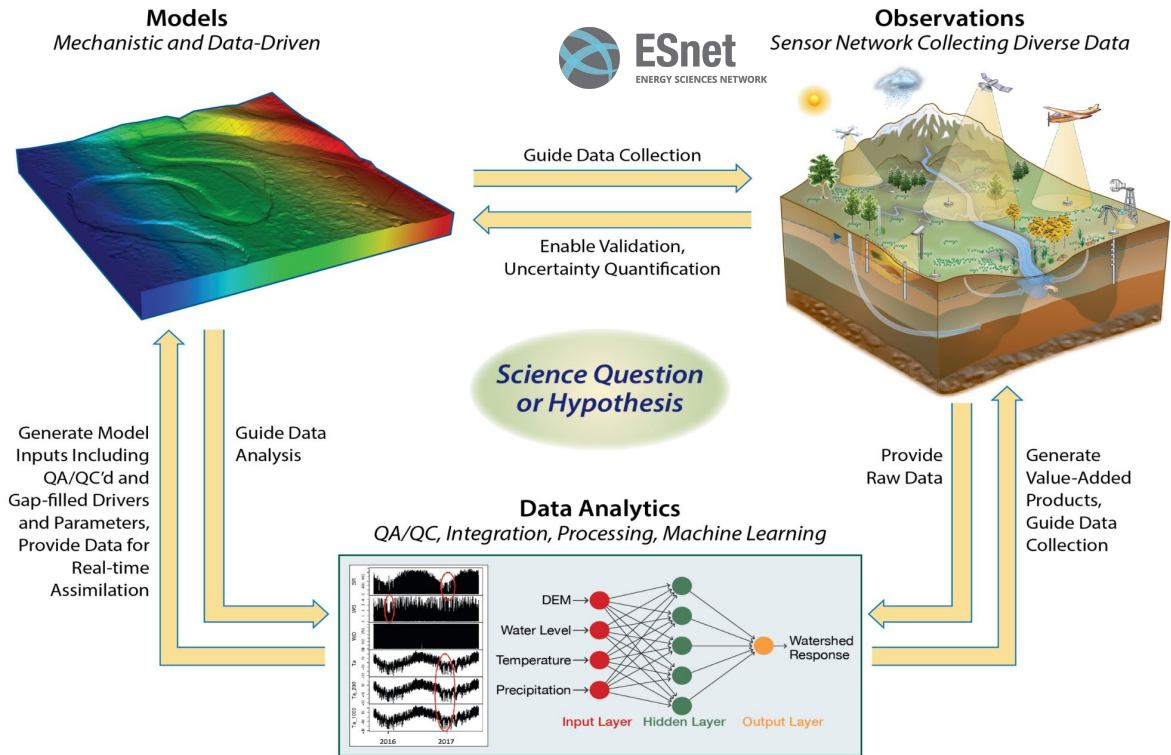
# IRI Use Case 3 - Real-time analysis for monitoring and control

## ITER (*originally* International Thermonuclear Experimental Reactor)

- First fusion device to produce net energy and maintain fusion for long periods of time with ten times the plasma volume of largest machine operating today.
- ITER is designed to produce **500 MW** of fusion power from 50 MW of input heating power.
- Real-time analysis and control is needed to flag potentially dangerous issues within the reactor and mitigate accordingly.



# Emerging Use Case: Distributed Sensor Networks, e.g., Watershed SFA, self-guided field observatories



Wireless technologies are enabling new science use cases.

Real-time HPC could enable:

- Rapid updates to AI models based on streaming data and real-time simulations on HPC
- More accurate automated decision making
- Optimised sensor placement and calibration

# Classes of key data patterns for data-intensive science informs Network-Application integration

## 1. Time-Sensitive patterns

- Requiring temporal end-to-end urgency. For instance, experiment steering, near real-time event detection, deadline scheduling to avoid falling behind.
- Examples including near-real-time analysis of to determine/predict plasma conditions in a fusion reactor, experimental steering using data analysis from a beamline

## 2. Data-Integration-Intensive patterns

- Requiring combining and analyzing data from multiple sources. For instance, data from multiple sites, experiments and/or simulations.
- Examples include combining multi-modal data from high-throughput sequencing, large scale sequence analysis, molecular scale imaging workflows

## 3. Long-term Campaign Patterns

- Requiring sustained access to resources over a long time to accomplish a well-defined objective. For instance, sustained simulation production, large data (re)processing for collaborative use.
- Examples include GRETA spectrometer data reconstruction in 1 - 5 day campaigns, night sky astronomical observations requiring a pipeline to identify interesting targets

# IRI Task Force has put together a white paper with abstract operational models and guiding principles.

ASCR Integrated Research Infrastructure Task Force

March 8, 2021

## Toward a Seamless Integration of Computing, Experimental, and Observational Science Facilities: A Blueprint to Accelerate Discovery

### About the ASCR Integrated Research Infrastructure Task Force

There is growing, broad recognition that integration of computational, data management, and experimental research infrastructure holds enormous potential to facilitate research and accelerate discovery.<sup>1</sup> The complexity of data-intensive scientific research—whether modeling/simulation or experimental/observational—poses scientific opportunities and resource challenges to the research community writ large.

Within the Department of Energy's Office of Science (SC), the Office of Advanced Scientific Computing Research (ASCR) will play a major role in defining the SC vision and strategy for integrated computational and data research infrastructure. The ASCR Facilities provide essential high end computing, high performance networking, and data management capabilities to advance the SC mission and broader Departmental and national research objectives. Today the ASCR Facilities are already working with other SC stakeholders to explore novel approaches to complex, data-intensive research workflows, leveraging ASCR-supported research and other investments. In February 2020, ASCR established the Integrated Research Infrastructure Task Force<sup>2</sup> as a forum for discussion and exploration, with specific focus on the operational opportunities, risks, and challenges that integration poses. In light of the global COVID-19 pandemic, the Task Force conducted its work asynchronously from April through December 2020, meeting via televideo for one hour every other week. The Director of the ASCR Facilities Division facilitated the Task Force, in coordination with the ASCR Facility Directors.

The work of the Task Force began with these questions: Can the group arrive at a shared vision for integrated research infrastructure? If so, what are the core principles that would maximize scientific productivity and optimize infrastructure operations? This paper represents the Task Force's initial answers to these questions and their thoughts on a strategy for world-leading integration capabilities that accelerate discovery across a wide range of science use cases.

***“Our vision is to integrate across scientific facilities to accelerate scientific discovery through productive data management and analysis, via the delivery of pervasive, composable, and easily usable computational and data services.”***

*B. Brown, C. Adams, K. Antypas, D. Bard, S. Canon,  
E. Dart, C. Guok, E. Kissel, E. Lancon, B. Messer, S.  
Oral, J. Ramprakash, A. Shankar, T. Uram*



# International efforts in Integrated Research Infrastructure are expanding too

中國科技器  
China Science & Technology Cloud

HOME | ABOUT | RESOURCES & SERVICES | USE CASES | GOSC | MEDIA | 中文CN

**CSTCLOUD IS A NATIONAL INFRASTRUCTURE FOR SCIENCE DISCOVERY**

EUROPEAN OPEN SCIENCE CLOUD

About Services & Resources Policy Use Cases Media For providers Using the Portal

**EOSC Portal - A gateway to information and resources in EOSC**

Home » About the EOSC Portal

### About the EOSC Portal

The EOSC Portal is part of the EOSC implementation roadmap as one of the expected "federating core" services contributing to the implementation of the "Access and Interface" action line. It has been conceived to provide a European delivery channel connecting the demand-side and the supply-side of the EOSC and all its stakeholders.

#### Content and structure

The EOSC Portal is a gateway to information and resources in EOSC, providing updates on its governance and players, the projects contributing to its realisation, funding opportunities for EOSC stakeholders, relevant European and national policies, important documents, and recent developments. The EOSC Portal Catalogue & Marketplace acts as an entry point to the multitude of services and resources for researchers.

For prospective users of the services, the Portal provides training materials and tutorials on how to use its features. The Portal also offers information for potential service providers on how to onboard their services to the EOSC Portal Catalogue & Marketplace.

The EOSC Portal also engages the EOSC community and stakeholders. The events and news sections cover relevant updates coming from the expanding EOSC ecosystem.

**LATEST NEWS**

Enhance your research with the EOSC Portal Marketplace

Enhance your research with the EOSC Portal Marketplace

iris

Home | What is IRIS? | Meetings | Partner Resources | Security | RSAP | Support

## A cooperative community creating digital research infrastructure to support STFC science

**Cutting edge science needs cutting edge digital infrastructure**

Scientific experiments, facilities and instruments require digital research infrastructure to manage, store, analyse and simulate their data.

IRIS is working with providers to create and develop the digital research infrastructure needed to allow UKRI to continue to play a leading role in global projects such as the Square Kilometre Array and Deep Underground Neutrino Experiment.

This infrastructure includes:

**UKRI** Science and Technology Facilities Council

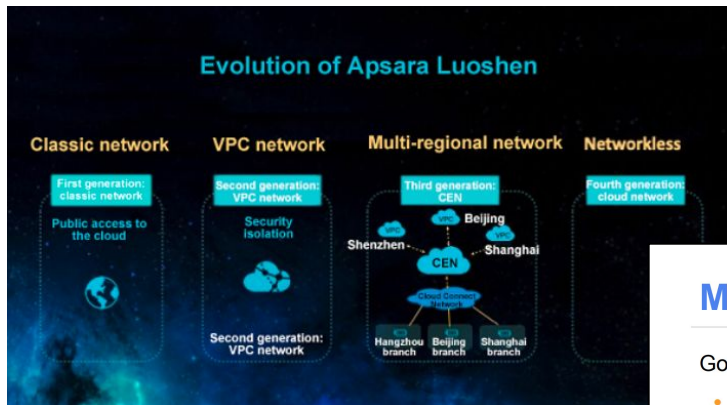
The Scientific Computing Department provides large scale HPC facilities, computing data services and infrastructure at both Daresbury Laboratory and Rutherford Appleton Laboratory. <https://stfc.ukri.org/about-us/where-we-work/daresbury-laboratory/scientific-computing-department/>

- China Science and Technology Cloud
- European Open Science Cloud
- IRIS UKRI STFC initiative



# Hyperscalers get it\*!

## Alibaba Cloud's Apsara Luoshen



Z. Zong, "Apsara Luoshen, a High Performance Network Engine that Drives Alibaba Cloud", GTNC 2018, Nov 15, 2018

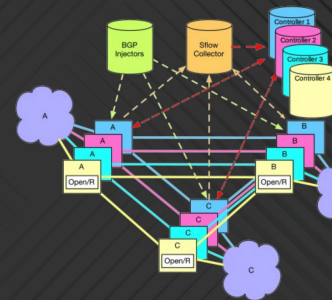
*Deploying a vanilla best-effort delivery network is not optimal!*

\*NB: Solutions deployed are only within a single administrative domain

## Facebook Express Backbone (EBB)

### Network Design

- Commodity switches
- Four parallel forwarding planes
- Open/R
- BGP injection
- Sflow collector
- Traffic-engineering controller



## Google B4 WAN

### More Than the Sum of Parts



Google Networking works together as an integrated whole

- **B4: WAN interconnect**
- GGC: edge presence
- Jupiter: building scale datacenter network
- Freedom: campus-level interconnect
- Andromeda: isolated, high-performance slices of the physical network

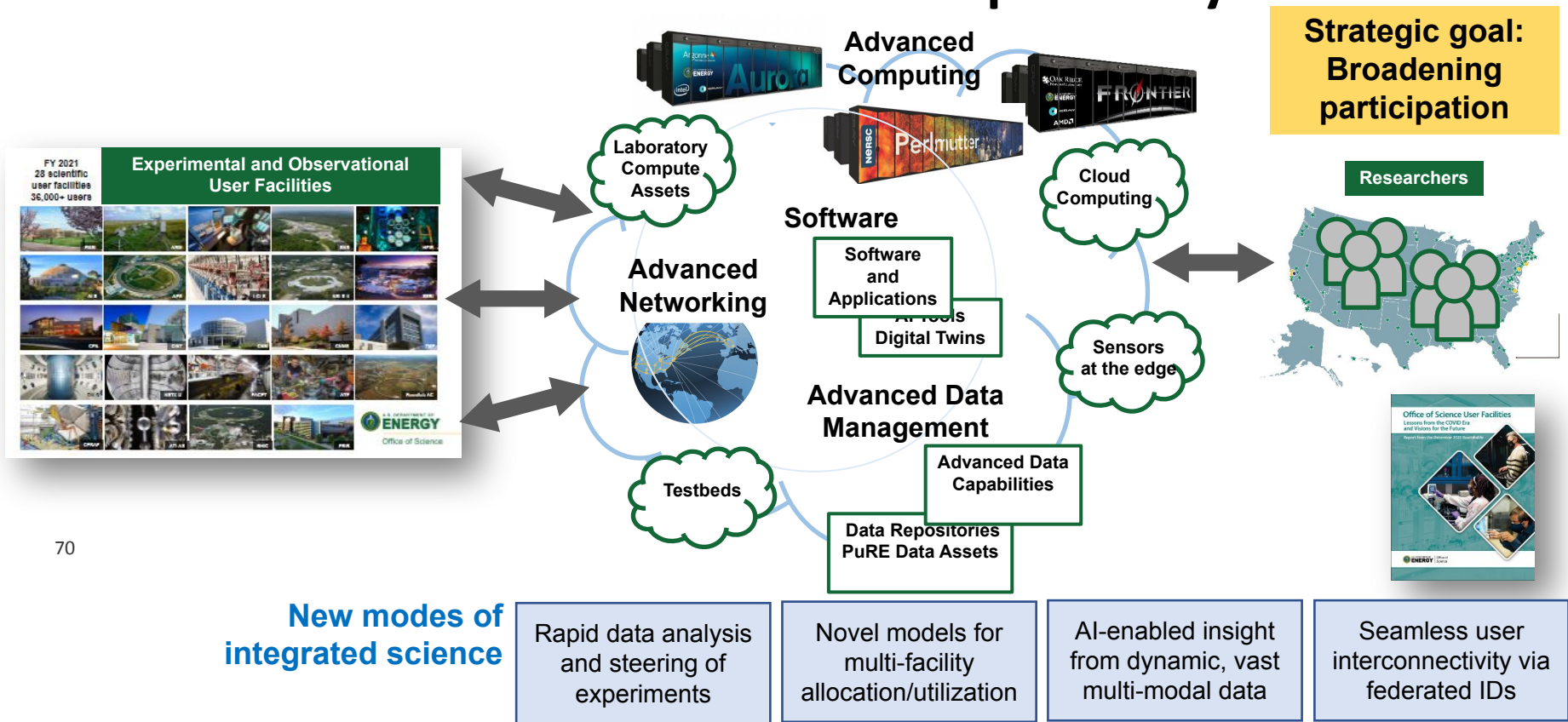
Publications in INFOCOM 2012, SIGCOMM 2013, SIGCOMM 2014, CoNEXT 2014, EuroSys 2014, SIGCOMM 2015



H. Kwok, "Express Backbone: Moving Fast with Facebook's Long-Haul Network", Networking @Scale 2017, July 9, 2015

S. Mandal, "Lessons Learned from B4, Google's SDN WAN", 2015 USENIX ATC'15, July 9, 2015

# The vision: A DOE/SC integrated research ecosystem that transforms science via seamless interoperability



# Talk Flow



Analogy: Case for a richer Network-Application interface

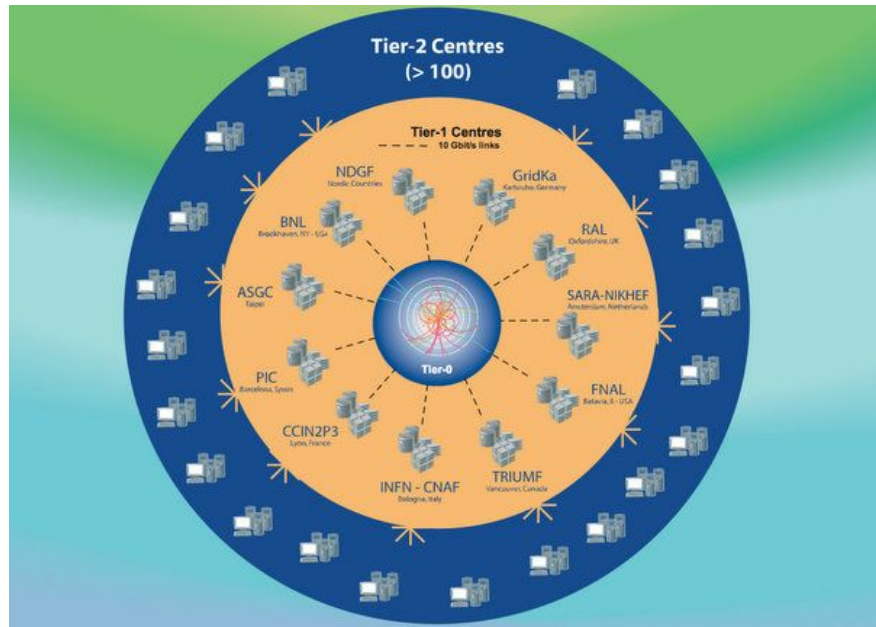


Data-intensive science: motivating the case for integrated research infrastructure

$$f(x)$$

Example of application-network integration

# The Large Hadron Collider and the global computing grid



# CERN's File Transfer Service (FTS) and Rucio

## 2021 IN NUMBERS



>1.0

EB transferred



1.15

billion files transferred



24

FTS instances



37

Virtual Organizations

---

## Rucio

The next generation  
of Distributed Data  
Management System



---

Vincent Garonne

- Discover data
- Transfer data to/from sites
- Delete data from sites
- Ensure data consistency at sites
- Enforce computing model

# CMS Workflow Considerations

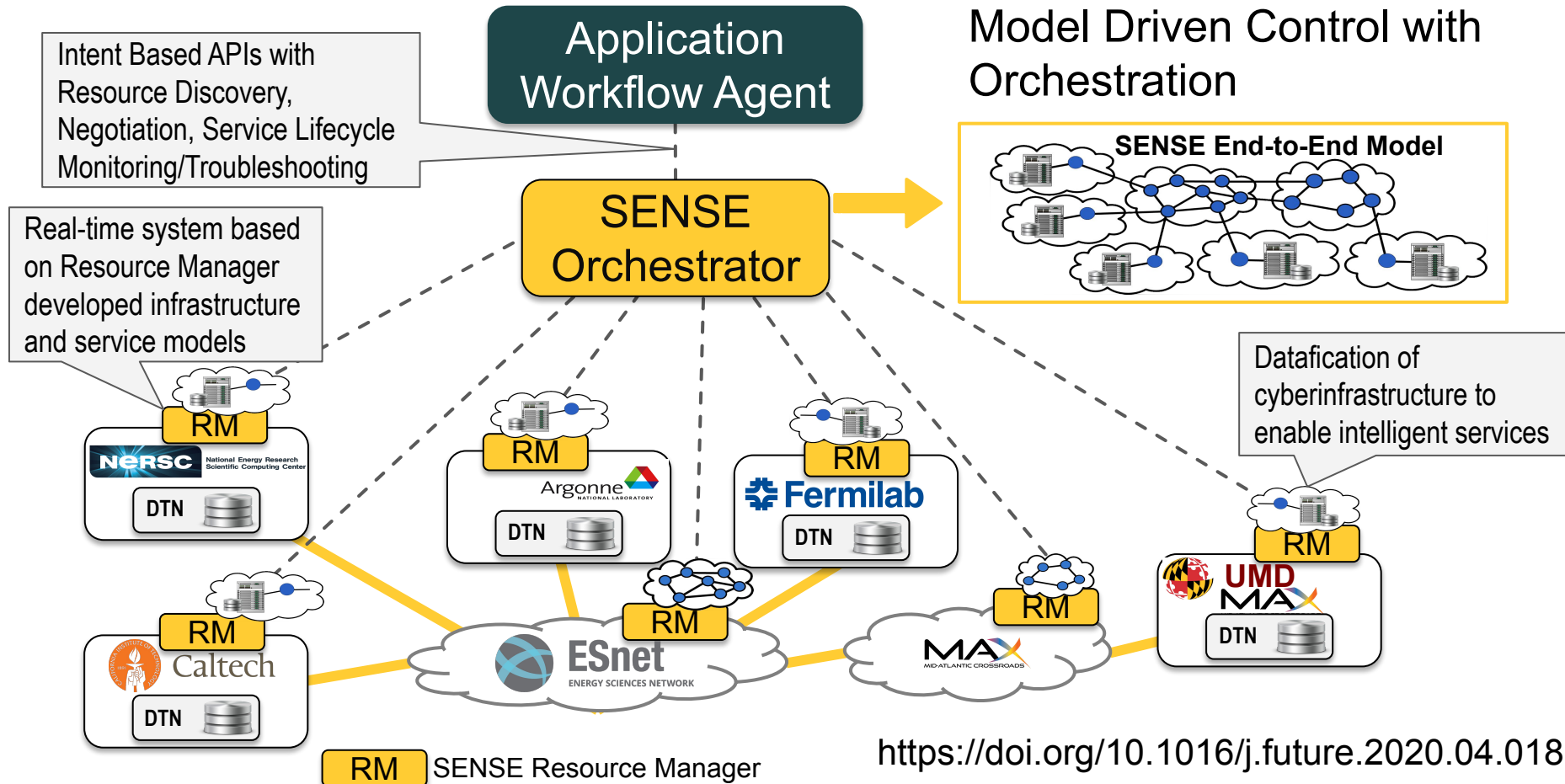
## CMS Annual Data Volume

	# of collisions	# of events simulated	RAW event size [MB]	AOD event size [MB]	Total per year [PB]
Today	9 Billion	22 Billion	0.9	0.35	~20
HL-LHC (2029)	56 Billion	64 Billion	6.5	2	~600

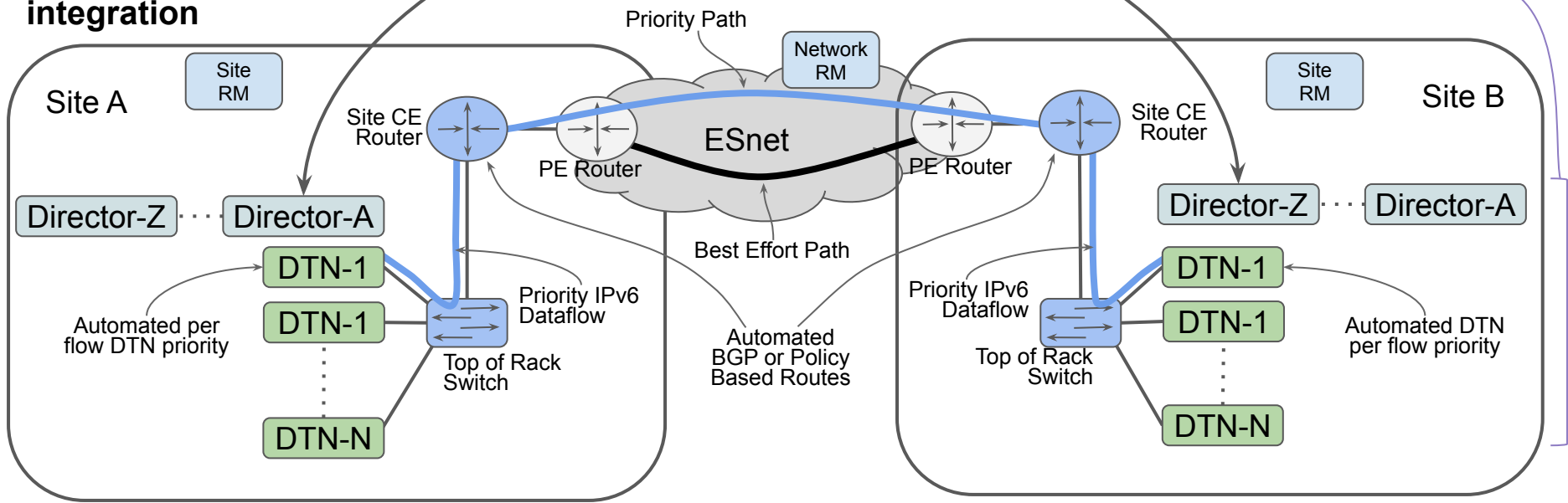
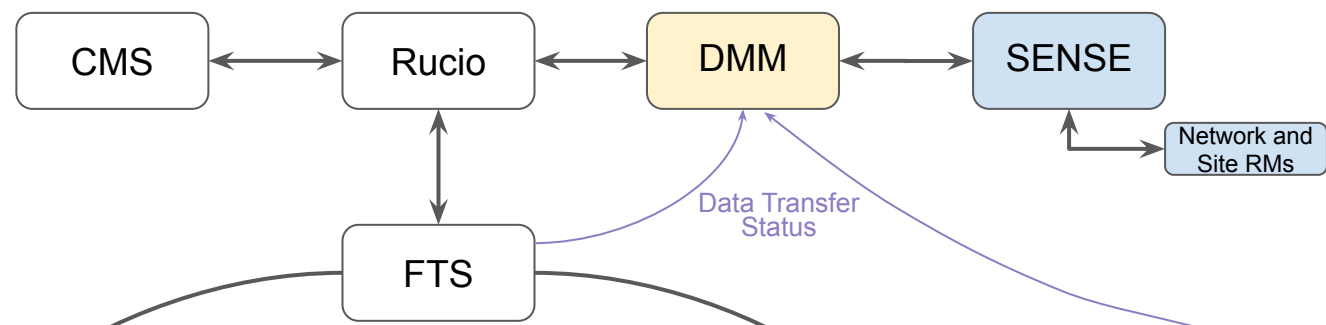
- Transfers that matter most are => 1PB@100Gbps ~ 1 day
- There is nothing “realtime” in this system.
- Relevant timescales for changes are O(10)min to O(10)h to O(10) days
  - Knowledge of what needs transferred exists well ahead of completion of transfer.
  - Expected that the available network bandwidth changes a few times during a flow of data that takes days or even weeks to complete.
- **Rucio should know when things at the network change such that the human operator of Rucio can learn about it from Rucio and/or Rucio can “plan around” slowdowns.**



# SENSE Architecture



**Integration of CMS experiment's Distributed Data Management and File Transfer Service with data movement APIs is a pathfinder to explore network-application integration**



# Network Integration allows new kinds of interaction

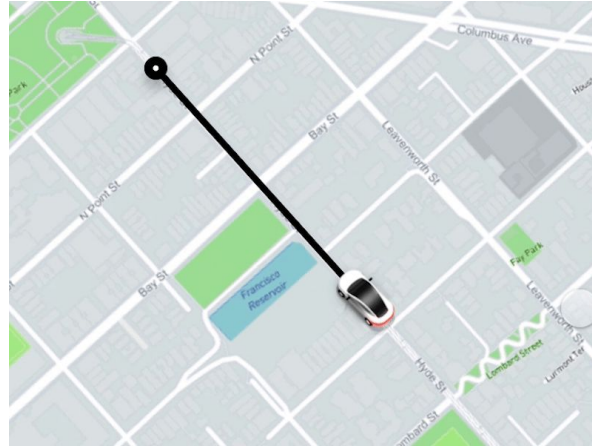
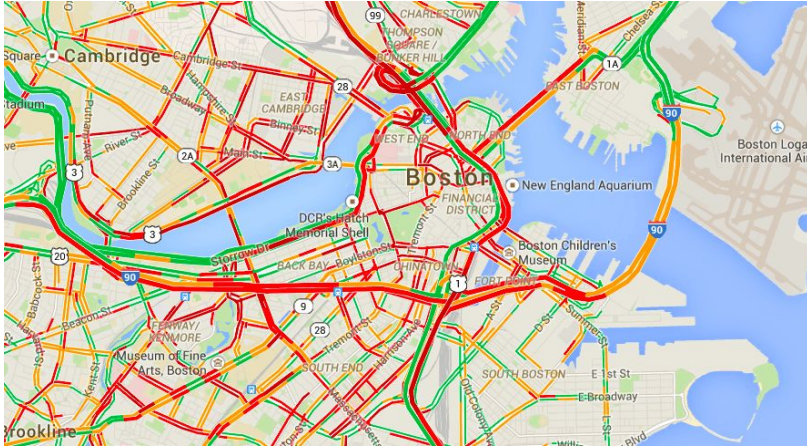
- "Application Workflow Agents" are encouraged to plan data movement operations to the maximum extent possible
  - Schedule in advance when possible
  - Decide which flows need priority service
  - Interact with the network to decide optimal time for data transfer from both the workflow and the network perspectives
  - Realtime adjustment of priority service assignment based on monitoring and workflow objective changes
- Allow the "Application Workflow Agents" to interact with the network and ask:"What is Possible?", "What is recommended?"
- These "interactions" can be tailored to optimize for specific workflow operations.

# Takeaways

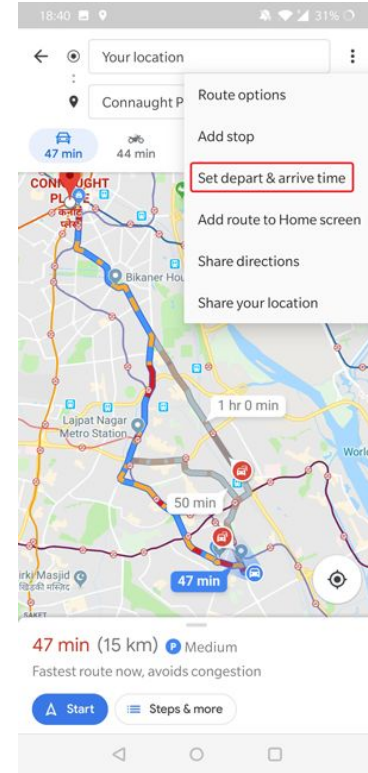
- DOE's IRI effort is motivating conversations between the science applications and the research infrastructure
- New Science use-cases and workflows are motivating discussions of 'API' interaction between applications, and research infrastructure, including networks
- Many such experiments will showcase the value of the integration, and highlight new challenges that need to be addressed

# Epilogue

# Early tools and components to build a network platform that enables deeper application integration



Real-time traffic and traffic prediction helps plan with just in time information, and features such as dynamic rerouting and updated accurate data on when the destination will be reached

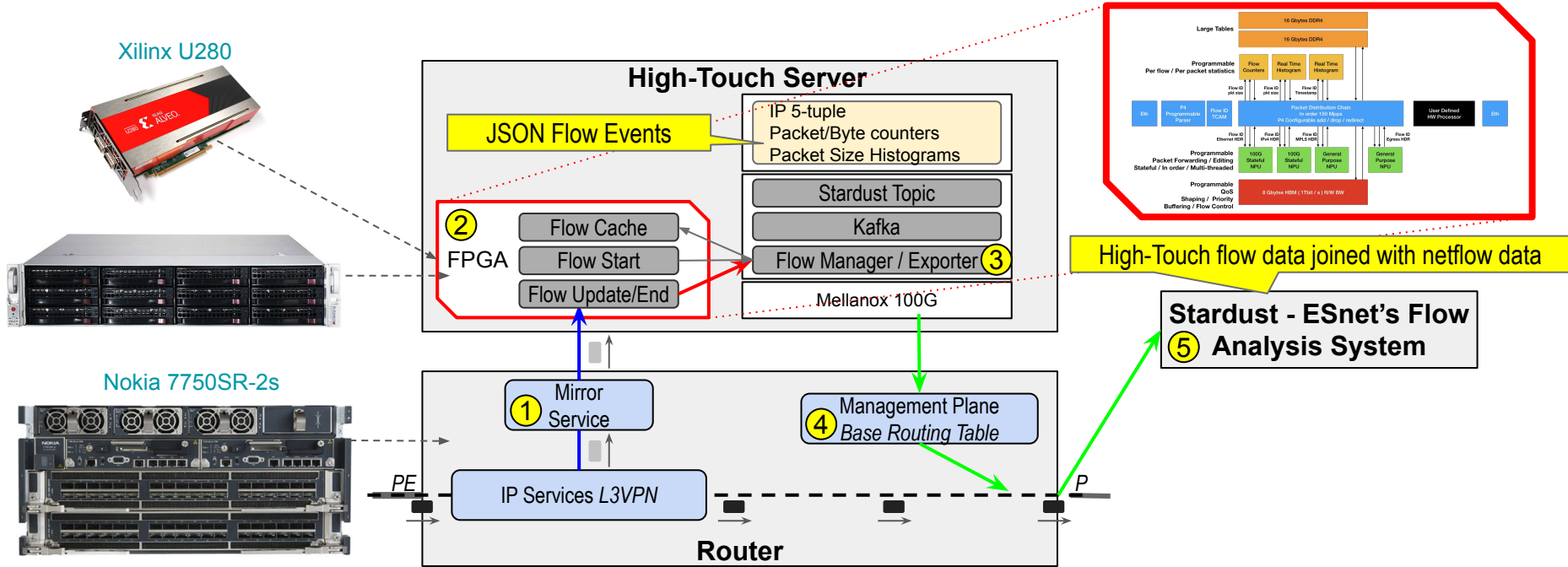




# Precision Telemetry Services

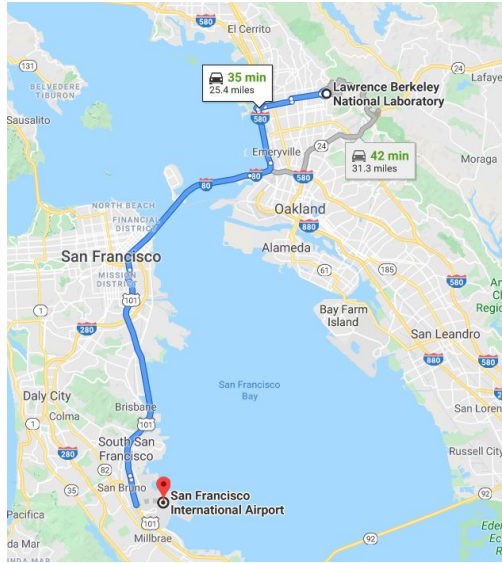
- Provide information on traffic passing over ESnet, for example:
  - Detailed flow summaries
  - Per-packet events on selected flows
- **What's notable about this...**
  - Process every packet of interest in real-time @ > 100Gbps (no sampling)
  - Accurate, precision timing (ns precision / accuracy)
  - Programmatically deployable and customizable
  - A flexible platform for applications and experiments
- Technology enablers
  - Programmable network dataplane hardware with accurate timestamps
  - High-speed packet processing libraries (DPDK, etc.)
- Part of ESnet6 production network in 40+ different locations

# Platform for Unprecedented Packet Visibility

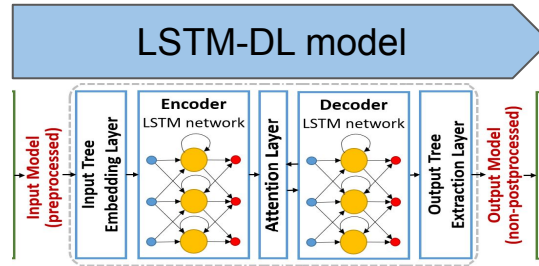
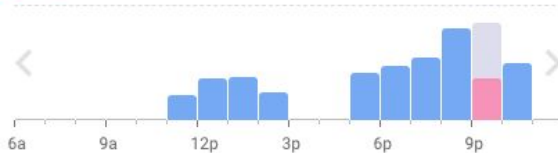


1. Mirror Service - Allows selective flows in the dataplane to be duplicated, truncated and sent to the FPGA for processing.
2. Programmable Dataplane (DP) - Every packet updates internal counters/flow state. Only flow start/end packets sent to SW.
3. Flow Exporter - Processes flow start/end events to update the Dataplane flow cache. Periodically collects flow state and publishes summary records into Kafka.
4. Management Plane Base Routing Table - Provides connectivity to Remote Servers.
5. Stardust Logstash - Subscribes to Hightouch Kafka Topic for Stardust and consumes flow event records, inserting the records into Elastic

# Predicting network congestion



**LIVE** Less busy than usual



*Planning for high-volume near-real-time streaming*



- Real-time Data
  - PerfSonar (Loss, Throughput)
  - Traffic: SNMP data
  - Flow behavior

# And 'potentially' using it for elephant flow routing

- Deployed on Google Cloud Platform
  - Different models can run at the same time to compute least congested paths
  - Estimates transfer completion time
- Trust dashboard
  - Real-time ML performance
  - Build engineer's confidence in predictions
- Still under active research/development

