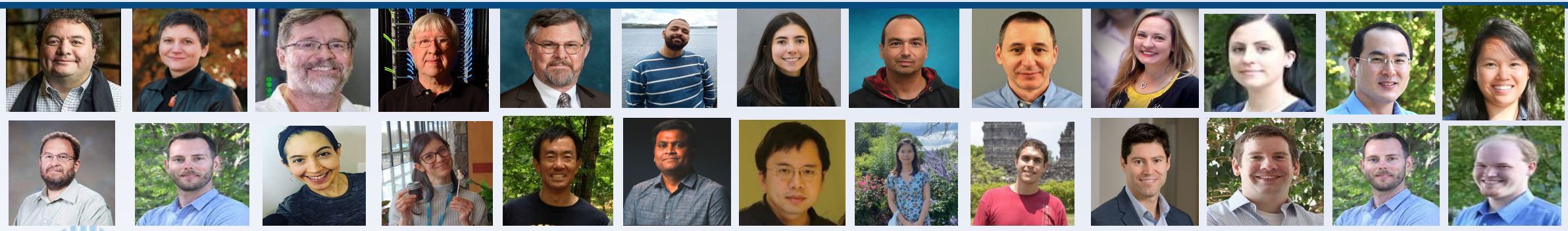
NATIONAL SCIENCE DATA FABRIC

A Platform Agnostic Testbed for Democratizing Data Delivery

Michela Taufer, University of Tennessee Knoxville

Support from NSF (awards: 2138811, 2103845, 2334945), SNL, and LLNL



NSDF National Science Data Fabric

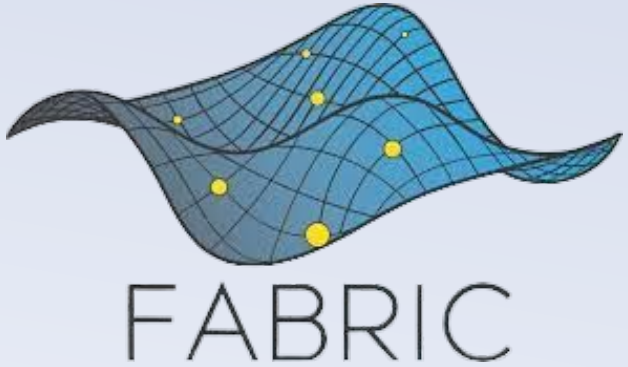
National Science Data Democratization Consortium: Engaging Industry Partners

IBM Cloud

Partnering with Existing NSF Initiatives

Chameleon

open storage network



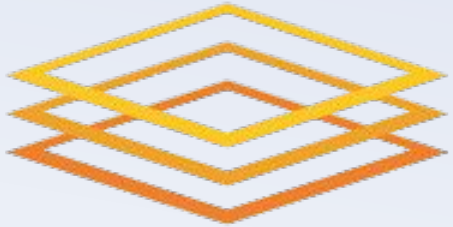
FABRIC

CYVERSE®

CICompass

ACCESS

CloudLab



CloudBank

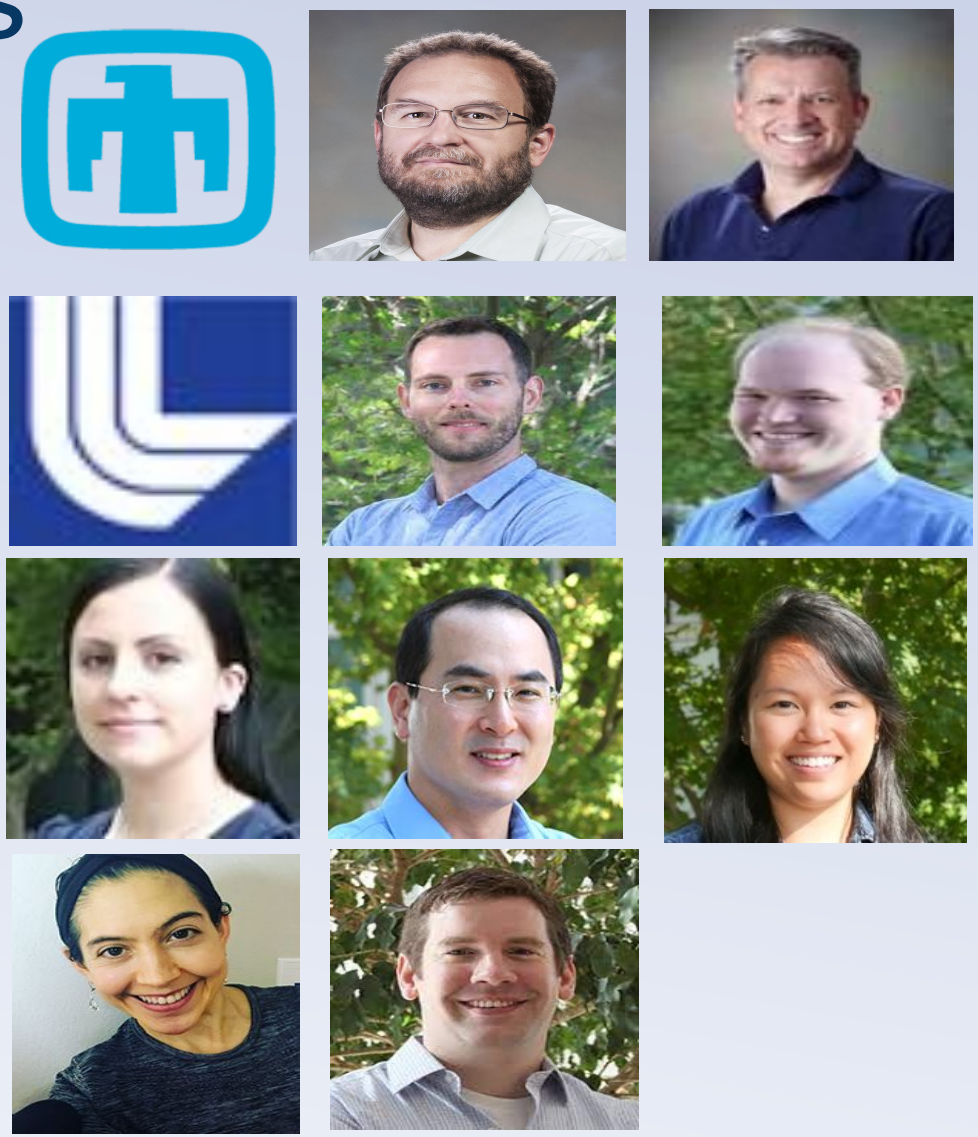
Open Science Grid

NSDF National Science Data Fabric

Partnerships with DoE Labs

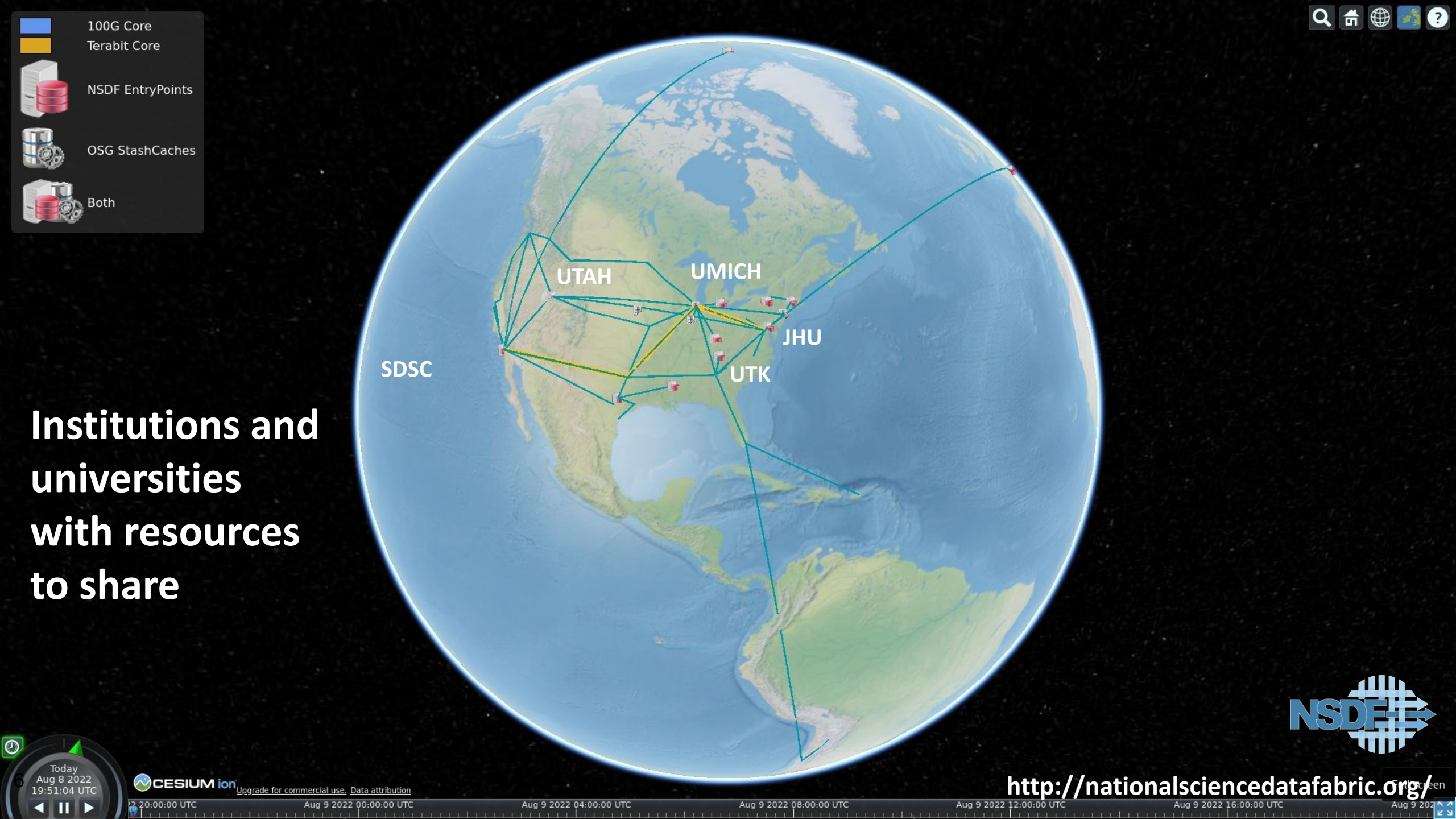
- **Sandia National Laboratories**
  - Workflow containerization (Trustworthy Computing; Data Democratization)

- **Lawrence Livermore National Laboratory**
  - Thicket project (Large Scale Computing and Performance)
  - Flux project (Scheduling and Resource Management)
  - Fractale (Convergence of HPC, Cloud, and Edge)



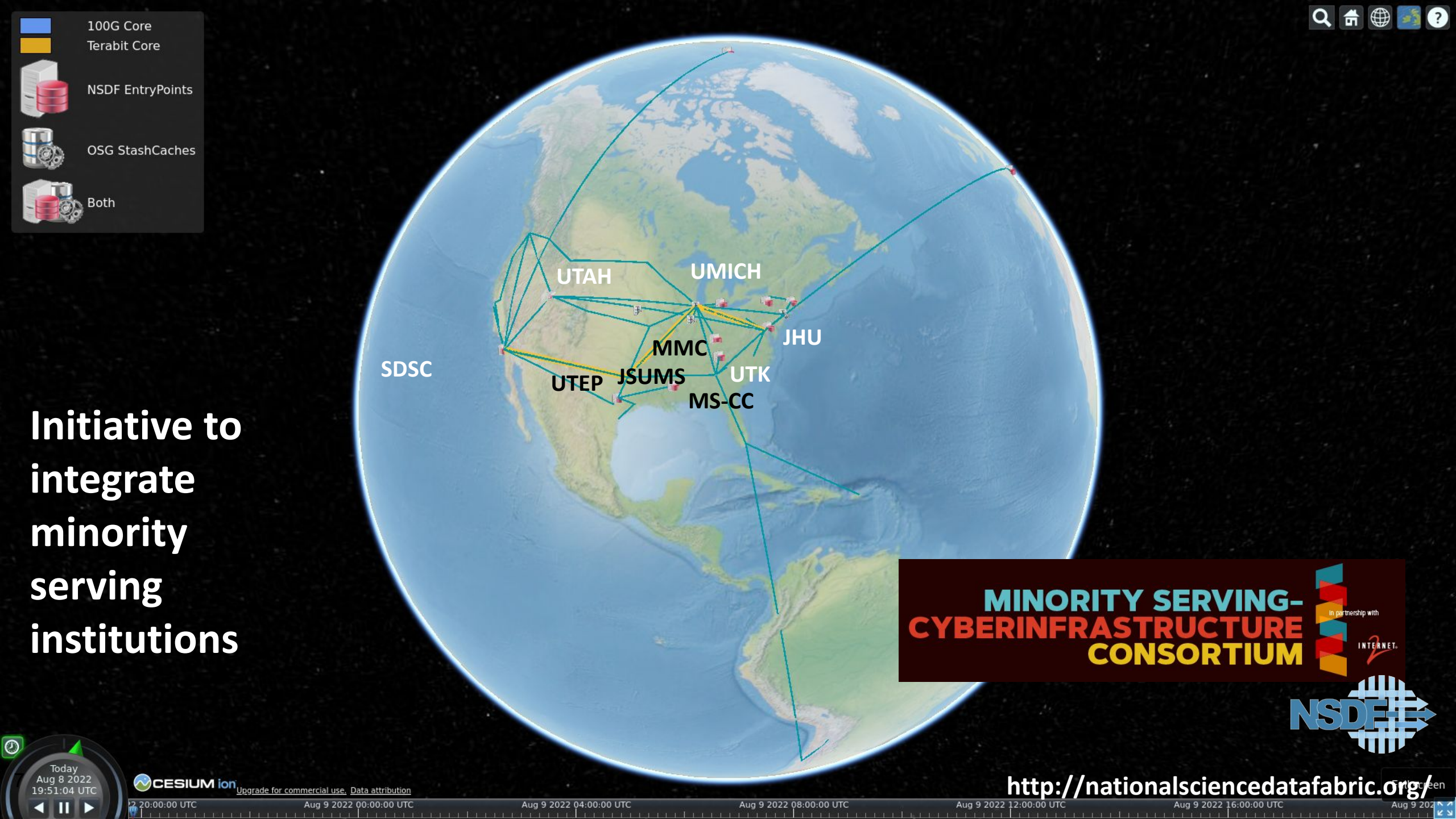
Mission of National Science Data Fabric (NSDF):
We are building a **holistic ecosystem** to **democratize data-driven scientific discovery** by **connecting an open network of institutions**, including minority serving institutions, with **a shared, modular, containerized data delivery environment**.

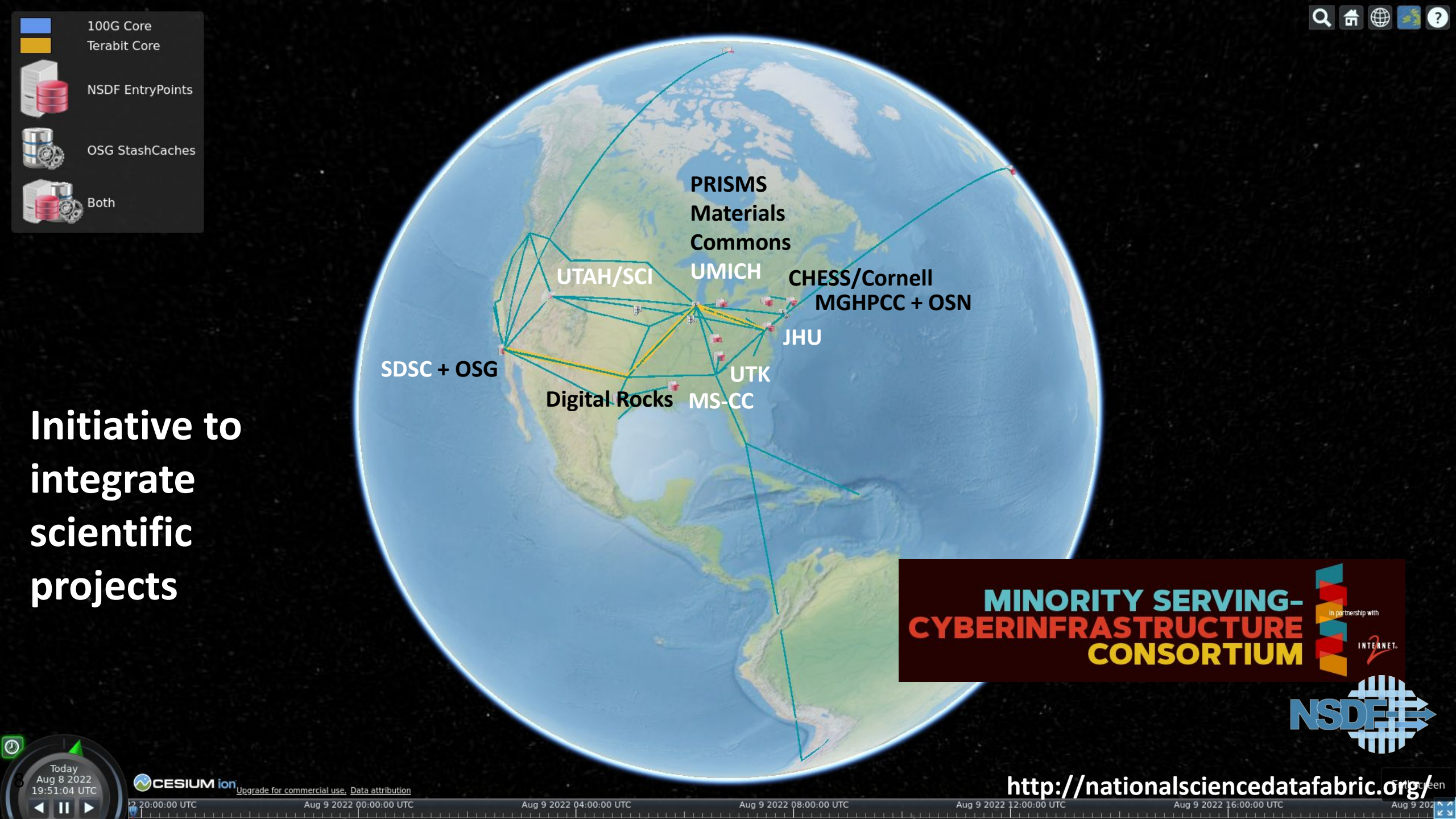
http://nationalsciencedatafabric.org/

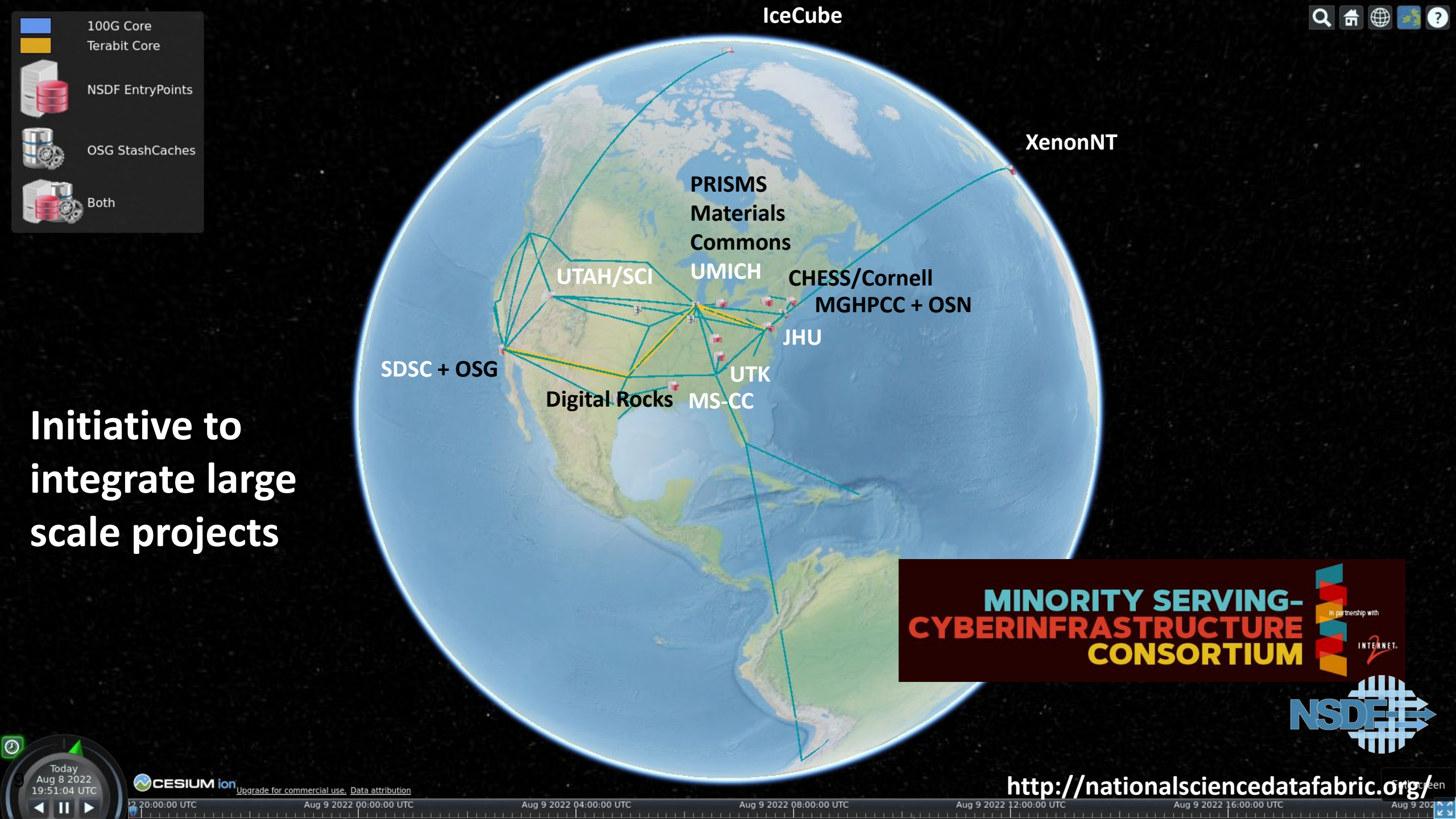


Institutions and universities with resources to share

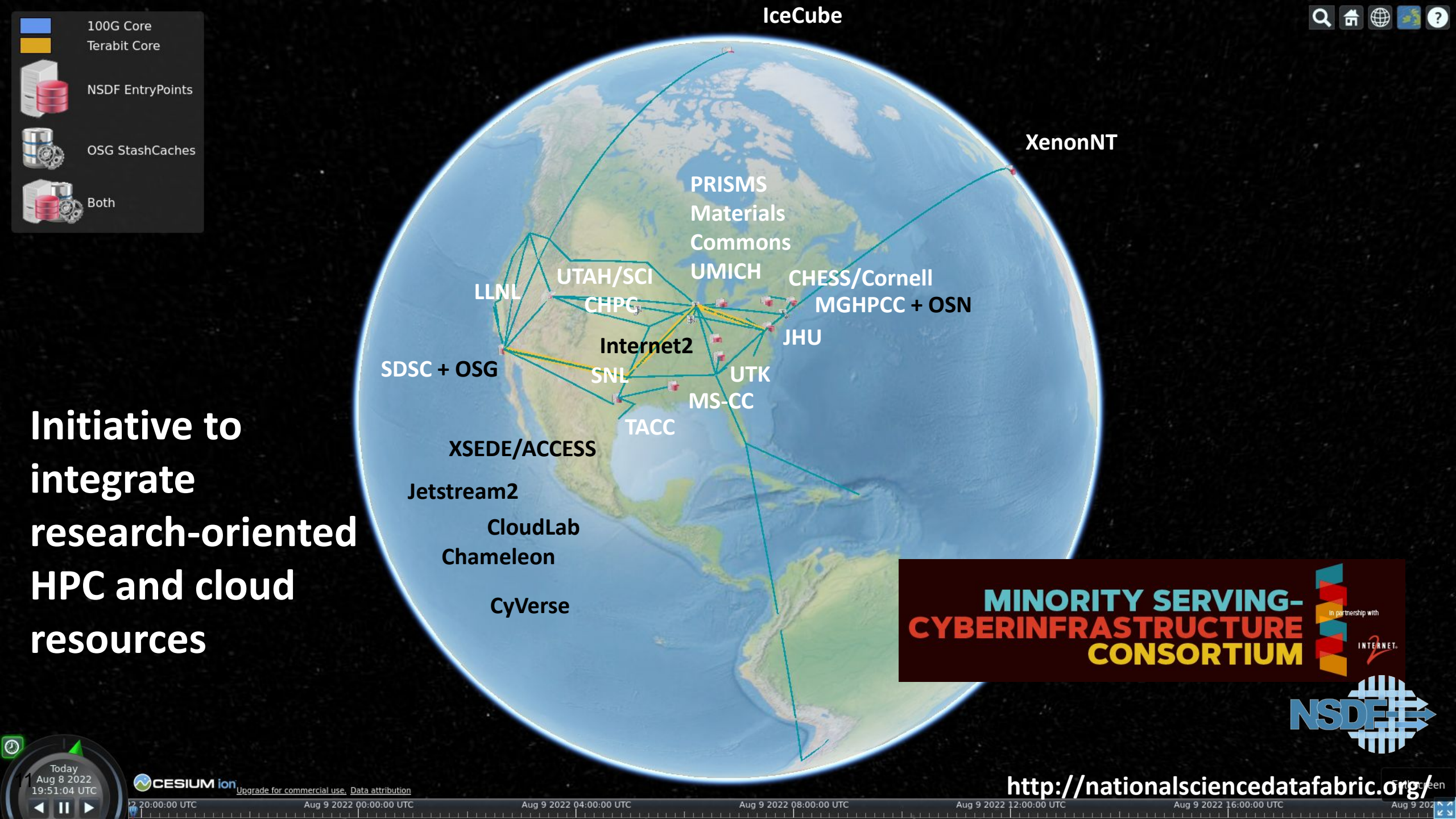
http://nationalsciencedatafabric.org/

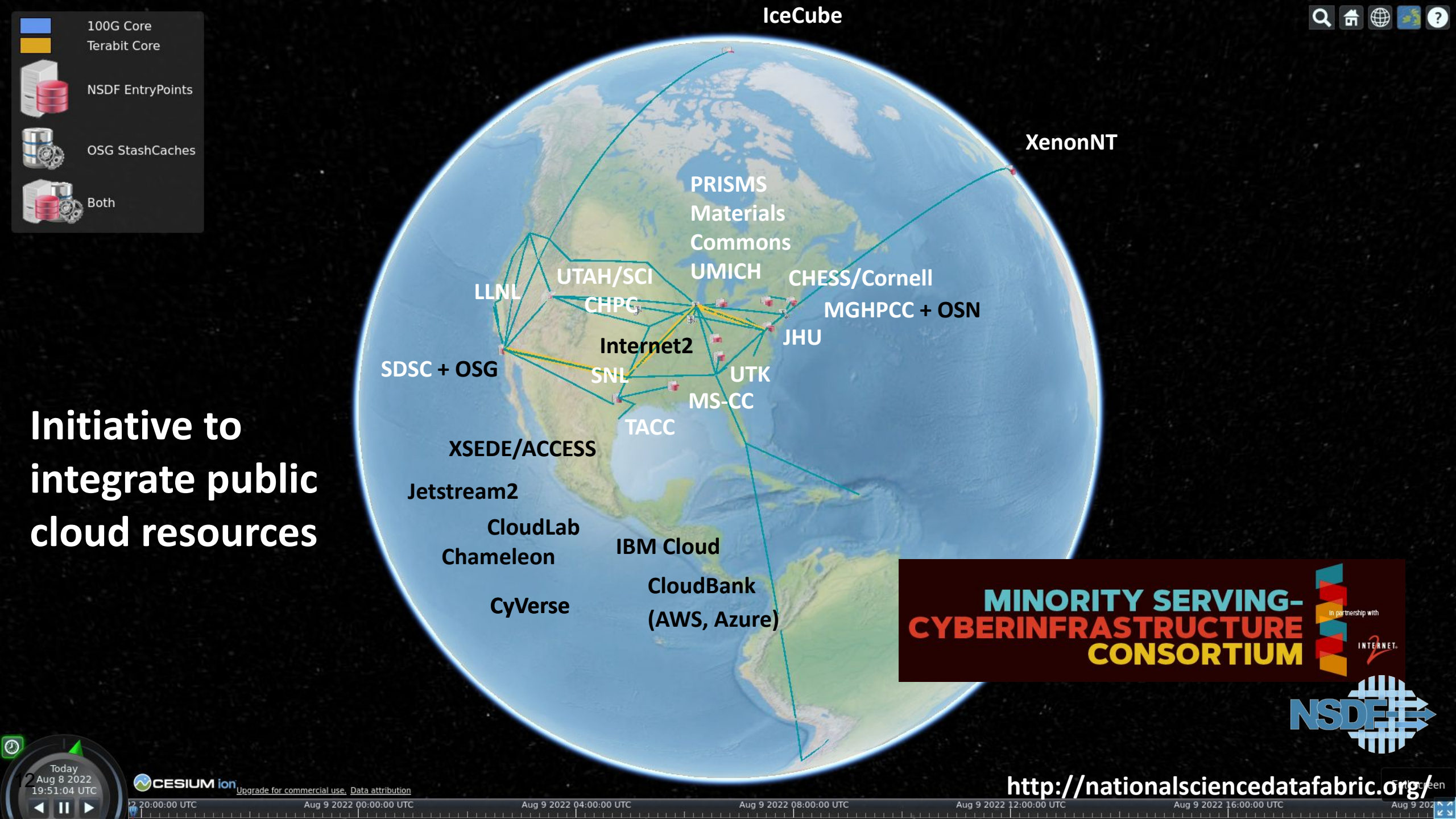

100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
UTAH
UMICH
JHU
MMC
SDSC
JSUMS
UTK
UTEP
MS-CC
Initiative to integrate minority serving institutions
MINORITY SERVING-CYBERINFRASTRUCTURE CONSORTIUM
In partnership with
INTERNET2
NSDF
http://nationalsciencedatafabric.org/
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use. Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC

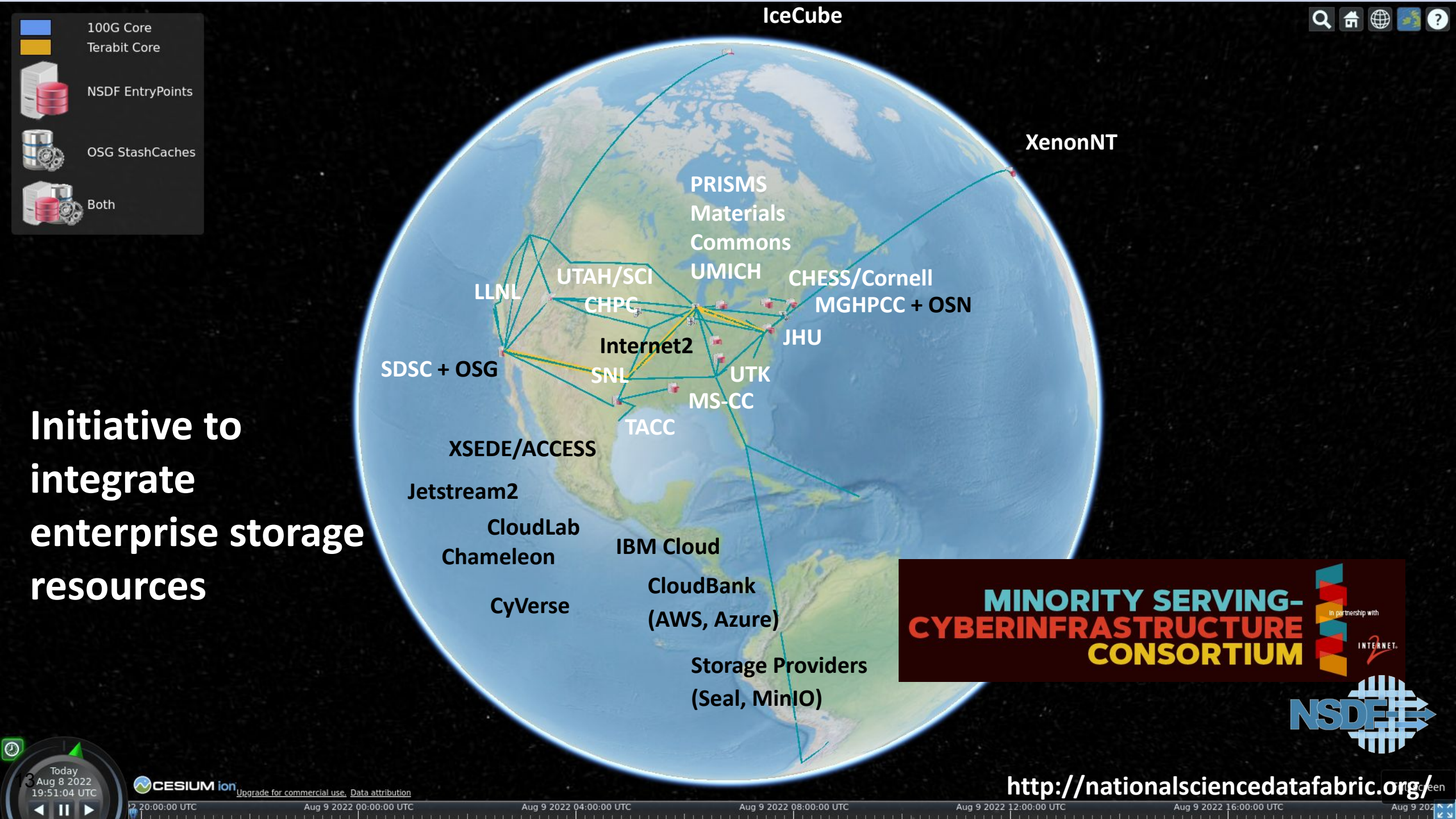

100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
PRISMS
Materials
Commons
UMICH
UTAH/SCI
CHESS/Cornell
MGHPCC + OSN
JHU
SDSC + OSG
UTK
Digital Rocks
MS-CC
Initiative to integrate scientific projects
MINORITY SERVING–
CYBERINFRASTRUCTURE
CONSORTIUM
In partnership with
INTERNET2
NSDF
http://nationalsciencedatafabric.org/
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use. Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
IceCube
XenonNT
PRISMS
Materials
Commons
UMICH
UTAH/SCI
CHESS/Cornell
MGHPCC + OSN
JHU
SDSC + OSG
UTK
Digital Rocks
MS-CC
Initiative to integrate large scale projects
MINORITY SERVING-CYBERINFRASTRUCTURE CONSORTIUM
In partnership with
INTERNET2
NSDF
http://nationalsciencedatafabric.org/
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use. Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
IceCube
XenonNT
PRISMS
Materials
Commons
UMICH
UTAH/SCI
CHESS/Cornell
LLNL
CHPC
MGHPCC + OSN
JHU
SDSC + OSG
SNL
UTK
Digital Rocks
MS-CC
TACC
Initiative to integrate HPC resources
MINORITY SERVING–
CYBERINFRASTRUCTURE
CONSORTIUM
In partnership with
INTERNET2
NSDF
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
http://nationalsciencedatafabric.org/


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
IceCube
XenonNT
PRISMS
Materials
Commons
UMICH
CHESS/Cornell
MGHPCC + OSN
JHU
UTAH/SCI
LLNL
CHPC
Internet2
SDSC + OSG
SNL
UTK
MS-CC
TACC
XSEDE/ACCESS
Jetstream2
CloudLab
Chameleon
CyVerse
Initiative to integrate research-oriented HPC and cloud resources
MINORITY SERVING-CYBERINFRASTRUCTURE CONSORTIUM
In partnership with
INTERNET2
NSDF
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
http://nationalsciencedatafabric.org/


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
IceCube
XenonNT
PRISMS
Materials
Commons
UMICH
CHESS/Cornell
UTAH/SCI
LLNL
CHPC
MGHPCC + OSN
JHU
Internet2
SDSC + OSG
SNL
UTK
MS-CC
TACC
XSEDE/ACCESS
Jetstream2
CloudLab
Chameleon
IBM Cloud
CloudBank
(AWS, Azure)
CyVerse
Initiative to integrate public cloud resources
MINORITY SERVING–CYBERINFRASTRUCTURE CONSORTIUM
In partnership with
INTERNET2
NSDF
http://nationalsciencedatafabric.org/
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
IceCube
XenonNT
PRISMS
Materials
Commons
UMICH
CHESS/Cornell
MGHPCC + OSN
JHU
UTAH/SCI
LLNL
CHPC
Internet2
SDSC + OSG
SNL
UTK
MS-CC
TACC
XSEDE/ACCESS
Jetstream2
CloudLab
Chameleon
IBM Cloud
CloudBank
(AWS, Azure)
CyVerse
Storage Providers
(Seal, MinIO)
Initiative to integrate enterprise storage resources
MINORITY SERVING-CYBERINFRASTRUCTURE CONSORTIUM
In partnership with
INTERNET2
NSDF
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC
http://nationalsciencedatafabric.org/

Our Strategy: Understanding and Addressing User's Pain Points

Democratizing Access and Use of Large-scale Data

Implementing the NSDF Vision: User Interviews

Identify Users

- Diverse roles: Domain scientists, CI professionals, developers
- Diverse domains: materials science, climate, earth sciences, astronomy, and more!
- Diverse institutions: R1 universities, teaching colleges, MSIs, national labs, experimental facilities

Target Questions

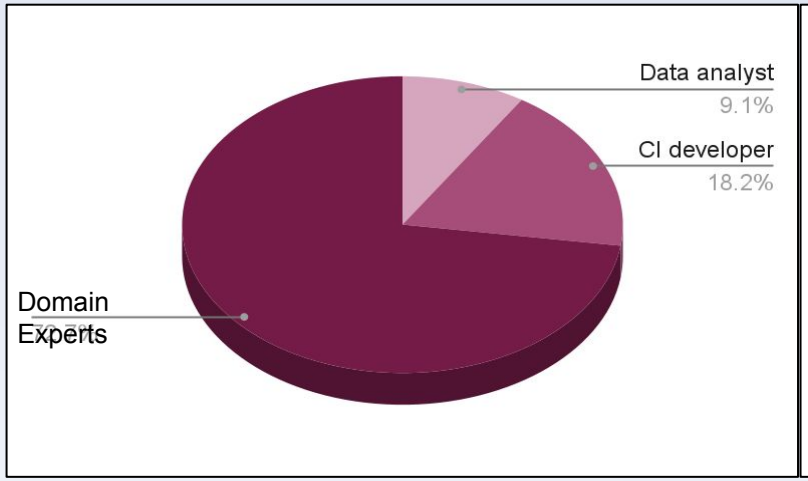
- General questions about data storage, data form, metadata storage, WMS, data catalogs, programming languages
- Specific questions about unique challenges related to role, domain, and institution

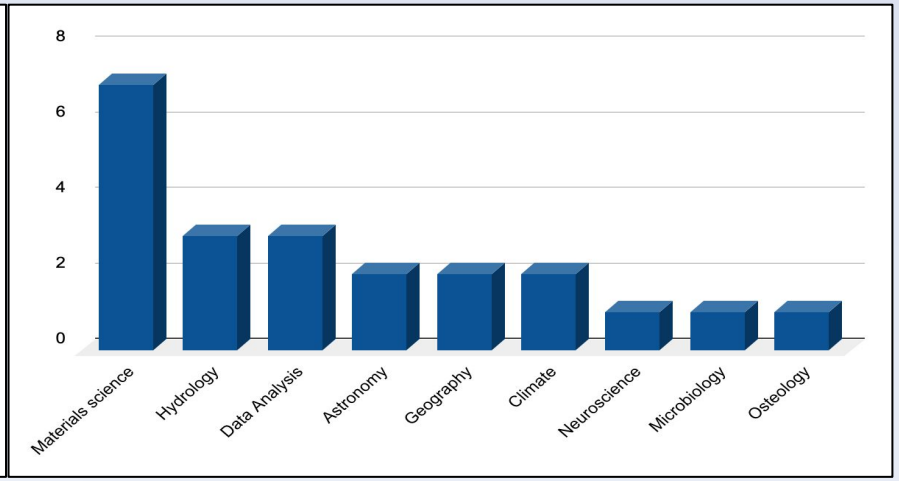
Analyze Results

- Identify cross-cutting concerns
- Identify concerns consistent for roles, domains, and institutions

Diagnose Pain Points

- Distill concerns into concrete problem statements
- Translate into objectives, actionable items, and milestones





Identifying Pain Points: Testimonials

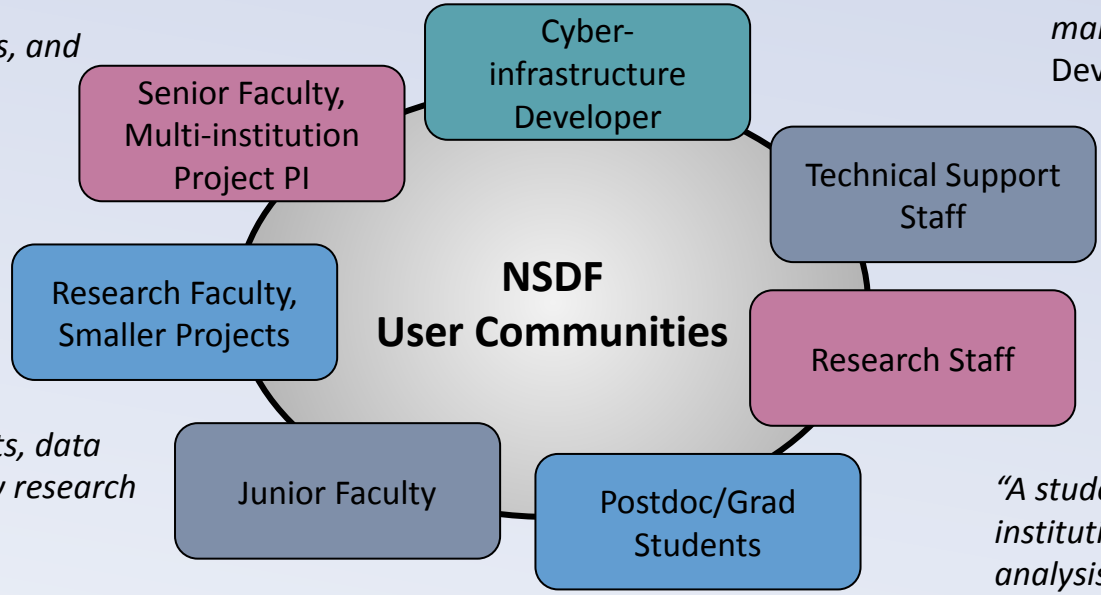
"If we make it easy, people will share their data. We need to extend the scalability of our infrastructure through community supplied storage." - Project PI, Materials Science

"We don't have a plan to scale storage if our cyberinfrastructure takes off." - Cyberinfrastructure developer, Hydrology

"We can't scale to PB data without massive funding or infrastructure. We need centralized point of access to federated data" - Cyberinfrastructure Developer, Materials Science

"We lack personnel to do basic development and maintenance of our systems" - Cyberinfrastructure Developer, Astronomy

"The time and effort for using public repositories, and limited realized gains limits our data sharing" - Faculty, Materials Science



"We can't hire enough system maintainers and *have research funding"* - Data Analyst Group Lead

"Remote quality control during acquisition would let us better use beamline time" - Faculty, Materials Science

"Jupyter notebook access to data (TBs) would reduce barrier to entry " - Research Staff, Climate Science

"Before I had funding to run my own experiments, data shared by a friend at a national lab launched my research career" - Faculty, Materials Science

"A student copies GBs of data from the scientist to my institution. I download to my laptop to prototype analysis. It is cumbersome and limits testing." - Research Staff, Data Analysis

"I'm perplexed by the lack of urgency around reproducible and replicable processes for data management" - Faculty, Geography

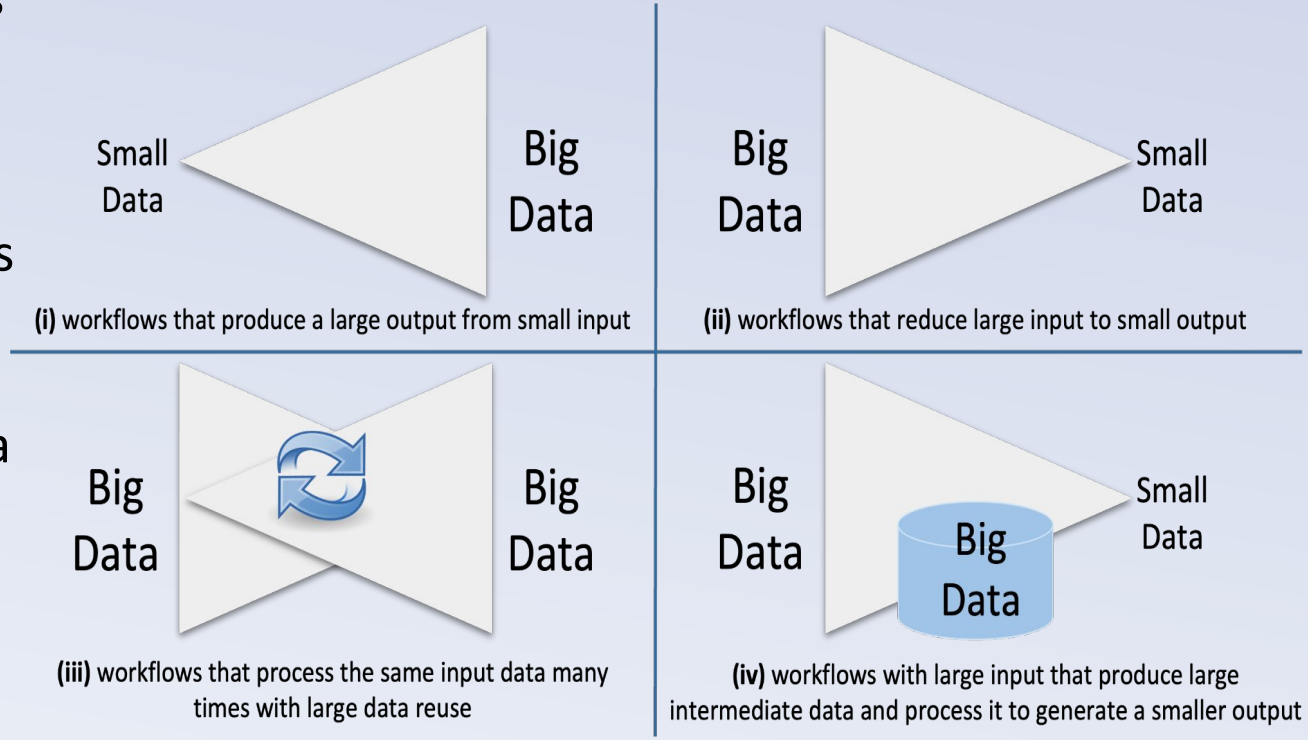
"Our old data is kept on (external) drives. It's hard to keep things organized" - Graduate Student, Materials Science

"Our long-term storage is a shelf of external hard drives" - Research Staff, Neuroscience

"We move data (from light source) by flying back with TB hard drives" - Graduate Student, Materials science

Pain Points Inform NSDF Strategy

Scarcity of Resources: Teams need to work with limited access to human and physical resources
Workforce Development: Limited access to trained personnel hampers CI development
Scalability: No path to scale domain-specific CI
Data Movement: Bottlenecks limit data movements
Data Management: Ad hoc data and metadata management tools result in replicated work
Longevity: Limitations in active storage require data to be shelved
Accessibility: Cumbersome data-sharing processes
Timeliness: Delayed access to rapid data slows science
Replicability: Programs/data versions are created but not maintained as environments change



Our Approach: Implementing an Accessible and Tightly Integrate Data Fabric

Democratizing Access and Use of Large-scale Data

A **data fabric** must be **accessible and tightly integrated** to coordinate data movement between geographically distributed teams or organizations

Develop a **FAIR, AI-ready, transdisciplinary software stack** that is easy to use, integrate, and scale

rated data fabric: a suite

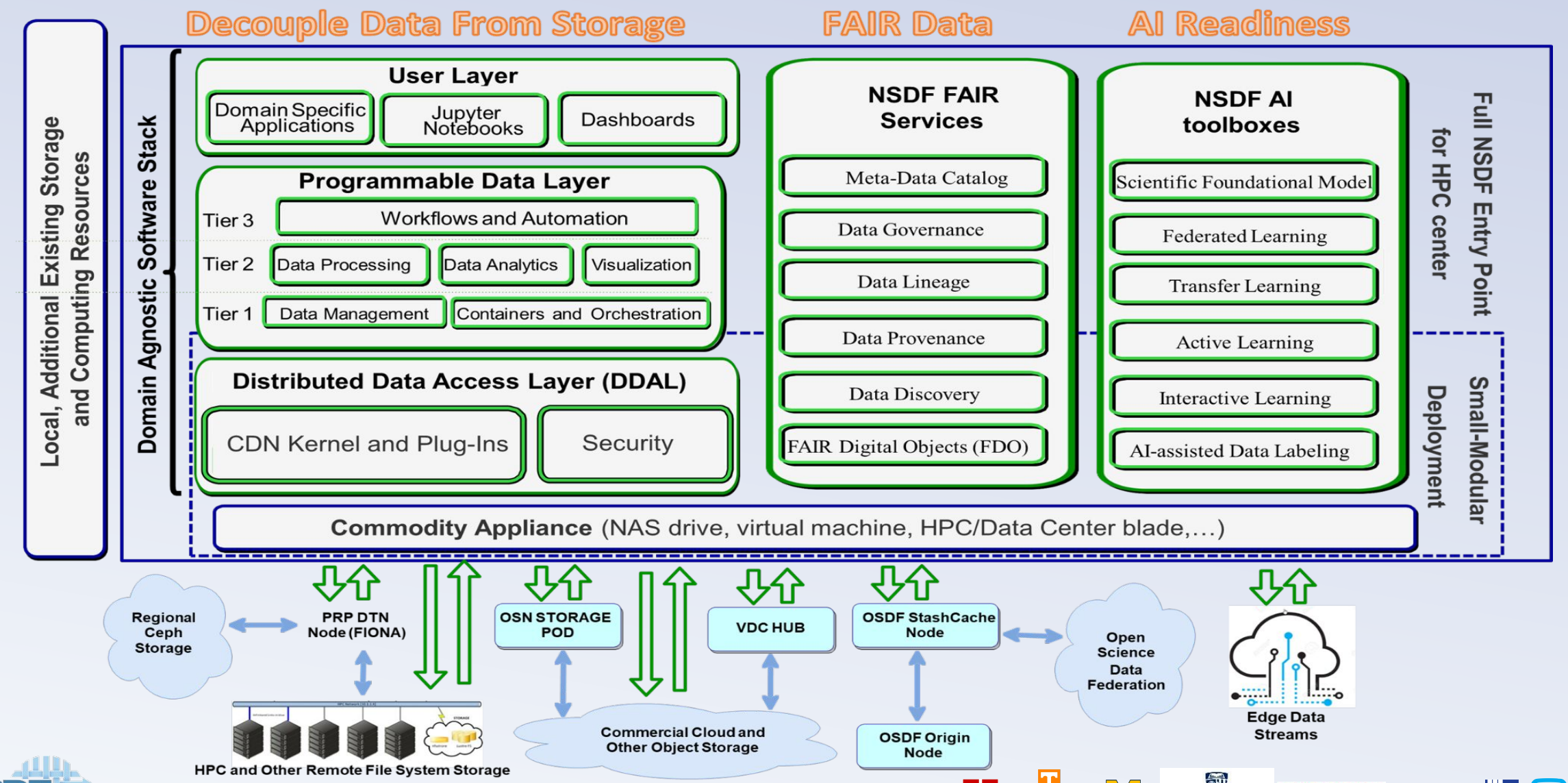
rage services interoper

and commercial cloud

NSDF

http://nationalsciencedatafabric.org/

FAIR, AI-ready, Transdisciplinary Software Stack

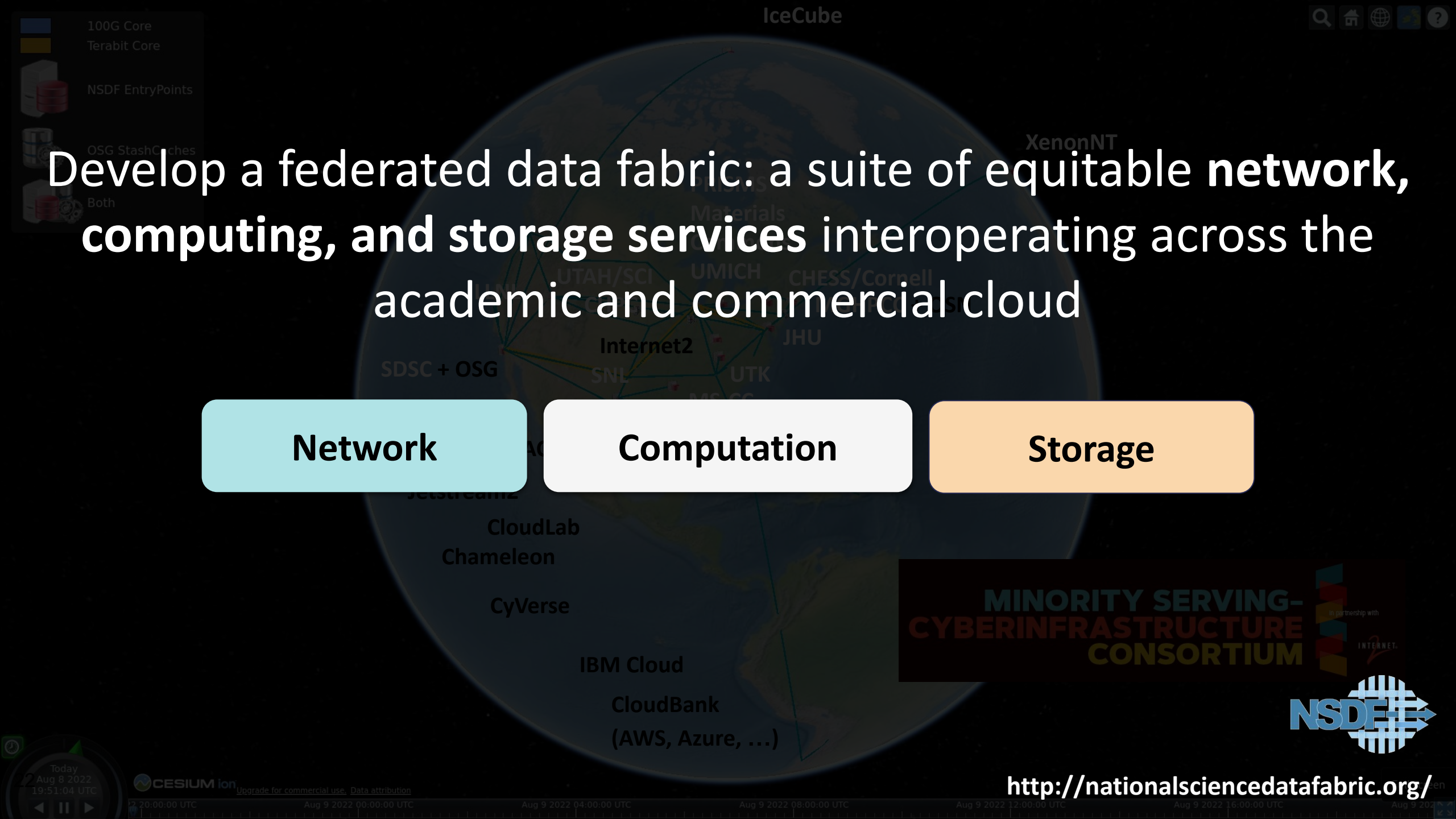


National Science Data Fabric

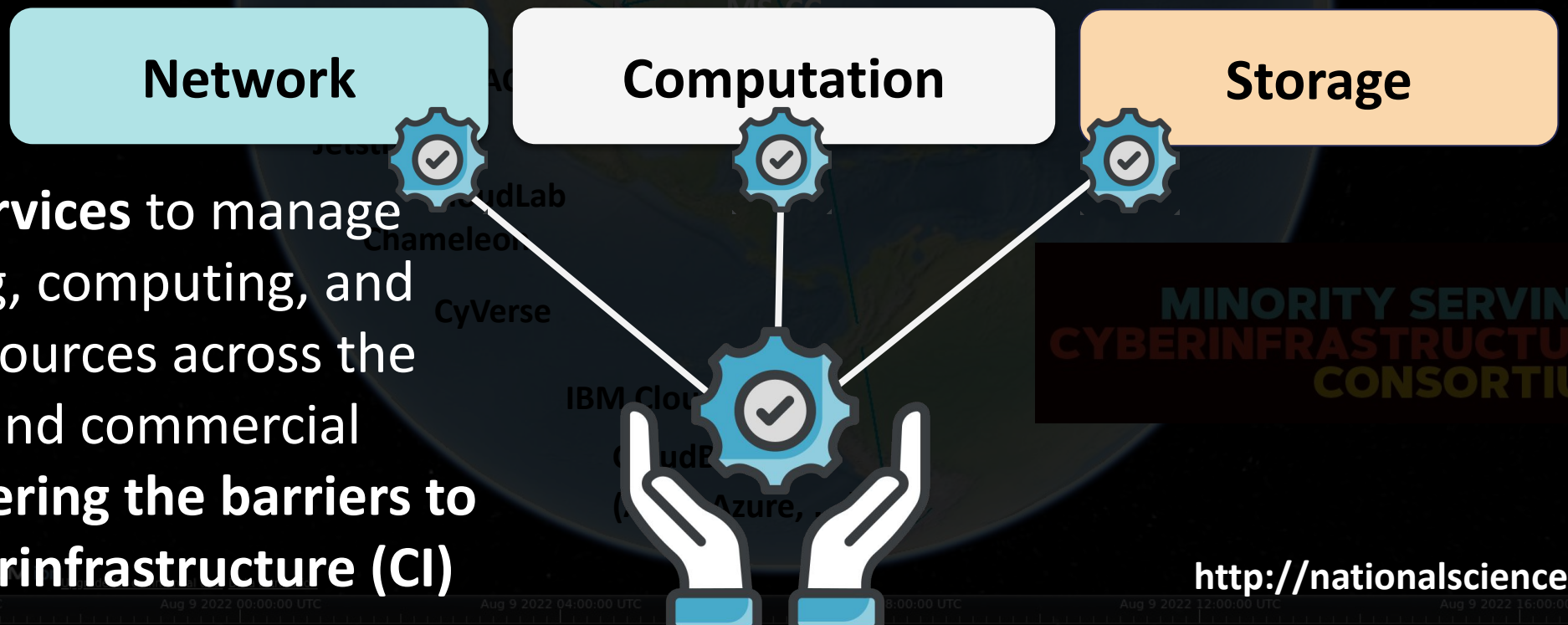
A **data fabric** must be **accessible and tightly integrated** to coordinate data movement between geographically distributed teams or organizations

Develop a **FAIR, AI-ready, transdisciplinary software stack** that is easy to use, integrate, and scale

Develop a federated data fabric: a suite of equitable **network, computing, and storage services** interoperating across the academic and commercial cloud


100G Core
Terabit Core
NSDF EntryPoints
OSG StashCaches
Both
IceCube
XenonNT
Materials
UTAH/SCI
UMICH
CHESS/Cornell
JHU
Internet2
SDSC + OSG
SNL
UTK
CloudLab
Chameleon
CyVerse
IBM Cloud
CloudBank
(AWS, Azure, …)
Develop a federated data fabric: a suite of equitable **network, computing, and storage services** interoperating across the academic and commercial cloud
Network
Computation
Storage
MINORITY SERVING–CYBERINFRASTRUCTURE CONSORTIUM
in partnership with
INTERNET2
NSDF
http://nationalsciencedatafabric.org/
Today
Aug 8 2022
19:51:04 UTC
CESIUM ion
Upgrade for commercial use.
Data attribution
Aug 9 2022 00:00:00 UTC
Aug 9 2022 04:00:00 UTC
Aug 9 2022 08:00:00 UTC
Aug 9 2022 12:00:00 UTC
Aug 9 2022 16:00:00 UTC

Develop a federated data fabric: a suite of equitable **network, computing, and storage services** interoperating across the academic and commercial cloud

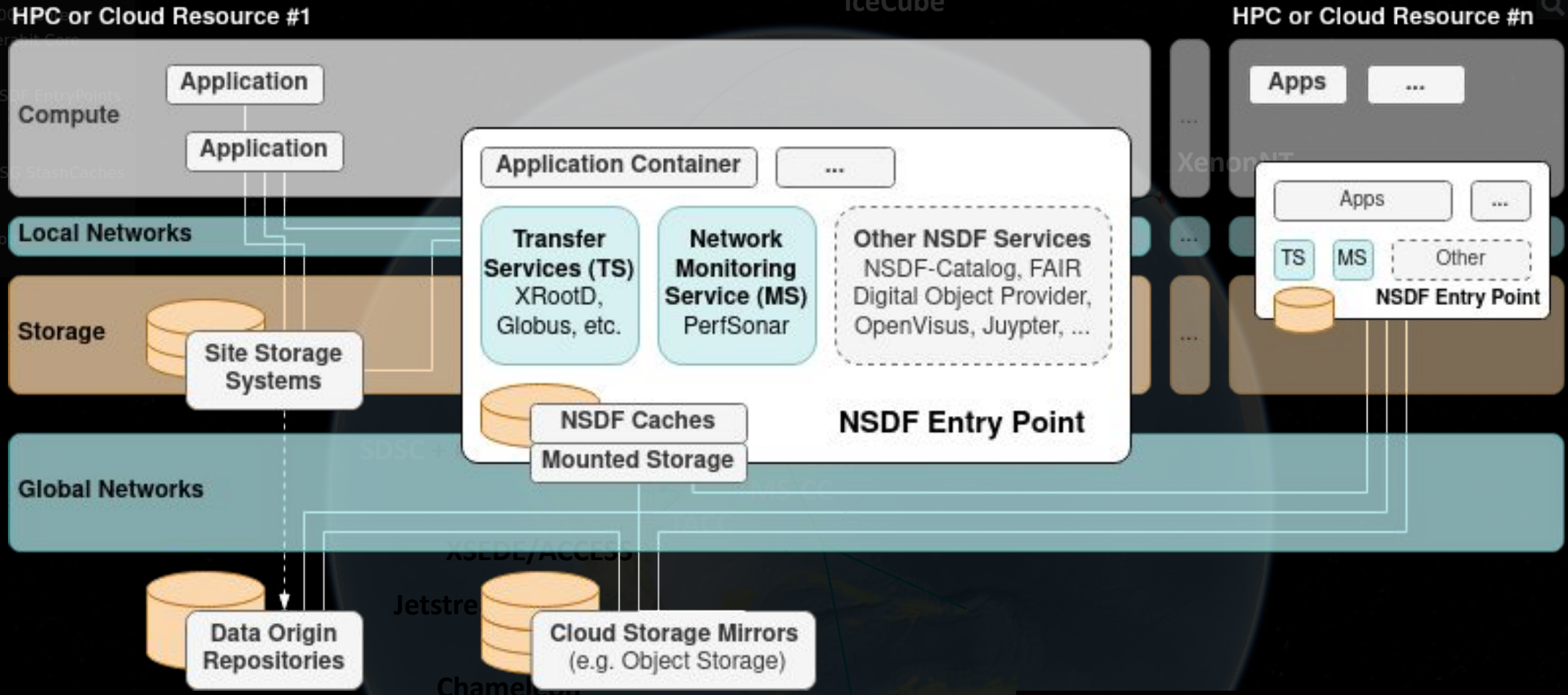


Suite of services to manage networking, computing, and storage resources across the academic and commercial cloud, **lowering the barriers to cloud cyberinfrastructure (CI)**

MINORITY SERVING–CYBERINFRASTRUCTURE CONSORTIUM

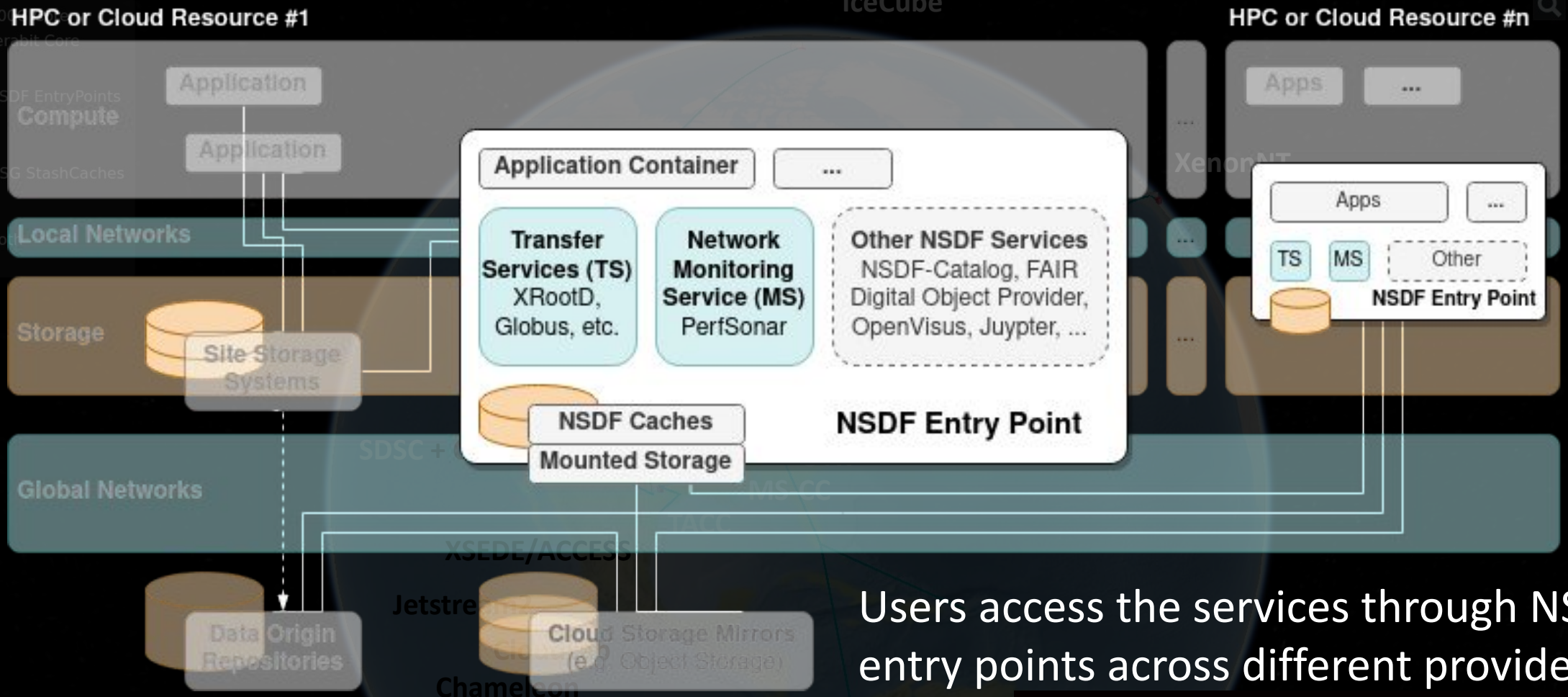
NSDF

http://nationalsciencedatafabric.org/



The NSDF architecture integrates a suite of **networking (both local and global), storage, and computing services**.

http://nationalsciencedatafabric.org/



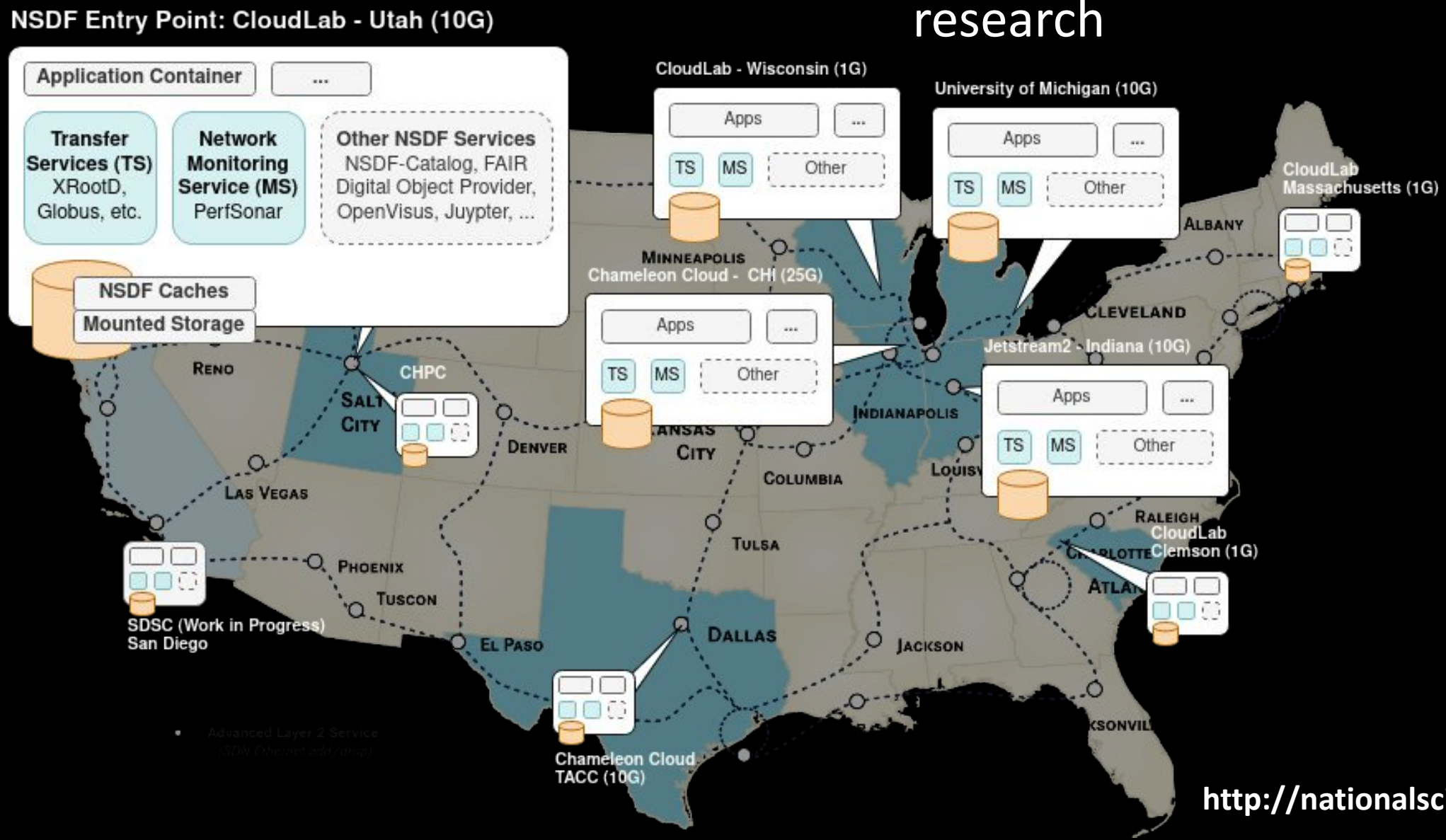
The NSDF architecture integrates a suite of **networking (both local and global), storage, and computing services**.

Users access the services through NSDF's entry points across different providers

The **entry points enable**

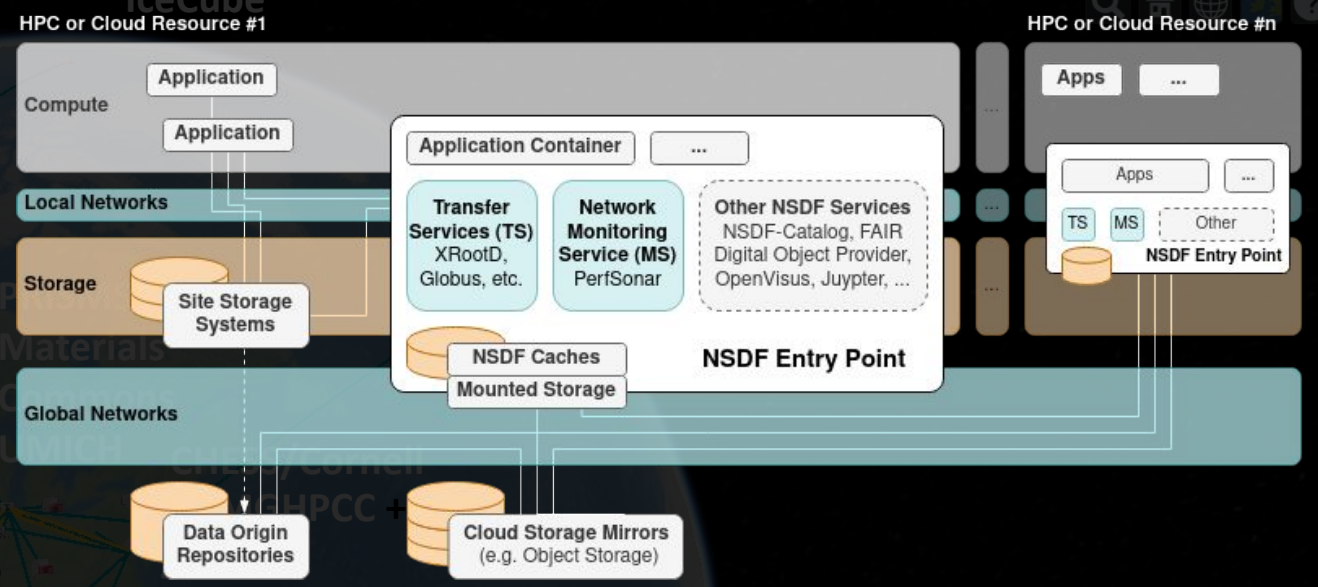
- interoperability of different applications and storage solutions
- fast data transfer and caching among data sources

The current **NSDF testbed** comprises 8 heterogeneous **entry points** in terms of their connections, type of institutions, and research



http://nationalsciencedatafabric.org/

Our **NSDF testbed** integrates networking, computing, and storage services that users access through entry points with different providers



Network

NSDF-Plugin enables efficient data sharing, transfer, and monitoring across networks while hiding the technical complexity of the process

Computation

NSDF-Cloud facilitates users at any entry level in the deployment of the cloud → **one single API** can generate a cluster of many VMs across multiple providers

Storage

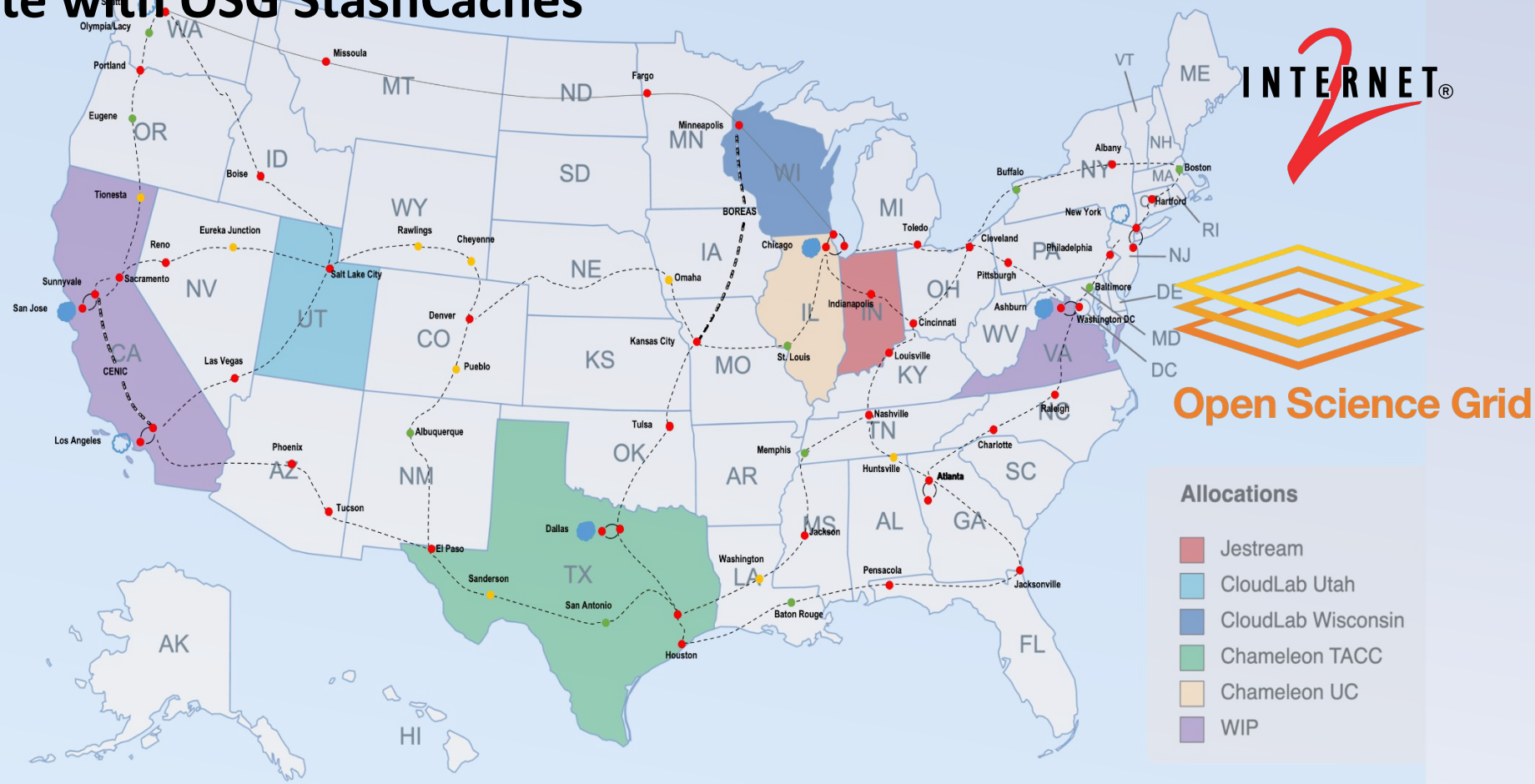
NSDF-FUSE allows the user to reach comprehensive conclusions about mapping packages given different data patterns and cloud platforms

http://nationalsciencedatafabric.org/

Network Services: NSDF-Plugin

Data connection through high-speed network backbone (Internet 2) and designed to interoperate with OSG StashCaches

Build a software stack that uses high-performance data transfer solutions



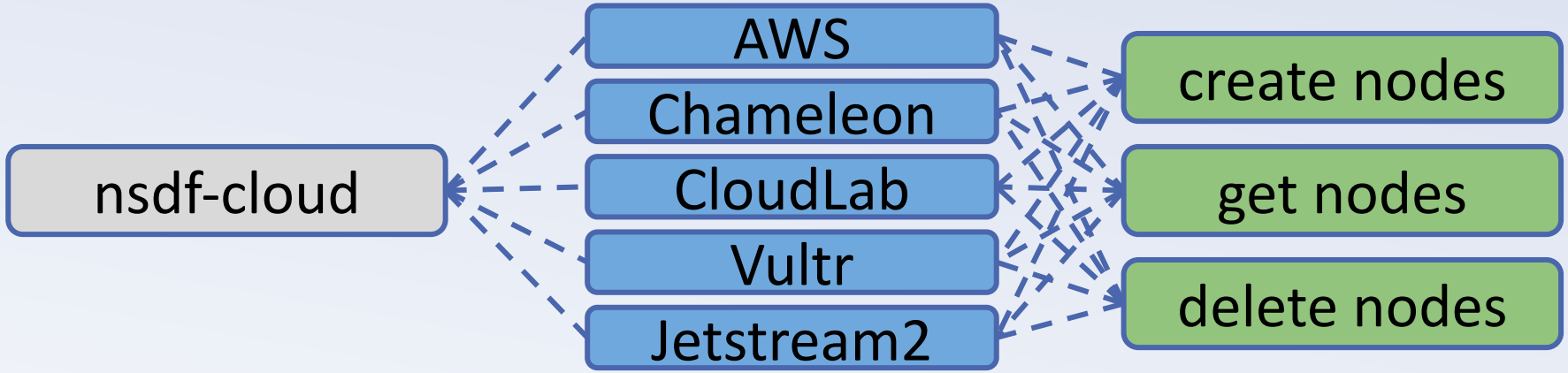
XRootD

globus

Computing Services: NSDF-Cloud

A unified API providing scalable resource management across different providers

- We design computing services built on a unified API for handling diverse jobs across platforms
  - Parallel creation/deletion of many VMs by using command-line tools
  - Automatic generation of Ansible inventory files
  - Integration of credentials for multiple providers via configuration file
- The NSDF-Cloud's unified APIs, both Python and CLI tools, consist of:

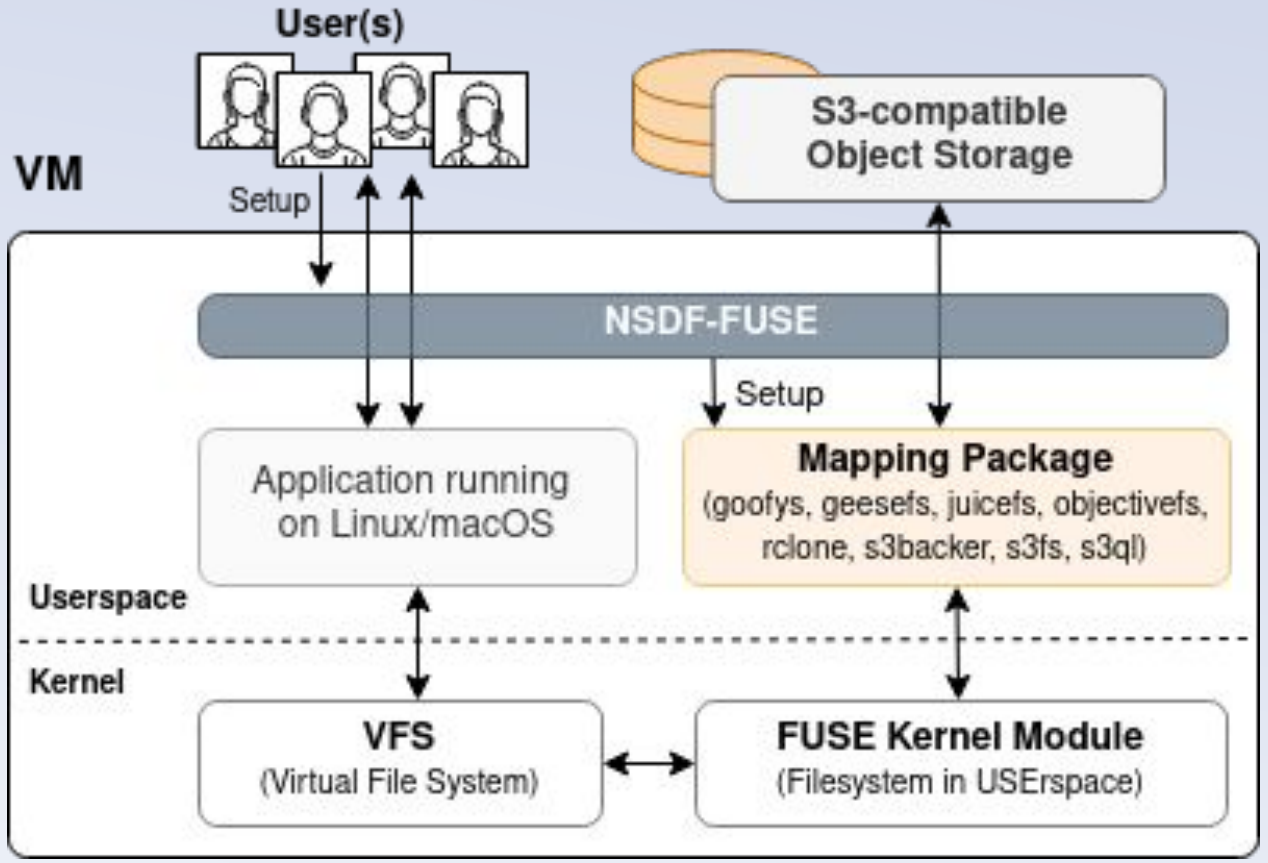


Storage Services: NSDF-FUSE Capabilities

A service for mapping object storage into POSIX namespaces for legacy applications

NSDF-FUSE Capabilities:

- **Creation/deletion of buckets**
- **Installation of mapping package**
- **Mount/unmount buckets as FS**
- **Evaluate I/O performance through I/O jobs**

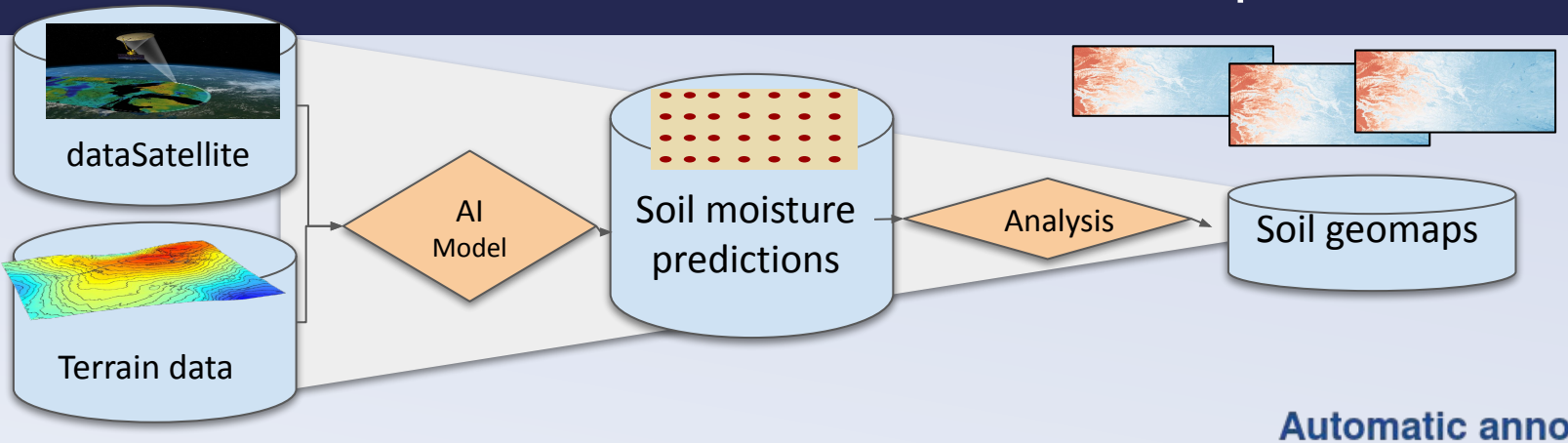


Our Vision: Enabling Scientific Discovery

Democratizing Access and Use of Large-scale Data

Enable Access to Reproducible Workflows

Automatic fine-grained workflow containerization for intermediate data preservation and reuse of elements of the workflow chain; deployment for earth science workflows for soil moisture predictions.

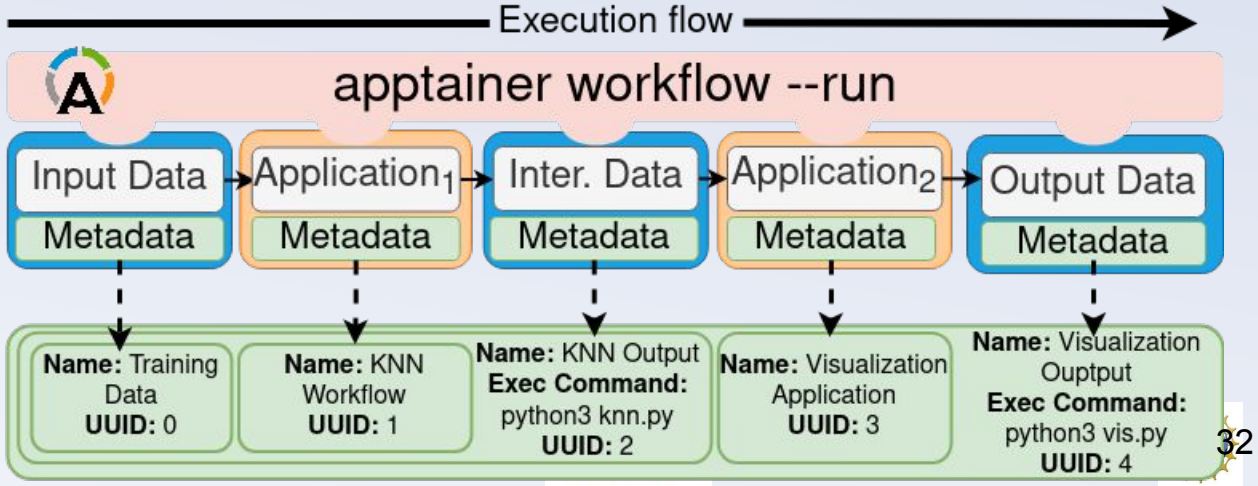


Paula Olaya
Student at UTK

Preservation of intermediate data in dataflow pipelines for reasoning, reproducibility, and replicability

→ Fine-grained containerization of workflow for automatic annotation and **preservation of intermediate data**

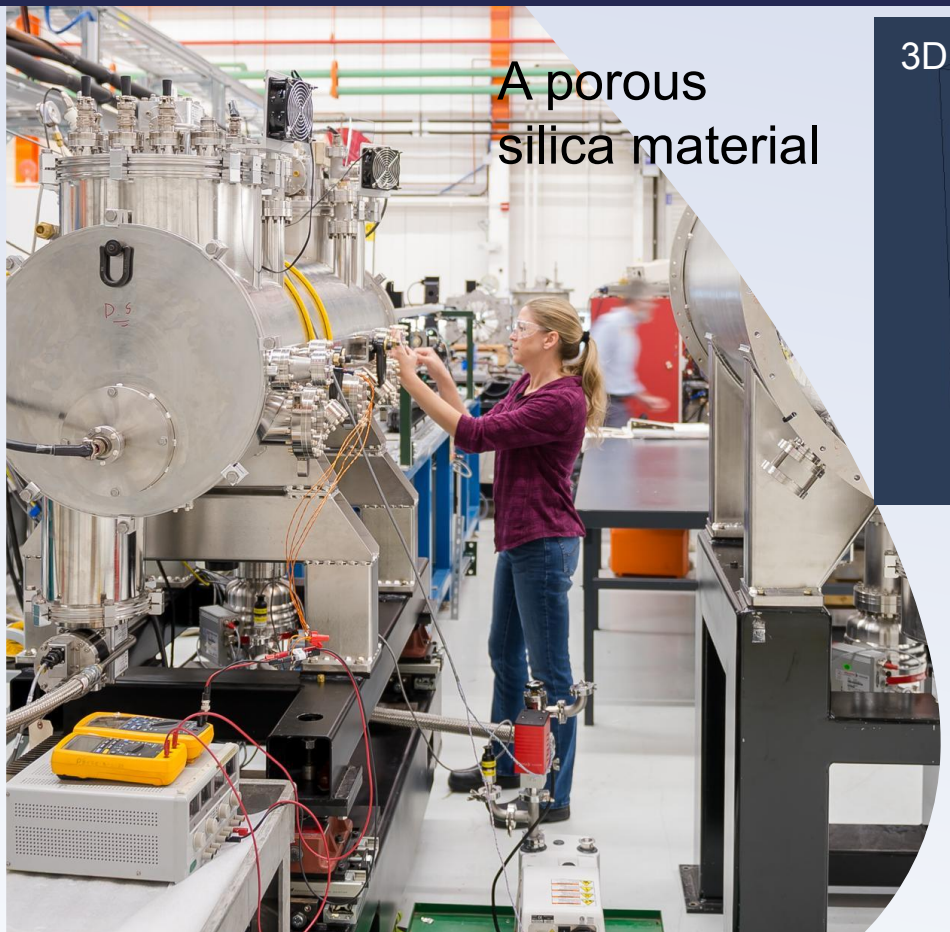
Automatic annotation of provenance metadata in our fine-grained containerized environment



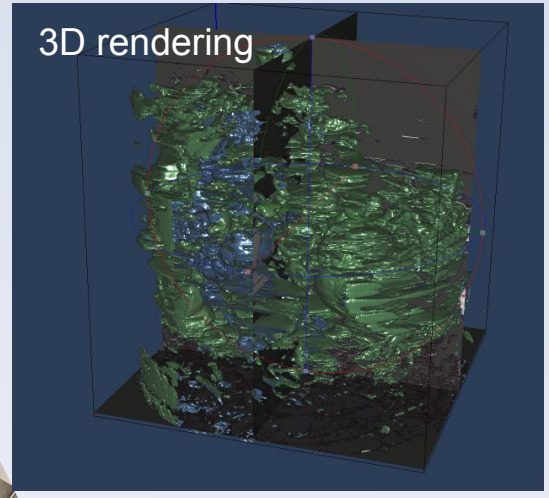
National Science Data Fabric

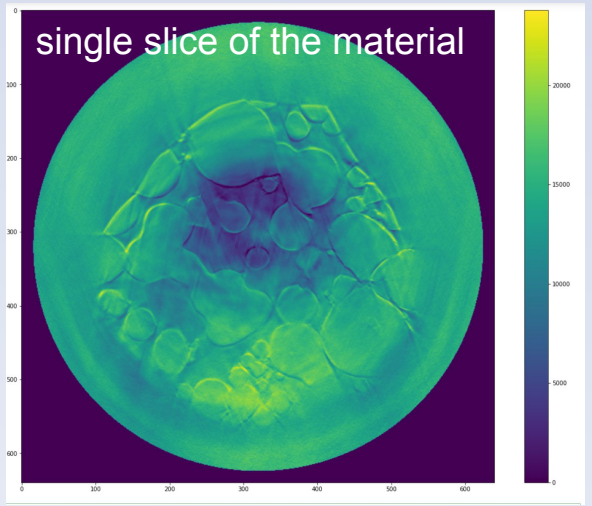
Enabling Material Sciences Research

Facilitating rapid processing of 100+ terabytes of data by compact virtual laboratories, achieving in days what would take months for moving data between national labs and universities.



A porous silica material





Pania Newell
Professor in Mechanical Engineering at UoU

THE UNIVERSITY OF UTAH
MECHANICAL ENGINEERING

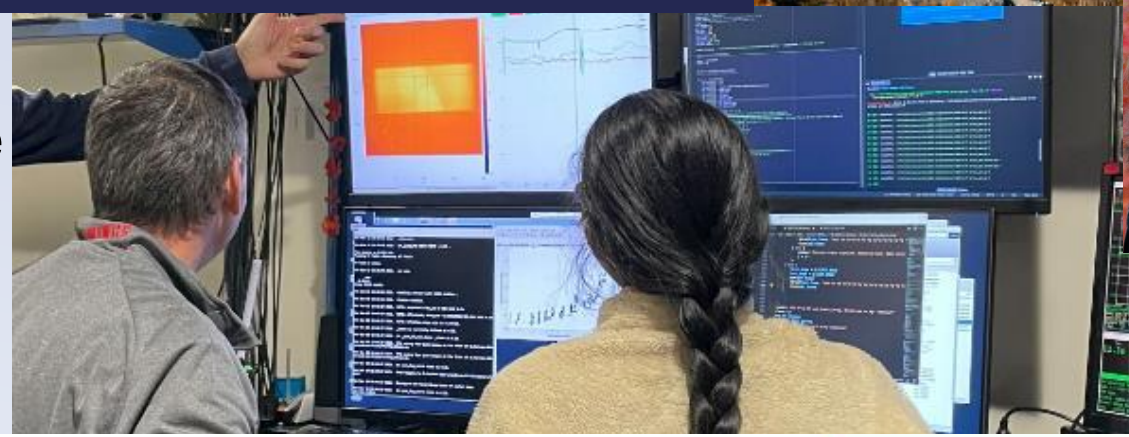
- AI-driven workflows for materials science
- Over 400TB of data generated
- More than 200 machines used on CloudLab, Chameleon, AWS, FluidStack, XSede, and VULTR

http://services.nationalsciencedatafabric.org/materialscience

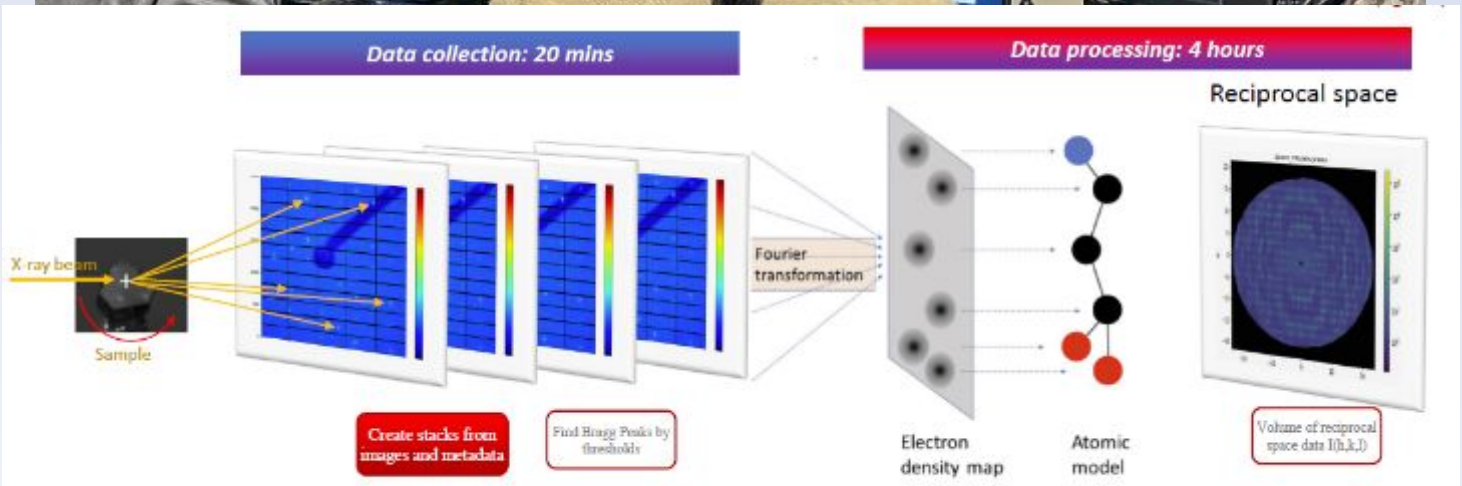
Enable Access to Experimental Facilities

Establishing a comprehensive workflow from experimental facilities to the end-user data analysis.

- Cornell High Energy Synchrotron Source: Quantum Materials Beamline (3 of the 8 lines)
- Real-time data access and sharing
- Steering experiments
- AI-driven workflows
- Remote team collaboration
- Optimize effective use of scientists' time
- Optimize the use of a national resource
- Publish data with no delay (e.g., Materials Commons) □ real-time

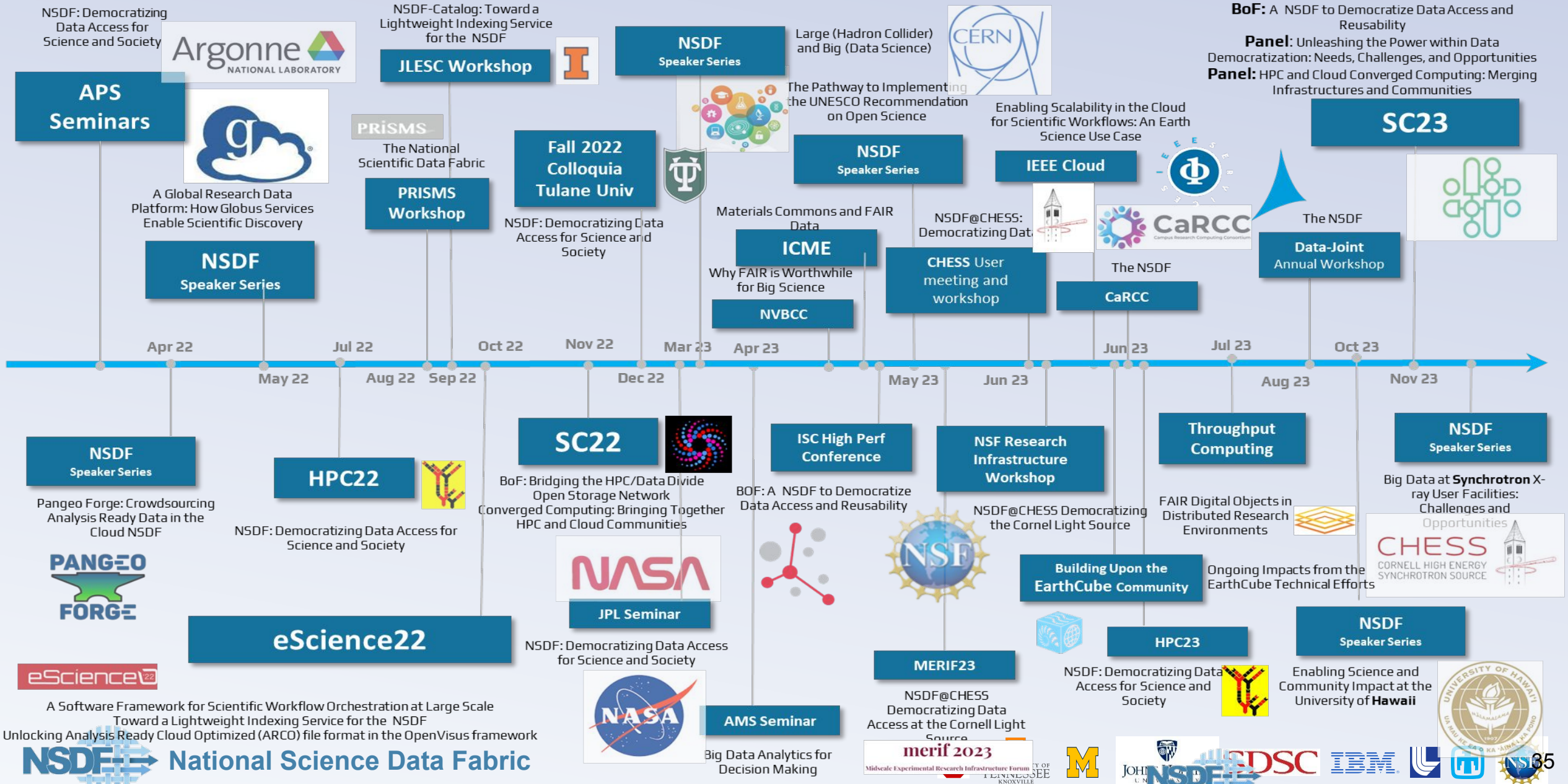


Kate Shanks
Staff Scientist
CHESS, Cornell



NSDF National Science Data Fabric

NSDF Accomplishments and Products



Where to find us

Contact:

info@nationalsciencedatafabric.org

Attendees to the Fourth National Science Data Fabric (NSDF) AHM in San Diego Feb 27-March 1, 2024.

Webpage:

- https://nationalsciencedatafabric.org/nsdf-ahm-2024-02