National Science Data Fabric: data democratization at a national level

V. Pascucci, G. Scorzelli, A. Gyulassy, O. Koppe, A. Panta, D. Hoang, P. Klacansky, S. Petruzza, P.-T., Bremer, University of Utah

M. Taufer, P. Olaya, J. Luettgau, H. Martinez, J. Marquez - University of Tennessee Knoxville

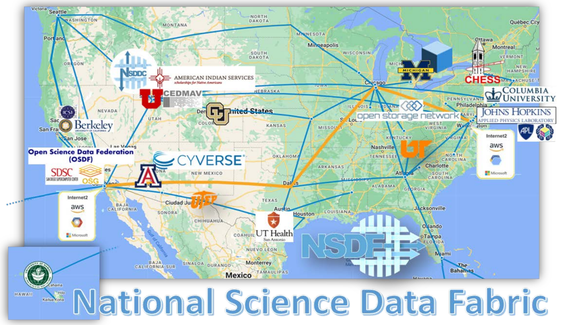
J. Allison, G. Tarcea – University of Michigan

W. Sun, D. Bougie, K. Soloway, R. Verberg, S. Sarker, L. Smieska, A. Woll, K. Nygren Cornell University

F. Wuerthwein, C. Kirkpatrick, F. Andrijauskas, K. Coakley, J. Christopher-Ohlin - San Diego Supercomputer Center

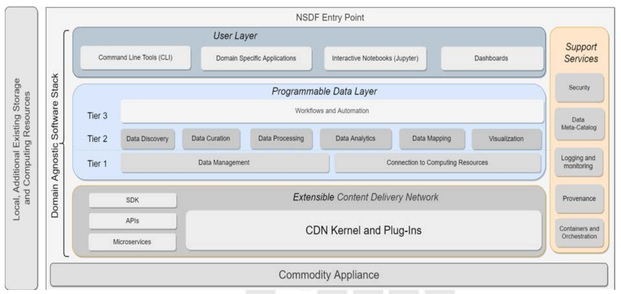
National Science Data Fabric

- *Democratized access* to scientific data
- Trans-disciplinary approach for integrated data delivery and access to shared storage, networking, computing, and educational resources
- **VISION**: globally connected science infrastructure
- **MISSION**: scalable solutions and community training to facilitate deployment on commodity hardware, cloud computing, and HPC resources
- **EXAMPLE**: fulfill the OSTP Guidance to Make Federally Funded Research Freely Available Without Delay



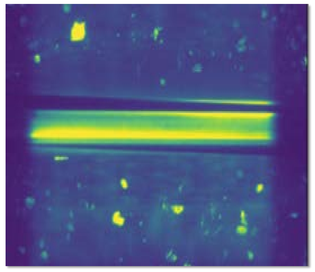
NSDF SOFTWARE STACK

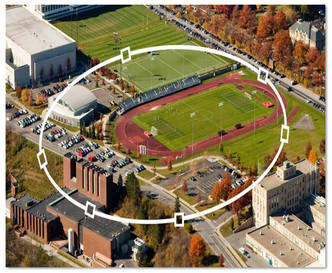
- catalog of best practices for configurations of servers, clients, middleware
- metadata transformation services
- *curation*: functionalities to describe, clean, enhance, and preserve data
- *Automation*: automate processes and speed up delivery of resources on a self-service basis.
- **Interactive** notebooks and dashboards
- data **access patterns** to optimize data sharing

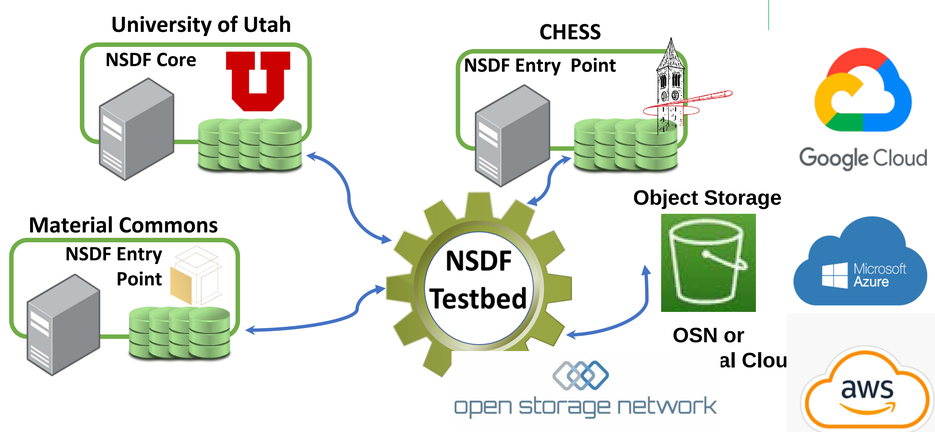


CHESS

- Cornell High Energy Synchrotron Source
- X-Ray user facility, 7 active state-of-the-art beam lines
- **~400TB** of raw data per year
- Serves 250 on-campus users and 1000 visitors yearly
- Stores ~2 petabytes of data

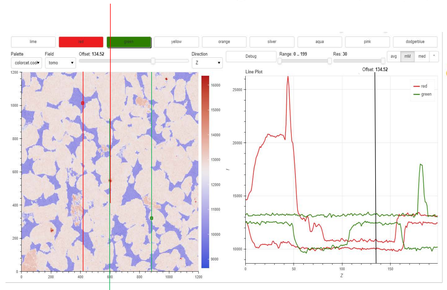


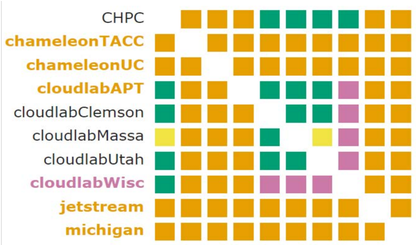




NSDF ENTRYPOINT@CHESS

- Scientific Linux 7 OS install
- **131TB** RAID6 from twelve 16TB drives
- Data accessed locally over NFS, transferred between sites using Globus
- Two INTEL XEON GOLD 6252N 2.3GHZ processors
- 512GB PC4-23400 2933MHz memory
- 25GbE network connection - 10Gb connection to CLASSE
- Accessible through firewall (https, ssh, sftp, and scp)



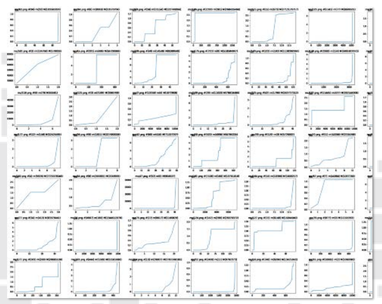


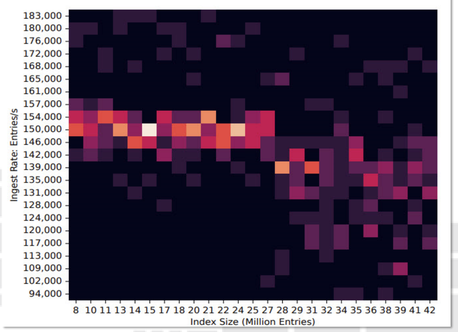
NSDF CATALOG

- multi-federation catalog **1.59 billion** records
- **68 community repositories**
- Answers at scale
- real-time search over 75 petabytes of data
- How many objects are tagged X in Repository Y
- What is the most used file format in all catalogs
- How much “scientific data” is available
- What is the file size distribution of objects in the X catalog
- Are repositories democratic and FAIR

NSDF SERVICES@CHESS

- Publish on-site storage capacity to external users
- Software libraries, tools etc. to handle large datasets from x-ray diffraction, scattering, microscopy, and tomography experiments
- *stream services* for on-the-fly access and fast feedback for experimenters during **remote data acquisition**
- as soon as a beamline produces data, remote users can access it without delay
- exploring large imaging datasets in real time without the need for bulk data transfer
- **GOAL**: streamline workflows by leveraging investments in NSF and open-science CI.





TAKE AWAY MESSAGE

NSDF helps science teams improve **data access performance** and match the resources available with the requirements of their teams and workflows.

Find out more about NSDF on nationalsciencedatafabric.org

NSDF FEDERATED DATA

- Chameleon Cloud
- CloudLab
- CHPC (Center For High Performance Computing)
- PRISM Center UMich
- Open Science Data Federation
- Open Storage Network
- Globus
- SciServer JHU
- Material Commons
- Virtual Data Collaboratory

THE UNIVERSITY OF UTAH · THE UNIVERSITY OF TENNESSEE KNOXVILLE · JOHNS HOPKINS UNIVERSITY · SDSC · NSF

www.sci.utah.edu