Towards a National Storage Fabric
Open Storage Network

Midwest Big Data Hub
Accelerating the Big Data Innovation Ecosystem

Hubs based on Census Regions of the United States

Alaska & Hawaii are part of the West Region
US Territories can participate in any region
Goals

1. Address unmet storage needs around active data sets
2. Proximity to other resources (e.g. computational)
3. Leverage, leverage, leverage!!!
“Big Data”

Large quantities of data
Large varieties of data

“Long-Tail”

The “Long-Tail” of “Big Data”


**Shedding Light on the Dark Data in the Long Tail of Science**

P. Bryan Heidorn

**Abstract**

One of the primary outputs of the scientific enterprise is data, but many institutions such as libraries that are charged with preserving and disseminating scholarly output have largely ignored this form of...
<table>
<thead>
<tr>
<th>Project</th>
<th>Average size of data entities</th>
<th>Total data volume</th>
<th>Storage problem being solved</th>
<th>Use case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical Zone Observatories</td>
<td>10 MB</td>
<td>50 TB</td>
<td>Provide storage space and access to CZO datasets and community-generated data</td>
<td>Community long-tail data</td>
</tr>
<tr>
<td>TerraFusion</td>
<td>10 GB</td>
<td>1 PB</td>
<td>Transport datasets across the US at high speed, obtain data slices with high probability of reutilization</td>
<td>Experiment-to-site, Slice-and-compute</td>
</tr>
<tr>
<td>HathiTrust Research Center collection</td>
<td>200 MB</td>
<td>500 TB</td>
<td>Provide storage space and access to the HTRC dataset and further community-generated derivatives</td>
<td>Common resource access</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>10 GB</td>
<td>1 PB</td>
<td>Make available a well-curated dataset for testing machine learning algorithms</td>
<td>Dataset-as-benchmark</td>
</tr>
<tr>
<td>Large Synoptic Survey Telescope</td>
<td>2 TB</td>
<td>100 PB</td>
<td>Transport datasets across the US at high speed, obtain data slices with high probability of reutilization, facilitate inter-site data processing</td>
<td>Experiment-to-site, Slice-and-compute, Workflow staging space</td>
</tr>
<tr>
<td>Combined Array for Research in Millimeter Astronomy</td>
<td>50 MB</td>
<td>50 TB</td>
<td>Transport datasets across the US at high speed, obtain data slices with high probability of reutilization</td>
<td>Experiment-to-site, Slice-and-compute</td>
</tr>
</tbody>
</table>
Use Case Types

OSN use cases

Multi-institution research collaborations (80% data volume)
- Independent claim verification
- Dataset-as-benchmark
- Breakthrough datasets
- Experiment-to-site (fan-out)

Single-institution projects (20% data volume)
- Independent claim verification
- Dataset-as-benchmark
- Breakthrough datasets
- Community long-tail data
- Data for new applications
- Common resource access
- Workflow staging space
- Slice-and-compute
Proximity between Data and Computation

• Avoid moving data
• Provide analysis capabilities near data whenever possible
  • Gateways, Workbenches, ...
Leverage, leverage, leverage
Big Data Hubs
Stimulate regional grassroots partnerships focused on Big Data

CC*
Coordinated campus-level cyberinfrastructure components of data, networking, and computing infrastructure across 200+ universities
Research Data Challenges

• Storage
• Everything else!!!
  • The bytes are not enough on their own
  00110100 00110010
  • Metadata, curation tools, indexes, storage abstraction, replication, data transfer, authentication, access control, transformation, analysis, tools, computation, …
• Develop a deep symbiotic relationship between science and engineering users and developers of cyberinfrastructure to simultaneously advance new research practices and open transformative opportunities across all science and engineering fields

• Provide an integrated and scalable cyberinfrastructure that leverages existing and new components across all areas of CIF21 and establishes a national data infrastructure and services capability

• Ensure long-term sustainability for cyberinfrastructure, via community development, learning and workforce development in CDS&E and transformation of practice

http://www.nsf.gov/cif21/
Data Infrastructure Building Blocks

- **Data Applications**
  - Particles, Materials, Astro, Geo, GIS, Bio, Social, Environ, Ag, Medical, Sensors, Education, Psychology, Hydrology, etc.

- **Data Cyberinfrastructure**
  - Computing, Storage, Federation, Clouds, Networking, SDN

- **Data Trust, Security, and Privacy**

- **Data Curation**
  - Capture, Annotation, Documentation, Archiving, Libraries, Management, Publishing, Quality, Validation

- **Data Discovery and Exploration**
  - Semantics, Ontology, Metadata, Data Mining, Web, Search, Visualization, Validation

- **Data Sharing Middleware**
  - Accessibility, Collaboration, Hubs, Repositories

- **Data Workflows & Analytic Notebooks**

- **Data Analytics and Analysis**
  - Data-Intensive Computing, Machine Learning, NLP, Statistics, Simulation

https://dibbs18.ucsd.edu
Architecture

Leverage NSF Data Software Investments e.g. DataNet, DIBBs, CSSI

OSN Pod

Leverage NSF XSEDE Capabilities

- AMIE
- XRAS

<table>
<thead>
<tr>
<th>GloBus</th>
<th>Clowder</th>
<th>WholeTale</th>
<th>iRODS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S3 API

- Ceph + radosgw
- Remote Operations, Boot, Monitoring, Management
- OSN Hardware
### OSN Nodes

<table>
<thead>
<tr>
<th></th>
<th>88-Disk OSN Node</th>
<th>24 Disk OSN Node (128GB)</th>
<th>24 Disk OSN Node (256GB)</th>
<th>36 Disk OSN Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Sockets</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Per-CPU Cores (HT)</td>
<td>6 (12)</td>
<td>12 (24)</td>
<td>12 (24)</td>
<td>18 (36)</td>
</tr>
<tr>
<td>Total CPU Cores (HT)</td>
<td>12 (24)</td>
<td>12 (24)</td>
<td>12 (24)</td>
<td>18 (36)</td>
</tr>
<tr>
<td>CPU Speed</td>
<td>3.4 GHz</td>
<td>2.1 GHz</td>
<td>2.1 GHz</td>
<td>2.3 GHz</td>
</tr>
<tr>
<td>Total GFLOPS</td>
<td>652.8</td>
<td>403.2</td>
<td>403.2</td>
<td>662.4</td>
</tr>
<tr>
<td>Total RAM (GB)</td>
<td>256</td>
<td>128</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>RAM Speed (MHz)</td>
<td>2400</td>
<td>2400</td>
<td>2400</td>
<td>2666</td>
</tr>
<tr>
<td>Drive Count (AKA OSD)</td>
<td>88</td>
<td>24</td>
<td>24</td>
<td>36</td>
</tr>
<tr>
<td>Per-Node Data Storage (TB)</td>
<td>704</td>
<td>192</td>
<td>192</td>
<td>288</td>
</tr>
<tr>
<td>~1.5PB Node Count</td>
<td>2</td>
<td>8</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>Total Data Storage</td>
<td>1408</td>
<td>1536</td>
<td>1536</td>
<td>1440</td>
</tr>
<tr>
<td>Node Cost</td>
<td>$51,492.00</td>
<td>$15,528.00</td>
<td>$16,672.00</td>
<td>$17,845.00</td>
</tr>
<tr>
<td>Total Cost</td>
<td>$102,984.00</td>
<td>$124,224.00</td>
<td>$133,376.00</td>
<td>$89,225.00</td>
</tr>
<tr>
<td>$ per TB</td>
<td>$73.14</td>
<td>$80.88</td>
<td>$86.83</td>
<td>$61.96</td>
</tr>
<tr>
<td>Utilized Rack Space (RU)</td>
<td>24</td>
<td>36</td>
<td>36</td>
<td>24</td>
</tr>
<tr>
<td>Max Power Consumption (Watts)</td>
<td>7600</td>
<td>9600</td>
<td>9600</td>
<td>6800</td>
</tr>
<tr>
<td>Ceph RAM:Storage Cap Ratio (Nearest to 1 is best)</td>
<td>0.36</td>
<td>0.67</td>
<td>1.33</td>
<td>0.89</td>
</tr>
<tr>
<td>Ceph CPU Thread: OSD Ratio (Nearest to 1 is best)</td>
<td>0.27</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
OSN Node Breakdown

NETWORKING
- 32 Ports Non-blocking 100Gb Ethernet
- 16 ports allocated to Internet VLAN
- 16 ports allocated for internal cluster VLAN
- Gigabit BMC/LOM Connections

STORAGE
- Five Storage Nodes
- 1.44PB Raw storage
  - 36 8TB Drives per Node
  - 12Gb Seagate Helium SAS Drives

Rack and Power
- 24 Rack Units
- 6.8kW maximum power consumption
- 4.9kW typical (estimate)
What are the site requirements?

Network, space, power cooling

Pods need:

- WAN Network
  - Two high speed network connections (Data)
  - 16 site domain IP addresses (one /28 address range)
- Out of Band Management
  - Three management network connections
- 12 Sq. feet of floor space
- Two x 208V 3P, 60A power feeds
- Appropriate environmental support systems

https://www.openstoragenetwork.org/documentation/hardware-interconnects/
**Site Owner/Operator**
- Fundamental rack (POD) installation, power and network cabling at delivery
- Site maintenance and “remote hands” POD repair work

**Command Center**
- Single point of contact
- Handles day-to-day management of OSN storage resources for all sites
- Central IDM, security and resource management
- Centralized logging, monitoring, and reporting
- Centralized provisioning and updates
How does a pod become a storage node?

*Software installation: On Site Admin*

1. Connect and power on gateway node
How does a pod become a storage node?

**Software installation: Command Center**

1. Connect and power on gateway node

**Local – On Site Admin**

1. Bootstrap remaining nodes via ansible from command center
   1. Configs pulled from project GitHub repo
   2. Images pushed to gateway node
   3. IPMI / PXE used to bootstrap remaining nodes

---

**Command Center – OSN Admin**

1. Bootstrap remaining nodes via ansible from command center
   1. Configs pulled from project GitHub repo
   2. Images pushed to gateway node
   3. IPMI / PXE used to bootstrap remaining nodes
How does a pod become a storage node?

**Software installation: Command Center**

1. **Connect and power on gateway node**

2. **Execute baseline system services installation**

**Local – On Site Admin**

1. Bootstrap remaining nodes via ansible from command center
   1. configs pulled from project GitHub repo
   2. Images pushed to gateway node
   3. IPMI / PXE used to bootstrap remaining nodes

**Command Center – OSN Admin**

1. Bootstrap remaining nodes via ansible from command center
   1. Configs pulled from project GitHub repo
   2. Images pushed to gateway node
   3. IPMI / PXE used to bootstrap remaining nodes

2. Execute baseline system services installation
How does a pod become a storage node?

**Software installation: Command Center**

**Local – On Site Admin**

1. Connect and power on gateway node

**Command Center – OSN Admin**

1. Bootstrap remaining nodes via ansible from command center
   1. configs pulled from project GitHub repo
   2. Images pushed to gateway node
   3. IPMI / PXE used to bootstrap remaining nodes
2. Execute baseline system services installation
3. Provision storage system (Ceph)
4. Provision user service stack
   1. S3, Globus, etc.
Then what?

Remote monitoring and management: Command Center

Pod monitoring is done using the “TIG” stack. Collected data is available onsite and forwarded to the Command Center.

• Consistent, pervasive infrastructure
• Telegraf, InfluxDB & Grafana
• Dashboards are customizable to suit many scenarios
• Command Center receives system alerts and notifies local site of events and if actions are required
<table>
<thead>
<tr>
<th>Capability</th>
<th>Origin</th>
<th>OSN Nodes Minimal Viable Product (MVP)</th>
<th>OSN Community Portals Reference Implementation(s)</th>
<th>OSN Web Outreach / Marketing</th>
<th>Post Pilot Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3 API</td>
<td>Proposal</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Authentication</td>
<td>Proposal</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Central System Upgrades</td>
<td>Proposal</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>User Interface: Search/Download</td>
<td>Proposal</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td>Proposal</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allocations Management</td>
<td>Architecture'18</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Naming/Namespace Management</td>
<td>Architecture'18</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Replication</td>
<td>Architecture'18</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Location Abstraction</td>
<td>Architecture'18</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Provenance of data</td>
<td>Architecture'18</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>List of OSN Node Sites</td>
<td>Wishlist</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSN Example Use Cases</td>
<td>Wishlist</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Globus

Clowder (2013-Present)

NSF Innovative Systems and Software: Applications to NARA Research Problems (OCI-0525308)

- Medici 2009-2013, ONR TRECC (N00014-04-1-0437)
- Clowder 2013, NSF Innovative Systems and Software: Applications to NARA Research Problems (OCI-0525308)
Sustainability

• Pure Project Funding Model -> Mixed Business Model
  • Service model packaging development of capabilities with value add services such as hosting, support, training, consultation
  • Institutional support, e.g. data-centric resources
  • Science driven project funding
Clowder Community

Object Store Support

<table>
<thead>
<tr>
<th></th>
<th>Biology</th>
<th>Civil Engineering</th>
<th>Comp. and Inf. Science</th>
<th>Education</th>
<th>Geoscience</th>
<th>Humanities</th>
<th>Industry</th>
<th>Materials-Science</th>
<th>Medicine</th>
<th>Social Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brown Dog</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BRAE CLET</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chicapomoriraza</td>
<td>University of Michigan</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countermeasures against Radiation</td>
<td>BARDA</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crowd-Sourced Green Infrastructure</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DataNet SEAD</td>
<td>NSF, NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EarthCube GeoSemantics</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eCam</td>
<td>European Commission</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IMCL-COD</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunomodulatory Effects of MSC</td>
<td>NIH</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Great Lakes Monitoring</td>
<td>Illinois-Indiana SeaGrant</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Great Lakes to Gulf</td>
<td>NGRECC, Walton Foundation</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Groupscope</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KISTI</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LinkSEEM</td>
<td>European Commission</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NARA</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NIST</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P6An</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIVeL3D</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIMDAS</td>
<td>European Commission</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southern Methodist University</td>
<td>SMU</td>
<td>X</td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syngenta</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T2-C2</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Taiwan NCHC</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TERRA-REF</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRECCE</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vector-Borne Disease</td>
<td>CDC</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VI-SEEM</td>
<td>NSF</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XSEDE Deccomposing Bodies</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XSEDE Image Analysis of Rural Photography</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XSEDE Large Scale Video Analytics</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XSEDE Real Stories of Bad Kids</td>
<td>NSF</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ILLINOIS NCSA
OSN Clowder Interface

- IML-CZO use case data
Increased reliance on computation across domains
  ○ new skill requirements for researchers

Open Science changing norms and expectations
  ○ increased emphasis on sharing data & code
  ○ ... with transparency and reproducibility in mind!
  ○ => from sharing data to sharing research objects
  ○ FAIR principles
Whole Tale & the Elements of a Reproducible Computational Research Platform

Develop
- Easy-to-access cloud-based computational environments

Analyze
- Transparent access to research data

Share
- Collaborate and share with others

Package
- Export or publish executable research objects

Reproduce
- Re-execute Review Verify Re-use

Coming soon
Tale Creation Workflow

"Analyze in WT" or register data by URL or digital object identifier:

Create a Tale, entering a name and selecting interactive environment

A container is launched based on selected environment with an empty workspace and external data mounted read-only

Customize environment adding special packages/software dependencies

Create/upload code and scripts

Create/upload code and scripts

Execute code/scripts to generate results/outputs

Export the Tale in compressed BagIt-RO format to run locally for verification.

Enter descriptive metadata including authors, title, description, and illustration image

Publish the tale to a supported repository, generating a persistent identifier.

Re-execute in Whole Tale

schema:author
schema:name
schema:category
pav:createdBy
schema:license
Whole Tale Platform Overview

- **Authenticate** using your institutional identity
- **Access** commonly-used **computational environments**
- Easily **customize** your environment (via repo2docker)
- Reference and access externally **registered data**

- **Create or upload** your **data and code**
- Add **metadata** (including **provenance** information)
- Submit code, data, and environment to **archival repository**
- Get a **persistent identifier**
- **Share** for **verification and re-use**
Whole Tale + Invenio + OSN

- Users transfer data to OSN pod using Globus
- Whole Tale exposes data and interactive environment (e.g., Jupyter) using Globus or S3
- Invenio would support publishing data/research objects that can also be exposed via WT
What’s Next

- Working more closely with use cases
- Application layer and additional Command Center features
- Partnership with XSEDE, TrustedCI Review, OSG, NRP/PRP
- Refine governance and shared administration
- Progress on policies, procedures around data lifecycle
- Performance tuning and evaluation of POD hardware recommendations
- Planning for future OSN at increasing scales
Demand for OSN

- Army Research Lab (DoD) – private, distributed storage, data sharing
- BDHubs – incubating collaborations, multisector data sharing
- Center for Applied Internet Data Analysis (CAIDA)
  - 100TB internet telescope data, need to push geographically to collaborators
- California Digital Library (CDL) – build out of UC Data Network, data sharing, data discovery
- Data Science Institutes/HDR Initiatives – Tanya Berger-Wolf @UIC
- Lei Wang @Northwestern (MBDH Neuroscience Spoke) SchizConnect - large database for researchers studying schizophrenia
  - OSN “receive, hold, and release” after the researcher retrieved the dataset.
- NIH – onramp to commercial cloud
- NIH PIs – sharing data, commercial cloud cache, e.g. R. Vanguri @ Colombia
- OSG – allocate storage with compute, scavenge for storage, etc.
- UIUC cancer research group producing and analyzing infrared imagery.
  - Need centralized service to hold their large datasets and “pass back and forth” between hold, retrieve, analyze, put back, repeat.
- University of Arizona, academic CIO community – ready to buy a pod, expand OSN
- …