https://www.openstoragenetwork.org

The Open Storage Network: Distributed Storage Cyberinfrastructure for Data-Driven Science

Santiago Nuñez-Corrales[1], Melissa Cragin[1,3], Kenton McHenry[2], Michael Norman[3], Christine Kirkpatrick[4], John Goodhue[5], Stanley Ahalt[6], Lea Shanley[7], Derek Simmel[8], Alex Szalay[9]

[1]MBDH NCSA UIUC; [2]NCSA UIUC; [3]SDSC UCSD; [4]NDS and WBDH; [5]GCHPCC MIT; [6]RENCI; [7]UNC Chapel Hill; [8]PSC UP-CM; [9]IDIES JHU. Corresponding authors: { nunezco2@illinois.edu; mcragin@ucsd.edu }

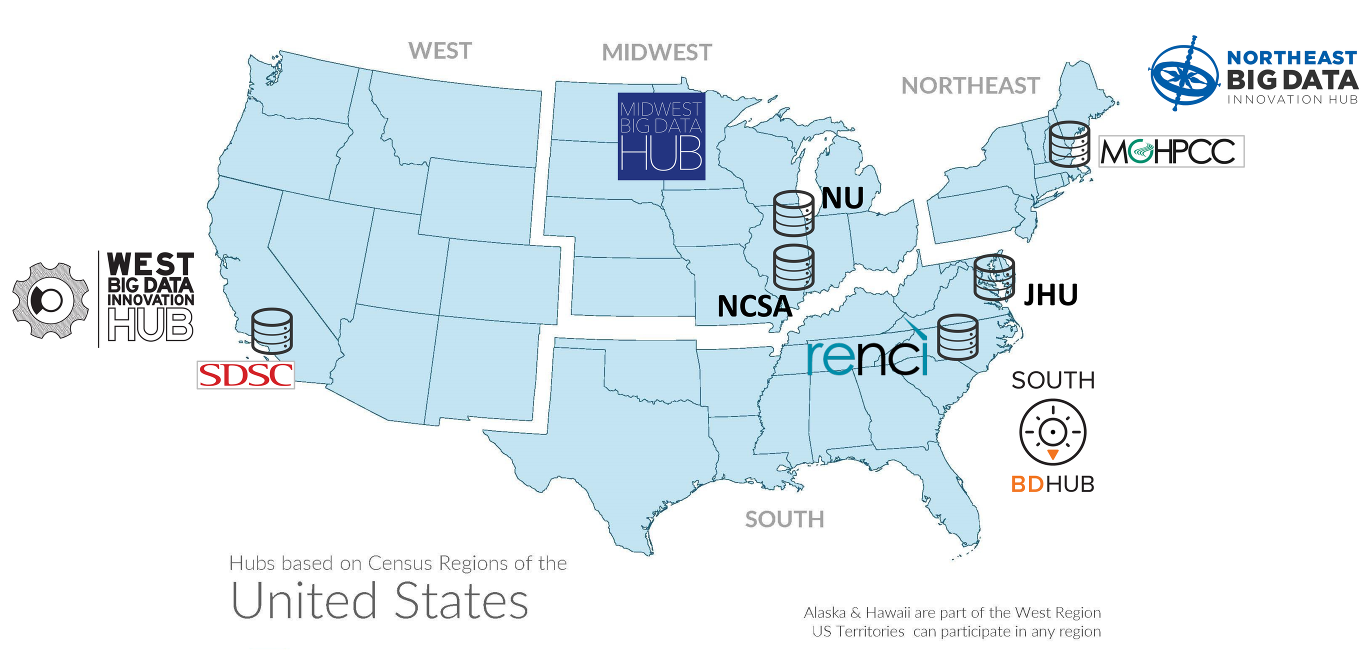
The challenge

- Increasing amounts of scientific data emerging from research projects on all scales is spurring research universities to invest in multi-petabyte (PB) storage systems.[1,4]
- More than 200 US academic institutions have access to high-speed network connectivity for research purposes through NSF CC*NIE awards.[8]
- Data storage and transfer for scientific research remain largely balkanized, without standard requirements and without nation-scale cyberinfrastructure such as XSEDE for computation.[5]

Our goals[6]

- Demonstrate the potential of a distributed storage infrastructure capable of leveraging high speed links to provide a transparent multi-petabyte data storage and access layer.
- Build a scalable substrate composed of storage appliances that are robust and secure, intended to be simple to manage while supporting various data access patterns.
- Enable and enhance science-driven collaborations across universities, and facilitate broad access for actively used data.

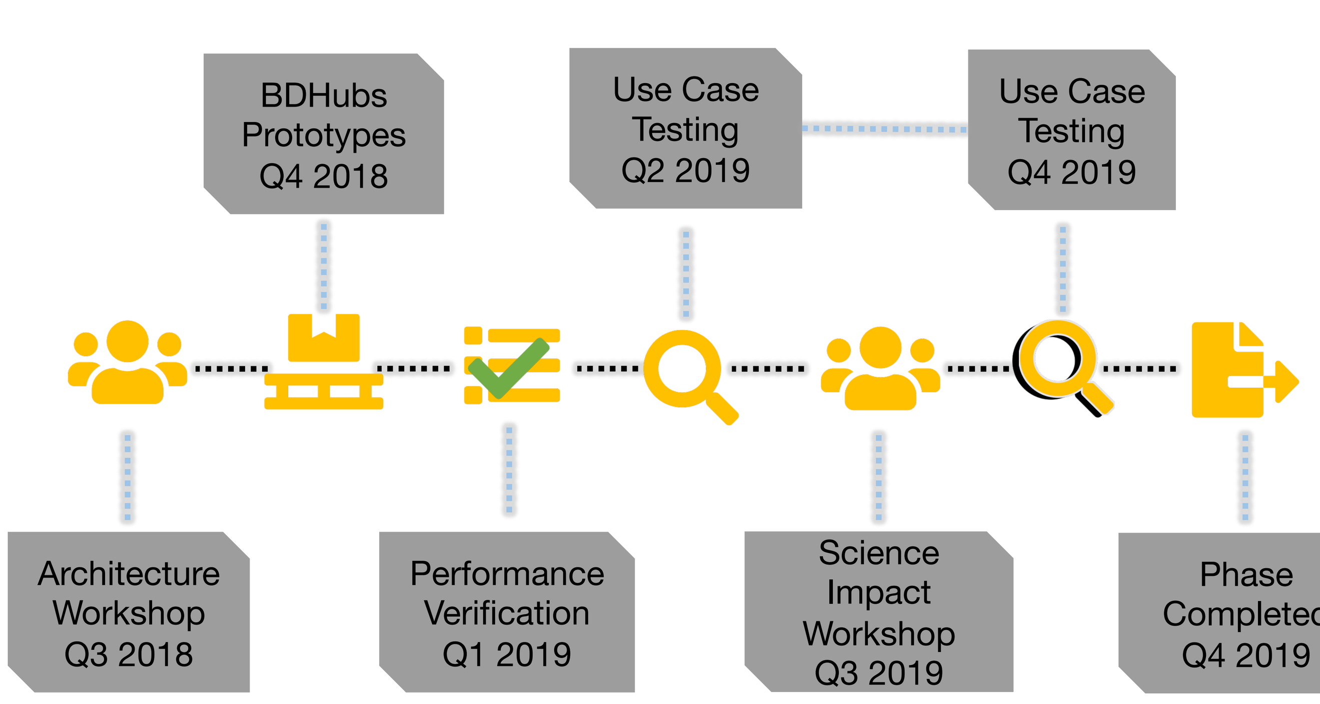
Prototype deployment sites



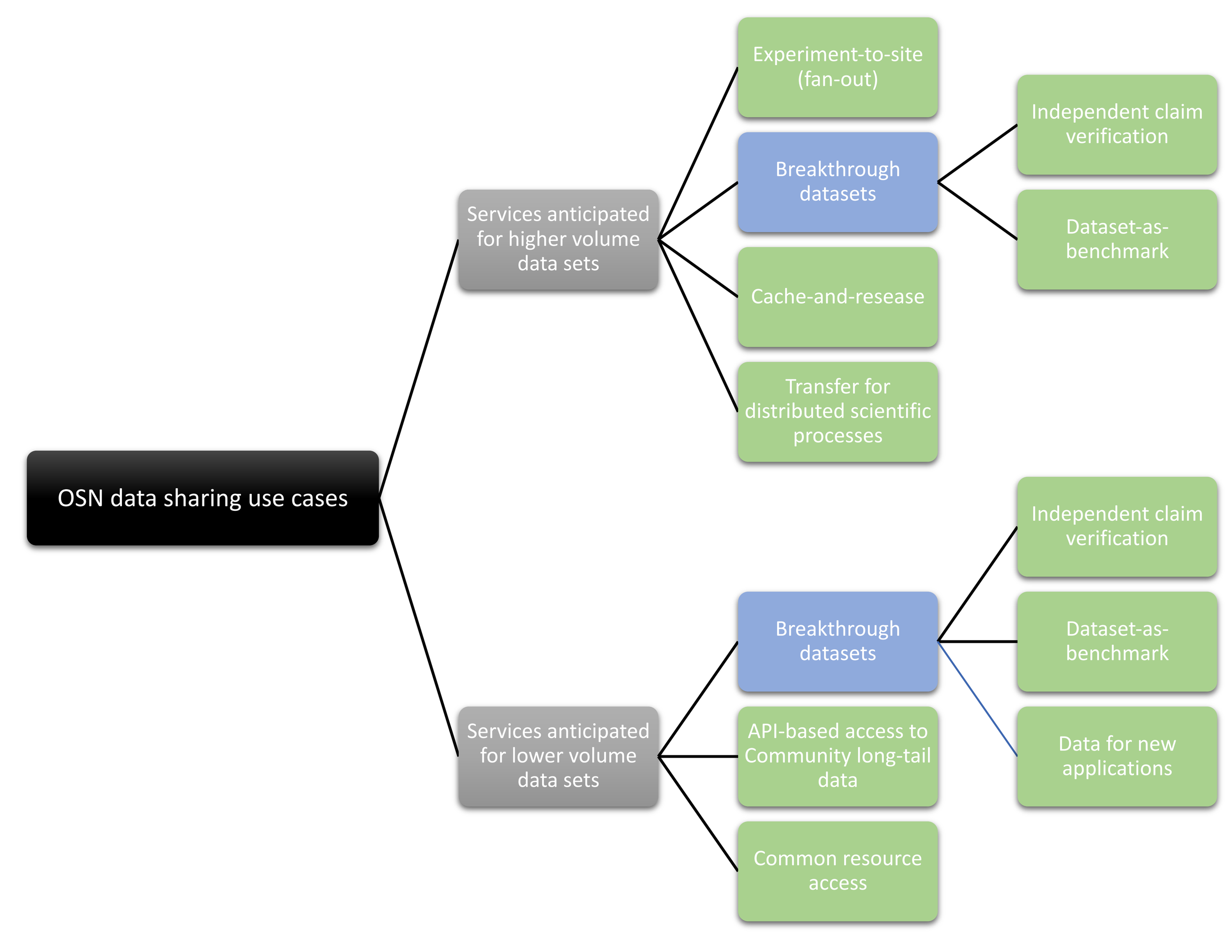
Science use cases for demonstration phase

Project	Research area	Average size of data entities	Total data volume
Connectomics	Neuroscience	10 GB	2 PB
Critical Zone Observatories (CZO)	Earth Sciences	10 MB	50 TB
TerraFusion	Earth Sciences	10 GB	1 PB
Global ocean modeling	Climatology and Oceanography	5 GB	4 PB
HathiTrust Research Center collection	Digital Humanities	200 MB	500 TB
Machine Learning	Neuroscience, Computer Science	10 GB	1 PB
Sloan Digital Sky Survey	Astronomy	15 MB	70 TB
Large Synoptic Survey Telescope (LSST)	Astronomy	2 TB	100 PB
Combined Array for Research in Millimeter Astronomy (CARMA)	Astronomy	50 MB	50 TB
Watershed Models at the Process Scale	Earth Sciences	1 GB	2 TB
Collaborative Gene Matching	Bioinformatics	1 GB	1 PB

Project timeline

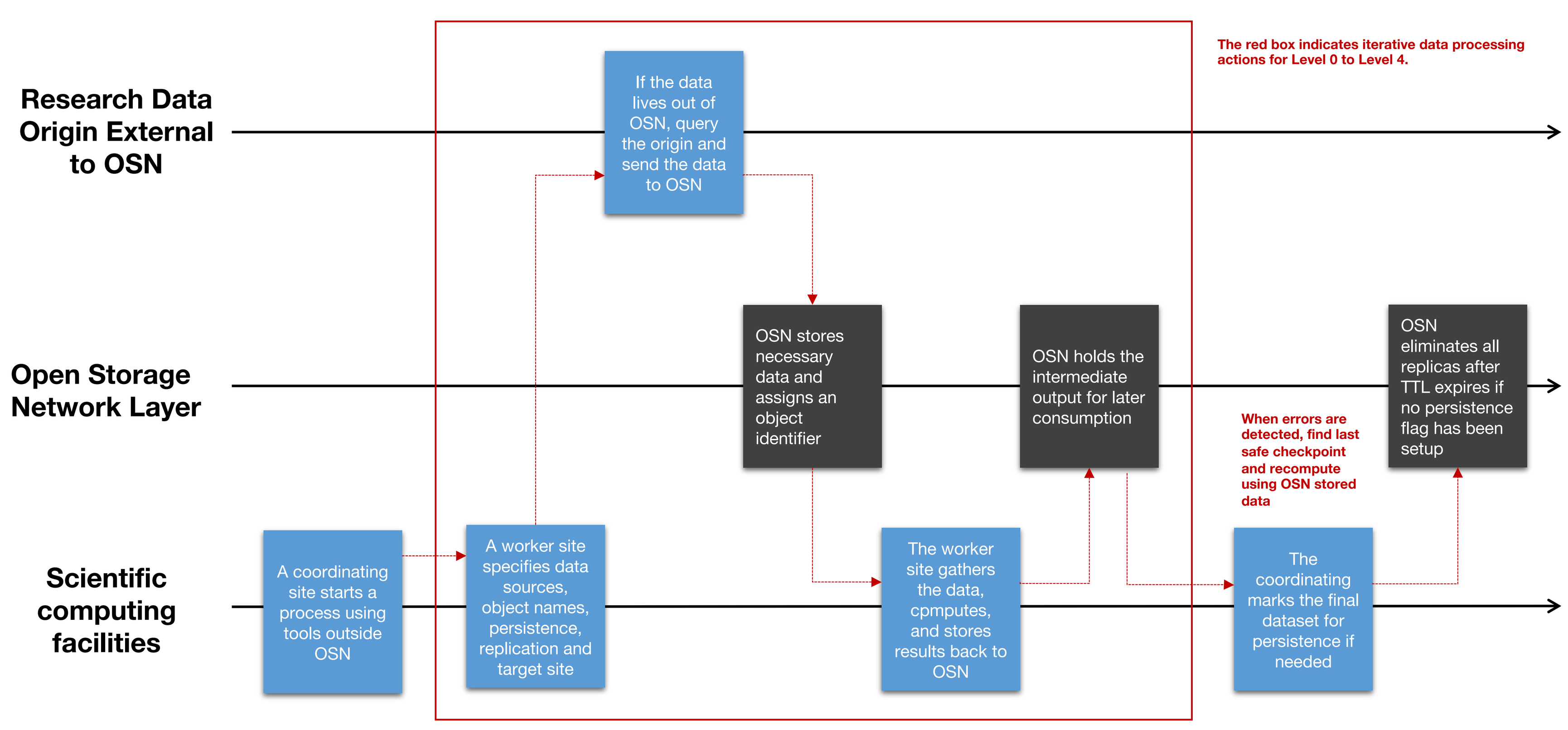


Typology of data storange and transfer use cases



Our use case typology abstracts and generalizes relevant data storage, transport and sharing patterns[7] represented by a wide variety of scientific domains and research exemplars, ranging from large-scale scientific collaborations to long-tail data. The typology was inspired in work performed by Bose & Frew (2005)[2].

OSN service example: transferring data to support complex, distributed scientific computing[3]



Anticipated applications of Midwest use cases

Project	Storage problem being solved	Applicable typology classes
CZO	Provide storage space and access to CZO datasets and community-generated data	Community long-tail data
TerraFusion	Transport datasets across the US at high speed, obtain data slices with high probability of reutilization	Experiment-to-site; Cache-and-release
HathiTrust Research Center Extracted Feature Dataset	Provide storage space and access to the HTRC dataset and further community-generated derivatives	Common resource access
Machine Learning Data	Availability of well-curated datasets for ML R+D and education	Common resource access; Dataset-as-benchmark
LSST	Transport datasets across the US at high speed, obtain data slices with high probability of reutilization, facilitate inter-site data processing	Experiment-to-site; Cache-and-release; Transfer for distributed processes
CARMA	Transport datasets across the US at high speed, obtain data slices with high probability of reutilization	Experiment-to-site; Cache-and-release

Next steps

- Performance testing and tuning of storage pod network across participating institutions
- Implementation of the software and service architectures for the OSN
- Engage science use case groups and prepare for moving data to OSN

References

1. **Biffard, B., Valenzuela, M., Conley, P., MacArthur, M., Tredger, S., Guillemot, E., & Pirenne, B. (2016).** Oceans 2.0: Interactive tools for the Visualization of Multi-dimensional Ocean Sensor Data. In AGU Fall Meeting Abstracts.

2. **Bose, R., & Frew, J. (2005).** Lineage retrieval for scientific data processing: a survey. ACM Computing Surveys (CSUR), 37(1), 1-28.

3. **Deelman, E., & Chervenak, A. (2008).** Data management challenges of data-intensive scientific workflows. In 2008 Eighth IEEE International Symposium on Cluster Computing and the Grid (CCGRID) (pp. 687-692). IEEE.

4. **Kiran, A., Gupta, P. K., Jha, A. K., & Saran, S. (2018).** Online Geoprocessing Using Multi-Dimensional Gridded Data. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 45, 29-36.

5. **Kowalczyk, S., & Shankar, K. (2011).** Data sharing in the sciences. Annual review of information science and technology, 45(1), 247-294.

6. **Open Storage Network.** National Science Foundation. Available at: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1747493

7. **Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., & Nolan, G. P. (2010).** Computational solutions to large-scale data management and analysis. Nature reviews genetics, 11(9), 647.

8. **Thompson, K. (2012).** Campus Cyberinfrastructure–Network Infrastructure and Engineering (CC-NIE). National Science Foundation, December 2012.

Funded by NSF (IIS 1747552, IIS 1747493, IIS 1747507, IIS 1747490, IIS 1747483) and the Schmidt Futures Foundation.

NSF CC*NIE: https://bit.ly/2qhQmwe
XSEDE: https://www.xsede.org
Connectomics: https://lichtmanlab.fas.harvard.edu
CZO: http://criticalzone.org/national/
Terra Fusion: https://go.nasa.gov/2ql6NIm
Global ocean modeling: https://bit.ly/2Oe3jRi

HathiTrust Research Center: https://wiki.htrc.illinois.edu
Machine learning: http://chemimage.illinois.edu
SDSS: https://www.sdss.org
LSST: https://www.lsstcorporation.org
CARMA: http://carma-server.ncsa.uiuc.edu:8181
Watershed models: https://www.hydroshare.org