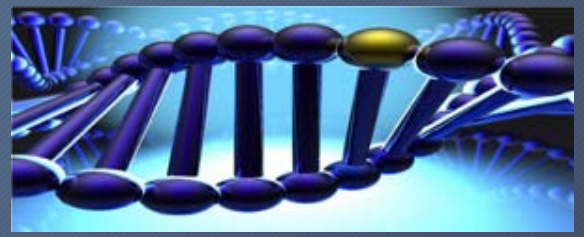
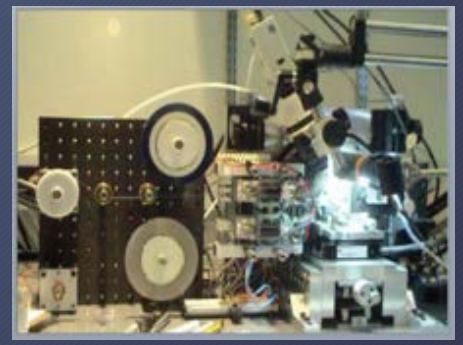
www.openstoragenetwork.org

The Data Deluge

Capture | Store | Share | Analyze





US Research Cyber-Infrastructure Today

Computation	Networking	Storage
Local, Regional and National Resources	*Over 200 universities with 40/100Gb Connectivity*	*Largely Balkanized*
Standardized	*Standardized*	*Many Standards to Choose From*

The Open Storage Network

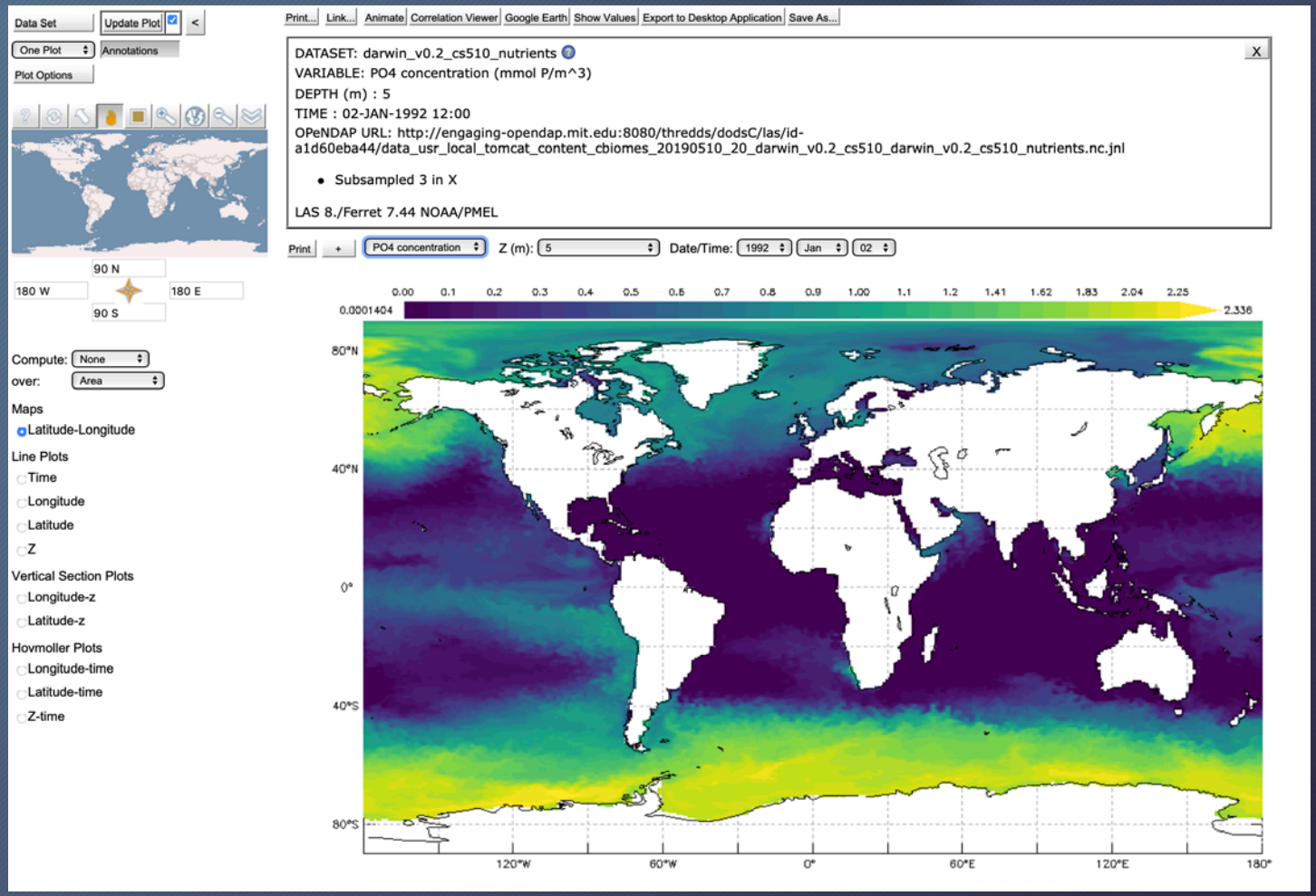
- National resource for sharing open scientific data
- Distributed Infrastructure and Governance

Six Deployment Sites

- Johns Hopkins University
- MGHPCC
- San Diego Supercomputing Center
- NCSA at University of Illinois
- RENCI UNC Chapel Hill
- Northwestern University (Starlight)

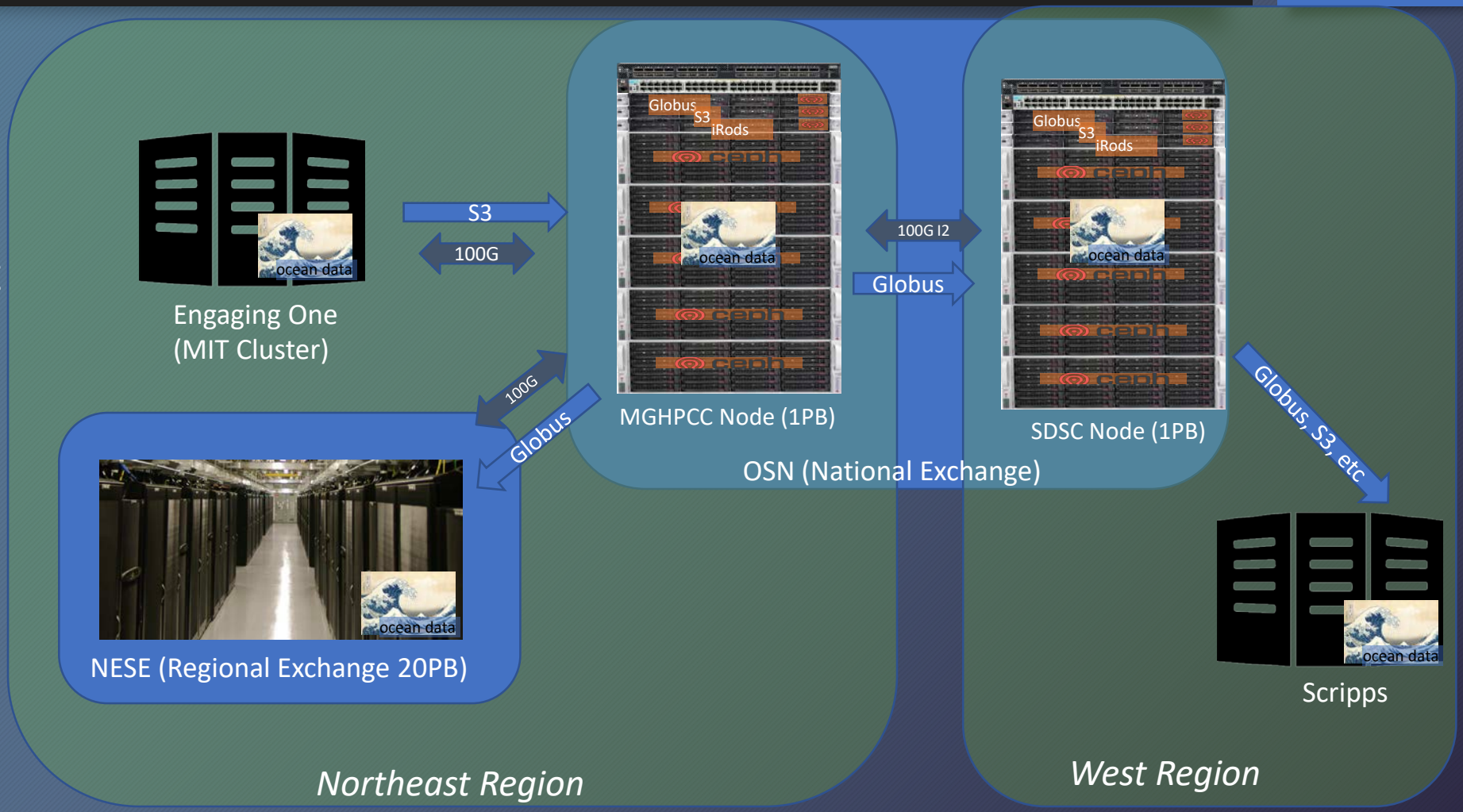
Science Use Case demonstration
MIT Resident Ocean Data

- Datasets are from global estimate of time-dependent marine microbial dynamics 1992-2016.
- As observations and compute improve we can start to do variety of holistic analysis (e.g. clustering; other structure) across entire corpus.
- Can serve slices of data, but for broader use want to use OSN to allow easy transfer of full data (PB+) for unrestricted analysis/study.



Serving Slices of Ocean Data on OSN

- Data stored on NFS storage on Engaging 1 Cluster (NFS)
- S3 transfer to MGHPCC OSN Site (rclone)
- Globus Distribution to Northeast Storage Exchange for use in NE (ceph)
- Globus Distribution to SDSC OSN site for use at Scripps



Near Term Users

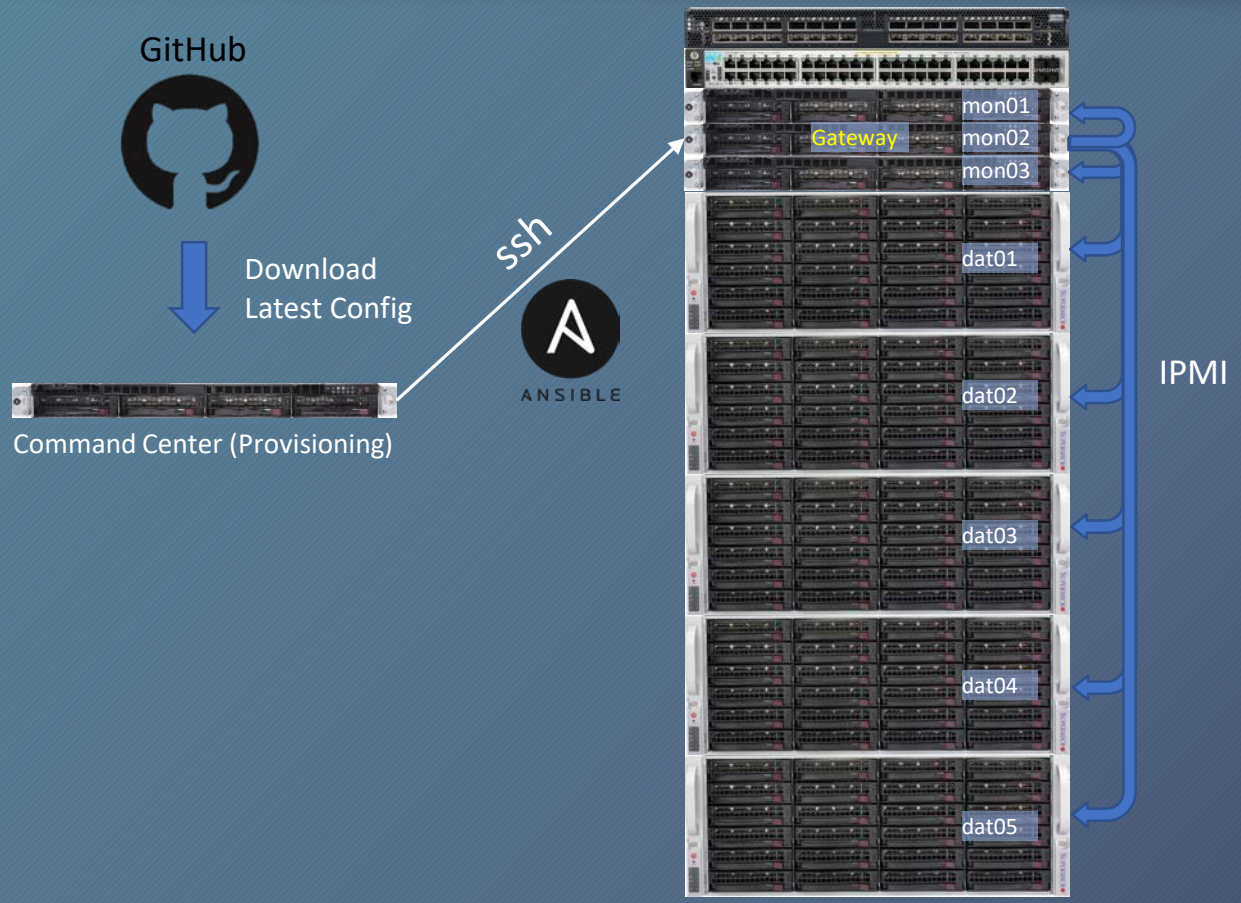
Project	Average size of data entities	Total data volume	Storage problem being solved	Use case
Critical Zone Observatories	10 MB	50 TB	Provide storage space and access to CZO datasets and community-generated data	Community long-tail data
TerraFusion	10 GB	1 PB	Transport datasets across the US at high speed, obtain data slices with high probability of reutilization	Experiment-to-site, Slice-and-compute
HathiTrust Research Center collection	200 MB	500 TB	Provide storage space and access to the HTRC dataset and further community-generated derivatives	Common resource access
Machine Learning	10 GB	1 PB	Make available a well-curated dataset for testing machine learning algorithms	Dataset-as-benchmark
Large Synoptic Survey Telescope	2 TB	100 PB	Transport datasets across the US at high speed, obtain data slices with high probability of reutilization, facilitate inter-site data processing	Experiment-to-site, Slice-and-compute, Workflow staging space
Combined Array for Research in Millimeter Astronomy	50 MB	50 TB	Transport datasets across the US at high speed, obtain data slices with high probability of reutilization	Experiment-to-site, Slice-and-compute

How to Access Data Sets Stored in the OSN

- S3
- rclone
- iRods
- Globus
- Clouder
- Cyberduck
- ownCloud
- …
- Experimental Overlays

How We Deploy

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

Monitoring and Management - *OSN 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 the operations team

How to Join

- Active Data Sets Welcome
  - Load and go
- New Sites Welcome
  - Get a pod
  - Join the team