OSiRIS

Distributed Ceph and Software Defined Networking for Multi-Institutional Research

NSF Award 1541335: CC*DNI DIBBs: Multi-Institutional Open Storage Research InfraStructure (MI-OSiRIS)
Shawn McKee (Principal Investigator), Douglas Swany (Co-Principal Investigator),
Patrick Gossman (Co-Principal Investigator), Kenneth Merz (Co-Principal Investigator)

ATLAS Event Service

ATLAS jobs will use the OSiRIS Rados Gateway (S3) to read/write single events.

This approach, different than the usual ATLAS data transfer model, allows easier leveraging of computing resources that may not be specifically configured for ATLAS or may be transient.

The next step with ATLAS is to test operations at scale with hundreds of thousands of ATLAS jobs using OSiRIS as an object store. The nature of the Ceph RGW means we should be able to spin up more gateways as needed to support ATLAS and other groups.
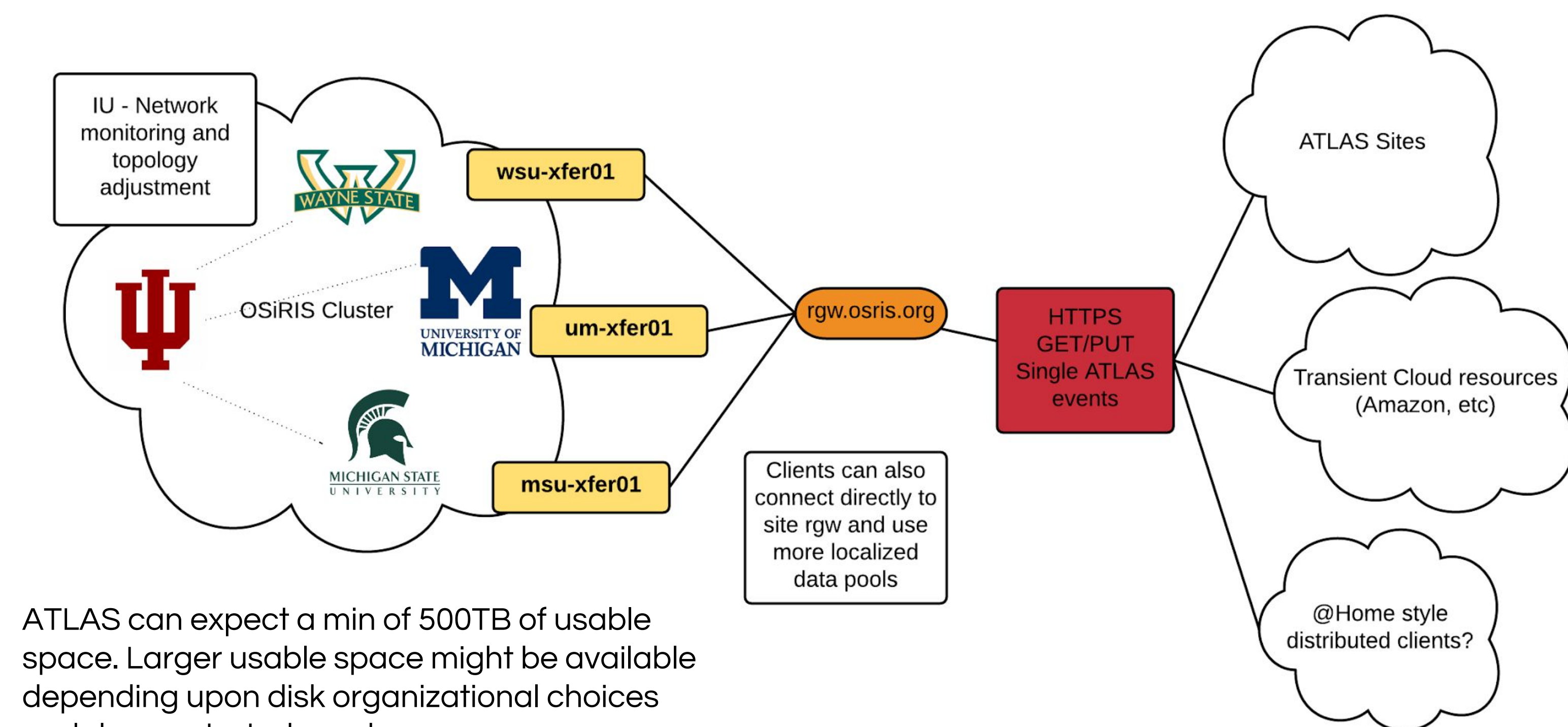
ATLAS can expect a min of 500TB of usable space. Larger usable space might be available depending upon disk organizational choices and demonstrated need.

Watching how OSiRIS scales as we ramp up ATLAS will be important.

We have detailed monitoring of our RGW throughput and ops at storage level, network level, or Ceph service level. Example at right of requests/second to one RGW node during ATLAS proof-of-concept tests.
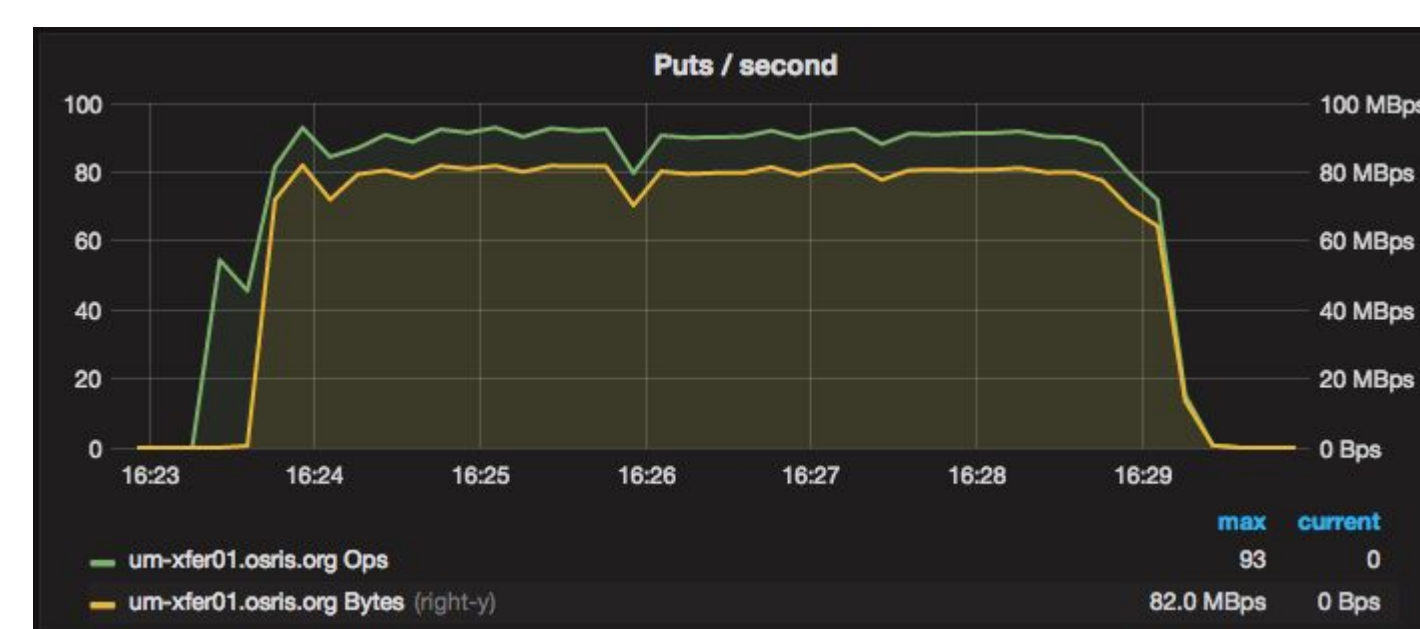
Ceph AAA Enhancements

OSiRIS Authentication, Authorization, and Accounting services are made up of three key components to bridge Ceph and institutional federations.

- stpd, short for St. Peter Daemon, is a Ceph Gatekeeper Daemon for OSiRIS resources that runs on the CephMon nodes
- oakd, short for OSiRIS Access Key Daemon, used to create, store, and retrieve access keys which grant access to specific resources
- mount.osiris, A filesystem utility that facilitates mounting of OSiRIS resources via web interfaces
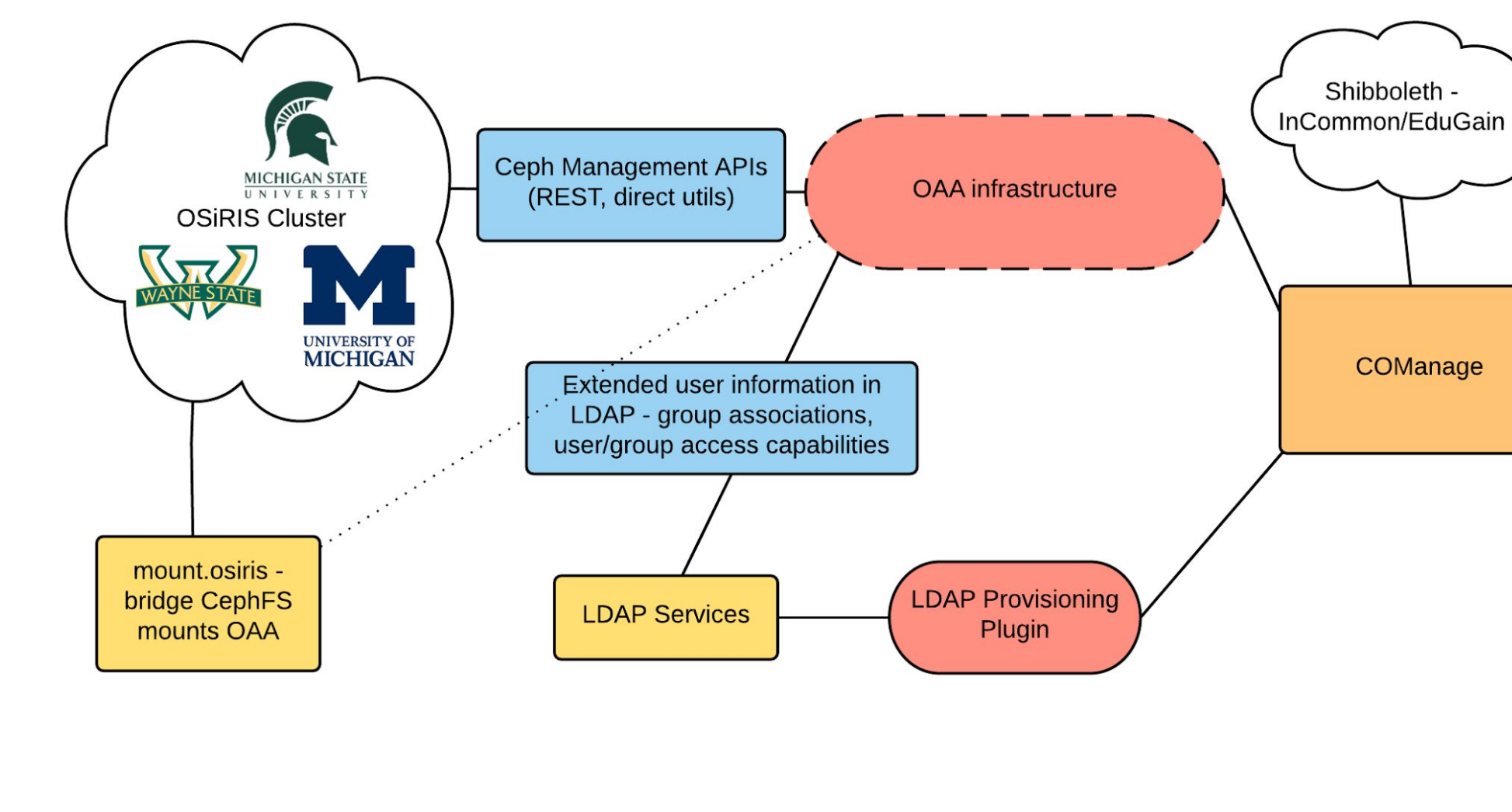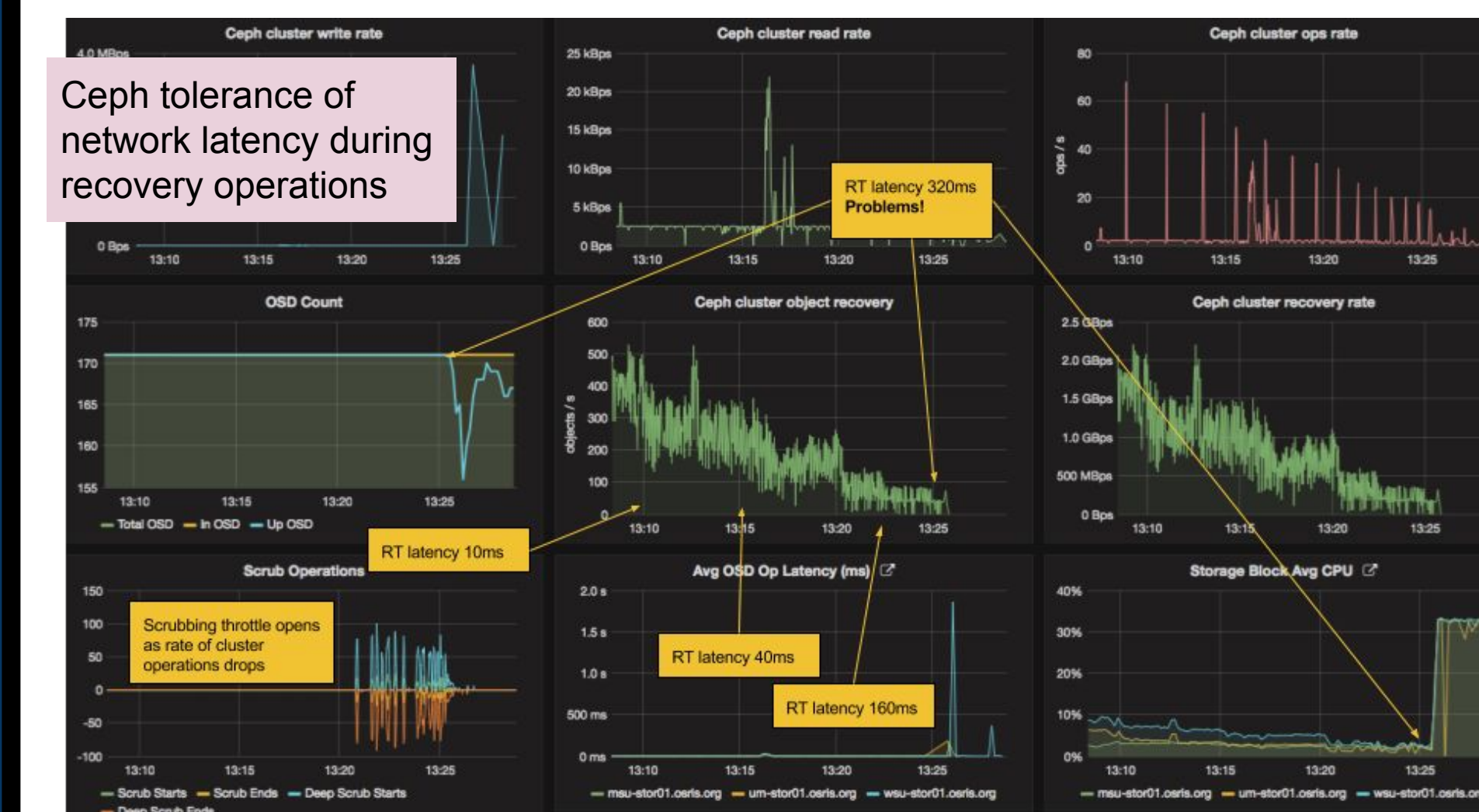

Ceph Capabilities

One of our grant responsibilities is exploration of the limits of Ceph and our architecture. Resilience to network latency is a key determinant in how far we can scale geographically and whether projects like OSiRIS can be successful using Ceph as we are.

At the recent SC16 conference we deployed and operated a 4th OSiRIS site on the floor at Salt Lake City. Not only were we successful deploying the site repeatedly as a demonstration but we did so without disrupting ongoing data transfers in our other demos.

Supercomputing was a real-world verification of simulated latency testing we did with several Ceph components. The annotated dashboard below shows the cluster's capability to continue background data replication (recovery) even as we ramp up simulated latency on a storage block (60 OSD effected).

Ceph tolerance of network latency during recovery operations

NMAL: Advanced Network Management

OSiRIS Network Management Abstraction Layer
Advanced network monitoring and control services

NMAL extends perfSONAR capabilities to include automated network topology discovery and tracking in the Unified Network Information Service (UNIS), and incorporate that into overall operations of the OSiRIS infrastructure.

Potential integrations include SDN control at the hardware switch and host vswitch level, Ceph cluster data map control, and data transfer node routing control.
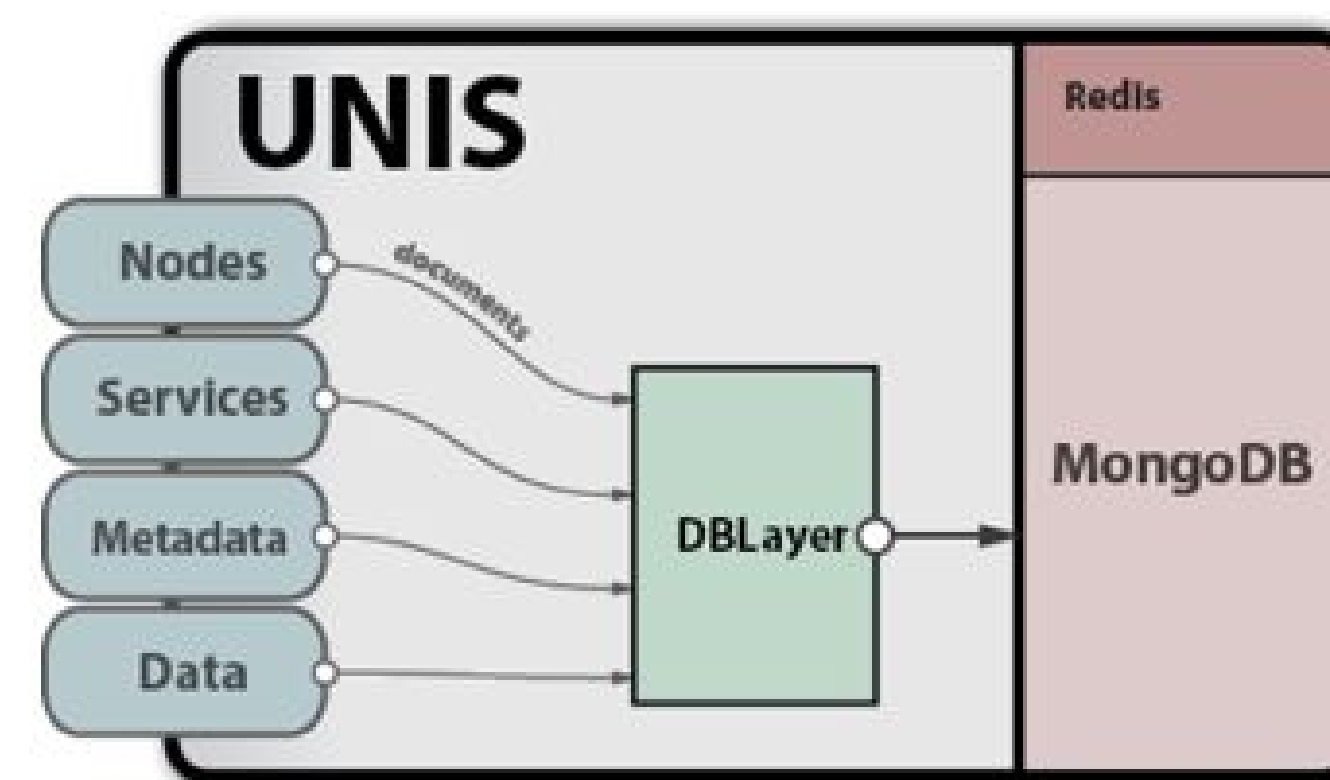

Project-wide we've deployed perfSONAR Periscope agents capable of reporting a variety of network and host information back to the UNIS datastore. The Periscope agent extensions to perfSonar not only report perfSonar host measurements to UNIS but also open up new APIs to leverage those measurements.

For example, we can use these APIs to pull scheduled perfSonar test results into our instances of Influxdb for visualization over time.
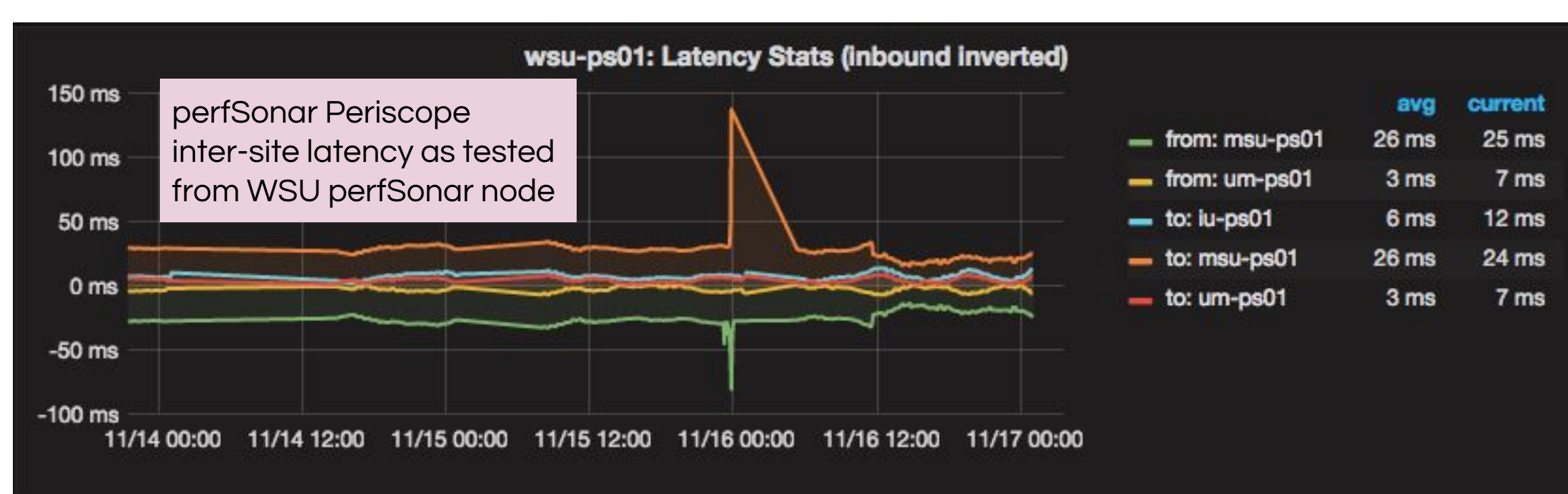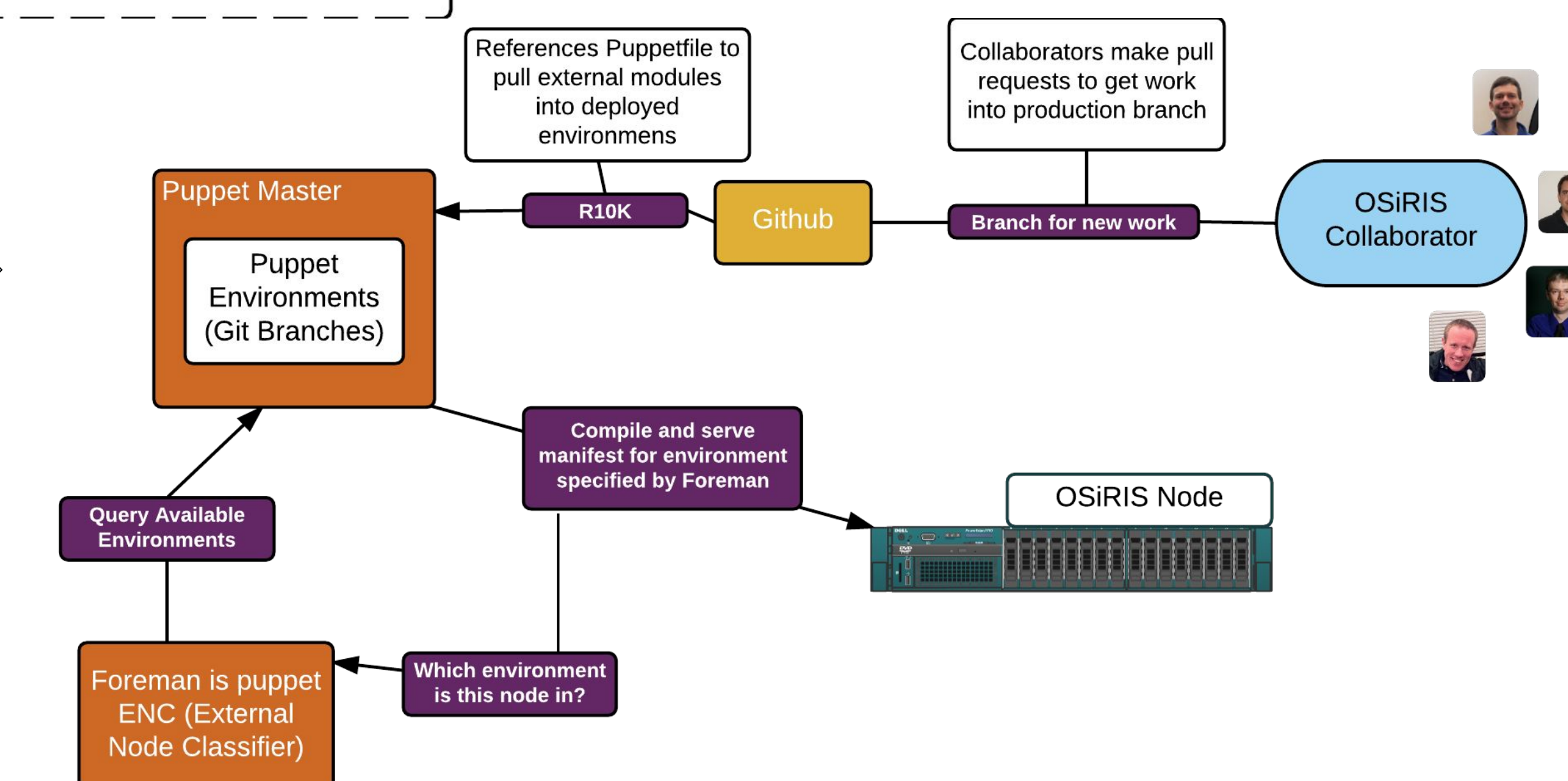
Multi-Institute Deployment

Using Foreman and Puppet we've successfully built an infrastructure for efficiently deploying new OSiRIS sites and managing configuration across the project.
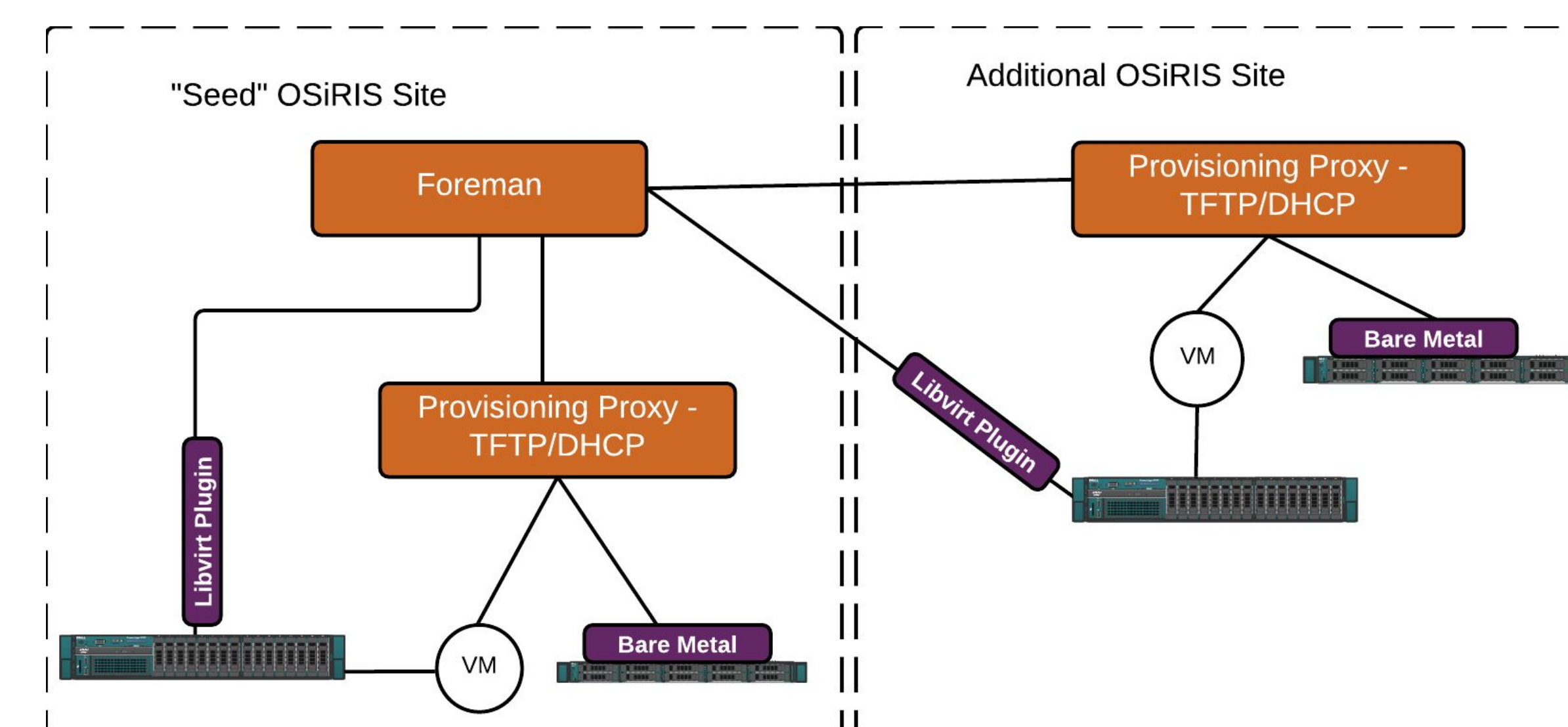
From a Foreman instance we can control provisioning 'smart proxies' at each of our sites. Initial site setup requires only getting this proxy running from a pre-configured template we provide.

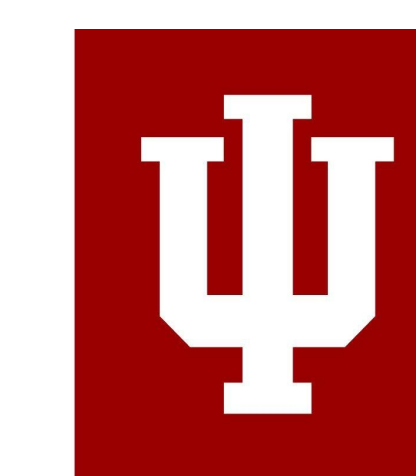Collaborators at all of our institutions can deploy and test new changes in private environments without disrupting regular OSiRIS production.