OSiRIS

Distributed Ceph and Software Defined Networking for Multi-Institutional Research

Benjeman Meekhof
University of Michigan
Advanced Research Computing – Technology Services
May 11, 2016

- About OSiRIS
  - Project Team
  - Overview
  - Challenges
- Technology
  - Ceph
  - Networking/NMAL
  - Monitoring
  - Orchestration
- Status Today
  - Hardware Deployment
  - Test and Production Ceph clusters
  - Baseline Metrics
- Next Steps

OSiRIS Summary

We proposed to design and deploy MI-OSiRIS (Multi-Institutional Open Storage Research Infrastructure) as a pilot project to evaluate a software-defined storage infrastructure for our primary Michigan research universities.

Our goal is to provide transparent, high-performance access to the same storage infrastructure from well-connected locations on any of our campuses.

By providing a single data infrastructure that supports computational access "in-place" we can meet many of the data-intensive and collaboration challenges faced by our research communities and enable them to easily undertake research collaborations beyond the border of their own universities.

OSiRIS Team

OSiRIS is composed of scientists, computer engineers and technicians, network and storage researchers and information science professionals from **University of Michigan**, **Michigan State University**, **Wayne State University**, and **Indiana University** (focusing on SDN and net-topology)

We have a wide-range of science stakeholders who have data collaboration and data analysis challenges to address within, between and beyond our campuses:

High-energy physics, High-Resolution Ocean Modeling, Degenerative Diseases, Biostatics and Bioinformatics, Population Studies, Genomics, Statistical Genetics and Aquatic Bio-Geochemistry

Multi Institutional Data Challenges

Scientists working with large amounts of data face many obstacles in conducting their research

Typically the workflow needed to get data to where they can process it becomes a substantial burden

The problem intensifies when adding in collaboration across their institution or especially beyond their institution

Institutions have sometimes responded to this challenge by constructing specialized and expensive infrastructures to support specific science domain needs

OSiRIS is Better

Scientists get customized, optimized data interfaces for their multi-institutional data needs

Network topology and perfSONAR-based monitoring components ensure the distributed system can optimize its use of the network for performance and resiliency

Ceph provides seamless rebalancing and expansion of the storage

A single, scalable infrastructure is much easier to build and maintain

Allows universities to reduce cost via economies-of–scale while better meeting the research needs of their campus

Eliminates isolated science data silos on campus:

- Data sharing, archiving, security and life-cycle management are feasible to implement and maintain with a single distributed service.
- Data infrastructure view for each research domain can be optimized for performance and resiliency.

Project Challenges

Deploying and managing a fault tolerant multi-site infrastructure

Resource management and optimization to maintain a sufficient quality of service for all stake-holders

Enabling the gathering and use of metadata to support data lifecycle management

Research domain customization using CEPH API and/or additional services

Authorization which integrates with existing campus systems

Authentication and Authorization

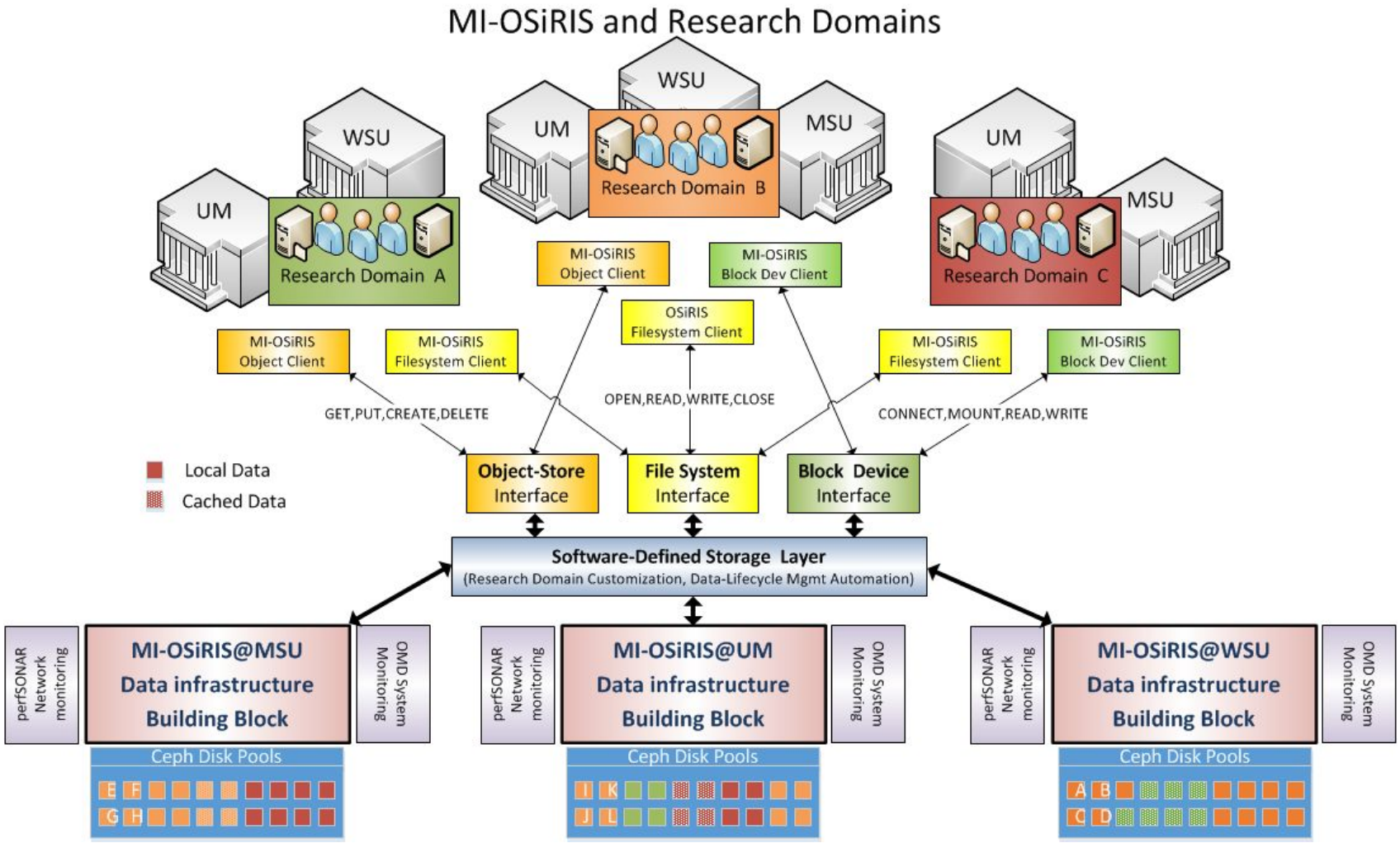
We are working with Von Welch and Jim Basney from the Center for Trusted Scientific CyberInfrastructure to find the best way forward: http://trustedci.org/who-we-are/

Using InCommon Federation attributes is not necessarily straightforward

- There are widely varying levels of InCommon participation and attribute release
- OSiRIS is registered as an InCommon Research and Scholarship entity. Participating sites release more attributes by default to registered entities
- Often have to contact institute identity teams to request needed attributes

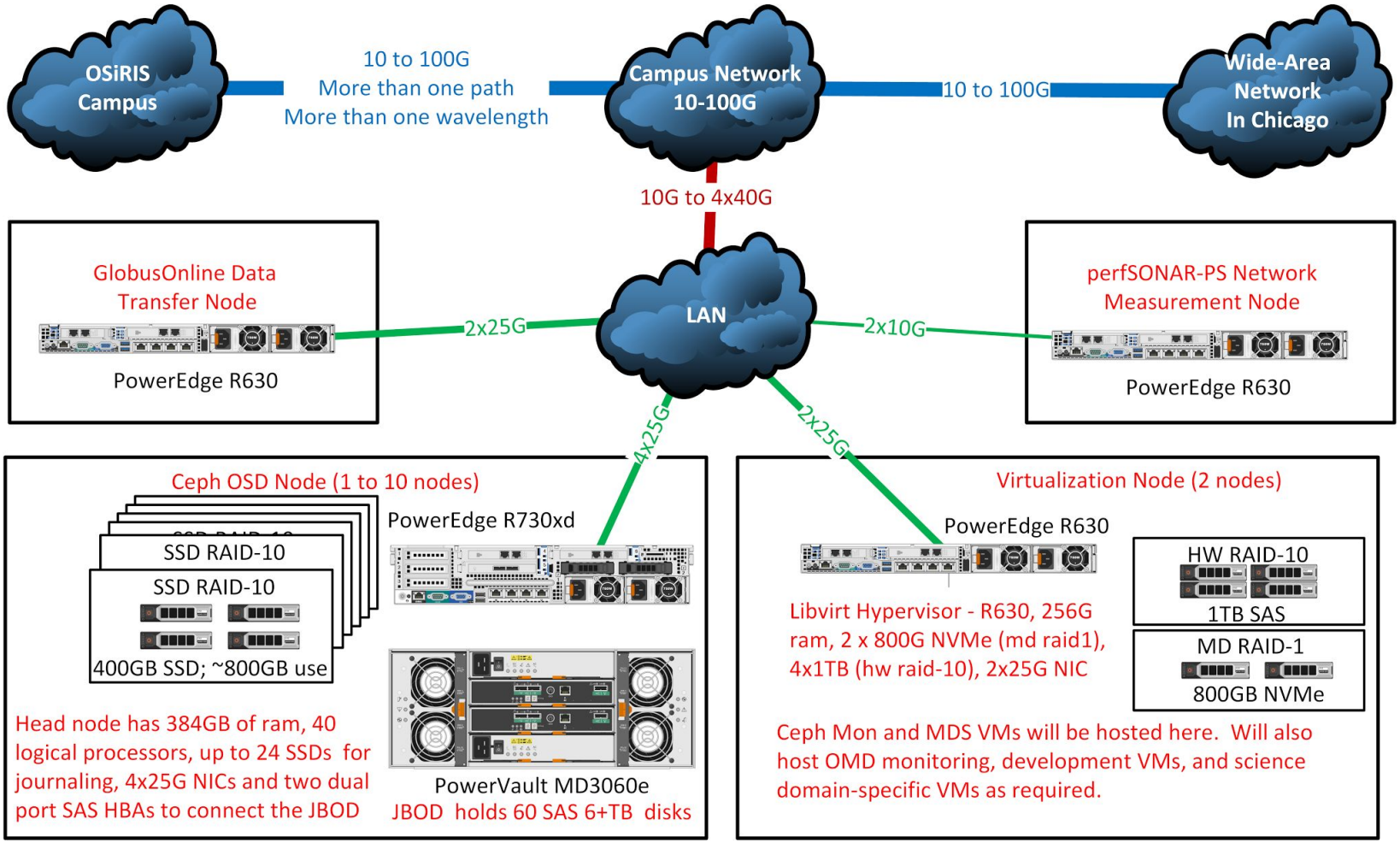
Augmenting Ceph for fine grained authorization from institutional and VO attributes is one of our major challenges

Logical View



Site View

OSiRIS Data Infrastructure Building Block



Ceph in OSiRIS

Ceph gives us a robust open source platform to host our multi-institutional science data

- Self-healing and self-managing
- Multiple data interfaces
- Rapid development supported by RedHat

Able to tune components to best meet specific needs

Software defined storage gives us more options for data lifecycle management automation

Sophisticated allocation mapping (CRUSH) to isolate, customize, optimize by science use case

Ceph overview:
https://umich.app.box.com/s/f8ftr82smlbuf5x8r256hay7660soafk

Deploying Ceph

Our Ceph cluster components are all deployed with puppet

We forked from Openstack Puppet module

- https://github.com/MI-OSiRIS/puppet-ceph
- needed support for provisioning multiple clusters on same hardware or clients with multiple cluster config
- Mon service init needed modification for > Infernalis + systemd and non-default cluster names
- Sufficiently re-organized that we're not following (all of) upstream anymore

Ceph keys/keyrings are deployed by puppet, secrets are kept in hiera-eyaml

Puppet prepares/activates OSD from resources in hiera (done as needed by setting trigger fact before run)

Deploying additional/replacement Mon, OSD, etc can be done quickly and consistently

Issues Deploying Ceph

Wanted to use software (mdraid) RAID-1 devices for Ceph journal - 2 x 400GB NVMe supporting 30 OSD journal per md

- Udev rule supplied with Ceph to create /dev/disk/by-partuuid/ ignored md devices - had to modify
  - Is someone saying that md raid1 for journal is a bad idea? Maybe!

As installed, Ceph systemd units for OSD do not support multiple cluster on same host.

- Can set "CLUSTER=name" in sysconfig/ceph to have one or the other work
- Copied test-osd@.service from ceph-osd@.service and set default cluster, then link to separate systemd target test-osd.target

Software Defined Networking

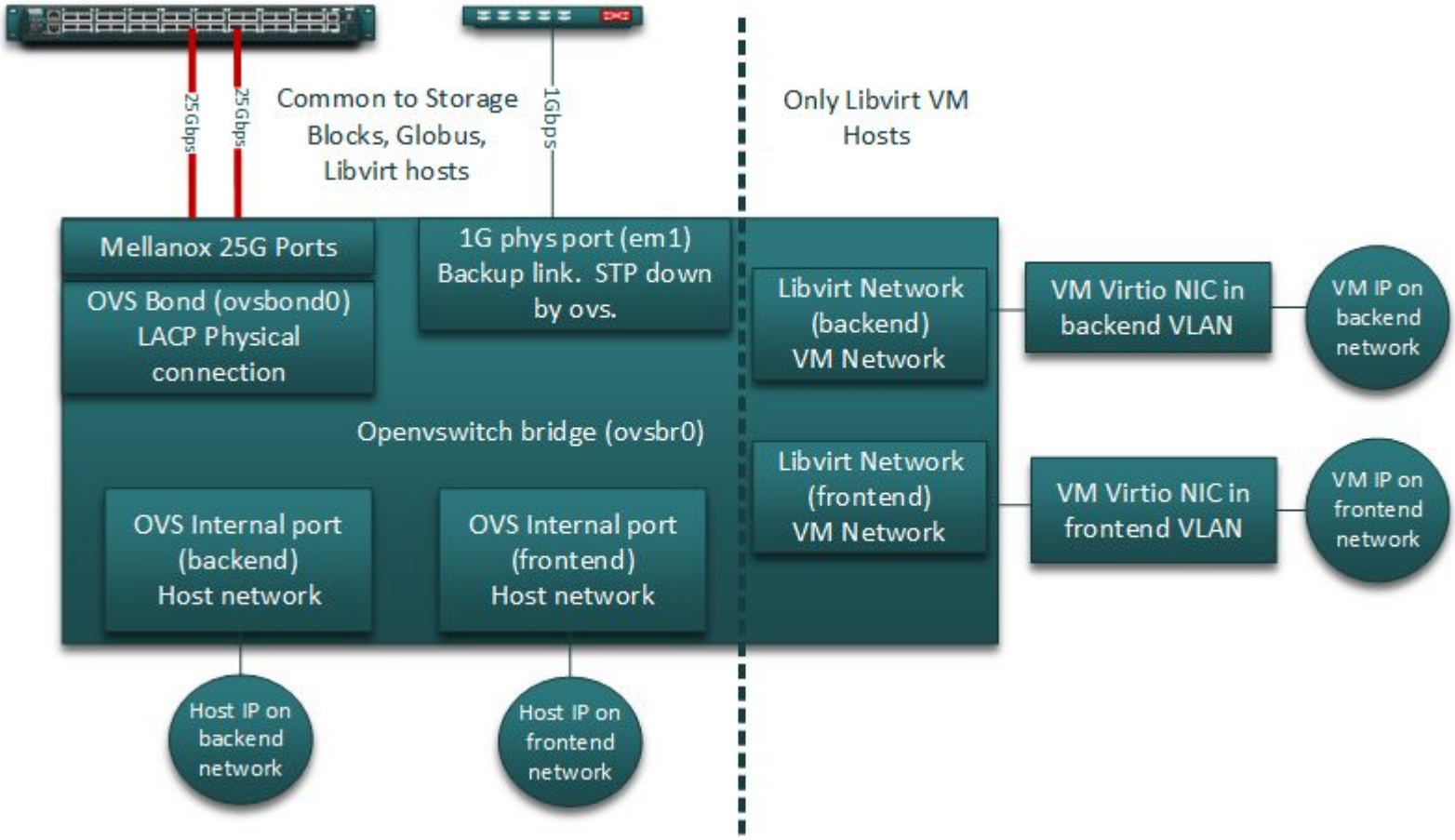
Software defined networking (SDN) changes traditional networking by decoupling the system that makes decisions about where traffic is sent (the control plane) from the underlying systems that forward traffic to the selected destination (the data plane).

Using SDN we can centralize the control plane and programatically update how the network behaves to meet our goals.

For OSiRIS the network will be a critical component, tying our multi-institutional users to our distributed storage components.

SDN - Open vSwitch

OSiRIS storage blocks, transfer gateways (S3, globus), and virtualization hosts incorporate Open vSwitch to allow fine-grained control dynamic network flows and integration with OpenFlow controllers



NMAL

The OSiRIS Network Management Abstraction Layer is a key part of the project with several important focuses:

Capturing site topology and routing information in UNIS from multiple sources: SNMP, LLDP, sflow, SDN controllers, and existing topology and looking glass services.

- Existing UNIS encoder is being extended to incorporate these new data sources.

Packaging and deploying conflict-free measurement scheduler (HELM) along with measurement agents (BLiPP).

Converge on common scheduled measurement architecture with existing perfSONAR mesh configurations.

Correlate long-term performance measurements with passive metrics collected via check_mk infrastructure.

Integrating Shibboleth to provide authentication/authorization for measurement and topology services. This includes extending existing perfSONAR toolkit components in addition to Periscope.

Defining best-practices for SDN controller and reactive agent deployments within OSiRIS.

Network Monitoring

Because networks underlie distributed cyberinfrastructure, monitoring their behavior is very important

The research and education networks have developed perfSONAR as a extensible infrastructure to measure and debug networks (http://www.perfsonar.net)

The CC*DNI DIBBs program recognized this and required the incorporation of perfSONAR as part of any proposal

For OSiRIS, we were well positioned since one of our PIs Shawn McKee leads the worldwide perfSONAR deployment effort for the LHC community: https://twiki.cern.ch/twiki/bin/view/LCG/NetworkTransferMetrics

We intend to extend perfSONAR to enable the discovery of all network paths that exist between instances

SDN can then be used to optimize how those paths are used for OSiRIS

BLiPP/UNIS

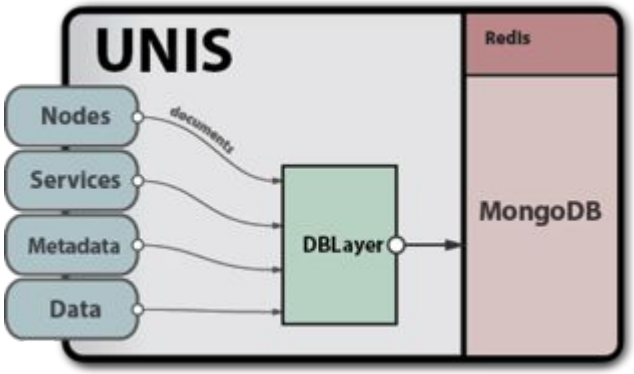
The monitoring and topology discovery components being worked on by Indiana University/CREST are key parts of OSiRIS NMAL SDN

UNIS Topology and Measurement Store

- Exposes a RESTful interface for information necessary to perform data logistics
  - Measurements from BLiPP
  - Network topology inferred through various agents
- Provides subscription endpoints for event-driven clients

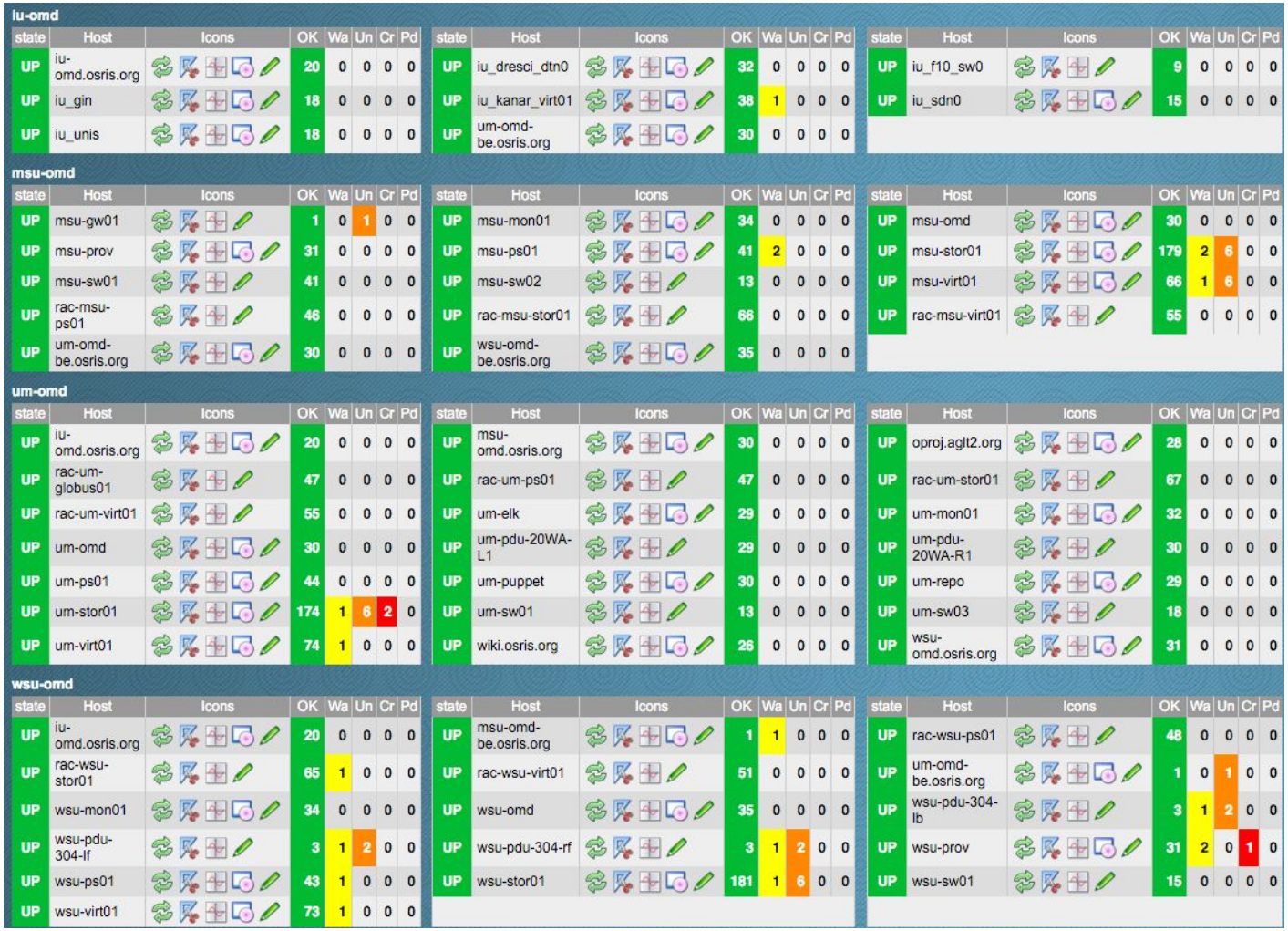
Basic Lightweight Periscope Probe (BLiPP)

- Distributed probe agent system
- BLiPP agents execute measurement tasks received from UNIS and report back results for further analysis.
- BLiPP agents may reside in both the end hosts (monitoring end-to-end network status) and dedicated diagnose hosts inside networks



Monitoring with Check_mk

Each site has an instance of Check_mk referencing the other instances for single dashboard status and centralized alerting



iu-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu-omd.osris.org		20	0	0	0	0
UP	iu_gin		18	0	0	0	0
UP	iu_unis		18	0	0	0	0
UP	iu_dresci_dtn0		32	0	0	0	0
UP	iu_kanar_virt01		38	1	0	0	0
UP	um-omd-be.osris.org		30	0	0	0	0
UP	iu_f10_sw0		9	0	0	0	0
UP	iu_sdn0		15	0	0	0	0

msu-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	msu-gw01		1	0	1	0	0
UP	msu-prov		31	0	0	0	0
UP	msu-sw01		41	0	0	0	0
UP	rac-msu-ps01		46	0	0	0	0
UP	um-omd-be.osris.org		30	0	0	0	0
UP	msu-mon01		34	0	0	0	0
UP	msu-ps01		41	2	0	0	0
UP	msu-sw02		13	0	0	0	0
UP	rac-msu-stor01		66	0	0	0	0
UP	wsu-omd-be.osris.org		35	0	0	0	0
UP	msu-omd		30	0	0	0	0
UP	msu-stor01		179	2	6	0	0
UP	msu-virt01		66	1	6	0	0
UP	rac-msu-virt01		55	0	0	0	0

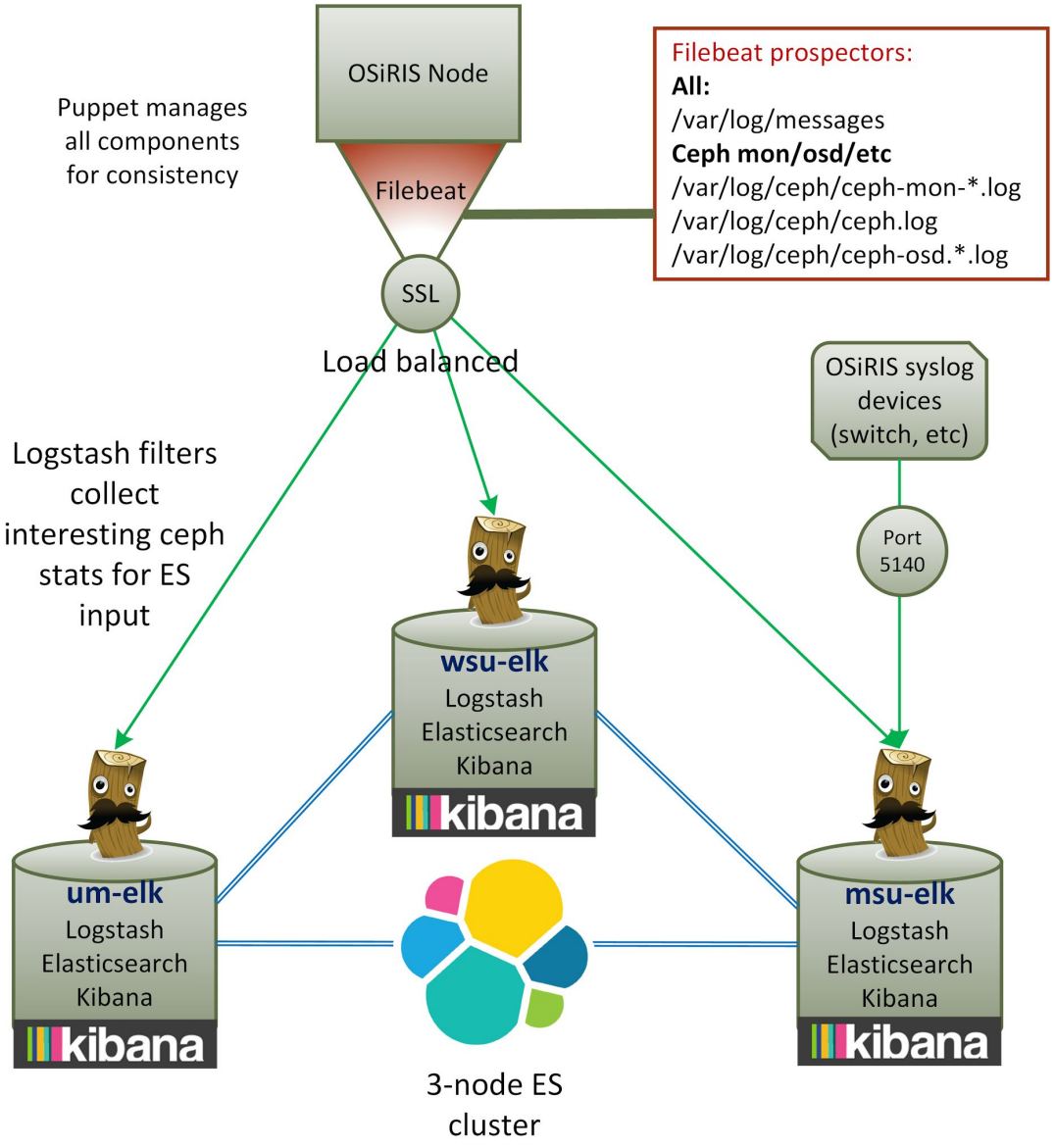
um-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu-omd.osris.org		20	0	0	0	0
UP	rac-um-globus01		47	0	0	0	0
UP	rac-um-virt01		55	0	0	0	0
UP	um-omd		30	0	0	0	0
UP	um-ps01		44	0	0	0	0
UP	um-stor01		174	1	6	2	0
UP	um-virt01		74	1	0	0	0
UP	msu-omd.osris.org		30	0	0	0	0
UP	rac-um-ps01		47	0	0	0	0
UP	um-elk		29	0	0	0	0
UP	um-pdu-20WA-L1		29	0	0	0	0
UP	um-puppet		30	0	0	0	0
UP	um-sw01		13	0	0	0	0
UP	wiki.osris.org		26	0	0	0	0
UP	oproj.aglt2.org		28	0	0	0	0
UP	rac-um-stor01		67	0	0	0	0
UP	um-mon01		32	0	0	0	0
UP	um-pdu-20WA-R1		30	0	0	0	0
UP	um-repo		29	0	0	0	0
UP	um-sw03		18	0	0	0	0
UP	wsu-omd.osris.org		31	0	0	0	0

wsu-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu-omd.osris.org		20	0	0	0	0
UP	rac-wsu-stor01		65	1	0	0	0
UP	wsu-mon01		34	0	0	0	0
UP	wsu-pdu-304-lf		3	1	2	0	0
UP	wsu-ps01		43	1	0	0	0
UP	wsu-virt01		73	1	0	0	0
UP	msu-omd-be.osris.org		1	1	0	0	0
UP	rac-wsu-virt01		51	0	0	0	0
UP	wsu-omd		35	0	0	0	0
UP	wsu-pdu-304-rf		3	1	2	0	0
UP	wsu-stor01		181	1	6	0	0
UP	rac-wsu-ps01		48	0	0	0	0
UP	um-omd-be.osris.org		1	0	1	0	0
UP	wsu-pdu-304-lb		3	1	2	0	0
UP	wsu-prov		31	2	0	1	0
UP	wsu-sw01		15	0	0	0	0

Monitoring with ELK



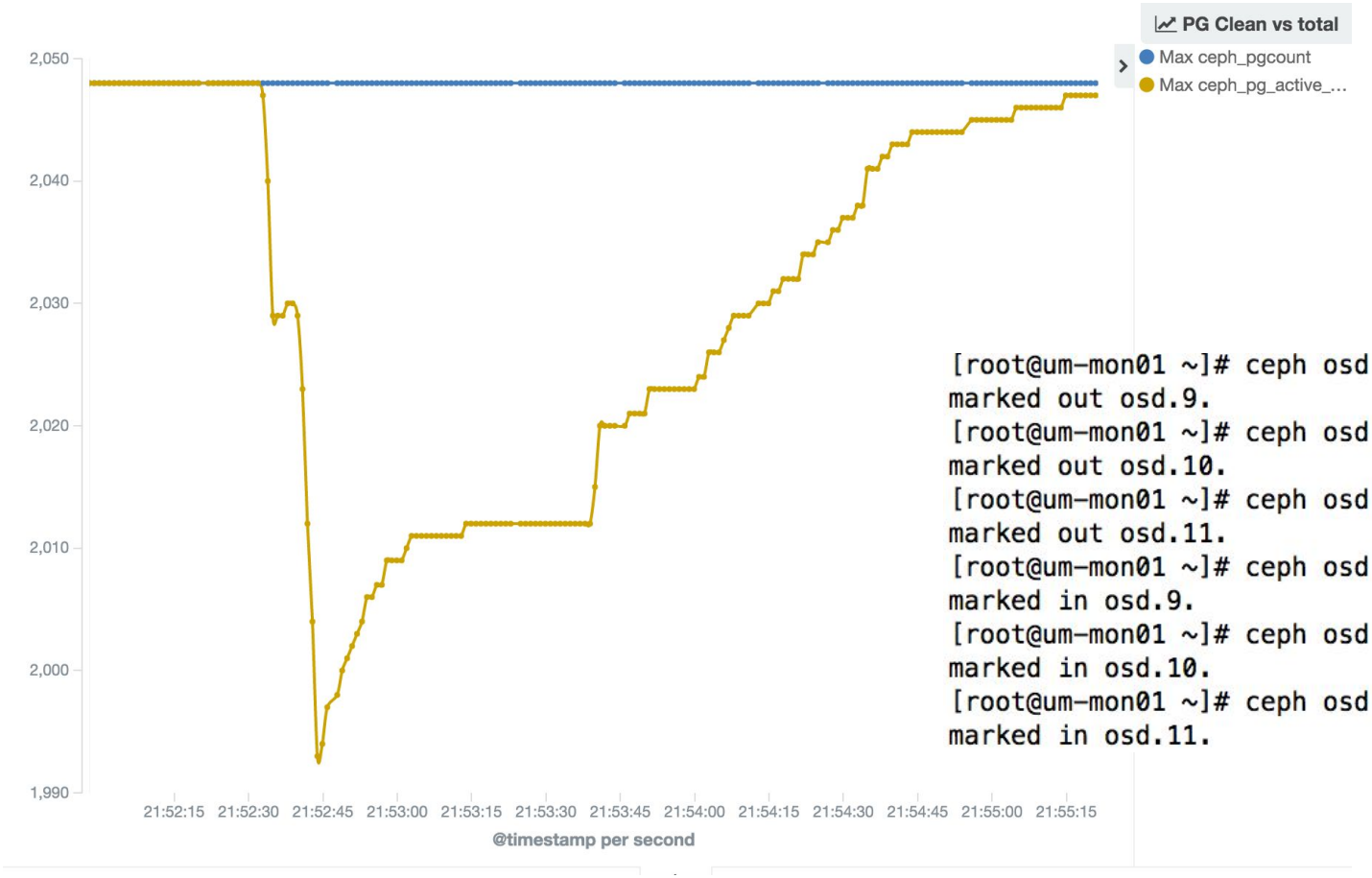
A resilient logging infrastructure is important to understand problems and long-term trends

The 3 node arrangement means we are not reliant on any one or even two sites being online to continue collecting logs

Ceph cluster logs give insights into cluster performance and health we can visualize with Kibana

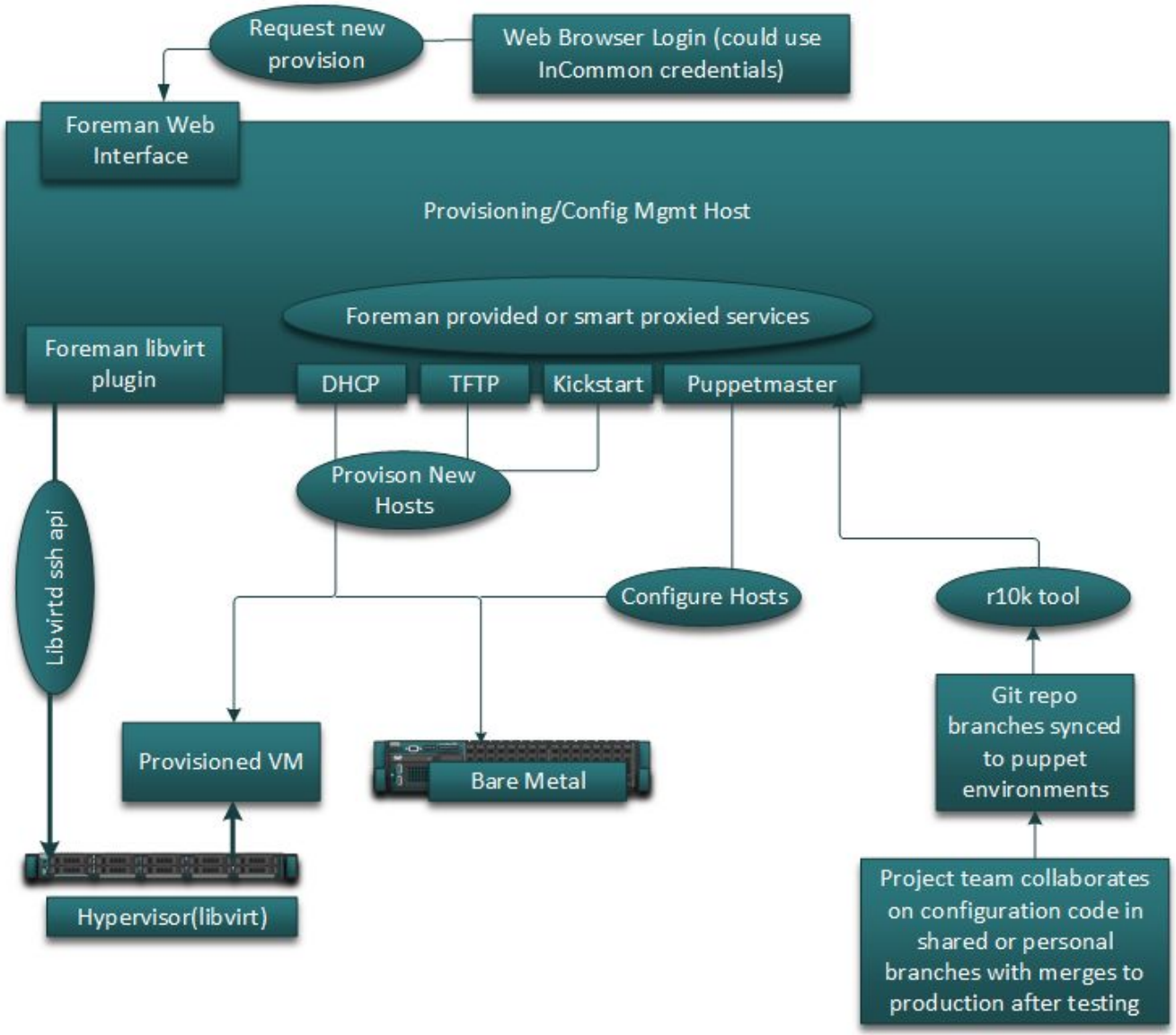
Monitoring with ELK

Simple example: The Ceph cluster regularly writes to <clustername>.log placement group status. Logstash pulls certain status out to our Elasticsearch index so we can use that as an integer in a date-range histogram




```
[root@um-mon01 ~]# ceph osd out 9
marked out osd.9.
[root@um-mon01 ~]# ceph osd out 10
marked out osd.10.
[root@um-mon01 ~]# ceph osd out 11
marked out osd.11.
[root@um-mon01 ~]# ceph osd in 9
marked in osd.9.
[root@um-mon01 ~]# ceph osd in 10
marked in osd.10.
[root@um-mon01 ~]# ceph osd in 11
marked in osd.11.
```

Orchestration



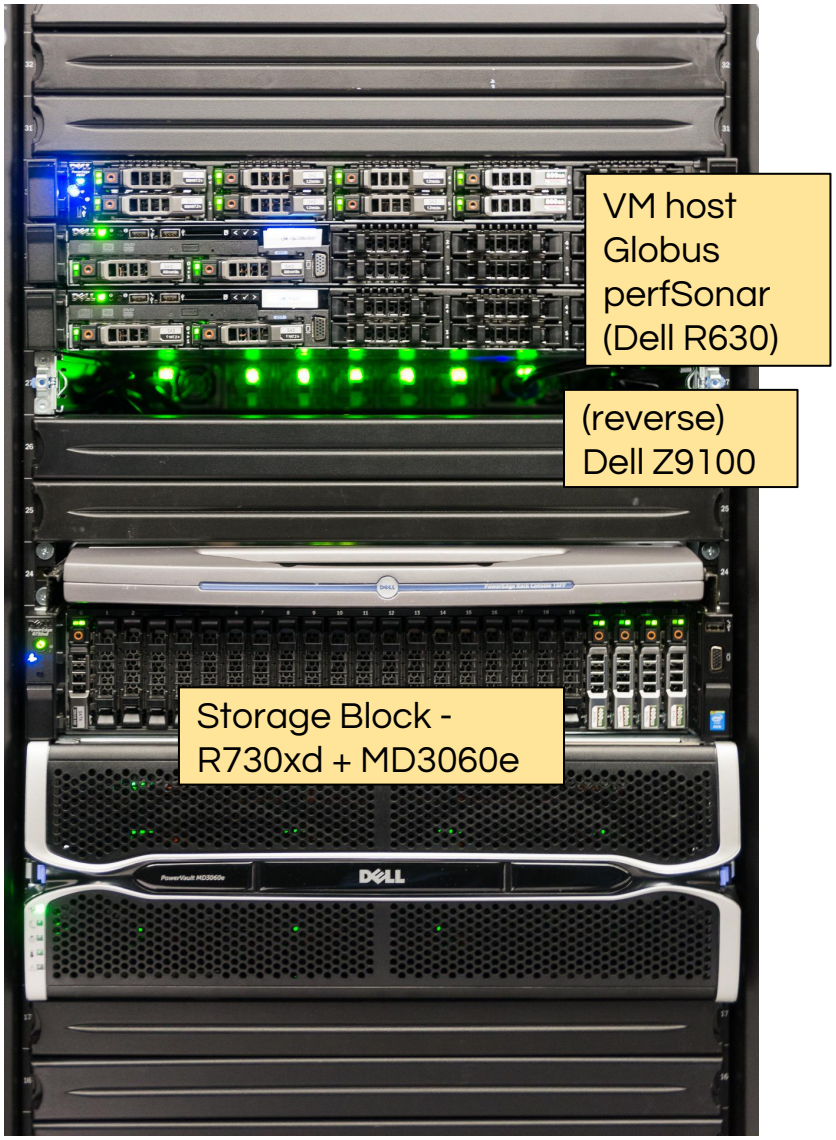
Deploying and extending our infrastructure relies heavily on orchestration with Puppet and Foreman

We can easily deploy bare-metal or VMs at any of the three sites and have services configured correctly from the first boot

Except: OSD activation requires a manual step

Openvswitch (scripted setup)

Status



The OSiRIS project requested proposals to meet our hardware needs in October 2015 (9 bids)

November 2015 we decided on Dell servers, HGST 8TB drives, Mellanox ConnectX 4 NICs

Orders out in December 2015

Equipment arrived in January/February 2016

Sites are all fully operational

Problems with Fiberstore 40GBASE-LR optics for Z9100 at UM - switch compatibility issues still in progress (though...we are running at full speed with borrowed Fiberstore Juniper-coded optics...)

Status - cluster

We deployed both production and test clusters running Infernalis later updated to Jewel

Production and test mons are different, isolated VMs (test mons have no interaction with production mons)

Production cluster and test cluster OSD reside on same hardware

- Test cluster takes 3 disks from each storage block
- Had to manually create systemd units for test cluster - by default the units packaged with Ceph can only deal with one cluster as defined in /etc/sysconfig/ceph (or default 'ceph')

First application of test cluster - update from Infernalis to Jewel

- Of all the updates we're likely to do, this one was really trivial and probably could have skipped testing but it doesn't hurt

Since it mirrors production cluster config we can also experiment with CRUSH maps and other items requiring full setup to test

Next Steps

Establishing baseline performance and evaluating/tuning as needed

Ceph has some benchmark tools built in
Have compiled results for lower level components (network, disk)
http://tracker.ceph.com/projects/ceph/wiki/Benchmark_Ceph_Cluster_Performance

Tuning our CRUSH map (data allocation map) to ensure we have resiliency at the level of site, rack, host

Default CRUSH map treats hosts as a failure domain, that's ok today since 1 host == 1 site

Tuning CRUSH map for cache overlay pools to read/write from local sites for better performance

Next Steps

This summer we will bring onboard our first science domains

ATLAS Great Lakes Tier 2 - processing ATLAS events read from Rados Gateway object store (S3 protocol)

Oceaning Modeling at UM - discussions underway to move US Navy oceanic models to OSiRIS for wider collaboration. Access protocol yet to be determined.

Our Goal: Enabling Science

The OSiRIS project goal is enable scientists to collaborate on data easily and without (re)building their own infrastructure

The science domains mentioned all want to be able to directly work with their data without having to move it to their compute clusters, transform it and move results back

Each science domain has different requirements about what is important for their storage use-cases: capacity, I/O capability, throughput and resiliency. OSiRIS has lots of ways to tune for these attributes.

Summary

- There are **significant challenges** in providing infrastructures that transparently enable scientists to quickly and easily extract meaning from large, distributed or diverse data.
- OSiRIS will incorporate a number of cutting edge technologies to build this infrastructure.

We have a talented collaboration prepared to meet the challenges and unanswered questions inherent to our goals.