OSiRIS Overview and Status

CNSECCS Symposium 2018, University of Michigan

Open Storage Research Infrastructure

Shawn McKee
University of Michigan
for the OSiRIS Collaboration

Mission Statement

OSiRIS is a pilot 5-year project funded by the National Science Foundation 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. We enable this via a combination of **network discovery**, **monitoring** and **management** tools and through the creative use of **CEPH** features.

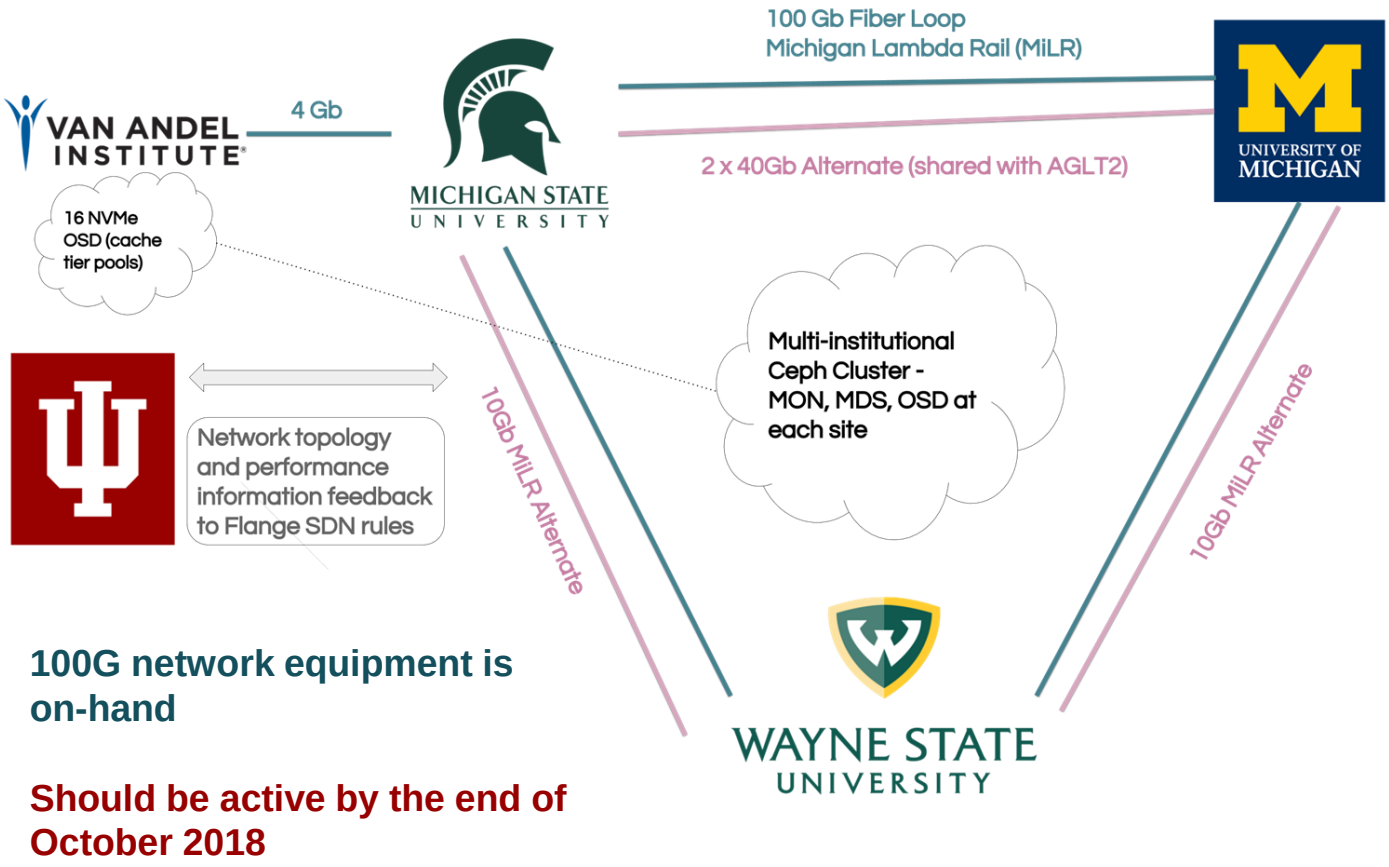
By providing a single data infrastructure that supports computational access on the data "in-place", we can meet many of the data-intensive and collaboration challenges faced by our research communities and enable these communities to easily undertake research collaborations beyond the border of their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

High Level Overview

Single Ceph cluster (Mimic 13.2.1) spanning **UM, WSU, MSU** - 600 OSD, 5 PiB (soon **840** OSD / **7.4** PiB)

Network topology store (UNIS) and SDN rules (Flange) managed at IU

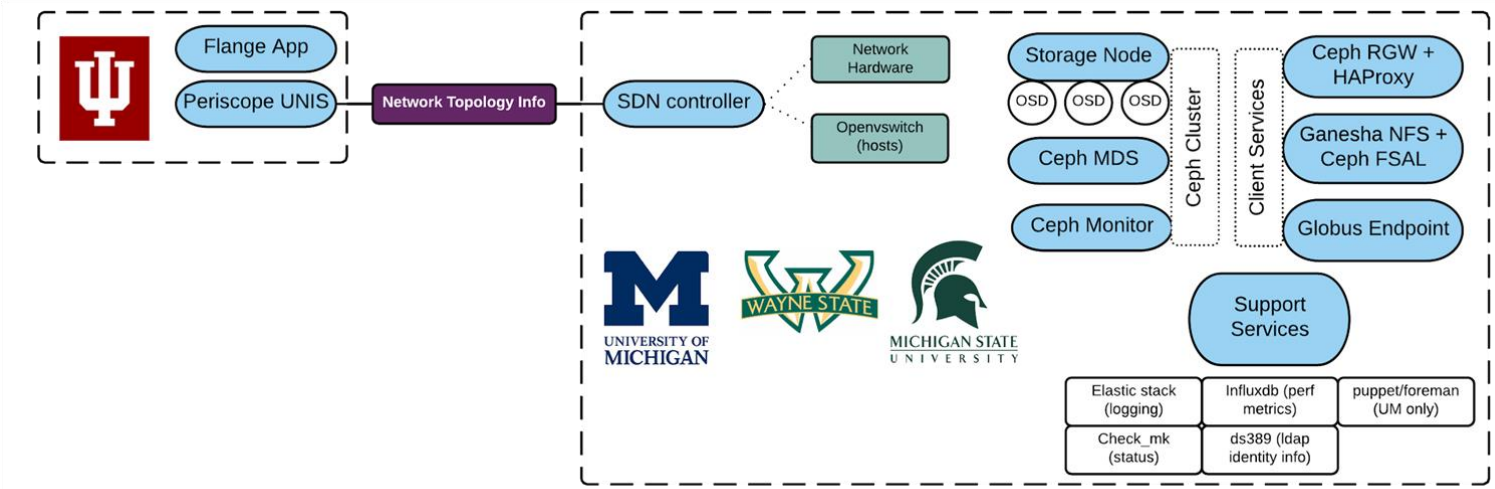
NVMe nodes at VAI used for Ceph cache tier (Ganesha NFS export to local cluster)



100G network equipment is on-hand

Should be active by the end of October 2018

Site Overview



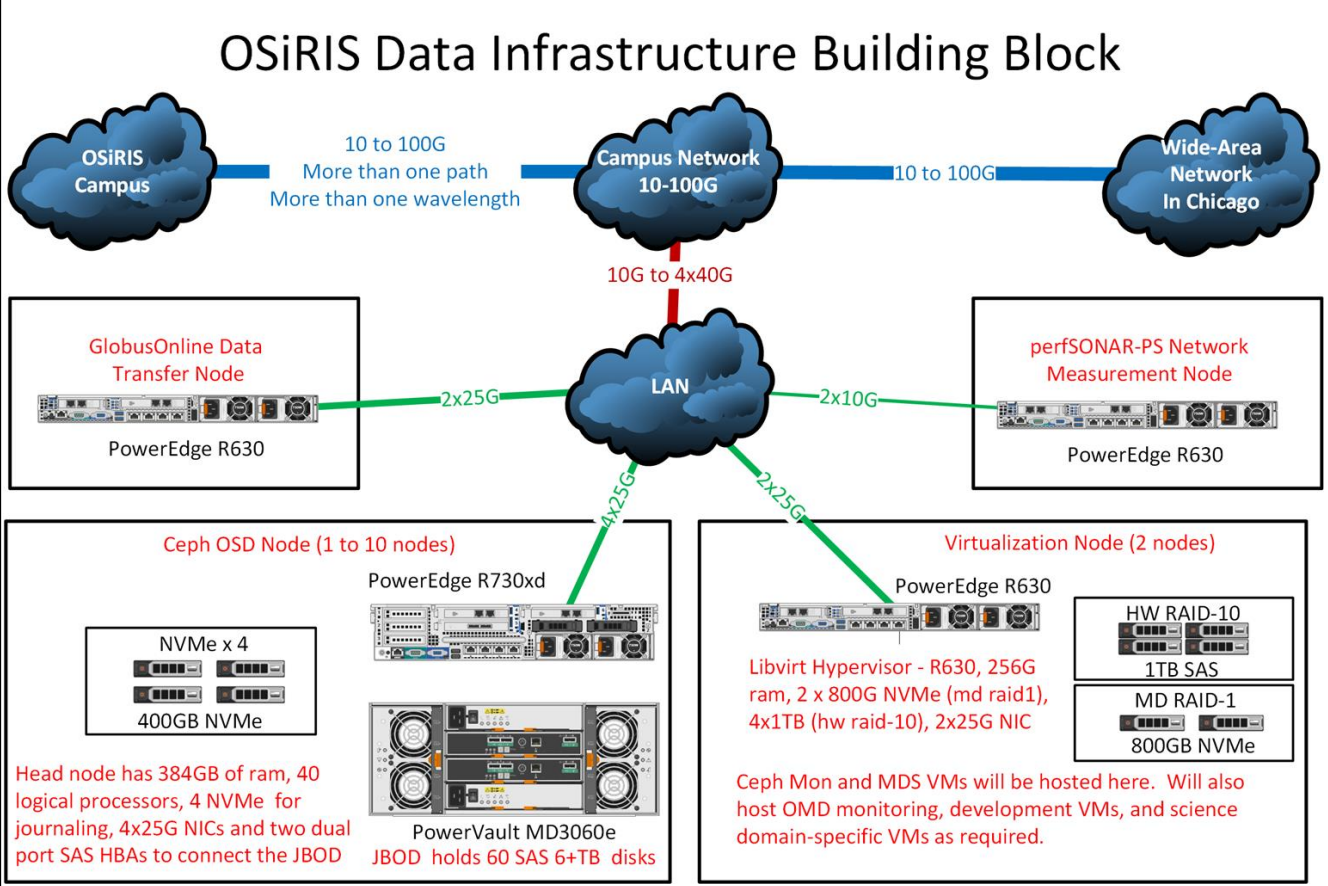
Core Ceph cluster sites share **identical** config and similar numbers / types of OSD

Any site can be used for S3/RGW access (HAproxy uses RGW backends at each site)

Any site can be used via Globus endpoint (same shared FS)

Users at each site can mount NFS export from Ganesha + Ceph FSAL. NFSv4 idmap umich_ldap scheme used to map POSIX identities.

Site Overview - hardware



Example hardware models and details shown in the diagram on the left.

This year's purchases used R740 headnodes and 10TB SAS disks and Intel P3700 PCIe NVMe devices

Quick Hardware Overview (current)

CPU

- 56 hyper-thread cores (28 real cores) to 60 disks in one node
- can see 100% usage on those during startup, normal 20-30%

Memory

- 384 GB for 600 TB disk (about 650 MiB per TB)
- 6.5GB per 10TB OSD

NVMe OSD Bluestore DB (Intel DC P3700 1.6TB)

- 110GB per 10TB OSD
- 11GB per TB of space
- about 1% of block device space (docs recommend 4%....not realistic for us)

VAI Cache Tier

- 3 nodes, each 1 x 11 TB Micron Pro 9100 NVMe
- 4 OSD per NVMe
- 2x AMD EPYC 7251 2Ghz 8-Core, 128GB

OSiRIS Science Domain Users

Building on existing collaboration between MSU and the VAI, OSiRIS has installed NVMe-based Ceph OSD nodes and an NFS gateway at the institute to enable direct access **bioinformatics research data**. This facilitates data access for **VAI researchers** to leverage the computational resources at MSU ICER.

U.S. Naval Research Lab is collaborating with researchers at UM to share their **high-resolution ocean models** with a global community. This unclassified data was stored on US Navy computers that were not easily accessible to many researchers. OSiRIS enables scientists worldwide to leverage these models.

The **ATLAS Event Service** is designed to leverage object stores like OSiRIS for fine grained physics event data which can be retrieved and computed in small chunks and leverage transient compute resources.

OSiRIS Science Domain Users

Homed at **UM ISR**, the NIH 5 year project 'Effect of the Placental Epigenome on Stunting in a Longitudinal African Cohort' uses OSiRIS to store and share data to a wider community

At WSU the Microscopy, Imaging & Cytometry Resources (MICR) core leverages OSiRIS to enable wider access to imaging data.

The JETSCAPE collaboration at WSU is an NSF funded multi-institutional effort to **design event simulators for ultra-relativistic heavy-ion collisions**. OSiRIS provides a universal storage platform for collaborative access.

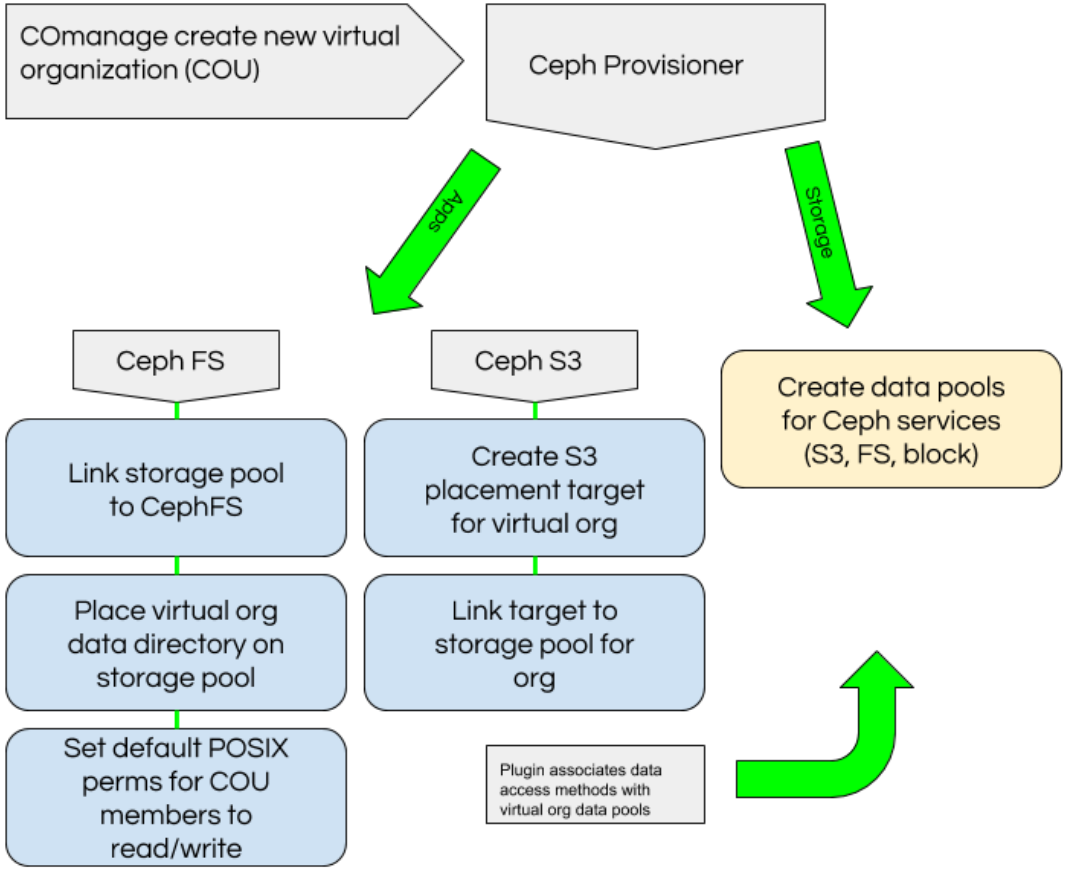
Global Nightlights at UM is the **only complete archive of NOAA nighttime imagery** from 2 different satellite programs: **DMSP** (1993-2016) and **VIIRS** (2012-ongoing). By keeping portions of this archive on OSiRIS we enable wider usage of the datasets by researchers outside the institution.

Provisioning Ceph VO Storage

Supporting multiple virtual organizations required us to write some glue to provision resources and users

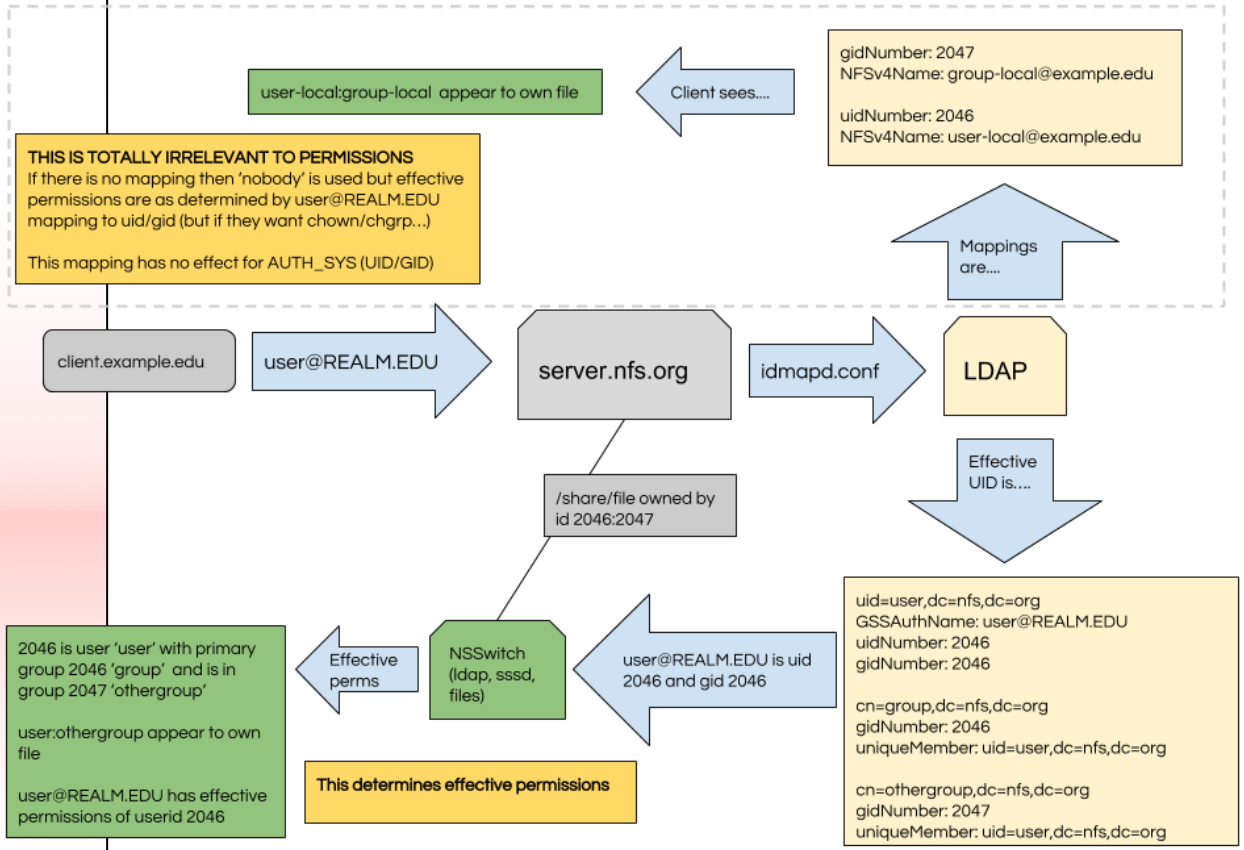
We use Internet2 **COmanage** to manage virtual orgs and enroll users based on their institutional identities (**Shibboleth** + federations like **InCommon, eduGain**)

When we create new VO, COmanage creates Ceph resources for them and associates directories and S3 placement targets



https://github.com/MI-OSiRIS/comanage-registry/tree/ceph_provisioner

Leveraging NFSv4 idmap



Ganesha NFS server with **Ceph FSAL** and **idmapd** to map Kerberos identities to our POSIX information

Idmapd uses umich_ldap config to lookup identities stored with nfsv4 LDAP schema (NFSv4Name, GSSAuthName, NFSv3RemotePerson, etc)

Campus users can access OSiRIS via these NFS gateways, automounted on compute clusters at UM / MSU

Container with baked in config: https://hub.docker.com/r/miosiris/nfs-ganesha-ceph

CephFS Client Access

We explored approaches to allow non-root users to individually use fuse for CephFS mounts and map client POSIX ID to our internal OSiRIS uid/gid/groups.

- Modified MDS server using client key as identity and doing POSIX info lookup in LDAP to then modify incoming client requests
- Modified ceph-fuse client allowing user setting of any given uid/gid/grouplist

Server approach did not pan out for us. We couldn't ever get permissions to work as expected in normal operations

- Thought it might be because Fuse client is interfering? We're not really FS experts
- Code still lives here: https://github.com/MI-OSiRIS/ceph/commits/mds_idmap

Working on a simpler approach that modifies fuse client so it does operations with arbitrary user-provided list of POSIX info.

- Security relies on uid,gids restriction of ceph access keys
- This is an ugly hack and we continue to look for a more elegant solution...

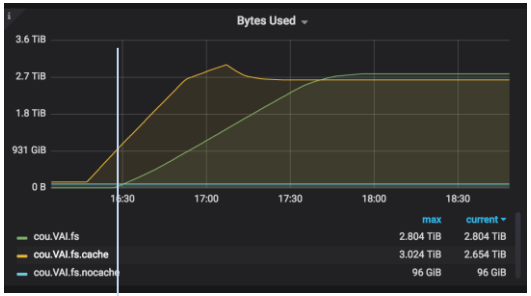
Ceph Cache Tier at VAI

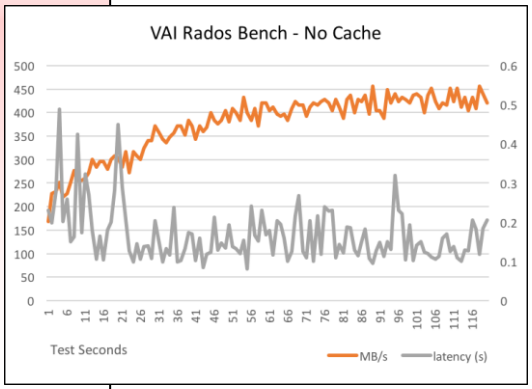
We setup Ceph cache tier pool mapped to NVMe OSD on 3 hosts at the Van Andel Institute in Grand Rapids (about 1 hr drive west of the MSU campus)

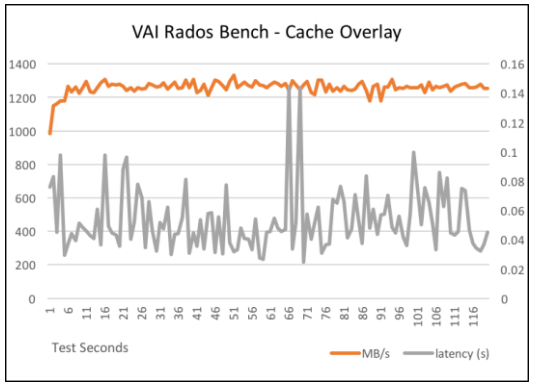
- **Note: will demo OSiRIS caching at SC18 in Dallas in November**

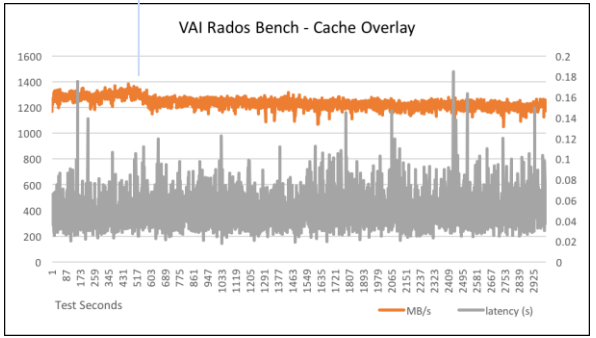
Benchmarks show **significant increase** in performance for VAI clients, and no traffic back to main pool until tier flush params reached (set fairly low for this test).

```
cache_target_dirty_ratio .1
cache_target_dirty_high_ratio .2
cache_target_full_ratio .3
```

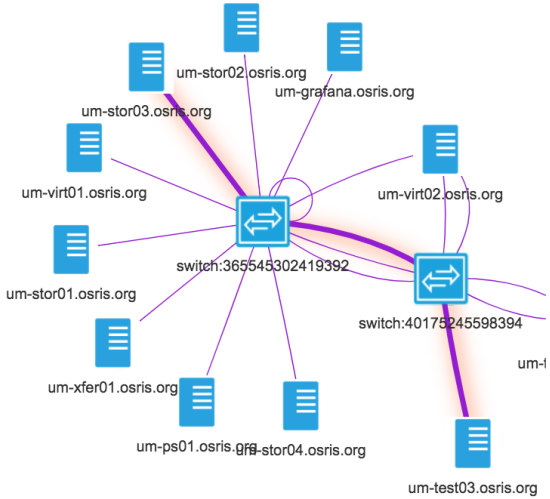






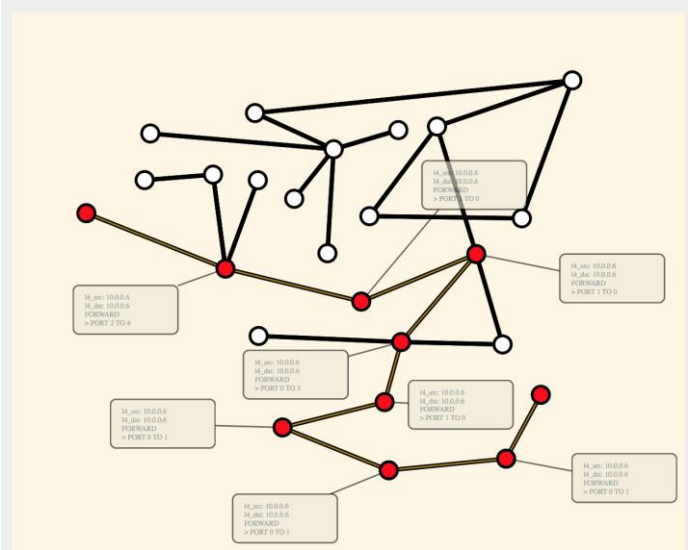


Flange and Openflow



Discover topology by listening to **LLDP** via **OF controller** and **SNMP**.

Path creation and instantiation using **NMAL** orchestration service.



The flow orchestration DSL - *Flange* - provides language support for modifying network configurations. *Flange* reacts to topological and performance changes; reconfiguring flows to respond to the conditions in the network.

High-level goal: Orchestrate system and science user traffic to minimize congestion and maximize use of **all** available network capacity.

Ceph + WAN Challenges

A too-frequent issue we encounter is 'asymmetrical' network problems

- For example: If UM has reliable connectivity to WSU, and to MSU, but MSU does not have connectivity to WSU then OSD issue conflicting 'down' and 'up' reports and flap (eg, one leg of the connectivity 'triangle' between 3 sites is unreliable)
- Net result is that cluster I/O is flaky, either because some OSD can't reliably reach others to replicate or they are busy deciding whether to be up or down
- Also causes monitors to trigger frequent elections and cluster commands hang while they figure it out

Similar/related: If connectivity is not completely lost but instead have intermittent packet loss it **causes worse problems than if the link is just dead.**

We have a lot of cluster reliability issues because of WAN link issues. Some ongoing issues needing diagnosis, but sometimes a guy with a backhoe is your bad luck.

- Often use reweight-subtree 0 to take a site out of commission temporarily
- Maybe helps: set 'mon_osd_down_out_subtree_limit' to our site crush buckets

Longer term fix will be to use UNIS+FLANGE to orchestrate consistent conflict-free connectivity for OSiRIS (both Ceph and our science domain users)

Status and Plans

We are just starting our **fourth year** with the project and will soon have about 7.4 PB (raw) in Ceph and our new 100G network in place between our three Ceph storage locations

The main goals for year-4:

- Integrate two more OSiRIS science domains
  - Bioinformatics
  - Acquatic Bio-Geo-Chemistry
- Continue to augment, improve and harden our “client toolkit”
- Enable OSiRIS network orchestration in production
- Experiment with different Ceph pool configurations to better support specific science domain use-cases

ATLAS is interested in how dCache over Ceph behaves for production. We intend to test about 1 Petabyte of OSiRIS storage into two or more pools that AGLT2 dCache will use.

There is also a new NSF funded project in the US call Open Storage Network (OSN), PI: Alex Szalay at John’s Hopkins.

- OSiRIS will be collaborating with OSN, adding 1 PB into the OSN initial prototype

Summary

The **OSiRIS Project** is progressing well and is supporting numerous diverse science communities.

- We have been surprised that more science domains have not focused on using OSiRIS to provide in-place use-and-transformation of their data.

Our main challenges are around incorporating networking orchestration into normal operations and improving our client toolkit to provide easier client access and capabilities.

There are a number of research areas we will be exploring, covering Ceph optimization and use, network monitoring and orchestration and user “ease-of-use”

Questions?

Any questions?

Resources

- OSiRIS Website: http://www.osris.org
- Github: https://github.com/MI-OSiRIS