OSiRIS Overview and Challenges

Ceph BOF, Supercomputing 2018, Dallas

Open Storage Research Infrastructure

Ben Meekhof
University of Michigan ARC-TS
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.

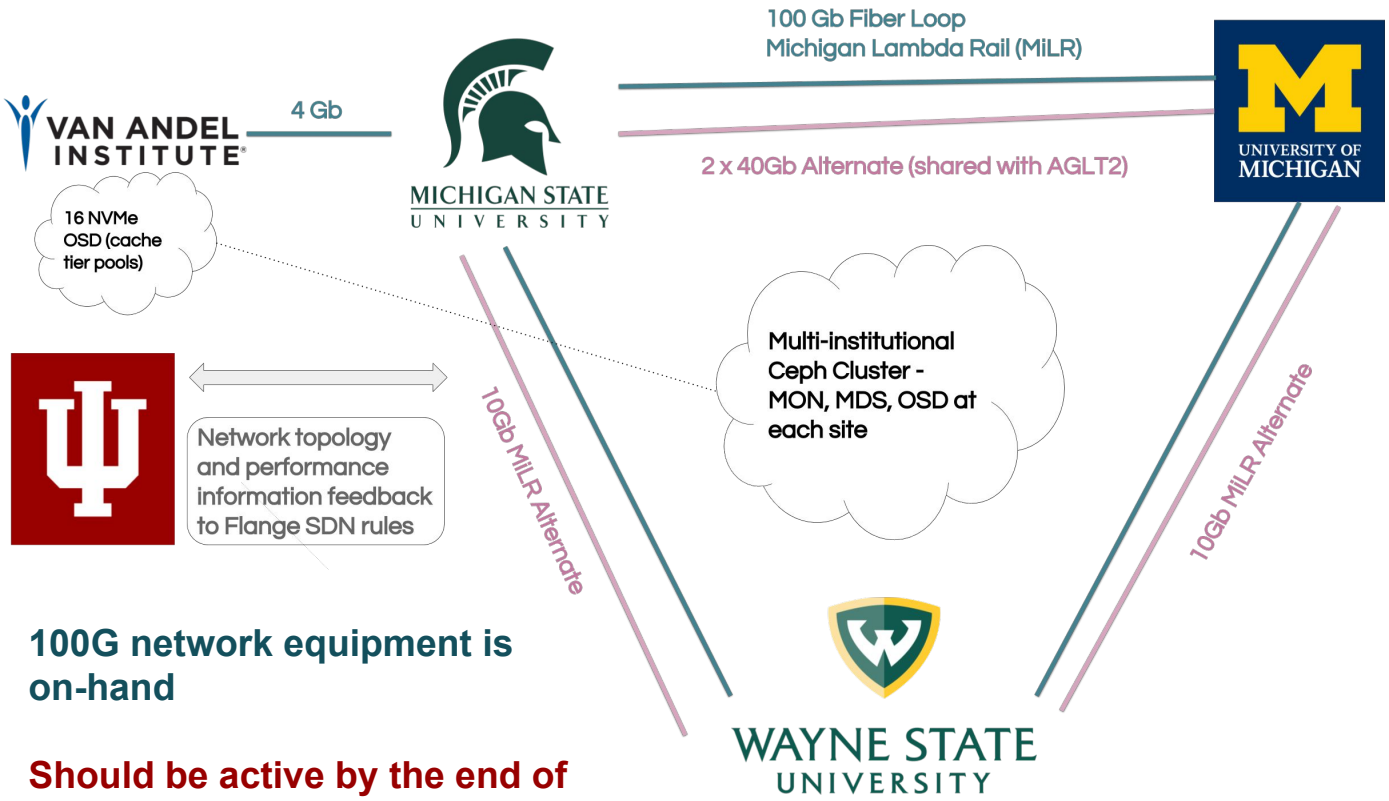
By providing a single data infrastructure that supports computational access on the data "in-place", we can meet many of the data-intensive and collaboration challenges faced by our research communities and enable these communities to easily undertake research collaborations beyond the border of their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

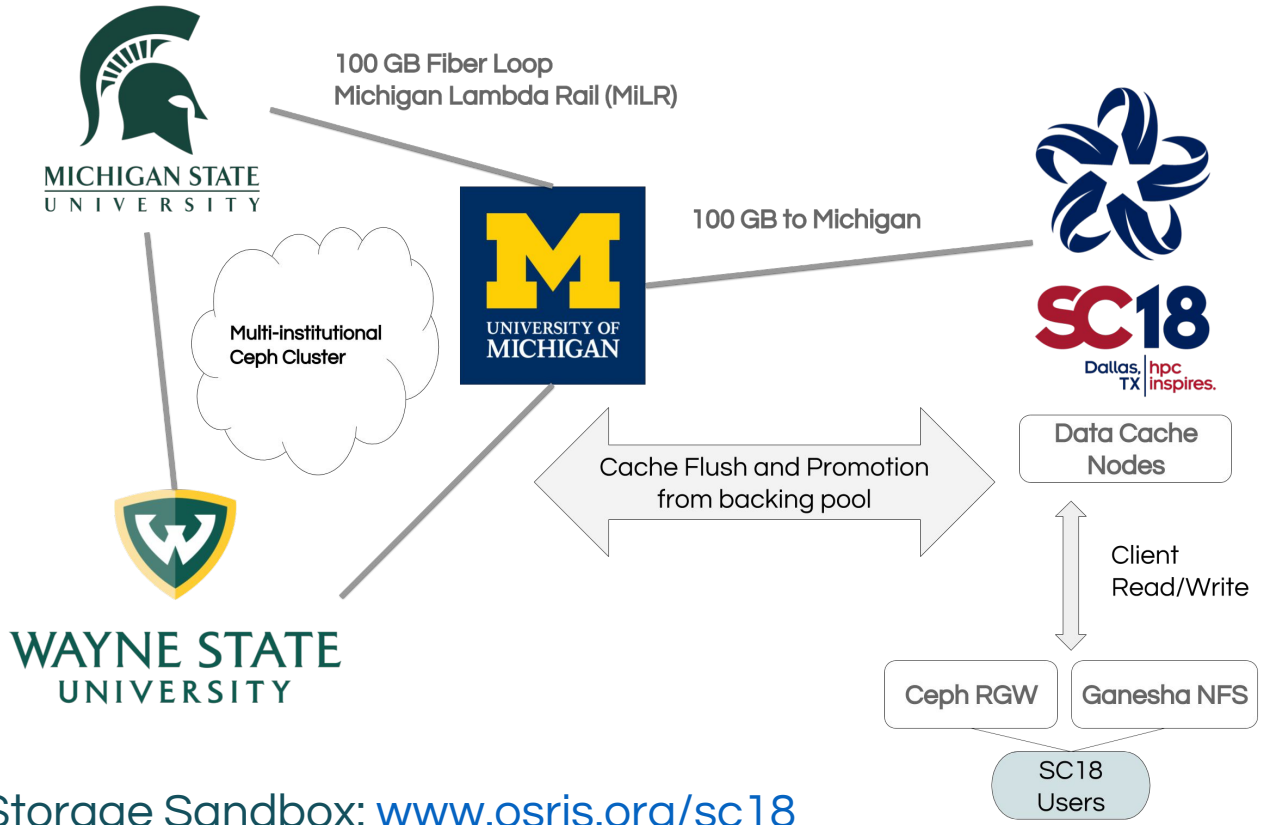
Should be active by the end of November 2018 (links in place)

Cache tier at SC18

Storage node at SC hosting several pools overlaid as cache tier

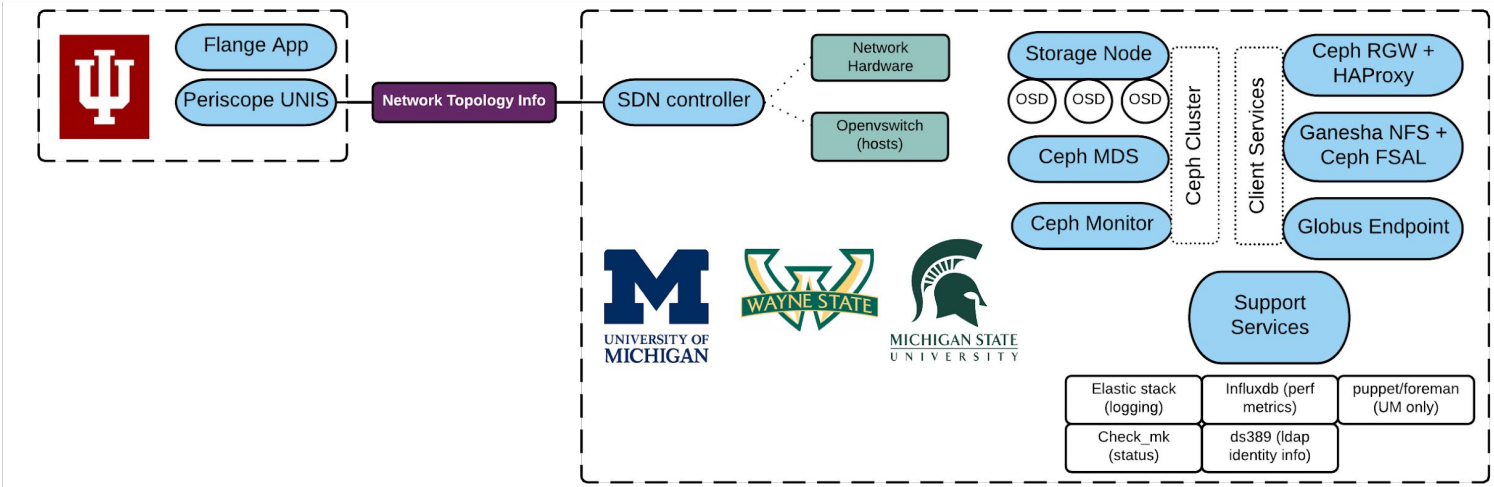
Also experimenting with pools placing primary OSD at SC18 (helps read speeds locally)

Encourage folks to use the 'sandbox' storage resources available to all SCInet hosts



Storage Sandbox: www.osris.org/sc18

Site Overview



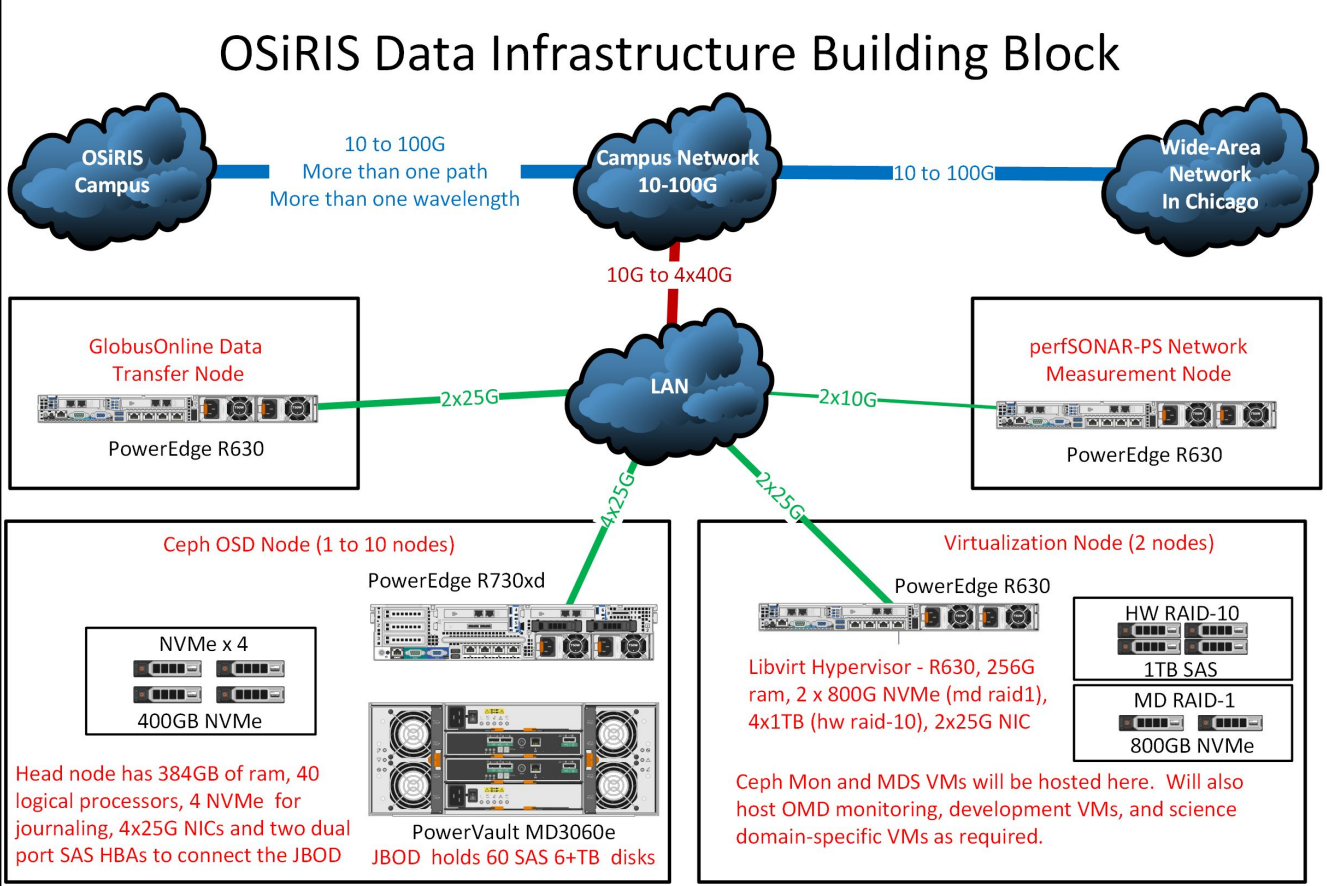
Core Ceph cluster sites share **identical** config and similar numbers / types of OSD

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 1.6TB 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

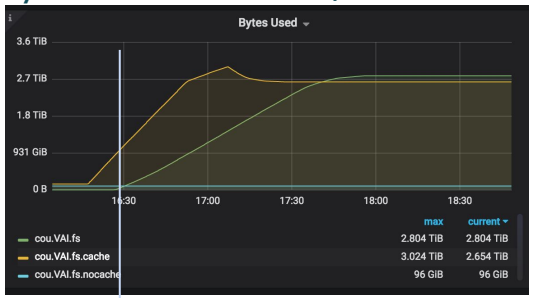
- 3 nodes, each 1 x 11 TB Micron Pro 9100 NVMe
- 4 OSD per NVMe
- 2x AMD EPYC 7251 2Ghz 8-Core, 128GB

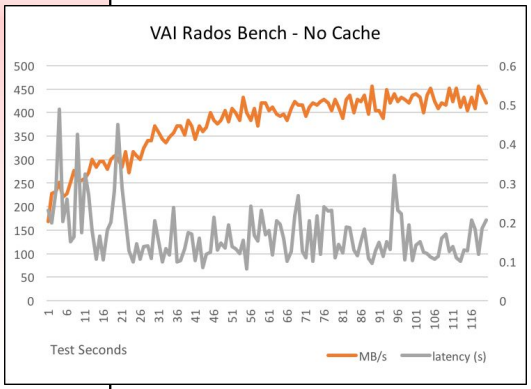
Ceph Cache Tier at VAI

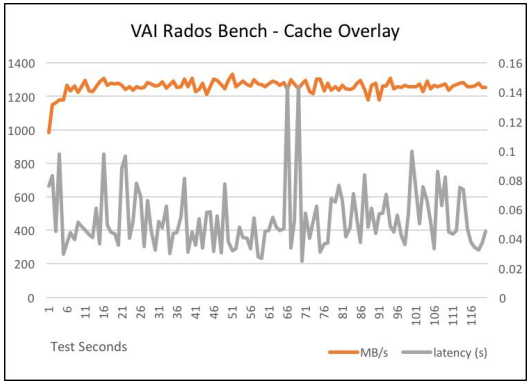
We setup Ceph cache tier pool mapped to NVMe OSD on 3 hosts at the Van Andel Institute in Grand Rapids (about 1 hr drive west of the MSU campus)

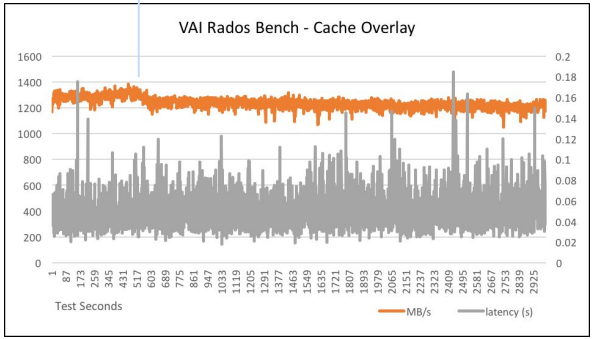
Benchmarks show significant increase in performance for VAI clients, and no traffic back to main pool until tier flush params reached (set fairly low for this test).

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







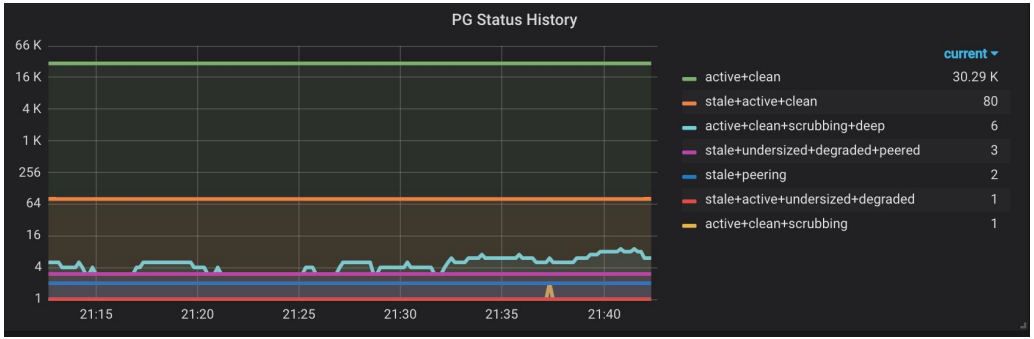


Ceph Mgr Influxdb module

First version of module was written and submitted as PR by UM OSiRIS undergrad student

- We were using collectd-ceph previously, almost entirely replaced by mgr module for mon, osd, rgw, mds stats

Same student later contributed the PG status feature



Also submitted PR to enable sending to multiple influxdb destinations - a feature we liked about collectd.

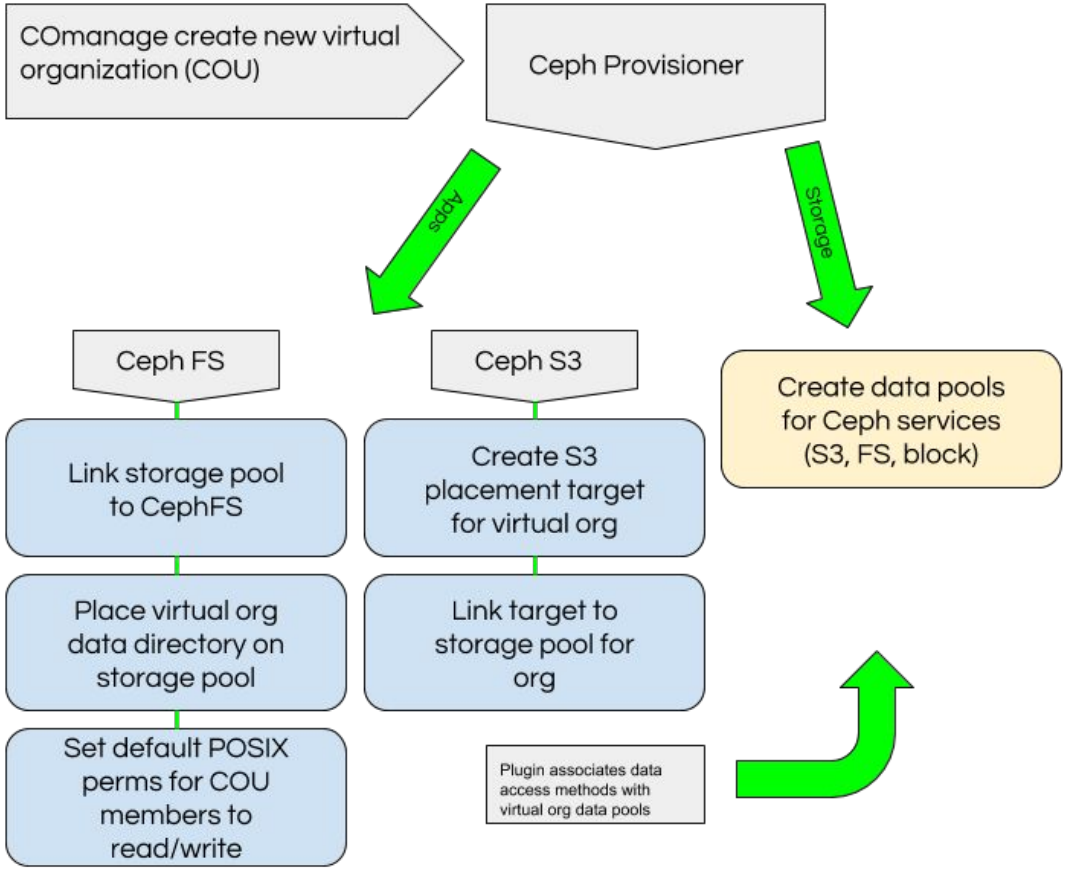
- It was (rightly) suggested that the configuration was not user-friendly
- Another student picked up that idea this past summer, basing on most recent upstream and incorporating previous suggestions. PR waiting on test/cleanup.

Provisioning Ceph VO Storage

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

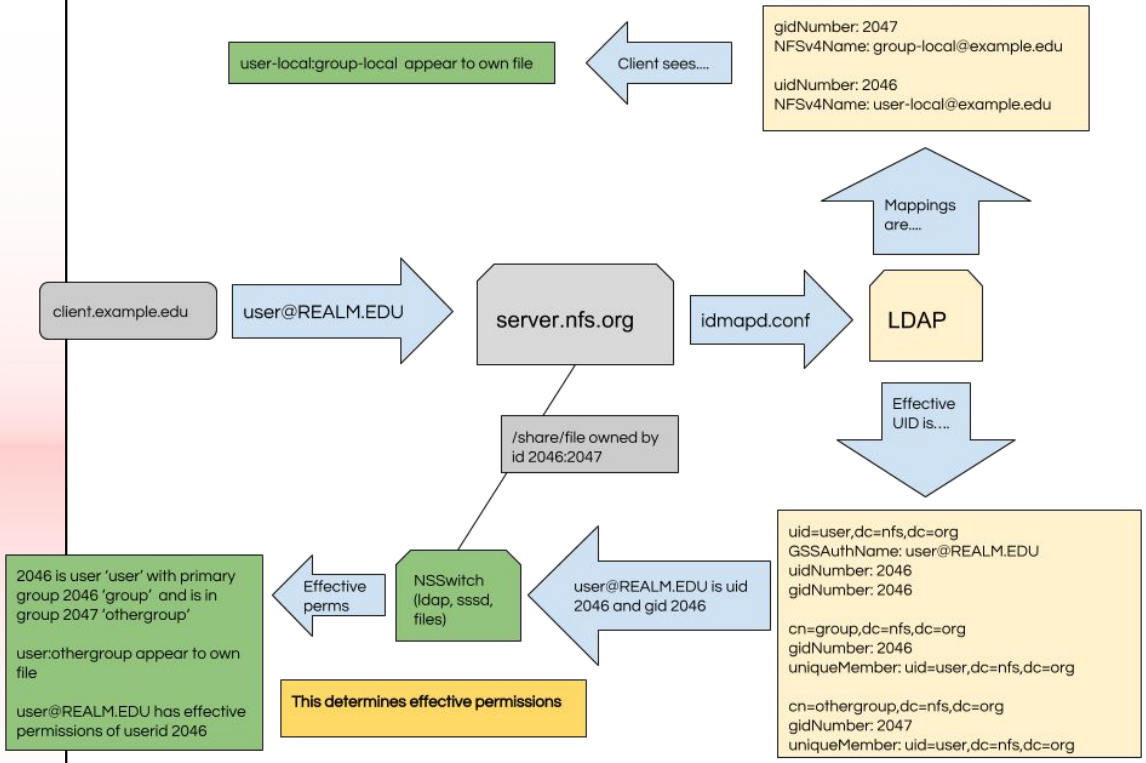
We use Internet2 **COmanage** to manage virtual orgs and enroll users based on their institutional identities (**Shibboleth** + federations like **InCommon**, **eduGain**)

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 Troubles

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

Server approach did not pan out. We couldn't make it work for non-root users.

- Works if you are root, apparently some client interactions we didn't understand
- Code still lives here: https://github.com/MI-OSiRIS/ceph/commits/mds_idmap

Tried 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
- Permissions all appear to work correctly, create/delete/chown/etc...but
- Data never gets written out to files...write isn't denied, but no data is put into the file?
- https://github.com/MI-OSiRIS/ceph/commits/mds_idmap_client

All of this work was done by undergrad at UM (since moved on)

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. Reasons vary - campus network issues, blind backhoe drivers, etc.

Longer term fix will be to use UNIS+FLANGE components being worked on by IU team (not covered in this talk) to orchestrate consistent conflict-free connectivity for OSiRIS

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

Questions?

Any questions?

Resources

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