

OSiRIS Overview for ARC-TS and Unit IT



Open Storage Research Infrastructure

Ben Meekhof

University of Michigan

Advanced Research Computing

OSiRIS Technical Lead

OSiRIS Summary

OSiRIS is a pilot project funded by the NSF to evaluate a software-defined storage infrastructure for our primary Michigan research universities and beyond.

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

- Leveraging CEPH features such as CRUSH, cache tiers to place data
- Radosgw/S3 behind HAproxy, public and campus local endpoints
- Globus access to S3 or mounted CephFS
- Identity establishment and provisioning of federated users (CManage)

UM, driven by OSiRIS, recently joined Ceph Foundation:

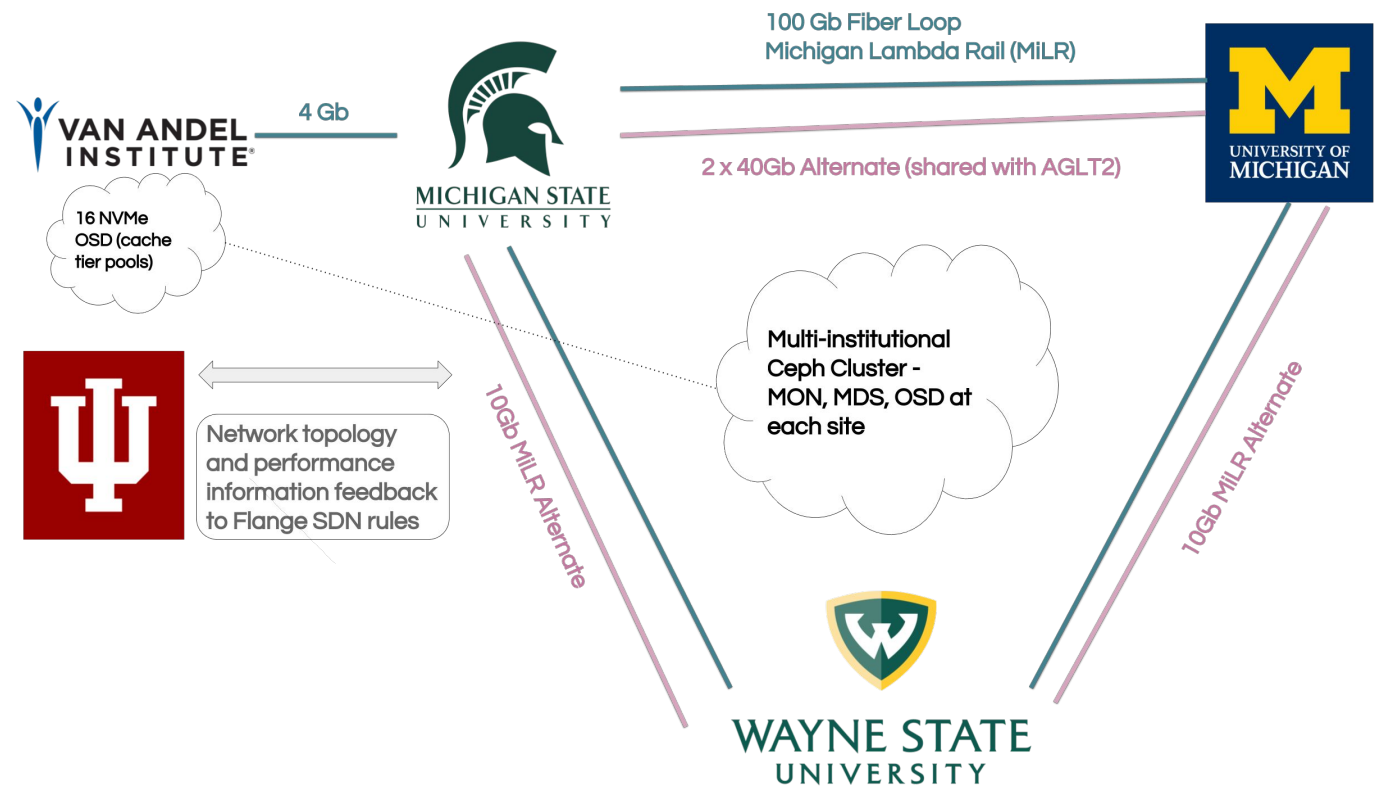
<https://ceph.com/foundation>

OSiRIS Summary - Structure

Single Ceph cluster
(**Mimic 13.2.x**)
spanning **UM, WSU,**
MSU - 792 OSD, 7 PiB
(soon 1300 OSD, 13
PiB)

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

NVMe nodes at VAI
used for Ceph cache
tier only



OSiRIS Identity Onboarding

OSiRIS relies on other identity providers to verify users

- InCommon and eduGain federations

Users enroll into Virtual Organizations (COU, COmanage Organizational Unit)

- The first step for a new group/project/etc to use OSiRIS is talking with the OSiRIS team to work out use case, space, and potential workflows
- We then establish a new VO / COU and users can enroll and use

Users authenticate and enroll via COmanage (Shibboleth)

- Users choose their COU (virtual org) at enrollment
- Designated virtual org admins can approve new enrollments, OSiRIS admins don't need to be involved for every enrollment

Once enrolled COmanage feeds information to provisioning plugins.

- LDAP, Grouper are core plugins included with COmanage
- We wrote a Ceph provisioner for the rest

CManage - Virtual Org Provisioning

When we create CManage COU (virtual org):

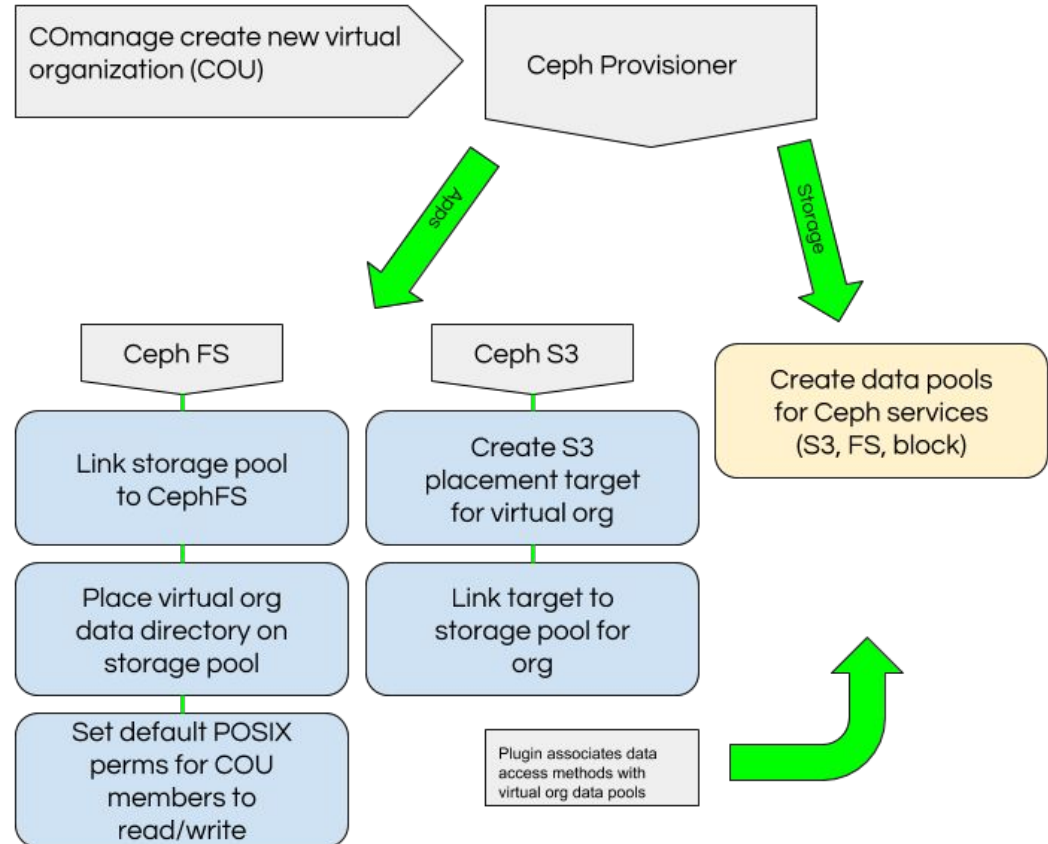
Data pools created

RGW placement target defined to link to pool
cou.Name.rgw

CephFS pool create and added to fs

COU directory created and placed on CephFS pool

Default perms/ownership set to COU all members group, write perms for admins group (as a default, can be modified)



Grouper - VO Group Self Management

Virtual Orgs are provisioned from
COmanage as Grouper stems

VO admins are given capabilities to
create/manage groups under their
stem

Groups become Unix group objects in
LDAP usable in filesystem permissions

Every COU (VO) has the CO_COU
groups available for use by default,
COmanage sets membership in these

The screenshot displays the Grouper web interface for managing groups. On the left, a sidebar contains a 'Quick links' section with items like 'My groups', 'My folders', and 'My favorites'. Below it is a 'Browse folders' section showing a tree structure starting from 'Root' and navigating through 'OSIRIS' to 'OsirisAdmin'. The main content area shows the 'OsirisAdmin' group details, including a breadcrumb trail 'Home > Root > OSIRIS > OsirisAdmin', a 'More' dropdown menu, and a 'Filter for:' input field. A list of group members is displayed below, including 'Up one folder', 'CO_COU_OsirisAdmin_admins', 'CO_COU_OsirisAdmin_members_active', 'CO_COU_OsirisAdmin_members_all', and various system groups like 'admin', 'lu-devel', 'mca', 'ps', 'puppet', 'sdn', 'sudo-ceph-cmd', 'sudo-mca', 'sudo-mon', 'sudo-msu-virt02-m', 'sudo-nfs', 'sudo-pkg-cmd', 'sudo-ps', and 'sudo-puppet-cmd'.

CManage Credential Management

Home > OSIRIS > Ceph Credentials

Ceph Credentials

Description	User ID	Credentials
S3/RGW Access Key	bmeekhof	access_key: 4R7GOZ3XRFS807F3D9WM secret_key: 9sFmGDa63t1ox77BY44moovbdGNkElFPtQBal7qP
Enter New Id:	<input type="text" value="bad%#\$example"/>	Can only contain alphanumeric, hyphen (-), or underscore (_)
S3/RGW Access Key	myservice	access_key: 05RBS3I78HP7URS9E806 secret_key: 6rKmKoomRzjGrXSPEZ17XC6xxofd6jq5AbFQdC81
S3/RGW Access Key	myservice	access_key: C136YO80W76TGUFKG2X6 secret_key: 68O33P10Salg2lj8rR4t8b4E3QICMoIIONFXGLpt
Ceph Client Key	osiris.bmeekhof	[client.osiris.bmeekhof] key = AQACon1cZSs4FBAAVDWHZLuUboF

Add a new userid

Add additional access/
secret pair to this userid

Reset main access/
secret pair for userid

Remove this id and all
keypairs associated

Remove this access/
secret pair

Click here to open
selector and choose
new bucket placement

CManage Ceph Provisioner plugin
provides user interface to
retrieve/manage credentials

S3/RGW Access Key

myservice

Default Bucket Data Placement:

- OsirisAdmin
- ATLAS**
- ARCTS

Ceph Client Key

osiris.bmeekhof

Globus and gridmap

We provide Globus access to CephFS and S3 storage

- For now separate endpoints, future Globus version will support multiple storage connectors
- Ceph connector uses radosgw admin API to lookup user credentials and connect to endpoint URL with them

Credentials: CILogon + globus-gridmap

- We keep CILogon DN in LDAP voPerson CoPersonCertificateDN attribute
- We wrote a Gridmap plugin to lookup DN directly from LDAP (thanks to our undergraduate student at UM, Raul Dutta)
- <https://groups.google.com/a/globus.org/forum/#!topic/admin-discuss/8D54FzJzS-o>

Puppet

We manage everything with puppet, deployment with Foreman

- foreman-bootdisk for external deployments such as Van Andel
- r10k git environments

Define a site and role (sub-role for storage) from hostname, use these in hiera lookups

- Example: um-stor-nvm01 becomes a Ceph 'stor' node using devices as defined in 'nvm' nodetype to create OSD
- site, role, node, nodetype are hiera tree levels
- At the site level define things like networks (frontend/backend/mgmt), CRUSH locations, etc

Ceph deployment and disk provisioning managed by Puppet module

- Storage nodes lookup Ceph OSD devices in hiera based on hostname component
- Our module was forked from openstack/puppet-ceph
- Supports all the ceph daemons, bluestore, multi-OSD devices
- <https://github.com/MI-OSiRIS/puppet-ceph>

Foreman makes our deployment really easy with the use of host groups, templates, puppet integration, and GUI or CLI tools

For example, simple CLI leveraging common host group, we just script this in a loop:

```
hammer host create --hostgroup BOSS --name um-stor-ds01 --mac=E4:43:4B:9B:DE:1E \  
--ip=141.211.169.24 --interface identifier=em3 --managed True \  
--operatingsystem "Scientific Linux 7.7"
```

<input type="checkbox"/>	Power	Name	Operating system	Puppet Environment	Model	Host group	Last report
<input type="checkbox"/>	🔌	✔️ um-stor-ds06.osris.org	🌀 Scientific Linux 7.7	production		Base/Metal/Storage/BOSS	8 minutes a...
<input type="checkbox"/>	🔌	✔️ um-stor-ds07.osris.org	🌀 Scientific Linux 7.7	production		Base/Metal/Storage/BOSS	7 minutes a...
<input type="checkbox"/>	🔌	✔️ um-stor-ds08.osris.org	🌀 Scientific Linux 7.7	production		Base/Metal/Storage/BOSS	4 minutes a...
<input type="checkbox"/>	🔌	✔️ um-stor-ds09.osris.org	🌀 Scientific Linux 7.7	production		Base/Metal/Storage/BOSS	29 minutes ...
<input type="checkbox"/>	🔌	✔️ um-stor-ds10.osris.org	🌀 Scientific Linux 7.7	production		Base/Metal/Storage/BOSS	7 minutes a...
<input type="checkbox"/>	🔌	⚠️ um-stor-lnvm01.osris.org	🌈 CentOS Linux 7.5.1804	production	SYS-1018R-...		2 months ago
<input type="checkbox"/>	🔌	✔️ um-stor-test01.osris.org	🌀 Scientific Linux 7.6	production	um-virt01-m	Base/VM	22 minutes ...
<input type="checkbox"/>	🔌	✔️ vai-stor-nvc01.osris.org	🌀 Scientific Linux 7.6	production	PowerEdge ...	Base/Metal/Storage	25 minutes ...

Round Up: How can we use OSiRIS?

Have a use case for OSiRIS? Get in touch with osiris-help@umich.edu and let us know.

What is a use case for OSiRIS?

- Needs to compute with off campus resources - accessing data directly with S3 tools is a perfect fit here
- Collaborates off-campus, esp at WSU or MSU. Any person from any InCommon / eduGain institution can establish identity with OSiRIS (there are open identity providers for non-edu people as well)
- Just needs a place to store and share data and use std Unix tools/groups - sure we can do that, use Globus or shell access to our CephFS xfer nodes
- Globus to S3 gives users a familiar tool for moving data and then there is the option to start leveraging S3 tools with that data (even if they aren't interested at first).

There's no particular requirement to establish a VO and start using OSiRIS. Especially if you have someone who wants to use S3 we're a good on-campus option, reachable from campus clusters directly without proxy (S3 endpoints in the same data centers)

Round Up: How can we access OSiRIS?

We have transfer nodes at each university with CephFS mounted and shell access

Globus endpoints exporting all CephFS storage

S3 endpoints at each university, DNS names to reach specific institution or RR between all

- S3 client libs such as Python boto
- CLI tools such as s3cmd or awscli
- FUSE mount s3fs-fuse
- Many S3 tools default to Amazon URL, but easy to specify ours
- We also have a 'client bundle' which attempts to simplify the FUSE use case and will be expanded to make CLI usage/config as easy as possible

Globus endpoints exporting S3 storage (users see buckets they own)

All of these are covered on documentation page: <http://www.osris.org/documentation/>

This is the 4th year of OSiRIS.

- Grant period is 5 years
- A no-cost extension is planned for year 6
- Potential campus support after that

We'd like to get more data on the platform, have a number of queued up users or new engagements (Brainlife, Oakland University, IceCube, Open Storage Network, U-M NeuroImaging Initiative, more)

More utilization of S3 services as a more practical path to working in-place on data sets

- Good option for OSG users
- Globus connector for Ceph gives people a familiar way to move data and have the option to use S3 clients and tools
- We can scale S3 (Ceph Radosgw) infinitely

Questions?

OSiRIS Team Contact: osiris-help@umich.edu

Website: <http://www.osris.org/documentation>

OSiRIS Contacts at UMICH:

Project PI: Shawn McKee, smckee@umich.edu

Soundararajan Rajendran, rajends@umich.edu

Muhammad Akhdhor, muali@umich.edu

Reference / Supplemental

Internet2 CManage: <https://spaces.at.internet2.edu/display/CManage/Home>

Internet2 Grouper: <https://www.internet2.edu/products-services/trust-identity/grouper/>

OSiRIS CephProvisioner:

https://github.com/MI-OSiRIS/comange-registry/tree/ceph_provisioner/app/AvailablePlugin/CephProvisioner

OSiRIS Docker (Ganesha, NMAL containers): <https://hub.docker.com/u/miosiris>

OSiRIS Docs: <https://www.osris.org/documentation>

Some Numbers (current hw purchase)

Dell PowerEdge R7425 / AMD EPYC 7301 2.2GHz/2.7GHz, 16 core

128GB Memory

16 x 12TB 7.2K RPM NLSAS 12Gbps 512e 3.5in hard drive

4 x 512GB Samsung 970 Pro NVMe in ASUS Hyper M.2 X4 Expansion Card (DB/WAL device, 4 per NVMe)

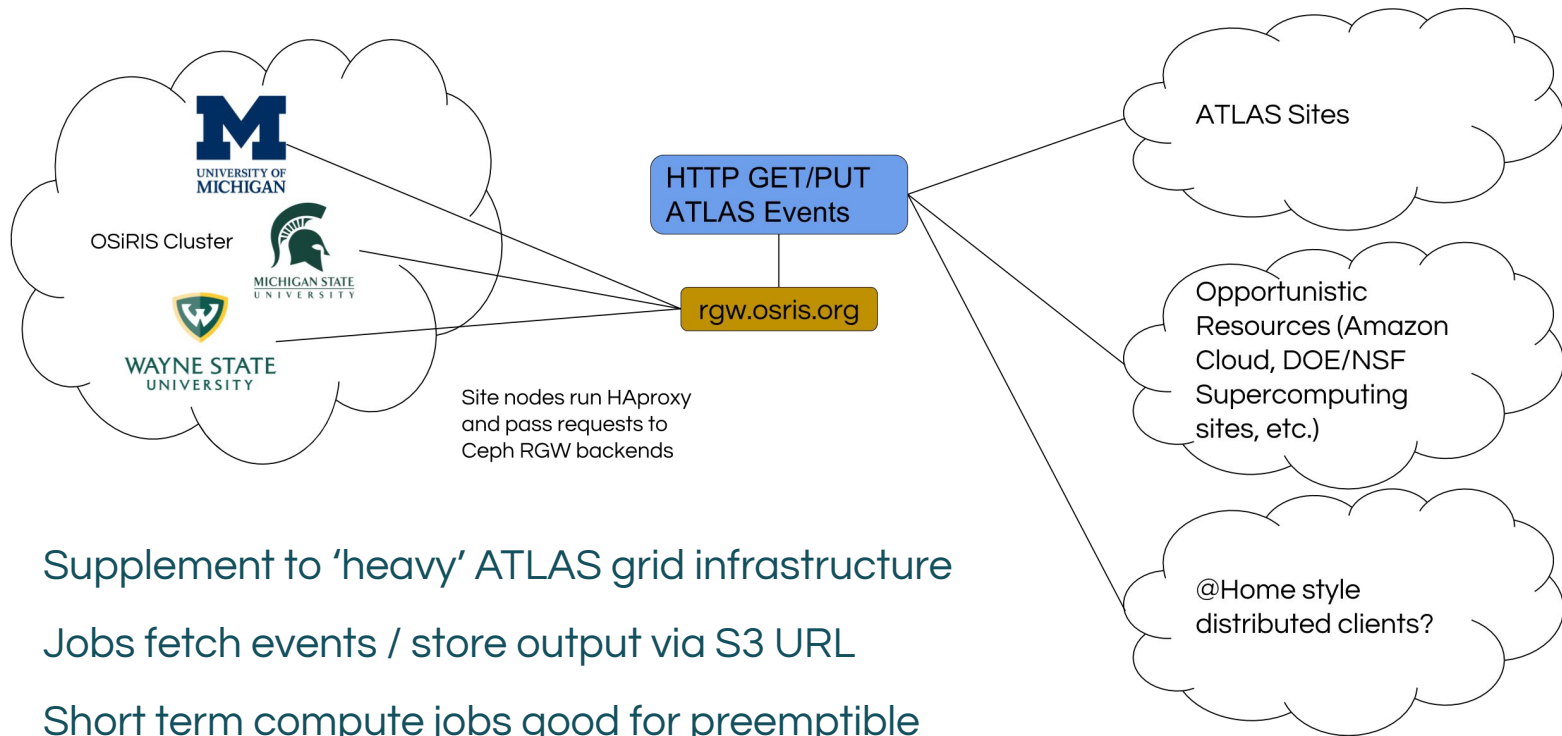
Mellanox ConnectX-4 LX Dual Port 10/25GbE SFP28

Net Result: 1 core per OSD / disk, 128GB DB volume per OSD, 8GB RAM per OSD (minus OS needs), 50 Gbps connectivity (OVS bond)

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

ATLAS Event Service

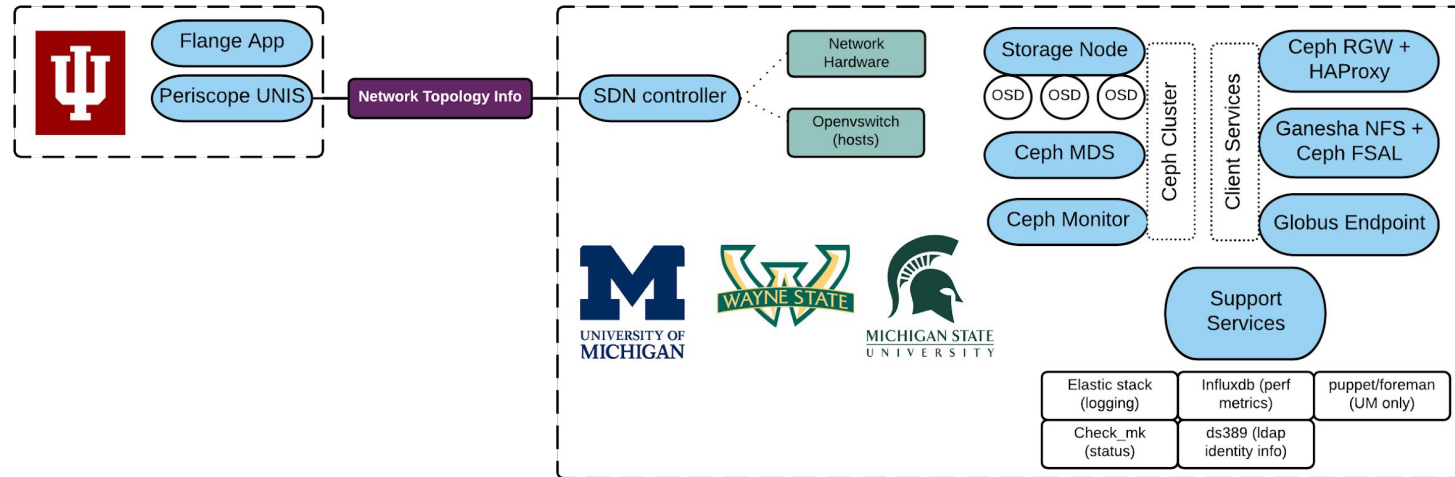


Supplement to 'heavy' ATLAS grid infrastructure

Jobs fetch events / store output via S3 URL

Short term compute jobs good for preemptible resources

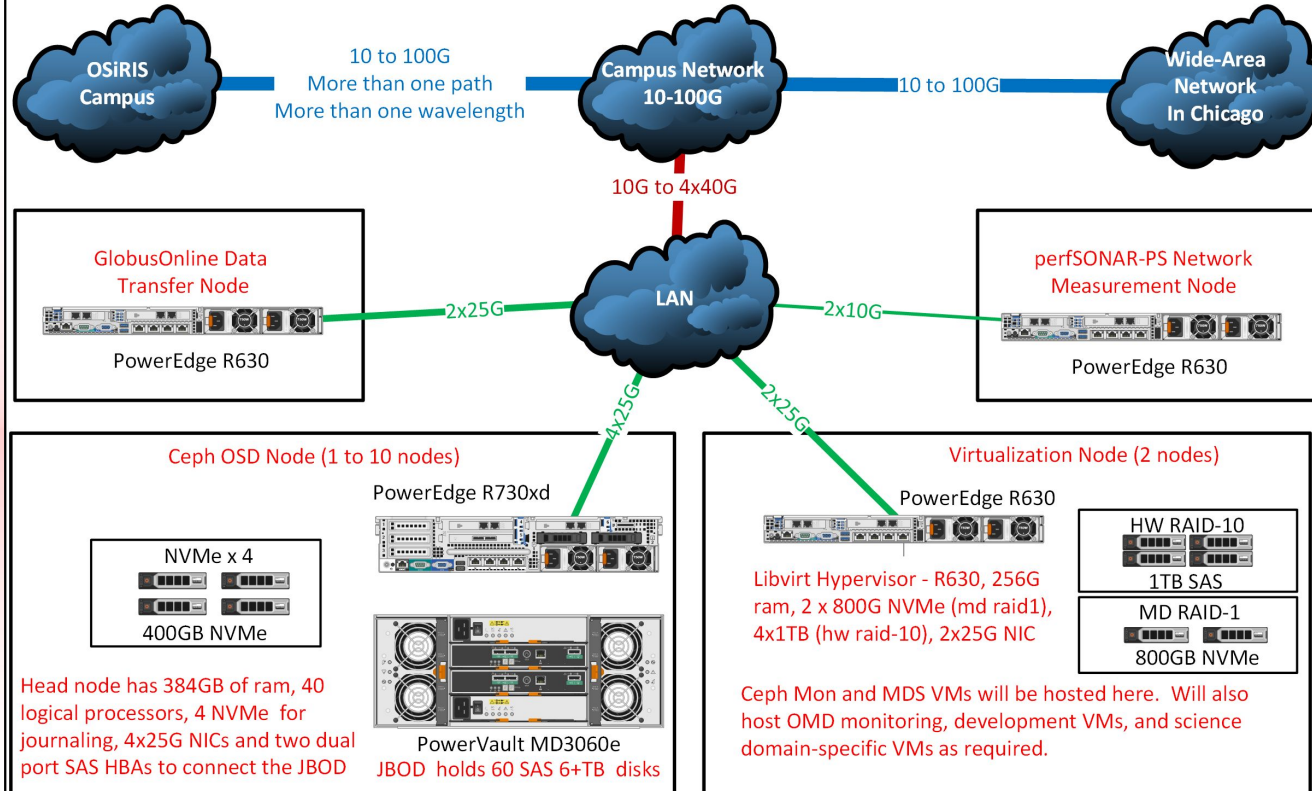
Site Overview



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 for FS or S3
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 (existing)

OSiRIS Data Infrastructure Building Block

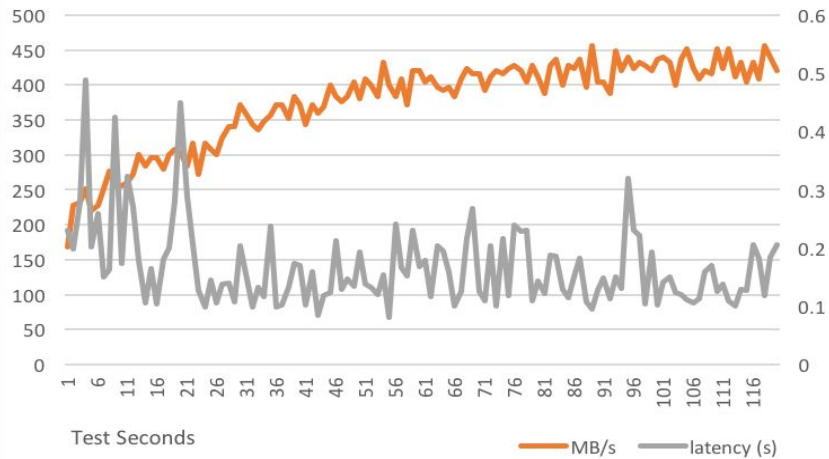


Example hardware models and details shown in the diagram on the left.

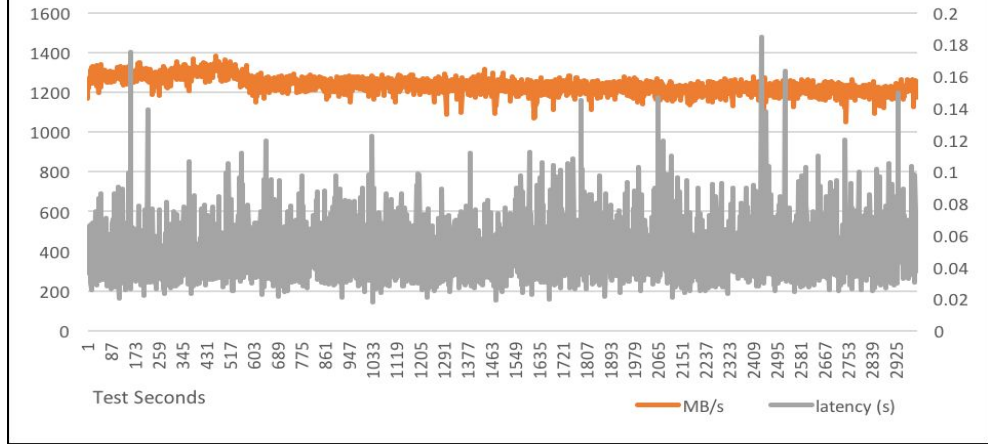
This year's purchases used R740 headnodes and 10TB SAS disks and Intel P3700 PCIe NVMe devices

Cache Tier Benchmarks - RADOS (VAI)

VAI Rados Bench - No Cache



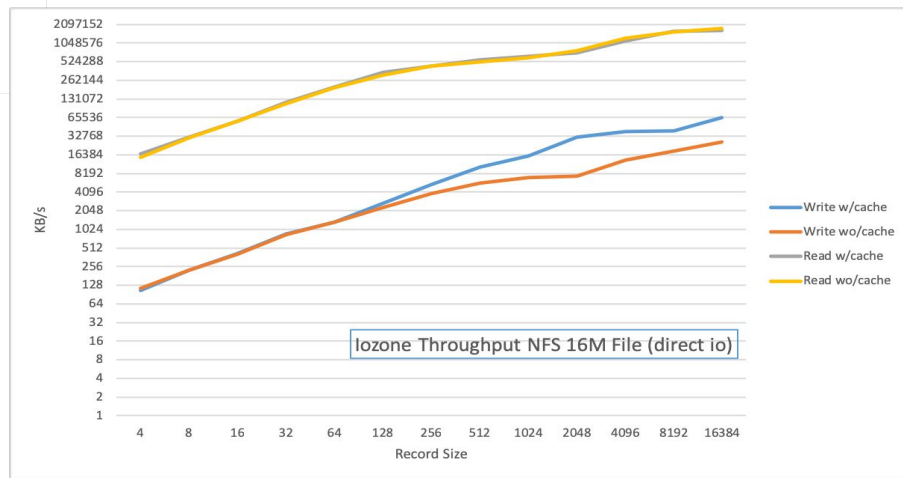
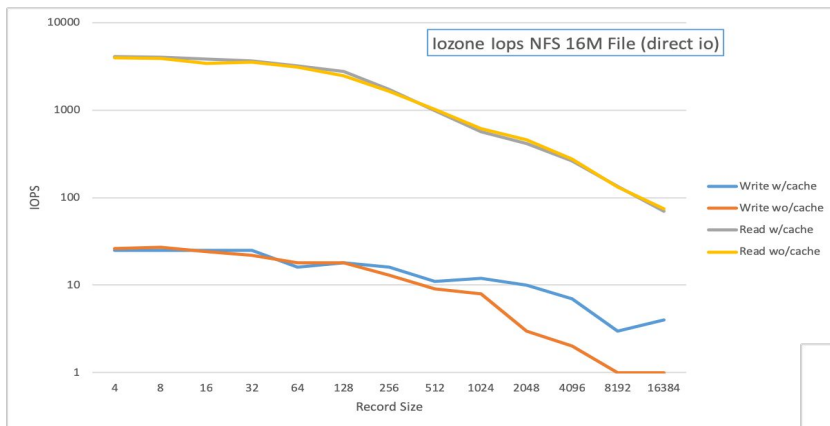
VAI Rados Bench - Cache Overlay




<http://www.osris.org/domains/vai.html>

Cache Tier Benchmarks - NFS / lozone

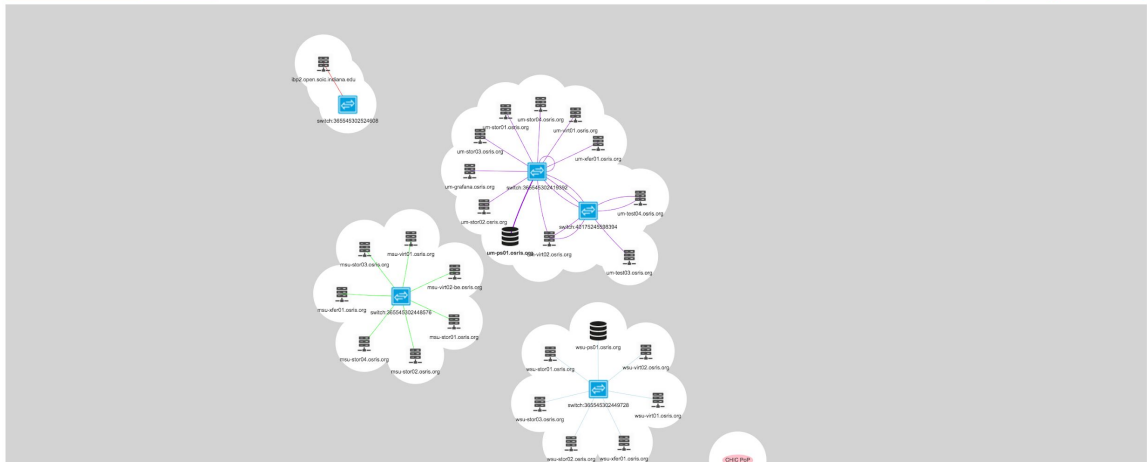
From SC18: <http://www.osris.org/article/2018/11/15/ceph-cache-tiering-demo-at-sc18>



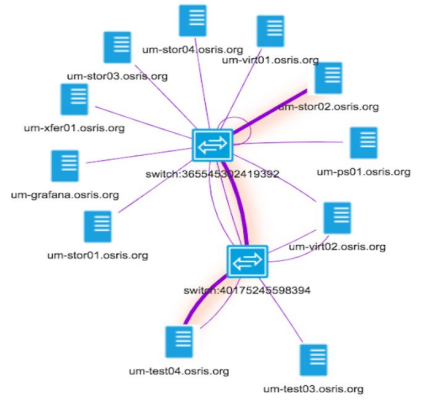
NMAL - Topology discovery (viz)

 OSiRIS Topology Monitor prototype version 0.2.0

SELECT TOPOLOGY CLUSTER CLEAR SHOW PATH



Visualization can also display computed paths through topology



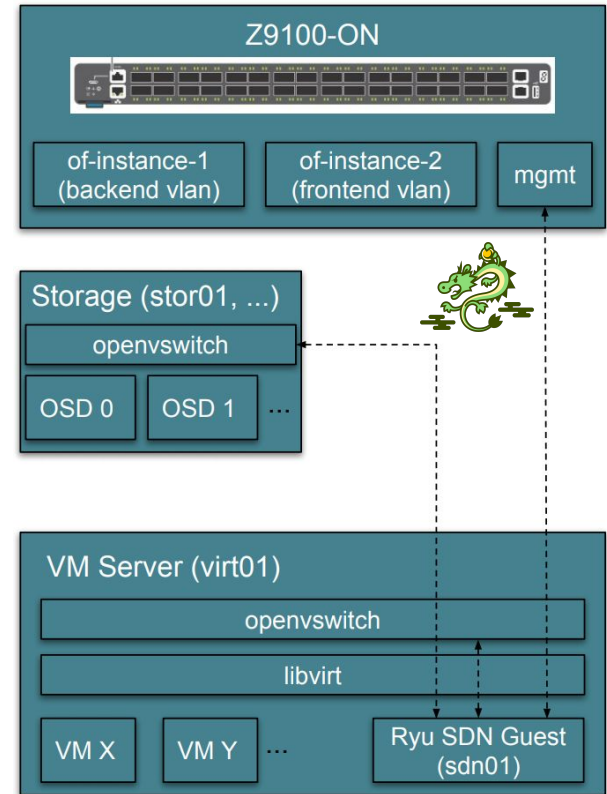
NMAL SDN Deployment - Ryu controller

Ryu SDN framework (<https://osrg.github.io/ryu/>)

Simple to deploy and integrate with our Python-based tools

Ryu in a VM through OVS required some planning to **separate control from dataplane** with common physical LAG on all hosts

Services managed by Puppet



NMAL - Bringing topology and monitoring together

Regular perfSONAR testing (via MCA mesh) results are exposed via topology visualization

The goal is to create an OSiRIS **monitoring dashboard** to quickly analyze and troubleshoot performance issues

Near-term plan:

Integrate passive measurements

Make topology elements dynamically update to highlight **current network and device state**

Long-term goal:

Integrate analysis engine based on UNIS-RT to perform **change point detection** and reporting

SOURCE	DESTINATION	VIEW	THROUGHPUT (avg)	LOSS	QW-LATENCY (ms)
um-pd01-be.osiris.org - 141.211.124.135	wsu-pd01-be.osiris.org - 204.24.195.86	SHOW	7.69 gbit/s	0.00000 %	30
um-pd01-be.osiris.org - 141.211.124.135	msu-pd01-be.osiris.org - 207.73.217.74	SHOW	7.40 gbit/s	0.00000 %	68

SOURCE	DESTINATION	VIEW	THROUGHPUT (avg)	LOSS	QW-LATENCY (ms)
um-pd01.osiris.org - 141.211.169.7	uwpd01.osiris.org - 129.79.51.136	SHOW	0.54 gbit/s	0.00000 %	2
um-pd01.osiris.org - 141.211.169.7	wsu-pd01.mang.osiris.org - 141.211.136.3	SHOW	0.99 gbit/s	0.00000 %	2
um-pd01.osiris.org - 141.211.169.7	wsu-pd01.osiris.org - 204.24.195.2	SHOW	7.53 gbit/s	0.00000 %	46
um-pd01.osiris.org - 141.211.169.7	msu-pd01.osiris.org - 207.73.217.10	SHOW	7.45 gbit/s	0.00000 %	98
um-pd01.osiris.org - 141.211.169.7	207.73.217.74 - 207.73.217.74	SHOW	7.41 gbit/s	0.00000 %	31

