OSiRIS: A Distributed Storage and Networking Project Update

Shawn McKee[1]*, *Benjeman* Meekhof[1], *Ezra* Kissel[2], *Andrew* Keen[3], *Kenneth M.* Merz, Jr.[4], and *Micheal* Thompson[5]for the OSiRIS Project

[1]Physics Department, University of Michigan, Ann Arbor, MI, USA
[2]Department of Intelligent Systems Engineering, Indiana University, Bloomington, USA
[3]Institute for Cyber-Enabled Research, Michigan State University, East Lansing, USA
[4]Department of Chemistry and the Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, USA
[5]Computing and Information Technology Department, Wayne State University, Detroit, USA

Abstract. We report on the status of the OSiRIS project (NSF Award #1541335, UM, IU, MSU and WSU) after its fourth year. OSiRIS is delivering a distributed Ceph storage infrastructure coupled together with software-defined networking to support multiple science domains across Michigan's three largest research universities as well as the Van Andel Institute. The project's goal is to provide a single scalable, distributed storage infrastructure that allows researchers at each campus to work collaboratively with other researchers across campus or across institutions. The NSF CC*DNI DIBBs program which funded OSiRIS is seeking solutions to the challenges of multi-institutional collaborations involving large amounts of data and we are exploring the creative use of Ceph and networking to address those challenges. We will present details on the current status of the project and its various science domain users and use-cases. In the presentation we will cover the various design choices, configuration and the tuning and operational challenges we have encountered in providing a multi-institutional Ceph deployment interconnected by a monitored, programmable network fabric. We will conclude with our plans for the final year of the project and its longer term outlook.

1 Introduction

The OSiRIS project[1] has successfully connected four campuses and a smaller edge site with a software defined networking and storage system that allows the seamless sharing of large datasets. By the end of the fourth year of the project, we have an established rapid deployment infrastructure, automated virtual organization provisioning, self-service user enrollment with delegated approval, and AAA (Authentication, Authorization, and Accounting) infrastructure allowing for role-based fine grained access to resources. Early in the project we completed an engagement with CTSC to externally evaluate our security model[2]. We manage our own access and usage of project resources using the very same system as our users. Automation and orchestration are fully leveraged for stable and well-monitored services. All of this is

*e-mail: smckee@umich.edu

linked to 13 PiB of Ceph storage accessible through gateway filesystem mounts or through scalable and universally accessible S3 protocols.

OSiRIS currently serves approximately 14 science virtual organizations housed across 6 US institutions or labs with collaborators worldwide. The scalable, flexible nature of OSiRIS leaves the door open for more users and more collaboration with other research platforms. The remainder of the paper will cover our accomplishments and future goals in detail.

2 The OSiRIS Project

OSiRIS (Open Storage Research InfraStructure) is a collaboration of scientists, computer engineers and technicians, network and storage researchers and information science professionals from University of Michigan/ARC-TS (UM), Michigan State University/iCER (MSU), Wayne State University (WSU), and Indiana University (IU) (focusing on SDN and network topology). Recently we have also collaborated with the Van Andel Institute (VAI) in Grand Rapids, MI to extend our Ceph cluster with a small, fast cache at their site which is backed by the much larger amount of storage at the primary sites.

We are one of four NSF "Campus Computing: Data, Networking, Innovation: Data Infrastructure Building Blocks" (CC*DNI DIBBs) projects funded in 2015. OSiRIS has been prototyping and evaluating a software-defined storage infrastructure, initially for our primary Michigan research universities, designed to support many science domains. 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.

A single scalable infrastructure is easier to build and maintain than isolated campus data silos. Data sharing, archiving, security, and life-cycle management can all be implemented under one infrastructure. At the same time, our architecture will allow the configuration for each research domain to be optimized for performance and resiliency.

3 Ceph in OSiRIS

Ceph is a distributed object storage system that gives us a robust open source platform to host scientific data used in multi-institutional collaborations. The core of Ceph is the Reliable Autonomic Distributed Object Store. RADOS is self healing, self manages replication, and has excellent scalability and performance[3]. RADOS supports multiple data interfaces including POSIX, S3 compatible object storage, and kernel block devices. Ceph has sophisticated allocation mapping using the Controlled Replication Under Scalable Hashing (CRUSH) algorithm to allow us to customize data placement by use-case and resources[4].

Our Ceph deployment is distributed across sites at WSU, MSU, UM, and VAI. We have also had several experiences distributing the deployment to sites geographically farther away with a slight loss of performance while retaining functionality[5] and/or augmenting the distributed deployment with local caches[6][7]. Ceph allows us to choose the level of replication among these sites based on the needs of participating science domains. Typically our highest level of data resiliency would be provided by having one or more replicas at each site. Ceph also has options for creating Erasure Coded data pools which provide configurable redundancy similar to RAID. Our most recent storage purchases were specified to increase overall node count at each site so that we will be better able to leverage EC functionality in Ceph[8].

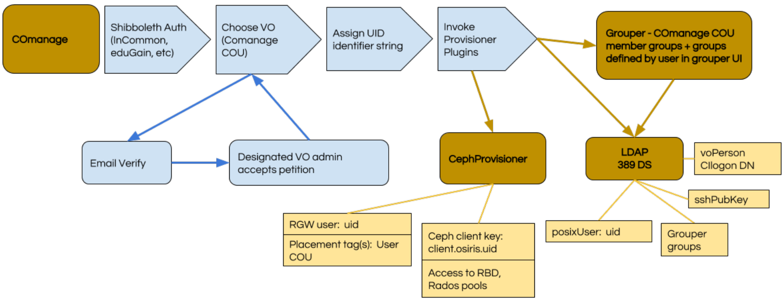
COmanage
Shibboleth Auth (InCommon, eduGain, etc)
Choose VO (Comanage COU)
Assign UID identifier string
Invoke Provisioner Plugins
Grouper - COmanage COU member groups + groups defined by user in grouper UI
Email Verify
Designated VO admin accepts petition
CephProvisioner
LDAP 389 DS
voPerson CIlogon DN
sshPubKey
RGW user: uid
Placement tag(s): User COU
Ceph client key: client.osiris.uid
Access to RBD, Rados pools
posixUser: uid
Grouper groups

Figure 1. COmanage flow for a new user authenticating to our gateway and joining a virtual organization. Following verification and approval the COmanage provisioning plugins translate identity and access information as appropriate for services including LDAP, Grouper, and Ceph.

4 OSiRIS Network Management Abstraction Layer (NMAL)

Another important part of the OSiRIS project is active network monitoring, management and network orchestration via the NMAL. Network topology and perfSONAR Periscope monitoring components deployed to hosts and switches to ensure that our distributed system can optimize the network for performance and resiliency through SDN (Software Defined Networking) control.

Main components in NMAP include BLiPP, UNIS, and an SDN controller.

- BLiPP - Basic Lightweight Periscope Probe. BLiPP agents may reside in both the end hosts (monitoring end-to-end network status) and dedicated diagnostic hosts inside networks.
- UNIS - Unified Network Information Store. The Periscope UNIS data store exposes a RESTful interface for information necessary to perform data logistics. The data store can hold measurements from BLiPP or network topology inferred through various agents.
- SDN Controller - Driven by information collected in UNIS, an SDN controller can dynamically modify network topologies to enable the best path between clients and data and between internal OSiRIS components (i.e., for Ceph replication).

5 OSiRIS Authentication, Authorization, and Auditing

The OSiRIS approach to authentication is to use identity federations and avoid managing authentication accounts. Federation participants will use their local providers to verify an identity and begin a self-service enrollment process which creates an OSiRIS identity belonging to one or more OSiRIS virtual organizations. We leverage InCommon[9] and eduGAIN[10] federations, Shibboleth[11], Grouper[12] and COmanage [13] to enable our science domain users to self-enroll, self-organize, and control access to their own storage within OSiRIS.

Virtual organizations are known as ’COU’ internally to COmanage (CO organizational units). Once a user is approved by designated VO admins or OSiRIS project admins provision their identity is established in COmanage and linked to their institutional identity. From there access to OSiRIS services is provisioned as shown in Figure 1. Access to service credentials is via the COmanage gateway. Should an individual move organizations we can simply link their new organization to existing OSiRIS identity. Multiple federated identities can be linked to a single OSiRIS identity as well.

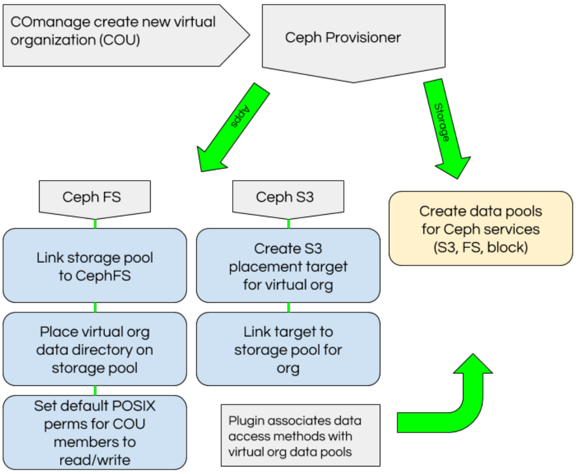
COmanage create new virtual organization (COU)
Ceph Provisioner
Apps
Storage
Ceph FS
Ceph S3
Create data pools for Ceph services (S3, FS, block)
Link storage pool to CephFS
Create S3 placement target for virtual org
Place virtual org data directory on storage pool
Link target to storage pool for org
Set default POSIX perms for COU members to read/write
Plugin associates data access methods with virtual org data pools

Figure 2. COmanage provisioning steps for Ceph VO storage include new pools for CephFS, S3, Rados, and steps to link those pools to directories or S3 placement locations

Virtual OSiRIS organizations can self-organize and manage members and roles via OSiRIS services such as COmanage and Grouper. They can further control access to data via service specific tools such as S3 ACL, Globus shares, or Posix permission tools.

5.0.1 COmanage Ceph Provisioner Plugin

COmanage has no built-in capability to provision users to Ceph, but it is designed with a plugin-based architecture. Different identity management events in COmanage trigger calls to configured plugins with information about the event. Events might include new user identifier, new user groups (new COU membership), request to reprovision a user, and more. OSiRIS created a new plugin under this architecture to handle provisioning storage for virtual orgs, link storage pools to CephFS directories or S3 placement targets, create users, and assign user capabilities. The plugin is freely available from our Github repository which is linked along with instructions on our website[14]. The structure of the plugin is shown in Figure 2.

5.0.2 Globus Gridmap LDAP Callout

OSiRIS provides Globus access to our data storage on CephFS and S3. We configure Globus to use CIlogon for authentication with Gridmap for certificate DN mapping to local S3 or POSIX users. However, we use the voPerson LDAP schema[15] to store certificate DN where Globus as-is relies on a text mapfile. For a brief period we used our own utility to generate the mapfile from LDAP[16] but thanks to the work of an undergraduate student at U-M we now use a Gridmap callout module which directly looks up the DN to username mapping in our LDAP directory. Documentation and source code for the module is available from the OSiRIS website[17].

6 OSiRIS Service Monitoring

In the past year we've made improvements to our monitoring and alerting infrastructure. We now consolidate node health, service health, performance metrics, ceph metrics, and alerts using Prometheus for metric gathering and alerting with Grafana for visualization. Our architecture is fully deployed and managed by Puppet. This also includes orchestration which

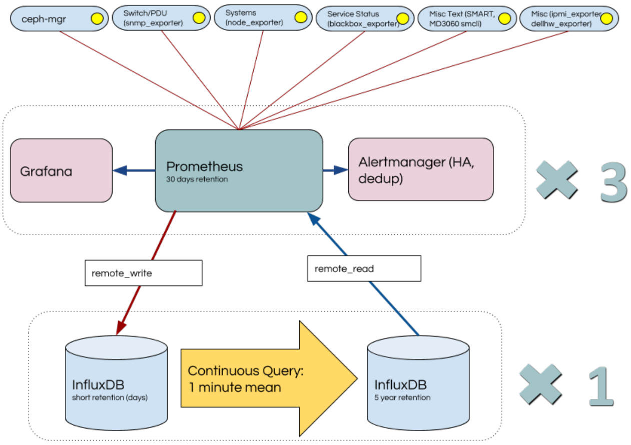

Figure 3. Structural diagram showing Prometheus instances at each OSiRIS site with a variety of metric feeds and long term storage in InfluxDB using remote read/write capabilities

defines new monitoring targets automatically. For example, every node we build automatically becomes a Prometheus target providing a range of performance metrics and monitored for basic functionality such as availability, ssh access, access to relevant services for type of node, etc. Through the use of Puppet 'exported resources' this configuration occurs without administrative action. A variety of alert rules notify us via email, Slack, or summary Grafana dashboards should any host or service have an issue. Our alert rules and dashboards are collected for reference in our Github repository[18].

We've deployed Prometheus (Figure 3) in a redundant configuration with an instance of the Prometheus server and Alertmanager at each site. Alertmanager is clustered for redundancy and alert deduplication. Prometheus is not well suited to long-term data storage so we continue to use InfluxDB to store and downsample data for the long term (currently 5 years).

7 Science Domain Engagements

OSiRIS plays an active role for multiple science domains and institutions. Some of these are just beginning to come up to full scale usage whereas as others have been incorporating OSiRIS since our first year. We also have discussions open with other projects such as the Open Science Network[19] and FABRIC[20]. Some recent engagements are highlighted below:

- Oakland University: Two groups at Oakland University in Rochester, Michigan are leveraging OSiRIS storage for their research. The Battistuzzi research lab focuses on long-term evolutionary patterns of microbial life, and the OU Genomics group aims to use and promote next-generation sequencing and bioinformatics technologies for research and education at the OU Biological Sciences Department.
- Brainlife.io: An online platform to accelerate scientific discovery by automated data management, large-scale analyses, and visualization. Brainlife plans to switch over to OSiRIS as their primary archival storage system before the end of this year.
- Building on existing collaboration between MSU and the Grand Rapids based Van Andel Institute, OSiRIS has installed NVMe-based Ceph OSD nodes running S3 instances to enable direct access to bioinformatics research data (Figure 4). OSIRIS at VAI will enable VAI bioinformaticians to work with MSU researchers to better understand Parkinson's disease and cancer. OSiRIS also facilitates data access for VAI researchers to leverage the computational resources at MSU's Institute for Cyber Enabled Research (iCER).

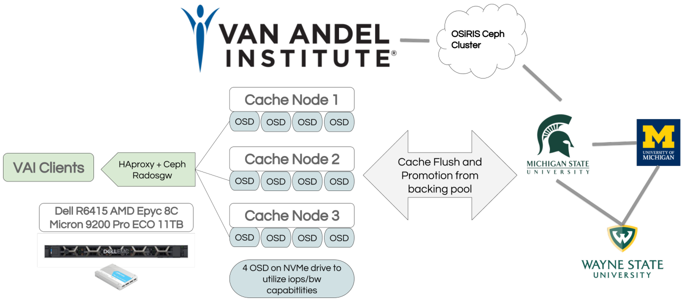

Figure 4. Van Andel Institute site architecture showing the arrangement of cache node hardware in relation to the OSiRIS cluster at other sites

Other users of OSiRIS include Naval Research Labs (tidal data), U-M Institute for Social Research (genomics studies), WSU Microscopy, Imaging & Cytometry Resources, the ATLAS experiment (physics event data), Global Nightlights at U-M (NOAA nightime imagery), Jetscape at WSU (heavy-ion physics event data), MSU Research Technology Support Facility, and Michigan Neuroimaging Initiative (NII).

8 Next Steps

To reach our goals the project faces a number of interesting challenges:

- Working with more scientific domains to leverage the strengths of OSiRIS as an worldwide-accessible object storage platform, especially interfacing with distributed compute/storage like OSG, XSEDE, Open Storage Network, and more.
- Building up a tool-kit of client interface options spanning from laptop to cluster systems. OSiRIS needs to be simple and easy to use even for those unfamiliar with object storage.
- Increasing the resiliency and scale of our object storage (S3) infrastructure to support IO at scale with no single points of failure. Ceph makes this kind of scaling is relatively straightforward in combination with commonly used open source components and only requires implementing resilient IP services on our existing proxy-backend architecture.
- Implementing software-defined networking (SDN) orchestration of both science-user and OSiRIS infrastructure network connectivity.
- Enabling science domain specific metrics to track, manage and optimize use of OSiRIS.
- Developing automated data life-cycle meta-data creation for users of OSiRIS.

9 Conclusions

The OSiRIS project goal is enabling scientists to collaborate on data easily and without building their own infrastructure. Scientists should be able to use our infrastructure by leveraging their existing institutional identities for authentication and self-management of resources. We aim not only to provide a scalable shared storage infrastructure, but to enable the most efficient use of that infrastructure with active network management via our NMAL layer. Users of OSiRIS should be able to get science done with their data instead of becoming bogged down in the details of data management and access.

10 Acknowledgements

We gratefully acknowledge the National Science Foundation (grant 1541335) which supported the work in this paper.

References

[1] S. McKee, B. Meekhof, C. Miller, E. Kissel, M. Swany, M. Gregorowicz, *OSiRIS: a distributed Ceph deployment using software defined networking for multi-institutional research*, in *J. Phys. Conf. Ser.* (2017), Vol. 898, p. 062045

[2] J. Basney, J.A. Candadai, T. Fleury, P. Gossman, M. Gregorowicz, S. Koranda, S. McKee, B. Meekhof, E. Kissel, Tech. rep. (2017), `http://hdl.handle.net/2022/21307`

[3] S.A. Weil, S.A. Brandt, E.L. Miller, D.D.E. Long, C. Maltzahn, *Ceph: A Scalable, High-Performance Distributed File System*, in *Proceedings of the 7th Symposium on Operating Systems Design and Implementation* (USENIX Association, USA, 2006), OSDI '06, p. 307–320, ISBN 1931971471

[4] S.A. Weil, S.A. Brandt, E.L. Miller, C. Maltzahn, *CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data*, in *Proceedings of the 2006 ACM/IEEE Conference on Supercomputing* (Association for Computing Machinery, New York, NY, USA, 2006), SC '06, p. 122–es, ISBN 0769527000, `https://doi.org/10.1145/1188455.1188582`

[5] OSiRIS, (2016), *Osiris at supercomputing 2016*, Retrieved from `http://www.osris.org/article/2016/11/18/OSiRIS-at-supercomputing-2016`

[6] OSiRIS, (2018), *Osiris at supercomputing 2018*, Retrieved from `http://www.osris.org/article/2018/11/19/OSiRIS-at-supercomputing-2018`

[7] OSiRIS, (2019), *Osiris at supercomputing 2019*, Retrieved from `http://www.osris.org/article/2019/11/18/OSiRIS-at-supercomputing-2019`

[8] OSiRIS, (2019), *New storage deployed*, Retrieved from `http://www.osris.org/article/2019/12/17/new-equipment-deployment`

[9] Internet2, (2017), *Incommon*, Retrieved from `https://www.incommon.org/`

[10] eduGAIN, (2019), *What is edugain*, Retrieved from `https://edugain.org/about-edugain/what-is-edugain`

[11] S. Cantor, (2005), *Shibboleth architecture: Protocols and profiles*, Retrieved from `https://wiki.shibboleth.net/confluence/download/attachments/2162702/internet2-mace-shibboleth-arch-protocols-200509.pdf`

[12] C. Hyzer, (2020), *Grouper software: An enterprise group and access management system*, Retrieved from `https://www.incommon.org/software/grouper/`

[13] S. Olshansky, (2017), *Comanage*, Retrieved from `https://spaces.internet2.edu/display/COmanage`

[14] B. Meekhof, (2018), *Ceph provisioner plugin for comanage*, Retrieved from `http://www.osris.org/components/cephprovisioner.html`

[15] NCSA, (2018), *voperson: Attribute management within a virtual organization*, Retrieved from `https://voperson.org`

[16] B. Meekhof, (2018), *ldaputil: Misc ldap utilities used in osiris*, Retrieved from `https://github.com/MI-OSiRIS/ldaputil`

[17] B. Meekhof, (2020), *Globus ldap gridmap callout*, Retrieved from `http://www.osris.org/components/globus_ldap`

[18] B. Meekhof, (2019), *Osiris monitoring contributions for grafana, prometheus, and more*, Retrieved from `https://github.com/MI-OSiRIS/OSiRIS-monitoring`

[19] A. Szalay, (2019), *The open storage network*, Retrieved from `https://www.openstoragenetwork.org`

[20] FABRIC, (2019), *Fabric: Adaptive programmable research infrastructure for computer science and science applications*, Retrieved from `https://fabric-testbed.net`