The OSiRIS Project: Collaborative Access to Data
CNSECCS Symposium – May 18, 2017

Project Outline
- rationale
- general goals

Project Components
- technical goals
- science users

Future Plans
- current science domains
- science domain roadmap
- technical focus areas

Community Benefits
- technical contributions
- scientific advancement
Those responsible...

The **OSiRIS** project engineering is coordinated by University Of Michigan Advanced Research Computing - Technology Services (**ARC-TS**).

### OSiRIS PIs
- **Shawn McKee** Lead PI University of Michigan
- **Patrick Gossman** Co-PI Wayne State University
- **Kenneth Merz** Co-PI Michigan State University
- **Martin Swany** Co-PI Indiana University

### OSiRIS Staff
- **Jayashree Candadai** IU Principal Research Engineer
- **Joseph Cottam** IU Research Scientist
- **Ezra Kissel** IU Lead Research Scientist
- **Jeremy Musser** IU Student Engineer
- **Jeff Goeke-Smith** MSU Network Engineer
- **Xiaoxing (Adele) Han** MSU OSiRIS Logo Design
- **Andy Keen** MSU Lead Engineer
- **Charlie Miller** MSU Engineer
- **Lillian Huang** UM Student Engineer (Summer 2016)
- **Jerome Kinlaw** UM Engineer
- **Dan Kirkland** UM Network Engineer
- **Benjeman Meekhof** UM Lead Engineer
- **Michael Gregorowicz** WSU Systems Architect
- **Matt Lessins** WSU Network Engineer
- **Carlo Musante** WSU Network Engineer
- **Michael Thompson** WSU Lead Engineer
THE PROBLEM: Storing, managing, transforming, sharing, and copying large research data sets is costly both in terms of dollars and time and impedes research and research collaboration.

OUR SOLUTION: Create an affordable and extensible, high-performance research storage cloud with properties not currently available commercially or in the open-source community. Create OSiRIS -- Open Storage Research InfraStructure.
OSiRIS Project Goals

- **Make data access and sharing easy** by creating a new form of data infrastructure with the following properties:
  - Direct access to storage across sites, eliminating the need for copy, compute, copy-back workflows.
  - Support for **block**, **object** and **filesystem** storage in the same cloud.
  - Streamlined data discovery, group management tools and underlying security to augment sharing and collaboration capabilities.
  - Automated network optimization and monitoring across storage sites to ensure quality of service.
- **Contain costs** - use off the shelf hardware and open source software.
- **Enable authentication/authorization** tools already in use at our campuses.
- Make it **easy to extend/replicate** the infrastructure regionally and nationally.
- **METAGOAL:** *Enable scientists to collaboratively pursue their research goals without requiring significant infrastructure expertise.*
OSiRIS Overview

- Logical View of OSiRIS Concept (from our Proposal)

**Science Domains**

**Multiple Interfaces**

**Distributed Ceph**
OSiRIS Site View

OSiRIS Data Infrastructure Building Block

Example

OSiRIS Site

Michigan State University

University of Michigan

Wayne State University

10 to 100G
More than one path
More than one wavelength

10G to 4x40G

GlobusOnline Data Transfer Node
PowerEdge R630

Ceph OSD Node (1 to 10 nodes)
PowerEdge R730xd

NVMe x 4
400-800GB Each

Head node has 384GB of ram, 40 logical processors, up to 4 NVMe for journaling, 4x25G NICs and two dual port SAS HBAs to connect the JBOD

Libvirt Hypervisor - R630, 256G ram, 2 x 800G NVMe (md raid1), 4x1TB (hw raid-10), 2x25G NIC

Virtualization Node (2 nodes)
PowerEdge R630

Ceph Mon and MDS VMs will be hosted here. Will also host OMD monitoring, development VMs, and science domain-specific VMs as required.
To implement the OSiRIS concepts requires an extensive amount of technical work in many areas:

- System provisioning and version control
- Ceph deployment, configuration and tuning
- Network monitoring, measuring and management
- Infrastructure monitoring and management
- Science-user management and interfacing
- Security and science domain organization and authorization

In the slides that follow I provide information about our efforts to-date framed in the context of specific goals, the status of the project, our results from our early adopters and our near and long term plans.
Project High Level Technical Goals

Provide templated provisioning for future OSiRIS-like deployments
- Tools to build, configure, manage services across diverse sites
- Optimizations to use Ceph across WAN deployments with good performance
- Documentation of components, services and design choices with examples

Develop OSiRIS Access Assertions (OAA) - Leverage existing institutional credentials via auth federations like InCommon for access to project resources including storage, network information, and other potential services

Develop Network Management Abstraction Layer (NMAL) - Extend perfSONAR/Leverage software defined networking (SDN) tools for efficient and flexible network orchestration
- Create tools that others can leverage
- Dynamic insight into network topology
- Using network topology to efficiently route data

Deploy detailed monitoring and auditing of services for internal troubleshooting, usage analysis, security and science user optimization.
**GOAL:** A distributed Ceph deployment as the basis of a storage infrastructure

- We implemented a multi-institutional Ceph storage platform reachable via Posix/CephFS mount, S3/Object store interface, and block device map.
- Ceph cluster has been operational for more than 1 year
- **Now:** 660 Ceph OSD (Object Storage Daemon; 1/disk), 5.2PB raw storage.
**GOAL:** Scalable and cooperative management of our deployment Puppet, Foreman with Git based workflow for project admins

- **Service Management**
- **Technical Goals**

- **GOAL:** Scalable and cooperative management of our deployment Puppet, Foreman with Git based workflow for project admins

- **Puppet Master**
  - Puppet Environments (Git Branches)
  - References Puppetfile to pull external modules into deployed environments

- **Github**
  - Branch for new work
  - Collaborators make pull requests to get work into production branch

- **R10K**

- **Query Available Environments**

- **Foreman is puppet ENC (External Node Classifier)**

- **Which environment is this node in?**

- **OSiRIS Node**

- **Compile and serve manifest for environment specified by Foreman**
**GOAL:** A framework to consistently seed future OSiRIS sites or manage nodes at non-OSiRIS sites

For non-OSiRIS sites lacking the 'Provisioning Proxy' component we can generate independent media to bootstrap them and connect to our management.
Technical Goals - Performance

**GOAL:** High performance for our Ceph deployment between participating campuses and for clients outside the project

User experience with Naval ocean model data: "Surprisingly the transfer from your machine to our machine (overseas) was not slower than locally moving the files between machines"

ATLAS S3 Object store testing: "Your system is responding much faster and with fewer errors. I would suggest letting others in USATLAS know of your settings."

Results so Far
**GOAL:** Understand how far can we stretch an OSiRIS deployment. Demo at SC16 augmented existing OSiRIS cluster with new site at Salt Lake City, Utah. Showed we could scale to ~42 ms RTT with acceptable performance.

As we added OSD in Salt Lake City the cluster was able to maintain good write and ‘backfill’ recovery of data replicas to the new OSD.
Technical Goals - Support Science Domains

**GOAL:** Capability to easily incorporate new science domain storage

- Our automation and provisioning tools make adding storage for new domains straightforward.

- We create **dedicated data pools** for each science domain and enable access to only the relevant pool(s) via access tokens.
  - Domains will control the visibility of their domain data (ie, browse directories in CephFS or access objects via S3) via OAA.
  - At this phase of the project we are easily able to manage manual administration of tokens but OAA will allow us to remove that burden and scale to larger numbers of self-managed domains and users.

- Our S3 object gateways run multiple separate instances of the “Rados Gateway” client
  - Each science domain is assigned its own instance(s) allowing us to tune/throttle each to fair-share resources and provide QoS.
**Technical Goals - Topology Discovery**

**GOAL**: Network topology discovery

- We have developed a discovery application integrated with the Ryu SDN controller
- Network resource information derived from both LLDP and SNMP
- We have extended the Periscope/UNIS schemas to support visualization and programmatic use of the topology within NMAL.
- **Next steps**: integrating techniques for capturing inter-site layer-3 topology
GOAL: Software Defined Networking (SDN) component control

- While our initial networking focus has been on understanding the topology and load within our networks, we are targeting the ability to orchestrate our network components to meet the needs of our OSiRIS users.
- We completed initial testing to ensure OpenFlow configuration works as expected across OSiRIS switching hardware and Open vSwitch instances.
- A Ryu library has been developed to inspect and modify OpenFlow device configuration, both programmatically and via a command line interface.
- Next steps: Integrate Ryu library with our UNIS runtime (UNIS-RT) to effect dynamic topology changes based on observed network state.

- Long term goal: The ability to orchestrate the use of all available network paths to simultaneously satisfy the requirements of multiple science domains and optimize the overall performance of the infrastructure.
Technical Goals - Identity

**GOAL:** Utilize federated campus identities to access OSiRIS and enable fine grained access control for services and resources.

- Developed **OSiRIS Access Assertions (OAA)** -- SAML compliant middleware to enable secure federated access to Ceph storage as well as secure access for network management.
- NSF’s **Center for Trusted Scientific Cyberinfrastructure (CTSC)** was engaged early in the OAA development process (July 2016)
  - CTSC strongly influenced the design of OAA
  - CTSC review of the completed software very positive
  - CTSC report published March 31, 2017 (review packet has a copy)
- **OAA** software published as open source and is expected to have much broader use than just with OSiRIS
- User interfaces and OSiRIS integration now in development.
**GOAL:** Science domains must be able to self-organize their own users and groups for Authentication, Authorization, Accounting (AAA) to Ceph and to network components

- NSF’s **CoManage** envisioned as the main interface to Ceph
- Multi-Master LDAP deployed March 2017 for **CoManage/OAA** framework
- User interfaces now under development
**Technical Goals: Securing OSiRIS Infrastructure**

**GOAL:** Securing and hardening OSiRIS infrastructure.

- All operating system installs use **SELinux** in enforcing mode.
- We install **fail2ban** ([www.fail2ban.org](http://www.fail2ban.org)) which blocks brute force attackers.
- **Network ACLs** in place to isolate backend network access
  - Backend network is for Ceph and administration use
- We only allow ssh login on the **backend** network except for site admin VMs (e.g., um-admin01).
- All logs are sent to a distributed central syslog using the **ELK stack** for monitoring and analysis.
Technical Goals - Monitoring

GOAL: Monitor and visualize OSiRIS storage, network, and related services.

Multi-faceted approach to monitor status, logs, and metrics. You can’t manage what you can’t see! We currently have deployed:

- Check_mk monitors host and service status

- Collectd/Influx/Grafana gather a variety of time-series data - network bytes/errors, Ceph service/cluster stats, host resources, etc

- Filebeat/Logstash/Elasticsearch ships, parses/normalizes logs, and indexes them for searching

Monitoring is an area that will evolve with the project. We anticipate adding new sources of monitoring data as well as improving our ability to benefit from our existing monitoring through targeted analytics.
**Check_mk** gives us a detailed overview of hosts and services.

Alerts sent when hosts or services not reachable, services cross thresholds (high CPU, too much disk, etc).

Alerts configurable by rules to alert groups depending on resource classifications.

Email or other alerts (we get them in Slack for example)
System and service specific logs are shipped to an elasticsearch cluster at UM, WSU, MSU.

Logstash pipeline lets us filter out specific strings later usable as time-series data for Grafana to visualize usage data from logs.

Kibana web gui provides flexible text searching and sorting for problem diagnosis.
Collectd/Grafana: Monitoring System Performance

System Performance Monitoring

**Collectd**
- Load
- CPU
- Disk

**InfluxDB**
- um-grafana
- msu-grafana
- wsu-grafana

Metrics are sent to Influxdb instances at each site where they are stored and downsampled over time. Influxdb internally supports transformations on data like derivatives, means, sums, etc.

**Periscope script**
Metrics are collected using collectd plugins. Periscope network data is collected from PS nodes using a python script.

**Grafana**
Grafana queries Influxdb for data with various operations applied (mean, sum, etc) and graphs the resulting time-series.

Sample: OSD op latency average for storage blocks from collectd-ceph

---

23
So far we have covered the technical work to enable OSiRIS. What is important is how all this enables us to support and expedite scientific research.

We have already begun engaging with two of our science domains to find out: physical ocean modeling and high-energy physics

These two were selected as early adopter communities for OSiRIS because:

- Both were willing to work collaboratively with the OSiRIS team to test and develop an initial storage infrastructure, understanding things might not start out at “production quality”
- Both had near-term needs for the capabilities OSiRIS was targeting
- Both were interested in use of “Object-store” capabilities

These two domains have helped us better understand the challenges and requirements needed to effectively support diverse science users.
What is the ATLAS use case for OSiRIS? ATLAS seeks to leverage commodity S3-style object storage such as OSiRIS. Our engagement includes providing assistance optimizing job pilots to efficiently access this type of storage.

OSiRIS event service gateways for ATLAS are balanced across 3 sites or optionally clients can target a specific OSiRIS gateway site.
So far our testing with ATLAS grid clients has resulted in improved tunings for the Ceph Rados Gateway (S3 provider) allowing an instance to process more simultaneous client ops.

40 requests / second

- Total number of connections - 18298
- Total number of Thread Exceptions - 2

rgw_thread_pool_size = 100
rgw_num_rados_handles = 1 (defaults)

150 requests / second

- Total number of connections - 58359
- Total number of Thread Exceptions - 273

rgw_thread_pool_size = 800
rgw_num_rados_handles = 8
ATLAS wants to work closely with people who know about object stores.

- ATLAS is especially interested in monitoring and visibility into how the object stores behave. We have a lot of expertise here.

- ATLAS is providing two people to drive the ATLAS use-cases for object stores but they want experts on the other end to work with.

- At a meeting at CERN with the ATLAS distributed data management team they specifically requested help from OSiRIS in the areas of monitoring and optimization of object stores.

- We participate in collab email list atlas-comp-event-service@cern.ch

- Join bi-weekly meetings of ATLAS event service and object store experts

- **GOAL:** Have OSiRIS as a “production” ATLAS event store by Fall 2017
US Navy transferred HYCOM tides data to OSiRIS to enable wider scientific collaboration. Storage is on CephFS posix filesystem accessible via direct mount or our general transfer gateways providing scp, FDT, and Globus.

Current users:
- Karlsruhe Institute of Technology (Germany) - Project: Temporal variation of tidal parameters. Funded by DFG.
- University of Michigan, collaboration with Navy (Arbic)
- University of Washington, collaboration with Navy
- SRI International (Ann Arbor, does work for ONR)
We will move this domain to self-managed OSiRIS Access Assertion system when fully implemented later this year.

Allows this community to define and manage their own membership

**Coming next:** Improving the ability to work *directly* with their data. We will work with them to enable direct mount of OSiRIS storage from relevant clusters instead of using transfer gateways to move data in/out of OSiRIS.

They have more users wanting to collaborate on existing data as well as plans to move significantly more data into OSiRIS (~100s of TB)

**Longer-term** we are discussing using S3 (object) storage as an indexed data retrieval method...the object namespace could reference objects by data resolution, time, location and/or type.
OSiRIS has a roadmap of science domain engagements planned beyond High-energy physics and High-Resolution Ocean Modeling.

As discussed in our proposal we will be engaging a diverse set of science “customers” who indicated their timelines for joining OSiRIS at our kick-off meeting in September 2015:

- **Aquatic Bio-Geochemistry** *(year-2)*
- **Biosocial methods and Population Studies** *(year-2)*
- **Neuro-degenerative Diseases** *(year-3)*
- **Bioinformatics** *(year-4)*
- **Genomics** *(year-4)*
- **Statistical Genetics** *(year-4)*
- Additional recruited science domains (years 2-5)

Since the proposal we have been contacted by Faculty from Biomedical Engineering and Mechanical Engineering (focus on turbulence and multiphase flows) asking to join OSiRIS and we intend to incorporate them as we have the effort available.
**Project Goal:** Understand the WAN-related limitations of Ceph; how far can we “stretch” a single Ceph cluster in terms of latency and what are the latency-related impacts?

Using ‘netem’ we simulated degrees of latency to a single storage block to quantify the effect on cluster throughput and ops capability: [http://www.osris.org/performance/latency](http://www.osris.org/performance/latency)

**Answer:** Unusable at 160 ms, recovers at 80 ms  
Our guess is 80 ms is a practical limit

---

Technical Contributions – Ceph Cache Testing

**Project Goal:** Explore Ceph’s caching capabilities to increase performance on a regional basis for science domains

Ceph Cache tiers can be ‘overlaid’ on top of distributed data pools so writes and some reads are redirected to storage elements closer to the users needing the data most.

Our tests found it can improve both latency and throughput for users near the cache.

- **Lower latency, similar to site-local pool**
- **Higher throughput than distributed pool**

Technical Focus - next year

Complete implementation of OSiRIS Access Assertions
- Integration with Ceph services
- Provisioning flow from Comanage sign-in with InCommon federation

Develop client bundle for CephFS Posix mounts
- Integrates with OAA credentials obtained via InCommon sign-in
- Packages OSiRIS Ceph configuration and client software
- Handles POSIX UID mapping
- Very likely to be a container-based release, Docker or Singularity
Initial implementation of NMAL components into “production”
- Self-adjusting conflict-free test meshes as new sites are added or existing site configurations are modified
- NMAL components integrated within OAA architecture to provide access restrictions
- All SDN resources enabled, discovered, and managed

Enable initial science domain onboarding automation process
- CoManage self-registration flow with institutional credentials
- OAA Resource self-provisioning (Ceph data pools, client access keys, etc)
Technical Focus - long term

Highlights of our longer term goals for the remaining years of the OSiRIS project

- **Network orchestration** - OSiRIS intends to discover and control the networks between OSiRIS resources and science users
  - This enables us to avoid network use conflict, steering traffic on distinct paths
  - We intend to shape traffic to optimize QoS for our users.

- **Data Lifecycle Management** - Working with our institutional library and information science colleagues we intend to automate the collection of suitable science-domain specific meta-data
  - The goal is to enrich the data being stored in OSiRIS, making it both discoverable and better managed.

- **Project Whitepaper** (How to Create your Own OSiRIS Deployment)
  - Details of how to replicate OSiRIS with tuning and deployment considerations and options explained.
OSiRIS is a 5-year project to develop and validate the concept of “multi-institutional, computable storage”. We anticipate a significant number of benefits.

- 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, conflict avoidance and resiliency.
- Ceph provides seamless rebalancing and expansion of the storage.
- A single, scalable infrastructure is much easier to build and maintain.
  - Allows institutions 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.
  - Configuration for each research domain can be optimized for performance and resiliency.

We have agreements from each of our institutions that they will continue to maintain and evolve their components of the infrastructure at the end of the project, assuming there are institutional science users benefiting from OSiRIS. Our goal is to benefit as many science domains as possible!
We have covered a lot of technical detail and explored OSiRIS from a high-level in this presentation.

Questions or comments?
Resources

OSiRIS Website:  
http://www.osris.org

OSiRIS Internal Wiki (InCommon sign in and authorization required, contact us):  
https://wiki.osris.org

OSiRIS Github Organization:  
https://github.com/MI-OSiRIS/

Periscope-PS Github Organization  
https://github.com/periscope-ps

Performance monitoring dashboards (read-only, public: kopakibana/OSiRIS#2):  
https://grafana.osris.org

Status monitoring dashboards (read-only, public: OSiRIS-guest/Guest)  
http://um-omd.osris.org/OSiRIS
Anticipated Project Challenges

- OSiRIS dynamic **SYSTEM optimization** to maintain a sufficient quality of service for all stakeholders.
- Infrastructure management capable of delivering a *production* system on *research* infrastructure.
- Enabling and automating the gathering and use of *metadata* to support data lifecycle management.
- Research domain customization toolkit using Ceph API and/or additional services.
- Network visibility and control using NMAL tools.
- Management of *quotas* and *ACLs* with Ceph and OSiRIS.
- Authorization which transparently integrates with each campuses existing systems as well as all OSiRIS services and components.