Multi-Institutional Open Storage Research InfraStructure

DOMA Workshop – November 16, 2017

Open Storage Research Infrastructure

Project Outline

- rationale
- general goals

Project Components

- technical details
- 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 (Sum '16)
My Do	UM Student Engineer (Sum '17)
Jerome Kinlaw	UM Engineer
Dan Kirkland	UM Network Engineer
Benjeman Meekhof	UM Lead Engineer
Michael Gregorowicz	WSU Security Architect
Matt Lessins	WSU Network Engineer
Carlo Musante	WSU Network Engineer
Michael Thompson	WSU Lead Engineer

Rationale for OSiRIS

- **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** -- **O**pen **S**torage **R**esearch **I**nfra**S**tructure.

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.

- META-GOAL: *Enable scientists to collaboratively pursue their research goals without requiring significant infrastructure expertise.*

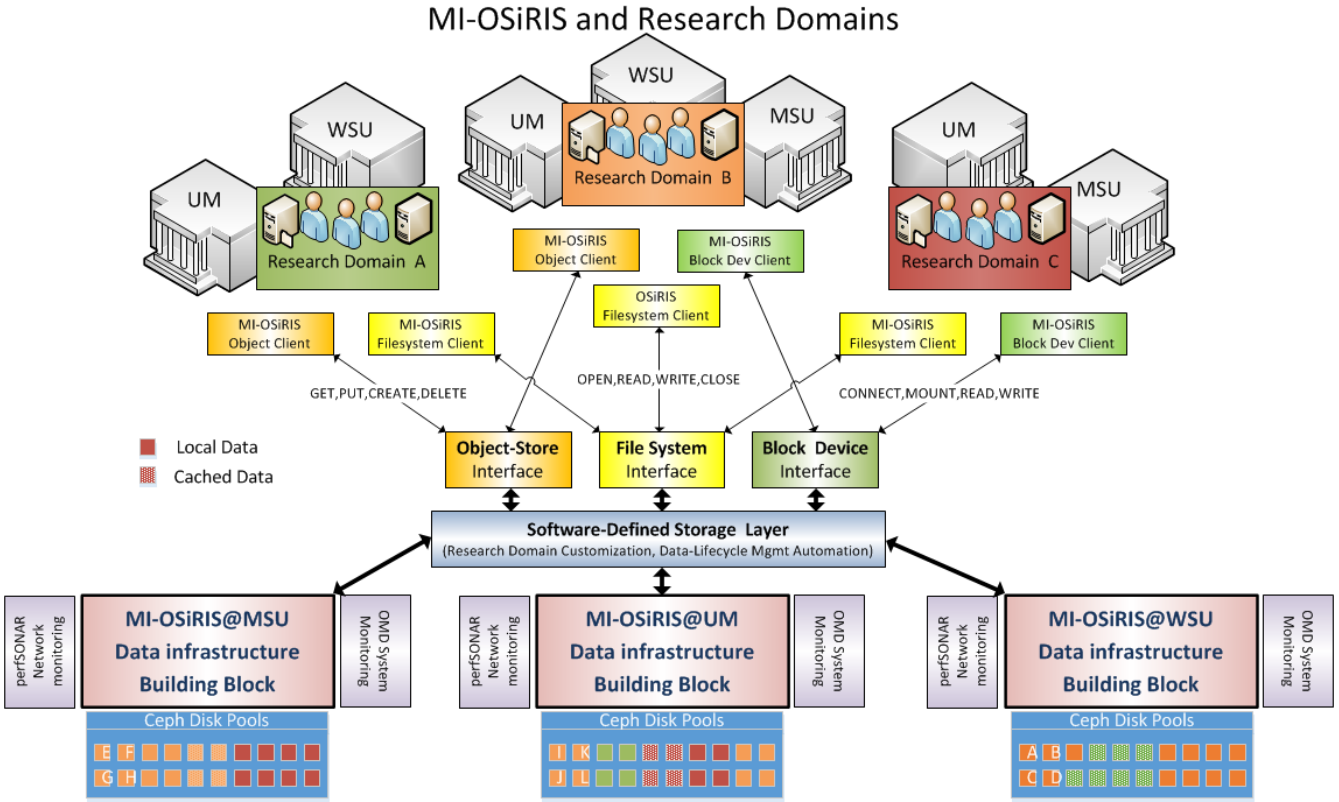
OSiRIS Overview

- Logical View of OSiRIS Concept (from our Proposal)

Science Domains

Multiple Interfaces

Distributed Ceph



OSiRIS Site View

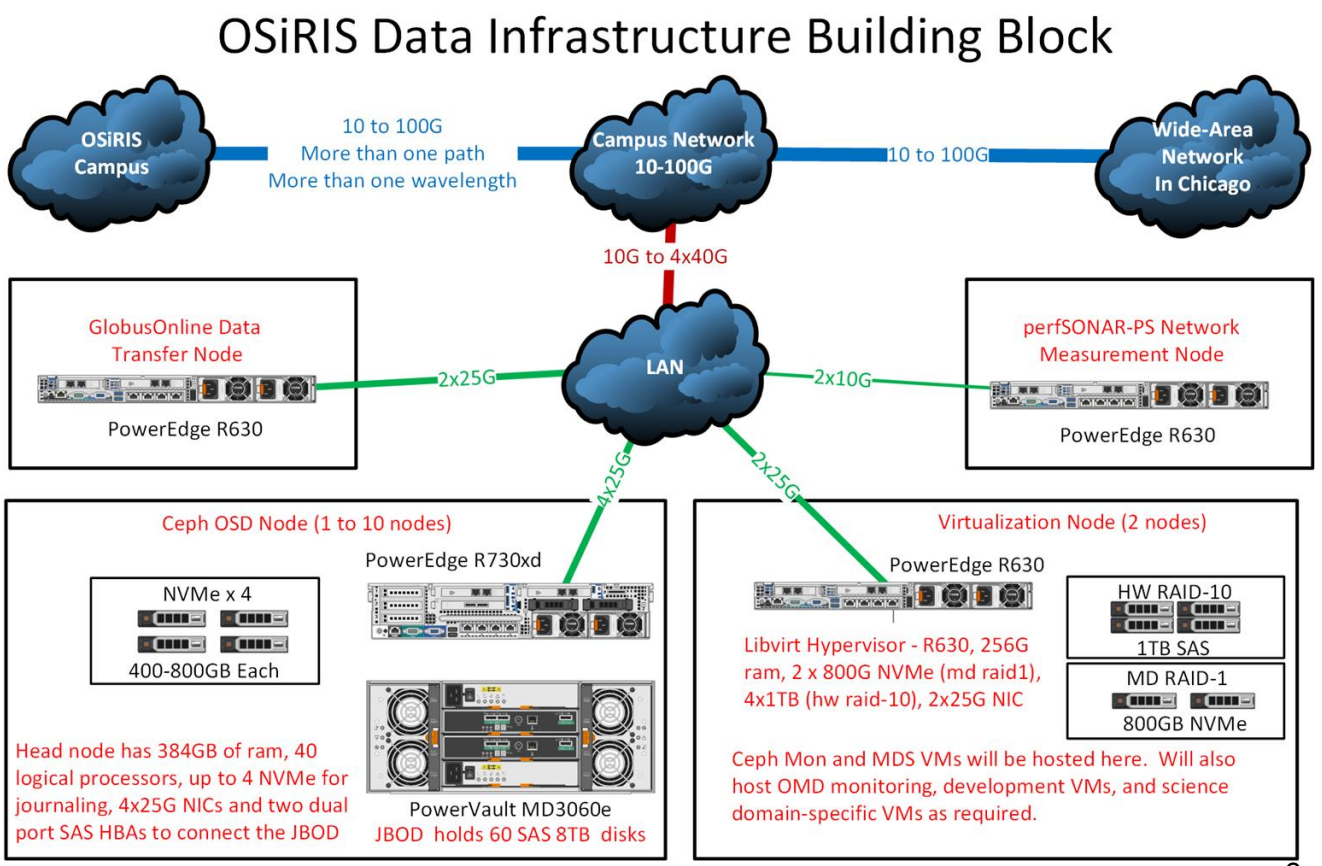
Example
OSiRIS Site

Michigan State University

University of Michigan

Wayne State University

OSiRIS now:
5.2 PB raw
660 OSDs



Scaling for the User

OSiRIS is designed to scale in a number of ways to support the needs of scientists collaborating on large, distributed or diverse data.

Ceph, with its multiple interface options, provides a robust and scalable basis to build a storage infrastructure upon

- Storage nodes can be easily added for increased capacity or performance
- Service components can scaled to both increase and control OSiRIS resource use

OSiRIS is designed to scale across institutions, providing a common storage platform able to support computing on the data, in place.

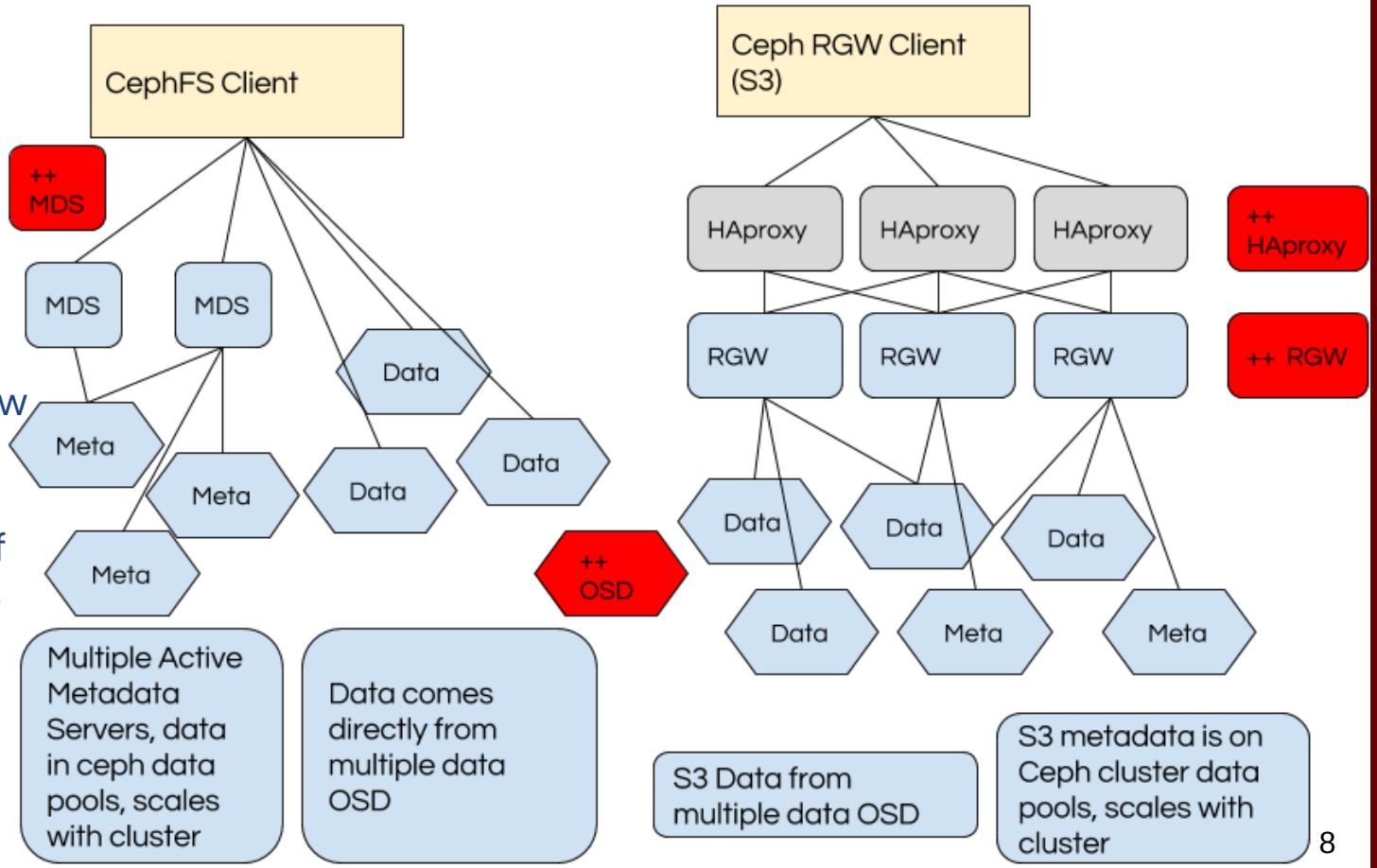
- Enabled by distributed or federated Ceph clusters coupled via SDN between institutions

Two important scaling areas for OSiRIS are covered in the next two slides

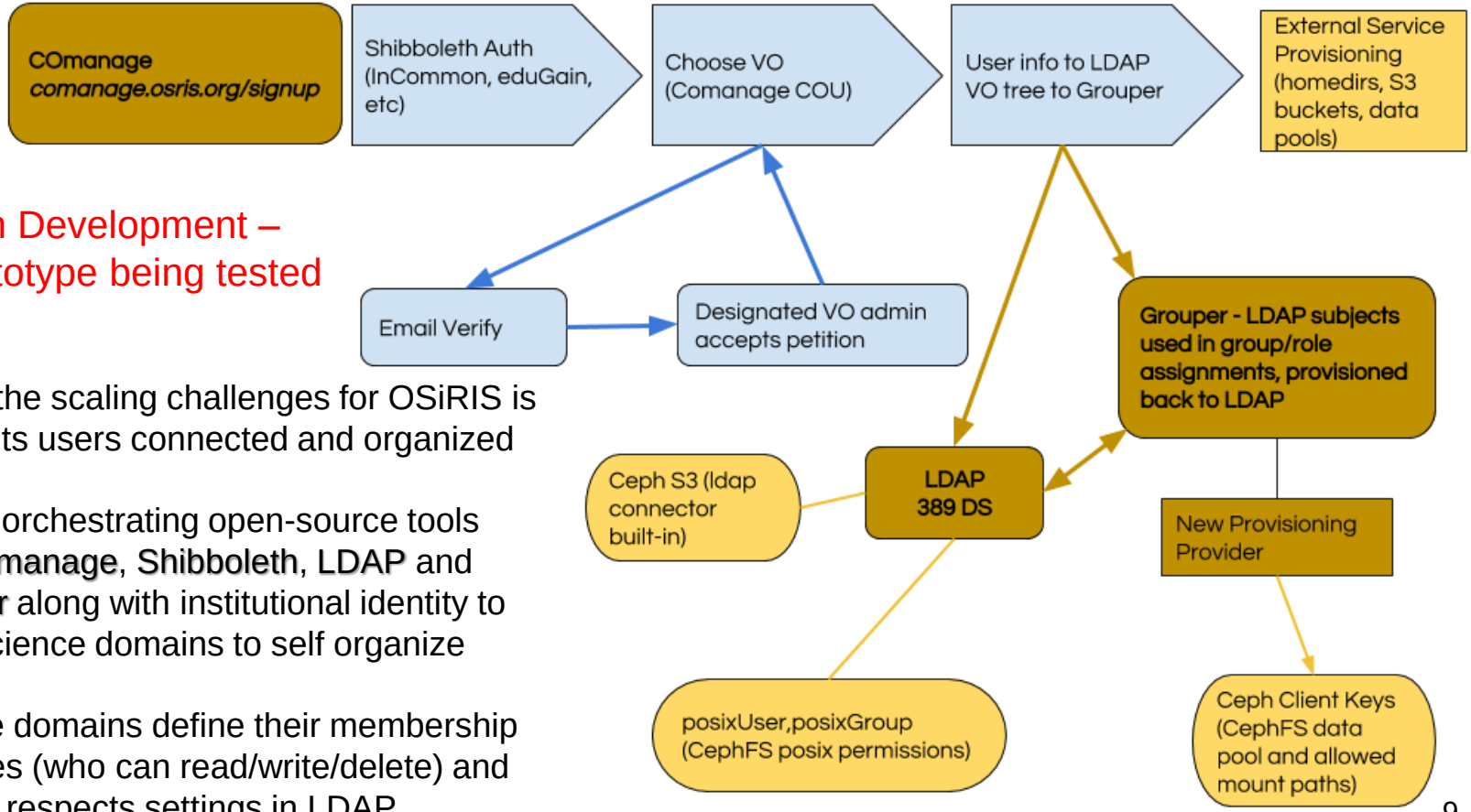
Scaling for the User: Storage Services

The red boxes indicate OSiRIS components we can easily scale out.

The storage services can grow as required to support the storage needs of the set of OSiRIS clients.



Scaling for the User: Automating Enrollment



In Development – Prototype being tested

One of the scaling challenges for OSiRIS is getting its users connected and organized

We are orchestrating open-source tools like COmanage, Shibboleth, LDAP and Grouper along with institutional identity to allow science domains to self organize

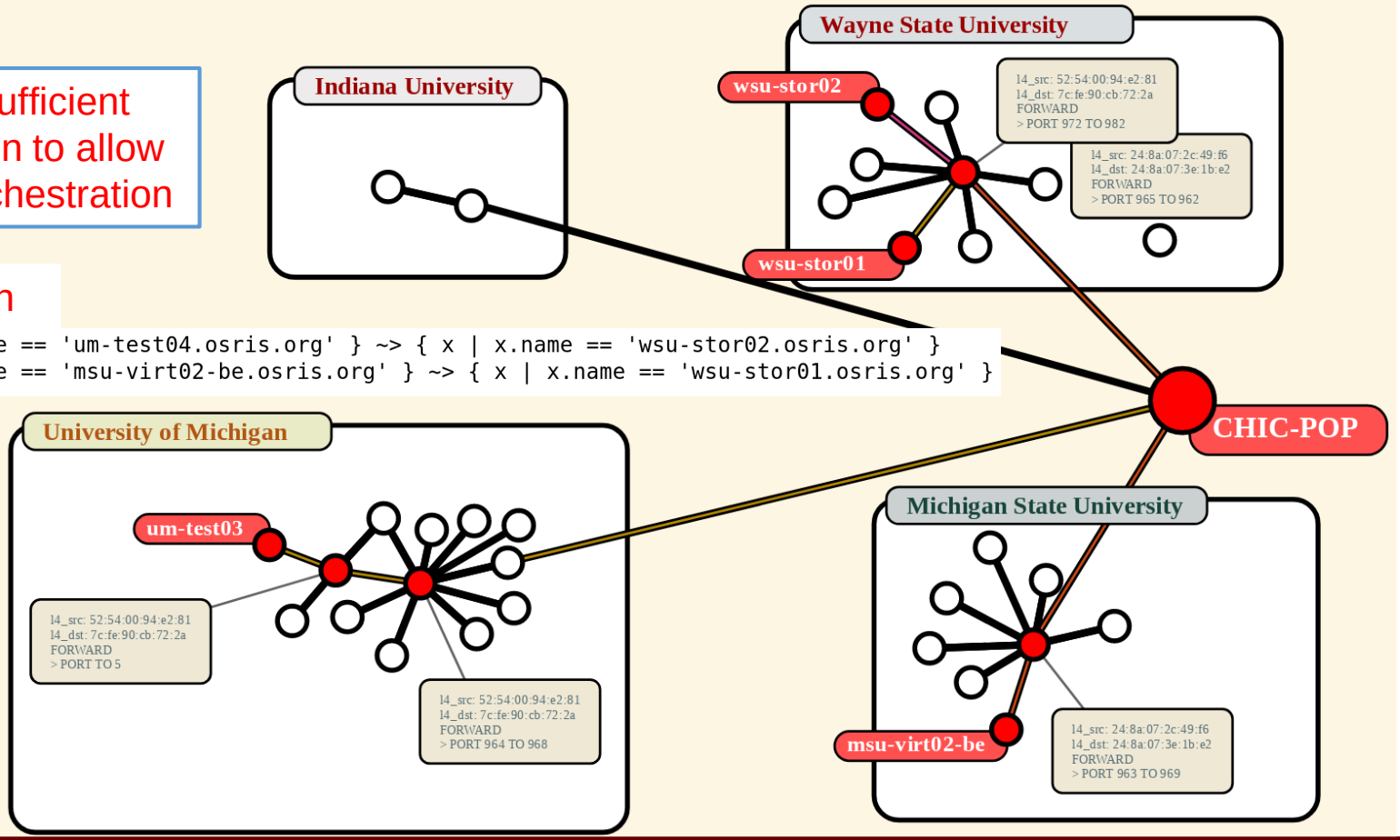
Science domains define their membership and roles (who can read/write/delete) and OSiRIS respects settings in LDAP

Network Topology Visualization and Path-finding

Goal is sufficient visualization to allow network orchestration

Configuration

```
exists { x | x.name == 'um-test04.osris.org' } ~> { x | x.name == 'wsu-stor02.osris.org' }
exists { x | x.name == 'msu-virt02-be.osris.org' } ~> { x | x.name == 'wsu-stor01.osris.org' }
```



OSiRIS Science Domain Engagement

So far we have covered the *technical* aspects of OSiRIS. What is important is how all this enables us to support and expedite scientific research.

We have been 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.

Science Domain Roadmap

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; beginning on-boarding)
- **Biosocial methods and Population Studies** (year-2 ; beginning on-boarding)
- **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.

Technical Focus - next year

Complete integration of authentication systems

- Integration with and automated provisioning of Ceph services
- Provisioning flow from COmanage sign-in with InCommon federation
- Self-management procedures for scientific users; possibly integration of Grouper: http://www.internet2.edu/products-services/trust-identity/grouper/

Develop client bundle for CephFS Posix mounts

- Packages OSiRIS Ceph configuration and client software
- Handles POSIX UID mapping (translation)
- Focusing on 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 / modified
- NMAL components integrated with authorization to provide access control
- All SDN resources enabled, discovered, and managed

Overview and Post-Project

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!

Conclusion / Discussion ?

We are exploring the usability of the combination of Ceph, SDN and institutional identity to support multiple science domains across multiple institutions.

We believe institutions can benefit from a versatile, compute-in-place storage platform that allows easy lateral scaling.

Thank-you!

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

Backup Slides

Backup slides follow

Technical Focus - long term

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

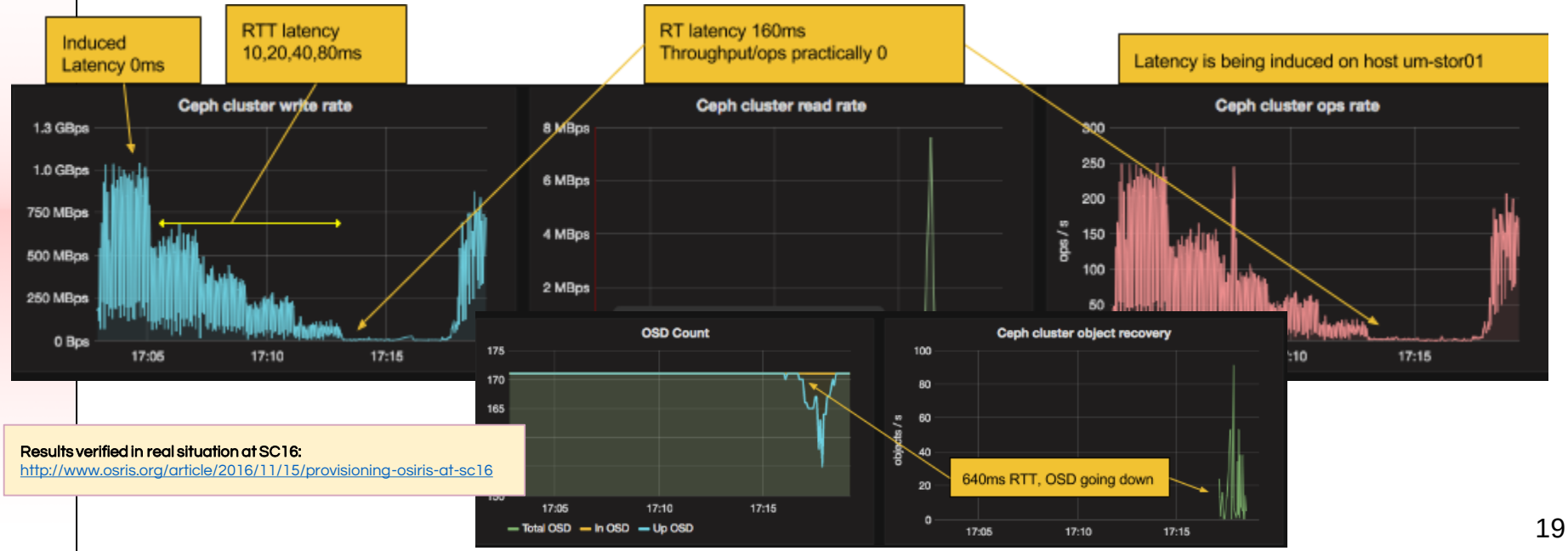
- **Network orchestration** - OSiRIS intends to discover and control the networks between OSiRIS resources and science users
  - This enables us to avoid network use conflict, steering traffic on distinct paths
  - We intend to shape traffic to optimize QoS for our users.
- **Data Lifecycle Management**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

Technical Contributions – Ceph Latency Tests

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

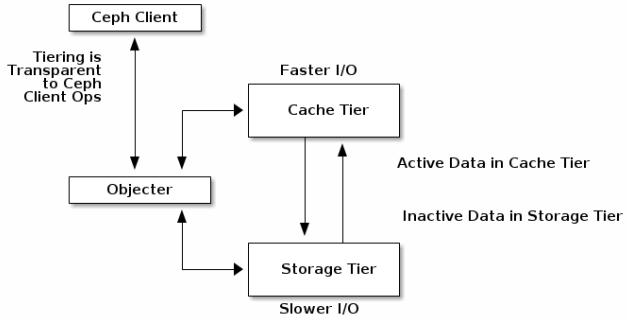
Answer: Unusable at **160 ms**, recovers at **80 ms** **Our guess is 80 ms is a practical limit**



Results verified in real situation at SC16:
http://www.osris.org/article/2016/11/15/provisioning-osiris-at-sc16

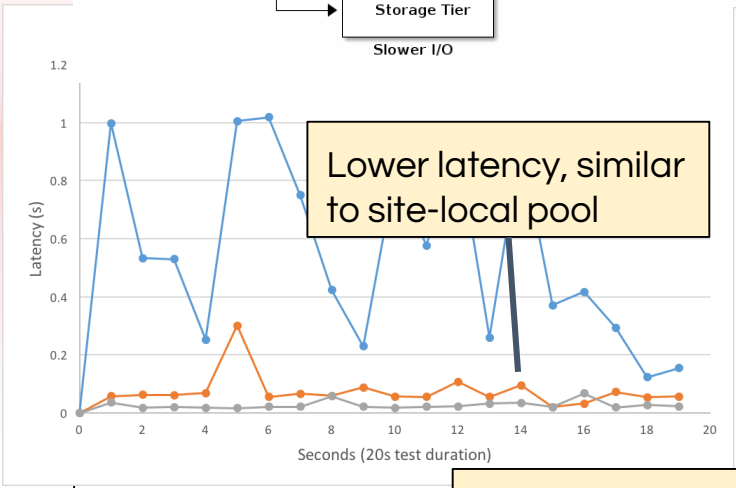
Technical Contributions – Ceph Cache Testing

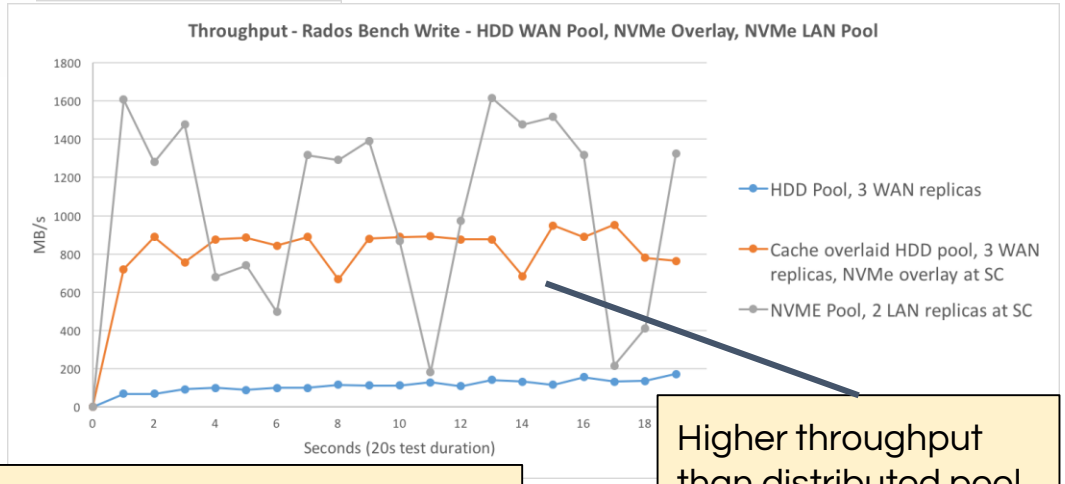
Project Goal: Explore Ceph's caching capabilities to increase performance on a regional basis for science domains



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*





http://www.osris.org/article/2016/11/16/ceph-cache-tiering-with-liqid-nvme-at-sc16

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.

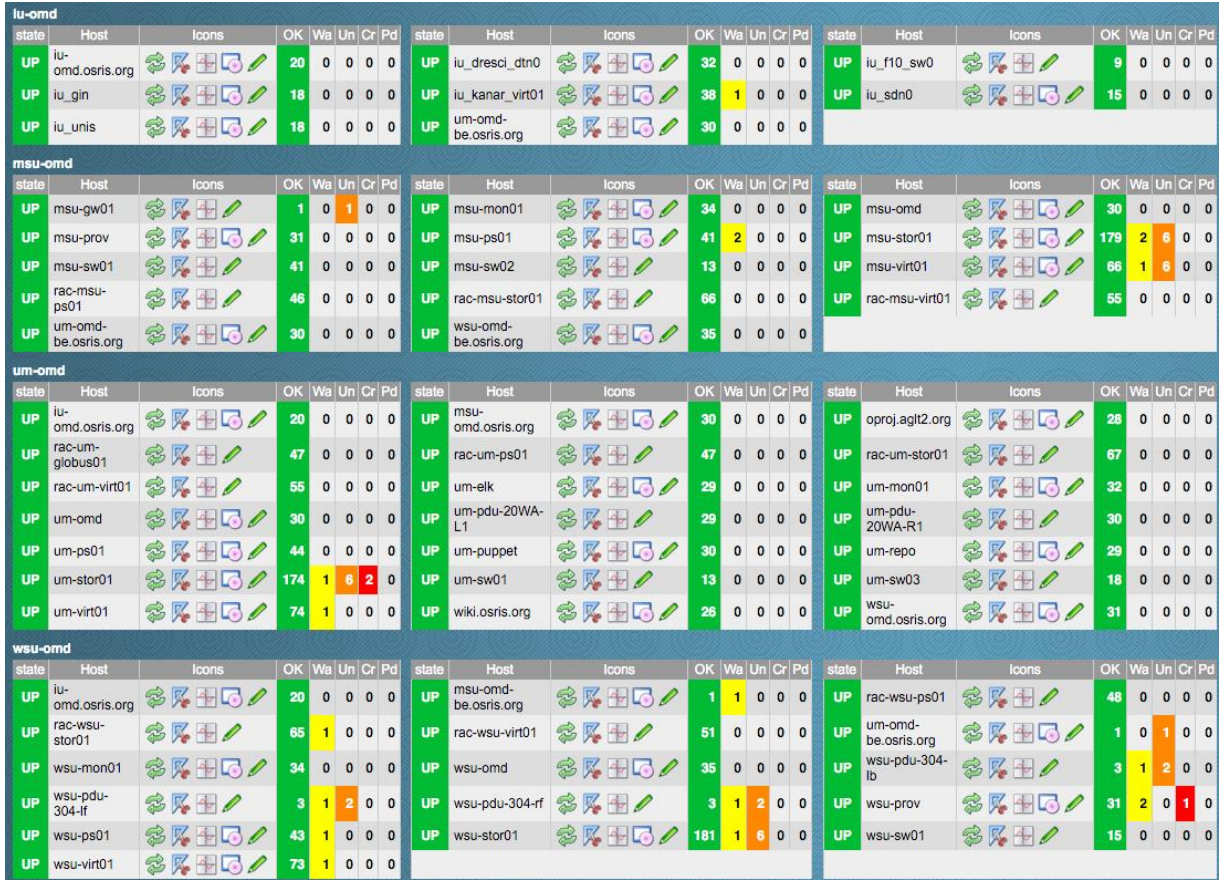
Check_mk Monitoring Services

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)



iu-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu-omd.osris.org		20	0	0	0	0
UP	iu_gin		18	0	0	0	0
UP	iu_unis		18	0	0	0	0

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu_dresci_dtn0		32	0	0	0	0
UP	iu_kanar_virt01		38	1	0	0	0
UP	um-omd-be.osris.org		30	0	0	0	0

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu_f10_sw0		9	0	0	0	0
UP	iu_sdn0		15	0	0	0	0

msu-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	msu-gw01		1	0	1	0	0
UP	msu-prov		31	0	0	0	0
UP	msu-sw01		41	0	0	0	0
UP	rac-msu-ps01		46	0	0	0	0
UP	um-omd-be.osris.org		30	0	0	0	0

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	msu-mon01		34	0	0	0	0
UP	msu-ps01		41	2	0	0	0
UP	msu-sw02		13	0	0	0	0
UP	rac-msu-stor01		66	0	0	0	0
UP	wsu-omd-be.osris.org		35	0	0	0	0

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	msu-omd		30	0	0	0	0
UP	msu-stor01		179	2	6	0	0
UP	msu-virt01		66	1	6	0	0
UP	rac-msu-virt01		55	0	0	0	0

um-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu-omd.osris.org		20	0	0	0	0
UP	rac-um-globus01		47	0	0	0	0
UP	rac-um-virt01		55	0	0	0	0
UP	um-omd		30	0	0	0	0
UP	um-ps01		44	0	0	0	0
UP	um-stor01		174	1	6	2	0
UP	um-virt01		74	1	0	0	0

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	msu-omd.osris.org		30	0	0	0	0
UP	rac-um-ps01		47	0	0	0	0
UP	um-elk		29	0	0	0	0
UP	um-pdu-20WA-L1		29	0	0	0	0
UP	um-puppet		30	0	0	0	0
UP	um-sw01		13	0	0	0	0
UP	wiki.osris.org		26	0	0	0	0

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	oproj.aglt2.org		28	0	0	0	0
UP	rac-um-stor01		67	0	0	0	0
UP	um-mon01		32	0	0	0	0
UP	um-pdu-20WA-R1		30	0	0	0	0
UP	um-repo		29	0	0	0	0
UP	um-sw03		18	0	0	0	0
UP	wsu-omd.osris.org		31	0	0	0	0

wsu-omd

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	iu-omd.osris.org		20	0	0	0	0
UP	rac-wsu-stor01		65	1	0	0	0
UP	wsu-mon01		34	0	0	0	0
UP	wsu-pdu-304-lf		3	1	2	0	0
UP	wsu-ps01		43	1	0	0	0
UP	wsu-virt01		73	1	0	0	0

state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	msu-omd-be.osris.org		1	1	0	0	0
UP	rac-wsu-virt01		51	0	0	0	0
UP	wsu-omd		35	0	0	0	0
UP	wsu-pdu-304-rf		3	1	2	0	0
UP	wsu-stor01		181	1	6	0	0

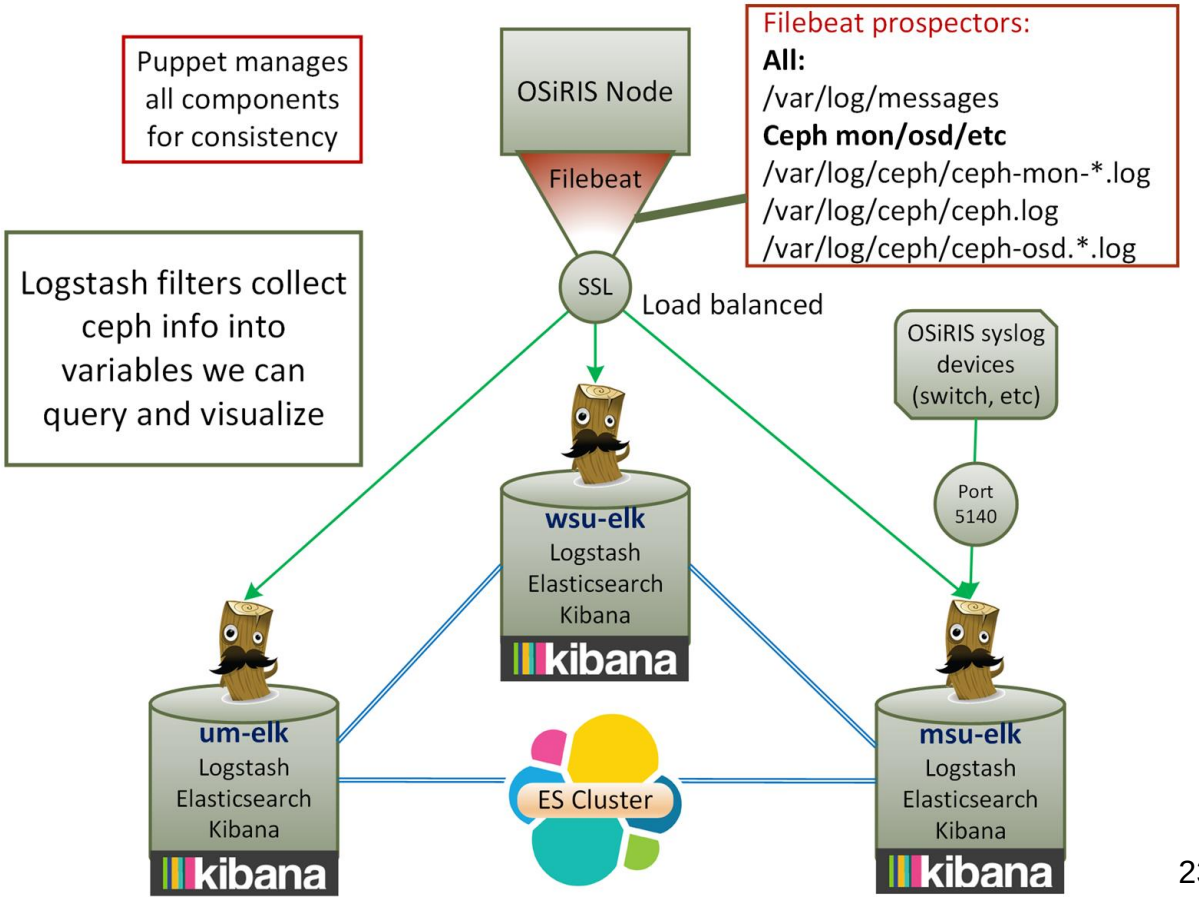
state	Host	Icons	OK	Wa	Un	Cr	Pd
UP	rac-wsu-ps01		48	0	0	0	0
UP	um-omd-be.osris.org		1	0	1	0	0
UP	wsu-pdu-304-lb		3	1	2	0	0
UP	wsu-prov		31	2	0	1	0
UP	wsu-sw01		15	0	0	0	0

ELK for Monitoring Logging Data

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
- Processes
- Memory
- CPU
- Network
- Ceph
- Disk
- Ethstat
- SNMP

Periscope script

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

InfluxDB

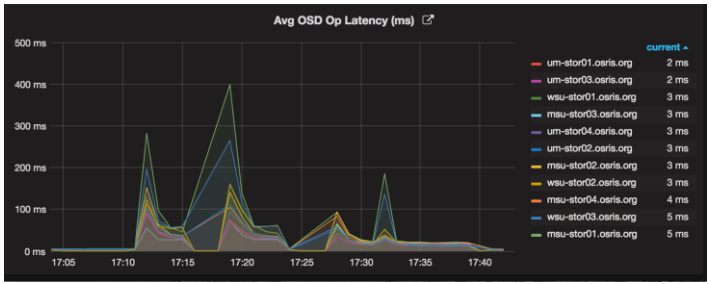
- um-grafana
- msu-grafana
- wsu-grafana

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.

Grafana

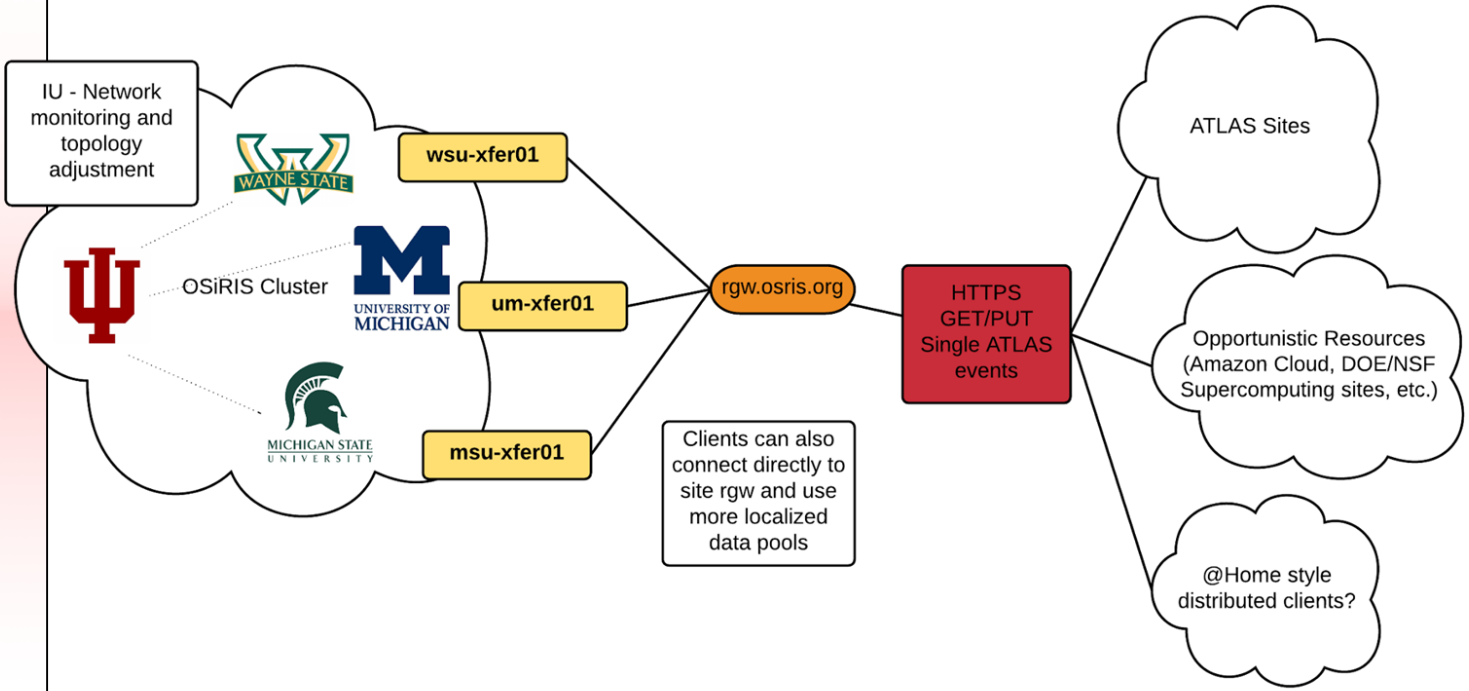
Grafana queries Influxdb for data with various operations applied (mean, sum, etc) and graphs the resulting time-series.

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



Science Domain - ATLAS

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

Next Steps - ATLAS

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
- Initial testing with ATLAS allowed us to increase our S3 IOPS performance by a factor of 3.5 via Ceph and OS tuning
- 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 MET**: OSiRIS is a "production" ATLAS event store in Nov 2017

Science Domain - Physical Ocean Modeling

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)

Next Steps - Physical Ocean Modeling

We will move this domain to our self-managed science domain enrollment and management system when fully implemented later this year.

Allows this community to completely define and manage their own membership and associated access control.

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

During this summer they added additional users interested in existing data and moved significantly more data into OSiRIS (~100 TB) for sharing.

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.