Data Lakes, Data Caching for Science and the OSIRIS Distributed Storage System

Open Storage Research Infrastructure

DOMA: Data Organization, Management and Access

The 1st Global Research Platform Workshop

September 18, 2019

Introduction

I believe one of the most challenging areas for any Global Research Platform will be developing cost-effective, performant methods for sharing and accessing data.

I have recently been involved in developing and testing capabilities in this area from both via my **OSiRIS** project and my participation in the **IRIS-HEP DOMA** area (both NSF funded)

In this talk I want to briefly cover OSiRIS and IRIS-HEP/DOMA and then discuss what we have learned so far and what we are working on.

IRIS-HEP

The Institute for Research and Innovation in Software in High Energy Physics (**IRIS-HEP**) project has been funded by National Science Foundation in the US as grant OAC-1836650 as of 1 September, 2018.

The institute focuses on preparing for **High Luminosity (HL) LHC** and is funded at **$5M** / year for 5 years. There are three primary development areas:

- Innovative algorithms for data reconstruction and triggering;
- Highly performant analysis systems that reduce time-to-insight & maximize HL-LHC physics potential;
- **D**ata **O**rganization, **M**anagement and **A**ccess systems for the community's upcoming Exabyte era.

The institute also funds the **LHC part of Open Science Grid, including the networking area** and will create a new integration path (the Scalable Systems Laboratory) to deliver its R&D activities into the distributed and scientific production infrastructures.   **Website for more info:** http://iris-hep.org/

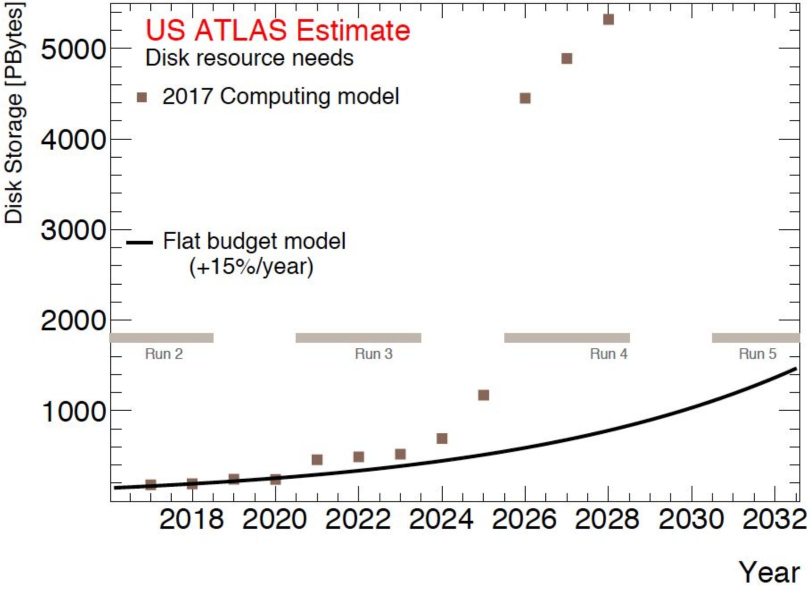
DOMA

IRIS-HEP has a number of areas of work but **DOMA** (**D**ata **O**rganization, **M**anagement and **A**ccess) is the most relevant for this talk

The **HL-LHC** data volume will exceed what our community can reasonably afford (even after accounting for technology) by a factor of 5-10

- We need to find ways to do more with the storage we will have
- Access to data and associated workflows need rethinking

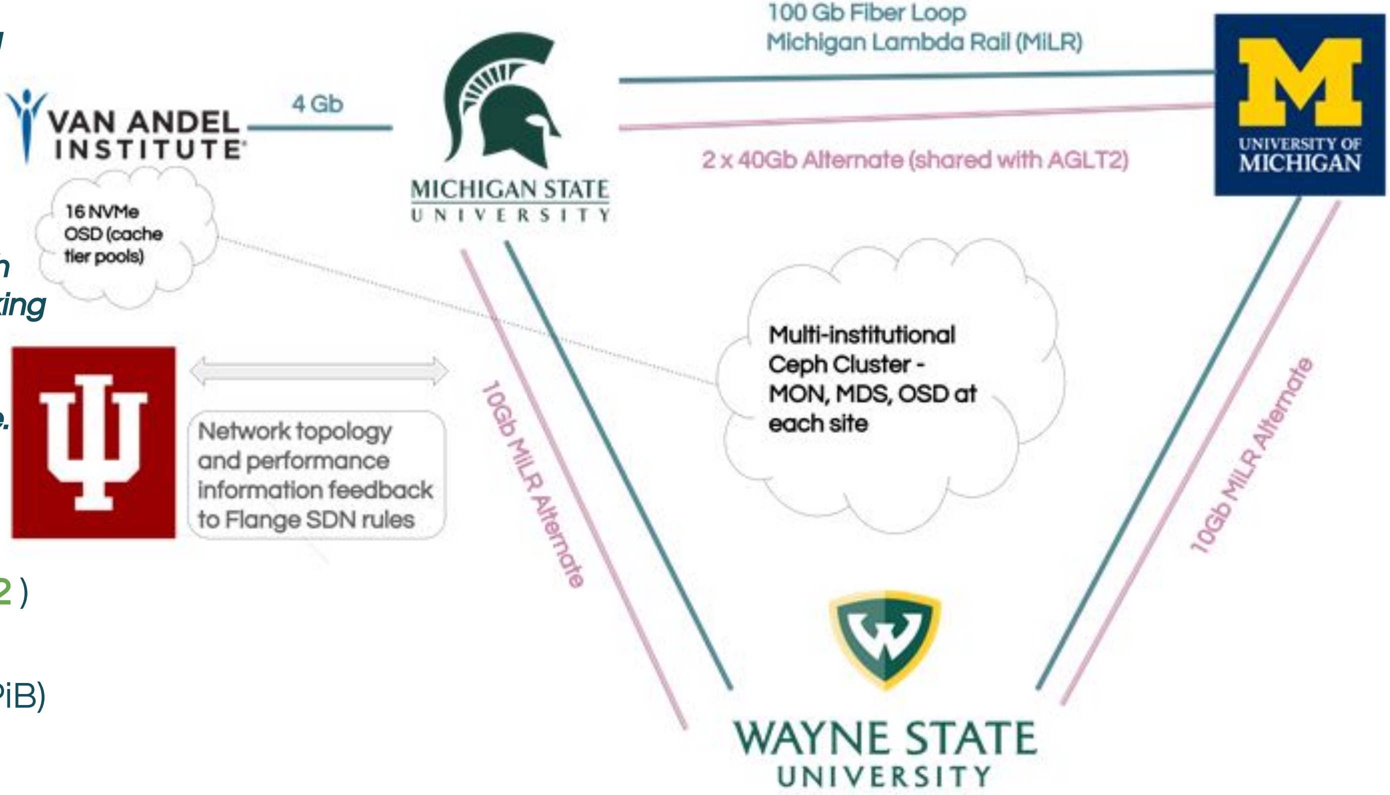
The DOMA starting point was a strawman called a **Data Lake**, a reorganization of our grid tiered hierarchy for our storage infrastructure. *More on this in a bit*



OSiRIS Overview

The OSiRIS proposal targeted the creation of ***a distributed storage infrastructure, built with inexpensive commercial off-the-shelf (COTS) hardware, combining the Ceph storage system with software defined networking to deliver a scalable infrastructure to support multi-institutional science.***

Current: Single Ceph cluster (Nautilus 14.2.2) spanning **UM, WSU, MSU** - **840** OSD / **7.4** PiB)



VAN ANDEL INSTITUTE
4 Gb
16 NVMe OSD (cache tier pools)
MICHIGAN STATE UNIVERSITY
100 Gb Fiber Loop
Michigan Lambda Rail (MiLR)
2 x 40Gb Alternate (shared with AGLT2)
UNIVERSITY OF MICHIGAN
Network topology and performance information feedback to Flange SDN rules
10Gb MiLR Alternate
Multi-institutional Ceph Cluster - MON, MDS, OSD at each site
10Gb MiLR Alternate
WAYNE STATE UNIVERSITY

OSiRIS Science Domains

The primary driver for OSiRIS was a set of science domains with either big data or multi-institutional challenges.

OSiRIS is supporting the following science domains:

- **ATLAS** (high-energy physics), **Bioinformatics**, **Jetscape** (nuclear physics), **Physical Ocean Modeling**, **Social Science** (via the Institute for Social Research), **Molecular Biology**, **Microscopy, Imaging & Cytometry Resources**, **Global Night-time Imaging**
- We are currently “on-boarding” new groups in **Multiphase Engineering Simulations** and **Cryo-EM**
- **Primary use-case is sharing working-access to data**

Summary of the OSiRIS Deployment

We have **deployed 7.4 pebibytes (PiB) of raw Ceph storage** across our **three research institutions** in the state of Michigan.

- Typical storage node is a 2U headnode and SAS attached 60 disk 5U shelf with either 8 TB or 10 TB disks
- Network connection is 4x25G links on two dual port cards
- Ceph components and services are virtualized
- **Year-4 hardware coming**: **33** new servers (11/site) adding 9.5 PiB (for EC)

The **OSiRIS infrastructure is monitored by Check_MK** and configuration control is provided by Puppet

Institutional identities are used to authenticate users and authorize their access via CoManage and Grouper

Augmented perfSONAR is used to **monitor and discover the networks** interconnecting our main science users.

Ideal Facilities

If we could have our way, we would have **ideal facilities**:

- CPUs would always be busy running science workflows
- Any data required would always be immediately available to the CPU when needed
- (Oh, and the facilities would be free and self-maintaining and use negligible power!)

As we all know, it is hard to create efficient infrastructures that manage access to large or distributed data effectively

Approaching “ideal” becomes very expensive (in $’s and effort)

So we need to make progress as best we can.

DOMA Data Lakes

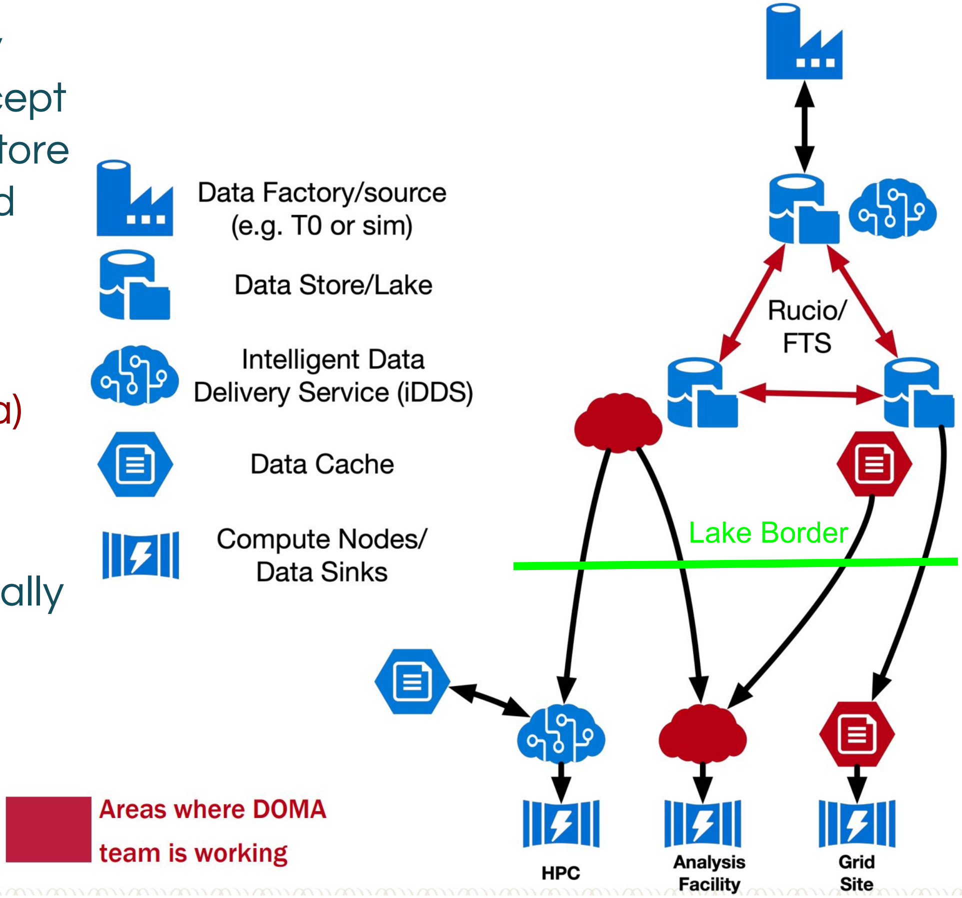
Data Lakes have been discussed in many contexts. Within DOMA we used the concept to provide a boundary between how we store and manage our data **cost-effectively** and how we access and consume that data.

The primary attributes

- Cost-effective (reduce copies of data)
- Provide internal intelligence
- Optimize QoS (performance/cost)

Our current HEP grid infrastructures typically uses various redundancy mechanisms at each site AND we store multiple copies across sites. This is expensive.

Data Lakes provide a way to optimize


Data Factory/source (e.g. T0 or sim)

Data Store/Lake

Intelligent Data Delivery Service (iDDS)

Data Cache

Compute Nodes/ Data Sinks

🟥 Areas where DOMA team is working

Caching and XCache

One of the critical methods to both reduce the amount of storage we use AND to provide acceptable performance is to utilize **caching**

- However, doing caching right is “hard”
- You need to carefully consider where to deploy caches and how to optimize them
- If you can’t afford a cache big enough to cover the work-set size you have challenges
- Managed caches take effort to operate well
- Caches that are transparent to clients and servers are best but may not be able to be easily optimized

While **GridFTP** is still our most used data transport mechanism, HEP has been exploring alternative protocols like **http** and **xrootd**.

Xrootd is a low latency storage protocol developed at SLAC and supports redirection

DOMA in conjunction with the **Xrootd** and **OSG** teams has developed XCache (think “Squid” for Xrootd) to help provide caching capability for our current grid and future data lake infrastructure.

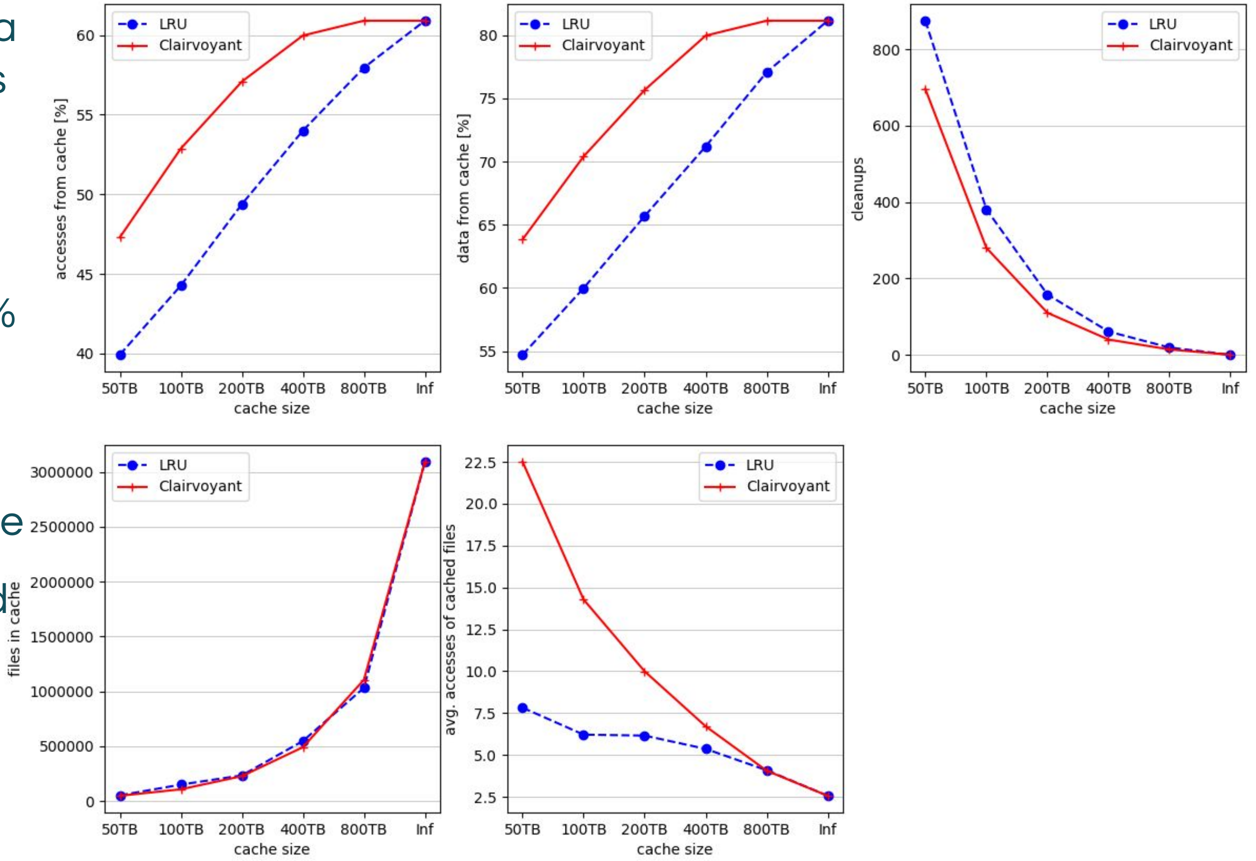
Simulating Caching Impact

We track all ATLAS file data for **each** site. This allows us to simulate the impact of a cache

- Full file caching
- Hi water: 95%, Lo: 85%
- Method either Least Recently Used or Clairvoyant (file with the longest time till use

In August 2018 MWT2 read 9.5 PB of files

40% of accesses and 55% of traffic could have been served from 50TB cache.

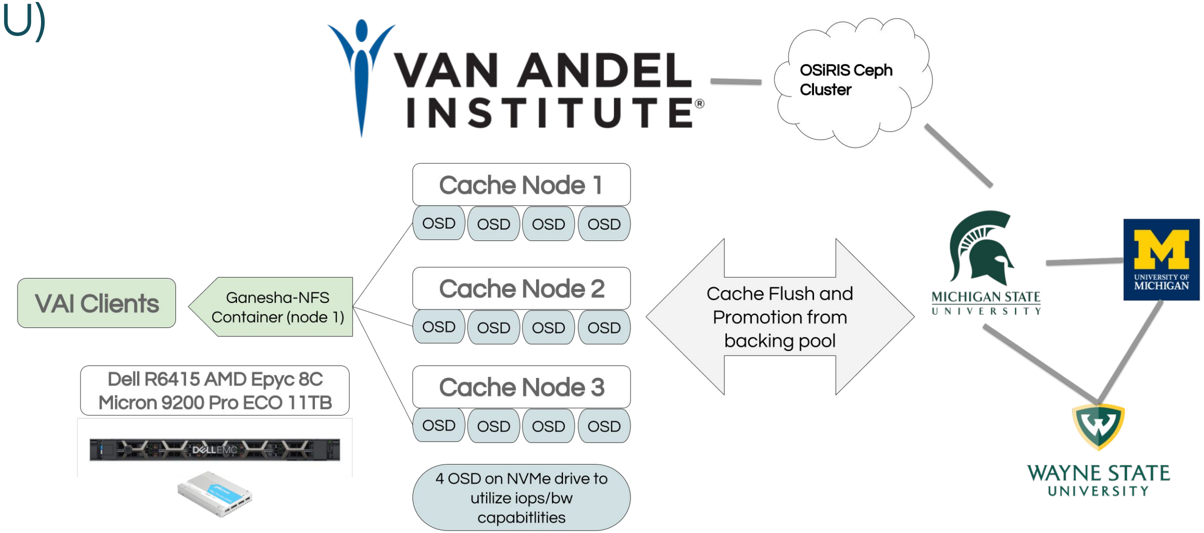


OSiRIS Ceph Cache Tiering

At Supercomputing conferences (2016/17/18) we've experimented with Ceph cache tiering to work around higher latency to core storage sites

- Deploy smaller edge storage elements which intercept reads/writes and flush or promote from backing storage as needed

Have edge OSiRIS site leveraging this technique at Van Andel Institute (primarily led by MSU)



VAN ANDEL INSTITUTE

OSiRIS Ceph Cluster

VAI Clients

Ganesha-NFS Container (node 1)

Dell R6415 AMD Epyc 8C
Micron 9200 Pro ECO 11TB

Cache Node 1

OSD OSD OSD OSD

Cache Node 2

OSD OSD OSD OSD

Cache Node 3

OSD OSD OSD OSD

4 OSD on NVMe drive to utilize iops/bw capabitlities

Cache Flush and Promotion from backing pool

MICHIGAN STATE UNIVERSITY

UNIVERSITY OF MICHIGAN

WAYNE STATE UNIVERSITY

OSiRIS Topology Discovery and Monitoring

UNIS-Runtime release integrated into ZOF-based discovery app

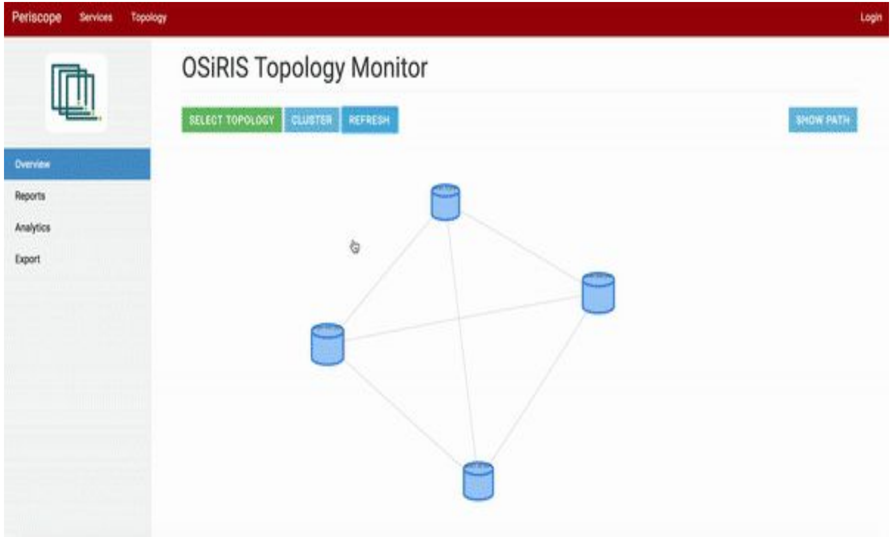
- Increased stability and ease of deployment
- Added extensions for Traceroute and SNMP polling

Web Development has focused on bringing measurements to dashboard

- Link and node highlighting with thresholds determined by link capacities
- Overlay for regular testing results to bring “at-a-glance” diagnostics

Filtering to show layer-2 topology versus layer-3 and virtualized components

- Fault localization, clustering, and zoom are work-in-progress



OSIRIS: Quality of Service for Ceph

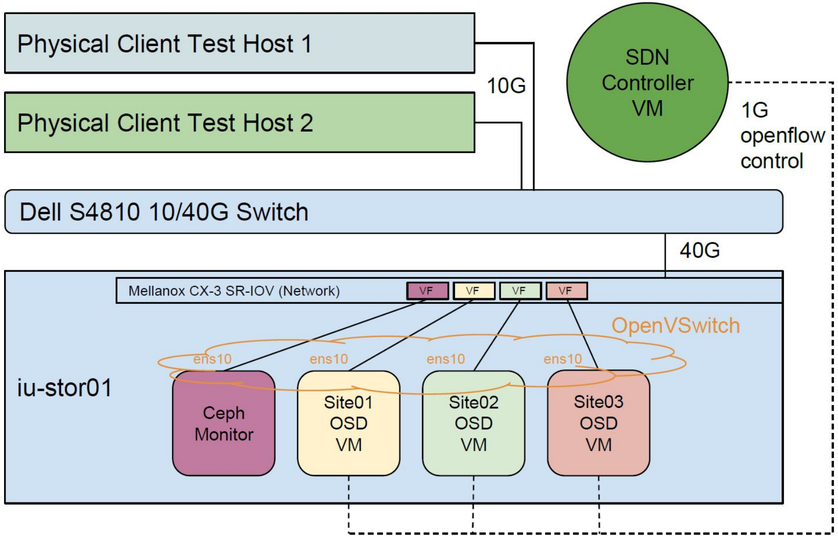
Testbed created to develop QoS functionality

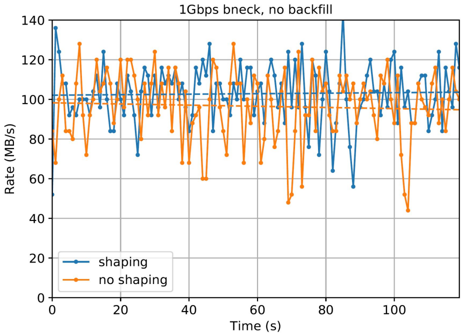
- Explicit control of operations, no noise
- Reduce risk of breaking production

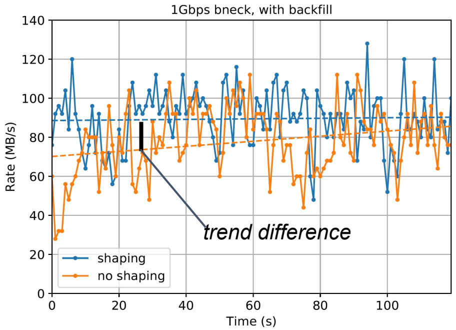
Apply priority queues to ensure that adequate bandwidth exists for Ceph client operations to prevent timeouts and delayed read/write performance

Apply traffic shaping to provide better transport protocol performance between sites with asymmetric link capacities. This is of particular importance when latency between sites is increased

Preliminary results: shaping from sites towards bottleneck can improve client performance, approx 5-10% in early testing.







OSiRIS Lesson's Learned

OSiRIS works very well on a regional scale (networking RTT ~< 10 ms)

We explored scaling for a single Ceph cluster at SC16 where we dynamically added a new site on the exhibition floor 42 ms RTT from the rest of OSiRIS

- The benchmark work-flow data access dropped from **1.2 GB/sec** to **0.45 GB/sec**
- The infrastructure continued to work without problems

Using 'netem' we were able to programmatically add arbitrary delay into the network stack of one of our Ceph servers.

- As we increased the latency we saw the expected impact in throughput
- When we reached **160 ms**, our (untuned) Ceph cluster stopped working
- We needed to decrease the latency back down to **80 ms** to recover

To reach more distributed deployments, OSiRIS would need to start using Ceph Federations (with associated costs) or employ caching to "hide" the latency as much as possible. In DOMA terms, OSiRIS would be appropriate as an element of a data lake.

DOMA Lesson’s Learned

Reworking the LHC HEP storage models to meet the requirements of the HL-LHC era will be a significant challenge.

The strawman of having a few data lakes to more fully optimize our use of storage while maintaining performance has allowed us to make significant progress

Reducing the amount of data we store is critical and we are working from the both ends: data model and organization as well as reducing the number of copies we keep

- We are able to effectively trade network use for storage by decreasing the number of copies and provide WAN access to specific data when needed
- Caching allows us to both recover performance for WAN data and reduce the amount of network bandwidth required for our typical workflows.

We are still in the early days for addressing our challenges for HL-LHC

Summary

There are some compelling challenges in assembling capabilities to support a GRP. In OSiRIS and IRIS-HEP we are trying to provide some components and information that could play a role in a future global infrastructure

Thanks to my OSiRIS, WLCG, IRIS-HEP/DOMA colleagues for providing content for this brief presentation

QUESTIONS?

This material is based upon work supported by the National Science Foundation under OAC-1541335 (OSiRIS) and OAC-1836650 (IRIS-HEP). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Further Information

OSIRIS

- http://www.osris.org project website
- Details in various presentations at http://www.osris.org/publications

IRIS-HEP

- https://iris-hep.org/ project website
- Details in various presentations at https://iris-hep.org/presentations/bymonth

DOMA

- https://iris-hep.org/doma.html sub-project website
- DOMA Presentations are available at the above URL

Some Caching studies

https://indico.cern.ch/event/770307/contributions/3301625/attachments/1807559/2952167/Scheduling_with_Virtual_Placement_for_Site_Jamboree.pdf

Additional Slides Follow

Backup Slides

OSiRIS NMAL Progress

Network Management Abstraction Layer (NMAL) capabilities

Admin and packaging:

- Deployed perfSONAR pSconfig Web Admin (PWA) to manage mesh configurations
- Migrated to latest perfSONAR Toolkit with Puppet code for rebuilds, replicate to central MA
- Refactored topology discovery application using ZOF to simplify code dependencies +
- All NMAL services packaged as Docker containers for simplified use on OSiRIS and elsewhere
- Technical documentation for NMAL components completed and published online

Feature developments:

- Demonstrated network orchestration service manipulating paths in SDN-driven topology +
- Exposed measurements and paths within topology and monitoring dashboard +
- Released runtime software that forms the basis of NMAL services (discovery, NOS, etc.) +
- Added quality of service (QoS) capabilities to SDN CLI tools
- Created QoS testbed at IU to develop traffic shaping and prioritization policies
- Extended IU network presence through collaboration with SLATE

+ Student-driven research efforts

NMAL Components

Unified Network Information Service (UNIS)

- Topology and measurement store

Measurement Store (MS)

- no-SQL DB for timeseries data

Flange Network Orchestration Service

- DSL and frontend for SDN control

UNIS-Runtime

- Library for interacting with UNIS and MS

Topology Discovery

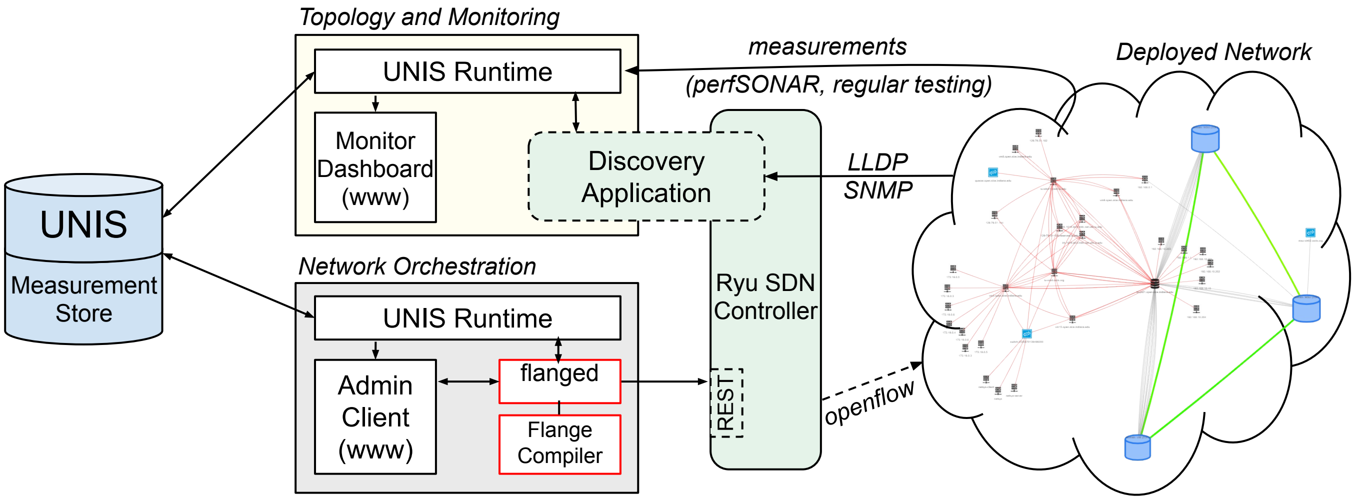
- Ryu/ZOF SDN Application: LLDP, SNMP, other

Periscope Web

- Dashboard for visualization

perfSONAR Deployments

- Regular testing data source



Topology and Monitoring
UNIS Runtime
Monitor Dashboard (www)
Discovery Application
UNIS
Measurement Store
Network Orchestration
UNIS Runtime
Admin Client (www)
flanged
Flange Compiler
REST
Ryu SDN Controller
measurements
(perfSONAR, regular testing)
LLDP
SNMP
openflow
Deployed Network