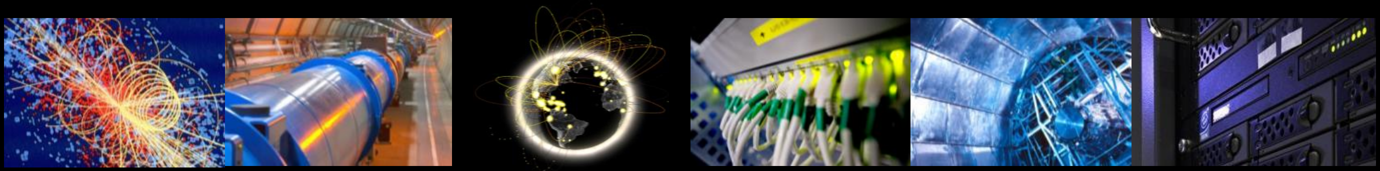
Scientific Data Lifecycle: Perspectives from an LHC Physicist

Shawn McKee, University of Michigan Physics Dept

For the MAGIC Meeting March 6, 2019



WLCG Worldwide LHC Computing Grid

ATLAS Great Lakes Tier 2 UNIVERSITY OF MICHIGAN AGLT2 MICHIGAN STATE UNIVERSITY

Open Science Grid

Introduction and Overview

Brief Introduction: I am a Research Scientist in the University of Michigan Physics department who has worked in the cyberinfrastructure space for ~25 years

- **Professional interests**: searching for dark matter using the ATLAS detector and creating new capabilities for enabling large-scale science

Today I want to talk about the challenges associated with distributed, data-intensive science and some of the experiences I have had working in this space.

The challenge for many such science domains is due to the growth in the volume, variety and velocity of the data they produce and the corresponding impact on the **network** and the **resulting requirements**.

I will start with some concepts and terminology, describe our tools and activities in LHC, discuss our challenges and problems and, finally cover the projects underway to address those challenges and problems.

Disclaimer: What is presented is my take on these issues and not official WLCG/ATLAS/CMS policy. Additionally I am not able to cover all the activities ongoing!

Data Life-Cycle Components

A good summary: https://datascience.columbia.edu/data-life-cycle

Generation->Collection->Processing->Storage->Management->
Analysis->Visualization->Interpretation

For LHC, here are the specifics:

Generation: For the LHC experiments, we generate Petabytes/sec of potential data

Collection: Because of cost and technical limitations, we collect only about a few Gigabytes/sec

Processing: Data from the online trigger system is sent to a cluster of computers for quick reconstruction, selection and formatting.. Data is tagged by its content and source

Storage: This is distributed: while CERN is the source, data is quickly distributed globally

Management: LHC experiments have setup distributed data management systems, to track, distribute and manage data

Analysis: The core of the physicist/s task is to transform, filter and analyze the relevant data for **new physics**

Visualization: This is required in many cases to understand the data and what it can tell us.

Interpretation: Here is where the “physics” happens. What does the data mean and what does it tell us?

LHC Data Lifecycle Challenges

The scale of **LHC** physics collaborations, ~3000 globally distributed physicists requiring access to 10’s of Petabytes of data, generates significant challenges.

How do physicists get access to the data? What systems, infrastructures and applications can they use to actually get to the “Interpretation” phase?

The initial answer was the LHC “grid” which was developed starting in 1999 and was ready for LHC turn-on in 2009.

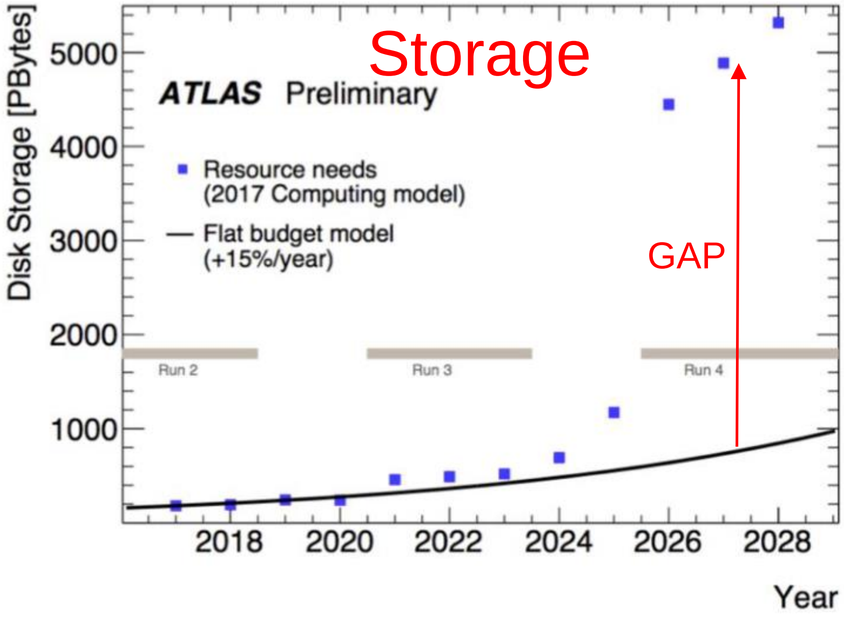
- The grid we constructed in the 2000-2010 was hierarchical, primarily x86 computers & commodity storage systems coupled together via R&E networks

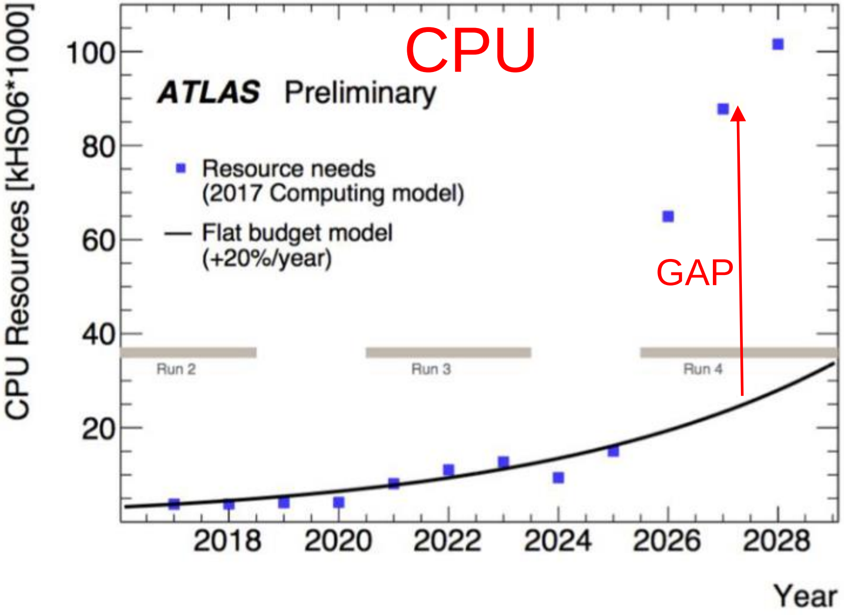
As technology marches forward and the LHC and associated experiments are upgraded, we are continually challenged to deliver an effective infrastructure

- Resources are evolving and being augmented with Cloud, HPC and new architectures (ARM, GPUs, FPGAs, SSD, NVMe, etc), SDN, etc

The HL-LHC Challenge

One daunting challenge coming in 2025: the needs for both **ATLAS** and **CMS** in the **HL-LHC era** are far beyond what we can expect to have assuming flat-budgets and (+15%/yr, +20%/yr) technology evolution.





How Are We Addressing Our Challenges?

There are numerous projects and working groups trying to both prepare for **LHC Run-3** and **High Luminosity (HL)-LHC**. I will cover the following to give you a flavor of the components (NOTE: this is biased towards projects I work on :))

Example LHC Components **NSF Research** **Large Projects**

AGLT2: My ATLAS Tier-2 (regional) center, to show an example "grid" site.

Rucio: The software ATLAS (and others) are using for distributed data mgmt.

OSiRIS: An NSF 5-year DIBBs award implementing a multi-institutional storage infrastructure, combining Ceph+SDN with data-lifecycle options

SLATE: AN NSF CIF21 DIBBs aware implementing "cyberinfrastructure as code", augmenting high bandwidth science networks with a programmable "underlayment" edge platform

OSG: Middleware and services for distributed science

IRIS-HEP: A new NSF institute exploring software/cyberinfrastructure for LHC

AGLT2 as an Example ATLAS/WLCG Site

ATLAS Great Lakes Tier 2
UNIVERSITY OF MICHIGAN AGLT2 MICHIGAN STATE UNIVERSITY

The ATLAS Great Lake Tier-2 (AGLT2) is a distributed LHC Tier-2 for ATLAS spanning between UM/Ann Arbor and MSU/East Lansing.

- **11176** logical cores, slots 1125 (dynamic) + 10 (static)
- Additional **936** Tier-3 job slots usable by Tier-2
- Average 10.55 HS06/slot
- **6.9 Petabytes** of storage
- Total of **117 kHS06**
- Tier-2 services virtualized in VMware 6.7

2x40 Gb inter-site connectivity, UM has **100G** to WAN, MSU has **10G** to WAN, lots of 10Gb internal ports and 20 x 40Gb ports, 32x100G/40G or 64x50G/25G ports

Middleware from OSG, WLCG, USATLAS Tier-2s have ~2 FTEs and provide MOUs for computing and storage with ATLAS. They all run opportunistic jobs from OSG, others

This is one of **4** US ATLAS Tier-2s; there are **7** US CMS Tier-2s; approximately **100** globally

Rucio (Distributed Data Management)

Rucio provides a complete and generic scientific data management service

- Designed with more than 10 years of operational experience in large-scale data management!
  - In use by the **ATLAS**, **AMS** and **Xenon1T** Collaborations
  - Website https://rucio.cern.ch/index.html
- Rucio manages multi-location data in a heterogeneous distributed environment
  - Creation, location, transfer, and deletion of replicas of data
  - Orchestration according to both low-level and high-level driven data management policies (usage policies, access control, and data lifetime)
  - Interfaces with workflow management systems
  - Supports a rich set of advanced features, use cases, and requirements
  - Large-scale and repetitive operational tasks can be automated

See workshop announcement at https://home.cern/news/news/computing/managing-scientific-data-exascale-rucio

OSiRIS

OSiRIS Open Storage Research Infrastructure

- In 2015 we proposed to design and deploy MI-OSiRIS (Multi-Institutional Open Storage Research Infrastructure) as a pilot project to evaluate a **software-defined storage infrastructure** for our primary Michigan Research Universities. OSiRIS combines a number of innovative concepts to provide a distributed, multi-institutional storage infrastructure

  The goal: to provide transparent, high-performance access to the same storage infrastructure from well-connected locations on any of our campuses via a combination of network discovery, monitoring and management tools and through the creative use of CEPH features

- **By providing a single data infrastructure that supports computational access on the data "in-place", we can meet many of the data-intensive and collaboration challenges faced by our research communities and enable these communities to easily undertake research collaborations beyond the border of their own Universities.**

OSiRIS (NSF Grant Number 1541335)

Logical view of OSiRIS

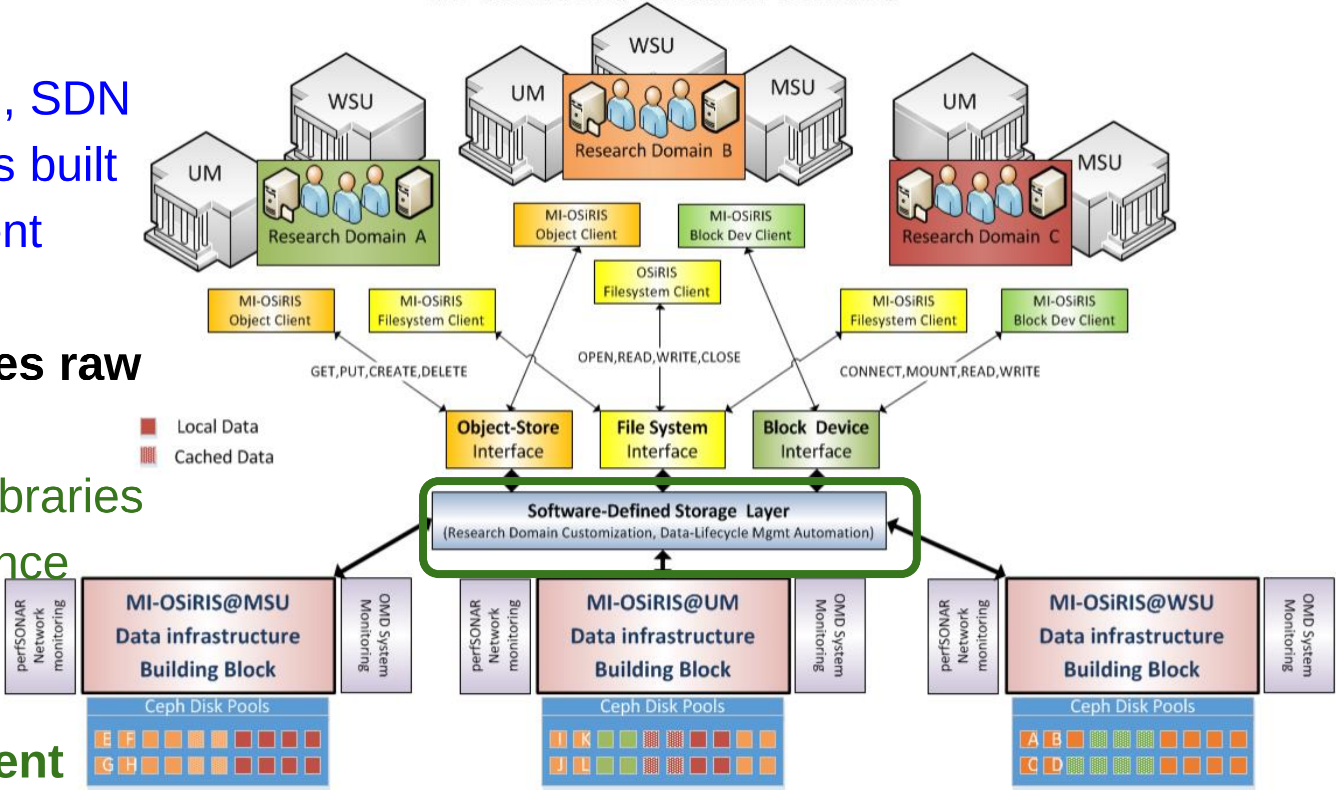
http://www.osris.org

MI-OSiRIS and Research Domains

Combination of Ceph, SDN and AAA components built upon COTS equipment

Currently 7 Petabytes raw

Exploring, with our Libraries and Information Science Departments, **how to enhance Data Lifecycle Management**



WSU
UM
MSU
Research Domain B
WSU
UM
Research Domain A
UM
MSU
Research Domain C
MI-OSiRIS Object Client
MI-OSiRIS Block Dev Client
OSiRIS Filesystem Client
MI-OSiRIS Object Client
MI-OSiRIS Filesystem Client
MI-OSiRIS Filesystem Client
MI-OSiRIS Block Dev Client
GET,PUT,CREATE,DELETE
OPEN,READ,WRITE,CLOSE
CONNECT,MOUNT,READ,WRITE
Local Data
Cached Data
Object-Store Interface
File System Interface
Block Device Interface
Software-Defined Storage Layer
(Research Domain Customization, Data-Lifecycle Mgmt Automation)
perfSONAR Network monitoring
MI-OSiRIS@MSU Data infrastructure Building Block
OMD System Monitoring
perfSONAR Network monitoring
MI-OSiRIS@UM Data infrastructure Building Block
OMD System Monitoring
perfSONAR Network monitoring
MI-OSiRIS@WSU Data infrastructure Building Block
OMD System Monitoring
Ceph Disk Pools
E
F
G
H
Ceph Disk Pools
I
K
J
L
Ceph Disk Pools
A
B
C
D

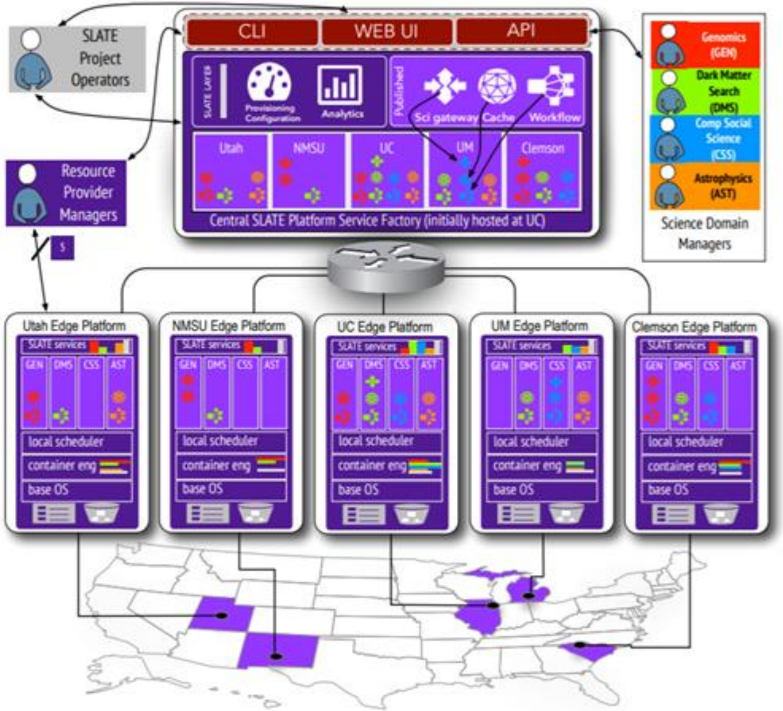
SLATE (Services Layer At The Edge) NSF Grant 1724821

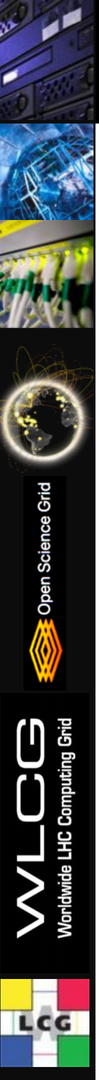
An NSF DIBBs award 724821, "SLATE and the Mobility of Capability"

Equip the SciDMZ with a service orchestration platform, potentially federated to create scalable, multi-campus science platforms

SLATE is building upon Kubernete to provide a way for scientists to define their science process **centrally**

The same platform allows those providing computing, storage and networking resources to control the amount of those resources any science domain will get





Open Science Grid (OSG)

The **O**pen **S**cience **G**rid (**OSG**) provides common service and support for **resource providers** and **scientific institutions** using a distributed fabric of high throughput computational services. The **OSG does not own resources** but provides **software** and **services** to users and resource providers alike to enable the opportunistic usage and sharing of resources. The OSG is funded through a diverse portfolio of awards from the National Science Foundation and the Department of Energy

- While **OSG** has many scientific stakeholders, **LHC** plays a prominent role
- The new **IRIS-HEP** Institute funds the LHC parts of OSG i
- The networking area is a good example of LHC efforts (actually USATLAS) that grew to provide a set of tools and infrastructure broadly used to support distributed science domains via OSG.

IRIS-HEP

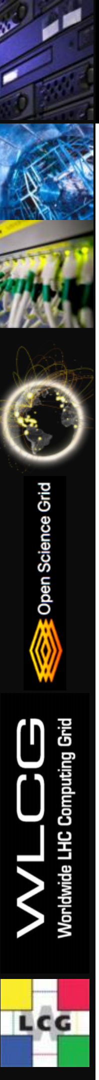
Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP) project has been funded by National Science Foundation in the US as grant OAC-1836650 as of 1 September, 2018.

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' and maximize the HL-LHC physics potential;
- **Data organization, management and access (DOMA)** 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/



Working Collaboratively

The high-energy physics community has to work collaboratively for at least the last 40 years to be able to undertake their science at the required scale.

This is not only **within** large collaborations but **between** them as well:

- **ATLAS** and **CMS** sharing, and jointly researching, networking technologies, data distribution technologies, workflow management systems and infrastructure components.

More recently we find it makes sense to work collaboratively beyond HEP: astronomy/astrophysics collaborations are reaching HEP scales and facing similar challenges and others science domains are soon to follow.

- The **network** has been an obvious place to work together since it is foundational for any distributed or data-intensive science domain; **LHCOPN/LHCONE** is a good example
- Sharing workflow management systems (**PanDA**), DDD (**Rucio**) and others

LHC Tool Trends and Technologies

Here I try to summarize some important changes since LHC startup

Reducing arbitrary boundaries and definitions: Tier-1 and Tier-2 sites

Allow sites to provide resources that aren’t limited by predefined roles

Improvements in job workflows accounting for new architectures and trends

- Pilot job system provides late binding and task optimization opportunities
- Use of CVMFS+SQUID (think filesystem over http) provides centralized applications to make levering HPC and Cloud resources easier
- XCache working to provide similar capabilities for datasets

Increasing use of the network, beyond just the growth with data volume

Has been one of LHC’s most reliable components

WAN access to data may help alleviate storage requirements

Could improve efficiency of tasks, allowing mix-match of storage and CPU

Software refactoring to take better advantage of technology trends: increasing core counts, GPUs, FPGAs, ARM systems, SSDs, NVMes, etc.

Data Lakes (A Placeholder for a Storage-Optimized LHC)

I want to close with a concept the HEP community is using a placeholder for how we might address the significant gap in storage capacity we foresee for the HL-LHC: Data Lakes

The input: 1) storage is hard manage and optimize across a large number of sites of varying capacity and capability and 2) we can't afford the capacity we need

The idea: create a few, large (potentially continental) Data Lakes that provide a simple external interface while handling complexity and optimization of use and performance internally. To do this focus on **Quality of Service, Smart Caching and Data Lifecycle management**

- The User can specify the number of replicas and the QoS associated for each of them, i.e. one on fast storage (disks on SSDs) and two on tape in three different locations. The system should be able to automatically maintain in time that policy verified.
- The User can specify that certain datasets always have a mirror, checking the replicas status in real time or quasi-real time.
- Users can specify that a number of replicas are created and they have to be accessed with different protocols, i.e. http, xrootd, srm)
- The user can specify movements between QoS and/or changes in access controls based on data age (i.e quarantine periods, move to Tape old data)
- Provide smart caching mechanisms to support the remote extension of a site to remote locations and to provide alternative models for large data centers. Data stored in the original site should be accessible in a transparent way from the remote location.

Status: The concept is actively being discussed and prototyped. Very challenging to save money while providing capacity and capability!!

Summary and Next Steps

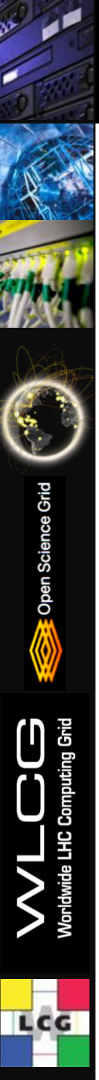
For HEP there are significant challenges preparing for LHC Run-3 and especially HL-LHC

There are a number of projects and activities working to address these challenges

- **Where possible HEP is trying to enable what we have found to work for the broader scientific community facing similar challenges.**
- **Likewise we are trying to also adopt and benefit from the work of other communities**

Understanding and evolving our scientific data lifecycle components and methodologies will be **critical** to how successful we will ultimately be in Run-3 and HL-LHC

QUESTIONS?



Useful References

AGLT2 website https://www.aglt2.org/

OSiRIS website http://www.osris.org

Open Science Grid (OSG) https://opensciencegrid.org/

IRIS-HEP website http://iris-hep.org/

HEP Software Foundation (HSF) https://hepsoftwarefoundation.org/

Simone Campana presentation on Data Lakes (2018) https://indico.cern.ch/event/738796/contributions/3174573/attachments/1755785/2846671/DataLake-ATCF.pdf

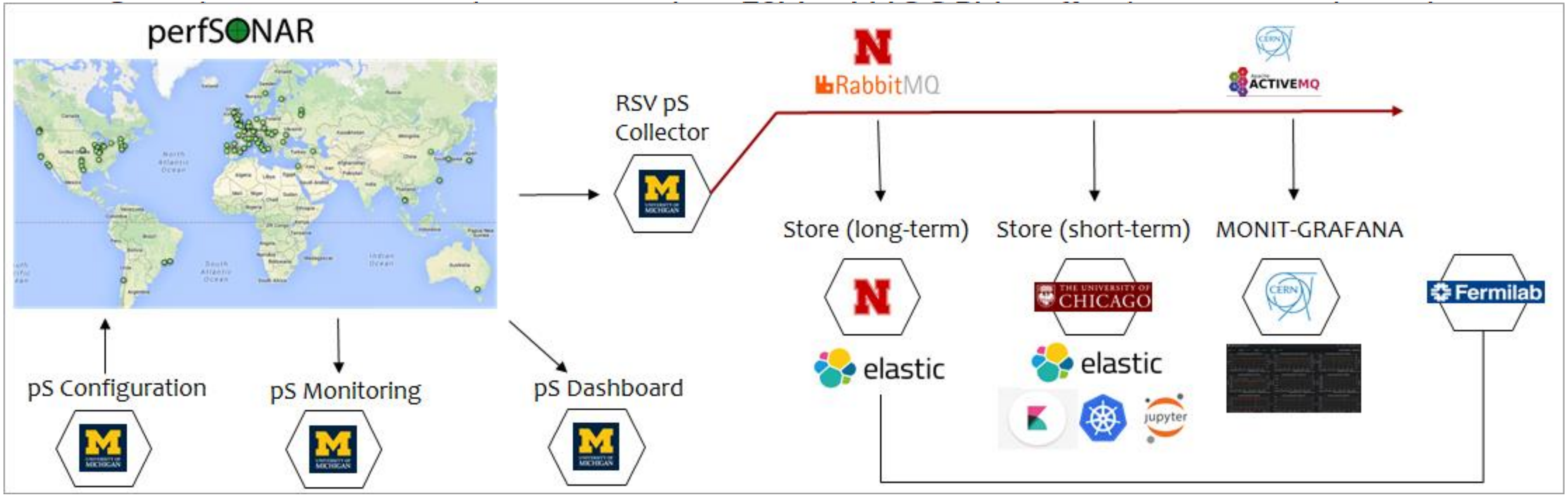
Frank Wuerthwein presentation on CMS HL-LHC storage requirements (2018) https://indico.cern.ch/event/764786/contributions/3221336/attachments/1756276/2847549/domaAccess112018.pdf

Ian Fisk presentation on evolution of the LHC Computing Model (2014) https://indico.egi.eu/indico/event/1994/session/73/contribution/200/material/slides/0.pdf

Additional Slides

OSG Network Monitoring Platform Overview

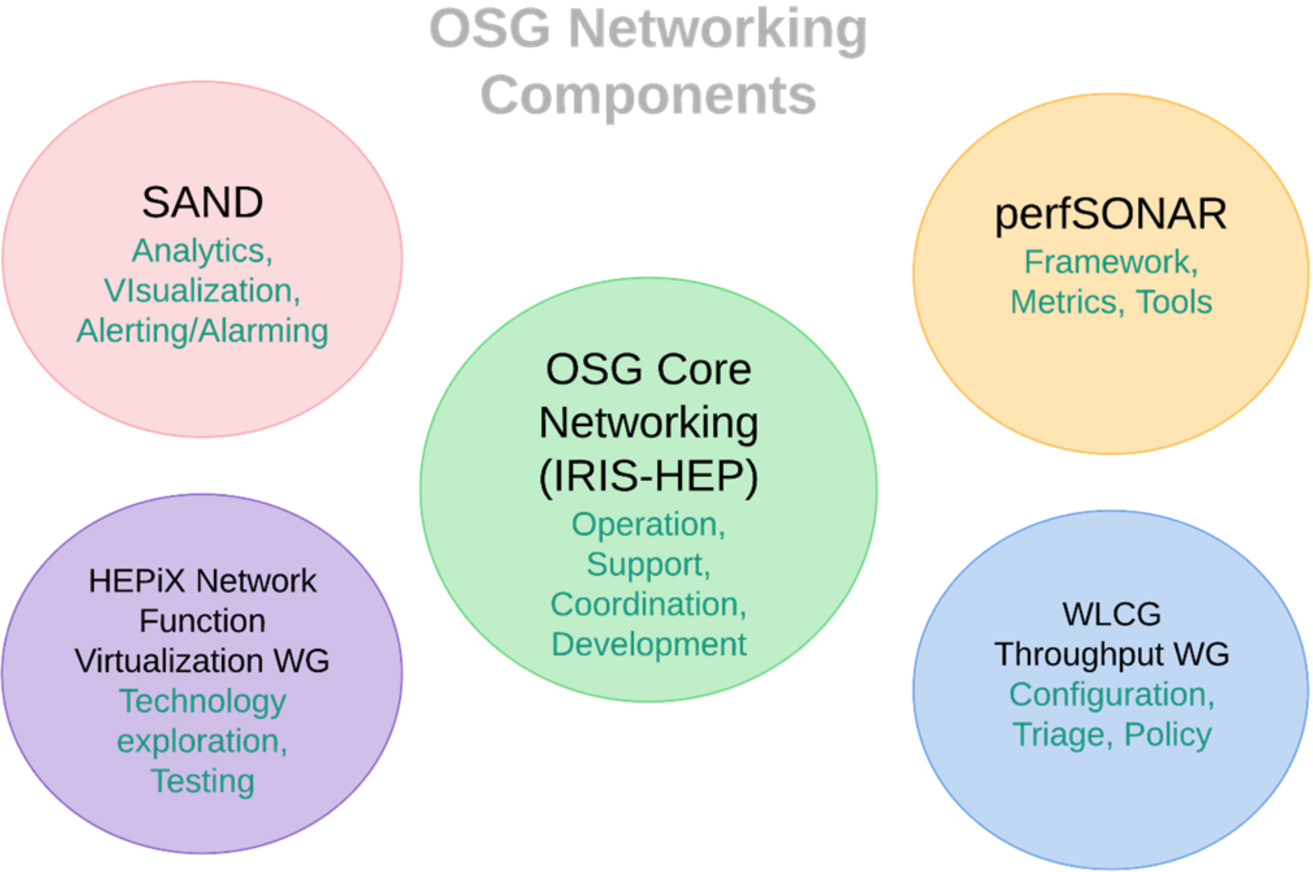
- Collects, stores, configures and transports all network metrics
  - Distributed deployment - operated in collaboration
- All perfSONAR metrics are available via API, live stream or directly on the analytical platforms



OSG Networking

There are 4 coupled projects around the core OSG Net Area

1. SAND (NSF) project for analytics
2. HEPiX NFV WG
3. perfSONAR project
4. WLCG Throughput WG



OSG Networking Components

SAND
Analytics, VIsualization, Alerting/Alarming

perfSONAR
Framework, Metrics, Tools

OSG Core Networking (IRIS-HEP)
Operation, Support, Coordination, Development

HEPiX Network Function Virtualization WG
Technology exploration, Testing

WLCG Throughput WG
Configuration, Triage, Policy

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

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov