The OSiRIS Project: Meeting the Multi-Institutional Data Collaboration Challenge

April 18th, 2016

HEPiX Spring 2016 - Zeuthen

Shawn McKee / University of Michigan

For the OSiRIS Collaboration

The OSiRIS Project Summary

- 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 will combine a number of innovative concepts to provide a distributed, multi-institutional storage infrastructure that will allow researchers at any of our three campuses to read, write, manage and share their data directly from their computing facility locations.
  - Our goal is to provide transparent, high-performance access to the same storage infrastructure from well-connected locations on any of our campuses. We intend to enable this via a combination of network discovery, monitoring and management tools and through the creative use of CEPH features as described below.
- **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.**

The Multi-Institutional Data Challenge

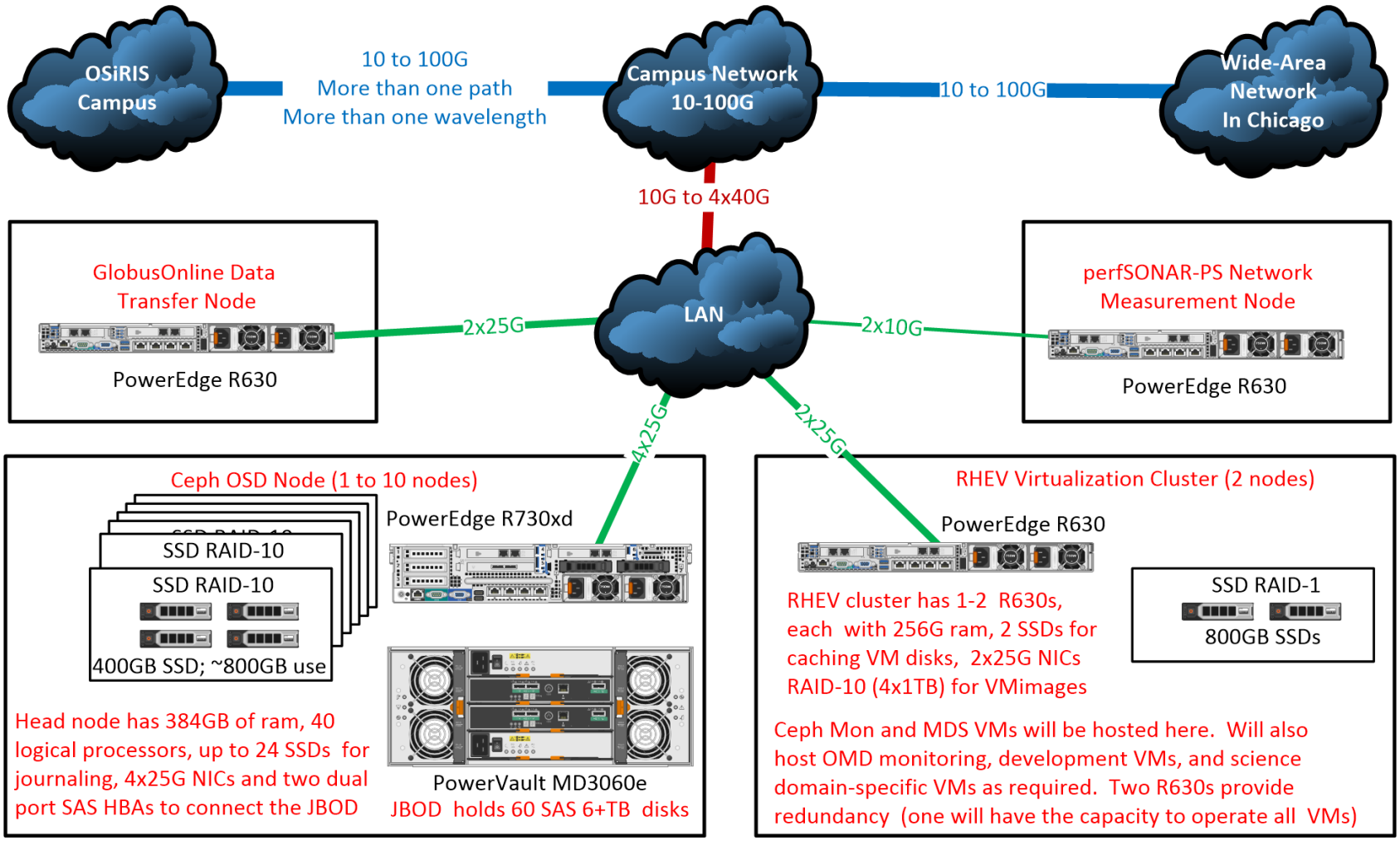
- Scientists working with large amounts of data face many obstacles in conducting their research
  - Typically the workflow needed to get data to where they can process it becomes a substantial burden (along with the bookkeeping)
- The problem **intensifies** when adding in collaboration across their institution or especially **beyond their institution**.
- Institutions have sometimes responded to this challenge by constructing specialized infrastructures to support specific science domain needs.
  - This doesn't scale and can be expensive (in many ways)
- The OSiRIS team proposed a research project to investigate a possible solution

Who is the OSiRIS Team?

- OSiRIS is composed of scientists, computer engineers and technicians, network and storage researchers and information science professionals from the 3 main research Universities in Michigan: University of Michigan, Michigan State University and Wayne State University, as wells as Indiana University (focusing on SDN and net-topology)
- We have a wide-range of science "stakeholders" who have data collaboration and data analysis challenges to address within, between and beyond our campuses:
  - *High-energy physics, High-Resolution Ocean Modeling, Degenerative Diseases, Biostatics and Bioinformatics, Population Studies, Genomics, Statistical Genetics and Aquatic Bio-Geochemistry*

An OSiRIS Institutional Deployment

OSiRIS Data Infrastructure Building Block



OSiRIS Building Block Details

- **Ceph Server Building Block** (6U, 480TB raw)
  - Dell R730xd, 384GB, E5-2650v3x2, 4x400GB NVMe, 2x120 SSD, 250G SATA, 24-disk capacity, 2x50G Connect-X4 NICs, 2 SAS HBA)
  - Dell MD3060e, 60x8TB HGST Helium disks, dual SAS interfaces, dual P/S
- **Switch:** Dell Z9100, 32x100G QSFP28 (10-50G)
- **GlobusOnline:** Dell R630, 2xE5-2650v3, 128GB, 2x25G Dual Port Connect-X4
- **perfSONAR:** Dell R630, Intel X520 2x10G NIC, 2xE5-2620v3, 32GB,2x1TB NLSAS
- **Virtualization:** Dell R630, 2xE5-2695, 256GB, 2x500G NLSAS, 2x800G NVMe, 4x1.2TB 10K SAS

OSiRIS Hardware is Deployed

- The OSiRIS project requested proposals to meet our hardware needs in October 2015 (9 bids)
- We decided on Dell+HGST+Mellanox Cx4 NICs in November
  - Orders out in December
- Equipment arrived in January/February 2016
- All sites are now racked and cabled.
  - Picture of UM Install

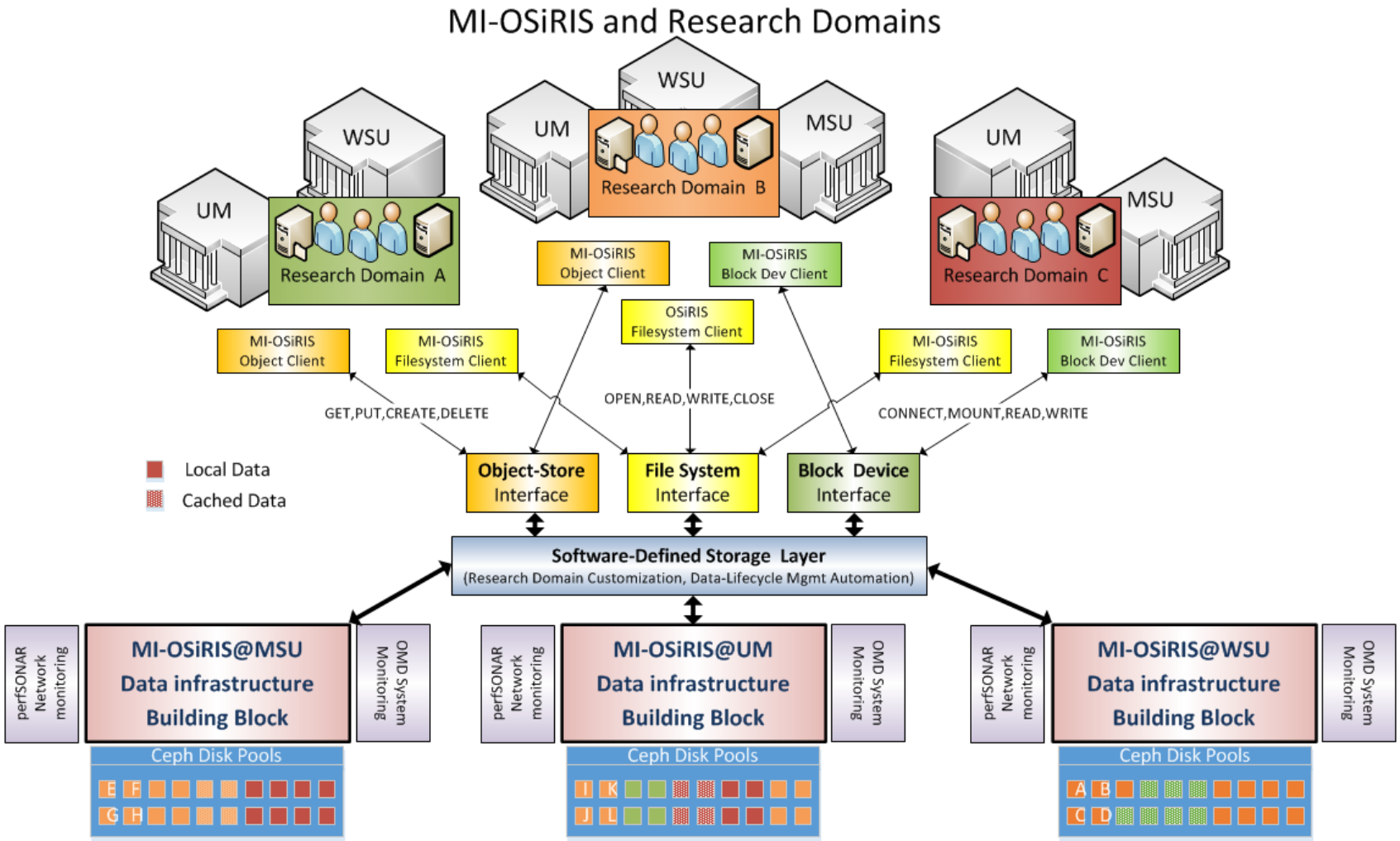
Why OSiRIS?

- 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 and resiliency.
- OSiRIS, via CEPH, provides seamless rebalancing and expansion of the storage.
- **A single, scalable infrastructure** is much easier to build and maintain
- Allows universities 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.
  - Data infrastructure view for each research domain can be optimized

Why Ceph for OSiRIS?

- Ceph gives us an Open Source platform to host our multi-institutional science data
  - Able to tune each science domains components to best meet their task
  - Multiple interfaces between users and data are possible
  - Has aspects of Software Defined Storage built-in which give us options for data lifecycle management automation
- The combination of self-healing and self-managing make it very attractive to us.
- It allows us to assign each science domain sets of disks which isolates science users from one another while allowing us to customize and optimize the storage for each science use-case.
- Ben Meekhof/UM ARC-TS has a nice online presentation of the Ceph details at https://umich.app.box.com/s/f8ftr82smlbuf5x8r256hay7660soafk
- We have installed Ceph via Openstack Puppet module https://github.com/openstack/puppet-ceph at each site (Infernalis). Forked version with minor fixes at https://github.com/MI-OSiRIS/puppet-ceph
  - Adds custom 'cephmajor', 'cephminor' facts
  - Fixes init issue with mons in Infernalis

Logical View of OSiRIS



Software Defined Networking?

- Software defined networking (SDN) changes traditional networking by decoupling the system that makes decisions about where traffic is sent (the control plane) from the underlying systems that forward traffic to the selected destination (the data plane).
- Using SDN we can centralize the control plane and, using software, programmatically update how the network behaves to meet our goals.
- For OSiRIS the network will be a critical component, tying our multi-institutional users to our distributed storage components.

Network Monitoring & perfSONAR

- Because networks underlie distributed cyberinfrastructure, monitoring their behavior is very important
- The research and education networks have developed perfSONAR as a extensible infrastructure to measure and debug networks (http://www.perfsonar.net/)
- The CC*DNI DIBBs program recognized this and required the incorporation of perfSONAR as part of any proposal.
- For OSiRIS, we were well positioned since I lead the worldwide perfSONAR deployment effort for the LHC community: https://twiki.cern.ch/twiki/bin/view/LCG/NetworkTransferMetrics
  - We intend to extend perfSONAR to enable the discovery of all network paths that exist between instances
  - SDN can then be used to optimize how those paths are used for OSiRIS

OSiRIS Network Management Abstraction Layer(NMAL) Progress

- OSiRIS is working on network management in a number of areas
- **Capturing site topology and routing information** in UNIS from multiple sources: SNMP, LLDP, sflow, SDN controllers, and existing topology and looking glass services.
  - Existing UNIS encoder is being extended to incorporate these new data sources.
- **Packaging and deploying conflict-free measurement scheduler** (HELM) along with measurement agents (BLiPP).
- Converged on common scheduled measurement architecture with existing perfSONAR mesh configurations.
  - Correlate long-term performance measurements with passive metrics collected via check_mk infrastructure.
- Integrating Shibboleth to provide authentication/authorization for measurement and topology services. This includes extending existing perfSONAR toolkit components in addition to Periscope.
- Defining best-practices for SDN controller and reactive agent deployments within OSiRIS.

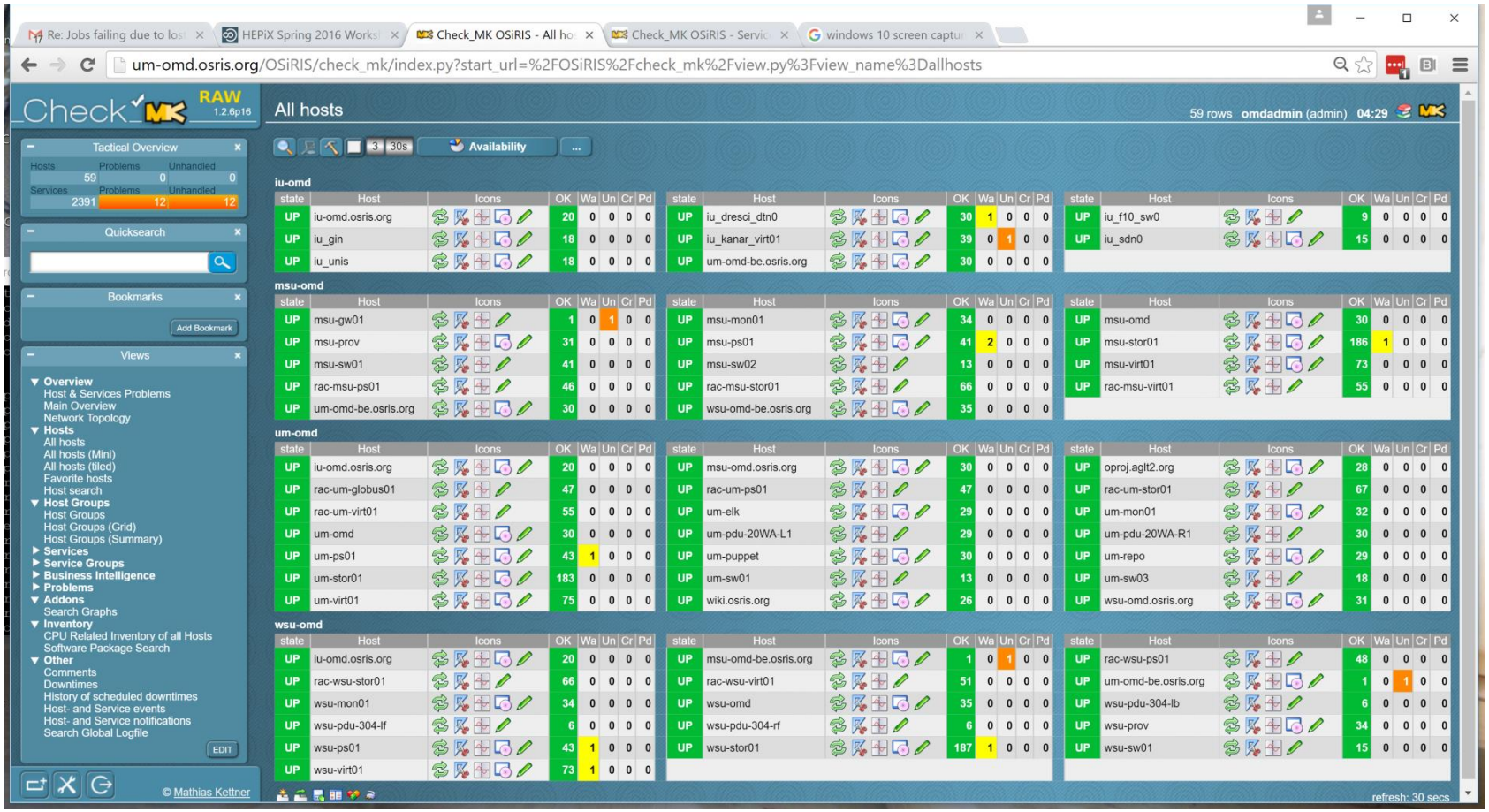
OSiRIS Security/Authorization Challenges

- We need to provide an infrastructure that leverages existing institutional credentials.
  - We are working with Von Welch and Jim Basney from the Center for Trusted Scientific CyberInfrastructure to find the best way forward http://trustedci.org/who-we-are/
- So far using **InCommon Federation** attributes is not necessarily straightforward
  - There are widely varying levels of InCommon participation and attribute release
  - Becoming a Research and Scholarship entity grants more attributes from sites that participate (but not all sites participate). **OSiRIS has done this**
- Augmenting Ceph will for fine grained authorization from institutional and VO attributes is one of our major challenges. We are investigating options…

OSiRIS Challenges

- SYSTEM optimization to maintain a sufficient quality of service for all stake-holders.
- Enabling the gathering and use of metadata to support data lifecycle management.
- Research domain customization using CEPH API and/or additional services.
- Management of "quotas" and ACLs? How best to control data space and services?
- Authorization which integrates with each campuses existing systems.
- **To meet these challenges we are using a number of tools to organize our effort and information…**

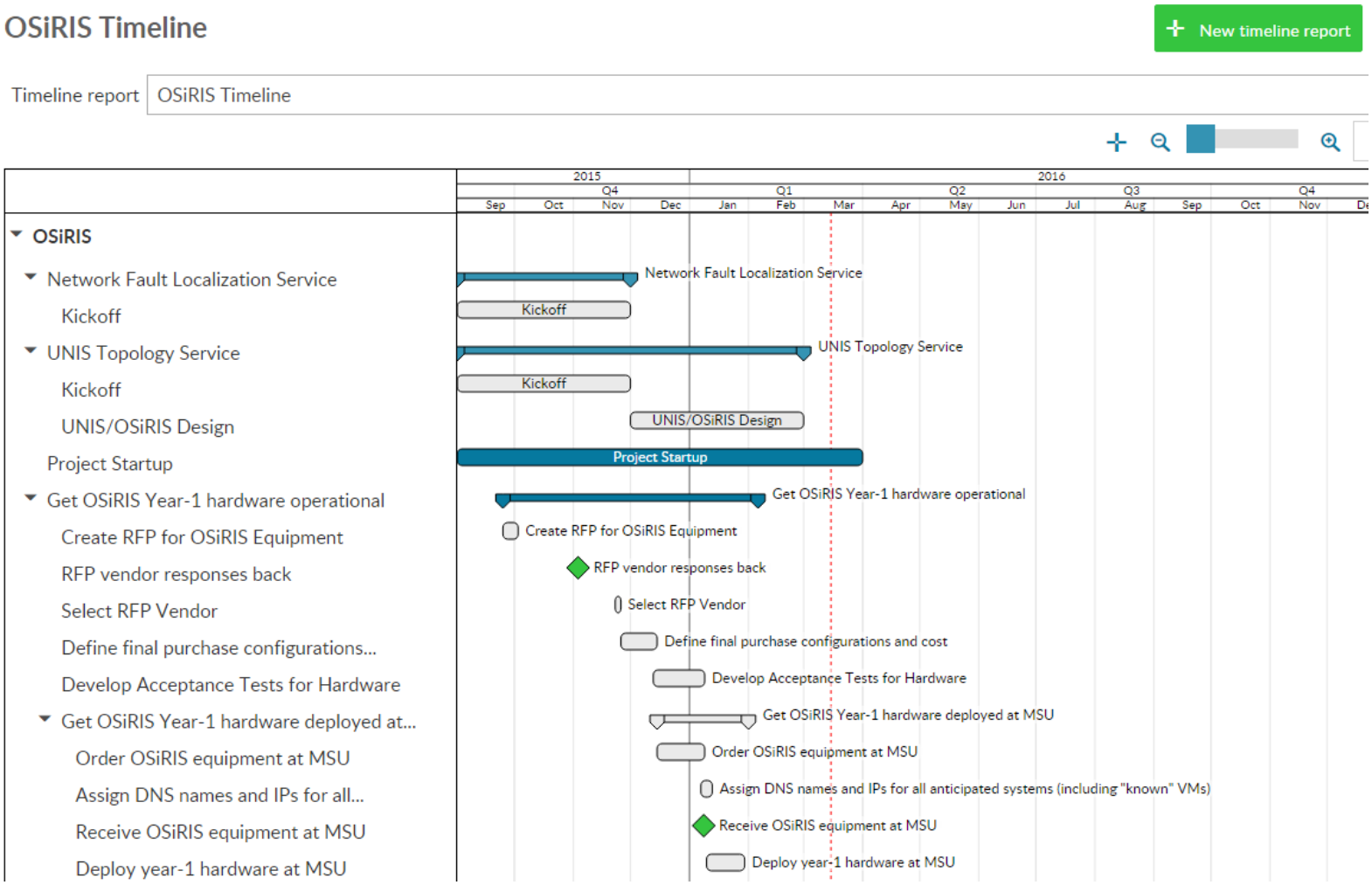
Check_mk for Monitoring



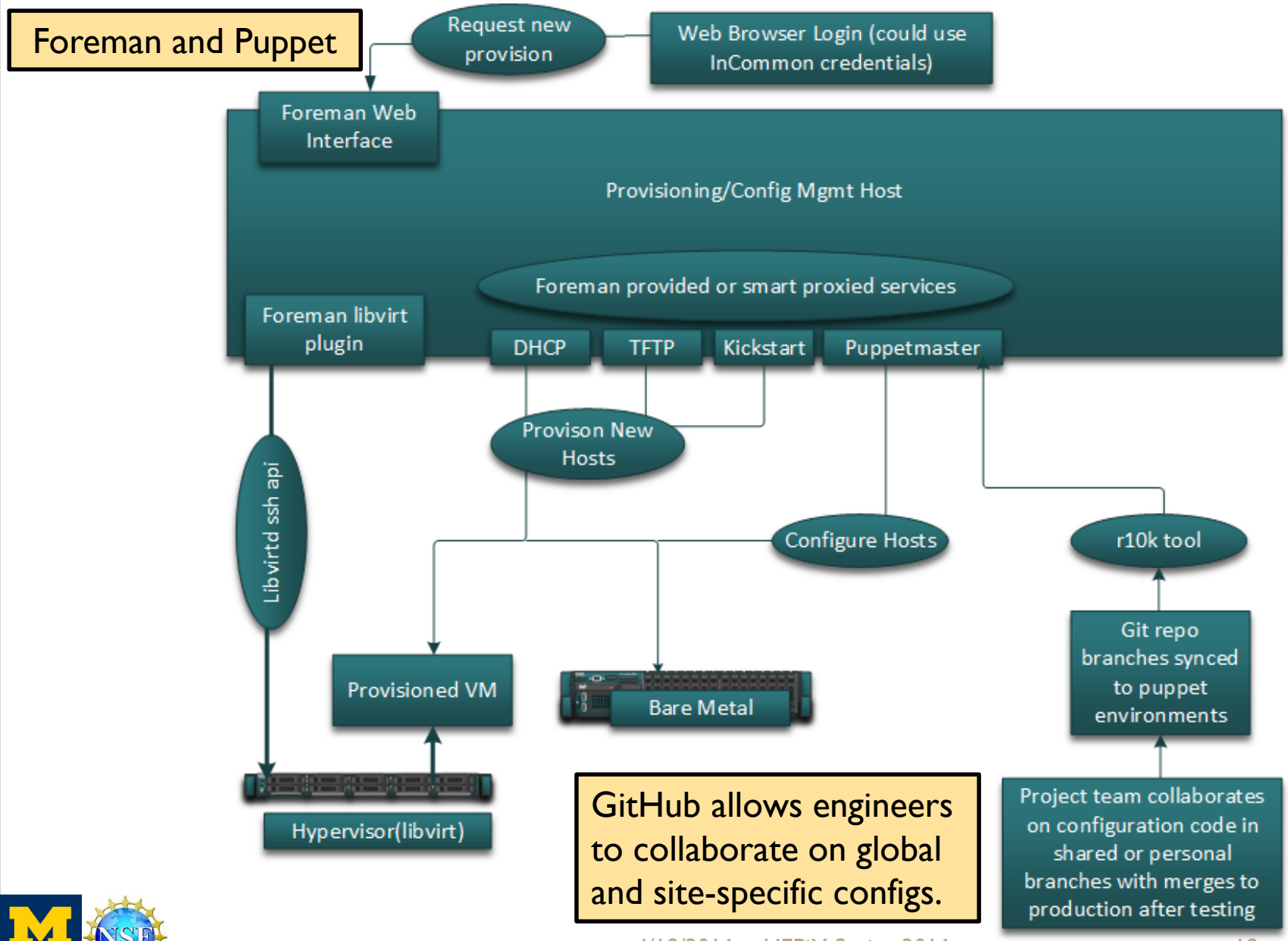
Monitoring 59 "servers" (VM+physical) with almost 2400 checks/minute at 4 sites

OSiRIS OpenProject

Project management via open-source http://www.openproject.org



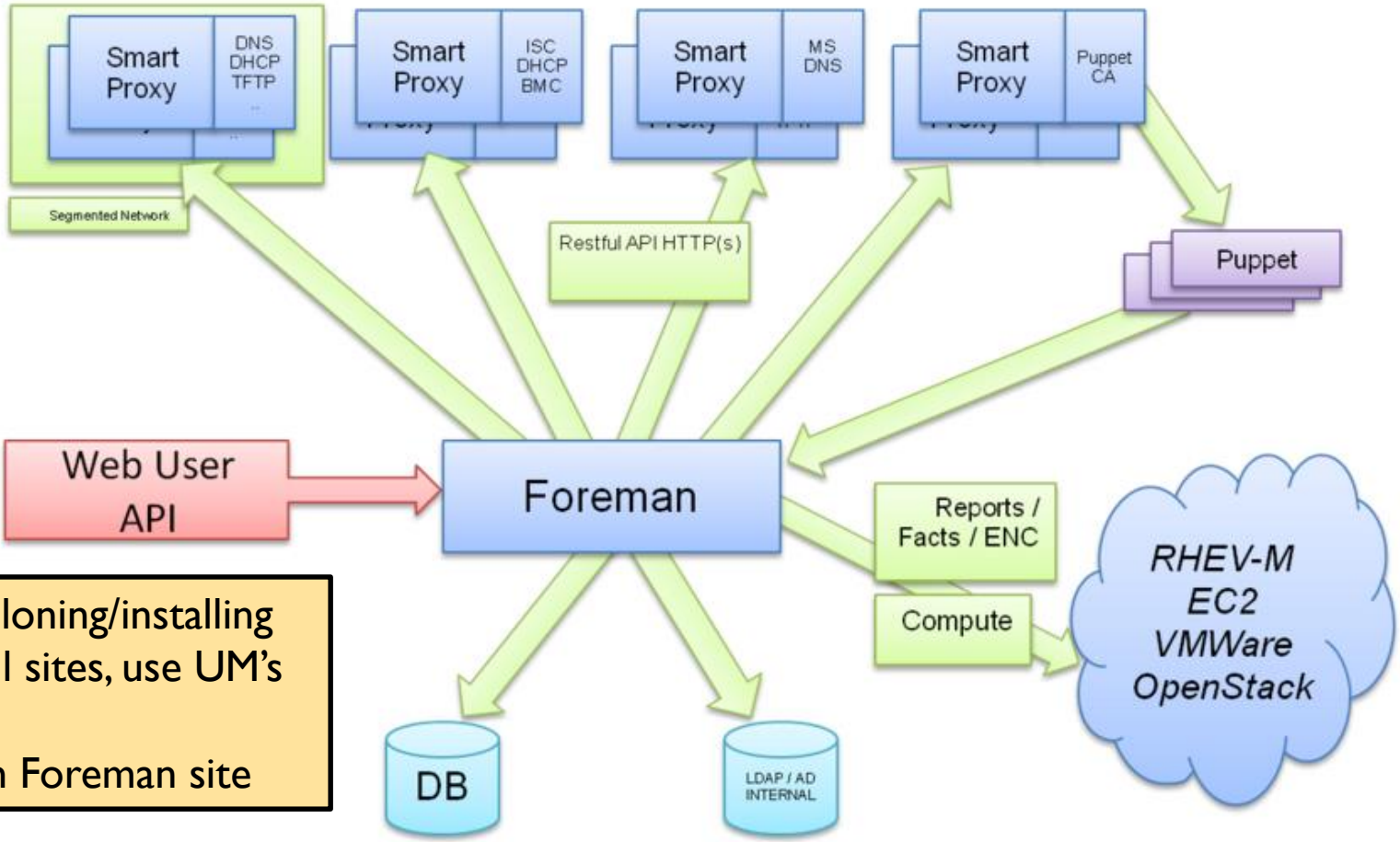
OSiRIS Provisioning Infrastructure



Provisioning MSU, WSU

Smart-Proxy

A Smart-Proxy is located on or near a machine that performs a specific function and helps foreman orchestrate the process of commissioning a new host. Placing the proxy on or near to the actual service will also help reduce latencies in large distributed organizations.

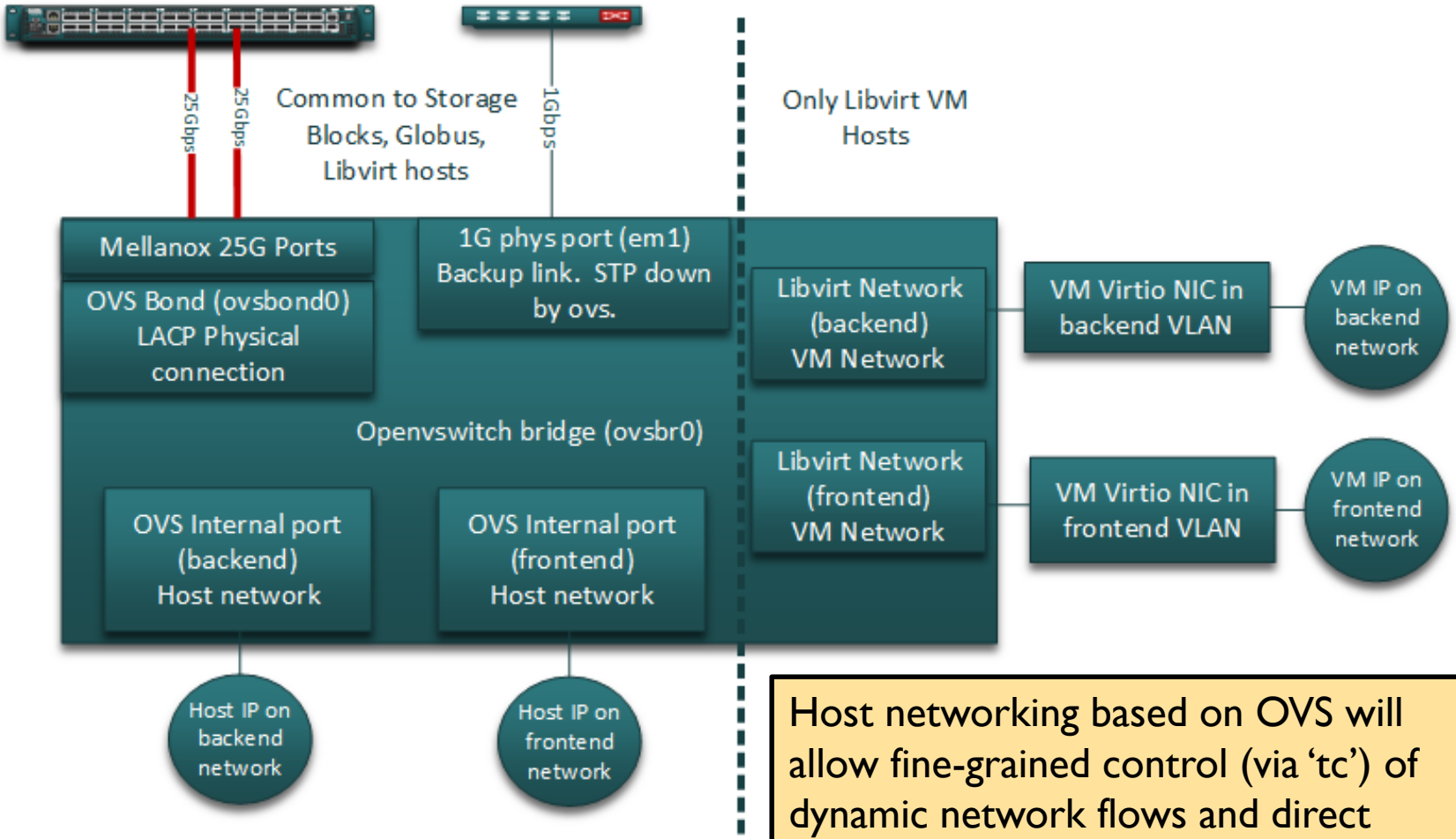


Rather than cloning/installing Foreman at all sites, use UM's

Diagram from Foreman site

OSiRIS OVS Setup

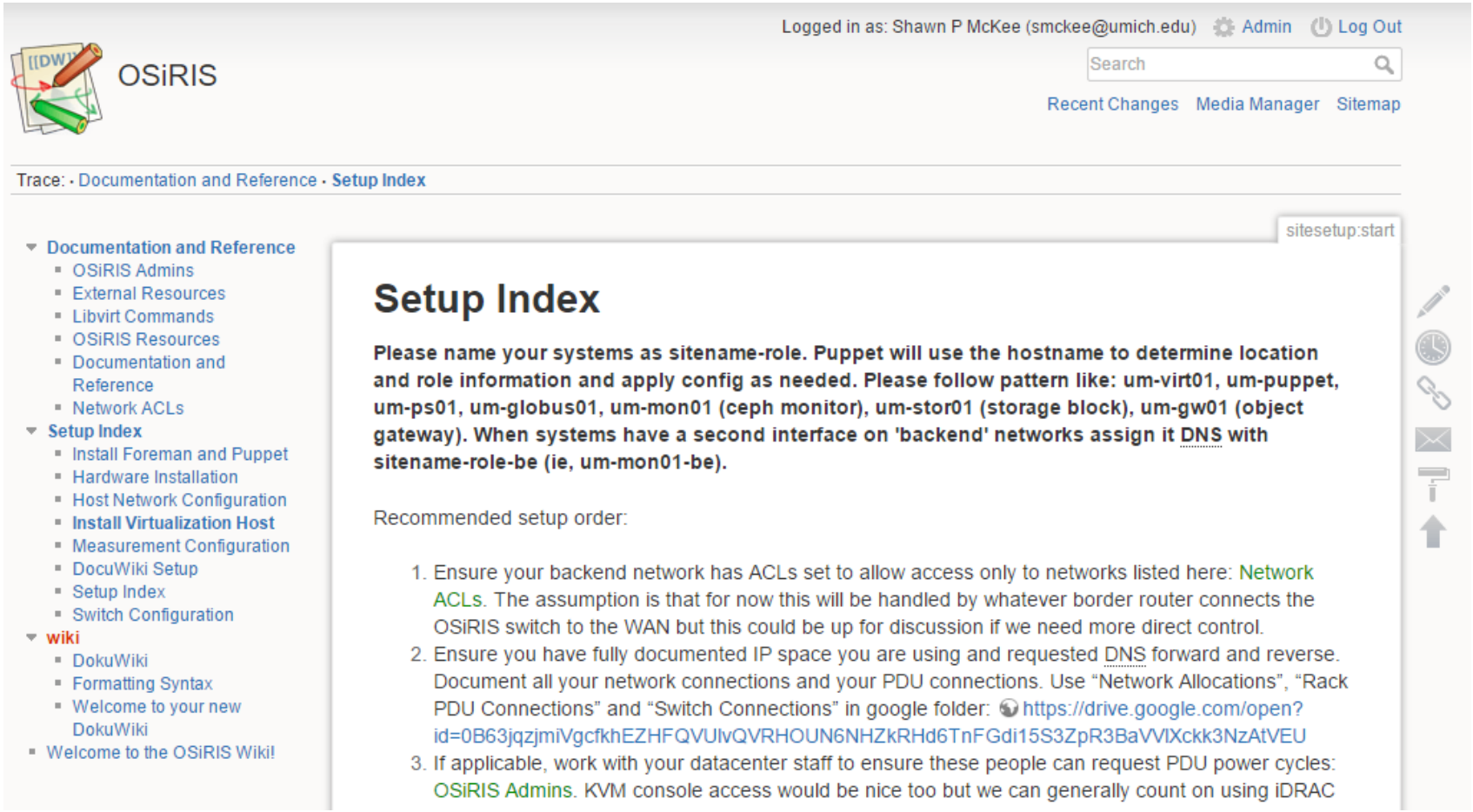
Open vSwitch (OVS)
http://openvswitch.org/



Host networking based on OVS will allow fine-grained control (via 'tc') of dynamic network flows and direct integration with OpenFlow controllers

OSiRIS DokuWiki

The OSiRIS Wiki uses DokuWiki https://www.dokuwiki.org/
Enables the use of Shibboleth and InCommon (matches our security plans)



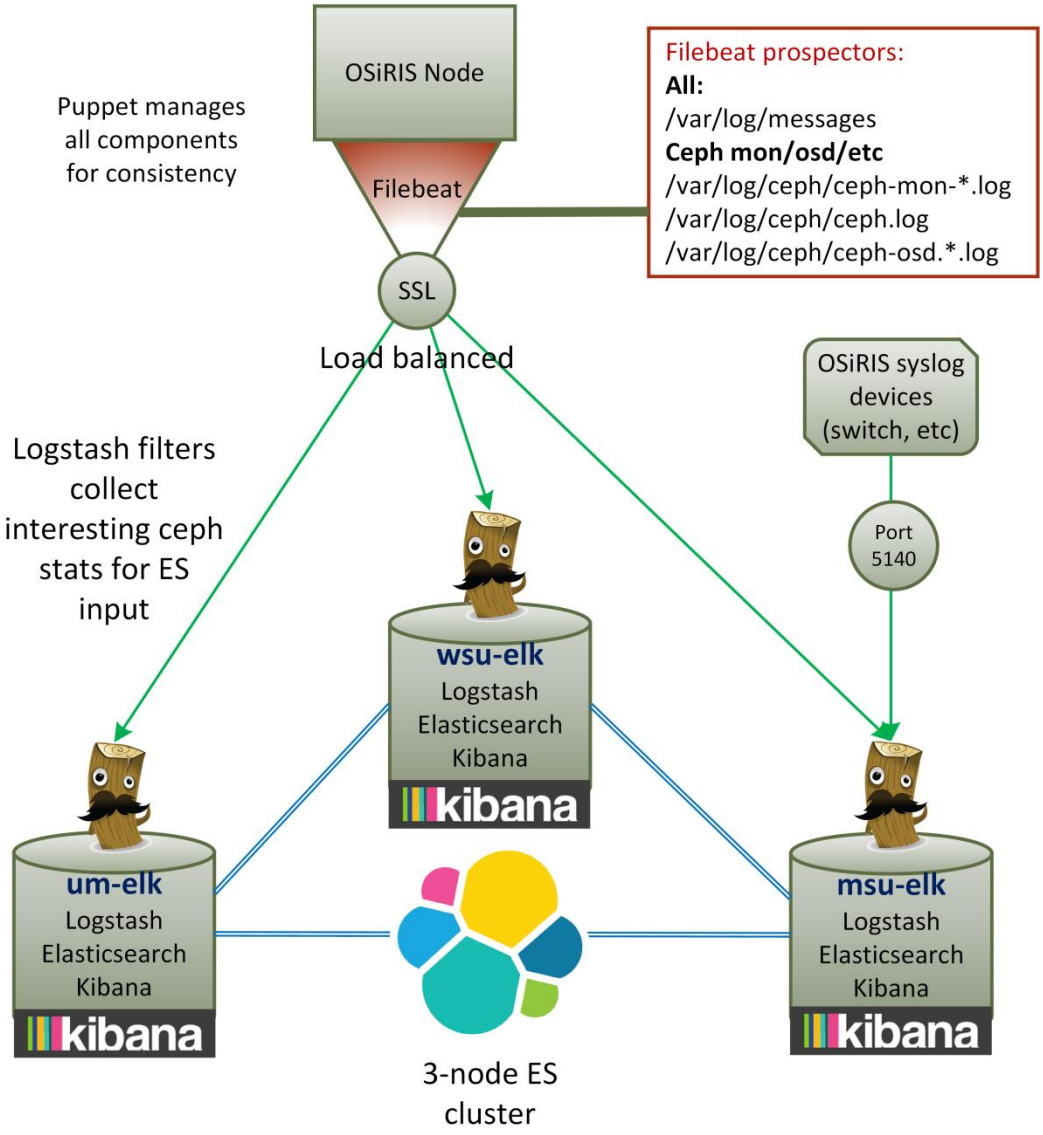
Integrating ELK for OSiRIS

- We need to make logfile contents avialable across our infrastructure
- ELK (ElasticSearch, Logstash, Kibana) is a proven way to this
- We setup a 3-node ES cluster (one instance per site)
  - Tuned for WAN

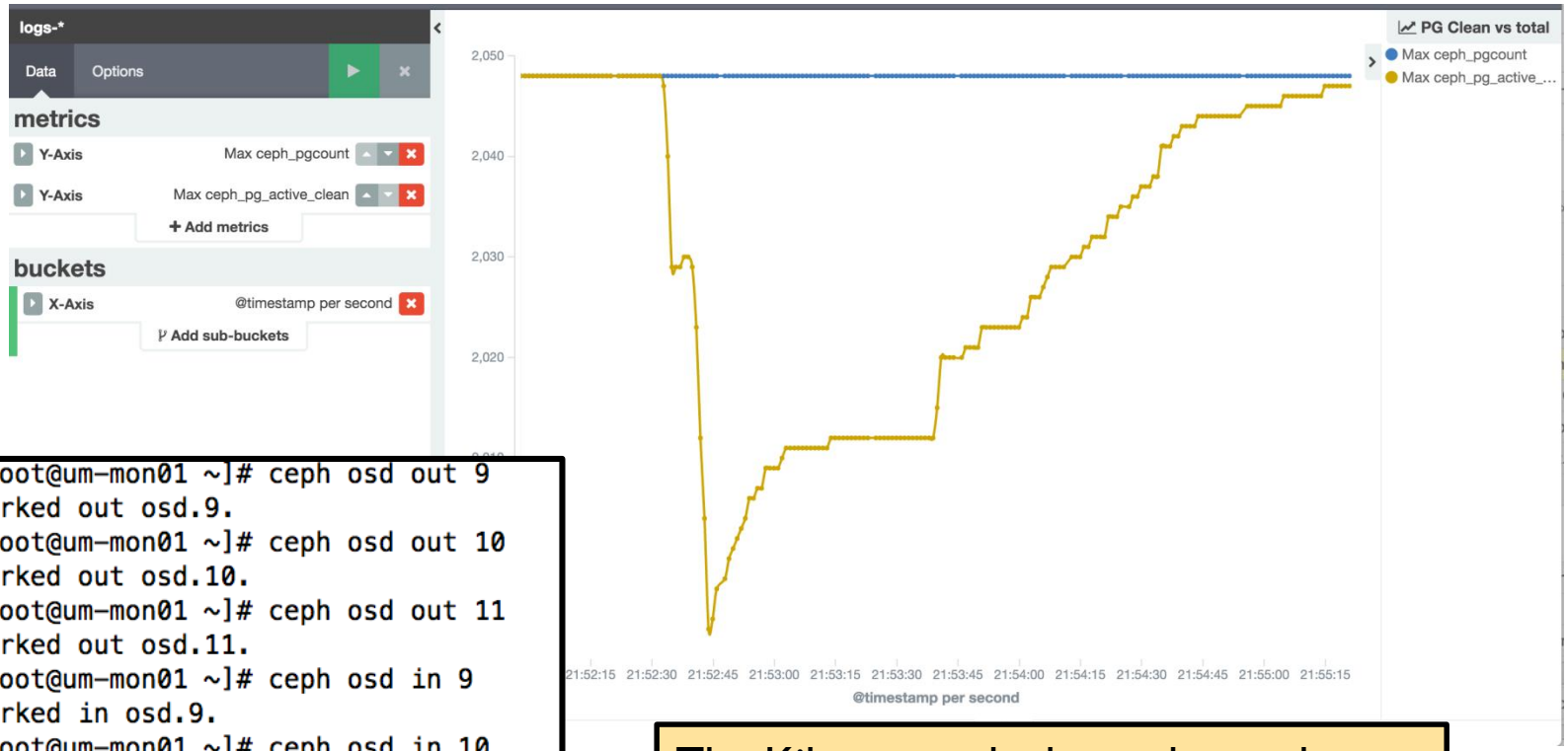
  ping_interval: 15s

  ping_retries: 5

  ping_timeout: 60s



Testing OSD Remove/Replace




```
[root@um-mon01 ~]# ceph osd out 9
marked out osd.9.
[root@um-mon01 ~]# ceph osd out 10
marked out osd.10.
[root@um-mon01 ~]# ceph osd out 11
marked out osd.11.
[root@um-mon01 ~]# ceph osd in 9
marked in osd.9.
[root@um-mon01 ~]# ceph osd in 10
marked in osd.10.
[root@um-mon01 ~]# ceph osd in 11
marked in osd.11.
```

The Kibana graph above shows the impact of removing and re-adding 3 OSTs in our Ceph instance. Y-axis is either pgcount (blue) or active+clean pg (yellow)

Remember the Goal

- The OSiRIS project is one attempt to try to address better enabling scientists to more easily collaborate without having to focus on the "how".
  - The science domains mentioned **all** want to be able to directly work with their data without having to move it to their compute clusters, transform it and move results back
  - Each science domain has different requirements about what is important for their storage use-cases: capacity, I/O capability, throughput and resiliency. OSiRIS has lots of ways to tune for these attributes (just not all of them at once!)

Summary

- There are significant challenges in providing infrastructures that transparently enable scientists to quickly and easily extract meaning from large, distributed or diverse data.
- OSiRIS is targeting doing exactly this and intends to incorporate a number of cutting edge technologies to provide such an infrastructure.

Questions or Suggestions?