OSiRIS Site Deployment

Leveraging puppet and foreman to build a distributed ceph cluster

Shawn McKee / Ben Meekhof
University of Michigan / ARC-TS
Michigan Institute for Computational Discovery and Engineering
Supercomputing - November 2016

What is OSiRIS?

OSiRIS combines a multi-site Ceph cluster with SDN and AAA infrastructure enabling scientific researchers to efficiently access data with federated institution credentials.

The current OSiRIS deployment spans Michigan State University, University of Michigan, and Wayne State University. Indiana University is also a part of OSiRIS working on SDN network management tools.

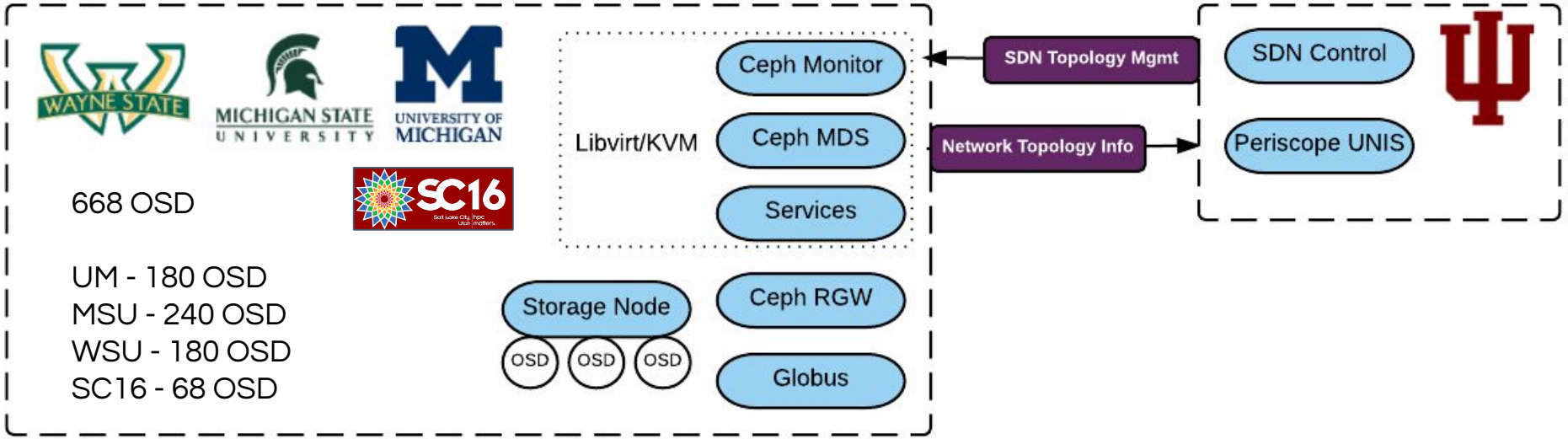
OSiRIS Goals

The OSiRIS project goal is enable scientists to collaborate on data easily and without building their own infrastructure.

We have a wide-range of science stakeholders who have data collaboration and data analysis challenges to address within, between and beyond our campuses.

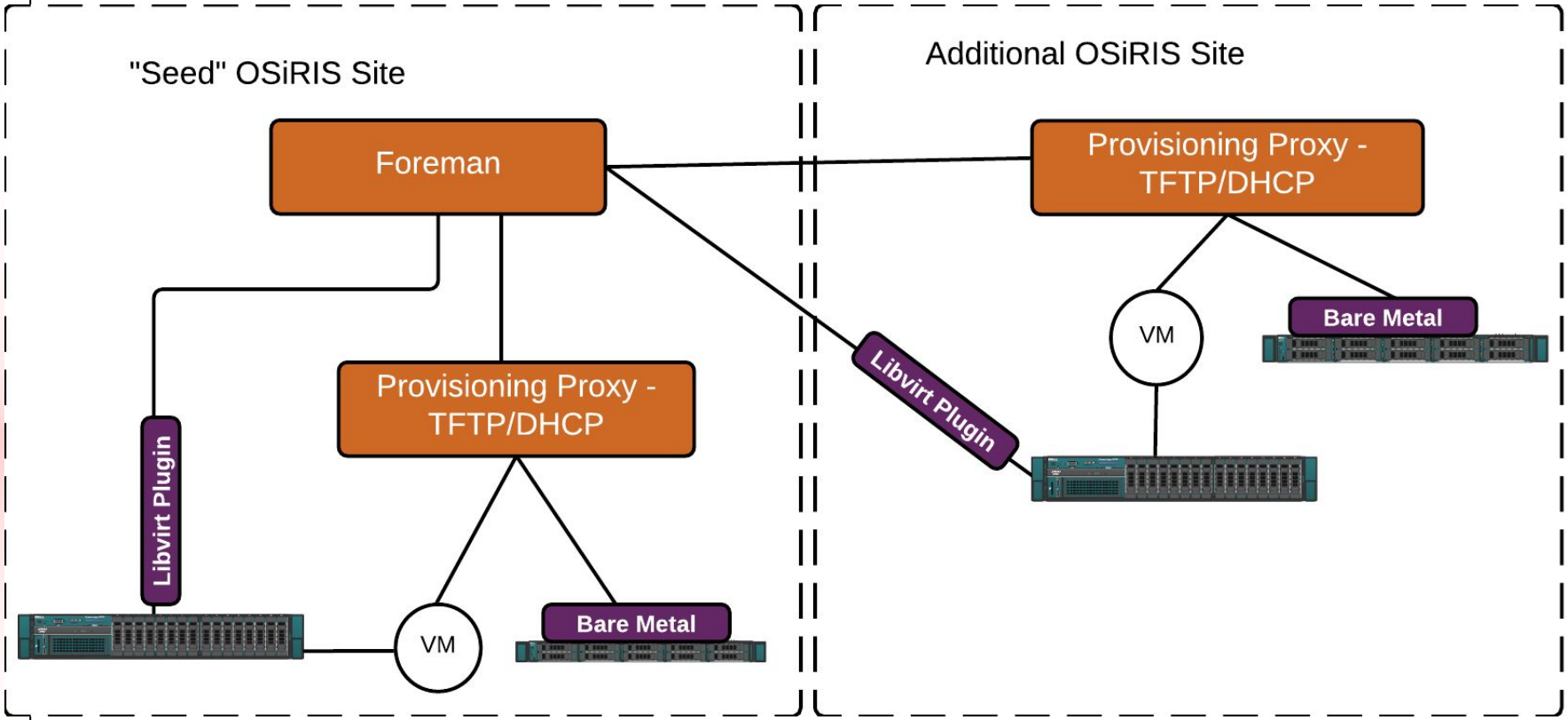
High-energy physics, High-Resolution Ocean Modeling, Degenerative Diseases, Biostatics and Bioinformatics, Population Studies, Genomics, Statistical Genetics and Aquatic Bio-Geochemistry

Our Deployment

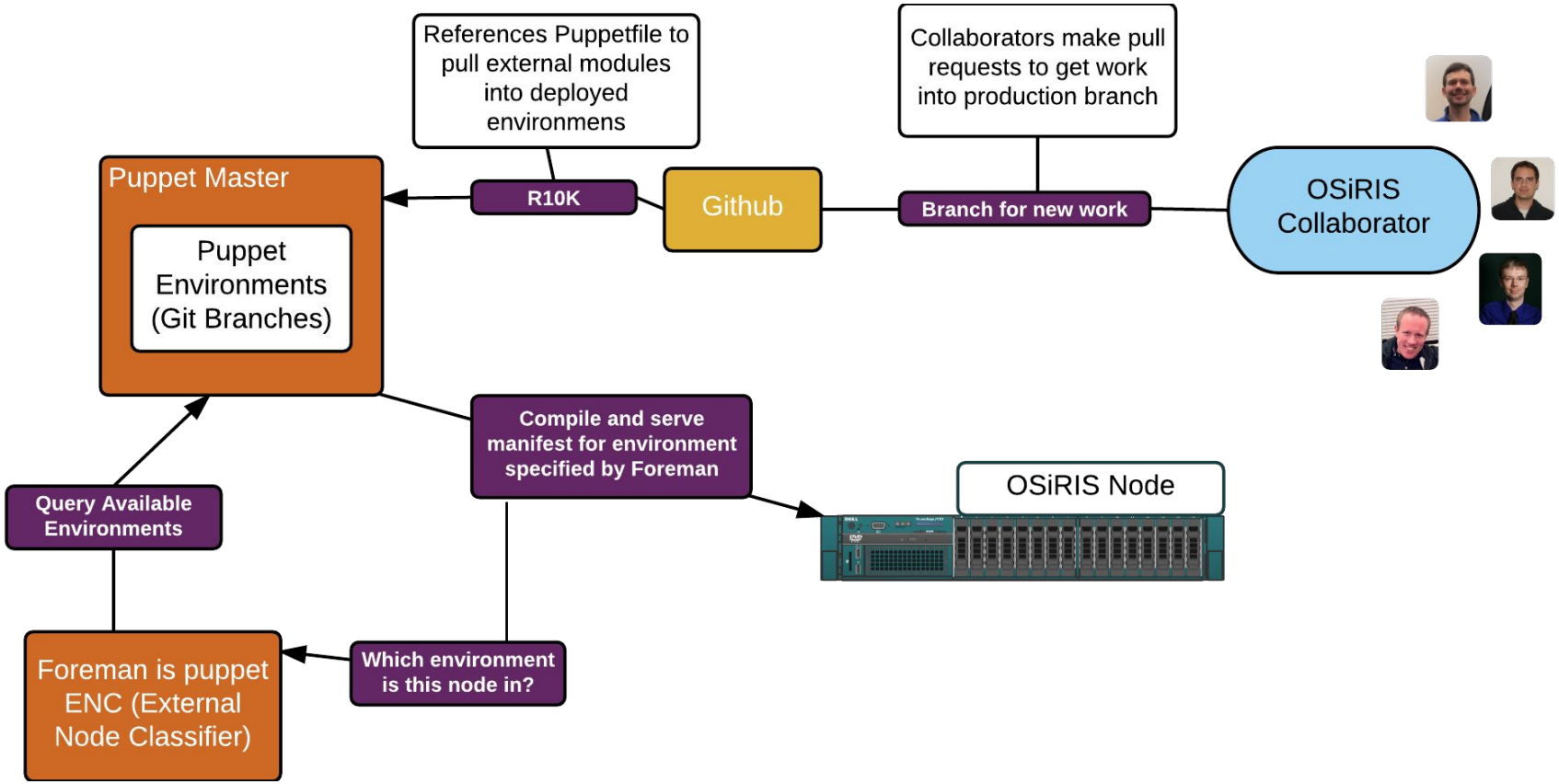


Our first site required manual steps to bring up VM host, and Foreman/Puppet installation. The rest, including Ceph components, is automated from there.

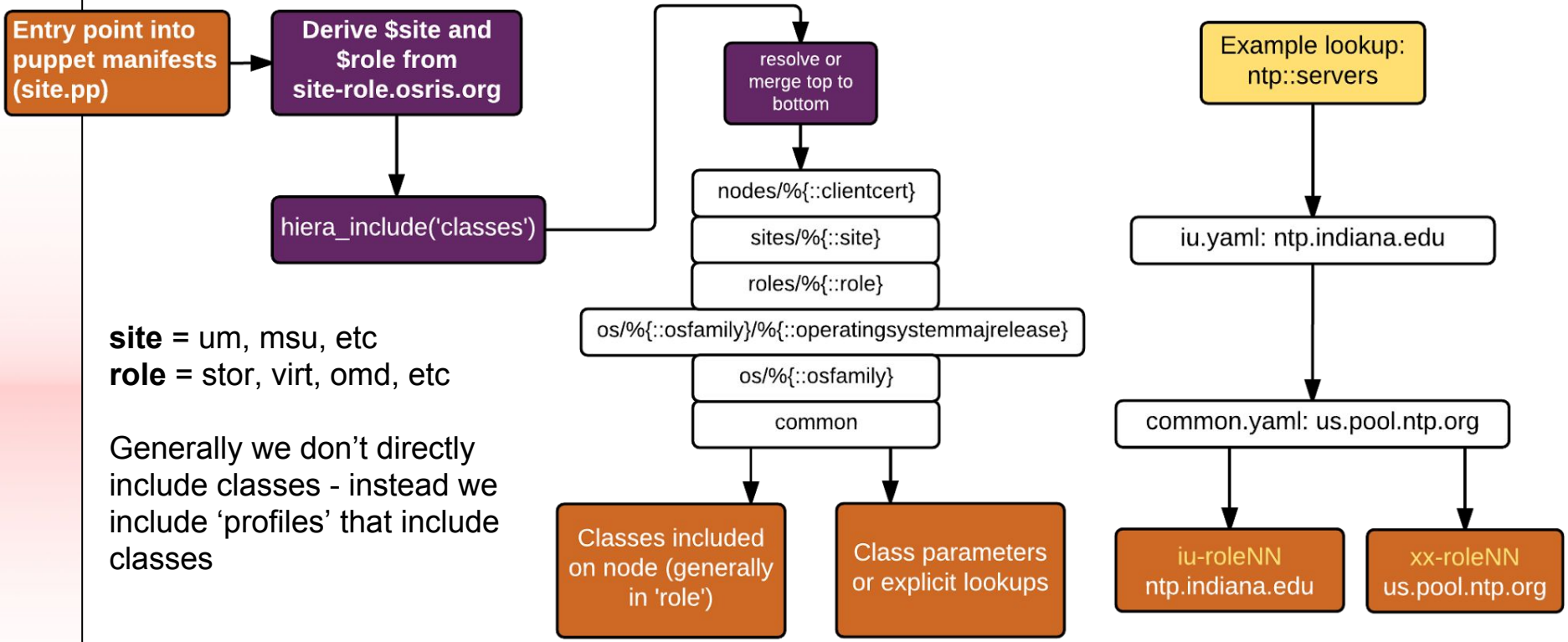
How we deploy



How we manage



How we organize



site = um, msu, etc
role = stor, virt, omd, etc

Generally we don't directly include classes - instead we include 'profiles' that include classes

Deploying a new site

Step 1: Define site specific information in site/sitename.yaml (hiera)

- Network information for provisioning (subnet info, dhcp ranges, etc)
- Ceph CRUSH location
- NTP, DNS, etc

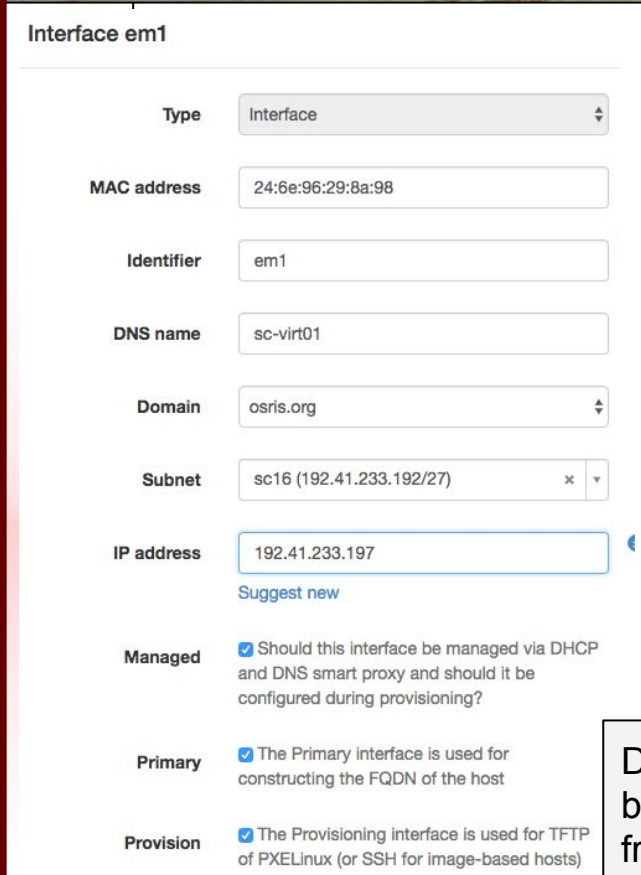
Deploying a new site

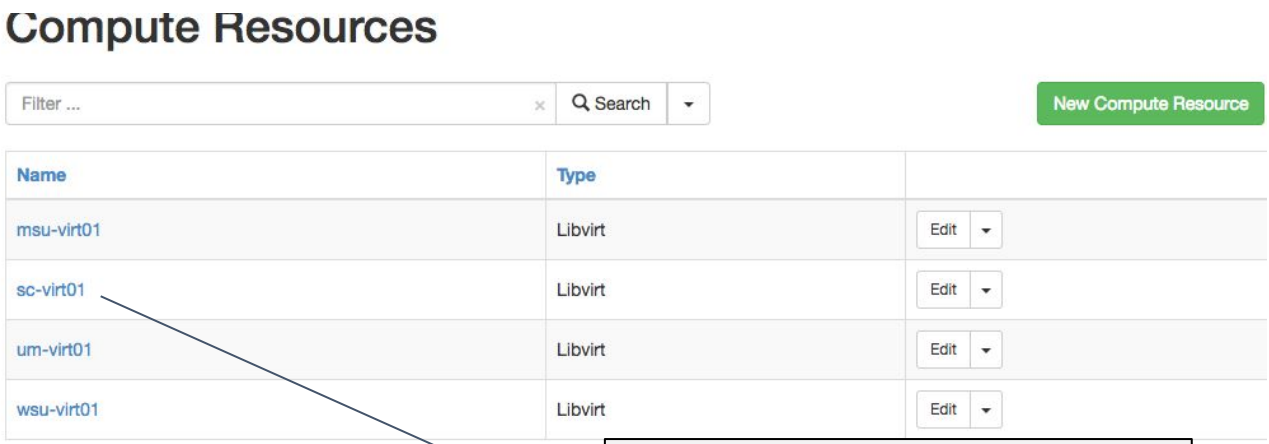
Deploying a new site

Step 2:

- Create a new host in Foreman for the site virtualization host
- Export bootable image
- Install virtualization host, puppet configures necessary packages/services
- Register compute resource in Foreman

Deploying a new site





After build we can define as a compute resource in Foreman

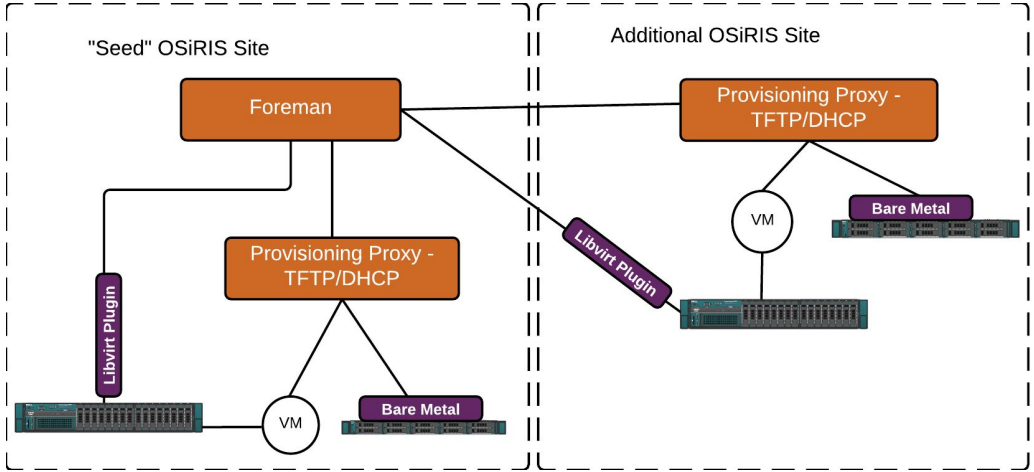
Define host network interface, build by exporting boot image from Foreman

Deploying a new site

Step 3:

- Download VM template for provisioning proxy
- Run VM, configure network

- run puppet to complete configuration and register with master Foreman instance



Deploying a new site

Smart Proxies

Filter ... × Search

New Smart Proxy | Documentation

Name	URL	Features	
msu-prov-be.osris.org	https://msu-prov-be.osris.org:8443	Templates, TFTP, and DHCP	Import subnets
sc-prov.osris.org	https://sc-prov.osris.org:8443	Templates, TFTP, and DHCP	Import subnets
um-puppet.osris.org	https://um-puppet.osris.org:8443	TFTP, Puppet, Puppet CA, and DHCP	Certificates
wsu-prov-be.osris.org	https://wsu-prov-be.osris.org:8443	Templates, TFTP, and DHCP	Import subnets

Puppet triggers provisioning host to register itself as a ‘smart proxy’ in foreman (auth info propogated in configuration)

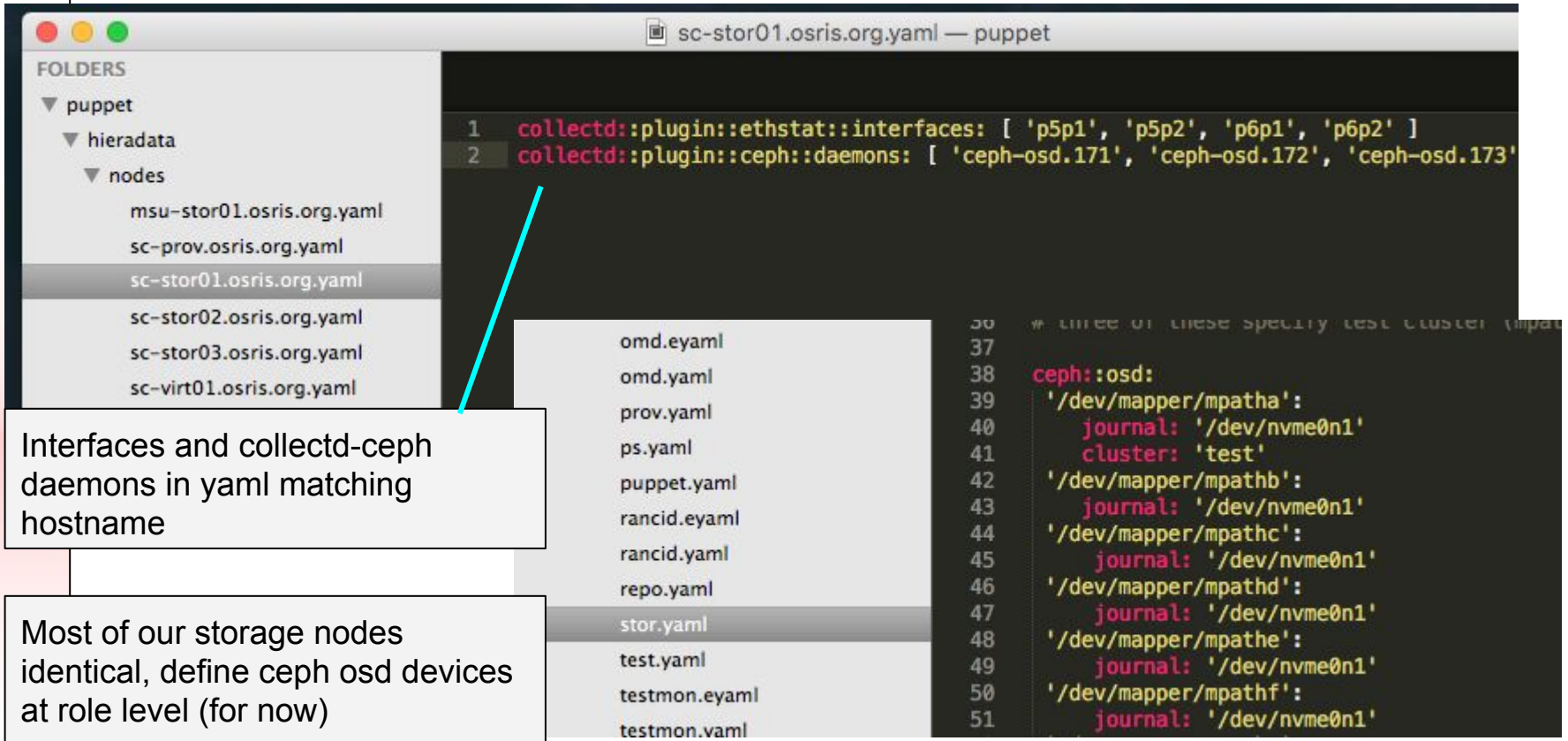
Smart proxy can provide kickstart templates, tftp, dhcp to local network at site

Deploying a new site OSD

In hiera:

- Define the OSD devices used for storage block(s)
- Define the network interfaces to collect stats to Influx/Grafana (collectd-ethstat)
- Define OSD id to collect stats (collectd-ceph)

Deploying a new site OSD



Interfaces and collectd-ceph daemons in yaml matching hostname

Most of our storage nodes identical, define ceph osd devices at role level (for now)

Deploying a new site

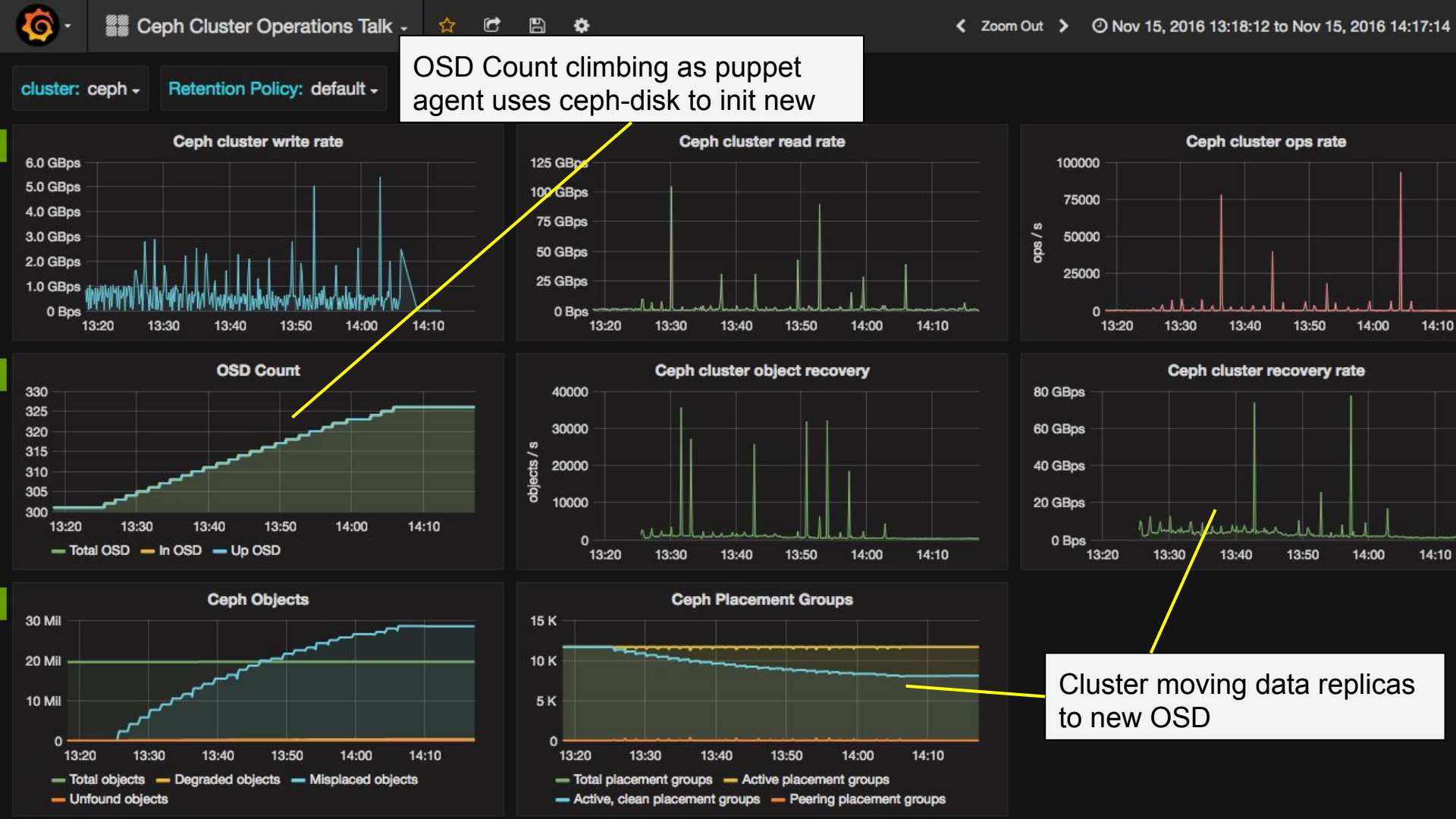
From this point we're ready to build new storage blocks, monitor, mds, grafana, omd, etc.

All of the above automated with puppet, and with Foreman groups defining appropriate partitions or data volumes

Dynamic and Scalable

While OSD are initializing and coming online we have a client data transfer ongoing

You can see the impact on the transfer and the progress of the OSD addition on our monitoring dashboard


Ceph Cluster Operations Talk
Zoom Out
Nov 15, 2016 13:18:12 to Nov 15, 2016 14:17:14
cluster: ceph
Retention Policy: default
OSD Count climbing as puppet agent uses ceph-disk to init new
Ceph cluster write rate
6.0 GBps
5.0 GBps
4.0 GBps
3.0 GBps
2.0 GBps
1.0 GBps
0 Bps
13:20
13:30
13:40
13:50
14:00
14:10
Ceph cluster read rate
125 GBps
100 GBps
75 GBps
50 GBps
25 GBps
0 Bps
Ceph cluster ops rate
100000
75000
50000
25000
0
ops / s
OSD Count
330
325
320
315
310
305
300
Total OSD
In OSD
Up OSD
Ceph cluster object recovery
40000
30000
20000
10000
objects / s
Ceph cluster recovery rate
80 GBps
60 GBps
40 GBps
20 GBps
Ceph Objects
30 Mil
20 Mil
10 Mil
Total objects
Degraded objects
Misplaced objects
Unfound objects
Ceph Placement Groups
15 K
10 K
5 K
Total placement groups
Active placement groups
Active, clean placement groups
Peering placement groups
Cluster moving data replicas to new OSD

DLT Demo

Ongoing during our talk is a demo of live data movement leveraging the Data Logistics Toolkit created at Indiana University.

This demo showcases the movement of USGS earthsat data from capture to storage not only in of the main OSiRIS Ceph cluster but also a dynamic OSiRIS Ceph cluster deployment built at Cloudlab.

Activity can be seen on the Periscope dashboard
http://dev.crest.iu.edu/map/

Questions?

Questions or comments?