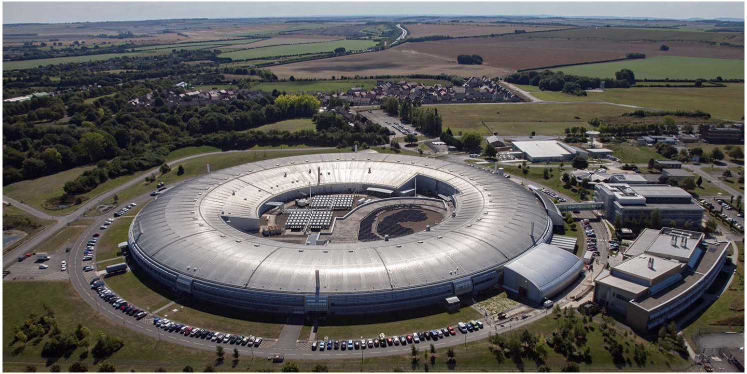
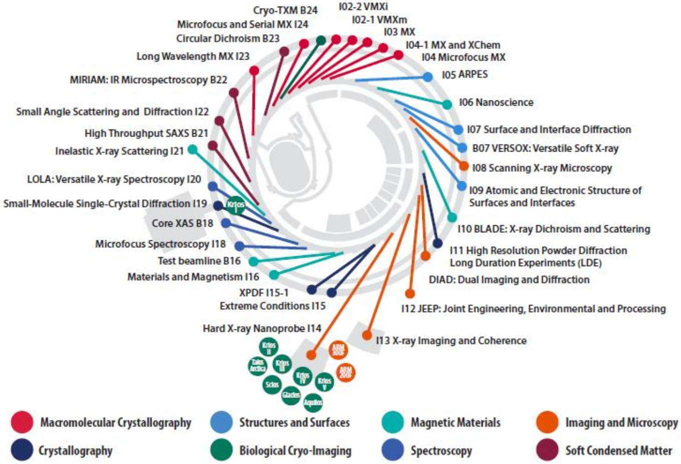
Kubernetes at Diamond Light Source

Thomas Hartland

About Diamond





Cryo-TXM B24

Microfocus and Serial MX I24

Circular Dichroism B23

Long Wavelength MX I23

MIRIAM: IR Microspectroscopy B22

Small Angle Scattering and Diffraction I22

High Throughput SAXS B21

Inelastic X-ray Scattering I21

LOLA: Versatile X-ray Spectroscopy I20

Small-Molecule Single-Crystal Diffraction I19

Core XAS B18

Microfocus Spectroscopy I18

Test beamline B16

Materials and Magnetism I16

XPDF I15-1

Extreme Conditions I15

Hard X-ray Nanoprobe I14

I02-2 VMXi

I02-1 VMXm

I03 MX

I04-1 MX and XChem

I04 Microfocus MX

I05 ARPES

I06 Nanoscience

I07 Surface and Interface Diffraction

B07 VERSOX: Versatile Soft X-ray

I08 Scanning X-ray Microscopy

I09 Atomic and Electronic Structure of Surfaces and Interfaces

I10 BLADE: X-ray Dichroism and Scattering

I11 High Resolution Powder Diffraction Long Duration Experiments (LDE)

DIAD: Dual Imaging and Diffraction

I12 JEEP: Joint Engineering, Environmental and Processing

I13 X-ray Imaging and Coherence

● Macromolecular Crystallography
● Structures and Surfaces
● Magnetic Materials
● Imaging and Microscopy
● Crystallography
● Biological Cryo-Imaging
● Spectroscopy
● Soft Condensed Matter

What's on Kubernetes

- General web services
  - 82 "project namespaces"
  - Gitlab runners, Jupyterhub, k8s stack etc..
- Some data processing that doesn't require HPC cluster
- (Moving towards) beamline controls software

On prem infra

- $\sim$2000 CPU cores (25 node) production cluster
  - 4x V100 GPUs
- Baremetal, 100Gb/s ethernet interconnects + IB
- NVMe storage exposed as Persistent Volumes
- Multi tenancy cluster
  - Self service “personal” namespaces
  - On request “project” namespaces for production deployments

Clusters

- Main class of cluster
  - Argus - production
  - Pollux - pre-preduction
  - Telamon - testing
  - Castor - testing (VMs)

Clusters

- Main class of cluster
  - Argus - production
  - Pollux - pre-preduction
  - Telamon - testing
  - Castor - testing (VMs)
- Special workers
  - Hylas - workers in controls & primary network
  - p38 & i22 - workers are located at beamlines

Clusters

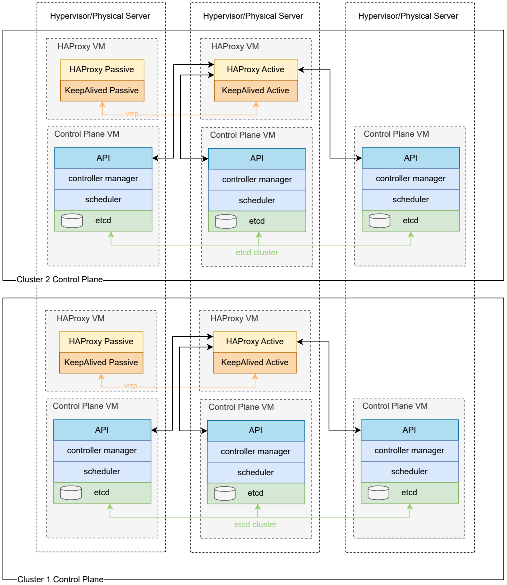
- Main class of cluster
  - Argus - production
  - Pollux - pre-preduction
  - Telamon - testing
  - Castor - testing (VMs)
- Special workers
  - Hylas - workers in controls & primary network
  - p38 & i22 - workers are located at beamlines
- Off-prem STFC/IRIS openstack cloud
  - Orpheus - ~100 node cluster in STFC/IRIS openstack cloud
  - Cepheus - testing cluster for Orpheus

Control plane

- Shared control plane for all on-prem clusters
- Three physical hypervisors running, per cluster:
  - 3 VM kubernetes master nodes
  - 2 VM HAproxy loadbalancers

Control plane

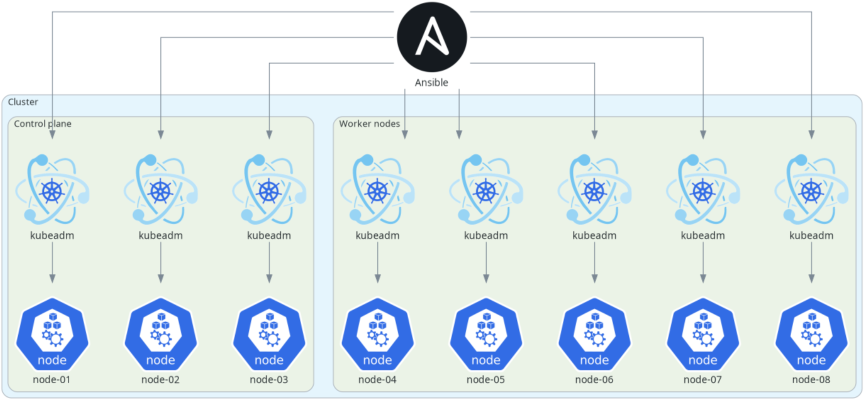
- Shared control plane for all on-prem clusters
- Three physical hypervisors running, per cluster:
  - 3 VM kubernetes master nodes
  - 2 VM HAproxy loadbalancers
- We had a hypervisor memory DIMM failure
  - No users noticed, no API downtime


Hypervisor/Physical Server
Hypervisor/Physical Server
Hypervisor/Physical Server
HAProxy VM
HAProxy Passive
KeepAlived Passive
HAProxy VM
HAProxy Active
KeepAlived Active
vrrp
Control Plane VM
API
controller manager
scheduler
etcd
Control Plane VM
API
controller manager
scheduler
etcd
Control Plane VM
API
controller manager
scheduler
etcd
etcd cluster
Cluster 2 Control Plane
HAProxy VM
HAProxy Passive
KeepAlived Passive
HAProxy VM
HAProxy Active
KeepAlived Active
vrrp
Control Plane VM
API
controller manager
scheduler
etcd
Control Plane VM
API
controller manager
scheduler
etcd
Control Plane VM
API
controller manager
scheduler
etcd
etcd cluster
Cluster 1 Control Plane

Deployment

- Full stack managed with Ansible
  - Managing hardware/VMs
  - Deploying kubernetes (kubeadm)
  - Deploying application stack onto cluster
- Cloud team were the first Ansible users at Diamond

Kubernetes deployment



Stack deployment

- Ansible kubernetes/helm modules
- Monitoring
  - Prometheus, Grafana, Alertmanager, k8s dashboard, fluentd
- Networking
  - Weave CNI, MetalLB, Ingress Nginx, Istio (beta state)
- Policy
  - ResourceQuotas, Kyverno

Highlights of ansible

- Ansible vault for managing secrets
- Coordination between nodes (e.g drain)
- Community modules/roles

Off-prem clusters

- ~8000 CPUs and ~60 A100 GPUs
- VMs provisioned in STFC/IRIS cloud
- Then we deploy kubernetes with ansible as usual

Off-prem clusters

- Cluster not directly exposed to users
- We run htcondor on these nodes
- Users can submit from head nodes in Argus
- Handles "offline" processing
    - Non-realtime, post visit processing

Future clusters

- Cluster per beamline
  - Failure/admin/security domain
  - Downside: many more clusters to manage
- Primary network (airgapped) cluster
  - Needs special consideration
  - Recently set up Harbor as a container image proxy/cache

I’m here to learn

- Multi cluster management (ClusterAPI etc.)
    - How well do they work
    - How well do they handle physical hardware
- Any experience running airgapped clusters?

“The cloud team”:

- Chris Reynolds
- Richard Parke
- Thomas Hartland