
K8s Distribution, Deployment and Monitoring @UChicago



Fengping Hu on behalf of the UChicago team

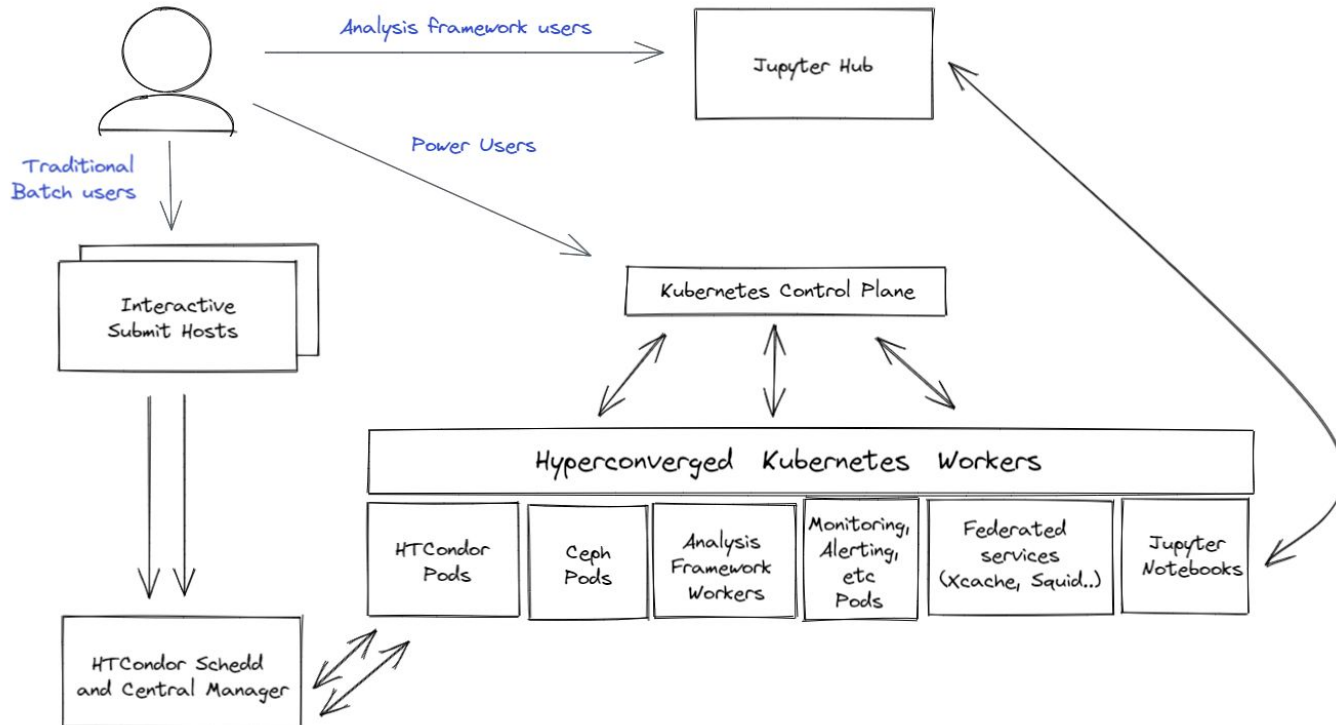
US ATLAS Face-to-Face Computing Facilities Meeting at SLAC
2022.12.01



ENRICO FERMI
INSTITUTE



Kubernetes based AF at UChicago



Why Kubernetes

- A number of interesting new analysis frameworks are being developed, many of which have adopted Kubernetes as an enabling technology
- Run traditional batch on top of kubernetes has proven to work well for us
 - a synergy of cloud native and traditional technology
- We use Rook to orchestrate the Ceph storage solution
- The future is cloud-native



HTCondor on Kubernetes – a synergy of cloud native system and traditional batch scheduling system

- Kubernetes is for containerized workload and it minimize the time and effort required to provision and manager infrastructure resource, but it has limited support of batch workloads
- HTCondor as deployments in Kubernetes
- Two layers of resource allocation
 - Horizontal Pod Autoscaling(HPA) dynamically scale HTCondor deployments to match demands
 - HTCondor job claim resources from the htcondor pod
- Kubernetes monitoring stacks offer realtime metrics

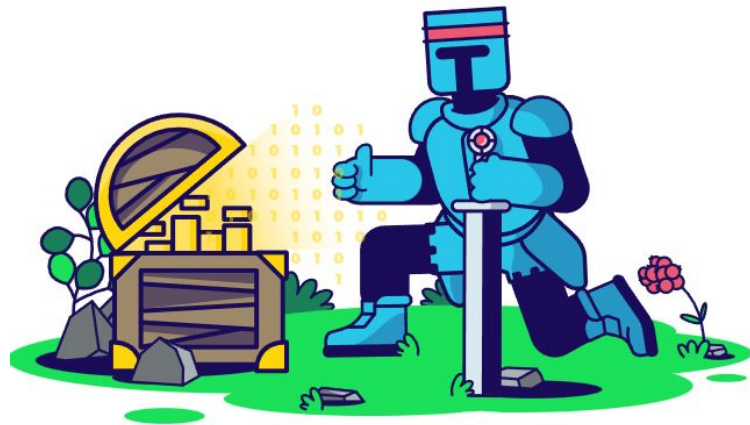


Prometheus



Rook

- Kubernetes-flavored Ceph, closely tracks upstream releases
- Fully manages Ceph cluster on the AF via Kubernetes Operator
- Very popular storage option in the Kubernetes community
- A "graduated" project in the Cloud Native Computing Foundation ecosystem.
 - i.e., multiple organizations committing code, completed security audit, explicitly defined project governance, etc.



Rook/Ceph configuration on the UC AF

- **Goal**: Provide a 1PB shared filesystem (\$DATA) for users of the AF
- 228x 16TB HDDs configured for 3x replication
 - EC is tantalizing for the capacity gains, but we haven't had a good experience with it elsewhere.
- Each node has a dedicated NVMe for Bluestore database (Metadata).
- Each node has a second dedicated NVMe for CephFS metadata
- CephFS configured with 6 MDSes
 - 3 active, 3 standby – floating on the Hyperconverged nodes.
- Filesystem mountable within Kubernetes and outside.



Kubernetes installation

- Installation

- Bootstrapping clusters with kubeadm
- Installing Kubernetes with kOps
- Installing Kubernetes with Kubespray

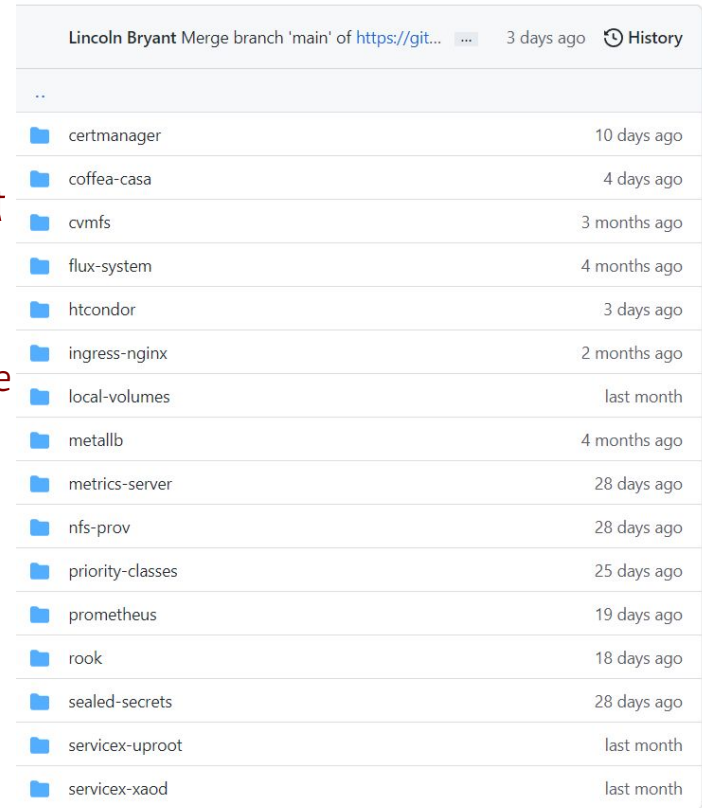
- Learning

- [Kubernetes The Hard Way](#)
-



Tools for declarative deployment – FluxCD

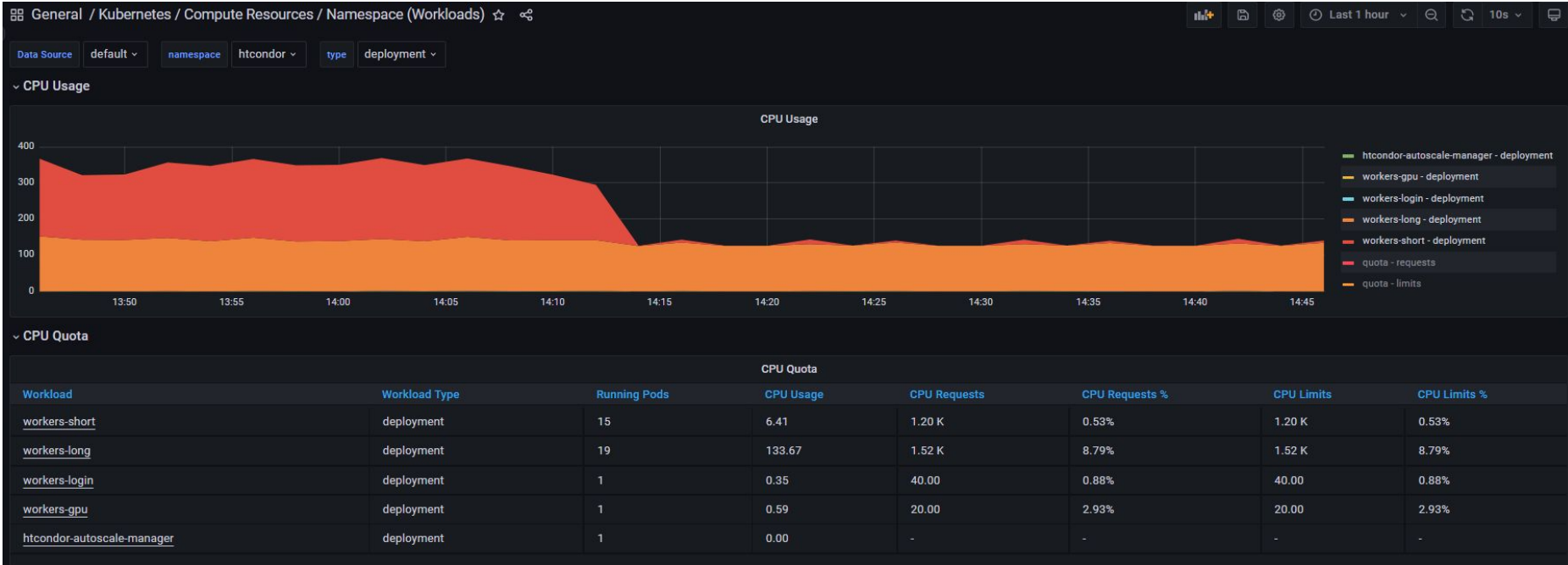
- **Flux CD** Graduated yesterday
 - The 18th CNCF project to graduate
- **Flux CD** – “GitOps” style application deployment
 - All configuration lives in GitHub, installation/updates/removal all happen via the Flux operator that uses Git as a single source of truth for the cluster.
 - All of the basic Kubernetes extensions are loaded into the Flux repo (Ingress, Load Balancer, monitoring, certificate management, etc)
 - Ceph, HTCondor, etc are also managed by Flux



Lincoln Bryant Merge branch 'main' of https://git... 3 days ago History

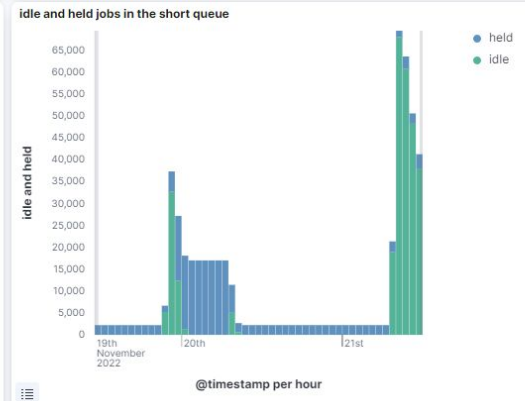
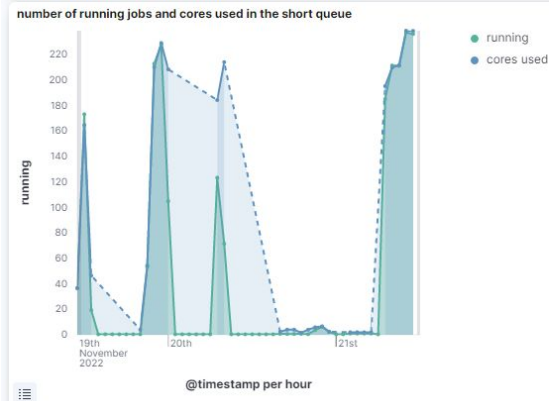
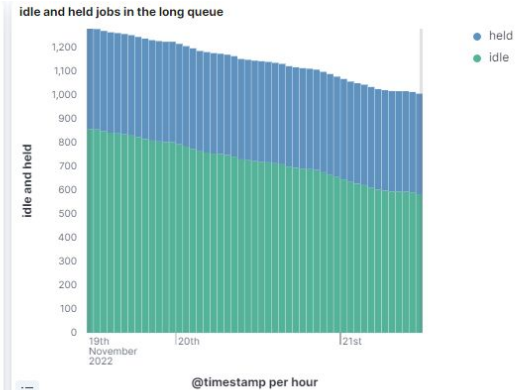
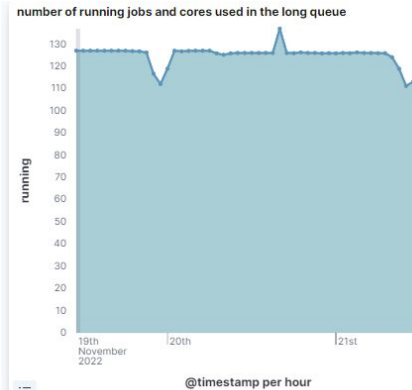
..	
certmanager	10 days ago
coffea-casa	4 days ago
cvmfs	3 months ago
flux-system	4 months ago
htcondor	3 days ago
ingress-nginx	2 months ago
local-volumes	last month
metallb	4 months ago
metrics-server	28 days ago
nfs-prov	28 days ago
priority-classes	25 days ago
prometheus	19 days ago
rook	18 days ago
sealed-secrets	28 days ago
servicex-uproot	last month
servicex-xaod	last month

Monitoring – prometheus + grafana



Monitoring – Elastic Search+ Kibana

- Metricbeat
- Logstash
- Custom metrics



Multi tenancy support

Faire share between kubernetes native workloads like servicex

Kueue?

