Kubernetes experiences @ CNAF

Francesco Sinisi (INFN-CNAF)

On behalf of many people & groups at CNAF & INFN:

Federico Fornari, Cristina Duma, Stefano Bovina, Enrico Vianello, Alessandro Costantini, Diego Ciangottini, Daniele Cesini, Lucia Morganti...

pre-GDB on Kubernetes, CERN, 7 June 2022

Overview

Experiences at CNAF in using K8S for various purposes:

- Storage: deploy EOS+CEPH and IBM Spectrum Scale (GPFS)
- Testbeds for different experiments (i.e VIRGO low latency analysis)
- Software Development: C.I./C.D. and testing tbs
- K8S HA cluster for hosting production services
- INFN Information System Department

Kubernetes at CNAF: why and how

Why:

- It allows to deploy containerized apps anywhere and manage them exactly in the same way everywhere

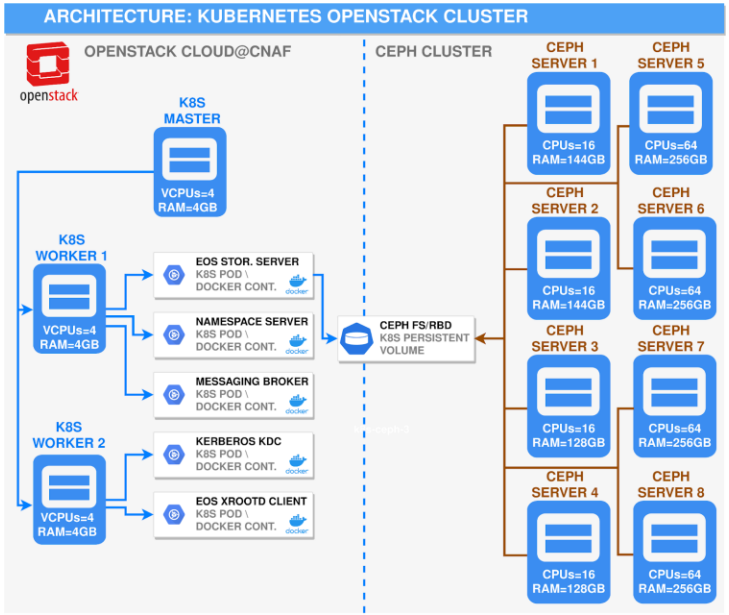
How:

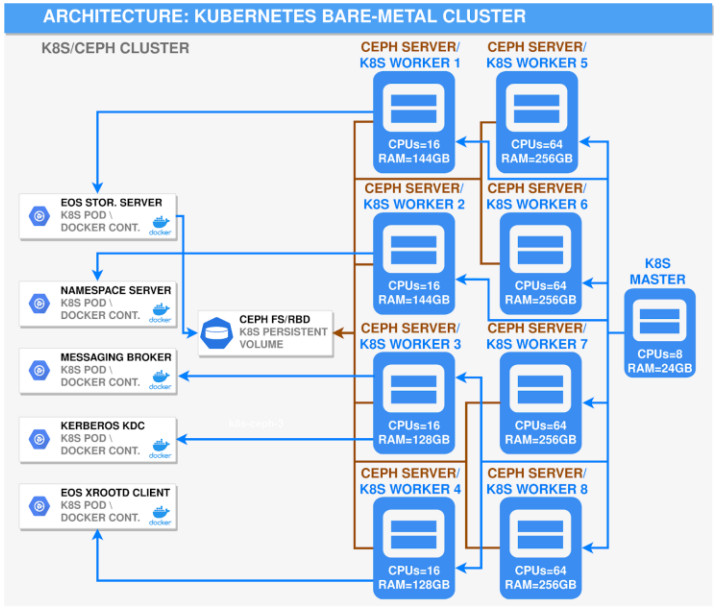
- Several solutions exist to configure and deploy Kubernetes.
- Open-Source networking, storage and monitoring plugins.
- Deploy K8s applications with Helm: the package manager for Kubernetes.
- Each cluster is tailored according to the needs of communities requesting it.

Kubernetes Deployment Tools

- **Kubeadm** is the least automated tool, but provides a simple way to try out Kubernetes (possibly for the first times).
- **Rancher Kubernetes Engine** (RKE) simplifies and automates installation and operation of Kubernetes cluster.
- **Kubespray** ansible playbook automates the installation of dependencies and creation of the cluster.
- **Rancher** is designed to deploy and manage multiple Kubernetes clusters regardless of the location or provider.
- **OpenShift** is an enterprise-ready Kubernetes container platform built for an open hybrid cloud strategy

Storage - EOS & Ceph integration with K8s



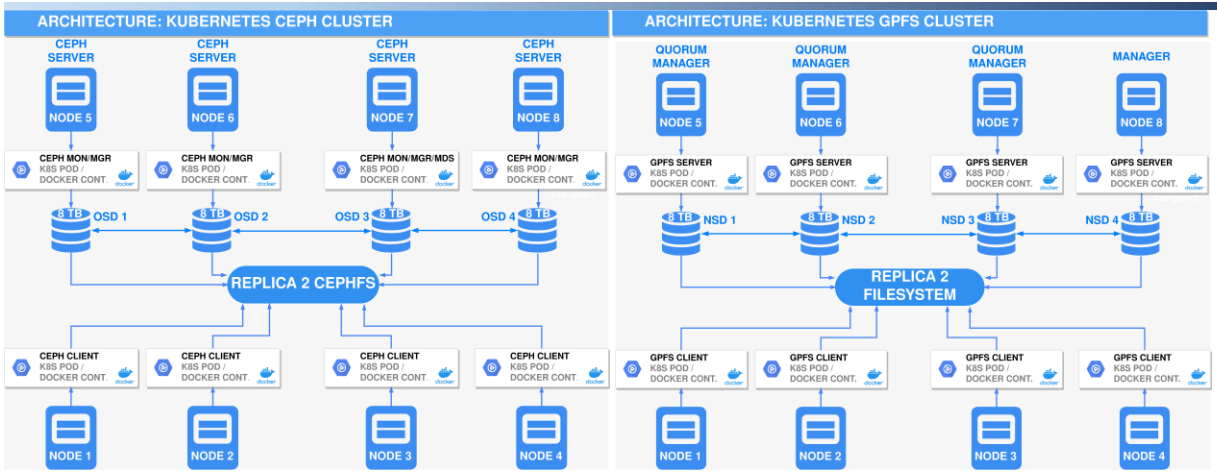


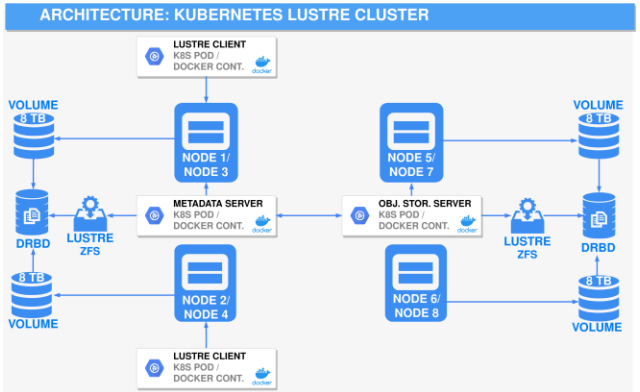
Different deployment scenarios on **cloud** and **bare-metal** to test integration between **Ceph** (backend) and **EOS** (frontend) with **K8s**.

EOS services in K8s Pods with storage provided by Ceph Persistent Volumes. K8s facilitates failover, scalability and management of EOS services, yielding good performance results (see here for further details).

Slide courtesy of Federico Fornari (Storage, INFN-CNAF)

Storage - GPFS/Ceph/Lustre K8s deployment





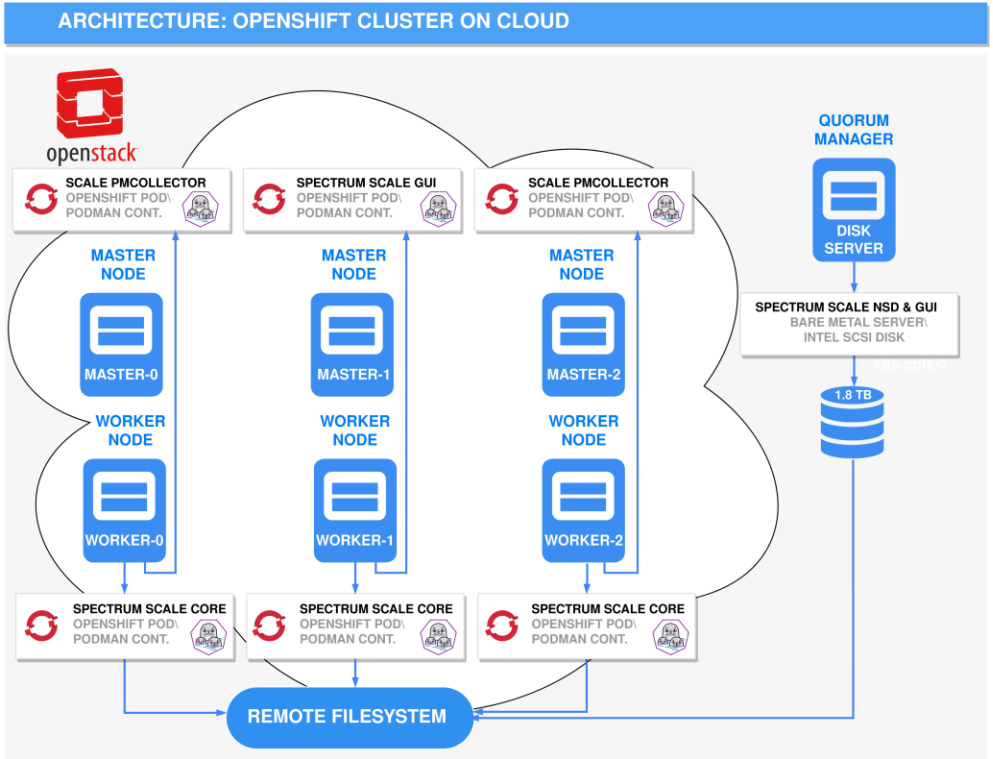
Different types of **distributed file systems clusters** have been deployed using **K8s on bare-metal**, both on server and client side, evaluating performance differences.

Tests proved feasibility of the containerized setups:

- **no performance loss** w.r.t. non-containerized setups
- reciprocally **comparable throughput scores**
- no more than one container/server (file system kernel modules not loadable by multiple containers, FUSE required but affects performance)
- See here and here for further details.

Slide courtesy of Federico Fornari (Storage, INFN-CNAF)

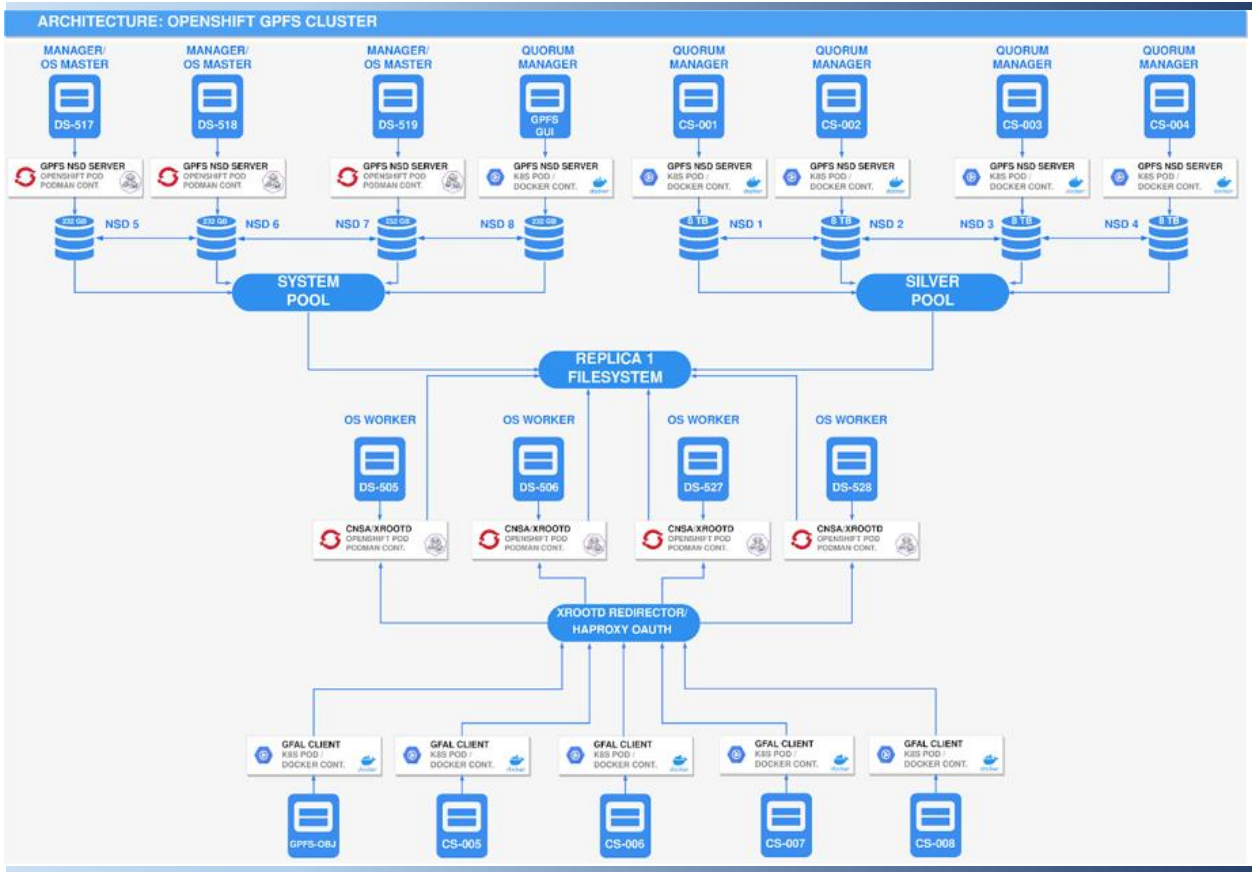
Storage - IBM Spectrum Scale/OpenShift (cloud)



- IBM Spectrum Scale introduced **CNSA (Container Native Storage Access)**
- A remote GPFS file system can be accessed in OpenShift Pods (clients)
- OpenShift leverages Podman as container engine and the cluster can be deployed on cloud or bare-metal
- Data management services(WebDAV, XRootD) can be instantiated in Pods (easily scalable)
  - Openstack allows to implement autoscaler functionality for OpenShift workers
  - The number of OpenShift worker nodes is automatically adjusted based on load

Slide courtesy of Federico Fornari (Storage, INFN-CNAF)

Storage - IBM Spectrum Scale/OpenShift (bare-metal)



- OpenShift cluster can also be deployed on bare-metal resources:
  - no Openstack, see here (setup procedure)
- Within CNAF Storage team some **performance tests** for XRootD and StoRM WebDAV have been carried out with IBM CNSA Pods on OpenShift bare-metal cluster yielding **positive results**
- OpenShift on bare-metal has no autoscaler

Slide courtesy of Federico Fornari (Storage, INFN-CNAF)

Testbeds/clusters for experiments/projects - VIRGO case

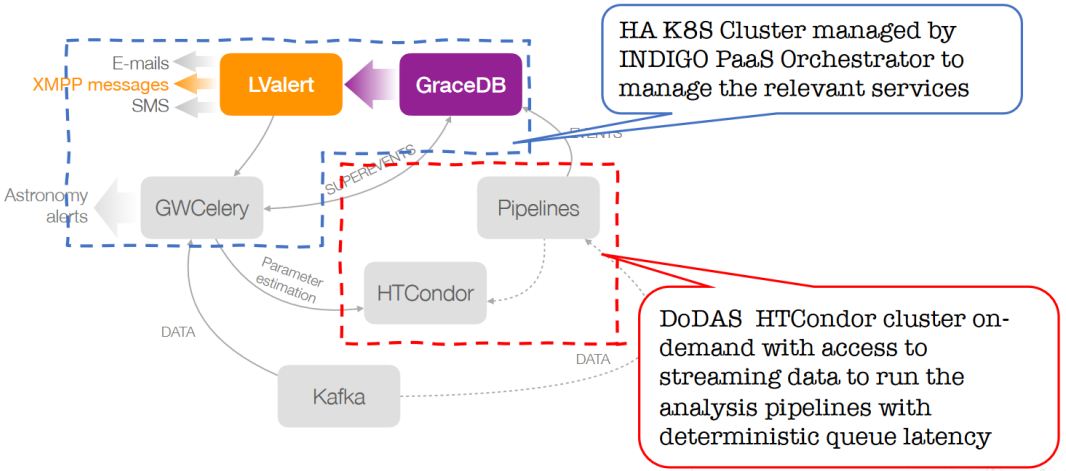
3 K8s clusters dedicated to VIRGO project (for a total of 176 of CPU cores, 352 GB of memory and 3.68 TB of storage):

- Testbed VIRGO:
  - 1 master + 12 workers deployed with INFN Cloud Orchestrator and imported in Rancher
  - Plug-in (Grafana, Prometheus, Longhorn, alerting)
- Testbed for storage benchmarking:
  - 1 master + 3 workers deployed with Kubespray
  - Plug-in (K8s dashboard, Grafana, Prometheus, Longhorn)
  - Benchmarking persistent disk performance (especially latency) with FIO
- Gitlab-runner:
  - 1 master + 10 workers deployed with Kubeadm
  - Plug-in (Grafana, Prometheus, Kubernetes Dashboard)

K8S cluster for VIRGO Testbed

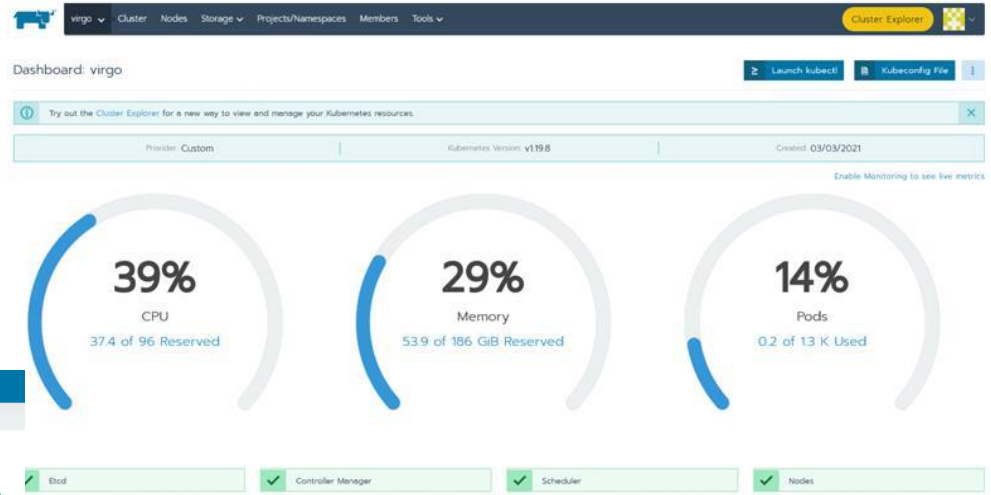
The relevant services to be deployed for the Virgo (and more generally IGWN) low-latency alert generation infrastructure are:

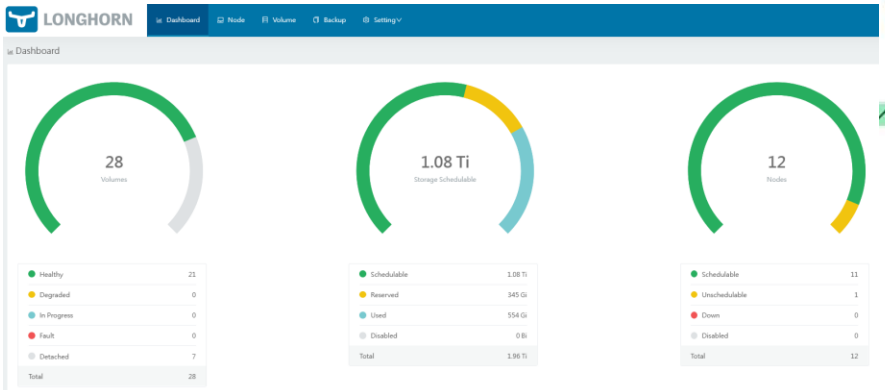
- The Gravitational-Wave Candidate Event Database (**GraceDB**): it provides a centralized location for aggregating and retrieving information about candidate gravitational-wave events.
- The LIGO-Virgo Alert Network (**LVAlert**): is a prototype notification service built on XMPP to provide a basic notification tool which allows multiple producers and consumers of notifications.
- **GWCelery**: is a service for annotating and orchestrating IGWN alerts, built on top of the Celery distributed task queue.



Testbed VIRGO

- Managed by Rancher:
  - 1 master + 12 workers
  - Plug-in (Grafana, Prometheus, Longhorn, alerting)
  - Scaling, resizing...

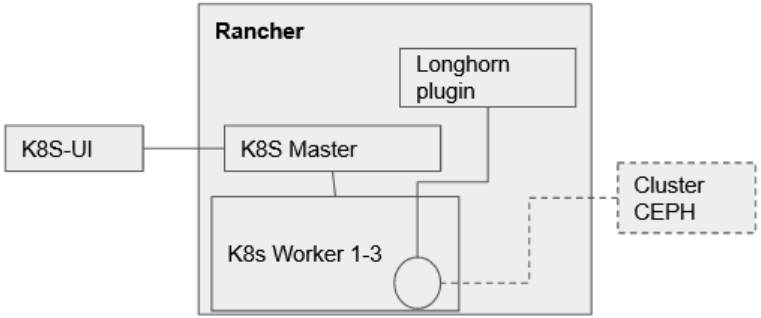


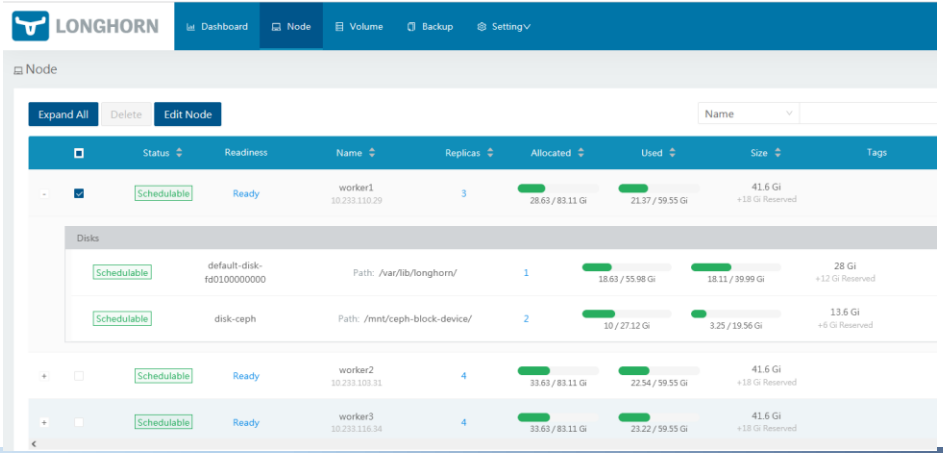


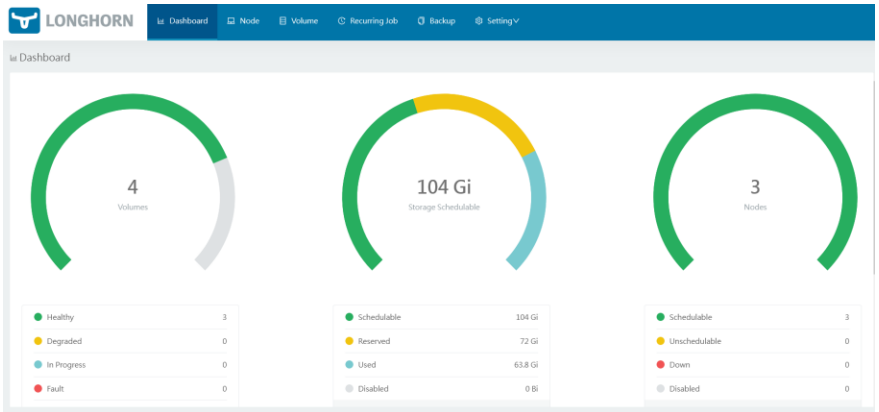
Testbed for storage benchmarking

- Longhorn with:
  - local storage
  - CEPH storage mounted via RBD
  - NFS provisioner (local node storage)
- CEPH storage provisioned via RBD

Deployed with Kubespray

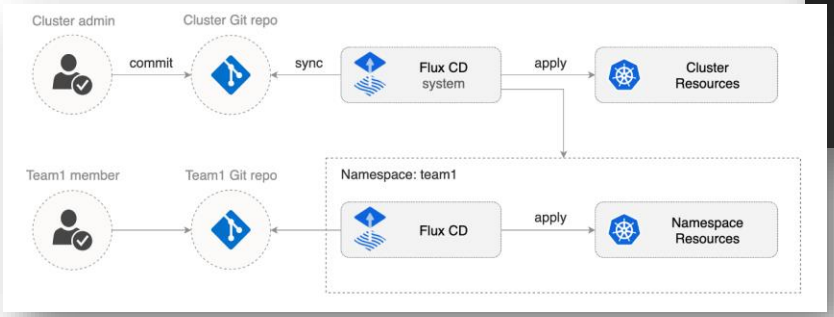


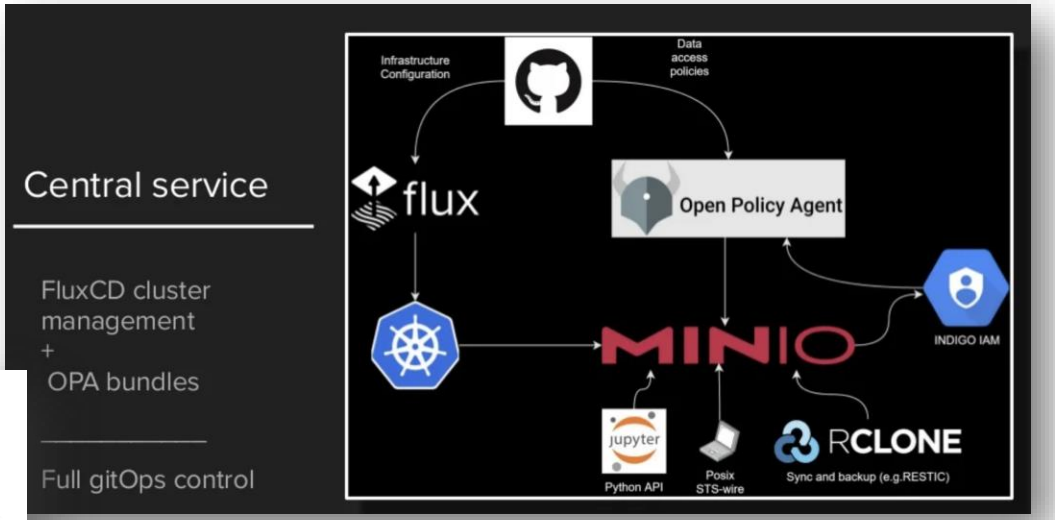




MinIO & FluxCD

- For one of the projects present at CNAF we are using another INFN Cloud solution: **MinIO** on top of a K8S cluster, created with RKE, managed using **FluxCD**.





Slide courtesy of Diego Ciangottini (INFN-PG)

Software Development k8s cluster

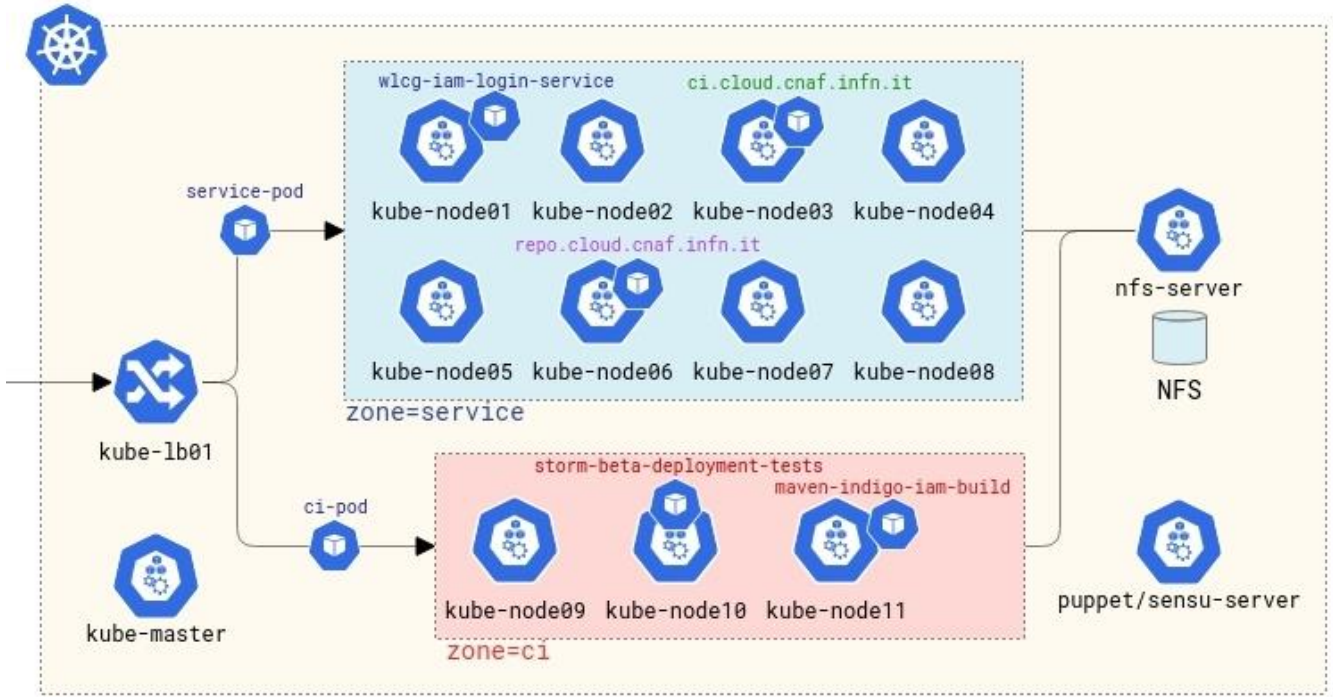
CNAF Software Development started using a k8s cluster since 2016 for the following activities:

- Service deployment
  - Mainly dedicated to INDIGO IAM deployments
  - ~24 namespaces dedicated to different INDIGO IAM instances
  - 8 dedicated nodes for services (identified via node label "zone=service")
- Continuous Integration support
  - Jenkins server deployed (https://ci.cloud.cnaf.infn.it/)
  - Each Jenkins job triggers the provisioning of an agent which is a k8s pod where job is run
  - 3 dedicated nodes for CI jobs (identified via node label "zone=ci")
- Public rpm repository + Maven public repository
  - Nexus server deployed (https://repo.cloud.cnaf.infn.it/)
  - StoRM, VOMS, Argus and INDIGO IAM released rpms are stored here

Software Development k8s cluster

The SD k8s cluster components:

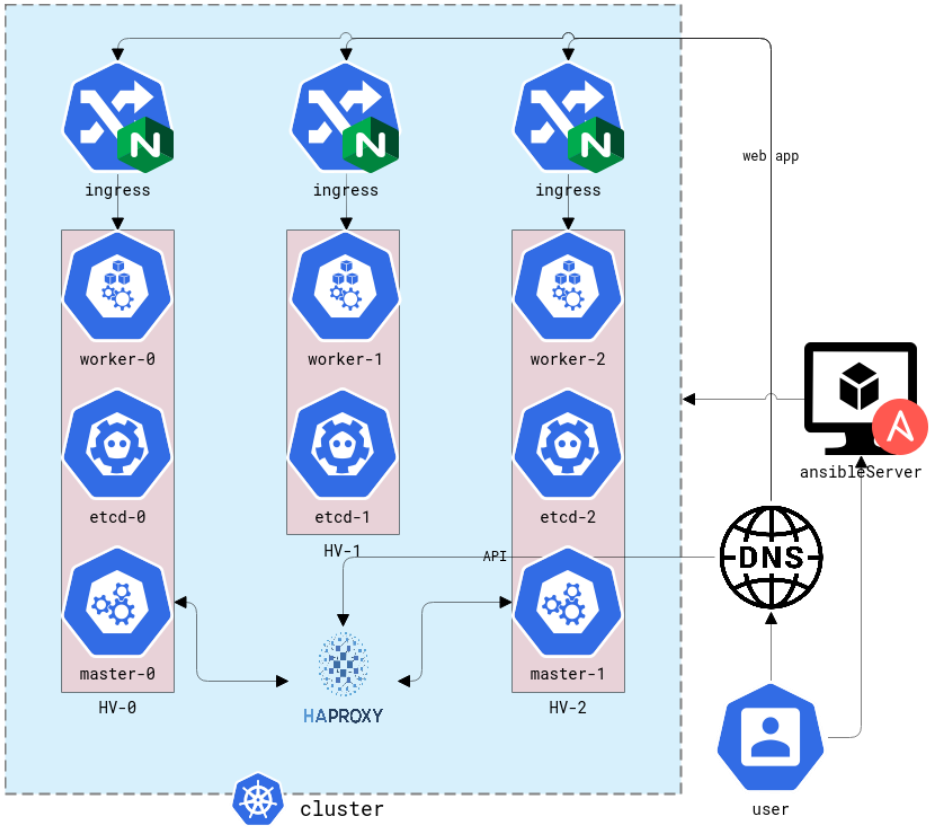
- Ingress node
- Master node
- Puppet/Sensu (internal) server node
- 11 worker nodes
- NFS server node



Slide courtesy of Enrico Vianello (SD, INFN-CNAF)

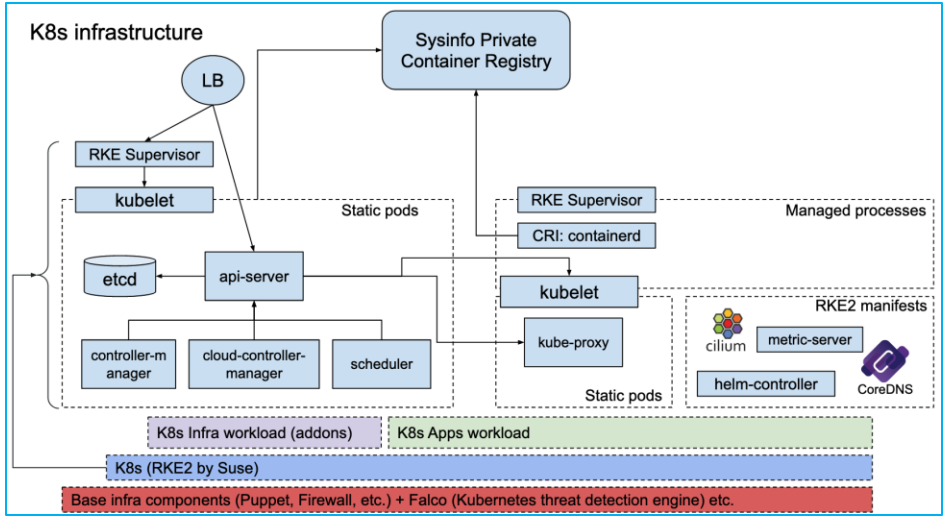
Kube HA for deploying production services

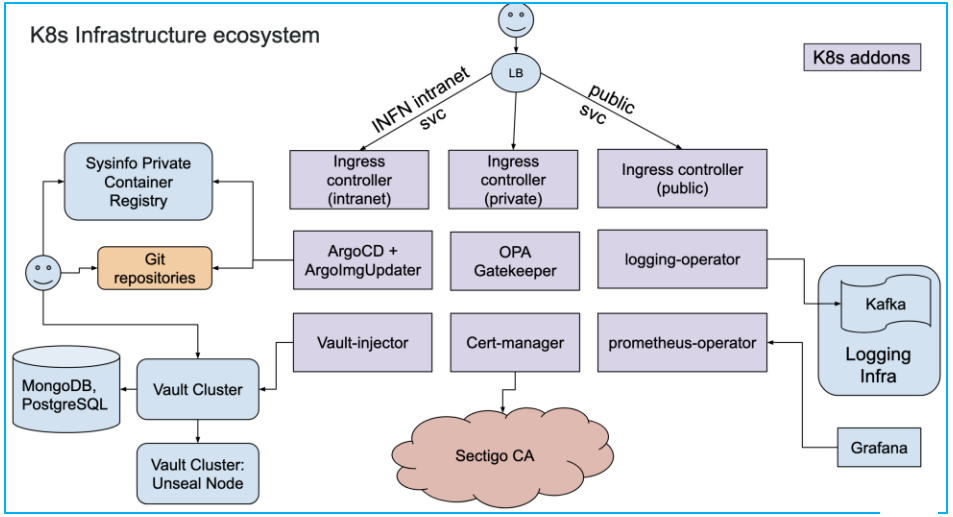
- Cluster is deployed from an "external" *ansibleServer* with Kubespray
- The VMs run on separate HVs
- Web applications (IAM, Grafana, etc.) are available, until a worker with ingress is active
- APIs are available, even without 1 Master
- The cluster is available, until the ETCD quorum is guaranteed



Software development infra for INFN Info System

- New infrastructure for the SysInfo software development group activities
- Interacting with monitoring/accounting, container registry (Harbor), secret management tool (Vault)





Slides courtesy of Stefano Bovina (SysInfo, INFN-CNAF)

Conclusions

- R&D activites have been performed, although we still working on it, but for what regards the "cloud-world" we are doing a massive use of K8S for all the situations and collaborations that request it.
- Steep learning curve to apply and tune solutions for different use-cases
  - the INFN Cloud solutions act as smoother start
  - we intend to continue on the road of adopting the use of K8S for other services, including the grid ones also following the other WLCG community expriences.