Multi-Cloud Federated Kubernetes at CERN

Clenimar Filemon @clenimar
clenimar@lsd.ufcg.edu.br

Ricardo Rocha @ahcorporto
ricardo.rocha@cern.ch
What is 96% of the universe made of?

What was the state of matter just after the Big Bang?

Why isn’t there anti-matter in the universe?
Collisions

~40 MHz
~ 1PB/sec

Huge Data

Trigger

L1

~ 100 kHz

Hardware Filter

Still Big

Trigger

HL

~ 1 kHz

Software Filter

Still Big

Raw Data

~ 1-10 GB/s
Distributed Computing

200+ Sites
700,000 Cores
~400,000 Jobs
~30 GiB/s

CERN
T1
T2

Reconstruction
Calibration
Simulation
Analysis
Motivation for Federation

Periodic Load Spikes

*International Conferences, Reconstruction Campaigns*

Simplification

*Monitoring, Lifecycle, Alarms*

Deployment

*Uniform API, Replication, Load Balancing*
OpenStack Magnum

An OpenStack API Service that allows creation of container clusters

- Use your keystone credentials
- You choose your cluster type
- Multi-Tenancy
- Quickly create new clusters with advanced features such as multi-master
OpenStack Magnum

Single command cluster creation

$ openstack coe cluster create --cluster-template kubernetes --node-count 100 ... mycluster

$ openstack cluster list
+-----------------+----------------+------------+--------------+-----------------+
| uuid | name   | node_count | master_count | status          |
+-----------------+----------------+------------+--------------+-----------------+
| .... | mycluster | 100        | 1            | CREATE_COMPLETE |
+-----------------+----------------+------------+--------------+-----------------+

$ $(magnum cluster-config mycluster --dir mycluster)

$ kubectl get pod

$ openstack coe cluster update mycluster replace node_count=200
Kubernetes
Kubernetes

Multiple type os Resources
- Pod, Service, Deployment, DaemonSet, Job, ...
- Requests and Limits
- Retrial Policies
- Taints and Tolerations
- And much more...

```yaml
apiVersion: batch/v1
kind: Job
metadata:
  name: pi-with-timeout
spec:
  backoffLimit: 5
  activeDeadlineSeconds: 100
template:
  spec:
    containers:
      - name: myjob
        image: python
        command: ["/myjob.py"]
        resources:
          limits:
            cpu: "1"
        restartPolicy: Never
```
Use Case

CERN Large Scale Batch Systems - HTCONDOR
**Sched**

**Collector**

**StartD**

**Matchmaking with ClassAds**

**Extensive Experience in HEP**

**Fair Share**

**Running Virtualized**

**Preemption**

**External Storage and Networking**

AcctGroup = "ATLAS"
JobPrio = 0
RequestCpus = 2
RequestMemory = 4260

CERNEnvironment = “production”
Datacenter = “meyrin”
HasMPI = true
TotalCpus = 8
TotalMemory = 22500
Sched → Collector → Negotiator → StartD

AcctGroup = "ATLAS"
JobPrio = 0
RequestCpus = 2
RequestMemory = 4260
...

CERNEnvironment = “production”
Datacenter = “meyrin”
HasMPI = true
TotalCpus = 8
TotalMemory = 22500
...

Matchmaking with ClassAds
Fair Share
Preemption

Extensive Experience in HEP
Running Virtualized
External Storage and Networking
kubefed init cern-condor --host-cluster-context=condor-host ...

openstack coe federation create --host-cluster condor-host cern-condor
kubefed join --host-cluster-context ... --cluster-context ... atlas-recast-y

openstack coe federation join cern-condor atlas-recast-x atlas-recast-y
apiVersion: apps/v1
kind: DaemonSet
metadata:
  name: {{ template "condor-startd.fullname" . }}

spec:
  hostNetwork: true
  containers:
    - name: {{ .Chart.Name }}
      image: "{{ .Values.image.repository }}:{{ .Values.image.tag }}"
      securityContext:
        privileged: true
      livenessProbe:
        exec:
          command:
            - condor_who

https://gitlab.cern.ch/helm/charts/tree/master/condor-startd
Storage

- Building on well established deployments
- Software distribution handle by CVMFS (hierarchical squid caches)
- Access to physics data done directly
1. An existing Magnum cluster in an OpenStack environment is to be extended using external resources. An external cluster endpoint (deployed in AWS, Azure, GKE, another OpenStack or cloud) can be added to an existing Magnum federated cluster, including the complex setup and management of cluster credentials.

2. A project has several existing clusters which it would like to expose to a set of users in a single endpoint, without disrupting existing users of each cluster.

3. A set of Magnum clusters is created, each with different characteristics: node flavor, storage setup, etc. Federating them together forms a heterogeneous cluster.

*API and Persistence Layer already merged, kubernetes support ongoing*
Kubernetes SIG Multi-Cluster

- Home of the Federation work
- Currently working on Federation v2, Cluster Registry, Multi Cluster Ingress

https://github.com/kubernetes/community/tree/master/sig-multicluster
Demo

Reusable Analysis Workflows - RECAST

Summary

• Federation support in Kubernetes is ready
  • Ongoing development for the v2 API, with significant changes

• OpenStack Magnum support coming in Rocky

• Already in use at CERN
  • Started with a legacy application, limited integration
  • Expanded to a cloud native implementation, with great results

• Great support from **OpenStack** and **Kubernetes** communities
Questions?

Clenimar Filemon
clenimar@lsd.ufcg.edu.br
@clenimar

Ricardo Rocha
ricardo.rocha@cern.ch
@ahcorporto