OSG use of the PRP Kubernetes cluster

Pre-GDB @ CERN
Presented by Igor Sfiligoi, UCSD

Dec 2019
Kubernetes as a resource manager

**Industry standard**
- Large and active development and support community

**Container based**
- More freedom for users

**Flexible scheduling**
- Allows for easy mixing of service and user workloads

Dec 2019
HTCondor as the OSG helper

PRP wanted to give opportunistic resources to Open Science Grid (OSG) users
- Since they can tolerate preemption

But OSG does not have native support for Kubernetes
- Supports only resources provided by batch systems

We thus instantiated an HTCondor pool
- As a fully Kubernetes/Containerized deployment
HTCondor in a (set of) container(s)

Putting HTCondor in a set of containers is not hard
- Just create an image with HTCondor binaries in it!
- Configuration injected through Kubernetes pod config

HTCondor deals nicely with ephemeral IPs
- The Collector must be discoverable – Kubernetes service
- Everything else just works from there

Persistency needed for the Schedd(s)
- And potentially for the Negotiator, if long term accounting desired
- Everything else can live off ephemeral storage

Dec 2019
HTCondor in a (set of) container(s)

Putting HTCondor in a set of containers is not hard

• Just create an image with HTCondor binaries in it!
• HTCondor deals nicely with ephemeral IPs
• Configuration injected through Kubernetes pod config
• Everything else just works from there
• The Collector must be discoverable – Kubernetes service
• HTCondor deals nicely with ephemeral IPs
• And potentially for the Negotiator, if long term accounting desired
• Everything else can live off ephemeral storage

Dec 2019

OSG HTCondor-CE runs alongside the Schedd
Service vs Opportunistic

Collector and Schedd(s) deployed as high priority service pods

- Should be running at all times
- Few pods, not high CPU or GPU users, so OK
- Using Kubernetes Deployment to re-start the pods in case of HW problems and/or maintenance
- Kubernetes Service used to get a persistent routing IP to the collector pod

Startds deployed as low priority pods

- Hundreds of pods in the Kubernetes queue at all times, many in Pending state
- HTCondor Startd configured to accept jobs as soon as it starts and forever after
- If pod preempted, HTCondor gets a SIGTERM and has a few seconds to go away

And OSG HTCondor-CE

Dec 2019
Kubernetes priorities

- Low priority pods only start if no demand from higher priority ones
- Low priority pods killed the moment a high priority pod needs the resources
- Just keep enough low-priority pods in the system

https://kubernetes.io/docs/concepts/configuration/pod-priority-preemption/

Dec 2019
• Most OSG users run their workloads inside containers
• But OSG is running “generic HTCondor worker node pods” in Nautilus
• How do we launch another container from inside a running container?
  • Can we do it without requiring any privileges?
  • Turns out, it is quite trivial with recent Linux Kernels (4.15+)
• Singularity can run from inside an (unprivileged) Docker container – like in Nautilus k8s
  • Most Nautilus nodes have the updated Kernel
CVMFS and Unprivileged Containers

OSG Pods run only unprivileged containers
- Like all other users in the k8s cluster
- To minimize risk

CVMFS cannot be mounted from inside an unprivileged container
- Using FUSE is a privileged operation
- Unless (maybe) using properly configured latest kernel, which we cannot assume

Installing CVMFS bare-metal not an option
- We want to have only k8s at bare-metal level
Kubernetes Container Storage Interface (CSI)

- Provides a standard way to add custom filesystems

Driver deployed by cluster admin

- Privileged operation
- But admin controls and can inspect the container images

User Pods see it as an additional mount option

- No privileges needed

https://kubernetes.io/blog/2019/01/15/container-storage-interface-ga/

Dec 2019
Kubernetes CSI to the Rescue

Kubernetes Container Storage Interface (CSI)
- Provides a standard way to add custom filesystems

Driver deployed by cluster admin
- Privileged operation
- But admin controls and can inspect the container images

User Pods see it as an additional mount option
- No privileges needed

https://kubernetes.io/blog/2019/01/15/container-storage-interface-ga/
Some stats - GPUs

https://gracc.opensciencegrid.org/dashboard/db/gpu-payload-jobs-summary?orgId=1&from=now-6M&to=now&var-ReportableVOName=All&var-Project=All&var-Facility=SDSC-PRP&var-User=All&var-ExitCode=All&var-Probe=All&var-interval=1M
Some stats - CPUs

https://gracc.opensciencegrid.org/dashboard/db/payload-jobs-summary?orgId=1&from=now-6M&to=now&var-ReportableVOName=All&var-Project=All&var-Facility=SDSC-PRP&var-User=All&var-ExitCode=All&var-Probe=All&var-interval=1M&var-Organization=All
Acknowledgments

This work was partially funded by US National Science Foundation (NSF) awards MPS-1148698, CNS-1456638, CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC 1450871, OAC-1659169 and OAC-1841530.