Kubernetes-native batch computing at CA-VICTORIA-WESTGRID-T2

Ryan Taylor, Jeff Albert, Danika MacDonell,
Fernando Barreiro Megino, Frank Berghaus

University of Victoria

compute canada | calcul canada

Deployment with KUBESPRAY

- Reasons to use Kubespray
  - Infrastructure as Code
  - Familiarity with Ansible, Terraform
  - Leverages kubeadm
  - Supports many OSs, cloud providers, CNI providers, CRI runtimes, deployment options, addons ...
  - Maximum flexibility, extensibility, control
  - Not locked in, freedom to change mind
  - Scrutable
- Reasons to not use Kubespray
  - You want simplicity (or at least complexity hiding)
  - You want an opinionated deployment method
  - You want a one-click cluster and don't care about the details

openstack®

Arbutus

UVic Kubernetes deployment

- UVic Kubespray modifications
  - integration of additional Ansible roles/inventory
  - reworked for multi-cluster management
  - fixes/improvements contributed upstream as much as possible
  - clusters can be deployed or rebuilt from scratch in ~ 5-10 minutes
- CVMFS installed on hosts, mounted to pods via hostPath volume
- `atlas-grid-centos7` published to `/cvmfs/unpacked.cern.ch` with DUCC
- Containers launched with CVMFS Docker Graph Driver plugin
  - Docker loads data on demand from CVMFS instead of pulling entire image
  - Significant startup acceleration and bandwidth savings for scalability

Mounting CVMFS securely

Pod Security Policy

```
---
kind: PodSecurityPolicy
apiVersion: policy/v1beta1
metadata:
  name: restricted-cvmfs
spec:
  privileged: false
  allowPrivilegeEscalation:
false
  volumes:
  - 'hostPath'
  - 'secret'
  allowedHostPaths:
  - pathPrefix: '/cvmfs'
      readOnly: true
```

Role

```
---
kind: Role
apiVersion:
rbac.authorization.k8s.io/v1
metadata:
  name: test-role
  namespace: test
rules:
- apiGroups: ['policy']
  resources: ['podsecuritypolicies']
  verbs:   ['use']
  resourceNames:
  - 'restricted-cvmfs'
```

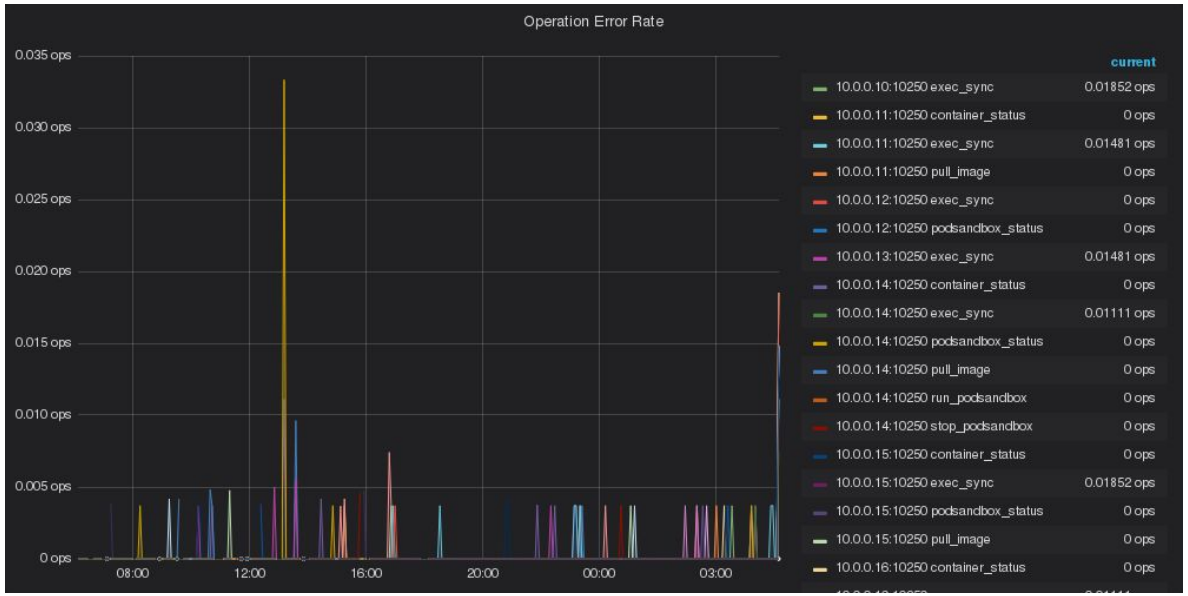
Pod

```
---
apiVersion: v1
kind: Pod
metadata:
  name: cvmfs-test
spec:
  volumes:
  - name: cvmfsvol
      hostPath:
      path: /cvmfs
      type: Directory
  containers:
  - name: fedora
      image: fedora
      command: [ "sh", "-c", "sleep 1h" ]
      volumeMounts:
      - name: cvmfsvol
      mountPath: /cvmfs
      readOnly: true
```

- Apply role binding, enable PSP admission controller
- Unprivileged container can mount only /cvmfs

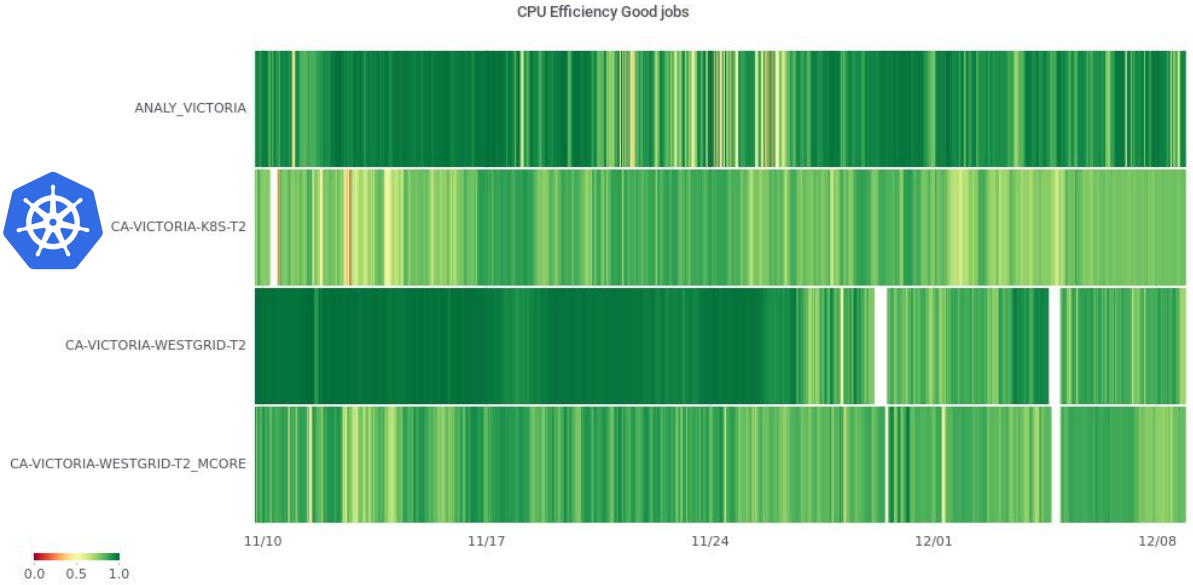
Monitoring & Operational experience

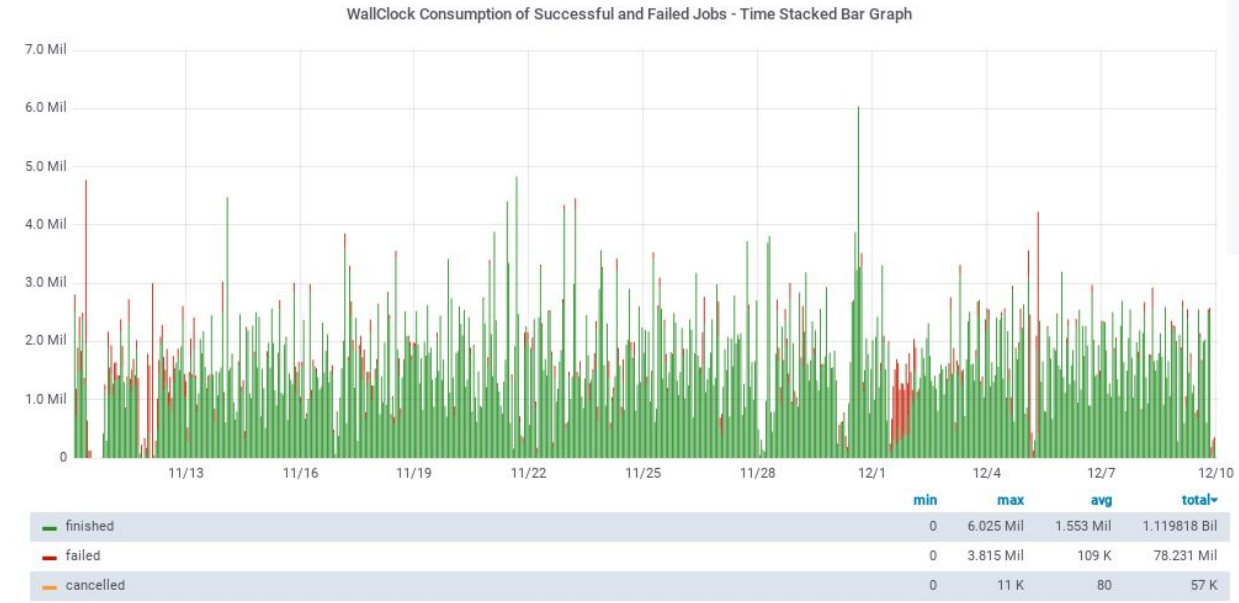
- Prometheus Helm operator provides real-time, self-configuring observability of containers and infrastructure
- Initial problem with CVMFS (autofs) configuration, repos appear unavailable

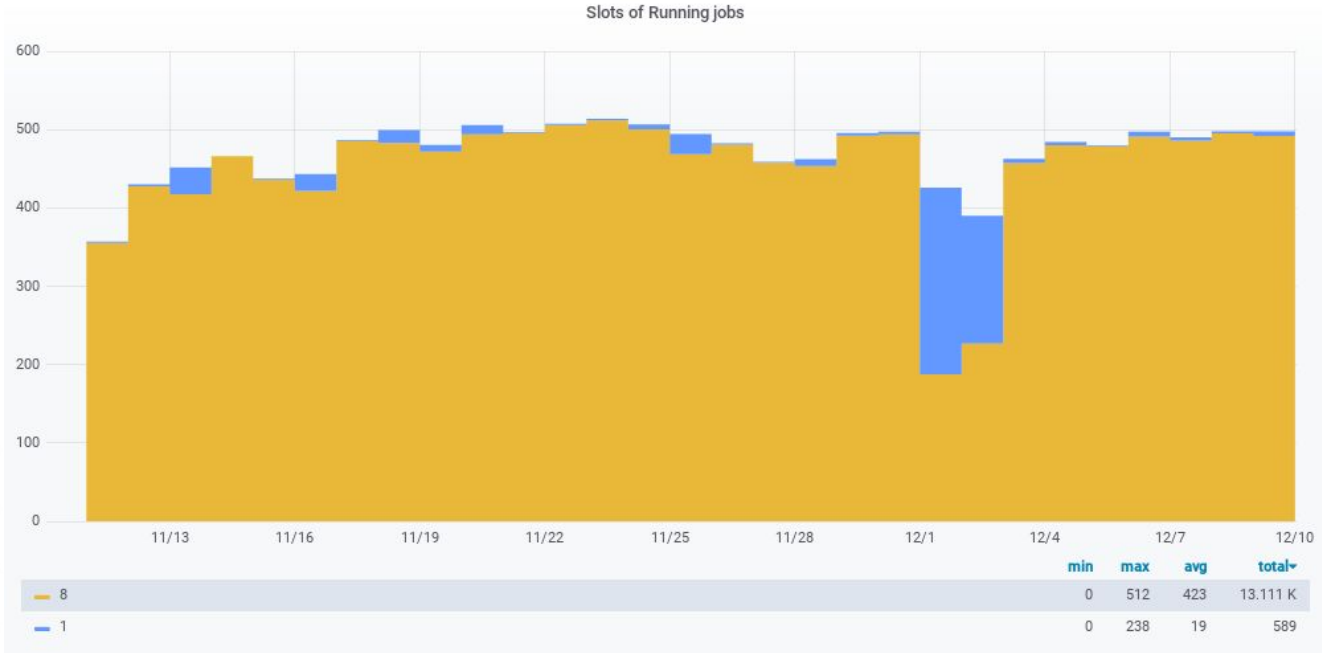


- Pods isolated from failure because they explicitly require CVMFS repo
- Problematic nodes automatically drained, impact was self-limiting, self-correcting

CA-VICTORIA-K8S-T2







- Oct 25: started running production jobs
- Nov 3: scaled up to 512 cores
- Cluster stable and untouched since CHEP

Next Steps

- Site
  - Cluster Autoscaler (and/or Virtual Kubelet?)
  - caching forward proxy deployment on k8s
  - bare metal with Openstack Ironic
  - switch to containerd, with CVMFS integration (containerd #3731)
  - publish images to `/cvmfs/images.computecanada.ca`
- Middleware
  - publish APEL accounting for k8s (?)
  - site A/R monitoring
- Workflow
  - Improve Harvester-k8s integration
  - Adjust pilot model to avoid lost heartbeats
- All: leverage k8s-native functionality

Discussion