Using Cluster API to provide managed Kubernetes on OpenStack

Matt Pryor, Senior Tech Lead, StackHPC OpenInfra Days Europe @ CERN, 6th June 2024







Kubernetes is 10 today!

StackHPC Company Overview



- Formed 2016, based in Bristol, UK
 - Based in Bristol with presence in Oxford, Cambridge, France and Poland
 - Currently around 30 people
- Founded on HPC expertise
 - Software Defined Networking
 - Systems Integration
 - OpenStack Development and Operations
- Motivation to transfer this expertise into Cloud to address HPC & HPDA (AI)
- "Open" Modus Operandi
 - Upstream development of OpenStack capability
 - Consultancy/Support to end-user organizations in managing HPC service transition
 - Scientific-WG engagement for the Open Infrastructure Foundation
- Hybrid Cloud Enablement

| StackHPC | |
|----------|---------------------|
| | Open Infrastructure |
| | SILVER MEMBER |

StackHPC Three Pillars









Reconfigurable and isolated infrastructure

Performance to extract maximum value

Self-service platforms





Science and Technology Facilities Council

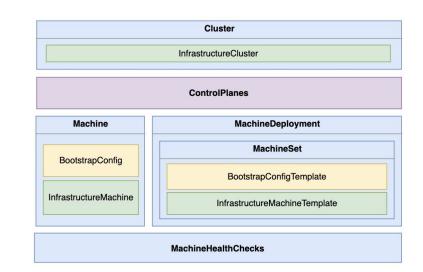


DiRA(;

What is Cluster API (CAPI)?



- Declarative API to manage Kubernetes clusters
- Multiple infrastructure providers
 - Provision machines, load balancers, networks
 - OpenStack, public cloud, bare metal
- Kubernetes deployed using kubeadm
- Provider-agnostic auto-healing, auto-scaling and rolling upgrade
- Management cluster runs CAPI controllers
- Workload clusters managed by CRUD operations on management cluster



Handling cluster addons

- Cluster needs addons to be fully functional
 - Networking (CNI)
 - Cloud integrations (CCM, CSI)
 - Ingress controller
 - Monitoring and logging
 - GPU and NIC drivers
- StackHPC addon provider
 - Open-source
 - Declarative interface for addons
 - HelmRelease and Manifests CRDs
 - Templating of properties from CAPI resources





CAPI Helm charts

- Cluster API clusters composed of several resources with references to each other
- Addons need to be wired up correctly
- Use Helm to template resources for a cluster
 - Knowledge encapsulated in charts
 - Charts are open-source
 - Simplified interface for consumers
 - Can be reused in multiple contexts
 - Tested on Kubernetes versions up to 1.30
- Cluster API now has ClusterClass
 - Extremely promising but still officially experimental
 - Will be integrated with CAPI Helm charts in future



kubernetesVersion: 1.29.4
machineImageId: <id>

machineSSHKeyName: <name>

clusterNetworking: externalNetworkId: <id>

controlPlane:
 machineFlavor: <name>

nodeGroups:

- name: md-0
machineFlavor: <name>
machineCount: 2

Magnum integration





What is Magnum?

- OpenStack project for managing container orchestration engines (COEs)
- Allows users to provision COEs in their OpenStack project to manage container-based workloads
 - Cluster templates define the available configurations
 - Clusters are where containers can be scheduled
- Designed to support multiple COEs using drivers
 - Existing drivers for Kubernetes, Swarm and Mesos
 - Swarm and Mesos drivers now deprecated
- REST server and conductor that communicate via AMQP







Existing Kubernetes driver issues



- Uses Heat to provision resources
 - Heat project is increasingly unloved
- Kubernetes deployed using non-standard tools
- Configuration using bash scripts
 - Brittle and difficult to debug
- Bespoke auto-healing and auto-scaling
- Upgrading a cluster almost never works
- Difficult to maintain
- Slow to support new Kubernetes versions



Magnum CAPI Helm driver



- Magnum architecture allows for multiple drivers
- Implement brand new driver that speaks Cluster API
 - Under OpenStack governance on OpenDev
 - Uses CAPI Helm charts to template resources
 - Benefit from upstream development and testing
 - Greatly reduce the amount of brittle, bespoke code
 - Not tied to Magnum release cycle (can evolve faster)
 - Deployed in production, e.g. Catalyst Cloud Kubernetes service
- Existing interfaces will continue to work
 - OpenStack CLI, Horizon plugin, Magnum Terraform plugin





User-friendly Kubernetes in Azimuth





What is Azimuth?

- Web portal for self-service platforms
- Configurable catalogue of curated platforms
 - StackHPC reference platforms
 - Apply site-specific optimisations
 - Automation using standard tools
- Platform services exposed using Zenith
 - Tunneling application proxy
 - No public IP required
 - SSO and TLS
- Manage platform users with Keycloak



Kubernetes



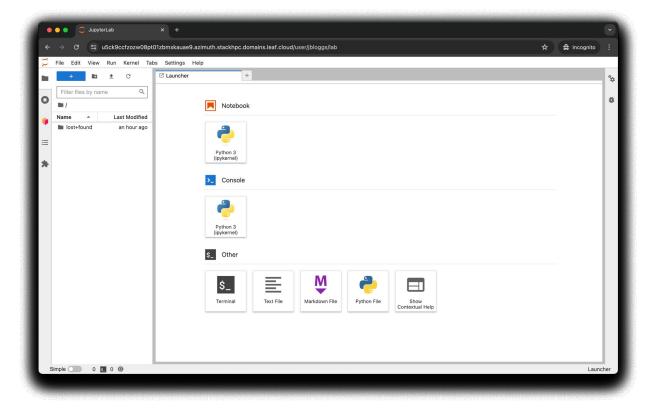
- Built on Cluster API and CAPI Helm charts
- Simple user interface
- HA control plane
- Multiple node groups
- Autoscaling, autohealing, rolling upgrades
- NVIDIA GPU and NIC support
- Kubernetes dashboard
- Monitoring and logging
- Secure access via Zenith

| | hpc-dev Le: 🗙 🧑 Loki / Pod Logs - Dashboa | ırds × + | | | | | | | * |
|---|--|---|---|---|---|--|---|---|---|
| → C 🖙 m20hcd1 | dfj58rr2s6ldcvvx7e7d9.azimuth.stackhpc.do | omains.leaf.cloud/d/209fd89b771c31 | 18dd442225414a50b59/ | loki-pod-logs?orgId=1&va | -namesp 🕁 | ð 📓 🖳 | Ď | M | : |
| | | Q Search or jump to | 🕮 cmd+k | | | + > | 0 | ٣ (| 4 |
| Home > Dashboards > | Loki / Pod Logs 🛛 🏠 | | | Share | Last 15 minut | es ~ Q | S | | |
| mespace kube-system ~ | pod etcd-mattp-k8s-control-plane-1 | tbxlw + etcd-mattp-k8s-control-pla | · search Enter | r variable value | | | | | |
| | | | | | | | | | |
| | | | | | | | | | |
| 3:06 18:07 1 | 8:08 18:09 18:10 18:11 | 11 18:12 18:13 | al 18:14 18:15 | 18:16 18:17 | 18:18 18 | 3:19 | 18:20 | | |
| | | | | | | | | | |
| gs Panel | | | | | | | | | |
| 2024-05-07 18:18:01.934 | {"level":"info","ts":"2024-05-07T17:18: | 01.934561Z","caller":"mvcc/hash. | go:137","msg":"storing | new hash","hash":242943 | 7070,"revision":33 | 51,"compact | -revisio | n":-1} | |
| 2024-05-07 18:18:01.934 | {"level":"info","ts":"2024-05-07T17:18: | 01.934382Z","caller":"mvcc/kvsto | re_compaction.go:68"," | msg":"finished scheduled | l compaction","comp | act-revisio | n":3351, | "too | |
| | k":"298.572711ms","hash":2429437070,"cu | rrent-db-size-bytes":36069376,"c | urrent-db-size":"36 MB | ","current-db-size-in-u | e-bytes":17883136, | "current-db | -size-in | | |
| | e":"18 MB"} | | | | | | | | |
| | {"level":"info","ts":"2024-05-07T17:18: | | | | | | | | |
| 2024-05-07 18:18:01.734 | {"level":"info","ts":"2024-05-07T17:18: | | | | | | | | |
| | k":"99.126727ms", "hash":2429437070, "cur | rent-db-size-bytes":35840000,"cu | rrent-db-size":"36 MB" | ,"current-db-size-in-use | e-bytes":17825792," | current-db- | size-in- | us | |
| | e":"18 MB"} | | | | | | | | |
| | {"level":"info","ts":"2024-05-07T17:18: | | | | | | | | |
| 2024-05-07 18:18:01.727 | {"level":"info","ts":"2024-05-07T17:18: | | | | | | | | |
| | k":"90.911022ms","hash":2429437070,"cur e":"18 MB"} | rent-db-size-bytes :35991552, cu | rrent-db-size : 36 MB | , current-db-size-in-use | e-bytes :17895424, | current-db- | size-in-l | JS | |
| 2024-05-07 10-10-01 624 | {"level":"info","ts":"2024-05-07T17:18: | -01 6245747" "coller":"mycc/index | | t trae index" "revision" | .22511 | | | | |
| | {"level":"info", "ts":"2024-05-07T17:18: | | | | | | | | |
| | {"level":"info","ts":"2024-05-07T17:18: | | | | | | | | |
| | {"level":"warn", "ts":"2024-05-07T17:14: | | ingoineit , mog i compuo | | | | | | |
| | | | rcentor go:197" "msg": | | | 7.14.33 573 | 4877" "t | ime so | |
| | | | | "request stats","start 1 | ime":"2024-05-07T1 | | | | |
| | ent":"485.187893ms","remote":"127.0.0.1 | 1:49688","response type":"/etcdse | rverpb.KV/Range", "requ | "request stats","start f est count":0,"request s: | ime":"2024-05-07T1 | | | | |
| 2024-05-07 18:14:34.058 | | 1:49688","response type":"/etcdse leases/openstack-system/external- | rverpb.KV/Range","requ resizer-cinder-csi-ope | "request stats","start 1 est count":0,"request s: nstack-org\" "} | ime":"2024-05-07T1 ze":77,"response c | ount":1,"re | sponse s: | ize":5 | |
| 2024-05-07 18:14:34.058 | ent":"485.187893ms","remote":"127.0.0.1 70,"request content":"key:\"/registry/1 | 1:49688","response type":"/etcdse leases/openstack-system/external- :34.05747Z","caller":"traceutil/t | rverpb.KV/Range","requ resizer-cinder-csi-ope race.go:171","msg":"tr | "request stats","start 1 est count":0,"request s: nstack-org\" "} ace[1127342216] range",' | ime":"2024-05-07T1 ze":77,"response c detail":"{range_be | ount":1,"re gin:/regist | sponse s: ry/lease: | ize":5 s/open | |
| 2024-05-07 18:14:34.058 | ent":"485.187893ms","remote":"127.0.0.1 70,"request content":"key:\"/registry/J {"level":"info","ts":"2024-05-07T17:14: | 1:49688","response type":"/etcdse leases/openstack-system/external- :34.057472","caller":"traceutil/t si-openstack-org; range_end:; res | rverpb.KV/Range","requ resizer-cinder-csi-ope race.go:171","msg":"tr ponse_count:1; respons | <pre>"request stats","start 1 est count":0,"request s: nstack-org\" "} ace[1127342216] range",' e_revision:4250; }","dua</pre> | :ime":"2024-05-07T1 .ze":77,"response c 'detail":"(range_be ration":"483.970228 | ount":1,"re gin:/regist ms","start" | sponse s: ry/lease: :"2024-0 | ize":5 s/open 5-07T1 | |
| 2024-05-07 18:14:34.058 | ent":"485.187893ms","remote":"127.0.0.1 70,"request content":"key:\"/registry/J {"level":"info","ts":"2024-05-07T17:14: stack-system/external-resizer-cinder-cs | 1:49688","response type":"/etcdse leases/openstack-system/external- :34.057472","caller":"traceutil/t si-openstack-org; range_end:; res | rverpb.KV/Range","requ resizer-cinder-csi-ope race.go:171","msg":"tr ponse_count:1; respons | <pre>"request stats","start 1 est count":0,"request s: nstack-org\" "} ace[1127342216] range",' e_revision:4250; }","dua</pre> | :ime":"2024-05-07T1 .ze":77,"response c 'detail":"(range_be ration":"483.970228 | ount":1,"re gin:/regist ms","start" | sponse s: ry/lease: :"2024-0 | ize":5 s/open 5-07T1 | |
| 2024-05-07 18:14:34.058 2024-05-07 18:14:34.058 | <pre>ent':'485.187893ms', 'remote':'127.0.0.1 70, 'request content':'key:\'/registry/3 ('level':'info','ts':'2824-68-69717:'14 stack-system/external-resizer-cinder-cs 7:14:33.5734922', 'mad':'2824-65-87117: s)'], 'step_count':1' ('level':'warn','ts':'2824-65-87117:14</pre> | 1:49688","response type":"/etcdse leases/openstack-system/external- 34.857472","caller":"traceutil/t si-openstack-org; range_end:; res 14:34.0574632","steps":["trace[11 :34.0584362","caller":"V3rpc/inte | rverpb.KV/Range", "requ resizer-cinder-csi-ope race.go:171", "msg":"tr ponse_count:1; respons 27342216] 'agreement a rceptor.go:197", "msg": | "request stats","start m est count":0, "request s: nstack-org\" "} e_revision:4250; }","dun mong raft nodes before ; "request stats","start m | :ime":"2024-05-07T1 .ze":77,"response c 'detail":"{range_be :ation":"483.970228 .inearized reading' :ime":"2024-05-07T1 | ount":1,"re gin:/regist ms","start" (duration 7:14:33.563 | sponse s: ry/lease: :"2024-0 : 483.80 0552","t: | ize":5 s/open 5-07T1 8557m ime sp | |
| 2024-05-07 18:14:34.058 2024-05-07 18:14:34.058 | <pre>ent':445.18793as', remote':127.0 8.1 70, request content':'key:\'registry/J ('level':'info','ts':'2924-65-0717):44 stack-system/external-resizer-cinder-cs 7):14:33.5734922','ned':'2024-65-07117):14 s)'','step_count':'1 ('level':'warn','ts':'2924-05-07117):14 ent':'495.37210s','remote':'127.0.0.1;</pre> | 1:49688","response type":"/etcdse leases/openstack-system/external- 34.857472", caller":"traceutil/t si-openstack-org; range_end:; res 14:34.8574632","steps":["traceutil :34.058462","caller":"Varpc/inte 46688","response type":/etcdser | rverpb.KV/Range", "requ resizer-cinder-csi-ope race.go:171", "msg":'tr ponse_count:1; respons 27342216] 'agreement a rceptor.go:197", "msg": verpb.KV/Range', "reque | "request stats","start t est count":0, "request s: nstack-org\" "} ace[1127342216] range",' e_revision:4250; }","dun mong raft nodes before ; "request stats","start f | :ime":"2024-05-07T1 .ze":77,"response c 'detail":"{range_be :ation":"483.970228 .inearized reading' :ime":"2024-05-07T1 | ount":1,"re gin:/regist ms","start" (duration 7:14:33.563 | sponse s: ry/lease: :"2024-0 : 483.80 0552","t: | ize":5 s/open 5-07T1 8557m ime sp | |
| 2024-05-07 18:14:34.058 2024-05-07 18:14:34.058 2024-05-07 18:14:34.058 | <pre>ent':'485.187893ms', 'remote':'127.0.0.1 70, 'request content':'key:\'/registry/3 ('level':'info','ts':'2824-68-69717:'14 stack-system/external-resizer-cinder-cs 7:14:33.5734922', 'mad':'2824-65-87117: s)'], 'step_count':1' ('level':'warn','ts':'2824-65-87117:14</pre> | 1:49688","response type":"/etcdse Leases/openstack-system/okternal- 34.057472", coller":"traceutil/t si-openstack-org; range_end:; res 14:34.0574632","steps':["trace[1 134.0534632","caller":"v3rpc/inte :49688","response type":"/etcdser eases/openstack-system/inder-cal | <pre>rverpb.KV/Range", "requ resizer-cinder-csi-ope race.go:171", "msg":"tr ponse_count:1; respons 27342216] 'agreement a rceptor.go:197", "msg": verpb.KV/Range", "reque -openstack-org\" "}</pre> | <pre>"request stats", "start t est count":0, "request s: nstack-org/* "} ace[1127342216] range", e_revision:4250; }","du mong raft nodes before 3 "request stats","start t st count":0,"request si;</pre> | <pre>ime":"2024-05-0711 ze":77,"response c 'detail":"(range_be ation":"483.970228 inearized reading' ime":"2024-05-07T1 te":60,"response co</pre> | ount":1,"re gin:/regist ms","start" (duration 7:14:33.563 unt":1,"res | sponse s: ry/lease: :"2024-0: : 483.80 055Z","t: ponse si; | ize":5 s/open 5-07T1 8557m ime sp ze":53 | |

DaskHub

StackHPC

- Runs on Kubernetes
- Apps managed using HelmRelease addon resources
- Each user gets their own notebook server
- Secure access via Zenith
- Grant access to external users using tenancy Keycloak realm
- Dask clusters for parallel computing using Dask Gateway



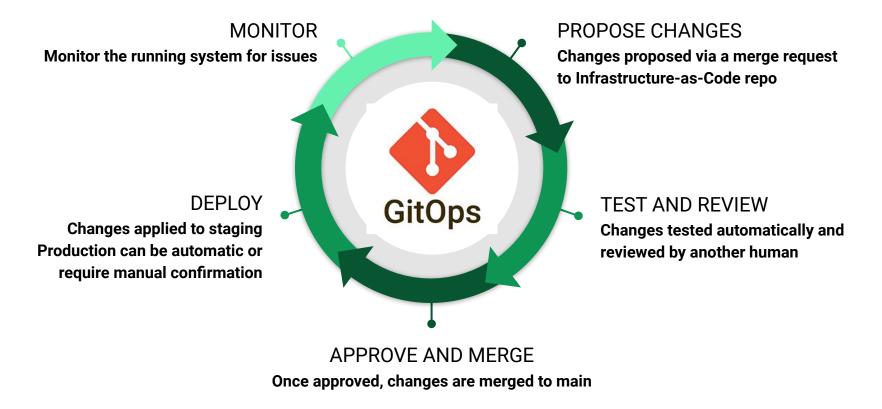
GitOps-managed Kubernetes





What is GitOps?





Self-managed Cluster API cluster



- Recap: Cluster API deploys clusters by creating resources in a management cluster
- But Cluster API clusters can manage themselves!
- Must be bootstrapped using a "pivot" process
 - Provision ephemeral Kubernetes cluster, e.g. kind
 - Install Cluster API controllers on ephemeral cluster
 - Provision a cluster using Cluster API resources on the ephemeral cluster (e.g. using CAPI Helm charts)
 - Install Cluster API controllers on Cluster API cluster
 - Pause reconciliation on ephemeral cluster
 - Move Cluster API resources to Cluster API cluster
 - Resume reconciliation on Cluster API cluster





GitOps-managed Kubernetes cluster



- Cluster API resources are just Kubernetes resources
- Bootstrap a self-managed Cluster API cluster
- Resources created using CAPI Helm charts
- Install Flux CD (or Argo) on cluster
 - Can also be self-managed
- Manage Helm release for cluster using Flux Helm resources or an Argo Application
- Changes to cluster configuration in git applied to cluster by Flux and reconciled by Cluster API



How to get started?





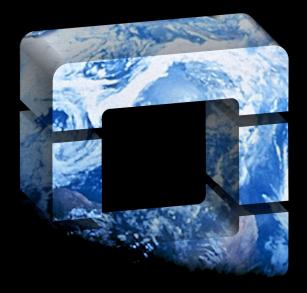
How to get started?

- Azimuth
 - Open-source (Apache 2.0)
 - Can be installed on any OpenStack cloud
 - <u>https://stackhpc.github.io/azimuth-config/try/</u>
- Magnum CAPI Helm driver
 - Open-source, under OpenStack governance
 - Install driver into Magnum Python environment
 - Needs a Cluster API management cluster to point to
 - Requires suitable images to be available
- StackHPC can help!









Thank You

https://www.stackhpc.com

StackHPC

The Rise of the HPC Cloud

23