OpenStack, RDMA and K8s

Bread, Oil and Vinegar?

John Garbutt, Principal Engineer
StackHPC Company Overview

- Formed 2016, based in Bristol, UK
  - Based in Bristol with presence in Oxford, Cambridge, France and Poland
  - Currently around 25 people
- Founded on HPC expertise
  - Software Defined Networking
  - Systems Integration
  - OpenStack Development and Operations
  - Growing Staff in AI/ML
- Motivation to transfer this expertise into Cloud to address HPC & HPDA (AI)
- “Open” Modus Operandi
  - Upstream development of OpenStack capability
  - Scientific-WG engagement for the Open Infrastructure Foundation
- Hybrid Cloud Enablement
Thank you!
Why **OpenStack** and **Kubernetes**?
Sharing Diverse Infrastructure

https://unsplash.com/photos/WcV2YkM3Di4
Reconfigurable Infrastructure (with Isolation)

https://unsplash.com/photos/4nHnP80Jtc
Many Apps built for Kubernetes

https://unsplash.com/photos/SIhqhxUA2_c
Why RDMA networking?
Why Remote Direct Memory Access?
OpenFOAM via kube-perftest

https://github.com/stackhpc/kube-perftest
OpenFOAM via kube-perftest

https://github.com/stackhpc/kube-perftest
KubeCon: Five ways with a CNI

https://github.com/stackhpc/kube-perftest
Repeatable on-demand: Azimuth using K8s Cluster API
Create Kubernetes using Cluster API
Kubernetes based Platforms
JupyterHub with Zenith Powered SSO
Kubernetes using Cluster API
Slurm with Open OnDemand
Bigger Laptop via Guacamole

Create a new platform

Platform name
Linux Workstation
Linux workstation (Ubuntu 20.04) accessible via a web browser

Workstation Size
Select a size...

Data volume size (GB)
10

The size of the data volume for the workstation.
The data volume will be available at /data

+ Create platform

Apache Guacamole

RECENT CONNECTIONS

desktop
shell

...revcbjjx0u.apps.hpc.cam.ac.uk wants to
See text and images copied to the clipboard

Allow
Block

ALL CONNECTIONS

johng-big-laptop
Linux Workstation

Web console
Monitoring

Details

Updated 8 days ago
How do you get **RDMA** in **LOKI**?
Three Steps to RDMA in LOKI

LOKI = Linux, OpenStack and Kubernetes Infrastructure

1. RDMA inside OpenStack servers
2. Kubernetes clusters on OpenStack
3. RDMA inside K8s pods
Step 1:
RDMA inside OpenStack servers
How to get RDMA?

- Baremetal servers (via Nova and Ironic)
  - ... it is just a physical server, all options are possible
  - But you don’t want untrusted users having root
- Virtual Machines need drivers for a real NIC
  - Dedicated NIC (PF passthrough)
- PCI passthrough a real NIC
  - Dedicated NIC(s) for SR-IOV
  - Multiple VFs on provider VLAN
- Legacy SR-IOV
  - Dedicate NIC(s) for SR-IOV
  - Multiple VFs on provider VLAN
- Mellanox VF-LAG
  - bond shared with hypervisor host
  - Full bond bandwidth of the bond in VM
  - OVS hardware offload flows
- Live migration isn’t ideal (ignoring vDPA and mdevs, for now)

PF = Physical Function
VF = Virtual Function
SR-IOV = Single Root I/O Virtualization
Step 2: Create K8s on OpenStack
Kubernetes on OpenStack

● K8s Cluster API
  ○ Describe Cluster in K8s CRDs
  ○ CAPI - OpenStack Provider
  ○ Bootstrap and Management Clusters
  ○ New OpenStack Magnum driver soon

● K8s Cloud Provider
  ○ CAPO - OpenStack provider
  ○ Cinder CSI
  ○ Octavia Load balancers

● Add ons
  ○ CNI, monitoring, GPU drivers …

https://github.com/stackhpc/capi-helm-charts
Step 3: RDMA inside k8s pods
RDMA Pods in OpenStack VMs

- OpenStack VM
  - Primary CNI network, using virtual NIC
  - Second SR-IOV VF-LAG
- Multus CNI
  - Pods opt into an additional network
- MACVLAN CNI
  - Additional MAC and IPs on VF-LAG NIC
  - Whereabouts IPAM address range
  - ... currently port security off
- Future ideas
  - Automatic addition of allowed address pairs

https://github.com/Mellanox/network-operator#macvlannetwork-crd
How to get involved?
Contributions are very welcome!

Please try Azimuth:
https://stackhpc.github.io/azimuth-config/try/
https://github.com/stackhpc/azimuth
https://github.com/stackhpc/zenith

K8s Cluster API Helm charts:
https://github.com/stackhpc/capi-helm-charts

kube-perftest:
https://github.com/stackhpc/kube-perftest

To read our blog or get in touch:
https://www.stackhpc.com/
Questions?

john.garbutt@stackhpc.com