Ceph In Compute Canada

Mike Cave
Senior Unix Systems Administrator
Research Computing Support
University of Victoria
Canada

What is Compute Canada?

What is Compute Canada?

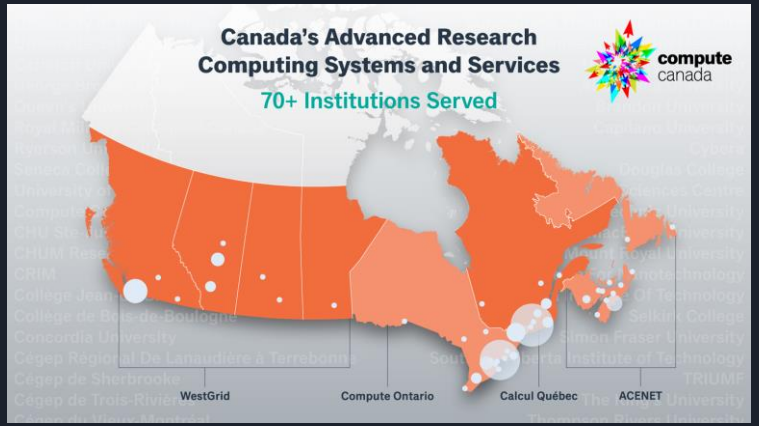
Canada's national computing research platform

Providing researchers:

- HPC
- Cloud
- Storage
- Backup

Combined funding model

Connected via 100Gb coast-to-coast network



Traditional HPC at Compute Canada

Four active HPC sites:

- Beluga - Montreal - ~35K cores
- Cedar - Vancouver - ~60K cores
- Graham - Waterloo - ~30K cores
- Niagara - Toronto - ~60K cores

Cloud for Compute Canada

Four cloud sites:

- Arbutus - Victoria - ~10,000 cores
- Cedar - Vancouver - ~500 cores
- East - Sherbrooke - ~600 cores
- Graham - Waterloo - ~900 cores

All sites are general purpose clouds

- Web portals
- Prototype

Smaller cloud installations attached to HPC

Ceph Deployments

All of the OpenStack installations use Ceph as the backend

Varying sizes:

- Arbutus: ~4.5 PB usable (triple replicated/erasure coded)
- East: 100 TB usable (triple replicated)
- Cedar: 700 TB useable (triple replicated)
- Graham: 100 TB usable (triple replicated)

Ceph at Arbutus

What is Arbutus?

First Steps
Expansion
New Cluster
Monitoring
The Near Future

Ceph at Arbutus: What is Arbutus?

Arbutus is the largest cloud installation for Compute Canada

On offer:

- ~10,000 physical cores
- 4.5PB of storage (usable)

Usage:

- Prototyping of experimental systems
- Web portals
- Atlas tier 2 and 3 job processing
- Training/workshops
- Virtual clusters

Ceph at Arbutus

What is Arbutus?

First Steps

Expansion

New Cluster

Monitoring

The Near Future

Ceph at Arbutus - First Steps

Ceph was initially deployed to support West Cloud phase 1

Started as 500GB of triple replicated storage available as RBD

Deployed CephFS to support persistent instances

Ceph at Arbutus - First Steps

Initial hardware offering:

- 13 OSD nodes
- 146 OSDs
- 500 TB raw capacity
- Journals co-located with OSDs
- 10Gb fibre network

Ceph at Arbutus - First Steps

Deployment:

- Largely completed manually from individual packages
- Some bash scripts for deploying configuration

Ceph at Arbutus

What is Arbutus?

First Steps

Expansion

New Cluster

Monitoring

The Near Future

Ceph at Arbutus - Expansion

Phase 2 for West Cloud was approved in 2016

The approval came with money to greatly expand our storage offering

New tools have been developed since phase one

Ceph at Arbutus - Expansion

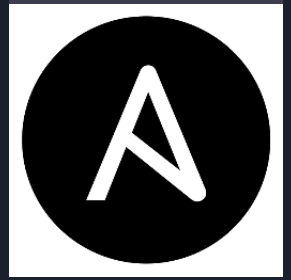
Expanded Cluster:

- 18 OSD nodes
- 260 OSDs
- SSDs for Journals
- 2.2 PB of raw capacity
- 10Gb fibre network

Ceph at Arbutus - Expansion

Deployment:

- Packages used for installation
- Some handcrafted bash scripts
- Introduction of Ansible with some handcrafted plays



Ceph at Arbutus

What is Arbutus?

First Steps

Expansion

New Cluster

Monitoring

The Near Future

Ceph at Arbutus - New Cluster

2018: New round of funding for hardware

Significant expansion opportunity

Enough to build a completely new ‘green field’ cluster

Ceph at Arbutus - New Cluster

The plan:

- Deploy a new control plane
- Deploy all the new hardware in a new cluster
- Migrate from West Cloud
- Spin down and migrate hardware to Arbutus

Ceph at Arbutus - New Cluster

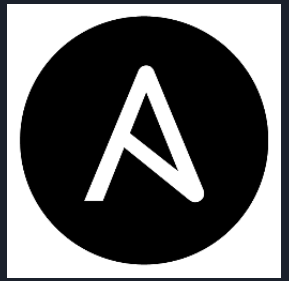
New Arbutus Ceph Cluster:

- 32 OSD Nodes
- 640 OSDs
- 5.3PB raw capacity
- SSDs for WAL/RocksDB
- Bonded Dual 10Gb Fibre Network

Ceph at Arbutus - New Cluster

Deployment:

- XCat for host discovery and imaging
- Ansible for host configuration
- Ceph-Ansible for ceph deployment



Ceph at Arbutus - New Cluster

Some notable changes and improvements:

- “Clean slate”
- Deploy Mimic with new install - West ran Jewel
- Changed to BlueStore
- Primary RBD pool employs erasure code profile (8-3)
- OpenStack images (glance) also using erasure coding
- Split client and replication traffic into separate vlans

Ceph at Arbutus

What is Arbutus?

First Steps

Expansion

New Cluster

Monitoring

The Near Future

Monitoring Ceph at Arbutus

Several layers for monitoring:

- In-house systems statistics (UVStats)
- Prometheus
- In-house alert management (FLARE)

Visualization with Grafana/UVStats

Monitoring Ceph at Arbutus

UVStats and Prometheus together watch all of the pieces of our system

UVStats stores system information for 7 years for every host in the environment

All messages forwarded to syslog-ng

FLARE watches our syslog streams and generates alerts based on a pattern matching engine built into syslog-ng

Visualization of Ceph at Arbutus

Grafana:

- Pulls from the Prometheus/InfluxDB datastream
- Stats back to the start of the new cluster
- Customized dashboards
  - Overview of cluster
  - Specific views of OSDs and Nodes
  - Rack level views

Ceph at Arbutus

What is Arbutus?

First Steps

Expansion

New Cluster

Monitoring

The Near Future

Ceph at Arbutus: The Near Future

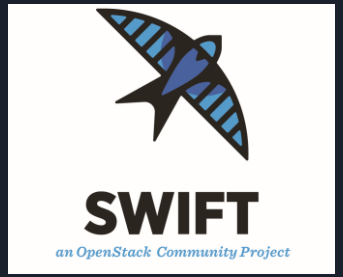
Fall 2019 through Spring 2020:

Addition of Object Storage

- Adding ~10PB of raw object storage with Ceph-RadosGW+ Swift+S3

Addition of Ceph FS

- Adding ~2PB of CephFS through Manila in OpenStack
- New cluster



Questions?

Contact:
mike.cave@computecanada.ca