CERN Cloud Service Update

Luis Pigueiras
On behalf of the CERN Cloud Team
CERN Cloud Recap

• Policy: servers in CERN IT shall be virtual

• Based on OpenStack
  • Production service since July 2013
  • Currently between Mitaka/Newton (depending on the component)
CERN Cloud Architecture

- Composed by two data centers
  - 1 region (1 API) and 50+ cells
  - Cells map different use cases
    - Shared: 5 AVZ
    - Project cells: special requirements (hardware, location)
    - Batch cells: optimized for batch workload (90% of HW)
CERN Cloud Architecture
CERN Cloud in Numbers (1)

- **7000** hypervisors in production + **~2000** to be added (5.8K 1yr ago)
- **220K** cores + **86K** to be added (155K)
- **530** TB of RAM (350 TB)
- **3.7K** volumes with **1.2 PB** allocated (Cinder) (2.8K)
- **3.8K** images/snapshots (Glance) (2.7K)
- **27** fileshares with **18 TB** allocated (Manila) (new)
- **71** container clusters (Magnum) (new)
CERN Cloud in Numbers (2)

- ~400 operations/min in Nova API
- Mainly creation/deletion of VMs
- ~27K VMs at the moment (+5k 6 months ago)
Operations and new features
Operations: Nova

- Upgrade from Liberty to Mitaka (done) and then from Mitaka to Newton (next week)
  - Required to have all nodes in CC7
  - Done in two steps due to DB migration constraints

- Steps:
  - Block APIs
  - Backup DBs
  - Upgrade Nova control plane
  - Upgrade DB schema
  - Validate
  - Enable API
  - Upgrade compute nodes (via Puppet and Yum)
Operations: Nova

**API requests (testing)**

<table>
<thead>
<tr>
<th>Time</th>
<th>08:00</th>
<th>10:00</th>
<th>12:00</th>
<th>14:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0</td>
<td>500</td>
<td>1300</td>
<td>1000</td>
</tr>
</tbody>
</table>

**API requests (users)**

<table>
<thead>
<tr>
<th>Time</th>
<th>08:00</th>
<th>10:00</th>
<th>12:00</th>
<th>14:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0</td>
<td>250</td>
<td>500</td>
<td>750</td>
</tr>
</tbody>
</table>

Testing

API enabled to all users
Operations: Nova

• Hypervisors upgrade from CC7.2 to CC7.3
• Upgrade done per cell
  • Limit potential impact
  • First batch cells, then project cells and finally shared cells
• Upgrade went fine, some minor issues:
  • Few VMs lost network connectivity
  • Kernel BUG soft lockup
    • Around ~200 AMD nodes affected (1 cell)
  • Various SELinux problems
Operations: Nova

- Consolidation of KVM under CentOS on the compute nodes (on-going)
  - Ceph as only backend for Volumes (NetApp decommission)
  - RHEL and HyperV decommission
  - Reduction of complexity
  - Remove dependency on Windows expertise
Operations: Neutron

- Networking service
- Started as a pilot in Q3 2015
- Available in 10 cells
- Deployed with a custom driver
  - Based on the linux bridge driver
  - Integration with CERN’s network infrastructure
- Migration from deprecated nova-network to Neutron networks (during this year)
- Neutron server upgraded from Mitaka to Newton
  - Neutron server moves but agents stay in Mitaka as they follow the Nova upgrade cycle
What’s new? Keystone

- Fernet tokens deployment
  - Much better performance and less load on the server
  - No need to store tokens in a database
- Additional project attributes
  - Accounting group and project type for accounting purposes
  - Cell mapping of the project (being prepared)
What’s new? Accounting

- Collecting cloud accounting data via cASO
- Storing data into S3
- Processing data with Spark
- Accounting data visualization through webpage
What’s new? Rally

- Cloud benchmarking system
  - Used to check status of the cloud since 2015
  - Cherry-pick to run as a non-admin user in the cloud
- Current release: Newton
- Stress-test the orchestration service and all the other services
- Dashboards with cell or service status
What’s new? Rally

<table>
<thead>
<tr>
<th>Time</th>
<th>gva_shared_016</th>
<th>gva_shared_017</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 18, 2017 2:00 PM</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>April 18, 2017 1:00 PM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>April 18, 2017 12:00 PM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>April 18, 2017 11:00 AM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>April 18, 2017 10:00 AM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>April 18, 2017 9:00 AM</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>April 18, 2017 8:00 AM</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
What’s new? Magnum

- OpenStack project to treat orchestration engines as 1st class resources
  - Docker Swarm, Kubernetes, Mesos and DC/OS
- Current release is in Newton (with cherry-picks from master)
- Timeline
What’s new? Magnum

• Interaction with Magnum client to create the cluster
• Docker client to interact with the created client (or kubectl, REST API...)

```
$ magnum cluster-create --name myswarmcluster --cluster-template swarm --node-count 100
$ magnum cluster-list
+-----------------+-----------------+-----------------+-----------------+-----------------+
| uuid           | name            | node_count      | master_count    | status          |
+-----------------+-----------------+-----------------+-----------------+-----------------+
| ....           | myswarmcluster  | 100             | 1               | CREATE_COMPLETE |
+-----------------+-----------------+-----------------+-----------------+-----------------+
$ $(magnum cluster-config myswarmcluster --dir magnum/myswarmcluster)

$ docker info / ps / ...
$ docker run --volume-driver cvmfs -v atlas.cern.ch:/cvmfs/atlas -it centos /bin/bash
[root@32f4cf39128d /]#```
What’s new? Magnum

- Several use cases at CERN
  - Gitlab CI
  - ROOTaaS / Swan (Jupyter Notebooks)
  - FTS
  - Batch
  - ATLAS RECAST (Analysis Catalog) – Next slide
  - ... and many others
What’s new? Magnum

- ATLAS RECAST use case for Magnum
  - Software distribution in HEP
  - Development environment parity with distributed/batch systems
  - Reproducible, interactive development environment for personal development/software tutorials
  - Continuous integration and release testing
  - Analysis preservation and reusability
Magnum feature plans

• Rolling upgrades of clusters
  • Upgrade to new versions of Kubernetes, Swarm, ...
• Heterogeneous clusters
  • Mix of VMs and baremetal, spread across AVZs
• Container Monitoring (cAdvisor, Prometheus and Grafana)
  • Deployed for K8S and WIP for Swarm
• LBaaS for HA
• Better storage support (EOS, CVMFS, ...)
Upcoming services

• Baremetal service Ironic
  • API server, conductor and first node deployed last month

• Workflow service Mistral
  • Will simplify operations
  • Already deployed with testing workflow prototypes
  • Will play along with Rundeck for workflows
Upcoming services

- Fileshare service Manila
  - Pilot since Q4 2016
  - CephFS as the backend
  - Off-the-shelf integration with Kubernetes, Swarm...
  - Need for a highly available FS (to replace NFS filer service)
    - Collaboration with FILER service
    - Share configuration, certificates, etc
Summary

• Cloud service continues to grow
  • Adding 2000 compute nodes with ~86K cores
  • New components added to our cloud like with several use cases like Magnum and Manila
  • Expansion planned (bare metal provisioning)
• Confronting some major operational challenges
  • Scaling all of our services with the new HW additions
  • Replacement of network layer

http://openstack-in-production.blogspot.com