Running an OpenStack Cloud for several years and living to tell the tale

Alexandre Maumené

Gaëtan Trellu

Tokyo Summit, November 2015

About the speakers

Alexandre Maumené

- OpenStacker since 2012, Red-Hatter *(CIP)* based in Paris

Gaëtan Trellu

- OpenStacker since 2013, Red-Hatter *(CIP)* based in Montréal

Technical Cloud Consultants

About this talk

Deployment software

- RHEL OSP-director, TripleO, Puppet modules

Standard HA architecture

- Networks, software, OpenStack components

Logging, monitoring, backup

Tips 'n tricks

redhat.

Deployment process

Undercloud

- RHEL server *(physical or virtual)*
- TripleO deployed via rdomanager-oscplugin *(Kilo)*

Overcloud

- Handled by the undercloud *(Ironic, TripleO, eDeploy)*
- Customized with TripleO templates
- Deployed as a Heat stack

redhat.

Undercloud

Ironic

- Provisioning
  - JSON file containing IPMI information, MAC address…
- Introspection
  - Boot on discovery image
  - Upload hardware profile to Swift
  - Update database
- Profile Matching
  - Assign a Nova flavor to an Ironic node

redhat.

Undercloud

Heat

- Customize TripleO templates
  - Network isolation and configuration
  - Storage configuration
  - Pre/Post deploy actions
- Deploy Overcloud
  - Image stored in Glance
  - Assign a Nova instance to an Ironic node
  - Configure Overcloud with Puppet

redhat.

Puppet

- Puppet upstream modules
- Puppet masterless

```
# puppet apply /var/lib/heat-config/xxx.pp
```

- Puppet modules already present on Glance image
- Define variables in Hiera
- Override them in TripleO templates *(Heat)*
- Orchestrated by steps

redhat.

Our ref-arch *(software)*

Controller *(3+2n)*:

- Databases: Galera, MongoDB & Redis
- HA: HAProxy & Pacemaker
- Messaging: RabbitMQ
- OpenStack services: Keystone, Cinder, Glance, Heat, Ceilometer, Horizon, Neutron, Swift Proxy, Ceph MON

Compute *(x)*

Storage *(x Ceph OSD and/or Swift)*

redhat.

Our ref-arch *(network)*

Network isolation:

- Provisioning/Management
- Internal API
- Tenant Networks *(VxLAN by default)*
- Storage
- Storage Management *(Replication)*
- External *(API and Floating IP)*

redhat.

Logging

Logging

- Centralized logging done by:
  - Fluentd
  - Kibana
  - Elasticsearch

redhat.

Monitoring

Monitoring

- Sensu with Uchiwa dashboard
- System checks *(CPU, RAID, RAM, Swap, Disk, NTP, ping)*
- OpenStack checks
  - AMQP and API for all services
  - Ceph disk usage and health
  - Galera replication and MongoDB
  - Swift dispersion, object, upload and ring usage
  - Upload image, launch instance with volume from this image, associate floating IP and test network connectivity

redhat.

Status page “à la Amazon”

- Use whiskerboard
- Sensu handler to update the dashboard

redhat.

Canada - Montréal

Status for eNoCloud Montréal

Service	Current	May 11	May 10	May 09	May 08	May 07
OpenStack Compute - API	♥	♥	♥	♥	♥	♥
OpenStack Compute - Instance	♥	♥	♥	♥	♥	♥
OpenStack Dashboard - HTTP	♥	♥	♥	♥	♥	♥
OpenStack Identity - API	♥	♥	♥	♥	♥	♥
OpenStack Image Service - API	♥	♥	♥	♥	♥	♥
OpenStack Image Service - Image	♥	♥	♥	♥	♥	♥
OpenStack Networking - API	♥	♥	♥	♥	♥	♥
OpenStack Networking - IP	♥	♥	♥	♥	♥	♥
OpenStack Telemetry - API	♥	♥	♥	ℹ	ℹ	♥
Openstack Block Storage - Volume	♥	ℹ	♥	♥	♥	✖

Legend

♥	The service is up and running
ℹ	The service had some issues
✖	The service is currently down

Backup

- Run from an external server
- Only Ceph volumes
- `XFS`, `ext3` and `ext4` supported

```
~# cinder metadata vol001 set stackup=True
```

- SSH key imported on the backup server
- Full and incremental backups

redhat.

Tips ’n tricks

Tips ’n tricks

Synchronize time with NTP servers

- Galera → Replication between nodes
- RabbitMQ → Synchronization *(tokens)*
- Ceph → Synchronization *(mon)*

redhat.

Tips ’n tricks

Network

- MTU *(9000 if the hardware supports jumbo frame)*
- Disable TSO, GSO in `qr`/`qbr`/`qvo`/`qvb` interfaces
- Disable `rp_filter` *(if needed)*
- Disable GRO, GSO, LRO for physical interfaces when using VxLAN *(depends on kernel version)*

redhat.

Tips ’n tricks

HAProxy

- Increase `maxconn`
- Increase Galera timeout
- Increase RabbitMQ timeout
- `/etc/haproxy/haproxy.cfg`

redhat.

Tips ’n tricks

RabbitMQ

- Limits

```
~# rabbitmqctl status | grep file_descriptors -A4
  {file_descriptors,
   [{total_limit,3996},
    {total_used,228},
    {sockets_limit,3594},
    {sockets_used,226}]},
```

redhat.

Tips ’n tricks

RabbitMQ

- Set `rabbit_durable_queue` in OpenStack components
  - Queues survive when RabbitMQ crashed
- Set `rabbit_ha_queue` in OpenStack components
  - Set "`x-ha-policy: all`" flag on queues related to the components
- Set an expire policy to avoid amount of orphans queue
  - `'{"expires":3600000, "ha-mode":"all", "ha-sync-mode":"automatic"}'`

Fish ’n chips

MySQL

- `open_files_limit = 131070` in `/etc/mysql/my.cnf`
- `LimitNOFILE=131070` in `/etc/systemd/system/mariadb.service.d/`
- `max_connections = 4096` in `/etc/mysql/my.cnf`
- Monitor the number of active connections

redhat.

Tips ’n tricks

Keystone

`~# keystone-manage token_flush` before Juno, was deleting expired tokens one by one *(and could get stuck)*

```
~# pt-archiver --source h=host,u=user,p=pass,D=db,t=table
--charset utf8 --where "expires < UTC_TIMESTAMP()" --purge
--txn-size 500 --statistics --primary-key-only
```

redhat.

Tips ’n tricks

MongoDB & Ceilometer

- `replicaSet` will not be sufficient after few weeks.
  - Use sharding from the beginning

- Distribute on `counter_name` for example:

```
~# sh.shardCollection("ceilometer.meter", {'counter_name': 1})
```

Tips ’n tricks

Ceph

- Avoid killing your cluster when a node crashes

```
~# ceph tell osd.* injectargs '--osd-max-backfills 1'
~# ceph tell osd.* injectargs '--osd-recovery-threads 1'
~# ceph tell osd.* injectargs '--osd-recovery-op-priority 1'
~# ceph tell osd.* injectargs '--osd-client-op-priority 63'
~# ceph tell osd.* injectargs '--osd-recovery-max-active 1'
~# ceph tell osd.* injectargs '--osd-scrub-load-threshold 1'
```

- Watch out: be sure to use the BIOS performance hardware profile!

redhat.

Tips ’n tricks

Glance

- Expose image direct URL in Glance *(COW)*
  - `show_image_direct_url=True` in glance-api.conf
  - Can be a security risk

redhat.

Links & Questions

- https://www.rdoproject.org/
- https://access.redhat.com/documentation/en/red-hat-enterprise-linux-openstack-platform/7/
- https://github.com/awheeler/whiskerboard
- https://github.com/AlexandreNo/sensu_to_whiskerboard
- https://github.com/openstack/monitoring-for-openstack
- https://github.com/goldyfruit/stackup

- Alexandre Maumené - amaumene@redhat.com
- Gaëtan Trellu - gtrellu@redhat.com

redhat.