Automated Deployment of an HA OpenStack Cloud with SUSE® Cloud

Adam Spiers
Senior Software Engineer
aspiers@suse.com

Vincent Untz
Project Manager
vuntz@suse.com
Agenda

Start **building** a cloud!

Quick intro to **SUSE® Cloud architecture**

Learn about **HA in OpenStack and SUSE Cloud**

Build an **HA cluster**

Build an **HA OpenStack cloud** on the cluster

**Break things!**
Workshop environment
Workshop environment

Relax ;-)  
- We have plenty of time  
- Whole build is also automated and idempotent  
- You can take home the entire environment afterwards (available online)  
- You can run on any machine with at least 16GB RAM  
  - ... or 8GB at a push  
  - (although that comes with limitations)
Workshop environment

• We'll build a miniature cloud on a single machine
• VirtualBox hypervisor
• 4 VMs
  - Administration Server (Crowbar)
  - 2 Control Nodes in an HA cluster
  - 1 Compute Node
• Vagrant for rapid deployment
What is Vagrant?

"Creates and configures lightweight, reproducible, and portable development environments."

https://www.vagrantup.com/

- Not just for development
- Perfect for "kicking the tyres", demoing, testing etc.
- Cross-platform (Linux, MacOS X, Windows)
- Providers for libvirt, VirtualBox, VMware, Hyper-V, Docker, OpenStack, ...
Vagrant inputs

• 1 or more Vagrant "box" – pre-built virtual appliances

• Vagrantfile: Ruby DSL file which defines:
  - which box(es) to use
  - virtual hardware required
  - virtual network topology
  - network ports to forward
  - hypervisor-specific settings
  - files to inject into appliance
  - commands to run in appliance

• files to inject
Using Vagrant: crash course

- `vagrant box add suse/cloud4-admin`
  - [https://vagrantcloud.com/suse/](https://vagrantcloud.com/suse/)
  - Also possible to add local boxes

- `vagrant up admin`

- `vagrant up controller1`

- `vagrant halt controller2`

- `vagrant destroy compute1`

- [https://docs.vagrantup.com/v2/getting-started/index.html](https://docs.vagrantup.com/v2/getting-started/index.html)
Workshop Vagrant environment

- https://github.com/SUSE-Cloud/suse-cloud-vagrant
  - demos/HA/
  - vagrant/
    - Vagrantfile and configs/2-controllers-1-compute.yaml
- VirtualBox pre-installed
- 2 boxes pre-installed
  - suse/cloud4-admin and suse/sles11sp3
- 4 VMs
  - admin: SUSE Cloud 4 Administration Server
  - controller1, controller2 (will form an HA cluster)
  - compute1
Exercise #1: start the build!

- Start up VirtualBox GUI
- `cd` to local copy of git repository
- `cd vagrant/`
- `vagrant up`
- All 4 VMs will be booted in sequence:
  - admin
  - controller1
  - controller2
  - compute1
SUSE® Cloud Overview
SUSE® Cloud

Enterprise OpenStack distribution that rapidly deploys and easily manages highly available, mixed hypervisor IaaS Clouds

- Increase business agility
- Economically scale IT capabilities
- Easily deliver future innovations
Why an Install Framework?

1229 Parameters

11 Components

1 Hour
Why an Install Framework?

SCARY AS HELL!
Introduction to Crowbar
Crowbar
What is Crowbar?

- Crowbar is a platform for planning and deployment from bare metal.
- It provides server discovery, upgrades, and operating system installation using PXE Boot.
- It deploys applications on servers and operating systems using Crowbar.
- It has a scary logo.
- It could have been worse.
SUSE® Cloud Administration Server
SUSE® Cloud Control Node

- PostgreSQL database
- Image Service (Glance) for managing virtual images
- Identity (Keystone), providing authentication and authorization for all SUSE Cloud services
- Dashboard (Horizon), providing the Dashboard, which is a user Web interface for the SUSE Cloud services
- Nova API and scheduler
- Message broker (RabbitMQ)
SUSE® Cloud Compute Nodes

• Pool of machines where instances run
• Equipped with RAM and CPU
• SUSE Cloud Compute (nova) service
  - Setting up, starting, stopping, migration of VMs
sledgehammer
barclamp
Interlude
Exercise #2: assign aliases to nodes

- Connect to admin node
  - vagrant ssh admin or
  - ssh root@192.168.124.10 or
  - use VM console in VirtualBox
- Root password is vagrant
- Type q then y to accept the beta EULA
- Run setup-node-aliases.sh
- Point a browser at the Crowbar web UI
  - http://localhost:3000
- Check the 4 nodes are registered, named correctly, and in Ready state (green)
High Availability and Cloud
Why High Availability?

“I can't have my systems go down. We lose $1,000,000 for every minute that we're down, and upper-management gets really 'excited' when that happens.”
High Availability for OpenStack

What might we want to protect?

• Admin server
  - core infrastructure: DNS, NTP, provisioning capabilities

• Controller node
  - OpenStack services

• Compute nodes
  - Hypervisor
  - VM instances (i.e. guests in the cloud)
Component failure impact

• Admin server
  - New cloud nodes require manual addition and configuration
  - Currently no ability to rediscover existing nodes on restart
  - No impact on currently operating cloud

• Control node
  - Cannot start or stop guest instances
  - No ability to rediscover existing nodes or guest VMs on restart
  - No impact on currently deployed instances
Pets vs. cattle metaphor

- Pets are given names like mittens.mycompany.com
- Each one is unique, lovingly hand-raised and cared for
- When they get ill, you spend money nursing them back to health

- Cattle are given names like vm0213.cloud.mycompany.com
- They are almost identical to other cattle
- When one gets ill, you shoot it and get another one
Component failure impact (continued)

• Compute node
  - Loss of VMs on that node
  - Recovery is by restart and re-provisioning of physical server
  - Can be mitigated through application design

• VM instances
  - Loss of workload
  - Recovery is by booting a replacement instance (cattle)
  - Can be mitigated through application design
Component failure assessment

• Control Node
  – Highest priority
  – Recovery realistically requires complete cloud restart

• Compute Node & VM instances
  – Application level recovery is normal practice for existing clouds
  – Not existing “enterprise” expectation, but workaround exists for new workloads

• Admin Server
  – Least impact on deployed system
  – Operation can continue with no impact on end users
Status Quo of HA in OpenStack

• Community is now mostly converged on a standard architecture for an HA control plane involving Pacemaker and HAproxy
  – SUSE® was first vendor to release a supported implementation of this, via an update to SUSE Cloud 3 (May 2014)

• No one has yet implemented a full solution for HA of compute nodes and VM guests
  – However community discussion in the last month has generated proposals which look quite promising.

• HA for storage and network nodes is also still ongoing work
High Availability in SUSE Cloud
HA in SUSE® Cloud (high level)

- **Administration Server**
  - No longer a SPoF (Single Point of Failure)
  - Can have multiple DNS / NTP servers
  - Backup / restore script for cold or warm standby

- **Control Plane**
  - Run services in a cluster
    - to ensure availability of data and service
  - Some OpenStack services are stateless
  - Some can run active/active, e.g. API endpoint services
  - The load balancer still needs protecting
  - Database and message queue need shared storage
HA Control Plane in SUSE® Cloud
HA Control Plane in SUSE® Cloud
HA Control Plane in SUSE® Cloud

- Fully automated cluster setup through Pacemaker barclamp
- Simple, intuitive web UI
- Allows choice of cluster size and quantity
- Supports multiple strategies for STONITH and storage
- Uses SLE HAE components
  - Pacemaker, HAproxy, DRBD
- Architecture consistent with OpenStack community recommendations
HA Control Plane in SUSE® Cloud

- Active/passive for PostgreSQL and RabbitMQ
  - choice of replicated (DRBD) or shared storage
- Active/active for other services via HAproxy load balancer
  - HAproxy itself is active/passive
  - Innovative approach to Neutron L3 agent
Setting Expectations

- *Not* fault tolerance
- Small outage of services is tolerated
- Automated recovery within small number of minutes
- “4 nines” availability (99.99% = ~53 mins/year)
- Maybe even five 9s achievable (4.32 mins/year) depending on context
- Some manual intervention may be necessary to repair a degraded (but still functioning) cluster
Simple cluster architecture

Control Node 1
Orchestration
Telemetry
Dashboard
Nova
Glance
Neutron
Cinder
Keystone
RabbitMQ
PostgreSQL

Control Node 2

Pacemaker Cluster
DRBD
Recommended architecture

Services Cluster

- Node 1
  - Orchestration
  - Telemetry
  - Dashboard
  - Nova
  - Glance
  - Cinder
  - Keystone

- Node 2

- Node 3

Network Cluster

- Node 1
  - Neutron

- Node 2

- Node 3

Database Cluster

- Node 1
  - RabbitMQ
  - PostgreSQL

- Node 2
  - DRBD or shared storage
Building a Pacemaker Cluster
Cluster fencing

- Every HA cluster needs an out-of-band **fencing** mechanism. This is not optional!

- Simplistically, if cluster communications break down, consensus is lost, and multiple nodes may contend for the same data/service ("split brain" syndrome)

- Solution: **STONITH** (Shoot The Other Node In The Head)

- Popular fencing devices include IPMI, IBM RSA, HP iLO, Dell DRAC

- We'll use **SBD** (Storage-Based Death) which allows sending "poison pill" messages via shared block storage device.
Exercise #3: build a Pacemaker cluster

• In VirtualBox GUI, view settings for controller1 and controller2
  - Observe locations of extra disks (SBD and DRBD)
  - Which disk is shared?

• Point a browser at the Crowbar web UI
  - http://localhost:3000

• Follow instructions for deploying a Pacemaker cluster
  - SBD device is /dev/sdc

• On admin node:
  - Run tail -f /var/log/crowbar/chef_client/*
Exercise #4: check cluster health

- Wait for chef-client to complete on both controller nodes
- Pacemaker proposal should finish applying and go green
- Connect to controller1 node
  - vagrant ssh controller or
  - connect to admin node and ssh controller1, or
  - use VM console in VirtualBox
- Root password (vagrant) and EULA as before
- Run crm_mon and check cluster has two nodes online
- Visit Hawk web UI on https://localhost:7630
Exercise #5: build HA cloud

- Follow remaining instructions for deploying the remaining barclamps
- This will take quite a long time (at least 30 minutes)
- If you are feeling lazy, you can use a tool to do this automatically:
  - `crowbar batch --timeout 1200 \ build HA-cloud.yaml`
- Watch Hawk web UI and `crm_mon` output as Crowbar/Chef automatically add resources to the cluster
Exercise #6: simulate failures

- Follow instructions for testing cluster failover
Exercise #7: recover degraded cluster

- Follow instructions for recovering a degraded cluster
If you made it this far, well done!
Unpublished Work of SUSE LLC. All Rights Reserved.
This work is an unpublished work and contains confidential, proprietary and trade secret information of SUSE LLC. Access to this work is restricted to SUSE employees who have a need to know to perform tasks within the scope of their assignments. No part of this work may be practiced, performed, copied, distributed, revised, modified, translated, abridged, condensed, expanded, collected, or adapted without the prior written consent of SUSE. Any use or exploitation of this work without authorization could subject the perpetrator to criminal and civil liability.

General Disclaimer
This document is not to be construed as a promise by any participating company to develop, deliver, or market a product. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. SUSE makes no representations or warranties with respect to the contents of this document, and specifically disclaims any express or implied warranties of merchantability or fitness for any particular purpose. The development, release, and timing of features or functionality described for SUSE products remains at the sole discretion of SUSE. Further, SUSE reserves the right to revise this document and to make changes to its content, at any time, without obligation to notify any person or entity of such revisions or changes. All SUSE marks referenced in this presentation are trademarks or registered trademarks of Novell, Inc. in the United States and other countries. All third-party trademarks are the property of their respective owners.