SUSE OpenStack Cloud HA

February 2019

Arthur Yang
Sales Engineer

HA概述

标准

最小数据恢复时间　　最小宕机时间

稳定性	宕机时间/年	宕机时间/月	宕机时间/周	宕机时间/天
99%（两个9）	3.65天	7.2小时	1.68小时	14.4分钟
99.9%	8.67小时	43.8分钟	10.1分钟	1.44分钟
99.99%	52.56分	4.38分钟	1.01分钟	8.66秒
99.999%	5.26分	25.9秒	6.05秒	864.3毫秒
99.9999%(六个9)	31.5秒	2.59秒	604.8毫秒	86.4毫秒

避免单点故障

网络设备的冗余

APP 冗余以及服务自主迁移

电源模块的冗余

共享存储

Stateful与Stateless

状态	描述	示例
有状态(stateful)	服务之间没有依赖 无需数据同步	OpenStack-APIS,Scheduler
无状态(stateless)	一个操作设计多个请求 在冗余节点间，数据需要同步与恢复	MySQL, RabbitMQ

其他概念

名称	描述
failover	服务从master节点迁移到slave节点
failback	服务从master节点迁移回至master节点
fencing	将出错节点进行关机隔离
Active/Passive	主从模式 只有一个master
Active/Active	双活模式 有多个master

OpenStack HA

Pacemaker

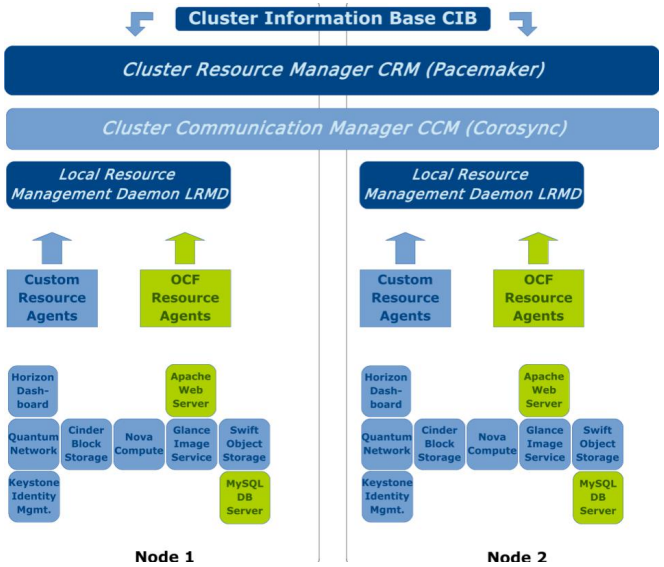
集群资源管理工具

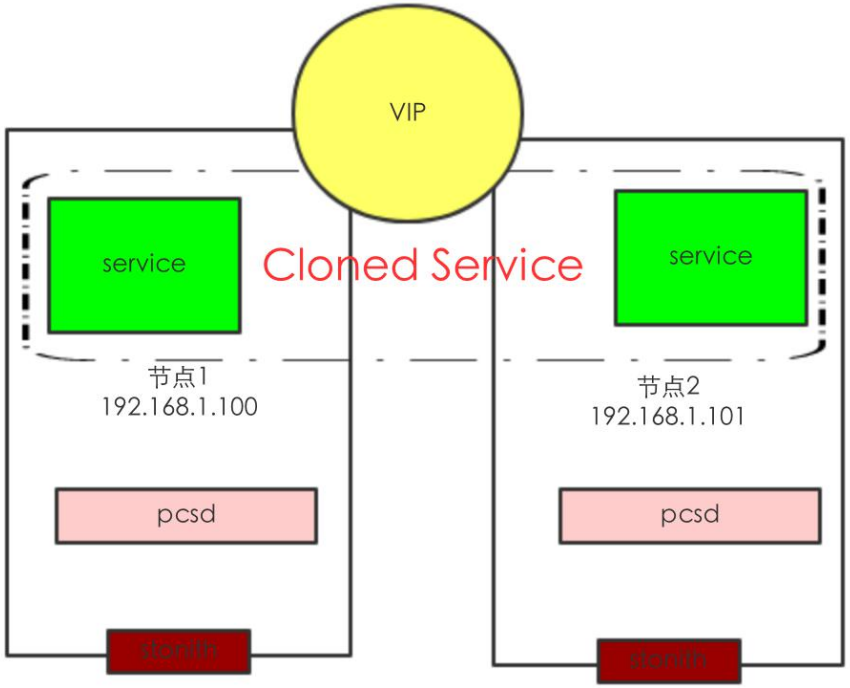
使用corosync作为集群通信的工具

浮动的VIP

Cloned Service (active/active)

Stonith-通过电源管理实现fencing





Stonith-一般为IPMI或者共享裸盘

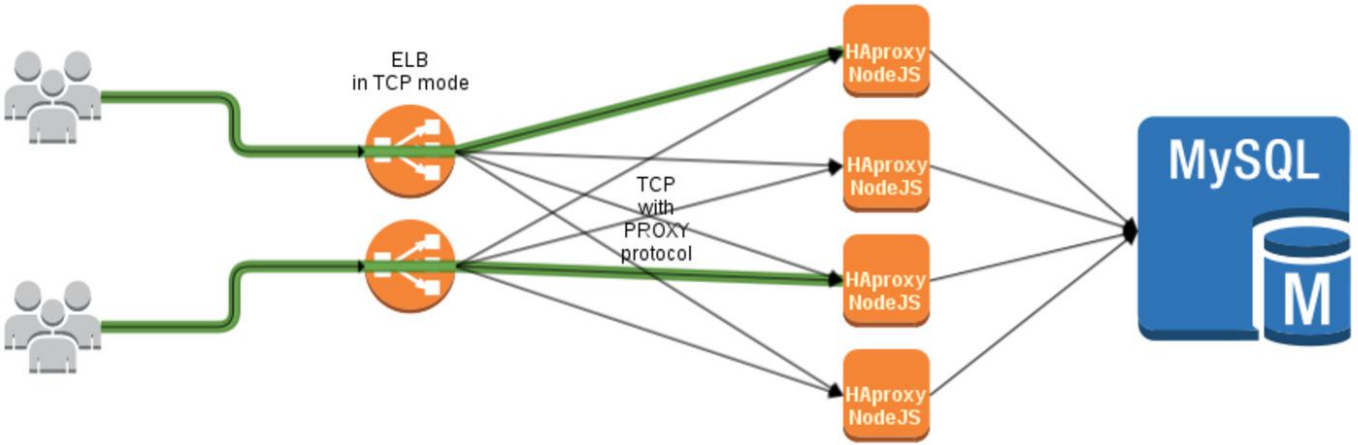
HAProxy应用

提供高可靠、负载均衡的代理

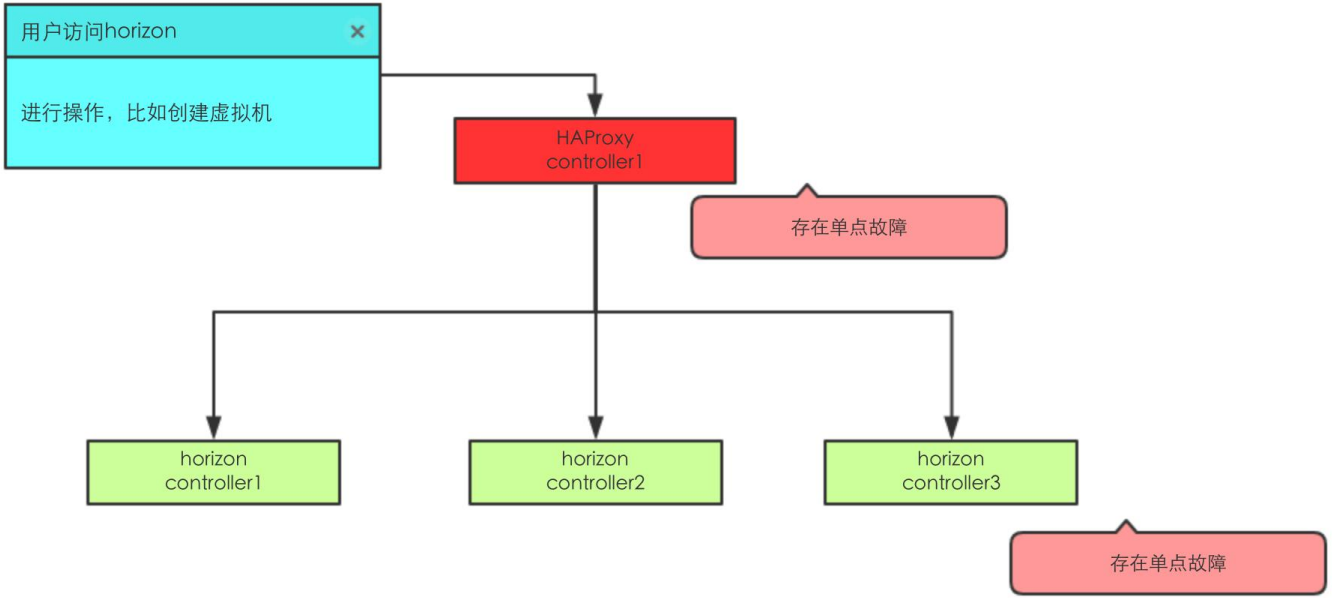
主流的web application负载均衡器

健康检测

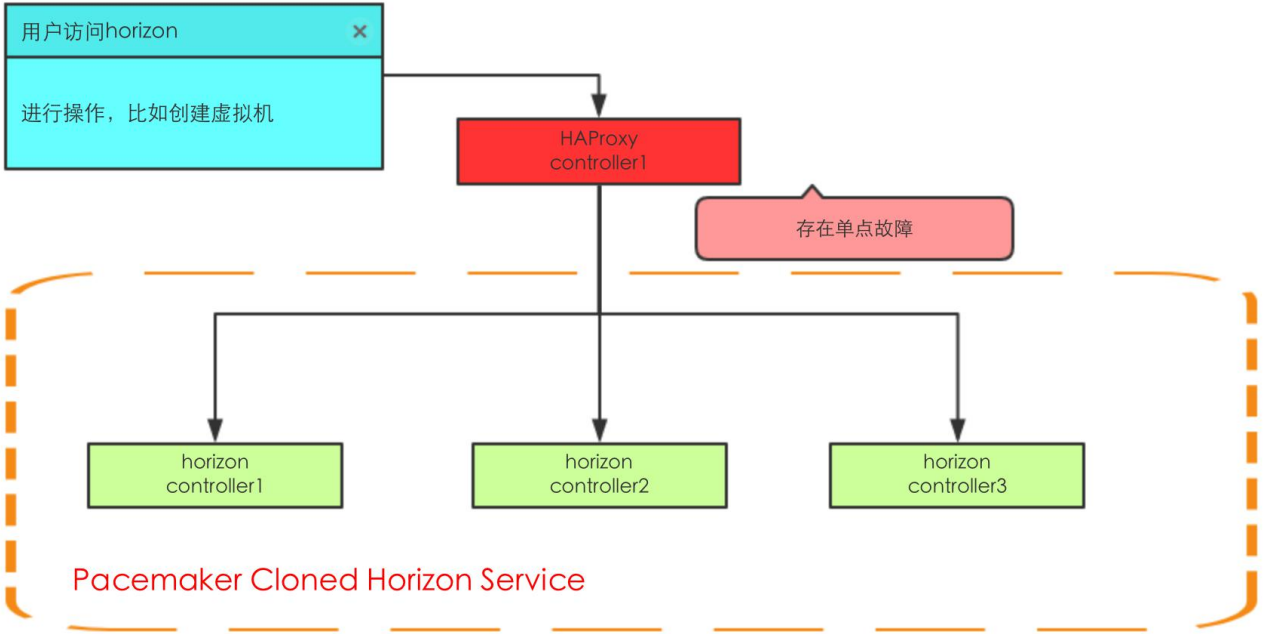
基于ACL的持久性



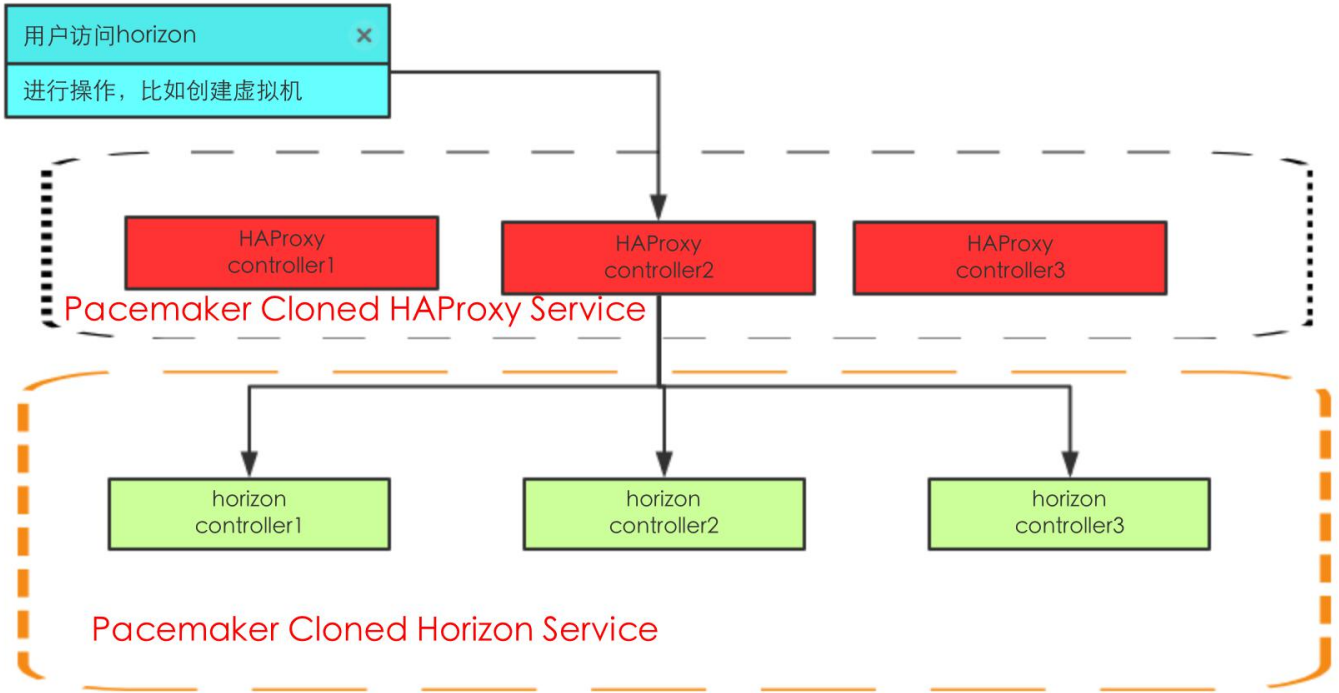
Pacemaker应用



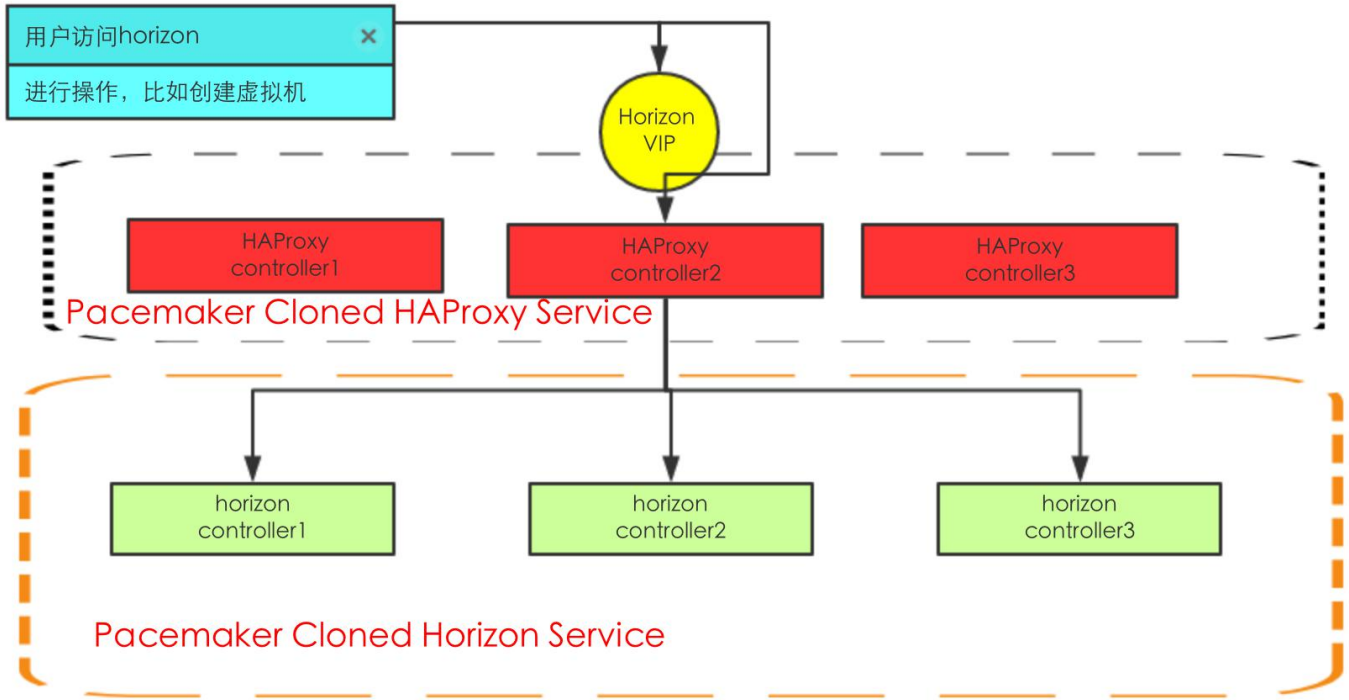
Pacemaker应用



Pacemaker应用



Pacemaker应用

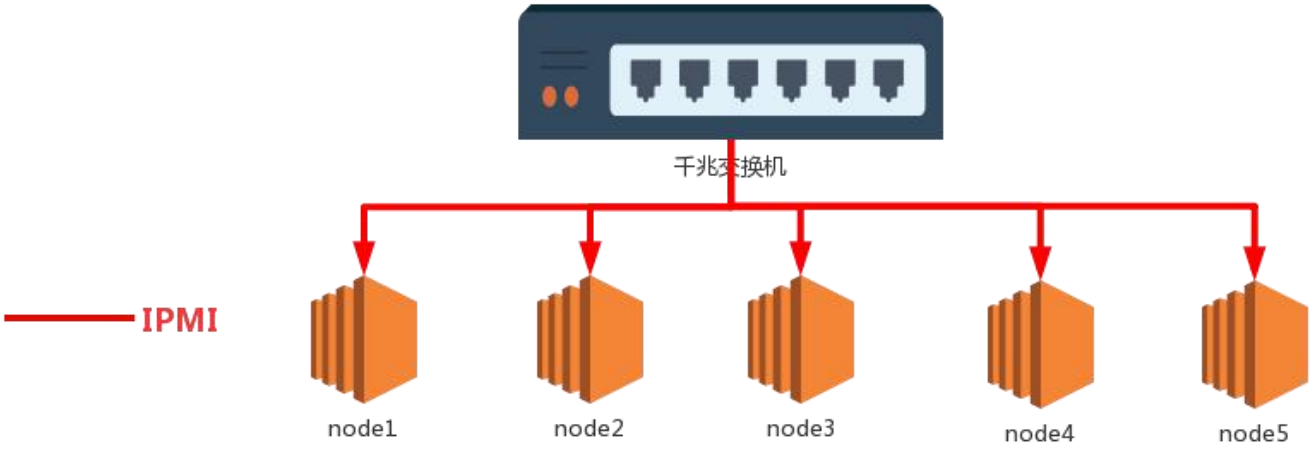


Stonith

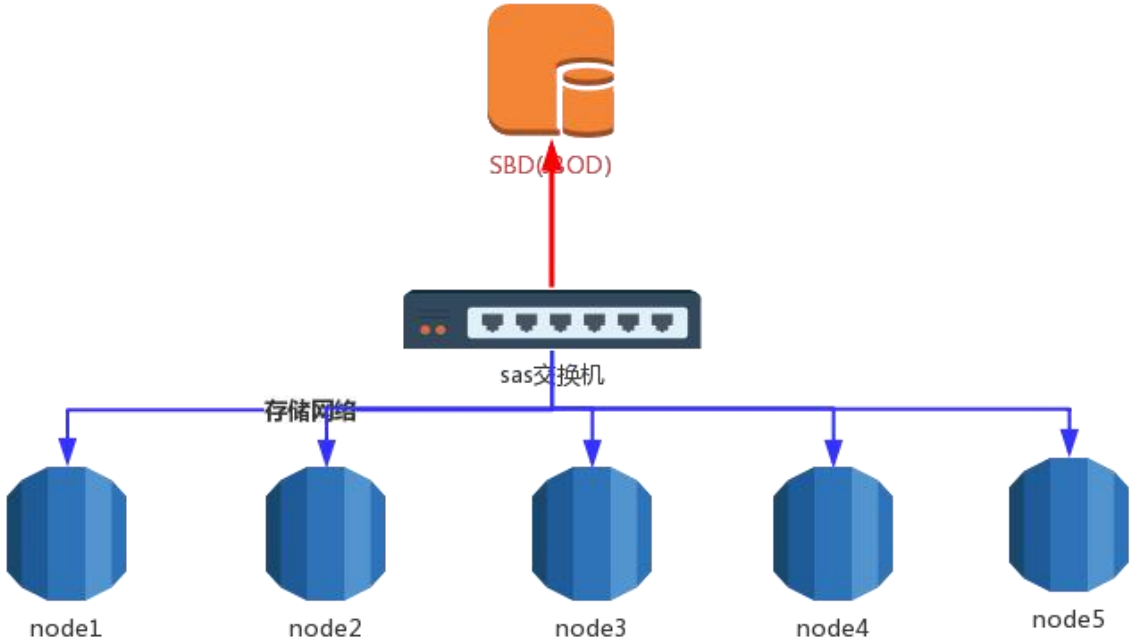
(Shoot-The-Other-Node-In-The-Head)

采用电源或者裸盘隔离节点

stonith实现-IPMI



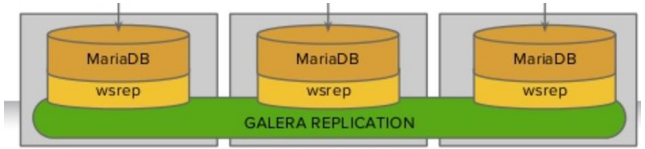
stonith实现-SBD



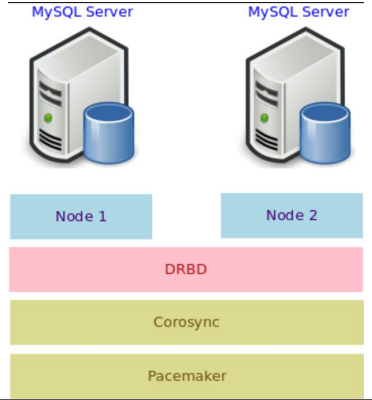
DB MQ高可用实现

DB、MQ HA

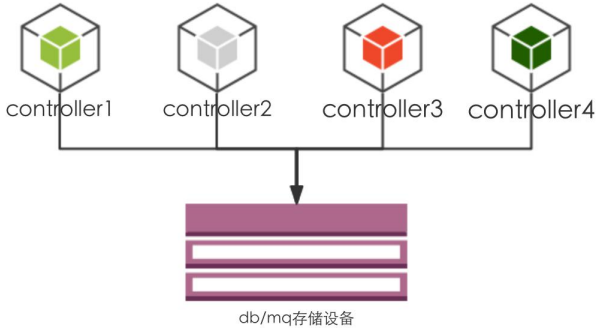
mysql可采用现有成熟方案实现ha，比如galera



对于集群中只有两个controller节点，db/mq可采用drbd的方式

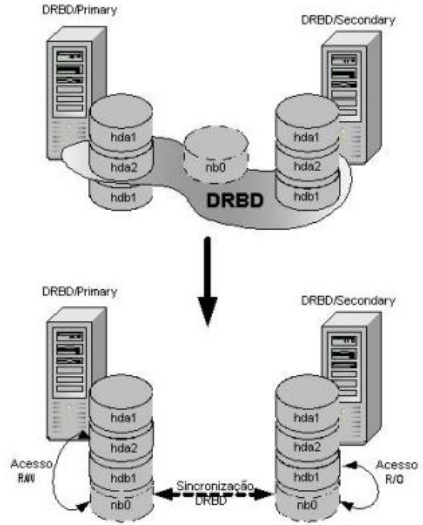


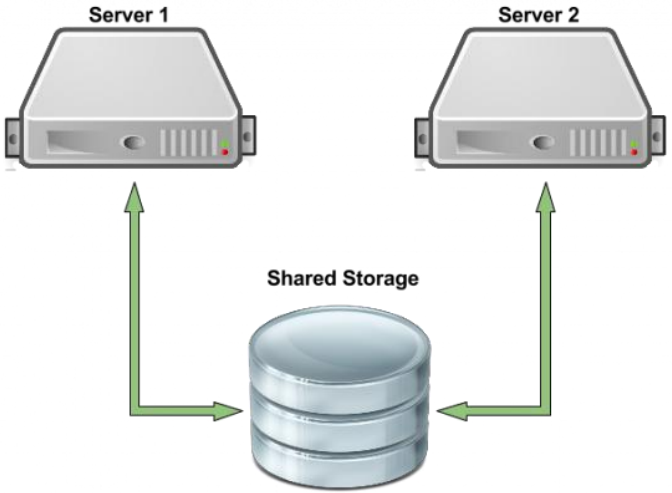
对于集群中只有多个个controller节点，db/mq可采用共享存储的方式



DB、MQ存储

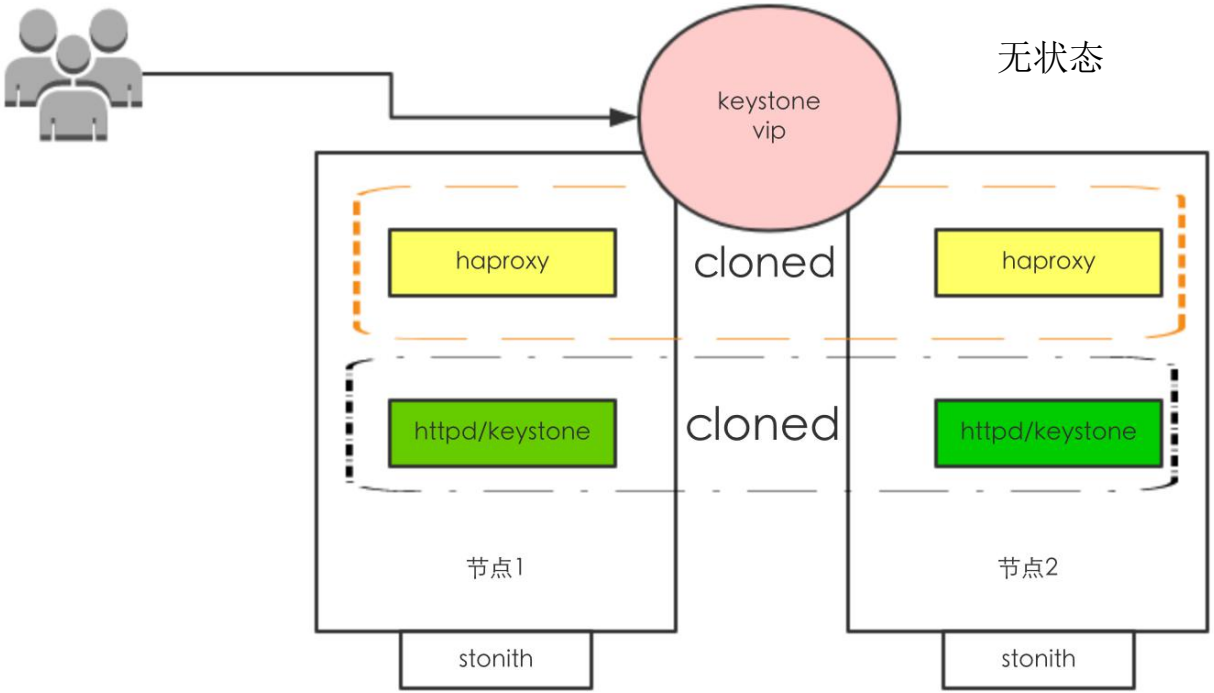
DRBD


DRBD/Primary
DRBD/Secondary
hda1
hda2
hdb1
nb0
DRBD
DRBD/Primary
DRBD/Secondary
Acesso
R/W
Sincronização
DRBD
Acesso
R/O

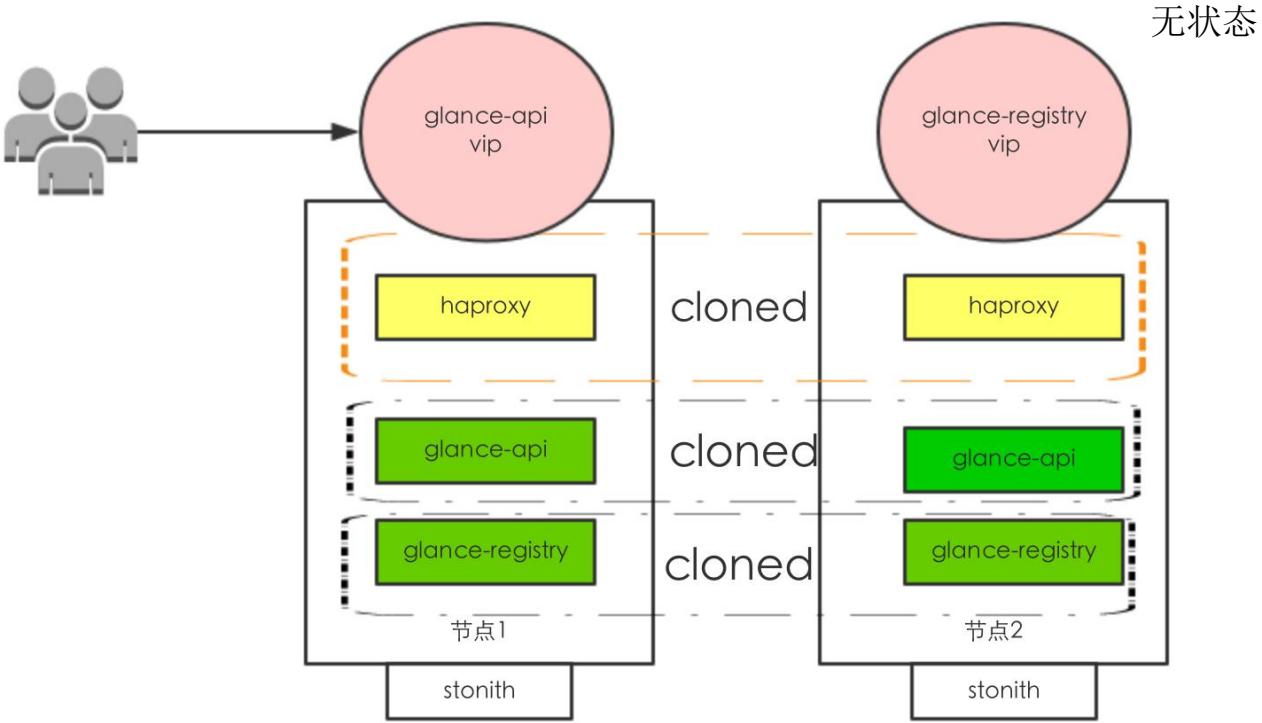

Server 1
Server 2
Shared Storage

OpenStack Services高可用

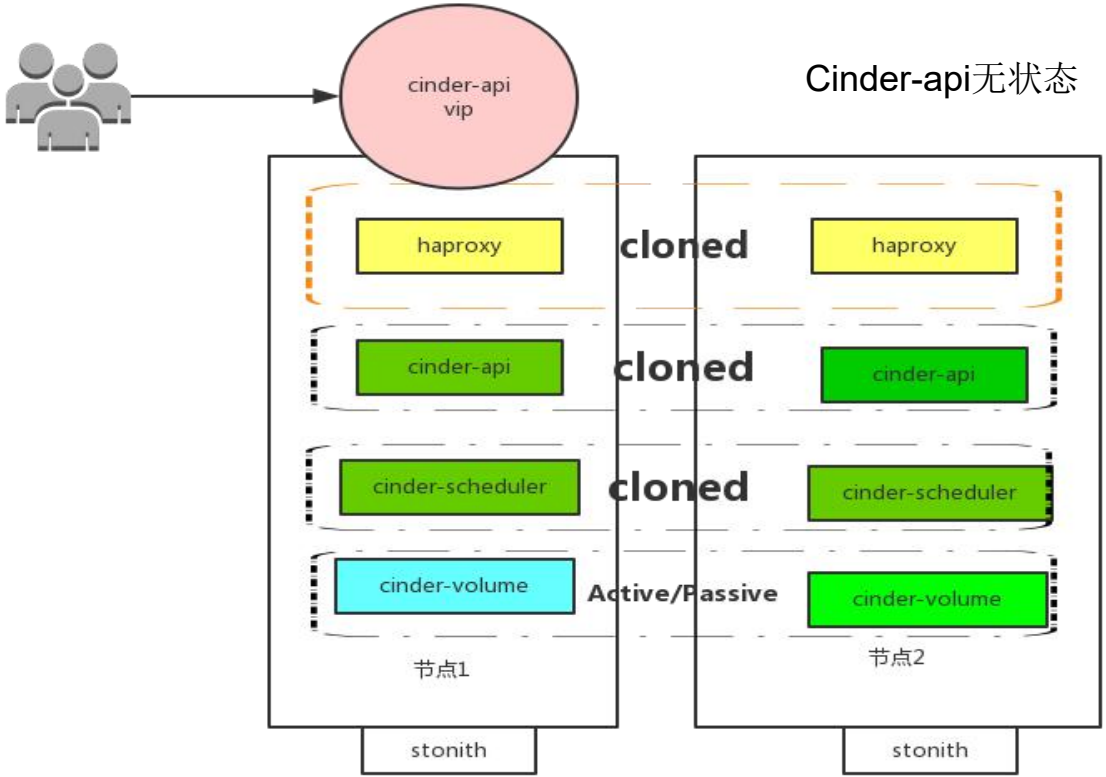
Keystone


keystone
vip
无状态
haproxy
cloned
haproxy
httpd/keystone
cloned
httpd/keystone
节点1
节点2
stonith
stonith

Glance

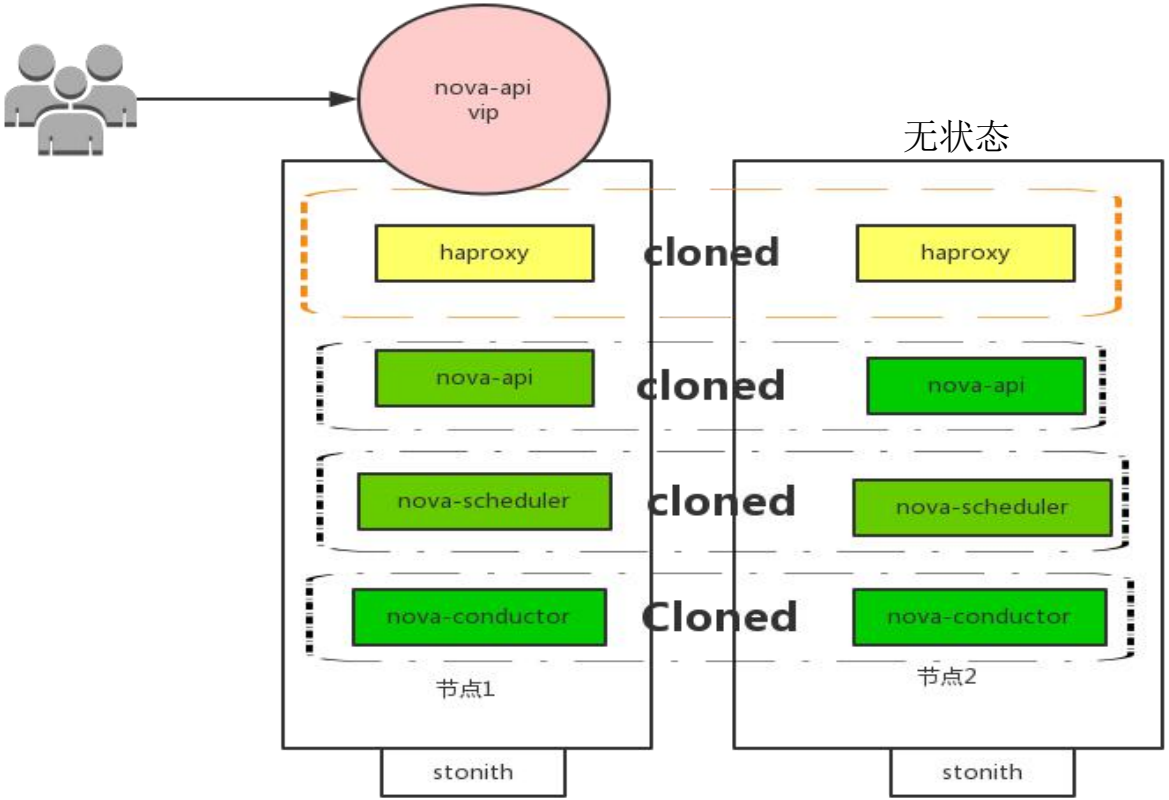

无状态
glance-api
vip
glance-registry
vip
haproxy
cloned
haproxy
glance-api
cloned
glance-api
glance-registry
cloned
glance-registry
节点1
节点2
stonith
stonith

Cinder

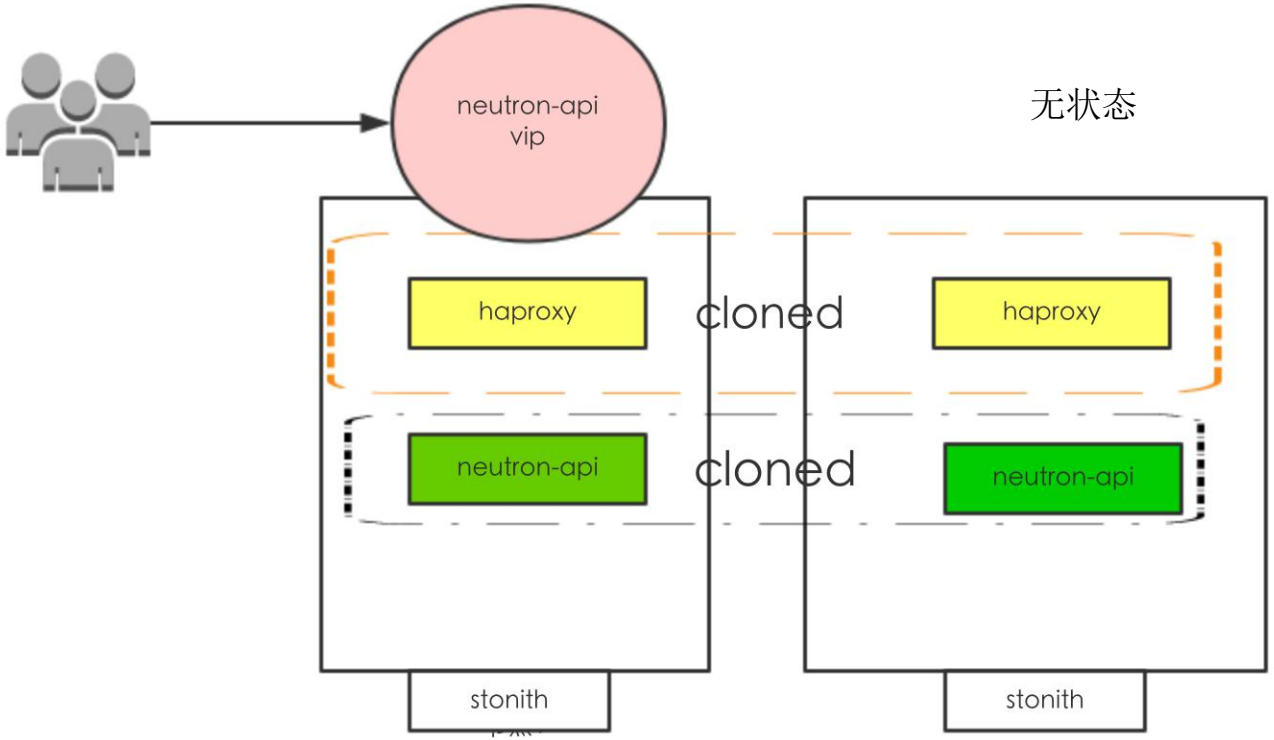


Cinder-api无状态

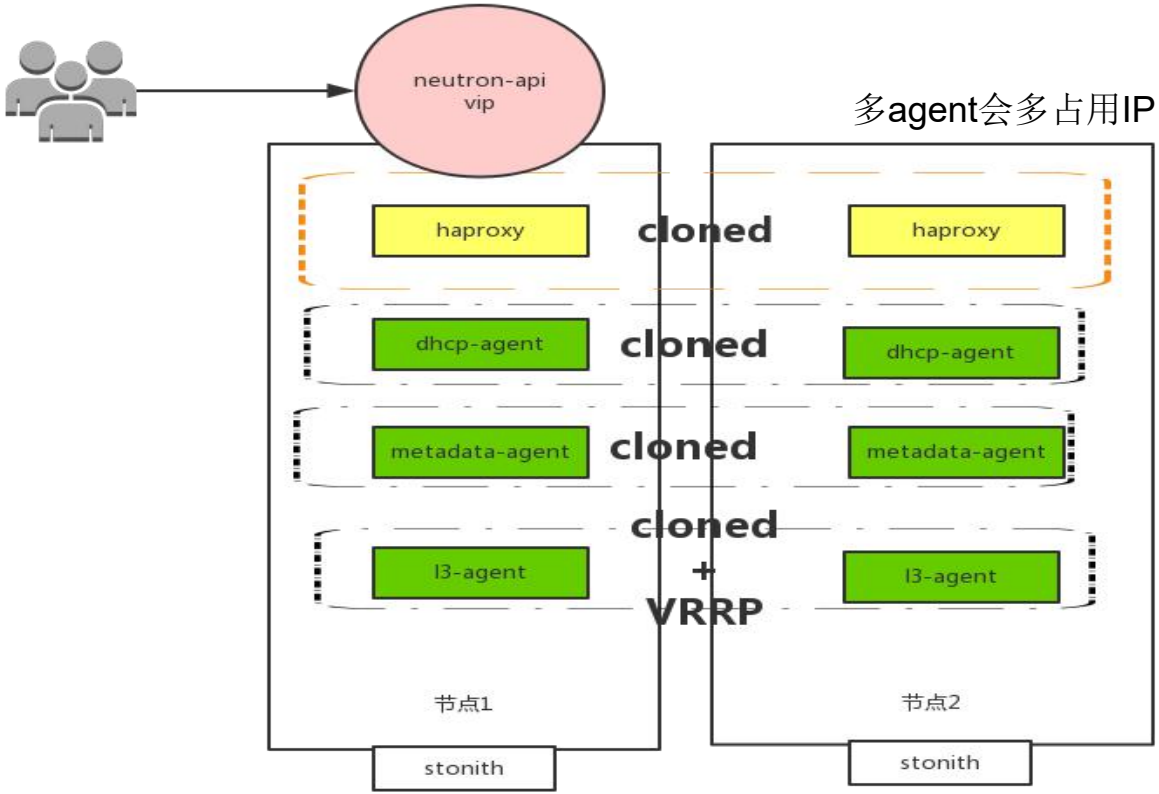
Nova


nova-api
vip
无状态
haproxy
cloned
haproxy
nova-api
cloned
nova-api
nova-scheduler
cloned
nova-scheduler
nova-conductor
Cloned
nova-conductor
节点1
节点2
stonith
stonith

Neutron API


neutron-api
vip
无状态
haproxy
cloned
haproxy
neutron-api
cloned
neutron-api
stonith
stonith

Neutron Agent



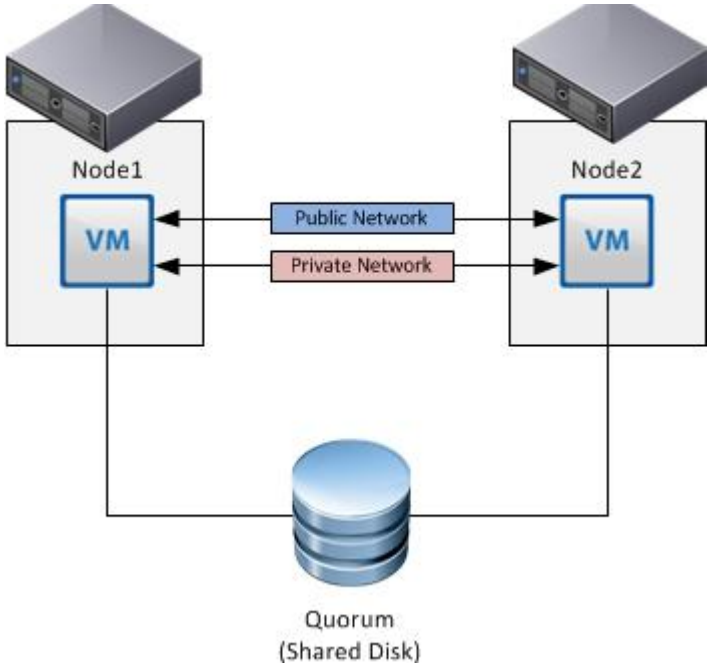
计算节点高可用

Pacemaker_remote

Pacemaker_remote自恢复

虚拟机自动迁移

短暂的不可访问

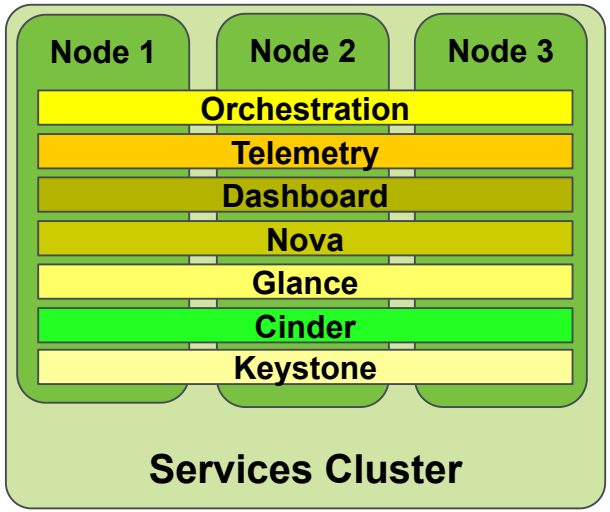


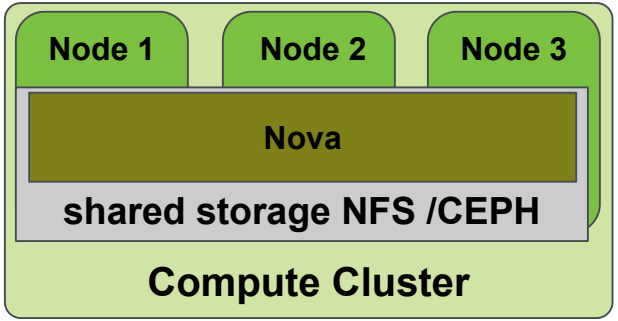
SUSE OpenStack Cloud高可用

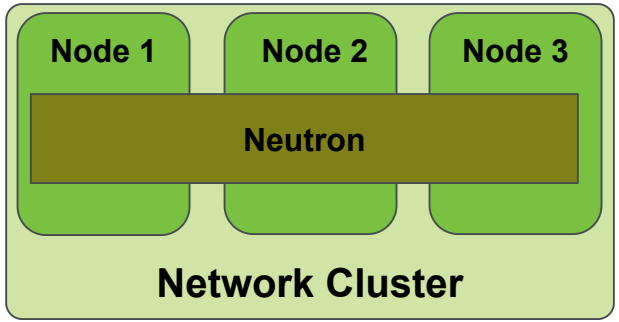
Crowbar Base

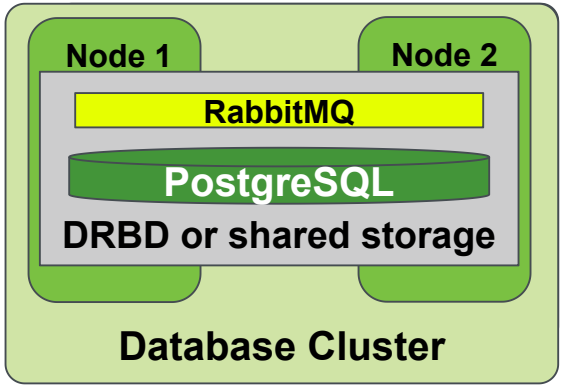
- 由SUSE主要维护
- 基于成熟的SUSE HA套件: Corosync+Pacemaker
- Many Bugs Fixed
- HA组件更好协同
- DRBD支持（不需要共享存储）
- 计算节点HA增强
- 容易部署
- 容易管理和监控Web UI

Crowbar Base HA概览

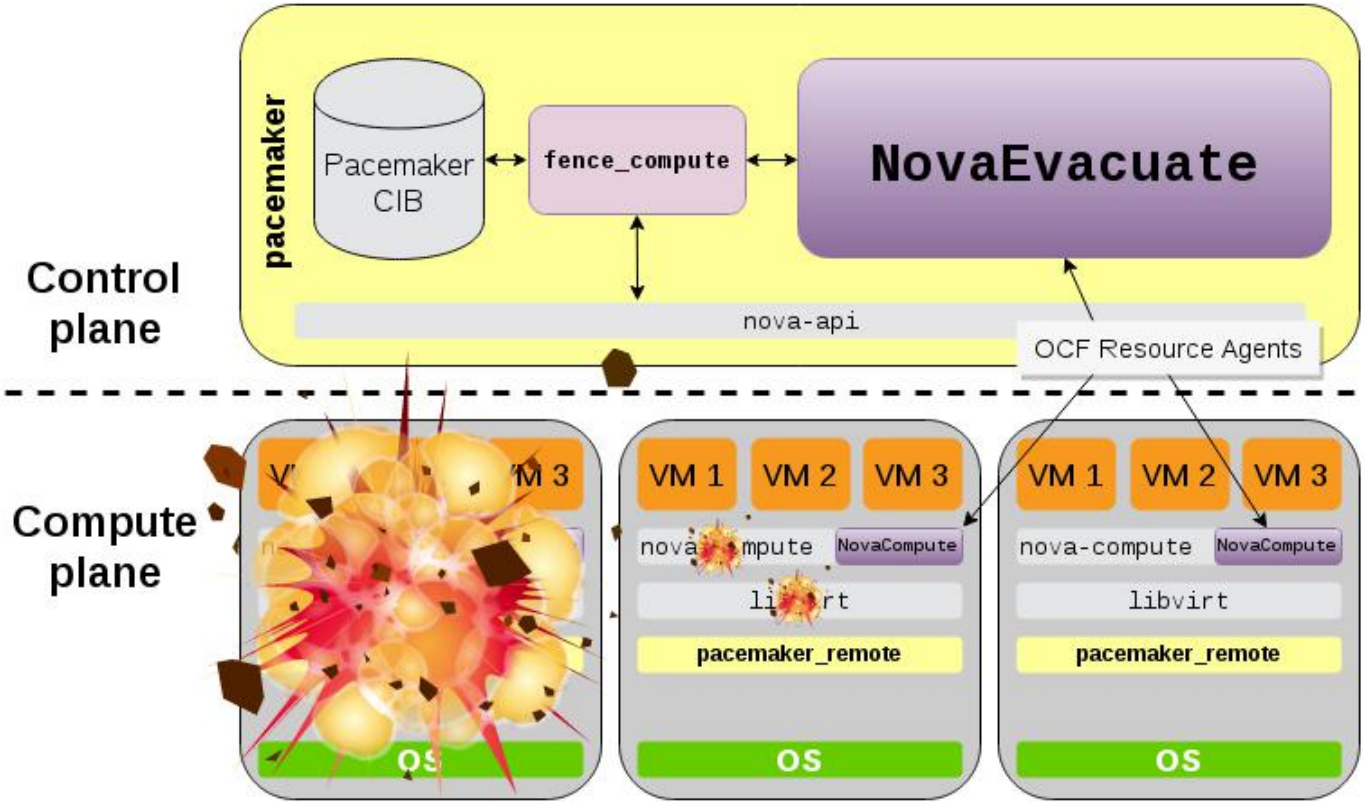




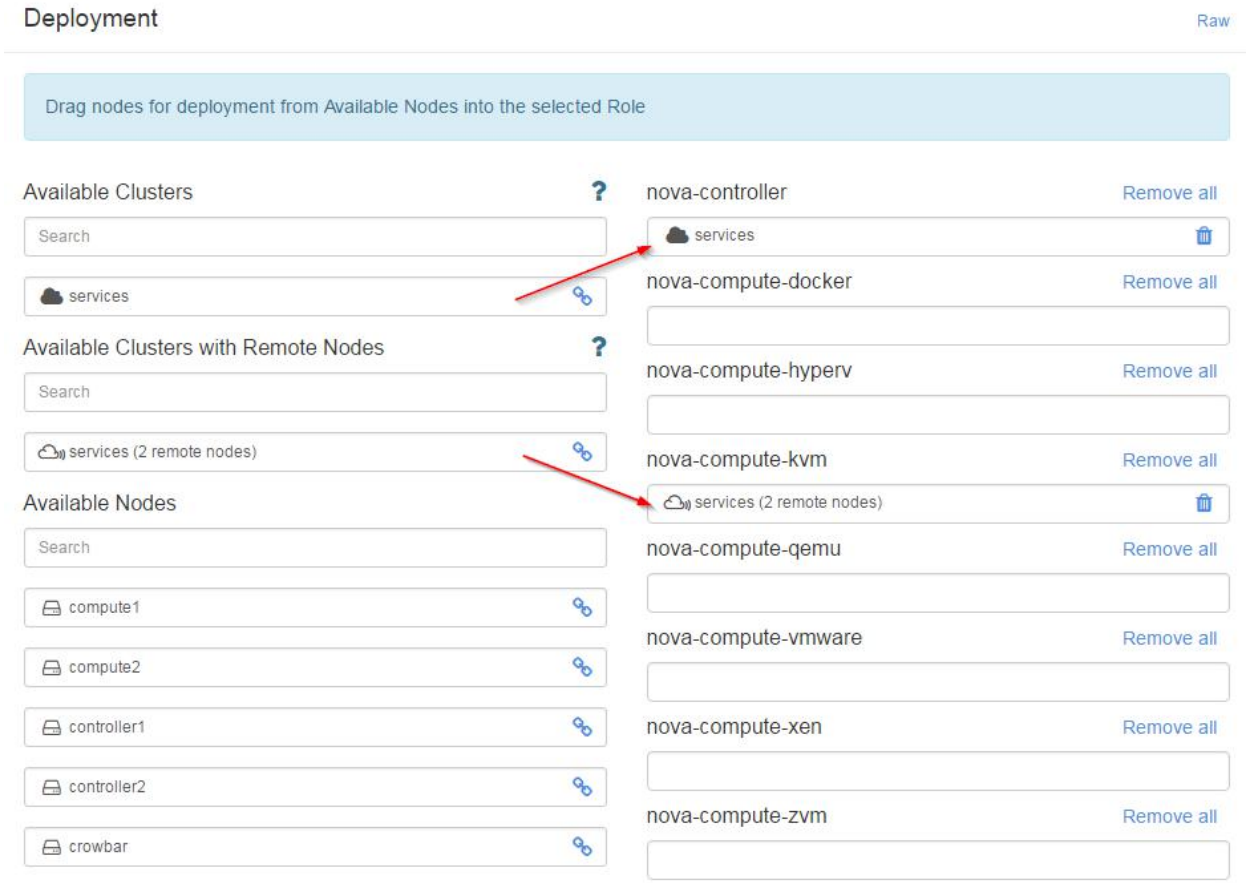




计算节点HA

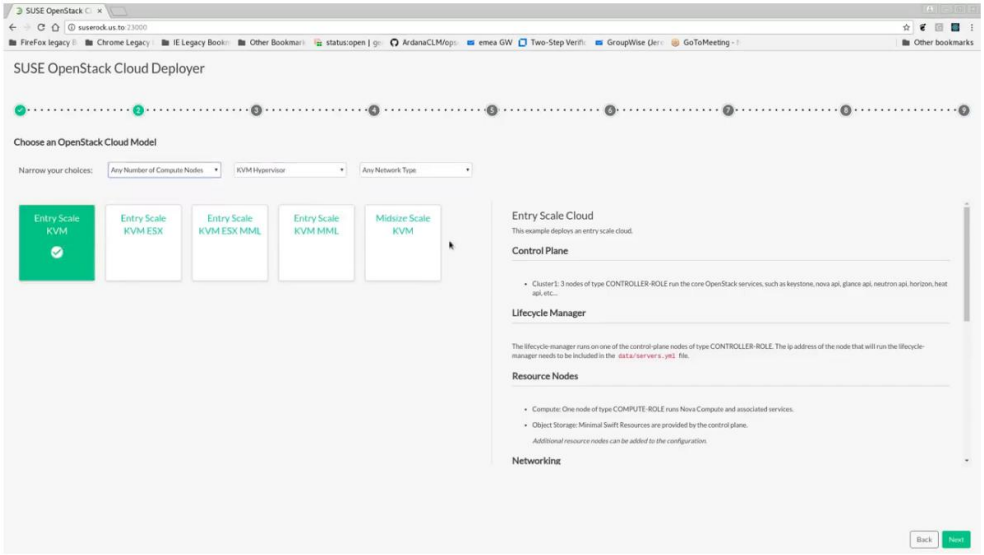

pacemaker
Pacemaker CIB
fence_compute
NovaEvacuate
Control plane
nova-api
OCF Resource Agents
Compute plane
VM 1
VM 2
VM 3
nova-compute
NovaCompute
libvirt
pacemaker_remote
OS

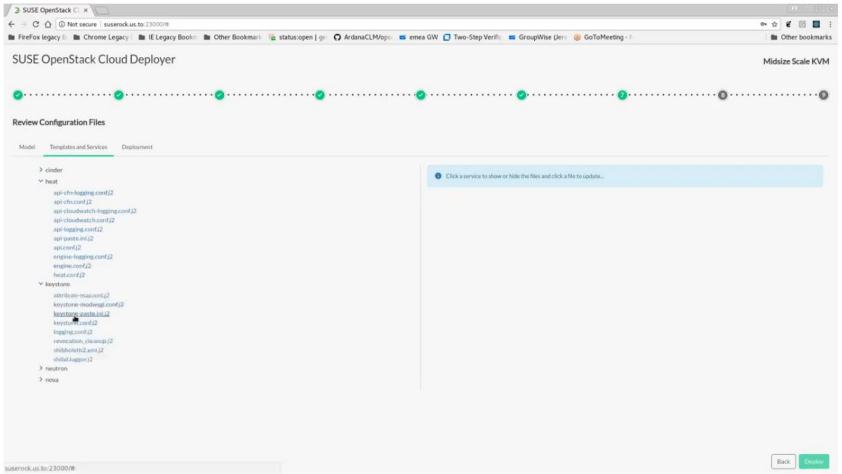
计算节点HA部署举例


Deployment
Raw
Drag nodes for deployment from Available Nodes into the selected Role
Available Clusters
Search
services
Available Clusters with Remote Nodes
Search
services (2 remote nodes)
Available Nodes
Search
compute1
compute2
controller1
controller2
crowbar
nova-controller
Remove all
services
nova-compute-docker
Remove all
nova-compute-hyperv
Remove all
nova-compute-kvm
Remove all
services (2 remote nodes)
nova-compute-qemu
Remove all
nova-compute-vmware
Remove all
nova-compute-xen
Remove all
nova-compute-zvm
Remove all

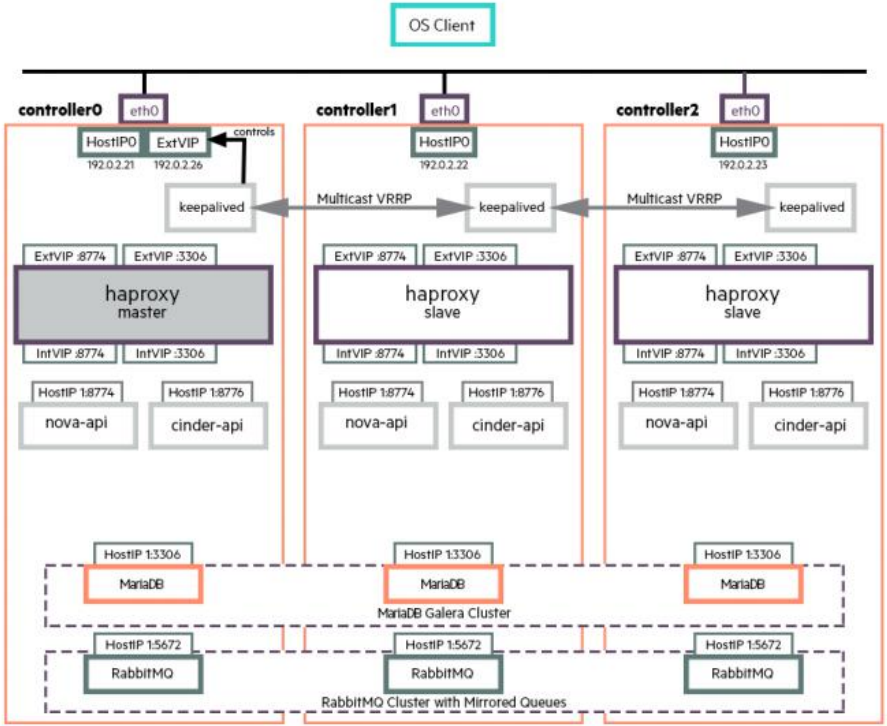
Ardana Base

- 由SUSE主要维护
- 基于成熟的开源HA套件: Haproxy+ Keepalived
- Many Bugs Fixed
- HA组件更好协同
- 容易部署
- 容易管理和监控Web UI

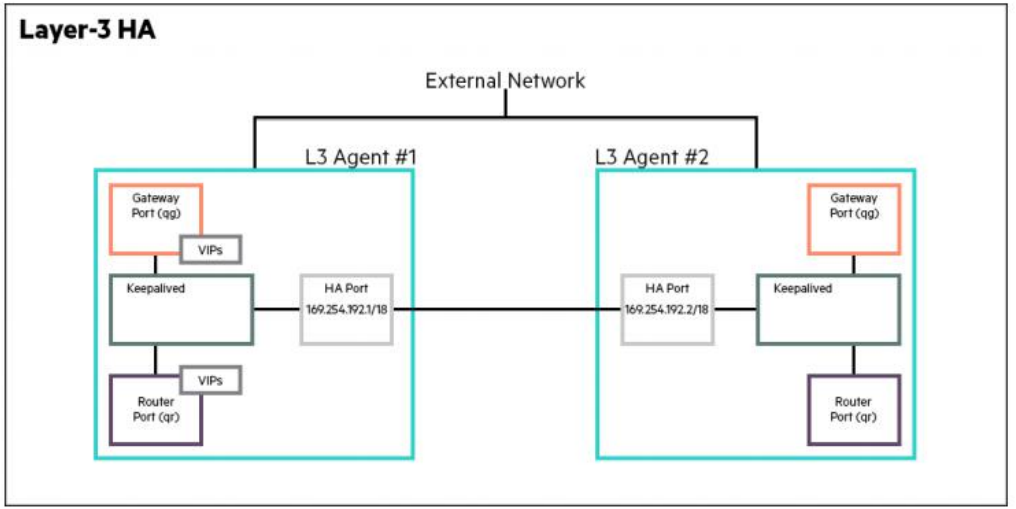




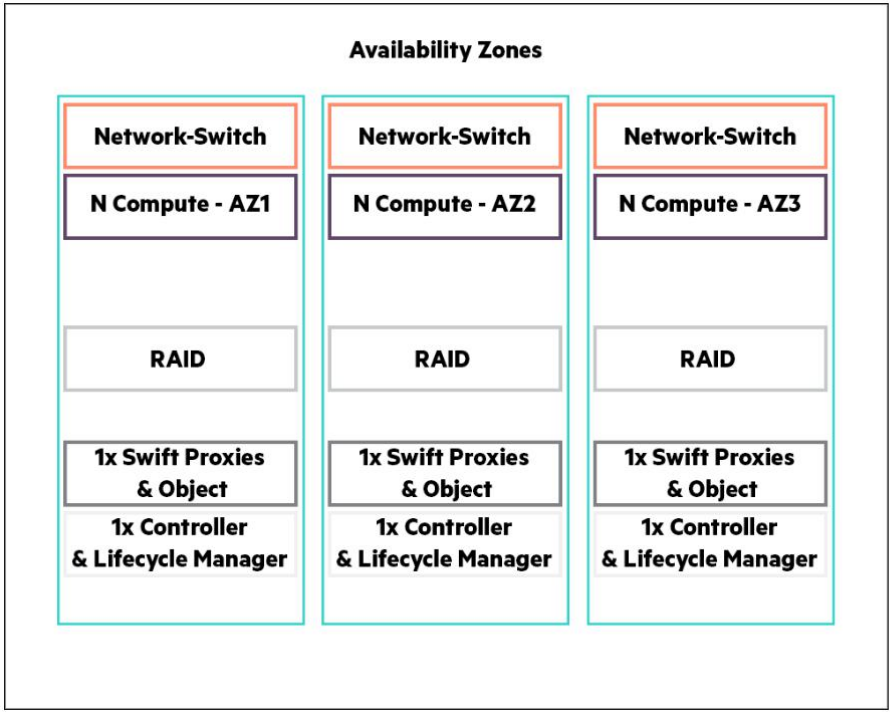
Ardana Base 控制节点HA



Ardana Base L3 Agent HA



Ardana Base HA AZ



Questions