基于 OpenStack 的高可用私有云的实施案例①

唐飞雄[1], 张 利[2], 杨 宁[1]

[1](南京南瑞集团公司 北京中电普华信息技术有限公司, 北京 100192)
[2](江苏省电力公司信通分公司, 南京 210024)

摘 要: 所述的 OpenStack 高可用生产环境案例作为互联网电子商务软件的硬件平台, 实现的架构包括设备和物理网络的冗余设计, 以及基于 Active/Active 的双活模式, 提供了负载均衡能力和系统高可扩展性. 并为生产环境部署了必要的告警监控软件和日志管理工具.

关键词: 基础实施即服务; OpenStack 高可用

OpenStack High-Availability Production Case in Private Cloud

TANG Fei-Xiong[1], ZHANG Li[2], YANG Ning[1]

[1](Beijing China Power Information Technology Co. Ltd, State Grid Electric Power Research Institute, Beijing 100192, China)
[2](Information and Communication Branch, State Grid Jiangsu Electric Power Company, Nanjing 210024, China)

Abstract: This article introduces a case of OpenStack High-Availability production that worked as platform of e-business in Internet. The HA architecture support fault redundancy of compute machines and physical network infrastructure, and run in active/active mode to provide load balance capacity and high scalability. Basically it is installed with monitoring software for system alarm, also management tool for log aggregation.

Key words: IaaS; OpenStack high-availability

当前云计算已成为数据中心 IT 资源的新交付模式, 交付的类型分为 IaaS(Infrastructure as a Service, 基础设施即服务), PaaS(Platform as a Service, 平台即服务), 和 SaaS(Software as a Service, 软件即服务).

在 IaaS 领域, 亚马逊公司的公有云(AWS)技术上被认为是事实的标准. AWS 的弹性计算云(EC2, Elastic Compute Cloud)、弹性块存储(EBS, Elastic Block Storage)和简单存储服务(S3, Simple Storage Service)是基于虚拟化技术的互联网操作系统, 为租户(Tenant)提供可伸缩的虚拟计算资源, 如 vCPUs、vMEM、vNET、vDISK 等.

开源 IaaS 项目大多借鉴了 AWS 的成功经验, 其中 OpenStack 得到业界最广泛的支持, 相关企业如 Canonical、HP、IBM、Red Hat、Cisco、Dell、华为、Intel、NetApp 等都是 OpenStack 基金会的高级会员. 据互联网统计, 在企业私有云和学术研究领域, OpenStack 拥有了最多的用户群. 如欧洲核子研究组织(CERN)在瑞士日内瓦的 OpenStack 已达到 50000 个处理器核规模, 后续还将在匈牙利布达佩斯增加 35000 个处理器核.

1 应用背景

该实施案例是 OpenStack 和高可用技术用于互联网电子商务生产环境的实践. 之所以选择该技术, 基于以下几个优点:

① 与传统生产环境相比较, OpenStack 充分体现了 IaaS 的架构力度和布局灵活性, 将来不必在设备扩容、系统安装、网络配置等各个方面运维管理投入更多资源与时间.

② 弹性伸缩方面, 可以根据虚拟机运行负荷, 动态增加或调整虚拟 CPU 的核数、虚拟内存容量、虚拟硬盘大小等. 传统服务器生产环境中, 必须通过关机, 更换内存, 甚至更换主机等方式来实现, 这往往带来技术风险.

① 收稿时间:2014-10-19;收到修改稿时间:2014-12-02

③ 成本控制方面, 尽管生产环境初期需要相当数量配置很高的计算机硬件, 但虚拟机可以充分利用 OpenStack 基础设施, 长远相比将低于传统上服务器、数据库、中间件所投入的总体成本.

④ 运维管理方面, OpenStack 环境中虚拟机不仅能够在线迁移, 便于对基础设施进行离线维护, 而且还可以对虚拟机进行快照备份, 在发生故障时可快速还原. 在传统的服务器上实现快照和备份需要额外硬件资源, 维护单个设备可能会影响整个系统的运行.

⑤ 高可扩展方面, OpenStack 把基础设施视为资源栈, 扩容时只需进行水平扩展(Scale out); 传统方式往往是通过垂直扩展(Scale up)实现硬件升级和性能提升, 必然花费更多人力和物力去购置、安装和上线调试.

2 OpenStack体系结构

OpenStack 核心是计算虚拟化, 软件定义网络(SDN, Software Defined Network), 软件定义存储(SDN, Software Defined Storage). 软件使用 Apache 2.0 许可证, 由 Nova、Neutron、Cinder、Swift、Keystone、Glance、Horizon 等项目组成. 还集成了关系数据库(如 MySQL), 消息队列(如 RabbitMQ), Apache Httpd 等第三方服务组件.

OpenStack 架构由计算节点, 网络节点, 存储节点, 控制节点等集群组成.

① 计算节点由 Nova 和 Hypervisor(如 KVM, Kernel-based Virtual Machine)组成. Nova 是 OpenStack 计算服务(Compute Service)的项目代号, 管理 Hypervisor 上虚拟机的生命周期.

② 网络节点由 Neutron 服务器实现, 是租户网络的虚拟网络基础设施(VNI, Virtual Network Infrastructure). Neutron 即 OpenStack 网络服务(Network Service), 由软件实现的网络 L2 层和 L3 层代理及插件, 为租户提供基于 VLan、GRE、VxLan 的高级组网技术, 使虚拟机可连通真实世界的物理网络基础设施.

③ 存储节点分别是 Cinder 项目提供的块存储服务(Block Storage Service)和 Swift 项目提供的对象存储服务(Object Storage Service). Cinder 为 Nova 提供虚拟机的数据卷或系统卷, 支持开源的 iSCSI、NFS、Ceph 存储系统, 以及 EMC、Netapp、VMware、华为等商业存储. Swift 是软件实现的具有自愈可靠特点的冗余系统, 可用于虚拟机的安全备份和文件云存储.

④ 控制节点对计算、网络、和存储节点集群进行控制, 提供租户身份认证服务(Identity Service, Keystone 项目)和虚拟机镜像服务(Image Service, Glance 项目), 控制节点管理关系数据库和消息队列. 仪表盘是 Horizon 项目提供的 Web UI.

3 OpenStack的高可用

在数据中心, 高可用的重要性是消除设备单点故障, 使系统能够达到 99.99%的无故障运行, 即一年累计停机时间不超过 1 小时.

OpenStack 高可用提供 Active/Active 双活模式及负载均衡, 能在设备出现故障时自动切换主机和从机, 确保不会发生业务停止或数据丢失的情况.

3.1 数据库高可用

OpenStack 支持 Galera 作为数据库的同步多主(Multi-Master)集群工具. Galera 使用底层并行机制对数据库集群进行读写, 实现同步复制.

3.2 消息队列高可用

OpenStack 的消息队列双活模式支持 RabbitMQ Clustring 和 RabbitMQ Mirrored Queue 高可用技术. 集群内任一节点生产的消息同步到其它节点的镜像队列, 当源节点故障时消息由同步节点接管.

3.3 负载均衡

OpenStack 高可用支持开源的 HAProxy 和 Keepalived 软件技术. HAProxy 作为安全可靠的反向代理软件, 用于快速处理 HTTP 和 TCP 大规模并发连接; Keepalived 提供基于 Linux 内核技术 IPVS(IP Virtual Server)的负载均衡和基于 VRRP(Virtual Router Redundancy Protocol)网络标准的路由故障转移能力.

在 HAProxy 和 Keepalived 的高可用场景中, 集群有一个 Master 节点和至少一个 Backup 节点组成. Master 节点由 Keepalived 通过 IPVS 配置 VIP(Virtual IP), 由 HAProxy 将 VIP 上的请求负载均衡到目标服务器(Back End). Backup 节点的 Keepalived 侦听 VRRP 心跳包对 Master 节点进行检查健康状态, 当原 Master 节点故障时即自动继任为新的 Master 节点.

3.4 OpenStack 控制节点集群的高可用

OpenStack API 是无状态的 REST 服务, 其它控制组件如 Nova 的 conductor 和 scheduler, Cinder 的 scheduler, Glance 的 registry 等也是无状态的服务, 因

此, OpenStack 控制节点集群可配置为 HAProxy 的后端, 实现双活模式.

3.5 OpenStack 网络节点集群的高可扩展

OpenStack 网络节点上的代理组件如 L3 Agent, Metadata Agent 可配置为 Active/Passive 故障转移实现高可扩展模式.

3.6 OpenStack 存储节点集群的高可用

OpenStack 块存储节点的后端如 iSCSI 通过 RAID 技术实现了冗余备份. OpenStack Swift 对象代理节点由 HAProxy 实现双活模式, Swift 对象存储节点自身具备数据复制能力, 因而确保了其存储的 Glance 镜像, Nova 和 Cinder 快照等文件的冗余.

3.7 OpenStack 计算节点集群的高可扩展

OpenStack Nova 项目提供可用区(Availability Zone)技术, 可以将 Hypervisor 按机房, 动力环境, 物理网络架构的不同划分为不同可用区. 可用区还可按 Hypervisor 的类型分为不同的主机聚合(Host Aggregate). 虚拟机可在可用区或主机聚合内备份或迁移, 可用区还能用于配置虚拟机高可用.

4 私有云建设

4.1 硬件环境

4.1.1 机架服务器

私有云采用了 Dell 第 12 代 Power Edge R620 和 R720xd 服务器, 如表 1. 主机配置冗余电源, 通过 Dell 的 iDRAC 专有技术可远程监控主机, 并远程控制电源开关.

表 1 服务器型号及数量表

名称	数量	机型	标识
控制节点	3	R620	controller01/02/03
网络节点	2	R620	network01/02
计算节点	9	R620	compute01/…/09
块存储节点	2	R720xd	blockstorage01/02
Swift 代理节点	2	R620	swiftproxy01/02
Swift 存储节点	6	R720xd	swiftstorage01/…/06
负载均衡节点	2	R620	loadbalance01/02
管理节点	1	R620	admin01

服务器配置 2 个 Intel 至强 E5-2630 v2 CPU, 每个 CPU 含 6 个原生处理核心; 内存相应配置 1600MHz 和 8 倍数据带宽的单列 RDIMM; 以太网卡使用 4 端口 Intel I350 QP 1Gib 千兆网卡和 2 端口 Intel x520 DP 10Gib DA/SFP+ 万兆网卡, 如表 2.

表 2 服务器配置表

节点	配置
控制节点	2 个至强 CPU，12 条 4GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，2 块 240GB 固态盘(RAID-1，数据卷)，1 块 4 端口千兆网卡，冗余电源
网络节点	2 个至强 CPU，12 条 4GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，1 块 4 端口千兆网卡，1 块 2 端口万兆网卡，冗余电源
计算节点	2 个至强 CPU，24 条 8GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，2 块 800GB 固态盘(RAID-6，Nova ephemeral 卷)，1 块 4 端口千兆网卡，2 块万兆网卡，冗余电源
块存储节点	2 个至强 CPU，12 条 4GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，16 块 800GB 固态盘(RAID-6，Cinder LVM)，1 块 4 端口千兆网卡，1 块万兆网卡，冗余电源
对象代理节点	2 个至强 CPU，12 条 4GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，1 块 4 端口千兆网卡，冗余电源
对象存储节点	2 个至强 CPU，12 条 4GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，22 块 900GB SAS(无 RAID，Swift 数据卷)，2 块 4 端口千兆网卡，冗余电源
负载均衡节点	2 个至强 CPU，12 条 4GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，1 块 4 端口千兆网卡，冗余电源
管理节点	2 个至强 CPU，12 条 4GB 内存，2 块 120GB 固态盘(RAID-1，操作系统卷)，2 块 240GB 固态盘(RAID-1，数据卷)，1 块 4 端口千兆网卡，冗余电源

4.1.2 网络交换机

私有云采用了 Dell Force10 S4810P 万兆交换机和 Dell Force10 S50N 千兆交换机, 如表 3.

表 3 交换机配置表

名称	数量	配置
S4810P	2	48 个 10Gib SFP+光以太网口，4 个 40Gib QSFP+中继网口，冗余电源
S50N	3	48 个 1Gib RJ45 以太网口，2 个 12GB 栈链网口，2 个 10Gib XFP 网口，冗余电源

4.2 架构设计

4.2.1 网络设计

网络交换机方面, 除 1 台 S50N 网络交换机单独用于 Dell 服务器的 iDRAC 管理网络外, 其余 4 台网络交换机作为 ToR(Top of Rack)设备. 如图 1, S4810P 交换机之间配置为 VLT(Virtual Link Trunking)专利技术下的链路聚合, S50N 交换机配置栈链模式后上行到 VLT 域. ToR 交换机创建 IEEE 802.3ad 端口组, 实现物理网

络冗余环境.

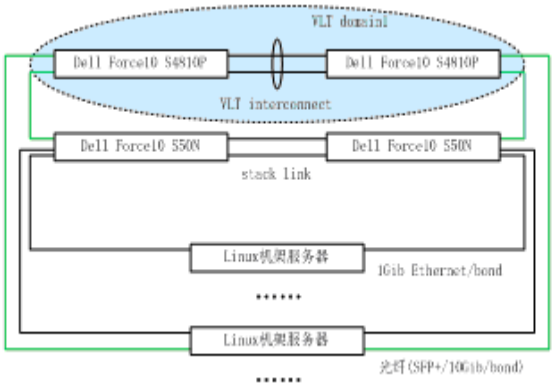

图 1　ToR 链路聚合示意图

服务器使用 CentOS 6.5 操作系统, 以太网端口设置为通道绑定(Channel Bond)模式. 集群被安装到 3 个机架上, 如图 2 为第一个机架安装的节点.

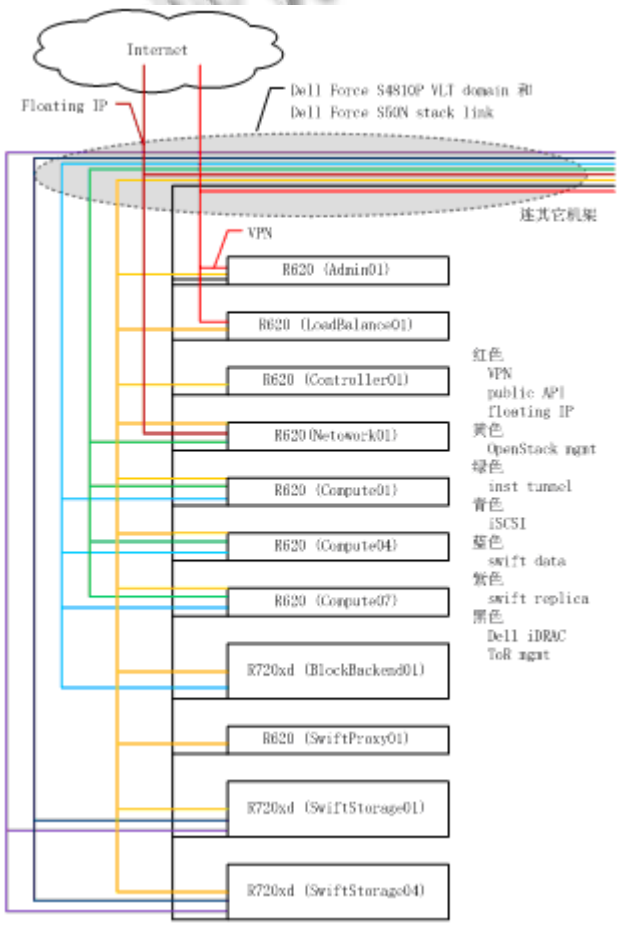

图 2　网络架构图

4.2.2 VLAN 和 IP 划分

① 设备管理网络: VLAN100, 10.0.x.0/24. 用于 Dell 服务器的 iDRAC, 以及网络交换机的远程管理. 管理节点的第四个千兆网口(em4)设置在该网段, 为设备管理网络提供 IP 转发.

② Internet 网络: VLAN101. 用于虚拟机的 Internet 地址段(OpenStack Floating IP)分配到 2 个网络节点的第 3 个千兆网口(em3); 用于 IPSec VPN 服务器提供远程安全访问的 1 个 Internet 地址分配到管理节点的第 3 个千兆网口; 用于 DevOps 环境的 Internet 地址分配到 2 个负载均衡节点的第 3 个千兆网口, 是 OpenStack API 的外部网络.

③ OpenStack 管理网络: VLAN103, 10.3.x.0/24. 分配给全部服务器的千兆以太网通道绑定端口组.

④ 虚拟机网络的隧道: VLAN104, 10.4.x.0/24. 分配给 2 个网络节点和 9 个计算节点的万兆以太网通道绑定端口组.

⑤ iSCSI 存储网络: VLAN105, 10.5.x.0/24. 分配给 9 个计算节点和 2 个块存储节点的万兆以太网通道绑定端口组.

⑥ ○6 OpenStack 对象存储网络: VLAN106, 10.6.x.0/24. 分配给 2 个对象代理节点和 6 个对象存储节点的千兆以太网通道绑定端口组.

⑦ OpenStack 对象复制网络: VLAN107, 10.7.x.0/24. 分配给 6 个对象存储节点的千兆以太网通道绑定端口组.

表 4　VLAN 和网络端口表

节点	VLAN	端口	IP
控制节点	100	iDRAC	10.0.x.20~10.0.x.22
	103	bond0	10.3.x.20~10.3.x.22
网络节点	100	iDRAC	10.0.x.30, 10.0.x.31
	101	em3	Internet
	103	bond0	10.3.x.30, 10.3.x.31
计算节点	100	iDRAC	10.0.x.40~10.0.x.48
	103	bond0	10.3.x.40~10.3.x.48
	104	bond2	10.4.x.40~10.4.x.48
	105	bond3	10.5.x.40~10.5.x.48
块存储节点	100	iDRAC	10.0.x.60, 10.0.x.61
	103	bond0	10.3.x.60, 10.3.x.61
	105	bond3	10.5.x.60, 10.5.x.61
对象代理节点	100	iDRAC	10.0.x.70, 10.0.x.71
	103	bond0	10.3.x.70, 10.3.x.71
	106	bond2	10.6.x.70, 10.6.x.71
对象存储节点	100	iDRAC	10.0.x.80~10.0.x.85
	103	bond0	10.3.x.80~10.3.x.85
	106	bond2	10.6.x.80~10.6.x.85

	107	bond3	10.7.x.80~10.7.x.85
负载均衡节点	100	iDRAC	10.0.x.11, 10.0.x.12
	101	em3	Internet
	103	bond0	10.3.x.11, 10.3.x.12
管理节点	100	iDRAC	10.0.x.10
		em4	10.0.x.1
	101	em3	Internet
	103	bond0	10.3.x.10
S4810P 交换机	100	管理端口	10.0.x.2, 10.0.x.3
S50N 交换机	100	port45	10.0.x.4, 10.0.x.5

4.2.3 VIP 设置

负载均衡节点由 Keepalived 管理 3 个 VIP, 作为 HAProxy 的前端服务地址.

10.3.x.100: OpenStack Keystone, Glance, Nova, Cinder, Neutron 的 API 地址, 同时也为 OpenStack Horizon, MySQL Garela, RabbitMQ HA 等的服务地址.

10.3.x.200: OpenStack Swift 的 API 地址.

Internet 地址: OpenStack API 的外部网络服务地址, 用于 DevOps 和持续集成环境.

4.3 管理工具

① 使用 Puppet 作为服务器自动化工具, 管理节点安装 Puppet Server, 其它节点安装 Puppet Agent.

② 服务器均安装网络时间服务(CentOS 的 ntpd 服务). 管理节点上的网络时间服务作为其它节点的时间同步源.

③ 管理节点安装 DNS 服务(CentOS 的 named 服务), 为 OpenStack 管理网络提供内部域名解析. 另外, 安装 OpenVPN 服务, 提供 IPSec VPN 访问.

④ 使用 Zabbix 作为生产环境的告警和监控工具, 管理节点为 Zabbix Server, 其它节点为 Zabbix Agent. 另外使用 Logstash 作为日志聚合工具, 管理节点安装 Logstash Server, 其它节点安装 Logstash Agent.

4.4 冒烟测试

① 断开任一台 ToR 交换机, 此时, 网络容错管理环境, Zabbix, Logstash 均显示和记录告警. 检查 OpenStack 工作能照常进行, 因此确认生产环境无单点故障. 恢复交换机, 告警结束.

② 断开 OpenStack 节点的任一条网线, 检查到告警, 但生产环境能工作正常.

③ 断开任一负载均衡节点, 确认生产环境无单点故障.

④ 断开任一 OpenStack 控制节点, 确认 OpenStack 组件, 数据库, 消息队列集群均能工作. 然后依次测试网络节点, 计算节点, 和块存储节点.

⑤ 断开任一 OpenStack 对象代理节点, 确认无单点故障. 再断开任一对象存储节点, 存取操作没有受到影响, 检查数据完整性, 确认正常.

5 问题分析与解决

在本案例实施前, 首先部署了一套由 4 台网络交换机和 18 台机架服务器组成的 OpenStack Staging 环境, 该环境帮助发现了设计存在的不足并在之后的生产系统中进行了改进. 具体包括:

① 用网段隔离管理所需的 Internet 地址与分配给虚拟机的 Internet 地址, 避免虚拟机网络发生异常时(如网络风暴)影响到 OpenStack 远程管理.

② 使用独立的负载均衡节点安装 HAProxy 和 Keepalived. 在 Staging 环境中用 OpenStack 控制节点安装负载均衡服务, 控制节点即为 OpenStack API 的负载均衡后端, 同时又是负载均衡前端. 为了使控制节点不暴露在互联网上, 需要在管理节点上配置 NAT 进行路由.

6 总结

经历了为期 4 个月的实验, 并在 Staging 环境的实施中得到验证和改进, 该案例所述的 OpenStack 高可用生产环境已交付使用并运行互联网电子商务平台至今, 可为企业使用开源技术建设 IaaS 环境提供参考.

参考文献

1 Fifield T, Fleming D, Gentle A, Hochstein L, Proulx J, Toews E, Topjian J. OpenStack Operations Guide: Set Up and Manage Your OpenStack Cloud. California: O'Reilly Media, 2014.

2 Arnold J, et al. OpenStack Swift: Using, Administering, and Developing for Swift Object Storage. California: O'Reilly Media, 2014.

3 Olups R. Zabbix 1.8 Network Monitoring. Birmingham: Packt Publishing, 2010.

4 Turnbull J. The Logstash Book. Kindle Ed., 2014.

5 Uphill T. Mastering Puppet. Birmingham: Packt Publishing, 2014.

6 Hobson J. CentOS 6 Linux Server Cookbook. Birmingham: Packt Publishing, 2013.

7 Fifield T. Introduction to OpenStack. Linux Journal, 2013, 11: 68–69.

8 Loschwitz M. Ceph and OpenStack join forces. Linux Magazine, 2014, 6(165).