



支持不同存储方式的 OpenStack 云平台设计与实现

冯伟斌, 张云勇, 王志军

(中国联合网络通信有限公司研究院, 北京 100032)

摘要:随着云计算的发展,用开源软件实现云平台的搭建日益便捷,目前云平台通常采用本地硬盘或共享存储两种方式来存储虚拟机文件,针对两种存储方式存在的缺点,提出了一种支持两种存储方式的 OpenStack 云平台搭建方法,搭建的云平台有效实现了对两种存储方式的支持。用户通过指定不同的参数,创建相应存储方式的虚拟机,从而灵活满足了不同业务的存储需求。所提方法解决了云平台对不同存储方式的支持,有效促进了现网系统向 OpenStack 云平台的迁移,对业务平台云化具有重要的推动作用。

关键词:云计算;虚拟机;OpenStack;本地硬盘;共享存储

中图分类号:TP393

文献标识码:A

doi: 10.11959/j.issn.1000-0801.2016097

Design and implementation of OpenStack cloud platform for different storage modes

FENG Weibin, ZHANG Yunyong, WANG Zhijun

Research Institute of China United Network Communications Group Co., Ltd., Beijing 100032, China

Abstract: With the development of cloud computing, it is more convenient to implement cloud platform by using open source software, cloud platforms usually use local hard disk or shared storage to store virtual machine file. In view of the advantages and disadvantages of these two kinds of storage modes, a method supporting two kinds of storage modes in cloud platform building based on OpenStack was presented. Users specified different parameters to create virtual machine in a corresponding storage mode, so as to meet the needs of different storage business. The method solved the support of cloud platform for different storage modes, promoted the existing network system to migrate to OpenStack cloud platform and played an important role in cloud service platform.

Key words: cloud computing, virtual machine, OpenStack, local disk, shared storage

1 引言

OpenStack 是目前业内最流行的开源云平台^[1],能够快速帮助用户实现类似于 Amazon EC2 和 S3 的云基础设施服务。OpenStack 最新版本 Kilo 已经于 2015 年 4 月 30 日正式发布,这是 OpenStack 的第 11 个版本^[2]。与 OpenStack 云平台的虚拟机存储架构不同,虚拟机可以创建在 PC 服

务器的本地硬盘上,也可以创建在基于 PC 服务器搭建的分布式共享存储上,两者各有优缺点。本地硬盘虚拟机的性能更为稳定,但不支持实时迁移。共享存储虚拟机支持实时迁移,但性能受到网络带宽的限制。

本文充分考虑了基于本地硬盘或共享存储的虚拟机的优缺点,基于 OpenStack 实现了一种能够在本地硬盘或共享存储上创建虚拟机的云平台。用户通过指定不同参数,就可以创建



相应存储方式的虚拟机,从而灵活满足业务的不同需求。

2 OpenStack 组件简介

目前 OpenStack 主要包括 8 个核心组件,分别是用户界面 (Horizon)、计算 (Nova)、认证 (Keystone)、网络 (Neutron)、监控 (Ceilometer)、镜像 (Glance)、块存储 (Cinder)、对象存储 (Swift),如图 1 所示。每个组件都是多个服务的集合,一个服务意味着运行着一个进程。

(1) 用户界面

为所有 OpenStack 服务提供一个模块化的基于 Django 的界面,通过这个界面,不论是最终用户还是运维人员都可以完成大多数的操作,如启动虚拟机、分配 IP 地址、动态迁移等。

(2) 计算

类似于 Amazon 的 EC2 服务,为用户提供虚拟机的增、删、改、查服务,如虚拟机的创建、重启、销毁、迁移等。

(3) 认证

为所有 OpenStack 服务提供身份验证和授权功能,跟踪用户权限,提供一个可用服务及 API 列表。

(4) 网络

提供云计算环境下的虚拟网络功能,允许用户创建自己的虚拟网络并连接各种网络设备接口。

(5) 监控

将 OpenStack 内部发生的几乎所有的事件都收集起

来,然后为计费 and 监控以及其他服务提供数据支撑。

(6) 镜像

提供虚拟机镜像的存储、查询和检索服务,通过提供一个虚拟磁盘映像的目录和存储库,为 Nova 虚拟机提供镜像服务。

(7) 块存储

类似于 Amazon 的 EBS 服务,为用户提供块存储服务,能够创建数据卷,并挂载到虚拟机上。

(8) 对象

存储类似于 Amazon 的 S3 服务,为用户提供对象存储服务,能够以低成本的方式管理大量无结构数据。

3 OpenStack 云平台存储架构

与虚拟机存储直接相关的 OpenStack 组件包括 Nova、Glance、Cinder,如图 2 所示,目前 OpenStack 已经支持三者在 Ceph、GlusterFS 等分布式共享存储上的统一存储。其中 Nova 负责虚拟机整个生命周期的管理,例如创建、重启、关机、销毁等,虚拟机自身的磁盘文件由 Nova 负责生成,属于本地易失性存储,即当虚拟机销毁后,相应的磁盘文件也会被销毁;Glance 提供虚拟机镜像服务,负责虚拟机镜像文件的存储、查询和检索;Cinder 为虚拟机提供持久块存储服务,虚拟机创建完成后,可以通过 Cinder 挂载块设备,并具备快照、备份等功能,类似于 AWS 的 EBS 服务^[3]。

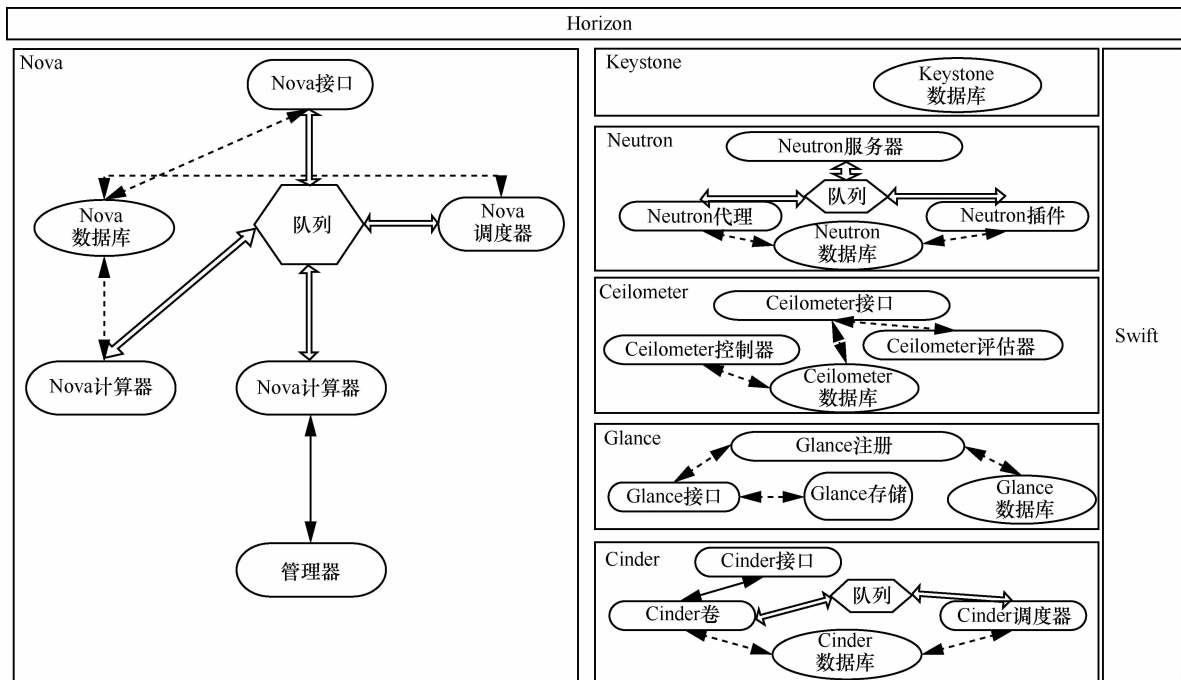


图 1 OpenStack 主要组件架构

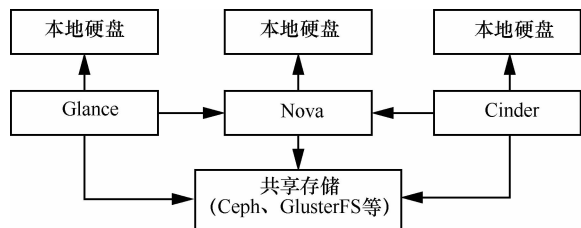


图2 OpenStack 云平台的虚拟机存储架构

(1) 本地硬盘虚拟机即虚拟机本身的文件位于 PC 服务器的本地硬盘上

主要优点在于服务器硬盘通过做 RAID (redundant arrays of independent disk), 除非发生硬件故障, 否则不会出现软件问题, 因此, 虚拟机性能更稳定。同时, 由于磁盘 I/O 不涉及网络流量交换, 性能也更为高效。主要缺点在于虚拟机在不同服务器之间迁移时, 需要将虚拟机文件从一台服务器通过网络复制到另外一台服务器, 导致不支持实时迁移, 同时还需对虚拟机文件做额外备份。

(2) 共享存储虚拟机即虚拟机本身的文件位于分布式共享存储上

目前业内流行基于 Ceph 或 GlusterFS 开源软件实现^[4]。由于所有计算服务器都可以访问共享存储, 因此, 自然支持虚拟机的实时迁移, 同时分布式共享存储自带对文件进行多个备份, 可靠性更高。主要缺点在于虚拟机进行读写操作时, 涉及整个云平台中的网络流量交换, 因此, 磁盘 I/O 会受到网络带宽的限制, 同时分布式共享存储在软件实现层面较为复杂, 与服务器硬盘相比, 稳定性略差。

4 OpenStack 云平台的设计实现

OpenStack 云平台中虚拟机的创建由 Nova 组件负责, 整个服务调用过程如图 3 所示, 每个服务都是单独的进程实例, 之间通过 RPC 相互调用^[5]。其中 Nova-API 负责验证和处理 HTTP 请求信息, 对外提供 REST API 供用户调用; Nova-conductor 负责与数据库交互, 提高数据库访问的安全性; Nova-scheduler 负责服务调度, 决定虚拟机实例创建在哪个服务器上; Nova-compute 作为虚拟机管理程序, 具

体完成虚拟机的创建和控制。

用户调用 OpenStack 云平台的 Keystone 服务得到认证 token 后, 通过调用 REST API, 向 Nova-API 服务发送创建虚拟机的 HTTP 请求。虚拟机创建的函数调用过程见算法 1, 算法 1 为 OpenStack 虚拟机创建的函数调用过程的伪代码, OpenStack 最终调用 nova/virt/libvirt/driver.py 中的 _create_image 和 _create_domain_and_network 函数具体创建虚拟机实例, 表示控制虚拟机文件存储位置的是 _create_image 函数。

算法 1 虚拟机创建的函数调用过程

```

Nova.api.openstack.compute.servers.create()
nova.compute.api.create()
nova.compute.api._create_instance()
nova.scheduler.reapi.run_instance()
nova.scheduler.manager.run_instance()
nova.scheduler.multi.manager.run_instance()
nova.scheduler.filter_scheduler.schedule_run_instance()
nova.scheduler.filter_scheduler._provision_resource()
nova.scheduler.filter_scheduler.compute_rpcapi.
run_instance()

nova.computer.reapi.run_instance()
nova.compute.manage.run_instance()
nova.compute.manage._run_instance()
nova.compute.manage._run_instance()
nova.compute.manage._spawn()
nova.virt.driver.spawn()
nova.virt.libvirt.driver.spawn()
nova.virt.libvirt.driver._create_image()

```

因此, 通过修改 _create_image 函数可以支持不同的存储方式, 伪代码简单实现见算法 2, 其中, 当 stor_type=1 时, 创建本地硬盘的虚拟机; 当 stor_type=2 时, 创建共享存储的虚拟机。同样在虚拟机迁移或快照函数里, 也需要配置以下代码。如果云平台想继续增加对不同共享存储系统的支持, 只需简单增加 stor_type 参数的值即可。

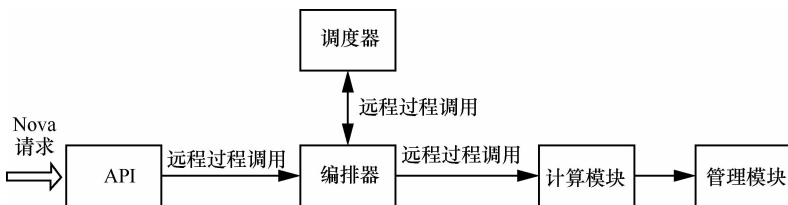


图3 不同存储方式的虚拟机创建流程



算法 2 伪代码简单实现

```
create_image ()
    if stor_type == 0
        instances_path = LocalStorage
    if stor_type == 1
        instances_path = SharedStorage1
    if stor_type == 2
        instances_path = SharedStorage2
    ...
```

本文充分考虑基于本地硬盘或共享存储的虚拟机的优缺点,通过分析和修改 OpenStack 源码,设计和实现了一种能够同时在本地硬盘或共享存储上创建虚拟机的云平台。前台通过指定不同参数,可以使后台创建相应存储形态的虚拟机,具体设计流程如图 4 所示。

具体创建步骤如下所述:

- 前台通过参数输入模块指定虚拟机创建的具体参数;
- 触发请求模块向后台系统发出虚拟机创建请求;
- 服务调度模块根据虚拟机创建的具体参数,将虚拟机创建请求调度到 1 台具体的物理服务器上;
- 虚拟机创建模块调用路径转换模块,根据前台指

定的参数,得到本地磁盘和共享存储对应的路径,然后调用资源配置模块,在物理服务器上配置硬件资源;

- 资源配置完成后,虚拟机创建模块在相应的存储路径上创建出虚拟机。

5 OpenStack 云平台验证环境

目前 OpenStack 云平台支持 GlusterFS、Ceph 等多种共享存储系统,本文通过搭建 OpenStack 实验室环境,采用 GlusterFS 分布式文件存储系统^[9],对以上想法进行了验证。如图 5 所示,nn20 为 OpenStack 云平台的一个计算节点,其中/var/lib/nova/instances_local 为本地硬盘的挂载点,/var/lib/nova/instances 为 GlusterFS 共享存储系统的挂载点,虚拟机自身的磁盘文件会分别存储在两个挂载点下的 private_dev_sdz 目录。

通过在虚拟机创建脚本中指定不同的 stor_type 值,如图 6 所示,就可以创建不同存储方式的虚拟机。

如图 7 所示,当 stor_type=1 时,虚拟机 test0-nn20 所在的目录 instance-00000971 位于/var/lib/nova/instances_local/private_dev_sdz/目录下,属于本地硬盘;当 stor_type=2 时,虚拟机 test1-nn20 所在的目录 instance-0000096f 位于/var/

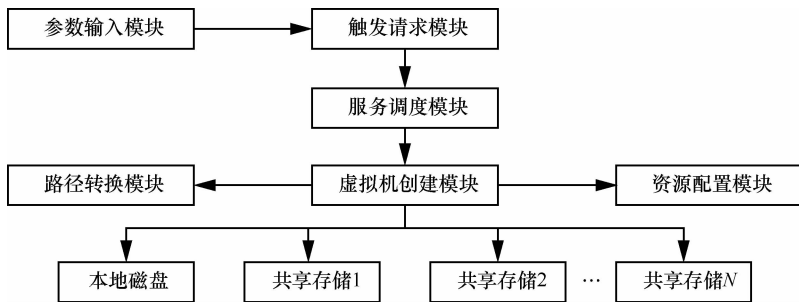


图 4 不同存储方式的虚拟机创建流程

```
[root@nn20 ~]# df -Th
```

文件系统	类型	容量	已用	可用	已用%%	挂载点
/dev/sda2	ext4	99G	8.3G	86G	9%	/
tmpfs	tmpfs	127G	0	127G	0%	/dev/shm
/dev/sda1	ext4	194M	32M	153M	18%	/boot
/dev/sda4	ext4	2.4T	7.9G	2.2T	1%	/var/lib/nova/instances_local
172.16.21.24:/gv0	fuse.glusterfs	32T	14T	16T	47%	/var/lib/nova/instances
172.16.21.28:/gv0	fuse.glusterfs	32T	14T	16T	47%	/var/lib/cinder/volumes/aa94baff1dde36d5f9325a53bf564188

图 5 OpenStack 云平台中本地硬盘和共享存储的挂载点

```
NET_ID=c2054cc4-fcc4-4ba4-a382-af1d642d8c05
POT_ID=$(quantum port-create --attached_phyport $IF $NET_ID|grep 'id'|awk 'print $4')
NAME="test0-nn20"
DIR=/private_dev_sdz
nova --debug boot --image $IMG_ID --flavor $FLR_ID --nic port-id=$POT_ID $NAME --hint stor_type= --availability-zone nova:nn20
```

图 6 OpenStack 云平台中的虚拟机创建脚本

```
[root@cc01 ~]# nova list --fields name,status,host,stor_type,instance_name|grep nn20
| 679b9c74-f324-4740-aff5-1cf31235de3d | test0-nn20 | ACTIVE | nn20 | 1 | instance-00000971 |
| 95a657fe-a01a-4e70-9469-63115e985d2e | test1-nn20 | ACTIVE | nn20 | 2 | instance-0000096f |
[root@cc01 ~]# ssh nn20
Last login: Thu Dec 24 17:43:49 2015 from cc01
[root@nn20 ~]# find / -name instance-00000971
/var/lib/nova/instances_local/private_dev_sdz/instance-00000971
[root@nn20 ~]# find / -name instance-0000096f
/var/lib/cinder/volumes/aa94baff1dde36d5f9325a53bf564188/private_dev_sdz/instance-0000096f
/var/lib/nova/instances/private_dev_sdz/instance-0000096f
```

图 7 OpenStack 云平台中虚拟机文件所在目录

```
[root@cc01 ~]# nova list --host=nn20
+-----+-----+-----+-----+
| ID | Name | Status | Networks |
+-----+-----+-----+-----+
| 679b9c74-f324-4740-aff5-1cf31235de3d | test0-nn20 | ACTIVE | IaaS Portal=192.168.27.8 |
| 95a657fe-a01a-4e70-9469-63115e985d2e | test1-nn20 | ACTIVE | IaaS Portal=192.168.27.7 |
+-----+-----+-----+-----+
[root@cc01 ~]# nova live-migration 679b9c74-f324-4740-aff5-1cf31235de3d nn22
ERROR: local stor_type forbidden live-migrate (HTTP 400) (Request-ID: req-c4d8e6d9-dbf5-4e43-a858-ec4970761dce)
[root@cc01 ~]# nova live-migration 95a657fe-a01a-4e70-9469-63115e985d2e nn22
[root@cc01 ~]# nova list --host=nn22
+-----+-----+-----+-----+
| ID | Name | Status | Networks |
+-----+-----+-----+-----+
| 95a657fe-a01a-4e70-9469-63115e985d2e | test1-nn20 | ACTIVE | IaaS Portal=192.168.27.7 |
+-----+-----+-----+-----+
```

图 8 不同存储方式的虚拟机迁移结果

lib/nova/instances/private_dev_sdz/目录下,属于 GlusterFS 共享存储。

最后 test0-nn20 和 test1-nn20 两台虚拟机分别向计算节点 nn22 进行实时迁移,结果如图 8 所示。对本地硬盘虚拟机 test0-nn20 执行 nova live-migration 命令后,系统提示不支持迁移,对共享存储虚拟机 test1-nn20 执行 Nova live-migration 命令后,则成功迁移到计算节点 nn22。

6 结束语

云平台通常采用本地硬盘或共享存储来存储虚拟机自身文件,两种存储方式各自存在优缺点。本文基于 OpenStack 开源软件,实现了云平台对两种存储方式的同時支持。用户通过指定不同参数,就可以创建相应存储方式的虚拟机,从而灵活满足业务的不同需求。本文能够促进现网系统向 OpenStack 云平台进行迁移,对业务平台云化具有重要的推动作用。

参考文献:

- [1] Open source cloud computing technology options [J]. Silicon Valley, 2012(1): 66.
- [2] OpenStack community website [EB/OL]. [2015-10-01]. <http://www.openstack.org/>.
- [3] Amazon EBS website[EB/OL]. [2015-10-01]. <https://aws.amazon.com/cn/ebs/>.
- [4] GlusterFS and Ceph performance test report [EB/OL]. [2015-10-01]. <http://www.openstack.cn/?p=2215>.
- [5] A OpenStack access request between the components of the call

process[EB/OL]. [2015-10-01]. <http://www.infoq.com/cn/articles/openstack-access-request-calling-process>.

- [6] Gluster community website[EB/OL]. [2015-10-01]. <http://www.gluster.org/>.

[作者简介]



冯伟斌(1982-),男,中国联合网络通信有限公司研究院工程师,主要研究方向为云计算、智慧城市、行业信息化等。



张云勇(1976-),男,博士后,中国联合网络通信有限公司研究院高级工程师,中国通信学会、电子学会、计算机学会高级会员,主要研究方向为下一代网络、网络融合、云计算、大数据等。曾作为主要研究人员主持和参与总装备部项目、教育部博士点基金项目、“863”计划、“973”计划、国家自然科学基金项目、国家重大专项项目。



王志军(1976-),男,中国联合网络通信有限公司研究院高级工程师,平台与产品实验室、云计算实验室主任,主要研究方向为业务平台及支撑系统体系架构、面向移动互联网的开放平台、新业务及下一代网络等。