Precentor 0.7:
开源云参考架构

关于本文：

本文介绍了 Precentor 项目，一个在英特尔平台上探索如何搭建基于 OpenStack 的 IaaS 云的项目。本文描述了我们基于 OpenStack Grizzly 版本的初步工作，其中包括软硬件的搭配和对扩展性及性能的调优与优化，以期为云平台的搭建提供参考。更多信息敬请登录 https://github.com/intel-cloud/Precentor。

由于受资源和知识的限制，目前的工作或多或少存在纰漏，很多地方有待改进。我们期待您的宝贵意见和建议，同时也欢迎更多的公司或者个人团体参与我们的工作，共同提高开源软件质量，一起加快生态环境的建设。

关于作者：

云架构技术实验室(CIT，Cloud Infrastructure Technology)是英特尔软件与服务部在位于上海的亚太研发中心的一个子部门，主要在英特尔平台上从事开源云技术的测试、分析、开发和优化。更多信息敬请 https://github.com/intel-cloud。

目录

简介

作为构建私有云的一项重要技术，OpenStack 在这些年里发展迅猛。然而搭建一个基于 OpenStack 的云平台却不是那么容易：在软件方面，OpenStack 云平台包含大量软件组件（大多是开源的），需要足够的专业知识去对软件进行选择、配置以及调优；在硬件方面，配置一个合理的硬件平台也需要掌握足够的软件和测试等专业知识。因此我们相信提供一个参考设计对大多数想要搭建自己的云平台的人而言会很有帮助。

Precentor 项目是英特尔平台上 OpenStack IaaS 云的参考设计方案，包含软硬件的选取、配置以及我们做出的优化，所有的内容都是开源的。

本文给展示了该项目的早期成果：Precentor 0.7，这是一个基于 OpenStack Grizzly 版和二十多台至强®双核服务器的一个参考架构。在这个版本中，我们关注的是核心组件的扩展性和性能问题，而不是整个产品的完整解决方案。由于资源和能力有限，当前版本或多或少存在纰漏，许多地方有待改进。我们会不断修改错误，优化系统方案，同时也欢迎各位宝贵的意见和建议。如果想了解联系方式、邮件列表、wiki 主页等信息，敬请登录 https://github.com/intel-cloud/Precentor。

硬件架构

设计目标

我们基本目标是搭建一个小规模的（1 到 2 个机架）并且配置合理的云环境。我们最终使用了 16 个计算节点，并且按比例添加其他辅助节点。

计算节点（Compute Node）

计算节点采用两路的双核 Xeon® E5-2670 服务器，其中配有 128 GB 内存和 6 块企业级 SATA 硬盘：1 块为系统盘，5 块组成 RAID 5 作为虚拟机本地存储。另外每个计算节点配备两块网卡，分别是作为管理网络的 1Gb 网卡和作为数据网络的 10Gb 网卡。

表 硬件配置汇总

名称	数量	硬件					
		Processor	1 Gb Port	10 Gb Port	O/S Disk	Data Disk (HDD*)	Data Disk (SSD**)
Compute Node	16	Dual processor Xeon® E5-2670	1	1	1 HDD	5 HDD in RAID5	0
Storage Node	4		1	1		20 HDD	4 SSD
API Node	1		2	0		0	0
Image Node	1		1	2		0	0
Network Node	1		1	2		0	0

* 7200 RPM 1TB enterprise grade SATA HDD

** Intel DC S3700

存储节点（Storage Node）

本参考架构的存储服务只包括 Cinder，而不包括 Swift。Cinder volume 服务由专门的存储集群组成，集群中每个存储节点采用双核 Xeon® E5-2670 服务器，配有 128GB 内存、1Gb 网卡、10Gb 网卡、20 块企业级 SATA 硬盘用以存储数据，以及 4 块英特尔 DC S3700 SSD 用以存储 Ceph 日志（系统盘未计算在内）。

辅助节点（Supporting Node）

辅助节点包括一个 API Node 用以提供除 Glance API 外的其他所有面向用户的 API 服务；一个 Image Node 用以提供虚拟机镜像服务（Glance API 亦在该节点上）；还有一个 Network Node 用以提供通向云外的 L3 路由。

网络

本参考架构中有三种类型的网络。首先是管理网络(Management Network)，每个节点都有一个 1Gb 的接口连接到该网络。

其次是数据网络(Data Network)。这是一个纯 10 Gb 网络，用作虚拟机之间的通信、虚拟机与云外的通信、镜像节点与计算节点之间的镜像数据传输、以及计算节点访问 volume 之用。

第三个是外部网络(External Network)。API 节点有一个 1Gb 连接到这个网络，另外 Image 节点和 Network 节点由于需要较高的外部带宽因此各有一个 10Gb 的连接到这个网络。

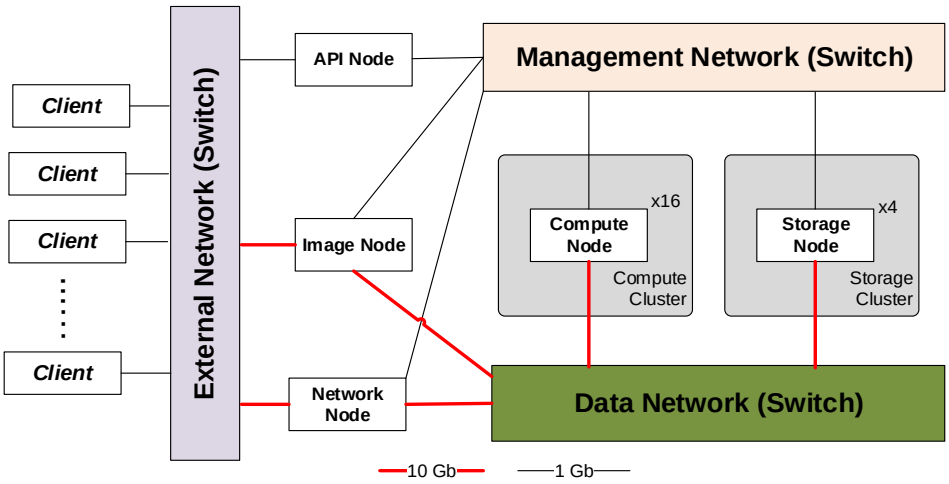

图　物理拓扑结构

小结

图 1 说明了整个参考方案的物理拓扑结构。

软件架构

软件组件

表 2 列举了参考方案中所有关键组件。本节的其他部分将会说明每种类型的节点的软件配置。

表　软件组件列表

Base O/S	CentOS 6.4
OpenStack	2013.1.2 (the 2nd Grizzly update release)
General: Database	mysql-server-5.1.69-1
General: AMQP Broker	qpid-cpp-server-0.14-22
Compute: Hypervisor	KVM
Compute: Host Kernel	2.6.32-358.111.1.openstack (with ns enabled)
Compute: QEMU	qemu-kvm-0.12.1.2-2.355 (with Ceph RBD support enabled)
Compute: libvirt	libvirt-0.10.2-18
Storage: Linux Kernel	3.8.13 (upstream)
Storage: Ceph	Ceph 0.61
Storage: Ceph librbd	librbd1-0.61
Storage: Ceph librados	librados2-0.61
Network: Open vSwitch	openvswitch-1.10.0-1
Keystone Token Store	memcached-1.4.4-3

基本环境

作为 Red Hat EL6.4 的克隆，CentOS 在 IPDC 数据中心有着广泛的应用，因此我们使用了 CentOS6.4 作为所有节点的操作系统。在 OpenStack 的版本方面我们选择了 2013.1.2，即 Grizzly 的第二个修复版。Red Hat 的 RDO 是一个对 upstream OpenStack 的简单封装，可以看成等同于 upstream，所以我们采用了 RDO 作为安装源。

计算节点

所有的计算节点采用 KVM 作为 hypervisor。另外，CentOS6.4 缺省的内核不支持 namespace，所以需要进行升级。CentOS 中缺省的 QEMU 也不支持 Ceph RBD，所以也需要另外安装。

以下服务运行在计算节点上：

- nova-compute
- quantum-openvswitch-agent
- openvswitch
- libvirtd

另外，在每个节点上我们用了 5 块硬盘做成一个 RAID5 来存放 Nova instance 镜像。

存储节点

我们使用 Ceph 作为 Volume Service (Cinder)和 Image Service (Glance)的统一存储后端。在本参考架构中有着一个 4 个节点的 Ceph 集群，运行着以下服务：

- cinder-volume (只在第一个节点上)
- ceph-mon (只在第一个节点上)
- ceph-osd (所有节点)

这个集群中所有服务器的 10Gb 的网络接口被配置到一个专门的子网下。所有 Ceph 服务之间的通信以及 Ceph 与客户端之间的通信都是在这个子网上完成，用 Ceph 的术语来讲就是 Public 网络和 Cluster 网络是同一个网络。与此相应的是在每一个计算节点的 10Gb 网口都需要配置一个同一子网下的 IP 地址，因为只有这样计算节点才能访问 Ceph 集群里的 volume。

Ceph 集群中的每个存储节点配置有 20 个 Ceph OSD daemon 来分别管理 20 块数据硬盘(HDD)。每块数据硬盘被格式化成 XFS。4 块 SSD 被划分成 20 个分区作为 20 块数据硬盘的 journal。

网络节点

网络节点上运行着除 quantum-api 之外的所有其它 quantum 服务，quantum-api 运行在 API 节点上。

- quantum-dhcp-agent
- quantum-l3-agent
- quantum-openvswitch-agent
- openvswitch

镜像节点

镜像节点上运行着 glance-api 和 glance-registry 两个服务。Ceph RBD 被用作镜像的存储后端，因此该节点的 10Gb 网络接口需要配置成存储集群子网的 IP 地址。

- glance-api
- glance-registry

API 节点

API 节点运行除 glance-api 外的所有面向用户的 API 服务。它也负责运行 Dashboard 和其他辅助服务，包括数据库和消息队列服务。

值得一提的是出于性能的考虑 Keystone 服务使用了 memcached 作为 token 的存储后端。更多的调优细节请参考下一章节。

- nova-api
- nova-cert
- nova-scheduler
- nova-conductor
- keystone
- quantum-server
- cinder-api
- cinder-scheduler
- dashboard
- mysqld
- httpd
- qpidd
- memcached

调优与优化

MySQL 基本调优

缺省的 MySQL 配置是不足以处理大负载的。尽管 MySQL 调优可以很复杂，但一些很小的调整就能带来较明显的性能改进。在次我们建议至少做一处改动：把 MySQL 的引擎切换到 innodb，并把缓冲池大小调整到较大的数值，比如 2GB。

```
[mysqld]
default-storage-engine=innodb
innodb_buffer_pool_size = 2G
```

使用 memcached 作为 Keystone 的 token 存储后端

默认情况下，Keystone 会将所有的信息存储在数据库中，其中也包含 token 信息。然而由于 Keystone 从来不会移除失效的 token，token 表会随着时间的推移而变大。更糟糕的是，token 表并没有建立索引，这会导致查询操作随着表变大而明显变慢(此 bug 在 Havana 版本已解决)。作为解决办法，我们使用 memcached 做为 Keystone 的 token 存储后端，memcached 的 TTL 特性会自动移除所有失效的 token，从而提高 token 查询效率。Keystone 的 memcached 的配置如下所示：

```
[token]
driver = keystone.token.backends.memcache.Token

[memcache]
servers = 127.0.0.1:11211
```

禁用 rootwrap 脚本

rootwrap 脚本允许 OpenStack 服务对需要 root 权限的命令进行细粒度的控制。然而由于 Python 版本实现的问题，它会带来明显的性能开销。禁用 rootwrap 之后，我们发现 OpenStack 的某些操作性能提高了 5 倍以上。因此，我们建议最好禁用 Quantum 服务的 rootwrap，这里需要改动的地方有两处：

1. 在 `quantum.conf` 设置 `root_helper=sudo`
2. 添加 `quantum` user 到`/etc/sudoers`

增大 quantum-server 的 SQLAlchemy QueuePool

quantum-server 的 SQLAlchemy QueuePool 默认大小为 5。由于这个值太小，系统在处于较大压力状态下时会出现一些错误。我们建议在 ovs_quantum_plugin.ini 中将此值增大到 60：

```
sqlalchemy_pool_size=60
```

但是上述的配置项在 Grizzly 中还不存在，相关的补丁 24986 已经合并到了 Havana，为了实现这个调优，我们需要将它移植到 Grizzly 版本中。

增大 Quantum agent 状态报告时间间隔和超时间隔

Quantum agent 以特定的时间间隔向 server 报告他们的状态，如果 Quantum-server 在指定的时间内没有收到一个 agent 的状态报告信息，它会认为此 agent 已经 DOWN。我们发现由于某些原因来自 agent 的状态报告消息会被阻塞并超时，从而导致 agent 的状态被错误地被标记为 DOWN。我们通过增加状态报告的时间间隔来减少 agent 和 quantum 的通信流量，同时增加超时时间来防止 agent 被错误标记成 DOWN 状态。`/etc/quantum/quantum.conf` 的变更如下：

```
[DEFAULT]
agent_down_time = 300
[AGENT]
report_interval = 40
```

增大 dhcp-lease-max

dpch-lease-max 的默认值为 150，这会限制所能使用 IP 的最大数目，建议将其调整为足够大的数值。

/etc/quantum/quantum.conf ：

```
dnsmasq_config_file= /etc/dnsmasq.conf
```

/etc/dnsmasq.conf：

```
dhcp-lease-max=1000
```

Ceph 调优

Ceph 核心的配置如下所示：

```
[osd]
osd mkfs type = xfs
osd mount options xfs = rw,noatime,inode64,logbsize=256k,delaylog
osd mkfs options xfs = -f -i size=2048
filestore max inline xattr size = 254
filestore max inline xattrs = 6
osd_op_threads=20
filestore_queue_max_ops=500
filestore_queue_committing_max_ops=5000
journal_max_write_entries=1000
journal_queue_max_ops=3000
objecter_inflight_ops=10240
filestore_queue_max_bytes=1048576000
filestore_queue_committing_max_bytes=1048576000
journal_max_write_bytes=1048576000
journal_queue_max_bytes=1048576000
ms_dispatch_throttle_bytes=1048576000
objecter_infilght_op_bytes=1048576000
filestore_max_sync_interval=10
filestore_flusher=false
filestore_flush_min=0
filestore_sync_flush=true
```

Ceph OSD 的存储节点上的 Linux 调优：

```
echo "2048" > /sys/block/${device}/queue/read_ahead_kb
```

计算节点 Ceph.conf 的修改：

```
objecter_infilght_op_bytes=1048576000
objecter_inflight_ops=10240
```

展望

到目前为止我们的工作还只是 Precentor 项目的一个开端，我们承诺会继续构建、优化和分享 Intel 平台上的 OpenStack 云参考架构。下一阶段我们需要解决目前发现的若干问题，并且一如既往的关注扩展性和性能问题。

我们也会保持跟进最新的 OpenStack 软件和 Intel 硬件平台。下一个版本的 Precentor 参考架构将会基于 OpenStac Havana 版，我们预计很多问题将会被解决同时新的问题又会出现。Intel 的新一代产品也会被集成进参考架构。

再次说明，Precentor 是研究项目，旨在和社区一起改善开源软件的质量，加快整个生态环境的建设；我们期待您的意见和建议，也欢迎更多的公司和个人团体参与我们的工作。

声明：

英特尔公司 2013 年版权所有。所有权保留。

英特尔、英特尔图标、英特尔 Atom、英特尔处理器和英特尔 Xeon 是英特尔在美国和/或其他国家的商标。*其他的名称和品牌可能是其他所有者的资产。

本结果基于英特尔内部分析估算得出，仅作参考之用。任何系统硬件、软件的设计或配置的不同均可能影响实际性能。

英特尔®超线程技术支持部分英特尔®酷睿处理器。此项技术需要安装兼容英特尔®超线程技术的系统，请与您的电脑生产商核实。系统运行性能将取决于具体的硬件与软件环境。英特尔®酷睿 i5-750 处理器暂不支持此项技术。更多信息（包括哪些处理器支持超线程技术）敬请登陆 http://www.intel.com/info/hyperthreading。

对本文件中包含的软件源代码的提供均依据相关软件许可而做出，任何对该等源代码的使用和复制均应按照相关软件许可的条款执行。

本文件中包含关于英特尔产品的信息。本文件不构成对任何知识产权的授权，包括明示的、暗示的，也无论是基于禁止反言的原则或其他。除英特尔产品销售的条款和条件规定的责任外，英特尔不承担任何其他责任。英特尔在此作出免责声明：本文件不构成英特尔关于其产品的使用和/或销售的任何明示或暗示的保证，包括不就其产品的(i)对某一特定用途的适用性、(ii)适销性以及(iii)对任何专利、版权或其他知识产权的侵害的承担任何责任或作出任何担保。

“临界任务应用”是任何因英特尔产品故障而直接或间接造成人身伤亡的应用。如果你在任何此类应用中购买或使用了英特尔产品，你有权向英特尔及其附属公司、分包商和供应商、董事、高级员工索取此类应用程序造成的个人损失赔偿，其中包括诉讼费、律师费和伤亡费等，不管英特尔或者其附属部门有没有疏忽该产品或者产品部件的设计、生产或者警告。

英特尔有权随时更改产品的规格和描述而毋需发出通知。设计者不应信赖任何英特产品所不具有的特性，设计者亦不应信赖任何标有“保留权利”或“未定义”说明或特性描述。对此，英特尔保留将来对其进行定义的权利，同时，英特尔不应为因其日后更改该等说明或特性描述而产生的冲突和不相容承担任何责任。此处提供的信息可随时改变而毋需通知。请勿根据本文件提供的信息完成一项产品设计。

本文件所描述的产品可能包含使其与宣称的规格不符的设计缺陷或失误。这些缺陷或失误已收录于勘误表中，可索取获得。

在发出订单之前，请联系当地的英特尔营业部或分销商以获取最新的产品规格。