

**联想LOE企业版**

**V4.3**

**运维手册**

**目 录**

[1. 概要 5](#_Toc1138460)

[2. LOE云平台术语 5](#_Toc1138461)

[3. 平台状态检查 9](#_Toc1138462)

[3.1. 硬件状态检查 9](#_Toc1138463)

[3.1.1. 系统状态检查 9](#_Toc1138464)

[3.1.2. CPU状态检查 9](#_Toc1138465)

[3.1.3. 内存状态检查 10](#_Toc1138466)

[3.1.4. 网络状态检查 11](#_Toc1138467)

[3.1.5. 磁盘状态检查 11](#_Toc1138468)

[3.2. 服务状态检查 14](#_Toc1138469)

[3.2.1. mysql集群状态检查 14](#_Toc1138470)

[3.2.2. rabbitmq集群状态 15](#_Toc1138471)

[3.2.3. nova组件状态检查 15](#_Toc1138472)

[3.2.4. neutron状态检查 16](#_Toc1138473)

[3.2.5. cinder状态检查 17](#_Toc1138474)

[3.2.6. ceph状态检查 17](#_Toc1138475)

[3.2.7. pacemaker管理的各服务集群状态检查 18](#_Toc1138476)

[3.2.8. l3-agent的namespace 19](#_Toc1138477)

[3.2.9. 查看dhcp-agent的namespace 19](#_Toc1138478)

[4. 常见故障处理 20](#_Toc1138479)

[4.1. 通用故障 20](#_Toc1138480)

[4.1.1. 控制节点宕机 20](#_Toc1138481)

[4.1.2. 操作系统盘损坏 21](#_Toc1138482)

[4.1.3. 服务状态异常 21](#_Toc1138483)

[4.1.3.1. Rabbitmq-server 21](#_Toc1138484)

[4.1.3.2. hwmgmt-api 21](#_Toc1138485)

[4.1.3.3. collectd 21](#_Toc1138486)

[4.1.4. ntp server不工作 22](#_Toc1138487)

[4.2. 计算 22](#_Toc1138488)

[4.2.1. 虚拟机状态错误 22](#_Toc1138489)

[4.2.2. 同一个vm运行在多个宿主机上 23](#_Toc1138490)

[4.2.3. vm（Linux系统）用户自己升级失败导致系统无法启动 23](#_Toc1138491)

[4.3. 存储 24](#_Toc1138492)

[4.3.1. ceph osd down 24](#_Toc1138493)

[4.3.2. ceph 'error removing image' 25](#_Toc1138494)

[4.3.3. mon.node-x store is getting too big! 25](#_Toc1138495)

[4.3.4. Full osd或near full osd 25](#_Toc1138496)

[4.4. 网络 27](#_Toc1138497)

[4.4.1. 虚机ping 不通，无法远程登录 27](#_Toc1138498)

[5. 系统定制Tips 27](#_Toc1138499)

[5.1. 修改配置以启/停相关功能 27](#_Toc1138500)

[5.1.1. 强制计费功能失效 27](#_Toc1138501)

[5.1.2. 开启包月包年模式 27](#_Toc1138502)

[5.1.3. 支付宝帐号修改 28](#_Toc1138503)

[5.1.4. 激活邀请码充值 28](#_Toc1138504)

[5.1.5. 强制申请与审批功能失效 28](#_Toc1138505)

[5.1.6. 配置Host HA功能 28](#_Toc1138506)

[5.1.7. 配置VM HA功能 33](#_Toc1138507)

[5.1.8. 启用物理网络拓扑 34](#_Toc1138508)

[5.1.9. 修改云硬盘1TB上限 35](#_Toc1138509)

[5.1.10. 启用flat网络 36](#_Toc1138510)

[5.1.11. 允许域账户登录 37](#_Toc1138511)

[5.1.12. 开启Ceph存储空间回收通知机制 37](#_Toc1138512)

[5.2. 系统优化选项 38](#_Toc1138513)

[5.2.1. 为10GB网络修改MTU到9000 38](#_Toc1138514)

[5.2.2. 调整Ceph pg\_num 39](#_Toc1138515)

[5.3. 系统运维常用操作 41](#_Toc1138516)

[5.3.1. 安全开关云平台 41](#_Toc1138517)

[5.3.2. 通过命令行上传大镜像文件 43](#_Toc1138518)

[5.3.3. 替换Ceph OSD数据盘 43](#_Toc1138519)

[5.4. 更新LOE许可证 45](#_Toc1138520)

# 概要

联想LOE企业版（简称LOE,后续将以LOE代替）是用户按需分配计算／存储／网络等资源的弹性云计算平台，用户只需通过浏览器轻点鼠标，就可以快速完成整个虚拟数据中心的搭建。LOE云平台上的各类资源可灵活扩展或缩减，云主机秒级启动，云存储稳定可靠，秒级备份，网络隔离及防火墙确保云主机安全，对各类资源秒级计费，可灵活调整计价策略，所有数据采用多副本保证高可靠性，分钟级别的监控告警为运维人员及时全面掌握系统状态提供有力保障，分层的用户管理方便企业级管理员对各级权限的管控。

本手册将将对LOE日常运维操作进行详细的介绍，以便您和您的团队更好地管理基于LOE的云平台。

# LOE云平台术语

为了方便理解文中的内容，本节将介绍LOE云平台涉及到的术语及概念。

* **用户**

LOE云平台提供四个级别的用户，每个级别具有不同的权限，适应企业及用户管理。权限由高到低：云管理员->企业管理员->项目管理员->普通用户。所有用户在整个LOE云平台内唯一、统一使用E-mail地址登录。

1. 云管理员：云管理员具有管理整个云平台的权限，在【权限管理】中，可以管理企业／项目／用户，并且具有【全局管理】菜单，管理整个平台的各项资源／计费定价等。系统内置帐户admin@example.org为云管理员帐号。其它云管理员帐户由云管理员创建。
2. 企业管理员：云平台针对每个企业计费。企业管理员由云管理员创建。除了具有普通用户的所有权限外，企业管理员在【权限管理】中，可以管理项目／用户，并查看费用使用情况。
3. 项目管理员：具有管理当前项目配额的功能。项目管理员由企业管理员创建。除了具有普通用户的所有权限外，企业管理员在【权限管理】中，可以管理用户。
4. 普通用户：具有操作所属项目资源的权限。普通用户可以被云管理员或企业管理员创建。用户通过登录后，可以操作所属项目的各项资源，如云主机／云硬盘等。

* **云主机**

运行在LOE云平台上的虚拟机，相当于数据中心的一台物理服务器。用户可以通过选择合适的CPU／内存／操作系统磁盘空间，网络，安全组等配置创建云主机。系统根据云主机使用的配置和时长计费。

* **对象存储**

用户创建对象存储，用于保存大规模的非结构化数据，例如图片文件、视频文件、虚拟机镜像文件、文档文件等。用户对属于自己权限的数据可以进行检索、下载、复制、删除、更新等操作，同时支持显示对象大小、访问地址、散列值等信息。

* **云硬盘**

为云主机提供块级存储设备，相当于一台物理机的硬盘。云硬盘是独立的资源，它的生命周期独立于云主机，可以被挂载到任何云主机上，也可以从云主机卸载，然后挂接到其他云主机。云平台根据云硬盘的类型／容量／使用时长计费。

* **镜像**

操作系统的安装模版，用户可以选择合适的操作系统镜像创建所需要的云主机。只有admin用户具有上传镜像操作权限，其他权限的用户只能使用和查看。但用户可以通过云主机快照创建新的镜像，并在启动云主机时选择“云主机快照”类型来使用新的镜像。

* **快照**

用户可以对云主机和云硬盘创建快照，保存当时状态下的云主机和云硬盘数据。云主机快照会被保存为镜像，用户可以基于这个镜像创建新的云主机。云硬盘快照保存当时状态下的硬盘数据，并可以基于快照创建新的云硬盘。云平台根据快照的数量和使用时长计费。

* **安全组**

一系列防火墙规则组成安全组，创建云主机时，用户可以选择合适的安全组来保障云主机的安全。安全组对主机上的所有网卡生效，新增网卡也将应用已有的安全组。

* **公网IP**

独立的IP地址资源，用户可以将申请的公网IP绑定到自己的云主机上，客户就可以通过这个公网IP来访问云主机提供的服务了。公网IP也绑定到路由器上，帮助内部网络连接外网或网络之间连接。云平台根据公网IP的个数／带宽／使用时长计费。

* **SSH密钥对**

基于密钥的安全验证登录方法，保证云主机安全。LOE推荐使用密钥对登录云主机。

* **网络**

网络与现实世界的交换机／路由器／服务器／连线组成的基础设施网络类似，创建网络后，用户可以在网络内创建子网，创建云主机时选择网络，组建服务器集群。

LOE提供的基础网络包含共享网络（share\_net）和外部网络（public\_net），创建在共享网络上的云主机处于同一个网络内，通过安全组保障云主机访问安全。外部网络主要用于公网IP地址的分配。

用户可以为项目创建内部网络，并在内部网络中创建子网。如同在物理网络上通过交换机将服务器连接到一起的局域网，服务器通过交换机连接到子网中。不同的内部网络之间是完全隔离的，因此不同的网络中可以配置相同的IP地址而不会产生冲突。同一个网络内可以创建多个子网，以适应业务的需求。

* **路由器**

用户创建路由器，为不同的子网提供三层路由，从而让子网内的云主机与其他子网的云主机互联互通。也可以将用户创建的内部网络连接到外部网络，让内部网络的云主机访问internet。路由器配置公网IP后，还可以为内网的云主机做端口转发，以节约公网IP地址资源。

* **负载均衡**

用户创建负载均衡，能够将所收到的网络流量分配给若干个提供相同处理功能的虚拟机，并按照特定的算法保证每台虚拟机工作在最优的负载状态，从而达到更高效的使用计算资源的目的。这些虚拟机构成了一个集群，负载均衡会为集群设置一个对外提供服务的地址Virtual IP，外部用户通过Virtual IP实现对集群的访问。负载均衡必须能够连接外部网络，如果不关联Floating IP，平台将会随机为负载均衡分配一个公网地址。

* **防火墙**

防火墙提供网络间的访问控制功能，通过防火墙策略中的过滤规则对当前项目中的网络流量进行过滤。防火墙必须与一个防火墙策略相关联，防火墙策略是防火墙规则的集合，防火墙规则支持多种网络协议。每个项目只允许配置一个防火墙。

* **虚拟专用网**

虚拟专用网是在云网络上建立一个临时的、安全、稳定的连接。虚拟专用网是对云网的扩展，可以帮助不同租户间建立可信的安全连接，并保证数据的安全传输。

* **网络拓扑**

展示用户当前所在项目的网络结构图。点击各个设备可以展示详细配置。

* **告警**

用户对资源（云主机／云硬盘等）的监控数据设置告警条件，当监控数据达到阈值就会发送告警到通知列表中的邮件。

* **企业**

企业通过云管理员直接创建生成。企业创建成功后，LOE自动为企业创建企业账户用户计费，创建企业管理员用于企业管理。LOE对企业账户进行计费，企业管理员管理项目和用户，可以查看其管理范围内项目的消费清单。在其管理范围内的项目名称唯一，但可以与其他企业用户管理的项目名称重合。

* **项目**

项目是定义资源所有权的基本单元，所有资源（如云主机等）都要隶属于某个项目中。项目必须隶属于一个企业。项目名称在单个企业的管理范围内是唯一的，但在整个云平台中可以不唯一。

* **邀请码**

云管理员创建／管理邀请码。其它用户可通过邀请码给帐户充值。

# 平台状态检查

## 硬件状态检查

硬件状态检查主要关注服务器的各硬件组件的可用状态，如磁盘、主板、CPU、内存、网卡、RAID卡、交换机等等。

检查方式分为现场巡检与硬件监控。现场巡检主要检查磁盘状态等是否变黄，如果变黄表明磁盘损坏，需尽快进行更换。服务器前面板状态等是否变黄，如果变黄，表明主板或CPU等组件存在硬件问题。需尽快联系技术支持进行更换。网卡是否正常闪亮，交换机端口是否正常闪亮。等等。

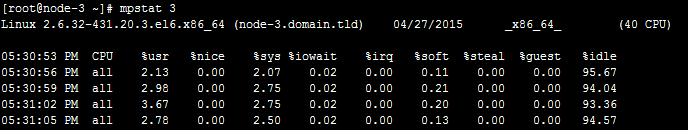
也可登录各硬件厂商提供的远程管理端口来查看各硬件组件的状态。根据厂商不同，管理与访问方式也有所不同。详细内容，请参考厂商提供的硬件管理文档。

### 系统状态检查

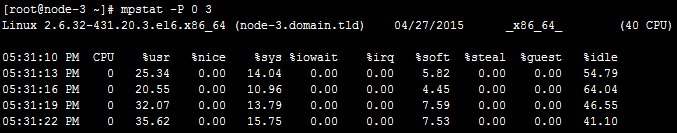
系统状态检查，主要关注操作系统的主要计算资源。如CPU，内存，磁盘，网络等等。

### CPU状态检查

通过运行mpstat 3, 命令来查看系统所有CPU加权平均的资源使用状态。其中3表示采样时间间隔为3秒。详细内容请参考下图。

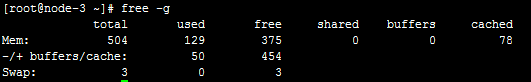


如果想要查看具体的CPU Core的使用状态，可以使用参数-P， 后跟要查看的CPU核编号。如0， 表示CPU的一个Core。详细内容请参考下图。

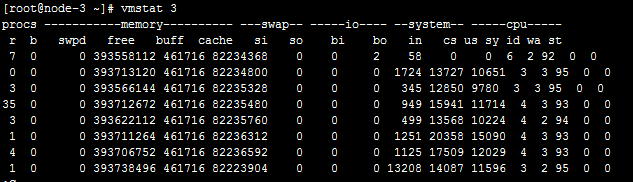


### 内存状态检查

通过运行free –g命令可以查看系统的内存与SWAP使用率。



通过运行vmstat 3命令，可以查看系统的内存与SWAP状态。3为采样时间间隔，单位为秒。如果系统中的swpd 值比较大，而free值较小，说明物理内存已经不够，需要添加内存或者采取其他资源释放的操作。

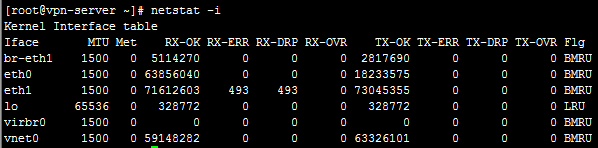


可以结合vmstat命令来判断系统是否繁忙，其中：

|  |  |  |
| --- | --- | --- |
| 项 | 值 | 说明 |
| proc | r | 等待运行的进程数。一般负载超过了3就比较高，超过了5就高，超过了10就不正常了，服务器的状态很危险。 |
| b | 处在非中断睡眠状态的进程数 |
| memeory | swpd | 虚拟内存使用情况，单位为KB。如果大于0，表示你的机器物理内存不足了 |
| free | 空闲的内存，单位为KB |
| buff | Linux系统用来存储目录索引/权限等的缓存，单位为KB |
| cache | 被用来作为缓存的内存数，如已打开文件的内容,单位为KB |
| swap | si | 从磁盘交换到内存的交换页数量，单位为KB。如果大于0，表示你的机器物理内存不足了 |
| so | 从内存交换到磁盘的交换页数量，单位为KB。如果大于0，表示你的机器物理内存不足了 |
| io | bi | 块设备每秒接收的块数，单位为KB。正常情况应该接近0，否则表明IO过于频繁 |
| bo | 块设备每秒发送的块数，单位为KB。正常情况应该接近0，否则表明IO过于频繁 |
| system | in | 每秒的中断数，包括时钟中断 |
| cs | 每秒的环境切换次数，如系统函数调用，线程切换等。这个值要越小越好，太大了，要考虑调低线程或者进程的数目 |
| cpu  (百分比) | us | 用户空间cpu使用时间 |
| sy | 内核空间cpu使用时间 |
| id | 空闲cpu时间。一般地，id,sy和us之和为100 |
| wt | 等待IO cpu时间 |

### 网络状态检查

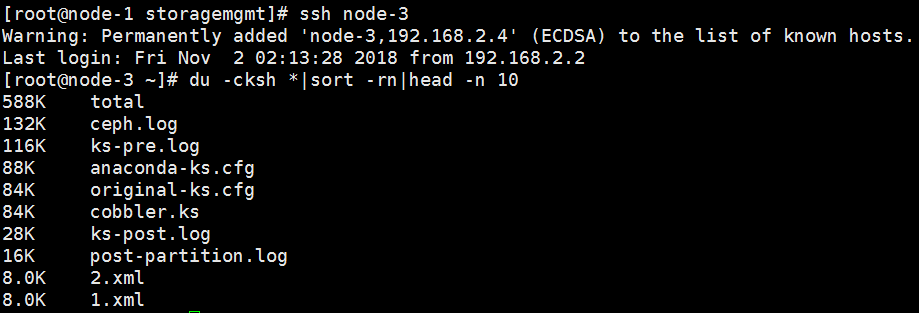
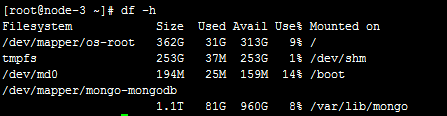
通过运行netstat –i 命令，可以查看每个网络接口上的接收与发送的数据包数。并且可以查看MTU值与有错误或者丢弃的包数，有助于网络故障的诊断。



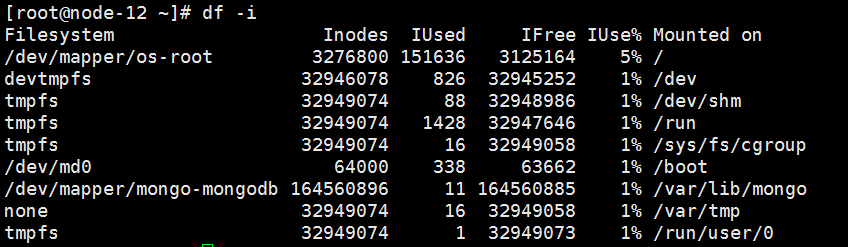
### 磁盘状态检查

通过运行df –h 命令，可以查看系统中的磁盘使用率。如果系统中磁盘的使用率超过80%，需要采取进一步的操作，如添加新磁盘或者删除不需要的大文件等等。如发现某个分区空间接近用完，可以进入该分区的挂载点，用以下命令找出占用空间最多的文件或目录，然后按照从大到小的顺序，找出系统中占用最多空间的前10个文件或目录：

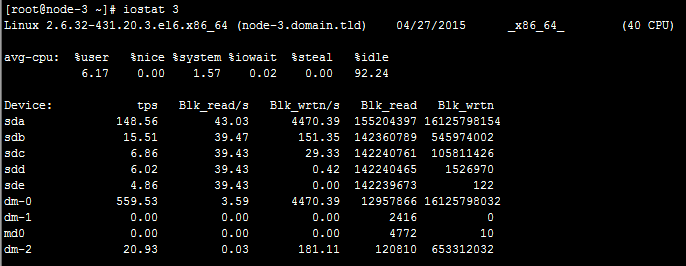
du -cksh \*|sort -rn|head -n 10



通过运行df –i 命令，可以查看系统中的inode使用率。每个文件都对应一个inode。系统中的inode数是有限的，如果系统中inode的使用率超过80%，需要采取进一步的操作，如添加新磁盘或者删除不需要的文件等等。可以通过命令find /var/log -size -5k|wc –l看小文件数，删除掉无用的小文件。



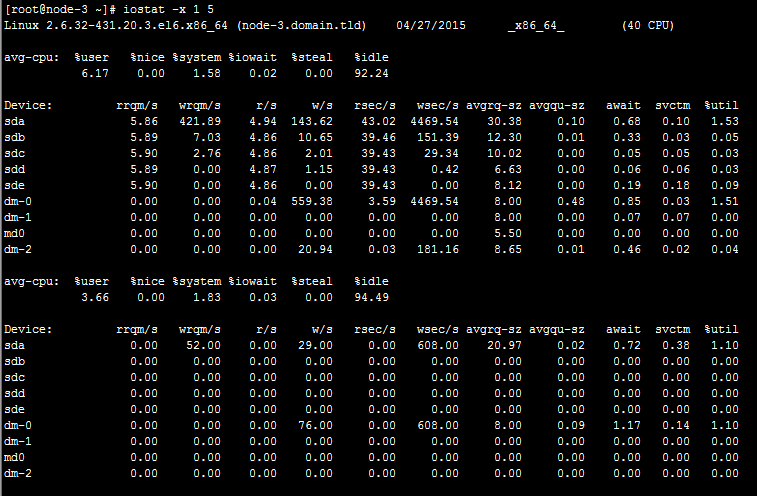
通过运行iostat 3 命令，可以查看磁盘的IO情况。其中3为采样时间间隔，单位为秒。



通过运行iostat –x 1 5命令来查看系统中是否存在IO性能瓶颈。iostat是含在套装systat中的,可以用yum -y install systat来安装。

常关注的参数：

如果%util接近100%,说明产生的I/O请求太多，I/O系统已经满负荷，该磁盘可能存在瓶颈。如果idle小于70%，I/O的压力就比较大了，说明读取进程中有较多的wait。



## 服务状态检查

可以通过systemctl list-units --all --type=services |grep KEYWORDS查看系统关键服务状态,系统关键服务前缀如下:

ceph-osd

ceph-mon

collectd

hwmgmt

httpd

libvirtd

openvswitch

rabbitmq-server

ntpd

neutron

openstack

若有服务状态异常, 需要进一步根据其日志检查错误并修复.

### mysql集群状态检查

#确保所有mysql集群都在

[root@node-1 ~]# mysql -e "show status" | grep wsrep\_incoming\_addresses

wsrep\_incoming\_addresses 192.168.0.3:3307,192.168.0.2:3307,192.168.0.4:3307

#确保mysql是同步的

[root@node-1 ~]# mysql -e "show status" | grep wsrep\_local\_state\_comment

wsrep\_local\_state\_comment Synced

#确保wsrep同步完成

[root@node-1 ~]# mysql -e "show status" | grep wsrep\_ready

wsrep\_ready ON

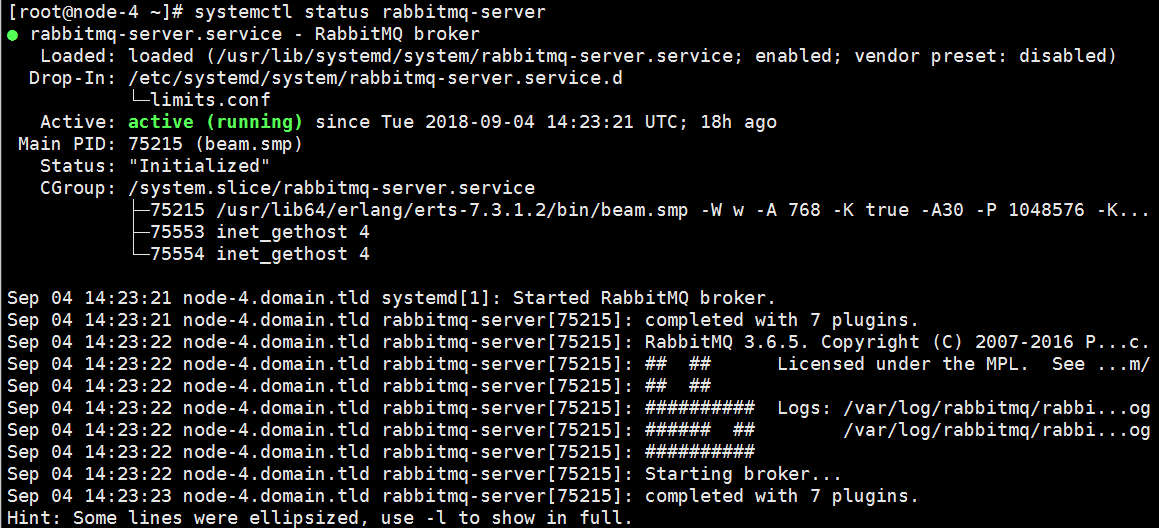
！！！禁止手动重起mysql集群，比如用命令

crm resource restart clone\_p\_mysql

mysql集群要求必须至少有一个活的节点，所以不能将mysql集群全部关闭，如果要关闭，要做相应设置才能恢复。

### rabbitmq集群状态

#确rabbitmq-server服务正常



#确保所有的节点都在running

[root@node-1 ~]# rabbitmqctl cluster\_status

Cluster status of node 'rabbit@node-1' ...

[{nodes,[{disc,['rabbit@node-1','rabbit@node-2','rabbit@node-3']}]},

{running\_nodes,['rabbit@node-2','rabbit@node-3','rabbit@node-1']},

{partitions,[]}]

...done.

#确保所有的节点能list queues

[root@node-1 ~]# rabbitmqctl list\_queues

！不建议手动重起rabbitmq. 若rabbitmq状态异常且无其它方法使其恢复, 可在所有控制节点上尝试以下方法之一:

1. systemctl restart rabbitmq-server
2. rabbitmqctl stop\_app && rabbitmqctl start\_app

### nova组件状态检查

[root@node-1 ~]# nova service-list

+----+------------------+-------------------+----------+---------+-------+----------------------------+-----------------+

| Id | Binary | Host | Zone | Status | State | Updated\_at | Disabled Reason |

+----+------------------+-------------------+----------+---------+-------+----------------------------+-----------------+

| 1 | nova-consoleauth | node-1.domain.tld | internal | enabled | up | 2014-12-02T11:23:07.000000 | - |

| 2 | nova-scheduler | node-1.domain.tld | internal | enabled | up | 2014-12-02T11:23:07.000000 | - |

| 3 | nova-conductor | node-1.domain.tld | internal | enabled | up | 2014-12-02T11:23:05.000000 | - |

| 4 | nova-cert | node-1.domain.tld | internal | enabled | up | 2014-12-02T11:23:05.000000 | - |

| 5 | nova-conductor | node-3.domain.tld | internal | enabled | up | 2014-12-02T11:23:05.000000 | - |

| 8 | nova-consoleauth | node-3.domain.tld | internal | enabled | up | 2014-12-02T11:23:10.000000 | - |

| 11 | nova-scheduler | node-3.domain.tld | internal | enabled | up | 2014-12-02T11:23:10.000000 | - |

| 14 | nova-cert | node-3.domain.tld | internal | enabled | up | 2014-12-02T11:23:08.000000 | - |

| 18 | nova-consoleauth | node-2.domain.tld | internal | enabled | up | 2014-12-02T11:23:08.000000 | - |

| 21 | nova-scheduler | node-2.domain.tld | internal | enabled | up | 2014-12-02T11:23:07.000000 | - |

| 24 | nova-conductor | node-2.domain.tld | internal | enabled | up | 2014-12-02T11:23:07.000000 | - |

| 27 | nova-cert | node-2.domain.tld | internal | enabled | up | 2014-12-02T11:23:09.000000 | - |

| 30 | nova-compute | node-4.domain.tld | nova | enabled | up | 2014-12-02T11:23:03.000000 | - |

| 33 | nova-compute | node-5.domain.tld | nova | enabled | up | 2014-12-02T11:23:03.000000 | - |

+----+------------------+-------------------+----------+---------+-------+----------------------------+-----------------+

### neutron状态检查

[root@node-1 ~]# neutron agent-list

+--------------------------------------+--------------------+-------------------+-------+----------------+

| id | agent\_type | host | alive | admin\_state\_up |

+--------------------------------------+--------------------+-------------------+-------+----------------+

| 01f2b261-4d34-4ff0-87bd-d1ee83d7c4ac | Open vSwitch agent | node-5.domain.tld | :-) | True |

| 15c8774d-190a-4741-8b5c-fef7dc413562 | DHCP agent | node-3.domain.tld | xxx | True |

| 212a6adb-394d-409f-9022-73f940086835 | Metadata agent | node-2.domain.tld | :-) | True |

| 2bd0cc46-fead-414a-93b7-624bd68f8053 | Open vSwitch agent | node-1.domain.tld | :-) | True |

| 309d33a5-211a-431a-8b48-35aecdb11f6b | Open vSwitch agent | node-2.domain.tld | :-) | True |

| 5dd63653-2024-4945-a823-dd967262a7f8 | L3 agent | node-3.domain.tld | :-) | True |

| 62c6e903-7f4e-4178-95fb-23c81977d4c8 | Metadata agent | node-1.domain.tld | :-) | True |

| 670cf505-cac0-46a4-a20c-3cc4bb88a914 | DHCP agent | node-1.domain.tld | :-) | True |

| 7d98f1fa-042c-494c-ae87-73305ef1be02 | DHCP agent | node-2.domain.tld | xxx | True |

| aadb6253-5dac-4c62-b7c9-54ffc6f39eee | Metadata agent | node-3.domain.tld | :-) | True |

| eceda162-20ee-4dda-b1ae-64bc2b0afbdb | Open vSwitch agent | node-4.domain.tld | :-) | True |

| f41489b4-8729-4fa5-b717-f1a223e5749c | L3 agent | node-1.domain.tld | xxx | True |

| fc286f57-47d3-4c46-9550-579cbd5a0440 | Open vSwitch agent | node-3.domain.tld | :-) | True |

+--------------------------------------+--------------------+-------------------+-------+----------------+

L3 agent和DHCP agent只要各有一个是正常的就行。

### cinder状态检查

[root@node-1 ~]# cinder service-list

+------------------+-------------------+------+---------+-------+----------------------------+

| Binary | Host | Zone | Status | State | Updated\_at |

+------------------+-------------------+------+---------+-------+----------------------------+

| cinder-scheduler | node-1.domain.tld | nova | enabled | up | 2014-12-02T11:24:27.000000 |

| cinder-scheduler | node-2.domain.tld | nova | enabled | up | 2014-12-02T11:24:28.000000 |

| cinder-scheduler | node-3.domain.tld | nova | enabled | up | 2014-12-02T11:24:30.000000 |

| cinder-volume | node-4.domain.tld | nova | enabled | up | 2014-12-02T11:24:32.000000 |

| cinder-volume | node-5.domain.tld | nova | enabled | up | 2014-12-02T11:24:32.000000 |

+------------------+-------------------+------+---------+-------+----------------------------+

### ceph状态检查

[root@node-1 ~]# ceph -s

cluster 478cd836-609f-4d00-96bb-ac58b813ca99

health HEALTH\_OK

monmap e3: 3 mons at {node-3=192.168.0.5:6789/0,node-4=192.168.0.6:6789/0,node-5=192.168.0.7:6789/0}

election epoch 12, quorum 0,1,2 node-3,node-4,node-5

osdmap e104: 24 osds: 24 up, 24 in

pgmap v32159: 704 pgs, 6 pools, 1096 MB data, 10009 objects

3380 MB used, 22308 GB / 22311 GB avail

704 active+clean

client io 505 kB/s rd, 317 kB/s wr, 837 op/s

若ceph状态为HEALTH\_WARN或HEALTH\_ERROR，可用ceph health detail查看详细信息，并且需要立刻马上检查ceph警告或错误原因，并修复。

### pacemaker管理的各服务集群状态检查

[root@node-1 ~]# crm status

Last updated: Tue Dec 2 11:40:19 2014

Last change: Tue Dec 2 11:40:13 2014 via crm\_attribute on node-1.domain.tld

Stack: classic openais (with plugin)

Current DC: node-1.domain.tld - partition with quorum

Version: 1.1.10-14.el6\_5.3-368c726

3 Nodes configured, 3 expected votes

25 Resources configured

Online: [ node-1.domain.tld node-2.domain.tld node-3.domain.tld ]

vip\_\_management\_old (ocf::es:ns\_IPaddr2): Started node-1.domain.tld

vip\_\_public\_old (ocf::es:ns\_IPaddr2): Started node-3.domain.tld

Clone Set: clone\_ping\_vip\_\_public\_old [ping\_vip\_\_public\_old]

Started: [ node-1.domain.tld node-2.domain.tld node-3.domain.tld ]

p\_openstack-ceilometer-central (ocf::es:ceilometer-agent-central): Started node-1.domain.tld

p\_openstack-ceilometer-alarm-evaluator (ocf::es:ceilometer-alarm-evaluator): Started node-2.domain.tld

Clone Set: clone\_p\_mysql [p\_mysql]

Started: [ node-1.domain.tld node-2.domain.tld node-3.domain.tld ]

Clone Set: clone\_p\_haproxy [p\_haproxy]

Started: [ node-1.domain.tld node-2.domain.tld node-3.domain.tld ]

p\_openstack-heat-engine (ocf::es:openstack-heat-engine): Started node-1.domain.tld

Clone Set: clone\_p\_neutron-openvswitch-agent [p\_neutron-openvswitch-agent]

Started: [ node-1.domain.tld node-2.domain.tld node-3.domain.tld ]

Clone Set: clone\_p\_neutron-metadata-agent [p\_neutron-metadata-agent]

Started: [ node-1.domain.tld node-2.domain.tld node-3.domain.tld ]

p\_neutron-dhcp-agent (ocf::es:neutron-agent-dhcp): Started node-1.domain.tld

p\_neutron-l3-agent (ocf::es:neutron-agent-l3): Started node-2.domain.tld

### l3-agent的namespace

从上面的状态可以看到l3-agent在node-2，所以在node-2上查看namespace

[root@node-2 ~]# ip netns

haproxy

qrouter-b6aa0473-1fa7-44af-bc67-49efefdeb209

[root@node-2 ~]# ip netns exec qrouter-b6aa0473-1fa7-44af-bc67-49efefdeb209 ip a

18: lo: <LOOPBACK,UP,LOWER\_UP> mtu 16436 qdisc noqueue state UNKNOWN

link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00

inet 127.0.0.1/8 scope host lo

inet6 ::1/128 scope host

valid\_lft forever preferred\_lft forever

19: qr-38c57255-95: <BROADCAST,UP,LOWER\_UP> mtu 1500 qdisc noqueue state UNKNOWN

link/ether fa:16:3e:51:cc:79 brd ff:ff:ff:ff:ff:ff

inet 192.168.111.1/24 brd 192.168.111.255 scope global qr-38c57255-95

inet6 fe80::f816:3eff:fe51:cc79/64 scope link

valid\_lft forever preferred\_lft forever

20: qg-93b57a01-80: <BROADCAST,UP,LOWER\_UP> mtu 1500 qdisc noqueue state UNKNOWN

link/ether fa:16:3e:fc:fc:36 brd ff:ff:ff:ff:ff:ff

inet 172.16.0.130/24 brd 172.16.0.255 scope global qg-93b57a01-80

inet6 fe80::f816:3eff:fefc:fc36/64 scope link

valid\_lft forever preferred\_lft forever

能看到qr，qg和对应的IP

### 查看dhcp-agent的namespace

从上面的状态可以看到dhcp-agent在node-1，所以在node-1上查看namespace

[root@node-1 ~]# ip netns

haproxy

qdhcp-3344c78a-d5a8-419c-bbc6-60424bf0ef14

[root@node-1 ~]# ip netns exec qdhcp-3344c78a-d5a8-419c-bbc6-60424bf0ef14 ip a

32: tap597c28fe-4b: <BROADCAST,UP,LOWER\_UP> mtu 1500 qdisc noqueue state UNKNOWN

link/ether fa:16:3e:ec:12:c4 brd ff:ff:ff:ff:ff:ff

inet 192.168.111.2/24 brd 192.168.111.255 scope global tap597c28fe-4b

inet6 fe80::f816:3eff:feec:12c4/64 scope link

valid\_lft forever preferred\_lft forever

33: lo: <LOOPBACK,UP,LOWER\_UP> mtu 16436 qdisc noqueue state UNKNOWN

link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00

inet 127.0.0.1/8 scope host lo

inet6 ::1/128 scope host

valid\_lft forever preferred\_lft forever

可以看到一个tap设备，ip是dhcp的IP。

如果虚机分配不到IP，请检查dhcp服务是否正常。

[root@node-1 ~]# ps aux | grep dhcp

neutron 13376 0.0 1.4 169052 34032 ? S 03:15 0:21 /usr/bin/python /usr/bin/neutron-dhcp-agent --config-file=/etc/neutron/neutron.conf --config-file=/etc/neutron/dhcp\_agent.ini

nobody 24012 0.0 0.0 12948 660 ? S 03:20 0:00 dnsmasq --no-hosts --no-resolv --strict-order --bind-interfaces --interface=tap597c28fe-4b --except-interface=lo --pid-file=/var/lib/neutron/dhcp/3344c78a-d5a8-419c-bbc6-60424bf0ef14/pid --dhcp-hostsfile=/var/lib/neutron/dhcp/3344c78a-d5a8-419c-bbc6-60424bf0ef14/host --addn-hosts=/var/lib/neutron/dhcp/3344c78a-d5a8-419c-bbc6-60424bf0ef14/addn\_hosts --dhcp-optsfile=/var/lib/neutron/dhcp/3344c78a-d5a8-419c-bbc6-60424bf0ef14/opts --leasefile-ro --dhcp-range=set:tag0,192.168.111.0,static,120s --dhcp-lease-max=256 --conf-file= --domain=openstacklocal

root 29846 0.0 0.0 103248 888 pts/1 S+ 13:02 0:00 grep dhcp

[root@node-1 ~]# cat /var/lib/neutron/dhcp/3344c78a-d5a8-419c-bbc6-60424bf0ef14/host

fa:16:3e:ad:7e:d2,host-192-168-111-3.openstacklocal,192.168.111.3

fa:16:3e:51:cc:79,host-192-168-111-1.openstacklocal,192.168.111.1

fa:16:3e:ec:12:c4,host-192-168-111-2.openstacklocal,192.168.111.2

fa:16:3e:8d:54:33,host-192-168-111-4.openstacklocal,192.168.111.4

如果正常这个host文件里会有虚机的mac地址和分配的IP的记录，如果没有这条记录，请及时处理。

# 常见故障处理

## 通用故障

### 控制节点宕机

ThinkCloud的OpenStack集群中拥有3台controller，任何一台controller宕机后。其余两台controller会自动接管宕机controller上的所有服务。运维人员需要做的，只需修复该宕机controller，并重启使其上线运行即可。

### 操作系统盘损坏

操作系统盘已经通过使用服务器RAID卡，配置了RAID 1 或则RAID5对操作系统数据进行了保护。并且服务器使用的磁盘，大多为支持热插拔的SATA或者SAS磁盘。直接将坏盘拔下，插入新盘，RAID卡会自动进行数据恢复。

### 服务状态异常

#### Rabbitmq-server

现象１：Rabbitmq-server状态正常，但OpenStack组件日志中报告amqp connection timeout，可能的原因有

１．网络异常，需先修复网络问题

２．使用消息队列的服务异常．可重启相关服务

３．rabbitmq自身异常，如消息阻塞等，可按1.2.2节所述重启rabbitmq-server服务．

现象2：Rabbitmq-server服务状态异常，可按1.2.2节所述重启rabbitmq-server服务．

#### hwmgmt-api

hwmgmt-api是系统关键服务，用于收集系统硬件信息及硬件告警．若hwmgmt-api不能正常工作，会导致OpenStack页面无法访问．可查看日志文件/var/log/hwmgmt/hwmgmt-api.log定位问题．

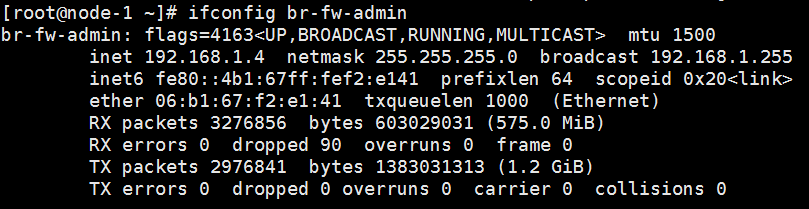
#### collectd

collectd是系统监控服务，需要内置OpenStack访问帐号于其配置文件/etc/gnocchi/collectd.conf, 当内置帐号密码修改后，需要同步到其配置文件．

### ntp server不工作

ThinkCloud OpenStack默认将部署节点作为ntp server, 若部署节点被移除后，可能导致云平台各节点时间不一致，进而导致OpenStack关键服务异常。在此情况下，可将某一控制节点设为ntp server,具体操作如下：

1. 获取该控制节点br-fw-admin网口的ip: ifconfig br-fw-admin



1. 编辑该控制节点的/etc/ntp.conf,在文件最后加上

server 127.127.1.0

fudge 127.127.1.0 stratum 8

1. 重启该控制节点上的ntp服务：systemctl restart ntpd
2. 在其它各节点上/etc/ntp.conf的server list中添加该控制节点，并重启ntpd服务。

server 192.168.1.4

## 计算

### 虚拟机状态错误

用户对虚拟机进行创建、迁移、重启操作时，如果操作失败虚拟机状态变为错误，会不断收到报警邮件。

处理方案：

1. 登录控制节点，通过nova show vm\_id 查看vm 所在的节点
2. 恢复虚拟机状态：nova reset-state vm\_id --active
3. 关闭虚拟机
4. 登录到虚拟机所在节点，查看日志信息（/var/log/nova/compute.log），查看错误信息找线索，通常原因如下：

4.1）如果是重启引发的多数是存储集群有问题

4.2）迁移引发的，从网络，云硬盘状态 、存储集群三方面分析

4.3）创建引发的，查看资源配额，网络，存储

### 同一个vm运行在多个宿主机上

同一个vm运行在两个不同的宿主机上，通常由迁移失败引发的。

解决方法：

1. 通过nova show vm\_uuid 查看虚拟机实际在哪个宿主机
2. 关闭虚拟机
3. 登录到另一台宿主机上，通过virsh destroy instance\_name

### vm（Linux系统）用户自己升级失败导致系统无法启动

解决方法：

1. 关闭虚拟机，查看该虚拟机的id
2. 新建一个vm,并找到该vm 所在节点
3. 登录新建虚拟机所在的节点
4. 新建attach.xml，内容如下：

<disk type='network' device='disk'>

<driver name='qemu' type='raw' cache='none'/>

<auth username='compute'>

<secret type='ceph' uuid='a5d0dd94-57c4-ae55-ffe0-7e3732a24455'/>

</auth>

<source protocol='rbd' name='compute/6f0ae0fe-150c-4837-9366-c4e54e2b8767\_disk'>

<host name='192.168.0.3' port='6789'/>

<host name='192.168.0.4' port='6789'/>

<host name='192.168.0.5' port='6789'/>

</source>

<backingStore/>

<target dev='vdb' bus='virtio'/>

</disk>

1. 用出问题的虚拟机的id 替换标红的uuid
2. 在新建vm 所在宿主机上执行virsh attach-device attach.xml
3. 等陆到vm 系统中 ，挂载云硬盘到/tmp 如：mount /dev/vdb /tmp
4. 编辑/tmp/boot/grub/grub.conf 修改启动选项，完成退出，并卸载该盘，umount /tmp
5. 在新建vm 所在宿主机上执行virsh detach-device attach.xml

10）重启问题虚拟机

### 虚机热迁移失败

解决方法：

1. 通过控制节点进入MySQL，执行

select host, launched\_on, node from nova.instances where uuid='74f743d0-0fb5-4855-9511-6428f9ae291d';

(注：uuid 为迁移失败的云主机的UUID)

update nova.instances set host='node-6.domain.tld', launched\_on='node-6.domain.tld', node='node-6.domain.tld' where uuid='74f743d0-0fb5-4855-9511-6428f9ae291d';

通过修改数据库，将云主机的宿主机改为迁移之前的宿主机

select a.port\_id, b.host from neutron.ipallocations a, neutron.ml2\_port\_bindings b where a.port\_id = b.port\_id and a.ip\_address='192.168.111.6';

通过云主机的内网IP地址，查询其neutron端口ID

update neutron.ml2\_port\_bindings set host='node-6.domain.tld' where port\_id= '49fcf468-3b60-4153-bec1-e10b95002f66';

通过修改数据库，将云主机的端口宿主机改为迁移之前的端口宿主机

nova reset-state 74f743d0-0fb5-4855-9511-6428f9ae291d

nova stop 74f743d0-0fb5-4855-9511-6428f9ae291d

1. 登录到thinkcloud UI页面，开启云主机。

### 计算节点加public网段，起br-ex,同时给虚机增加public网IP

1. 进入到计算节点执行：

ovs-vsctl add-br br-ex

起br-ex网桥并连接至以有网桥br-ethX

ovs-vsctl add-port br-ethX br-ethX--br-ex

ovs-vsctl add-port br-ex br-ex--br-ethX

ovs-vsctl set interface br-ex--br-ethX type=patch

ovs-vsctl set port br-ex--br-ethX trunks=0

ovs-vsctl set interface br-ex--br-ethX options:peer=br-ethX--br-ex

ovs-vsctl set interface br-ethX--br-ex type=patch

ovs-vsctl set port br-ethX--br-ex trunks=0

ovs-vsctl set interface br-ethX--br-ex options:peer=br-ex--br-ethX

给br-ex添加静态IP

1. 编辑 /etc/sysconfig/network-scripts/ifcfg-br-ex

配置示例：

BOOTPROTO=static

DEVICE=br-ex

ONBOOT=yes

TYPE=Ethernet

USERCTL=no

IPADDR=x.x.x.x

NETMASK=255.255.255.x

GATEWAY=x.x.x.x

删除默认路由，增加新的路由

ip r del default via x.x.x.x #fuel ip

ip r add default via x.x.x.x dev br-ex #public gateway

重启网络 service network restart

编辑/etc/neutron/plugins/ml2/openvswitch\_agent.ini

bridge\_mappings =physnet1:br-ex,physnet2:br-prv

重启systemctl restart neutron-openvswitch-agent

1. 进入到控制节点

重启systemctl restart neutron-l3-agent.service

. /root/ openrc.v2

给public网开启dhcp

neutron subnet-update xxx --enable-dhcp

给虚机增加网卡

nova interface-attach --net-id uuid0 uuid1 #uuid0为public网的ID，uuid1为虚机的ID

#注 如果是centos，再添加第二块网卡后，其默认路由仍旧是第一块网卡的网关，视情况删除默认路由，增加新的路由，命令参考

ip r del default via x.x.x.x #share subnet gateway

ip r add default via x.x.x.x dev ethx #public subnet gateway

## 存储

### ceph osd down

解决方法：

1. 通过ceph osd tree|grep down 查找osd 所在的节点
2. 登录到该节点，通过dmesg信息查看是否有关于磁盘方面的错误，并通过ipmi 确认是否为硬件问题
3. 如果是硬件问题，把该osd移出集群，具体方法如下：

*/etc/init.d/ceph stop osd.$osd\_num*

*ceph osd out osd.$osd\_num*

*ceph osd crush rm osd.$osd\_num*

*ceph osd rm osd.$osd\_num*

*ceph auth del osd.$osd\_num*

*umount /var/lib/ceph/osd/ceph-$osd\_num*

*rm -rf /var/lib/ceph/osd/ceph-$osd\_num*

***注：确认好是硬件问题，否则不要轻易移除osd***

4) 如果不是硬件问题，启动该osd , 启动命令 : /etc/init.d/ceph start osd.$osd\_num

### ceph 'error removing image'

虚拟机删除失败，如查看log 有ceph 'error removing image' 信息是可能vm 运行在两个宿主机上，解决方法参见[2.2.2](#OLE_LINK1)

### mon.node-x store is getting too big!

解决方法：

1. 登录到对应节点如node-1 执行ceph tell mon.node-1 compact

***注：压缩完成后,mon 状态为down ,等up后在进行另一个mon压缩***

1. 重启mon ，删除LOG.old文件, 如：重启node-1上mon

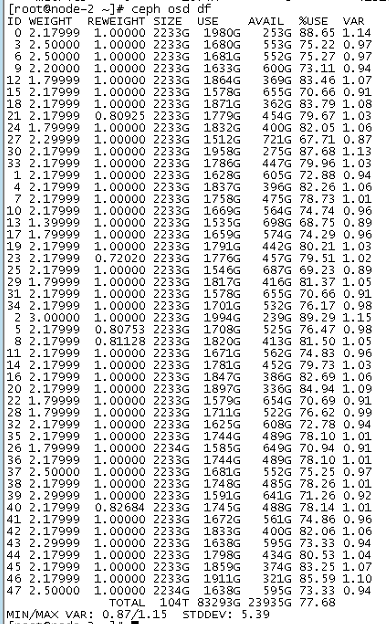
>> /etc/init.d/ceph stop mon.node-1

>> rm /var/lib/ceph/mon/ceph-node-1/store.db/LOG.old

### Full osd或near full osd

解决方法：

1. Ceph osd df查看各osd使用量，通过手动调整各osd权重看是否能恢复ceph到健康状态：



通过命令ceph osd crush reweight osd.x yyy将使用率偏高的weight下调，偏低的上调，这里yyy值参考WEIGHT.

若上述方法仍不能让ceph –s恢复到HEALTH\_OK或HEALTH\_WARN,则进一步适量增大调整:

ceph tell osd.\* injectargs '--mon\_osd\_full\_ratio 0.97' *#was 0.95*

ceph tell osd.\* injectargs '--mon\_osd\_nearfull\_ratio 0.9' *#was 0.85*

ceph tell osd.\* injectargs '--osd\_backfill\_full\_ratio 0.9' *#was 0.85*

ceph tell osd.\* injectargs '--osd\_failsafe\_nearfull\_ratio 0.92' *#was 0.9*

1. 检查pg\_num/pgp\_num是否合理

每个pool的pg num的经验计算公式: (100\*OSD盘数)/(副本数\*pool数),向上取最接近的2的倍数. 注: 这里的pool数仅计算真实数据的pool。

## 网络

### 虚机ping 不通，无法远程登录

解决方法：

1. 通过vnc登录虚拟机验证系统是否正常
2. 查看安全组，是否开放ICMP端口
3. 查看路由设置,有些需要设置默认路由。Linux通过ip route show ,windows 通过route print

3.1) 只有10.x.x.x网段查看默认路由

3.2) 配置了43.x.x.x 和10.x.x.x网段的，43.x.x.x 对应默认路由

10.x.x.x 需要单独添加路由规则

# 系统定制Tips

## 修改配置以启/停相关功能

### 强制计费功能失效

部署时选装了计费模块，若想让计费功能失效，可在所有控制节点上设置：

#vim /etc/openstack-dashboard/local\_settings

ENABLE\_BILLING = False

#service httpd restart

### 开启包月包年模式

要需要到控制节点上修改配置文件，具体如下：

在controller节点执行以下操作：

#vim /etc/openstack-dashboard/local\_settings

添加下面配置

PREBILLING = True #包年包月

然后重启服务

#service httpd restart

### 支付宝帐号修改

若需要修改支付宝帐号，可在所有控制节点上设置：

#vim /usr/share/openstack-dashboard/easystack\_dashboard/settings.py

*ALIPAY\_KEY = '* *xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'*

*ALIPAY\_PARTNER = 'xxxxxxxxxxxxxxxx' # partner ID*

*ALIPAY\_SELLER\_EMAIL = 'xxxxxxx' #alipay account*

#service httpd restart

### 激活邀请码充值

在所有控制节点上：

#vim /etc/openstack-dashboard/local\_settings

*INVCODE\_RECHARGE = True*

*PREBILLING = True*

#service httpd restart

### 强制申请与审批功能失效

部署时选装了申请与审批模块，若想让申请与审批功能失效，可在所有控制节点上设置：

#vim /etc/openstack-dashboard/local\_settings

APPROVAL\_ENABLED = False

#service httpd restart

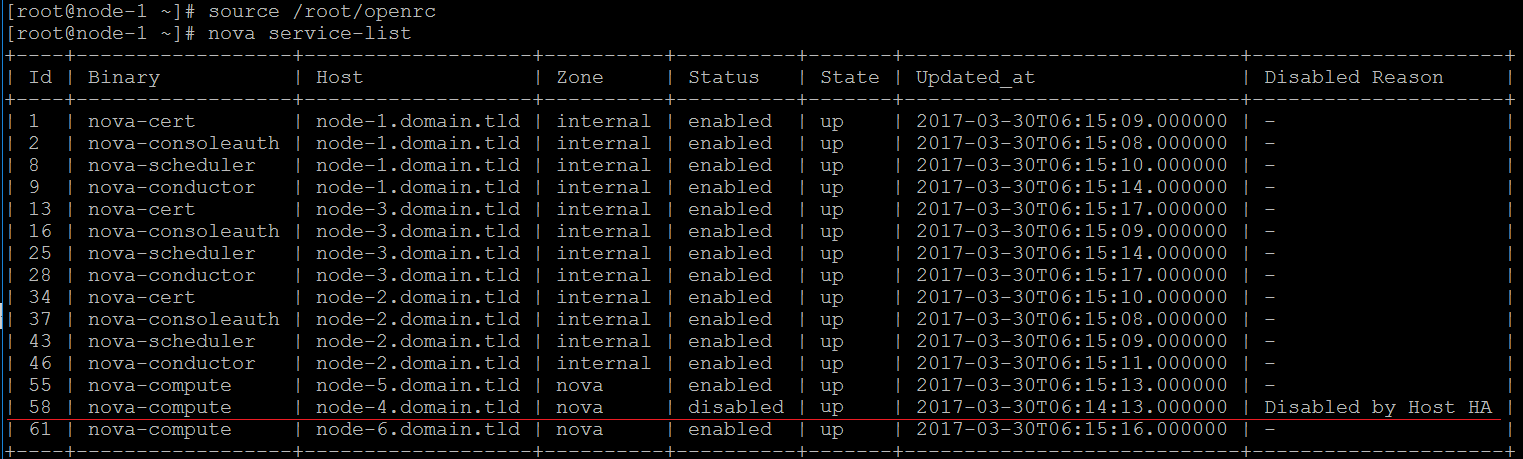
### 配置Host HA功能

* + **功能说明：**

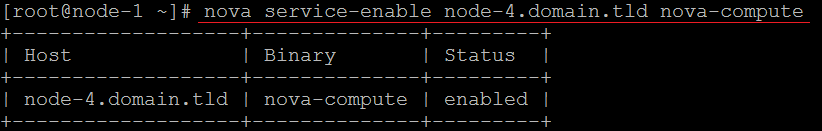
1. 人为关物理机或调整网卡设置前，要使用nova service-disable {node-name} nova-compute 将物理机置成维护模式，明确告诉Host HA不需要监管该物理主机。
2. 因为后台存储／网络等原因，虚机状态不正确，需要管理员介入，恢复虚机状态。
3. 人工介入时，需要将Host HA运行时失效，可创建一个空文件/etc/lenovo/enable\_ha.conf, 使Host HA结束当前监管，等待/var/log/nova/hagent.log 出现＂skip Host HA by user＂字样才可以人工介入。若需要重新使Host HA生效，可删除该文件。
4. 确认每个计算节点的/etc/nova/nova.conf　配置项my\_ip={管理网ＩＰ}

resize\_confirm\_window=600, 控制节点allow\_resize\_to\_same\_host=False, allow\_migrate\_to\_same\_host = False

1. 某个物理机断电，很短的时间内重新插电，仅当系统已经重新插电，IPMI状态显示系统为ON，由于系统没有完全启动，其他所有状态都是FAIＬ，Host HA会执行POWER OFF和Evacuate。这种情况下，虚机的业务已经受到影响，即使重新加电服务器，等到操作系统及服务起来还需要比较长时间，这期间用IPMI关物理机，把虚机迁移走也是必要的。
2. 极特殊情况下，在接连几个检测周期里，物理机生产网卡恰好依次出现问题，每次检测周期内都满足有问题的物理机总数小于或等于/etc/lenovo/hagent.conf中的配置项fault\_hosts\_number\_threshold时，按照目前的处理逻辑，每检查到一个失败，就会migrate一次，或者当物理机上虚机过多，没有物理机可以migrate时，需要管理员介入。
3. 当一台计算节点发生故障需要迁移其上的虚机时，系统会disable其nova-compute服务，如下图：



用户recover此节点时，需要手动enable其nova-compute服务，如下图：



1. 根据多元组信息采取对应的动作见下表：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Power | 虚机访问存储网络(ceph-public) | 管理网 | 生产网 | 当作存储节点 | 动作 | 备注 |
| OFF |  |  |  |  | Evacuate | 一个条件就够 |
| ON | ON | OFF | ON |  | None | 管理网不影响生产 |
| ON | ON | ON | OFF | ON/OFF | Migrate | Migrate虚机。需要Dashboard上云主机列表中确认Migrate还是revert. |
| ON | ON | OFF | Unknown |  | None | 管理网OFF，生产网卡状态未知，算作ON |
| ON | ON | OFF | ON |  | None | 这种情况，不能检测到虚机生产网卡down，不会出现这种状态 |
| ON | ON | OFF | OFF |  | None | 这种情况，不能检测到虚机生产网卡down，  不会出现这种状态． |
| ON | OFF | ON | ON | ON | None | Ceph-public 断，会导致nova-compute  为down状态，也是存储节点，不采取任何措施 |
| ON | OFF | ON | ON | OFF | Power off Evacuate | Ceph-public 断，会导致nova-compute  为down状态，只能Evacuate. |
| ON | OFF | OFF |  | OFF | Power off Evacuate | 不做存储节点，管理网／虚机访问后端存储，power off/evacuate |
| ON | OFF | OFF |  | ON | None | 当作存储节点　不做操作 |
| ON | OFF | ON | OFF | ON | None | Ceph-public 断，会导致nova-compute  为down状态，也是存储节点，不采取任何措施. |
| ON | OFF | ON | OFF | OFF | Power off Evacuate | Ceph-public 断，会导致nova-compute  为down状态，只能Evacuate. |

* + **启动Ｈost HA功能:**

控制节点HA: #crm resource start lenovo-hagent

控制节点非HA: #systemctl restart lenovo-hagent

需要修改以下配置文件使host ha生效:

所有controller节点上修改/etc/lenovo/下的hagent.conf、ipmi.conf及ping\_list.conf等文件。配置文件参考：



1. /etc/lenovo/hagent.conf

*[DEFAULT]*

*use\_rpc=False*

*debug = False*

*verbose=True*

*log\_file = /var/log/nova/hagent.log*

*check\_interval = 5　 #监控的最小间隔*

*on\_shared\_storage = True #目前只支持True，当为False时不会采取任何action．*

*evacuate\_interval=1*

*action\_in\_same\_aggregate=False*

*fault\_hosts\_number\_threshold=2 #指同一时刻有问题的计算节点总数超过这个值，说明发生了整个平台*

*#的故障，不采取任何动作，如设置为２，有三台出现问题就不处理．*

*service\_down\_time=60*

*dry\_run=False #当为true时，只检测不做任何action，可用于测试*

*mixed\_hosts=null #定义融合节点（计算＋存储）集合。当此项设置为null时，系统默认融合节*

*#点为计算节点，当某融合节点的存储网络不可用时系统会迁移其上虚机*

*#若用户定义了融合节点集合，则集合中某融合节点的存储网络不可用*

*#时系统不执行迁移操作。*

*[Auth]*

*username = admin*

*password = admin*

*auth\_url = http://192.168.2.2:5000/v2.0*

*project\_id = admin*

*region\_name = RegionOne*

*[ping]*

*ping\_list\_file\_path = /etc/lenovo/ping\_list.conf*

*packet\_count = 5 # 检测时的发包数目*

*packet\_interval = 1 # 检测时的发包间隔*

*ip\_item\_format=MANAGEMENT\_IP,CEPH\_PUBLIC\_IP*

*[ipmi]*

*conf\_file\_path = /etc/lenovo/ipmi.conf*

*separator=' '*

1. /etc/lenovo/ipmi.conf

*#please use IMM IP/username/password for this*

*#Host's IPMI access information the format like this*

*# {HC-node-name}={ipmi\_address} {ipmi\_username} {ipmi\_password}*

*node-4.domain.tld=10.240.220.7 USERID PASSW0RD*

*node-5.domain.tld=10.240.220.8 USERID PASSW0RD*

*node-6.domain.tld=10.240.220.9 USERID PASSW0RD*

1. /etc/lenovo/ping\_list.conf

*# The list for the ping inspector*

*# Before the '=' is the node name.*

*# The IPs are seperated by ','*

*# ip\_item\_format＝br-mgmt ip,br-storage ip*

*# The sequence of the IPs are specified by ip\_item\_format in the hagent.conf*

*node-4.domain.tld=192.168.2.6,192.168.4.5*

*node-5.domain.tld=192.168.2.7,192.168.4.6*

*node-6.domain.tld=192.168.2.8,192.168.4.7*

注：若客户环境无法配BMC网络，可在所有控制节点上修改/etc/hwmgmt/hwmgmt.conf，　置ha\_ipmi\_enabled = False

[ha]  
# for configure Host HA in dashboard  
  
# If this value is changed, you should restart hwmgmt-api.service  
# and lenovo-hagent.service on all controller nodes.  
ha\_ipmi\_enabled = False

并重启lenovo-hagent及hwmgmt-api服务：

控制节点HA:

# crm resource start lenovo-hagent // 任一控制节点

# systemctl restart hwmgmt-api // 所有控制节点

控制节点非HA: #systemctl restart lenovo-hagent & systemctl restart hwmgmt-api

* + **关闭Ｈost HA功能:**

#crm resource stop lenovo-hagent

### 配置VM HA功能

* + **功能说明：**

1. 在云平台中有些业务并不能通过集群等手段实现高可用，需要对这类型单个虚机进行监控，出现故障如虚机内部crash，意外关机，根据预先设定策略，采取措施恢复虚机，即针对单个虚机的高可用方案。
2. VM HA只运行在compute节点。
   * **启动VM HA功能:**

#service lenovo-vagent start

需要在compute节点上修改配置文件/etc/lenovo/vagent.conf。配置文件参考：



*[DEFAULT]*

*log\_file= /var/log/nova/vagent.log*

*# log\_level: CRITICAL, ERROR, WARNING, INFO, DEBUG*

*log\_level = DEBUG*

*# enable task current supports ha, maintain*

*enable\_tasks= ha*

*[Auth]*

*# Find configurations from /ect/nova/nova.conf on compute node*

*username = admin*

*password = admin*

*auth\_url=http://192.168.0.2:5000/v2.0*

*project\_id = admin*

*[HA]*

*on\_shared\_storage = True*

*check\_interval = 20*

*# detect strategy , current supports ping,crashed, norunning*

*available\_strategy = norunning*

*# how to combine the strategy ,supports all, any*

*strategy\_combine= all*

*# what pre-actions before taking actual actions*

*# supports pre\_notify, email*

*pre\_actions =*

*# what post-actions before taking actual actions*

*# supports post\_notify, email*

*post\_actions =*

*# take action when service is up , supports reboot, migrate, rebuild, start*

*up\_action = start*

*# take action when service is down, only supports evacuate*

*down\_action = evacuate*

*[MAINTAIN]*

*check\_interval = 30*

*flag\_file\_path = /etc/lenovo/agent\_flag.conf*

* + **关闭Ｈost HA功能:**

#service lenovo-vagent stop

### 启用物理网络拓扑

1. 配置交换机机ip

为环境中所有交换机配置一个controller节点可达的管理ip。

例如，在Lenovo Networking Operating System (NOS) G8272交换机上配置如下: ip address 192.168.10.1/24

1. 设置交换机lldp：确认所有的交换机已开启lldp。

例如，在Lenovo Networking Operating System (NOS) Version 10.4.2.0 G8272交换机上已默认开启lldp，可通过display lldp interface all查看各端口lldp信息。

其他交换机需根据官方交换机配置文档开启lldp（一般在config模式下使用feature lldp开启）

1. 设置交换机snmp：在所有交换机上配置snmp。

例如，在Lenovo Networking Operating System (NOS) Version 10.4.2.0 G8272交换机上：

1. snmp-server enable snmp
2. snmp-server version v1v2v3
3. snmp-server community public group network-operator
4. 配置proton配置文件

在controller节点上，修改/etc/proton/proton.conf.

* + 在[topology]中设置switch\_ip\_range，代表1内配置的所有交换机管理ip地址的范围，可以是一个列表，以逗号隔开，每个范围可以是一个ip地址，或者用掩码表示的ip地址范围，或者ip地址A到B的范围，例如：

switch\_ip\_range = 192.168.10.1, 192.168.20/24, 192.168.30.1-192.168.30.10

* + 在[topology]中设置snmp\_version
  1. 若snmp\_version为2，则需要配置snmp\_v2\_community = public
  2. 若snmp\_version为3，则需要配置如下项：

*snmp\_v3\_seclevel = authPriv*

*snmp\_v3\_authproto = MD5*

*snmp\_v3\_authpass = admin*

*snmp\_v3\_privproto = DES*

*snmp\_v3\_privpass = admin*

*snmp\_v3\_secname = admin*

1. 重启pronton-topology服务
   * 若ThinkCloud为HA环境，使用crm重启服务：

crm resource restart proton-topology

* + 若ThinkCloud为multi-node环境，使用systemctl重启服务：

systemctl restart proton-topology.service

### 修改云硬盘1TB上限

1. 在3个controller节点执行以下操作：  
   #vim /etc/openstack-dashboard/local\_settings  
   最下面添加下面这项配置  
   MAX\_VOLUME\_SIZE = 100000000 # Volume 大小, MB  
   然后重启服务  
   #service httpd restart
2. 在任一Controller节点执行以下操作：  
   #source openrc  
   #keystone tenant-list //获取所要修改配额的tenant id  
   #cinder quota-update --gigabytes -1 --backup-gigabytes -1 --volumes -1 --snapshots -1 --backups -1 xxxx //此xxxx为tenant id

### 启用flat网络

1. 在控制节点和计算节点上，修改 /etc/neutron/plugins/ml2/ml2\_conf.ini
   1. 将type\_drivers = vlan,flat,local 修改为 type\_drivers = flat,vlan,local
   2. 将tenant\_network\_types = vlan修改为 tenant\_network\_types = flat
   3. 将mechanism\_drivers=openvswitch,lenovo 修改为 mechanism\_drivers=openvswitch
   4. 在控制节点执行 systemctl restart neutron-server

在计算节点执行 systemctl restart neutron-openvswitch-agent

1. 配置修改完之后，便可以开始创建flat网络

进入某个controller执行以下命令:

# source /root/openrc

# neutron net-create flat\_network\_name --provider:network\_type flat --provider:physical\_network physnet2 --shared

对于参数 physnet2 需要根据 /etc/neutron/plugins/ml2/openvswitch\_agent.ini

中的 bridge\_mappings =physnet1:br-ex,physnet2:br-prv

physnet2指向了 br-prv，而我们的又将private网络配置为flat网络，因此 这个地方跟参为 physnet2

1. 修改保存后即可在UI上这个网络下创建子网，需要注意的是创建子网时，不要enable DHCP 和网关

### 允许域账户登录

系统默认用户邮箱登录，若需要通过域账户登录，可在所有控制节点上设置：

#vim /etc/openstack-dashboard/local\_settings

DOMAIN\_LOGIN\_ENABLE = True

#service httpd restart

若域账户登录过程中出现“MemcachedKeyLengthError: Key length is > 250”的错误，可编辑/etc/openstack-dashboard/local\_settings， 可替换memcached配置项如下:

**def hash\_key**(key, key\_prefix, version):  
 new\_key = ':'.join([key\_prefix, str(version), key])  
 **if** len(new\_key) > 250:  
 **import** hashlib  
 m = hashlib.md5()  
 m.update(new\_key)  
 new\_key = m.hexdigest()  
 **return** new\_key  
CACHES = {  
 'default': {  
 'BACKEND': 'django.core.cache.backends.memcached.MemcachedCache',  
 'LOCATION': '127.0.0.1:11211',  
 'KEY\_FUNCTION': hash\_key,  
 }  
}

### 开启Ceph存储空间回收通知机制

在OpenStack云环境的部署中，Ceph存储具备Thin provision的功能，这项功能实现了存储按需分配的能力。以RBD image为例，它本身是稀疏格式的，也就是说它所占用objects会随着用户写入数据的增加而增加（Thin provision）。当用户删除数据以后，这些obejct不再使用，但并没有被释放。因为从Ceph的角度讲，它并不知道文件系统中发生的事情。若要回收,需要开启回收通知机制，步骤如下：

1. 给所有镜像加hw\_scsi\_model和hw\_disk\_bus两个属性:

*#glance image-update --property hw\_scsi\_model=virtio-scsi --property hw\_disk\_bus=scsi xxxxx*

1. 编辑各计算节点上的nova.conf,在[libvirt]中添加：

*[libvirt]*

*…*

*hw\_disk\_discard = unmap*

并重启nova-compute服务

1. 在对接ceph的cinder节点（ThinkCloud OpenStack默认位于控制节点）的/etc/cinder/cinder.conf的ceph backend section里添加：

*…*

*volume\_driver=cinder.volume.drivers.rbd.RBDDriver*

*report\_discard\_supported=True*

*…*

并重启cinder-volume服务。

1. 在虚机中执行fstrim通知ceph回收空间：

*#fstrim –v mountpoint\_of\_ceph\_hdd*

计算节点上改了之后要重启nova-compute服务, cinder要重启cinder-volume

系统默认用户邮箱登录，若需要通过域账户登录，可在所有控制节点上设置：

#vim /etc/openstack-dashboard/local\_settings

DOMAIN\_LOGIN\_ENABLE = True

#service httpd restart

## 系统优化选项

### 为10GB网络修改MTU到9000

* + - 1. 在控制节点上，执行ovs-vsctl show 查看 所有的 management 网络 是通过bond 网卡是哪几块,假设是 bond0 绑定了eth4，eth6，则 编辑修改/etc/sysconfig/network-scripts/下的 eth4,eth6，以及br-mgmt 在其后面添加配置项 MTU=9000
      2. 修改完后，执行 systemctl restart network。查看ovs是否自动识别到新的mtu值，执行 ovs-vsctl get int br-mgmt mtu

返回结果为9000,否则执行ovs-vsctl set int br-mgmt mtu=9000

由于控制节点没有接入storage网络，所以storage无需配置。

* + - 1. 在计算节点和ceph 节点上 执行 ovs-vsctl show同样查看网卡绑定关系。假设得到关系为 br-mgmt为bond0绑定了eth4,eth6；br-storage为bond1绑定了eth5,eth7， 因此，需要修改MTU=9000的网卡为 eth4,eth6,eth5,eth7 br-storage,br-mgmt，修改完后执行 systemctl restart network 同样执行 ovs-vsctl get int br-mgmt mtu 和ovs-vsctl get int br-storage mtu 来确认 网桥的MTU是否被修改。
      2. 执行完以上操作，下面需要验证是否生效。通过在host上执行ping命令来查看：

ping -I TARGET\_IFC -M do -s 8972 TARGET\_IP 其中 TARGET\_IFC 代表网卡而 TARGET\_IP表示目标IP

例如：ping -I br-mgmt -M do -s 8972 192.168.2.8

### 调整Ceph pg\_num

当ceph集群中仅少量osd提示full或near full,可能是pg\_num设置不合理造成的.ceph的数据存储结构应该都很容易查到。就是 file->object->pg->OSD->physics disk 。因此，一旦这里的pg数设置过小，pg到OSD的映射不均匀就会造成OSD上分配到的数据不均匀。这种额解决方法就是重新调整 pg\_num 和 pgp\_num 。

1. 执行如下命令查看ceph osd 上pg分布情况

*ceph pg dump | awk '*

*/^pg\_stat/ { col=1; while($col!="up") {col++}; col++ }*

*/^[0-9a-f]+\.[0-9a-f]+/ { match($0,/^[0-9a-f]+/); pool=substr($0, RSTART, RLENGTH); poollist[pool]=0;*

*up=$col; i=0; RSTART=0; RLENGTH=0; delete osds; while(match(up,/[0-9]+/)>0) { osds[++i]=substr(up,RSTART,RLENGTH); up = substr(up, RSTART+RLENGTH) }*

*for(i in osds) {array[osds[i],pool]++; osdlist[osds[i]];}*

*}*

*END {*

*printf("\n");*

*printf("pool :\t"); for (i in poollist) printf("%s\t",i); printf("| SUM \n");*

*for (i in poollist) printf("--------"); printf("----------------\n");*

*for (i in osdlist) { printf("osd.%i\t", i); sum=0;*

*for (j in poollist) { printf("%i\t", array[i,j]); sum+=array[i,j]; poollist[j]+=array[i,j] }; printf("| %i\n",sum) }*

*for (i in poollist) printf("--------"); printf("----------------\n");*

*printf("SUM :\t"); for (i in poollist) printf("%s\t",poollist[i]); printf("|\n");*

*}'*

1. 推荐的ceph pool pg num

参考: https://ceph.com/pgcalc/

每个pool的pg num的经验计算公式: (100\*OSD盘数)/(副本数\*pool数),向上取最接近的2的倍数. 注: 这里的pool数仅计算真实数据的pool.

1. 调整数据同步参数，减少数据同步时对业务的影响

当调整 PG/PGP 的值时，会引发ceph集群的 backfill 操作，数据会以最快的数据进行平衡，因此可能导致集群不稳定。 因此首先设置 backfill ratio 到一个比较小的值，通过下面的命令设置：

# ceph tell osd.\* injectargs '--osd-max-backfills 1'

# ceph tell osd.\* injectargs '--osd-recovery-max-active 1'

# ceph tell osd.\* injectargs '--osd\_recovery\_max\_single\_start 1'

注: ThinkCloud OpenStack中已预设为以上值

1. 平滑调整ceph pool的pg num

（1）检查pool的pg num, pgp num,及复制size，执行如下命令：

# ceph osd dump |grep size|grep volumes

pool 2 'volumes' replicated size 3 min\_size 2 crush\_ruleset 0 object\_hash rjenkins pg\_num 128 pgp\_num 128 last\_change 45 flags hashpspool stripe\_width 0

（2）使用上述公式，根据OSD数量、复制size、pool的数量，计算出新的PG数量，假设是1024.

（3）按2的倍数逐渐平滑增大pool的pg\_num和pgp\_num: 256->512->1024：

# ceph osd pool set volumes pg\_num 256

---------------等待pg创建完成---------------

# ceph osd pool set volumes pgp\_num 256

-------ceph -s查看待recover进度完成--------

# ceph osd pool set volumes pg\_num 512

---------------等待pg创建完成---------------

# ceph osd pool set volumes pgp\_num 512

-------ceph -s查看待recover进度完成--------

# ceph osd pool set volumes pg\_num 1024

---------------等待pg创建完成---------------

# ceph osd pool set volumes pgp\_num 1024

-------ceph -s查看待recover进度完成--------

（4）如果有其他pool，同步调整它们的pg\_num和pgp\_num，以使负载更加均衡。

## 系统运维常用操作

### 安全开关云平台

1. 通过Horizon web管理界面或者命令行关闭虚拟机

亦可在控制节点命令行模式下参考以下命令批量关机：

*for ins in `mysql -e "select uuid from nova.instances where vm\_state='active'"|awk '{print $1}'`;do if [[ "$ins" =~ "uuid" ]];then continue;fi;nova stop $ins;done*

注意：请确认虚拟机之间是否有特殊的关闭顺序要求。

1. 关闭Ceph OSD 节点

设置noout标记，防止因为Ceph节点断电引起重平衡

（1）登陆到Controller节点或者Ceph OSD节点，执行：

ceph osd set noout

（2）确认集群状态，可以看到noout 标记

# ceph -s

cluster 7eb2b84a-c73e-4050-b735-3ba643ffc603

health HEALTH\_WARN

noout flag(s) set

（3）登陆到每个Ceph OSD节点，执行：init 0

1. 关闭计算节点

登陆每一个计算节点，执行: init 0

1. 关闭控制节点

对3个控制节点，判断哪一台是主控制节点，首先关闭其他2台控制节点，最后关闭主控制节点，并且记住关闭的顺序，后面启动的时候需要相反的顺序启动：

（1）首先判断哪一台是主控制节点：若web ip为172.16.1.2, 则在控制节点上执行 ip -a|grep 172.16.1.3，有输出即为主控节点。

（2）登陆到其他两台控制节点，分别执行：

init 0

每个节点关闭完成，多等几分钟，使集群有充足的时间在剩余的控制节点之间重新分配服务。

（3）最后登陆到controller01节点，执行：init 0

至此，集群已经安全关闭。

重启集群过程正好相反，先启动控制节点，再启动Ceph节点，最后启动计算节点，启动计算节点后再登录web管理界面将相应的虚拟机启动。

需要特别注意的是，启动控制节点的时候，按照关闭相反的顺序进行启动。

### 通过命令行上传大镜像文件

当镜像文件超大时，通过Horizon web可能上传不成功，可尝试通过下面的命令行上传：

# source /root/openrc.v2

# glance image-create --disk-format raw --container-format bare --disk-format qcow2 --name centos7 --is-public True --file CentOS7\_all\_tools.qcow2 –progress

### 替换Ceph OSD数据盘

1. 所有控制节点上停掉以下服务:

*#systemctl stop storagemgmt-agent*

*#systemctl stop openstack-gnocchi-metricd*

*#systemctl stop openstack-gnocchi-statsd*

1. 在某一控制节点上操作Ceph进入维护模式

*#**ceph osd set noout*

1. 在对应的ceph osd节点上移除待替换的osd

*#**/etc/init.d/ceph osd -a stop osd.x*

*#**ceph osd down osd.x*

*#ceph osd out osd.x*

*#ceph osd crush rm osd.x*

*#ceph auth del osd.x*

*#ceph osd rm osd.x*

注: 需要等待 ceph -s 为HEALTH\_OK再继续剩余步骤。

1. 检查待替换的盘上是否存在软RAID，若有，删除之

#mdadm --detail /dev/md0

#mdadm /dev/md0 --fail /dev/sdX3 --remove /dev/sdX3

1. 替换新盘并格式化

*#ceph-disk zap /dev/sdx*

*或者*

*#parted /dev/sdx*

*#>mklabel gpt*

可用ceph-disk list看状态

1. 在某一控制节点上操作创建ceph osd

*#ceph-deploy osd create node-x:/dev/sdx:/dev/sd?*

冒号前面为osd分区, 后面为日志分区(可选)，如ceph-deploy osd create node-5:/dev/sdc:/dev/sde7

1. 检查新的osd盘是否自动挂载

登录osd节点, df -h查看新的ceph-osd.x分区是否自动挂载, 若无,如下方式手动挂载:

*#mount /dev/sdx1 /var/lib/ceph/osd/ceph-x （注: 若替换ceph osd盘, 确保之前的分区先umount）*

1. 检查新的ceph osd分区服务是否启动

登录osd节点, systemctl list-units|grep osd.x.查看是否有相关服务running；若无，执行

*#/etc/init.d/ceph -a start osd.x*

*若执行ceph –a start osd.x时提示找不到osd.x, 可执行下面的命令激活此osd:*

*#* *ceph-disk activate --mark-init sysvinit --mount /dev/sdx1*

1. 确认以下信息:
   * ceph -s 查看ceph状态, 确保集群的状态为HEALTH OK
   * 所有PG都是 active + clean
   * 所有OSD都是 up 状态
2. 退出维护模式

*#ceph osd unset noout*

1. 恢复停掉的服务

*#systemctl restart storagemgmt-agent*

*#systemctl restar openstack-gnocchi-metricd*

*#systemctl restar openstack-gnocchi-statsd*

注: 若替换多张盘，必须one by one重复步骤3~9

## 更新LOE许可证

更新许可证方法如下：

1. 在Carrier节点上，利用上传License的方式上传。
2. 在所有控制节点上：
3. 使用下列命令更新license。

*wget http://${master\_ip}:8080/licenses/tcos.lic -O /etc/lenovo/tcos.lic*

1. 重启keystone服务:

*#service openstack-keystone restart*

1. 重启httpd服务:

*#service httpd restart*