CAS存储模块详解（上）

——OCFS2文件系统分布式锁介绍

2020.8

课程目标

学习完本课程，您应该能够：

- 了解OCFS2文件系统结构
- 掌握OCFS2分布式锁机制
- 了解OCFS2相关问题

2

目录

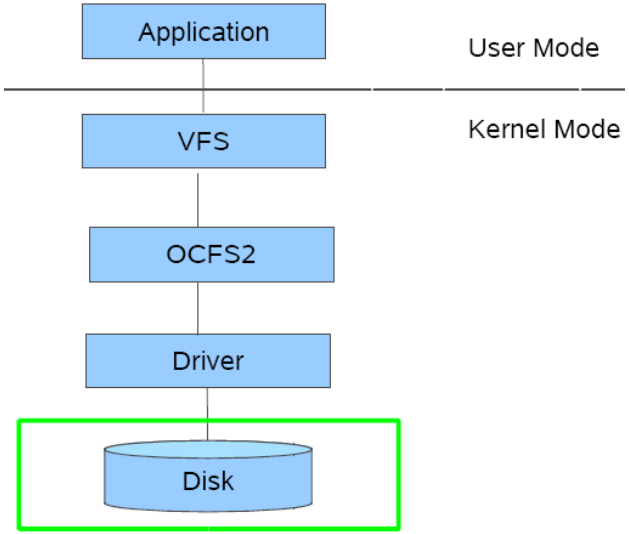
3

OCFS2文件系统

OCFS2：The Oracle Clustered File System, Version 2

OCFS2是一种可以跨节点共享的开源分布式文件系统，仅适用于Linux操作系统。

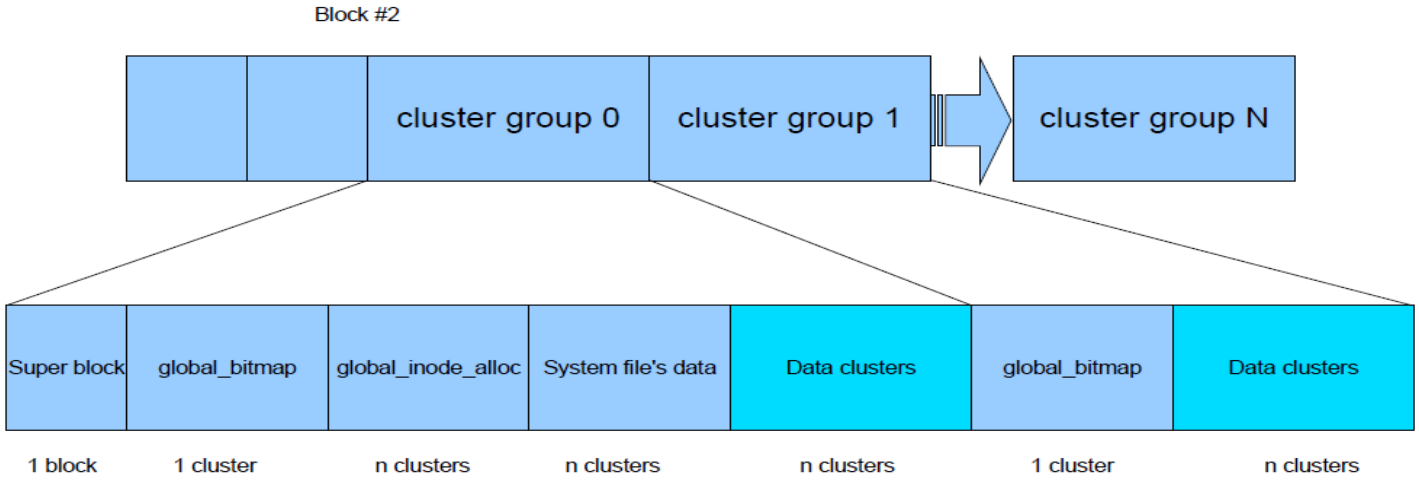
OCFS2对于虚拟机的HA，必不可少。



OCFS2文件系统结构

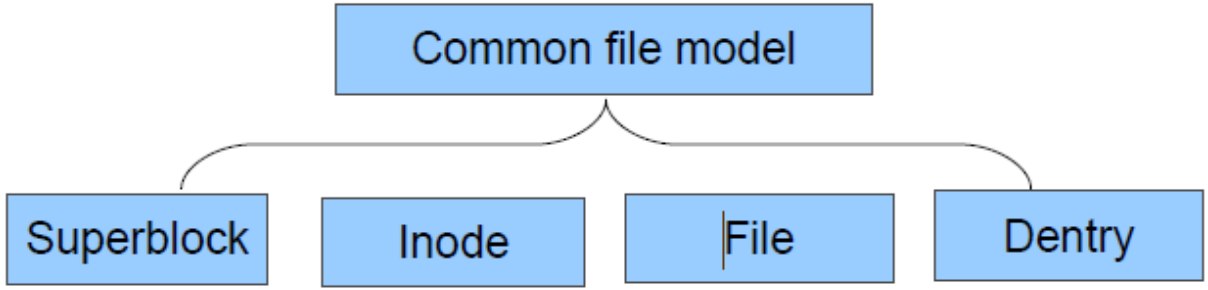
block：512B~4KB，存放元数据的最小单位。

cluster：4KB~1MB，存放数据的最小单位。



VFS——虚拟文件系统

VFS脱胎于EXT2，提供通用接口供系统调用，系统不用关心底层的存储介质和文件系统类型就可以工作。
系统调用接口：
mount/umount/mkdir/symlink/chown/stat/chmod/open/close/select/read/write/truncate



OCFS2文件系统结构

```
[root@UIS27 default]# hexdump -C /dev/dm-15 |more
00000000  02 02 02 02 02 02 02 02  74 68 69 73 20 69 73 20  |........this is |
00000010  61 6e 20 6f 63 66 73 32  20 76 6f 6c 75 6d 65 00  |an ocfs2 volume.|
00000020  02 02 02 02 02 02 02 02  02 02 02 02 02 02 02 02  |................|
*
00000080  02 02 02 02 02 02 02 02  74 68 69 73 20 69 73 20  |........this is |
00000090  61 6e 20 6f 63 66 73 32  20 76 6f 6c 75 6d 65 00  |an ocfs2 volume.|
000000a0  02 02 02 02 02 02 02 02  02 02 02 02 02 02 02 02  |................|
*
00000230  74 68 69 73 20 69 73 20  61 6e 20 6f 63 66 73 32  |this is an ocfs2|
00000240  20 76 6f 6c 75 6d 65 00  02 02 02 02 02 02 02 02  | volume.........|
00000250  02 02 02 02 02 02 02 02  02 02 02 02 02 02 02 02  |................|
*
00000280  02 02 02 02 74 68 69 73  20 69 73 20 61 6e 20 6f  |....this is an o|
00000290  63 66 73 32 20 76 6f 6c  75 6d 65 00 02 02 02 02  |cfs2 volume.....|
000002a0  02 02 02 02 02 02 02 02  02 02 02 02 02 02 02 02  |................|
*
```

前面8192个字节是OCFS2的特殊写入， 开始4096个字节，填充ocfs2的信息，在mount的时候，系统会校验。

```
00002000  4f 43 46 53 56 32 00 00  86 65 a1 1a ff ff ff ff  |OCFSV2...e......|
00002010  00 00 00 00 02 0c 3e 00  00 00 00 00 00 00 00 00  |......>.........|
00002020  00 00 00 00 00 00 00 00  00 00 00 00 31 00 00 00  |............1...|
00002030  00 00 00 00 00 00 00 00  e3 08 fb 5e 00 00 00 00  |...........^....|
00002040  e3 08 fb 5e 00 00 00 00  00 00 00 00 00 00 00 00  |...^............|
00002050  02 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
00002060  86 65 a1 1a 00 00 00 00  00 00 00 00 00 00 00 00  |.e..............|
00002070  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
*
000020a0  00 00 00 00 ff ff 01 00  00 00 00 00 00 00 00 00  |................|
000020b0  01 04 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
000020c0  00 00 5a 00 00 00 14 00  00 00 00 00 00 00 00 00  |..Z.............|
000020d0  e3 08 fb 5e 00 00 00 00  00 00 00 00 03 00 00 00  |...^............|
000020e0  50 37 01 00 01 00 00 00  01 02 00 00 00 00 00 00  |P7..............|
000020f0  02 02 00 00 00 00 00 00  0c 00 00 00 14 00 00 00  |................|
```

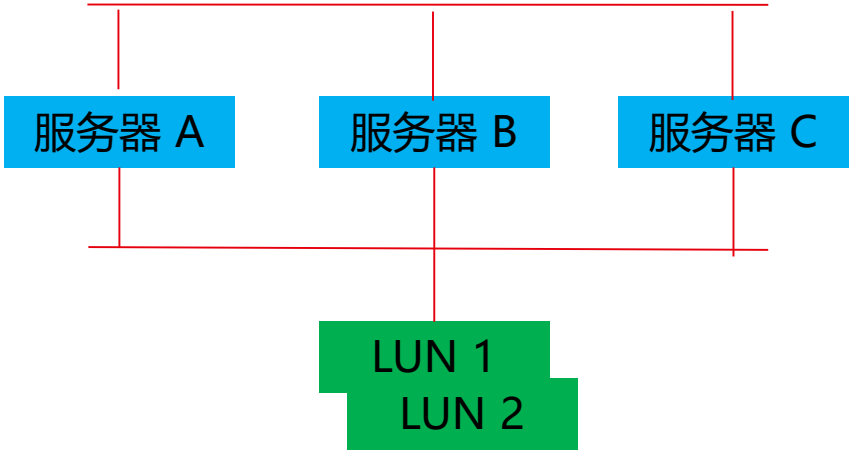
Super block在4096字节开始位置的一个block，同时在1G, 4G, 16G, 64G, 256G and 1T位置处有super block的备份。

目录

8

共享文件系统

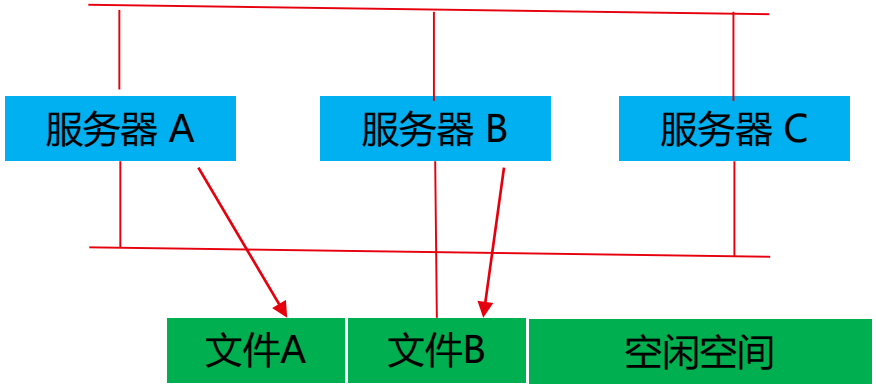
- 虚拟机迁移需要有共享存储，多个主机要能同时访问存储
- 共享存储提供的块/LUN，需要格式化为文件系统才能方便访问，就像是服务器的本地磁盘需要格式化为EXT4文件系统
- EXT4等常见的文件系统都是本地文件系统，不能同时被多个主机mount
- 共享文件系统/集群文件系统才能支持被多个主机同时mount使用，每个主机看到的内容相同



9

需加锁进行互斥

- 共享文件系统很多操作需要加锁进行互斥保护，防止多个服务器同时访问一个资源，造成冲突
- 加锁的操作可以通过DLM分布式锁管理实现，在访问需加锁的资源前，通过网络发一个消息给所有的活动服务器申请加锁，申请成功后才能进行访问
- 访问结束后，通过网络发一个消息给所有的服务器，通知已解锁
- 加锁到解锁的过程中，其他的服务器不能访问被锁的资源



举例：当文件A、B扩展空间时，需加锁才能从空闲空间申请新空间，否则可能申请到同一块空间

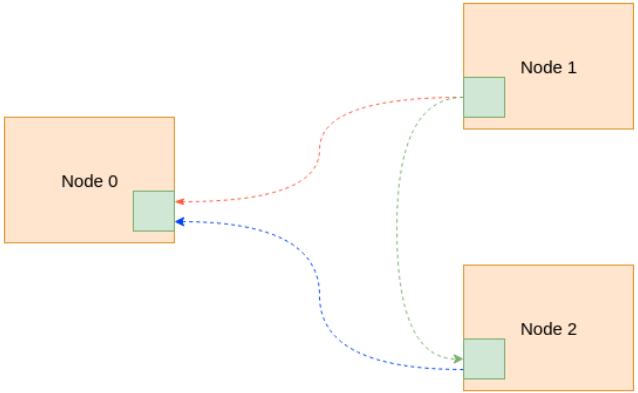
集群下的OCFS2文件系统

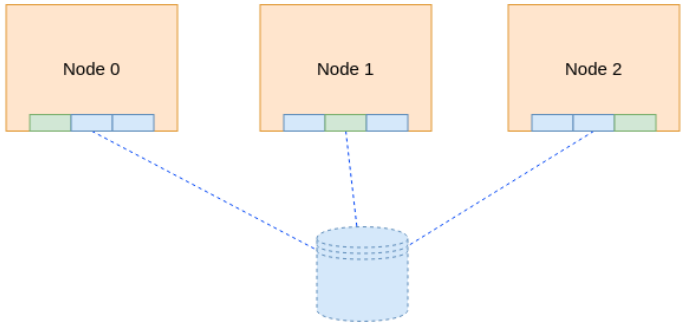
网络心跳

在同一个集群中，每个节点会根据配置文件中的节点信息向集群中其它节点发起连接请求，从而建立其网络通信的连接。任意两个节点之间都会建立连接。

磁盘心跳

OCFS2每隔2S更新磁盘上的时间戳，并且读取其他节点的时间戳。





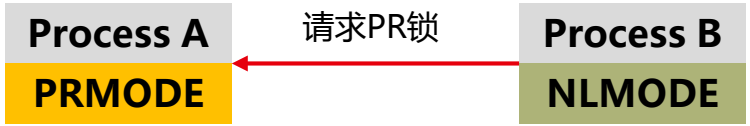
分布式锁

分布式锁是在集群环境下实现多个节点协同的，确保被访问的资源不会因为并发而出现不一致的情况。

Name	Access Type	Compatible Modes
EXMODE	Exclusive	NLMODE
PRMODE	Read Only	PRMODE, NLMODE
NLMODE	No Lock	EXMODE, PRMODE, NLMODE

被锁的项被定义为锁资源，在一个应用第一次请求对资源进行加锁时，锁管理器会创建一个锁资源，锁资源是与实际资源相关联的。需要注意的是，一个锁资源可能会与多个锁相关联，但一个锁只能与一个锁资源关联。

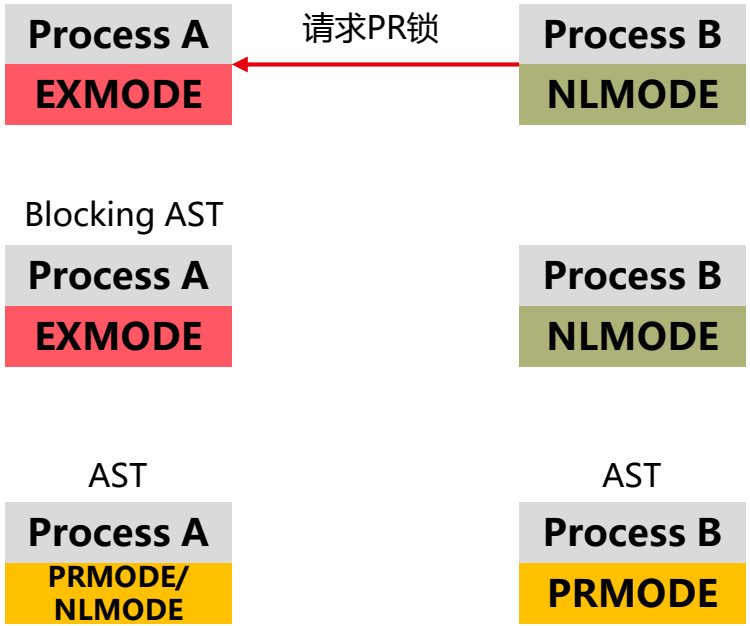
分布式锁



A进程持有PR锁，无锁的B进程请求PR锁

发现PR锁和PR锁可以兼容，直接获取PR锁
调用AST回调函数，返回结果

分布式锁



A进程持有EX锁，B进程请求PR锁

调用BAST函数触发A进程降级

调用AST回调函数，返回结果：
A进程降级完成
B进程获取到锁

目录

15

Fence机制

什么是FENCE

FENCE 是集群系统的一种用来隔离故障节点或者保护共享资源的机制。FENCE实现有杀死对方和自杀两种，CAS共享文件系统采用了后者。

CAS共享文件系统为什么要FENCE

当某一节点到共享LUN的访问故障或者该节点无法参与到集群的加锁/解锁流程中，这个节点就是集群系统里的故障节点（malfunctioning）。这样的节点必须被隔离，否则该节点将影响集群中的其他成员工作，并有可能损坏共享资源。

Fence的触发条件

管理网异常引起fence

当节点间网络故障，锁消息无法继续传递，节点加解集群锁阻塞，此时需要隔离出部分节点，保证集群可以继续工作。

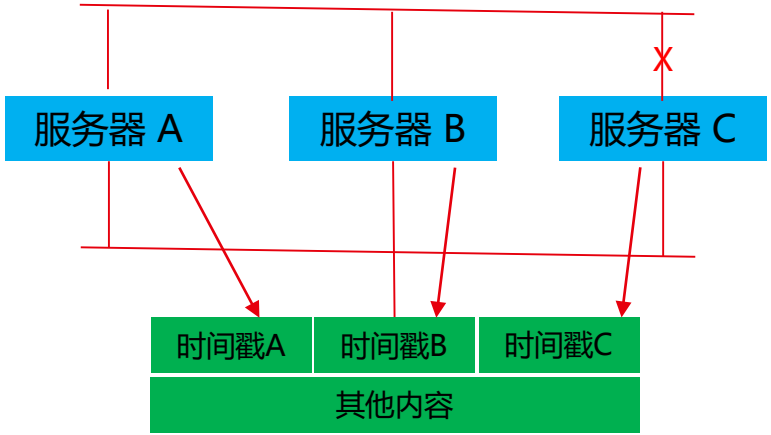
每个节点每10秒向其他节点发送保活(keep-alive)消息，当一个节点和另外一个节点90秒没有保活消息，开始启动一个126秒的定时器，在定时器超时前如果仍没有新的保活消息，则开始启动quorum判决，此时CVK会计算本节点到其他节点的连接位图，并和正在做心跳的位图做比较计算是否FENCE本节点。以此，保证集群可以继续工作。

如何判断是否fence：

1）心跳节点数为奇数：tcp连接数 < (心跳数 + 1) /2

2）心跳节点数为偶数：tcp连接数 < 心跳数/2 或 tcp连接数==心跳数/2 且最小节点号不通

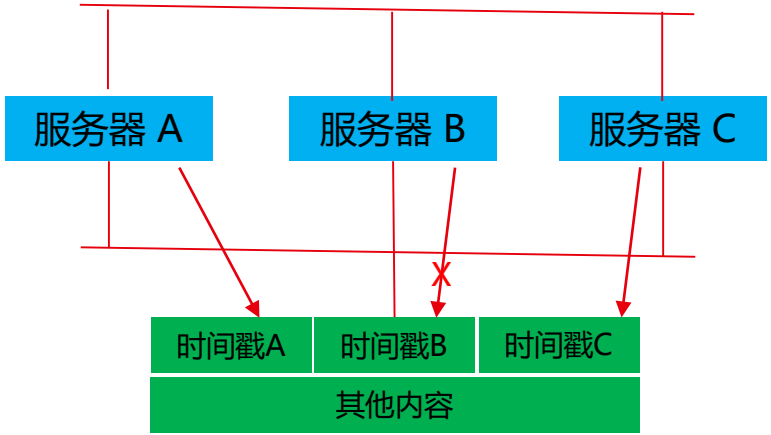
分布式锁机制是限制集群规模的重要因素。



Fence的触发条件

存储异常引起fence

- 每个成员（CVK）每2秒向共享磁盘指定心跳区域写时间戳。
- 每个成员（CVK）每2秒将整个心跳区域读到内存。并对比每个节点前后两次的时间戳是否有变化。
- 在上述读写磁盘心跳的过程中，启动一个120秒的定时器（(o2hb_dead_threshold - 1) * 2），在定时器超时后触发FENCE（FENCE自己）。
- 其他可以正常读写的节点，如果61次没有读到一个节点的心跳时间戳变化，就认为这个节点已经被故障隔离，开始抢占其持有的共享资源。



Fence动作优化

重启服务器

- 早期实现，简单粗暴，适用于单一业务，只有一个业务，即使重启也不影响其他业务
- 不适合虚拟化场景，因每个服务器通常挂多个LUN，一个LUN异常会影响所有LUN

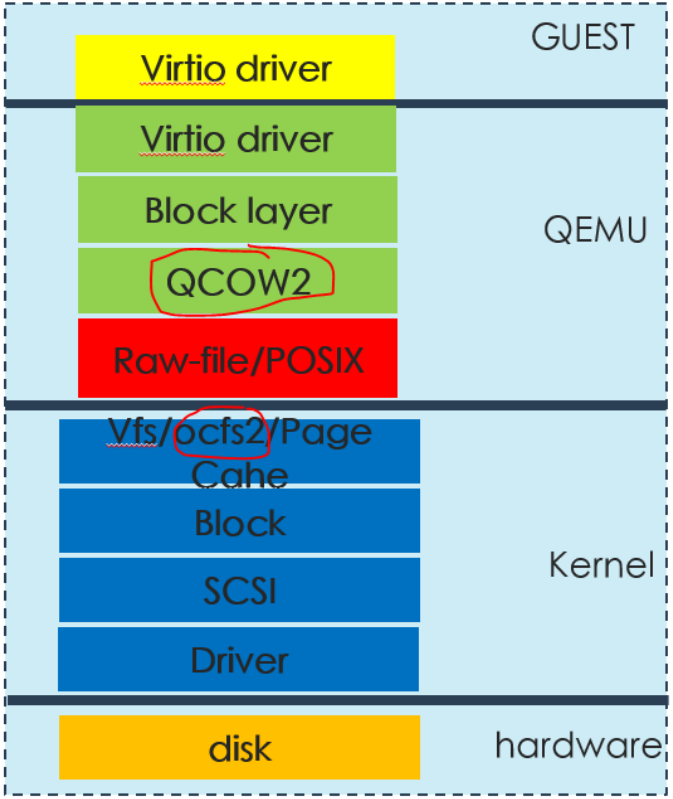
umount解挂

- 只对异常的LUN进行解挂，加挂前将虚拟机关闭
- 影响使用异常LUN的虚拟机，支持的LUN上的虚拟机不受影响
- 等存储恢复后，将LUN挂载，启动关闭的虚拟机

冻结虚机业务

- 将异常的LUN上虚拟机冻结，虚拟机进程仍在在
- 等存储恢复后，将冻结的虚机恢复

OCFS2文件系统对性能的影响



- 虚拟机内部的IO相比存储上的IO，会额外多出很多开销，其中占主要的是QCOW2层和ocfs2层。
- 对于高IO的业务虚拟机，建议使用直接挂载裸卷的方式，不通过QCOW2层和ocfs2层。

DLM加锁阻塞

当集群下某些节点间管理网闪断、存储链路闪断，但又不满足fence条件的异常情况时，DLM锁转换会被阻断，拿锁请求无法满足，进而导致文件操作就会阻塞。

当前CAS版本存在节点通过libvirt定时刷新存储池，若该请求阻断会导致该节点libvirt卡住，进一步CAS管理界面的主机及虚机状态异常。

判断是否存在DLM加锁阻塞：

```
    inode of lock: M000000000000002f409f000000000 is 0000000002f409f0, file is:
        49547760        /volume-5241d051-1754-4d27-a368-1f18eb5b93e6
    lock: M000000000000002f409f000000000 on local is:
Lockres: M000000000000002f409f000000000   Owner: 2    State: 0x8 Dirty
Last Used: -443648170  ASTs Reserved: 0    Inflight: 0    Migration Pending: No
Refs: 19   Locks: 16   On Lists: Dirty
Reference Map: 1 3 4 5 6 7 8 9 10 11 12 13 14 15 16
 Lock-Queue  Node  Level  Conv  Cookie          Refs  AST  BAST  Pending-Action
 Granted     4     NL     -1    4:7913850       2     No   No    None
 Granted     5     NL     -1    5:7912565       2     No   No    None
 Granted     15    NL     -1    15:7894330      2     No   No    None
 Granted     10    NL     -1    10:7892366      2     No   No    None
 Granted     13    NL     -1    13:7892240      2     No   No    None
 Granted     14    NL     -1    14:7893636      2     No   No    None
 Granted     12    NL     -1    12:2055548      2     No   No    None
 Granted     9     NL     -1    9:7889442       2     No   No    None
 Granted     11    NL     -1    11:7890964      2     No   No    None
 Granted     6     NL     -1    6:7909203       2     No   No    None
 Granted     3     NL     -1    3:7912349       2     No   No    None
 Granted     16    NL     -1    16:1236593      2     No   No    None
 Granted     7     NL     -1    7:6760856       2     No   No    None
 Granted     1     NL     -1    1:7916091       2     No   No    None
 Granted     8     PR     -1    8:7892416       2     No   No    None
 Converting  2     PR     EX    2:6156465       2     No   No    None

    Local host is the Owner of M000000000000002f409f000000000
```

Thanks!

新华三集团
www.h3c.com