Cinder创建卷失败原因及解决方案分析

***背景：***

Cinder后端存储的SVM配置了多个Pool时，如果Nova从镜像批量创建VM有概率性块设备映射失败。

***问题根因分析：***

原因一：Pool已经达到LUN数目上限

c-sch指定的Pool已经达到LUN数目上限，无法创建出新的LUN，当c-vol调用后端driver创建LUN时，抛出如下异常：

NaApiError: NetApp API failed. Reason - 9158:Clone operation failed to start: Number of LUNs per volume exceeds maximum limit..

从目前来看解决问题有如下困境：

1. NetApp后端的Pool中LUN的数目固定为512，不可修改或配置
2. NetApp没有提供专门的接口统计每个Pool中当前的lun数目，cinder需要自己通过代码实现。

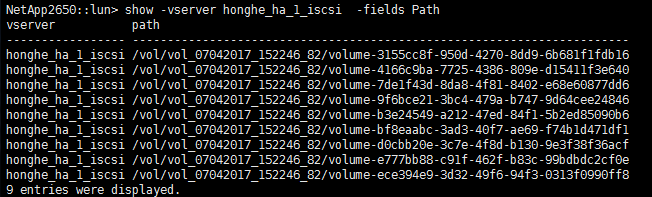
解决方案：需要添加Filter，过滤一些水位线过高的Pool

方案一：

cinder-volume获取整个SVM下的所有的lun，根据lun的path名称来完成统计，并上报给cinder-scheduler。

Lun的path格式如下：/vol/vol\_07042017\_152246\_82/volume-9f6bce21-3bc4-479a-b747-9d64cee24846，显然中间字段即是pool的名称。

在ontap系统上使用lun show命令可以返回所有的lun信息。



netapp的驱动同样提供相应的接口方法：

cinder. Volume.drivers.netapp.dataontap.client.client\_base.get\_lun\_list

cinder. Volume.drivers.netapp.dataontap.client.client\_base.get\_lun\_by\_args

方案二：

cinder-scheduler直接统计volumes表中的HOST字段，以此作为Pool中lun数目的依据。该方案假定，只有cinder使用pool。统计信息可能不太准确。

原因二：c-sch指定的Pool与Image的location不一致

c-sch指定的Pool与Image的location不是同一个，c-volume判断Image镜像不存在，并重新下载Image到本地，执行转换后再上传到后端。较大的镜像可能会因为该过程时间太长，导致nova判断创建磁盘失败。

目前考虑了下面几个规避、解决方案

方案一：修改nova创建卷的等待时间，目前是180秒。

相关参数：

block\_device\_allocate\_retries\_interval = 3

block\_device\_allocate\_retries = 60

方案二：修改cinder-volume侧代码，当出现这种情况时走卷Copy流程，而不是下载流程。

方案存在的问题：

Ontap后端实现了copy的控制台命令，可以跨卷进行lun拷贝：

lun copy start

Usage:

[-vserver] <vserver name>                       Destination Vserver Name

[-destination-path] <path>                       Destination Path

[-source-path] <path>                           Source Path

[[-source-vserver] <vserver name>]             Source Vserver Name

[ -promote-early [true] ]                        Promote Early

[ -max-throughput {<integer>[KB|MB|GB|TB|PB]} ]  Maximum Transfer Rate (per sec)

但是，Netapp驱动似乎没有提供Copy的接口。

Netapp驱动提供了move\_lun这个接口，但是好像cinder没有调用这个接口。

同时，migrate命令走的不是move\_lun，而走clone接口，如果clone不成功，会直接创建一个卷，挂载并进行文件拷贝。这就导致了跨卷迁移的速度特别慢（上面这个流程没有仔细分析，有待验证！）。

方案三：利用Glance的multi-location特性

在创建Image时准备多个location（每个pool一个），当后端增加pool时，cinder需要将所有的image在新增加的这个pool进行拷贝，并且通知glance添加一个location。

方案四：利用Glance的multi-location特性，并且动态的为image添加location

方案二、三的结合，c-sch可以添加两个Filter。

*Filter1：检查Lun数目未满的pool；*

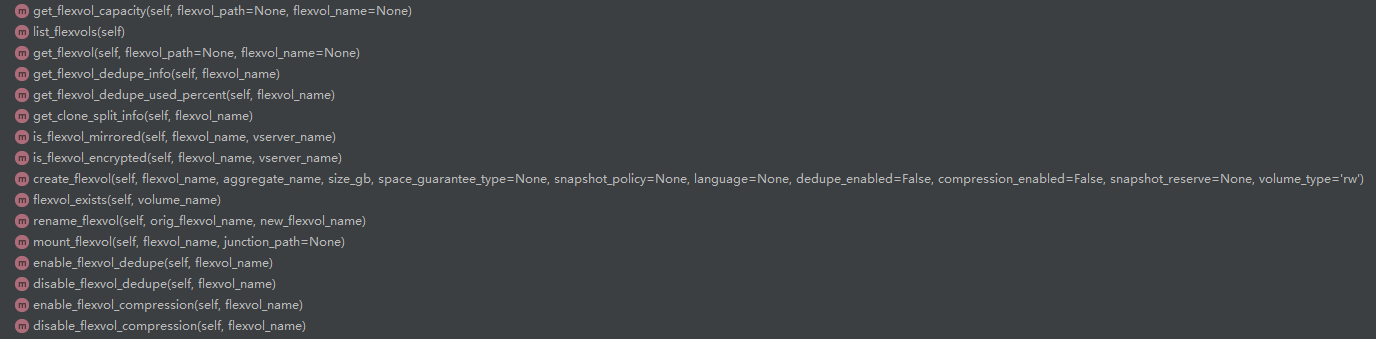
*Filter2：根据image的location信息过滤，找到合适的HOST。*

*这里Filter2需要处理这样一种情况：image所有的location对应的Pool全都已经满了，我们需要为image添加一个新的location。这里我们可以从未满的Pool中挑一个，执行migrate，或者直接clone一个pool，这个pool是上传镜像时准备好的，里面只有一个lun。****上面两种方案相比，migrate相当的慢，原因是跨pool迁移时是逐个文件复制的，并不是走的后端copy（需要确认！！），而clone pool的方式几乎依然是零拷贝，但是虽然netapp提供了驱动，实现起来似乎工作量颇大！***

由于cinder没有批量建卷的能力，而nova可以批量创建虚拟机，因此nova创建10虚拟机会调用10次cinder的接口，如果按照上面的方案可能会导致竞争，使image被添加了多个location。

附：

**1.**cinder.volume.drivers.netapp.dataontap.client.client\_cmode. Client提供的池操作接口：



2.手工规避问题的操作方案

调整后端pool的大小，保证c-sch不会调度该卷。

手工添加image的location：

手工操作步骤：

#注意创建时添加的metadata需要参考原镜像卷，下面列出的都是必须项

cinder create --name <target-volid> --source-volid <source-id> --metadata image\_owner=<image\_owner> readonly=True image\_size=<image-size> glance\_image\_id=<image-id> attached\_mode=ro

cinder migrate <target-volid> <target-pool>

glance location-add <image-id> --url cinder://<target-volid>