

Mars 200 檔案系統

Mars 200的檔案系統CephFS是相容於 POSIX標準的檔案系統。他使用Ceph儲存叢集為儲存池來達到分佈式高可用度(high availability)。

它具有以下特性

- 橫向擴展：可隨需求擴充或是縮小檔案系統容量。支援單一檔案系統容量大於數個petabytes (1024TB)。
- 檔案分享：具有檔案鎖定機制。支援同時多使用者同時讀寫，並可維持資料的一致性。
- 高可用性：CephFS使用多個metadata server (MDS)所形成的叢集。如果Active MDS失效後，其他的備援MDS可以成為active. 因此維持用戶的讀取，而不受影響。
- 用戶端配額管理：使用CephFS 的FUSE客戶，可以為其使用的子目錄設置的使用容量配額(Quota)
- 安全性： CephFS使用金鑰來管理使用者的認證。沒有金鑰的使用者無法進入子目錄取用檔案。具有不同金鑰的使用者無法存取其他使用者的目錄。並且可以限制使用者的存取權限。

CephFS 設定與管理

第一章: 流程

1.1 管理主機與管理者設定

- 1) 準備一台管理主機，此主機可以是虛擬機或是實體機，作業系統為Centos. 並安裝 ceph-fuse 套件(package)
- 2) 利用scp從Mars 200任一MON節點，將管理者client.admin的keyring 以及ceph.conf 拷貝(copy)到管理主機
- 3) 在管理主機上設立掛載目錄 (mount point)，利用Ceph admin帳號 將CephFS的根目錄掛載於管理主機
- 4) 設定開機時自動掛載Ceph Filesystem

1.2 創建及管理使用者帳號

- 5) 在管理主機之掛載目錄下創建用戶主機所需要的Ceph FS子目錄，這些子目錄要為分配用戶之工作目錄
- 6) ssh 至Mars 200之任一MON 節點, 創建Ceph用戶帳號, 設定存取權限, 並產生金鑰

1.3 用戶主機設定

- 7) 在用戶主機VM上安裝ceph-fuse 套件
- 8) 將Ceph用戶帳號金鑰以及ceph.conf 拷貝到用戶主機VM
- 9) 在用戶主機VM上設立掛載目錄 (mount point)
- 10) 利用Ceph用戶帳號 將CephFS的目錄掛載於用戶主機
- 11) 設定自動掛載Ceph Filesystem 並且確認
- 1.4 管理主機上設定/修改每個CephFS用戶子目錄的配額(cephfs quota)
- 12) 利用admin帳號在管理主機上設定/修改每個子目錄的配額

詳細設定方式參見以下第二章

- 13) 如何unmount, delete user
增加新的用戶主機重複5) 到 12)

第二章: Step by Step

2.1 管理主機與管理者設定

2.1.1 安裝 ceph-fuse 套件

準備一台管理主機，此主機可以是虛擬機或是實體機，作業系統為Centos 7
Mars 200的CephFS支援內核模組(kernel module)與用戶空間(Filesystem in Userspace)兩種方式，目前只有FUSE支援以目錄為基礎的Quota管理

首先須下載安裝ceph-fuse 套件

```
yum install ceph-fuse
```

或事先將需要之套件ceph-fuse-pkg.tgz存放於local directory

```
tar -zxvf ceph-fuse-pkg.tgz
```

```
cd ceph-fuse-pkg
```

```
yum install `ls | grep rpm`
```

2.1.2 利用scp從Mars 200任一MON節點，將管理者client.admin的keyring 以及ceph.conf 拷貝(copy)到管理主機

在/etc/下創建新的子目錄 /etc/ceph

```
mkdir -p /etc/ceph
```

```
scp root@192.168.14.10:/etc/ceph/ceph.client.admin.keyring /etc/ceph/
```

```
scp root@192.168.14.10:/etc/ceph/ceph.conf /etc/ceph/
```

確認金鑰檔與設定檔都有正確的權限設定

```
chmod 644 /etc/ceph/ceph.conf
```

```
chmod 644 /etc/ceph/ceph.client.admin.keyring
```

2.1.3利用Ceph admin帳號 將Ceph Filesystem的根目錄掛載於管理主機
先在管理主機上創建掛載點

```
mkdir -p /mnt/cephfs
```

掛載

```
ceph-fuse -k=/etc/ceph/ceph.client.admin.keyring /mnt/cephfs -r /
```

使用 "mount" 指令檢查 /mnt/cephfs是否掛載成功

```
使用 "fusermount -u /mnt/cephfs" 指令測試卸載 /mnt/cephfs
```

2.1.4設定開機時自動掛載Ceph Filesystem

在 /etc/fstab 中加入以下內容

```
id=admin /mnt/cephfs fuse.ceph defaults,_netdev 0 0
```

使用 "mount -a" 指令驗證 /etc/fstab 可以用來mount the Ceph file system

重開機以確認是否可自動掛載

2.2 創建及管理使用者帳號

2.2.1 在管理主機創建使用者目錄

在管理主機之掛載目錄下創建用戶所需要的子目錄，這些子目錄要為分配用戶之工作目錄

```
mkdir -p /mnt/cephfs/home/user1/volume1
```

建議目錄架構如下：

2.2.2 在Ceph Monitor 節點(192.168.14.10)創建CephFS 使用者，並且限制其只能存取 /home/user1/volume1

```
ssh root@192.168.14.10
```

```
ceph auth get-or-create client.x.user1 -o /etc/ceph/client.x.user1.keyring
```

```
ceph auth caps client.x.user1 mon 'allow r' mds 'allow r, allow rw
```

```
path=/home/user1/volume1' osd 'allow rw pool=cephfs_data'
```

Note: 設定後可使用 "ceph auth list"指令檢查設定，
可使用"ceph auth del client.x.user1" 刪除使用者。

2.3 用戶主機設定

2.3.1在用戶主機VM上安裝 ceph-fuse 套件

首先須下載安裝ceph-fuse 套件

```
yum install ceph-fuse
```

或事先將需要之套件ceph-fuse-pkg.tgz存放於local directory

```
tar -zxvf ceph-fuse-pkg.tgz
```

```
cd ceph-fuse-pkg
```

```
yum install `ls | grep rpm`
```

2.3.2 利用scp從Mars 200 MON節點(192.168.14.10)，將對應於此用戶主機的keyring 以及 ceph.conf 拷貝(copy)到此用戶主機

在/etc/下創建新的子目錄 /etc/ceph

```
mkdir -p /etc/ceph
```

```
scp root@192.168.14.10:/etc/ceph/ceph.client.x.user1.keyring /etc/ceph/
```

```
scp root@192.168.14.10:/etc/ceph/ceph.conf /etc/ceph/
```

確認金鑰檔與設定檔都有正確的權限設定

```
chmod 644 /etc/ceph/ceph.client.x.user1.keyring
```

```
chmod 644 /etc/ceph/ceph.conf
```

2.3.3利用client.x.user1帳號 將Ceph Filesystem的目錄掛載於管理主機

先在用戶主機上創建掛載點

```
mkdir -p /mnt/ volume1
```

掛載

```
ceph-fuse -n client.x.user1 -k=/etc/ceph/ceph.client.x.user1.keyring /mnt/volume1 -r /home/user1/volume1
```

使用 "mount "指令檢查 /mnt/volume1是否掛載成功

使用 "fusermount -u /mnt/volume1" 指令測試卸載 /mnt/volume1

2.3.4設定開機時自動掛載Ceph Filesystem

在 /etc/fstab 中加入以下內容

```
id=x.user1,client_mountpoint=/home/user1/volume1 /mnt/volume1 fuse.ceph defaults,_netdev 0 0
```

使用 "mount -a" 指令驗證 /etc/fstab 可以用來mount the Ceph file system

重開機以確認是否可自動掛載

2.4 管理主機上設定/修改每個CephFS用戶子目錄的配額(cephfs quota)

2.4.1在用戶主機VM上安裝 attr package 套件

如果attr package尚未安裝，請先安裝

```
yum install attr
```

或事先將需要之套件attr-2.4.46-12.el7.x86_64.rpm存放於local directory

```
rpm -ivh attr-2.4.46-12.el7.x86_64.rpm
```

2.4.2設定quota

設定 quota #100 MB

```
setfattr -n ceph.quota.max_bytes -v 100000000 /mnt/cephfs/home/user1/volume1
```

或是 設定檔案數 #10,000個檔案

```
setfattr -n ceph.quota.max_files -v 10000 /mnt/cephfs/home/user1/volume1
```

2.4.3取得Quota 設定

```
getfattr -n ceph.quota.max_bytes /mnt/cephfs/home/user1/volume1
```

```
getfattr -n ceph.quota.max_files /mnt/cephfs/home/user1/volume1
```

2.4.4取消 Quota 設定

```
setfattr -n ceph.quota.max_bytes -v 0 /mnt/cephfs/home/user1/volume1
```

```
setfattr -n ceph.quota.max_files -v 0 /mnt/cephfs/home/user1/volume1
```

附註：

QUOTAS

CephFS允許在系統中的任何目錄上設置配額。配額可以限制存儲在目錄層次結構中該點之下的用量大小(GB)或檔案數。

限制

- 配額是合作性的、非對抗性的。CephFS 的配額功能依賴於掛載它的客戶端的合作，在達到上限時要停止寫入；無法阻止篡改過的或者對抗性的客戶端，它們可以想寫多少就寫多少。在客戶端完全不可信時，用配額防止多佔空間是靠不住的。
- 配額不精確。正在寫入文件系統的進程將在達到配額限制後的短時間內停止。它們將不可避免地被允許在配置的限制上寫入一些數據量。他們能夠超過的配額有多遠，主要取決於時間量，而不是數據量。一般來說，作者將在超過配置限制的

十秒鐘內停止。

- 配額尚未在內核客戶端中支援。配額由用戶空間客戶端（libcephfs，ceph-fuse）支持，但尚未在Linux內核客戶端中支援。
- 基於路徑限制(path restriction) 掛載時必須謹慎地配置配額。客戶端必須能夠訪問配置了配額的那個目錄的索引節點，這樣才能執行配額管理。如果某一客戶端被MDS 能力限制成了只能訪問一個特定路徑（如/home/user ），而配額是設置於它們無權訪問的父目錄（如/home），這個客戶端就不會被配額限制。所以，基於路徑做訪問控制時，請確定在限制了客戶端的那個目錄（如 /home/user）、或者它下面的子目錄上設置配額。

設置 Quota

Ceph配額是用虛擬擴展屬性(Virtual Extended Attribute)來配置的：

ceph.quota.max_files - file limit

ceph.quota.max_bytes - byte limit (單位: Byte)

如果這些屬性出現在目錄索引節點(inode)裡，就意味著那裡設置了配額；如果不存在，那麼那個目錄就沒設置配額（然而父目錄仍然有可能設置了Quota）。

