

修訂紀錄：

版次	日期	變更內容	變更者
0.1	2019/03/05	- 初版	李金泓
1.0	2019/03/08	- 修改用字、改為網頁排版	魏嘉玟

Lustre 檔案系統使用說明

版次：20190308 v1.0

台灣杉一號主機主要有提供 /home, /work1 及 /project 三個 Lustre 平行檔案系統。這三個 Lustre 平行檔案系統在建置初期已有不同的使用規劃，本文件在此提供更詳細操作解說給用戶，請用戶根據您的計算與儲存需求，來選擇合適的檔案系統來使用。

三個平行檔案系統：

檔案系統掛載點	用途說明
/home	登入後的 Shell 空間，自行安裝的軟體請安裝家目錄下
/work1	計算工作執行過程暫時的儲存空間，適合高效能讀寫
/project	供專用申請計畫保存長期靜態資料的共享空間

- /work1 的儲存總頻寬與平行讀寫效能均優於 /home 與 /project，因此建議用戶去使用 /work1 當作計算工作運算過程的資料輸出空間，不建議用戶直接於 /home 或 /project 進行有大量讀寫檔案的計算工作。
- /project 的使用需要額外付費申請，/project 是專門用來存放需要保存較久的靜態資料。

1. 查詢額度與用量

用戶可以利用以下指令查詢各個 Lustre 檔案系統的儲存容量額度：

```

$ lfs quota -u username /home
Disk quotas for usr username (uid 10181):
Filesystem kbytes  quota  limit  grace  files  quota  limit  grac
/home 34176060  104857600  104857600  -  334315  0  0  -
    
```

- Quota 欄位的數值，為用戶在該平行檔案系統的儲存容量上限 (單位為 bytes)。
- kbytes 欄位的數值，為用戶在該平行檔案系統上，屬於其帳號擁有的所有檔案大小總和 (單位為 bytes)。
- 當使用過程出現 Disk quota exceeded 時，表示檔案總儲存容量超過申請的額度，請刪除較舊的資料或是增購額外的儲存空間，否則資料無法持續輸出，將影響計算工作進行。

當然您也可以使用基本的 Linux 作業系統 "du" 指令，去統計各子目錄下的檔案使用容量：

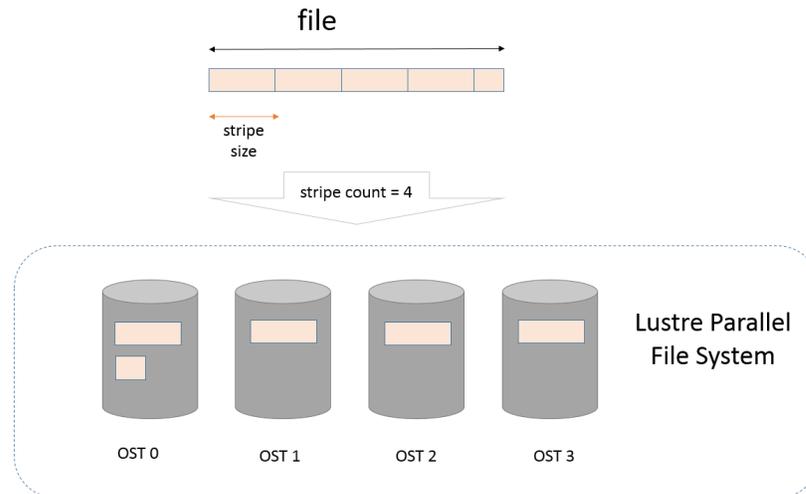
```
$ du -sh /home/username
```

2. 巨大檔案的讀寫與 Application I/O

當您輸出一個檔案，若沒有事先設定 stripe，檔案預設只會寫入到單一個 Lustre OST。當輸出超過容量達 100GB 以上的單一巨大檔案的過程往往非常耗時，因此請使用 Lustre 檔案系統提供的特殊指令，先規劃好專門用來儲存超過 100GB 檔案以上的一個子目錄。然後對這一個子目錄設定 Lustre stripe，以便將巨大檔案的實際資料內容，分散儲存於 Lustre 檔案系統底層的各個 Lustre OST 內，讓儲存底層更多的硬碟共同協助存取資料，增加大檔案的讀寫效率。指令使用方式如下：

```
$ lfs setstripe -c <stripe count> -S <stripe size> <file|directory>
```

```
$ lfs getstripe <file|directory>
```



許多文獻也都指出科學計算常使用到的 IO Library (ADIOS, PnetCDF, NetCDF, HDF5)，平行計算 MPI Library 支援的平行讀寫 (MPIIO, ROMIO) 機制等，都建議要對於計算工作的資料輸出目錄，先設定 Lustre striping，讓平行程式計算的過程搭配平行化的資料讀寫。

以下為一個子目錄設定 8 個 stripe 的範例 (每 32MB 就循序分散到下一個 OST 儲存)，即您寫出的檔案存放到這個子目錄之後，檔案實際上是分散儲存到 8 個 OST：

```
$ lfs setstripe -c 8 -S 32m big_file_dir/  
$ lfs getstripe big_file_dir/  
big_file_dir/  
stripe_count: 8 stripe_size: 33554432 stripe_offset: -1
```

```
$ dd if=/dev/zero of=./big_file_dir/100GB.dat bs=10M count=10240  
10240+0 records in  
10240+0 records out  
107374182400 bytes (107 GB) copied, 86.2875 s, 1.2 GB/s
```

```
$ lfs getstripe ./big_file_dir/100GB.dat  
./big_file_dir/100GB.dat  
lmm_stripe_count: 8  
lmm_stripe_size: 33554432  
lmm_pattern: 1  
lmm_layout_gen: 0  
lmm_stripe_offset: 54  
      obdidx      objid      objid      group  
      54      19473325      0x12923ad      0  
      9      19473037      0x129228d      0  
      63      19474579      0x1292893      0  
      27      19473253      0x1292365      0  
      45      19473700      0x1292524      0  
      19      19472431      0x129202f      0  
      1      19472305      0x1291fb1      0  
      37      19472366      0x1291fee      0
```

備註： stripe count (-c 的參數) 數值不可超過 144。

您無法對已經建立的大檔案的設定 stripe。請參考以下範例重新準備一個有設定 stripe 的空檔案，然後將原本的大檔案複製成這一個空檔案，來達成 stripe 設定：

```
$ lfs setstripe newfile -S 2m -c 8  
$ file newfile  
newfile: empty
```

```
$ cp oldfile newfile
```

```
$ lfs getstripe newfile  
newfile  
lmm_stripe_count: 8  
lmm_stripe_size: 2097152  
lmm_pattern: 1
```

```
lmm_layout_gen: 0
lmm_stripe_offset: 28
  obdidx      objid      objid      group
    28      20104223    0x132c41f    0
    46      20101885    0x132bafd    0
    20      20102663    0x132be07    0
     2      20101708    0x132ba4c    0
    38      20102452    0x132bd34    0
    56      20103079    0x132bfa7    0
    11      20101991    0x132bb67    0
    65      20102382    0x132bcee    0
```

3. 檔案搜尋

搜尋檔案是常見的檔案系統操作，檔案數量越多則搜尋檔案的時間自然就會越久，因此當您在執行計算工作時，不建議進行大量檔案的搜尋，以免計算過程中處理器閒置，大量檔案搜尋也容易造成無法預估合理的計算完成時間。

"lfs" 指令是 Lustre 檔案系統本身所提供給一般使用者可以使用的指令，使用 "lfs find" 來尋找檔案，會比用 Linux 作業系統預設提供 "find" 搜尋指令來得更有效率。以下提供幾個 "lfs find" 使用參考範例：

1. 遞迴搜尋所有子目錄，分行顯示全部

```
$ lfs find /work1/username/DIR
```

2. 只搜尋一層，分行顯示

```
$ lfs find /work1/username/DIR -maxdepth 1
$ lfs find /work1/username/DIR -maxdepth 1 --print
```

3. 只搜尋一層，同一行顯示

```
$ lfs find /work1/username/DIR -maxdepth 1 -print0
```

4. 檔案列表

任何檔案系統的使用過程，都不建議在單一目錄之下，儲存超過執行 "ls -al" 能正常反應時間的檔案數量。

就資料管理的角度，當您有更多的資料檔案需要儲存或輸出時，每一次輸出超過 256 個檔案，就需考慮儲存到其他目錄或子目錄，若您直接對有擁大量檔案的目錄執行列表查看，或者是進行搜尋檔案，都將影響到您計算工作的效率。

根據 Lustre 官方文件，建議您可以使用以下命令，取代 "ls" 指令，將更有效率地顯示列表：

```
$ lfs find -D 0 *
```