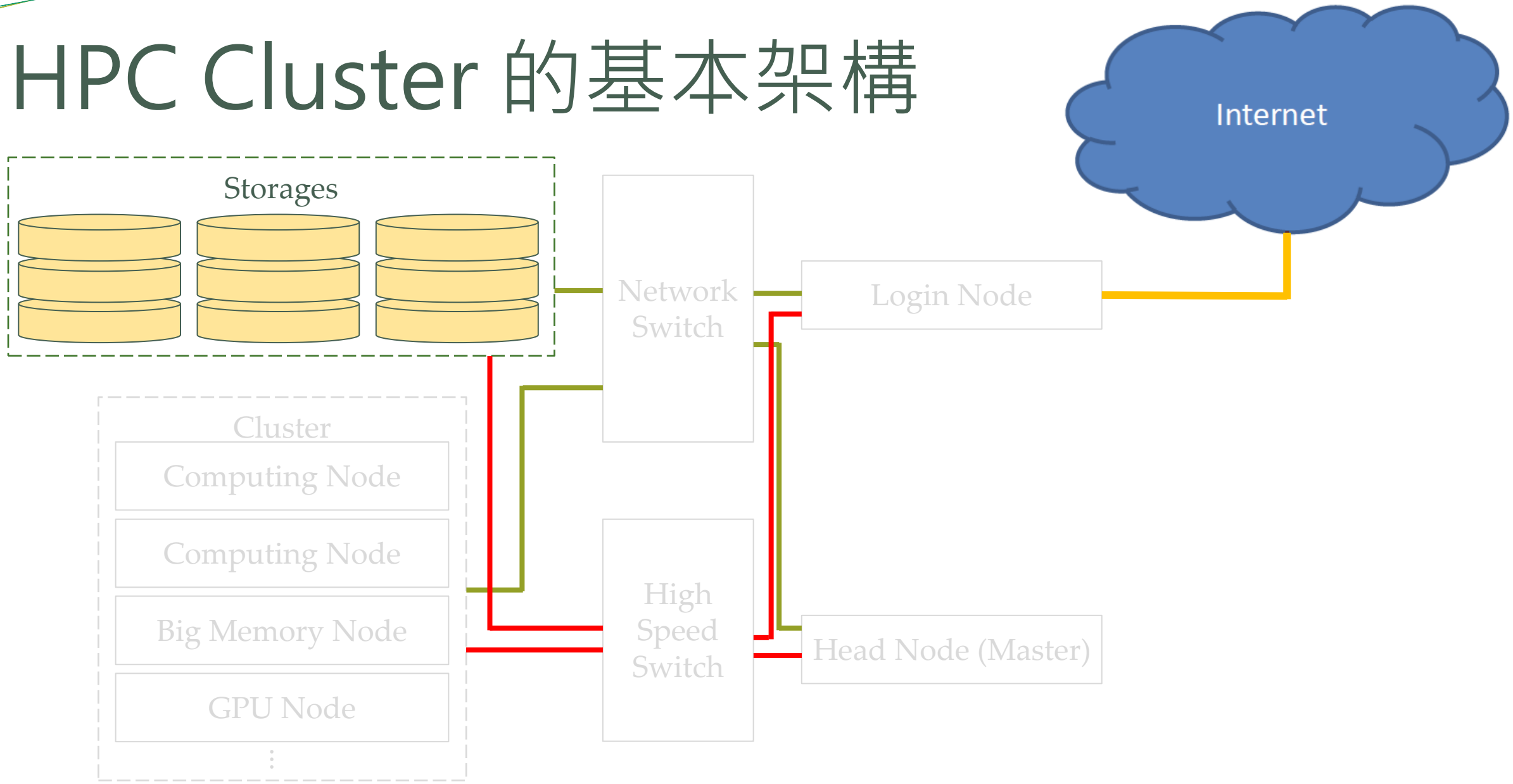
平行檔案系統

國立臺灣師範大學物理學系 陳俊明

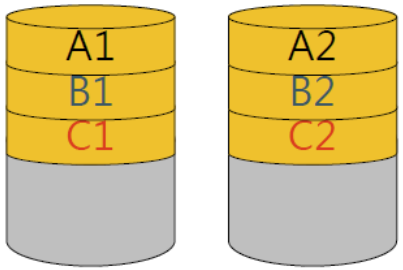
chunming@ntnu.edu.tw

HPC Cluster 的基本架構



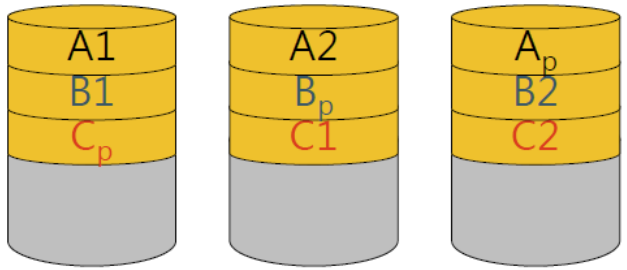
磁碟陣列-RAID(基本型)

RAID0



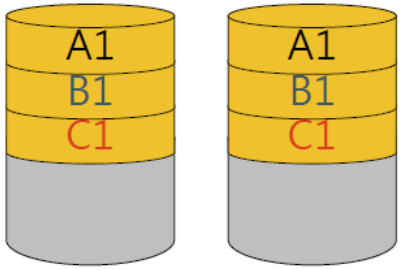
$Size = \mathrm{N} \times \min(S_1, S_2 \ldots S_N)$

RAID5



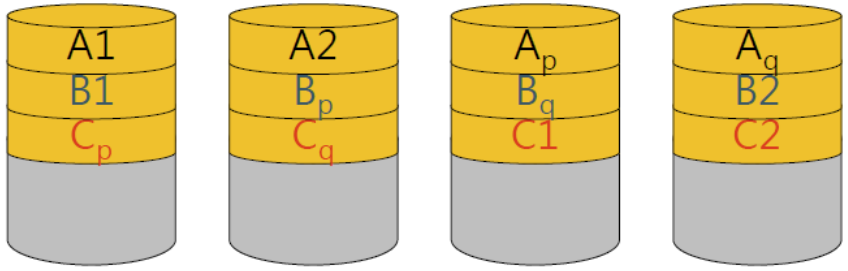
$Size = (N-1) \times \min(S_1, S_2, S_3, \ldots S_N)$

RAID1



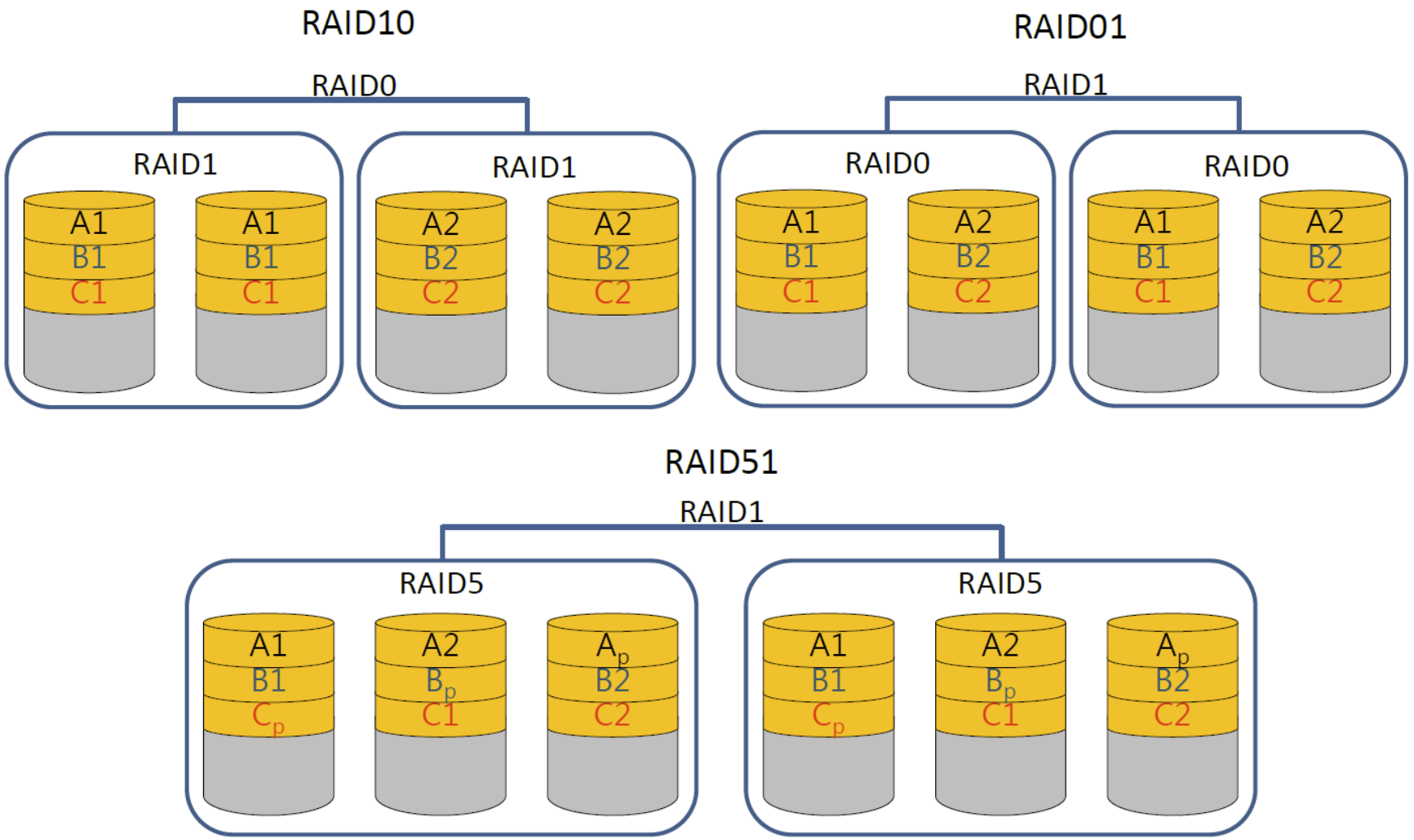
$Size = \min(S_1, S_2 \ldots S_N)$

RAID6



$Size = (N-2) \times \min(S_1, S_2, S_3, S_4 \ldots S_N)$

磁碟陣列-RAID (混合型)



平行檔案系統

- 以分散式讀寫資料的方式，避免所有的磁碟讀寫集中在單一硬碟或是單一伺服器上，通常由數個儲存節點所組成。

Lustre

GlusterFS

BeeGFS

OrangeFS

ceph

- Open Source
  - Lustre
  - Glusterfs
  - BeeGFS
  - OrangeFS
  - Ceph

IBM Spectrum Scale

- Enterprise
  - IBM GPFS
  - DELL Isilon OneFS

Lustre

- 名字源自於 Linux 和 Cluster 這兩個字的混合。
- 符合POSIX(Portable Operating System Interface)
- 基於 GNU GPLv2.0 的開放原始碼授權。
- 檔案系統主要由 MDS、OSS 及 Client 三個部份所組成。
- 支援容量的橫向擴充。

Lustre®

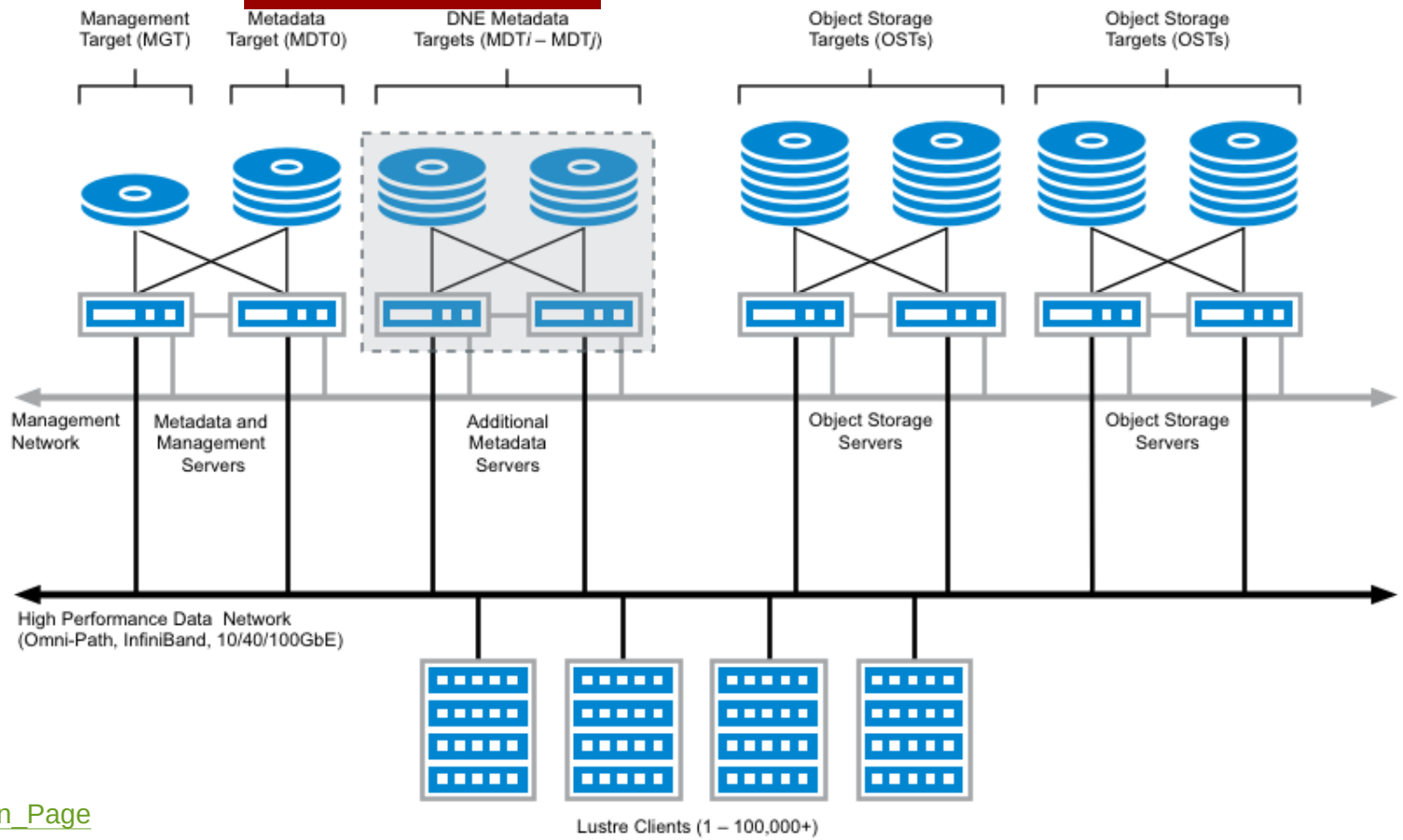
http://lustre.org/

Lustre 架構圖

MDT : where is which files ?

OST-1, OST-2, OST-n : data files

MGT : System Management Info about OSTs and clients



圖片來源：http://wiki.lustre.org/Main_Page

Lustre基本架構

- 管理伺服器 MGS（Management Server)
- Lustre文件系统架構
  - 描述資料伺服器 MDS (Metadata Servers)
  - 物件儲存伺服器 OSS（Object Storage Servers)
  - Lustre 用戶端
- Lustre 網路通訊 LNet（Lustre Networking)

Lustre - MGS (Management Server)

- 可提供多個 lustre 檔案系統（掛載點）
- 儲存一個或多個 lustre 檔案系統的組態資訊，提供資訊給其他的 lustre 主機
- 伺服器及掛載節點連到 MGS 開始使用 lustre
- 通知伺服器及掛載節點檔案系統任何異動包含伺服器重啟
- 該服務的組態資訊寫於 MGT 內

Lustre - MDS (Metadata Server)

- 主要存放資料的部份稱之為 MDT (Metadata Target)
- MDT 中存放所有檔案的元資料 (metadata)，例如檔案名稱、目錄、權限以及檔案位置
- MDT 通常位於 MDS 本機磁碟上
- 與其它 Block-based 的平行分散式檔案系統（如GPFS）不同，Lustre 的 MDS 不參與所有的 Block 分配，僅涉及路徑名和權限檢查，不參與任何文件 I/O 操 作，從而避免了可能的 I/O 瓶頸
- 從 Lustre 2.4 版開始支援在單一系統中具有多個 MDT

Lustre - OSS (Object Storage Server)

- 主要存放資料的部份稱之為 OST (Object Storage Target)
- OST 中存放所有檔案的真實數據
- 1 個 OSS 可存在 1 至 8 個 OST
- OST 通常位於外接的磁碟陣列，並以 DAS (Direct Attached Storage) 的方式連接。也有把內接的磁碟陣列規劃成 OST，一台機箱內有兩台 OSS，能互相備援讀寫 OST，常見於高端設備
- 所有 Block 的分配由個別的 OSS 負責
- Lustre 可用的總容量就是所有 OST 容量的總和

Lustre - Client

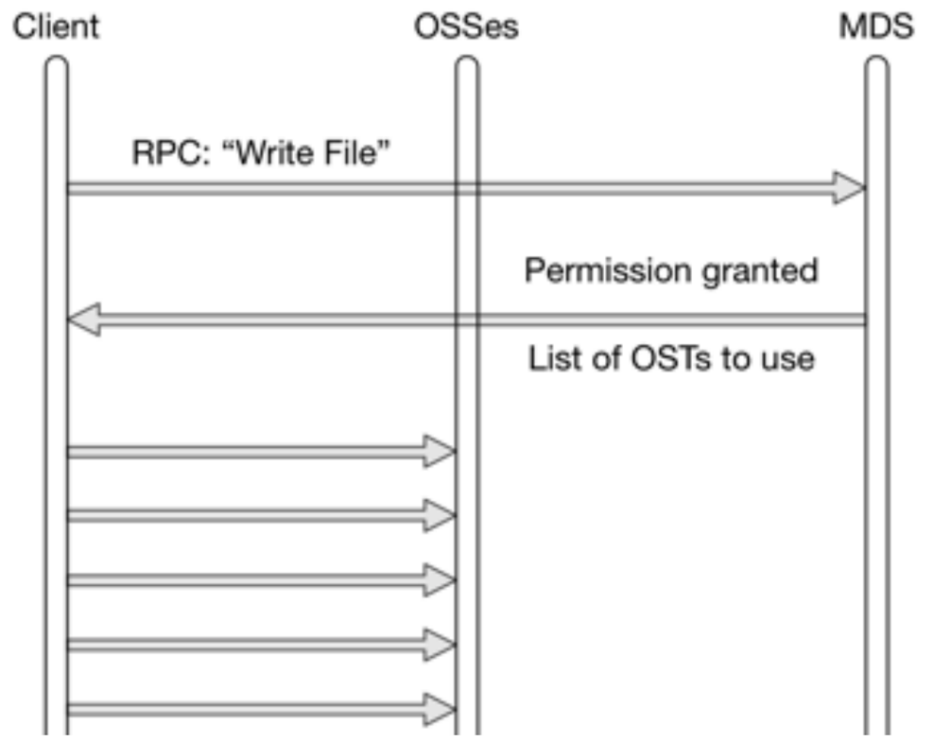
- 存取和使用資料的客戶端。
- Lustre 使用標準 POSIX (Portable Operating System Interface) 為所有客戶端提供檔案系統中所有檔案和數據的統一命名空間，並允許對檔案系統中的檔案同步進行讀寫。

Lustre - LNET

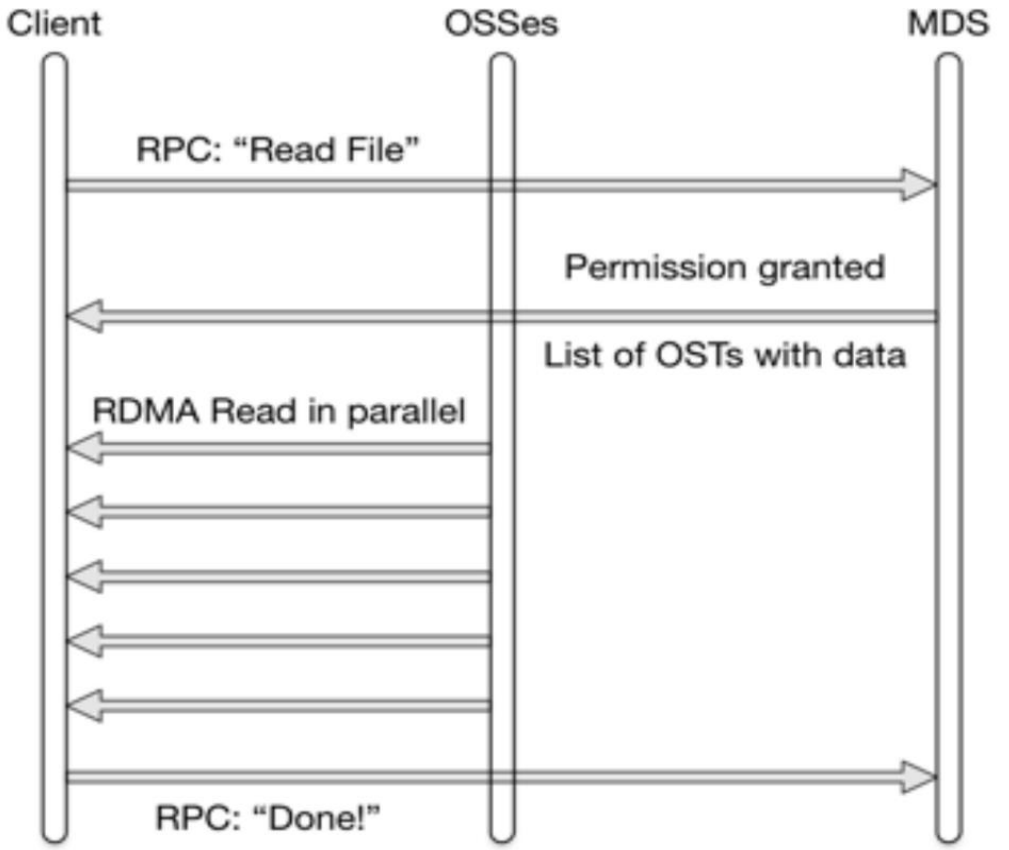
- 不同網路給予獨特名稱
  - o2ib0, tcp0, tcp1
- Lustre Network Identifier (NID) 定義介面
  - 10.1.145.16@o2ib0
- 藉由 Lustre Network Driver (LND) 包含原生支援多網路介面
  - Infiniband via o2ib verbs interface, with RDMA support
  - Ethernet via TCP/IP interface
- Lustre -> Network RPC API, LNet -> LND -> Linux Driver
- 專門為了大規模的運算叢集而設計
  - 對非常大的節點數 (100,000+) 最佳化並且提供高吞吐量
- 能運作於大多數的網路型態及支援 RDMA
  - Ethernet, Infiniband, Omni-Path XC/XT/XE, ELAN, Myrinet*, etc.
- LNet 是獨立於 Lustre 檔案系統
  - 從 lustre 抽離出網路部分，實作成一系列 kernel modules

Lustre資料讀寫

Write



Read



圖片來源：Inside The Lustre File System

Lustre 建制策略

- MGT, MDT需做RAID1以擁有最安全的資料保護機制
- MGT, MDT可採用SSD或NVMe以降低存取瓶頸
- OST需做RAID6+Hot Spare以保護資料
- Client最好使用低延遲的高速網路

Lustre 建立

- 利用Rocky8.ova建立MDS、OSS
  - 變更hostname為mds、oss
  - 關閉Selinux
  - 停止firewalld服務
  - 設定內部網路IP
    - MSD: 192.168.1.100
    - OSS: 192.168.1.101
  - MDS增加2個10 GB的硬碟
  - OSS增加2個20 GB的硬碟

Lustre伺服器以 yum 安裝範例

- 新增 lustre.repo 檔案

```
[root@mds ~]# vi /etc/yum.repos.d/lustre.repo
[lustre-server]
name=lustre-server
baseurl= https://downloads.whamcloud.com/public/lustre/lustre-2.15.3/el8.8/server/
gpgcheck=0

[e2fsprogs-w]
name=e2fsprogs-wc
baseurl=https://downloads.whamcloud.com/public/e2fsprogs/latest/el8/
gpgcheck=0
```

Lustre Server以 yum 安裝範例

- 安裝某特定版本的 kernel。到先前新增的 lustre.repo 找 baseurl 網址，進入其子目錄 RPMS/x86_64 ，查看 kernel 版本

Index of /public/lustre/lustre-2.15.3/el8.8/server/RPMS/x86_64

Name	Last modified	Size	Description
Parent Directory		-	
bpftool-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:05	10M	
bpftool-debuginfo-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:05	11M	
kernel-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:04	9.4M	
kernel-core-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:05	42M	
kernel-cross-headers-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:04	15M	
kernel-debuginfo-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:07	651M	
kernel-debuginfo-common-x86_64-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:04	75M	
kernel-devel-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:05	34M	
kernel-headers-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:04	11M	
kernel-ipaclones-internal-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:07	25M	
kernel-modules-4.18.0-477.10.1.el8_lustre.x86_64.rpm	2023-06-20 01:05	34M	

```
[root@mds ~]# yum install kernel-*
[root@mds ~]# reboot
```

用 **uname -r** 檢查 **kernel** 版本

Lustre伺服器以 yum 安裝範例

- 安裝 lustre 套件

```
[root@mds ~]# yum install lustre
```

```
================================================================================
 Package                 Arch     Version              Repository        Size
================================================================================
Installing:
 lustre                  x86_64   2.12.5-1.el7         lustre-server     801 k
Installing for dependencies:
 kmod-lustre             x86_64   2.12.5-1.el7         lustre-server     3.9 M
 kmod-lustre-osd-ldiskfs x86_64   2.12.5-1.el7         lustre-server     468 k
 lustre-osd-ldiskfs-mount x86_64  2.12.5-1.el7         lustre-server      14 k
Updating for dependencies:
 e2fsprogs               x86_64   1.45.6.wc1-0.el7     e2fsprogs-wc      996 k
 e2fsprogs-libs          x86_64   1.45.6.wc1-0.el7     e2fsprogs-wc      241 k
 libcom_err              x86_64   1.45.6.wc1-0.el7     e2fsprogs-wc       42 k
 libss                   x86_64   1.45.6.wc1-0.el7     e2fsprogs-wc       47 k

Transaction Summary
================================================================================
Install  1 Package  (+3 Dependent packages)
Upgrade             ( 4 Dependent packages)

Total download size: 6.5 M
Is this ok [y/d/N]:
```

以上步驟 MGS, MDS, OSS 都要個別安裝

Lustre伺服器設定

- 設定 lustre module 檔案

```
[root@mds ~]# vi /etc/modprobe.d/lustre.conf
options lnet networks=tcp0(enp0s3 or eth0)
```

  - InfiniBand “optins lnet networks=o2ib0(ib0)”
- 載入lnet, lustre module

```
[root@mds ~]# modprobe lnet
[root@mds ~]# modprobe lustre
[root@mds ~]# lctl list_nids
192.168.1.100@tcp
```

以上步驟 MGS, MDS, OSS 都要個別安裝

Lustre 伺服器設定範例 - MGS / MDS

- MGS 建立 MGT

```
[root@mds ~]# mkfs.lustre --fsname lustre --mgs /dev/sdb
```

- MDS 建立 MDT

```
[root@mds ~]# mkfs.lustre --fsname lustre --mgsnode MGS_SERVER_IP_OR_HOSTNAME@tcp --mdt /dev/sdc
```

- 建立掛目錄

```
[root@mds ~]# mkdir -p /lustre/mgt
[root@mds ~]# mkdir -p /lustre/mdt
```

建議將 MGS 與 MDS 分開建置

Lustre 伺服器設定範例 - MGS / MDS

- 掛載 MGT 開始提供服務

```
[root@mds ~]# mount -t lustre /dev/sdb /lustre/mgt
```

- 掛載 MDT 開始提供服務

```
[root@mds ~]# mount -t lustre /dev/sdc /lustre/mdt
```

Lustre 伺服器設定範例 - MGS / MDS

- 設定 fstab 使用 UUID 掛載，但建議用手動掛載

```
[root@mds ~]# blkid /dev/sdb
/dev/sdb: LABEL="MGS" UUID="da7da54e-e621-4c7f-bcad-a584f0bad13c" TYPE="ext4"
# blkid /dev/sdc
/dev/sdc: LABEL="scratch-MDT0000" UUID="9ec31bd9-88a6-4507-bcc2-48366513f80b" TYPE="ext4"
[root@mds ~]# vi /etc/fstab
UUID=da7da54e-e621-4c7f-bcad-a584f0bad13c /lustre/mgt   lustre  defaults        0 0
UUID=9ec31bd9-88a6-4507-bcc2-48366513f80b /lustre/mdt   lustre  defaults        0 0
# mount /lustre/mgt
# mount /lustre/mdt
```

Lustre 伺服器設定範例 - OSS

- 建立 OST0，預設是用 ext4 進行格式化

```
[root@oss ~]# mkfs.lustre --fsname lustre --ost --index=0 --mgsnode=MGS_SERVER_IP@tcp /dev/sdb
```

- 建立 OST1

```
[root@oss ~]# mkfs.lustre --fsname lustre --ost --index=1 --mgsnode=MGS_SERVER_IP@tcp /dev/sdc
```

- 建立掛目錄

```
[root@oss ~]# mkdir -p /lustre/ost0
[root@oss ~]# mkdir -p /lustre/ost1
```

- 掛載 OST0 及 OST1 開始提供服務

```
[root@oss ~]# mount -t lustre /dev/sdb /lustre/ost0
[root@oss ~]# mount -t lustre /dev/sdc /lustre/ost1
```

Lustre 伺服器設定範例 - OSS

- 設定 fstab 使用 UUID 掛載，但建議用手動掛載

```
[root@mds ~]# blkid /dev/sdb
/dev/sdb: LABEL="MGS" UUID="5af6018d-e39f-48f4-afe3-79e1c021743f" TYPE="ext4"
# blkid /dev/sdc
/dev/sdc: LABEL="scratch-MDT0000" UUID="2616c799-2735-44ee-8b54-4ff62a2ecb72" TYPE="ext4"
[root@mds ~]# vi /etc/fstab
UUID=5af6018d-e39f-48f4-afe3-79e1c021743f /lustre/ost0   lustre  defaults        0 0
UUID=2616c799-2735-44ee-8b54-4ff62a2ecb72 /lustre/ost1   lustre  defaults        0 0
# mount /lustre/ost0
# mount /lustre/ost1
```

Lustre 用戶端以 yum 安裝範例

- 新增 lustre.repo 檔案

```
[root@master ~]# vi /etc/yum.repos.d/lustre.repo
[lustre-client]
name=lustre-client
baseurl= https://downloads.whamcloud.com/public/lustre/lustre-2.15.3/el8.8/client/
gpgcheck=0
```

- 安裝 lustre-client 套件

```
[root@master ~]# yum install lustre-client
```

- 若有更新kernel需重新啟動

```
[root@master ~]# reboot
```

Lustre 用戶端設定範例

- 編輯 lustre module 檔案

```
[root@master ~]# vi /etc/modprobe.d/lustre.conf
options lnet networks=tcp0(enp0s8)
```

  - InfiniBand “optins lnet networks=o2ib0(ib0)”
- 建立掛目錄並且掛載

```
[root@master ~]# mkdir /lustre
[root@master ~]# mount -t lustre MGS_SERVER_IP_OR_HOSTNAME@tcp:/lustre /lustre
[root@master ~]# vi /etc/fstab
MGS_SERVER_IP_OR_HOSTNAME@tcp:/lustre  /lustre  lustre  defaults,_netdev 0 0
```

- 檢查掛載狀

```
[root@master ~]# df -ht lustre
```

一般使用者帳號使用lustre空間須在MGS/MDS上安裝NIS服務

Lustre - 常用指令

- modprobe lustre: 載入Lustre模組
- lustre_rmmod: Lustre模組卸載
- lctl: 控制Lustre屬性，可調整相關的配置與屬性
  - lctl lustre_build_version：顯示Lustre版本
  - lctl list_nids：顯示網路ID
  - lctl dl：顯示檔案系統組件
  - lctl device_list:列表 lustre 設備狀態

Lustre - 常用指令

- lfs: 主要對於檔案相關屬性進行配置及查詢
  - lfs df: 用戶端上執行時，顯示各MDT、OST等空間使用情況
  - lfs mdts /lustre: 用戶端上執行時，顯示/lustre使用MDT情況
  - lfs osts / lustre: 用戶端上執行時，显示/lustre使用OST情況
  - lfs quota -h -u ‘USERID’ /lustre: 查詢使用者在/lustre目錄下使用的容量大小及檔案數

Lustre - stripe 設定

- 取得目前 stripe 設定，以下範例為預設的設定

```
[root@master ~]# lfs getstripe /lustre
/lustre
stripe_count:  1 stripe_size:   1048576 pattern:       0 stripe_offset: -1
```

- 設定 stripe 大小為 4MB，使用 2 個 OSTs，起始 OST 順序不指定

```
[root@master ~]# lfs setstripe -S 4M -i -1 -c 2 /lustre/data
[root@master ~]# lfs getstripe FILENAME
lmm_stripe_count:  2
lmm_stripe_size:   4194304
lmm_pattern:       raid0
lmm_layout_gen:    0
lmm_stripe_offset:  0
        obdidx           objid           objid           group
             0              35            0x23               0
             1              35            0x23               0
```

Lustre - stripe 設定

- 讓所有 OST 都抄寫一份，小於 1GB 的檔案可以得到保護

```
[root@master ~]# lfs setstripe -S 1G -i -1 -c -1 /lustre/vip
[root@master ~]# lfs getstripe FILENAME
lmm_stripe_count:  2
lmm_stripe_size:   1073741824
lmm_pattern:       raid0
lmm_layout_gen:    0
lmm_stripe_offset:  1
        obdidx           objid           objid           group
             1              34            0x22               0
             0              34            0x22               0
```

Lustre - quota 設定

- 設定只能使用 1GB 最多不能超過 10,000 檔案

```
[root@master ~]# lfs setquota -u USERID -B 1224M -b 1G -I 12000 -i 10000 /lustre
[root@master ~]# lfs quota -uh USERID /lustre
Disk quotas for usr USERID (uid 1000):
    Filesystem    used   quota   limit   grace   files   quota   limit   grace
         /lustre      4k      1G  1.195G       -       1   10000   12000       -
```

- 取消 quota 設定

```
[root@master ~]# lfs setquota -u USERID -B 0 -b 0 -I 0 -i 0 /lustre
[root@master ~]# lfs quota -uh USERID /lustre
Disk quotas for usr USERID (uid 1000):
    Filesystem    used   quota   limit   grace   files   quota   limit   grace
         /lustre      4k       0k      0k       -       1       0       0       -
```

Lustre 維護

- 更新 Kernel 並重新編譯安裝 Lustre Client
- 日常檢查
  - 檢查 MDS 及 OSS 的硬體狀態
  - 檢查 MDS 及 OSS 上的負載、記憶體及系統容量 (lfs df -h, lfs df -hi)
  - 檢查 MDS 及 OSS 的網路狀態
  - OSS 負載過高時，登入 OSS 找尋負載來源 (lctl get_param ost.OSS.ost_io.req_history)
- 強制重開機時（負載過高導致無法登入MDS、OSS或當機）
  - 需在 MDS 或 OSS 上執行 e2fsck -f /dev/sdX 檢查 MDT 或 OST
  - 掛載 MDT 或 OST（lctl dl, 使用 dmesg 確認 kernel 訊息）
  - 等待 Client 端回復連線

Lustre 維護 - Client 端更新 Kernel

- 卸載 lustre 掛載點用 umount /MOUNT_POINT 指令，如果無法正常卸載用 lsof | grep MOUNT_NAME 取得正在 lustre 使用的程序，把它清除 kill -9 PID
- lustre_rmmod 指令卸載 kernel module
- 將 /etc/fstab 註解掉 lustre 掛載設定
- 移除目前使用 lustre 套件 rpm -e lustre-client kmod-lustre-client
- 更新 Kernel 重新開機
- 重新編譯 lustre client 的 rpm 檔案
- 安裝編譯好的 rpm 檔案
- 將 /etc/fstab 取消註解
- 掛載 lustre

Lustre 維護 - 開機順序

- 登入 MGS 掛載 mgt
- 登入 MDS 掛載 mdt
- 登入 OSS 掛載 ost[0-n]
- 登入運算節點，掛載 lustre client 的掛載點

Lustre 維護 - 關機順序

- 登入運算節點，卸載 lustre client 的掛載點
- 卸載 MDS 的 mdt 的掛載點
- 卸載 OSS 的 ost[0-n] 的掛載點
- 卸載 MGS 的 mgt 的掛載點（非必要）

Lustre 維護 - OSS 維運

- OST1 資料過滿，可以使用 lfs_migrate 指令搬動資料

```
[root@oss ~]# lfs find --obd lustre-OST0001_UUID /lust | lfs_migrate -y
```

- 暫時把 OST1 停用

```
[root@oss ~]# lctl set_param osc.lustre-OST0001-*.active=0
```

- 恢復 OST1 服務

```
[root@oss ~]# lctl set_param osc.lustre-OST0001-*.active=1
```

其他平行檔案系統

- DDN - EXAScaler, Cray - ClusterStor (Lustre)
- IBM Spectrum Scale (GPFS)
- BeeGFS - The Leading Parallel Cluster File System
- Panasas / pNFS
- Dell Isilon
- Red Hat Gluster Storage
- Ceph

Lustre 編譯

- 準備 build rpms 安裝環境

```
[root@master ~]# yum groupinstall "Development Tools"
[root@master ~]# yum config-manager --set-enabled powertools
[root@master ~]# yum install -y gcc autoconf libtool which make patch diffutils file binutils-devel python38 python3-devel elfutils-devel libselinux-devel libaio-devel dnf-plugins-core bc bison flex git libyaml-devel libnl3-devel libmount-devel json-c-devel redhat-lsb libssh-devel libattr-devel libtirpc-devel libblkid-devel openssl-devel libuuid-devel texinfo texinfo-tex
[root@master ~]# yum -y install audit-libs-devel binutils-devel elfutils-devel kabi-dw ncurses-devel newt-devel numactl-devel openssl-devel pciutils-devel perl perl-devel python2 python3-docutils xmlto xz-devel elfutils-libelf-devel libcap-devel libcap-ng-devel llvm-toolset libyaml libyaml-devel kernel-rpm-macros kernel-abi-whitelists
[root@master ~]# yum install epel-release
[root@master ~]# yum install -y ccache
```

- 準備建立kernel相關套件

```
[root@master ~]# yum install -y bpftool dwarves java-devel libbabeltrace-devel libbpf-devel libmnl-devel net-tools rsync
[root@master ~]# yum install rocky-sb-certs --enablerepo devel
```

Lustre 編譯

- 建立 build 帳戶

```
[root@master ~]# useradd -m build
```

- 用 build 帳戶下載Lustre原始碼

```
[root@master ~]# su - build
[build@master ~]$ git clone --branch 2.15.3 git://git.whamcloud.com/fs/lustre-release.git
```

- 進入lustre-release目錄，產生 configure 檔案

```
[build@master ~]$ cd lustre-release
[build@master lustre-lelease]$ sh autogen.sh
configure.ac:10: installing 'config/config.guess'
configure.ac:10: installing 'config/config.sub'
configure.ac:12: installing 'config/install-sh'
configure.ac:12: installing 'config/missing'
libcfs/libcfs/autoMakefile.am: installing 'config/depcomp'
```

Lustre 編譯

- 準備編譯含有 lustre 修補 (patch) 的 kernel 原始碼

```
[build@master ~]$ mkdir -p ~/kernel/rpmbuild/{BUILD,RPMS,SOURCES,SPECS,SRPMS}
[build@master ~]$ cd ~/kernel
[build@master kernel]$ echo '%_topdir %(echo $HOME)/kernel/rpmbuild' > ~/.rpmmacros
```

- 獲得kernel原始碼

```
[build@master kernel]$ rpm -i https://download.rockylinux.org/pub/rocky/8/BaseOS/source/tree/Packages/k/kernel-4.18.0-477.21.1.el8_8.src.rpm 2>&1 | grep -v exist
```

- 利用rpmbuild準備kernel原始碼

```
[build@master kernel]$ cd ~/kernel/rpmbuild
[build@master rpmbuild]$ rpmbuild -bp --target=`uname -m` ./SPECS/kernel.spec
```

Lustre 編譯

- 複製 kernel config file 到lustre目錄下

```
[build@master rpmbuild]$ cp ~/kernel/rpmbuild/BUILD/kernel-4.18.0-477.21.1.el8_8/linux-4.18.0-477.21.1.el8.x86_64/configs/kernel-4.18.0-x86_64.config ~/lustre-release/lustre/kernel_patches/kernel_configs/kernel-4.18.0-4.18-rhel8.8-x86_64.config
```

- 修改kernel-4.18.0-4.18-rhel8.8-x86_64.config

```
[build@master rpmbuild]$ vi ~/lustre-release/lustre/kernel_patches/kernel_configs/kernel-4.18.0-4.18-rhel8.8-x86_64.config
搜尋字串 '# IO Schedulers'並於插入下列兩行:

CONFIG_IOSCHED_DEADLINE=y
CONFIG_DEFAULT_IOSCHED="deadline"
```

Lustre 編譯

- 在kernel中加入luster修補(patch)

```
[build@master rpmbuild]$ cd ~/lustre-release/lustre/kernel_patches/series
[build@master series]$ for patch in $(<"4.18-rhel8.8.series"); do \
>patch_file="$HOME/lustre-release/lustre/kernel_patches/patches/${patch}"; \
>cat "${patch_file}" >> "$HOME/lustre-kernel-x86_64-lustre.patch"; \
>done
[build@master series]$ cp ~/lustre-kernel-x86_64-lustre.patch ~/kernel/rpmbuild/SOURCES/patch-4.18.0-lustre.patch
```

- 編輯kernel spec檔案

```
[build@master ~]$ vi ~/kernel/rpmbuild/SPECS/kernel.spec
```

Lustre 編譯

1. 搜尋字串：‘find $RPM_BUILD_ROOT/lib/modules/$KernelVer‘ 並插入下列兩行

```
cp -a fs/ext4/* $RPM_BUILD_ROOT/lib/modules/$KernelVer/build/fs/ext4
rm -f $RPM_BUILD_ROOT/lib/modules/$KernelVer/build/fs/ext4/ext4-inode-test*
```

2. 搜尋字串： '# empty final patch to facilitate testing of kernel patches'並並於上一行插入下列兩行

```
# adds Lustre patches
Patch99995: patch-%{version}-lustre.patch
```

3. 搜尋字串： ‘ApplyOptionalPatch linux-kernel-test.patch’並於上一行插入下列兩行

```
# lustre patch
ApplyOptionalPatch patch-%{version}-lustre.patch
```

4. 儲存spec檔案退出編輯。

Lustre 編譯

- 將lustre-release/lustre/kernel_patches/kernel_configs/kernel-4.18.0-4.18-rhel8.8-x86_64.config檔覆蓋原本kernel的config檔

```
[build@master ~]$ echo '# x86_64' > ~/kernel/rpmbuild/SOURCES/kernel-x86_64.config
[build@master ~]$ cat ~/lustre-release/lustre/kernel_patches/kernel_configs/kernel-4.18.0-4.18-rhel8.8-x86_64.config >>
~/kernel/rpmbuild/SOURCES/kernel-x86_64.config
```

- 編譯kernel並打包成RPM檔

```
[build@master ~]$ cd ~/kernel/rpmbuild
[build@master rpmbuild]$ buildid="_lustre"
[build@master rpmbuild]$ rpmbuild -ba --with firmware --target x86_64 --with baseonly --without kabichk --define "buildid ${buildid}"
~/kernel/rpmbuild/SPECS/kernel.spec
```

- 複製 lustre kernel 並至~/lustre/kernel

```
[root@master ~]# cd /home/build/kernel/rpmbuild/RPMS/x86_64
[root@mds x86_64]# mkdir -p ~/lustre/kernel
[root@mds x86_64]# cp *.rpm ~/lustre/kernel
```

Lustre 編譯

- 安裝e2fsprogs

```
[root@master ~]# vi /etc/yum.repos.d/e2fsprogs.repo
[e2fsprogs-el8-x86_64]
name=e2fsprogs-el8-x86_64
baseurl=https://downloads.whamcloud.com/public/e2fsprogs/latest/el8/
enabled=1
priority=1
gpgcheck=0

[root@master ~]# yum install e2fsprogs-devel
```

- 安裝kernel相關工具

```
[root@master ~]# yum install kernel-rpm-macros kernel-abi-whitelists
```

Lustre 編譯

- 編譯 lustre server 原始碼，編譯好之後以 rpm 方式安裝

```
[root@master ~]# cd /home/build/lustre-release/
[root@master lustre-release]# ./configure --with-linux=/home/build/kernel/rpmbuild/BUILD/kernel-4.18.0-477.21.1.el8_8/linux-4.18.0-477.21.1.el8_lustre.x86_64/ --enable-quota
```

  - --enable-quota 表示啟動硬碟配額
  - -with-o2ib=“/usr/src/kernels/KERNEL-VERSION” 如有使用InfiniBand
- 打包成RPM檔

```
[root@master lustre-release]# make rpms
```

- 複製 lustre server 套件至~/lustre/server/

```
[root@master lustre-release]# cp *.x86_64.rpm ~/lustre/server/
```

Lustre 編譯

- 編譯 lustre client 原始碼，編譯好之後以 rpm 方式安裝

```
[root@master ~]# cd /home/build/lustre-release/
[root@master lustre-release]# make distclean
[root@master lustre-release]# ./configure --with-linux=/home/build/kernel/rpmbuild/BUILD/kernel-4.18.0-477.21.1.el8_8/linux-4.18.0-477.21.1.el8_lustre.x86_64/ --enable-quota  –disable-server
```

  - --enable-quota 表示啟動硬碟配額
  - -with-o2ib=“/usr/src/kernels/KERNEL-VERSION” 如有使用InfiniBand
- 打包成RPM檔

```
[root@master lustre-release]# make rpms
```

- 複製 lustre client 套件至~/lustre/client/

```
[root@master lustre-release]# cp *.x86_64.rpm ~/lustre/client/
```

Lustre 伺服器安裝

- 安裝 lustre kernel並重新開機

```
[root@mds ~]# cd ~/lustre/kernel
[root@mds kernel]# yum localinstall {kernel,kernel-devel,kernel-core,kernel-modules}-4.18.0-477.21.1.el8_lustre.x86_64.rpm
[root@mds kernel]# reboot
```

- 升級e2fsprogs套件

```
[root@mds ~]# vi /etc/yum.repos.d/e2fsprogs.repo
[e2fsprogs-el8-x86_64]
name=e2fsprogs-el8-x86_64
baseurl=https://downloads.whamcloud.com/public/e2fsprogs/latest/el8/
enabled=1
priority=1
gpgcheck=0
[root@mds ~]# yum update e2fsprogs
```

- 安裝 lustre server

```
[root@mds ~]# cd ~/lustre/server
[root@mds ~]# yum localinstall *.x86_64.rpm
```

Lustre 用戶端安裝

- 安裝 lustre kernel並重新開機

```
[root@master ~]# cd ~/lustre/kernel
[root@master kernel]# yum install {kernel,kernel-devel,kernel-headers}-4.18.0-477.21.1.el8_lustre.x86_64.rpm
[root@msd kernel]# reboot
```

- 升級e2fsprogs套件

```
[root@master ~]# vi /etc/yum.repos.d/e2fsprogs.repo
[e2fsprogs-el8-x86_64]
name=e2fsprogs-el8-x86_64
baseurl=https://downloads.whamcloud.com/public/e2fsprogs/latest/el8/
enabled=1
priority=1
gpgcheck=0
[root@mds ~]# yum update e2fsprogs
```

- 安裝 lustre client

```
[root@master ~]# cd ~/lustre/client
[root@master ~]# yum install lustre-client-2.15.3-1.el8.x86_64.rpm kmod-lustre-client-2.15.3-1.el8.x86_64.rpm
```