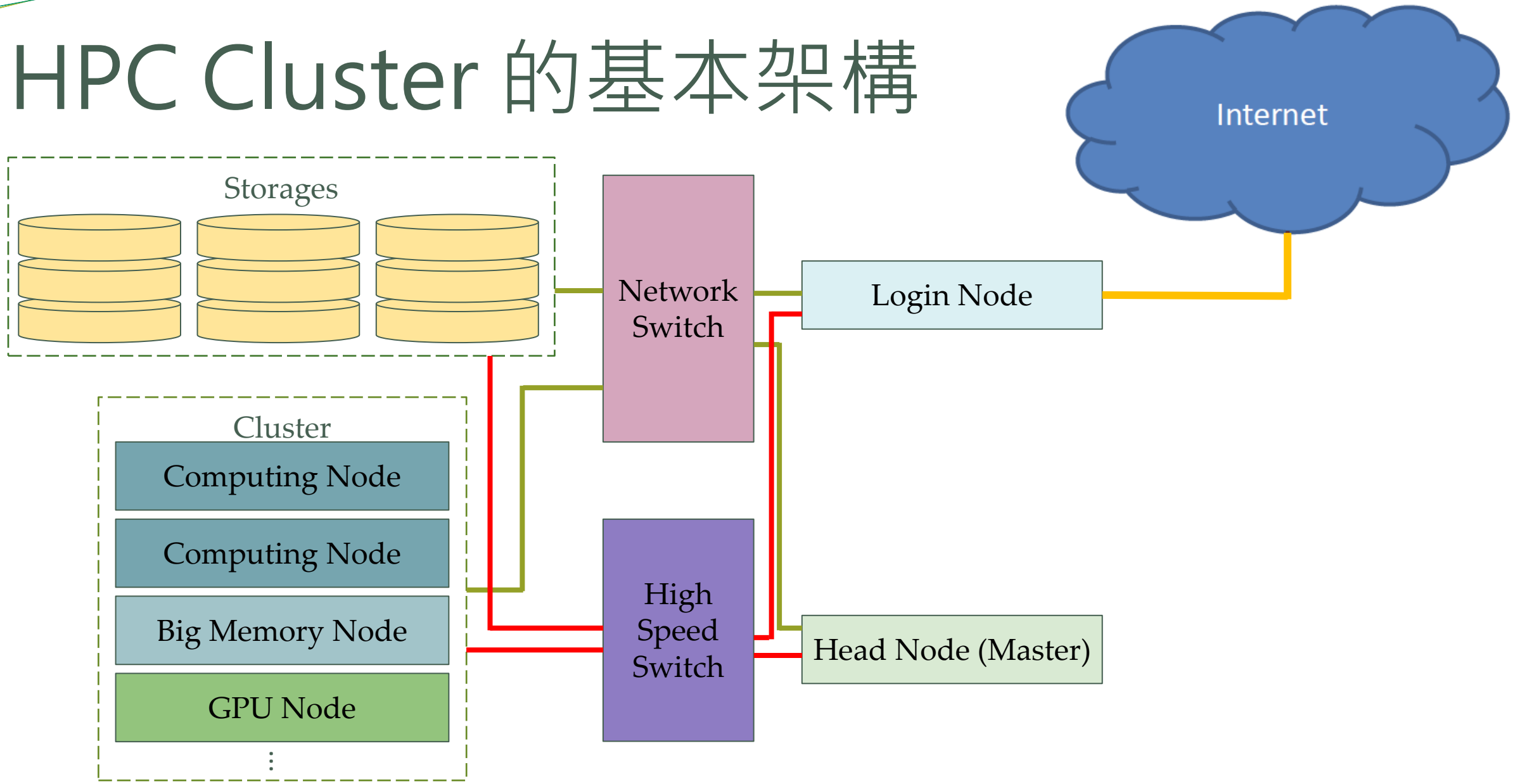
Cluster 建立

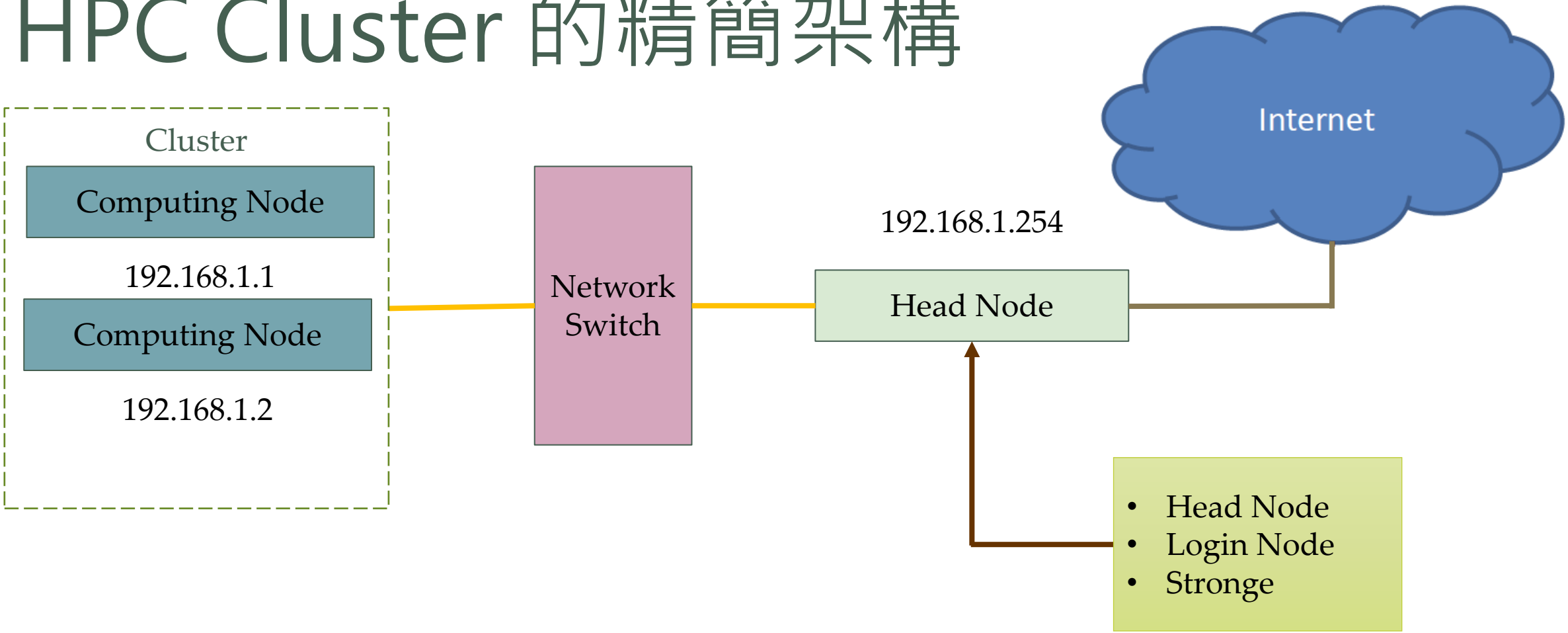
國立臺灣師範大學物理學系 陳俊明

chunming@ntnu.edu.tw

HPC Cluster 的基本架構



HPC Cluster 的精簡架構

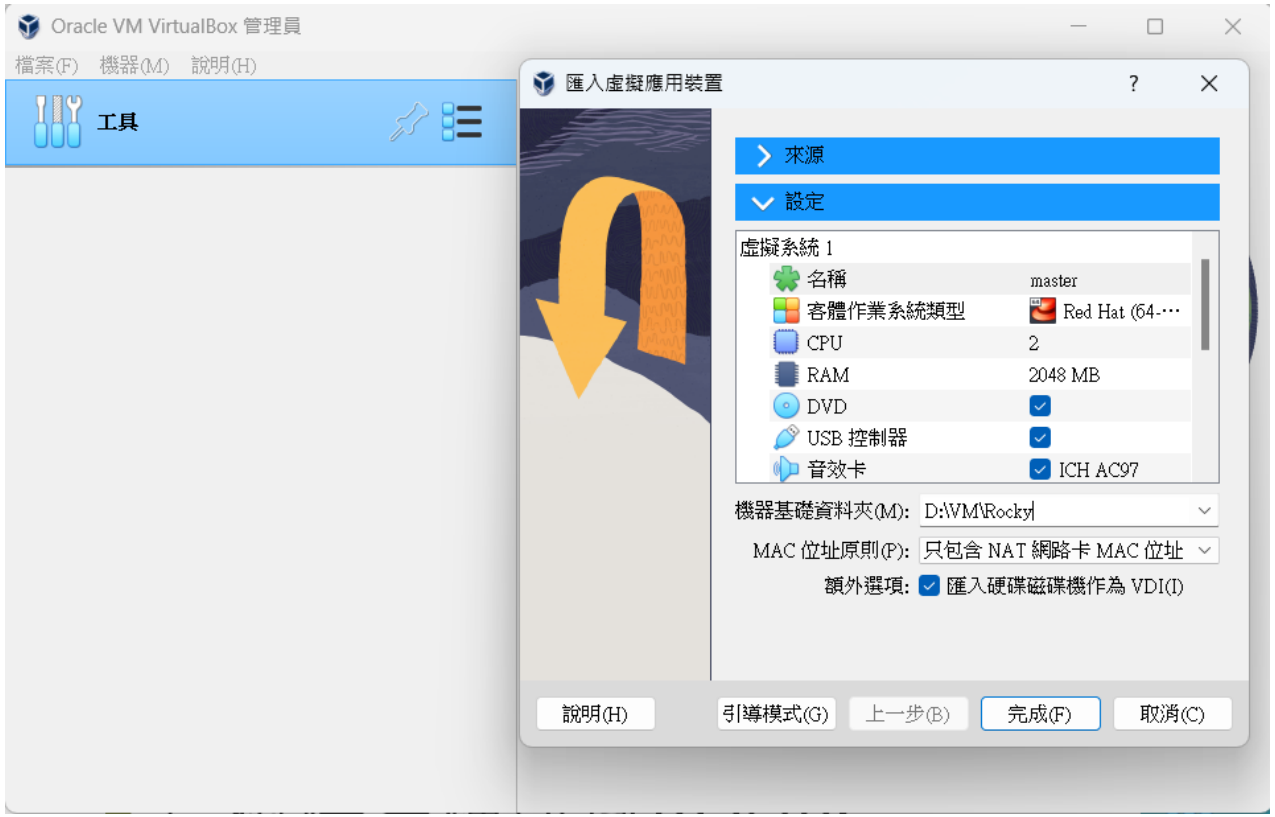


HPC Cluster的必要服務

功能	軟體元件
網路磁碟	NFS, Lustre, BeeGFS...etc
帳號	NIS (ypserver / ypbind), LDAP
校時	Chroncy
排程	PBS Pro, Torque, Slurm...etc

建立Cluster –虛擬機 (Head Node)

- 下載 Oracle VM VirtualBox (https://www.virtualbox.org/) 並安裝
- 下載 Rocky8.ova (https://reurl.cc/QXkAAq)
- 匯入 Rocky8.ova
- 虛擬機名稱改成master



建立Cluster – 關閉SELinux

- HPC Cluster需關閉SELinux

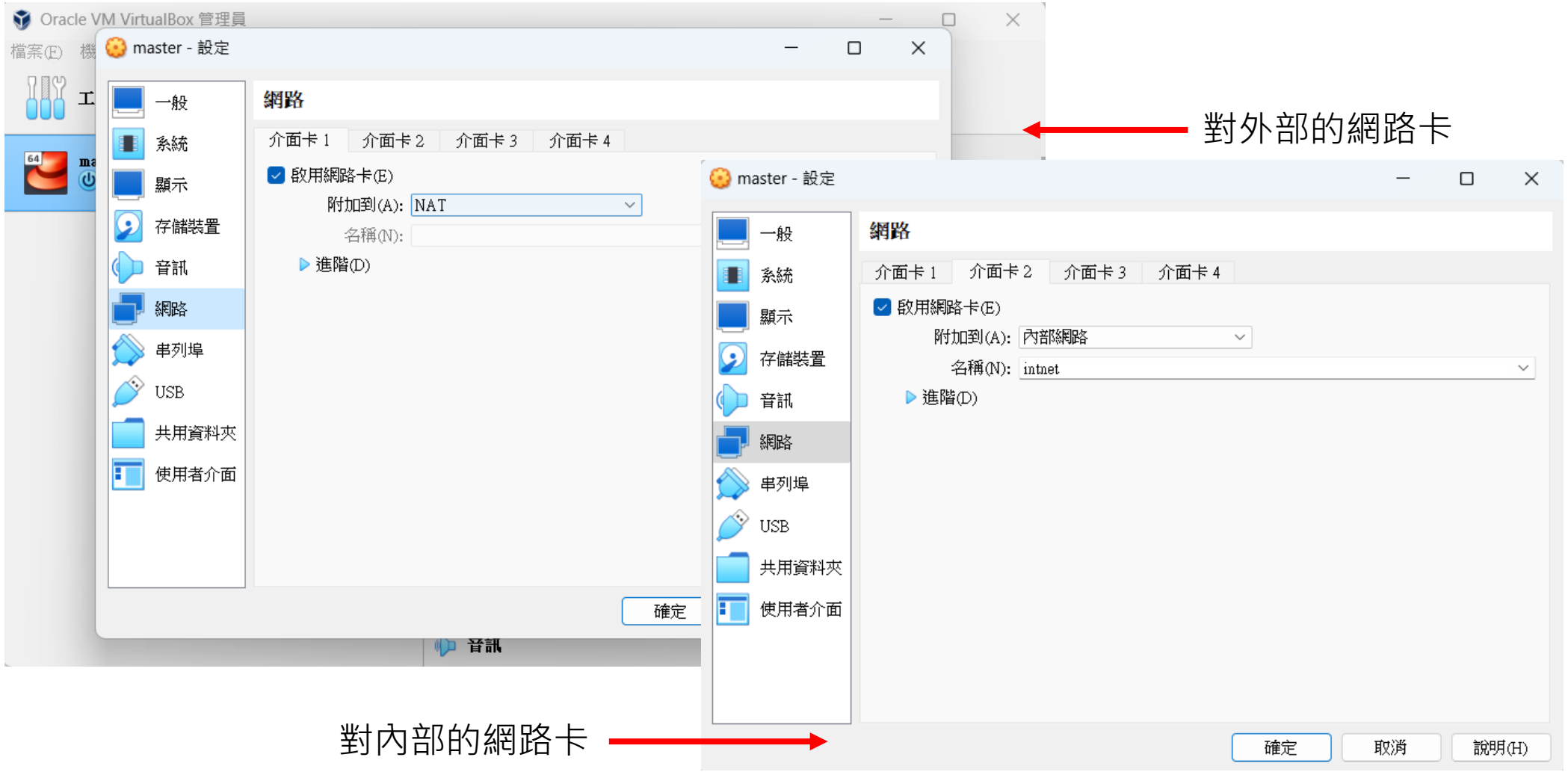
關閉SELinux, 重啟系統後生效

編輯：/etc/selinux/config
SELINUX=disabled

```
# This file controls the state of SELinux on the system.
# SELINUX= can take one of these three values:
#     enforcing - SELinux security policy is enforced.
#     permissive - SELinux prints warnings instead of enforcing.
#     disabled - No SELinux policy is loaded.
SELINUX=enforcing
# SELINUXTYPE= can take one of these three values:
#     targeted - Targeted processes are protected,
#     minimum - Modification of targeted policy. Only selected processes are protected.
#     mls - Multi Level Security protection.
SELINUXTYPE=targeted

~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
"/etc/selinux/config" 14L, 548C
```

建立Cluster - VM網路設定(Master)



建立Cluster – VM網路設定(Master)

查詢網路裝置訊息： “ip add”

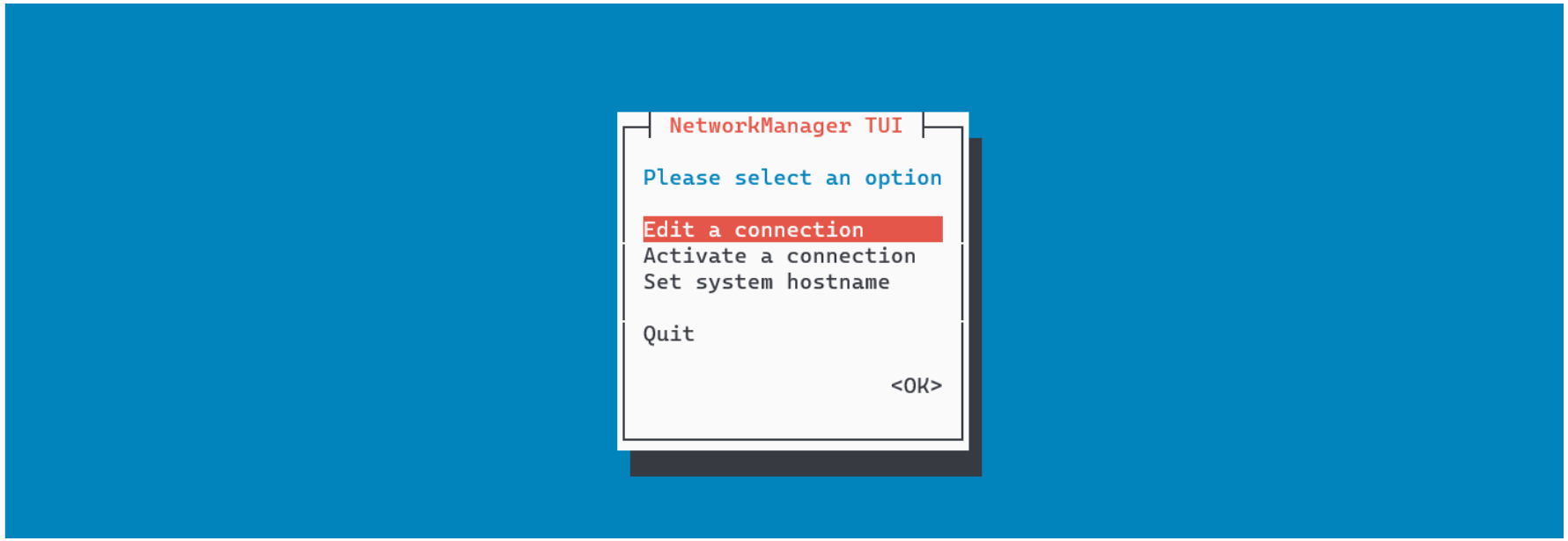
```
[root@Rocky8 ~]# ip add
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp0s3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP group default qlen 1000
    link/ether 08:00:27:37:eb:7b brd ff:ff:ff:ff:ff:ff
    inet 10.0.2.15/24 brd 10.0.2.255 scope global dynamic noprefixroute enp0s3
       valid_lft 86077sec preferred_lft 86077sec
    inet6 fe80::a00:27ff:fe37:eb7b/64 scope link noprefixroute
       valid_lft forever preferred_lft forever
3: enp0s8: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP group default qlen 1000
    link/ether 08:00:27:a3:2b:0b brd ff:ff:ff:ff:ff:ff
[root@Rocky8 ~]#
```

DEVICE_NAME

建立Cluster – VM網路設定(Master)

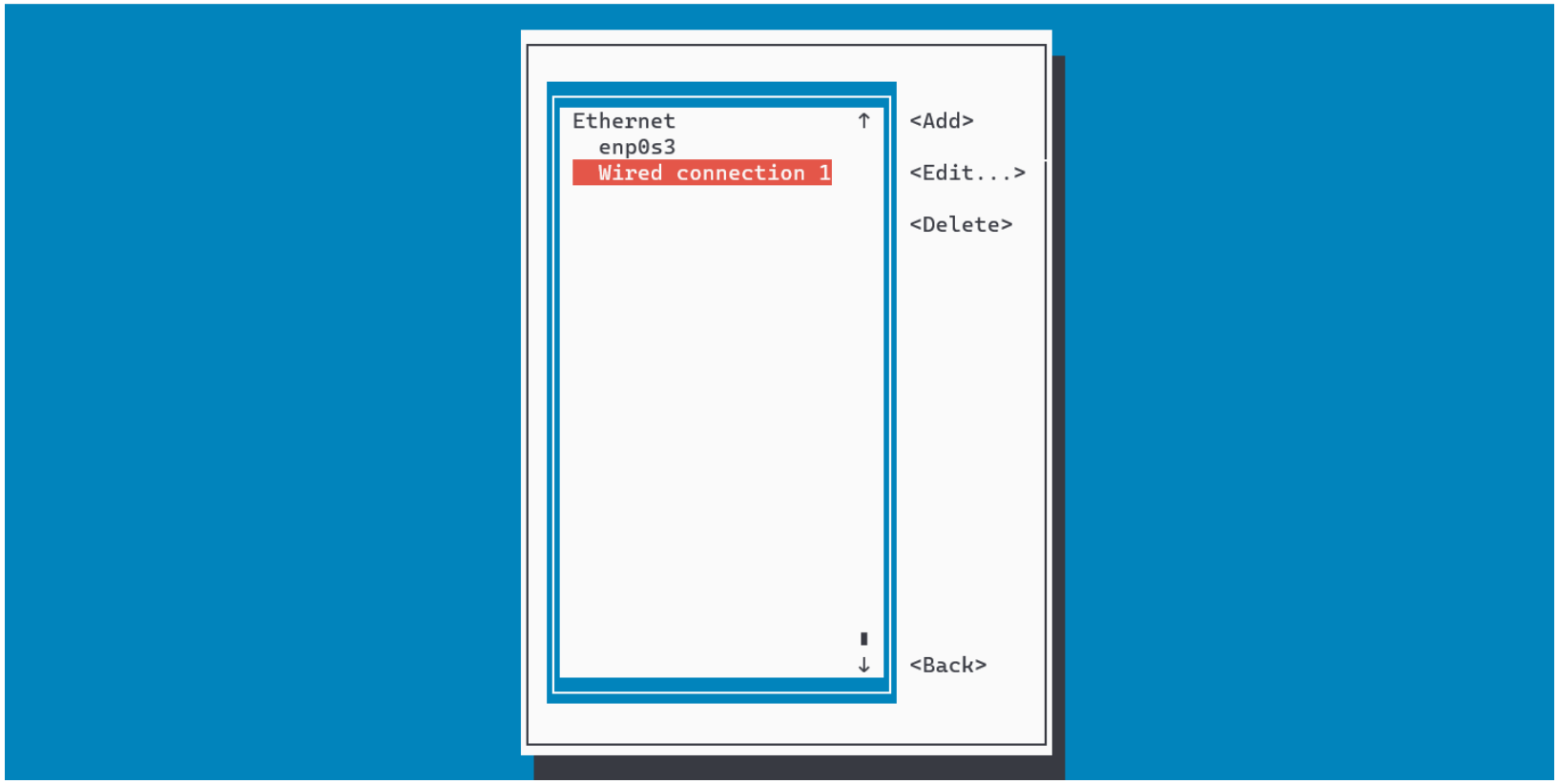
- 設定網路

```
[root@Rocky8 ~]# nmtui
```



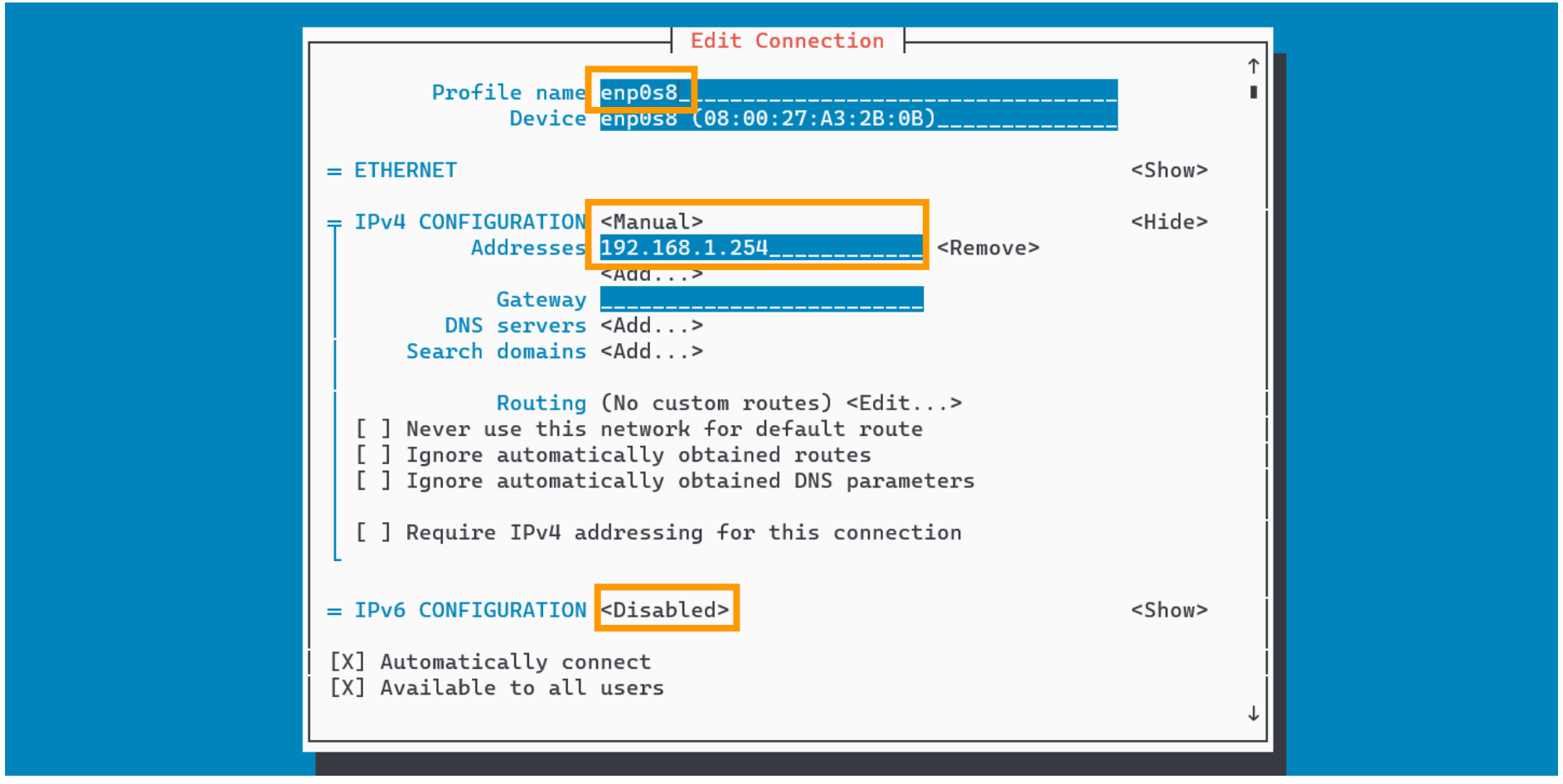
建立Cluster – VM網路設定(Master)

- 設定IP
  - 選擇網路卡 Wired connection 1



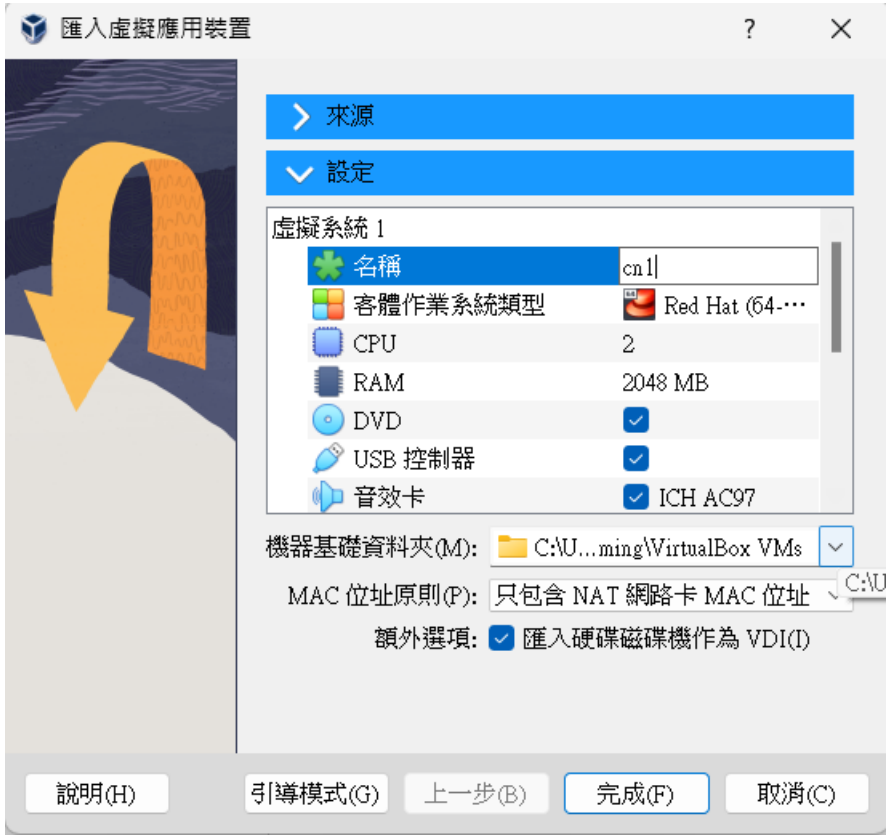
建立Cluster – VM網路設定(Master)

- 設定IP



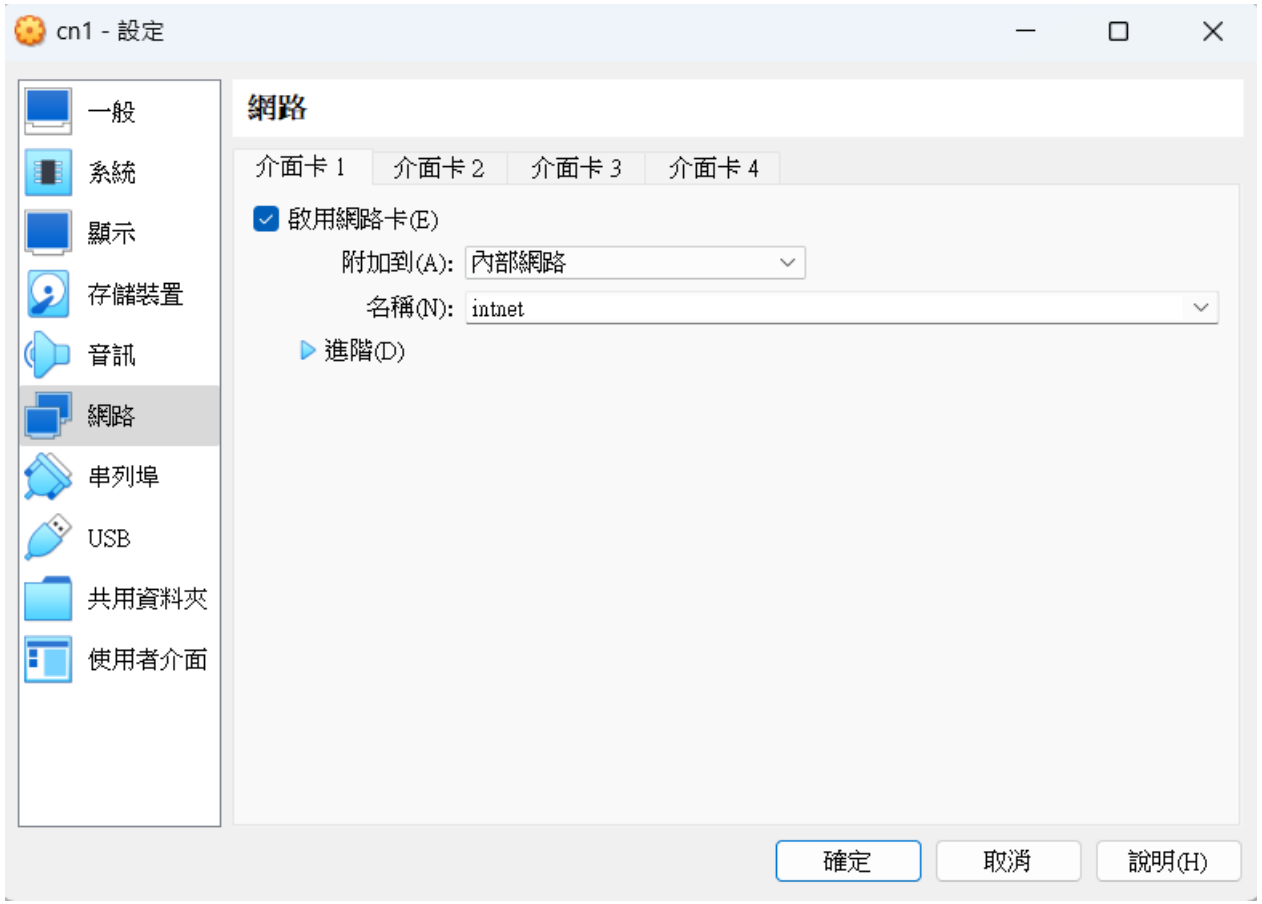
建立Cluster –虛擬機 (Computing Node)

- 匯入 Rocky8.ova
- 虛擬機名稱改成cn1
  (可重複建立不同Computing Node)
- 關閉SELinux



建立Cluster –VM網路設定(cn1)

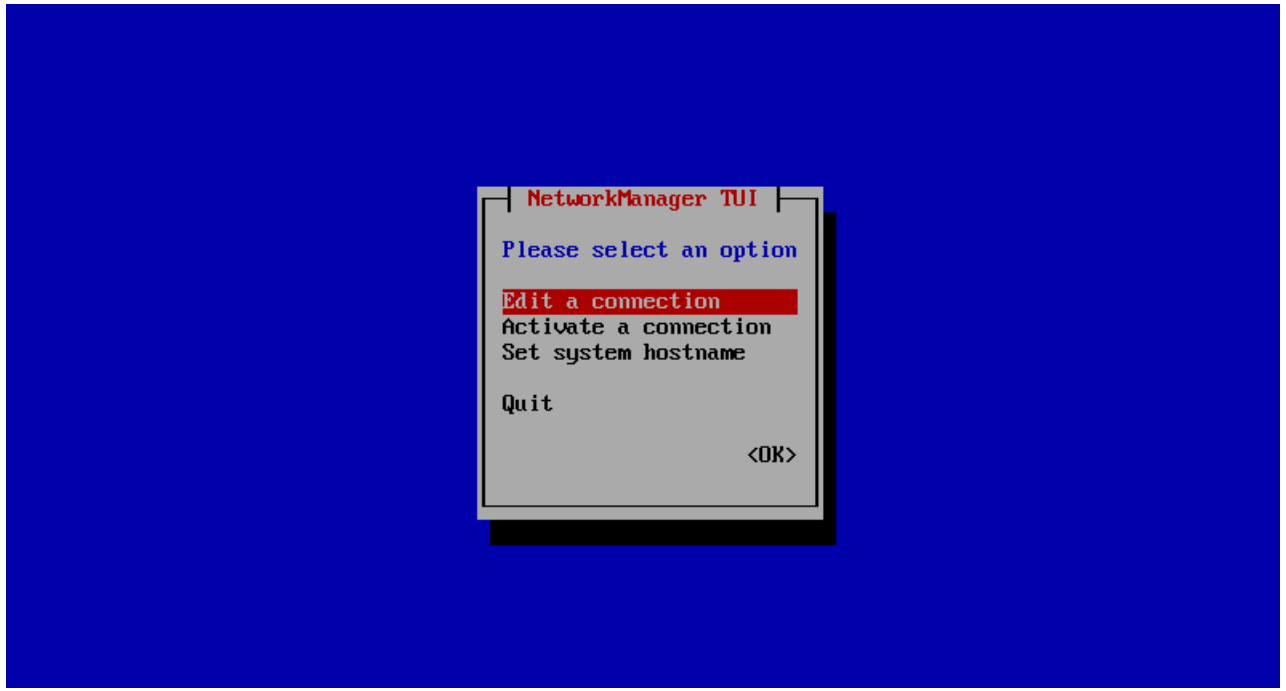
Computing node 的網路卡只對內部連接



建立Cluster –VM網路設定(cn1)

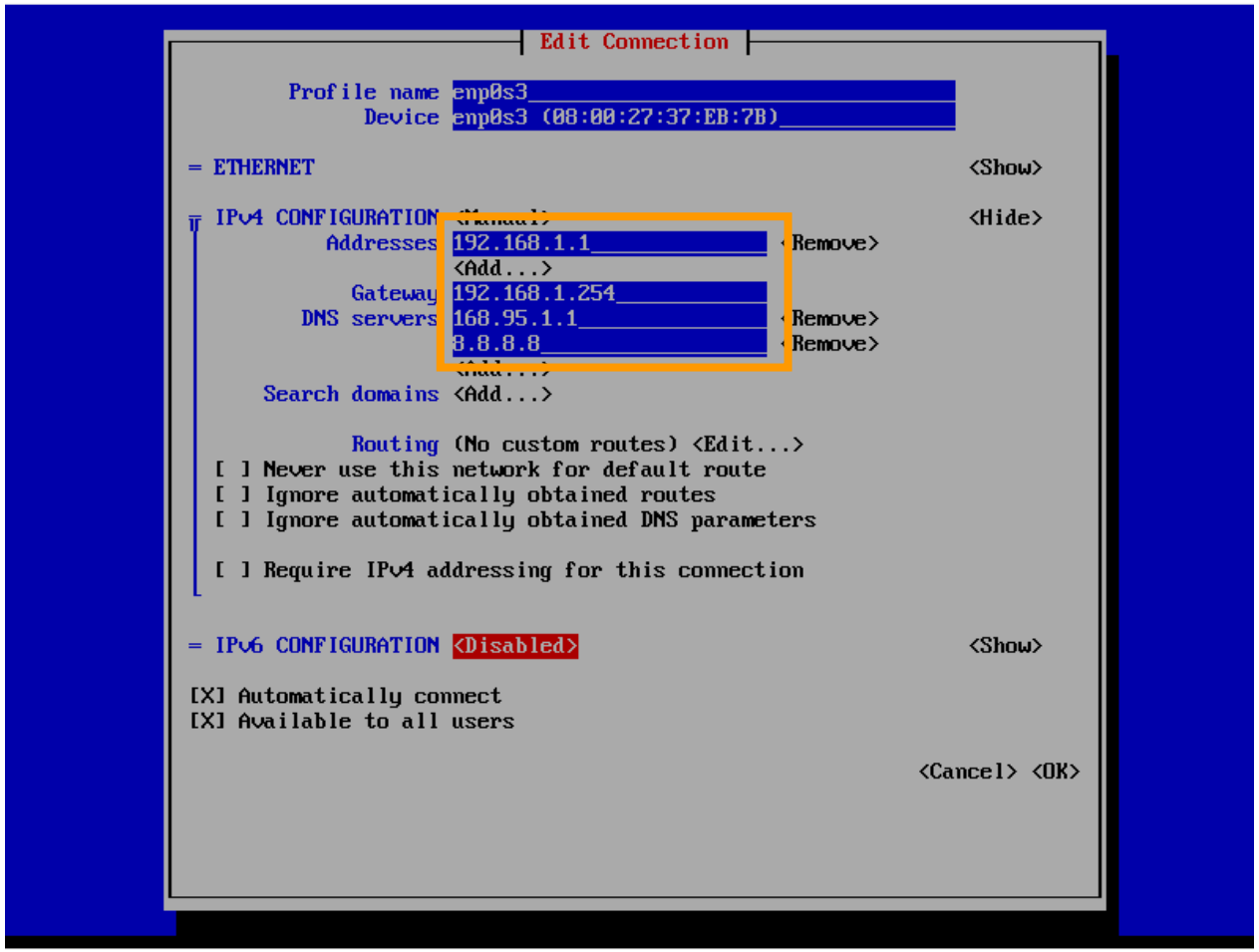
- 設定IP

```
[root@Rocky8 ~]# nmtui
```



建立Cluster –VM網路設定(cn1)

- 設定IP



虛擬機主機名稱設定

- 查詢主機名稱：hostnamectl

```
[root@Rocky8 ~]# hostnamectl
   Static hostname: Rocky8
         Icon name: computer-vm
           Chassis: vm
        Machine ID: f721a3f1401e4116b862c18e45604695
           Boot ID: 20cb3c35922747d4b939e2d1ba40461f
    Virtualization: oracle
  Operating System: Rocky Linux 8.8 (Green Obsidian)
       CPE OS Name: cpe:/o:rocky:rocky:8:GA
            Kernel: Linux 4.18.0-477.10.1.el8_8.x86_64
      Architecture: x86-64
[root@Rocky8 ~]#
```

- 設定主機名稱：hostnamectl set-hostname <HOST_NAME>

```
[root@Rocky8 ~]# hostnamectl set-hostname master
```

- 重新登入後便可看到主機名稱改變

Head / Computing Node 虛擬機網路設定

- 停止並關閉 Head / Compute Node 的預設Firewalld

```
[root@master ~]# systemctl stop firewalld
[root@master ~]# systemctl disable firewalld
```

- 暫時設定 Head Node 的 NAT 服務

NAT 只開起於 Computing Node 軟體更新，平常運作建議關閉

```
[root@master ~]# sysctl net net.ipv4.ip_forward=1
[root@master ~]# iptables -t nat -F
[root@master ~]# iptables -t nat -A POSTROUTING -s 192.168.1.0/24 -j MASQUERADE
```

- 確認Computing Node 可以連線到網際網路

```
[root@cn1 ~]# ping -c 2 8.8.8.8
PING 8.8.8.8 (8.8.8.8) 56(84) bytes of data.
64 bytes from 8.8.8.8: icmp_seq=1 ttl=113 time=8.93 ms
64 bytes from 8.8.8.8: icmp_seq=2 ttl=113 time=5.08 ms

--- 8.8.8.8 ping statistics ---
2 packets transmitted, 2 received, 0% packet loss, time 1003ms
rtt min/avg/max/mdev = 5.077/7.004/8.931/1.927 ms
[root@cn1 ~]# _
```

Head / Computing Node 虛擬機網路設定

- 停止 Head Node 的 NAT 服務

```
[root@master ~]# sysctl net.ipv4.ip_forward=0
[root@master ~]# iptables -t nat -F
or
[root@master ~]# iptables -t nat -D POSTROUTING -s 192.168.1.0/24 -j MASQUERADE
```

Head Node 安裝設定 NIS 服務

- 安裝需要的套件

```
[root@master ~]# yum install ypserv yp-tools
```

- 設定

```
[root@master ~]# nisdomainname cluster
```

```
[root@master ~]# vi /etc/hosts
192.168.1.254 master master.cluster
192.168.1.1 cn1
```

```
[root@master ~]# vi /etc/sysconfig/network
NISDOMAIN=cluster
```

Head Node 安裝設定 NIS 服務

- 設定

```
[root@master ~]# vi /etc/ypserv.conf
192.168.1.0/255.255.255.0  : *      : *            : none
*                          : *      : *            : deny
```

- 啟動服務並設定開機時自動啟動服務

```
[root@master ~]# systemctl start ypserv
[root@master ~]# systemctl start yppasswdd
[root@master ~]# systemctl enable ypserv
[root@master ~]# systemctl enable yppasswdd
```

- 建立NIS資料庫

```
[root@master ~]# /usr/lib64/yp/ypinit -m
```

Head Node 安裝設定 NIS 服務

- 確認 rpcbind 啟動

```
[root@master ~]# rpcinfo -p
   program vers proto   port  service
    100000    4   tcp    111  portmapper
    100000    3   tcp    111  portmapper
    100000    2   tcp    111  portmapper
    100000    4   udp    111  portmapper
    100000    3   udp    111  portmapper
    100000    2   udp    111  portmapper
    100004    2   udp    943  ypserv
    100004    1   udp    943  ypserv
    100004    2   tcp    946  ypserv
    100004    1   tcp    946  ypserv
    100009    1   udp    987  yppasswdd
[root@master ~]# _
```

Computing Node 設定 NIS 服務

- 安裝需要的套件

```
[root@cn1 ~]# yum install ypbind yp-tools
```

- 設定

```
[root@cn1 ~]# nisdomainname cluster
```

```
[root@cn1 ~]# vi /etc/hosts
192.168.1.254 master master.cluster
192.168.1.1  cn1
```

```
[root@cn1 ~]# vi /etc/sysconfig/network
NISDOMAIN=cluster
```

Computing Node 設定 NIS 服務

- 設定

```
[root@cn1 ~]# vi /etc/yp.conf
domain cluster server master
```

- 設定認證機制

```
[root@cn1 ~]# vi /etc/sysconfig/authconfig
USENIS=yes
```

```
[root@cn1 ~]# vi /etc/pam.d/system-auth
password    sufficient pam_unix.so try_first_pass use_authtok nullok sha512 shadow nis
```

Computing Node 設定 NIS 服務

- 設定認證機制

```
[root@cn1 ~]# vi /etc/nsswitch.conf
passwd:     files sss nis systemd
shadow:     files sss nis
group:      files sss nis systemd
hosts:      files nis dns myhostname
```

- 啟動服務並設定自動啟用

```
[root@cn1 ~]# systemctl start ypbind
[root@cn1 ~]# systemctl enable ypbind
```

Computing Node 設定 NIS 服務

- 確認 rpcbind 啟動

```
[root@cn1 ~]# rpcinfo -p
   program vers proto   port  service
    100000    4   tcp    111  portmapper
    100000    3   tcp    111  portmapper
    100000    2   tcp    111  portmapper
    100000    4   udp    111  portmapper
    100000    3   udp    111  portmapper
    100000    2   udp    111  portmapper
    100007    2   udp    714  ypbind
    100007    1   udp    714  ypbind
    100007    2   tcp    717  ypbind
    100007    1   tcp    717  ypbind
[root@cn1 ~]#
```

- 測試 NIS : “yptest”
- 確認連接的NIS Server : “ypwhich”
- 取得NIS資料庫的內容 ： “ypcat <NIS_MAP>”

NIS Client無法正常連接Server? 看看防火牆是否關閉

Head Node 建立使用者

- 新增使用者，用 -c 設定 Full Name

```
[root@master ~]# useradd -c “User 1” user1
```

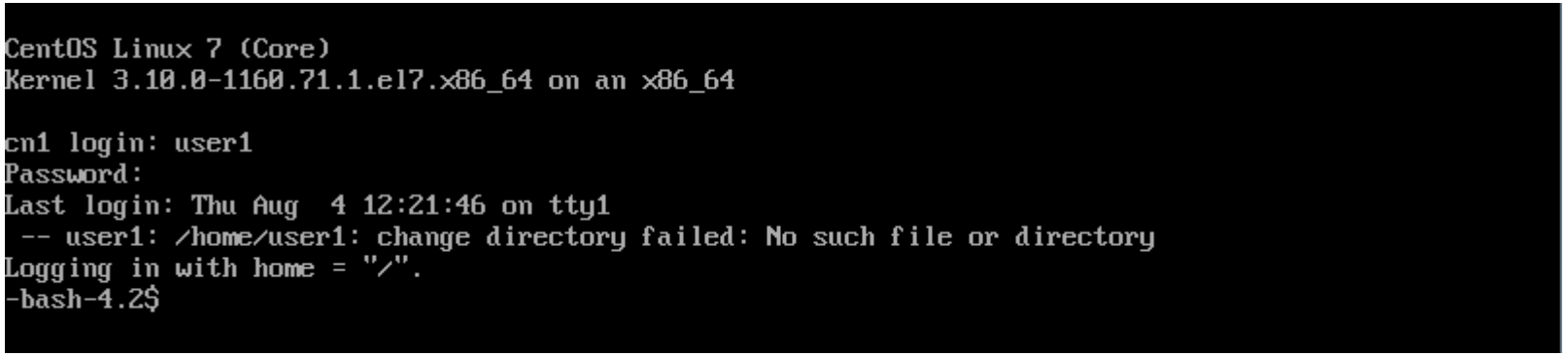
- 設定使用者密碼

```
[root@master ~]# passwd user1
```

- 更新 NIS 資料庫

```
[root@master ~]# make -C /var/yp
```

登入使用者到 Computing Node

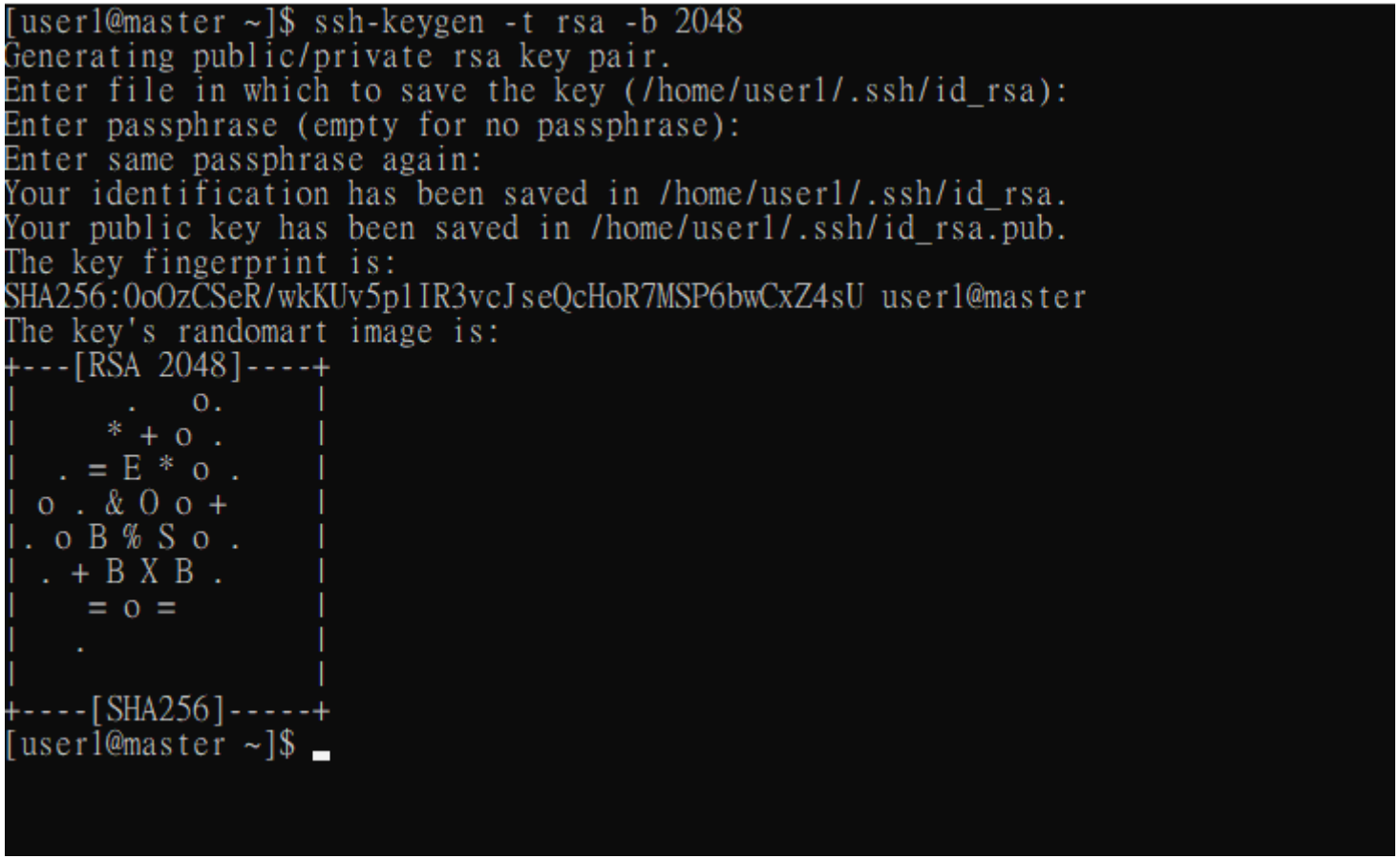


- 使用者更改密碼

```
[user1@cn1 ~]$ yppasswd
```

建立免敲密碼登入

- 建立一對 rsa 密鑰：ssh-keygen -t rsa -b 2048



建立免敲密碼登入

- 複製 rsa 公鑰 (public key) 至 authorized_keys

```
[user1@master ~]# cp ~/.ssh/id_rsa.pub ~/.ssh/authorized_keys
```

- 建立 root 免敲密碼：
  - 建立一對 rsa 密鑰：ssh-keygen -t rsa -b 2048
  - 複製 rsa 公鑰 (public key) 至 authorized_keys
  - 複製整個 ~/.ssh/ 到 Compute Node

```
[root@master ~]# scp -r ~/.ssh/ cn1:~/
```

Head Node 安裝設定 NFS 服務

- 安裝需要的套件

```
[root@master ~]# yum install nfs-utils
```

- 設定 NFS 分享的掛載點

```
[root@master ~]# mkdir /software
[root@master ~]# vi /etc/exports
/home          192.168.1.0/24(insecure,rw,async,no_root_squash)
/software      192.168.1.0/24(insecure,rw,async,no_root_squash)
/opt           192.168.1.0/24(insecure,rw,async,no_root_squash)
```

常用參數	意義	預設值
rw, ro	讀寫模式(rw: read-write, ro: read-only)	rw
async, sync	記憶體磁碟同步模式	async
no_root_squash, root_squash	是否壓縮client端的root身份為nfsnobody	root_squash
all_squash	一律把client端的使用者壓縮成nobody	null

Head Node 安裝設定 NFS 服務

- 啟動服務

```
[root@master ~]# systemctl start nfs-server
```

- 設定開機時自動啟動服務

```
[root@master ~]# systemctl enable nfs-server
```

Computing Node 設定 NFS 服務

- 安裝需要的套件

```
[root@cn1 ~]# yum install nfs-utils
```

- 手動掛載 NFS 載點

```
[root@cn1 ~]# mount -t nfs master:/home /home
```

- 開機自動掛載 NFS 載點

```
[root@cn1 ~]# vi /etc/fstab
master:/home      /home      nfs   defaults    0 0
master:/software  /software  nfs   defaults    0 0
master:/opt       /opt       nfs   defaults    0 0
```

更多的掛載參數：man mount

Head Node 安裝設定 NTP 服務

- 安裝需要的套件

```
[root@master ~]# yum install chrony
```

- 設定 NTP 服務

```
[root@master ~]# vi /etc/chrony.conf
# Use public servers from the pool.ntp.org project.
# Please consider joining the pool (http://www.pool.ntp.org/join.html).
#server 2.rocky.pool.ntp.org iburst
server tock.stdtime.gov.tw iburst
server clock.stdtime.gov.tw iburst
server tick.stdtime.gov.tw iburst
server time.stdtime.gov.tw iburst

# Allow NTP client access from local network
# allow 192.168.0.0/16
allow 192.168.1.0/24
```

Head Node 安裝設定 NTP 服務

- 重新啟動服務

```
[root@master ~]# systemctl restart chronyd
```

- 觀察校時目的server

```
[root@master ~]# chronyc sources
```

- 手動自動校時

```
[root@master ~]# chronyc -a makestep
```

- 查看校時的詳情

```
[root@master ~]# chronyc tracking
```

Head Node 安裝設定 NTP 服務

- 顯示時間相關設定

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

- 設定時區

```
[root@master ~]# timedatectl set-timezone Asia/Taipei
```

- 手動設定時間

```
[root@master ~]# timedatectl set-time “YYYY-mm-dd H:m:s”
```

Computing Node 安裝設定 NTP

- 設定 NTP 服務

```
[root@cn1 ~]# vi /etc/chrony.conf
# Use public servers from the pool.ntp.org project.
# Please consider joining the pool (http://www.pool.ntp.org/join.html).
#server 2.rocky.pool.ntp.org iburst
server master iburst
```

- 重新啟動服務

```
[root@cn1 ~]# systemctl restart chronyd
```

Computing Node 安裝設定 NTP

- 觀察校時目的server

```
[root@cn1 ~]# chronyc sources
```

- 手動對 Head Node 校時

```
[root@cn1 ~]# chronyc –a makestep
```

- 查看校時的詳情

```
[root@cn1 ~]# chronyc tracking
```

回家作業

- 請在自己的電腦從頭開始架設 VM 與 Linux 環境
  - Rocky Linux 8 (https://rockylinux.org/zh_TW/)
  - Ubuntu Server 22.04 (https://www.ubuntu-tw.org/)
- 練習第一天上課的Linux指令至少 10 個
- 請下載最新版 GCC、 intel® oneapi base & hpc toolkit (離線版本)
  - https://gcc.gnu.org/
  - https://www.intel.com/content/www/us/en/developer/tools/oneapi/overview.html