Linux Cluster System

2024.03.24

Ref: Tzu-Wen Wang
Chih-Yu Hsieh, Shih-Rou Cheng, Wei-Chih Huang

NARLabs 財團法人國家實驗研究院
國家高速網路與計算中心
National Center for High-performance Computing
國立清華大學叢集電腦競賽團隊
Student Cluster Competition Team of NTHU

Outline

- BMC Introduction
- PVE Introduction
- Cluster Topology
- Linux Installation
  - Prepare images
  - Partition
- Network Setup
- Linux File Hierarchy
- NFS
- User Authentication

BMC Introduction

Baseboard Management Controller

- An independent chip on server’s motherboard
- Power control
- Fan speed control
- Virtual media: iso redirect
- **Console redirection**

Proxmox Virtual Environment (PVE)

- Open-source platform for virtualization
- Built-in web interface
- Easily manage VMs and containers, software-defined storage and networking.
- We use PVE in this training camp.
- Each group will have five virtual machine host on our PVE server.

VPN

Connect to PVE

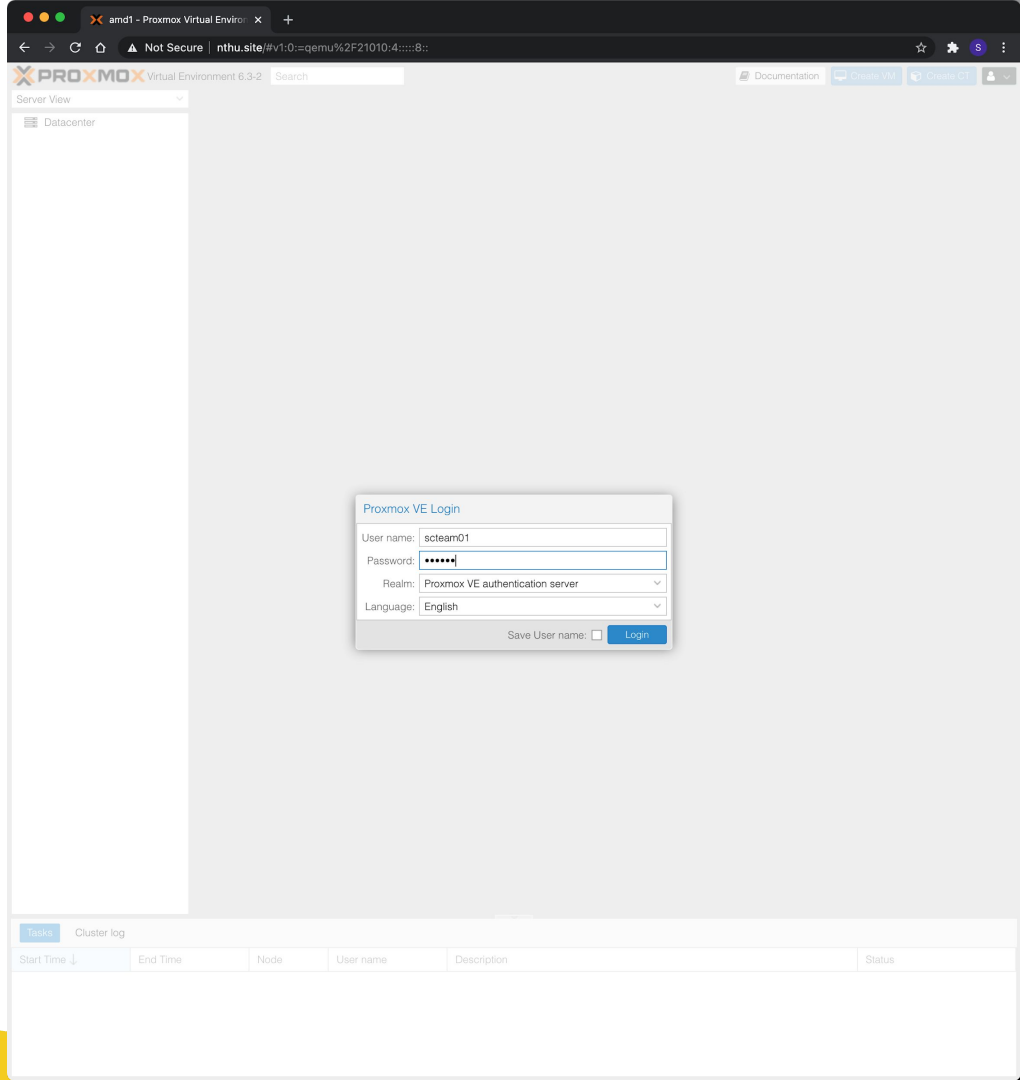
wifi: HPC Camp 1 5G

Password : 20240127

Visit pve.nthu.site

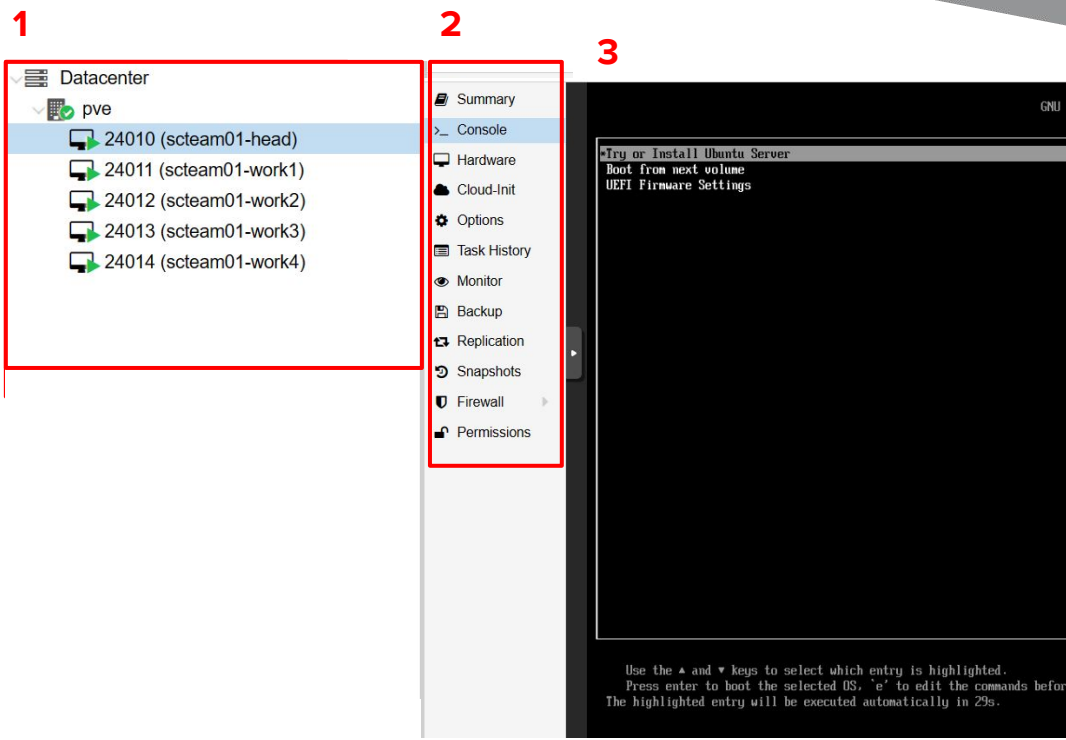
Login to PVE

- Account: scteam0X
  - X is your team number
- Password: nthusc24



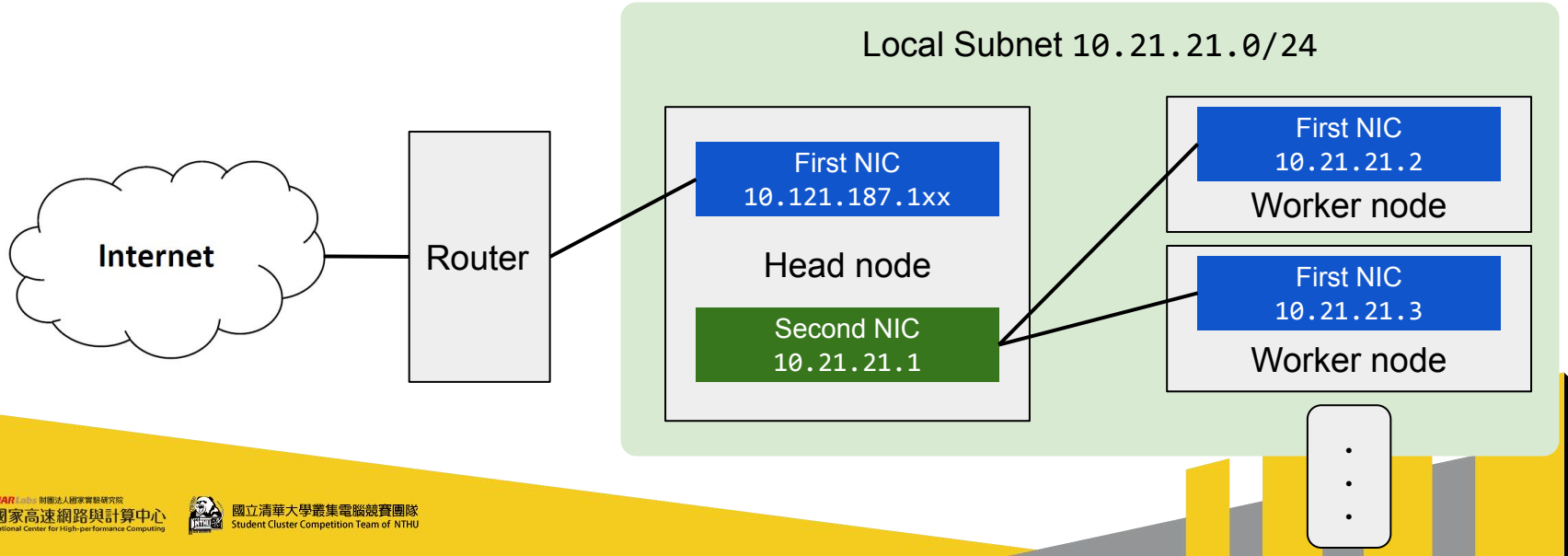
Interface

1. Your five VM here
2. VM's detail
   a. Console: VNC console
   b. Hardware: Hardware spec
3. VNC console: We will use this interface to install Linux before we setup ssh.



Cluster Topology

- There are two node each group.
- Head node is the login node. It's responsible for hold userdata (`/home`), NIS server and forward IP.



Linux Installation

Linux Installation

- We have already prepare the image for you.
- Please select English. This will save time debugging language issues.
- You can choose your own hostname (E.g. `bigmac01, bigmac02`) or simply use VM id (E.g. vm23XX0, vm23XX1)

Steps

1. Try to install ubuntu
2. Select English
3. Continue without updating
4. No need to change “keyboard configuration”
5. No need to change “type of install”
6. Set up network
    a. Head node
    b. Work node

Network Configuration

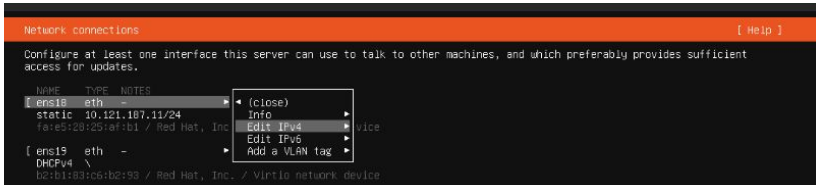
- Use static IPv4
- There are two NIC on head node (net0, net1), one on worker node (net0).
- For head node, please set net0 IP to `10.121.187.11X`
  - Netmask: `255.255.255.0`
  - Gateway: `10.121.187.254`
  - X is your team number. (2~7)
  - For example, team2 should set head node net0 IP to `10.121.187.12`
- Set head node net1 IP to `10.21.21.1`
- Set worker node net0 IP to `10.21.21.2~10.21.21.5`

Network-head

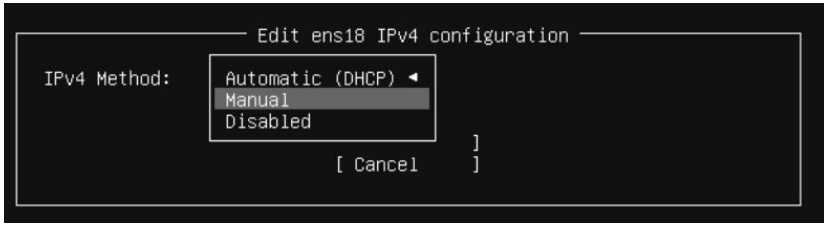
Network configuration

ens18 -> Edit IPv4 -> Manual

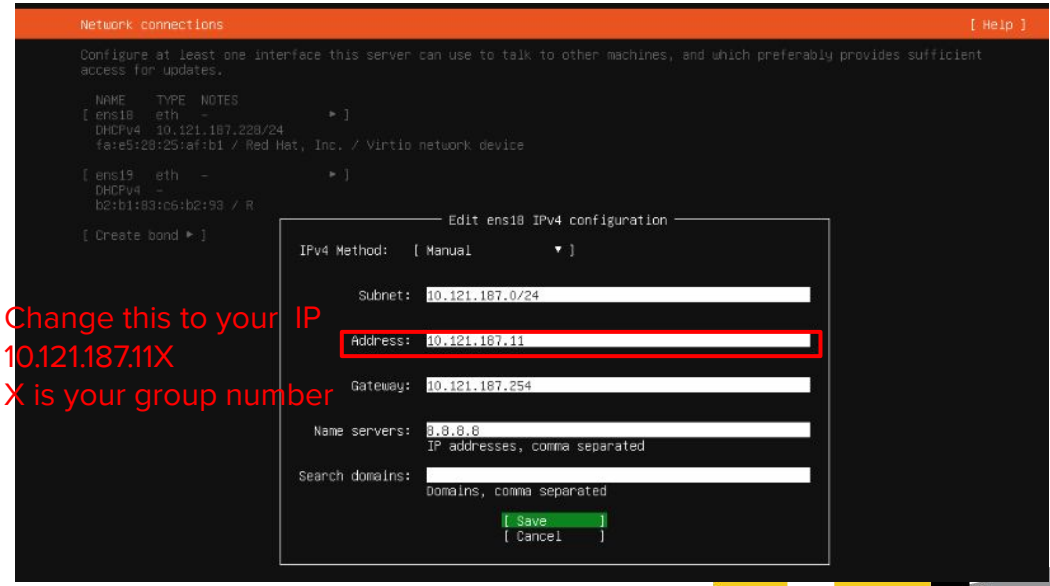
1



2

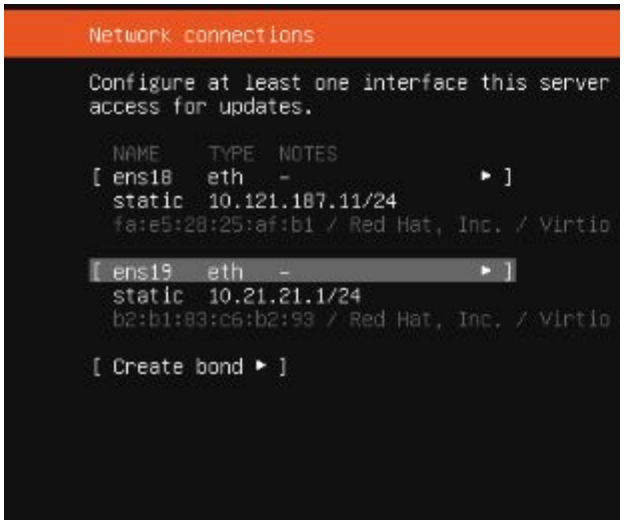


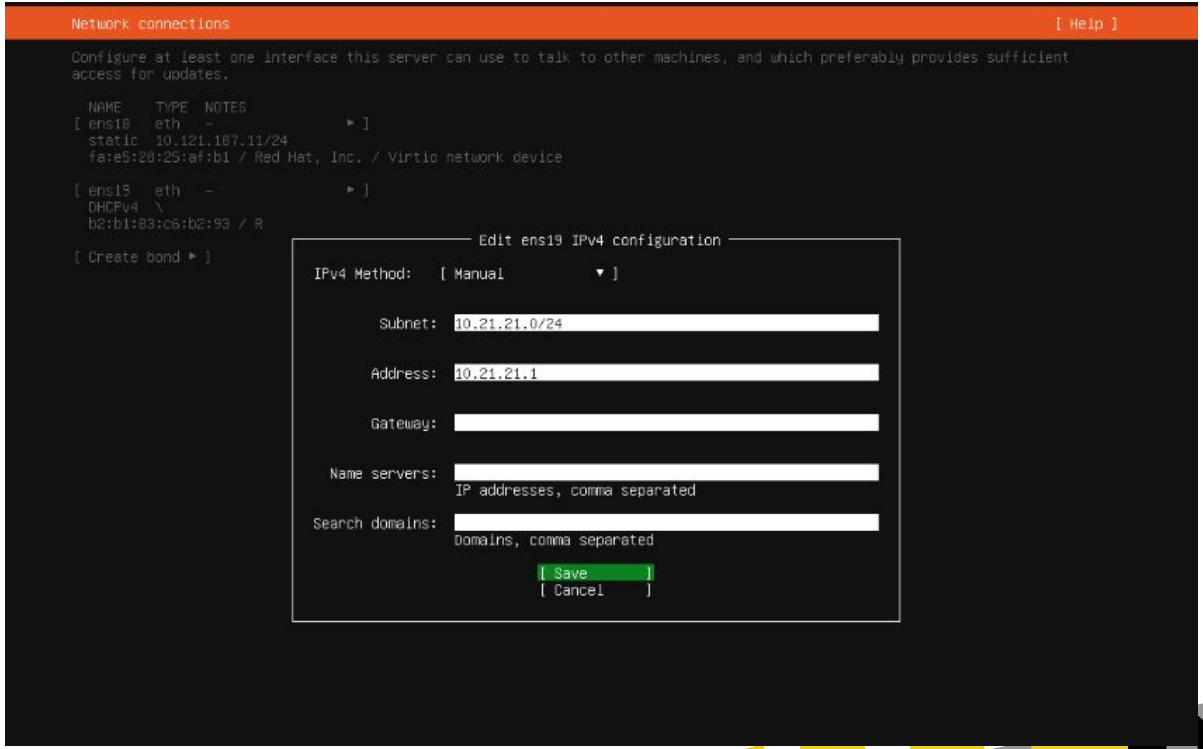
3



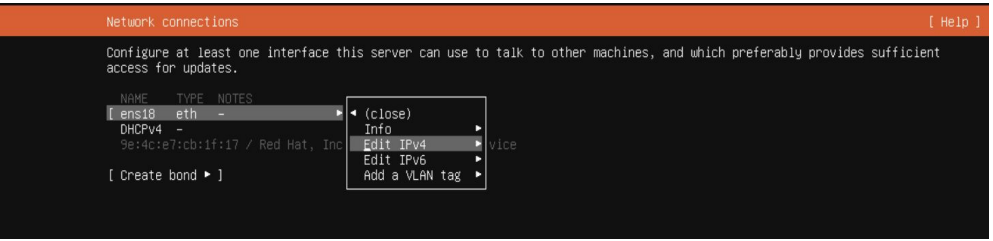
Change this to your IP
10.121.187.11X
X is your group number

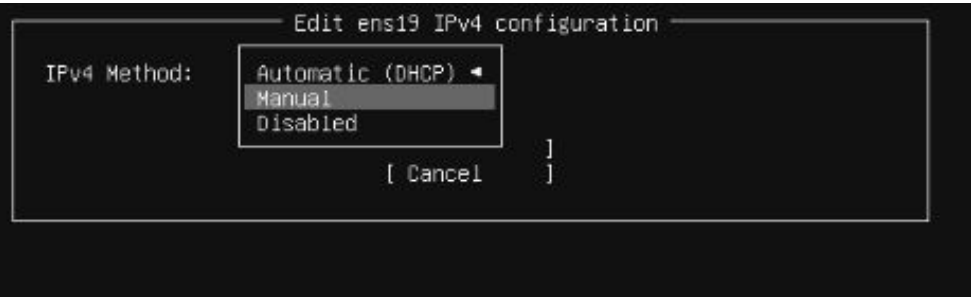
ens19-> Edit IPv4 -> Manual

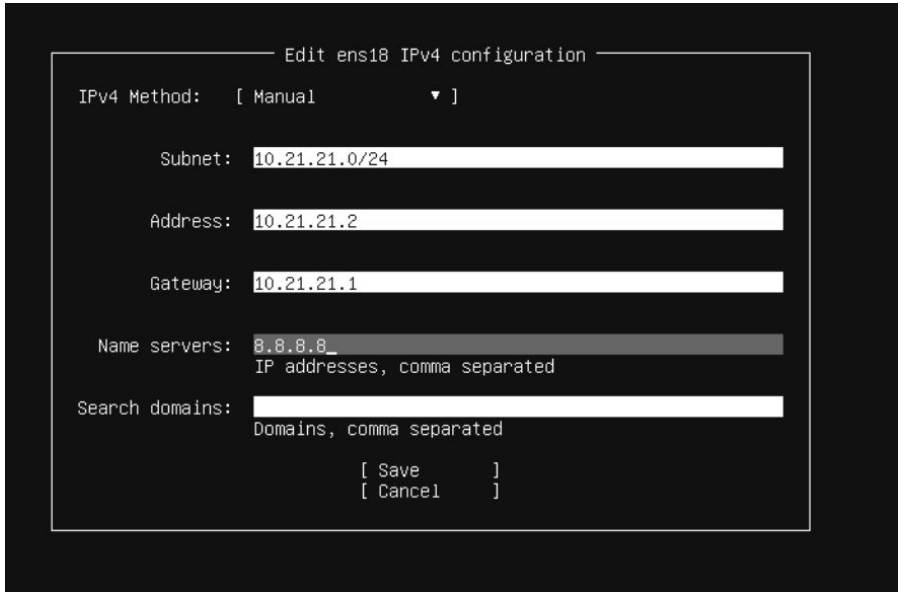




Network-work

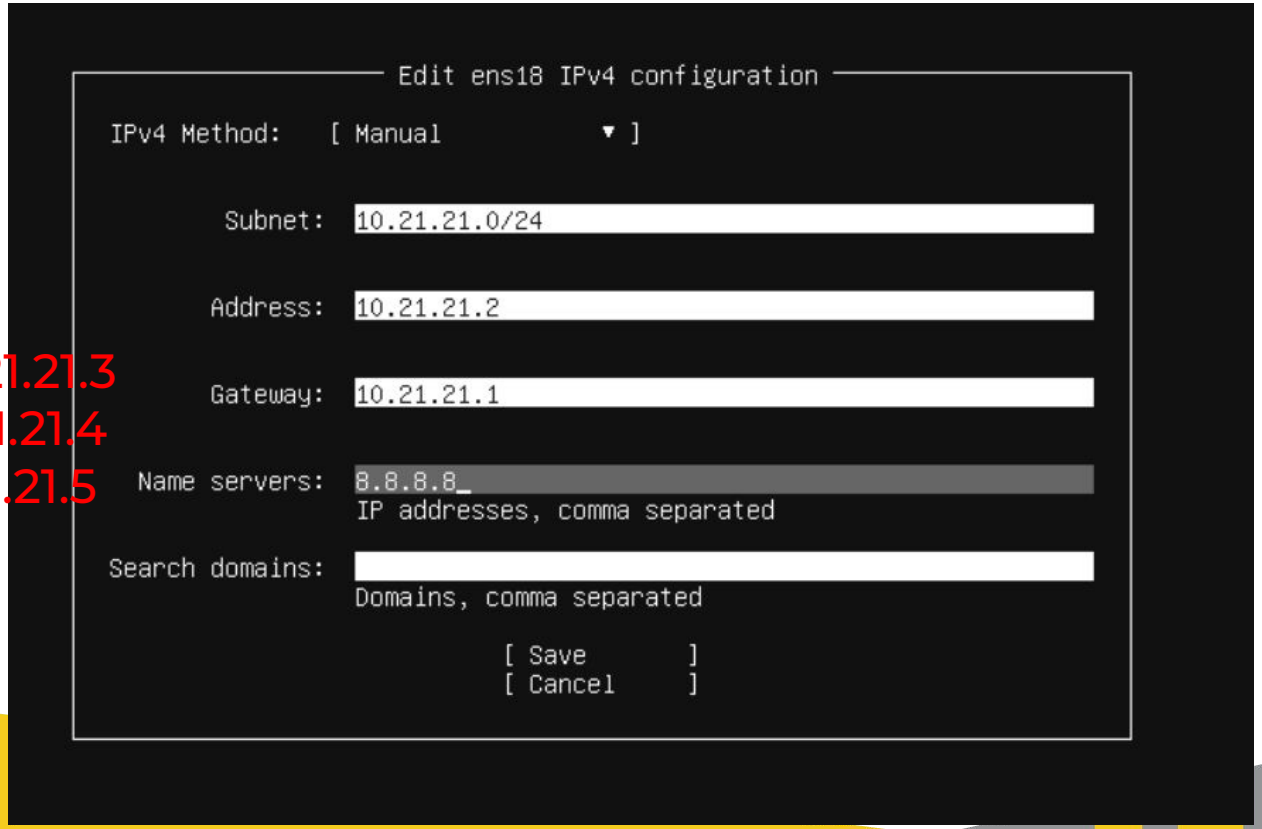

Network connections
[Help]
Configure at least one interface this server can use to talk to other machines, and which preferably provides sufficient access for updates.
NAME TYPE NOTES
[ens18 eth -
DHCPv4 -
9e:4c:e7:cb:1f:17 / Red Hat, Inc
[Create bond ▸]
◂ (close)
Info ▸
Edit IPv4 ▸
Edit IPv6 ▸
Add a VLAN tag ▸


Edit ens19 IPv4 configuration
IPv4 Method:
Automatic (DHCP) ◂
Manual
Disabled
]
[Cancel]


Edit ens18 IPv4 configuration
IPv4 Method: [Manual ▾]
Subnet: 10.21.21.0/24
Address: 10.21.21.2
Gateway: 10.21.21.1
Name servers: 8.8.8.8_
IP addresses, comma separated
Search domains:
Domains, comma separated
[Save]
[Cancel]

Net0 in work node(Take work1 for example)

Address
Work2 :10.21.21.3
Work3:10.21.21.4
work4:10.21.21.5

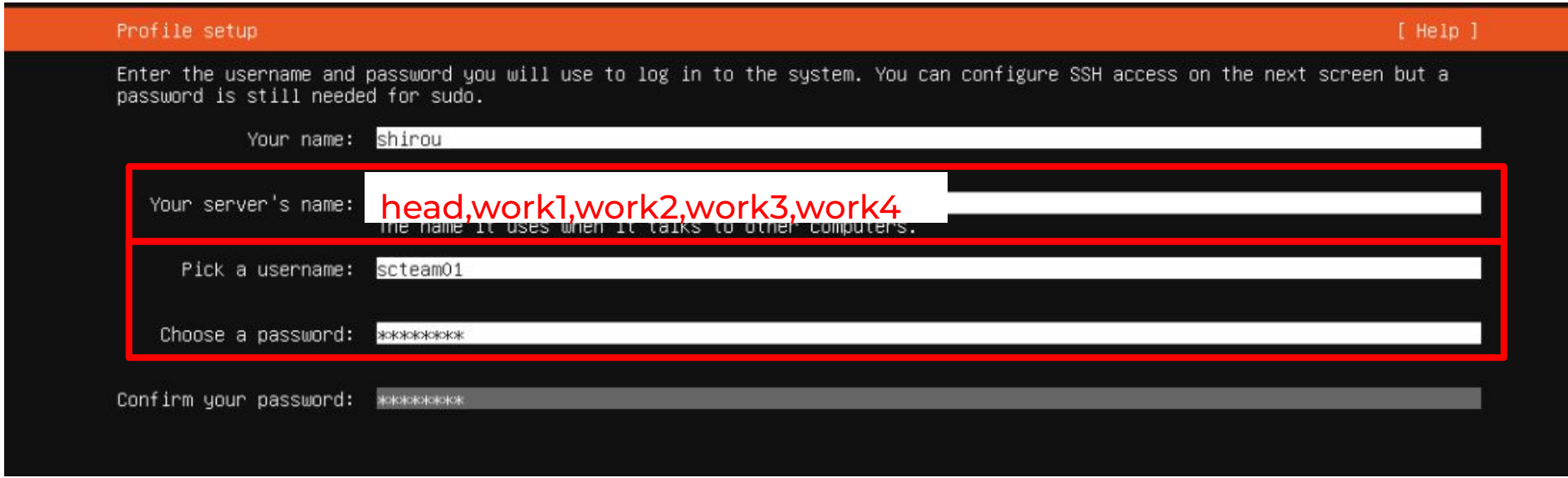


Steps

1. Select English
2. Continue without updating
3. Set up network
4. No need to set proxy.
5. Use default mirror site.
6. Guided storage configuration: Use an entire dick (default). Or you can set your own partition if you want.
7. NO need to change storage configuration
8. Confirm destructive action =>continue

Profile setup

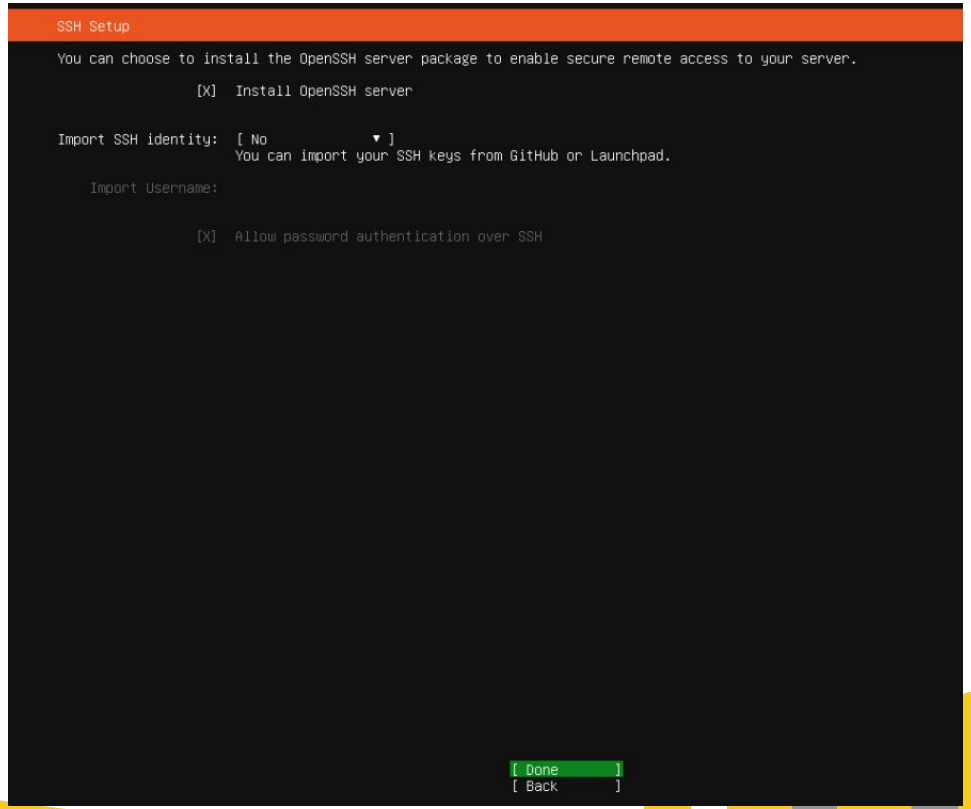
{username}@{server's name}



Set username:scteam0X
X is your team number

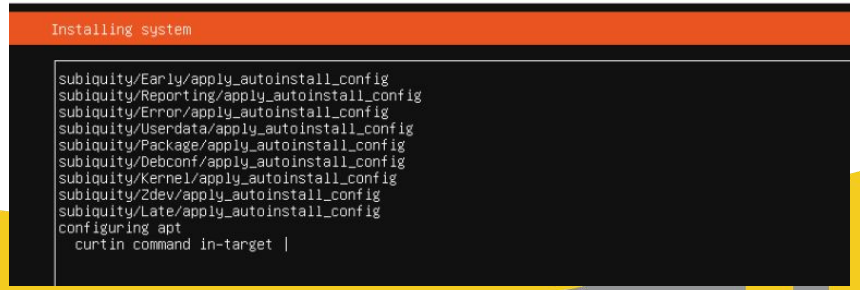
SSH Setup

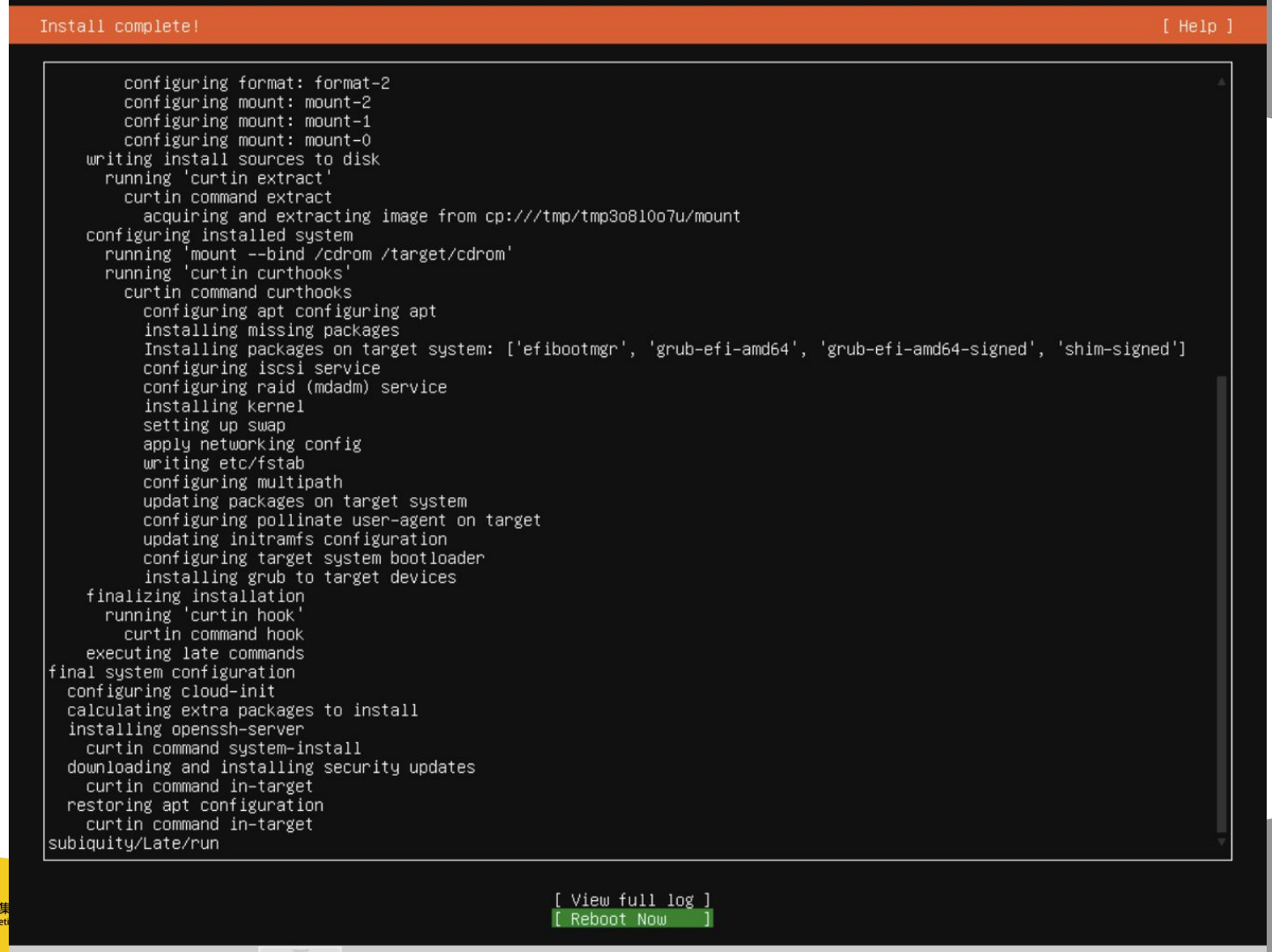
ens18 -> Edit IPv4 -> Manual



Steps

1. Select English
2. Continue without updating
3. Set up network
4. No need to set proxy.
5. Use default mirror site.
6. Guided storage configuration: Use an entire dick (default). Or you can set your own partition if you want.
7. NO need to change storage configuration
8. Confirm destructive action =>continue
9. No need to change Third-parity drivers
10. No need to set Features Server Snaps
11. Start installing

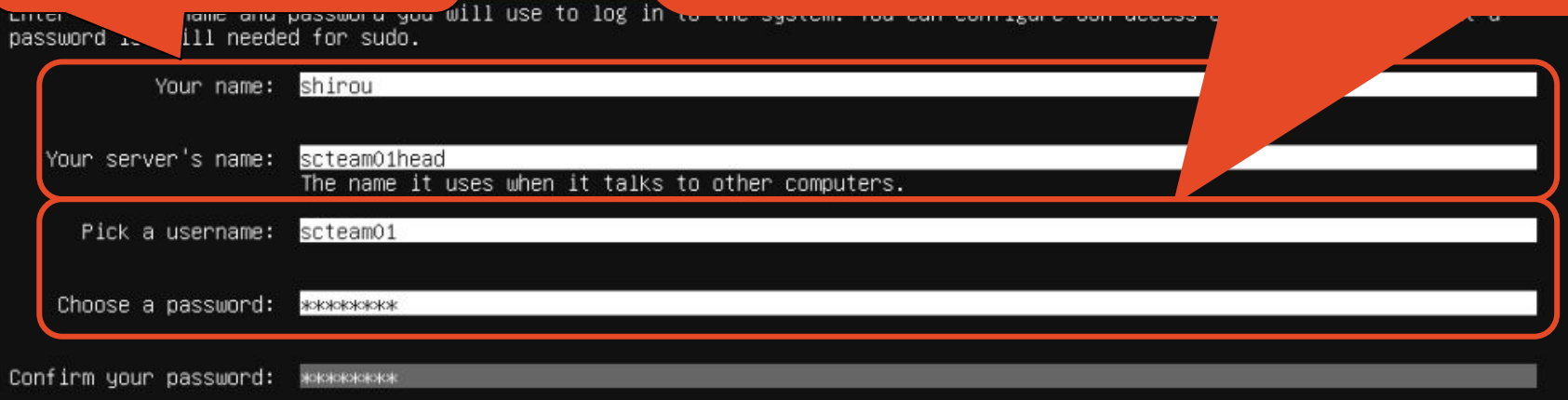



Install complete!
[Help]
configuring format: format-2
configuring mount: mount-2
configuring mount: mount-1
configuring mount: mount-0
writing install sources to disk
running 'curtin extract'
curtin command extract
acquiring and extracting image from cp:///tmp/tmp3o8l0o7u/mount
configuring installed system
running 'mount --bind /cdrom /target/cdrom'
running 'curtin curthooks'
curtin command curthooks
configuring apt configuring apt
installing missing packages
Installing packages on target system: ['efibootmgr', 'grub-efi-amd64', 'grub-efi-amd64-signed', 'shim-signed']
configuring iscsi service
configuring raid (mdadm) service
installing kernel
setting up swap
apply networking config
writing etc/fstab
configuring multipath
updating packages on target system
configuring pollinate user-agent on target
updating initramfs configuration
configuring target system bootloader
installing grub to target devices
finalizing installation
running 'curtin hook'
curtin command hook
executing late commands
final system configuration
configuring cloud-init
calculating extra packages to install
installing openssh-server
curtin command system-install
downloading and installing security updates
curtin command in-target
restoring apt configuration
curtin command in-target
subiquity/Late/run
[View full log]
[Reboot Now]

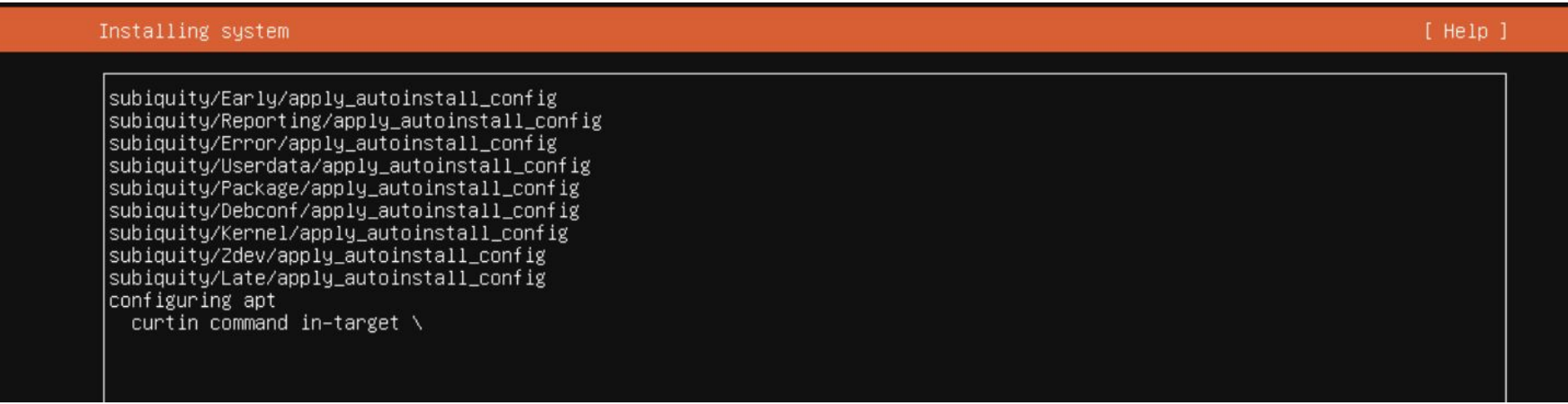
Profile Setup – Take head for example

"Your name" is optional

- You can choose whatever you want for your **"server's name", "username", and "password"**
- However, for your convenience, you might want to **set username and passwd to be the same** :D

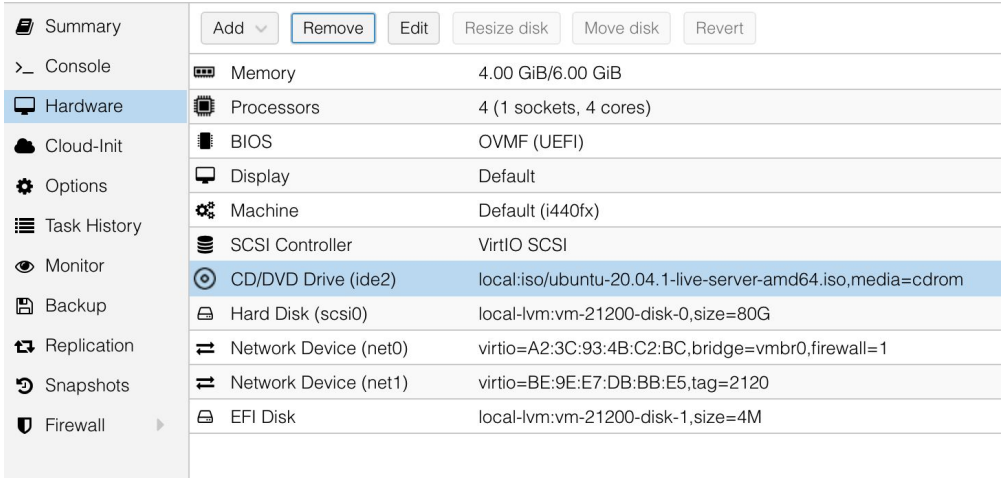


It’s installing!

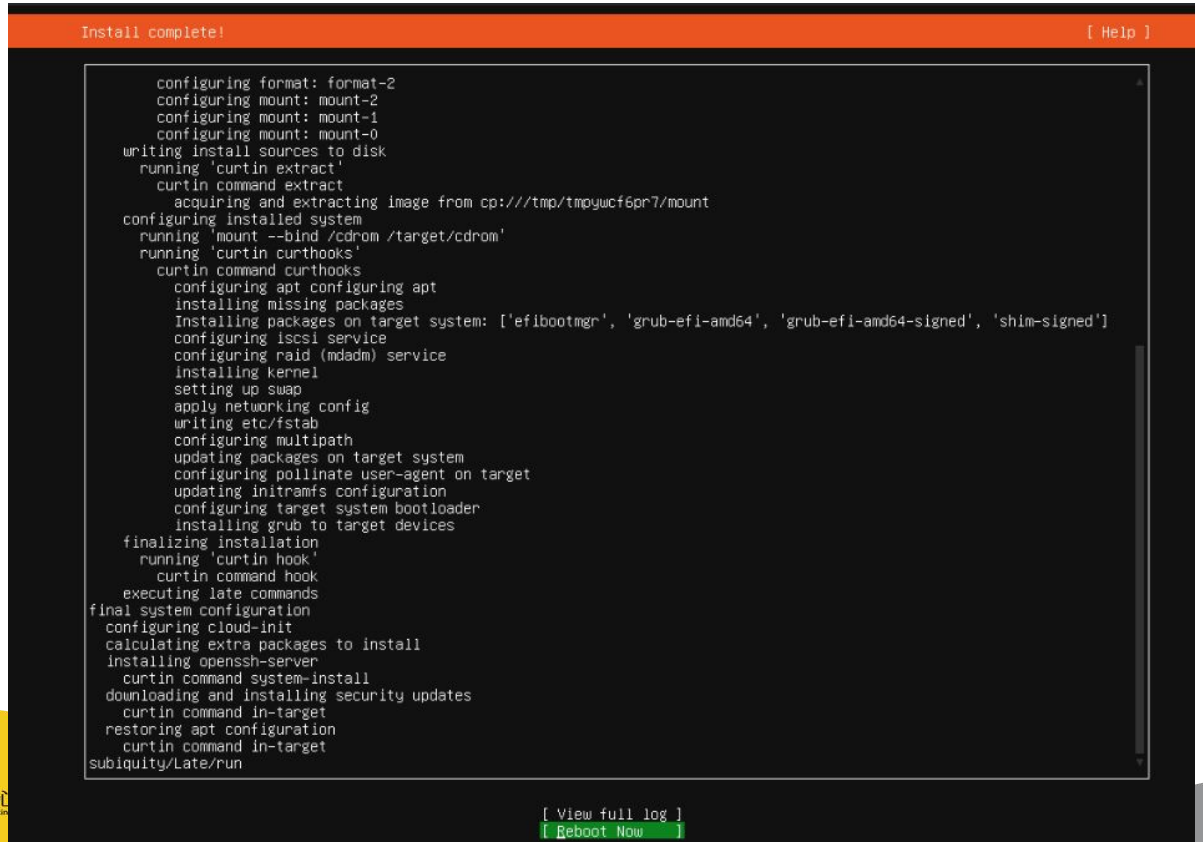


Remove Disk

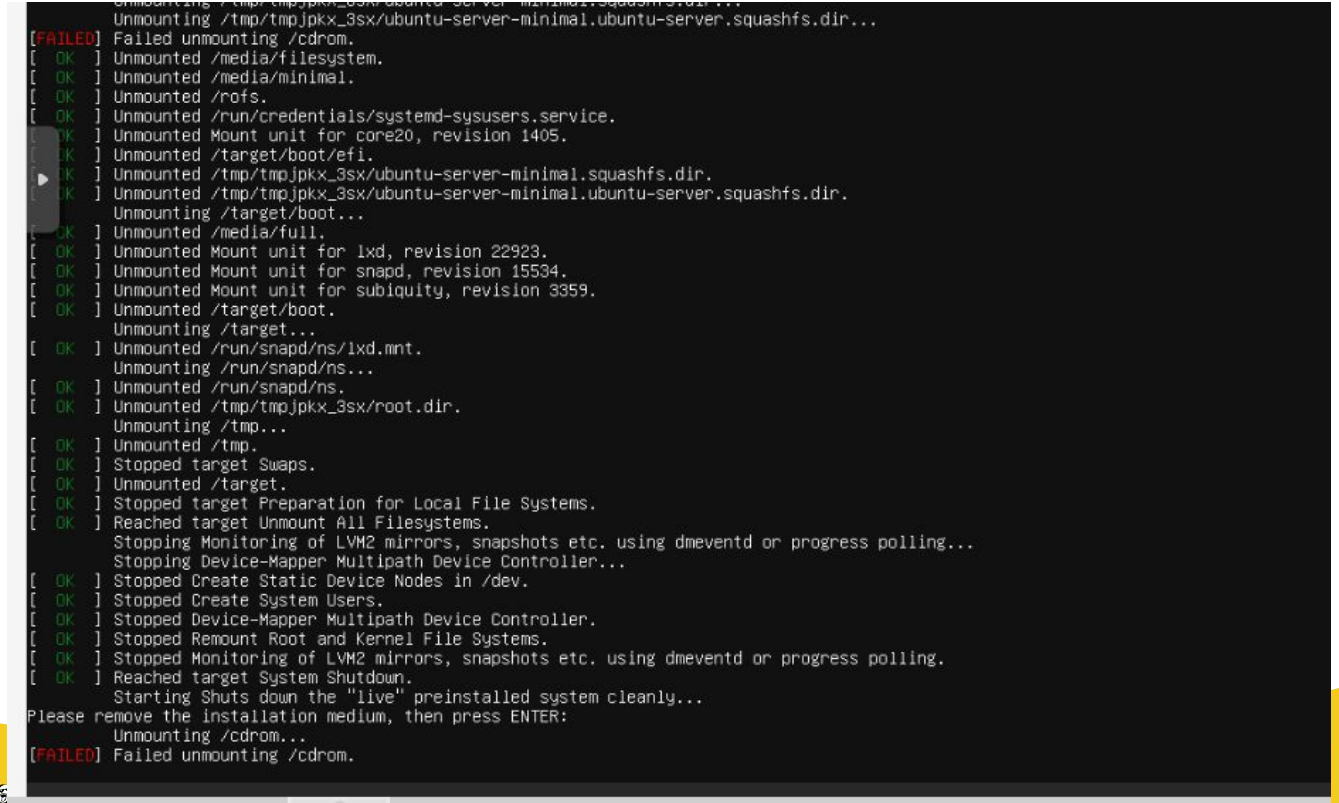
- In Hardware page.
- Double click CD/DVD.
- Choose Do not use.
- Then reboot.



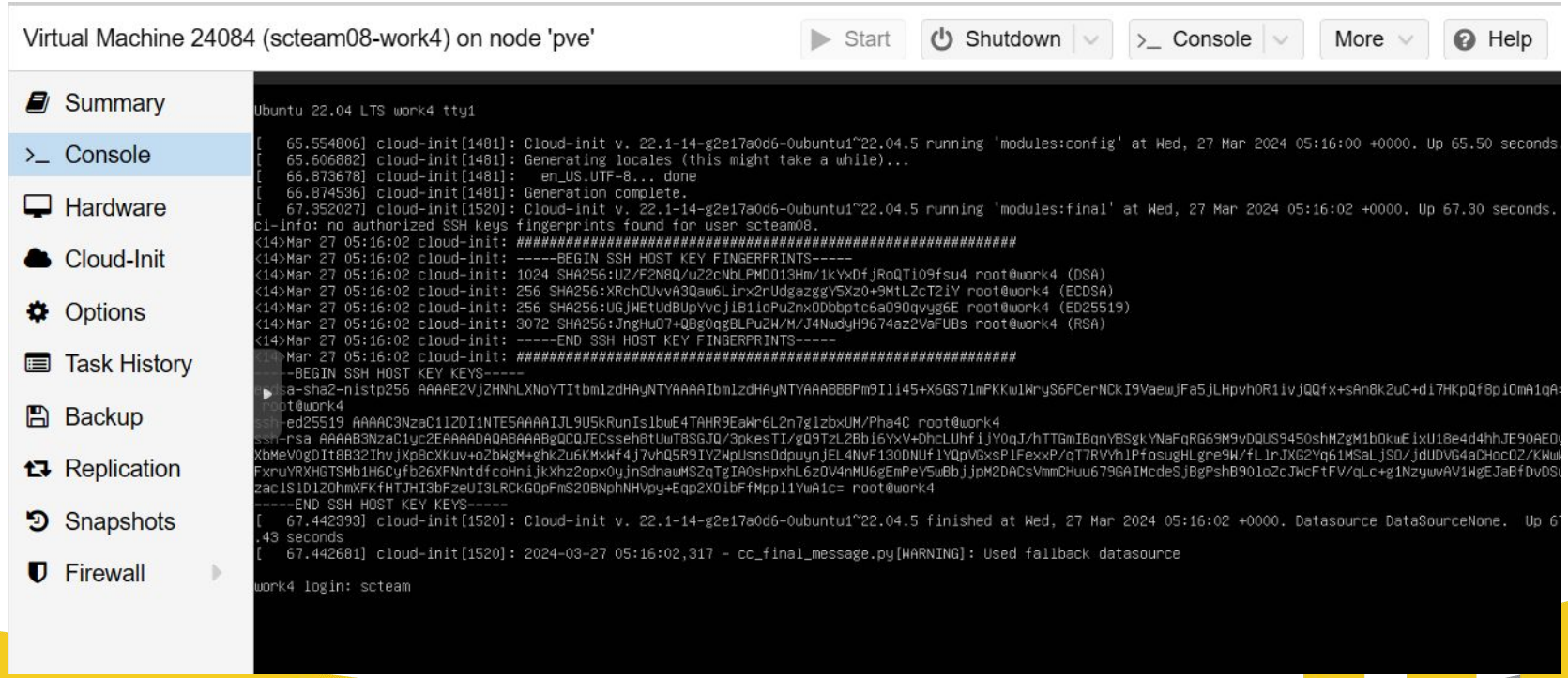
click”reboot”



Click “enter”

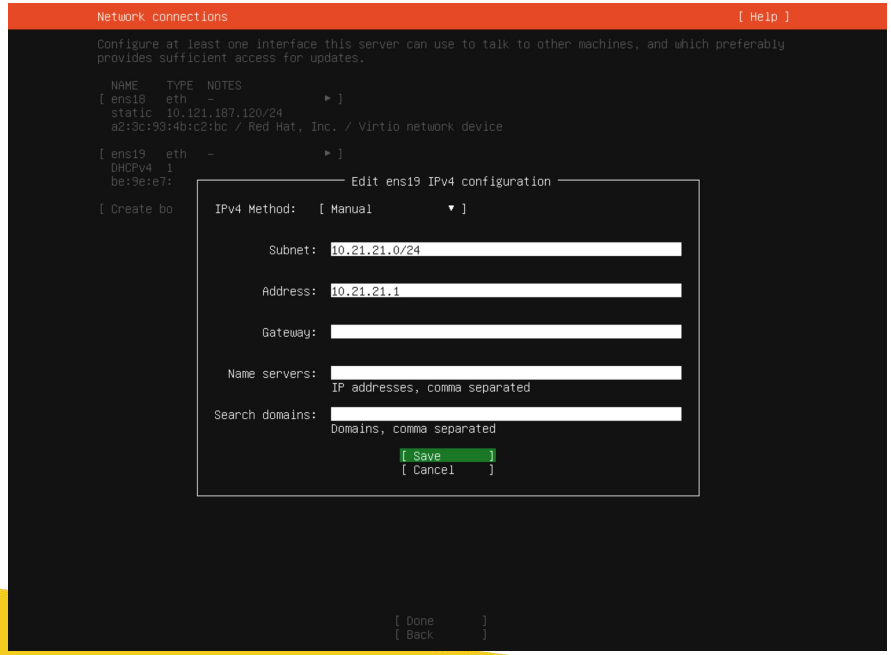

Unmounting /tmp/tmpjpkx_3sx/ubuntu-server-minimal.ubuntu-server.squashfs.dir...
[FAILED] Failed unmounting /cdrom.
[OK] Unmounted /media/filesystem.
[OK] Unmounted /media/minimal.
[OK] Unmounted /rofs.
[OK] Unmounted /run/credentials/systemd-sysusers.service.
[OK] Unmounted Mount unit for core20, revision 1405.
[OK] Unmounted /target/boot/efi.
[OK] Unmounted /tmp/tmpjpkx_3sx/ubuntu-server-minimal.squashfs.dir.
[OK] Unmounted /tmp/tmpjpkx_3sx/ubuntu-server-minimal.ubuntu-server.squashfs.dir.
Unmounting /target/boot...
[OK] Unmounted /media/full.
[OK] Unmounted Mount unit for lxd, revision 22923.
[OK] Unmounted Mount unit for snapd, revision 15534.
[OK] Unmounted Mount unit for subiquity, revision 3359.
[OK] Unmounted /target/boot.
Unmounting /target...
[OK] Unmounted /run/snapd/ns/lxd.mnt.
Unmounting /run/snapd/ns...
[OK] Unmounted /run/snapd/ns.
[OK] Unmounted /tmp/tmpjpkx_3sx/root.dir.
Unmounting /tmp...
[OK] Unmounted /tmp.
[OK] Stopped target Swaps.
[OK] Unmounted /target.
[OK] Stopped target Preparation for Local File Systems.
[OK] Reached target Unmount All Filesystems.
Stopping Monitoring of LVM2 mirrors, snapshots etc. using dmeventd or progress polling...
Stopping Device-Mapper Multipath Device Controller...
[OK] Stopped Create Static Device Nodes in /dev.
[OK] Stopped Create System Users.
[OK] Stopped Device-Mapper Multipath Device Controller.
[OK] Stopped Remount Root and Kernel File Systems.
[OK] Stopped Monitoring of LVM2 mirrors, snapshots etc. using dmeventd or progress polling.
[OK] Reached target System Shutdown.
Starting Shuts down the "live" preinstalled system cleanly...
Please remove the installation medium, then press ENTER:
Unmounting /cdrom...
[FAILED] Failed unmounting /cdrom.

Install successfully!Login!

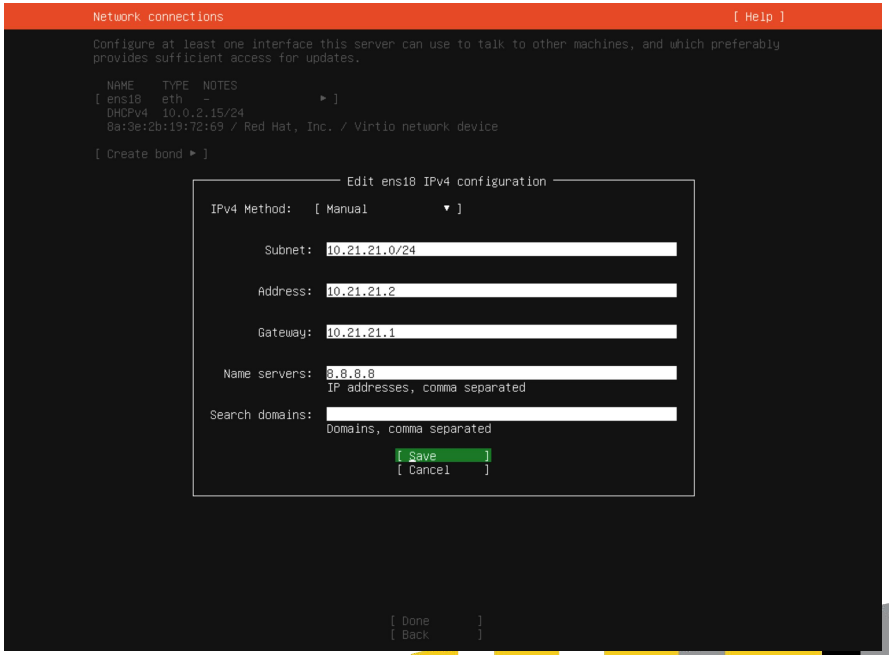


Network (con't)

Head node net1

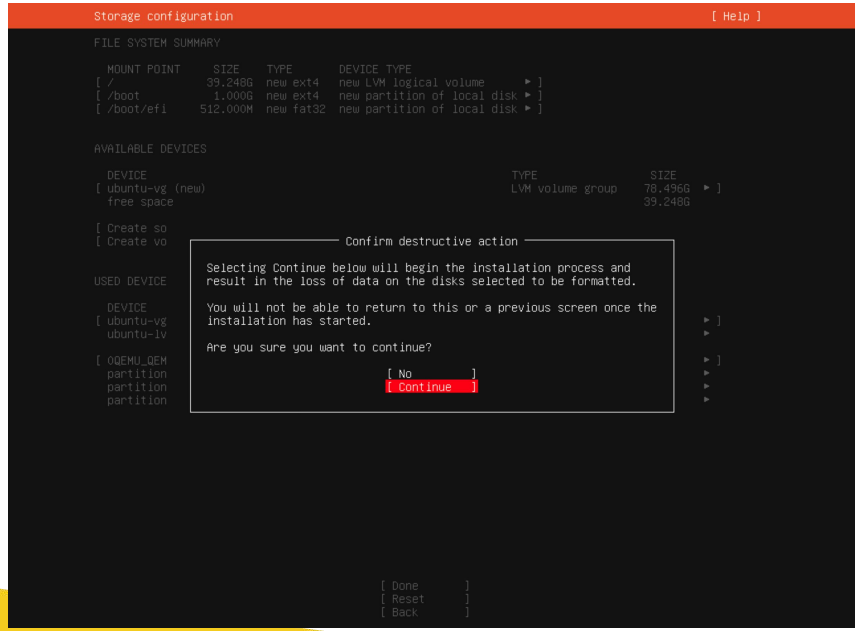


Worker node net0

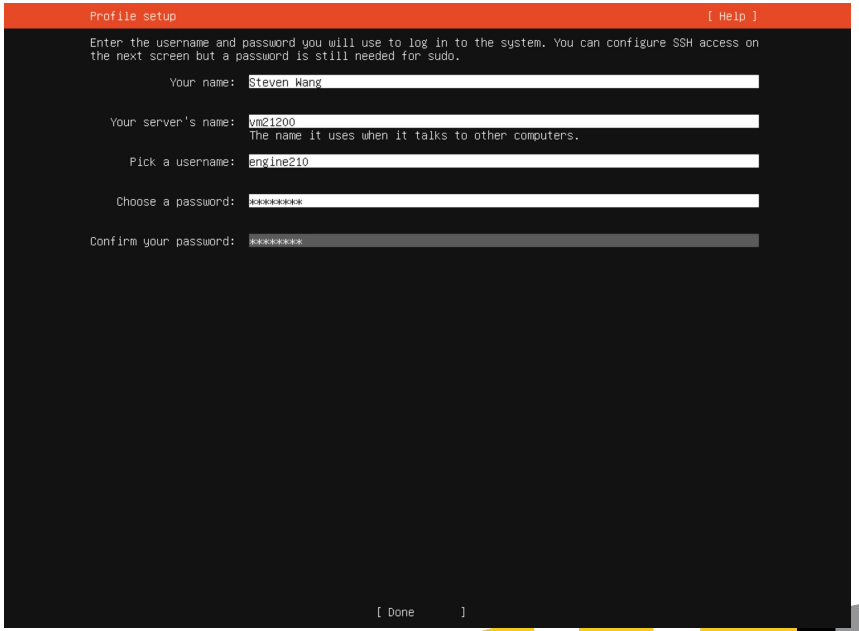


Ubuntu Server (con’t)

Format disk



Set username and password.

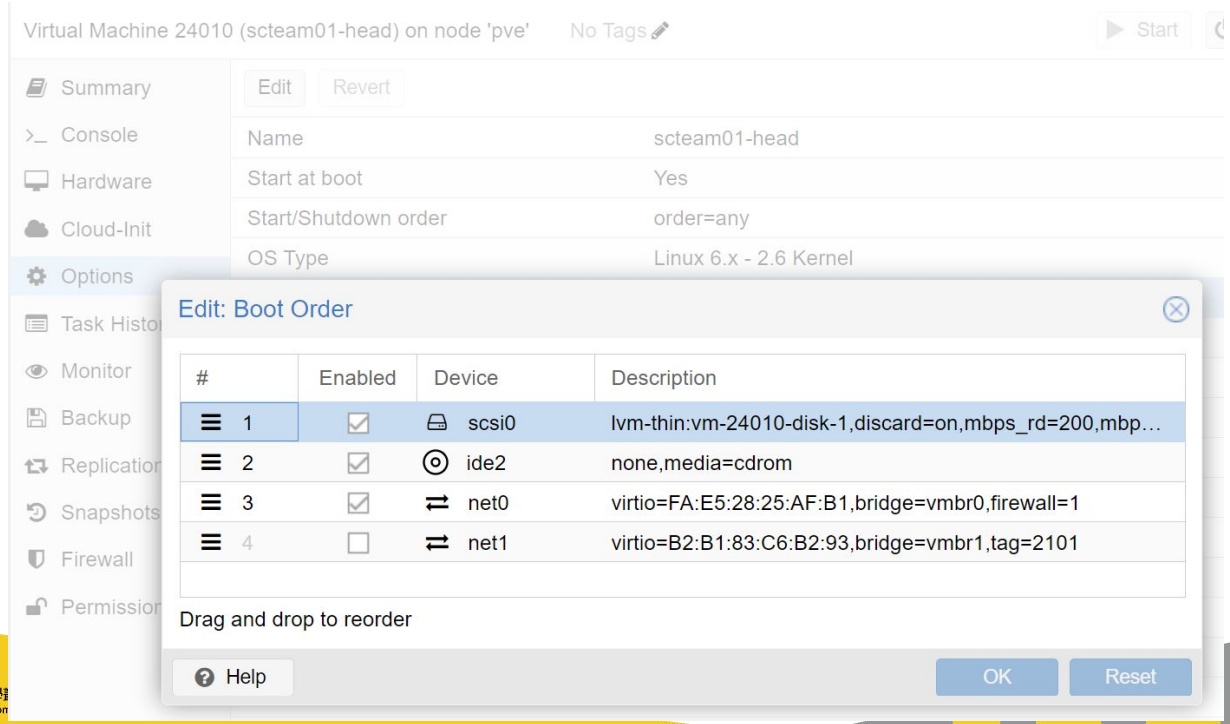


Steps (con’t)

1. Select English
2. Continue without updating
3. Set up network
4. No need to set proxy.
5. Use default mirror site.
6. Storage configuration: Use an entire dick (default). Or you can set your own partition if you want.
7. Check install OpenSSH server.
8. Do not select anything on Feature Server Snap.
9. Remove disk
10. Reboot

Change Boot Order (if needed)

- Options -> Boot Order



Troubleshooting – failed unmounting /cdrom

- It’s fine, just press enter

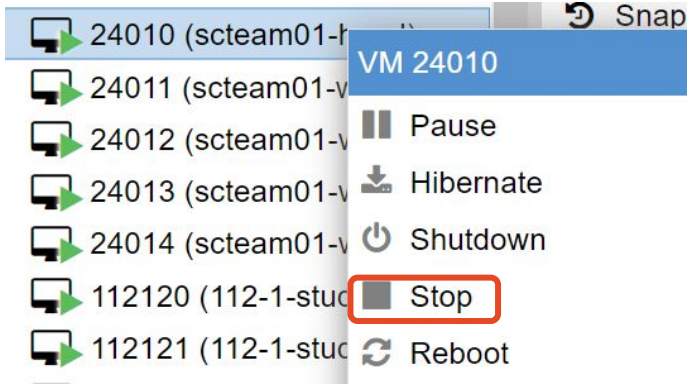
```
         Starting Shuts down the "live" preinstalled system cleanly...
Please remove the installation medium, then press ENTER:
         Unmounting /cdrom...
[FAILED] Failed unmounting /cdrom.
```

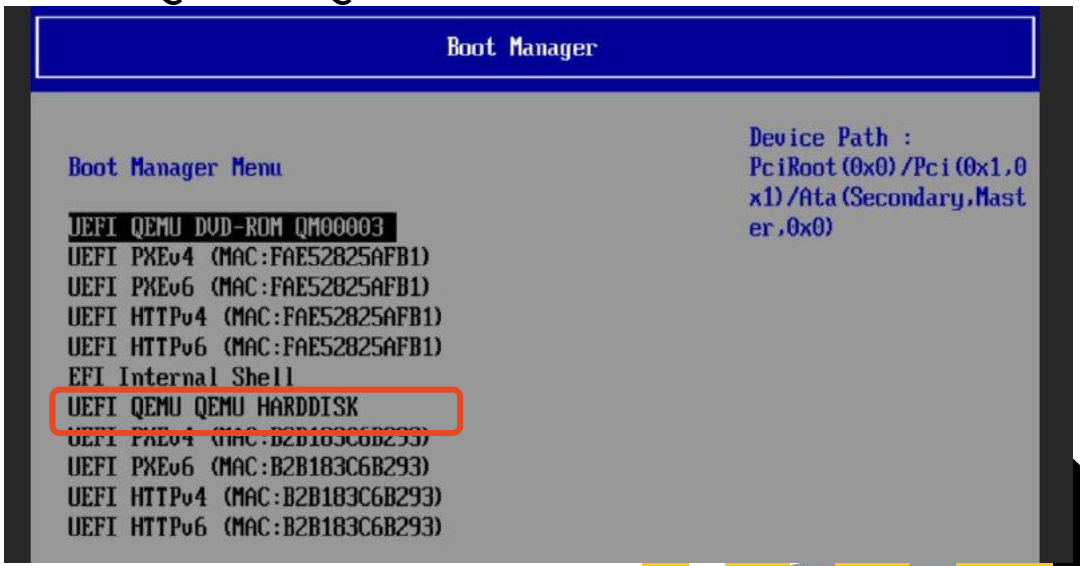
Troubleshooting – UEFI

```
UEFI Interactive Shell v2.2
EDK II
UEFI v2.70 (EFI Development Kit II / OVMF, 0x00010000)
Mapping table
      FS0: Alias(s):HD1a0b::BLK2:
           PciRoot(0x0)/Pci(0x5,0x0)/Scsi(0x0,0x0)/HD(1,GPT,B0EA7463-B8AB-43AD-9EEA-9186C38A7F33,0x800,0x219800)
     BLK0: Alias(s):
           PciRoot(0x0)/Pci(0x1,0x1)/Ata(0x0)
     BLK1: Alias(s):
           PciRoot(0x0)/Pci(0x5,0x0)/Scsi(0x0,0x0)
     BLK3: Alias(s):
           PciRoot(0x0)/Pci(0x5,0x0)/Scsi(0x0,0x0)/HD(2,GPT,FC878FA7-F180-4F7D-8544-ACEEB0DFF914,0x21A000,0x400000)
     BLK4: Alias(s):
           PciRoot(0x0)/Pci(0x5,0x0)/Scsi(0x0,0x0)/HD(3,GPT,F2ECD095-027D-4330-B5FF-A73C25002A17,0x61A000,0x5DE5800)
Press ESC in 1 seconds to skip startup.nsh or any other key to continue.
Shell> _
```

Troubleshooting – UEFI (con't)

- Restart your VM (stop -> start)
- When booting, keep pressing ESC (like 狂按) to enter BIOS
- Select Boot Manager -> UEFI QEMU QEMU HARDDISK





Other Settings

Steps

1. IP Forward (head)

IP Forward-Head node

- Worker node access internet through head node.
- Head node need to setup IP forward rule.
- We use iptables here (or you can choose anything you familiar with)
- Sudo -i

```
#!/bin/sh
EIF="ens18" # WAN interface (use "ip a" to check)
IIF="ens19" # LAN interface (use "ip a" to check)
INNET="10.21.21.0/24" # LAN subnet
# forwarding
echo "net.ipv4.conf.all.forwarding = 1" >> /etc/sysctl.conf
sysctl -p

# NAT
iptables -t nat -A POSTROUTING -o $EIF -s $INNET -j MASQUERADE
apt-get install iptables-persistent
iptables-save > /etc/iptables/rules.v4
systemctl enable iptables.service
```

Steps

1. IP Forward (head)
2. NFS
   a. Head Node
   b. Work Node

Linux File Hierarchy

- `/`: Root directory
- `/bin`: Essential binary executables. e.g. cat, ls, cp.
- `/dev`: Device files. e.g. `/dev/sda`
- **`/etc`: Configuration file, most application configuration can be found here.**
- `/home`: Home directory, user data.
- `/lib`: Libraries for /bin and /sbin.
- `/mnt`: Temporarily mounted filesystem. We usually mount USB here.
- `/opt` : Optional application software packages.

NFS (Network File System)

File sharing across nodes

- normal files
- environment config files (.bashrc, .vimrc)
- libraries & executables (MPI)

NFS Installation

- Install NFS packages
  - Ubuntu: https://ubuntu.com/server/docs/service-nfs
  - CentOS: https://www.opencli.com/linux/rhel-centos-7-install-nfs-server
  - Arch Linux: https://wiki.archlinux.org/index.php/NFS
- Share /home
- Configure NFS server (edit `/etc/exports`) then restart nfs server.
  - `/home 10.21.21.0/24(rw,async,no_root_squash)`
  - `/opt 10.21.21.0/24(rw,async,no_root_squash,no_subtree_check)`
- Client mount NFS (edit /etc/fstab) then reboot.
  - `10.21.21.1:/home  /home  nfs4 soft,intr,bg  0 0`
  - `10.21.21.1:/opt  /opt  nfs4 soft,intr,bg  0 0`
- Change subnet IP according to your cluster's setting!
- Do not simply copy-and-paste!

NFS Installation (Head)

- Install some package (help check whether successfully implement NFS)
  - wget https://download.open-mpi.org/release/open-mpi/v5.0/openmpi-5.0.1.tar.gz
- Install NFS packages
  - Ubuntu: `sudo apt install nfs-kernel-server`
- Create /opt and set the permission
  - `sudo mkdir /opt`
  - `sudo chmod -R 777 /opt`
- Configure NFS server (sudo vim /etc/exports)
  - `/home 10.21.21.0/24(rw,async,no_root_squash)`
  - `/opt 10.21.21.0/24(rw,async,no_root_squash,no_subtree_check)`
- Then restart nfs server.
  - sudo systemctl enable nfs-server
  - sudo systemctl restart nfs-server
- Do not simply copy-and-paste!

NFS Installation (Worker)

- Install NFS packages
  - Worker: `sudo apt install nfs-common`
  - CentOS: https://www.opencli.com/linux/rhel-centos-7-install-nfs-server
  - Arch Linux: https://wiki.archlinux.org/index.php/NFS
- Client mount NFS (edit /etc/fstab)
  - `10.21.21.1:/home  /home  nfs4 soft,intr,bg,timeo=600  0 0`
  - `10.21.21.1:/opt  /opt  nfs4 soft,intr,bg,timeo=600  0 0`
- Mount the filesystem(NFS)
  - `sudo mount -a`
  - `cd`
- Do not simply copy-and-paste!

if you don’t want to type IP when ssh

- sudo vim /etc/hosts

```
# /etc/hosts
# ....
10.21.21.1 head
10.21.21.2 work1
10.21.21.3 work2
10.21.21.4 work3
1021.21.5 work4
```

Steps

1. IP Forward (head)
2. NFS
   a. Head Node
   b. Work Node
3. SSH KEY

SSH (Laptop to head node)

After you finish installation and configure network, you can use SSH to connect to your VM’s terminal.

- Linux and Max:
  ssh username@10.121.187.11x (x is your team number)
- Windows: MobaXterm / VScode / …

Useful tips:

- ssh-keygen -m PEM -t rsa
- ssh-copy-id <username>@<head_node_ip>
  - In Windows:
    ssh public key is in C:/Users/user/.ssh/id_rsa.pub

Now, we can connect to head node without password

SSH (head node -> worker node, say work1)

On headnode:

- Generate key
  - ssh-keygen -m PEM -t rsa
  - ssh-keygen -t ed25519
- Copy public key in other machine
  - ssh-copy-id <username>@10.21.21.2 # work 1's ip
  - cat ~/.ssh/id_ed25519.pub > ~/.ssh/authorized_keys(after NFS)

Now, we can connect to worker node without password
(It's important step for OpenMPI.)

User Authentication

- User account information sharing
- Objective:

```
[root@node1 ~] useradd -m alice
[alice@node1 ~] ssh alice@node1
[alice@node1 ~] pwd
/home/alice
[alice@node1 ~]
```

Solution

- NIS
- LDAP
- Arch Linux: minos

- Or... Just create user for each node in the same sequence

Install software on Linux

1. Package Manager provide by Linux distribution (APT) -> binary-install
2. Install-script provided by software development company
3. Package Manager for HPC (Spack) -> source-code-install
4. Compile from source

安裝方便:1 > 3 = 2 > 4

管理方便:3 = 1 > 2 > 4

自由度高:4 > 3 > 2 > 1

Week 5 homework

https://hackmd.io/@9jNKJjQJQ-GI9NPpBTJO3A/HkUjfvpAa