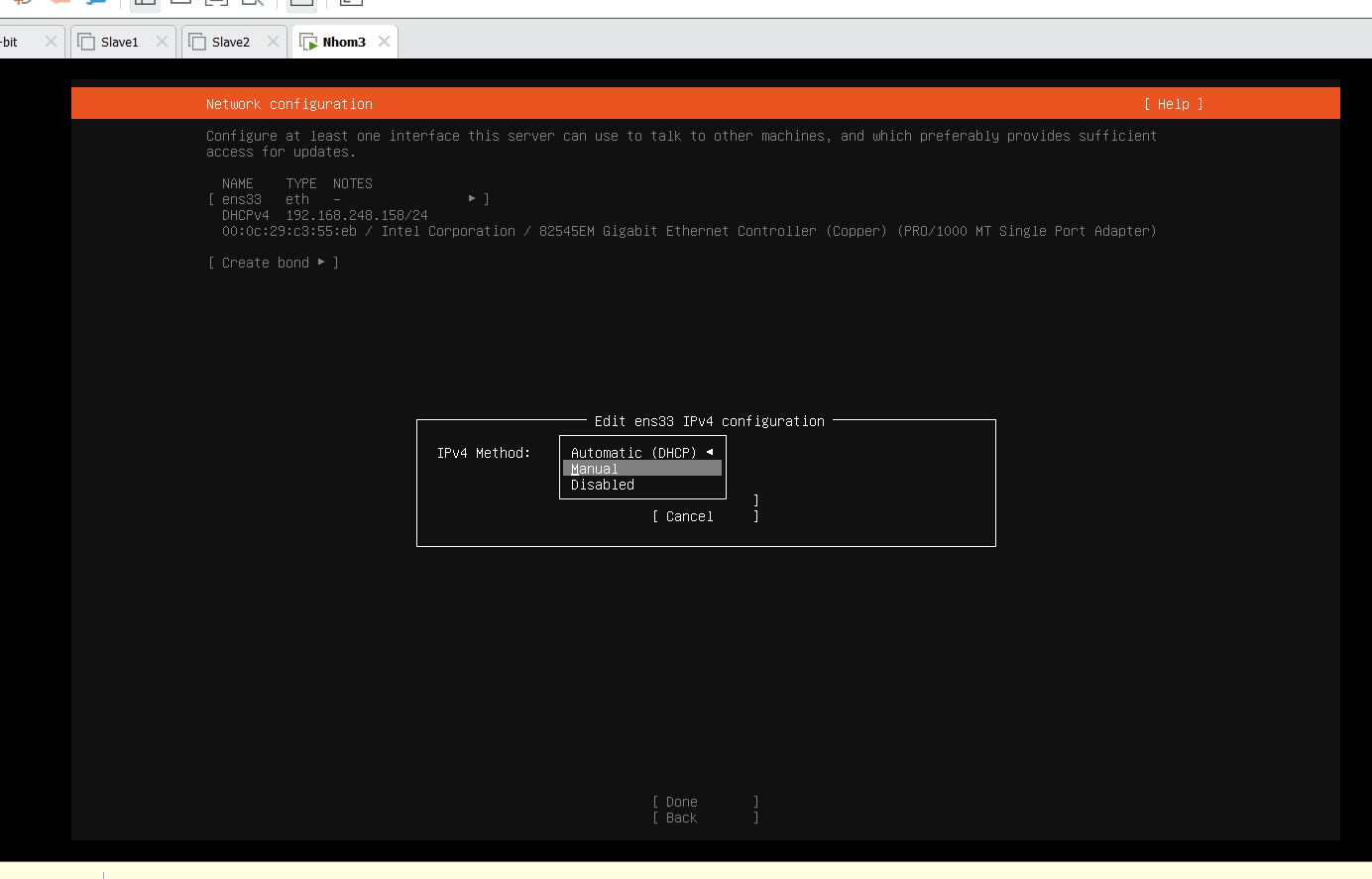
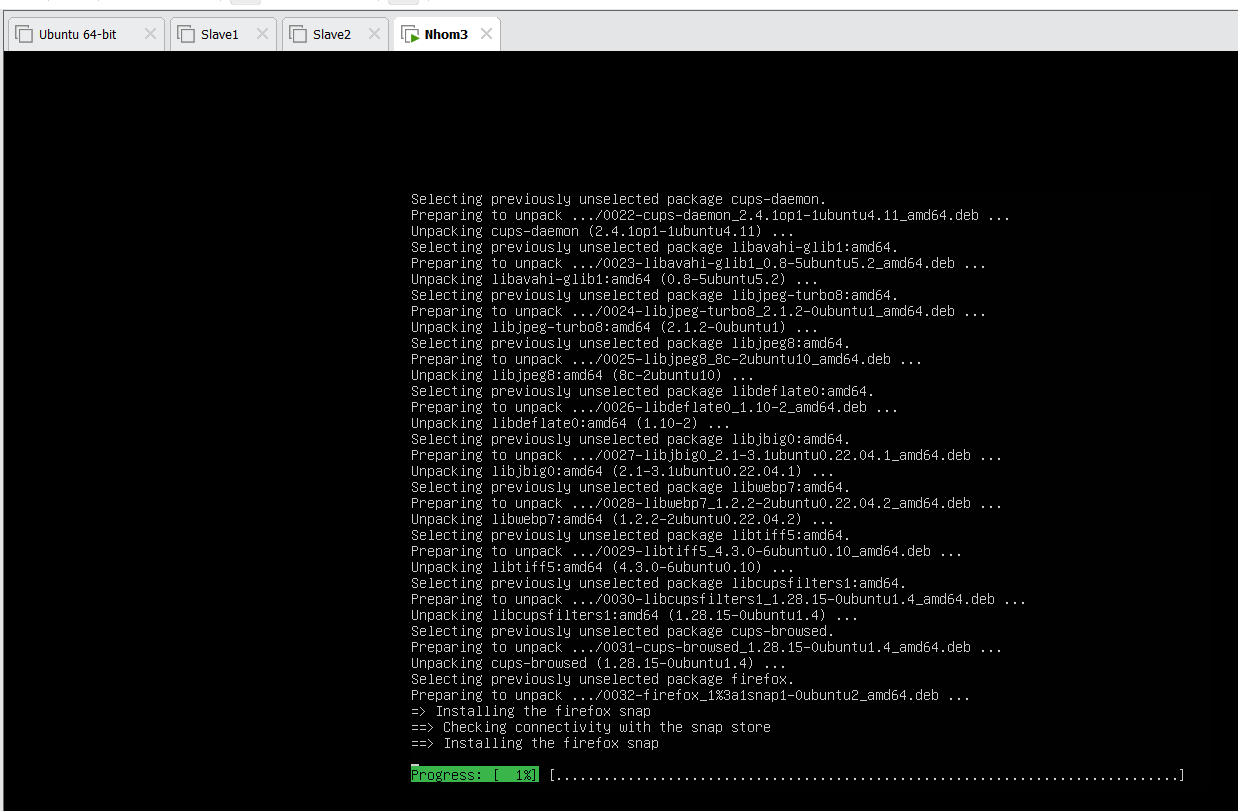
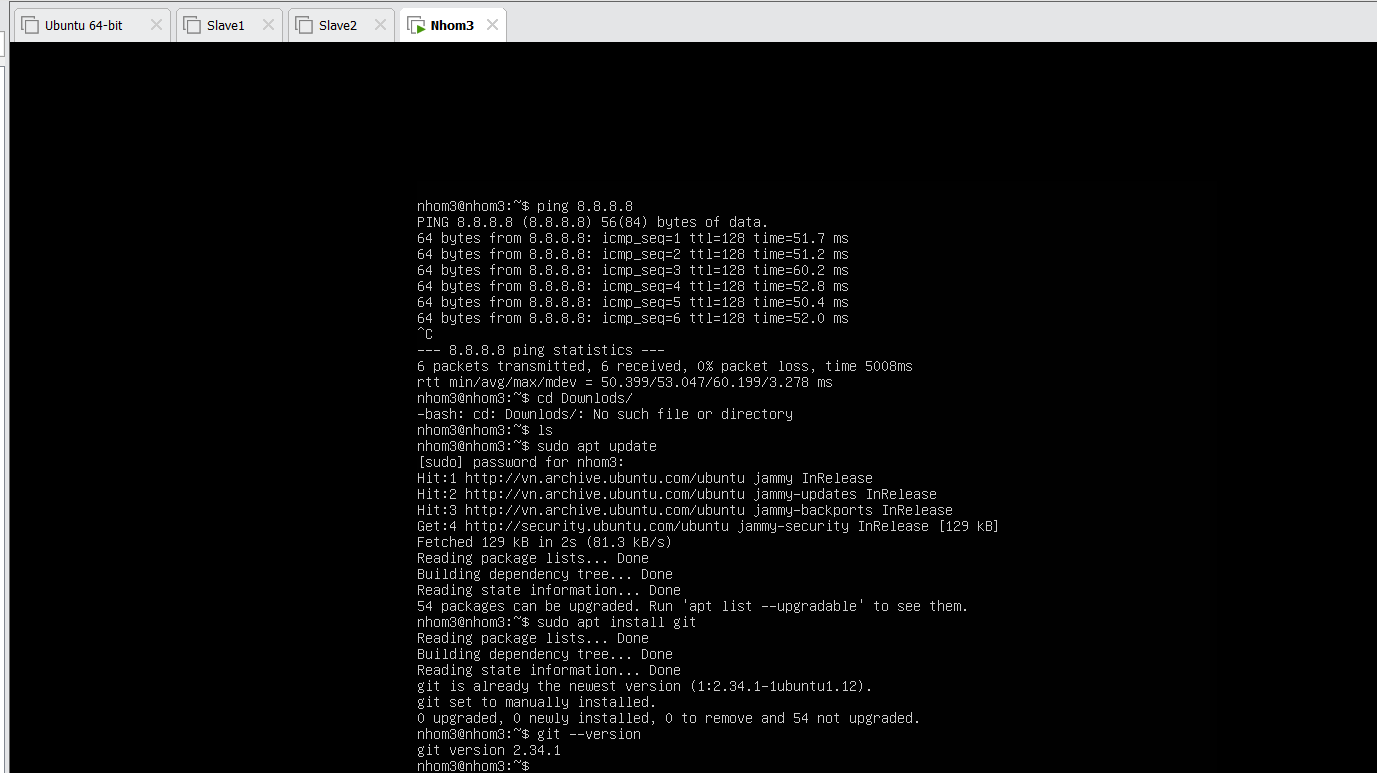
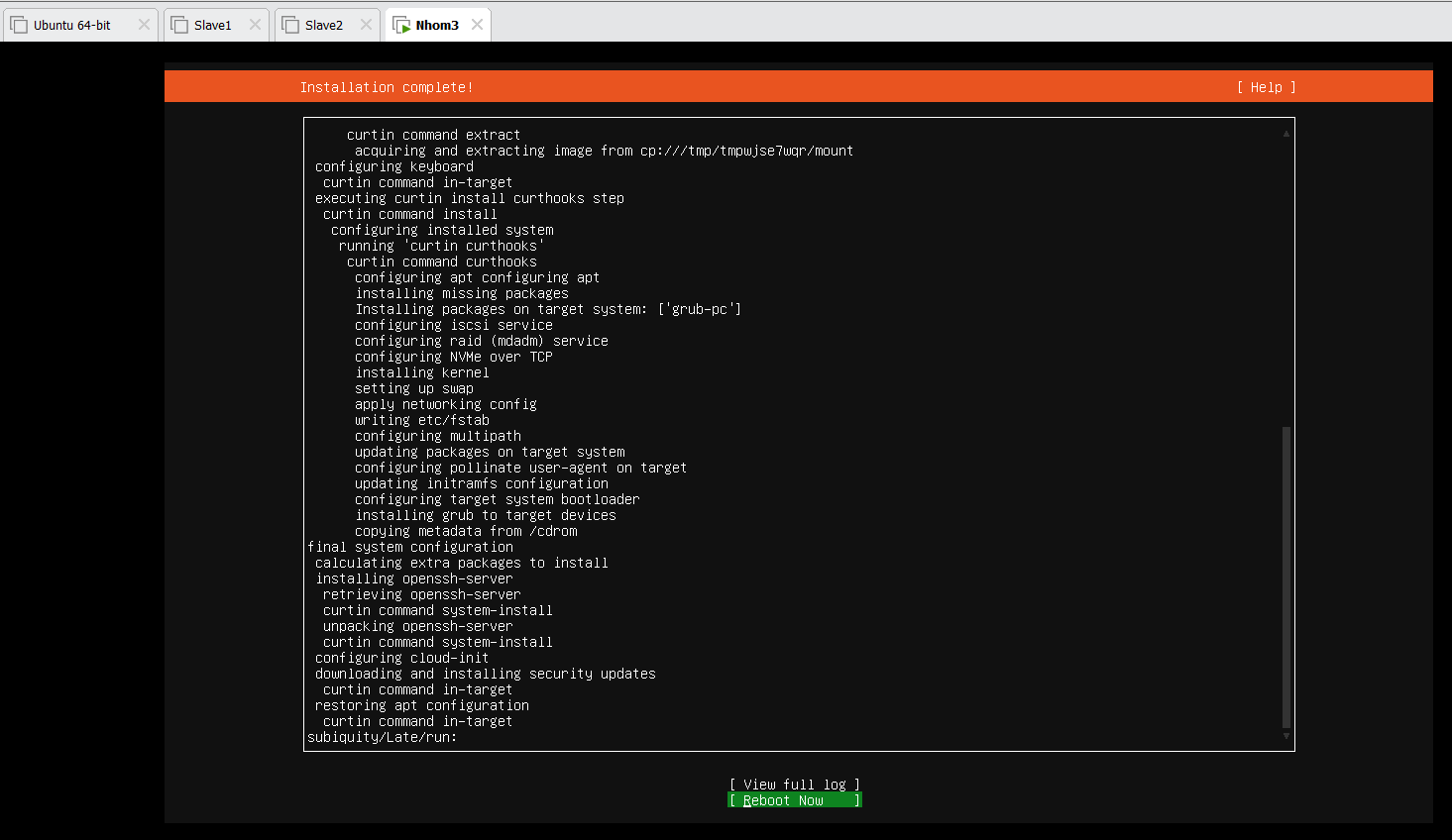
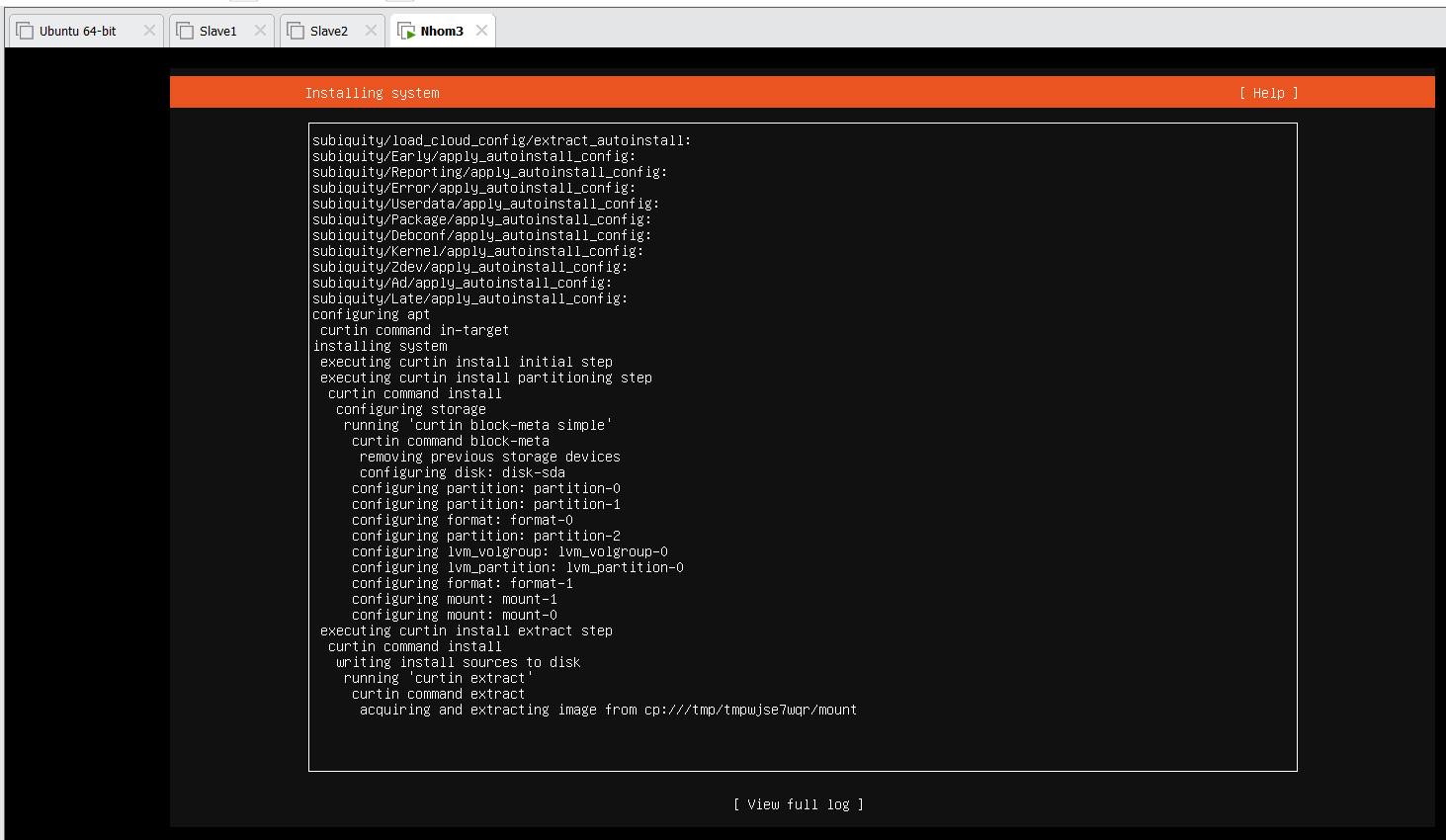
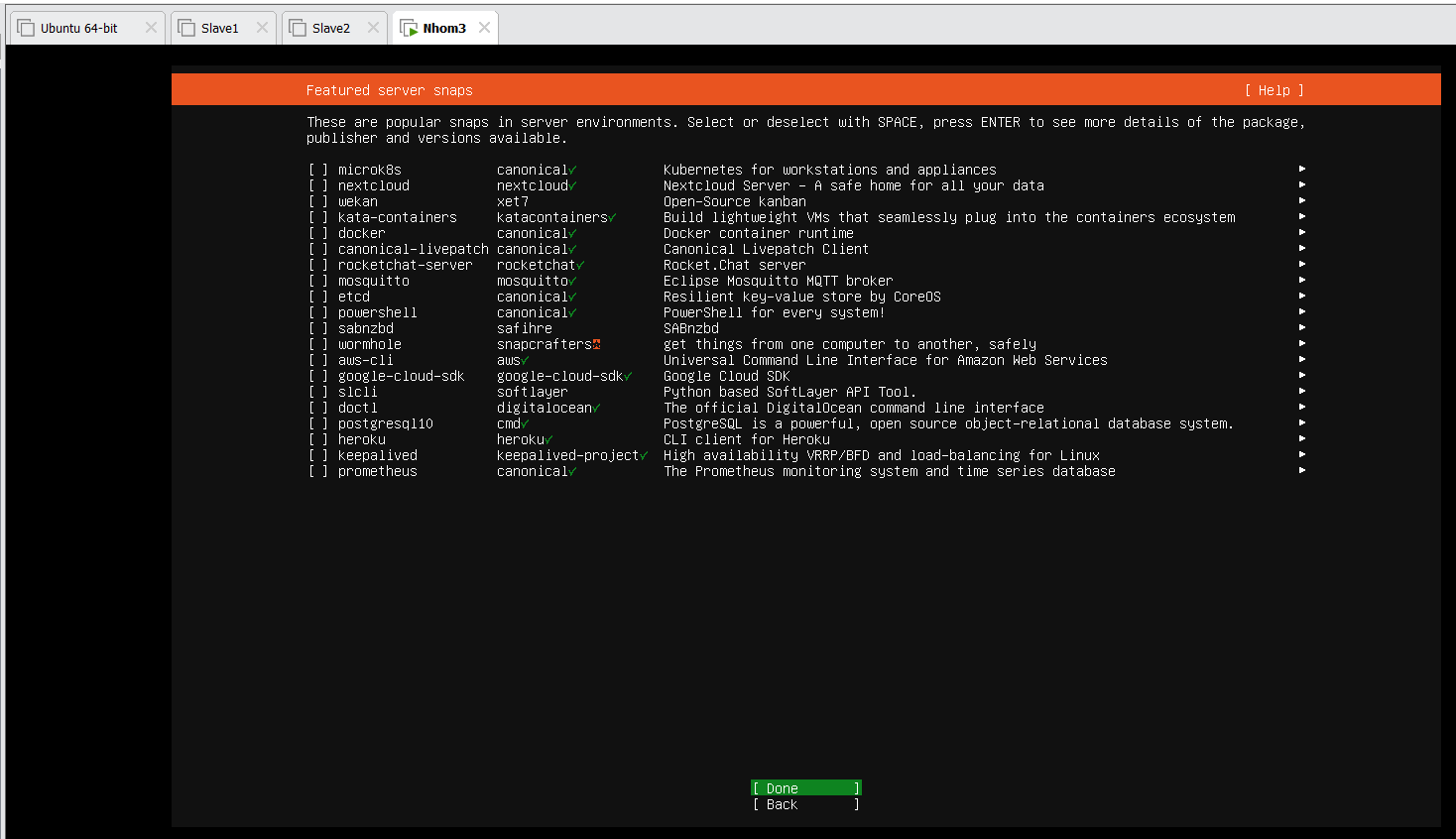
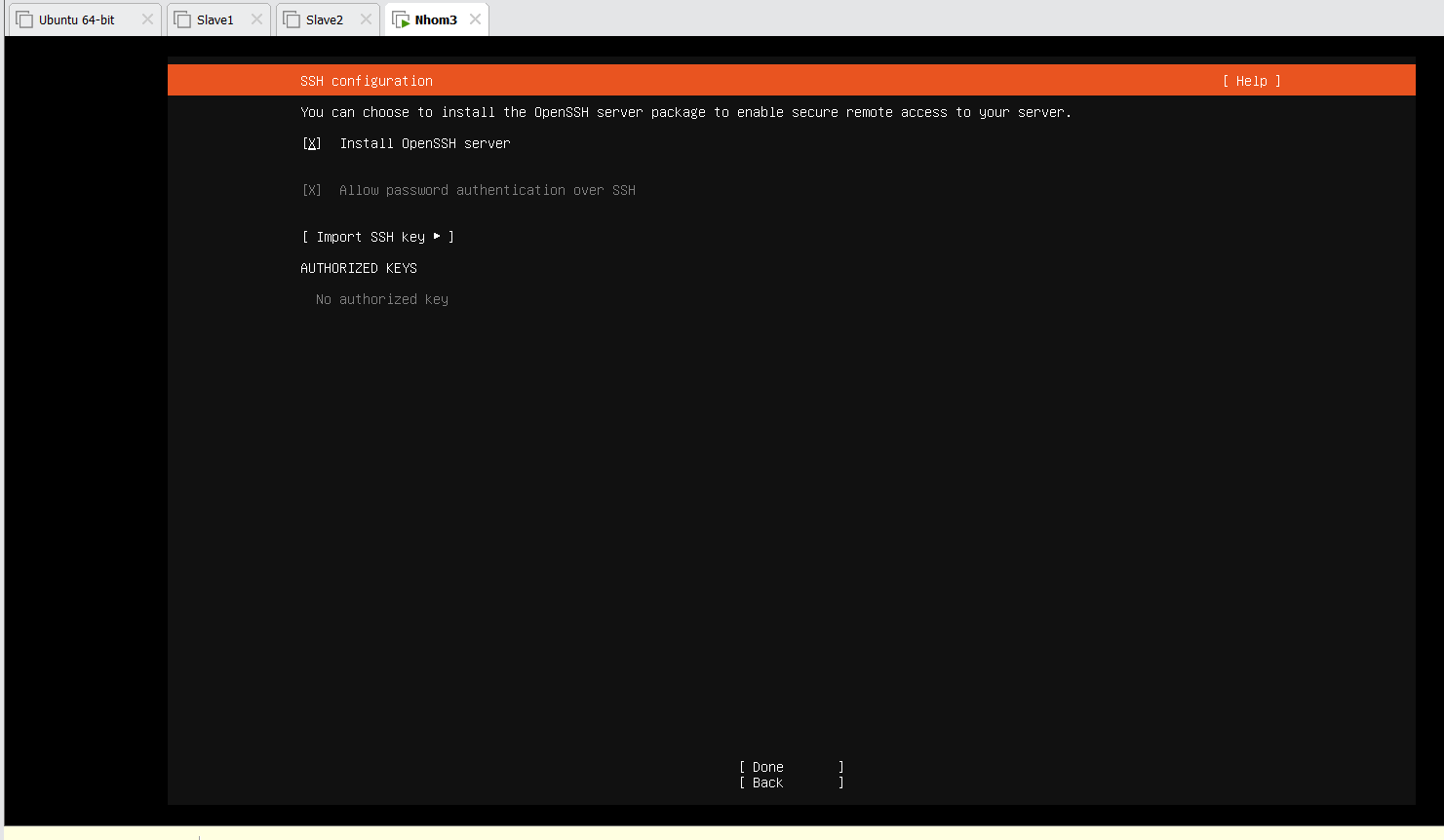
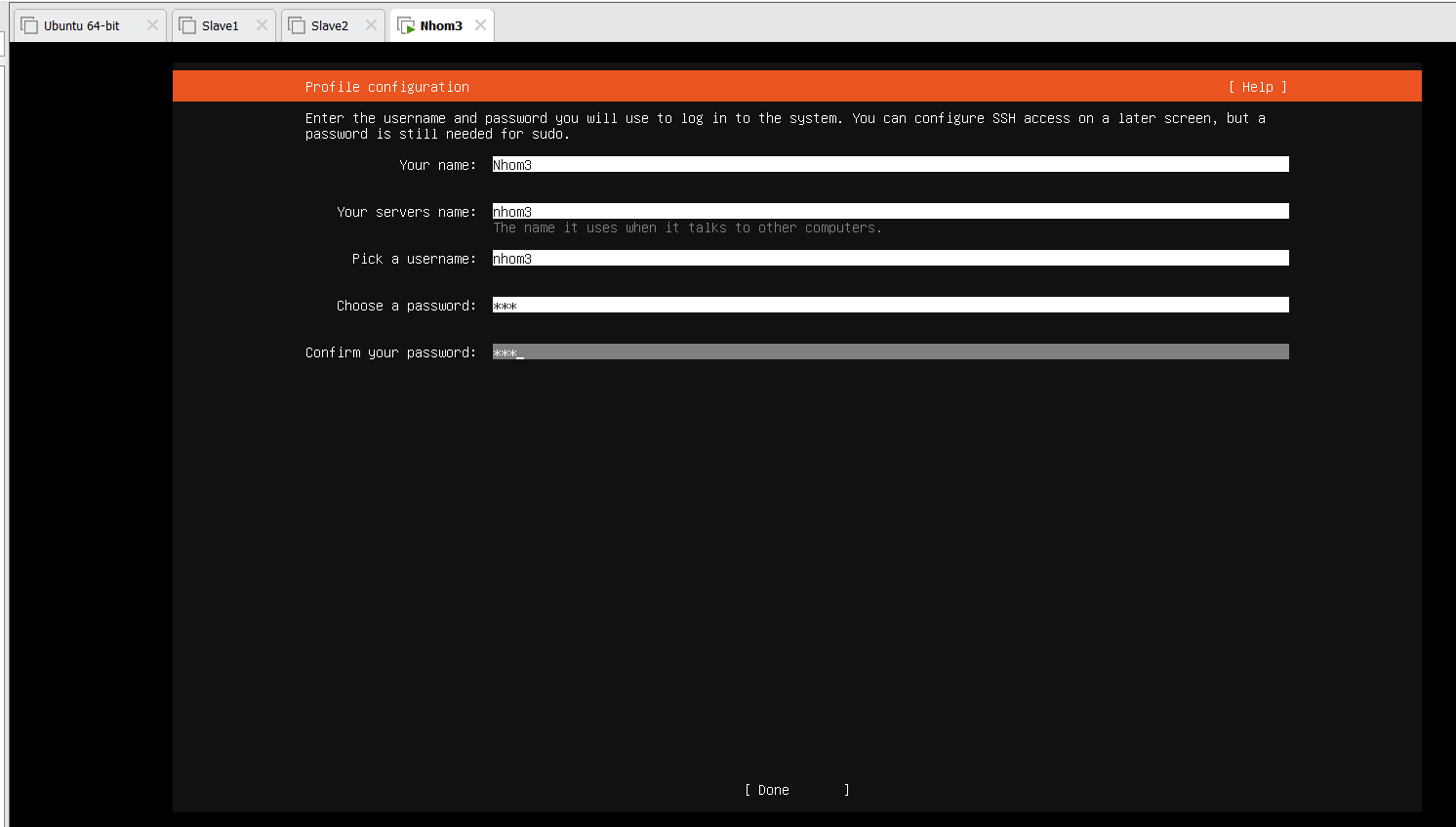
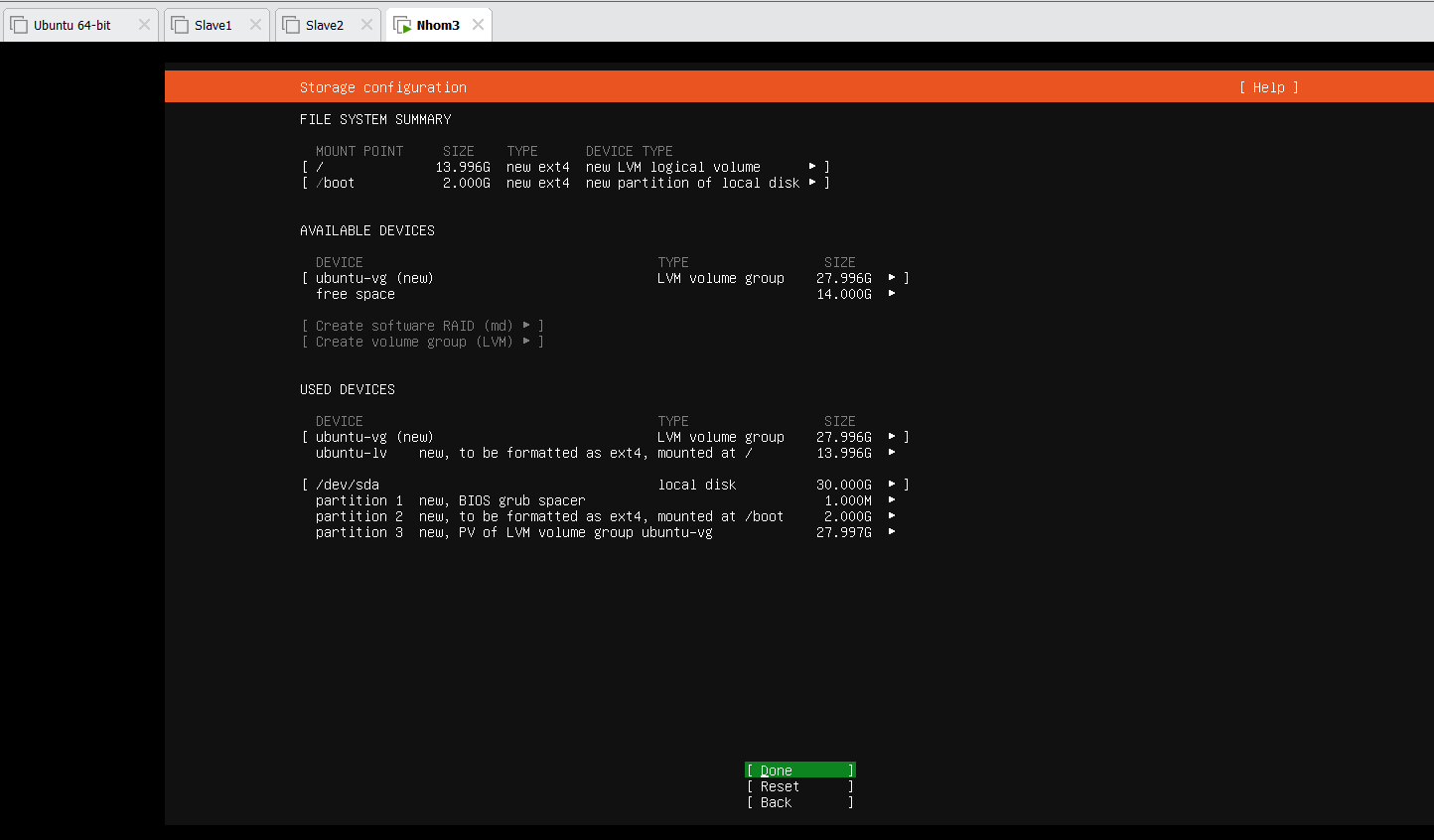
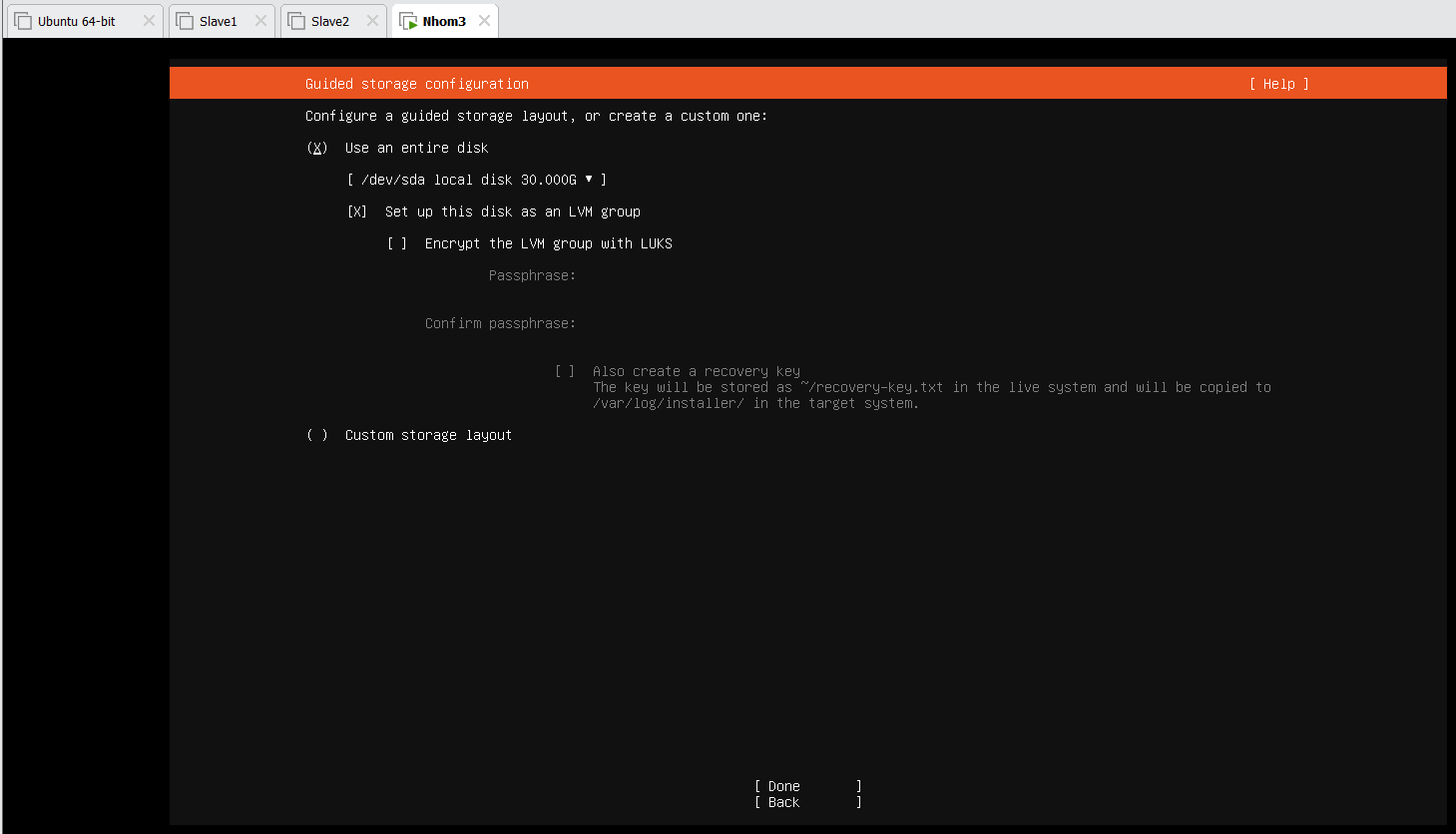
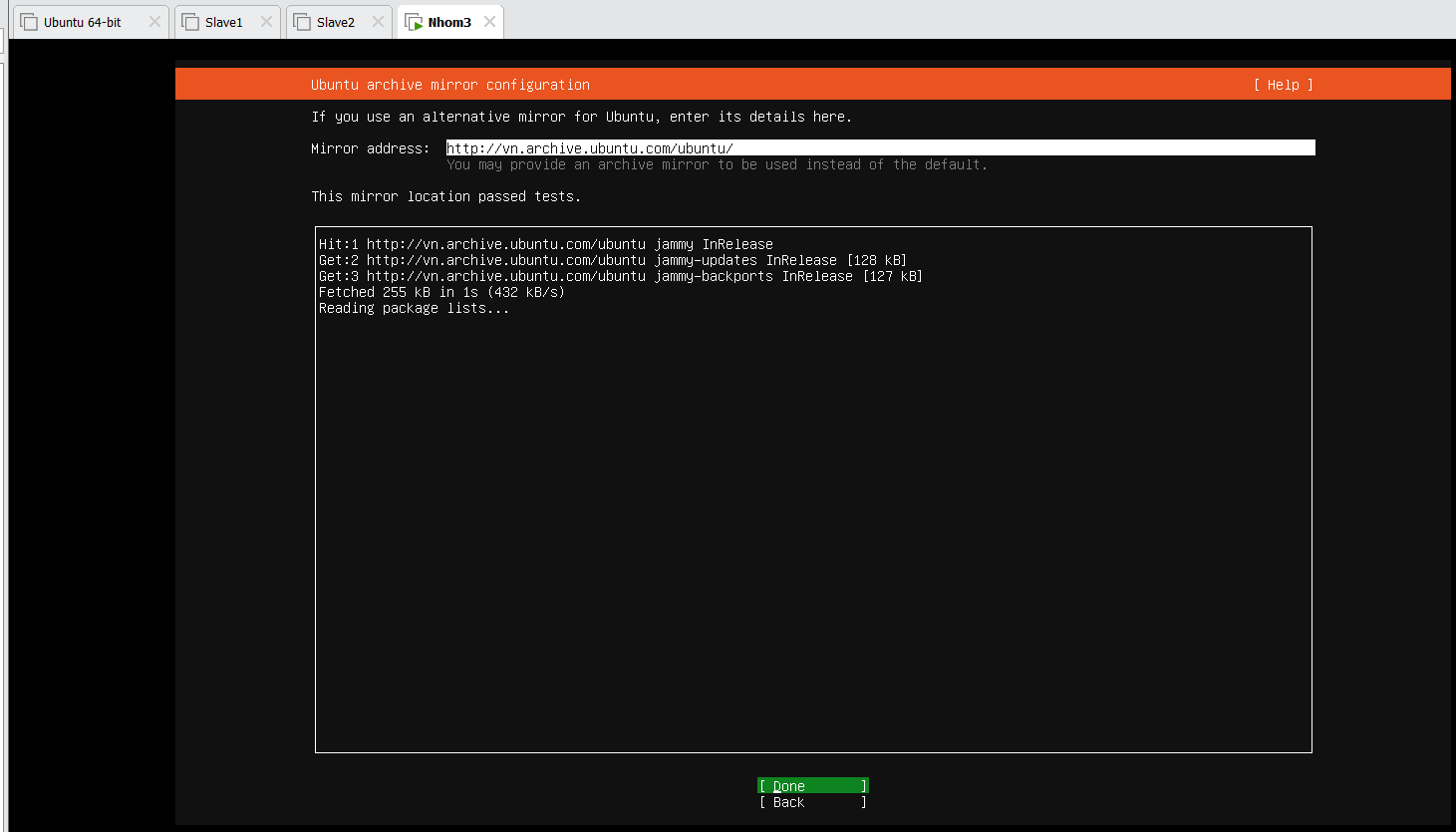
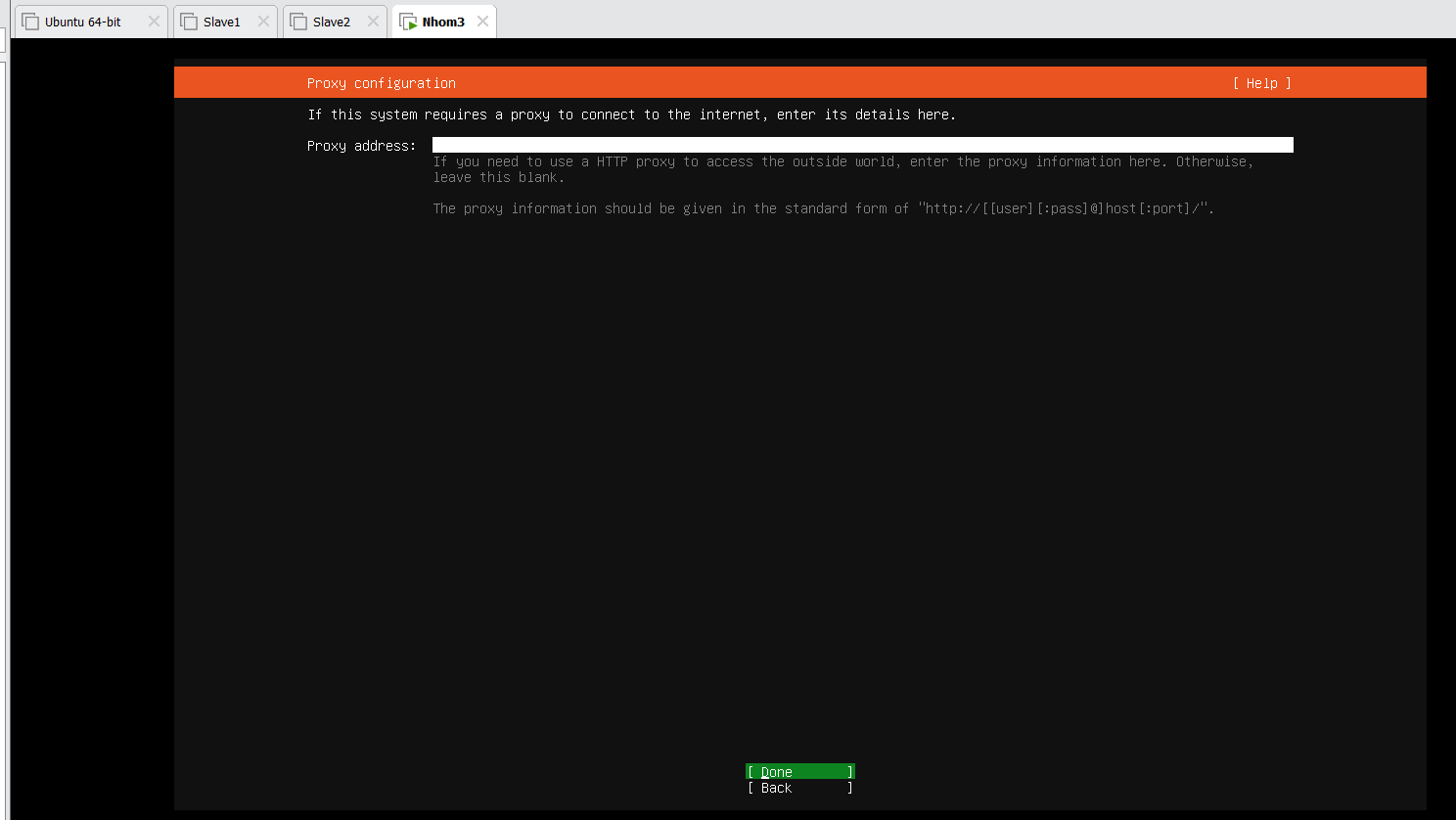
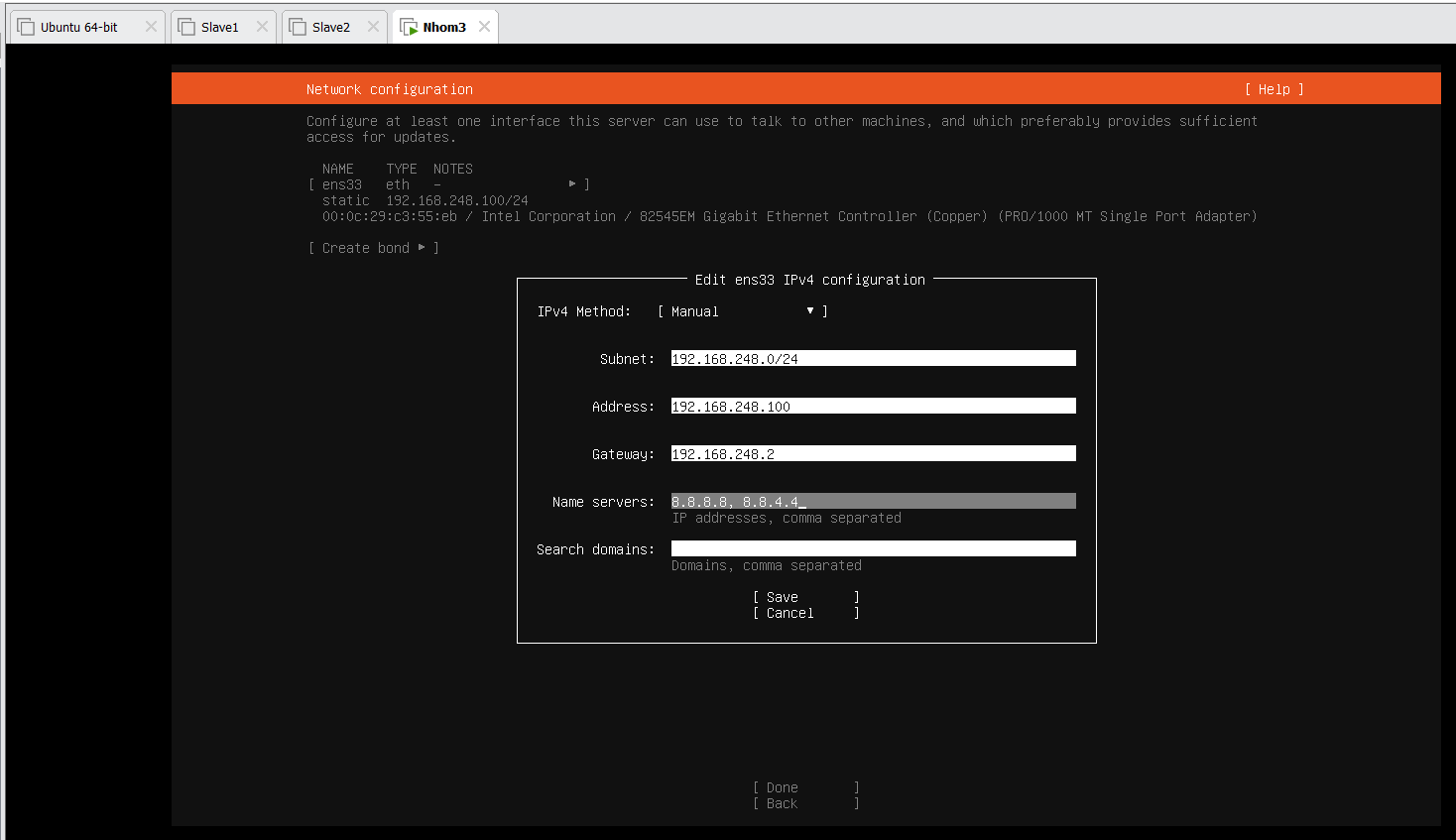
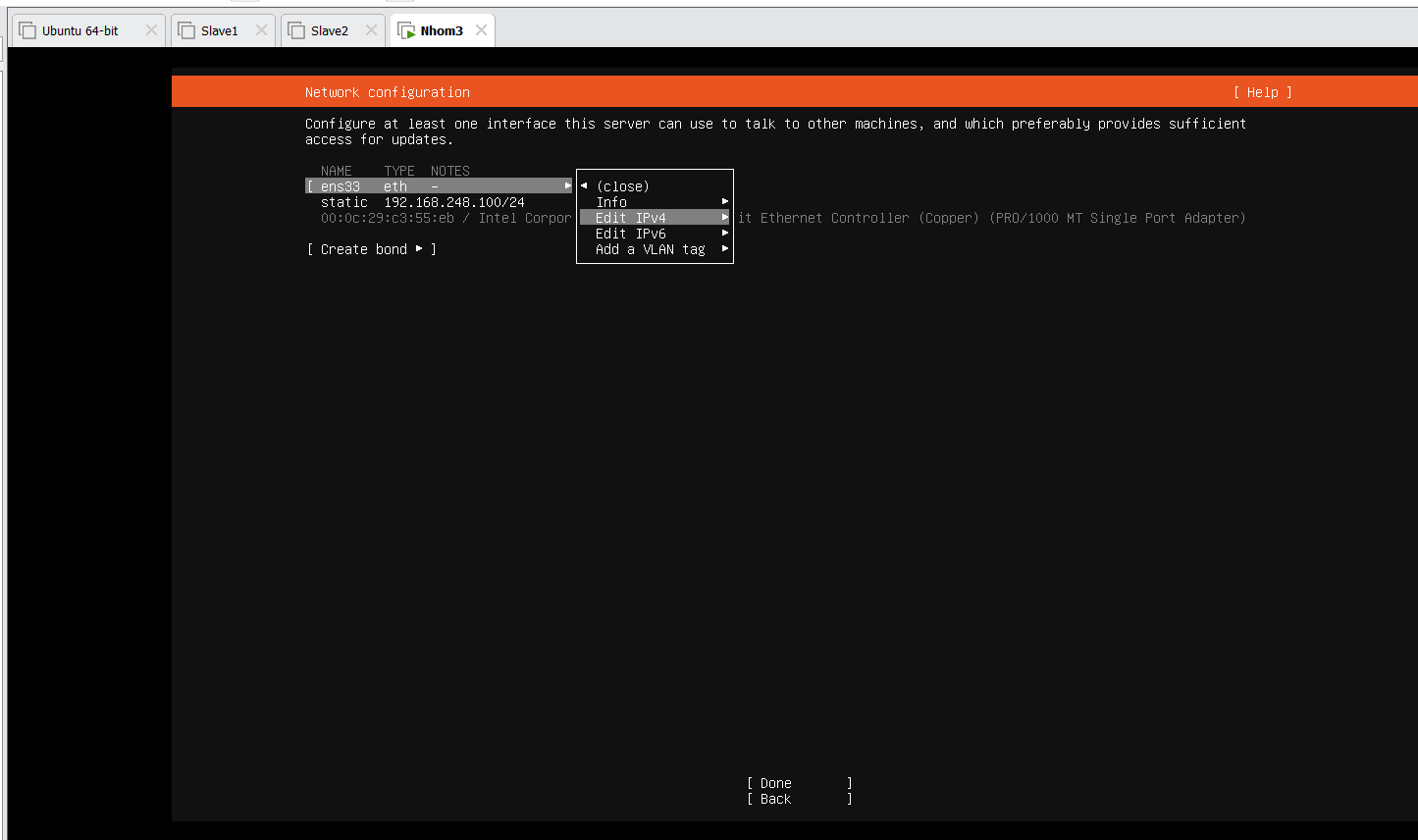
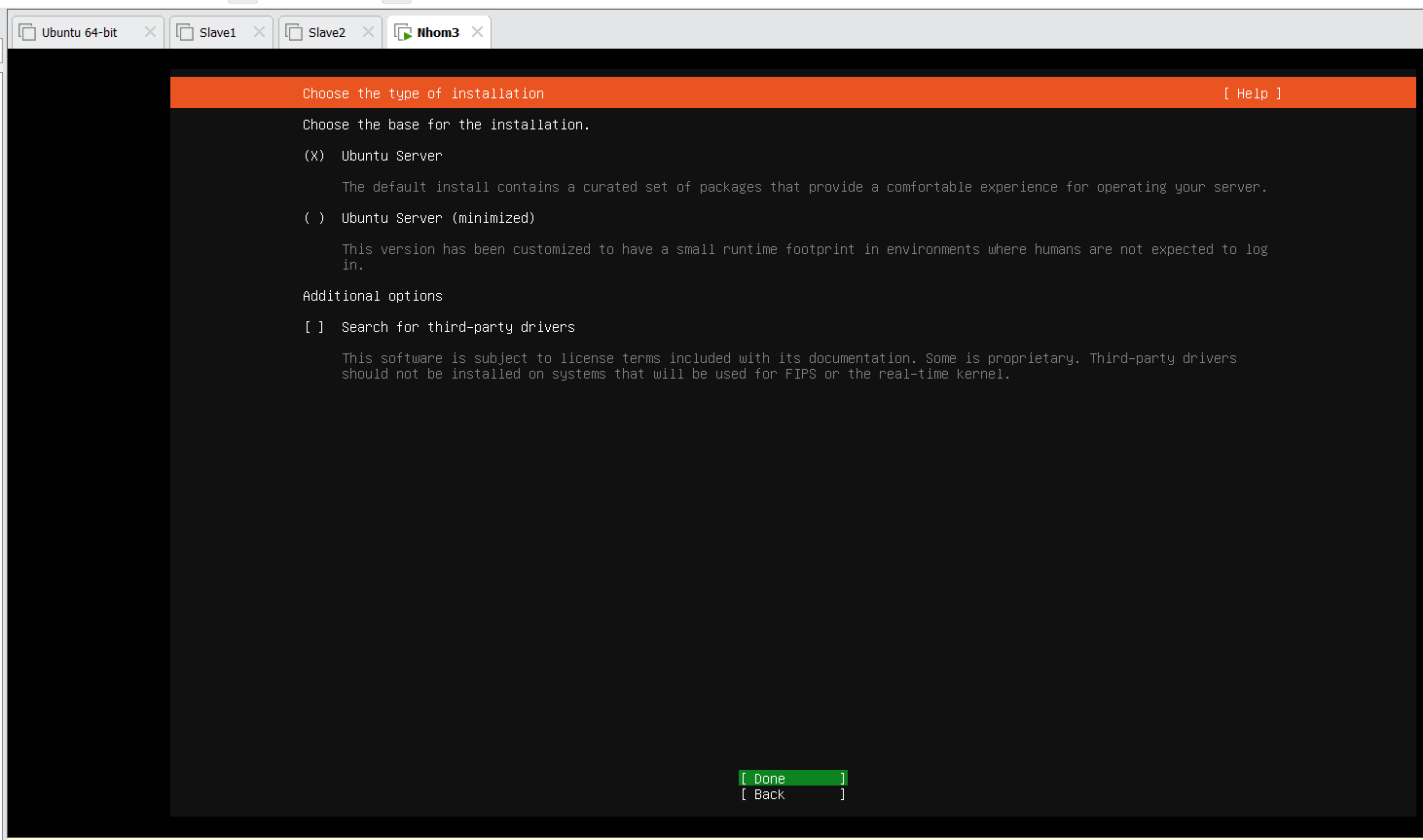
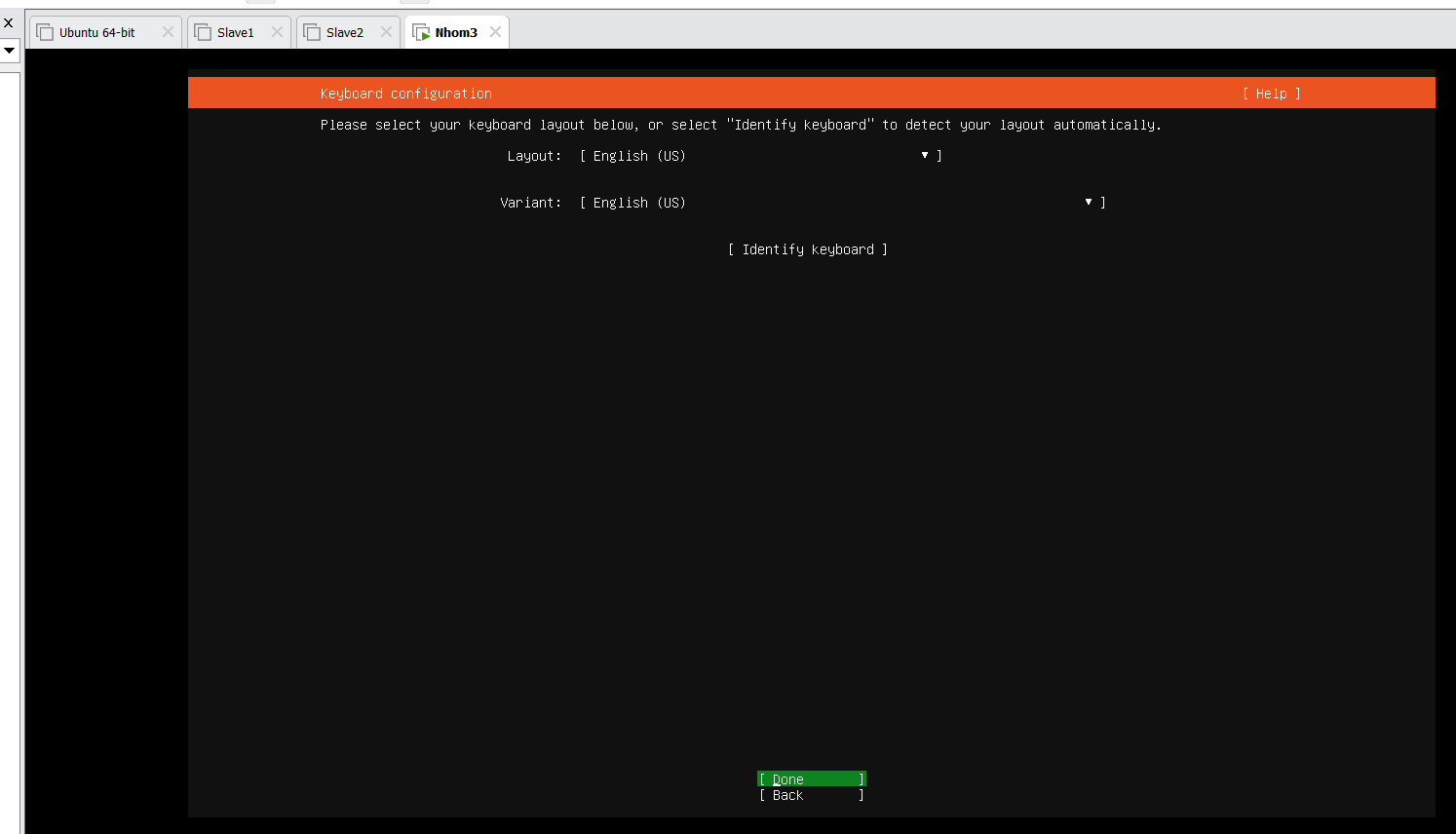
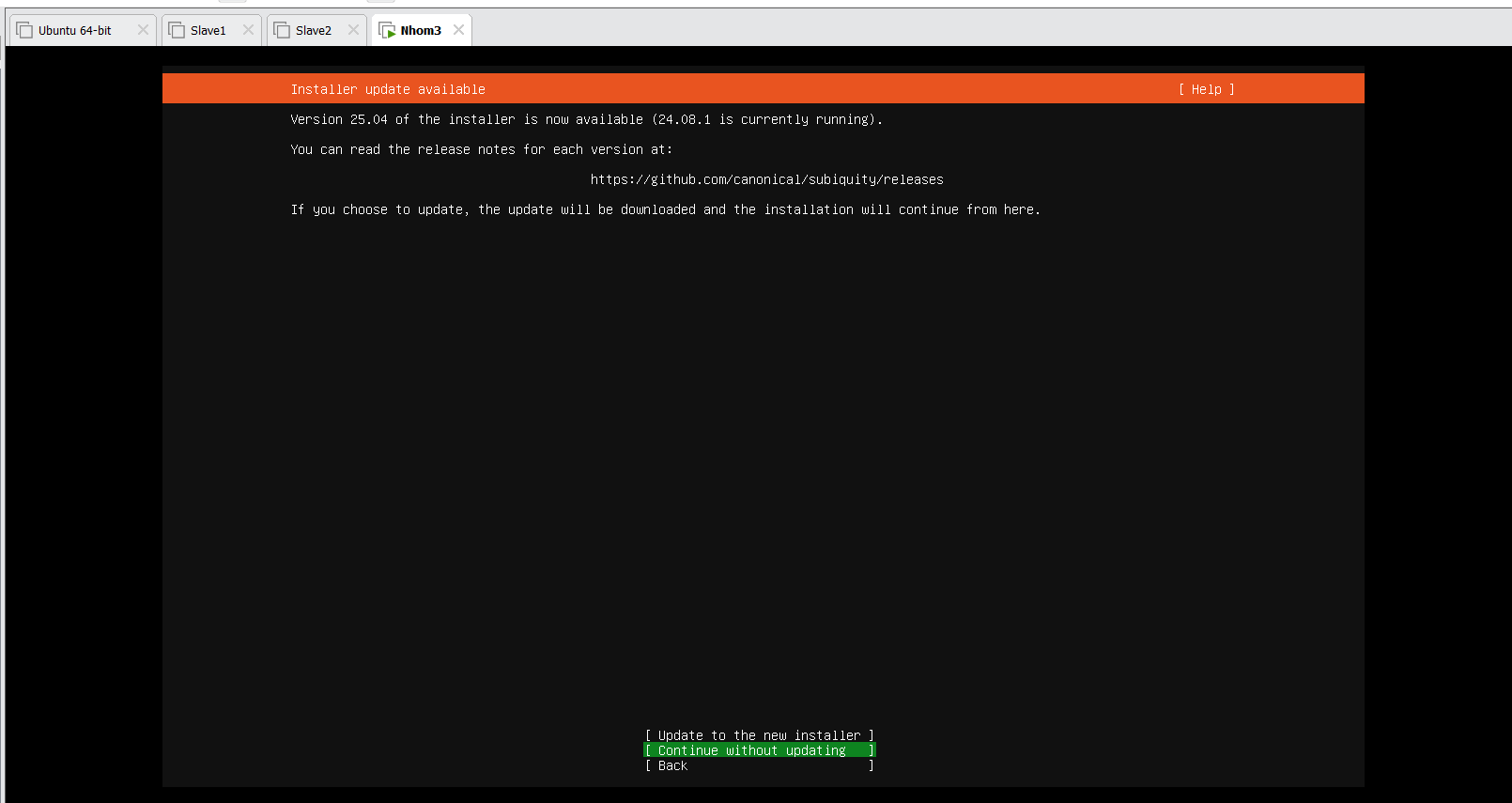
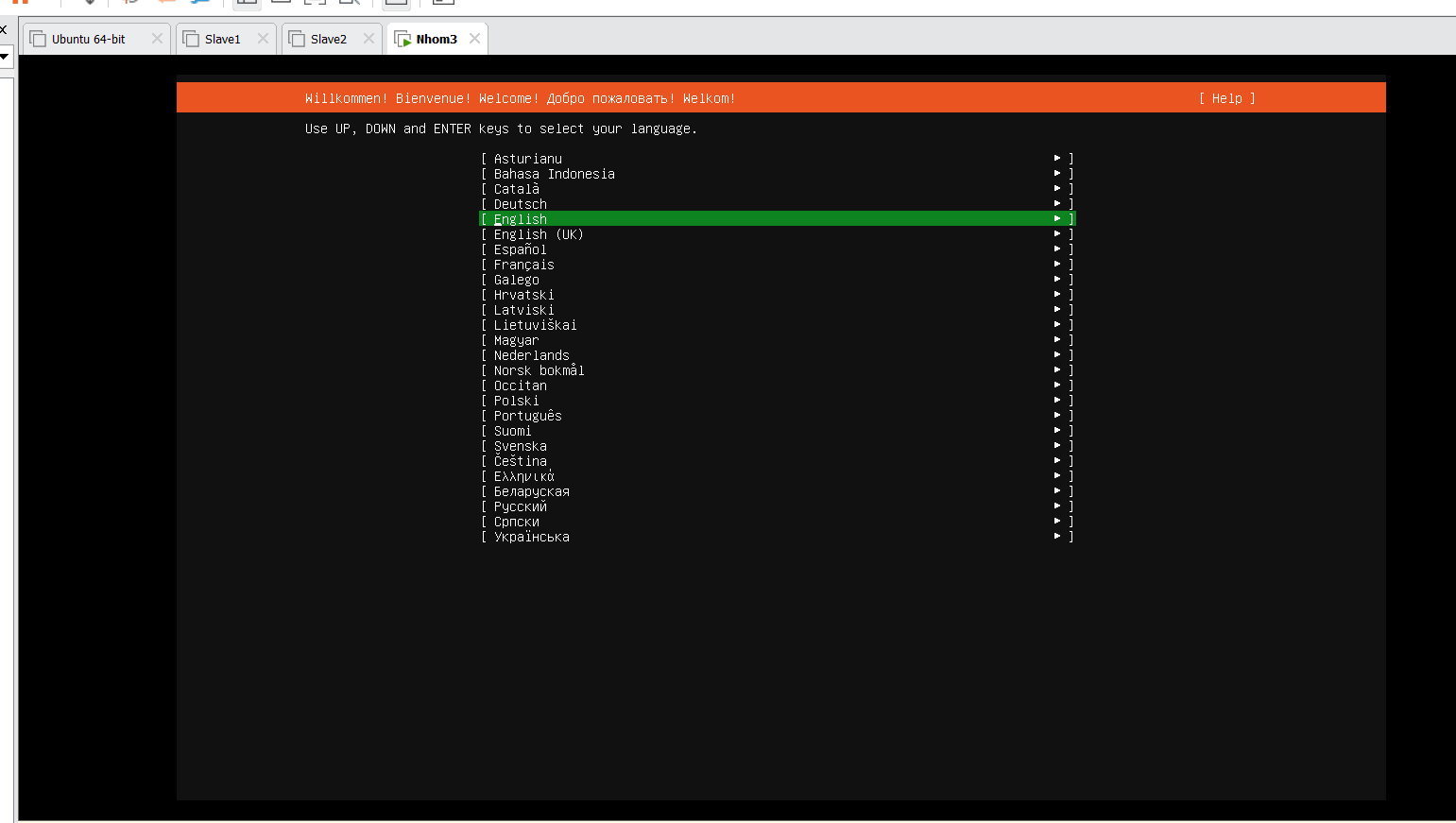
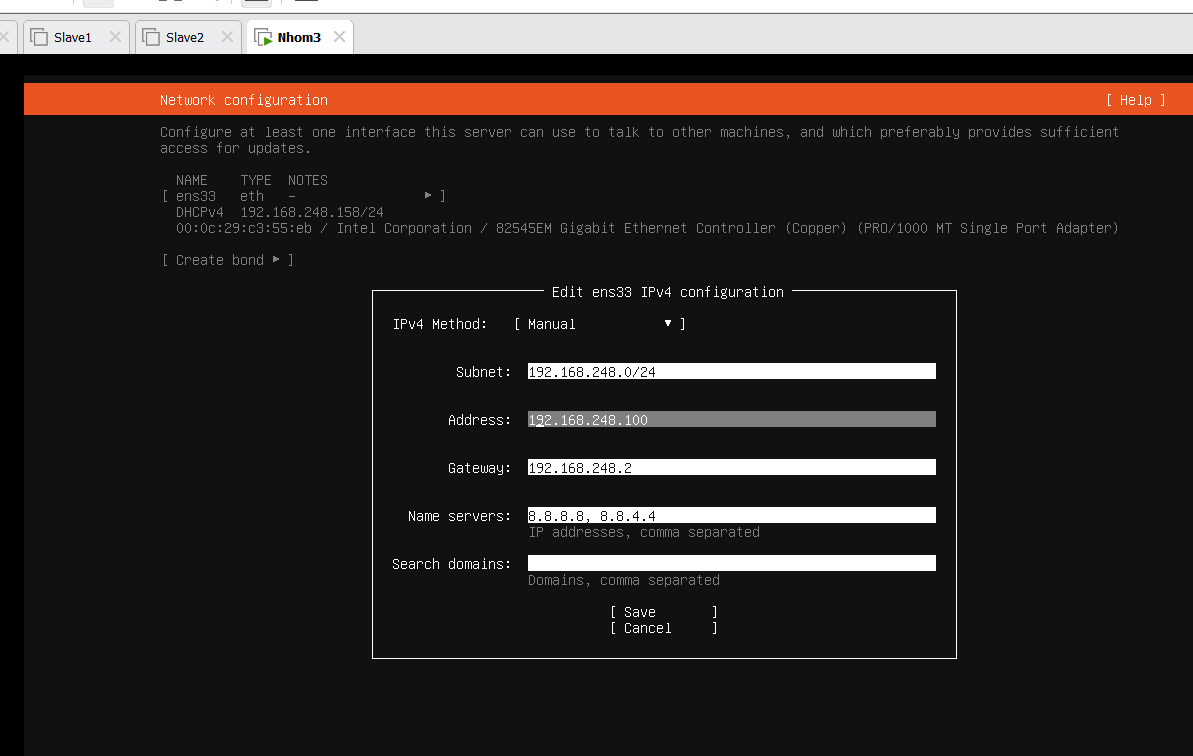
**Chương 1:CÀI ĐẶT UBUNTU SERVER**

****

****

**Chương 2: CÀI ĐẶT HADOOP CLUSTER**

**1.Cài đặt phụ trợ**

**Cài đặt VMWare Tools:**

# sudo apt update

# sudo apt install open-vm-tools-desktop

**Cài đặt Vim:**

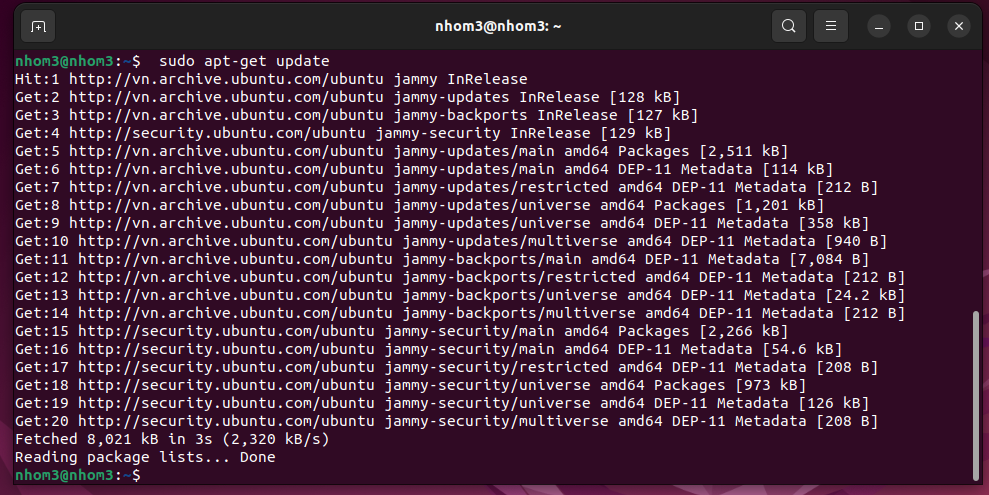
# sudo apt install vim

**Cài đặt git:**

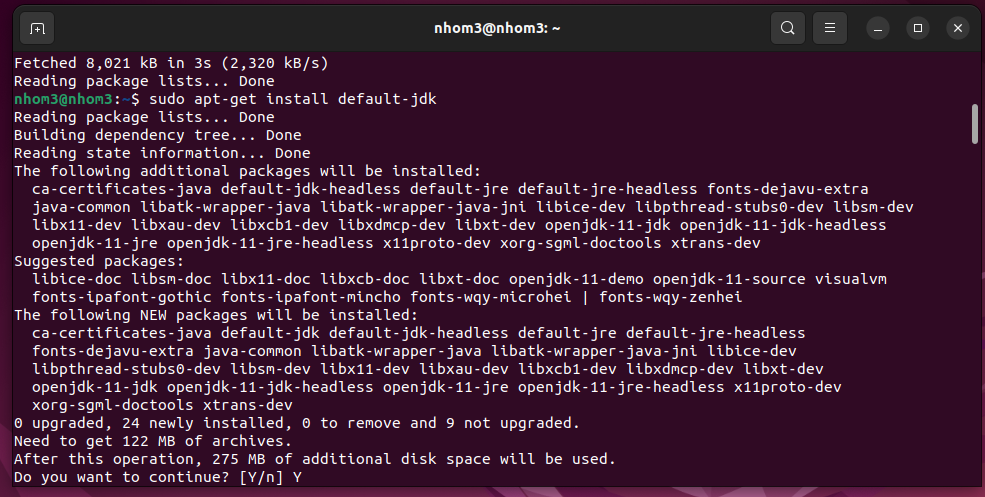
# sudo apt install git

**Cài đặt OpenJDK**

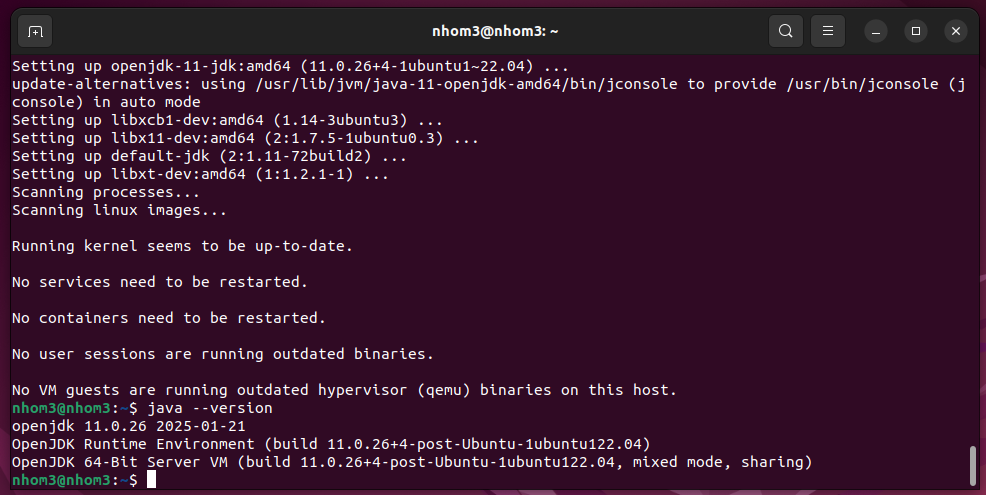
# sudo apt-get update



# sudo apt-get install default-jdk

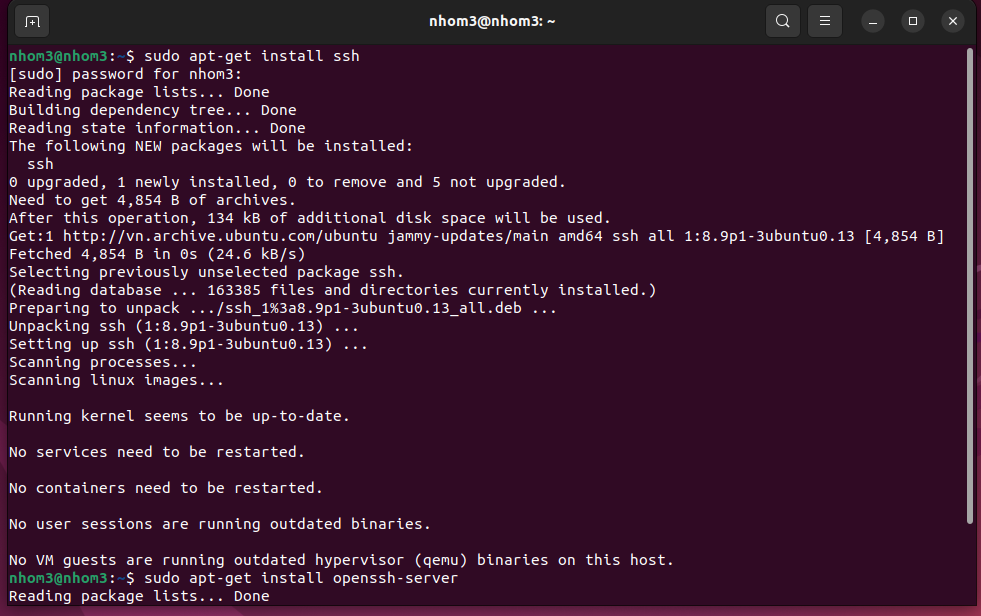


# java --version



**Cài SSH:**

# sudo apt-get install ssh



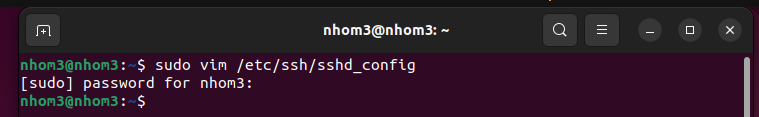
# sudo apt-get install openssh-server

# reboot



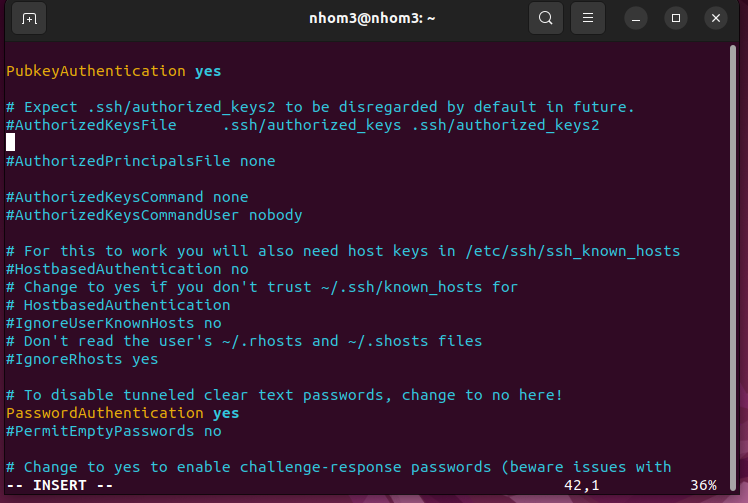
**2. Cấu hình SSH**

# sudo vim /etc/ssh/sshd\_config



Tìm đoạn # PubkeyAuthentication yes. Bỏ dấu # phía trước thành ... PubkeyAuthentication yes ...

Tìm đoạn # PasswordAuthentication yes. Bỏ dấu # phía trước thành ... PasswordAuthentication yes …



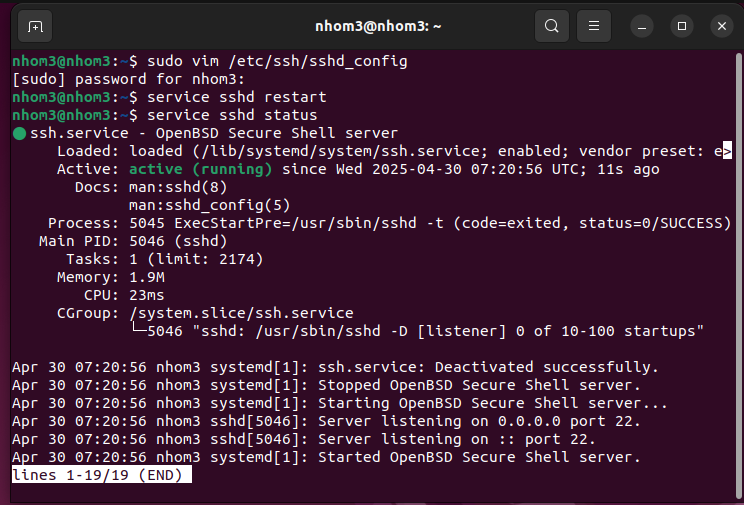
Sau khi sửa thì nhấn phím ESC, nhập :wq để lưu và thoát khỏi vim. :w !sudo tee %

Khởi động lại SSH

# service sshd restart

Xem trạng thái dịch vụ SSH

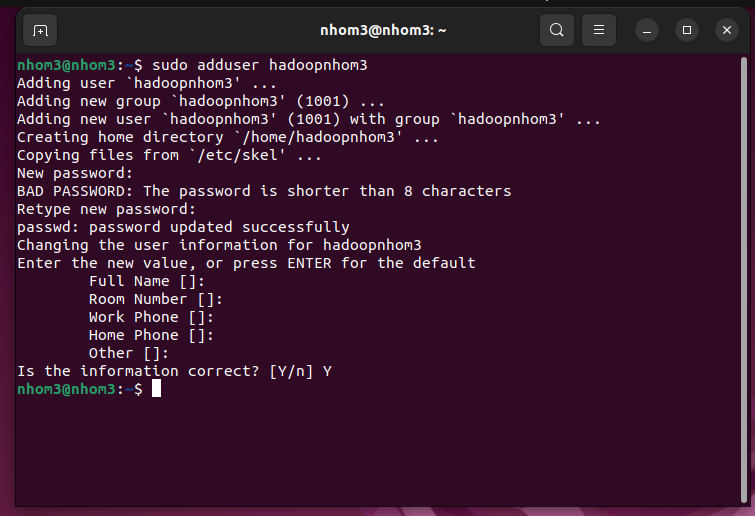
# service sshd status



**3. Tạo user hadoop**

Tạo user hadoopnhom3 để quản lý các permission cho đơn giản

# sudo adduser hadoopnhom3



**4. Cài đặt Hadoop 3.3.4**

Chuyển qua hadoopuser

# su hadoopnhom3



Chuyển qua thư mục /home/hadoopnhom3

# cd /home/hadoopnhom3/

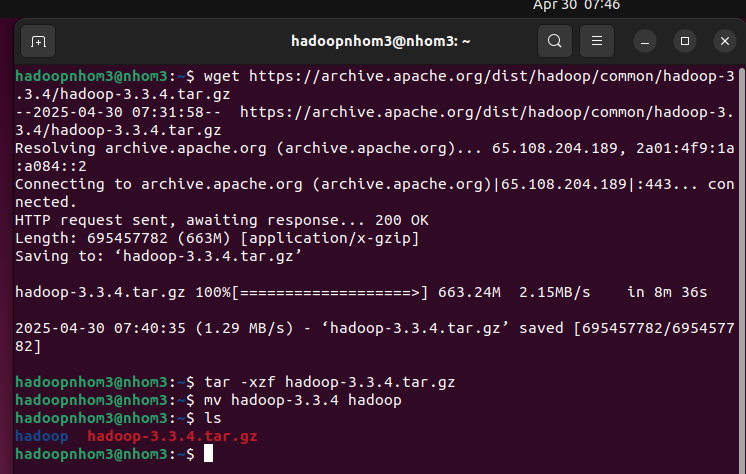
#wget [https://archive.apache.org/dist/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz](https://www.google.com/search?q=https://archive.apache.org/dist/hadoop/common/hadoop-3.3.4/hadoop-3.3.4.tar.gz" \t "_blank)

Giải nén file

# tar -xzf hadoop-3.3.4.tar.gz

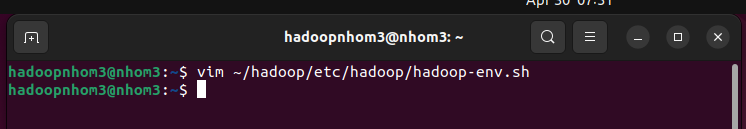
Đổi tên thư mục giải nén thành hadoop cho dễ quản lý

# mv hadoop-3.3.4 hadoop



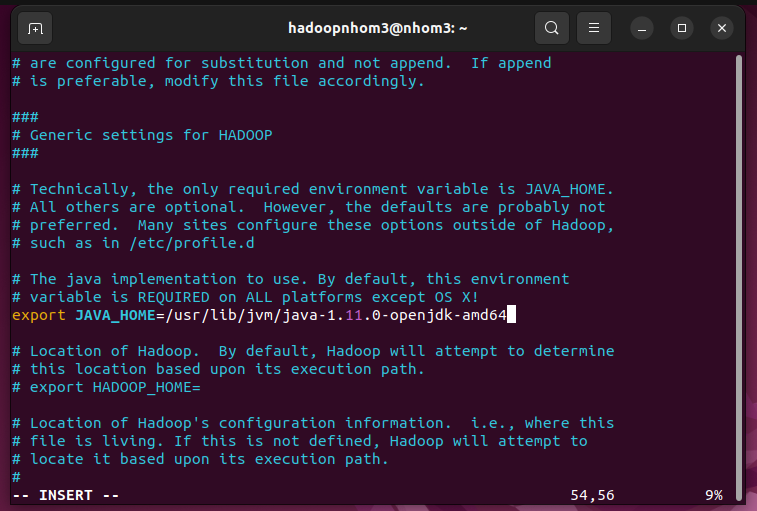
**5. Thiết lập JAVA\_HOME**

# vim ~/hadoop/etc/hadoop/hadoop-env.sh



Tìm đoạn export JAVA\_HOME=... sửa thành như sau:

export JAVA\_HOME=/usr/lib/jvm/java-1.11.0-openjdk-amd64



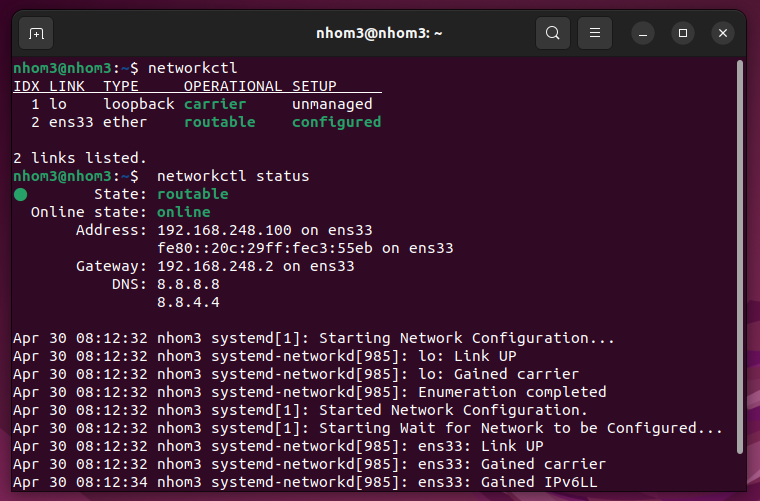
1. **Thiết lập IP tĩnh cho master:**

- Kiểm tra các thiết bị mạng

# networkctl

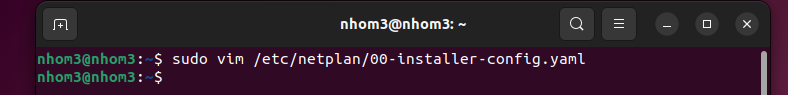
- In trạng thái từng địa chỉ ip trên hệ thống:

# networkctl status



- Cấu hình ip tĩnh

# vim /etc/netplan/00-installer-config.yaml



Thêm nội dung sau:

#this is the network config written by 'subiquity'

network:

ethernets:

ens33:

dhcp: false

dhcp6: false

addresses: [192.168.248.100/24]

routes:

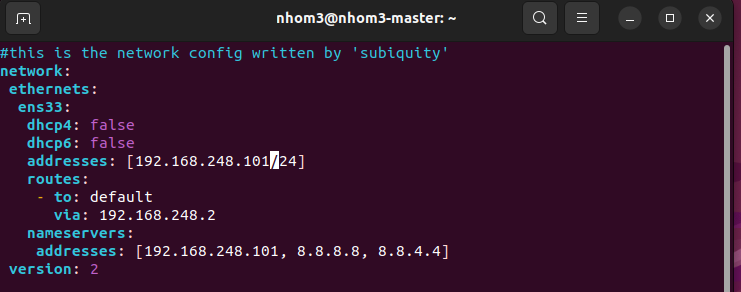
- to: default

via: 192.168.248.2

nameservers:

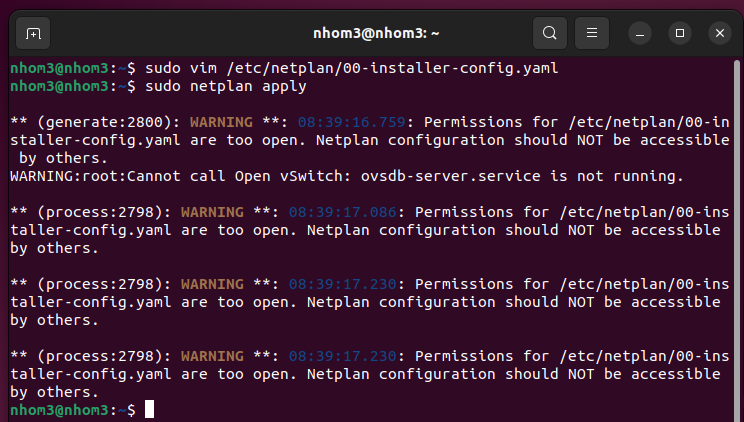
addresses: [192.168.248.1, 8.8.8.8, 8.8.4.4]

version: 2



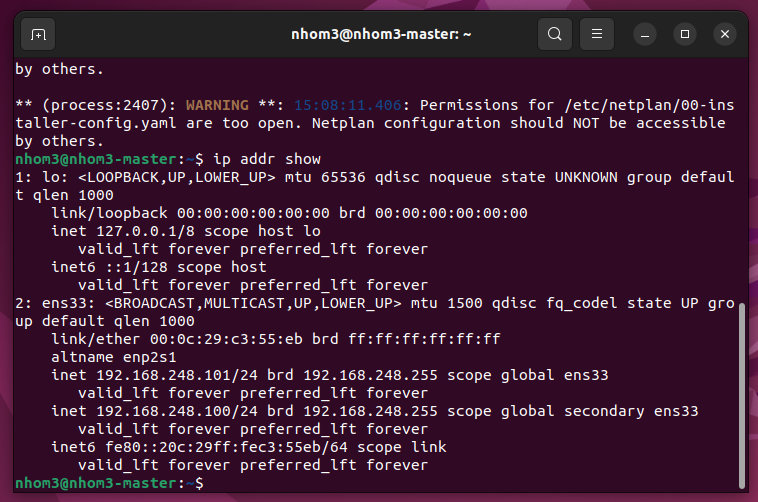
- lưu file và chạy lệnh sau

# sudo netplan apply



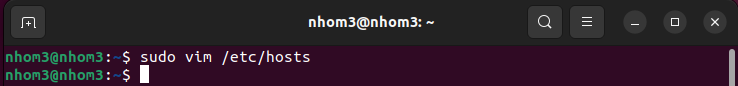
- kiểm tra cấu hình ip mới

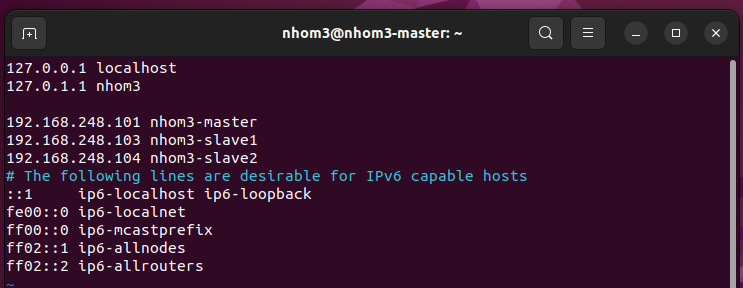
# ip addr show



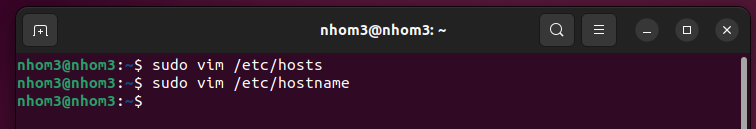
1. Cài đặt OpenJDK Đã hoàn thành trong Phần 1
2. Cài đặt SSH Đã hoàn thành trong Phần 1
   1. Cấu hình SSH Đã hoàn thành trong Phần 1
3. Cấu hình host/hostname cho máy master:

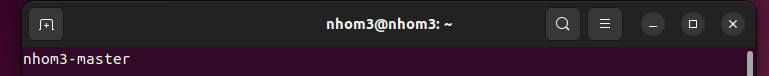
# sudo vim /etc/hosts





# sudo vim /etc/hostname





1. **Cấu hình thông số cho hadoop:**

**10.1 File .bashrc:**

# vim ~/.bashrc

export JAVA\_HOME=/usr/lib/jvm/java-1.11.0-openjdk-amd64

export HADOOP\_HOME=/home/hadoopnhom3/hadoop

export PATH=$PATH:$HADOOP\_HOME/bin

export PATH=$PATH:$HADOOP\_HOME/sbin

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

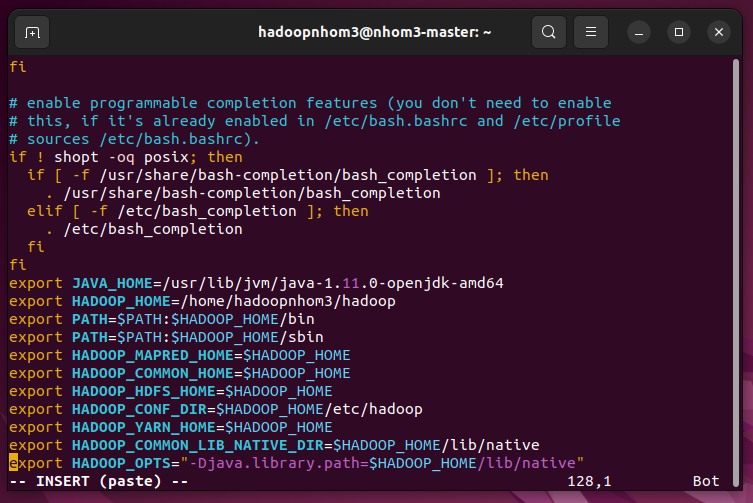
export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

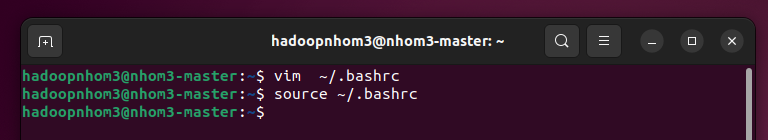
export HADOOP\_CONF\_DIR=$HADOOP\_HOME/etc/hadoop

export HADOOP\_YARN\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"

# source ~/.bashrc



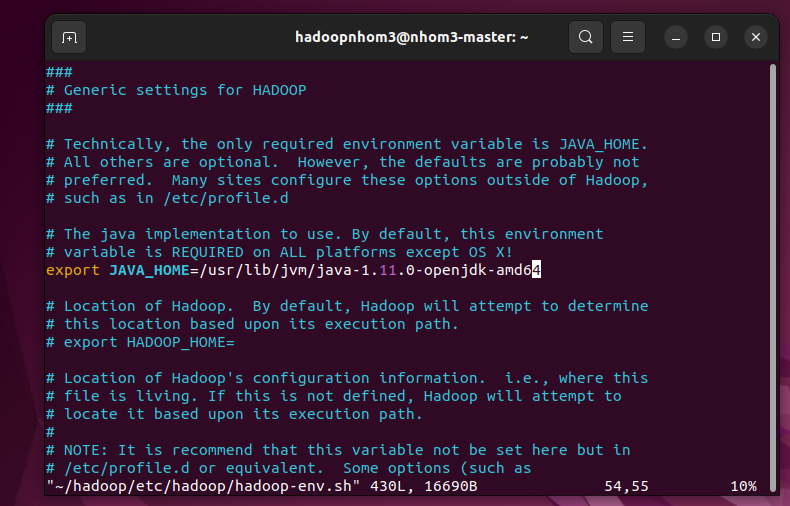


**10.2 File hadoop-env.sh**

# vim ~/hadoop/etc/hadoop/hadoop-env.sh

- Tìm đoạn export JAVA\_HOME=... sửa thành như sau:

export JAVA\_HOME=/usr/lib/jvm/java-1.11.0-openjdk-amd64



**10.3 File core-site.xml**

# vim ~/hadoop/etc/hadoop/core-site.xml

<configuration>

<property>

<name>fs.defaultFS</name>

<value>hdfs://khai-master:9000</value>

<description>Use HDFS as file storage engine</description>

</property>

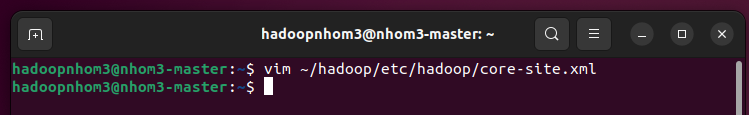
<property>

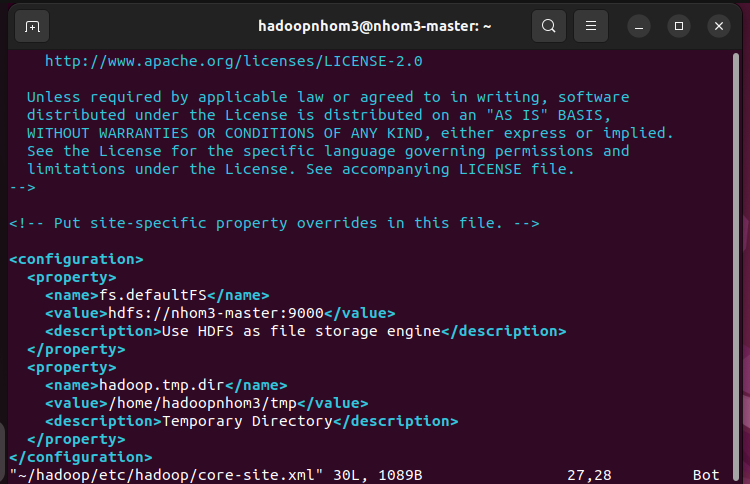
<name>hadoop.tmp.dir</name>

<value>/home/hadoopnhom3/tmp</value>

<description>Temporary Directory</description>

</property>

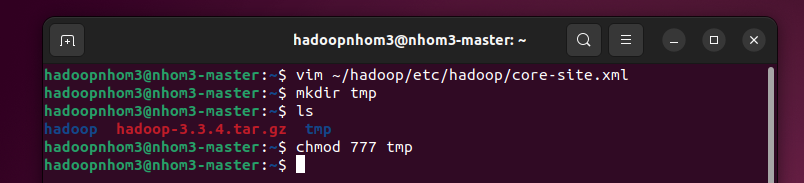
</configuration>



# mkdir tmp

# ls

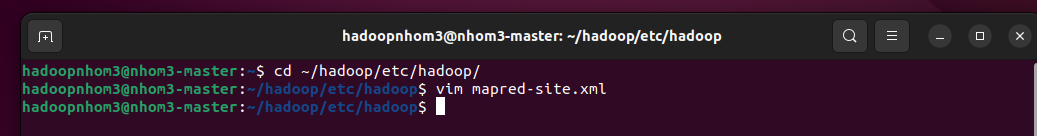
# chmod 777 tmp



**10.4 File mapred-site.xml (**chỉ cấu hình ở máy master**)**

# cd ~/hadoop/etc/hadoop/

# vim mapred-site.xml



<configuration>

<property>

<name>mapreduce.application.classpath</name> <value>$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/\*:$HADOOP\_MAPRED\_HOME/share/hadoop/mapreduce/lib/\*</value>

</property>

<property>

<name>mapreduce.jobtracker.address</name>

<value>nhom3-master:9001</value>

<description>

The host and port that the MapReduce job tracker runs at. If “local”, then jobs

are run in-process as a single map and reduce task.

</description>

</property>

<property>

<name>mapreduce.framework.name</name>

<value>yarn</value>

<description>The framework for running mapreduce jobs</description>

</property>

<property>

<name>yarn.app.mapreduce.am.env</name>

<value>HADOOP\_MAPRED\_HOME=/home/hadoopnhom3/hadoop</value>

</property>

<property>

<name>mapreduce.map.env</name>

<value>HADOOP\_MAPRED\_HOME=/home/hadoopnhom3/hadoop</value>

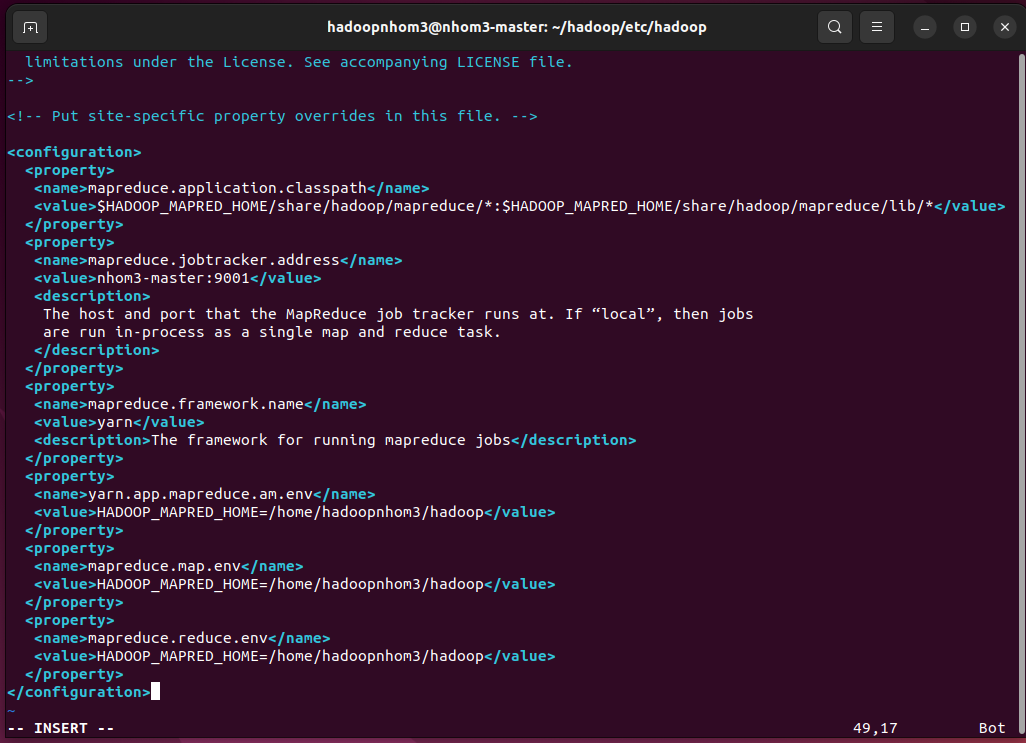
</property>

<property>

<name>mapreduce.reduce.env</name>

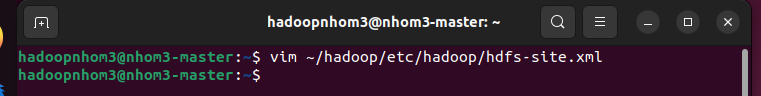
<value>HADOOP\_MAPRED\_HOME=/home/hadoopnhom3/hadoop</value>

</property>



**10.5 File hdfs-site.xml**

# vim ~/hadoop/etc/hadoop/hdfs-site.xml



<configuration>

<property>

<name>dfs.replication</name>

<value>2</value>

<description>

Default block replication. The actual number of replications can be specified when the file is created. The default is used if replication is not specified in create time.

</description>

</property>

<property>

<name>dfs.namenode.name.dir</name>

<value>/home/hadoopnhom3/hadoop\_data/hdfs/namenode</value>

<description>

Determines where on the local filesystem the DFS name node should store the name table (fsimage). If this is a comma-delimited list of directories then the name table is replicated in all of the directories, for redundancy.

</description>

</property>

<property>

<name>dfs.datanode.data.dir</name>

<value>/home/hadoopnhom3/hadoop\_data/hdfs/datanode</value>

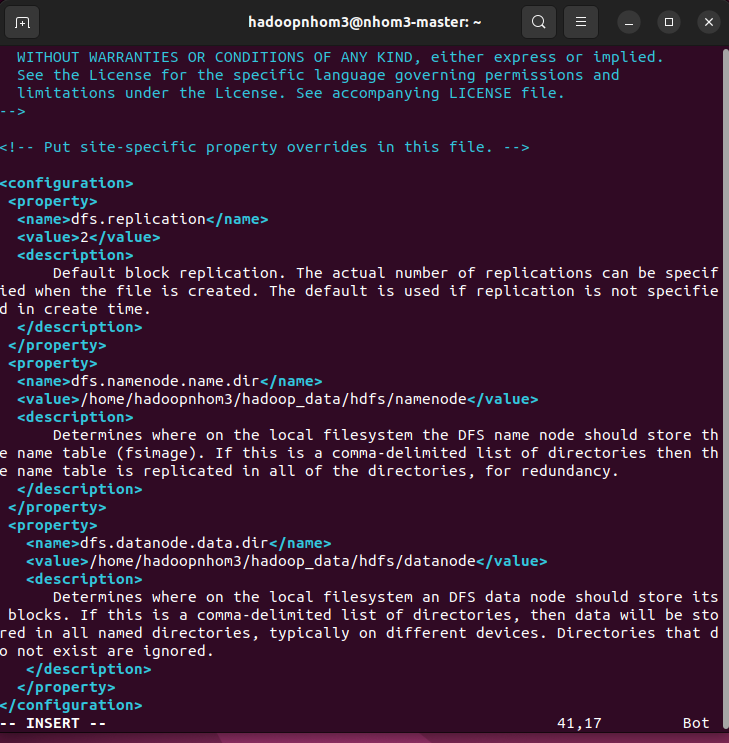
<description>

Determines where on the local filesystem an DFS data node should store its blocks. If this is a comma-delimited list of directories, then data will be stored in all named directories, typically on different devices. Directories that do not exist are ignored.

</description>

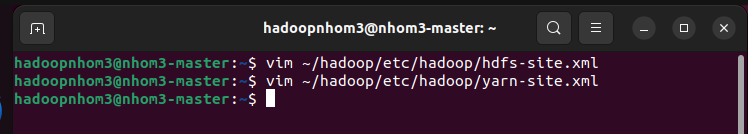
</property>

</configuration>



**10.6 File yarn-site.xml**

# vim ~/hadoop/etc/hadoop/yarn-site.xml



<configuration>

<property>

<name>yarn.nodemanager.aux-services</name>

<value>mapreduce\_shuffle</value>

</property>

<property>

<name>yarn.nodemanager.env-whitelist</name> <value>JAVA\_HOME,HADOOP\_COMMON\_HOME,HADOOP\_HDFS\_HOME,HADOOP\_CONF\_DIR,CLASSPATH\_PREPEND\_DISTCACHE,HADOOP\_YARN\_HOME,HADOOP\_MAPRED\_HOME</value>

</property>

<property>

<name>yarn.resourcemanager.scheduler.address</name>

<value>nhom3-master:9002</value>

</property>

<property>

<name>yarn.resourcemanager.address</name>

<value>nhom3-master:9003</value>

</property>

<property>

<name>yarn.resourcemanager.webapp.address</name>

<value>nhom3-master:9004</value>

</property>

<property>

<name>yarn.resourcemanager.resourcetracker.address</name>

<value>nhom3-master:9005</value>

</property>

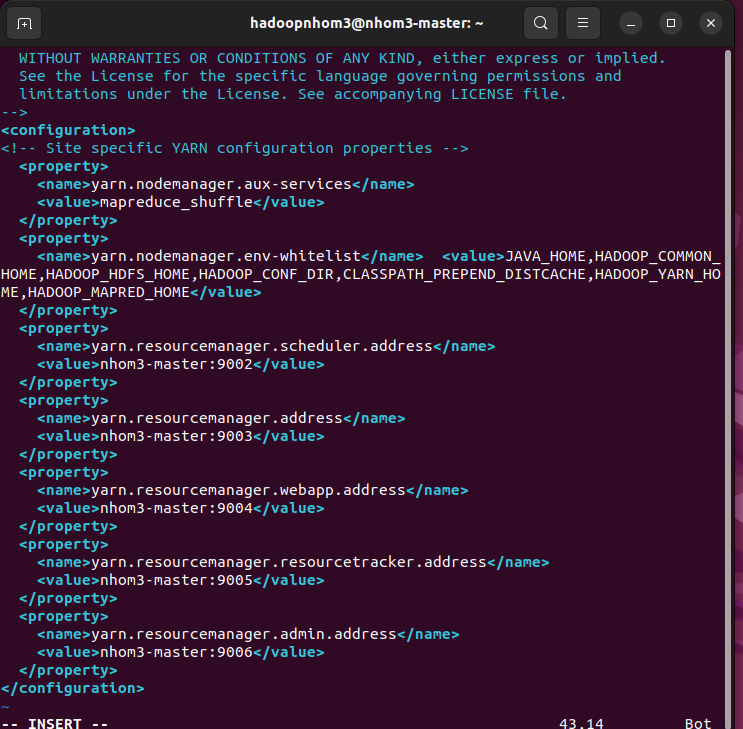
<property>

<name>yarn.resourcemanager.admin.address</name>

<value>nhom3-master:9006</value>

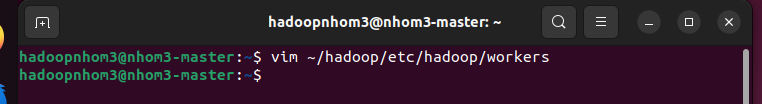
</property>

</configuration>



1. **Chỉ ra các máy slave( chỉ ở máy master)**

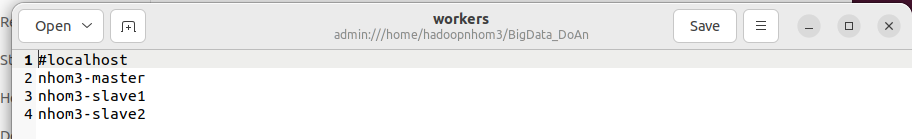
# vim ~/hadoop/etc/hadoop/workers



nhom3-master

nhom3-slave1

nhom3-slave2



1. **Tạo máy các máy slave**

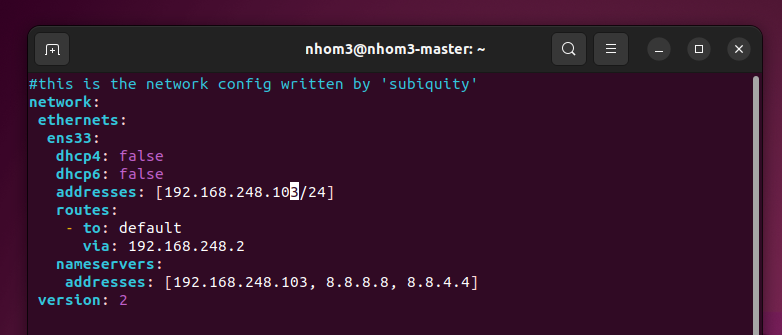
**12.1 Máy slave1**

Tắt máy Master.

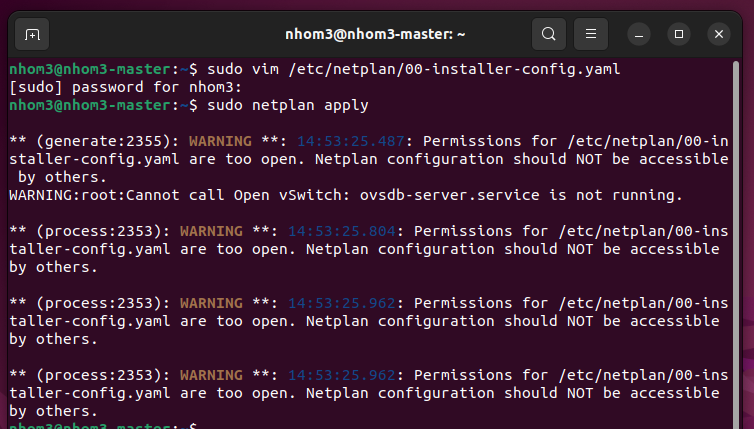
Copy Master ra, đổi tên thành Slave1

Mở máy slave, chỉnh lại IP tĩnh và các thông số cho phù hợp: hosts, hostname…

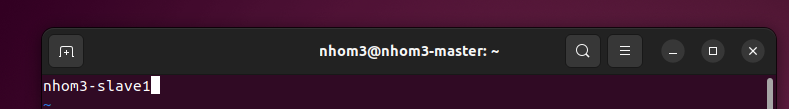
# sudo vim /etc/netplan/00-installer-config.yaml



# sudo neyplan apply



# sudo vim /etc/hostname



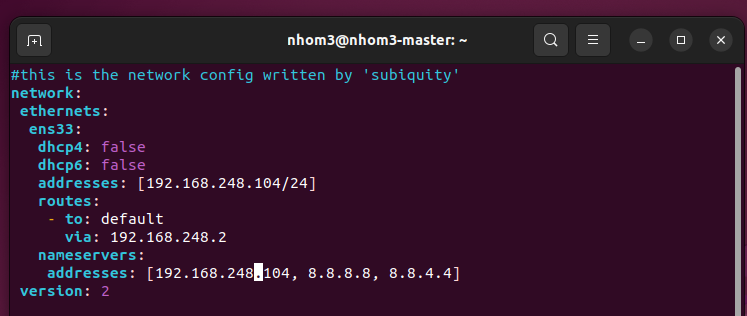
**12.2 Máy slave2**

Tắt máy Master.

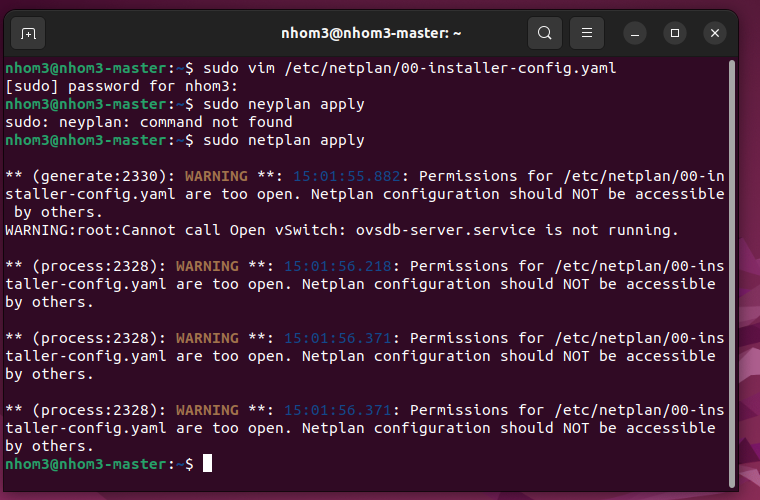
Copy Master ra, đổi tên thành Slave2

Mở máy slave, chỉnh lại IP tĩnh và các thông số cho phù hợp: hosts, hostname…

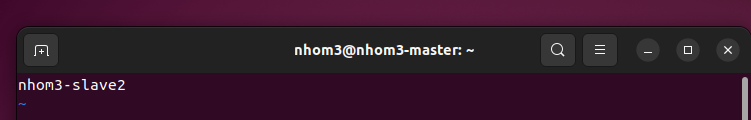
# sudo vim /etc/netplan/00-installer-config.yaml



# sudo netplan apply



# sudo vim /etc/hostname



1. **Cài đặt ssh key giữa các node**

**Thao tác này chỉ thực hiện trên master**

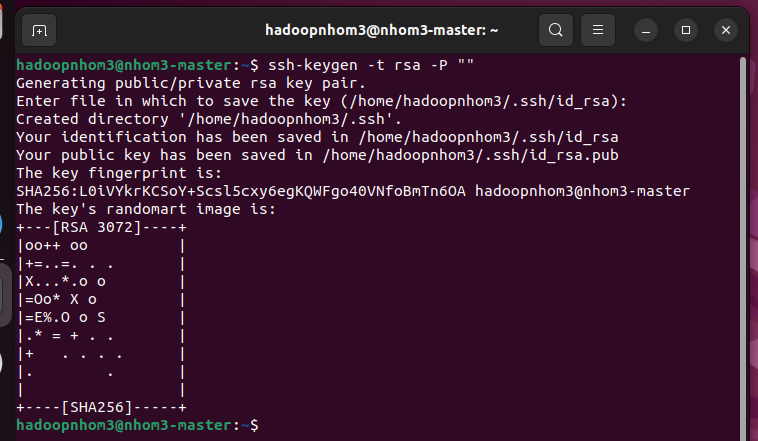
Khởi động máy nhom3-master

Đăng nhập với hadoopnhom3

Tạo ssh key

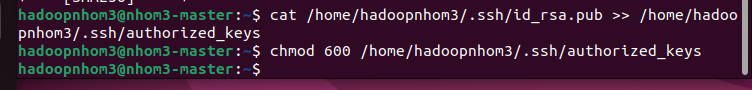
# ssh-keygen -t rsa -P ""

Nhấn Enter để chấp nhận giá trị mặc định



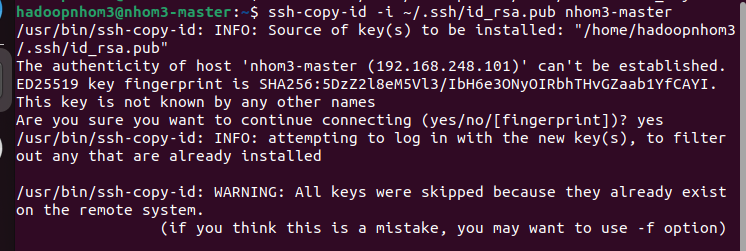
# cat /home/hadoopnhom3/.ssh/id\_rsa.pub >> /home/hadoopnhom3/.ssh/authorized\_keys

# chmod 600 /home/hadoopnhom3/.ssh/authorized\_keys



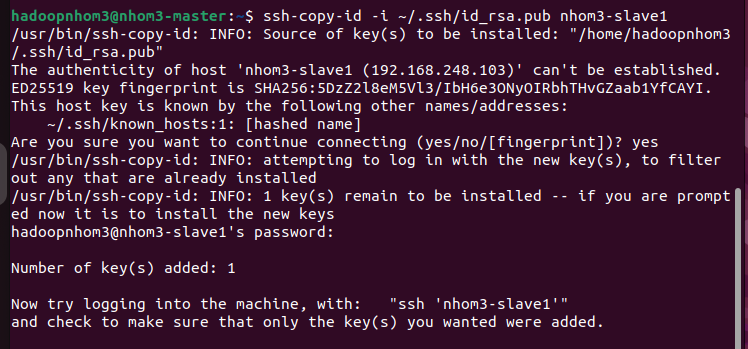
Share ssh key giữa master - master

# ssh-copy-id -i ~/.ssh/id\_rsa.pub nhom3-master



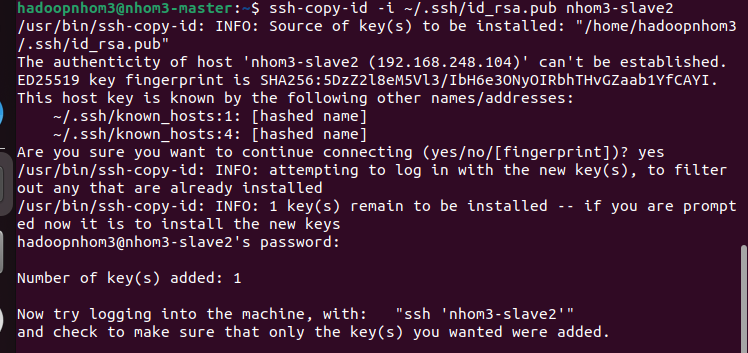
Share ssh key giữa master - slave1

# ssh-copy-id -i ~/.ssh/id\_rsa.pub nhom3-slave1



Share ssh key giữa master - slave2

# ssh-copy-id -i ~/.ssh/id\_rsa.pub nhom3-slave2



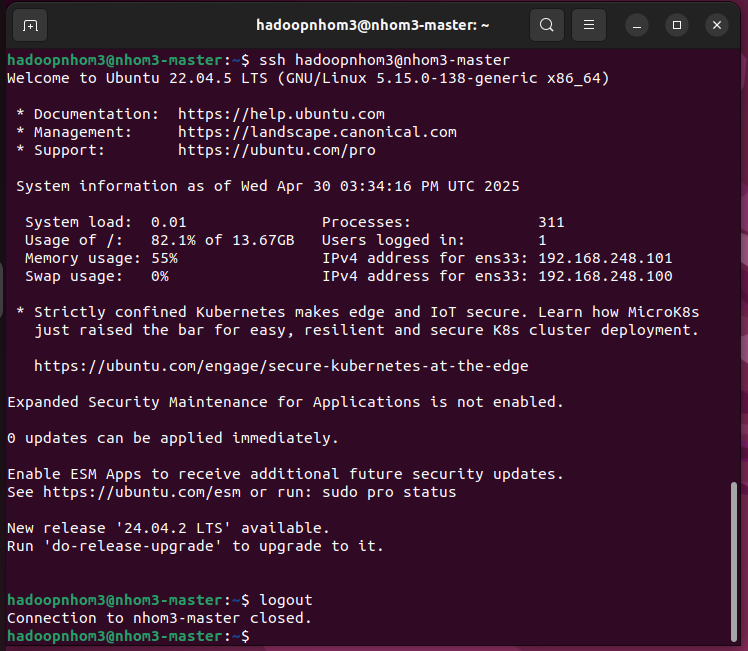
**10.1 Test kết nối ssh**

Test kết nối tới master

# ssh hadoopnhom3@nhom3-master

Đăng xuất

# logout

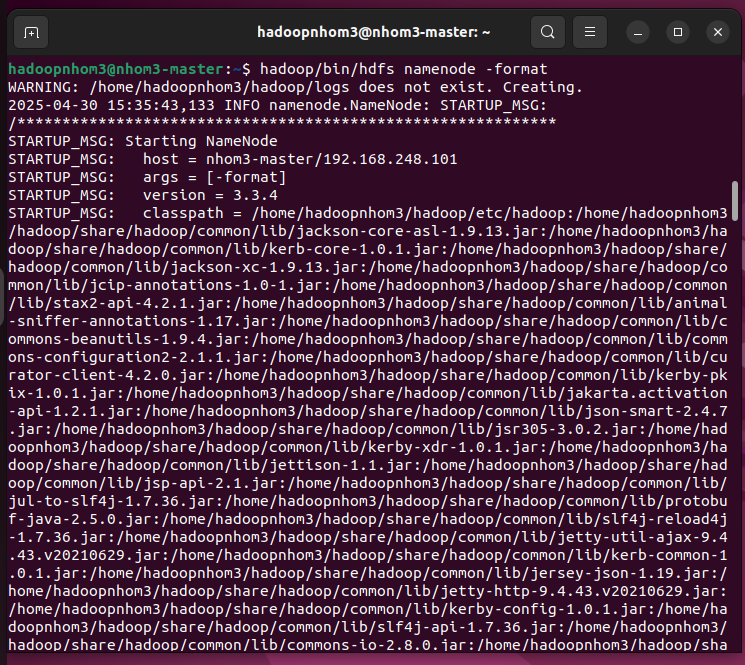


**11. Format namenode**

**Thao tác này chỉ thực hiện trên master và chỉ làm 1 lần.**

Cập nhật lại các thông tin cấu hình của master

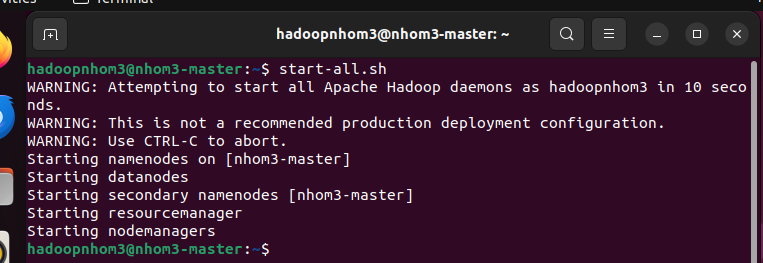
$ hadoop/bin/hdfs namenode -format



**12. Kiểm tra xem mọi thứ đã ổn**

Trên master chúng ta chạy lệnh sau để khởi động các thành phần có trong Hadoop

# start-all.sh (chỉ cần chạy trên Master, không cần chạy trên Slave)



Kiểm tra các thành phần có chạy đủ bằng lệnh sau

# jps

