TITLE TEXT

SUBTITLE TEXT

Contributors

Table of contents

1. Home

1.1 step_into_llm_docs

Online docs for mindspore-courses/step_into_llm .

1.2 MindNLP

Please visit MindNLP to view the full API documentation for MindNLP.

2. step_into_chatgpt

2.1 Transformer

2.2 BERT

3. step_into_llm

3.1 ChatGLM

4. others

4.1 Ascend physical machine installation Cann and MindSpore environment guidance

4.2 Pre -reading and resource preparation

1. Log in to the fortress machine or physical machine of the Shengteng chip

   Log in to a fortress machine needs to contact the administrator to get the corresponding resources

1. Download the dependency package in advance (see the correspondence and download address below for the version of the version)
2. Ubuntu 20.04 ISO installation package
3. Cann and its kernel
4. Anaconda installation supporting MindSpore environment

4.2.1 VPN opens the network of the fortress machine

(1). VPN connecting the Shengteng community can access the fortress machine in Hangzhou

according to Shengteng Ecological Intelligence Laboratory Network Connection Guidance, Import configuration files to connect to the first layer of VPN.

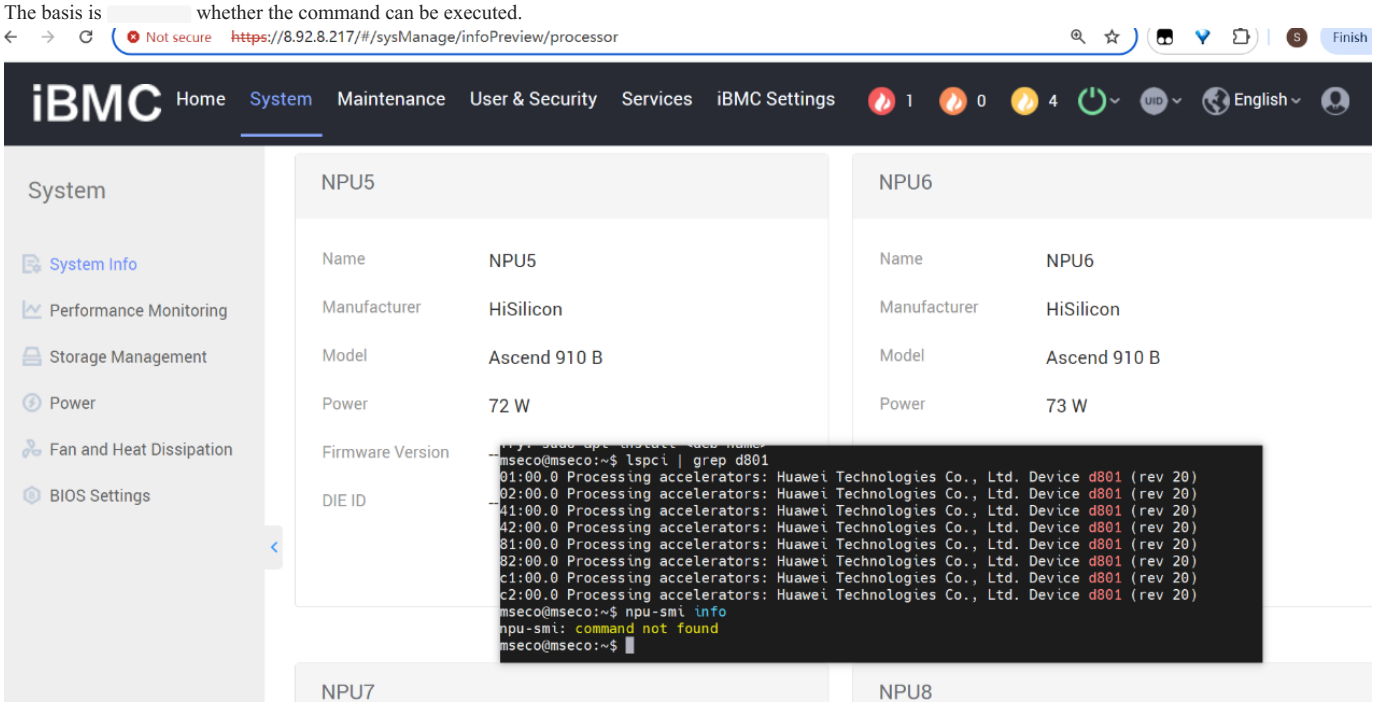
If the CMD can Ping Fortress IP indicates that the network has been opened, you can use MobaxTerm to log in to the fortress machine.

(2). Dongguan-Tuanbo Waka's fortress machine also needs to connect the second layer of VPN

according to The network environment VPN connection guidelines for Dongguan-Tuanpububi Machine after completing the network environment configuration, you can log in to the fortress machine.

4.3 IBMC installation operating system and NPU driver firmware

The basis is whether the command can be executed.



Try to use the operating system version and the corresponding firmware that HDK teams have been developed, check the details Official website document "Version Support Form" there is no guarantee outside the "Version of the Facial Form". The user needs to install the front dependencies by themselves to resolve the conflict and the source code to compile the NPU driver and firmware. As of 2024.10.11, the kernel 5.15.122-GENERIC Ubuntu22.04 source code compilation and adaptation problem, other systems are similar to similar.According to Altas800 (9000) official tutorial installation documentation it is recommended to use the operating system version corresponding to the NPU -driven firmware in the "Version of the Edition" uses binary installation. It is not recommended to install the source code.

```
root@mseco:~/temp# ./Ascend-hdk-910-npu-driver_6.0.0_linux-aarch64.run --full --install-for-all
Verifying archive integrity...  100%   SHA256 checksums are OK. All good.
Uncompressing ASCEND DRIVER RUN PACKAGE  100%
[Driver] [2024-10-10 03:15:55] [INFO]Start time: 2024-10-10 03:15:55
[Driver] [2024-10-10 03:15:55] [INFO]LogFile: /var/log/ascend_seclog/ascend_install.log
[Driver] [2024-10-10 03:15:55] [INFO]OperationLogFile: /var/log/ascend_seclog/operation.log
[Driver] [2024-10-10 03:15:55] [INFO]base version is none.
[Driver] [2024-10-10 03:15:55] [WARNING]Do not power off or restart the system during the installation/upgrade
[Driver] [2024-10-10 03:15:55] [INFO]set username and usergroup, HwHiAiUser:HwHiAiUser
/usr/local/Ascend/driver/tools/upgrade-tool: error while loading shared libraries: libc_sec.so: cannot open shared object file:
 No such file or directory
^CSignal caught, cleaning up
root@mseco:~/temp# find / -name libc_sec.so
/usr/local/Ascend/driver/lib64/common/libc_sec.so
root@mseco:~/temp# export LD_LIBRARY_PATH=/usr/local/Ascend/driver/lib64/driver:/usr/local/Ascend/driver/lib64/common:/lib:/usr
/lib:usr/lib64:/usr/local/lib:$LD_LIBRARY_PATH
root@mseco:~/temp# ./Ascend-hdk-910-npu-driver_6.0.0_linux-aarch64.run --full --install-for-all
Verifying archive integrity...  100%   SHA256 checksums are OK. All good.
Uncompressing ASCEND DRIVER RUN PACKAGE  100%
[Driver] [2024-10-10 03:16:43] [INFO]Start time: 2024-10-10 03:16:43
[Driver] [2024-10-10 03:16:43] [INFO]LogFile: /var/log/ascend_seclog/ascend_install.log
[Driver] [2024-10-10 03:16:43] [INFO]OperationLogFile: /var/log/ascend_seclog/operation.log
[Driver] [2024-10-10 03:16:43] [INFO]base version is none.
[Driver] [2024-10-10 03:16:43] [WARNING]Do not power off or restart the system during the installation/upgrade
[Driver] [2024-10-10 03:16:43] [INFO]set username and usergroup, HwHiAiUser:HwHiAiUser
[Driver] [2024-10-10 03:17:52] [INFO]driver install type: DKMS
[Driver] [2024-10-10 03:17:52] [INFO]upgradePercentage:10%
[Driver] [2024-10-10 03:17:54] [INFO]upgradePercentage:30%
[Driver] [2024-10-10 03:17:54] [INFO]upgradePercentage:40%
[Driver] [2024-10-10 03:18:11] [ERROR]Dkms install failed, details in : /var/log/ascend_seclog/ascend_install.log
[Driver] [2024-10-10 03:18:11] [ERROR]Driver_ko_install failed, details in : /var/log/ascend_seclog/ascend_install.log
[Driver] [2024-10-10 03:18:11] [INFO]Failed to install driver package, please retry after uninstall and reboot!
[Driver] [2024-10-10 03:18:11] [INFO]End time: 2024-10-10 03:18:11
```

4.3.1 Recommended NPU driver and firmware on Ubuntu 20.04

Install ubuntu20.04 operating system

Most of the installation tutorial in Ubuntu can be operated

```
mseco@mseco:~$ sudo vim /etc/ssh/sshd_config
mseco@mseco:~$ sudo service sshd restart
```

```
#LoginGraceTime 2m
#PermitRootLogin prohibit-password
PermitRootLogin yes
#StrictModes yes
#MaxAuthTries 6
#MaxSessions 10

#PubkeyAuthentication yes

# Expect .ssh/authorized_keys2 to be disregarded by default in future.
#AuthorizedKeysFile     .ssh/authorized_keys .ssh/authorized_keys2

#AuthorizedPrincipalsFile none

#AuthorizedKeysCommand none
#AuthorizedKeysCommandUser nobody

# For this to work you will also need host keys in /etc/ssh/ssh_known_hosts
#HostbasedAuthentication no
# Change to yes if you don't trust ~/.ssh/known_hosts for
# HostbasedAuthentication
#IgnoreUserKnownHosts no
# Don't read the user's ~/.rhosts and ~/.shosts files
#IgnoreRhosts yes

# To disable tunneled clear text passwords, change to no here!
#PasswordAuthentication no
PasswordAuthentication yes
#PermitEmptyPasswords no

# Change to yes to enable challenge-response passwords (beware issues with
# some PAM modules and threads)
```

After the MOBAXTERM connection

Install NPU driver and firmware

IBMC can directly view the NPU model,

then go to the rising page to find the corresponding version of the installation document.

The kernel version will be automatically upgraded, and the current kernel version is not adapted to the current kernel version upgrade. Therefore, the function of automatic upgrade kernel can be turned off.

```
root@tridu33:~/done# ./Ascend-hdk-910-npu-driver_6.0.0_linux-aarch64.run  --full --install-for-all
Verifying archive integrity...  100%   SHA256 checksums are OK. All good.
Uncompressing ASCEND DRIVER RUN PACKAGE  100%
[Driver] [2024-10-11 12:40:06] [INFO]Start time: 2024-10-11 12:40:06
[Driver] [2024-10-11 12:40:06] [INFO]LogFile: /var/log/ascend_seclog/ascend_install.log
[Driver] [2024-10-11 12:40:06] [INFO]OperationLogFile: /var/log/ascend_seclog/operation.log
[Driver] [2024-10-11 12:40:06] [WARNING]Do not power off or restart the system during the installation/upgrade
[Driver] [2024-10-11 12:40:06] [INFO]set username and usergroup, HwHiAiUser:HwHiAiUser
[Driver] [2024-10-11 12:40:07] [INFO]driver install type: DKMS
[Driver] [2024-10-11 12:40:07] [INFO]upgradePercentage:10%
[Driver] [2024-10-11 12:40:11] [INFO]upgradePercentage:30%
[Driver] [2024-10-11 12:40:11] [INFO]upgradePercentage:40%
[Driver] [2024-10-11 12:40:30] [INFO]upgradePercentage:90%
[Driver] [2024-10-11 12:40:30] [INFO]Waiting for device startup...
[Driver] [2024-10-11 12:40:32] [INFO]Device startup success
[Driver] [2024-10-11 12:40:44] [INFO]upgradePercentage:100%
[Driver] [2024-10-11 12:40:55] [INFO]Driver package installed successfully! The new version takes effect immediately.
[Driver] [2024-10-11 12:40:55] [INFO]End time: 2024-10-11 12:40:55
root@tridu33:~/done# ./Ascend-hdk-910-npu-firmware_6.0.0.run --full
Verifying archive integrity...  100%   SHA256 checksums are OK. All good.
Uncompressing ASCEND-HDK-910-NPU FIRMWARE RUN PACKAGE  100%
[Firmware] [2024-10-11 12:41:31] [INFO]Start time: 2024-10-11 12:41:31
[Firmware] [2024-10-11 12:41:31] [INFO]LogFile: /var/log/ascend_seclog/ascend_install.log
[Firmware] [2024-10-11 12:41:31] [INFO]OperationLogFile: /var/log/ascend_seclog/operation.log
[Firmware] [2024-10-11 12:41:32] [WARNING]Do not power off or restart the system during the installation/upgrade
[Firmware] [2024-10-11 12:42:05] [INFO]upgradePercentage: 0%
[Firmware] [2024-10-11 12:42:14] [INFO]upgradePercentage: 0%
[Firmware] [2024-10-11 12:42:23] [INFO]upgradePercentage: 0%
[Firmware] [2024-10-11 12:42:29] [INFO]upgradePercentage: 100%
[Firmware] [2024-10-11 12:42:29] [INFO]Firmware package installed successfully! Reboot now or after driver installation for the installation/upgrade to take effect.
[Firmware] [2024-10-11 12:42:29] [INFO]End time: 2024-10-11 12:42:29
root@tridu33:~/done# uname -r
5.4.0-196-generic
root@tridu33:~/done#
```

report an error, according to the tutorial https://bbs.huaweicloud.com/blogs/423686 Set white

```
root@tridu33:~# uname -r
5.4.0-196-generic
root@tridu33:~#
root@tridu33:~# cat  /etc/default/grub
# If you change this file, run 'update-grub' afterwards to update
# /boot/grub/grub.cfg.
# For full documentation of the options in this file, see:
#   info -f grub -n 'Simple configuration'

GRUB_DEFAULT="Advanced options for Ubuntu>Ubuntu, with Linux 5.4.0-125-generic"
GRUB_DEFAULT=0
GRUB_TIMEOUT_STYLE=hidden
GRUB_TIMEOUT=0
GRUB_DISTRIBUTOR=`lsb_release -i -s 2> /dev/null || echo Debian`
GRUB_CMDLINE_LINUX_DEFAULT=""
GRUB_CMDLINE_LINUX=""

# Uncomment to enable BadRAM filtering, modify to suit your needs
# This works with Linux (no patch required) and with any kernel that obtains
# the memory map information from GRUB (GNU Mach, kernel of FreeBSD ...)
#GRUB_BADRAM="0x01234567,0xfefefefe,0x89abcdef,0xefefefef"

# Uncomment to disable graphical terminal (grub-pc only)
#GRUB_TERMINAL=console

# The resolution used on graphical terminal
# note that you can use only modes which your graphic card supports via VBE
# you can see them in real GRUB with the command `vbeinfo'
#GRUB_GFXMODE=640x480

# Uncomment if you don't want GRUB to pass "root=UUID=xxx" parameter to Linux
#GRUB_DISABLE_LINUX_UUID=true

# Uncomment to disable generation of recovery mode menu entries
#GRUB_DISABLE_RECOVERY="true"

# Uncomment to get a beep at grub start
#GRUB_INIT_TUNE="480 440 1"
```

Then restart

```
root@tridu33:~# npu-smi info
+------------------------------------------------------------------------------------------+
| npu-smi 22.0.4                   Version: 22.0.4                                         |
+----------------------+---------------+----------------------------------------------------+
| NPU   Name           | Health        | Power(W)    Temp(C)           Hugepages-Usage(page)|
| Chip                 | Bus-Id        | AICore(%)   Memory-Usage(MB)  HBM-Usage(MB)        |
+======================+===============+====================================================+
| 0     910B           | OK            | 71.2        44                0    / 0             |
| 0                    | 0000:C1:00.0  | 0           605  / 15137      0    / 32768         |
+======================+===============+====================================================+
| 1     910B           | OK            | 66.2        39                0    / 0             |
| 0                    | 0000:81:00.0  | 0           1240 / 15137      0    / 32768         |
+======================+===============+====================================================+
| 2     910B           | OK            | 68.0        38                0    / 0             |
| 0                    | 0000:41:00.0  | 0           2409 / 15137      0    / 32768         |
+======================+===============+====================================================+
| 3     910B           | OK            | 67.3        43                0    / 0             |
| 0                    | 0000:01:00.0  | 0           2345 / 15039      0    / 32768         |
+======================+===============+====================================================+
| 4     910B           | OK            | 70.2        42                0    / 0             |
| 0                    | 0000:C2:00.0  | 0           751  / 15137      0    / 32768         |
+======================+===============+====================================================+
| 5     910B           | OK            | 66.2        38                0    / 0             |
| 0                    | 0000:82:00.0  | 0           2297 / 15137      0    / 32768         |
+======================+===============+====================================================+
| 6     910B           | OK            | 68.8        39                0    / 0             |
| 0                    | 0000:42:00.0  | 0           2254 / 15137      0    / 32768         |
+======================+===============+====================================================+
| 7     910B           | OK            | 68.1        44                0    / 0             |
| 0                    | 0000:02:00.0  | 0           1302 / 15039      0    / 32768         |
+======================+===============+====================================================+
```

This shows that there is no problem with the driver, you can continue to check

```
```

4.3.2 Supporting Cann and Kernel

Establishing necessary dependent components

```
```

After the installation, you can print the environment variables to see if the value is written correctly. If not, you need to write it yourself or the

```
```

You can install Python3.7.5 according to the tutorial, or you can install the environment of anaconda or miniconda. I use Anaconda as an example here.

```
```

Configure Huawei source as follows

```
```

You can install these front tools.

4.3.3 Install MindSpore

It is recommended to use anaconda, or you can use Ubuntu native PIP+Python environment to install MindSpore.

<! ---->

1. Verify Ubuntu, Cann, MindSpore version supporting relationship
2. Check Can MindSpore use the GPU normally?,

   the figure below is the command to check whether the NPU is used normally:

```
(base) root@tridu33:~/temp# conda env list
# conda environments:
#
base                  *  /root/anaconda3
pt1.8.1                  /root/anaconda3/envs/pt1.8.1
py37                     /root/anaconda3/envs/py37
py37ms1.10.1             /root/anaconda3/envs/py37ms1.10.1
tf1                      /root/anaconda3/envs/tf1
tf2                      /root/anaconda3/envs/tf2

(base) root@tridu33:~/temp# conda activate py37ms1.10.1
(py37ms1.10.1) root@tridu33:~/temp# python -c "import mindspore;mindspore.run_check()"
MindSpore version:  1.10.1
The result of multiplication calculation is correct, MindSpore has been installed successfully!
(py37ms1.10.1) root@tridu33:~/temp# touch test_ms.py
(py37ms1.10.1) root@tridu33:~/temp# vim test_ms.py
(py37ms1.10.1) root@tridu33:~/temp# python test_ms.py
[[[[2. 2. 2. 2.]
   [2. 2. 2. 2.]
   [2. 2. 2. 2.]]

  [[2. 2. 2. 2.]
   [2. 2. 2. 2.]
   [2. 2. 2. 2.]]

  [[2. 2. 2. 2.]
   [2. 2. 2. 2.]
   [2. 2. 2. 2.]]]]
(py37ms1.10.1) root@tridu33:~/temp# cat test_ms.py
import numpy as np
import mindspore as ms
import mindspore.ops as ops

ms.set_context(device_target="Ascend")
x = ms.Tensor(np.ones([1,3,3,4]).astype(np.float32))
y = ms.Tensor(np.ones([1,3,3,4]).astype(np.float32))
print(ops.add(x, y))
```

'Device_target' parameters are \ ['cpu', 'gpu', 'ascend', 'davinci'], Davinci is the old name of Ascend.

4.3.4 Q & A

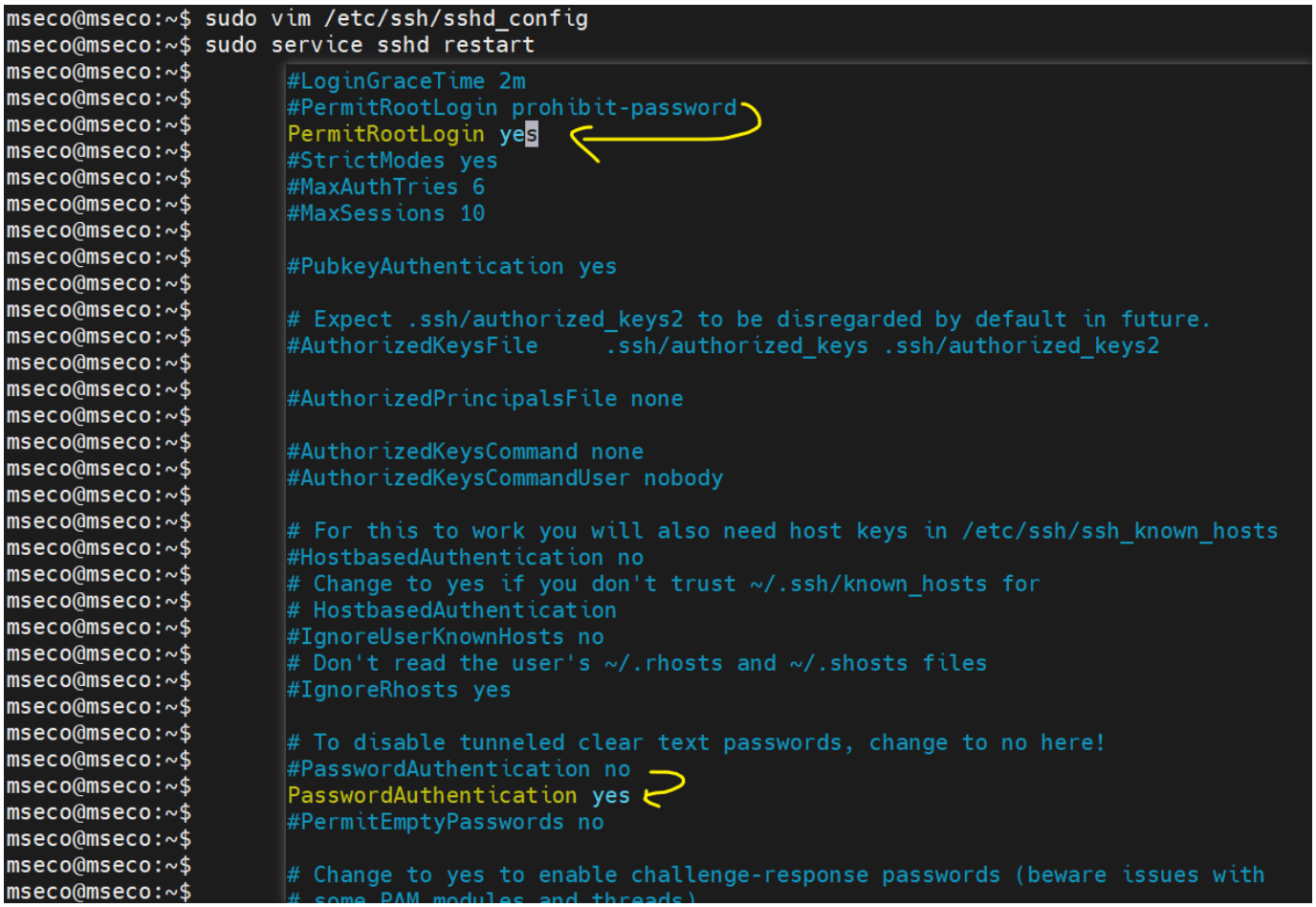
1) When the third -party SSH client logs in Linux instance, it prompts "Access denied" the reason why the error is possible: - SSH login account password input error;

• MacBook or Windows keyboard layout or input method of lax and special character transition causes password errors.

```
mseco@mseco:~$ cat /etc/passwd
root:x:0:0:root:/root:/bin/bash
daemon:x:1:1:daemon:/usr/sbin:/usr/sbin/nologin
bin:x:2:2:bin:/bin:/usr/sbin/nologin
sys:x:3:3:sys:/dev:/usr/sbin/nologin
sync:x:4:65534:sync:/bin:/bin/sync
games:x:5:60:games:/usr/games:/usr/sbin/nologin
man:x:6:12:man:/var/cache/man:/usr/sbin/nologin
lp:x:7:7:lp:/var/spool/lpd:/usr/sbin/nologin
mail:x:8:8:mail:/var/mail:/usr/sbin/nologin
news:x:9:9:news:/var/spool/news:/usr/sbin/nologin
uucp:x:10:10:uucp:/var/spool/uucp:/usr/sbin/nologin
proxy:x:13:13:proxy:/bin:/usr/sbin/nologin
www-data:x:33:33:www-data:/var/www:/usr/sbin/nologin
backup:x:34:34:backup:/var/backups:/usr/sbin/nologin
list:x:38:38:Mailing List Manager:/var/list:/usr/sbin/nologin
irc:x:39:39:ircd:/run/ircd:/usr/sbin/nologin
gnats:x:41:41:Gnats Bug-Reporting System (admin):/var/lib/gnats:/usr/sbin/nologin
nobody:x:65534:65534:nobody:/nonexistent:/usr/sbin/nologin
_apt:x:100:65534::/nonexistent:/usr/sbin/nologin
systemd-network:x:101:102:systemd Network Management,,,:/run/systemd:/usr/sbin/nologin
systemd-resolve:x:102:103:systemd Resolver,,,:/run/systemd:/usr/sbin/nologin
messagebus:x:103:104::/nonexistent:/usr/sbin/nologin
systemd-timesync:x:104:105:systemd Time Synchronization,,,:/run/systemd:/usr/sbin/nologin
pollinate:x:105:1::/var/cache/pollinate:/bin/false
syslog:x:106:113::/home/syslog:/usr/sbin/nologin
uuidd:x:107:114::/run/uuidd:/usr/sbin/nologin
tcpdump:x:108:115::/nonexistent:/usr/sbin/nologin
tss:x:109:116:TPM software stack,,,:/var/lib/tpm:/bin/false
landscape:x:110:117::/var/lib/landscape:/usr/sbin/nologin
fwupd-refresh:x:111:118:fwupd-refresh user,,,:/run/systemd:/usr/sbin/nologin
sshd:x:112:65534::/run/sshd:/usr/sbin/nologin
mseco:x:1000:1000:mseco:/home/mseco:/bin/bash
lxd:x:999:100::/var/snap/lxd/common/lxd:/bin/false
tridu33:x:1001:1001::/home/tridu33:/bin/bash
mseco@mseco:~$ useradd -d /home/tridu33 -s /usr/bin/bash -m tridu33 && passwd tridu33 # 输入小写账号密码
```

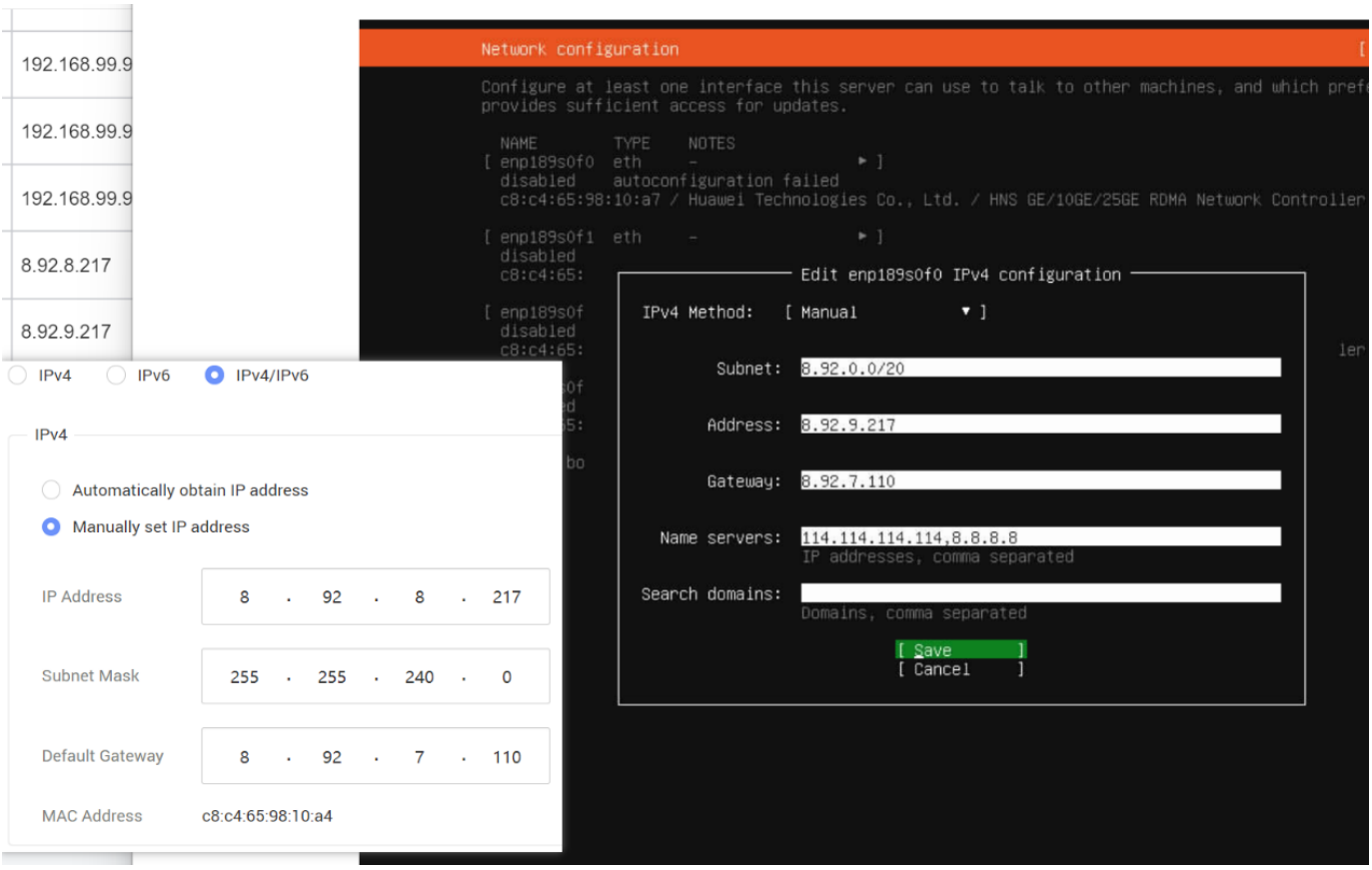

try set a simple password to try whether to log in;

• SSH_CONFIG configuration is not set correctly



2) After installing Ubuntu

Without the correct setting of the network, refer to the IBMC network configuration when it is recommended to reinstall OS. Correctly set the network card:

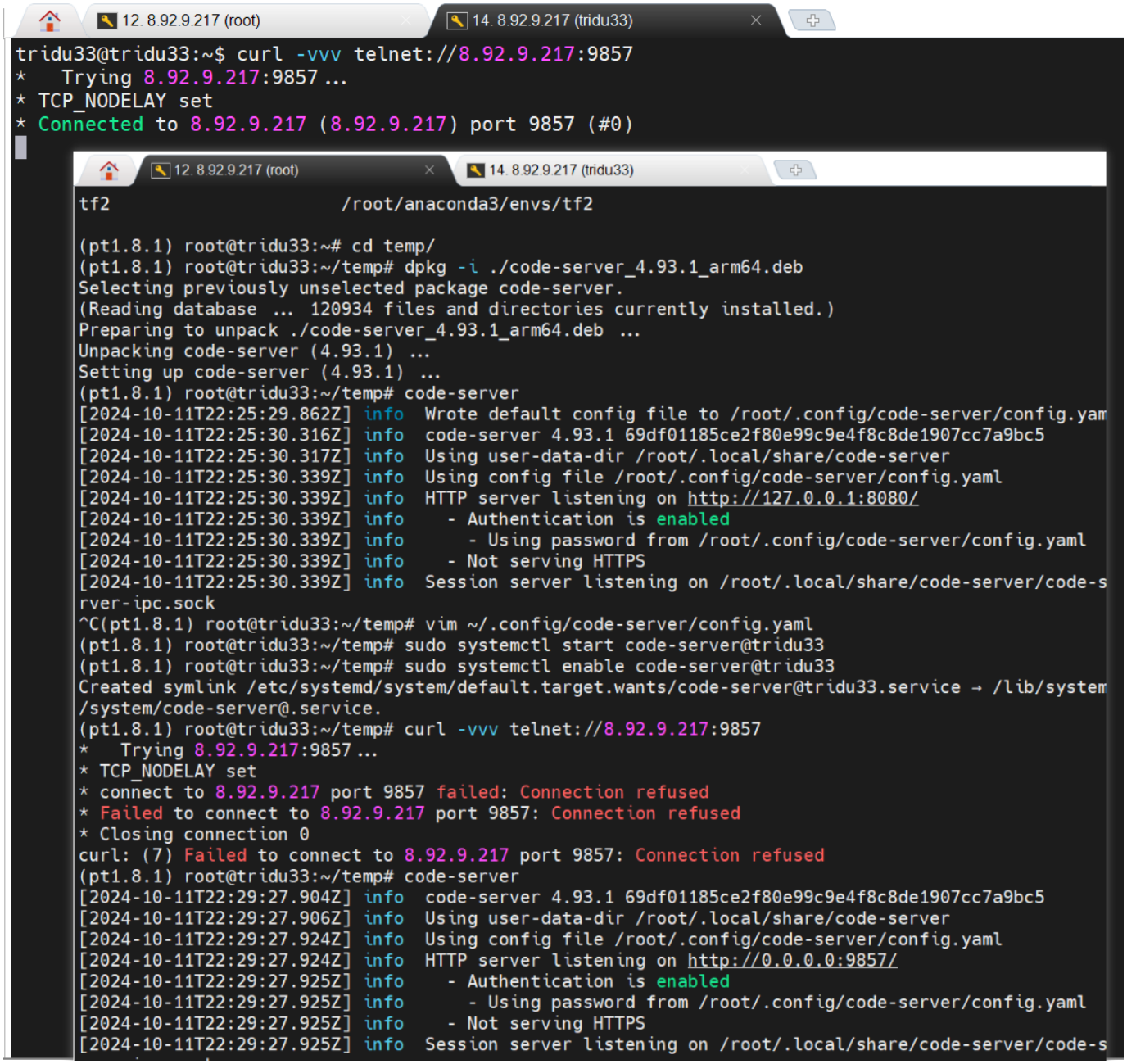


After turning on, you can check whether the IP of the network card is set correctly

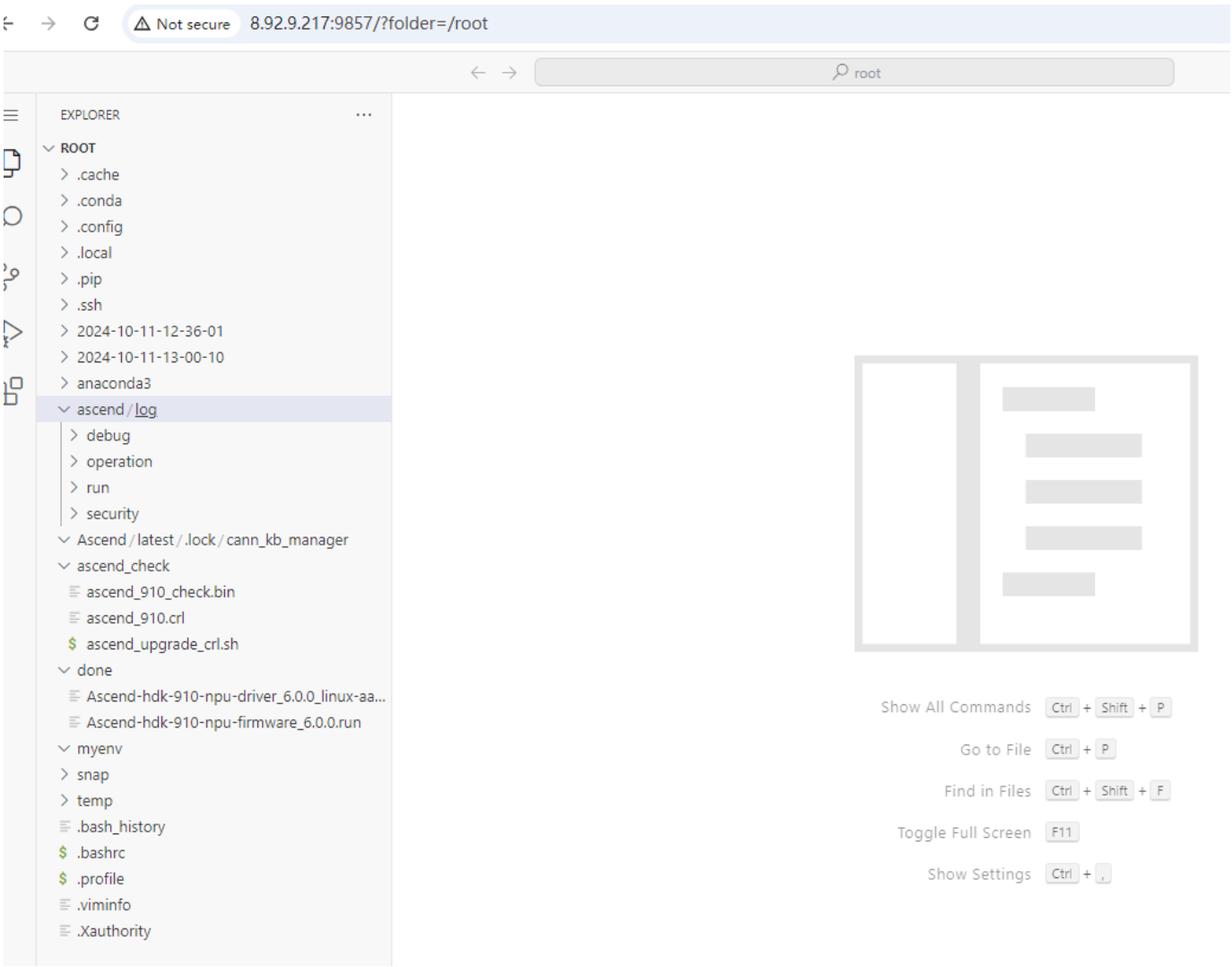
```
mseco@mseco:~$ sudo apt-get install -y net-tools pciutils
Reading package lists ... Done
Building dependency tree ... Done
Reading state information ... Done
net-tools is already the newest version (1.60+git20181103.0eebece-1ubuntu5).
pciutils is already the newest version (1:3.7.0-6).
0 upgraded, 0 newly installed, 0 to remove and 2 not upgraded.
mseco@mseco:~$ ifconfig
enp189s0f0: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 8.92.9.217  netmask 255.255.240.0  broadcast 8.92.15.255
        inet6 fe80::cac4:65ff:fe98:10a7  prefixlen 64  scopeid 0x20<link>
        ether c8:c4:65:98:10:a7  txqueuelen 1000  (Ethernet)
        RX packets 283584  bytes 531537732 (531.5 MB)
        RX errors 0  dropped 34674  overruns 0  frame 0
        TX packets 103291  bytes 7705040 (7.7 MB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

lo: flags=73<UP,LOOPBACK,RUNNING>  mtu 65536
        inet 127.0.0.1  netmask 255.0.0.0
        inet6 ::1  prefixlen 128  scopeid 0x10<host>
        loop  txqueuelen 1000  (Local Loopback)
        RX packets 363  bytes 36464 (36.4 KB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 363  bytes 36464 (36.4 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0
```

3) Remote development Try to install Code-Server, or use Pycharm remote development.



At this time, the local can be connected and developed remote Visual Code



Do not set the SystemCTL automatic background startup (it will automatically restart many background services), every time Mobaxterm is used in TMUX, manually enter local development.**Not recommended "Version Supporting Table" Os Outside OS Source Code Compile NPU Driver and firmware**.

https://github.com/mindspore-courses/step_into_llm