TITLE TEXT

SUBTITLE TEXT

Contributors

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

Table of contents

1. Home 3
1.1 走进LLM的殿堂 3
1.2 MindNLP 4
2. step_into_chatgpt 5
2.1 Transformer简介 5
2.2 BERT简介 6
3. step_into_llm 7
3.1 ChatGLM简介 7
4. others 8
4.1 Ascend物理机安装CANN和MindSpore环境指导 8
4.2 前置阅读和资源准备 8
4.3 iBMC安装操作系统和NPU驱动固件 8

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

1. Home

1.1 走进LLM的殿堂

项目 mindspore-courses/step_into_llm 的在线文档。

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

1.2 MindNLP

请访问 MindNLP 查看MindNLP完整API文档。

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

2. step_into_chatgpt

2.1 Transformer简介

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

2.2 BERT简介

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

3. step_into_llm

3.1 ChatGLM简介

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

4. others

4.1 Ascend物理机安装CANN和MindSpore环境指导

4.2 前置阅读和资源准备

1. 登录昇腾芯片的堡垒机或者物理机

   登录堡垒机需要联系管理员获取相应资源

1. 提前下载依赖包(版本对应关系和下载地址见下文)
2. Ubuntu 20.04 iso安装包
3. CANN及其kernel
4. Anaconda安装配套MindSpore环境

4.2.1 VPN打通堡垒机的网络

(1) . 连接昇腾社区的VPN即可访问杭州的堡垒机

根据昇腾生态众智实验室网络连接指导，导入配置和根据管理员给的资源连接第一层VPN。

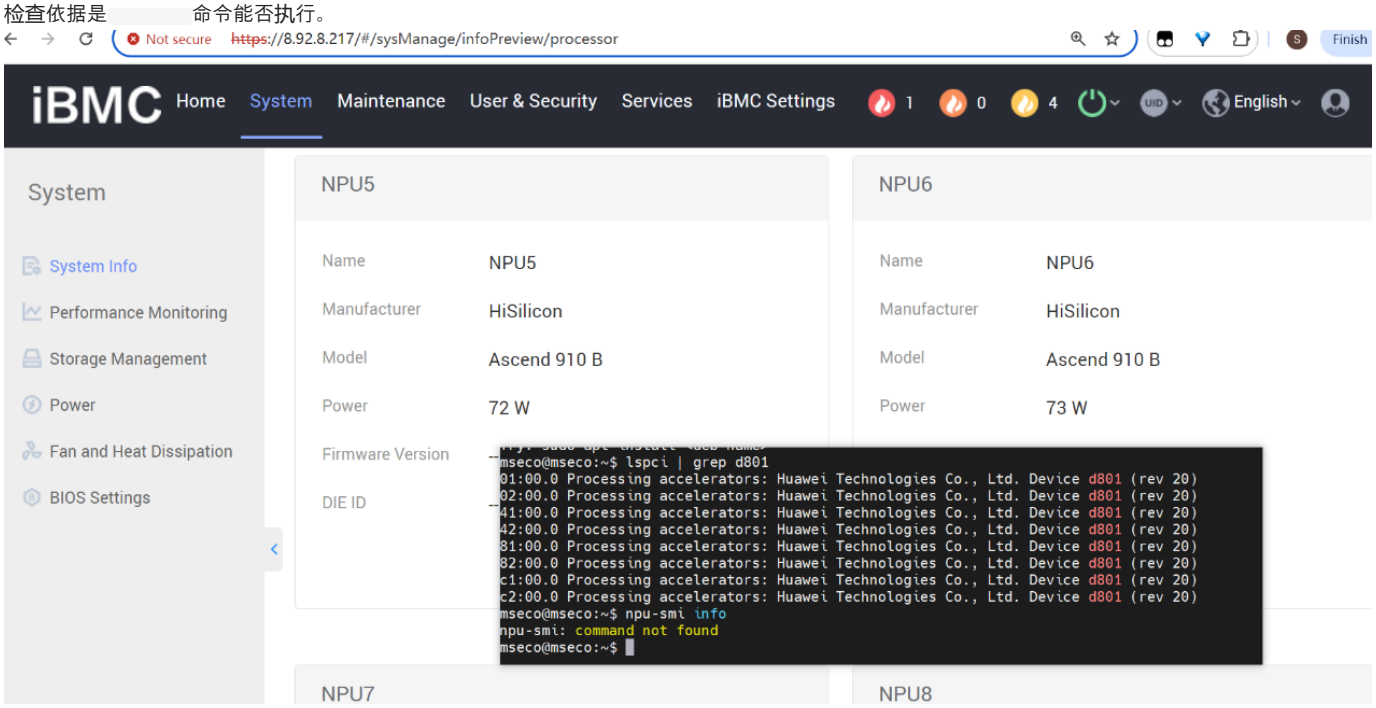
如果cmd能够ping 堡垒机IP 说明网络已打通，可以使用MobaXterm登录堡垒机。

(2) . 东莞-团泊洼的堡垒机还需连接第二层VPN

根据东莞-团泊洼堡垒机的网络环境VPN连接指引完成网络环境配置,即可登录堡垒机。

4.3 iBMC安装操作系统和NPU驱动固件

检查依据是 命令能否执行。



Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

尽量使用hdk团队已经调通的操作系统版本和对应的固件，详情查看官网文档《版本配套表》，《版本配套表》外不做保证，需要用户自行安装前置依赖，解决冲突并源码编译NPU驱动和固件。截止到2024.10.11，内核5.15.122-generic的Ubuntu22.04源码编译还有适配问题，其他系统对照关系类似。 按照Altas800(9000)官方教程安装文档，推荐使用《版本配套表》中NPU驱动固件对应的操作系统版本使用二进制安装，不推荐源码安装。

```
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 推荐Ubuntu20.04上配套NPU驱动和固件

安装Ubuntu20.04操作系统

大部分按照Ubuntu安装教程操作即可

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

```
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)
```

mobaXterm连接之后

安装NPU驱动和固件

iBMC可以直接查看NPU型号，

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

然后去昇腾页面寻找对应版本安装文档即可。

会自动升级内核版本，由于内核版本升级导致当前内核版本未适配驱动。因此可以关闭自动升级内核的功能。

查看有多少张

重启令驱动固件生效

```
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#
```

报错，根据教程https://bbs.huaweicloud.com/blogs/423686 设置白色这行

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

```
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"
```

然后重启

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

```
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         |
+=======================+===============+====================================================+
```

这样说明驱动没有问题，还可以继续检查下

```
```

4.3.2 配套CANN和kernel

安装必备依赖组件

```
```

安装完可以打印环境变量看看值是否正确写入。没有的话，需要自己写入一下到 或 ：

```
```

可以根据教程安装python3.7.5也可以安装Anaconda或者miniconda创建py37的环境，我这里使用Anaconda举例

```
```

配置华为源如下

```
```

可以安装这些前置工具。

```
```


Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

4.3.3 安装MindSpore

推荐使用Anaconda，也可以用Ubuntu原生pip+python环境安装MindSpore。

1. 验证Ubuntu、CANN、MindSpore版本配套关系
2. 检查MindSpore能否正常使用GPU,

下图是检查NPU是否正常使用的命令：

```
(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'参数有 ['CPU', 'GPU', 'Ascend','Davinci']这几种情况，Davinci是Ascend旧叫法。

4.3.4 Q&A

1) 第三方SSH客户端登录Linux实例时，提示"Access denied"错误可能的原因： - SSH登陆账号的密码输入错误；

- MacBook或者Windows键盘布局或者输入法的大小写和特殊字符转义导致密码错误，

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

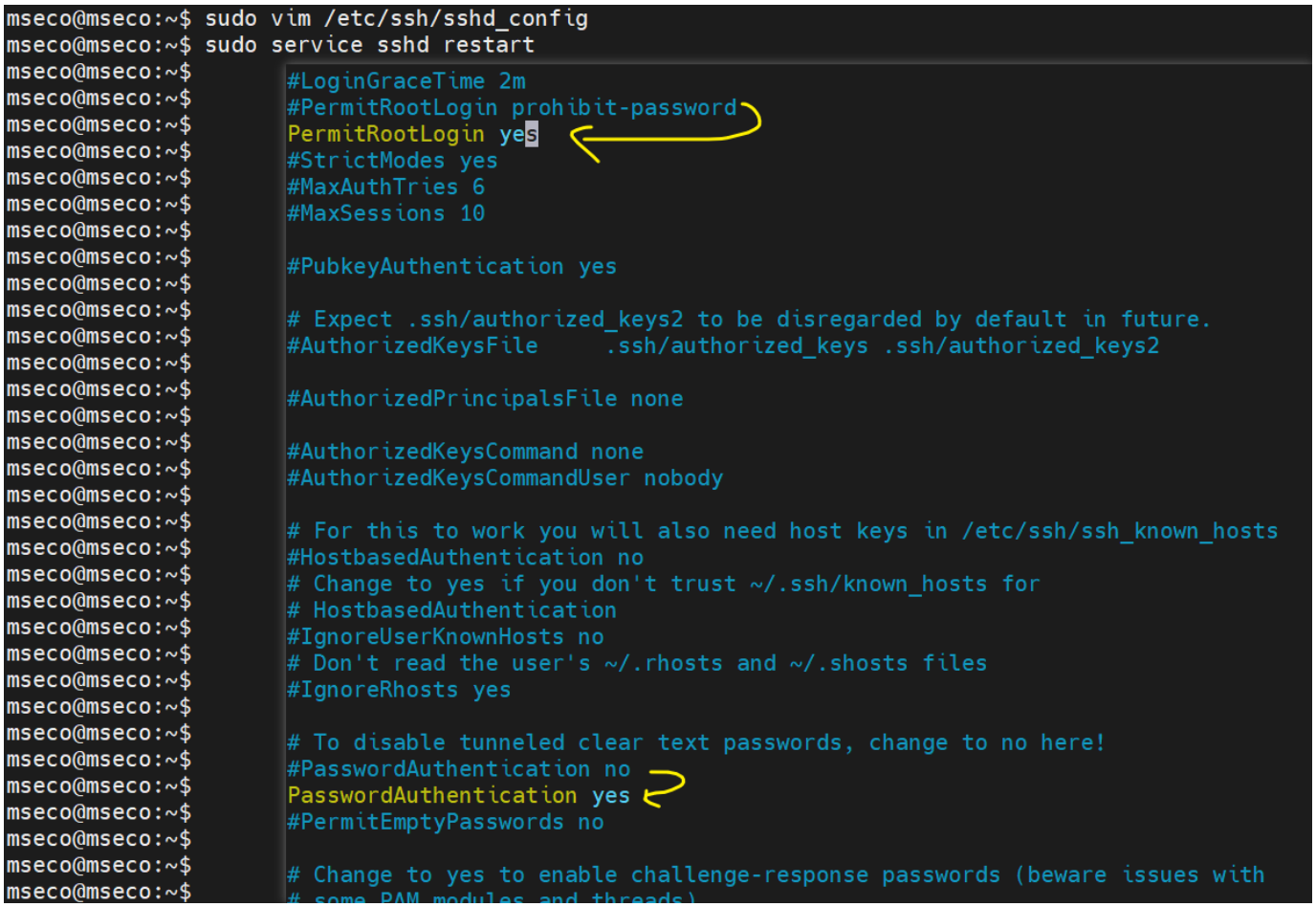
```
mseco@mseco:~$ cat /etc/passwd
root:x:0:0:root:/root:/bin/bash
daemon:x:1:1:daemon:/usr/sbin:/usr/sbin/nologin
bin:x:2:2:bin:/bin:/usr/sbin/nologin
sys:x:3:3:sys:/dev:/usr/sbin/nologin
sync:x:4:65534:sync:/bin:/bin/sync
games:x:5:60:games:/usr/games:/usr/sbin/nologin
man:x:6:12:man:/var/cache/man:/usr/sbin/nologin
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
```

密码有大写字母会出现Permission Denied密码错误，登录失败的情况

尝试 设置简单密码先试试能否登录；

- ssh_config配置没有正确设置

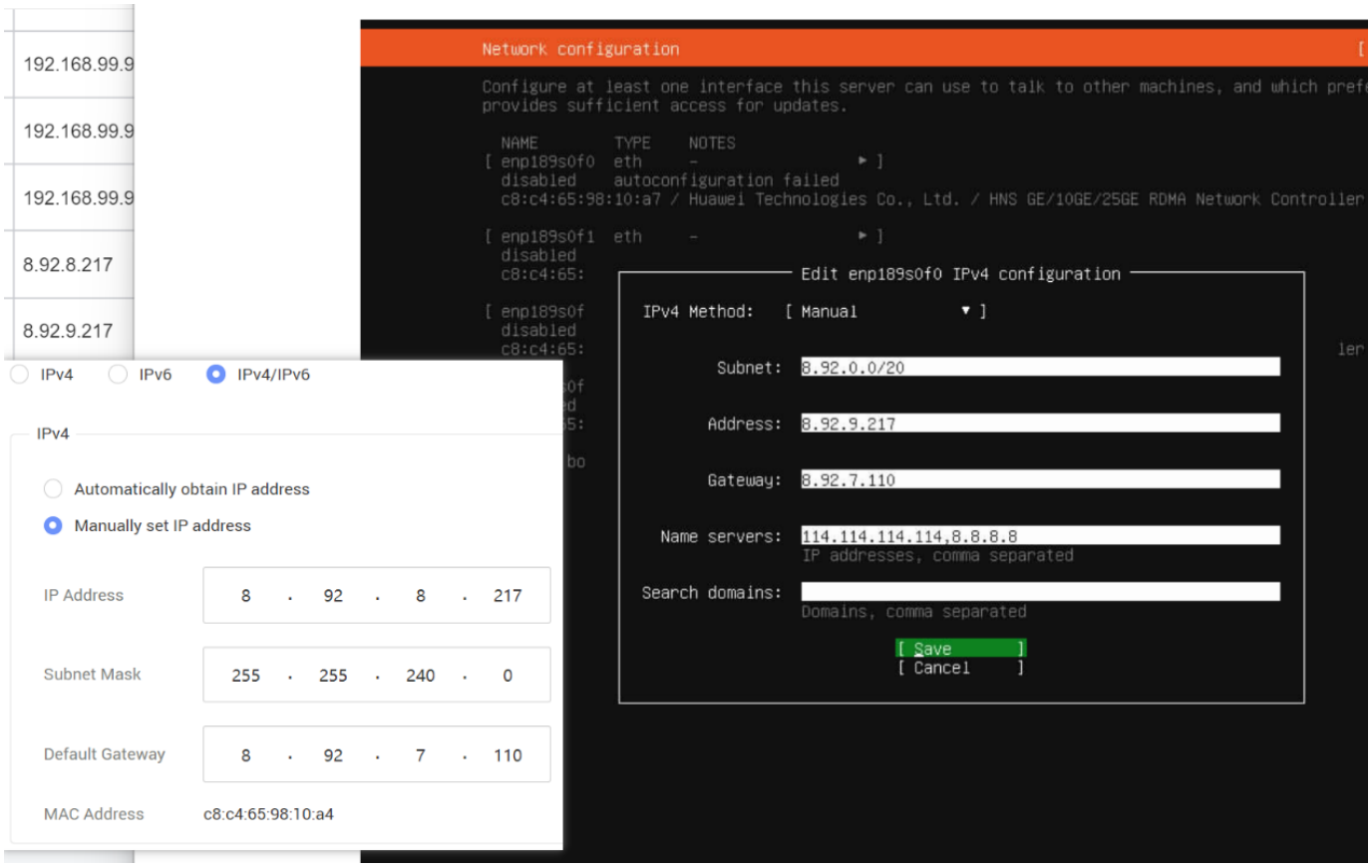
Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.



2) 安装Ubuntu之后能ping通iBMC IP无法ping通物理机ip

没有正确设置网络，推荐重装OS时参照iBMC网络配置，正确设置网卡：

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.



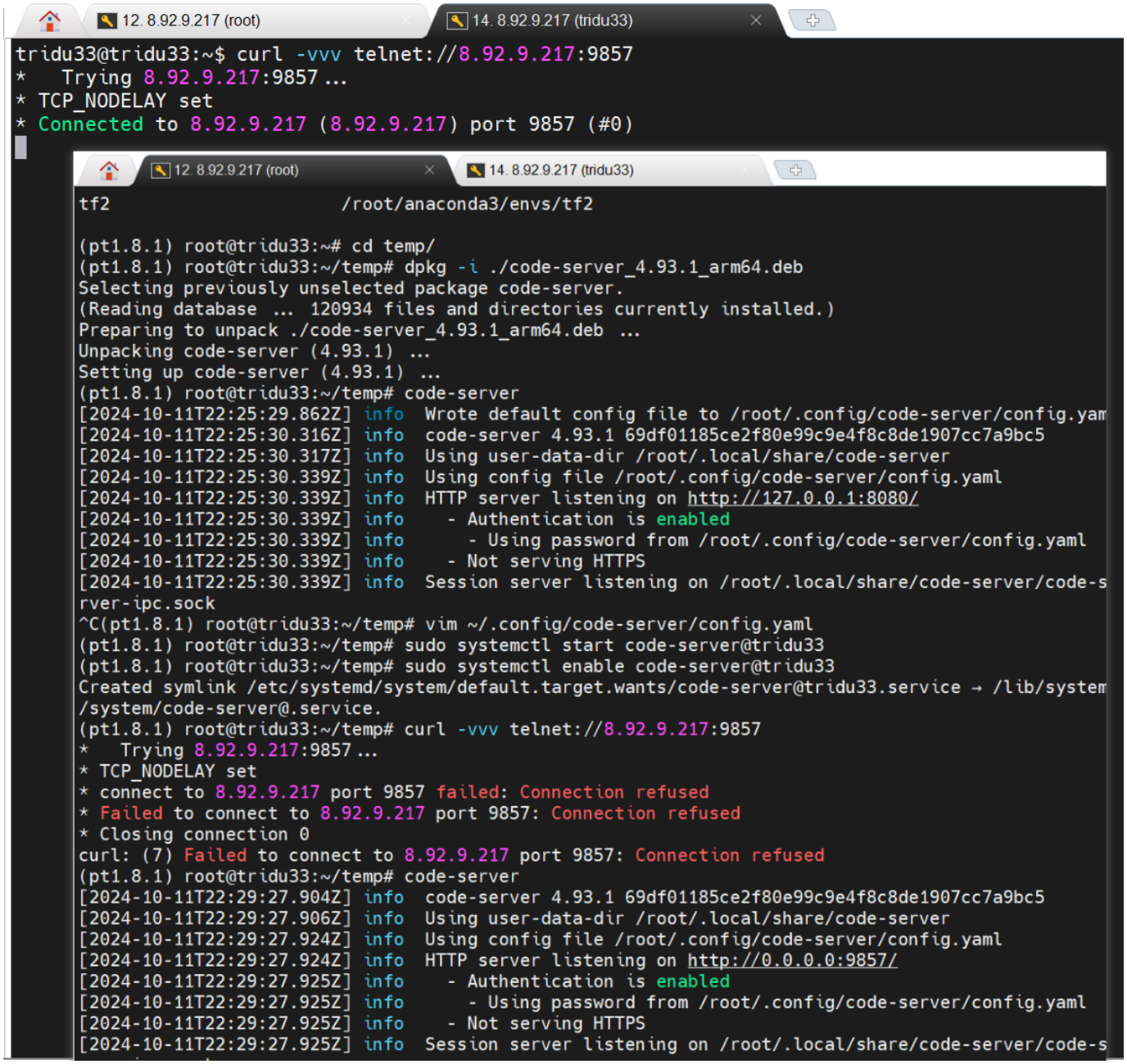
开机之后可以检查网卡的ip是否正确设置

```
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
```

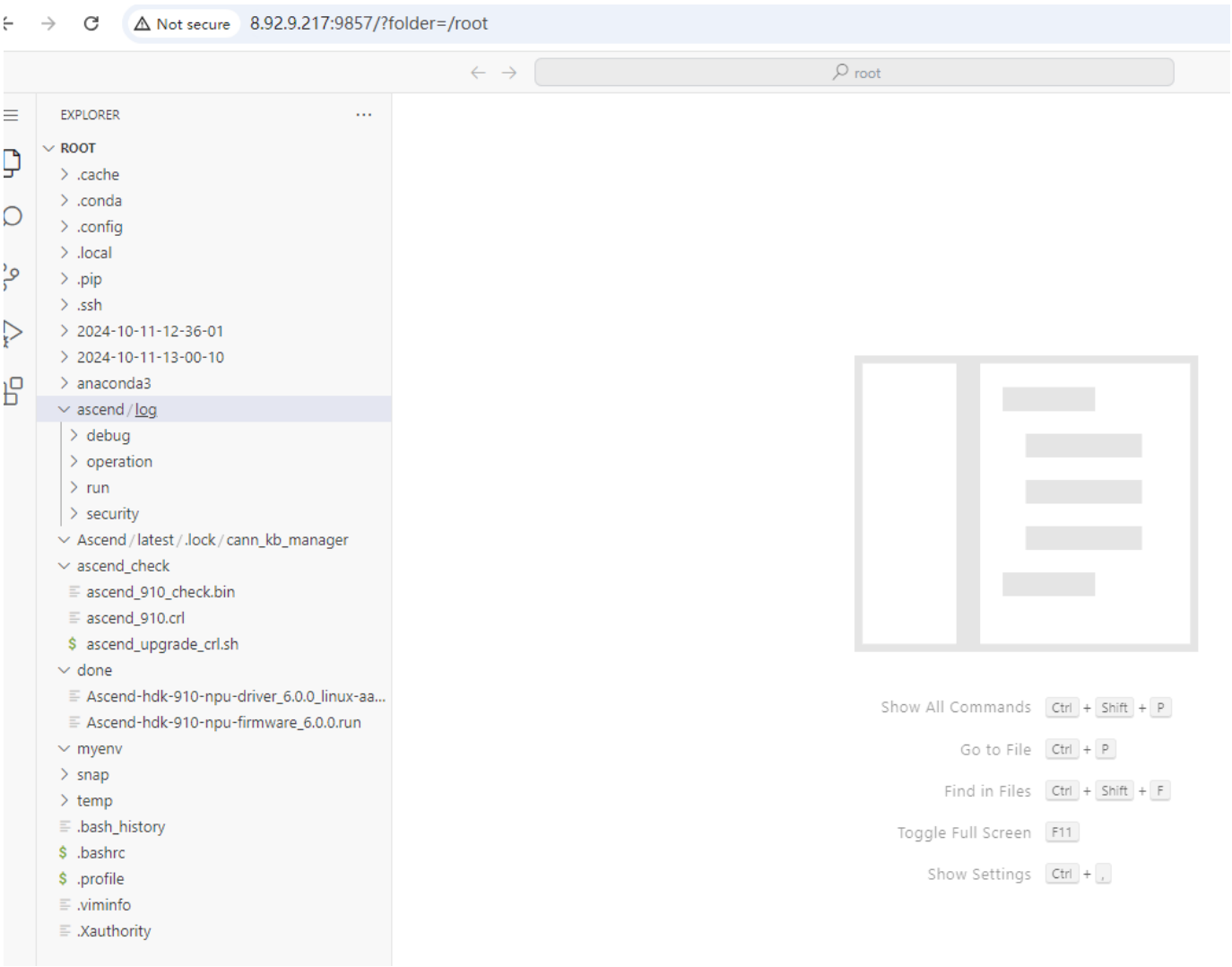

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

3) 远程开发答疑 可以安装code-server或者使用Pycharm远程开发。



此时本地就能远程visual code连接并开发

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.



不要设置systemctl自动后台启动（会自动重启很多个后台服务），每次用的时候mobaxterm在tmux会话手动输入 本地开发即可。**不推荐《版本配套表》外OS上源码编译NPU驱动和固件**.

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.

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

Licensed under Creative Commons Attribution-NonCommercial 4.0 International License.