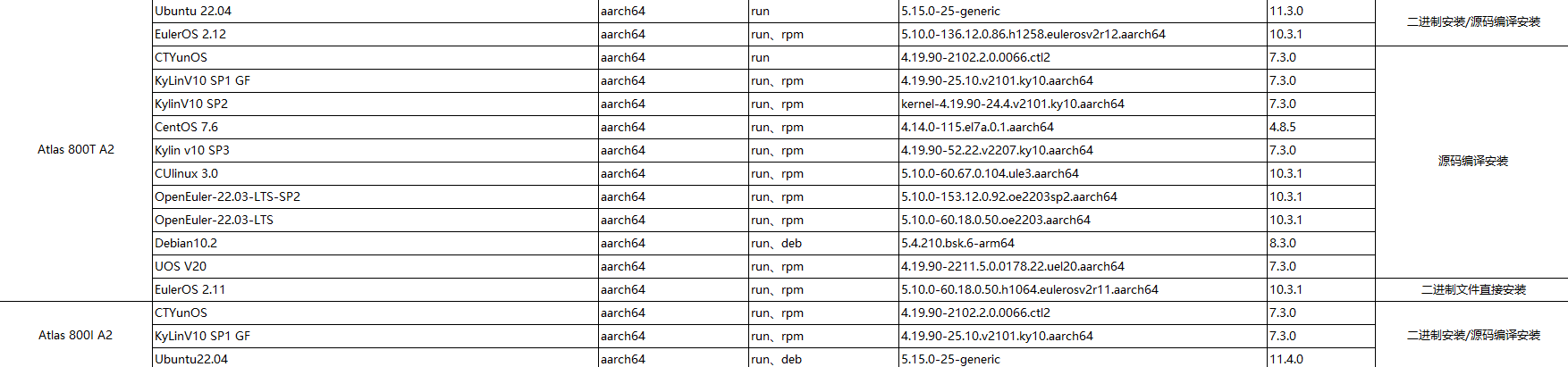
800I/T A2服务器环境配置与服务部署

# 操作系统安装和准备

参考[Ascend HDK 24.1.RC2 版本配套表.xlsx](https://support.huawei.com/enterprise/zh/doc/EDOC1100388239) 中的**产品操作系统内核版本要求**安装服务器对应操作系统，注意内核版本对齐！！！

**uname -m && cat /etc/\*release 查询操作系统架构及版本**

如果查询的操作系统版本不在兼容性列表中，请替换为兼容性列表中支持的操作系统。



# 安装系统依赖

确保环境联网，或者能够apt或yum直接下载依赖包

先apt-get update或yum makecache

以ubuntu22.04为例：

|  |
| --- |
| apt-get install gcc g++ make cmake zlib1g zlib1g-dev openssl libsqlite3-dev libssl-dev libffi-dev libbz2-dev libxslt1-dev unzippciutils net-tools libblas-dev gfortran libblas3 |

以openEuler22.03为例：

|  |
| --- |
| yum install -y gcc gcc-c++ make cmakeunzipzlib-devel libffi-devel openssl-devel pciutils net-tools sqlite-devel lapack-devel gcc-gfortran python3-devel |

根据不同操作系统自行下载安装依赖：gcc g++ make cmake等

# 安装驱动和固件：

1、以root用户登录服务器

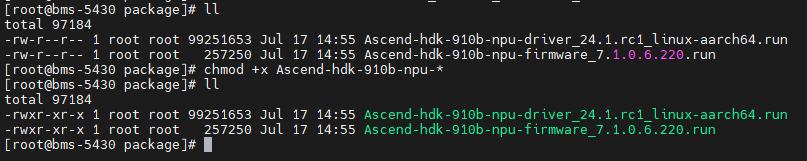
2、从华为support网站下载驱动和固件包（超链接）

[Ascend-hdk-910b-npu-driver\_24.1.rc2\_linux-aarch64.run](https://support.huawei.com/enterprise/zh/software/262029946-ESW2001108426)

[Ascend-hdk-910b-npu-firmware\_7.3.0.1.231.run](https://support.huawei.com/enterprise/zh/software/262029946-ESW2001108439)

并上传到服务器

3、chmod +x Ascend-hdk-\* 修改可执行权限



4、创建运行用户HwHiAiUser

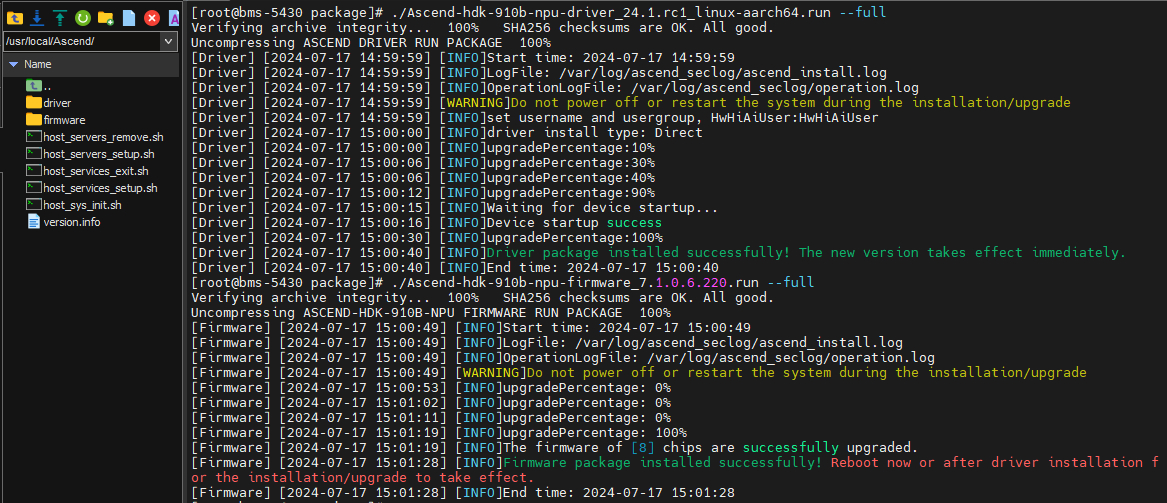
|  |
| --- |
| groupadd HwHiAiUser  useradd -g HwHiAiUser -d /home/HwHiAiUser -m HwHiAiUser -s /bin/bash |



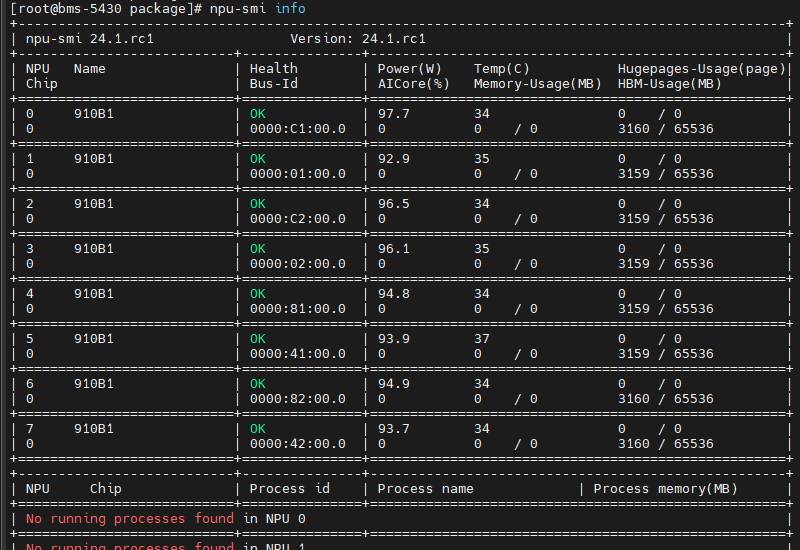
5、安装驱动和固件：

|  |
| --- |
| ./Ascend-hdk-xxx-npu-driver\_x.x.x\_linux-{arch}.run –full  ./Ascend-hdk-xxx-npu-firmware\_x.x.x.x.X.run –full |

（先装驱动后装固件）



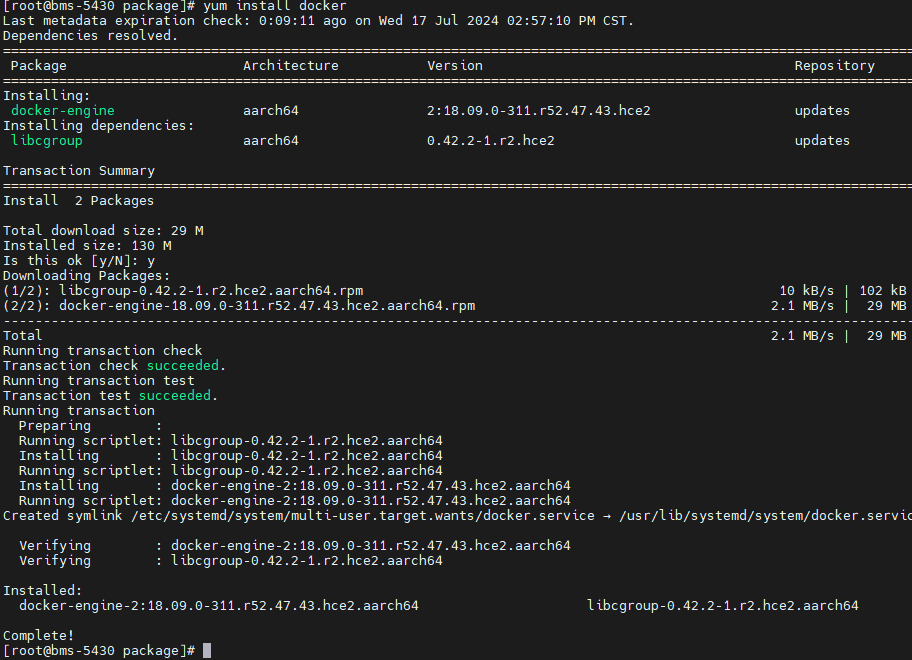
6、npu-smi info命令验证是否安装成功



# 安装docker和ascend docker runtime容器引擎插件

1、安装docker：apt或yum install docker ##Docker版本需要为18.03及以上版本！！

确保docker能够正常使用和启动普通镜像：例如docker run -it ubuntu20.04:latest



2、下载ascend docker runtime容器引擎插件：

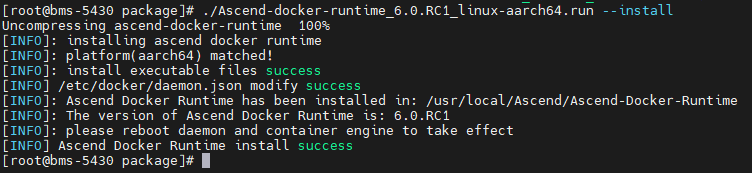
<https://gitee.com/ascend/ascend-docker-runtime/releases/tag/v6.0.0-RC1>

3、修改可执行权限：chmod +x Ascend-docker-runtime\_{version}\_linux-{arch}.run



4、安装：./Ascend-docker-runtime\_{version}\_linux-{arch}.run –install

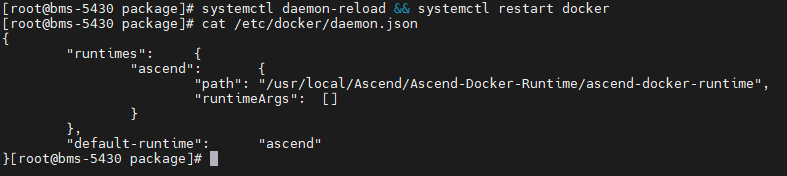
回显示例如下，表示安装成功。



5、执行以下命令，使Ascend Docker Runtime生效！！！

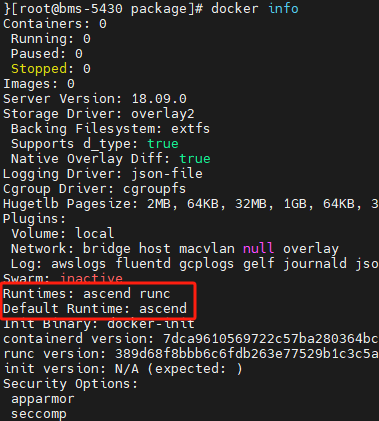
|  |
| --- |
| systemctl daemon-reload && systemctl restart docker |

6、确认组件状态：Ascend Docker Runtime安装后，会修改配置文件daemon.json中的代码，不需要手动修改配置！！！

执行docker info查看默认Runtime是否为ascend：

Runtimes: **ascend** runc

**Default Runtime: ascend**



# mindie镜像部署

1、下载mindie镜像并启动mindie容器：

mindieT59版本镜像获取：

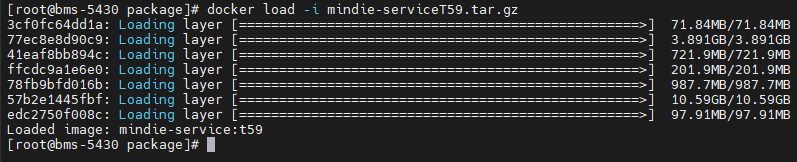
<https://driver.obs.cn-east-292.mygaoxinai.com/24.1.rc2/T59/mindie-serviceT59.tar.gz> （浏览器打开下载）

mindieRC2版本镜像获取：（推荐）

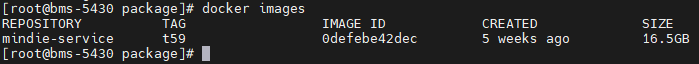
<https://driver.obs.cn-east-292.mygaoxinai.com/24.1.rc2/RC2/mindieRC2-910B.tar.gz>（浏览器打开下载）

<https://driver.obs.cn-east-292.mygaoxinai.com/24.1.rc2/RC2/mindieRC2-310P.tar.gz>

2、将压缩包上传到服务器，使用docker load -i mindie-serviceT59.tar.gz 命令导入镜像



3、docker images查看镜像是否被导入

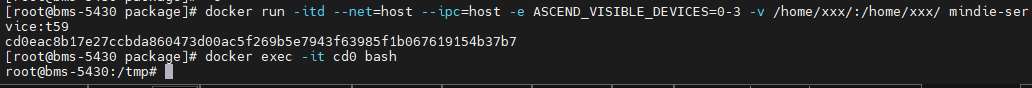


4、启动mindie镜像并挂载指定芯片

（1）如果你的环境中已经安装了ascend docker runtime，可以使用简化版启动命令：

|  |
| --- |
| docker run -itd --net=host --ipc=host --security-opt --seccomp=unconfined \  -e ASCEND\_VISIBLE\_DEVICES=0-3 \  -v /home/xxx/:/home/xxx/ \  mindie-service:t59 |

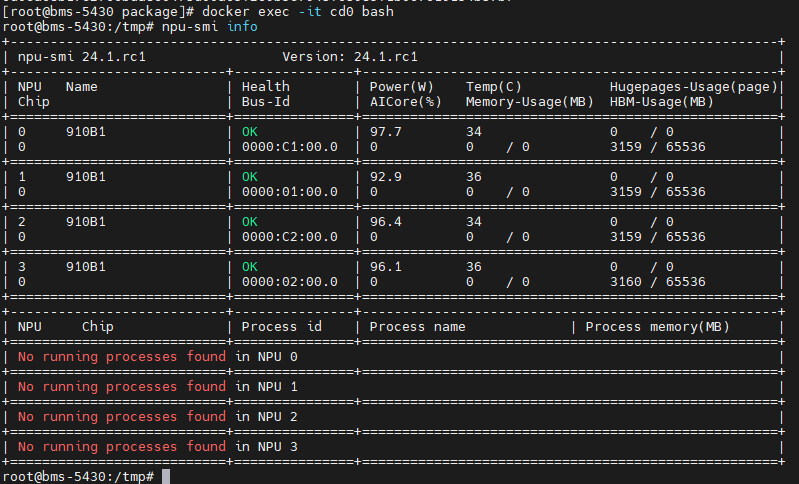
其中--seccomp=unconfined按需添加，也可改成--privileged=true



（2）第二种容器启动方法：

|  |
| --- |
| docker run -itd --name=mindie\_server --net=host --ipc=host \  --device=/dev/davinci0 \  --device=/dev/davinci1 \  --device=/dev/davinci2 \  --device=/dev/davinci3 \  --device=/dev/davinci4 \  --device=/dev/davinci5 \  --device=/dev/davinci6 \  --device=/dev/davinci7 \  --device=/dev/davinci\_manager \  --device=/dev/devmm\_svm \  --device=/dev/hisi\_hdc \  -v /usr/local/Ascend/driver:/usr/local/Ascend/driver \  -v /usr/local/Ascend/add-ons/:/usr/local/Ascend/add-ons/ \  -v /usr/local/sbin/:/usr/local/sbin/ \  -v /var/log/npu/slog/:/var/log/npu/slog \  -v /var/log/npu/profiling/:/var/log/npu/profiling \  -v /var/log/npu/dump/:/var/log/npu/dump \  -v /var/log/npu/:/usr/slog \  -v /etc/hccn.conf:/etc/hccn.conf \  -v /home/xxxx/:/home/xxxx/ \  mindie-service:t59 \  /bin/bash |

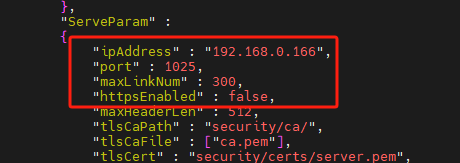
5、进入容器查看挂载芯片是否成功：npu-smi info



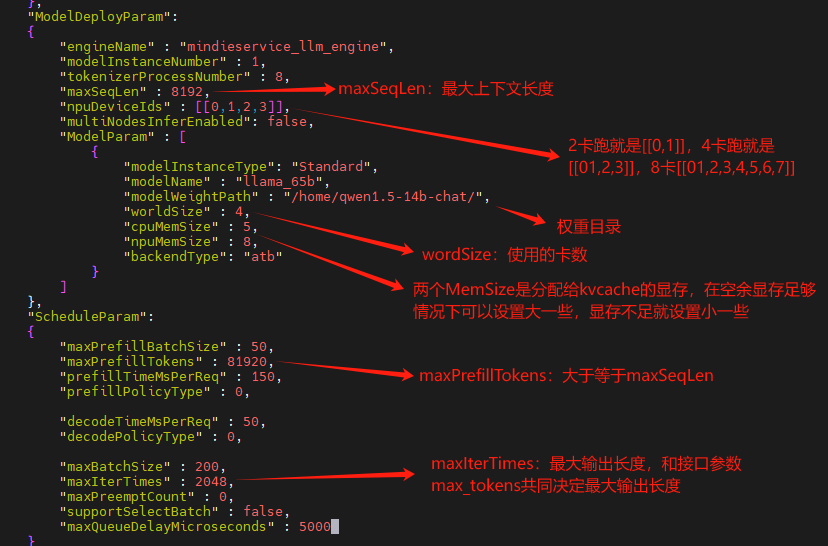
5 修改配置并启动服务：

|  |
| --- |
| /usr/local/Ascend/mindie/latest/mindie-service/  vim conf/config.json |

修改IP和端口



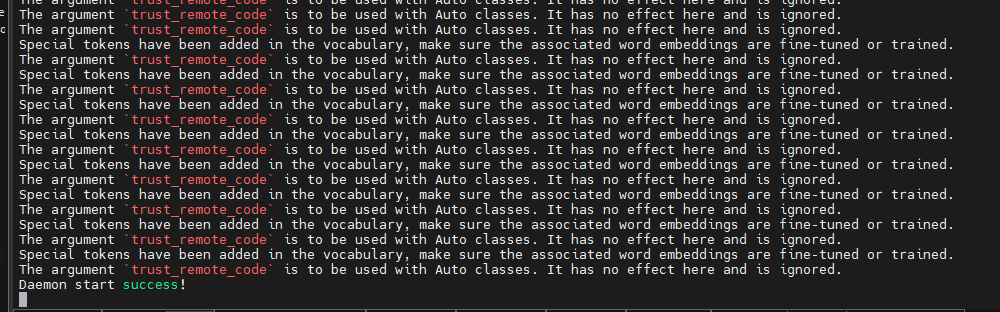
修改模型路径和最大上下文长度



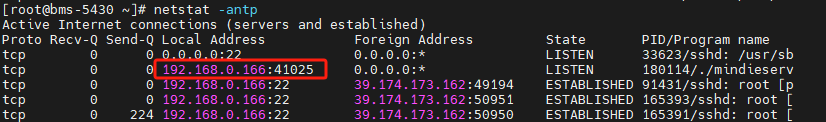
启动服务

进入/usr/local/Ascend/mindie/latest/mindie-service/bin

执行./mindieservice\_daemon

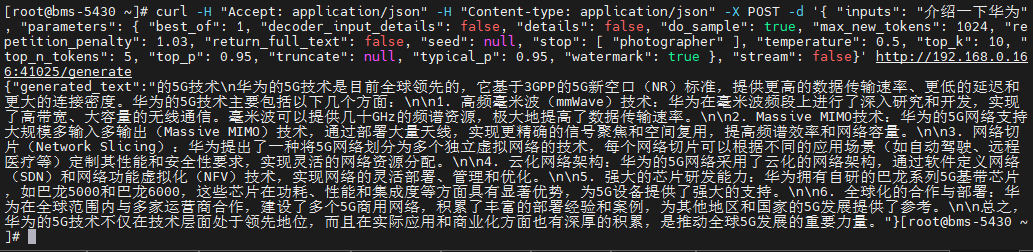


netstat -antp查看端口是否正常



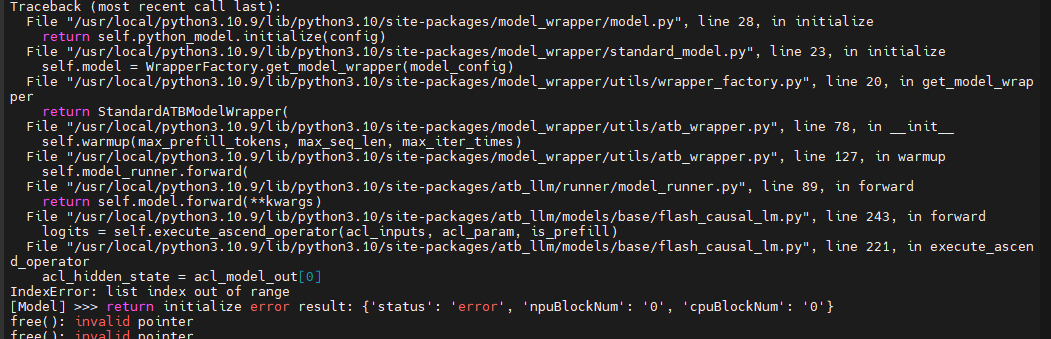
curl测试

|  |
| --- |
| curl -H "Accept: application/json" -H "Content-type: application/json" -X POST -d '{ "inputs": "介绍一下华为", "parameters": { "best\_of": 1, "decoder\_input\_details": false, "details": false, "do\_sample": true, "max\_new\_tokens": 1024, "repetition\_penalty": 1.03, "return\_full\_text": false, "seed": null, "stop": [ "photographer" ], "temperature": 0.5, "top\_k": 10, "top\_n\_tokens": 5, "top\_p": 0.95, "truncate": null, "typical\_p": 0.95, "watermark": true }, "stream": false}' http://192.168.0.166:41025/generate |



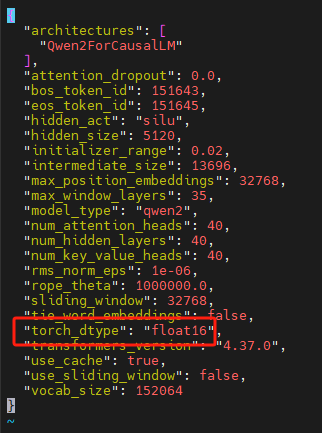
# 常见报错：

1、list index out of range列表索引超出范围

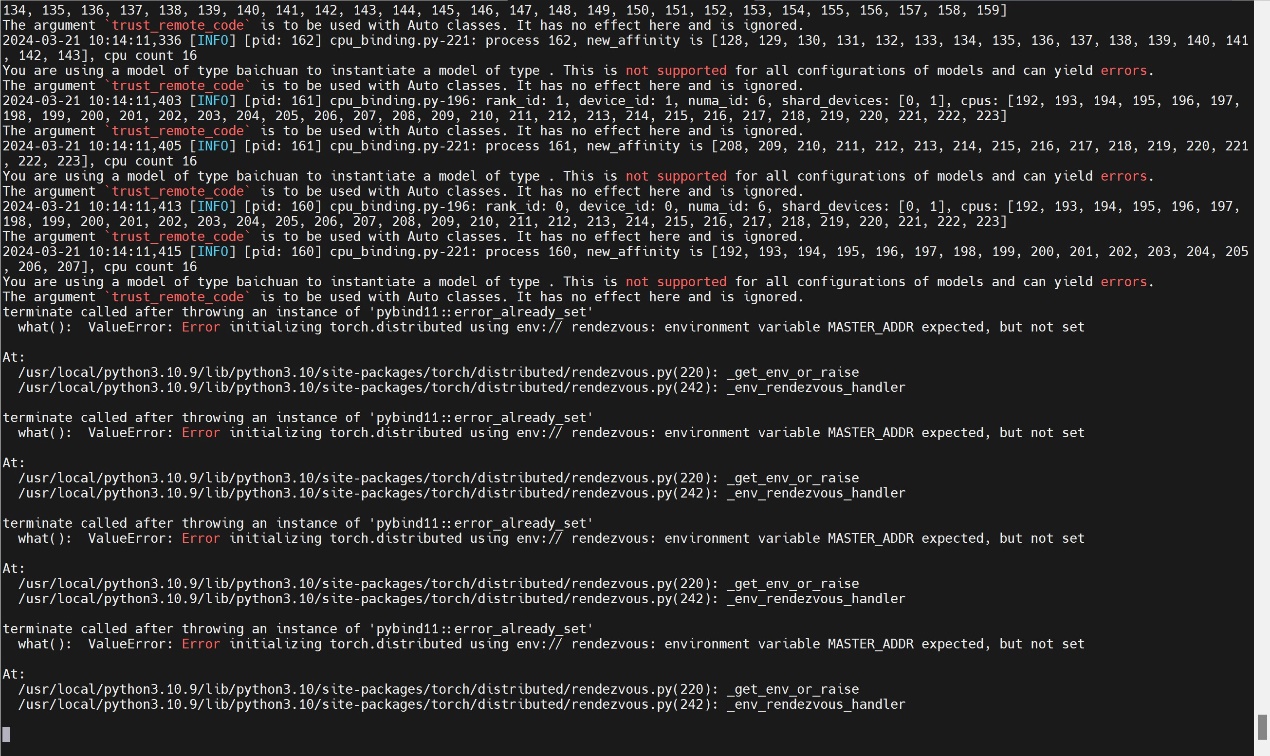


解决方法，修改权重目录下的config.json文件

将torch\_dtype参数值改为float16



2、未设置MASTER\_ADDR

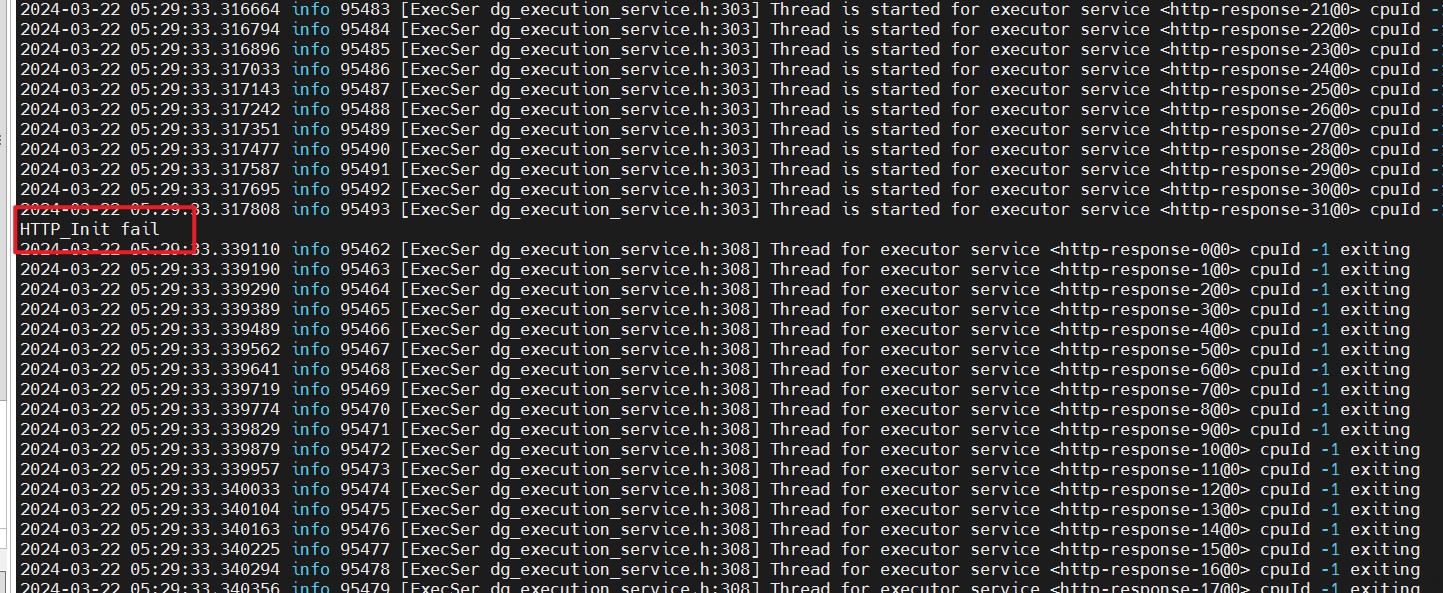


解决方案：设置以下环境变量

export MASTER\_ADDR=127.0.0.1

export MASTER\_PORT=9090

3、http初始化失败



解决方案：可能是端口冲突，设置环境变量 MASTER\_PORT为别的端口