



天数智芯
Iluvatar CoreX

天数智芯

天数智芯加速卡常见问题分析手册

版本：SDK V1.0 及以上

日期：2024.11.18

适用产品：天垓 100 | 天垓 150 | 智铠 50 | 智铠 100

1 声明

1.1 版权声明

版权所有。未经天数智芯书面许可，不得以任何形式或方式将本文档的任何部分复制，传播，转录或翻译成任何语言。

1.2 免责声明

天数智芯可以随时对本文档或本文档中描述的产品进行改进和/或更改。本文档包括与天数智芯产品有关的信息，作为说明典型应用的一种方式，因此，不一定提供足以进行生产设计的完整信息。对于本文档中内容的准确性或完整性，天数智芯不做任何陈述或保证。

1.3 联系方式

地址：上海市闵行区陈行公路 2168 号 3 幢

电话：021-68886607

网址：www.iluvatar.com

Contents

1 声明	2
1.1 版权声明	2
1.2 免责声明	2
1.3 联系方式	2
2 天数智芯加速卡常见问题分析手册	4
2.1 修订记录	4
2.2 硬件状态相关	4
2.2.1 问题 1: 如何连接电源线?	4
2.2.2 问题 2: 如何确认加速卡已经正确安装?	7
2.2.3 问题 3: 如何监控加速卡状态?	9
2.2.4 问题 4: 如何解决掉卡问题?	9
2.2.5 问题 5: 为什么带宽测试结果偏低?	10
2.3 固件相关	10
2.3.1 问题 1: 升级固件后, 固件版本号未发生变化, 且重启后仍然存在该问题	11
2.4 软件栈安装相关	11
2.4.1 问题 1: 无法安装天数智算软件栈驱动	11
2.4.2 问题 2: 无法调用天数智算软件栈相关命令	11
2.4.3 问题 3: 生成 Docker 镜像后启动容器时, 如何指定设备?	12
2.4.4 问题 4: 哪些 Python 包必须安装天数智芯适配版本? 为什么?	12
2.4.5 问题 5: 为什么不能直接使用 pip install torch ?	13
2.4.6 问题 6: 如果已有代码指定的 PyTorch 版本不是天数智芯支持的版本, 怎么处理?	13
2.4.7 问题 7: 对于不支持的大模型推理引擎, 我是否可以自己移植?	13
2.4.8 问题 8: 机器无法访问网络, 无法 pip 在线安装	14
2.4.9 问题 9: 使用 corex-docker-installer 的.run 包生成包含天数智算软件栈的 Docker 镜像失败, 遇到 error code:100 报错	14
2.5 错误码	14
3 商标声明	18

2 天数智芯加速卡常见问题分析手册

本文档旨在为用户提供常见的天数智芯加速卡使用相关问题，并提供问题根因分析和对应的解决方法。

2.1 修订记录

- COREX-GPUFAQ-RF01-03: 2024/11/18
 - 更新**错误码**，修改对错误代码 11 的说明
- COREX-GPUFAQ-RF01-02: 2024/11/8
 - 更新**问题 1: 如何连接电源线?**，补充对天数智芯加速卡电源线的重要事项说明
- COREX-GPUFAQ-RF01-01: 2024/8/26
 - 更新**问题 1: 如何连接电源线?**，新增天垓 150S 型号
 - 新增**问题 1: 升级固件后，固件版本号未发生变化，且重启后仍然存在该问题**及解决方法
 - 更新**问题 2: 无法调用天数智算软件栈相关命令**，提供使环境变量永久生效的方法
 - 更新**问题 2: 如何确认加速卡已经正确安装?**，提供更多可能出现的问题及解决方法
 - **问题 4: 如何解决掉卡问题?**，更新查询固件版本命令
 - 更新**问题 4: 哪些 Python 包必须安装天数智芯适配版本? 为什么?**，建议用户先安装天数智芯适配版本的 Python 包
 - 新增**问题 5: 为什么带宽测试结果偏低?**及解决方法
 - 新增**问题 8: 机器无法访问网络，无法 pip 在线安装**及解决方法
 - 新增**问题 9: 使用 corex-docker-installer 的.run 包生成包含天数智算软件栈的 Docker 镜像失败，遇到 error code:100 报错**及解决方法
 - 更新**错误码**
- COREX-GPUFAQ-RF01-00: 2024/6/30

文档首次对外发布

2.2 硬件状态相关

2.2.1 问题 1: 如何连接电源线?

问题描述

我拿到天数智芯的加速卡后不知道电源线的种类和使用方法。

问题分析

天数智芯加速卡有天垓 (天垓 100、天垓 150 和天垓 150S) 和智铠 (智铠 100 和智铠 50) 两个产品线，对应的接口也有区别，详情如下表所示：

加速卡	电源线接口要求
天垓 100	ATX 8-pin CPU type
天垓 150/天垓 150S	ATX 8-pin CPU type
智铠 50	无需电源线
智铠 100	ATX 8-pin PCIe type

具体区别如下图所示。

天垓加速卡服务器主板电源接口：

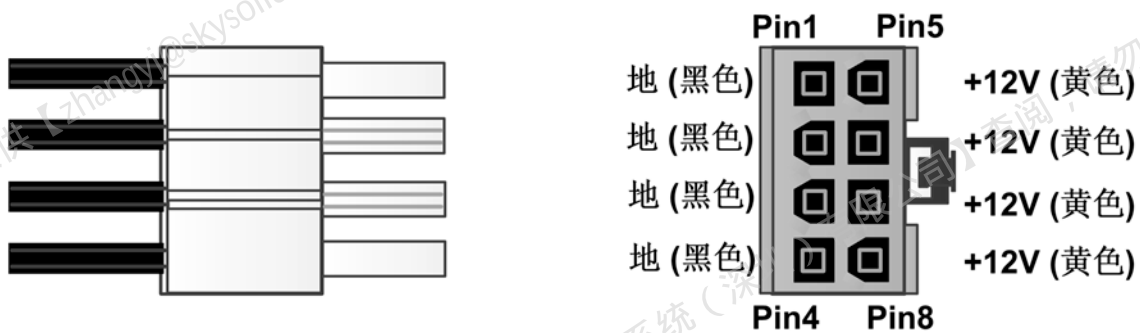


Figure 1: 天垓加速卡服务器主板电源接口示意图

智铠 100 加速卡服务器主板电源接口：

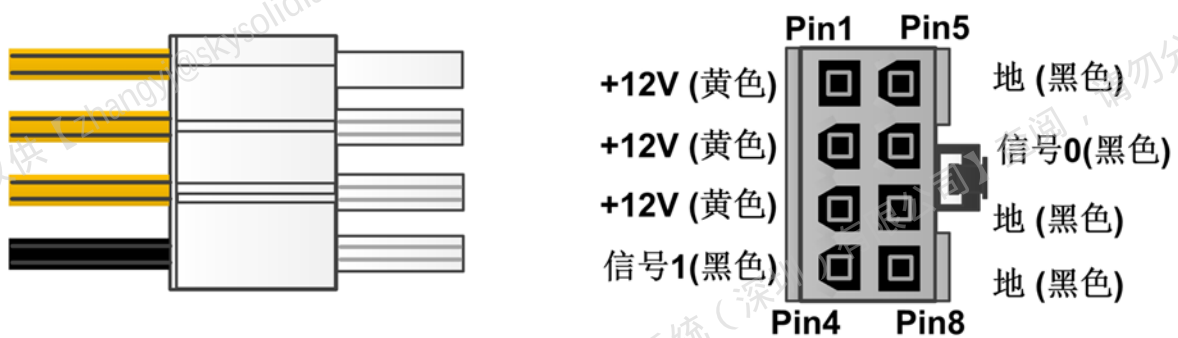


Figure 2: 智铠 100 加速卡服务器主板电源接口示意图

解决方法

请访问 [天数智芯官网文档中心](#) 并下载产品对应的《硬件说明及安装指南》了解详情；也可以通过扫描下方二维码观看安装视频文件(天垓 100 加速卡)，了解安装过程。其它加速卡，请联系天数智芯应用工程师获取支持。



Figure 3: 天垓 100 加速卡安装视频

另外，各家服务器厂商的加速卡电源线接口会有不同的定义。详情请联系服务器原厂并获得加速卡电源线的购买支持。

Important

- 请尽可能使用服务器厂商提供的原厂电源线，以确保电源线质量和与服务器的匹配度。
- 电源线接口必须为天数智芯加速卡指定接口，具体可查阅各加速卡的《硬件说明及安装指南》。
- **重要提示：**原厂线常见有 1 转 1，1 转 2 等形式；**对于 1 转 2 电源线，如果同时接 2 张加速卡，需要注意计算同时供电是否充足。**
- 如果无原厂电源线或原厂电源线接口不匹配天数智芯加速卡要求，才考虑使用转接线。
 - 测量转接线接口类型，请使用万用表实测 pin 脚电压，请勿根据接口形状判断电源线类型。
 - 确认转接线最大载流量可为天数智芯加速卡提供充足供电。
 - 确保转接线质量合格。
- 如果使用转接线，需要注意接线情况：
 - 观察捆绑线束的胶条，注意以下标注，请勿插错两端：
 - * 标注 RISER 端用以接服务器

* 标注 GPU 端用以接加速卡

- 观察接口形状，每个 pin 脚需要公头和母头对应，请勿翻转插错。

注意事项

- 请确保服务器供电、电源线接口和电源线线缆能够支持天数智芯加速卡要求的规格。
- 如选择不匹配的服务器和电源线，可能造成服务器主板、卡槽、加速卡、电源线的损坏。
- 请确保服务器供电满足天数智芯加速卡的功耗要求。

服务器对加速卡的供电由以下两部分组成，每部分的实际功耗根据加速卡负荷动态调整：

- 服务器连接加速卡的供电线 (cable)
- PCIe 插槽供电

Tip

推荐：cable 对每张加速卡的供电需要大于加速卡的 TDP 功耗，即不考虑 PCIe 插槽供电 (最多 75W 来自 PCIe 插槽，实际根据加速卡负荷动态调整，不能以固定值预测)，也能确保加速卡的供电充足。

Important

- 供电不足可能会损坏服务器主板、卡槽、加速卡、电源线。
- 建议：优先选择具有过流保护，且过流保护的电流设置合理的硬件，过大的电流设置可能无法起到过流保护效果。
- 在连接电源线之前，请先检查如下项目：
 - 天数智芯加速卡：了解天数智芯加速卡的功耗需求，可查看产品规格书，如天垓 150 TDP 功耗为 350W，天垓 150S TDP 功耗为 450W。

Important

重要提示：注意加速卡迁移情况。例如，服务器已适配天垓 100 的功耗要求 (250W)，从天垓 100 迁移到天垓 150，还需要重新计算是否满足天垓 100/150S (350W/450W) 的功耗要求。

- 天数智芯加速卡供电线规格：电源线的供电规格与 pin 脚电压、最大电流、供电 pin 脚数量有关。例如，电源线规格 12V*9A*4pin，则电源线可提供 432W 功率。
- 天数智芯加速卡供电线形态：原厂线常见有 1 转 1，1 转 2 等形式；**对于 1 转 2 电源线，如果同时接 2 张加速卡，需要注意计算同时供电是否充足。**

2.2.2 问题 2：如何确认加速卡已经正确安装？

问题描述

我需要知道天数智芯加速卡是否能正常工作。

问题分析

从硬件方面，可以通过卡端末尾的电源指示灯来判断加速卡是否连接正常。电源未接入或者电源接入异常，指示灯将显示红色；正常工作的加速卡，指示灯显示绿色。

从系统方面，可以通过 BIOS 设置、加速卡 I/O 访问能力、Memory 访问能力等方面进行问题排查。

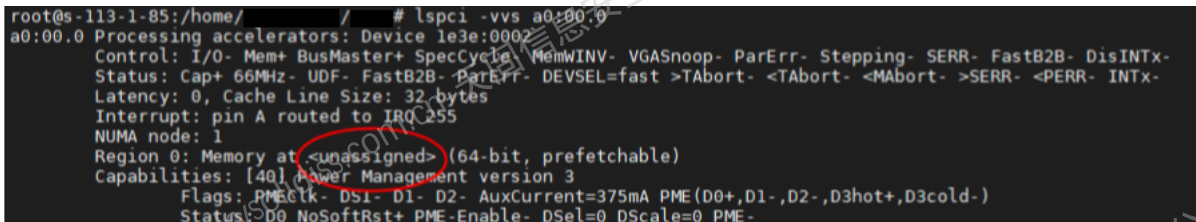
从软件方面，安装好软件栈后，可以通过 ixSMI 工具来查询设备的详情并观察设备的工作情况。

解决方法

下面的解决方法将从系统方面指导用户进行逐一排查。

1. 请先确认主机 BIOS 设置中 **Above 4G Decoding** 选项已开启。如果该选项未开启，可能存在以下两种异常情况：

- 系统命令 **lspci** 无法识别到天数智芯加速卡，使用 **lspci -v | grep 1e3e** 命令将无任何输出。
- 天数智芯加速卡内存没有分配到地址空间，显示“unassigned”，详情如下图所示：



```

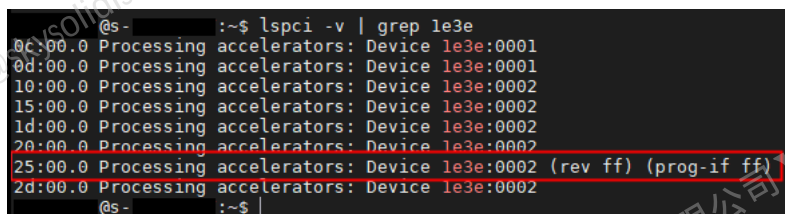
root@es-113-1-85:/home/ / # lspci -vvs a0:00.0
a0:00.0 Processing accelerators: Device 1e3e:0002
Control: I/O- Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR- FastB2B- DisINTx-
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >Abort- <Abort- <MAbort- >SERR- <PERR- INTx-
Latency: 0, Cache Line Size: 32 bytes
Interrupt: pin A routed to IRQ 255
NUMA node: 1
Region 0: Memory at <unassigned> (64-bit, prefetchable)
Capabilities: [40] Power Management version 3
Flags: PMEClk- DSI- D1- D2- AuxCurrent=375mA PME(D0+,D1-,D2-,D3hot+,D3cold-)
Status: D0 NoSoftRst+ PME-Enable- DSel=0 DScale=0 PME-
    
```

Figure 4: 加速卡未分配到地址空间

以上两种情况都无法正常使用天数智芯加速卡。

2. (如果 **Resize Bar** 选项存在) 请确认主机 BIOS 设置中 **Resize Bar** 选项已开启。如果该选项未开启，可能存在的异常情况同 1。
3. 如果 BIOS 设置中 **Above 4G Decoding** 和 **Resize Bar** 选项都已开启，但天数智芯加速卡内存没有分配到地址空间，这种情况大多出现在 10 卡的天垓 150 环境中。原因是 BIOS 版本或者 BIOS 中的设置导致系统没有足够的地址空间分配。此时需要联系服务器厂商确认 BIOS 版本。
4. 请使用系统命令 **lspci -v | grep 1e3e** 查询是否可以正常识别到天数智芯加速卡。

下图示例表示无法正常识别标红框的加速卡：



```

@es- :~$ lspci -v | grep 1e3e
0c:00.0 Processing accelerators: Device 1e3e:0001
0d:00.0 Processing accelerators: Device 1e3e:0001
10:00.0 Processing accelerators: Device 1e3e:0002
15:00.0 Processing accelerators: Device 1e3e:0002
1d:00.0 Processing accelerators: Device 1e3e:0002
20:00.0 Processing accelerators: Device 1e3e:0002
25:00.0 Processing accelerators: Device 1e3e:0002 (rev ff) (prog-if ff)
2d:00.0 Processing accelerators: Device 1e3e:0002
@es- :~$
    
```

Figure 5: 掉卡情况

如果存在上图中“(rev ff) (prog-if ff)”的情况，说明存在掉卡现象。此时，需要关闭主机电源并重新插拔加速卡。

5. 使用系统命令 **lspci -v | grep 1e3e** 查询到的设备数量与实际的设备 (插卡) 数量不符，可能是接触不良导致。此时，建议关闭主机电源后重新插拔加速卡。
6. 请确认天数智芯加速卡有 I/O 访问能力和 Memory 访问能力 (一般情况下，在加载驱动时，驱动会使能)。

下图示例表示加速卡没有 I/O 访问能力和 Memory 访问能力：

```
@ ~$ sudo lspci -vvs 68:00.0
68:00.0 Processing accelerators: Device 1e3e:0001
Control: I/O- Mem- BusMaster- SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR- FastB2B- DisINTx-
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx-
Interrupt: pin A routed to IRQ 37
NUMA node: 0
Region 0: Memory at 1780000000 (64-bit, prefetchable) [disabled] [size=32G]
Region 2: Memory at d8d00000 (32-bit, non-prefetchable) [disabled] [size=256K]
```

Figure 6: 加速卡没有 I/O 访问能力和 Memory 访问能力

如果存在上图中显示的 Region 0 和 Region 2 都是“disabled”状态，表示天数智芯加速卡没有 I/O 访问能力和 Memory 访问能力，需要先使能。

Tip

使用 **lspci -vvs <bus-id>** 命令前，需要先确认待查询的加速卡的 bus-id，参考命令 **lspci -v | grep 1e3e;** 也可以使用 **lspci -vvv -d '1e3e:'** 命令查询所有加速卡的状态。

以设备 bus-id 为 68:00.0 举例，通过以下命令使能：

```
$ sudo setpci -s 68:00.0 COMMAND=0x06
```

Note

如果查询到多张加速卡缺少上述能力，需要逐一对其使能。

使能后，使用 **lspci -vvv -d '1e3e:'** 命令查询天数智芯加速卡状态，确认所有加速卡有 I/O 访问能力和 Memory 访问能力。

更多信息，请访问 [天数智芯官网文档中心](#) 并下载产品对应的《硬件安装说明文档》和《软件栈工具使用指南》。如有未提及的问题，请咨询应用工程师进行答疑。

2.2.3 问题 3：如何监控加速卡状态？

问题描述

我需要知道天数智芯加速卡是否能正常工作。

问题分析

需要通过工具或者方法来判断和指导加速卡工作的实时状态。

解决方法

通过 ixSMI 工具，可以实时查看加速卡的工作状态；另外，可以利用 BMC 固件在 K8s 系统内实时返回加速卡的状态值。

ixSMI 工具的使用方法，请访问 [天数智芯官网文档中心](#) 并下载对应产品的《软件栈工具使用指南》了解详情。

有关 BMC 固件，请访问 [天数智芯官网文档中心](#) 并下载对应产品的《BMC 用户文档》了解详情。

2.2.4 问题 4：如何解决掉卡问题？

问题描述

我在使用天数智芯加速卡过程中，出现设备不见的情况。

问题分析

请提供更详细的使用情况以便天数智芯支持团队进行分析。

解决方法

请分别查询加速卡的固件版本和驱动版本，并收集 dmesg，将查询到的所有信息提供给天数智芯应用工程师用于问题排查和解决。

- 查询固件版本命令：**bash ./ixfw-{gpu_type}-{fw_version}_{ARCH}.run -l**
- 查询驱动版本命令：**ixsmi**

2.2.5 问题 5：为什么带宽测试结果偏低？

问题描述

我在测试天数智芯加速卡的 H2D/D2H/D2D/P2P 带宽性能时，发现部分测试结果低于预期。

问题分析

以下服务器环境可能影响带宽测试性能：

- 在 PCIe 架构中，开启 IOMMU 可能会对 PCIe 带宽产生影响，包括地址转换开销、中断处理、额外的数据复制、PCIe 通道占用等。
- PCIe 交换机开启 ACS 也可能会对 PCIe 带宽产生一些影响，包括额外的转发延迟、带宽利用率、性能开销等。
- 检查 Switch 和天数加速卡的 LnkSta 速率，如果存在 downgraded 的情况，会降低 P2P 带宽速率。
- 检查 PCIe extend 扩展功能的配置，包括 MaxPayloadSize、ExtTag、RlxdOrd、NoSnoop 等，它们是和 PCIe 设备在发送报文或者数据请求时最大有效负载、传输事务数量、数据暂存、cache 一致检查、TLP 线序等相关，可能会影响 P2P 带宽速率。
- NUMA 结构也会影响传输带宽，在不同的 NUMA node 下的 PCIe 设备进行内存读写时速度与响应时间存在差异，相同的 NUMA node 节点下的设备与内存的读写效果会更好。
- 检查拓扑结构，如果连接通过 PCIe host bridge，通常会导致速率降低。
- BIOS 某些版本也可能导致 P2P 性能偏低。
- 某些主板 Switch 的模式可能会影响 P2P 的通路和结果。

解决方法

在测试之前，请检查以下项目，以避免带宽测试结果异常：

- 关闭主机 BIOS 中的 IOMMU/SMMU 设置，并且 grub iommu off。
- 关闭所有 Switch 的 ACS。
- 检查卡所在的 Switch 的 LnkSta 速率是否为 PCIe 4.0 x16。
- 启用天数智芯加速卡对应的 PCIe 设备的 ExtTag 功能。
- NUMA 结构会影响传输带宽，相同的 NUMA node 节点下的设备与内存的读写效果会更好。
- 某些 BIOS 版本会导致带宽性能偏低。

2.3 固件相关

2.3.1 问题 1：升级固件后，固件版本号未发生变化，且重启后仍然存在该问题

问题描述

参考《固件升级工具使用指南》升级固件后，固件版本号未发生变化，且根据提示重启后仍然存在该问题。

解决方法

升级固件后，固件版本号未发生变化的场景下，您需要关闭并重新打开主机电源，再次验证固件是否升级成功。

2.4 软件栈安装相关

2.4.1 问题 1：无法安装天数智算软件栈驱动

问题描述

安装天数智算软件栈出现报错 Corex Driver installation failed。

解决方法

1. 确认您的 Linux kernel 版本属于《附录：支持的 Linux kernel 版本》中天数智芯已验证的支持版本。
2. 确认已安装 Linux kernel 版本相应的头文件。

- a. 您可以通过以下命令进行检查：

```
$ uname -r
3.10.0-1160.42.2.el7.x86_64
```

- b. 检查该 kernel 版本对应的头文件是否已安装：

- CentOS/Kylin:

```
$ ls -l /usr/src/kernels/$(uname -r)
```

- Ubuntu:

```
$ ls -l /usr/src/linux-headers-$(uname -r)
```

- c. 若对应的 Linux kernel 头文件未安装，请查阅以下参考文档：

- CentOS/Kylin: <https://www.tecmint.com/install-kernel-headers-in-centos-7>
- Ubuntu: <https://www.tecmint.com/install-kernel-headers-in-ubuntu-and-debian>

3. 确认已安装与 Linux kernel 版本匹配的 GCC，详见《软件栈安装指南》或《软件栈安装快速入门》中的相关内容。

2.4.2 问题 2：无法调用天数智算软件栈相关命令

问题描述

调用天数智算软件栈相关命令，如 **ixsmi**，提示 Command 'ixsmi' not found。

问题分析

当前用户未设置环境变量。

解决方法

在宿主机上安装天数智算软件栈后，您需要设置 PATH 和 LD_LIBRARY_PATH 环境变量。以软件栈默认安装路径 /usr/local/corex-{v.r.m}/ 为例，您需要：

- 为 PATH 环境变量加上 /usr/local/corex-{v.r.m}/bin
- 为 LD_LIBRARY_PATH 环境变量加上 /usr/local/corex-{v.r.m}/lib

参考如下方法设置环境变量永久生效：

1. 将以下 export 写入 ~/.bashrc 文件：

```
$ export LD_LIBRARY_PATH=/usr/local/corex-{v.r.m}/lib
$ export PATH=/usr/local/corex-{v.r.m}/bin:$PATH
```

2. 保存后执行 **source ~/.bashrc** 命令即可生效。

Note

- 无论是 root 用户还是非 root 用户使用软件栈，都需要设置环境变量。
- 如果在使用软件栈中途切换用户 (如使用 root 用户安装软件栈后，切换 user 用户使用测试工具)，必须要为切换后的用户重新设置环境变量。

2.4.3 问题 3：生成 Docker 镜像后启动容器时，如何指定设备？

解决方法

您在使用如下命令启动 Docker 容器时，

```
$ sudo docker run --shm-size="32g" -it -v /usr/src:/usr/src \
-v /lib/modules:/lib/modules -v /dev:/dev \
--privileged --cap-add=ALL --pid=host corex:{v.r.m}
```

可以通过 **--device** 参数传递指定的设备来启动容器：

1. 需要用带冒号的格式进行映射，不能直接使用 **--device=/dev/iluvatar2**，正确的写法是 **--device=/dev/iluvatar2:/dev/iluvatar0**，冒号前面是宿主机的设备号，冒号后面是映射到容器内部的设备号。
2. 冒号后面是映射到容器内部的设备号，必须是 0-n 按照从小到大连续的顺序，例如 **device=/dev/iluvatar2:/dev/iluvatar0 --device=/dev/iluvatar5:/dev/iluvatar1**。对宿主机的卡号没有限制要求。

2.4.4 问题 4：哪些 Python 包必须安装天数智芯适配版本？为什么？

依赖 CUDA 的 Python 包，都需要安装天数智芯适配的版本，详细列表可参考《版本发布说明》中的“发布 whl 包说明”小节。不依赖 CUDA 的 Python 包，您都可以直接使用 **pip** 命令在线安装，比如 **pip install transformers**，因为 Transformers 依赖 PyTorch，并不直接依赖 CUDA。

Important

为了避免安装过多不必要的包，在安装 Python 包时，建议您先安装天数智芯适配的版本。例如，先安装天数智芯适配版 PyTorch，再 **pip install transformers**。

针对天数智芯硬件，您必须使用天数智芯适配版框架，不应使用官方开源版本，以避免编译和运行报错。例如，使用 **pip install torch** 会采用在线方式下载并安装 PyTorch 官方开源版本，编译和运行可能会报错。天数智芯适配版的 .whl 包是针对天数智芯加速卡和天数智算软件栈优化的。使用这些定制包可以确保深度学习框架和其它相关库能够适配适配天数智芯的硬件指令，充分利用天数智芯加速卡的性能，提供更好的计算效率和优化的运行时体验。

由于天数智芯适配版 .whl 包目前没有在线发布，故不支持在线安装。您需要联系您的应用工程师获取所需 .whl 包并使用 **pip3** 命令进行安装。

2.4.5 问题 5：为什么不能直接使用 pip install torch?

问题原因

使用 **pip install torch** 会采用在线方式下载并安装官方开源版本的 PyTorch。针对天数智芯硬件，您必须使用天数智芯适配版框架，不应使用官方开源版本，以避免编译和运行报错。

解决方法

您需要修改 requirements.txt 文件，删除 PyTorch 等相关依赖的开源条目，并通过离线方式安装从天数智芯获取的适配版 .whl 包。参考《版本发布说明》查看天数智芯适配版 .whl 包列表。

2.4.6 问题 6：如果已有代码指定的 PyTorch 版本不是天数智芯支持的版本，怎么处理？

解决方法

如果已有的代码指定的 PyTorch 版本不是天数智芯适配版，您可以按照以下步骤操作：

1. 检查兼容性：

首先，确认代码是否能与天数智芯建议的 PyTorch 版本兼容。例如，天数智芯提供的 PyTorch 适配版为 v2.1.1，如果代码调用的是 PyTorch v2.0.1，由于 PyTorch 可以向前兼容，您可以修改代码依赖的 PyTorch 版本为 v2.1.1 并使用天数智芯适配版 PyTorch。

2. 修改代码：

如果代码依赖的 PyTorch 版本高于天数智芯提供的适配版，您需要甄别并修改/删除代码中不支持的 PyTorch 新特性。

2.4.7 问题 7：对于不支持的大模型推理引擎，我是否可以自己移植？

对于暂不支持的推理框架/引擎，可与天数智芯团队沟通，不建议用户自行移植开源版本。这是因为天数智芯硬件与开源版本基于的国际主流 GPU 并非完全相同，直接使用开源版本的 kernel，无法最大化天数智芯硬件的性能，因此不建议用户自行适配。

2.4.8 问题 8：机器无法访问网络，无法 pip 在线安装

解决方法

以 vLLM 离线安装为例：

1. 在可以联网的机器上运行以下命令：

```
$ pip3 download vllm-0.3.3+corex.{v.r.m}-py3-none-any.whl -d "./whl" # 将 {v.r.m} 替换为对
↳ 应的天数智算软件栈版本号
```

此时，将会下载 vLLM 依赖的相关 whl 包到 whl 文件夹中。

2. 将 whl 文件夹拷贝至离线的机器上并运行以下命令即可离线安装 vLLM：

```
$ pip3 install --no-index --find-links="./whl" vllm-0.3.3+corex.{v.r.m}-py3-none-any.whl #
↳ 将 {v.r.m} 替换为对应的天数智算软件栈版本号
```

2.4.9 问题 9：使用 corex-docker-installer 的.run 包生成包含天数智算软件栈的 Docker 镜像失败，遇到 error code:100 报错

问题描述

使用 corex-docker-installer 的 .run 包生成包含天数智算软件栈的 Docker 镜像失败，遇到类似以下报错信息：

```
Uncompressing Corex Ubuntu Docker Installer 100%
Docker daemon is not running. exit now.
Corex Docker Installer installation failed. error code:100
```

问题分析

报错 error code:100，表示 dockerd 服务未启动，使用 corex-docker-installer 的 .run 包需要启动 dockerd 服务。

解决方法

1. 执行 **systemctl status docker** 命令，查看 Docker daemon 是否运行中。
2. 如果 Docker daemon 未运行，可能是环境中没安装 Docker Engine。
 - Ubuntu/CentOS：参考 [Docker 官方文档](#) 并选择对应的平台获取安装指南。
 - Kylin：参考 [Kylin 平台安装 Docker 操作指南](#)。

2.5 错误码

错误代码 0

安装成功。

错误代码 1

拒绝 EULA 导致安装失败。

解决方法：接受 EULA。

错误代码 2

退出 UI 界面。

解决方法：无需处理。

错误代码 7

无执行权限。

解决方法：以 root 权限执行。

错误代码 11

安装配置错误，一般情况是未在环境中找到必要头文件。

解决方法：联系您的应用工程师获取安装软件栈的必要头文件，详见《软件栈安装指南》。

错误代码 12

选择同时安装 Driver 和 Toolkit 时，Driver 安装失败，且跳过 Toolkit 安装。

解决方法：建议检查驱动安装日志。

错误代码 31

使用了不正确的 GCC 版本。

解决方法：安装符合要求的 GCC 版本，GCC 版本要求详见《软件栈安装指南》或《软件栈安装快速入门》。

错误代码 32

内核源码未安装正确。

解决方法：检查 `/lib/modules/${uname -r}/build` 是否存在，详见《软件栈安装指南》。

错误代码 33

缺失必要的依赖 (glibc, util-linux, grep, coreutils, binutils, kmod, gcc, make, binutils 等)。

解决方法：检查安装日志，安装缺失的必要依赖。

错误代码 34

驱动编译失败。

解决方法：尝试排查以下情况：

- 当前编译器不支持编译驱动使用的编译参数，检查编译器版本与内核版本是否匹配。
- 使用错误版本的 Mellanox InfiniBand RDMA driver。
- 使用的内核版本未经适配，联系您的应用工程师获取帮助。

错误代码 35

驱动安装失败。

解决方法：建议检查驱动安装日志或 `dmesg`。

错误代码 38

驱动编译对应的内核版本与当前运行的内核版本不匹配。

解决方法：建议重新安装内核源码。

错误代码 39

复制失败错误。

解决方法：建议检查驱动安装日志。

错误代码 43

驱动被占用，无法卸载。

安装程序将尝试卸载当前的天数智芯加速卡驱动，如果驱动正在被使用 (有程序占用驱动)，安装将失败。

解决方法：关闭相关程序后尝试重新安装。

错误代码 45

DKMS 编译驱动失败。

解决方法：尝试排查以下情况：

- 当前编译器不支持编译驱动使用的编译参数，检查编译器版本与内核版本是否匹配。
- 使用错误版本的 Mellanox InfiniBand RDMA driver。
- 使用的内核版本未经适配，联系您的应用工程师获取帮助。

错误代码 49

加载驱动失败，建议检查安装日志或 dmesg，可能存在以下情况：

- 驱动已存在，安装过程中应避免潜在的相关驱动加载；
- 没有相关权限，请检查 CAP_SYS_MODULE 是否有权限，并检查 /proc/sys/kernel/modules_disabled 是否为 0。

解决方法：检查环境中天数智芯加速卡固件是否符合要求。

错误代码 50

由于签名验证失败，导致驱动安装失败。

解决方法：检查安装日志及 dmesg，详情请参考《软件栈安装指南》中“准备并注册密钥对”小节。

错误代码 51

插入内核模块失败。

解决方法：参考错误码 49。

错误代码 52

DKMS 注册驱动模块失败。

解决方法：检查驱动安装日志。

错误代码 53

DEPMOD 失败。

解决方法：检查驱动安装日志。

错误代码 54

固件版本检查失败。

解决方法：建议检查软件栈版本和板卡固件版本。

错误代码 70

文件丢失。

解决方法：建议检查安装包完整性。

错误代码 71

文件夹丢失。

解决方法：建议检查安装包完整性。

错误代码 74

安装磁盘空间不足。

解决方法：检查 /tmp 及安装目录 (默认为 /usr/local/corex) 的空间是否足够。

错误代码 100

dockerd 服务未启动，docker-installer 安装包需要启动 dockerd 服务。

解决方法：检查 **systemctl status docker** 结果，docker-installer 安装包需要 Docker daemon 运行。

错误代码 101

加载 Docker 镜像失败，建议清理 installer 中间镜像后重试安装。

解决方法：执行 **docker rmi \$(docker image ls installer -q)** 命令后尝试重新安装。

错误代码 102

构建 Corex Docker 镜像失败。

解决方法：建议检查安装日志。

3 商标声明

- 天数智芯、天数智芯 logo、Iluvatar CoreX 等商标、标识、组合商标为上海天数智芯半导体有限公司之注册商标或商标，受法律保护。
- 除了天数智芯的注册商标外，本内容中使用的其他产品名称及标志仅用于识别目的，该等名称及标志可能是归属于其各自公司的商标。我们否认对该等名称及标志的所有权利。
- CentOS 标识为 Red Hat 公司的商标。
- Docker 为 Docker 公司在美国和其他国家的商标或注册商标。
- Linux 为 Linus Torvalds 在美国和其它国家的注册商标。
- NVIDIA 和 CUDA 为 NVIDIA 公司在美国和/或其它国家的商标和/或注册商标。
- PyTorch 为 Facebook 公司的商标。
- TensorFlow 为 Google 公司的商标。
- Ubuntu 为 Canonical 公司的注册商标。