

腾讯云GPU 云服务器

操作指南

产品文档



腾讯云

【版权声明】

©2013-2017 腾讯云版权所有

本文档著作权归腾讯云单独所有，未经腾讯云事先书面许可，任何主体不得以任何形式复制、修改、抄袭、传播全部或部分本文档内容。

【商标声明】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。

【服务声明】

本文档意在向客户介绍腾讯云全部或部分产品、服务的当时的整体概况，部分产品、服务的内容可能有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或模式的承诺或保证。

文档目录

文档声明.....	2
操作指南.....	4
最佳实践	4
登录实例	5
重启实例	6
安装 NVIDIA 驱动指引	7
安装 AMD 驱动指引	11
安装 CUDA 驱动指引	12

操作指南

最佳实践

安全组与网络

- 安全组是一种有状态的提供包过滤功能的虚拟防火墙，用户可通过设置安全组允许受信任的地址访问实例，达到限制访问的目的。创建不同的安全组规则应用于不同安全级别的实例组上，确保运行重要业务的实例无法轻易从外部触达。有关更多信息，请参阅 [安全组](#)。
- 定期修补，更新和保护实例上的操作系统和应用程序。
- 借助弹性公网 IP 地址，快速将地址重新映射到账户中的另一个实例（或 NAT 网关实例），从而屏蔽实例故障。有关更多信息，请参阅 [弹性IP地址](#)。
- 尽量使用 [SSH 密钥](#) 方式登录用户的 Linux 类型实例。使用 [密码登录](#) 的实例需要不定期修改密码。
- 选择使用 [私有网络](#) 进行逻辑区的划分。

存储

- 对于可靠性要求极高的数据，请使用 [腾讯云云硬盘](#) 以保证数据持久可靠存储。
- 对于访问频繁、容量不稳定的数据库，可使用 [腾讯云云数据库](#)。
- 利用 [对象存储 COS](#)，存储静态网页和海量图片、视频等重要数据。

备份与恢复

- 通过 [云主机控制台](#) 回滚备份好的 [自定义镜像](#) 恢复等方式。
- 跨多个可用区部署应用程序的关键组件，并适当地复制数据。
- 定期查看监控数据并设置好适当的告警。有关更多信息，请参阅 [云监控产品文档](#)。

登录实例

在购买并启动了 GPU 实例后，您可以连接并登录它。根据您的本地的操作系统、GPU 实例操作系统和 GPU 实例是否可被 Internet 访问，不同情况下可以使用不同的登录方式。

先决条件

- 使用密码登录到 GPU 云服务器时，需要使用管理员帐号和对应的密码；
- 使用密钥登录到GPU 云服务器时需要创建并下载私钥。

登录指引

若 GPU 实例为 Linux 实例，具体登录指引可参考[登录 Linux 实例](#)

若 GPU 实例为 Ubuntu 实例，具体登录指引可参考[登录 Windows 实例](#)

重启实例

重启操作是维护 GPU 云服务器的一种常用方式，重启实例相当于本地计算机的重启操作系统操作。

概述

- 重启准备：重启期间实例将无法正常工作，因此在重启之前，请确保 GPU 云服务器已暂停业务请求。
- 重启操作方式：建议使用腾讯云提供的重启操作进行实例重启，而非在实例中运行重启命令（如 Windows 下的重新启动命令及 Linux 下的 Reboot 命令）。
- 重启时间：一般来说重启操作后只需要几分钟时间。
- 实例物理特性：重启实例不改变实例的物理特性。实例的公网 IP、内网 IP、存储的任何数据都不会改变。
- 计费相关：重启实例不启动新的实例计费时间。

使用控制台重启实例

1. 登录 [云主机控制台](#)。
2. 重启单个实例：勾选需要重启的实例，在列表顶部，单击【重启】按键。或在右侧操作栏中，单击【更多】-【云主机状态】-【重启】。
3. 重启多个实例：勾选所有需要重启的实例，在列表顶部，单击【重启】按键。即可批量重启实例。不能重启的实例会显示原因。

使用 API 重启实例

请参考 [RebootInstances 接口](#)。

安装 NVIDIA 驱动指引

GPU 云服务器必须具备相应的 NVIDIA 驱动，必须针对您的编译环境安装合适的 NVIDIA 驱动程序。

Linux 安装驱动

1. 登录 [NVIDIA 驱动下载](#) 或打开链接 <http://www.nvidia.com/Download/Find.aspx>。
2. 选择操作系统和安装包。以 M40 为例，选择要下载的驱动版本：

Product Type:

Tesla ▼

Product Series:

M-Class ▼

Product:

M40 24GB ▼

Operating System:

Linux 64-bit ▼

CUDA Toolkit:

8.0 ▼

Language:

English (US) ▼

Recommended/Beta:

All ▼

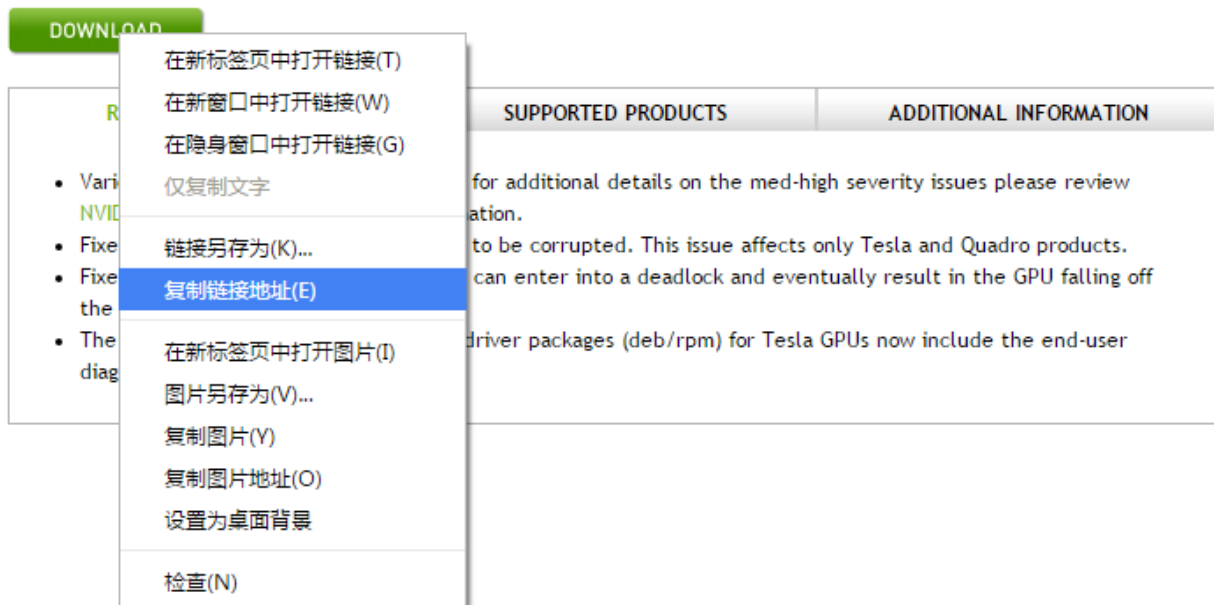
SEARCH

Name	Version	Release Date	CUDA Toolkit
+ Tesla Driver for Linux x64 	384.66	August 14, 2017	8.0
+ Tesla Driver for Linux x64 	375.74	July 31, 2017	8.0
+ Tesla Driver for Linux x64 	384.59	July 28, 2017	8.0
+ Tesla Driver for Linux x64 	375.66	May 9, 2017	8.0
+ Tesla Driver for Linux x64 	367.92	April 14, 2017	8.0
+ Tesla Driver for Linux x64 	375.51	April 5, 2017	8.0
+ Tesla Driver for Linux x64 	375.39	February 15, 2017	8.0
+ Tesla Driver for Linux x64 	375.20	December 9, 2016	8.0
+ Tesla Driver for Linux x64 	367.55	October 24, 2016	8.0
+ Tesla Driver for Linux x64 BETA 	361.93.02	September 26, 2016	8.0

3. 跳转后，右击【DOWNLOAD】 - 【复制链接地址】。

TESLA DRIVER FOR LINUX X64

Version: 384.66
Release Date: 2017.8.14
Operating System: Linux 64-bit
CUDA Toolkit: 8.0
Language: English (US)
File Size: 97.77 MB



4. 登录 GPU 实例，使用

wget

命令，粘贴上述步骤复制的链接地址下载安装包；或通过在本机系统下载 NVIDIA 安装包，上传到 GPU 实例的服务器。

```
[root@ip-10.10.10.10 ~]# wget http://www.nvidia.com/content/DriverDownload-March2009/confirmation.php?url=/XFree86/Linux-x86_64/384.66/NVIDIA-Linux-x86_64-384.66.run&lang=us&type=Tesla
```

5. 对安装包加执行权限。比如对文件名为

NVIDIA-Linux-x86_64-352.99.run

加执行权限

chmod +x NVIDIA-Linux-x86_64-352.99.run

6. 安装当前系统对应的 gcc 和 kernel-devel 包

```
sudo yum install -y gcc kernel-devel-xxx
```

xxx

是内核版本号，可以通过

```
uname -r
```

查看。

7. 运行驱动安装程序

```
sudo /bin/bash ./NVIDIA-Linux-x86_64-352.99.ru
```

按提示进行后续操作。

8. 安装完成后，运行 nvidia-smi 如果有类似如下的 GPU 信息显示出来，说明驱动安装成功。

NVIDIA-SMI 352.99 Driver Version: 352.99									
GPU	Name	Persistence-M		Bus-Id	Disp.A	Volatile	Uncorr. ECC		
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage		GPU-Util	Compute M.		
0	Tesla M40 24GB	Off	0000:03:00.0	Off		Off			
N/A	27C	P0	55W / 250W	88MiB / 24575MiB		0%	Default		
1	Tesla M40 24GB	Off	0000:82:00.0	Off		Off			
N/A	28C	P0	57W / 250W	88MiB / 24575MiB		0%	Default		
Processes:									
GPU	PID	Type	Process name	GPU Memory Usage					
No running processes found									

Windows 安装驱动

1. 登录 [NVIDIA 驱动下载官网](#)。
2. 选择操作系统和安装包。以 M40 为例，选择如下驱动程序：

NVIDIA Driver Downloads

Advanced Driver Search

Product Type:

Tesla ▼

Product Series:

M-Class ▼

Product:

M40 24GB ▼

Operating System:

Windows Server 2012 R2 64 ▼

CUDA Toolkit:

8.0 ▼

Language:

English (US) ▼

Recommended/Beta:

All ▼

SEARCH

3. 打开下载驱动程序的文件夹，然后双击安装文件以启动它。按照说明安装驱动程序并根据需要重启实例。要验证 GPU 是否正常工作，请检查设备管理器。

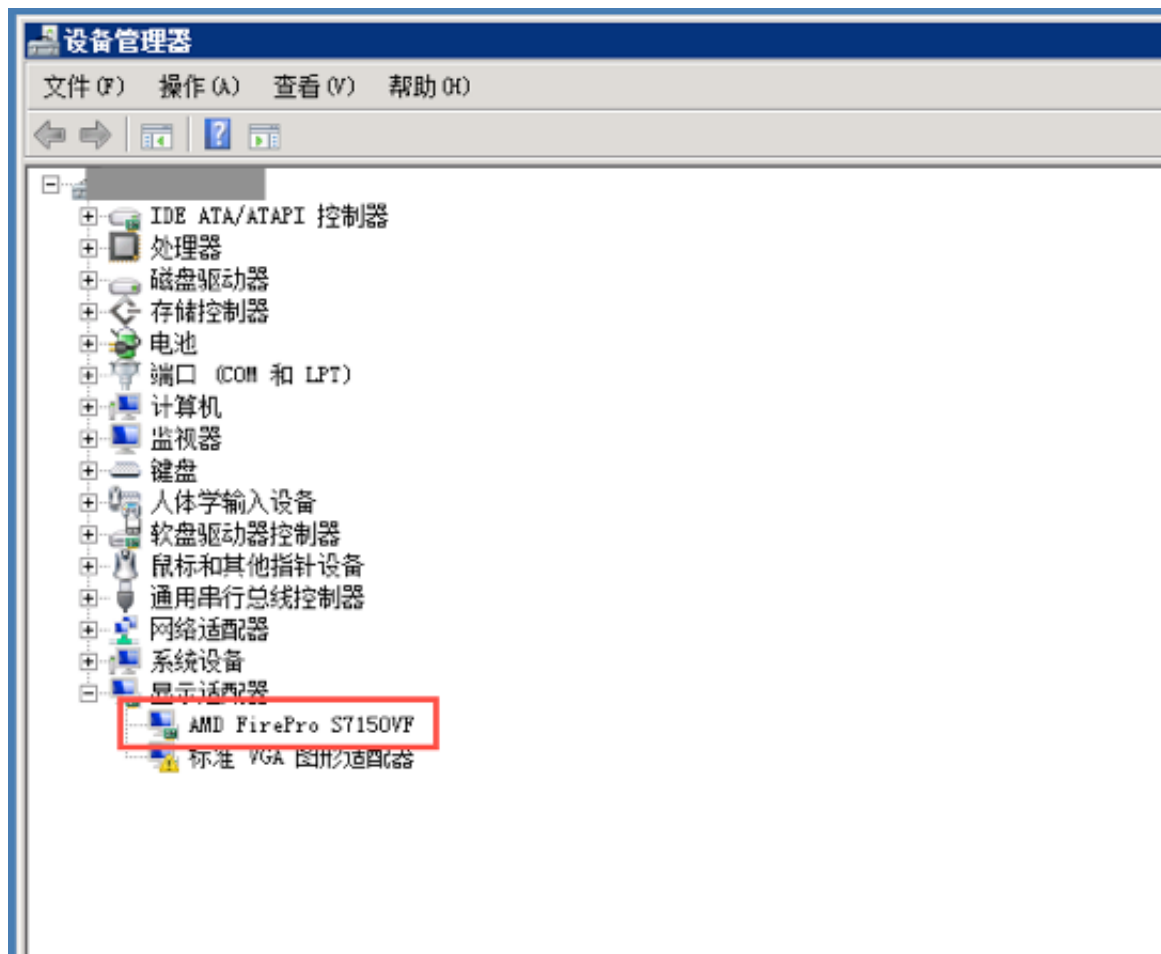
安装 AMD 驱动指引

GPU 云服务器必须具备相应的 GPU 驱动才可正常运用，对于AMD GPU 云服务器必须针对您的编译环境安装合适的 AMD GPU 驱动程序。

GA2 实例驱动安装

Windows 安装驱动

1. 请从您购买的GPU云服务器内访问该链接下载 AMD GPU 驱动
http://mirrors.tencentyun.com/install/windows/s7150_guest_driver.7z
2. 打开下载驱动程序的文件夹，然后双击安装文件以启动它。要验证 GPU 是否正常工作，请检查设备管理器
3. 安装完成后，到设备管理器内查看，如下显示表明安装成功



安装 CUDA 驱动指引

CUDA (Compute Unified Device Architecture) 是显卡厂商 NVIDIA 推出的运算平台。CUDA™ 是一种由 NVIDIA 推出的通用并行计算架构，该架构使 GPU 能够解决复杂的计算问题。它包含了 CUDA 指令集架构 (ISA) 以及 GPU 内部的并行计算引擎。开发人员现在可以使用 C 语言, C++, FORTRAN 来为 CUDA™ 架构编写程序，所编写出的程序可以在支持 CUDA™ 的处理器上以超高性能运行。

GPU 云服务器采用 NVIDIA 显卡，需要安装 CUDA 开发运行环境。以目前最常用的 CUDA 7.5 为例，可参照以下步骤进行安装。

Linux 系统指引

1. 登录 [CUDA驱动下载](#) 或复制链接 <https://developer.nvidia.com/cuda-75-downloads-archive>。
2. 选择操作系统和安装包。以 CentOS 7.2 64 位为例，可按如下方式进行选择：

Select Target Platform ⓘ

Click on the green buttons that describe your target platform. Only supported platforms will be shown.

Operating System	Windows	Linux	Mac OSX
Architecture ⓘ	x86_64	ppc64le	
Distribution	Fedora	OpenSUSE	RHEL
	SLES	SteamOS	Ubuntu
Version	7	6	
Installer Type ⓘ	runfile (local)	rpm (local)	rpm (network)

注意：

Installer Type 推荐选择 rpm (network)。

network：网络安装包，安装包较小，需要在主机内联网下载实际的安装包。

local：本地安装包。安装包较大，包含每一个下载安装组件的安装包。

3. 右击【Download】 - 【复制链接地址】。



4. 登录 GPU 实例，使用

wget

命令，粘贴上述步骤复制的链接地址下载安装包；或通过在本机系统下载 CUDA 安装包，上传到 GPU 实例的服务器。

```

$ wget http://developer.download.nvidia.com/compute/cuda/7.5/Prod/local_installers/cuda-repo-rhel7-7-5-local-7.5-18.x86_64.rpm
Resolving developer.download.nvidia.com (developer.download.nvidia.com)... 111.202.43.167, 111.202.43.168, 111.202.43.169, ...
Connecting to developer.download.nvidia.com (developer.download.nvidia.com):111.202.43.167:80... connected.
HTTP request sent, awaiting response... 200 OK
Length: 1211912140 (1.1G) [application/x-rpm]
Saving to: 'cuda-repo-rhel7-7-5-local-7.5-18.x86_64.rpm'

```

5. 在 CUDA 安装包所在目录下运行如下命令：

```
sudo rpm -i cuda-repo-rhel7-7.5-18.x86_64.rpm
```

```
sudo yum clean all
```

```
sudo yum install cuda
```

6. 在

```
/usr/local/cuda-7.5/samples/1_Uutilities/deviceQuery
```

目录下，执行

```
make
```

命令，可以编译出 deviceQuery 程序。

7. 执行 deviceQuery 正常显示如下设备信息，此刻认为 CUDA 安装正确。

```
./deviceQuery Starting...

CUDA Device Query (Runtime API) version (CUDA RT static linking)
Detected 1 CUDA Capable device(s)
Device 0: "Tesla M40 24GB"
  CUDA Driver Version / Runtime Version      8.0 / 7.5
  CUDA Capability Major/Minor version number: 5.2
  Total amount of global memory:             24505 MBytes (25695092736 bytes)
  (24) Multiprocessors, (128) CUDA Cores/MP: 3072 CUDA Cores
  GPU Max Clock rate:                        1112 MHz (1.11 GHz)
  Memory Clock rate:                         3004 Mhz
  Memory Bus Width:                          384-bit
  L2 Cache Size:                            3145728 bytes
  Maximum Texture Dimension Size (x,y,z)     1D=(65536), 2D=(65536, 65536), 3D=(4096, 4096, 4096)
  Maximum Layered 1D Texture Size, (num) layers 1D=(16384), 2048 layers
  Maximum Layered 2D Texture Size, (num) layers 2D=(16384, 16384), 2048 layers
  Total amount of constant memory:            65536 bytes
  Total amount of shared memory per block:    49152 bytes
  Total number of registers available per block: 65536
  Warp size:                                 32
  Maximum number of threads per multiprocessor: 2048
  Maximum number of threads per block:        1024
  Max dimension size of a thread block (x,y,z): (1024, 1024, 64)
  Max dimension size of a grid size    (x,y,z): (2147483647, 65535, 65535)
  Maximum memory pitch:                      2147483647 bytes
  Texture alignment:                          512 bytes
  Concurrent copy and kernel execution:       Yes with 2 copy engine(s)
  Run time limit on kernels:                  No
  Integrated GPU sharing Host Memory:         No
  Support host page-locked memory mapping:    Yes
  Alignment requirement for Surfaces:         Yes
  Device has ECC support:                     Disabled
  Device supports Unified Addressing (UVA):    Yes
  Device PCI Domain ID / Bus ID / location ID: 0 / 0 / 7
  Compute Mode:
    < Default (multiple host threads can use ::cudaSetDevice() with device simultaneously) >

deviceQuery, CUDA Driver = CUDART, CUDA Driver Version = 8.0, CUDA Runtime Version = 7.5, NumDevs = 1, Device0 = Tesla M40
Result = PASS
```

Windows 系统指引

要在 Windows 实例上安装 CUDA，请使用远程桌面以管理员的身份登录您的 Windows 实例。

1. 在 [CUDA 驱动官网](#) 下载 CUDA 安装包。
2. 选择操作系统和安装包。以 Win Server 2012 R2 64 位为例，可按如下方式进行选择:

Select Target Platform ⓘ

Click on the green buttons that describe your target platform. Only supported platforms will be shown.

Operating System

WindowsLinuxMac OSX

Architecture ⓘ

x86_64

Version

108.17Server 2012 R2Server 2008 R2

Installer Type ⓘ

exe (network)exe (local)

Download Target Installer for Windows Server 2012 R2 x86_64

cuda_7.5.18_windows_network.exe (md5sum: bd166ba460e7507e2ddab3e324653263)

Download (8.5 MB)

Installation Instructions:

1. Double click cuda_7.5.18_windows_network.exe
2. Follow on-screen prompts

For further information, see the [Installation Guide for Microsoft Windows](#) and the [CUDA Quick Start Guide](#).

3. 启动安装程序，按提示进行安装，如果最后出现完成对话框，则安装成功。

