GPU 文档

2019-12-29



百度智能云

目录

1 产品描述 1

1.1 介绍 1

1.2 优势 1

1.2.1 计算能力 1

1.2.2 镜像 2

1.2.3 网络环境 2

1.2.4 存储系统 2

1.2.5 监控系统 2

1.3 GPU卡详情及使用场景 3

1.3.1 GPU卡详情 3

1.3.2 GPU使用场景 3

2 产品定价 5

2.1 包年包月计费 5

2.2 按需计费 7

3 操作指南 10

3.1 快速入门 10

3.1.1 实名认证及注册 10

3.1.2 操作流程介绍 10

3.2 创建GPU实例 11

3.2.1 操作步骤 11

3.2.2 管理GPU实例 12
3.3 镜像使用 12
3.3.1 GPU专用镜像使用说明 12
3.3.2 公共镜像使用步骤 13
3.4 数据上传 14
3.4.1 数据上传 14
3.4.2 BOS设置 14
3.4.3 BOS在GPU服务器上的使用功能方法 15
3.4.4 BOS的其他功能及计费 15
3.5 申请GPU物理服务器 16
3.5.1 GPU物理服务器特点 16
3.5.2 申请步骤 16
4 常见问题 17
4.1 一般类问题 17

第1章　产品描述

1.1　介绍

GPU服务器是百度智能云中配备独享GPU卡的高性能云计算服务，可以帮助您快速、便捷的获得高质量的GPU计算资源，能够大幅提高机器学习及科学计算等大规模计算框架的运行速度，为搭建人工智能及高性能计算平台提供基础架构支持。

GPU服务器提供两种形态的计算资源：虚机形态的GPU云服务器和物理机形态的GPU物理服务器。

GPU云服务器提供与使用标准BCC云服务器一致的管理方式，您可以快速灵活的申请并管理GPU云服务器，结合私有网络VPC、负载均衡BLB构建云环境中高可用的GPU集群。

当普通GPU云服务器无法满足您的业务要求时，百度智能云可以为您提供租用独享的GPU物理服务器的服务。并在您计划搭建超高性能的GPU集群时，为您提供配置更多GPU卡（8卡以上）和100G互联带宽的超高性能的GPU物理服务器集群，立即申请GPU物理服务器 。

1.2　优势

1.2.1　计算能力

** 高配计算节点**

- 将GPU云服务器上的CPU、内存、本地磁盘配置与GPU数量按照最优比例组合，确保GPU保持最优计算性能。
- 每台GPU云服务器均独享物理GPU卡，确保GPU性能稳定可靠。

** 多种GPU卡**

- 提供NVIDIA Tesla V100、Tesla P40、Tesla P4、K40M 等多种高性能计算卡供用户自由选择，了解详细性能参数。
- 每台GPU云服务器可选配置1、2、4块GPU卡，CPU、内存、磁盘配置随GPU数量线性增加。

1.2.2 镜像

** 集成GPU驱动**

在Ubuntu16.04镜像基础上，集成最新的GPU驱动和CUDA 8.0，帮助用户解决底层兼容问题，加速业务上云。同时也会陆续支持更多镜像。

公共镜像

提供更多公共镜像支持，用户可自行选择发行版公共镜像，并根据需要自行安装GPU驱动和相应软件。

自定义镜像 可基于已部署好服务的GPU云服务器的系统盘生成自定义镜像，再通过自定义镜像创建更多GPU云服务器，从而实现业务的快速复制和部署。

1.2.3 网络环境

私有网络VPC

私有网络VPC，基于百度智能云构建一个安全隔离的网络环境，支持用户可自定义网络地址空间、多VPC之间（同城、跨城）对等高速连接，通过VPN/专线的方式，与用户的数据中心构建安全、定制的混合云网络，实现原有业务轻松、安全的迁移到云端。

100G高性能物理集群

面向超高性能GPU集群客户，可提供100G物理服务器间的互联网络，提升GPU集群间的数据并行能力。

1.2.4 存储系统

云磁盘CDS

在GPU云服务器本地磁盘容量不能满足使用需求时，可通过购买CDS云磁盘来实现存储资源的扩展。 了解详情>>

对象存储BOS

当需要将本地大量需要处理的数据上传云端时，可使用更具性价比的对象存储BOS。在使用数据时可选择通过存储网关BSG或BOS FS功能，将BOS数据作为共享数据在多台GPU云服务器间共享使用。

1.2.5 监控系统

GPU监控

针对GPU云服务器，提供细粒度的GPU卡状态监控和报警机制，包括GPU使用率、内存使用

率、温度、状态等一系列监控项。

1.3 GPU卡详情及使用场景

1.3.1 GPU卡详情

GPU服务器使用的NVIDIA GPU卡基本参数信息如下表所示：

GPU卡型号	CUDA Cores	显存容量	单精度浮点性能	双精度浮点性能	INT8性能	混合精度
NVIDIA Tesla V100-32G	5120	32GB	15.7 Tflops	7.8 Tflops	–	125 Tflops
NVIDIA Tesla V100-16G	5120	16GB	15.7 Tflops	7.8 Tflops	–	125 Tflops
NVIDIA Tesla P40	3840	24GB	12 Tflops	–	47 Tops	–
NVIDIA Tesla P4	2560	8GB	5.5 Tflops	–	22 Tops	–
NVIDIA Tesla K40M	2880	12GB	4.29 Tflops	1.43 Tflops	–	–
NVIDIA 深度学习开发卡	3584	12GB	11 Tflops	–	44 Tops	–

1.3.2 GPU使用场景

针对深度学习的离线训练场景，如果您是深度学习的初学者或初创公司，我们向您推荐性价比最高的NVIDIA深度学习开发卡，在创建GPU云服务器时再配合集成GPU驱动镜像进行使用，能帮助您快速完成环境部署和Training实验。

针对深度学习的离线训练场景，如果您是深度学习的重度用户，周期性有大量数据需要训练，我们向您推荐性能和稳定性更好的NVIDIA Tesla P40，如果您对性能有极致的要求，更可选择支持Tensor Core功能的NVIDIA Tesla V100英伟达最新GPU卡。并可以选择按需使用，训练结束后即可将资源释放节省成本，同时通过自定义镜像功能，快速完成GPU云服务器的服务部署和搭建。

针对深度学习的在线预测场景，相比离线训练对GPU性能的要求降低，但对运行稳定性要求更高，对服务器响应延时也有了更高要求，所以我们向您推荐NVIDIA Tesla P4 在满足

性能要求的同时，提供更具性价比的选择。

当然，没有最好的云资源，只有最适合业务使用的才是最好的。所以我们为您提供了多种GPU类型选择，帮助您依据自身业务特点和资源投入预算来选择不同的云资源。当您对训练能力没有太高要求时，NVIDIA Tesla K40M也是个不错的选择；让您对离线训练能力的要求超高时，也可以来申请GPU物理服务器，可以为您提供具备100G网络互联的GPU集群，帮您提速训练任务。

第2章　产品定价

2.1　包年包月计费

GPU包年包月服务采用预付费方式，价格较按需计费更低，并赠送系统盘（Linux送20GB，Windows送40GB）。

如需公网访问，请购买弹性公网IP EIP，可至EIP价格查看带宽价格。

> 说明：购买前需保证账户无欠款。

收费标准

- GPU云服务器套餐配置及价格（北京）

GPU卡型号	GPU数量（块）	CPU（核）	内存（GB）	磁盘（GB）	价格（元/月）	价格(元/年)8.3折
NVIDIA Tesla V100-32G	1	12	56	400	5298	52768.08
NVIDIA Tesla V100-32G	2	24	112	800	10596	105536.16
NVIDIA Tesla V100-32G	4	48	224	1600	21192	211072.32
NVIDIA Tesla V100-32G	8	96	448	3200	42384	422144.64
NVIDIA Tesla V100-16G	1	12	56	400	5129	51084.84
NVIDIA Tesla V100-16G	2	24	112	800	10258	102169.68

GPU 卡型号	GPU 数量（块）	CPU（核）	内存（GB）	磁盘（GB）	价格（元/月）	价格(元/年)8.3折
NVIDIA Tesla V100-16G	4	48	224	1600	20516	204339.36
NVIDIA Tesla V100-16G	8	96	448	3200	41032	408678.72
NVIDIA Tesla P40	1	12	40	450	4017	40009.32
NVIDIA Tesla P40	2	24	80	900	8034	80018.64
NVIDIA Tesla P40	4	48	160	1800	16068	160037.28
NVIDIA Tesla P4	1	12	40	450	2880	28684.8
NVIDIA Tesla P4	2	24	80	900	5760	57369.6
NVIDIA Tesla P4	4	48	160	1800	11520	114739.2
NVIDIA Tesla K40	1	6	40	200	3150	31374
NVIDIA Tesla K40	2	12	80	400	6300	62748
NVIDIA Tesla K40	4	24	160	800	12600	125496
NVIDIA 深度学习开发卡	1	6	40	200	2500	24900
NVIDIA 深度学习开发卡	2	12	80	400	5000	49800
NVIDIA 深度学习开发卡	4	24	160	800	10000	99600

说明： GPU云服务器包年优惠规则如下： 购买一年价格为总价的8.3折、两年为7折、三年为5折。

到期提醒和处理

- 到期提醒：GPU服务到期前7天，系统会给您发送即将到期提醒通知。
- 到期后处理：到期后立即停服，系统会发送欠费停服通知。数据为您保留7天，期间不收取费用，7天内未充值则释放，释放前1天和释放时系统都会发送释放通知。

2.2 按需计费

GPU按需计费服务按小时计费，并赠送系统盘（Linux送20GB，Windows送40GB）。具体规则如下：

- 按小时扣费，即北京时间整点扣费并生成账单。出账单时间是当前计费周期结束后1小时内。例如，10:00-11:00的账单会在12:00之前生成，具体以系统出账时间为准。
- 购买实例前需保证账户无欠款，且保证账户余额和可用代金券总和大于或等于100元。

如需公网访问，请购买弹性公网IP EIP，可至EIP价格查看带宽价格。

收费标准

- GPU云服务器套餐配置及价格（北京）

GPU卡型号	GPU数量（块）	CPU（核）	内存（GB）	磁盘（GB）	按需计费（元/小时）
NVIDIA Tesla V100-32G	1	12	56	400	11.04
NVIDIA Tesla V100-32G	2	24	112	800	22.08
NVIDIA Tesla V100-32G	4	48	224	1600	44.15
NVIDIA Tesla V100-32G	8	96	448	3200	88.30
NVIDIA Tesla V100-16G	1	12	56	400	10.69
NVIDIA Tesla V100-16G	2	24	112	800	21.37
NVIDIA Tesla V100-16G	4	48	224	1600	42.74

GPU卡型号	GPU 数 量（块）	CPU（核）	内存（GB）	磁盘（GB）	按 需 计 费（元/小时）
NVIDIA Tesla V100-16G	8	96	448	3200	85.48
NVIDIA Tesla P40	1	12	40	450	8.37
NVIDIA Tesla P40	2	24	80	900	16.74
NVIDIA Tesla P40	4	48	160	1800	33.48
NVIDIA Tesla P4	1	12	40	450	6.00
NVIDIA Tesla P4	2	24	80	900	12.00
NVIDIA Tesla P4	4	48	160	1800	24.00
NVIDIA Tesla K40	1	6	40	200	6.56
NVIDIA Tesla K40	2	12	80	400	13.13
NVIDIA Tesla K40	4	24	160	800	26.25
NVIDIA 深度学习开发卡	1	6	40	200	5.21
NVIDIA 深度学习开发卡	2	12	80	400	10.42
NVIDIA 深度学习开发卡	4	24	160	800	20.83

余额不足提醒和欠费

- 余额不足提醒：
  - 根据您最近3天的账单金额来判断您的账户余额（含可用代金券）是否足够支付未来3天的费用，若不足以支付，系统发送续费提醒。
  - 根据您最近1天的账单金额来判断您的账户余额（含可用代金券）是否足以支付未来1天的费用，若不足以支付，系统发送续费提醒。
- 欠费处理：
  - 北京时间整点检查您的账户余额是否足以支付本次GPU账单的费用（如北京时间11点整检查账户余额是否足以支付10点至11点的账单费用），若不足以支付，即

为欠费，欠费时系统会发送欠费通知。

- 欠费后立即停服，系统会发送欠费停服通知。数据为您保留7天，期间不收取费用，7天内未充值则释放，释放前1天和释放时系统都会发送释放通知。

第3章　操作指南

3.1　快速入门

GPU云服务器是基于NVIDIA GPU卡的高性能云计算服务。我们采用成熟的Passthrough技术将GPU卡透传给高配云服务器，使每位用户都能独享一块或多块GPU卡，确保GPU性能完美发挥。

在操作GPU云服务器前，应先完成以下任务：

1. 实名认证及注册百度智能云帐号
2. 购买GPU服务

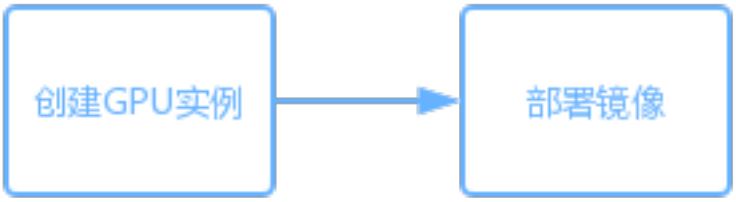
3.1.1　实名认证及注册

请根据自身情况，进行”企业认证”或者“个人认证”，具体请参考认证流程。

如果您在创建GPU 实例的时候仍未完成实名认证，你可以点击页面上的『认证』提示按钮到实名认证页面完成相关操作。

注册百度智能云账号，请参考注册百度账号，完成注册操作。

3.1.2　操作流程介绍



1. 创建GPU实例
2. 选择适合的镜像类型，部署镜像

3.2 创建GPU实例

3.2.1 操作步骤

1. 登录BCC管理控制台主界面。
2. 登录成功后，选择“产品服务>云服务器 BCC”，进入“实例列表”页面。
3. 点击“创建实例”，选择配置信息，然后设置实例名称和管理员密码，选择购买时长和数量。

配置信息	说明
付费方式	预付费（包年包月）、后付费（按需购买）
当前地域	华北-北京、华南-广州
可用区	可用区是指在同一区域下，电力和网络互相独立的区域，故障会被隔离在一个可用区内。
类型	GPU 实例
GPU型号	NVIDIA TeslaV100、 P40、 P4、 K40、NVIDIA 深度学习开发卡
GPU数量	1、2、4块
操作系统	Ubuntu 16.04 LTS amd64 (64bit)-CUDA9.2 等 及 BCC公共镜像
镜像类型	提供集成GPU驱动和CUDA开发环境的Ubuntu16.04 GPU专用镜像、普通公共镜像和自定义镜像
本地磁盘	提供固定容量的本地磁盘（暂不支持自主选择）
网络类型	当前服务器所属的虚拟私有网络，缺省情况下系统默认私有网络。

4. 点击“下一步”，确认订单并支付。
5. 购买成功后，可以在“服务器实例列表”页查看。

3.2.2 管理GPU实例

您可以对当前账户中的GPU开启、停止、重启、修改名称以及绑定公网IP等操作。

1. 登录控制台主界面。
2. 登录成功后，选择“产品服务>云服务器 BCC”，显示“实例列表”页面。
3. 查看当前云服务器详情。
4. （可选）绑定公网IP，点击当前GPU 实例的公网IP图标，在弹框中选择需要绑定的弹性公网IP。如还未创建公网IP，请先购买，参见创建EIP流程。

3.3 镜像使用

3.3.1 GPU专用镜像使用说明

GPU专用镜像中，会集成GPU驱动、CUDA、cuDNN、主流深度学习框架等，具体信息如下：

GPU专用镜像	CUDA版本	cuDNN	深度学习框架版本
Ubuntu 16.04 LTS amd64 (64bit)-CUDA9.2	CUDA 9.2	cuDNN 7.1.4	无
Ubuntu 16.04 LTS amd64 (64bit)-CUDA8.0	CUDA 8.0	cuDNN_v5.1、cuDNN_v6	无
16.04 LTS amd64 (64bit)-CUDA9.0-深度学习开发镜像	CUDA 9.0	cuDNN 7.1.4	TensorFlow_gpu 1.10.1、PaddlePaddle_gpu 0.14.0、Caffe2
Ubuntu 16.04 LTS amd64 (64bit)-CUDA8.0-paddlepaddle_0.11.0	CUDA 8.0	cuDNN 7.0	PaddlePaddle_gpu 0.11.0
Ubuntu 16.04 LTS amd64 (64bit)-CUDA8.0-tensorflow_1.3.0	CUDA 8.0	cuDNN 7.0	TensorFlow_gpu 1.3.0
CentOS 7.5 x86_64 (64bit)-CUDA10.0	CUDA 10.0	无	无
CentOS 7.5 x86_64 (64bit)-CUDA9.2	CUDA 9.2	cuDNN 7.1.4	无

GPU专用镜像	CUDA版本	cuDNN	深度学习框架版本
CentOS 6.8 x86_64 (64bit)-CUDA9.2	CUDA 9.2	cuDNN 7.1.4	无

使用说明：

- 如果您购买的是英伟达Tesla V100 GPU卡，请选择CUDA 9.2版本的镜像创建GPU实例，使用CUDA 8.0版本镜像会导致GPU实例中无法识别到V100的GPU卡。
- 在您使用16.04 LTS amd64 (64bit)-CUDA9.0-深度学习开发镜像时，我们帮您预装了多个框架，在使用时需要通过source 沙箱方式使用，例如如需使用Caffe2环境，需要采用如下方式 # source /home/Caffe2-py27/bin/activate 切换。
- 没有安装深度学习框架的GPU专用镜像中，cuDNN已下载到 /home/work/cudnn/ 目录中，需要您手动安装，可参考英伟达官方文档：https://docs.nvidia.com/deeplearning/sdk/cudnn-install/index.html。如您需要其他版本的cuDNN，也可自行下载：https://developer.nvidia.com/rdp/cudnn-download。

3.3.2 公共镜像使用步骤

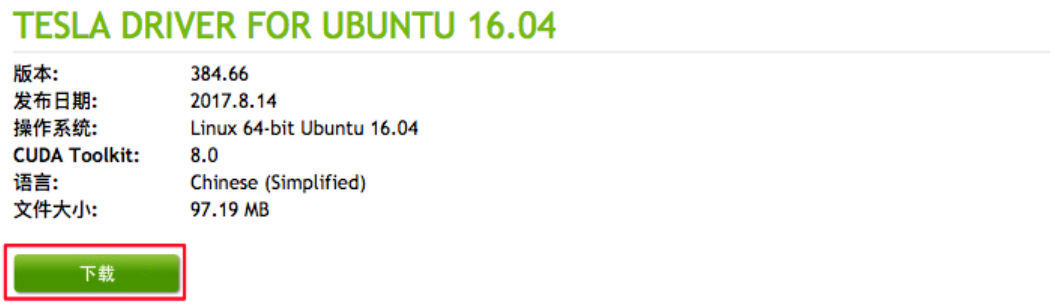
BCC公共镜像，不包含GPU驱动的官方发行版本，需要用户自行安装NVIDIA GPU驱动，可参考如下方式：

下载GPU驱动：

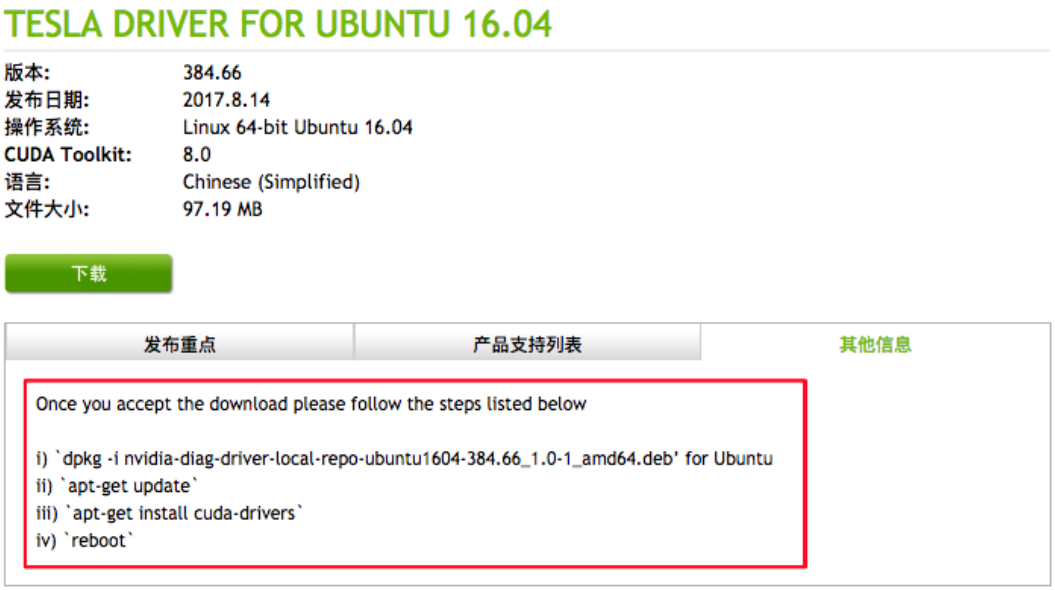
1. 在NVIDIA官方网站下载与实例包含GPU型号相对应的驱动程序。地址：http://www.nvidia.com/Download/index.aspx?lang=cn。
2. 手动查找适用于GPU实例的驱动程序，主要注意GPU型号和操作系统版本：



3. 点击搜索按钮进入下载页面。
4. 核实信息后，点击下载，将驱动程序下载到本地。



5. 将驱动程序上传到云主机中，并按照NVIDIA官网提示进行安装：



6. 自2017年12月19日，暂停提供GPU实例对于Windows进行的支持，用户将无法选择Windows镜像创建新的GPU实例，后续开放Windows镜像支持时间请关注官方公告。”

3.4 数据上传

3.4.1 数据上传

百度智能云提供对象存储BOS服务，这是一个类似于百度网盘的存储空间，提供简单的客户端和CLI工具来进行使用。

对于GPU使用过程中需要用到的数据集，您可以先将他们统一上传到您的对象存储中，在需要用GPU作训练时，再将数据拉取到GPU云服务器的本地进行处理。

3.4.2 BOS设置

1. 访问[BOS产品页](https://cloud.baidu.com/product/bos.html)，点击立即购买登录控制台；

2. 在Bucket管理页面点击新建bucket，命名后选择所属地域（建议选择与GPU服务器相同的地域，目前GPU服务器实例的所属地域均在北京），并设置读写权限为私有；
3. 创建完成后，在Bucket管理页面点击创建完成的Bucket名字进入Bucket，即可在web上查看和上传相关数据文件；
4. 文件上传完成后，点文件名即可获取文件地址，在GPU服务器平台上直接通过wget方式获取。

3.4.3 BOS在GPU服务器上的使用功能方法

BOS CLI是BOS调用的命令行工具，以下介绍如何在GPU云服务器中使用BOS CLI工具。

1. 从BOS CLI工具文档页 获取linux下BOS CLI工具的压缩包。
2. 解压CLI工具包:

```
\$ unzip bce-cli-0.8.3.zip
```

3. 将bcecli的库安装到系统的python目录下:

```
\$ python setup.py install
```

4. 使用BOS CLI工具之前，推荐先设置Access Key、Secure Key、Region；可以通过 -c/–configure 来设置AK、SK、Region，CLI会在配置后自动写入当前用户主目录；AK、SK可以从登录BOS控制台后，点击右上角按钮悬窗中Access Key链接获得。Region则应与BOS设置中，所属地域的选择一致，选择北京填bj，选择广州则填gz，后续选项均可以直接回车填默认值。

```
\$ bce -c
\$ BOS Access Key ID [None]: Enter Your AK
\$ BOS Secret Access Key [None]: Enter Your SK
\$ Default region name [bj]: Enter Your region
\$ Default domain [bj.bcebos.com]: Enter Your host
```

5. 以将GPU云服务器上的 text.txt文件上传至mybucket为例，上传命令如下：

```
\$ bce bos cp text.txt bos:/mybucket/text.txt
```

3.4.4 BOS的其他功能及计费

参见BOS文档

3.5 申请GPU物理服务器

3.5.1 GPU物理服务器特点

- 通过BBC物理服务器形式提供GPU物理裸机服务，使整体性能达到最佳。
- 转为Training打造多GPU卡的高配物理机，可大幅提高Training效率。
- 可提供100G网络互联的GPU集群，支持动态扩展，用户独享高性能专有集群。

3.5.2 申请步骤

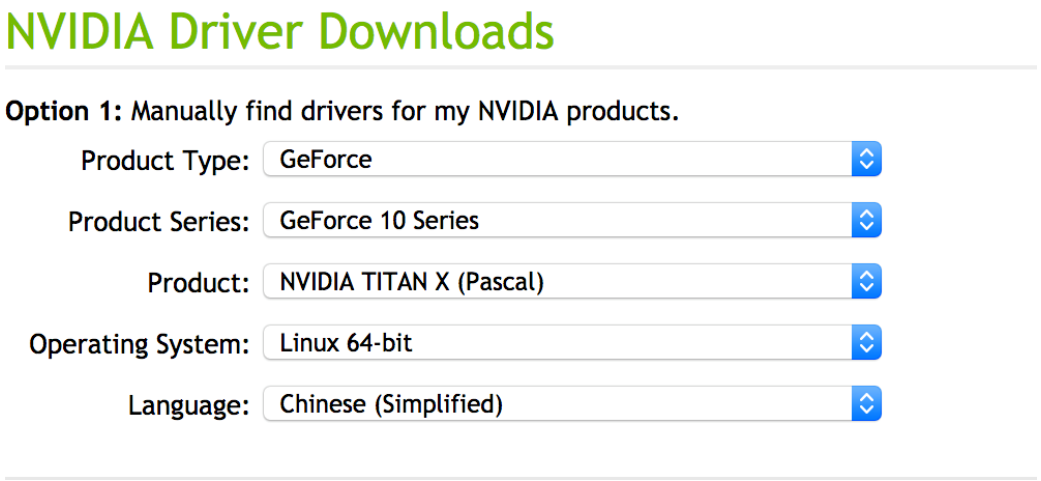
1. 登录百度智能云GPU官网。
2. 填写GPU物理服务器开通申请。
3. 收到您的申请后，我们将进入审核阶段，如通过审核，我们会根据您提供的联系方式及需求细节派专人与您联系，洽谈需求细节。如您已被告知开通GPU物理服务器的购买权限，可登录控制台界面，选择BCC物理服务器，点击创建实例后即可在购买实例环节购买GPU物理服务器。

第4章　常见问题

4.1　一般类问题

NVIDIA 深度学习开发卡应该使用哪种驱动？

请在Nvidia官网按照下图的选项进行驱动下载：



GPU服务器显卡型号有哪些？

关于GPU的显卡型号，您可以参考此文档：https://cloud.baidu.com/doc/GPU/ProductDescription/30.5C.32.78.79.86.1A.14.FD.99.3E.2A.D1.02.90.4E.67.B5.html#GPU.E4.BD.BF.E7.94.A8.E5.9C.BA.E6.99.AF

GPU服务器是否有GPU卡状态监控和报警机制，包括GPU使用率、内存使用率、温度、状态等？

使用状态监控前，需要您手动安装BCM-agent，安装完成后按照以下步骤找到GPU卡的监控信息 1. 登录百度智能云控制台选择云服务器BCC，点击实例进入实例列表页面。 2. 选择需要监控的实例名称，进入实例详情页面，然后点击监控按钮进入监控详情页面。 3. 在监控详情页面的最下侧，找到扩展监控项，然后选择GPU后，即可看到GPU卡的监控信息。

使用windows环境的虚机，远程桌面登录，Nvidia控制面板问题无法打开，会报如下错误



此问题是由于windows的远程桌面协议不支持GPU的信息查询，Windows 自带的远程连接（RDP）协议不支持 DirectX、OpenGL 等相关应用，您需要使用控制台的 VNC ，或其它支持的协议进行远程访问