GPU Cloud Computing

Introduction

Product Documentation

Contents

Introduction
Overview

Last updated：2018-11-26 18:03:28

GPU Cloud Computing is a rapid, stable and flexible computing service based on GPU, and is applicable to deep learning training/reasoning, graphic/image processing and scientific computing. It is managed easily in the same way as with standard CVM. With its powerful computing capability of processing mass data in a rapid manner, GPU Cloud Computing can effectively relieve the user's computing pressure, improving the efficiency and competitiveness of business processing.

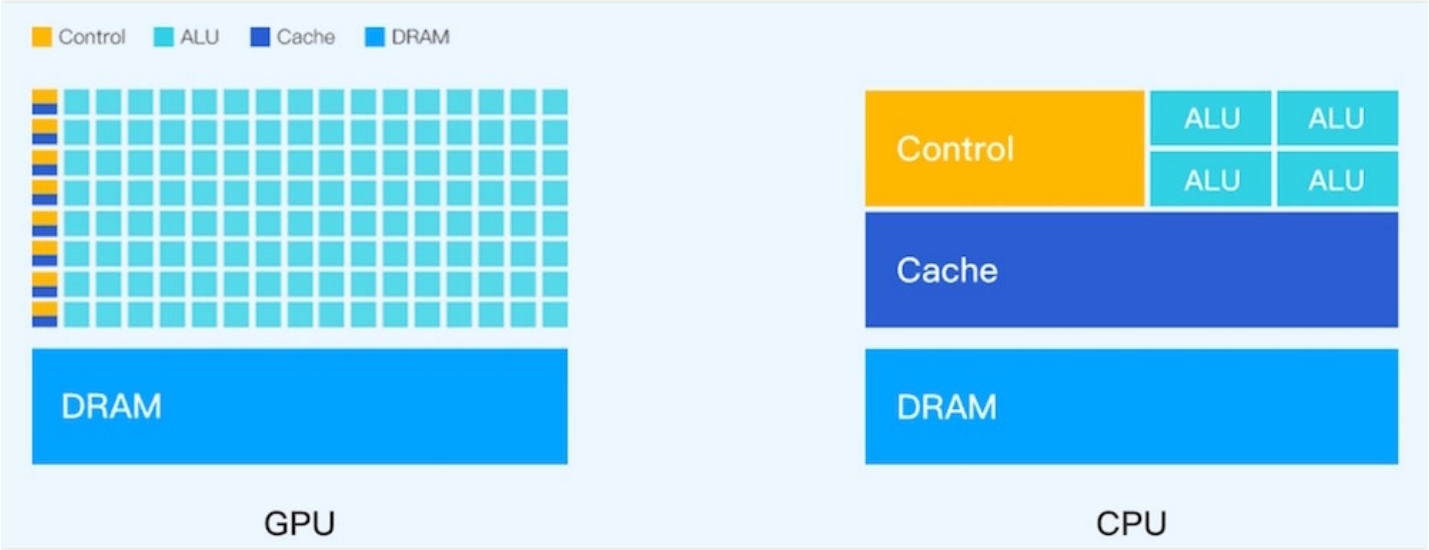
Why GPU Cloud Computing?

Comparison between GPU Cloud Computing and Self-built GPU Cloud Computing:

Advantages	Tencent Cloud GPU Instance	Custumer GPU-based Servers
Elastic	• In just a few minutes, you can easily obtain one or more high-performance computing instances. • It can be customized flexibly as needed, and upgraded to an instance specification with higher performance and larger capacity with just one click, to achieve rapid, smooth expansion, and satisfy the requirement for fast business development.	Fixed configuration makes it hard to satisfy the ever-changing requirements.
High-performance	• It supports GPU pass-through to make the best use of GPU performance. • Peak computing capacity for single machine: 125.6T Flops for single-precision floating point computing and 62.4T Flops for double-precision floating point computing.	• Users have to perform disaster recovery manually, depending on the robustness of hardware. • Single point of failure may occur on physical server. Data security is uncontrollable.

Advantages	Tencent Cloud GPU Instance	Custumer GPU-based Servers
Ease of Use	• Seamless connection with CVM, CLB and many other Tencent Cloud products. Private network traffic is free of charge. • Designed for ease of use, it is managed in the same way as with CVM, without the need to use jump server for login. • It provides clear guides on installation and deployment of GPU driver to make it easier for users to get started with it.	• Users must purchase installation management service to achieve automatic hardware expansion and driver installation. • Jump server is required for login with complicated operation procedures.
Secure	• Resources are completely isolated among different users to ensure the data security. • Complete security groups and network ACL settings allow you to control and securely filter the inbound and outbound network traffic to or from instances and subnets. • It can be seamlessly connected to Cloud Security, and has the basic protection and high defense services of Cloud Security equivalent to that of CVM.	• Resources are shared among different users, and data is not isolated. • Additional security protection services must be purchased.
Low-cost	• It supports the prepaid billing method. You can purchase physical servers without the need to make a huge one-off investment. • Hardware is updated with the mainstream GPU, eliminating the need to replace the hardware after each update. • With low server OPS cost, you can effectively reduce investment in infrastructure construction without the need to purchase and prepare hardware resources in advance.	• The cost for the server OPS is high. • Due to high power consumption of devices, hardware modification is required. • Higher IT OPS cost is required to guarantee the stability of service.

Comparison between GPU Instance and CPU Instance



Dimension	GPU	CPU
Kernels	Thousands of accelerated kernels (dual ENI, M40, and up to 6,144 accelerated kernels)	Dozens of kernels
Product features	1. Numerous efficient arithmetic logic units (ALU) support parallel processing 2. Massively parallel throughput can be achieved using multiple threads 3. Simple logic control	1. Complex logic control unit 2. Powerful ALUs 3. Simple logic control
Use Cases	Compute-intensive applications that support parallel processing	Applications with logic control and serial arithmetic

Benefits

Last updated：2019-09-20 15:53:27

Excellent and Reliable Performance

Accelerate computing in real time
GPU Cloud Computing provides superior computing capabilities:

- It adopts mainstream GPUs and CPUs.
- It offers a powerful single/double-precision floating point computing feature. Peak computing for a single machine is 125.6T Flops for single-precision floating point computing and 62/4T Flops for double-precision floating point computing.

Stable and Secure Services

GPU Cloud Computing provides secure and reliable network environment and perfect protection services: GPU Cloud Computing resides in a **25 GB** (or 10 GB) network environment, and provides a private network environment with low latency, offering outstanding computing capabilities.

- It can be integrated with CVM, VPC, CLB and other businesses, without additional management and OPS costs. Private network traffic is free of charge.
- Complete Network ACL settings allow you to control and securely filter the inbound and outbound network traffic to or from instances and subnets.
- It can be seamlessly connected to Cloud Security, and has the basic protection and high defense services of Cloud Security equivalent to that of CVM. For more information, see Learn more about network and security >>.

Rapid Deployment of Instances

The payment process is easy and ready to use for GPU Cloud Computing.
It is easy to get started with GPU Cloud Computing. Designed for ease of use, a GPU instance can be quickly built and managed in the same way as with CVM, without the need to use the jump server for login. For more information, see Quick Start >>.
GPU Cloud Computing can be seamlessly connected to multiple Tencent Cloud products, such as CLB and SSD. With clear guide on deployment and Installation of Nvidia Graphics Card Driver, you don't need to manually implement hardware expansion and driver installation.

Use Cases
GPU Computing Instances

Last updated：2019-08-06 17:01:09

Mass Computing Processing

GCC instances provide powerful computing capability to perform operations on mass data processing, such as search, big data recommendation, intelligent input method:

• With GCC instances, the data operation that used to take several days now only takes few hours.

• Cluster computing that used to be implemented using dozens of CCC instances is now completed with a single GCC instance.

Deep Learning Model

GCC instance serves as a training platform for deep learning:

1. GCC instance can directly accelerate the computing service and communicate externally.
2. GCC instance can be used in combination with CVM which provides computing platform for GCC instance.
3. COS provides GCC instance with cloud storage service for massive data.

GPU Rendering Instances

Last updated：2019-08-06 16:59:47

GCC instances powered by NVIDIA GPU (except GN2) not only support generic GPU computing but also can be used to do graphic process with NVIDIA GRID driver installed. It is suitable for GPU rendering scenarios, such as non-linear editing, video encoding/decoding, graphics acceleration visualization, 3D design.

Non-linear Editing

Non-linear editing is a modern editing method in film and TV post-production. To handle heavy graphic/image processing load, visualization GPU service is required for picture processing and visualization design. In addition, massive computing capacity, memory and storage are needed to store and process media asset data. With media asset data stored in the cloud, a project can be shared in the network editing environment. So, multiple users can work on the same project on their local machines at the same time, and perform tasks separately, such as editing, subtitling, adding special effects, coloring, and packaging.

Rendering

Rendering is the process of generating images from a model using a software, and is widely used in video, simulation, film and TV production and other fields. Rendering scenarios require GPU graphics card for graphics acceleration and real-time rendering, and also need massive computing capacity, memory and storage. High-performance computing and graphics rendering capabilities allow online graphics rendering, so as to greatly shorten the production cycle and improve the overall efficiency.

Remote Graphics Workstations

A remote graphics workstation connects to a server through private network for day-to-day work by separating server and client. Generally, the server is centrally deployed in the information data center to handle graphics workload using GPU graphics cards. Client terminals connect to the server by means of keyboard, mouse, monitor through private network for day-to-day work.

Instructions

Last updated：2018-11-26 18:04:17

GPU instance, as a special type of CVMs, is purchased, operated and maintained in the same way as with CVM. For more information, see CVM Documentation

To make better use of GPU instance, **read the following notes carefully**:

1. Back up Data

GPU Cloud Computing provides powerful computing capacities. GN2 and GN8 are mounted with local SSD. Make sure to back up data periodically to ensure data security and prevent data loss in extreme circumstances.
To ensure data security, you can also purchase and mount elastic cloud disks separately.

2. Renew in Time

You will receive an expiration notice 7 days before the expiration of the GPU instance. Renew in time if you want to continue using it. Otherwise, the instance will be shut down, disconnected and placed in the Recycle Bin once it expires. Be sure to renew in time or back up your data before expiration date.

3. Connect External Devices

Currently, you cannot use external hardware devices (such as hardware protection dongles, USB disks, external disks, and U-keys for banks) directly on GPU instances.

4. Upgrade Configuration

You cannot upgrade or degrade specification of GPU instances.

5. Prohibition Notes

- **Traffic traversal** is prohibited. The highest penalty may be termination, lockup and clean-up of your instances.
- DO NOT use GPU Cloud Computing to perform these activities against e-commerce websites such as Taobao.com: click farming (order, sales volume or advertisement), making **bogus website transactions**.