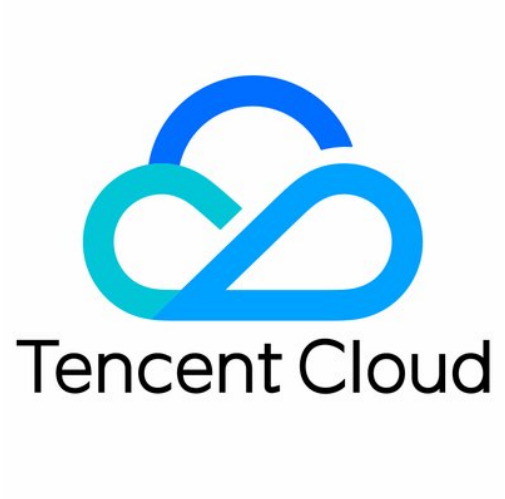
Cloud GPU Service

Product Introduction

Product Documentation



Contents

Product Introduction

Overview

Last updated：2024-01-11 17:11:13

Cloud GPU Service is a fast, stable and elastic GPU-based computing service. It's applicable to training or reasoning of deep learning, graphics and image processing, and scientific computing, etc. Cloud GPU Service is easily managed in the same way as the CVMs. With the powerful computing performance of processing massive data, it can effectively relieve the computing pressure of users and improve the efficiency and competitiveness of business processing.

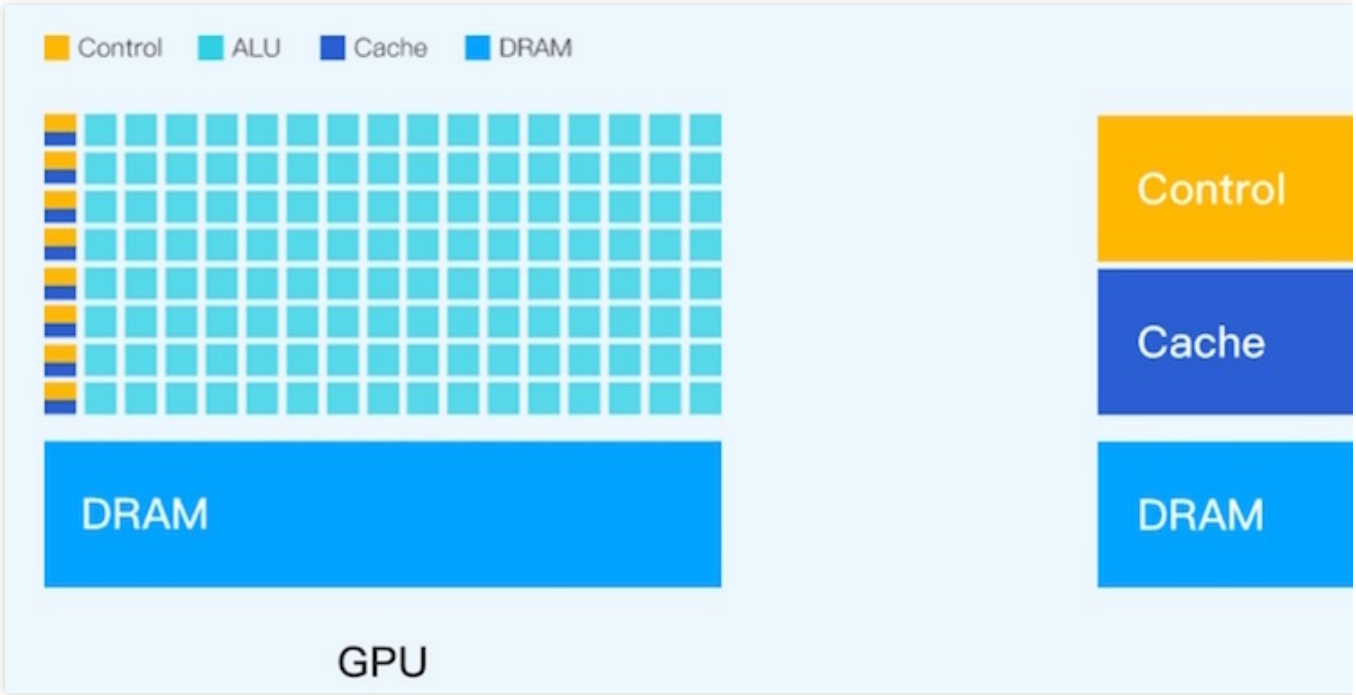
Why Cloud GPU Service?

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

Advantages	Cloud GPU Service	Self-built GPU Computing
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 most advantage of GPU. Peak computing capacity for a 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. A single point of failure may occur on physical servers. Data security is uncontrollable.
Easy-to-use	It can be seamlessly connected to 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 CVMs, 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 Tencent Cloud security serices, just the same as CVM.	Resources are shared among different users, and data is not isolated. Additional security protection services must be purchased.
Low-cost	It supports monthly subscription. 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.	High server investment and operating costs. 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 Cloud GPU Service and CPU Cloud Computing:



Dimension	GPU	CPU
Number of 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
Application Scenario	Compute-intensive applications that support parallel processing	Applications with logic control and serial arithmetic

Benefits

Last updated：2024-01-11 17:11:13

Excellent and Reliable Performance

Accelerate computing in real time

Cloud GPU Service 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

Cloud GPU Service provides secure and reliable network environment and perfect protection services:

Cloud GPU Service 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 Cloud GPU Service.

It is easy to get started with Cloud GPU Service. 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.

Cloud GPU Service 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：2024-01-11 17:11:13

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

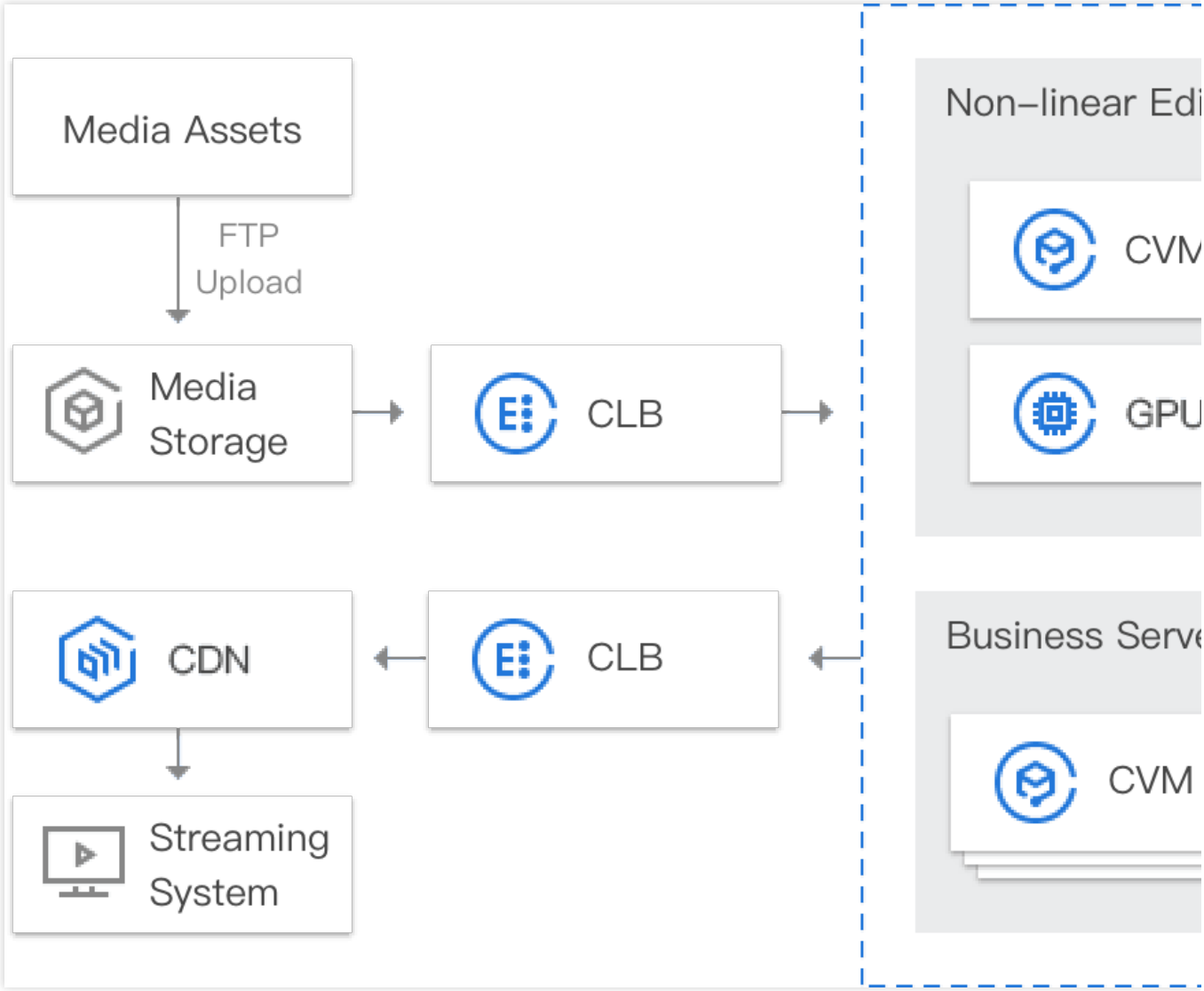
GCC instance serves as a training platform for deep learning:
GCC instance can directly accelerate the computing service and communicate externally.
GCC instance can be used in combination with CVM which provides computing platform for GCC instance.
COS provides GCC instance with cloud storage service for massive data.

GPU Rendering Instances

Last updated：2024-01-11 17:11:13

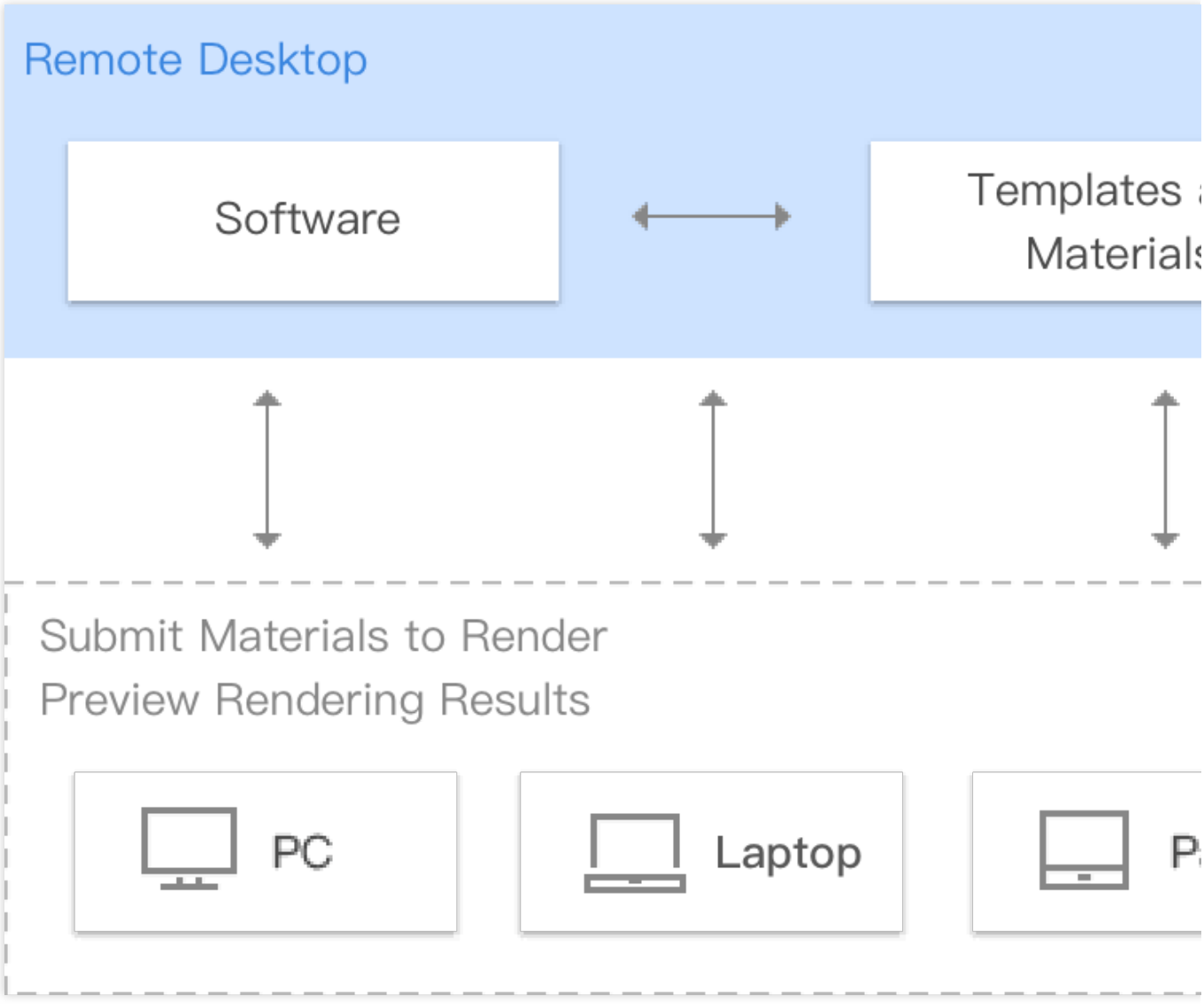
Non-linear Editing

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



Rendering

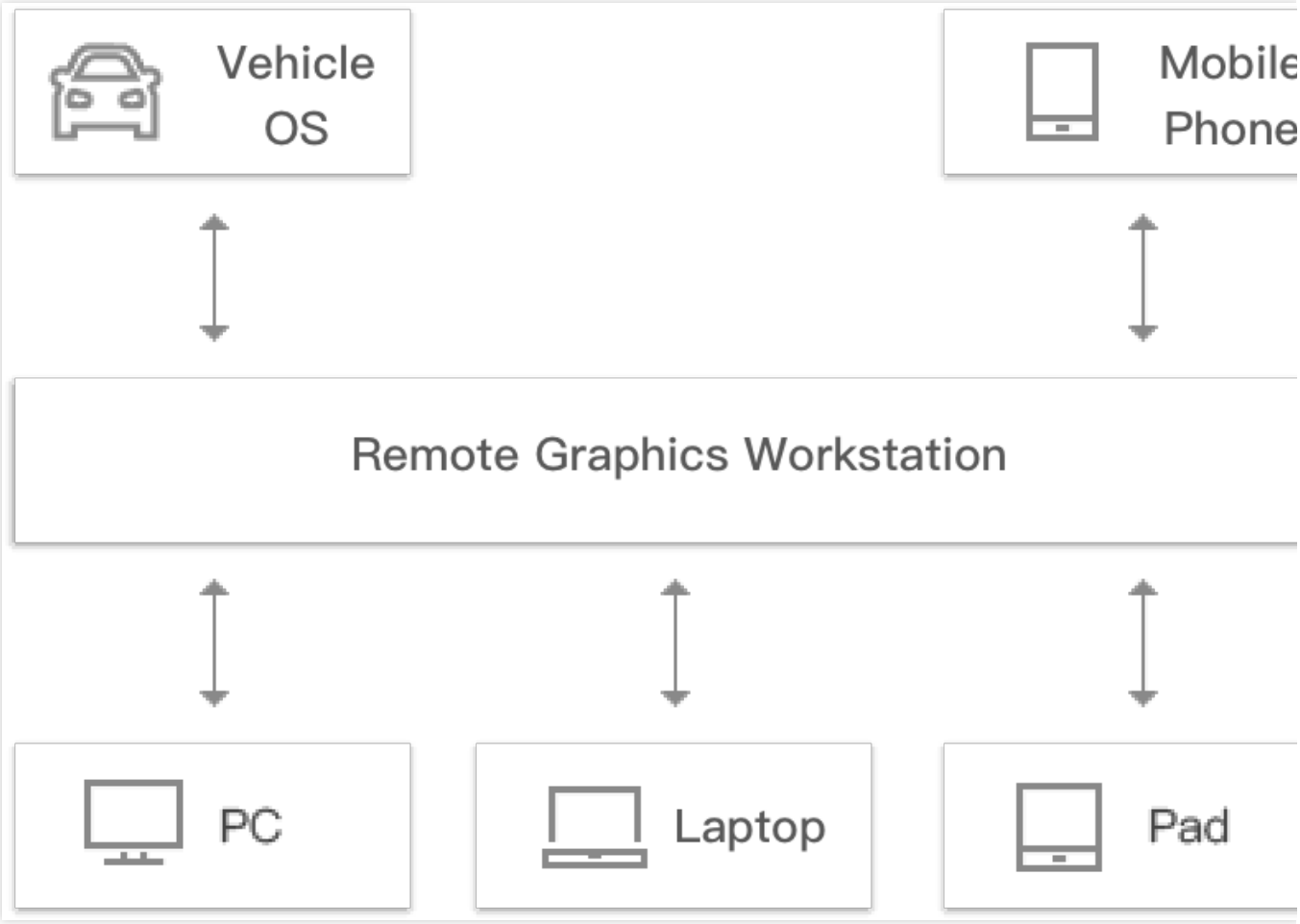
Rendering is the process of generating images from a model using software, and is widely used in fields such as video, simulation, and film/TV production. Rendering scenarios require GPUs for graphics acceleration and real-time rendering, and also require massive computing capacities, memory, and storage. High-performance computing and graphics rendering capabilities enable online graphics rendering, greatly shortening the production cycle and improving the overall efficiency.



Remote Graphics Workstation

A remote graphics workstation adopts the client/server (C/S) architecture. A client equipped with a keyboard, mouse, and monitor connects to the server through a dedicated network for daily work. Generally, the server is centrally

deployed in an information data center to handle graphics workload using GPUs.



Considerations

Last updated：2024-01-11 17:11:13

Note:

Cloud GPU Service allows you to purchase, manage, and maintain GPU instances in the same way as CVM. For more information, see CVM Documentation.

Note the following before using GPU instances.

1. Data Backup

Cloud GPU Service provides powerful computing capabilities. GN8 instances support local SSDs. Please back up data periodically to ensure data security and prevent data loss.

You can also purchase elastic cloud disks separately to further enhance data security and reliability.

2. Renewal

You will receive an alert 7 days before the expiry date of GPU instances. Make sure your account balance is sufficient. GPU instances will be shut down, disconnected and moved to the recycle bin after the expiration.

3. External Devices Mounting

You cannot mount external hardware devices (such as hardware dongle, USB flash drives, external disks, and U-keys) to a GPU instance.

4. Configuration Adjustment

GPU instances cannot be upgraded or downgraded.

5. Prohibition

Do not use GPU instances for **traffic traversal**. Tencent Cloud may suspend your service and repossess the resources if related behaviors are detected.

Do not use GPU instances to engage in **fraudulent online transactions** such as click farming (order, sales volume or advertisement) on e-commerce websites.