Cloud GPU Service

FAQs

Product Documentation

Service Statement

This document is intended to provide users with general information about Tencent Cloud's products and services only and does not form part of Tencent Cloud's terms and conditions. Tencent Cloud's products or services are subject to change. Specific products and services and the standards applicable to them are exclusively provided for in Tencent Cloud's applicable terms and conditions.

FAQs

Last updated：2024-01-11 17:11:13

What is NVIDIA Tesla?

NVIDIA Tesla is a new product line introduced by NVIDIA following the launch of professional acceleration card QUADRO and the entertainment graphics card GeForce series, which is mainly used for scenarios that require high performance computing in a broad range of scientific research. With NVIDIA® Tesla® GPU accelerator, it can handle the workloads that require super strict HPC in ultra-large data centers faster.

What is computing acceleration?

Computing acceleration is used to perform floating-point computing and graphic processing with a hardware accelerator or a coprocessor, which is more efficient than using a software running on CPU. Tencent Cloud three two computing acceleration models: GPU computing (GN2, GN8) for generic computing, and GPU rendering GA2 for graphics-intensive applications.

What are the advantages of GPU over CPU?

GPU has more arithmetic logic units (ALU) than CPU and supports large-scale multi-threaded parallel computing.

When should I use GPU instances?

GPU instances are most suitable for parallel applications requiring high concurrency, such as workloads that use thousands of threads. When a great deal of computation is required for graphics processing where each task is relatively small, a group of operations to be performed form a pipeline. The throughput of this pipeline is more important than the latency of a single operation. To build an application that makes full use of this parallelism, you need to master the expertise of GPU devices, and to learn how to program for various graphical APIs (DirectX, OpenGL) or GPU computing programming models (CUDA, OpenCL).

How are GPU instances billed?

GPU instances are billed per usage. The bills are calculated down to the second and settled on an hourly basis. You can purchase and release the instances any time. GPU instances are applicable to scenarios where the demand for devices fluctuates dramatically, such as flash sale on an e-commerce site. For more information, see Pricing Overview.

Can I upgrade/degrade GPU instance configuration?

No. GPU instance upgrade and degrade are not supported for now.

What is local SSD?

Local SSD is a local storage on the physical machine where the CVM resides in. It provides instances with block-level data access capability with a low latency, high random IOPS, and high I/O throughput. As GPU instances are mounted with local SSDs, you cannot upgrade hardware (CPU and memory), but only the bandwidth.

Can GPU instances access CVM instances?

Yes. GPU instances have private IPs and public IPs, so they can communicate with other Tencent Cloud products such as CVMs.