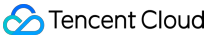
Cloud GPU Service

Instance Types

Copyright Notice

Trademark Notice



Service Notice

This document provides an overview of the as-is details of Tencent Cloud's products and services in their entirety or part. The descriptions of certain products and services may be subject to adjustments from time to time.

The commercial contract concluded by you and Tencent Cloud will provide the specific types of Tencent Cloud products and services you purchase and the service standards. Unless otherwise agreed upon by both parties, Tencent Cloud does not make any explicit or implied commitments or warranties regarding the content of this document.

Contact Us

We are committed to providing personalized pre-sales consultation and technical after-sale support. Don't hesitate to contact us at 4009100100 or 95716 for any inquiries or concerns.

Contents

Instance Types
Compute-Optimized Instances

Last updated: 2026-03-27 18:20:57

GPU compute-optimized instances deliver powerful computing capabilities, making them ideal for high real-time, high-concurrency, and large-scale computational scenarios. They are widely used in general GPU computing scenarios such as deep learning and scientific computing. Tencent Cloud GPU Service offers fast, stable, and flexible computing services, **managed in the same way** as CVM.

> **Note:**
> If your GPU instance is used for 3D graphics rendering tasks, it is recommended to use rendering instances pre-configured with vDWS/vWS licenses and installed GRID drivers. his eliminates the need to manually configure the GPU graphics and image processing environment.

Overview of Compute-Optimized Instances

The Cloud GPU Service compute-optimized series offers the following instances:

Purchase	Instance	GPU Type	Available Regions
Recommended	PNV4	NVIDIA A10	Guangzhou, Shanghai, Nanjing, and Beijing
	GT4	NVIDIA A100	Guangzhou, Shanghai, Nanjing, and Beijing
	GN10Xp	NVIDIA V100	Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Singapore, Bangkok, Seoul, Tokyo, Frankfurt, and Silicon Valley
	GN7	NVIDIA T4	Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Hong Kong (China), Singapore, Bangkok, Jakarta, Seoul, Tokyo, Frankfurt, Silicon Valley, Virginia, and São Paulo
	GN7vi	NVIDIA T4	Shanghai and Nanjing
	PTX1	Tencent Zixiao	Beijing, Shanghai, and Nanjing

		C100	
Available	GI3X	NVIDIA T4	Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, and Chongqing
	GN10X	NVIDIA V100	Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Singapore, Frankfurt, and Silicon Valley
	GN8	NVIDIA P40	Guangzhou, Shanghai, Beijing, Chengdu, Chongqing, Hong Kong (China), and Silicon Valley
	GN6 GN6S	NVIDIA P4	• GN6: Chengdu • GN6S: Guangzhou, Shanghai, and Beijing

Recommendations of Compute-Optimized Instance Selection

Tencent Cloud offers a diverse range of GPU compute instances to meet the needs of various business use cases. See the table below and select the appropriate compute instance based on your specific requirements.
The **recommendations of Cloud GPU Service compute-optimized instances selection** are shown in the table below, where ✓ indicates support and ★ indicates a recommendation.

Feature/Instance	PNV4	GT4	GN10Xp	GN7	GN7vi	GI3X
Graphics and image processing	✓	–	✓	✓	✓	✓
Video encoding and decoding	✓	–	✓	★	★	★
Deep learning training	✓	★	★	✓	✓	✓
Deep learning inference	★	✓	★	★	★	★

Scientific computing	–	★	★	–	–	–

Note:

- The above recommended use cases are for reference only. Choose according to your actual needs.
- For NVIDIA series GPU instances used for general computing, you need to install the Tesla Driver and CUDA. For installation instructions, see Manually Installing Tesla Driver – Linux and Installing CUDA Driver.
- For NVIDIA series GPU instances used for 3D graphics rendering tasks (such as high-performance graphics processing and video encoding/decoding), you need to install the GRID Driver and configure the License Server. For installation instructions, see Applying for a License and Installing the GRID Driver.

Instance Specifications

Compute-Optimized PNV4

Compute-optimized PNV4 is suitable not only for general GPU computing scenarios such as deep learning but also for graphics and image processing tasks, including 3D rendering and video encoding/decoding.

Applicable Scenarios

High cost-effectiveness, suitable for the following scenarios:

- Reasoning scenarios and small-scale training scenarios for deep learning. For example:
  - AI reasoning for large-scale deployment
  - Small-scale training for deep learning
- Graphics and image processing scenario. For example:
  - Graphics and image processing
  - Video encoding and decoding
  - Graphic databases

Hardware Specifications

- **GPU:** NVIDIA® A10 (FP32 31.2 TFLOPS, TF32 62.5 TFLOPS, FP16 125 TFLOPS, INT8 250 TOPS).
- **CPU:** 2.55 GHz AMD EPYC™ Milan processor with a turbo boost of 3.5 GHz.

- **Memory:** Equipped with eight-channel DDR4.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
PNV4.7XLARGE116	NVIDIA A10 * 1	24GB * 1	28	116	13	2.3 million	28
PNV4.14XLARGE232	NVIDIA A10 * 2	24GB * 2	56	232	25	4.7 million	48
PNV4.28XLARGE466	NVIDIA A10 * 4	24GB * 4	112	466	50	9.5 million	48
PNV4.56XLARGE932	NVIDIA A10 * 8	24GB * 8	224	932	100	19 Million	48

Compute-Optimized GT4

Compute-optimized GT4 is suitable for general GPU computing scenarios such as deep learning and scientific computing.

Applicable Scenarios

GT4 offers powerful double-precision floating-point computing capabilities, making it ideal for large-scale deep learning training, inference, and scientific computing scenarios. For example:

- Deep learning
- High-performance databases
- Computational fluid dynamics
- Computational finance
- Seismic analysis

- Molecular modeling
- Genomics and more

Hardware Specifications

- **GPU:** NVIDIA® A100 NVLink 40GB (FP64 9.7 TFLOPS, FP32 19.5 TFLOPS, 600GB/s NVLink).
- **CPU:** 2.6 GHz AMD EPYC™ ROME processor with a turbo boost of 3.3 GHz.
- **Memory:** Equipped with eight-channel DDR4.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Supports up to 50 Gbps private network bandwidth with exceptional packet processing capabilities. Network performance aligns with instance specifications. public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Per Second (PPS)	Number of Queues
GT4.4XLARGE96	NVIDIA A100 * 1	40GB * 1	16	96	5	1.2 million	4
GT4.8XLARGE192	NVIDIA A100 * 2	40GB * 2	32	192	10	2.35 million	8
GT4.20XLARGE474	NVIDIA A100 * 4	40GB * 4	82	474	25	6 million	16
GT4.41XLARGE948	NVIDIA A100 * 8	40GB * 8	164	948	50	12 million	32

Note:
GPU driver: NVIDIA A100 series requires NVIDIA Tesla drivers version 450 or later. For driver version details, see NVIDIA official documentation.

Compute-Optimized GN10Xp

Compute-optimized GN10Xp is suitable not only for general GPU computing scenarios such as deep learning and scientific computing but also for graphics and image processing tasks, including 3D rendering and video encoding/decoding.

Applicable Scenarios

GN10Xp offers powerful double-precision floating-point computing capabilities, making it suitable for the following scenarios:

- Large-scale deep learning training, inference, and scientific computing scenarios. For example:
  - Deep learning
  - High-performance databases
  - Computational fluid dynamics
  - Computational finance
  - Seismic analysis
  - Molecular modeling
  - Genomics and more
- Graphics and image processing scenario. For example:
  - Graphics and image processing
  - Video encoding and decoding
  - Graphic databases

Hardware Specifications

- **CPU:** The GN10Xp is configured with an Intel® Xeon®Platinum 8255C CPU with a CPU clock speed of 2.5 GHz.
- **GPU:** NVIDIA® Tesla® V100 NVLink 32GB (15.7 TFLOPS single-precision floating-point performance, 7.8 TFLOPS double-precision floating-point performance, 125 TFLOPS Tensor Core for deep learning acceleration, 300 GB/s NVLink bandwidth).
- **Memory:** DDR4 with a memory speed of up to 2666 MT/s.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory	Private Network	Packet Tx/Rx (PPS)	Number of Queues

				(GiB)	Bandwidth (Gbps)		
GN10Xp.2XLARGE40	NVIDIA V100 * 1	32GB * 1	10	40	3	800 thousand	2
GN10Xp.5XLARGE80	NVIDIA V100 * 2	32GB * 2	20	80	6	1.5 million	5
GN10Xp.10XLARGE160	NVIDIA V100 * 4	32GB * 4	40	160	12	2.5 million	10
GN10Xp.20XLARGE320	NVIDIA V100 * 8	32GB * 8	80	320	24	4.9 million	16

Compute-Optimized GN7

NVIDIA GN7 instances are suitable not only for general GPU computing scenarios such as deep learning but also for graphics and image processing tasks, including 3D rendering and video encoding/decoding.

Applicable Scenarios

High cost-effectiveness, suitable for the following scenarios:

- Reasoning scenarios and small-scale training scenarios for deep learning. For example:
  - AI reasoning for large-scale deployment
  - Small-scale training for deep learning
- Graphics and image processing scenario. For example:
  - Graphics and image processing
  - Video encoding and decoding
  - Graphic databases

Hardware Specifications

- **CPU:** Intel® Xeon® Platinum 8255C CPU with a CPU clock speed of 2.5 GHz.
- **GPU:** NVIDIA® Tesla® T4 (8.1 TFLOPS single-precision floating-point performance, 130 INT8 TOPS, 260 INT4 TOPS).
- **Memory:** DDR4 with a memory speed of up to 2666 MT/s.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.

- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GN7.2XLARGE32	NVIDIA T4 * 1	16GB * 1	8	32	3	600 thousand	8
GN7.5XLARGE80	NVIDIA T4 * 1	16GB * 1	20	80	7	1.4 million	10
GN7.8XLARGE128	NVIDIA T4 * 1	16GB * 1	32	128	10	2.4 million	16
GN7.10XLARGE160	NVIDIA T4 * 2	16GB * 2	40	160	13	2.8 million	20
GN7.20XLARGE320	NVIDIA T4 * 4	16GB * 4	80	320	25	5.6 million	32

Video-Enhanced GN7vi

NVIDIA GN7vi instances are built on the GN7 foundation and configured with Tencent's proprietary Media Video Fusion AI technology, including a top speed codec engine and image quality enhancement toolkit. These instances are ideal for on-demand and live streaming scenarios. With GN7vi, you can utilize Tencent Cloud's proprietary top speed codec and AI-based image quality enhancement features directly within the instance.

> **Note:**
> If you would like to learn more about the GPU Video-Enhanced GN7vi product, visit the product consulting page to provide your contact details. Our dedicated team will get in touch with you.

Hardware Specifications

- **CPU:** Intel® Xeon® Platinum 8255C CPU with a CPU clock speed of 2.5 GHz.

- **GPU:** NVIDIA® Tesla® T4 (8.1 TFLOPS single-precision floating-point performance, 130 INT8 TOPS, 260 INT4 TOPS).
- **Memory:** DDR4 with a memory speed of up to 2666 MT/s.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GN7vi.5XLARGE80	NVIDIA T4 * 1	16GB * 1	20	80	6	1.4 million	20
GN7vi.10XLARGE160	NVIDIA T4 * 2	16GB * 2	40	160	13	2.8 million	32
GN7vi.20XLARGE320	NVIDIA T4 * 4	16GB * 4	80	320	25	5.6 million	32

Inference-Optimized GI3X

NVIDIA GI3X instances are suitable not only for general GPU computing scenarios such as deep learning but also for graphics and image processing tasks, including 3D rendering and video encoding/decoding.

Applicable Scenarios

High cost-effectiveness, suitable for the following scenarios:

- Reasoning scenarios and small-scale training scenarios for deep learning. For example:
  - AI reasoning for large-scale deployment
  - Small-scale training for deep learning
- Graphics and image processing scenario. For example:
  - Graphics and image processing
  - Video encoding and decoding
  - Graphic databases

Hardware Specifications

- **CPU:** 2.6 GHz AMD EPYC™ ROME processor with a turbo boost of 3.3 GHz.
- **GPU:** NVIDIA® Tesla® T4 (8.1 TFLOPS single-precision floating-point performance, 130 INT8 TOPS, 260 INT4 TOPS).
- **Memory:** Equipped with eight-channel DDR4, delivering stable memory performance.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

GI3X instances offer the following configurations:

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GI3X.8XLARGE64	NVIDIA T4 * 1	16GB * 1	32	64	5	1.4 million	8
GI3X.22XLARGE226	NVIDIA T4 * 2	16GB * 2	90	226	13	3.75 million	16
GI3X.45XLARGE452	NVIDIA T4 * 4	16GB * 4	180	452	25	7.5 million	32

Compute-Optimized GN10X

Compute-optimized GN10X is suitable not only for general GPU computing scenarios such as deep learning and scientific computing but also for graphics and image processing tasks, including 3D rendering and video encoding/decoding.

Applicable Scenarios

GN10X offers powerful double-precision floating-point computing capabilities, making it applicable to the following scenarios:

- Large-scale deep learning training, inference, and scientific computing scenarios. For example:
  - Deep learning

  - High-performance databases
  - Computational fluid dynamics
  - Computational finance
  - Seismic analysis
  - Molecular modeling
  - Genomics and more
- Graphics and image processing scenario. For example:
  - Graphics and image processing
  - Video encoding and decoding
  - Graphic databases

Hardware Specifications

- **CPU:** The GN10X is configured with an Intel® Xeon® Gold 6133 CPU with a CPU clock speed of 2.5 GHz.
- **GPU:** NVIDIA® Tesla® V100 NVLink 32GB (15.7 TFLOPS single-precision floating-point performance, 7.8 TFLOPS double-precision floating-point performance, 125 TFLOPS Tensor Core for deep learning acceleration, 300 GB/s NVLink bandwidth).
- **Memory:** DDR4 with a memory speed of up to 2666 MT/s.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

GN10X instances offer the following configurations:

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GN10X.2XLARGE40	NVIDIA V100 * 1	32GB * 1	8	40	3	800 thousand	2

GN10X.4XLARGE80	NVIDIA V100 * 2	32GB * 2	18	80	7	1.5 million	4
GN10X.9XLARGE160	NVIDIA V100 * 4	32GB * 4	36	160	13	2.5 million	9
GN10X.18XLARGE320	NVIDIA V100 * 8	32GB * 8	72	320	25	4.9 million	16

Compute-Optimized GN8

NVIDIA GN8 instances are suitable not only for general GPU computing scenarios such as deep learning but also for graphics and image processing tasks, including 3D rendering and video encoding/decoding.

Applicable Scenarios

Applicable to the following scenarios:

- Deep learning inference and training scenarios.
  For example:
  - High-throughput AI inference
  - Deep learning
- Graphics and image processing scenario. For example:
  - Graphics and image processing
  - Video encoding and decoding
  - Graphic databases

Hardware Specifications

- **CPU:** Intel® Xeon® E5-2680 v4 CPU with a CPU clock speed of 2.4 GHz.
- **GPU:** NVIDIA® Tesla® P40 (12 TFLOPS single-precision floating-point performance, 47 INT8 TOPS).
- **Memory:** DDR4 with a memory speed of up to 2666 MT/s.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

GN8 instances offer the following configurations:

Specification	GPU	GPU Memor	vCPU	Memo	Private	Packet	Number of

		y		ry (GiB)	Network Bandwidth (Gbps)	Tx/Rx (PPS)	Queues
GN8.LARGE56	NVIDIA P40 * 1	24GB * 1	6	56	1.5	450 thousand	8
GN8.3XLARGE112	NVIDIA P40 * 2	24GB * 2	14	112	2.5	500 thousand.	8
GN8.7XLARGE224	NVIDIA P40 * 4	24GB * 4	28	224	5	700 thousand	14
GN8.14XLARGE448	NVIDIA P40 * 8	24GB * 8	56	448	10	700 thousand	28

Compute-Optimized GN6/GN6S

NVIDIA GN6/GN6S instances are suitable not only for general GPU computing scenarios such as deep learning but also for graphics and image processing tasks, including 3D rendering and video encoding/decoding.

Applicable Scenarios

High cost-effectiveness, suitable for the following scenarios:

- Reasoning scenarios and small-scale training scenarios for deep learning. For example:
  - AI reasoning for large-scale deployment
  - Small-scale training for deep learning
- Graphics and image processing scenario. For example:
  - Graphics and image processing
  - Video encoding and decoding
  - Graphic databases

Hardware Specifications

- **CPU:** GN6 is configured with Intel® Xeon® E5-2680 v4 CPU with a base frequency of 2.4 GHz. GN6S is configured with Intel® Xeon® Silver 4110 CPU with a CPU clock speed of 2.1 GHz.
- **GPU:** NVIDIA® Tesla® P4 (5.5 TFLOPS single-precision floating-point performance, 22 INT8 TOPS).
- **Memory:** DDR4 with a memory speed of up to 2666 MT/s.
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

GN6/GN6S instances offer the following configurations:

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GN6.7XLARGE48	NVIDIA P4 * 1	8GB * 1	28	48	5	1.2 million	14
GN6.14XLARGE96	NVIDIA P4 * 2	8GB * 2	56	96	10	1.2 million	28
GN6S.LARGE20	NVIDIA P4 * 1	8GB * 1	4	20	5	500 thousand.	8
GN6S.2XLARGE40	NVIDIA P4 * 2	8GB * 2	8	40	9	800 thousand	8

NPU Compute-Optimized PTX1

PTX1 is suitable for deep learning inference workloads and demonstrates excellent performance in scenarios such as CV, OCR, and ASR.

> ⚠ **Note:**

> This instance is temporarily available on an allowlist basis. Contact the pre-sales online to request access permissions for purchasing the instance.

Applicable Scenarios

- Cost-effective and ideal for deep learning inference scenarios. For example:
  - AI reasoning for large-scale deployment
  - Image analysis
  - Text recognition
  - Recognize audio content

Hardware Specifications

- **CPU:** 2.55 GHz AMD EPYCTM Milan processor with a turbo boost of 3.5 GHz.
- **NPU:** Tencent Zixiao C100 (120 TFLOPS FP16)
- **Storage:** You can select cloud block storage type. If you need to scale out, you can create an auto-scaling cloud disk and mount it.
- **Network:** Supports up to 100 Gbps private network bandwidth with exceptional packet processing capabilities. Network performance aligns with instance specifications. Public network can be configured as needed.

PTX1 instances offer the following configurations:

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
PTX1.7XLARGE116	Tencent Zixiao C100 * 1	16GB * 1	28	116	13	2.3 million	28
PTX1.14XLARGE232	Tencent Zixiao C100 * 2	16GB * 2	56	232	25	4.7 million	48
PTX1.28XLARGE464	Tencent Zixiao C100 * 4	16GB * 4	112	464	50	9.5 million	48

PTX1.56XLARGE928	Tencent Zixiao C100 * 8	16GB * 8	224	928	100	19 Million	48

Rendering Instances

Last updated: 2025-04-11 15:45:02

GPU rendering instances are designed for traditional GPU-based graphics and image processing use cases, such as 3D rendering. Tencent Cloud offers fast, stable, and auto scaling computing services **managed consistently** with CVM.

Applicable Scenarios

Suitable for high-performance graphics processing and 3D rendering. For example:

- Non-linear editing
- Game streaming
- Cloud phone
- Cloud virtual desktop
- CloudXR
- Graphics and image processing

Overview of Rendering Instances

The Cloud GPU Service rendering series offers the following instances:

Instance	GPU Types	Available Image	Available Regions
GA3	GA01	Windows Server 2019 Datacenter Edition 64-bit	Guangzhou
GNV4v	NVIDIA A10	• Windows Server 2019 Datacenter Edition 64-bit Chinese Edition with GRID 16.2 • Windows Server 2022 Datacenter Edition 64-bit Chinese Edition with GRID 16.2	Guangzhou, Shanghai, and Beijing
GNV4	NVIDIA A10	• CentOS 7.2 and above • Ubuntu 16.04 – 20.04 • Windows Server 2019 Datacenter Edition 64-bit Chinese Edition with GRID 16.2 • Windows Server 2022 Datacenter Edition 64-bit	Guangzhou, Shanghai, Beijing, Nanjing, and Chongqing

		Chinese Edition with GRID 16.2	
GN7vw	NVIDIA T4	• TencentOS Server 3.1 (TK4) GRID16.2 • Ubuntu Server 20.04 LTS 64-bit with GRID 16.2 • Ubuntu Server 22.04 LTS 64-bit with GRID 16.2 • Windows Server 2019 Datacenter Edition 64-bit Chinese Edition with GRID 16.2 • Windows Server 2022 Datacenter Edition 64-bit Chinese Edition with GRID 16.2	Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Hong Kong (China), Singapore, Bangkok, Frankfurt, Seoul, Tokyo, Silicon Valley, and Virginia
GI1	Intel SG1	• CentOS 7.6 64-bit + SG1-pv1.3 • CentOS 7.6 64-bit + SG1-pv1.4 • CentOS 7.6 64-bit + SG1-pv1.5 • CentOS 7.6 64-bit + SG1-pv1.6	Guangzhou, Nanjing, and Chongqing

Rendering Instance Selection Recommendations

Tencent Cloud offers a diverse range of GPU compute instances to meet the needs of various business use cases. See the table below and select the appropriate compute instance based on your specific requirements.
The **selection recommendations for Cloud GPU Service rendering instances** are shown in the table below, where ✓ indicates support and ★ indicates a recommendation.

Feature/Instance	GA3	GNV4v	GNV4	GN7vw	GI1
Graphics and image processing	★	★	★	★	★
Video encoding and decoding	★	★	★	★	★
Deep learning training	–	–	✓	–	–
Deep learning inference	–	–	✓	–	–
Scientific computing	–	–	–	–	–

⚠ **Notes:**

- The above recommended use cases are for reference only. Choose according to your actual needs.
- For NVIDIA-series GPU instances used for 3D rendering tasks, including high-performance graphics processing and video encoding/decoding, you need to install the GRID Driver and configure the License Server. For installation instructions, see Installing NVIDIA GRID Driver. GNV4v, GNV4, and GN7vw instances can opt for specific images with the GRID Driver pre-installed, eliminating the need to install the GRID Driver and configure the License Server separately.
- The GNV4v, GNV4, and GN7vw instance families provide vGPU instance types that support vDWS/vWS and are compatible with graphics APIs such as DirectX and OpenGL.

Instance Specifications

Rendering GA3

GA3 is suitable for graphics and image processing scenarios, including 3D rendering and video encoding/decoding.

⚠ **Notes:**

- This instance is temporarily available on an allowlist basis. Contact Pre-Sales Online Consultation to request access permissions for purchasing the instance.
- This instance requires a pre-installed GA01 driver image for proper functionality. On the purchase page, click **Image Market**, search for Pre-installed GA01 GPU Driver and select it. For more details, see Using Pre-installed GPU Driver Images.

Hardware Specifications

- **GPU:** Tencent Cloud Star Lake GA01.
- **CPU:** 2.55 GHz AMD EPYC™ Milan processor with a turbo boost of 3.5 GHz.
- **Memory:** Equipped with eight-channel DDR4.
- **Storage:** You can select Cloud Block Storage Type. If you need to scale out, you can create an auto scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GA3.LARGE10	GA01 * 1/6	32GB * 1/6	4	10	1.5	150 thousand	4
GA3.2XLARGE14	GA01 * 1/4	32GB * 1/4	8	14	2	350 thousand	8
GA3.4XLARGE30	GA01 * 1/2	32GB * 1/2	16	30	4	750 thousand	16
GA3.8XLARGE62	GA01 * 1	32GB * 1	32	62	7	1.5 million	32

Rendering GNV4v

NVIDIA GNV4v instances are rendering instances configured with a vDWS License server and pre-installed GRID drivers. They are well-suited for graphics and image processing scenarios, including 3D rendering and video encoding/decoding. These instances eliminate the need for manual GPU graphics and image processing environment configuration.

> **Notes:**
> This instance is currently available on an allowlist basis. Contact Pre-Sales Online Consultation to request access permissions for purchasing the instance.

Hardware Specifications

- **GPU:** NVIDIA® A10(FP32 31.2 TFLOPS, INT8 250 TOPS).
- **CPU:** 2.55 GHz AMD EPYC™ Milan processor with a turbo boost of 3.5 GHz.
- **Memory:** Equipped with eight-channel DDR4.
- **Storage:** You can select Cloud Block Storage Type. If you need to scale out, you can create an auto scaling cloud disk and mount it.

- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GNV4v.LARGE24	NVIDIA A10 * 1/4	24GB * 1/4	6	24	3	500 thousand.	6
GNV4v.3XLARGE58	NVIDIA A10 * 1/2	24GB * 1/2	14	58	7	1.1 million	14
GNV4v.7XLARGE116	NVIDIA A10 * 1	24GB * 1	Day 28	116	13	2.3 million	Day 28

Rendering GNV4

NVIDIA GNV4 instances are rendering instances configured with a vDWS License server and pre-installed GRID drivers. They include public images with GRID support, such as Windows Server 2022 Datacenter Edition 64-bit Chinese Edition with GRID 16.2. These instances are well-suited for graphics and image processing scenarios, including 3D rendering and video encoding/decoding. By using these instances, you can eliminate the need for manual GPU graphics and image processing environment configuration.

Hardware Specifications

- **GPU:** NVIDIA® A10 (31.2 TFLOPS single-precision floating-point performance, 250 INT8 TOPS, 500 INT4 TOPS).
- **CPU:** Intel® Xeon® Cooper Lake processor with a fundamental frequency of 3.4 GHz and a turbo frequency of 3.8 GHz.
- **Memory**: Equipped with six-channel DDR4 memory.
- **Storage:** You can select Cloud Block Storage Type. If you need to scale out, you can create an auto scaling cloud disk and mount it.

- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GNV4.3XLARGE44	NVIDIA A10 * 1	24GB	12	44	2	500 thousand.	12
GNV4.6XLARGE88	NVIDIA A10 * 2	24GB * 2	24	88	4	1 million	24
GNV4.12XLARGE176	NVIDIA A10 * 4	24GB * 4	48	176	7	2.1 million	32
GNV4.24XLARGE352	NVIDIA A10 * 8	24GB * 8	96	352	13	4.2 million	32
GNV4.48XLARGE704	NVIDIA A10 * 16	24GB * 16	192	704	25	8.5 million	32

Rendering GN7vw

NVIDIA GN7vw instances are rendering instances built on the GN7 foundation, configured with a vDWS License server and pre-installed GRID drivers. They are well-suited for graphics and image processing scenarios, including 3D rendering and video encoding/decoding. These instances eliminate the need for manual GPU graphics and image processing environment configuration.

> ⚠ **Note:**
> This instance is currently available on an allowlist basis. Contact Pre-Sales Online Consultation to request access permissions for purchasing the instance.

Hardware Specifications

- **GPU:** NVIDIA® Tesla® T4 (FP32 8.1 TFLOPS, 130 INT8 TOPS, 260 INT4 TOPS).

- **CPU:** Intel® Xeon® Platinum 8255C CPU with a CPU clock speed of 2.5 GHz.
- **Memory:** DDR4 with a memory speed of up to 2666 MT/s.
- **Storage:** You can select Cloud Block Storage Type. If you need to scale out, you can create an auto scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GN7vw.LARGE16	NVIDIA T4 * 1/4	16GB * 1/4	4	16	2	500 thousand.	8
GN7vw.2XLARGE32	NVIDIA T4 * 1/2	16GB * 1/2	8	32	4	800 thousand	8
GN7vw.4XLARGE64	NVIDIA T4 * 1	16GB * 1	16	64	7	1.5 million	8

Rendering GI1

GPU rendering GI1 instances are equipped with H3C XG310 accelerator cards, each featuring four Intel SG1 chips. These instances are ideal for scenarios such as Android game streaming, Android Cloud App, and video transcoding.

> ⚠ **Notes:**
> This instance is currently available on an allowlist basis. Contact Pre-Sales Online Consultation to request access permissions for purchasing the instance.

Applicable Scenarios

- Android cloud phone
- Android game streaming
- Android cloud App

- Video Transcoding

Hardware Specifications

- **GPU:** Intel® SG1, utilizing H3C XG310 accelerator cards, with each card containing four SG1 chips.
- **CPU:** Intel® Xeon® Platinum 8255c CPU with a CPU clock speed of 2.5 GHz.
- **Storage:** You can select Cloud Block Storage Type. If you need to scale out, you can create an auto scaling cloud disk and mount it.
- **Network:** Default network optimization is enabled, with instance network performance matching its specifications. Public network can be configured as needed.

Specification	GPU	GPU Memory	vCPU	Memory (GiB)	Private Network Bandwidth (Gbps)	Packet Tx/Rx (PPS)	Number of Queues
GI1.10XLARGE160	H3C XG310 * 1 (Intel SG1 chips x 4)	32GB * 1 (8GB * 4)	42	160	13	2.5 million	32
GI1.21XLARGE320	H3C XG310 * 2 (Intel SG1 chips x 8)	32GB * 2 (8GB * 8)	84	320	25	6 million	32