GPU Cloud Computing

Instance Types

Product Documentation

Contents

Instance Types
Computing Instance
Rendering Instance

Instance Types
Computing Instance

Last updated：2022-11-03 17:20:44

GPU Computing instances provide powerful computing capabilities to help you process a large number of concurrent computing tasks in real time. They are suitable for general computing scenarios such as deep learning and scientific computing. They provide a fast, stable, and elastic computing service and can be **managed just like CVM instances**.

Use Cases

They are suitable for AI computing and HPC scenarios, for example:

- AI computing
- Deep learning inference
- Deep learning training
- Scientific computing/HPC
- Fluid dynamics
- Molecular modeling
- Meteorological engineering
- Seismic analysis
- Genomics

> Note
> If your GPU instance is to be used for 3D rendering tasks, we recommend you use a rendering instance configured with a vDWs/vWs license and installed with a GRID driver. It eliminates the need to manually configure the basic environment for GPU-based graphics and image processing.

Overview

GPU Computing instances are available in the following types:

Availability	Resource Type	GPU Type	Available Image	AZ
Featured	PNV4	NVIDIA A10	• CentOS 7.2 or later • Ubuntu 16.04 or later • Windows Server 2016 or later	Guangzhou, Shanghai, and Beijing
	GT4	NVIDIA A100 NVLink 40 GB		Guangzhou, Shanghai, Beijing, and Nanjing
	GN10Xp	NVIDIA Tesla V100 NVLink 32 GB	• CentOS 7.2 or later • Ubuntu 14.04 or later • Windows Server 2012 or later	Guangzhou, Shanghai, Beijing, Nanjing, Chengdu, Chongqing, Singapore, Mumbai, Silicon Valley, and Frankfurt
	GN7	NVIDIA Tesla T4		Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Hong Kong, Singapore, Bangkok, Jakarta, Mumbai, Seoul, Tokyo, Silicon Valley, Virginia, Frankfurt, Moscow, and São Paulo
		vGPU - NVIDIA Tesla T4	• CentOS 8.0 64-bit GRID 11.1 • Ubuntu 20.04 LTS 64-bit GRID 11.1	Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Hong Kong, Silicon Valley, and São Paulo
	GN7vi	NVIDIA Tesla T4	• CentOS 7.2–7.9 • Ubuntu 14.04 or later	Shanghai and Nanjing
Available	GI3X	NVIDIA Tesla T4	• CentOS 7.2 or later • Ubuntu 14.04 or later • Windows Server 2012 or later	Guangzhou, Shanghai, Beijing, Nanjing, Chengdu, and Chongqing
	GN10X	NVIDIA Tesla V100 NVLink 32 GB		Guangzhou, Shanghai, Beijing, Nanjing, Chengdu, Chongqing, Singapore, Silicon Valley, Frankfurt, and Mumbai
	GN8	NVIDIA Tesla P40		Guangzhou, Shanghai, Beijing, Chengdu, Chongqing, Hong Kong, and Silicon Valley

Availability	Resource Type	GPU Type	Available Image	AZ
	GN6 GN6S	NVIDIA Tesla P4		• GN6: Chengdu • GN6S: Guangzhou, Shanghai, and Beijing

Note

AZ: Accurate to the city level. For more information, see the instance configuration information below.

Suggestions on Computing Instance Model Selection

Tencent Cloud provides NVIDIA GPU instances to meet business needs in different scenarios. Refer to the following tables to select an NVIDIA GPU instance as needed.

The table below lists **recommended GPU Computing instance models**. A tick (✓) indicates that the model supports the corresponding feature. A pentagram (★) indicates that the model is recommended.

Feature/Instance	PNV4	GT4	GN10Xp	GN7	GN7vi	GI3X	GN10X	GN8	GN6 GN6S
Graphics and image processing	✓	-	✓	✓	✓	✓	✓	✓	✓
Video encoding and decoding	✓	-	✓	★	★	★	✓	✓	✓
Deep learning training	✓	★	★	✓	✓	✓	★	★	✓
Deep learning inference	★	✓	★	★	★	★	★	✓	✓
Scientific computing	-	★	★	-	-	-	★	-	-

Note

- These recommendations are for reference only. Select an appropriate instance model based on your needs.

- To use NVIDIA GPU instances for general computing tasks, you need to install the Tesla driver and CUDA toolkit. For more information, see Installing NVIDIA Driver and Installing CUDA Driver.
- To use NVIDIA GPU instances for 3D rendering tasks such as high-performance graphics processing and video encoding and decoding, you need to install a GRID driver and configure a license server.

Service Options

- Pay-as-you-go billing is supported.
- Instances can be launched in a VPC.
- Instances can be connected to other services such as CLB, without additional management and Ops costs. Private network traffic is free of charge.

Instance Specification

Computing PNV4

Computing PNV4 supports not only general GPU computing tasks such as deep learning, but also graphics and image processing tasks such as 3D rendering and video encoding and decoding.

Use cases

GN6 and GN6S are cost-effective and applicable to the following scenarios:

- Deep learning inference and small-scale training scenarios, such as:
  - AI inference for mass deployment
  - Small-scale deep learning training
- Graphic and image processing scenarios, such as:
  - Graphic and image processing
  - Video encoding and decoding
  - Graph database

AZs

PNV4 instances are available in Guangzhou Zone 7, Shanghai Zones 4 and 5, and Beijing Zone 6.

Hardware specification

- **CPU:** AMD EPYCTM Milan CPU 2.55 GHz, with a Max Boost frequency of 3.5 GHz.

- **GPU:** NVIDIA® A10, providing 62.5 TFLOPS of single-precision floating point performance, 250 TOPS for INT8, and 500 TOPS for INT4.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

PNV4 instances are available in the following configurations:

Model	GPU (NVIDIA A10)	GPU Video Memory (GDDR6)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
PNV4.7XLARGE116	1	1 * 24 GB	28 cores	116 GB	13 Gbps	2.3 million	28
PNV4.14XLARGE232	2	2 * 24 GB	56 cores	232 GB	25 Gbps	4.7 million	48
PNV4.28XLARGE466	4	4 * 24 GB	112 cores	466 GB	50 Gbps	9.5 million	48
PNV4.56XLARGE932	8	8 * 24 GB	224 cores	932 GB	100 Gbps	19 million	48

Computing GT4

Computing GT4 instances are suitable for general GPU computing tasks such as deep learning and scientific computing.

Use cases

GT4 features powerful double-precision floating point computing capabilities. It is suitable for large-scale deep learning training and inference as well as scientific computing scenarios, such as:

- Deep learning
- High-performance database
- Computational fluid dynamics
- Computational finance
- Seismic analysis
- Molecular modeling
- Genomics and others

AZs

GT4 instances are available in Guangzhou Zones 3, 4, and 6, Shanghai Zones 4 and 5, Beijing Zones 5 and 6, and Nanjing Zone 1.

Hardware specification

- **CPU:** AMD EPYC™ ROME CPU, with a clock rate of 2.6 GHz.
- **GPU:** NVIDIA® A100 NVLink 40 GB, providing 19.5 TFLOPS of single-precision floating point performance, 9.7 TFLOPS of double-precision floating point performance, and 600 GB/s NVLink.
- **Memory:** DDR4 with stable computing performance.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Private network bandwidth of up to 50 Gbps is supported, with strong packet sending/receiving capabilities. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GT4 instances are available in the following configurations:

Model	GPU (NVIDIA Tesla A100 NVLink 40 GB)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GT4.4XLARGE96	1	1 * 40 GB	16 cores	96 GB	5 Gbps	1.2 million	4
GT4.8XLARGE192	2	2 * 40 GB	32 cores	192 GB	10 Gbps	2.35 million	8
GT4.20XLARGE474	4	4 * 40 GB	82 cores	474 GB	25 Gbps	6 million	16
GT4.41XLARGE948	8	8 * 40 GB	164 cores	948 GB	50 Gbps	12 million	32

Note

GPU driver: Drivers of NVIDIA Tesla 450 or later are required for NVIDIA A100 GPUs, and version 460.32.03 (Linux)/461.33 (Windows) are recommended. For more information on driver versions, see NVIDIA Driver Documentation.

Computing GN10Xp

Computing GN10Xp instances support not only general GPU computing tasks such as deep learning and scientific computing, but also graphics and image processing tasks such as 3D rendering and video encoding and decoding.

Use cases

GN10Xp features powerful double-precision floating point computing capabilities. It is suitable for the following scenarios:

- Large-scale deep learning training and inference as well as scientific computing scenarios, such as:
  - Deep learning
  - High-performance database
  - Computational fluid dynamics
  - Computational finance
  - Seismic analysis
  - Molecular modeling
  - Genomics and others
- Graphic and image processing scenarios, such as:
  - Graphic and image processing
  - Video encoding and decoding
  - Graph database

AZs

GN10Xp instances are available in Guangzhou Zones 3 and 4, Shanghai Zones 2 and 3, Nanjing Zone 1, Beijing Zones 4, 5, and 7, Chengdu Zone 1, Chongqing Zone 1, Singapore Zone 1, Mumbai Zone 2, Silicon Valley Zone 2, and Frankfurt Zone 1.

Hardware specification

- **CPU:** Intel® Xeon® Platinum 8255C CPU, with a clock rate of 2.5 GHz.
- **GPU:** NVIDIA® Tesla® V100 NVLink 32GB, providing 15.7 TFLOPS of single-precision floating point performance, 7.8 TFLOPS of double-precision floating point performance, 125 TFLOPS of deep learning accelerator performance with Tensor cores, and 300 GB/s NVLink.
- **Memory:** DDR4, providing memory bandwidth of up to 2,666 MT/s.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GN10Xp instances are available in the following configurations:

Model	GPU (NVIDIA Tesla V100 NVLink 32 GB)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN10Xp.2XLARGE40	1	1 * 32 GB	10 cores	40 GB	3 Gbps	0.8 million	2
GN10Xp.5XLARGE80	2	2 * 32 GB	20 cores	80 GB	6 Gbps	1.5 million	5
GN10Xp.10XLARGE160	4	4 * 32 GB	40 cores	160 GB	12 Gbps	2.5 million	10
GN10Xp.20XLARGE320	8	8 * 32 GB	80 cores	320 GB	24 Gbps	4.9 million	16

Computing GN7

NVIDIA GPU instance GN7 supports not only general GPU computing tasks such as deep learning, but also graphic and image processing tasks such as 3D rendering and video encoding and decoding.

Use cases

GN6 and GN6S are cost-effective and applicable to the following scenarios:

- Deep learning inference and small-scale training scenarios, such as:
  - AI inference for mass deployment
  - Small-scale deep learning training
- Graphic and image processing scenarios, such as:
  - Graphic and image processing
  - Video encoding and decoding
  - Graph database

AZs

GN7 instances are available in the following AZs:

- **GN7.LARGE20 and GN7.2XLARGE40** instances are available in Guangzhou Zones 3, 4, 6, and 7, Shanghai Zones 2, 3, 4, and 5, Nanjing Zones 1, 2, and 3, Beijing Zones 3, 5, 6, and 7, Chengdu Zone 1, Chongqing Zone 1, Hong Kong Zone 2, Silicon Valley Zone 2, and São Paulo Zone 1.

- **Other GN7** instances are available in Guangzhou Zones 3, 4, 6, and 7, Shanghai Zones 2, 3, 4, and 5, Nanjing Zones 1, 2, and 3, Beijing Zones 3, 5, 6, and 7, Chengdu Zone 1, Chongqing Zone 1, Hong Kong Zone 2, Singapore Zones 1, 2, and 3, Bangkok Zone 2, Jakarta Zone 2, Mumbai Zone 2, Seoul Zones 1 and 2, Tokyo Zone 2, Silicon Valley Zone 2, Frankfurt Zone 1, Moscow Zone 1, Virginia Zone 2, and São Paulo Zone 1.

Hardware specification

- **CPU:** Intel® Xeon® Platinum 8255C CPU, with a clock rate of 2.5 GHz.
- **GPU:** NVIDIA® Tesla® T4, providing 8.1 TFLOPS of single-precision floating point performance, 130 TOPS for INT8, and 260 TOPS for INT4.
- **Memory:** DDR4, providing memory bandwidth of up to 2,666 MT/s.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GN7 instances are available in the following configurations:

Model	GPU (NVIDIA Tesla T4)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN7.LARGE20	1/4	4 GB vGPU	4 cores	20 GB	1.5 Gbps	0.5 million	8
GN7.2XLARGE40	1/2	8 GB vGPU	10 cores	40 GB	3 Gbps	0.7 million	8
GN7.2XLARGE32	1	1 * 16 GB	8 cores	32 GB	3 Gbps	0.6 million	8
GN7.5XLARGE80	1	1 * 16 GB	20 cores	80 GB	7 Gbps	1.4 million	10
GN7.8XLARGE128	1	1 * 16 GB	32 cores	128 GB	10 Gbps	2.4 million	16
GN7.10XLARGE160	2	2 * 16 GB	40 cores	160 GB	13 Gbps	2.8 million	20
GN7.20XLARGE320	4	4 * 16 GB	80 cores	320 GB	25 Gbps	5.6 million	32

Note

vGPU:

- GN7 instance cluster provides vGPU-based instances. The vGPU type is vComputeServer, which only supports CUDA APIs but not DirectX or OpenGL APIs. In graphics and image processing scenarios such as 3D rendering and video encoding and decoding, we recommend you use rendering GN7vw instances configured with a vDWS license server and installed with a GRID driver.
- vCS instances require a GRID driver and don't support Windows.

Video enhancement GN7vi

NVIDIA GN7vi instances are GN7 instances configured with Tencent's proprietary MPS technology and integrated with AI. They include the TSC encoding and decoding engine and image quality enhancement toolkit and are suitable for VOD and live streaming scenarios. This type of instance allows you to leverage Tencent Cloud's proprietary TSC encoding and decoding as well as AI image quality enhancement features.

AZs

GN7vi instances are available in Shanghai Zones 2, 3, 4, and 5 and Nanjing Zones 1 and 2.

Hardware specification

- **CPU:** Intel® Xeon® Platinum 8255C CPU, with a clock rate of 2.5 GHz.
- **GPU:** NVIDIA® Tesla® T4, providing 8.1 TFLOPS of single-precision floating point performance, 130 TOPS for INT8, and 260 TOPS for INT4.
- **Memory:** DDR4, providing memory bandwidth of up to 2,666 MT/s.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GN7vi instances are available in the following configurations:

Model	GPU (NVIDIA Tesla T4)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN7vi.5XLARGE80	1	1 * 16 GB	20 cores	80 GB	6 Gbps	1.4 million	20

Model	GPU (NVIDIA Tesla T4)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN7vi.10XLARGE160	2	2 * 16 GB	40 cores	160 GB	13 Gbps	2.8 million	32
GN7vi.20XLARGE320	4	4 * 16 GB	80 cores	320 GB	25 Gbps	5.6 million	32

Interference GI3X

NVIDIA GI3X supports not only general GPU computing tasks such as deep learning, but also graphics and image processing tasks such as 3D rendering and video encoding and decoding.

Use cases

GN6 and GN6S are cost-effective and applicable to the following scenarios:

- Deep learning inference and small-scale training scenarios, such as:
  - AI inference for mass deployment
  - Small-scale deep learning training
- Graphic and image processing scenarios, such as:
  - Graphic and image processing
  - Video encoding and decoding
  - Graph database

AZs

GI3X instances are available in Guangzhou Zone 3, Shanghai Zones 4 and 5, Nanjing Zones 1 and 2, Beijing Zones 5 and 6, Chengdu Zone 1, and Chongqing Zone 1.

Hardware specification

- **CPU:** AMD EPYC™ ROME CPU 2.6 GHz, with a Max Boost frequency of 3.3 GHz.
- **GPU:** NVIDIA® Tesla® T4, providing 8.1 TFLOPS of single-precision floating point performance, 130 TOPS for INT8, and 260 TOPS for INT4.
- **Memory**: Latest eight-channel DDR4 with stable computing performance.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GI3X instances are available in the following configurations:

Model	GPU (NVIDIA Tesla T4)	GPU Video Memory (GDDR6)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GI3X.8XLARGE64	1	1 * 16 GB	32 cores	64 GB	5 Gbps	1.4 million	8
GI3X.22XLARGE226	2	2 * 16 GB	90 cores	226 GB	13 Gbps	3.75 million	16
GI3X.45XLARGE452	4	4 * 16 GB	180 cores	452 GB	25 Gbps	7.5 million	32

Computing GN10X

Computing GN10X supports not only general GPU computing tasks such as deep learning and scientific computing, but also graphics and image processing tasks such as 3D rendering and video encoding and decoding.

Use cases

GN10X features powerful double-precision floating point computing capabilities. It is suitable for the following scenarios:

- Large-scale deep learning training and inference as well as scientific computing scenarios, such as:
  - Deep learning
  - High-performance database
  - Computational fluid dynamics
  - Computational finance
  - Seismic analysis
  - Molecular modeling
  - Genomics and others
- Graphic and image processing scenarios, such as:
  - Graphic and image processing
  - Video encoding and decoding
  - Graph database

AZs

GN10X instances are available in Guangzhou Zones 3 and 4, Shanghai Zones 2 and 3, Nanjing Zone 1, Beijing Zones 4, 5, and 7, Chengdu Zone 1, Chongqing Zone 1, Singapore Zone 1, Silicon Valley Zone 2, Frankfurt Zone 1, and Mumbai Zone 2.

Hardware specification

- **CPU:** GN10X is configured with an Intel® Xeon® Gold 6133 CPU, with a clock rate of 2.5 GHz.
- **GPU: **NVIDIA® Tesla® V100 NVLink 32GB, providing 15.7 TFLOPS of single-precision floating point performance, 7.8 TFLOPS of double-precision floating point performance, 125 TFLOPS of deep learning accelerator performance with Tensor cores, and 300 GB/s NVLink.
- **Memory:** DDR4, providing memory bandwidth of up to 2,666 MT/s.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GN10X instances are available in the following configurations:

Model	GPU (NVIDIA Tesla V100 NVLink 32 GB)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN10X.2XLARGE40	1	1 * 32 GB	8 cores	40 GB	3 Gbps	0.8 million	2
GN10X.9XLARGE160	4	4 * 32 GB	36 cores	160 GB	13 Gbps	2.5 million	9
GN10X.18XLARGE320	8	8 * 32 GB	72 cores	320 GB	25 Gbps	4.9 million	16

Computing GN8

NVIDIA GPU instance GN8 supports not only general GPU computing tasks such as deep learning, but also graphic and image processing tasks such as 3D rendering and video encoding and decoding.

Use cases

GN8 is applicable to the following scenarios:

- Deep learning training and inference scenarios, such as:
  - AI inference with high throughput
  - Deep learning
- Graphic and image processing scenarios, such as:
  - Graphic and image processing

- Video encoding and decoding
- Graph database

AZs

GN8 instances are available in Guangzhou Zone 3, Beijing Zones 2 and 4, Chengdu Zone 1, Hong Kong Zone 2, Shanghai Zone 3, Chongqing Zone 1, and Silicon Valley Zone 1.

Hardware specification

- **CPU:** Intel® Xeon® E5-2680 v4 CPU, with a clock rate of 2.4 GHz.
- **GPU:** NVIDIA® Tesla® P40, providing 12 TFLOPS of single-precision floating point performance and 47 TOPS for INT8.
- **Memory:** DDR4, providing memory bandwidth of up to 2,666 MT/s.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GN8 instances are available in the following configurations:

Model	GPU (NVIDIA Tesla P40)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN8.LARGE56	1	24 GB	6 cores	56 GB	1.5 Gbps	0.45 million	8
GN8.3XLARGE112	2	48 GB	14 cores	112 GB	2.5 Gbps	0.5 million	8
GN8.7XLARGE224	4	96 GB	28 cores	224 GB	5 Gbps	0.7 million	14
GN8.14XLARGE448	8	192 GB	56 cores	448 GB	10 Gbps	0.7 million	28

Computing GN6 and GN6S

NVIDIA GPU instances GN6 and GN6S support not only general GPU computing tasks such as deep learning, but also graphic and image processing tasks such as 3D rendering and video encoding and decoding.

Use cases

GN6 and GN6S are cost-effective and applicable to the following scenarios:

- Deep learning inference and small-scale training scenarios, such as:
  - AI inference for mass deployment
  - Small-scale deep learning training
- Graphic and image processing scenarios, such as:
  - Graphic and image processing
  - Video encoding and decoding
  - Graph database

AZs

GN6 and GN6S instances are available in the following AZs:

- **GN6**: Chengdu Zone 1.
- **GN6S**: Guangzhou Zone 3, Shanghai Zones 2, 3, and 4, and Beijing Zones 4 and 5.

Hardware specification

- **CPU:** GN6 is configured with an Intel® Xeon® E5-2680 v4 CPU, with a clock rate of 2.4 GHz. GN6S is configured with an Intel® Xeon® Silver 4110 CPU, with a clock rate of 2.1 GHz.
- **GPU:** NVIDIA® Tesla® P4, providing 5.5 TFLOPS of single-precision floating point performance and 22 TOPS for INT8.
- **Memory:** DDR4, providing memory bandwidth of up to 2,666 MT/s.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GN6 and GN6S instances are available in the following configurations:

Model	GPU (NVIDIA Tesla P4)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN6.7XLARGE48	1	8 GB	28 cores	48 GB	5 Gbps	1.2 million	14
GN6.14XLARGE96	2	16 GB	56 cores	96 GB	10 Gbps	1.2 million	28

Model	GPU (NVIDIA Tesla P4)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GN6S.LARGE20	1	8 GB	4 cores	20 GB	5 Gbps	0.5 million	8
GN6S.2XLARGE40	2	16 GB	8 cores	40 GB	9 Gbps	0.8 million	8

Rendering Instance

Last updated：2022-05-09 16:33:49

GPU Rendering instances support GPU-based traditional graphics and image processing such as 3D rendering. They provide a fast, stable, and elastic computing service and can be **managed just like CVM instances**.

Use Cases

High-performance graphics processing and 3D rendering, such as:

- Nonlinear editing
- Cloud game
- Cloud phone
- Cloud desktop
- CloudXR
- Graphics and image processing

Overview

GPU Rendering instances are available in the following types:

Instance	GPU Type	Available Image	AZ
GNV4v	NVIDIA A10	Windows Server 2019 Datacenter 64-bit Chinese GRID 13	Beijing, Shanghai, and Guangzhou
GNV4	NVIDIA A10	• CentOS 7.2 or later • Ubuntu 16.04 or later • Windows Server 2019 Datacenter 64-bit Chinese GRID 13	Beijing, Shanghai, Guangzhou, and Chongqing
GN7vw	NVIDIA Tesla T4	• CentOS 8.0 64-bit GRID 11.1 • Windows Server 2019 Datacenter 64-bit Chinese GRID 11.1	Beijing, Shanghai, Guangzhou, Nanjing, Chengdu, Chongqing, Hong Kong (China), Singapore, Mumbai, Silicon Valley, Virginia, and Frankfurt
GI1	Intel SG1	• CentOS 7.6 64-bit + SG1-pv1.3 • CentOS 7.6 64-bit + SG1-pv1.4	Beijing, Shanghai, Guangzhou, Nanjing, and Chongqing

Note

AZ: Accurate to the city level. For more information, see the instance configuration information below.

Suggestions on Rendering Instance Model Selection

Tencent Cloud provides diverse GPU Computing instances to meet business needs in different scenarios. Refer to the following tables to select a Computing instance as needed.

The table below lists **recommended GPU Rendering instance models**. A tick (✓) indicates that the model supports the corresponding feature. A pentagram (★) indicates that the model is recommended.

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

Note

>- These recommendations are for reference only. Select an appropriate instance model based on your needs.

- To use NVIDIA GPU instances for 3D rendering tasks such as high-performance graphics processing and video encoding and decoding, you need to install a GRID driver and configure a license server. For GNV4v, GNV4, and GN7vw instances, you can select a specified image with a GRID driver preinstalled and a license server preconfigured.
- GNV4v, GNV4, and GN7vw instance clusters provide vGPU instance types that support vDWs and vWs. They also support graphics APIs such as DirectX and OpenGL.

Service Options

- Spot instances and pay-as-you-go instances are supported.
- Instances can be launched in VPC.
- Instances can be connected to other services such as CLB, without additional management and Ops costs. Private network traffic is free of charge.

Instance Specification

Rendering GNV4v

A **NVIDIA GNV4v instance** is a rendering instance configured with a vDWS license server and installed with a GRID driver. It is suitable for graphics and image processing scenarios such as 3D rendering and video encoding and decoding. It eliminates the need to manually configure the basic environment for GPU-based graphics and image processing.

> Note
> This instance model is currently made available through an allowlist. To purchase it, submit a ticket for application.

AZs

GNV4v instances are available in Guangzhou Zone 7, Shanghai Zone 5, and Beijing Zone 6.

Hardware specification

- **CPU:** AMD EPYCTM Milan CPU 2.55 GHz, with a Max Boost frequency of 3.5 GHz.
- **GPU:** NVIDIA® A10, providing 62.5 TFLOPS of single-precision floating point performance, 250 TOPS for INT8, and 500 TOPS for INT4.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GNV4v instances are available in the following configurations:

Model	GPU (NVIDIA A10)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues

Model	GPU (NVIDIA A10)	GPU Video Memory (HBM2)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GNV4v.XLARGE24	1/4	6 GB vGPU	6 cores	24 GB	3 Gbps	0.5 million	6
GNV4v.3XLARGE58	1/2	12 GB vGPU	14 cores	58 GB	7 Gbps	1.1 million	14
GNV4v.7XLARGE116	1	1 * 24 GB	28 cores	116 GB	13 Gbps	2.3 million	28

Rendering GNV4

A **NVIDIA GNV4 instance** is a rendering instance configured with a vDWS license server and installed with a GRID driver. It is suitable for graphics and image processing scenarios such as 3D rendering and video encoding and decoding. It eliminates the need to manually configure the basic environment for GPU-based graphics and image processing.

> Note
> This instance model is currently made available through an allowlist. To purchase it, submit a ticket for application.

AZs

GNV4 instances are available in Beijing Zone 6, Shanghai Zone 5, Guangzhou Zone 6, and Chongqing Zone 1.

Hardware specification

- **CPU:** Intel® Xeon® Cooper Lake CPU, with a base clock of 3.4 GHz and a Max Turbo frequency of 3.8 GHz.
- **GPU:** NVIDIA® A10, providing 31.2 TFLOPS of single-precision floating point performance, 250 TOPS for INT8, and 500 TOPS for INT4.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GNV4 instances are available in the following configurations:

Model	GPU (NVIDIA A10)	GPU Video Memory (GDDR6)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of queues
GNV4.3XLARGE44	1	24 GB	12 cores	44 GB	2 Gbps	0.53 million	4

Rendering GN7vw

A **NVIDIA GN7vw instance** is a rendering instance configured with a vDWS license server and installed with a GRID driver on the basis of GN7. It is suitable for graphics and image processing scenarios such as 3D rendering and video encoding and decoding. It eliminates the need to manually configure the basic environment for GPU-based graphics and image processing.

> Note
> GPU Rendering GN7vw is offered with limited availability.

AZs

GN7vw instances are available in Guangzhou Zones 3 and 4, Shanghai Zones 2, 4, and 5, Nanjing Zones 1 and 2, Beijing Zone 5, Chengdu Zone 1, Chongqing Zone 1, Hong Kong Zone 2, Singapore Zone 1, Mumbai Zone 2 , Silicon Valley Zone 2, Virginia Zone 2, and Frankfurt Zone 1.

Hardware specification

- **CPU:** Intel® Xeon® Platinum 8255C CPU, with a clock rate of 2.5 GHz.
- **GPU:** NVIDIA® Tesla® T4, providing 8.1 TFLOPS of single-precision floating point performance, 130 TOPS for INT8, and 260 TOPS for INT4.
- **Memory:** DDR4, providing memory bandwidth up to 2,666 MT/s.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.
- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GN7vw instances are available in the following configurations:

Model	GPU (NVIDIA Tesla T4)	GPU Video Memory (GDDR6)	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of queues
GN7vw.LARGE16	1/4	4 GB vGPU	4 cores	16 GB	2 Gbps	0.5 million	8
GN7vw.2XLARGE32	1/2	8 GB vGPU	8 cores	32 GB	4 Gbps	0.8 million	8
GN7vw.4XLARGE64	1	1 * 16 GB	16 cores	64 GB	7 Gbps	1.5 million	8

Rendering GI1

GPU Rendering GI1 instances are equipped with H3C XG310 graphics cards, with each containing four Intel SG1 chips. They are suitable for Android cloud games and apps and video transcoding.

> Note
> This instance model is currently made available through an allowlist. To purchase it, submit a ticket for application.

Use cases

- Android cloud phone
- Android cloud game
- Android cloud app
- Video transcoding

AZs

GI1 instances are available in Beijing Zone 6, Shanghai Zone 5, Guangzhou Zone 7, Nanjing Zone 3, and Chongqing Zone 1.

Hardware specification

- **CPU:** Intel® Xeon® Platinum 8255c CPU, with a clock rate of 2.5 GHz.
- **GPU:** Intel® SG1, adopting H3C XG310 graphics cards with each containing four SG1 chips.
- **Storage:** Select the appropriate CBS cloud disk type. To expand the cloud disk capacity, create and mount an elastic cloud disk.

- **Network:** Network optimization is enabled by default. The network performance of an instance depends on its specification. You can purchase public network bandwidth as needed.

GI1 instances are available in the following configurations:

Model	GPU (Intel SG1)	GPU Video Memory	vCPU	Memory (DDR4)	Private Network Bandwidth	Packets In/Out (PPS)	Number of Queues
GI1.10XLARGE160	1 * H3C XG310 (four Intel SG1 chips)	32 GB (4 * 8 GB)	42 cores	160 GB	13 Gbps	2.5 million	32
GI1.21XLARGE320	2 * H3C XG310 (eight Intel SG1 chips)	64 GB (8 * 8 GB)	84 cores	320 GB	25 Gbps	6 million	32