GPU Cloud Computing

Instance Types

Product Documentation
Copyright Notice

©2013-2022 Tencent Cloud. All rights reserved.

Copyright in this document is exclusively owned by Tencent Cloud. You must not reproduce, modify, copy or distribute in any way, in whole or in part, the contents of this document without Tencent Cloud's prior written consent.

Trademark Notice

All trademarks associated with Tencent Cloud and its services are owned by Tencent Cloud Computing (Beijing) Company Limited and its affiliated companies. Trademarks of third parties referred to in this document are owned by their respective proprietors.

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.
Contents

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

<table>
<thead>
<tr>
<th>Feature/Instance</th>
<th>PNV4</th>
<th>GT4</th>
<th>GN10Xp</th>
<th>GN7</th>
<th>GN7vi</th>
<th>GI3X</th>
<th>GN10X</th>
<th>GN8</th>
<th>GN6</th>
<th>GN6S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphics and image processing</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Video encoding and decoding</td>
<td>✓</td>
<td>-</td>
<td>✓</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Deep learning training</td>
<td>✓</td>
<td>★</td>
<td>★</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
</tr>
<tr>
<td>Deep learning inference</td>
<td>★</td>
<td>✓</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Scientific computing</td>
<td>-</td>
<td>★</td>
<td>★</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>★</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**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:**

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

**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:

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

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:

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

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:

<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA Tesla T4)</th>
<th>GPU Video Memory (HBM2)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Private Network Bandwidth</th>
<th>Packets In/Out (PPS)</th>
<th>Number of Queues</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN7.1XLARGE20</td>
<td>1/4</td>
<td>4 GB vGPU</td>
<td>4 cores</td>
<td>20 GB</td>
<td>1.5 Gbps</td>
<td>0.5 million</td>
<td>8</td>
</tr>
<tr>
<td>GN7.2XLARGE40</td>
<td>1/2</td>
<td>8 GB vGPU</td>
<td>10 cores</td>
<td>40 GB</td>
<td>3 Gbps</td>
<td>0.7 million</td>
<td>8</td>
</tr>
<tr>
<td>GN7.2XLARGE32</td>
<td>1</td>
<td>1 * 16 GB</td>
<td>8 cores</td>
<td>32 GB</td>
<td>3 Gbps</td>
<td>0.6 million</td>
<td>8</td>
</tr>
<tr>
<td>GN7.5XLARGE80</td>
<td>1</td>
<td>1 * 16 GB</td>
<td>20 cores</td>
<td>80 GB</td>
<td>7 Gbps</td>
<td>1.4 million</td>
<td>10</td>
</tr>
<tr>
<td>GN7.8XLARGE128</td>
<td>1</td>
<td>1 * 16 GB</td>
<td>32 cores</td>
<td>128 GB</td>
<td>10 Gbps</td>
<td>2.4 million</td>
<td>16</td>
</tr>
<tr>
<td>GN7.10XLARGE160</td>
<td>2</td>
<td>2 * 16 GB</td>
<td>40 cores</td>
<td>160 GB</td>
<td>13 Gbps</td>
<td>2.8 million</td>
<td>20</td>
</tr>
<tr>
<td>GN7.20XLARGE320</td>
<td>4</td>
<td>4 * 16 GB</td>
<td>80 cores</td>
<td>320 GB</td>
<td>25 Gbps</td>
<td>5.6 million</td>
<td>32</td>
</tr>
</tbody>
</table>
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:

<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA Tesla T4)</th>
<th>GPU Video Memory (HBM2)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Private Network Bandwidth</th>
<th>Packets In/Out (PPS)</th>
<th>Number of Queues</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN7vi.5XLARGE80</td>
<td>1</td>
<td>1 * 16 GB</td>
<td>20 cores</td>
<td>80 GB</td>
<td>6 Gbps</td>
<td>1.4 million</td>
<td>20</td>
</tr>
</tbody>
</table>
### 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:

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

**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:

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

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:

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

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:

<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA Tesla P4)</th>
<th>GPU Video Memory (HBM2)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Private Network Bandwidth</th>
<th>Packets In/Out (PPS)</th>
<th>Number of Queues</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN6.7XLARGE48</td>
<td>1</td>
<td>8 GB</td>
<td>28</td>
<td>48 GB</td>
<td>5 Gbps</td>
<td>1.2 million</td>
<td>14</td>
</tr>
<tr>
<td>GN6.14XLARGE96</td>
<td>2</td>
<td>16 GB</td>
<td>56</td>
<td>96 GB</td>
<td>10 Gbps</td>
<td>1.2 million</td>
<td>28</td>
</tr>
<tr>
<td>Model</td>
<td>GPU (NVIDIA Tesla P4)</td>
<td>GPU Video Memory (HBM2)</td>
<td>vCPU</td>
<td>Memory (DDR4)</td>
<td>Private Network Bandwidth</td>
<td>Packets In/Out (PPS)</td>
<td>Number of Queues</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------</td>
<td>-------------------------</td>
<td>-------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>GN6S.LARGE20</td>
<td>1</td>
<td>8 GB</td>
<td>4</td>
<td>20 GB</td>
<td>5 Gbps</td>
<td>0.5 million</td>
<td>8</td>
</tr>
<tr>
<td>GN6S.2XLARGE40</td>
<td>2</td>
<td>16 GB</td>
<td>8</td>
<td>40 GB</td>
<td>9 Gbps</td>
<td>0.8 million</td>
<td>8</td>
</tr>
</tbody>
</table>
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:

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

<table>
<thead>
<tr>
<th>Feature/Instance</th>
<th>GNV4v</th>
<th>GNV4</th>
<th>GN7vw</th>
<th>GI1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphics and image processing</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
</tr>
<tr>
<td>Video encoding and decoding</td>
<td>★</td>
<td>★</td>
<td>★</td>
<td>★</td>
</tr>
<tr>
<td>Deep learning training</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Deep learning inference</td>
<td>-</td>
<td>✓</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Scientific computing</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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

**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:**
<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA A10)</th>
<th>GPU Video Memory (GDDR6)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Private Network Bandwidth</th>
<th>Packets In/Out (PPS)</th>
<th>Number of queues</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNV4.3XLARGE44</td>
<td>1</td>
<td>24 GB</td>
<td>12 cores</td>
<td>44 GB</td>
<td>2 Gbps</td>
<td>0.53 million</td>
<td>4</td>
</tr>
</tbody>
</table>

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

**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.

- **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:**

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