GPU Cloud Computing

Instance Types

Product Documentation
Copyright Notice

©2013-2019 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 the 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
  NVIDIA GPU Instances
Instance Types
NVIDIA GPU Instances

Last updated: 2021-05-14 11:35:27

NVIDIA GPU GN* instances provide powerful computing capabilities to help you process a large number of concurrent computing tasks in real time. NVIDIA GPU instances are not only suitable for general computing scenarios such as deep learning and scientific computing, but also graphic and image processing such as 3D rendering and video encoding and decoding. Tencent Cloud GPU Cloud Computing (GCC) provides a fast, stable, and elastic computing service and allows you to manage NVIDIA GPU instances the same way you manage CVM instances.

Overview

NVIDIA GPU instances are categorized into rendering and computing types:

- Rendering: suitable for 3D rendering, video encoding and decoding, computer-aided design (CAD), etc.
- Computing: suitable for deep learning, scientific computing, computer-aided engineering (CAE), etc.

GCC provides the following types of NVIDIA GPU instances:

<table>
<thead>
<tr>
<th>Type</th>
<th>NVIDIA GPU instance</th>
<th>GPU type</th>
<th>GPU performance</th>
<th>Availability zone</th>
</tr>
</thead>
</table>
| Computing | GN10X or GN10Xp      | Tesla V100 NVLink 32GB | • 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 through Tensor cores  
• 300 GB/s NVLink | • GN10X: Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Singapore, and Silicon Valley  
• GN10Xp: Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, and Chongqing |
| Computing | GN8                  | Tesla P40 | • 12 TFLOPS of single-precision floating point performance  
• 47 TOPS for INT8 | Hong Kong (China), Guangzhou, Shanghai, Beijing, Chengdu, Chongqing, and Silicon Valley |
| Computing | GN7                  | Tesla T4  | • 8.1 TFLOPS of single-precision floating point performance  
• 130 TOPS for INT8  
• 260 TOPS for INT4 | Guangzhou, Shanghai, Nanjing, Beijing, Chengdu, Chongqing, Singapore, and Silicon Valley |
| Rendering | GN6 or GN6S          | Tesla P4  | • 5.5 TFLOPS of single-precision floating point performance  
• 22 TOPS for INT8 | • GN6: Chengdu  
• GN6S: Guangzhou, Shanghai, and Beijing |
| Rendering | GN7vw                | Tesla T4  | • 8.1 TFLOPS of single-precision floating point performance  
• 130 TOPS for INT8  
• 260 TOPS for INT4 | - |
Availability zone: accurate to the city level. For more information, please see the instance configuration information below.

Recommended NVIDIA GPU instances

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 NVIDIA GPU instances provided by GCC. A tick (✓) indicates that the NVIDIA GPU instance supports the corresponding feature. A pentagram (★) indicates that the NVIDIA GPU instance is recommended.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Instance</th>
<th>GN6 or GN6S</th>
<th>GN7</th>
<th>GN8</th>
<th>GN10X or GN10Xp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graphic and image processing</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>
</tr>
<tr>
<td>Deep learning training</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>
</tr>
<tr>
<td>Scientific computing</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>★</td>
<td></td>
</tr>
</tbody>
</table>

Video encoding and decoding

Recommended instance: GN7. GN7 uses T4 GPU. It features high performance and the lowest cost for transcoding a single video, making it suitable for video encoding and decoding.

Deep learning training

Recommended instance: GN8, GN10X, or GN10Xp. GN8 and GN10X use P40 or V100 mid- to high-end GPU. They feature powerful single-precision floating-point computing capabilities and large on-board memories, making them ideal for deep learning training.

Deep learning inference

Recommended instance: GN6, GN6S, GN7, or GN8. These instances use P4, T4, or P40 GPU. They feature INT8 computing capabilities and cost-efficiency, making them suitable for mass deployment.

Scientific computing

Recommended instance: GN10X or GN10Xp. GN10X and GN10Xp use V100 GPU. They feature powerful double-precision floating-point computing capabilities and provide optimal acceleration for computational science and engineering applications.

Note:
- These recommendations are for reference only. Please select an appropriate instance based on your needs.
- To use NVIDIA GPU instances for general computing tasks, you must install the Tesla driver and CUDA toolkit. For more information, please see Installing NVIDIA Driver and Installing CUDA Toolkit.
- To use NVIDIA GPU instances for 3D rendering tasks such as high-performance graphic processing and video encoding and decoding, you must install a GRID driver and configure a license server.
Service options

- NVIDIA GPU instances can be launched in Virtual Private Cloud (VPC).
- NVIDIA GPU instances support services such as Cloud Load Balancer (CLB) with no additional management and OPS costs. Private network traffic is free of charge.

GN10X and GN10Xp for computing tasks

**NVIDIA GPU instances GN10X and GN10Xp** support not only general GPU computing tasks such as deep learning and scientific computing, but also graphic and image processing tasks such as 3D rendering and video encoding and decoding.

Application scenarios

GN10X and GN10Xp feature powerful double-precision floating-point computing capabilities. They are applicable to 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

Hardware specifications

- **CPU:** GN10X is configured with an Intel® Xeon® Gold 6133 CPU with a clock rate of 2.5 GHz. GN10Xp is configured with an 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 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.

<table>
<thead>
<tr>
<th>GN10X and GN10Xp instances provide the following configurations:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
</tr>
<tr>
<td>-------</td>
</tr>
</tbody>
</table>

©2013-2019 Tencent Cloud. All rights reserved.
<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA Tesla V100 NVLink 32GB)</th>
<th>GPU memory (HBM2)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Private network bandwidth</th>
<th>Packet forwarding rate (PPS)</th>
<th>Number of queues</th>
<th>Availability zone</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>4 Gbps</td>
<td>800,000</td>
<td>2</td>
<td>Guangzhou Zone 3, Guangzhou Zone 4, Shanghai Zone 2, Shanghai Zone 3, Nanjing Zone 1, Beijing Zone 4, Beijing Zone 5, Chengdu Zone 1, Chongqing Zone 1, Singapore Zone 1, and Silicon Valley Zone 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>
<td></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>
<td>Chongqing Zone 1, Singapore Zone 1, and Silicon Valley Zone 2</td>
</tr>
<tr>
<td>GN10X.4XLARGE80</td>
<td>2</td>
<td>2 * 32 GB</td>
<td>18</td>
<td>80 GB</td>
<td>7 Gbps</td>
<td>1.5 million</td>
<td>4</td>
<td>Guangzhou Zone 3, Guangzhou Zone 4, Nanjing Zone 1, Chengdu Zone 1, and Chongqing Zone 1</td>
</tr>
<tr>
<td>GN10X.MEDIUM10</td>
<td>1/4</td>
<td>8 GB vGPU</td>
<td>2</td>
<td>10 GB</td>
<td>1 Gbps</td>
<td>200,000</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>GN10X.LARGE20</td>
<td>1/2</td>
<td>16 GB vGPU</td>
<td>4</td>
<td>20 GB</td>
<td>2 Gbps</td>
<td>400,000</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>GN10Xp.2XLARGE40</td>
<td>1</td>
<td>1 * 32 GB</td>
<td>10</td>
<td>40 GB</td>
<td>4 Gbps</td>
<td>800,000</td>
<td>10</td>
<td>Guangzhou Zone 3, Guangzhou Zone 4, Shanghai Zone 2, Shanghai Zone 3, Nanjing Zone 1, Beijing</td>
</tr>
<tr>
<td>GN10Xp.5XLARGE80</td>
<td>2</td>
<td>2 * 32 GB</td>
<td>20</td>
<td>80 GB</td>
<td>7 Gbps</td>
<td>1.5 million</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>GN10Xp.10XLARGE160</td>
<td>4</td>
<td>4 * 32 GB</td>
<td>40</td>
<td>160 GB</td>
<td>13 Gbps</td>
<td>2.5 million</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>GPU (NVIDIA Tesla V100 NVLink 32GB)</td>
<td>GPU memory (HBM2)</td>
<td>vCPU</td>
<td>Memory (DDR4)</td>
<td>Private network bandwidth</td>
<td>Packet forwarding rate (PPS)</td>
<td>Number of queues</td>
<td>Availability zone</td>
</tr>
<tr>
<td>--------------------</td>
<td>------------------------------------</td>
<td>-------------------</td>
<td>------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>----------------------------</td>
<td>-----------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td>GN10Xp.20XLARGE320</td>
<td>8</td>
<td>8 * 32 GB</td>
<td>80</td>
<td>320 GB</td>
<td>25 Gbps</td>
<td>4.9 million</td>
<td>16</td>
<td>Zone 5, Chengdu Zone 1, and Chongqing Zone 1</td>
</tr>
</tbody>
</table>

**vGPU**: GN10X instance cluster provides vGPU-based instances, which are currently in beta. The vGPU type is vComputeServer, which only supports CUDA APIs.

**GN8 for computing tasks**

**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.

**Application scenarios**

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

**Hardware specifications**

- **CPU**: Intel® Xeon® E5-2680 v4 CPU, 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 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 instance provides the following configurations:**
GN7 for computing tasks

**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.

**Application scenarios**
GN7 is 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

**Hardware specifications**

- **CPU**: Intel® Xeon® Platinum 8255C CPU, 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.

**GN7 instance provides the following configurations:**

<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA Tesla P40)</th>
<th>GPU memory (GDDR5)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Packet forwarding rate (PPS)</th>
<th>Number of queues</th>
<th>Availability zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN8.LARGE56</td>
<td>1</td>
<td>24 GB</td>
<td>6</td>
<td>56 GB</td>
<td>1.5 Gbps</td>
<td>450,000</td>
<td>6</td>
</tr>
<tr>
<td>GN8.3XLARGE112</td>
<td>2</td>
<td>48 GB</td>
<td>14</td>
<td>112 GB</td>
<td>2.5 Gbps</td>
<td>500,000</td>
<td>8</td>
</tr>
<tr>
<td>GN8.7XLARGE224</td>
<td>4</td>
<td>96 GB</td>
<td>28</td>
<td>224 GB</td>
<td>5 Gbps</td>
<td>700,000</td>
<td>8</td>
</tr>
<tr>
<td>GN8.14XLARGE448</td>
<td>8</td>
<td>192 GB</td>
<td>56</td>
<td>448 GB</td>
<td>10 Gbps</td>
<td>700,000</td>
<td>8</td>
</tr>
</tbody>
</table>

Hong Kong Zone 2, Guangzhou Zone 3, Shanghai Zone 3, Beijing Zone 2, Beijing Zone 4, Chengdu Zone 1, Chongqing Zone 1, and Silicon Valley Zone 1
<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA Tesla T4)</th>
<th>GPU memory (GDDR6)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Private network bandwidth</th>
<th>Packet forwarding rate (PPS)</th>
<th>Number of queues</th>
<th>Availability zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN7.LARGE20</td>
<td>1/4</td>
<td>4 GB vGPU</td>
<td>4</td>
<td>20 GB</td>
<td>1.5Gbps</td>
<td>500,000</td>
<td>4</td>
<td>Guangzhou Zone 3, Guangzhou Zone 4, Shanghai Zone 2, Shanghai Zone 4, Nanjing Zone 1, Nanjing Zone 2, Beijing Zone 3, Beijing Zone 5, Chengdu Zone 1, Chongqing Zone 1, and Silicon Valley Zone 2</td>
</tr>
<tr>
<td>GN7.2XLARGE40</td>
<td>1/2</td>
<td>8 GB vGPU</td>
<td>10</td>
<td>40 GB</td>
<td>3Gbps</td>
<td>700,000</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>GN7.2XLARGE32</td>
<td>1</td>
<td>1 * 16 GB</td>
<td>8</td>
<td>32 GB</td>
<td>3Gbps</td>
<td>600,000</td>
<td>8</td>
<td>Guangzhou Zone 3, Guangzhou Zone 4, Shanghai Zone 2, Shanghai Zone 4, Nanjing Zone 1, Nanjing Zone 2, Beijing Zone 3, Beijing Zone 5, Chengdu Zone 1, Chongqing Zone 1, and Silicon Valley Zone 2</td>
</tr>
<tr>
<td>GN7.5XLARGE80</td>
<td>1</td>
<td>1 * 16 GB</td>
<td>20</td>
<td>80 GB</td>
<td>7 Gbps</td>
<td>1.4 million</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>GN7.8XLARGE128</td>
<td>1</td>
<td>1 * 16 GB</td>
<td>32</td>
<td>128 GB</td>
<td>10Gbps</td>
<td>2.4 million</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>GN7.10XLARGE160</td>
<td>2</td>
<td>2 * 16 GB</td>
<td>40</td>
<td>160 GB</td>
<td>13 Gbps</td>
<td>2.8 million</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>GN7.20XLARGE320</td>
<td>4</td>
<td>4 * 16 GB</td>
<td>80</td>
<td>320 GB</td>
<td>25 Gbps</td>
<td>5.6 million</td>
<td>16</td>
<td></td>
</tr>
</tbody>
</table>

说明：
vGPU: GN7 instance cluster provides vGPU-based instances. The vGPU type is vComputeServer, which only supports CUDA APIs.

GN6 and GN6S for computing tasks

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.

Application scenarios

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

Hardware specifications

- **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 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 provide the following configurations:

<table>
<thead>
<tr>
<th>Model</th>
<th>GPU (NVIDIA Tesla P4)</th>
<th>GPU memory (GDDR5)</th>
<th>vCPU</th>
<th>Memory (DDR4)</th>
<th>Private network bandwidth</th>
<th>Packet forwarding rate (PPS)</th>
<th>Number of queues</th>
<th>Availability zone</th>
</tr>
</thead>
<tbody>
<tr>
<td>GN6.7XLARGE48</td>
<td>1</td>
<td>8 GB</td>
<td>28 cores</td>
<td>48 GB</td>
<td>5 Gbps</td>
<td>1.2 million</td>
<td>16</td>
<td>Chengdu Zone 1</td>
</tr>
<tr>
<td>GN6.14XLARGE96</td>
<td>2</td>
<td>16 GB</td>
<td>56 cores</td>
<td>96 GB</td>
<td>10 Gbps</td>
<td>1.2 million</td>
<td>16</td>
<td>Guangzhou Zone 3, Shanghai Zone 2, Shanghai Zone 3, Shanghai Zone 4, Beijing Zone 4,</td>
</tr>
<tr>
<td>GN6S.LARGE20</td>
<td>1</td>
<td>8 GB</td>
<td>4 cores</td>
<td>20 GB</td>
<td>7 Gbps</td>
<td>500,000</td>
<td>2</td>
<td>Shanghai Zone 4, Beijing Zone 4,</td>
</tr>
<tr>
<td>Model</td>
<td>GPU   (NVIDIA Tesla P4)</td>
<td>GPU memory (GDDR5)</td>
<td>vCPU</td>
<td>Memory (DDR4)</td>
<td>Private network bandwidth</td>
<td>Packet forwarding rate (PPS)</td>
<td>Number of queues</td>
<td>and Beijing Zone 5</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------------------------</td>
<td>--------------------</td>
<td>------</td>
<td>---------------</td>
<td>---------------------------</td>
<td>-----------------------------</td>
<td>------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>GN6S.2XLARGE40</td>
<td>2</td>
<td>16 GB</td>
<td>8 cores</td>
<td>40 GB</td>
<td>13 Gbps</td>
<td>800,000</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>