ALIBABA CLOUD

Alibaba Cloud

Elastic GPU Service

Product Introduction

Document Version: 20210122

Alibaba Cloud

Legal disclaimer

Alibaba Cloud reminds you to carefully read and fully understand the terms and conditions of this legal disclaimer before you read or use this document. If you have read or used this document, it shall be deemed as your total acceptance of this legal disclaimer.

1. You shall download and obtain this document from the Alibaba Cloud website or other Alibaba Cloud-authorized channels, and use this document for your own legal business activities only. The content of this document is considered confidential information of Alibaba Cloud. You shall strictly abide by the confidentiality obligations. No part of this document shall be disclosed or provided to any third party for use without the prior written consent of Alibaba Cloud.

2. No part of this document shall be excerpted, translated, reproduced, transmitted, or disseminated by any organization, company or individual in any form or by any means without the prior written consent of Alibaba Cloud.

3. The content of this document may be changed because of product version upgrade, adjustment, or other reasons. Alibaba Cloud reserves the right to modify the content of this document without notice and an updated version of this document will be released through Alibaba Cloud-authorized channels from time to time. You should pay attention to the version changes of this document as they occur and download and obtain the most up-to-date version of this document from Alibaba Cloud-authorized channels.

4. This document serves only as a reference guide for your use of Alibaba Cloud products and services. Alibaba Cloud provides this document based on the "status quo", "being defective", and "existing functions" of its products and services. Alibaba Cloud makes every effort to provide relevant operational guidance based on existing technologies. However, Alibaba Cloud hereby makes a clear statement that it in no way guarantees the accuracy, integrity, applicability, and reliability of the content of this document, either explicitly or implicitly. Alibaba Cloud shall not take legal responsibility for any errors or lost profits incurred by any organization, company, or individual arising from download, use, or trust in this document. Alibaba Cloud shall not, under any circumstances, take responsibility for any indirect, consequential, punitive, contingent, special, or punitive damages, including lost profits arising from the use or trust in this document (even if Alibaba Cloud has been notified of the possibility of such a loss).

5. By law, all the contents in Alibaba Cloud documents, including but not limited to pictures, architecture design, page layout, and text description, are intellectual property of Alibaba Cloud and/or its affiliates. This intellectual property includes, but is not limited to, trademark rights, patent rights, copyrights, and trade secrets. No part of this document shall be used, modified, reproduced, publicly transmitted, changed, disseminated, distributed, or published without the prior written consent of Alibaba Cloud and/or its affiliates. The names owned by Alibaba Cloud shall not be used, published, or reproduced for marketing, advertising, promotion, or other purposes without the prior written consent of Alibaba Cloud. The names owned by Alibaba Cloud include, but are not limited to, "Alibaba Cloud", "Aliyun", "HiChina", and other brands of Alibaba Cloud and/or its affiliates, which appear separately or in combination, as well as the auxiliary signs and patterns of the preceding brands, or anything similar to the company names, trade names, trademarks, product or service names, domain names, patterns, logos, marks, signs, or special descriptions that third parties identify as Alibaba Cloud and/or its affiliates.

6. Please directly contact Alibaba Cloud for any errors of this document.

Document conventions

Style	Description	Example
Danger	A danger notice indicates a situation that will cause major system changes, faults, physical injuries, and other adverse results.	Danger: Resetting will result in the loss of user configuration data.
Warning	A warning notice indicates a situation that may cause major system changes, faults, physical injuries, and other adverse results.	Warning: Restarting will cause business interruption. About 10 minutes are required to restart an instance.
Notice	A caution notice indicates warning information, supplementary instructions, and other content that the user must understand.	Notice: If the weight is set to 0, the server no longer receives new requests.
Note	A note indicates supplemental instructions, best practices, tips, and other content.	Note: You can use Ctrl + A to select all files.
>	Closing angle brackets are used to indicate a multi-level menu cascade.	Click **Settings**> **Network**> **Set network type**.
Bold	Bold formatting is used for buttons , menus, page names, and other UI elements.	Click **OK**.
`Courier font`	Courier font is used for commands	Run the `cd /d C:/window` command to enter the Windows system folder.
Italic	Italic formatting is used for parameters and variables.	`bae log list --instanceid` *Instance_ID*
[] or [a|b]	This format is used for an optional value, where only one item can be selected.	`ipconfig [-all|-t]`
{} or {a|b}	This format is used for a required value, where only one item can be selected.	`switch {active|stand}`

Table of Contents

1.What is Elastic GPU Service? ---- 05
2.Benefits ---- 06
3.Scenarios ---- 07
4.Regions and zones ---- 09
5.Instance families with GPU capabilities ---- 12
5.1. GPU-accelerated compute optimized instance families ---- 12
5.2. GPU-accelerated virtualization instance families ---- 24

1.What is Elastic GPU Service?

Elastic GPU Service provides GPU-accelerated computing capacity to implement the ready availability and auto scaling of GPU computing resources. As an elastic computing service provided by Alibaba Cloud, Elastic GPU Service combines the computing power of GPUs and CPUs to meet the challenges of scenarios such as AI, high-performance computing, and professional graphics and image processing.

Elastic GPU Service platform

As a computing chip, GPU provides real-time, high-speed parallel computing and floating-point computing capacity. Elastic GPU Service combines ECS with high-speed parallel heterogeneous accelerators of GPUs, delivering both ECS features and GPU acceleration capabilities.

Based on Elastic GPU Service, Alibaba Cloud launched instances with GPU capabilities, which can be operated in the same manner as common ECS instances while providing GPU acceleration capabilities. To use instances with GPU capabilities, select an enterprise-level heterogeneous computing instance type. For more information, see Instance families.

Features

- High elasticity

  Provides serial instance families. Instances with GPU capabilities can be created within minutes and support horizontal scaling as well as instance type changes within the same instance family.

- High performance and high security

  Supports point-to-point communication between GPUDirect and GPUs. GPUs can communicate with each other directly through NVLink with high bandwidth, low latency, and no CPU interventions. GPU provides elastic security isolation among tenants and authorizes and manages systems by using hypervisors. You can configure high speed communication between secure, isolated GPUs.

- Easy deployment

  Deeply integrated with the Alibaba Cloud ecosystem. You can build applications by combining Elastic GPU Service with other Alibaba Cloud products. For example, you can combine GPUs with Object Storage Service (OSS) and Apsara File Storage NAS to meet storage requirements and with E-MapReduce (EMR) to preprocess deep learning data. Elastic GPU Service supports cloud-native applications such as Alibaba Cloud Kubernetes, facilitating delivery.

- Easy monitoring

  Provides comprehensive GPU monitoring data, including GPUs, instances, and group dimensions, eliminating your O&M pressure. For more information, see GPU monitoring.

2.Benefits

Elastic GPU Service provides world-leading coverage, superior computing capacity, excellent network performance, and flexible purchase methods.

- World-leading coverage

  With large-scale deployment in 17 regions, Elastic GPU Service provides world-leading coverage. Based on delivery methods such as auto supply and auto provisioning, Elastic GPU Service is able to handle emergent needs of your business.

- Superior computing capacity

  Elastic GPU Service is equipped with industry-leading GPU processors. Combined with a high-performance CPU platform, a single instance with GPU capabilities can provide mixed precision computing performance of up to 1,000 TFLOPS.

- Excellent network performance

  VPCs of instances with GPU capabilities support up to 4,500 Kpps and 32 Gbit/s internal bandwidth. In addition, Super Computing Cluster (SCC) products provide an additional RDMA network of up to 50 Gbit/s between nodes to meet the low-latency and high-bandwidth requirements for data transmission between nodes.

- Flexible purchase methods

  Elastic GPU Service supports a variety of billing methods including subscription, pay-as-you-go, preemptible instances, and storage capacity units (SCUs). You can select the billing method that best suits your needs to avoid inefficient use of resources.

  > **Note** Reserved instances cannot be used for instances with GPU capabilities.

3.Scenarios

Elastic GPU Service is suitable for scenarios such as video transcoding, image rendering, AI training, AI inference, and cloud graphics workstations.

Transcoding for real-time videos

During Double 11 Global Shopping Festival gala in 2019, instances with GPU capabilities and FPGAs were used to support video transcoding at resolutions of 1080P, 2K, and 4K in real-time while consuming minimal bandwidth. Instances with FPGAs transcoded videos in 720P in real time based on the H.265 standard with a 21.6% reduction in bandwidth consumption. Instances with GPU capabilities supported high-concurrency real-time video streaming of more than 5,000 channels, gradually rose to the peak of 6,200 channels per minute, and smoothly handled the traffic peak. Instances with GPU capabilities also took part in services such as generating real-time rendering images of households. For the first time, a large number of ebmgn6v bare metal instances with powerful computing capacity are provided to support Taobao renderers to improve performance by dozens of times. Real-time rendering in seconds was achieved, and more than 5,000 household images were rendered. The FPGA image transcoding service used a super-large cluster of over 3,000 nodes to provide processing capabilities of up to millions of QPS for the Taobao Image Space, and handled 85% of the traffic of the Taobao images on Double 11.

AI training

gn6v and gn6e instances provide excellent general-purpose GPU acceleration capabilities and are suitable for providing acceleration engines for deep learning.

gn6v and gn6e instances are equipped with NVIDIA V100 GPU processors with 16 GB and 32 GB memory respectively and can provide mixed precision computing capacity of up to 1,000 TFLOPS per node. gn6v and gn6e instances can be seamlessly integrated into an elastic computing ecosystem to provide solutions that are ideal for either online or offline computation scenarios. Additionally, making full use of Container Service can help simplify deployment and O&M, and provide resource scheduling services.

AI inference

gn6i provides excellent AI inference capabilities.

gn6i instances are equipped with NVIDIA Tesla T4 GPU processors, providing single-precision floating-point computing capacity of up to 8.1 TFLOPS and int8 fixed-point processing capabilities of up to 130 TOPS. gn6i instances support mixed precision and meet requirements on computing power in deep learning (especially inference) scenarios. Additionally, a single processor only consumes 75 W of power while maintaining a high-performance output. gn6i instances can be seamlessly integrated into an elastic computing ecosystem to provide solutions that are ideal for either online or offline computation scenarios. Additionally, making full use of Container Service can help simplify deployment and O&M, and provide resource scheduling services. Alibaba Cloud Marketplace provides a gn6i instance image that is equipped with an NVIDIA GPU driver and a deep learning framework for simplified development.

Cloud games, cloud-based Internet cafes, and cloud graphics workstations

vgn6i and gn6i instances are equipped with NVIDIA Tesla T4 GPU accelerators based on the Turing architecture and provide excellent graphics computing capacity. vgn6i instances contain virtual GPUs generated from GPU slice virtualization, provide 1/2, 1/4, and 1/8 of T4 GPU computing capacity, and excellent 3D image rendering capabilities. vgn6i instances are suitable for scenarios such as cloud games and cloud-based Internet cafes. vgn6i and gn6i instances can be combined with Cloud Desktop products to provide cloud graphics workstation services and can be applied to scenarios such as film and television animation design, industrial design, medical imaging, and high-performance computing result presentation.

4.Regions and zones

This topic provides a complete list of regions and zones of Alibaba Cloud.

Regions are completely independent. Zones are completely isolated. However, zones in the same region can be connected with low-latency links.

Region

A region is a geographic area where a data center resides. The region of an Alibaba Cloud resource cannot be changed after the resource is created. The following table describes the information about all regions of Alibaba Cloud, including the region IDs and the cities where the regions reside.

- Regions in mainland China

Region	City	Region ID	Number of zones
China (Qingdao)	Qingdao	cn-qingdao	2
China (Beijing)	Beijing	cn-beijing	8
China (Zhangjiakou)	Zhangjiakou	cn-zhangjiakou	3
China (Hohhot)	Hohhot	cn-huhehaote	2
China (Ulanqab)	Ulanqab	cn-wulanchabu	2
China (Hangzhou)	Hangzhou	cn-hangzhou	8
China (Shanghai)	Shanghai	cn-shanghai	7
China (Shenzhen)	Shenzhen	cn-shenzhen	5
China (Heyuan)	Heyuan	cn-heyuan	2
China (Guangzhou)	Guangzhou	cn-guangzhou	2
China (Chengdu)	Chengdu	cn-chengdu	2

- Region outside mainland China

Region	City	Region ID	Number of zones
China (Hong Kong)	Hong Kong	cn-hongkong	2
Singapore (Singapore)	Singapore	ap-southeast-1	3
Australia (Sydney)	Sydney	ap-southeast-2	2
Malaysia (Kuala Lumpur)	Kuala Lumpur	ap-southeast-3	2
Indonesia (Jakarta)	Jakarta	ap-southeast-5	2

Region	City	Region ID	Number of zones
India (Mumbai)	Mumbai	ap-south-1	2
Japan (Tokyo)	Tokyo	ap-northeast-1	2
US (Silicon Valley)	Silicon Valley	us-west-1	2
US (Virginia)	Virginia	us-east-1	2
Germany (Frankfurt)	Frankfurt	eu-central-1	2
UK (London)	London	eu-west-1	2
UAE (Dubai)	Dubai	me-east-1	1

When you select a region, you must consider the following factors:

- Geographical location

  Select a region based on the geographical location and of you and your target users.

  - Mainland China

    In mainland China, we recommend that you select a region that is the closest to the geographical location of your target users to speed up the access. However, in terms of network infrastructure, Border Gateway Protocol (BGP) network quality, quality of service (QoS), and ease of use and configuration on Elastic Compute Service (ECS) instances, Alibaba Cloud regions in mainland China are almost the same. BGP networks ensure fast access to all regions in mainland China.

  - Outside mainland China
    - If your target users are located in China (Hong Kong) or Southeast Asia, you can select the following regions: China (Hong Kong), Singapore (Singapore), Malaysia (Kuala Lumpur), and Indonesia (Jakarta).
    - If your target users are located in Japan or Korea, you can select the Japan (Tokyo) region.
    - If your target users are located in India, you can select the India (Mumbai) region.
    - If your target users are located in Australia, you can select the Australia (Sydney) region
    - If your target users are located in America, you can select the US (Silicon Valley) and US (Virginia) regions.
    - If your target users are located in Continental Europe, you can select the Germany (Frankfurt) region.
    - If your target users are located in Middle East, you can select the UAE (Dubai) region.

- Connection between Alibaba Cloud products

  If you use multiple Alibaba Cloud products together, note the following items:

  - ECS instances, ApsaraDB for RDS instances, and Object Service Storage (OSS) buckets that are created in different regions cannot communicate with each other through internal networks.
  - Server Load Balancer (SLB) cannot balance requests from ECS instances deployed in different regions. ECS instances that you purchased in different regions cannot be deployed under the same SLB instance.

- Resource price

The price of resources may vary with regions. For more information, see Pricing.

- ICP license filing

  When you select a region, you must consider the special requirements of some areas. For example, if you purchase an ECS instance in a region in mainland China and use the instance as a web server, you must apply for an ICP license.

  If you need to apply for an ICP license, note the following items:

  - If you want to apply for an ICP license for services in Beijing, select the **China (Beijing)** region.
  - If you want to apply for an ICP license for services in Guangdong, select the **China (Shenzhen)** region.

  > **Note** The Communications Administration of each province in China has specific approval requirements for ICP licenses. For the latest requirements, see the content published on the ICP license application website of the local Communications Administration.

Zone

A zone is a physical area with independent power grids and networks in a region. The network latency for access between instances within the same zone is shorter.

Zones within the same region have access to each other, but faults within a single zone will not affect the others. We recommend that you choose a deployment method based on your business requirements for disaster recovery and network latency.

- If your application requires high disaster recovery capabilities, we recommend that you choose multi-zone deployment to create your instances in different zones of the same region.
- If your application requires low network latency, we recommend that you choose single-zone deployment to create your RDS instances in the same zone.

5.Instance families with GPU capabilities

5.1. GPU-accelerated compute optimized instance families

This topic describes the features of GPU-accelerated compute optimized instance families and lists the instance types of each family.

- Recommended instance families
  - gn6i, GPU-accelerated compute optimized instance family
  - gn6e, GPU-accelerated compute optimized instance family
  - gn6v, GPU-accelerated compute optimized instance family
  - ebmgn6e, GPU-accelerated compute optimized ECS Bare Metal Instance family
  - ebmgn6v, GPU-accelerated compute optimized ECS Bare Metal Instance family
  - ebmgn6i, GPU-accelerated compute optimized ECS Bare Metal Instance family
  - sccgn6, GPU-accelerated compute optimized SCC instance family
- Other available instance families
  - gn5, GPU-accelerated compute optimized instance family
  - gn5i, GPU-accelerated compute optimized instance family

gn6i, GPU-accelerated compute optimized instance family

Features

- Is an instance family in which all instances are I/O optimized.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Supports enhanced SSDs (ESSDs) that deliver millions of IOPS, standard SSDs, and ultra disks.
- Uses NVIDIA T4 GPU computing accelerators that feature:
  - New NVIDIA Turing architecture
  - 16 GB memory (320 GB/s bandwidth) per GPU
  - Up to 2,560 CUDA cores per GPU
  - Up to 320 Turing Tensor cores per GPU
  - Mixed-precision Tensor cores that support 65 FP16 TFLOPS, 130 INT8 TOPS, and 260 INT4 TOPS
- Provides high network performance based on large computing capacity.
- Suitable for the following scenarios:
  - AI (deep learning and machine learning) inference for computer vision, speech recognition, speech synthesis, natural language processing (NLP), machine translation, and recommendation systems
  - Real-time rendering for cloud games
  - Real-time rendering for AR and VR applications

- Graphics workstations or overloaded graphics computing
- GPU-accelerated databases
- High-performance computing

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.gn6i-c4g1.xlarge	4	15.0	None	T4*1	16	4.0	500	Yes	2	2	10
ecs.gn6i-c8g1.2xlarge	8	31.0	None	T4*1	16	5.0	800	Yes	2	2	10
ecs.gn6i-c16g1.4xlarge	16	62.0	None	T4*1	16	6.0	1,000	Yes	4	3	10
ecs.gn6i-c24g1.6xlarge	24	93.0	None	T4*1	16	7.5	1,200	Yes	6	4	10
ecs.gn6i-c24g1.12xlarge	48	186.0	None	T4*2	32	15.0	2,400	Yes	12	6	10
ecs.gn6i-c24g1.24xlarge	96	372.0	None	T4*4	64	30.0	4,800	Yes	24	8	10

> **Note**
>
> - You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
> - For more information about these specifications, see Description of instance specifications.

gn6e, GPU-accelerated compute optimized instance family

Features

- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Uses NVIDIA V100 (32 GB NVLink) GPU processors.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Uses NVIDIA V100 GPU computing accelerators (SXM2-based) that feature:
  - New NVIDIA Volta architecture
  - 32 GB HBM2 GPU memory (900 GB/s bandwidth) per GPU
  - Up to 5,120 CUDA cores per GPU
  - Up to 640 Tensor cores per GPU
  - Support for up to six NVLink connections and a total bandwidth of 300 GB/s (25 GB/s per connection)
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Deep learning applications such as training and inference applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition
  - Scientific computing applications such as fluid dynamics, finance, molecular dynamics, and environmental analysis

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.gn6e-c12g1.3xlarge	12	92.0	None	V100 *1	32	5.0	800	Yes	8	6	10

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.gn6e-c12g1.12xlarge	48	368.0	None	V100*4	128	16.0	2,400	Yes	8	8	20
ecs.gn6e-c12g1.24xlarge	96	736.0	None	V100*8	256	32.0	4,800	Yes	16	8	20

> **Note**
>
> - You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
> - For more information about these specifications, see Description of instance specifications.

gn6v, GPU-accelerated compute optimized instance family

Features

- Is an instance family in which all instances are I/O optimized.
- Supports ESSDs, standard SSDs, and ultra disks.
- Uses NVIDIA V100 GPU processors.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Uses NVIDIA V100 GPU computing accelerators (SXM2-based) that feature:
  - New NVIDIA Volta architecture
  - 16 GB HBM2 GPU memory (900 GB/s bandwidth) per GPU
  - Up to 5,120 CUDA cores per GPU
  - Up to 640 Tensor cores per GPU
  - Support for up to six NVLink connections and a total bandwidth of 300 GB/s (25 GB/s per connection)
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Deep learning applications such as training and inference applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition

- Scientific computing applications such as fluid dynamics, finance, molecular dynamics, and environmental analysis

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.gn6v-c8g1.2xlarge	8	32.0	None	1 * NVIDIA V100	1 * 16	2.5	800	Yes	4	4	10
ecs.gn6v-c8g1.8xlarge	32	128.0	None	4 * NVIDIA V100	4 * 16	10.0	2,000	Yes	8	8	20
ecs.gn6v-c8g1.16xlarge	64	256.0	None	8 * NVIDIA V100	8 * 16	20.0	2,500	Yes	16	8	20
ecs.gn6v-c10g1.20xlarge	82	336.0	None	8 * NVIDIA V100	8 * 16	32.0	4,500	Yes	16	8	20

Note

- You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
- For more information about these specifications, see Description of instance specifications.

ebmgn6e, GPU-accelerated compute optimized ECS Bare Metal Instance family

ebmgn6e is in invitational preview. To use ebmgn6e, submit a ticket.

Features

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Is an instance family in which all instances are I/O optimized.

- Supports ESSDs, standard SSDs, and ultra disks.
- Uses NVIDIA V100 (32 GB NVLink) GPU processors.
- Offers a CPU-to-memory ratio of 1:8.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Uses NVIDIA V100 GPU accelerators (SXM2-based) that feature:
  - New NVIDIA Volta architecture
  - 32 GB HBM2 memory (900 GB/s bandwidth) per GPU
  - 5,120 CUDA cores per GPU
  - 640 Tensor cores per GPU
  - Support for up to six NVLink connections for a total bandwidth of 300 GB/s per GPU (25 GB/s per connection)
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Deep learning applications such as training and inference applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition
  - Scientific computing applications such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analysis

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.ebmgn6e.24xlarge	96	768.0	None	V100*8	256	32.0	4,800	Yes	16	15	20

> **Note**
> - You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
> - For more information about these specifications, see Description of instance specifications.

ebmgn6v, GPU-accelerated compute optimized ECS Bare Metal Instance family

Features

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Is an instance family in which all instances are I/O optimized.

- Supports ESSDs, standard SSDs, and ultra disks.
- Uses NVIDIA V100 GPU processors.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Uses NVIDIA V100 GPU accelerators (SXM2-based) that feature:
  - New NVIDIA Volta architecture
  - 16 GB HBM2 memory (900 GB/s bandwidth) per GPU
  - 5,120 CUDA cores per GPU
  - 640 Tensor cores per GPU
  - Support for up to six NVLink connections for a total bandwidth of 300 GB/s per GPU (25 GB/s per connection)
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Deep learning applications such as training and inference applications of AI algorithms used in image classification, autonomous vehicles, and speech recognition
  - Scientific computing applications such as computational fluid dynamics, computational finance, molecular dynamics, and environmental analysis

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.ebmgn6v.24xlarge	96	384.0	None	V100*8	128	30.0	4,500	Yes	8	32	10

> **Note**
> - You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
> - For more information about these specifications, see Description of instance specifications.

ebmgn6i, GPU-accelerated compute optimized ECS Bare Metal Instance family

Features

- Provides flexible and powerful software-defined compute based on the SHENLONG architecture.
- Is an instance family in which all instances are I/O optimized.

- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Supports ESSDs that deliver millions of IOPS, standard SSDs, and ultra disks.
- Uses NVIDIA T4 GPU accelerators that feature:
  - New NVIDIA Turing architecture
  - 16 GB memory (320 GB/s bandwidth) per GPU.
  - 2,560 CUDA cores per GPU
  - Up to 320 Turing Tensor cores per GPU
  - Mixed-precision Tensor cores that support 65 FP16 TFLOPS, 130 INT8 TOPS, and 260 INT4 TOPS
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - AI (deep learning and machine learning) inference for computer vision, speech recognition, speech synthesis, natural language processing (NLP), machine translation, and recommendation systems
  - Real-time rendering for cloud gaming
  - Real-time rendering for AR and VR applications
  - Graphics workstations or overloaded graphics computing
  - GPU-accelerated databases
  - High-performance computing

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.ebmgn6i.24xlarge	96	384.0	None	T4*4	64	30.0	4,500	Yes	8	32	10

> **Note**
> - You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
> - For more information about these specifications, see Description of instance specifications.

sccgn6, GPU-accelerated compute optimized SCC instance family

Features

- Is an instance family in which all instances are I/O optimized.

- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors for consistent computing performance.
- Provides all features of ECS Bare Metal Instance.
- Storage:
  - Supports ESSDs, standard SSDs, and ultra disks.
  - Supports a high performance CPFS.
- Networking:
  - Supports VPCs.
  - Supports the RoCE v2 network, which is dedicated to low-latency RDMA communication.
- Uses NVIDIA V100 GPU accelerators (SXM2-based) that feature:
  - New NVIDIA Volta architecture
  - 16 GB HBM2 GPU memory
  - Up to 5,120 CUDA cores per GPUUp to 640 Tensor cores per GPU
  - A GPU memory bandwidth of up to 900 GB/s
  - Support for up to six NVLink connections and total bandwidth of 300 GB/s (25 GB/s per connection)
- Applies to the following scenarios:
  - Ultra-large-scale machine learning training on a distributed GPU cluster
  - Large-scale high performance scientific computing and simulations
  - Large-scale data analysis, batch processing, and video encoding

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	RoCE (Gbit/s)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.sccgn6.24xlarge	96	384.0	None	V100*8	30	4,500	25*2	Yes	8	32	10

> **Note**
> - You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
> - For more information about these specifications, see Description of instance specifications.

gn5, GPU-accelerated compute optimized instance family

Features

- Is an instance family in which all instances are I/O optimized.
- Supports standard SSDs and ultra disks only.
- Uses NVIDIA P100 GPU processors.
- Offers multiple CPU-to-memory ratios.
- Supports high-performance local NVMe SSDs.
- Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Deep learning
  - Scientific computing applications such as fluid dynamics, finance, genomics, and environmental analysis
  - Server-side GPU compute workloads such as high-performance computing, rendering, and multi-media encoding and decoding

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.gn5-c4g1.xlarge	4	30.0	440	1 * NVIDIA P100	1 * 16	3.0	300	No	1	3	10
ecs.gn5-c8g1.2xlarge	8	60.0	440	1 * NVIDIA P100	1 * 16	3.0	400	No	1	4	10
ecs.gn5-c4g1.2xlarge	8	60.0	880	2 * NVIDIA P100	2 * 16	5.0	1,000	No	2	4	10
ecs.gn5-c8g1.4xlarge	16	120.0	880	2 * NVIDIA P100	2 * 16	5.0	1,000	No	4	8	20

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.gn5-c28g1.7xlarge	28	112.0	440	1 * NVIDIA P100	1 * 16	5.0	1,000	No	8	8	20
ecs.gn5-c8g1.8xlarge	32	240.0	1760	4 * NVIDIA P100	4 * 16	10.0	2,000	No	8	8	20
ecs.gn5-c28g1.14xlarge	56	224.0	880	2 * NVIDIA P100	2 * 16	10.0	2,000	No	14	8	20
ecs.gn5-c8g1.14xlarge	54	480.0	3520	8 * NVIDIA P100	8 * 16	25.0	4,000	No	14	8	20

Note

- You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
- For more information about these specifications, see Description of instance specifications.

gn5i, GPU-accelerated compute optimized instance family

Features

- Is an instance family in which all instances are I/O optimized.
- Supports standard SSDs and ultra disks only.
- Uses NVIDIA P4 GPU processors.
- Offers a CPU-to-memory ratio of 1:4.
- Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Deep learning inference

- Server-side GPU compute workloads such as multi-media encoding and decoding

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.gn5i-c2g1.large	2	8.0	None	1 * NVIDIA P4	1 * 8	1.0	100	Yes	2	2	6
ecs.gn5i-c4g1.xlarge	4	16.0	None	1 * NVIDIA P4	1 * 8	1.5	200	Yes	2	3	10
ecs.gn5i-c8g1.2xlarge	8	32.0	None	1 * NVIDIA P4	1 * 8	2.0	400	Yes	4	4	10
ecs.gn5i-c16g1.4xlarge	16	64.0	None	1 * NVIDIA P4	1 * 8	3.0	800	Yes	4	8	20
ecs.gn5i-c16g1.8xlarge	32	128.0	None	2 * NVIDIA P4	2 * 8	6.0	1,200	Yes	8	8	20
ecs.gn5i-c28g1.14xlarge	56	224.0	None	2 * NVIDIA P4	2 * 8	10.0	2,000	Yes	14	8	20

Note

- You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
- For more information about these specifications, see Description of instance specifications.

5.2. GPU-accelerated virtualization instance families

This topic describes the features of GPU-accelerated virtualization instance families and lists the instance types of each family.

- Recommended instance families

  vgn6i, lightweight GPU-accelerated compute optimized instance family

- Other available instance families

  vgn5i, lightweight GPU-accelerated compute optimized instance family

To use a GPU-accelerated virtualization instance, you must install a GRID driver on the instance. Click here to go to the NVIDIA official website and purchase a GRID license. After you create an instance, you can manually install the GRID driver and activate the license.

vgn6i, lightweight GPU-accelerated compute optimized instance family

vgn6i is in invitational preview. To use vgn6i, submit a ticket.

Features

- Is an instance family in which all instances are I/O optimized.
- Supports standard SSDs and ultra disks only.
- Uses NVIDIA T4 GPU computing accelerators.
- Contains virtual GPUs generated from GPU slice virtualization.
  - Supports the 1/4 and 1/2 computing capacity of NVIDIA Tesla T4 GPUs.
  - Supports 4 GB and 8 GB of GPU video memory.
- Offers a CPU-to-memory ratio of 1:5.
- Uses 2.5 GHz Intel® Xeon® Platinum 8163 (Skylake) processors.
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Real-time rendering for cloud games
  - Real-time rendering for AR and VR applications
  - AI (deep learning and machine learning) inference for elastic Internet service deployment
  - Educational environment of deep learning
  - Modeling experiment environment of deep learning

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.vgn6i-m4.xlarge	4	23.0	None	T4*1/4	4	3.0	500	Yes	2	4	10
ecs.vgn6i-m8.2xlarge	10	46.0	None	T4*1/2	8	4.0	800	Yes	4	5	20

> **Note**
> - You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
> - For more information about these specifications, see Description of instance specifications.

vgn5i, lightweight GPU-accelerated compute optimized instance family

Features

- Is an instance family in which all instances are I/O optimized.
- Supports standard SSDs and ultra disks only.
- Uses NVIDIA P4 GPU computing accelerators.
- Contains virtual GPUs generated from GPU slice virtualization.
  - Supports the 1/8, 1/4, 1/2, and 1/1 computing capacity of NVIDIA Tesla P4 GPUs.
  - Supports 1 GB, 2 GB, 4 GB, and 8 GB of GPU video memory.
- Offers a CPU-to-memory ratio of 1:3.
- Uses 2.5 GHz Intel® Xeon® E5-2682 v4 (Broadwell) processors.
- Provides high network performance based on large computing capacity.
- Applies to the following scenarios:
  - Real-time rendering for cloud games
  - Real-time rendering for AR and VR applications
  - AI (deep learning and machine learning) inference for elastic Internet service deployment
  - Educational environment of deep learning
  - Modeling experiment environment of deep learning

Instance types

Instance type	vCPU	Memory (GiB)	Local storage (GiB)	GPU	GPU memory (GB)	Bandwidth (Gbit/s)	Packet forwarding rate (Kpps)	IPv6 support	NIC queues	ENIs (including one primary ENI)	Private IP addresses per ENI
ecs.vgn5i-m1.large	2	6.0	None	P4*1/8	1	1.0	300	Yes	2	2	6
ecs.vgn5i-m2.xlarge	4	12.0	None	P4*1/4	2	2.0	500	Yes	2	3	10
ecs.vgn5i-m4.2xlarge	8	24.0	None	P4*1/2	4	3.0	800	Yes	2	4	10
ecs.vgn5i-m8.4xlarge	16	48.0	None	P4*1	8	5.0	1,000	Yes	4	5	20

Note

- You can go to the ECS Instance Types Available for Each Region page to view the instance types available in each region.
- For more information about these specifications, see Description of instance specifications.