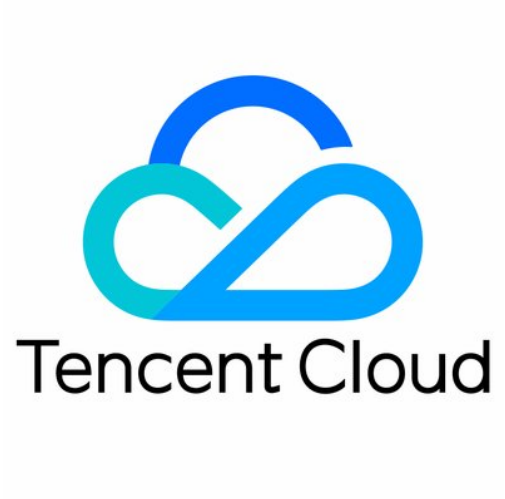
Hyper Computing Cluster

Instance Specifications

Product Documentation



Instance Specifications

Last updated：2024-08-20 17:33:24

Hyper Computing Cluster takes high-performance CVM instances as nodes and interconnects them through RDMA. It provides network services with high bandwidth and ultra-low latency, significantly improves network performance, and meets the parallel computing demands of large-scale high-performance computing, AI, big data recommendation, and other applications.

Instance Overview

Hyper Computing Cluster provides instances of the following specifications:

Purchase	Instance	Instance Type	GPU	Available Image
Recommended	HCCPNV5	GPU	Nvidia H800	TencentOS Server 3.1 (TK4) UEFI Edition
	HCCPNV5v	GPU	Nvidia H800	TencentOS Server 2.4 (TK4)
	HCCPNV4sne	GPU	Nvidia A800	TencentOS Server 2.4 (TK4)
	HCCPNV4sn	GPU	Nvidia A800	TencentOS Server 2.4 (TK4)
	HCCPNV4h	GPU	Nvidia A100	TencentOS Server 2.4 (TK4) Ubuntu Server 18.04 LTS CentOS 7.6
Available	HCCG5vm	GPU	Nvidia V100	TencentOS Server 2.4 (TK4) Ubuntu Server 18.04 LTS CentOS 7.6
	HCCG5v	GPU	Nvidia V100	TencentOS Server 2.4 (TK4) Ubuntu Server 18.04 LTS CentOS 7.6
	HCCS5	Standard	-	TencentOS Server 2.4

				(TK4) Ubuntu Server 18.04 LTS CentOS 7.6
	HCCIC5	Compute	-	TencentOS Server 2.4 (TK4) Ubuntu Server 18.04 LTS CentOS 7.6

Instance Specifications

Refer to the introduction below to choose the instance specifications that meet your business needs, especially the minimum requirements for CPU, memory, GPU, and other resources.

GPU Hyper Computing ClusterPNV5

The GPU Hyper Computing ClusterPNV5 instance is the latest instance equipped with NVIDIA® H800 Tensor Core GPU. GPUs support 400 GB/s NVLink interconnection, and instances support 3.2 Tbps RDMA interconnection, offering high performance.

Note:

To purchase this instance, certain permissions are required. Please contact your pre-sales manager to obtain the permissions.

Application Scenario

Hyper Computing ClusterPNV5 boasts strong floating-point computing capability and applies to large-scale AI and scientific computing scenarios:

Large-scale deep learning training and big data recommendations.

HPC applications, such as computational finance, quantum simulation of materials, and molecular modeling.

Hardware Specifications

CPU: 2.6GHz Intel® Xeon® Sapphire Rapids, with turbo boost up to 3.1GHz.

GPU: 8 × NVIDIA® H800 NVLink® 80GB (FP32 64 TFLOPS, TF32 494 TFLOPS, BF16 989 TFLOPS, 400GB/s NVLink®).

Memory: 8-channel DDR5.

Storage: 8 × 6,400 GB NVMe SSDs for high-performance local storage. CBS disks can be used as system and data disks, supporting scaling out on demand.

Network: Support 100 Gbps private network bandwidth and 3.2 Tbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, and ENIs can be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	GPU	GPU Memory	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)	I
HCCPNV5	192	2048	2.6/3.1	Nvidia H800 × 8	80GB × 8	3.2 Tbps RoCEv2	100	4

Note

GPU driver: It is recommended to install the NVIDIA Tesla driver of 535 or later versions for NVIDIA H800 series GPUs. 535.54.03 (Linux) and 536.25 (Windows) are recommended. For driver version details, see the NVIDIA official documentation.

GPU Hyper Computing ClusterPNV5v

The GPU Hyper Computing ClusterPNV5v instance is the latest instance equipped with NVIDIA® H800 Tensor Core GPU. GPUs support 400 GB/s NVLink interconnection, and instances support 3.2 Tbps RDMA interconnection, offering high performance.

Note:

To purchase this instance, certain permissions are required. Please contact your pre-sales manager to obtain the permissions.

Application Scenario

Hyper Computing ClusterPNV5v boasts strong floating-point computing capability and applies to large-scale AI and scientific computing scenarios:

Large-scale deep learning training and big data recommendations.

HPC applications, such as computational finance, quantum simulation of materials, and molecular modeling.

Hardware Specifications

CPU: 2.6GHz Intel® Xeon® Sapphire Rapids, with turbo boost up to 3.1GHz.

GPU: 8 × NVIDIA® H800 NVLink® 80GB (FP32 64 TFLOPS, TF32 494 TFLOPS, BF16 989 TFLOPS, 400GB/s NVLink®).
Memory: 8-channel DDR5.
Storage: 8 × 6,400 GB NVMe SSDs for high-performance local storage. CBS disks can be used as system and data disks, supporting scaling out on demand.
Network: Support 100 Gbps private network bandwidth and 3.2 Tbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, and ENIs can be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	GPU	GPU Memory	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)	I
HCCPNV5v	172	1939	2.6/3.1	Nvidia H800 × 8	80GB × 8	3.2 Tbps RoCEv2	100	

Note
GPU driver: It is recommended to install the NVIDIA Tesla driver of 535 or later versions for NVIDIA H800 series GPUs. 535.54.03 (Linux) and 536.25 (Windows) are recommended. For driver version details, see the NVIDIA official documentation.

GPU Hyper Computing ClusterPNV4sne

The GPU Hyper Computing ClusterPNV4sne instance is a new instance equipped with NVIDIA® A800 Tensor Core GPU. GPUs support 400 GB/s NVLink interconnection, and instances support 1.6 Tbps RDMA interconnection, offering high performance.
Note:
To purchase this instance, certain permissions are required. Please contact your pre-sales manager to obtain the permissions.

Application Scenario

Hyper Computing ClusterPNV4sne boasts strong floating-point computing capability and applies to large-scale AI and scientific computing scenarios:
Large-scale deep learning training and big data recommendations.

HPC applications, such as computational finance, quantum simulation of materials, molecular modeling, and gene sequencing.

Hardware Specifications

CPU: 2.7GHz Intel® Xeon® Ice Lake, with turbo boost up to 3.3GHz.

GPU: 8 × NVIDIA® A800 NVLink® 80GB (FP64 9.7 TFLOPS, TF32 156 TFLOPS, BF16 312 TFLOPS, 400GB/s NVLink®).

Memory: 8-channel DDR4.

Storage: 4 × 6,400 GB NVMe SSDs for high-performance local storage. CBS disks can be used as system and data disks, supporting scaling out on demand.

Network: Support 100 Gbps private network bandwidth and 1.6 Tbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, and ENIs can be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	GPU	GPU Memory	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)
HCCPNV4sne	124	1929	2.7/3.3	Nvidia A800 × 8	80GB × 8	1.6 Tbps RoCEv2	100

Note:

GPU driver: It is recommended to install the NVIDIA Tesla driver of 450 or later versions for NVIDIA A800 series GPUs. 460.32.03 (Linux) and 461.33 (Windows) are recommended. For driver version details, see the NVIDIA official documentation.

GPU Hyper Computing ClusterPNV4sn

The GPU Hyper Computing ClusterPNV4sn instance is a new instance equipped with NVIDIA® A800 Tensor Core GPU. GPUs support 400 GB/s NVLink interconnection, and instances support 800 Gbps RDMA interconnection, offering high performance.

Note:

To purchase this instance, certain permissions are required. Please contact your pre-sales manager to obtain the permissions.

Application Scenario

Hyper Computing ClusterPNV4sn boasts strong floating-point computing capability and applies to large-scale AI and scientific computing scenarios:
Large-scale deep learning training and big data recommendations.
HPC applications, such as computational finance, quantum simulation of materials, molecular modeling, and gene sequencing.

Hardware Specifications

CPU: 2.55GHz AMD EPYC™ Milan, with turbo boost up to 3.5GHz.
GPU: 8 × NVIDIA® A800 NVLink® 80GB (FP64 9.7 TFLOPS, TF32 156 TFLOPS, BF16 312 TFLOPS, 400GB/s NVLink®).
Memory: 8-channel DDR4.
Storage: 2 × 7,680 GB NVMe SSDs for high-performance local storage. CBS disks can be used as system and data disks, supporting scaling out on demand.
Network: Support 100 Gbps private network bandwidth and 800 Gbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, and ENIs can be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	GPU	GPU Memory	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)
HCCPNV4sn	232	1929	2.55/3.5	Nvidia A800 × 8	80GB × 8	800 Gbps RoCEv2	100

Note:
GPU driver: It is recommended to install the NVIDIA Tesla driver of 450 or later versions for NVIDIA A800 series GPUs. 460.32.03 (Linux) and 461.33 (Windows) are recommended. For driver version details, see the NVIDIA official documentation.

GPU Hyper Computing ClusterPNV4h

The GPU Hyper Computing ClusterPNV4h instance is a new instance equipped with NVIDIA® A100 Tensor Core GPU. It uses NVMe SSDs as instance storage with low latency, ultra-high IOPS, and high throughput, offering high

performance.

Application Scenario

Hyper Computing ClusterPNV4h boasts strong double-precision floating-point computing capability and applies to large-scale AI and scientific computing scenarios:
Large-scale machine learning training and big data recommendations.
HPC applications, such as computational finance, quantum simulation of materials, molecular modeling, and gene sequencing.

Hardware Specifications

CPU: 2.6GHz AMD EPYC™ ROME, with turbo boost up to 3.3GHz.
GPU: 8 × NVIDIA® A100 NVLink® 40GB (FP64 9.7 TFLOPS, TF32 156 TFLOPS, BF16 312 TFLOPS, 600GB/s NVLink®).
Memory: 8-channel DDR4.
Storage: 1 × 480 GB SATA SSD as local system disk and 4 × 3,200 GB NVMe SSDs for high-performance local storage. CBS disks cannot be mounted.
Network: Support 25 Gbps private network bandwidth and 100 Gbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, but ENIs cannot be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	GPU	GPU Memory	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)	I
HCCPNV4h	192	1024	2.6/3.3	Nvidia A100 × 8	40GB × 8	100 Gbps RoCEv2	25	

Note:
GPU driver: It is recommended to install the NVIDIA Tesla driver of 450 or later versions for NVIDIA A100 series GPUs. 460.32.03 (Linux) and 461.33 (Windows) are recommended. For driver version details, see the NVIDIA official documentation.

GPU Hyper Computing ClusterG5vm

The GPU Hyper Computing ClusterG5vm instance is equipped with NVIDIA® Tesla® V100 GPU and uses NVMe SSDs for instance storage with low latency, ultra-high IOPS, and high throughput, offering high performance.

Application Scenario

Large-scale machine learning training and big data recommendations.
HPC applications, such as computational finance, quantum simulation of materials, molecular modeling, and gene sequencing.

Hardware Specifications

CPU: 2.5GHz Intel® Xeon® Cascade Lake, with turbo boost up to 3.1GHz.
GPU: 8 × NVIDIA® Tesla® V100 GPU (FP64 7.8 TFLOPS, FP32 15.7 TFLOPS, 300GB/s NVLink®).
Memory: 6-channel DDR4.
Storage: 1 × 480 GB SATA SSD as local system disk and 4 × 3,200 GB NVMe SSDs for high-performance local storage. CBS disks cannot be mounted.
Network: Support 25 Gbps private network bandwidth and 100 Gbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, but ENIs cannot be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	GPU	GPU Memory	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)	I
HCCG5vm	96	768	2.5/3.1	Nvidia V100 × 8	32GB × 8	100 Gbps RoCEv2	25	

GPU Hyper Computing ClusterG5v

The GPU Hyper Computing ClusterG5v instance is equipped with NVIDIA® Tesla® V100 GPU and uses NVMe SSDs for instance storage with low latency, ultra-high IOPS, and high throughput, offering high performance.

Application Scenario

Large-scale machine learning training and big data recommendations.
HPC applications, such as computational finance, quantum simulation of materials, molecular modeling, and gene sequencing.

Hardware Specifications

CPU: 2.5GHz Intel® Xeon® Cascade Lake, with turbo boost up to 3.1GHz.
GPU: 8 × NVIDIA® Tesla® V100 GPU (FP64 7.8 TFLOPS, FP32 15.7 TFLOPS, 300GB/s NVLink®).
Memory: 6-channel DDR4.
Storage: 1 × 480 GB SATA SSD as local system disk and 4 × 3,200 GB NVMe SSDs for high-performance local storage. CBS disks cannot be mounted.
Network: Support 25 Gbps private network bandwidth and 100 Gbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, but ENIs cannot be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	GPU	GPU Memory	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)	
HCCG5v	96	384	2.5/3.1	Nvidia V100 × 8	32GB × 8	100 Gbps RoCEv2	25	

Standard Hyper Computing ClusterS5

The Standard Hyper Computing ClusterS5 instance is equipped with 2.5 GHz CPU and applies to general multi-core batch processing, multi-core high-performance computing, and other compute-intensive applications.

Application Scenario

Large-scale high-performance computing applications.
HPC applications, such as fluid dynamics analysis, industrial simulation, molecular modeling, gene sequencing, and meteorological analysis.

Hardware Specifications

CPU: 2.5GHz Intel® Xeon® Cascade Lake, with turbo boost up to 3.1GHz.
Memory: 6-channel DDR4.
Storage: 1 × 480 GB SATA SSD. CBS disks cannot be mounted.
Network: Support 25 Gbps private network bandwidth and 100 Gbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, but ENIs cannot be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)	Packet Tx/Rx (pps)	Number of Queues
HCCS5	96	384	2.5/3.1	100 Gbps RoCEv2	25	10 million	16

Compute Hyper Computing ClusterIC5

The high-I/O Compute Hyper Computing ClusterIC5 instance is equipped with 3.2 GHz CPU featuring high single-core computing performance. It uses NVMe SSDs for instance storage with low latency and ultra-high IOPS. It applies to batch processing, fluid dynamics analysis, structural simulation, and other compute-intensive and IO-intensive applications.

Application Scenario

Large-scale high-performance computing applications.
HPC applications, such as fluid dynamics analysis, industrial simulation, molecular modeling, gene sequencing, and meteorological analysis.

Hardware Specifications

CPU: 3.2GHz Intel® Xeon® Cascade Lake, with turbo boost up to 3.7GHz.
Memory: 6-channel DDR4.

Storage: 2 × 480 GB SATA SSDs (RAID1) as local system disks and 2 × 3,840 GB NVMe SSDs for high-performance local storage. CBS disks cannot be mounted.
Network: Support 25 Gbps private network bandwidth and 100 Gbps low-latency RDMA network dedicated to internal communication of Hyper Computing Cluster instances, with strong packet transporting and receiving capabilities. The public network can be configured as needed, but ENIs cannot be mounted.

Specification	vCPU	Memory (GiB)	Clock Speed/Turbo Boost (GHz)	RDMA Configuration	Private Network Bandwidth Capacity (Gbps)	Packet Tx/Rx (pps)	Number of Queues
HCCIC5	64	384	3.2/3.7	100 Gbps RoCEv2	25	10 million	16