

目录

目录	1
产品概述	2
基本概念	2
生命周期	2
产品优势	2
选型丰富	2
强大算力	2
方便易用	2
安全高效	2
高性价比	2
弹性扩容	2
功能概览	3
快速创建	3
Web化实例管理	3
多种类型	3
自定义镜像	3
VPC支持	3
统计和告警	3
产品类型	3
实例类型概览	3
GPU通用计算型P4V	3
GPU高效计算型P5V	4
GPU推理 II 型GN6I	4
GPU通用计算型GN7 I	4
GPU虚拟化vGN6	5

产品概述

金山云GPU云服务器（GPU Elastic Compute，简称GEC）提供通用GPU加速计算，可以用于科学计算，深度学习，图形图像渲染与基于GPU的音视频编解码等诸多应用场景。为用户提供稳定，快速与弹性的计算服务与便捷统一的云服务器管理方式。

GPU云服务器GEC典型的应用场景包括深度学习的离线训练和在线预测等。

利用GPU的强大计算能力，GPU云服务器GEC可作为深度学习的训练和预测平台。同时，可结合对象存储KS3 提供的云存储服务，云数据库KRDS提供的在线数据库服务、大数据平台KMR提供的海量分布式处理服务，您可以搭建一个功能完备的深度学习系统，帮助您安全、高效的进行各种深度学习的模型训练和在线服务需求。

基本概念

使用GPU云服务器之前，您需要了解以下几个概念：	概念	说明
实例	等同于一台虚拟机，能提供高性能的计算资源，包含CPU、内存、操作系统、网络等基础的计算组件。	
实例启动模板	实例启动模板保存了一系列启动参数配置，帮助用户免去创建云服务器的重复配置，可用于快速创建实例。	
镜像	指GEC实例运行环境的模板，一般包括操作系统和预装的软件。操作系统支持多种Linux发行版本和不同的Windows版本。	
快照	指某一个时间点上一块弹性块存储的数据备份。可以用于备份或恢复弹性块存储的数据。	
VPC	金山云构建的一个虚拟的网络环境，与其他资源是相互隔离的，您可以完全掌控自己的专有网络。	
弹性IP	专为动态网络设计的静态公网 IP，可以绑定到用户的任何一台云服务器中。	
安全组	一种虚拟防火墙，是一个逻辑上的分组，可以将同一地域具有相同网络安全隔离需求的云服务器加到同一个安全组内。	
容灾组	是指一个保证云服务器绝对分散到不同的物理机上的集合，从而保证云服务器绝对的物理分散，提高服务的高可用性，可动态加入和移除云服务器。	
地域和可用区	地域是指金山云的物理数据中心所处的不同地理区域。不同地域之间完全隔离，以保证所提供服务的稳定性与容错性。可用区是指同一地域内，电力和网络等基础设施互相隔离的物理数据中心。	

生命周期

金山云GPU云服务器的生命周期如下图所示：

产品优势

选型丰富

GPU云服务器 (GEC) 针对不同类型的GPU推出多种套餐配置，满足各类行业场景的需求。

强大算力

GPU超强的并行计算力与强大的数据吞吐力，可显著提高大规模运算、复杂AI模型计算、及视频图像处理等应用性能。

方便易用

GPU云服务器（GEC）采用和云服务器（KEC）一致的操作和管理，无需额外学习。

安全高效

您可以随时掌握GPU等资源使用情况；即时运维，不同用户之间资源安全隔离，有效保障您的数据安全。

高性价比

GPU云服务器 (GEC) 免去您采购、运维、频繁升级硬件带来的成本压力。GEC支持包年包月和按日配置月结，您可以根据需要选择合适的计费模式。

弹性扩容

支持配置的弹性扩容，按需购买。为客户节省运营成本，提升资源有效利用率。

功能概览

快速创建

一键式创建，分钟级部署。

Web化实例管理

通过Web控制台可实现对GPU加速型云服务器实例的新建、查看、续费和开关机等全生命周期管理操作。

多种类型

GPU云服务器支持多类型与多配置，适于不同应用场景下的不同业务规模。

自定义镜像

支持用户自主创建镜像，以及使用该镜像创建实例。

VPC支持

原生支持VPC虚拟专有网络，提供灵活的网络规划支持，便利用户使用VPC内的各种资源。

统计和告警

提供丰富的检测功能，实时检测CPU、GPU、磁盘性能、网络流量等，业务负载一目了然，还可以自定义阈值配置和检测告警，让用户灵活掌控业务变化。

产品类型

GPU云服务器针对典型应用场景，提供多种产品类型供用户选择。各类型产品所采用的硬件（GPU、CPU、内存和硬盘）及网络资源配置各有不同。

本节将详细介绍产品适用场景、型号及配置信息。

实例类型概览

GPU云服务器可分为两大类，详见下表：

GPU云服务器	实例类型	适用场景
直通（Passthrough）	GPU通用计算型P4V GPU高效计算型P5V GPU推理 II 型GN6I GPU通用计算型GN7I	深度学习、语音、图形/图像学习等常见训练和推理场景
vGPU	GPU虚拟化vGN6	云端渲染和小规模、弹性、灵活的AI应用场景

GPU通用计算型P4V

- 计算：
 - 处理器：2.6 GHz主频的Intel® Xeon® Processor E5-2690 v4
 - 基于NVIDIA Tesla V100，每GPU具备16GB HBM2 GPU内存、15TFLOPS的单精度（FP32）计算能力和125TFLOPS的混合精度计算能力。
- 存储：
 - 支持系统盘类型：本地SSD
 - 支持数据盘类型：本地SSD、EBS3.0
- 使用场景：

- 该实例适用于深度学习的训练场景和推理场景。
- P4V实例包括的型号和参数规格如下表所示：

型号	GPU (Tesla V100)	GPU显存 (HB M2)	vCPU (核)	内存 (DDR 4)	数据盘 (本地 SSD)	网络收发包能力 (万PPS)	网络带宽能力 (Gbit/s)
P4V. 8A1	1颗	16GB*1	8	32GB	240GB	20	1.5
P4V. 16B2	2颗	16GB*2	16	64GB	480GB	30	3
P4V. 28C4	4颗	16GB*4	28	128GB	960GB	30	6
P4V. 56D8	8颗	16GB*8	56	256GB	1920GB	40	8

GPU高效计算型P5V

- 计算：
 - 处理器： 2.6 GHz主频的Intel Xeon Platinum 8358P Processor
 - 基于NVIDIA A100，每块GPU具备**80GB GDDR6**显存，8.1TFLOPS的单精度（FP32）计算能力和130 TOPS的INT8计算能力，多卡之间以NVSwitch实现两两互联。
- 存储：
 - 支持系统盘类型：EBS3.0
 - 支持数据盘类型：EBS3.0
- 使用场景：
 - 可用于深度学习 、高性能数据分析和高性能计算应用场景。
 - 深度学习，例如：无人驾驶、对象检测、语音翻译识别等人工智能算法训练。
 - 高性能数据分析和高性能计算，例如石油勘探、生命科学、气象环境分析等场景。
- P5V实例包括的型号和参数规格如下表所示：

型号	GPU	GPU显存 (GDDR 6)	vCPU (核)	内存 (GiB)	网络收发包能力 (万P PS)	网络带宽能力 (Gbit /s)	多队列
P5V. 14A1	A100*1	80GB*1	14	112	60	5	8
P5V. 28B2	A100*2	80GB*2	28	224	120	10	16
P5V. 56C4	A100*4	80GB*4	56	448	250	25	16
P5V. 112D8	A100*8	80GB*8	112	896	500	25	32

GPU推理 II 型GN6I

- 计算：
 - 处理器： 2.6 GHz主频的Intel® Xeon® Gold 6240 Processor
 - 基于NVIDIA Tesla T4，每GPU具备16GB GDDR6显存、8.1TFLOPS的单精度（FP32）计算能力和130 TOPS的INT8计算能力。
- 存储：
 - 支持系统盘类型：EBS3.0
 - 支持数据盘类型：EBS3.0
- 使用场景：
 - 该实例适用于推理场景，以及简单的训练场景。
- GN6I实例包括的型号和参数规格如下表所示：

型号	GPU	GPU显存 (GDDR 6)	vCPU (核)	内存 (GiB)	网络收发包能力 (万P PS)	网络带宽能力 (Gbit /s)	多队列
GN6I. 4A1	T4*1	16GB*1	4	16	50	4	2
GN6I. 8A1	T4*1	16GB*1	8	32	80	5	2
GN6I. 16A1	T4*1	16GB*1	16	64	120	6	4
GN6I. 16B2	T4*2	16GB*2	16	64	120	6	4
GN6I. 32B2	T4*2	16GB*2	32	128	240	8	8
GN6I. 32C4	T4*4	16GB*4	32	128	240	8	8

GPU通用计算型GN7I

- 计算：

- 处理器：2.6 GHz主频的Intel Xeon Platinum 8358P
- 基于NVIDIA A10，每GPU具备24GB GDDR6显存、31.2TFLOPS的单精度（FP32）计算能力和250 TOPS的INT8计算能力。
- 存储：
 - 支持系统盘类型：SSD3.0和高效云盘EHDD
 - 支持数据盘类型：SSD3.0和高效云盘EHDD
- 使用场景：
 - 该实例适用于推理场景，以及简单的训练场景。
- GN7I实例包括的型号和参数规格如下表所示：

型号	GPU	GPU显存（GDDR6）	vCPU（核）	内存（GiB）	网络收发包能力（万PPS）	网络带宽能力（Gbit/s）	多队列
GN7I.16C1	A10*1	24GB*1	16	64	150	6	16
GN7I.16D1	A10*1	24GB*1	16	96	150	6	16
GN7I.32D1	A10*1	24GB*1	32	192	300	12	16
GN7I.64D2	A10*2	24GB*2	64	384	500	20	16
GN7I.112D4	A10*4	24GB*4	112	719	500	20	16

GPU虚拟化vGN6

- 计算：
 - GPU：采用NVIDIA T4 GPU
 - 处理器：2.6 GHz主频的Intel® Xeon® Gold 6240 Processor
- vGPU类别
 - vCS：专门用于深度学习，提供1/2*Tesla T4、1/4*Tesla T4两种实例
 - vPC：图形/图像处理场景，提供1/8*Tesla T4实例
- 存储：
 - 支持系统盘类型：EBS3.0
 - 支持数据盘类型：EBS3.0
- 使用场景：
 - 云游戏的云端实时渲染
 - AR/VR的云端实时渲染
 - AI（深度学习DL/机器学习ML）
- vGN6实例包括的型号和参数规格如下表所示：

型号	GPU（Tesla T4）	GPU显存（GDDR6）	vCPU（核）	内存（DDR4）	网络收发包能力（万PPS）	网络带宽能力（Gbit/s）
vGN6.vCS-10B2	1/2颗	8GB	10	40GB	80	3
vGN6.vCS-4C4	1/4颗	4GB	4	20GB	50	2
vGN6.vPC-2D8	1/8颗	2GB	2	10GB	30	1

其中vPC适用于图形图像处理，vCS适用于CUDA计算，如AI推理等。 关于vGN6的具体配置方法，可以参考[vGPU用户指南](#)。