目录

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。

产品功能与优势

GPU云服务器提供易于用户管理的功能，包括：

- **快速创建：** 一键式创建，分钟级部署。
- **Web化管理：** 通过Web控制台可实现对GPU云服务器实例的创建、查看、续费和开关机等生命周期管理操作。
- **VPC支持：** GPU云服务器实例原生支持VPC专有网络，提供灵活的网络规划选择，便捷用户使用VPC内的各种资源。
- **监控统计：** 提供实时详细的实例监控，性能高低一目了然。具体介绍请参考云监控。
- **挂载云硬盘：** GPU云服务器支持挂载云硬盘。更多挂载云硬盘的内容，请参考挂载云硬盘。

GPU云服务器可满足用户高性能的需求，并具备轻资产、全面的安全性等优势。

- **高性能：** 单个GPU实例总计可以提供8颗NVIDIA Tesla V100加速器，配备256GB内存，总计提供40960CUDA cores、5120Tensor cores和最高112TFLOPS的单精度浮点计算能力，同时采用最新的Volta架构，为深度学习和高性能计算应用提供了卓越的性能。
- **轻资产：** 低服务器投资风险，低服务器运维成本，始终用到最新GPU加速卡硬件。
- **安全性：** 用户私有网络（VPC）隔离，全面支持金山云安全产品。
- **易用性：** 图形化控制台管理，可灵活与金山云其他产品配合使用，包括KEC、VPC、RDS、Redis、KS3等。

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P3 • GPU推理计算型P3I • GPU推理计算型P3IN • GPU通用计算型P4V • GPU推理Ⅱ型GN6I	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vGN5 • GPU虚拟化vGN6	云端渲染和小规模、弹性、灵活的AI应用场景

GPU推理Ⅱ型GN6I

该实例适用于推理场景，以及简单的训练场景。

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。

实例特点包括：

- 处理器：2.6 GHz主频的Intel® Xeon® Gold 6240 Processor
- 支持系统盘类型：EBS3.0
- 支持数据盘类型：EBS3.0

GN6I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
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P3

该实例适用于深度学习的训练场景和推理场景。

基于NVIDIA Tesla P40，每GPU具备24GB DDR5 GPU内存、12TFLOPS的单精度（FP32）计算能力和46TOPS的INT8计算能力。

实例特点包括：

- 处理器：2.6 GHz主频的Intel® Xeon® Processor E5-2690 v4
- 支持系统盘类型：本地SSD
- 支持数据盘类型：本地SSD、EBS3.0

P3实例包括的型号和参数规格如下表所示：

型号	GPU（Tesla P40）	GPU显存（GDDR5）	vCPU（核）	内存（DDR4）	数据盘（本地SSD）	网络收发包能力（万PPS）	网络带宽能力（Gbit/s）
P3.28A1	1颗	24GB*1	28	56GB	1000GB	30	3
P3.56B2	2颗	24GB*2	56	112GB	2000GB	40	6
P3.56C4	4颗	24GB*4	56	224GB	4000GB	40	8

GPU推理计算型P3I

该实例适用于语音识别、语音合成、图像识别等推理预测场景。

基于NVIDIA Tesla P4，每GPU具备8GB DDR5 GPU内存、5.5TFLOPS的单精度（FP32）计算能力和22TOPS的INT8计算能力，单GPU实例在深度学习的推理预测场景下相比于CPU延时降低15倍，吞吐增加60倍。

实例特点包括：

- 处理器：2.6 GHz主频的Intel® Xeon® Processor E5-2690 v4
- 支持系统盘类型：本地SSD
- 支持数据盘类型：本地SSD、EBS3.0

P3I实例包括的型号和参数规格如下表所示：

型号	GPU（Tesla P4）	GPU显存（GDDR5）	vCPU（核）	内存（DDR4）	数据盘（本地SSD）	网络收发包能力（万PPS）	网络带宽能力（Gbit/s）
P3I.14B1	1颗	8GB*1	14	120GB	500GB	20	3
P3I.28C2	2颗	8GB*2	28	240GB	1000GB	30	6

GPU推理计算型P3IN

实例特点包括：

- 处理器：2.6 GHz主频的Intel® Xeon® Processor E5-2690 v4
- 支持系统盘类型：本地SSD
- 支持数据盘类型：本地SSD、EBS3.0

该实例的适用场景以及采用的硬件与P3I一致，包括的型号和参数规格如下表所示：

型号	GPU（Tesla P4）	GPU显存（GDDR5）	vCPU（核）	内存（DDR4）	数据盘（本地SSD）	网络收发包能力（万PPS）	网络带宽能力（Gbit/s）
P3IN.4A1	1颗	8GB*1	4	16GB	120GB	10	1.5
P3IN.8B1	1颗	8GB*1	8	32GB	180GB	20	1.5
P3IN.16C2	2颗	8GB*2	16	64GB	360GB	30	3

P3IN.32D4	4颗	8GB*4	32	128GB	720GB	40	6

GPU通用计算型P4V

该实例适用于深度学习的训练场景和推理场景。

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。

实例特点包括：

- 处理器：2.6 GHz主频的Intel® Xeon® Processor E5-2690 v4
- 支持系统盘类型：本地SSD
- 支持数据盘类型：本地SSD、EBS3.0

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（HBM2）	vCPU（核）	内存（DDR4）	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虚拟化vGN5

该实例的适用场景包括：

- 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）

实例特点包括：

- GPU：采用NVIDIA P4 GPU
- 处理器：2.6 GHz主频的Intel ® Xeon ® E5-2690 v4 (Broadwell)
- 支持系统盘类型：本地SSD
- 支持数据盘类型：本地SSD、EBS3.0
- vGPU类别
  - vCS：专门用于深度学习，提供1*Tesla P4、1/2*Tesla P4两种实例
  - vPC：图形/图像处理场景，提供1/4*Tesla P4、1/8*Tesla P4两种实例

vGN5实例包括的型号和参数规格如下表所示：

型号	GPU（Tesla P4）	GPU显存（GDDR5）	vCPU（核）	内存（DDR4）	数据盘（本地SSD）	网络收发包能力（万PPS）	网络带宽能力（Gbit/s）
vGN5.vCS-8B2	1/2颗	4GB	8	48GB	400GB	20	2
vGN5.vPC-4C4	1/4颗	2GB	4	24GB	200GB	10	1
vGN5.vPC-2D8	1/8颗	1GB	2	12GB	100GB	10	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vGN5的具体使用方法，可以参考vGPU用户指南。

GPU虚拟化vGN6

该实例的适用场景包括：

- 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）

实例特点包括：

- GPU：采用NVIDIA T4 GPU
- 处理器：2.6 GHz主频的Intel® Xeon® Gold 6240 Processor
- 支持系统盘类型：EBS3.0

- 支持数据盘类型：EBS3.0
- 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

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。