什么是GPU云服务器

更新时间：2020-11-23 16:37:43

[**本页目录**](javascript:void(0))

* [EGS平台介绍](https://help.aliyun.com/document_detail/163791.html?spm=a2c4g.11174283.6.544.6cd62d155im6zY#title-1sq-woo-x7l)
* [功能特性](https://help.aliyun.com/document_detail/163791.html?spm=a2c4g.11174283.6.544.6cd62d155im6zY#title-enf-4ec-i7t)
* [相关工具](https://help.aliyun.com/document_detail/163791.html?spm=a2c4g.11174283.6.544.6cd62d155im6zY#title-bvr-v0t-ry4)

GPU云服务器提供GPU加速计算能力，实现GPU计算资源的即开即用和弹性伸缩。作为阿里云弹性计算家族的一员，GPU云服务器结合了GPU计算力与CPU计算力，满足您在人工智能、高性能计算、专业图形图像处理等场景中的需求。

**EGS平台介绍**

GPU作为一种计算芯片，其优势非常明显，具有实时高速、强并行计算能力和强浮点计算能力等特点。阿里云弹性GPU服务EGS（Elastic GPU Service）将阿里云的弹性计算和GPU高速的并行异构加速器组合在一起，兼具弹性计算的特点和GPU的加速能力。

阿里云基于EGS平台推出了GPU云服务器，在提供GPU加速能力的同时，保留了与普通ECS实例一致的使用体验。您在创建ECS实例时，选择企业级异构计算规格即可。实例规格的详细信息，请参见[实例规格族](https://help.aliyun.com/document_detail/25378.html" \l "concept-sx4-lxv-tdb" \o "实例是能够为您的业务提供计算服务的最小单位，不同的实例规格可以提供的计算能力不同。本章节为您介绍在售的所有ECS实例规格族，包括每种实例规格族的特点、在售规格和适用场景。)。

**功能特性**

* 高弹性

提供系列化的规格族，分钟级创建GPU实例，支持水平扩容和垂直变配。

* 高性能高安全性

支持GPUDirect，GPU之间点对点通信。GPU可以直接通过NVLink总线互联通信，具备高带宽低时延的特点，无需CPU干预。多租户间弹性GPU安全隔离 ，并通过Hypervisor授权和管理。允许灵活配置多块GPU实现互相高速通信的同时，还具备隔离带来的安全性。

* 易部署

和阿里云生态深度融合，您可以轻松搭配其他阿里云产品构建应用，例如搭配OSS、NAS等产品满足存储要求，搭配EMR进行深度学习数据预处理等。支持云原生（阿里云Kubernetes），交付更加便捷。

* 易监控

提供全面的GPU监控数据，包括GPU、实例和分组维度，免去您的运维压力。详情请参见[GPU监控](https://help.aliyun.com/document_detail/84885.html#concept-ysn-byq-32b)。

应用场景

更新时间：2020-04-29 17:14:19

[**本页目录**](javascript:void(0))

* [直播实时视频转码](https://help.aliyun.com/document_detail/163794.html?spm=a2c4g.11186623.6.546.d346739dPu1UQT#title-pfv-bjy-kof)
* [AI训练](https://help.aliyun.com/document_detail/163794.html?spm=a2c4g.11186623.6.546.d346739dPu1UQT#title-qz0-vx8-jm0)
* [AI推理](https://help.aliyun.com/document_detail/163794.html?spm=a2c4g.11186623.6.546.d346739dPu1UQT#title-v61-pic-ful)
* [云游戏、云网吧和云端图形工作站](https://help.aliyun.com/document_detail/163794.html?spm=a2c4g.11186623.6.546.d346739dPu1UQT#title-p2q-28s-ljg)

GPU云服务器适用于视频转码、图片渲染、AI训练、AI推理、云端图形工作站等场景。

**直播实时视频转码**

阿里云异构GPU和FPGA服务器重点支持2019年双11猫晚直播的实时视频转码，以高画质、低带宽、高分辨率、实时的综合优势服务猫晚当天直播业务4K、2K、1080P等各个分辨率的转码。其中FPGA H.265高清编码、720P节省带宽21.6%，GPU云服务器支持高并发实时视频流5000路以上，并逐步上升到峰值6200路每分钟，且顺利度过流量洪峰。异构GPU云服务器还参与实时家居渲染图片生成等业务，首次提供了大量算力强劲的ebmgn6v裸金属实例，支持淘宝渲染方提升几十倍的渲染性能，第一次实现秒级实时渲染，完成总计超过5000张大型家居渲染图。异构FPGA图片转码业务则以3K+片的超大集群，为淘宝图片空间提供高达数百万QPS的处理能力，承担了双十一淘宝图片85%的流量，预计节省计算成本3亿。

**AI训练**

gn6v和gn6e具有优异的通用GPU计算加速能力，适合为深度学习提供加速引擎。

gn6v和gn6e实例分别配备具有16 GB显存和32 GB显存的NVIDIA V100 GPU计算卡，单节点可提供高达1000TFLOPS的混合精度计算能力。与弹性计算生态的完美结合，为在线和离线场景提供了通用的解决方案。搭配容器服务使用，可以简化部署和运维的复杂度，并提供资源调度服务。

**AI推理**

gn6i具有优异的AI推理能力。

gn6i实例基于配备NVIDIA Tesla T4 GPU计算卡，单精度浮点计算能力最高可达8.1 TFLOPS，int8定点运算处理能力最高可达130 TOPS，支持混合精度，满足了深度学习（尤其是推理）场景下的算力需求。同时单卡功耗仅75 W，具有极高的性能功耗比。与弹性计算生态的完美结合，为在线和离线场景提供了通用的解决方案。搭配容器服务使用，可以简化部署和运维的复杂度，并提供资源调度服务。镜像市场提供预装NVIDIA GPU驱动和深度学习框架的镜像，简化您的部署操作。

**云游戏、云网吧和云端图形工作站**

vgn6i和gn6i均采用基于Turing架构的NVIDIA Tesla T4 GPU加速器，具有极佳的图形计算能力。vgn6i提供分片虚拟化后的虚拟GPU，支持T4 GPU计算能力的1/2、1/4、1/8，并提供极佳的3D图像渲染能力，适用于云游戏、云网吧等场景。vgn6i和gn6i实例可以结合云桌面产品提供云端图形工作站服务，应用于影视动画设计、工业设计、医疗成像、高性能计算的结果呈现等场景。