ALIBABA CLOUD

阿里云

弹性加速计算实例

产品简介

文档版本：20210119

阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。

2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。

3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。

4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。

5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。

6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	危险 重置操作将丢失用户配置数据。
警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	警告 重启操作将导致业务中断，恢复业务时间约十分钟。
注意	用于警示信息、补充说明等，是用户必须了解的内容。	注意 权重设置为0，该服务器不会再接受新请求。
说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击**设置**> **网络**> **设置网络类型**。
粗体	表示按键、菜单、页面名称等UI元素。	在**结果确认**页面，单击**确定**。
Courier字体	命令或代码。	执行 `cd /d C:/window` 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	`bae log list --instanceid` *Instance_ID*
[] 或者 [a|b]	表示可选项，至多选择一个。	`ipconfig [-all|-t]`
{} 或者 {a|b}	表示必选项，至多选择一个。	`switch {active|stand}`

目录

1.什么是弹性加速计算实例EAIS

弹性加速计算实例EAIS（Elastic Accelerated Computing Intances）是一款阿里云提供的性能卓越、成本优化、弹性扩展的IaaS（Infrastructure as a Service）级别弹性计算服务。EAIS可以将CPU资源与GPU资源成功解耦，帮助您将GPU资源附加到ECS实例上，构建成您希望得到的GPU实例规格，用于推理场景下的弹性使用，从而提高资源利用率，降低成本。

为什么选择EAIS

选择EAIS，您可以轻松构建具有以下优势的异构计算资源：

- 无需自建机房，无需采购以及配置硬件设施。
- 分钟级交付，快速部署，缩短应用上线周期。
- 快速接入部署在全球范围内的数据中心和BGP（Border Gateway Protocol，边界网关协议）机房。
- 成本透明，按需使用，支持根据业务波动随时扩展和释放资源。
- 提供任意ECS实例与GPU、NPU、FPGA等异构计算搭配的实例类型，满足您的多种需求。
- 支持通过内网访问其他阿里云服务，形成丰富的行业解决方案，降低公网流量成本。
- 提供虚拟防火墙、角色权限控制、内网隔离、防病毒攻击及流量监控等多重安全方案。
- 提供性能监控框架和主动运维体系。
- 提供行业通用标准API，提高易用性和适用性。

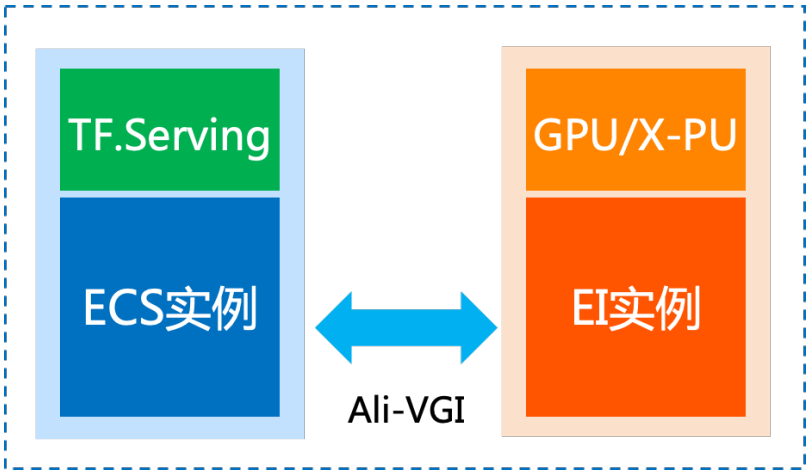
更多信息，请参见产品优势和应用场景。

产品架构

EAIS主要包含以下功能组件：

- 云服务器ECS包含的所有功能组件。更多信息，请参见产品架构。
- 异构计算资源：一块或1/N块GPU或NPU。

以下为EAIS的产品组件架构图，图中涉及的功能组件的详细介绍请参见相应的帮助文档。



产品定价

EAIS支持按量付费计算模式。如需使用，请前往弹性加速计算实例控制台购买，收费标准以控制台实际定价为准。

地域及可用区

EAIS目前仅支持国内部分地域。更多信息，请参见地域和可用区。

其他

EAIS所绑定的云服务器ECS实例需要遵循其自身的部署建议，同时也可以选择ECS支持的相关服务。

2.产品优势

与常规的异构实例相比，阿里云弹性加速计算实例EAIS提供的异构实例具有解耦性、低成本、弹性和多适配性的优势。

解耦性

常规的GPU实例，其CPU、内存和GPU是部署在同一台物理机中。EAIS实例可以将CPU与GPU成功解耦，其CPU、内存和GPU可以存在于不同的物理机中。您可以根据对CPU和内存的需求选择一款ECS实例，然后再匹配一个EAIS实例，即可生成一款满足您需求的新规格GPU实例。

低成本

EAIS能够将推理成本降低多达50%。您可以单独制定所需的推理加速量级，无需超额预置GPU资源，选择最合适您应用的实例类型即可。

例如，您需要一个内存超过128GB，且只带一块GPU的实例用于推荐系统，那么在常规的GPU实例规格中，能刚好满足您需求的实例规格有如下几种选择：

实例规格	vCPU	内存（GiB）	GPU	GPU显存（GB）
ecs.gn6i-c24g1.12xlarge	48	186.0	T4*2	32
ecs.gn6v-c8g1.8xlarge	32	128.0	V100*4	64
ecs.gn6e-c12g1.12xlarge	48	368.0	V100*4	128

如果您使用EAIS，则仅需要购买如下计算资源：

产品	实例规格	指标数据
云服务器ECS	ecs.r6.6xlarge	24vCPU 192GB
弹性计算加速实例EAIS	eais.ei-a6.4xlarge	FP32 16TFlops 32GB

综上所述，如果您购买GPU实例，则只能在已有的固定规格中进行选择，并且您需要为该实例的全部资源付费。而使用EAIS，您只需选择刚好满足您业务场景的实例规格，更灵活地解决相同的问题场景，且成本降低优势明显。

弹性

EAIS可以准确获取您所需的资源，为您灵活匹配GPU资源。您可以轻松扩展和缩减推理加速量级以满足您的业务需求，不会过度投资预置资源。当ECS增加实例以满足不断增长的需求时，您可以为每个ECS实例扩展EAIS实例。当需求降低时，您也可以随时释放任意ECS实例连接的EAIS实例。这有助您为所需资源灵活付费。

多适配性

EAIS具有极强的适配性，能够支持GPU、NPU、FPGA多种异构硬件的适配，种类多样，适配性强。

3.应用场景

弹性加速计算实例EAIS能够支持您所有的推理场景。本章节为您介绍几种主要的推理场景供您参考。

目标检测

检测出图片中多个目标的坐标位置，并给出目标的分类标签。目标检测可应用在：

- 视频检测：对视频中每一帧目标的精准定位。如车载地图对过往车辆、行人、车道线、红绿灯位置等周边环境进行智能检测，为驾驶员提供跟车距离预警、压线预警、红绿灯监测与提醒、提前变道提醒等驾驶安全辅助。
- 图像识别：将图像检测目标剪裁后配合图像识别提升识别精度。
- 目标定位：对海量图片进行分类、打标签。

图片分类

通过识别图片信息实现分类管理，得出正确结果。EAIS在推理场景中支持更灵活的配置和丰富的网络访问。

自然语言处理

支持对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作和加工功能。能够提供：

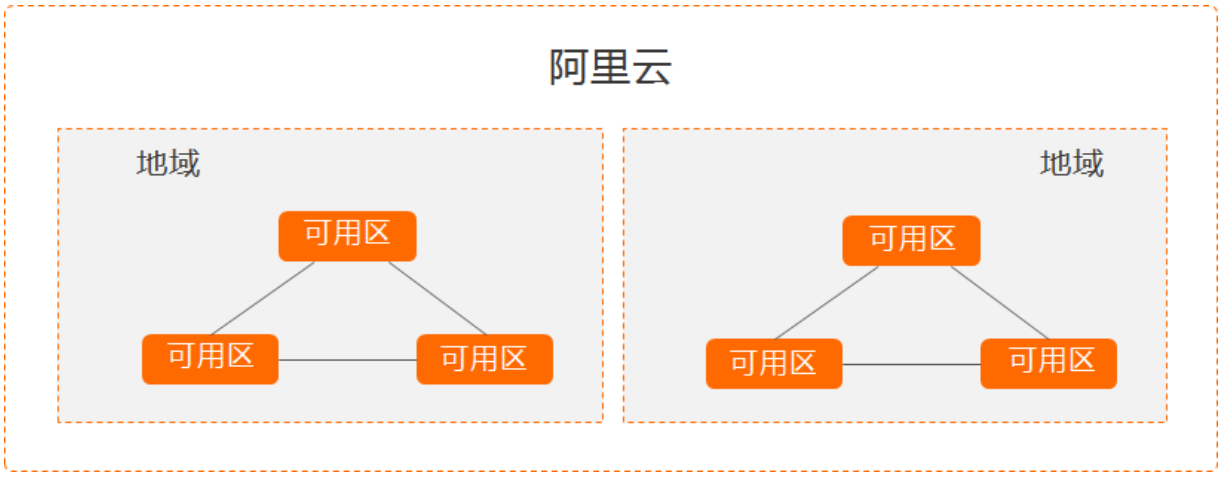
- 内容推荐：通过关键词提取、短文本相似度等技术，提取关键语义信息，精准匹配出语义相似的内容，快速构建推荐场景。
- 翻译：通过文本语言分析，精准翻译语句，帮助用户跨语言沟通。
- 舆情分析：挖掘突发事件、舆论导向，进行话题发现、趋势发现、舆情分析，多维度分析情绪、热点、趋势，及时全面的掌握舆情动态。
- 智能问答系统：通过中文分词、短文本相似度、命名实体识别等相关技术解决问答、对话、语料挖掘、知识库构建等问题。

4.地域和可用区

本章节为您提供弹性加速计算实例EAIS目前支持的阿里云地域和可用区。

每个地域完全独立。每个可用区完全隔离，但同一个地域内的可用区之间使用低时延链路相连。

地域和可用区之间的关系如下图所示。



地域

地域是指物理的数据中心。资源创建成功后不能更换地域。目前EAIS支持的地域、地域所在城市和Region ID的对照关系如下表所示。

地域名称	所在城市	Region ID	可用区数量
华北 2	北京	cn-beijing	8
华东 1	杭州	cn-hangzhou	8
华东 2	上海	cn-shanghai	7
华南 1	深圳	cn-shenzhen	5
西南 1	成都	cn-chengdu	2

选择地域时，您需要考虑以下几个因素：

- 地理位置

  一般情况下建议选择和您目标用户所在地域最为接近的数据中心，可以进一步提升用户访问速度。不过，在基础设施、BGP网络品质、服务质量、云服务器操作使用与配置等方面，阿里云中国内地地域没有太大区别。BGP网络可以保证中国内地全部地域的快速访问。

- 阿里云产品之间的关系

如果多个阿里云产品一起搭配使用，需要注意：

- 不同地域的云服务器ECS、关系型数据库RDS、对象存储服务OSS内网不互通。
- 不同地域之间的云服务器ECS不能跨地域部署负载均衡，即在不同的地域购买的ECS实例不支持跨地域部署在同一负载均衡实例下。

- 资源的价格

  不同地域的资源价格可能有差异，请参见弹性加速计算实例EAIS产品详情页。

可用区

可用区（Availability Zone，简称AZ）是指在同一地域内，电力和网络互相独立的物理区域。同一可用区内实例之间的网络延时更小。

在同一地域内可用区与可用区之间内网互通，可用区之间能做到故障隔离。是否将实例放在同一可用区内，主要取决于对容灾能力和网络延时的要求。

- 如果您的应用需要较高的容灾能力，建议您将实例部署在同一地域的不同可用区内。
- 如果您的应用要求实例之间的网络延时较低，建议您将实例创建在同一可用区内。

5.使用限制

本文问您介绍弹性加速计算实例EAIS在产品功能和使用上的不同限制，帮助您更好的应用EAIS。

限制项	限制说明
用户限制	创建EAIS实例的用户必须完成实名认证。
资源限制	创建按量付费资源的限制，账户余额、代金券和信用度之和不得小于100元。
推理框架	目前仅支持基于TensorFlow框架的推理业务，其他框架暂不支持。
付费方式	目前仅支持按量付费的方式，其他购买方式暂不支持。
售卖地域	目前仅只支持国内5个地域：华北2（北京）、华东1（杭州）、华东2（上海）、华南1（深圳）以及西南1（成都）。暂不支持国际业务售卖。
ECS规格	ECS实例的vCPU最小为1vCPU，内存最小为2GiB。
绑定关系	EAIS实例仅支持绑定一台ECS实例，且ECS实例仅支持绑定一个EAIS实例。
地域限制	具有关联性的ECS实例与EAIS实例必须在同一个地域。

6.实例规格

本章节为您介绍弹性加速计算实例EAIS的具体实例规格。

EAIS目前提供以下两款实例规格。

实例规格	FP32（TFLOPS）	FP16（TFLOPS）	显存（GB）
eais.ei-a6.4xlarge	16	128	32
eais.ei-a6.2xlarge	8	64	16

> **说明**
>
> EAIS目前不支持升降配操作。

7.实例生命周期介绍

本文介绍EAIS实例的生命周期，即从实例创建（购买）开始到释放结束的可能状态。

实例状态

在一个生命周期中，实例有其固有的几个状态，如下表所示。

控制台状态	API状态	状态属性	状态解释	是否计费	控制台上可见
准备中	Pending	中间状态	在控制台或通过API接口执行创建实例操作后，在进入**创建中**之前的中间状态。	否	否
创建中	Starting	中间状态	在控制台或通过API接口执行创建实例操作后，进入**可绑定**之前的中间状态。 如果长时间处于该状态，说明出现异常。	否	是
可绑定	Available	稳定状态	实例完成创建或解绑操作后，将进入**可绑定**的稳定状态。	是	是
绑定中	Attaching	中间状态	在控制台或通过API接口执行绑定操作后，实例进入**已绑定**之前的中间状态。 如果长时间处于该状态，说明出现异常。	是	是
已绑定	InUse	稳定状态	实例处于正常运行状态。 **说明** 当且仅当EAIS实例处于该状态时，才能提供加速服务。	是	是
解绑中	Detaching	中间状态	在控制台或通过API接口执行解绑操作后，实例进入**可绑定**之前的中间状态。 如果长时间处于该状态，说明出现异常。	是	是

API状态的转换流程

您可以调用DescribeEais查看实例状态。具体的API状态转换如下图所示。

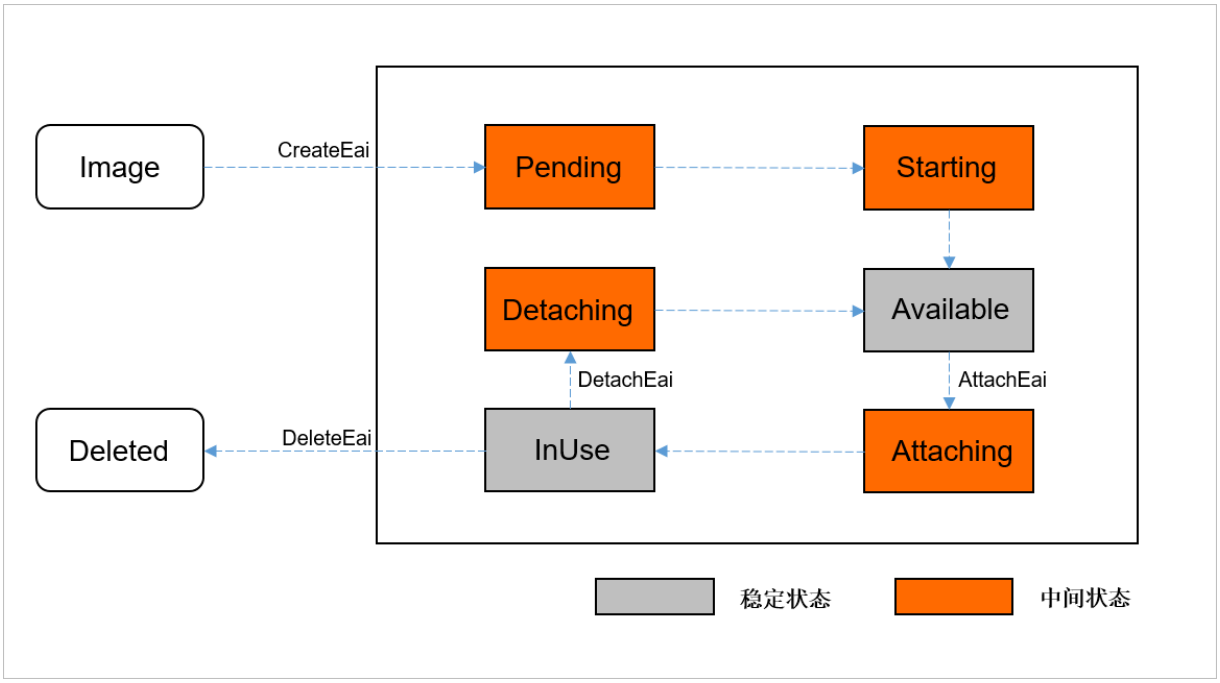


Image
CreateEai
Pending
Starting
Detaching
Available
DetachEai
AttachEai
Deleted
DeleteEai
InUse
Attaching
稳定状态
中间状态