

高性能应用服务 HAI

产品简介



腾讯云

【 版权声明 】

©2013–2025 腾讯云版权所有

本文档（含所有文字、数据、图片等内容）完整的著作权归腾讯云计算（北京）有限责任公司单独所有，未经腾讯云事先明确书面许可，任何主体不得以任何形式复制、修改、使用、抄袭、传播本文档全部或部分内容。前述行为构成对腾讯云著作权的侵犯，腾讯云将依法采取措施追究法律责任。

【 商标声明 】



及其它腾讯云服务相关的商标均为腾讯云计算（北京）有限责任公司及其关联公司所有。本文档涉及的第三方主体的商标，依法由权利人所有。未经腾讯云及有关权利人书面许可，任何主体不得以任何方式对前述商标进行使用、复制、修改、传播、抄录等行为，否则将构成对腾讯云及有关权利人商标权的侵犯，腾讯云将依法采取措施追究法律责任。

【 服务声明 】

本文档意在向您介绍腾讯云全部或部分产品、服务的当时的相关概况，部分产品、服务的内容可能不时有所调整。您所购买的腾讯云产品、服务的种类、服务标准等应由您与腾讯云之间的商业合同约定，除非双方另有约定，否则，腾讯云对本文档内容不做任何明示或默示的承诺或保证。

【 联系我们 】

我们致力于为您提供个性化的售前购买咨询服务，及相应的技术售后服务，任何问题请联系 4009100100或 95716。

文档目录

产品简介

产品概述

产品优势

产品对比

应用场景

套餐类型

产品简介

产品概述

最近更新时间：2024-08-09 17:35:11

快速了解高性能应用服务 HAI

高性能应用服务（Hyper Application Inventor，HAI）是一款面向 AI 和科学计算的 GPU/NPU 应用服务产品，提供即插即用的强大算力和常见环境。它可以帮助中小企业和开发者快速部署语言模型（LLM）、AI 绘图、数据科学等高性能应用，原生集成配套的开发工具和组件，大大提升应用层的开发生产效率。

与 GPU 云服务器的区别

若需了解高性能应用服务 HAI 与 GPU 云服务器的区别，详情可参见 [产品对比](#)。

如何使用高性能应用服务 HAI

腾讯云为您提供如下方式进行高性能应用服务 HAI 的配置和管理：

- **控制台**：高性能应用服务 HAI 可使用 [控制台](#)，对算力资源进行管理。
- **API**：腾讯云也提供了 API 接口方便您管理高性能应用服务 HAI。关于 API 说明，请参见 [API 概览](#)。

ⓘ 说明

如果您未使用过高性能应用服务 HAI，可参见 [通过高性能应用服务 HAI 一键创建应用](#) 开始使用。

计费概述

有关高性能应用服务 HAI 的计费相关说明，详情请参见 [计费概述](#)。

产品优势

最近更新时间：2025-02-18 17:52:12

简单易用

通过简化计算、网络和存储等基础设施的配置流程，大幅降低了云服务操作和管理的复杂度。

应用环境快速部署

支持多种 AI 环境快速部署，如 ChatGLM-6B、StableDiffusion 等，使用户可专注业务及应用场景创新。

高灵活性

支持用户登录实例，对 AI 模型及实例环境进行灵活配置。可进行内部开发、业务测试，或对外提供业务服务。

多种登录方式

除传统连接方式外，支持通过 jupyterlab、WebUI 等方式一键启动，提供更贴合使用场景的登录方式。

算力种类丰富

提供多种算力套餐选择，未来还会将加入更多种类供用户选择。目前在售算力套餐，详细配置请参见 [套餐类型](#)。

产品对比

最近更新时间：2024-12-19 10:19:53

高性能应用服务 HAI 相比传统 GPU 云服务器的主要区别和优势请参考下表：

功能类别	GPU 云服务器	高性能应用服务 HAI
交付形态	基础的虚拟机	即插即用的应用
机型选择	需要了解 GPU 型号，自行选择合适机型，有不匹配风险	基于 AI 应用，自动匹配合适套餐
环境部署	需要自行部署驱动、CUDA、Python、Notebook 等环境依赖	分钟级快速启动，直接交付可用应用环境
资源配置	需要额外购买合适的云硬盘、带宽或流量	打包 GPU、云硬盘、带宽及网络，一键启动
产品入口	需要具备一定的运维知识，登录命令行界面进行操作	提供 WebUI 等可视化连接方式，一键进入服务，可视化配置
模型甄别	各类模型版本繁多，难以自行挑选	预置最新版本的主流模型，适配套餐机型
资源下载	部分访问可能遇到网络拥塞问题	跨境线路自动择优，支持学术资源平台访问、下载加速

应用场景

最近更新时间：2024-08-09 17:35:11

AI 作画/设计

设计师和开发者可以使用高性能应用服务 HAI 快速地部署和优化 AI 绘画模型。高性能应用服务 HAI 预置 Stable Diffusion 等主流 AI 作画模型及常用插件，提供 GUI 图形化界面即开即用，大幅降低上手门槛。

AI 对话/写作

研究者和开发者可以使用高性能应用服务 HAI 快速部署和运行大型语言模型，如 LLAMA2、ChatGLM 等，进行自然语言处理任务，如文本生成、情感分析、文本分类等。高性能应用服务 HAI 提供的算力支持和优化环境确保了语言模型可以在最短的时间内进行部署，同时还能保持高稳定性和可靠性。

AI 开发测试

高性能应用服务 HAI 的预配置环境支持大多数流行的 AI 框架和工具，如 TensorFlow、PyTorch 等，使得开发者可以专注于算法设计和模型优化。AI 研究者可以在高性能应用服务 HAI 上进行模型的开发、训练、测试和优化，无需担心硬件兼容性和软件配置问题。如新算法的原型开发、模型微调与迁移学习、深度学习框架的交叉测试等。

数据科学

数据科学家们可使用高性能应用服务 HAI，快速进行数据分析和图标处理。高性能应用服务 HAI 预置了 Notebook、Python 环境，以及主流分析软件。

套餐类型

最近更新时间：2025-03-05 14:56:42

高性能应用服务 HAI 算力方案提供丰富的算力方案套餐选择，致力于提供更易用的 GPU 智算服务。创建高性能应用服务 HAI 时，选择的套餐决定了算力、显存、CPU 和内存配置。这些配置参照套餐配置标准，由符合要求的 CPU、内存和 GPU 硬件构成，不支持指定硬件型号。

说明：

邀测套餐处于白名单邀测状态，请联系您的售前经理开通实例购买权限。

售卖情况	算力方案	每实例显存	每实例算力	每实例其他配置	性能参考（以 INT4 量化为基准）
在售	GPU 基础型	16GB +	8+TFlops SP(FP32)	CPU: 8 核 内存: 32GB	对标 T4 GPU 及以上，推荐运行 8B 及以下参数模型
	GPU 性能型	24GB +	30+TFlops SP(FP32)	CPU: 28 核 内存: 116GB	对标 A10 GPU 及以上，推荐运行 32B 及以下参数模型
	GPU 进阶型	32GB +	15+TFlops SP(FP32)	CPU: 8~10 核 内存: 40GB	对标 V100 GPU 及以上，推荐运行 32B 及以下参数模型
	CPU 算力	-	-	CPU: 8/16/32/64 核 内存: 16/32/64/256GB	<ul style="list-style-type: none">8 核 16GB: 推荐运行 1.5B 及以下参数模型，适合进行轻量任务测试体验16 核 32GB: 推荐运行 8B 及以下参数模型，可进行中等难度任务，如文本生成等32 核 64GB: 推荐运行 14B 及以下参数模型，可进行文本摘要等任务64 核 256GB: 推荐运行 32B 及以下参数

					模型，可进行复杂任务，如知识库问答等
邀 测	GPU 高阶型	48GB +	59+TFlops SP(FP32)	CPU: 48 核 内存: 192GB	对标 L20 GPU 及以上， 推荐运行 70B 及以下参数 模型
	GPU 旗舰型	96GB +	44+TFlops SP(FP32)	CPU: 16 核 内存: 160GB	推荐运行 70B 及以下模型
	八卡 GPU 旗 舰型	768G B+	352+TFlops SP(FP32)	CPU: 384 核 内存: 2304GB	推荐运行 671B 及以下模 型
	紫霄基础型	16GB +	30+TFlops FP16	CPU: 28 核 内存: 116GB	对标 T4 GPU 及以上