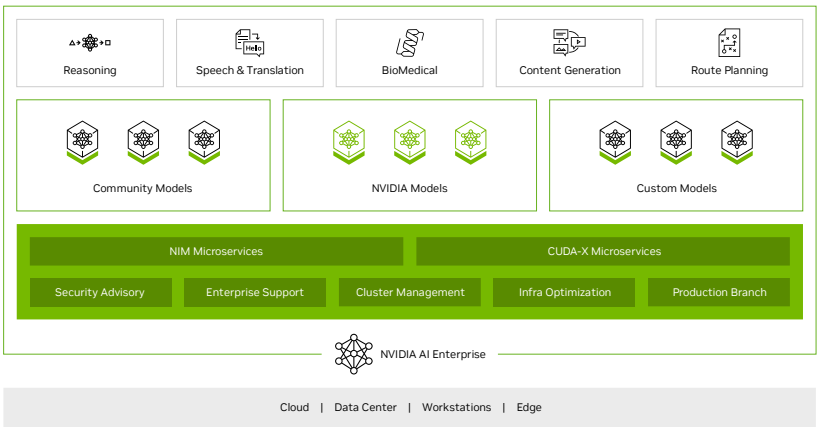
解决方案概要

NVIDIA AI Enterprise

使用企业级的端到端云原生软件平台构建生产级的 AI。

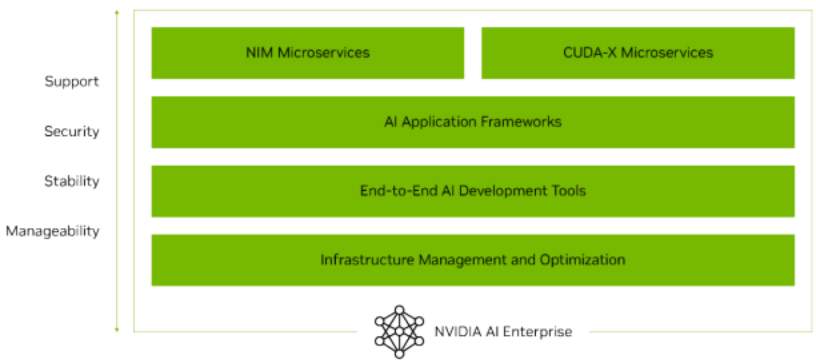
NVIDIA AI Enterprise 是一个端到端云原生软件平台，可加速数据科学工作流，简化生产级协作驾驶和其他生成式 AI 应用的开发和部署。易于使用的微服务优化了模型性能，可提供企业级的安全性、支持服务和稳定性，能够确保以 AI 为基础开展业务的企业从原型到生产的平稳过渡。



适用于生产级 AI 的端到端软件平台

NVIDIA AI Enterprise 不仅提供适合 AI 从业者的出色开发工具、框架和预训练模型，而且能可靠地满足 IT 专业人员在管理和编排方面的要求。这意味着，您可以在性能、高可用性和安全性上获得全面保障。利用全栈 NVIDIA AI 软件加速您的 AI 工作流：

- 使用面向 Apache Spark 的 NVIDIA RAPIDS™ 加速数据处理。
- 使用 NVIDIA NeMo 开发自定义生成式 AI，这是一个端到端平台，可提供具有精确数据管护、先进定制、RAG 和加速性能的企业就绪型模型。
- 使用基于 NVIDIA TensorRT™、TensorRT-LLM 和 NVIDIA Triton™ 推理服务器构建的 NIM 完成大规模推理。
- 使用 NVIDIA Base Command™ Manager Essentials 在边缘和数据中心进行大规模 AI 集群管理。



特性和优势

性能优化

NVIDIA NIM 和 CUDA-X 微服务提供了优化的 runtime 并可轻松使用基础模组，简化了生成式 AI 的开发。

放心部署

通过持续监控安全漏洞和模型定制所有权来保护公司数据和知识产权。

运行位置不受限制

基于标准的容器化微服务已经过认证，可以运行于云端、数据中心和工作站之上。

适合企业级需求

凭借可预测的 API 稳定性生产分支、管理软件和 NVIDIA Enterprise Support，帮助确保项目保持平稳进行。

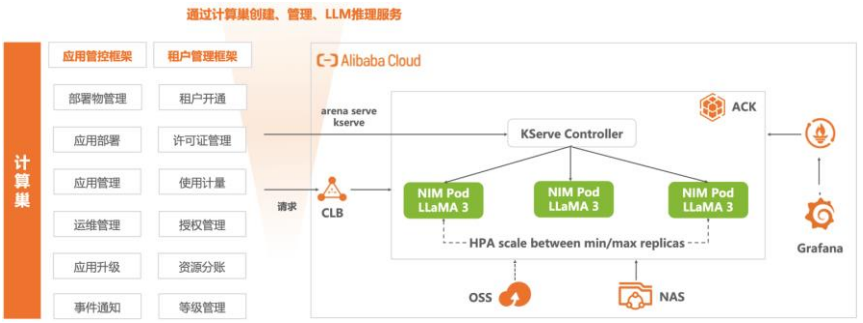
通过阿里云云市场获取 NVIDIA AI Enterprise License

阿里云云市场提供的 NVIDIA AI Enterprise license，分为概念验证 （PoC）测试服务和购买两个下单页面。

> PoC 测试服务提供 90 天 NVIDIA AI Enterprise 试用，使用时须支付阿里云计算资源的费用。
> 购买 NVIDIA AI Enterprise License 提供技术、售前咨询，页面价格仅供展示。通过线下沟通采购和线上推送 Private Offer，经由阿里云云市场交付 NVIDIA AI Enterprise License。

通过阿里云计算巢快速部署 NVIDIA NIM 的架构图

在阿里云容器服务 ACK （容器服务 Kubernetes 版）集群上，使用云原生 AI 套件，集成开源推理服务框架 Kserve 来部署 NVIDIA NIM。
同时，结合阿里云的 Prometheus 和 Grafana 服务，快速搭建监控大盘，实时观测推理服务状态；利用 NVIDIA NIM 提供丰富的监控指标，如 num_requests_waiting，配置推理服务弹性扩缩容策略。
此处列出的云上资源，以及阿里云弹性计算服务 ECS、专有网络 VPC 等基础资源，都可以通过计算巢来轻松配置，一键拉起，最终形成一个云上的 LLM 推理服务。用户只需要根据该服务创建实例，便可部署云上弹性的 LLM 推理服务。



阿里云计算巢：专为服务商打造的云集成 PaaS 平台

阿里云计算巢服务是一个开放给服务商（包括：企业应用服务商、IT 集成服务商、交付服务商和管理服务提供商等）和用户的服务管理 PaaS 平台，提供上云的“一站式”解决方案。
通过阿里云计算巢服务，能够实现软件的交付、部署、运维流程标准化，支持软件和资源的一体化交付，真正实现了软件的开箱即用。

NVIDIA NIM：易于使用的预构建容器工具

NVIDIA NIM ™ 是 NVIDIA AI Enterprise 的一部分，是一套易于使用的预构建容器工具，目的是帮助企业客户在云，数据中心和工作站上安全，可靠的部署高性能的 AI 模型推理。

开始上手

> 在 PC 网页浏览器免费试用 NVIDIA NIM，访问: ai.nvidia.com
> 点击此处，在阿里云云市场试用 90 天 NVIDIA AI Enterprise
> 点击此处，在阿里云云市场购买 NVIDIA AI Enterprise
> 点击此处，了解关于 NVIDIA AI Enterprise 的企业级支持服务

以 Llama 3.1 作为技术演示，阿里云计算巢服务快速部署 NVIDIA NIM 的流程

> 参考 NVIDIA NIM 文档，生成 NVIDIA NGC API Key，用于访问需要部署的模型镜像。以 Llama-3-8B-Instruct为 例，通过 NVIDIA NGC，或者 API Catalog 获取。同时，请阅读并承诺遵守 Llama 模型的自定义可商用开源协议。

> 在计算巢服务目录中找到“基于 NVIDIA NIM 快速部署 LLM 模型推理服务”，并进入实例部署页面。 主要配置服务的基本信息和云上资源，并填入第一步中获取的 NVIDIA NGC API Key。因服务部署在阿里云 ACK 集群上，这步也需要配置 Kubernetes。

> 按照页面提示完成所有配置，点击下一步“确认订单”，获得服务实例信息和价格预览。

> 部署中需要创建和访问阿里云资源，当阿里云账号属于 RAM 账号时，需开通以下权限，页面有权限开通入口。

权限策略名称	备注
AliyunECSFullAccess	管理云服务器服务（ECS）的权限
AliyunBSSReadOnlyAccess	只读访问费用中心（BSS）的权限
AliyunCSFullAccess	管理容器服务（CS）的权限
AliyunVPCFullAccess	管理专有网络（VPC）的权限
AliyunROSFullAccess	管理资源编排服务（ROS）的权限
AliyunComputeNestUserFullAccess	管理计算巢服务（ComputeNest）的用户侧权限
AliyunECSFullAccess	管理云服务器服务（ECS）的权限

> 点击立即创建。部署过程中会涉及阿里云资源的创建、NIM 模型镜像拉取等。拉取过程的日志可通过“点击资源 tab -> 找到 ACK 集群 ->页面左侧的工作负载 ->无状态”查看。

> 部署完成后，进入服务实例详情查看使用说明。通过 curl 发送 HTTP 请求访问推理服务，修改 content 字段，便可自定义和推理服务交互的内容。