INFINIGENCE
无问芯穹

Infinigence AI

释放无穹算力
让AGI触手可及

01.

公司介绍

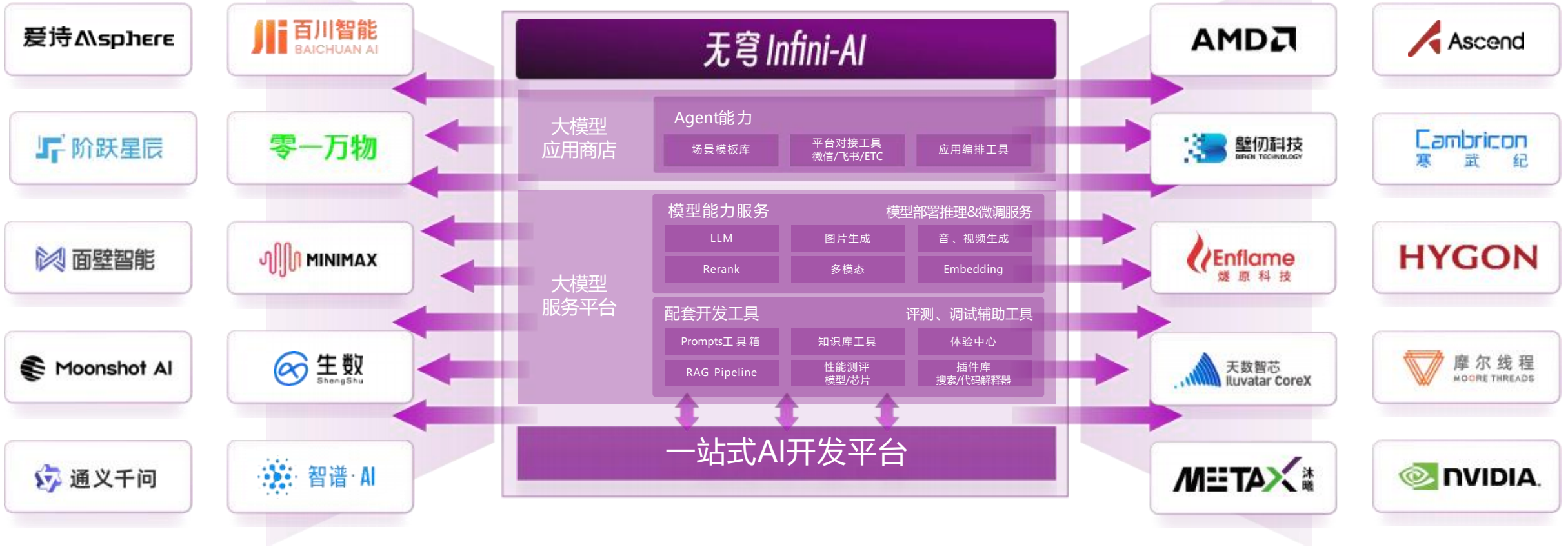
关于无问芯穹

无问芯穹（Infinigence AI）依托行业领先且经过验证的AI计算优化能力与算力解决方案，追求大模型落地的极致能效。打造“M 种模型”和“N 种芯片”间的“M×N”中间层产品，实现多种大模型算法在多元芯片上的高效、统一部署。链接上下游，共建AGI时代大模型基础设施，加速AGI落地千行百业。



创始团队	清华大学电子工程系 Department of Electronic Engineering, Tsinghua University 推动成立
核心成员	夏立雪、颜深根、戴国浩 Alibaba Cloud大模型压缩加速 生成式AI模型芯片、上海AI超算原型机 国家自然科学基金等多个项目负责人
研发团队 *35%以上来自清华大学	Apache、ONNX、TensorFlow、PyTorch、PyG 等知名项目重要贡献者

无穹Infini-AI：面向大模型应用开发者的企业级服务平台



与2家以上闭源模型合作
兼容100+个模型

完整实现模型到芯片的
M×N自动路由

支持
10余家国产芯片

整体产品架构图

MaaS：灵活易用大模型服务

敏捷开发企业级大规模应用

- ✓ 一键式模型训练与部署；
- ✓ 丰富开源模型生态支持；
- ✓ 极致软硬协同算力优化；

PaaS：云端协同算力底座

让大模型开发&研究者拎包入住

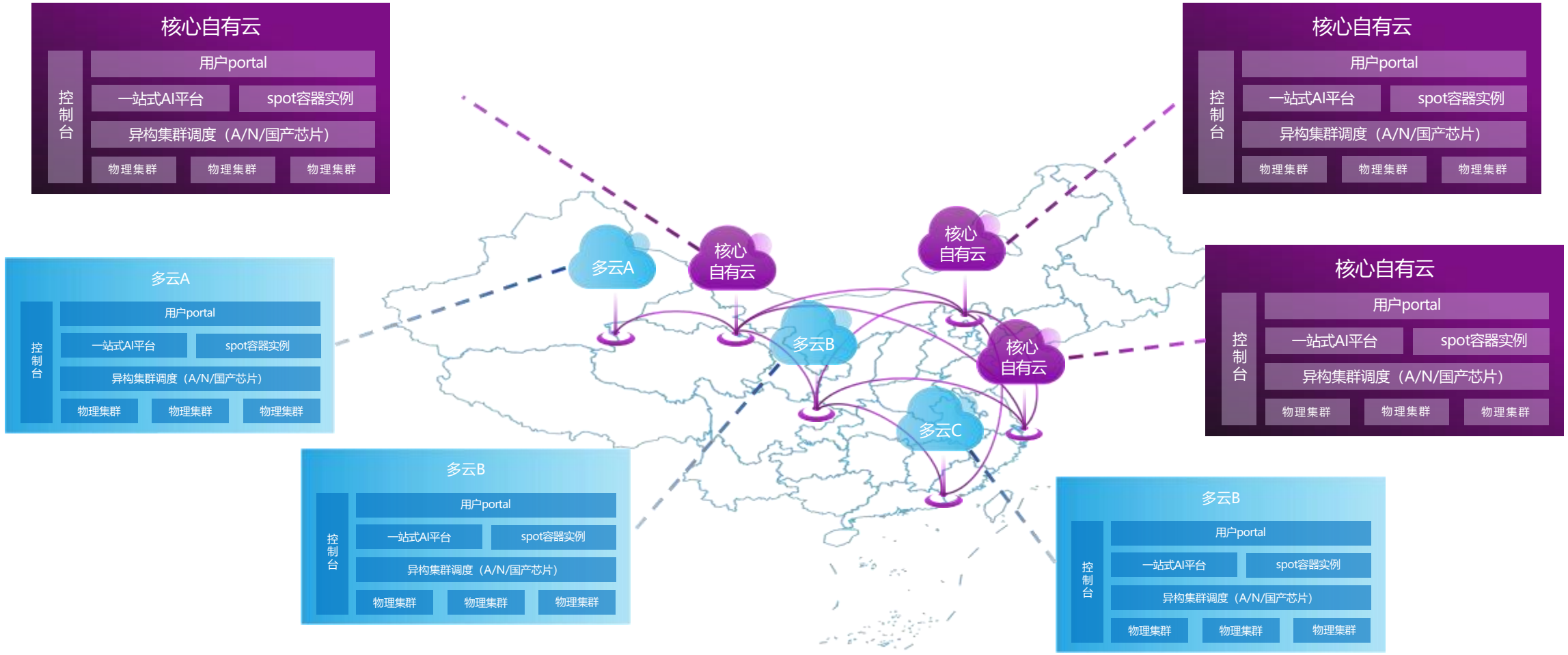
- ✓ AI算力集群智能运维；
- ✓ 适配异构算力硬件；
- ✓ 云产品私有化部署一体机；

IaaS&引擎：高效底层能力支持

为用户释放无穷算力

- ✓ 业内领先的训推优化算法；
- ✓ 极致底层算力支撑；
- ✓ 端侧AI定制化场景；

跨地域万卡集群纳管能力



异构集群调度引擎

多集群调度

- 多租户多集群调度和资源监控
- 国产等多异构芯片统一管理
- 小时级资源池动态分配管理
- 大模型训推微分布式任务支持

20% 算力使用成本

92% 集群资源利用率

集群内弹性调度

- 反碎片化、动态租借抢占等10余调度策略
- 多存储及多算力网络纳管

30% 任务等待时间

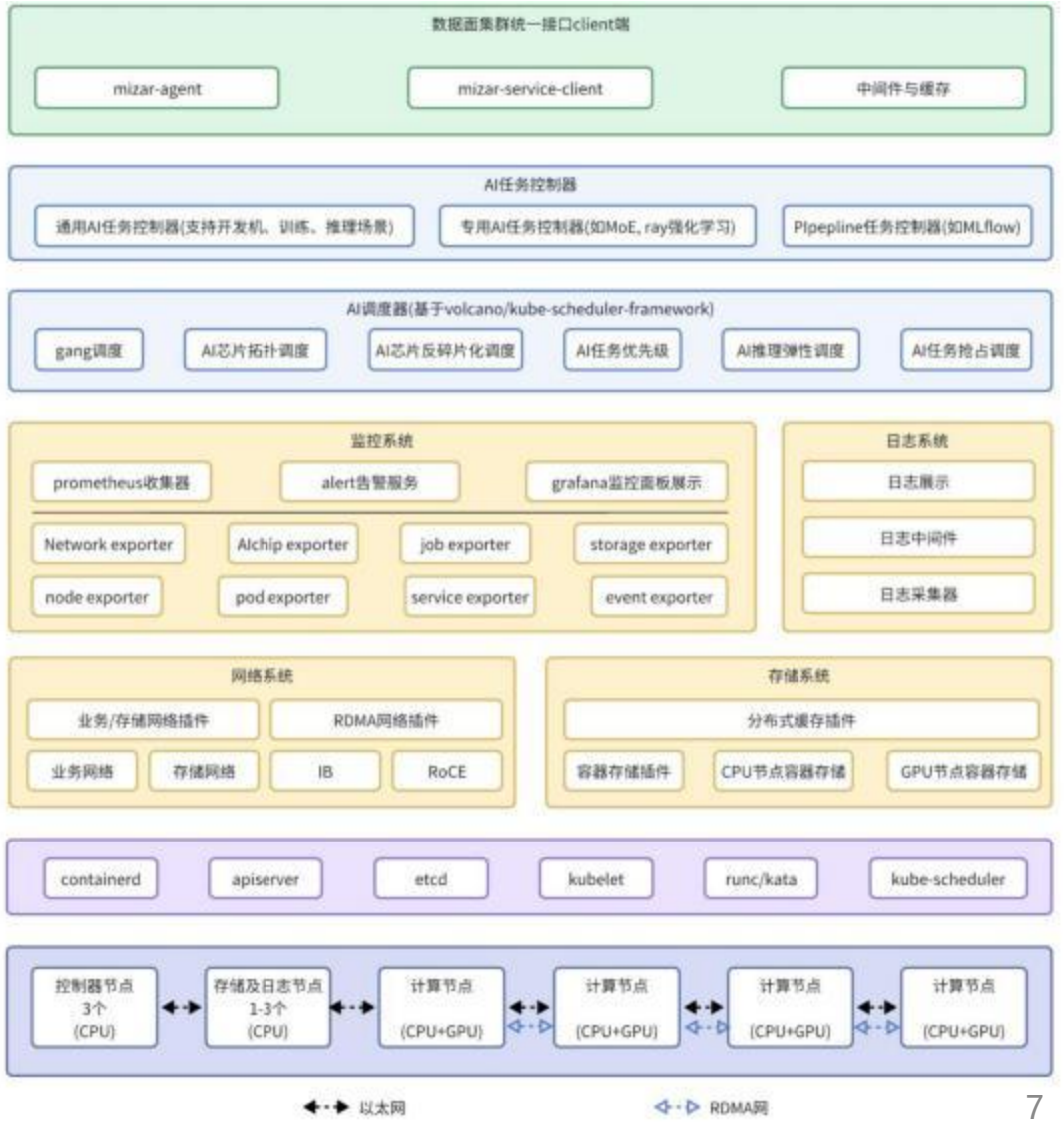
270 workers/s 调度吞吐量

AI容器基座增强

- 开发场景的rootfs秒级高性能持久化
- GPU运行时的资源隔离增强

5% 相比裸机，容器化性能损耗

99.99% 集群SLO



02.

产品介绍

无问芯穹一站式AI 平台

无问芯穹一站式AI 平台是面向机器学习开发者，提供开发机、任务等功能的企业级开发平台，支持从数据托管、代码开发、模型训练、模型部署的全生命周期工作流

最懂开发者的
共享开发机解决方案

最稳定的万卡训练
容错保障解决方案

最高效的多芯片
推理服务解决方案

产品优势

一站式

涵盖 AI 开发全流程，包含数据管理、模型开发、训练、推理

高性价比

提供高性价比的预付费计算资源，额外提供辅助功能帮助提升资源利用率

开发调试工具

平台内置多种主流机器学习框架的镜像，极大提升开发和调试环境一致性

分布式训练

预置多种分布式框架，可稳定、高效运行超大规模的分布式训练任务

高性能推理

部署多种框架的模型到异构硬件，提供高吞吐、低延时、实时扩缩容、容错等特性

一站式 AI 平台-优势

多样灵活的开发机配置

公网访问：支持公网通过 SSH 远程访问开发机

共享文件挂载：可挂载租户下的共享文件存储，采用多副本机制，确保数据安全可靠

Docker支持 ：在开发机内也可以使用 Docker 命令创建容器，推送镜像

资源共享：允许多个开发机共用GPU资源，确保算力高效利用

丰富的任务功能

分布式训练：预置分布式框架，支持 PyTorch-DDP 和 MPI

可视化日志：预置 TensorBoard，可对训练过程中保存的 TensorBoard 日志进行可视化

容错保障：保证长时间稳定运行，异常检测自动恢复

任务监控：对于任务及Worker进行多维度的资源指标监控

日志可视化：支持对用户在训练过程中保存的 TensorBoard 日志进行可视化

便捷的推理服务

推理服务：支持推理服务的创建、克隆、停止、启动、升级、删除、调用等功能

业务线上化：支持快速便捷地将训练好的模型部署成线上服务，接入实际业务场景

自动扩缩容：支持根据流量和负载情况对于实例进行动态扩缩容

滚动升级：支持在不影响线上业务的同时进行升级

服务监控：支持对于资源、业务及大模型场景的监控

全面的镜像中心

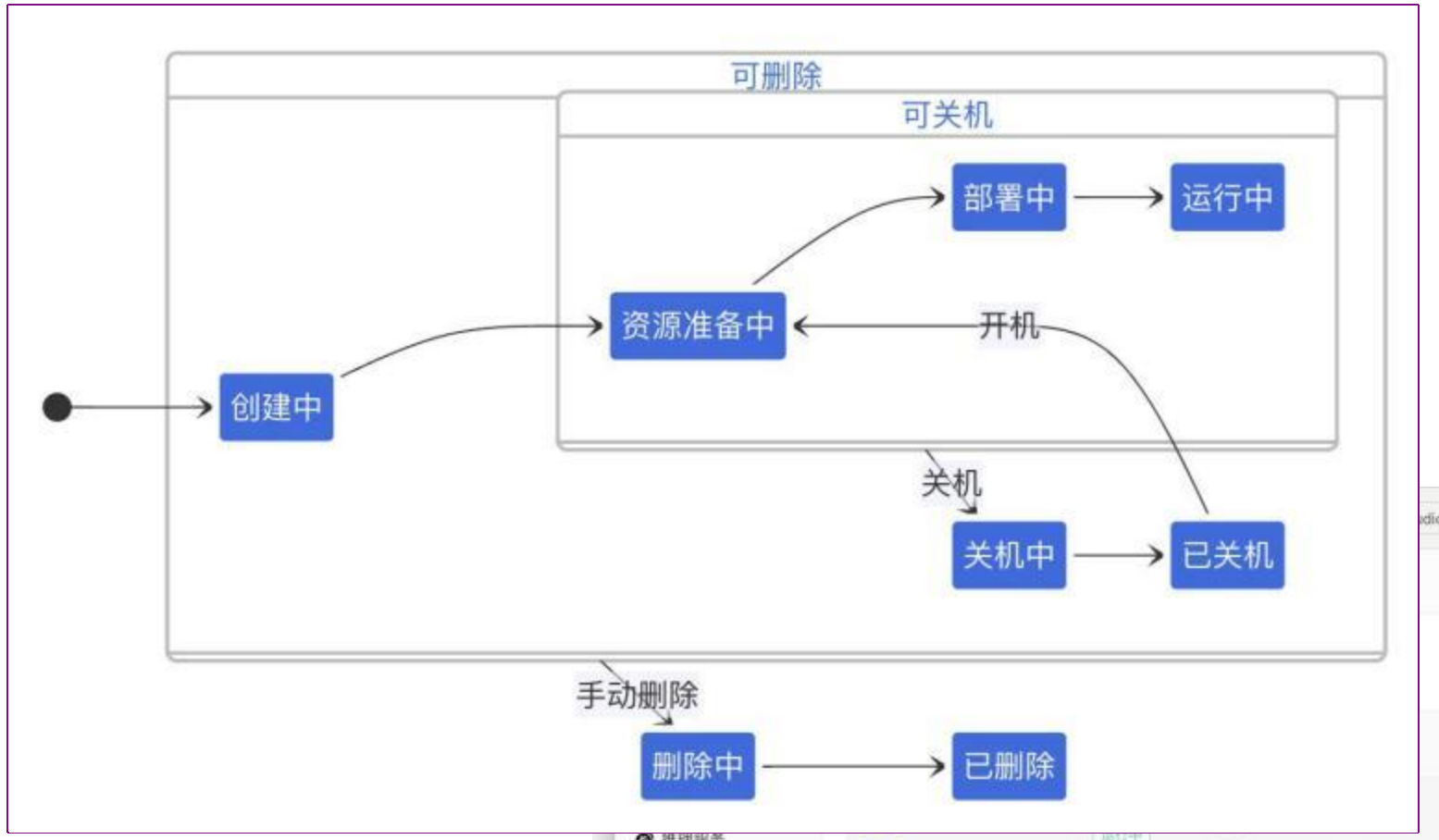
预置镜像：预置多种可直接使用基础镜像，可基于预置镜像构建自定义镜像

镜像仓库：创建租户后，平台自动为租户生成租户的镜像仓库，可供用户使用

快速构建镜像：通过自定义镜像标签，可快速构建自有镜像

多可用区同步：镜像可实现多可用区的同步，实现业务快速开展

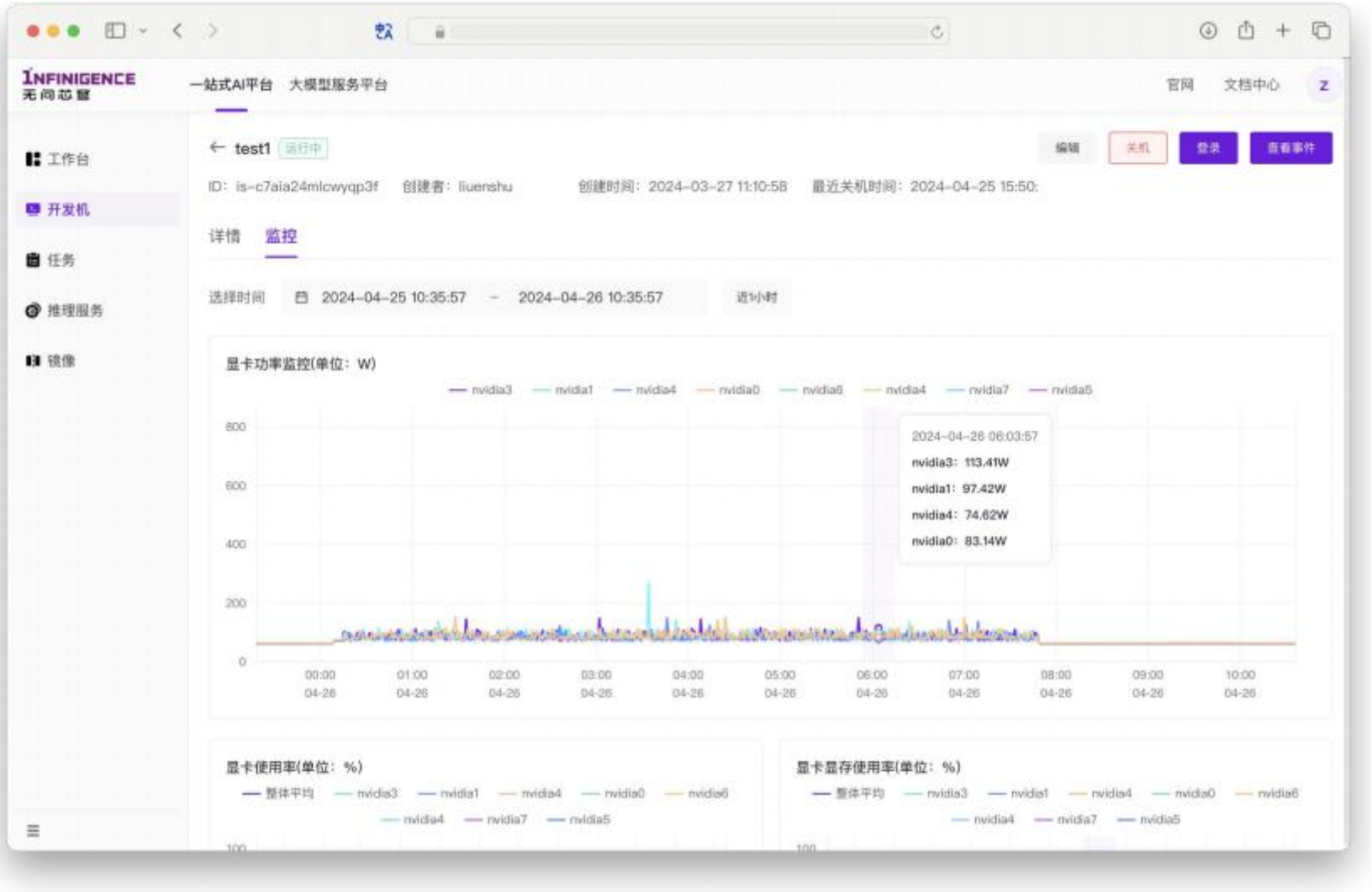
无问芯穹一站式AI 平台功能



开发机

开发机是 AI Studio 中用于在线编译、调试代码和模型开发的模块。用户只需要拥有一台电脑或者移动设备，并连接到互联网，就可以开始在线开发和调试代码。

无问芯穹一站式AI 平台功能



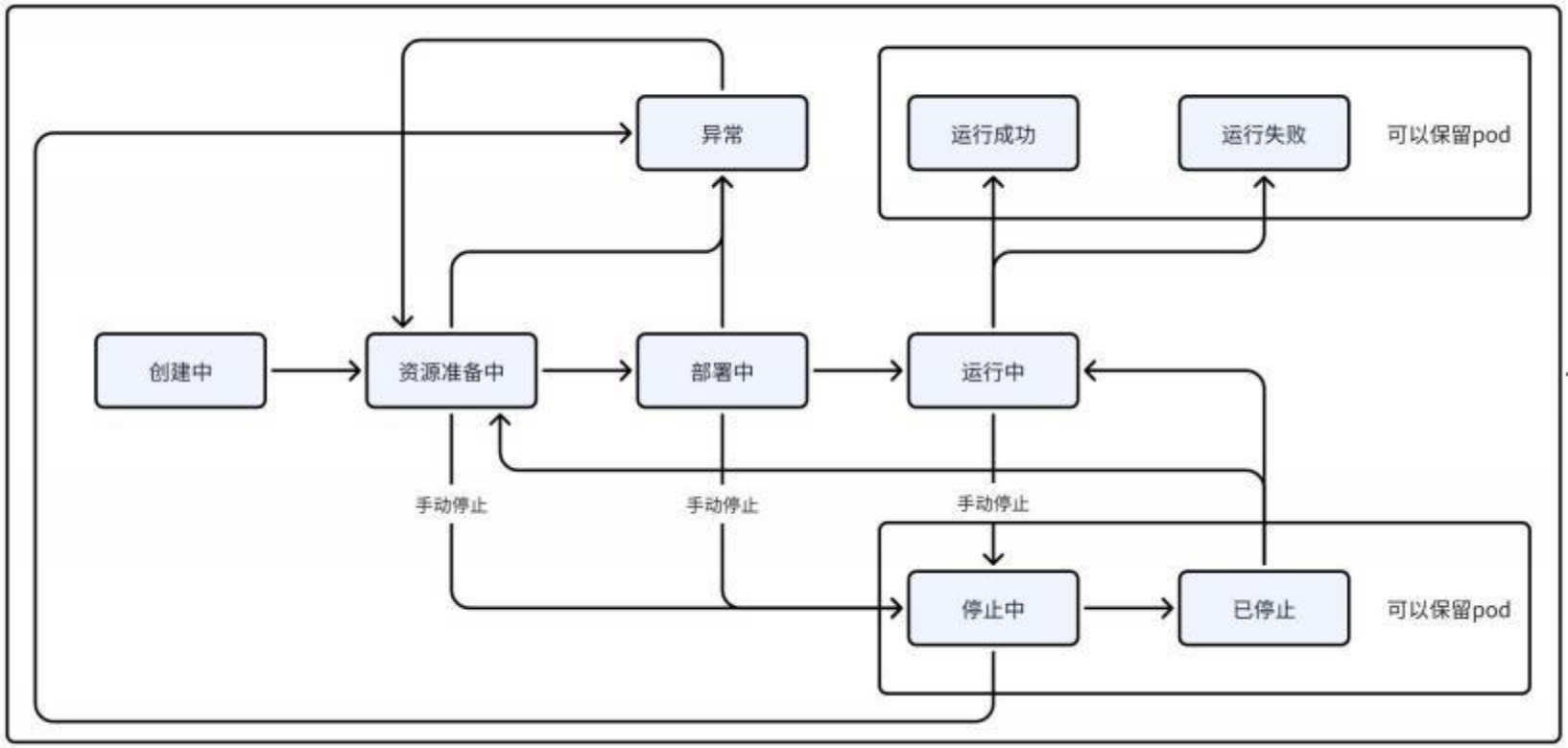
资源监控

AIStudio 平台支持监控开发机的显卡、CPU、内存资源使用情况。

事件监控

智算云平台会记录开发机在生命周期中的所有事件，可在开发机详情页面点击查看事件。

无问芯穹一站式AI 平台功能



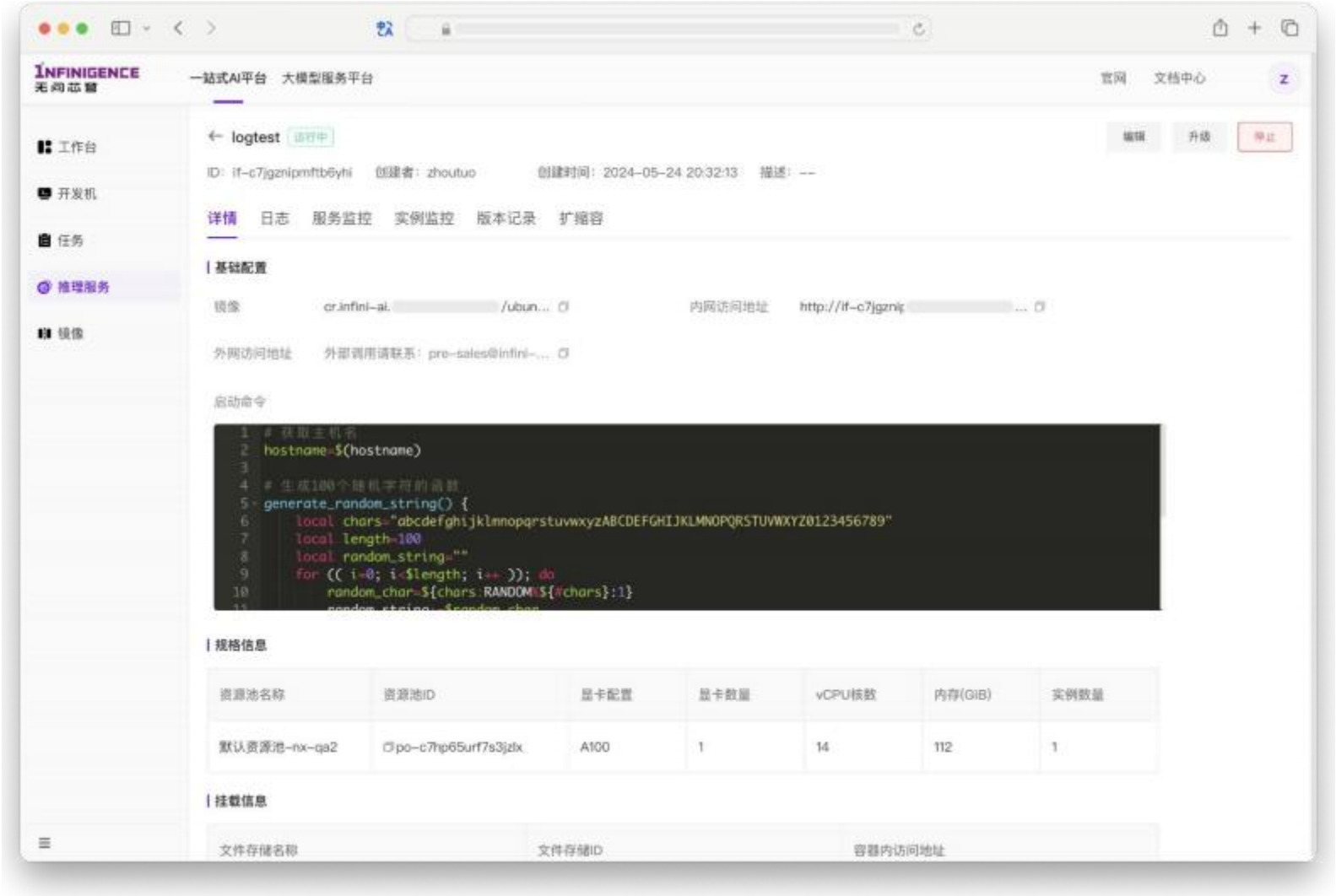
任务功能

平台已具备预置镜像，同时支持多种分布式框架，用户可便捷的发起分布式训练任务，极大地提高了工作效率。

任务容错

具备多种异常检测机制，支持千卡分布式训练及快速恢复；容错系统确保训练任务连续不间断，有效降低大模型企业任务训练成本。

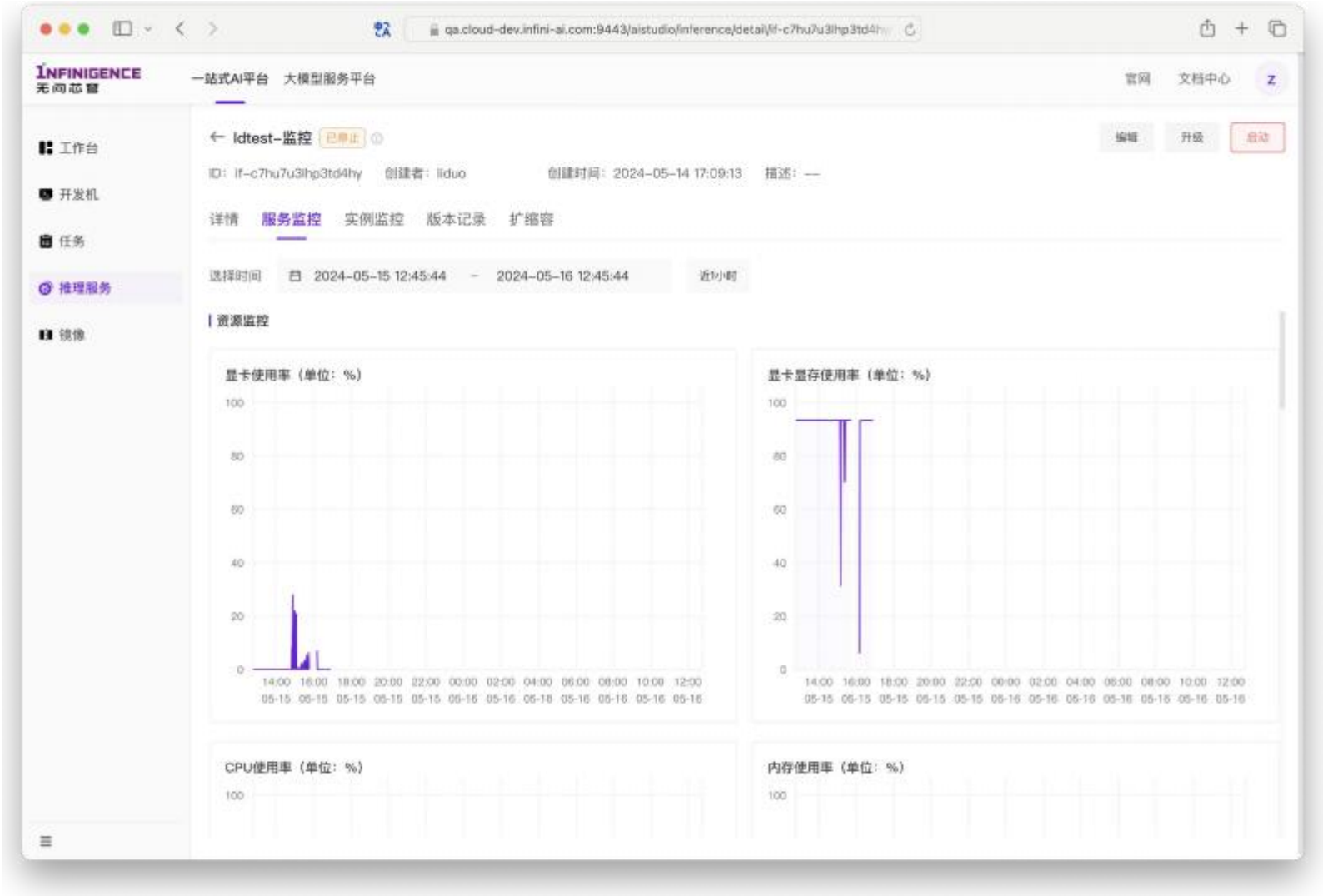
无问芯穹一站式AI 平台功能



推理服务

一站式 AI 平台（AI Studio）的推理服务，可快速便捷地将训练好的模型部署成线上服务，接入实际业务场景。

无问芯穹一站式AI 平台功能



推理服务

推理服务监控可提供服务整体与实例级别的资源使用情况。

资源监控：适用于所有推理服务，反映推理服务的显卡、显存、内存、CPU 的使用情况。

业务监控：适用于选择了「特定预置镜像」的推理服务，提供每秒请求数、流量等推理业务通用指标。

LLM 场景业务监控：如果使用「大模型专用镜像」，可提供 TTFT（生成首个 Token 的时间）、总 Token 数等 LLM 业务指标。

大模型服务平台

无问芯穹大模型服务平台基于无问芯穹的智算云平台，
针对生成式大模型的应用落地的多种场景需求，为应用开发者提供高性能、易上手、安全可靠的大模型服务，覆盖从大模型开发到大模型服务化部署的全流程。

最便捷的
模型微调使用平台

最客观的
模型芯片评测平台

最强大的
技术生态兼容平台

产品优势

模型种类齐全

预置多种来源、参数规模、不同类型的大模型，用户可根据需要快速选择

模型使用简单

预置丰富的各类大模型，用户注册后，无需部署可一键调用

模型微调便捷

大力降低微调门槛，更贴合落地需求，用户通过很低的成本和技术要求对预置大模型进行微调

异构芯片适配

内置无穹自研推理优化能力，快速适配M*N的最好的组合。针对主流及多种国产芯片在训练、推理场景适配和优化

模型对比客观

提供多种大模型评测对比工具，可根据需求快速选择合适自己需求的模型，并将模型在不同的硬件环境上进行效果使用比较

开发者生态兼容

Web UI 支持即来即用、同时支持简单易用的 API、SDK 便于快速集成

强大底层技术支撑

基于大模型计算优化引擎，提供训练和推理优化能力，基于调度引擎保障

优质agent灵活调用

提供较好的agent模版能，并能有效地与其他Agent协同工作

大模型服务平台功能

各类模型开箱即用

多模型的便捷使用：平台提供语言、视觉、多模态等主流模型的接口调用

精细化的模型筛选：用户可根据标签，快速筛选出符合自己需求的模型

先用后付的服务模式：用户可以先使用相关的服务，平台根据实际的用量（如token数）计费

客观的模型预先精选：平台提供通过筛选的开源顶级模型和闭源主流模型，确保平台模型好用

业务模型快速自定义

快速的模型定制：可快速创建符合用户需求的特定任务定制化模型

便捷的模型微调与部署：用户无需考虑复杂算法和硬件细节，可快速微调与部署微调后的模型

全面的国产与主流硬件适配：用户可根据需要自主选择多种主流和国产芯片

便捷的落地应用模式：用户可根据部署方式自主选择计费模式

M×N全场景评测选型

模型与芯片全场景适配：支持用户自主选择模型和芯片进行评测比较

模型的实际效果评测：支持固定芯片类型快速对比不同模型或同一模型不同大小的输出效果

芯片的实际能力评测：支持固定模型类型快速对比不同芯片的实际性能

评测数据集的自定义：用户可上传评测数据集，帮助用户自动评测模型的业务效果

Agent快速上手

预置多个调优验证的agent：平台已有多个经过调优和业务验证后的agent

提供全链条工具：平台提供prompts工具、知识增强等工具，用户可便捷自定义agent

智能辅助优化：平台提供提示词自动优化增强、自主业务流规划等智能辅助功能

灵活易用的多模型、多芯片直观交互

模型广场

- 预置丰富的模型
- 无需部署一键调用
- 按token/并发/资源多种计费模型
- 内置无穹推理优化能力

体验中心

- 多模型自由交互
- 多模型效果对比

云管平台——专属云软件部署·公共云模式运营

智算中心客户

开通、测试、购买、使用、续费、售后
全线上化流程，开箱即用

智算云统一portal

智算云管理平台

- 统一运维
- 统一运营
- 财务系统

智算中心管理者

智算中心硬件集群状态、客户运营情况、
财务收入账单一目了然

云管平台——功能

无问芯穹云管平台基于完整的运营管理体系，
为管理者提供账号管理、云审计体系、访问控制、后台管理等功能
涵盖用户管理、用户审计、用量监控等全流程。

最便捷的
算力管理平台

最易用的
用户管理平台

最丰富的
模型管理平台

产品优势

多云设计

- 每朵云包含一个或多个集群（集群内部共享存储，每朵云有自己独立的管理台和控制台
- 隔离各个云的控制台、各个云的管理台独立部署，域名独立
- 联通 infini-ai控制台可以售卖所有集群（集群客户同意）的资源

管理台账号和权限

- 管理台部署后自动生成一个admin账号
- admin账号可以创建管理台的子账号
- 登陆平台使用账号+密码的形式

用户使用管理

- 事件查询列表：操作审计事件产讯列表中按时间或其他字段查询你需要追踪的日志
- 事件查询详情： 日志的详细信息
- 事件查询导出：以excel形式导出操作日志

Thanks

释放无穷算力，让AGI触手可及

INFINIGENCE
无问芯穹

Follow Infinigence AI

Contact us： sales@infini-ai.com

上海市徐汇区龙台路180号（200232 ）

北京市海淀区东升大厦裙楼三层301 （100083）