智算加速
实现人与AI共创

智算时空（SegmAi Cloud）产品推介

智算加速（杭州）科技有限公司

2024年7月

V1.0

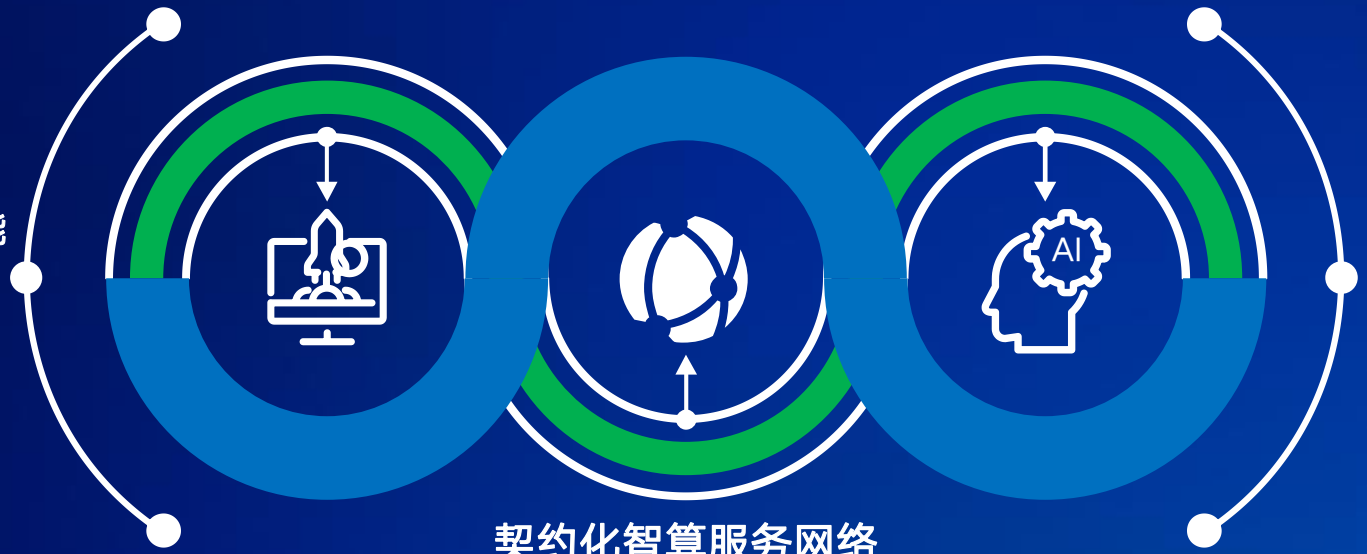
我们在做什么--以技术及模式创新，加速催化AI生产力

智算时空SegmAi Cloud

聚能算PlasmAi

面向算力底座的支撑场景
进行优化提效、统筹调度赋能

- 智算中心运营商
- 电信运营商
- 智算节点资源方
- 算力交易平台方



契约化智算服务网络

集大智SegmAi

面向AI应用能力的构建场景
提供一体化GenAI Ops使能

- 垂直行业大模型服务商
- 企业级自建场景模型需求方
- 科研院校
- AI应用服务商

智算时空 —— GenAI生产力融合服务网络

通过契约化将不同算力、模型、技术与服务等融入到应用中，形成一个可信的多方协作体系，支撑算力方、模型方以及应用方的高质量资源管理与服务保障，让参与方可以更好更多的享受AI发展红利！

智算时空的核心使能平台

赋能　连接　协作

基于GPU的
算力底座

面向大模型的
技术底座

企业级GenAI应用使能与催化

智算时空的核心使能平台将算力、模型、数据、加速组件与服务等在可信环境按需选择，一键装配，亚秒级就绪，按使用付费，让客户的AI需求与保障一站满足，让AI触手可及。

智算时空之装配服务

装配服务为算力及AI服务的需求放提供快速部署快速就绪能力，用户仅需简单操作就可以较为准确定义需求、部署方案智能化与部署过程的自动化，节约用户50～90%的部署时间，数百个差异化环境的部署也可在分钟级完成。

极高的性价比

- ✓ 智能化生成优化装配方案，屏蔽繁杂技术细节
- ✓ 10余秒完成装配
- ✓ 装配结果的自主诊断与调优

- ✓ 单人操作，无需专门技能亦可轻松上手，无需处理繁琐细节与调试
- ✓ 覆盖训练、推理及应用等多种场景，满足不同需求
- ✓ 支持复合式装配，快速构建面向特定领域的体系化AI应用

智算时空之IP资产库

智算加速资产库实现AI核心知识与能力的的标准化、智能化以及组件化，并在实现装配引擎中实现自动化装配。IP资产库包含算力、模型、数据、应用服务等加速包，主要为AGI技术研发、工程化以及大规模AI应用中的存在算力与环境的适配、资源调度、模型训练与推理性能、模型与应用间的协同等问题解决提供支撑。

IP资产库的价值

- ✓ vllm 加速库推理速度，相对HF、TGI处理推理请求量单位时间分别提升15、3.5倍，支持A100、H100、4090以及部分国产卡
- ✓ 应用级别kv cache加速组件，节省90%输入token的计算量，在智能体使用频繁的场景实现数10倍以上的加速。
- ✓ nerf数字人的渲染帧数提升300%，实80ms以内的时延
- ✓ ...

*我们可为您关注的模型、加速组件等IP资产提供一键装配部署服务，快速加速客户业务场景落地。

智算时空之AIGC集成服务

提供文生图、大模型、多模态大模型以及智能体等的AIGC服务，实现以算力池、模型池应对应用的需求池，支持单个应用10万用户以上的并发，并支持自动水平扩展，支持按使用次数或token付费，具有可靠性高、性能好以及成本低等三大特点。

主要特点

- ✓ 便捷的API接入，采用行业或事实标准
- ✓ 高吞吐、低延时、高可靠，可以满足每日千万次访问运营级应用需求。
- ✓ 极低的初置成本和使用成本，极高的性价比
- ✓ 安全程度高

* 通过算力与平台优化，让通用的服务更稳、更快、更便宜、更便捷，让优质的适配的模型更多、支撑更好！

智算时空之基础算力服务

提供容器化、虚拟化以及裸金属等多种类型的基础算力服务，支持高速IB网络互联的算力集群，实现专属算力设备与专属集群快速就绪与自动化管理。在数据安全、模型性能以及资源高效利用等方面提供高质量的支持，为企业级用户的不同需求构建最佳性价比曲线。

主要特点

- ✓ 算力可信，支持自有和可信的第三方算力接入，为用户提供最佳算力配比。
- ✓ 安全加固，最大程度保证模型安全、数据安全和通信安全
- ✓ 面向SLA服务保障，支持自动化运维

智算时空框架图


大模型训练
大模型推理
AI教学实训
科研
医疗agent
蛋白质预测
陪伴数字人
AI设计
.....
模型
加速组件
数据集
智算时空租户专属工作区
JupyterLab
大模型
数字人
智能体
...
AIGC服务
加速服务
工具/服务
公共算力池
租户专属算力池
智算中心 A
智算中心 B
智算中心 C
租户自有算力中心

跨区域资源共享与调度引擎

③ 装配引擎，自动理解用户需求构建高质量的部署方案，支持多元化的算力、多元化的模型以及服务的一键装配；

④ 服务感知与修复引擎，为大规模的服务运维提供支持；

② 高级算力计划（APC），实现基于算力资源拓扑感知的算力池化与容器编排优化，支持队列和Petri.Net的资源调度，支持单智算中心和多智算中心；

⑤ 契约化规则和记账引擎，让运营方可以自证清白，让参与方可以安心，实现50毫秒内完成多方记账；

① 可信节点引擎（trustnode@PlasmAI），让算力节点在可信环境下安全调度，支持细粒度GPU切分（MIG/MPS）与协同。

⑥ 契约化资源管理与调度平台；

价值主张

智算时空以技术及模式创新，加速、催化GenAI生产力服务，致力于构建并运营契约化AI智算价值融合网络，是专注于GenAI复杂和密集工作负载的开放式契约化GPU云服务平台，链接广泛的多元算力，为用户提供极具性价比的GPU算力最佳选项，为运营合作伙伴提供运营赋能，为IP资产合作伙伴提供IP资产投放。

智算时空用户

为企业、科研机构、开发者提供极具性价比的GPU算力最佳选项，内置典型加速库、工具链、模型库和应用集成，简化开发、训练与部署，加速生成式AI开发、生产部署以及集成到应用场景。

运营合作伙伴

基于闲置算力的可信接入和运营赋能，满足自用算力纳管、调度和加速的同时，支持契约化算力开放共享，通过弹性调度提高闲置算力利用率，以便获得可持续的收益。

IP资产合作伙伴

提供开放式GenAI生态平台，支持合作伙伴优质模型、加速组件等IP资产投放，助其充分释放IP资产价值，从而获得收益。

智算加速
实现人与AI共创

THANKS

FOR WATCHING

实现人与AI的共创