打造模型-硬件最佳适配平台
加速大模型落地

清昴智能 联合创始人 姚航
mlguider@tsingmao.com

让大模型触达世界每个角落

AIGC时代：人工智能的强大价值 vs 大模型落地的高门槛要求

- 大模型已成为新时代的技术底座，底层技术革新带来结构性机遇
- 大模型落地门槛高，参数量大带来的是开发、微调、部署、推理的成本提高
- 未来通用/垂直大模型数量、场景繁多，硬件环境分散且复杂

- **模型参数量大、显存与资源占用高**

以OPT-175B模型为例，部署一个模型仅推理就需要**5张80GB显存版本的A100**

- **推理延迟高、成本高**

据推测，ChatGPT运行每天花费至少70万美元；OpenAI推理集群对GPU的需求量近**10万张。**实际测算，大模型对算力的需求是现有小模型基础设施的**30倍**

- **芯片卡脖子问题日益突出**

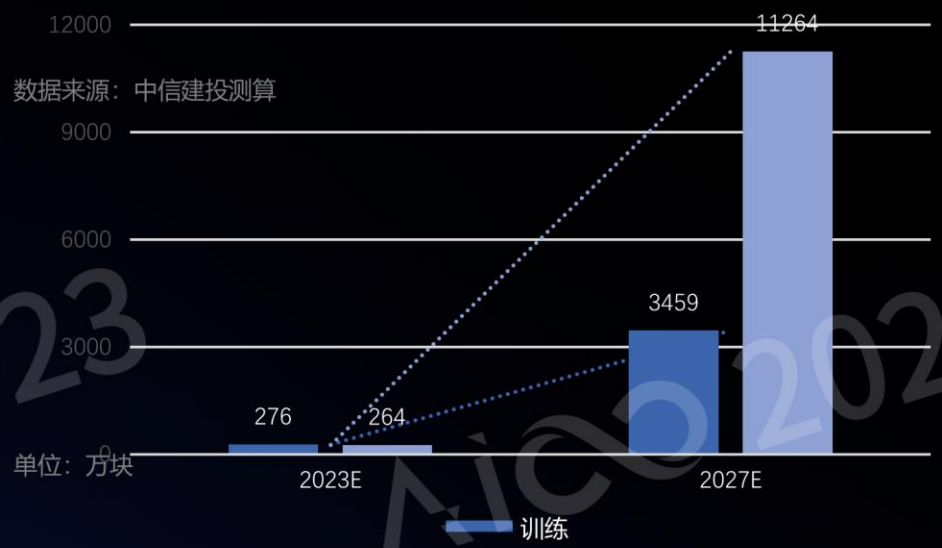
英伟达溢价高，货源受限；

非N卡工具链和库资源匮乏、产品迭代和兼容性的代价大；

- **模型种类多，硬件环境碎片化**

基础模型、垂直模型层出不穷，并不断更新。云边端各场景硬件环境碎片化分布，AI部署环境与大模型分布高度耦合，开发部署周期长

算力需求量预估（A100）



清昴智能
QINGMAO INTELLIGENCE

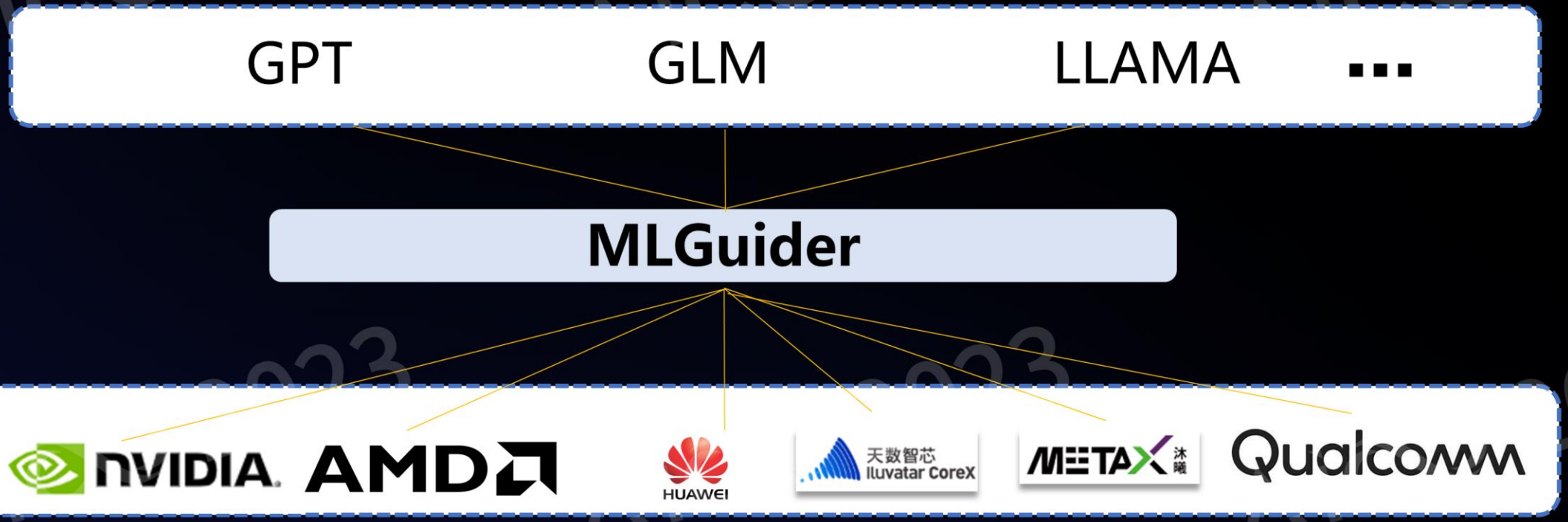
解决方案

自研的自动机器学习算法和优化系统，打造面向基础模型的自动优化工具链MLGuider

用AI构建AI，自动寻找模型和硬件适配的最优解

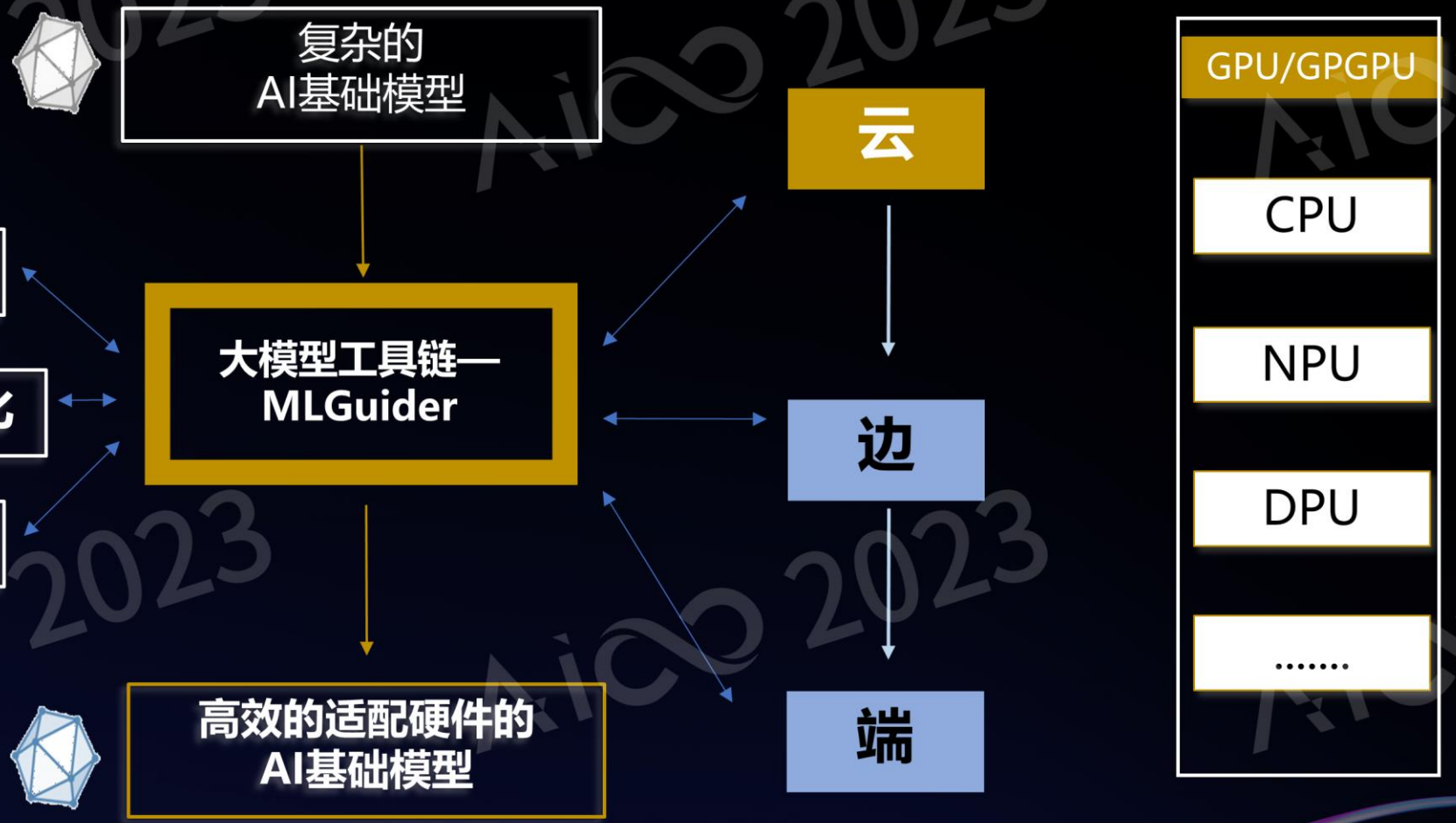
AI Infra：衔接模型/算法与硬件的工程化一环

硬件峰值性能 x Infra x 算法 = 最终利用率



打造链接大模型、底层硬件、行业用户及开发者的枢纽平台

MLGuider —— 统一的大模型软件工具链



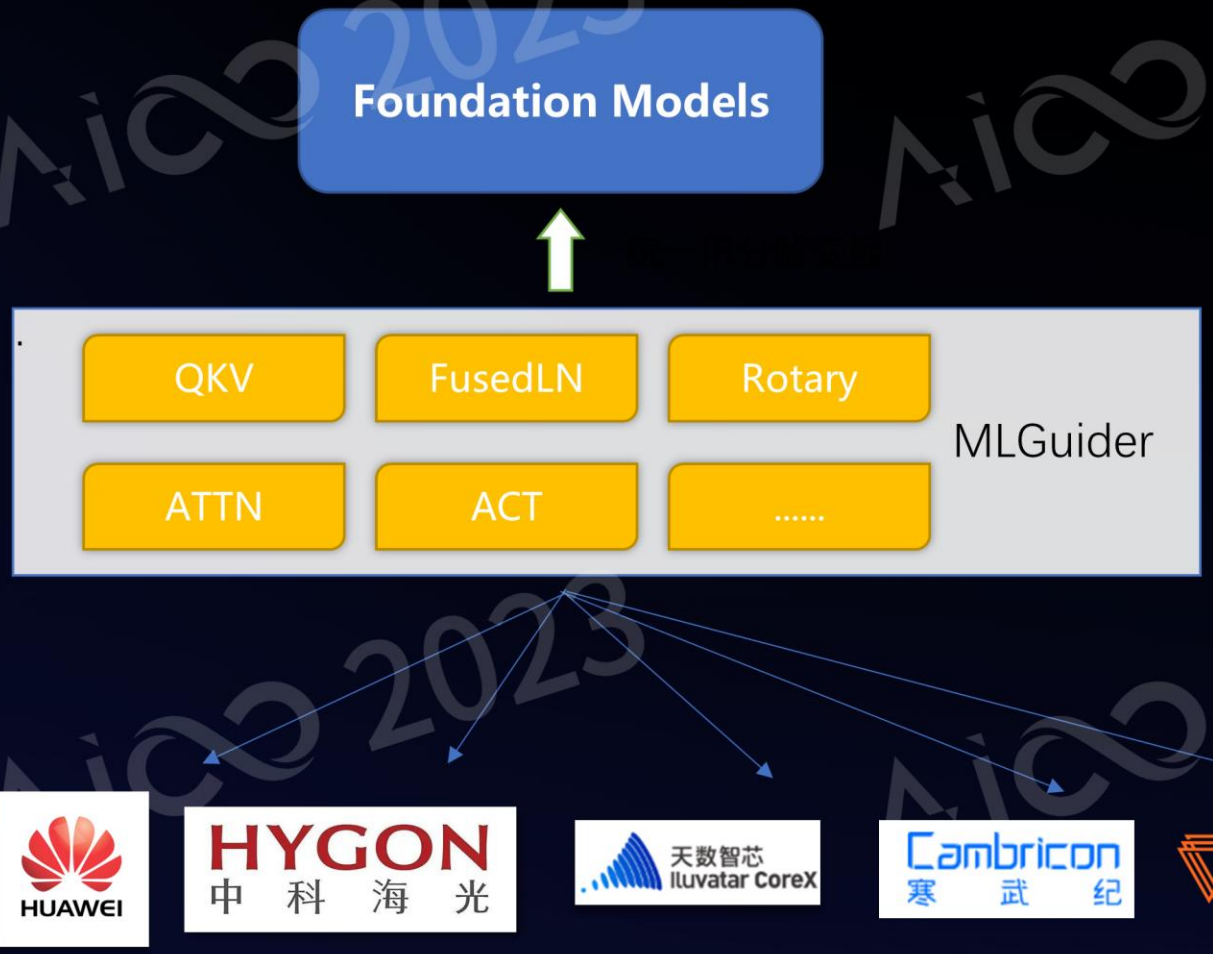
MLGuider —— 统一的大模型软件工具链

策略一：整合硬件环境，解决非N卡适配，实现多硬件无缝迁移

策略二：硬件感知的大模型压缩，实现硬件效率最大化

策略三：高并发、高吞吐、规模化部署，优化服务体验

策略一：整合硬件环境，解决非N卡适配，实现多硬件无缝迁移



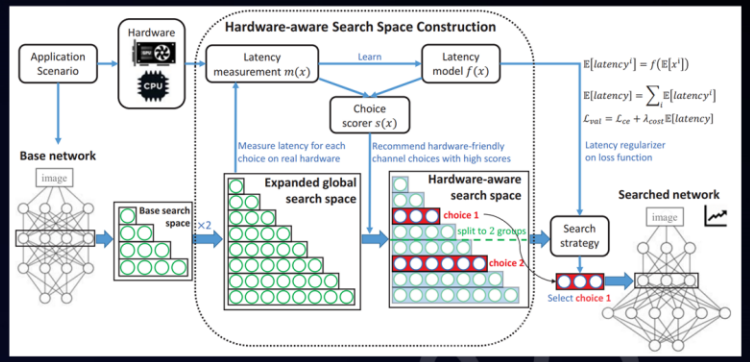
- 构建**兼容各类硬件**的算子库
- 支持上层**各类模型**的灵活构建

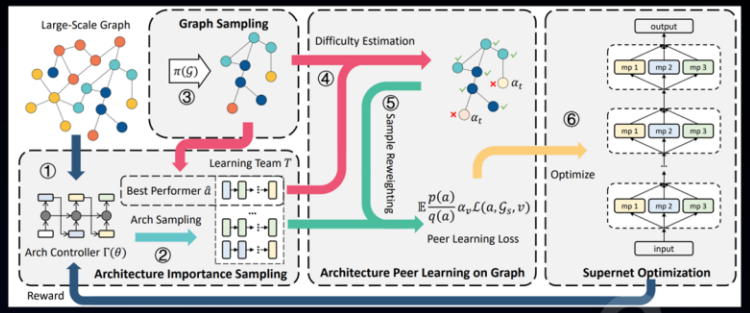
AutoML: 自动搜索最佳实现方式

适配效率↑ 执行速度↑

策略二：硬件感知的大模型压缩，实现硬件效率最大化

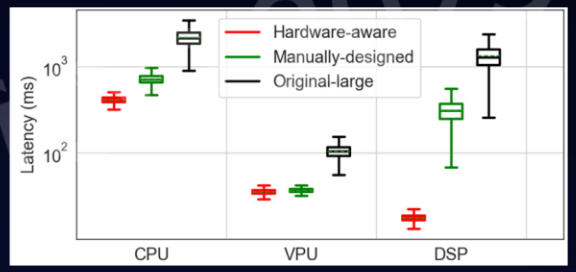
- 不同的底层硬件，最适合执行的**算子种类、参数**也不同，**执行效率**也不同
- 基于**硬件感知的架构搜索**对模型进行剪枝量化，实现**硬件感知的模型部署**



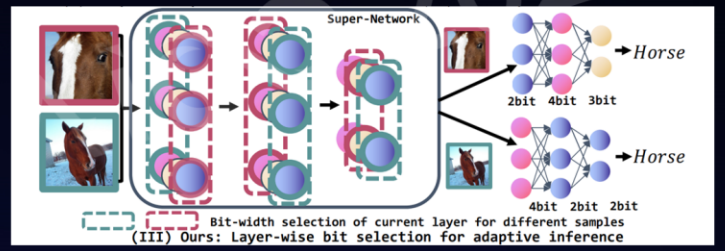


[ICML 22] 面向复杂空间的高效搜索

首个能在**1GPU天**内完成**亿级**网络的搜索框架



[ICME 20] 跨硬件可迁移的结构搜索

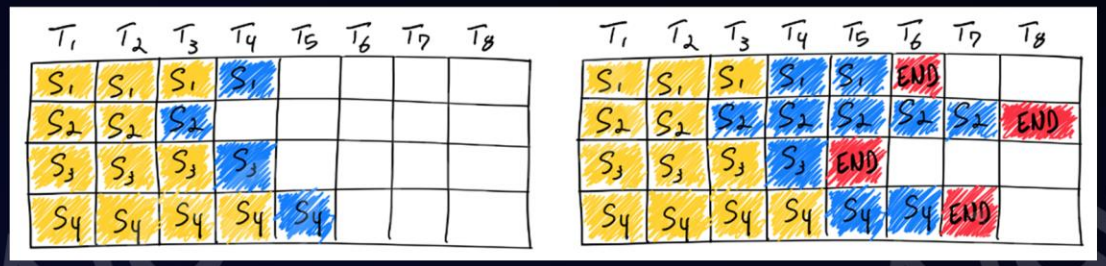


[ECCV 22] 全球首创动态量化网络

首次实现量化推理的**逐层动态路由**

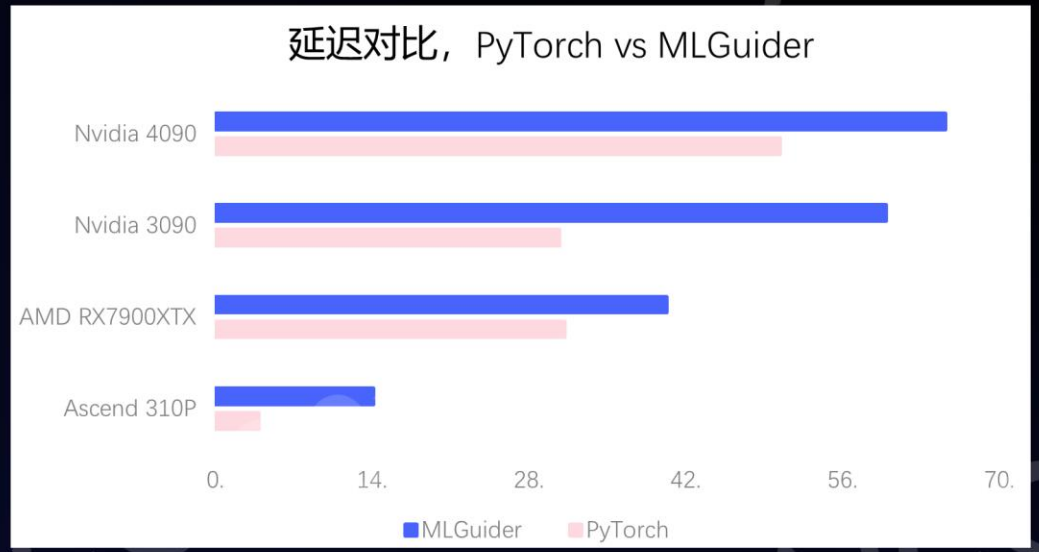
策略三：高并发、高吞吐、规模化部署，优化服务体验

- 统一显存管理方式，**自动按需分配/复用显存**，有效提升显存利用率
- 结合paged attention、token attention等，进一步**提升计算效率**
- 通过**多进程流水线管理**，提升CPU-GPU协同效率

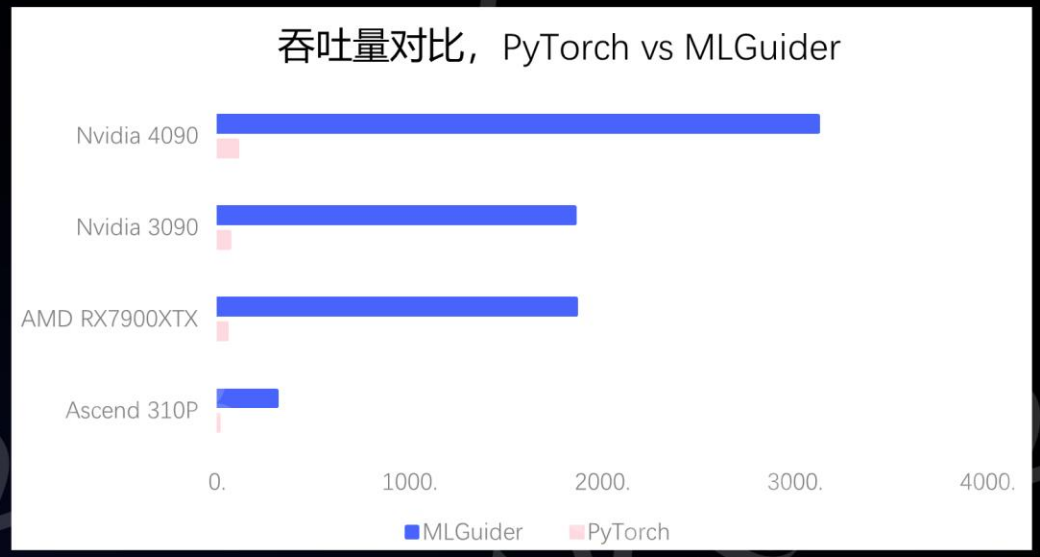


释放极致速度和吞吐量

推理速度 2-4x



吞吐量 10 - 30x



Quantized Cached Keys & Values

Inflight Batching & Continous Batching

Paged Attention & Flash Attention

Highly Optimized & Fused Kernels

清昴智能
QINGMAO INTELLIGENCE

硬件到软件全栈产品矩阵，加速企业大模型投产落地

硬件算力

大模型服务

企业挑战		清昴方案	
高端硬件短缺	场上高性能的计算服务器、GPU加速器等硬件资源可能供不应求，导致采购难度增加	算力生产 企业算力基础设施补足	01 高性能训推一体机 提供最优性价比算力支持和多种硬件选择
硬件成本压力	市场紧俏的硬件资源，导致成本的不可控制性增加		
非N卡适配差	不同硬件之间的适配性差异可能导致一些技术上的挑战，需要进行针对性的调整和优化	算力调度 企业技术支撑补足	02 一云多芯算力平台 打造算力基座，整合硬件环境，衔接上层业务
资源调度难	算力资源的运维调度，多场景混部，多硬件兼容		
开发部署难	部署大模型需要复杂的技术实现，包括模型的训练、优化、部署等环节，团队技术要求高	价值释放 企业应用能力补足	03 大模型开发部署平台 高性能推理部署，服务全周期管理
推理成本高	模型推理带来的高昂的能源成本、资源功耗...		
低功耗高显存	边缘端最看重功耗和续航，功耗就是成本问题	场景拓展 企业落地边界拓宽	04 大模型边端部署方案 极致模型轻量化小型化
模型本地化	企业数据隐私与安全问题		

大模型训推一体机方案

- AI产业的“Iphone”时刻已经到来，急需大量高能效的易用算力，轻松运行各类大模型

- 全流程覆盖：针对AMD GPU的大模型workload，涵盖从模型训练、微调到优化、推理、部署全流程支持

- 易用高兼容：无缝兼容PyTorch、Huggingface等主流框架，完美支持Llama、ChatGLM，并预装Stable Diffusion、CPM-Bee、Falcon等开源可商用模型

- 可扩展性强：已验证可支持近千卡（~400台服务器）规模的并行训练LLM能力

- 高效低延迟：从6B到176B LLM的高效推理能力，多场景性能对标甚至超越Nvidia A100 Pytorch + CUDA

基于AMD MI210
预装MLGuider全链路工具链的训推一体机

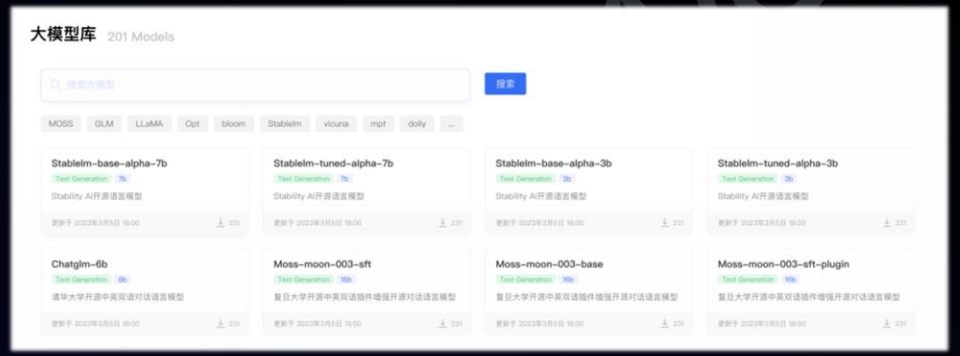
完全兼容AI工作流

PyTorch TensorFlow FastAPI docker mxnet deepspeed Flask

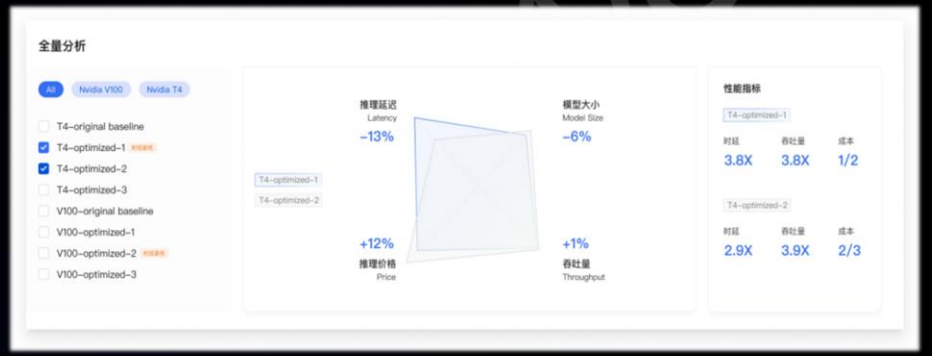
大模型开发部署LLMOps平台

基础模型服务构建与微调触手可及

基础模型构建



基础模型优化

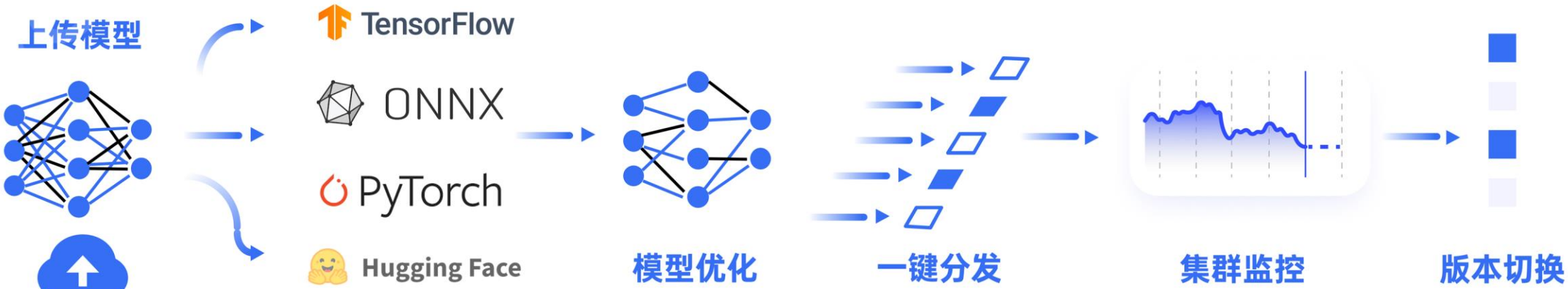


基础模型微调

基础模型部署

- 上线监控的**全流程管理** 版本控制、访问控制、 按量计费

一云多芯纳管平台



多种模型支持

自动推理优化

高可用高并发

性能监控分析

一云多芯

任务自适应调度

屏蔽底层硬件细节

统一纳管

大模型本地化及边端部署方案

清昂智能 QINGMAO INTELLIGENCE

本地化集群部署

数据隐私与安全 | 去中心化

易部署和易用 | 低延时高吞吐

企业私有数据 | 开源基础模型

LLMGuider | 算法工程师

本地化集群搭建 | LLM全生命周期管理

企业私有化大模型

加速企业数据到企业私有化大模型

边端部署

低延时高吞吐 | 低功耗高可用

资源占用低 | 易部署

- 适配高通芯片：已实现在高性能手机上，使用3G内存推理70亿参数大模型
- 适配瑞芯微芯片：已实现6B大模型高效推理部署

Qualcomm

Rockchip 瑞芯微电子

智能手机

智能机器人

智能音箱

总结

- 模型参数量大、显存与资源占用高
- 推理延迟高、成本高
- 开发与部署周期长
- 非N卡适配性差

→

- 大模型一体化整机方案
- 大模型开发部署LLMOps平台
- 大模型本地化及边端部署方案

→

- 极致的模型优化
- 高性价比的训推服务
- 低门槛的模型开发
- 云边端全栈的硬件适配

面向基础模型的软件工具链，让大模型触达每个角落

清昂智能科技（北京）有限公司

- 清昂智能是一家**AI模型推理部署解决方案提供商**，旨在为各行业客户提供顶尖的AI优化和工程化能力，致力于解决AIGC、自动驾驶、AIoT 等领域复杂AI模型的落地难、性能差、资源耗费高等问题

- 清昂智能创始团队来自于**清华大学计算机系**，团队成员来自清华、交大、新国立、爱丁堡、华为、阿里等海内外知名高校和大厂，成立一年内获得来自多家一线基金的**数千万投资**

- 公司已发布第一款**面向LLM的自研优化与推理工具链与平台产品**，填补LLM落地优化的空白，成为Nvidia、AMD、ARM、华为昇腾在内的多家硬件厂商的重要合作伙伴。致力于大模型的快速高效落地，帮助企业客户完成智能化升级

奇绩创坛 MIRACLEPLUS

綠洲資本 VITALBRIDGE

HUAWEI

Tencent

谢　谢