AI Infra研究报告

行业分析

公司分析

BUCT

01 行业分析

市场需求

竞争格局

行业机会：中国特色的AI大模型生态带给中间层服务商巨大的市场机遇

——三大机会点：国产芯片软件适配、算力集群供需匹配、终端客户低成本解决方案

产业链	1 芯片厂商	2 算力资产方	3 大模型企业	终端客户
美国生态	英伟达成熟生态	三大云厂商市场份额70%+	大厂领衔、头部格局基本确定	
中国生态	格局分散、生态尚未成熟	算力供需错配、实用性不足	格局分散、交付能力有限	
中国产业链现状	• **国产芯片格局分散**，从物理模型到架构逻辑差异性大 • 国产芯片厂商**缺乏足够的人力财力建设软件生态**，甚至难以完成高效的硬件编译	• 英伟达集群：全国各地分散的中小型集群（32~128卡）缺少获取稳定下游需求订单的能力，造成**算力闲置** • 国产芯片集群：**实用性弱**、无法发挥效率，下游政企客户无法作为主力集群使用	• 上游仍在活跃迭代的本土大模型企业仍有超过十余家 • 目前面对下游巨量的需求爆发，**缺乏足够的产品能力、服务能力和交付效率**	
服务商机会点	• **1）**协助国产芯片厂商**进行硬件编译及软件适配**（交付物举例：协助编译器、预装软件） • **2）**协助国产芯片厂商或其客户**针对下游应用场景进行算法优化加速**（交付物举例：加速卡）	• 1）**建立大规模上游算力纳管体系并统一调度**（交付物举例：算力云平台） • 2）针对国产芯片集群或英伟达+国产组合集群**进行同构/异构技术优化，提高集群可用性和商业化能力**（交付物举例：算力云平台）	• 1）**协助大模型客户进行推理训练/侧算力优化**（交付物举例：项目制；或针对底层不同芯片的软件接口） • 2）**协助终端客户**（如：车企、教育企业等）**开发一站式、低成本的AI大模型解决方案**（交付物举例：算力云平台/应用模块）	

行业机会1：国产芯片软件适配

——1）国产芯片厂商普遍软件生态搭建缺失；2）异构进一步加剧算力使用适配痛点

大模型时代爆发算力需求，国产化需求迫切但软件生态搭建缺失。进入大模型时代后，在Scaling Law驱动下训练和推理算力迅速增长，在美国芯片管制下海外核心高端 AI 芯片无法进入大陆市场，国产替代需求迫切性高，但国产芯片厂商普遍软件生态搭建缺失，国产卡的类CUDA生态搭建情况较差。

国产异构芯片与基于 NVIDIA 生态的错配加剧算力使用痛点。海外软件中间层由 NVIDIA 垄断，CUDA 提供了从算子库到编译器再到性能分析工具的全套解决方案，并有PyTorch、TensorFlow等主流深度学习框架的加速适配，形成了包含大量优秀代码的成熟生态社区。而国产算力生态混乱，各家芯片厂商独立圈地形成异构生态，在国产芯片原生生态尚未建立之前对接 CUDA 生态需要投入大量研发和时间成本，导致了异构算力在使用上的困难，为构建从国产异构中间软件生态提供了发展窗口。

	NVIDIA生态		国内算力技术生态
AI应用方	第三方公司 Runway, Midjourney, Microsoft, Character.ai		第三方公司+芯片企业 互联网厂商，行业ISV，芯片厂商A，芯片厂商B
模型研发方	第三方公司（模型+算子） OpenAI, Google, Stability.AI, Mistral AI		第三方公司+芯片企业 互联网厂商，大模型厂商，芯片厂商A，芯片厂商B
框架研发方	以第三方公司为主 Meta, Google, Microsoft, ….		芯片企业为主 芯片厂商A，芯片厂商B，芯片厂商C，芯片厂商D
算子提供方	Nvidia 400万+CUDA开发者		芯片企业为主 芯片厂商A，芯片厂商B，芯片厂商C，芯片厂商D
	基于Nvidia CUDA的统一生态		硬件异构下的烟囱式隔离型软件生态

海外模型研发、应用公司层均以各头部大厂、活跃度较高的初创公司为核心，在芯片、算子、框架搭建完善的基础上可以更方便的做基础模型及应用开发

海外框架研发层均以头部大厂或开源社区自发维护，受研究院所、校园、企业方所依赖，壁垒极深

海外芯片及算子层生态基本以英伟达显卡+CUDA为核心，性能、生态搭建均占绝对优势

国内模型研发、应用公司层均以各互联网大厂、行业解决方案公司为主，因底层建设尚有缺憾，目前无法匹及海外的生态、模型性能、应用落地效率

国内框架研发层也以国产芯片厂或少部分开源社区为主，缺少领军人物做好定一、统一生态

国内芯片及算子层生态基本以各国产芯片厂自研为主，性能参差不齐、生态搭建缓慢且分散精力和资源

国内烟囱式生态带来限制国内AI产业落地的关键问题

迁移难
-缺少需要的算子
-内存不够，OOM错误
-运行结果和NVIDIA不同

部署难
-需要维护多个版本的应用
-部署过程与硬件平台耦合
-难以获得及时有效的支持

利用率低
-算力只能到峰值的10%甚至更低
-难以锁定瓶颈
-算子运行效率低

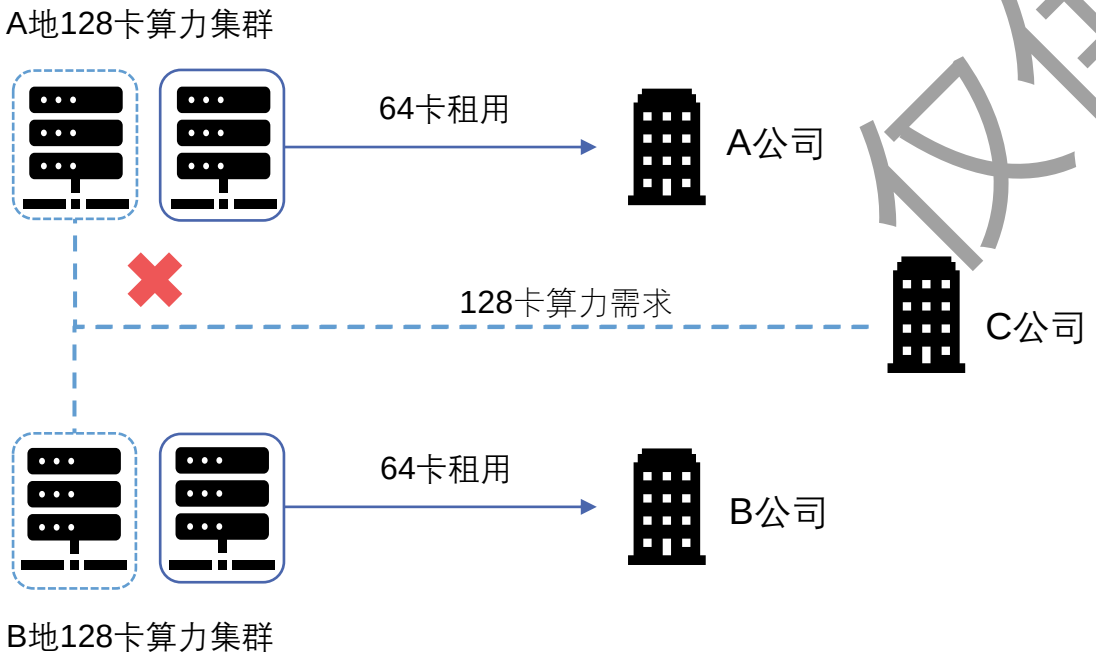
数据来源：内部资料，东方证券研究所，中金研究所

行业机会2：算力集群供需匹配

——1）分散的中小型算力存在供需错配，2）国产芯片集群可用性弱、亟需针对性调适

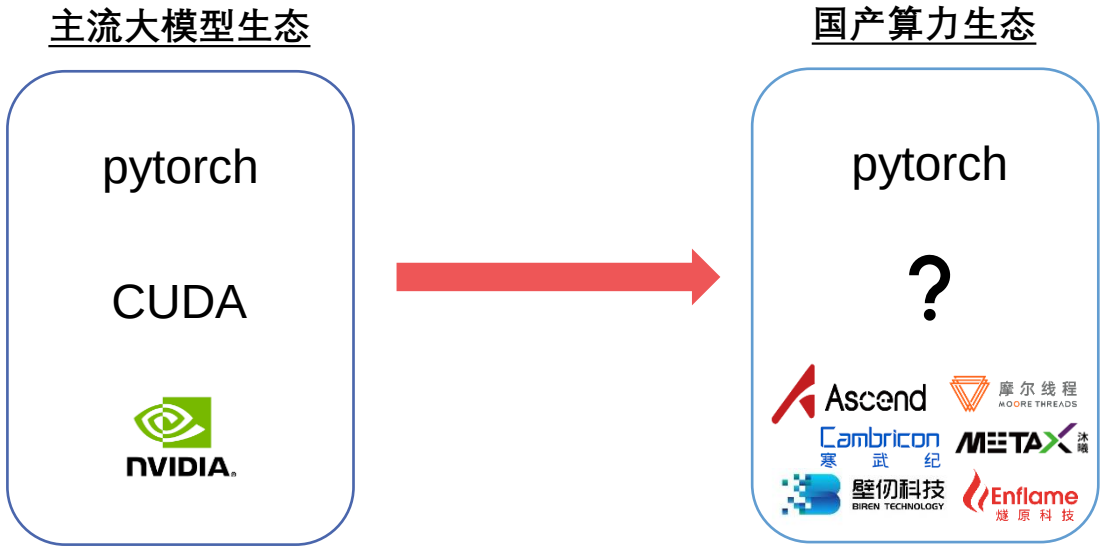
痛点一：上下游供需错配

- 国内目前智算中心超过2xx座，其中已运营的xx座，剩余处于在建中或已签约状态。根据有规划算力的项目统计，平均每个智算中心规划算力约为xxP，大约为xxx卡H100集群算力规模。
- 大量中小型算力中心处于xx卡的规模上下，对应的客户群体为有大模型推理落地、开源模型微调、小模型训练等需求的企业，需求集中在xx卡或以下的集群规模。
- 鉴于目前GPU算力租用市场先到先得和较为零散的交易特点，很容易出现算力中心可用资源和客户需求资源的不匹配。因此上下游均需要一个中游平台型企业进行上下游资源调配和对接。



痛点二：国产算力集群可用性弱

- 主流大模型生态以英伟达算力、CUDA、加上pytorch等成熟软件框架构成。由于英伟达算力供给问题，国产算力硬件不得不逐渐成为算力市场的重要组成部分。而从英伟达生态向国产算力生态迁移的过程中主要存在两大问题：
  - ✓ 国产硬件厂商众多，没有类似CUDA的通用中间层软件生态，导致迁移过程中软硬件适配出现问题，导致算力近乎不可用
  - ✓ 成熟上层框架如pytorch等无法针对各硬件厂商进行一一适配优化，导致即使硬件算力可用，效率相比英伟达也极其低下
- 因此，大量智算中心采购了国产硬件，但终端客户却无法使用国产硬件运行其大模型任务，导致国产算力集群建设进度缓慢、空置率高。



行业机会3：低成本解决方案

——1）大模型企业算力优化，2）终端企业低成本解决方案

大模型运用重构下游产品形态，下游算力需求巨大，算力使用方亟需低成本获取大算力的解决方案。主要机会包括：

- **1. 大模型企业：**业务增速高但算力不匹配、推理成本高，虽然掌握一定的AI技术能力但算力成本控制和获取算力能力不足。
  - ✓ **1）算力需求大：**当下大模型向多模态、长文本演进，模型推理需要5000+P的超大规模算力需求；
  - ✓ **2）算力不好用：**大规模算力推理要求多芯片、多集群，但在目前国产芯片异构局面下物理通讯、分布式并行、加速调优等造成额外适配成本。
- **2. 终端企业：**使用大模型进行现有成熟产品的升级或产品服务流程的重构，在AI技术能力、算力成本控制能力和获取算力能力方面都较弱。
  - ✓ **1）算力需求无法满足：**难以找到高性能、高容错的算力供给；
  - ✓ **2）模型推训缺乏技术团队：**行业客户缺乏训练大模型的技术团队，应用推理成本高，自建算力ROI太高且灵活度低。

	大模型企业	行业应用公司	解决方案公司
需求	需要与业务增速相匹配的推理方案	在成熟产品中内建AI大模型助手	基于大模型重构产品服务流程
痛点	• 5000+P的超大规模算力需求 • 单一集群、单一芯片难以满足 • 亿级别的年推理成本	• 难以找到短期阶段性500P+的算力需求 • 每100万DAU的应用完整运行成本每年高达1000万	• 无法承担千万级的算法人员成本 • 希望0成本启动，按效果增加算力预算
能力	• AI技术掌控能力强 • AI算力成本控制能力中 • 获取算力能力弱	• AI技术掌控能力弱 • AI算力成本控制能力中 • 获取算力能力弱	• AI技术掌控能力极弱 • AI算力成本控制能力极弱 • 获取算力能力弱

竞争格局：

——算力、异构、模型能力多方面因素导致AI Infra在中国有不同角度的市场机会

	A公司	B公司	C公司	D公司	E公司	F公司
核心产品						
业务定位						
切入角度						
核心能力						
客户群体						
产品形态						
商业模式						
资本投入水平						
潜在竞对						

BUCT

02

公司分析

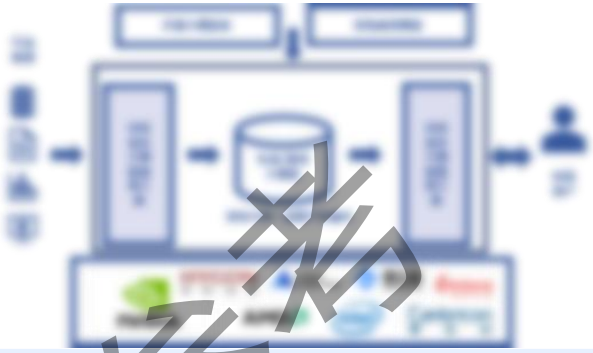
产品

技术

财务

标的综述：2B公司

——产品能力：面向芯片、服务器、智算中心和大型算力终端用户

	芯片厂商（某芯片厂）	服务器厂商（某服务器厂）	大模型公司（某大模型公司）
			
合作时间	YY-QQ	YY-QQ	YY-QQ
需求痛点	目前国内芯片市场处于早期，各芯片厂商专注硬件抢占市场份额，自行搭建软件生态的研发资金和时间投入太高，需要第三方软件服务商帮助构建生态	国产异构芯片集群需要一个统一的软件基础设施（管理+调度+开发+推理）减少适配成本、优化芯片算力性能	解决用户在异卡算力迁移和建立异构芯片算力中心时的软件基础设施
切入角度	对标 CUDA 的软件工具链和算子库 推理引擎	推理引擎	英伟达卡到华为卡的大模型迁移与优化 未来支持公司混合算力中心的软件基建
产品形态	lisence in，收费为单卡售价的 xx%，按出货量阶梯计价，后收费模式	作为预装软件 OEM in，参与一些优化模块，收费为裸卡售价的xx%，根据出货量进行采购	偏项目制
合同金额	xx元	xx元	xx元
合同账期	月度/季度等	月度/季度等	月度/季度等
毛利率水平	~xx%	xx%	xx%
潜在竞对	xx	xx	xx
商业模式判断	**优势：**编译研发积累、行业认可度；竞争少 **风险：**国产芯片厂收敛趋势不明朗	**优势：**依托异构编译基础做推理加速 **风险：**竞争激烈、优势不明显；通用infra平台与追求性能相悖	**优势：**异构编译基础良好，切入训练迁移和混合算力中心 **风险：**竞争激烈；目前订单较少、发展不清晰
管理层规划	用户侧构筑壁垒；未来可能被新芯片或互联网厂收编	未来AI小型算力需求上升的重点拓展方向	目前以训练迁移项目为主，未来连同算力底座做算力基建

数据来源：内部资料，客户访谈

标的综述：2B公司

——技术能力：覆盖了算子层、编译层，搭建基于芯片底层的多卡适配、性能优化平台

主要目的是做好国产卡适配与推理加速，主要技术基于：

1）算子层－xx技术：除了GPU厂自研的通用算子，B公司另编写、优化全套复杂算子库，快速补齐生态能力。

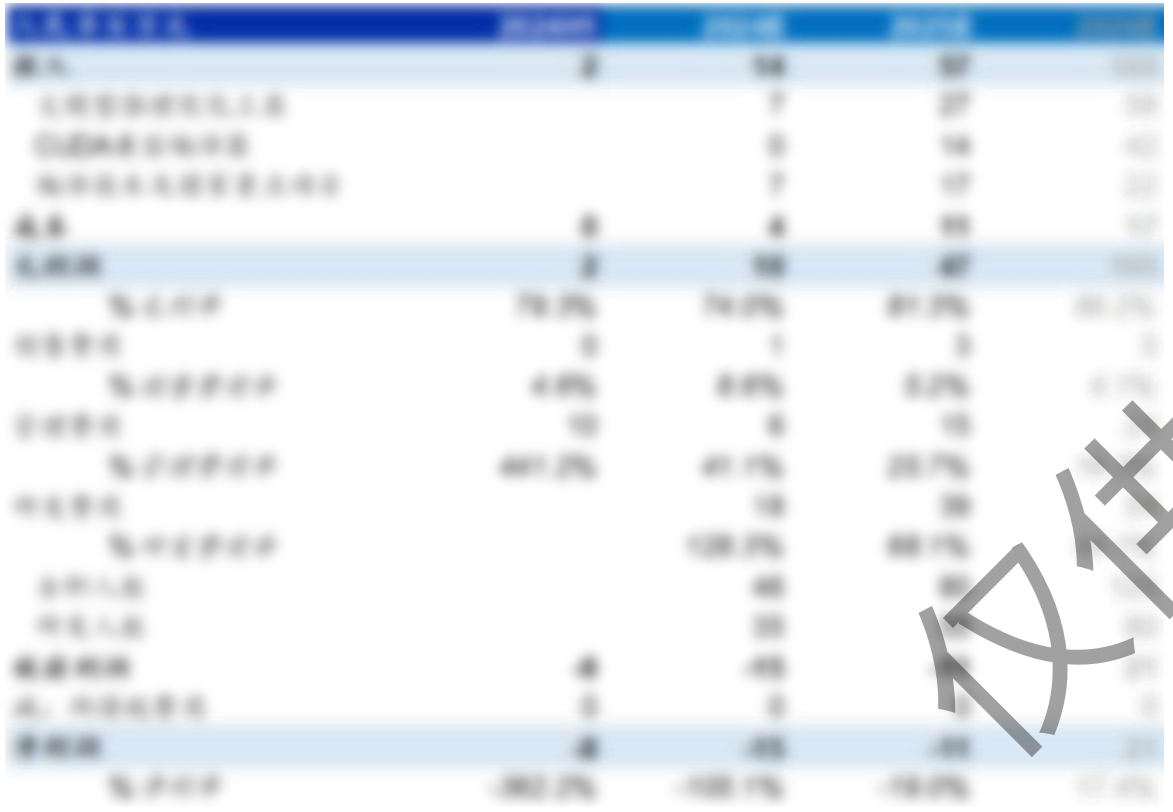
2）编译层－xx技术：团队做编译器出身，有丰富的国产卡异构架构适配能力，支持 CUDA MLIR、TVM 和 Triton 编译语言的转化。此外，在多卡的通信间做了优化调度，从底层就能使模型推理提升效率。

3）框架层－xx技术：基于上述两层的基础搭建，形成推理/微调引擎，供应给算力需求方、服务器厂商等使用，以更好的在英伟达卡+国产卡的异构框架下起到加速模型推理的能力。

数据来源：内部资料

标的综述：②B公司
——财务面

利润表



资产负债表

在手订单