

时势造英雄，宜谋定而后动

——AI算力租赁行业研究框架

行业评级：看好

2023年10月8日

分析师 刘雯蜀
邮箱 liuwenshu03@stocke.com.cn
证书编号 S1230523020002

分析师 刘静一
邮箱 liujingyi@stocke.com.cn
证书编号 S1230523070005

分析师 李佩京
邮箱 lipeijing@stocke.com.cn
证书编号 S1230522060001

并购家 免费行业报告网

IPOIPO.CN 每日更新 不用注册会员 无广告 永久免费

1、商业逻辑——为什么AI算力租赁具有商业价值

- AI算力租赁业务聚焦于解决大模型训练的算力需求，连接多方资源。其商业价值为实现算力资源配置效率的最优化，即设备调试和网络搭建的know-how得以最大化复用。AI算力租赁业务涉及到多方实体，包括项目投资方、项目建设方、项目运营方、模型研发商以及终端模型用户。
- 短期来看**，我们考虑国内15家头部大模型厂商对标GPT-3模型的训练需求，测算需要1920台A100/A800服务器，**对应15360张GPU**；**长期来看**，我们考虑国内5家头部大模型厂商对标GPT-4模型的训练需求，测算额外需要13705台A100/A800服务器，**对应近11万张GPU**。

2、盈利模型——AI算力租赁业务的盈利能力

- 从价格来看**，GPU的租金价格随着配置性能的提升呈现上升趋势（不考虑CPU、存储等其他参数的影响），其中单就A100算力而言，**最高配置约为最低配置价格的1.8倍**。**从成本来看**，AI算力租赁业务的运营成本主要包括**设备折旧、数据中心日常运营、以及人员成本**，其中设备折旧为非现金支出。

3、相关标的——哪些上市公司在布局AI算力租赁

- 布局AI算力的上市公司可分为四类**：IaaS云服务厂商、传统IDC服务厂商、AI算力用户向上游扩张、跨界布局第二生长曲线。
- 从建设方式来看**，分为**自建和共建**；**从经营方式来看**，分为**自用和出租**。

4、海外映射——GPU云

- 短期来看**，海外云厂商大量囤积英伟达A100/H100芯片；**长期来看**，头部厂商推进自研AI芯片。
- GPU云是算力租赁业务的长期进阶方向，具有更高的价值量和技术壁垒，市场想象空间更大。

- 1、报告中对于商业模式的总结以及对于盈利模型的测算包含部分主观假设
- 2、AI算力租赁业务落地不及预期
- 3、上游AI服务器供应持续短缺，或受到美国商务部禁令影响，中国厂商无法获得英伟达相关服务器设备
- 4、AI算力产业政策发生重大变化
- 5、布局AI算力租赁业务厂商持续增加，市场竞争风险加剧

目录

CONTENTS

01

商业逻辑——
为什么AI算力租赁具有商业价值

02

盈利模型——
AI算力租赁业务的盈利能力

03

相关标的——
哪些上市公司在布局AI算力租赁

04

海外映射——
GPU云

01

商业逻辑—— 为什么AI算力租赁具有商业价值

大模型训练的爆发带动高端AI服务器（以A100、H100为主）需求的涌现：

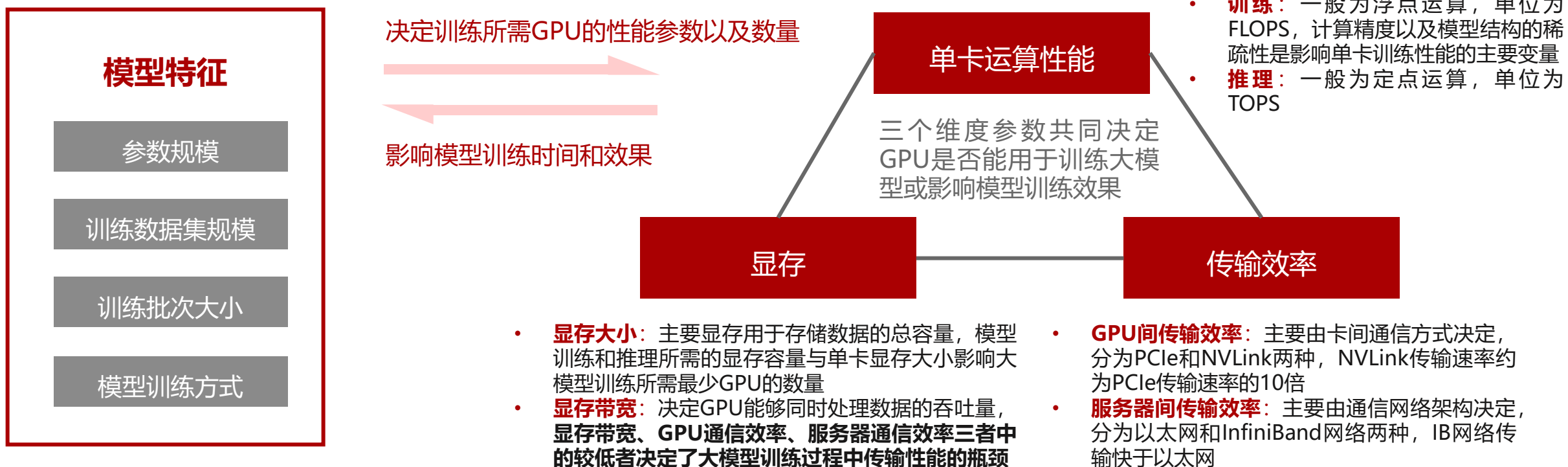
- 随着模型参数量的增加，需要的GPU数量非线性增加。根据英伟达和微软联合发布的论文，给出一个模型的参数量P、GPU数n、每个GPU的吞吐量X、以及需要训练的总token数T、训练时间的估计公式为： $\text{Training time} \approx 8TP/nX$ 。
- 以GPT-3到GPT-4的迭代为例，参数规模增大10倍，对应训练计算量增加至少60倍。
 - LLaMA-65B模型——使用2048张A100，训练21天（基于32k大小的词表，在1.4T的Tokens上）
 - GPT3-175B模型——使用1024张A100（80G显存），训练34天
 - GPT4-1800B模型（预估）——使用约25000个A100GPU，训练90-100天（利用率约32%-36%）
- TrendForce持续上调人工智能服务器出货量。根据TrendForce集邦咨询，2023年AI服务器（包含搭载GPU、FPGA、ASIC等）出货量近120万台，年增38.4%，占整体服务器的出货量有望从2023年近9%提升至2026年15%，同时上修2022~2026年AI服务器出货量年复合成长率至22%，而AI芯片2023年出货量将成长46%。

供需失衡导致GPU价格上涨，同时交货周期拉长：

- 英伟达A100价格2022年12月份至2023年4月上半月期间，5个月价格累计涨幅达到37.5%，根据IDC，2023年5月17日A100 GPU市场单价达15万元
- 英伟达A800价格2022年12月份至2023年4月上半月，5个月价格累计涨幅达20.0%，根据IDC，2023年5月17日A800 GPU市场单价达9.5万元
- 交货周期拉长，根据集微网，截至2023年5月13日，交货周期由此前一个月左右至三个月甚至更长

单卡运算性能、显存和传输效率是影响GPU训练大模型效果的三个关键参数

- 从算力供给侧来看，单卡每秒运算次数、显存、传输效率从三个维度共同影响大模型的训练效果
- 从模型需求侧来看，模型参数规模、训练数据集规模、训练批次大小以及模型训练方式决定了模型训练所需的总计算次数、训练和推理阶段所需的显存大小，从而进一步决定了大模型训练所需最少GPU数量以及模型训练时间。



英伟达A100-SXM和H100-SXM为目前训练大模型的首选GPU

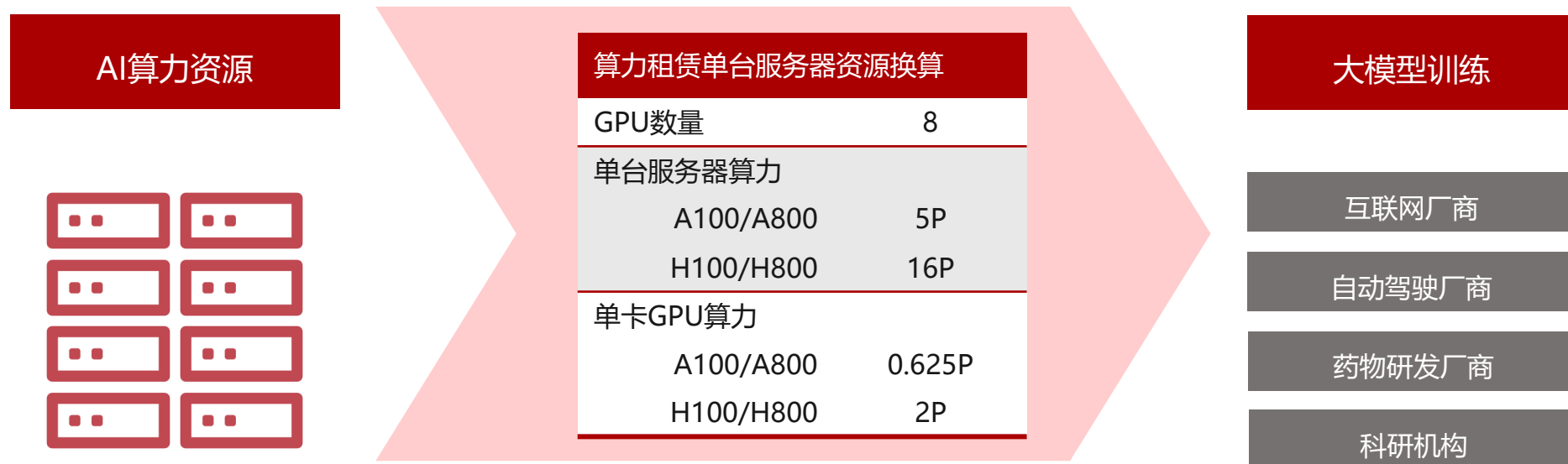
- 英伟达GPU根据使用场景分为多种类别，其中以RTX4090等为代表的消费级显卡主要用于游戏场景，以Tesla T4等为代表的工业级显卡主要用于图形处理和人工智能推理领域，而以V100、A100、H100等为代表的加速计算卡主要用于人工智能训练场景。
- 英伟达2017年5月将Tensor Core引入V100**，大幅提升GPU进行混合精度矩阵乘加运算的速度，可大幅缩短AI场景下大规模矩阵计算的时间，为人工智能场景下的加速计算奠定了基础。
- 随着模型参数的增加，对于GPU运算性能以及通信传输性能的逐步提升，**目前搭载NVLink和IB网络的A100-SXM和H100-SXM成为大模型训练的首选GPU。**

表：英伟达不同类别GPU性能梳理

		游戏	图形处理/AI推理	人工智能加速计算						
		4090	T4	V100		A100			H100	
		RTX4090	Tesla T4	V100-SXM	V100S-PCle	A100-SXM	A100-PCle	A100-PCle	H100-SXM	H100-PCle
运算性能	FP32-cuda core (TFLOPS)		8.1	15.7	16.4	19.5			67	51
	FP16-tensor core (TFLOPS)	/	65	125	130	312/624*			1979	1513
显存	容量 (GB)	24 (GDDR6X)	16 (GDDR6X)	16 (HBM2)	32 (HBM2)	80 (HBM2)	40 (HBM2)	80 (HBM2)	80	
	带宽 (GB/s)	1008	300	900	1134	2039	1555	1935	3430	2048
传输	卡间 (GB/s)	PCle 4.0	32 (PCle 3.0)	300 (NVLink)	32 (PCle 3.0)	600 (NVLink)	64 (PCle 4.0)	64 (PCle 4.0)	900	128 (PCle 5.0)
	服务器间 (GB/s)	/	/	500 (双向IB网络)		500 (双向IB网络)			900 (双向IB网络)	
半精度(FP16)单卡算力 (P)			0.07	0.13	0.13	0.62*			1.98	1.51
单卡平均售价 (万元)		1.45	0.67	2.70	5.60	11.00	8.40	11.50	19.00	19.00

AI算力租赁业务聚焦于解决大模型训练的算力需求，连接多方资源

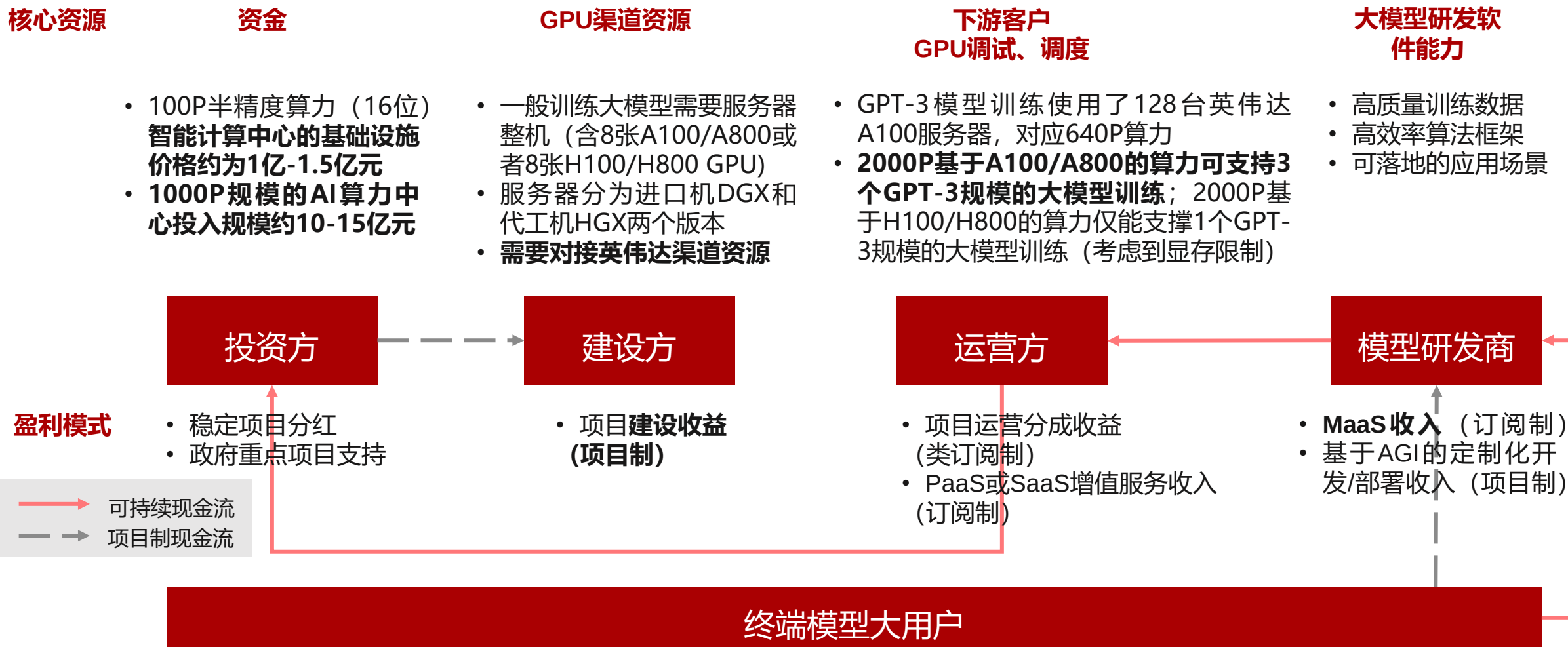
- AI算力租赁业务产生的两个催化条件：1) 可用于大模型训练的算力资源和大模型训练需求供需失衡，**短期算力需求远高于算力供给**（尤其针对用于大模型训练的英伟达A100-SXM和H100-SXM两类GPU）；2) 时间对于大模型研发厂商而言是较为稀缺的资源，即**先行完成大模型研发的厂商有望获得更多的先发优势**。
- AI算力租赁的商业本质为具有大模型训练需求的软件研发厂商向具有GPU资源的厂商租赁GPU算力，按月或按年支付租金，市场上常见的租金计量方式包括：1) 按整台服务器租赁（每台服务器含8张GPU），租金按照每台每月计量；2) 按算力规模租赁，租金按每P每年计量；3) 按单张GPU租赁，租金按照每GPU每小时计量。以上三类计量方式可相互换算。



AI算力租赁业务的商业价值为实现算力资源配置效率的最优化，即设备调试和网络搭建的know-how得以最大化复用

- **若只比较服务器采购成本与租金成本，AI算力租赁并不具备性价比优势：**
 - 根据阿里云网站，8卡英伟达A100-NVLink（80GB显存）的GPU服务器的月租金约为13.34万元，对应全年租金约为160万元
 - 根据京东商城数据，8卡英伟达HGX加速显卡A100 SXM模组售价为160万元，基本与同等规格GPU服务器月租金相当
- **但若考虑到服务器的等待、调试、运维成本，以及软件研发的试错成本，AI算力租赁对于大模型研发厂商来说仍极具性价比：**
 - **硬件交付等待周期拉长：**硬件交货周期在3个月以上，而28天足够训练一款大模型
 - **GPU服务器调试难度增大，GPU利用率低，间接增大GPU算力成本：**根据趋动科技，多数用户GPU利用率仅达到10%-30%（根据头部科技，OpenAI 在GPT-4模型训练的GPU利用率约为32%-36%）。一方面，GPU的任务运行需要与CPU协作，两者资源需要合理调配；另一方面，随着模型参数量增大，单机无法完成运算，引入分布式任务的过程中机器间通信效能再次对服务器利用率造成影响
 - **算力运维需要额外的专业团队，进一步增大算力使用及管理的成本：**涉及到资源管理审批、调配、监控、排错、管理，以及底层Driver、CUDA等各种API、AI框架等工作，同时部分公司还需考虑到支持异构算力芯片以报证供应链安全
 - **考虑到模型参数调整和架构修改的时间成本，自有算力成本进一步提升**
- **互联网厂商为例外：互联网大厂自建AI算力具有较高复用价值，可形成规模化效应，故而自建AI算力为最优解**
 - 一方面，互联网厂商拥有庞大的数据中心运维团队，自建AI算力可进一步提升运维团队的利用效率，提升人均产出
 - 另一方面，互联网厂商可将GPU资源池化后，将未占用部分打包出租，进一步提升单位算力建设的投入产出比

AI算力租赁业务涉及到多方实体，包括项目投资方、项目建设方、项目运营方、模型研发商以及终端模型用户



地方政府是AI智算中心建设的主要规划与投资方

- 根据IDC圈，全国有超过30个城市正在建设或规划智算中心，其中一些已经投入运营或即将投入运营，总规划算力达到了数十EFLOPS。若假设未来全国智算中心算力达到50EFLOPS（对应50000PFLOPS），对应AIDC建设规模空间为500-750亿元。
- 据不完全统计，目前已经明确公开宣布规划或建设中的地方智算中心规模已超26000P（统一换算成FP16口径）。

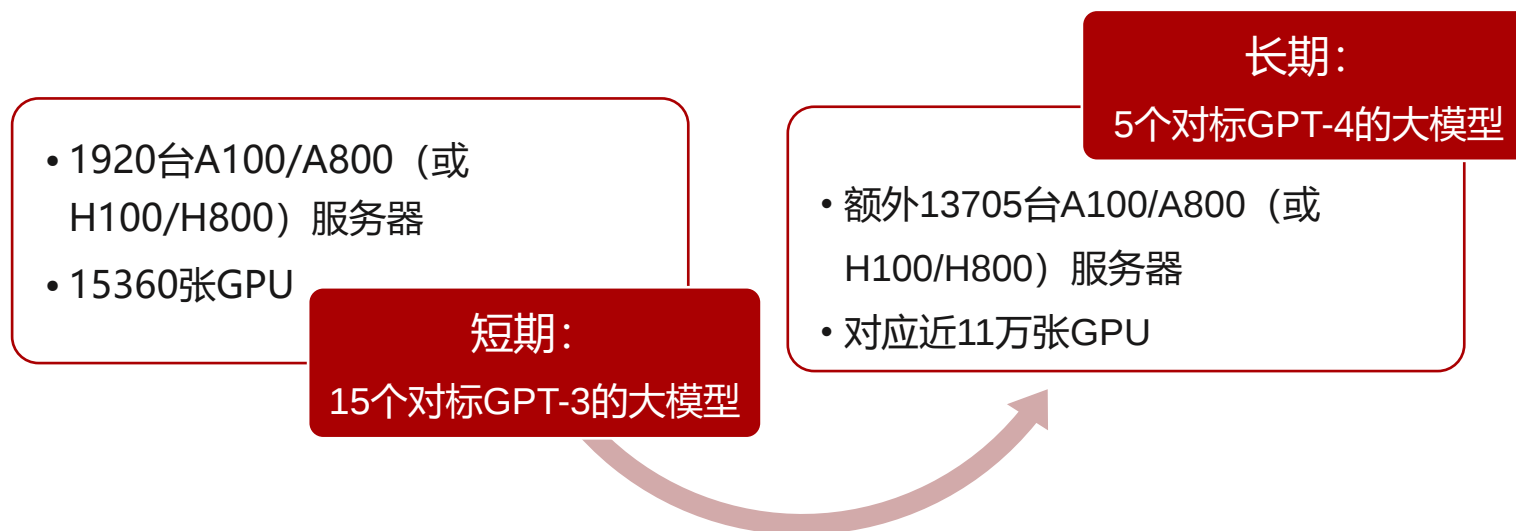
表：地方政府智算中心建设规划（基于公开资料不完全统计）

地区	规模 (PFLOPS)	服务器	合作方	状态
北京亦庄	1000			规划
广州	900	鲲鹏+昇腾	华为	规划
成都	300 (FP16)	昇腾	华为	建设中
沈阳	200			规划
大连	200	昇腾	德泰控股	规划
福建	295			规划
南京市浦口区	900			建设中
贵阳	2000		摩尔线程、威星智能	规划
芜湖	3000			规划
宁波	200			规划
青田	100		浪潮信息、谷梵科技	规划
昆山	500			建设中
杭州	400			建设中
长沙	1000			建设中
北京朝阳	1000			规划
北京中关村科学城	4000		电信、京能	规划
长春	200		神州控股	规划
上海松江			上海仪电、云赛智联	规划
上海临港	5000 (FP32)	英伟达A800		建设中

资料来源：IDC圈、北京亦庄、网信辽宁、福州新区、杭州、上海、深圳、成都、贵阳、芜湖、宁波、青田、昆山、杭州、长沙、北京朝阳、北京中关村科学城、长春、上海松江、上海临港、

对标GPT-3和GPT-4模型，算力需求非线性增长，受限于显存，单卡算力的升级不会减少模型训练所需GPU的数量

- GPT-3模型训练使用了128台英伟达A100服务器（训练34天），对应640P算力，GPT-4模型训练使用了3125台英伟达A100服务器（训练90-100天），对应15625P算力。从GPT-3至GPT-4模型参数规模增加约10倍，但用于训练的GPU数量增加了近24倍（且不考虑模型训练时间的增长）。
- 短期来看，我们考虑国内15家头部大模型厂商对标GPT-3模型的训练需求（百度、腾讯、阿里、字节、京东、美团、讯飞、网易、360、商汤、云从、百川、智谱、minimax、深言），则需要1920台A100/A800服务器（考虑到A100和H100的单卡显存容量相同，使用H100理论上也需要相同数量的服务器，但可以大幅缩短训练时间），对应15360张GPU。
- 长期来看，我们考虑国内5家头部大模型厂商对标GPT-4模型的训练需求，则额外需要13705台A100/A800服务器，对应近11万张GPU。



02

盈利模型—— AI算力租赁业务的盈利能力

支持NVLink传输的A100算力资源主要集中在互联网大厂，但仍较为稀缺

- 从供给端来看，各大云厂商尚未推出基于H100/H800的云端GPU实例，目前的可租用资源以A100为主，但**支持NVLink传输以及单卡达到80GB显存的GPU算力资源**，目前只有火山引擎能提供可供租用的资源。
- 从价格来看，GPU的租金价格随着配置性能的提升呈现上升趋势（不考虑CPU、存储等其他参数的影响），其中单就A100算力而言，最高配置约为最低配置价格的1.8倍。

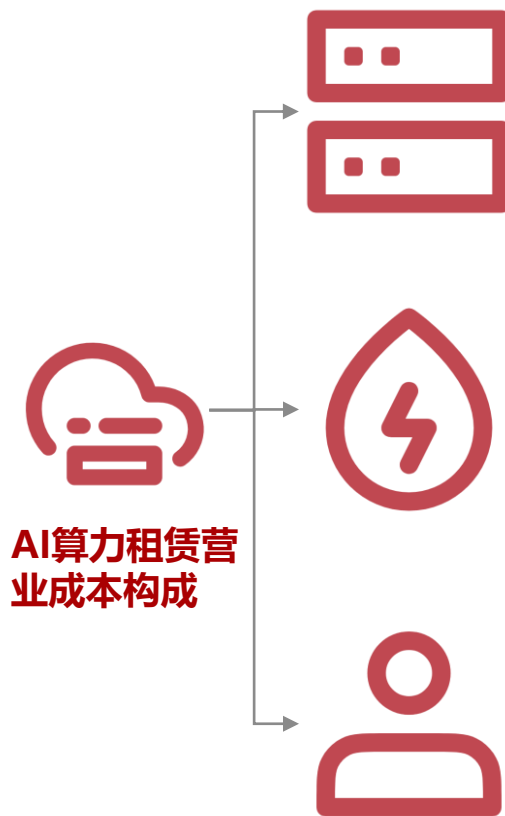
表：各类云厂商网站GPU实例售价对表（单位：万元/月，均按照8卡GPU换算，截至2023.9.20报价）

GPU型号	显存	卡传输	腾讯云	阿里云	火山引擎	天翼云	优刻得	青云	均值
A100	80GB	NVLink		13.34*（缺货）	15.77				14.6
A100	40GB	NVLink	11.40						11.4
A100	未披露							10.16	10.2
A100	40GB	PCIe				8.18			8.2
V100	32GB	NVLink	4.8	6.83	5.19				5.6
V100	16GB	未披露					1.99		2.0
年折扣			半年	8.8折	/	/	/	8.8折	
			1年	8.3折	8.5折	8.3折	8.3折	8.3折	
			2年	7折	7折	7折	7折		
			3年	5折	5.5折	5.5折	5折	5.3折	

注释：以上报价对应的CPU、存储等配置或有不同，我们仅列示GPU参数作为价格参考

AI算力租赁业务的运营成本主要包括设备折旧、数据中心日常运营、以及人员成本，其中设备折旧为非现金支出

- **设备折旧**：设备折旧在AI算力租赁成本中占比最高，其中既包括服务器也包括网络设备等，且设备折旧年限对毛利率影响较大
 - 仅以服务器为例：以市场7月A800服务器成交价140万元为例，若按3年摊销对应月折旧成本约为3.9万元，按5年摊销对应月折旧成本约为2.3万元，参考市场8卡A800-80GB-NVLink实例月租金14.6万元，对应成本占比分别为27%和16%，对毛利率影响11pct。
- **数据中心日常运营**：主要包括数据中心运营所需的成本以及部分情景下对于机房改造的成本
 - 数据中心运营：能源功耗成本（水电等）、散热成本、房屋租金成本等
 - 数据中心改造成本（或有）：英伟达DGX H100服务器系统功耗约为10.2kW，而传统数据中心每个机架的功耗约7kW，故而若采用H100/H800建设AI算力集群，还需对传统数据中心机房进行改造
- **人员成本**：参考奥飞数据2022年报，人工成本在IDC服务业务成本中占比约3%，占IDC服务收入比例约2%。



设备折旧：

- 服务器+网络设备+安全设备+存储设备等
- 非现金支出成本
- 折旧摊销年限对毛利率影响大

数据中心运营：

- 能源功耗+散热+房屋租金等
- 自有机房：按照原有成本计量
- 采用第三方机房：可按照机柜租金进行估算（默认租金包含所有数据中心运营成本）

人员成本：

- 参考奥飞数据，人工成本在IDC服务营业成本中占比约3%，占IDC服务营业收入比例约2%

03

相关标的—— 哪些上市公司在布局AI算力租赁

布局AI算力的上市公司可分为四类

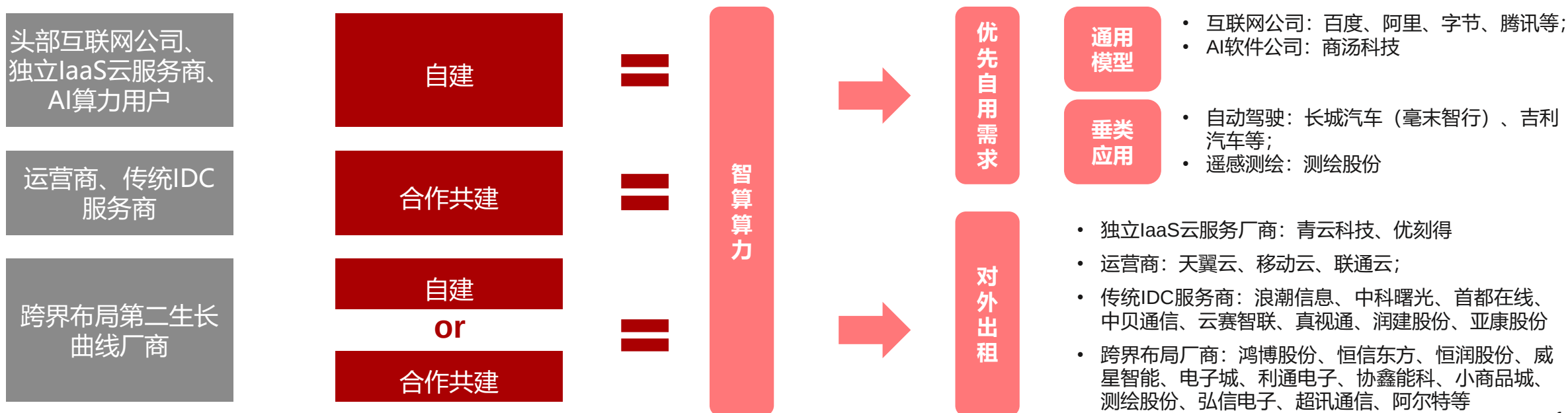
- **IaaS云服务厂商**：包括BBAT及三大运营商等云计算IaaS服务商，其中头部互联网厂商高端算力资源更充足
- **传统IDC服务厂商**：从传统IDC服务业务切入，在数据中心建设运营和能耗成本方面更具有优势
- **AI算力用户向上游扩张**：由AI算力需求方延伸为AI算力供给方，天然具备应用场景，AI算力运维能力可复用提升投入产出比
- **跨界布局第二生长曲线**：传统主业保持平稳或增长乏力，布局AI算力以期拉动业绩增长

表：布局AI算力上市公司分类

	代表厂商	优势	劣势
IaaS云服务厂商	• BBAT：字节、百度、腾讯、阿里 • 三大运营商：天翼云、移动云、联通云 • 独立IaaS厂商：优刻得、青云科技	• 头部互联网厂商高端GPU资源充足 • 云资源平台化管理与调度能力强 • 客户基础良好	• 部分厂商高端GPU资源不足 • 头部互联网厂商以满足自用为先，对外供应的智算算力的易用性、规模相对较少
传统IDC服务厂商	• IDC设备厂商：浪潮信息、中科曙光 • IDC服务：首都在线、中贝通信、云赛智联、真视通、润建股份、亚康股份	• 熟悉数据中心建设运营，数据中心机房资源充足 • 能耗成本管控方面更具有优势	• GPU设备调试与运维能力不足
AI算力用户向上游扩张	• AI软件公司：商汤科技 • 车厂：吉利汽车、长城汽车	• 熟悉AI模型框架 • 天然具备应用场景 • AI算力运维能力可复用提升投入产出比	• 以满足自用为先 • 内外兼修，资源管理难度增大
跨界布局第二生长曲线	鸿博股份、恒信东方、恒润股份、威星智能、电子城、利通电子、协鑫能科、小商品城、测绘股份、弘信电子、超讯通信等	资金实例充足	跨界转型具有一定壁垒，需从头积累行业knowhow

从建设方式来看，分为自建和共建；从经营方式来看，分为自用和出租

- 具备AI训练需求和AI应用场景的公司以自建AI算力中心为主，以头部互联网公司、AI算力用户为代表，且所建设的AI算力以满足自身业务需求为先，其次再为提供给外部客户使用。此类模式一方面可提升公司算力基础设施的利用效率，另一方面也可基于软、硬件实力构建生态圈，赋能合作伙伴。
- 主业涉及算力中心产业链条的公司以共建AI算力中心为主，以运营商、独立IaaS云服务商、传统IDC服务厂商为代表，合作对象主要为政府（或政府授权的公司主体）。此类模式受益于政策红利以及政府支持，区域属性较强。
- 跨界布局第二生长曲线的公司业务模式相对更加灵活，各类业务模式均有涉及，在共建模式下合作方也更加多元。



上市公司公告/公开披露信息统计（不完全统计）

上市公司	算力规模 (P)	总投资规模 (亿元)	售价	预计年收入 (亿元)	净利率	预计年净利润 (亿元)	项目内容
真视通		1					设立子公司，原主业数据中心业务升级为新一代智算中心和新一代绿色节能数据中心业务作为智算中心业务的承建及运营平台
恒润股份		1					计划于上海、福州经开区、安徽芜湖、山东济宁等地合作建立算力中心，规划的项目领域包括：（1）AIDC（人工智能计算中心）一站式服务；（2）新型计算中心建设与运维；（3）算力对外租赁
威星智能	2000	30(5+10+15)	3.75万元/P/月	9	66%	5.93	分三期建设2000P以上算力的智算中心，一期不少于370P算力。建成后可用于3D图形渲染和视觉设计、高性能科学计算、HDTV高速视频解编码、AI训练和推理等场景的智算需求。
中贝通信	500	2					提供不低于500P的算力运营服务能力
润建股份	172	2	T4:25万/台/年 云存储：120万/台/年	3.4	64%	2.20	最高2533Pops（Int8）定点算力或43Pflops(FP32)浮点算力及配套云存储，建成后主要提供AI大模型训练、推理算力、图形渲染算力服务，服务于人工智能大模型、行业模型等。
弘信电子		10					国产AI算力服务器智能制造基地：年产10万台，一期建设AI算力服务器工厂；二期建设AI算力服务器二期工厂、国产AI算力服务租赁基地
电子城		3.5					购置算力服务器及软件建设算力中心，构建以CPU+AI加速芯片为主体的计算集群，提供数据中心服务以及人工智能应用所需的算力服务
云赛智联		20					一是综合管理板块；二是技术研发板块，包括算力调度平台开发、平台运维等；三是项目运营板块，包括算力平台运维保障，商务合作和市场拓展、客户服务等
超讯通信		7.6			IRR 13.68%	1.05	宁淮绿色数字经济算力中心项目：含数据中心、云计算中心于一体的数字经济算力中心，预计可实现3,960个IDC机柜
首都在线		3.7			IRR 12.87%	0.54	渲染一体化智算平台项目:建设算力高质量供给、数据高效率流通的GPU算力资源池
利通电子		5					设立合资公司世纪利通，提供 AI算力租赁服务

上市公司公告/公开披露信息统计（不完全统计）

上市公司	算力规模 (P)	总投资规模 (亿元)	售价	预计年收入 (亿元)	净利率	预计年净利润 (亿元)	项目内容
协鑫能科		50					设立协鑫能源算力中心全球总部，计划2024年底前在全球建立15个能源算力中心
小商品城	50	18					国际数据中心项目一期规划机柜3000架，计算能力约12万VCPU,智算能力约50Pflops，互联网总带宽约2T和构建与北美、欧洲、东南亚、阿拉伯等国内外主流云服务商、数据中心的国际新型互联网数据专用通道。
恒信东方	55	0.5	18万/台/月	0.26	38%	0.10	配置30台服务器，11台训练（24.86P）+5台国产推理（昇腾310，560T）+16台NV推理（T4，200T）
测绘股份		1.5					购置办公楼及各项软硬件设备建设算力中心，在现有算力水平基础上增加算力节点
亚康股份	以CPU服务器为主	2.65					拟在庆阳、怀来、简阳、芜湖、贵安、韶关建设算力集群与节点支持服务站点
青云科技		1.79		0.9	10%	0.09	采购部分硬件设备自建超算和智算领域的公有云
鸿博股份	3000		A800: 16万/月/台; H800: 30万/月/台				北京AI创新赋能中心
百度	200	47.08					百度智能云-昆仑芯（盐城）智算中心 百度阳泉智算中心
阿里	12000 3000		13万/月/台				阿里云张北超级智算中心 阿里云乌兰察布智算中心
腾讯			12万/月/台				腾讯长三角人工智能先进计算中心 腾讯智慧产业长三角(合肥)智算中心
商汤科技	3740						商汤科技智能计算中心项目
吉利汽车	1000	10					吉利星睿智算中心
中国联通							
浪潮信息		2					克拉玛依浪潮智算中心
长城汽车							毫末智行：智算中心雪湖绿洲（MANA OASIS）
阿尔特							推进“阿尔特AI创新赋能中心”项目，探索人工智能+汽车设计研发解决方案开发

04

海外映射—— GPU云

短期来看，海外云厂商大量囤积英伟达A100/H100芯片；长期来看，头部厂商推进自研AI芯片

- 根据硅谷风投机构A16Z，生成式AI所产生总收入的10%~20%最终流向了云服务商
- 短期来看**，英伟达A100和H100芯片是大模型训练与推理的最佳选择，生成式AI爆发之后，云服务商对于英伟达GPU的采购进一步加大
- 长期来看**，头部厂商加速推进自研AI芯片计划，削减英伟达的“GPU税”。如谷歌、亚马逊、微软先后在内部启动自研AI芯片项目——谷歌的TPU系列，亚马逊的Inferentia和Trainium系列，以及微软的Athena芯片，但从通用计算数据中心到加速计算数据中心的过渡仍需要一定时间

表：云服务商拥有的英伟达H100 GPU数量

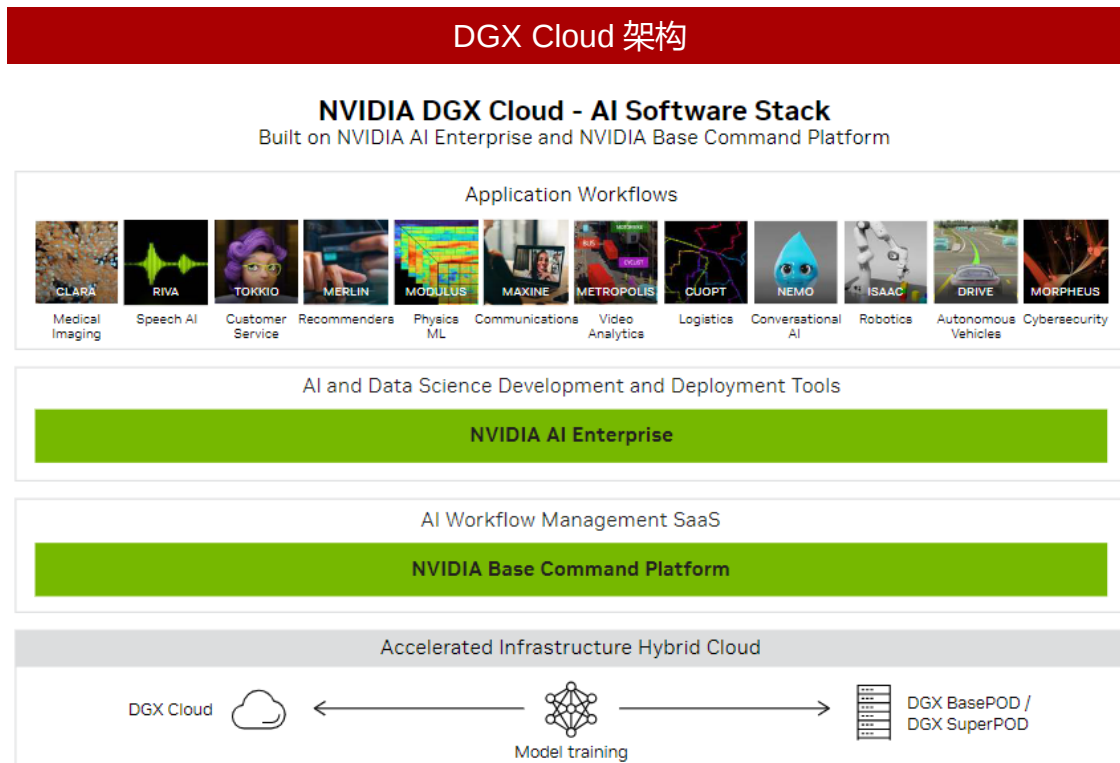
公司	英伟达H100数量
微软	8 (28.5万块A100)
Oracle	16000
亚马逊AWS	20000
谷歌GCP	26000
CoreWeave	35000 (预估)

表：海外云服务商GPU出租价格（截至2023.9.17）

公司	A100 40GB	A100 80GB	H100
Lambda Labs	/	/	/
FluidStack	\$1.79/hour	\$2.13/hour	/
Runpod	/	\$1.69/hour	\$3.99/hour
CoreWeave	\$2.06/hour	\$2.21/hour	\$4.76/hour
OCI	\$3.05/hour	\$4.00/hour	

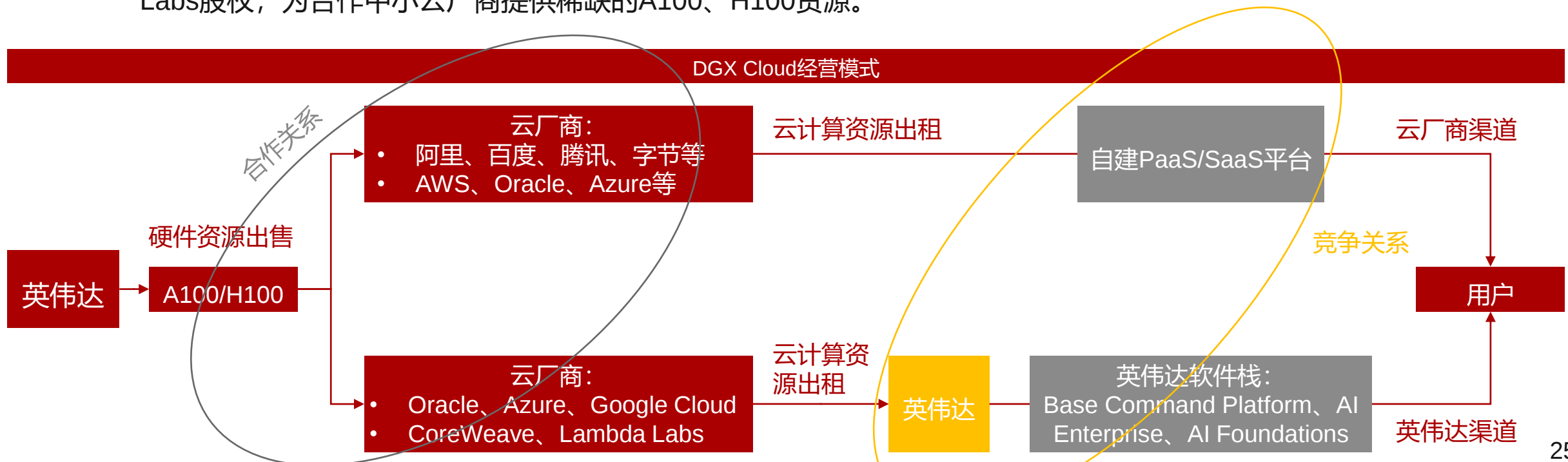
英伟达DGX Cloud是为客户打造的“软硬件一体及服务”，售价为每实例3.7万美元/月起

- DGX Cloud是2023年3月英伟达推出的一项人工智能超级计算服务，可以让企业快速访问为生成式人工智能和其他开创性应用训练高级模型所需的基础设施和软件，**价格为每实例3.7万美元/月起**
- 与传统购买英伟达AI服务器相比，DGX Cloud还提供丰富的软件栈服务，如Base Command Platform（基础命令平台）、AI Enterprise、AI Foundations等，可为客户提供全面的AI算力支持及解决方案
 - Base Command Platform（基础命令平台）是一个管理与监控软件，不仅可以用来记录云端算力的训练负载，提供跨云端和本地算力的整合，还能让用户直接从浏览器访问 DGX Cloud；
 - AI Enterprise是英伟达 AI 平台中的软件层，高达数千个软件包提供了各种预训练模型、AI 框架和加速库，从而简化端到端的 AI 开发和部署成本；
 - AI Foundations是模型铸造服务，让企业用户可以使用自己的专有数据定制属于自己的垂直大模型；



英伟达自身定位为算力服务商，采用托管模式经营，与传统云服务商既合作又竞争

- 英伟达自身定位于AI算力服务商，而非GPU厂商，我们认为这是在谷歌、亚马逊、微软等厂商先后在内部启动自研AI芯片项目的背景下，英伟达进行的前瞻性布局，但英伟达并未从零搭建云基础设施：
 - 一方面，英伟达将DGX Cloud托管在各家云服务厂商的云平台上，先将基础硬件设施出售给云厂商，再向他们购买云计算资源，最后把云服务出售给企业客户并自留全部收入，目前合作的云厂商包括Oracle、微软Azure和谷歌云；
 - 另一方面，英伟达扶持中小云厂商，2023年4月跟投中小云服务商CoreWeave，并拟收购另一家云服务商Lambda Labs股权，为合作中小云厂商提供稀缺的A100、H100资源。



持续布局AI软硬件，提升产品竞争力

- 从Bing Chat，到跨Microsoft 365应用程序组合的CoPilot内容创建体验，使用GitHub Copilot进行自然语言编码等等，现在这些大型语言模型都在Azure中运行。Azure OpenAI 服务提供对GTP-4、 GPT-3、 Codex 和 Embeddings 模型的访问权限。
- Microsoft Azure 和 NVIDIA 使云中的企业能够利用 NVIDIA 加速计算和 NVIDIA 按需网络的组合功能，以满足人工智能、机器学习、数据分析、图形、虚拟桌面和高性能计算（HPC）应用程序的各种计算要求。客户可在Azure上使用ND A100 v4 VM、NDm A100 v4 VM、NC A100 v4 VM、NV A10 v5 VM四类NVIDIA GPU虚拟机以满足不同情景下的需求。

微软 Azure AI+机器学习产品与服务

 Azure AI 服务 将高质量 AI 模型部署为 API	 Azure AI 文档智能 加速从文档中提取信息	 机器学习 构建和训练模型，并将其从云端部署到边缘	 Azure 认知搜索 适用于应用开发的企业级搜索
 Azure AI 机器人服务 为客户构建对话式 AI 体验	 Azure Databricks 使用基于 Apache Spark™ 的分析设计 AI	 Kinect DK 使用具有高级 AI 传感器的开发人员工具包生成计算机视觉和语音模型	 Azure OpenAI 服务 将高级编码语言模型应用于各种用例

Microsoft Azure 上的 NVIDIA GPU 加速虚拟机

ND A100 v4 VM Powered by eight NVIDIA A100 40GB Tensor Core GPUs and a dedicated NVIDIA Quantum InfiniBand 200 Gb/s connection per VM for scale-out, multi-node, multi-GPU distributed computing Learn More >	NDm A100 v4 VM Powered by eight NVIDIA A100 80GB Tensor Core GPUs and NVIDIA Quantum InfiniBand 200 Gb/s connection per VM for scale-out, multi-node, multi-GPU distributed computing Learn More >	NC A100 v4 VM Flexibility to select one, two, or four NVIDIA A100 80GB Tensor Core GPUs per VM to leverage the right-size GPU acceleration for your workload Learn More >	NV A10 v5 VM Flexibility to provision partial GPU partitions to two full NVIDIA A10 Tensor Core GPUs per VM. Powered by Microsoft Azure GPU-partitioning capabilities built on top of NVIDIA RTX Virtual Workstation technology Learn More >
---	--	---	--

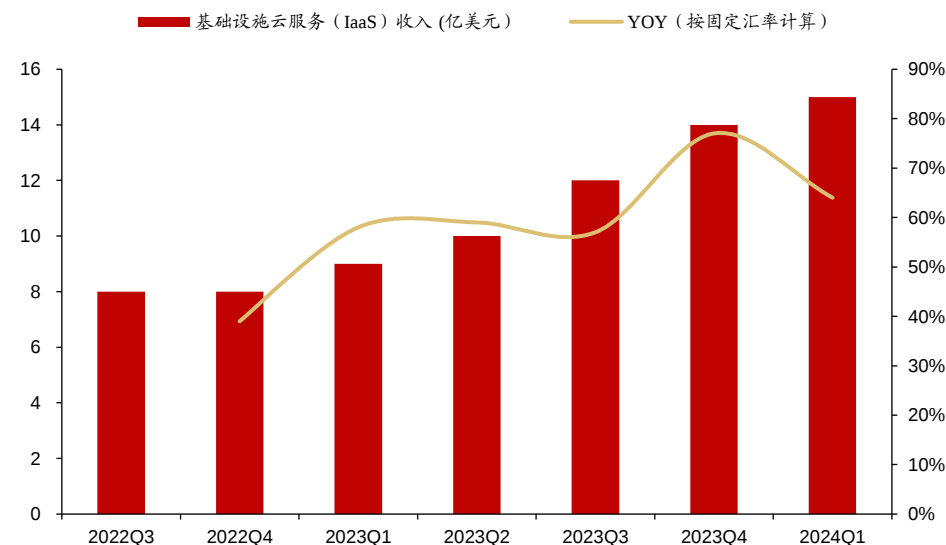
凭借AI算力布局加速赶超，云基础设施IaaS收入连续多个季度增长超过50%

- Oracle Cloud Infrastructure (OCI) 在一个全球云平台上提供 100 多个云服务和先进的行业特定 SaaS 应用。
- 甲骨文公司与英伟达 (NVIDIA) 合作持续加深：
 - 全新的 Oracle 云基础设施远程软件服务 (Oracle Cloud Infrastructure, OCI) Supercluster™ 上运行战略性 NVIDIA AI 应用
 - 英伟达选择 OCI 作为该企业的超大规模云技术提供商，提供大规模的AI超级计算服务 NVIDIA DGX Cloud™
 - 英伟达基于 OCI 的 DGX Cloud 提供生成式 AI 云服务 NVIDIA AI Foundations

OCI 上的 NVIDIA GPU 加速虚拟机及收费标准

Shape	GPUs	Architecture	GPU Interconnect	GPU Memory	CPU Cores	CPU Memory	Storage	Network	GPU Price Per Hour (US\$)	Server Price Per Hour (US\$)
Large scale-out AI training, data analytics, and HPC										
BM.GPU.A100-v2.8	8x NVIDIA A100 80GB Tensor Core	Ampere	NVIDIA NVLINK	640 GB	128	2,048 GB	4x 6.8TB NVMe	8x2x100 Gb/sec RDMA*	\$4.00	\$32
BM.GPU4.8	8x NVIDIA A100 40GB Tensor Core	Ampere	NVIDIA NVLINK	320 GB	64	2,048 GB	4x 6.8TB NVMe	8x2x100 Gb/sec RDMA*	\$3.05	\$24.4

甲骨文OCI云基础设施收入及增速（年份对应公司财年）



AI先行者，已形成较为完整的服务矩阵，CloudTPU+TensorFlow的软硬结合打造差异化优势

- 谷歌先后于2015年推出TensorFlow、2018年推出AutoML、2021年推出Vertex AI，在AI与机器学习领域深度布局模型、工具、平台等环节。在面向开发者的产品方面，目前Google Cloud平台已形成较为完整的服务矩阵，包含用于语音转文字的Speech-to-Text、用于图片识别的Vision AI以及用于视频识别的Video AI等；在AI解决方案方面，目前已有智能客服解决方案Contact Center AI以及商品搜索解决方案Discovery AI for Retail等产品。
- Compute Engine 以直通模式提供 NVIDIA GPU服务：**
 - Compute Engine 提供了可添加到虚拟机实例的图形处理单元 (GPU)，用户可以使用这些 GPU 加速虚拟机上的特定工作负载，例如机器学习和数据处理
 - NVIDIA AI Enterprise 现已在Google Cloud Marketplace上推出
- Cloud TPU加速 AI 开发：**
 - TPU 是 Google 专为神经网络设计的应用专用集成电路 (ASIC)，针对大型 AI 模型的训练和推断进行了优化，非常适合用于加快 AI 训练和推断速度
 - Cloud TPU可与谷歌的AI软件栈结合，加速AI研发进程

Google Cloud 上提供的可供选择GPU类型

对于计算工作负载，可在以下阶段使用 GPU 模型：

- NVIDIA L4: **正式版**
- NVIDIA A100
 - NVIDIA A100 40GB: **正式版**
 - NVIDIA A100 80GB: **正式版**
- NVIDIA T4: `nvidia-tesla-t4` (**正式版**)
- NVIDIA V100: `nvidia-tesla-v100` (**正式版**)
- NVIDIA P100: `nvidia-tesla-p100` (**正式版**)
- NVIDIA P4: `nvidia-tesla-p4` (**正式版**)
- NVIDIA K80: `nvidia-tesla-k80` : **已正式发布**。请参阅 [NVIDIA K80 服务终止 \(EOL\)](#)。

GPU云是算力租赁业务的长期进阶方向，具有更高的价值量和技术壁垒，市场想象空间更大

- 算力租赁业务的本质是AI算力固定资产变现，就其商业模式而言，可挖掘的增量价值空间有限：
 - 从收入端来看，AI算力的出租价格受到市场供需关系以及市场竞争的影响：供不应求时，AI算力租赁厂商具有较高的议价权；而当算力资源紧缺程度缓解之后，AI算力租赁厂商议价权减弱，存在租金下行的风险
 - 从成本端来看，给定算力租金水平和折旧年限，AI算力租赁的毛利率基本固定，可提升空间有限：由于AI算力租赁的成本由设备折旧摊销、数据中心能耗成本、人工运维成本构成，对于大部分成本AI算力租赁厂商处于被动接受的状态，议价能力弱
 - 基于以上，我们认为**算力租赁业务的利润规模量级基本由投资规模决定，增厚利润的最有效方式为增大投资，扩张算力规模**
- GPU云的本质是算力资源分配优化，同时提供AI软件开发相关的增值服务，壁垒高且易行程规模化优势：
 - 从收入端来看，给定算力规模和算力租金水平的情况下，算力的调度和优化能力可进一步增加GPU云厂商的收入天花板；同时，围绕AI软件开发相关的增值服务（PaaS层或SaaS层），可为GPU云厂商额外贡献增量收入，且收入天花板不受到算力规模的限制
 - 从成本端来看，算力调度与软件增值服务的研发投入体现在费用端，随着对应营收规模的增长，盈利能力有望持续提升
 - 基于以上，我们认为**GPU云相较于算力租赁业务而言具有更高的技术壁垒以及成长性，我们看好AI算力租赁厂商向GPU云的迭代转型**

05

风险提示

- 1、报告中对于商业模式的总结以及对于盈利模型的测算包含部分主观假设
- 2、AI算力租赁业务落地不及预期
- 3、上游AI服务器供应持续短缺，或受到美国商务部禁令影响，中国厂商无法获得英伟达相关服务器设备
- 4、AI算力产业政策发生重大变化
- 5、布局AI算力租赁业务厂商持续增加，市场竞争风险加剧

行业的投资评级

以报告日后的6个月内，行业指数相对于沪深300指数的涨跌幅为标准，定义如下：

- 1、看好：行业指数相对于沪深300指数表现 + 10%以上；
- 2、中性：行业指数相对于沪深300指数表现 - 10% ~ + 10%以上；
- 3、看淡：行业指数相对于沪深300指数表现 - 10%以下。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。

建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。

本公司的客户作参考之用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本报告仅反映报告作者的出具日的观点和判断，在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

本报告版权均归本公司所有，未经本公司事先书面授权，任何机构或个人不得以任何形式复制、发布、传播本报告的全部或部分内容。经授权刊载、转发本报告或者摘要的，应当注明本报告发布人和发布日期，并提示使用本报告的风险。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

浙商证券研究所

上海总部地址：杨高南路729号陆家嘴世纪金融广场1号楼25层

北京地址：北京市东城区朝阳门北大街8号富华大厦E座4层

深圳地址：广东省深圳市福田区广电金融中心33层

邮政编码：200127

电话：(8621)80108518

传真：(8621)80106010

浙商证券研究所：<http://research.stocke.com.cn>