证 券 研 究 报 告

AI云计算新范式：规模效应+AI Infra+ASIC芯片

——GenAI系列报告之五十四

2025.03.28

重点报告回顾

- 我们近期已发布多篇深度报告，围绕重点标的AI布局及进展，从底层硬件至上层应用进行全方位梳理：

1. **腾讯AI详细梳理：**《腾讯控股（00700）点评：AI应用+云业务有望迎来价值重估》
2. **阿里云深度：**《阿里巴巴-W（09988）深度：AI开启阿里云新成长（阿里巴巴深度之三暨GenAI系列报告之39）》
3. **字节AI详细梳理：**《豆包大模型升级，字节AI产业链梳理—— GenAI之四十四》
4. **金山云深度：**《金山小米生态核心云厂，AI+智驾乘风而上》
5. **美股云行业季度总结：**《云厂Capex指引仍乐观，AI应用ROI路线清晰或将迎来催化——美股云计算和互联网巨头24Q4总结》、《北美云厂Capex加速，AI降本增效初步体现—— 美股云计算和互联网巨头24Q3总结》
6. **谷歌深度：**《谷歌：AI征途换档提速，云业务驱动成长》
7. **META深度：**《Meta Platforms （META）：广告推荐应用+开源模型+算力，AI布局解析》
8. **博通深度：**《博通：软硬一体的AI卖铲人》
9. **AI应用深度：**2024年总结-《AI应用：商业化初露锋芒——AI应用深度之二暨GenAI系列报告之三十九》、2023年总结-《AI应用：从生产力工具到交互体验升级——生成式AI2024年投资策略》

核心观点：

- **AI云计算新范式：规模效应+AI Infra能力+算力自主化。**云计算在AI收入拉动下营收增速回暖、Capex增长加速已成为市场共识。（详见此前相关报告总结。）但对于AI云时代竞争格局以及云厂利润率还有分歧，也是本报告的重点。**1）更强的规模效应；2）AI infra能力；3）算力自主化为云厂中长期降本方向。**
- **规模效应：更高的初始投入，更高的算力利用率。（1）AI云更高的资本密集度。（2）AI服务器/网络设备使用年限更短、成本占比明显提升。**多租户+多场景（含自有场景）+自有模型平抑需求峰谷，降低产能空置率、摊薄单位计算成本，实现更高的ROI。以腾讯、阿里、谷歌等为代表的大型云厂商/互联网巨头具备庞大的内部工作负载禀赋+AI大模型的优势，有望降低单位计算成本。
- **AI Infra：实现计算性能挖潜。**AI Infra定位于算力与应用之间的“桥梁”角色的基础软件设施层，体现在：**1）硬件集群的组网构建、算力调度系统；2）大模型+AI开发工具，增强大模型对于算力计算效率的挖潜；3）针对应用的定向优化等工作。**尽管模型开源，但针对特定模型推理的优化能力、AI工具丰富度差异仍会放大云厂对同一开源模型优化后的推理成本差距。以谷歌、字节火山引擎、阿里云、DeepSeek等为代表的厂商已在AI Infra领域发布训练/推理侧工具。
- **算力自主化：海外ASIC芯片趋势启示。强大的工程能力或有望弥补ASIC和GPU硬件生态差距。**ASIC架构：基于脉动阵列的定制架构为重要路线；ASIC开发生态：谷歌和AWS均基于XLA，Meta MTIA v2软件堆栈基于Triton。ASIC芯片的确定性来自：（1）供给端，芯片设计制造专业分工：降低ASIC与GPU在代工制造、后端封装设计上的差距，ASIC辅助设计博通、迈威尔等崛起。（2）需求端：牧本摆动，有望从标准化到定制化：架构创新，催生新的定制化芯片，并再度基于新的芯片进行算法创新升级，以实现芯片性价比优势；商业上可行：具备庞大算力需求的云厂可覆盖开发定制化芯片的成本。ASIC制造模式：云厂前端设计+IC辅助设计支持。
- **推荐（1）互联网云计算：腾讯控股，阿里巴巴，金山云；谷歌、微软、META、亚马逊；（2）ASIC辅助设计：博通。**
- **风险提示：内容和互联网平台监管环境变化风险；大模型性能进步不及预期；AI应用落地进展不及预期风险**

主要内容

1. AI云计算新范式：规模效应+AI Infra能力+算力自主化
2. 规模效应：资本密集度+多租户+内部负载的削峰填谷
3. AI Infra：实现计算性能挖潜
4. 算力自主化：海外ASIC芯片趋势启示
5. 重点标的：互联网云厂+ASIC芯片
6. 重点公司估值表及风险提示

1.1 云计算：计算资源公共化，AI云聚焦于AI算力+工具

- **云计算是将计算资源变成可租用的公共服务，**强调集中管理和动态分配虚拟化计算资源，以按需自助服务、弹性扩展和按使用量计费为核心特征的标准化服务模式，实现相对企业自建数据中心的性价比优势。
- **传统云计算指基于CPU服务器，主要为传统工作负载提供支持。AI云的区别在于，硬件平台基于GPU服务器，主要提供包括MaaS层在内的各环节AI工具及服务。**

图：云计算按服务方式的分层

1.1 云计算：AI时代云需求明确提升，重点关注未来竞争

- **AI对于算力基础设施的需求明确提升，**各云厂在AI云收入拉动下营收增速回暖、Capex将增长加速已成为市场共识。
- **本报告则旨在聚焦于未来的AI云竞争，在规模效应、AI Infra能力、算力自主化三大层面讨论AI云竞争格局变化和未来利润率趋势。**

表：国内及海外主要云厂商营收增速回暖（单位：美股标的为亿美元，其他标的为亿人民币）

公司	2023年云收入	2023年YoY	云收入占比	2024年云收入	2024年YoY	云收入占比	云经营利润率
亚马逊	908	13%	16%	1,076	19%	17%	37%
微软智能云	797		35%	956	20%	37%	40%以上
谷歌	331	26%	11%	432	31%	17%	14%
阿里巴巴	994	2%	11%	1,135	8%	12%	9%
金山云	70	-14%	100%	78	10%	100%	-6%
中国移动	833	66%	8%	1,004	20%	10%	
中国联通	510	42%	14%	686	17%	18%	
中国电信	972	68%	19%	1,139	17%	22%	

表：国内及海外主要云厂商Capex同比增速大幅提升

公司	23Q3	23Q4	24Q1	24Q2	24Q3	24Q4
微软	70%	69%	79%	78%	79%	97%
亚马逊	-24%	-12%	5%	54%	81%	91%
Meta	-30%	-15%	-2%	36%	41%	94%
谷歌	11%	45%	91%	91%	62%	30%
阿里巴巴	-57%	28%	221%	75%	240%	259%
腾讯控股	237%	33%	226%	121%	114%	386%
百度	61%	90%	57%	-22%	-53%	-36%

1.2 AI云新范式：更多竞争要素，看好互联网云/大型云

- **对于云计算而言，云服务工具/资源的丰富度、计算资源的利用率为云厂商盈利核心。**
- 相对传统云，**AI云计算出现新范式：**云技术重新进入快速迭代阶段、资本更为密集，对云厂商的**资本密集度、产能利用率、云基础设施能力、工具和生态的丰富度、自研芯片布局**等维度均提出新要求。
- AI云实现盈利的门槛将进一步提升，看好拥有技术能力、云多租户、内部负载规模效应的互联网云/大型云。

图：AI云相对传统云，在资本开支、产能利用率、技术能力等方面提出更高要求

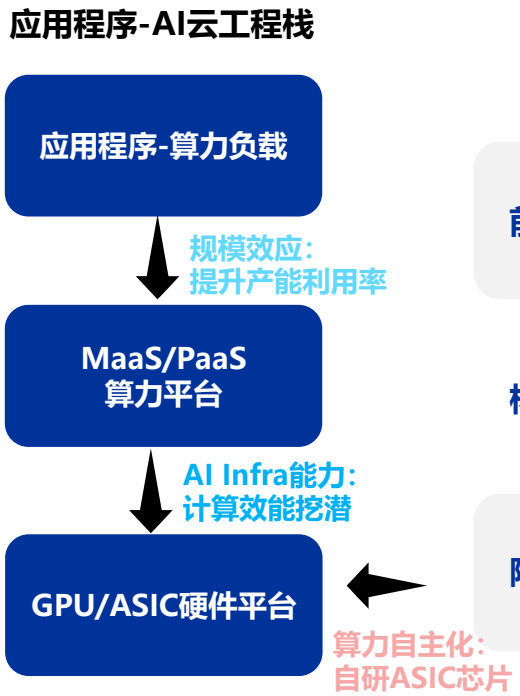
核心竞争要素	传统云计算	AI云计算
规模效应-资本密集度	资本密集型，以CPU服务器为主 2023年海外大型云/互联网巨头Capex总和为1534亿美元	以GPU服务器为主，单集群**Capex更庞大** 2024年海外大型云/互联网巨头Capex总和为2504亿美元
规模效应-产能利用率	IT设备成本占比约为40%-50%，基建、运维占比相对高，对产能空置有一定容忍度	IT设备成本占比约为60%-70%，折旧压力更大,对产能空置容忍度很低，**拥有内部负载/自研大模型/云多租户十分重要**
AI Infra-云基础设施能力	传统云基础设施强调稳定性及服务，虚拟化、容器等技术已趋于成熟	AI云的基础设施能力仍在快速迭代，**且各厂商能力差异大**，体现在大模型推理成本等领域
AI Infra-工具/生态丰富度	主要提供数据分析工具、Devops平台等服务，工具丰富度	除传统工具外，竞争格局还**受到云厂商提供的AI工具丰富度影响**
算力自主化自研ASIC芯片	主要云厂商亚马逊、阿里巴巴、谷歌等均在自研CPU等芯片	主要云厂商谷歌、亚马逊、阿里巴巴等均在针对特定工作负载**自研AI ASIC芯片**

资料：申万宏源研究
注：本表内海外大型云/互联网巨头包括微软、亚马逊、谷歌、META

1.2 AI云ROI：更强的规模效应、AI Infra能力、算力自主化

- **AI云利润率将由三大方向影响，不同能力、规模间的AI云利润率或将拉开较为明显的差距。**
- **1）需求侧-规模效应提升算力利用率：**增加工作负载保证集群满负载、实现算力需求削峰填谷；
- **2）供给侧-AI Infra能力提升硬件计算效能：**对应用程序/大模型至硬件间的组网、软件算法进行优化；
- **3）长期供给侧-算力自主化降低硬件成本：**中长期维度降本途径。

图：AI云的ROI主要由规模效应、AI Infra优化、算力自主化带来



	规模效应	AI Infra能力	算力自主化
前提条件	软件技术、业务运营导向	**软硬件技术**、研发导向	**硬件技术**、研发导向
核心因素	➢ **自研/投资大模型** ➢ **云多租户需求量** ➢ **庞大而稳定的AI内部工作负载**	➢ **AI Infra工程能力**	➢ ASIC芯片设计能力 ➢ 开发生态构建能力
降本方式	**提升产能利用率：削峰填谷，平稳地工作负载，摊薄折旧成本**	**提升计算效能，提升同等芯片在单位时间内可完成的训练/推理任务量**	降低硬件采购成本，提升单位资本开支可获取的算力

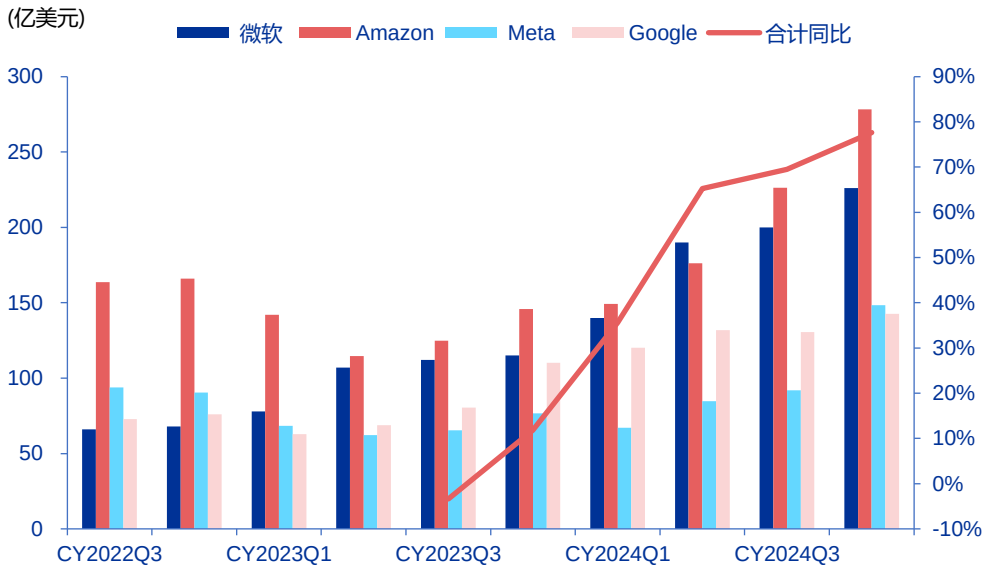
主要内容

1. AI云计算新范式：规模效应+AI Infra能力+算力自主化
2. 规模效应：资本密集度+多租户+内部负载的削峰填谷
3. AI Infra：实现计算性能挖潜
4. 算力自主化：海外ASIC芯片趋势启示
5. 重点标的：互联网云厂+ASIC芯片
6. 重点公司估值表及风险提示

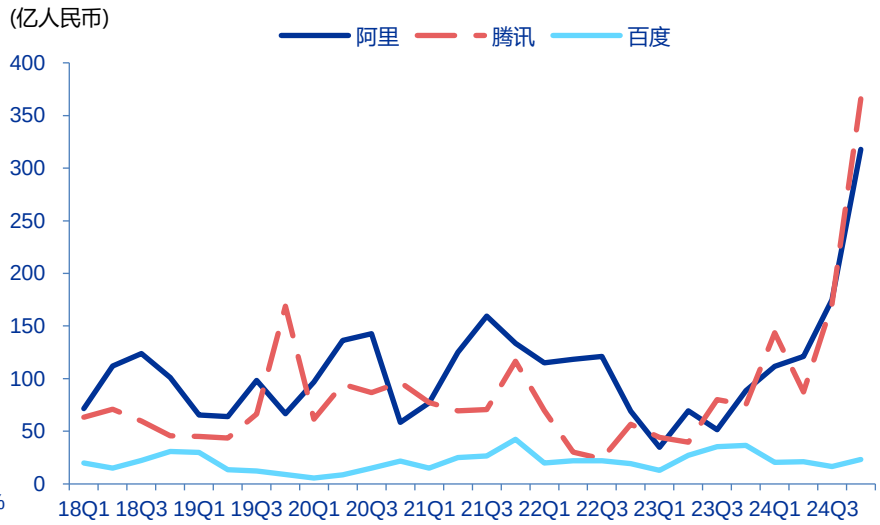
2.1 资本密集度：构建AI云集群的支出量级仍在不断扩大

- 海外：根据各企业指引，2024年谷歌、微软、亚马逊、META的Capex总计2504亿美元；若假设2025年（即FY25Q3-FY26Q2）微软保持FY25Q2的资本开支水平，**则四家巨头的Capex预计将接近3400亿美元，同比增速有望达到35%。**随着各家Capex已达到较高基数水平，预计26年增速或有所放缓。
- 国内：阿里巴巴指引25-27年资本开支将达到3800亿元，年均将接近1300亿元；腾讯指引Capex将占营收的低两位数百分比（Low Teens）。

图：海外主要互联网云巨头资本开支快速增长



图：国内主要互联网云巨头资本开支快速增长



资料 ：谷歌、微软、亚马逊、META财报，Bloomberg，申万宏源研究

2.1 资本密集度：AI视频/Agent到来将提升算力需求量级

- **AI应用即将走向AI Agent、视频、3D等模态，对算力的消耗量级将进一步提升：**文字交互的推理单次请求目前仅为数百Tokens的计算量，但AI Agent的复杂任务规划、多步推理，以及视频和3D工具的单次推理，消耗Tokens的量级将相对文字交互明确提升。
- **此外，AI有望拉动国内企业上云需求，进一步带动云计算Capex提升。**

表：图片/视频生成及AI Agent预计将带来更高量级算力需求

功能	模型	价格	具体消耗
文字对话	谷歌 Gemini 2.0 Flash	输入：0.1美元/百万Tokens；输出：0.4美元/百万Tokens	4字符/Token，100Tokens大约相当于60-80英文单词，每轮对话生成300个单词，则消耗大约500Tokens
图片生成	谷歌 Imagen3	生成图片：0.04美元/图片	按同等价格算约等同于10万Tokens文字输出算力
视频生成	谷歌 Veo2	生成视频：0.5美元/s	8s视频价格为4美元，按同等价格算约等同于1000万Tokens文字输出算力
AI Agent	基于基础大模型	参照文字对话消耗	越复杂的任务需要的大模型推理步数更多。AI Agent完成某一简单代码开发需要约20步，则算力消耗为单步推理的20倍以上（多步推理还需考虑状态维持开销、动态规划损耗等算力消耗），复杂代码开发则需要更多推理步数。
3D模型生成	Meshy	生成模型+纹理：0.4美元/个	按同等价格算，约等同于100万Tokens文字输出算力

资料：谷歌、Meshy官网，申万宏源研究

2.2 产能利用率：AI云IT设备折旧压力大，空置容忍度更低

- **对比传统云计算，AI云厂将面临更大的折旧压力，利润率将对产能利用率更为敏感，将形成更强规模效应。**
- **1）AI云的IT设备在建设成本的占比提升：**AI服务器+网络设备折旧周期更短，通常折旧年限在5-6年，而基础设施折旧年限通常超过15年；短折旧项占比更高，AI云厂面临更大的折旧压力。
- **2）AI服务器实际折旧周期更短：**不同于发展成熟的CPU，GPU/ASIC仍处于高速更新迭代阶段，可能加速折旧。以亚马逊FY24Q4财报为例，重新将部分IT设备折旧年限从6年缩短至5年。

表：折旧期限更短的IT设备在自建AIDC成本占比重中更高，产能空置的容忍度大幅降低

	典型传统数据中心建设成本占比	典型AI数据中心建设成本占比
基础设施	30%-40%	25%-35%
IT设备	40%-50%	60%-70%
服务器/IT设备：	*60%-70%*	*80%-90%*
存储及网络/IT设备：	*30%-40%*	*10%-20%*
运维及人工	10%-20%	5%-10%

表：FY24Q4亚马逊缩短部分服务器及网络设备折旧年限至5年，季度折旧摊销成本环比加速增加

单位：百万美元	3Q22A	4Q22A	1Q23A	2Q23A	3Q23A	4Q23A	1Q24A	2Q24A	3Q24A	4Q24A
亚马逊	10327	12081	11123	11589	12131	13114	11684	12038	13442	15631
QoQ		17.0%	-7.9%	4.2%	4.7%	8.1%	-10.9%	3.0%	11.7%	16.3%
谷歌	3933	3602	2635	2824	3171	3316	3413	3708	3,985	4205
QoQ		-8.4%	-26.8%	7.2%	12.3%	4.6%	2.9%	8.6%	7.5%	5.5%
微软	2790	3648	3549	3874	3921	5959	6027	6380	7383	6827
QoQ		30.8%	-2.7%	9.2%	1.2%	52.0%	1.1%	5.9%	15.7%	-7.5%
META	2130	2329	2524	2623	2858	3134	3374	3637	4027	4460
QoQ		9.3%	8.4%	3.9%	9.0%	9.7%	7.7%	7.8%	10.7%	10.8%

资料：谷歌、微软、亚马逊、META财报，Bloomberg，申万宏源研究
注：微软的折旧中包含收购暴雪动视后的无形资产折旧等

2.2 产能利用率：短期GPU供不应求利润率向好，供需平衡后产能利用率影响将凸显

- **AI云计算需求供不应求，拉动云厂营业利润率自23Q3后明确回暖。**H100等GPU租赁价格保持在较高水平，为核心云厂带来了较为丰厚的投资回报率；此外北美云厂叠加北美宏观经济从23Q3后从悲观预期中逐渐修复。
- 尽管当前云厂营业利润率对折旧成本抬升仍不敏感，**但仍需关注，随着台积电COWOS产能逐渐释放，GPU将从紧缺逐渐转向平衡，GPU租赁价格或有所回落，届时云厂AI算力产能利用率对利润率影响将更明确体现。**

表： AI算力供不应求+需求回暖，主要云厂利润率持续提升后仍保持较高水平

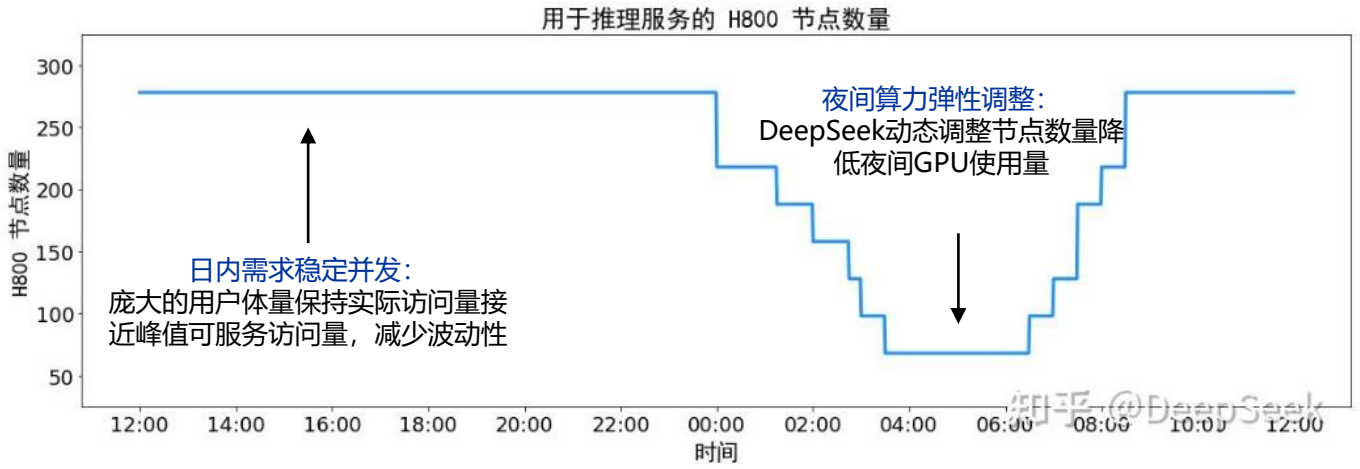
单位：亿美元		CY23Q1	CY23Q2	CY23Q3	CY23Q4	CY24Q1	CY24Q2	CY24Q3	CY24Q4
谷歌云	营收	74.54	80.31	84.11	91.92	95.74	103.47	113.53	119.55
	同比增速	28.1%	28.0%	22.5%	25.7%	28.4%	28.8%	35.0%	30.1%
	营业利润率	2.6%	4.9%	3.2%	9.4%	9.4%	11.3%	17.1%	17.5%
亚马逊AWS	营收	213.54	221.40	230.59	242.04	250.37	262.81	274.52	287.86
	同比增速	15.8%	12.2%	12.3%	13.2%	17.2%	18.7%	19.1%	18.9%
	营业利润率	24.0%	24.2%	30.3%	29.6%	37.6%	35.5%	38.1%	36.9%
Azure	营收增速			30.0%	31.0%	35.0%	35.0%	34.0%	31.0%
微软智能云	营收	182.44	198.89	200.13	215.25	221.41	237.85	240.92	255.44
	同比增速			18.5%	20.1%	21.4%	19.6%	20.4%	18.7%
	营业利润率			44.5%				43.6%	42.5%
阿里云	营收（亿人民币）	185.82	251.23	276.48	280.66	255.95	265.49	296.10	317.42
	营收YoY	-2.1%	4.1%	2.3%	2.6%	3.4%	5.9%	7.1%	13.1%
	EBITA Margin	2.1%	1.5%	5.1%	8.4%	5.6%	8.8%	9.0%	9.9%

资料　　：谷歌、微软、亚马逊、阿里巴巴财报，Bloomberg，申万宏源研究
注：微软智能云业务在Azure之外还包括服务器产品等其他业务

2.3 如何实现规模效应？多租户+内部负载均衡算力需求

- **对于大模型/云厂商而言，应用访问需求在日内呈现明显周期性和波动性：1）日间算力需求高峰期：**尽可能实现访问请求量相对稳定**减少波动性，**避免峰值需求过高偏离可服务量，拥有**云多租户/大规模用户的AI应用**至关重要。**2）夜间算力需求低谷期：**尽可能增加**时效性要求偏低的任务负载**，平抑需求**周期性。**

图：DeepSeek应用推理节点数量按需弹性变化，日间需求平稳并跑满产能，夜间实现弹性调整



2.3 如何实现规模效应？多租户+内部负载均衡算力需求

- **云多租户/大规模AI应用平抑波动性：**以互联网云为代表的云厂，对AI布局较早并已吸引众多AI初创公司客户，旗下拥有用户规模较大的AI应用（豆包、腾讯元宝）以及内部AI负载，可实现日内需求的稳定性。
- **内部负载调度均衡平抑周期性：**互联网云厂拥有较为旺盛的非实时算力需求，包括大模型/多模态工具/推荐系统的训练迭代需求、数据分析处理需求等，可以运行于算力需求低谷期，可平抑需求的周期性。

表：多租户/应用+非实时内部负载将帮助AI云算力实现削峰填谷

	整体需求	日间需求波动	夜间需求填补
对AI云的要求	较长时间维度内对客户需求的准确估算	**拥有云多租户、大规模AI应用**	**拥有云多租户、自有业务的非实时AI算力需求**
提升产能利用率方式	根据云客户或自身需求设计集群规模，减少因租户不足而带来的产能空置	实际满足算力需求的大数定律，拥有云多租户、应用用户数量大的AI应用，可以保持在大部分时间段的负载相对稳定，而租户、应用用户少的情况下更可能出现的需求波动性，导致算力空载。	由于夜间推理访问量较少， 1）可运行时效性要求较低的AI工作负载，包括模型训练、离线推理、推荐系统训练等，填补夜间算力空闲时间。 2）可通过大幅降价吸引云租户业绩运行工作负载。

2.4 互联网云：闭源大模型将影响云竞争格局、算力需求量

- **闭源模型仍为主要模式，云厂商可通过自研大模型+投资大模型厂商形成模型独占，获取更大市场份额，增加云客户数量、提升对于云厂的算力需求量。**海外TOP3闭源厂商（OpenAI-微软+甲骨文、谷歌、Anthropic-亚马逊）+以阿里为代表国内大模型云厂。
- **但开源模型亦逐渐走向繁荣，一定程度上缩小大模型能力差距对云厂竞争格局的影响力。**DeepSeek接力META的Llama系列大模型，领导开源生态逐渐走向繁荣，此外阿里、谷歌等厂商也开源部分模型构建开发者生态，预计闭源与开源两大路径将共存。

表：主要大模型性能排名

Arena Score排名	模型	Arena分数	模型厂商	是否开源
1	Grok-3-Preview-02-24	1412	xAI	闭源
2	GPT-4.5-Preview	1411	OpenAI	闭源
3	chocolate (Early Grok-3)	1402	xAI	闭源
4	Gemini-2.0-Flash-Thinking-Exp-01-21	1384	谷歌	闭源
5	Gemini-2.0-Pro-Exp-02-05	1380	谷歌	闭源
6	ChatGPT-4o-latest (2025-01-29)	1377	OpenAI	闭源
7	DeepSeek-R1	1363	**DeepSeek**	开源
8	Gemini-2.0-Flash-001	1357	谷歌	闭源
9	o1-2024-12-17	1352	OpenAI	闭源
10	Qwen2.5-Max	1336	**阿里巴巴**	闭源
13	DeepSeek-V3	1318	**DeepSeek**	开源
14	GLM-4-Plus-0111	1311	**智谱AI**	闭源
16	Claude 3.7 Sonnet	1309	Anthropic	闭源
18	Step-2-16K-Exp	1305	**阶跃星辰**	闭源
28	Hunyuan-Large-2025-02-10	1271	**腾讯**	闭源
34	Meta-Llama-3.1-405B-Instruct-bf16	1269	Meta	开源

资料：Hugging Face，申万宏源研究
注：时间截止2025年3月，并未包括字节跳动与百度等公司的大模型

2.4 互联网云：庞大的工作负载+潜在AI应用将摊薄成本

- **互联网云公司拥有庞大的可迁移至AI芯片的内部工作负载，以META为例，2022年开始将推荐系统负载转移至GPU服务器上，此外搜索引擎、大模型训练推理、潜在爆款AI应用均可运行于AI芯片，具备规模效应。**
- **内部负载/全球性应用可调节算力芯片工作峰谷。**1）任务调整：将时效性要求更低的负载（例如大模型/推荐系统训练迭代、数据分析处理）用于闲时。2）全球布局的企业，日间与夜间工作负载的时差可以被平抑。

表：国内互联网云厂商拥有庞大工作负载，可有效摊薄成本

	AI芯片布局	大模型及AI开发框架	已推出的核心AI应用	可在AI芯片上运行的内部工作负载
字节跳动	外购：根据Omdia，2024年公司购买了23万片H100	➢ 大模型：豆包；多模态BuboGPT ➢ 开发平台：Coze AI平台	➢ AI视频工具：即梦 ➢ AI Chatbot：豆包 ➢ AI Agent平台：小悟空	➢ 云计算：火山引擎 ➢ **推荐系统：应用矩阵抖音、TikTok、剪映、今日头条等的AI推荐算法**
阿里巴巴	外购：采购英伟达芯片 自研AI芯片：12nm 含光800（推理）等 自研CPU：倚天系列	➢ 大模型：24年5月发布通义千问2.5 ➢ 开发平台：百炼AI平台	➢ AI Chatbot：通义 ➢ 电商助手：淘宝问问（ToC）、AI生意助手（ToB） ➢ 开源大模型社区：魔塔社区	➢ 云计算：阿里云 ➢ **推荐系统：电商平台淘宝、阿里国际站等的AI推荐算法** ➢ **AI助手：承担Apple Intelligence的大模型/算力支持**
腾讯	外购：根据Omdia，2024年公司购买了23万片H100 自研AI芯片：紫霄（推理）等	➢ 大模型：24年11月推出Huanyuan large 389B MoE开源模型 ➢ 开发平台：腾讯云AI平台	➢ AI Chatbot：混元助手、腾讯元宝 ➢ AI视频平台：腾讯智影 ➢ AI Agent平台：腾讯元器 ➢ AI笔记：Ima copilot	➢ 云计算：腾讯云 ➢ **推荐系统：微信视频号、腾讯视频等的AI推荐算法** ➢ **搜索引擎：微信搜一搜的AI搜索算法**
百度	外购：采购英伟达芯片 自研AI芯片：7nm 昆仑芯二代	➢ 大模型：24年6月发布文心4.0 Turbo ➢ 深度学习框架：飞桨 ➢ 开发平台：千帆	➢ AI搜索：百度AI智能问答 ➢ AI Chatbot：文心一言 ➢ AI Agent平台：文心智能体 ➢ 自动驾驶：萝卜快跑	➢ 云计算：百度云 ➢ **搜索引擎：百度搜索的AI搜索算法** ➢ **推荐系统：应用矩阵百度地图、[redacted]等的AI推荐算法**

资料：字节跳动、阿里巴巴、腾讯、百度官网及财报，Omdia，申万宏源研究

2.4 互联网云：庞大的工作负载+潜在AI应用将摊薄成本

表：海外互联网巨头/大型云厂商拥有多租户/庞大内部工作负载，可有效摊薄成本

	AI芯片布局	大模型及开发框架	AI研发布局模式	已推出的核心AI应用	现有业务生态协同
微软	外购：根据Omdia，24年购买约48.5万张H100芯片 自研：2023年11月发布Maia 100芯片	大模型：OpenAI推出GPT系列模型，2023年3月推出GPT-4，24年5月推出GPT-4o，24年9月推出GPT-o1 开发平台：Azure AI Studio，包括GPT系列独家模型及第三方大模型	大比例持股体外公司+深度合作。2023年向OpenAI投资100亿美元，为OpenAI主要的算力提供商 自研：招揽Inflection AI核心团队，布局大模型	办公：推出Microsoft 365 Copilot CRM/ERP：推出Dynamic 365 copilot 编程工具：Github Copilot 搜索引擎：必应集成ChatGPT	云计算：Microsoft Azure 办公软件：Microsoft 365、Office 操作系统：Windows 浏览器：Edge 搜索引擎：Bing
谷歌	外购：根据Omdia，24年购买约16.9万张H100； 自研：2016年推出第一代TPU，TPU v6 Trilium已上线谷歌云，性能出色。TPU芯片可基本支撑自研大模型的训练和推理 通信：自研OCS通信系统，通信性能出色	大模型：2023年12月推出首个多模态大模型Gemini，24年底开始发布Gemini 2.0系列 深度学习框架：TensorFlow（两大主流框架之一）、JAX 开发平台：Vertex AI	旗下部门自研：此前有Google Brain、Deepmind等多个AI研发部门/全资子公司，分立运营；2023年4月起整合为单一AI研发部门Google Deepmind	办公：推出Duet AI，定价30美元/月 搜索：AI搜索功能AI Overview，至24年10月，已覆盖10亿用户 应用：NotebookLM 其他：编程工具Alphacode等	云计算：Google Cloud 办公软件：Workspace 操作系统：安卓 浏览器：Chrome **搜索引擎：Google** 应用矩阵：谷歌地图、Youtube、Play store、Gmail
Meta	外购：根据Omdia，2024年购买约22.4万张H100芯片；计划在25年底拥有130万块GPU 自研：2024年发布MTIA v2芯片，陆续应用于推荐系统等的推理负载中，26年将应用于训练及推理负载	大模型（开源）：2023年7月开源Llama2，2024年推出Llama3，Llama4正在10万卡集群上训练，Llama4 mini已完成训练 深度学习框架：Pytorch（两大主流框架之一）	旗下部门自研：AI业务均由旗下AI部门进行研发，为直属部门模式	AI推荐系统升级：截至24年10月，AI全年已提升Facebook /Ins使用时长8%/6% META AI助手：已集成于社交软件中，至24Q4 MAU超7亿 广告创意及投放：推出辅助广告内容生成工具、AI广告投放工具	社交应用：Facebook、Instagram等 元宇宙：旗下VR设备品牌Quest以及内容平台
亚马逊	外购：根据Omdia，2024年购买约19.6万张H100 自研：2020年推出Trainium，23年推出Trainium2，Rainier项目正构建数十万卡Tranium2集群；Tranium3将于25年底发布	自研大模型：2023年12月推出Titan系列AI模型 大模型（Anthropic）：24年开始持续更新Claude3.5系列 开发平台：Bedrock AI搭载自研及第三方模型	旗下部门自研+持股重点公司：旗下AI部门完成自研大模型研发；重点投资Anthropic，2023-24年投资80亿美元，并提供算力支持；谷歌也参与Anthropic多轮投资	电商：为电商运营提供一系列AI功能支持，以及导购助手Rufus； 生成式助手：面向企业端的AmazonQ； 广告：辅助广告内容生成工具；通过AI实现广告智能投放提升效率	云计算：AWS 电商平台：亚马逊商城

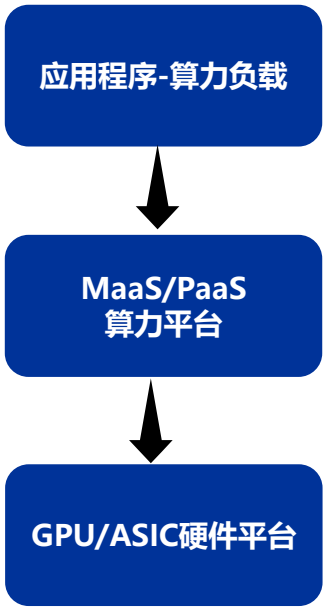
主要内容

1. AI云计算新范式：规模效应+AI Infra能力+算力自主化
2. 规模效应：资本密集度+多租户+内部负载的削峰填谷
3. AI Infra：实现计算性能挖潜
4. 算力自主化：海外ASIC芯片趋势启示
5. 重点标的：互联网云厂+ASIC芯片
6. 重点公司估值表及风险提示

3.1 AI Infra：从算力到应用的基础设施软件/工具

- **AI Infra定位于算力与应用之间的“桥梁”角色的基础软件设施层，包括：1）算力硬件层面的组网、算力资源调度等，实现集群高效率；2）模型层面提供的工具库、框架库的丰富度及有效性，帮助云客户实现高效资源调用；3）针对具体应用的定向优化。**
- **各厂商间AI Infra能力有较大差距。**不同于开发生态十分成熟、潜能已充分挖掘的CPU，GPU/ASIC硬件的开发生态仍在不断迭代丰富中，不同AI Infra工程能力的团队对于算力硬件的利用率有较明显差距。

表：AI Infra从硬件平台到软件工具



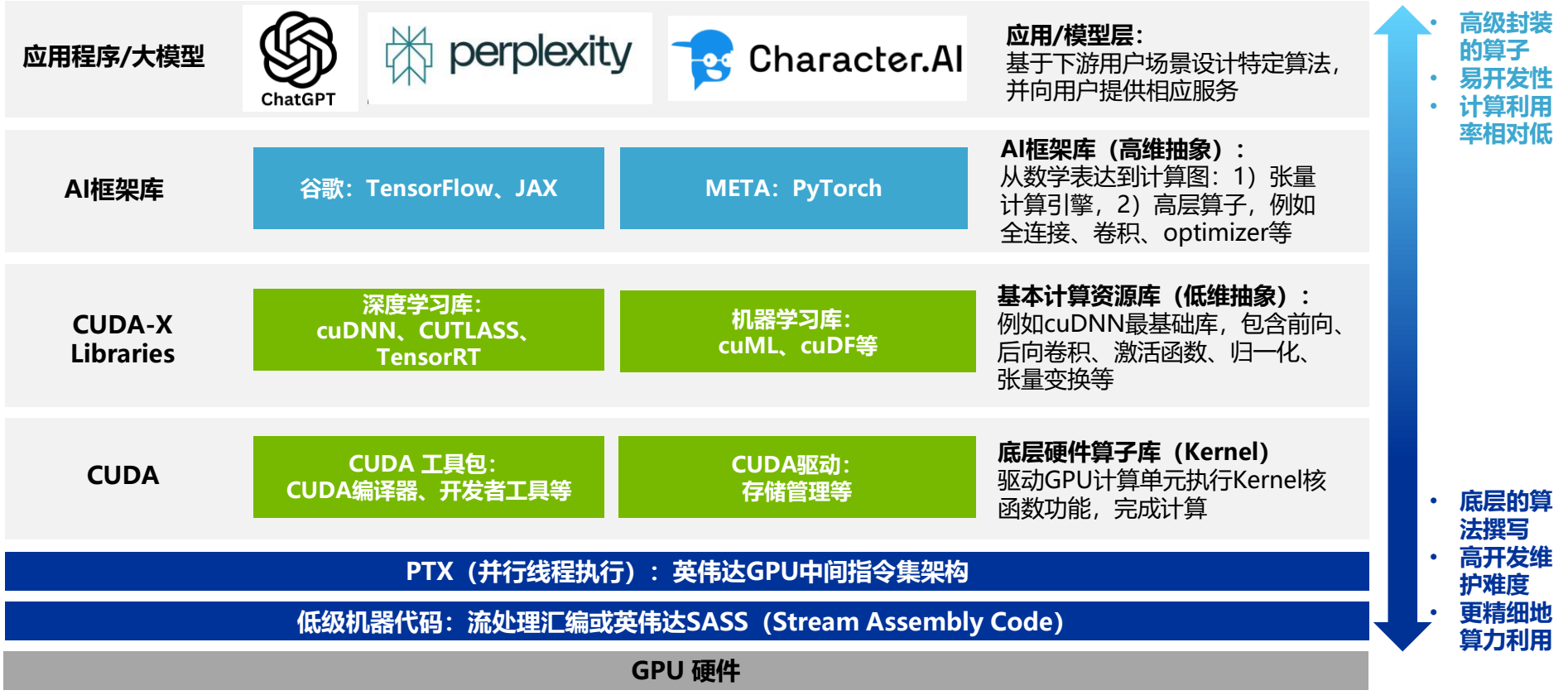
AI Infra能力层	所处层次主要工作	AI Infra具体能力/实现方式	以谷歌/DeepSeek为例的典型工作
应用管理层	提供资源管理、运营管理、运维管理等运营能力	针对具体的应用进行定向优化，降低推理成本等	➢ 谷歌：根据具体使用场景，基于大模型能力开发AI Agent、AI应用（NotebookLM）等
模型管理层	提供模型开发和应用所需的各种基础工具和组件	主要为软件、算法能力。 1）提供AI框架库、开发资源库、工具库； 2）针对大模型进行计算效率的算力优化、负载均衡、拥塞控制等	➢ 谷歌：1）提供Tensorflow深度学习框架库以及众多工具；2）针对大模型进行定制化优化。 ➢ DeepSeek：针对大模型进行专家并行、数据并行等方面的优化
算力管理层	提供计算、存储、网络、安全等基础资源和服务	主要为通信优化、算力资源调度、管理能力。 包括通信组网、异构计算协调、容器管理、弹性部署等	➢ 谷歌：1）组网：通过OCS组建TPU集群；2）通过Pathway实现异构计算资源大规模编排调度； ➢ DeepSeek：构建Fire-Flyer AI-HPC集群，在组网、通信方面定向优化；

资料：谷歌、DeepSeek官网，申万宏源研究

3.1 AI Infra：优化主要由云厂/互联网/大模型厂商完成

- 具体看，从硬件到大模型的训练推理仍有AI框架库、AI资源库、底层算子等生态层次，英伟达CUDA生态提供众多AI Infra工具，能够提供较好的计算利用率，但以出售硬件产品为目的的英伟达，在AI Infra优化上进一步算力挖潜的动机略显不足。因此**云厂商/互联网/大模型厂商将承担主要的AI Infra优化、计算效能挖潜任务。**

图：基于英伟达GPU的开发工程栈，DeepSeek自PTX层定制算子优化算法工程

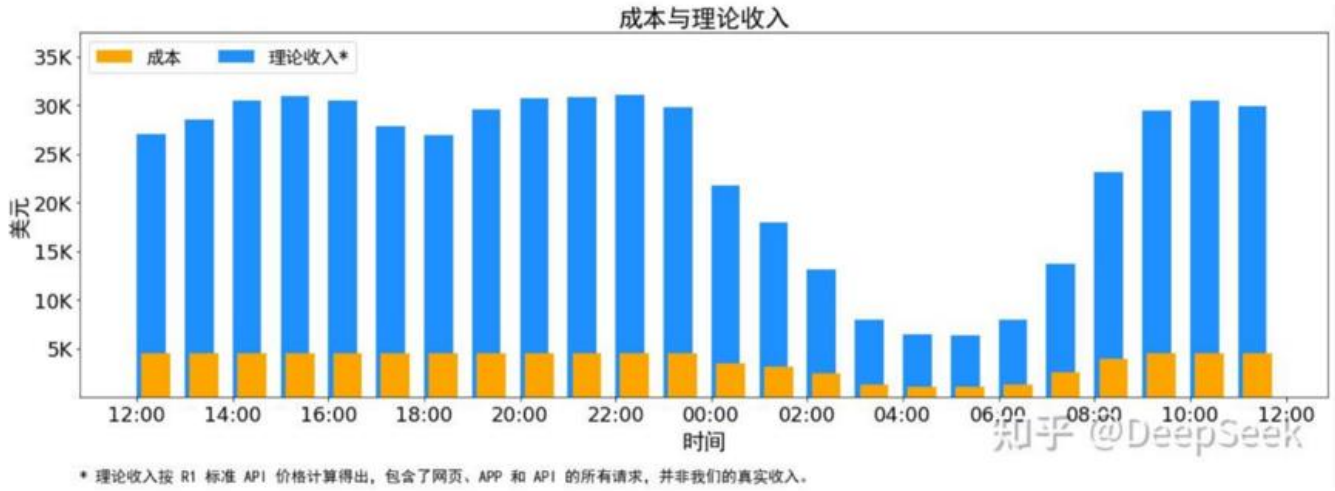


 资料 ：英伟达官网，申万宏源研究

3.2 DeepSeek启示：AI Infra能力对推理成本影响重大

- **AI Infra能力正拉开AI应用/大模型API的单次推理成本差距。**英伟达GPU提供的开发工具适用于标准化通用需求，易开发性出色，但大模型至硬件调用间仍有多个步骤可实现成本优化，优化与否将拉开成本差距。
- DeepSeek测算的应用理论利润率出色，一大核心在于其针对特定DeepSeek R1大模型进行充分优化。**而同为DeepSeek R1模型搭载于第三方大模型平台，若未进行充分优化，则其推理成本仍将相对较高。**例如大模型平台公司潞晨科技停用DeepSeek R1 API接口，或为成本侧难以复制DeepSeek的优化措施，成本仍较高。

图：DeepSeek列举的DeepSeek应用理论收入及成本对比，可实现利润/成本=545%的理论比例

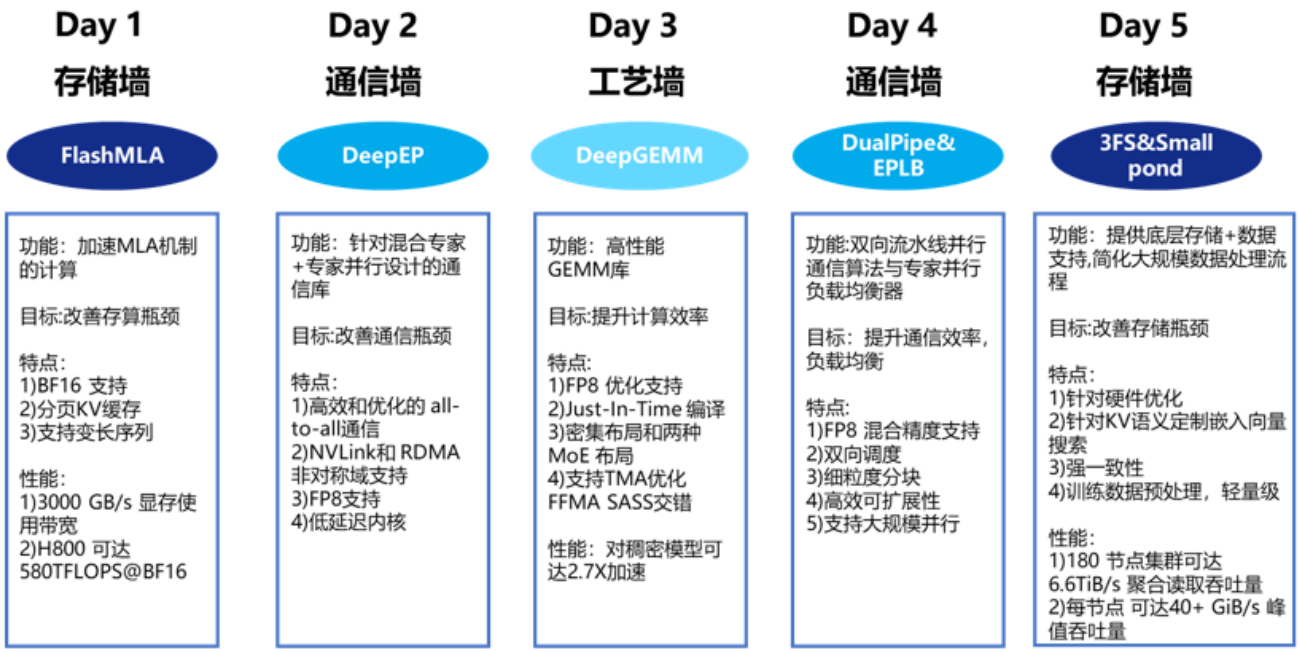


资料 ：《DeepSeek-V3/R1推理系统概览》，申万宏源研究
注：仅为理论测算，未考虑收费服务的折扣以及实际收费Tokens占比，实际场景中比例将低于此

3.2 DeepSeek：AI Infra优化深入AI工程栈全环节

- 从算力硬件到大模型的API调用，**其中的众多环节可均有较大优化空间**，AI Infra能力体现在针对改善存储瓶颈、提升通信效率、提升计算单元效率等方面，实际上是对已有GPU性能的进一步发掘：**1）让大模型推理/训练中计算、通信、存取方式更简洁，减少算法粗糙下的算力浪费；2）根据具体的GPU（如英伟达H100）的微架构设计，针对性实现优化。**

图：DeepSeek开源周发布了各环节算法工程优化的工具



资料 ：DeepSeek官网，中存算，申万宏源研究

3.3 互联网云：在AI Infra领域已有较深技术积累

- AI Infra能力的积累通常需要具备前沿大模型开发经验，即完成了**构建AI算力集群→基于集群的大模型训练→提供大模型API推理服务→构建上层AI应用的全工作栈。**
- 大模型厂商/互联网云已积累较强的AI Infra能力，发布较多AI Infra成果，包括**实现万卡集群的高利用率、提供丰富的大模型训练和推理工具提升开发效率等**，已具备较为明确的优势。

表：字节、腾讯、阿里巴巴、DeepSeek在AI Infra上的主要工作

平台	IaaS重要AI Infra工作	MaaS/PaaS重要AI Infra工作
字节跳动 火山引擎	**Gödel实现万卡集群的资源调度** 自 2022 年开始在字节跳动内部各数据中心批量部署，Gödel 调度器已经被验证可以在高峰期提供 >60%的 CPU 利用率和 >95%的 GPU 利用率。	**MegaScale大模型训练框架** MegaScale系统在12,288个GPU上训练175B LLM模型时模型FLOPs利用率（MFU）达到了55.2%，**比起英伟达的Megatron-LM，提升了1.34倍。**
腾讯 腾讯云平台	**高性能网络IHN** 单集群支持万卡规模，单机支持3.2T大带宽，通信占比低至 6%，训练效率提升 20%。	**TACO大模型推理加速套件** 同样以 Llama-3.1 70B 为例，使用 TACO-LLM 部署的成本低至 <$0.5/1M tokens，相比直接调用 MaaS API 的成本节约超过60%+，且使用方式、调用接口保持一致，支持无缝切换。
阿里巴巴 阿里云	**灵骏计算集群+HPN 7.0组网架构** 灵骏计算集群提供可扩容到 10 万张 GPU 卡规模的能力，同时在万卡的规模下性能线性增长率达到了 96%，性能网络吞吐的有效使用率也达到了99%。	**训练框架PAI-ChatLearn** 相比于当前的SOTA 系统，ChatLearn在 7B+7B 规模有 115% 的加速，在 70B+70B规模有 208% 的加速。同时ChatLearn可以扩展到更大规模，如：300B+300B(Policy+Reward)。
DeepSeek	**Fire-Flyer AI-HPC集群** 在DL训练中部署含1万个PCIe A100 GPU的Fire-Flyer 2，实现了接近NVIDIA DGX-A100的性能，同时将成本降低近一半，能源消耗降低了40%。	**HAI LLM训练框架** 包括HAI Scale算子库等，针对专家并行、流水线并行、张量并行等领域的通信、计算能力进行大量优化。

资料 ：阿里云、火山引擎、腾讯云、DeepSeek官网，申万宏源研究

3.3 字节：MegaScale针对万卡集群训练大幅提升MFU

- **模型训练两大挑战：1）实现高训练效率：**体现在MFU（模型计算利用率），即实际吞吐量/理论最大吞吐量，与集合通信、算法优化、数据预处理等相关，**2）保持高训练效率：**体现在降低初始化时间和容错修复能力。
- **字节算法优化：Transformer Block 并行、滑动窗口的Attention、LAMB优化器。**实现初始化时间大幅优化，2048卡GPU集群初始化时间从1047秒下降到5秒以下。实现高效容错管理：自动检测故障并实现快速恢复工作。
- **网络优化：**1）基于博通Tomahawk 4的交换机，优化网络拓扑结构；2）降低ECMP哈希冲突：将数据密集型节点都安排在一个ToR交换机上；3）拥塞控制：将往返时延精确测量与显式拥塞通知的快速拥塞响应能力结合。

图：字节在2024年2月提出的MegaScale训练框架的MFU相对英伟达的Megatron-LM大幅优化，万卡集群MFU达到55.2%

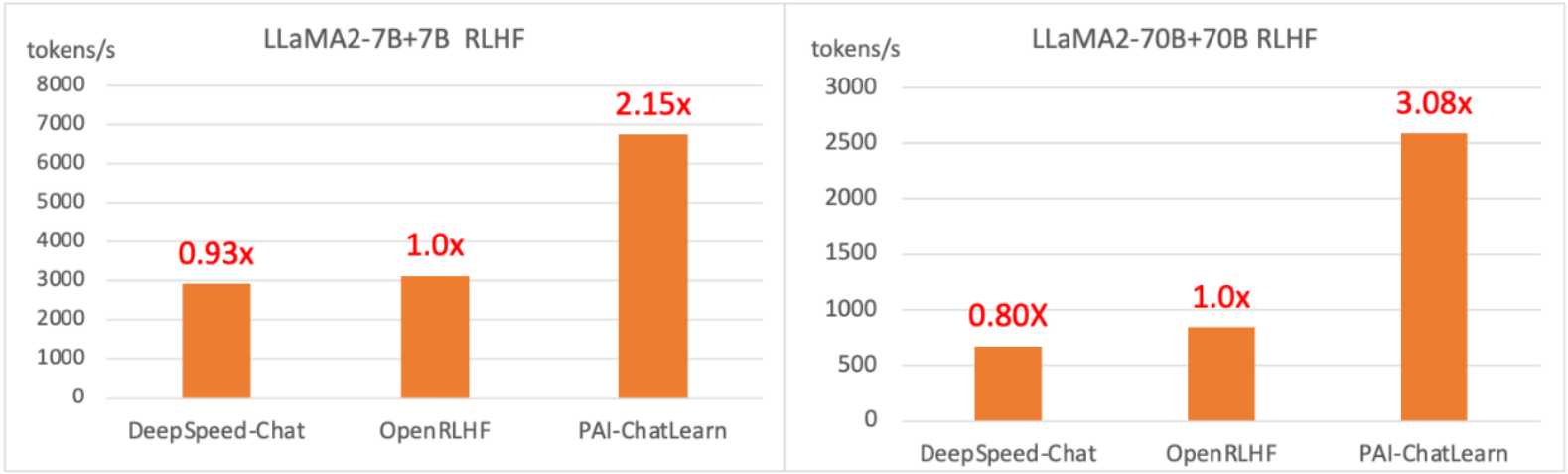
Batch Size	Method	GPUs	Iteration Time (s)	Throughput (tokens/s)	Training Time (days)	MFU	Aggregate PFlops/s
768	Megatron-LM	256	40.0	39.3k	88.35	53.0%	43.3
		512	21.2	74.1k	46.86	49.9%	77.6
		768	15.2	103.8k	33.45	46.7%	111.9
		1024	11.9	132.7k	26.17	44.7%	131.9
	MegaScale	256	32.0	49.0k	70.86	65.3%(**1.23**×)	52.2
		512	16.5	95.1k	36.51	63.5%(**1.27**×)	101.4
		768	11.5	136.7k	25.40	61.3%(**1.31**×)	146.9
		1024	8.9	176.9k	19.62	59.0%(**1.32**×)	188.5
6144	Megatron-LM	3072	29.02	433.6k	8.01	48.7%	466.8
		6144	14.78	851.6k	4.08	47.8%	916.3
		8192	12.24	1027.9k	3.38	43.3%	1106.7
		12288	8.57	1466.8k	2.37	41.2%	1579.5
	MegaScale	3072	23.66	531.9k	6.53	59.1%(**1.21**×)	566.5
		6144	12.21	1030.9k	3.37	57.3%(**1.19**×)	1098.4
		8192	9.56	1315.6k	2.64	54.9%(**1.26**×)	1400.6
		12288	6.34	1984.0k	**1.75**	55.2%(**1.34**×)	2166.3

资料：《MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs》，申万宏源研究

3.3 阿里云：PAI-ChatLearn实现RLHF训练效率提升

- **PAI-ChatLearn 是阿里云 PAI 团队自研的、灵活易用的、支持大规模 Alignment 高效训练的框架。**
- ChatLearn通过对 Alignment 训练流程进行合理的抽象和解耦，提供灵活的资源分配和并行调度策略。ChatLearn提供了RLHF、DPO、OnlineDPO、GRPO等对齐训练，同时也支持用户自定义大模型训练流程。相比于当时的SOTA 系统，ChatLearn在7B+7B规模有115%的加速，在70B+70B规模有208% 的加速。

图：阿里巴巴2024年8月开源的大规模对齐训练框架PAI-ChatLearn在Llama2模型 RLHF训练中实现更高效率



资料 ：GitHub，阿里云官网，申万宏源研究

主要内容

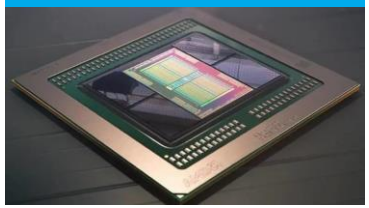
1. AI云计算新范式：规模效应+AI Infra能力+算力自主化
2. 规模效应：资本密集度+多租户+内部负载的削峰填谷
3. AI Infra：实现计算性能挖潜
4. 算力自主化：海外ASIC芯片趋势启示
5. 重点标的：互联网云厂+ASIC芯片
6. 重点公司估值表及风险提示

4.1 ASIC VS GPU：架构、生态、成本对比

- **从IC设计思路来看，GPU为自下而上，即基于已设计的硬件平台作工具丰富、生态适配工作支持上层应用；ASIC（专用集成电路）则是自上而下，基于现有应用/工作负载进行芯片架构设计，**通过更定制化、针对性的架构设计匹配算法提升计算效能，但将牺牲通用性，完成非特定任务的效率较差。
- 但云客户更倾向于使用开发生态成熟、具备易开发性的英伟达GPU，预计在较长时间内仍将为云服务的首选。**有望形成英伟达GPU仍占据公有云市场、ASIC芯片在巨头内部负载形成替代的并行格局。**

图：主要的AI算力芯片分类

通用性 ← → 计算效能

	CPU	GPU	FPGA	ASIC
				
芯片架构	• 冯诺依曼架构，串行计算为主 • 计算单元占比较低，重在控制	• **冯诺依曼架构，并行计算为主** • **计算单元占比很高**	• 哈佛架构，无须共享内存 • 可重构逻辑单元	• **非冯诺依曼架构** • **计算单元占比高**
应用构建	标准化硬件，用户基于架构固定的硬件构建应用/工作负载	标准化硬件，用户基于架构固定的硬件构建应用/工作负载	可编程硬件，可灵活根据应用/工作负载在使用过程中改变硬件架构	定制化硬件，根据应用/工作负载特点设计硬件架构
开发生态	十分成熟	**仅英伟达的CUDA较成熟，其他GPU厂商生态成熟度较低**	可适用主流编程语言	**生态成熟度相对较低**
相对优劣势	• 通用性最强，编程难度低 • 计算能力弱，不适用于AI计算	• 通用性较强，并行计算能力出色适用于AI • **功耗较高，编程难度中等**	• 灵活性好，多用于推理环节 • 峰值计算能力较弱	• **计算效能出众功耗低，成本更低** • 仅在特定类别的工作负载表现出色，灵活性差，**编程难度高**

资料 ：寒武纪招股说明书，申万宏源研究