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

证券分析师：林起贤 A0230519060002 夏嘉励 A0230522090001
研究支持：黄俊儒 A0230123070011
联系人：黄俊儒

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

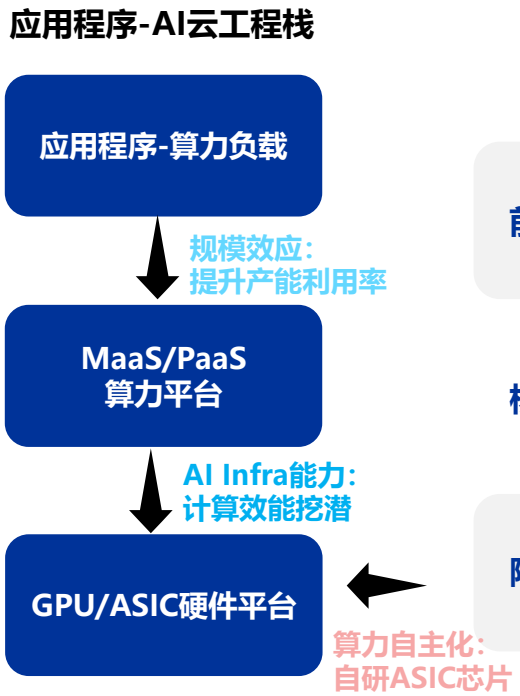
核心竞争要素	传统云计算	AI云计算
规模效应-资本密集度	资本密集型，以CPU服务器为主 2023年海外大型云/互联网巨头Capex总和为1534亿美元	以GPU服务器为主，单集群**Capex更庞大** 2024年海外大型云/互联网巨头Capex总和为2504亿美元
规模效应-产能利用率	IT设备成本占比约为40%-50%，基建、运维占比相对高，对产能空置有一定容忍度	IT设备成本占比约为60%-70%，折旧压力更大,对产能空置容忍度很低，**拥有内部负载/自研大模型/云多租户十分重要**
AI Infra-云基础设施能力	传统云基础设施强调稳定性及服务，虚拟化、容器等技术已趋于成熟	AI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**

资料来源：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

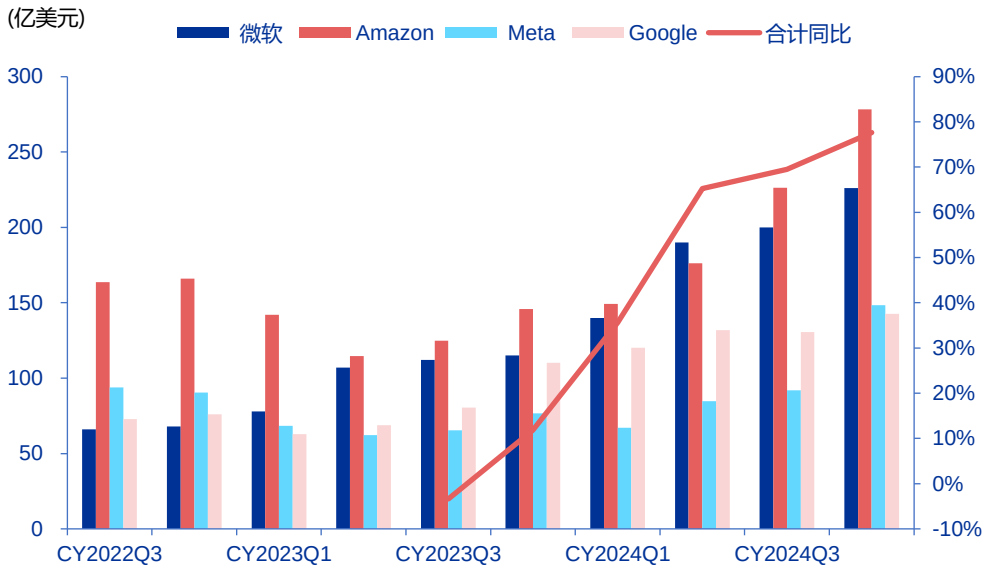
主要内容

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

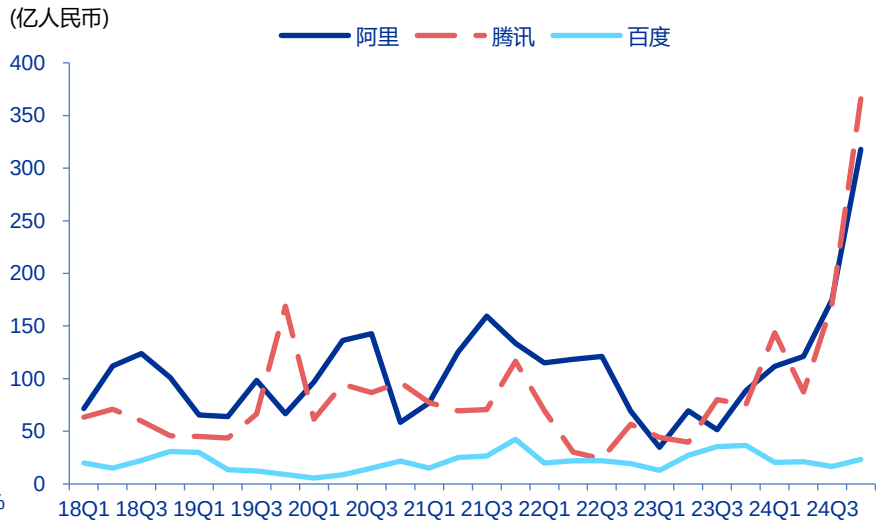
2.1 资本密集度：构建AI云集群的支出量级仍在不断扩大

- 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



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

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%

资料来源：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

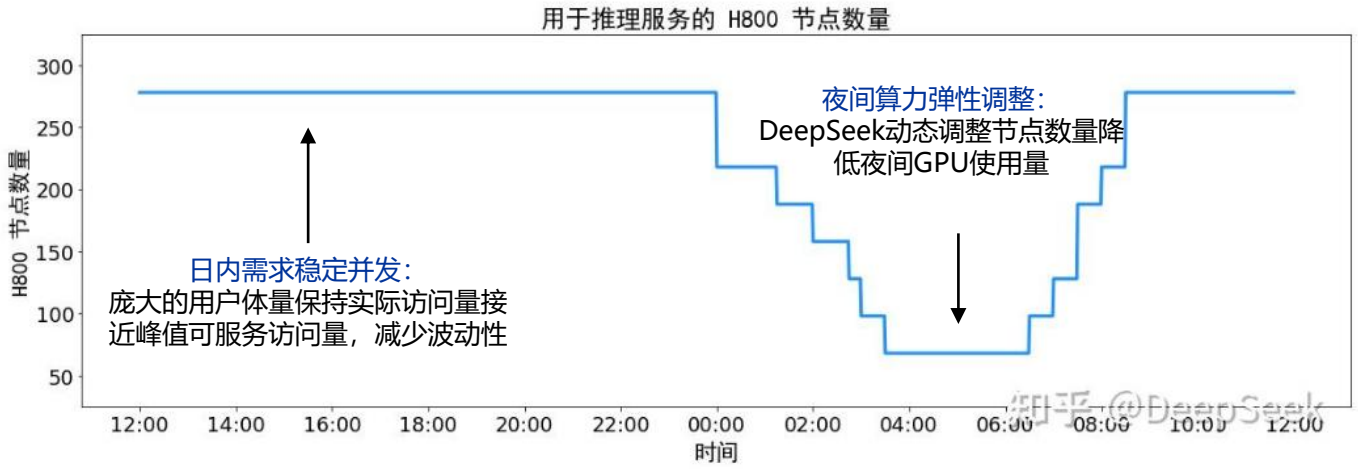
单位：亿美元		CY23Q1	CY23Q2	CY23Q3	CY23Q4	CY24Q1	CY24Q2	CY24Q3	CY24Q4
谷歌云	营收	74.54	80.31	84.11	91.92	95.74	103.47	113.53	119.55
	同比增速	28.1%	28.0%	22.5%	25.7%	28.4%	28.8%	35.0%	30.1%
	营业利润率	2.6%	4.9%	3.2%	9.4%	9.4%	11.3%	17.1%	17.5%
亚马逊AWS	营收	213.54	221.40	230.59	242.04	250.37	262.81	274.52	287.86
	同比增速	15.8%	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%

资料来源：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

资料来源：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、爱奇艺等的AI推荐算法**

资料来源：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

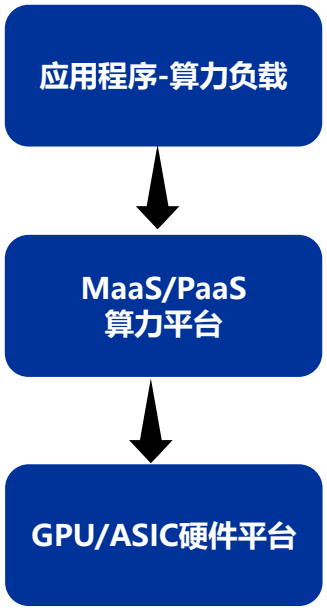
主要内容

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



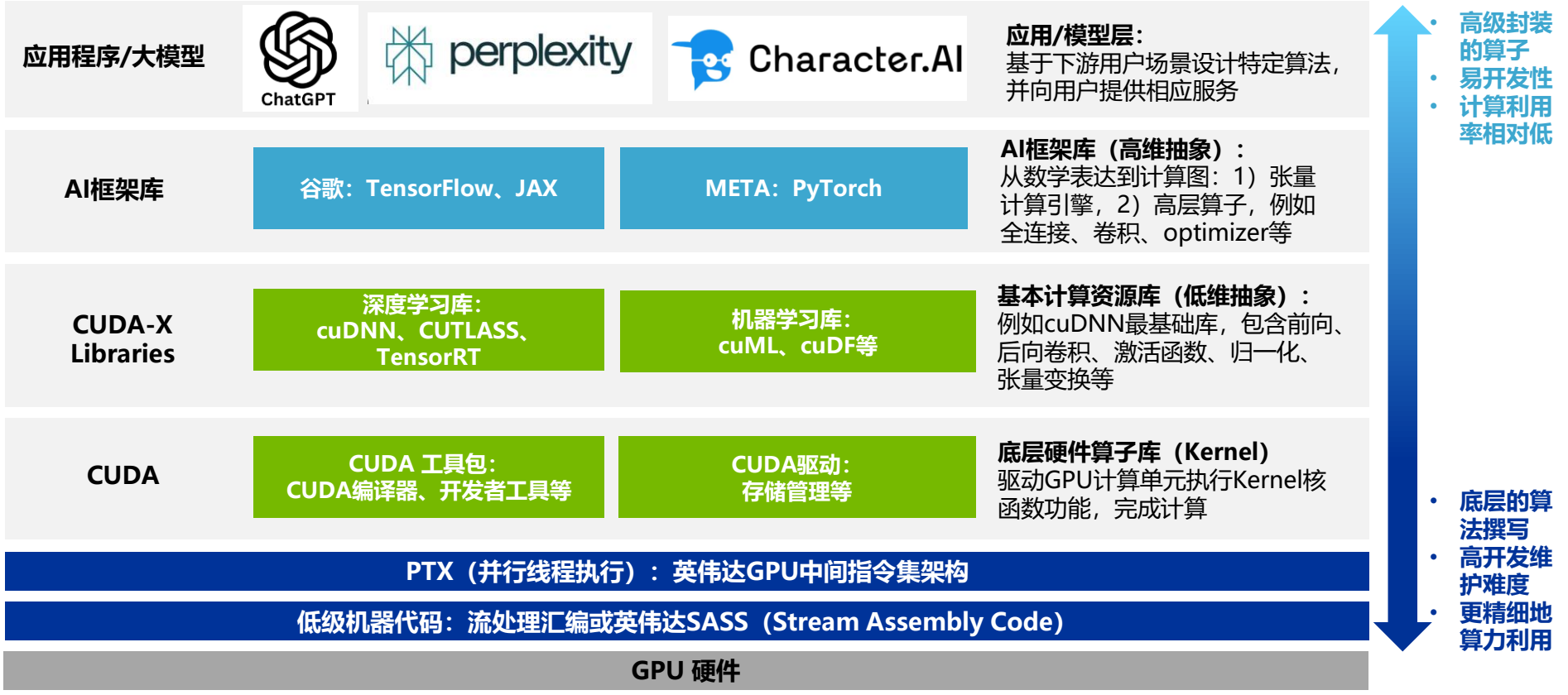
AI Infra能力层	所处层次主要工作	AI Infra具体能力/实现方式	以谷歌/DeepSeek为例的典型工作
应用管理层	提供资源管理、运营管理、运维管理等运营能力	针对具体的应用进行定向优化，降低推理成本等	➢ 谷歌：根据具体使用场景，基于大模型能力开发AI Agent、AI应用（NotebookLM）等
模型管理层	提供模型开发和应用所需的各种基础工具和组件	主要为软件、算法能力。 1）提供AI框架库、开发资源库、工具库； 2）针对大模型进行计算效率的算力优化、负载均衡、拥塞控制等	➢ 谷歌：1）提供Tensorflow深度学习框架库以及众多工具；2）针对大模型进行定制化优化。 ➢ 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；

资料来源：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。**

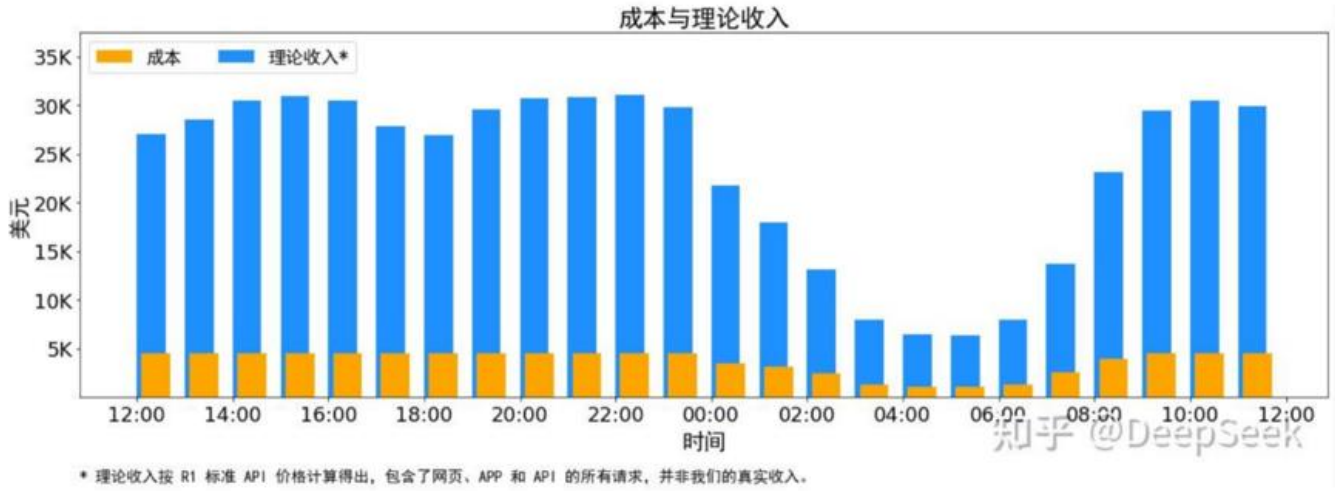
图：基于英伟达GPU的开发工程栈，DeepSeek自PTX层定制算子优化算法工程



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

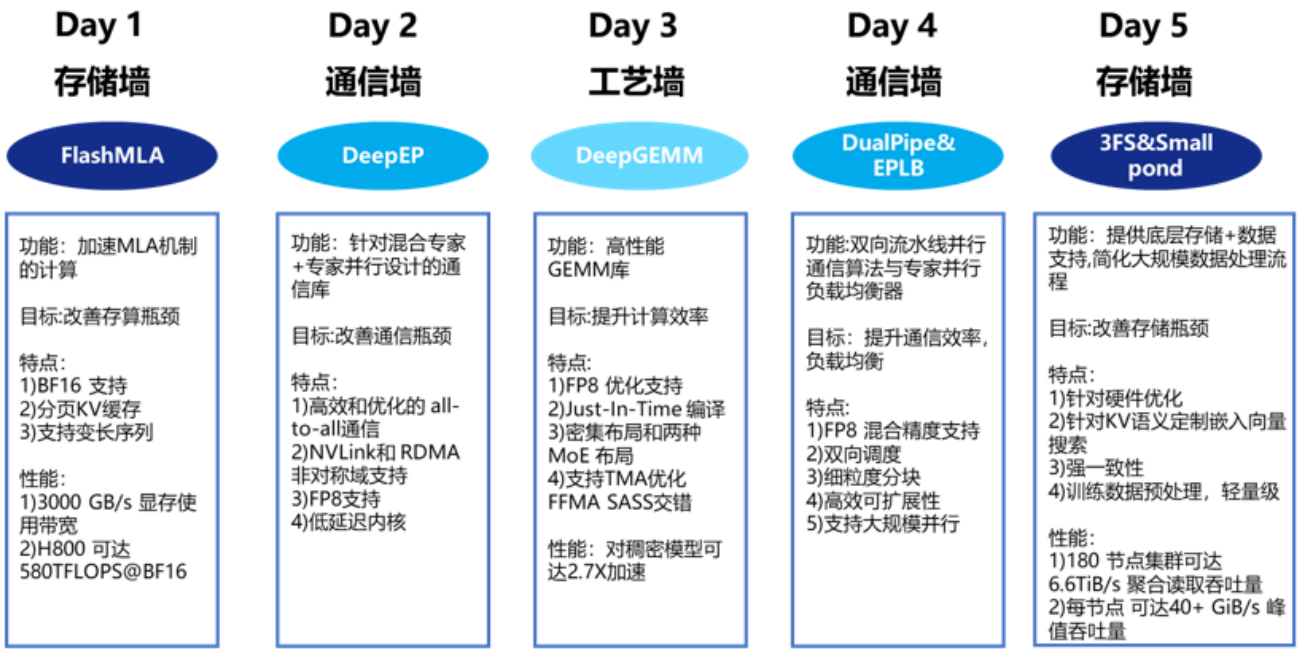


资料来源：《DeepSeek-V3/R1推理系统概览》，申万宏源研究
注：仅为理论测算，未考虑收费服务的折扣以及实际收费Tokens占比，实际场景中比例将低于此

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



资料来源：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。

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。

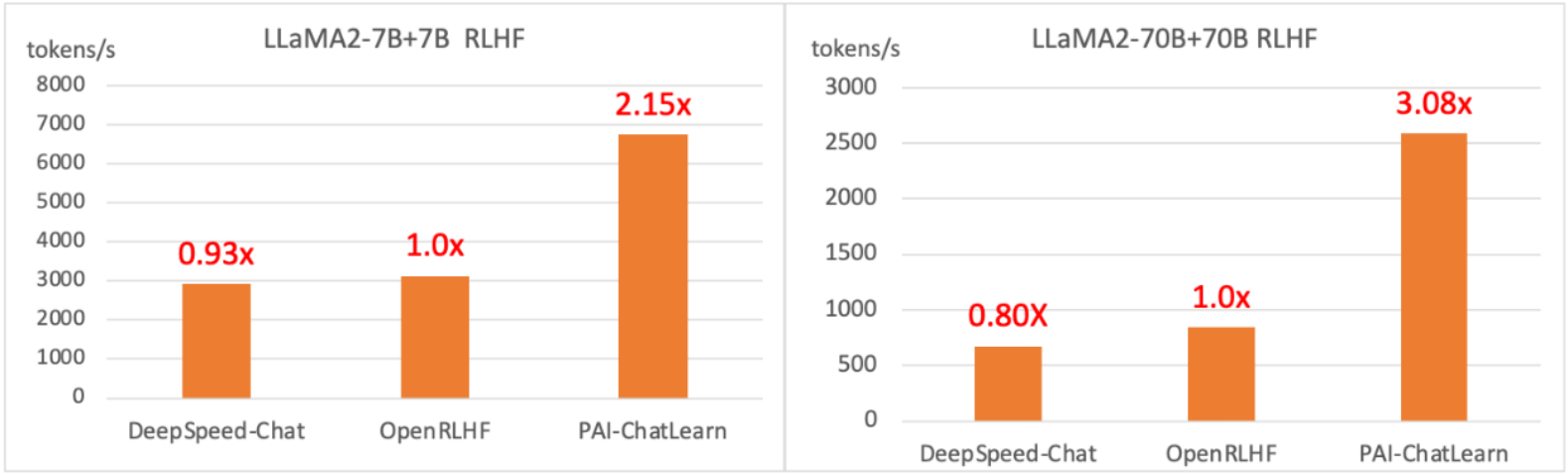
图：字节在2024年2月提出的MegaScale训练框架的MFU相对英伟达的Megatron-LM大幅优化，万卡集群MFU达到55.2%

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

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



主要内容

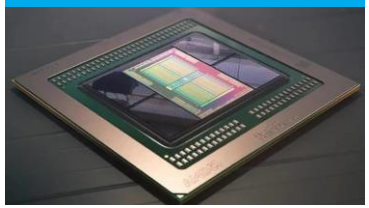
1. AI云计算新范式：规模效应+AI Infra能力+算力自主化
2. 规模效应：资本密集度+多租户+内部负载的削峰填谷
3. AI Infra：实现计算性能挖潜
4. 算力自主化：海外ASIC芯片趋势启示
5. 重点标的：互联网云厂+ASIC芯片
6. 重点公司估值表及风险提示

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**

4.2 ASIC：架构+生态大相径庭，将成为GPU的有力补充

GPU与ASIC在架构及开发生态上有着明确差异：

- **架构存在差异：**GPU基于通用并行计算向AI转变，内部设计通常为大量并行计算核+小型AI加速单元TensorCore；TPU等则为仅针对AI算力需求场景直接设计内部架构，代表架构有大型脉动阵列等。
- **开发生态存在差异：**英伟达具备完整成熟的CUDA开发生态，AMD GPU/ASIC厂商开发生态均不完善。

图：各家GPU/ASIC芯片对比

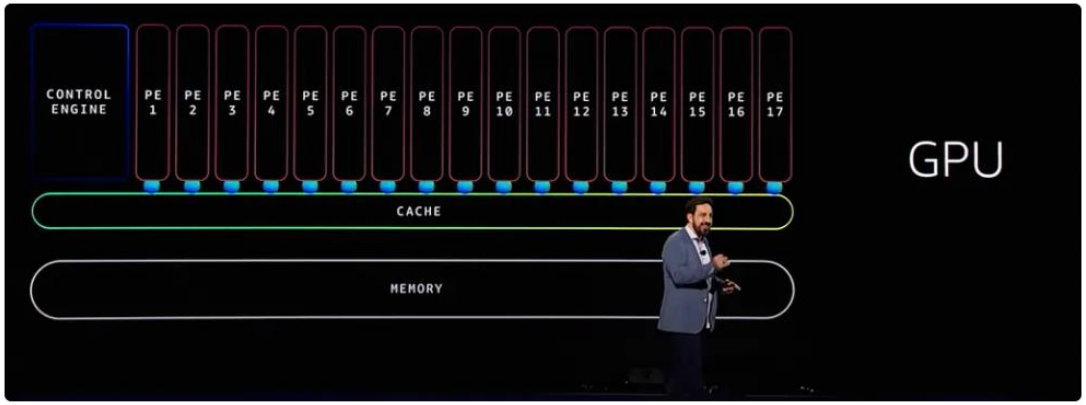
	NVIDIA H100	NVIDIA B200	NVIDIA B300	AMD MI325x	TPU v5p	TPU v6e	Trainium 2	META MTIA v2	微软 Maia 100
推出时间	2022	2024	2024	2024	2023	2024	2023	2024	2023
芯片制程	4nm	4nm	4nm	5nm	5nm	4nm	5nm	5nm	5nm
峰值计算性能-BF/FP16（TFlops）	990	2250	3375	1300	459	926	431	177	800
功耗	700W	1000W	1200W	1000W	-	-	-	90W	860W
存储类型	HBM3	HBM3e	HBM3e	HBM3e	HBM2e	HBM3	HBM3	LPDDR5	HBM3e
存储（GB）	80	192	288	256	96	32	96	128	64
内存带宽	3.35TB/s	8TB/s	8TB/s	6TB/s	2765GB/s	1640GB/s	4000GB/s	204.8GB/s	1600GB/s
卡间通信带宽	NVLink 900GB/s	NVLink 1800GB/s	NVLink 1800GB/s	Infinity Fabric Link 896GB/s	ICI Links 600GB/s	ICI Links 3584GB/s	NeuronLink 768GB/s	-	600GB/s
计算强度-FP16峰值性能/存储（Flops/GB）	12.4	11.7	11.7	5.1	4.8	28.9	4.5	1.4	12.5
芯片架构+开发生态									
Compute Die数量	1	2	2	4	1	1	2	1	1
HBM Stacks数量	6	8	8	8	6	2	4	4	4
计算单元微架构	**大量并行CUDA核+TensorCore**	**大量并行CUDA核+TensorCore**	**大量并行CUDA核+TensorCore**	**大量并行运算核+Matrix Core**	**少量大型脉动阵列单元**	**少量大型脉动阵列单元**	**少量大型脉动阵列单元**	**多核CPU+多核AI加速单元**	**多核AI加速单元**
开发生态	**CUDA**	**CUDA**	**CUDA**	**Rocm**	**XLA**	**XLA**	**XLA**	**Triton**	**Triton**

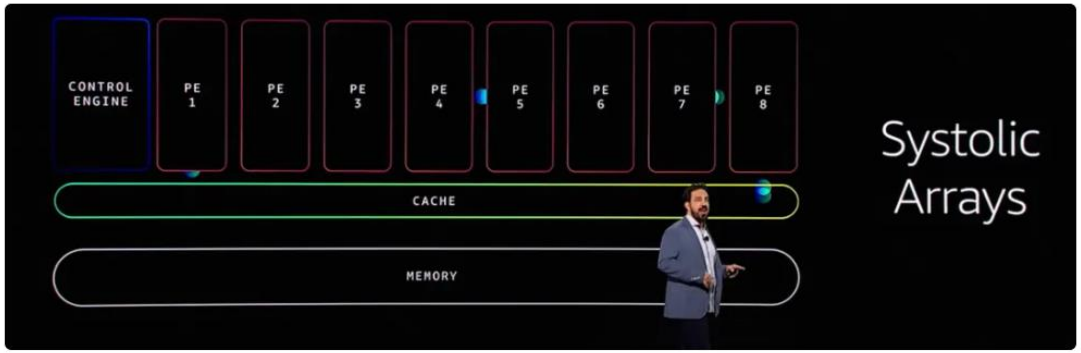
资料来源：英伟达、谷歌、AMD、亚马逊、META、微软官网，申万宏源研究

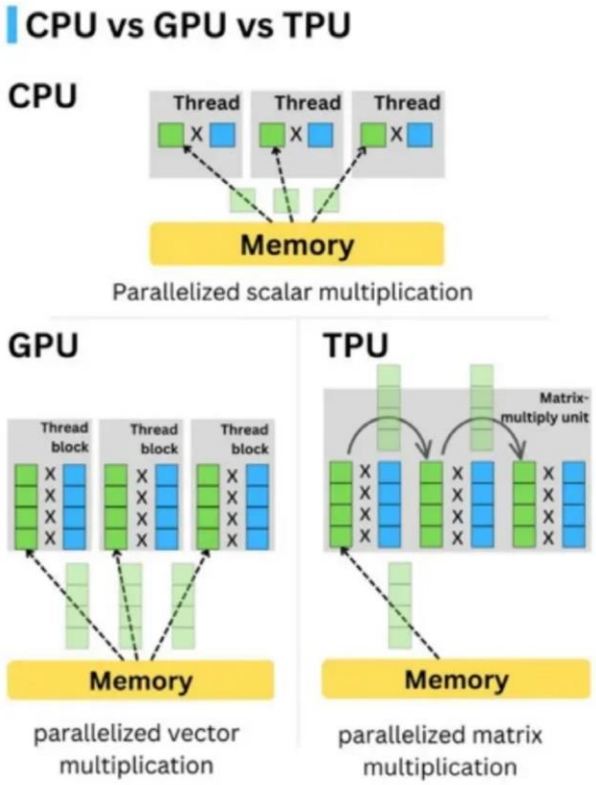
4.2 ASIC架构：基于脉动阵列的定制架构为重要路线

- **GPU为冯诺依曼架构，运算中与寄存器需要高频数据交换，对存储容量要求较高。**GPU主要是针对数据并行执行，控制单元较小，执行单元众多，同时有大量的寄存器文件用于在多个执行线程上隐藏延迟。
- **谷歌TPU、AWS Tranium2均基于脉动阵列架构，**专为矩阵计算设计，计算结果可以直接向下一个计算单元递推，直到该维度的矩阵结果计算完毕，再与寄存器作数据存取，减少不必要的全局数据交换等。

图：脉动阵列架构专用于大型矩阵计算，可降低存储消耗



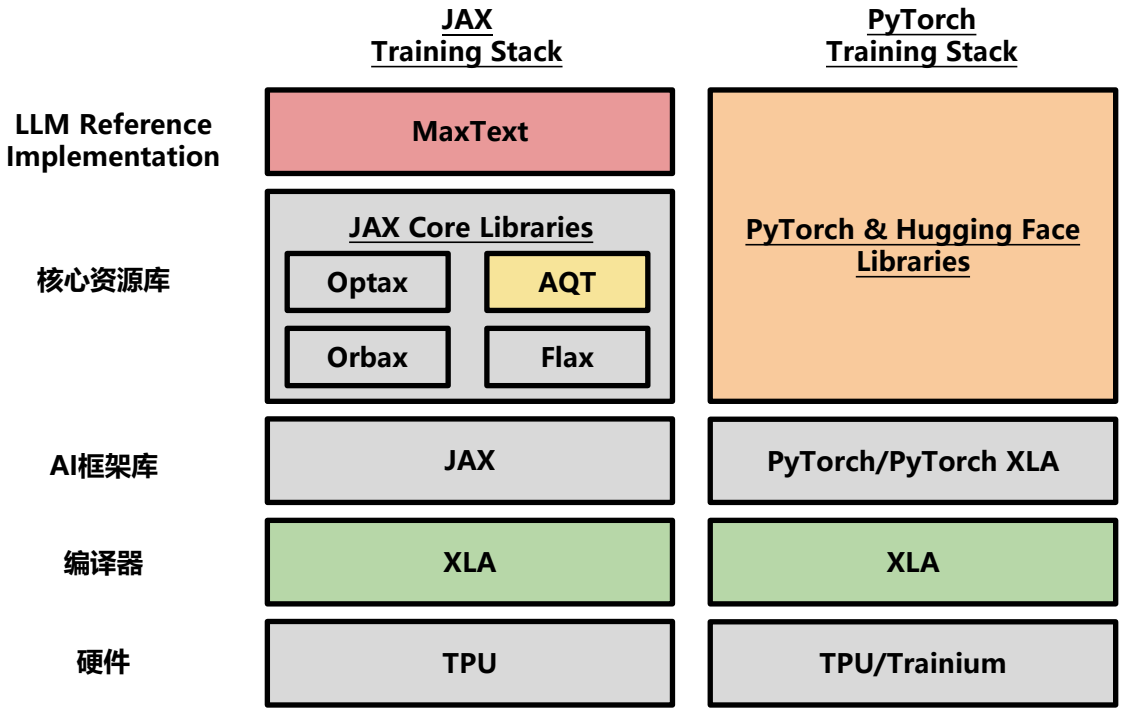




4.2 ASIC开发生态：谷歌和AWS均基于脉动阵列+XLA

- **开发生态应在硬件架构/计算架构ROI提升的方向逐渐成熟，国内AI算力+海外云厂ASIC芯片等均具备潜力。**DeepSeek实际证明拥有强大的工程团队，有能力为其他AI芯片构建相对可用的开发生态（但易开发性预计仍有明显差距。
- XLA为谷歌为TPU构建的编译器，并陆续结合JAX等AI框架形成开发生态，逐渐走向成熟，同为脉动阵列架构的AWS Tranium2同样采用XLA，将加速相关生态更新迭代。

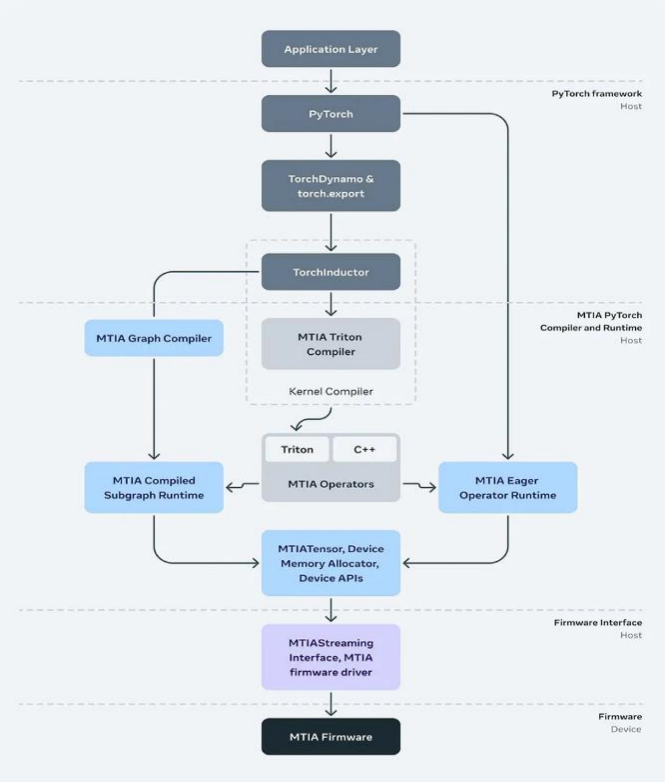
图：谷歌TPU/亚马逊Trainuim基于XLA的开发生态栈



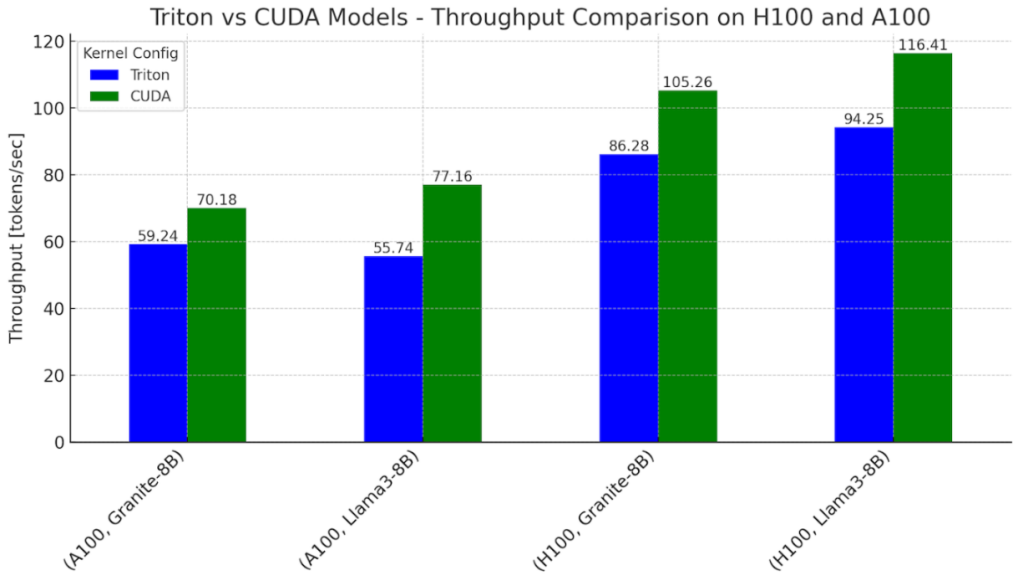
4.2 ASIC开发生态：META MTIA v2软件堆栈基于Triton

- Triton为OpenAI于2021年7月推出的类似Python的开源编程语言，旨在降低GPU的编程难度，但Triton并不非常依赖单一厂商的GPU，可拓展至MTIA v2等非GPU硬件架构。
- **Pytorch正致力于推广Triton，已经在英伟达GPU上实现无CUDA条件下较高的硬件效率。MTIA v2基于Triton，并提供Triton-MTIA编译器进一步优化软件堆栈。**

图：MTIA v2软件堆栈主要基于Triton编程语言



图：Pytorch使用无CUDA的Triton编译语言实现较高的GPU调用效率



资料来源：Pytorch官网，申万宏源研究
注：LLM模型选用Meta的Llama3-8B和IBM的Granite-8B代码模型，推理计算使用FP16

4.2 ASIC成本：具备性价比，但使用范围相对局限

- **我们简单测算各家芯片的制造成本，**主要根据各芯片具体的存储容量、晶圆尺寸等进行测算，并根据英伟达、博通、Marvell/AIChip大致的毛利率进行估计，大致推测各家芯片的价格。
- **ASIC芯片在特定任务部署中实际具备性价比，**但受限于开发生态：1）开发过程中，生态不成熟存在开发效率损失，一定程度提升隐性成本。2）场景限于云厂内部负载，云客户基于其开发的难度较大。

表：各家GPU/ASIC芯片预计的成本拆分对比测算

单位：美元	H100	B200	TPU v5p	TPU v6e	Trainium2
厂商	英伟达	英伟达	谷歌-博通	谷歌-博通	亚马逊-AIChip/Marvell
制程	4nm	4nm	5nm	4nm	5nm
峰值计算性能-BF16/FP16（TFlops）	990	2250	459	926	431
存储（GB）	96	192	96	32	96
存储类型	HBM3	HBM3e	HBM2e	HBM3	HBM3
预计存储成本	1150	2700	1000	400	1150
预计制造、封测等成本	1350	2150	800	550	1000
总成本	2500	4850	1800	950	2150
预计业务毛利率	85%~90%	85%~90%	65%~70%	65%~70%	47%~53%
估算的各家芯片单价	18000	33000	6000	3100	4400

资料来源：英伟达、谷歌、AMD、亚马逊、META、微软官网，申万宏源研究
注：成本及单价仅为估算，实际价格与成本为动态变化

4.2 ASIC成本：典型训练场景具备性价比

芯片产品	NVIDIA H100	NVIDIA B200 GPU	NVIDIA GB200 Superchip	Google/博通 TPU v5p	AWS/Marvell Trainium 2
训练LLama3 405B模型所需的FP16算力总量 (ZFLOPS)	18000	18000	18000	18000	18000
单卡FP16峰值性能(TOPS)	990	2250	5000	459	650
计算性能使用效率 (%)	40%	40%	40%	40%	40%
平均计算性能 (TFLOPS)	396	900	2000	184	260
单卡单日算力（PFLPOS）	34214	77760	172800	15863	22464
集群单日算力(ZFLOPS)	821	1866	4147	381	539
大模型训练所需的运行天数 (24000卡集群)	22	10	4	47	33
算力芯片硬件成本					
AI加速器芯片数量-算力集群	24000	24000	24000	24000	24000
AI加速器芯片数量-每台服务器	8	8	36	8	16
服务器数量-算力集群	3000	3000	667	3000	1500
AI加速器芯片价格(美元/片)	18000	33000	70000	6000	4400
AI加速器芯片+CPU价格（万美元）-每台服务器	144	264	2524	48	71
AI加速器芯片+CPU的成本（亿美元）-算力集群	43.3	79.3	168.3	14.5	10.6
折旧年限	4	4	4	4	4
算力集群中AI算力硬件年折旧费用（亿美元）	**10.8**	**19.8**	**42.1**	**3.6**	**2.7**
能源成本					
AI加速器设计功耗	700W	1000W	2400W	700W	700W
每瓦特AI计算性能 (TFLOPS/Watt)	1.4	2.3	2.1	0.7	0.9
服务器中AI计算单元功耗（千瓦）	6.2	8.6	97.2	6.2	11.8
电源使用效率（PUE）	1.5	1.4	1.2	1.3	1.5
AI服务器电力功耗（千瓦）	9.3	12.0	116.6	8.1	17.7
AI集群电力功耗（千千瓦）	27.9	36.1	77.8	24.2	26.6
AI算力集群运行成本 (不包括网络和基建等)					
电价（千瓦时/美元）	0.10	0.10	0.10	0.10	0.10
耗电量 (万千瓦时)	587	334	324	1097	851
能源成本（万美元）	59	33	32	110	85
集群算力硬件折旧成本（万美元）	650	524	500	468	243
AI计算总成本（万美元）	**708**	**557**	**533**	**578**	**328**

资料来源：《Scaling Laws for Neural Language Models》，英伟达、谷歌、亚马逊官网，申万宏源研究
注：标红为非官方公布的假设数据，仅用以本表格进行计算

4.3 为什么ASIC增长趋势明显？从供需两端出发

- **上游供给：芯片设计制造分工化：**全球芯片设计制造分工化以及ASIC辅助设计的成熟，大幅降低了ASIC与GPU之间在代工制造、后端封装设计等领域的差距，差异集中在前端设计和软件开发生态。
- **云厂需求：1）技术/架构演进：牧本摆动**本质为针对通用芯片的算法演进迭代陷入停滞后，需要在架构上进行创新，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。当前正处于重要节点。**2）商业化驱动：**算力需求量级提升，具备庞大算力需求的厂商足以覆盖开发定制化芯片的成本。

图：云厂开发自研ASIC芯片已具备商业化、技术驱动力

上游供给：芯片设计制造行业的分工化

芯片制造与IC设计的分工：
半导体产业正从Sharing foundry到Sharing design house转变，共享代工厂+后端设计封装等链路。

AI ASIC辅助设计企业逐渐成熟：
博通、AIChip为代表的企业已经在TPU、Trainium的设计和制造中积累了丰富经验和制造流程管理能力。

云厂需求：技术/架构演进驱动

牧本摆动：
按需定制化设计算力的端到端方案在算力中的效能、利用率正体现优势

推理侧对算力芯片要求或更低：
大模型推理算法相对训练更固定，对于芯片性能、开发生态灵活性要求可略降低

云厂需求：商业化驱动

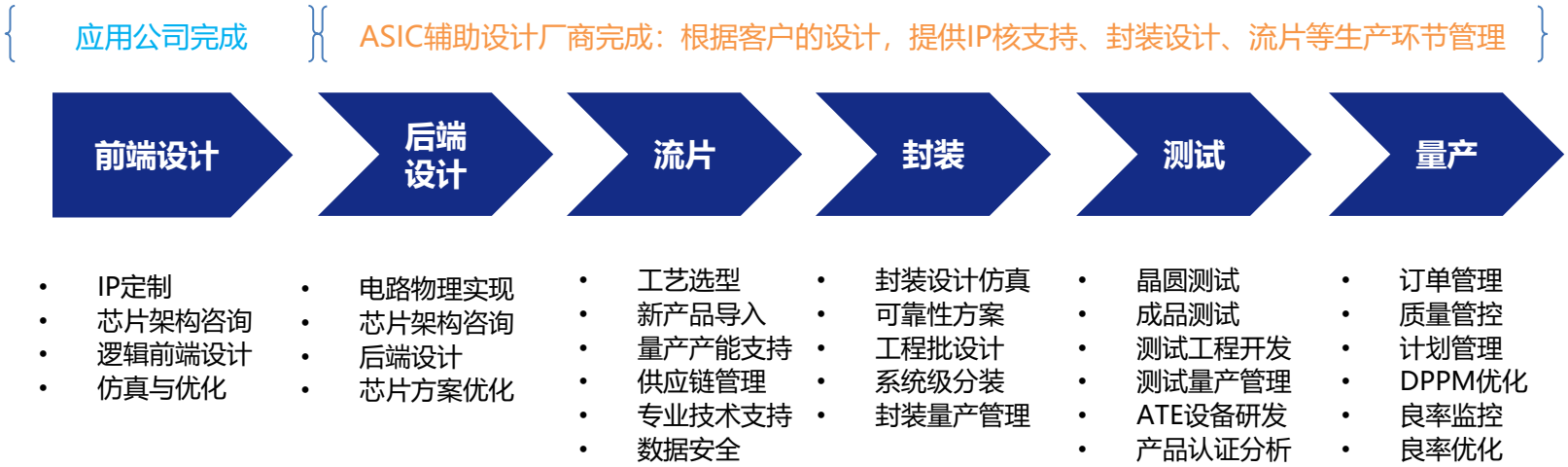
算力开支量级上升：
达到重新设计一款芯片的综合成本临界点。

单芯片的价格更低：
成本侧后续有望实现量产爬坡。

4.4 如何设计制造ASIC？云厂前端设计+IC辅助设计支持

- **云厂：仅维持相对精简的IC设计团队，无须困扰于庞杂的芯片制造流程。**云厂可根据自有业务场景的算力需求进行前端设计（逻辑设计、仿真验证等）等环节，并避免在主业之外形成庞大半导体业务部门。
- **IC辅助设计：提供后端设计、制造流片等环节支持。**博通、Marvell、AIChip通常提供IC设计所需的IP核等，并完成后端设计、封装测试设计等，全流程跟踪、优化晶圆厂代工流片的制造流程，最终向云厂交付。

图：ASIC芯片设计流程，辅助设计厂商可辅助完成后端设计到流片管理等环节



4.4 ASIC落地路线图：海外云厂硬件成本优化进行时

- **ASIC技术难度较大，目前已验证能力的产品主要为谷歌TPU。**亚马逊Tranium2进入大规模爬坡投产阶段，仍待Anthropic等厂商进一步验证性能。
- **ASIC开发周期通常为3-4年，2026年将进入密集落地期。**META指引25H2 推理卡MTIA系列将用于推荐系统等领域，26年将推出训练卡。根据路透社，OpenAI的ASIC芯片有望在25年流片测试，并计划在26年爬坡。另外，包括微软、ARM在内的多家公司已开启ASIC项目，后续将密集落地。

表：预计各公司ASIC芯片投产路线图

	2023		2024		2025E	
	1H	2H	1H	2H	1H	2H
亚马逊			**Trainium 2**			**Trainium 3**
存储			HBM3 96GB			HBM
制程			5nm			3nm
ASIC辅助设计			MRVL/AIchip			AIChip等
谷歌	**TPU v5e**	**TPU v5p**			**Trillium v6**	
存储	HBM2 16GB	HBM2e 96GB			HBM3 32GB	
制程	5nm	5nm			4nm	
ASIC辅助设计	博通	博通			博通	
Meta				**MTIA v1**		**MTIA v2**
存储				LPDDR5 64GB		LPDDR5 128GB
制程				7nm		5nm
ASIC辅助设计				博通		博通
微软			**Maia 100**			
存储			HBM3e 64GB			
制程			5nm			
ASIC辅助设计			创意电子			

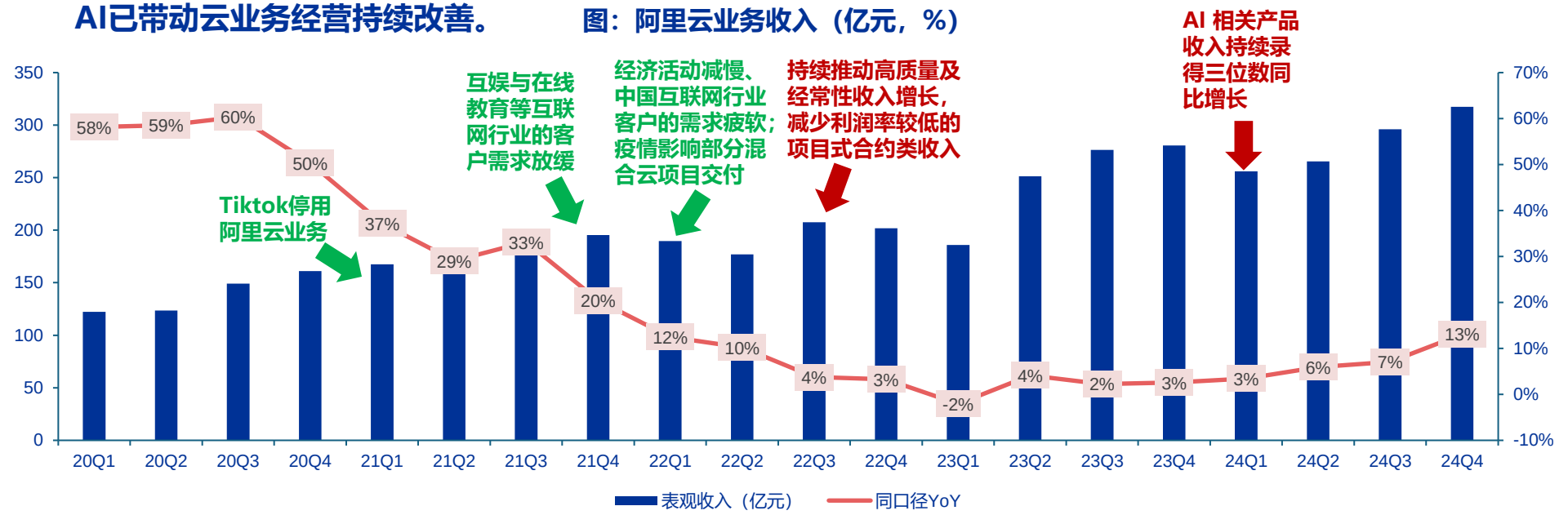
主要内容

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

5.1 阿里云：国内云厂领军，强技术能力，云增长高确定性

- 阿里云为国内云厂领军；集团CEO吴泳铭在23年底接管阿里云后，确定新战略为“AI 驱动，公共云优先”；AI已带动云业务经营持续改善。

图：阿里云业务收入（亿元，%）



图：阿里云调整后EBITA利润率（%）

资料来源：阿里巴巴财报，申万宏源研究

5.1 阿里云：国内云厂领军，强技术能力，云增长高确定性

- **阿里云核心竞争优势为强技术能力：1）芯片，全资半导体芯片业务主体——平头哥；2）模型能力国内一梯队。**

苹果确认选择阿里巴巴作为国内iPhone的AI合作商，再证阿里云强技术能力.

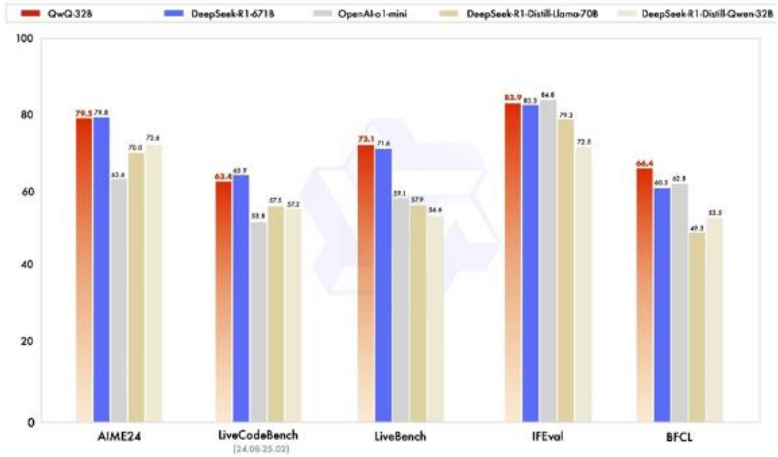
表：平头哥芯片产品

芯片名称	发布时间	芯片类型	关键参数	主要应用场景
含光800	2019.9	人工智能芯片	12nm制程，集成170亿晶体管，性能峰值算力达820，192M本地存储，采用自研TPU	云计算服务、电商智能搜索、电商营销
羽阵600	2021.1	RFID电子标签芯片	满足EPC Global Class-1 Generation-2 UHF RFID协议，读取灵敏度达-21dBm，96-bit出厂预编程 EPC区(只读)	智慧零售、智慧物流、航空包裹跟踪、库存管理
倚天710	2021.1	Arm服务器芯片	5nm制程，Armv9架构，128核心数，主频最高达3.2GHz	倚天云服务器、AI推理、大数据、视频编解码、电商
羽阵611	2022.11	RFID电子标签芯片	满足EPC global G2 V2和ISO/IEC 18000-6C协议，读取灵敏度-23dBm，写入灵敏度-20dBm，128-bit EPC，96-bit TID 永久锁定，32-bit 访问密码和灭活密码共享	鞋服、快消品零售、智慧物流、供应链管理、动态资产管理
镇岳510	2023.11	SSD主控芯片	12nm制程，IO处理能力达到3400K IOPS，数据带宽达到14GByte/s，能效比达到420K IOPS/Watt	电商、大数据、虚拟化、软件定义存储、边缘计算

图：2/25开源视频生成大模型万相2.1，VBench测评中分数超越Sora、Pika等位居榜首

Model Name (clickable)	Sampled by	Evaluated by	Accessibility	Date	Total Score
Wan2.1(2025-02-24)	Wan Team	VBench Team	API	2025-02-24	86.22%
MiracleVision V5	MVV Team	VBench Team	API	2025-01-21	85.23%
Wan2.1	Wan Team	VBench Team	API	2025-01-08	84.70%
Sora	VBench Team	VBench Team	API	2025-01-14	84.28%
CausVid(2025-01-02 5s)	CausVid Team	VBench Team	Close Source	2025-01-02	84.27%
CausVid	CausVid Team	VBench Team	Close Source	2024-12-07	83.88%
Luma	VBench Team	VBench Team	API	2025-01-14	83.61%
EasyAnimateV5.1	EasyAnimate Team	VBench Team	Open Source	2025-01-22	83.42%
MiniMax-Video-01	VBench Team	VBench Team	API	2024-10-01	83.41%
STIV (Apple)	Apple Team	VBench Team	Close Source	2024-12-19	83.35%
HunyuanVideo (Open-Source	VBench Team	VBench Team	Open Source	2024-12-16	83.24%
Gen-3 (2024-07)	VBench Team	VBench Team	API	2024-07-25	82.32%

图：3/6开源通义千问QwQ-32B模型，性能比肩671B参数的DeepSeeK-R1



资料来源：平头哥半导体官网，通义万相Wan公众号，阿里云公众号，申万宏源研究

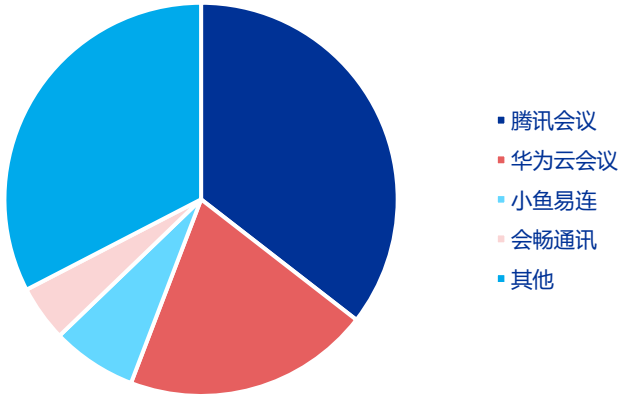
5.2 腾讯控股：云SaaS差异化竞争，AI应用后发优势突出

- 腾讯云业务自22年起转为聚焦高质量增长，资源集中于视频云、网络安全等PaaS，腾讯会议、企业微信等SaaS。**投资燧原+SaaS差异化竞争，将增强腾讯云竞争力：**
  - 腾讯为AI芯片公司燧原科技第一大股东，根据爱企查，持股比例为21%。
  - 腾讯办公SaaS产品与微信打通，具备强用户优势；23年起加强变现。

表：燧原科技大事记

	产品	融资
2018		Pre-A轮融资3.4亿，腾讯领投
2019	第一代训练产品云燧T10发布	A轮融资3亿，红点领投
2020	第一代推理产品云燧i10发布	B轮融资7亿，武岳峰领投
2021	第二代训练产品云燧T20/T21发布，第二代推理产品云燧i20发布	C轮融资18亿，中信产业基金、中金资本旗下基金、春华资本领投
2022	人工智能加速集群产品一云燧智算机发布	C+轮融资，国家集成电路产业投资基金投资
2023		D轮融资20亿，上海国际集团子公司及旗下基金领投
2024	新一代推理产品燧原S60发布	

图： 2023年中国云会议市场份额（%）



表：腾讯会议收费

	账号数量	定价	部分功能区别
免费版	1		单场会议40分钟限制
专业版	1-5	82.34元/月/账号，按年购买；或按每月98元购买	不限时会议，AI小助手Pro可提供会议纪要等功能
商业版	6-255	115.84元/月/账号，按年购买；或按每月139元购买	
企业版	≥256个	联系销售	
教育版		学生与公益人群特惠折扣	

注：23年4月，单场会议人数上限从疫情期300人缩减至100人，会议时长从“不限时”调整为60分钟（2人会议不限时）；24年10月，免费版会议时长从60分钟缩短至40分钟（仅2人会议不限时）。

资料来源： 燧原科技官网，IDC，腾讯会议官网，申万宏源研究

5.2 腾讯控股：云SaaS差异化竞争，AI应用后发优势突出

- **腾讯应用场景、数据优势突出，叠加云+推理算力（投资燧原）储备，有望在“开源模型+私有数据” AI应用构建阶段具备后发优势。**
- 腾讯目前已经有超过10个应用接入DeepSeek，包括社交搜索、办公提效、金融服务、娱乐内容等场景。微信超级入口提升粘性和变现价值，对广告游戏业务降本增效。

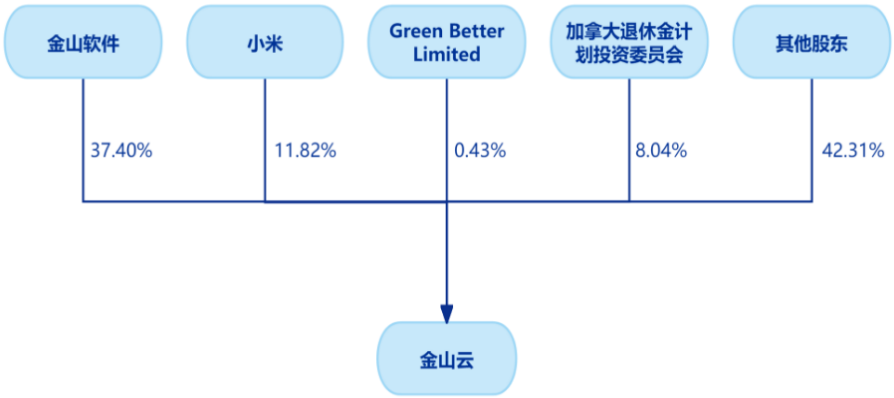
表：腾讯旗下接入AI大模型的相关应用及可实现的AI功能

类型	产品	大模型	接入AI后实现的功能
社交+搜索	腾讯元宝	混元+DeepSeek R1	日常通用AI搜索、总结
	微信搜一搜	混元+DeepSeek R1	日常通用AI搜索、总结
办公	腾讯文档	混元+DeepSeek R1	它可以生成文档、表格、幻灯片、思维导图、智能文档等
	ima.copilot	混元+DeepSeek R1	基于知识库的AI搜索、总结
	企业微信	混元+DeepSeek R1	包括AI Agent在内的多种AI生产力效率工具
	腾讯AI代码助手	混元+DeepSeek R1	AI代码生成、代码补全、代码解释、代码检查等
AI Agent平台	腾讯元器	混元+DeepSeek R1	可构建定制化的AI Agent
浏览器+AI Agent	QQ浏览器	混元+DeepSeek R1	支持深度思考、联网搜索、多轮对话、历史纪录回溯等，包括搜索、翻译、记笔记等功能
金融+AI助手	理财通	混元+DeepSeek R1	应用于理财通社区、智能客服等场景，可整合专业金融信息数据、微信公众号文章等资源
社交+AI助手	QQ音乐	混元+DeepSeek R1	精准的歌曲推荐、专业的音乐知识问答，AI助手都能提供更准确的回应。
游戏+AI	和平精英	DeepSeek R1	为数字代言人“吉莉”注入人工智能
地图+AI助手	腾讯地图	混元+DeepSeek R1	AI助手辅助规划形成、推荐景点美食等

5.3 金山云：金山小米生态核心云厂

- **雷军系的金山软件和小米集团为前两大股东，核心高管亦有交叉。**金山云22年管理层更换，邹涛先生接任CEO；一方面进行收入结构调整及人工智能转型，另一方面与金山软件、小米生态加强业务合作，成效明显。

图：金山云股东结构（截至2024年6月30日）

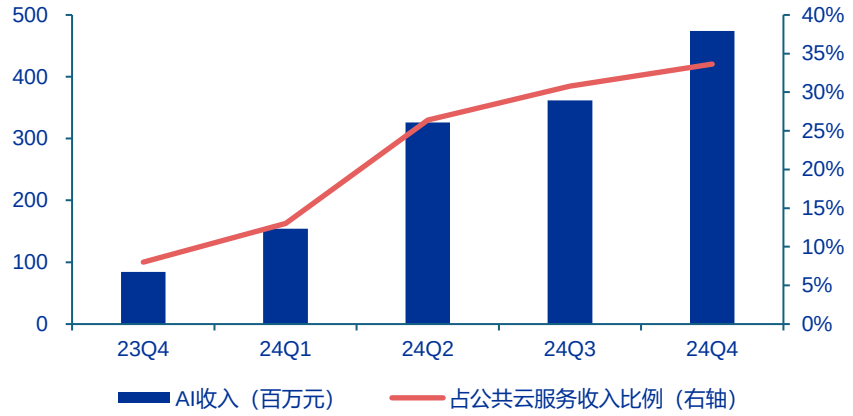


注：Green Better Limited为小米间接全资子公司

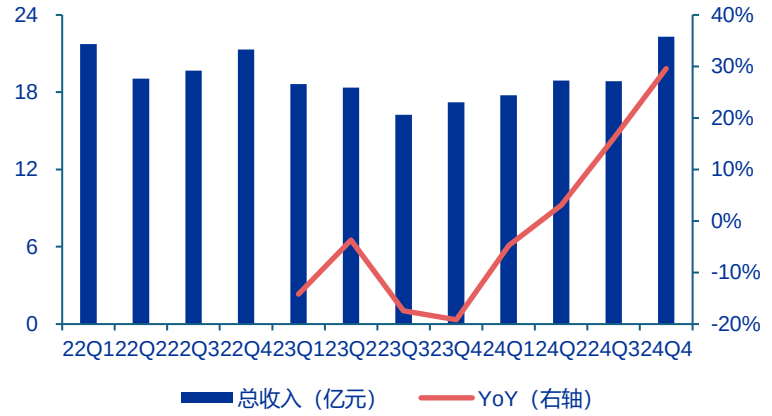
表：金山云核心高管

	雷军	邹涛
金山云	董事长、非执行董事	副董事长、执行董事、CEO
金山软件	董事长、非执行董事	执行董事、CEO
金山办公	董事	董事长
小米集团	创始人、董事长、CEO	

图：金山云AI收入（百万元，%）



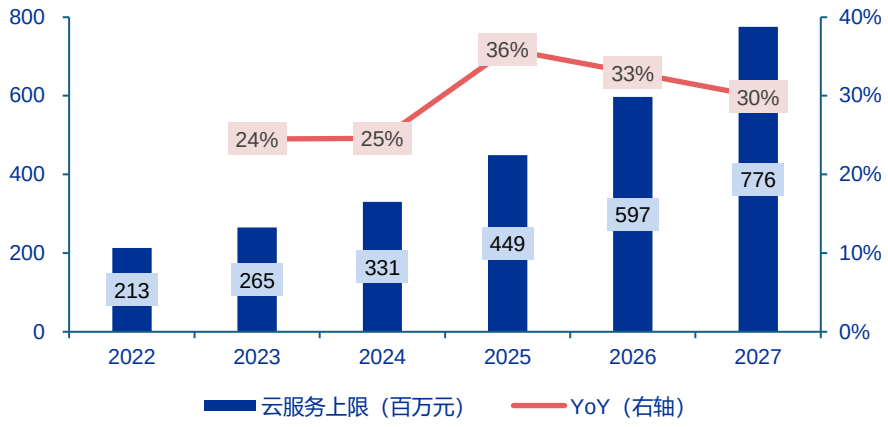
图：金山云总收入（亿元）



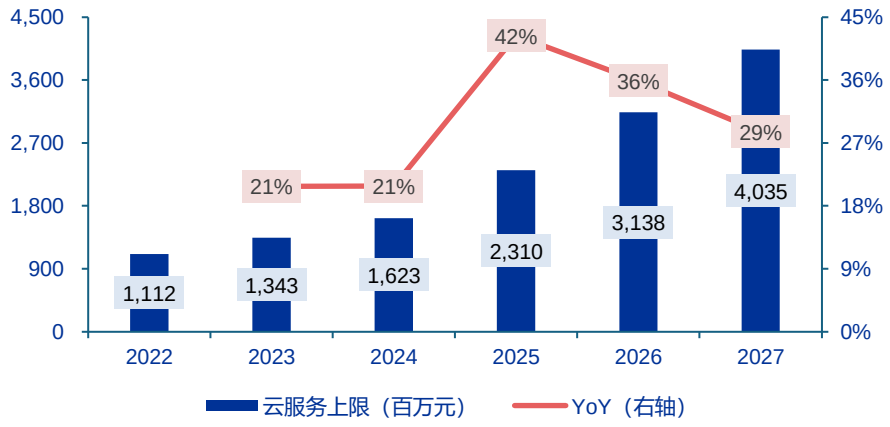
5.3 金山云：金山小米生态核心云厂

- **金山云与金山软件、小米集团续签25-27年战略框架协议，保证公司收入增长基本盘和资本开支投入。**

图：金山云与金山软件签订的云服务上限（百万元）



图：金山云与小米集团件签订的云服务上限（百万元）



表：金山云与小米集团2022-2024融资服务框架（亿元）

	2022	2023	9M24
小米集团所提供租赁融资之最高未偿还融资租赁结余及利息	7.5	8.1	7.2
上限	14	14	14 （2024年全年度）

表：金山云与小米集团2025-2027融资服务框架（亿元）

	2025	2026	2027
融资租赁（包括售后回租融资租赁及直接融资租赁）：最高未偿还结余	12	12	12
保理：最高未偿还结余	12	12	12
抵押贷款：最高未偿还结余	20	20	20

5.4 亚马逊：长于云基础设施能力，Anthropic补足模型层

- 亚马逊陆续向Anthropic投资80亿美元，并持续补足AI云能力，推出AmazonQ助手、BedRock模型平台、SageMaker等AI云服务，**布局重心聚焦于完善AI云基础设施，对上层SaaS应用的布局相对较少。**
- **自Graviton起，亚马逊长期坚持自研云芯片，AI时代正大力推动AI算力芯片自主化，**目前推进Rainier项目，为Anthropic构建数十万卡Tranium2集群，并计划于25Q4发布Tranium 3预览版本。

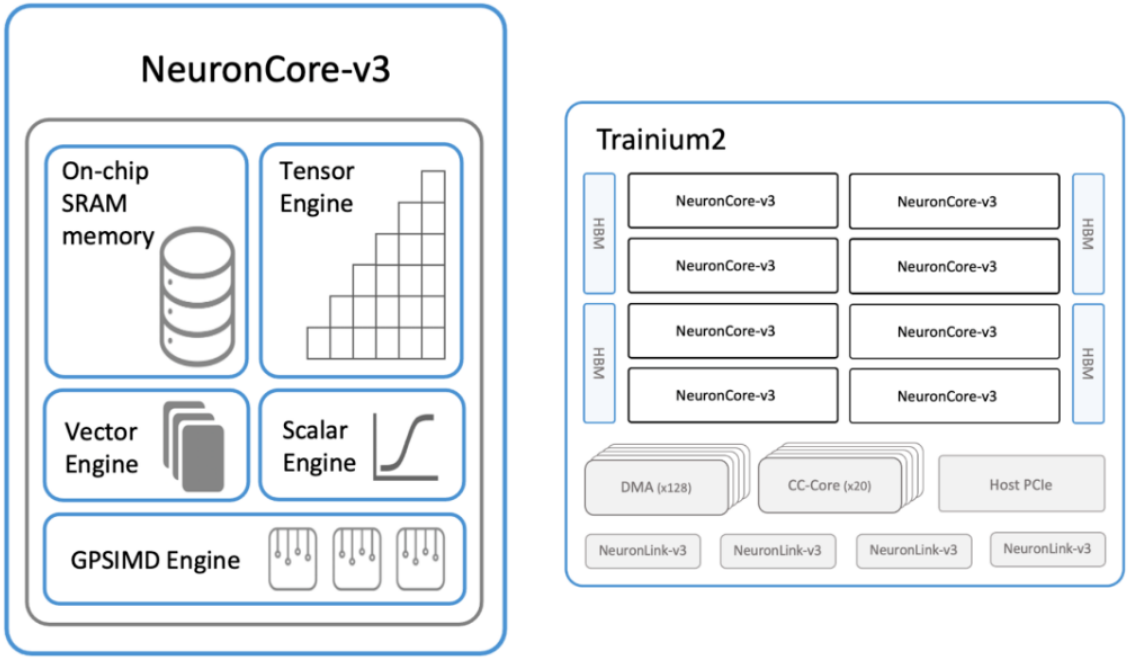
表：亚马逊AWS AI相关布局梳理

AWS布局层次	产品名称	具体产品	主要用途
应用层	AmazonQ	Q Developer	开发场景，研发提效，应用架构移植等
		Q Bussiness	商业场景，跨数据源链接和索引
模型层	Nova自研模型系列	Micro/Lite/Pro/Premier	基础大模型系列
		Nova Canvas	图像生成模型
		Nova Reel	视频生成模型
	BedRock模型平台	Claude、Llama3等	提供包括Anthropic Claude在内的第三方模型
数据层	SageMaker	SageMaker Unified Studio	统一的数据和AI开发环境
		SageMaker LakeHouse	数据编织服务，提供数据接入和访问控制
		SageMaker HyperPod	创建训练计划，高效跨团队、跨项目利用计算资源
基础设施层	自研芯片系列	Tranium	系列训练芯片
		Inferentia	系列推理芯片
		Graviton	系列通用CPU芯片
		Nitro	系列网卡

5.4 亚马逊Tranium2：AWS算力自主化的重要产品

- **AWS Trainium2主要由8个NeuronCore组成，每个NeuronCore包括：1）张量引擎：**128×128 脉动阵列核心，与TPUv4的MXU规格类似，主要用于加速矩阵乘法计算。**2）向量引擎：**用于向量加速计算。**3）标量引擎：**执行科学计算等标量计算；**4）GPSIMD引擎：**可编程计算单元，基于C++，可由开发人员进行自定义运算。**张量引擎为算力核心来源，每个NeuronCore中，张量引擎提供79 FP16 Tflops，向量引擎提供1 FP16 TFLops，标量引擎则提供1.2 FP16 TFlops。**

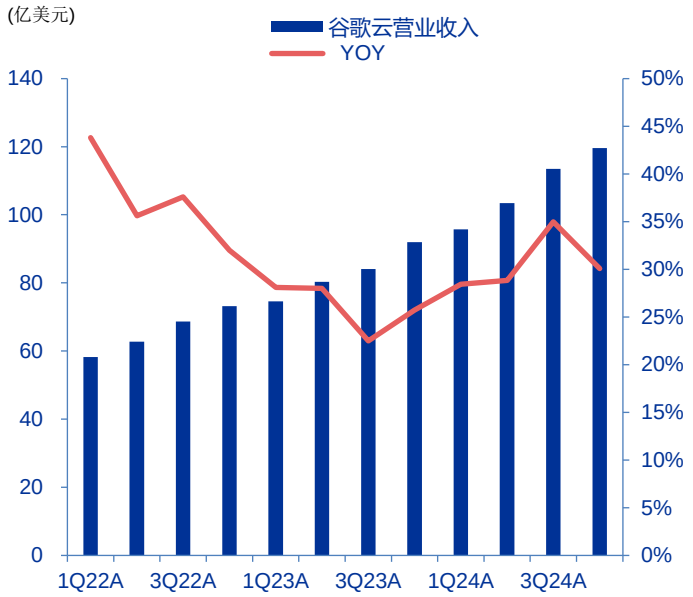
图：亚马逊Trainium2架构，核心算力单元NeuronCore-v3



5.5 谷歌：拥有从芯片、公有云到应用的AI全生态栈

- 谷歌拥有芯片硬件TPU、谷歌云、大模型Gemini、AI应用的AI全生态栈，相比其他海外互联网/云巨头布局更完整。关注：1）谷歌搜索AI化后竞争力能否提升：2）大模型竞赛下，能否在C端应用、尤其AI Agent领域取得明确进展。
- 谷歌具备优秀的AI Infra能力，云业务在公司营收占比已从22Q1的8.6%上升至24Q4的12.4%。

图：谷歌云营收及同比增速



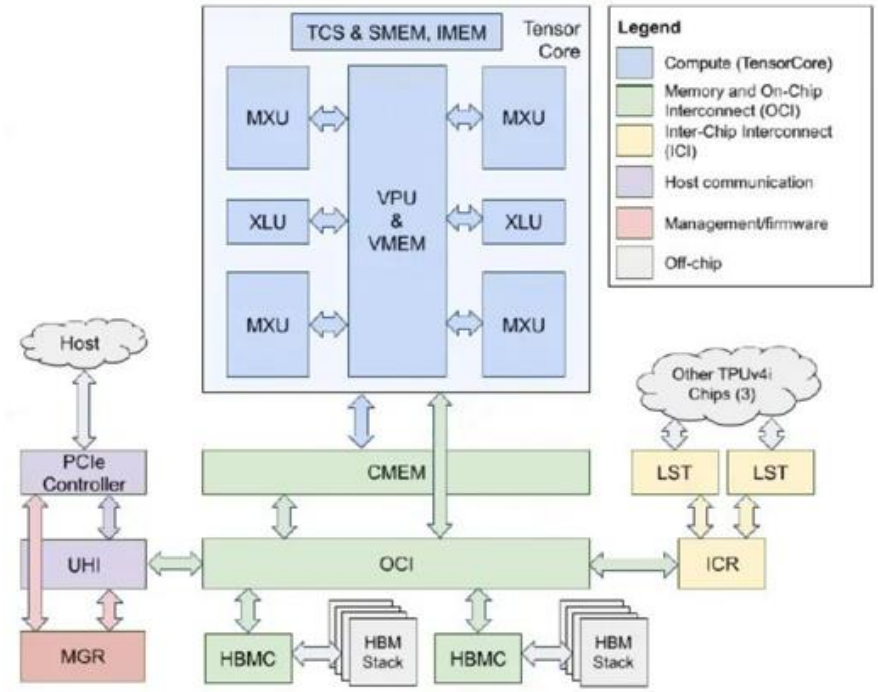
图：谷歌AI布局全工程栈

资料来源： 谷歌官网，谷歌财报，申万宏源研究

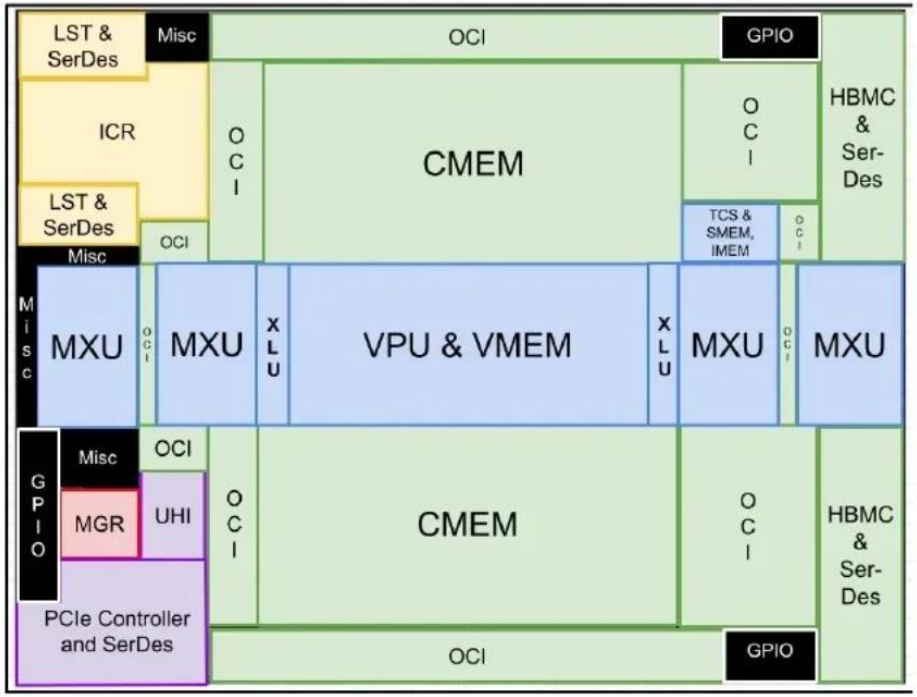
5.5 谷歌TPU：大型脉动阵列核心，专用于矩阵计算

- **以谷歌TPUv4i架构为例，核心算力单元均为矩阵计算而设计，非矩阵计算能力相对有限。**TPU v4i为推理卡，训推卡为TPUv4，**TPUv4包含2个TensorCore。**TPU v4i TensorCore包括：**1）4×MXU：**Matrix Multiple Unit，128×128 大型脉动阵列单元，专用于执行大型矩阵计算；**2）1×VPU：**专用于执行向量计算，128通道，每个通道16个ALU；**3）XLU：**用于执行交叉计算；
- **TPU的核心为MXU，与向量引擎VPU一同占据了芯片的大部分算力单元面积，相比英伟达H100拥有576个小型并行TensorCore，TPU的MXU为规格更为庞大的大型AI加速单元，以针对大型矩阵计算需求。**

图：TPU v4i架构，脉动阵列MXU单元为算力核心部分



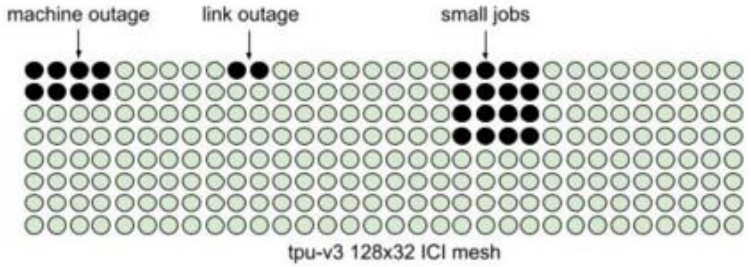
图：TPU v4i拓扑结构，MXU及VPU



5.5 谷歌TPU：大规模弹性部署，集群实现高可用性

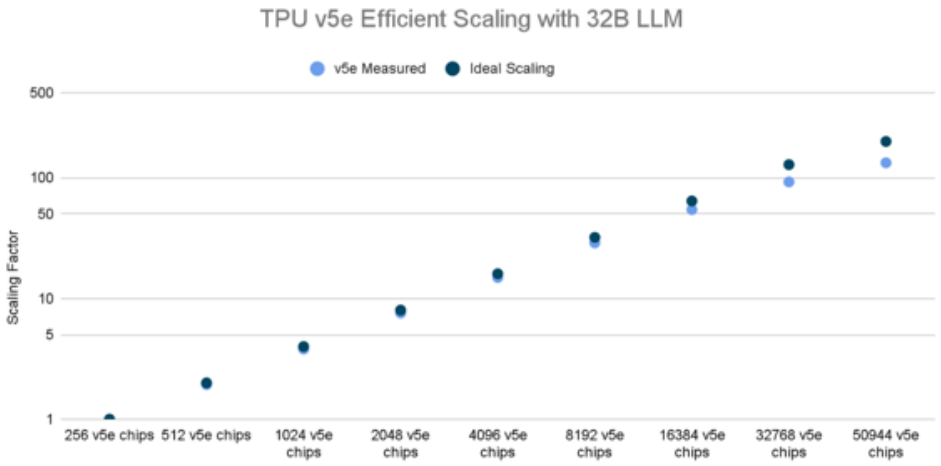
- **自研芯片优势+AI Infra能力加持，谷歌TPU在集群性能上有着较强积淀。**谷歌TPUv3集群为计算资源静态配置，计算集群算卡数量增加后可用性将快速下降，导致集群低效。自TPU v4开始，大规模动态配置下实现ICI容错管理，降低了集群因为错误而导致的性能损失。TPU v5e可实现的动态大规模弹性部署，使得集群在5万卡时接近线性性能提升。

图：谷歌TPUv3计算资源为静态配置



图：谷歌TPUv4动态配置可解决小型计算任务、芯片故障等带来的计算性能下降，保证集群高可用性

图：谷歌TPU v5e可实现5万卡集群接近线性性能提升



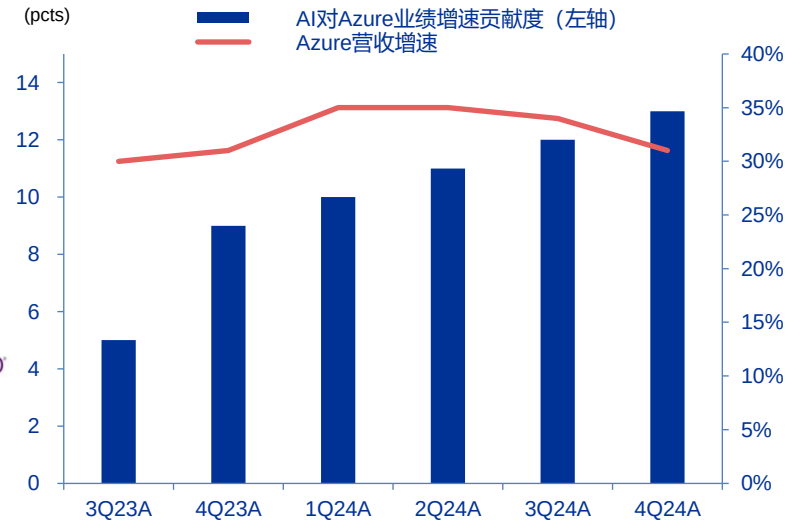
资料来源：谷歌官网，《Resiliency at Scale: Managing Google’s TPUv4 Machine Learning Supercomputer》，申万宏源研究

5.6 微软：核心看B端AI SaaS的落地

- **微软Azure布局AI较早，AI收入持续兑现，CY24Q4 AI在Azure增速的贡献度已达到13pcts。**
- **微软的禀赋优势在于SaaS业务庞大的云租户客群，**尽管与OpenAI的合作关系紧密度有所降低，但开源模型的繁荣，将填补微软在自研大模型领域的劣势。
- **微软已推出Azure OpenAI服务、M365 Copilot、多个AI Agent等多个AI产品，企业级AI卡位明确。**

图：微软AI布局，核心看AI SaaS落地

图：AI对Azure云业务增速贡献持续提升

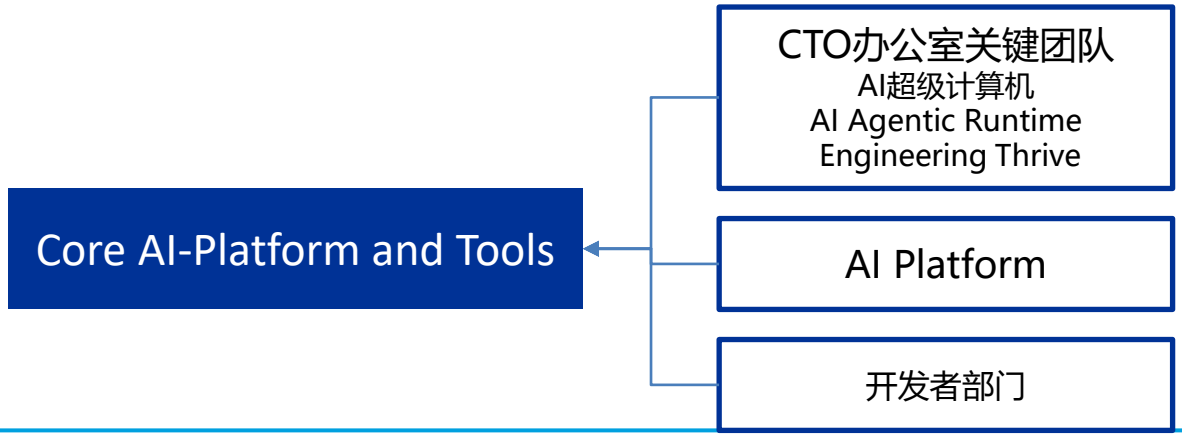


资料来源： 微软官网，微软财报，申万宏源研究
注：时间轴为自然年时间

5.6 微软：成立CoreAI，打造Copilot&AI stack

- **微软CEO：AI agent将重塑SaaS**
  - SaaS 应用程序将从简单的 CRUD 数据库演变为由AI agent编排的业务逻辑层，AI agent跨多个 SaaS 应用程序协调工作，无需直接访问单个应用程序的数据。（纳德拉25年1月访谈）
- **An Agentic world有赖于AI三种呈现指数级增长的能力：**多模态通用界面、推理和规划能力、记忆和丰富的上下文能力（微软2024）
- **组织架构变革：25年1月成立CoreAI-平台和工具，目标打造Copilot 和 AI 堆栈，以支持各类 AI 应用程序和AI Agents构建与运行。**
  - 整合微软的开发者部门、AI平台、CTO办公室的关键团队，由Meta 前工程主管Jay Parikh 领导
  - 突破组织边界，促进跨部门协作；
  - 构建平台-工具-基础设施的反馈链路，推动 AI 技术和产品迭代

图：微软成立CoreAI-平台和工具



资料来源：微软官网，申万宏源研究

5.6 微软：成立CoreAI，打造Copilot&AI stack

- **微软Copilot&AI堆栈方便用户构建自己的Copilot和智能代理：推出Azure AI Foundry一站式创建AI应用，在Microsoft 365、Dynamics等集成AI Agent服务。**
- **在订阅模式之外推出即用即付模式：**
  - 办公SaaS及AI升级后的Copilot产品主要采用经典的月付费模式，Microsoft Copilot Studio在包月模式外提供了即用即付模式-每条消息0.01美元。

表：微软Copilot产品及定价

产品	介绍	商业化进展
Microsoft 365 Copilot Chat	OpenAI旗舰模型GPT-4o支持的免费安全AI聊天；使用自然语言创建Agent来自动化重复性业务流程；IT控制，包括企业数据保护和智能体管理。	即用即付，每条消息收费0.01美元，25年1月推出。
Microsoft 365 Copilot	面向企业版Microsoft 365办公套件中使用AI助手提高效率，Copilot Studio：帮助构建企业自定义AI助手。	30 美元/月；23年11月1日开始正式商业化。
Copilot pro	面向个人用户，可自定义GPT助手	20美元/月，24年1月15日推出。
Copilot for Sales	CRM领域的智能助理，客户洞察&分析、个性化互动，简化销售流程	50美元/月，24年2月开始推出
Copilot for Service	打造智能化客户服务中心，还能智搜全网内容，辅助客服团队为客户提供个性化回复。	50美元/月，24年2月开始推出
Microsoft Copilot Studio	面向企业用户，图形化、低代码工具，可用自然语言创建智能助理	200美元/月（含25000信息）+即用即付模式，每条消息0.01美元
Github Copilot	使用自然语言提示的代码在代码编辑器中实时获取由 AI 生成的数十种语言的代码建议。	个人版：每用户按月订阅 10美元/月，按年订阅 100 美元/年；商业版：每用户19 美元/月，2022年6月开始商业化

表：微软Copilot&AI Stack

开发者工具及应用服务	Github Copilot Microsoft 365 Copilot Copilot for Sales/Service
Azure AI Foundry	用于构建AI应用，1800+模型； Azure AI agent服务 Azure AI Studio-构建部署AI解决方案
数据	Fabric数据管理分析
基础设施	60多个数据中心区域 自研芯片Maia+外购英伟达AMD芯片

资料来源：微软官网，申万宏源研究

5.7 META：AI赋能广告已落地，关注新的AI应用孵化

- **对互联网巨头来看，AI对广告效率改进为目前核心落地方式，META 推荐系统已部署于AI基础设施上。**
- **AI未来增量需关注社交生态孵化的爆款AI应用，**META AI助手的月活用户在FY24Q4已达到7亿。

表：META 各季度法说会对于AI推荐系统与Advantage+的表态，AI正深入META广告产品线

财季	AI推荐系统/AI应用	Advantage+
22Q2	**AI会推荐大约15%的Facebook动态内容，以及略多于15%的 Instagram动态内容，**这些内容来自用户不关注的人、群组或账户。我们预计到23年年底，这些数字将增加一倍以上。	
22Q3		最近对一批广告商进行的一项测试发现，使用 Advantages+ Shopping广告系列的广告商的广告支出回报率提高了32%。
23Q1	Facebook动态消息中超过20%的内容是由AI从用户不关注的人群或帐户推荐的。在整个Instagram上，这大约占您看到内容的40%。 自我们推出 Reels 以来，AI 推荐已使Instagram用户使用时间增加了24%以上。 我们的AI工作也提高了盈利能力。与上一季度相比，Reels在 Instagram上的盈利效率提高了30%以上，在Facebook上的盈利效率提高了40%以上。	过去六个月，Advantage+购物活动的每日收入增长了7倍。
23Q2	自从推出AI推荐以来，它们已经**推动了平台总使用时间增加7%。** **我们还部署了Meta Lattice，**这是一种新的模型架构，可以学习预测各种数据集和优化目标中的广告效果。	
23Q3	仅今年一年，由于推荐改进，我们使用户在Facebook上花费的时间就增加了7%，在Instagram上花费的时间增加了6%。	为广告商提供的AI工具推动了Advantage+购物广告系列的成果，达到了100亿美元的年化收入，超过一半的广告商使用我们的 Advantage+创意工具来优化其广告创意中的图片和文字。
23Q4	我们看到Reels和视频整体持续增长，因为在第四季度，所有视频类型的每日观看时间同比增长超过25%，这得益于排名的持续改善。	
24Q1	**Facebook信息流中约30%的帖子是由我们的AI推荐系统提供的。Instagram上人们看到的50%以上的内容都是AI推荐的。** **最近，我们一直在开发一种新的模型架构，**旨在为多种推荐产品提供支持。23年，我们开始部分验证该模型，使用它来支持 Facebook Reels，部署该模型后，**观看时间增加了8%到10%。**	根据我们进行的测试，使用 Advantage+受众定位的广告系列的每次点击费用或每个目标费用平均下降了28%。
24Q2	在Facebook上，我们于6月在全球推出统一的视频播放器和排名系统，看到了令人鼓舞的早期成果。	今年进行的一项研究表明，美国广告客户在采用 Advantage+购物广告系列后，广告支出回报率提高了22%。
24Q3	META AI助手MAU已经达到5亿； AI 驱动信息流和视频推荐仅在今年就使Facebook上的用户使用时间增加了 8%，Instagram上的用户使用时间增加了6%。	仅24年9月，超过一百万广告商使用我们的 Gen AI 工具制作了超过 1500 万条广告，我们估计使用图像生成的公司的广告转化率增加了 7%。
24Q4	META AI助手MAU已经达到7亿；	Advantage+ 购物活动的采用率持续增长，第四季度收入超过 200 亿美元的年运行率，同比增长 70%。

资料来源：META官网，申万宏源研究
注：内容均取自META季报公开法说会表述，仅作翻译后参考，具体以公司实际表述为准

5.8 博通：高确定性，AI ASIC合作伙伴已达到7个

- **博通为ASIC芯片辅助设计领先厂商，在AI ASIC 的IP核生态、后端设计、封装能力上具备明显优势。**
- FY25Q1重申FY2027三大现有客户AI SAM（可服务潜在市场规模）达600-900亿美元，有望对应博通超500亿美元AI收入。公司披露除目前三大现有客户外，AI ASIC业务已有四个新合作伙伴，预计包括OpenAI、Arm等潜在用户。

图：博通拥有完整的ASIC开发核生态

Architecture &Design	System Knowhow	Power Efficiency	Design for Cost	MCM/2.5D /3D Packaging	Fast Ramp (NPI)	Low DPPM Quality
Processing	ARM CPUs	DSP Cores	Offload Engines	Packet Processing	SwitchingCores	Security
Memory and Protocol	Multiport SRAMs	HBM I/F Cores	DDR I/F Cores	PCIe/CXL Cores	SAS/SATA Cores	Ethernet Cores
Signal Processing	Analog VGA	High Res ADC/DACs	High-Rate ADC/DACs	Modulation	Filtering	Error Correction
Connectivity	PCIe SerDes	SAS/SATA/FC SerDes	EthernetSerDes	VSR SerDes	MCM Chiplet	High B/W Chiplet

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

6.1 重点公司估值表

表：重点公司估值表（美股标的单位：亿美元）

标的	代码	总市值	营业收入			净利润			PS			PE		
			2024E	2025E	2026E	2024E	2025E	2026E	2024E	2025E	2026E	2024E	2025E	2026E
微软	MSFT.O	28,990	2,768	3,146	3,602	982	1,121	1,299	10	9	8	30	26	22
谷歌	GOOGL.O	20,235	3,442	3,763	4,143	1,149	1,293	1,492	6	5	5	18	16	14
亚马逊	AMZN.O	21,315	6,994	7,717	8,489	774	936	1,133	3	3	3	28	23	19
脸书	META.O	15,480	1,886	2,146	2,420	653	751	874	8	7	6	24	21	18
苹果	AAPL.O	33,278	4,093	4,416	4,677	1,099	1,200	1,293	8	8	7	30	28	26
英伟达	NVDA.O	27,757	2,047	2,539	2,903	1,119	1,409	1,552	14	11	10	25	20	18
博通	AVGO.O	8,429	625	730	830	326	385	449	13	12	10	26	22	19
港股标的（单位：亿人民币）														
腾讯控股	0700.HK	43,226	7199	7794	8333	2513	2822	3113	6	6	5	17	15	14
阿里巴巴	9988.HK	22,816	10014	10999	12032	1559	1773	1902	2	2	2	15	13	12
金山云	3896.HK	280	93	110	132	(5)	(1)	5	3	3	2	-	-	-

资料来源：Wind，Bloomberg，申万宏源研究；
注：美股标的市场数据选自2025/3/28盘后；收入和净利润预测来自彭博一致预期，博通利润为Non-Gaap净利润，其余为GAAP净利润
注：港股标的市场数据选自2025/3/28盘后；腾讯、阿里巴巴收入和净利润预测来自申万宏源研究，金山云来自彭博一致预期，利润均为Non-Gaap净利润

6.2 风险提示

- **大模型性能进步不及预期。**目前中美大模型技术仍存在差异，部分技术尚处于早期实验室阶段。
- **内容和互联网平台监管环境变化风险。**行业监管政策边际变化风险，互联网和文化产业受到国家相关政策和职能部门监管，因此国内外相关监管政策如果收紧可能会影响相关公司的经营。
- **AI应用落地进展不及预期风险。**AI算力需求受到下游应用实际落地进展影响。

信息披露

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告,并对本报告的内容和观点负责。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

与公司有关的信息披露

本公司隶属于申万宏源证券有限公司。本公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司关联机构在法律许可情况下可能持有或交易本报告提到的投资标的，还可能为或争取为这些标的提供投资银行服务。本公司在知晓范围内依法合规地履行披露义务。客户可通过compliance@swsresearch.com索取有关披露资料或登录www.swsresearch.com信息披露栏目查询从业人员资质情况、静默期安排及其他有关的信息披露。

机构销售团队联系人

华东A组	茅炯	021-33388488	maojiong@swhysc.com
华东B组	李庆	18017963206	liqing3@swhysc.com
华北组	肖霞	15724767486	xiaoxia@swhysc.com
华南组	张晓卓	13724383669	zhangxiaozhuo@swhysc.com
华北创新团队	潘烨明	15201910123	panyeming@swhysc.com
华东创新团队	朱晓艺	18702179817	zhuxiaoyi@swhysc.com
华南创新团队	夏苏云	13631505872	xiasuyun@swhysc.com

A股投资评级说明

证券的投资评级：

以报告日后的6个月内，证券相对于市场基准指数的涨跌幅为标准，定义如下：

买入（Buy）	：相对强于市场表现20%以上；
增持（Outperform）	：相对强于市场表现5%～20%；
中性（Neutral）	：相对市场表现在－5%～＋5%之间波动；
减持（Underperform）	：相对弱于市场表现5%以下。

行业的投资评级：

以报告日后的6个月内，行业相对于市场基准指数的涨跌幅为标准，定义如下：

看好（Overweight）	：行业超越整体市场表现；
中性（Neutral）	：行业与整体市场表现基本持平；
看淡（Underweight）	：行业弱于整体市场表现。
本报告采用的基准指数	：沪深300指数

港股投资评级说明

证券的投资评级：

以报告日后的6个月内，证券相对于市场基准指数的涨跌幅为标准，定义如下：

买入（BUY）：	：股价预计将上涨20%以上；
增持（Outperform）	：股价预计将上涨10-20%；
持有（Hold）	：股价变动幅度预计在-10%和+10%之间；
减持（Underperform）	：股价预计将下跌10-20%；
卖出（SELL）	：股价预计将下跌20%以上。

行业的投资评级：

以报告日后的6个月内，行业相对于市场基准指数的涨跌幅为标准，定义如下：

看好（Overweight）	：行业超越整体市场表现；
中性（Neutral）	：行业与整体市场表现基本持平；
看淡（Underweight）	：行业弱于整体市场表现。
本报告采用的基准指数	：恒生中国企业指数（HSCEI）

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。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。申银万国使用自己的行业分类体系，如果您对我们的行业分类有兴趣，可以向我们的销售员索取。

法律声明

本报告由上海申银万国证券研究所有限公司（隶属于申万宏源证券有限公司，以下简称“本公司”）在中华人民共和国内地（香港、澳门、台湾除外）发布，仅供本公司的客户（包括合格的境外机构投资者等合法合规的客户）使用。本公司不会因接收人收到本报告而视其为客户。有关本报告的短信提示、电话推荐等只是研究观点的简要沟通，需以本公司http://www.swsresearch.com网站刊载的完整报告为准，本公司并接受客户的后续问询。本报告首页列示的联系人，除非另有说明，仅作为本公司就本报告与客户的联络人，承担联络工作，不从事任何证券投资咨询服务业务。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为作出投资决策的惟一因素。客户应自主作出投资决策并自行承担投资风险。本公司特别提示,本公司不会与任何客户以任何形式分享证券投资收益或分担证券投资损失，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。市场有风险，投资需谨慎。若本报告的接收人非本公司的客户，应在基于本报告作出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。

简单金融·成就梦想

A Virtue of Simple Finance

申万宏源研究微信订阅号

申万宏源研究微信服务号

上海申银万国证券研究所有限公司
(隶属于申万宏源证券有限公司)