ASIC芯片崛起：云厂、芯片设计及光互连投资机遇

行业研究 · 行业专题

互联网 · 互联网II

投资评级：优于大市（维持）

证券分析师：张伦可
0755-81982651
zhanglunke@guosen.com.cn
S0980521120004

证券分析师：张昊晨
zhanghaochen1@guosen.com.cn
S0980525010001

报告摘要

- **模型商业模式逐步闭环，算力成为核心生产要素。**2026年以来，模型最大的变化是跨越了可持续执行任务拐点，编程能力、长程任务、工具调用成为模型重点提升的方向。随着模型价值量不断释放，推理需求进入快速增长阶段，Anthropic或将实现单季度盈利标志着模型公司商业模式逐步形成闭环，算力正成为企业创造价值的核心生产要素。
- **伴随推理侧需求爆发，各公司自研芯片能力验证成熟，ASIC出货迎来加速期。**我们预计26年ASIC芯片出货量大约770万片，份额为45%，并将在27年超过GPU份额达到58%。对比来看，英伟达GPU拥有更好的TPS，但是从TCO或者Cost Per Token角度，ASIC的优势更显著，因此也更加适合推理场景。
- **自研ASIC正在成为云厂商构建竞争优势的重要抓手。**一方面，自研芯片有助于缓解高端GPU供给约束，提升算力供给能力；另一方面，相较于外部采购GPU，自研ASIC能够显著优化成本结构，提升云业务盈利能力。随着自研芯片成熟度持续提升，我们预计具备ASIC能力的云厂商有望在收入增长与利润释放两个维度持续受益。

① **谷歌：自研芯片进展领先同业，26年起对外出售，加速放量带动云收入增长。**自研芯片已推动26Q1云收入同比+63%，预计26/27年出货量达到500/1500张，并于26年开始正式对外出售，预计27年贡献云收入占比达到18%。

② **亚马逊：26年起Trainium外部客户认可度提升，芯片业务整体ARR超过200亿美元，带动云收入加速。**26年起Anthropic、OpenAI等更多外部客户逐步与公司签订合同，表明芯片能力已达到成熟可用阶段，伴随出货量持续增长，我们测算Trainium将带动AWS27年约4%的收入增速。

③ **阿里巴巴：真武系列芯片已累计出货56万片，26年云业务有望迎来利润率拐点。**

报告摘要

- **ASIC崛起对产业链投资的启示：1）ASIC设计与制造产业链。**随着ASIC需求快速增长，云巨头正在逐步收回芯片设计主导权，并主动推动供应链多元化，其核心目标是实现更快迭代、更快交付以及更低TCO。例如Google正在弱化对博通的依赖，引入联发科、Marvell等设计伙伴；AWS也在持续优化Trainium产业链分工，通过提升自主掌控能力降低成本。**2）网络与互联基础设施。随着ASIC规模化部署，产业瓶颈正逐步从算力转向连接。**①Scale-up方向，交换芯片、CPO等技术有望受益于柜内连接升级。以Trainium3为代表的新一代机柜方案，NeuronSwitch-v1带来了大量的增量PCIe芯片需求。②Scale-out方向，Google采用的OCS光交换方案以及持续扩大的大规模网络架构，均将带动OCS芯片、光模块以及相关光通信器件需求增长。

① **AVGO：同时受益于ASIC放量和连接需求增加，但ASIC份额受冲击。**博通此前在高端ASIC设计领域拥有超过80%市场份额，充分受益于TPU放量以及Meta、OpenAI等公司的ASIC需求，但由于谷歌分散供应链后续份额将受到冲击。

② **MRVL：ASIC放量、XPU Attach崛起、连接需求爆发。**公司在ASIC设计层面尽管面临丢失Trainium份额的风险，但积极与谷歌洽谈合作同时大力发展XPU Attach业务。同时公司在光互联 DSP 市场占据绝对主导地位，充分受益于连接需求增加。

③ **COHR：核心受益于光模块需求增加。**从800G到1.6T以及后续的3.2T，ASP和需求量持续增长。

④ **LITE：OCS贡献弹性，光模块需求放量。**公司作为谷歌的核心供应商，目标是在2027年实现10亿美元的OCS芯片年化销售额。由于原本收入体量有限，因此TPU爆发释放的OCS需求有望充分贡献业绩弹性。

⑤ **ALAB：PCIe交换芯片放量。**从PCIe Retimer（Aries）单品垄断向PCIe Switch（Scorpio）及光学连接平台跨越，PCIe交换芯片成为重要增量。

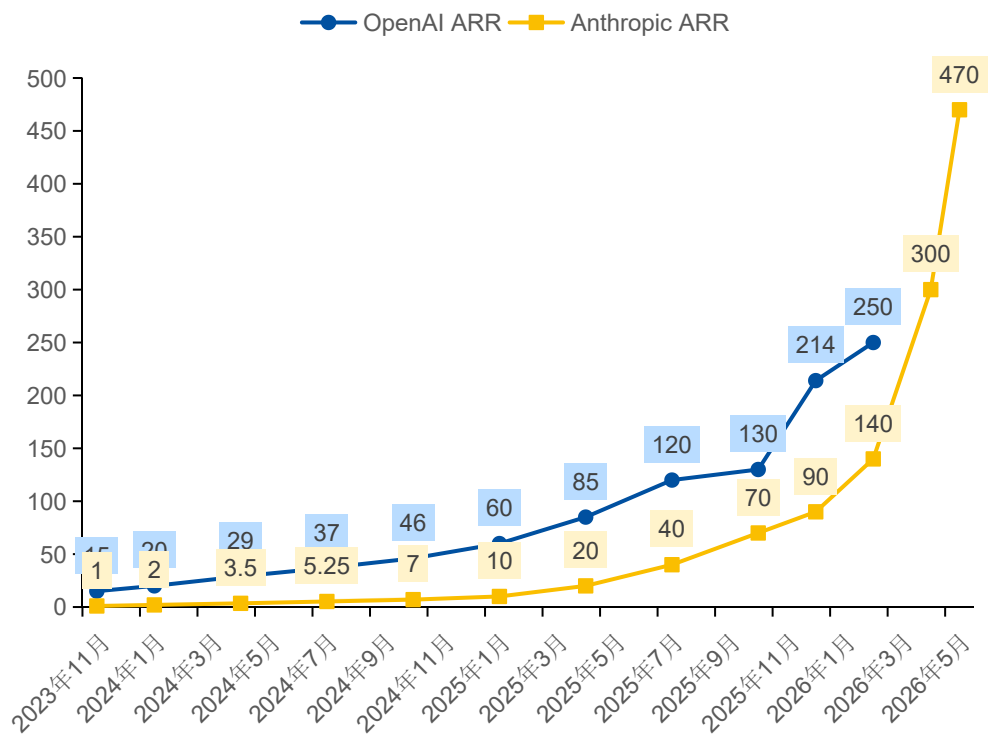
风险提示：宏观经济波动、下游需求不及预期、核心技术水平升级不及预期的风险、AI快速迭代平权化下竞争加剧。

目录

Anthropic或实现单季度盈利，模型厂商迎来商业化闭环

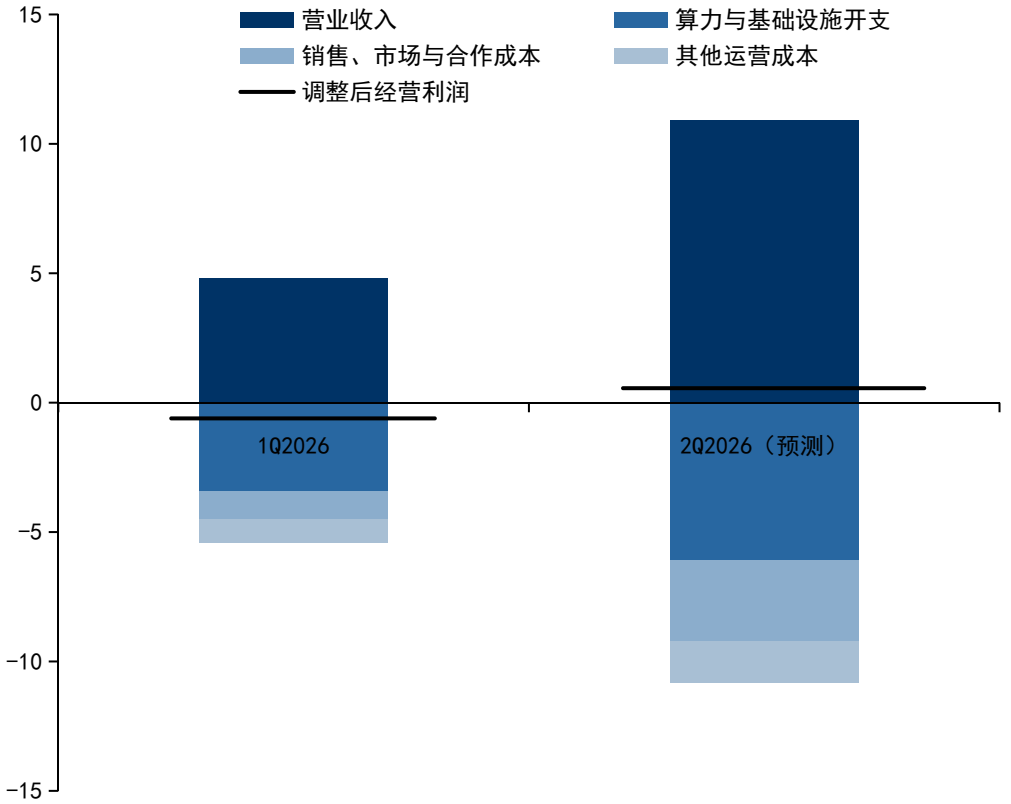
- **Anthropic 26Q2或将实现盈利，ARR接近500亿美元。**根据The information和WSJ的数据，26Q1 Anthropic收入48亿美元，OpenAI 57亿美元，但Q2预计Anthropic将达到109亿，并且实现单季度OP盈利（5.6亿），而根据公司官网，5月初其ARR已达到470亿美元。

图：OpenAI和Anthropic ARR变化情况（亿美元）



资料来源：openai、Anthropic、国信证券经济研究所整理

图：WSJ预计Anthropic或将实现单季度经营利润转正

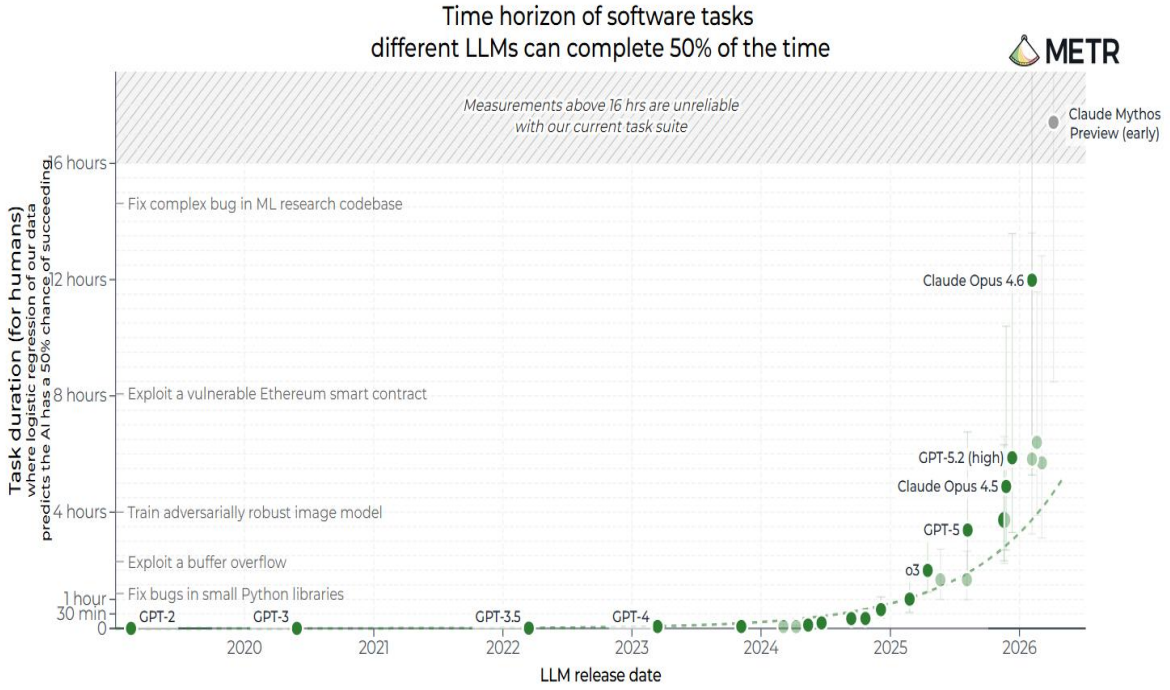


资料来源：WSJ、国信证券经济研究所整理

模型能力：跨过可持续执行的拐点

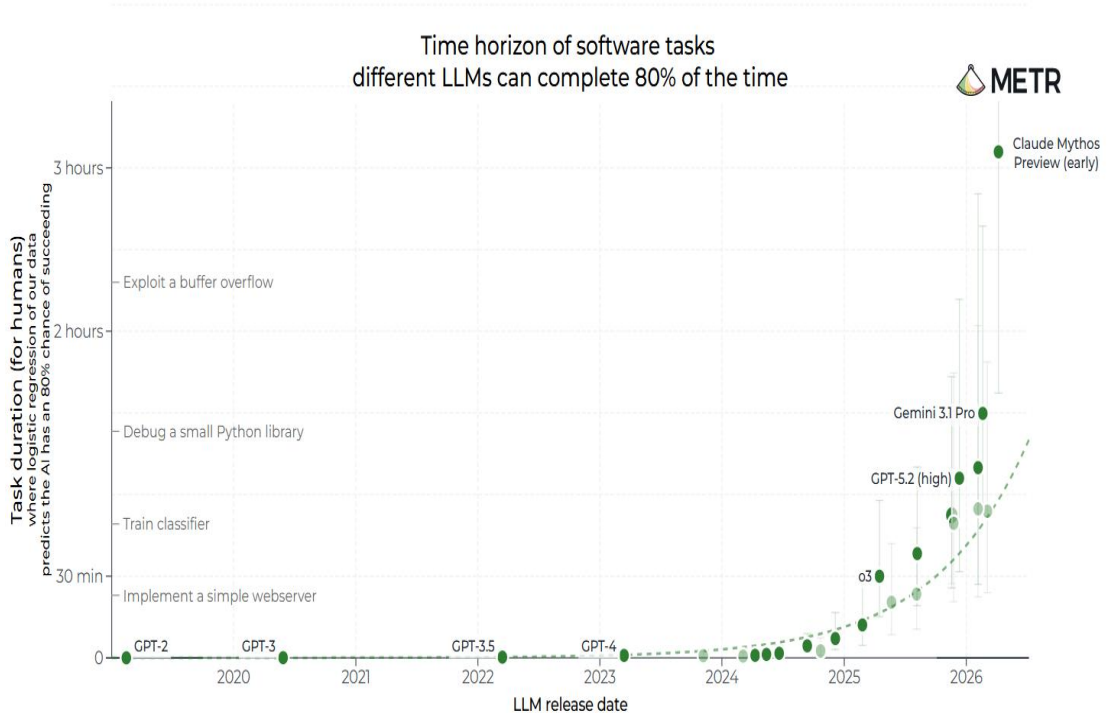
- **今年以来模型能力发生的核心变化是，跨过了可持续执行的拐点，编程能力、长程任务、工具调用成为模型重点提升的方向。**用户可以放心将大段时间交给模型自行完成任务而不需要过多人为介入，模型发展的方向从更好回答问题向更高效更准确完成任务进化。根据METR，过去六年里，这一指标持续呈指数级增长，平均每七个月翻一番，目前Opus 4.6、Gemini 3.1等新一代前沿模型以80%准确度完成任务的时间已经突破了1小时，即将正式发布的Claude Mythos甚至突破了3小时。

图：通用前沿模型智能体能够以50%的可靠性自主完成的任务长度



资料来源：METR、国信证券经济研究所整理

图：通用前沿模型智能体能够以 80% 的可靠性自主完成的任务长度



资料来源：METR、国信证券经济研究所整理

Token的价值量提升，ROI清晰被测算

- **企业对AI的使用量迅速增长，部分公司甚至不设置预算上限，开启Tokenmaxxing。**Tokenmaxxing指的是把 AI Token 消耗量当作生产力指标，并尽可能把 Token 用到极致。企业为了鼓励模型使用，甚至会在内部推出消耗排行榜。SemiAnalysis 的年化 Token 支出已达到员工薪酬的约 30%，在Anthropic Claude token 上的年度支出率已高达 1095 万美元，每位员工每月消耗的Token 数量略低于 50 亿个。有些团队成员每月的消耗量超过 1000 亿个 Token。
- **Tokenmaxxing同时招致了部分员工的Token滥用，企业开始寻求低成本替代方案。**持续的Tokenmaxxing导致部分企业员工出现浪费或无意义的Token消耗，近期Uber等公司开始设置限额或选择性价比模型支持简单任务，微软CEO在近期的访谈中表示，针对确定且可重复的业务流程，应部署小参数模型进行定向优化，如何合理利用模型，针对不同任务分配不同的模型成为企业新的关注方向，国产开源模型受到更多关注。

图：Semi Analysis部分工作场景AI ROI

Token Spend vs Labor Cost on Real SemiAnalysis Workflows

Task	Token Cost	Human Cost	Human Time Taken	ROI (Human Cost / Token Cost)
Chart Cursor vs Anthropic ARR Growth (past 4 months)	$4.66	$50	1 hour	10.7x
Build tokenomics disclosure Slack bot	$58.02	$1,000	20 hours	17.2x
Set up DigitalOcean droplet for Terminal Bench	$35.97	$500	10 hours	13.9x
Initiate on Hewlett Packard Enterprise: roadmap, balance sheet, and capex sustainability	$21.33	$1,000	20 hours	46.9x
Perfomed keyword analysis on, and thematically organized full OCP 2024 conference	$16.69	$450	9 hours	27.0x
Search past conferences for CPU trends, demand, pricing, and attach ratios	$7.61	$500	10 hours	65.7x
Marvell earnings recap with CXL and CPO updates, matched to prior TEL recap format	$2.82	$200	4 hours	70.9x
Search past conferences for optical networking and optical circuit switching trends	$8.20	$500	10 hours	61.0x
Pull 5 years + YTD financials, email results with Excel tables attached	$1.87	$150	3 hours	80.2x

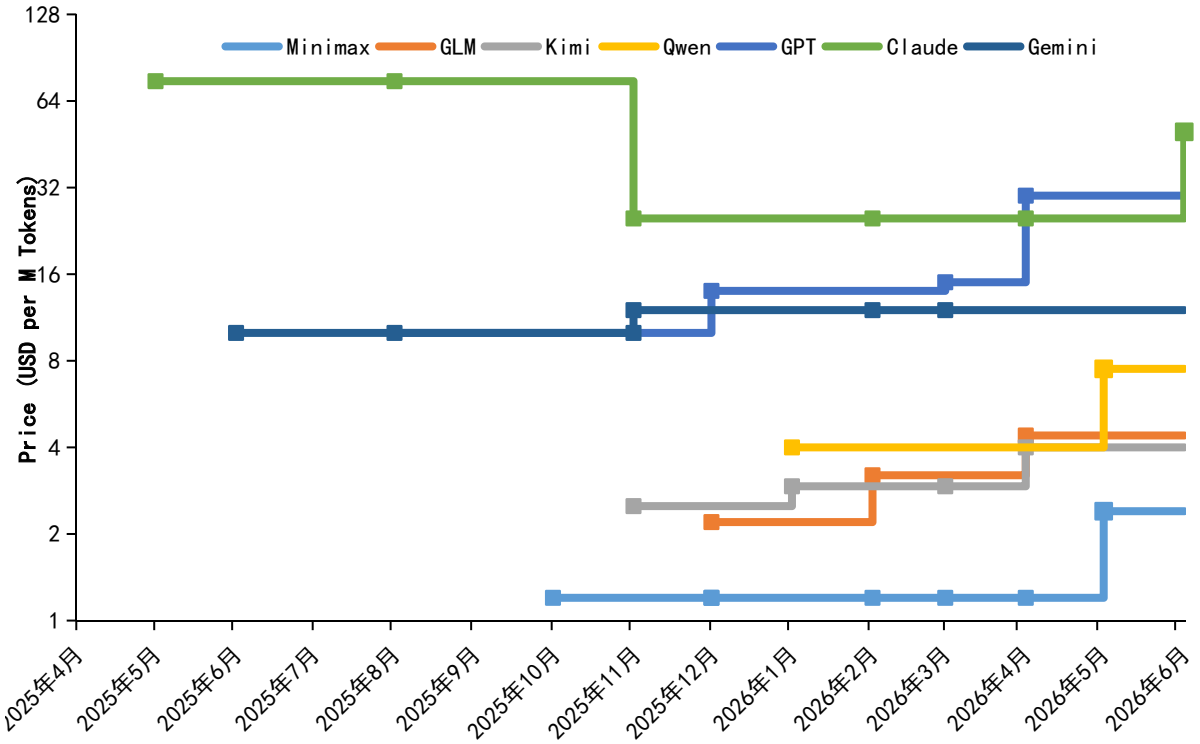
Assumptions: Human paid at ~$50/h hourly rate.

资料来源：Semi Analysis、国信证券经济研究所整理

Token的价值量提升，价格持续上涨

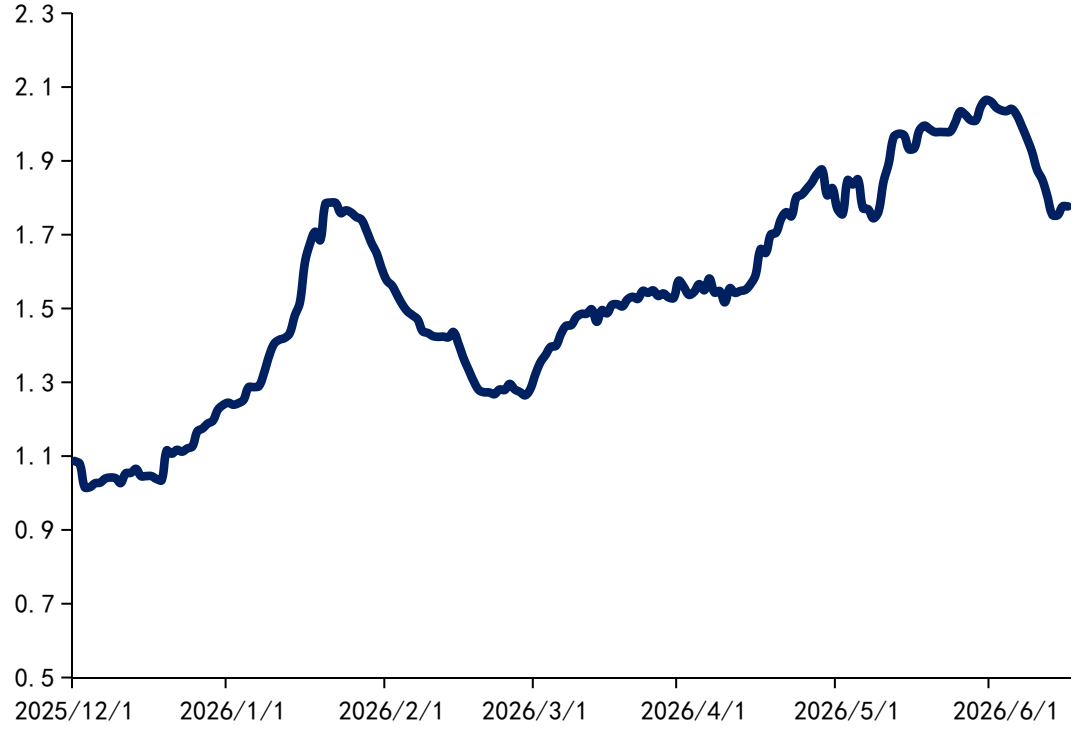
- **Token价值量提升驱动价格持续上涨。**去年以来，SOTA模型的价格普遍呈现上涨趋势，尤其是智谱、OpenAI、谷歌，智谱26Q1累计提价幅度接近翻倍。长期来看，随着模型能力的进步，现在的单位Token和过去的单位Token已经是不同的商品，能完成的任务价值量明显增加，因此，Token价格的上涨可能是长期趋势。Sam Altman在访谈中曾提到，未来模型收费的商业模式可能变为按照完成的任务进行付费。

图：目前主流模型公司去年以来前沿模型输出价格变化情况



资料来源：Artificial Analysis、国信证券经济研究所整理

图：海外Token平均支付价格



资料来源：Silicon Data、国信证券经济研究所整理

算力成为模型公司增长的关键

国信证券 GUOSEN SECURITIES

- **AI时代，算力已成为企业创造价值的核心生产要素。**黄仁勋认为，过去购买算力是为了运行软件，是支持性成本，现在算力就是生产力，算力规模直接决定token产量，进而决定企业收入。
- 26年4月开始，随着Claude用户量快速增长，社区大量反馈 Max 订阅也会频繁 hit limit，5月在与SpaceX 达成算力合作后（300MW+、22万+ NVIDIA GPU 新容量），立刻5小时限额翻倍，对于Pro、Max、Team及按席位计费的Enterprise计划，Claude Code的5小时使用限额翻倍，同时取消高峰时段限速。

图：Anthropic去年末以来加速算力建设和采购

合作方	采购/自建	具体细节
谷歌+博通	**采购**	• 23年双方开始合作，谷歌一直是Anthropic的主要算力提供方。 • 25年10月，双方宣布扩展**1GW**的计算容量，26年上线，Anthropic将获得100万个TPU芯片 • 26年4月，与谷歌和博通签署了一项新协议，获得**5GW**的TPU容量，27年开始上线
亚马逊	**采购**	• AWS与Anthropic合作开展Project Rainier，25年10月已投入使用，到年底采用超过50万颗 Trainium 2 芯片进行训练和推理任务。其中仅印第安纳州项目就斥资110亿美元，发电量超过2.2GW。 • 26年4月，与AWS达成协议，未来10年投入1000亿美元，确保**5GW**的容量，投入涵盖 Graviton 和 Trainium2 至 Trainium4 芯片，包括到26年底上线的接近**1GW**的 Trainium2 和 Trainium3
微软+英伟达	**采购**	25年11月，Anthropic宣布采购 300 亿美元 Azure计算容量，由英伟达Blackwell和Rubin提供支持，购买算力容量**上限达到1GW**，以获得英伟达和微软150亿美元的投资
Fluidstack	自建	2025年11月，Anthropic宣布将投资500亿美元，与 Fluidstack 合作在德克萨斯州和纽约州建设数据中心
Coreweave	采购	26年4月，达成多年期协议，今年晚些时候开始启用计算服务
SpaceX	采购	26年5月，与 SpaceX 达成协议，使用其 Colossus 1 数据中心的所有计算能力，一个月内获得**300MW+**，22万+ NVIDIA GPU

资料来源：Anthropic、国信证券经济研究所整理

图：OpenAI数据中心建设计划及算力采购计划

时间	地点	负责方与容量	建设说明
2025-2026	德克萨斯州阿比林市	Oracle负责，目前0.7GW，后续可达1.4GW	两栋建筑完工，每栋约60,000个GB200 NVL 72服务器
2025-2026	俄亥俄州洛兹敦、德克萨斯州米拉姆县	软银负责，未来18个月内可出售1.5GW	Vera Rubin 芯片
2026-	阿联酋的Stargate UAE 、挪威的 Stargate 等	阿联酋2026年上线200MW，后续计划规模5GW	星际之门阿联酋
2026-2027	德克萨斯州沙克尔福德县、新墨西哥州多纳阿纳县、美国中西部	Oracle负责，超4GW	
总计	**原计划未来四年内Stargate计划容量达到近10吉瓦（GW），投资超5000亿美元；最新26年4月已经超过10GW规模 英伟达投资1000亿美元锁定未来10GW数据中心的芯片供应权。 AMD授权最多1.6亿股权（AMD的10%）绑定未来6GW芯片供应权。**		
AWS	• **25年11月，采购380亿美元的英伟达GPU，26年底部署完成** • **26年2月，新增8年1000亿美元，约2GW AWS Trainium容量，覆盖Trainium3 和下一代 Trainium4**		

数据来源：公司官网、华尔街新闻、彭博新闻、国信证券经济研究所整理

目录

推理需求爆发，自研ASIC能力成熟，GPU/加速器芯片迎来加速期

国信证券 GUOSEN SECURITIES

- **伴随推理侧需求爆发，以及各公司自研芯片能力验证成熟，ASIC出货迎来加速期。**根据Gartner数据，到 2026 年，终端用户在推理方面的支出将超过训练密集型工作负载。另一方面，供给端各公司的ASIC芯片产品经历持续迭代目前也已逐步验证成熟。
- 我们根据各个企业对于其自研芯片产品路线与产能测算得：**26年ASIC芯片出货量大约770万片，份额为45%，并将在27年超过GPU份额达到58%。**并且未来在2030年占比达到约73%，谷歌贡献主要增量。

表：全球GPU/ASIC 出货量与YoY

加速器/千颗*	2024	2025	2026E	2027E	2028E	2029E	2030E
GPU	**5,234**	**6,110**	**9,570**	**11,500**	**13,300**	**14,600**	**16,000**
NVDA	4,850	5,700	9,000	10,500	12,000	13,000	14,000
AMD	384	410	570	1,000	1,300	1,600	2,000
ASIC	**2,130**	**2,850**	**7,700**	**15,900**	**35,200**	**39,500**	**43,000**
TPU	1,280	1,700	5,000	12,000	30,000	33,000	35,000
Trainium	750	750	1,700	2,400	3,200	4,000	5,000
其他ASIC	100	400	1,000	1,500	2,000	2,500	3,000
Total	**7,364**	**8,960**	**17,270**	**27,400**	**48,500**	**54,100**	**59,000**
YoY		22%	93%	59%	77%	12%	9%
占比							
GPU	**71%**	**68%**	**55%**	**42%**	**27%**	**27%**	**27%**
NVDA	66%	64%	52%	38%	25%	24%	24%
AMD	5%	5%	3%	4%	3%	3%	3%
ASIC	**29%**	**32%**	**45%**	**58%**	**73%**	**73%**	**73%**
TPU	17%	19%	29%	44%	62%	61%	59%
Trainium	10%	8%	10%	9%	7%	7%	8%
其他ASIC	1%	4%	6%	5%	4%	5%	5%

资料来源：各公司业绩会、各公司官网、Google Cloud Next 26、SemiAnalysis、JPR，国信证券经济研究所整理　*根据各家产品迭代路线与代工产能测算

26年起ASIC厂商占比台积电CoWoS产能份额崛起

国信证券 GUOSEN SECURITIES

- **从制造和封装环节来看，ASIC厂商在台积电的份额快速崛起。**根据我们测算，预计27年台积电CoWoS产能266万Wafer，同比+87%，其中ASIC厂商的份额将从28%提升至32%。

图：台积电26/27年CoWoS产能预测

AI芯片厂商	产品	2026			2027			Compute die size	制程	Compute die颗数
		CoWoS产能(k wafers)	每片CoWoS芯片数	隐含出货(k)	CoWoS产能(k wafers)	每片CoWoS芯片数	隐含出货(k)			
NVIDIA	B300	390	14	5,460	40	14	560	850	4nm	2
NVIDIA	Vera CPU	90	23	2,070	250	23	5,750	850	3nm	1
NVIDIA	Spectrum / CPX	60								
NVIDIA	Rubin R200	260	8	2,080	740	8	5,920	850	3nm	2
NVIDIA	Rubin Ultra				130	8	1,040	850	3nm	2
NVIDIA	H200	20	27	540				814	4nm	1
AMD	MI300	3	12	36				110	5nm	8
AMD	MI350/375	7	12	84	24	12	288	110	3nm	8
AMD	MI400	65	10	650	192	10	1,920	110/850	2nm	8/2
AMD	MI500				24	4	96		2nm	
AMD	Venice	50	25	1,250					2nm	
AMD	Venice GPU				270	25	6,750		2nm	
Google	TPU v7p	145	16	2,320	13	16	208	700	3nm	2
Google	TPU v8i	80	12	960	330	12	3,960	800	3nm	2
Google	TPU v8t	40	20	800	180	20	3,600	800	3nm	2
Google	TPU v9						400	850	2nm	2
AWS	Trainium 3	100	17	1,700	140	17	2,380	700	3nm	2
GUC's customers		10	20	200	60	20	1,200	645	4nm	1
Microsoft	Maia 200	4	29	116				700	3nm	1
Microsoft	Maia 300	5	11	55	50	11	550	850	2nm	1
Meta	MTIA 3	15	10	150	55	10	550	850	3nm	2
Others					18					
Total		1,424			2,664					

资料来源：台积电、国信证券经济研究所整理测算

单芯片TCO来看，ASIC芯片更有优势

国信证券 GUOSEN SECURITIES

图：GPU、TPU、Trainium芯片 TCO比较

单位：美元	单位	GB200 NVL72 (Spectrum)	GB300 NVL72 (Spectrum)	Trainium2 NL16 2D Torus	Trainium2 NL32x2 3D Torus	Trainium3 NL16 2D Torus	Trainium3 NL32x2 3D Torus	Trainium3 NL32x2 Switched	Trainium3 NL72x2 Switched	TPU v7 - 3D Torus - Internal	TPU v7 - 3D Torus - External
服务器成本	USD	$3,179,652	$3,997,587	$154,020	$655,793	$170,554	$693,183	$757,415	$2,020,274		
服务器维护服务	USD	$72,000	$72,000	$16,000	$64,000	$16,000	$64,000	$64,000	$144,000		
网络成本	USD	$296,733	$296,733	$26,300	$105,202	$26,300	$105,202	$105,202	$236,700		
存储成本	USD	$42,021	$42,021	$5,800	$23,200	$5,800	$23,200	$23,200	$52,200		
软件授权及其他成本	USD	$20,677	$23,057	$2,448	$9,838	$2,912	$11,684	$12,956	$30,826		
其他安装成本	USD	$72,000	$72,000	$16,000	$64,000	$16,000	$64,000	$64,000	$144,000		
单服务器总前期集群资本开支	USD	$3,683,083	$4,503,398	$220,569	$922,032	$237,566	$961,269	$1,026,772	$2,627,998		
单服务器GPU数	#	72	72	16	64	16	64	64	144		
单 GPU 总前期资本开支	USD	$51,154	$62,547	$13,786	$14,407	$14,848	$15,020	$16,043	$18,250		
WACC		9.4%	9.4%	9.4%	9.4%	9.4%	9.4%	9.4%	9.4%		
折旧周期	年	4	4	4	4	4	4	4	4		
单服务器每月总资本成本	USD/mth	$92,311	$112,871	$5,528	$23,109	$5,954	$24,093	$25,734	$65,867		
单GPU 每小时资本成本	**USD/hr/GPU**	**$1.76**	**$2.15**	**$0.47**	**$0.49**	**$0.51**	**$0.52**	**$0.55**	**$0.63**		
电力成本	USD/kWh	$0.09	$0.09	$0.09	$0.09	$0.09	$0.09	$0.09	$0.09		
利用率	%	80%	80%	80%	80%	80%	80%	80%	80%		
电源使用效率（PUE）	Ratio	1.35	1.35	1.35	1.35	1.35	1.35	1.35	1.35		
每月每 kW IT 负载电力成本	USD/kW/mth	$68.60	$68.60	$68.60	$68.60	$68.60	$68.60	$68.60	$68.60		
机房托管成本	USD/kW/mth	$110.00	$110.00	$110.00	$110.00	$110.00	$110.00	$110.00	$110.00		
每月每 kW 总托管成本	USD/kW/mth	$178.60	$178.60	$178.60	$178.60	$178.60	$178.60	$178.60	$178.60		
整机功耗	kW	143.05	160.05	12.46	50.16	15.77	63.34	72.43	174.89		
每月服务器托管总成本	USD/mth	$25,548	$28,584	$2,225	$8,958	$2,817	$11,312	$12,935	$31,233		
远程运维与支持工程师成本	USD/mth	$1,179	$1,179	$262	$1,048	$262	$1,048	$1,048	$2,358		
网络连接成本	USD/mth	$351	$351	$78	$312	$78	$312	$312	$702		
单服务器每月总运营成本	USD/mth	$27,078	$30,114	$2,565	$10,318	$3,157	$12,672	$14,295	$34,293		
单 GPU 每月运营成本	USD/mth	$376	$418	$160	$161	$197	$198	$223	$238		
单 GPU 每小时运营成本	**USD/hr/GPU**	**$0.52**	**$0.57**	**$0.22**	**$0.22**	**$0.27**	**$0.27**	**$0.31**	**$0.33**		
单GPU 每小时成本TCO	**USD/hr/GPU**	**2.28**	**2.72**	**0.69**	**0.71**	**0.78**	**0.79**	**0.86**	**0.96**	**1.28**	**1.60**

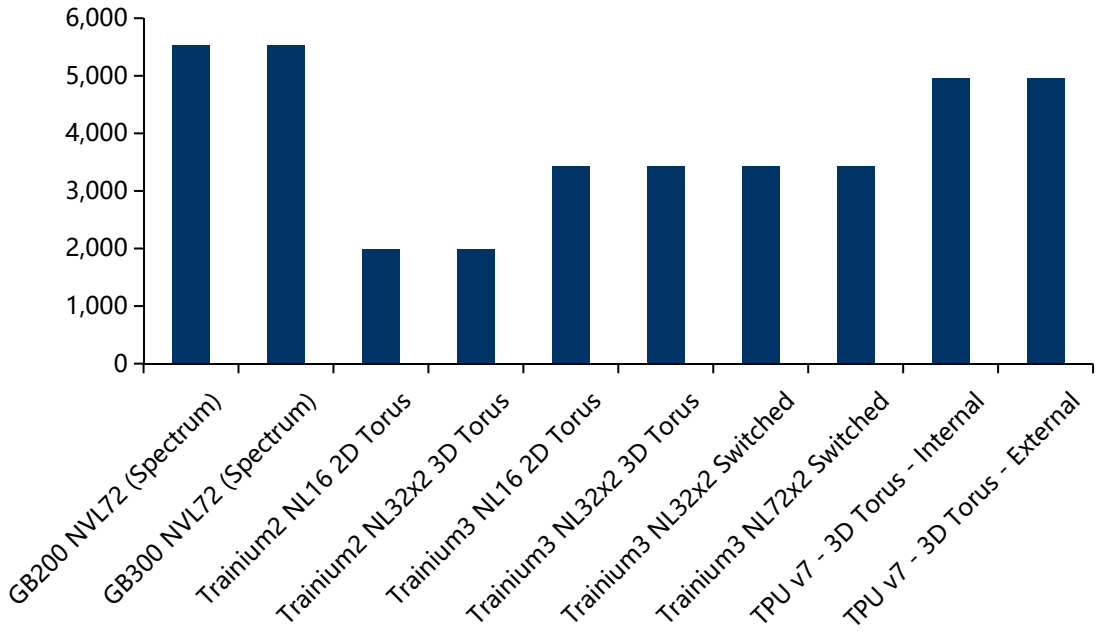
资料来源：Semi Analysis、国信证券经济研究所整理

GPU拥有更好的TPS（Token Per Second），但ASIC的Cost Per Token更优

国信证券 GUOSEN SECURITIES

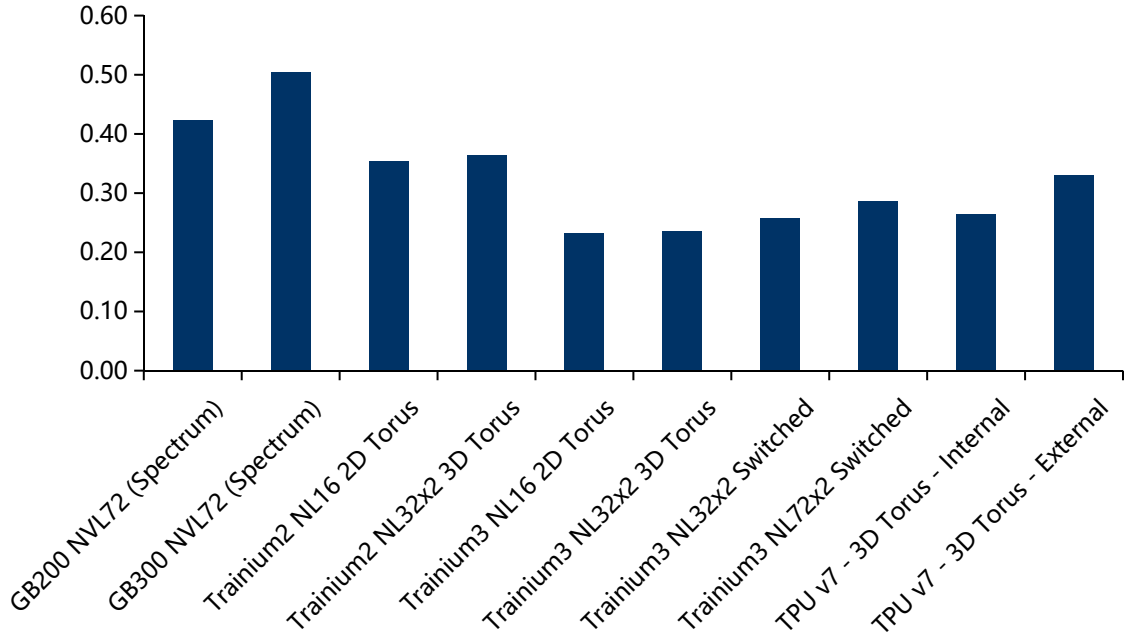
- **在推理场景中，客户更关注Cost Per Token，ASIC的成本优势更加明显。**根据我们测算，在相同模型下，GPU单位时间能够产生更多Tokens，这可能意味着更加适合训练的场景，而ASIC的成本优势则更明显，也因此在推理场景下更受青睐。具体来说，我们采用DeepSeekR1模型，激活专家数：9 out of 257，模型大小：671GB，Model layer：61。假设Input token：1,024；Output token：1,024，Batch size：1，Computing power：FP8。并且假设芯片利用率为100%。根据测算结果，ASIC的Cost Per Token更优
- **训练阶段，成本不是核心要素，每秒的Token吞吐量、集群的稳定性更加关键。**根据我们测算，GPU拥有更好的TPS（Token Per Second），叠加万卡集群的成熟度，GPU依然是更好的选择。

图：GPU、TPU、Trainium芯片Token Per Second



资料来源：Semi Analysis、国信证券经济研究所整理测算

图：GPU、TPU、Trainium芯片Cost Per Token



资料来源：Semi Analysis、国信证券经济研究所整理测算

亚马逊Trainium、谷歌TPU、GPU基础参数对比

图：Trainium与TPU、GPU对比

核心规格对比	Trn3	Trn2	Trn2-Ultra	TPUv7	TPUv6e	GB300	GB200	H100
理论 BF16 密集算力（TFLOPS/芯片）	671	667	667	2307	918	2500	2500	989
理论 8-bit 密集算力（TFLOPS/芯片）	2517	1300	1300	4614	1836	5000	5000	1978
HBM 容量（GB/芯片）	144	96	96	192	32GB HBM3	288GB HBM3e	192	80
HBM 带宽（GB/s/芯片）	4900	2900	2900	7380	1640	8000	8000	3350
算术强度（BF16 FLOP/Byte）	136.9	230	230	312.6	559.8	625	312.5	295.2
单位 HBM 容量算力（FLOP/Byte）	4659.7	6947.9	6947.9	12015.6	28687.5	17361.1	13020.8	12362.5
结构化稀疏支持	2:4, 4:8, 4:16	2:4, 4:8, 4:16	2:4, 4:8, 4:16	未披露	无	2:04	2:04	2:04
芯片功耗（W）	700	500	500	980	600	1400	1200	700
散热方式	风冷/液冷	风冷	风冷	液冷	风冷	直连液冷（DLC）	直连液冷（DLC）	主要风冷 / 部分液冷
Scale Up 网络								
Scale Up 互连技术	NeuronLinkV4	NeuronLinkV3 (PCIe Gen5)	NeuronLinkV3	TPU ICI	TPU ICI	NVLink5	NVLink5	NVLink4
Scale Up 带宽（GB/s 单向）	1200	512	640	600	448	900	900	450
Scale Up 最大规模（芯片数）	16	16	64	9216	256	72	72	8
Scale Up 拓扑结构	二维/三维互连域	4×4 二维环形 (Torus)	4×4×4 三维环形	3D Torus	16×16 二维环形	18轨优化 Fat Tree	18轨优化 Fat Tree	4轨优化 Fat Tree
每芯片 Scale Up 邻居数	4	4	6	6	4任意互连	任意互连	任意互连	
Scale Up 总 HBM 容量	2304	1536	6144	1769472	8192	20736	13824	640
Scale Up 物理服务器数量	1	1	4		32	18	18	1
Scale Out 网络								
Scale Out 网络技术	EFAv4	EFAv3	EFAv3	Google 以太网	Google 以太网	InfiniBand / 以太网	InfiniBand / 以太网	InfiniBand / 以太网
Scale Out 带宽（Gb/s 单向）	400最高 800		200	400	100	800	800	400
Scale Out 拓扑	ToR Fat Tree	ToR Fat Tree	ToR Fat Tree	ToR / OCS	ToR Fat Tree	4轨 Fat Tree	4轨 Fat Tree	8轨 Fat Tree
Scale Up / Scale Out 带宽比	24	5.12	25.6	12	35.84	9	9	9

资料来源：SemiAnalysis、国信证券经济研究所整理

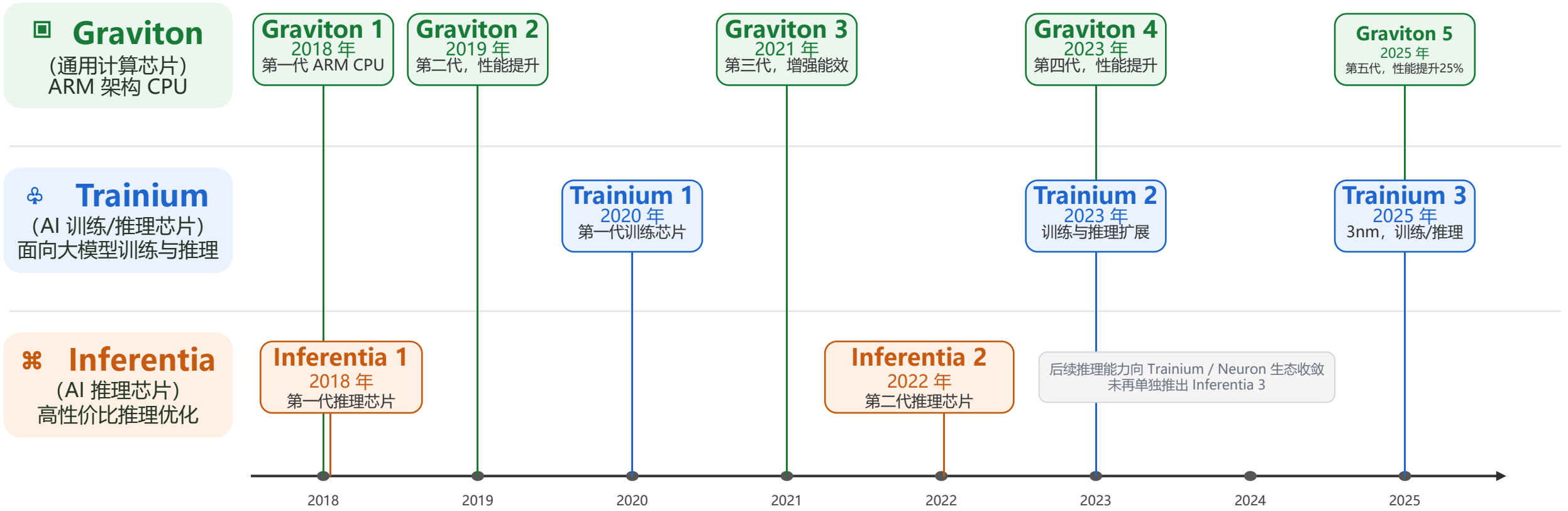
目录

AWS自研芯片历程：技术班底来自Marvell，从网络芯片拓展至AI算力

国信证券 GUOSEN SECURITIES

- **技术班底来自Marvell，从网络芯片拓展至AI算力。**2015年亚马逊以约3.5亿美元收购以色列芯片设计公司 Annapurna Labs，两位核心创始人Nafea Bshara和Bilic Hrvoje在Marvell工作约10年，亚马逊收购的最初目的是为降低成本，2017年推出Nitro把原来CPU负责的网络、存储、虚拟化、安全隔离全部卸载到专用芯片上；2018年推出CPU芯片AWS Graviton采用ARM架构；2019年发布推理芯片AWS Inferentia进入AI领域；21年推出训练芯片AWS Trainium，并且自23年开始将推理芯片与训练芯片合并。

图：AWS自研芯片发展历程



注：时间口径为 AWS 对外首次公开宣布/发布芯片代际的时间，非实例 GA 或大规模商用时间。

资料来源：AWS、国信证券经济研究所整理

亚马逊Trainium：单一芯片开始逐渐构建系统级集群能力

国信证券 GUOSEN SECURITIES

- Trainium目前一共推出三代，从成本替代，到具备单卡能力和大模型训练能力，再到拥有系统级互联能力挑战GPU集群。

图：Trainium每代参数对比

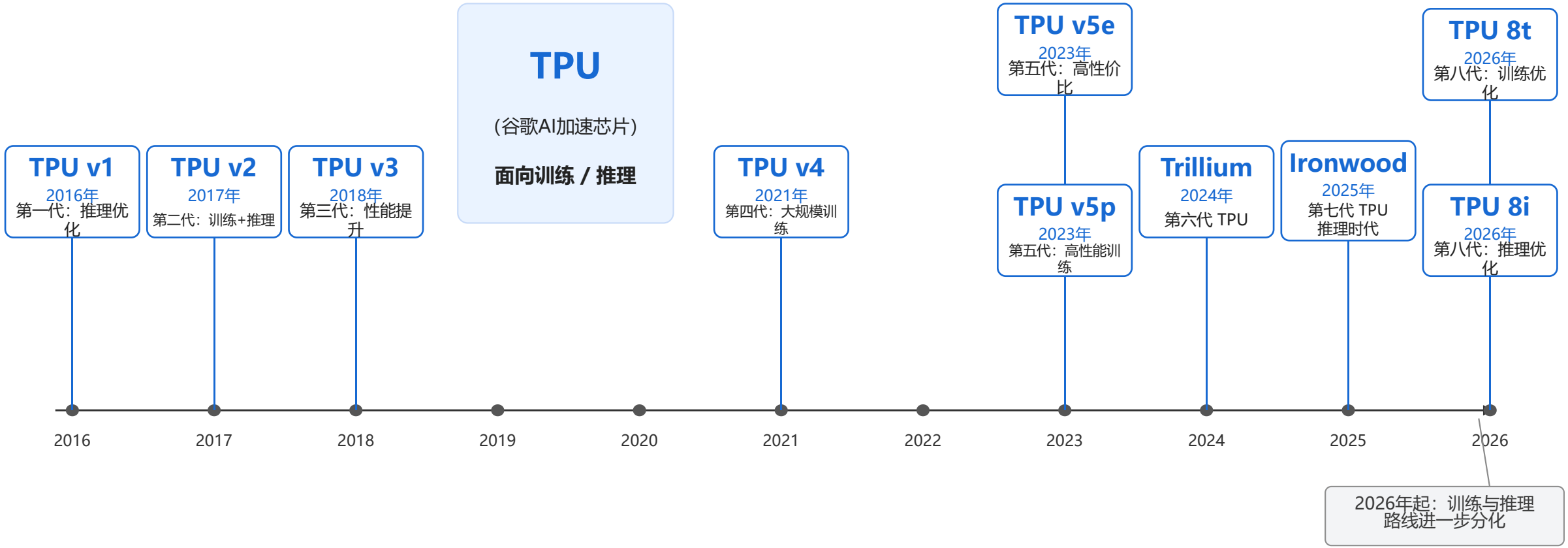
	Trainium	Trainium2 NL16	Trainium2 NL32x2	Trainium3	Trainium4
逻辑与算力					
逻辑芯片配置	1×计算芯片 + 2×HBM2E	2×计算芯片 + 4×HBM3E	同左	2×计算芯片 + 4×HBM3E	4×计算芯片 + 8×HBM4 + 2×IO Die
制程节点	N7	N5	N5	N3P	N2
FP8 密集算力（TFLOPS/封装）	191	1299	1299	2517	TBD
BF16 密集算力（TFLOPS/封装）	191	667	667	671	TBD
FP16 密集算力（TFLOPS/封装）	191	667	667	671	TBD
TF32 密集算力（TFLOPS/封装）	191	667	667	671	TBD
FP32 密集算力（TFLOPS/封装）	48	181	181	183	TBD
内存系统					
HBM 类型	HBM2E 8-Hi	HBM3E 8-Hi	HBM3E 8-Hi	HBM3E 12-Hi	HBM4 12-Hi
HBM 堆叠数量	2	4	4	4	8
单堆容量（GB）	16GB	24GB	24GB	36GB	36GB
总 HBM 容量	32GB	96GB	96GB	144GB	288GB
总线位宽（bit）	2048	4096	4096	4096	16384
引脚速率（Gb/s）	3.2	5.7	5.7	9.6	9.6
内存带宽	0.8 TB/s	2.9 TB/s	2.9 TB/s	4.9 TB/s	19.6 TB/s
封装与散热					
封装技术	CoWoS-S	CoWoS-R	CoWoS-R	CoWoS-R	TBD
功耗（TDP）	-	~500W	~500W	~700W	TBD
散热方式	风冷	风冷	风冷	风冷 / 液冷	TBD
Scale Up 网络					
Scale Up 互连技术	NeuronLinkV2	NeuronLinkV3	NeuronLinkV3	NeuronLinkV4	TBD
有效通道数（Lanes）	-	128	160	144	TBD
单通道速率（Gb/s）	-	32	32	64	TBD
Scale Up 带宽（TB/s）	-	0.5	0.6	1.2	TBD
Scale Out 网络					
Scale Out 网络技术	EFAv2	EFAv3	EFAv3	EFAv4	TBD
最大 Scale Out 带宽（Gb/s）	170	最高 800	200	400	TBD

资料来源：SemiAnalysis、国信证券经济研究所整理

谷歌TPU发展历程：从内部推理加速器，到AI基础设施平台

- **从谷歌内部推理加速器，逐步演进成“训练+推理+大规模互联+云服务化”的AI基础设施平台。**谷歌的TPU设计团队是由内部基础设施团队牵头成立的，2015年内部部署TPUv1，服务内部搜索、广告、翻译、推荐等业务，v2开始支持训练，并通过Cloud TPU对外提供；v3进一步提高算力和内存带宽，并引入更大规模Pod。TPUv4是关键分水岭，在Pod中引入了OCS，变成“芯片+光互联+系统调度”的数据中心级产品。v5分拆训练和推理线，v6则是面向Transformer时代的一次大升级，v7 Ironwood则面向生成式AI推理时代。

图：谷歌TPU发展历程



资料来源：Google、国信证券经济研究所整理

谷歌TPU：训推分离，演进重心转向互联架构

国信证券 GUOSEN SECURITIES

- **2026年4月的谷歌CloudNext大会上披露了TPUv8的更多信息，最大变化是训练和推理分拆。**TPUv8t（训练）：单Pod可扩展至9600芯片，FP4峰值算力121 ExaFlops（前代3倍），芯片间ICI带宽翻倍至19.2 Tb/s；采用Virgo两层无阻塞网络，单集群可连超100万芯片，Scale-out带宽升至400Gb/s，空载延迟降40%。TPU8i（推理/智能体）：单Pod1152芯片，搭载288GBHBM+384MB片上SRAM（前代3倍），HBM带宽达8601GB/s；改用Boardfly拓扑，将1024芯片集群网络直径从16跳减至7跳，尾部延迟降50%，片上CAE引擎将全局操作延迟降低5倍。

图：TPU每代参数对比

项目	单位	TPU v5e (Viperlite)	TPU v5p (Viperfish)	TPU v6 (Trillium)	TPU v7 (Ironwood)	TPU v8t	TPU v8i
逻辑裸片配置	-	1x Compute Die, 1x IOD, 2x HBM2E 4-Hi	1x Compute Die, 1x IOD, 6x HBM2E 8-Hi	1x Compute Die, 1x IOD, 2x HBM3 8-Hi	2x Compute Dies, 1x IOD, 8x HBM3E 8-Hi		
系统可用时间	-	4Q23	1Q24	4Q24	4Q25		
计算芯片制程节点	-	N/A	N5	N5P	N3E		
HBM 容量	GB	16GB HBM2E	95GB HBM2E	32GB HBM3	192GB HBM3E	216GB	288GB
HBM 带宽	GB/s	819	2,765	1,640	7,380	6,528	8601
封装技术	-	-	CoWoS-S	-	-		
FP8 稠密算力	TFLOPs	-	459	918	4,614		
INT8 稠密算力	TFLOPs	394	918	1,836	4,614		
BF16 稠密算力	TFLOPs	197	459	918	2,307		
芯片热设计功耗（TDP）	W	~300	~550	~390	~980		
Scale up网络							
Scale Up 拓扑结构	-	2D Torus	3D Torus	2D Torus	2D/3D Torus	3D torus	Boardfly
芯片间单向互联带宽	Gb/s	1,600	4,800	3,200	4,800	9600	9600
TPU Pod 集群规模	#	256	8,960	256	9,216	9600	1152

资料来源：SemiAnalysis、国信证券经济研究所整理

预计27年谷歌TPU和亚马逊Trainium出货量将达到1500/220万张

国信证券 GUOSEN SECURITIES

- **亚马逊：**根据公司25年股东信，亚马逊26Q1芯片业务ARR已超过200亿美元，同比三位数增长（包括Graviton, Trainium, and Nitro），占云收入的约15%。如果作为独立业务对外销售则将达到500亿美元。
  1. **Graviton：23-25年，连续3年AWS 新增 CPU 容量的一半以上由 Graviton 提供支持，**排名前 1000 的 EC2 客户中有 98%已经受益于 Graviton 的性价比优势。
  2. **Trainium：**公司主要面向Anthropic和OpenAI提供服务，25年主要面向Anthropic，预计25年Trainium芯片为公司贡献约30亿+美元收入。
- **谷歌：**TPU由于更成熟的使用效果，目前谷歌已经开始向特定客户群体直接交付TPU硬件，首批客户包括Thinking Machines Lab、Hudson River Trading和Boston Dynamics等AI实验室、资本市场公司和高性能计算应用客户，今年晚些时候开始确认收入，27年产生绝大部分收入，我们预计今年TPU的出货量将达到500万张。

图：AWS AI芯片保有量预测（存量）、谷歌TPU出货量预测

单位：万张	2022年	2023年	2024年	2025年	2026年	2027年
Trainium1		10				
Trainium2			10	120		
Trainium3					170	220
合计		10	10	120	170	220

单位：万张	2023年	2024年	2025年	2026E	2027E
v5	50	240	180		
v6			120	110	
v7				200	
v8t（博通负责）				60	
v8i（联发科负责）				130	
合计	50	240	300	500	1200

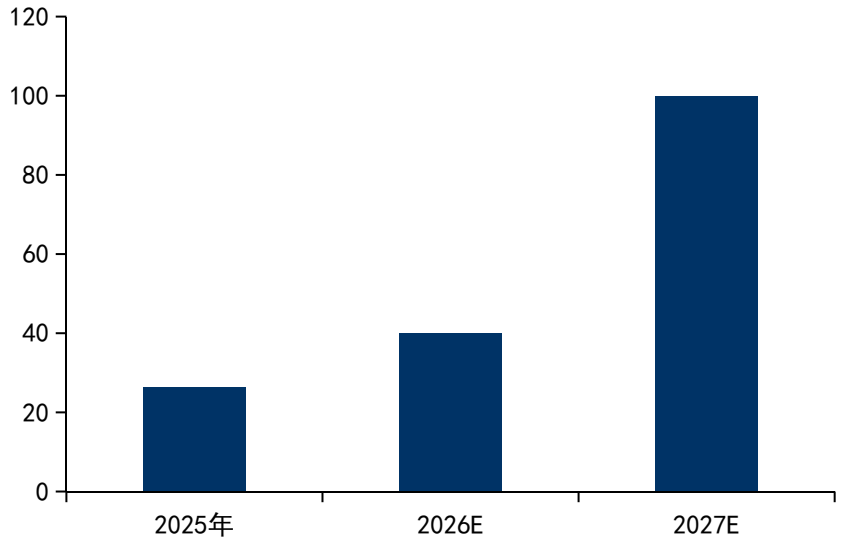
资料来源：Omdia、Amazon、The Information、Wolfe Research、国信证券经济研究所整理预测

阿里云：平头哥芯片出货量达56万，真武M890发布

- 26年5月20日，阿里发布自研芯片的路线图，平头哥发布了新一代训推一体AI芯片真武M890，144GB显存，片间互联带宽800GB/s，性能是上一代真武810E的3倍。配套发布的ICN Switch 1.0互联芯片，可以将128张AI芯片组成一台超节点服务器，P2P时延低于150纳秒。未来两年将陆续推出算力更强的真武V900、真武J900两代芯片。
- **根据阿里技术，截止5月，真武系列芯片已累计出货56万片。**服务了中国电信、中国一汽、浦发银行等20多个行业的400多家客户。截至26Q1，超过10万卡真武PPU已部署于阿里云公共平台，超过30家车企及自动驾驶公司正基于此芯片开展智能驾驶研发。

图：真武芯片发展路线&平头哥芯片出货量

	真武810E	真武M890	真武V900	真武J900
发布时间	2024Q2	2026Q2	2027Q3	2028Q3
核心性能	基线（1X）	3X	3X	-
显存容量	96GB	144GB	216GB	-
片间互联带宽	700GB/s	800GB/s	1200GB/s	-
架构特点	自研并行计算架构，高易用训推一体AI芯片	全面升级自研并行计算架构	深度迭代自研并行计算架构	自研并行计算架构跨越革新
累计出货量	截至2026年2月，真武（810E）已经累计规模化交付47万片			

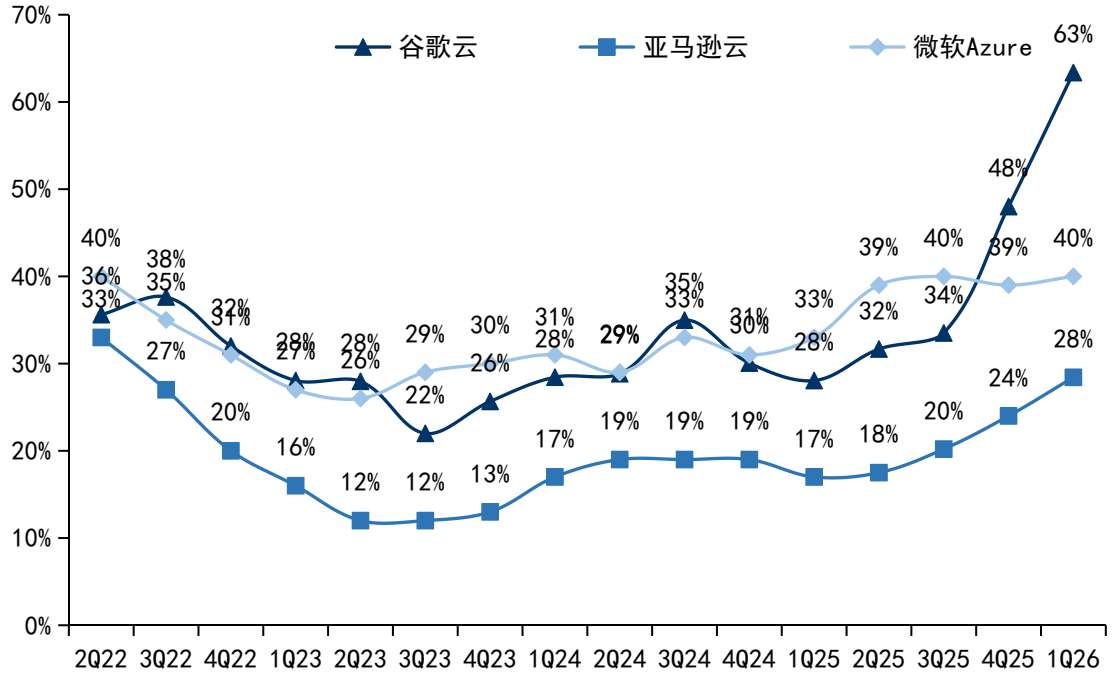


资料来源：阿里巴巴、国信证券经济研究所整理测算

海外云收入与利润率：研芯片产能爬坡带动云厂利润率优化

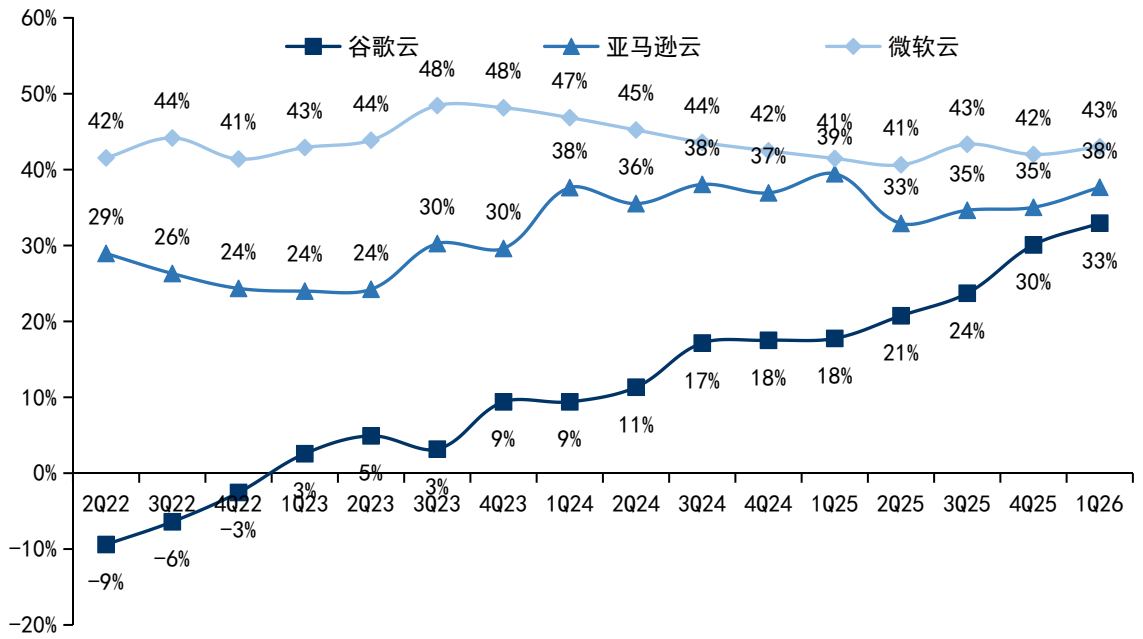
- **研芯片产能爬坡带动云厂利润率优化。**26Q1海外云厂收入加速增长，其中谷歌、亚马逊表现亮眼，谷歌云收入增速已达到63%，AWS也连续多个季度加速，但Azure则保持平稳，我们认为核心原因在于TPU和Trainium芯片得到验证后极大缓解了两家公司的供给压力，但微软的自研芯片尚未获得大范围认可。利润端，自研芯片也能够缓解云厂的成本压力（根据公司财报，Trainium合作的世芯毛利率仅20%+，博通在60%，因此谷歌寻求了联发科合作，而英伟达毛利率高达75%）。

图：海外云厂收入增速



资料来源：公司财报、国信证券经济研究所整理

图：海外云OPM



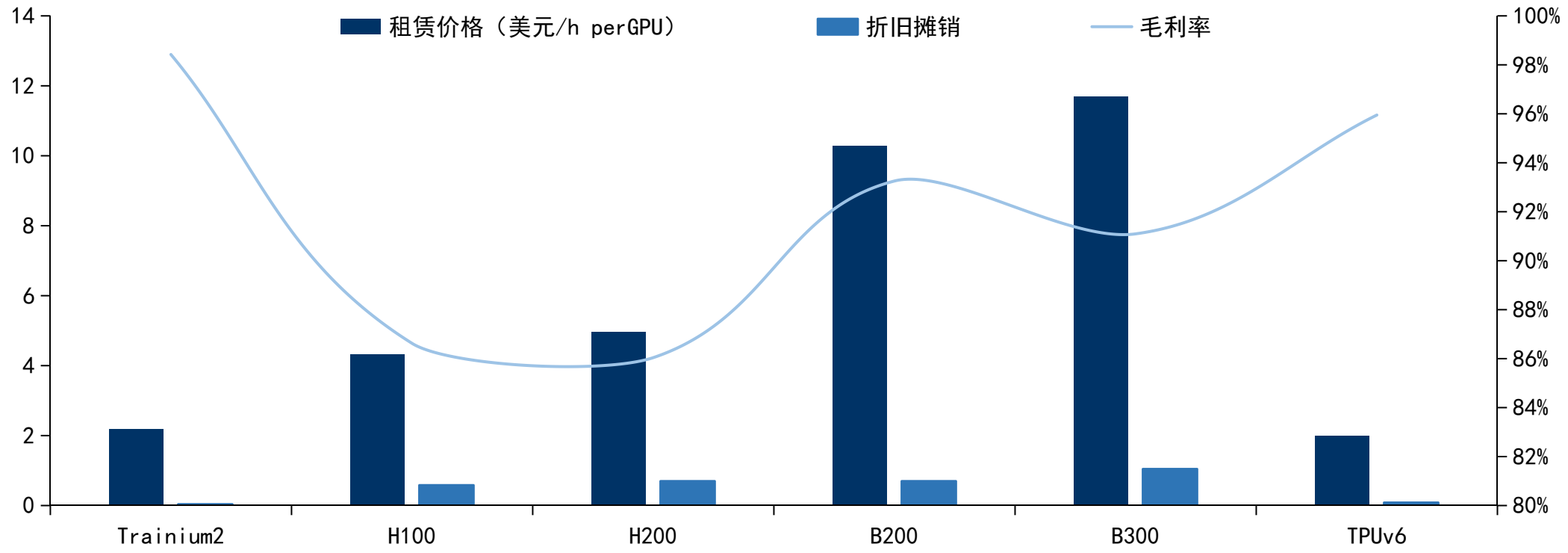
资料来源：公司财报、国信证券经济研究所整理

云厂自研芯片能够有效降低成本，带来利润率的提升

国信证券 GUOSEN SECURITIES

- **云厂自研芯片能够有效降低成本，带来利润率的提升。**我们结合Trainium2、H100、H200、B200、B300、TPUv6的租赁价格和售价，仅考虑芯片成本，不考虑租赁价格的折扣，测算毛利率能够得到Trainium2和TPUv6的毛利率都好于英伟达芯片。

图：Trainium、GPU、TPU的毛利率对比（注：仅考虑芯片成本，租赁价格未考虑折扣）



资料来源：AWS、GCP、国信证券经济研究所整理测算

亚马逊云：Trainium将带动AWS27年约4%的收入增速

国信证券 GUOSEN SECURITIES

- 我们预计25年亚马逊Trainium芯片贡献收入34亿美元，27年在云收入中占比提升至7%，AI IaaS收入2027年占比有望达到18.5%，Trainium将带动约4%的收入增速

图：AWS Trainium和GPU收入

$MMs	2022	2023	2024	2025E	2026E	2027E
AWS资本开支	27,755	24,843	53,267	90,554	117,720	141,264
AWS服务器资本开支占比	60.0%	60.0%	60.0%	60.0%	60.0%	60.0%
AWS服务器资本开支	16,653	14,906	31,960	54,332	70,632	84,758
Trainium出货量测算	**2022**	**2023**	**2024**	**2025E**	**2026E**	**2027E**
Trainium期初保有量	0.20	0.30	0.40	0.50	1.70	3.40
加：Trainium单位净新增	0.10	0.10	0.10	1.20	1.70	2.20
Trainium期末保有量	**0.3**	**0.4**	**0.5**	**1.7**	**3.4**	**5.6**
Trainium预计单芯片成本	$1,200	$1,200	$1,500	$8,071	$8,071	$8,071
Trainium资本开支	**106**	**120**	**150**	**9,686**	**13,721**	**17,757**
占AWS资本开支比例	1%	0.8%	0.5%	17.8%	19.4%	20.9%
Trainium3芯片	0.00	0.00	0.00	0.00	1.70	2.20
Trainium2芯片	0.00	0.00	0.10	1.30	1.30	1.30
Trainium1芯片	0.30	0.40	0.40	0.40	0.40	0.40
Trainium3占比	--	--	--	--	50%	56%
Trainium2占比	--	--	20%	76%	38%	33%
Trainium1占比	100%	100%	80%	24%	12%	10%
Trainium收入测算	**2022**	**2023**	**2024**	**2025E**	**2026E**	**2027E**
Trainium每小时成本	$0.60	$0.60	$2.24	$2.24	$2.24	$2.24
Trainium综合每小时成本	**$0.60**	**$0.60**	**$0.92**	**$0.82**	**$0.91**	**$0.92**
批量折扣	58%	57%	56%	56%	56%	56%
内部工作负载	35%	33%	31%	26%	21%	16%
外部工作负载	65%	67%	69%	74%	79%	84%
利用率	85%	85%	85%	85%	85%	85%
Trainium收入	$363	$511	$1,044	$3,430	$8,088	$14,336
同比增速		*40.7%*	*104.2%*	*228.5%*	*135.8%*	*77.3%*
占AWS收入比例	0.5%	0.6%	1.0%	2.7%	5.0%	7.0%

NVDA GPU出货量测算（百万颗）	2022	2023	2024	2025E	2026E	2027E
NVDA GPU期初保有量	**0.06**	**0.13**	**0.23**	**0.48**	**0.79**	**1.04**
加：NVDA GPU净新增	0.06	0.11	0.25	0.31	0.43	0.54
减：退役NVDA GPU					0.19	0.16
NVDA GPU期末保有量	**0.13**	**0.23**	**0.48**	**0.79**	**1.04**	**1.42**
NVDA芯片收入贡献（百万美元）	**2022**	**2023**	**2024**	**2025E**	**2026E**	**2027E**
NVDA GPU每小时成本						
A100	1.475	1.475	1.48	0.00	0.00	0.00
H100	0.00	3.93	3.93	3.93	3.93	3.93
H200	0.00	0.00	0.00	4.33	4.33	4.33
GB200	0.00	0.00	0.00	5.29	5.29	5.29
NVDA GPU综合每小时成本	**0.74**	**1.47**	**1.72**	**2.12**	**2.18**	**2.23**
A100	100%	40%	20%	0%	0%	0%
H100	0%	60%	80%	60%	55%	50%
H200	0%	0%	0%	25%	20%	15%
GB200	0%	0%	0%	15%	25%	35%
批量折扣	50%	50%	50%	50%	50%	50%
相对Trainium溢价	**1.24x**	**2.47x**	**1.86x**	**2.57x**	**2.39x**	**2.43x**
内部工作负载	5%	5%	5%	5%	5%	5%
外部工作负载	95%	95%	95%	95%	95%	95%
利用率	90%	90%	90%	90%	90%	90%
NVDA GPU收入	$694	$2,556	$6,176	$12,595	$16,870	$23,714
同比增速		*268.4%*	*141.6%*	*103.9%*	*33.9%*	*40.6%*
占AWS收入比例	0.9%	2.8%	5.7%	9.8%	10.4%	11.5%
Trainium + NVDA GPU收入	$1,057	$3,067	$7,220	$16,025	$24,958	$38,050
同比增速		*190.1%*	*135.4%*	*122.0%*	*55.7%*	*52.5%*
占AWS总收入比例	1.3%	3.4%	6.7%	12.5%	15.3%	18.5%

资料来源：AWS、国信证券经济研究所整理测算

谷歌云：TPU收入在26年占比将达到18%，贡献近170亿美元

国信证券 GUOSEN SECURITIES

- **AI云已成为谷歌云增长主要动力。**25年谷歌云收入同比+36%，我们测算其中AI云有望实现同比+102%，收入占比达到27%。GPU和TPU租赁相关收入同比+72%，26年将延续这一趋势。
- **当前算力的需求方仍以头部公司为主，大型互联网公司是主要客户。**我们预计TPU的收入主要来自苹果和Anthropic，小部分来自Safe Superintelligence和Physical Intelligence等人工智能实验室，GPU租赁收入则主要来自Meta、字节、腾讯等公司。**根据我们测算，预计TPU在云收入中的占比在26年将达到18%，贡献近170亿美元**

图：谷歌云收入预测

	2022	2023	2024	2025	2026E	2027E
谷歌云收入	**26,280**	**33,088**	**43,229**	**58,705**	**93,521**	**142,582**
YoY	37%	26%	31%	36%	59%	52%
传统云	**24,510**	**29,496**	**35,431**	**42,953**	**53,692**	**63,356**
YoY		20%	20%	21%	25%	18%
AI云	**1,770**	**3,592**	**7,798**	**15,752**	**39,829**	**79,225**
YoY		103%	117%	102%	153%	99%
占云收入比例	7%	11%	18%	27%	43%	56%
AI IaaS收入	**1,770**	**3,592**	**7,798**	**13,404**	**28,089**	**44,005**
YoY		103%	117%	72%	110%	57%
TPU收入	**1,340**	**2,109**	**3,832**	**6,220**	**16,907**	**30,142**
YoY	121%	56%	82%	52%	66%	46%
占云收入比例	5%	6%	9%	11%	18%	21%
TPU 数量	4.4	6.1	8.1	11.1	16.1	24
TPU Blended Cost/HR	0.92	0.86	1.00	1.03	1.11	1.18
折扣率	70%	70%	70%	70%	70%	70%
外部使用占比	14%	17%	20%	23%	40%	45%
利用率	90%	90%	90%	90%	90%	90%
GPU收入	**430**	**1483**	**3966**	**7184**	**11182**	**13863**
YoY		245%	168%	81%	56%	24%
占云收入比例	2%	4%	9%	12%	12%	10%
MaaS收入				**2,348**	**11,740**	**35,220**
YoY					400%	200%

资料来源：谷歌、国信证券经济研究所整理测算

阿里云：自研芯片带动收入加速增长及EBITA Margin改善

国信证券 GUOSEN SECURITIES

- **自研芯片放量驱动云收入加速增长。**根据我们测算，AI云收入在阿里外部云收入中占比在FY28将超过50%，核心驱动来自于GPU租赁和MaaS的高速增长，其中，自研芯片的放量将带动GPU租赁收入的加速增长，保持三年超过50%的增速。
- **利润端，自研芯片将提升GPU租赁业务的毛利率。**结合算力租赁的涨价，我们预计FY29GPU租赁的毛利率将提升至约36%。结合MaaS的高毛利率，预计到FY29阿里云整体的经调EBITA利润率将达到16%。

图：阿里云收入拆分

单位：百万元人民币	FY2023	FY2024	FY2025	FY2026	FY2027E	FY2028E	FY2029E
收入	**103,268**	**106,432**	**118,028**	**158,132**	**234,078**	**344,923**	**492,590**
YoY	38%	3%	11%	34%	48%	47%	43%
外部	**77,203**	**75,957**	**83,698**	**111,043**	**168,153**	**252,628**	**363,378**
yoy		-2%	10%	33%	51%	50%	44%
AI云收入	**800**	**4,800**	**11,260**	**28,263**	**70,291**	**140,086**	**236,206**
yoy		500%	135%	151%	149%	99%	69%
占外部云收入	1%	6%	13%	25%	42%	55%	65%
GPU租赁				**19,687**	**38,437**	**61,499**	**92,248**
YoY					95%	60%	50%
占AI云比例				70%	55%	44%	39%
毛利率				25%	31%	35%	36%
MaaS（含百炼+API调用）				**5,658**	**25,000**	**66,250**	**121,750**
YoY					342%	165%	84%
占AI云比例				20%	36%	47%	52%
毛利率				30%	40%	45%	48%
传统云收入	**76,403**	**71,157**	**72,438**	**82,780**	**97,862**	**112,542**	**127,172**
yoy		-7%	2%	14%	18%	15%	13%
经调EBITA	**2281**	**5592**	**10556**	**14265**	**24,899**	**47,955**	**80,404**
YoY	99%	145%	89%	35%	75%	93%	68%
经调EBITA Margin	**2.2%**	**5.3%**	**8.9%**	**9.0%**	**10.6%**	**13.9%**	**16.3%**

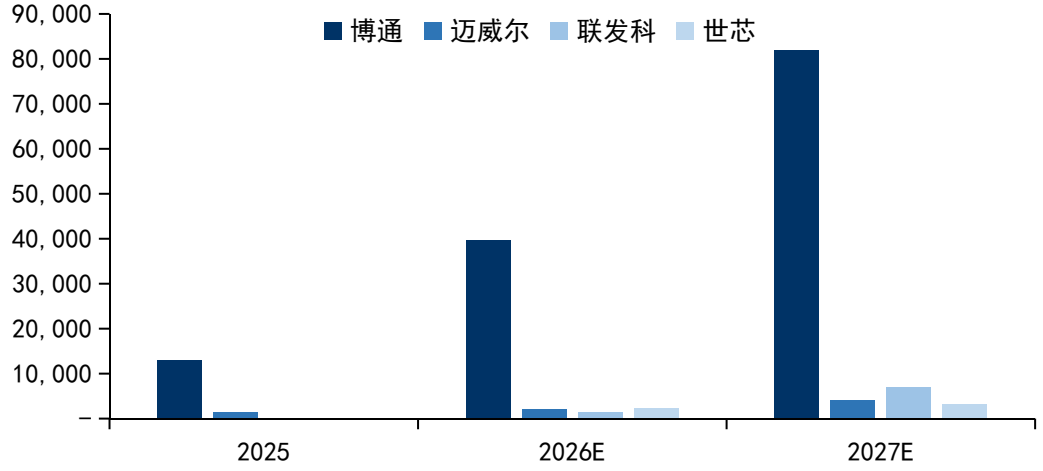
资料来源：公司财报、国信证券经济研究所整理测算

目录

谷歌TPU：寻找更多设计合作方，推动成本降低

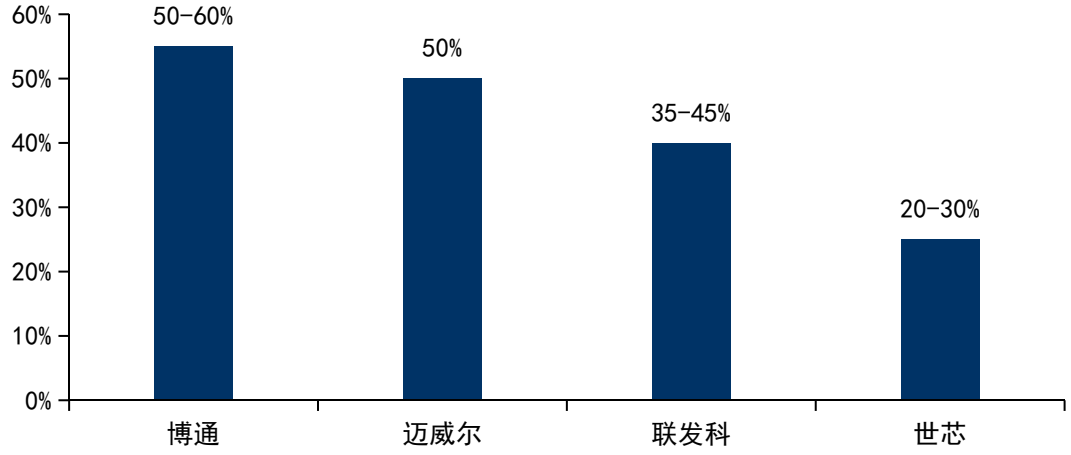
- **根据公司财报，博通在芯片设计中的毛利率达到接近60%，在中间环节加价约15%-20%。**博通是谷歌TPU的主要合作方，与谷歌在芯片设计中分工明确，谷歌负责整体芯片架构、算法及结构验证；博通承担芯片集成、电路实现（含互联、内存、电源管理相关电路设计等前端环节）及后端的布局布线、流片测试、封装交付等全流程工作。由于博通在其中打包销售了高毛利的 IO Chiplet（高速接口芯粒）和专用交换机芯片（如 Tomahawk 5/6 系列和分布式 X-Bar 架构），这部分网络 IP 的毛利率高达 75%+
- **26年4月，博通与谷歌签订合作，将为谷歌未来几代的张量处理单开发和供应定制的TPU。**此外，根据公司公告，博通还达成了一项供应保证协议，博通将为谷歌的下一代AI机架提供网络和其他组件，直至2031年。同时，博通、谷歌和Anthropic扩大了现有的战略合作。根据该合作，Anthropic将从2027年开始，通过博通接入约3.5GW的计算能力。
- **谷歌积极寻求摆脱对博通的依赖，与联发科、Marvell洽谈合作。**26年初官宣的TPUv8分为推理和训练两个版本，其中推理芯片TPUv8i由联发科负责。此外，根据The information，谷歌还在正与Marvell洽谈开发两款新型AI芯片：一是内存处理单元（MPU），用于协同TPU卸载内存计算任务，缓解高并发推理下的内存带宽瓶颈；二是针对推理场景优化的专用TPU，旨在降低单次AI响应的算力成本。

图：博通、迈威尔、联发科、世芯 ASIC业务收入



资料来源：公司财报、国信证券经济研究所整理测算

图：博通、迈威尔、联发科、世芯 ASIC业务毛利率



资料来源：公司财报、国信证券经济研究所整理测算

谷歌TPU：积极寻求台积电外产能

国信证券 GUOSEN SECURITIES

- **谷歌主要依赖台积电提供的CoWoS封装和晶圆生产，但由于需要同时支持自身模型业务发展、主业运行、以及云业务对外提供服务，谷歌存在较大的算力缺口，因此近两年积极寻求其他公司产能。**

1. **Intel已获得谷歌300万颗订单：**根据The information报道，谷歌在对英特尔的先进封装技术进行数月测试后，已向英特尔订购了超过300万颗TPU，预计将于2028年交付。
2. **Amkor、ASE承接CoWoS体系的外溢产能合作：**Amkor24年将台积电成熟的先进封装技术——晶圆基板芯片封装（CoWoS）和集成扇出封装（InFO）——引入美国市场。
3. **三星：**根据The Information，谷歌正在与三星电子洽谈，计划由三星承接第十代TPU芯片的部分组件，负责制造内存I/O Die，即连接计算核心与HBM等内存的桥梁，采用2nm GAA工艺制程。I/O Die主要负责SerDes、PCIe/CXL 等接口，以及与HBM的互联，同时相对更标准化，更适合三星。核心计算引擎（计算Die）则是由台积电代工，采用其最尖端的1.4nm先进制程。分开制造再拼接，既释放了台积电的产能压力，又能大幅降低整颗TPU的综合制造成本。

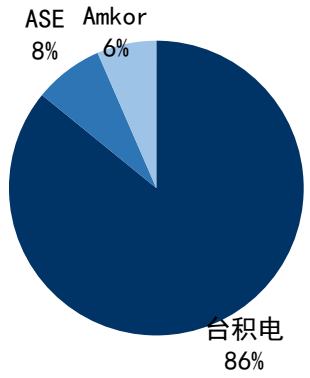
- **我们预计26年谷歌TPU接近500万颗产能中，台积电与ASE合计贡献接近450万颗，其余部分则由Amkor承接。**

图：半导体产业链价值量分布

环节	代表公司	价值量占比	特点
设计（Design）	英伟达、博通、高通、联发科	25%–40%	轻资产，高毛利，核心壁垒在于人才、IP积累和生态绑定。
制造（Fabrication）	台积电、三星电子、中芯国际	40%–60%	重资产、资本密集、规模效应显著
封装（Packaging）	日月光、安靠、长电科技、通富微电、华天科技、台积电	5%–20%	劳动密集，技术壁垒中等，AI时代先进封装价值量在提升
测试（Testing）	京元电子、日月光、长电科技、通富微电、华峰测控	2%–8%	劳动密集，技术壁垒中等

资料来源：各公司官网、国信证券经济研究所整理

图：TPU26年封装产能结构

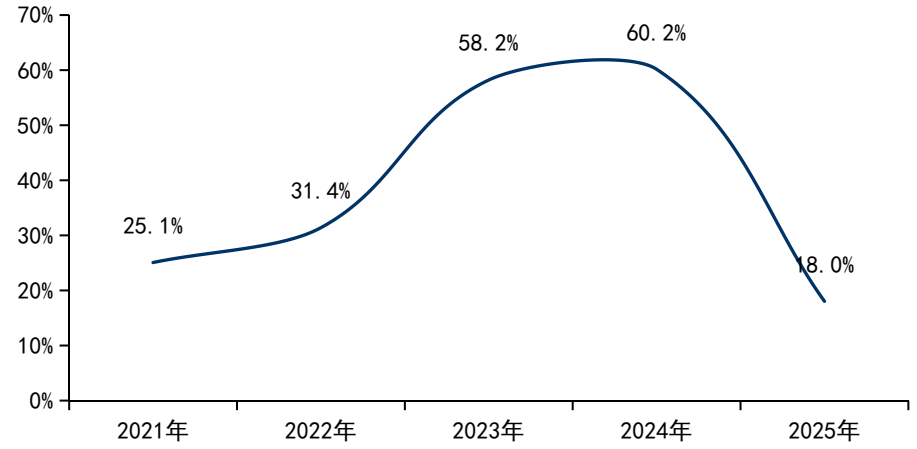


资料来源：公司财报、国信证券经济研究所整理测算

亚马逊Trainium：与Marvell和Alchip深度合作，逐渐收拢价值量

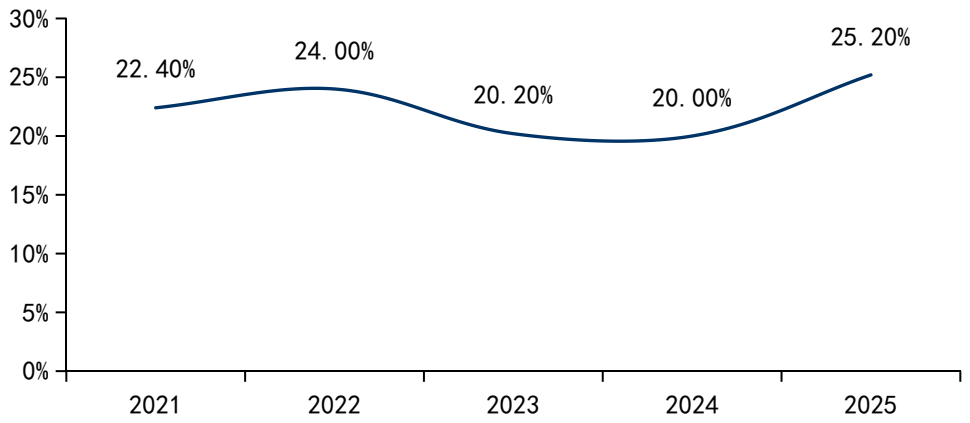
- **从世芯到Marvell，Trn1+Trn2 芯片设计和互联能力逐渐成熟。**根据SemiAnalysis，Trainium系列的前端设计由亚马逊自己完成，第一代版本后端设计和全套管理工作，包括台积电的流片均由Alchip完成，根据公司财报，预计2024年其收入的60%都由亚马逊贡献。采用台积电 5nm 工艺制造的Trainium2 芯片由Marvell负责，但其在Trainium2 项目上执行不佳，开发周期拖太久，并且在封装 RDL interposer 设计上遇到问题，后来需要 Alchip 介入，帮助交付可工作的方案。
- **Trn3世芯重新主导，但价值量进一步被亚马逊收拢。**根据SemiAnalysis，Trainium3 前端继续由亚马逊设计，PCIe SerDes 部分则获得了 Synopsys 的授权。后端物理设计和封装设计由 Alchip 负责。SemiAnaylsis认为 Trainium3 可能继承了 Marvell 设计的 Trainium2 的部分接口 IP。在Trn3中，预计存在两个流片项目，一个由 Alchip 拥有（称为“Anita”），另一个则由 Annapurna 直接拥有（称为“Mariana”）。在 Anita 项目中，Alchip 直接从台积电采购芯片组件；而在 Mariana 项目中，Annapurna 也直接采购芯片组件。大部分产量将用于 Mariana 项目。虽然 Alchip 在 Mariana 项目中的设计参与度与 Anita 项目类似，但他们从 Mariana 项目获得的收入应该会低于 Anita 项目。就每总拥有成本（TCO）的性能而言，亚马逊更加重视降低 TCO。

图：世芯电子来自亚马逊的收入占比



资料来源：公司财报、国信证券经济研究所整理

图：世芯电子毛利率



资料来源：公司财报、国信证券经济研究所整理

博通：同时受益于ASIC放量和连接需求增加，但ASIC份额受冲击

- **博通的主要收入驱动因素来自AI Semis（FY25占比大约30%），其中包括AI Networking，主要是以太网AI网络芯片，以及AI Compute（ASIC）。**
- **AI Compute（ASIC）受益于TPU放量以及Meta、OpenAI等公司的ASIC需求，但由于谷歌分散供应链份额将受到冲击。**大约90%收入来自TPU，预计FY26/27伴随TPU出货加速将迎来爆发期，但考虑到从TPUv8开始谷歌仅将训练芯片交给博通负责设计，我们预计FY26-FY28博通的TPU出货量为370/550/700万。预计博通ASIC芯片业务FY26-FY28收入分别为433/932/1328亿美元。
- **AI Networking受益于连接需求增加。**Tomahawk交换芯片是博通AI网络收入的主要来源，预计FY26-FY28出货量为69/110/130万张，对应产生的收入为104/198/234亿美元，此外其他部分中，PCIe交换芯片、DSP芯片、NIC也将迎来持续放量

图：博通收入利润预测

单位：百万美元	FY2024	FY2025	FY2026E	FY2027E	FY2028E
收入	51,574	63,887	105,807	173,088	226,142
YoY %	44%	24%	66%	64%	31%
Semis	30,096	36,858	72,084	137,988	191,354
YoY %	7%	22%	96%	91%	39%
AI Semis	12,252	20,138	56,542	119,255	164,474
YoY %		64%	181%	111%	38%
AI Networking	3,841	7,168	13,231	26,033	31,685
YoY %		87%	85%	97%	22%
交换芯片		5,808	10,350	19,800	23,400
YoY %			78%	91%	18%
出货量		0.53	0.69	1.1	1.3
单价（美元）		11,000	15,000	18,000	18,000
其他芯片		1,360	2,881	6,233	8,285
YoY %			112%	116%	33%
AI Compute	8,411	13,056	43,311	93,222	132,789
YoY %		55%	232%	115%	42%
TPU		11,750	40,700	88,000	126,000
出货量		3.0	3.7	5.5	7.0
单价（美元）		3,917	11,000	16,000	18,000
其他ASIC		1,306	2,611	5,222	6,789
			100%	100%	30%
Non-AI Semis	17,844	16,720	15,542	18,733	26,880
YoY %		-6%	-7%	21%	43%
Infra Software	21,478	27,029	31,946	37,228	37,789
YoY %	181%	26%	18%	17%	2%
毛利	32,509	43,294	72,969	119,693	162,195
毛利率	63%	68%	69%	69%	72%
OP	13,463	25,484	53,240	95,975	132,688
OPM	26%	40%	50%	55%	59%
Non-Gaap OP	30,736	41,997	70,115	113,940	151,061
Non-Gaap OPM	60%	66%	66%	66%	67%
Semis OP	16,759	21,232	44,366	79,840	111,053
Semis OPM	56%	58%	62%	58%	58%
Infra Software OP	13,977	20,765	26,227	36,332	37,033
Infra Software OPM	65%	77%	82%	98%	98%
net income	5,895	23,126	45,320	82,721	111,536
NPM	11%	36%	43%	48%	49%

资料来源：公司财报、彭博、国信证券经济研究所整理测算

迈威尔：ASIC放量、XPU Attach崛起、连接需求爆发

- **ASIC业务放量。**根据The Information，ASIC除了传统的 Amazon Trainium 项目，Microsoft 的 Maia 3nm 项目预计在明年进入大批量生产，同时公司正在与谷歌洽谈相关的合作。
- **XPU Attach（如 CXL 控制器、SmartNIC 等配套芯片） 的崛起。**公司预计Custom XPU 市场 2023-2028 CAGR 为 47%，XPU Attach 市场 CAGR 高达 90%，到 2028 年，XPU Attach TAM 将从 2023 年的约 6 亿美元增长至约 146 亿美元，占整个定制计算市场的约 26%。未来每个 AI Tray 内部的很多芯片都会逐渐定制化，因此 Attach 的设计机会数量将显著高于 XPU 本身。目前已有超过 20 个 Custom + XPU Attach design wins。
- **Marvell 在光互联 DSP 市场占据绝对主导地位，受益于连接需求增加。**在 800G 时代，其市场份额高达 70%，尽管博通布局1.6T对其形成一定竞争压力，但连接需求的爆发依然能够带动公司相关收入增长。

图：迈威尔收入利润预测

单位：百万美元	FY2023	FY2024	FY2025	FY2026	FY2027E	FY2028E	FY2029E
收入	5,920	5,508	5,767	8,195	11,518	16,763	23,361
		-7%	5%	42%	41%	46%	39%
Data center	2,409	2,217	4,164	6,100	9,210	14,313	20,959
		-8%	88%	46%	51%	55%	46%
AI				3,370.0	5,416.2	9,184.6	13,835.6
					61%	70%	51%
光				1,950.6	3,686.4	5,582.7	7,119.1
					89%	51%	28%
ASIC/XPU attach			666.5	1,419.4	1,777.5	3,601.9	6,716.5
						103%	86%
其他非AI				2,730	3,794	5,129	7,123
					39%	35%	39%
Comms and Other	3,511	3,291	1,603	2,094	2,221	2,361	2,492
		-6%	-51%	31%	6%	6%	6%
毛利	2,988	2,294	2,382	4,181	6,023	8,911	12,753
毛利率	50%	42%	41%	51%	52%	53%	55%
OP	238	-568	-720	1,323	2,172	4,383	7,731
OPM	4%	-10%	-12%	16%	19%	26%	33%
net income	-164	-933	-885	2,670	1,514	3,631	6,110
NPM	-3%	-17%	-15%	33%	13%	22%	26%

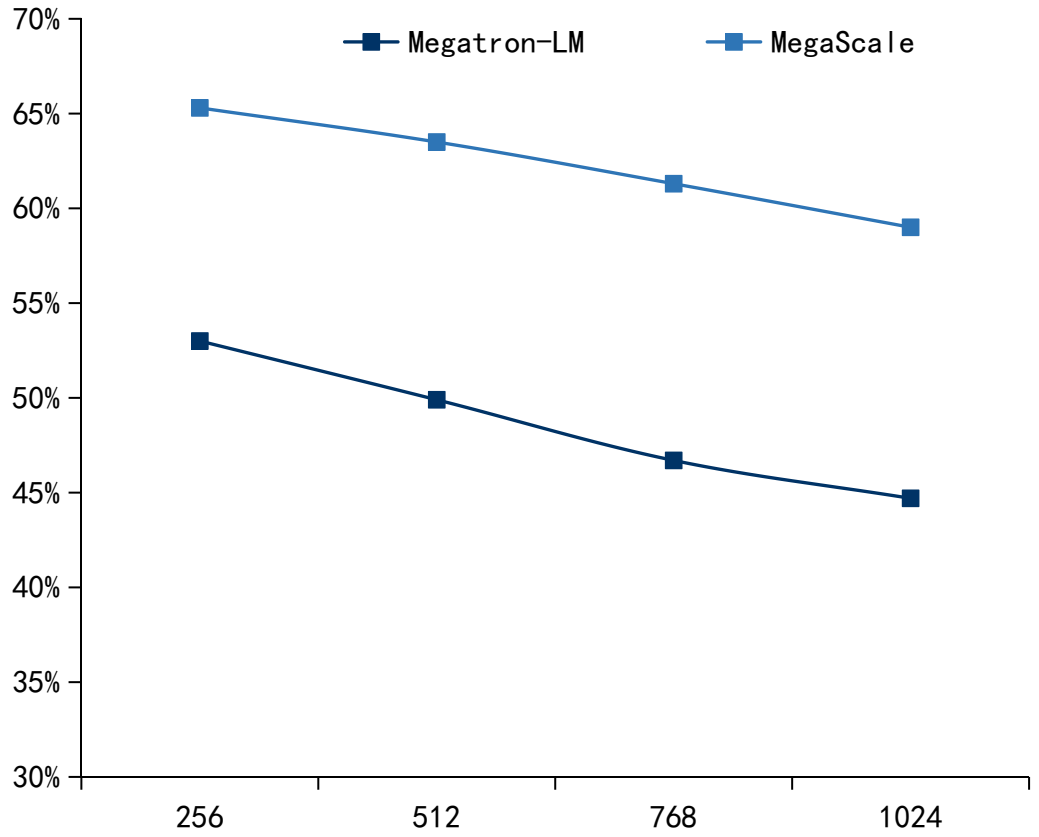
资料来源：公司财报、彭博、国信证券经济研究所整理测算

目录

网络与互联基础设施：算力规模扩大后，连接成为关键

- Marvell CEO Murphy 6月COMPUTEX大会上提出了一条核心结论：**AI基础设施的瓶颈正在从算力和存储转向连接，而铜缆正在逼近物理极限，光学替代不可逆。具体来看，AI基础设施的瓶颈正在经历第三次迁移。**
- 第一次瓶颈是算力。行业需要远超传统的计算能力来支撑AI训练和推理。
- 第二次瓶颈是存储。更大的模型需要更大的存储容量和更高的带宽。HBM高带宽存储器成为关键资源，它把存储颗粒垂直堆叠后紧贴GPU封装，提供远超传统内存的带宽。
- **第三次瓶颈是连接。**现代AI需要数万乃至数百万个处理器协同工作，构成一台超大规模计算引擎。当计算问题被拆解成无数子任务分散到整个数据中心执行时，连接就是一切。 推理模型、MoE混合专家架构（一个大模型被拆成多组"专家"子网络，每次推理只激活一小部分，但不同专家可能分布在不同芯片上）、agentic AI（能自主规划任务、调用工具、持续执行的AI系统）都在推高对连接带宽和延迟的要求。

图：当集群规模扩大时（横轴是GPU个数），集群利用效率会下降，连接的重要性凸显



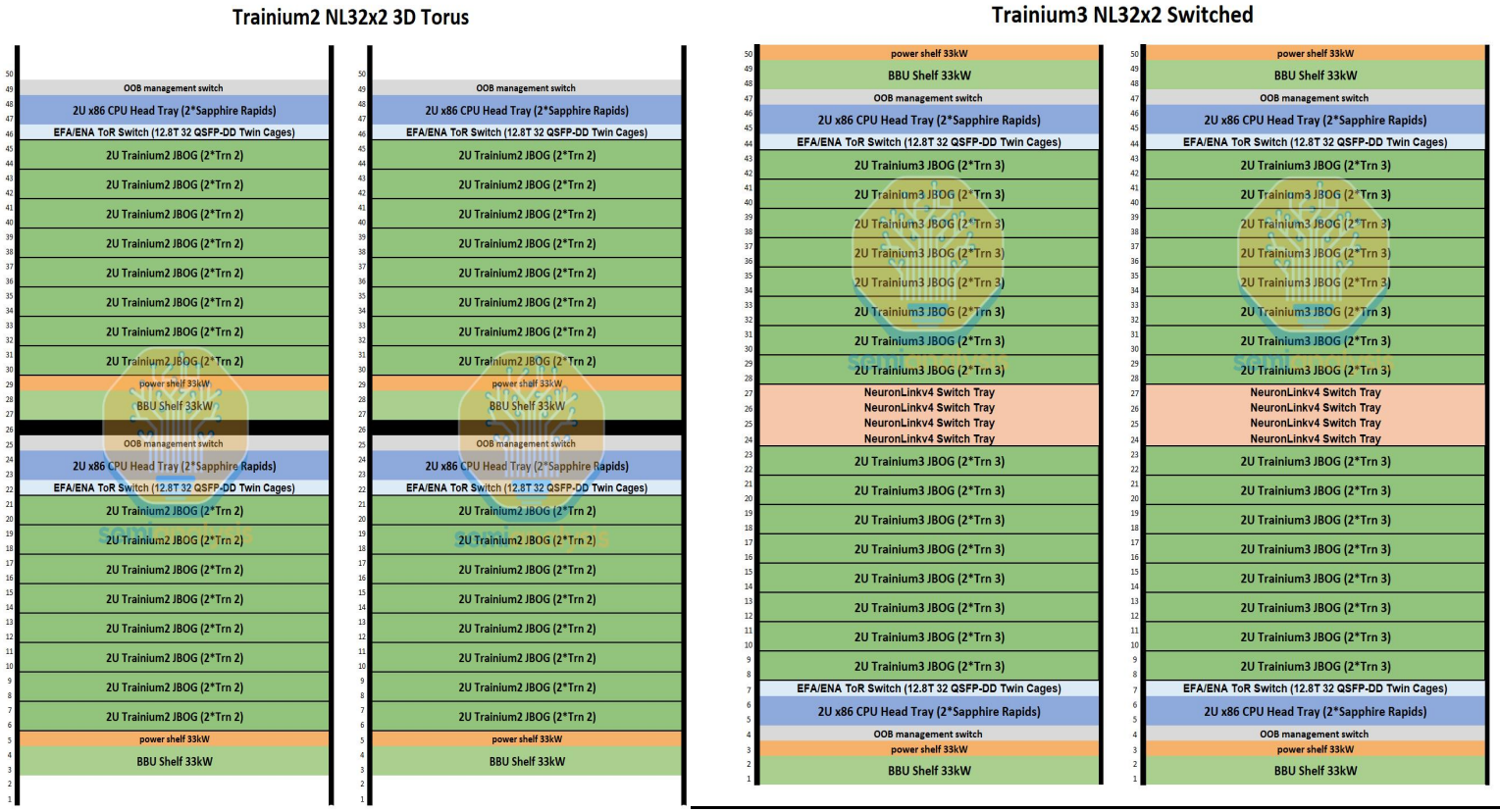
资料来源：《MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs》、国信证券经济研究所整理

Scale Up：亚马逊Scale Up方案较谷歌更激进，PCIe交换芯片需求增加

国信证券 GUOSEN SECURITIES

- **亚马逊Scale Up方案较谷歌更激进，PCIe交换芯片需求增加。**在Scale up网络中，Trainium2 使用的是2D/3D Torus，与谷歌的思路类似，但环面拓扑结构并不是最适合MOE模型的网络结构，采用3D Torus 架构时，扩展网络会因过载而突然面临带宽瓶颈。为应对以上问题，Trainium3从 2D/3D 环面架构转向了交换式架构，引入 NeuronSwitch-v1，这带来了大量交换芯片需求的增加。

图：Trn2和Trn3 NL32X2 机柜方案



资料来源：Semi Analysis、国信证券经济研究所整理测算

Trainium3带来PCIe交换芯片需求量增加

国信证券 GUOSEN SECURITIES

- **采用NeuronSwitch-v1带来了大量的增量PCIe芯片需求。**Trainium3 的生命周期内，共推出了三代Scale up交换机，而每代升级意味着需要大量的 PCIe芯片。第一代是 160 通道的 Scorpio X PCIe 6.0 交换机，第二代是 320 通道的 Scorpio-X PCIe 6.0 交换机，第三大还可以选择升级到更高基数的 72+ 端口 UALink 交换机。。
- **第一代交换机需求数量与芯片数量比例为1:2左右。**以第一代版本为例，在32x2的服务器节点中需要32个交换芯片，但在72x2的服务器节点中泽需要368个交换芯片（其中288个36通道的pcie交换芯片），如果仅看160通道的PCIe芯片，需求数量与芯片数量比例为1:2左右，在第二代中的需求量有所减少，约为1:4。

图：Trn3 Scale-up交换机路线图

	单位	Gen1(Scorpio X 160通道)		Gen2(Scorpio X 320通道)		Gen3(Scorpic X UALink)	
Trainium3服务器SKU		NL32x2 Switched	NL72x2 Switched	NL32x2 Switched	NL72x2 Switched	NL32x2 Switched	NL72x2 Switched
交换机规格对比（按代际）							
PCIe代际	无	6.0		6.0		7.0	
PCIe物理通道数（PHY）	无	160		320		?	
每通道带宽（单向）	Gbit/s	64		64		128	
每端口通道数	无	8		8		?	
端口数量	无	20		40		72+	
单交换机带宽（单向）	**Gbit/s**	**10240**		**20480**		**?**	
每机架拓扑结构（按具体SKU）							
每机架交换平面数	无	2	4	1	2	1	1
每交换平面交换机数量	无	8	10	8	10	?	?
每机架Switch Tray交换机总数	无	16	40	8	20	?	?
每机架PCB上32通道交换机数	无	0	144	0	144	0	144
每机架横向扩展交换机总数	无	16	184	8	164	-	?+144
单服务器横向扩展交换机总数	**无**	**32**	**368**	**16**	**328**	**?x2**	**?x2+288**
160通道交换机连接率（每Tm3）	无	0.50	0.56	-	-	-	-
320通道交换机连接率（每Tm3）	无	0.50	0.56	0.25	0.28	-	-
32通道交换机连接率（每Tm3）	无	-	2.00	-	2.00	-	2.00
机箱内最大跳数	**无**	**2**	**3**	**1**	**3**	**1**	**1**
横向扩展域内最大跳数	**无**	**3**	**4**	**2**	**4**	**2**	**2**

资料来源：Semi Analysis、国信证券经济研究所整理测算

Scale out：Google TPU采用OCS架构，打开AI光互联新价值量

- Google在TPU集群中率先规模化采用OCS。OCS通过直接调度光路实现端口连接，避免传统EPS的光电/电光转换，具备低功耗、低时延、速率无关等优势，适合训练场景中相对稳定的大规模互联流量。
- OCS价值量主要集中在MEMS芯片、校准系统和2D准直阵列。以176x176 MEMS OCS为例，BOM约13k美元，MEMS芯片/校准系统/2D collimator array占比分别约35%/30%/15%。
- TPUv7以4x4x4、64颗TPU cube为ICI基本单元，每颗TPU在3D Torus中连接6个邻居，其中2个固定通过PCB走线连接，其余4条按位置采用DAC或光模块；face/edge/corner TPU分别需要1/2/3个光模块，推导出整柜96个光模块、**1.5个/TPU attach ratio**，并通过OCS实现cube间连接与环回。

图：176x176 MEMS OCS BOM拆分

MEMS-176x176平台	成本($)	制造成本占比	BOM占比
MEMS芯片	4.5	32%	35%
校准系统	4.0	28%	31%
2D准直阵列	2.0	14%	16%
2D MEMS阵列	0.3	2%	2%
分束器	0.5	4%	4%
其他组件	1.4	10%	11%
制造	1.4	10%	11%
总制造成本	**14.1**	**100%**	
BOM成本	**12.7**		**100%**
销售价格		**32**	
毛利率		**56%**	

资料来源：Semi Fundamental、国信证券经济研究所整理测算

图：Google TPU v7 3D Torus连接数量测算

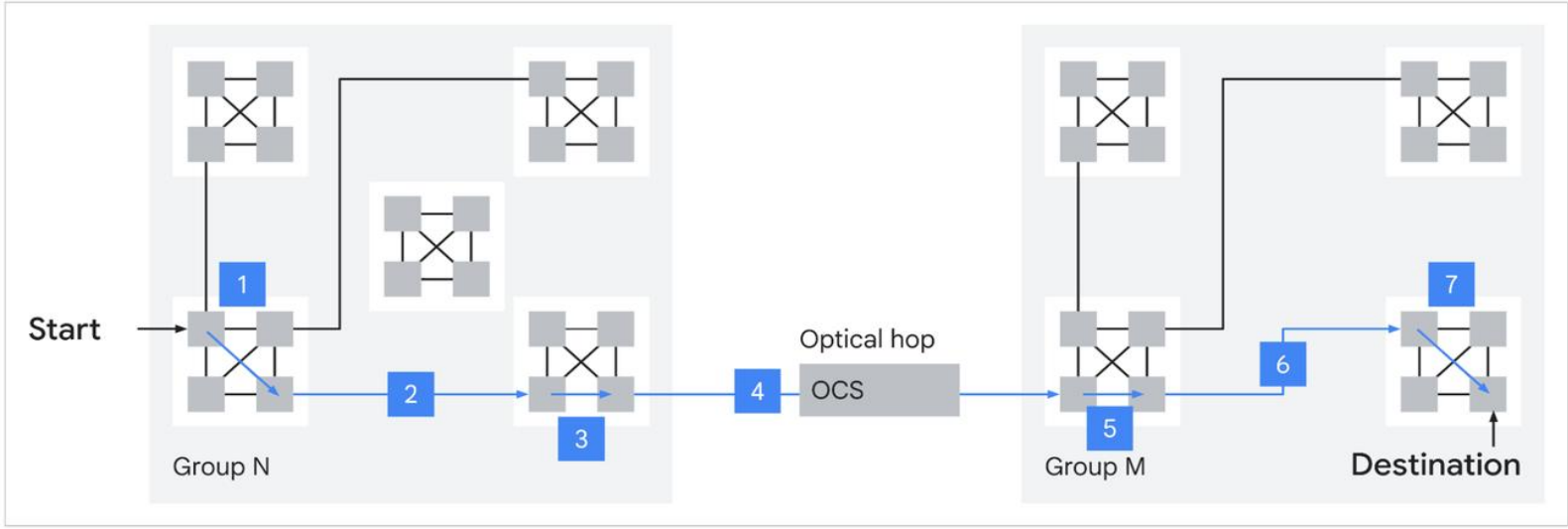
TPU位置	数量	铜缆/TPU	PCB/TPU	光模块/TPU	光模块合计
内部TPU	8	4	2	0	0
角部TPU	8	1	2	3	24
边缘TPU	24	2	2	2	48
面部TPU	24	3	2	1	24
合计/均值	**64**	**1.25**	**1.00**	**1.50**	**96**

资料来源：Semi Analysis、国信证券经济研究所整理测算

TPUv8：v8i采用Boardfly 拓扑，OCS进入scale up

- **TPUv8t：沿用3D Torus + ICI + OCS 的训练型架构，同时谷歌推出 scale-out架构Virgo Network，**通过增加每个交换机的端口数量来减少网络层数，并采用扁平化的两层无阻塞拓扑结构。可以在单个结构中连接 134,000 个芯片（TPU 8t），提供高达 47 petabits/秒的无阻塞双向带宽。
- **TPUv8i：为推理/采样/MoE 场景引入了新的 Boardfly 拓扑，放弃了环形结构，将OCS引入scale up。**虽然环面对于密集训练中常见的邻居间通信非常高效，但它会给所有芯片间的通信模式带来额外的延迟。在推理模型和迭代优化（MoE）时代，任何芯片都可能需要与其他任何芯片通信来路由令牌，因此跳数至关重要。通过增加板组之间的直接光纤长距离链路数量，同样 1,024 颗芯片，如果用 8×8×16 的 3D Torus，最远通信需要 16 hops；Boardfly 把网络直径降到 7 hops，网络直径减少 56%，通信密集型工作负载延迟最多改善 50%。
- **根据SemiAnalysis，TPUv7的9216TPU POD中，对应需要48个OCS交换机，TPU:OCS=192:1，但在TPUv8i中，TPU与OCS的配比关系为58：1，OCS的比例明显增加。**

图：TPU 8i pod 上通过光路交换机实现的最大七跳 ICI 网络直径的可视化表示



资料来源：Google Cloud、国信证券经济研究所整理

OCS市场28年有望超过60亿美元，LITE、COHR有望受益

- **仅考虑TPU采用的OCS方案，我们测算，假设28年TPU出货量达到3000万，OCS市场将突破60亿美元。**
- 早期谷歌部署基于自主研发的MEMS OCS交换机，但近年来已开始转向Lumentum等商业MEMS OCS供应商。Lumentum在最近一次财报电话会议上表示，其OCS硬件订单积压总额达4亿美元，这些订单来自三家不同的客户，均为公司新增订单。Coherent公司也在不断增加客户参与度，其OCS订单积压量也逐季度增长。

图：OCS供应链公司

公司	OCS业务布局
Lumentum (LITE)	主要生产基于MEMS的OCS芯片，但也具备一定的LCoS芯片生产能力。其主要商业产品是为谷歌（主要）和甲骨文提供的300x300 OCS芯片。此外，该公司还向微软和英伟达发送了64×64 OCS芯片样品进行测试。Lite的OCS生产基地位于泰国，预计将于2026年第四季度开始量产。目标是在2027年实现10亿美元的OCS芯片年化销售额。
Coherent (COHR)	Coherent预计将于2026年向谷歌交付约3000台基于LCoS的OCS设备，量产计划将于2026年上半年启动。公司表示已收到来自谷歌和甲骨文超过3亿美元的订单。
Silex	赛微电子持有Silex 45%的股份。谷歌OCS系统的主要MEMS系统供应商

资料来源：公司财报、国信证券经济研究所整理

图：OCS 市场规模预测

	2024	2025	2026	2027	2028
TPU出货量（百万）	2.4	3	5	12	30
OCS端口数/TPU端口	1	1.5	1.5	1.5	1.5
OCS端口总数（百万个）	2.4	4.5	7.5	18	45
OCS设备ASP（美元/台，600端口）	100,000	100,000	95,000	90,250	81,225
OCS市场规模（百万美元）	400	750	1,188	2,708	6,092

资料来源：Google、国信证券经济研究所整理测算

Scale out：集群规模扩大增加拓扑网络层数，光芯比提升

- 英伟达在GB300的SuperPOD参考架构中给出了从1152张GPU到9216张GPU的计算网络和存储网络，根据我们测算，**当前计算网络的光模块需求与GPU比例为2.5:1，加上存储网络则接近4:1，**区别在于计算网络在交换机侧用1.6T光模块，服务器侧和存储网络都使用的是800G光模块。
- **Scale Out网络的层数增加，英伟达案例中最大接近万卡集群都用的是2层fabric，但如果进一步扩大，则可能需要三层结构，需要更多交换机和光模块。**

图：DGX GB300 Super POD光芯比测算

		DGX GB300 Super POD			
计算网络	fabric层数	2	2	2	2
	GPU 数	1152	2304	4608	9216
	SU 数	2	4	8	16
	每个 SU 的 IB leaf	8	8	8	8
	IB leaf 总数	16	32	64	128
	IB spine 总数	8	18	36	72
	服务器侧光模块数量	1152	2304	4608	9216
	交换机侧光模块数量（Q3400-RD，72个cage）	1728	3600	7200	14400
	光模块总数	2880	5904	11808	23616
	计算网络光芯比	**2.5**	**2.6**	**2.6**	**2.6**
存储网络	Spine数（SN5600D，64个cage，下同）	4	8	12	24
	计算leaf	8	16	32	64
	存储leaf	2	2	4	8
	管理leaf	2	2	2	2
	OOB spine（SN2201，48个cage）	2	2	2	2
	交换机侧光模块数量	1120	1888	3296	6368
	服务器侧光模块数量	576	1152	2304	4608
	光模块总数	1696	3040	5600	10976
	合计光模块总数	4576	8944	17408	34592
	光芯比	**4.0**	**3.9**	**3.8**	**3.8**

资料来源：英伟达，国信证券经济研究所整理测算

COHR：光模块需求持续释放

国信证券 GUOSEN SECURITIES

- **Coherent的核心增长驱动来自于Datacom部分的光模块以及光器件、OCS等，其中贡献最多的仍然是光模块，**受益于连接需求增加，从800G到1.6T以及后续的3.2T，ASP和需求量持续增长。

图：Coherent收入利润预测

单位：百万美元	FY2022	FY2023	FY2024	FY2025	FY2026E	FY2027E	FY2028E	
收入		**3,317**	**5,160**	**4,708**	**5,810**	**7,060**	**9,615**	**12,725**
YoY %			56%	-9%	23%	22%	36%	32%
Datacenter&Communications					**3,756**	**5,190**	**7,578**	**10,030**
YoY %						38%	46%	32%
Datacom			1180.5	1573.9	2594.6	3820.4	5882.0	8479.2
YoY %				33%	65%	47%	54%	44%
<400 Gbps			1015.9	758.3	916.9	966.8	870.1	783.1
YoY %				-25%	21%	5%	-10%	-10%
出货量				3.0	3.7	3.9	3.5	3.1
单价				250.0	250.0	250.0	250.0	250.0
800+ Gbps			0.0	605.0	1461.2	2631.7	4783.3	7460.7
YoY %					142%	80%	82%	56%
出货量				1.2	2.9	5.3	9.6	14.9
单价				500.0	500.0	500.0	500.0	500.0
其他			164.6	210.6	216.5	221.9	228.6	235.5
Other					1161.4	1369.6	1696.0	1550.8
YoY %						18%	24%	-9%
Industrial					2,054	1,873	1,957	2,044
YoY %						-9%	4%	4%
分拆二：								
Networking		2,197	2,341	2,296	3,421	4,365	5,871	8,427
						28%	35%	44%
Materials		1,119	1,350	1,017	954	1,250	1,461	1,642
						31%	17%	12%
lasers		0	1,469	1,395	1,435	1,454	1,755	1,726
						1%	21%	-2%
毛利		1266	1618	1456	2043	2,650	3,774	5,181
毛利率		38%	31%	31%	35%	38%	39%	41%
OP		414	-37	96	290	880	1,542	2,355
OPM		12%	-1%	2%	5%	12%	16%	19%
net income		235	-404	-159	49	806	1,327	2,082
NPM		7%	-8%	-3%	1%	11%	14%	16%

资料来源：公司财报、彭博、国信证券经济研究所整理测算

LITE：OCS贡献弹性，光模块需求放量

- **连接需求增加，光模块放量，同时从800T到1.6T升级ASP翻倍。**1.6T 模块的 ASP 是 800G 的两倍，但成本和裸片尺寸相似，具有显著更高的毛利率。随着 1.6T 在今年夏季开始规模出货，叠加内部 CW 激光器集成比例的提升，指引OPM持续提升。
- **OCS获得大单。**LITE 获得了一份多年期、价值数十亿美元的 OCS 采购协议，预计仅在 2026 年下半年，OCS 的出货量就将超过 4 亿美元。目标是在2027年实现10亿美元的OCS芯片年化销售额。

图：LITE收入利润预测

单位：百万美元	FY2022	FY2023	FY2024	FY2025	FY2026E	FY2027E	FY2028E
收入	1,713	1,767	1,359	1,645	2,997	5,374	8,272
YoY %		3%	-23%	21%	82%	79%	54%
Cloud & Networking		1,322.5	1,084.9	1,410.8	2,744.3	5,117.9	8,002.6
YoY %			-18%	30%	95%	86%	56%
EML		150.0	163.6	272.7	627.1	983.3	1,173.6
YoY %			9%	67%	130%	57%	19%
Cloud Light			199.5	310.0	793.8	1,462.2	1,992.2
YoY %				55%	156%	84%	36%
OCS/CPO			-	-	153.9	1,100.0	2,950.0
YoY %						615%	168%
Telecom		832.1	721.8	815.4	1,169.5	1,572.4	1,886.8
YoY %			-13%	13%	43%	34%	20%
Industrial Tech		444.5	274.3	234.2	252.7	256.4	269.9
YoY %			-38%	-15%	8%	1%	5%
Components				1,116	1,992	3,312	5,013
YoY %					78%	66%	51%
Systems				529	1,005	2,062	3,259
YoY %					90%	105%	58%
毛利	789	569	252	460	1,219	2,572	4,285
毛利率	46%	32%	19%	28%	41%	48%	52%
OP	303	-116	-434	-180	504	1,780	3,263
OPM	18%	-7%	-32%	-11%	17%	33%	39%
Non-Gaap OP	527	339	38	160	873	2,170	3,586
Non-Gaap OPM	31%	19%	3%	10%	29%	40%	43%
net income	199	-132	-547	26	456	1,508	2,644
NPM	12%	-7%	-40%	2%	15%	28%	32%

资料来源：公司财报、彭博、国信证券经济研究所整理测算

ALAB：PCIe交换芯片放量

- ALAB是一家AI连接半导体解决方案提供商，主要依靠Aries Retimer拿下了PCIe 4.0/5.0时代90%以上的市场份额。
- **从PCIe Retimer（Aries）单品垄断向PCIe Switch（Scorpio）及光学连接平台跨越，PCIe交换芯片成为重要增量。**随着Scorpio P（Scale-out）和Scorpio X（Scale-up）在2026年下半年加速出货，管理层预计到2026年底，Scorpio将超越Aries成为公司最大的产品线。此外，通过收购aiXscale，公司正将触角延伸至2027-2028年的NPO/CPO（共封装光学）市场，试图在多机架集群的光互联时代提前卡位。

图：ALAB收入利润预测

单位：百万美元	FY2023	FY2024	FY2025	FY2026E	FY2027E	FY2028E
收入	115.8	396.3	852.5	1,546.20	2,196.10	2,817.10
		242%	115%	81%	42%	28%
Aries				729.9	787.5	874.7
					8%	11%
Taurus				183.6	159.3	230.9
					-13%	45%
Leo				19.9	57.3	134.7
					188%	135%
Scorpio				536.6	1,197.60	1,567.50
					123%	31%
毛利	79.8	302.7	645.3	1,125.30	1,601.10	2,077.30
毛利率	69%	76%	76%	73%	73%	74%
OP	-29.5	-116.1	173.4	367	661.6	780.6
OPM	-25%	-29%	20%	24%	30%	28%
net income	-26.3	-83.4	219.1	395.1	643	852.2
NPM	-23%	-21%	26%	26%	29%	30%

资料来源：公司财报、彭博、国信证券经济研究所整理测算

风险提示

第一，宏观经济波动。若宏观经济波动，公司业务、产业变革及新技术的落地节奏或将受到影响。

第二，下游需求不及预期。若下游AI需求不及预期，相关的AI研发投入增长或慢于预期，致使行业增长不及预期。

第三，核心技术水平升级不及预期的风险。AI大模型研发进度落后，AIGC相关产业技术壁垒较高，核心技术难以突破，影响整体进度。

第四，AI快速迭代、平权化下竞争加剧，影响云业务利润率。

免责声明

国信证券 GUOSEN SECURITIES

国信证券投资评级

投资评级标准	类别	级别	说明
报告中投资建议所涉及的评级（如有）分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后6到12个月内的相对市场表现，也即报告发布日后的6到12个月内公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。A股市场以沪深300指数（000300.SH）作为基准；新三板市场以三板成指（899001.CSI）为基准；香港市场以恒生指数(HSI.HI)作为基准；美国市场以标普500指数(SPX.GI)或纳斯达克指数（IXIC.GI）为基准。	股票投资评级	优于大市	股价表现优于市场代表性指数10%以上
		中性	股价表现介于市场代表性指数±10%之间
		弱于大市	股价表现弱于市场代表性指数10%以上
		无评级	股价与市场代表性指数相比无明确观点
	行业投资评级	优于大市	行业指数表现优于市场代表性指数10%以上
		中性	行业指数表现介于市场代表性指数±10%之间
		弱于大市	行业指数表现弱于市场代表性指数10%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道；分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求独立、客观、公正，结论不受任何第三方的授意或影响；作者在过去、现在或未来未就其研究报告所提供的具体建议或所表述的意见直接或间接收取任何报酬，特此声明。

重要声明

本报告由国信证券股份有限公司（已具备中国证监会许可的证券投资咨询业务资格）制作；报告版权归国信证券股份有限公司（以下简称“我公司”）所有。本报告仅供我公司客户使用，本公司不会因接收人收到本报告而视其为客户。未经书面许可，任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。

本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司不保证本报告所含信息及资料处于最新状态；我公司可能随时补充、更新和修订有关信息及资料，投资者应当自行关注相关更新和修订内容。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券并进行交易，还可能为这些公司提供或争取提供投资银行、财务顾问或金融产品等相关服务。本公司的资产管理部门、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意见或建议不一致的投资决策。

本报告仅供参考之用，不构成出售或购买证券或其他投资标的要约或邀请。在任何情况下，本报告中的信息和意见均不构成对任何个人的投资建议。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。投资者应结合自己的投资目标和财务状况自行判断是否采用本报告所载内容和信息并自行承担风险，我公司及雇员对投资者使用本报告及其内容而造成的一切后果不承担任何法律责任。

证券投资咨询业务的说明

本公司具备中国证监会核准的证券投资咨询业务资格。证券投资咨询，是指从事证券投资咨询业务的机构及其投资咨询人员以下列形式为证券投资人或者客户提供证券投资分析、预测或者建议等直接或者间接有偿咨询服务的活动：接受投资人或者客户委托，提供证券投资咨询服务；举办有关证券投资咨询的讲座、报告会、分析会等；在报刊上发表证券投资咨询的文章、评论、报告，以及通过电台、电视台等公众传播媒体提供证券投资咨询服务；通过电话、传真、电脑网络等电信设备系统，提供证券投资咨询服务；中国证监会认定的其他形式。

发布证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券
GUOSEN SECURITIES

国信证券经济研究所

深圳

深圳市福田区福华一路125号国信金融大厦36层

邮编：518046　　总机：0755-82130833

上海

上海浦东民生路1199弄证大五道口广场1号楼12楼

邮编：200135

北京

北京西城区金融大街兴盛街6号国信证券9层

邮编：100032