DEEP TECH & CAPITAL NARRATIVE

跨越硅基鸿沟

中国AI算力底座突围与超节点时代

国产AI芯片替代 · 超节点工程 · 全球算力供应链深度研究



Yavin质观

硬科技与资本叙事研究者，长期关注技术、产业、政策与资本化逻辑，
书写技术时代、现代生活与精神处境中的缓慢观察。

扫码关注公众号

EXECUTIVE SUMMARY · 核心执行摘要

六大核心产业判断与结论锚点

Yavin质观 · 旗舰智库专刊

66%	50.6%	75%–95%	1,024卡	400亿元
2026H1中国AI加速卡本土品牌份额	非GPU出货占比（+80%首次反超GPU）	国产卡推理单Token与硬件成本降幅	首套昇腾Atlas 950液冷超节点点亮	字节跳动2026年锁定昇腾950PR大单

01
市场与架构倒挂

市场与架构“双倒挂”：本土份额斩获66%，非GPU首次反超通用GPU

2026H1中国AI芯片出货 **200.8万颗**，昇腾（**28%**）与英伟达（**34%**）差距缩至 **6个百分点**；推理主导下 **非GPU（占50.6%）首次反超GPU**，本土四强包揽 **95%** 份额。

02
算力经济学双轨

算力经济学双轨制：推理降本95%接管 vs. 预训练4.5倍堆卡与4倍电费

推理侧国产卡达H100逾 **80%** 性能、单Token成本直降 **75%–95%**；**预训练侧**受40%–50% MFU约束，对齐B200算力需承受 **4.5倍堆卡**（约284亿元）与 **4倍电费溢价**。

03
超节点与算法奇点

系统与算法双奇点：千卡液冷超节点代偿单卡，“模型迁就芯片”破壁CUDA

首套 **1024卡昇腾950超节点** 点亮（**1 EFLOPS** 算力、**200ns** 互联）；DeepSeek开源 **TileLang-Ascend** 达手写算子 **98%** 性能，**TransMLA** 削减 **62.5%–94.7%** 显存。

04
商业锁单与ASIC

商业兑现与全球ASIC共振：字节400亿锁单，本土双雄率先规模盈利

字节超400亿大单锁死950PR近半产能，平头哥PPU出货 **47万颗**，**海光**（H1净利17.98亿）与 **寒武纪**（营收+108%）规模盈利；定制ASIC增速（**+44.6%**）首超GPU。

05
晶圆存储封装突围

底层制造三级突围：40%晶圆缺口下的“三维架构换平面代差”

中芯7nm对420万颗需求存 **40%缺口**；**长鑫**奇袭 **4F² DDR5** 并推进 **12层HBM3/3E联调**，**长江存储** eSSD居全球第三（**14%**），**盛合晶微**占2.5D封装 **85%**。

06
终局推演与风险标尺

终局推演与三大验金石：封装与HBM定上限，高集中度暗藏波动

“封装+HBM+互连” 取代制程成第一瓶颈，近90%客户集中度暗藏风险；未来18个月成败取决于 **万卡预训练MFU破55%**、**智算亮灯率超60%** 与 **晶圆/HBM良率爬坡**。

TABLE OF CONTENTS · 报告目录

全文八大核心章节与研究导览

15 幅数据图表 · 11 张全景矩阵表 · 320 项溯源信源

01 MARKET LANDSCAPE & INFLECTION

破局与市场洗牌——当200.8万张加速卡完成历史性倒挂，量变在哪里发生?

- 1.1 市场全局与结构倒挂：从单极垄断迈向66%份额的非对称国产替代
- 1.2 非GPU首次反超GPU背后：算力主战场从预训练军备竞赛转向“推理吞噬一切”
- 1.3 自用闭环盘与外部市场盘的真相：左手倒右手的生态闭环与公开市场的残酷搏杀

02 WORKLOADS & BENCHMARKS

客户选择、工作负载与实测对标——真正的替代发生在哪些业务？算力经济学的两本账怎么算?

- 2.1 工作负载的金字塔替代梯度：推理侧攻城略地，预训练仍是无人区
- 2.2 中美头部算力芯片实测工程剖析：训练与推理的剪刀差及三大本土技术流派
- 2.3 算力经济学的两本账：推理单Token的极致降本与预训练TCO的隐性通胀

03 SUPERNODE ARCHITECTURE

系统级能力突围——“单卡不够系统来凑”，超节点（SuperPoD）的暴力工程学与物理商业边界

- 3.1 暴力工程学：用光电互联与统一内存池击穿MoE通信墙与显存墙
- 3.2 超节点的三大硬约束：当物理制程短板转嫁为机房改造、电力黑洞与生态绑定

04 SOFTWARE STACK & ALGORITHMS

国产算力软件栈破冰与算法革命——从千万行代码开源、TileLang-Ascend 1:1映射到“模型主动迁就芯片”与MLA显存通缩

- 4.1 传统CANN高昂迁移账单与国产软件栈三条进化路径分野
- 4.2 跨越CUDA护城河的奇点：DeepSeek五大开源算子矩阵、TileLang-Ascend编译破壁与“模型迁就芯片”
- 4.3 MLA与TransMLA / DoseRank / MLAfication后训练无损迁移：算法效率十倍跃升对远期HBM与高端GPU弹性的重塑

05 DOMESTIC CHIP CONTENDERS

国产GPU与ASIC四大阵营分化——财务兑现、产品矩阵与中国科技巨头自研芯片全景

- 5.1 本土算力芯片四梯队财务兑现、研发强度与盈利收敛实绩
- 5.2 中国四大互联网科技大厂自研ASIC/GPGPU布局：阿里平头哥、腾讯紫霄、百度昆仑芯与字节跳动
- 5.3 核心本土厂商替代路径全景对比矩阵

06 SUPPLY CHAIN & CUSTOMER ORDERS

全球与中国AI算力芯片上下游供应链及客户生态深度剖析

- 6.1 产业供应链重构与“上下游集中度双向极化”特征
- 6.2 华为昇腾：全栈国产化供应链垂直整合与互联网、运营商双轮驱动（含字节400亿锁单与三大运营商集采明细）
- 6.3 海光信息与寒武纪：晶圆封装底座绑定与服务器OEM关联生态
- 6.4 海外AI算力巨头与定制化ASIC供应链全景对照：英伟达Blackwell闭环、博通与谷歌TPU Turnkey联盟及吉瓦级长单

07 FOUNDRY, MEMORY & PACKAGING

底层制造、AI存储三级突围、先进封装与光互联卡点——决定算力供给上限的四大物理支点

- 7.1 先进晶圆制造：中芯国际7nm DUV的极限压榨与良率爬坡
- 7.2 AI存储三级突围：长鑫存储 4F² DDR5 RDIMM、12层HBM3/HBM3E试产联调与长江存储PCIe 5.0企业级eSSD全栈替代
- 7.3 先进封装与CoWoS替代：盛合晶微独占国内硅基2.5D八成份额，碳化硅（SiC）中介层蓄势待发
- 7.4 集群光互联：中国光模块统治全球市场，但上游高端光芯片（国产化率仅5%）亟待补短板

08 ENDGAME & LEADING INDICATORS

终局推演与产业风险研判——算力产业链演进趋势及未来12至18个月检验替代成败的三大先行指标

- 8.1 算力产业链演进三大深层趋势与结构性财务风险研判
- 8.2 未来12至18个月检验国产替代成败的三个核心先行指标

APPENDIX DATA METHODOLOGY & 320 VERIFIED SOURCES

附录：参考文献与核心信源索引

- 涵盖市场格局与超节点、供应链与客户集采、晶圆制造与AI存储、中美算力实测与算法、全球大厂定制ASIC五大领域共 320 项可溯源原始信源

01 CHAPTER 01 · MARKET LANDSCAPE & INFLECTION
破局与市场洗牌——当200.8万张加速卡完成历史性倒挂，量变在哪里发生？

1.1 市场全局与结构倒挂：从单极垄断迈向66%份额的非对称国产替代

在外部先进制程与高端算力芯片出口管制持续收紧、国内大模型步入万亿参数MoE（混合专家模型）与高并发推理商业化落地的双重驱动下，中国AI算力基础设施产业在2025年至2026年间完成了从“政策性备胎验证”向“规模化商业实交付”的历史性跨越。早期的算力国产化主要依赖政务与运营商信创工程的政策兜底，而当前阶段的核心特征则是互联网大厂与头部大模型企业在真实业务负载下的主动采购与深度耦合。

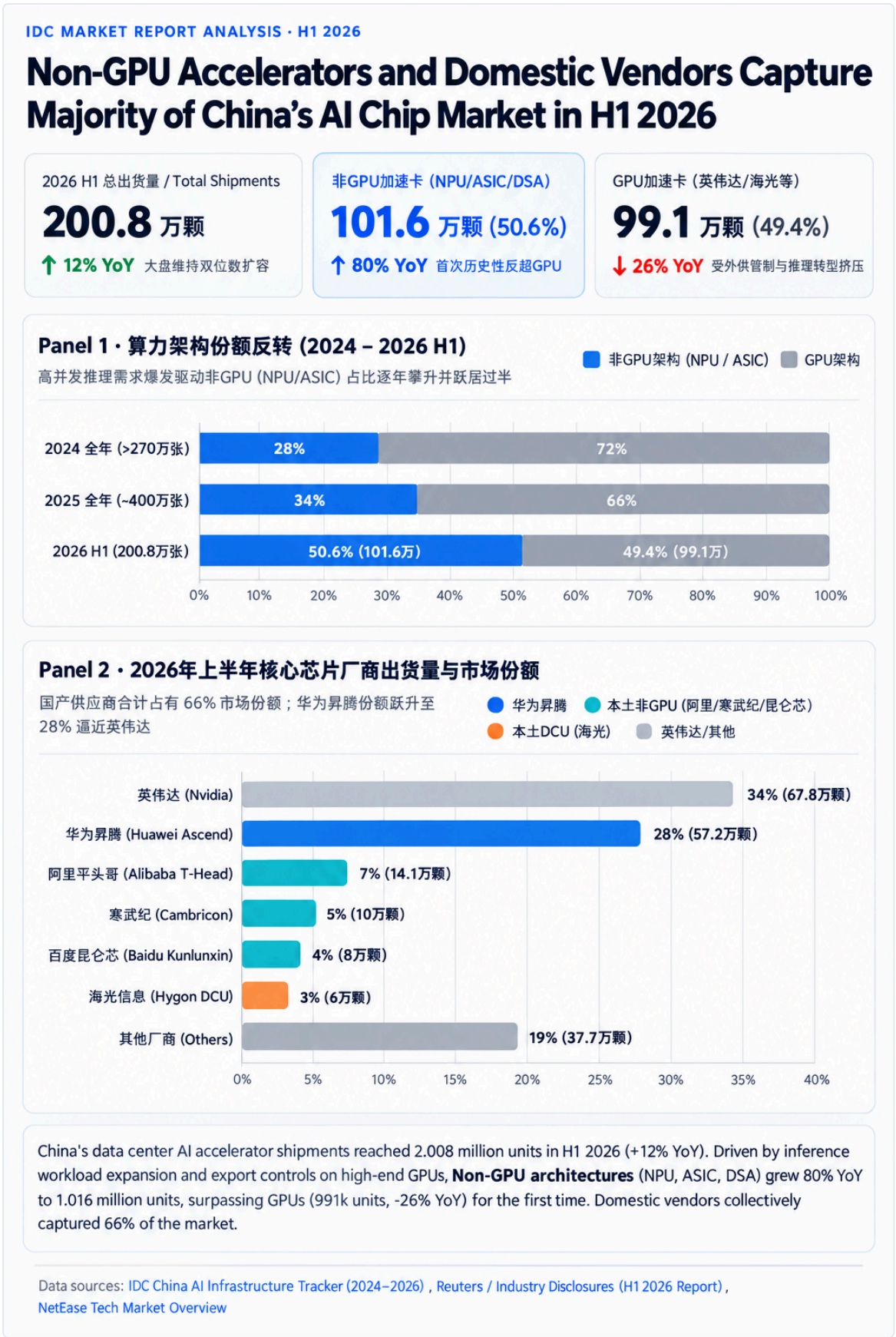
市场份额的剧烈重构直接印证了这一趋势：

- **2024年（单极垄断期）**：中国市场加速卡全年出货量超过270万张。按销售金额口径统计（Bernstein Research数据），英伟达在中国AI加速器市场仍占据约 **66%** 的绝对主导地位（按早期峰值出货口径更曾超过90%），华为海思市占率约为 **23%**，AMD占 **5%**，而寒武纪、摩尔线程、沐曦股份等本土初创与上市企业各自份额仅在 **1%** 左右徘徊。彼时通用GPU占据着72%以上的绝对统治地位，非GPU加速卡份额仅为28%。
- **2025年（份额破局期）**：中国AI加速卡市场总出货量同比增长47%，攀升至约 **400万张**。其中国产AI芯片合计出货约 **165万张**，整体市场份额首次突破四成并达到 **41%**，英伟达的市场份额则回落至 **55%**（全年在华出货约220万张）。从架构看，非GPU占比爬升至34%，而GPU依然握有66%的主导权。其中，仅华为昇腾910C单品在2025年全年出货量便达到约81.2万颗（另有供应链交付口径统计2025年昇腾910B+910C实际确认出货约63万–64万颗），并在2026年第一季度继续交付15万颗，覆盖了国内37家头部大模型企业。
- **2026年上半年（历史性倒挂期）**：2026年10月13日，路透社披露了市场研究机构IDC发布的《中国半年度加速计算追踪报告》，将中国人工智能基础设施领域持续两年的暗流涌动推向台前。2026年1月至6月，中国数据中心AI加速器总出货量达到 **200.8万颗**，在同比增长 **12%**（IDC预计2026全年总出货量将同比增长19%）的稳健大盘下，两组关键指标完成了历史性的结构倒挂：

01 **本土品牌份额反超海外**：受美国商务部工业与安全局（BIS）在2025年4月将中国特供版H20芯片重新纳入严格许可证管控、B200/B300等旗舰卡禁售以及国内对进口合规芯片履行完备安全审批程序的共同影响，英伟达2026年上半年在华AI加速器出货量同比骤降 **44%** 至 **67.8万颗**，市场份额进一步萎缩至 **34%**，其来自中国市场的数据中心营收占比也从三年前的逾20%骤降至约9%（甚至在部分单季财报口径中萎缩至占总数据中心收入不足1%）。与此同时，以华为昇腾（出货 **57.2万颗**，较上年同期的30余万颗几乎翻倍，市场份额跃升至 **28%**，与英伟达差距从

一年前的31个百分点急速收窄至仅 **6个百分点**）为首的本土芯片阵营合力拿下了 **66%** 的市场份额。阿里平头哥（**7%**，约14.1万颗）、寒武纪（**5%**，约10.0万颗）、百度昆仑芯（**4%**，约8.0万颗）、海光信息（**3%**，约6.0万颗）紧随其后；IDC持续追踪的11家中国本土AI芯片厂商中有9家在2026年上半年实现了出货量同比正增长。

02 **非GPU架构首次反超通用GPU**：以NPU、ASIC及DSA（领域专用架构）为代表的非GPU加速卡出货量达到 **101.6万颗**，同比暴增 **80%**，以 **50.6%** 的占比在绝对数量上首次历史性地反超了同比萎缩 **26%** 的传统通用GPU加速卡（**99.1万颗**，占 **49.4%**，IDC预测2026全年传统GPU出货量将进一步萎缩9%）。在101.6万颗非GPU芯片中，华为昇腾NPU、阿里平头哥PPU、寒武纪思元MLU以及百度昆仑芯XPU这四家本土企业合计包揽了高达 **95%** 的市场份额。



■ 图01：2026年上半年中国AI芯片市场非GPU加速卡首次反超与核心厂商份额格局 — 2026H1中国数据中心AI加速卡总出货200.8万颗，非GPU架构（101.6万颗，占50.6%）首次反超通用GPU（99.1万颗，占49.4%），本土品牌合计拿下66%市场份额。



■ 图02：2024–2026H1中国AI加速芯片市场份额演进与技术架构分化对比 — 2024–2026H1中国AI加速芯片市场份额100%结构堆叠演进，以及华为昇腾与英伟达在华份额差距收窄至6个百分点。

这两组数字在资本市场与产业界激起了截然不同的回响。乐观者从中看到了国产算力跨越“可用”分水岭、全面接管中国AI底座的终局信号；而身处大模型训练一线与云厂商机房深处的架构师们则保持着审慎的冷静。穿透宏观出货数据的表象可以发现，2026年中国AI芯片的国产替代并非均匀推进的平面扩张，而是一场高度分化的结构性渗透——其真实深度严格受制于**客户工作负载的容忍度、超节点系统工程对单卡短板的代偿边界**，以及**底层软件与算法对开发摩擦力的消解程度**。

TABLE 01 · 第一章 · 核心审视矩阵 Yavin质观 · 核心数据矩阵

2026H1 中国AI芯片国产替代五大核心审视维度

从市场份额反超、架构反转、负载梯度到超节点与软件栈的全局量化锚点

审视维度	2026年核心量化锚点	替代深度评级	核心驱动与底层约束
市场总量与格局	H1总出货 **200.8万颗 (+12%)** • 英伟达 **34%** (67.8万, **-44%**) • **昇腾28%** (57.2万)、平头哥7%、寒武纪5%、昆仑芯4%、海光3%	份额实质反超 **本土合计 66%** 距NV仅差6个百分点	B200/B300禁售、H20性价比受挫，倒逼本土算力填补真空
硬件架构演变	• **非GPU出货 101.6万颗 (+80%)** • 通用GPU出货 99.1万颗 (-26%) • 昇腾/平头哥/寒武纪/昆仑芯包揽 **95%**	架构重心迁移 **非GPU占 50.6%** 首次反超通用GPU	算力主战场转向高并发推理；DSA/ASIC特定算子能效与成本更优
工作负载梯度	• **成熟区**：常规推理/政企 (**降本75%–95%**) • **攻坚区**：600B MoE长文本推理/RL微调 • **未替代区**：万亿基座从0到1预训练	推理全面接管 **训练局部渗透** V4预训练仍用H800	推理可异步切分；预训练属全局同步，受制木桶效应与单卡故障率
系统级超节点	• 昇腾384超节点商用 **750+套** • 首套 **1024卡昇腾950** 液冷点亮 (**1 EFLOPS**) • **25.6万张** 950DT正部署测试	系统代偿单卡 **物理边界凸显** MoE MFU约45%	光电互联+256TB统一内存破通信墙；但需 **4.5倍卡数 (284亿)** 与 **4倍电费**
软件栈与生态	• 华为CANN开源 **1244万行代码** • Step 5开源24h内4家国产卡 **Day-0适配** • DeepSeek开源 **TileLang-Ascend**	推理生态拉平 **训练范式重构** 达手写算子98%性能	推理接入vLLM/SGLang；训练开启 **“模型主动迁就芯片”** (UE8M0 FP8)

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

1.2 非GPU首次反超GPU背后：算力主战场从预训练军备竞赛转向“推理吞噬一切”

非GPU加速卡（101.6万颗，占50.6%）首次反超通用GPU，表面源于高端通用GPU进口受阻，深层产业动因则是**中国AI算力消耗的主航道已从“百模大战预训练军备竞赛”，彻底转向“高并发推理、垂直微调与传统业务AI重塑”**。

2026年，国内仍在持续投入十万亿参数级从0到1预训练的机构已收敛至 **6到7家前沿AI实验室**；与此同时，企业级AI Agent、百万Token超长文档解析与多模态搜广推重构，使推理调用频次呈指数级累加。Silicon Analysts统计显示，2026年全球数据中心加速器市场规模突破2,000亿美元（较2023年的550亿美元增长逾2.6倍），其中推理工作负载在总资本开支中占比已达 **67%**，在AI系统全生命周期成本中占比更高达 **80%至90%**。

传统通用GPU脱胎于图形渲染，若本土芯片商在缺乏EUV光刻机与顶级HBM条件下强行复制庞大的全功能GPGPU架构（兼顾光栅化渲染管线、FP64科学计算与复杂动态调度器），将在中芯国际N+2（等效7nm级）节点面临严重的晶圆面积浪费与功耗失控。相反，专用AI加速卡（NPU/ASIC/DSA）通过裁撤图形管线与冗余控制逻辑，将晶圆面积集中投向**矩阵乘加单元（如达芬奇Cube核、思元MLUarch）**、**向量计算单元（Vector）**与**大容量片上缓存/显存控制器**，在处理架构收敛的Transformer与MoE推理时，能在受限的7nm/12nm工艺下榨取更高的有效算力密度与能效比，展现出远优于阉割版合规GPU（如H20）的实测性价比。

1.3 自用闭环盘与外部市场盘的真相：左手倒右手的生态闭环与公开市场的残酷搏杀

将上半年逾130万颗国产芯片出货量置于显微镜下审视，市场正清晰裂变为两类生态：**以云大厂“自产自销、内部消化”为核心的自用闭环盘**，以及**依赖政企招投标、运营商集采和商业云租赁的外部市场增量盘**。

TABLE 02 · 第一章 · 厂商格局拆解 Yavin质观 · 核心数据矩阵

中国AI加速卡核心厂商出货、财务与自用/市场盘拆解

2025–2026H1 七大主力阵营：云厂自产自销闭环盘 vs. 公开市场增量盘搏杀

厂商与架构	2026H1份额	2025–2026财务与交付实绩	自用盘 vs. 市场盘	主力代际与核心约束
英伟达 Hopper / H20	**34%** 67.8万颗 同比 -44%	2025在华出货220万张(55%)；中国区占数据中心营收由>20%骤降至**9% (单季<1%)**	**100%外部市场盘** 集中于大厂与头部实验室预训练底座	B200/B300禁售；H20在推理侧遭国产卡价格战挤压
华为昇腾 达芬奇 NPU	**28%** 57.2万颗 同比近翻倍	2025出货63–81万颗；384超节点部署**750+套**；**字节400亿**锁单950PR近半产能	**外部市场盘 >80%** 覆盖互联网大厂、三大运营商及央国企	910C → **950PR/DT**；锁中芯**43%**产能，超节点卡数膨胀4.5倍
阿里平头哥 真武 PPU	**7%** 约14.1万颗 本土第二	2025出货26.5万颗，累计**47万颗**；建成多个**1.6万卡**集群，服务400+客户	**自用与云内 >70%** 支撑淘天/千问及阿里云API算力输出	810E → **M890(144G)** → **V900(216G)**；晶圆受限暂不裸卖板卡

厂商与架构	2026H1份额	2025–2026财务与交付实绩	自用盘 vs. 市场盘	主力代际与核心约束
寒武纪 思元 MLU	**5%** 约10.0万颗 本土第三	2025营收64.97亿(扭亏)；**2026H1营收59.96亿 (+108%)**，在手订单**超80亿元**	**100%外部市场盘** 渗透运营商、金融及互联网第二供应商	**思元590/370** 主力放量；国资协同锁中芯北方 **3万片/月** 7nm产能
百度昆仑芯 自研 XPU	**4%** 约8.0万颗 冲刺科创IPO	2025出货11.6万颗，**3.2万卡** 集群落地；2026年5月启动IPO辅导，年收冲**65亿**	**自用+外拓 5:5** 内部撑文心5.1，外部斩获中移动十亿大单	**昆仑芯P800 + 天池256超节点**；同业竞争顾虑影响部分外拓
海光信息 类CUDA DCU	**3%** 约6.0万颗 盈利规模榜首	2025营收143.8亿；**2026H1营收90.99亿 (+66.5%)、净利17.98亿**；年交付25万片	**100%外部市场盘** CPU+DCU双轮驱动，信创市占 **超50%**	**深算三号** 主力、**深算四号(FP8)** 上市；CUDA算子覆盖 **>99%**
摩尔线程等 独立GPU阵营	**合计19%** 9家本土厂 同比正增长	**摩尔H1营收17.36亿 (+147%)** 亏损缩至1156万；**沐曦** 扣非亏0.49亿；**壁仞** 毛利5.27亿	**100%外部市场盘** 夸娥智算集群与曦云系列多地批量交付	MTT S4000/S5000、曦云C系列；面临晶圆争抢与机房周期错配

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

- **自用闭环盘（“云芯一体”内部消化与独立外拓博弈）**：**阿里平头哥**（2026H1份额7%，截至1月累计出货47万颗）依托3年超3,800亿元AI基建投入，构建了从倚天710 CPU、磐脉网卡、ICN互联芯片、镇岳510 SSD主控到真武AI芯片的“通云哥”全栈底座，既消化自身超2万亿参数模型在线推理，又将国产算力封装为标准化API向外部输出，规避了直接裸卖板卡的软件适配泥潭。**百度昆仑芯**（百度控股57.67%，2026H1份额4%）则在支撑内部文心5.1与搜索重构的同时，于2026年5月启动科创板IPO辅导（传闻估值超400亿美元，2026年营收冲刺65亿元），在中国移动集采多个标段斩获70%–100%份额并打入招商银行、南方电网，但向存在同业竞争的互联网大厂外拓时仍面临股东背景顾虑。
- **外部市场增量盘（运营商百亿兜底与上游产能饥饿游戏）**：华为昇腾、寒武纪、海光信息及摩尔线程、沐曦等在四大买方阵营展开贴身肉搏。三大运营商服务器国产化率已从2020年的40%升至63.7%以上；中国电信15.6万台服务器集采国产G系列占67.5%，而2026年6月启动的 **115亿元、4万台高性能服务器集采国产化率更达100%**。但决定市场盘座次的终极变量已是**上游晶圆与HBM配额争夺战**——2026年国内AI芯片真实需求达 **420万颗**，而中芯国际7nm等效产线全年有效供给仅 **260万–300万颗（刚性缺口40%）**。其中**华为**通过长期协议锁定约 **43%** 产能，**寒武纪**依托北京国资协同拿下中芯北方约 **9%–11%（3万片/月）** 配额（支撑其超80亿元在手订单率先释放利润），其余十余家厂商只能在剩余产能中争抢，超半数中尾部企业被迫转向“12nm + HBM3 + Chiplet”折中路径。

CHAPTER 02 · WORKLOADS & BENCHMARKS

02 客户选择、工作负载与实测对标——真正的替代发生在哪些业务？算力经济学的两本账怎么算？

2.1 工作负载的金字塔替代梯度：推理侧攻城略地，预训练仍是无人区

客户并非按照芯片宣传册上的峰值TFLOPS分配任务，而是根据工作负载对通信延迟、容错率及单Token经济账的敏感度，形成了清晰的**四级替代梯度**（详见表2-1）：

01 **成熟替代区（渗透率60%–90%+，中小模型推理/RAG/搜广推/政企私有化）**：计算图固定、权重可装入单卡。如招商银行引入昆仑芯P800承接智能客服全量流量实现推理成本直降95%，数坤科技基于昇腾完成全人体医疗大模型全离线部署，政务央国企普遍确立“国产芯片+DeepSeek/Qwen开源模型”本地化标准。

02 **新兴扩张层（渗透率40%–60%，具身智能与垂直Agent强化学习）**：随着2026年8–10月开源框架 **RLinf** 完成对昇腾、摩尔线程、壁仞、沐曦与昆仑芯的多节点适配，智元、自变量、华为引望、蚂蚁灵波等团队已在国产集群规模化运行VLA模型微调。

03 **深水突破区（渗透率25%–45%，千亿MoE长文本推理与后训练）**：2026年10月15日阶跃星辰开源600B总参、27B激活、1M上下文旗舰 **Step 5 Preview** 当天，华为昇腾（1024卡Atlas 950超节点）、摩尔线程、沐曦与壁仞在 **24小时内同步完成Day-0推理适配**；DeepSeek最新 **V4-Flash** 后训练与线上推理亦已实质性迁入昇腾集群（注：当日Day-0口径严格限定于推理部署，不含全参数预训练）。

04 **核心未替代区（渗透率<15%，万亿参数基座从0到1预训练）**：除华为盘古（8192卡昇腾910）与百度文心5.1等自研闭环外，市场化头部实验室的旗舰基座预训练（如 **DeepSeek V4主模型**）仍运行在英伟达H800/H100存量集群上，形成**“存量GPU预训练 + 国产NPU/ASIC大规模推理”双轨制**。其物理根源在于：推理可异步切分，而万卡预训练是严格的**全局同步过程**，受制于“木桶效应”，任何单卡互联抖动、精度溢出或硬件故障都会导致数千张卡空转等待。

TABLE 03 · 第二章 · 工作负载替代梯度 Yavin质观 · 核心数据矩阵

AI工作负载四级替代梯度、代表实证与工程瓶颈

客户按通信延迟、容错率与单Token经济账切分算力：推理攻城略地，预训练仍是深水区

工作负载层级	国产渗透率	2026年标杆落地实证	决策因子与工程瓶颈
1. 成熟替代区 中小模型推理 / RAG / 搜广推 / 政企私有化	极高 **60%–90%+**	• **招商银行** 用昆仑芯P800智能客服**推理降本95%** • **中移动** 集采昆仑芯中标70%–100% • **数坤科技** 医疗大模型基于昇腾全离线部署	**驱动**：单Token极致TCO、信创合规 **瓶颈**：存量CUDA代码迁移惰性，少量长尾算子需调优

工作负载层级	国产渗透率	2026年标杆落地实证	决策因子与工程瓶颈
2. 新兴扩张区 具身智能 / 垂直Agent仿真强化学习	中高 **40%–60%**	• 开源强化学习框架 **RLinf** 完成昇腾、摩尔、壁仞、沐曦、昆仑芯多节点适配 • **智元、自变量、华为引望、蚂蚁灵波** 规模化采用	**驱动**：“渲染-训练-推理”异构调度 **瓶颈**：物理仿真器 (Isaac Lab) 具英伟达生态惯性
3. 攻坚突破区 千亿MoE长文本推理 / 后训练 (RL/SFT)	中等 **25%–45%**	• **昇腾950超节点、摩尔、沐曦、壁仞** 24小时内完成600B **Step 5 Preview (1M上下文)** Day-0推理适配 • **DeepSeek V4-Flash** 后训练跑在昇腾集群	**驱动**：大显存装KV Cache、低时延通信 **瓶颈**：Day-0仅限推理，大MoE训练MFU (40%–45%) 折损
4. 核心未替代区 万亿Frontier基座从0到1预训练	极低 **<15%** (仅大厂自研)	• **DeepSeek V4** 主模型预训练仍依赖 **H800/H100** • 海思正部署测试 **25.6万张昇腾950DT**，备战未来6–7家实验室10T–40T基座训练	**驱动**：全局同步极致MFU、万卡MTBF **瓶颈**：需 **4.5倍卡数** 弥补，故障率倍增、电费与机房成本高企

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

2.2 中美头部算力芯片实测工程剖析：训练与推理的剪刀差及三大本土技术流派

当前中国本土算力供给已形成**DSA/NPU专用流派**（昇腾、寒武纪、昆仑芯、紫霄）、**去图形化科学计算增强型GPGPU流派**（平头哥真武PPU、海光深算DCU）与**全功能GPU流派**（摩尔线程MTT系列）三大阵营。



■ 图03：中美主力算力芯片关键工程指标与DeepSeek-R1实测效能对标 — 横向解构板载显存容量（面板一）、显存吞吐带宽（面板二）与300W–400W低功耗区间下的DeepSeek-R1实测吞吐表现（面板三）。



■ 图04：昇腾910C与英伟达H100/B200全维度实测与迁移成本对比 — 昇腾910C在DeepSeek-R1推理中达H100逾80%性能，但千亿参数预训练有效吞吐仅22.2%（落后4.5倍），同时展示了单卡采购、整机部署、CANN迁移人力与千卡故障恢复开销。

TABLE 05 · 第二章 · 中美八大芯片实测 Yavin质观 · 核心数据矩阵

中美八大主力算力芯片规格与 DeepSeek-R1 实测表现

三大本土流派（DSA/NPU、去图形GPGPU、全功能GPU）对标英伟达 H100 / H20 真实工程效能

芯片型号与定位	制程与峰值算力	显存与互联带宽	功耗与参考价	千亿训练与 DeepSeek-R1 实测
NVIDIA H100 北美全能基准	TSMC 4nm **FP16: 989/1979T** FP64: 34 TFLOPS	**80GB HBM3** 带宽 3.35 TB/s NVLink 900 GB/s	700W 单卡 18–40万 整机 180–200万	预训练MFU基准 (100%)；**DeepSeek实测吞吐 14,200 tokens/s，首字延迟 (TTFT) 85ms**
NVIDIA H20 在华合规基准	TSMC 4nm **FP16: ~148 TFLOPS** INT8: 296 TOPS	**96GB HBM3** 带宽 **4.0 TB/s** NVLink 900 GB/s	550W 单卡 >14万元	算力被砍85%致预训练重挫；**实测TTFT 95ms，Batch=64 吞吐落后平头哥PPU约15%**
华为昇腾 910C 国产NPU主力	SMIC 7nm 双Die **FP16: 780–800T** INT8: 1560 TOPS	**96GB / 128GB** 带宽 1.8 TB/s 灵衢UB / HCCS	310W–600W 单卡 10–18万 910B整机20–30万	千亿预训练为H100的 **22.2% (差4.5倍)**；但 **R1推理Prefill达H100的82%，Decode达81% (8,500 tps)**
平头哥 真武 810E 云厂自研PPU	5nm/7nm 2.5D **FP16: 400–600T** 超越A800	**96GB HBM2e** 带宽 **2.76 TB/s** ICN互联 **700 GB/s**	实测 **350W** 阿里云万卡商用	零代码修改迁移；**实测TTFT 82ms (胜H20)，Batch=64 吞吐 2,800 tps (+18%)，能效领先H20超30%**
腾讯 紫霄 2.0 云厂训推NPU	第六代自研架构 **训练提升4倍** **推理提升3倍**	**48GB 显存/卡** 带宽 ~1.2 TB/s 星脉高速网络	~320W 年底部署 **10万卡**	覆盖1B–700B模型训推，**赋能混元大模型实际训练提速50%**，新增自研占比达30%
寒武纪 思元590 第五代训推DSA	7nm Chiplet **FP16: 314/512T** INT8: 512 TOPS	**64GB / 80GB** 带宽 1.6–2.0 TB/s MLU-Link 3.0	350W 单卡 **5.6–8万元** 整机 <100万元	综合性能达A100的70%–90%；**中科院实测能效比 52.3 TFLOPS/W，历史性超越H20 (49.8)**
海光 深算三号 类CUDA DCU	7nm级 x86兼容 **FP64: 30 TFLOPS** FP16: 480 TFLOPS	**64GB HBM2e** 带宽 **1.6 TB/s** 光海HSL总线	**~300W** 单卡 ~10万元 (H100的1/3)	CUDA算子覆盖 **>99%**；**FP64追平H100，千亿训练达A100的80%–90%**；DeepSeek发布当日零感适配
摩尔 MTT S4000 全功能MUSA GPU	第三代MUSA架构 **FP16: 250 TFLOPS** 支持FP64–INT8	**48GB 显存** 带宽 768 GB/s MTLink **240 GB/s**	350W 夸娥(KUAE)集群	**夸娥千卡集群预训练线性加速比 >91%** (30天交付)；通过信通院 **DeepSeek-R1 671B满血推理认证**

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

从图03、图04与表2-2的实测对标中，可提炼出三大核心工程结论：

01 **昇腾910C在DeepSeek-R1推理中实现“以40%标称算力打出81%–82%真实性能”的越级表现**：在7B–13B单卡推理中，910C达H100 100%等效可用度（8卡Atlas 800I整机报价仅20万–30万元，较H100整机降85%）；在671B满血DeepSeek-R1推理中，910C在Prefill阶段达H100的 **82%**、Decode阶段达 **81%**（8,500 tokens/s），核心得益于 **96GB–128GB超大容量HBM**减少了跨机通信，配合CANN针对MLA的算子融合将单步延迟从10ms压至6ms。但在千亿至万亿参数预训练下，受制于卡间通信与约55%的MFU天花板，910C有效训练吞吐回落至H100的 **22.2%**（落后4.5倍，落后B200达5.6倍）。

02 **云巨头自研芯片对英伟达H20形成净替代**：阿里平头哥真武810E凭借 **2,765 GB/s** 显存带宽与 **700 GB/s** ICN片间互联，实测首Token延迟（82ms）与Batch=64高并发吞吐（2,800 tokens/s，较H20高18%）全面反超H20，350W功耗下综合能效比领先超30%。

03 **第三方商用GPU/DCU锚定差异化刚需**：海光深算三号BW1000以 **30 TFLOPS** 双精度FP64算力追平H100，千亿模型训练达A100的80%–90%；寒武纪思元590以不到H20一半的售价（5.6万–8万元）切入高性价比训推市场；摩尔线程MTT S4000在千卡集群上实现>91%预训练线性扩展率。

2.3 算力经济学的两本账：推理单Token的极致降本与预训练TCO的隐性通胀



■ 图05：国产超节点预训练规模惩罚、分布式训练MFU与企业端推理降本三层拆解 — Layer 1展示昇腾950DT超节点对齐英伟达B200峰值算力需承受4.5倍卡数（25.6万张/284亿元）与4.0倍电费（380MW/10亿元/年）惩罚；Layer 2展示Dense/MoE训练MFU（40%–60%）与全国智算中心实际平均利用率（30%）；Layer 3展示企业端推理部署成本直降75%–95%。

- **第一本账（推理侧单Token极致降本账）**：① **硬件替换降本95%**（昆仑芯P800将每百万Token推理成本从H20的数十元压至几元；DeepSeek配合昇腾MindIE三年可省100万–1200万美元TCO）；② **本土推理引擎再砍75%硬件门槛**（清程极智"赤兔"引擎将满血DeepSeek本地部署门槛从4台8卡服务器压缩至单台8卡服务器）；③ **Mooncake PD分离盘活异构资源**（月之暗面联合清华开源的Mooncake承载Kimi线上80%以上流量，以KV Cache为中心解耦Prefill与Decode，掩盖单卡带宽偏科）。
- **第二本账（预训练侧系统TCO与MFU隐性通胀账）**：受制于制程代差，一颗昇腾950DT（报价约11.1万元）单价虽仅为英伟达B200（约4.2万美元/30万元）的三分之一，但**折算FP8峰值算力，1颗B200相当于4至5颗昇腾950DT**。海思披露目前正在部署测试的 **25.6万张Atlas 950 SuperPoD计算卡**（旨在对齐约10万张B200算力，支撑国内6–7家前沿实验室训练10万亿–40万亿参数基座），暴露出三大规模惩罚：① **284亿元初始CapEx**（卡数乘以4.5倍带动光模块、交换机与液冷机柜成倍膨胀）；② **380MW电力黑洞与4倍电费**（年电费高达 **10亿元**，而同算力B200仅需95MW、年电费2.5亿元）；③ **MFU折损与空转损耗**（国产超节点Dense模型MFU约50%–60%，MoE模型MFU仅 **40%–45%**，近半算力在跨节点同步中空转；而全国地方智算中心平均真实利用率更仅有 **30%** 左右）。

03 CHAPTER 03 · SUPERNODE ARCHITECTURE
系统级能力突围——“单卡不够系统来凑”，超节点（SuperPoD）的暴力工程学与物理商业边界

既然单颗芯片短期内无法突破EUV封锁，中国架构师在2026年达成了高度共识：**放弃单卡指标纠缠，用Scale-up柜内超高速互联与Scale-out全光网络将成百上千张芯片捆绑成“逻辑上的一台超级计算机”**。

3.1 暴力工程学：用光电互联与统一内存池击穿MoE通信墙与显存墙

传统8卡服务器跨机通信单跳延迟高达 **2微秒（2 μs）**，多层路由导致30%–40%闲置惩罚。本土超节点（SuperPoD）通过全光互联与统一内存编址实现了系统级破壁：

- **华为昇腾384超节点（CloudMatrix 384）**：通过“灵衢（UnifiedBus 1.0）”与全光背板耦合384张910C（300 PFLOPS算力、192 TB/s带宽），训练效率提升 **45%**、成本降1/3，10万卡集群MFU提升 **2.75倍**，可用度达 **98%**（年无中断超358天），截至2026Q3累计部署 **超750套**。
- **首套1024卡昇腾Atlas 950 SuperPoD点亮（2026.10.15）**：16柜拼成 **1024卡**（**1 EFLOPS FP8 / 2 EFLOPS FP4算力**，最高扩至8,192卡/8 EFLOPS）；依托 **灵衢2.0** 全光Mesh将带宽提15倍、**单跳延迟从2微秒压缩至200纳秒（降10倍）**；构建业界最大 **256TB统一内存空间**（可扩展至1,152TB），点亮当天支撑阶跃星辰600B旗舰 **Step 5 Preview（1M上下文）** Day-0适配，全液冷PUE压至 **1.15**。
- **本土群雄跟进**：**阿里平头哥真武V900**（2027Q1量产，216GB显存、1,200 GB/s互联，性能提3倍，支持50万卡集群）；**百度昆仑芯天池256卡超节点**（有效训练率97%，落地3.2万卡集群）；**天数智芯**（单柜144卡）、**沐曦**（64卡曦景S600）与**燧原**（64卡）同步入局。

3.2 超节点的三大硬约束：机房改造、可靠性反噬与折旧焦虑

01 **亿元级门槛与机房改造**：千卡液冷超节点动辄耗资上亿元（384超节点单套8,000万–9,000万元），传统风冷机房必须重建承重、冷源管网与高压变电（类似海外800VDC演进），故首批买家集中于字节、阿里等巨头，区域智算多妥协于64卡/16卡方案。

02 **可靠性反噬与周期错配**：4.5倍芯片与光模块使故障率成倍增加，迫使华为攻坚“30小时无故障、5分钟恢复、95%检测覆盖”，天数智芯在单柜144卡中预留16张热冗余卡；同时灵衢、ICNLink等私有协议带来3–5年强绑定，而国产芯片“一年一代、性能翻三倍”的迭代速度与地方智算“两年以上”基建周期严重错配，极易陷入“点亮即落后”困局。

CHAPTER 04 · SOFTWARE STACK & ALGORITHMS

04 国产算力软件栈破冰与算法革命——从千万行代码开源、TileLang-Ascend 1:1映射到“模型主动迁就芯片”与MLA显存通缩

如果说硬件出货量决定了国产替代的“广度”，超节点系统工程决定了算力的“上限”，那么软件栈与开发者生态则直接决定了国产替代的“留存率与真实深度”。2026年9月至10月，困扰国产芯片多年的软件生态瓶颈迎来了实质性破冰。

4.1 传统CANN高昂迁移账单与国产软件栈三条进化路径分野

早期国产软件栈（如早期CANN）原生算子库覆盖度仅约为CUDA的 **20%**（约2,000个算子对比CUDA的10,000+个）。对于拥有约50个自定义算子的中等规模大模型训练团队而言，将项目从CUDA迁移至传统CANN，仅20–40个算子重写（10–40人月）、分布式通信调试（4–8周）与性能调优（8–12周）的**单次纯人力成本即高达160万–720万元人民币（22万–100万美元），耗时6至18个月**，很大程度上抵消了硬件采购价格低30%–37%的成本优势。

为撕掉“软件难用”的标签，国内主流算力厂商在2025–2026年演化出**三条互补的软件生态构建路径**：

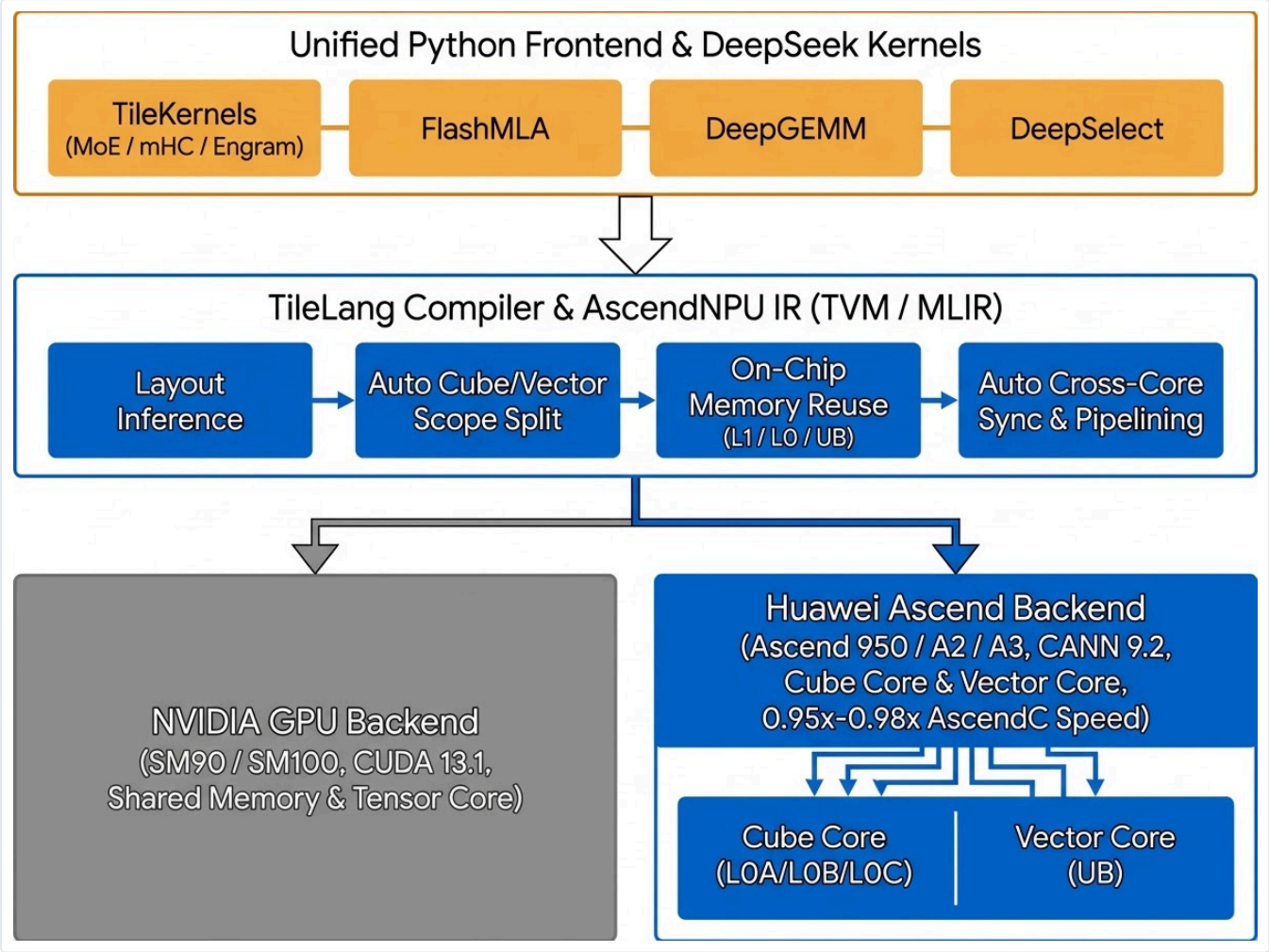
01 **“类CUDA/全功能GPU兼容路线”（海光、摩尔线程、沐曦）**：在指令集或运行时高度对齐CUDA/ROCm。**海光信息**DCU直接兼容ROCm，CUDA算子覆盖率超 **99%**（迁移效率85%–95%）；**沐曦股份**依托MACA软件栈连续完成 **43个主流旗舰大模型的Day-0适配**；**摩尔线程**基于MUSA与MUSIFY工具实现零成本迁移并深度融入SGLang，在阶跃星辰Step 5 Preview开源当天即完成“3层滑窗+1层全注意力”混合稀疏结构的FP8推理与MTP3投机解码适配。

02 **“全栈自研 + 彻底开源开放路线”（华为昇腾）**：CANN与Mind系列基础软件全面开源，已上线 **67个社区项目、累计开源代码超1,244万行**（月活开发者3,500+、深度贡献者6,000+）；开源 **Ascend C** 与四层解耦的 **CATLASS算子模板库**，将Matmul算子开发周期从4人周缩短至2人周（带动30余家核心客户整网性能提升超10%）；开放毕昇编译器 **AscendNPU IR** 接口并推出 **Triton-Ascend** 无感对接Python算子生态；在HCCL中开源 **NB 2.0通信算法**（非对称通信性能提升 **20%+**），并联合社区创立官方维护项目 **vLLM-Ascend** 与打通 **SGLang大EP集群推理方案**，实现主流模型开箱即用。

03 **“云原生封装与异构混训屏蔽路线”（平头哥、昆仑芯、无问芯穹、智源）**：平头哥与昆仑芯将底层驱动与算子调优隐藏在阿里云与百度智能云PaaS层之下；第三方平台 **无问芯穹** 已汇聚调度全国 **3.7万P算力、覆盖16款不同架构芯片**；**智源研究院（众智FlagOS）** 推动的跨品牌异构混训更实现了 **89.5%–92.81%** 的混训效率，与纯英伟达基线的精度差异控制在 **0.62%以内**。

4.2 跨越CUDA护城河的奇点：DeepSeek五大开源算子矩阵、TileLang-Ascend编译破壁与“模型迁就芯片”

2026年9月30日至10月2日，DeepSeek正式向全球开源一整套面向华为昇腾平台的底层训练与基础设施组件（基于昇腾950的128卡超节点联合打磨），且全部与DeepSeek此前面向英伟达CUDA开发的开源版本实现了**接口层面的一一对应（1:1映射）**。



■ 图06： DeepSeek开源算子矩阵与TileLang-Ascend跨架构编译架构图 — Unified Python Frontend & DeepSeek Kernels （TileKernels、FlashMLA、DeepGEMM、DeepSelect） 通过TileLang Compiler & AscendNPU IR四大编译遍，实现NVIDIA GPU Backend与Huawei Ascend Backend（Cube核+Vector核）的零重写1:1映射，达手写AscendC 95%–98%性能。

TABLE 06 · 第四章 · 跨越CUDA护城河 Yavin质观 · 核心数据矩阵

DeepSeek 五大核心开源算子库与昇腾 1:1 映射加速表现

仅百行代码榨干硬件极限，联合华为昇腾128卡超节点实现CUDA向NPU零重写迁移

开源算子项目	核心解决瓶颈	底层编译与微架构创新	实测加速比与昇腾适配
DeepGEMM & Ascend版	稠密与MoE分组矩阵乘(GEMM) 及分页Logit计算	仅 **300行核心代码**；完全即时编译 (JIT) 省寄存器；重叠TMA搬运与Tensor/Cube计算	小批量实时推理 **加速 2.7倍**，MoE分组乘 **加速 1.2倍**；已开源独立C++库 **DeepGEMM-Ascend**

开源算子项目	核心解决瓶颈	底层编译与微架构创新	实测加速比与昇腾适配
DeepEP & Ascend版	MoE跨节点/节点内专家并行 (EP) All-to-All通信	针对NVLink/HCCS与RDMA/RoCE非对称域宽，定制高吞吐令牌分发 (Dispatch) 与聚合原语	彻底打通MoE预训练与解码跨卡阻塞，**支撑数十万卡级分布式专家路由**，通信计算无感重叠
FlashMLA	多头潜在注意力 (MLA) 解码与稀疏注意力内核	动态智能调配计算资源；引入 **Split-KV机制** 并行计算再合并，配合共享内存交错消除冲突	逼近物理显存带宽极限；在 **TileLang中仅需80行Python代码** 复现，已支持 **昇腾与MI300X**
TileKernels (v2.1.0)	生产级大模型训推融合算子与 **V4流形超连接 (mHC)**	基于TileLang涵盖MoE路由、Engram门控、Sinkhorn归一化；自动完成 **昇腾Cube/Vector分核与双缓冲**	2026.09合入 **昇腾950/A2/A3后端**；同一套Python API横跨NV与昇腾，**达手写算子 96%–98% 性能**
DeepSelect	稀疏注意力 (DSA) 动态Token筛选与Top-K采样	针对百万长上下文动态Token筛选，深度优化底层显存访问模式与并行归约路径	在稀疏注意力Top-K基准测试中，访存优化直接带来 **1.9倍性能跃升**；Sparse FA耗时 **347μs → 127μs**

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

TileLang-Ascend 编译破壁与“模型主动迁就芯片”的范式逆转

不同于英伟达GPU的SIMT统一流多处理器，昇腾达芬奇架构采用物理分离的 **Cube矩阵核** 与 **Vector向量核**，直接移植代码极易产生流水线气泡。`TileLang-Ascend` 在TVM/MLIR之上引入“开发者模式”，通过四大核心编译遍（**自动剥离Cube/Vector作用域并插入跨核硬件同步、片上L1/L0/UB存储层级自动映射与L1常驻策略、MTE2/Compute/MTE3核内与C/V核间双缓冲流水线并行、以及标量循环自动SIMD向量化**），让开发者仅需编写统一Python/GPU风格代码即可直达昇腾硬件极限。实测显示，在昇腾A2/A3及Ascend 950上，TileLang编写的 **GEMM算子达手写AscendC专家代码的98%，流形超连接（mHC）达96%，FlashAttention达95%–100%，Sparse Flash Attention执行耗时从347微秒大幅压缩至127微秒**。

更深远的变化在于**“模型主动迁就芯片”**的软硬协同范式确立：**DeepSeek V3.1** 专门引入 **UE8M0 FP8数据格式** 以契合下一代国产AI芯片微架构；**阶跃星辰 Step 5 Preview** 则以27B极小激活参数量与8-Token Region稀疏注意力精准匹配国产超节点与大显存GPU。当顶尖基座模型实验室亲自下场用TileLang重写底层算子并调整模型数值格式适配国产硬件时，CUDA长达十年的生态铁幕已被撕开结构性裂口。

4.3 MLA与TransMLA / DoseRank / MLAfication后训练无损迁移：算法效率十倍跃升对远期HBM与高端GPU弹性的重塑

如果说 `TileLang-Ascend` 拆除了跨芯片编译藩篱，那么以 **MLA（多头潜在注意力）** 为核心的注意力革命，以及2025–2026年成熟的 **TransMLA、DoseRank 与 MLAfication** 后训练无损迁移技术，则从算法源头引爆了数据中心对HBM容量与高端GPU采购弹性的需求通缩。

1. MLA 吸收操作（Absorb）与三代后训练无损转换框架

传统GQA架构中，长序列KV Cache对显存带宽的吞噬是制约推理吞吐的头号瓶颈。MLA通过低秩潜在向量压缩键值特征，并在推理阶段利用矩阵乘法结合律实施**“吸收操作（Absorb Operation）”**——将上投影矩阵直接合并至查询与输出投影矩阵中，使查询头直接与微小共享潜在向量交互。在典型配置下，**MLA吸收路径的计算访存比高达约 295 FLOPs/byte，逼近H100 BF16屋脊平衡点（约 242 FLOPs/byte）**，将长文本解码从“显存带宽受限”扭转为“计算受限”。针对存量GQA模型受RoPE与QKNorm阻断无法直接转为MLA的难题，三代后训练框架完成了接力突破：

- **TransMLA（NeurIPS 2025 Spotlight）**：通过RoRoPE正交旋转、FreqFold频率折叠与Balance-KV范数平衡，将旋转位置能量集中至首个头从而解耦Absorb操作。在**免训练**下即可砍掉LLaMA-2 **68.75%**、Qwen2.5-7B **71.875%** 的KV Cache；配合轻量微调在LLaMA-2-7B上实现 **93% KV Cache压缩** 与 **10.6倍推理加速**（吞吐从48.5飙升至 **4,151.17 tokens/s**）。
- **DoseRank（ICLR 2027审稿）**：提出跨层动态秩分配，在 `gpt-oss-120B` 上砍掉 **62.5% KV Cache** 的同时，校准成本较TransMLA降低 **177至462倍**。
- **MLAfication（2026年10月复旦大学开源）**：提出校准型KNorm攻克Qwen3等新一代模型的QKNorm障碍，仅训练1.6%–11.1%参数抑制MoE路由漂移，在 `Qwen3-30B-A3B` 上实现 **5.35倍** 服务吞吐量，叠加4-bit量化后 **将KV Cache显存惊人砍掉94.73%！**

2. 财务报表印证与对国产AI芯片的二阶红利

算法效率的数量级跃升已直接兑现于财务报表：**DeepSeek在维持全行业最低梯队API定价的同时，实现了高达82.9%的API毛利率**（年化营收突破10亿美元），远超依赖传统算力堆叠的OpenAI（2026Q1毛利率 **39%**）与Anthropic（2026全年预期 **63%**）。对国产芯片而言，当存储芯片在2026年吞噬全球云厂约 **30%** 资本开支（2023/2024年仅8%）时，后训练MLA将KV Cache无损砍掉62.5%–94.73%，使得**原本需要8张192GB B200才能装下的高并发长文本推理批次，现在仅用1–2张64GB–128GB的国产加速卡（如96GB平头哥真武810E、128GB昇腾910C、48GB腾讯紫霄2.0或摩尔线程S4000）即可轻松承载**，从算法侧抹平了国产芯片在极限HBM容量与带宽上的物理短板。



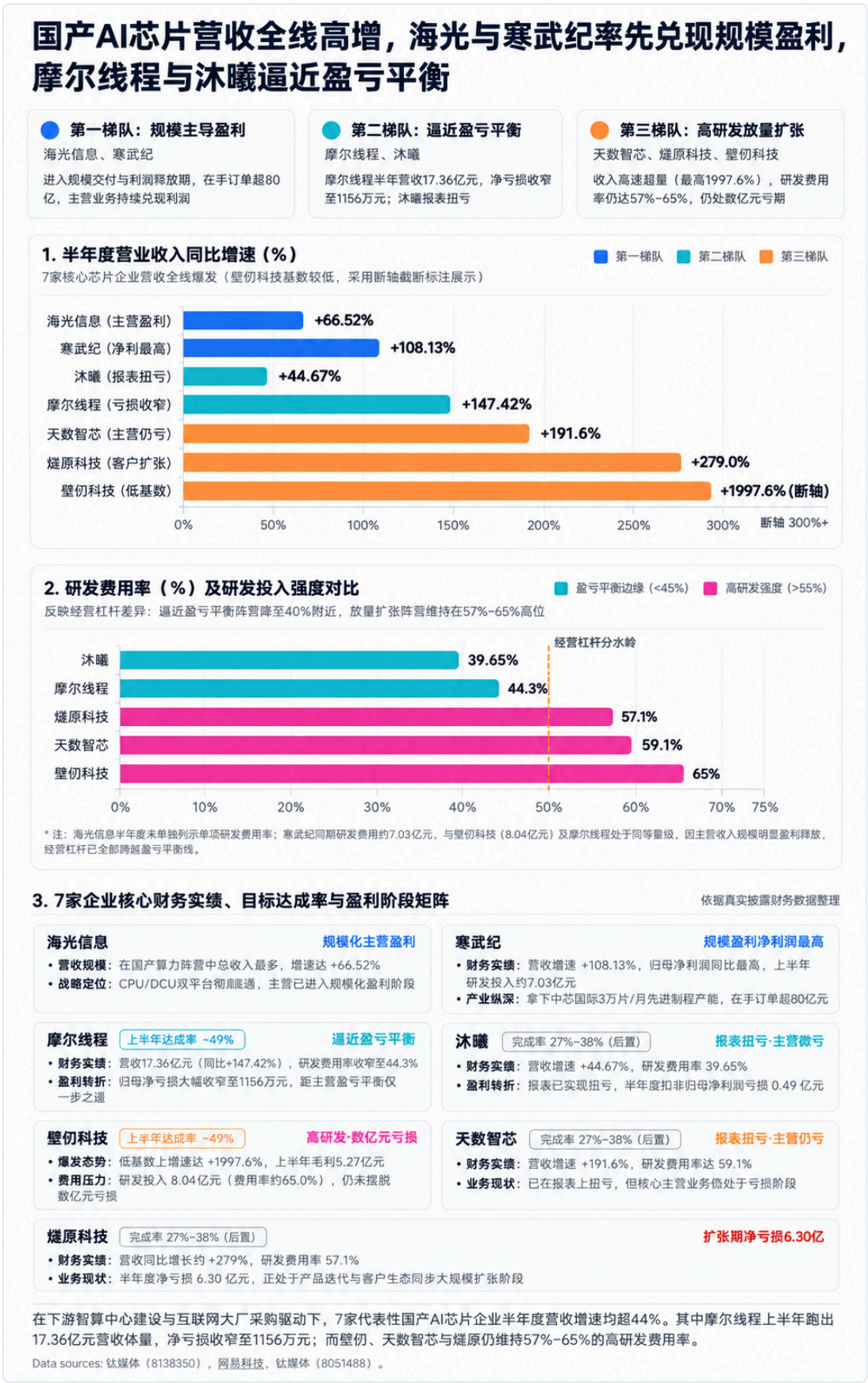
■ 图07：后训练MLA转换对KV Cache的指数级削减与大模型API毛利率分化 — 面板1展示TransMLA、DoseRank与MLAfication将存量GQA模型的KV Cache显存削减62.5%至94.73%并实现5.35x–10.6x吞吐跃升；面板2展示DeepSeek API推理毛利率高达82.9%，远超Anthropic（63%）与OpenAI（39%）。

05 CHAPTER 05 · DOMESTIC CHIP CONTENDERS
国产GPU与ASIC四大阵营分化——财务兑现、产品矩阵与中国科技巨头自研芯片全景

除华为海思以断层优势领跑外，中国本土AI算力芯片行业在2025至2026年经历了深刻的商业化洗牌，正式告别单纯依赖融资输血的早期阶段，分化出**云厂自研ASIC/GPGPU、规模化主营盈利龙头、逼近盈亏平衡独角兽以及高研发攻坚企业**四个层次分明的梯队格局。

5.1 本土算力芯片四梯队财务兑现、研发强度与盈利收敛实绩

2026年上半年，7家代表性本土算力芯片企业迎来营收集体爆发，并依据研发费用率（45%分水岭）分化为三大商业化梯队：**第一梯队（海光信息、寒武纪）**已率先兑现规模化主营盈利，海光2026H1营收达90.99亿元（+66.52%）、归母净利17.98亿元，寒武纪2026H1营收暴增108.13%至59.96亿元、在手订单突破80亿元；**第二梯队（摩尔线程、沐曦股份）**研发费用率降至39.65%–44.3%，半年度亏损收窄至千万元级并实质性触及盈亏平衡线；**第三梯队（壁仞科技、天数智芯、燧原科技）**在实现191%–1,997%营收高增长的同时，维持57%–65%的高研发强度攻坚下一代旗舰架构。详细财务数据与主力型号对比见下图08与表07：



■ 图08：国产AI芯片营收全线高增，海光与寒武纪率先兑现规模盈利，摩尔线程与沐曦逼近盈亏平衡 — 7家代表性国产AI芯片企业半年度营业收入同比增速（%）、研发费用率（%）及研发投入强度对比，以及核心财务实绩与盈利阶段矩阵。

TABLE 07 · 第五章 · 本土算力主力矩阵 Yavin质观 · 核心数据矩阵

国产 GPU / ASIC 四大阵营主力型号与商业化兑现对比

海光与寒武纪率先跨过盈亏平衡大规模释利，摩尔线程与沐曦逼近盈亏平衡线

厂商阵营	主力型号与制程	关键性能与生态兼容度	核心客户群与2026财务兑现
海光信息 688041.SH	**深算二号 / 三号** **深算四号 (SG2040)** 7nm+ Chiplet / 64G	· **FP64 30 TFLOPS、FP16 480 TFLOPS** · 带宽 1.6–1.8 TB/s，兼容ROCm · **CUDA算子兼容 >99%**，四号支持FP8	· 信创市占 **>50%**，智算中标率 **45%** · **2026H1营收 90.99亿 (+66.5%)** · **H1归母净利 17.98亿**，居本土之首
寒武纪 688256.SH	**思元 590 (MLU590)** 思元 370 / 270 7nm MLUarch05	· 综合实测达 **A100的 70%–90%** · 实测能效比 **52.3 TFLOPS/W 超越H20** · 全面适配DeepSeek/Qwen训推一体	· 单卡售价 5.6–8万元，获字节/阿里/运营商采购 · **2026H1营收 59.96亿 (+108%)** · **在手订单突破 80亿元**
摩尔线程 全功能GPU	**MTT S4000 / S5000** 夸娥 (KUAE) 集群 第三代MUSA架构	· 48GB显存 (S5000升级HBM)，768 GB/s · MTLink 240 GB/s，MUSIFY零成本迁CUDA · **千卡训练线性加速比 >91%**	· 兼顾大模型千卡训推与图形渲染 · **2026H1营收 17.36亿 (+147%，超去年全年)** · **净亏损大幅收窄至 1,156万元**
互联网自研 云芯一体	**平头哥真武 810E/V900** **百度昆仑芯 P800** **腾讯紫霄 2.0**	· 深度耦合自有云底座与大模型框架 · 平头哥/寒武纪/昇腾正与 **长鑫存储开展HBM3E系统级联调**	· **平头哥市占 7%** (累计出货47万颗) · **昆仑芯市占 4%** (冲刺IPO，年收冲65亿) · **紫霄2.0** 年底部署达 **10万卡**

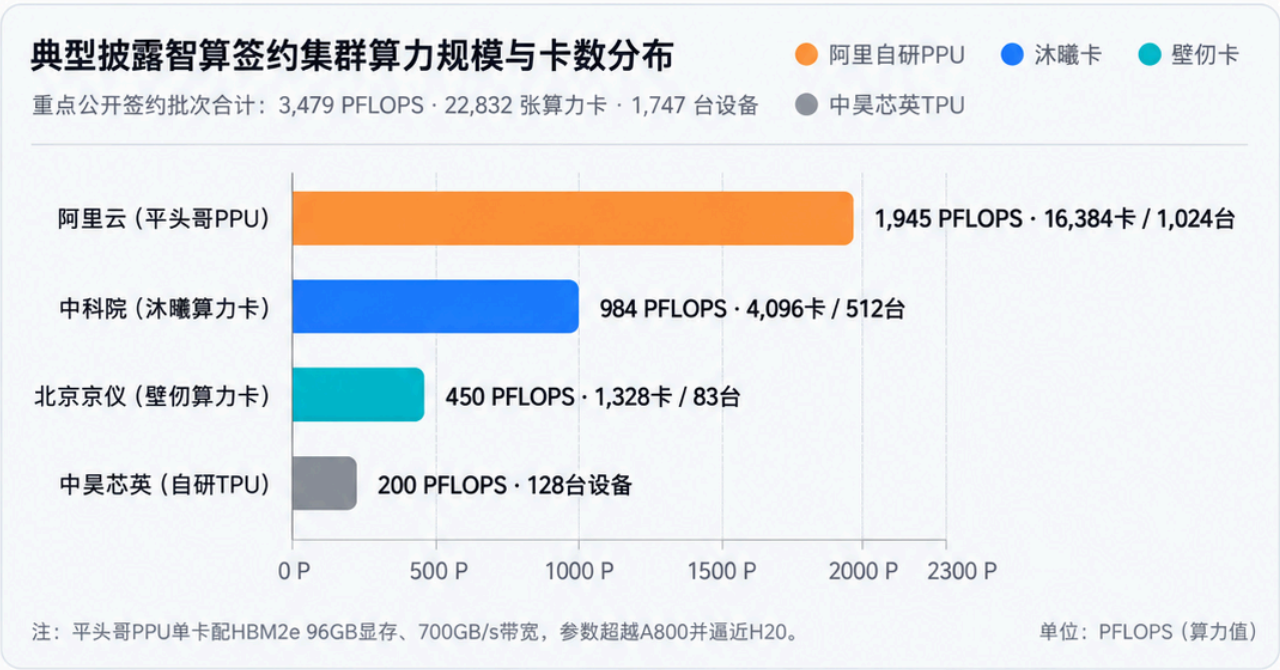
资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

5.2 中国四大互联网科技大厂自研ASIC/GPGPU布局：阿里平头哥、腾讯紫霄、百度昆仑芯与字节跳动

与北美大厂主要出于TCO与能源经济账自研ASIC不同，中国科技巨头（阿里巴巴、腾讯、百度、字节跳动）的自研芯片落地叠加了应对地缘供应链限制与保障算力底座生存安全的刚性诉求。据Bernstein数据，**中国国内AI芯片市场需求规模已跨入395亿美元量级，本土化率正从2023年的17%加速跃升至2027年预计的55%。**

算力芯片国产化落地实证 2025–2027 战略周期

国产算力替代实质提速：阿里平头哥PPU万卡集群落地，2027年本土化率预计达55%



宏观市场演进（Bernstein） +38 pct 增幅

中国AI芯片市场本土化率预期跃升

2023年（替代初期基准） 17%

2027E（预测本土化目标） 55%

据分析机构研报预测，2025年国内AI芯片总需求将攀升至395亿美元，本土替代率4年内预计跃升逾3倍。

互联网大厂自研落地 混元业务实证

腾讯云紫霄2.0芯片 2026落地里程碑

2026年新增自研占比 **30%** 新增算力自给率目标

2026年底部署目标 **10万卡** 自研集群规模量级

实测业务效能：已率先承载混元大模型训练，训练性能较上代提升4倍，推理吞吐提升3倍，实际训练提效达50%。

在公开披露的合计22,832张算力卡、总算力3,479 PFLOPS的重点签约智算项目中，阿里云自研平头哥PPU芯片单家即部署1,024台服务器、16,384张算力卡（提供1,945 PFLOPS算力），成为国内出货量最高的自研AI芯片之一。与此同时，腾讯云紫霄2.0在2026年将新增自研芯片占比提升至30%并冲击年底10万卡规模，共同推动中国AI芯片本土化率向2027年的55%迈进。

数据来源：央视《新闻联播》 / Bernstein研报，腾讯云 WAIC 2026 官方发布，腾讯云开发者社区，36氪行业报道

■ 图09：国产算力替代实质提速——阿里平头哥PPU万卡集群落地与腾讯紫霄2.0里程碑 — 公开披露的重点智算签约集群算力规模与卡数分布（阿里云平头哥PPU 16,384卡/1,945 PFLOPS、中科院沐曦4,096卡/984 PFLOPS、北京京仪壁仞1,328卡/450 PFLOPS、中昊芯英TPU 128台/200 PFLOPS），以及2027年本土化率55%目标与腾讯云紫霄2.0年底10万卡部署里程碑。

TABLE 08 · 第五章 · 科技巨头自研全景 Yavin质观 · 核心数据矩阵

中国四大互联网科技巨头自研 AI 芯片方案与落地实况

中国AI芯片需求跨入395亿美元量级，本土化率从2023年17%加速跃升至2027年55%

国内巨头	自研芯片演进序列	核心硬件规格与生态模式	实际落地规模与业务赋能 (截至2026.10)
阿里巴巴 平头哥 T-Head	含光800 → **真武810E** → **真武M890** → **真武V900 (2027Q1)** 配倚天CPU+磐脉网卡	• **810E**：96G HBM2e，**2.76 TB/s**，700 GB/s互联，实测350W • **M890**：144G显存；**V900**：**3倍性能、216G显存、1.2 TB/s互联**	• 部署多个万卡集群（单批签约**16,384卡 / 1,945 PFLOPS**） • 累计出货 **47万颗**，占阿里采购 **30%–40%+** • 服务国家电网、小鹏等 **400+外部客户**
腾讯 蓬莱实验室	紫霄v1 → **紫霄2.0 (C200 NPU)** → 紫霄v3 (试产) 携燧原(持股20.5%)协同	• **紫霄2.0**：训练性能较上代**提升4倍**，推理吞吐 **提升3倍** • 单卡48GB显存，配合星脉网络与动态分桶调度	• **赋能混元大模型实际训练提速50%** • 支撑混元Hy3首周调用暴涨 **68倍** • 占新增芯片 **30%**，年底总部署达 **10万卡**
百度 昆仑芯科技	昆仑芯1/2代 → **3代P800 (HBM)** → X3代际 **天池256卡超节点**	• 自研XPU架构，**P800升级64G/96G HBM** (1.6–2.0 TB/s) 与原生FP8 • 天池256超节点 **有效训练率97%**	• 内部落地 **3.2万卡集群** 撑搜索与文心5.1 • 外部斩获 **中移动集采70%–100%标段 (十亿级)**，招行推理降本95%，2026营收冲 **65亿元**
字节跳动 自研+百亿外采	**两款定制 5nm ASIC** (训/推分离) + 百亿外采国产卡	• 联合 **博通 (Broadcom)** 基于台积电5nm定制ASIC，算法固化极致降本 • 应对豆包高频调用与 **700亿美元** 基建规划	• 一方面预订数十万枚台积电5nm自研ASIC • 另一方面 **2026年向华为下达超400亿元 (56亿美元) 昇腾950PR创纪录大单**

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

5.3 核心本土厂商替代路径全景对比矩阵

基于需求结构、工作负载、系统工程与软件生态的四层穿透分析，当前主导中国数据中心AI加速卡市场的五大本土核心厂商已分化出清晰的生态位与商业边界：

TABLE 09 · 第五章 · 本土五强路径对比 Yavin质观 · 核心数据矩阵

中国五大本土核心 AI 芯片厂商替代路径全景对比矩阵

昇腾、平头哥、昆仑芯、寒武纪、海光信息在超节点、软件栈与商业模式上的生态位分化

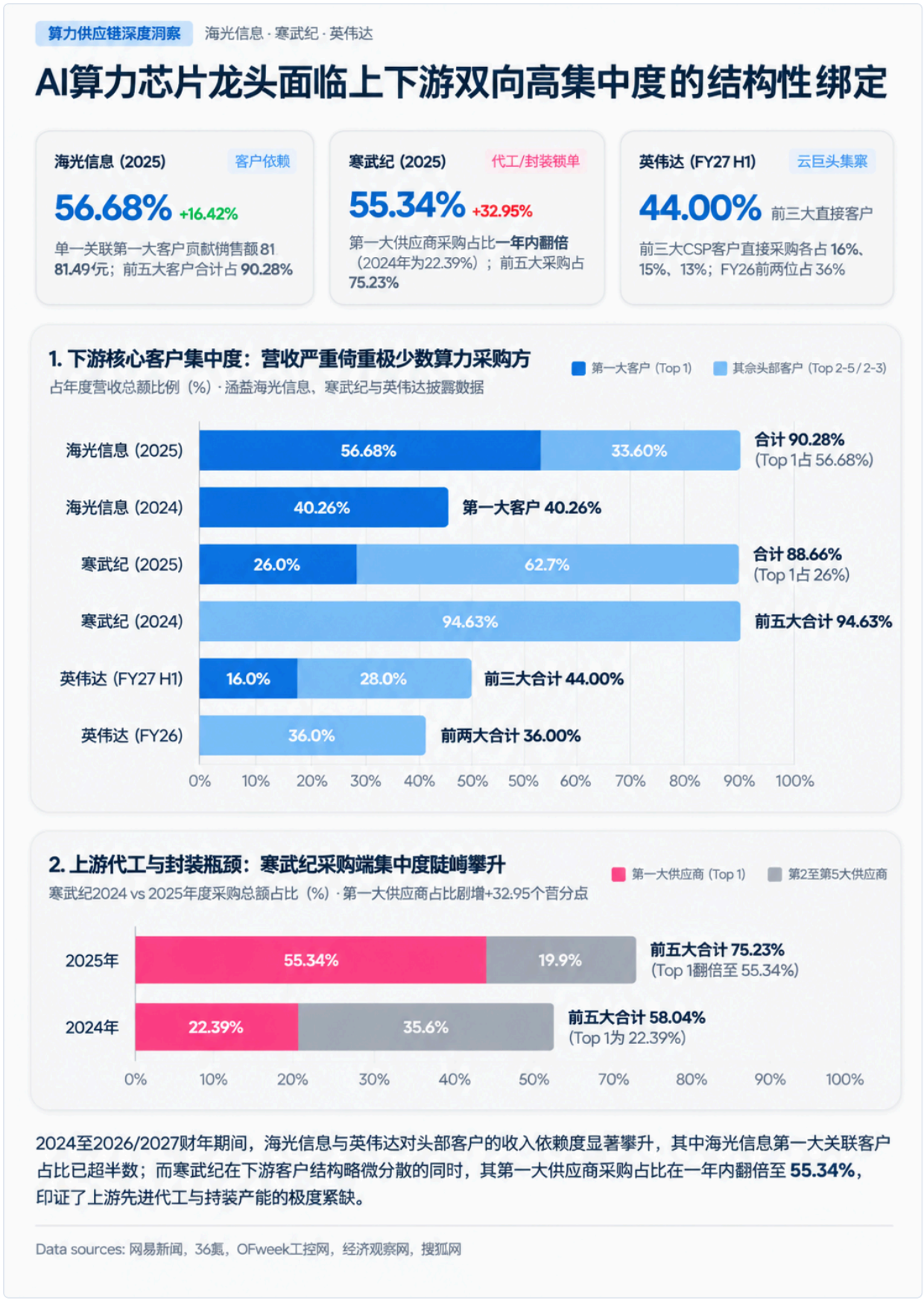
评估维度	华为昇腾 (28%)	阿里平头哥 (7%)	百度昆仑芯 (4%)	寒武纪 (5%)	海光信息 (3%)
主力芯片与迭代	910B/910C主力；**950PR/DT (FP8，11.1万)** Q4上市	810E / **M890(144G)** 主力；**V900(216G)** 27Q1量产	**昆仑芯P800** 规模放量；X3代际推进中	**思元590**、思元370规模量产交付	**深算三号(64G)** 主力；**深算四号 (+40%，FP8)** 上市
超节点与集群	**行业断层领先**：384商用750+套；**1024卡950液冷点亮**	**云原生超节点**：M890上线，V900支持 **50万卡** 扩展	落地 **3.2万卡** 集群；上市 **天池256超节点** (训练率97%)	高密8卡及千/万卡RoCE集群，支持MLU-Link 16卡节点	依托x86 C86 CPU+DCU异构协同，信创中标率 **45%**
软件栈与生态	CANN开源 **1244万行**；获 **DeepSeek TileLang** 1:1支持	耦合阿里云PaaS与通义生态，对租户屏蔽硬件复杂性	耦合飞桨 (**一行代码切换**)；完成RLinf具身智能适配	NeuWare软件栈全面适配DeepSeek/Qwen训推	**类CUDA (ROCm) 算子覆盖 >99%**，几乎零修改迁移
核心落地负载	**全场景**：政企推理、V4-Flash后训练，冲刺10T预训练	淘天搜广推、2万亿模型推理、阿里云外部API租赁	百度搜索与文心5.1训练；运营商与银行高并发推理	互联网推理分流、运营商及省级智算中心底座	党政信创、金融电信私有化训推与 **科学计算 (FP64)**
自用 vs 外销	**外销 >80%**：绑定运营商、央国企与字节400亿大单	**云内消化 >70%**：以阿里云服务输出，暂不裸卖板卡	**内生外拓 5:5**：百度托底+中移动大单，冲刺独立IPO	**100%纯外销**：H1营收59.96亿，在手订单 **超80亿**	**100%纯外销**：H1营收90.99亿、**净利17.98亿 (第一)**
核心演进瓶颈	同等算力需 **4.5倍卡数与4倍电费**；MoE MFU待提升	晶圆产能受限致供不应求；外部机房无法单采板卡	百度控股57.67%带来同业竞争顾虑，影响同行渗透	晶圆配额(9%–11%)少于华为；地市级下沉市场待拓	单卡峰值与超节点互联密度较昇腾950略显保守

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

06 CHAPTER 06 · SUPPLY CHAIN & CUSTOMER ORDERS
全球与中国AI算力芯片上下游供应链及客户生态深度剖析

6.1 产业供应链重构与“上下游集中度双向极化”特征

当前全球及中国AI算力产业链正呈现显著的**“上下游双向极化”**特征：在上游制造与核心元器件端，受制于先进制程代工、2.5D/3D先进封装（CoWoS、EMIB）及HBM的物理产能瓶颈，采购份额急剧向极少数底座供应商聚拢；在下游应用端，由于万卡集群CapEx门槛呈指数级攀升，订单高度集中于超大规模云厂商、互联网巨头、电信运营商及头部服务器整机代工商（OEM/ODM）。



■ 图10：AI算力芯片龙头面临上下游双向高集中度的结构性绑定 — 海光信息（2025第一大关联客户占56.68%，前五大占90.28%）、寒武纪（2025第一大供应商采购占比翻倍至55.34%，前五大客户占88.66%）与英伟达（FY27 H1前三大直接客户合计占44.00%）的上下游集中度实证。

三大代表性龙头的最新财报数据深刻印证了这一双向绑定格局：

01 **寒武纪（上游代工封装集中度陡峭攀升）**：2025年下游前五大客户营收占比微降至 **88.66%**（贡献营收57.60亿元，第一大客户占比超26%），但**上游前五大供应商采购占比从2024年的58.04%飙升至75.23%，其中第一大供应商采购占比从22.39%暴增至55.34%（+32.95个百分点）**，反映出国内高良率先进制造与2.5D封装底座的极度稀缺。

02 **海光信息（下游关联整机大客户占比超半数）**：2025年前五大客户销售占比达 **90.28%**，**对第一大关联方客户（中科曙光等关联整机体系）销售占比从2024年的40.26%跃升至56.68%（对应销售额81.49亿元）**，期末第一大客户应收账款占比亦超 **50%**。

03 **英伟达（全球云巨头客户集中度空前攀升）**：FY23尚无单一客户占比超10%，FY26前两大客户合计占比已达 **36%**，而在 **FY27 H1前三大直接客户分别贡献16%、15%和13%收入，合计占比高达44%**（配套光模块龙头2025年前五大客户占比亦高达 **78.7%**）。

TABLE 10 · 第六章 · 全球与中国供应链 Yavin质观 · 核心数据矩阵

全球与中国核心 AI 芯片企业上下游供应链及客户对照

产业链呈"上下游双向极化"：上游晶圆/CoWoS/HBM高度垄断，下游百亿与吉瓦级大单集中

核心算力企业	上游晶圆代工与先进封装	上游存储、载板与高速互连	下游整机/终端客户与集中度硬指标
华为昇腾 910B/910C/950	**中芯国际** (7nm DUV) **盛合晶微** (2.5D占其营收74.4%)、**长电科技**	**长鑫科技** (HBM2/3)、**兴森科技** (ABF占>60%)、**深南/沪电**、**华丰科技** (224G连接器占60.5%)	• **整机**：超聚变(昆仑)、华鲲振宇、长江计算 • **客户**：**字节豪掷400亿锁950PR近半产能**；中移动集采CANN生态占 **66.1% (33.8亿)**
海光信息 C86 + 深算DCU	**中芯国际** (晶圆底座) 盛合晶微、长电科技	高集中度IP授权、EDA及本土高阶基板配套	• **整机**：中科曙光、浪潮(独中工行30亿)、新华三 • **集中度**：前五大客户占 **90.28%**，**第一大关联客户占 56.68% (81.49亿元)**
寒武纪 思元590/370	**中芯北方** (锁3万片/月 7nm+) 盛合晶微、长电科技	核心HBM存储与高阶基板供应商高度集中	• **整机**：浪潮(元脑)、联想、曙光、科大讯飞 • **集中度**：前五大客户占 **88.66% (57.6亿)**；**第一大供应商采购占比飙升至55.34%**
平头哥 & 昆仑芯 真武 / P800	5nm/7nm先进代工 2.5D/3D先进封装	96G–216G HBM2e/3、自研ICN互联 (**正联调长鑫HBM3E**)	• **平头哥**：累计出货 **47万颗**，占阿里采购30%–40%+ • **昆仑芯**：中移动集采斩获 **70%–100%标段**，获比亚迪与国新基金战投
英伟达 Blackwell / Rubin	**台积电** 4nm/3nm CoWoS-S / CoWoS-L	SK海力士/美光/三星 HBM3e/4、NVLink Fusion、**安费诺** 连接器	• **ODM/客户**：富士康、广达；微软/Meta/谷歌/AWS • **集中度**：**FY27 H1前三大直接客户分别占 16%、15%、13% (合计高达44%)**

核心算力企业	上游晶圆代工与先进封装	上游存储、载板与高速互连	下游整机/终端客户与集中度硬指标
博通 & 谷歌 TPU v7 Ironwood / v8	**设计**：博通(70%–80%)、联发科 **封装**：台积电、Amkor、日月光、**Intel EMIB-T**	**HBM**：海力士/三星；**载板/PCB**：欣兴/金像电/台光；**光模块**：**中际旭创**、**新易盛**、Lumentum	· **吉瓦级长单**：谷歌续签至2031 (**2026贡献210亿美元**)；向 **Anthropic供应3.5GW TPU**；与 **Meta签 >1GW** 定制XPU至2029年

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

6.2 华为昇腾：全栈国产化供应链垂直整合与互联网、运营商双轮驱动

华为昇腾及Atlas超节点系统已构建起国内最完整的本土硬件供应链。**在昇腾AI服务器及超节点硬件成本中，CPU与存储（特别是HBM）合计占比超70%，高速背板连接器等互连硬件占10%–15%**。

1. 上游核心元器件与制造端的本土深度绑定

- **ABF载板与高阶PCB**：**兴森科技**（国内唯一规模量产3微米线宽ABF载板、良率>92%）在昇腾供应链中份额超 **60%**；**深南电路**承担ABF载板、FCBGA基板与高阶PCB核心供应；**沪电股份**为主力高阶服务器PCB供应商。
- **超节点224G高速连接器与线模组**：华为哈勃入股的 **华丰科技** 为昇腾224G高速背板连接器及线模组主力核心供应商（**对华为销售占比达60.52%，对应金额近15.3亿元，同比翻倍不止**），其面向下一代超节点的 **224G高速线模组已于2026年6月正式量产，良率达85%**；**中航光电** 亦提供高速线缆与连接器配套。
- **先进封装与国产HBM突破**：**盛合晶微**（华为占其营收高达 **74.4%**）与 **长电科技** 构成核心封装底座；于 **2026年7月27日刷新科创板大型IPO最快纪录（165天过会）** 挂牌上市的 **长鑫科技（长鑫存储）** 已实现HBM2量产及12层HBM3试产，将IPO募集的 **近600亿元资金** 投向HBM产线建设（规划月产6万片HBM晶圆，远期目标月产60万片晶圆）。



■ 图12：华为昇腾下游客户分配结构、字节跳动400亿锁单与中国移动AI推理集采格局 — 2025年昇腾63–64万颗核心客户出货分配量、2026年字节跳动超400亿元（56亿美元）锁定75万颗950PR近半产能，以及中国移动2025–2026年AI推理集采（7,058台/51.12亿元）中CANN生态占66.1%预算（5,000套/33.81亿元）与整机OEM分食明细。

2. 下游客户端："政企运营商基本盘"与"互联网巨头百亿级锁单"双轮驱动

- **2025年客户出货分配（总计63万–64万颗，含910C约33万–34万颗、910B近30万颗）**：政企与基建刚性底座合计 **34万颗**（地方超算14万、地方运营商10万、云厂及金融能源央企10万）；互联网及民营厂商市场化增量合计 **近30万颗**（字节跳动10万、腾讯5万、阿里3万、百度2万、美团1.5万及金山云、深信服等），且互联网大厂采购910C占比超90%。
- **2026年字节跳动超400亿元创纪录锁单**：受2025年4月H20断供催化，**字节跳动在2026年向华为下达总额超400亿元人民币（约56亿美元）的昇腾芯片采购大单（首批交付百亿级）**，直接锁定华为2026年 **近半数昇腾950PR产能**，带动阿里、腾讯紧随预购，使 **2026全年75万颗950PR产能一经释放即被抢购一空**。
- **三大运营商与金融集采标包落地**：**中国移动**在2024–2025年超150亿元（8,054台）智算集采后，于2025–2026年启动 **7,058台（51.12亿元含税）** 推理型集采，其中 **CANN（昇腾）生态独占5,000套（套数占70.8%，金额33.81亿元占66.1%）**，由河南昆仑技术、华鲲振宇、长江计算等昇腾整机伙伴分食，其余 **2,058套类CUDA设备（17.31亿元，占33.9%）** 由新华三、中兴、浪潮中标；**中国电信** 15.6万台集采中国产化G系列占 **67.5%**（ARM占45%、海光C86占8.7%）；**工商银行** 30亿元海光芯片服务器集采则由浪潮信息独家拿下。

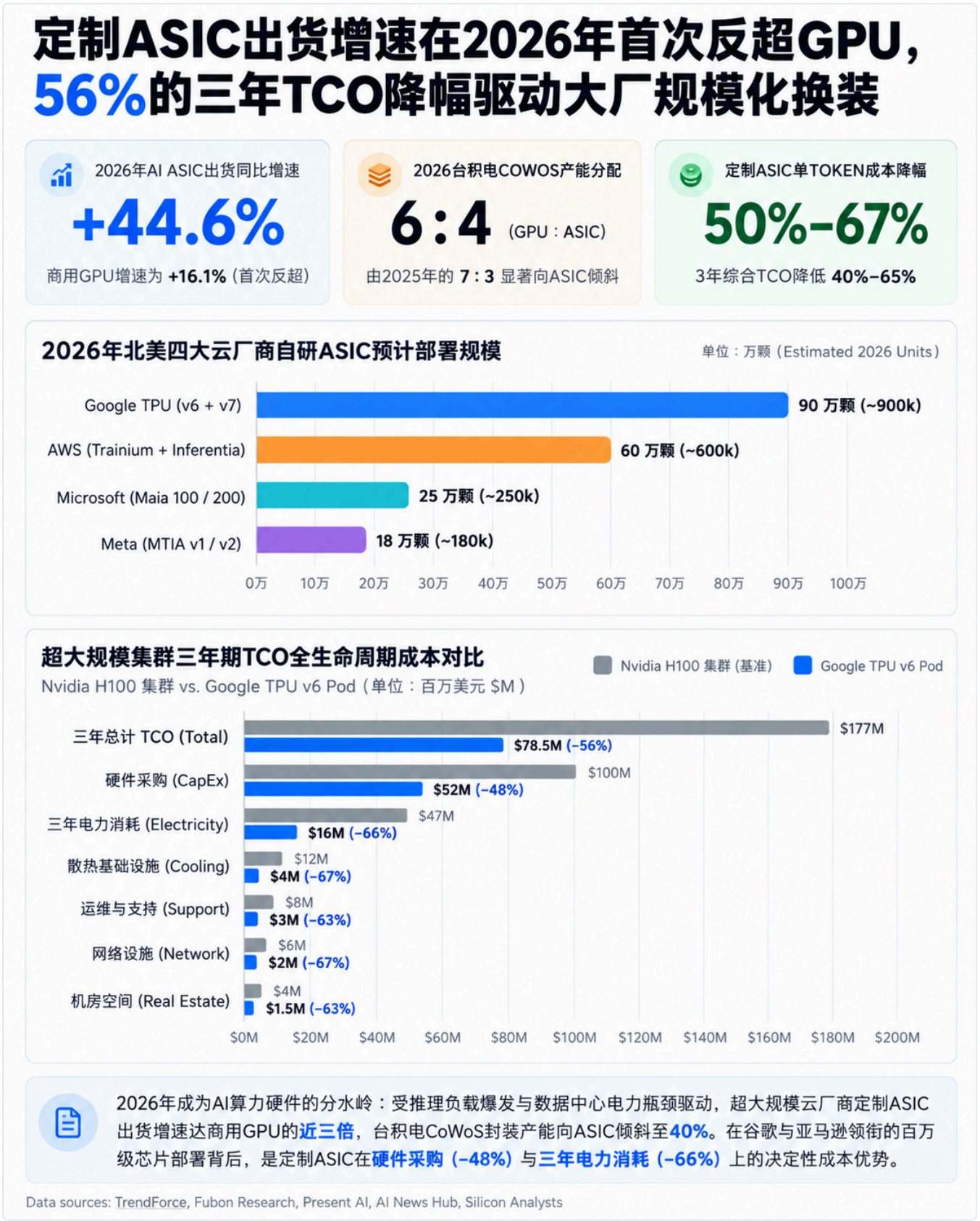
6.3 海光信息与寒武纪：晶圆封装底座绑定与服务器OEM关联生态

海光信息与寒武纪深刻体现了国产Fabless设计龙头对本土制造底座与头部服务器OEM渠道的双向依托：

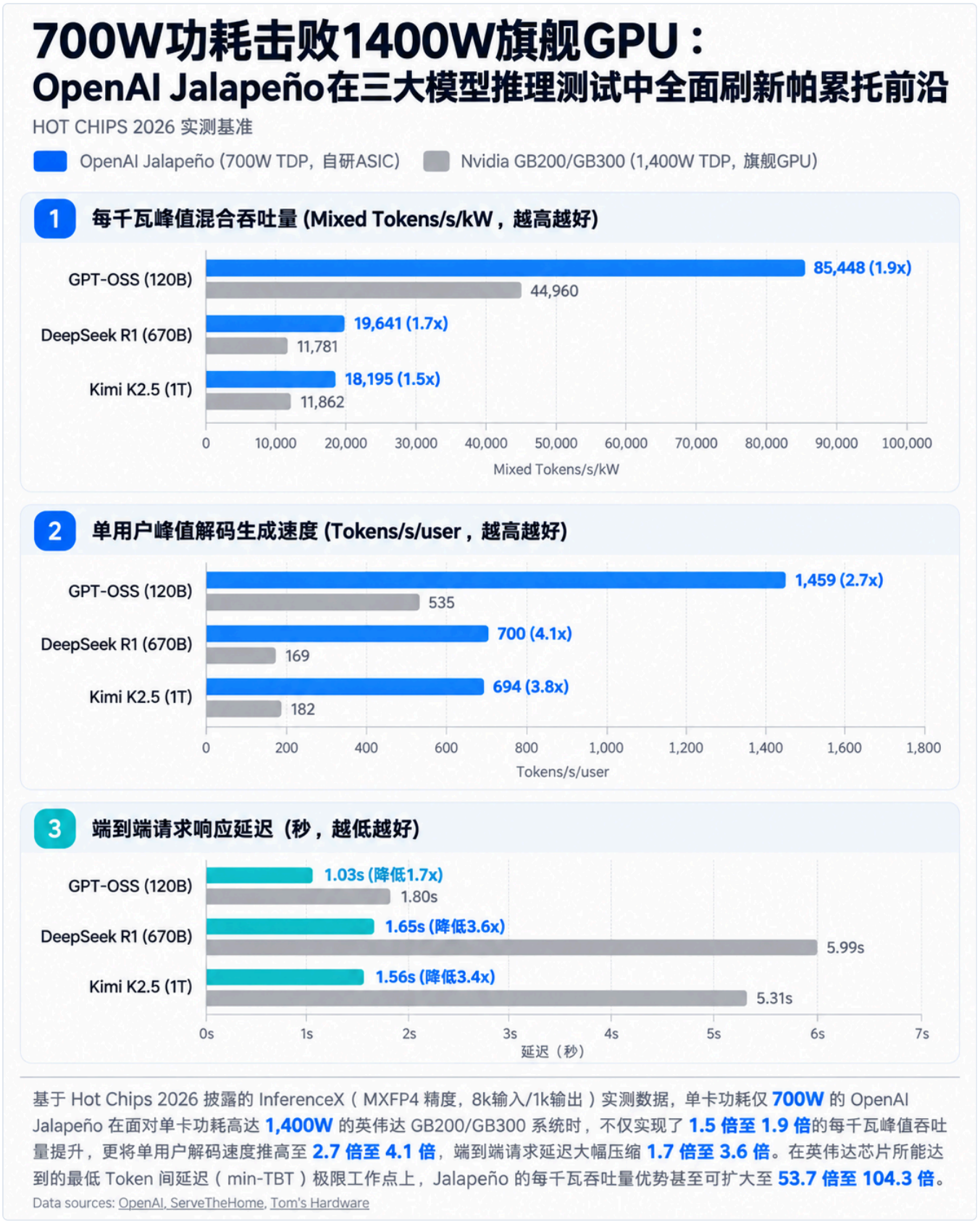
- **海光信息**：依托早年x86技术授权构建了C86 CPU与"类CUDA"深算DCU双产品线，上游与中芯国际、盛合晶微、长电科技深度协同，下游除第一大关联整机厂（中科曙光体系占56.68%）外，全面覆盖 **浪潮信息、新华三、中兴通讯**，在金融（工行30亿集采）、运营商及互联网大厂非自研份额中稳步放量。
- **寒武纪**：在中短期国产单卡及通用生态热度排序中，**思元590、昆仑芯P800、海光深算三号、昇腾910B依次递减**（昇腾则在超节点系统级能力上领跑）。寒武纪通过与 **浪潮信息**（元脑生态适配NF5468M5/NF5280M5）、**联想集团**、**中科曙光** 及 **科大讯飞** 紧密合作，奠定了思元590在互联网与智算中心的整机交付基础。

6.4 海外AI算力巨头与定制化ASIC供应链全景对照：英伟达Blackwell闭环、博通与谷歌TPU Turnkey联盟及吉瓦级长单

在全局视野下，海外正同步上演"Blackwell机架级闭环 vs. 博通/云厂定制ASIC大倒戈"的路线决战：



■ 图13：定制ASIC出货增速在2026年首次反超GPU，56%的三年TCO降幅驱动北美大厂规模化换装 — 2026年全球AI ASIC出货同比增速达+44.6%（商用GPU为+16.1%），台积电CoWoS产能分配达6:4（GPU:ASIC），北美四大云厂2026年部署规模（Google TPU 90万颗、AWS 60万颗、Microsoft 25万颗、Meta 18万颗），以及Google TPU v6 Pod较Nvidia H100集群三年TCO降低56%（$78.5M vs $177M）。



■ 图14：700W功耗击败1400W旗舰GPU——OpenAI Jalapeño在三大模型推理测试中全面刷新帕累托前沿 — Hot Chips 2026披露的InferenceX实测数据，OpenAI携手博通打造的700W 3nm自研推理ASIC Jalapeño在GPT-OSS (120B)、DeepSeek R1 (670B)、Kimi K2.5 (1T) 上全面击败1400W英伟达GB200/GB300（每千瓦吞吐高1.5x–1.9x，单用户解码快2.7x–4.1x，端到端延迟降1.7x–3.6x，极限min-TBT点每千瓦吞吐优势达53.7x–104.3x）。

1. 英伟达：Blackwell机架级液冷闭环与200亿美元并购Groq反围剿

英伟达通过Blackwell（NVL72机架级液冷、MGX模块化架构及向第三方CPU/ASIC开放的**NVLink Fusion**）深度绑定台积电CoWoS、**工业富联/广达** ODM与 **安费诺** 高速互连；同时为反击云厂推理ASIC蚕食，于2025年12月斥资 **200亿美元** 收购推理芯片独角兽 **Groq**（2026年3月发布Groq 3 LPU），推出物理分离Prefill/Decode的 **Rubin CPX / VR200 NVL144 CPX** 与 **CuTile** 编程模型，全方位锁死软硬件生态。

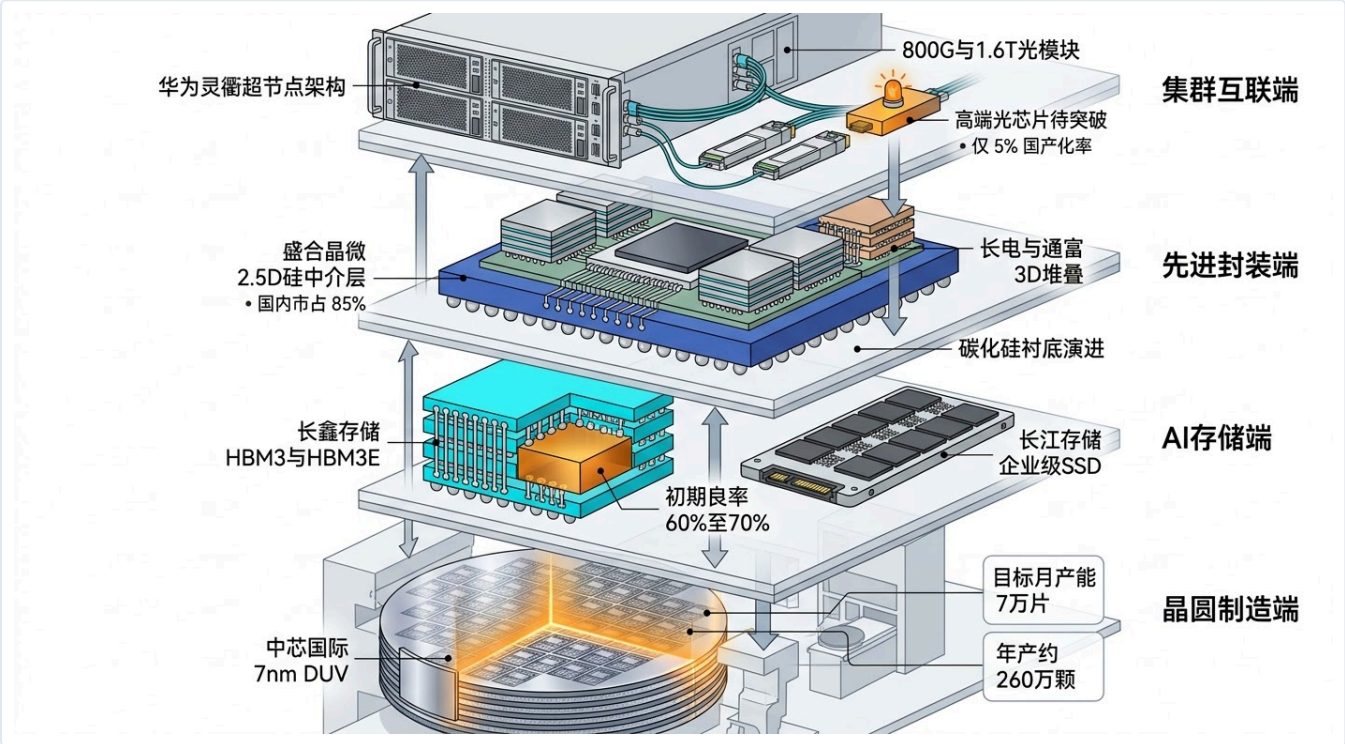
2. 博通与谷歌TPU Turnkey联盟、中国光模块嵌入及吉瓦级长单

博通（Broadcom）与美满电子（Marvell）垄断了全球约95%的定制AI ASIC联合设计市场（博通FY2026 AI营收指引高达 **560亿美元**，同比暴增180%）。在谷歌TPU研发中，谷歌负责脉动阵列MXU、OCS光交换与软件栈，博通则承担底层IP集成（SerDes/HBM/PCIe）、台积电晶圆与CoWoS产能锁定、良率风险缓释及搭配Tomahawk 6交换芯片的**全流程交钥匙（Turnkey）交付**。其五层全球供应链呈现三大核心动向：

- **设计多元化与封装外溢**：博通占谷歌TPU **70%–80%** 份额（TPU v7 Ironwood单卡FP8达4,614 TFLOPS、配192GB HBM3e；TPU 8i已开始出货），谷歌同时引入 **联发科** 参与TPU v8设计并自研将Gemini蚀刻进硅片的 **“Frozen v2”芯片**（预计每瓦Token效率提升6–10倍）；封装端除台积电CoWoS、日月光（ASE）与艾克尔（Amkor）外，**英特尔（Intel）已通过EMIB-T先进封装切入谷歌TPU供应链**。
- **中国光模块制造深度嵌入全球算力链**：在TPU及海外超大规模集群中，除安费诺、泰科电子（TE）、Lumentum外，**中国光模块龙头中际旭创（主供400G/800G模块）与新易盛成为海外算力链不可或缺的中国制造力量**。
- **2026年4月博通连续锁定三份吉瓦级长单**：① **4月6日与谷歌**续签至2031年的TPU长期供应协议（2026年预计贡献约 **210亿美元** 收入）；② **4月7日披露从2027年起向Anthropic** 供应约 **3.5吉瓦（GW）** 的Google TPU算力（Anthropic另与AMD签署2GW MI450协议）；③ **4月14日与Meta** 延长定制芯片（MTIA系列）合作至2029年，初始承诺超 **1吉瓦（GW）** 并向多吉瓦扩展。此外，AWS在2026年携手世芯电子（Alchip）与纬颖出货60万–150万颗Trainium芯片，微软规模部署25万颗3nm **Maia 200**，但海外ASIC同样面临 **静态AoT编译器高墙**（定制Pallas/NKI内核比自动编译器快3.7x–19x）与云平台强绑定约束。

07 CHAPTER 07 · FOUNDRY, MEMORY & PACKAGING
底层制造、AI存储三级突围、先进封装与光互联卡点——决定算力供给上限的四大物理支点

将视角下沉至半导体物理制造底层，可以清晰看到2026年中国AI算力产业所依赖的**晶圆制造、AI存储、2.5D/3D先进封装以及集群光互联**四大物理支点及其真实的产能天花板。



■ 图15：中国AI算力底层物理底座四维全景图——晶圆制造、AI存储、先进封装与集群互联 — 自下而上解构中国AI算力的四大物理底座——晶圆制造端（中芯国际7nm DUV，目标月产能7万片、年产约260万颗）、AI存储端（长鑫存储HBM3/HBM3E初期良率60%–70% + 长江存储企业级SSD）、先进封装端（盛合晶微2.5D硅中介层国内市占85% + 长电与通富3D堆叠 + 碳化硅衬底演进）与集群互联端（华为灵衢超节点架构 + 800G与1.6T光模块 + 国产化率仅5%的高端光芯片待突破）。

7.1 先进晶圆制造：中芯国际7nm DUV的极限压榨与良率爬坡

国内 **85%以上** 的国产AI芯片产能高度依赖 **中芯国际（SMIC）** 的 **N+2工艺（等效7nm DUV多重曝光）**：

- **月产能与全年产出上限**：2026年2月实际月产能尚不足 **2万片晶圆**（台积电同代际超15万片/月）；即便按2026年底扩产至 **7万片/月** 的极限目标计算，全国全年高端AI芯片总产能也仅约 **260万至300万颗**（面对420万颗需求存在约40%刚性缺口）。
- **良率与单片制造成本**：受DUV多重曝光套刻精度约束，昇腾910C（双Die合封）初期晶圆综合良率仅 **35%–36%**（目标2026年底升至 **40%**），导致 **910C单片制造成本高达14.5万元人民币，**

显著高于910B的9万元，这也是字节、阿里扫空库存、现货二级市场溢价三成以上的物理根源。

7.2 AI存储三级突围：长鑫存储 4F² DDR5 RDIMM、12层HBM3/HBM3E试产联调与长江存储PCIe 5.0企业级eSSD全栈替代

面对海外存储三巨头推动的"存储超级周期"（美光单季毛利率飙至84.9%，SK海力士于2026年3月向ASML下达80亿美元约30台EUV史上最大单笔订单）及断供限制，**长鑫存储（CXMT）** 与 **长江存储（YMTC）** 在无EUV条件下，通过**"三维立体结构重构 + 晶圆级混合键合"**实现了服务器DRAM、HBM与企业级eSSD的历史性三级突围：

TABLE 11 · 第七章 · AI存储三级突围 Yavin质观 · 核心数据矩阵

中国 AI 存储产业链三大核心赛道突围实绩与全球冲击

在缺少EUV光刻机约束下，长鑫与长江存储以"三维立体结构 + 晶圆级混合键合"实现换道超车

存储细分赛道	本土突围主体与架构路线	产能规划、良率与设备配套	对全球存储格局与价格冲击评估
服务器主存 (Server DRAM) **长鑫存储**	• 新一代 **4F² 架构 DDR5 RDIMM** (2026年底推出) • **垂直环绕栅极 (VSG)** 缩减33%面积 + **晶圆对晶圆(WoW)混合键合**	• 公开专利族 **超300件**，实际有效密度 **增幅约40%** • 直接拉动 **拓荆科技 (键合机)** 与 **中微公司 (高深宽比刻蚀)**	• 较三星/海力士商业化路线图(2028年推4F²) **名义提前两年** • 倒逼三星10月15日紧急确认完成4F²开发，压制远期DDR5暴利
高带宽内存 (HBM2 / 3 / 3E) **长鑫存储**	• **12层堆叠 HBM3** (2026.05规模试产) + **HBM3E系统级联调** • 16/17nm DUV多重曝光 + TSV + MR-MUF，单栈带宽 **819 GB/s**	• 科创板募资 **近600亿**，划拨 **月产6万片晶圆**；良率 **60%–70%** (年底冲80%) • **长电(良率98.5%)、沛顿(98.2%)、盛合晶微(>99.5%)** 封测闭环	• 已向 **昇腾、寒武纪、平头哥、海光** 供货联调 • **2027年产 300–400万堆栈，独立支撑 80–100万颗国产AI芯片**，终结海外95%垄断
企业级 SSD (PCIe 5.0 eSSD) **长江存储**	• PCIe 5.0家族：**PE501 (122.88TB)**、**PE522 (3100K IOPS)**、PE511 • 自研 **晶栈 Xtacking 4.0**，量产267层，**超300层下半年量产**	• 前两厂月产20万片，三期年底完工 **迈向 30万片/月** • 联合 **平头哥、大普微、得瑞领新、英韧** 实现主控+闪存全国产	• **eSSD出货两年暴增 8倍！2026Q2全球NAND份额达14%** • **超越铠侠与美光跃居全球第三；三星反向购买其Xtacking专利** 用于第10代V-NAND

资料来源：上市公司财报/招股书、IDC、TrendForce、SemiAnalysis 及 Yavin质观 整理（2026.10）

01 **服务器DRAM（长鑫存储以立体架构换平面微缩，奇袭 4F² DDR5 RDIMM）**：三星与SK海力士仰仗EUV沿用 6F² 推进微缩并将 4F² 推迟至2028年；长鑫存储则在2026年10月10日正式披露基于 **垂直环绕栅极（VSG）晶体管（单元面积缩减1/3）+ 晶圆级混合键合** 的 **4F² DDR5 RDIMM** 路线图（2026年底前推出），较韩系巨头领先两年，直接带动 **拓荆科技**（键合设备）与 **中微公司**（高深宽比刻蚀/CMP）收获晶圆加工强度翻倍红利。

02 **HBM（跨越1,300万颗进口库存悬崖，HBM3试产与HBM3E系统级联调并行）**：随着2025年支撑百万颗出货的 **约1,300万颗进口HBM库存** 耗尽，长鑫存储大幅前移时间表，于2025年底交付 **12层堆叠HBM3工程样品**，2026年5–6月完成首轮验证并步入小批量试产（划拨月产能 **6万片晶圆**，在大基金三期3,440亿元与50亿元存储专项基金支持下推动国产HBM渗透率破30%，华为亦推出HiBL 1.0与HiZQ 2.0自研方案）；2026年下半年，长鑫存储 **HBM3E样片** 已与 **阿里平头哥、寒武纪、华为昇腾** 展开深度系统级联调，冲刺2027年量产 **300万–400万个12-hi HBM3/HBM3e堆栈**（足以独立支撑 **80万–100万颗** 国产高端AI芯片出货）。

03 **企业级SSD（长江存储Xtacking 4.0跃居全球第三）**：中国存储产业已打通“长江存储eTLC颗粒 + 英韧/平头哥/大普微/得瑞领新主控 + 长鑫DDR缓存”全国产链路；长江存储于2026年3–4月密集发布三款 **PCIe 5.0企业级SSD**——超大容量QLC型号 **PE501（122.88TB，读能效比为HDD的19倍）**、AI训练旗舰 **PE522（随机读3,100K IOPS，4K写延迟5微秒）** 与通用主力 **PE511**，推动eSSD出货两年暴增8倍，2026Q2全球NAND份额达 **14%**，超越铠侠与美光跃居全球第三。

7.3 先进封装与CoWoS替代：盛合晶微独占国内硅基2.5D八成份额，碳化硅（SiC）中介层蓄势待发

- **盛合晶微领跑硅基2.5D量产**：作为中国大陆唯一实现硅基2.5D大规模量产的企业（全球仅四家），**独占国内85%市场份额**（华为占其营收74.4%），TSV转接板良率>99.5%、HBM封装良率**>95%**，并已启动 **HBM4封装小批量试产**。
- **多层次封测与材料设备梯队**：**长电科技**（XDFOI 8层HBM3E封装良率98.5%）、**通富微电**、**沛顿科技**（HBM3堆叠良率98.2%）与 **武汉新芯** 构成封测梯队；**赛腾股份**（检测）、**深南电路/兴森科技**（基板）、**鼎龙股份**（PSPI与抛光垫）及 **安集科技**（抛光液/电镀液）补齐材料设备短板。
- **碳化硅（SiC）中介层前瞻替代**：针对高功耗下传统硅中介层（热导率仅148 W/m·K）易热膨胀失配开裂的瓶颈，国内正依托 **天岳先进**（12英寸SiC衬底已量产，全球第二）、**三安光电** 及 **晶盛机电/晶升股份** 的全球领先优势，推进 **SiC中介层替代硅中介层与混合键合技术**，目标将互连密度再提升 **10倍以上**。

7.4 集群光互联：中国光模块统治全球市场，但上游高端光芯片（国产化率仅5%）亟待补短板

- **光模块制造端的全球统治力**：以 **中际旭创、新易盛** 为代表的中国厂商占全球TOP 10光模块厂商 **7席**。2026年国内智算中心光模块需求从2025年的270万只飙升至 **2,500万只**（预计实际出货1,000万–1,500万只，供需缺口约 **40%**）；代际上 **800G占70%为主力**，超节点 **1.6T光模块已商业化放量**（渗透率从15%向60%演进），**3.2T** 与 **CPO/NPO光引擎** 加速推进。
- **核心隐忧与补短板窗口**：**上游高端光芯片（高速EML激光器、CW光源及探测器芯片）国产化率目前仅约5%**，这一“模块封装强、上游芯片弱”的软肋已成为继HBM之后的下一个核心国产替代窗口。

08 CHAPTER 08 · ENDGAME & LEADING INDICATORS
终局推演与产业风险研判——算力产业链演进趋势及未来12至18个月检验替代成败的三大先行指标

8.1 算力产业链演进三大深层趋势与结构性财务风险研判

综合海内外核心算力企业的上下游图谱与工程实测，可以提炼出当前AI算力产业链演进的三大深层趋势与结构性风险：

01 **上游“先进封装 + HBM + 高阶互连”已全面取代单纯的逻辑制程，成为决定芯片出货上限的第一性瓶颈**：无论是谷歌TPU引入Intel EMIB-T封装以缓解台积电CoWoS紧缺，还是寒武纪第一大供应商采购占比飙升至55.34%，抑或是华为昇腾依托384/950超节点拉动华丰科技224G高速线模组与兴森科技ABF载板翻倍增长，均表明算力竞争已演变为系统级工程与核心瓶颈产能的争夺。在国内市场，中芯国际7nm DUV 7万片/月产能、盛合晶微2.5D封装产能与长鑫科技每月6万片HBM晶圆产能的良率爬坡进度，将直接决定未来三年国产AI芯片的整体供给天花板。

02 **下游需求结构从“分散的政企智算工程”全面转向“头部互联网与大模型巨头的百亿级与吉瓦级锁单”**：在海外市场，算力采购已演进为Anthropic 3.5吉瓦、Meta超1吉瓦乃至xAI计划到2027年将数据中心容量扩大7倍的“电力级”长期承诺；在国内市场，2026年字节跳动超400亿元（56亿美元）锁定华为950PR近半产能的创纪录订单，以及阿里、腾讯的紧随抢购，标志着国产算力芯片已在商业化要求最严苛的互联网核心训推场景完成正循环验证。

03 **高客户集中度与应收账款积压带来的财务脆弱性不容忽视**：海光信息前五大客户占比90.28%（第一大关联客户占56.68%且应收账款占比超50%）、寒武纪前五大客户占比88.66%以及英伟达前三大客户占比升至44%，意味着芯片原厂与极少数下游寡头形成了深度捆绑。一旦下游云巨头资本开支节奏放缓（如北美五大云厂2026年合计7,250亿–7,750亿美元CapEx面临自由现金流折现大考），或内部自研芯片（如阿里平头哥真武系列、腾讯紫霄2.0、谷歌Frozen v2、Meta/OpenAI定制XPU）替代比例超预期上升，第三方商用芯片原厂将面临订单波动与估值乘数下修的双重冲击。

8.2 未来12至18个月检验国产替代成败的三个核心先行指标

站在2026年10月的产业节点向前看，国产AI芯片已经赢下了“份额破局（66%）”与“推理抢滩（非GPU占50.6%）”的上半场战役。但在接下来的2027年至2028年下半场，随着英伟达Vera Rubin等更新一代架构在海外拉开代差，国产算力能否从“政策与缺货驱动的被动替代”走向“TCO与生态驱动的主动替代”，取决于以下三个可量化的核心先行指标：

指标一：万亿参数基座模型从0到1预训练能否在国产万卡超节点上跑通并收敛

推理侧的Day-0适配已经成为行业及格线，不再构成差异化壁垒。真正的试金石在于：目前海思披露正在部署测试的 **25.6万张昇腾950DT超节点计算卡**，以及即将在2027年第一季度量产的**平头哥真武V900集群**，能否在未来12个月内，真正支撑起DeepSeek、阶跃星辰、通义千问或字节跳动等头部实验室完成下一代旗舰基座模型（如DeepSeek V5或10万亿–40万亿参数级多模态基座）**从0到1的完整预训练**。观察这一指标的微观锚点有两个：

- 一是DeepSeek于9月30日开源的 `TileLang-Ascend`、`DeepGEMM-Ascend` 与 `DeepEP-Ascend` 工具链，能否将万卡级稀疏MoE预训练的真实 **MFU从目前的40%–45%拉升至55%以上**，并将单次故障恢复时间稳定控制在 **5分钟以内**；
- 二是外部独立大模型团队是否开始实质性削减对存量H100/H800算力租赁的依赖，将核心预训练预算投向国产超节点。

指标二：外部商业化复购率与地方智算中心的真实亮灯率能否跨过生死线

2025年至2026年上半年的出货放量中，包含了大量互联网大厂“预防性战略备货”以及地方政府智算园区的“首期基建采购”。但首单交付不等于商业闭环——当前**全国智算中心约30%的平均GPU/NPU实际利用率**是一道刺眼的警报。在未来12至18个月内，随着首批建设的国产智算中心进入为期两年的折旧与运营考核期，必须紧密跟踪**商业化客户的二期复购率（Repurchase Rate）**与**地方智算中心的真实机柜亮灯率（Utilization Rate）**。唯有当独立GPU/ASIC厂商（如寒武纪、摩尔线程、沐曦、天数、燧原、壁仞）不仅能把板卡卖给国资平台，还能协同清程极智、无问芯穹等软件伙伴将当地特色产业应用跑起来，把机房利用率推升至 **60%以上的盈亏平衡线**，国产算力的市场盘才能摆脱对地方财政补贴的依赖，形成健康的现金流造血循环。

指标三：先进制程晶圆与HBM供应链能否支撑超节点路线4.5倍的芯片消耗膨胀

“超节点”路线在供应链逻辑上潜藏着一个核心悖论：**用系统能力换取单卡性能的代价，是同等算力输出下芯片消耗量乘以4.5倍**，将本土先进制程晶圆（2026年需求420万颗 vs. 供给260万颗，缺口40%）与HBM防线逼至极限。未来一年半观察国产替代天花板的最硬指标在于制造与封测端：① 中芯国际N+2工艺晶圆良率（35%–36%向40%+爬坡）与月产能扩充能否弥合40%刚性缺口；② 长鑫存储12层HBM3与HBM3E能否如期实现60%–70%良率向80%+及年产300万–400万堆栈的爬坡，支撑昇腾950DT与平头哥真武V900的激进吞吐；③ 盛合晶微、长电科技与天岳先进推进的12英寸SiC中介层及5%国产化率的高端光芯片能否完成量产攻关。唯有当晶圆、HBM与高密封装不再成为交付锁链，且单瓦特有效算力产出（Tokens/Watt）通过架构与MLA/TileLang算法协同压缩掉4倍电费溢价时，这场始于2026年的国产AI芯片放量潮，才能真正完成从“特殊时期的备胎突围”向“全球算力新极点”的历史跨越。

APPENDIX & METHODOLOGY · 资料口径与全量信源

附录：四级信源说明与320项参考资料索引

DATA & REFERENCES

资料口径与四级信源分级说明： （1）**上市公司财报与招股书（最高优先级）**：海光信息、寒武纪、摩尔线程、沐曦股份、长鑫科技、盛合晶微、NVIDIA、Broadcom、TSMC、SMIC等官方定期报告与运营商阳光采购网公示； （2）**权威第三方机构**：IDC、TrendForce、Counterpoint Research、SemiAnalysis、Bernstein、摩根士丹利、高盛等出货量与市场份额统计； （3）**顶会论文与开源工程**：华为CANN、DeepSeek开源周（FlashMLA/DeepEP/DeepGEMM/DualPipe/3FS）、TileLang-Ascend、NeurIPS 2025（TransMLA）及ICLR 2027（DoseRank）； （4）**供应链交叉验证**：未独立上市主体（海思、平头哥、昆仑芯、蓬莱实验室）之良率与锁单规模均经多源交叉核验。

一、国产AI芯片市场格局、超节点与软件栈核心信源

[01] IDC：英伟达在华AI芯片份额降至34% - 报价宝 (baojiabao.com/a717563)

[02] 国产GPU竞逐超节点谁来买单？ - QQ News (news.qq.com/rain/a/20261016A0CG...)

[03] IDC 报告：英伟达中国AI 芯片份额跌至34%，华为追到28% - 搜狐 (m.sohu.com/a/1087287600_211762)

[04] IDC 重磅揭晓2025 国产AI芯片出货量TOP8 - 金股讯 (jinguxun.com/article/7806627)

[05] 英伟达在中国AI芯片市场份额降至34% 华为升至28% - 网易 (163.com/dy/article/L94GR155...)

[06] 英伟达34%：中国AI 算力第一次「换轨」 - 网易新闻 (c.m.163.com/news/a/L99BSCNN0511...)

[07] 逛完WAIC才明白，国产算力不跟英伟达比单卡了 - OFweek电子工程网 (ee.ofweek.com/2026-07/ART-8420-28...)

[08] 超节点算力集群，能托住国产模型吗？ - 虎嗅网 (huxiu.com/article/4897125.html)

[09] 国产AI芯片算力实测：昆仑芯、摩尔线程能否正面硬刚英伟达H20? (post.m.smzdm.com/p/aggr2qed)

[10] 昆仑芯份额约3.5%居国产第三，3.2万卡集群落地，估值超400亿美元 (finance.ifeng.com/c/8vKG1edMWy3)

[11] 2026上半年中国市场数据中心AI加速器出货量：英伟达下降44% - 搜狐 (m.sohu.com/a/1089335516_120468...)

[12] 中国人工智能加速卡市场份额，2025 - IDC-Login (my.idc.com/getdoc.jsp)

[13] 当推理吃掉七成算力，国产芯片等来一个窗口期 - 南方财经网 (sfccn.com/2026/5-29/wMMDE1MjB...)

[14] WAIC 2026终极盘点：15大核心趋势，看透AI下半场 - 36氪 (m.36kr.com/p/3904188372485763)

[15] 阶跃星辰开源Step 5 Preview权重，华为昇腾等首发支持 - 什么值得买 (post.smzdm.com/p/a70pdm7d)

[16] 昇腾950超节点真机亮相2026世界人工智能大会- 华为 - Huawei (huawei.com/cn/news/2026/7/atla...)

[17] 华为昇腾Atlas 950 SuperPoD亮相WAIC 共计1024张昇腾卡 - 搜狐 (sohu.com/a/1051456277_115831)

[18] 全球领先！千卡超节点算力巨无霸，成功点亮 - 智源社区 (hub.baai.ac.cn/view/58734)

[19] 业界最大规模！首套1024卡华为昇腾950液冷超节点正式点亮 - 凤凰网 (feng.ifeng.com/c/8xFaWV9uAHL)

[20] MindIE-昇腾社区 (hiascend.com/developer/software/...)

[21] 告别CUDA？DeepSeek 给昇腾造了一套工具链：已开源 - 网易 (163.com/dy/article/L831FU95...)

[22] DeepSeek开源昇腾！全套工具替换英伟达，国产芯片接住前沿大模型 (news.qq.com/rain/a/20261002A047VQ)

[23] 昇腾vLLM-Ascend Skill实战 - 昇腾社区 (hiascend.com/developer/blog/deta...)

[24] 中国加速计算服务器市场规模将突破千亿美元，行业生态即将全面重塑 (my.idc.com/getdoc.jsp)

[25] 中国芯，杀破英伟达！ - 财经号- 中金在线 (mp.cnfol.com/8832/article/179214...)

[26] 国产AI芯片份额逼近英伟达，大厂转向硬件变现与产能争夺 - 火猫网络 (firecat-web.com/daily-news/17598)

[27] China Archives - Page 2 of 18 - IDC (idc.com/resource-center/blo...)

[28] China Archives - IDC (idc.com/resource-center/blo...)

[29] 36氪产业深度研究报道 (36kr.com/p/4028460994777479)

[30] 华为昇腾950超节点有多强？1024卡集群算力相当于3205张A100！ (hwcomputing.csdn.net/6a5bac0010ee7a33f28...)

[31] 昇腾950超节点斩获2026世界人工智能大会SAIL最高奖- 华为企业业务 (e.huawei.com/cn/news/2026/soluti...)

[32] 阿里发布AI芯片真武V900，算力较M890 提升3 倍 - CFM闪存市场 (chinaflashmarket.com/a/184294)

[33] 平头哥发布新一代AI芯片真武V900，明春上市 - 新闻- 科学网 (news.sciencenet.cn/htmlnews/2026/9/571...)

[34] 阿里推新款真武AI芯片！吴泳铭：3倍性能，国产最强 - 智东西 (m.zhidx.com/p/596203.html)

[35] 出货量超寒武纪，阿里系芯片公司揭开神秘面纱 - 36氪 (m.36kr.com/p/3680138297257857)

[36] 国产GPU/AI算力芯片现状（修订完整版· 截至2026年9月 (hwcomputing.csdn.net/6abe4994ad85947e68c...)

[37] 如何用一支指数覆盖“光芯”双主线？科创创业人工智能ETF ... - QQ News (news.qq.com/rain/a/20260515A066...)

[38] IDC：本土对手崛起英伟达在华数据中心AI芯片份额降至34% - QQ News (news.qq.com/rain/a/20261013A074...)

[39] 百度旗下昆仑芯启动上市辅导！2026年收入有望达65亿元 - 新浪财经 (finance.sina.com.cn/wm/2026-05-08/doc-i...)

[40] 国产AI芯片“三强”并起 - 中国工业新闻网 (cinn.cn/2026/05-25/GDBjZNKD...)

[41] AI时代百度终于押对了？ - 业界- 亿邦动力 (ebrun.com/20260709/686286.shtml)

[42] 三季报丨国产算力竞速，海光信息业绩稳增 - 网易 (163.com/dy/article/L9CH49QA...)

[43] 海光信息（688041）：国产CPU领军企业，DCU卡位算力芯片千亿蓝海 (pdf.dfcfw.com/pdf/H3_AP2025122318...)

[44] 国产算力基础设施CPU+DCU 双芯底座+ 生态枢纽海光信息从芯片 (m.eefocus.com/article/2011414.html)

[45] 原生开发 - 昇腾社区 (hiascend.com/developer/native-de...)

[46] RLinf/README.zh-CN.md at main - GitHub (github.com/RLinf/RLinf/blob/ma...)

[47] 昆仑芯完成RLinf适配验证，拓展具身智能强化学习开源生态 (kunlunxin.com/news/5190.html)

[48] RLinf：面向具身智能的“渲训推一体化”开源强化训练框架 - 青稞社区 (qingkeai.online/blog/RLinf)

[49] 阶跃星辰放开Step 5权重，国产芯片当天接住：一文看懂"权重"与"算力" (post.m.smzdm.com/p/agg54e2w)

[50] 昇腾万里，开放创新，共赢智能时代 (hiascend.com/activities/dynamic-...)

[51] 一文看懂Mooncake 原创 - CSDN博客 (blog.csdn.net/daydayup858/article...)

[52] Mooncake: A KVCache-centric Disaggregated Architecture for LLM (hub.baai.ac.cn/paper/c17f9070-9c6e...)

[53] 从真武V900看国产AI芯片：性能三倍跃升背后，算力生态如何重塑 (m.sohu.com/a/1087814474_128075)

[54] 华为Atlas 950 SuperCluster 50 万卡超集群2026 年Q4 上市- 产业链 (c114.net.cn/industry/22365.html)

[55] 阿里平头哥发布AI芯片真武V900 算力达真武M890的3倍 - 凤凰网 (feng.ifeng.com/c/8wd4io6s2JV)

[56] 华为云昇腾AI云服务 (pdf.dfcfw.com/pdf/H3_AP2024100716...)

[57] CANN全面开放，解锁创新范式，定义智能未来-官方技术文章-昇腾社区 (hiascend.com/developer/techArtic...)

[58] triton-ascend:基于昇腾平台的Triton编译框架项目- AtomGit - GitCode (gitcode.com/Ascend/triton-ascend)

[59] vLLM原生支持昇腾，加速大模型推理创新，社区尝鲜版本首发 (tech.ifeng.com/c/8h9WiwLsIfd)

[60] SGLang全面支持昇腾，使能大EP高性能推理-官方技术文章 (hiascend.com/developer/techArtic...)

[61] 同构+异构千卡——众智FlagOS以统一技术栈完成AI训练“全要素”验证 (hub.baai.ac.cn/view/53417)

[62] 对标英伟达平台，DeepSeek开源华为昇腾全套基础设施组件 - 财富号 (caifuhao.eastmoney.com/news/20261002180930...)

[63] DeepSeek开源昇腾基础组件，涵盖对标英伟达CUDA的高级语言编译 (finance.sina.com.cn/roll/2026-09-30/doc...)

二、AI算力芯片上下游供应链与客户集采核心信源

[01] 重磅！昇腾960超节点正式发布！NPO光引擎带来算力革新（附相关 (caifuhao.eastmoney.com/news/20260917134506...)

[02] 未知机构:《VIP 昇腾调研-1：2025年910B及910C出货量估计和核心 (fxbaogao.com/detail/5081717)

[03] 液冷服务器成为AI黄金赛道！华为力挺，两大国际芯片巨头下场 (lctexpo.com/news/669.html)

[04] Broadcom: Sideways for Nearly Two Months After Earnings, Is Now (tradingkey.com/research/stocks/avg...)

[05] 寒武纪的利润狂奔现金流还在追 - 网易 (163.com/dy/article/L74I6VPT...)

[06] 长电科技vs 盛合晶微：谁能主导中国高算力芯片的封装赛道？ - 与非网 (m.eefocus.com/article/192689.html)

[07] 商业头条No.135 | 解码长鑫逆袭：国产存储如何跨越周期 (m.jiemian.com/article/14831122.html)

[08] 海光信息(688041)_公司公告_海光信息：2025年年度报告新浪财经 (money.finance.sina.com.cn/corp/view/vCB_AllBu...)

[09] GOOGL's TPU supply chain. We know Broadcom and AMKR are key (reddit.com/r/TradingEdge/comme...)

[10] 51亿！2025最大AI服务器大单被这9家瓜分 - 智东西 (m.zhidx.com/p/498166.html)

[11] 员工套现27亿，股东浮盈80亿：海光信息AI盛宴绕不开中科曙光？ (znyj.ofweek.com/news/2026-08/ART-23...)

[12] 英伟达客户集中度持续上升，前三大客户收入占比达44% - 搜狐 (m.sohu.com/a/1075993031_313745)

[13] Broadcom Locks In AI Chip Deals with Google, Meta & Anthropic in (tomorrowaccess.com/en/news/18688)

[14] 除了杀回来的英伟达，寒武纪还要防着自己人 - 36氪 (m.36kr.com/p/3728263213920774)

[15] 芯片里大佬，海光信息VS中芯国际：谁是真龙？ - 网易 (163.com/dy/article/KRET6462...)

[16] 一对国产“算力搭子”的2025年- 经济观察网 – 专业财经新闻网站 (eeo.com.cn/2026/0415/839830.sh...)

[17] 是中国半导体史上最大的单笔AI芯片采购单。不仅锁了华为2 - 东方财富 (emcreative.eastmoney.com/app_fortune/article...)

[18] 华为昇腾950PR芯片的产业链涵盖了从上游芯片制造、核心元器件 (caifuhao.eastmoney.com/news/20260914151542...)

[19] 华为昇腾384超节点概念股受益相关龙头 - 布乐秀攻略网 (bulexiu.com/n/31443.html)

[20] 华丰科技深度绑定华为224G量产加速AI算力放量 - 中财网 (cfi.net.cn/newspage.aspx)

[21] 传字节跳动将以400亿人民币采购华为晶片 - 联合早报 (zaobao.com.sg/finance/china/story...)

[22] 运营商超百亿服务器集采大单：ARM处理器占比超70% - 观察者网 (guancha.cn/economy/2026_06_25_...)

[23] 海光信息（688041.SH）：国产CPU+DCU龙头 (pdf.dfcfw.com/pdf/H3_AP2023121516...)

[24] 中国本土算力企业收到工行、联通的100亿国庆“红包”| 海斌访谈 (yicai.com/news/102852852.html)

[25] 寒武纪发布云端AI芯片，进军服务器市场 - 搜狐 (sohu.com/a/230299306_114986)

[26] 浪潮与寒武纪签署元脑战略合作协议，推进领先AI方案落地 (vision-systems-china.com/deinews.asp)

[27] 阿里平头哥真武PPU - 阿里巴巴Alibaba | AI算力卡百科 - MirrorFrog (mirrorfrog.com/docs/gpu/t-head-zhe...)

[28] 刚刚，阿里AI芯片“真武810E”来了！已实现多个万卡集群部署 - 智东西 (m.zhidx.com/p/531878.html)

[29] 阿里首次证实自研AI芯片PPU存在，大规模出货已有一年 - 财经杂志APP (mycaijing.com/article/detail/563211)

[30] 昆仑芯与百度联合发起飞桨+文心大模型硬件生态共创计划 (kunlunxin.com/news/3550.html)

[31] 百度昆仑芯的“家底”和“包袱” - 澎湃新闻 (m.thepaper.cn/newsDetail_forward_...)

[32] 2025年Blackwell将占英伟达高端GPU出货量的80%以上 (user.guancha.cn/main/content)

[33] Broadcom to supply Anthropic with 3.5 gigawatts of Google TPU (tomshardware.com/tech-industry/broad...)

[34] Google Speeds Up Next-Gen TPU Mass Production - Businesskorea (businesskorea.co.kr/news/articleView.html)

[35] With Google's awakening, the price of "allies" who worked together (mk.co.kr/en/it/11476894)

[36] 华丰科技：部分224G高速连接器已根据客户需求实现批量交付 - 凤凰网 (feng.ifeng.com/c/8wnBFaXVTAd)

[37] 应对算力需求增长，传字节跳动明年将采购超400亿元华为昇腾芯片 (chinaflashmarket.com/a/183486)

[38] 三大运营商相继开标，最大规模AI服务器订单要来？ - 中国基金报 (chnfund.com/article/AR91c2f108-...)

[39] 中移动启动新一轮智算中心采购AI服务器成最大亮点 - 财联社 (cls.cn/detail/1653817)

[40] 中国移动165亿元服务器集采开标上下游厂商纷纷切换至算力赛道 (stcn.com/article/detail/1203...)

[41] 谁是大赢家！中国移动50亿AI推理服务器集采落地 - 新浪财经 (finance.sina.com.cn/tech/roll/2026-01-2...)

[42] 从集采看需求:电信市场需要怎样的算力服务器？ - 流媒体 (lmtw.com/mzw/content/detail/...)

[43] 寒武纪与浪潮签署元脑战略合作，加速AI布局-立创商城 (szlcsc.com/info/12788.html)

[44] 海光芯正上市聆讯通过：营收12亿增41%，净亏损却扩大近5倍至1亿 (finance.ifeng.com/c/8tt1dmaMipW)

[45] What role does Broadcom play in semi-custom ASIC (like google (forums.anandtech.com/threads/what-role-d...)

三、中国AI芯片制造、存储突围与先进封装核心信源

[01] 万亿参数模型DeepSeek-V4“解锁”国产芯片 昇腾适配意味着什么？ (finance.sina.com.cn/stock/t/2026-04-24/...)

[02] 英伟达与华为的6个百分点：中国AI芯片市场竞争格局生变 - 36氪 (m.36kr.com/p/4024302190728839)

[03] 英伟达在中国AI芯片市场份额降至34% 华为升至28% - 凤凰网 (i.ifeng.com/c/8xD8cmftcEn)

[04] 华为AI芯片中国市场出货量逼近英伟达 - 网易 (163.com/dy/article/L96NM4FS...)

[05] 6000亿寒武纪，可以开香槟了吗？-钛媒体官方网站 (tmtpost.com/7676298.html)

[06] 长鑫据报开始生产HBM3E，国产AI芯片有了第一个本土高带宽内存 (163.com/dy/article/L5O2Q8PB...)

[07] 国产GPU龙头回应一切！ - 智东西 (m.zhidx.com/p/500170.html)

[08] 暴跌44%！英伟达中国AI芯片份额缩水至34%，华为追至28%国产 (163.com/dy/article/L96NJQ97...)

[09] 2026 美國晶片管制新政下的「紅線」與「生機」 - taiwan-industry.com (taiwan-industry.com/2026-%E7%BE%8E%E5%9...)

[10] 大厂醒悟了，卷模型不如卖芯片 - 搜狐 (m.sohu.com/a/1087733440_121948...)

[11] AI芯片算力等技术参数分析- 吴建明wujianming - 博客园 (cnblogs.com/wujianming-110117/p...)

[12] 2025-2026国产AI芯片内存带宽瓶颈突破：自研HBM量产在即 (post.m.smzdm.com/p/a502r9w7)

[13] TileLang-Ascend 算子性能优化方法与实操_昇腾 - CANN开发者社区 (cann.csdn.net/69f02d2c0a2f6a37c5a...)

[14] 亿欧智库| 2024信创产业新发展趋势及百强报告_企业服务行业研究报告 (iyiou.com/research/202501271439)

[15] 电信服务器百亿集采出炉！联想、中兴、浪潮信息等入围 - 财联社 (cls.cn/detail/1838130)

[16] 昇腾910B与910C的性能与定位全景对比 - 算力租赁 (omniyq.com/sys-nd/153.html)

[17] 华为昇腾910C不是最好的AI芯片，但它是中国AI产业唯一的选择 (cloud.tencent.com/developer/article/2...)

[18] 昇腾910B vs NVIDIA A100/H100：国产GPU算力租赁选型与迁移实践 (cann.csdn.net/6a4b7009662f9a54cb8...)

[19] 未知机构:《HW昇腾专家调研纪要20240810 | 研报》 - 发现报告 (fxbaogao.com/detail/4441341)

[20] 英伟达市值蒸发创纪录，DeepSeek掀起AI算力革命意味着什么？ (jazzyear.com/article_info.html)

[21] DeepSeek-V3 技术报告 (alphaxiv.org/zh/abs/2412.19437)

[22] DeepSeek-V3 推进技术突破与创新，降低模型开发门槛 (download.s21i.co99.net/13115299/0/1/ABUIAB...)

[23] 灵衢（UB）：打破算力垄断，中国算力互联的「终极答案」 - 存储在线 (dostor.com/p/99982.html)

[24] 昇腾算力集群高可靠性特点 (hiascend.com/developer/blog/deta...)

[25] 十万卡集群算力利用率仅约两成，华为的解法是灵衢互联 - 华尔街见闻 (wallstreetcn.com/articles/3782011)

[26] 算法创新Lab_NEWS_昇腾实现万卡级算力集群突破_华为云 (huaweicloud.com/lab/algorithm/news/...)

[27] IDC 最新AI芯片出货量解读：国产AI算力大洗牌，国产GPU“四小龙”谁 (163.com/dy/article/L94VS1O9...)

[28] 国产AI芯片大卖，上半年谁真正赚到了钱？ - 钛媒体 (tmtpost.com/8138350.html)

[29] 海光信息的芯片技术突破曙光意义重大 - 东方财富 (emcreative.eastmoney.com/app_fortune/article...)

[30] 国产AI芯片之争才刚刚开始 (m.cyzone.cn/article/731528.html)

[31] 华为43%，寒武11%：2026年国产AI芯片的产能饥饿游戏 - 钛媒体 (tmtpost.com/8051488.html)

[32] MTT S4000 | 大模型智算加速卡 - 摩尔线程 (mthreads.com/product/S4000)

[33] Moore Threads 摩尔线程MTT S4000 (2023) | AI 算力卡百科 (mirrorfrog.com/docs/cards/others/m...)

[34] 长鑫存储HBM3E进入小规模试生产，平头哥、寒武纪等国产AI芯片正 (caifuhao.eastmoney.com/news/20260902091626...)

[35] 7nm产能严重不够，10多家企业，在抢中芯国际的先进芯片 - 网易 (163.com/dy/article/L1KUQMPG...)

[36] 华为43%，寒武11%：2026年国产AI芯片的产能饥饿游戏 - 新浪财经 (finance.sina.cn/stock/jdts/2026-07-...)

[37] 摩尔线程MTT S4000 AI GPU助力30亿参数大模型训练 - CSDN博客 (blog.csdn.net/zhuzongpeng/article...)

[38] DeepSeek开源面向华为昇腾算力平台的基础设施组件 - 凤凰网科技 (tech.ifeng.com/c/8wpwu6MVfRV)

[39] 国产AI的“伟大闭环”！DeepSeek“不起眼的一句话”可能改变整个GPU (wallstreetcn.com/articles/3756531)

[40] 为什么说DeepSeek这次开源，可能和R1一样重要？ - 网易 (m.163.com/dy/article/L91KR1FM...)

[41] DeepSeek修好了，此前崩了近5小时 - 证券时报 (stcn.com/article/detail/4219...)

[42] SSD国产替代时机已至，PCIe 5.0和AI需求爆发成最大推动力 (memorys.com/2024/a/13)

[43] 产业丨国产企业级SSD群雄并起，存储爆发迎突围时刻 (ee.ofweek.com/2026-04/ART-8420-28...)

[44] 国产HBM量产破局！长鑫12层HBM3试产，良率爬坡成关键 (post.m.smzdm.com/p/apq4wnzx)

[45] 国产HBM3突破！ - 与非网 (m.eefocus.com/article/1885937.html)

[46] 长鑫存储向华为交付HBM3样品，国产高带宽存储取得突破 (dram.com.cn/page/views.asp)

[47] 长鑫HBM3样品通过验证,2027年大规模投产 (huizhan086.com.cn/zhxw/282.html)

[48] 先进封装竞速：台积电守擂、英特尔突围、封测六强抢跑 - 网易 (163.com/dy/article/L4IH2DQJ...)

[49] 光通信产业：AI算力点火的黄金赛道——1.6 T放量 - 东方财富 (finance.eastmoney.com/a/202610153896640 73...)

[50] 国产GPU：为什么今明两年肯定不够用 - 腾讯云 (cloud.tencent.com/developer/article/2...)

[51] 高端先进封装：AI 时代关键基座，重视自主可控趋势下的投资机会 (pdf.dfcfw.com/pdf/H3_AP202508151 7...)

[52] 深度剖析HBM千亿蓝海，AI算力激战下供需新格局 - 长城证券 (cgws.com/cczq/ggdt/ccyj/2025...)

[53] AI算力破局关键，先进封装板块暴涨，风口来了？ - 36氪 (m.36kr.com/p/3643303514918530)

四、中美算力双轨制、实测对标与算法演进核心信源

[01] AI GPU Market Share NVIDIA vs AMD vs Intel: 2026 Statistics (commandlinux.com/statistics/ai-gpu-m...)

[02] Nvidia Customers Become Rivals: Custom Chip Race [2026] (tech-insider.org/nvidia-customers-ri...)

[03] Custom AI Chips Outpace Nvidia GPU Growth in 2026 - Tech Times (techtimes.com/articles/317225/202...)

[04] Google, Amazon and Meta Are Out-Growing Nvidia With Their Own (startupfortune.com/google-amazon-and-m...)

[05] Anthropic Is Designing Its Own Chips. Why Every AI Lab Wants Silicon (medium.com/@kidscarebridge/ant...)

[06] Amazon May Sell Its AI Chips: The Nvidia Challenge - Digital Applied (digitalapplied.com/blog/amazon-custom-...)

[07] 国产AI 芯片迎来关键转折点：英伟达份额暴跌 华为上升 - 时间线- 搜狐 (timeline.sohu.com/news/B9ZAWejl4O)

[08] 英伟达中国GPU 出货量暴跌44%！仍居第一 - 搜狐 (m.sohu.com/a/1087222340_121948...)

[09] 英伟达中国AI芯片份额跌至34% 国产厂商加速进入规模化交付阶段 (caifuhao.eastmoney.com/news/20261014110523...)

[10] ASIC Set to Outpace GPU? NVIDIA's Scale-Up and Beyond (trendforce.com/insights/nvidia-sca...)

[11] AI Accelerator Chips - 2026 Outlook (assets.bbhub.io/professional/sites/...)

[12] The custom AI ASIC state of play (May 2026) - Tom's Hardware (tomshardware.com/tech-industry/semic...)

[13] $775 billion—who do the orders ultimately go to? - Binance (binance.com/en-AE/square/post/3...)

[14] Meta做了一个决定，存储暴跌 - 36氪 (m.36kr.com/p/3878176206516225)

[15] The Nine-Month ASIC: How Jalapeño Rewrites the... - Silicon Analysts (siliconanalysts.com/analysis/custom-ai-...)

[16] 中国芯，杀破英伟达！ - 搜狐 (sohu.com/a/1088668304_121118...)

[17] 华为昇腾910C深度解析：规格、部署与性能全览 (mirrorfrog.com/blog/ascend-910c-sp...)

[18] 阿里真武810E芯片性能超英伟达A800 - U深搜 - UniFuncs (unifuncs.com/s/mEkpV2E9)

[19] GitHub - tile-ai/tilelang-ascend (github.com/tile-ai/tilelang-as...)

[20] 国家超算中心免费算力海光深算三号BW1000（即异构加速卡BW (blog.51cto.com/u_16120231/14561096)

[21] 昇腾部署deepseek教程 - 兼一书虫 (study.hycbook.com/article/54084)

[22] 爆火的DeepSeek一体机，更多满足的是情绪价值 - 甲子光年 (jazzyear.com/article_info.html)

[23] Cambricon 寒武纪思元590 (国产AI 训练/推理) | AI 算力卡百科 (mirrorfrog.com/docs/cards/others/c...)

[24] 腾讯的“AI芯事” - 时间线- 搜狐 (timeline.sohu.com/news/IUphs6XFsd)

[25] 國海證券：AI ASIC推理場景優勢顯著與GPU共築算力新格局 (hk.investing.com/news/stock-market-n...)

[26] 海光深算三号BW1000 | AI 算力卡百科| 233 款AI 芯片规格对比 (mirrorfrog.com/docs/cards/others/h...)

[27] 寒武纪MLU590 (思元590) | AI 算力卡百科| 233 款AI 芯片规格对比 (mirrorfrog.com/docs/cards/others/c...)

[28] 华为昇腾DeepSeek一体机深度拆解- 53AI-AI知识库 (53ai.com/news/zhinengyingjia...)

[29] 2026年5月2日：米乐m6苹果app最新版揭示新一代AI芯片技术路线图 (eoffcn.com/wiki/articles/20261...)

[30] PPU简介 - 阿里云文档 (help.aliyun.com/zh/document_detail/...)

[31] 真武810E - 平头哥半导体 (t-head.cn/product)

[32] Hygon 海光DCU K100 AI 版(2024) | AI 算力卡百科 - MirrorFrog (mirrorfrog.com/docs/cards/others/h...)

[33] 海光信息：存储芯片涨价反映AI服务器需求爆发深算三号产品进展顺利 (cls.cn/detail/2202869)

[34] 产品规格书 - 摩尔线程文档中心 (docs.mthreads.com/s4000/s4000-doc-onl...)

[35] 摩尔线程首个千卡智算中心落地，从算力到生态全面加速大模型发展 (mthreads.com/news/146)

[36] 完美运行DeepSeek-R1 671B：摩尔线程MTT S4000通过中国信通院 (finance.sina.com.cn/tech/roll/2025-05-0...)

[37] 华为昇腾+DeepSeek：国产AI推理引擎的破局之战 - DOIT (doit.com.cn/p/527695.html)

[38] DeepSeek：助力中国芯突围 - Supplyframe 四方维 (cn.supplyframe.com/article/8309.html)

[39] TileKernels in the DeepSeek-V4 Era - Manuel Blog (manumartinm.dev/blog/tile-kernels-d...)

[40] GitHub - deepseek-ai/TileKernels: A kernel library written in tilelang (github.com/deepseek-ai/TileKer...)

[41] DeepSeek Huawei AI Tools: TileLang and the Ascend Ecosystem (suyashpachauri.com/post/deepseek-huawe...)

[42] DeepSeek - GitHub (github.com/deepseek-ai)

[43] tile-ai/tilelang-mlir-ascend - GitHub (github.com/tile-ai/tilelang-ml...)

[44] DeepSeek开源周Day 3：DeepGEMM——300行代码性能飙升2.7倍 (pingwest.com/a/302626)

[45] 回顾DeepSeek“开源周”：越是开源，越能扩大生态 - 36氪 (m.36kr.com/p/3186220751724681)

[46] fix a typo of tilelang ref (#11) - Hugging Face (huggingface.co/deepseek-ai/DeepSee...)

[47] DeepSeek“开源周”收官，连续五天到底都发布了什么？ - 证券时报 (stcn.com/article/detail/1553...)

[48] DeepSeek 开源周（三）：DeepGEMM 性能实测及NV 往事- 文章 (developer.volcengine.com/articles/7477422974...)

[49] tilelang/examples/deepseek_mla/README.md at main - GitHub (github.com/tile-ai/tilelang/bl...)

[50] Write High Performance FlashMLA with TileLang on Hopper (tilelang.com/deeplearning_operat...)

[51] What is DeepSeek TileKernels? - Medium (medium.com/data-science-in-you...)

[52] tilelang · PyPI (pypi.org/project/tilelang)

[53] GitHub - tile-ai/tilelang: Domain-specific language designed to (github.com/tile-ai/tilelang)

[54] NPU算子"智能编译"：TileLang Developer模式来了 - QQ News - 腾讯 (news.qq.com/rain/a/20251230A06X...)

[55] DOSERANK: MEASURED RANK ALLOCATION FOR TRAINING (openreview.net/pdf)

[56] TransMLA: Migrating GQA Models to MLA with Full DeepSeek (proceedings.neurips.cc/paper_files/paper/2...)

[57] MLAfication: Efficiently Enabling High-Fidelity MLA for Modern LLMs (arxiv.org/html/2610.17202)

[58] TransMLA: Multi-Head Latent Attention Is All You Need - arXiv (arxiv.org/pdf/2502.07864)

[59] TransMLA: MLA is ALL You Need (cocoxu.github.io/CS8803-LLM-spring20...)

[60] GQLA: Group-Query Latent Attention for Hardware-Adaptive ... - arXiv (arxiv.org/pdf/2605.15250)

[61] TransMLA: Migrating GQA Models to MLA with Full DeepSeek (openreview.net/forum)

[62] TransMLA: Migrating GQA Models to MLA with Full DeepSeek (neurips.cc/virtual/2025/poster...)

[63] DeepSeek 今年前七个月营收约4.75 亿元 - 开源中国 (oschina.net/news/502134)

[64] DeepSeek Revenue Hits $1B Run Rate, Eyes $7.5B Shanghai Raise (aiweekly.co/alerts/deepseek-rev...)

[65] Everyone keeps yelling "AI bubble just like dotcom/housing" but zero (reddit.com/r/stocks/comments/1...)

[66] 【半导体】长鑫科技近期4F2技术进展的影响——野村 - QQ News (news.qq.com/rain/a/20261015A076...)

[67] 存储格局生变！长江存储eSSD出货狂飙：两年暴增8倍 - 新浪财经 (finance.sina.com.cn/tech/roll/2026-10-1...)

[68] 存储大战升级！长鑫刚宣布年底推4F2 DDR5 三星后脚确认 ... - 凤凰网 (feng.ifeng.com/c/8xFGkufkf59)

[69] "比三星早两年"中国挑战DRAM……长鑫存储年内推出下一代存储器- MK (mk.co.kr/cn/business/12173993)

[70] 長鑫4F2架構是什麼？DDR5技術突破，南亞科、華邦電受衝擊？ (bnext.com.tw/article/92623/cxmt-...)

[71] [新聞] 長鑫科技新突破！年底推4F2 架構DDR5伺服- 看板Stock (ptt.cc/bbs/Stock/M.1791812...)

[72] 存储大战升级！长鑫刚宣布年底推4F2 DDR5 三星后脚确认 ... - 凤凰网 (feng.ifeng.com/c/8xFI2cynWUR)

[73] 長鑫存儲預告年底推4F² DDR5 伺服器記憶體 或較三星早兩年 (unwire.hk/2026/10/11/cxmt-4f2...)

[74] DDR5、LPDDR6之后下一站国产HBM内存2年就能自给自足 - 快科技 (news.mydrivers.com/1/1147/1147673.htm)

[75] 长鑫科技上市史，在巨头裂缝里照进光 (tmtpost.com/8092230.html)

[76] 疾速14200MB/s！长江存储PE501 61.44TB图赏 - 新浪财经 (finance.sina.com.cn/tech/roll/2026-04-0...)

[77] 持久不衰的PCIe 5.0硬实力！长江存储PE511企业级固态硬盘首测 (news.qq.com/rain/a/20260814A0BL...)

[78] NVIDIA Q2 revenue jumps to $96.2B on AI demand - Stock Titan (stocktitan.net/sec-filings/NVDA/10...)

[79] Six months ago we called Microsoft overvalued at $490 and got (reddit.com/r/ValueInvesting/co...)

[80] Second Quarter, 2026 - Chevy Chase Trust (chevychasetrust.com/second-quarter-2026)

[81] NVIDIA Leases a Texas Data Center for $50 Billion, Hyperscalers (agentsflare.com/blog/-500--2026--72...)

[82] 800V高压直流供电来了！英伟达量产在即，AI数据中心将迎来巨变 (post.m.smzdm.com/p/ad742ork)

[83] Q&A with Investment Strategy - Raymond James (raymondjames.com/-/media/rj/dotcom/f...)

[84] Nvidia AI Spending Boom 2026 | Tech Stock Trading Guide (tradersuite.online/blog/nvidia-65-bill...)

五、全球科技巨头定制ASIC方案与落地核心信源

[01] What Is an AI ASIC? The Complete Guide - HashrateIndex (hashrateindex.com/blog/what-is-an-ai-...)

[02] The Rise of AI ASICs as GPU Alternatives for Deep Learning - Reddit (reddit.com/r/AIProgrammingHard...)

[03] 2025 ASIC AI Server Market Outlook | PDF | Broker Dealer - Scribd (scribd.com/document/886944415/...)

[04] NVIDIA AI GPU Market Share 2026: ~80% of AI Accelerators (siliconanalysts.com/analysis/nvidia-ai-...)

[05] Hot Chips 2026: OpenAI's Jalapeño AI ASIC unpacked (tomshardware.com/tech-industry/artif...)

[06] OpenAI's Jalapeño: What a Custom AI Inference Chip Actually (pinggy.io/blog/openai_jalapen...)

[07] OpenAI's 2026 Hardware Strategy Overview | PDF - Scribd (scribd.com/document/984277478/...)

[08] Morgan Stanley: U.S. data center power shortfall reaches 34% (news.futunn.com/en/post/1000580509/...)

[09] Nvidia to Google TPU Migration 2025: The $6.32B Inference Cost (ainewshub.org/post/nvidia-vs-goog...)

[10] Google TPU: The AI Chip for the AI Inference Era - NADDOD Blog (naddod.com/ai-insights/google-...)

[11] OpenAI Jalapeño AI Chip: Complete Guide 2026 (theainavigatorhub.com/2026/06/openai-jala...)

[12] Google TPU Chip Ironwood Technology Explained (nevsemi.com/blog/google-tpu-chi...)

[13] AI Data Center Value Chain: Every Layer from Chips to Cloud (2026) (siliconanalysts.com/research/ai-data-ce...)

[14] AI Chip Market: where's the money now? (newmarketpitch.com/blogs/news/ai-chip-...)

[15] Broadcom set to retain leadership position as AI server compute (seekingalpha.com/news/4542867-broadc...)

[16] AI Server Compute ASIC Shipments to Triple by 2027 as Custom (counterpointresearch.com/en/insights/AI-Serv...)

[17] OpenAI unveils its first custom chip, built by Broadcom - Hacker News (news.ycombinator.com/item)

[18] Meta Broadcom AI Chip Deal 2029: 35B MTIA Extension - Tech Insider (tech-insider.org/meta-broadcom-ai-ch...)

[19] OpenAI and Broadcom announce strategic collaboration to deploy (openai.com/index/openai-and-br...)

[20] AI Server Compute ASIC Shipments to Triple by 2027 as Custom (counterpointresearch.com/cn/insights/AI-Serv...)

[21] Hyperscaler Custom Silicon Tracker 2026 - Presenc AI (presenc.ai/research/hyperscale...)

[22] Gemini 2.5 Inference Infrastructure Explained for Bunnies - note (note.com/taku_sid/n/nefe4d49...)

[23] Amazon's Latest AI Data Center: Suppliers and Builders Complete (developer.tenten.co/amazons-latest-ai-d...)

[24] AI & Machine Learning Semiconductors — Sector Supply Chain (semiconductorx.com/sector-ai.html)

[25] Microsoft Unveils Maia 300 AI Chip to Break Nvidia's Stranglehold (innotechinsider.com/ai/microsoft-maia-3...)

[26] 阿里云自研AI 芯片主要包括平头哥PPU和含光800两大产品线 (hansenfluid.com/news/hangguang_800....)

[27] 阿里或拆分旗下AI芯片企业平头哥独立上市 - 财联社 (cls.cn/detail/2267016)

[28] 阿里平头哥启动上市计划，已布局全栈AI芯片 - 36氪 (m.36kr.com/p/3650412256731265)

[29] 腾讯云发布紫霄2.0芯片混元大模型训练性能提升50% - 淘AI工具 (bot.taobao.com/tool-today_discover...)

[30] 腾讯云2026 年7 月：AGI 时代的全栈AI 基础设施 (developer.cloud.tencent.com/article/2712907)

[31] 腾讯紫霄(Zixiao) | AI 算力卡百科| 233 款AI 芯片规格对比 - MirrorFrog (mirrorfrog.com/docs/cards/others/t...)

[32] 字节跳动联合博通开发专用芯片？官方回应来了 - 第一财经 (yicai.com/news/102163385.html)

[33] 豆包收费，能赚多少钱？ - 36氪 (m.36kr.com/p/3862321545902729)

[34] 豆包收费，能赚多少钱？ - QQ News (news.qq.com/rain/a/20260621A07U...)

[35] 与台积电合作生产AI芯片？字节：报道不实|台积电|处理器|字节|平均售 (163.com/dy/article/JCCSHKUN...)

[36] gemini-samples/examples/interactions-deep-research-getting (github.com/philschmid/gemini-s...)

[37] TrainSpotting: Improving Inference on AWS Trainium2 through End (makora.com/blog/trainium2-opt)

[38] Axon: A Synthesizing Superoptimizer for Tensor Programs - arXiv (arxiv.org/html/2606.26344v1)

[39] AI Accelerator - AWS Trainium (aws.amazon.com/ai/machine-learning...)

[40] Inside Maia 100: Revolutionizing AI Workloads with Microsoft's (techcommunity.microsoft.com/blog/azureinfrastru...)

[41] Microsoft MAIA 100 AI Accelerator for Azure - ServeTheHome (servethehome.com/microsoft-maia-100-...)

[42] Microsoft Unveils New Details on Maia 100, Its First Custom AI Chip (techpowerup.com/326105/microsoft-un...)

[43] TPU vs GPU in 2026 — the real Ironwood versus Blackwell face-off (labo-llm.fr/en/infrastructure/t...)

[44] OpenAI details Jalapeño AI chip, with 700W TDP - DCD (datacenterdynamics.com/en/news/openai-deta...)

[45] 阿里神秘芯片对标英伟达？业内人士：或是含光800下一代产品 (finance.sina.com.cn/jjxw/2025-09-19/doc...)

[46] Jan Gray - FPGA CPU News (fpga.org/author/jan)

[47] OpenAI Jalapeno Custom AI ASIC at Hot Chips 2026 - ServeTheHome (servethehome.com/openai-jalapeno-asi...)

[48] AWS-Trainium2-Instances-Now-Generally-Available (press.aboutamazon.com/2024/12/aws-trainiu...)

[49] AI Chip Architectures - Jacob Peake (jepeake.com/ai-chip-architectures)

[50] AI Chip Manufacturers 2026: GPUs vs Custom ASICs - TechDirectory (techdirectory.sg/insights/ai-chip-ma...)

[51] AWS' custom chip strategy is showing results, and cutting into (reddit.com/r/hardware/comments...)

[52] Agentic Operator Generation for ML ASICs - arXiv (arxiv.org/html/2512.10977v1)

[53] Hot Chips 2024 - (hc2024.hotchips.org)

[54] Tesla Dojo - Wikipedia (en.wikipedia.org/wiki/Tesla_Dojo)

[55] Tesla Dojo Chip Wiki - SemiWiki (semiwiki.com/wikis/semiconductor...)

[56] AWS unveils next-gen Trainium3 custom AI chips and cloud (siliconangle.com/2024/12/03/aws-unve...)

[57] OpenAI and Broadcom announce strategic collaboration to deploy (investors.broadcom.com/news-releases/news-...)

[58] OpenAI and Broadcom unveil LLM-optimized inference chip (openai.com/index/openai-broadc...)

[59] Jalapeño's first results show industry-leading speed and efficiency in (openai.com/index/jalapeno-firs...)

[60] OpenAI diversifies AI chip supply chain with Nvidia, Broadcom and (facebook.com/rcrwireless/posts/o...)

[61] Microsoft Maia 200 Beats Blackwell's Efficiency - XPU.pub (xpu.pub/2026/01/30/microsof...)

[62] The Terafab: Why One Chip Needs Two Factories? The "shocking (facebook.com/hardworkercraft/pos...)

[63] Hardware Strikes Back in the AI Era | Bain & Company (bain.com/insights/hardware-s...)

[64] 自研芯片露真容阿里“通云哥”角逐全球AI赛场 - 证券时报 (stcn.com/article/detail/3621...)

[65] 腾讯自研推理芯片紫霄已落地腾讯会议 - 深圳市人工智能产业协会 (szaicx.com/gd/16602.html)

[66] 信创GPU显卡算力对比 - CSDN博客 (blog.csdn.net/qq_14845119/article...)

[67] 率先全面开放，文心一言为何一骑绝尘？ - 新浪财经 (finance.sina.com.cn/tech/2023-09-01/doc...)

[68] 中信建投-人工智能行业：AI新纪元，砥砺开疆·智火燎原 (hibor.com.cn/data/bdc9612a283861...)

[69] 传字节跳动两款自研AI芯片2026年前量产，将由台积电代工 - Moomoo (moomoo.com/hans/news/post/4365...)

[70] 颠覆传统依赖？阿里、腾讯亮出抖音企业号官网入口，英伟达回应了 (eoffcn.com/wiki/document/vW2VU...)

[71] AWS Trainium Runs Real-Time AI Video Generation Without NVIDIA (techtimes.com/articles/328113/202...)

[72] TPU Developer Hub | Google Cloud (cloud.google.com/products/tpu/tpu-de...)

[73] Patrick C. Toulme — AI Compilers & Systems (patricktoulme.com)

[74] What about TVM, XLA, and AI compilers? (Democratizing ... - Modular (modular.com/blog/democratizing-...)

[75] vLLM TPU: A new unified-backend supporting Pytorch and JAX (youtube.com/watch)

ABOUT YAVIN 质观

Yavin质观

硬科技与资本叙事研究者，长期关注技术、产业、政策与资本化逻辑，
书写技术时代、现代生活与精神处境中的缓慢观察。

微信扫码关注「Yavin质观」公众号

获取更多硬科技、产业、政策与资本化逻辑深度研究报告

研究声明与免责条款：本报告由「Yavin质观」独立研究编制，报告所引用的财务数据、市场出货量、工程技术参数及供应链信息均来源于公开财报、招股说明书、第三方研究机构（IDC、TrendForce、Counterpoint、SemiAnalysis等）、学术论文及行业公开信息。针对非上市企业或未公开披露之产线良率、成本模型与情景推演，均属于研究假设与工程估算，仅供参考，在任何情况下均不构成对任何证券、芯片产品或投资行为的建议。未经许可，不得用于商业篡改或二次售卖。

© 2026 Yavin质观 · 保留所有权利