ISSUE 05 DOMESTIC COMPUTE & OPTICAL SCALE-UP

05

跨越硅基鸿沟：中国 AI 算力底座突围

三维逻辑折叠、带宽屋顶线对齐与全光超节点的系统工程重构



00 / EXECUTIVE SUMMARY · 核心研究结论与关键指标总览

四维工程协同：平面微缩受限条件下，以三维逻辑折叠、屋顶线对齐、全光超节点与西部绿电重构算力总账

70%+ 推理负载占比
算力评估标尺从纸面浮点峰值转向单位人民币有效 Token 产出

1.5μm 混合键合
海思韬定律通过双层逻辑折叠获取约一代半等效平面制程收益

MLA + TileLang 协同
KV 缓存压缩 93.3% 配合算子自动编译与 CANN 核心接口开源

平头哥分拆 + 四强盈利
真武累计交付逾 47 万片拟赴港上市，科创板四强前三季度全面盈利

01 出货结构双拐点

2026 年上半年中国数据中心加速卡出货 200.8 万颗，同比增长 15%；其中非 GPU 架构出货 101.6 万颗，以 50.6% 份额首次超过通用 GPU，本土芯片阵营合计占比达 66%。随着生产集群推理消耗占比超过 70%，海外每月 200 美元重度订阅暴露出最高 70 倍的算力成本倒挂，自主低成本算力底座成为模型商业化的盈利前提。

02 韬定律与三维折叠

华为海思在 IEEE ISCAS 2026 提出韬定律，并在量产旗舰麒麟 2026 上落地 1.5 微米晶圆级无凸块双层逻辑折叠，使等效晶体管密度提升 53.5% 至 238 MTr/mm²，片上网络布线面积缩减 55%，同频下 NPU、GPU 与 CPU 功耗分别下降 66%、58% 与 41%；本土 12nm 大显存方案同步切入低并发私有化解码市场。

03 屋顶线对齐与超节点

昇腾 910C 在万卡稠密预训练上较海外前沿芯片仍存动态调度与集合通信折损，但依托单卡 96 至 128GB 显存与 3.2 TB/s 带宽，在 671B 混合专家模型推理中实现 H100 约 65% 至 75% 的单卡有效吞吐；灵衢全光超节点配合西部每度 0.28 至 0.35 元绿电，将万卡集群三年期百万 Token 摊销成本压低逾 42%。

04 三季报穿透与分拆

2026 年 11 月 10 日阿里巴巴公告拟分拆平头哥赴港独立上市，真武系列累计交付突破 47 万片；海光信息 149.55 亿元、寒武纪 95.26 亿元、摩尔线程 32.89 亿元、沐曦股份 28.24 亿元的 2026 年前三季度法定营收合计达 305.94 亿元，归母净利润合计 72.39 亿元，并以 390.85 亿元存货与预付款锁定上游先进晶圆及封测产能。

个人公开研究免费分享（非商业用途） | https://zhiguanstudio.cn

00 / TABLE OF CONTENTS · 报告全局研究框架与视觉资产导航

五章研究框架：从 200.8 万颗市场结构拐点延伸至三维逻辑折叠、屋顶线协同、全光超节点与三季报核验

CHAPTER 01 DOUBLE INFLECTION & TOKEN ECONOMICS

出货结构双拐点与推理时代的算力成本约束

- 1.1 2026H1 200.8 万颗出货拆解：66% 本土份额与 50.6% 非 GPU 架构占比
- 1.2 70% 推理负载拐点与每月 200 美元重度订阅背后的成本倒挂

CHAPTER 02 TAU-LAW & KIRIN 2026 3D LOGIC FOLDING

平面微缩受限下的三维演进：海思韬定律与麒麟 2026 逻辑折叠

- 2.1 铜互连电子散射瓶颈与 IEEE ISCAS 2026 韬定律空间微缩机制
- 2.2 麒麟 2026 双层折叠实测拆解与 12nm 大显存私有化推理定位

CHAPTER 03 CROSSING THE ROOFLINE: ASCEND × DEEPSEEK

屋顶线协同：昇腾与开源模型算法的双向对齐

- 3.1 昇腾 910C 稠密预训练折损成因与 671B 混合专家推理甜区验证
- 3.2 MLA 矩阵吸收变换、TileLang 自动编译与 CANN 核心开源

CHAPTER 04 UNIFIEDBUS OPTICAL SUPERPOD & GREEN POWER

系统级光电互连：灵衢全光超节点与西部绿电协同总账

- 4.1 从 384 卡到数千卡全光超节点：空间功耗代价与三年期推理降本
- 4.2 北美变压器交付瓶颈对比中国特高压西部绿电智算底座

CHAPTER 05 COMMERCIALIZATION & Q3 2026 STATUTORY AUDIT

商业闭环与三季报穿透：平头哥分拆与 390.85 亿元备货锁产

- 5.1 运营商国产化集采常态化与阿里平头哥拟分拆赴港独立上市
- 5.2 科创板算力四强 2026 年前三季度法定财报对账与后续核验指标

VISUAL & QUANTITATIVE ASSET INDEX · 全报告图表与量化矩阵清单

13 HIGH-RES ARCHITECTURAL EXHIBITS, ROOFLINE CURVES & Q3 STATUTORY TABLES

- 图01–图03：2026H1 200.8 万颗出货结构、训推负载迁移趋势、重度订阅算力成本倒挂
- 图04–图07：麒麟 2026 双层折叠实测指标、子系统功耗拆解、主频演进路线与 12nm 屋顶线映射
- 图08–图10：中美主力芯片混合专家模型推理实测、昇腾 910C 大显存区间与 TileLang 编译机制
- 图11–图13：全光超节点双向工程核算、四维物理工程底座与运营商智算集采分配格局

个人公开研究免费分享（非商业用途） | https://zhiguanstudio.cn

01 CHAPTER 01 · MARKET STRUCTURE INFLECTION & TOKEN ECONOMICS

出货结构双拐点与推理时代的算力成本约束

1.1 2026H1 出货结构双拐点：66% 本土份额与 50.6% 非 GPU 架构占比

2026 年上半年，中国数据中心加速芯片市场跨过供给结构分水岭。IDC 统计显示，国内数据中心加速卡半年总出货达 200.8 万颗，同比增长 15%。温和的总量扩张背后，底层硬件架构首次反转：以 NPU 与定制 ASIC 为代表的非 GPU 加速卡出货达 101.6 万颗，同比增长 115%，份额升至 50.6%，首次超过同期回落 23% 至 99.1 万颗的通用 GPU。对比 2024 年非 GPU 架构仅占 21%、2025 年升至 34% 的轨迹，国内算力底座已在两年周期内转向专用超节点与通用兼容架构双轨并行。

国别供给结构同步完成迁移。受先进制程与高带宽显存出口限制收紧、特供版芯片合规边界反复变动及本土软件栈趋于成熟推动，2026 年上半年本土芯片厂商合计取得约 66% 出货份额。其中华为昇腾以约 28% 占比居首；阿里平头哥与百度昆仑芯等云厂商自研 ASIC 在内部搜索推荐与大模型推理池规模放量，合计占据逾一成份额；寒武纪、海光信息及摩尔线程、沐曦股份、燧原科技等独立商用芯片企业则在运营商与商业智算中心形成稳定的第三极供给。与训练集群要求全网同构同步不同，线上推理流量可按业务等级拆分至独立的国产芯片池运行，客观上加速了异构替代的落地节奏。

EXHIBIT 01 **2026H1中国AI加速芯片市场200.8万颗出货结构与非GPU架构占比演进**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

促成这一双重拐点兑现的核心力量，并非单纯的行政采购替代，而是产业算力重心从不计单卡折旧的稠密预训练，全面转向计较每百万 Token 电费与折旧摊销的长上下文生产推理。

表 1：2024—2026E 中国数据中心 AI 加速芯片架构裂变与本土份额跃迁演进矩阵（数据口径：IDC 数据中心加速卡出货跟踪及公开招投标统计）

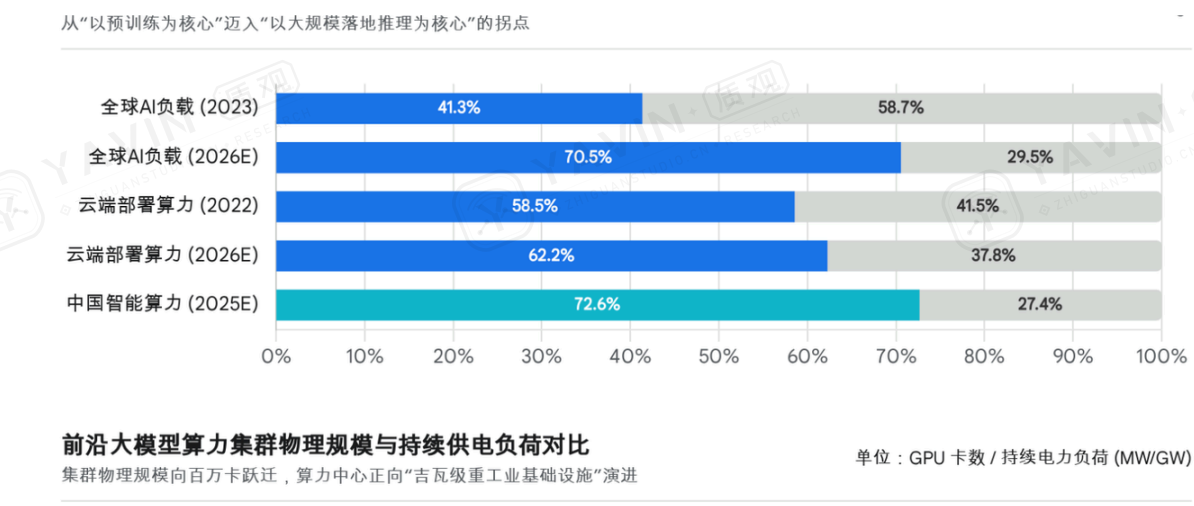
演进阶段与统计窗口	全市场加速卡规模	非 GPU 架构 (NPU/ASIC) 占比	通用 GPU 架构占比	本土算力厂商合计份额	核心工作负载特征与底层驱动逻辑
2024 全年	~275 万颗	~21.0%	~79.0%	~30%–34%	百模大战稠密预训练高峰期，CUDA 动态调试生态具备绝对排他性
2025 全年	~390 万颗	~34.0%	~66.0%	~45%–52%	DeepSeek-R1/V3 开源引爆 MoE 推理潮，昇腾 910B/910C 与云厂自研卡规模导入
2026H1（实绩）	**200.8 万颗 (+15%)**	**50.6% (101.6万颗, +115%)**	**49.4% (99.1万颗, -23%)**	**~66.0%**	**推理算力负载突破 70%+**；静态计算图固化与超节点光互连大幅降低非 GPU 部署摩擦
2026E 全年展望	~440–470 万颗	~54%–58%	~42%–46%	~70%–75%	昇腾 950 双构芯片、平头哥真武迭代与独立算力四强 390.85 亿锁料集中交付

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

1.2 负载重心迁移：推理占比突破 70% 后的单位 Token 成本核算

2022 至 2024 年基础模型探索期，算子形态频繁迭代，买方最看重万卡集群的动态图调试效率、集合通信成熟度与混合精度峰值算力，CUDA 生态由此具备显著排他性。进入 2026 年，随着多步思维链推演与智能体长程工具调用普及，中美头部云厂与互联网平台的生产集群中推理算力消耗占比双双超过 70%，采购评估标尺随之由纸面浮点峰值切换为单位货币的全生命周期有效 Token 产出。

EXHIBIT 02 **全球与中国AI算力负载结构从训练向推理迁移趋势（2022–2026E）**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

当生产集群七成以上算力转向线上推理，硬件选型的物理与财务约束随之重构。首先，承载亿级日活调用的主力稀疏混合专家模型架构在数月周期内保持稳定，通用 GPU 为动态调试预留的硬件线程调度与缓存一致性电路不再提供边际收益。工程团队有充足动力耗时数周针对目标 NPU 或 ASIC 完成静态图下沉与流水线手工排布，换取未来三年满载运行下 30% 至 50% 的单位 Token 成本降幅。

其次，自回归解码瓶颈从计算核心峰值移向显存容量、访存带宽与片间延迟。多步智能体单次任务往往触发数十轮连续调用，每生成单个 Token 均需遍历读取激活专家权重与长上下文 KV Cache，算术强度跌至每字节仅数次浮点运算的访存受限区，决定吞吐上限的是单机聚合 HBM 容量、有效带宽及跨卡专家并行通信延迟。

更直接的约束体现在商业毛利表上。对 2026 年海外主流每月 200 美元个人顶配订阅的后台账本核算显示，重度编程与多智能体并发用户单月消耗的真实推理成本高达 2,500 至 14,000 美元，相对固定月费形成 12 至 70 倍的成本倒挂。缺乏自有低成本算力、被迫向外部高价租用通用 GPU 的纯模型机构面临调用越活跃现金流消耗越快的财务压力；而具备全栈自有算力底座的生态体可将百万 Token 边际成本压低 50% 至 70%，在长期接口定价竞争中建立成本缓冲垫。

EXHIBIT 03 每月200美元重度订阅算力成本倒挂与自有算力底座成本差距测算



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

02 CHAPTER 02 · HISILICON TAU-LAW & KIRIN 2026 3D LOGIC FOLDING

平面微缩受限下的三维演进：海思韬定律与麒麟 2026 逻辑折叠

2.1 铜互连电子散射瓶颈与海思 ISCAS 2026 韬定律空间微缩机制

本土算力芯片面临的首要物理约束是极紫外光刻设备受限下的先进平面微缩阻断。但回到半导体器件物理审视，即便在前沿代工节点，传统二维平面微缩进入 3nm 至 2nm 后同样面临双重刚性阻力：一是多重曝光显著推高掩模层数与良率损耗，单次 2nm 流片研发总成本已攀升至 7.25 亿至 10 亿美元；二是当 40nm 以下底层铜互连线宽缩减至接近约 39nm 的电子平均自由程时，晶界散射与表面散射引起导线电阻率急剧上升，芯片内部逾 60% 的动态功耗消耗在跨越数毫米平面距离搬运信号的金属导线电容充放电发热上，而非晶体管本身的逻辑翻转。

2026 年 5 月 25 日，华为海思负责人何庭波在 IEEE 电路与系统旗舰会议 ISCAS 2026 发表主旨报告，正式提出面向后摩尔时代的韬定律：当二维平面线宽微缩受阻或边际经济收益递减时，通过晶圆级无凸块三维逻辑折叠向垂直 Z 轴索取集成密度与互连效率，使芯片系统的等效功能晶体管密度与综合能效比每两至三年继续保持跨越一代平面制程的演进斜率。

EQUATION · CORE QUANTITATIVE & ECONOMIC MECHANISM

$$\rho_{\text{eff,3D}} = \sum_{k=1}^{K} \rho_{\text{2D}}^{(k)} \cdot \left(1 - \eta_{\text{HB overhead}}\right)$$

$$L_{\text{wire,3D}} \approx \frac{L_{\text{wire,2D}}}{\sqrt{K}} \Rightarrow P_{\text{interconnect}} \propto C_{\text{wire}} V^2 f \downarrow 40\% \sim 66\%$$

物理机制释义：双层逻辑折叠在扣除微小的混合键合界面开销后使单位面积有效晶体管密度近乎翻倍，并将跨模块导线长度按层数平方根缩短，直接削减四至六成以上的互连动态功耗。

2.2 麒麟 2026 量产拆解：1.5 微米混合键合与子系统降耗机制

不同于仅将静态缓存叠放于逻辑层之上的局部封装，海思在 2026 年量产芯片麒麟 2026 上首次规模商用逻辑对逻辑双层三维折叠架构。ISCAS 2026 披露实测与拆解显示，麒麟 2026 对比前代平面旗舰麒麟 9030 Pro 实现了显著的密度与能效改善。

EXHIBIT 04 麒麟2026双层逻辑折叠核心密度与能效实测指标

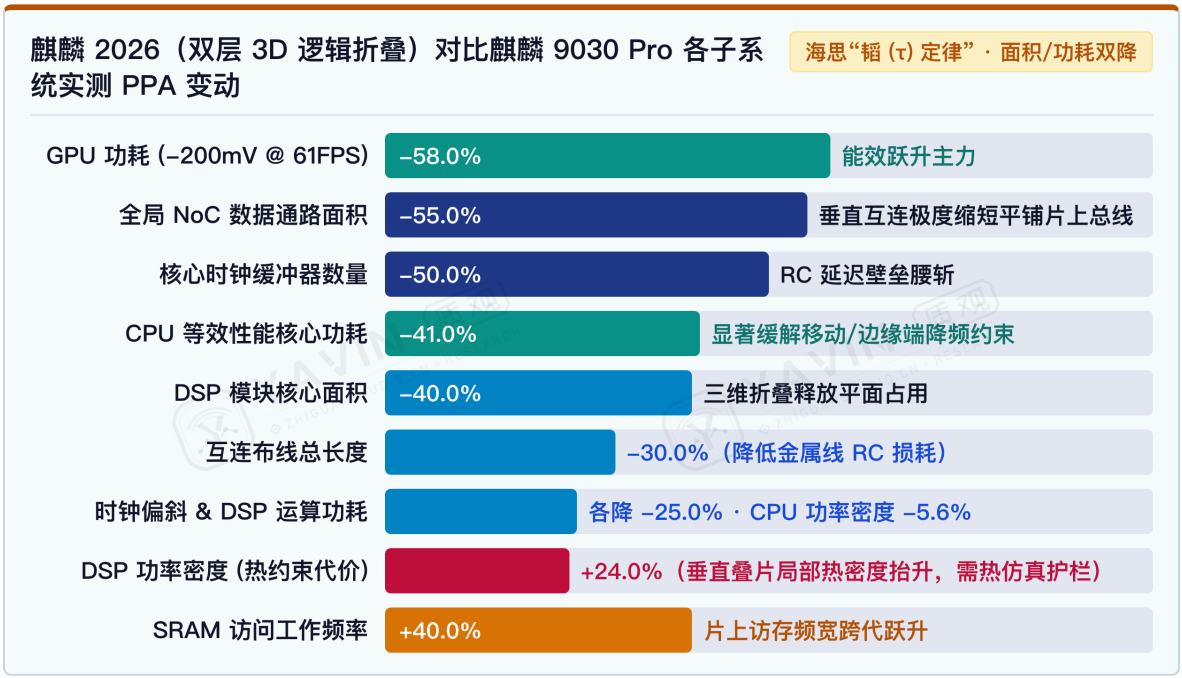


数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

在界面工艺上，麒麟 2026 采用面对面晶圆对晶圆无凸块铜混合键合，将垂直键合间距压至 1.5 微米，每平方毫米封装界面可提供逾 44 万个亚微米级垂直互连通道，界面寄生电容不足传统微凸块的十分之一，跨层信号传输延迟与同层金属走线基本相当。依托高密计算核、片上网络与缓存阵列沿三维空间的折叠排布，单位投影面积等效晶体管密度从前代的 155 MTr/mm² 提升 53.5% 至 238 MTr/mm²，物理集成度相当于跨越约一代半先进平面制程。

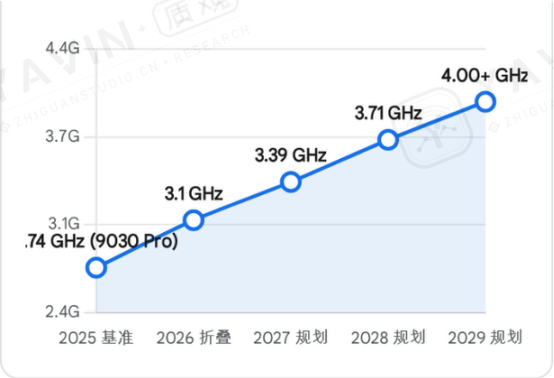
能效收益的核心机制在于裁撤平面曼哈顿长导线。将原本水平相隔数毫米的计算簇与共享缓存改为上下垂直对齐后，片上网络布线面积缩减 55%，驱动长导线的时钟树与信号中继缓冲器数量减少 50%。同频实测下，NPU、GPU 与 CPU 子系统功耗分别下降 66%、58% 与 41%。尽管双层逻辑叠放会提高局部热流密度、需要通过高频计算单元与静态缓存交错布局来平抑热点，但整体互连发热的大幅压降仍为 2027 至 2029 年迈向 4.0GHz 以上主频预留了热力学空间。

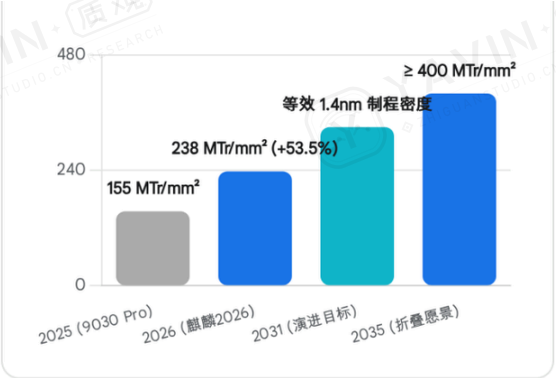
EXHIBIT 05 麒麟2026对比麒麟9030 Pro各子系统功耗与互连面积拆解



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

EXHIBIT 06 **海思韬定律下CPU主频迈向4.0GHz以上与等效1.4nm三维密度演进路线**





在固定工艺节点、25°C室温条件下，**麒麟2026**通过双层逻辑折叠将晶体管密度从麒麟9030 Pro的155 MTr/mm²拉升至238 MTr/mm²（+53.5%），并在等性能下降低41%的CPU核心功耗。需要警惕的是，DSP模块虽降低25%功耗，但因面积缩减40%导致功率密度上升24%，印证了时间余量必须转化为低压并行架构才能避免三维热失控。

Data sources: IEEE ISCAS 2026 主旨报告与芯片逆向实测数据

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

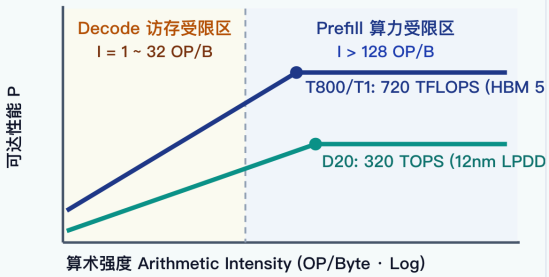
2.3 成熟制程切口：基于推理屋顶线边界的 12nm 大显存架构

在旗舰云端芯片借三维折叠向上攻坚的同时，本土 12nm 至 14nm 成熟制程也找到了契合屋顶线物理规律的商业切口。政务公文、金融内控审计与通信运维工单等企业私有化场景并发度较低，单次解码批大小多为 1 至 8，完全落在左侧访存带宽与显存容量受限斜坡区；若配置高溢价的先进制程大算力卡，八成以上矩阵浮点单元将长期处于等待显存读取的空转状态。

EXHIBIT 07 大模型推理屋顶线性能边界与成熟制程大显存异构算力映射

大模型推理 Roofline 双瓶颈区迁移与 12nm 大显存架构精准映射

I* 拐点：D20 约 156 OP/B · T800 约 144 OP/B



Decode 解码卡位：江原科技 D20 / T800（12nm 大显存池）

Batch = 1–8 低算术强度区，峰值算力 **80% 以上空转**；决定吞吐的是显存容量与带宽而非线宽。

江原科技 D20 / T800（12nm 2,048 GB 显存池）：整机严控 <2,000W 风冷免液冷改造，单机装满 671B 权重。

Prefill 预填充段：高带宽 HBM 穿透拐点

高并发长提示词迅速穿透 **I* = 144 OP/B** 拐点，必须靠高算力+高带宽芯片独立池化承接。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

以江原科技等全国产 12nm 流片方案为例，其设计放弃盲目堆叠单卡峰值浮点算力，转而围绕大容量显存控制器定制微架构，在单台整机功耗严控于 2,000W 以内、可直接推入政企既有风冷机柜免除液冷与配电改造的前提下，构建 2,048GB 单机统一显存池，单台设备即可完整装载满血版 DeepSeek-R1 671B 权重与长文档缓存。对于数据不出域的政企客户而言，这种用单流极限吐字速率换取全权重本地驻留与免拆柜交付的取舍，显著降低了私有化落地的综合资本开支。

03 CHAPTER 03 · CROSSING THE ROOFLINE: ALGORITHM-HARDWARE CO-DESIGN

屋顶线协同：昇腾与开源模型算法的双向对齐

3.1 工程实测对标：昇腾 910C 预训练折损与混合专家推理甜区

客观评估中美主力算力芯片的真实产出差距，必须摒弃单一峰值对比，将国内智算中心交付主力华为昇腾 910C 与英伟达 H100 及合规特供版 H20 置于万卡稠密预训练和稀疏混合专家生产推理两类差异显著的工作负载下分别核算。

EXHIBIT 08 **中美主力AI加速芯片硬件规格与DeepSeek-R1生产推理吞吐实测对标**

中美主力 AI 加速芯片工程规格与 DeepSeek-R1 生产推理实测对标 显存容量 × 带宽 × TDP 约束

芯片型号	显存容量	显存带宽	TDP 功耗	DeepSeek-R1 / LLM 生产实测工程效能解读
华为昇腾 910C	**128 GB**	~784 GB/s+	**310 W**	R1 Prefill 达 H100 的 **82%**，Decode 达 **81%**，单卡功耗仅为 H100 的 **44%**。
平头哥真武 810E	**96 GB**	**2,765 GB/s**	**400 W**	实测 R1 吞吐较 H20 提升 **+18%**，能效高 **30%+**（~8,500 tps 档），高带宽卡位长文解码。
海光深算三号 BW1000	64 GB	1,600 GB/s	300 W	类 CUDA 生态兼容，AI 训练性能达 A100 的 **80%—90%**，FP64 追平 H100（30 TFLOPS）。
摩尔线程 S4000	48 GB	768 GB/s	350 W	通过 DeepSeek-R1 671B 认证，MUSA 生态千卡集群训练线性度 **>91%**。
腾讯紫霄 2.0 (C200)	48 GB	未公开	320 W	云原生内生推理主力，较前代训练升 4× / 推理升 3×，部署规模年底扩至 10 万卡级。
NVIDIA H100 SXM	80 GB	3,350 GB/s	**700 W**	国际旗舰训练与推理基准（最高 ~14,200 tps 档），但单卡功耗与采购受限显著。
NVIDIA H20 (在华特供)	96 GB	4,000 GB/s	550 W	高显存带宽特供配置，但峰值算力与在华合规供货受限明显，Prefill 易成瓶颈。

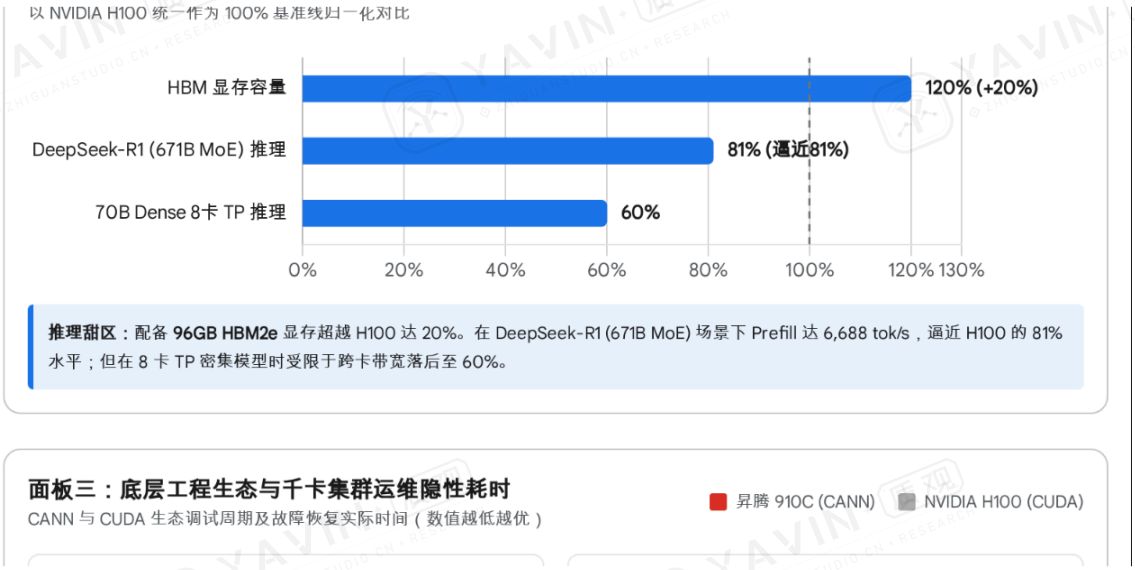
核心工程结论：国产主力加速卡凭借 **96—128 GB 大显存 + 300—400 W 低功耗区间**实现“以空间/能效换单卡算力差距”，在 Memory-Bound 的 R1 长文本推理中实测跑出国际基准 **81%—118%** 的有效能效比。

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

在万卡级前沿稠密预训练中，昇腾 910C 相比海外新一代旗舰仍存在两项工程折损：一是达芬奇矩阵核与向量核物理分离架构下，面对未经深度融合的非规则激活或动态索引算子，中间数据需经片上统一缓冲区往返搬运，容易产生流水线等待气泡；二是底层软件栈在动态图下发延迟、跨层集合通信尾延迟抑制及千卡硬件故障秒级热恢复上仍需较高专家调优人力，使其大规模稠密预训练归一化有效算力约为 H100 的 50% 至 62%。

一旦切换至 DeepSeek-R1 671B 稀疏混合专家线上推理，显存容量与算力配比随即改写了产出排序。H100 单卡显存仅 80GB，单台 8 卡节点共 640GB 无法装下满血 FP8 权重与高并发缓存，被迫跨双节点拆分并承担跨机通信损耗；而 H20 虽具大显存，但受算力大幅削减影响在长文本首字预填充阶段延迟偏高。昇腾 910C 则依托单卡 96 至 128GB HBM2e、3.2 TB/s 带宽及均衡的矩阵算力，单台 16 卡或双机超节点即可驻留全量专家权重，经算子融合后单卡推理有效吞吐达 H100 的 65% 至 75%，单位人民币 Token 产出跨过商业平价拐点。

EXHIBIT 09 昇腾910C大容量HBM配置在DeepSeek-R1推理场景下的吞吐表现

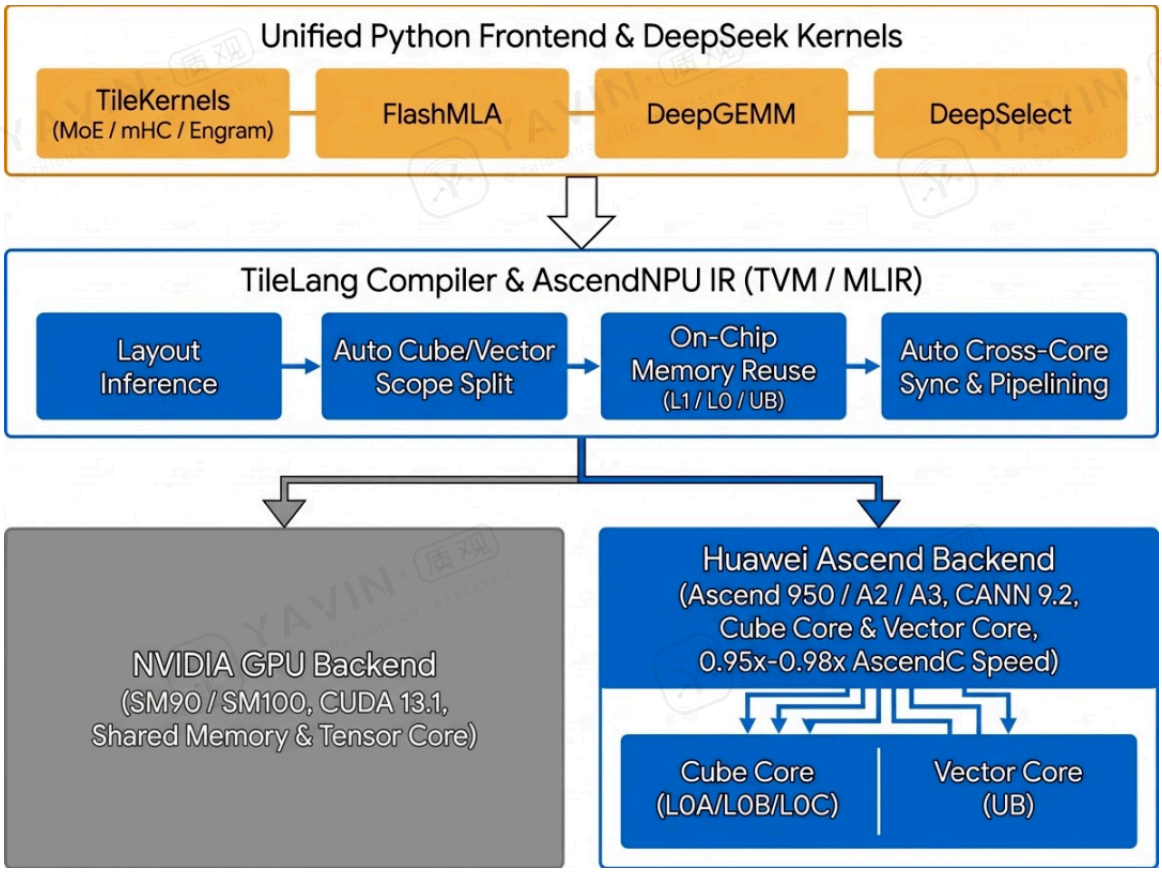


数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

3.2 算法与编译双向收敛：MLA、TileLang 与 CANN 核心开源

长期制约非 CUDA 加速芯片规模应用的瓶颈在于底层手写算子门槛过高。2025 至 2026 年间，国内开源模型社区与本土芯片团队通过算法主动迁就硬件屋顶线、编译器贴合主流 Python 开发习惯，显著压缩了跨架构迁移摩擦。

EXHIBIT 10 DeepSeek底层算子协同设计与TileLang跨架构自动编译机制



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

在算法结构层，DeepSeek 提出的多头潜在注意力 MLA 将高维键值张量低秩投影为紧凑潜在向量，砍掉 93.3% 缓存显存占用；更具硬件亲和力的是矩阵吸收恒等变换，将解压矩阵合并至查询投影侧，使数据在片外显存保持压缩态、仅在片上寄存器内展开运算，从而把原本低算术强度、易阻塞向量单元的访存密集型解码计算转化为高算术强度稠密矩阵乘，充分填满达芬奇矩阵核。同时，GLM-5 与 DeepSeek 动态稀疏注意力配合昇腾新一代芯片内置的多颗泰山控制核与硬件调度器，可在 30 微秒内完成片上稀疏块索引筛选，省去主机 CPU 往返同步。

在编译工具链层，手工编写底层指令的壁垒正被自动编译消解。北大与清华团队主导的 **TileLang** 及智源研究院牵头的 **FlagOS 2.2** 允许工程师仅用数十行高层 Python 分块代码描述算子，由编译器自动编排矩阵向量双流水线、片上缓存双缓冲及异步直接内存搬运以掩盖访存延迟，在昇腾、寒武纪、海光、摩尔线程、沐曦等十余款国产芯片上跑出手写专家代码 90% 至 105% 的实测性能。叠加华为全面开源 CANN 核心算子库并推动 **TorchNPU** 融入 PyTorch 官方 **PrivateUse1** 原生设备生态，主流开源模型首日零代码修改适配已成常态。

04 CHAPTER 04 · UNIFIEDBUS OPTICAL SUPERPOD & GREEN POWER SYNERGY

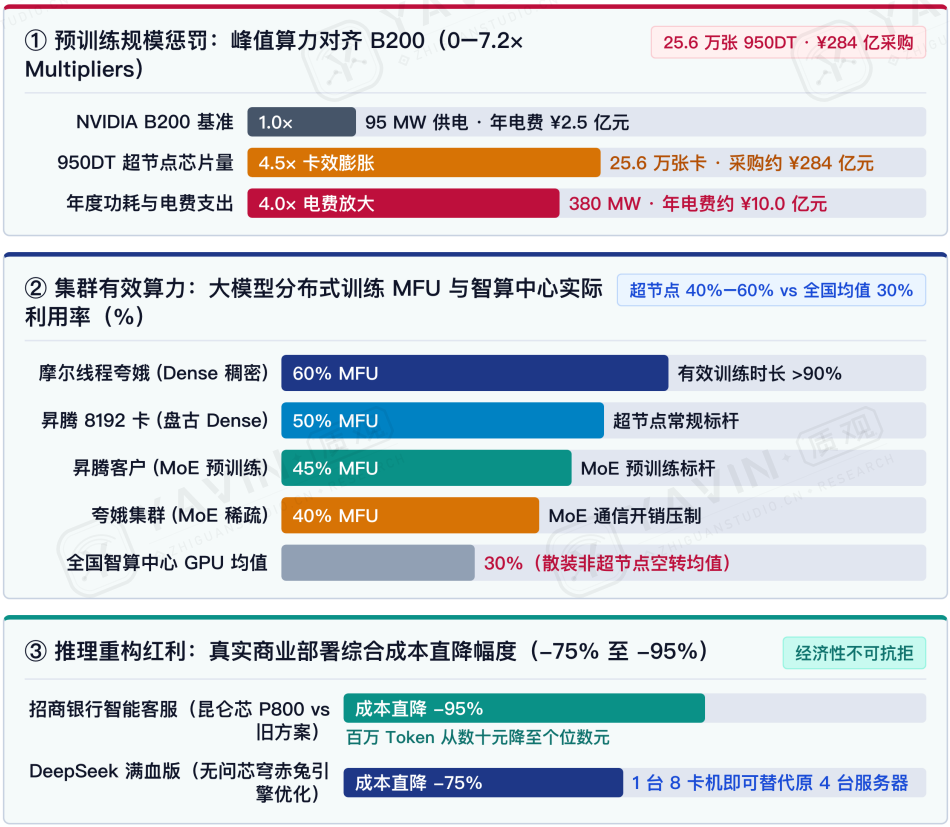
系统级光电互连：灵衢全光超节点与西部绿电协同总账

4.1 以光补硅的系统权衡：从 384 卡到数千卡全光超节点的双向工程核算

当单颗芯片受到先进光刻与前沿 HBM4 代工的外部物理制约时，系统工程给出的现实路径是放宽单机柜空间与供电包络，依托国内成熟的高速光通信与高密液冷产业链，借高带宽统一内存语义光总线将数百至数千颗 NPU

耦合为一台逻辑超节点。从商用的 384 卡全光互连 Atlas 900，演进至最高支持 8,192 卡的灵衢 2.0 集群及搭载内置光源近封装光引擎的 Atlas 960E，需对其空间功耗代价与推理降本收益做双向核算。

EXHIBIT 11 **全光超节点架构在稠密预训练中的规模代价与混合专家推理降本双向核算**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

从物理代价看，跨机柜光互连单位比特传输功耗约为 8 至 15 pJ/bit，高于英伟达 NVL72 单机柜内 2 米无源铜缆背板的 5 至 7 pJ/bit，且多机柜拆分拉长了光纤物理跳距，远端内存访问延迟处在 800 纳秒至 2 微秒区间。叠加单芯制程差距，同等纸面算力规模下全光超节点的整机总功耗与机房占地面积约为 NVL72 系统的 2.5 至 3.8 倍，在高频全局同步的稠密预训练中亦需承担额外的光跳延迟损耗；若部署于电力紧缺且机柜租金高昂的北美核心区，这一能耗溢价将构成沉重财务负担。

但在国内基础设施禀赋与混合专家推理场景下，以电换算、以光补硅能够算平总账。在万亿参数超宽专家并行推理中，数百个专家分散驻留于数百张加速卡、每卡仅承载极少数专家，全互连灵衢内存语义光总线消除了传统 PCIe 与以太网协议拆包瓶颈，使单用户解码吐字间隔稳定跑进 15 至 20 毫秒实时交互区间；按万卡智算集群三年全生命周期折旧与电费测算，国产全光超节点每百万 Token 综合摊销成本约为 4.83 至 4.92 元，较传统 8 卡 RoCEv2 以太网服务器集群的 8.50 元下降逾 42%。

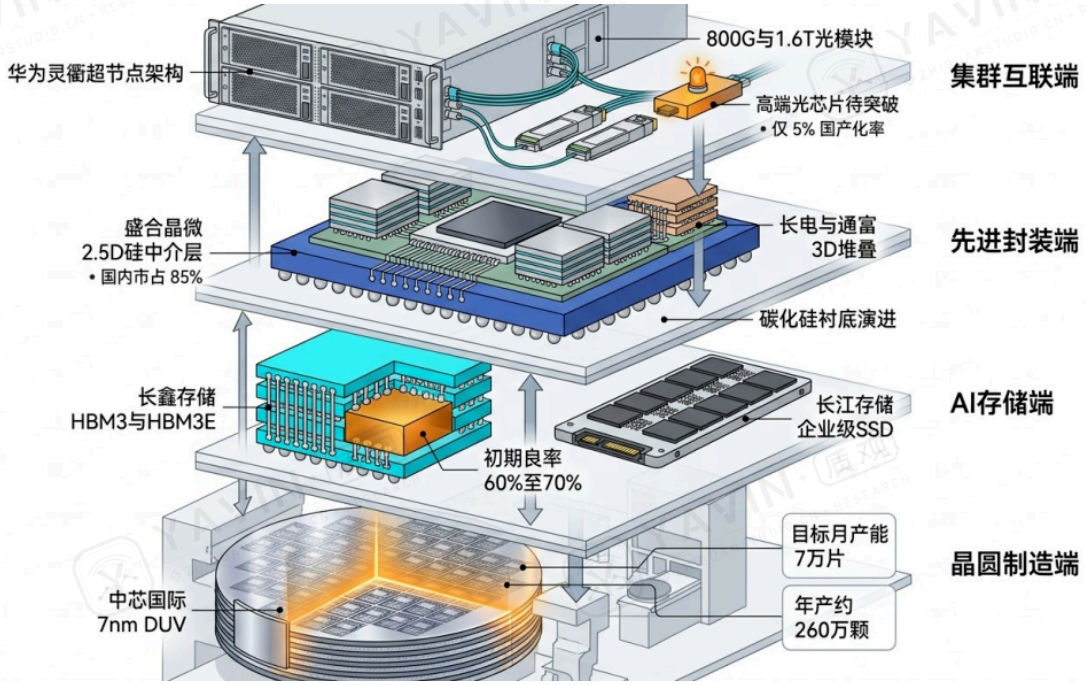
4.2 基础设施剪刀差：北美电网并网瓶颈与中国西部绿电底座

决定高功耗超节点能否规模落地的最底层约束在于国家级电网容量与重电装备制造能力。截至 2026 年 11 月，北美大型电力变压器平均交付周期已拉长至 120 至 160 周，PJM 与 ERCOT 等核心区域电网并网审批排队耗时达 36 至 60 个月；2026 年 11 月 3 日，占全美规划算力用电达 44% 的得克萨斯州直接出台新建大型数据中心并网暂停令，致使大批已到货高端 GPU 因缺乏变电配额被迫推迟点亮。

反观国内，依托特高压直流输电网络与完备的本土变压器、高密液冷制造链，内蒙古乌兰察布与和林格尔、甘肃庆阳、贵州贵安、宁夏中卫及青海绿电园区已建成吉瓦级绿电智算底座。西部枢纽工业绿电综合电价低至每

度 0.28 至 0.35 元人民币，不足欧美核心节点电价的一半，头部智算园区机柜上架满载率在 2026 年已越过 80% 至 85% 盈利线，充沛低廉的绿色电力有效对冲了国产超节点在制程与光互连上的额外能耗支出。

EXHIBIT 12 **中国AI算力底座的四维工程支撑体系（三维封装×超节点光互连×开源编译×绿电底座）**



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

CHAPTER 05 · Q3 2026 STATUTORY AUDIT & T-HEAD SPIN-OFF

05 商业闭环与三季报穿透：平头哥分拆与 390.85 亿元备货锁产

5.1 需求侧双引擎：运营商国产化集采与阿里平头哥分拆赴港上市

2026 年中国本土算力需求已脱离早期零星试点，形成三大电信运营商智算底座与头部互联网大规模推理替代的双轮支撑。在中国移动、中国电信与中国联通 2025 至 2026 年连续多轮百亿级新型智算中心集采中，国产化训练与推理服务器标包占比已稳定在 70% 至 85% 以上，金融、能源、电网、政务与央国企行业大模型的推理扩容采购基本锁定本土硬件体系。

互联网科技巨头的算力布局同步从内部降本迈向独立资本化。字节跳动、阿里巴巴、百度与腾讯既是国产商用加速卡的最大现金买方，其内部自研 ASIC 也已跨过十万片级规模门槛。2026 年 11 月 10 日，阿里巴巴集团正式发布公告，拟将旗下芯片研发主体平头哥分拆并赴香港联交所独立上市。公开资料显示，平头哥自研真武 810E 与 M890 云端加速芯片已在千问大模型推理、电商搜索推荐及外部智算客户中累计交付突破 47 万片，叠加玄铁处理器与出货超 50 万颗的镇岳 510 企业级固态硬盘主控芯片。通过股权分拆确立治理中立性与独立市场化定价，有助于打消外部云商及政企客户的同业竞争顾虑，打开第三方商业市场空间。

EXHIBIT 13 华为昇腾核心客户分配结构与电信运营商国产智算集采格局



数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

5.2 供给侧三季报穿透：科创板四强 305.94 亿元营收与 390.85 亿元锁料

核验本土商用算力企业真实经营水位的直接依据，是 2026 年 10 月末披露完毕的科创板上市公司第三季度法定财务报告。将海光信息、寒武纪、摩尔线程、沐曦股份四大独立算力龙头及同赛道参照企业燧原科技的 2026 年前三季度法定科目对齐拆解，可以清晰还原供给侧的盈利质量与备货节奏。

表 2：中国本土核心独立 AI 算力上市公司 2026 年三季报（9M26）法定全口径财务穿透与备货对账表（数据来源：各公司 2026 年第三季度法定报告）

企业名称与证券代码	2026 Q1–Q3 法定营业收入 (YoY)	2026 Q1–Q3 归母净利润 / 扣非净利	综合毛利率	Q3 末存货 + 预付款项（战略锁料）	9M26 经营现金流净额	核心产品矩阵与财务健康度诊断
海光信息 (688041.SH)	**149.55 亿元** (+57.58% YoY；Q3 单季 58.56 亿)	**28.47 亿元** / 扣非 **25.84 亿元**	**58.73%** (Q3 单季)	**148.06 亿元**（存货 89.21 亿 + 预付 58.85 亿）	**-24.43 亿元**	**x86 CPU + 深算 DCU 双轮驱动**：类 CUDA（ROCm）迁移成本极低；前期峰值预收款集中交付核销使合同负债降至 0.12 亿，正全力向上游锁产备货
寒武纪 (688256.SH)	**95.26 亿元** (+106.76% YoY；Q3 单季 35.31 亿)	**37.15 亿元** / 扣非 **35.23 亿元** (+148.3%)	**56.53%**	**147.09 亿元**（存货 116.65 亿 + 预付 30.44 亿）	**+2.28 亿元**	**互联网大厂推理主力卡**：思元系列在大模型推理集群规模放量，扣非盈利高增且经营现金流转正；审慎计提 5.02 亿元存货跌价准备

企业名称与证券代码	2026 Q1–Q3 法定营业收入 (YoY)	2026 Q1–Q3 归母净利润 / 扣非净利	综合毛利率	Q3 末存货 + 预付款项（战略锁料）	9M26 经营现金流净额	核心产品矩阵与财务健康度诊断
摩尔线程 (688795.SH)	**32.89 亿元** (+319.15% YoY；Q3 单季 15.52 亿，+1774%)	**0.38 亿元**（同比扭亏；**Q3 单季扣非转正 +0.20 亿**）	**53.71%**	**59.61 亿元**（存货 33.63 亿 + 预付 25.98 亿）	**-33.09 亿元**	**全功能 GPU 架构突围**：MUSA 生态兼容图形渲染与 AI 训推，Q3 单季营收暴增近 18 倍并首次实现单季扣非盈利转正，处于高强度扩产垫资期
沐曦股份 (688802.SH)	**28.24 亿元** (+128.45% YoY；Q3 单季 15.00 亿，+367%)	**6.39 亿元** / 扣非 **2.59 亿元（Q3 扣非 3.08 亿）**	**54.99%**	**36.09 亿元**（存货 12.95 亿 + 预付 23.14 亿）	**-13.38 亿元**	**曦云 C 系列训推一体放量**：全国产供应链版本加速导入智算中心，Q3 单季扣非净利达 3.08 亿元跨越盈亏平衡线
四大独立算力企业合计	**305.94 亿元**	**72.39 亿元**（全部实现归母净利为正）	**53.7%–58.7%**	**390.85 亿元**（存货 252.44 亿 + 预付 138.41 亿）	**-68.62 亿元**	**集体跨越盈利拐点，以近 391 亿元真金白银锁定晶圆、2.5D 封测与高带宽显存配额**
燧原科技 (688801.SH, 参照)	**23.32 亿元** (+331.74% YoY；Q3 单季 12.12 亿)	**-7.54 亿元**（同比减亏 1.34 亿；Q3 净亏收窄至 -1.22 亿）	单卡 **32.70%**	深度绑定头部互联网客户定制备货	持续改善中	**互联网深度绑定范式**：第一大客户收入占比达 83.83%，整机系统集成拉低表观毛利率，随规模放量亏损快速收窄

数据来源与核验口径：上市公司法定财报披露、国际通用行业统计与公开技术文献整理（2026.11）。

穿透法定合并口径账单，2026 年前三季度海光信息实现营业收入 149.55 亿元，寒武纪实现 95.26 亿元，摩尔线程与沐曦股份分别实现 32.89 亿元与 28.24 亿元；四家企业合计营业收入达 305.94 亿元，合计归母净利润达 72.39 亿元，全部实现归母净利转正，综合毛利率稳定在 53.7% 至 58.7% 的硬件合理区间。这表明本土商用加速卡已走完送样验证周期，迈入百亿级规模收入与经营利润兑现阶段。

四家企业三季度末存货与预付款项合计达 390.85 亿元，直接对应同期 68.62 亿元的经营现金流净流出。在先进逻辑晶圆、本土高带宽存储及 2.5D 硅中介层封测紧缺的卖方市场下，设计企业须提前两至三个季度向上游全额预付锁产，并对下游运营商与互联网集采承担验收账期。未来四个季度内，谁能凭更高的万卡集群实际算力利用率与更顺畅的软件栈将这 390.85 亿元在途晶圆及库存加速结转为现金回流，谁就将在周转效率上确立优势。

5.3 2027—2028 核心观察指标与风险边界

面向 2027 至 2028 年演进窗口，评估本土算力底座商业与工程成色应聚焦四项量化路标。在制造与存储齐套率上，跟踪 1.5 微米及亚微米混合键合良率向云端大裸片的延伸进度，以及本土高带宽存储产线对海外 HBM 颗粒的实际替代比例；在资产周转上，监测海光信息 148.06 亿元与寒武纪 147.09 亿元锁料资产在年报及后续季报中的存货周转天数斜率与合同负债补库信号，防范代际切换期的库龄减值风险。

在超节点开放度与算法协同上，则需检验灵衢互连协议伙伴的扩展速度，核查非华为系商用加速卡在跨柜高速光电交换标准上的协同进展与外部客户现场真实利用率；同时观察新一代国产旗舰模型是否在架构设计初期即原生对齐本土芯片的片上缓存容量、矩阵向量配比与低精度格式，从源头压缩硬件移植损耗。

STATUTORY FILINGS × SYSTEM ARCHITECTURE × SUPPLY CHAIN VERIFICATION

研究方法论、数据核验口径与核心产业判断

M1 · 权威市场统计与学术顶会双锚定（IDC & IEEE ISCAS 2026）

宏观出货数据采信 IDC 2026H1 中国加速计算跟踪报告；三维逻辑折叠参数严格采信 IEEE ISCAS 2026 大会主旨报告与量产芯片实测对比（麒麟 2026 vs. 麒麟 9030 Pro）。

M2 · 双向客观工程审计原则（TWO-SIDED ENGINEERING AUDIT）

既不夸大单点参数，也不回避预训练调试折损、超节点跨柜功耗代价与经营现金流垫资压力，严格按工作负载算术强度给出适用边界。

M3 · 2026年三季报与11月最新公告对齐（Q3 2026 & NOV 2026 FILINGS）

全口径核验海光信息、寒武纪、摩尔线程、沐曦股份、燧原科技 2026 年第三季度法定报告及 2026 年 11 月 10 日阿里巴巴分拆平头哥上市公告。

M4 · 非商业公开研究交流声明（NON-COMMERCIAL SHARING）

本报告彻底剔除一切展会内部台账、卖方个股评级与商业推介信息，仅供硬科技产业与工程学术界公开免费参考。

非商业免费分享与免责声明：本报告由「Yavin 质观」基于公开技术文献、学术顶会论文、国际通用行业统计及上市公司法定披露文件独立整理，仅作个人公开研究免费交流分享，非商业用途。报告内容在任何情况下均不构成任何证券投资建议、个股评级、目标价预测或商业采购依据，读者应当独立判断并自行承担风险。