

内存墙的五层解法：真内存池是 NVLink，不是 CXL——机架内存互连的价值分配与国产卡位

内存墙的五层解法——真内存池是 NVLink 不是 CXL：机架内存互连的价值分配与国产卡位

AI 推理把内存推成系统第一瓶颈（内存占服务器 TCO 40–50%、VR200 整机架 BOM 中内存占 26%、64GB RDIMM 半年翻倍破千美元），"把内存从单台服务器中解放出来"正在成为 AI 基础设施新主线。但解法分五层，价值分配极不均匀：真正大规模在用的内存池是 NVLink 机架体系（GB300 单柜 20TB HBM/130TB/s，VR200 将至 75TB/260TB/s），而市场默认的 CXL 跨机池化仍在公有云 preview 与数据库单点商用阶段、TAM 八成是内存颗粒；最有效的软件 KV-cache 卸载没有纯标的，价值最高的交换芯片被平台方独占。可投卡位收敛于四处：模组侧接口芯片（澜起——MRDIMM 二代唯一批量出货）、机架内铜互连（沃尔核材线材全球第二/立讯组件）、整机（工业富联）、以及 OCS 商业化外溢（Lumentum/Coherent）。铜 vs 光是本主题最大的可验证变量。

涌现资本产业研究部 · 2026-07-07

🕒 30秒结论卡

- **一句话判断：**内存正在取代算力成为 AI 系统的第一瓶颈，但"内存池化"的五层解法里，**真池是 NVLink 机架体系**（当下唯一大规模在用、2030 年约占 scale-up 互连市场八成的口径）；CXL 跨机池化是被叙事放大的远期期权（公有云仍在 preview、平台底座要等 2027）；**可投的钱集中在模组侧接口芯片、机架内铜互连、整机组装与 OCS 外溢四处。**
- **机会状态：**🟢 结构主线确立（2026H2 三拐点齐至：C2C 带宽一次性翻倍、VR200 上量、MRDIMM 二代平台适配启动）× 🟡 板块情绪高位（AI 硬件十倍股 81% 同链、涨幅与利润脱钩）——选纯度、等回调、按代际配仓。
- **市场错在哪：**① 把"内存池化"默认等同 CXL 并按 158 亿美元大 TAM 给 CXL 概念股估值——该口径八成是内存颗粒，且出自 2023 年陈旧预测，真正的第三方控制器盘子至今是千万元级/季；② 忽略模组侧解法 MRDIMM 的确定性——接口硅从 1 颗变 11 颗、二代 12800MT/s 随 2026H2-2027 新 CPU 平台放量，全球只有两家能做而只有一家在批量出货；③ 低估平台方的价值独占——交换芯片与高速接口 100% 自研、软件卸载开源自用，产业链只剩"物理载体"可分。
- **重点标的：**澜起科技（688008，模组侧主档）、沃尔核材（002130，线材全球第二 24.9%，不在册）、工业富联（601138，机架组装约 45%）、Lumentum（LITE，OCS 商业化最纯）、博通（AVGO，开放阵营事实赢家）。
- **最大反证：**机架内 NVLink 互连若在 2027-2028 提前转向光背板/CPO，铜缆-连接器这条最实的国产卡位线将被直接削弱（当前证据：Rubin Ultra 跨机架将首用 CPO，但机架内仍是铜背板；Kyber 延期反而延长铜方案生命周期）。

投资要点

一句话框架：钱流从"云厂商 7000 亿美元资本开支"经"整机架价值 780 万美元、其中内存 200 万"流向五层解法；价值链上平台方独占交换芯片与软件层，受益链真正可分的卡点只有四个——模组侧接口硅（数量 1→11 颗）、机架内铜互连（5184 根线缆/柜）、整机组装、OCS 外溢——逐一落到公司卡，情景按 2026H2 拐点、2027 放量、2028 光化风险三段管理。

#	核心观点	方向
1	内存是整机最大单项成本：占服务器 TCO 40–50%，VR200 机架 BOM 780 万美元中内存约 200 万（26%）；64GB RDIMM 合约价半年 \$450→\$900+，Q2 破千	●
2	机架内存池量价齐升：NVL72 机架 2025 约 2.9 万台→2026 预期 7–8 万台；单柜快内存 37TB→75TB、聚合带宽 130→260TB/s（VR200，2026Q3 首发）	●
3	MRDIMM 是模组侧确定性最高的解法：接口硅 1→11 颗、二代速率为主流 RDIMM 两倍，2026H2–2027 平台适配——澜起是唯一已批量出货者	●
4	CXL 按大 TAM 估值是陷阱：八成 TAM 归颗粒、公有云仅 preview、原厂自研控制器挤压第三方——CXL 仓位只配期权	●

一、钱流：内存为什么成了第一瓶颈

三个数字定义问题。**其一**，全球四大云厂 2026 年资本开支指引合计约 7000 亿美元（同比约 +70%，约四分之三投向 AI）——这是分子。**其二**，内存占 AI 服务器总成本已达 40–50%，是整机最大单项：VR200 NVL72 整机架 BOM 约 780 万美元中内存约 200 万美元、占 26%，若云厂自采内存可把整柜成本降至约 670 万——内存直接决定算力的采购经济学。**其三**，推理时代的内存需求是结构性的：单次长文本请求的 KV-cache 占用几十到上百 GB 高带宽内存，而 DRAM 涨价周期（64GB RDIMM 合约价从 2025Q4 的 \$450 到 2026Q1 的 \$900+、Q2 破 \$1000）把“每 GB 内存的利用率”变成了所有推理厂的成本生死线。

需求端要多内存，供给端内存暴涨——于是“把内存从单台服务器中解放出来、池化共享、分层卸载”成为 AI 基础设施 2026–2028 年最确定的工程主线。问题只在：五层解法里，哪一层有可投的钱。

二、价值链：五层解法全景图

按作用域从内到外分五层（本报告的价值链主图）：

层	解法	规格锚（口径见来源节）	成熟度	可投性
① 封装 内存	HBM 堆叠	单 GPU 最高 288GB（HBM3e）	成熟放量	与本部存储系列并轨（HBM/HBF 已有专题）
② 节点 内存	NVLink-C2C（CPU↔GPU 一致内存）	Grace 三代均 900GB/s；Vera 世代一次性翻倍至 1.8TB/s（2026H2）	大规模在用	平台自研，无直接标的、是拐点信号
③ 机架 架内存	NVLink+NVSwitch 显存池	GB200：13.4TB HBM/130TB/s；GB300：20TB HBM+37TB 快内存/130TB/s；VR200 NVL144：75TB/260TB/s	大规模在用 =真内存池	物理载体可投：18 颗 NVSwitch（自研）+5184 根铜缆+连接器+整机
④ 跨 机	CXL 解耦池	直连延迟 170–250ns（约本地 2 倍）；跨机实测延迟 +60%、带宽仅本地约 1/10	preview/单点商用	期权：控制器（澜起 MXC/Astera Leo）均未成收入量级
⑤ 软 件	KV-cache 分层卸载	缓存命中场景 TTFT 降 3–10 倍；DeepSeek 命中价降至 1/10	已放量（生产环境）	2026H2 起有硬件载体：BlueField-4 ICMS 推理存储、HBF（送样）、Groq LPX

核心判断：当前大规模在用、市场最大的内存池是 NVLink 体系——英伟达新一代平台的统一内存走 C2C 而非 CXL，这是带宽量级（TB/s vs 百 GB/s）与耦合方式的本质区别。层③的机架级出货就是本主题的总量指针：**NVL72 机架 2025 年约 2.9 万台 → 2026 年预期 7–8 万台（2026 年 2 月单月已达 6500 台，年化互洽）**。2030 年机架内 scale-up 互连市场超 300 亿美元、其中 NVLink 约 250 亿（单一机构预测口径，作边界不作基线）。

三、层④现实检查：CXL 何时才值得下注

CXL 的采用拐点确实出现了——但形态与叙事完全不同：

- 公有云首例是**"扩展"不是"池化"**：Azure M 系列 VM 于 2025-11 上线业界首个公有云 CXL 内存（preview 非正式商用），用的是单机内存扩展（CXL 2.0、每控制器 2TB），不是跨机共享池。
- 真正的跨机池化只有一个商用单点**：阿里云 PolarDB 的数据库内存池（全球首个 CXL 交换商用化，10TB 级、对比 RDMA 性能 +2.1 倍）——场景是数据库缓存，不是 AI 训练。
- Meta 的用法最诚实**：自研控制器把退役 DDR4 回收进推理服务器（省 25% 服务器），CXL 在它手里是降本工具而非新市场。
- 平台底座未到位**：在售主流平台仅支持 CXL 2.0 Type-3；CXL 3.x 的 CPU 底座要等 2026–2027 的下一代平台——**狭义"跨机架共享池规模商用"确实是 2027 年之后的事**。

5. **TAM 拆解是陷阱**：流传的"CXL 硬件 2028 年 158 亿美元"出自 2023 年的机构预测，其中约八成是内存颗粒；现实是第三方控制器至今收入千万元级/季（澜起 MXC 出货占约九成但绝对额极小、仅送样至量产过渡），且**内存原厂自研挤压已坐实**（SK 海力士首颗自研控制器明确减少第三方依赖）——第三方控制器的窗口在 2.0 世代，3.x 世代将被进一步压缩。

下注/推翻信号（可验证）：做多信号=Azure 从 preview 转正式商用并扩至第二 VM 家族（窗口 2026H2–2027H1）、Astera 电话会首次把 Leo 单列为收入驱动、澜起 MXC 年收入过 5 亿元或单独拆分；推翻信号=原厂自研控制器量产自用扩散、英伟达继续把内存扩展锁在 NVLink Fusion（GPU 侧无 CXL）、大容量 DIMM 价格回落使直插重新便宜。**当前结论：CXL 不构成任何标的的独立买入理由，一律按期权仓位处理。**

四、开放标准的真相：第二供给是以太网，不是 UALink

- **NVLink Fusion（2025–05 发布）是攻不是守**：把 NVLink 物理层开放给第三方 CPU/ASIC（富士通/高通/Arm/三星代工相继加入，并向 Marvell 注资 20 亿美元）——用"发门票"换取整个 scale-up 生态锁进自家物理层，层③的垄断由封闭升级为事实标准。
- **UALink 短期无硬件**：1.0 规范 2025–04 发布，但原生交换芯片 2026 年底前无法就绪——**AMD 自家旗舰 Helios/MI400 机架实际跑的是"UALink over Ethernet"**，交换芯片用博通 Tomahawk 6；原生 UALink 要等 2027 年 MI500 世代。
- **开放阵营 2026 年的事实赢家是博通**：Tomahawk Ultra（51.2T/250ns，2025–07 出货）+SUE 规范是当下唯一可出货的非 NVLink scale-up 方案，急于出机架的云厂商全部转投以太网系。

对组合的含义：给"开放标准挑战 NVLink"定价的正确标的是 AVGO（已兑现出货），不是 UALink 概念（ALAB/MRVL 的该项收入 2027 年前不存在）。

五、受益链与公司卡

在册标的市值/PE(TTM) 绑定涌现资本雷达（2026–07–07 快照）；不在册者标注。同一主题内按"卡位硬度×兑现时点"排序。

5.1 全球标的

标的	层	卡位	雷达绑定	判断
英伟达 NVDA	基础	层②③垄断者+Fusion 事实标准；层⑤全栈（Dynamo/ICMS/LPX）	\$49440 亿/PE 30	主题对其是增量非主逻辑
博通 AVGO	基础	开放阵营事实赢家：TH6+SUE 是 2026 唯一非 NVLink 方案	\$18492 亿/PE 63	scale-up 以太网+ASIC 双引擎
安费诺 APH	甜蜜	层③物理载体：铜背板 cartridge（5184 根/柜）主供	\$1947 亿/PE 47	Kyber 世代铜转 PCB 是远期变量
Lumentum LITE	甜蜜	OCS 商业化最纯放量（约五成份额轨迹、2026 底约 \$1 亿/季度口径）	\$550 亿/PE 148	买家集中（谷歌占九成台数）；卡位 A/时点等回调
Coherent COHR	甜蜜	OCS 第二供应商（液晶路线）+高损耗预 算收发器	\$620 亿/PE 185	OCS 占营收极小，主题纯度低
Astera Labs ALAB	观察	当期引擎是 Scorpio/Aries；Leo（CXL）与 UALink 均为期权	\$674 亿/PE 270	期权已被 270 倍 PE 预付
SanDisk SNDK	观察	HBF 高带宽闪存与 SK 海力士共同标准化——层⑤下沉闪存最直接载体	\$2558 亿/PE 67	2026H2 仅送样、设备 2027 后；深周期峰值护栏适用
Rambus RMBS	观察/ 期权	三寡头第二席；MRDIMM 套片 2026 底初 始出货、2027 放量	\$119 亿/PE 55	比澜起晚约一年；监管悬项未决（与本部 6-28 深报口径一致）

5.2 国产卡位

标的	层	卡位	雷达绑定	判断
澜起科技 688008	甜蜜 (主档)	模组侧核心：全球唯一全套内存条芯片商，MRDIMM 二代唯一批量出货	\$482 亿/PE 120	见公司卡；估值高位分批不追
沃尔核材 002130	甜蜜 (线材层)	高速铜缆全球第二（24.9%，2024 收入口径）、国内唯一量产 224G	不在册	见公司卡；十层深报曾判透支，等回调
工业富联 601138	基础	NVL72 整机组装约 45% 第一（GB300 2026 年约 45%，产业口径）	\$2031 亿/PE 38	全链估值最低的机架β；组装净利率薄
立讯精密 002475	甜蜜	NVDA 链国产电连接最强：224G 铜缆/MCIO 批量交付，NVL72 单柜方案约 209 万元（机构测算）	\$692 亿/PE 32	消费电子主业稀释 AI 弹性，估值合理
鼎通科技 688668	观察	连接器组件二供（安费诺/莫仕），224G 已认证批量供货	\$70 亿/PE 207	卡位证据硬但价格已透支（库内判避雷口径）
兆龙互连 300913	观察	800G DAC 过英伟达认证（产业口径）；224G 线缆验证中未量产	不在册	落后沃尔一个身位
华丰科技 688629	观察	国产分支：112G 背板连接器批量进昇腾链；224G 仅样品	不在册核验 (07-05 已入雷达)	属国产算力链标的，NVLink 体系内暂无实绩

排除/警示：意华股份（224G 尚在开发、2026H1 才出样，且光伏支架占营收约六成、AI 纯度全组最低）；CXL 概念股按第三节结论一律不作为独立逻辑；美光（无键合/池化特异卡位，另见存储系列）。

5.3 重点公司卡

澜起科技 688008 —— 模组侧解法的唯一先行兑现者 ●

- 卡位：内存接口芯片全球前二（与 Rambus 争第一、DDR5 RCD 份额过半口径）；MRCD/MDB 全球唯二、二代（12800MT/s）近两季出货显著提升、获全球主要模组厂批量采用——比 Rambus 早约一年；PCIe 6.x Retimer 已推出；CXL MXC 送样（期权）。
- 数字：2025 营收 54.56 亿（+49.9%）、互连类 51.39 亿（毛利率 65.6%）；2026Q1 营收 +19.5%、归母 +61.3%、互连类毛利率 71.5%；四款新品合计 2.69 亿（+93.8%）。
- 逻辑：MRDIMM 把单条模组的接口硅从 1 颗变 11 颗（1 MRCD+10 MDB，架构级事实；模组级金额为 RCD 的 4 倍——单一管理层口径），二代随 2026H2–2027 新平台适配放量；RDIMM 涨价反而凸显 MRDIMM 性价比（公司口径）。
- 反证：新 CPU 平台延期；"规模试用"到"规模量产"的斜率；PE 120 已预付一部分。

沃尔核材 002130（乐庭智联） —— 线材层的全球第二 ●

- 卡位：高速通信线材全球第二（24.9%，2024 年收入口径、招股材料引第三方）、中国最大；**国内唯一量产 224G**，直接客户为安费诺/莫仕/泰科/立讯等连接器厂（线材层上游、非英伟达直供）。
- 数字：高速铜缆收入 2025H1 同比 +397.8%、前三季度招股书口径 +242%（IR 口径约 +300%，双口径并列）；2025 前三季总营收 60.82 亿（+26%）。
- 反证：448G 仅样品（下一代若掉队即让位）；机架内互连转光；本部十层深报曾判其透支——卡位真、时点等回调。

工业富联 / 立讯精密 / Lumentum：分别是机架β（组装 45%、PE 38 全链最低）、组件层最强国产（单柜 209 万元方案价值）、OCS 商业化最纯载体（谷歌外第一个规模买单者出现即上修）——判断见表内，不再展开。

六、情景与节奏

情景	关键假设	组合动作
基准	2026H2 三拐点兑现：C2C 翻倍+VR200 上量（Q3 首发 Q4 放量）+MRDIMM 二代平台适配启动；机架 7-8 万台	澜起/工业富联/立讯为主仓，沃尔核材等回调，LITE 观察谷歌外订单
乐观	MRDIMM 2027 提前放量+Azure CXL 转正式商用+机架出货超 8 万台	加模组侧与铜互连弹性，CXL 期权（澜起 MXC/ALAB）升层
保守	新 CPU 平台延期拖 MRDIMM、机架出货 6 万台以下、DRAM 价格见顶回落	收缩到工业富联/APH 等份额型，高 PE 弹性先杀

节奏：2026H2 看三拐点落地与澜起新品收入环比；2027 看 MRDIMM ramp、原生 UALink 首硬件、CXL"规模商用元年"成色；2028 看 HBF 设备、Feynman 世代 NVLink 光化（CPO）——**铜 vs 光是贯穿全程的头号变量**：当前证据是机架内仍铜背板（Kyber 延期反而延长铜方案生命周期）、跨机架先光化，一旦"机架内转光背板"出现，铜缆/连接器线全部重写。

七、风险与反证

#	反证信号	冲击	监控
1	机架内 NVLink 宣布转光背板/CPO 提前	铜缆-连接器国产卡位线重写	英伟达路线图、448G 电通道量产成败
2	新 CPU 平台（含二代 MRDIMM 适配）延期	澜起/Rambus 放量后移	平台发布节奏、模组厂拉货
3	DRAM 涨价周期见顶回落	"内存太贵→池化省钱"的紧迫性下降	合约价环比、原厂资本开支
4	CXL 意外加速（Azure GA+第二云跟进）	本报告对层④的低配结论需上修	云厂商公告、澜起 MXC 收入拆分
5	开放阵营原生硬件早产（UALink 交换 2026 内就绪）	NVLink 独占溢价折价	Marvell/Astera 芯片流片进展
6	板块情绪逆转（同链十倍股集体回撤）	高 PE 标的戴维斯双杀	附录体温计指标

八、附录：行情体温计与仓位纪律

三组市场统计（2026-06-30 口径）作为操作层背景：①2025 年以来 A 股 74 只十倍股中 **81% 集中在同一条 AI 算力硬件链**，且 31% 已从高点回落至 10 倍以下——同链标的的 β 高度同源，名单本身不构成分散；②2026 上半年涨幅 TOP100 的 PE 中位数 174 倍（全 A 约 18 倍）、Q1 净利中位仅 0.46 亿——市场只为弹性叙事付钱、不为便宜付钱，涨幅与利润明显脱钩；③上一轮核心资产（2021 年白酒）顶部买入者五年仍深套约五至八成，当下 AI 硬件龙头市值已坐到当年泡沫王座的同一量级。**含义：本报告标的池普遍不便宜（澜起 PE 120、鼎通 207、ALAB 270），产业逻辑的确定性不能替代入场时点的纪律——分批、等回调、按代际切换，把附录当刹车用。**

九、口径与来源置信度

类别	代表	置信度
官方规格	英伟达产品页/路线图（13.4TB/20TB/75TB、130/260TB/s、18 颗 NVSwitch、5184 根铜缆、C2C 900GB/s→1.8TB/s）、Broadcom 出货新闻稿、UALink 1.0 规范、CES/GTC 2026（ICMS、CPX 撤出、Groq 交易）	A
公司财报/公告/招股材料	澜起 2025 年报+2026Q1+投关记录、沃尔核材招股书与 IR、Rambus 电话会、Coherent 披露	A
机构口径	机架出货（供应链月度追踪）、VR200 BOM 780 万、650 Group 2030 拆解（单一机构）、OCS 市场（两源交叉）、Azure/阿里/Meta 案例（论文+官方微博客）	B
产业口径	组装份额约 45%、立讯单柜 209 万、“批量交付英伟达”类表述	B/C（已逐处标注）
推演	情景与节奏、“CXL 下注信号”	D

口径说明：五层框架源自一份匿名产业研究材料，本报告对其全部关键数字做了独立重验——其中“机架 9 颗交换芯片”（实为 18 颗）、“C2C 逐代翻倍”（实为 Grace 三代同为 900GB/s、Vera 一次性翻倍）、“单机架 700 万美元+”（仅 VR200 世代成立）、“接口硅价值 10 倍”（11 颗是数量口径、金额约 4 倍）等多处已按核验结果修正后采用；“2030 年 scale-up 超 300 亿美元”为单一机构预测，只作边界。

数字三道关自评：①量级一致性——18 颗 NVSwitch×7.2TB/s=129.6≈130TB/s、72×3.6TB/s=259.2≈260TB/s、2 月单月 6500 台年化≈7.8 万台与全年预期互洽，交叉闭合；②单位显式——TB/s 与 GB/s、颗数与金额、preview 与正式商用逐处区分；③关键数字≥2 源，单源（650 Group、BOM 拆解、专家会份额）与公司自报口径已逐处标注。

免责声明：本报告由涌现资本产业研究部出品，仅供内部研究与学习交流，不构成任何投资建议。英伟达供应链与云厂商部署细节多为非公开信息，文中产业口径可能与实际存在偏差；技术路线（铜/光、CXL 节奏）随平台方决策可能显著变化。市场有风险，决策需独立。

—— 涌现资本产业研究部