# GPU四小龙进入竞争新阶段

图为2026长三角机器人及自动化展览会摩尔线程展台

本报记者 张心怡 实习生 刘璐

9月2日，燧原科技启动网上、网下申购。至此，摩尔线程、沐曦股份、壁仞科技、燧原科技四家国产算力芯片企业全部完成与资本市场的对接。这意味着国产高端GPU完成了初步的技术验证和资本化，下一阶段将是规模化商用能力的直接比拼。

市场习惯把四家企业合称为“国产GPU四小龙”，但翻开2026年半年报和燧原的招股书就能发现，这只是一个通俗标签。四家走的技术路线、做的产品、面对的客户，差异远大于相似。与其说这是一个赛道里的四家竞对，不如说它们各自站在国产算力产业链的不同节点上，用不同的方式回答同一个问题：在市场高度集中的AI算力产业里，国产算力企业如何找到自己的位置？

## 摩尔线程：全功能GPU向下延伸，算力从云端走向终端

摩尔线程不止布局云端大模型训推，同时将产品线向下延伸至边缘SoC、AI算力本等。

2026年上半年，摩尔线程营收17.36亿元，同比增长147%。相比亮眼的增长数字，收入结构变化更值得关注：公司经营正从早期项目交付模式向规模化交付演进——2026年3月签订的6.6亿元夸娥智算集群合同已在报告期内完成交付，意味着国产算力正在从信创示范项目走向常态化采购。

目前，基于MTT S5000构建的夸娥万卡集群已成功部署并上线服务，核心产品MTT S3000云电脑方案单卡支持32路并发。当前公司营收高度集中云端，云端智算产品贡献营收占比达97.5%。归母净利润亏损大幅收窄近96%，但扣除非经常性损益后依旧亏损1.51亿元，尚未实现经营性盈利。高研发投入与存货攀升仍是现实存在的压力。

在国产GPU大多优先深耕云端训推市场的背景下，摩尔线程选择了一条最有挑战性的路线：不只布局云端大模型训推，同时将产品线向下延伸至桌面显卡、边缘SoC、AI算力本乃至家庭AI中枢。在边缘与终端智算领域，以“长江”为代表的智能SoC系列产品，正广泛赋能边缘智能、具身智能、工业智造等多元场景。摩尔线程希望通过与云端智算产品的深度协同，打通“云-边-端”一体化解决方案，满足各类场景下的AI训推需求。

这套策略能否走通，关键在于：当AI能力持续从云端向边缘、终端扩散时，这种既懂图形又懂AI计算的通用芯片，能否在行业与消费市场获得真实、可持续的订单。

分业务看，云端业务放量显著；边缘与终端产品尚处于起步阶段，贡献了2.33%的收入，商业化考验才刚刚开始。

## 沐曦股份：全国产供应链的盈利路径，正在被验证

仅有硬件成本优势还不够，沐曦同时将“使用国产算力的隐形成本”压到了临界点以下。

沐曦上半年营收13.24亿元，同比增长44.67%，归母净利润6.12亿元，公司也因此成为“国产GPU四小龙”中率先实现盈利的公司。但这笔盈利并非来自芯片销售：其中有8.87亿元来自公允价值变动收益，公司明确标注该项收益“不具有可持续性”，扣非净利润仍为-4886万元。更值得关注的数据来自二季度，单季度扣非净利润为5407万元，环比由亏转盈，显示其商业化进程取得阶段性进展。

产品及服务持续获得下游客户的认可，是该公司上半年业绩增长的主因。报告期内，公司首款基于全国产工艺的高性能通用GPU产品曦云C600顺利进入量产交付，从芯片设计到制造、封装，再到配套软件栈，实现了国产供应链闭环。在财务侧，全国产供应链意味着成本端的可控——上半年毛利率达57.2%，同比提升1.1个百分点，远超单纯硬件销售的毛利水平；在客户侧，则意味着供应交付的确定性与可预期性。今年5月，曦云C600首批通过国家《安全可靠测评》，为产品进入金融、能源等关键基础设施行业提供了准入基础。

仅有硬件成本优势还不够，沐曦同时将“使用国产算力的隐形成本”压到了临界点以下。国产GPU长期面临一个尴尬：芯片单价虽低，但代码迁移、算子补全、性能调优等隐性投入高昂，导致综合使用成本高于国外方案。沐曦半年报里的几个动作，指向同一目标：拆除这座迁移成本的墙。沐曦股份自研的MX-MACA软件栈高度兼容国际主流生态，已支持超过6000个CUDA应用，并与超过1000个模型原生适配，开发者无需过多修改代码即可迁移至沐曦GPU运行，为公司构建了深厚的生态护城河。

此外，报告期内公司还发布了面向AI4S（科学智能）的曦索X系列GPU，由此形成覆盖人工智能训练与推理、图形渲染与科学智能的完整产品矩阵，进一步拓宽了市场空间。

## 壁仞科技：港股“GPU第一股”，接受多元市场检验

壁仞选择港股上市，意味着它必须用持续的商业化成果和技术迭代，证明自己的长期价值。

壁仞科技是四家中唯一选择港股上市的企业。8月28日晚，壁仞科技交出上市后的首份中报：营收12.36亿元，同比增长近20倍，超过去年全年；净亏损同比收窄76.4%，毛利率提升10.8个百分点至42.7%，但离盈利还有距离。

商业化方面，壁仞科技的壁砺系列训练及推理产品销售持续扩大，客户结构与场景覆盖进一步拓展。根据公告披露，壁仞科技客户已覆盖头部互联网企业、AI大模型开发商、国家级AI算力平台、AI数据中心、电信运营商等多个行业。壁仞科技表示，已完成互联网大客户供应商引入认证并开始批量交付产品，为收入持续放量增长打开更大空间。

产品层面，壁仞有几个值得关注的信号。自研BLink 2.0互连协议最高支持1024卡单域扩展，是目前国内GPGPU路线中走得比较远的集群互联方案之一；“光跃”超节点已实现数千卡规模商用部署，MoE模型端到端训练性能提升超过30%。公司同时在WAIC 2026上发布了NPO（近封装光学）光互连方案。

港股市场对硬科技公司的估值逻辑与科创板不同，壁仞选择这条路，意味着它必须用持续的商业化成果和技术迭代，在一个更加注重资本逻辑和国际化的市场中证明自己的长期价值。

## 燧原科技：DSA路线闯关IPO，市场认可多技术并行逻辑

燧原用专注AI推理的DSA架构拿到了市场通行证，多路线并行已是产业现实。

摩尔线程走“全功能GPU”路线，生态挑战最大但天花板最高；沐曦与壁仞走“GPGPU”路线，主攻数据中心高性能计算；燧原则选择DSA架构，专注云端AI训推，在推理场景追求极致能效。

与其他国产GPU厂商类似，燧原目前仍处在尚未盈利的状态。按照招股书的解释，亏损源于行业的前置性投入：云端AI芯片硬件需要跟随先进工艺快速迭代，独立的软件栈需要持续打磨，还需要与供应链伙伴联合研发，保障供应链稳定。研发投入居高不下。同时AI芯片产业链预付采购比例高，也是行业共性难题。本次IPO拟募资60亿元，重点投向第五代、第六代AI芯片研发及产业化。

燧原作为四家中唯一采用DSA架构的企业，招股书这样概括其技术取舍：“公司AI芯片采用DSA架构，舍弃了非AI并行计算领域的通用性。”这意味着它放弃了部分通用性，把硬件资源全部倾注于AI计算，追求推理场景的极致能效。产品线也全部围绕云端加速卡和智算集群，应用场景相对集中。

在IPO进程中，燧原推理产品比例较高的成因曾被上交所问询。燧原表示，公司总体采用“训练产品逐步探索，推理产品迅速推进”的产品迭代路径。训练领域技术复杂度与系统要求极高，而推理场景商业化落地更快，是能快速实现自我造血以反哺训练产品持续投入的务实选择。未来，随着公司第四代训推一体产品L600规模化量产及超节点系统的交付，燧原将持续拓展训练领域的应用。

但燧原的IPO落定本身就是一个信号：资本市场对“多路线并行”格局的认可。它的上市说明DSA这条技术路线，正在逐步得到资本市场的认可。

四家企业，四条路。摩尔线程用“全功能GPU”拓展应用场景，沐曦用全国产供应链验证盈利路径，壁仞用港股打开国际化窗口，燧原用专注AI推理的DSA架构拿到了市场通行证。GPGPU与DSA也并非非此即彼的竞争关系，多路线并行已是产业现实。

尽管各家路径各不相同，但行业仍要共同面对产能波动、客户结构集中、生态建设漫长的现实考验，规模化交付不等于能够兑现可持续盈利。它们反映出一个行业现实：国产算力已走过样机验证阶段，正在进入规模化交付和全栈体系竞争的新阶段。虽然资本市场提供了更加充足的资金支持，但企业的实力，终究要经由现实场景中的落地应用来检验。

## 我国研发出相干量子路由系统 破解QRAM扩展行业难题

**本报讯** 近日，记者从安徽省量子计算芯片重点实验室获悉，本源量子联合中国科学技术大学、合肥综合性国家科学中心人工智能研究院等科研团队，在“本源悟空”自主超导量子计算机上成功研发出面向桶式量子随机存储器（QRAM）的“相干量子路由系统”，相当于为数据中心架设了“快速路”，有效解决了QRAM扩展的行业难题。

如果把量子计算机比作“智慧城市”，那么量子随机存取存储器（QRAM）就是其中存储和调度数据的“数据中心”。QRAM是实现量子搜索、量子机器学习等算法的重要功能模块，其核心是“量子路由器”——它负责根据地址指令将数据准确送达目标位置。但传统方案随路由网络规模增大，每增加一个节点就需大量标准量子门层层分解指令，导致线路深度剧增、误差累积，限制了QRAM的扩展。

针对这一难点，研究团队提出了一种基于辅助能级的量子路由器构建方案。“相干量子路由系统”利用超导量子比特的辅助能级，将复杂的受控交换操作转化为由少量双比特跃迁门构成的浅线路等效实现，相当于在拥堵的数据通道上方架设了“快速路”，让数据绕过冗余节点直达目的地，大幅削减线路深度并降低误差。科研人员在“本源悟空”自主超导量子计算机上实际搭建了三个独立量子路由器和一套两层量子路由网络进行测试。结果显示，单个量子路由器的信息传输效率高达98%，两层路由网络整体传输效率也能达到93%。在随机访问测试中，单个量子路由器的平均保真度达到94.8%，两层路由网络的平均保真度达到82.4%。数据表明，该方案能够实现由量子地址控制的相干、可逆路由，并具备向更大规模网络扩展的潜力。

研究人员表示，此次突破不仅为在现有超导硬件上构建可扩展QRAM提供了新路径，更标志着我国量子计算研发从单一性能优化迈向复杂量子功能在真实硬件上的落地验证阶段，依托量子器件本身特性可实现更低成本、更高效率的运算，为量子芯片与线路协同优化开辟了新思路。（文 编）

## 长鑫宣布LPDDR6正式量产 小米18 Fold首发搭载

**本报讯** 记者**张心怡**报道：8月29日，长鑫科技在社交平台宣布，长鑫LPDDR6内存正式量产，将在小米18 Fold新折叠旗舰手机首发搭载。小米集团创始人雷军同步在社交平台表示，祝贺长鑫实现LPDDR6内存量产，玄戒O3是首家支持长鑫LPDDR6的芯片，将在小米18 Fold首发搭载，并称“我相信，这只是开始，将来还会有更多优秀的中国科技企业，强强联合，勇攀高峰”。

LPDDR6是当前国际最新一代低功耗内存标准，由JEDEC于2025年7月发布。该标准将移动端和AI内存速率区间提升至10667~14400MT/s，相当于约28.5~38.4GB/s带宽，并针对移动设备和前沿AI系统所需的更低功耗预算进行优化，有利于解决端侧大模型的带宽瓶颈。

具体来看，LPDDR6采用双子通道架构，每个裸片带有一个24位通道，每个通道又分为两个12位子通道，这种配置缩短了访问路径、降低了延迟，并保持了32字节的最小粒度。同时，它降低了核心电压，并增加了片上ECC、奇偶校验和自检功能，能够满足汽车和数据中心环境更严格的要求。

长鑫科技在8月28日发布的半年报称，公司自主研发的LPDDR6芯片性能较上一代LPDDR5X实现全面跃升。通过架构创新及数据传输优化技术，充分释放数据传输潜力，峰值速率达12800Mbps，芯片最高容量16GB。目前，该产品已向关键客户送样验证，正加速推进量产，未来预计在移动设备、服务器、智能汽车等领域广泛运用。

8月24日，小米发布新一代玄戒芯片，包含为小米最高端产品打造的AI旗舰SoC玄戒O3、1.22TB/s的高带宽AI加速芯片玄戒O100、3nm智驾高算力AI芯片玄戒D100。其中玄戒O3将在小米18 Fold首次亮相，玄戒O100和D100将于2027年正式商用。

小米称，玄戒O3是行业首款支持LPDDR6的移动SoC，单通道位宽从LPDDR5X的16bit提升至24bit，内存带宽提升至113.8GB/s，增长近50%，使手机的日常使用更流畅、端侧AI更快、功耗更低。结合缓存配置，能够实现82ns超低访存时延。

目前，SK海力士、三星均规划在2026年下半年实现LPDDR6量产出货，美光已向OEM和生态伙伴送样，预计2027年全面商用。这意味着长鑫科技的LPDDR6量产节奏与三星、海力士同步。长鑫科技与小米的协同创新成果，及长鑫LPDDR6产品在小米18 Fold的应用细节预计将在9月份的小米发布活动上进一步揭晓。

## 英伟达35亿美元投资联发科 在三大领域合作

**本报讯** 记者**张心怡**报道：8月31日，英伟达与联发科宣布，将共建从AI边缘到云端的下一代AI计算平台，英伟达已斥资35亿美元购买了联发科发行的可转债券。

具体来看，双方将在三个领域开展合作：

一是AI基础设施，联发科将与英伟达NVLink Fusion生态系统合作，助力客户开发能够与英伟达机架级系统和AI工厂集成的定制化AI基础设施。NVLink Fusion平台为多裸片（multi-die）XPU（可拓展处理单元）开发提供了预构建、预认证且经过系统预验证的基础框架，从而加速了从芯片、先进封装到机架级系统的落地进程，该平台还整合了NVLink Fusion芯粒、NV-Link-C2C、NVHBM等定制XPU的周边技术。

二是本地AI计算，双方将在GB10 Grace Blackwell超级芯片上展开合作，为边缘系统带来更强大的AI能力。双方还将合作延伸至英伟达RTX Spark芯片，为AI时代重新定义的下一代消费级PC提供算力支持。

三是汽车领域，联发科天玑Auto平台集成了英伟达技术，为智能汽车座舱提供AI与英伟达RTX图形能力，并可搭配英伟达DRIVE AGX协同工作。双方将继续深化合作基础，为日益智能化的AI定义汽车打造可扩展架构。

公开资料显示，在4G通信时代，按照手机基带/SoC出货量计算，联发科是仅次于高通的芯片设计公司，拥有强大的移动端处理器CPU及SoC芯片设计能力、方案定制能力。5G时代，联发科凭借中低端5G机型及天玑旗舰高端突破，在手机处理器出货量份额上于2020年第三季度首次反超高通，之后多个季度双方交替领先。当前，联发科战略重心从性价比转向“高端化+全域AI+多领域定制”。

此前，在将超算能力落地本地AI计算领域，英伟达已经与联发科合作开发了GB10 Grace Blackwell超级芯片（于2025年8月正式发布），用于DGX Spark产品。

在此基础上，双方将进一步推进RTX Spark项目，面向AI时代新一代消费PC。

联发科还将天玑汽车平台集成英伟达技术，为智能座舱提供先进AI能力与RTX图形能力，可与英伟达DRIVE AGX协同工作。