巨头打响“推理芯片战争”

中经记者 顾梦轩 李正豪
广州 北京报道

大语言模型已从研发走向应用，AI产业的重心也已经从训练阶段转向推理环节。今年以来，随着华为、英伟达和谷歌三大巨头相继发布各自的推理芯片，一场关于AI推理芯片的战争悄然打响。

根据麦肯锡报告，全球AI推理市场规模预计在2028年将达1500亿美元，年复合增长率超40%，远高于训练市场的20%。

南开大学金融发展研究院院长田利辉在接受《中国经营报》记者采访时指出，推理芯片将重塑人类生活，形成云端、边缘、终端三元共存格局，自动驾驶、智能摄像头、语音识别等应用将普及，真正实现“AI赋能千行百业”。

华为、寒武纪等中国企业正在积极布局推理芯片领域，并在国内外市场获得了一定的认可。 视觉中国/图

三大巨头各显“神通”

三家公司的推理芯片在技术路径与市场定位上呈现显著差异，在成本、效能以及应用场景中各有千秋。

在今年9月的2025年华为全联接大会上，华为宣布了昇腾芯片的规划和进展。未来3年，华为开发和规划了三个系列，分别是Ascend950系列、Ascend960系列、Ascend970系列。据悉，华为AI芯片将以几乎一年一代、算力翻倍的速度，围绕更易用、更多数据格式、更高带宽等方向持续演进。

同样是在今年9月，英伟达推出专为大规模上下文处理设计的GPU——RubinCPX，预计于2026年年底上市。

今年4月，谷歌在Google Cloud Next 25大会上推出了其首款Google TPU推理芯片Ironwood。据了解，Ironwood根据AI工作负载需求提供两种尺寸：256芯片配置和9216芯片配置。后者总算力达到42.5Exaflops（百亿亿次），是ElCapitan超算的24倍，单芯片峰值4.614Exaflops。

记者在采访中获悉，上述三家公司的推理芯片在技术路径与市场定位上呈现显著差异，在成本、效能以及应用场景中各有千秋。英伟达凭借CUDA生态与全场景适配能力稳居行业龙头，谷歌TPU以ASIC架构实现云端推理极致能效，华为通过集群技术与存储优化突破制程限制。

在成本控制上，各家公司“各显神通”。星图金融研究院研究员张思远向记者指出，英伟达通过存储技术创新降低单位成本，华为依赖系统级优化分摊成本；效能表现上，谷歌在专用场景领先，英伟达全场景性能均衡。

张思远向记者表示，华为通过UCM推理记忆数据管理器构建三级存储架构，避免重复计算，降低推理成本。但受限于制程工艺，单芯片硬件成本较英伟达产品更高，需通过规模化部署摊薄成本。谷歌聚焦于ASIC架构与云端规模化降本。

“英伟达的推理芯片以强大的计算能力和成熟的CUDA生态系统著称，广泛应用于各种AI场景中。但其产品价格较高，增加了使用成本。”经济学家、新金融专家余丰慧向记者指出。

在效能表现方面，张思远指出，英伟达全场景性能均衡，长上下文推理领先。华为的效能优势则体现在集群算力突破和行业场景深度优化。谷歌则通过HBM（高宽带内存）容量与互联带宽驱动云端效能。

在应用场景方面，张思远表示，英伟达实现全场景覆盖，消费级与企业级并重。华为则聚焦国内行业，在国内政务、金融、医疗场景市场占有率较大，依托昇腾生态，参与多地智算中心建设，但海外市场拓展受限，消费级场景渗透率不足。

对谷歌而言，张思远指出，谷歌以云端服务为主导，搜索与AI模型协同。其中搜索业务依赖TPU推理加速，Cloud业务提供Gemini推理服务，支持企业级MoE模型部署。但硬件仅通过云端开放，企业本地化部署需求难以满足。

“谷歌的TPU以其高度定制化的硬件设计和出色的机器学习性能占据一席之地，不过，谷歌TPU的应用范围较窄，主要针对自家服务和特定合作伙伴。”余丰慧补充道。

谈及华为推理芯片的发展状态，田利辉指出，华为实质上形成了“算力积木”架构，进而实现灵活扩展，成本效率平衡。

行业规模将突破3000亿元

中商产业研究院分析师预测，2025年中国AI推理芯片相关产品及服务行业市场规模将达到3106亿元。

推理芯片的发展将极大地促进人工智能技术在日常生活中的普及与深化。张思远指出，首先，推理芯片的发展可能带来效率革命，重构服务响应范式。推理芯片的能效提升将推动AI应用从“实验室”走向“日常生活”。例如，金融客服系统通过华为UCM技术实现通话分析时间从120秒缩短至10秒；医疗领域，推理加速方案使医学影像分析效率提升6倍，基层医院也能快速获取诊断支持。

其次，推理芯片可能带来成本普降，降低AI应用门槛。例如，搭载国产推理芯片的AI学习机大量出货，使优质教育资源向三、四线城市渗透。

“最后，可能带来产业升级，”张思远表示，推理芯片与边缘设备结合，推动消费电子形态革新。如AI眼镜通过低功耗推理芯片实现实时翻译、视觉识别，人形机器人依赖高能效推理芯片完成环境感知与运动控制。

全球范围内，推理芯片正处于快速发展阶段，各大科技公司纷纷加大研发投入，试图在这一新兴市场分得一杯羹。

中国推理芯片市场前景广阔，发展潜力巨大。中商产业研究院发布的《2025—2030年人工智能芯片行业市场调研及投资前景预测报告》显示，中国AI推理芯片相关产品及服务行业市场规模由2020年的113亿元增至2024年的1626亿元，期内复合年增长率为94.9%。中商产业研究院分析师预测，2025年中国AI推理芯片相关产品及服务行业市场规模将达到3106亿元。

以华为为例，华为轮值董事长徐直军曾在全联接大会上公开表示：“有了昇腾芯片为基础，我们就能够打造满足客户需求的算力解决方案。从大型AI算力基础设施建设的技术方向看，超节点已经成为主导性产品形态，并正在成为AI基础设施建设的新常态。”“但仅有强大的单芯片远远不够，如何将成千上万张芯片高效地连接起来，形成一个协同工作的‘超级大脑’才是挑战。”诺安基金科技组基金经理刘慧影表示，为此，华为重磅预告了三款超节点产品，并面向超节点创新性地推出了“灵衢”全光互联协议，且宣传将其技术规范开源。这一举措被认为是算力互联领域的颠覆性的突破。据悉，“灵衢”采用光传输技术，可实现数据高速流转。更重要的是，基于超节点与该协议打造的Atlas950超节点，其算力水平在未来数年内有望保持全球领先地位。

先进制程有待突破

国内的整体技术水平与国际领先企业相比仍存在一定差距。

受访人士向记者表示，随着国家政策的支持以及市场需求的增长，本地企业也在积极布局该领域，如华为、寒武纪等企业已推出多款自研推理芯片，并在国内外市场上获得了一定的认可。然而，整体技术水平与国际领先企业相比仍存在一定差距。

张思远指出，第一，在技术指标方面，制程与单芯性能仍有巨大进步空间。英伟达RubinCPX采用3nm制程，华为昇腾910B仍依赖7nm工艺，单芯片算力差距约3倍。第二，在生态建设方面，开发者生态壁垒显著。英伟达CUDA生态积累超15年，全球超400万开发者支持；华为CANN架构开发者数量突破50万，但工具链完善度仍需提升，部分企业因迁移成本高而选择继续使用英伟达方案。第三，在市场渗透方面，中国企业国际份额与场景覆盖不足，相比之下，英伟达推理芯片全球市场占比超70%，覆盖云厂商、消费电子等多领域。

“中国芯片主要集中在国内政务、安防等toG场景，海外市场拓展缓慢，且高端消费级市场仍以进口为主。”张思远说。

北京社科院副研究员王鹏表示，国产推理芯片在政企、安防领域的渗透率较高，但高端训练芯片与复杂模型支持能力不足。

“中国芯片企业正通过‘应用场景驱动—数据积累—算法优化—芯片迭代’闭环加速追赶，本土品牌渗透率从30%持续提升，2025年市场规模将突破1530亿元。”田利辉说，一些国内企业正在探索存算一体和3Dmemory技术，未来将突破大规模集群互联瓶颈，实现从“跟跑”到“并跑”的跨越，真正成为全球AI基础设施的核心力量。

萨摩耶云科技集团首席经济学家郑磊向记者表示，中国推理芯片正在从“可用”向“好用”阶段过渡，但在先进工艺、存储带宽、软件栈与极致性能场景上仍落后全球顶尖水平，下一步，行业需在RISC-V开源指令集、Chiplet国产封装线、AI编译器框架及行业芯片协同定义上加速迭代，方能真正与世界头部公司在同一梯队竞争。

海南华铁36.9亿元算力大单终止：跨界转型的市场考验

中经记者 秦枭 北京报道

在国庆中秋长假前夕，海南华铁（603300.SH）宣布，因市场环境和供需状况出现重大变化，其子公司华铁大黄蜂与杭州X公司签订的价值36.9亿元（含税）的算力设备合同已终止。这一事件引起了市场和监管机构的关注。为了增强投资者信心，10月8日晚间，海南华铁第二大股东兼董事总经理胡丹锋宣布，将停止先前公布的股份减持计划。

《中国经营报》记者致电致函海南华铁询问具体原因，截至发稿，对方未作回复。多位业内人士在接受记者采访时表示，海南华铁作为算力领域的跨界新兵，在算力积累方面相对薄弱，面对市场环境的急剧变化和供需状况的重大调整，难以迅速做出有效的应对策略。此外，算力市场竞争激烈，老牌企业占据着比较大的市场份额和资源优势，海南华铁这样的跨界新兵想要分一杯羹本来就困难重重。

从热捧到悄然终止

在2025年的算力市场浪潮中，海南华铁的入局可谓声势浩大。3月，其全资子公司华铁大黄蜂与杭州X公司签署的《算力服务协议》成为市场焦点。这份长达5年、总金额预计达36.9亿元的协议，就像一颗投入平静湖面的石子，在算力市场上激起千层浪。要知道，这一金额几乎相当于海南华铁2025年中期收入的1.32倍。

公告一经披露，海南华铁的股价飙升，在3月5日至7日，收盘价格涨幅偏离值累计超过20%。其市值也随之水涨船高，一度突破250亿元大关，成功跻身市场热捧的“算力黑马”行列。当时，投资者纷纷对海南华铁的未来充满期待，认为它在算力领域的这一重大布局，将为公司带来全新的发展机遇，一时间，海南华铁成了算力赛道上备受瞩目的“新星”。

然而，好景不长，这场算力领域的“美梦”在短短半年之后便戛然而止。2025年9月30日晚，海南华铁发布公告，宣布终止与杭州X公司的合作。公告中称，由于协议所涉交易及设备的市场环境、供需情况发生了较大变化，并且自协议签订以来，未收到任何采购订单，基于这些因素，双方最终决定终止合作。

“公告中提到的合作终止原因‘市场环境、供需情况变化’，反映出算力租赁行业深层次的市场矛盾。”新智派新质生产力会客厅联合创始发起人袁帅表示，“当前全球算力硬件供应链方面，高端GPU供应紧张，受到芯片制造技术、地缘政治等多种因素影响，供应的不稳定导致算力租赁企业获取硬件的难度加大，进而导致成本增加，最终影响服务价格和供应能力。”

上海证券交易所也迅速做出反应，同日下发监管工作函，直指重大合同终止相关事项，要求海南华铁就合同终止的原因、对公司经营的影响等问题作出详细说明。值得注意的是，今年3月，曾有投资者询问关于该订单取消的市场传闻，公司当时回复称“正常履行中”。

受上述消息影响，10月9日开盘，海南华铁直接跌停至8.71元/股。

天使投资人、人工智能专家郭涛分析道，传统企业跨界高科技领域常陷入技术认知偏差与资源错配的困境。海南华铁的案例显示，其作为建筑设备租赁商，缺乏算力行业技术沉淀、运营经验及产业链协同能力，仅凭概念驱动就签约巨额订单，暴露对出算力重资产属性（硬件迭代投入）、技术门槛（异构计算架构适配）及客户需求（政企定制化需求超70%）的预估不足。

跨界算力的信任危机

实际上，智算业务正在被海南华铁塑造成“第二增长曲线”，自2024年布局以来，承载着公司转型发展的厚望。

2024年5月7日，公司发布《关于投资智算中心建设的公告》，创新拓展裸金属算力服务模式，正式布局算力业务。公司主要通过融资租赁方式采购硬件设备，并将相关设备部署至指定的智算中心机房后，并提供客户业务所需的各类运营运维服务，以此向客户收取相关服务费用。

2024年，海南华铁在算力领域积极拓展，累计签订算力服务金额达到24.75亿元，这一成绩在当时无疑给市场带来了极大的想象空间。

不只是海南华铁，随着人工智能技术的全面爆发，算力需求更是呈现出指数级增长态势，算力租赁产业顺势崛起，成为数字经济领域的新风口。众多企业纷纷布局，除运营商、云厂商外，不乏一些跨界玩家，比如“味精大王”莲花控股（600186.SH）、房地产开发商大名城（600094.SH）、染料生产商锦鸡股份（300798.SZ）等，均宣布布局算力新赛道，推进智算中心项目。

袁帅表示，算力租赁作为新兴领域，具有广阔的市场前景和增长潜力，传统企业跨界进入可能获得新的利润增长点，提升企业的竞争力和市场地位。

然而，算力租赁并不是“建”和“租”那么简单，“现在很多地方的智算中心商业模式是单一地购买了GPU，组件集群，出租算力。”一地（广州）智算中心的运营总监张勇对记者表示，“这是一种非常粗放的商业模式，尤其是在当地没有AI应用场景的情况下，完全不具备抵抗市场波动的能力。”

郭涛认为，当前，算力租赁行业结构性矛盾凸显，供给端受英伟达H系列GPU交付延期、台积电CoWoS产能瓶颈限制，实际可交付算力远低于合同理论峰值；需求端碎片化，东部沿海城市自建集群挤压第三方租赁商空间；定价体系崩塌，头部云厂商通过“自研芯片+规模效应”将FP16算力单价压至0.8元/PFLOPS（每秒执行1千万亿次浮点运算）以下，中小租赁商陷入“量价齐跌”的困局。

今年以来，多家上市公司宣布终止此前签订的算力租赁服务。2025年4月29日，莲花控股的控股孙公司浙江莲花紫星智算科技有限公司与上海国企签订《高性能算力服务合同补充协议》，双方一致同意解除原合同，原合同中莲花紫星向上海X国企提供1545PFLOPS（BFLOAT16）高性能算力服务及算力运维服务，服务期限为五年，合同总价为5.5亿元。

2025年9月26日，京源环保公告称，公司全资子公司京源云计算2025年5月与客户R公司签署的算力集群建设项目承包合同因在执行过程中存在理解差异及外部客观条件限制，双方经协商一致签署了终止协议，原合同预计的3.2亿元营业收入将不再纳入公司未来业绩预期。

袁帅认为，上述案例为众多寻求向算力、AI等新兴领域转型的传统企业提供了重要警示。传统企业在转型新兴科技赛道时，不能仅看到新兴领域的光明前景而盲目跟风，要充分认识到转型过程中面临的巨大挑战和风险。