专用芯片破解Token高成本难题

图为清微智能展示可重构超节点系统

本报记者 姬晓婷 实习生 贺紫依

打开AI聊天软件、代码助手、智能办公工具，每一次问答、每一段回答生成，背后都在消耗一种叫Token（词元）的“AI流量”。Token作为计量AI输出的最小单元，正逐渐成为AI产业的基础计价单位，其单位算力成本直接决定了AI服务商的盈利空间。

在刚刚结束的2026世界人工智能大会（WAIC）上，记者观察到，“降低Token成本”正在成为众多芯片与服务器厂商的共识。“性价比”正在“上桌”，成为继算力指标、适配模型规模之后的竞争新维度。

高昂算力成本待破局

提升单块芯片算力效能和降低整机机房建设成本，可以助力AI应用商降低算力成本。

随着AI智能体、大模型规模化落地，词元消耗量大幅攀升，不少公司甚至陷入“用得越多亏得越多”的困境。中昊芯英创始人、CEO杨龚轶凡在接受采访时表示，智能体调用API时会携带全部历史对话上下文打包传输，即便用户提问量仅千词左右，请求包带来的Token量也可能达到20万。

现阶段，算力支出已然成为AI公司总成本中最高的一项，多数AI软件服务商毛利率不足40%。“如果算力成本居高不下，很多AI业务会陷入亏损、商业模式无法闭环。”杨龚轶凡坦言，“但如果能把输出词元的算力成本减半，客户利润将成倍增长，毛利率将从40%提升至70%。”

从AI推理市场来看，算力负载可被划分为两类底层逻辑完全差异化的算力需求，对应两类市场服务：高实时交互的高速词元（Fast Token）、离线批量处理的慢速词元（Slow Token）。各行业、各应用场景对AI服务的要求差异显著。

国内某算力芯片企业在接受《中国电子报》记者采访时也表示，若针对不同推理业务分别搭建独立算力集群，会出现硬件长期闲置、硬件与电力投入翻倍的问题，还会抬高多架构配套软件的运维成本。

为助力AI应用商降低算力成本，国产算力芯片厂商摸索出两条实现路径：

一是提升单块芯片算力效能。重新设计芯片内部计算结构，减少无效运算与耗电，在同等电力条件下产出更多词元；同时完成芯片与大模型软件深度适配，避免算力空转造成的资源消耗。

二是降低整机机房建设成本。优化服务器之间的数据传输架构，减少交换机、高端网卡等高价配套设备投入。

此外，为控制芯片本身的生产成本，部分算力芯片企业也在通过远期锁单、提前备货等方式，平抑由晶圆、存储、封装原材料涨价带来的成本波动；采用Chiplet芯粒方案，将大芯片拆解为多颗小型芯粒分开流片，以应对大尺寸单片晶圆流片难度高、良品率偏低等问题。

革新芯片架构

TPU架构的行业价值持续放大，相比通用GPU拥有更大的性能和性价比提升空间。

现阶段，GPU仍然是数据中心数量最多、应用最广泛的芯片类型。而在AI时代，杨龚轶凡对GPU发挥的作用有着不同的看法：“通用GPU架构如同万能工具箱，虽然什么都能干，但处理基于Transformer架构的大语言模型任务的效率偏低。”在他看来，GPU技术经过30年迭代，底层核心架构很难再做颠覆性革新；同时完整CUDA软件生态已经成为GPU厂商最坚固的护城河，后来者很难实现弯道超车。

相比之下，TPU的创新空间更为广阔。“只要AI产业持续扩张，Transformer与Attention算子仍是大模型底层核心计算单元，TPU架构的行业价值就会持续放大，相比通用GPU拥有更大的性能和性价比提升空间。”杨龚轶凡表示。

未来推理算力会做专业化拆分，一类是读取全部历史对话的预填充（Prefill），另一类是输出内容的解码（Decode），二者运算逻辑割裂。杨龚轶凡表示，通用GPU依靠一套硬件兼顾两类运算，两端性能均无法做到最优。芯片存储层面，通用GPU内置高速缓存容量有限，需要频繁读写外部显存调取数据，既拖慢运算速度，又产生额外能耗。“中昊芯英推出的二代‘须臾’芯片通过Prefill、Decode硬件分离适配，可采用1颗预填充芯片搭配1~2颗解码芯片协同运算，高度适配智能体长对话场景，能够显著降低重复计算带来的电力消耗。”杨龚轶凡说道。

清微智能自研可重构数据流架构，实现数据流转与运算同步执行，硬件资源利用率大幅提升。

降低基建开支

大型AI算力集群中，服务器间数据传输依赖交换机、高速网卡设备等硬件。在一个数据中心中，数据交换相关硬件的成本约占到整个系统的40%。

在2026世界人工智能大会上，记者注意到，清微智能首次展出了自研的REX81 Supernode国产4K超节点方案。该方案采用独创的算力网格互联技术，通过光纤实现芯片间、服务器间无交换机直接连接。这样的调整使集群互联硬件的综合成本较传统交换机架构下降90%。

“硬件参数不再是唯一竞争力，软件生态、落地规模、客户口碑成为更重要的竞争维度。”清微智能董秘、CFO张磊在接受《中国电子报》记者采访时介绍道，“清微智能自研可重构数据流架构，实现数据流转与运算同步执行，硬件资源利用率大幅提升；单芯片可同时支撑时延要求不同的推理任务，降低硬件重复部署成本。”

三星HBM5将采用2纳米GAA工艺 速度较HBM4E提升50%以上

本报讯 7月27日，三星电子半导体透露，下一代HBM5内存的基础芯片将采用GAA（全环绕栅极）结构的2纳米制程，并实现更高集成度的TSV硅通孔互连。据悉，HBM5的运行速度目标较HBM4E提升50%以上，因此在基础芯片上必须引入能在速度与电力效率两个维度均实现大幅跃升的最先进制程。这是三星迄今对外阐述的最明确的HBM5技术规格与制程路线图。

三星同时披露，其代工部门已于2025年下半年启动2纳米第一代工艺量产，计划2026年下半年量产性能与良率进一步提升的第二代2纳米工艺，已获得特斯拉车规级2纳米产品订单。三星将HBM产品定位为“交钥匙”解决方案——核心芯片由内存部门制造，基础芯片由代工部门生产，先进封装由TSP部门完成，全流程内部协同。以HBM4基础芯片为例，从首批样品便同时达成性能与良率目标，整个过程仅历时约三个月。

HBM5路线图的明确化，恰逢AI算力供应链从“按需采购”转向“超长期产能锁定”的关键节点。同日，Anthropic正式向SK海力士提出存储半导体供应需求，用于制造其自研芯片。此外，SK集团还计划在未来五年内额外寻求2500亿美元的全球存储供应合作，进一步巩固其在AI存储供应链中的核心地位。

（李 平）

芯和半导体与诺瓦星云宣布战略合作

本报讯 7月26日，全球规模最大的电子设计自动化（EDA）行业盛会DAC 2026在美国加利福尼亚州长滩市开幕。本届大会以“AI For EDA”为核心议题，全球主要EDA厂商均在会上展示了AI智能体（Agent）技术的最新应用。国产EDA企业代表芯和半导体在现场联合诺瓦星云科技股份有限公司发布战略合作声明，共同致力于AI+EDA赋能LED显控系统设计。

本次合作聚焦LED显控产品的板级（PCB）AI设计与多物理场协同仿真，覆盖诺瓦星云端到端全链路视频显控解决方案中核心产品的PCB设计、高速信号完整性（SI）仿真优化、电源完整性（PI）与电源分配网络设计、板级热仿真、散热设计。通过AI驱动的板级多物理场协同仿真，全面提升产品在复杂应用场景下的信号质量、供电稳定性、散热可靠性。

此次合作的核心价值，在于芯和半导体与诺瓦星云各自领域AI能力的融合：芯和半导体以多物理场AI智能体驱动板级设计仿真效率提升，诺瓦星云则在LED显控产品定义与工程经验上有深厚积累。双方合力，将推动AI技术在LED显控产品板级设计验证中的落地应用，助力诺瓦星云在新一代控制器、接收卡等产品的板级设计效率与可靠性上进一步提升，也为国产EDA在新兴应用领域的拓展提供又一实证案例。

（张 宇）

英特尔与蓝思科技达成战略合作 玻璃基板封装取得新进展

本报讯 近日，英特尔宣布与蓝思科技（Lens Technology）达成战略合作，探索基于玻璃基板的先进半导体封装技术，旨在支持下一代人工智能、数据中心、高性能计算系统。

此次合作将英特尔在先进芯片封装领域的专长与蓝思科技在精密玻璃加工和规模化量产制造能力相结合。两家公司表示，此次合作将探索新的封装方案，以期提升未来计算平台的性能、互连密度、电源效率。

据了解，此次合作凸显了先进封装技术对于半导体关键领域创新的重要性，人工智能算力需求爆发，倒逼芯片向高性能、高能效迭代。这也反映出材料工程和制造领域的合作对下一代半导体技术发展的重要作用。

半导体公司不再仅仅关注晶体管尺寸的缩小，而是更多地投资于封装技术，以提升系统综合性能。玻璃基板有望取代传统有机基板，因为它具有更好的尺寸稳定性、更精细的互连能力、更优异的电气性能，能够满足日益复杂的芯片封装需求。

英特尔和蓝思科技计划结合双方在先进封装、精密制造、玻璃基板研发、工艺创新方面的优势，推动相关技术的发展。

英特尔首席执行官陈立武表示：“随着人工智能不断突破性能、规模、效率瓶颈，先进封装正成为半导体创新的关键驱动力。英特尔在半导体架构和先进封装技术方面拥有深厚积累，而蓝思科技则贡献了世界一流的精密玻璃加工、激光制造、量产能力。我们携手合作，探索先进封装的新方法，将助力人工智能时代下一代系统级创新。”

同时，两家公司也看到了将合作拓展到封装技术以外的领域的机会。未来的合作重点可能包括人工智能个人电脑组件、数据中心服务器、散热、结构解决方案，以及用于人工智能机器人、智能工业设备、边缘计算的硬件平台。

蓝思科技以其在玻璃材料和精密制造方面的专业性而闻名，该公司认为其能力可以与英特尔的半导体设计和封装技术形成互补。

本次战略合作协议落地也印证了人工智能正在重塑芯片架构和制造标准，先进封装技术正成为半导体企业战略竞争的战场。

（王 浩）