HBM承压，四大技术破解AI算力瓶颈

本报记者 张心怡

HBC、NVHBM、zHBM、HBF……近期，高通、英伟达等头部芯片设计厂商（Fabless）与三星、SK海力士、闪迪等存储原厂持续“输出”新的存储概念。乍一看，它们的名称似乎都与HBM存在着若隐若现的关联，事实上也确实如此——它们的出现源于半导体厂商对于内存的一种“集体焦虑”：在智能体、大模型引领的计算变局中，HBM作为当前主流的内存方案，正在面临越来越大的缓存容量和词元（Token）成本压力。产业链上下游的企业正在以各自的方式探索破局之道。

图为三星zHBM概念图

HBM是AI算力集群的主流内存方案，正在承受超长上下文和算力规模“爆炸式”增长的压力。

AI系统的能力标尺变了

在ChatGPT登场的2022年，“算力”是AI部署乃至半导体增长的叙事主线，谁的芯片算得快、谁的集群堆料足，谁就更容易在资本市场和流量声势上占据上风。然而，当AI从尝鲜期走向实干期、聊天的机器人变成干活的智能体，AI系统的能力标尺，已经从“算得快不快”的单一逻辑，走向实际的工程问题。

首先是数据搬运正在成为AI系统的新瓶颈。在AI推理阶段，模型需要经历两个阶段来回应用户提问：在预填充阶段，模型读取提示词，此阶段需要完成大量的算术运算，适合在密集矩阵引擎上执行；但在解码阶段，模型一个一个地生成Token，每生成一个Token，都需要从内存中读取模型参数和累积的上下文。当大量数据在计算单元与内存之间来回搬运，计算组件就会浪费大量时间、能耗和成本，空置着等待操作数（数据）从内存传过来。

其次是KV Cache（模型为避免重复计算而临时存储的键值对）急速膨胀。随着上下文窗口增长、对话轮次增多，KV Cache呈指数级增长。中信证券研报显示，AI从“简单对话”向“智能体”演进，驱动上下文需求从8K激增至1M Tokens，单用户FP16精度下的KV Cache显存占用从5GB飙升至640GB以上。尽管AI模型和云计算厂商持续推出减少每Token生成KV Cache量的技术，但最长上下文窗口正在以每年约30倍的速度攀升（Epoch AI数据），带来显著的显存容量压力。

最后是Token成本。数据搬运带来的功耗、封装集成度上升带来的成本压力、系统复杂度的持续提升，都在影响Token生产的成本效益。

而以上趋势，正在从不同层面考验AI系统的内存。当前，HBM是AI算力集群的主流内存方案。HBM将DRAM芯片垂直堆叠，再通过先进封装与计算引擎（GPU、XPU等）紧密互联，使计算与内存通过数据总线直连，从而在带宽、功耗、集成度上取得较好的平衡。但也需看到，HBM正在承受超长上下文和算力规模“爆炸式”增长的压力。

比如，智能体、大模型的参数规模和上下文长度“狂飙”，正在“考验”HBM的能效。尽管HBM通过TSV实现各层DRAM连接，并与GPU封装在一起，大幅缩短了数据的物理传输距离，且持续提升接口宽度，已经是当前存储产品在带宽上的“最优解”，但依然无法避免数据在算与存之间来回搬运带来的功耗、热量和时间成本。

KV Cache还在“压力”内存容量，HBM的优势是高带宽、低时延，但受限于层数和封装面积，在单卡容量上不占优势。

Token成本，也在“牵制”HBM的演进。虽然HBM可以通过提升接口带宽来改善数据搬运瓶颈，但更宽的接口会增加I/O面积和封装尺寸，并增加物理引脚和信号线，从而影响Token成本。

针对数据搬运带来的能效问题，头部Fabless和存储原厂开始对HBM进行“优化”或“补充”。

芯片和存储厂商纷纷“想辙”

面向AI系统的一系列新需求和HBM的“吃力”之处，头部Fabless和存储原厂开始对HBM进行“优化”或“补充”。

HBC：不是把数据搬向计算，而是让计算走向数据

针对数据搬运带来的能效问题，高通给出的解法是HBC（高带宽计算）。其核心理念是“近存计算”，将计算单元部署在数据近旁，让计算在数据所在之处执行，进而省去了数据离开内存、跨越芯片边界的环节。

在物理实现形式上，HBC将近内存加速器（NMA）从SoC中分离出来，直接放在垂直堆叠的LPDDR DRAM下方，通过硅通孔（TSV）连接各层，取代了传统长边HBM接口。

如此一来，数据密度最高的计算直接在内存封装中执行，仅经过处理后的精简结果需要向外传递。SoC则与HBC各展所长：SoC继续承担复杂、灵活的调度任务，HBC单元接手传统方式下数据搬运成本较高的运算。同时，近存加速器可以直接使用LPDDR堆栈的带宽，而不是通过外部接口的传输带宽。高通称，与竞品已公布的板卡级标准化产品参数相比，HBC技术支持的每瓦特带宽相比HBM技术提升6倍。

当前，一个Token的生成涉及前馈网络（FFN）层和注意力机制主导的两种计算，前者的数据相对静态、规律，后者则需要不断读取上下文，占用空间随序列长度而增长。高通对几种主流存储方案进行评估后认为，片上SRAM在小批量FFN处理中领先，HBM是全能型选手，而HBC有望成为注意力密集型工作负载的理想方案，并且随着上下文长度持续增长而越来越有优势。

在产品化进度方面，搭载第一代HBC技术的高通飞龙AI250预计于2027年年中商用出样，其单卡带宽速率可达133TB/s，与采用LPDDR5X的AI200相比，有效内存带宽提升18倍。

NVHBM：给计算单元腾出更多面积

8月26日，英伟达提出NVHBM，并宣布亚马逊Annapurna Labs将成为首个在NVLink Fusion基础上协作开发NVHBM技术的合作伙伴。

不同于高通将近存加速器“拆出来”放在内存堆栈下，英伟达的“换位”对象是内存控制器。

在主流的XPU（异构计算加速器）+HBM模式下，内存控制器大多集成在XPU中。NVHBM的策略是将英伟达定制的内存控制器集成到HBM基底裸片。英伟达称，相比标准HBM4E，NVHBM每堆栈可提供最高30%的内存带宽提升，从而使计算单元更快速地获得数据供给，提升用户的Token吞吐量。在能效方面，HBM功耗降低最高达15%，这意味着其能够在数据中心的XPU集群中节省更多功耗。

NVHBM的潜力还体现在能为计算芯片腾出更多面积（可为XPU计算芯片释放最高25%的面积），从而为AI芯片的功能布局预留更多空间。标准HBM的带宽提升，多基于更宽的接口连接，却增加了整体封装尺寸。NVHBM采用定制基底裸晶，通过将内存控制器移至3D HBM堆栈并集成定制PHY，将PHY和支持电路面积减少了最多67%（相比JEDEC HBM4E）；更窄的接口还简化了中介层布线，在整个布局上提供最多80%的更多可用硅面积。英伟达表示，随着AI工作负载的多样化，腾出的芯片面积为超大规模数据中心运营商提供了更大的灵活性，使其能够针对推理服务、推荐系统、多模态流水线或内部训练工作负载优化XPU。

在产品化路径上，英伟达正在建立一套标准的NVHBM实现方案（可由多家内存供应商提供），并通过NVLink Fusion为用户提供与内存制造商共同验证的NVHBM基底裸晶，以减少集成和验证环节的瓶颈。

zHBM：用垂直堆叠缩短传输路径

在芯片厂商思索破局之道的同时，存储原厂也看到了AI系统在数据搬运上的掣肘。为了减少AI处理器与内存之间的数据传输距离，三星在今年8月举办的FMS 2026上，提出了全新内存架构zHBM。

不同于传统设计中将HBM与处理器并排放置，zHBM将HBM垂直堆叠在AI加速器上方。

三星在FMS 2026上称，采用zHBM的下一代接口系统预计将实现约8倍于HBM5的性能。借助下一代晶圆键合技术，zHBM可实现超过HBM5 10倍的内存密度，同时能效提升3倍、热阻降低一半以上。

在定制化设计层面，三星支持在内存与AI加速器之间的中介层集成客户定制IP，以扩展内存容量并增强加速器性能。三星还计划通过zHBM进一步优化AI处理器与内存的集成，以最大化AI系统性能。

HBF：补充HBM容量短板

随着KV Cache膨胀，内存容量的压力与日俱增。对此，SK海力士与闪迪正在牵头推进HBF（高带宽闪存）的标准化进程，其目标是在保持类似HBM的高传输带宽优势的同时，将存储容量做上去。

从架构来看，HBF采用了类似HBM的封装方式，同时，将堆叠对象从DRAM芯片换成了NAND，也就是用TSV将NAND垂直堆叠，并通过中介层与GPU连接。如此一来，它既继承了HBM靠近计算核心的“区位优势”，又保留了NAND高存储密度、非易失的特点。闪迪称，在封装尺寸、堆叠高度、功耗与HBM4基本相当的前提下，HBF容量可达HBM的8~16倍，且成本相近。

目前，SK海力士、闪迪等存储原厂对HBF的定位，更像是HBM的“容量补丁”。比如SK海力士提出了H3混合架构，将HBM与HBF整合在同一中介层并同时连接到GPU。其中，HBF负责读取模型权重和预计算的KV缓存，HBM负责存储动态的、频繁更新的数据，从而将HBM从容量负担中解放出来，专注实时任务处理。

与此同时，美光也在尝试将NAND“推向”计算单元。产业资讯显示，美光正在探索近存NAND方案，其核心理念是让NAND以更短的路径连接GPU，比如将NAND与GPU共同封装或放置在GPU附近的板卡上，用于处理LLM权重、可复用的推理数据以及KV缓存等。

各家芯片和存储厂商围绕内存设计理念的一系列动作，反映了AI基础设施构建的底层逻辑正在迭代更新，纸面算力不再是核心标尺，数据搬运能力、Token吞吐能力、每瓦特性能等指标正变得更加关键和紧迫。但哪种新的存储路径能够落地，还要看各家企业与产业链的协同能力，如何构建易于部署的生态系统，以及整体的系统效能和成本。并且，AI的形态还远未稳定，或许未来几年又会有新的AI范式，继续倒逼存储产业的转型升级——但可以肯定的是，这已经不仅是存储原厂要面对的课题。

华为再次发布韬定律论文 回应堆叠芯片发热问题

本报讯 9月4日，华为半导体负责人何庭波在中国科学院科技论文预发布平台ChinaXiv上发布了论文《Huawei's τ Chip Was Supposed to Melt?》（华为韬芯片本应熔化吗？）。这是继5月25日发布韬定律V1版本、7月3日更新V2版本之后，何庭波在不到四个月时间内第三次就韬定律发表学术论文。

此次论文更新，是华为试图向外展示“τ芯片”是如何通过重构电路结构、缩短信号传输距离、压缩传输延迟，将“时间维度革新”转化为芯片性能、功耗、温控的突破点，同时，回应堆叠芯片“高密度等于高发热”的行业质疑。

在传统认知中，将更多晶体管堆叠到同一面积内，单位面积的功率密度必然上升，热量难以散出。但何庭波的核心观点并不是3D堆叠天然更节能，而是认为过去对芯片功耗的关注，可能过多集中在“晶体管计算”本身，却低估了数据在芯片内部移动所消耗的能量。

“我们能预见的τ的每一个未来，都成于功耗，亦败于功耗。时间余量是手段，能量才是锦标。这条定律未来十年的评判标准，不会是折叠堆叠跑得多快，而是它们奔跑时燃耗多低。那颗本该烧毁的τ芯片，反而运行得更冷，不是因为折叠才变冷，而是因为折叠本身。归根结底，τ（韬）的用武之地不止一处。”何庭波在论文中表示。

何庭波认为，芯片消耗的能量，并不只来自“计算”，还来自“移动”。她在论文中用了一个很形象的比喻：一个人在工作日里消耗的能量，并不主要来自坐在办公桌前工作的那几个小时，通勤同样需要消耗大量能量。芯片也是如此，晶体管完成计算需要能量，而数据在晶体管、模块和不同电路之间长距离传输，同样需要消耗能量。

按照论文中的描述，在典型的智能手机工作负载中，动态功耗可能占总功耗的约90%；而动态功耗除了晶体管开关，还包括信号在金属连线中的传输。随着制程不断发展，互连电容对能耗的影响越来越重要，在某些情况下，真正昂贵的并不是晶体管“思考”，而是数据在芯片内部“通勤”。

韬定律与LogicFolding（逻辑折叠）技术的破局点，正在于此。何庭波表示，τ定律，并不是简单地把芯片做成多层，而是希望改变信号传播的路径。在传统平面芯片中，一些关键连接需要横向跨越较长距离；LogicFolding则把部分电路重新设计成上下两层，让原本较长的水平连接转化成更短的垂直连接。

论文称，在一些典型核心中，这种方式可以将连线长度缩短约20%，关键路径上的部分连线最多可以缩短70%。在一个处理模块中，时钟布线长度缩短28%，时钟缓冲器数量也从约4.36万个减少到1.9万个。从这个角度看，所谓“折叠”并不是简单地把更多晶体管堆在一起，而是在重新安排晶体管之间的关系：少让数据跑远路。 （华 讯）

激光技术新突破 有望助力光子计算发展

本报讯 近日，英国南安普敦大学发布新闻公报称，该校研究人员领衔的国际团队在室温下实现多个微型激光器自发同步，形成一种相干且可重构的光源。这项成果可能成为迈向成本更低、可重构且更易普及的光子计算技术的第一步。

公报介绍，光子计算机利用光子而非电子处理数据，具有速度快、可同时处理更多数据以及能耗较低等优势。

研究团队在一种基于液晶和有机染料分子的光学微腔中生成彼此空间分离的微型激光器，并使这些激光器自发同步，在相同频率和相位下作为统一的相干光发射体工作，由此形成被称为“超模”的宏观集体状态。

此前，这类激光控制现象大多是在低于零下250摄氏度运行的专用半导体微腔中观测到，新平台则可在室温下运行。新平台的另一个关键优势在于其电可调谐性，施加微小电压可以重新定向微腔内的液晶分子，进而调节光的传播方式和激光器的运行模式。此外，该平台还具有光学可重构性，微型激光器在同一器件中的相对位置可以重新进行调整。

论文第一作者、南安普敦大学光电研究中心的德米特里·多夫任科说，研究结果表明，实现这种光的“集体行为”无须复杂的光—物质耦合态或极低温环境。新平台更简单实用，并具有光学可重构、电可调谐以及常温环境下稳定运行等特点。

据介绍，该系统采用成熟材料，为未来开发实用器件提供了一条有潜力的路径。 （文 编）

第二季度全球半导体设备市场 同比大增23%

本报讯 当地时间9月3日，国际半导体产业协会（SEMI）发布的最新数据显示，2026年第二季度，全球半导体设备开票销售额达到405.3亿美元，同比大幅上涨23%，环比增速同样达到11%，行业销售额连续两个季度刷新历史最高纪录，AI算力基础设施建设正成为本轮设备采购热潮的核心驱动力。

SEMI总裁兼首席执行官阿吉特·马诺查（Ajit Manocha）表示，连续两个季度设备销售额创下新高，代表着全球晶圆制造企业正在持续加大先进产能布局，以此匹配人工智能芯片快速上涨的市场需求。跨大洲同步增长的数据也证明，算力建设已经不是单一地区的产业红利，半导体制造能力将成为下一阶段全球AI创新竞争的底层基石。

本轮半导体设备景气上行，本质上是全球AI投资热潮沿着产业链自上而下传导的结果。过去两年，各大云厂商、芯片设计企业持续上调资本开支，谷歌、微软、亚马逊等科技巨头纷纷公布千亿美元级别的算力建设计划，AI服务器、高性能GPU、高带宽内存HBM订单持续放量。而芯片产能的扩张，最终将压力传导至上游半导体设备端，先进逻辑制程、3D先进封装、HBM存储生产线成为当下晶圆厂采购设备的三大热门方向。

下游晶圆制造大厂的资本开支计划已经印证了这一趋势。作为全球最大的晶圆代工厂，台积电将2026全年资本支出上调至520亿~560亿美元，重点投向2纳米先进制程与AI芯片先进封装产线。韩国方面，三星、SK海力士启动大规模存储扩产计划，目标瞄准高附加值HBM产品；美光科技同样上调长期资本开支，加码新一代DRAM产能建设。英特尔在美国、欧洲同步推进先进晶圆厂落地，全球范围内新建晶圆厂项目集中开工，共同推高设备市场整体需求。 （程 迅）

社长：张立 社址：北京市海淀区紫竹院路66号赛迪大厦18层 邮编：100048 每周二、五出版 周二8版 周五8版 零售4.50元 全年定价420元 广告部：010-88558848/8808 发行部：010-88558777 广告许可证：京石工商广登字20170003号 发行单位：中国电子报社 印刷单位及地址：经济日报印刷厂 北京市西城区白纸坊东街2号