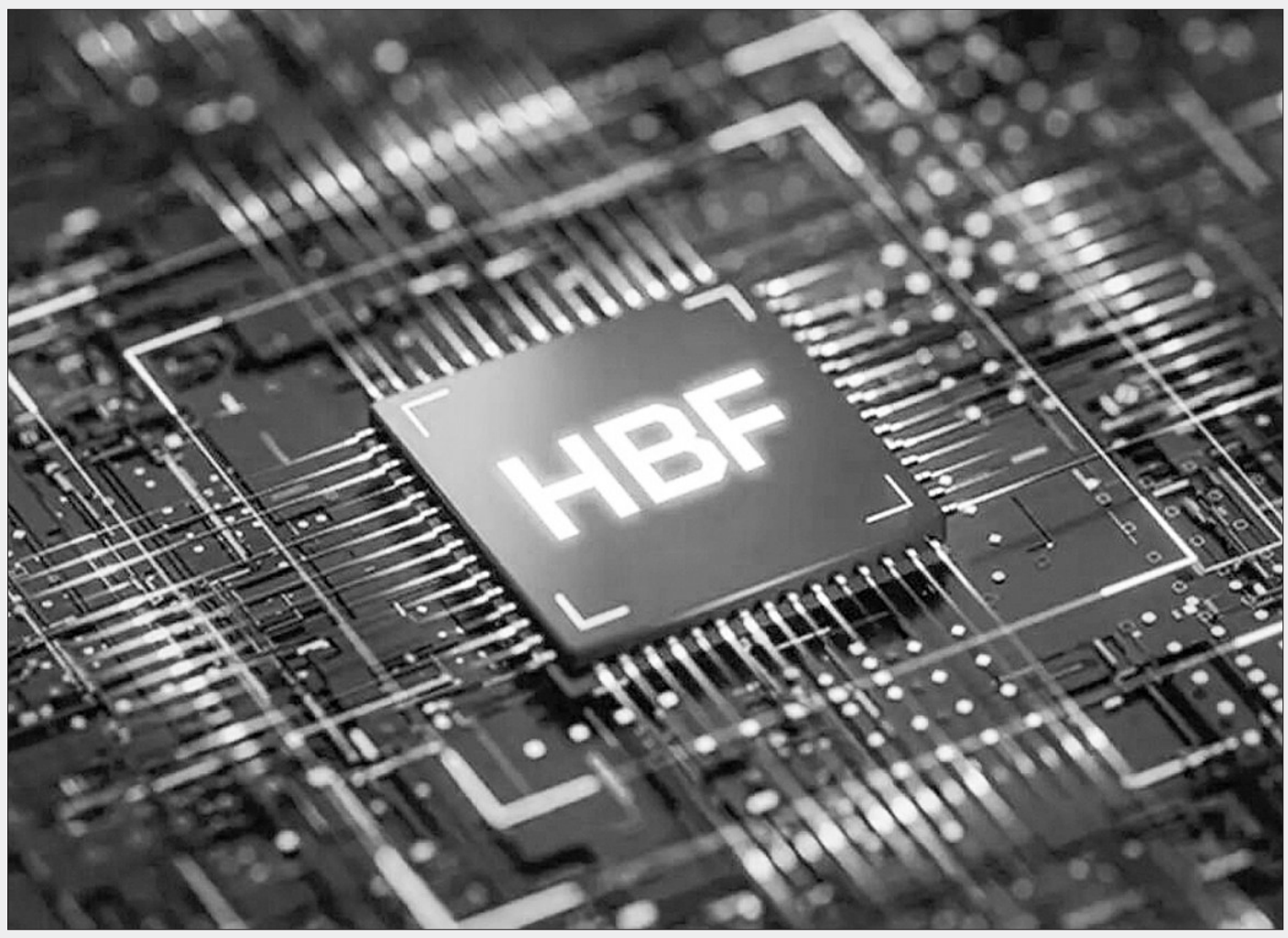
HBF，AI存储的下一站



本报记者 张心怡

HBF（高带宽闪存）已有山雨欲来之势。今年8月初，在美国加利福尼亚州举行的2026年全球闪存峰会（FMS 2026）上，SK海力士联手闪迪，通过开放计算项目（OCP）发布了HBF首个标准规范。谷歌、Tenstorrent作为HBF联盟成员参与了此次标准化进程，该联盟由SK海力士、闪迪牵头在2026年2月成立。

作为介于HBM（高带宽内存）与SSD（固态硬盘）之间的存储层级，HBF是存储厂商面向AI推理内存瓶颈的一种破题思路：当智能体导致数据规模激增、KV Cache膨胀，单一存储产品已难以满足AI推理需求，需要整合特性互补的多种存储器、构建多层次的存储体系。

HBF有望解决什么问题？

HBF不仅具备与HBM类似的高速数据传输能力，还可基于NAND闪存大幅提升容量。

智能体的规模化部署，正在推动AI产业进入“超长上下文”时代。相比“听令执行”的大模型，“自主规划”的智能体会对多轮对话和历史信息进行分析，并调用各种工具，带来数倍的上下文膨胀。为此，智能体厂商不断增加上下文窗口的容量，以增强智能体的“记性”和处理能力。

在这一过程中，KV Cache（键值缓存）是智能体流畅处理长上下文的关键。当用户输入文字时，AI会将文字拆解为模型能识别的最小单位Token（词元），上下文越长，生成的Token就越多。而KV Cache能够将已有Token计算得到的Key和Value缓存起来，在后续生成时直接复用，避免重复计算，从而提升AI推理的响应能力和经济性。

HBF聚焦的，就是AI推理中KV Cache的存储需求。

数据显示，当上下文长度从64K Token增长到1000万Token时，KV Cache容量需求从百GB级跳升至TB级。而当前的存储系统中，HBM带宽高、延时低但容量有限、成本较高，SSD容量高、成本低却带宽不足，均难以满足KV Cache快速膨胀的需求。针对这一瓶颈，产业界也尝试了各种方案。一种方案是将KV Cache“卸载”到SSD上，让GPU通过PCIe通道读取——若采用传统路径，则需经过CPU中转。这虽然比HBM容量耗尽后重新计算Token和向量更快，但会增加AI推理延迟。另一种方案是扩展GPU数量以增加HBM总容量，但会进一步增加计算集群能耗和成本。

HBF的出发点，就是将HBM的高带宽与NAND的高容量结合起来。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。SK海力士称，HBF不仅具备与HBM类似的高速数据传输能力，还可基于NAND闪存大幅提升容量。根据SK海力士的仿真结果，在Blackwell B200 GPU上，采用8组HBM3E堆栈与8组HBF堆栈的组合配置，相比纯HBM配置，每瓦性能可提升至2.69倍。这意味着，如果HBF能够批量部署，有利于改善AI基础设施效率和Token开支效率。

与HBM联手而非替代

HBF之所以能在保持读取带宽的同时将存储容量做上去，关键在于其“借鉴”了HBM的封装方式：一方面采用硅通孔（TSV）技术将多颗芯片垂直堆叠，同时将堆叠对象从DRAM芯片换成了NAND；另一方面，HBF像HBM一样通过中介层与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。

虽然HBF具备了NAND的种种优势，但也不可避免地“背负”了NAND的局限性，比如访问延迟较长、写入耐久度有限等。

因而SK海力士、闪迪等主力厂商，暂未产生用HBF替代HBM的激进念头，而是将其作为HBM面向AI“大考”的工作“搭子”。

在2026年2月发表于IEEE的论文中，SK海力士提出了H3混合架构，也就是将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，两者采用级联方式连接，实现统一地址空间并支持GPU直接访问。这种方式能够将HBM从容量负担中解放出来，专注实时任务处理，从而加速AI推理。

根据该论文数据，在对1000万Token的KV缓存进行测试时，HBM+HBF组合可同时处理的查询数量（即批处理规模）相比纯HBM配置提升了18.8倍。通过采用HBF，原本可能需要32颗GPU及其HBM来处理的工作负载，仅需2颗GPU即可完成，从而显著降低电力消耗。

而闪迪在2026年8月的投资者日活动上，展示了HBF的多种部署方式，包括在相似的封装尺寸内直接替代HBM，与HBM混合部署，与HBM分层部署（HBF承载模型权重和KV Cache、容量较小的HBM层作为缓存）等。说明其策略并非简单取代HBM，而是以更加灵活的方式嵌入存储系统。

HBF之所以能在保持读取带宽的同时将存储容量做上去，关键在于其“借鉴”了HBM的封装方式。

折射AI时代存储破题新思路

整合特性互补的多种存储器、构建多层次存储体系，或许是面向未来需求的破题方式。

相比一种新的存储产品，HBF更值得咀嚼的产业信号，或许是存储厂商在AI时代的另一种布局思路：当单一存储产品越来越难“扛住”数据规模与KV Cache的膨胀，整合特性互补的多种存储器、构建多层次存储体系，或许是面向未来需求的破题方式。相比“单点性能”，系统级的分层与生态或将成为未来产业的另一个竞争维度。

本次SK海力士、闪迪等企业发布的HBF技术规范，就为HBF与HBM等存储器的衔接预留了空间。该规范定义了围绕HBF技术交互和使用的系统设计指南，包括系统接口、电气等技术要求，涵盖基本性能预期、xPU-HBF主机接口、HBF裸晶堆叠的可靠性与封装指南等。闪迪称，该规范为AI计算系统设计人员提供了更高的灵活性，使其能够构建HBF技术与HBM共存共用的系统，有助于推动生态系统的成熟。

HBF+HBM，或许只是这种分层存储体系的起点。在去年的一档视频节目中，被誉为“HBM之父”的金正浩（Kim Jung-Ho）对未来的多层存储系统进行了设想。在他看来，那时的存储如同“智能图书馆系统”，GPU内部的SRAM将充当桌面上的笔记本，速度最快但容量最小；HBM是近旁的书架，用于快速访问和计算；HBF是地下图书馆，存储深厚的AI知识，并持续向HBM输送数据；云端存储是公共图书馆，通过光网络连接数据中心。

按照目前的时间表，HBF的规模化量产预计在2028年。虽然标准和规范已经“抢跑”在芯片样品之前，但能否如约跨越样品、量产、规模化部署的道道关卡，既要看存储厂商本身的研发和制造能力，又要看下游的AI加速器、云计算厂商是否愿意采纳。毕竟，目前大部分头部AI加速器厂商尚未对HBF具体表态。HBF能否兑现厂商设想的技术定位和市场空间，还需要产业界的紧密衔接和同步探索。

北京经开区发布AI4Chip专项政策 以AI赋能集成电路产业发展

本报讯 近日，北京经济技术开发区（以下简称“北京经开区”）正式发布《“AI4Chip”人工智能赋能集成电路产业跨越式发展行动计划（2026—2028年）》（以下简称“AI4Chip20条”）。作为AI4Chip专项政策，“AI4Chip20条”以五大行动、20条重点任务为抓手，在北京亦庄全域人工智能之城战略引领下，聚力推进人工智能赋能集成电路产业跨越式发展的芯智协同。

所谓AI4Chip，即人工智能赋能集成电路，旨在将AI技术深度嵌入芯片设计、制造、封测、设备材料及产业生态全链条，以智能算法驱动工艺优化、以数据闭环加速研发迭代，从根本上重塑传统集成电路产业依赖人工经验与物理试错的发展模式。

面对这一技术变革浪潮，北京经开区选择主动破题、先行一步。“打造新质生产力典范区，必须下好AI与集成电路深度融合的先手棋。”北京经开区有关负责人表示，“我们将按照‘坚持AI原生筑基、全链阶跃发展、生态资源融通、安全自主可控’总体思路，加快构建‘根植AI原生，助推AI赋能’产业发展体系。”

“AI4Chip20条”聚焦AI原生全链赋能、“AI+智能设计”焕芯、“AI+制造封测”筑芯、“AI+设备材料”强芯、“AI+产业生态”育芯五大行动精准发力。

在AI原生全链赋能方面，聚焦搭建AI原生驱动的设计、制造封测、验证等创新体系，筑牢AI4Chip原创技术创新底座，以AI原生技术贯通集成电路全产业链。

在“AI+智能设计”焕芯方面，围绕AI芯片前沿架构攻关、设计全流程AI赋能、芯片全生命周期管理、芯片设计与制造封测智能化协同、智能体芯片设计助手构建等方面，全方位实现芯片设计环节AI深度渗透与降本增效。

在“AI+制造封测”筑芯方面，通过打造AI驱动的集成电路制造产线、智能封装技术体系、全流程智能化测试体系以及推动AI+半导体CIM系统全域布局，实现集成电路产线、封装、测试全流程智能化改造与工厂自动化能力升级。

在“AI+设备材料”强芯方面，依托数字孪生体系，运用AI边端盒子、边缘智能体及虚拟仿真等技术赋能半导体设备研发攻坚与效率提升、关键零部件自主优化与性能提升以及半导体材料工艺升级与研发创新，强化集成电路产业链关键配套支撑能力。

在“AI+产业生态”育芯方面，围绕专用算力建设、产业数据集搭建、垂类模型矩阵研发、行业标准制定、复合人才引育以及公共服务能力优化六大维度，完善AI赋能集成电路产业底层支撑体系，夯实全链条AI融合发展生态基础。

为确保各项任务落地见效，“AI4Chip20条”明确了组织机制、政策保障和要素供给三方面内容。

北京经开区有关负责人表示：“在资金保障方面，我们争取国家专项资金支持，统筹用好市区两级政策和‘揭榜挂帅’项目支持，重点向AI4Chip领域倾斜，同时用好专项政策和‘算力券’‘模型券’等普惠性政策，支持经开区集成电路企业开展AI赋能器件工艺、关键设备、EDA工具及IP、先进封装、存储器等方向研发，创新数据共享、模型推广扶持方式，降低企业AI赋能转型成本。同步我们也将强化要素供给，统筹算力、模型、数据要素保障，集聚产业融合高端人才，引导金融机构创新信贷产品，助力企业推进AI赋能相关业务。”

作为首都高精尖产业主阵地，北京经开区已聚集超400家集成电路相关企业，产业规模超1300亿元，形成了以北方华创、中芯国际、长鑫集电为龙头，涵盖设计、制造封测、装备零部件等环节的集成电路全产业链。

根据“AI4Chip20条”目标蓝图，力争到2028年，经开区基本建成集成电路产业AI融合发展体系，培育3~5家具有国际影响力的AI4Chip生态型领军企业，形成10个以上具备行业推广能力的标杆应用，持续推动集成电路产业“AI原生”创新与“AI赋能”升级，构建“软硬互驱、芯智共生”的产业发展新范式。 （刘 璐）

苹果新款Mac mini 首搭2纳米芯片

本报讯 8月25日，苹果公司官网更新了搭载M5 Pro和M6芯片的新款Mac mini，以及搭载M5 Max和M5 Ultra芯片的新款Mac Studio。

苹果公司表示，M6是苹果首款采用2纳米制程工艺的尖端芯片，M5 Ultra则是“迄今为止苹果最强大的芯片”。

据介绍，苹果M6芯片包含12核CPU、12核GPU和双16核神经网络引擎，Mac mini成为首款搭载该芯片的产品。相比采用M4芯片的产品，搭载M6芯片的Mac mini可提供高达4倍的AI性能、2倍的存储容量与图形处理速度，CPU性能增加40%。Mac mini将可运行即将推出的macOS 27系统和下一代Apple Intelligence。

据悉，苹果将在9月举行秋季新品发布会，重点产品包括iPhone 18 Pro系列、首款折叠iPhone，以及Apple Watch、AirPods等。随着M6芯片亮相，这款2纳米制程的芯片有望应用于更多产品中。 （时 简）

SK海力士 拟在韩国清州投资扩产

本报讯 近日，SK海力士计划在韩国忠清北道清州市追加高达100万亿韩元的巨额投资。韩国相关部门正展开深入讨论，拟依据新近生效的《半导体特别法》，正式将清州设立为国家级半导体产业集群。

具体而言，其中约80万亿韩元将投入新建NAND闪存晶圆厂M17，约20万亿韩元用于强化先进封装能力的P&T7及其他设施建设。SK海力士代表理事社长郭鲁正在此前公布该中长期投资规划时表示，随着人工智能服务全面铺开，不仅高带宽存储器和服务器DRAM需求激增，企业级固态硬盘等NAND需求也正爆发式增长，甚至达到供不应求的程度，因此有必要立即进行一定规模以上的扩产。他强调，清州是能够快速、高效建设NAND晶圆厂的成熟据点，土地、电力、用水等核心基础设施已在相当程度上完备。

在具体的投资落地方面，8月7日，SK海力士董事会批准了合计约54.3万亿韩元的本土投资方案。其中，清州M17工厂获投19.1万亿韩元，投资周期至2031年4月30日。M17工厂总占地面积约68万平方米，计划于明年2月动工，首个洁净室预计于2028年12月启用。P&T7先进封装厂预计将于2027年年底完工。 （苏 砚）

英伟达推出 面向定制AI加速器的NVHBM

本报讯 近日，英伟达宣布为NV-Link Fusion定制芯片生态推出全新组件NVHBM，这是一款面向定制AI加速器的定制HBM解决方案。

英伟达宣称，NVHBM单栈带宽相比标准HBM4E最高提升30%，同时功耗更低、主芯片占用面积更小。

NVHBM的核心设计在于将内存控制器从主芯片转移到HBM堆栈的基片之中。传统方案将HBM内存控制器集成在主芯片上，占用本可用于计算的宝贵硅面积。NVHBM将控制器下沉至HBM基片后，主芯片计算可用面积最多可增加30%。

功耗方面，NVHBM比标准商用HBM4E低15%。在数千上万颗芯片规模下，单颗芯片节省的功耗累积效应显著，可反哺计算单元提升性能，或在固定功耗预算内部署更多加速器。

需要指出的是，NVHBM并非用来取代通用HBM产品，而是英伟达专门向潜在合作伙伴提供的基础组件。

英伟达已与领先内存厂商完成NVHBM的设计与验证。亚马逊旗下Annapurna Labs成为首家合作伙伴。

英伟达同时还向NVLink Fusion合作伙伴提供小型定制PHY物理层IP，供其集成到自有芯片中。 （何 浩）