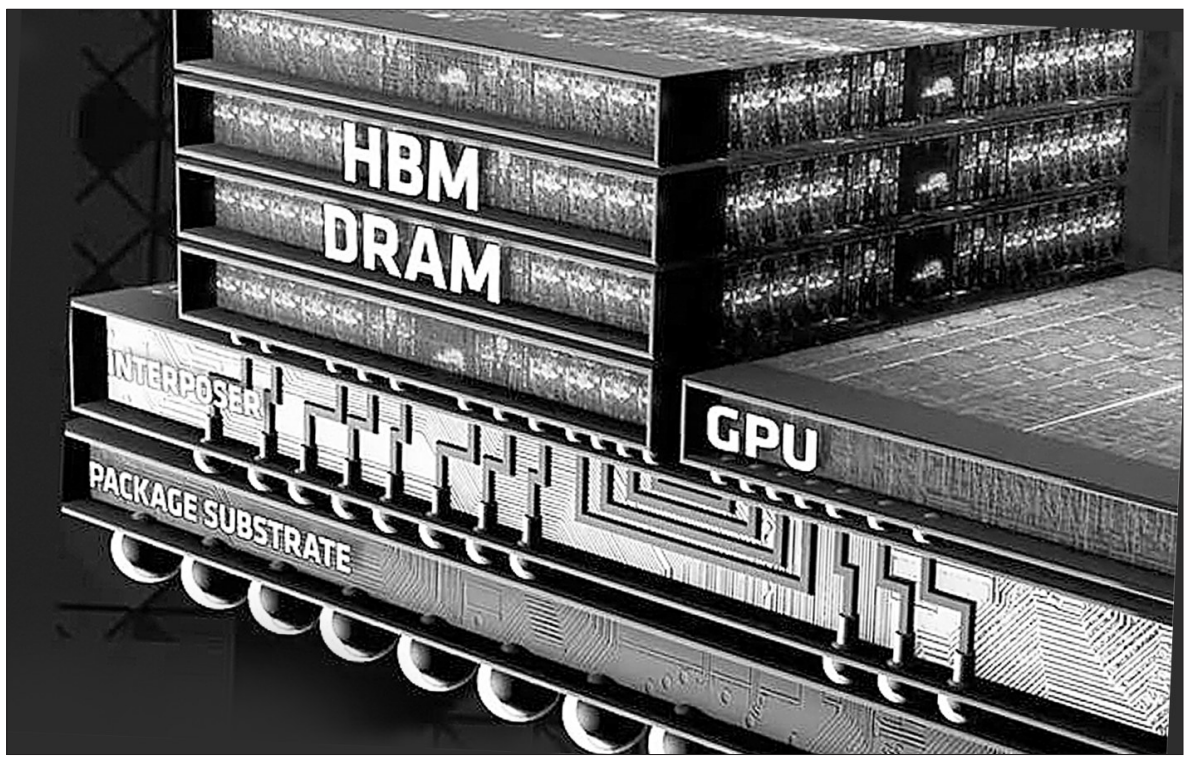
需求预期生变，HBM企业“忐忑不安”



本报记者 许子皓

近日，韩国替代数据平台KED Aicel的最新数据显示，2025年1月，HBM缔造者、也是英伟达最大HBM供应商的SK海力士位于利川和清州的芯片工厂多芯片封装（MCP）的出口额同比大幅增长105.7%，出口额为12.9亿美元，但环比下跌29.8%，这也是自2023年4月该公司开发出全球12层HBM3芯片以来，环比跌幅最大的一次。业内专家表示，SK海力士第一季度的HBM出货量很可能比2024年第四季度下降10%以上。此外，三星电子位于平泽、龙仁、水原、天安和牙山的芯片工厂，1月的MCP出口总量较上月也下降62.3%。

HBM厂商在高歌猛进时突然急转直下，多少和DeepSeek的惊艳登场有关系。毕竟，它曾在1月重创美股科技板块，英伟达股价一度暴跌约17%，创下美股单日最大跌幅纪录。因为与GPU的深度绑定，HBM也成为了除GPU以外，受到DeepSeek影响最大的芯片品类。

HBM的需求预期发生变化

HBM即高带宽存储，由多层DRAM Die垂直堆叠，每层Die通过硅通孔（TSV）技术实现与逻辑Die连接，使得8层、12层Die封装于小体积空间中，从而实现小尺存与高带宽、高传输速度的兼容，优秀的特性使其成为高性能AI服务器GPU显存的主流解决方案。随着AI市场持续增长，AI服务器对于HBM的需求量越来越高，据Mordor Intelligence预测，从2024年到2029年，HBM市场规模将从约25.2亿美元激增至79.5亿美元，年复合增长率高达25.86%。三星、SK海力士等各大存储芯片企业都在加紧研发，推出更高级别的HBM产品，产能持续紧张，这也成为了正处于下行周期的存储芯片市场中的一盏明灯。

但DeepSeek掀起的波澜，让这盏希望之灯摇摇欲坠。因为DeepSeek采用的H800 GPU，其性能仅为H100 GPU的一半的主要原因在于其HBM的带宽降低。尽管H800使用与H100相同的80GB HBM3内存，但带宽大约降低了16%。这一选择表明，在一定程度上，AI模型可以在较低规格的HBM支持下，依然实现较高的性能，也让市场对HBM的需求预期发生了变化，并很快体现在股市上。

2025年韩国农历新年后的第一个交易日，SK海力士的股价下跌12%，三星电子股价下跌4%。美光科技股价在1月27日开盘也下跌7.93%。三星设备解决方案（DS）部门执行副总裁Kim Jae-joon在第四季度财报电话会议上告诉投资者：“由于我们向多家供应商提供GPU的HBM，我们正在密切关注行业趋势并考虑各种情况。由于市场中长期机遇和短期风险并存，我们将确保对市场的快速变化做出迅速、及时的反应。”

业内专家表示，随着DeepSeek的崛起，对英伟达和其他人工智能科技公司造成打击，三星、SK海力士和美光预计在快速增长的人工智能内存芯片业务中面临越来越多的不确定性。如果越来越多的AI企业效仿DeepSeek，采用低成本、低规格的芯片来实现高性能的AI模型，那么高端HBM的市场需求将受到更严重的冲击。比如HBM的市场价格会下跌，可能直接影响到企业的营收和利润，企业可能需要通过降低成本、提高生产效率等方式来维持盈利能力，而由于现在各大企业都在增加产能，市场很可能会从供不应求变成供大于求，这无疑增加了企业的运营难度。

低规格HBM的需求量有望提高

但DeepSeek并非不需要HBM，业内人士普遍认为市场只是对高端HBM的需求预期发生了改变。

赛迪顾问集成电路中心副总经理杨俊刚表示，以HBM3E不同层数产品为例，原本市场对高性能、高规格的12层HBM3E需求旺盛，然而随着DeepSeek等采用低规格GPU的发展，对这种高端产品的需求增速可能会放缓。而相对低规格的HBM产品，由于其成本优势，可能会在市场上获得更多的关注和需求。尽管短期内HBM企业面临着需求预期变化和价格承压的风险，但从长期来看，AI技术的持续发展仍为HBM企业带来了潜在的机遇。随着AI应用场景的不断拓展，如智能驾驶、智能家居、工业自动化等领域对AI技术的需求日益增长，对HBM的整体需求也有望随之上升。即使云服务提供商减少了对高端GPU的投资，HBM的整体供应量预计仍会保持增长。

DeepSeek的出现可能促使HBM市场格局发生重塑。在过去，HBM市场主要由少数几家企业主导，市场竞争相对稳定。然而，DeepSeek的技术创新，可能会吸引更多的企业进入HBM市场，加剧了市场竞争。这将促使HBM企业不断创新，提高产品的性价比，以满足市场的需求。在这个过程中，市场份额会重新分配，这对于传统的HBM厂商来说，也是一个巨大的挑战。

定制化是HBM未来发展新趋势

随着AI技术的不断发展，不同的应用场景对HBM的性能和规格提出了多样化的需求。杨俊刚认为，传统的通用HBM产品，越来越难以满足客户在性能、功率、价格和占地面积等方面的具体需求。因此，HBM厂商纷纷加大技术创新力度，推出定制化的HBM产品。

据了解，SK海力士计划通过半定制化方案为不同客户提供个性化服务，以满足特定应用的需求。这种定制化服务不仅提升了存储器的适用性，也使其在市场上更具竞争力。在新技术的支持下，HBM的增强封装技术将使得产品在功能上更加多样化，为客户提供了更大程度的选择自由；三星电子通过将HBM与定制逻辑芯片进行3D堆叠，在保证性能的同时，有效地降低了功耗和占地面积。这种技术创新，不仅满足了客户对高性能、低功耗HBM产品的需求。通过不断推出定制化、多样化的产品，HBM厂商能够更好地适应市场变化，满足不同客户的需求，从而在激烈的市场竞争中保持领先地位；美光则是会在存储密度和带宽方面寻求突破，研发更高层数的存储堆栈，以增加存储容量和带宽，并在功耗优化上发力，通过改进电路设计和封装技术，降低HBM的功耗，满足AI服务器对低功耗内存的需求。

在英伟达股价因受DeepSeek影响而经历了一轮过山车般的起伏后，黄仁勋终于站了出来。

2月21日，一段黄仁勋的受访视频出现在其合作伙伴DataDirect Networksg公司举办的线上活动。在访谈中，他首次公开回应了DeepSeek是否利好英伟达的相关话题。

黄仁勋首次公开回应DeepSeek影响

本报记者 杨鹏岳

“R1让实际情况恰恰相反”

“DeepSeek-R1作为全球首个开源的推理模型，令人十分兴奋。R1模型的开源，让全球的热情也变得非常高涨。”黄仁勋表示。

他首先否定了一个既有认知：从投资者的角度来看，有这样一种思维定式——AI世界就是先做预训练，然后就是推理。而“推理”就是给AI提出一个问题，它立刻给出答案。

“虽然不知道这种思维源自哪里，但它显然是错误的。”黄仁勋强调。

为此，黄仁勋给出了详细的解释：正确的模式是先进行预训练，让模型对信息有一个基础的理解，预训练后要持续保持严谨。第二个阶段是最为重要的后训练（post training），也就是模型学会解决问题的过程。在这一阶段，模型已经有了基础的信息，而后运用这些基础知识去解决实际问题。后训练阶段与一系列不同的学习范式相关。在这些范式的推动下，AI技术在过去五年发展得非常迅猛，计算需求也因而变得极为密集。“大家会觉得预训练要少得多，但是他们忘记了预训练之后的后训练的计算需求量是相当大的。”他表示。

黄仁勋提到了第三条缩放定律。推理越多，回答问题前思考得就越多，推理效果就会越好，这是一个计算量相当大的过程。

“所以市场对DeepSeek R1问世的反应是‘天哪，AI到头了！’，就好像有了它，我们便不再需要进行任何计算了，但实际情况恰恰相反。”黄仁勋表示。

市场是否“买单”？

2025年1月，DeepSeek推出低算力成本的开源大语言模型R1，以仅560万美元的训练成本实现了与OpenAI o1等闭源模型相当的性能。这一成果挑战了传统AI训练依赖高算力芯片堆料的商业模式。

受此影响，英伟达市值曾一度下跌，尤其在1月27日跌去近17%，一夜蒸发近6000亿美元，创美股单日市值蒸发纪录。当时，投资者们担忧AI行业对英伟达芯片的需求可能大幅减少，导致市场对算力泡沫的恐慌。

事实上，英伟达官方在股价暴跌后迅速发声，指出DeepSeek的成功恰恰证明了其芯片的实用性，并强调“推理阶段需要大量英伟达GPU和高性能网络”，未来市场需求将随着AI应用扩展而增长。但当时的发声并未对其股价回升起到实际效果。

随后市场逐渐认识到，DeepSeek的创新可能推动更多中小型企业进入AI领域，反而刺激对GPU等算力芯片的增量需求。英伟达股价在接下来近一个月里震荡回升，目前已逐渐“收复失地”。

值得关注的是，此次黄仁勋的发声在英伟达财报公布之前，2月26日美股盘后，英伟达将公布四季度财报。虽然此次黄仁勋正面表达了对于DeepSeek问世将利好英伟达未来的信心，但市场的真正反应将有赖于其财报给出的关键数据。

5G-A渐入佳境

（上接第1版）在哈尔滨中国联通营业厅内，工作人员告诉记者：“在最受游客青睐的中华巴洛克等热门景区，我们已经部署了5G-A 3CC网络和内生智能技术，可为街区内的用户提供超高速率、超低时延以及大连接容量的网络服务，支持游客观看高清视频和直播等多种需求。”

5G-A网络的“星星之火”，早已被播撒到了全国的各个角落。近日，中国电信宣布将从2025年2月10日至12月31日开展友好客户5G-A体验活动，首批体验活动将在上海、北京、广东、江苏、浙江、四川、贵州等地开展；中国联通也表示，将在39个重点城市主城区、其他300余城市重点场景启动5G-A业务；中国移动则早在去年6月就表示，将完成300个城市的5G-A网络部署……记者了解到，截至目前，国内已有超330个城市完成了5G-A网络部署。

而北上广深等重点区域则更是运营商开展5G-A商用部署的“兵家必争之地”。上海移动已建成5G-A基站超1.4万个，实现了上海主城区及重点区域的全覆盖；北京联通在北京四环内及城市副中心等主要核心区域以及知名地标实现3CC全面覆盖，站点规模超4000个，核心城区及重点区域5G-A生效比超70%；深圳电信在深新增建设5G-A通感基站150个、通信基站1500个，实现全市起降场（点）和航线全覆盖……

除“硬装”的新型信息基础设施建设外，面向差异化人群的“软装”——5G-A流量套餐包也做好了面向大众的准备。据了解，运营商已在100余个城市发布了面向C端用户的5G-A套餐，其中，中国移动、中国电信均将直播、游戏、商旅三大人群作为5G-A个人用户的“重点目标”，聚焦不同场景需求推出了差异化的5G-A套餐；中国联通发布的5G-A套餐则包含青春版、商务版、云加速版、算力宽带1000M提速包等多档5G-A算力卡套餐，力图为个人用户打造“量身定制”的5G-A使用体验。

要真正让5G-A融入每个人的日常生活，支持5G-A网络的终端是不可或缺的媒介。联通华盛董事长王启明表示，从终端层面看，5G-A同样已经做好了迎接大众考验的准备：“芯片方面，目前高通、联发科技、三星、紫光、展讯等芯片厂商推出的26款中高端芯片均可支持5G-A；终端方面，截至目前，主流品牌、主流价格段的162款5G终端已经全面具备5G-A能力，2025年，支持5G-A的手机将超过300款。”

赛迪顾问预测，“十五五”期间，中国5G-A用户量预计将接近13亿户，占全球5G-A用户总数的比例超过50%，网络覆盖将实现深度和广度的双重飞跃。可以看到，以“5GA”标识“入驻”手机为标志，5G-A技术普惠化进程全面提速，5G-A网络走入千家万户已成大势所趋。

5G-A点亮全新生产生活方式

在哈尔滨冰球馆，5G-A“赛事通信高速路”为全球观众带来了8K超高清直播体验，视线随着一颗小小的冰球辗转挪腾，淋漓尽致地体验比赛快节奏的攻防转换；在浙江慈城古县城景区，在5G-A网络和XR技术、大空间厘米级定位算法等核心技术赋能下，游客得以“亲身”体验“江南小长安”的独特韵味；在行进中的上海地铁18号线列车上，5G-A网络让乘客再也不必为卡顿的视频、加载不出的网页烦恼；在八达岭长城景区，只需手机扫码下单，无人机就能将商品及时送到游客手中……点滴细节，昭示着5G-A网络正真真切切地为人们的生活方式带来“新气象”。

对此，华为公司副总裁、无线产品线总裁曹明认为，5G-A网络的发展，无疑将带来更加差异化的用户体验，并重塑传统业务边界：“一方面，5G-A凭借强大的网络能力，能够改变以往不同等级用户服务体验趋同的状况，为不同用户、不同场景提供差异化服务；另一方面，5G-A不仅强化了网络能力，还极大拓展了连接边界，能够实现一网连接生活和工作中的所有移动终端；此外，5G-A能够实现连接与感知的结合，从而开辟新的业务领域。”

从实例上看，除裸眼3D、混合现实等高度依赖网络带宽和稳定性的应用领域外，车路协同、低空经济等依托5G-A网络通感一体特性的新兴产业也正逐渐走出“实验室”，为越来越多的人的生活带来新的改变。工信部数据显示，截至2024年9月，全国共建设17个国家级测试示范区、7个车联网先导区、16个智慧城市与智能网联汽车协同发展试点城市，开放智能网联汽车测试道路3.2万公里，发放测试示范牌照超过7700张，测试里程超过1.2亿公里。低空经济方面，北京、上海、常州等15个城市已宣布联合共建低空经济生态圈，计划到2025年打造100个示范项目。据赛迪顾问预测，2025年，我国低空经济规模将达到8591.7亿元。

此外，5G-A在生产制造中的比重也正逐渐扩大。在济南市莱芜区某汽车工业厂区内，山东移动建成省内首个5G-A工业基站商用试点，通过应用5G多层网、4.9G基站大上行时隙改造、基站内置算力引擎单板部署等技术，结合5G-A工业基站多种确定性技术，为工业生产提供内生确定性的网络保障；在河北保定徐水厂区焊接车间，河北联通与长城精工自动化打造5G-A柔性汽车产线，智能化设备和机器人通过5G-A网络默契协作，真正实现“黑灯”生产；在上海宝钢轧钢产线上，上海电信建设5G-A专网，实现高架起重机远程控制、堆垛取料机远程控制、5G加渣机器人、重载AGV自动驱动、钢表面质量检测等一系列应用升级，完成多个场景的智能化蜕变……

基于此，行业专家向记者表示：“就像3G网络开启社交媒体时代、4G网络引爆直播产业、5G为工业互联网注入新血液，5G-A发展催生出的新技术、新产业也必将在潜移默化间改变人们的生产生活方式。”

“5G-A”+AI成发展“主基调”

从今年MWC各家企业放出的前瞻消息来看，如何实现5G-A与AI的深度融合，将成为2025年通信产业发展的“主基调”：华为宣布将在MWC2025展示其全系列、全场景的5G-A产品解决方案，帮助全球运营商构筑AI-Centric 5G-A无线网络；中国联通将在展会现场展示AI在云手机、智能家居、低空经济、智能汽车、工业自动化等方面的深度应用；中兴通讯将联合中国移动发布5G-A×AI创新成果发布会；高通等芯片制造商则倾向于重点展示5G-A芯片在移动终端的AI应用……

中国移动研究院首席专家王大鹏告诉记者：“5G-A可以从三方面为AI提供能力支撑：在连接方面，5G-A强大的连接能力能够确保AI服务的质量，并且可以提供泛在连接的AI服务；在计算方面，基于5G-A提供的边、端、云的算力底座，能够进一步保障AI的计算能力；在数据方面，5G-A可以通过通感一体、无源物联等新技术，为AI提供更加丰富的数据来源。”

曹明指出：“移动AI时代的到来带来了30亿AI助理‘入端’，要抓住这一广阔的市场，就要深化5G-A与AI的融合，提升移动网络谱效和能效，在提升网络性能和用户体验、降低网络比特成本的同时，基于数字化站点和无线智能化助力运营商网络实现L4高阶自智，赋能网络提质增效。”

记者观察到，提到AI与5G-A的融合应用，工业生产和AI智能家居仍然是较为主流的应用场景。而今年的MWC上，华为宣称“业界首次”将5G-A与AI技术深度融合，在无线网络中全栈引入智能化能力，在满足移动AI时代的差异化需求的同时，实现网络智绿与运维智简。正逐步全方位融入人们生产生活、应用渐入佳境的5G-A网络将与AI碰撞出怎样的火花？让我们拭目以待。

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