ChatGPT带火高带宽内存HBM

本报记者 陈炳欣

ChatGPT已经从下游AI应用火到了上游芯片领域，在将GPU等AI芯片推向高峰的同时，也极大带动了市场对新一代内存芯片HBM（高带宽内存）的需求。据悉，2023年开年以来，三星、SK海力士的HBM订单就快速增加，价格也水涨船高。有市场人士透露，近期HBM3规格DRAM价格上涨了5倍。这也为寒冬中的存储器市场带来了一抹春色。

ChatGPT爆红提升HBM需求

作为一款自然语言类AI应用，ChatGPT不仅对芯片算力有着巨大需求，对内存技术的要求也极高，比如ChatGPT上线应用以来，拥有高速数据传输速度的HBM内存芯片几乎成为ChatGPT的必备配置，市场需求激增。有消息称，2023年开年以来，三星、SK海力士HBM订单就快速增加，价格也随之提升。

有市场人士透露英伟达已经将SK海力士的HBM3安装到其高性能GPU H100上，而H100已开始供应ChatGPT服务器所需。近期，最新一代高带宽内存HBM3的价格上涨了5倍。

SK海力士HBM设计团队项目负责人朴明宰撰文介绍，HBM是一种可以实现高带宽的高附加值DRAM产品，适用于超级计算机、AI加速器等对性能要求较高的计算系统。随着计算技术的发展，机器学习的应用日渐广泛，而机器学习的基础是自20世纪80年代以来一直作为研究热点的神经网络模型。作为速度最快的DRAM产品，HBM在克服计算技术的局限性方面发挥着关键作用。

此前，尽管HBM具有出色的性能，但与一般DRAM相比，其应用较少。这是因为HBM需要复杂的生产过程，平均售价至少是DRAM的3倍。但AI服务的扩展正在扭转这样的局面。据悉，目前已经有超过2.5万块英伟达计算卡加入到了深度学习的训练中。未来，随着不断接入ChatGPT等生成式AI需求大增，HBM的需求也将呈现暴增态势。三星内存副总裁Kim Jae-joon便指出，ChatGPT等基于自然语言技术的交互式AI应用的发展有利于提升内存需求。高效且大量的运算能力、高容量的内存，是AI学习与推论模型的根基。

> 此前，尽管HBM具有出色的性能，但与一般DRAM相比，其应用较少。

> 去年以来，受消费电子产品需求下滑等多种因素影响，存储行业进入了下行周期。

封装存储巨头齐争高性能市场

去年以来，受消费电子产品需求下滑等多种因素影响，存储行业进入了下行周期。SK海力士全年的净利润降至2.4万亿韩元，同比下滑75%。三星电子存储业务的营收与营业利润，在去年第三季度和第四季度也是同比环比双双下滑。TrendForce集邦咨询预估2023年第一季度DRAM价格仍将持续下探。其中，PC及服务器用DRAM跌幅仍是近两成。HBM的需求火爆不啻于一剂“强心针”。这种情况下，SK海力士、三星电子、美光等内存厂商均表示将致力于HBM的开发。企业之间的产品开发竞争也随之升温。

2013年SK海力士将TSV技术应用于DRAM，在业界首次成功研发出第一代HBM芯片。此后，SK海力士相继推出HBM2、HBM2E、HBM3数代产品。据悉，SK海力士正在研发HBM4，预计新一代产品将能够更广泛地应用于高性能数据中心、超级计算机和AI等领域。SK海力士CFO金祐贤表示，公司正关注着市场上出现以AI为基础的新服务器替换存储器需求的积极信号。公司计划持续投资DDR5/LPDDR5、HBM3等主力产品的量产和未来高成长领域。

三星电子虽然在HBM的开发上落后一步，但也在寻找新的突破口。据悉，三星电子正在开发集成AI处理器新一代存内计算芯片HBM-PIM。北京大学信息科学技术学院微纳电子学系副教授叶乐表示，存内计算芯片是将AI引擎引入存储库，将处理操作转移到HBM本身，将有效加大数据传输带宽，提高内存的处理速度。存内计算技术正在实现产品化。目前基于SRAM的存内计算，已经进入产品化的前夜，DRAM存内计算适用于大算力AI芯片，因此还需要解决其他一系列的技术难题。

> 与HBM相关的上下游企业也在纷纷发力，以期抢占市场先机。

上下游企业发力抢占先机

与HBM相关的上下游企业也在纷纷发力，以期抢占先机。AMD在HBM的诞生与发展过程中功不可没。最早是AMD意识到DDR的局限性并产生开发堆叠内存的想法，其后与SK海力士联手研发了HBM，并在其Fury显卡采用全球首款HBM。据ISSCC 2023国际固态电路大会上的消息，AMD考虑在Instinct系列加速卡已经整合封装HBM高带宽内存的基础上，在后者之上继续堆叠DRAM内存，在一些关键算法内核可以直接在整合内存内执行，而不必在CPU和独立内存之间往复进行通信传输，从而提升AI处理的性能，并降低功耗。

英伟达同样重视处理器与内存间的协同，一直在要求SK海力士提供最新的HBM3内存芯片。据悉，目前已经有超过2.5万块英伟达计算卡加入到了深度学习的训练中。如果所有的互联网企业都在搜索引擎中加入ChatGPT这样的机器人，那么计算卡以及相应的服务器的需求量将达到50万块，也将连同带动HBM的需求量大幅增长。

IP厂商亦已先行布局HBM3。去年，Synopsys推出首个完整的HBM3 IP解决方案，包括用于2.5D多芯片封装系统的控制器、PHY（物理层芯片）和验证IP。HBM3 PHY IP基于5nm制程打造，每个引脚的速率可达7200 Mbps，内存带宽最高可提升至921GB/s。Rambus也推出支持HBM3的内存接口子系统，内含完全集成的PHY和数字控制器，数据传输速率达8.4 Gbps，可提供超过1TB/s的带宽。

Rambus IP核产品营销高级总监Frank Ferro此前在接受采访时指出，HBM现在依旧处于相对早期的阶段，其未来还有很长的一段路要走。而可预见的是，随着越来越多的厂商在AI和机器学习等领域不断发力，内存产品设计的复杂性正在快速上升，并对带宽提出了更高的要求，不断上升的宽带需求将持续驱动HBM发展。

英伟达2023财年营收269.74亿美元

本报讯 记者**张心怡**报道：北京时间2月23日凌晨，英伟达公布了截至1月29日的2023财年第四季度及全年财报。信息显示，英伟达第四季度营收60.51亿美元，同比下降21%；2023财年营收269.74亿美元，和2022财年的269.14亿美元基本持平。

供过于求导致净利润大幅下降。虽然英伟达2023财年营收与2022财年基本持平，但净利润同比下降55%，毛利率下降了8个百分点。英伟达表示，由于2023财年基于Ampere架构的游戏和数据中心产品供过于求，产生了21.7亿美元的库存费用，影响了毛利表现。

终止收购Arm损失13.6亿美元。在财报的运营费用中，有一笔13.53亿美元的“收购终止支出”。结合此前披露的信息，这笔支出是英伟达终止收购ARM的相关运营费用。2022年2月，英伟达宣布终止对Arm的收购，原因是交易面临重大监管挑战。按照英伟达与软银的协议，在收购终止后，软银保留了英伟达12.5亿美元的预付款，英伟达则保留了对Arm架构20年的使用许可。

库存金额同比增长98%。截至2023年1月29日，英伟达的库存金额为51.59亿美元，相较去年同期增长98%。英伟达表示，这些库存主要用于支撑市场对数据中心和游戏新品的需求增长。截至统计期，英伟达库存采购总额和长期供应义务金额为49.2亿美元，预付供应协议金额为34.5亿美元。

数据中心全年营收激增。2023财年，英伟达数据中心业务营收150.05亿美元，较2022财年激增41%。2023财年数据中心业务的增长主要得益于超大规模客户的强劲增长，以及数个与英伟达有多年合作协议的云服务提供商对英伟达新款AI云服务产品的采购。虽然第四季度数据中心业务出现小幅度的环比下降，较去年同期仍有11%的增长。

游戏业务在第四季度停止下跌。经历了连续3个季度的营收下降之后，英伟达游戏业务终于在第四季度小幅回升。该季度的环比增长主要得益于新款GeForce RTX GPU的带动。2022年9月，英伟达发布新一代图形架构Ada Lovelace，以及采用该架构的GeForceRTX40系列显卡。

汽车业务小步快跑。英伟达2023财年汽车业务收入创历史新高，较去年增长60%。第四季度汽车业务营收同比增长135%、环比增长17%。英伟达表示，汽车业务的营收表现反映了自动驾驶解决方案、电动汽车制造商计算解决方案以及人工智能驾驶舱解决方案的销售增长。不难看出，英伟达的汽车业务与英特尔财报中的Mobileye业务类似，正处于体量小、增长快的爬坡期。

预计下一季度营收、毛利将小幅增长。英伟达预计2024财年第一季度营收65亿美元，毛利率为64.1%，环比均小幅上升。

关注AICG和大模型。针对第四季度业绩，英伟达创始人兼首席执行官黄仁勋表示，从初创公司到大型企业，对于生成式AI的多功能性与能力的兴趣越来越浓厚。英伟达将帮助客户从生成式AI和大型语言模型技术的突破中获取优势。

DDR5能否拉动存储市场上扬？

本报记者 许子皓

当下，存储芯片市场仍处于低迷状态。各大存储芯片企业的最新财报显示，企业利润均在下跌。在产业下行周期，新技术、新产品的诞生往往会成为振奋市场的关键点。DDR5 DRAM作为存储领域的新产品，逐渐成为各大企业竞逐的焦点。例如，近期SK海力士就宣布，公司已经开发出了当前速度最快的移动DRAM“LPDDR5T”。但就目前情况来看，DDR5的市场渗透率还未达到预期，DDR5能否拉动存储行业上扬，仍有待观察。

上下游厂商“惺惺相惜”猛推新品

2020年7月15日，JEDEC固态技术协会正式发布了下一代主流内存标准DDR5 SDRAM的最终规范（JESD79-5）。DDR5作为最新的高带宽电脑存储器规格，被业界视为具备“革命意义”的内存架构。与DDR4相比，DDR5具备更高速度、更大容量与更低能耗。全新DDR5内存的最高传输速率达6.4Gbps，比DDR4内存提升了一倍。这些特性有助于缓解每个核心的带宽紧张难题，可进一步推动CPU内核数量增加，并逐年提升计算能力。由于近期内存市场需求低迷，DDR5被认为是提振市场需求、扩大内存厂商营收的“利器”。

SK海力士一直在加速研发DDR5新产品，以刺激客户淘汰旧产品，扩大市场需求。1月12日，SK海力士宣布，公司研发的第四代10纳米级DDR5服务器DRAM，获得了英特尔全新第四代Xeon服务器处理器兼容认证。SK海力士表示，英特尔的第四代Xeon服务器处理器是存储器半导体行业反弹的关键，SK海力士将把握此次机会，尽早克服存储器半导体的低迷市况。1月25日，SK海力士再次宣布，公司已经开发出当前速度最快的移动DRAM（内存）“LPDDR5T”，并已向客户提供了样品。据介绍，本次产品的速度比现有产品快13%，运行速度高达9.6Gbps。

SK海力士预测，虽然今年上半年的半导体市场情况可能持续低迷，但下半年会出现好转。随着市场需求逐渐回升，IT企业将增加价格大幅下降的存储芯片使用量。

赛迪顾问集成电路高级分析师杨俊刚对《中国电子报》记者表示，SK海力士推出第四代1α工艺级别的DDR5，单芯片容量高达24Gb，在性能上比前一代DDR5要提升至少70%，功耗最多可降低20%。同时，SK海力士1α工艺级别的DDR5获得了英特尔中央处理器的兼容认可。在SK海力士收购英特尔的存储厂后，双方将进一步加强合作。

三星作为DRAM领域的领先企业，占据DRAM市场40%以上份额，研发DDR5的速度也比较快。早在2020年2月，三星就宣布成功研发出DDR5芯片。2022年12月21日，三星宣布利用12纳米制程工艺成功开发出16Gb DDR5 DRAM，近期与AMD完成了兼容性测试。三星表示，这款产品是业界最先进的高性能且低能耗的DDR5 DRAM。

据悉，此次他们推出的16Gb DDR5 DRAM所取得的技术突破是通过使用一种新的高介电（high-k）材料来增加电池电容，以及改进关键电路特性的专利设计技术来实现的。结合先进的多层EUV光刻技术，这款产品拥有三星最高的DDR5 Die密度，可使晶圆生产率提高20%，功耗有望节省约23%，最高支持7.2Gbps的运行速度，这意味着它可以在一秒钟内处理两部30GB超高清电影。2022年12月22日，三星表示，为了抢占逐渐扩大的DDR5市场，计划从2023年开始批量生产，并向数据中心和人工智能等领域客户供货。

美光则在1月19日宣布推出新一代DDR5内存模块，产品覆盖DDR5-5200/5600，最高拥有48GB容量的版本。据悉，美光新一代DDR5内存可以支持5200MT/s和5600MT/s的数据传输速率，以及1.1V电压下的CL46延迟，同时兼容AMD EXPO和英特尔XMP 3.0配置文件。该模块在与英特尔处理器兼容上性能提升了49%，配置最高可达48GB。

记者了解到，美光的新DDR5存储模块已在英特尔处理器芯片上完成了认证。美光还与AMD在奥斯汀建立联合服务器实验室，以缩短存储芯片的验证周期。杨俊刚指出，从合作对象来看，三大存储芯片厂商在DDR5方面主要与英特尔和超威半导体服务器处理芯片供应商合作，能够高效提升处理器芯片和存储器芯片之间的兼容性。

服务器DRAM市场复苏速度低于预期

DDR5作为新一代DRAM存储技术，与DDR4相比在传输速率和容量方面出现较大突破。因此，厂商抢先布局DDR5，是构筑自身DRAM技术优势的战略需求。

但业界对DDR5 DRAM的市场预期不是十分乐观。受到全球经济前景不明朗、市场持续低迷、客户库存水位较高等因素影响，服务器DRAM市场的复苏速度将低于预期。Omdia预计，今年服务器DRAM中采用DDR5 DRAM芯片的速度将比预期慢。此前，市场预计DDR5将占据服务器DRAM市场的28%，但当前这一数字已被调整为13%。Omdia的数据显示，2023年第一季度，DDR5 DRAM将仅占服务器DRAM市场的3%，第二季度为8%，这一比例将在第三季度跃升至15%，在第四季度将达到24%。

服务器DRAM预估DDR5第一季价格跌幅约18%~23%，略高于DDR4，但DDR5第一季度导入率仅约10%，故服务器DRAM价格下跌幅度主要还是由DDR4决定，预估跌幅约15%~20%。

目前存储器市场主要表现为市场需求疲软、库存仍处于高位，导致内存芯片价格下滑，存储器厂商整体营收下降。目前，市场下滑趋势还未终结，预计将持续一段时间。专家认为，这对DDR5 DRAM的市场渗透速度有较大影响。此外，由于目前DDR5成本较高，主要应用领域是对算力要求较大的数据中心、元宇宙、人工智能等领域的服务器产品，目前DRAM市场主流产品还是以DDR4为主，DDR5替代DDR4产品还需要一定时间。

芯谋研究分析师张先扬向《中国电子报》记者表示，DDR5并非是存储芯片市场的救命稻草。一方面，受服务器和PC终端市场萎靡影响，去年DRAM存储器市场承压明显，整体市场出现较大下滑，DDR5对比DDR4利润率和溢价空间更大，DDR5渗透率增长将抬升DRAM整体市场ASP；另一方面，从实际市场情况来看，2022年DDR5对比利基型DRAM（DDR4及以下）市场整体降幅更大，主要原因还是终端需求的萎靡限制了高价格DDR5的渗透和应用推广。同时，DDR5的应用需要CPU芯片的支持，目前已发布的可支持DDR5的CPU产品还比较少，这也是限制DDR5渗透的重要原因。

但也有很多专家认为，随着DDR5产品性能的大幅度提升，以及技术不断成熟、成本持续降低，DDR5产品未来应用领域将继续扩张。

TrendForce集邦咨询分析师吴雅婷就十分看好DDR5的发展。在她看来，预计自2023年起，服务器端将逐步导入DDR5，在服务器新平台的带动下，将抬高DDR5比重。DDR5有望取代DDR4，DDR5 DRAM将迎来快速普及期，成为市场中采用的主流产品。TrendForce集邦咨询认为，DDR5与DDR4成为主流应用的交汇点应该在2024年年底前或2025年年初。

相信在产品单价、产能达到要求，英特尔和AMD等厂商的积极应用推动下，DDR5会取代DDR4成为DRAM的主流产品，三家企业在DDR5 DRAM领域的竞争将变得更加激烈。

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