今日头条 (China)

In-depth report on storage industry: Computing power demand drives HBM market to grow several times – November 22, 2023

存储行业深度报告：算力需求推动HBM市场数倍增长

2023-11-23 11:20·未来智库

（报告出品方/作者：招商证券，鄢凡、曹辉）

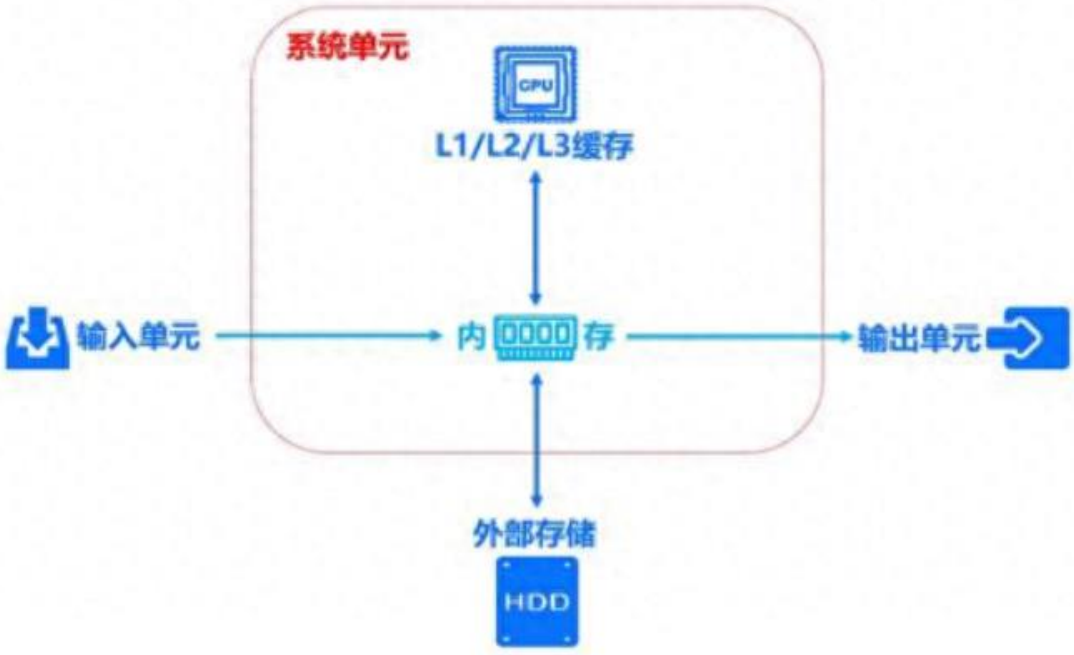
一、AI 服务器相较普通服务器算力明显提升，存储 量价均有数倍增长空间

1、传统服务器采用 CPU 作为算力来源，一般配置 DDR4 内存和 SATA/SAS 等类型 SSD

传统服务器以 CPU 作为算力来源，采用提高核心数来提升算力。传统服务器组 件包括 CPU、内存、硬盘、RAID 卡、网卡等，搭配 BMC 管理系统，以及电源、 主板、机箱等基础硬件。按照外形分类，服务器可分为塔式、刀片、机架和高密 度服务器，按照处理器数量，可分为单路、双路及多路。传统服务器以 CPU 提 供算力，擅长逻辑和浮点型计算等，服务器的算力提升主要依靠 CPU 增加核心 数量。

CPU 和内存是普通服务器逻辑架构中最重要的部分，等效频率是内存的核心指 标。在服务器中，CPU 负责对数据进行逻辑运算，随着 CPU 主频、核心和线程 数量的不断提升，CPU 运算处理的速度已经远超硬盘的读写速度，因此需要 CPU 内存作为桥梁来暂时存放 CPU 中的运算数据，并与硬盘等外部存储器交换数据，CPU 和内存之间通过内存总线进行通信。由于内存是能够传输和储存数据的中 转站，传输速度和处理性能尤其重要，核心指标为频率。内存的频率分为三种， ①核心频率：指内存颗粒核心的时钟频率，核心频率是固定的；②工作频率：指 内存颗粒总线的时钟频率；③等效频率：指内存实际的数据传输率，也被称为内 存速率。

图 2：CPU 内存工作原理



服务器的 DRAM 类型伴随 CPU 而升级，目前服务器标配 DDR4 或 DDR5。 SDRAM 最早出现，为了与 CPU 的计时同步而设计，在一个周期内，SDRAM 只能在升的时候进行数据传输，即只能读写一次；改进后的 DDR 能够在升和降 **两个**阶段进行数据传输，所以等效频率相较 SDRAM 增加 1 倍；DDR1 到 DDR3 的升级主要在于预读取的能力，预读取从 1bit 上升至 8bit，**每**进步一代，预读取 能力增加一倍，相当于一次读写的数据数量增加一倍，即工作频率翻倍；相较 DDR3，DDR4 采用了 2-4 个 Bank Group 组的设计，即在同一脉冲工作周期内， DDR4 最多可以处理 2-4 组数据，相当于等效工作频率提升 2-4 倍；DDR5 将预 读取数提升至 16bit，等效频率再次翻倍。从当前服务器型号来看，传统服务器 的 CPU 内存一般标配 DDR4，最先进的 AI 服务器 CPU 则配备了 DDR5 或 LPDDR5。

GDDR 在 DDR 的基础上发展形成，主要用于 PC 端的显卡。GDDR/GDDR2 基 于原有的内存技术，但无法满足 GPU 对高性能存储的需求，进而推动了 GDDR3 作为第一款专为 GPU 而设计的存储器的诞生。GDDR3 采用 4-bit 预读取，能够 实现读写操作的快速切换；GDDR4 采用 8-bit 预读取技术，但由于 NVIDIA 不愿 跟进，最终并未得到广泛应用；GDDR5 使用了双 DQ 总线，实现了双倍的接口 带宽，自 GDDR5 **开始**，GDDR 广泛用于 PC 显卡中，但由于传统服务器中不配 备 GPU，因此 GDDR 并未在传统服务器中得到使用。

服务器硬盘主要包括 HDD 和 SSD，SSD 能够大幅提升服务器性能。目前服务 器的硬盘主要包括 SATA/SAS 等机械硬盘（HDD）以及固态硬盘（SSD）等， HDD 主要由机械驱动，包括马达、盘片、磁头摇臂等部件，尽管容量不断提升， 但磁盘记录方式并未改变，存储系统的发展越来越落后于 CPU 的发展速度；SSD 固态硬盘没有机械结构，主要由主控-内存-固件算法**构成**，SSD 以区块写入和擦 除的方式实

现读写的功能，具备低功耗、稳定性高、耐震耐低温等特点，可以大 幅提升服务器性能，更适合工业自动化、嵌入式移动设备和 AI 推理/训练服务器 等应用。

RAID 卡可实现对固态硬盘数据的冗余保护，已成为 AI 服务器的重要配置。RAID （Redundant Array of Independent Disks），即独立硬盘冗余阵列，意为很多块 独立磁盘构成的磁盘组，主要具备如下功能：①通过对硬盘上数据的条带化，实 现对数据成块存取，减少硬盘的机械寻道时间，提高数据存取速度；②通过对一 个阵列的几块硬盘同时读取，减少硬盘的机械寻道时间，提高数据存取速度；③ 通过镜像或存储奇偶校验信息的方式，实现对数据的冗余保护。由于硬件 RAID 是采用集成的阵列卡或专用的阵列卡来控制硬盘驱动器，因此可以极大节省服务 器系统 CPU 和操作系统的资源，目前 RAID 卡已成为 AI 服务器的标配。

2、AI 服务器中 CPU+GPU 协同工作，存储容量和价值量相 较传统服务器均有数倍提升

AI 训练模型算力提升速度突破极限，目前英伟达训练型 AI 服务器一般配备 8 个 GPU。随着以 chatGPT 为代表的 AI 的发展，训练 GPT-3、Megatron-Turing NLG 530B 等超大语言模型所要求的算力提升速度已经突破了后摩尔定律算力提升速 度的极限，尽管 CPU 不断升级，但 CPU 制程以及单个 CPU 和核心数量接近极 限，仅依靠 CPU 无法满足算力需求。CPU 的内核数量大约数十个，但 GPU 具 备成千上万个 CUDA 核心，因此 GPU 多个内核决定了其能够在相同的价格和功 率范围内，比 CPU 提供更高的指令吞吐量和内存带宽，GPU 能够并行执行成千 上万个线程（摊销较慢的单线程性能以实现更大数据吞吐量）。在训练 AI 模型的 过程中，需要同时对所有样本数据执行几乎相同的操作，GPU 架构设计能够很 好满足 AI 场景需求。AI 服务器相较通用服务器的一个明显差别之一是增加了 GPU，通用服务器一般含有 1-2 个 CPU、不含 GPU，而当前英伟达训练型 AI 服务器一般搭载 8 个 GPU。

AI 服务器 GPU 需要 CPU 来进行指令，模型算力提升带动 CPU 核心、主频等提 升。在 AI 服务器的 GPU 模式下，模型训练一般分为 4 步，①将输入数据从系统 内存拷贝到显存；②CPU 指示 GPU 处理数据；③GPU 并行完成一系列计算； ④将计算结果从显存拷贝到内存。虽然 GPU 并行能力优异但无法单独工作，必 须由 CPU 进行控制调用，CPU 可以独立工作并直接访问内存数据完成计算。因 此在 AI 服务器中，GPU 和 CPU 需要协同工作，训练模型所需算力升级也将带 动 CPU 技术升级，例如在英伟达 DGX-2 服务器中，采用英特尔第三代至强处理 器 8168，主频大约 2.7GHz，核心数量为 24 个；在英伟达 DGX H100 服务器中， 搭载英特尔第四代至强处理器 8480C，主频提升至最高 3.8GHz，CPU 核心数量 提升至大约 56 个。

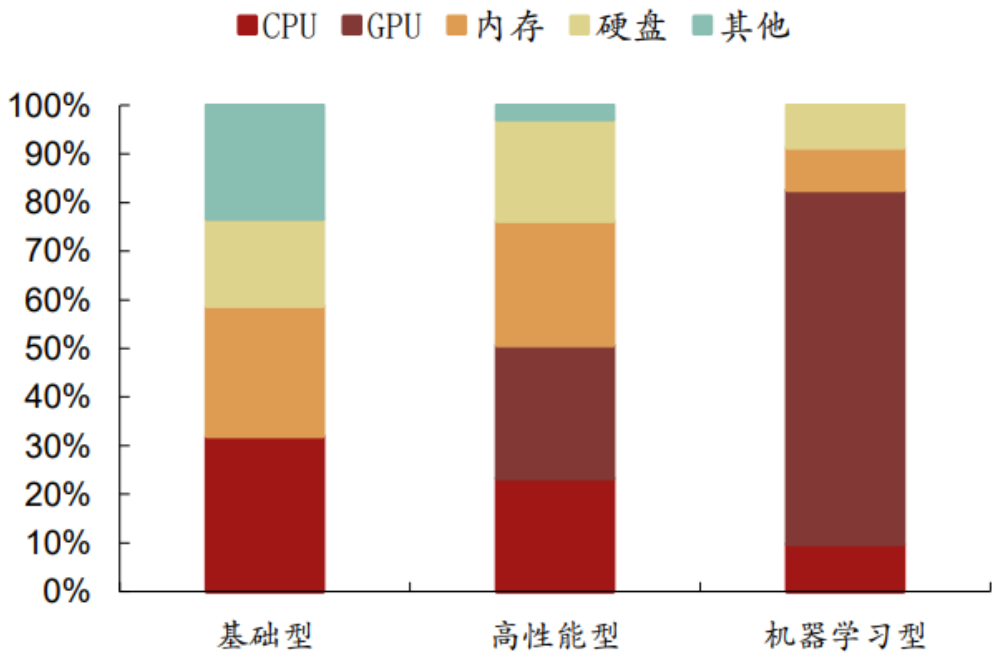
AI 服务器存储器容量伴随 CPU/GPU 的升级而提升，相较传统服务器有数倍提 升。最先进的 AI 服务器尽管增加了大量 GPU 需求，但存储器的数据存储方式、 总线连接方式均和普通服务器相近，CPU 的运行数据写入 DRAM 中，CPU 和 GPU 产生的数据共同写入 NAND 中。AI 服务器将提升内存、显存的工作频率和 带宽等，带动存储容量明显上升。

DRAM：AI 服务器 CPU+GPU 相较普通/高性能服务器对 DRAM 容量大约 有 3-8 倍的拉动。当前服务器一般配 1-2 个 CPU，每个 CPU 一般标配 DDR4 内存，普通双路服务器用于打印、网络管理、数据记录等场景，一般配备 1-2 个 CPU，每个 CPU 周围搭配大约 10 个内存插槽，按照单个 DDR4 32GB 容量计算，每个 CPU 搭配的内存容量大约不超过 320GB，系统容量不超过 640GB；根据浪潮官网，高性能服务器用于涉及高密度计算的企业端，例如 搭载双路 CPU 的 NF5280M6 型号服务器单个 CPU 插槽数量为 16 个，按 照每个 DDR4 32GB、插满内存条计算，系统的内存容量可达 1TB；英伟达 AI 服务器 CPU 均标配 2TB 的 DDR4/DDR5；另外，AI 服务器增加了 GPU 的数量，每个 DGX A100 和 H100 均需要搭配 80GB 容量的 HBM，系统整 体需要 640GB 的容量。综上所述，从容量上看，按照普通服务器 CPU 内 存 320-640GB 容量计算，AI 服务器对 DRAM 容量的提升大约为 4-8 倍； 按照高性能服务器 CPU 内存 1TB 容量计算，AI 服务器对 DRAM 容量的提 升大约为 2.5-3 倍。

NAND：AI 服务器相较普通/高性能服务器对 NAND 容量大约有 2-4 倍的拉 动。由于机械硬盘持续进行成本优化，普通服务器依然会配备较多的机械硬 盘，而高性能服务器和 AI 服务器对于存储速度、准确性等提出更高要求， 因此主要配备固态硬盘。从容量上看，传统服务器 NP5570M5 最多支持 16 块 2.5 英寸 SAS 硬盘或 4 块 3.5 英寸 SATA 硬盘，SAS 硬盘支持 300GB/600GB/1TB/1.2TB/1.8TB/2.4TB 容量，SATA 硬盘支持 1/2/4/6/8TB 容量，按照中间容量测算，系统硬盘容量大约共 8-16TB；高性能 NF5280M6 型号最大支持 20 块 3.5 英寸硬盘，系统硬盘容量大约 20TB；AI 服务器采 用的硬盘容量大约 30TB，相较传统服务器 NAND 容量提升大约 2-4 倍。

AI 服务器的 GPU 和存储价值量均明显提升，GPU 价值量占比最高。1）基础型服务器：根据 IDC，基础型服务器中 CPU、内存和硬盘的占比最高， 根据英特尔，典型的 x86 服务器—E5 高配服务器的成本中硬盘、CPU 和内存成 本占比分别为 31%、28%和 21%。例如浪潮 NF5270M5 售价大约 39000 元，采 用 2 颗 Xeon Silver 4214 CPU，每颗售价大约 1000 美金，CPU 成本占比大约 32%；配置 16 个 32GB DDR4，按单 GB 成本 3 美金计算，DRAM 成本占比大 约 26%；最大支持 25 块 2.5 英寸硬盘，按照每块硬盘 512GB、1TB SATA 40 美元计算，硬盘总价值量占比大约 20%； 2）AI 服务器：以英伟达 DGX A100 系统为例，售价大约 19.5 万美金，含有 2 颗 AMD Rome 7742，每颗售价大约 7000 美金，价值量占比大约 7%；含有 8 颗A100 GPU，每颗GPU售价大约1-1.5万美元，GPU价值量占比大约40-50%； CPU 采用 2TB 的 DDR4，价值量占比大约 3-4%；GPU 配置共 640GB HBM2E， 假设单 GB 价值量 15-20 美金，价值量占比大约 5-8%；操作系统配备 2 块 1.9TB SSD，内部配备 8 块 3.84TB SSD，按照 1 块 1.92TB 数据中心 SSD 650 美元、 1 块 3.84TB 数据中心 SSD 1200 美元测算，硬盘价值占比大约 6%。 综上所述，从价值量来看，AI 服务器中 CPU 内存和 SSD 的价值量相较于普通 双路服务器预计均提升 5 倍左右，GPU 显存为纯增量市场。

图 14：不同类型服务器成本占比



3、存算一体突破存储和功耗“双墙”瓶颈，催生 2.5/3D 封 装及新型存储器需求

传统服务器计算芯片与存储器采用并行架构，“存储墙”和“功耗墙”分别限制 数据访问速度提升和数据访问能耗下降。在冯-诺依曼架构中，计算和存储功能 分别由 CPU 和存储器完成，数据从存储单元外的存储器中获取，处理完毕后再 写回存储器，计算核心与存储器之间有限的总带宽直接限制了交换数据的速度， 处理器速度和访问速度的差异进一步减缓处理速度。一方面，处理器和存储器制 程、工艺不同，性能差距逐渐增大，存储器数据访问速度远低于 CPU 的数据处 理速度，即形成“存储墙”；**另一方面，数据搬运的能耗比浮点**计算高 1-2 个数 量级，DRAM 的访问功耗可达芯片内一级功耗的 50-100 倍，进一步增加了数据 访问功耗。

存算一体架**构可部分解决**“双墙”问题，主要包括近存计算和存内计算。1）近 存计算：主要包括光互联、2.5/3D 封装等形式，其中 2.5/3D 封装可以提高系统 集成度，增大并行带宽或利用串行传输提升存储带宽。典型的产品形态为高带宽 内存（High Bandwidth Memory，HBM），将内存颗粒通过硅通孔（TSV）多层堆叠实现存储容量提升，同时基于硅中介板的高速接口与计算单元互联提供高 带宽存储服务；2）存内计算：主要包括①存储内计算（In-Storage Computing， ISC），即在固态硬盘中等加入 FPGA、ARM 处理核等计算单元，承担如数据压 缩、搜索、视频文件转码等本地处理，减少远端处理器如 CPU 的负载；②内存 内计算（In-Memory Computing，IMC），数据直接在内存中计算后返回，通过将 存储层和逻辑层堆叠实现高带宽计算。典型产品形态包括 HBM-PIM 和 PIM-DIMM，在 DRAM die 中内置 AI 等处理单元，提供大吞吐低延迟片上处理 能力。

存内计算电路可基于易失性和非易失性存储器实现，未来 RRAM、MRAM 等新 型存储器件有望得到更多应用。易失性存储器包含 SRAM 等，非易失存储器包 含 NOR Flash、阻变随机存储器（Resistive Random Access Memory，RRAM）、 磁性随机存储器（Magnetoresistive Random Access Memory，MRAM）、相变 存储器（Phase Change Memory，PCM）等。 SRAM：常用于 CPU 中的缓存，其读写速度快、无读写次数限制，且工艺 成熟，器件的一致性和稳定性较好，缺点是价格相对昂贵、多晶体管单元结 **构下存**储密度较低、需要通电以保持数据，因此芯片面积大、功耗较高； NOR：基于 NOR 的存内计算芯片具有存储密度大、工艺成熟、成本低等优 点，业内主要利用其多比特存储特性进行模拟存内计算，相关产品在智能终 端与边缘场景实现小批量商用。目前 NOR 的问题在于工艺微缩存在一定瓶 颈，可能需要 2.5/3D 封装工艺配合扩展性能； 阻变随机存储器（RRAM）：基本存储单元一般为金属-绝缘体-金属/半导体 结构，通过在电极层施加电压/电流，电阻转变层的电阻值可以实现高阻态 和低阻态的切换，并可实现多级电阻状态，从而存储多比特信息。基于 RRAM 的存内计算芯片制备工艺简单、工艺成本低、支持多比特存储，但目前在器 件一致性和准**确性等方面需要提高**； 磁性存储器（MRAM）：利用磁隧道磁阻效应实现电阻状态改变，从而存储 信息，具有非易失、访问速度快、读/写次数高等优点，具备较高的可靠性和稳定性，但目前成熟度较低，功耗、速度和耐久性等指标有待提升； 相变存储器（PCM）：可以介于完全多晶态和完全非晶态之间的多种状态， 以此实现多值存储，具备存储密度高、多比特存储、支持 3D 堆**叠等**优点， 但 PCM 存在写入功耗较大、擦写次数较少等问题。

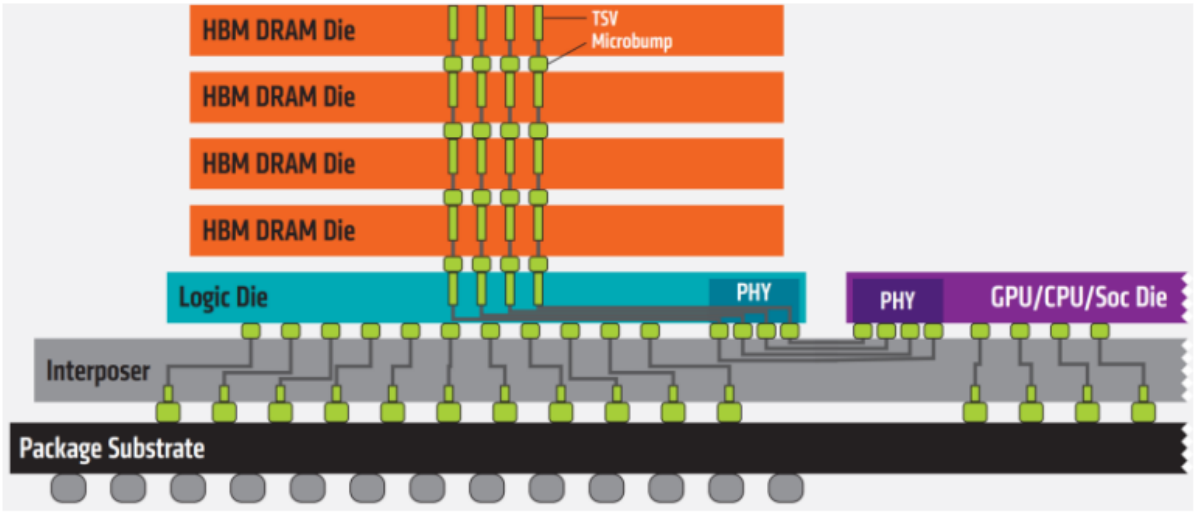
目前存内计算主要采用 NOR、SRAM、DRAM 等成熟存储介质，新型存储器在 大算力领域应用尚不成熟。目前 HBM 等近存计算已经规模商用，但大部分存内 计算采用模拟计算或使用 DRAM 等成熟存储介质，新型存储器落地场景仍相对 有限。在低功耗、低精度和中小算力场景，多使用 NOR、SRAM 等工艺；在高 精度、大算力的 AI 芯片中，目前 RRAM、MRAM 等新型存储器技术指标尚在提 升，因此新型存储器尚未规模应用，应用方案以 HBM、HBM-PIM 等为主，例如 三星、海力士等均采用 2.5D 封装形式，将 DRAM die 堆**叠形成** HBM，进而提高 集成度；三星于 2021 年推出结合 DRAM 的高带宽内存-内存内处理（HBM-PIM）， 将 AI 计算能力引入内存，通过将经过 DRAM 优化的 AI 引擎置于**每个存**储子单 元内，将处理能力直接引入到数据的存储位置，从而实现并行处理并尽可能减少 数据移动。

二、HBM 新型存储器突破带宽极限，训练型 AI 服务 器推动超百亿美金增量市场

1、HBM 用于高带宽和高容量场景中，已普遍搭配主流 AI 训练芯片使用

HBM 是基于 2.5/3D 封装技术的新型存储器，满足高带宽、高速度等需求。HBM （High Bandwidth Memory）指高带宽存储器，是一款新型的 CPU/GPU 内存芯 片，基于 2.5D/3D 封装技术将 DRAM Die 垂直堆叠，具备高带宽、高速度等特 点。DRAM Die 之间通过 TSV 的方式连接，逻辑控制单元对 DRAM 进行控制， GPU 和 DRAM 之间通过 uBump 和 Interposer（起互联功能的硅片）连通。目 前最先进的 HBM 为第五代 HBM3 以及第六代 HBM3E，封装的 DRAM Die 层数 达到 12 层。

图 17：HBM 结构图



GPU 显存一般采用 GDDR 或 HBM 两种方案，HBM 能够在实现高带宽和高容 量的同时节约芯片面积和功耗，非常适合在功率受限但又需要高带宽的场景中使 用。 1）传统的 GDDR 方案：主要通过提升 GDDR 的带宽实现性能的提升，但存在 如下问题：①GDDR 不断增长的功耗一定程度阻碍了图形性能的表现，未来可能 导致GDDR的性能表现无法匹配GPU的性能要求；②为了实现GPU的高带宽， GPU 需要搭配大量的 GDDR5/6 芯片，芯片面积不断增大，同时也需要更高的电 压。 2）HBM 方案：将原本在 PCB 板上的 DDR 内存颗粒和 GPU 芯片同时集成到 SiP 封装中，使内存更加靠近 GPU，DDR Die 采用堆叠方式，通过增加单个 DDR 容量和层数来实现内存容量和带宽的提高。HBM 具备如下优势，①大大节约芯 片面积：根据 AMD，1GB HBM 相较 1GB GDDR5 能够节省大约 94%的芯片面 积；②能够在更低的频率和更少的功耗下提供超越常规内存的带宽：HBM 通过 同一封装内的硅中介层与 SoC 集成在一起，能够克服数据 I/O 封装管脚限制的 最大数量，进而突破内存带宽的限制。例如，典型的 GDDR5 使用 32 位总线， 在 1750MHz、1.5V 电压下，最高带宽为 28GB/s，而一个 HBM 包使用 1024 位 总线，仅在 500MHz、1.3V 的电压下，就能达到超过 100GB 的带宽。

HBM 已普遍搭配主流 AI 训练芯片使用，相较 GDDR 带宽显著提升。HBM 最早 于 2013年由SK 海力士首先制造，第一颗采用HBM存储的 GPU是 2015 年AMD 的 Radeon R9 Fury X，第一颗应用 HBM2 的 GPU 是英伟达的 Tesla P100；目 前，先进的 AI 训练 GPU 芯片均搭载 HBM 存储芯片，例如英伟达的 DGX A100 单 GPU 搭载了总计 80GB容量的HBM2E，DGX H100 单 GPU 搭载了总计80GB 容量的 HBM3 存储，单个 HBM3 存储带宽最高可达 819GB/s，较单个 GDDR 带 宽提升 10 倍以上，H200 更是搭载了共 141GB 的 HBM3e 内存，总 HBM 容量 较 H100 提升 76%，总带宽提升约 43%。

2、HBM 容量、带宽、速度持续升级，训练型 AI 服务器将 带来 HBM 超百亿美元增量市场

HBM 目前最主要搭配 AI 的 GPU 使用，训练型 AI 服务器是最主要的增量市场。 根据 SK 海力士，受益于 HPC、AI、CPU 等应用，HPC 市场复合增长率达到 40% 左右，由于随着数据量呈指数级增长，以及 AI/ML 训练等高级工作负载的快速增 长，预计 AI 服务器将是几年内 HBM 最大的下游市场。 训练型 AI 服务器对于 HBM 市场的拉动体现在：①AI 服务器渗透率的提升；② 单个 AI 服务器搭载 GPU 数量的增加；③每个 GPU 搭载的 HBM Stack 数量增 加；④单个 HBM 堆叠的 DRAM die 层数（Hi）增加，以及 DRAM die 容量增 加；⑤HBM 代际进步带来单 GB 价值量的提升。

AI 训练服务器渗透率提升：根据 Trendforce，2022 年全球服务器出货大约 1430 万台，其中训练侧 AI 服务器占比仅为 1%；由于四大 CSP 陆续下调 采购量，Dell 及 HPE 等 OEM 厂商在 2-4 月期间下调全年出货量至同比分 别减少 15%和 12%，Trendforce 预估 2023 年全球服务器整机出货量将同 比减少 2.85%至 1383.5 万台。同时，考虑到云服务厂商积极投入生成式 AI， AI 服务器渗透率预计将快速提升； 1）GPU 基板搭载 HBM 数量提升：搭载 HBM1 的 GPU 基板上仅采用 4 个 HBM1 芯片，而英伟达 AI 服务器 DGX A100 和 H100 的基板搭载了 8 个 GPU，每个 GPU 搭配 6 个 HBM2e； 2）单个 DRAM die 密度增加、HBM 堆叠层数和容量提升：根据 SK 海力 士官网信息，从 HBM1 进步到 HBM3，单个 DRAM Die 密度从 2Gb 提升至 16Gb，堆叠高度从 4Hi 提升至最高 12Hi，单个 HBM 容量从最高 1GB 提升 至 24GB；3）HBM 性能提升：HBM1 提升至 HBM3，最高速度从 1Gbps 提高到高于 6Gbps，带宽从 128GB/s 提升至最高 4.8TB/s。例如，SK 海力士的 HBM3 将 DRAM 芯片高度减少至大约 30 微米，相当于 A4 纸厚度的 1/3，使用 TSV 技术（在 DRAM 芯片上打上数千个细微孔并通过垂直贯通的电极连接上下 芯片）垂直连接 12 个芯片，相较于 HBM2e，HBM3 的速度提高了大约 78%， 容量提升了 1.5 倍，并且大大降低了工作温度； 4）价值量较传统 DRAM 提升：传统 DDR4 每 GB 价格大约数美元，HBM 价格大约为其数倍，HBM3、HBM3E 相较前几代产品价值量也大大提升。

2024 年训练型 AI 服务器对 HBM 增量空间预计超百亿美元。由于推理型 AI 服 务器对 GPU 和 HBM 需求量相对较少，因此我们仅考虑训练型 AI 服务器带来的 HBM 市场增量。假设 2024 年训练型 AI 服务器渗透率达 4%，单个训练型 AI 服 务器搭载 8 个 GPU，目前 DGX A100 和 H100 的每个 GPU 配 6 个 HBM，总容 量为 80GB；考虑到改进前的 GH200 搭载的 HBM 总容量为 96GB，H200 搭载 的 HBM 总容量升至 141GB，考虑到 GPU 算力持续提升，因此我们合理假设平 均每个 GPU 搭载 HBM 容量持续提升；目前 HBM 持续供不应求，单价快速增长， 但目前 HBM 产品以 HBM2 及 HBM2E 为主，考虑到 2024 年更高价值量的 HBM3 将持续放量，因此假设 2024 年价格继续上涨；考虑到 2025 年 HBM 产能逐步开 出，成本有望持续下降，我们假设 2025 年价格有所回落。综上所述，我们预计 2024 年 AI 服务器有望带动 HBM 超百亿美金增量市场。

英伟达在 COMPUTEX 2023 大会上公布了 GH200 Grace Hopper，每个 GPU 使用 96GB 的 HBM3，搭载的 HBM 容量继续提升。由 NVIDIA GH200 Grace Hopper 超级芯片和 NVIDIA NVLink Switch System 驱动的 NVIDIA DGX 超级计 算机，旨在助力开发面向生成式 AI 语言应用、推荐系统和数据分析工作负载的巨型、下一代模型。NVIDIA DGX GH200 的超大共享内存空间通过 NVLink 互连 技术以及 NVLink Switch System 使 256 个 GH200 超级芯片相连，使它们能够 作为单个 GPU 运行。其提供 1 exaflop 的性能和 144 TB 的共享内存——相较 2020 年推出的上一代 NVIDIA DGX A100 内存大了近 500 倍。H200 单 GPU 总 共支持 576GB 的内存，其中 480GB 为 LPDDR5X，从 HBM 容量来看，A100 和 H100 单 GPU 最大支持 80GB HBM 容量，H200 单 GPU 最大支持 96GB 容 量的 HBM3。

英伟达于 11 月 13 日上午在“Supercomputing 23”会议上正式发布了全新的 H200 GPU 和改进后的 GH200，每个 H200 GPU 搭载 HBM3e 容量进一步提升 至 141GB。H200 依然是建立在现有的 HopperH100 架构之上，但增加了更多高 带宽内存（HBM3e），全新的 H200 提供了总共高达 141GB 的 HBM3e 内存，有 效运行速度约为6.25Gbps，六个HBM3e堆栈中每个GPU的总带宽为4.8TB/s。 与上一代的 H100（具有 80GB HBM3 和 3.35TB/s 带宽）相比，HBM 容量提升 了超过 76%。改进后的 GH200 预计将包含总计 624GB 内存，相较改进前的 GH200（总计 576GB 内存），总内存容量提升了 8.3%。

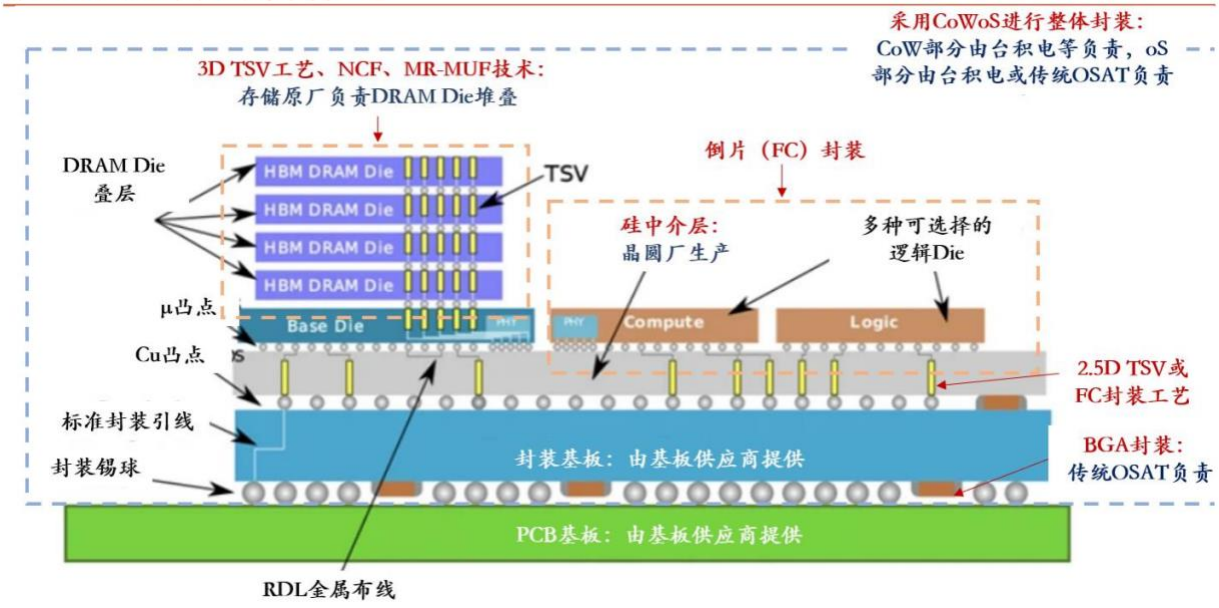
长期来看，HBM 在 AR/VR 设备、汽车自动驾驶域等领域也将有更多应用。1） AR/VR 设备：以 HTC VIVE pro2 为例，其连接 GTX1060，当前并不需要 HBM等高带宽产品，未来有望伴随显示需求增多逐步产生高带宽需求；2）汽车：当 前自动驾驶域并不需要 HBM 的高带宽，比如英伟达 Jetson AGX Orin 采用 256 位的 LPDDR，带宽最高为 204.8HB/s。但长期来看，伴随 L3/L4 等级别自动驾 驶汽车逐渐落地，也将提升 HBM 需求，例如 L3 级别自动驾驶带宽要求一般为 600GB/s-1TB/s 之间，单车可用 2 个 HBM2e 或者 1 个 HBM；L4 级别自动驾驶 带宽要求一般为 1-1.5TB/s 之间，单车可用 3 个 HBM2e 或 2 个 HBM3 。

三、HBM 采用 2.5D+3D 封装，核心工艺包括 TSV、 CoWoS 等

1、HBM 和 Si 中介层封装远期市场预计数倍增长，核心封 装工艺包括 Bumping、TSV、CoWoS 等

HBM 采用 2.5D+3D 封装工艺，采用的核心封装工艺包括 Bumping、RDL、FC、 TSV、CoWoS 等。封装工艺主要有四项功能：①保护芯片免受外部冲击或损坏； ②将外部电源传输至芯片，保证芯片的正常运行；③为芯片提供线路连接，以便 执行信号输入和输出操作；④合理分配芯片产生的热量，确保其稳定运行。HBM 采用“2.5+3D”封装工艺，即分别在 DRAM die 上和硅中介层上制作 TSV，主 要使用的封装工艺如下： 1）HBM 由 DRAM Die 堆叠而成，Die 之间借助 TSV（硅通孔）和微凸块 （microbump）相连，并借助 TC（Thermal Compression，热压）-NCF （non-conductive film，非导电薄膜）或 MR-MUF（批量回流模制底部填充）技 术，用 NCF 或 LMC（Liquid Molding Compound，液态塑封料）对芯片形成互 连和保护； 2）HBM 通过 TSV 和微凸块连接下方的 HBM 逻辑控制 die（Base Die），逻辑 控制 die 再通过凸块（μbump）连接下方的硅中介基板（Si interposer）；3）HBM Stack 通过 CoWoS 等 2.5D 封装工艺，和 CPU/GPU 等并排铺设在硅 中介基板上方，CPU/GPU 等逻辑 die 采用倒片封装（FC）形式和硅中介基板连 接，存储器和 GPU 等逻辑芯片之间通过 RDL 实现通信； 4）硅中介基板通过 2.5D TSV 工艺和 Cu Bump 连接至封装基板（Package Substrate）上；4）最后，封装基板通过封装锡球（package balls）和下方的 PCB 基板（Circuit Board）相连，这一过程采用传统的球栅阵列（Ball Grid Array，BGA）封装工艺。

图 27：HBM 结构图及用到的封装工艺



2021 年针对 HBM 和 Si Interposer 的高端高性能封装市场规模合计约 14 亿美 元，预计 2027 年增至约 35 亿美元。根据 Yole，高端高性能封装主要包括 UHD FO、嵌入式硅桥、硅中介层（Si Interposer）、3D 堆栈存储器和 3D SoC，嵌入 式硅桥主要解决方案包括台积电的 LSI 和英特尔的 EMIB；硅中介层通常由台积 电、三星和联电提供，还包括英特尔的 Foveros 方案；3D 堆栈存储器分为 HBM、 3DS 和 3D NAND 堆栈。Yole 预计 2021 年高端高性能封装市场规模约 27.4 亿 美元，其中 HBM 和硅中介层市场封装规模分别为 4.35 亿美元和 9.64 亿美元； 到 2027 年，整体市场规模预计增至 78.7 亿美元，其中 UHD FO、HBM、3DS 和硅中介层封装将在 2027 年占据超 50%的市场份额，其中 HBM 和硅中介层封 装预计分别增至 16.3 亿美元和 18.8 亿美元，CAGR 分别为 25%和 12%。

2、凸块制造（Bumping）是扇入/扇出、FC 等封装必要环 节，推动沉积、涂胶显影、电镀等前道工艺需求增长

晶圆级封装采用凸块（Bump）取代键合引线，凸块可以分布在整个芯片表面形 成信号触点。凸块工艺即在晶圆切割成单个芯片之前，在晶圆上以整个晶圆的形 式形成由焊料组成的“凸块”或“球”，这些凸块是芯片和基板互连形成单个封 装的基本互连组件。传统的引线键合工艺中，接合焊盘/引脚放置在外围区域， 但用于凸块的 I/O 焊盘可以分布在芯片的整个表面，每个凸块都是一个信号触点， 从而可以缩小芯片尺寸并优化电气路径，因此凸块工艺广泛用于倒装芯片封装、 扇入或扇出型封装工艺。

凸块可以由共晶、无铅、高铅材料或晶圆上的铜柱组成，高端应用多采用铜柱作 为凸块。含铅焊料作为凸块有多年历史，但已经被锡基无铅焊料取代并用于倒装 封装等工艺，而在 AI、HPC、基带、高性能存储等应用场景，铜柱凸块（Cu Pillar Bump，CPB）逐渐取代无铅锡球，主要系 CPB 的精细节距（Fine Pitch）更小， 无铅锡球的 Fine Pitch 一般为 130-250um，铜柱凸块的 Fine Pitch 可缩小至130-40um 左右，采用铜柱凸块的芯片尺寸较锡球能够缩小 5-10%，衬底层可从 6 层减少到 4 层，基板成本下降约 30%；铜柱凸块还具有改善的电子迁移阻力， 能够减少功率损耗和信号延迟，同时实现更好的散热性能。无铅锡球整体由锡金 属**构成**，铜柱则分为三层结构，包括铜金属柱、一层薄镍金属作为铜柱的扩散屏 蔽层、锡/银金属作为焊帽。

凸块工艺（Bumping）流程为制备 UBM 和形成凸块，主要使用 PVD 设备、涂 **胶**显影机、电镀机、刻蚀机、回流炉等设备。在制作凸块前，首先需要在芯片表 面需要生长凸块的区域制作凸块下金属（Under Bump Metallization，UBM）， 用于阻挡凸点金属扩散至下层金属焊盘中；在完成制备 UBM 后，凸块的制备可 以采用三**种方式**：电镀、印刷锡膏固化和植球方式，电镀方式形成铜柱等凸块， 在电镀焊料、去除光刻胶和刻蚀掉溅射金属后，通过回流进行固化；印刷锡膏固 化方式通过将含焊锡料的改性锡膏通过贴片、回流焊等加热固化；植球方式则通 过植球机将预成型的凸块印制在 UBM 上。

具体工艺步骤如下： 1）**沉**积阻挡层、溅射 Ti/Cu：采用溅射或其他物理气相沉积方式在晶圆表面沉 积一层钛或钛钨作为阻挡层，再沉积一层铜或其他金属作为后续电镀所需的种子 层。阻挡层用于阻挡芯片表面焊盘金属与种子层间金属相互扩散。在沉积金属前， 晶圆先进入溅射机台的预清洁腔体，用氩气等离子去除焊盘金属表面的氧化层， 以提高溅射金属层与芯片表面的结合力； 2）**涂胶**显影：在晶圆表面旋涂一定厚度的光敏光刻胶（UV **胶**，Ultraviolet Rays）， 并应用光刻曝光工艺，对光刻胶进行选择性曝光，光刻系统在 UV **胶**顶部形成预 定凸块尺寸的图案； 3）电镀、刻蚀：前一步预定凸块尺寸的图案被刻蚀，形成一个小间隙；使用电 化学**沉**积（ECD）或电镀设备在间隙填充或镀铜；再剥离 UV **胶并刻**蚀结构； 4）回流或加热：最后，形成的凸块结构在回流炉中回流和加热，形成凸块。

3、TSV 用于 DRAM Die 之间及 HBM 和 Interposer 之间的 互联，为电镀、键合等工艺带来较多增量

TSV 技术专为 2.5/3D 封装而生，能够以最低的能耗提供极高的带宽和密度。TSV （Through-Silicon Via）即硅通孔技术，是目前最先进的封装技术之一，与传统 的 SIP 等封装技术相比，TSV 的垂直连接可以允许更多数量的连接，因此具备 更好的电势能、更低的功耗、更宽的带宽、更高的密度、更小的外形尺寸、更

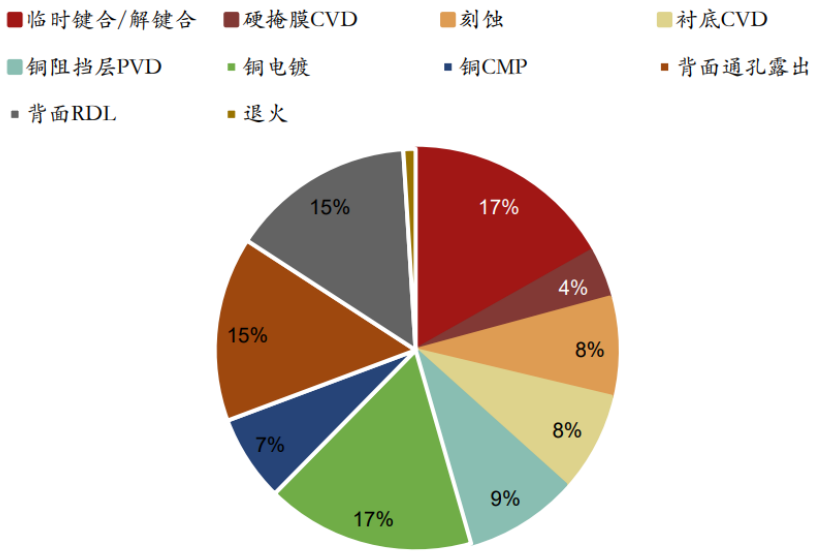
轻 的质量等优势，是实现电路小型化、高密度、多功能化的首选解决方案。2.5/3D TSV 技术已经广泛用于 AI GPU 基板上的 HBM 中，DRAM 各层 Die 之间的连接 采用 3D TSV 工艺，HBM 芯片和下方的金属凸块之间的连接采用 2.5D TSV 工 艺。

TSV 工艺价值量在 HBM 3D 封装工艺中占比最高。3D TSV 是 HBM 堆叠工艺中 成本占比最高的工艺，根据 3DinCites，考虑 4 层 DRAM die 和 1 层逻辑 die 堆 叠的 HBM 结构，在 99.5%和 99%的芯片键合（die bonding）良率下，TSV 制 造和 TSV 通孔露出工艺分别占其成本的 30%和 29%。

TSV 工艺主要包括前段的成型制造工艺和中段的露出及金属化工艺。晶圆厂一 般负责 TSV 的成型工艺，主要包括刻蚀和沉积，封装厂一般负责中段制程 （MEOL），即硅通孔露出和背面金属化工艺，主要包括减薄、钝化和键合等， 以及最后的封装工艺。1）刻蚀：主要使用深反应离子刻蚀（DRIE，Deep Reactive Ion Etching）法，又称博世法，该技术在真空系统中利用分子气体等离子的诱导 化学反应来实现各向异性刻蚀；2）沉积：在刻蚀完成后，需要利用 PECVD 法 沉积二氧化硅或氮化硅，以形成绝缘层对硅衬底进行电气隔离；之后，需要制作 金属扩散阻挡层来防止铜原子扩散、制作种子层方便后续铜的沉积，一般采用 PVD 沉积 Ti/Ta、采用 CVD 沉积 TiN/TaN 等；3）电镀：在沉积完成后，需要电 镀方法形成铜，一般采用自下而上的方法（Bottom-up）；4）减薄抛光：在电镀 完成后，需要对晶圆进行减薄抛光来使 TSV 背面的铜层快速露出，一般采用机 械研磨等方法将晶圆减薄至 50μm 甚至更薄，并采用 CMP 等技术露出铜通孔中 的铜柱；5）晶圆键合：用来完成通孔金属化和连接端子的晶片之间的互联。

TSV 工艺中临时键合/解键合、铜电镀、背面通孔露出成本占比最高。在 via-middle 的 TSV 工艺中，临时键合/解键合和铜电镀成本占比均为 17%左右，背面通孔露 出（主要包括背面减薄和抛光）和背面 RDL（主要包括刻蚀、CVD、PVD、电 镀等）成本占比约为 15%左右，其他关键工艺包括刻蚀、CVD、铜阻挡层 PVD 等。

图 37：TSV 制造成本构成（via-middle 工艺）



4、CoWoS 用于英伟达 AI GPU 的整体封装，生成式 AI 有 望推动 2024 年 CoWoS 需求翻倍以上增长

HBM 和 GPU 等芯片采用 TSMC CoWoS 技术进行整体封装，能够减少封装体 积和功耗等。目前立体封装主要包括 2.5/3D 封装，2.5D 封装主要指将存储、记 忆体等其他芯片，并排铺设在硅中介板（Silicon Interposer）上，经过微凸块（Micro Bump）连接，以实现硅中介板上不同芯片之间的通信；再通过硅穿孔（TSV） 来连接下方的金属凸块，再由导线载板连接外部；3D封装致力于实现芯片如HBM 中 DRAM die 的垂直堆叠。目前，英伟达 AI GPU DGX A100、H100、H200 等 均采用台积电的 CoWoS（Chip on Wafer on Substrate）技术，首先将芯片通过 CoW（Chip on Wafer）工艺连接至硅中介层（Interposer），再通过 OS（On Subtrate）工艺将 CoW 芯片和基板连接，整合成 CoWoS。

按照中介基板（interposer）材料的不同，CoWoS 可分为 CoWoS-S/R/L 三种 形式。CoWoS（Chip on Wafer on Substrate，晶圆级封装）是一种高集成度封 装技术，能够在中介基板（Interposer）上密集防止多颗芯片。从工艺流程上来 看，已知良品芯片（Know Good Die，KGD）首先通过 Chip on Wafer（CoW） 的封装制程连接至硅晶圆，再将 CoW 芯片与基板相连并整合成 CoWoS。台积 电根据中介基板的不同，将 CoWoS 技术分为三种类型。 1）CoWoS-S：是基础的 CoWoS 技术，使用 Si 衬底作为中介层，可以支持超 高集成密度，提供不超过两倍掩膜版尺寸的硅中介层，是 HBM 等高速高带宽内 存芯片使用的工艺； 2）CoWoS-R：引入 RDL 作为中介层，RDL（Redistribution Layer，重新分配 层）是集成电路上形成的额外金属布线层，旨在重新排列 I/O 焊盘，将焊盘重塑 到所需位置，以便于在必要时操作焊盘。这种工艺具备灵活性，增加了封装连接 的可靠性，允许新封装可以扩大尺寸以满足更复杂的功能需求，有效支持多个 Chiplet 之间进行高度可靠互联。例如，芯片中心的凸块阵列可重新分配到靠近 芯片边缘的位置。重新分配焊盘有助提高接触密度；3）CoWoS-L：在 CoWoS-S 和 InFO 技术的基础上，引入 LSI（Local Silicon Interconnect）技术，LSI 芯片在每个产品中可以具有多种连接结构（例如从 SoC 到 SoC、SoC 到小芯片、SoC 到 HBM 等），提供更灵活和可复用的多芯片互联 架构。

台积电 CoWoS 需求持续强劲，展望 2024 年产能将增至目前 2 倍。台积电在 CoWoS 工艺处于领先地位，根据 Yole，2021 年全球领先厂商在高端高性能封 装上的资本支出总投入约 119 亿美元，其中 TSMC 资本支出为 30.5 亿美元，主 要包括①InFO 工艺，用于 UHD FO；②CoWoS 工艺，用于 RDL 或硅中介层等； ③LSI 平台，和英特尔的 EMIB 直接竞争。由于 Google TPU、英伟达 GPU 和 AMD MI300 等均导入生成式 AI，台积电 AIGC 订单增多，先进封装需求持续提 升，其中英伟达是 A100 和 H100 级数据中心 CoWoS 封装的最大客户，谷歌通 过博通成为 TPUv4 和 v5 的第二大客户，AMD 在 MI300 上采用 CoWoS，亚马 逊的 Trainium through AI Chip 及微软最新 AI 芯片也使用 CoWoS 工艺，国内壁 仞科技于 2022 年 8 月发布首款通用 GPU 芯片 BR100，采用 7nm 工艺，也使用 Chiplet 与 CoWoS 2.5D封装工艺。台积电总裁魏哲家于 6 月 6 日股东大会表示， 自去年以来 CoWoS 需求几乎翻倍增长，明年需求将持续强劲。由于目前市场需 求大于产能，台积电表示当前首要任务是增加 CoWoS 产能，必要情况可能会通 过转厂方式生产；台积电于 23Q3 法说会上表示，2024 年 CoWoS 产能将翻倍。

5、FC 工艺用于逻辑 die 与硅中介层连接，相较传统引线键 合方式大大缩短信号路径

倒片封装（Filp Chip，FC）将芯片翻转，在先进封装工艺中取代引线键合。倒 装键合（Flip Chip Bonding）指通过在芯片顶部形成凸点，并将芯片倒转过来， 实现芯片的有源区域和基板间的电气和机械连接。倒片封装在先进封装工艺中取 代引线键合，主要系①引线键合对于可进行电连接的输入/输出（I/O）引脚的数 量和位置有限制，而倒片封装不存在此种限制；②倒片封装的电信号传输路径短 于引线键合。在引线键合方法中，金属焊盘在芯片表面采用一维方式排列，无法 出现在芯片边缘或中心位置，而倒片键合方法在键合至基板或形成焊接凸点的过 程中不存在任何工艺方面的限制，因此在倒片封装中，金属焊盘可以采用二维方 式全部排列在芯片的一个侧面，增加金属焊盘的数量；另外，用于形成凸点的焊 盘可以布置在芯片顶部的任何位置，用于供电的焊盘可以布置在靠近需要供电的 区域，进一步提升电气性能。

FC 封装采用回流焊或热压工艺形成芯片上的凸点与基板上的焊盘连接。倒装封 装首先将焊球附着在芯片焊盘上，再将芯片正面朝下放置在基板上，焊球通过回 流焊（Reflow）或热压缩（Thermo Compression）工艺实现和基板的连接。回 流焊工艺是倒装封装中最主流的工艺，又称为 C4 工艺（Controlled

Collapse Chip Connection），通过在回流炉的高温下融化结合处的凸点；热压缩工艺通过向接 合处施加热量和压力，一般用于 I/O 密度更高的铜柱凸点。热压工艺的成本高、 装片工艺也要求更高的精度，导致目前耗时较长。

四、HBM 多层堆叠结构提升前后道工艺价值量，量 测、电镀、键合等设备需求倍增

HBM 工艺流程包括晶圆测试、中段制造、后段封测等环节，一方面带动部分传 统封装工艺环节的设备价值量提升或带来传统封装工艺变革，另一方面，HBM 引入凸块、RDL、TSV 等先进封装工艺，使得越来越多的检/量测、电镀等前道 设备融入中后段工艺中。

1）前道检/量测：HBM 中全部的 DRAM die 均需要 100%检测，并且由于 HBM 中的铜凸点、铜柱、TSV 、UBM（凸点下金属层）的结构极其复杂，其质量影响 整个 HBM 及 CoWoS 封装的良率，因此需要对每个关键结构的大小、情况、精 确程度等进行量测，此环节使用的设备包括前道检测和量测设备； 2）凸块（Bump）制造：介于前道晶圆制造和后道封装工艺之间，是扇入（Fan-in） 封装、芯片级封装（CSP）、系统级封装（SiP）、芯粒封装（Chiplet）等工艺的 必要环节，TSV、晶圆级封装（WLP）等均是凸块制造工艺的延伸。在 HBM 中， 微凸块（uBump）用于 HBM 之间 DRAM Die、HBM 和硅中介层之间的连接， C4 Cu 凸点用于硅中介层和封装基板之间的连接，实现电气互连和应力缓冲。如 前文所述，Bumping 工艺主要使用 PVD 设备、涂胶显影机、电镀机、刻蚀机、 植球机、回流炉等设备； 3）TSV 制造：如前文所述，TSV 工艺主要包括前段的通孔成型工艺和中段的通 孔露出和背面金属化工艺，使用的设备主要包括刻蚀机、PECVD、PVD、电镀 机、减薄机、CMP、键合机等；4）硅中介层（Interposer）制造：主要包括形成 TSV、双面 RDL 和凸块制造、 临时键合和解键合、晶圆减薄、切片等工序。RDL（Redistribution Layer，重新 分配层）是额外金属布线层，通过重新排列 I/O 焊盘，满足封装工艺更复杂的功 能需求，例如将 HBM 堆栈与 SoC 等逻辑芯片相连。凸块工艺可以看做是点与点 的连接，RDL 工艺则是面阵连接，RDL 制备使用的设备与凸块类似，主要包括 光刻机、刻蚀机、溅射、CVD、电镀设备等，完整的硅中介层制造还主要需要临 时键合/解键合机、背面减薄机等；

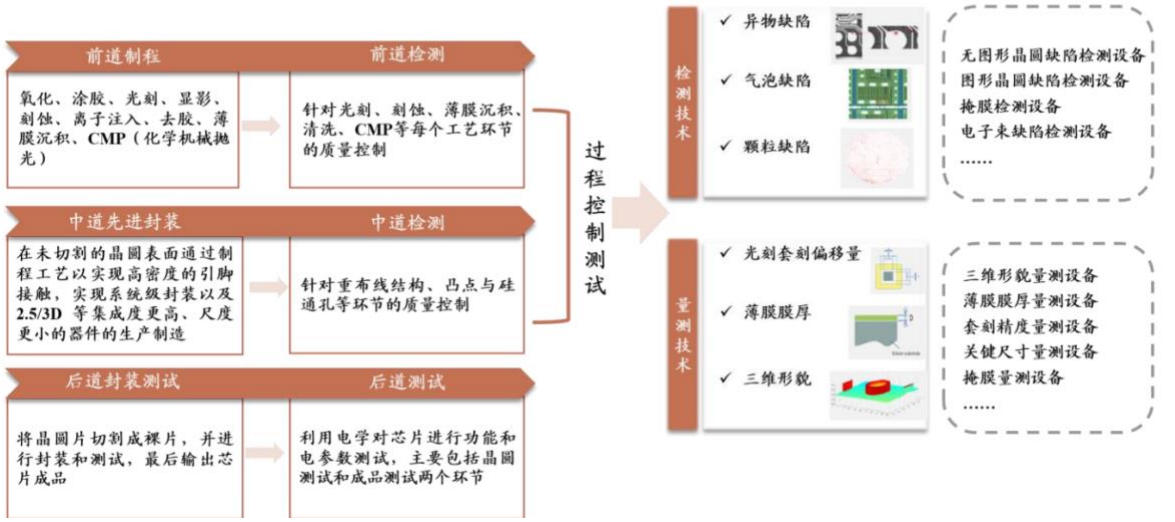
5）后道封装：传统封装工艺流程包括晶圆背面减薄、划片/切割、贴片/固晶、互 连、模塑、激光打标、切筋成型等步骤，HBM 的多层堆叠结构显著提升晶背减 薄、键合等设备的需求，并且对注塑机提出更高要求； 6）后道测试：在晶圆制造之后、封装开始之前，需要首先使用探针台、测试机 等将不良芯片剔除，探针台接触裸芯片上的 PAD 点，并连接测试机完成测试过 程，再将不良芯片进行打点标记，形成晶圆 Map 图；在封装完成后，需要分选 机配合测试机进行成品芯片的终测，该环节主要使用的设备包括探针台、分选机、 测试机； 7）板卡级组装和整机组装：为 HBM 封装的最后一道工序，主要将引脚通过 PCB 上的导线和其他器件建立连接，工艺涉及通孔插装技术、表面贴装技术等，为了 实现 PCB 组装工艺，使用的设备主要包括焊膏涂覆设备、丝网印刷机、点胶机、 贴片机、回流炉、清洗机、自动光学检测设备等；为了提供 PCB 基板，使用的 设备主要包括真空层压机、钻孔机、通孔电镀设备、涂胶机、光刻机、显影机、刻蚀机、丝网印刷机、电镀铜设备、自动光学检测仪等。

1、测试：用于 Bump、TSV、KGSD 等环节，HBM 工艺步 骤增多带来前、后道测试需求提升

（1）前道检/量测：核心增量来自 Bumping、TSV、RDL 工艺

检测（Inspection）和量测（Metrology）两大工艺分别用于检测晶圆异质情况 和对晶圆结构做出量化描述。检测指在晶圆表面上或电路结构中，检测其是否出 现异质情况，如颗粒污染、表面划伤、开短路等对芯片工艺性能具有不良影响的 特征性结构缺陷；量测指对被观测的晶圆电路上的结构尺寸和材料特性做出的量 化描述，如薄膜厚度、关键尺寸、刻蚀深度、表面形貌等物理性参数的量测。根 据检测类型不同，半导体质量控制设备也相应分为检测设备和量测设备两大类。

图 48：半导体质量控制环节根据集成电路制造工艺的划分



2022 年全球检测和量测设备市场空间大约 125 亿美元。根据 VLSI Research 数 据，2016 年全球检测和量测设备合计规模为 47.6 亿美金，2020 年增长至 76.5 亿美金，CAGR 达 12.6%。根据 Gartner，2021 年和 2022 年全球工艺控制设备 市场占比大约为 11%和 12%，结合全球前道设备市场空间，2021 和 2022 年全 球半导体检测和量测设备市场规模预计分别为 100 亿美元和 125 亿美元。

HBM 对前道检/量测设备的增量主要来自 Bumping、TSV、RDL 工艺。 Bumping 工艺中，检/量测设备需要识别凸点的缺失、大小不一致等问题。 在 HBM 中，铜凸点（Bump）、凸点下金属（UBM）、RDL、TSV 等的结构 极其复杂，其形状、质量、精确程度直接影响 HBM 甚至整体 CoWoS 封装 的良率。例如，Bump 用于各 DRAM die、HBM 和硅中介层、GPU 和硅中 介层的连接，C4 凸点用于硅中介层和封装基板之间的连接，均需要检/量测 设备来检测 Bump 是否缺失、大小是否不一致等问题。

针对 RDL 工艺，更精细的电路增加检/量测设备需求。HBM 中硅中介层上 排列多层 RDL 布线，支持极其精细的线路，可以互连芯片上间距为 55um 甚至 40um 的微凸点。随着需要互联的凸点增多，RDL 互连的凸点间距进 一步缩小，增加更加精细和先进的检/量测设备需求。

（2）后道测试：针对堆叠晶圆增加预键合测试，针对 KGSD 增加多 种封装测试流程

根据 SEMI，2021 年全球半导体后道测试设备市场规模大约 78 亿美元，国内市 场大约 22.5 亿美元，其中测试机、分选机、探针台分别占比 63%、17.4%、15.2%。 目前，测试设备主要被泰瑞达、爱德万等海外厂商垄断，探针台/分选机等主要 被东京精密、东京电子等垄断，国产化提升空间较大。

测试机：主要实现晶圆的功能测试，通过一台工作站来做外部的控制，透过测试程序的定义来产生待测器件上所需要的电压、电流及时序信号等的输入 和读取，依次来判断待测器件的好坏； 探针台：主要用于半导体制造晶圆检测（CP，Circuit Probing）环节，也 用于设计验证和成品测试环节。在晶圆检测环节中，探针测试主要对封装前 晶圆上的裸芯片进行性能测试，裸芯片上有多个 PAD 点用于信号传输，测试时探针与之接触并 输入、输出测试信号。裸芯片经 PAD 引出引脚、覆盖封装材料等工序后即 成为成品芯片，稳定性与尺寸均大幅提升； 分选机主要用于 IC 设计阶段的验证环节和封测阶段的成品测试环节，一般 和探针台配合使用。分选机将待检测的芯片自动传送至测试工位（用于将芯 片与测试机连接并进行测试的位

置），待检测芯片的引脚通过测试工位上的 专用连接线与测试机的功能模块进行连接，测试机在进行检测之后将测试结 果传送给分选机，分选机根据测试结果将检测过的芯片进行标记、分类、收 料。

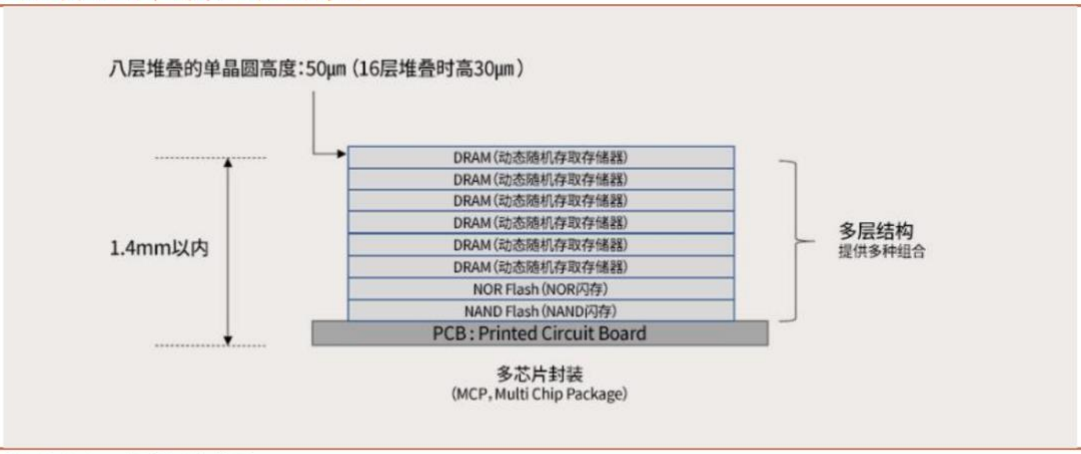
2、减薄：全球减薄设备市场空间超 8 亿美元，HBM 多层堆 叠结构提升设备需求

用来提高晶圆强度的衬底材料 90%以上均需要被去除，背面**减薄工**艺可将晶圆 **减薄至** 250um-20um。伴随硅片和晶圆尺寸的增大，其厚度也需要相应增厚来 保证硅片或晶圆在制造过程中具有足够的强度，例如 6 英寸和 8 英寸的晶圆的厚 度分别为 625um 和 725um，12 英寸硅片的平均厚度为 775um。对于 725um 厚度的 12 英寸晶圆，其电路层的有效厚度一般为 5-10um，总厚度 90%以上的衬 底材料是为了保证晶圆在制造、测试和运送过程中具备足够的强度。随着芯片集 成化程度提高，晶圆厚度一般越薄越好，因此，在晶圆制造完成之后，需要首先 通过晶圆背面减薄（Back Grinding）工艺来降低厚度，再进行划片、贴片等后 续封装工艺。对于单芯片封装来说，经过背面减薄的晶圆厚度一般会从 700-800um **减少到** 200-250um ；对于多层堆叠芯片封装来说，晶圆一般需减薄 至 70-80um，通过两次研磨工艺甚至可以减薄至 20um。2022 年全球晶圆减薄市场约 8.2 亿美元，主要为 DISCO、东京精密等海外厂商 垄断。根据 QYResearch ，2022 年全球**减薄机市**场规模约 8.2 亿美元，2018-2022 年CAGR约为18.7%，预计到2029年将增长至接近13.2亿美元，未来6年CAGR 大约 6.5%。全球主要**减薄机厂商包括** Disco、东京精密（TOKYO SEIMITSU）、G&N、Okamoto Semiconductor Equipment Division 等，CR3 占有 85%的**份**额。

先进封装、Chiplet 等技术需要厚度更低的超薄晶圆，将大幅提升对减薄设备的 需求。芯片不断追求更高的集成度和更小的体积，3D IC 等工艺得以发展，通过 硅穿孔（TSV）等技术实现 IC 堆**叠**，**可以有效减小** IC 之间互连的长度，将芯片 整合成效能最佳、体积最小的状态，目前大部分的 3D NAND、背照型 CMOS 图 像传感器、智能手机 SoC 等先进芯片均使用 2.5 或 3D IC 技术。传统的减薄工 艺一般只需要将晶圆减薄至 100-200um，但在 3D 封装中，需要将多层芯片进行 堆**叠**，**往往需要将晶**圆厚度减薄至 50-100um 甚至 50um 以下。根据 Yole，2019 年全球**减薄后晶**圆出货量为 1 亿片，预计到 2025 年上升至 1.35 亿片，其中 100-200um 晶圆仍占据主要份额，约为 8200 万片；30-50um 晶圆增速最快， 2019-2025 年 CAGR 为 98%，在 2025 年预计出货量约 170 万片。

减薄抛光是 TSV 工艺中介于电镀和键合之间的重要工序，帮助实现 HBM 多层 DRAM die 堆**叠**。**晶**圆厂一般负责 TSV 的成型工艺，主要包括刻蚀和沉积，封 装厂一般负责中段制程（MEOL），即硅通孔露出和背面金属化工艺，主要包括 **减薄**、钝化和键合等，以及最后的封装工艺。在沉积完成后，需要电镀方法形成铜，一般采用自下而上的方法（Bottom-up）；在电镀完成后，需要对晶圆进行减 薄抛光来使 TSV 背面的铜层快速露出，一般采用机械研磨等方法将晶圆减薄至 50μm 甚至更薄，并采用 CMP 等技术露出铜通孔中的铜柱。减薄后的晶圆将大 大降低 TSV 阻抗，还会增加数据带宽、降低热阻，最终增加互连密度。在 HBM结构中，减薄和键合工艺配合使用，DRAM die 之间不再需要导电凸块，芯片的 厚度将薄数倍，整体堆**叠高度得以降低**。经过背面减薄的晶圆厚度一般从 700-800um 降至 70-80um，**减薄到十分之一厚度的晶**圆能够堆叠 4-6 层，经过 **两次减薄工**艺，晶圆可以减薄至大约 20um，从而实现未来 HBM 堆**叠** 16-32 层。

图 57：多芯片封装结构中减薄后的芯片厚度



资料来源：SK 海力士，招商证券整理

3、互连：超薄晶圆增加临时键合/解键合需求，混合键合市 场有望指数级增长

键合（Bonding）属于后道封装过程，按键合的材料类型分为晶圆-晶圆键合 （Wafer-to-Wafer，W2W）和芯片-晶圆键合（Die-to-Wafer，D2W）。键合主 要指将**两片表面清**洁、原子级平整的同质或异质半导体材料经表面清洗和活化处 理，在一定条件下直接结合，通过范德华力、分子力甚至原子力使两片半导体材 料成为一体的技术。W2W 是指通过化学或物理反应将晶圆与晶圆、晶圆与玻璃 基板或其他材料圆片永久结合起来的工艺。键合过程为，在外能量的作用下，两 个晶圆接合界面上的原子相互反应形成共价键，从而使晶圆接合并达到一定的界 面键合强度；D2W 指在**划片工**艺之后，将从晶圆上切割的芯片黏贴在封装基板（引线框架或印刷电路板）上。芯片键合的传统方法为引线键合（Wire Bonding）， 先进方法采用倒装键合（Flip Chip Bonding）、混合键合（Hybrid Bonding）等。

按键合的方式，键合分为临时键合和永久键合，键合精度、材料选择等是工艺成 功的**关**键。晶圆键合工艺是半导体器件物理、材料物理化学、精密机械设计、高 精度自动控制等多学科交叉的领域，对对位精度、键合温度均匀性、键合压力范 围及控制精度等要求较高，另外用于固定薄晶圆的键合胶也是工艺成功的关键。 1）临时键合：超薄晶圆需要临时键合至较厚的载体上，方便后续进行减薄等一 些列工艺。柔性、易碎、翘曲是减薄后晶圆的特点，因此需要通过临时键合工艺， 用中间材料将薄晶圆键合到较厚的载体片上，再经过背面减薄、TSV **开孔、重布** 互联等工艺后，再输入外界能量（光、电、热、力等）使粘层失效，无损地将晶 圆与载体片分离； 2）永久键合：指通过化学或物理反应将晶圆与晶圆、晶圆与玻璃基板或其他材 料圆片永久结合起来的工艺。在外能量的作用下，两个晶圆接合界面上的原子相 互反应形成共价键，从而使晶圆接合并达到一定的界面键合强度。在 TSV 封装 工艺中，为了形成良好的电学互联，精度要达到连接柱直径的 10%，对连接精度 要求很高，TSV 连接柱的直径一般为 5um 以下，那么连接精度就要求在 0.5um 以下，这时候采用晶圆对晶圆的对准方式就能实现很好的对准精度。 混合键合技术能够实现芯片模块间快速通信，成为最先进的新一代键合技术。在 混合键合之前，2D、2.5D、3D 封装均使用焊球凸点或微凸点来实现芯片与基板、 芯片与中介层之间的连接，而由于凸点间距逐渐缩小，芯片模块间的通信速度不 断提升，原来的锡焊球无法满足工艺要求；铜-铜混合键合（Hybrid Bonding）通 过精密间隔的铜焊盘垂直连接 D2W 或 W2W，信号丢失率可忽略不计，在高数 据量、高性能计算领域优势明显。

HBM 的多层堆叠结构带动临时键合、混合键合等需求。在传统工艺中，每一代 SoC 上集成更多的功能，但芯片缩放越来越困难，而使用 Chiplet 技术，大型 SoC 被分割成更小的 Dies 或 IP 块，并重新聚合成一个全新的设计，对临时键合、混 合键合等需求均有拉动。 1）临时键合需求：在 Chiplet 技术中，为了缩小芯片体积、提高芯片散热性能和 传导效率等，晶圆减薄工艺会被大量应用，为了不损伤减薄中以及减薄后

晶圆， 需要将晶圆片与玻璃基板临时键合并在完成后续工艺后最终解键合。同时在 Chiplet 技术路线下，Fan-out、CoWoS 等封装工艺路线都要经过单次或多次的 临时键合及解键合工艺来实现芯粒互联； 2）混合键合需求：随着台积电、英特尔、三星等厂商开始采用 3D 封装工艺， Chiplet 应用范围扩大，混合键合越来越多用于 CPU、GPU 及高性能计算中。最 早实现 Hybrid Bonding 量产的产品是索尼的图像传感器，下部电路芯片和上部 像素芯片之间增加了铜连接焊盘，同时建立物理和电气连接；又如在 AMD 的一 个方案中，AMD 堆叠了 SRAM 和一个处理器芯片，形成了 3D 封装形式，并在 顶部结合了高性能的 MPU 和高速缓存，使用混合键合连接各个 Die。

半导体晶圆键合设备全球市场空间大约 9 亿美元，混合键合设备细分市场有望迎 来指数级增长。根据 thebrainyinsights 数据，2021 年全球半导体键合设备市场 空间大约 8.74 亿美元，当前晶圆键合设备主要为海外厂商垄断，行业龙头包括 奥地利的 EVG Group、德国的 SUSS、日本 TEL 等，国内尚无多模块集成的晶 圆键合设备，技术与国外差距较大，国内主要厂商为拓荆科技、上海微电子 （SMEE），苏州芯图（芯睿科技）、华卓精科等，芯源微正在研临时键合机、解 键合机等新型设备。伴随着“后摩尔时代”来临，三维集成领域进入成长期，混 合键合设备细分市场有望迎来指数级增长。

4、模塑：用于对芯片进行塑封保护，HBM 采用注塑或压缩 式工艺进行塑封

模塑（Molding）是指一**种将芯片或器件模塑料**进行保护的封装工艺，传统封装 采用注射或传递式成型塑封法。模塑使裸露于外界的芯片、器件及连接线路通过 外部塑封体得到保护，免受外界环境对半导体器件的侵蚀，避免产品失效。对于 传统的 TSSOP、DFN、QFN 及 BGA 等封装工艺，半导体塑封主要采用注射式 成型塑封（Injection Molding）或传递成型塑封（Transfer Molding），过程为： 将引线键合连接芯片的基板放在两个模具上，同时将环氧树脂模塑料片放在中间， 然后通过注塑头施加热量和压力，使环氧树脂塑料熔化为液态，流入模具并填充 间隙。

SK 海力士的 HBM 采用 MR 方式将液态 LMC 注入 DRAM die 之间。考虑到批 量回流工艺的效率高、可靠性高，因此 SK 海力士在其 HBM3、HBM3e 堆叠中 并未采用 TC（热压）方式，而是沿用 MR（批量回流）工艺进行 HBM 间模塑材 料的填充。具体过程为，直接将液态的 LMC 材料注入 DRAM die 之间的空间， 并进行固化，从而达到芯片保护的作用。SK 海力士采用新技术来控制晶圆变薄， 使其不会弯曲；在 12 层及以上堆叠过程中，瞬时施加强烈热量，以确保连接芯 片的凸块均匀拼接；将 LMC 材料置于真空下，并施加 70 吨压力来填充芯片之间 的狭小空间。

五、HBM 提升 EMC、电镀液、PSPI 等材料需求， CXL 搭配 HBM 有望突破 CPU 内存极限

1、环氧塑封料：用于保护芯片，GMC 和 LMC 及原材料均 受益于 HBM 需求增长

EMC起到保护芯片的功能，在传统和先进封装中均广泛应用。环氧塑封料（Epoxy Molding Compound，简称 EMC）全称为环氧树脂模塑料，属于包装材料，是用 于半导体封装的一种热固性化学材料，由环氧树脂为基本树脂，以高性能酚醛树 脂为固化剂，加入硅微粉等填料，以及添加多种助剂加工而成，主要功能是保护 半导体芯片不受外界环境（水汽、温度、污染等）的影响，并实现导热、绝缘、 耐湿、耐压、支撑等复合功能。EMC 在传统引线键合封装和先进封装中均广泛 应用，FOWLP、多层堆叠封装对塑封料提出了更高的性能要求。

EMC 主要包括液态塑封料（LMC，Liquid Molding Compound）与颗粒状环 **氧塑封料**（GMC，Granular Molding Compound）。1）GMC 指采用均匀撒粉 的方式，在预热后变为液态，将带有芯片的承载板浸入到树脂中而成型，具有操 作简单、工时较短、成本较低等优势；2）LMC 指通过将液态树脂挤压到产品中央，在塑封机温度和压力的作用下增强液态树脂的流动性，从而填满整个晶圆。 LMC 具备可中低温固化、低翘曲、模塑过程无粉尘、低吸水率及高可靠性等优 点，是目前用于晶圆级封装的相对成熟的塑封材料。EMC 的填充料成本占比最高，主要无机填料为球形硅微粉和球形氧化铝。目前 常见的环氧塑封料主要组成

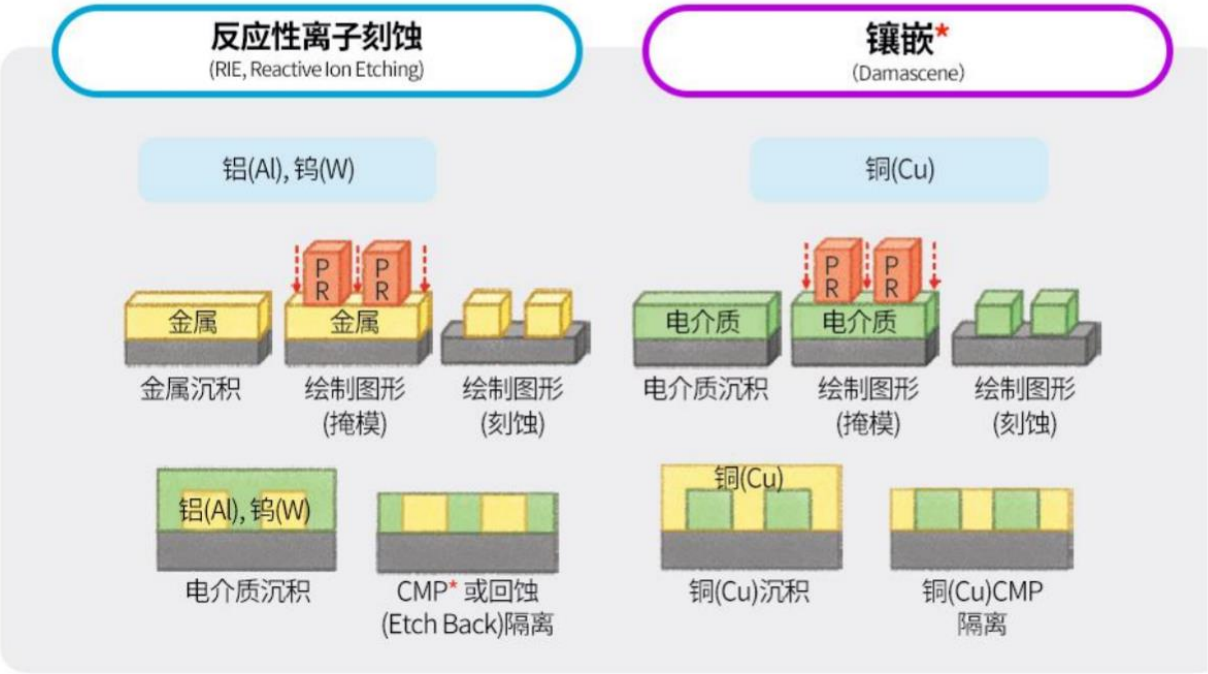
为填充料（60-90%）、环氧树脂（18%以下）、固化 剂（9%以下）、添加剂（约 3%）。在先进封装中，环氧塑封料的主要要求为高耐 潮、低应力、低α射线、耐浸焊和回流焊，需要保证塑封性能好，因此环氧塑封 料必须在无机树脂基体内掺杂无机填料，现有的无机填料基本均为二氧化硅球形 微粉，具有降低塑封料的线性膨胀系数。增加热导，降低介电常数，环保、阻燃， **减小内**应力，防止吸潮，增加塑封料强度，降低封装料成本等作用。**另外**，针对 HBM 封装等高导热存储芯片封装领域，颗粒封装材料（GMC）中一般将 TOP CUT20um 以下球形硅微粉和 Low α球形**氧化**铝复配混用，散热要求越高的场 景，Low-α球铝的占比会越高。

2021 年国内 EMC 市场规模约 66 亿元，HBM 带动 EMC 和球形硅微粉等市场增 长。根据华海诚科招股书，2021 年中国包封材料市场规模为 73.6 亿元，EMC 占比大约 90%，即市场规模大约 66.24 亿元。在传统封装领域，内资 EMC 厂商 **份**额逐渐提升，在 SOP、QFP 等领域仍存在一定替代空间；在先进封装领域， 市场份额基本被住友电木、蔼司蒂、京瓷等外资厂商占据，内资厂商多处于客户 验证阶段，少数厂商产品实现小批量产。

2、电镀液：用于形成 HBM 铜凸块、TSV 等结构，全球市 场空间约 10 亿美元

电镀液及添加剂主要用于铜互联工艺，全球市场空间接近 10 亿美元。电镀液在 IC 制造和先进封装中用于铜互联工艺，该工艺贯穿整个芯片制造过程，随着先进 封装对镀铜材料需求快速增加，大马士革铜互联、先进封装凸块电镀（Cu Pillar/Bump/RDL/UBM）、硅通孔（TSV）电镀等材料市场不断扩大，在先进芯 片中互联材料首尾连接可长达约 30 英里。根据 TECHCET，铜互联材料是电镀 材料最大的细分市场，2022 年全球半导体用电镀材料市场规模约 10.2 亿美元， 预计到 2026 年增加至 13.8 亿美元。 IC 制造工艺中金属布线用于连接电子元器件层，铜互联采用电镀方式实现铜的 填充。IC 最初采用铝作为导体，二氧化硅作为绝缘体来构造互联层，整个互联 过程从在晶圆表面沉积铝开始，然后通过选择性刻蚀形成布线图案，沉积氧化物 绝缘体，并利用 CMP 使晶圆表面平坦化；随着器件特征尺寸缩小，越来越薄的 铝线无法实现所需的速度和电性能，铜互联结构逐步取代铝互联；然而由于铜不 易形成挥发性化合物，因此使用等离子干法刻蚀铜的方法并不可行，因此工程师 选择采用大马士革铜互联镶嵌工艺（即借鉴大马士革的珠宝行业，先在基底金属 上刻蚀图案，再将贵金属嵌入图案中），先沉积和刻蚀电介质材料，再将铜填充 到图案之中。为了实现高深宽比图案特征，必须采用电镀而非 PVD 或 CVD 来填 充铜金属。

图 66：铝互联和铜互联需要的不同工艺流程



HBM 引入铜互连工艺，电镀液主要用于形成铜柱凸块、TSV 等结构。由于 HBM 中需要的 Bumping、RDL 、TSV 等先进封装工艺引入前道电镀工艺环节，自然 带来电镀液需求提升。1）在 Bumping 工艺中，电镀步骤位于涂胶显影工艺环节 之后，用于制备一定厚度的金属层作为 UBM；2）在 RDL 工艺中，每层 RDL 布 线均需要电镀来形成铜凸块；3）TSV 技术的核心是在晶圆上打孔，并在硅通孔 中进行镀铜填充，从而实现晶圆的互联和堆叠，在无需继续缩小芯片线宽的情况 下，提高芯片的集成度和性能。和芯片制造铜互连工艺相比，TSV 电镀的尺寸 更大，通常需要更长的沉积时间、更高的电镀速率以及多个工艺步骤，铜互连电 镀液及添加剂成本占 TSV 工艺的总成本比重也更高。

3、PSPI：用于 HBM 的硅中介层，形成 RDL 的再钝化层

聚酰亚胺具有最高的阻燃等级，高端光敏聚酰亚胺主要用作光敏光刻胶和先进封 装树脂。聚酰亚胺（Polymide，PI）指分子结构主链中含有酰亚胺结构的高分子 聚合物，高性能 PI 的主链大多以芳环和杂环为主要结构单元。PI 具有最高的阻 燃等级（UL-94），良好的电气绝缘性能、机械性能、化学稳定性、耐老化性能、 耐辐照性能、低节点损耗等，这些性能在很宽的温度范围（-269℃-400℃）内不 会发生显著变化。光敏聚酰亚胺（Photosensitive Polyimide，PSPI）是一类在 高分子链上兼有亚胺环和光敏基因，具备良好的感光性能，主要用于光刻胶和电 子封装领域。

4、封装基板：用于 HBM 硅中介层下方，实现内部芯片与外 部电路间的电气连接

封装基板（Package Substrate）可分为有机基板和陶瓷基板。有机基板由有机 树脂和玻璃纤维布为主要材料，一般采用铜箔作为导体，有机树脂包括：环氧树 脂（FR4）、BT 树脂（耐高温双马来酰亚胺三嗪）、PPE 树脂（聚苯醚树脂）、 PI 树脂（聚酰亚胺树脂）等；陶瓷基板具备更好的机械和热性能，通常包括：HTCC、LTCC、氮化铝等。英特尔主导研发了一种以 ABF（Ajinomoto Build-up film）作为原材料的基板

，相较于 BT 基板，ABF 材质可用于线路较细、高讯息 传输的 IC，比如 CPU、GPU 等芯片。ABF 作为基板的优势在于，铜箔基板上面 附着 ABF 增厚薄膜就可以直接电镀铜来形成金属布线，不需要热压过程。在 HBM 中，硅中介层一般采用 FC-BGA（球栅阵列封装）和下方封装基板相连， 采用锡球来代替引线框架，锡球附着于基板底部，引线与基板顶部连接，基板中 心位置由名为“芯板（Core）”的材料**构成**，**通**过将铜箔与浸渍过有机树脂的玻 璃纤维粘合在一起，金属引线在铜箔表面形成，之后在铜箔上涂覆阻焊剂，露出 作为保护层的金属焊盘/引脚。

5、CXL：能够突破 CPU 内存极限，未来将越来越多搭配 HBM 使用

HBM 目前直接搭配 CPU 的应用较少，主要系访问延迟等限制。由于目前 HBM 访问延迟较高，在访问某个特定位置的数据后，需要若干周期的时间以后才能抵 **达**该位置并执行控制器发出的指令，同时由于 HBM 和主芯片在同一系统中进行 SiP 封装，**灵活性欠佳**，HBM 显存容量基本不可以后续进行扩展。因此，当前 HBM 在 PC 等 CPU 中应用较少，2022 年 11 月，英特尔发布全球首款配备 HBM 内存的 x86 CPU“至强 Max”Sapphire Rapids，搭载 4 个 HBM2e Stack 供 64GB， **每个** HBM2e Stack 容量为 16GB。 CXL 使 CPU 内存空间和连接设备上的内存之间保持一致性，进而突破 CPU 内 存**极限**。AI 等应用促进处理的数据指数级增长，服务器越来越多转向异构计算架 **构**，**即使用**专门的 GPU 等加速器从 CPU 卸载专门的工作负载。CXL（Compute Express）是一**种新的开放式互**联标准，是基于 PCIe 物理层的高速、低延迟 CPU 设备互联技术，可以在主机 CPU 和互联设备（例如加速器和存储器扩展设备） 之间提供高效连接，能够简化加速器和内存扩展的互联和可扩展性。CXL 本质特 点是①**确保直接挂**载的 CPU 存储器与 CXL 设备上的存储器保持一致，使主机和 CXL 设备之间可以无缝读取相同的数据；②允许通过 CXL 接口将各**种**类型（例 如易失性、持久性等）的存储器挂载到主机，这与底层存储器技术无关；③支持 交换和内存池，交换功能可实现存储器扩展，池化功能允许动态分配和释放存储 器资源，从而提高整体系统效率。

CXL 可以作为 HBM 的补充，未来 CXL 将越来越多搭配 HBM 使用。根据 CXL 联盟，目前确定了 CXL 互联的三类主要设备：①智能 NIC 等加速器通常缺少本 地内存，CXL 可以实现 NIC 等设备和主机 CPU 的 DDR 内存进行通信；②GPU、 ASIC 和 FPGA 等都配备了 DDR 或 HBM 内存，可以使用 CXL 使主机 CPU 的内 存在本地供加速器使用，加速器的内存也可在本地供 CPU 使用，并且加速器和 CPU 的内存位于同一个缓存的一致域中，有助于提升异构工作负载；③连接内 存缓冲区（buffer），为主机 CPU 提供额外的带宽和容量。

六、海外厂商较早布局 HBM，三大存储原厂占据 HBM 主要市场份额

HBM 市场份额集中于三大存储原厂，国内厂商尚不具备 HBM DRAM Die 生产 和堆**叠能力**。HBM 的 DRAM Die 由存储原厂负责生产和堆叠，目前份额集中于 三大原厂 SK 海力士、三星、美光，根据集邦咨询，2022 年三大原厂 HBM 市占 率分别为 SK 海力士 50%、三星约 40%、美光约 10%，其中 SK 海力士是 HBM3 产品的领先生产商，是英伟达 AI 服务器 GPU 的主要供应商，三星主要满足其他 云端服务商的订单。

1、SK 海力士：开发 MR-MUF 技术，HBM 市场份额领先

HBM 市场先行者和领军者，全面推进第五代 HBM3E。2014 年，SK 海力士和 AMD 合作**开**发第一代硅通孔 HBM 产品，还联合开发了高带宽三维堆叠存储器技 术和相关产品。HBM1 带宽高于 DDR4 和 GDDR5 产品，同时以较小的外形尺寸 消耗较低的功率，更能满足 GPU 等带宽需求较高的处理器；SK 海力士凭借 HBM1 占据市场领先地位，并在 2018 年发布第二代产品——HBM2，其中一项 **关**键的改进是伪通道模式（Pseudo Channel Model），将一个通道分为两个单独 的 64bit I/O 子通道，为每个存储器的读写访问提供 128 位预取，从而优化内存 访问并降低延迟，从而有效提供带宽；2020 年，SK 海力士发布第三代产品——HBM2E，作为 HBM2 的扩展版本，HBM2E 堆**叠** 8 个 16Gb 芯片，总容量是 HBM2 的**两倍**，处理速度高达 3.6Gbps；2021 年 10 月，SK 海力士成功**开**发出第四代 产品——HBM3，并于 2022 年 6 月**开始生**

产；2023 年 4 月，公司宣布，在全球 率先研发出 12 层堆叠的 HBM3 内存，单颗容量最高可达 24GB；2023 年 8 月， 公司成功开发出面向 AI 超高性能 DRAM 新品 HBM3E，预计 24H1 量产；公司 预计 2026 年量产 HBM4。

SK 海力士开发 MR-MUF 等堆叠技术，巩固 HBM 市场领先地位。SK 海力士的 CoC（芯片内建芯片）技术将凸块互联（Bump Interconnection）与引线键合（Wire Bonding）相结合，在提高运行速度和降低成本方面实现突破，目前专门用于 SK 海力士的高密度模块的量产与生产；SK 海力士还开发了 MR-MUF 技术，用于其 HBM2E、HBM3 等产品中，确保了 HBM 10 万多个微凸块互连，另外该技术还 提供更出色的散热性能；SK 海力士还正在开发 Cu-to-Cu（Copper-to-Copper， 铜-铜）键合工艺替代焊接，Cu-to-Cu 可在完全不使用凸块的情况下将间距缩小 至 10 微米及以下；另外，SK 海力士还在研究用于 Fan-out RDL（扇出型重新分 配层）技术等，计划将该技术用于以 Chiplet 为基础的集成封装，计划到 2025 年实现 1um 及以下水平的 RDL 技术。

图 72：SK 海力士最新封装技术

MR-MUF 技术工艺效率高并且散热性能更好，难点在于热翘曲和芯片偏移问题。 对比 TC-NCF 技术，MR-MUF 技术的优点为：1）提高工艺效率：NCF 在每堆 叠一个芯片时铺上一层薄膜型材料，MR-MUF 在注入 LMC 材料之后，批量加热 每个芯片以实现瞬时互连；2）散热性能更好，提高凸块互连的质量：MR-MUF 由于采用了高导热率的模制底部填充（MUF）材料，散热效果较 NCF 技术更好, 提高了 HBM 超过 10 万个微凸块互连的质量。MR-MUF 工艺难点在于热翘曲和 芯片偏移问题，由于模塑是一个热反应过程，芯片堆叠过程中需要对每个芯片施 加瞬时高热，进而熔化凸点实现连接，但由于环氧树脂和晶圆之间的热膨胀系数 （CTE）相互不匹配，回流（MR）工艺会导致热翘曲、非接触性断开、局部桥 接等问题，进而导致焊接凸点失效；另外，料盘/载具在传输过程中容易受振动 影响，回流过程中芯片也会自由偏移，因此芯片的位置也容易发生偏移。

2、三星：开创 I-Cube 及 X-cube 先进封装工艺，2024 年 HBM 产能将提高 2.5 倍

24H1 HBM3 产量将占 HBM 总产量一半以上，计划将 2024 年 HBM 产能提高 2.5 倍。2020 年，三星推出 HBM2；2021 年 2 月，三星推出 HBM-PIM（存算 一体）芯片；2022 年，三星开始量产 HBM3，在 23Q3 量产并开始供应 8 层和 12 层产品，计划在 23Q4 进一步扩大生产规模，预计到 24H1，HBM3 将占公司 HBM 总产量的一半以上；公司计划 24H1 推出 HBM3E，24GB 8 层样品已经推 出，计划 24H1 量产，

36GB 12 层产品计划于 24Q1 提供样品。为了保持行业最 高的 HBM 产能，公司计划将 2024 年 HBM 产能提高 2.5 倍。 三星自研 I-Cube 技术和 X-Cube 技术，分别用于 2.5D 和 3D 封装。

I-Cube 2.5D 封装技术通过并行水平芯片放置，处理热量积存并扩展性能。 三星 I-Cube 技术为 2.5D 封装工艺的一种，包括 I-Cube S 和 I-Cube E， I-Cube S 将逻辑芯片与一组 HBM 裸片水平放置在硅中介层上，通过调整材 料和厚度来控制硅中介层翘曲和热膨胀问题；I-Cube E 采用硅嵌入式结构， 采用大尺寸、无 TSV 结构的 RDL 中介层。目前三星已开发出 I-Cube 8（搭 载 8 组 HBM）技术，计划 2024 年量产；

X-Cube 3D 封装技术包括微凸块和铜混合键合两种方案。X-Cube 技术中， 上下层逻辑 die 通过微凸块（X-Cube TCB 方案）或铜混合键合（HCB 方案） 连接，在铜混合键合方案中，堆叠精度进一步降低，三星正在开发低于 4um 等更精细的 HCB 技术。

三星 HBM4 路线图计划变革 HBM 封装方式。在 HBM4 技术路线图中，三星计 **划使用** FinFET 节点生产逻辑 die，封装方式计划将从基于凸点的 CoW（chip on wafer）变为基于 Pad 连接的 Bumpless 形式，并且三星正**开**发针对高温热特性 优化的非导电粘合膜（NCF）组装与混合键合（HCB）等技术。

3、美光：计划通过 HBM3E 实现弯道超车，正开发 HBM Next 产品

HBM3E 产品正在英伟达验证，预计 2024 财年产生 7 亿美元收入。美光于 2020 年**开始提供** HBM2 产品，用于高性能显卡、服务器处理器等领域；美光跳过 HBM3 直接提供 HBM3E 样品，在 8 层 die 堆**叠的情况下**，单颗 HBM 容量提升至 24GB， 美光 HBM3E 采用 1βnm DRAM 和 TSV 技术，正在获得英伟达的产品资格认证， 预计 2024 年初量产，2024 财年产生 7 亿美元收入。 美光预计于 2026 年推出 HBM Next 产品。美光正在开发 HBM Next 产品，预计 2026 年推出，内存容量预计提升至 36-64GB，带宽为 1.5-2TB/s。

（本文仅供参考，不代表我们的任何投资建议。如需使用相关信息，请参阅报告原文。）

精选报告来源：【未来智库】。「链接」

https://www.toutiao.com/article/7304491922679693833/?wid=1702367155029