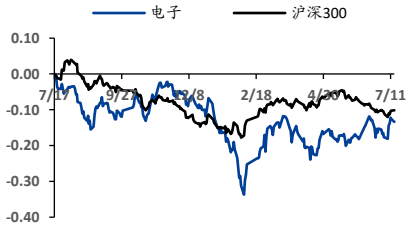
电子

逐鹿顶尖工艺，HBM4 的三国时代——HBM 专题研究二

强于大市（维持评级）

一年内行业相对大盘走势



团队成员

分析师： **陈海进(S0210524060003)**
chj30590@hfzq.com.cn
分析师： **徐巡(S0210524060004)**
xx30511@hfzq.com.cn
联系人： **谢文嘉(S0210124040078)**
xwj30510@hfzq.com.cn

相关报告

1、【华福电子杨钟】20240714 周报：折叠屏+AI 双剑合璧，三星揭开 AI 终端新篇章——2024.07.14
2、半导体 24H1 基本面持续改善，关注中报行情-半导体系列跟踪——2024.07.14
3、电子月报（台股）2024-6：关注普通服务器复苏及 H2 端侧 AI 机遇——2024.07.12

投资要点：

➢ **算力需求澎湃催化 HBM 技术快速迭代。**目前 HBM 已然成为 AI 服务器、数据中心、汽车驾驶等高性能计算领域的标配，未来其适用市场仍在不断拓宽。2024 年的 HBM 需求位元年成长率近 200%，2025 年可望再翻倍。受市场需求催化，当前 HBM 的开发周期已缩短至一年。针对 HBM4，各买方也开始启动定制化要求，未来 HBM 或不再排列在 SoC 主芯片旁边，亦有可能堆叠在 SoC 主芯片之上。垂直堆叠技术在散热，成本，分工等方面也带来了新的挑战。受先进制程技术和资金投入规模的限制，目前，只有 SK 海力士、美光和三星有能力生产兼容 H100 等高性能 AI 计算系统的 HBM 芯片。23 年海力士市场份额为 53%，三星市场份额为 38%，美光市场份额为 9%。海力士具有先发优势，但三星有望通过其一站式策略抢占市场份额。

➢ **HBM 具有三大堆叠键合工艺：MR-MUF，TC-NCF 与混合键合。**目前只有 SK 海力士使用 MR-MUF 先进封装工艺。三星与美光目前采用的为 TC-NCF 技术。各家厂商也在考虑在新一代 HBM 产品上应用铜-铜混合键合技术。在设备端，混合键合设备在单机价值量上为所有固晶机中最高，行业头部领先优势明显。预计存储领域未来贡献混合键合设备明显增量，保守预计 2026 年市场需求量超过 200 台。混合键合设备国内起步较晚，距国际领先水平仍有 5-6 年差距。

➢ **HBM4 技术路线：海力士优势明显，三星/美光发力追赶。**根据海力士最新披露的数据，公司 HBM3E 产品上的良品率以达到 80%，远远超出此前行业预期的 60%-70%，同时也大幅领先竞争对手三星与美光的良率。由于混合键合技术非常复杂，需要控制键合层的平整度和键合强度，粒子控制也需要在纳米级别进行，这将导致 HBM 在生产效率与良品率上有所欠缺。同时随着 HBM 标准限制的放宽，预计海力士仍将在 HBM4 上采用成熟的 MR-MUF 技术。三星目前的产品良率不如海力士，但三星表示 HBM 在最多 8 个堆叠时，MR-MUF 的生产效率比 TC-NCF 更高，一旦堆叠达到 12 个或以上，后者将具有更多优势，而未来 HBM4 高度的放宽势必将增加 HBM4 堆叠层数至 12-16 层，这将为 TC-NCF 工艺带来更大发挥的可能。美光研发的 HBM3E 今年和海力士一同通过了英伟达验证。美光的 HBM3E 在功耗上比竞争 HBM3E 产品低约 30%，在性能上有约 10%的提升。美光在海力士产能不足情况下，未来有望承接英伟达更多订单。

➢ **建议关注行业重点公司：**

HBM 设备：芯源微、拓荆科技、盛美上海、赛腾股份等
HBM 材料：华海诚科、联瑞新材、强力新材、飞凯材料等
HBM 封测：长电科技、通富微电、佰维存储等
海力士分销：香农芯创等

➢ **风险提示**

AI 需求不及预期风险、技术迭代不及预期、行业竞争加剧。

正文目录

图表目录

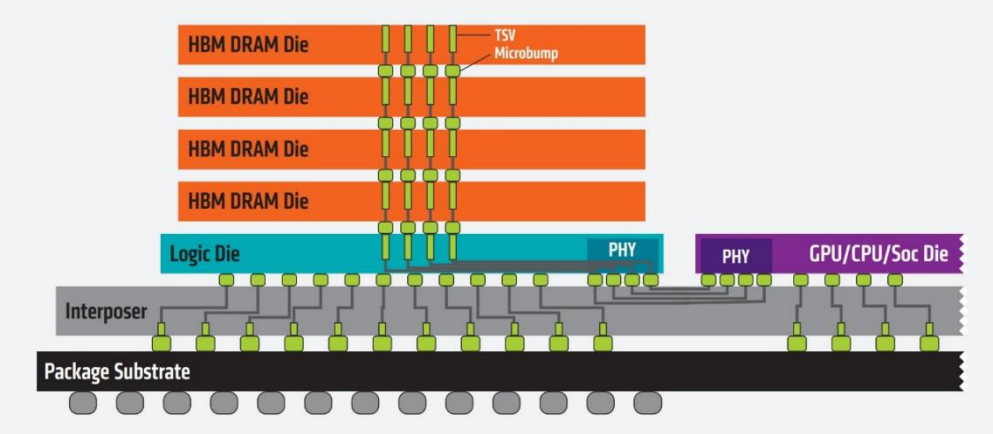
1 算力需求澎湃催化 HBM 技术快速迭代

1.1 HBM：高带宽低功耗的全新一代存储芯片

HBM（High Bandwidth Memory）即高带宽内存，作为全新一代的 CPU/GPU 内存芯片，其本质上是指基于 2.5/3D 先进封装技术，把多块 DRAM Die 堆叠起来后与 GPU 芯片封装在一起，实现大容量，高位宽的 DDR 组合阵列。

在结构上，HBM 是由多个 DRAM 堆叠而成，主要利用 TSV（硅通孔）和微凸块（Micro bump）将裸片相连接，多层 DRAM die 再与最下层的 Base die 连接，然后通过凸块（Bump）与硅中阶层（interposer）互联。同一平面内，HBM 与 GPU、CPU 或 ASIC 共同铺设在硅中阶层上，再通过 CoWoS 等 2.5D 先进封装工艺相互连接，硅中介层通过 CuBump 连接至封装基板上，最后封装基板再通过锡球与下方 PCB 基板相连。

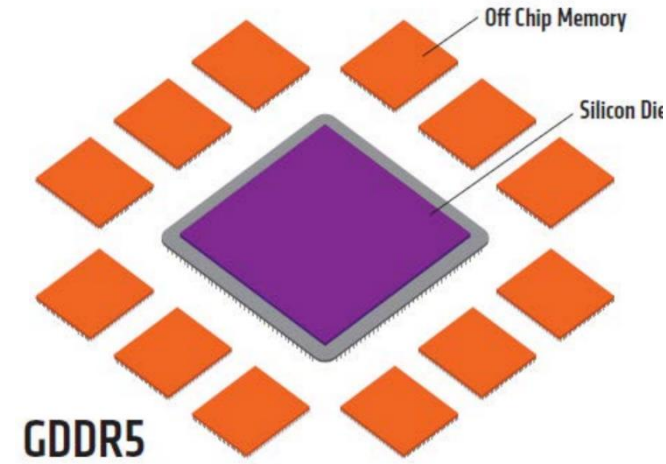
图表1：HBM 结构图



数据来源：EEPW，华福证券研究所

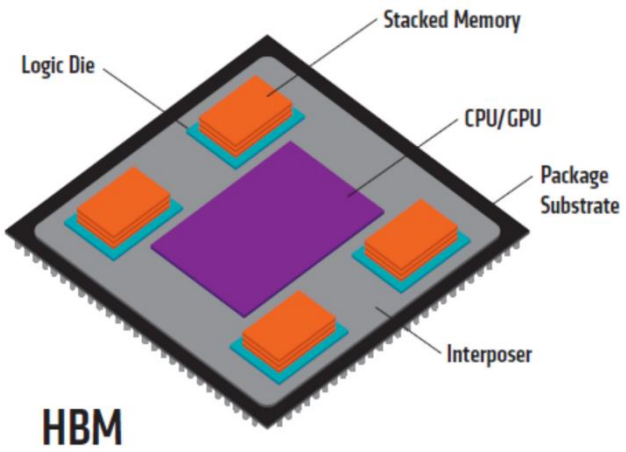
和传统的 DRAM 相比，HBM 具有高带宽、低功耗、小尺寸三大特点。1）高带宽：HBM 堆栈没有以物理方式与 CPU 或 GPU 集成，而是通过中介层紧凑而快速地连接，同时，HBM 通过堆栈结构的改变来增加引脚数量达到每颗 1024bit I/O，以实现更高带宽。**2）低功耗：**HBM 通过 TSV 技术实现走线更短，同时 I/O 数据的传输速度慢，通过重新调整内存的功耗效率，使每瓦带宽比 GDDR5 高出 3 倍。即功耗降低 3 倍。**3）小尺寸：**HBM 由于与 GPU 封装在一块，从而大幅度减少了显卡 PCB 的空间，相比于 GDDR5，HBM 单位容量表面积减少了 94%。

图表2：GDDR5 与处理器内部空间位置



数据来源：EEPW，华福证券研究所

图表3：HBM 与处理器内部空间位置



数据来源：EEPW，华福证券研究所

AI 服务器需求驱动，HBM 加速迭代。目前 HBM 已然成为 AI 服务器、数据中心、汽车驾驶等高性能计算领域的标配，未来其适用市场还在不断拓宽。目前大多数 AI 训练芯片都用到 HBM，以英伟达 H100 为例，1 每颗英伟达 H100 PICe 需要通过台积电 CoWoS-S 封装技术将 7 颗芯片（1 颗 GPU+6 颗 HBM）封在一起。而随着最新的 B200 等芯片发布，对 HBM 的需求也将逐渐增加。

市场需求催化，HBM 研发周期已缩短至一年。自 2013 年 SK 海力士推出第一代 HBM 以来，在三大原厂的竞合下，至今已历经第二代（HBM2）、第三代（HBM2E）、第四代（HBM3）、第五代（HBM3E）产品。而第六代（HBM4）也已经在研发当中。据此前数据来看，自从海力士 2014 年推出全世界第一颗 HBM 后，从 HBM2 开始大概每两年 HBM 会更新一代。但随着英伟达等主要客户的需求以及技术的发展，SK 海力士技术长表示，未来 HBM 的开发周期已缩短至大约 1 年。

1.2 方兴未艾，HBM3E 市场需求稳步增速

在 HBM3E 方面：三大存储芯片原厂美光、SK 海力士和三星在 2023 年下半年陆续向英伟达（NVIDIA）送去了 8 层垂直堆叠的 24GB HBM3E 样品以供验证。三星旗下的 12 层 HBM3E 产品在 24 年 GTC 大会上被英伟达 CEO 签下“Jensen Approved”，但随或由于发热以及功耗问题，产品未能通过英伟达效能验证。

海力士的 HBM3E 在 1024 位接口上拥有 9.2GT/s 的数据传输速率，单个 HBM3E 内存堆栈可提供 1.18TB/s 的理论峰值带宽。**三星**在 2023 年第四季度，具有 8 层堆栈的下一 HBM3E 样品已提供给客户，并计划于今年上半年开始量产。据悉，三星 HBM3E 12H DRAM 高达 1280GB/s 带宽，数据传输速度为每秒 9.8GT，领先于 SK 海力士的 9GHz 和美光的 9.2GHz。加上 36GB，较前代八层堆叠提高 50%。**美光**于今年 2 月率先宣布实现 8 层 24GB HBM3 的量产，并确认供货英伟达 H200，该产品数据传输速度为每秒 9.2GT、峰值存储带宽超越每秒 1.2TB。

HBM3E 市场需求 25 年或可翻倍。展望 2025 年，由主要 AI 解决方案供应商的角度来看，HBM 规格需求大幅转向 HBM3E，且将会有更多 12hi 的产品出现，带动单芯片搭载 HBM 的容量提升。

图表4：HBM 需求测算

	2023	2024E	2025E	2026E
台积电CoWoS产能年末月产能(k/月)	15	40	45	60
台积电CoWoS年产能（k/年）	120	330	510	630
单片CoWoS可封装GPU数（颗）	29	29	29	29
全球GPU+ASIC出货量（万颗）	348	957	1479	1827
单GPU中HBM需求颗数（颗）	6.0	6.0	6.1	6.3
HBM堆栈层数	8	9	10	11
单GPU中HBM需求量（GB）	96	108	140	165
合计HBM需求量（万颗）	2088	5742	9022	11510
合计HBM需求量（亿GB）	3.3	10.3	20.8	30.2
HBM单GB价格（美元）	15	15	15	15
HBM需求价值量（亿美元）	50	155	311	453

数据来源：电子发烧友、半导体芯情、腾讯科技、Trendforce、半导体行业观察、芯智讯、华福证券研究所测算

1.3 未来可期，三大厂构想 HBM4 蓝图

HBM4 研发进度：海力士 25 年量产，三星与美光预计 26 年量产。随着人工智能工作负载发展，内存上的创新也必须跟上步伐。三大厂在 HBM4 市场份额的争夺战上竞争激烈，**海力士**预计 25 年提供 HBM4 样品并于当年实现 12 层堆叠 DRAM 的 HBM4 量产，在 26 年实现 16 层 DRAM 的量产，比预期提前一年。同时**三星与美光**也表示将于 26 年实现 HBM4 的量产。

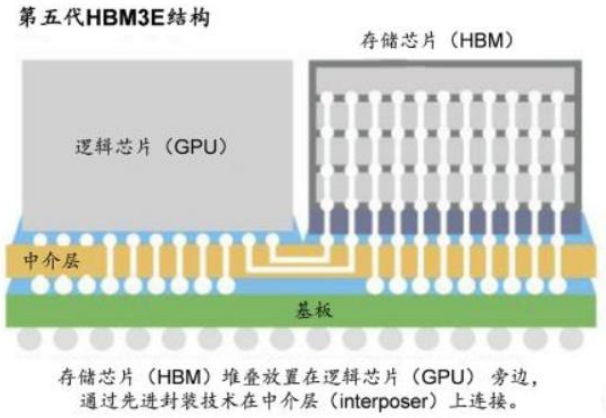
HBM4 潜在排列方式：存储芯片垂直堆叠于逻辑芯片上。根据 TrendForce 观察，针对 HBM4，各买方也开始启动定制化要求，除了 HBM 可能不再仅是排列在 SoC 主芯片旁边，亦有部分讨论转向堆叠在 SoC 主芯片之上。SK 海力士考虑将 HBM4 堆栈直接放置在 GPU 上，从而将存储芯片和逻辑半导体集成在同一芯片上。

目前，HBM 的垂直堆叠通常位于 CPU 或 GPU 的邻近中介层之上，并通过 1024 位的接口与处理器逻辑芯片相连。SK 海力士提出了一个目标，即直接将 HBM4 的存储堆叠置于处理器之上，以此来免去 HBM3E 设计中围绕逻辑芯片堆栈所带来的中介层复杂布线需求。这种方法在概念上与 AMD 的 3D V-Cache 技术相似，后者将缓存直接集成在 CPU 上。这样的技术带来的好处包括减小封装的体积、增加存储容量以及提升整体性能。

然而这种垂直堆叠技术在散热，成本，分工等方面也带来了新的挑战。1）在散热上：以 AMD 的采用 V-Cache 技术的 CPU 为例，它通过降低热设计功耗(TDP)和处理器频率来抵消由于 3D 缓存带来的额外热量。相比之下，像英伟达 H100 这样的 GPU 在数据中心中使用的 HBM 存储容量达到 80-96GB，无论是在存储容量还是发热量方面，都远远超过了 V-Cache。目前，数据中心内的计算卡可能消耗数百瓦的电力，HBM 组件本身的功耗也相当高，因此需要在 EMC（特种环氧树脂）和

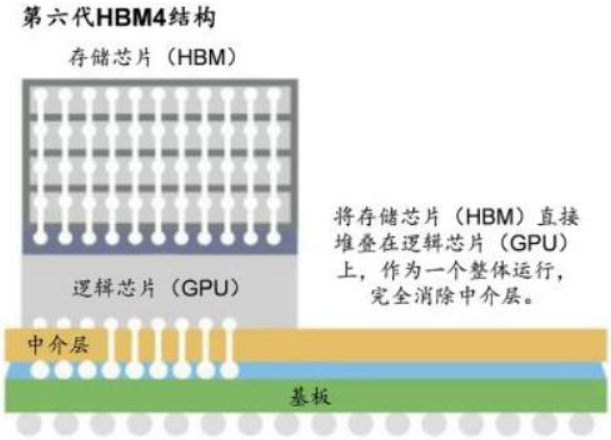
芯片间 PMIC 等方面改进现有的散热方案。**2）在分工上：**此外，采用这种集成方法还将改变芯片设计和制造流程。存储芯片和逻辑芯片需要使用相同的制造工艺，并在同一晶圆厂内生产，以确保最终产品的性能。**3）在成本上：**更高级的集成方式也将大大增加 HBM 的生产成本。

图表5：当前 HBM 与 GPU 排列方式



数据来源：Tom's Hardware，华福证券研究所

图表6：潜在的 HBM 与 GPU 排列方式



数据来源：Tom's Hardware，华福证券研究所

HBM4 性能相对 HBM3E 提升：

1）存储容量：HBM4 的容量预计将达到 36-48GB，相较于 HBM3E 的 24/36GB，这是一个显著的提升。若未来每个 GPU 搭载 HBM 数量从 6 个升级到 8 个，一个 GPU 的 HBM 搭载容量将会达到 8*36 或 8*48GB。

2）带宽：HBM4 将采用 2048 位接口或更高，比 HBM3E 的 1024 接口数量增加一倍，**同时 HBM4 预计将提供 1.5-2TB/s 的带宽**，而 HBM3E 的带宽为 1.2TB/s。为了控制功耗，**HBM4 的数据传输速率预计保持在 6GT/s 左右。**更高的带宽有助于处理更大量的数据，满足高性能计算和 AI 应用的需求。不过，2048 位接口需要更复杂的布线设计，这将导致 HBM4 的成本高于 HBM3 和 HBM3E。

3）堆叠层数：可实现 16 层 DRAM 堆叠。国际半导体标准组织（JEDEC）的主要参与者最近同意将 HBM4 产品的标准定为 **775 微米**（μm），比上一代的 720 微米更厚。这表示使用现有的键合技术就可以充分实现 **16 层 DRAM** 堆叠 HBM4。但更多的层数意味着更高的功耗和热量产生，这需要更有效的散热解决方案来保持芯片的性能和可靠性。

4）单个 GPU 搭载 HBM 数量：可搭载 8 颗。英伟达下一代 AI 芯片 R 系列 R100 芯片将搭载 HBM4 芯片，该芯片或将于 2025 年第四季度在台积电 3 纳米代工厂进入量产。据悉 R100 搭载 HBM 数量将超过此前产品的 6 颗，达到 **8 颗**。

5）制程工艺：目前海力士与美光均采用 1-β 制程工艺，领先于三星的 1-α 技术一代。同时美光预期在 HBM4 上继续采用先进的制程技术，以提升产品性能，并计划在 2025 年率先量产下一代 1-γDRAM。

6）处理能力：有望达到每颗 576GB。台积电此前宣布将把处理单元和 12 层 HBM 芯片整合到一个 AI 芯片中，将其尖端封装技术“CoWoS”（Chip-on-Wafer-on-Substrate）升级为“CoWoS-L”和“CoWoS-R”。当该技术商业化后，搭载 HBM4 的下一代 AI 半导体的数据处理能力有望达到每颗芯片 576GB。

图表7：HBM 性能参数演变

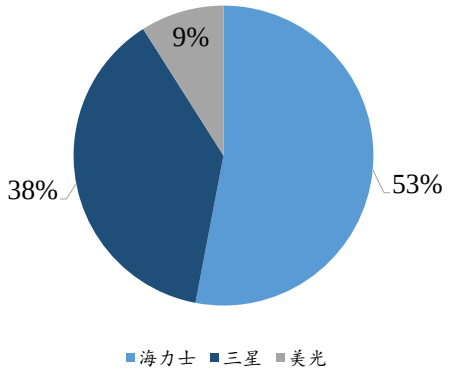
HBM各代产品参数							
产品名称	时间	芯片密度	带宽GB/s	堆叠高度	容量	I/O速率	内存接口
HBM1	2014	2Gb	128GB/s	4层	1GB	1Gbps	1024位
HBM2	2018	8Gb	307GB/g	4/8层	4/8/16GB	2.4Gbps	1024位
HBM2E	2020	8Gb/16Gb	460GB/s	4/8层	8/16Gb	3.6/3.2Gbps	1024位
HBM3	2022	16Gb	819GB/s	8/12层	16/24GB	6.4Gbps	1024位
HBM3E	2024	24Gb	1.2TB/s	12/16层	24/36GB	9.2Gbps	1024位
HBM4	2025/2026	48Gb	1.25/2TB/s	12/16层	36/64GB	未知	2048位

数据来源：三星，海力士，美光官网，华福证券研究所，

1.4 竞争激烈，三大厂各自积极布局供应链合作、

受先进制程技术的和资金投入规模的限制，目前，目前只有 SK 海力士、美光和三星有能力生产兼容 H100 等高性能 AI 计算系统的 HBM 芯片。23 年海力士市场份额为 53%，三星市场份额为 38%，美光市场份额为 9%。

图表8：三大原厂 23 年 HBM 市场份额



数据来源：闪存市场公众号，华福证券研究所

SK 海力士先发优势明显，与英伟达合作紧密。SK 海力士于 AMD 共同开发了第一代 HBM，并将其用于 AMD Fiji 系列游戏的 GPU。随后海力士于 2021 年推出了世界上首款 HBM3，并于 22 年量产后独家供应于英伟达的 H100 芯片，相对于其他两家有明显的先发优势。由于海力士和三星相比，自家没有晶圆代工厂。因此未来随着 HBM 产品在性能和功效上的多样化，海力士将进一步优化海力士 HBM 产品和台积电 CoWoS 技术融合，甚至 SK 海力士和英伟达有望从一开始就共同设计芯片，并委托台积电来生产半导体。

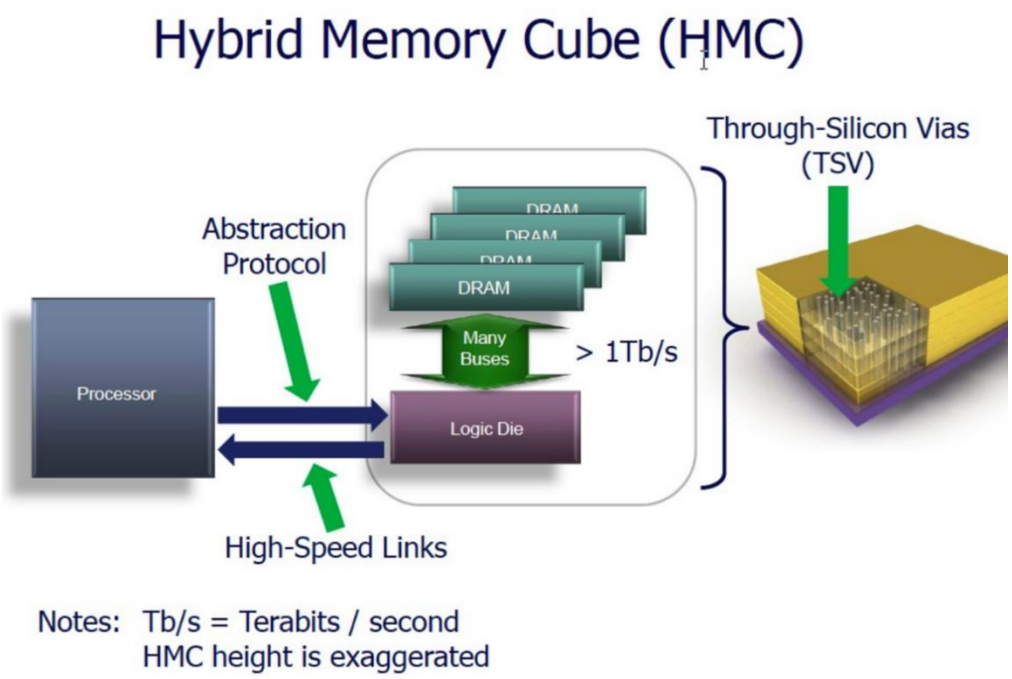
图表9：英伟达 H100 芯片

数据来源：英伟达官网，华福证券研究所

三星同时具有生产存储芯片和晶圆代工的能力，其一站式策略在争夺 HBM4 的订单上或许具有优势。三星电子日前投资 7000-10000 亿韩元用于从三星显示（Samsung Display）购买天安厂区内的部分建筑物和设备，以此来建设新的 HBM 封装线，三星存储与封装部门协同将大大缩短 HBM4 从研发到生产的中间环节，并在未来的量产中缩短从内存芯片制造、封装到交付的周期，从而能占得 HBM4 及后续产品先机。同时三星也在寻求与英伟达等全球半导体公司合作来共同为半导体设计赋能。

由于此前自研的 HMC 并未广泛应用，美光在 HBM 工艺上布局较晚，并尝试通过 1-β 制程的 HBM3E 弯道超车。由于未获市场采纳，美光于 18 年放弃 HMC 的自研转而研究 HBM，致使公司在 HBM 的研发进度上明显落后于竞争对手。目前美光绕过 HBM3 的研发，采用和海力士相同的 1-β 制程来研发 HBM3E，试图在工艺上弯道超车。在供应链方面，美光积极与包括台积电在内的中国台湾供应商合作，共同商讨 HBM 与 GPU 的整合方案。

图表10：美光自研 HMC 技术

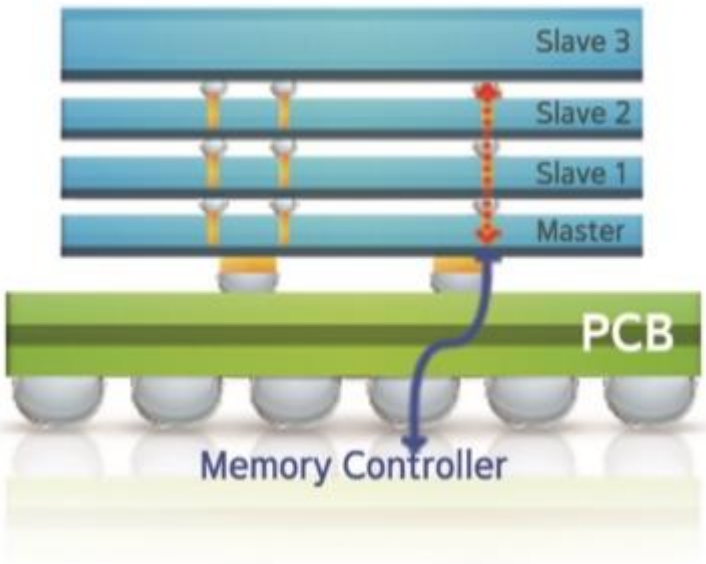


数据来源：美光官网，华福证券研究所

2 HBM 三大堆叠键合工艺：MR-MUF，TC-NCF 与混合键合

相较于传统的 DRAM，HBM 具有三大关键工艺：TSV、Micro bump 和堆叠键合。其中成本占比最高、最核心的技术便是硅通孔工艺。**硅通孔技术**（TSV，Through Silicon Via）是通过在芯片和芯片之间、晶圆和晶圆之间制作垂直导通，实现芯片之间互连的技术，是 2.5D/3D 封装的关键工艺之一。通过垂直互连减小互连长度、信号延迟，降低电容、电感，实现芯片间低功耗、高速通讯，增加带宽和实现小型化。**涉及的设备和材料**有光刻机（光刻胶）、深孔刻蚀设备（电子特气）、PVD（靶材）、CVD、电镀设备（电镀液）、抛光机（抛光液）、减薄机（减薄液）等。

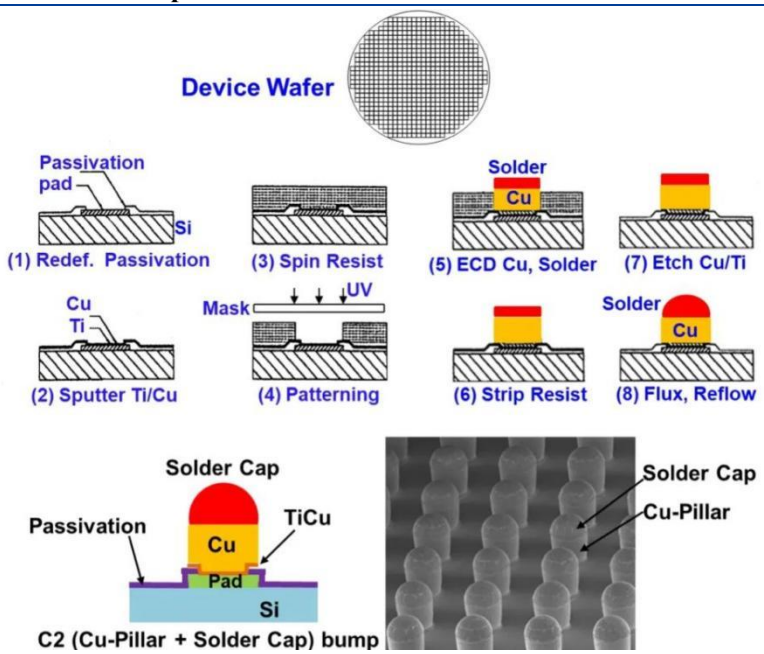
图表11：TSV 结构



数据来源：公司官网，华福证券研究所

Micro bump 是铜柱微凸点，主要制备方法是电镀。通过此项技术可以实现芯片与基板，芯片与中介层(interposer),芯片与芯片间的电连接。**涉及的设备和材料有** PVD（靶材）、涂胶显影机、光刻机（光刻胶）、电镀设备（金属、焊料）、去胶设备（剥离液）、刻蚀设备（电子特气）、回流焊设备等。

图表12：Micro bump 工艺流程图

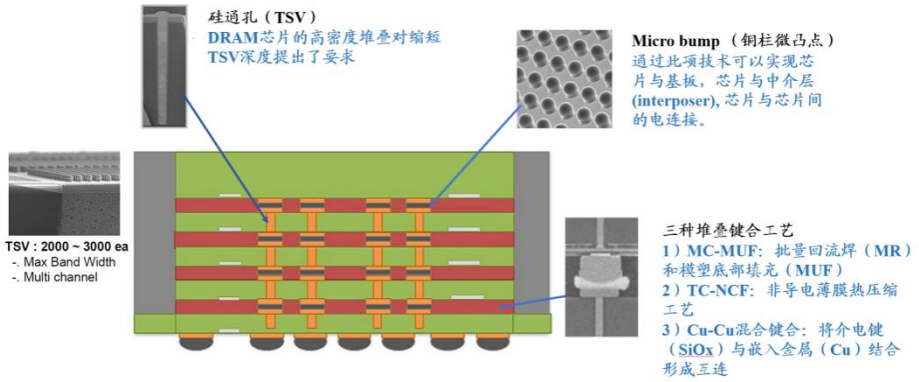


数据来源：John Lau，Unimicron，华福证券研究所

堆叠键合主要包括三种类型:MR-MUF 技术,TC-NCF 技术以及混合键合技术。

其中 MR-MUF 技术为海力士独家所有，凭借这一技术海力士得以远超于竞争对手的良品率占据市场大量份额。美光和三星目前则使用 TC-NCF 技术。而随着先进封装技术的不断开发，混合键合也成为可行的封装方案。

图表13：HBM 核心工艺

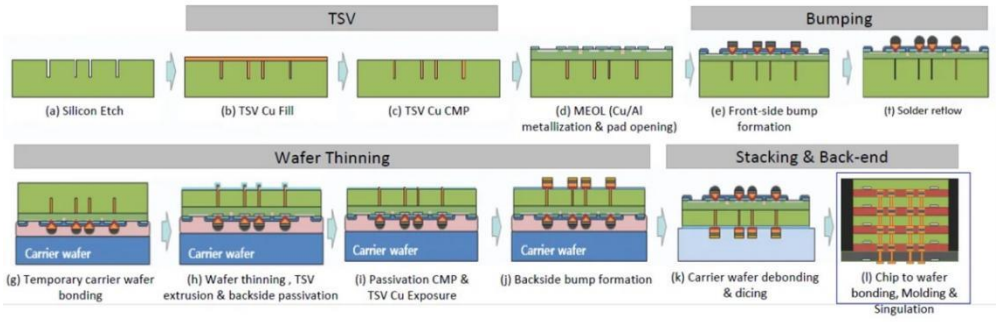


数据来源：semianalysis，华福证券研究所

2.1 MR-MUF 技术：熔化凸点+注入环氧树脂，兼具散热与生产效率

SK 海力士以其专有的批量回流模制底部填充（Mass Reflow-Molded Underfill，简称 MR-MUF）先进封装工艺为核心，迅速占据领先地位。MR-MUF 技术结合了批量回流焊（MR）和模塑底部填充（MUF）两个关键步骤。批量回流焊通过熔化堆叠芯片间的凸块实现芯片间的电气连接。随后，模塑底部填充在芯片堆叠之间注入保护材料，增强了结构的耐久性和散热效果。具体到技术流程，DRAM 芯片下方设有用于连接芯片的铅基“凸块”。MR 技术通过加热熔化这些凸块完成焊接。焊接完成后，进行 MUF 步骤，此时注入以优异散热性能著称的环氧树脂密封剂，填充芯片间的空隙并封装。通过加热和加压使组件硬化，完成 HBM 的封装过程。SK 海力士表示，MR-MUF 工艺确保了 HBM 中超过 10 万个凸点互连的高质量，增加了散热凸点的数量，实现了更佳的散热效果。这些优势巩固了 SK 海力士在 HBM 市场的竞争力，并助其在 HBM3 市场占据了领先地位。

图表14：MR-MUF 处理流程



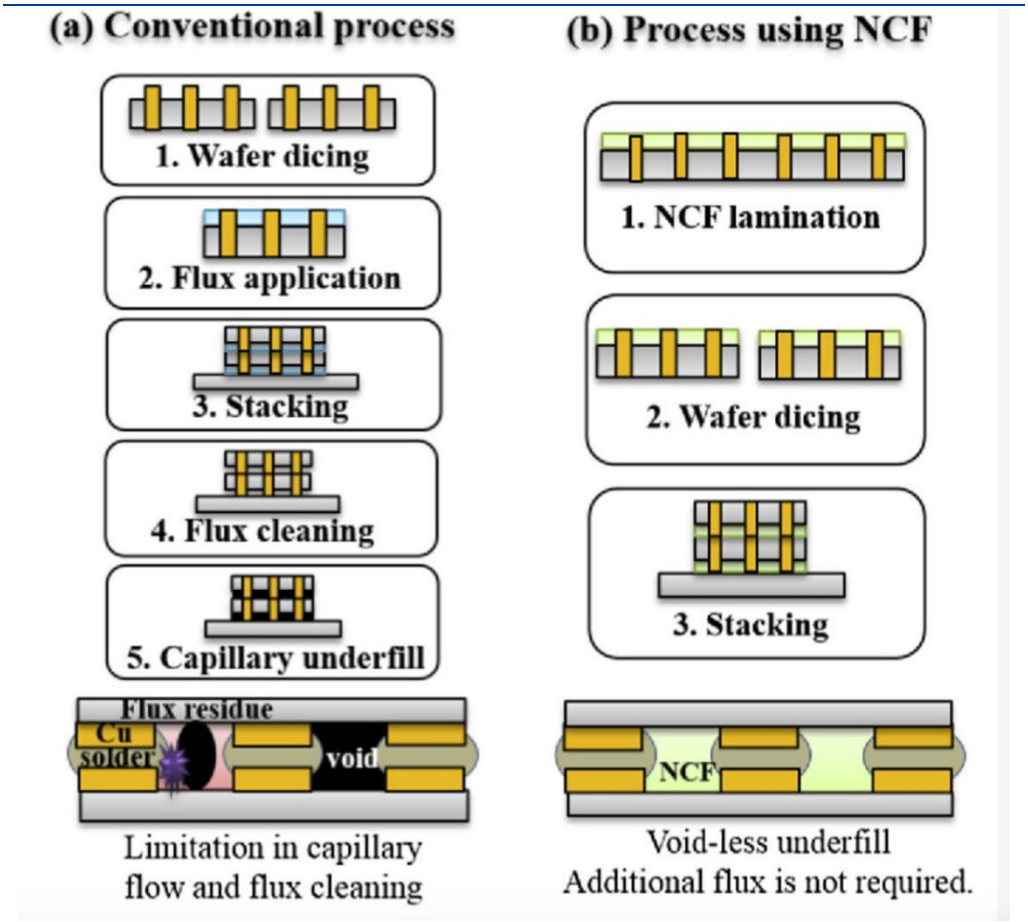
数据来源：semi analysis，华福证券研究所

2.2 TC-NCF 技术：高度持续降低，适合 12-16 高层堆叠

和海力士不同，在 HBM 封装上，三星采用的是 TC-NCF（thermal compression with non-conductive film）技术，也就是非导电薄膜热压缩工艺。该过程需要在高温

高压环境下进行。而在每次堆叠芯片时，都会在各层之间放置一层不导电的薄膜。该薄膜是一种聚合物材料，用于使芯片彼此绝缘并保护连接点免受撞击。随着发展，三星逐渐减少了 NCF 材料的厚度，将 12 层第五代 HBM3E 的厚度降至 7 微米(μm)。公司表示："这种方法的优点是可以最大限度地减少随着层数增加和芯片厚度减小而可能发生的翘曲，使其更适合构建更高的堆栈。"

图表15：NCF 工艺示意图



数据来源：Jiwon shin(2015.02):《Non-conductive film with Zn-nanoparticles (Zn-NCF) for 40 μm pitch Cu-pillar/Sn-Ag bump interconnection》，华福证券研究所

在三星看来，HBM 的热阻主要受芯片间距的影响，而三星拥有先进的高密度堆叠芯片控制技术，减少芯片之间 NCF 材料的厚度，并利用热压缩技术使芯片更加紧密。这种创新方法实现了业界最小的 7 微米(um)芯片间距。此外，在芯片键合过程中，三星策略性地设计了需要信号传输的小凸块和散热至关重要的大凸块。这种优化增强了散热和产量。此外，应用工艺技术在有限的封装尺寸内最小化单个 DRAM 芯片的尺寸，确保了卓越的量产能力和可靠性，从而提供了显着的竞争优势。

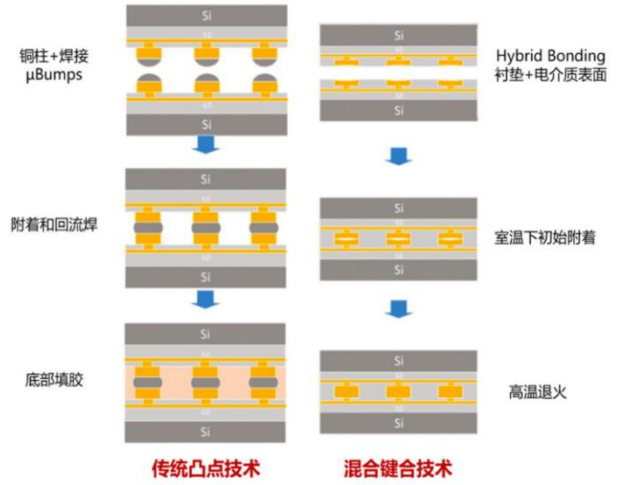
2.3 混合键合技术：无需凸点，进一步降低高度

2.3.1铜铜-混合键合：兼具低间距、多接点、低厚度等特质

铜铜-混合键合（Cu-Cu hybrid bondind）是一种将介电键（SiOx）与嵌入金属（Cu）结合形成互连的工艺技术。混合键合无需通过芯片间上下凸点的焊接实现互

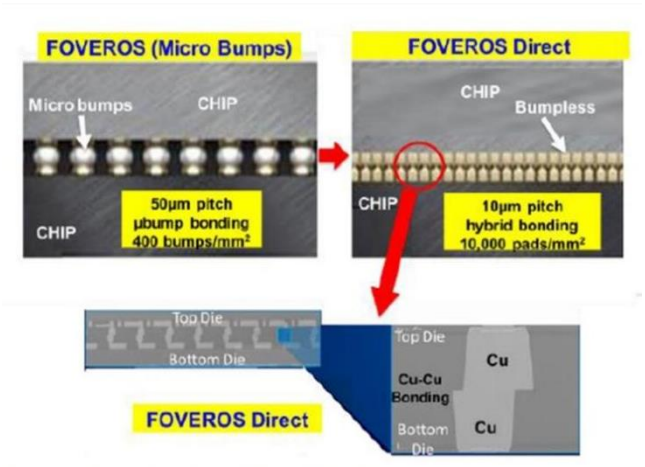
连，因为不依赖焊料，混合键合可实现超细间距和更小的接点尺寸，从而实现单位面积上更多的接点数量。此项技术不仅可以使芯片节距达到 10μm 及以下，未来有望缩小至 2μm 及以下，在散热效率上相较微凸点提升约 20%。

图表16：微凸点技术与混合键合技术对比



数据来源：公众号“SiP 与先进封装技术”，华福证券研究所

图表17：Intel 混合键合接点与微凸点横截面比较图



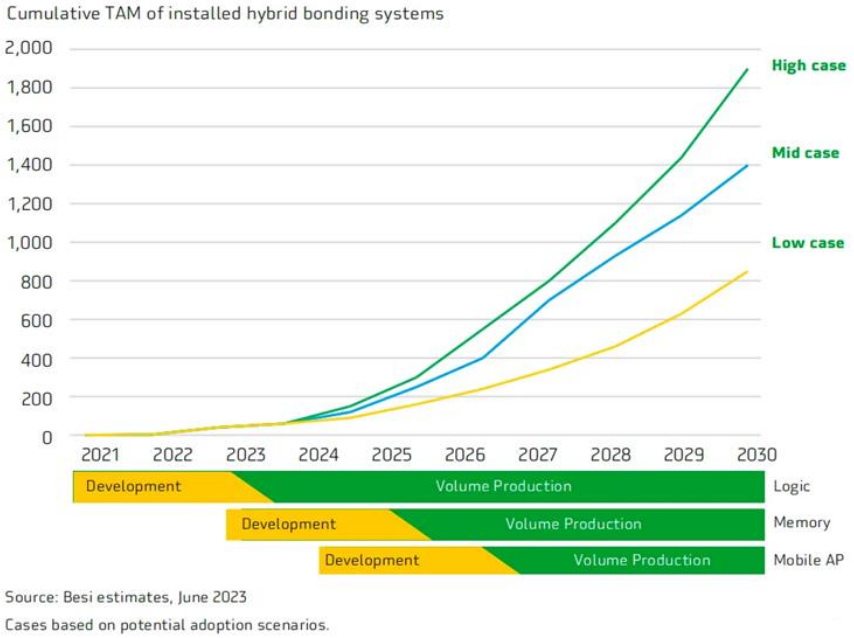
数据来源：R. Mahajan and S. Sane(2021.8):《Advanced packaging technologies for heterogeneous integration》，华福证券研究所

2.3.2混合键合设备的引入：国内外设备水平仍有差距

混合键合设备单机价值量高，行业头部领先优势明显。目前混合键合设备分为两类：一种是基于 wafer to wafer 技术的，代表性公司有奥地利的 EVG 与德国的 SUSS，另一种是基于 die to die 技术，此项技术可以用于支持 CoWos 先进封装，代表性公司为荷兰的 Besi。同时由于在贴片机上存在精度越高设备价格越高的情况，因此混合键合设备在定价上也将显著高于此前的 Flip chip（倒装芯片）或 TCB 键合系统，据 Besi 估计，键合设备价格将达到 200-250 万欧元每台。

应用领域广泛，混合键合设备预期需求增加。目前，混合键合已经成功用于商业生产数据中心和其他高性能计算应用的高端逻辑设备。AMD 作为第一家推出采用铜混合键合芯片的供应商。在 AMD Ryzen 7 5800x 的小芯片设计中，就采用了台积电的混合键合技术 SoIC，将 7nm 64MB SRAM 堆叠并键合到 7nm 处理器上，使内存密度增加了两倍。Meta 在 2024 IEEE 国际固态电路会议(ISSCC)介绍了其最新的采用 3D 堆叠芯片的 AR 处理器，也采用了混合键合技术并成功地在动作追踪上相较此前产品速度提升了 40%。Yole 也指出，芯片到晶圆混合键合技术即将渗透到服务器、数据中心以及未来的移动应用处理器（APUs）系统中。设备厂商 Besi 表示，混合键合有潜力在未来十年成为 3 纳米以下器件的领先组装解决方案。预计存储领域未来贡献混合键合设备明显增量，**保守预计 2026 年需求量超过 200 台**。

图表18：混合键合潜在的市场应用



数据来源：Besi 官网，华福证券研究所

混合键合设备国内起步较晚，距国际领先水平仍有 5-6 年差距。目前国内的设备厂商与海外的差距大约 5-6 年。要缩短这一时间差，国内企业首先需要与能够成熟进行该工艺的企业（如日月光、台积电、矽品）合作，共同打磨设备，以实现设备与工艺的匹配，随后可尝试逐渐与苹果，三星等终端厂商进行合作。

国内多家厂商正积极布局混合键合设备，目前国内混合键合上最具领先优势的公司为拓荆科技。拓荆科技研发的晶圆对晶圆键合设备 Dione300 已成功通过验证并投入商业使用，该设备的性能和产能指标均已达到国际领先水平。而其芯片对晶圆键合的表面处理设备 Pollux 也已发送至客户处进行测试验证。芯源微公司生产的临时键合设备和解键合设备已经获得了国内多家客户的青睐，并且订单量不断增加。华卓精科推出的 UPHBS300 晶圆级键合机旨在与国际知名企业 EVG 竞争。此外，去年 12 月，芯睿科技这家国内设备制造商在完成一轮超过亿元人民币的融资后，专注于半导体晶圆键合设备的研发，目前 wafertowafer 混合键合技术的开发正在稳步推进。国内企业在混合键合技术领域的迅猛进步，将极大地促进我国半导体产业的技术革新和产业升级。

图表19：拓荆科技 PEVCD 设备（NF-300HTEOS）

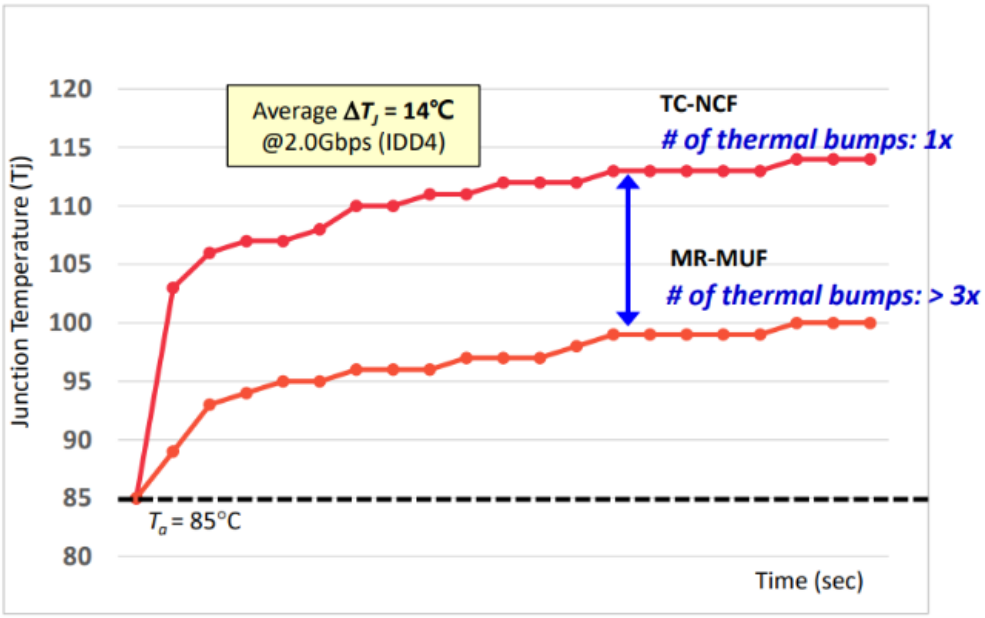
数据来源：拓荆科技官网，华福证券研究所

3 HBM4 技术路线：海力士优势明显，三星/美光发力追赶

3.1 SK 海力士：先发优势明显，MR-MUF 良率遥遥领先

海力士目前 HBM3E 良品率已达 80%。三星与美光在 HBM 封装上均采用 TC-NCF（基于热压的非导电薄膜 Thermal Compression - Non Conductive Film）工艺，该过程需要高温高压环境将凸点（bumps）推入非导电薄膜，在单个 DRAM 高度减少的环境下更易导致芯片翘曲。而海力士所采用的先进的 MR-MUF 技术通过在芯片间注入 EMC（液态环氧树脂模塑料 Epoxy Molding Compound）填充芯片之间或芯片与凸块之间间隙。由于 EMC 材料本身具备中低温固化、低翘曲、低吸水率等优点，无需借助高温高压，可有效解决芯片翘曲从而提升良率。相比于 NCF，MUF 具有更高的热导性，在一定条件下，MUF 材料温度要低 14℃，也即散热效果更好。根据海力士最新披露的数据，公司 HBM3E 产品上的良品率以达到 80%，远远超出此前行业预期的 60%-70%，同时也大幅领先竞争对手三星与美光的良率。

图表20：MR-MUF 对比 TC-NCF 温度下降 14℃，散热效果更好



数据来源：《A study on the adavanced chip to wafer stack for better thermal dissipation of high bandwidth memory》，华福证券研究所

海力士在 HBM4 上仍将采用 Advanced MR-MUF 工艺，还致力于 Fan-out RDL（扇出型重新分配层）及混合键合（Hybrid bonding）等下一代先进封装技术的开发。当中，混合键合也是被看作是 HBM 封装的又一个新选择。但由于混合键合技术非常复杂，需要控制键合层的平整度和键合强度，粒子控制也需要在纳米级别进行，这将导致 HBM 在生产效率与良品率上有所欠缺。同时随着 HBM 标准限制的放宽。我们预计，海力士仍将在 HBM4 上采用成熟的 MR-MUF 技术。

海力士将为新一代 HBM 产品兴建封装厂。为巩固在 AI 半导体技术与客户合作领域的领先地位，海力士决定在美国印第安纳州西拉法叶市投资约 38.7 亿美元，兴建一座尖端的封装生产设施，专注于 AI 存储器生产。海力士还将与当地机构携手开展研发工作。**预计从 2028 年起**，该设施将大规模生产包括 HBM 在内的下一代 AI 存储器产品。此举不仅将使海力士能够为客户提供更多定制化的存储器产品，满足他们日益增长的需求和期望，还将在全球 AI 半导体供应链中发挥领导作用，应对 HBM 需求的迅猛增长。

3.2 三星：万亿韩元投入，一站式方案争夺市场份额

投入规模：万亿韩元投入，激进扩产计划。在各大厂商积极扩产 HBM 的情况下，相比于竞争对手，三星在资金上优势更为明显。三星在 HBM 上的扩产计划显然更为激进，三星于 23 第三季度宣布计划在 2024 年将 HBM 年产能扩大 2.5 倍以上，并投资 7000-10000 亿韩元从子公司三星显示（SDC）处收购天安工厂的建筑和设施用于建设新的 HBM 封装线。当前 HBM 市场仍处于供不应求的阶段，三星 8 层与 12 层的 HBM3 均已通过 AMD Instinct MI300 系列的验证,未来预计随着 AI 算力需求的进一步增加，三星的扩产计划有助于帮助公司获得更多订单。

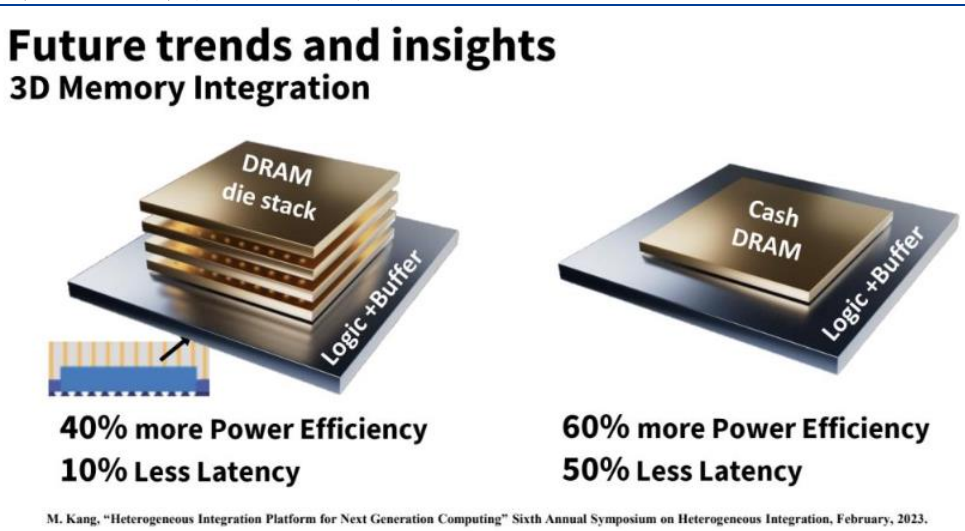
封装方案：自研 2.5D 以及 3D 封装方案。三星同时作为存储厂商和晶圆代工厂，提供了集存储、AI 芯片设计、晶圆代工和封装的一站式服务。公司提供了包括 I-CubeS（2.5D）、I-CubeE(2.5D)、X-Cube(TCB)(3D)和 X-Cube(HCB)(3D)四种不同的先进封装方案。

图表21：三星封装方案

数据来源：OPC 演讲《AI/HPC: Advanced package technologies for chiplet adoption and memory integration in HPC/AI applications》，华福证券研究所

针对未来潜在的封装方式，三星提出了两种构想。第一种与海力士的方案相同，通过将 DRAM 芯片堆叠在 GPU 上，可以在提升 40%的功耗效率降低 10%的延迟。第二种是通过将 Cash DRAM 堆叠在 GPU 上，在提升 60%的功耗效率降低 50%的延迟。

图表22：三星未来 3D 封装方案计划



数据来源：OPC 演讲《AI/HPC:Advanced package technologies for chiplet adoption and memory integration in HPC/AI applications》，华福证券研究所

我们认为，上述方案对封装工艺提出了更高的要求，而海力士由于没有晶圆代工厂，因此选择与台积电共同合作。过于依赖台积电的产能同时在生产过程中也存在时间和空间上的错配。三星存储与封装部门协同有望大大缩短 HBM4 从研发到生产的中间环节，并在未来的量产中缩短从内存芯片制造、封装到交付的周期，从而能占得 HBM4 及后续产品先机。根据芯智讯的报道，在当前台积电产能不足的情况下，三星

的先进封装（AVP）团队将为英伟达提供 Interposer（中间层）和基于 I-Cube 技术的 2.5D 先进封装产能，在长期来看有利于三星争夺 HBM 市场份额。

制程上：1-α制程落后竞争对手一代。目前海力士与美光在 HBM3E 上均采用 1-β（第五代 10nm）制程，对于未来的 HBM4E，美光计划使用 32GB DRAM 芯片，并首次采用 10nm 级的 6 代(1-γ)制程。SK 海力士也表示正在基于第六代 10nm 级 1-γ 制程 32Gb DRAM 裸片构建 HBM4E 内存。而三星目前在 HBM3E 上使用的仍是 1-α（第四代 10nm）制程。随着未来 HBM 内存密度要求的提升，以垂直方式来堆叠芯片也必然增加散热上的负担。三星未来能否突破制程上的缺点也为其争夺大客户市场份额带来新的挑战。

图表23：各厂商 HBM 制程工艺

各厂商HBM制程工艺			
产品名称	三星	海力士	美光
HBM3E	1-α	1-β	1-β
HBM4	计划1-γ	计划1-γ	计划1-γ
HBM4E	计划1-γ	计划1-γ	计划1-γ
注：DRAM制程迭代顺序为1y、1z、1α、1β和1γ			

数据来源：公司官网，华福证券研究所

现有键合技术：TC-NCF 在更高堆叠层数上或更具优势。尽管三星目前采用的 TC-NCF 技术可以通过减少芯片之间 NCF 材料的厚度，并利用热压缩技术使芯片更加紧密，在 HBM3E 上实现了业界最小的 7 微米(um)芯片间距。但由于工艺中不可避免的高温高压环境，在将凸点（bumps）推入非导电薄膜时更容易导致芯片翘曲，单位 HBM 的损耗大，良率相比于海力士 MR-MUF 技术落后较多。但三星也表示 HBM 在最多 8 个堆叠时，MR-MUF 的生产效率比 TC-NCF 更高，**一旦堆叠达到 12 个或以上，后者将具有更多优势**，而未来 HBM4 高度的放宽势必将增加 HBM4 堆叠层数至 12-16 层，这将为 TC-NCF 工艺带来更大发挥的可能。同时考虑到三星雄厚的资金优势，由于 HBM 具有高价值、高毛利的特点，凭借生产效率与规模的优势可以抵消一部分良率不高带来的负面影响。

混合键合的引入：受键合设备配置成本，键合良率以及 HBM4 高度限制放宽影响，短期预计仍采用现有技术，但长期来看混合键合技术系大势所趋。目前三星在 HBM4 内存键合技术方面采取了两条腿走路的策略，同时开发混合键合和传统的 TC-NCF 工艺。现有键合工艺需要在 DRAM 内存层间添加凸块，而混合键合无需使用填充凸块进行连接的硅通孔(TSV)，上下两层直接铜对铜连接，显著提高了信号传输速率，同时降低 DRAM 层间距，进而减少 HBM 模块整体高度，适应了 AI 计算对高带宽的需求。三星高管在今年 4 月表示该公司已成功制造了采用 16 层混合键合技术的 HBM3 内存样品，并且该内存工作正常。他表示，未来 16 层堆叠混合键合技术将用于 HBM4 内存量产。这表明在混合键合技术上三星相较于竞争对手或具有一定领先

优势。

图表24：EVG 晶圆键合设备

数据来源：EV Group 官网，华福证券研究所

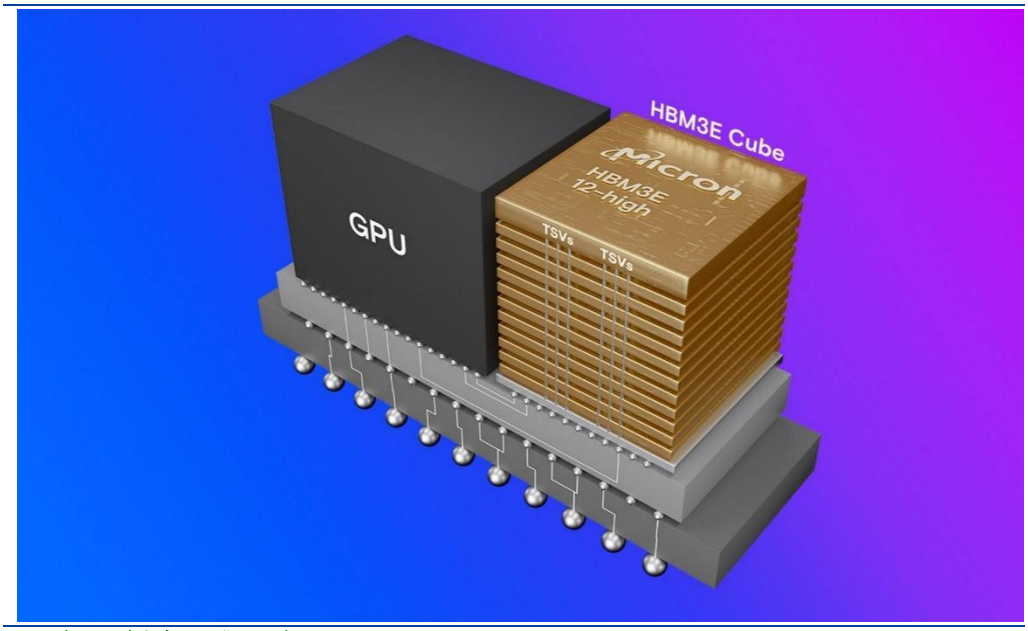
但从设备端来看，混合键合技术所需设备与当前三大厂并不匹配，混合键合的工艺流程涉及许多传统上仅由晶圆代工厂专用的工具，例如 CVD、CMP 和表面离子活化等，存储厂商较难在短期内配备大规模生产条件，同时其工艺要求较高，如清洗工艺要求需要 ISO3 及以上。此外混合键合所需要的设备价格要显著高于目前最先进的 Flip chip（倒装芯片）或 TCB 键合系统，在配置成本上存在一定压力。同时由于混合键合的工艺难度极高以及 JEDEC 将 HBM4 的封装厚度标准放宽至 775 微米，在 HBM4 项目初期预计三大厂商还是求稳为主，使用现有键合技术来保证生产效率与良率。

3.3 美光：破釜沉舟，越过 HBM3 力求弯道超车

在市场定位上：美光目前以承接台积电与三星溢出份额为主，有望成为英伟达的第二选择。美光并非 HBM 市场的先行者，但预期将成为 HBM 增长的主要受益者。美光此前自研的 HMC 未获广泛的应用，导致公司在 HBM 赛道上起步较晚，2023 年 HBM 市场份额仅占 9%。为了追赶海力士与三星，美光越过 HBM3 直接开始与台积电共同研发基于 1-β 制程的 HBM3E，并于今年 2 月率先宣布 HBM3E 的量产，确认供货英伟达 H200。中短期来看，HBM3E 处于供不应求的阶段。此前美光主要承接台积电与三星溢出的份额，但随着美光与海力士一同得到英伟达 HBM3E 认证，受益于英伟达对 HBM 大量需求，未来有望成为平替。

在性能上：采用 1-β 先进制程，HBM3E 功耗行业最低。美光采用了和海力士先进的 1-β 制程，并计划在未来产品上采用 1-γ 制程。目前美光通过 36GB 12 层 HBM3E 样品扩大了其领先地位，该产品预计将提供超过 1.2TB/s 的性能和卓越的能效，在功耗上比竞争 HBM3E 产品低约 30%，在性能上有约 10%的提升。

图表25：美光 12 层 36GB HBM3E 产品



数据来源：美光官网，华福证券研究所

在供应链上：积极布局中国台湾供应链。1）美光与台积电 OIP 3D-Fabric 联盟合作，加速客户验证和纠错过程。2）美光 HBM3E 封测和出货在中国台湾完成，与当地供应链紧密合作。3）与 IP 供应商合作提供 GPU 与 HBM 快速交互技术。

面对的挑战：1）相比于海力士多年耕耘积攒下的丰富合作商资源，**目前美光 HBM 产品过于单一**，仅供货于英伟达。在市场份额争夺上仍有很大压力。2）**美光目前采用的 TC-NCF 键合技术良率偏低**，导致制取 DRAM 芯片成本较高。3）由于资金上的劣势，在各大厂商扩产的趋势下，**美光生产规模相对有限，研发重心转移到 HBM 上势或将挤占其他产线上 DRAM 产能。**

4 建议关注行业重点公司

我们认为，HBM4 性能参数与工艺均有较大革新，同时国内先进存储厂商亦有望陆续布局 HBM 产线，带动相关设备、材料、封测厂商业绩增长。建议关注：

HBM 设备：芯源微、拓荆科技、盛美上海、赛腾股份等

HBM 材料：华海诚科、联瑞新材、强力新材、飞凯材料等

HBM 封测：长电科技、通富微电、佰维存储等

海力士分销：香农芯创等

5 风险提示

AI 需求不及预期风险。总体来看，人工智能芯片技术仍处于发展阶段，技术迭代速度较快，若后续应用落地需求不及预期，或将影响上游 GPU 与 HBM 需求量。

技术迭代不及预期风险。目前三大原厂均持续进行 HBM4 等相关产品的研发，

但由于 HBM 涉及 2.5D/3D 封装、TSV 等先进封装技术，本身具有较高的技术难度，而 HBM4 对芯片堆叠、容量、速率均提出更高要求，研发难度大幅提升，未来技术迭代存在不及预期风险。

行业竞争加剧风险。目前 HBM 市场为三星、美光、海力士三大原厂所占据，但由于 HBM 具备高价值量，未来或为存储原厂竞争主战场。三大原厂未来在英伟达等核心客户的产品导入、HBM 产能、HBM 价格等维度或存在竞争加剧风险。

分析师声明

本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

一般声明

华福证券有限责任公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。本报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。

本报告的信息均来源于本公司认为可信的公开资料，该等公开资料的准确性及完整性由其发布者负责，本公司及其研究人员对该等信息不作任何保证。本报告中的资料、意见及预测仅反映本公司于发布本报告当日的判断，之后可能会随情况的变化而调整。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息及资料保持在最新状态，对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

在任何情况下，本报告所载的信息或所做出的任何建议、意见及推测并不构成所述证券买卖的出价或询价，也不构成对所述金融产品、产品发行或管理人作出任何形式的保证。在任何情况下，本公司仅承诺以勤勉的职业态度，独立、客观地出具本报告以供投资者参考，但不就本报告中的任何内容对任何投资做出任何形式的承诺或担保。投资者应自行决策，自担投资风险。

特别声明

投资者应注意，在法律许可的情况下，本公司及其本公司的关联机构可能会持有本报告中涉及的公司所发行的证券并进行交易，也可能为这些公司正在提供或争取提供投资银行、财务顾问和金融产品等各种金融服务。投资者请勿将本报告视为投资或其他决定的唯一参考依据。

投资评级声明

类别	评级	评级说明
公司评级	买入	未来 6 个月内，个股相对市场基准指数指数涨幅在 20%以上
	持有	未来 6 个月内，个股相对市场基准指数指数涨幅介于 10%与 20%之间
	中性	未来 6 个月内，个股相对市场基准指数涨幅介于-10%与 10%之间
	回避	未来 6 个月内，个股相对市场基准指数涨幅介于-20%与-10%之间
	卖出	未来 6 个月内，个股相对市场基准指数涨幅在-20%以下
行业评级	强于大市	未来 6 个月内，行业整体回报高于市场基准指数 5%以上
	跟随大市	未来 6 个月内，行业整体回报介于市场基准指数-5%与 5%之间
	弱于大市	未来 6 个月内，行业整体回报低于市场基准指数-5%以下

备注：评级标准为报告发布日后的 6~12 个月内公司股价（或行业指数）相对同期基准指数的相对市场表现。其中 A 股市场以沪深 300 指数为基准；香港市场以恒生指数为基准，美股市场以标普 500 指数或纳斯达克综合指数为基准（另有说明的除外）

联系方式

华福证券研究所 上海
公司地址：上海市浦东新区浦明路 1436 号陆家嘴滨江中心 MT 座 20 层
邮编：200120
邮箱：hfyjs@hfzq.com.cn