

HBM: AI的内存瓶颈, 高壁垒高增速

分析师: 刘双锋

liushuangfeng@csc.com.cn

SAC 编号: S1440520070002

分析师: 章合坤

zhanghekun@csc.com.cn

SAC 编号: S1440522050001

分析师: 孙芳芳

sunfangfang@csc.com.cn

SAC 编号: S1440520060001

研究助理: 何昱灵

heyuling@csc.com.cn

发布日期: 2024年3月9日



摘要

- HBM是限制当前算力卡性能的关键因素，海力士、三星、美光正加大研发投入和资本开支，大力扩产并快速迭代HBM，预计2024年HBM3e 24GB/36GB版本将量产/发布，内存性能进一步提高。HBM供需将持续紧俏，市场规模高速增长。通过分析生产工艺（TSV、键合等）和技术演进方向（先进制程、叠层），我们认为封装测试、前道和后道先进封装的设备和材料将是HBM主要受益方向。
- HBM是当前算力的内存瓶颈。存储性能是当下制约高性能计算的关键因素，从存储器到处理器，数据搬运会面临带宽和功耗的问题。为解决传统DRAM带宽较低的问题，本质上需要对单I/O数据速率和位宽进行提升。HBM由于采用了TSV、微凸块等技术，DRAM裸片、计算核心间实现了较短的信号传输路径、较高的I/O数据速率、高位宽和较低的I/O电压，因此具备高带宽、高存储密度、低功耗等优势。即便如此，当前HBM的性能仍然跟不上算力卡的需求。
- 三大原厂持续加大研发投入，HBM性能倍数级提升。随着技术的迭代，HBM的层数、容量、带宽指标不断升级，目前最先进的HBM3e版本，理论上可实现16层堆叠、64GB容量和1.2TB/s的带宽，分别为初代HBM的2倍、9.6倍和4倍。从Trendforce公布的HBM Roadmap来看，2024年上半年，海力士、三星、美光均会推出24GB容量的HBM3e，均为8层堆叠。2024年下半年，三家厂商将推出36GB版本的HBM3e，或为12层堆叠。此外，HBM4有望于2026年推出。
- HBM制造集成前道工艺与先进封装，TSV、EMC、键合工艺是关键。HBM制造的关键在于TSV DRAM，以及每层TSV DRAM之间的连接方式。目前主流的HBM制造工艺是TSV+Micro bumping+TCB，例如三星的TC-NCF工艺，而SK海力士则采用改进的MR-MUF工艺，在键合应力、散热性能、堆叠层数方面更有优势。目前的TCB工艺可支撑最多16层的HBM生产，随着HBM堆叠层数增加，以及HBM对速率、散热等性能要求的提升，HBM4开始可能引入混合键合工艺，对应的，TSV、GMC/LMC的要求也将提高。

摘要

- **AI刺激服务器存储容量扩充，HBM需求强劲。**随着云计算厂商将更多资本开支投入AI基础设施，AI服务器ODM对2024年展望乐观，预计2024年AI服务器出货量继续大幅增长。相较于一般服务器而言，AI服务器多增加GPGPU的使用，以NVIDIA A100/H100 80GB配置8张计算，HBM用量约为640GB，超越常规服务器的内存条容量，H200、B100等算力卡将搭载更高容量、更高速率HBM。我们测算，随着算力卡单卡HBM容量提升、算力卡出货量提升、技术迭代带来单GB HBM单价提升，2023年HBM市场规模为40亿美元，预计2024年增长至148亿美元，2026年增长至242亿美元，2023~2026年CAGR为82%。
- **目前HBM供应链以海外厂商为主，部分国内厂商打入了海外存储/HBM供应链。**国产HBM正处于0到1的突破期，HBM供应主要为韩系、美系厂商，国内能获得的HBM资源较少。随着国产算力卡需求快速增长，对于算力卡性能至关重要的HBM也有强烈的供应保障诉求和国产化诉求。建议关注：封测、设备、材料等环节。
- **相关标的：**
 - ✓ 封测：通富微电、长电科技、深科技；
 - ✓ 设备：中微公司、北方华创、拓荆科技、芯源微、赛腾股份、华海清科、精智达、新益昌；
 - ✓ 材料：雅克科技、联瑞新材、华海诚科、强力新材、天承科技、飞凯材料、壹石通、兴森科技；
 - ✓ 代理：香农芯创。
- **风险提示：**AI技术落地不及预期；国际环境变化影响；宏观环境的不利因素；市场竞争加剧风险。

一、HBM：算力的内存瓶颈

二、SK海力士HBM工艺分析：TSV、EMC、混合键合成趋势

三、市场测算：未来三年CAGR超80%

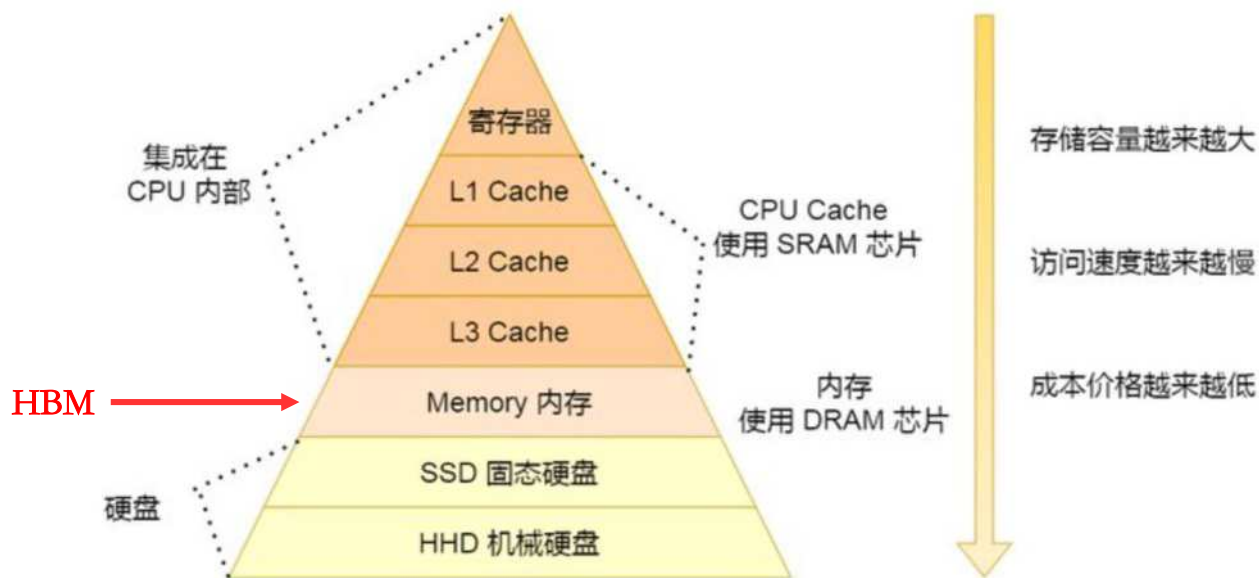
四、相关标的：设备、材料迎来新成长

五、风险提示

CPU拥有多级缓存架构，HBM属于内存环节

- 现代CPU为了提升执行效率，减少CPU与内存的交互(交互影响CPU效率)，一般在CPU上集成了**多级缓存架构**。
- CPU缓存即**高速缓冲存储器**，是位于CPU与主内存间的一种容量较小但速度很高的存储器。由于CPU的速度远高于主内存，CPU直接从内存中存取数据要等待一定时间周期，Cache中保存着CPU刚用过或循环使用的一部分数据，当CPU再次使用该部分数据时可从Cache中直接调用，减少CPU的等待时间，提高了系统的效率。
- 从CPU到用户数据，需经历“寄存器>SRAM>内存>SSD或HDD”的路径，HBM属于内存的一种。

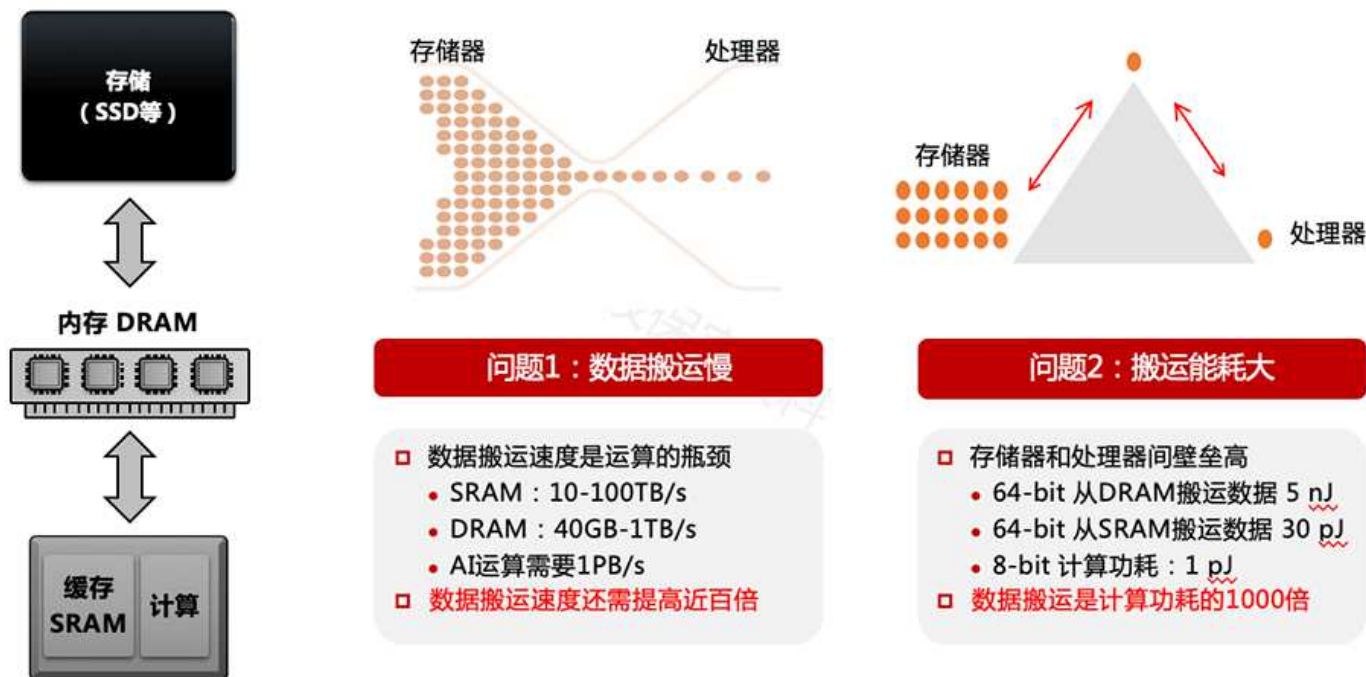
图表：CPU的多级缓存架构



存储性能是当下制约高性能计算的关键因素

- 内存墙（传输带宽慢或容量有限）是算力提升的重要瓶颈。
- 从存储器到处理器，数据搬运会面临2个问题：（1）数据搬运慢；（2）搬运能耗大。

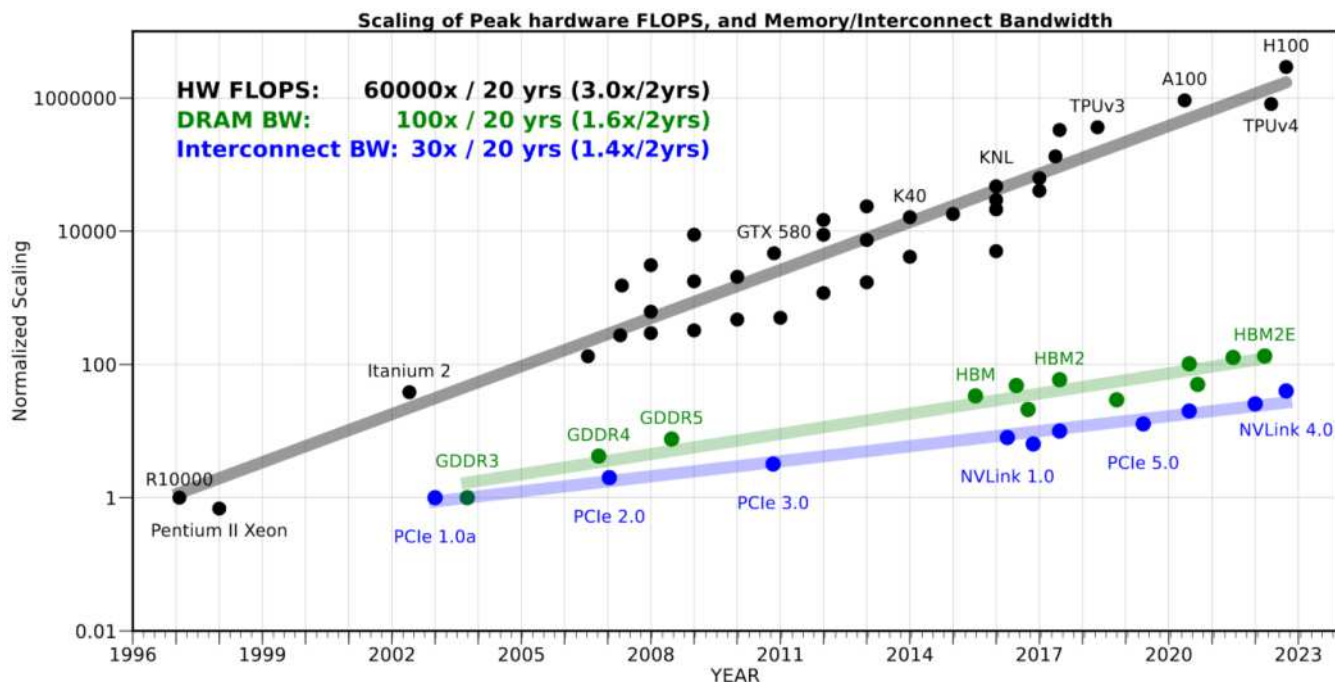
图表：数据传输的速率及能耗限制算力性能发挥



存储性能是当下制约高性能计算的关键因素

- 算力的增速远大于存储的增长速度。目前绝大部分的计算系统基于冯诺依曼计算机体系，而该体系下存储与计算单元分离，需要通过总线不断在存储与处理器之间传输数据，因此存储的带宽制约了算力的利用效率。此外，从外部处理器到内存之间不断进行数据的搬运，搬运时间往往是运算时间的成百上千倍，产生的无用能耗过多。

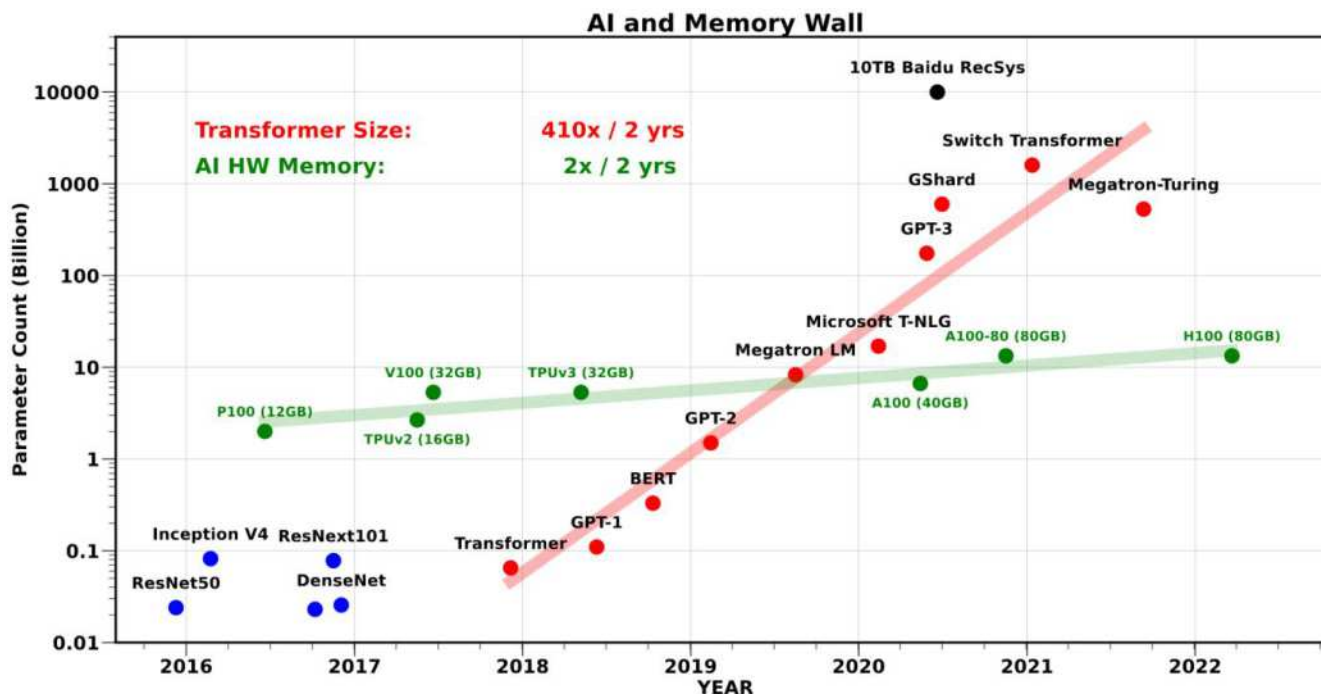
图表：算力的增速远超存储与互连带宽的增速



存储性能是当下制约高性能计算的关键因素

- 模型体量的增速远大于算力卡存储容量的增速。随着 Transformer 模型的大规模发展和应用，模型大小每两年平均增长了240倍，而单个GPU内存容量仅以2年2倍的速度扩大。为了摆脱单一算力芯片内存有限的问题，可以将模型部署于多颗GPU上运行，但在算力芯片之间移动数据，仍然比单一芯片内部移动数据低效，因此算力芯片内存容量的缓慢增速制约了更大规模的模型应用。

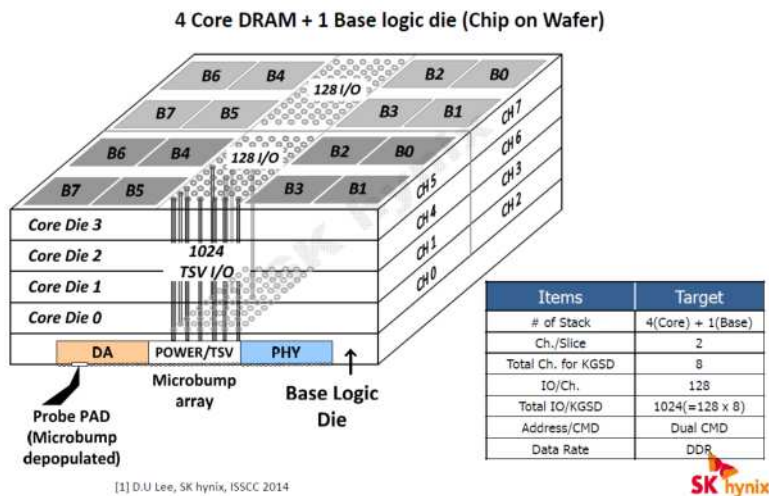
图表：大模型体积的增速远高于算力芯片存储容量的增速



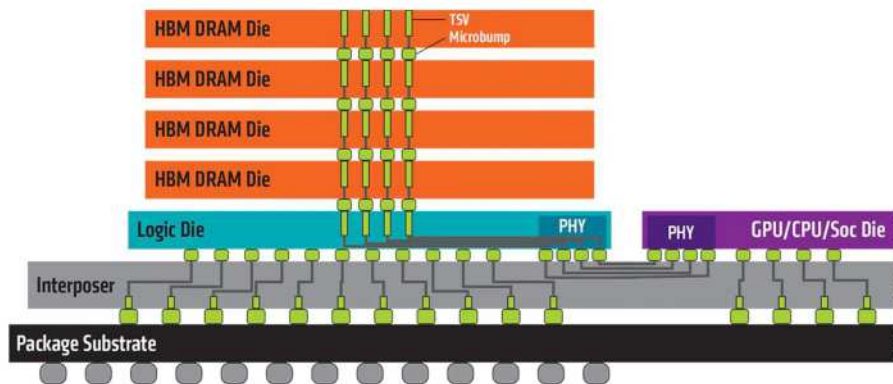
HBM: 基于TSV技术获得的高带宽内存，已成为高性能计算的首选

- 如同闪存从2D NAND向3D NAND发展，DRAM也正在从2D向3D技术发展，HBM（High Bandwidth Memory，高带宽存储）为主要代表产品。从结构上看其特点有：（1）3D堆叠结构并由TSV互连：HBM由多颗DRAM die堆叠成3D结构，使用TSV技术实现信号的共享与分配；（2）高I/O数量带来高位宽：HBM的每颗DRAM Die包含多个通道，可独立访问。每个通道又包含多个I/O口，位宽64/128bit，使HBM的总位宽高达1024bit。
- HBM性能优异，主要用于高性能计算芯片。HBM的结构特点为其带来存储密度更大、功耗更低、带宽更高的优势。HBM通常采用CoWoS等先进封装技术与计算核心进行互连，多用于与数据中心算力芯片GPU/FPGA/ASIC等配合工作。

图表： HBM结构示意图



图表： HBM采用CoWoS封装与计算核心互连



数据来源: SK Hynix, 中信建投

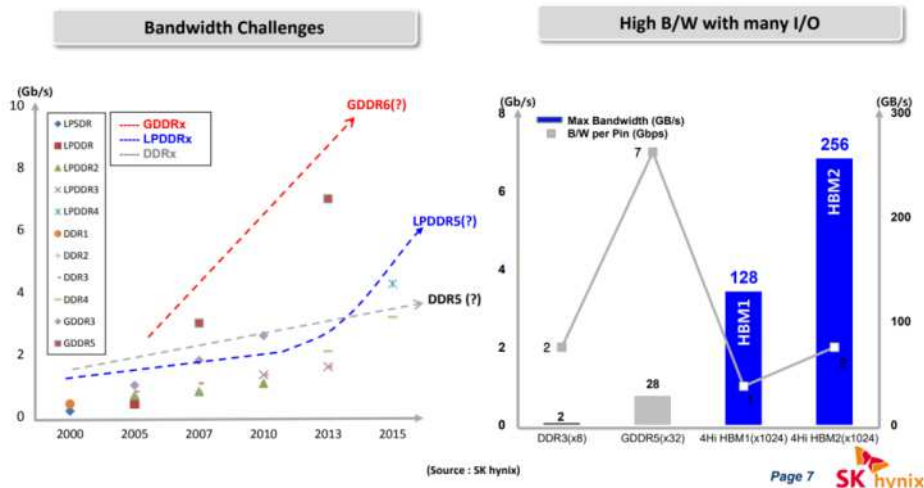
数据来源: AMD, 中信建投

HBM优势：相比GDDR，大幅增加计算核心可用的带宽

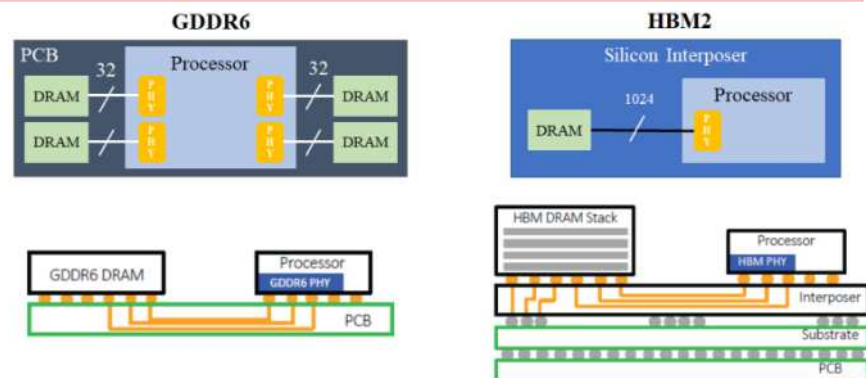
- 总带宽=I/O数据速率 (Gb/s) *位宽/8。为解决DDR带宽较低的问题，本质上需要对单I/O的数据速率和位宽 (I/O数*单I/O位宽) 进行提升，着重于不同方向发力的方案分别为GDDR和HBM。
- GDDR提升单I/O的数据速率，总带宽提升有限：GDDR采取大幅提升单I/O数据速率的手段来改善总带宽，GDDR5和GDDR6的单I/O数据速率已达到7 Gb/s到16Gb/s，超过HBM3的6.4 Gb/s。虽然GDDR的位宽相比DDR也有提升，但由于GDDR仍然是通过PCB与计算核心进行互连，总位宽的提升受到限制。
- HBM利用TSV技术提升I/O数，总带宽提升明显：HBM利用TSV技术在维持较低的单I/O数据速率的情况下，大幅提升了位宽进而获得了远优于GDDR的总带宽表现。

图表：HBM利用高位宽克服带宽的限制

HBM utilizes TSV technology to overcome DRAM bandwidth challenges



图表：决定GDDR、HBM带宽差异的核心因素对比

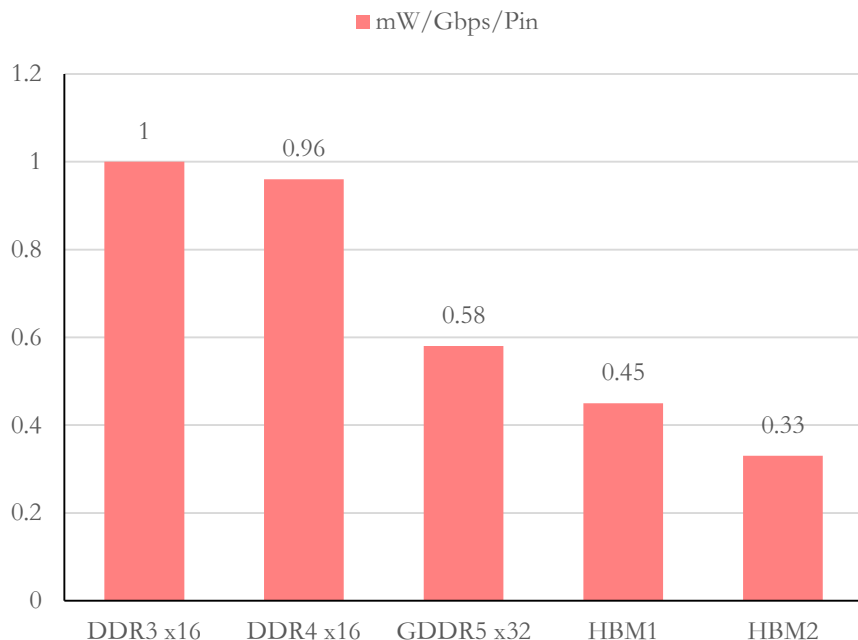


最新标准	GDDR6	HBM2e	HBM3
最高I/O速率(Gb/s)	16	3.6	6.4
每颗DRAM通道数	2	8	16
单通道位宽 (bit)	16	128	64
最高位宽 (bit)	32	1024	1024
最高带宽 (GB/s)	64	460	819

HBM优势：相比GDDR，功耗、集成度更具优势

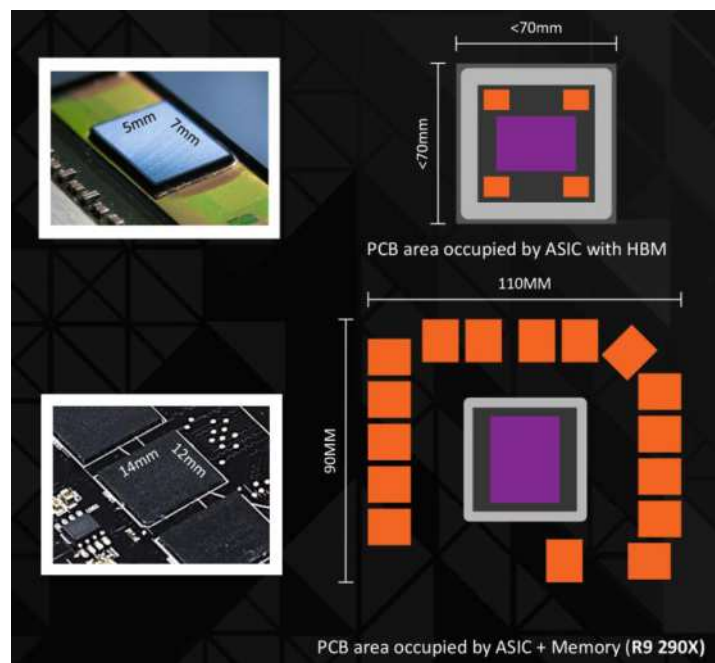
- HBM的高I/O速率和低电压使其获得优异的能效比。由于采用了TSV、微凸块等技术，DRAM裸片、计算核心间实现了较短的信号传输路径、较高的数据速率和较低的I/O电压，使HBM的能效比优于DDR和GDDR。
- HBM采用DRAM堆叠获得高存储密度，有利于提高集成度。相比于1GB GDDR5（4颗256M颗粒），1GB HBM1的表面积节省达到了94%。对比AMD使用4GB HBM1的GPU R9 Fury X与使用4GB GDDR5的R9 290X，HBM方案的面积可以控制在4900 mm²以内，而使用GDDR5方案的面积达到9900 mm²。

图表：HBM能效比优于GDDR与DDR



数据来源：Sk hynix, 中信建投

图表：HBM有效减少芯片的面积占用



数据来源：AMD, 中信建投

HBM标准历经多次升级，I/O速率、带宽获得明显提升

- **三大DRAM原厂推动技术迭代，性能表现持续提升。**HBM高带宽的特点使其适合数据中心GPU等高性能算力芯片，全球三大DRAM原厂陆续加强HBM的技术投入。随着技术的迭代，HBM的层数、容量、带宽指标不断升级，目前最先进的HBM3e版本，理论上可实现16层堆叠、64GB容量和1.2TB/s的带宽，分别为初代HBM的2倍、9.6倍和4倍。
- **容量（GB）：**由单颗DRAM颗粒的容量、DRAM颗粒的堆叠层数共同决定。
- **总带宽（GB/s）：**从HBM1到HBM3e，尽管不同代际之间的I/O数量和单I/O位宽存在差异，但总位宽均保持为1024bit，因此I/O速率的提升推动了总带宽提升。根据DigiTimes援引Seoul Economy的消息，HBM4有望将总位宽提升至2048bit，即使I/O速率维持不变，总带宽也将有翻倍增长。

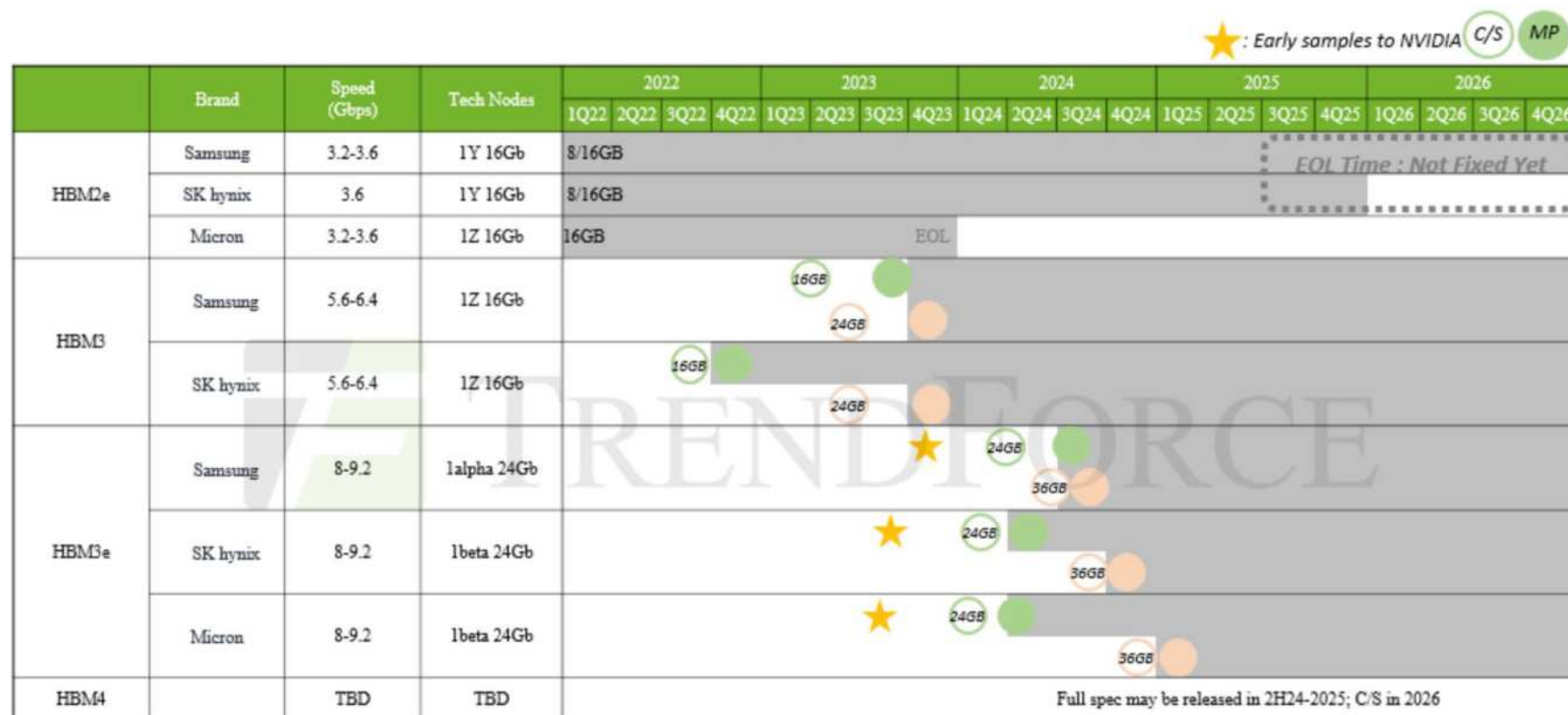
图表：各代HBM参数对比

代际	制程 (nm)	最大I/O 速率 (Gb/s)	通道数	单通道位 宽 (bit)	总位宽 (bit)	最大带宽 (GB/s)	可堆叠高 度	DRAM颗 粒最大容 量 (Gb)	HBM最 大容量 (GB)	Sk Hynix	Samsung	Micron
HBM	20nm	1	8	128	1024	128	8	16	16	√		
HBM2	20nm	2	8	128	1024	256	8	16	16	√	√	√
HBM2e	1y/1z 16Gb	3.6	8	128	1024	461	12	24	36	√	√	
HBM3	1z 16Gb	6.4	16	64	1024	819	16	32	64	√	√	
HBM3e	1α/1β 24Gb	9.6	16	64	1024	1229	16	32	64	√	√	√

2024年HBM3e 24GB版本和36GB版本将量产

- 自2023年ChatGPT发布以来，AI服务器的强劲需求正推动HBM快速迭代。从Trendforce公布的HBM Roadmap来看，2024年上半年，海力士、三星、美光均会推出24GB容量的HBM3e，按照1 alpha/beta 24Gb的工艺计算，均为8层堆叠。2024年下半年，三家厂商将推出36GB版本的HBM3e，或为12层堆叠。

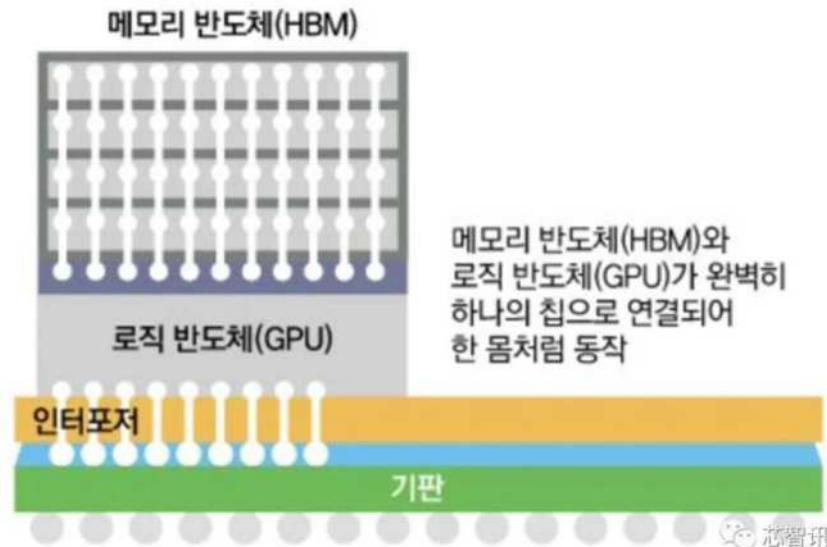
图表：各家HBM Roadmap



HBM4展望：除配置升级外，还可能朝客制化发展

- 当前供应商聚焦于迭代HBM3e，单die容量和堆叠层数仍有提升空间，但HBM4已经列入研发计划：
- 1、**堆叠层数**：HBM4在堆叠层数上，除了现有的8/12层外，将再往16层发展，更高层数也将带动新堆叠方式Hybrid bonding的需求。HBM4 12层将于2026年推出，而16层则预计于2027年问世。
- 2、**逻辑die**：受到GPU/HBM规格更往高速发展带动，Trendforce预计HBM最底层的Logic die（Base die）将采用12nm制程，该部分将由晶圆代工厂提供，使得单颗HBM需要结合晶圆代工厂与存储器厂的合作。
- 3、**客制化**：根据韩国中央日报报道，SK海力士正在招募CPU、GPU等逻辑芯片的设计人员，目标是将未来的HBM4以3D堆叠的形式堆叠在英伟达、AMD等公司的逻辑芯片上，预计该HBM4内存堆栈将采用2048位接口。该方案未定，只是在讨论中。

图表：HBM4概念图

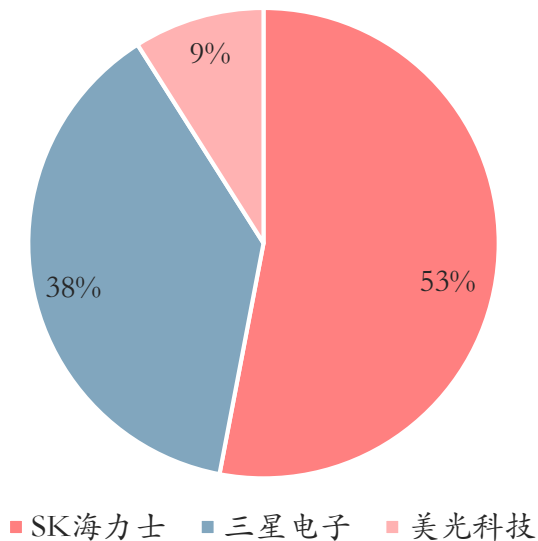


数据来源：芯智讯，韩国中央日报，中信建投

HBM市场寡头垄断，SK海力士占过半份额

- **HBM市场格局集中，SK海力士占有主导地位。**根据TrendForce，2022年三大原厂HBM市占率分别为SK海力士53%、三星电子约38%、美光约9%。
- **HBM市场竞争白热化。**2023年市场主要HBM代际是HBM2、HBM2e和HBM3，算力卡性能提升刺激HBM产品更迭，2023年下半年伴随NVIDIA H100与AMD MI300的搭载，HBM3渗透率提升。2024年伊始，SK海力士完成HBM3e开发，并送样英伟达测试，有望于上半年量产。预计三星电子和美光科技即将送样HBM3e，也有望于上半年量产，其中美光科技跳过了HBM3，直接研发HBM3e。三家原厂在HBM领域的竞争日趋白热化。

图表：2022年HBM市场竞争格局

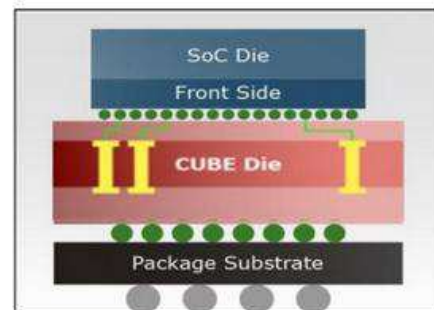
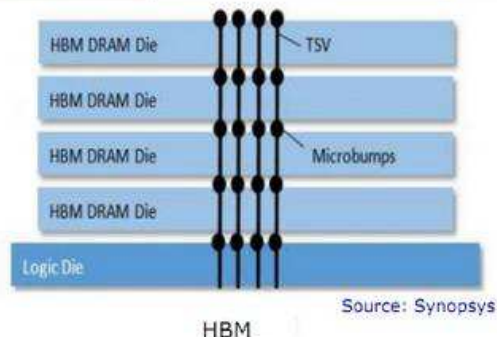


更多DRAM厂商正切入HBM赛道，国产HBM有望突破

- 二线、三线DRAM厂商也正在切入HBM赛道。华邦电于2023年8月介绍了其类HBM高带宽产品CUBEx，采用1~4层TSV DRAM堆叠，I/O速度500M~2Gbps，总带宽最高可达1024GB/s，颗粒容量为0.5~4GB，功耗低至不足1pJ/bit。这种比常规HBM拥有更高带宽的CUBEx可用于AR、VR、可穿戴等领域。
- 国产DRAM厂商有望突破HBM。目前一线厂商DRAM制程在1alpha、1beta水平，国产DRAM制程在25~17nm水平，中国台湾DRAM制程在25~19nm水平，国内DRAM制程接近海外。且国内拥有先进封装技术资源和GPU客户资源，有强烈的国产化诉求，未来国产DRAM厂商有望突破HBM。

图表：华邦电的类HBM高带宽产品CUBEx

Item	HBM1	HBM2	HBM2E	HBM3	CUBEx
Stack height	4HI	8HI	8HI/12HI	8HI/12HI/16HI	1 - 4HI
I/O Transfer rates	1 Gbps	2.4 Gbps	3.6 Gbps	6.4 Gbps	500M-2 Gbps
Overall package bandwidth	128 GB/s	256 GB/s	460 GB/s	819 GB/s	Up to 1024 GB/s
Package capacity	4GB	8GB	16GB/24GB	16GB/24GB/32GB	0.5 - 4GB
I/O number(R+W)	1024	1024	1024	1024	256-1024
Power / bit	6pJ/ bit	<5pJ/ bit	<5pJ/ bit	<4pJ/ bit	<1pJ/bit
Application	AI / Machine Learning in data center, HPC, accelerator system				SoC L4, AI NMC, AR/VR,...



一、HBM：算力的内存瓶颈

二、SK海力士HBM工艺分析：TSV、EMC、混合键合成趋势

三、市场测算：未来三年CAGR超80%

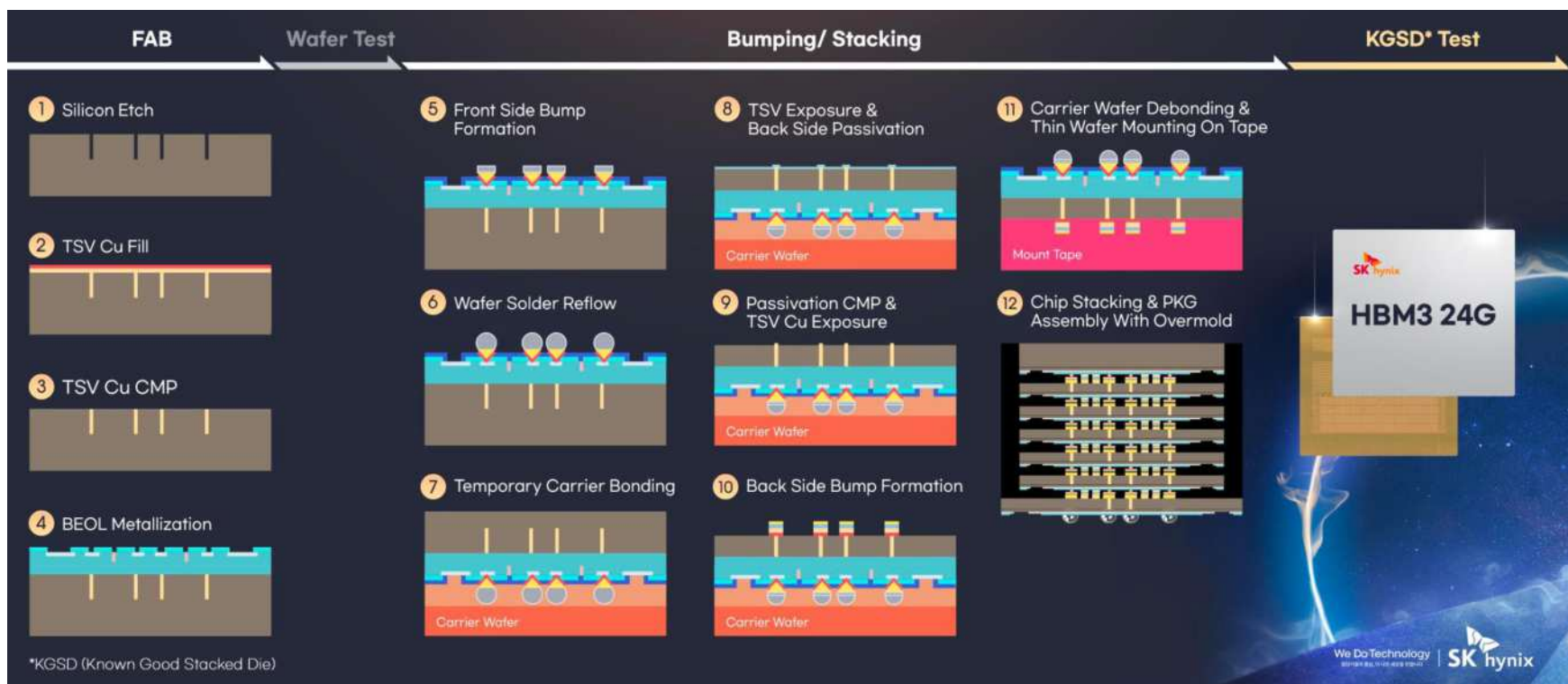
四、相关标的：设备、材料迎来新成长

五、风险提示

SK海力士HBM工艺分析

- SK海力士在HBM领域的成功归因于：布局较早、绑定大客户英伟达、先进的封装工艺。
- SK海力士的HBM采用定制的DRAM颗粒（TSV DRAM），核心工艺是MR-MUF工艺和EMC材料。

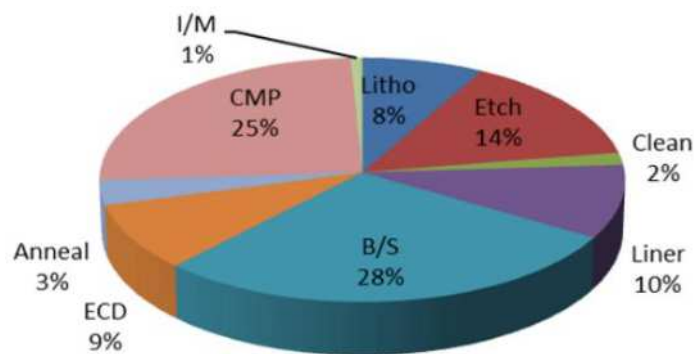
图表：SK海力士HBM生产工艺全流程



TSV DRAM

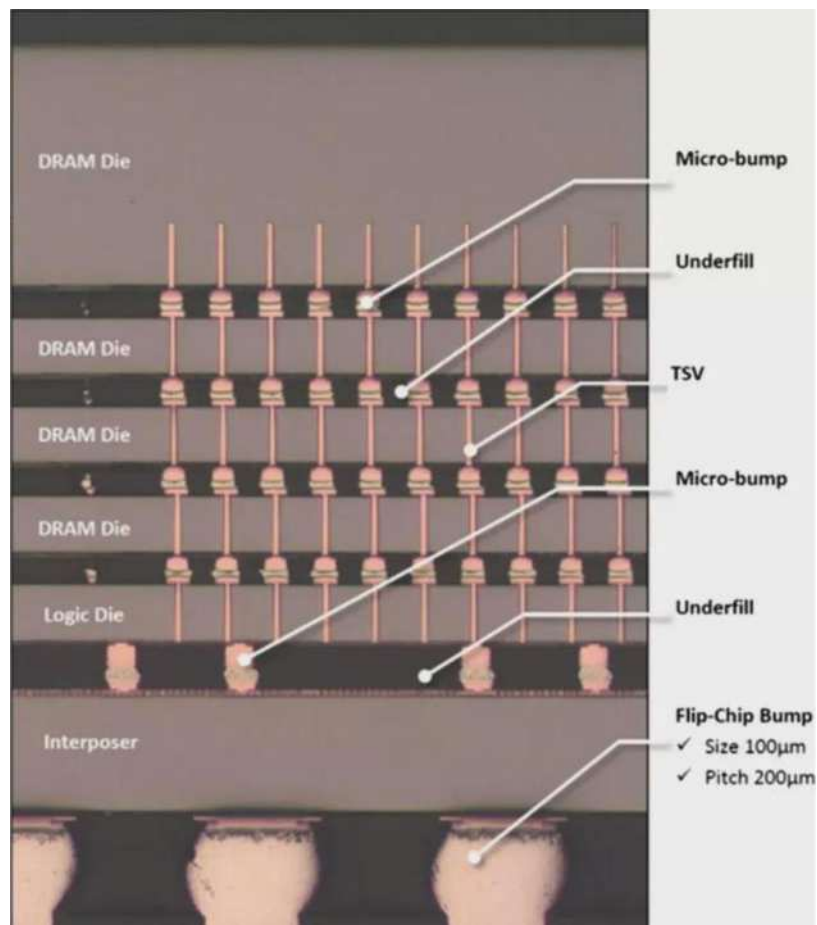
- **定制DRAM颗粒：**用于生产HBM的DRAM颗粒需要进行TSV打孔以便堆叠，因而需要在生产DRAM颗粒时，预留不含存储电路的位置用于TSV打孔。电路布局涉及存储厂商的商业机密，一般的封测厂商难以获得，也就难以参与HBM后道的生产，因此存储IDM在前道后道工艺磨合方面具备先天优势。
- TSV属于前道工艺，由存储厂商完成，主要用到刻蚀设备、薄膜沉积设备、电化学沉积设备ECD、CMP、清洗设备、退火设备、减薄机、划片机等，相关供应商包括：AMAT、TEL、Lam Research等。

图表：TSV的成本结构



数据来源：《Cost Analysis of TSV Process and Scaling Options》，
中信建投。B/S指Barrier and seed

图表：HBM中的TSV结构

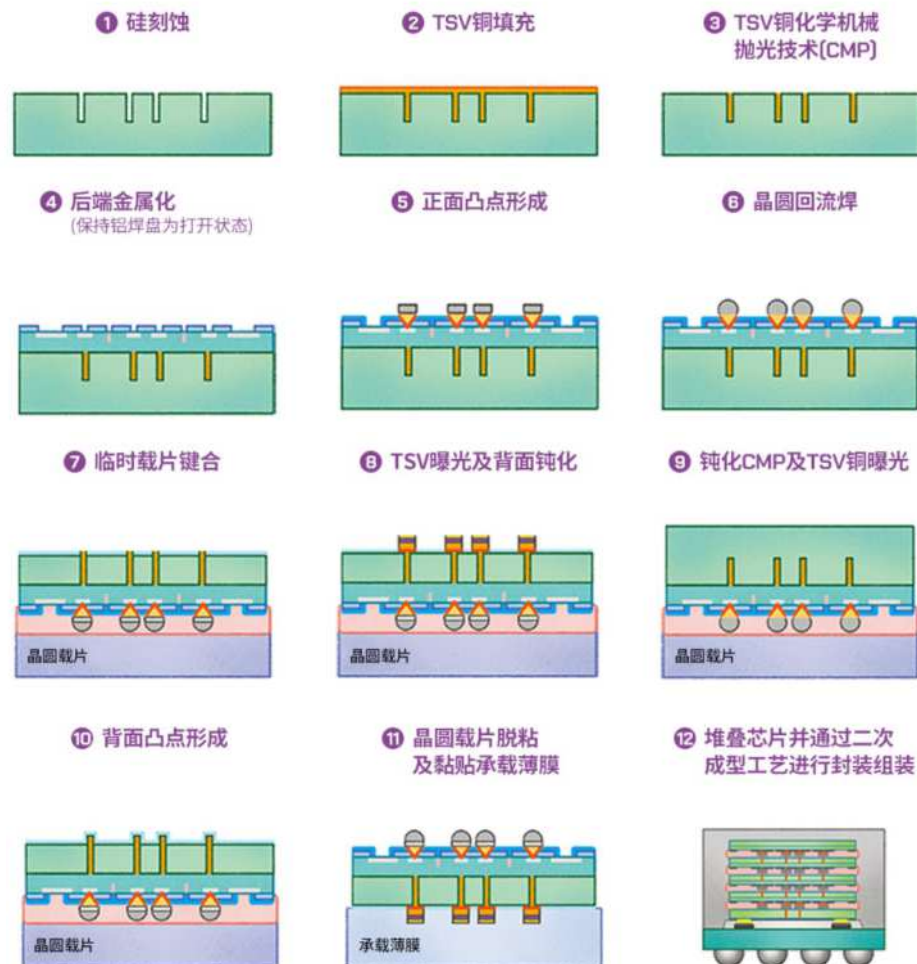


数据来源：半导体行业观察，中信建投

TSV DRAM

- TSV是目前唯一的垂直电互连技术，是实现2.5D/3D先进封装的关键技术之一。
- **1、高密度集成：**通过先进封装，可以大幅度地提高电子元器件集成度，减小封装的几何尺寸，和封装重量。
- **2、提高电性能：**由于TSV技术可以大幅度地缩短电互连的长度，从而可以很好地解决出现在SOC技术中的信号延迟等问题，提高电性能。
- **3、多种功能集成：**通过TSV互连的方式，可以把不同的功能芯片(如射频、内存、逻辑、数字和MEMS等)集成在一起实现电子元器件的多功能。
- **4、降低制造成本：**TSV三维集成技术虽然目前在工艺上的成本较高，但是可以在元器件的总体水平上降低制造成本。

图表：TSV的工艺流程

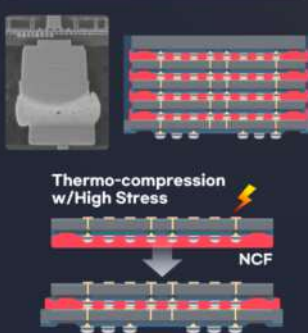
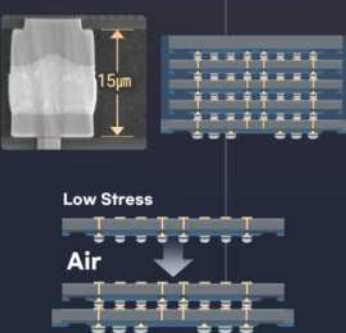
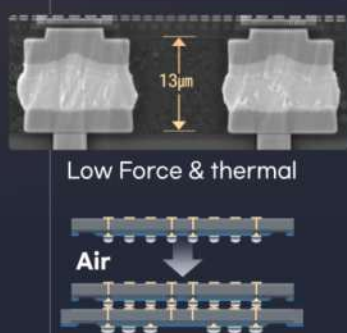
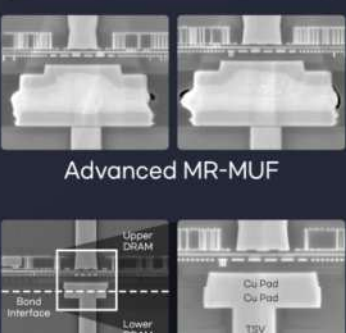


数据来源：TechSugar，中信建投

MR-MUF工艺

- SK海力士HBM领先三星电子和美光的关键在于MR-MUF工艺。HBM制造的关键在于每层TSV DRAM之间的连接方式，三星电子从HBM生产之初就一直采用热压缩非导电薄膜（TC-NCF）工艺，而SK海力士采用的则是批量回流模制底部填充（MR-MUF）工艺。
- 技术路线图显示，从HBM2e开始，海力士放弃了TC-NCF工艺，改用MR-MUF工艺，实现了更低的键合应力和更优的散热性能。海力士HBM3e将采用改进的MR-MUF工艺，降低键合应力，提升散热性能，增加堆叠层数。HBM4有可能采用混合键合（Hybrid bonding）工艺，进一步降低TSV DRAM层与层之间的间隙，实现更高层数堆叠。

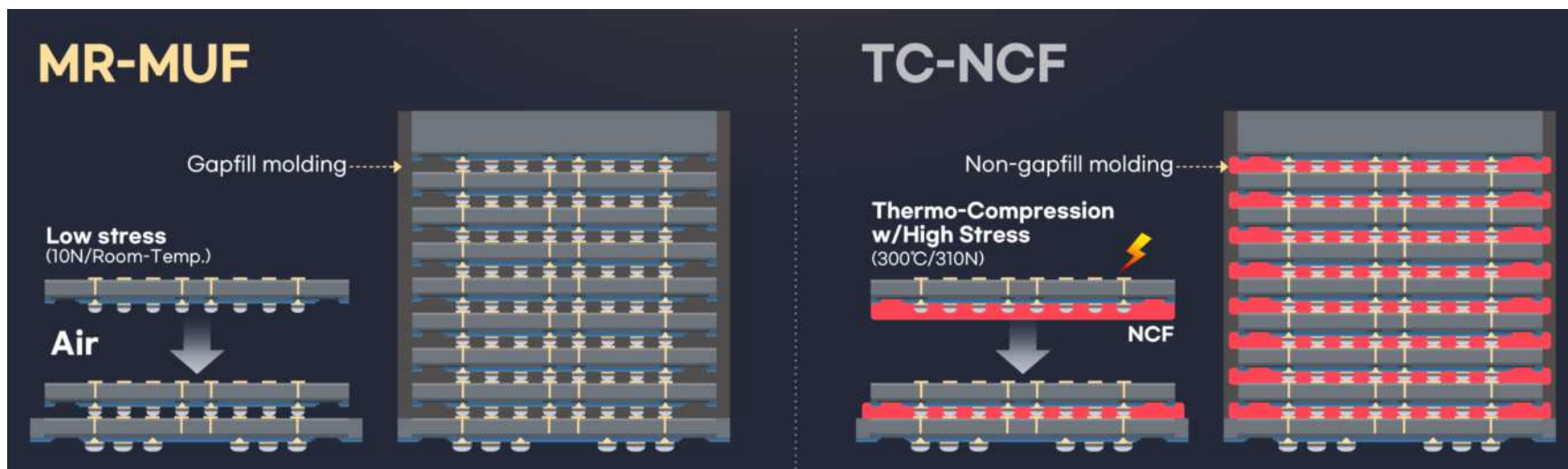
图表：SK海力士堆叠技术路线图

	HBM2	HBM2E	HBM3	HBM3 (12Hi) / HBM3E	HBM4
Stacking Tech.	TC-NCF	MR-MUF	MR-MUF	Advanced MR-MUF	TBD
Remark	 Thermo-compression w/High Stress NCF ✓ World 1st TSV chip stack	 Low Stress Air ✓ Low bond force & Robust joints: Higher Bump portion (thermal dissipation ↑)	 Low Force & thermal Air ✓ More Enhanced thermal dissipation: Lower gap height & thermal resistance ↓	 Advanced MR-MUF Hybrid Bonding Upper DRAM Lower DRAM Bond Interface Cu Pad TSV	
Achievable Stack Height	4Hi / 8Hi	4Hi / 8Hi	4Hi / 8Hi	8Hi / 12Hi	12Hi / 16Hi
Thermal R (Relative)	○ (1.0)	○ (0.65)	○ (0.55)	⊙ (0.5)	⊙ (0.4 ~ 0.5)

MR-MUF工艺

- MR-MUF全称为批量回流模制底部填充（Mass reflow molded underfill），是将多个芯片放置在下层基板上，通过回流焊一次性粘合，然后用模塑料填充芯片之间或芯片与基板之间间隙的方法，该方法主要用于FC封装和TSV芯片堆叠。
- 工艺流程：（1）在DRAM层与层之间，微凸点接触助焊剂，并进行横向引线键合层叠；（2）一次熔融全部的微凸点和助焊剂，并室温下施加压力，微凸点冷后即完成芯片和线路连接；（3）选用环氧树脂模塑料（EMC）填充芯片之间或基材之间的缝隙，同步进行绝缘和成型。该加工工艺的核心是处理层叠芯片中产生的热胀冷缩难题，及其处理芯片中间部位的空隙填充。
- 相比于TC-NCF，（1）MR-MUF不依靠高温和高压，减少了芯片键合产生的应力；（2）环氧树脂模塑料的导热系数是非导电薄膜的2~4倍，散热效率更高。

图表：海力士MR-MUF工艺和三星TC-NCF工艺比较



填充料EMC

- MR-MUF工艺的关键是：（1）晶圆翘曲度的控制（Chip warpage control）；（2）填充料的选取（Gapfill-MUF material）。
- 晶圆翘曲度：受加工条件（温度、压力等）影响，用量测设备检查，供应商有Camtek、ONTO。
- 填充料：EMC是先进封装常用的填充料，传统DDR或者Flash用的EMC一般是粉末状或圆柱体，而HBM中Die与Die之间的间距不足20微米（HBM2e 15微米，HBM3e 13微米），且有微凸点阻挡，粉末或者六面体流不进去。目前HBM主要使用GMC（Granular Molding Compound）或者LMC（Liquid Molding Compound）。目前海力士的GMC主要由日本Namics供应。

图表：MR-MUF工艺的微凸点键合和EMC填充流程

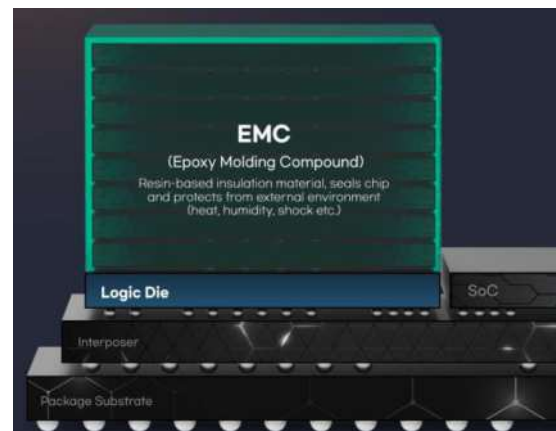
- 芯片堆叠，添加助焊剂



- 室温加压，完成芯片、线路连接



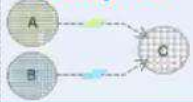
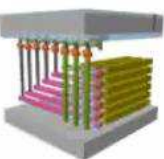
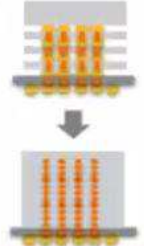
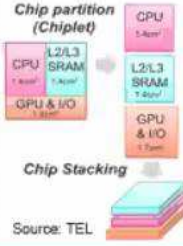
- EMC填充间隙



混合键合Hybrid Bonding

- 混合键合是将两片以上不相同的晶圆通过金属互连的工艺，不使用芯片堆叠之间的微凸点（球形），而是用铜与铜（片状，copper pad）互连，极大提升了芯片之间的信号传输速率。混合键合并非HBM专用的技术，CIS、3D NAND、逻辑芯片等均已大规模使用混合键合工艺。
- 技术优势：（1）允许无焊料键合，减少键合层厚度、缩短电气路径，并降低电阻；（2）通过直接将铜与铜接合，显著减小凸块间距。目前使用焊料很难实现10微米或更小的间距，而铜对铜直接键合可以将间距减小到不足1微米，提高芯片设计的灵活性；（3）拥有绝佳的散热性能；（4）极薄的粘合层和小间距会影响封装的形状因数，可以大大减小封装尺寸。

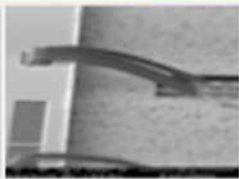
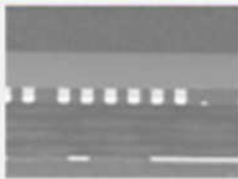
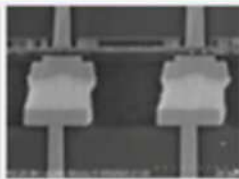
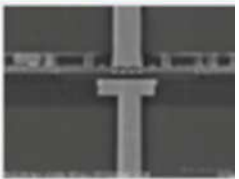
图表：混合键合工艺在各类芯片中的使用

Application	CIS	NAND	DRAM	Logic		Advanced Package	
	BSI	3D NAND	3D DRAM	Backside PDN	Sequential CFET	HBM	Disaggregation / Chiplets
Stacking Device	Sensor + Memory + Logic	Cell + Peripheral	Cell + Peripheral	Logic + Bare Si	Logic + Logic	DRAM + DRAM + Logic	
Bonding	W-W Cu Hybrid	W-W Cu Hybrid	W-W Cu Hybrid	W-W Ox Fusion	W-W Ox Fusion	D-W u Bump → Cu Hybrid	D-W / D-D Cu Hybrid
3D I/O Pitch	3um →1um	1um →0.5um	1um →0.5um	Sub um (nTSV)	Sub um (nTSV)	40um →25um	10um →1um
Structure			 Source: TEL 3D DRAM structure				 Source: TEL
Status	HVM	R&D~HVM	R&D	R&D	R&D	R&D	R&D~HVM

混合键合Hybrid Bonding

- 传统DRAM使用wire bonding、Flip Chip等工艺进行电气互连，I/O数量为4/8/16个，触点密度不足200，互连长度50-2000微米，可堆叠至多8层；而TSV和Hybrid的I/O数量多达1024个，触点密度可达3.5~10k，互连长度低至5~20微米，堆叠层数可达16层甚至更高。因此，用Hybrid和TSV工艺生产的DRAM拥有高带宽、高存储密度，SK海力士预计HBM4开始使用Hybrid bonding，三星电子、美光也可能转向该技术。
- 混合键合设备的供应商主要有Besi、EV Group，其中Besi擅长Die to Wafer bonding，EV Group擅长Wafer to Wafer bonding。

图表：DRAM的不同互连工艺

Interconnection	Wire	Flip Chip	TSV	Hybrid ³
Image				
Memory Application	DRAM (Mobile), NAND	DRAM (Computing, Graphics)	DRAM (HPC/Server, Graphics)	DRAM (HPC/Server, Graphics)
Number of I/Os	× 4, × 8, × 16	× 4, × 8, × 16	× 1,024	× 1,024
Number of Physical Interconnections	50-150 ea.	150-200 ea.	Bottom die: 5-8 K ea. Upper die: 3.5-10 K ea. (Up to 200 K ea. with dummy bumps)	Bottom die: 5-8 K ea. Upper die: 3.5-10 K ea. (Up to 200 K ea. with dummy bumps)
Interconnection Length	200 μm-2,000 μm	50 μm	20 μm	5 μm
Number of Stacks	2 / 4 / 6 / 8 (DRAM)	1 / 2 (Planar)	4 / 8 / 12	4 / 8 / 12 / 16
Max. Capacity / PKG	16 GB	4 GB	24 GB	32 GB

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/128130066112006044>