中泰证券
ZHONGTAI SECURITIES

证券研究报告
报告日期：2025年2月20日

【中泰电子】AI系列之存储：近存计算3D DRAM，AI应用星辰大海

分析师：

王芳 S0740521120002

杨旭 S0740521120001

中泰证券研究所
专业 | 领先 | 深度 | 诚信

目 录

目 录

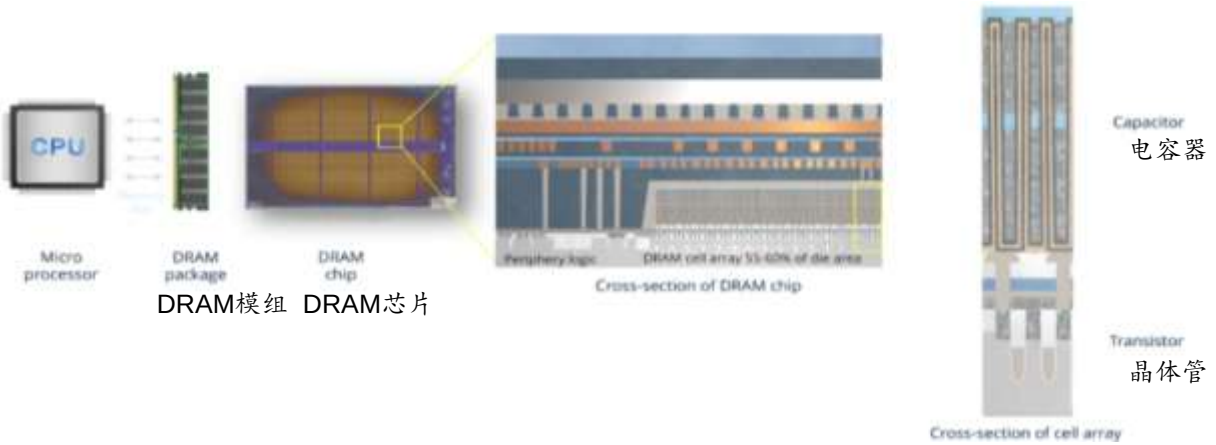
1.1 2D DRAM制程瓶颈凸显，3D是大趋势

■ **DRAM是易失性存储器，与CPU/GPU等计算芯片直接交互，可以快速存储每秒执行数十亿次计算所需的信息。**

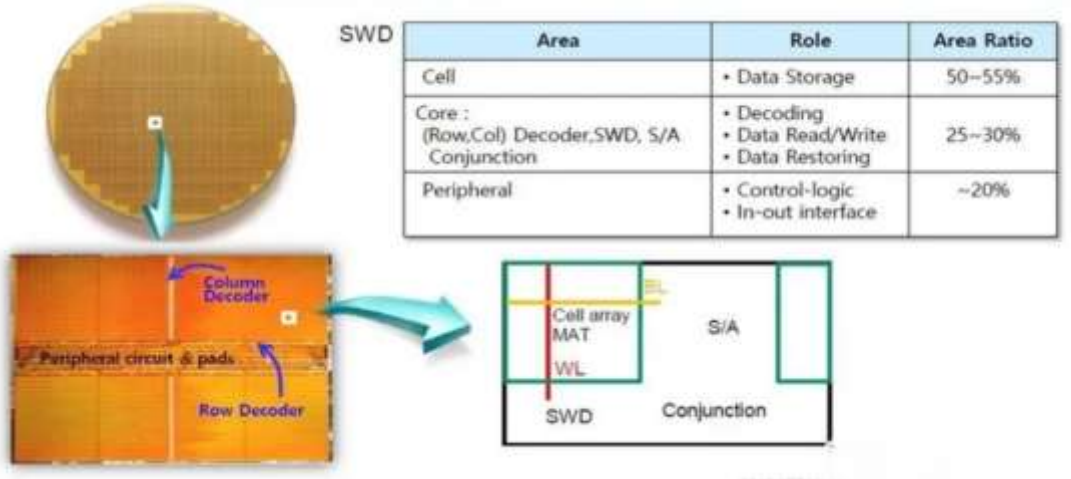
➢ **DRAM三构成：**1）存储单元（Cell），占据50%-55%面积：存储单元是DRAM芯片存储数据的最小单元，每个单元存储1bit数据（二进制0或1），单颗DRAM芯片的容量拓展主要是通过增加存储单元的数量实现（即提高单位面积下的存储单元密度），存储单元基本占据了DRAM芯片50-55%的面积，是DRAM芯片最核心的组成部分。1个存储单元由1个晶体管和1个电容器构成（1T1C结构），晶体管控制对存储单元的访问，电容器存储电荷来表示二进制0或1。2）外围逻辑电路（Core），占据25-30%面积：由逻辑晶体管和连接 DRAM 各个部分的线路组成，从存储单元中选择所需存储单元，并读取、写入数据，包括感应放大器（ Sense Amplifiers ）和字线解码器（Word Line Decoders）等结构，如感应放大器被附加在每个位线的末端，检测从存储单元读取非常小的电荷，并将信号放大信号，强化后的信号可在系统其他地方读取为二进制1或0。3）周边线路（Peripheral），占据20%左右面积：由控制线路和输出线路构成。控制线路主要根据外部输入的指令、地址，让DRAM内部工作。输出/输入线路负责数据的输入（写入）、输出（读取）。

➢ **DRAM工作原理：**存储电容器会泄漏电荷，因此需要频繁进行刷新（大约每 32 毫秒一次），以维持存储的数据。每次刷新都会读取存储单元的内容，将位线上的电压提升至理想水平，并让刷新后的值流回电容器，刷新完全在 DRAM 芯片内部进行，没有数据流入或流出芯片。这虽最大限度地减少了浪费的电量，但刷新仍会占据 DRAM 总功耗的 10% 以上。

图表：DRAM结构图



图表：DRAM三个构成的占比



Area	Role	Area Ratio
Cell	• Data Storage	50~55%
Core : (Row,Col) Decoder,SWD, S/A Conjunction	• Decoding • Data Read/Write • Data Restoring	25~30%
Peripheral	• Control-logic • In-out interface	~20%

来源：AMAT官网、Semianalysis、36氪，中泰证券研究所

1.1 2D DRAM制程瓶颈凸显，3D是大趋势

■ **容量、带宽和功耗是DRAM三大关键参数。**

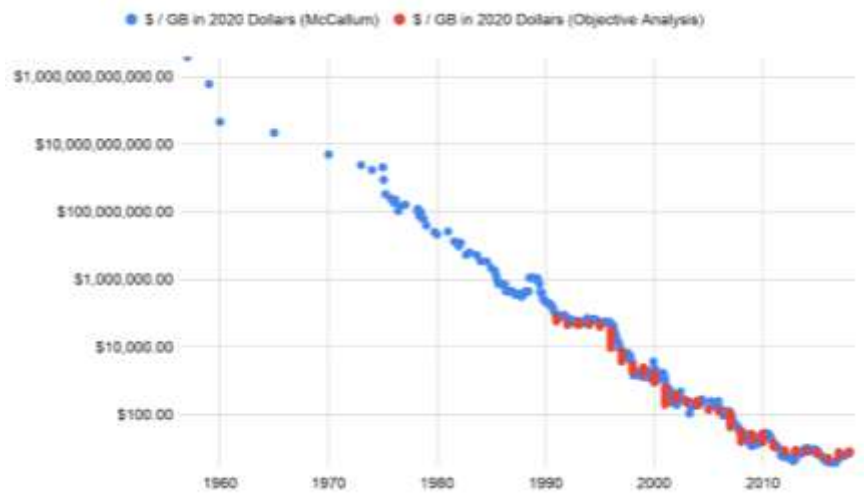
➢ 1）容量：指存储数据的多少，存储容量最小单位是1bit，即表示存储单个二进制（0或1），另外有B、KB、MB、GB、TB等存储容量单位，关系如下：1B（Byte，B）= 8bit，1KB=1024B，1MB = 1024KB，1GB = 1024MB，1TB = 1024GB。单位面积下，存储单元数量越多、存储容量越高，制程是决定单位面积下存储容量的主导因素。

➢ 2）带宽：指每秒钟的数据吞吐量，单位TB/s、GB/s，内存带宽 = 最大时钟速频率 (MHz) × 总线宽度 (bits) ×每时钟数据段数量/ 8。

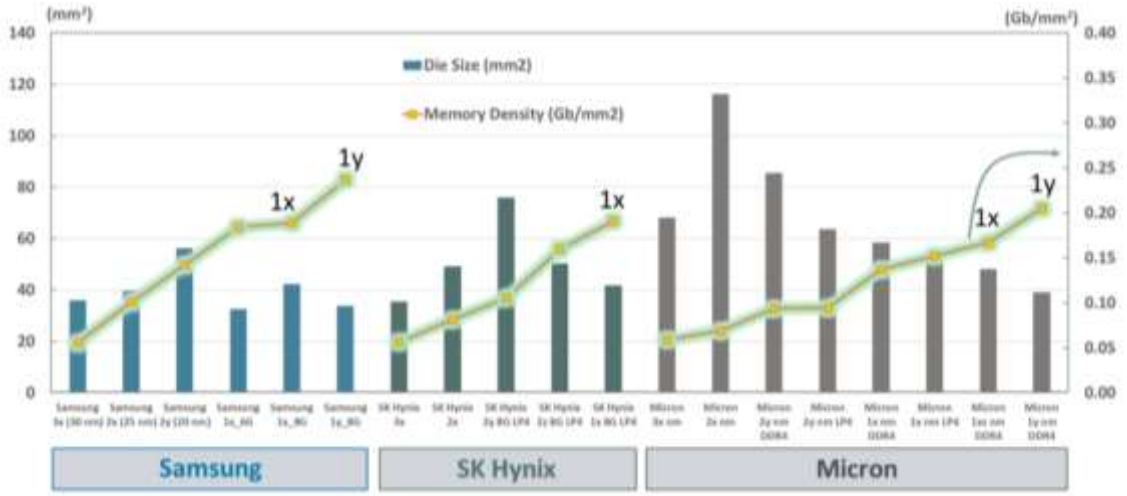
➢ 3）功耗：数据的传输需要的功耗，功耗越低越好。

■ **DRAM制程微缩，带来DRAM成本下降和容量密度提升。**

图表：DRAM单位容量价格处于下降趋势



图表：DRAM通过制程迭代提升容量密度



来源：Techinsights、AI IMPACTS，中泰证券研究所

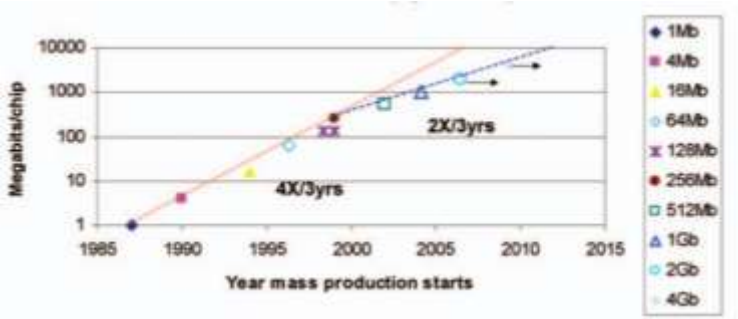
1.1 2D DRAM制程瓶颈凸显，3D是大趋势

■ **DRAM此前符合摩尔定律，后面摩尔定律失效，制程微缩放缓。**

➢ DRAM通过制程微缩（晶体管、电容器、逻辑电路等微缩）实现单位面积内更多的存储单元，即实现单位面积下更高存储容量。

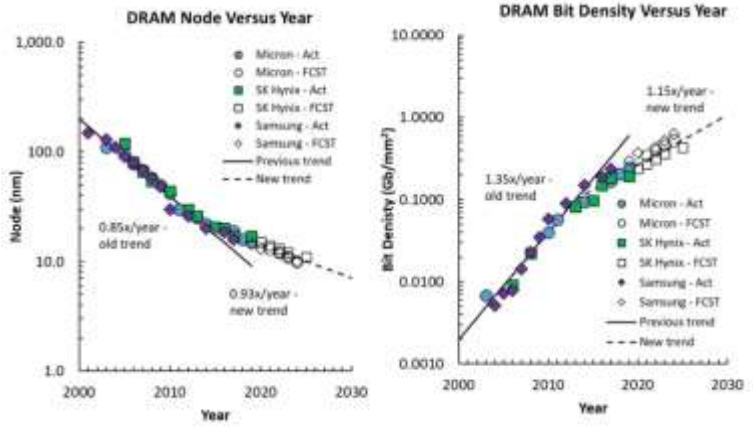
➢ 1970-2005年，DRAM以每颗芯片的容量每3年增加4倍的速度升级，后续迭代速度不断放缓，带来单位密度提升速度放缓，存储单元微缩放缓。

图表：DRAM总位元出货量/DRAM芯片出货量

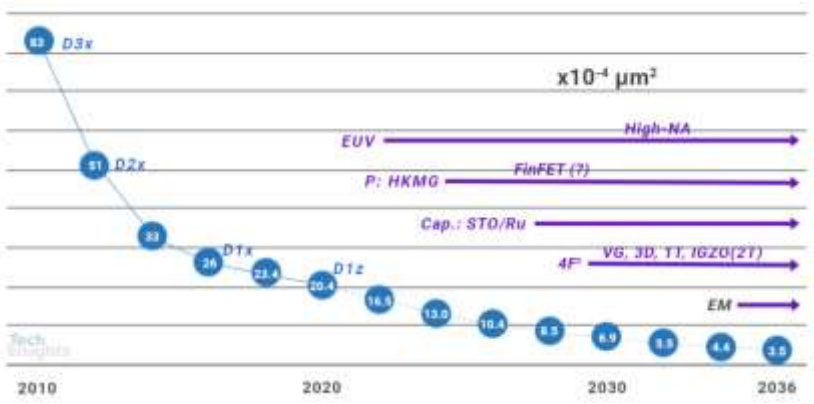


图表：DRAM容量升级的速率放缓

图表：2D DRAM的制程微缩和单位密度提升速度放缓



图表：DRAM 存储单元面积（Cell Szie）微缩放缓



来源：IC Knowledge、IEEE、Techinsights，中泰证券研究所

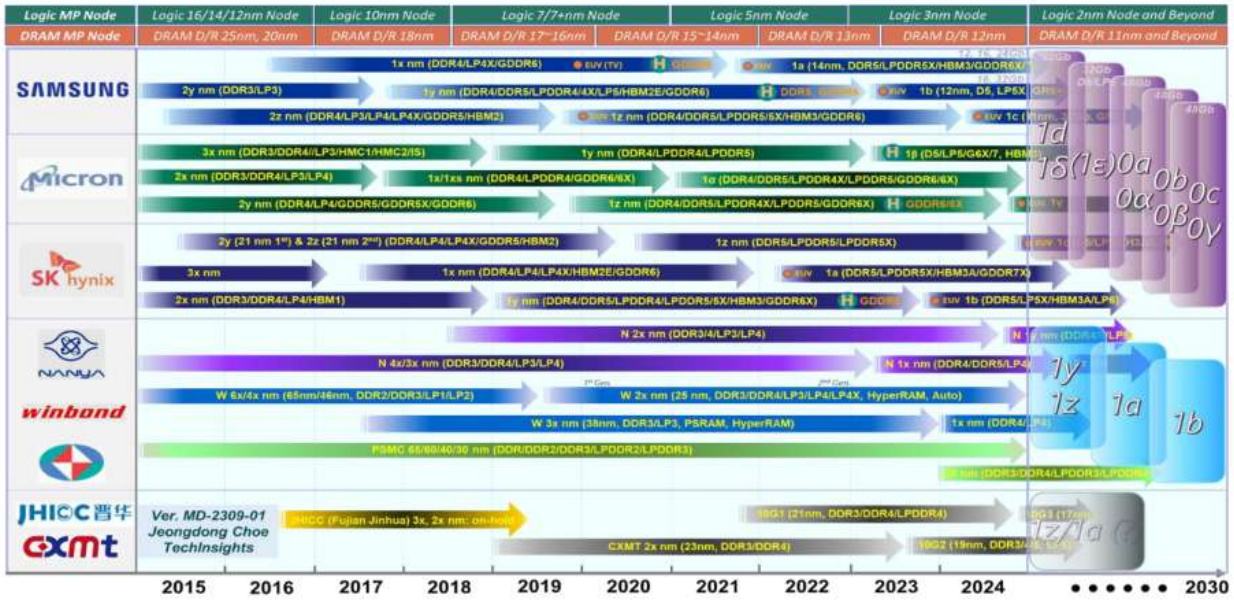
1.1 2D DRAM制程瓶颈凸显，3D是大趋势

■ **DRAM制程微缩难度大，目前制程迭代逼近10nm（1γnm），必须使用EUV光刻机。**

➢ 目前DRAM最新量产制程是1b，10-12nm左右：DRAM制程迭代速度放缓，10nm级别（10-20nm），使用1x、1y、1z、1a、1b和1c指代，另外美光使用罗马字母1α、1β、1γ对应1a、1b和1c。目前三星、海力士和美光三大家目前量产制程是1b（1β）制程，近两年将开始迭代1c（1γ）制程。

➢ EUV的使用：EUV是目前光刻机的天花板，2020年三星在1z节点开始首次使用EUV光刻机，后续的制程沿用EUV，2021年海力士在1a节点开始使用EUV光刻机，后续制程继续沿用，美光在1c（1γ）节点将使用EUV。

图表：DRAM制程迭代



图表：三大家EUV光刻机使用情况

节点		三星	海力士	美光
1x	16-19nm	Test vehicle		
1y	14-16nm			
1z	12-14nm	√		
1a（1α）	约13nm	√	√	
1b（1β）	10-12nm	√	√	
1c（1γ）	约10nm，1β的增强版	√		√

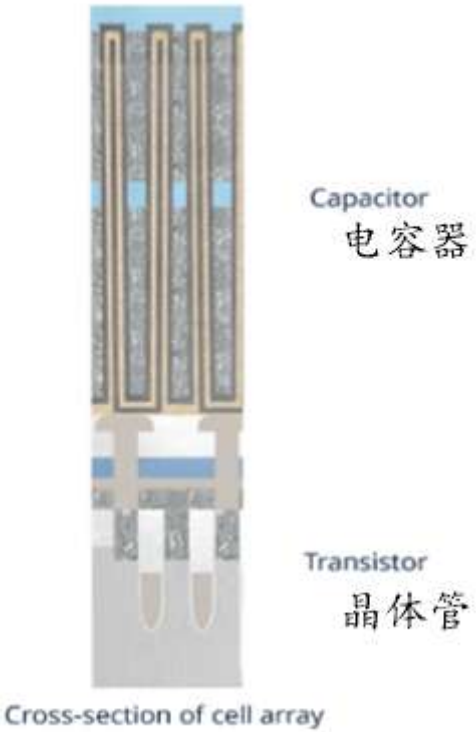
来源：Techinsights，Semiconductor degest，中泰证券研究所

1.1 2D DRAM制程瓶颈凸显，3D是大趋势

■ **DRAM制程微缩难度：微缩电容器和感应放大器面临挑战。**

- 三星在1z、海力士在1a工艺中采用了极紫外光刻（EUV），也未能显著提升密度。它们面临的主要挑战在于电容器与感应放大器。
- 1）电容器：
  - 电容器微缩，电容漏电风险、干扰问题变严重。DRAM依赖电容器来存储电荷，但当电容器变得更小，电荷泄漏的风险增加，从而导致数据的可靠性下降。为了解决这个问题，工程师们需要开发新的材料和设计方法，以减少漏电率并提高数据保持能力。另一个重大挑战是干扰问题。在高集成度的芯片上，不同存储单元之间的电场和磁场干扰变得更加频繁，这可能导致数据错误或损坏。为了应对这一问题，需要更加复杂的错误校正机制和抗干扰设计，这进一步增加了DRAM开发的难度。
  - 电容器制作难度极大。首先，电容器的图案化要求非常高，因为孔必须紧密排列，且具有极为良好的临界尺寸和覆盖控制，以便接触下方的访问晶体管并避免出现桥接或其他缺陷。电容器与晶体管极为相似，已缩小至纳米级宽度，不过其纵横比也非常大，大约 1000 纳米高，而直径却只有数十纳米——纵横比接近 100:1，因此蚀刻出又直又窄的孔轮廓极为困难。此外，还需要更厚的硬掩模来实现更深的蚀刻，因为更厚的掩模需要更厚的光刻胶。接下来，必须在整个孔轮廓的壁上沉积几纳米厚的多个无缺陷层，以形成电容器。另外电容器即使微缩，电容器也需要存储一定量的电荷，如果电荷过少，“1”和“0”的区别就会变得模糊，会对存储功能产生影响。
- 2）感应放大器：必须进行面积缩放以匹配位线的缩小，感应放大器变得更不敏感，并且随着尺寸变小而更容易出现变化和泄漏。同时，较小的电容器存储的电荷较少，读取变得更加困难。

图表：DRAM存储单元结构



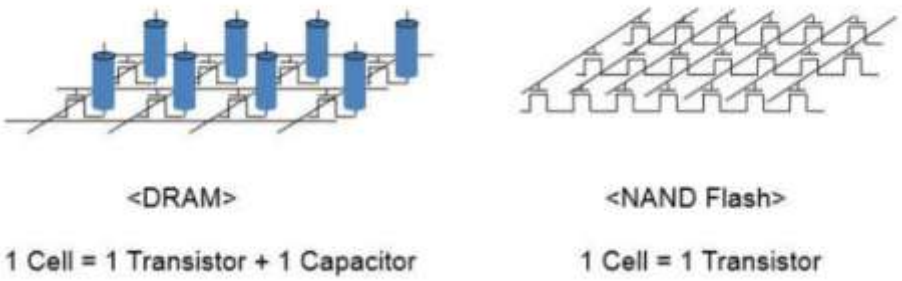
来源：semianalysis、AMAT官网，中泰证券研究所

1.1 2D DRAM制程瓶颈凸显，3D是大趋势

■ **NAND存储单元结构简单，已率先实现晶圆级3D结构，通过层数堆叠来提升容量。**

- NAND存储单元仅由一个晶体管构成，相对DRAM、结构简单。NAND从2014年开始进行晶圆级层面的从2D到3D的转换，成功解决了2D NAND在增加容量的同时性能降低的问题，实现容量、速度、能效及可靠性等全方位提升。NAND的2D平面制程微缩基本停留在2017年的14-15nm左右，后续的迭代升级是层数堆叠。
- 2019年，3D NAND的渗透率为72.6%，已远超2D NAND，预计2025年3D NAND将占闪存总市场的97.5%。2024年11月21日海力士宣布321层NAND样品，自2025年上半年开始交货，此前海力士量产产品为238层。
- **DRAM存储单元包含垂直方向的电容器，制程微缩难度高于NAND，同时晶圆级3D需要存储单元结构创新，难度大。**
- DRAM存储单元由1个晶体管和1个电容器构成，比NAND的存储单元结构更复杂，电容器增加了制程微缩难度，因此在2D NAND还在通过制程微缩时，DRAM的制程就落后于NAND，如2015年2D NAND进入17-18nm，而DRAM在20-30nm。
- DRAM具有较大的垂直方向电容器，电容器很高且难以分层堆叠，因此需要采用将电容器水平放置等创新的存储单元结构或者采用无电容DRAM来实现晶圆级3D，制造难度大幅提升。

图表：DRAM和NAND的存储单元结构



图表：DRAM、NAND和Logic的制程迭代

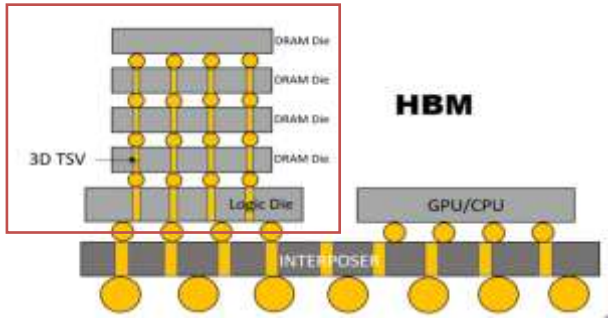
来源：三星、ASML 2018 Presentation，中泰证券研究所

1.1 2D DRAM制程瓶颈凸显，3D是大趋势

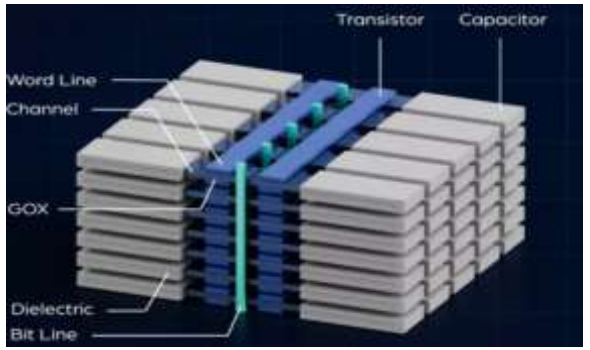
■ **DRAM 3D化趋势已现，封装级先行，晶圆级在研发阶段。**

- 3D DRAM分为封装级和晶圆级，封装级3D DRAM属于近存计算，突破内存墙瓶颈，已商业化量产，晶圆级3D DRAM突破2D DRAM制程微缩瓶颈，难度更大，目前仍处于研发阶段。
- 封装级3D DRAM：指通过封装工艺将多颗2D DRAM Die进行3D堆叠，HBM目前最高堆叠12层DRAM Die，每层Die之间通过TSV/Microbump等先进封装工艺实现电气连接，最后实现在单位面积下更高的存储容量密度。然后将封装级3D DRAM继续通过封装工艺与逻辑芯片封装在一起，实现近存计算，性能上实现更高的带宽、更低的功耗，缓解内存墙问题，契合AI芯片要求。典型产品如HBM、华邦CUBE和WoW 3D堆叠DRAM。
- 晶圆级3D DRAM：在晶圆结构层面实现3D结构，突破2D DRAM制程微缩瓶颈、实现更高容量密度，目前各家厂家处于探索阶段。

图表：封装级3D DRAM：HBM结构图



图表：晶圆级3D DRAM结构图



图表：封装级别3D DRAM的分类

		与计算芯片的封装形式	芯片之间的连接
封装级3D DRAM	HBM	2.5D	TSV+Microbump
	CUBE	3D	TSV+Microbump
	WOW 3D堆叠DRAM	3D	TSV+混合键合

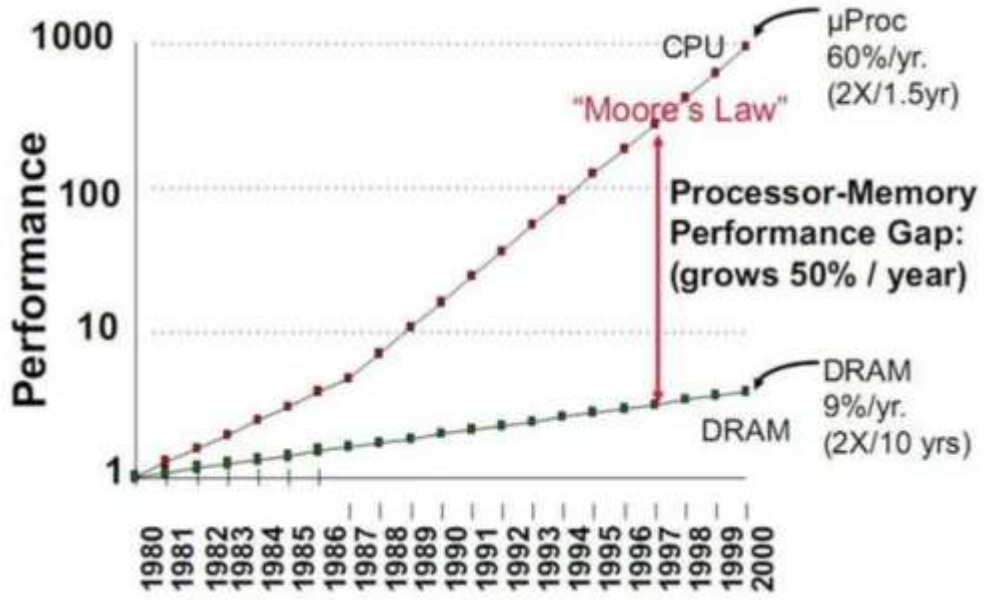
来源：NEO，中泰证券研究所

1.2存内计算难度大，近存计算和存内处理是重要方向

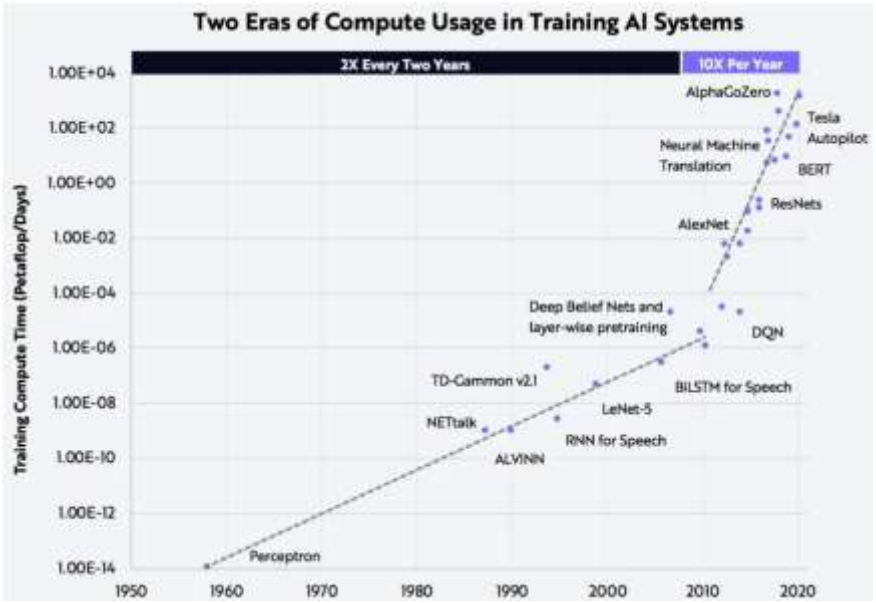
■ **存储速度滞后于计算器速度，AI时代存储带宽制约算力芯片性能发挥。**

- 在过去二十年，处理器性能以每年大约60%的速度提升，内存性能的提升速度每年只有9%左右。结果长期下来，不均衡的发展速度造成了当前的存储速度严重滞后于处理器的计算速度。
- 虽然多核（例如CPU）/众核（例如GPU）并行加速技术提升算力，AI时代处理器计算技术能力大幅提升，同时大型 Transformer 模型的参数数量呈指数级增长，每两年增加 410 倍，而单个 GPU 内存仅以每两年 2 倍的速度扩展。从峰值算力看，峰值算力在过去 20 年中增加了 60000 倍，而 DRAM带宽增加了100 倍，存储和计算的互连带宽增加了 30 倍。
- 随着近几年云计算和AI应用发展，面对计算中心的数据洪流，存算分离架构下数据搬运慢、搬运能耗大等问题成为了计算的关键瓶颈，“存储墙”问题更加显著。

图表：处理器和存储器速度失衡



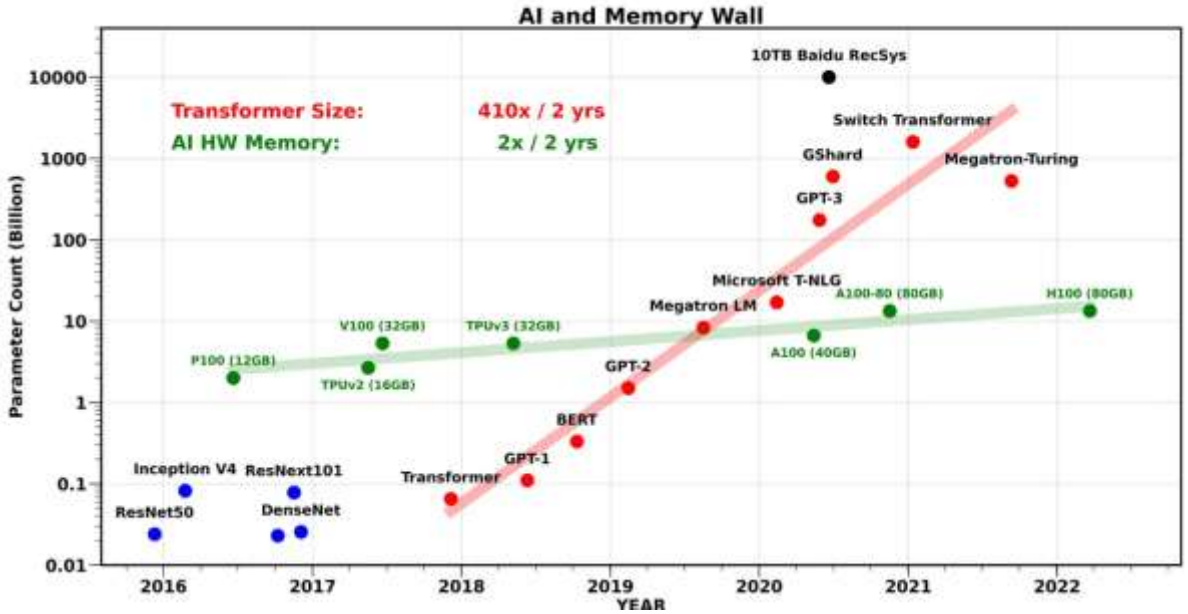
图表：1960～2020年人工智能计算复杂度变化



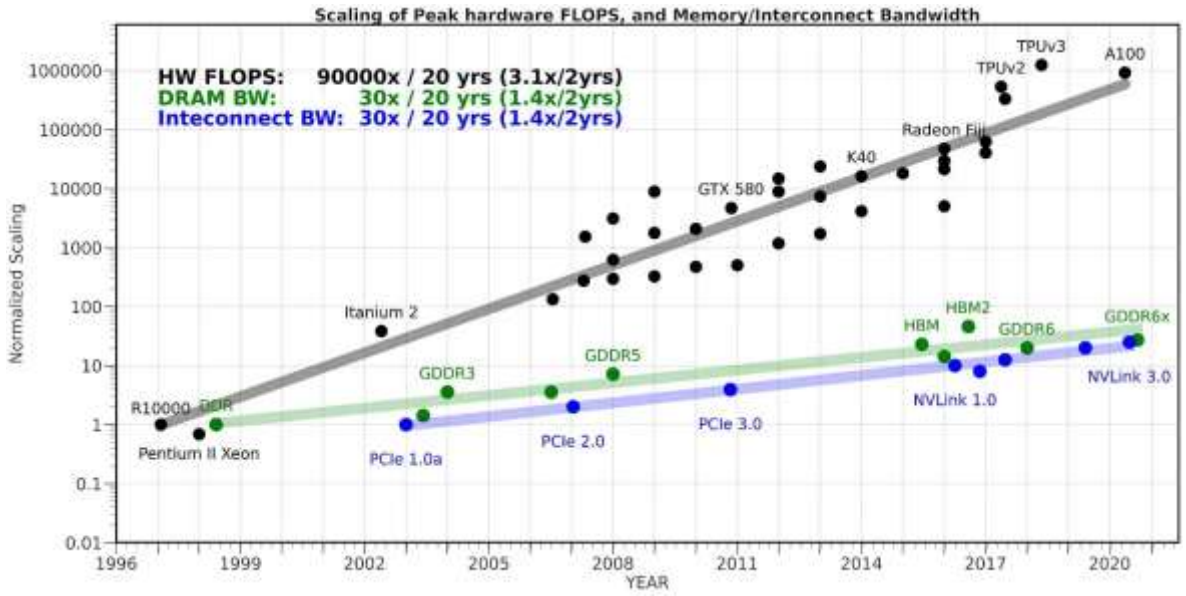
来源：elecfans，智东西，中泰证券研究所

1.2存内计算难度大，近存计算和存内处理是重要方向

图表：模型参数量增长趋势（红线）
VS 单GPU内存扩展趋势（绿线）



图表：不同代的内存带宽以及峰值算力



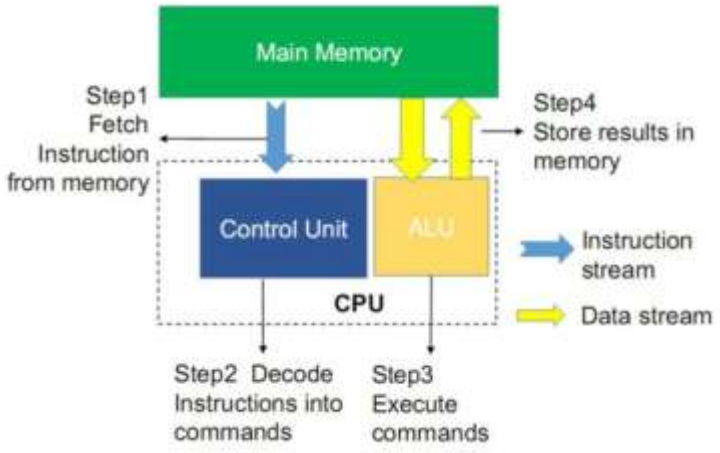
来源：Medium，中泰证券研究所

1.2存内计算难度大，近存计算和存内处理是重要方向

■ **传统存算分离架构带来存储墙问题。**

➢ 上世纪40年代开始计算机使用冯诺伊曼架构——存算分离，即处理器和存储器相互独立，两者通过总线连接。1）存算分离，数据存算间传输造成延迟。处理器从外部存储中调取数据，计算完成后再传输到内存中，一来一回都会造成延迟。2）数据在多级存储间传输。为了提升速度，冯诺依曼架构对存储进行分级，越往外的存储介质密度越大、速度越慢，越往内的存储密度越小，速度越快，因此数据需要在多级存储之间搬运，能耗大。通常第一级存储是速度最快、容量低，主要是SRAM片上缓存，第二级是传统DDR。3）存储制程推进慢于逻辑。目前DRAM制程最先进仍在10-15nm左右，而逻辑制程已进入3nm，主要是因存储器制程缩小难度更大。

图表：存算分离架构



图表：数据的传输速度慢

图表：数据的传输功耗大

Operation	Energy(pJ)
Computation Energy Cost	
Integer Add (32b)	0.1
Integer Multiply (32b)	3.1
Floating Point Add (32b)	0.9
Floating Point Multiply (32b)	3.7
Memory Access Energy Cost	
8KB SRAM (64b)	10
1MB SRAM (64b)	100
DRAM	2000

~650X

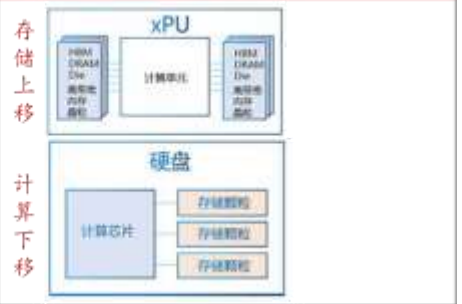
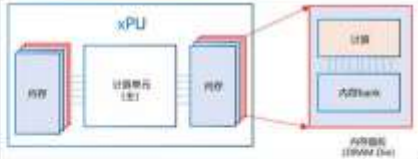
来源：知存科技等，中泰证券研究所

1.2存内计算难度大，近存计算和存内处理是重要方向

■ **存算一体可有效克服冯诺依曼架构，可有效提升带宽、缓解存储墙问题，迎合AI时代需求。**

- 存算一体是一种新的架构，其核心理念是将计算和存储融合，降低“存储墙”问题，实现计算能效的数量级提升。从广义而言，存算一体可分为三种：近存计算（PNM）、存内处理（PIM）、存内计算（CIM），狭义的存算一体主要指存内计算。目前近存计算和存内处理已开始商业化应用，但存内计算因设计等难度大，目前暂未商业化大规模使用。
- 近存计算：存算分离，通过封装拉近存储和计算单元的距离。
- 存内处理：在存储单元内加了部分计算单元，存储芯片有部分计算能力。
- 存内计算：真正的存算一体，存储单元和计算单位完全融合。

图表：存算一体三个类别

	类型	英文	简介	具体方法与原理	典型产品	示意图
广义的存算一体	近存计算	PNM, Processing Near Memory	通过芯片封装或板卡组装将存储芯片和计算芯片集成在一起	通过芯片封装和板卡组装，将存储单元和计算单元集成，增加访存带宽、减少数据“搬运”，提升计算效率。 又可细分为： →存储上移，用先进封装，使得存储器向处理器（CPU、GPU）靠近，增加计算和存储间的链路数量、增加带宽。 →计算下移，利用板卡集成技术，在存储设备引入计算引擎，承担如数据压缩、搜索、视频转码等本地处理，减少远端处理器负担。	存储上移：HBM（高带宽内存）；计算下移：CSD（可计算存储）；3D堆叠DRAM	
	存内处理	PIM, Processing In Memory	DRAM内集成计算单元	在芯片制造时，将存和算集成在同一颗晶粒（die）上，存储器本身即具备一定计算能力。比如在DRAM中内置处理单元，提供大吞吐低延迟片上处理能力。	HBM-PIM、PIM-DIMM	
狭义的存算一体	存内计算	CIM, Computing In Memory	存储和计算完全融合	芯片设计过程中，不再区分存储和计算单元，存储电路重新设计后同时具备存储和计算能力，达到计算能效数量级提升	存内计算（IMC）芯片	

来源：中国移动研究院，中泰证券研究所

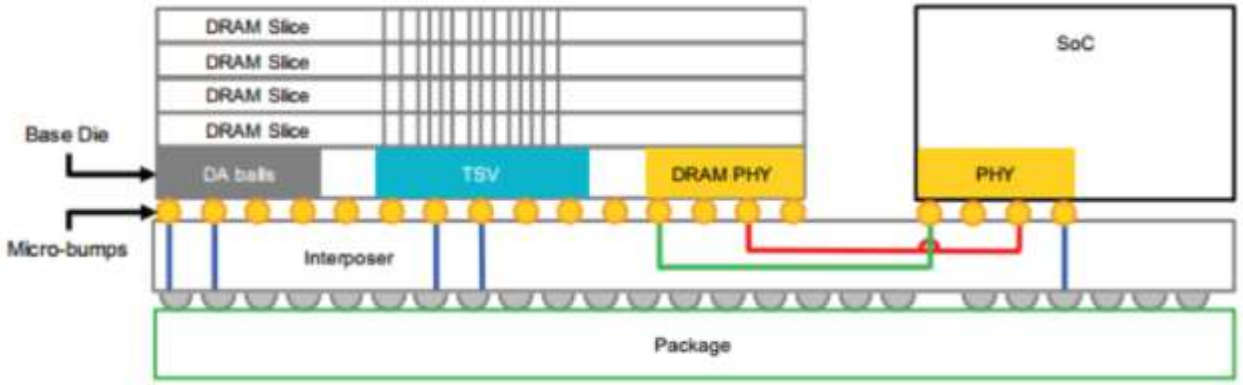
1.2存内计算难度大，近存计算和存内处理是重要方向

■ **近存计算：通过封装工艺拉近存储单元和计算单元距离，目前已大规模使用。**

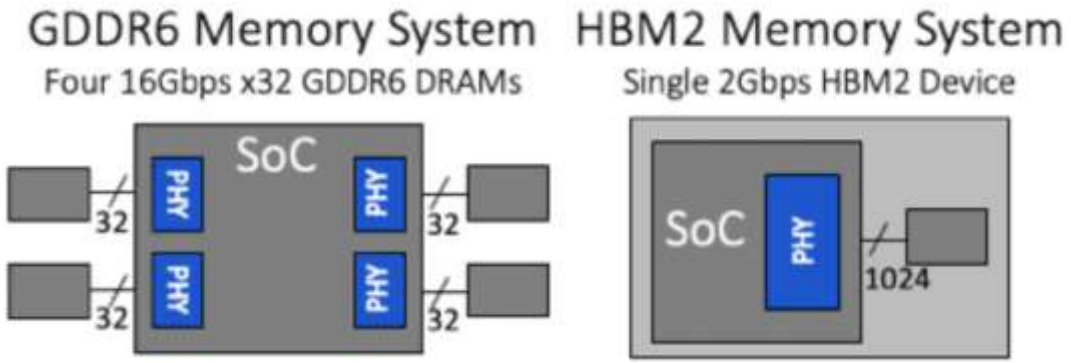
➢ 近存计算不改变计算单元和存储单元本身设计功能，通过采用先进的封装方式及合理的硬件布局和结构优化，通过芯片封装和板卡组装的方式，将存储和计算芯片封装在一起，使用系统级封装工艺，增加存储和计算芯片的信号连接通路，增强二者间带宽。近存计算本质上属于传统冯诺依曼的存算分离架构，通过拉近存储单元和计算单元的距离，对“存储墙”进行优化。

➢ 典型产品：HBM、3D堆叠DRAM和华邦CUBE产品均属于近存计算。

图表：HBM是近存计算



图表：HBM VS GDDR



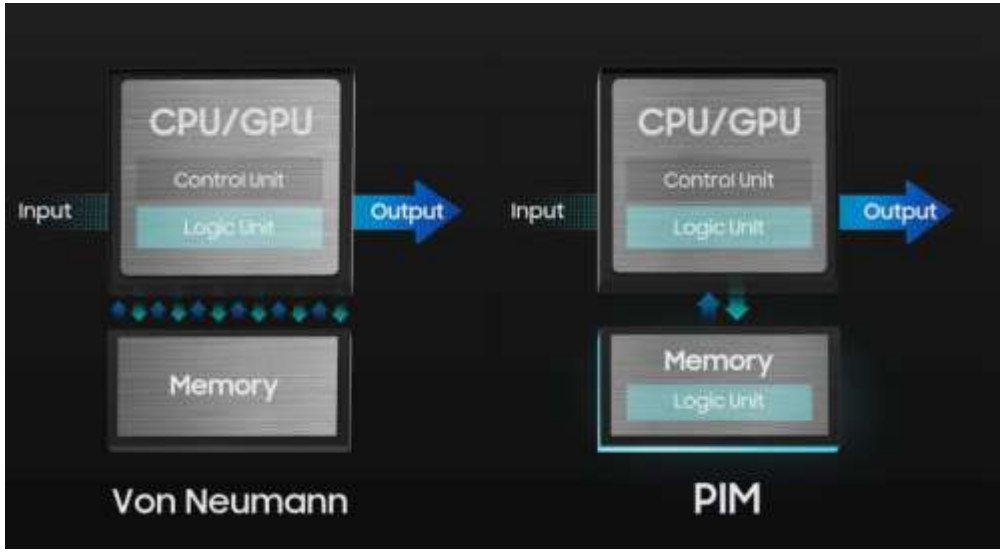
来源：Rambus，中泰证券研究所

1.2存内计算难度大，近存计算和存内处理是重要方向

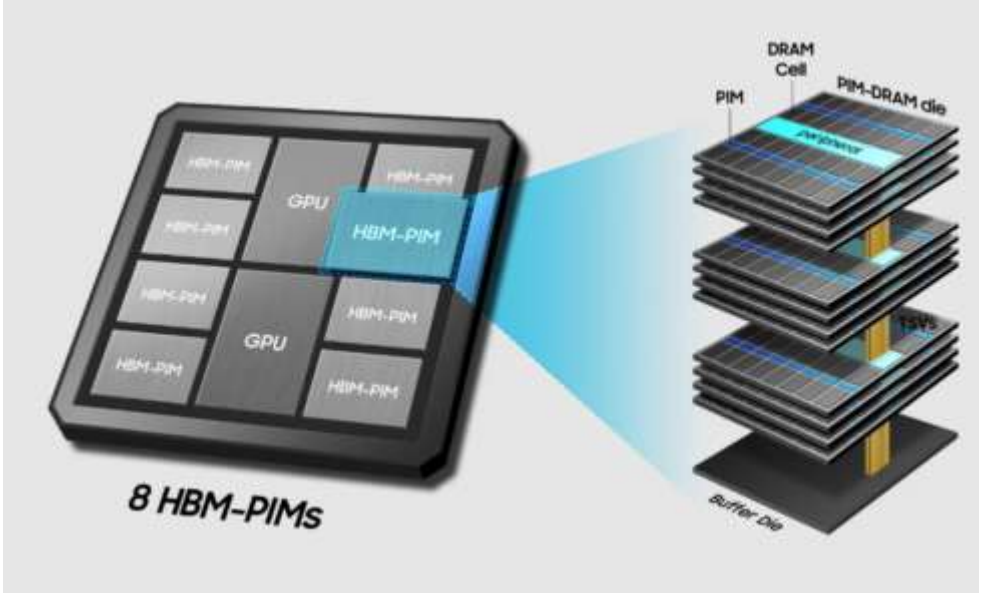
■ **存内处理：存储器具备一定计算能力，目前三星海力士已推出相关产品，但未大规模使用，LPDDR6-PIM新标准制定中。**

- 目前的存内处理方案主要通过在内存（DRAM）芯片中实现部分数据处理，芯片制造过程中，将存储和计算单元集成在同一颗die上，使得存储器本身具备一定计算能力，与近存计算相比，“存”与“算”之间的距离更为紧密。
- 2021年三星推出HBM2-PIM，2022年海力士推出GDDR6-PIM，但未大规模使用。
- 根据报道，目前三星电子和 SK 海力士正在合作标准化 LPDDR6-PIM 内存产品。

图表：从存算分离到存内处理



图表：三星HBM-PIM



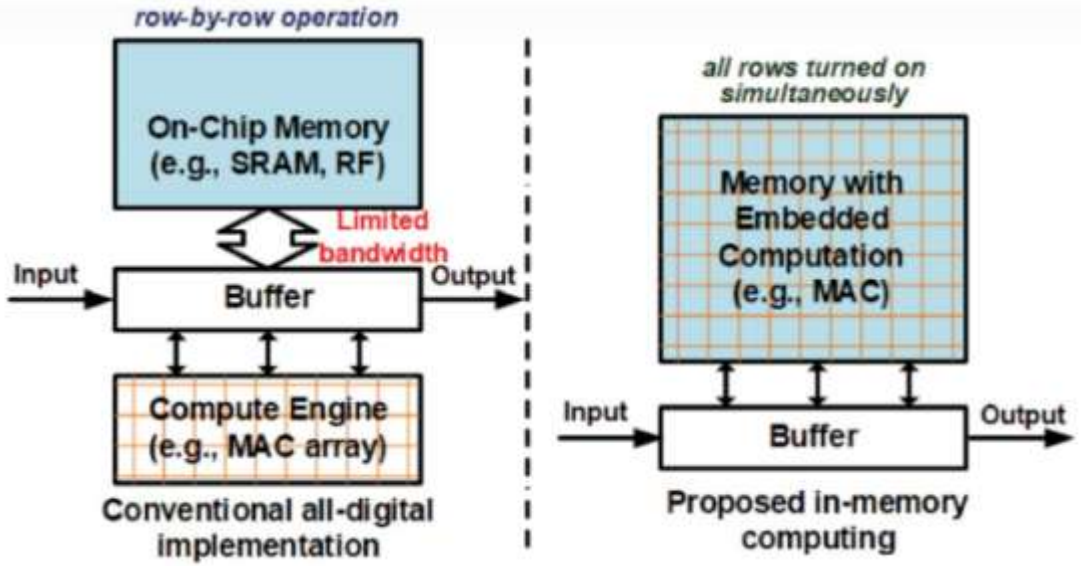
来源：三星、Bussiness Korea、芯语、财联社，中泰证券研究所

1.2存内计算难度大，近存计算和存内处理是重要方向

■ **存内计算：真正的存算一体，将存储和计算单元完全融合，计算效能大幅提升，但技术难度大。**

- 不区分存储单元和计算单元，真正意义上实现了同一个晶体管同时具备存储和计算能力，通过存储器颗粒上嵌入算法，存储电路同时具备存储和计算能力，计算由存储器芯片内部的存储单元完成全部计算操作，使得计算效能实现数量级提升，能耗可降至1/10-1/100，能效可提升 10-100TOPS/W。
- 存算一体的计算方式分为数字和模拟计算。数字存算一体主要以SRAM和RRAM为存储介质，采用先进逻辑工艺，具有高性能高精度的优势，且具备很好的抗噪声能力和可靠性。而模拟存算一体通常使用FLASH、RRAM、PRAM等非易失性介质作为存储介质，存储密度大，并行度高，但是对环境噪声和温度非常敏感。例如Intel和NVIDIA的算力芯片，尽管也可采用模拟计算技术提升能效，但从未有一颗大算力芯片采用模拟计算技术。因此数字存算一体适合大算力高能效的商用场景，而模拟存算一体适合小算力、不需要可靠性的民用场景。
- 存内计算芯片被认为是下一代芯片，但目前还处于起步阶段，受限于成熟度，应用范围不够广泛，面临着诸多挑战：1）在芯片设计方面，架构设计的难度和复杂度要求很高，同时市面上也缺乏成熟的存算一体软件编译器的快速部署、专用EDA工具辅助设计和仿真验证。2）在芯片测试方面，流片之后，同样缺乏成熟的工具协助测试。3）在生态方面，缺乏相应的与之匹配的软件生态。

图表：从存算分离到存内计算



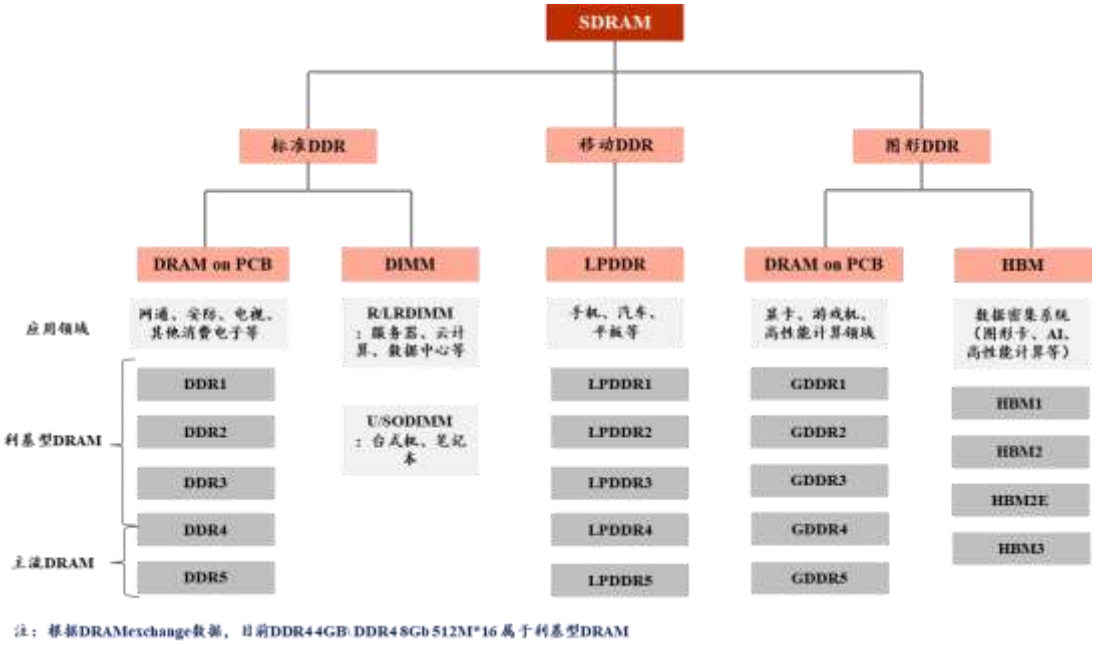
来源：芯语等，中泰证券研究所

目 录

2.1 HBM：AI大算力+高带宽存储解决方案

- **目前HBM属于标准化DRAM产品，是GDDR的一类。**DRAM是大宗产品，JEDEC（固态技术协会，微电子产业的领导标准机构）定义并开发了以下三类 SDRAM 标准，以帮助设计人员满足其目标应用的功率、性能和尺寸要求，从芯片本身来看，它们的差异主要体现在外围电路上，而存储单元本身在各类型中较为相似，制造工艺也基本一致。1）标准型DDR：Double Data Rate SDRAM，针对服务器、云计算、网络、笔记本电脑、台式机和消费类应用程序，允许更宽的通道宽度、更高的密度和不同的外形尺寸。2）LPDDR：Low Power Double Data Rate SDRAM，针对尺寸和功率非常敏感的移动和汽车领域，有低功耗的特点，提供更窄的通道宽度。3）GDDR：Graphics Double Data Rate SDRAM，适用于具有高带宽需求的计算领域，例如图形相关应用程序、数据中心和 AI等，HBM属于GDDR。详情请参考此前外发深度报告《AI系列之HBM：AI硬件核心，需求爆发增长》。
- **HBM主要应用在AI训练和部分AI推理。**AI训练需要处理大量并行数据，需要DRAM容量大和数据的传输速度快，同时模型训练耗时长，需要硬件的功耗低，相较传统的DRAM存储器，HBM高带宽、低功耗，容量拓展性好，目前云端训练卡全部使用HBM，部分云端推理卡有使用HBM，另外也有推理卡使用GDDR。

图表：标准DRAM分类



注：根据DRAMexchange数据，目前DDR4 4GB/DDR4 8Gb 512M*16属于利基型DRAM

图表：云端芯片的存储器使用情况

		定位（训练or推理）	峰值算力			所用内存	内存容量	内存位宽	峰值内存带宽
			FP8(TFLOPS)	FP16(TFLOPS)	FP32(TFLOPS)				
Intel数据中心GPU	GPU Flex 140	推理			8	GDDR6	12 GB	192bit	336GB/s
	GPU Flex 170	推理			16	GDDR6	16GB	256bits	576GB/s
英伟达数据中心GPU	B200	训练	9000	4500	80	HBM3E	192GB	8192bits	8 TB/s
	B100	训练	7000	3500	60	HBM3E	192GB	8192bits	8 TB/s
	H200	训练	3341	1671	60	HBM3E	141GB	6144bits	4.8TB/s
			3958	1979	67	HBM3E	141GB	6144bits	4.8TB/s
	H100	训练	3026	1513	51.2	HBM2E	80GB	5120bits	2TB/s
			3958	1978.9	66.9	HBM3	80GB	5120bits	3.35TB/s
			7916	3958	134	HBM3	188GB	6144bits	7.8TB/s
	L4	推理	485	242	30.3	GDDR6	24GB	192bits	300GB/s
	L40S	推理	1466	733	91.6	GDDR6	48GB	384bits	864GB/s
	L40	推理	724	362.1	90.5	GDDR6	48GB	384bits	865GB/s
	A100	训练		312	19.5	HBM2	80GB	5120bits	1935GB/s
				624					2039GB/s
	A2	推理		36	4.5	GDDR6	16GB	128bits	200GB/s
	A10	推理		250	31.2	GDDR6	24GB	384bits	600GB/s
	A16	推理		71.6	18	GDDR6	64GB	128bits	800GB/s
	A30	推理		330	10.3	HBM2	24GB	3072bits	933GB/s
	A40	-		299.4	37.4	GDDR6	48GB	384bits	696GB/s
AMD数据中心GPU	MI50	推理		26.5	13.3	HBM2	16GB	4096bits	1024GB/s
							32GB		
	MI60	推理		29.49	14.7	HBM2	32GB	4096bits	1024GB/s
	MI100	训练/推理		184.6	23.1	HBM2	32GB	4096bits	1.2TB/s
	MI250	训练/推理		362.1	45.3	HBM2E	128GB	8192bits	3.2TB/s
	MI250X			383	47.87	HBM2E	128GB	8192bits	3.2TB/s
	MI210			181	22.6	HBM2E	64GB	4096bits	1.6TB/s
	MI300A	训练				HBM3	128GB		3.2TB/s
	MI300X	训练				HBM3	192GB		5.2TB/s

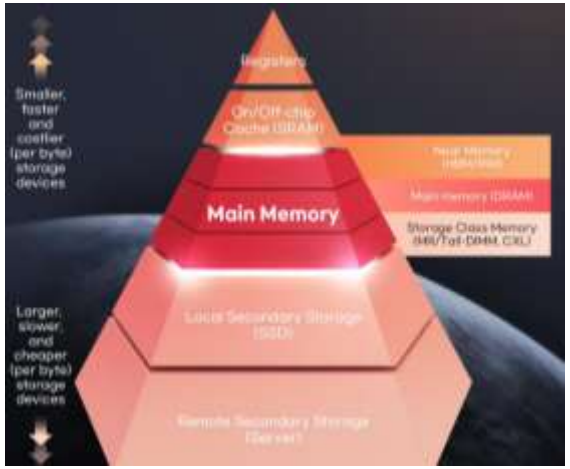
来源：Semiconductor Engineering、各公司官网等，中泰证券研究所

2.1 HBM：AI大算力+高带宽存储解决方案

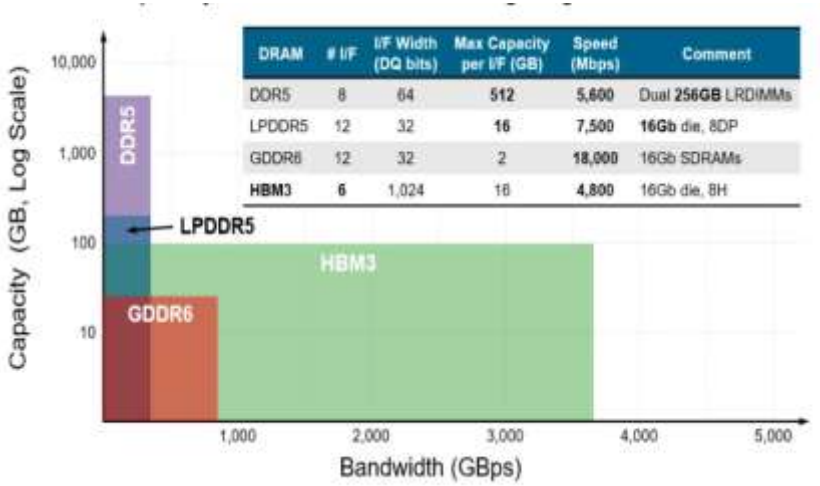
■ **HBM定位在片上缓存LLC和传统DDR中间，弥补带宽缺口，与GDDR等传统DRAM产品相比，兼顾带宽和容量。**

- HBM定位在CPU/GPU片上缓存（Last Level Cache，LLC，通常是SRAM）和DRAM之间，弥补处理器高带宽需求与主存储器最大带宽供应能力之间的带宽缺口，容量大于片上存储、小于传统DDR，但速度小于片上存储、大于传统DDR，成本低于片上存储、高于传统DDR。
- 以成本为例，1MB SRAM 价值$5~$10，1GB HBM价格$10-$20，根据集邦咨询，24年2月1GB DDR4合约价 $1.95（历史最高$4.1），1GB=1024MB，从单位存储成本看，SRAM成本是HBM的500倍+、普通DRAM的1000倍+，HBM是普通DRAM的5倍+。
- 从速度来看，在AI应用中，每个SoC的带宽需求（尤其是在训练应用中）都会超过几TB/s，但常规主存储器无法满足这个要求，具有3200Mbps DDR4 DIMM的单个主存储器通道只能提供25.6GB/s的带宽，具有4800Mbps DDR5 DIMM的单个主存储器通道提供38.4GB/s，即使是具有8个存储器通道的最先进的CPU平台，DDR4和DDR5对应速度也只能达到204.8GB/s、307GB/s，围绕单个SoC的4个HBM2堆叠可提供大于1TB/s的带宽，因而能够消除带宽差距。

图表：HBM定位在片上存储和普通DRAM之间

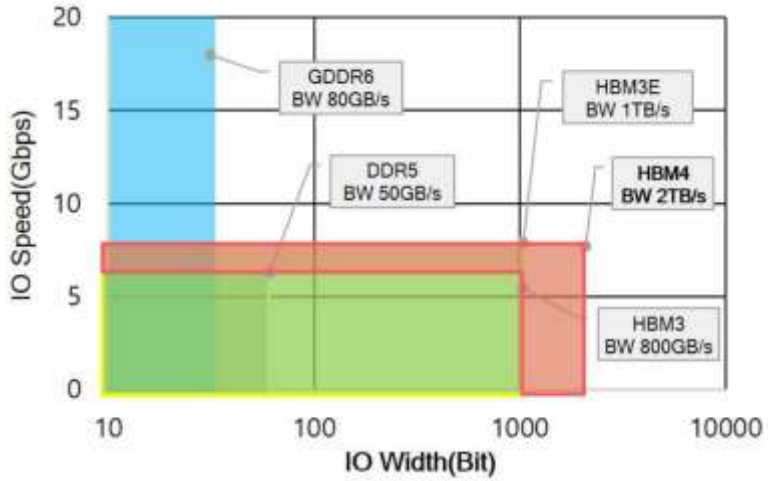


图表：HBM兼顾带宽和容量



DRAM	# I/F	I/F Width (DQ bits)	Max Capacity per I/F (GB)	Speed (Mbps)	Comment
DDR5	8	64	512	5,600	Dual 256GB LRDIMMs
LPDDR5	12	32	16	7,500	16Gb die, 8DP
GDDR6	12	32	2	18,000	16Gb SDRAMs
HBM3	6	1,024	16	4,800	16Gb die, 8H

图表：存储的带宽和速度



注：计算3200Mbps DDR4 DIMM的单个主存储器带宽：3200Mbps是等效传输效率，最大时钟频率=3200Mbps/2=1600MHz，总线宽带=64bits，每时钟数据段数量=2
内存带宽=最大时钟频率（MHz）×总线宽度（bits）×每时钟数据段数量÷8=1600×64×2=25600MB/s=256600MB/s÷1024GB/s=25.6GB/s

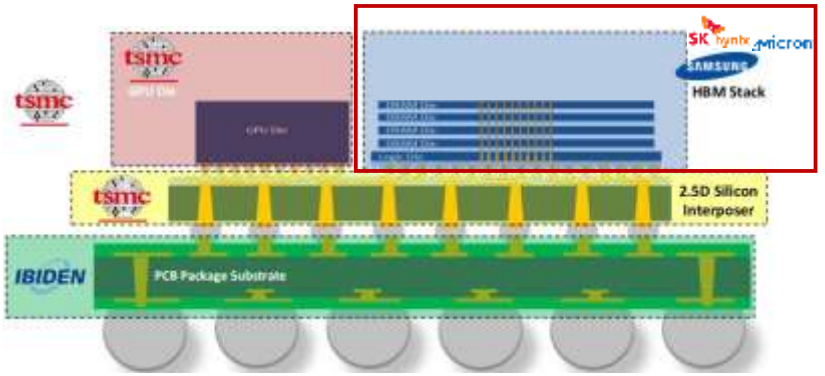
来源：海力士，Synopsys，中泰证券研究所

2.1 HBM：AI大算力+高带宽存储解决方案

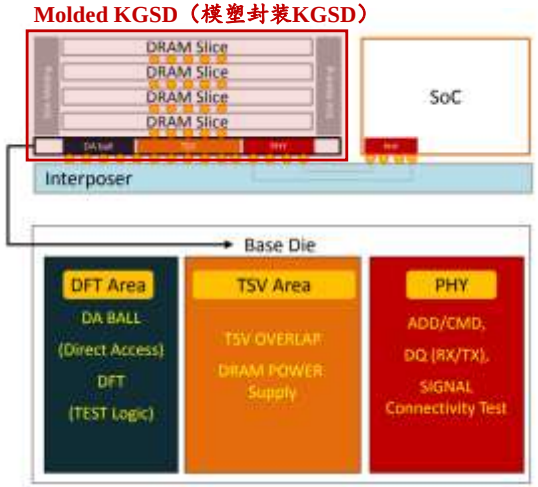
■ **HBM使用TSV、Microbump实现3D堆叠结构，并采用2.5D封装技术（CoWos）实现与GPU直接封装在一起，在不占用面积的前提下，实现容量拓展、高带宽和降低功耗。**

- 供应链：海力士、三星等存储原厂将HBM采用晶圆级封装，以KGSD（Known Good Die Stack，已知合格堆叠芯片）的封装形式交给台积电，台积电使用2.5D封装技术（包括CoWoS）将HBM与SoC（GPU等）封装在一起。关于CoWoS工艺的具体介绍，详情请参考此前外发深度报告《AI系列之先进封装：后摩尔时代利器，AI+国产化紧缺赛道》。
- 结构：1颗HBM KGSD = N 颗DRAM芯片（也称为Core Die） + 1颗逻辑芯片（也称为Logic Base Die）组合而成，目前N=4/8/12，预计HBM4将采用16颗DRAM芯片堆叠。将多片HBM DRAM Die堆叠在一颗Logic Die，DRAM Die之间、DRAM和Logic Die均通过硅通孔（TSV）和Microbump（微凸块）连接。DRAM与Logic Die放置在Interposer（中介层）上与GPU互联，中介层放置在ABF载板上，最后HBM与GPU使用2.5D封装技术封在一起。
- 逻辑芯片的三个功能区：①用于测试的区域（DFT Area），②TSV区域，TSV用于给DRAM 芯片传输信号和电力，③PHY芯片区域， HBM和SoC中的存储控制器之间的接口。PHY芯片区域和TSV区域中间有1024根信号传输线路，对应1024bit总线位宽。逻辑芯片的大小通常大于DRAM芯片，如海力士8层HBM3的逻辑芯片大小为10.8 mm x 9.8 mm，而DRAM芯片为10.5 mm x 9.5 mm，这是为了可以模塑封装（Mode晶圆模塑，一种扇出型晶圆级芯片封装工艺）以保护晶圆，通常使用环氧树脂模塑料（EMC）作为填充材料。

图表：供应链



图表：结构图



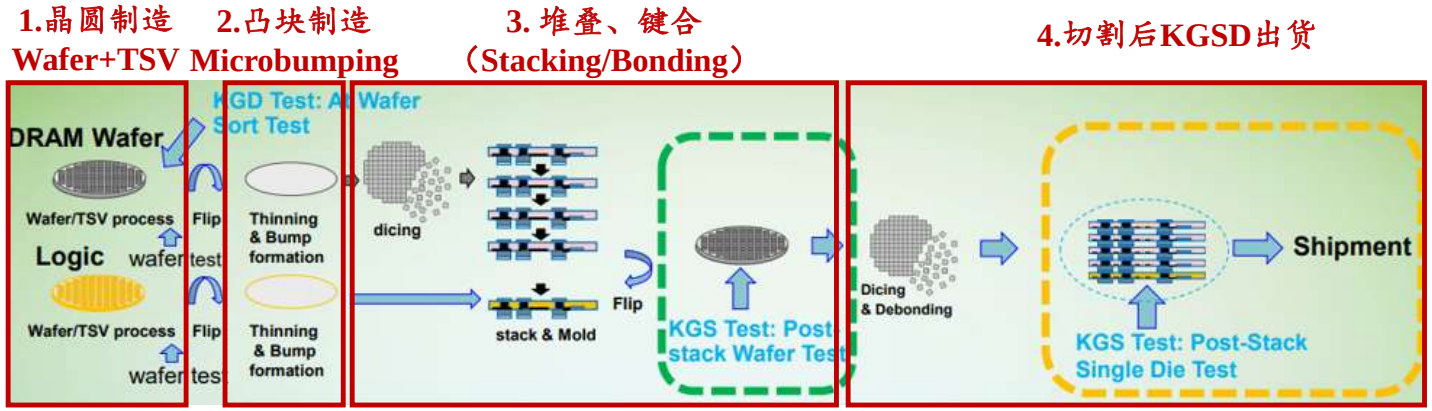
来源：System Plus Consulting，海力士，中泰证券研究所

2.1 HBM：AI大算力+高带宽存储解决方案

■ **制造：采用TSV、Microbump等先进封装工艺。**

- HBM制造流程分为四步，涉及TSV（硅通孔）、Microbump（凸点制造）、堆叠键合等技术。HBM从设计、制造和封测方式均与传统DRAM有较大区别，相较传统DRAM，HBM多了TSV、逻辑晶圆制备、凸点制造、堆叠键合等工艺，主要差异集中在封装测试部分，HBM KGSD的制备工艺包括扇出型晶圆级封装、TSV、Microbump等先进封装技术。
- 1）晶圆制造（包括TSV）：分别制造DRAM晶圆和逻辑晶圆，同时做好DRAM和逻辑晶圆的TSV硅通孔，TSV硅通孔需要晶圆制造工艺，包括深孔刻蚀、气相沉积、铜填充、CMP、晶圆减薄等工艺，此时DRAM和逻辑都是处于晶圆阶段，与传统DRAM主要差异是HBM晶圆需要制造TSV。
- 2）凸点制造（Microbum）：将硅通孔后的DRAM晶圆和逻辑晶圆倒装，然后进行减薄，在晶圆背面形成凸点，此时DRAM和逻辑都是处于晶圆阶段。
- 3）堆叠和键合（Stack&Bond），主要的差异化环节：在进行堆叠前，DRAM晶圆和逻辑晶圆的TSV通孔和凸点均已做好，DRAM晶圆切割成DRAM颗粒，DRAM颗粒一层一层堆叠在逻辑晶圆上，然后进行键合（此处为Die to wafer的键合），再进行晶圆模塑封装，最后获得模塑封装后的KGSD（Molded KGSD）。海力士和三星/美光主要是在键合工艺上有差异，三星/美光使用较为传统的TC-NCF（Thermo-Compression Bonding with None Conductive Film，热压缩-非导电薄膜），先在有TSV和凸点的晶圆上填充NCF，然后堆叠进行热压键合，后进行模塑封装，而海力士采用独创的MR-MUF工艺（Mass Reflow Bonding with Molded UnderFill，大规模回流焊-注塑底填充技术），不使用NCF，直接先堆叠，然后进行大规模回流焊做凸点的键合，然后使用以液体EMC为主要原材料的MUF使用模塑方式填充缝隙，工艺具体介绍详见后文。
- 4）切割KGSD晶圆获得KGSD颗粒：将模塑后的KGSD晶圆切割成颗粒，测试完成后出货给台积电继续做CoWoS封装。

图表：HBM Stack制造流程



来源：海力士，中泰证券研究所

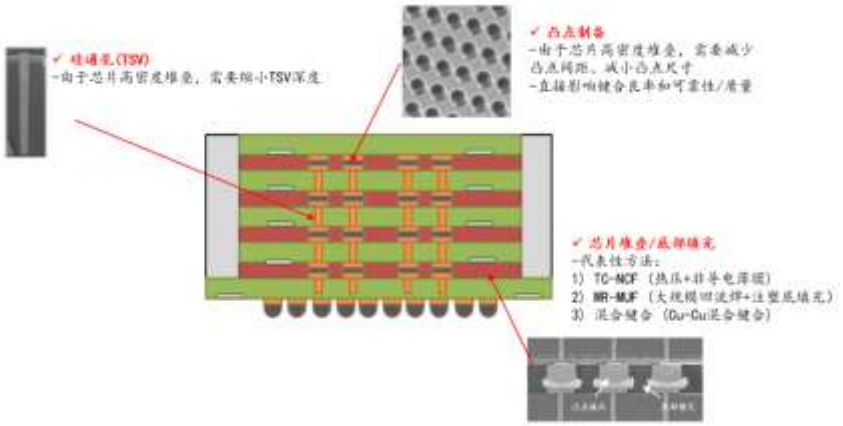
2.1 HBM：AI大算力+高带宽存储解决方案

■ **HBM三大关键工艺：TSV、Microbump和堆叠键合**

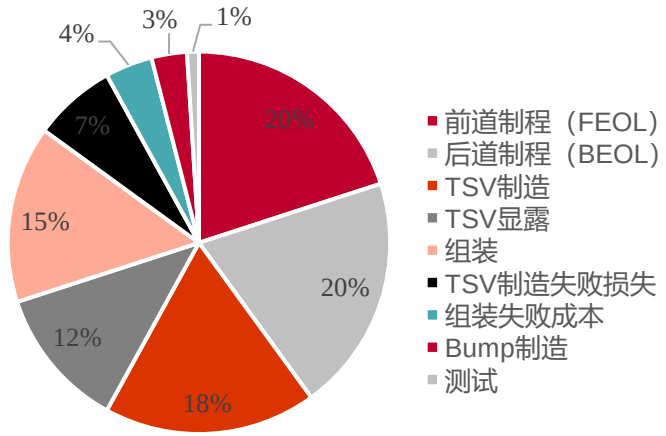
➢ TSV实现电气连接通路，在HBM成本中占比最高，约30%。HBM核心工艺主要是TSV、micro bump和堆叠键合，其中TSV工艺是HBM中成本占比最高、最核心的工艺，利用TSV才能实现DRAM芯片的3D堆叠和芯片间的快速传输。根据3D InCites 2016年数据，在4层DRAM和1层逻辑的HBM中，99.5%的键合良率下，TSV工艺所占的成本比重为30%，其中TSV制造（在正常晶圆厚度上制作TSV的过程）为18%，TSV显露（晶圆减薄等工艺使TSV触点露出）为12%。

➢ Microbump是芯片倒装的基础。Bump技术具备引脚密度高、低成本的特点，是构成倒装技术的基础。相较于传统打线技术（Wire Bond）的“线连接”，Bump技术“以点代线”，在芯片上制造Bump，连接芯片与焊盘，此种方法拥有更高的端口密度，缩短了信号传输路径，减少了信号延迟，具备了更优良的热传导性及可靠性，也是进行FC（Flip Chip）倒装工艺在内的先进封装工艺的技术基础。

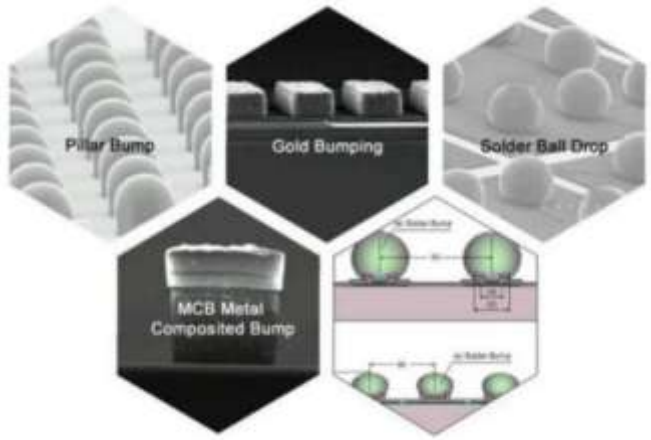
图表：HBM核心工艺：TSV、micro bump和堆叠键合



图表：HBM（4层DRAM+1层逻辑）3D封装成本划分（99.5%键合良率）



图表：Bump金属凸点



来源：海力士、3D Incites、《先进封装与异构集成》，中泰证券研究所

2.1 HBM：AI大算力+高带宽存储解决方案

■ **堆叠键合工艺主要包括：NCF、MUF、混合键合。**

- HBM2，Bump pitch（凸点间距）在55μm，三星和海力士共同使用TCB（热压合）技术，其中海力士采用的是TCB的分支TCB-NCF。
- HBM2/2E/3/3E，Bump pitch进展到25/22μm水平，三星继续采用TCB技术，而海力士独家采用MR-MUF（大规模回流焊-注塑底填充技术）。
- HBM4，规划12层和16层，目前12层明确不使用混合键合，16层方案暂未确定。 24年11月海力士使用MR-MUF工艺的16层HBM3E发布。

图表：不同代际HBM的Bump间距与互联技术

	HBM2	HBM2E/3	HBM3（12层）/3E
Bump pitch（μm）	55	25	22
层数（Hi）	4/8	4/8	8/12（HBM3E有8层、12层版本）
海力士的内部互联封装	TCB-NCF （热压合-非导电薄膜技术）	MR-MUF （大批量回流焊-注塑底填充技术）	Advanced MR-MUF
三星的内部互联封装	TCB（热压合）	TCB（热压合）	TCB（热压合）

来源：DAUM、NEWSIS，中泰证券研究所

2.1 HBM：AI大算力+高带宽存储解决方案

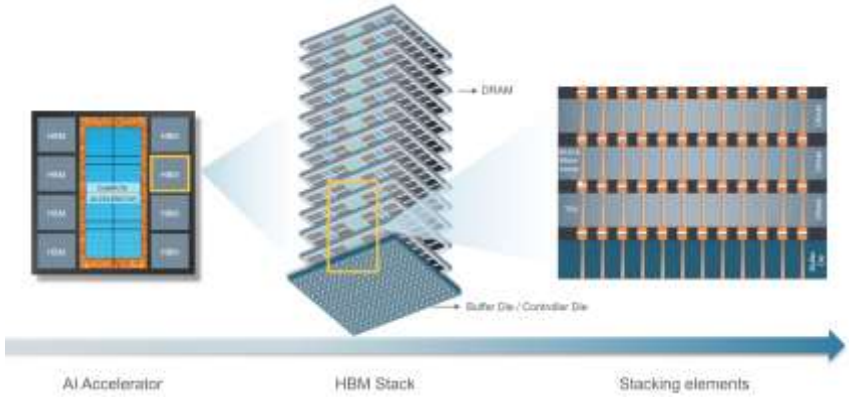
■ **性能特色：标准化产品，与GDDR等传统DRAM产品相比，HBM带宽高、功耗低，同时容量可拓展。**

➢ 1）高带宽：因为使用TSV和Microbump，在单位面积下可以创造更多的数据连接点，即数据的传输的I/O数量多，达到1024个IO数量，带宽=位宽×数据的传输速度。

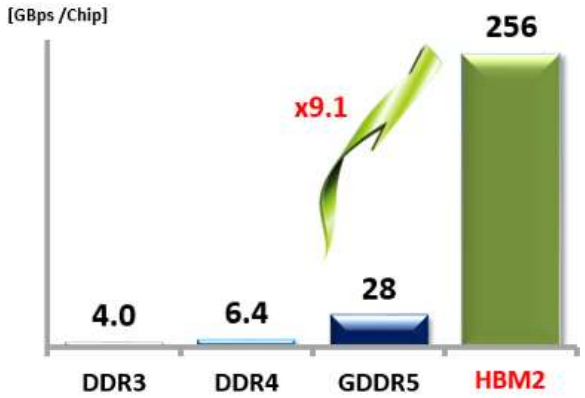
➢ 2）功耗低：GDDR采用正常2D结构，不需要中介层连接，总线位宽小，主要是通过体现数据的传输速率来提升位宽，因为数据的传输速率快，因此功耗也高于HBM，GDDR基本50%的功耗是用于高速的数据的传输（PCB走线传输），而HBM用TSV技术实现走线更短，同时I/O数据的传输速度慢，功耗低。

➢ 3）占用面积小、容量可拓展：HBM将多层DRAM进行3D垂直方向的堆叠，通过增加层数来扩展容量，GDDR为2D结构，因此HBM在实现相同容量下，占用的面积更小。同时HBM与GPU通过中介层连接，1个GPU旁边可以放置多颗HBM，中介层面积相对容易拓展。

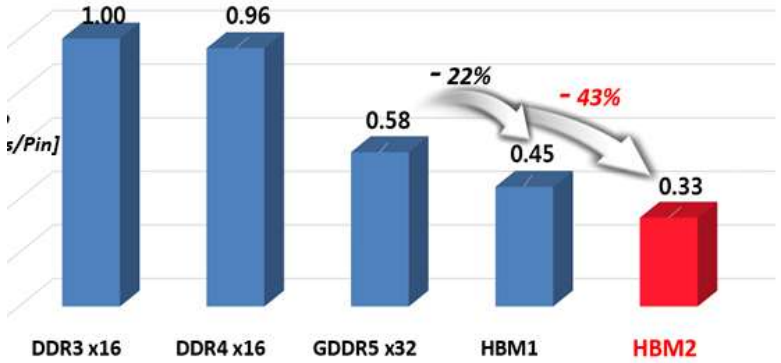
图表：HBM结构图



图表：HBM高带宽



图表：HBM低能耗



来源：AMAT，海力士，Synopsys，中泰证券研究所

2.1 HBM：AI大算力+高带宽存储解决方案

■ **性能特色：标准化产品，带宽高、功耗低，同时容量可拓展。**

➢ 4）标准化产品。HBM的标准由JEDEC指定，对HBM成品的长宽高、Microbump的位置形状、通道数量、数据的传输速度等参数均有明确要求。

图表：HBM3 Microbump参数

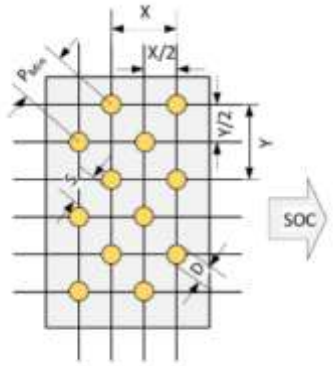


Figure 95 — Staggered MicroBump Pattern

Table 95 — Geometric Parameters of the Staggered MicroBump Pattern

Label	Nominal Value	Description
X	96 um	Horizontal pitch of two adjacent MicroBumps
Y	110 um	Vertical pitch of two adjacent MicroBumps
P_{Min}	73 um	Minimum pitch of the bump field
D	28 um	MicroBump diameter
S		Bump-to-bump air gap, $S = P_{Min} - D$

The HBM3 bump matrix is defined as shown in subsequent tables. Please refer to MO-349 for device dimensions.

Figure 96 — MicroBump Pillar Diameter

图表：HBM成品的长宽高参数

Table 96 — HBM3 Device Dimensions

Parameter	Symbol	Configuration	Minimum	Nominal	Maximum	Unit	Notes
Width	X			10.975		mm	3
Length	Y			10.975		mm	
Height	Z	4-High	695	720	745	µm	1,2
		8-High	695	720	745	µm	
		12-High	695	720	745	µm	
		16-High	TBD	TBD	TBD	µm	

NOTE 1 The configuration refers to the number of memory dies in the stack. The stack may include an additional base (interface) die.
NOTE 2 Refer to MO-349 for related package drawings.
NOTE 3 Refer to MO-349 for X and Y dimension min/max tolerances and related package drawings.

来源：JEDEC，中泰证券研究所

2.1 HBM：AI大算力+高带宽存储解决方案

■ **HBM方案下，GPU增加带宽和容量的方式主要是增加HBM颗数和提升单颗HBM的性能。**

➢ 容量增加：1）增加HBM颗数：目前1颗8层HBM3E可提供24GB，GPU增加1颗HBM3E，可增加24GB容量。但HBM的颗粒必须跟GPU对齐和封装在一起，是紧耦合的状态，受限GPU面积，HBM数量不能无限增加，同时还需考虑散热等问题。2）提升单颗HBM容量：提高单颗HBM的容量，HBM通常是100mm^2的面积，容量增加一方面来自单层容量密度提升，主要是由升级制程，另一方面来自堆叠层数的增加，但因HBM的高度需要与GPU高度相对平行，层数不能无限增加，因此需要通过升级键合工艺、晶圆减薄工艺等。

➢ 带宽增加：1）增加HBM颗数：目前1颗HBM3E可提供1024bit总线位宽，增加1颗HBM3E，可增加1024bit总线位宽。2）提升单颗HBM的带宽：带宽=位宽x数据的传输速度，位宽的增加，主要是通过创造更多I/O，即数据连接传输点，主要通过改进键合工艺，实现更小的pitch，而数据的传输速度的提升，主要是来自制程升级。

■ **HBM不断迭代，迭代方向为增加容量和带宽，目前量产的最高层数为12层HBM3E。**

➢ 从单颗容量看，堆叠层数和单层DRAM容量均有所增加，HBM1仅堆叠4层2Gb的DRAM，实现单颗HBM 8Gb（1GB），而HBM3E最高堆叠12层3GB的DRAM，实现单颗HBM 36GB，HBM4最高16层堆叠。从I/O数量看（总线位宽），HBM1到HBM3E均保持在1024bit，而数据的传输速率从HBM1的1Gb/s提升到HBM3E的9.2Gb/s，最终实现带宽从HBM1的128GB/s提升至HBM3E的1.2TB/s

图表：HBM迭代情况（参考海力士官网）

	HBM1	HBM2	HBM2E	HBM3	HBM3E	HBM4
年份	2014	2018	2020	2022	2024	2026
堆叠层数	4	4 or 8	4 or 8	8 or 12	8 or 12	12 or 16
单层DRAM容量	2Gb	1GB	2GB	2GB	3GB	4GB
容量	1GB	4GB OR 8GB	8GB OR 16GB	16GB OR 24GB	24GB OR 36GB	48GB OR 64GB
I/O数量（总线位宽，bit）	1024	1024	1024	1024	1024	2048
I/O速度（数据的传输速率）	1Gbps	2.4Gbps	3.6Gbps	6.4Gbps	9.8Gbps	6.4+Gbps
带宽	128GB/s	307GB/s	460GB/s	819GB/s	1.2TB/s	1.5 - 2.56 TB/s
电压	1.2V	1.2V	1.2V	1.1V	1.1V	1.05V

来源：海力士，中泰证券研究所

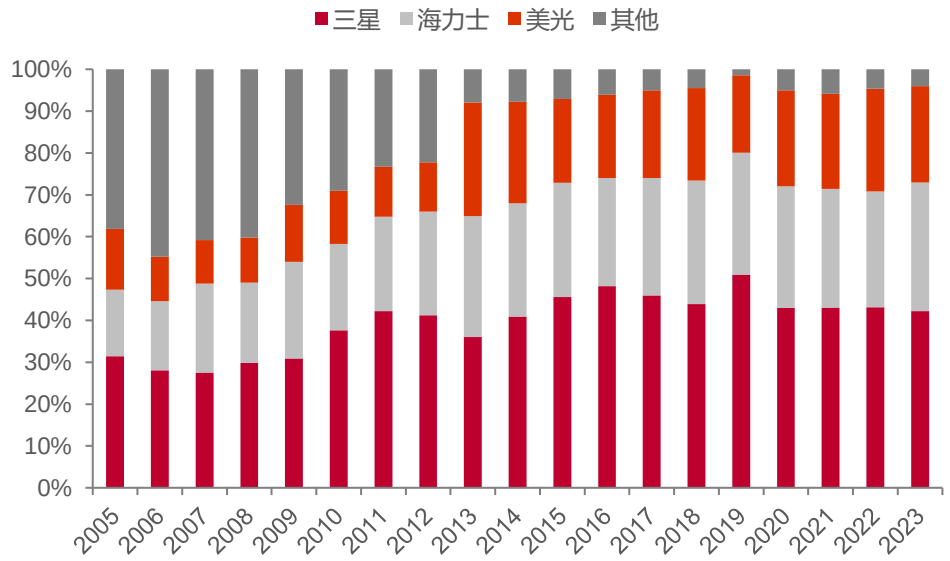
2.1 HBM：AI大算力+高带宽存储解决方案

■ **竞争格局：海力士、三星和美光垄断。**

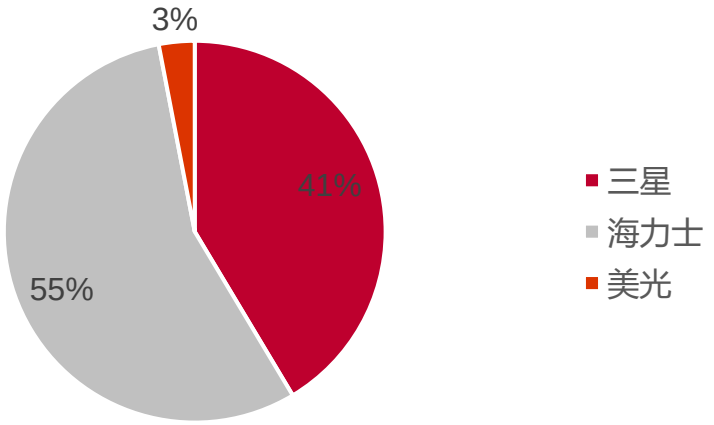
➢ AI驱动，HBM市场快速增长：美光预计24年HBM市场规模160亿美金，预计25年市场规模超过300亿美金，预计到2030年市场规模超过1000亿美金。

➢ DRAM市场由三家DRAM IDM 三星、海力士、美光垄断，2023年三家合计市占率96%，另外DRAM IDM还有中国台湾南亚科、华邦和力积电，大陆长鑫、晋华等。而HBM市场垄断效应更强，2023年海力士/三星/美光份额为55%/41%/3%。

图表：DRAM竞争格局



图表：HBM竞争格局（2023）



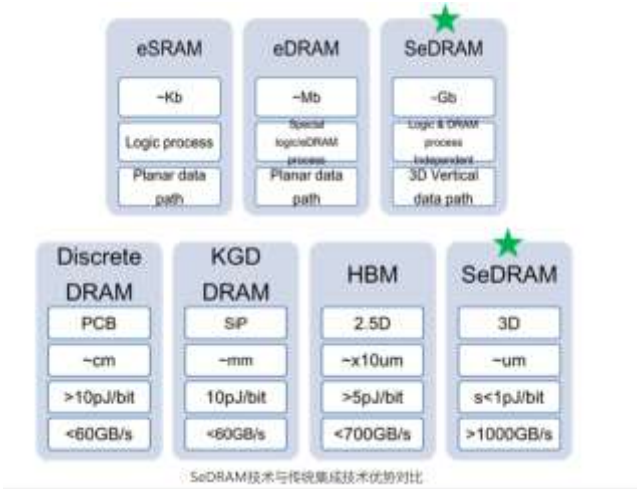
来源：Trendforce，Yole，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

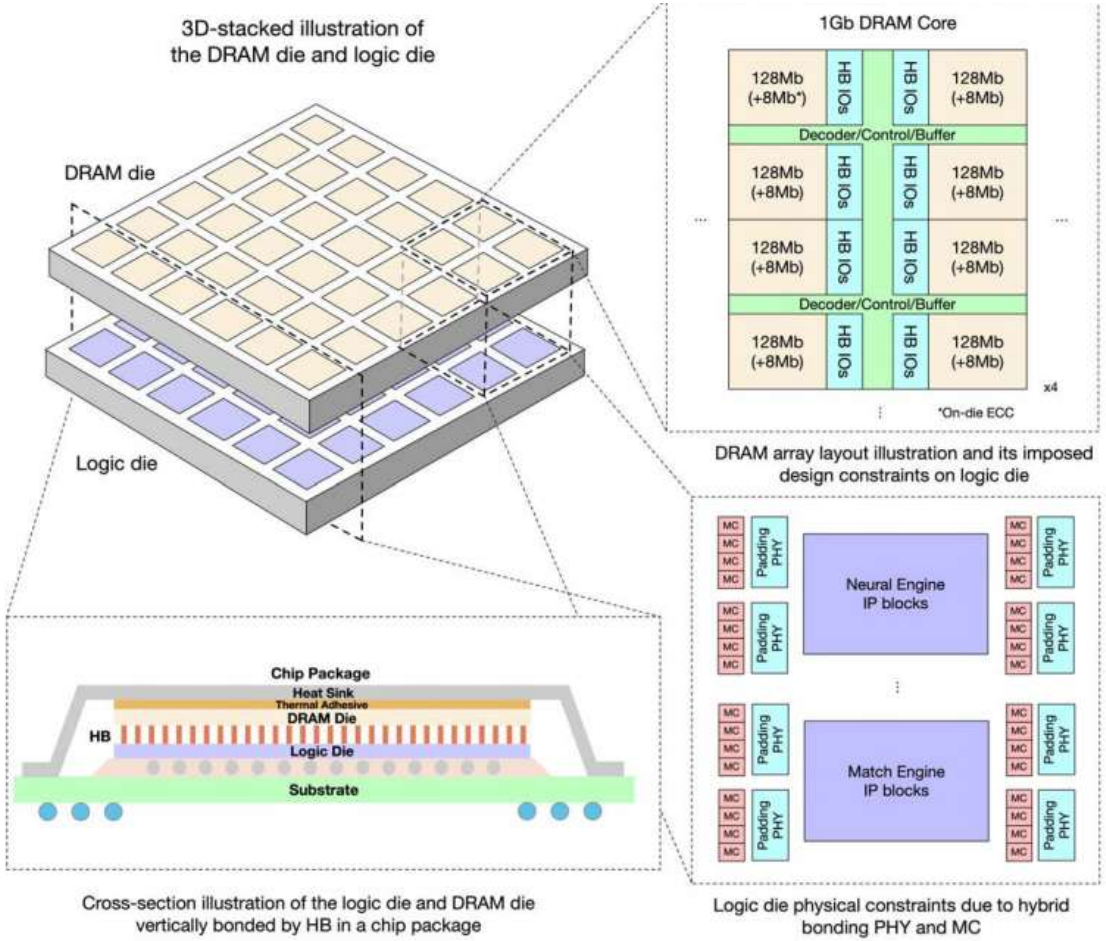
■ **WOW 3D堆叠DRAM与逻辑芯片是3D结构，属于近存计算。**

- 结构：属于近存计算，DRAM与逻辑芯片采用3D堆叠工艺封装在一起，在1片逻辑芯片上堆叠多层DRAM芯片，逻辑芯片指GPU、CPU、NPU等计算芯片、右图中为紫色的Logic Die，DRAM芯片图中仅只有1层，实际可堆叠多层。
- 技术：使用TSV硅通孔技术、Wafer on Wafer的混合键合工艺（Hybrid Bonding）实现多层芯片之间的电气连接。
- 性能特点：以紫光国芯的WOW 3D堆叠DRAM产品 SeDRAM为例，通孔间距（Pitch）达到10μm以内的级别，HBM的Pitch目前为几十微米，因此WoW 3D堆叠DRAM的带宽更高，另外功耗更低，属于定制化产品，容量拓展性一般。

图表：紫光国芯的WOW 3D堆叠DRAM性能特点



图表：紫光国芯的WOW 3D堆叠DRAM



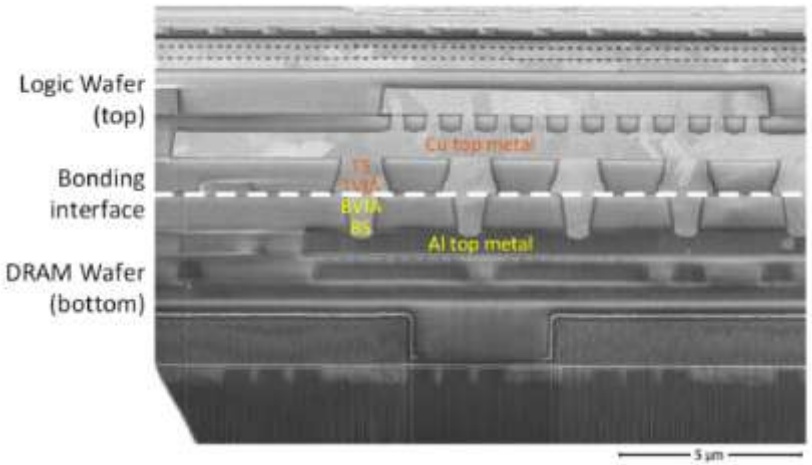
来源：IEEE，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

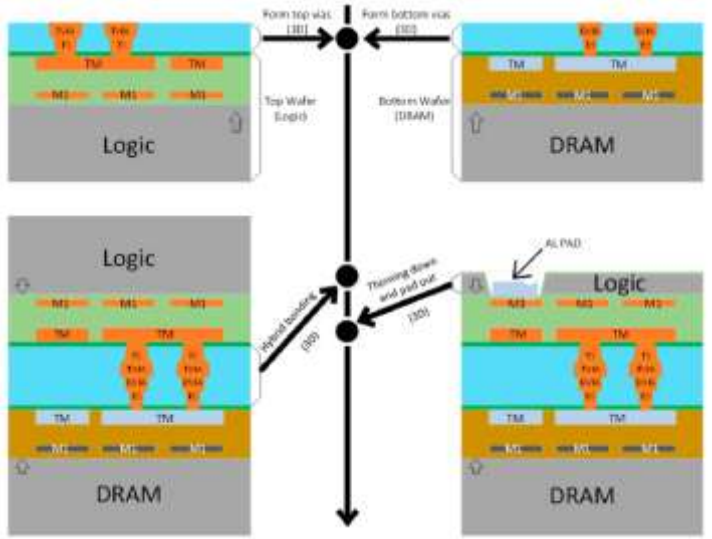
■ **制造：使用TSV、Wafer on Wafer混合键合等先进封装工艺。（参考紫光国芯SeDRAM制造工艺）**

- 1、制造Wafer：流片生产不同工艺下的DRAM晶圆（DRAM Wafer）和搭载有DRAM外围电路的逻辑晶圆（Logic Wafer）；
- 2、在晶圆上制造TSV通孔：通过平坦化、曝光和刻蚀等工艺，在DRAM和逻辑晶圆上分别制造接触通孔，顶部通孔为LTVIA，底部通孔为LBVIA；
- 3、多片晶圆的键合：
  - 1）多层DRAM晶圆的键合：以2层DRAM为例，将一片DRAM晶圆（DRAM1）正面键合到载体晶圆上，然后通过背面研磨和化学机械抛光（CMP）工艺将DRAM1的硅衬底研磨至几微米厚度，在减薄后进行TSV和混合键合工艺；在DRAM2上进行用于粘合铜焊盘的金属互连；将处理后的DRAM1和DRAM2晶圆通过混合键合Face to Back键合；最后移除载体晶圆，并利用顶部金属层工艺形成探测焊盘。
  - 2）逻辑和DRAM的键合：将逻辑晶圆翻转，通过Cu-Cu互连的方式，将逻辑Wafer的顶部和DRAM Wafer的底部进行Face to Face的混合键合（后续缩写为HB）；然后将逻辑晶圆减薄至约3um厚度，并从逻辑晶圆背面开口完成PAD制作。
- 4、传统的封测工艺：多层晶圆后键合后就相当于是一片晶圆，然后进行减薄、切割、测试等传统封装测试流程。

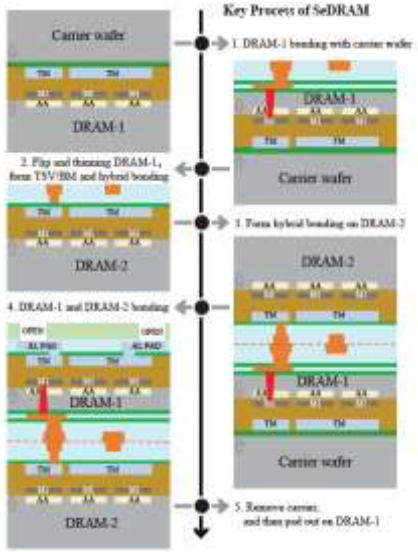
图表：3D堆叠DRAM的横截面TEM图像



图表：逻辑芯片和存储芯片的键合



图表：堆叠两层DRAM晶圆的关键工艺

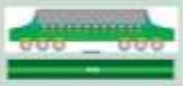
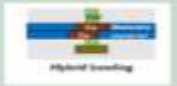


来源：MDPI、IEEE，中泰证券研究所

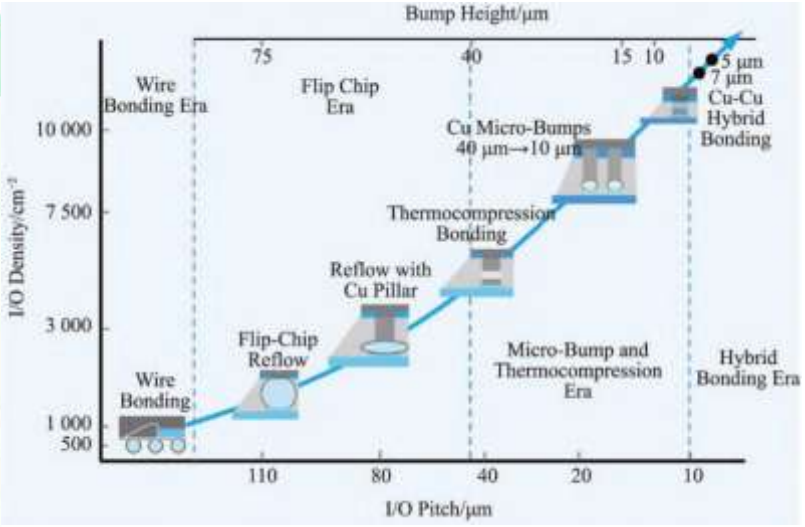
2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

- **WOW 3D堆叠DRAM的关键是混合键合工艺。混合键合的性能优势显著，是未来Bump技术的迭代方向。**
  - Microbump（连接是基于焊料）：在TSV铜通孔上生成焊球，如锡焊球，右下角图中的Microbump锡球，芯片之间通过焊料连接。
  - 混合键合（去掉焊料）：不再使用焊料，不同芯片或晶圆的互连直接通过铜通孔连接，直接铜连接可以降低电阻，从而在向各种芯片发送数据时降低功耗，另外去掉焊球后，铜通孔的间距可以做到小、通孔密度更高。
  - 混合键合用于10μm以下：Microbump很难缩小到10μm以下，混合键合用在10μm间距以下的领域。

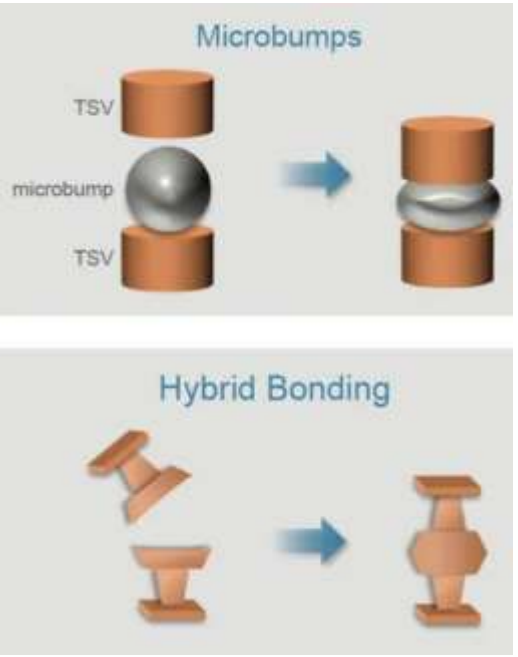
图表：键合技术的发展历史

	Wire Bond (1975)	Flip Chip (1995)	TCB Bonding (2012)	HD Fan Out (2015)	Hybrid Bonding (2018)
Architecture					
Contact Type	 Wire	 Solder ball or copper pillar	 Copper pillar	 RDL or copper pillar	 Copper to copper
Contact Density	 5-10/mm²	 25-400/mm²	 156-625/mm²	 500+/mm²	 10K-1MM/mm²
Substrate	Organic/leadframe	Organic/leadframe	Organic /Silicon	None	None
Accuracy	20-10μm	10-5μm	5-1μm	5-1μm	0.5-0.1μm
Energy/Bit	10pJ/bit	0.5pJ/bit	0.1pJ/bit	0.5pJ/bit	<.05pJ/bit

图表：Bump技术的发展趋势



图表：2种键合的示意图



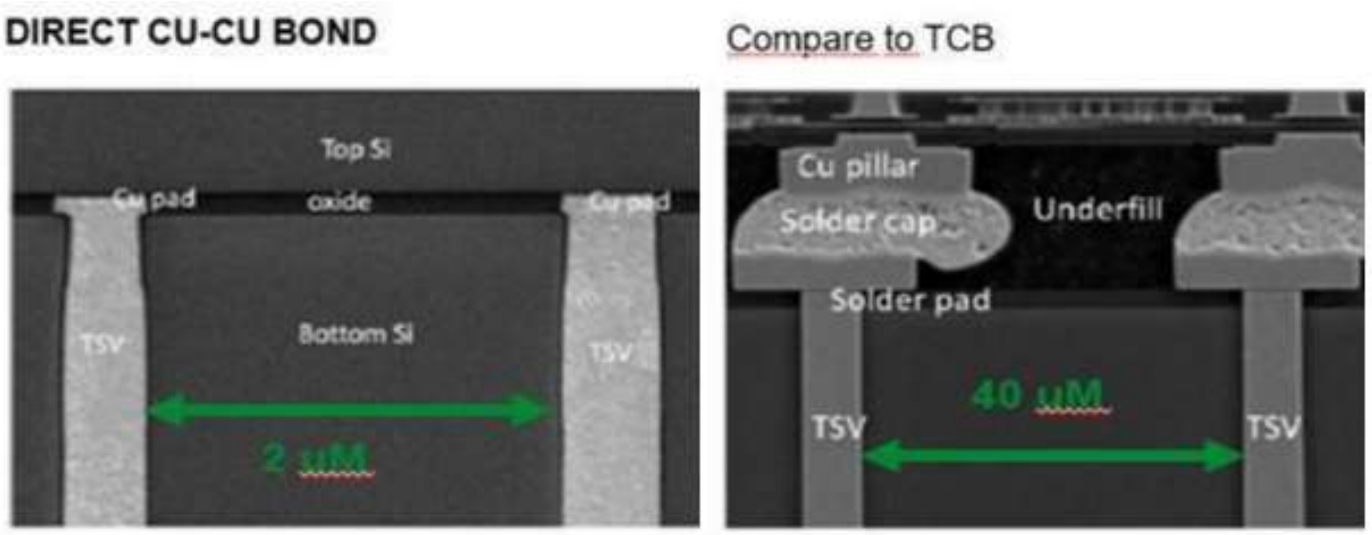
来源：《高端性能封装技术的某些特点与挑战》，Besi官网，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

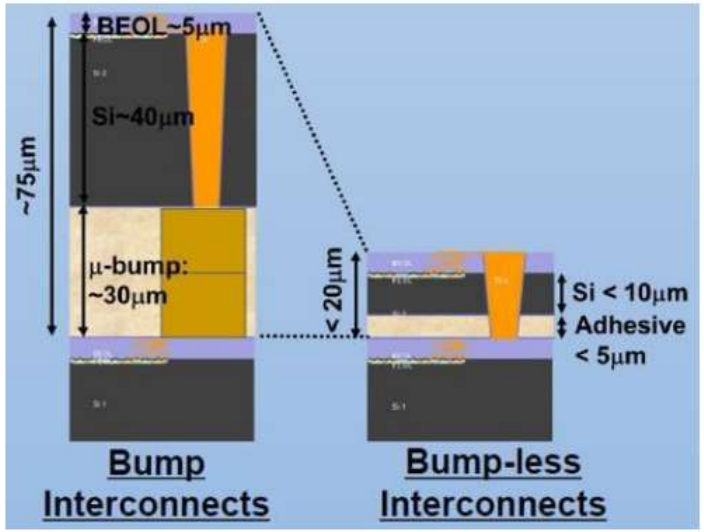
■ **混合键合改善互联结构，突破I/O密度瓶颈。**

- 1）I/O密度更高：混合键合金属焊盘（大小约为0.5um×0.5um方形）间距可以微缩到2um以下，极大地提升I/O密度；
- 2）走线距离更短：混合键合是直接键合，中间没有层间距，可以缩短小芯片间连线长度，从而改善总体性能、功率和成本，且相较于焊球键合约30um的层间厚度，混合键合封装的芯片会更薄。
- 3）省去底部填充成本：相较于倒装芯片键合，混合键合不需要在层间底部填充，可以省去相应材料成本。

图表：混合键合具有更高的I/O密度



图表：混合键合具有更短层间互联



来源：EVG，中泰证券研究所

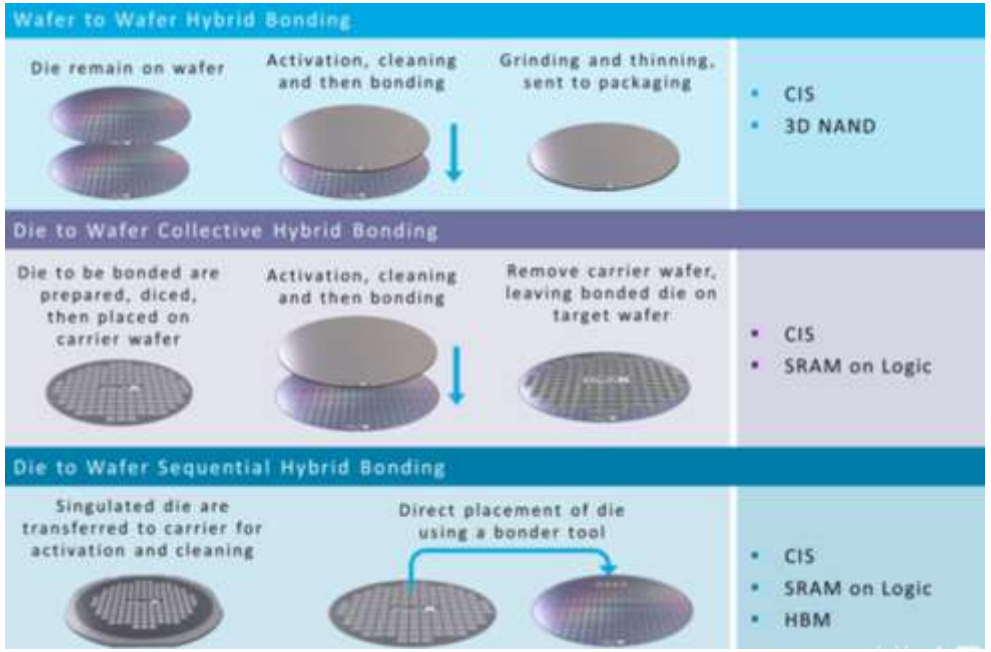
2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

■ **混合键合分为W2W（Wafer to/on Wafer，晶圆对晶圆）、D2W（Die to Wafer，芯片对晶圆）两大类，二者整体封装步骤相似，D2W涉及切片。**

➢ W2W是将两片晶圆直接键合，效率更高但良率较低，适用于高良率芯片的键合，目前应用在CIS/3D NAND等领域。W2W键合是指两个完整的晶圆进行键合，完成后再切割。因W2W键合前不需要晶圆切割，因此颗粒污染产生较少同时效率更高，根据贴装方式，可以进一步分为背对面键合与面对面键合。但是W2W键合无法筛选已知的良好芯片进行键合，这会导致有缺陷的芯片键合到合格芯片上，从而导致良率下降（约为两片晶圆的良率相乘）。对于尺寸较小的芯片，一片晶圆可以产出更多芯片，同样的缺陷面积造成的芯片损失率更小，其良率更高，一般来说更适合用W2W键合方式，因此其在CIS、3D NAND等高良率小型芯片上应用广泛。

➢ C2W良率更高，但因技术难度高，处于研发量产爬坡阶段。C2W是将晶圆切割后进行键合，整体工艺发展受限于清洁度与产能等因素限制，比W2W发展慢，但是C2W可以支持不同的芯片尺寸、晶圆类型，并可以将良好的芯片筛选出来进行键合，良率也会更高。

图表：混合键合工作流程



来源：应用材料官网，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

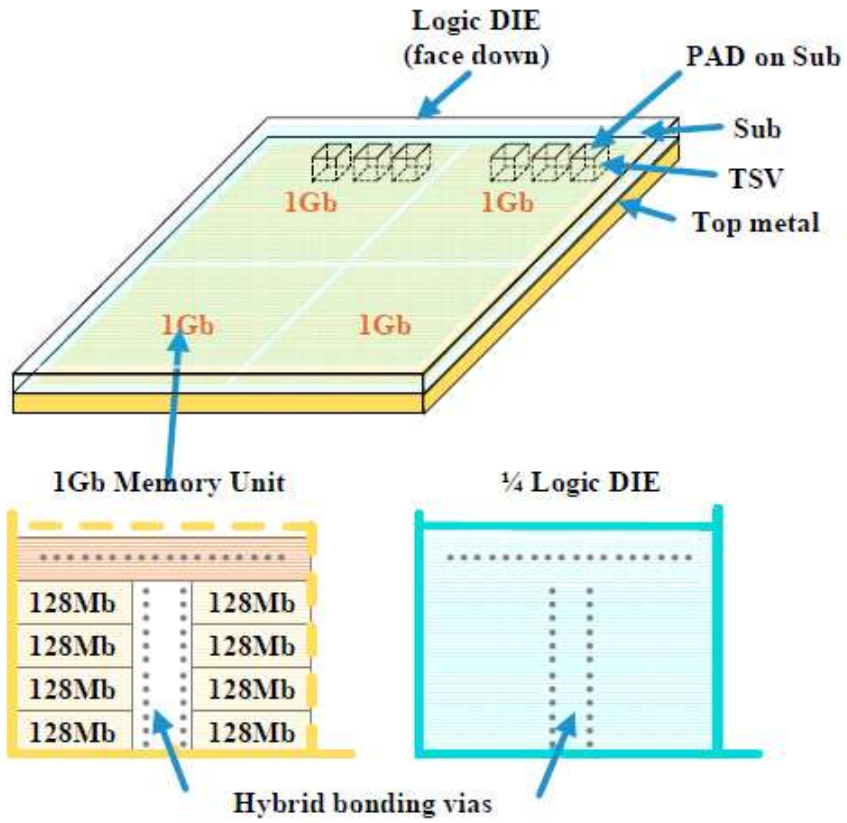
■ **WoW 3D堆叠DRAM是高度定制化产品，DRAM容量和层数可根据客户要求定制。**

- 拆解紫光国芯1层DRAM的SeDRAM产品：1层DRAM（4Gb）+1层Logic。
- 1）制程：DRAM 25nm，Logic 55nm
- 2）面积：DRAM和Logic都是25.24x23.86mm^2，面积相同。
- 3）DRAM和Logic连接的混合键合：
  - ①混合键合pitch是3μm，有超过6.4万个混合键合的孔，最大通孔密度110,000/mm^2。
  - ②Pad既是金属导线，同时也是DRAM和Logic之间的支撑材料。
  - ③混合键合的电阻小，因此逻辑到存储接口的能耗可以降低40%。
- 4）4Gb SeDRAM：容量4Gb，32个通道，4096个I/O（位宽4096bit），I/O速度为266 MHz，带宽136GBps。4Gb是由4个独立可扩展的1Gb存储单元阵列构成，根据需求SeDRAM容量可以组合成1Gb-48Gb。
- 5）1Gb SeDRAM规格：8个通道，1024个I/O（位宽1024bit），I/O速度266 MHz，带宽34GBps，0.88pJ/bit的功耗，1Gb的存储单元阵列是由8个128M的存储单元阵列和独立片上电源系统构成。每个128Mb存储单元阵列有128个I/O，每个128M都是一个独立的内存通道，具有单独的控制和数据信号，所有内存通道是可以同时访问，并行性高。
- 其他特点：
  - ①存储控制器（Memory Controller）、I/O电路等都在对应的Logic芯片上，SeDRAM去掉PHY结构。
  - ②logic芯片也是分区的，每个logic block都可以直接连接对应的dram block，同时可以通过on-chip bus连接所有其他memory blocks；
  - ③SeDRAM结构与传统DRAM制造工艺兼容。

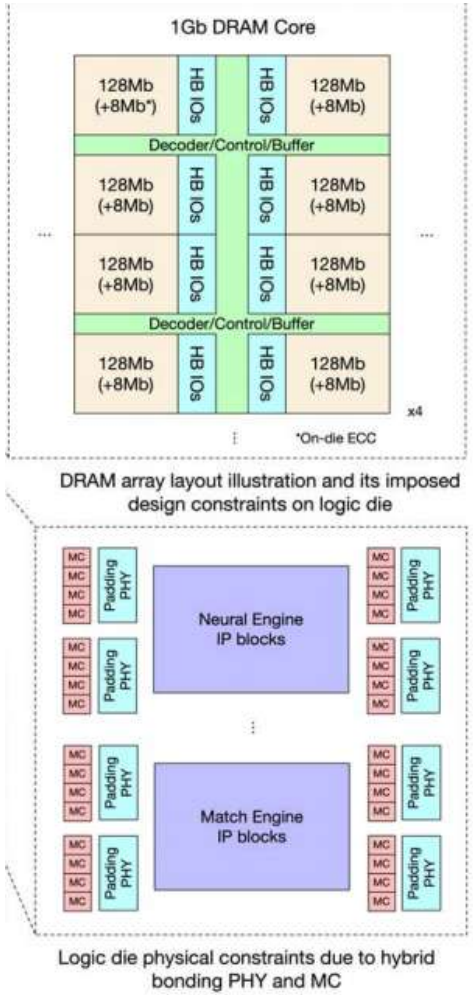
来源：IEEE，中泰证券研究所

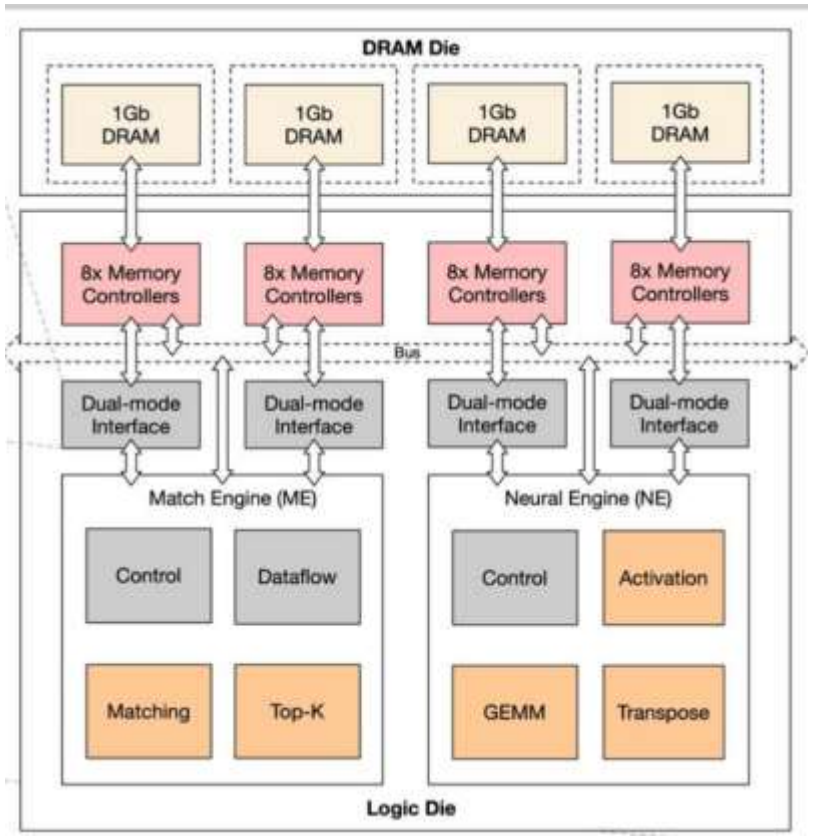
2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

图表：4Gb SeDRAM由4个1Gb构成，每个1Gb由8个128M构成



图表：逻辑电路、I/O等都在Logic芯片上

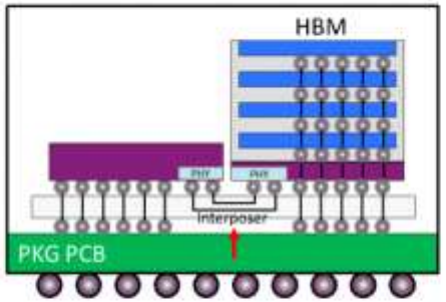
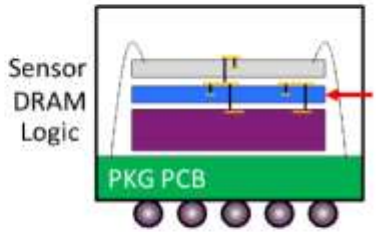
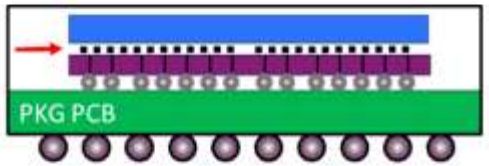




来源：IEEE，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

图表：4Gb SeDRAM的性能

	ISSCC14[3]	ISSCC20[4]	ISSCC17[5], IEDM17[6]	this proposal	
Structure diagram					
Connection,pitch(um)	**microbump, 48 x 55, interposer,**		**TSV,6.3 x 6.3, no interposer**	**Hybrid Bonding, 3 x 3, no interposer**	
Connection length	~5mm, microbump+wiring		~10um, TSV+wiring	2um, via thickness	→走线距离短，数据的传输快、功耗低
PHY needed	Yes		No	**No**	→不需要PHY，数据的传输快、功耗低
Energy efficiency(pJ/b)	~1.5[2]		N/A	**0.88**	→功耗低
Total Density	8Gb	128Gb	1Gb	4Gb	
# of Stack dies	4	8	1	1	
Density per die	2Gb	16Gb	1Gb	4Gb	
# of Channel	8	8x2	4	**32**	→并行的内存通道数量多
Data bus width	1024(128/ch)	1024(64/ch)	512(128/ch)	**4096(128/ch)**	→I/O多
Data rate per pin(Mbps/pin)	1000	4000	200	**266**	
Bandwidth(GBps)	128	512	12.8	136	
Bandwidth per die(GBps)	32	64	12.8	**136**	→带宽高

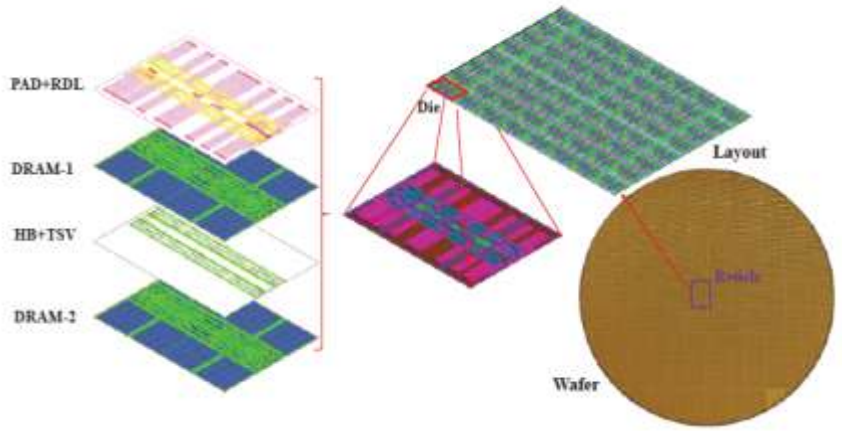
来源：IEEE，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

■ **WOW 3D堆叠DRAM 对比HBM：定制化产品，带宽更高，功耗更低，但容量拓展性不如HBM 。**

➢ 1、混合键合工艺的Pitch小，IO数量多，带宽较HBM有十倍以上提升。根据紫光国芯2023年发布论文中的2层SeDRAM方案，其使用WoW 混合键合工艺，DRAM和逻辑芯片的混合键合的通孔间距（Pitch）为 3um ，且每个过孔的电阻小于 0.5Ω，2层DRAM之间的Mini-TSV 的通孔间距缩小至 1.5um，能构建的IO数量更多，该2层DRAM产品64Gb（8GB，2层4GB），IO数量131072个，平均每Gb的IO数量达到2048个，而192Gb的HBM3（24GB，8层3GB）的IO数量为1024个，平均每Gb的IO数量为5.3个。紫光国芯的2层产品的IO速度为541Mbps（而HBM3 IO速度仅为7168Mbps），通过IO数量的提升，最终实现每Gb的带宽为135GB/s，而HBM3每Gb的带宽为4.7GB/s。

图表：紫光国芯2层DRAM的方案



图表：SeDRAM性能对比

	传统DRAM	HBM		WOW 3D堆叠DRAM		WOW 3D堆叠DRAM的特点
类型	GDDR6 ISSCC2018	HBM2E ISSCC2020	HBM3 ISSCC2022	SeDRAM IEDM2020	SeDRAM（2层）	
连接方式	-	ubump, TSV	ubump, TSV	Hybrid bonding	Hybrid bonding, Mini-TSV	
ubump / TSV pitch(um²)	-	48*55	96*110	-	1.5 * 1.5	Pitch更小
HB pitch	-	-	-	3	3	
是否需要PHY	-	Yes	Yes	No	No	去掉PHY，降低功耗
DRAM堆叠层数	-	8	8	1	2	堆叠层数少于HBM
每层DRAM容量（GB）	-	2	3	0.5	4	
存储容量（GB）	1	16	24	0.5	8	
IO数量	32	1024	1024	4096	131072	IO数量更多
IO速度(Mbps)	16384	4096	7168	266	541	IO速度慢，功耗低
总带宽(GBps)	64GBps	512GBps	896GBps	136GBps	8656GBps	
耗电量（相对值）	100%	80%	53%	12%	9%	低功耗性能显著
每Gb带宽（GBps/Gb）	8	4	4.7	34	135	单位容量的带宽更高

来源：IEEE，中泰证券研究所

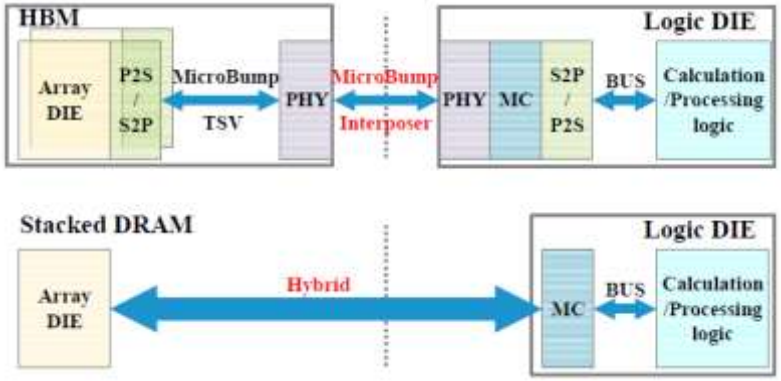
2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

■ **WOW 3D堆叠DRAM 对比HBM：定制化产品，带宽更高，功耗更低，但容量拓展性不如HBM。**

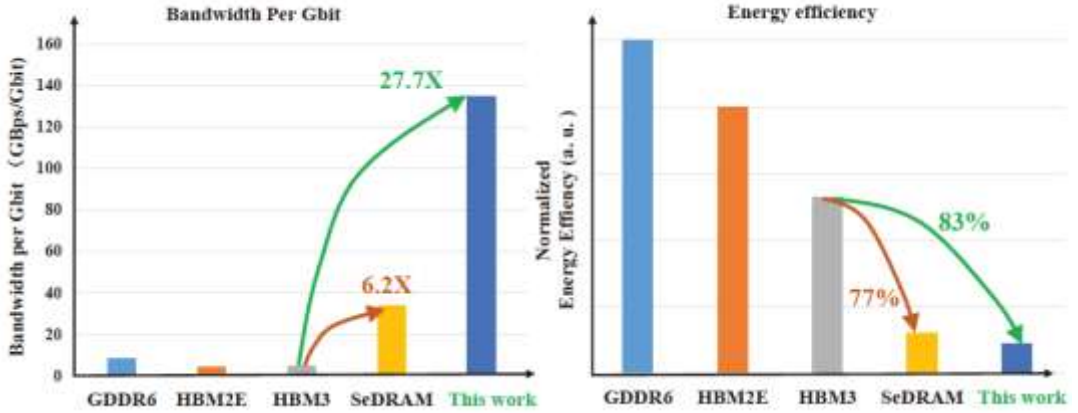
➢ 2、多方因素带来功耗更低。1）去掉PHY区域，减少时延和节省功耗：以紫光国芯SeDRAM为例，传统HBM互联结构中，DRAM和逻辑芯片中均有耗时且耗能的PHY，3D堆叠DRAM结构将此移除。2）IO速度慢：IO速度越大，传输信号的功耗越大，HBM3 IO速度达到7168Mbps，而紫光国芯的2层DRAM方案产品的IO速度为541Mbps。3）数据的传输路径短：相较2.5D封装结构，3D封装结构下存储和计算芯片之间的数据的传输路径变短，功耗低。4）混合键合功耗低：混合键合下直接进行铜对通的互连，没有锡焊球，直接铜导电，电阻更小，功耗更低。

➢ 3）HBM容量拓展性更好。每颗HBM、WOW 3D堆叠DRAM都可以通过堆叠层数和增加单层密度来提高容量；HBM与计算芯片采用2.5D封装，1颗计算芯片可以使用多颗HBM，如H100使用6颗HBM，而WOW 3D堆叠DRAM与计算芯片采用3D封装，1颗计算芯片只能配套1颗WOW 3D堆叠DRAM。

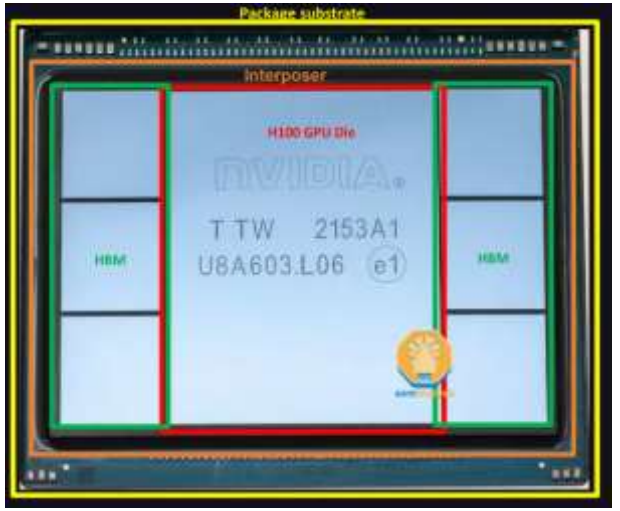
图表：3D堆叠DRAM中逻辑-DRAM的接口（对比HBM）



图表：带宽和功耗对比



图表：H100板卡图

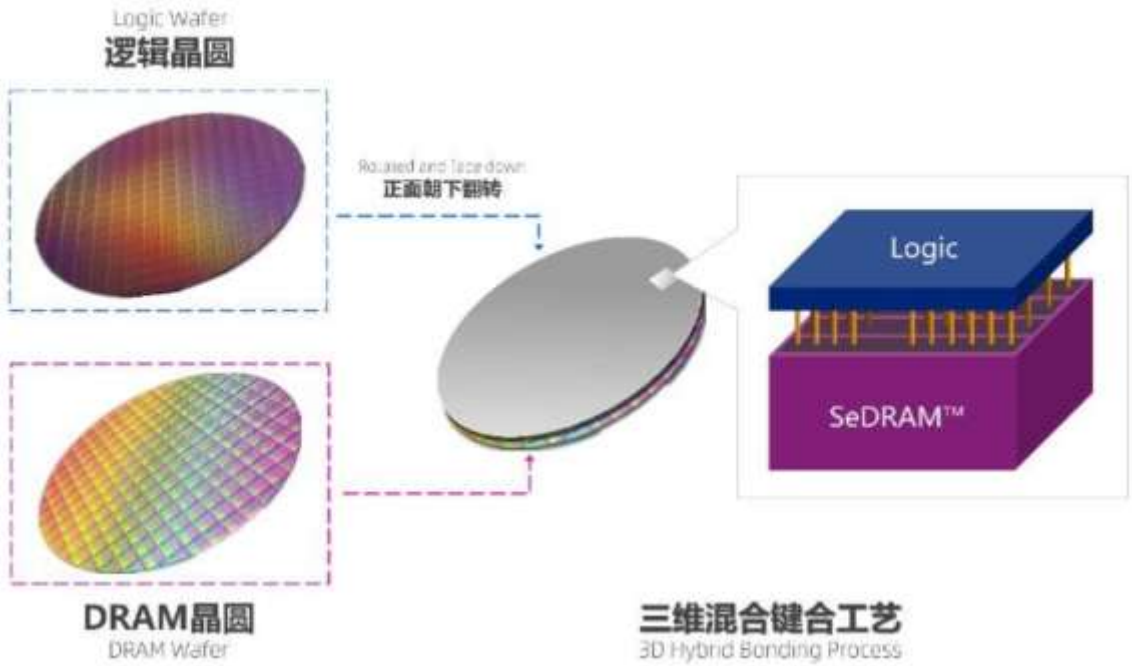


来源：IEEE，Semianalysis，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

- **中国台湾和大陆企业均有布局WOW 3D堆叠DRAM，矿机市场率先落地使用。**
- 大陆公司紫光国芯2020年量产WOW 3D堆叠DRAM。
  - 西安紫光国芯半导体股份有限公司前身为成立于2004年德国英飞凌西安研发中心的存储事业部，2006年分拆成为独立的奇梦达科技西安有限公司，2009年被浪潮集团收购转制成为国内公司并更名为西安华芯半导体有限公司。2015年，紫光集团紫光国芯微电子股份有限公司收购西安华芯半导体有限公司并更名为西安紫光国芯半导体有限公司。2019年12月，经过重组，西安紫光国芯半导体并入北京紫光存储科技有限公司。
  - 2020年紫光国芯WOW 3D堆叠DRAM产品（公司称为SeDRAM）量产问世，SeDRAM采用Wafer on Wafer混合键合工艺，相比HBM的MicroBump（微凸块）工艺，SeDRAM接触孔可达110,000个/mm2，实现了百倍量级的密度提升，而且连接电阻低至0.5欧姆。从而实现了从逻辑电路到存储阵列之间每Gbit高达34GB/s的带宽和0.88pJ/bit的能效。2021年已有产品量产。
  - 产品是基于西安紫光国芯SeDRAM™平台，由阿里巴巴达摩院计算技术实验室定制设计，西安紫光国芯SoC团队完成芯片实现以及系统级测试支持。供应链：DRAM芯片代工，力积电；混合键合，武汉新芯。

图表：紫光国芯的WOW 3D堆叠DRAM



来源：紫光国芯，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

- **中国台湾和大陆企业均有布局WOW 3D堆叠DRAM，矿机市场率先落地使用。**
- 中国台湾公司爱普存储2021年发布WOW 3D堆叠DRAM方案。
  - 中国台湾上市公司爱普存储，2021年宣布成功实现WOW 3D堆叠高带宽存储方案（VHM），即DRAM与逻辑晶片的真3D堆叠异质整合，透过WoW（Wafer on Wafer）的多点I/O连接，每GB DRAM可以提供超过4TB/s的带宽并有极出色的耗能表现，而全窗大小的VHM则可提供高达24TB/s的频宽供SoC的运算需求。
  - 供应链：爱普科技提供VHM TM，包含客制化DRAM设计及DRAM与逻辑晶片整合介面之VHM TM LInK IP；DRAM芯片代工，力积电；逻辑芯片的代工和3D堆叠键合工艺，台积电。

图表：爱普存储的WOW 3D堆叠DRAM

VHM™

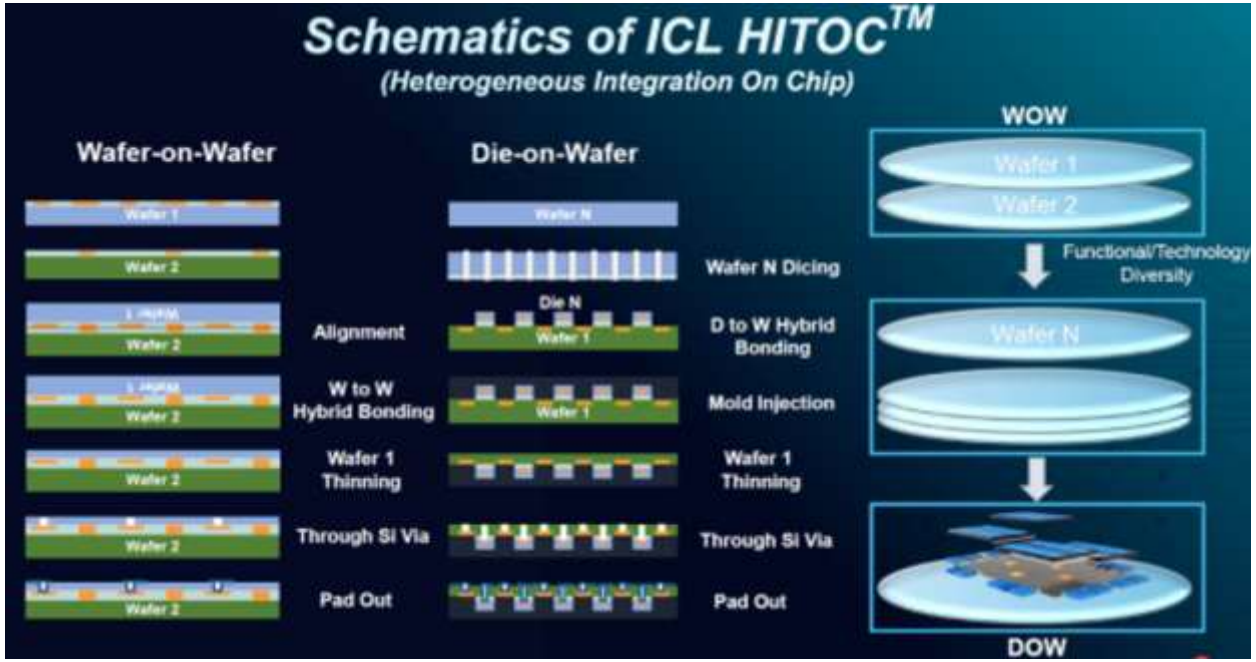
VHM™是第一个透过WoW 3D 堆叠的宽频记忆体解决方案。透过WoW 的多点IO 连结，及爱普自订的通讯协定，每GB 的VHM™可以提供超过4TB/s 的频宽并且有极为出色的耗能表现。而全窗大小的VHM™则可提供高达24TB/s 的频宽供SoC 的运算需求。

来源：爱普科技官网，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

- **中国台湾和大陆企业均有布局WOW 3D堆叠DRAM，矿机市场率先落地使用。**
- 大陆公司芯盟科技2022年宣布WOW 3D堆叠DRAM方案。
  - 2022年中国国际半导体技术大会中，芯盟科技宣布了基于 HITOC 技术的 3D DRAM 架构的问世，HITOC 技术（Heterogeneous Integration Technology on Chip）技术是运用Wafer-on-Wafer和Die-on-Wafer混合键合（Hybrid Bonding）制造工艺，将不同类型的 wafer 或 die 上下对准贴合，以实现真正的三维异构单芯片集成，2022年1层DRAM WOW的方案已导入市场，Die-on-Wafer 和多层WOW的3D 堆叠产品处于研发状态。

图表：芯盟科技的WOW 3D堆叠DRAM



来源：腾讯网，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

- **中国台湾和大陆企业均有布局WOW 3D堆叠DRAM，矿机市场率先落地使用。**
- 大陆公司兆易创新布局定制化DRAM。
  - 兆易创新是大陆存储龙头公司，在利基存储市场，布局NOR、SLC NAND和利基DRAM，同时积极布局定制化DRAM业务，2024年成立子公司青耘科技布局该定制化存储领域（兆易直接持股78%）。

图表：大陆存储公司布局情况

类别		主流DRAM	主流NAND	利基DRAM	SLC NAND	Nor Flash	EEPROM	SRAM	MCU
	市场规模（亿美金）	700	448	78	23	26	8	4	250
存储设计	兆易创新			√	√	√			√
	北京君正			√	√	√		√	√
	普冉股份					√	√		√
	东芯股份			√	√	√			
	聚辰股份					√	√		
	恒烁股份					√			√
	博雅科技（未上市）					√			√
	芯天下（未上市）				√	√			
存储IDM	长鑫存储	√							
	长江存储		√						
存储模组	江波龙	嵌入式存储（52%）、移动存储（24%）、固态硬盘（18%）、内存条（5%）							
	佰维存储	嵌入式存储(73%)、消费级存储（21%）、工业级存储（3%）、先进封测服务（2%）							
	德明利	移动存储（50%）、存储晶圆及晶圆封装片（32%）、固态硬盘（16%）							
	朗科科技	闪存应用产品（58%）、闪存控制芯片及其他（38%）、移动存储产品（1%）							
存储封测	深科技	高端制造（72%）、存储半导体（16%）、计量智能终端（11%）							
存储代理	香农芯创	80%海力士存储器产品、20%联发科产品							
内存接口芯片	澜起科技	74%互连芯片，26%津逮CPU							

注：收入占比为2023年数据；市场规模为2022年数据，利基DRAM市场规模按照DRAM市场规模乘以10%计算，NOR市场规模按照存储市场规模乘以2%计算

来源：各公司官网、SIA等，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

■ **中国台湾和大陆企业均有布局WOW 3D堆叠DRAM，矿机市场率先落地使用。**

➢ 以太币因其算法需要加载DAG数据包，需要配套高带宽存储以提升挖矿效率，WOW 3D堆叠DRAM方案应用其中。2021年中科声龙第一代高通量算力芯片（Jasminer X4，中文为茉莉X4）首次流片即一次性流片成功，至今量产晶圆已逾万片，该芯片采用堆叠技术（DRAM和逻辑芯片 3D堆叠），芯片面积裸Die 678平方毫米，存储带宽1TByte/s，存储容量5GB，处理能力达到65MH/s（一款高端显卡的处理能力），但功耗只有23W，而显卡功耗一般在150W以上。

图表：中科声龙的茉莉X4

图表：茉莉X4算力砖

来源：中科声龙，中泰证券研究所

2.2 WOW 3D堆叠DRAM：AI低算力+超高宽带存储解决方案

■ **端侧部署本地大模型是趋势，WOW 3D堆叠DRAM可定制，高带宽、低功耗性能显著，。**

➢ 端侧模型本地化趋势：模型部署在端侧，既有高隐私性、安全性、可靠性等优点，同时可提供个性化服务，目前AI手机、AI PC、汽车等都有模型部署在本地的趋势。

➢ 端侧模型有轻量化趋势，DeepSeek降本促进端侧AI渗透：Deepseek是全开源模型，模型可蒸馏，从DeepSeek-R1蒸馏出的较小模型有1.5B、7B、14B、32B、70B，蒸馏出来的小模型性能好，降低了模型部署在手机PC等端侧的难度，但使用传统存储器仍面临存储墙问题。

➢ WOW 3D堆叠DRAM存储容量可定制，超高带宽、低功耗性能瞩目，高带宽带来低延时、快响应，契合端侧场景，带来更好的用户体验，潜在场景手机、PC、汽车和机器人等。

图表：端侧AI的优势

图表：DeepSeek蒸馏后的小模型

模型	基础模型
DeepSeek-R1-Distill-Qwen-1.5B	Qwen2.5-数学-1.5B
DeepSeek-R1-Distill-Qwen-7B	Qwen2.5-Math-7B
DeepSeek-R1-Distill-Llama-8B	骆驼-3.1-8B
DeepSeek-R1-Distill-Qwen-14B	Qwen2.5-14B
DeepSeek-R1-Distill-Qwen-32B	Qwen2.5-32B
DeepSeek-R1-Distill-Llama-70B	Llama-3.3-70B-指导

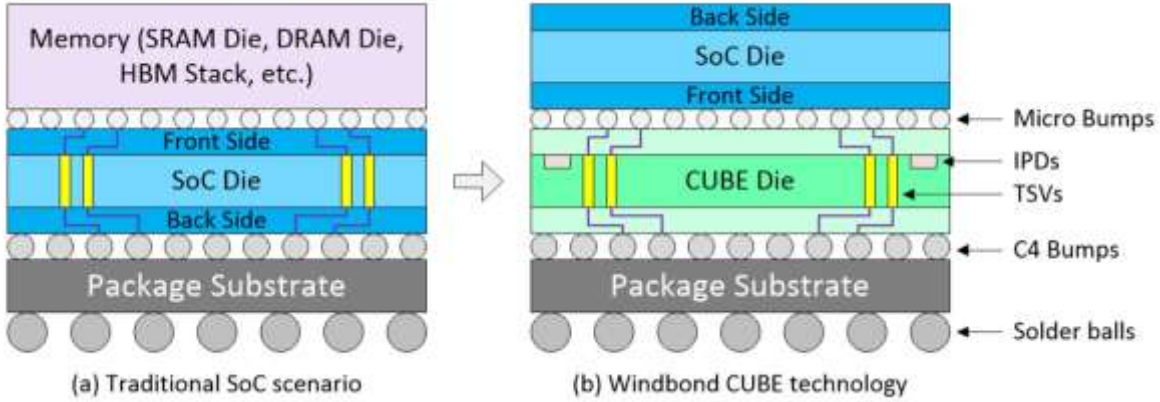
来源：AI赋能实验室等，中泰证券研究所

2.3 CUBE：AI低算力+高带宽存储解决方案

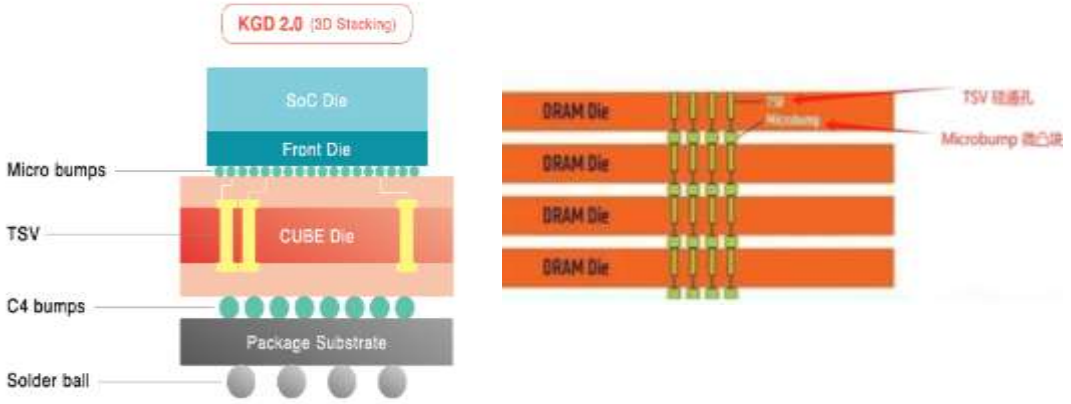
■ **华邦2023年宣布CUBE方案，定位边缘计算。**

- 2023年华邦宣布CUBE（Customized/Compact Ultra Bandwidth Elements）。
- 结构：属于近存计算，1层SOC和多层DRAM是上下堆叠结构，SoC放置在上面，DRAM芯片在下面，省去了SoC的TSV工艺，SOC无性能损失、系统成本更低，同时，3D DRAM TSV工艺可以将SoC信号引至外部，使它们成为同一颗芯片，进一步缩减了封装尺寸，同时SoC在上可以带来更好的散热效果。
- 技术：主要使用TSV和Microbump（微凸块）工艺，与目前HBM使用工艺相同。
- 供应链：联电负责CMOS晶圆制造和键合技术；华邦电导入客制化CUBE架构；智原提供全面的3D先进封装一站式服务，以及存储IP和ASIC小芯片设计服务；日月光则是提供晶圆切割、封装和测试服务，另外还有Cadence 负责晶圆对晶圆设计流程，提取TSV特性和签核认证。

图表：华邦CUBE方案的结构图



图表：华邦CUBE方案拆解



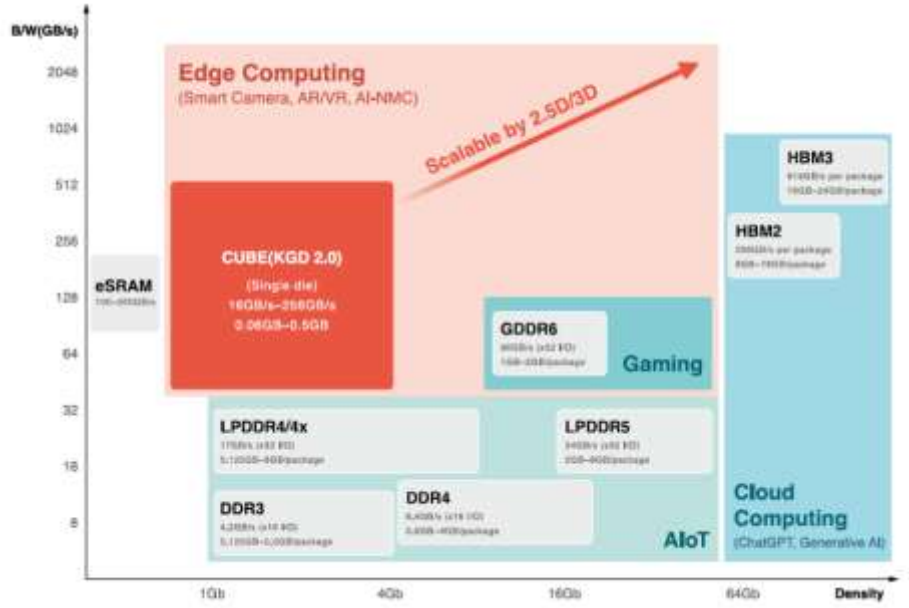
来源：华邦官网，Electronic Engineering，中泰证券研究所

2.3 CUBE：AI低算力+高带宽存储解决方案

■ **性能：功耗低于HBM，带宽小于目前的HBM3E。**

- 低功耗：CUBE 功耗低于 1pJ/bit，功耗优于HBM，CUBE能够确保延长运行时间并优化能源使用。
- 高带宽：CUBE的IO速度达到2Gbps，带宽提升主要来自IO数量提升，官网表示CUBE带宽32GB/s - 256GB/s（可以根据客户要求定制化），相当于HBM2带宽，也相当于4至32个LP-DDR4x 4266Mbps x16 IO，而公开演讲资料显示CUBEx的带宽可达到1TB/s，相当于HBM3E带宽。
- 面积小：SoC（不带TSV，置上）堆叠在 CUBE（带 TSV，置下）上，SOC去除 TSV 区域损失，其芯片尺寸可能会更小，能够为边缘AI设备带来更明显的成本优势。
- 散热好：SoC在上，散热会更好。
- 华邦DRAM制程较三星、海力士和美光落后，目前CUBE基于20nm制程，每片DRAM容量可以达到256Mb-1GB容量，2025年将有16nm。

图表：CUBE适用于边缘运算



图表：CUBE性能

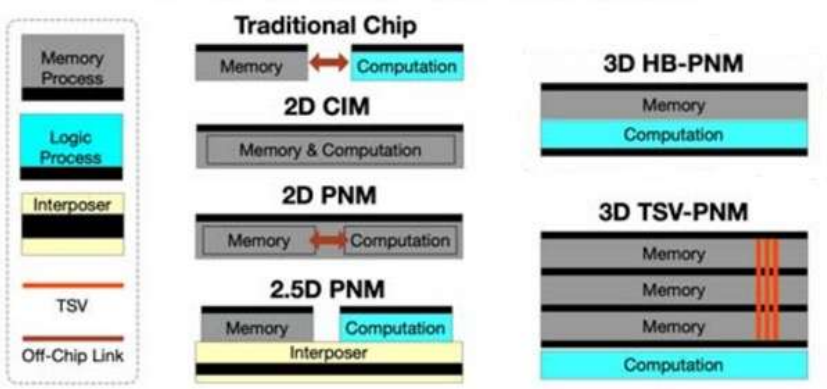
	HBM1	HBM2	HBM2E	HBM3	HBM3E	HBM4	CUBEx
层高	4层	8层	8层 or 12层	8层 or 12层	8层 or 12层	12层 or 16层	4层
I/O传输速率	1Gb/s	2.4Gb/s	3.6Gb/s	6.4Gb/s	9.2Gb/s	9.2Gb/s	2Gb/s
产品整体带宽	128GB/s	307 GB/s	460 GB/s	819 GB/s	1.2TB/s	2.4TB/s	1TB/s
单层颗粒 GB	0.25	1	2	2	3	4	0.5-1
产品容量	1GB	8GB	GB/16GB/24G	16GB/24GB	24GB/36GB	36GB/48GB	2-4GB
I/O数量（个）	1024	1024	1024	1024	1024	2048	4096
功率/bit	6pJ / bit	<5pJ / bit	<5pJ / bit	<4pJ / bit	<4pJ / bit	<4pJ / bit	<1pJ / bit
应用	AI/机器学习/HPC/数据中心等						边缘运算AI应用
制程							
三星				1znm	1anm		
海力士				1znm	1bnm		
美光					1bnm		
华邦							20nm

来源：华邦，中泰证券研究所

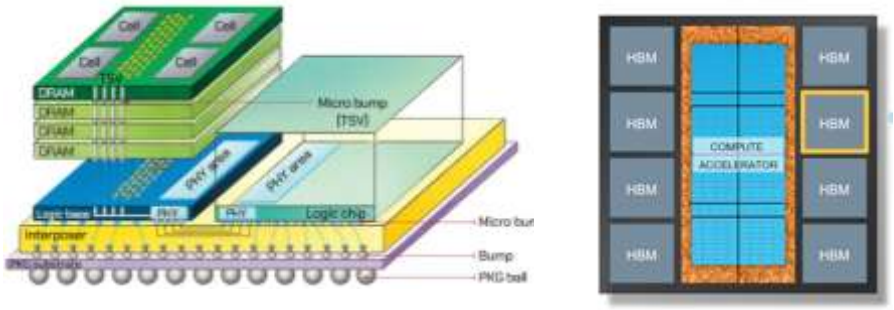
2.4 三种方案对比总结

- **封装级3D DRAM均属于近存计算，主要有HBM、WOW 3D堆叠DRAM和华邦CUBE三类。**以下为相同点：
  - 1）DRAM芯片3D堆叠：多层DRAM上下3D堆叠，通过TSV+键合方式等先进封装工艺实现电气连接；
  - 2）采用系统级封装：DRAM与逻辑芯片采用系统级封装工艺封装在一起，属于近存计算。但从封装工艺上，HBM与计算芯片是2.5D封装，右上角图中的2.5D PNM（Process near memory，近存计算），WOW 3D堆叠DRAM是3D封装，右上角图中的3D TSV-PNM，华邦CUBE是属于3D封装；
  - 3）相较传统DRAM均有高带宽、单位面积下容量更大、功耗更低的特点。
  - 4）HBM目前是标准化产品，而WOW 3D堆叠DRAM和华邦CUBE目前是定制化DRAM产品。

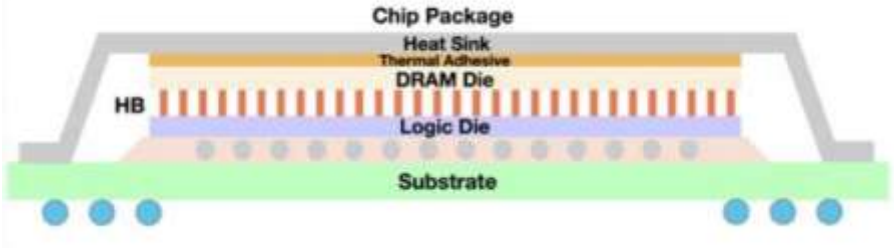
图表：近存计算的类型



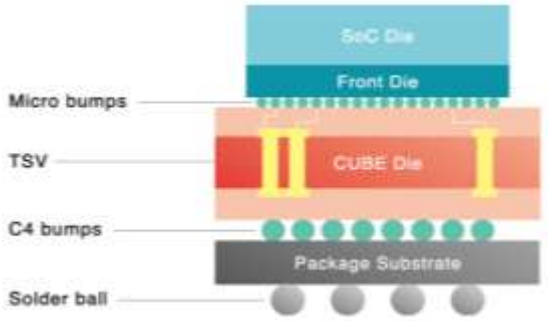
图表：HBM



图表：WOW 3D堆叠DRAM



图表：华邦CUBE方案



来源：海力士、紫光国芯和华邦，中泰证券研究所

2.4 三种方案对比总结

■ **在系统级封装工艺、键合方式等方面有差异，最终性能各有特色。HBM是标准化产品，兼顾高带宽和高容量，容量拓展性最好，而WOW 3D堆叠DRAM是高度定制化产品，带宽指标领先，容量拓展性不如HBM 。**

➢ 1、系统级封装工艺不同。计算芯片和存储芯片的位置关系：1）HBM与计算芯片放置在同一个水平平面，2.5D封装。2）WOW 3D 堆叠DRAM 堆叠在计算芯片上，DRAM在上、逻辑在下，3D封装。3）华邦CUBE方案是 逻辑芯片堆叠在DRAM上，DRAM在下、逻辑在上，3D封装。

➢ 2、HBM的容量拓展性更好。计算芯片可使用的DRAM CUBE颗数：1）1颗计算芯片可以使用多颗HBM（1:N配套，如H100使用8颗HBM），计算芯片可使用DRAM容量可以通过增加HBM颗数，或者通过增加单颗HBM容量实现（堆叠层数或增加单层容量密度，HBM目前是标准化产品），HBM的容量拓展性最好。2）1颗计算芯片使用1颗WOW 3D堆叠DRAM或华邦CUBE（1:1配套），计算芯片使用DRAM容量主要是通过增加单颗3D堆叠DRAM/华邦CUBE容量（堆叠层数或增加单层容量密度）。

➢ 3、是否定制化：HBM是标准化产品，标准由JEDEC确定；WOW 3D堆叠DRAM和华邦CUBE是与计算芯片直接上下堆叠，是高度定制化产品，DRAM面积、堆叠层数都可以根据客户要求定制，因此有稳定供货和积极配合的晶圆是关键。

图表：3类封装级3D DRAM的对比

类型	HBM	CUBE	WOW 3D堆叠DRAM
存储芯片与计算芯片的封装方式	2.5D（Cowos）	3D	3D
与逻辑芯片的连接方式	TSV+Microbump	TSV+Microbump	TSV+Hybird Bonding
是否需要中介层	是	否	否
逻辑芯片与GPU的对应关系	1:N	1:1	1:1
是否需要PHY	是	否	否
DRAM是否定制化	否	是	是
每层DRAM之间的连接方式	TSV+Microbump	TSV+Microbump	TSV+Hybrid Bonding
DRAM设计	三星、海力士、美光等	华邦	紫光国芯、芯盟科技、爱普科技、兆易创新等
DRAM制造	三星、海力士、美光等	华邦	力积电等
计算芯片的代工	台积电等	联电	-
系统级封装	台积电	联电	武汉新芯等
Pitch	>10μm	>10μm	<10μm
下游应用	云	端+云	端+云

来源：IEEE、海力士、三星、华邦、紫光国芯等，中泰证券研究所整理

2.4 三种方案对比总结

■ **在系统级封装工艺、键合方式等方面有差异，最终性能各有特色。HBM是标准化产品，兼顾高带宽和高容量，容量拓展性最好，而WOW 3D堆叠DRAM是高度定制化产品，带宽指标领先，容量拓展性不如HBM。**

➢ 4、WOW 3D堆叠DRAM，使用混合键合工艺，IO数量更多，带宽更高。1）HBM与GPU通过中介层+Microbump+TSV连接，HBM与GPU之间信号通过PHY转换。2）3D堆叠DRAM/华邦CUBE，信号不需要通过PHY转换，3D堆叠DRAM与计算芯片通过混合键合+TSV连接，华邦CUBE与计算芯片通过Microbump+TSV连接，均不需要中介层。

  ➢ 相较Microbump，混合键合的通孔间距更小（Pitch更小），混合键合工艺的Pitch小于10μm，Microbump的Pitch大于10μm，混合键合工艺创造的信号传输通道更多（IO数量多），虽然IO速度降低，但IO数量的量级提升更快，最终实现更高带宽。

➢ 5、性能差异：带宽，WOW 3D堆叠DRAM>HBM，华邦CUBE带宽在HBM2及HBM3E之间；功耗，3D堆叠DRAM、华邦CUBE<HBM；容量拓展性，HBM>3D 堆叠DRAM、华邦CUBE。对比WOW 3D堆叠DRAM和华邦CUBE，均是计算芯片与DRAM上下堆叠的结构同时定制化属性，但在带宽层面，因华邦CUBE仍使用Microbump，IO数量小，因此带宽小于WOW 3D堆叠DRAM；对比WOW 3D堆叠DRAM和HBM，WOW 3D堆叠DRAM的带宽和功耗均优于HBM，1颗计算芯片可以使用多颗HBM，1颗计算芯片只能配套1颗WOW 3D堆叠DRAM，因此HBM的容量拓展性更好。

图表：3类封装级3D DRAM的性能对比

	传统DRAM	HBM			WOW 3D堆叠DRAM		CUBE	说明
类型	GDDR6 ISSCC2018	HBM2E ISSCC2020	HBM3 ISSCC2022	HBM3E（标准化）	SeDRAM IEDM2020	SeDRAM（2层）	CUBE	
芯片面积（mm^2）		约100	约100	约100	-	-	约100	
连接方式	-	ubump, TSV	ubump, TSV	ubump, TSV	Hybrid bonding	Hybrid bonding, Mini-TSV	ubump, TSV	
ubump / TSV pitch(um²)	-	48*55	96*110		-	1.5 * 1.5		
HB pitch	-	-	-		3	3	-	
是否需要PHY	-	Yes	Yes	Yes	No	No	-	
DRAM堆叠层数	-	8	8	12	1	2	4	
每层DRAM容量（GB）	-	2	3	3	0.5	4	0.5~1	
存储容量（GB）	1	16	24	36	0.5	8	2~4	
IO数量	32	1024	1024	1024	4096	131072	4096	
IO速度(Mbps)	16384	4096	7168	9421	266	541	2048	
总带宽(GBps)	64GBps	512GBps	896GBps	1229GBps	136GBps	8656GBps	1024GBps	
耗电量（相对值）	100%	80%	53%		12%	9%		WoW 3D堆叠DRAM功耗低
功率/bit		<5pJ / bit	<4pJ / bit	<4pJ / bit			<1pJ / bit	
每Gb带宽（GBps/Gb）	8	4	4.7	4.3	34	135	-	WoW 3D堆叠DRAM带宽高

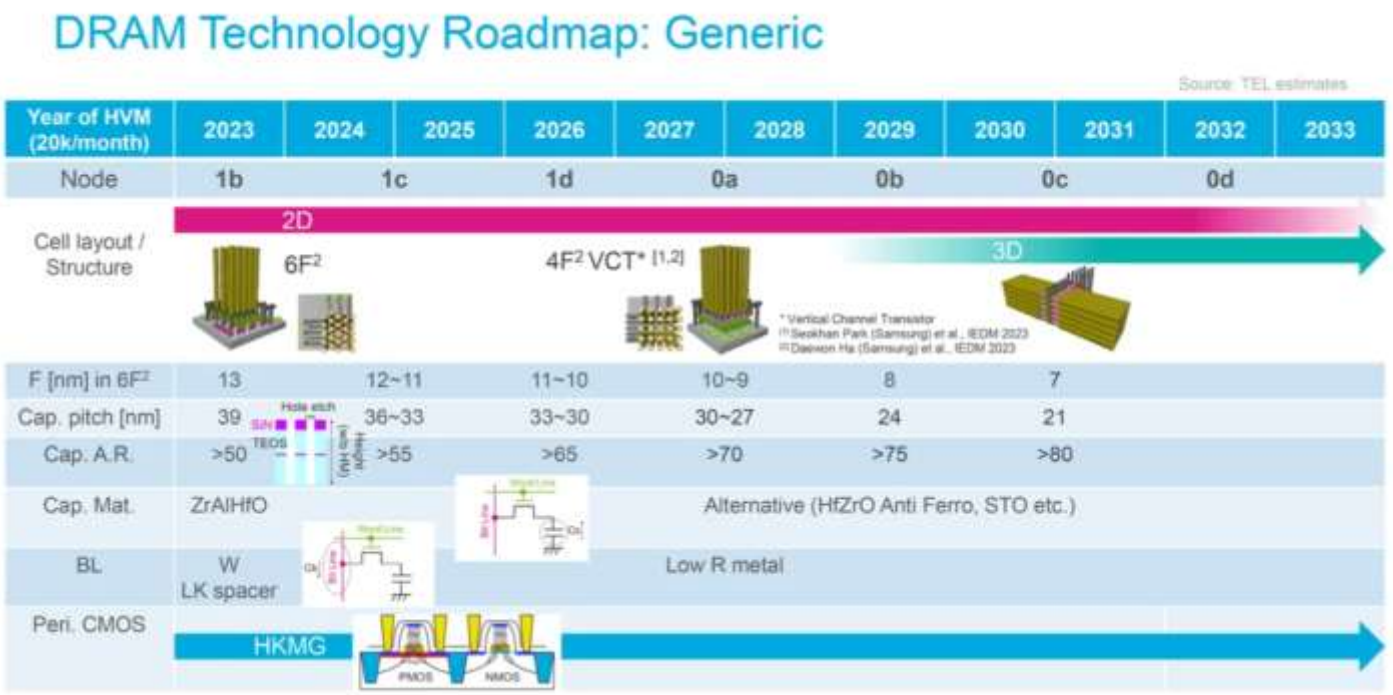
来源：紫光国芯、华邦，中泰证券研究所整理

目 录

3.1 三星&海力士：探索电容水平放置方案

- **2D DRAM主要通过水平方向的制程升级来提升单位面积下的存储密度，而晶圆级3D DRAM是通过堆叠层数来升级。目前DRAM制程迭代到12nm左右（1bnm），1cnm将到10nm，进入0nm级别后，预计DRAM开启晶圆级3D之路。3D DRAM目前各家处于实验室状态，探索多种技术路径，目前仅三星公布规划图。**
- **目前晶圆级3D DRAM仍处于研发阶段，主要是2个方案。**
  - 1）方案一：存储单元仍是基于1T1C结构（1个电容器+1个晶体管），主要改变存储单元各个组成部分的结构。传统2D DRAM的存储单元中，电容器是垂直方向，3D DRAM将垂直的电容水平放置，然后进行堆叠。三星、海力士和长鑫存储均探索这个方案。
  - 2）方案二，无电容方案：存储单元中去掉电容器，然后进行堆叠。美国公司NEO探索这个方案。

图表：DRAM发展路线



图表：3D DRAM的层数迭代

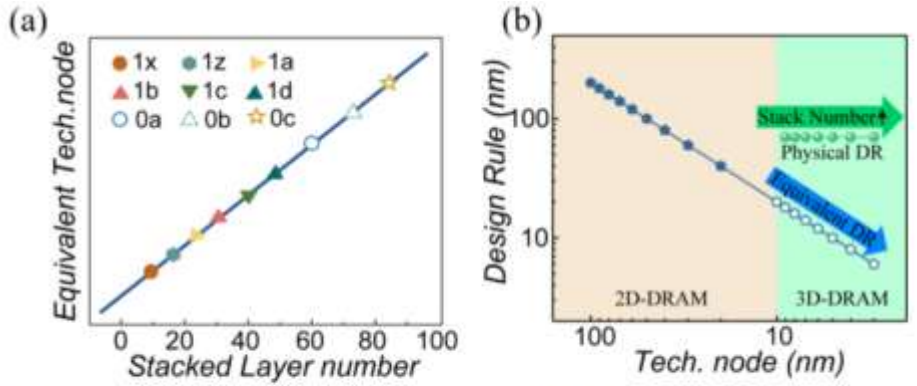
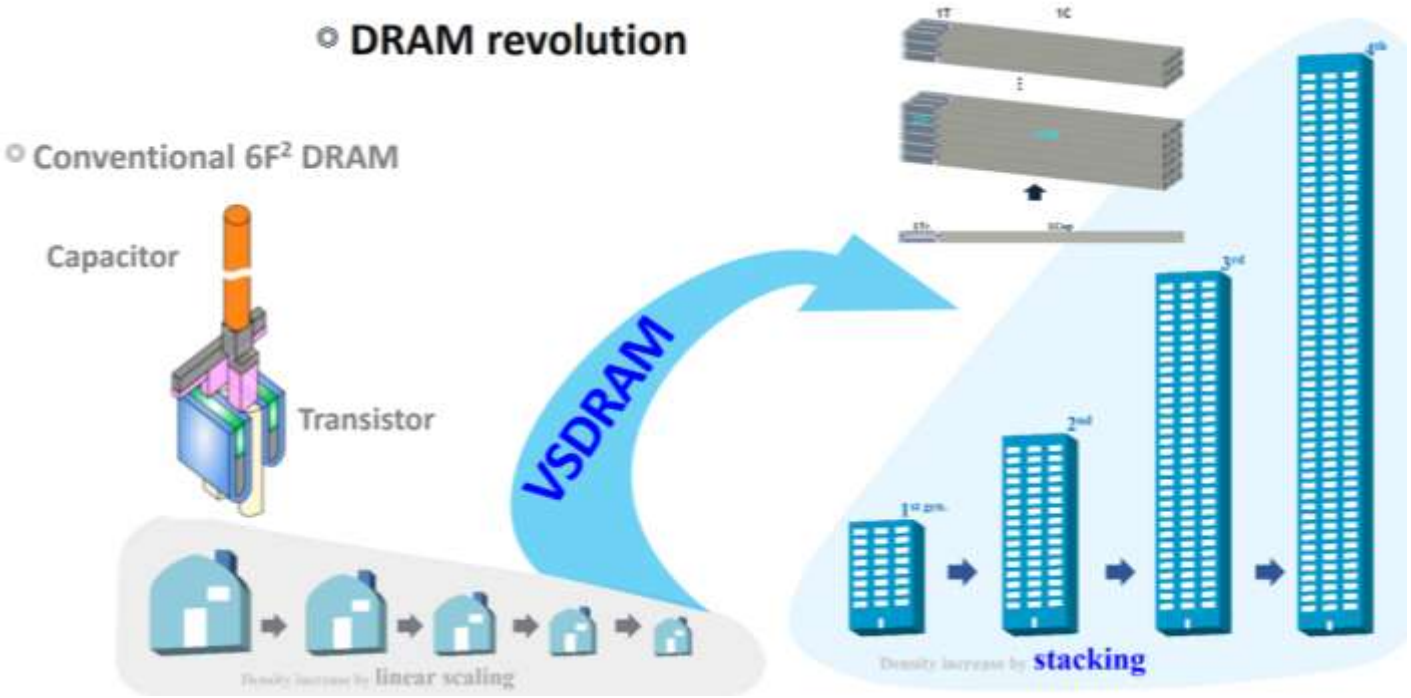


Fig. 2 (a). Equivalent technological node of 3D DRAM as stacked layers increasing; (b) Design rule vs. technology node for 2D-DRAM and 3D-DRAM.

来源：TEL、长鑫存储，中泰证券研究所

3.1 三星&海力士：探索电容水平放置方案

图表：DRAM从2D到3D（方案一）



来源：AMAT、TEL，中泰证券研究所

3.1 三星：探索电容水平放置方案

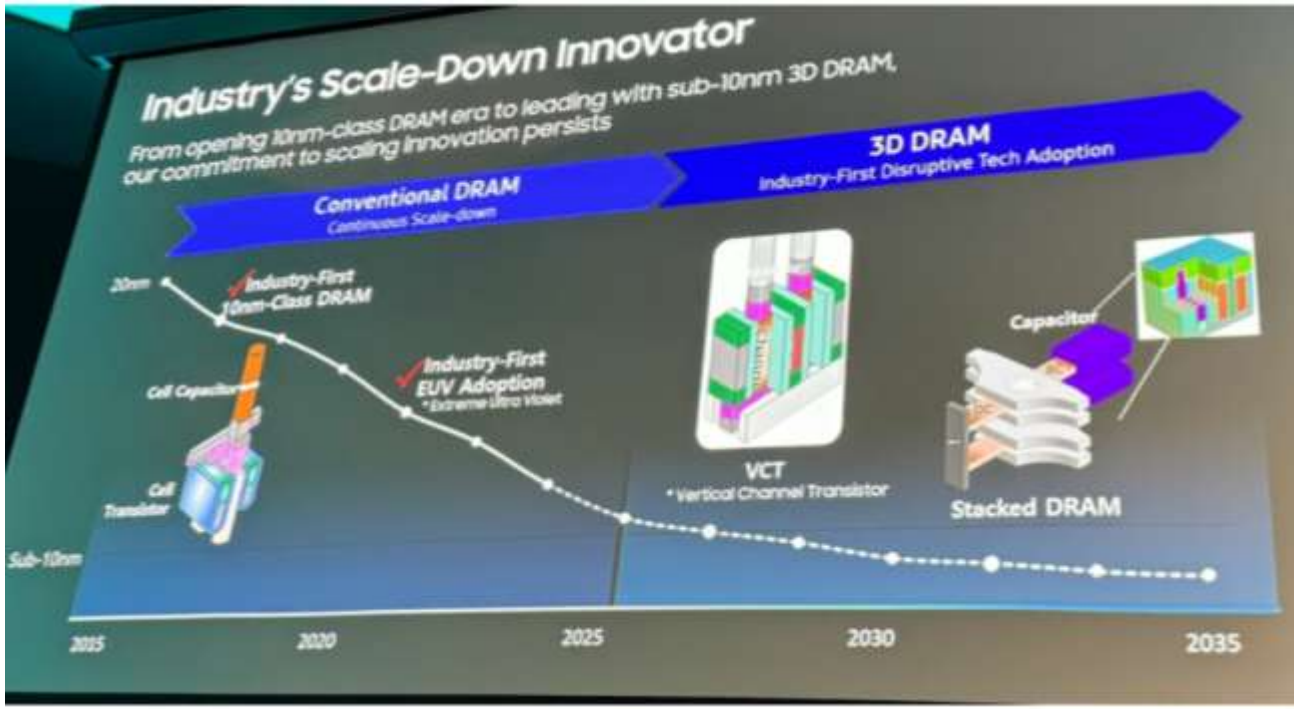
■ **三星2024年公布3D DRAM规划图，预计2030年采用晶圆级3D DRAM。**

➢ 三星在2024年的Memcon大会上正式公布3D DRAM技术路线图，3D DRAM首次被纳入规划之中，这标志着该技术正逐步从实验室走向实际生产的阶段。规划图上公布了三星在研的2种新DRAM结构： 4F2 VCT DRAM 和VS-CAT DRAM。

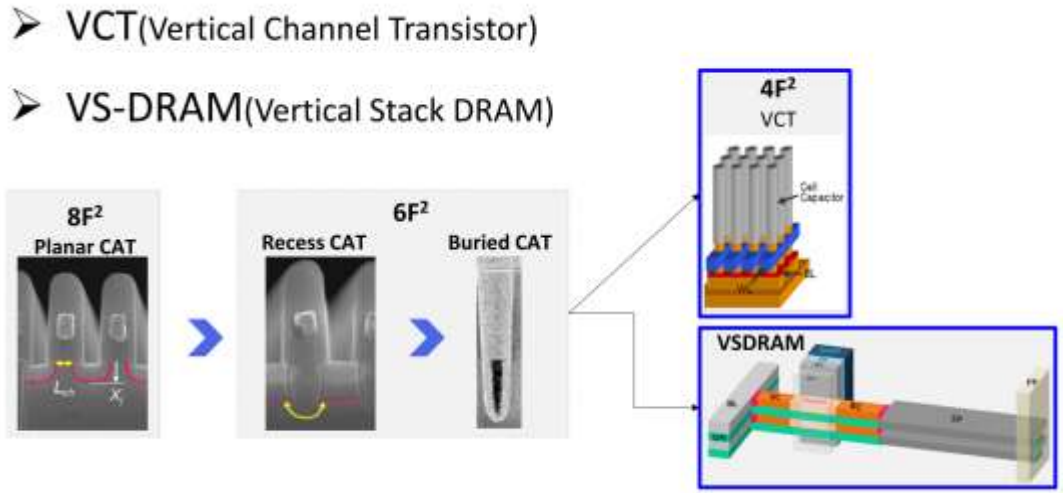
➢ 1）4F2 VCT DRAM，将晶体管从水平方向变为垂直方向，容量主要还是通过水平方向升级，预计2025年开始采用。

➢ 2）VS-CAT DRAM，将垂直的电容器变成水平方向，容量通过堆叠升级，预计2030年开始采用。

图表：三星DRAM发展路线



图表：三星的两类DRAM

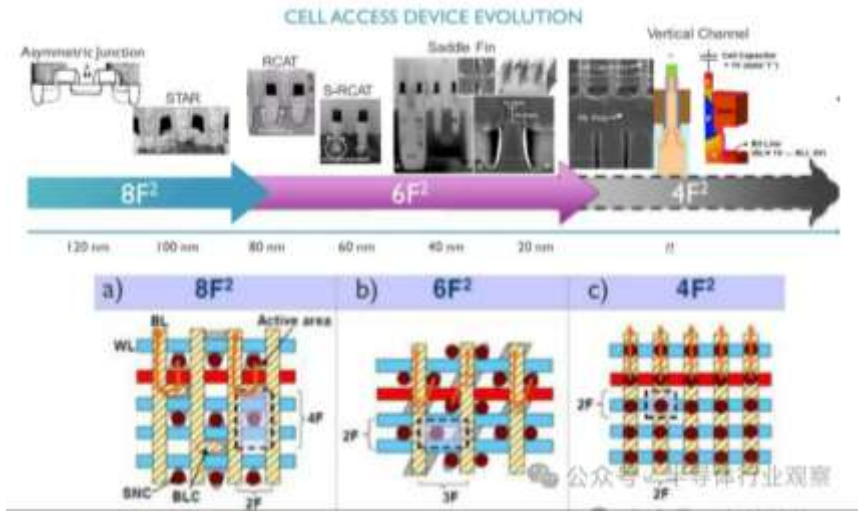


来源：Semiconductor Engineering，三星，中泰证券研究所

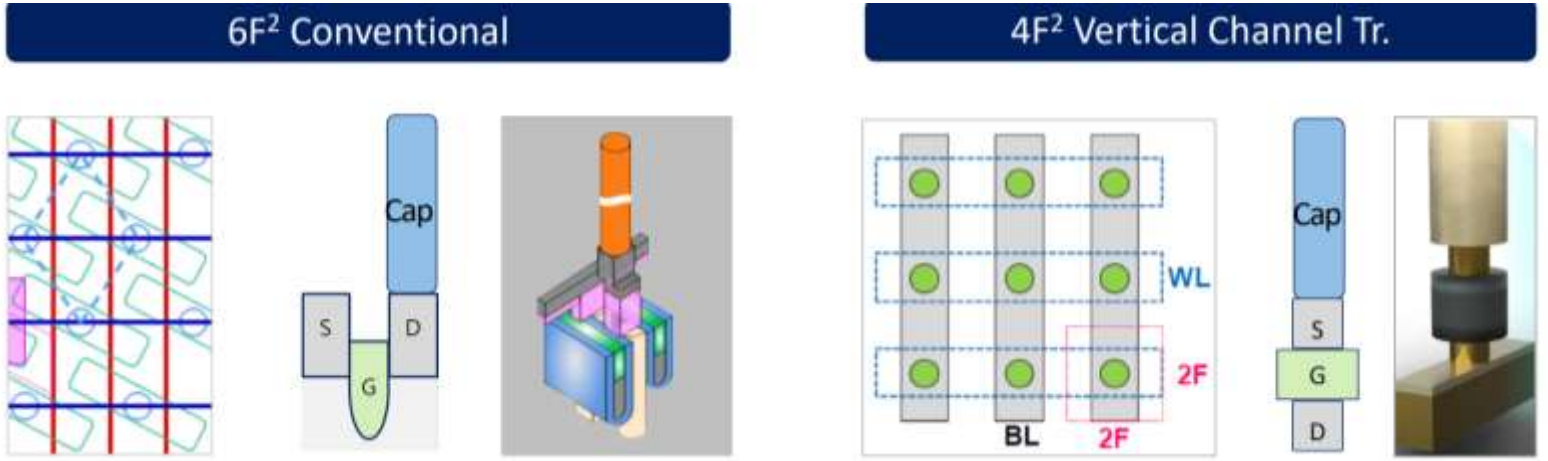
3.1 三星：探索电容水平放置方案

- **VCT DRAM，将晶体管垂直堆叠利用Z轴空间，是真正3D DRAM的过渡方案。**
  - 4F2 VCT DRAM：Vertical Channel Transistor，垂直通道晶体管，相较传统的水平方向的晶体管结构，VCT将晶体管变为垂直方向，存储单元结构向Z方向发展，4F2 DRAM单元尺寸比现有的6F2 DRAM减少约30%，在提高能效的同时大幅降低了单元面积，但同时提升了对刻蚀工艺精度的要求。三星预计2025年在内部发布并推进。

图表：存储单元的迭代路线



图表： 4F² VCT DRAM与传统6F² DRAM的对比

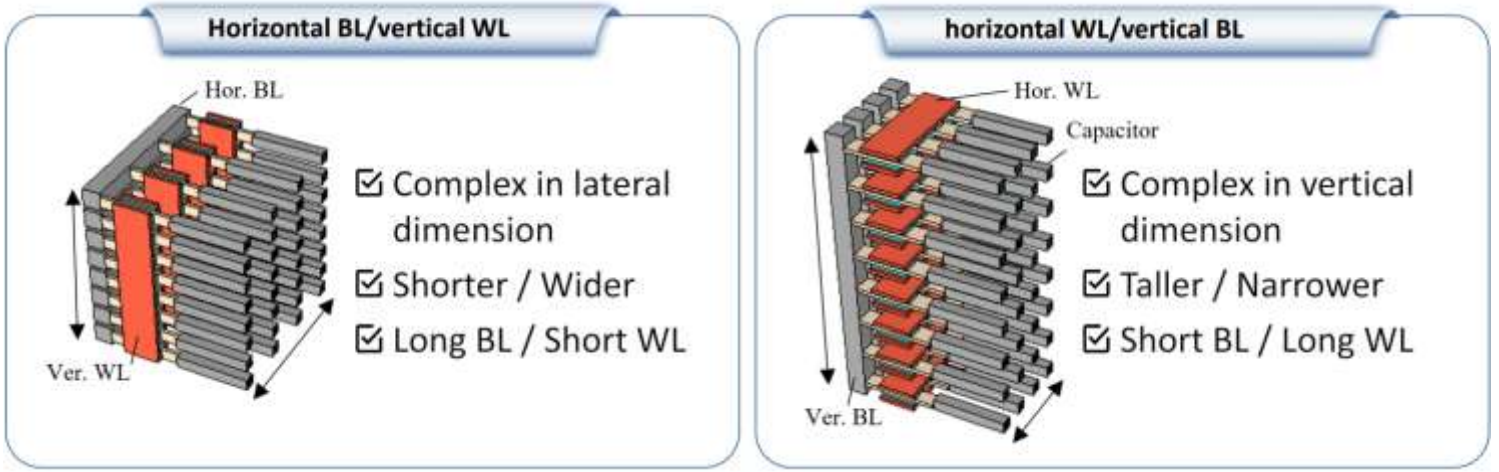


来源：三星，半导体行业观察，中泰证券研究所

3.1 三星：探索电容水平放置方案

- **VS DRAM，真正的3D DRAM，电容器水平放置，三星目前有2种细分方案（垂直Wordline和垂直Bitline），目前有16层的内部方案，预计2030年开始大规模采用。**
  - VS-CAT DRAM：Vertical Stacked-Cell Array Transistor，垂直堆叠单元阵列晶体管，类似 3D NAND 一样堆叠多层 DRAM。传统2D DRAM，电容器垂直放置，VS DRAM将电容器水平放置，三星目前展示了垂直Wordline（vertical wordline）和垂直Bitline（vertical bitline）两种潜在方案。另外预计采用存储单元和外围逻辑单元分离的双晶圆结构，在分别完成存储单元晶圆和逻辑单元晶圆的生产后，需要进行晶圆对晶圆（W2W）混合键合，然后得到 VS-CAT DRAM成品。根据新闻，三星2024年已在内部实现了16层堆叠的VS-CAT DRAM，三星预计2030年前推出市场。

图表： 三星的VS DRAM有两种潜在方案（垂直Wordline或者垂直Bitline）



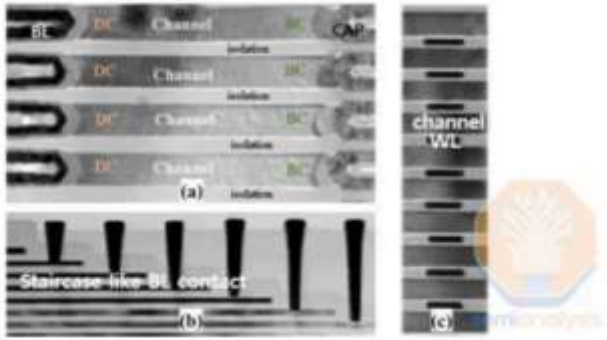


Fig. 5 TEM images of fabricated VS-DRAM. Crossectional view of (**a**) channel and (**b**) staircare BL region of vertical WL type, and (**c**) channel region of vertical BL type.

Source: Samsung VLSI 2023

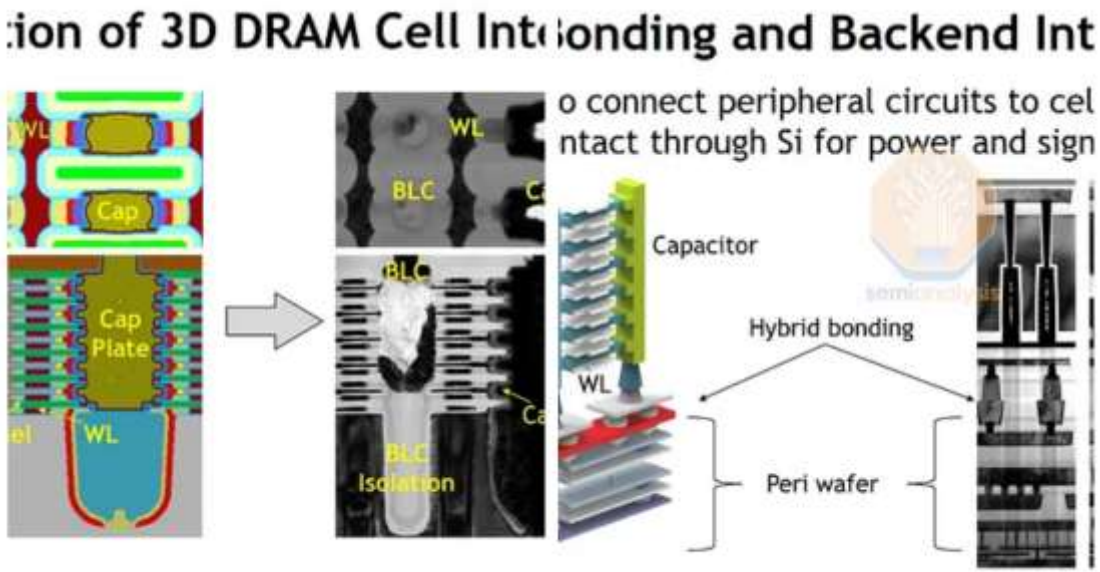
来源：三星， Semianalysis，中泰证券研究所

3.1 海力士：探索电容水平放置方案

■ **海力士的路线图暂未公布，目前有5层的堆叠产品方案，海力士选择垂直Bitline架构。**

➢ 海力士在VLSI 2024上详细展示了五层垂直Bitline（BL），并表示在即将到来的1c和1d节点之后，工艺整合和缩放的挑战将促使引入3D技术，预计这一转型需要大约5年时间。海力士认为垂直Bitline是更合理的架构选择，因为它能提供更大的感测裕度。海力士的5层3D DRAM也是采用混合键合连接的存储阵列和外围电路的结构。海力士认为，要达到广泛应用的目标，需要进一步提升3D DRAM的堆栈层数，实现32层至192层堆栈的存储单元。

图表： 海力士的5层3D DRAM



SK Hynix first demonstration of 5-layer 3D DRAM. Source: SK Hynix VLSI 2024:

来源： Semianalysis，中泰证券研究所

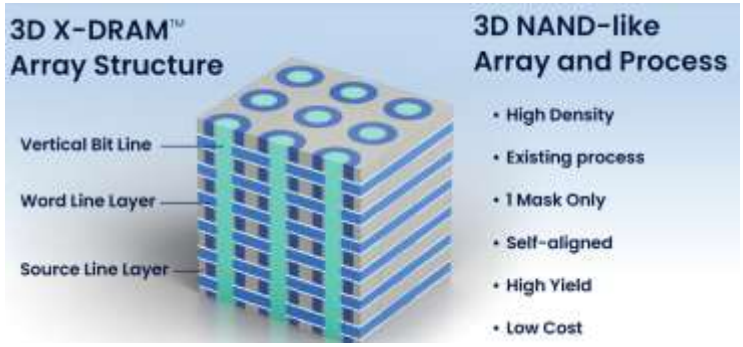
3.2 NEO公司：探索无电容方案

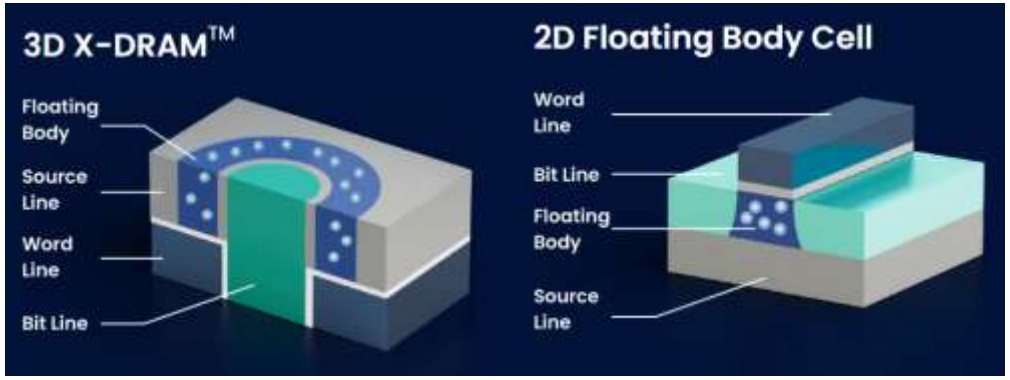
■ **NEO：采用无电容方案。**

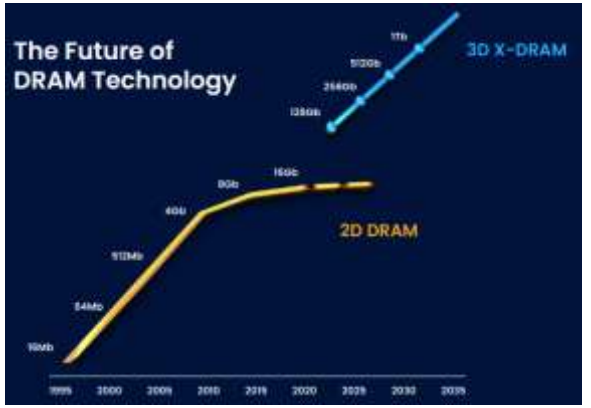
➢ 2023年美国存储公司NEO公布3D-X DRAM技术，3D X-DRAM具有基于无电容器浮栅极 (FBC) 技术的类 3D NAND DRAM 单元阵列结构。这种 FBC浮栅极技术使用一个晶体管和零个电容器将数据存储为电荷。NEO半导体表示它可以使用当前的 3D NAND 类工艺制造，并且只需要增加一层光罩掩模来定义位线孔并在孔内形成垂直结构，这提供了一种高速、高密度、低成本和高产量的制造解决方案。

➢ 据 Neo 的估计，3D X-DRAM技术可以实现 230层128 Gb 密度，这是当今 DRAM 密度的 8 倍。NEO提出，每10年容量提升8倍的目标，将在2030到2035年间实现1Tb的容量，较现DRAM核心容量达64倍提升，满足ChatGPT等AI应用对高性能和大容量存储器半导体的增长需求。

图表：NEO公司的3D DRAM







来源：全球半导体行业观察，中泰证券研究所

3.3 其他公司：积极探索

■ **美光3D DRAM的相关信息较少，但积极探索该方向。**

- 据TechInsights称，美光在2019年就开始了3D DRAM的研究工作。在2022年9月接受采访的时候，美光确认正在探索3D DARM的方案。
- 截止2022年8月，美光已获得了30多项3D DRAM专利。相比之下，美光专利数量是三星和SK海力士这两家韩国芯片制造商的两三倍。
- Yole强调，美光提交了与三星不同的3D DRAM专利申请。美光的方法是在不放置Cell的情况下改变晶体管和电容器的形状。

来源： Semianalysis，中泰证券研究所

3.3 其他公司：积极探索

■ **长鑫存储：电容水平放置，积极探索。**

➢ 2023长鑫存储发布3D DRAM相关研究工作的论文，《A 3D Stackable 1T1C DRAM: Architecture, Process Integration and Circuit Simulation》，其研究也是基于1T1C结构，电容水平放置。

图表：长鑫存储的3D DRAM

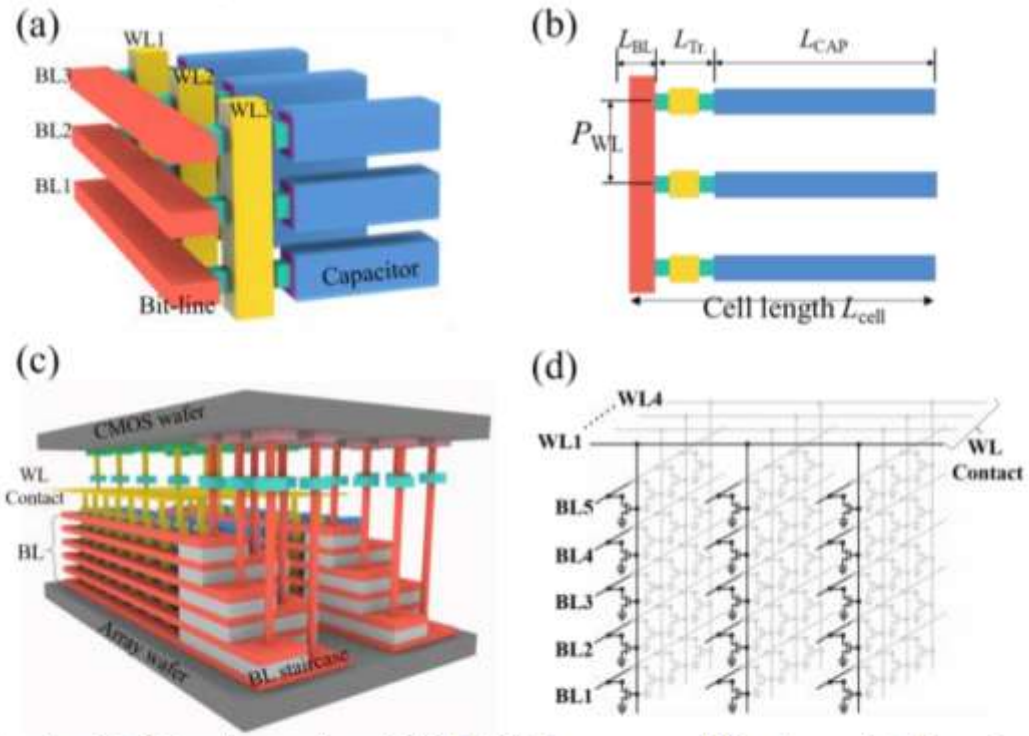


Fig. 1. (a) 3D schematic of 3D DRAM array architecture; (b) Top down view of 3D-DRAM; (c) Bird's eye view of 3D-DRAM; (d) Equivalent circuits of 3D-DRAM.

来源：长鑫存储，中泰证券研究所

目 录

四、投资建议

■ **近存计算的3D DRAM已成为产业趋势，高带宽、低功耗契合AI场景需求，建议关注产业链公司：**

- 存储：兆易创新、北京君正等
- SOC：瑞芯微、小米等
- 先进封装相关：长电科技、通富微电、甬矽电子、晶方科技、精智达、拓荆科技、芯源微、华海诚科、赛腾股份等

目 录

五、风险提示

- 行业需求不及预期的风险；
- 大陆厂商技术进步不及预期；
- 研报使用的信息更新不及时的风险等

重要声明

- **中泰证券股份有限公司（以下简称“本公司”）具有中国证券监督管理委员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。**
- 本报告基于本公司及其研究人员认为可信的公开资料或实地调研资料，反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响。本公司力求但不保证这些信息的准确性和完整性，且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，可能会随时调整。本公司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。本报告所载的资料、工具、意见、信息及推测只提供给客户作参考之用，不构成任何投资、法律、会计或税务的最终操作建议，本公司不就报告中的内容对最终操作建议做出任何担保。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。
- 市场有风险，投资需谨慎。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。
- 投资者应注意，在法律允许的情况下，本公司及其本公司的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并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。本公司及其本公司的关联机构或个人可能在本报告公开发布之前已经使用或了解其中的信息。
-