浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

从算力到存力：存储芯片研究框架
——AI行业系列报告

行业评级：看好

2023年4月3日

分析师	陈杭	研究助理	安子超
邮箱	chenhang@stocke.com.cn	邮箱	anzichao@stocke.com.cn
证书编号	S1230522110004	电话	18611396466

摘要

2023年3月31日，我国发起对美光在华销售产品的网络安全审查，体现出存储产业安全的重要性。此外，AI算力需求拉动高算力服务器出货，而AI服务器的存力需求更强，AI将驱动“从算力到存力”的中长期需求：

1、海外厂商占据绝对份额，国内存储安全重要性凸显

- **存力的底层支撑：**半导体存储器芯片（主流为DRAM+NAND Flash）。存力的体现形式：数据中心+存储服务器。
- **海外巨头垄断，国内存储安全重要性日益凸显。**全球DRAM市场几乎由三星、SK海力士和美光所垄断，CR3 超过 95%，全球NAND flash市场由前三大厂商分别为三星、铠侠和海力士，目前 CR3 市场份额达 65%，CR6 市场份额接近 95%。

2、国内数据圈庞大，AI驱动“从算力到存力”的长期需求

- 得益于人工智能、物联网、云计算等新兴技术的快速发展，中国数据正在迎来爆发式增长，驱动存储设备在数据中心采购占比进一步提升。据IDC预测，预计到2025年，中国数据圈将增长至48.6ZB，占全球数据圈的27.8%，成为全球最大的数据圈。
- AI技术革命推动高算力服务器等基础设施需求提升，**AI服务器所需的DRAM/NAND分别是常规服务器的8/3倍。**

3、存储周期拐点已至，库存改善、价格压力缓解

- 美光23Q1存货环比小幅回落，集邦咨询预测23Q2DRAM价格跌幅收窄至10%-15%（23Q1为20%），库存情况改善、价格压力缓解，存储行业周期迎来拐点。

4、先进存力的前进方向：存算一体、HBM/DRAM、3D NAND

- **存算一体：**将存储单元和计算单元合为一体，省去了计算的数据搬运环节，消除由于数据搬运带来的功耗，提升计算能效。
- **HBM/DRAM：**作为存储器主流之一的DRAM技术不断升级，衍生出HBM（高带宽内存），其是一款新型的CPU/GPU 内存芯片，将多个DDR芯片堆叠后与GPU封装在一起，实现大容量，高位宽的DDR组合阵列， 突破内存容量与带宽瓶颈。
- **3D NAND**（立体堆叠技术）：可以摆脱对先进制程工艺的束缚，不依赖于EUV技术，而闪存的容量/性能/可靠性也有了保障。

建议关注标的：兆易创新、北京君正、澜起科技、深科技、东芯股份、聚辰股份、普冉股份、江波龙、佰维存储、德明利
易华录、朗科科技、恒烁股份、同有科技、雅创电子

风险提示

1、宏观经济下行风险
2、上游晶圆紧缺加剧的风险
3、市场发展不及预期的风险
4、技术发展不及预期风险

01 存力概况

存力主要由数据中心和存储服务器驱动

数据中心

存储服务器

体现形式

存力

底层支撑

半导体存储器芯片

存储技术架构

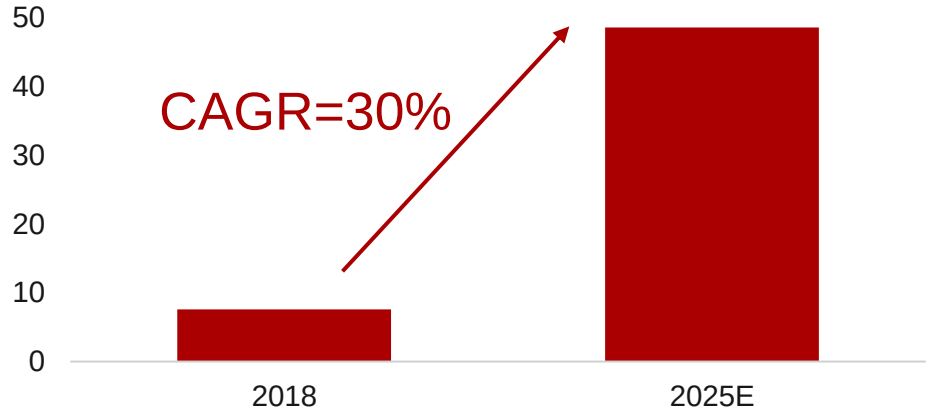
连接方式	SAN	NAS	DAS	
存储架构	集中式存储	分布式存储		
存储类型和协议	文件存储：NFS、CIFS、FTP	块存储：SCSI、iSCSI、NVMe	对象存储：S3、Swift	其他：HDFS、表存储
组网方式	IP组网	FC组网	IB组网	
存储介质	机械硬盘	固态硬盘	光盘	磁带

资料来源：下一代数据存储技术研究报告（2021 年），中国信息通信研究院，浙商证券研究所

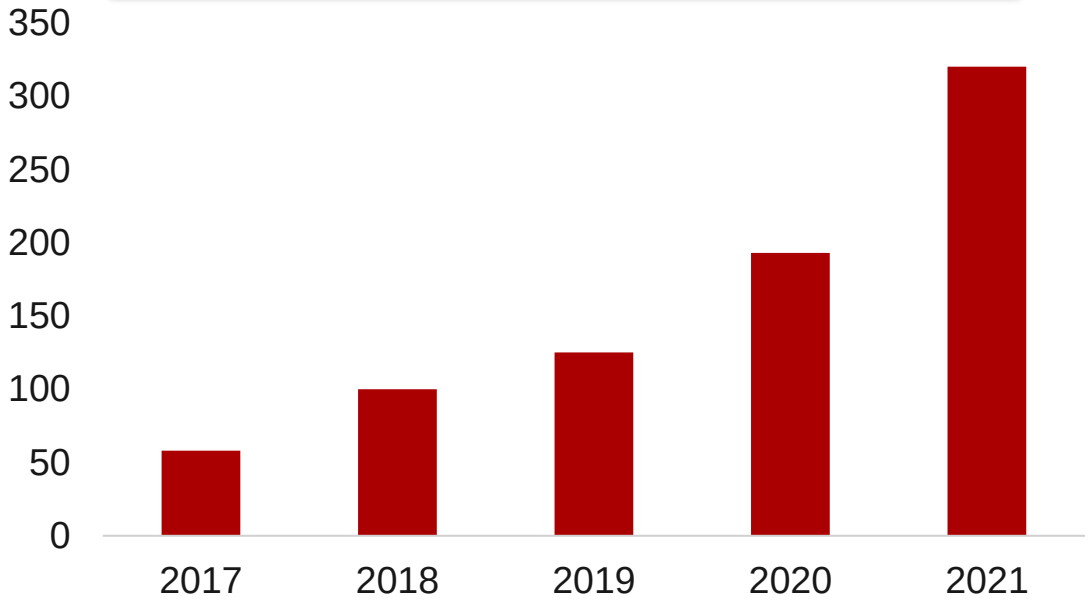
01 我国数据量爆发驱动存储需求增长

得益于人工智能、物联网、云计算、边缘计算等新兴技术在中国的快速发展，中国数据正在迎来爆发式增长。据此前IDC预测，预计到2025年，中国数据圈将增长至48.6ZB，占全球数据圈的27.8%，成为全球最大的数据圈。

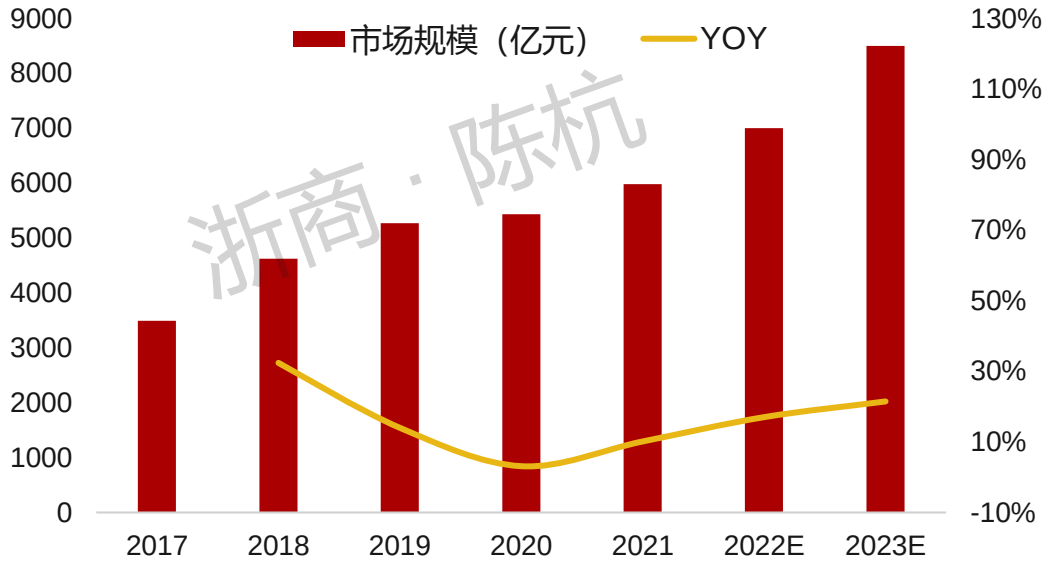
中国新增数据量统计及预测



我国存储容量增量



中国数据存储市场规模预测

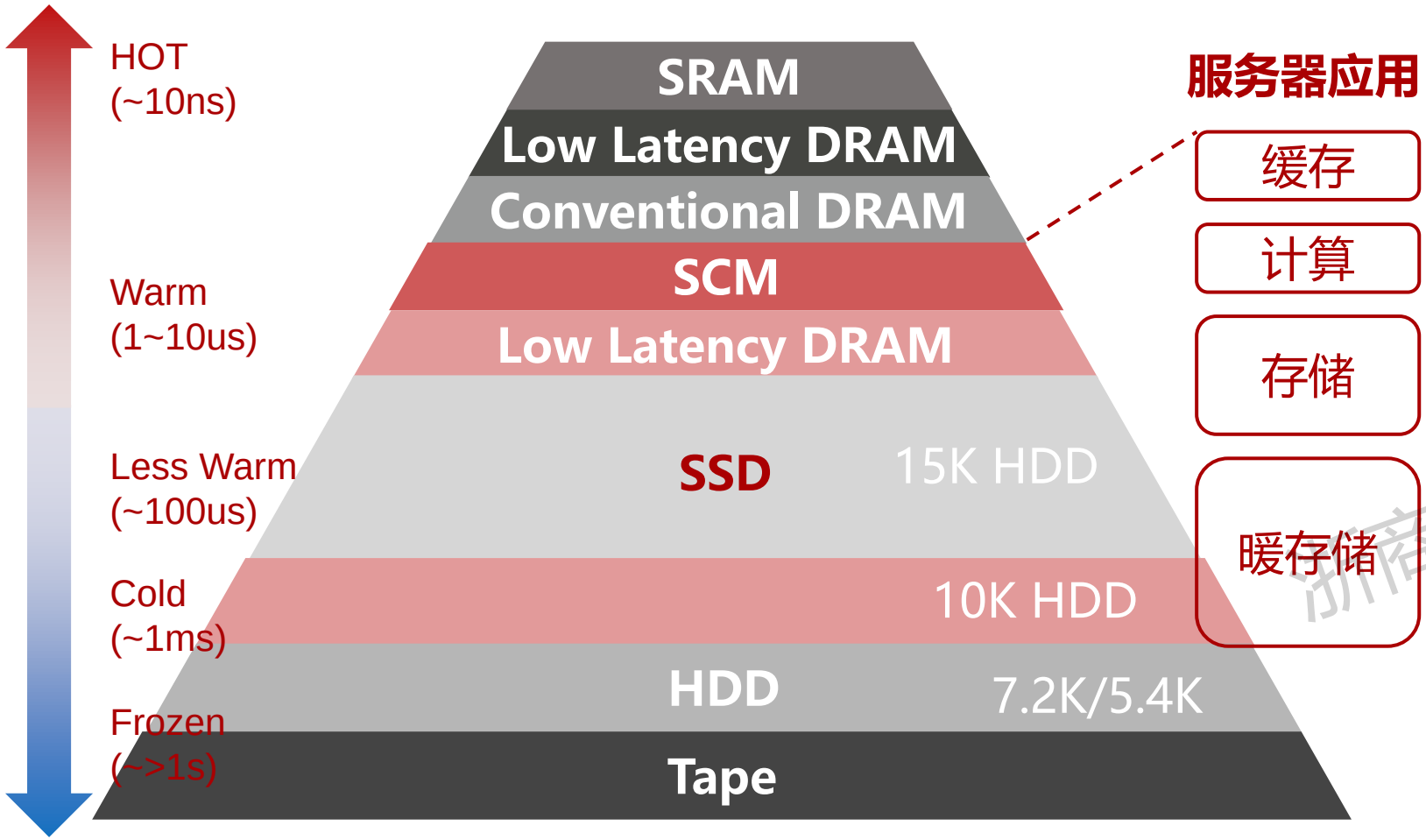


资料来源：联想、IDC、中国存力白皮书（2022年），中国信通院，中国数据存储行业发展深度调研与投资前景研究报告（2023-2030年），浙商证券研究所

01 数据中心驱动存储需求增长

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

随着数据量的大规模增长，存储设备在数据中心采购的BOM中占比进一步提升，美光曾提及，目前存储芯片在数据中心采购中比例约为40%，未来预计将提升至50%。数据中心将成为引领存储市场增长的重要引擎。



服务器应用

- 缓存
- 计算
- 存储
- 暖存储

联想存储服务器 ThinkServer DN8848 V2

处理器	两颗英特尔至强可扩展处理器，最大支持TDP 205W
内存	最多24个DDR4 DIMM 插槽；支持RDIMM和LRDIMM，内存频率最高3200 MHz；最多支持8根英特尔Optane持久内存200系列
扩展性（扩展插槽）	Up to 8 FHFL PCIe Gen4 Slots + 1 OCP 3.0 支持机箱内置Raid卡（通过机箱内置PCIe riser） 最多 4个单宽GPU（60W）
主板硬盘控制器	最多48+4个硬盘

资料来源：联想，闪存市场，浙商证券研究所

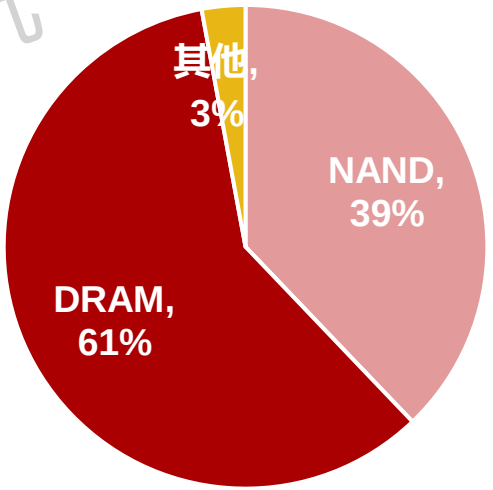
存储器芯片提供底层支撑

- 存储器
  - RAM 只读存储器
    - SRAM
    - DRAM
      - DDR
      - LPDDR
    - FRAM
  - ROM 随机存储器
    - FLASH
      - NAND
        - 2D
        - 3D
      - NOR
        - 串行
        - 并行
    - PROM
    - EPROM
    - EEPROM

应用存储：华邦、旺宏、兆易创新、Cypress、美光

密码存储：三星、SK海力士、美光、南亚、华邦

数据存储：三星、SK海力士、英特尔、美光、西数、铠侠

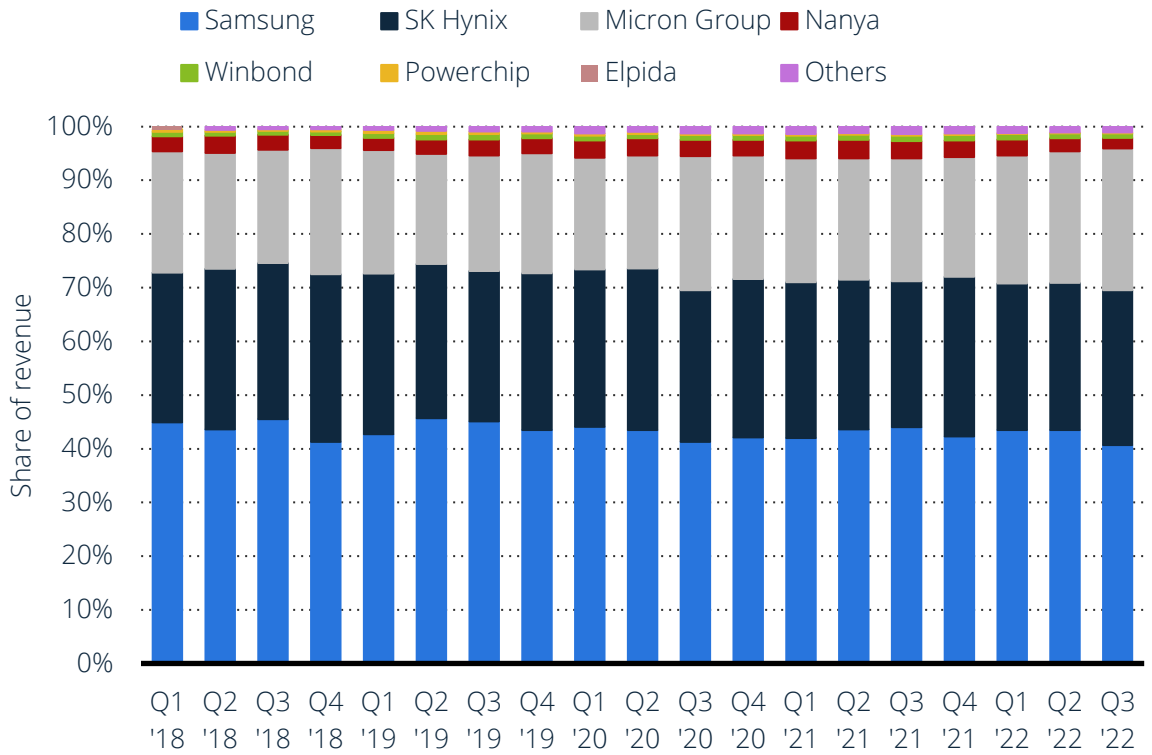


资料来源：CSDN，华经产业研究院，浙商证券研究所

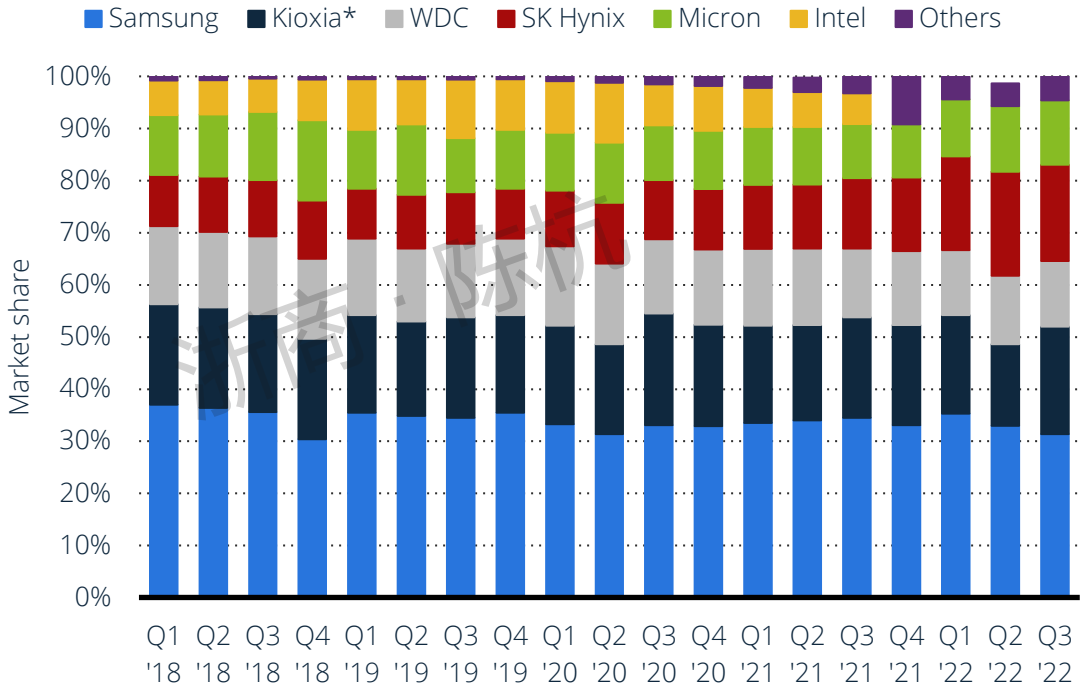
01 全球存储市场：国外厂商垄断，行业高度集中

- **全球存储市场绝大部分份额由国外厂商占有，呈现寡头垄断格局，行业集中度较高。**
- 根据statista数据，截至2022Q3，全球DRAM市场几乎由三星、SK海力士和美光所垄断，CR3 超过 95%，三星、海力士和美光分别占比 41%、29%和 26%。
- 全球NAND flash市场由前三大厂商分别为三星、铠侠和海力士，2022Q3 市场份额分别为 31.4%、20.6%和 13.0%，目前 CR3 市场份额达 65%，CR6 市场份额接近 95%。

全球市场DRAM竞争格局



全球市场NAND flash竞争格局



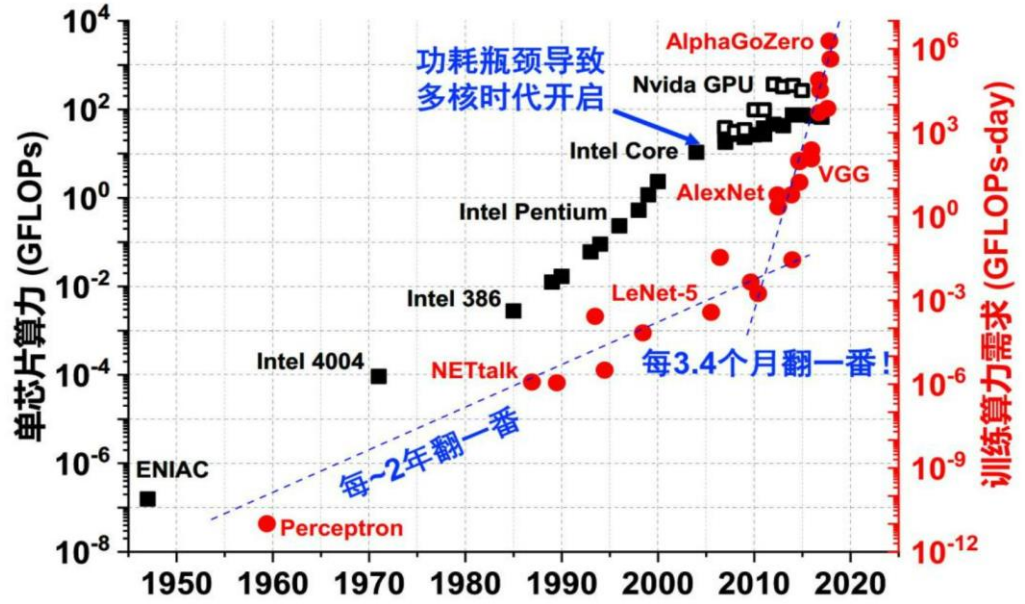
数据来源：statista、浙商证券研究所

02 存算一体

芯片的发展速度和人工智能的算力需求之间的矛盾加剧

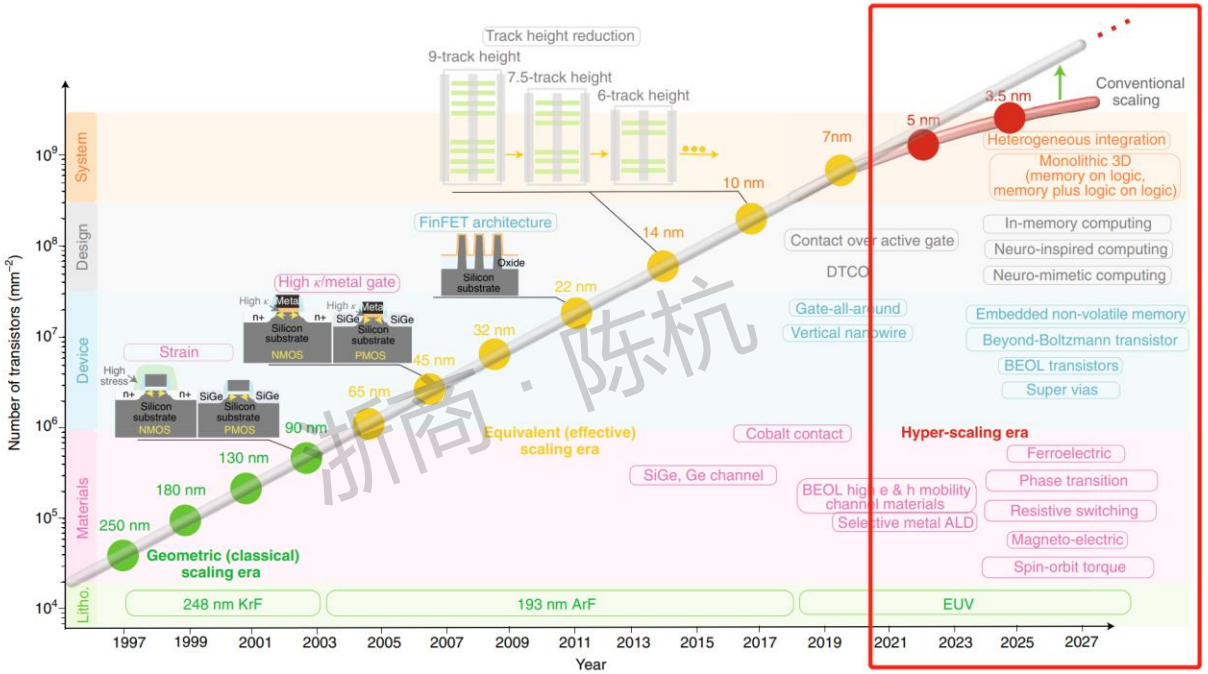
- **芯片的发展速度和人工智能的算力需求之间的矛盾加剧**：21世纪以来，信息爆炸式增长，算力需求大规模上升，提升算力成为芯片行业的共同目标。随着半导体发展放缓，摩尔定理逼近物理极限，依靠器件尺寸微缩来提高芯片性能的技术路径在功耗和可靠性方面都面临巨大挑战，芯片的发展速度无法满足人工智能需求。

算力发展无法满足人工智能的需求



数据来源：CSDN，Intel，Nvidia，OpenAI，清华大学微电子学研究所

摩尔定律陷入发展瓶颈

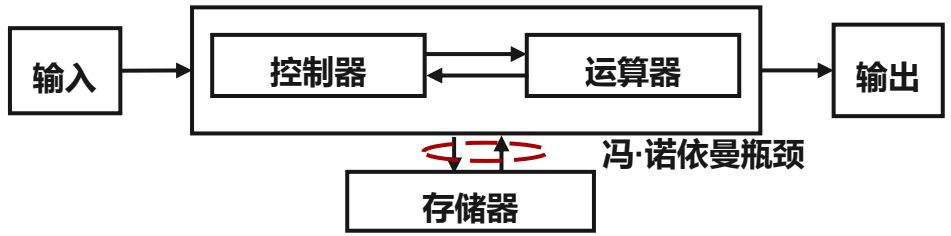


数据来源：Salahuddin, S., Ni, K. & Datta, S. (2018),

传统计算架构面临发展挑战

- **冯·诺依曼架构：**该架构以计算为中心，计算与内存是两个分离单元。计算单元根据指令从内存中读取数据，在计算单元中完成计算和处理，完成后再将数据存回内存。
- **先进制程的优势有限**：随着摩尔定理发展放缓，基于传统架构的芯片计算性能发展速度明显放缓。基于传统架构的先进制程工艺虽一定程度能够提升芯片的性能表现，但从投入产出比、芯片性能可靠性及应用场景的适配度角度考虑都面临较大挑战。

冯·诺依曼架构



投入成本与收益不匹配

- 在传统架构下，从7nm到5nm的研发成和生产成本约增加50%，但性能提升只有10%-20%；
- 联电、格罗方德等芯片制造商已经放弃先进制程，转而聚焦在14nm制程上

物理限制引发的芯片性能问题突显

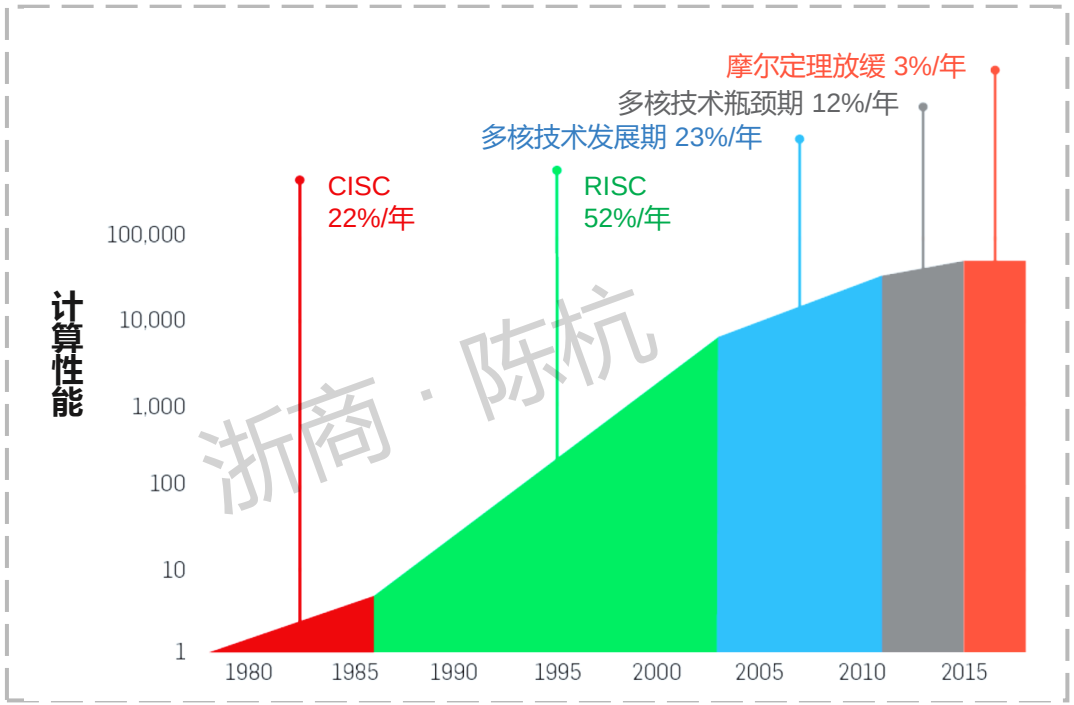
- 随着集成电路尺寸进一步缩小，芯片的性能可靠性受到挑战，由“短沟道效应”和“量子隧穿效应”等引发的芯片漏电，高功耗，产品发热问题引发关注。

先进工艺仅在有限的应用场景中有优势

- 先进工艺下尽管芯片拥有大算力，但同时也产生了高能耗，对于功耗敏感的应用场景，先进制程不占优势。

数据来源：Insights，量子位存算一体芯片深度报告，CSDN

传统构架下性能提升达到极限

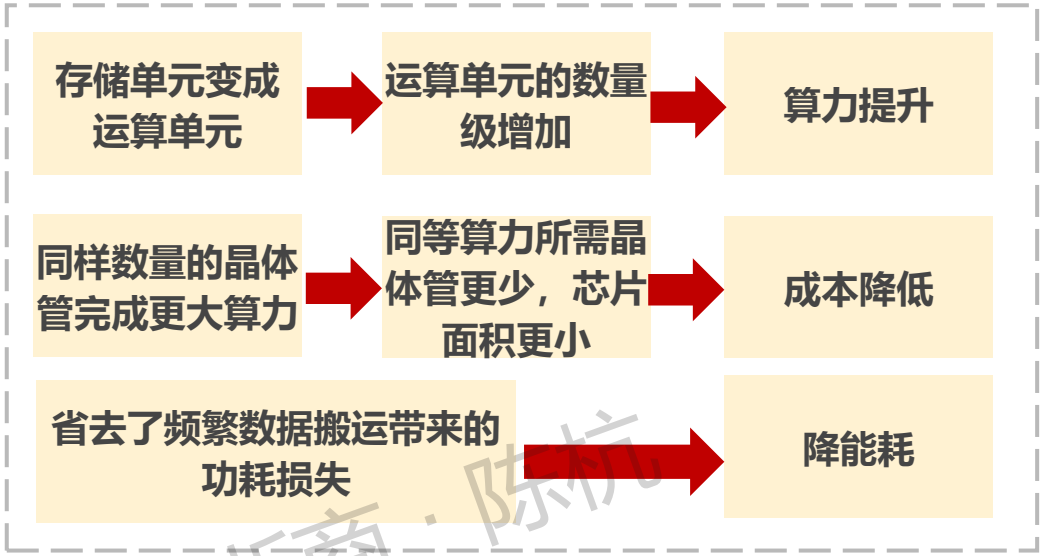


数据来源：清华大学微电子学研究所，J. Hennessy & D. Patterson, 2019

存算一体是先进算力的代表技术

■ **存算一体是先进算力的代表技术**：传统构架下性能提升达到极限，冯·诺依曼架构已成为发展芯片算力的桎梏，存算一体是一种新型计算架构，它是在存储器中嵌入计算能力，将存储单元和计算单元合为一体，省去了计算过程中数据搬运环节，消除了由于数据搬运带来的功耗和延迟，提升计算能效。

存算一体架构



提算力	算力提升 10-100倍	降成本	能耗降低至 1/10-1/100	降能耗	传输功耗约为 0

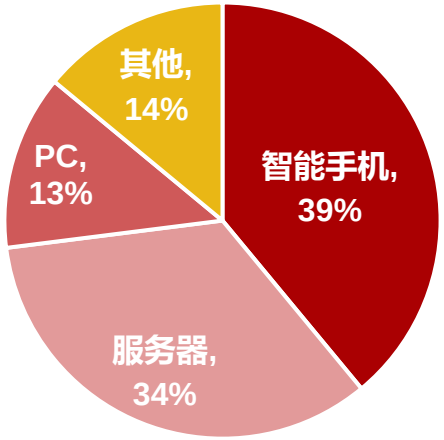
数据来源：量子位存算一体芯片深度报告

03 HBM/DRAM

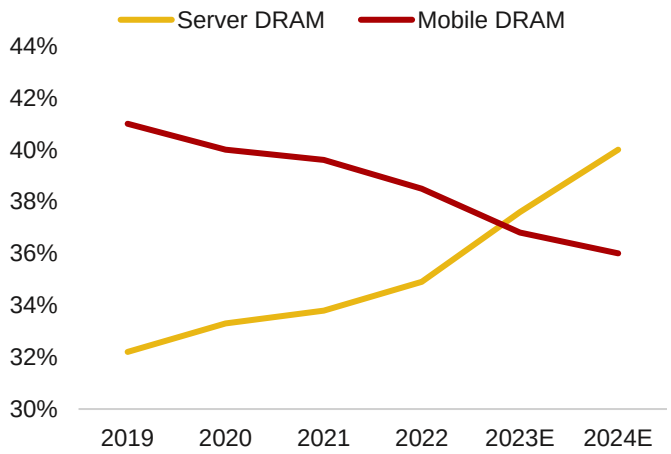
03 DRAM的应用场景主要有智能手机、PC、服务器

浙商证券股份有限公司 ZHESHANG SECURITIES CO.LTD

2021年全球DRAM下游应用占比



服务器与移动DRAM产出比重



DRAM市场前景

2018年

620亿美元

2026年

1219亿美元

市场驱动力

- 居家办公 在线学习
- 家居智能化
- 无人驾驶
- 5G
- 云计算

增量应用市场

- PC
- IoT
- 智能电视
- 机顶盒
- 汽车电子
- 智能手机
- 基站与网络设备
- 服务器

主要DRAM厂商

三巨头：SAMSUNG、SK hynix、Micron

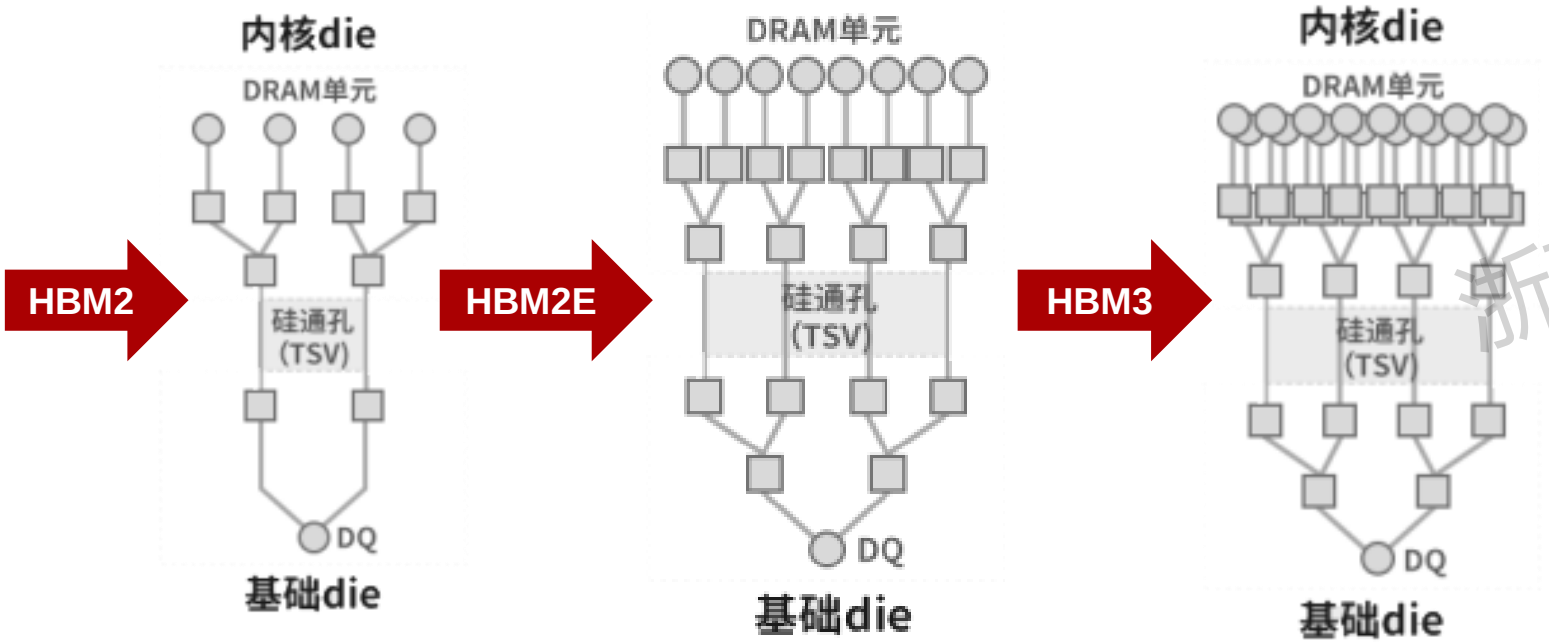
其他厂商：winbond、NANYA、ProMOS TECHNOLOGIES、Etron、力積電、大陆、cxmt、UniIC、GigaDevice、JHICC

资料来源：华经产业研究院，TrendForce，浙商证券研究所

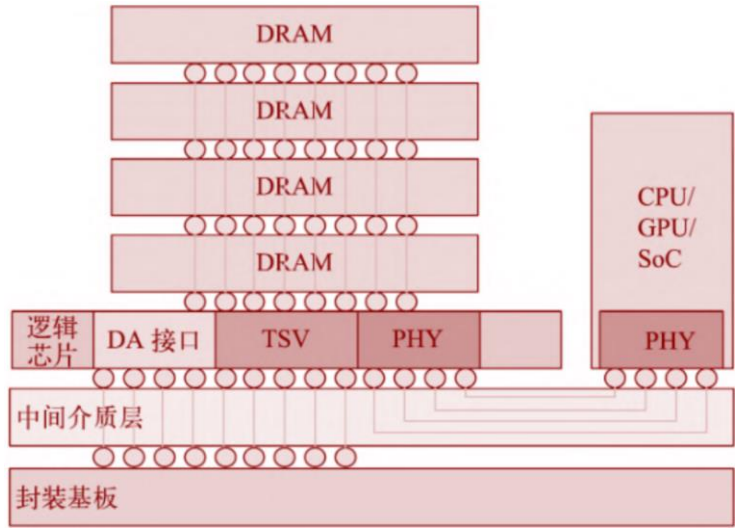
03 HBM是什么：属于DRAM的一种新技术

- HBM（High Bandwidth Memory，高带宽内存）是一款新型的CPU/GPU 内存芯片，其实就是将很多个DDR芯片堆叠在一起后和GPU封装在一起，实现大容量，高位宽的DDR组合阵列。
- 高速、高带宽HBM堆栈没有以外部互连线的方式与信号处理器芯片连接，而是通过中间介质层紧凑而快速地连接，同时HBM内部的不同DRAM采用TSV实现信号纵向连接，HBM具备的特性几乎与片内集成的RAM存储器一样。

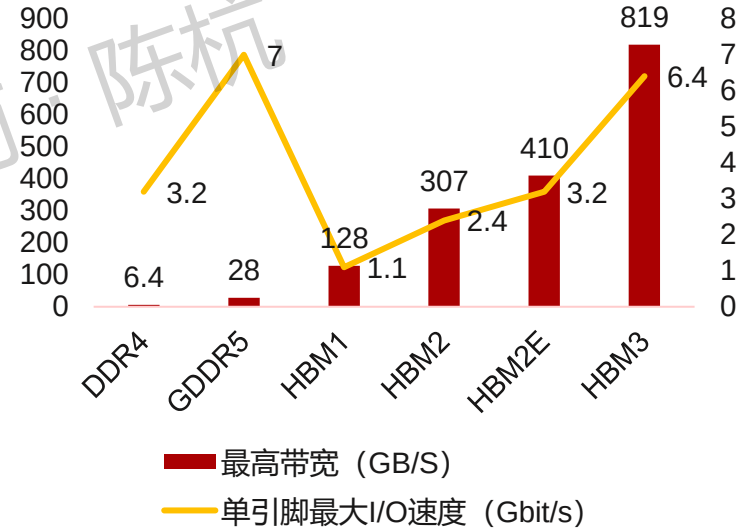
各代HBM产品的数据传输路径配置



HBM堆叠结构



HBM高速、高带宽性能指标



数据来源：半导体产业纵横、半导体行业观察、浙商证券研究所

03 HBM与DDR、GDDR、LPDDR等DRAM分支的参数对比

GDDR5内存每通道位宽32bit，16通道总共512bit; 目前主流的第二代HBM2每个堆栈可以堆至多8层DRAM die，在容量和速度方面有了提升。HBM2的每个堆栈支持最多1024个数据pin，每pin的传输速率可以达到2000Mbit/s，那么总带宽是256Gbyte/s; 在2400Mbit/s的每pin传输速率之下，一个HBM2堆栈封装的带宽为307Gbyte/s。

	DDR4	DDR5	HBM2	GDDR5	LPDDR4	LPDDR5
Applications	Servers→PCs →consumer	Servers→PCs →consumer	Graphics,HPC	Graphics	Mobile,auto, consumer	Mobile,auto, consumer
Typical interface (primary)	Server:64+8 bits	Server:dual channel,32+8 bits	Octal channel, 128-bit(1024 bits total)	Multi-channel, 32-bits	Mobile:quad channel,16-bit (64-bits total)	Mobile:quad channel,16-bit (64-bits total)
Typical interface (secondary)	Consumer:32 bits	Consumer:32 bits	None	None	Dual channel, 16-bit(32-bits total)	Dual channel, 16-bit (32-bits total)
Max Pin BW	3.2 Gb/s	6.4 Gb/s	2.0→2.4 Gb/s	8Gbs	4.267Gb/s	6.4Gb/s
Max I/F BW	25.6 GB/s	51 GB/s	307 GB/s	32 GB/s	34 GB/s	51 GB/s
#Pins/channel	~380 pins	~380 pins	~2,860 pins	~170 pins	~350 pins	~370 pins
Max capacity	3DS RDIMM: 128GB	3DS RDIMM: 256GB	4H Stack: 4GB	One channel: 1GB	4 channels:2GB	4 channels:4GB
Peak volumes	******	*****	**	*	*****	*****
Price per GB	$	$$	$$$$	$$$	$$	$$

数据来源：算力基建、浙商证券研究所

全球HBM厂商：传统DRAM巨头升级竞赛

- **全球HBM芯片市场目前以SK海力士与三星为主**
- SK海力士HBM技术起步早，2014年在业界首次成功研发HBM1，确立领先地位，2022年HBM3芯片供货英伟达，持续巩固其市场领先地位。三星紧随其后，2022年HBM3技术已经量产。
- 从HBM1到HBM3，SK海力士和三星一直是HBM行业的领军企业。目前，HBM4的相关预测数据已经出炉，预计新一代产品将能够更广泛地应用于高性能数据中心、超级计算机和人工智能等领域。

2014	2016	2018	2020	2021	2022
SK海力士:SK海力士与AMD联合开发了全球首款硅通孔（TSV, Through Silicon Via）HBM产品。两家公司还联合开发了高带宽三维堆叠存储器技术和相关产品。	**三星:**宣布开始量产4GB HBM2 DRAM，并在同一年内生产8GB HBM2 DRAM	**三星:**宣布开始量产第二代8GB HBM2“Aquabolt” **SK海力士**：发布第二代HBM产品HBM2	**SK海力士:**宣布成功研发出新一代“HBM2E” **三星**：宣布推出其16GB HBM2E产品“Flashbolt” **美光**：在财报会议上表示，将开始提供HBM2内存/显存，用于高性能显卡，服务器处理器产品	**SK海力士**：开发出全球首款HBM3 **三星**：开发出具有AI处理能力的高带宽内，将强大的AI计算能力引入高性能内存（HBM-PIM） **美光**：HBM2E产品上市，规划2022HBMNEXT产品	**SK海力士**：量产全球首款HBM3 DRAM芯片，并将供货英伟达，持续巩固其市场领先地位。 **三星**：HBM3技术已经量产，其单芯片接口宽度可达1024bit，接口传输速率可达6.4Gbps

数据来源：SKhynix NEWSROOM、美光官网、三星官网、半导体行业观察、浙商证券研究所

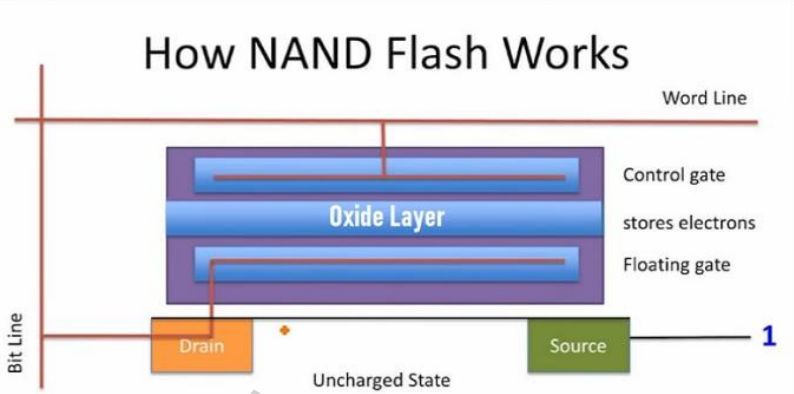
04 NAND

04 NAND：主流闪存芯片

闪存芯片是最主要的存储芯片，主要为 NOR Flash 和 NAND Flash 两种。NOR Flash 主要用来存储代码及部分数据，是手机、PC、DVD、TV、USB Key、 机顶盒、物联网设备等代码闪存应用领域的首选。**NAND Flash 可以实现大容量存储、高写入和擦除速度、相当擦写次数，多应用于大容量数据存储，例如智能手机、平板电脑、U 盘、固态硬盘等领域。**

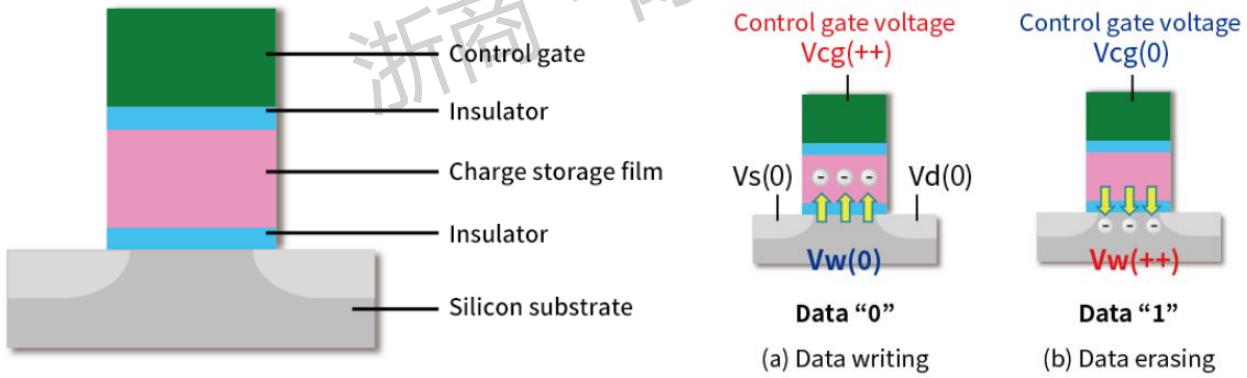
NAND Flash 工作方式

NAND 存储器使用浮栅晶体管，它能在没有电源的情况下存储信息。所有的电路都依赖于某种能量来使整个电池的电荷产生差异，这种能量迫使电子穿过栅极。随着这种电荷返回到关闭状态，随机存取存储器 (RAM) 等易失性类型的存储器会丢失其数据。但是Nand闪存就不同了，它的浮动栅极系统通过使用第二个栅极在电子穿过电池时收集和捕获一些电子，这使得粘在浮栅上的电子在没有电压的情况下保持原位，在这一过程中不管是否有电源连接，芯片都能继续存储下一个值。



NAND Flash 结构

闪存的存储单元结构（横截面）。存储单元是数据存储的最小单位，目前闪存已经由数千亿个存储单元组成。通过将电子移入和移出封闭在绝缘体中的电荷存储膜来存储数据。右图解释了电子是如何移入和移出该电荷存储膜，

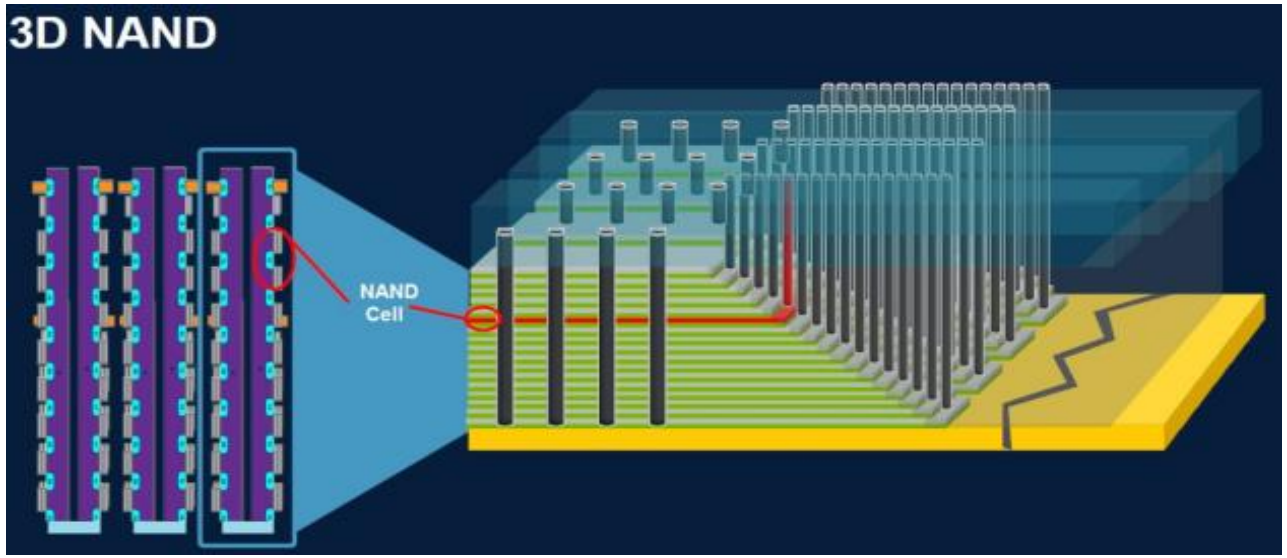


数据来源：兆易创新招股说明书、电巢、浙商证券研究所

3D NAND：高楼大厦平地起

3D NAND, 即立体堆叠技术，如果把2D NAND看成平房，那么3D NAND就是高楼大厦，建筑面积成倍扩增，理论上可以无限堆叠，可以摆脱对先进制程工艺的束缚，同时也不依赖于极紫外光刻（EUV）技术，而闪存的容量/性能/可靠性也有了保障。

三维 NAND 单元的阶梯通孔接触



日前三星宣布，已开始批量生产采用第8代V-NAND技术的产品，为1Tb（128GB）TLC 3D NAND闪存芯片，达到了236层，相比第7代V-NAND技术的176层有了大幅度的提高。三星称，新的闪存芯片提供了迄今为止业界内最高的位密度，可在下一代企业服务器系统中实现更大的存储空间。

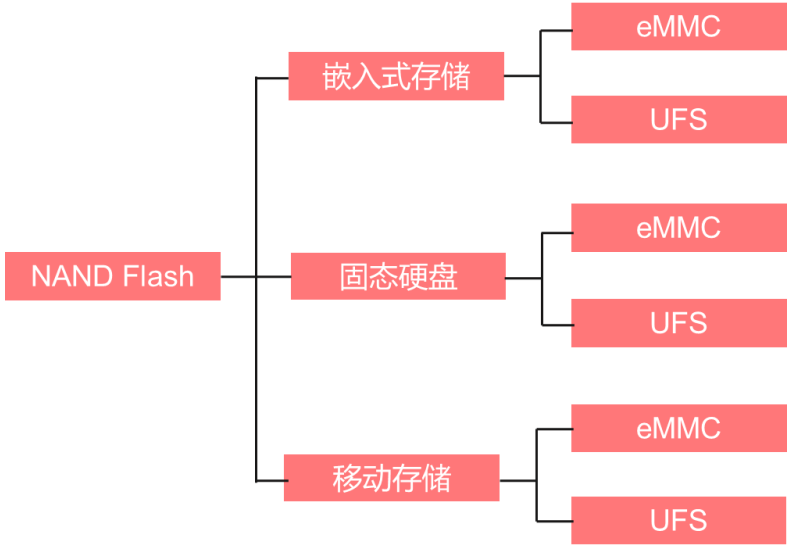
2D与3D NAND对比图

Product	2D NAND	3D NAND
Image		
Capacity per die	max. 128Gb	256Gb/512Gb (space for future increase)
Design	Floating Gate	Floating Gate or Charge Trap
Endurance (P/E Cycles)	Lower	Higher
	3D has less cell-to-cell interference which results in higher endurance	
Performance	Slower	Faster
Power consumption	High	Low

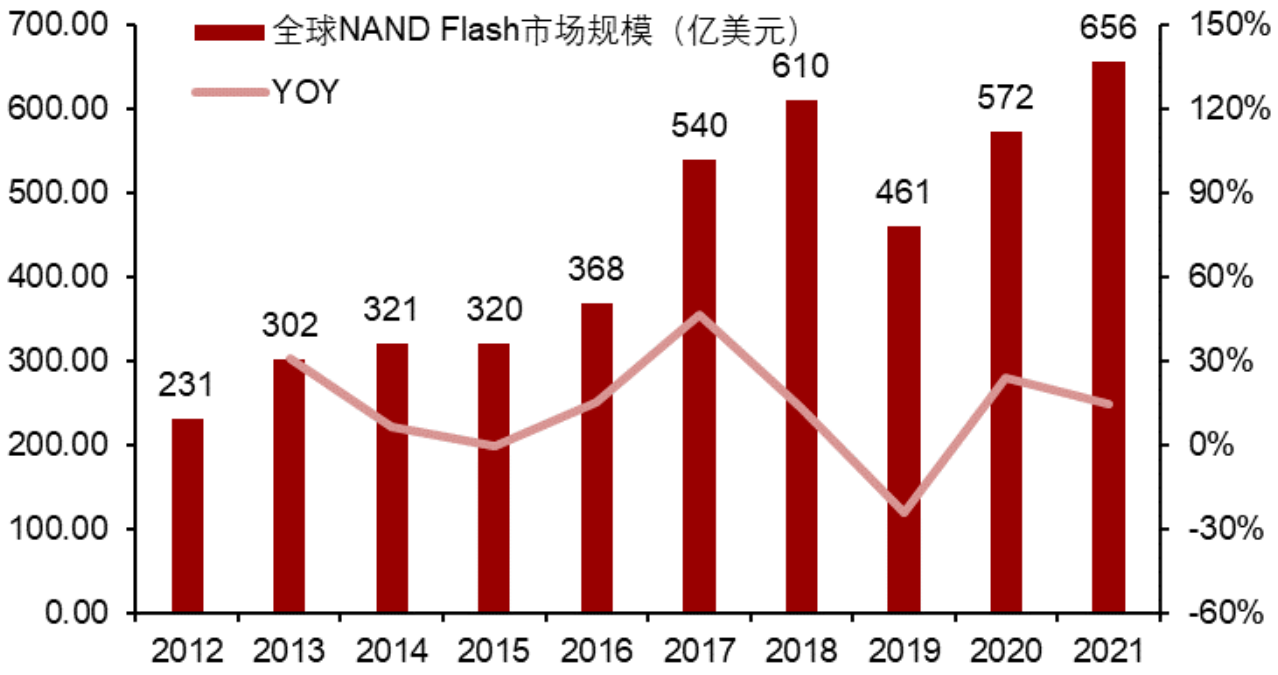
数据来源：电巢、peacat、浙商证券研究所

3D NAND是低成本/大容量非易失存储的主流技术方案

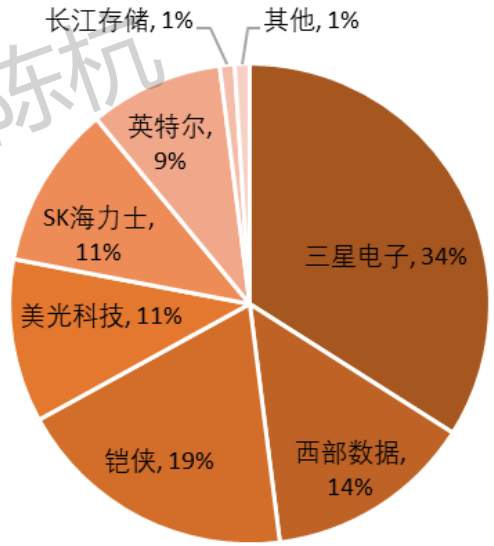
NAND Flash 为大容量数据存储的实现提供了廉价有效的解决方案，是目前全球市场大容量非易失存储的主流技术方案。NAND Flash 是使用电可擦技术的高密度非易失性存储，NAND Flash 每位只使用一个晶体管，存储密度， Flash 所存的电荷（数据）可长期保存；同时， NAND Flash 能够实现快速读写和擦除。



全球NAND Flash市场规模



2020年NAND Flash市场份额

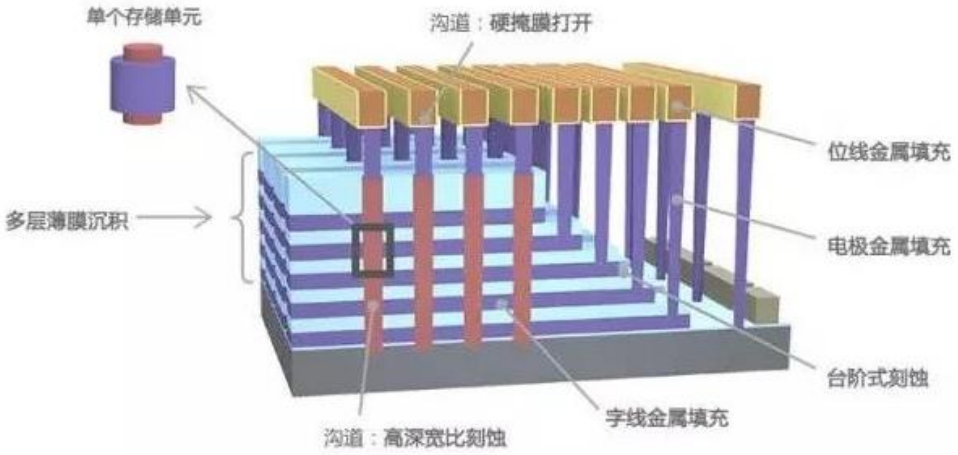


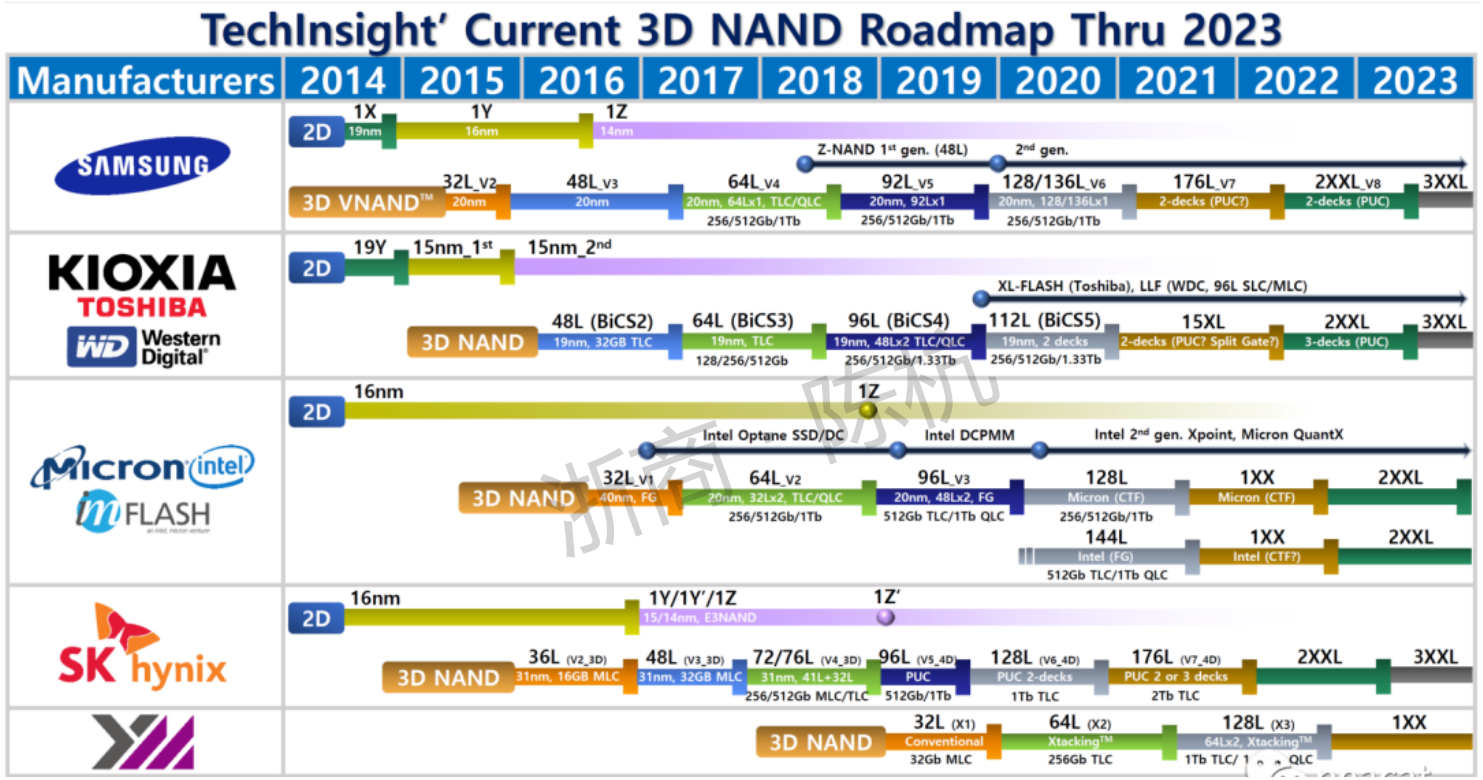
数据来源：江波龙招股说明书，浙商证券研究所

3D NAND持续追求高堆栈层数，多种工艺架构并存

与2D NAND缩小Cell提高存储密度不同的是，3D NAND只需要提高堆栈层数，目前多种工艺架构并存。从2013年三星推出了第一款24层SLC/MLC 3D V-NAND，到现在层数已经迈进200+层，并即将进入300+层阶段。目前，三星/西部数据/海力士/美光/铠侠等几乎垄断了所有市场份额，并且都具有自己的特殊工艺架构，韩系三星/海力士的CTF，美系镁光/英特尔的FG，国内长江存储的X-tacking。

3D NAND制造工艺的关键步骤





数据来源：长江存储官网，Techinsights，浙商证券研究所

04 高堆栈层数的3D NAND面临的工艺难度越来越高

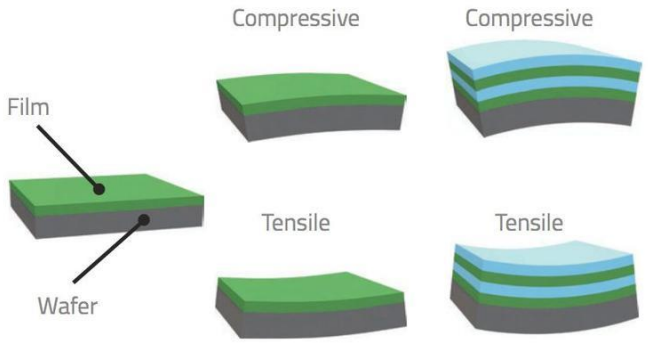
随着堆栈层数的增加，工艺也面临越来越多的挑战，对制造设备和材料也提出了更多的要求。主要包括以下几个方面：

（1）ONON薄膜应力：随着器件层数增加，薄膜应力问题越发凸显，会影响后续光刻对准精度；

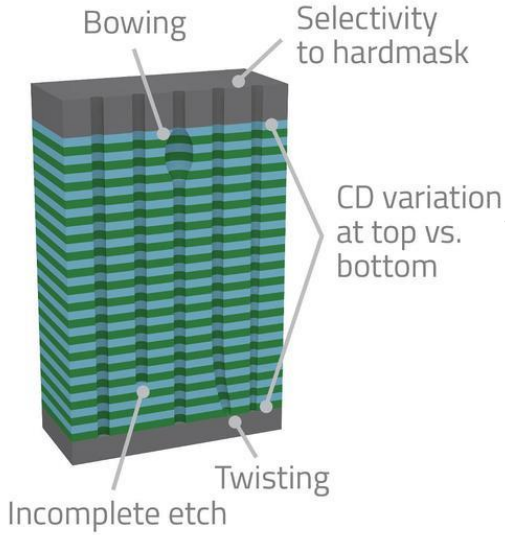
（2）高深宽比通孔刻蚀：随着深宽比增加，刻蚀难度会显著增加，容易出现刻蚀不完全、通孔结构扭曲等问题；

（3）WL台阶的设计与刻蚀：垂直管状环栅结构的器件需要刻蚀出精确的台阶结构，保障CT能打到对应位置，而随着层数增加，工艺难度加大，需要重新设计WL台阶结构。

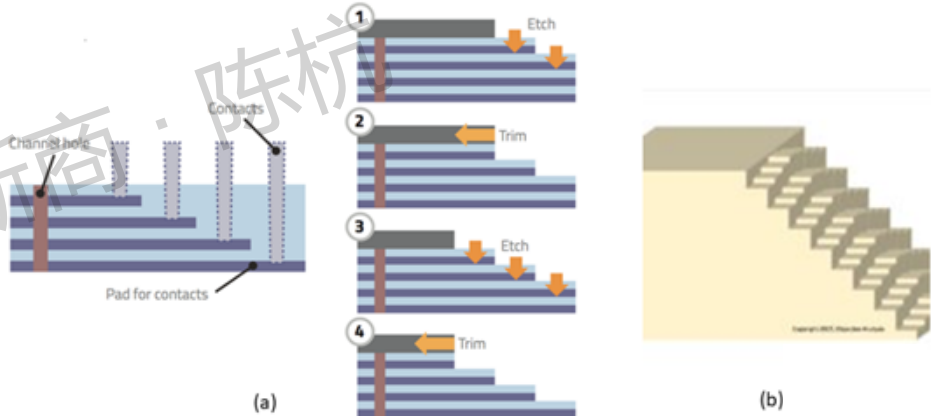
ONON/OPOP多层堆叠



高深宽比通孔刻蚀



WL台阶的设计与刻蚀



数据来源：Web of Science，国科大，CNKI，浙商证券研究所

风险提示

1、**宏观经济下行超预期风险**：若宏观经济下行并超出预期，将影响整个产业链；

2、**上游晶圆紧缺加剧的风险**：若本土晶圆厂扩产进度不及预期，将影响芯片供给；

3、**市场发展不及预期的风险**：若人工智能发展所带来的催化作用不及预期，将影响下游需求；

4、**技术发展不及预期风险**：若技术发展不及预期，将影响行业发展趋势。

行业评级与免责声明

浙商证券股份有限公司
ZHESHANG SECURITIES CO.LTD

行业的投资评级

以报告日后的6个月内，行业指数相对于沪深300指数的涨跌幅为标准，定义如下：

1、看好 ：行业指数相对于沪深300指数表现＋10%以上；

2、中性 ：行业指数相对于沪深300指数表现－10%～＋10%以上；

3、看淡 ：行业指数相对于沪深300指数表现－10%以下。

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重。

建议：投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者不应仅仅依靠投资评级来推断结论

行业评级与免责声明

法律声明及风险提示

本报告由浙商证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，经营许可证编号为：Z39833000）制作。本报告中的信息均来源于我们认为可靠的已公开资料，但浙商证券股份有限公司及其关联机构（以下统称“本公司”）对这些信息的真实性、准确性及完整性不作任何保证，也不保证所包含的信息和建议不发生任何变更。本公司没有将变更的信息和建议向报告所有接收者进行更新的义务。

本报告仅供本公司的客户作参考之用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本报告仅反映报告作者的出具日的观点和判断，在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求。对依据或者使用本报告所造成的一切后果，本公司及/或其关联人员均不承担任何法律责任。

本公司的交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告意见及建议不一致的市场评论和/或交易观点。本公司没有将此意见及建议向报告所有接收者进行更新的义务。本公司的资产管理公司、自营部门以及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。

联系方式

浙商证券研究所

上海总部地址：杨高南路729号陆家嘴世纪金融广场1号楼25层

北京地址：北京市东城区朝阳门北大街8号富华大厦E座4层

深圳地址：广东省深圳市福田区广电金融中心33层

邮政编码：200127

电话：(8621)80108518

传真：(8621)80106010

浙商证券研究所：http://research.stocke.com.cn