存算一体：继 CPU、GPU 架构之后的算力架构“第三极”

方正证券研究所证券研究报告

行业专题报告　　行业研究

半导体行业　　2023.04.14/推荐

分析师：　吴文吉

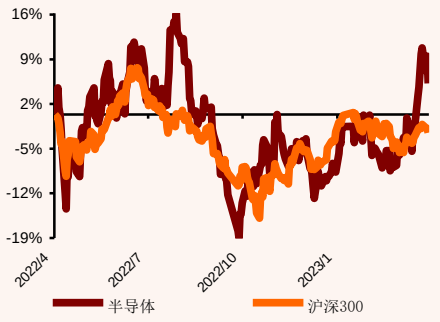
登记编号：　S1220521120003

联 系 人　万玮

重要数据：

上市公司总家数	114
总股本(亿股)	790.23
销售收入(亿元)	3637.16
利润总额(亿元)	581.86
行业平均 PE	-80.13
平均股价(元)	72.28

行业相对指数表现：



数据来源：wind 方正证券研究所

相关研究

《北京君正：营收逆势实现增长，车规存储巨头有望下半年业绩回升》2023.04.11

《瑞芯微：旗舰芯片放量拓宽应用领域，静待 SoC 龙头复苏》2023.04.11

《斯达半导：高歌猛进新能源，研发突破业绩辉》2023.04.10

《国芯科技：RAID 晋新程，助力 AI 锐意进取》2023.04.09

存算一体：解决冯·诺依曼计算架构瓶颈。

算力需求的指数级增长驱动大算力与大模型计算的瓶颈（带宽低、时延长、功耗高）亟待解决。在深度学习中，数据移动大量且频繁地存在于计算单元与存储单元之间，由于数据在 CPU 或 GPU 中频繁高速传递，整个过程的无用能耗大概在 60%-90%；同时由于外部 DRAM 的运行速度远远小于 CPU 或 GPU 的运算速度，冯·诺依曼架构也受到传输带宽瓶颈的限制（常称：存储墙瓶颈），因此系统的运算效率大打折扣。计算架构演进道阻且长，存算一体呼之欲出。

存算一体：继 CPU、GPU 架构之后的算力架构“第三极”。

提升算力的传统思路（ASIC/CPU/GPU/NPU）有待完善，存算一体的优势包括：1）具有更大算力（1000TOPS 以上）；2）具有更高能效（超过 10-100TOPS/W），超越传统 ASIC 算力芯片；3）降本增效（可超过一个数量级）。

存算一体：在云、边、端大有可为。

端侧单设备算力需求约为 0.1~64 TOPS；端侧设备对运行时间、功耗、便携性等有较高要求。边侧单设备算力需求约为 64~256 TOPS；边侧设备对时延、功耗、成本以及通用性等有较高要求。云侧大算力、高带宽、低功耗需求催涨 AI 芯片，存内计算或将成为智算中心下一代关键 AI 芯片技术。

存算一体技术三大驱动因素：

新型存储器的发展+来自应用侧的需求+产业侧的配合

存算一体技术三大应用方向：

AI 和大数据计算、感存算一体、类脑计算

存算一体公司竞争格局：

国外存算一体产业比国内起步早 3-5 年左右。

存算一体芯片市场规模：

基于存算一体技术的小算力芯片 2025 年约 125 亿人民币远期市场空间。2030 年，基于存算一体技术的中小算力芯片市场规模约为 1069 亿人民币，基于存算一体技术的大算力芯片市场规模约为 67 亿人民币，总市场规模约为 1136 亿人民币。

建议关注：

恒烁股份，知存科技（非上市）

风险提示：

1） 半导体下游需求不及预期；

2） 技术发展不及预期；

3） 行业竞争加剧。

目录

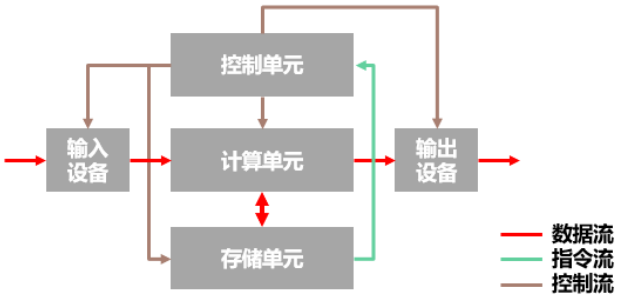
图表目录

1 存算一体：解决冯·诺依曼计算架构瓶颈

1.1 算力需求的指数级增长驱动大算力与大模型计算的瓶颈（带宽低、时延长、功耗高）亟待解决

传统的人工推理芯片解决方案将训练好的权重值存储在外部的存储器 DRAM 中，CPU 或 GPU 做推理运算时不停地调用 DRAM 中的数据，并将中间数据实时存回。这种架构被称为传统**冯·诺伊曼架构**，冯氏架构以计算为中心，计算和存储分离，二者配合完成数据的存取与运算。

图表1： 冯·诺伊曼计算架构



资料来源：《存算一体白皮书（2022 年）-中国移动研究院》，方正证券研究所

图表2： 数据搬运占 AI 计算的主要功耗

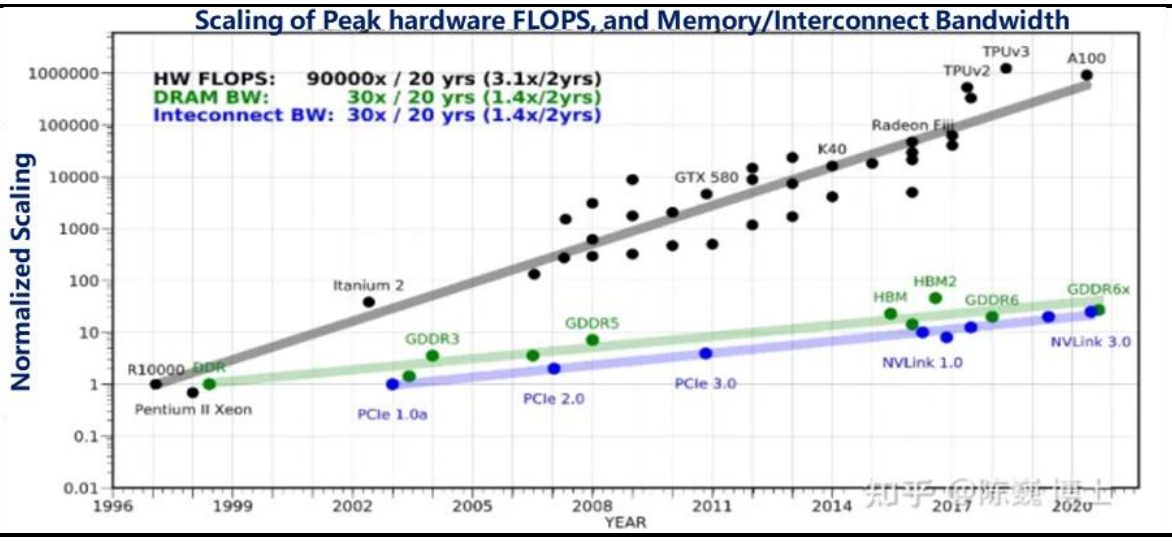
	数据带宽	数据搬运能耗
片外HBM	~960GB/s	~10nJ
片外DDR4	~40GB/s	~10nJ
片内SRAM	10-100TB/s	50pJ
计算功耗	-	5pJ

资料来源：知乎，陈巍谈芯，《先进存算一体芯片设计》（陈巍、耿云川等），方正证券研究所

1.2 优良的能效比为提升算力的关键

正如 CMOS 工艺凭借优良的能效比成为主流工艺的关键，优良的能效比亦是提升算力的关键。在深度学习中，数据移动大量且频繁地存在于计算单元与存储单元之间，由于数据在 CPU 或 GPU 中频繁高速传递，整个过程的**无用能耗大概在 60%-90%**；同时由于外部 DRAM 的运行速度远远小于 CPU 或 GPU 的运算速度，冯·诺依曼架构也受到传输带宽瓶颈的限制（常称：**存储墙瓶颈**），因此系统的运算效率大打折扣。

图表3： 存储计算性能“剪刀差”：算力发展速度远超存储器

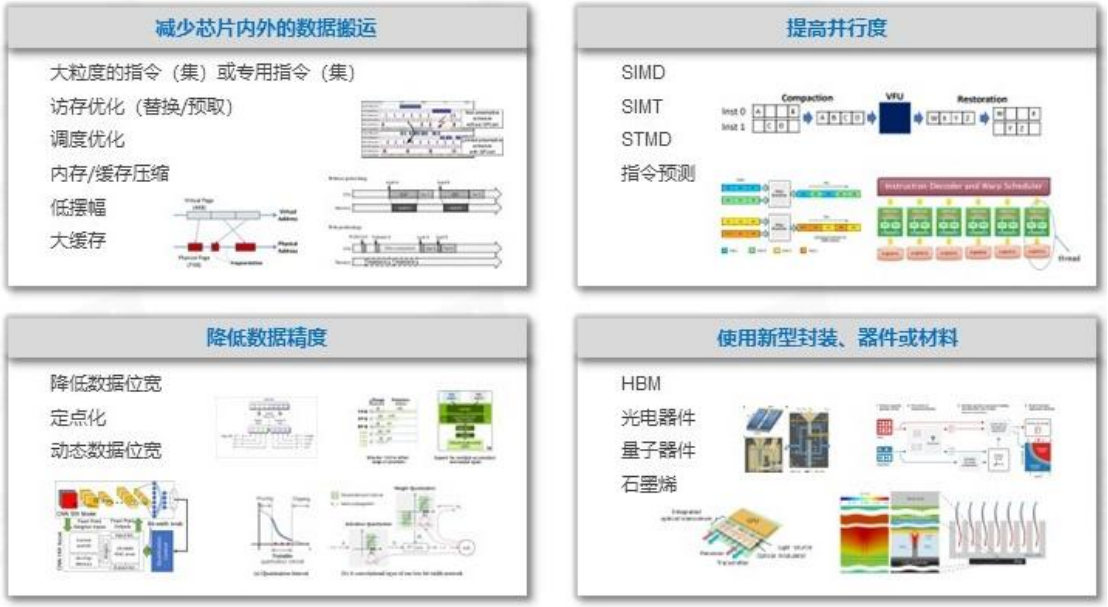


资料来源：知乎，陈巍谈芯，《先进存算一体芯片设计》（陈巍、耿云川等），github，方正证券研究所

1.3 提升算力的传统思路（ASIC/CPU/GPU/NPU）有待完善

目前集成电路的发展进入后摩尔时代，业界除了从“More Moore（深度摩尔）”、“More than Moore（超越摩尔）”与“Beyond CMOS（新器件）”这三大方向探索提升算力的技术路径，也在**通过变革当前的计算架构来实现算力的突破。目前，主流芯片如 CPU、GPU 以及 DPU 均按照冯·诺依曼架构设计。**

图表4： 提升算力的传统思路（ASIC/CPU/GPU/NPU）

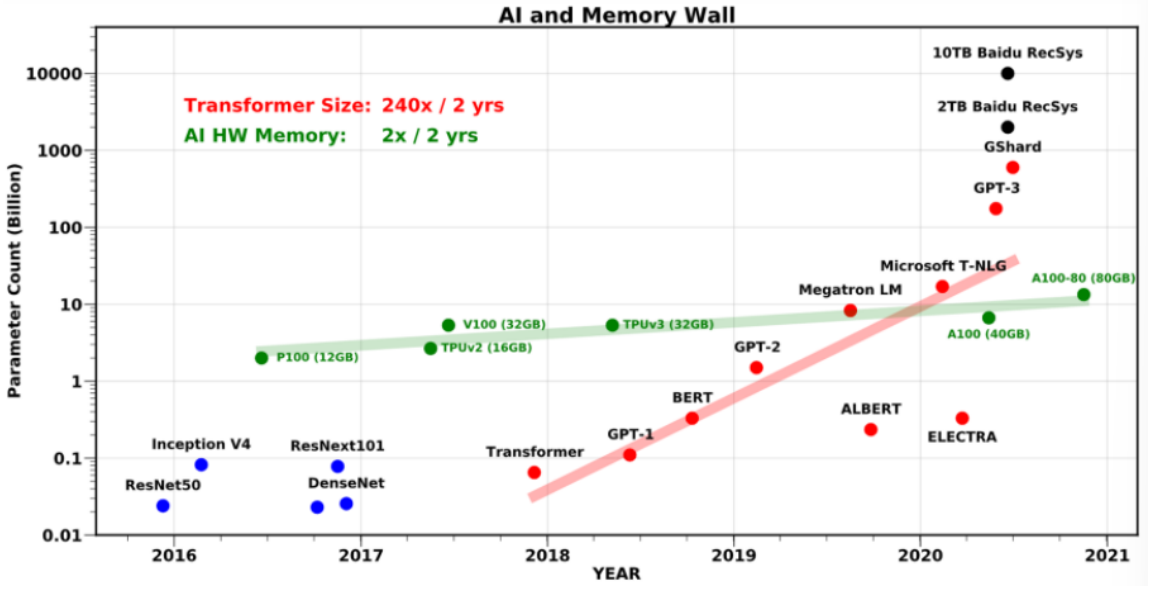


资料来源：知乎，陈巍谈芯，方正证券研究所

1.4 计算架构演进道阻且长，存算一体呼之欲出

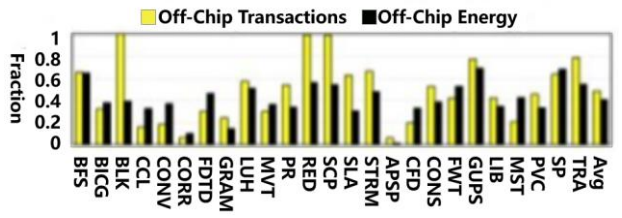
虽然多核（如 CPU）/众核（如 GPU）并行加速技术可以提升算力，但在后摩尔时代，存储带宽制约了计算系统的有效带宽，系统算力增长步履维艰。**GPU 的架构演进并未解决大算力和大模型的挑战。**

图表5： SOTA Transformer 模型参数量（红点）和 AI 硬件内存大小（绿点）增长趋势对比



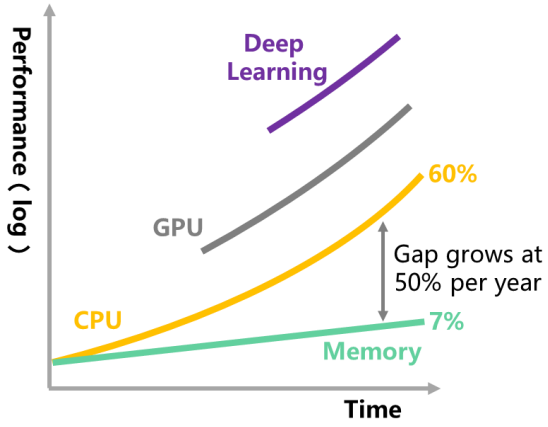
资料来源：英特尔官网，github，方正证券研究所

图表6： GPU 中数据传输引发功耗损失超过 60%



资料来源：知乎，陈巍谈芯，方正证券研究所

图表7： CPU、GPU 等处理数据的能效表现

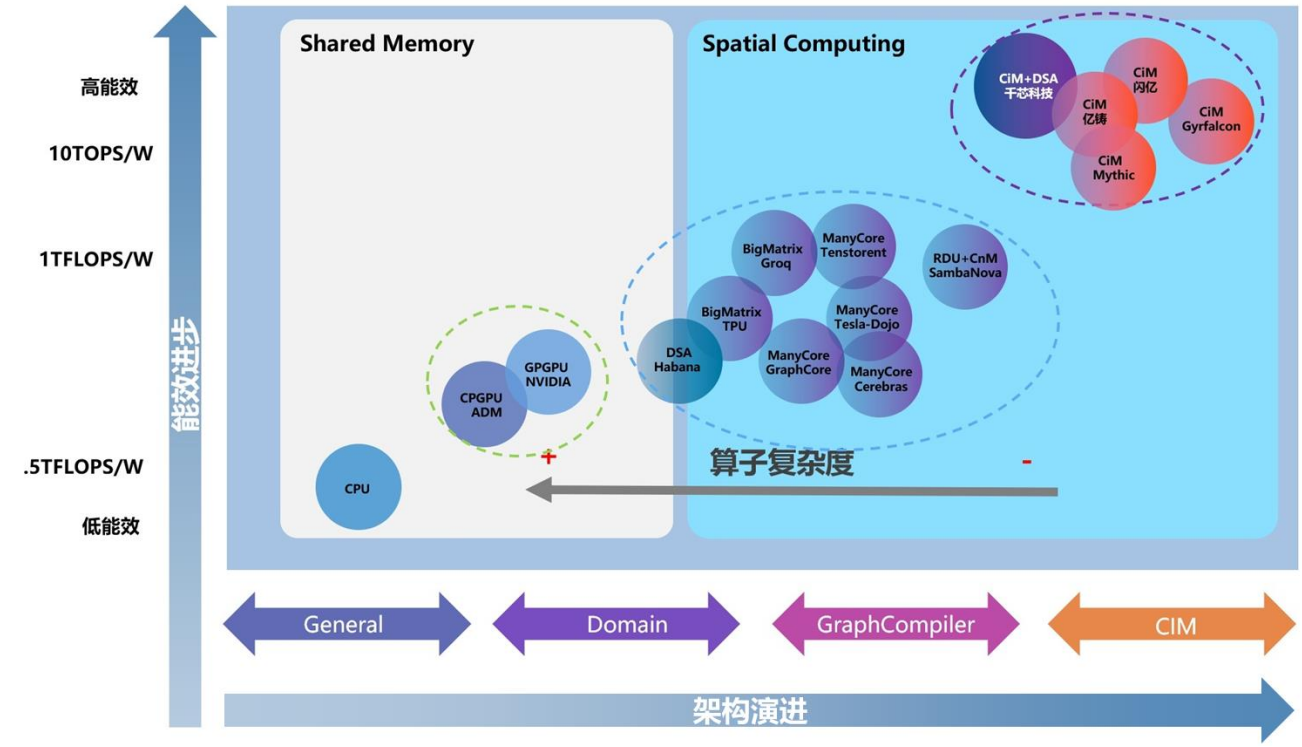


资料来源：知乎，陈巍谈芯，方正证券研究所

2 存算一体：继 CPU、GPU 架构之后的算力架构“第三极”

作为一种新的计算架构，存算一体被认为是最具有潜力的革命性技术，其核心是将存储与计算完全融合，存储器中叠加计算能力，以新的高效运算架构进行二维和三维矩阵计算，结合后摩尔时代先进封装、新型存储器件等技术，能有效克服冯·诺依曼架构瓶颈，实现计算能效的数量级提升。

图表8： 计算架构演进图示



资料来源：知乎，陈巍谈芯，方正证券研究所

2.1 存算一体三大优势

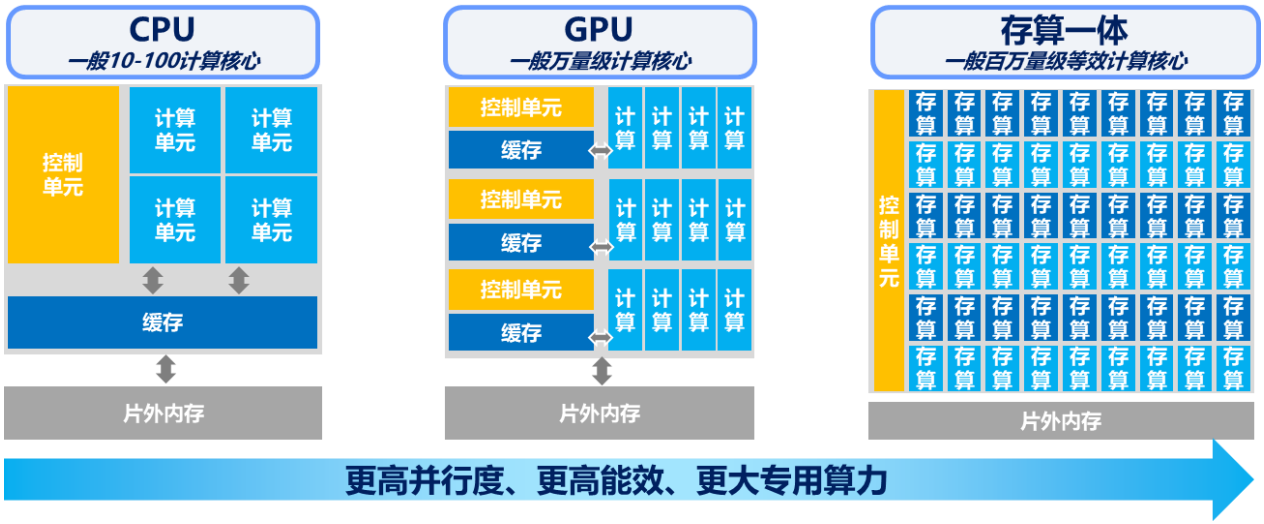
存算一体的优势包括 1）具有更大算力（1000TOPS 以上）；2）具有

更高能效（超过 10-100TOPS/W），超越传统 ASIC 算力芯片；3）降本增效（可超过一个数量级）。

2.2 存算一体技术三大底层特征

存算一体技术的技术底层特征包括：1）减少数据搬运（降低能耗至 1/10~1/100）；2）存储单元具备计算能力（等效于在面积不变的情况下规模化增加计算核心数，或者等效于提升工艺代）；3）单个存算单元替代“计算逻辑+寄存器”更小更快。

图表9： CPU、GPU 与存算一体结构比较



资料来源：知乎，陈巍谈芯，《先进存算一体芯片设计》（陈巍、耿云川等），方正证券研究所

2.3 存算一体行业趋势

图表10： 存算一体行业趋势

资料来源：《存算一体芯片深度产业报告—量子位智库》，方正证券研究所

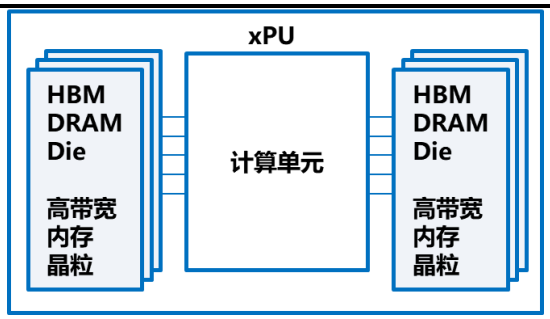
2.4 存算一体技术分类

2.4.1 近存计算（PNM）

近存计算通过芯片封装和板卡组装等方式，将存储单元和计算单元集成，增加访存带宽、减少数据搬移，提升整体计算效率。近存计算仍是存算分离架构，本质上计算操作由位于存储外部、独立的计算单元完成其技术成熟度较高，主要包括存储上移、计算下移两种方式。近存计算已应用于人工智能、大数据、边缘计算等场景因其基本保持原有计算架构，产品化方案可较快投入使用。

1）存储上移：采用先进封装技术将存储器向处理器（如 CPU、GPU）靠近，增加计算和存储间的链路数量，提供更高访存带宽。典型的产品形态为高带宽内存（High Bandwidth Memory，HBM），将内存颗粒通过硅通孔（Through Silicon Via，TSV）多层堆叠实现存储容量提升，同时基于硅中介板的高速接口与计算单元互联提供高带宽存储服务。

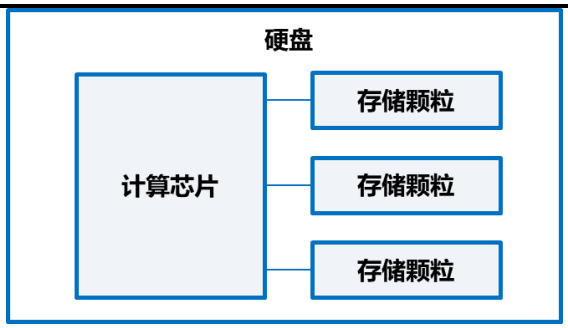
图表11： 高带宽内存方案



资料来源：《存算一体白皮书（2022 年）-中国移动研究院》，方正证券研究所

2）计算下移：采用板卡集成技术将数据处理能力卸载到存储器，由近端处理器进行数据处理，有效减少存储器与远端处理器的数据搬移开销。典型的方案为可计算存储（Computational Storage Drives，CSD），通过在存储设备引入计算引擎承担如数据压缩、搜索、视频文件转码等本地处理，减少远端处理器（如 CPU）的负载。

图表12： 可计算存储方案



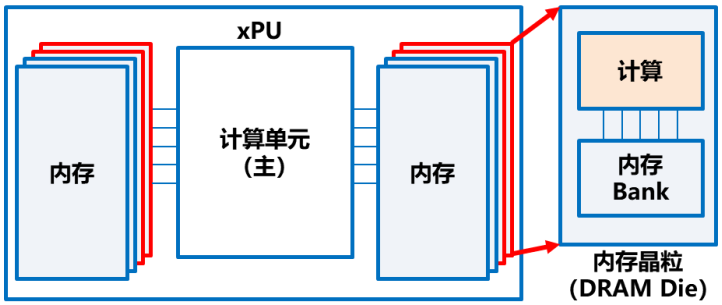
资料来源：《存算一体白皮书（2022 年）-中国移动研究院》，方正证券研究所

2.4.2 存内处理（PIM）

存内处理是在芯片制造的过程中，将存和算集成在同一个晶粒

（Die）中，使存储器本身具备了一定算的能力。存内处理本质上仍是存算分离相比于近存计算，“存”与“算”距离更近。当前存内处理方案大多在内存（DRAM）芯片中实现部分数据处理，较为典型的产品形态为 HBM-PIM 和 PIM-DIMM，在 DRAM Die 中内置处理单元，提供大吞吐低延迟片上处理能力，可应用于语音识别、数据库索引搜索、基因匹配等场景。

图表13：　　基于 DRAM 的 PIM 方案实例



资料来源：《存算一体白皮书（2022 年）-中国移动研究院》，方正证券研究所

2.4.3　存内计算（CIM）

存内计算即狭义的存算一体。在芯片设计过程中，不再区分存储单元和计算单元，真正实现存算融合。存内计算是计算新范式的研究热点，其本质是利用不同存储介质的物理特性，对存储电路进行重新设计使其同时具备计算和存储能力，直接消除“存”“算”界限，使计算能效达到数量级提升的目标。**存内计算最典型的场景是为 AI 算法提供向量矩阵乘的算子加速**，目前已经在神经网络领域开展大量研究，如卷积神经网络（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Network, RNN）等。

存内计算主要包含数字和模拟两种实现方式，二者适用于不用的应用场景。

1）模拟存内计算：适用于低精度、低功耗计算

模拟存算一体通常使用 FLASH/RRAM、PRAM 等非易失性存储截至作为存储器件，存储密度大，并行度高，但是对环境噪声和温度非常敏感，适用于低精度、低功耗计算场景，如端侧可穿戴设备等。

2）数字存内计算：适用于高精度、功耗不敏感计算

数字存算一体主要以 SRAM/RRAM/DRAM 作为存储器件，采用先进逻辑工艺，具有高性能高精度的优势，且具备良好的抗噪声能力和可靠性，适用于高精度、功耗不敏感计算场景，未来可应用于云边 AI 场景。一直以来，主流的存内计算大多采用模拟计算实现，近两年数字存内计算的研究热度飞速提升。

2.5　存内计算存储器件

存内计算电路可基于易失性存储器和非易失存储器件实现。易失性存储器在设备掉电之后数据丢失，如 SRAM 等；非易失性存储器在

设备掉电后数据可保持不变，如 NOR Flash、阻变随机存储器（Resistive Random Access Memory，RRAM)）、磁性随机存储器（Magnetoresistive Random Access Memory，MRAM)）、相变存储器（Phase Change Memory，PCM）等。

图表14： 五种主流存内计算器件性能对比分析

器件	SRAM	NOR FLASH	RRAM	MRAM	PCM
易失特性	易失	非易失	非易失	非易失	非易失
多值存储	否	是	是	否	是
现有工艺节点	5nm	28nm	28nm	16nm	28nm
理论工艺极限	2nm	14nm	5nm	5nm	5nm
单比特存储面积（F^2/bit[1]）	~300	~7.5	20~40	~30	~24
读写次数	无限	10^6	10^8	~10^15	10^8
应用场景	云侧和边侧的推理和训练	边侧和端侧的推理	云侧、边侧和端侧的推理	云侧和边侧的推理和训练	云侧、边侧和端侧的推理

资料来源：《存算一体白皮书（2022年）-中国移动研究院》，方正证券研究所

图表15： 五种主流存内计算器件的研究与应用进展

名称	特征	研究进展	应用进展
SRAM	✓ 通过开启阵列的多行字线来读取存储器数据，并进行计算。开启的字线数越多，计算并行度越高，系统能效越高，但计算精度会受到影响。 ✓ SRAM的存取速度是所有主流存储器中最接近CPU的，基于它进行存内计算开发，最容易解决内存墙问题。	✓ 基于传统6T SRAM的存内计算技术，为了实现更复杂的运算，研究者提出了不同结构的SRAM单元，如分列式字线的6TSRAM，用作转置单元的8TSRAM，能存储2 bit权重的双8TSRAM。 ✓ **2016年，Intel**基于SRAM实现了支持逻辑操作的可配置存储器，在此基础上实现了支持无进位乘法运算的计算型cache； ✓ **2018年，Intel**发布面向深度学习算法的神经 Cache，实现加法、乘法和减法操作； ✓ **2021年**的国际固态电路会议（ISSCC）上，**台积电**提出了一种基于数字改良的SRAM设计存内计算方案，可支持更大的神经网络。	✓ **九天睿芯：**基于神经拟态感存算一体架构的芯片已实现量产，应用于智能语音和视觉识别领域。 ✓ **后摩智能：**基于SRAM的存算一体大算力芯片，已成功点亮并跑通算法模型。 ✓ **苹芯科技：**开发实现多款基于SRAM的存内计算加速单元并实现流片，目前处于外部测试和demo阶段。产品应用于图像识别、无人机等领域。
DRAM	✓ 每次执行运算时，DRAM存储单元存储的数据会被破坏，每次运算后需要刷新，导致功耗较大； ✓ 单位面积容量大，适合大算力芯片。	✓ DRAM适用于大算力AI芯片，对于架构的改变最小，因此落地快。 ✓ **2017年，三星**联合圣芭芭拉大学推出DRISA架构，基于DRAM工艺实现了卷积神经网络的计算功能，提供大规模片上存储的同时也提供较高的计算性能； ✓ **2022年，SK海力士**公布了基于GDDR接口的浮点数DRAM近存计算的最新研究成果，用于减少GPU模组内的数据搬运。	✓ **阿里达摩院：**基于DRAM的3D键合堆叠存算一体AI芯片，应用于自身生态。
RRAM	✓ 器件结构简单、与CMOS工艺兼容性高，且器件尺寸可缩小； ✓ RRAM具备多组态，从而模仿生物大脑中神经突触功能，同样适合类脑计算； ✓ 适合片上存储和存内计算，数据无需在存储单元和片下存储器之间移动，避免“存储墙”问题，可并行处理大量数据，与神经网络运算的适度高。	✓ 目前研究主要集中在器件性质、小规模阵列的基本逻辑操作以及算法、架构优化，基于忆阻器的存算融合架构为近期研究热点；但其材料不稳定，预计5年内可达到成熟工艺。 ✓ **2016年，惠普**实验室设计了一种转换算法，将任意矩阵值以阻值的形式映射到交叉存储阵列的记忆电阻中，并用闭环脉冲来调节器件阻值的精度； ✓ **2016年，加州大学圣塔芭芭拉分校的谢源教授团队**提出利用RRAM构建基于存算一体架构的深度学习神经网络(PRIME)； ✓ **2019年，杜克大学李海教授与陈怡然教授联合中国台湾新竹清华大学张孟凡教授**完成首个基于RRAM的实际芯片CNN演示。	✓ **亿铸科技：**基于RRAM研发“全数字存算一体”大算力芯片，通过减少数据搬运提高能效比，同时利用数字存算一体保证运算精度，适用于云端AI推理和边缘计算。
PCM	✓ PCM通过相变材料相态的变化获得不同的电阻值，主要用于独立式存储； ✓ 相变存储器目前尚未有明确物理极限，当相变材料厚度到达2nm时，器件仍可发生相变。	✓ 很多难点有待攻克，大多机构研发进展并不顺利，能实现小规模量产的只有三星、美光等海外大公司。 ✓ **2016年，IBM苏黎世研究院**在《Nature》发文，称创造出了世界上首个人工纳米级的随机相变神经元，可用于创造人工神经元； ✓ 2018年，**IBM**在Nature期刊发表的论文提出了全新芯片设计的方案，通过PCM存储技术来加速全连接神经网络的训练，且该芯片可以达到 GPU 280 倍的能源效率，并在同样面积上实现 100 倍的算力。	-
MRAM	✓ 通过铁磁材料相对的磁化方向表现出高低两种阻值，从而实现信息的非易失存储； ✓ MRAM具备极快的开关速度、低功耗和无限写入次数特征，适用于消耗大量计算资源的神经网络计算。	✓ 三星使用电阻加和，降低了支路上的电流，解决了MRAM器件电阻较小的问题。 ✓ **2022年，三星**研究团队设计了一种名为“电阻总和”(resistance sum)的新型内存内计算架构，取代标准的“电流总和”(current-sum)架构，成功开发了一种能演示内存内计算架构的MRAM阵列芯片，命名为“用于内存内计算的磁阻内存交叉阵列”。	-
NOR FLASH	✓ 器件工艺成熟，研发成本低； ✓ 存储阵列大，能够实现大规模运算，适合人工智能和深度学习应用。	✓ **2018年，UCSB**的Dmitri B. Strukov教授发明了在浮栅晶体管技术完成类脑计算的电路； ✓ **2022年，Mythic**发布了基于Nor Flash的存内计算片上系统，用于人物动作识别。	✓ **恒烁半导体：**推出基于NOR Flash的存算一体AI推理芯片，聚焦边缘计算领域，适用于物联网终端设备。 ✓ **知存科技：**基于NOR Flash的存算一体SoC芯片实现量产，应用于智能可穿戴设备，智能安防等端侧小算力场景。

资料来源：《存算一体芯片深度产业报告—量子位智库》，方正证券研究所

3 存算一体：在云、边、端大有可为

图表16： 存算一体技术应用

	特征	优势	代表产品
智能可穿戴设备 2MB-100GOPS	✓ 可穿戴设备总是处于工作、待机或可存储状态。 ✓ 对于低功耗需求强烈，待机时间是产品竞争力的核心。	✓ 存算一体技术能够减少不必要的数据搬运，功耗相较传统的芯片降低10-20倍，符合可穿戴设备对低功耗的需求。 ✓ 在极低功耗的基础上，存算一体在人工智能加速上比当前芯片的效率提升几十到几百倍不等。	✓ **九天睿芯：**ADA100，功耗为同类芯片1/10； ✓ **Syntiant：**NDP102，与当前基于MCU的架构相比，效率和性能提高了100倍。
智能安防 **（智能摄像机）** 32MB-16TOPS	✓ 偏视觉类的垂直场景，算法已相对稳定，对于初创公司来讲能够以较小的成本突破传统大厂的生态壁垒。	✓ 存算一体的高并行计算能力使得计算的实时性比传统芯片高出很多。	✓ **闪易半导体：**闪锌石 HEXA01，计算效率比同类 芯片提升10倍。
移动终端 64MB-32TOPS	✓ 云端推理因网络延迟带来用户体验的问题；受制于手机电池，对芯片的功耗有严格限制。	✓ 存算一体在视觉信号处理上可以达到端侧产品低功耗要求	-
AR/VR 128MB-64TOPS	✓ AR需要处理目标识别、定位、跟踪和建模等人工智能和计算机视觉问题，且计算量大； ✓ 此外，AR/VR眼镜中的电池小、散热差，对低功耗都有较高的要求。因此，在SoC设计方法上需要做出改变以同时满足高性能和低功耗的需求。	✓ 轻薄是AR/VR眼镜的必然趋势。在电池技术没有突破的情况下，芯片功耗需要大幅下降，因此存算一体非常适合嵌入到SoC当中； ✓ AR/VR场景中会涉及较多的人工智能交互（如语音识别，手势识别），存算一体在计算效率和实时性上的优势也可以发挥出来，为用户提供 更真实通畅的交互场景。	**Mythic：**Mythic AMP，拥有四个模拟矩阵处理器， AI计算性能达100 TOPs，支持多达 3.2 亿个权重，以低于25W的功率处理复杂的AI工作负载。
自动驾驶 512MB-256TOPS及以上	✓ 对芯片的散热、实时性及可靠性有高要求。	✓ 存算一体技术低功耗和低延迟的特 性能够很好地匹配自动驾驶的需求；存算一体技术可以在较低的成本下把算力做大； ✓ 自动驾驶场景的算法演进没有那么快，对于初创公司来说能够以较小的代价突破芯片大厂的生态壁垒。	**后摩智能：**首款芯片，样片算力达20TOPS，可扩展至200TOPS，计算单元能效比高达20TOPS/W，在相同功耗下提供10倍算力。

注：2MB-100GOPS：把算法存在2MB的空间中，同时2MB空间可以提供一定的算力进行向量-矩阵运算

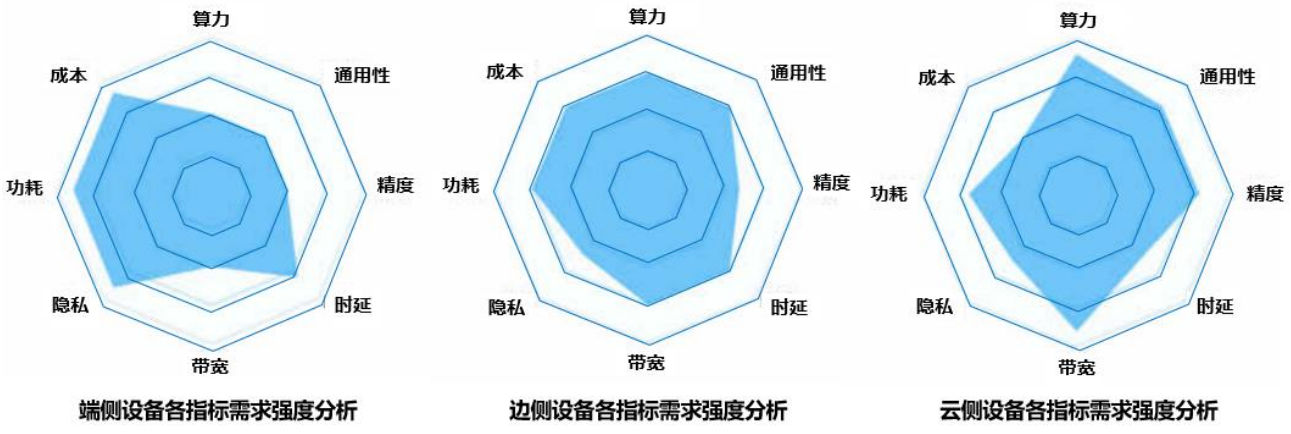
资料来源：《存算一体芯片深度产业报告—量子位智库》，方正证券研究所

根据陈巍谈芯，按算力大小划分：

1）针对端侧的可穿戴等小设备，对算力的要求远低于智能驾驶和云计算设备，但对成本、功耗、时延、开发难度很敏感。端侧竞品众多，应用场景碎片化，面临成本与功效的难题。存算一体技术在端侧的竞争力影响约占 30%。（例如 arm 占 30%，降噪或 ISP 占 40%，AI 加速能力只占 30%）；

2）针对云计算和边缘计算的大算力设备是存算一体芯片的优势领域。存算一体在云和边缘的大算力领域的竞争力影响约占 90%。边缘端存算一体芯片具有高算力、低功耗、高性价比的优势。

图表17： 端侧、边侧、云侧设备各指标需求强度分析



资料来源：《存算一体白皮书（2022 年）-中国移动研究院》，方正证券研究所

3.1 端侧应用场景

3.1.1 端侧单设备算力需求约为 0.1~64 TOPS

据 IDC 预测，2025 年全球物联网设备数将超过 400 亿台，产生的数据量接近 80ZB，智慧城市、智能家居、自动驾驶等诸多场景中超过一半的数据需依赖终端本地处理，单设备算力需求约在 0.1~64 TOPS 之间。

3.1.2 端侧设备对运行时间、功耗、便携性等有较高要求

如智能眼镜/耳机需保证满负荷待机时间超 16 小时，手机的最高运行功耗则不超 8W。端侧设备的未来发展将更加注重时延、功耗、成本和隐私性等需求特征。

3.1.3 存内计算在功耗与计算效率等方面具备明显优势

在相同制程工艺下，存内计算芯片可在单位面积下提供更高算力与更低功耗，进而延长设备工作时间。目前存内计算产品已成功在端侧初步商用，提供语音、视频等 AI 处理能力，并获得十倍以上的能效提升，有效降低端侧成本。

图表18： 端侧小算力企业概览

企业名称	场景	架构类型	存储器类型	主力产品	算力（TOPS）
闪存半导体/闪艺	端侧小算力	模拟存内计算	闪存/自主核心工艺	语音/图像HEXA01	未公布 能效比明显优于某家
SST/Cypress		模拟存内计算	闪存/SF	memBrain IP核	未公布
知存科技		模拟存内计算	闪存/SF	语音WTM2101	0.05@INT8(50GOPS)
每刻深思		模拟存内计算	SRAM	未公布	未公布
九天睿芯		模拟存内计算	SRAM	图像ADA20X	0.3-200@INT8
恒烁股份		模拟存内计算	闪存/ETOX	CiNOR	未公布
新忆科技		模拟存内计算	RRAM	Xuanwu	未公布
智芯科		模拟存内计算	SRAM	语音AT660x	未公布
苹芯科技		存内计算	SRAM	图像PIMCHIP-S200 语音PIMCHIP-S100	未公布

资料来源：陈巍谈芯，创业芯睿，方正证券研究所

3.2 边侧应用场景

3.2.1 边侧单设备算力需求约为 64~256 TOPS

随着车联网等边缘计算应用的快速兴起，海量数据将在边缘侧进行处理，流量模型逐渐从云侧扩展到边侧。

3.2.2 边侧设备对时延、功耗、成本以及通用性等有较高要求

比如智慧港口要求端到端时延 10~20ms，车联网场景要求端到端时延 3~100ms。此外，由于边侧设备通常部署在等靠近数据生产或使用的场所，对散热要求也比较高。

3.2.3 存算一体在深度学习等领域具备独特优势

与传统方案相比，存算一体在深度学习等领域可以提供比传统设备高几十倍的算效比，此外存内计算芯片通过架构创新能提供综合性能全面兼顾的芯片及板卡，预计将为广泛的边缘 AI 业务提供服务。

3.3 云侧应用场景

3.3.1 云侧大算力、高带宽、低功耗需求催涨 AI 芯片

以图像、语音、视频为主的非结构化数据呈现高速增长趋势，根据 IDC 预测，到 2030 年将带动智能算力需求增长 500 倍，以 AI 算力为核心的智算中心将成为算力基础设施主流，大规模的 AI 芯片集约化建设带来高功耗挑战，每机架平均功耗将由 3~5kw 逐渐升至 7~10kw。未来智算中心呼唤新型 AI 芯片，以满足云侧大算力、高带宽、低功耗等特性。

3.3.2 存内计算——智算中心下一代关键 AI 芯片技术

存内计算可通过多核协同集成大算力芯片，结合可重构设计打造通用计算架构，正面向大算力、通用性、高计算精度等方面持续演进，有望为智算中心提供绿色节能的大规模 AI 算力。

针对智能驾驶、数据中心等大算力应用场景，在可靠性、算力方面有较高要求，云计算市场厂商相对集中，存算一体芯片以其高能效大算力优势有望另辟蹊径抢占云计算市场。

图表19： 云和边缘大算力企业概览

企业名称	场景	架构类型	存储器类型	主力产品	算力（TOPS）
亿铸科技	边缘为主大算力（ADAS）	全数字存算一体	RRAM	未公布	未公布
千芯科技	云和边缘大算力	存内计算/存内逻辑	RRAM/SRAM	云计算卡G40710E；G41210E；F11610E；F12010	>1000-4000@INT8
后摩智能	边缘为主大算力（ADAS）	模拟存内计算	SRAM/MRAM/RRAM	智能驾驶芯片	20
中科声龙	云为主大算力	近存计算	SRAM	矿机	-
d-Matrix	Transformer加速	近存/存内	SRAM	计算卡	未公布

资料来源：陈巍谈芯，创业芯睿，方正证券研究所

4 存算一体技术三大驱动因素

4.1 新型存储器的发展

新型存储器件的物理性能更适合开发存内计算，在实现更高计算密度的同时具备成本优势。在新型存储器件上发展存算一体技术，能够带来更大的算力优势，从而开拓更多的人工智能应用场景。此外，新型存储器件的发展上限更高，现有存储器件再过 3-4 年将走向技术极限，而**新型存储器件还可以往前发展 10-20 年**。根据量子位智库，基于 RRAM 的新型存储器件有望在 5 年内在产品化上取得突破。新型存储器件的特点是在其开发过程中需要在传统 CMOS 工艺里增加特殊材料或工艺，这些特殊材料或工艺的开发需要经过大量实验及测试验证，而传统的 CMOS 代工厂在开发进度上相对缓慢。

因此，新型器件工艺的突破点主要是工艺的迭代速度，如果没有标准的 12 寸量产产线，新型存储器件很难走向量产。如果新型存储器发展受限，在传统存储器件走到成熟尽头后，开拓新应用场景的难度会极大。

4.2 应用侧需求

后摩智能认为存算一体的发展逻辑是由外向内的，当大量需求出现后，一项能够满足客户需求的新技术将迅速发展。在存算一体领域，AI、大数据分析这类数据密集型应用的出现，对能效比的需要迅速上升，推动了存算一体的发展。存算一体的底层逻辑是让很大一部分数据不需要搬出存储器便可参与计算，以此大幅提升计算效率。同时，随着深度学习被广泛应用，对算力的需求不仅仅是大算力，有效算力也成为企业关注的焦点。在传统的冯·诺依曼架构下，存储单元和计算单元分离，存储器读写速度慢产生的时延，在一定程度上造成算力浪费。尽管处理器的性能再优，依然需要平衡存储器的特性，存储器运行速度慢导致实际运算效率不及理论上所呈现的指标。而存算一体架构通过使存储器具备计算能力，实现在相同芯片面积下规模化增加计算核心数。

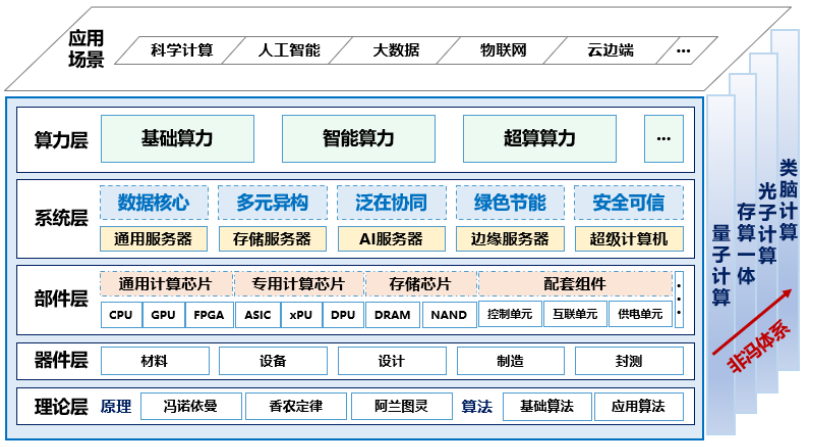
4.3 产业侧配合

存算一体技术在 0 到 1 的阶段已初步形成 IP 授权，定制开发，自定义产品多种商业模式，能够在特定应用场景中实现小规模量产。一旦产品出现可大规模量产的趋势或能够产生足够的收益，整个产业链便会积极加入，在生产制造的各个环节都将有相应公司专门基于存算一体做研发，共同推动整个产业发展。

根据量子位智库，从目前小规模量产到实现大规模量产大概有 10 年的时间，其中前 5 年存内计算将以 AI 计算为主，后 5 年将覆盖更多应用场景。在底层技术上，选择正确的方向和适配的场景决定了技术在未来是否有足够的潜力走向产业化。

5 存算一体技术三大应用方向

图表20： 先进计算技术产业体系框架



资料来源：《中国算力发展指数白皮书（2022 年）》，方正证券研究所

5.1 AI 和大数据计算

存内计算适用于 AI 的深度神经网络应用和基于 AI 的大数据技术。通过存算一体技术，可将带 AI 计算的大量乘加计算的权重部分存在存储单元中。对此，在电路设计方面，通过在存储单元的核心电路上做修改，从而在读取的同时进行数据输入和计算处理，在存储阵列中完成卷积运算，再加上大量乘加的卷积运算是深度学习算法中的核心组成部分，存内计算助力基于 AI 的大数据技术。

5.2 感存算一体

存算一体助力含 AI 存算一体芯片的传感器实现零延时和超低功耗的智能视觉处理能力。集传感、储存和运算为一体构建感存算一体架构，在解决冯·诺依曼架构的存储墙瓶颈的同时，与传感结合提高整体效率。1）在传感器自身包含的 AI 存算一体芯片上运算，可实现零延时和超低功耗的智能视觉处理能力；2）基于 SRAM 模数混合的视觉应用存内计算神经拟态芯片仅在检测到有意义的时间才会进行处理，大幅降低能耗。

5.3 类脑计算

存算一体为类脑计算的关键技术基石。类脑计算又被称为神经形态计算，是借鉴生物神经系统信息处理模式和结构的计算理论、体系结构、芯片设计以及应用模型与算法的总称。本质与存算一体类似，类脑计算旨在使计算机像人脑一样将存储和计算合二为一，从而高速处理信息。由于类脑计算属于大算力高能效领域，因此针对云计算和边缘计算的存算一体技术，是未来类脑计算的首选和产品快速落地的关键。

6 存算一体公司竞争格局：国外存算一体产业比国内起步早 3-5 年左右

从发展进程上讲，国外存算一体产业比国内起步早 3-5 年左右，并且基于存算一体的技术已普遍实现产品化。目前来看，**SST，Syntiant 和 Mythic 走在商业化前列**；SST 的 IP 授权数量最多，且许多芯片大厂愿意为其买单。

从芯片营收上讲，经量子位智库统计，美国超过 100 亿美元营收的芯片公司有 10 家左右，欧洲有 5 家左右，而中国只有 1-2 家（中国公司数量远超国外）。

从技术得到验证到产品化过程的前期，存算一体配套工具（如 EDA 软件）的研发公司尚处在探索阶段。缺乏成熟的配套工具导致基于存算一体技术的产品在短期内（5 年左右）以小规模量产为主。

图表21： 中国存算一体芯片公司

公司	定位	融资轮次	产品及解决方案	技术亮点	合作
九天睿芯	神经拟态感存算一体架构芯片	A轮	✓ ADA20X：低功耗中低算力视觉协处理器，用于 多视觉场景的数模混合AI视觉芯片； ✓ ADA10X：时序信号多传感器处理芯片&超低功耗传感器处理芯片，应用于可穿戴AIOT设备； ✓ ADC芯片：应用于通信（车载激光雷达）、仪器仪表行业	✓ 采用前端模拟预处理（ASP）+模数转换(ADC)+模拟加速器（ADA）架构，将感、存、算集合为一 体	歌尔、……
阿里达摩院	基于DRAM的3D键合堆叠存算一体AI芯片	-	-	✓ 存储芯片：采用异质集成嵌入式DRAM； ✓ 计算芯片：流式定制化加速器架构 ✓ 封装：3D混合键合技术	内部赋能
后摩智能	基于存算一体技术的大算力AI芯片	Pre A+	✓ 第一代芯片：基于SRAM-CIM技术快速构建存算一体核，并以此核搭建存算一体芯片； ✓ 第二代芯片：基于MRAM/RRAM等存储工艺，继续扩充模型容量	✓ 基于数字域存内计算的电路	-
知存科技	基于存算一体技术的人工智能芯片	B2	✓ WTM2101：国际首个量产存算一体SoC芯片； ✓ WTM1001：智能语音芯片； ✓ WTM3000：针对图像分析应用的存算核心技术，面向智能安防、移动终端场景	✓ 使用Flash存储器同时完成神经网络的存储和运算	科大讯飞、芯来科技、……
苹芯科技	存算一体芯片与非冯架构智能算力平台	A	✓ S230：存内计算智能感知决策芯片； ✓ S200：基于SRAM架构的存内计算加速器，可将深度学习算法中占主导的基本运算在存储器 内部完成 ✓ S100：基于存算一体技术的智能语音处理器	✓ 基于SRAM及新型存储器存内计算技术，打造非冯架构计算体系	-
千芯科技	大算力存算一体AI芯片	天使+轮	✓ AI计算卡：先进存算架构，深度优化存储墙与编译墙，提供更强更高效的大模型支持； ✓ 边缘AI计算板卡：支持边缘计算的灵活算法部署与客户自定义算子，为各类边缘计算提供算力支持； ✓ AI计算IP核：多种存算IP核解决方案	✓ 针对大算力的存内逻辑/存内计算创新架构； ✓ 支持CUDA语法	某互联网大 厂、兆易创新、阿里平头哥、芯来科技、昕原半导体
恒烁半导体	基于Nor Flash技术的存算一体终端推理芯片	上市	✓ 第一版CiNOR存算一体AI推理芯片已经成功流片， 并且搭载该芯片现场演示了一个人脸识别的深度学习算法	✓ 通过Flash阵列的模拟计算来高度并行化完成矩阵计算	-
杭州智芯科	面向边缘计算的大算存内计算SoC	天使轮	✓ AT680x：超低功耗智能语音芯片； ✓ AT700：超低功耗AI终端处理器芯片，用于离线语音和图像识别	✓ 先进的数据流架构； ✓ 由SDK驱动带来的通用性大算力模拟内存计算	-
闪易半导体	存算一体AI芯片	股权融资	✓ HEXA01：首款集成PLRAM忆阻器阵列的SoC芯片， 可支持多种神经网络模型，可应用于家电和物联网设备的智能控制	✓ 基于双向Fowler-Nordheim隧穿进行擦写的闪存忆阻器	-
新忆科技	新型存储器技术研发	股权融资	✓ 新型阻变存储器（RRAM）及其周边产品，包括独立式存储器、嵌入式存储器和周边的SOC产品	✓ 新一代非易失性存储器技术	SK海力士、集创北方、中芯国际、联华电子
中科声龙	存算一体高通量算力芯片	计划2024年在北交所上市	✓ 茉莉X4：高通量算力芯片，为区块链网络提供算力支持	✓ 基于片内超大规模全相联网络的高通量算力芯片，实现计算核心与数据通路之间的性能平衡	-
亿铸科技	基于RRAM的全数字存算一体大算力AI芯 片	天使轮	✓ 大算力、高能效比、高精度、易编译的存算一体PCIe加速卡; ✓ 高性价比、确定性时延自动驾驶存算一体Chiplet模组; ✓ 首套针对存算一体架构的软硬件协同EDA设计工具和应用开发平台	✓ 核心IP均为自研，软件-架 构-芯片-工艺-制造均可实现自主可控和国产化; ✓ 全数字路线，解决模拟计算精度不高等问题	昕原半导体
台积电	前沿技术探索	上市	✓ 2022年的ISSCC上合作发表了六篇关于存内计算存储器IP的论文，大力推进基于ReRAM的存内计算方案; ✓ 2021年初的国际固态电路会议（ISSCC）上提出了一种基于数字改良的SRAM设计存内计算方案，能支持更大的神经网络。	✓ 通过扩展常规SRAM阵列，提供了一种高面积效率的存内计算方法，支持可编程位宽、有符号或无符号以及4种不同位宽权重的输入激活	-

资料来源：《存算一体芯片深度产业报告—量子位智库》，方正证券研究所

图表22： 海外存算一体芯片公司

公司	定位	融资轮次	产品及解决方案	技术亮点	合作
Mythic	基于Nor Flash的存算一体芯片	C轮	✓ M1076模拟矩阵处理器； ✓ MP10304 Quad-AMP PCIe 卡； ✓ MM1076 M.2 M 钥匙卡； ✓ ME1076 M.2 A+E 钥匙卡； ✓ MNS1076 AMP 评估系统	✓ 数据流架构； ✓ 工具链的开发，包括优化套件和编译器	高通、……
Syntiant	使用模拟神经网络计算的深度神经网络处理器	D轮	✓ NDP100神经决策处理器：通过高度耦合的计算和内存实现突破性性能； ✓ NDP120：应用神经处理器以最小的功耗同时运行多个应用； ✓ Tiny ML开发工具包； ✓ 软件开发套件包含控制和操作 Syntiant NDP 设备所需的工具和 软件	✓ 基于模拟神经网络进行大规模并行乘法累加计算	亚马逊、瑞萨电子
d-Matrix	基于SRAM的AI推理芯片	A轮	✓ Nighthawk芯片：基于小芯片架 构，一种使用存内计算技术和小芯片级横向扩展互连进行数据中心AI推理的新方法； ✓ AI计算平台：结合了智能ML工具 和无摩擦软件方法，并结合类似乐高形式的小芯片，将多个编程引擎集成到一个通用封装中	✓ 内存计算定制数字电路； ✓ 开放、简单、可扩展且无摩擦，易于采用的软件； ✓ 机器学习算法及工具，可无缝映射到现有训练模型； ✓ Hetero-Modular封装技术	台积电、微软Azure
Crossbar	基于RRAM的存算一体芯片	D轮	✓ 高性能存储器非易失性存储器IP核	✓ 基于模拟神经网络进行大规模并行乘法累加计算	亚马逊、瑞萨电子
SST	高度可靠和通用的 NOR Flash 技术	被Microchip收购	✓ memBrain 神经形态内存产品：使用模拟内存计算，将突触权重存储在Flash存储器内，以显著改善系统延迟	✓ 基于 SuperFlash技术， 计算用于神经网络推理的 向量矩阵乘法 (VMM)， 通过模拟内存计算方法改进了VMM的系统架构实现，增强了边缘的AI推理	SK海力士 、联华电子……
IMEC	研究机构	-	✓ 基于DRAM和NAND-Flash的存内计算； ✓ 基于新型存储器的研究，如 MRAM，铁电等	✓ 模拟内存内计算（AiMC） 架构	格罗方德
IBM	相变存储技术(PCM)研究	上市	✓ 非易失存储器研究	✓ 基于相变存储器的芯片技术：像人脑一样在存储中执行计算任务，以超低功耗实现复杂且准确的深度神经网络推理	-
SK海力士	基于GDDR接口的DRAM存内计算	上市	✓ 基于PIM技术的产品：GDDR6-AiM	✓ 构建全新的存储器解决方案生态系统	-
三星	基于MRAM的存内计算研究	上市	✓ 通过构建新的MRAM阵列结构，用基于28nmCMOS工艺的MRAM阵列芯片运行了手写数字识别和人脸检测等AI算法，准确率分别为98%和93%	✓ 通过用新的“电阻和”存内计算架构替换标准的“当前和”存内计算架构来演示存内计算，解决了MRAM器件低电阻的问题	-

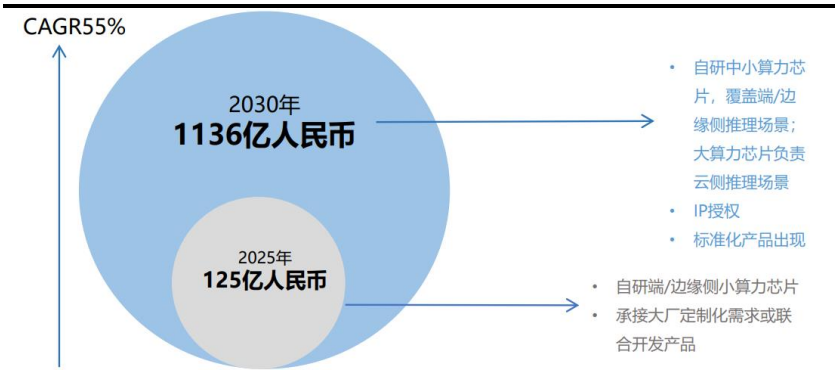
资料来源：《存算一体芯片深度产业报告—量子位智库》，方正证券研究所

7 基于存算一体技术的小算力芯片 2025 年约 125 亿人民币远期市场空间

根据量子位智库测算，**2025 年，基于存算一体技术的小算力芯片市场规模约为 125 亿人民币**。存算一体技术从实验室的研究成果到实现初步量产需要 5 年左右的时间，从初步量产到大规模量产则需要 10 年左右时间。国内存算一体公司从成立时间上看，集中在 2017-2020 年，其中实现量产的公司有 4 家左右，其余公司中进入测试阶段的有 2-3 家。**量子位智库预计，2025 年存算一体将迎来商业化转折点，应用场景从麦克风、智能手表和 TWS 耳机拓展到智能安防、移动终端和 AR/VR 等（从语音识别、唤醒到视觉处理）。**

2030 年，基于存算一体技术的**中小算力**芯片市场规模约为 **1069 亿人民币**，基于存算一体技术的**大算力**芯片市场规模约为 **67 亿人民币**，总市场规模约为 1136 亿人民币。大算力芯片和小算力芯片在底层的存算一体单元基本可以复用，但 NPU 架构和编译器需要做一定修改以支持更通用的场景。除了提升芯片设计能力，使用新型存储器也能够增加单个芯片的算力。RRAM 新型存储器技术具有高速、结构简单 的优点，有望成为未来发展最快的新型存储器，目前距离工艺成熟还有 2-5 年的时间。考虑到从技术突破到产品化还需要 2-3 年的时间，**量子位智库预计在 2030 年，基于存算一体的大算力芯片将实现规模量产**，应用场景覆盖大数据检索、蛋白质/基因分析、数据加密、图像处理等。

图表23： 中国存算一体芯片市场规模估计



资料来源：《存算一体芯片深度产业报告—量子位智库》，方正证券研究所

8 相关厂商

8.1 恒烁股份：CiNOR 存算一体 AI 推理芯片方兴未艾

高并行度和高能效计算催涨存算一体需求，CiNOR 存算一体 AI 推理芯片方兴未艾。2019 年公司研发的存算一体 AI 推理芯片（恒芯 1 号）流片和系统演示成功，目前在研 CiNOR V2 芯片（恒芯 2 号）。随着存算一体技术的深化应用，公司的 CiNOR 存算一体 AI 推理芯片前景可期。

估值：Wind 一致预期 23、24 年摊薄 EPS 分别为 1.00、1.77，对应 76X、43X PE。

8.2 知存科技（非上市）：深耕存内计算芯片领域，引领存内计算产业化

知存科技创立于 2017 年，专注存内计算芯片领域，创新使用 Flash 存储器完成神经网络的储存和运算，解决 AI 的存储墙问题，提高运算效率，降低成本。研发团队由王绍迪博士与郭昕婕博士联合多位学者、产业从业者组建，平均拥有 10 年以上产业工作经验。公司旗下 WTM2101 芯片适配低功耗 AIoT 应用，可使用微瓦到毫瓦级功耗完成大规模深度学习运算，可应用于智能语音、智能健康等市场领域，目前已完成批量生产和市场应用。WTM8 系列芯片面向 6-48Tops 算力产品，应用于 4K-8K 视频的实时处理。

2023 年 1 月，知存科技完成 2 亿元 B2 轮融资，累计融资近 8 亿元。未来，公司将继续专注存内计算芯片领域，引领存内计算产业化。

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格，保证报告所采用的数据和信息均来自公开合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地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。研究报告对所涉及的证券或发行人的评价是分析师本人通过财务分析预测、数量化方法、或行业比较分析所得出的结论，但使用以上信息和分析方法存在局限性。特此声明。

免责声明

本研究报告由方正证券制作及在中国（香港和澳门特别行政区、台湾省除外）发布。根据《证券期货投资者适当性管理办法》，本报告内容仅供我公司适当性评级为C3及以上等级的投资者使用，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。若您并非前述等级的投资者，为保证服务质量、控制风险，请勿订阅本报告中的信息，本资料难以设置访问权限，若给您造成不便，敬请谅解。

在任何情况下，本报告的内容不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求，方正证券不对任何人因使用本报告所载任何内容所引致的任何损失负任何责任，投资者需自行承担风险。

公司投资评级的说明：

强烈推荐：分析师预测未来半年公司股价有20%以上的涨幅；
推荐：分析师预测未来半年公司股价有10%以上的涨幅；
中性：分析师预测未来半年公司股价在-10%和10%之间波动；
减持：分析师预测未来半年公司股价有10%以上的跌幅。

行业投资评级的说明：

推荐：分析师预测未来半年行业表现强于沪深300指数；
中性：分析师预测未来半年行业表现与沪深300指数持平；
减持：分析师预测未来半年行业表现弱于沪深300指数。

地址	网址：https://www.foundersc.com	E-mail:yjzx@foundersc.com
北京	西城区展览馆路 48 号新联写字楼 6 层	
上海	静安区延平路 71 号延平大厦 2 楼	
深圳	福田区竹子林紫竹七道光大银行大厦 31 层	
广州	天河区兴盛路 12 号楼 隽峰苑 2 期 3 层方正证券	
长沙	天心区湘江中路二段 36 号华远国际中心 37 层	