

Compute in Memory

存算一体芯片

Qbitai Industry Insight

深度产业报告

- 技术篇
- 产业篇
- 展望篇

引言

2021年，全球半导体产业销售总额共计5,559亿美元，其中，中国以1,925亿美元的半导体销售额成为全球规模最大的区域市场。在这个千亿美元市场中，有价值数亿元的高端制造设备，也有价格几元到几十元不等的芯片。

今天，全球半导体市场的竞争格局相对稳定，在产业链各环节都由市场份额占绝对优势的企业主导，这些企业凭借着极高的技术壁垒，在击退新玩家的同时也掌握着行业的发展动向。

半导体产业“牵一发而动全身”的特性导致新玩家很难凭一己之力改变行业内的某些规则，改变往来自应用侧的真实需求。正是基于这个特性，我们所看到的半导体行业创新案例往往具备更高的可靠性。

半导体行业随着应用侧技术的发展而演进，不断往下做的先进工艺正是因为人工智能的发展，对于计算规模和计算速度的要求在不断提高。当先进工艺走到7nm以下，芯片在物理层面的缺点逐渐显现，随之而来的是持续走高的成本投入。

由于当前几乎所有芯片大厂都将主要精力放在先进工艺研发上，产业链上中下游的配合方也都围绕先进工艺开展相应研发。然而，在突破1nm极限后，摩尔定律将由此失效，基于冯诺依曼架构的芯片技术发展将不会再依托先进工艺。

在技术快要走到极限之时，必然会提前出现“掉头”的现象，在半导体领域，“掉头”意味着在芯片从0到1的各个环节寻找新的方法突破瓶颈。

芯片领域的创新包括先进封装技术，新型架构，新材料等各个方向。在其中，我们关注到基于存算一体架构的芯片研发。在采访了部分行业头部机构后，我们希望可以还原存算一体技术的本真，并且能够一探这一领域的真正价值。

感谢以下机构与个人参与深度访谈（按照首字母排序）：

达摩院、后摩智能、九天睿芯、苹芯科技、千芯科技、燕博南、亿铸科技

关键结论

- 存算一体芯片的关键在于存算一体架构，其核心是电路设计革新。
- 目前尚未形成完成的产业链生态，尤其在软件层面缺乏相应的研发公司配合完善技术链条。
- 在产业界和资本一致看好存算一体的现状下，完整的技术链条、对客户需求的把握以及全面的人才储备是初创公司在业内保持竞争力的关键，也是新玩家进入这个赛道需要具备的实力。
- 如果将基于存算一体的芯片放在半导体产业的大背景下，其尚处在发展的早期阶段，在这个阶段业界呈现出一种百花齐放的态势。
- 基于存算一体的产品从初步商业化到大规模商业化的过程中，主要有三点驱动因素：新型存储器的发展，来自应用侧的需求以及产业侧的配合。
- 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。
- 新玩家在选择是否进入这个赛道时，首先要明确目标市场，在此基础上要厘清技术与需求的匹配度，真正理解客户的痛点以及针对这个痛点，判断存算一体是否有足够的优势吸引客户。



目录



技术篇

- 1.1 技术简介
- 1.2 技术价值
- 1.3 技术路径
- 1.4 关键技术
- 1.5 技术挑战与展望



产业篇

- 2.1 行业现状与驱动力
- 2.2 市场价值
- 2.3 市场规模
- 2.4 产业链分布
- 2.5 主要玩家及中外竞争对比
- 2.6 进入门槛



展望篇

- 3.1 展望结论

技术篇

1.1 技术简介

• 研究背景

人工智能芯片是人工智能技术发展的硬件基础，在人工智能发展三大要素，数据、算法和算力中，算力主要由人工智能芯片支撑。人工智能芯片目前有两种发展路径：一种是在传统计算架构下的AI加速器/计算卡，主要以GPU, FPGA, ASIC等为代表；另一种路径是颠覆传统的冯诺依曼架构，采用新的架构来提升计算能力，以存算一体芯片为代表。

当前，摩尔定律已逼近极限，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。传统的冯诺依曼架构已无法适应如今AI计算对算力和低功耗的需求，存算一体芯片架构是需求变化中催生出的新型计算架构，在算力和能效比方面相比冯诺依曼架构具有绝对优势。

• 定义

存算一体是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，有望彻底解决传统冯·诺伊曼架构的存储墙问题，极大提高计算能效。由于实现形式不同，目前业内对于存内计算的概念并没有形成非常明确的定义。

1.2 技术价值

过去几十年，半导体行业都是按照摩尔定律在发展。摩尔定律的核心内容是“集成电路上可以容纳的晶体管数目大约每经过18个月便会增加一倍”。在摩尔定律能够持续往下走的时候，每一到两年换一代芯片工艺，整体性能便可提升数倍，成本也会自然降低。在性能提升速度非常快的前提下，产业界不需要进行架构创新便可以不断开拓新的市场空间。

到2010年以后，进入后摩尔时代，人们意识到摩尔定律会走到极限。自2012年以来，AI训练任务的算力需求每3.5个月就会翻倍，这个数字远超过摩尔定律的18月。为了满足算力需求，芯片需要更高的集成度，晶体管的体积变得越来越小。当小到一定程度时（逼近物理极限），便会引发现新现象，如量子隧穿效应。

在冯诺依曼架构下，即使处理器的算力能够做到非常大，但存储器的访问速度远比不上处理器的处理速度，导致处理器的实际性能受到严重制约。当前针对算力需求出现了很多解决方案，如先进工艺、3D堆叠技术等，但这些技术依旧是基于冯诺依曼架构下，仍无法从底层突破瓶颈，很快也会面临技术走到极限的问题。

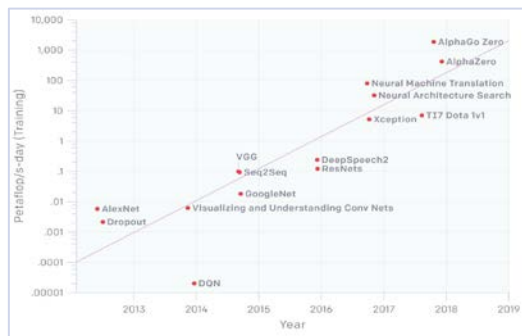


图1 算力需求的增速远超摩尔定律 (图片来源《AI与计算》分析报告)

当前最先进的计算机采用的都是冯诺依曼架构。在这种架构下，数据的处理和存储是分离的，分别由中央处理器CPU和存储器完成。每次执行运算时，需要把数据从存储器搬运到处理器，中间经过数据总线，当数据处理完之后再将其搬回存储器中。

冯诺依曼理论模型的重要假设之一是计算与存储速度相当，如果双方一旦在速度上不匹配，慢的一方将会制约整体计算效率。随着半导体产业的发展，处理器和存储器针对不同的用户需求形成了不同的工艺路线，速度快成为处理器的发展方向，存储器则强调大容量和低功耗，导致处理器的运行速度远快于存储器。

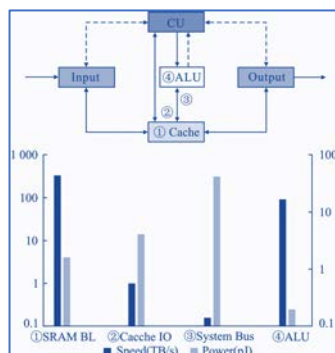


图2 冯诺依曼架构及冯诺依曼瓶颈

随着人工智能技术爆发，计算机每天要在处理器和存储器之间进行高频数据传递，产生大量功耗。谷歌2018年对其产品耗能情况展开调研，结果显示系统能耗的62.7%浪费在CPU和内存的读写传输上。

在数据传递过程中，内存的传输速度跟不上CPU性能，会导致实际算力受限，影响CPU运行处理速度。假设CPU处理运算一道指令的耗时为1ns，内存读取传输该指令的耗时大约在10ns。上述问题所导致的散热需求增加、用电成本上升以及算力浪费都是企业在人工智能技术发展中面临的瓶颈。

面对数据搬运产生的高能耗和存算分离导致的性能瓶颈，存算一体架构能够从根本上解决冯诺依曼瓶颈。存算一体是在存储器内嵌入计算能力，以新的架构进行二维或三维矩阵运算。这种直接利用存储器进行数据处理的方式，消除/缩小了计算单元和存储单元之间的距离，从而解决冯诺依曼瓶颈。

1.3 技术路径

目前，学术界和产业界对存算一体的技术路径尚未形成统一的分类，不同研究领域（器件、电路、架构等）对存算一体的称呼也不尽相同。目前主流的划分方式依照计算单元与存储单元的关系（距离），将其大致分为近存计算和存内计算两种技术路线。

存内计算又包含两种形式，第一种计算操作由位于存储器内部的独立计算单元完成，存储单元和计算单元相互独立存在。第二种是在内部存储中添加计算逻辑，直接在内部存储执行数据计算，这种架构数据传输路径最短，能同时满足大模型的计算精度要求。

近存计算广义上也被纳入存算一体架构，其通过将计算资源和存储资源拉近，实现能效和性能的大幅提升。由于近存计算不涉及改变计算单元和存储单元之间的关系，因此是现阶段最容易实现的技术手段。

分类	特征	实践
近存计算 (Process near Memory)	计算芯片与存储芯片分离； 计算部分通过存储芯片外部的计算芯片完成，将数据靠近计算单元，从而缩小数据移动的延迟和功耗	阿里达摩院于2021年研发出基于DRAM的3D键合堆叠存算一体芯片，在特定AI场景中，其性能提升10倍以上，效能比提升300倍
内存储计算 (Process in Memory)	计算单元和存储单元位于同一芯片中，但电路设计是分离的； 计算部分由存储器内部的独立计算单元完成	三星在2021年发布的HBM2-PIM，其使用Aquabolt-XL技术，围绕HBM2 DRAM存储介质进行内存储计算，可实现高达1.2 TFLOPS的计算能力
内存执行计算 (Process with Memory)	存储单元和计算单元完全融合； 没有独立的计算单元，直接通过在存储器颗粒上嵌入算法，由存储器芯片内部的存储单元完成计算操作	2010年，惠普实验室的Williams教授团队提出并验证利用忆阻器实现简单布尔逻辑功能。 2016年，美国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)。测试表明，相较于冯诺伊曼架构下的方案，PRIME可以实现功耗降低约20倍，速度提高约50倍

存算一体本质上属于计算类产品，通过存储器完成计算。

存算一体的计算方式分为数字计算和模拟计算，不同种类的存储器在特性和成熟度上都有差异，所能实现的计算程度也不尽相同。

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，且具备很好的抗噪能力和可靠性。模拟存算一体通常使用RRAM/SRAM/Flash等存储器件，存储密度大，并行度高，但对环境噪声和温度敏感。

基于易失性存储器SRAM和Flash的存算一体架构工艺成熟，已经实现商业化。目前，利用存算一体技术的公司中，实现量产的均以SRAM和Flash作为存储介质。

近年来，新型非易失存储器在计算密度上的巨大潜力（规模可以做大）和在存算融合性上的天然优势，使其成为未来发展趋势，其中RRAM和MRAM有广泛的应用前景。

名称	特征	研究进展	应用进展
SRAM (Static Random-Access Memory)	<ul style="list-style-type: none"> 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 	<p>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。</p> <p>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；</p> <p>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；</p> <p>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；</p>	<ul style="list-style-type: none"> 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 (Dynamic Random Access Memory)	<ul style="list-style-type: none"> 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。 	<p>DRAM适用于大算力AI芯片，对于架构的改变最小，因此落地快。</p> <p>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；</p> <p>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</p>	<p>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。</p>
RRAM (Resistive Random Access Memory)	<ul style="list-style-type: none"> 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 	<p>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。</p> <p>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；</p> <p>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)</p> <p>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</p>	<p>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。</p>

名称	特征	研究进展	应用进展
PCM (Phase-Change Memory)	<ul style="list-style-type: none"> 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。 	<p>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。</p> <p>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；</p> <p>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</p>	/
MRAM (Magnetoresistive RAM)	<ul style="list-style-type: none"> 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。 	<p>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。</p> <p>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”</p>	/
NOR FLASH	<ul style="list-style-type: none"> 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。 	<p>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</p> <p>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</p>	<ul style="list-style-type: none"> 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。

1.4 关键技术

芯片制造流程主要分为四步：芯片设计、晶圆生产、芯片封装和芯片测试。存算一体芯片的关键在于存算一体架构，其核心是器件和电路设计革新。

生产制造环节相对于设计环节的理论创新，更多聚焦落地。目前市场主流选择为基于Flash和SRAM的存算一体芯片技术，其在标准工艺（CMOS）下即可获得，流片门槛低。实现量产的公司中，九天睿芯使用的是SRAM存储器。封装技术上，知存采用WLCSP的2.6x3.2mm²极小封装，其他家尚未透露。

在近存计算中，主要考验的是先进封装技术，通过在内存单元集成DRAM拉近存储和计算距离，使用3D堆叠技术进行芯片封装。达摩院研发的芯片采用混合键合(Hybrid Bonding)的3D堆叠技术，将计算芯片和存储芯片face-to-face地用特定金属材质和工艺进行互联，拉近存储单元与计算单元的距离增加带宽，降低数据搬运的代价。

在存算一体芯片研发中，与传统芯片最大的不同在于计算单元与存储单元的融合。

从微观角度看，对模拟计算的掌握决定着技术成熟的速度。在这种计算模式下，存储器被视为电阻，乘法计算通过欧姆定律实现，加法计算通过电流的连续性（基尔霍夫电流定律）实现。模拟计算的精度会受到低信噪比的影响，其计算精度目前较低，因此对于阻值本身的精度有较高要求，技术突破就在于对阻值精确度的把握上。

从宏观角度看，存算一体芯片在器件、芯片、算法和应用等多层次的协同是技术落地的关键。细分应用场景的不同性能需求决定了神经网络算法与芯片的设计，算法依赖神经网络框架、编译、驱动、映射等工具与芯片架构的协同，芯片架构又依赖器件、电路与代工厂工艺。其中，尤为重要是新型存储介质，其物理原理、特性、集成工艺都不尽相同，需要跨层协同来实现性能（精度、功耗、时延等）与成本最优。

1.5 技术挑战与展望



挑战

• 计算的挑战：模拟计算精度

模拟计算本身的局限性：模拟抗干扰能力弱，此特性决定其计算精度会受到计算电路中噪声的限制，导致电路面积在电路精度接近噪声极限时快速增加。因此，模拟计算只在低精度的计算中存在优势，而更高精度的计算需要数字电路的辅助。

此外，模拟计算中电流、电压、电阻会随着事件而衰退，一旦芯片的可靠性不过关，芯片参数出现细微的浮动和干扰都会导致计算结果不准确。

外部环境局限：逻辑单元设计主要基于数字计算，模拟计算的理论和经验相对缺乏。另外，在硬件层面实现模拟计算的可控和大容量存储难度较高。

• 存储器的挑战：新型存储器研发

基于新型存储介质的存算一体技术，器件物理原理、行为特征、工艺都不尽相同，需要不同层级相互协同，使性能（精度、功耗、时延等）与成本达到最佳状态。

当前新型存储器的制造工艺还没有解决不同存储单元之间一致性的问题，良率有待提高，单位容量成本很高，存储密度优势不显著。

新型存储器需要开发新的软件配套。当前的成熟计算架构是基于DRAM和NAND Flash为基础构建，硬件电路和软件驱动均是根据传统存储的特点设计。新型存储在物理特性上迥异于传统存储，需要开发新的辅助芯片和对应的软件程序才能使其发挥性能。

• 存储与计算融合过程的挑战

存算一体架构的硬件开销只有控制在一定范围内（至少要明显低于分立的存储单元和逻辑单元的开销总和），才能体现其存在的意义。当前的半导体技术从设计、制造工艺、封装和软件使用等主要还是围绕存储和逻辑，分别进行优化。存内计算真正发展起来，需要将两个独立的生态整合到一起，其中需要投入的精力和资源是巨大的。在存内计算架构中，存储单元和逻辑单元的融合，使得单位逻辑单元可调用的存储信息规模下降，因此需要加强逻辑单元的复用能力，这将带来额外的片上开销。



展望

存算一体芯片的发展方向可以概括为在解决技术瓶颈的同时找到最容易落地的应用场景。由此我们总结出存算一体芯片在技术、应用场景和落地上的趋势：

• 新型存储器技术的发展

RRAM将与神经网络计算深度结合发展全新的计算架构。RRAM在传统存储领域的应用优势不明显，而新的存储架构将会是RRAM的发展机会。例如，将RRAM作为神经网络中突触节点的权重存储单元，或是应用于存算一体架构，从而大幅度提升神经网络芯片的性能并降低功耗。

MRAM将更加广泛地运用于嵌入式系统中。目前独立式的MRAM由于容量难以进一步增长，成本极高，市场应用空间有限。未来MRAM将主要针对嵌入式市场，逐步替代现有的嵌入式闪存技术，有望成为嵌入式系统中的主流存储器。

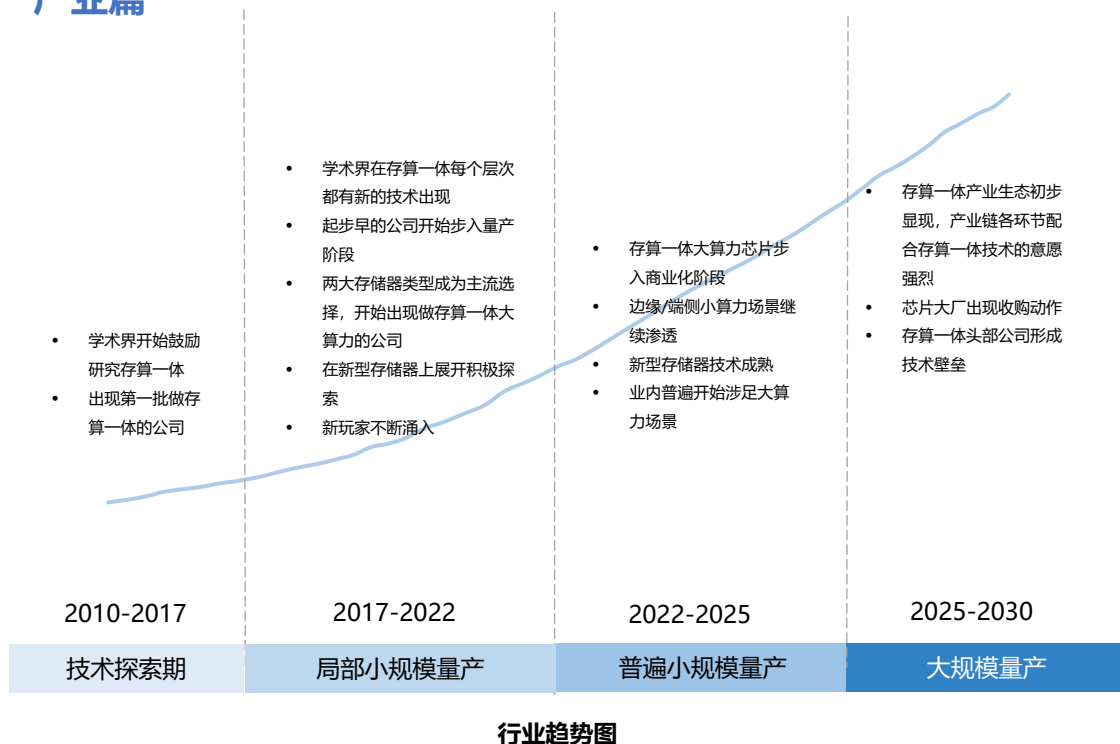
PCM将向更高层数的三维集成发展。目前，唯一商用的PCM产品英特尔傲腾存储器，第一代仅实现了两层三维集成，2020年发布的第二代也仅仅做到了四层堆叠。PCM在随机读写速度和寿命方面相比于NAND Flash都有数量级上的优势，三维集成层数是制约其容量快速发展的主要瓶颈。英特尔已经在着手研发相关三维制造技术。

• 终端推理将是主要趋势，物联网则为主要应用场景

首先，终端涉及任务较为固定，无需频繁变动架构，更适合存内计算的实现。其次，推理过程算法更新频率低，无需频繁修正计算，对存储器耐久性的要求较低。最后，推理过程可适当降低算法精度，减少计算复杂度，从而更容易实现存储和计算的融合。

与终端推理对应的场景是物联网。物联网涉及大量智能终端，其计算过程通常只涉及浅层人工智能。存内计算在浅层的算法实现过程中，简化的存算结构更易于硬件实现。同时，存内计算由于无需频繁搬运数据能耗将显著降低，符合物联网对低耗能的要求。

产业篇



2.1 行业现状与驱动因素

行业现状

• 玩家现状

存算一体技术早在2011年就引起学术界关注。根据国内早期入局者的观察，存算一体在2016-2017年成为学术界热议的话题，到2019年逐渐开始受到工业界和资本的关注，彼时大家的讨论主要集中在该项技术的可靠性上。从2020年开始，越来越多的玩家进入这个市场，并且大公司都开始在内存计算上发力，此时的内存计算已成为产业界“不得不跟进”的技术之一，大家的讨论聚焦在内存计算未来的市场空间上。

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。SST的memBrain神经形态内存产品基于SuperFlash 技术以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，增强了边缘侧AI推理。与传统的基于数字DSP 和SRAM/DRAM的方法相比，可将功耗降低10到20倍。Mythic M1076模拟矩阵处理器在单芯片中的算力高达25TOPS，预计将于2022年下半年出货。

在存算一体领域，美国Syntiant公司声称其芯片出货量已超2000万片，而国内具备量产能力的公司目前的出货量在百万级。出货量上的差距一方面由于国内起步时间较晚，另一方面国外存算公司相对完整的技术链以及更成熟的存算技术使其获得更多来自芯片大厂的合作，这些机会在助力存算一体技术迭代的同时也给存算公司带去更大的订单。

从芯片营收上讲，经统计，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（中国玩家数量远超国外）。

如果将基于存算一体的芯片放在半导体产业的大背景下，其尚处在发展的早期阶段，在这个阶段业界呈现出一种百花齐放的态势。各玩家根据自身的背景和擅长领域，选择不同的技术路线，在各自的应用场景中进行商业化探索。

• 技术进展

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。

• 应用场景布局

主要分为两大场景：面向端侧，对低功耗需求强烈的场景，以及面向云侧推理，对大算力需求强烈的场景。目前存内计算的计算效率提升已经得到业内的充分认可，业内80%的存算公司优先布局对能效比有高要求的端侧场景。

在端侧小算力场景中，目前实现量产的有三家，知存科技、九天睿芯和闪易半导体。

存算一体目前在大算力落地上并未做到优势明显，针对存算一体芯片的算力提升是业内共同的发展方向。

在做大算力存算一体芯片的公司中，千芯科技选择将存算一体与可重构计算融合，兼顾能效、精度和灵活性，以支持AI芯片的更大算力和可编程灵活性。后摩智能选择基于SRAM做大算力存算一体芯片，AI算力达到数十TOPS甚至更高，可支持大规模视觉计算模型的AI芯片。目前，后摩智能自主研发的存算一体大算力芯片已成功点亮，并跑通智能驾驶算法模型。亿铸科技选择研发基于ReRAM的全数字存算一体大算力AI芯片，面向云计算等中心侧和自动驾驶等边缘侧场景。

• 战略布局

初创公司的思路普遍是前期聚焦特定场景，在一个垂直领域站稳脚跟后，逐步进行技术外溢，扩展到更丰富的应用场景。目前实现落地的场景集中在小算力低功耗场景中，在新型存储器技术发展和形成完整工具链的基础上，将会走向更大算力的场景。

我们预计小算力场景将在2-3年后迎来一个爆发点，存算一体的头部公司在这个时间点基本会实现初步量产，届时这项技术将在智能可穿戴设备领域占据一定市场。云端大算力场景由于技术更加复杂以及更强调通用性，预计到2030年可实现初步量产，自动驾驶将成为存算一体大算力芯片率先涉足的场景。

驱动因素

存算一体已经完成了从学术界到工业界的转化，基于存算一体的产品在国内外都有具备量产能力的公司。下一阶段，基于存算一体的产品将聚焦如何实现从初步商业化到大规模商业化的跃迁。在这个过程中，分析师认为主要有三点驱动因素：新型存储器的发展，来自应用侧的需求以及产业侧的配合。

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还具备成本优势。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。业内观点认为，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非常大。

2. 应用侧的需求

后摩智能认为存算一体的发展逻辑是由外向内，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。

3. 产业侧的需求

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。根据受访者的反馈，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。

技术产业化的速度与技术对应的需求增长速度成正比，而需求增长取决于技术在企业降本增效上发挥的作用，以及技术提升带给用户的价值。因此，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。在技术路线正确且能够实现闭环的前提下，选择大场景能够带来更大的市场空间，从而触动产业界和资本持续投入，逐步形成良好的生态。

共同挑战

当前存算一体公司面临的共同挑战是如何突破技术及应用层面的壁垒，将客户的迁移成本降到最低，从而使客户在传统方案和存算一体方案中选择后者。

由于存算一体架构完全颠覆了传统的冯诺依曼架构，客户在选择是否使用基于这项技术的产品时，必定会将新技术下的解决方案与传统技术下的解决方案进行对比，然而存算一体目前在技术链的完整度上并不占优势。换句话说，客户选择存算一体技术会带来额外成本，例如技术人员的学习成本，产品适配过程中产生的成本等都会成为客户选择时的考虑因素。当迁移成本足够低且融合后产品在功耗或计算效率上有明显优势时，基于存算一体技术的产品或IP才会被市场广泛接受。

2.2 市场价值

目前主流芯片在性能提升上主要依靠先进的工艺制程，不论是芯片设计厂商还是制造厂商都在围绕先进工艺进行研发投入。然而在后摩尔时代，工艺节点的迭代需要付出的成本越来越大，迭代后芯片所性能提升的空间却在逐渐下降。具体来讲，先进工艺的发展面临以下三点挑战：

1. 先进工艺下投入成本与收益不匹配

根据IC Insights调研机构估算，如果想追赶上台积电在先进制程上的制造能力，起码需要五年时间外加近万亿元的投入。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。

2. 先进工艺下由物理限制引发的芯片性能问题突显

先进制程下的芯片存在漏电流的问题，导致产品发热问题严重。短沟道效应是先进工艺下芯片功耗高的主要原因：沟道的长度随着集成电路尺寸缩小而缩小，导致沟道管中的S和D（源和漏）之间距离越来越短。管道变短使得栅极对沟道的控制能力变差，表现为栅极电压夹断沟道的难度变大，因而出现严重的电流泄露现象，令芯片的发热和耗电失控。

此外，当制程走到1nm时，电子会产生“量子隧穿效应”而穿透绝缘层，导致晶体管出现漏电问题。

3. 先进工艺并非在所有应用场景中都有优势

随着人工智能持续发展，计算效率与能耗之间的权衡愈发重要。先进工艺下尽管芯片拥有大算力，但同时也产生了高能耗。当半导体工艺到达7nm，数据搬运功耗达到35pJ/bit，占比达63.7%。然而数据搬运并不产生额外价值，消耗越大，就会造成越多的浪费。在AIoT领域，对于算力的要求没有那么高，相反对于功耗却有着严格的限制，因此先进制程芯片在这个领域并不占优势。此外，先进制程芯片在可靠性上不及成熟制程芯片，使其在工业和军事领域同样缺乏优势。

先进工艺的局限性促使业界开辟新的研发思路，考虑到先进工艺距离物理极限越来越近，工业界在芯片突破上的思路逐渐转向更底层的架构和电路设计。

存算一体的优势

传统芯片在性能提升上的瓶颈主要是冯诺依曼架构下存储单元和计算单元分离，数据在传输过程中产生延迟且频繁的数据移动带来了过高的功耗（在此架构下无法突破）。从存储器提取数据，搬运时间是运算时间的百倍甚至千倍，整个过程的无用功耗占比高达60%-90%。随着AI计算的不断发展，越来越庞大的数据量以及人工智能对实时性的高要求使得人们不得不转变底层逻辑，通过改变芯片架构来满足应用需求。

在新型架构中，存算一体架构在功耗和计算效率上的天然优势使其成为近年来备受关注的发展方向。分析师将存算一体架构的优势归纳为技术的先进性和技术落地后的成本优势。

技术先进性

存算一体通过拉近或消除计算单元与存储单元之间的距离，从而增大它们之间的带宽。同时，存算一体也消除了数据搬运过程中产生的延迟和功耗。采用存算一体技术，算力能效可以提升数十倍到百倍，能耗降低至1/10-1/100。

存算一体技术把存储器中的存储单元变成运算单元，存储器容量越大，运算单元就越多（可做的运算就越多）。在存算一体下，衡量算力大小取决于存储器容量大小。举例来说，如果存储器容纳100-200万个存储单元，就能够同时完成100-200万个乘法和加法运算。同时，存算一体省去了把百万个数读出来，再放到计算单元中做百万次运算的过程，能够以很高的并行度去实现乘加运算。

目前市场上的GPU芯片里的计算单元在百万级别，而一个存储器当中可容纳几百亿到几千亿的存储单元，一旦存算一体技术能够将大比例的存储单元变成计算单元，完后大规模运算将不再局限于先进制程。在成熟制程下，采用存算一体技术也可以满足大算力需求。据业内人士透露，目前存算一体掌握了千万级的并行运算，未来有望实现几千亿的并行计算。

成本优势

1) 芯片成本优势：一个芯片上的晶体管数量增加，良率就会降低。如果一个芯片的良率达不到80%~90%，这个芯片的利润会非常低，甚至无法做到盈利。此外，芯片的面积效率决定了其价格优势（芯片的面积决定了成本的一大部分），在同样的面积下可以放的晶体管数量是固定的（工艺限制决定），存算一体技术可以做到用同样数量的晶体管完成更多的算力。换句话说，完成同样的算力，用存算一体技术所需的芯片面积（晶体管数量）更小，因而成本也更低。

2) 传输功耗几乎为零：存算一体架构省去了数据在存储单元和计算单元之间相互传输所产生的功耗，从而极大地降低访问存储器的成本。

3) 对比传统架构的优势：在传统的冯诺依曼架构下，7nm先进工艺的研发成本为3亿美元，到5nm的研发成本会再增加50%，但性能提升只有10%-20%；而在生产成本上，5nm芯片的生产成本也要比7nm高出50%。当高昂的研发和生产费用没有足够的利润支撑后，厂商便不会采用先进工艺生产芯片。在计算性能上，存算一体技术可以在成熟工艺下，通过技术优势达到与先进工艺同等的性能。

在不同场景中的优势（按不同应用场景需要的算力和容量区分）

存内计算在极低功耗和极大算力场景中相较于传统计算模式都能够形成绝对优势。据业内人士描述，在16TOPS节点之上的大算力场景，存内计算兼具算力高和功耗低的优势，能够以更低的成本实现更大的算力。在低功耗场景中，基于存算一体的芯片可以完成极低功耗的深度学习运算。

存算一体技术演进过程中，如果能够在扩大计算规模的同时满足精度和可靠性要求，将会在云端爆发出更大的潜力，应用场景可延伸至自动驾驶、大数据检索、蛋白质/基因分析、数据加密、图像处理等。

我们将应用场景按照算力大小进行划分，包含边缘/端侧小算力场景和云端推理大算力场景。在两大分类下，分别对应了五类细分场景，边缘/端侧小算力包括智能可穿戴设备，智能安防，移动终端，AR/VR；大算力场景目前主要是自动驾驶。

	定义	特征	优势	代表产品
智能可穿戴设备 2MB-100GOPS	应用穿戴式技术对人们的穿戴进行智能化配置，将各种传感、识别、连接和云服务等，植入到人们的眼镜、手表、手环等日常穿戴中。	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。	芯片技术是智能可穿戴设备发展的核心，芯片的技术成熟度会影响可穿戴设备的性能；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。	<ul style="list-style-type: none">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	基于智能视觉、多维感知、组网协同等技术，打造前段智能体系。	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	具备通信功能的微型计算机设备。	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/VR通过处理相机，激光雷达，ToF和音频传感器协同操作，为游戏等娱乐场景提供最佳体验。	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	无需人类操作即能感测其环境及导航，通过雷达、光学雷达、GPS及电脑视觉等技术感测环境。	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。	<ul style="list-style-type: none">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

在以上场景中，经过对各家公司场景布局的统计，我们对相关应用场景的分析如下：

- 端侧小算力场景已经实现初步量产，下一步将走向大规模量产。小算力场景的特点是碎片化，如果针对每个不同场景去提供不同种类的IP，需要投入的周期长。因此在小算力场景中，通过单芯片的形式提供存算一体算力是更好的选择。
- 在大算力场景中，由于更加强调通用性，想要做独立的芯片难度非常大。对于存算一体公司来讲，以IP授权或定制服务的形式赋能芯片大厂以及终端设备制造商是更好的选择。

目前，国内外存算一体公司在端侧小算力场景中的应用最为成熟。其中，智能语音是早期存算一体公司率先布局的场景，并且已实现产品化。在成功探索智能语音场景后，智能视觉场景是产品/技术迭代的方向（有些公司会在这两个场景同时发力）。

2.3 市场规模

存算一体芯片的价值可以类比为冰山，在大家能够看到的市场之下蕴藏着巨大潜力。现阶段存算一体技术主要应用在端侧小算力场景，相比于备受关注的大算力芯片，基于存算一体技术的芯片目前看似是低功耗小算力场景的替代性产品，实则其技术实力能够同时满足AI芯片对于算力和功耗的需求，并且在芯片性能上可实现指数级提升。当基于存算一体技术的产品得到市场验证后，其市场将逐步扩展到整个AI芯片领域。

短期内，行业将会呈现持续走高的态势，但整体市场规模不会很高。现阶段增长来源主要为定制开发费用和SoC芯片销售费用。国内当前实现量产的公司有限，近期产能可达千万级别预计仅有1-2家。分析师预计从2022年到2025年，还会有3-5家存算一体公司能够实现量产，在此期间存算一体芯片的市场规模相对有限。

到2030年，市场规模将实现高速增长。一方面，2030年，基于存算一体的芯片在低功耗小算力场景可以实现完全替代，且产能也满足这个市场的需求。另一方面，2025年后存算一体公司将逐渐从定制化产品过度到标准化产品研发上，针对端侧、边缘侧和云侧不同场景提供更具通用性的产品。在得到市场验证的基础上，基于存算一体打造标准化产品需要5年左右的时间，因此在2030年市场规模将大幅增长。

以下是我们的量化思路：

- **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**

根据受访者的描述，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四家左右，其余公司中进入测试阶段的有2-3家。分析师预计，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，分析师预计在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。

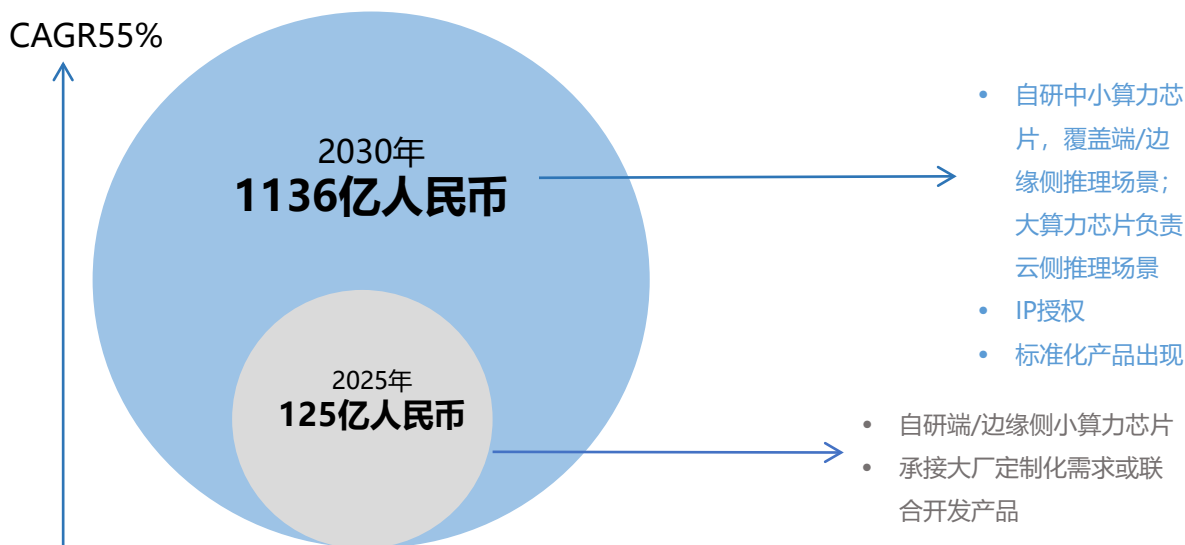


图3 我国存算一体芯片市场规模估算

2.4 产业链分布

• 目前尚未形成完成的产业链生态，尤其在软件层面缺乏相应的研发公司配合完善技术链条

存算一体技术在产业链的各环节都需要相应公司在现有技术能力上进行改变，但存算一体技术目前的应用规模尚不能支撑产业链中其他技术环节上的公司专门为其研发新的配套工具/软件技术，因此想要让存算一体真正发展起来，上游的存算一体公司往往需要具备全栈自研的能力。根据调查，目前国内存算一体头部公司（详见下表）除了无法自己流片，在产业链上游的其他环节，包括工具链开发、算法开发等都采用自研的模式。产业链中游的设计环节，存算一体公司负责芯片设计，芯片制造需要与代工厂合作完成，封测环节也需要与相关企业合作完成。

多家受访公司表示工具链和EDA设计软件是存算一体技术发展的关键，前者是促进产业快速发展的重要因素，后者是实现大规模量产不可或缺的工具。目前，国内外均未出现专门针对存算一体的工具链和EDA软件，各个公司都是在成熟的工具和软件上进行改造；

未来，基于存算一体技术的芯片必将走向大规模量产，而能够被大规模应用是产业生态的基本特征。由此可见，软件与硬件的协同是存算一体形成产业生态的必备条件。

在存储技术上，大多数公司采用购买国外成熟的存储技术IP来完成存内计算的研发。在这种模式下存在一定的风险，一旦在IP授权上发生问题（如停止授权），将会影响到存算一体技术的持续演进，甚至会阻碍技术成熟的速度。

上游

IP授权



EDA工具

- 原理图编辑工具
- 版图编辑工具
- 电路仿真工具
- 物理验证工具
- 寄生参数提取工具和可靠性分析工具等
(目前尚未出现针对存算一体的EDA工具)

流片代工厂



编译工具链

- 编译器：前端、IR、后端
- 量化器
- runtime

算法相关

- 算法压缩
- AI降噪算
- 各类算法开发
- 算法移植

芯片设计



SYNTIANT



芯片制造



芯片封装



中游

智能终端设备

智能手表/环



智能家居



智能耳机



AR/VR



智能安防



自动驾驶



下游

2.5 主要玩家

玩家现状

存算一体目前已是学术领域最热门的话题之一，也是产业界不得不发展的技术方向。随着越来越多新玩家涌入，存算一体也成为近两年芯片投资的热门赛道。

目前已知的商业模式主要分为三种：IP授权，定制/联合开发以及自主SoC芯片；

国内外存算一体初创公司倾向于SoC芯片研发；有存储器背景的存算公司则更倾向于存储技术IP授权。由此可以看出，以电路/架构设计出身的存算一体初创公司最终的竞争核心在SoC芯片性能上，其中包括基于存算一体IP核的设计能力以及SoC芯片设计规划能力。做存储器出身的公司的竞争核心是存储器技术的先进性和成熟度。

国内


从融资轮次来看，国内的头部初创公司（10家左右），知存目前到了B1轮，其余均在A轮前后。由此可见，存算一体处于行业较为早期的阶段，投资机会较多。做存算一体的公司之间尚未出现相互竞争的状况，各家之间对如何运用存算一体技术有各自的判断，并且在选定的技术路线和应用场景中都有自己的独特优势。

九天睿芯专注于神经拟态感存算一体芯片研发，后摩智能采用存算一体技术做大算力AI芯片，千芯科技聚焦大算力可重构存算一体芯片的架构设计，苹芯科技在基于SRAM做存内计算加速器。以上公司现阶段均基于传统的Nor Flash或SRAM做存算一体芯片，目前亿铸科技在开局便选择基于新型存储器RRAM进行存算一体研发，与昕原半导体深度合作，通过Fabless+IDM模式，实现从芯片设计到生产制造100%国产化。未来，随着新型存储器技术的成熟，直接选择新型存储器进行存算研发的公司将会越来越多。

除了在芯片设计上的探索，国内还出现了专门做新型存储器技术（RRAM）研发的公司，一旦在存储器技术上实现完全国产化，存算一体技术的发展将不会受到存储器技术IP授权的限制。

目前在存算一体领域实现量产的公司有九天睿芯、智芯科和闪易半导体，其余头部公司均已完成多次流片，其中至少有2-3家已进入量产前的测试阶段。

大中华地区代表玩家

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 九天睿芯	神经拟态感 存算一体架 构芯片	A轮	<ul style="list-style-type: none"> 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

公司	定位	融资轮次	产品及解决方案	技术亮点	合作
 阿里达摩院	基于DRAM的3D键合堆叠存算一体AI芯片	/	/	<ul style="list-style-type: none"> 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+	<ul style="list-style-type: none"> 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。 	<ul style="list-style-type: none"> 基于数字域内存计算的电路 	/
 知存科技	基于存算一体技术的人工智能芯片	B1	<ul style="list-style-type: none"> 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
 莘芯科技	存算一体芯片与非冯架构智能算力平台	A	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> AI计算卡：先进存算架构，深度优化存储墙与编译墙，提供更强大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。 	<ul style="list-style-type: none"> 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	科创板过会	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	<ul style="list-style-type: none"> 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。 	<ul style="list-style-type: none"> 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	/

注：标注“/”不代表没有，仅是我们未收集到准确信息

公司	定位	融资轮次	产品及解决方案	技术亮点	合作
 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	<ul style="list-style-type: none">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	<ul style="list-style-type: none">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	上市公司	<ul style="list-style-type: none">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。	/

国外

从融资轮次来讲，国外的存算一体公司集中在C轮和D轮，投资方包括SK 海力士、瑞萨电子等传统芯片大厂。



基于Nor Flash做存内计算的公司最多，其中Mythic与Syntiant已实现量产，前者与高通在无人机领域展开合作，后者与瑞萨、Canary Speech等公司在智能语音领域展开合作。

国外具备量产能力的玩家在工具链的完整度和成熟度上更高，且软件具备较高的适配度。

国外代表玩家

公司	定位	融资轮次	产品及解决方案	技术亮点	合作
 MYTHIC Mythic	基于Nor Flash的存算一体芯片	C轮	<ul style="list-style-type: none">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	<ul style="list-style-type: none">数据流架构；工具链的开发，包括优化套件和编译器	高通

公司	定位	融资轮次	产品及解决方案	技术亮点	合作
 Syntiant	使用模拟神经网络计算的深度神经网络处理器	D轮	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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。 	<ul style="list-style-type: none"> 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	<ul style="list-style-type: none"> 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	/	<ul style="list-style-type: none"> 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	上市公司	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	/

公司	定位	融资轮次	产品及解决方案/研究	技术亮点	合作
 SK海力士	基于GDDR接口的 DRAM存内计算	上市公司	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	上市公司	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	/

中外竞争格局对比

对比国内外存算一体技术的发展，可以发现以下特征：



成立时间不同会影响技术路线选择，国内外实现产品化的公司数量不多，离规模化还有一定距离

从成立时间上看，国外的初创公司普遍比国内早5年左右，也因此量产方面更具优势，头部初创公司均已实现量产。国内初创公司中量产水平可以与国外比肩的目前只有知存（前提是其能够达到2022年预计产量的目标）。由此可见，不论国内外，一个初创公司在保证路线正确且不走弯路的前提下，成立5年左右能够实现小规模量产（百万片）。

其次，从技术角度看，成立时间与选择的技术路线之间存在一定关联。举例来说，不论是国外的Mythic（2012年成立），还是国内的知存（2017年成立），都选择了端侧小算力场景，并且在存储器的使用上都选择了闪存技术。在2015年，闪存技术已经非常成熟，数据中心都在使用闪存，在这个时间段成立的初创公司会选择使用已经成熟的存储器技术做存算一体，尽量避免使用不成熟的器件工艺导致过高的研发成本。而到了2020年前后，SRAM技术已趋于成熟且新型存储器件的研发越来越深入，这个时间段成立的公司存储器件上的选择以SRAM为主。甚至，亿铸科技直接选择基于新型存储器件RRAM做全数字化存内计算研发。在看到存算一体技术在云端推理的优势后，部分公司（如后摩）会直接选择大算力场景发展自己的技术。

此外，技术的产品化进程也与成立时间有关。选择技术路线之后，更重要的是如何落地，以及进一步实现大规模量产。国外公司由于起步时间早，在技术上经历过1-2代的迭代后，基本上都可以形成自己的技术优势，并且在优势领域深耕的同时去开拓更多的场景。国外头部的初创公司已经具备全栈研发能力，从底层架构，到硬件，再到工具链以及软件层面，能够为客户提供完整的技术链条和配套工具，尽可能降低客户的迁移成本。

值得注意的是，国外头部存算一体公司已经与芯片大厂展开多次合作，说明存算一体技术在国外已经得到认可，并且传统芯片大厂愿意拿出自己的资源配合存算一体技术进一步发展。

近年来半导体产业缺芯的状况使得一些公司转向更容易采购的成熟工艺节点，这将进一步推动采用成熟制程的存算一体技术发展。在成熟制程的基础上，运用模拟计算又能够以更低的功耗获得更高的性能，进一步突显存算一体公司的优势。



技术路线：大公司选择最容易落地的，初创公司在确保技术先进性基础上选择最容易落地的

在不同的技术路线上，现阶段选择近存计算的主要以大企业为主，包括传统的芯片大厂如英特尔，IBM等；拥有丰富生态的综合型企业，如三星，阿里巴巴等。

大企业选择近存计算一种情况是自身应用的需要，如阿里巴巴拥有丰富的应用场景，这些场景继续往前走将可能同时面临算力和功耗上的瓶颈，需要一种能够快速落地的技术去解决现有问题。另一种情况是芯片大厂自己的客户出现对高效算力和低功耗的需求，使其需要开发出符合客户需求的技术。

以上两种情况都是以解决现有问题为出发点，落地速度和实用性是首要前提。因此，近存计算作为最接近工程落地的技术，成为大企业首先的研发方向。同时，大型企业也通过在内部进行学术研究，对存内计算展开积极探索。

选择以存内计算落地的公司则以初创公司为主，通过将创新技术应用在特定场景中，逐步实现工程化落地。目前做存内计算的初创公司在大算力场景均以合作形式赋能应用，通过定制或联合开发的形式进行技术落地。此外，也有公司通过IP授权的方式与SoC大厂合作。对于小算力低功耗的场景，存算一体公司更多会选择做SoC芯片来满足特定场景的需求。



国外已形成完整的自研技术链，大规模量产上国内外均未实现突破

在技术的完整性上，国外的初创公司已经实现从技术到产品化的完整闭环，而国内的初创公司在某些技术上还处在研发阶段，目前尚未有一家可以提供完整技术链条的公司出现。国外公司在编译器等配套工具上已有成熟产品，并且可以投入应用，而国内多数公司在编译器上仍处于研发阶段。

在芯片性能层面，相比于传统的处理器，国内外存算公司的芯片都能够达到数十倍到百倍的能效优势。

国内外面临的共同难题是大规模量产，虽然国外公司能够做到初步量产，但依然没有能够实现大规模量产的存算一体公司出现。



不同的业务场景均已呈现出各自的优势，在商业模式上国内外都处在探索阶段

从应用场景来看，国外公司已经能够在应用场景中发挥各自的优势。不论是国内外公司，边端和云端推理是大家一致的方向。自动驾驶是目前云端推理中最受欢迎的场景，做大算力存算一体的公司都会在这个场景中布局。边端的优势之一是离线功能，将原来需要上传到云端处理的数据在本地完成。边端场景对于极低功耗的硬性需求使其成为多数做存算一体公司最先发力的场景。

在商业落地层面，国外的头部初创公司均选择赋能芯片大厂，所覆盖的应用场景从汽车，到银行卡IC，到语音搜索，安防，AR/VR等都有所布局。目前与大厂的合作均在使用成熟制程的芯片产品上，合作案例的共性是能够开发出大厂可以直接用的工具链，从而最大程度打破技术融合过程中的壁垒。

由于SoC芯片的研发成本高且周期长，短期内存算一体技术想要快速落地，多会以AI加速器的形式嵌入完整的芯片内。但长期来看，存算一体赛道的竞争会逐步过度到SoC芯片设计能力上，原因在于合作方只会针对基于存算一体技术的SoC芯片进行考量，而不是只关注存算一体技术本身。

除了基于存算一体技术的SoC产品开发外，IP授权也将是存算一体领域非常有前景的商业模式。目前来看，存算一体技术在各环节都具有较高的研发门槛，且突破难度大。但只要学术界和产业界能够持续投入，在突破核心技术后，围绕存算一体技术的IP授权将成为存算公司的一大营收来源。



虽然业内尚未形成完整的生态，产业链部分环节已经出现针对存算一体进行技术研发的公司

从生态的角度看，目前国内外均处于生态探索阶段，在与大厂的合作中摸索适合自己的方式。对比国外成熟的半导体市场，国内半导体领域中萌发的新兴技术更有可能得到各方的支持与配合，在半导体国产化的目标下，国内的存算一体公司在生态建设上的外部环境优于国外。

目前，已有四家公司实现初步量产且多家公司进入测试阶段，产业界对于存算一体的信心在不断增强。现阶段，业内已出现专门针对存算一体做编译器开发的公司，但尚未建立统一标准。

未来，这些标准的制定可能出自现有的EDA大厂或编译器公司，通过在公司内部形成专门针对存算的研发部门或通过存算公司项目合作的方式，制定一套标准化方案，为大规模量产奠定基础。

2.6 进入门槛

存算一体技术考验的是计算系统和存储系统的整合能力，设计要求比标准模IP和存储器IP更复杂。从技术到产品化的过程中，由于缺乏EDA工具以及适配的工具链，需要通过多次存储器流片积累经验，对相关公司创始团队在存储器量产经验和路线认知上都有非常高的要求。分析师认为，在产业界和资本一致看好存算一体的现状下，完整的技术链条、对客户需求的把握以及全面的人才储备是初创公司在业内保持竞争力的关键，也是新玩家进入这个赛道需要具备的实力。

• 完整的技术链条

存算一体最大的挑战在于技术的未知性带来的高风险。想要真正将存算一体技术的价值发挥出来，需要在各个层次上都做出改变。从最底层的器件，到电路设计，架构设计，工具链，再到软件层的研发，在做相应改变的同时还要考虑各层级之间的适配度。



器件层面

如果想做到深度融合，需要在存储器上做修改。不论是在传统存储器还是新型存储器上进行电路设计，首先都需要存储器公司的IP授权。

器件选择是实现存算一体的基础：存储器设计决定芯片的良率，一旦方向错误将可能导致芯片无法量产。



电路层面

有了器件之后，需要用它做存储阵列的电路设计。目前在电路设计上，存内计算没有EDA工具指导，需要靠手动完成。



架构层面

有电路之后，需要做架构层的设计。每一个电路是一个基本的计算模块，整个架构由不同模块组成，存算一体模块的设计决定了芯片的能效比。

模拟电路会受到噪声干扰，芯片受到噪声影响后运转起来会遇到很多问题。这种情况下，需要架构师了解模拟存内计算的工艺特点，针对这些特点去设计架构，同时也要考虑到架构与软件开发的适配度。



软件层面

架构设计完成后，需要开发相应的工具链。工具链主要包括算法训练指导，编译器和芯片移植三部分，通过指导算法进行训练，把算法改造成适合存算一体应用的；编译器针对调整过的算法写相应的编译功能，最终将算法移植到芯片中。

编译器是帮助程序员理解硬件的重要工具，也是推动存算一体产品落地的重要环节。由于存算一体的原始模型与传统架构下的模型不同，编译器要适配完全不同的存算一体架构，确保所有计算单元能够映射到硬件上，并且顺利运行。

除了完整的技术链，存算一体是否能够持续演进还要考虑到外部因素。

厂商拥有的核心技术（如存储器工艺）将成为存算一体公司的护城河，如果厂商在关键技术上选择购买IP，则需要经过卖方授权。举例来说，未经授权的存储器只能使用其读取功能；授权后，厂商才能使用存储器阵列改进设计；

在外部环境不稳定的现状下，如果厂商未能掌握某些核心技术，一旦卖方停止授权，存算一体技术的发展也将因为这一技术环节的缺失而放缓甚至暂停。

• 客户需求是第一推动力

存算一体本身不能作为一个独立的产品，在工程落地上需要各家在所选的应用场景下做出具有通用性的，高性能NPU，再基于NPU做出完整的SoC。最终交付产品时，客户考量的并不仅仅是存算一体技术，而是整体SoC的能效比、面效比和易用性等指标是否有足够的提升。

分析师认为，新玩家在选择是否进入这个赛道时，首先要明确目标市场，在此基础上要厘清技术与需求的匹配度，真正理解客户的痛点以及针对这个痛点，判断存算一体是否有足够的优势吸引客户。从客户的角度出发，做选择时主要会考虑以下两点要素：迁移成本和性能指标。

1) 迁移成本是否在可承受范围内

如果选择新的芯片提升算法表现力需要重新学习一套编程体系，在模型迁移上所花的人工成本高出购买一个新GPU的成本，那么客户大概率不会选择使用新的芯片。基于存算一体的架构如果要实现大规模应用，需要尽可能降低客户的学习成本，让客户在熟悉的环境中进行应用开发。

此外，客户产品所覆盖的场景通常不会只涉及AI算法，同时会包含其他算法。如果只有AI算法部分编译得很好，其余部分无法完成迁移，客户只能选择放弃这个方案。因此，存算一体在落地过程中是否能够将迁移成本降到最低是客户在选择产品时的关键因素。

2) 性能提升是否达到客户的期望

客户会考虑计算效率提升带来的收益是否大于新技术与产品融合中产生的成本。新技术与产品的融合会产生额外的研发费用（如开发新的编译器）。如果这部分费用超出计算效率提升带来的成本优势，客户可能不会在短期内大批量使用基于存算一体技术的芯片。

其次，在低功耗场景中，除了存算一体解决方案，还有其他基于传统架构的解决方案。如果其他方案能够解决客户现有问题且成本可控，客户选择基于传统架构方案的概率将会远大于选择存算一体方案。因此，基于存算一体的解决方案需要在解决客户痛点的基础上，在能效比或功耗上呈现出绝对优势，将会极大地提高客户的使用意愿。

• 全面的人才储备

根据受访者的经验，能够做成存算一体的公司在人员储备上需要有以下三点特征：

1) 领导层要有清晰的目标来牵引团队，在存储器和计算模式的选择上要有清晰的思路，并且能够准确快速地带领团队往前走。

存算一体作为一项颠覆式技术，没有前人探路，试错成本极高。能够实现商业化的企业，创始人往往具备丰富的产业界经验和学术背景，能够带领团队快速完成产品迭代。

2) 团队需要具有深厚技术背景的人员，对技术方向有精准把握，尤其在新型存储器技术上的探索。

基于存算一体技术的芯片在技术的各个层级上都需要做出改变，因此相关技术人员也需要对技术链各环节有强感知力，能够辨别哪些技术可以率先实现工业化；

综上，对于初创公司来讲，有深厚技术背景的人员储备能够减少企业不必要的成本付出，保证在公司发展前期以最小的代价完成技术到产品化的蜕变。

3) 在核心团队中，需要在技术的各个层级中配备经验丰富的人才。

多家公司在采访中表明，架构师是团队的核心，架构师需要对底层硬件，软件工具有深厚的理解和认知，能够把构想中的存算架构通过技术实现出来，最终达成产品落地；

此外，国内缺乏电路设计的高端人才，尤其在混合电路领域。存内计算涉及大量的模拟电路设计，与强调团队协作的数字电路设计相比，模拟电路设计需要对于工艺、设计、版图、模型pdk以及封装都极度熟悉的个人设计师，符合要求的通常都是在业内有丰富经验的人才。

在此基础上，选择混合电路设计的存算一体公司，需要技术人员对数字电路和模拟电路都有很深的理解和经验，因此模数混合电路设计的相关人才最为稀缺。

展望篇

- 存内计算的应用能够以更低成本，更高的能效以及更低的功耗推动人工智能产业发展。在算力的发展上，存算一体架构能够在成本可控的范围内持续扩大算力，有望成为未来人工智能时代的基石之一。
- 一个计算类的芯片必须有好的生态基础，成熟的工具链才能被快速开发并真正应用起来。未来的存内计算，生态能不能建立好是其能否大规模落地最重要的一点。
- 存算一体技术参照人脑的运行模式，将存储和计算融合，非常适合应用在类脑计算领域。由于类脑计算大算力高能效的特点与存算一体所具备的技术能力高度适配，基于存算一体技术的产品将会成为未来类脑计算落地的关键。
- 存内计算能够为基因工程的发展提供更大的空间。随着基因工程的发展，生物信息数据指数级上升，存内计算能够大幅降低数据移动的功耗，适用于大规模生物数据处理。
- 存算一体技术在国内有更大的发展空间。中国市场在从非智能硬件向智能硬件转变的过程中，对端侧智能产品的兴趣远大于国外市场，因此在需求侧会出现更多元的机会。

这些差异化的需求是新型技术公司发展和成熟的好机会，在国内丰富的智能硬件场景中，具有技术优势的公司，在选对应用的前提下，能够发展的市场空间远大于国外同类公司。

- 在新型存储器研发上将呈现国产化趋势，未来基于存算一体技术的芯片在存储器的选择上有望做到自主可控。
- 存算一体技术的应用场景并不局限于AI计算，在实现更复杂的运算后，有望扩展到更丰富的应用场景。