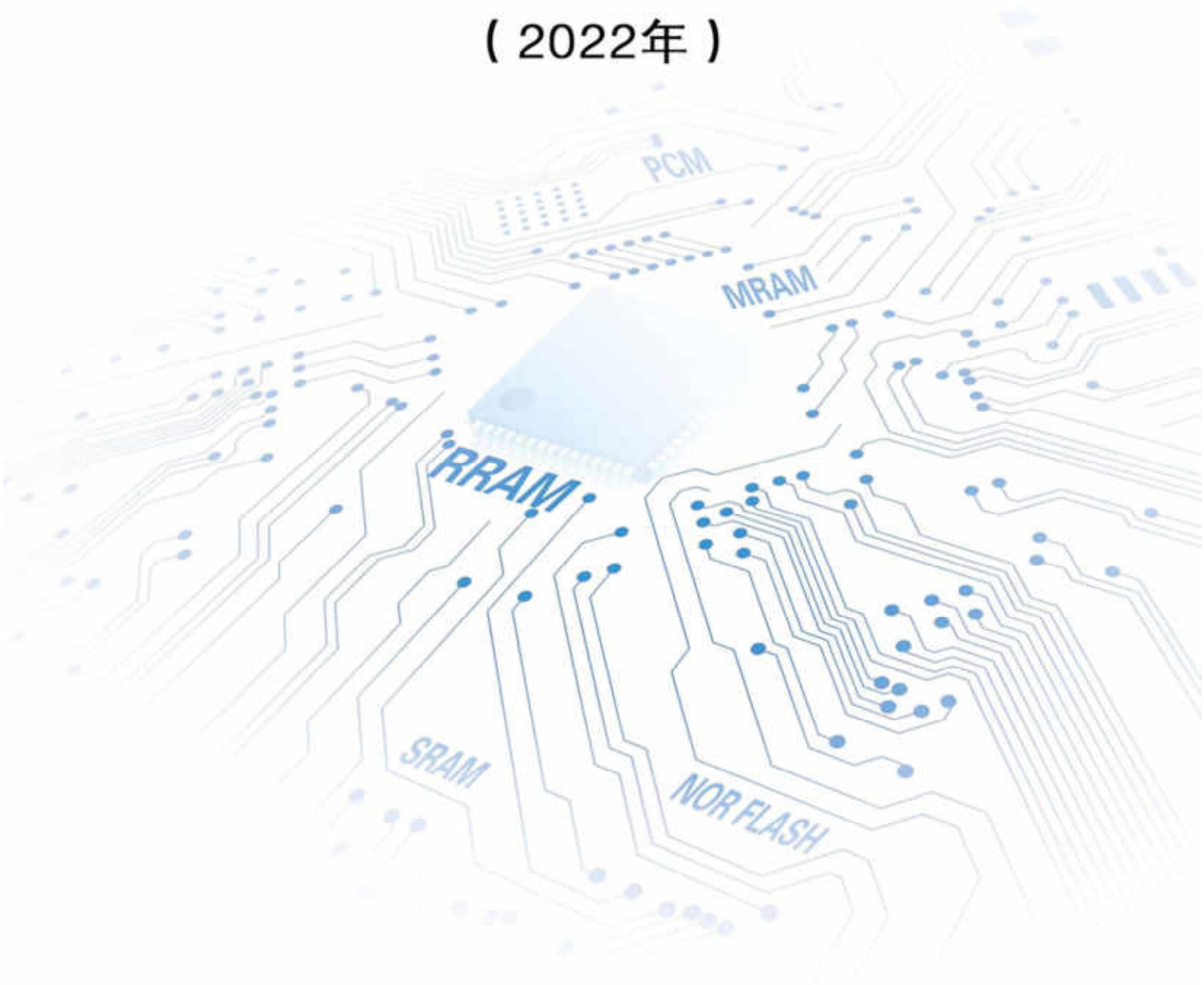
中国移动 研究院
China Mobile CMRI

存算一体白皮书

（2022年）



中国移动通信有限公司研究院

前言

当前数字经济已成为继农业经济、工业经济之后的主要经济形态。算力作为数字经济的核心生产力，将直接影响数字经济发展的速度，决定社会智能的发展高度。中国移动充分发挥运营商网络领先优势，积极承接国家“新基建”和“东数西算”战略，提出“算力网络”全新发展理念，旨在构建一个算力和网络深度融合的新型信息基础设施，助力数字经济、智慧社会高质量发展。

存算一体作为一种新型算力，有望解决传统冯·诺依曼架构下的“存储墙”、“功耗墙”问题，是中国移动极为关注的算力学科的突破性技术，已被确定为算力网络十大关键技术之一 。存算一体将存储与计算有机融合，以其巨大的能效比提升潜力，有望成为数字经济时代的先进生产力。

本白皮书全面阐释了存算一体的核心技术、发展路线、应用场景和产业链生态。希望产学研各界能凝聚共识、加强合作、协同发展，推动存算一体技术成熟和生态繁荣，加快存算一体产业化进程，助力我国在先进计算领域实现高水平自立自强。

编写说明

牵头编写单位：

中国移动通信有限公司研究院

联合编写单位：

中兴通讯股份有限公司　　华为技术有限公司

清华大学　　北京大学　　北京知存科技有限公司

曙光信息产业股份有限公司　　深圳亘存科技有限责任公司

目录

1. 存算一体是先进算力的代表性技术

回顾60多年计算行业的发展史，芯片的算力提升一直按照摩尔定律的节奏推进，但主流的计算范式始终遵循冯·诺依曼架构设计。进入二十一世纪，信息爆炸式增长，大规模数据处理成为挑战，算力的需求呈现指数级提升，业界需要从各种维度探索芯片算力提升的方法。

1965年，戈登·摩尔归纳了晶体管的发展规律-摩尔定律，成为了丈量半导体行业技术进步、产品迭代和企业发展的标尺。然而，随着半导体工艺逼近物理极限，摩尔定律的节奏明显放缓，集成电路的发展进入后摩尔时代，业界主要从三大方向探索算力提升的技术路径：“More Moore（深度摩尔）”、“More than Moore（超越摩尔）”、“Beyond CMOS（新器件）”，其中深度摩尔沿着摩尔定律的道路继续推进，通过新型技术持续微缩晶体管提升计算密度；超越摩尔则是发展摩尔定律演进过程中未开发的部分，如先进封装技术扩展计算性能；新器件则是探索除传统硅基路线之外的芯片材料在计算产业的应用。三大方向为半导体行业延续高速发展的节奏提供了可能。

除了上述维度，业界也在通过变革当前计算架构来实现算力的突破。目前，主流芯片如CPU、GPU（Graphics Processing Unit）、DPU（Data Processing Unit）均按照冯·诺依曼架构设计。冯氏架构以计算为中心（如图1-1），计算和存储分离，二者配合完成数据的存取与运算。然而，由于处理器的设计以提升计算速度为主，存储则更注重容量提升和成本优化，“存”“算”之间性能失配（如图1-2），从而导致了访存带宽低、时延长、功耗高等问题，即通常所说的“存储墙”和“功耗墙”。访存愈密集，“墙”的问题愈严重，算力提升愈困难。随着以人工智能为代表的访存密集型应用快速崛起，访存时延和功耗开销无法忽视，计算架构的变革显得尤为迫切。

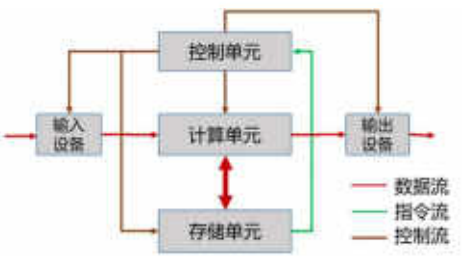


图 1-1 冯·诺依曼计算架构

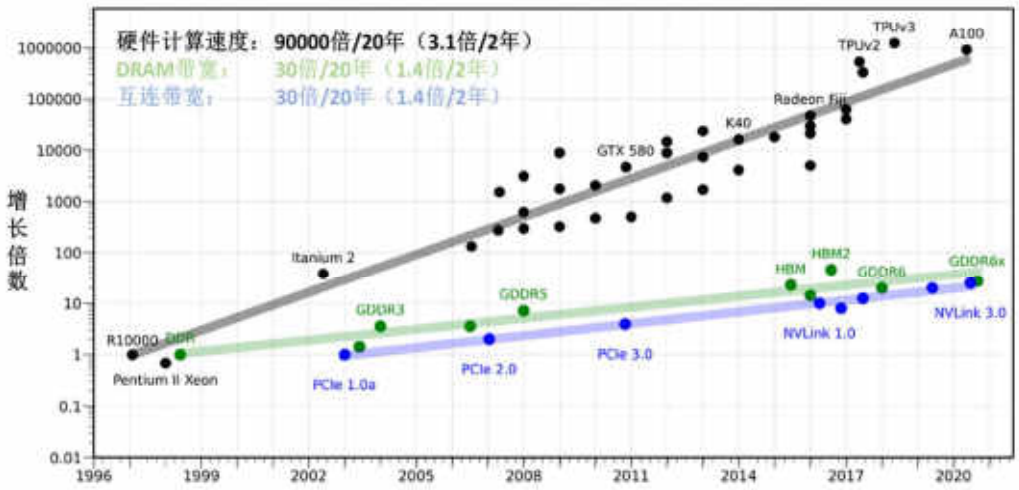


图 1-2 存储计算性能“剪刀差”[1]

存算一体作为一种新的计算架构，被认为是具有潜力的革命性技术，受到国内外的高度关注。**核心是将存储与计算完全融合，有效克服冯•诺依曼架构瓶颈，并结合后摩尔时代先进封装、新型存储器件等技术，实现计算能效的数量级提升。**

2. 存算一体技术路线建议

由于“墙”的问题存在已久，业界已形成多种解决思路，包括对计算或存储部件本身的性能提升，以及存与算的协同优化，即存算一体技术。目前学术界和工业界均在开展存算一体技术研究，学术界主要关注狭义的存算一体，即利用存储介质进行计算；工业界关注商用化进程，着重宣传广义存算一体概念，但分类方法尚未完全统一。本章节将对广义存算一体技术进行分类，望达成广泛共识。

2.1 存算一体技术分类建议

根据存储与计算的距离远近，我们将广义存算一体的技术方案分为三大类，分别是近存计算（Processing Near Memory, PNM）、存内处理（Processing In Memory, PIM）和存内计算（Computing in Memory, CIM）。存内计算即狭义的存算一体。

2.1.1 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：

（一）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，如图2-1。

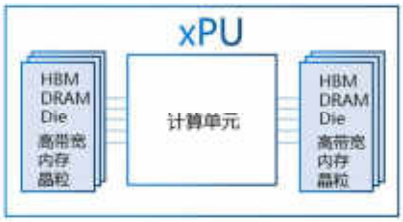


图 2-1 高带宽内存方案

（二）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，如图2-2。

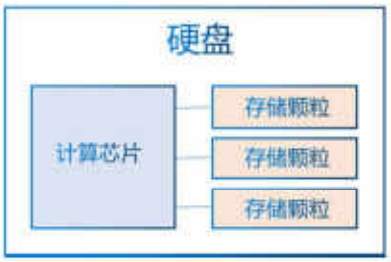


图 2-2 可计算存储方案

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。

2.1.2 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[2]和PIM-DIMM[3]，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，如图2-3。

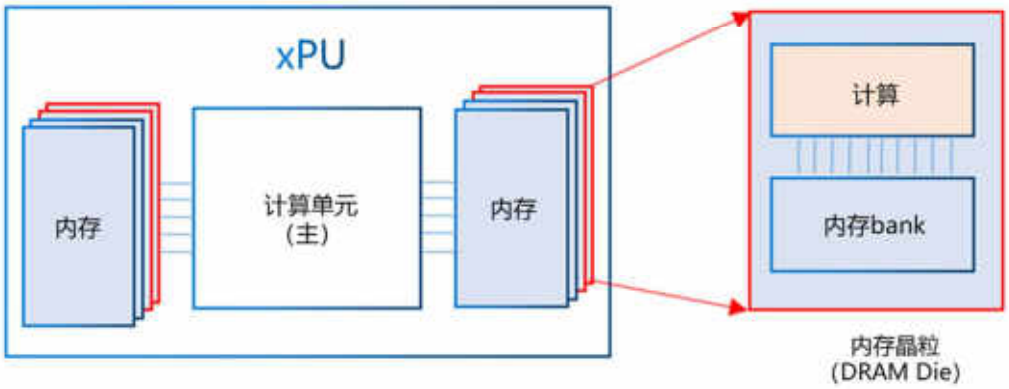


图 2-3 基于DRAM的PIM方案示例

2.1.3 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，如图2-4。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。

图2-4 CIM存内计算

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。存内计算有望激发人工智能领域的下一波浪潮，是广义存算一体技术的攻关重点，本白皮书后续章节将围绕存内计算展开分析。

2.2 存内计算分析

2.2.1 存内计算原理

存内计算主要包含数字和模拟两种实现方式，二者适用于不同应用场景。模拟存内计算能效高，但误差较大，适用于低精度、低功耗计算

场景，如端侧可穿戴设备等。相比之下，数字存内计算误差低，但单位面积功耗较大，适用于高精度、功耗不敏感的计算场景，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也在飞速提升。

模拟存内计算

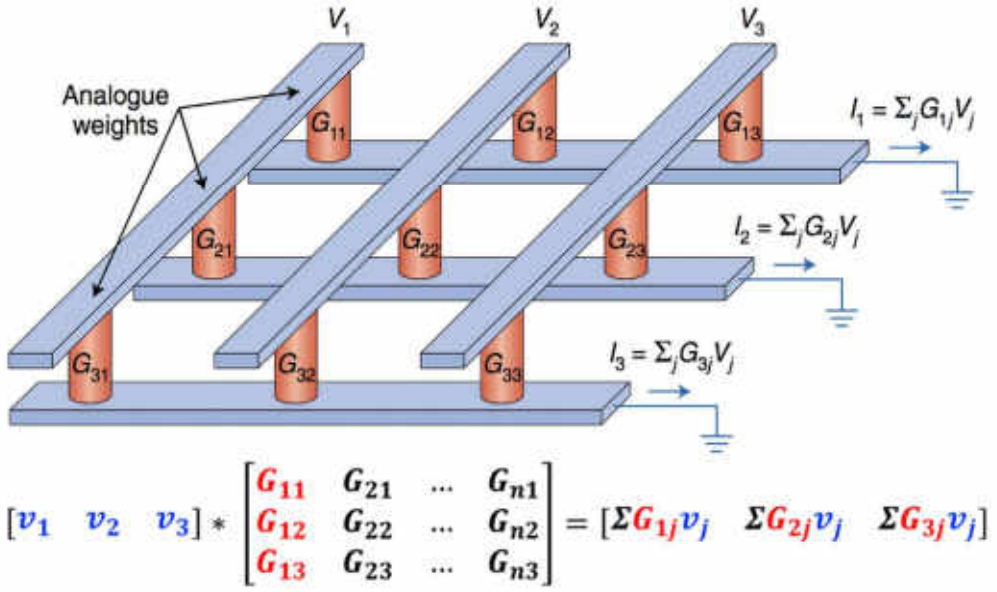


图 2-5 基于RRAM的模拟存内计算阵列

模拟存内计算主要基于物理定律（欧姆定律和基尔霍夫定律），在存算阵列上实现乘加运算[4]。我们以存内计算介质材料之一阻变随机存储器（Resistive Random Access Memory，RRAM，又名忆阻器）为例，来描述存内计算如何实现在数据写入的同时完成计算。

忆阻器电路可以做成阵列结构，与矩阵形状类似，利用其矩阵运算能力，可以广泛应用于AI推理场景中。在AI推理过程中，通过输入矢量与模型的参数矩阵完成乘加运算，便可以得到推理结果。

以矩阵乘加运算为例（如图2-5所示），将模型的输入数据设为矩阵[V]，模型的参数设为矩阵[G]，运算后的输出数据设为矩阵[I]。运算前，先将模型参数矩阵按行列位置存入忆阻器（即[G]），在输入端给定不同电压值来表示输入矢量（即[V]），根据欧姆定律（电流=电压/电阻），便可在输出端得到对应的电流矢量，再根据基尔霍夫定律将电流相加，即得到输出结果（即[I]）。多个存算阵列并行，便可完成多个矩阵乘加计算。

由于整个运算过程无需再从存储器中反复读取大量模型参数，绕开了冯·诺依曼架构的瓶颈，能效比得到显著提升。除忆阻器外，其他存储介质也可通过不同的物理机制满足同样的并行计算需求。

数字存内计算

数字存内计算通过在存储阵列内部加入逻辑计算电路，如与门和加法器等，使数字存内计算阵列具备存储及计算能力[5]。我们以静态随机存储器（Static Random-Access Memory，SRAM）为例，来描述数字存内计算基本原理。

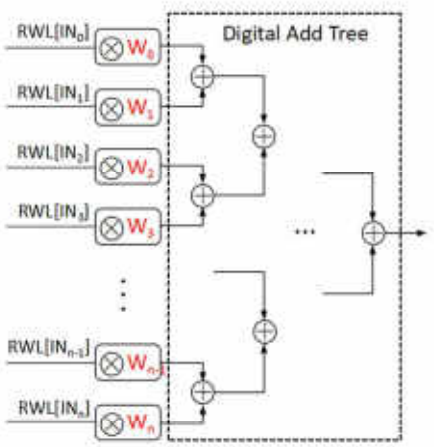


图 2-6 基于SRAM的数字存内计算加法树

如图2-6，输入数据为向量[IN_0,IN_1...IN_n]，存储单元中依次存入模型参数[W_0,W_1...W_n]，通过控制存储器的读字线（Read Word Line, RWL），实现输入数据与存储单元内模型参数的乘法操作，然后通过数字加法树（Digital Add Tree）实现累加，即可完成向量乘加运算。对多个向量重复以上过程，便可实现矩阵乘加计算。

数字存内计算的存储单元只能存储单比特数据，且需增加部分传统逻辑电路，一定程度上限制了面积及能效优势。因此，当前业界多采用可兼容先进工艺的SRAM来实现数字存内计算。

2.2.2 存内计算存储器件分析与建议

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。本章主要对五种主流的存储器件及其存内计算进行描述。

静态随机存储器（SRAM）

SRAM是应用范围最广的易失性存储器之一，常用于CPU中的缓存，基本存储单元由晶体管搭建而成，常见有6晶体管（6T）、8晶体管（8T）结构形式，图2-7为6T SRAM基本单元结构。SRAM通过形成互锁结构的两个反相器来存储数据，在设备供电时可保持存储数据不变，掉电后存储数据丢失，呈现易失性。

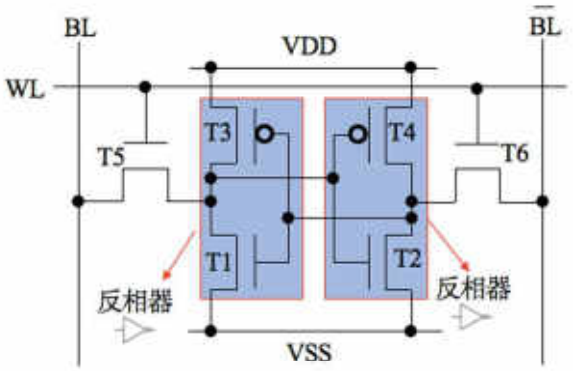


图 2-7 6T SRAM 结构

SRAM读写速度快、无读写次数限制，且其工艺成熟，器件的一致性和稳定性较好，基于SRAM的全数字存内计算可以做到较高的计算精度，并且可以较快地实现技术落地与量产。但SRAM价格相对昂贵、多晶体管单元结构下存储密度较低、需要通电以保持数据，因此芯片面积较大，功耗较高，不适用于对成本和功耗敏感的场景。

NOR FLASH

NOR Flash是一种非易失闪存器件，基本存储单元为浮栅晶体管（如图2-8所示），NOR Flash通过热电子注入/隧穿效应控制浮栅中的电荷数量[6]，每个单元可以存储多比特信息。NOR Flash中浮栅被绝缘层分离以避免电荷泄露，供电消失后浮栅层仍能保持电荷数量不变，存储信息不丢失，呈现非易失性。

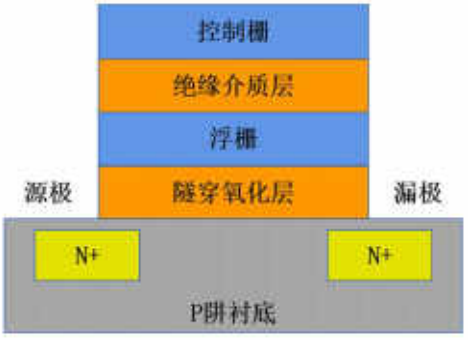


图 2-8 Nor Flash浮栅晶体管

基于NOR Flash的存内计算芯片具有存储密度大、工艺成熟、成本低等优点，业界主要利用其多比特存储特性进行模拟存内计算，相关产品已在智能终端与边缘场景实现小批量商用，带来数十倍的计算能效提升。然而，NOR Flash工艺微缩存在一定挑战，为保证其性能与扩展性，需展开NOR Flash与先进工艺的3D/2.5D集成技术研究。

阻变随机存储器（RRAM）

RRAM又被称为忆阻器，是一种极具潜力的新型非易失存储器件，基本存储单元为金属-绝缘体-金属或者金属-绝缘体-半导体的三明治结构[7]。如图2-9所示，上下为电极层，中间为绝缘的电阻转变层。通过在电极层施加电压/电流，电阻转变层的电阻值可以实现高阻态和低阻态的切换，且电阻转变层可以实现多级电阻状态，使其可存储多比特信息。

图 2-9 RRAM结构示意图

基于RRAM的存内计算芯片具有制备简单、工艺成本低、时延低、支持多比特存储、兼容先进工艺、支持3D堆叠等诸多优点，被普遍认为拥有广阔的发展前景。当前业界主要利用RRAM的模拟多比特特性进行模拟存内计算，可以达到较高的计算能效。然而，RRAM目前在器件一致性和准确性等指标方面还有继续提高的空间。

磁性随机存储器（MRAM）

MRAM是一种基于自旋电子学的新型非易失存储器件，以磁隧道结（Magnetic Tunneling Junction，MTJ）为核心结构，利用隧道磁阻效应实现电阻状态改变，达到存储信息的目的[8]。如图2-10所示，MTJ是自由层-隔离层-固定层三明治结构。固定层的磁场方向保持不变，施加电压可改变自由层的磁场方向，当自由层和固定层磁场方向一致时，器件呈现低阻态，代表逻辑“0”；当自由层和固定层磁场方向相反则为高阻态，代表逻辑“1”。

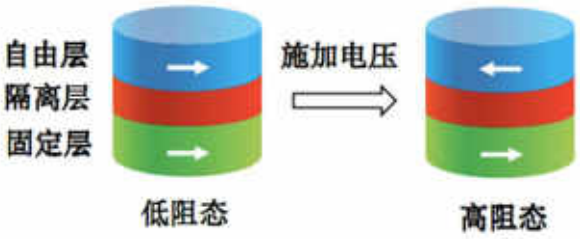


图 2-10 磁隧道结结构

基于MRAM的存内计算芯片具有非易失、访问速度快、读/写次数高等优点，且具备较高的可靠性和稳定性，但MRAM器件成熟度较低，功耗、速度和耐久性等指标离理论预期尚有一定差距。当前业界基于MRAM的存内计算研究较少，需要推动器件成熟，同步挖掘其在存算一体领域的潜在场景。

相变存储器（PCM）

PCM是一种由硫族化合物材料构成的非易失存储器件。如图2-11所示，PCM器件的典型结构为上电极、硫族化合物、电阻加热器、下电极，通过在两电极间施加不同形式的脉冲，对硫族化合物加热使其在晶态和非晶态之间转变，来表征高低阻特性，从而实现数据的存储和控制[9]。PCM的状态可以是介于完全多晶态和完全非晶态之间的多种状态，并以此实现多值存储。PCM断电后状态可保持不变，呈现非易失性。

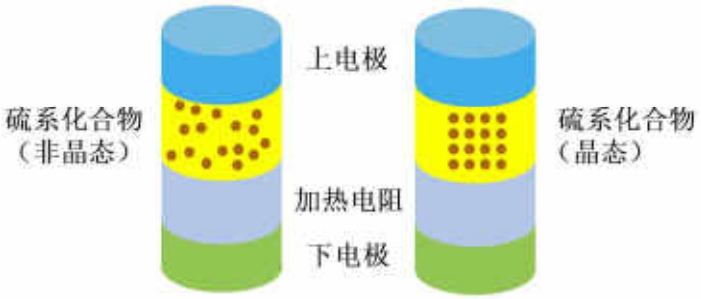


图 2-11 PCM结构及阻态变化原理

PCM有非易失、存储密度高、多比特存储、支持3D堆叠等优点，但PCM存在写入功耗较大、擦写次数较少等问题尚需攻关，因此当前PCM主要还是作为大容量存储器发展，基于PCM的存内计算研究相对较少，待产业进一步发展。

表1 存内计算器件对比分析

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	云侧和边测的推理和训练	云侧、边侧和端侧的推理

结合当前研究情况，表1对五种主流存储器件的主要参数特征以及其存内计算适用场景进行了对比分析。整体来看，五种主流存储器件各有优缺点，产品化选择时需综合考虑器件的成熟度、存储密度、寿命、读写性能、能耗等多方面指标。当前NOR Flash、SRAM等传统器件相对成熟，可率先开展存内计算产品化落地推动。新型器件中RRAM各指标综合表现较好，MRAM寿命和读写性能较好，均有各自独特优势与发展潜力，可持续推动器件成熟，同步进行存内计算探索。PCM新器件成熟度相对较高，当前已可应用于近存计算研究，不过其寿命、能耗指标较RRAM无优势，预计存内计算潜力稍弱，未来可能更多作为存储器辅助存算一体整体技术发展。建议产业未来展开多路径探索，实现各方案优势互补，推动整体产业发展。

[1] F^2/bit，F的是芯片制造工艺的最小特征尺寸（Feature Size），部分存储器可单比特存储多比特数据，此处为等效对比结果.

3. 存内计算在云边端具有广泛的应用场景

存内计算产品基于其不同的器件特性和计算方式，可为云边端应用提供推理、训练等多种AI能力，提升运算效率、降低系统功耗以及设备成本。

3.1 端侧应用场景

据IDC预测，2025年全球物联网设备数将超过400亿台，产生数据量接近80ZB[10]，在智慧城市、智能家居、自动驾驶等诸多场景中，超过一半的数据需要依赖终端本地处理，单设备算力需求约在0.1~64TOPS之间。此外，各类终端设备对运行时间、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，如图3-1。

与传统方案相比，存内计算在功耗、计算效率等方面具有明显优势，在相同制程工艺下，存内计算芯片能在单位面积下提供更高的算力，更低的功耗，进而延长设备工作时间，将在端侧具有广阔应用前景，将广泛应用于家庭网关、工业网关、摄像头、可穿戴设备等场景。

图 3-1 端侧设备各指标需求强度分析

当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，有效降低了端侧成本。

3.2 边侧应用场景

随着云游戏、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。边缘计算场景下对单设备算力需求约在64~256TOPS之间，时延要求高，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。整体来看，边侧设备的未来发展将更加注重时延、功耗、成本和通用性等需求特征，如图3-2。

图 3-2 边侧设备各指标需求强度分析

与传统方案相比，存算一体在深度学习等领域有独特优势，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可以提供综合性能全面兼顾的芯片及板卡，预计将在边侧推理场景中有着广泛的应用，为广泛的边缘AI业务提供服务。

3.3 云侧应用场景

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，如图3-3。

图 3-3 云侧应用各指标需求强度分析

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，存内计算作为智算中心下一代关键AI芯片技术，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。

4. 存内计算五大技术挑战

广义存算一体技术正由学术研究逐步走向商业应用，其中近存计算和存内处理在产品实现阶段面临制造和封装技术门槛高的挑战，在落地阶段需要解决近、远端处理器协同引起的软件重构问题，但整体技术成熟。存内计算技术成熟度较弱，从**器件研发及制造、电路设计、芯片架构、EDA工具链到软件算法生态**等诸多方面均需加强，对产业链各环节提出了更密切的协作需求。

4.1 新器件成熟度低，制造工艺难升级

存内计算在新器件成熟度方面问题突出。采用传统及新型器件是实现存内计算的两种重要方式。其中，NOR Flash、SRAM等传统器件相对成熟，但RRAM、PCM、MRAM等新型器件在器件一致性、擦写次数、功耗、可靠性等方面存在差异化问题，影响存内计算产品在计算精度、寿命、能耗等方面的表现。

针对新器件引入，现有制造产线无法实现无缝切换，且现有工艺水平尚有提升空间。在芯片制造阶段，需要制造商在已有产线流程基础上进行改造，如在掩膜、设备调参等环节进行持续优化。此外，面向新器件的制程微缩无法完全沿用现有晶体管工艺路线经验，新器件工艺兼容先进制程时，难以全面兼顾高可靠性、高精度等要求。

4.2 电路设计影响芯片算效提升

电路设计是存内计算芯片能效优势的核心决定因素，整体技术尚未成熟。电路设计主要分为存算计算核（Macro）以及周边电路两大部分。不同计算核的存算单元、电路连接设计存在不同，诸多前沿研发成果能效水平不一，尚未完成技术沉淀。周边电路提供输入输出衔接处理、计算核处理结果累加计算等能力，帮助芯片实现完整计算能力，该部分需要结合计算核进行适配设计，并保证较低的能耗和面积消耗。此外，模拟存内计算还涉及复杂的模数转换器（ADC）、数模转换器（DAC）、跨阻放大器（TIA）等模块，也为电路面积和能耗带来技术挑战。

4.3 芯片架构场景通用性及规模扩展能力较差

当前少量商用存内计算芯片产品的芯片算力较小，且主要面向端侧特定领域实现，尚无成熟大算力芯片架构，无法为存内计算产品向云边场景推动提供有效支撑。一方面，当前存内计算芯片支持的算子种类有限，难以满足诸多神经网络算法丰富的计算需求，缺乏场景通用性。另一方面，缺乏成熟多核协同机制以及统一的片上互联、片间互联协议及标准，难以实现大算力芯片。

4.4 EDA工具链尚未健全

存内计算芯片设计与常规芯片有较大差异，当前EDA工具辅助设计与仿真验证尚未成熟。具体表现在：

缺乏标准单元库与快速组装工具。不同存储器件的存内计算芯片使用不同的存算单元结构，现有的EDA工具无法全面提供标准单元库

以供芯片设计者使用，只能依赖手工绘制完成。此外，当前存内计算芯片产品化效率低，缺乏自动化工具实现大规模存算阵列的快速组装。

缺乏功能与性能仿真验证工具。当前没有面向存内计算场景进行仿真效率优化的工具，需要花费大量时间对存内计算的功能与性能进行仿真验证，实现大规模存算阵列仿真时难度更高。

缺乏建模与误差评估工具。建模与误差评估的不准确会导致实际计算结果与理想结果产生偏差，如对器件的电路噪声的模拟可以帮助开发者在设计阶段进行方案评估并及时进行调整。当前存内计算研究缺乏相关工具来模拟器件ADC/DAC/TIA相关电路噪声，给芯片设计方案评估和芯片可用性带来挑战。

4.5 软件及算法生态不完善

缺乏通用开发环境和编译器支持。为有效发挥存内计算芯片算力，编译器需要将神经网络模型算子映射到底层存算单元上，当前存内计算编译器多为针对专用领域产品的个性化实现，缺少面向存内计算的通用开发环境和编译器，以便达到向上对接不同算法，向下屏蔽底层存内计算硬件差异的目标。

神经网络算法匹配问题存在挑战。业界存在多种主流神经网络模型量化方案，因模型特性而异，而当前存内计算所支持的量化方案较为单一，需要运用更多的训练样本，更多的迭代次数，更复杂的模型等途径来弥补量化带来的精度损失。此外，存内计算适合高并行处理场景，但部分神经网络应用将矩阵下的乘累加计算变得碎片化，其算法和芯片的计算特性不匹配，会导致硬件利用率低等问题。

5. 存内计算五大发展建议

中国移动结合算力网络业务发展诉求，提出存内计算发展建议，与业界共进，加速产业化进程。

5.1 建议一、协同先进封装技术，实现不同方案相结合

各种存储器件的存内计算方案各有优势，且可与近存计算、存内处理方案结合，如协同2.5D/3D/Chiplet等先进封装技术，将不同工艺、器件的存内计算芯片高度集成，实现优势互补，兼顾成本、能效、性能、精度和通用性等方面指标，如图5-1所示。在此过程中，需推动RRAM、PCM、MRAM等新型器件成熟并向先进工艺兼容，以充分发挥其能耗低、密度大等优势。

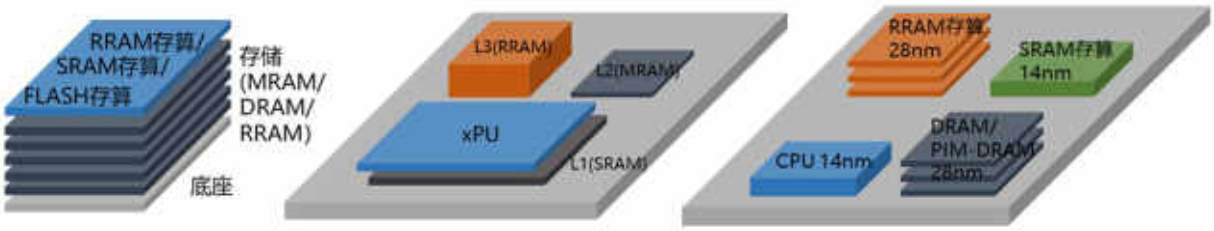


图 5-1 未来先进芯片示例：
3D堆叠（左），多级新型存储（中），异构小芯粒合封（右）

5.2 建议二、优化电路与芯片架构，保障能效优势和演进能力

电路设计和芯片架构对存算一体芯片实现高能效和通用性至关重要。一方面，需加强存算阵列以及周边模块的电路设计能力，保障芯片整体的高并行度、低功耗优势，另一方面，应构建可持续演进的通用存内计算芯片架构，来支持更大规模算力需求、更多算法及应用场景。

5.3 建议三、加速EDA工具孵化，缩短芯片研发周期

存内计算工业化进程需要EDA等产业链上游企业的广泛支持。为保证芯片规模量产，需要芯片设计、EDA、制造商通力协作，打造涵盖单元仿真、可靠性设计、低功耗设计、计算模块设计等诸多环节的配套EDA工具，为存内计算芯片设计和仿真验证提供有力辅助。此外，以存算一体为契机，可反哺推动国产EDA产业发展。

5.4 建议四、构建开发生态与编程框架，加速应用规模发展

为推动存内计算规模应用，相应开发环境与编译平台的建立成为必然诉求，需要业界共同发力，推进开源及标准生态，搭建面向存内计算的编程框架，健全自动化算法开发、仿真和编译工具，构建针对存内计算并行计算特性的算法设计与开发生态。

5.5 建议五、产学研紧密协同，推动端侧到云侧演进

随着存内计算应用范围由边侧向云侧逐步演进，需要推动产学研紧密协同，建立端到端技术栈。存内计算适用于音频、视频、自动驾驶、决策分析等诸多应用场景，目前商用的NOR Flash、SRAM存内计算芯片主要用于中小算力需求的端侧语音和视频场景，未来可进一步实现通用大算力芯片，为云边提供通信、自然语言理解、自动驾驶等场景高效算力服务。因此需要产学研紧密协同工作，构建链式合作平台，拉通器件与芯片研发、工具链构建、软件生态构建、产业发展、方案测试与应用的全链接。

6. 产业发展倡议

针对狭义存算一体发展面临的挑战和问题，中国移动作为算力网络新发展理念的引领者和实践者，希望与合作伙伴通力合作，围绕技术、产业、生态三个方面开展工作，打通存算一体各环节产业链条，推动生态发展，加速产业化进程，真正释放存算一体技术在性能与成本方面的巨大潜力，助力国家实现计算领域的原创科技创新和引领。

共同攻关存算一体核心技术。共同攻关新型材料、芯片架构、编译器等领域的关键技术，共同挖掘存算一体应用场景，支撑国家新型算力基础设施全新发展路径，助力网络强国、数字中国、智慧社会发展战略落地。

共同加快存算一体产业成熟。协同攻关存算一体产业链共性问题，推动产业链上下游、产供销有效衔接，提升产业链韧性，加强新技术对产业渗透的深度和广度，探索存算一体试验示范，协同推动加强产业链创新、健康发展。

共同推动存算一体生态繁荣。通过标准制定、开源推动、产业合作等多种方式和手段，加速推动存算一体技术成熟，加快推动上层应用迁移，共同构建从低层芯片到上层应用的繁荣产业生态。

缩略语列表

缩略语	英文全称	中文解释
AI	Artificial Intelligence	人工智能
CIM	Computing in Memory	存内计算
CNN	Convolutional Neural Network	卷积神经网络
CPU	Central Processing Unit	中央处理器
CSD	Computational Storage Drives	可计算存储
DAC	Digital Analog Converter	数模转换器
DIMM	Dual-Inline-Memory-Modules,	双列直插式存储模块
DPU	Data Processing Unit	数据处理单元
DRAM	Dynamic Random Access Memory	动态随机存取存储器
EDA	Electronic design automation	电子设计自动化
FPGA	Field Programmable Gate Array	现场可编程逻辑门阵列
GPU	Graphics Processing Unit	图形处理器
HBM	High Bandwidth Memory	高带宽内存
MRAM	Magnetoresistance Random Access Memory	磁性随机存储器
MTJ	Magnetic Tunnel Junctions	磁隧道结
NN	Neural Networks	人工智能神经网络
PCM	Phase Change Memory	相变存储器
PIM	Processing in Memory	存内处理
PNM	Processing Near Memory	近存计算
RNN	Recurrent Neural Network	循环神经网络
RWL	Read Word Line	读字线
RRAM	Resistive Random Access Memory	阻变随机存储器
SRAM	Static Random-Access Memory	静态随机存储器
TSV	Through Silicon Via	硅通孔

参考文献

[1] Amir Gholami, “AI and Memory Wall” [Online] . Available: https://medium.com/riselab/ai-and-memory-wall-2cb4265cb0b8

[2] Kim J H, Kang S, Lee S, et al. Aquabolt-XL: Samsung HBM2-PIM with in-memory processing for ML accelerators and beyond[C]//2021 IEEE Hot Chips 33 Symposium (HCS). IEEE, 2021: 1-26

[3] Devaux F. The true processing in memory accelerator[C]//2019 IEEE Hot Chips 31 Symposium (HCS). IEEE Computer Society, 2019: 1-24

[4] 林钰登，基于新型忆阻器的存内计算[J]，微纳电子与智能制造，2019

[5] 尹勋钊，存算一体电路与跨层次协同设计优化：从SRAM到铁电晶体管[J],中国科学，2022

[6] 丁士鹏，基于NOR Flash的存算一体模拟乘加电路设计[J]，信息技术与网络安全，2021

[7] Tanachutiwat S , Ming L , Wei W . FPGA Based on Integration of CMOS and RRAM[M]

[8] Tehrani S , Engel B , Chen E , et al. Recent developments in magnetic tunnel junction MRAM[C]// IEEE International Magnetics Conference. IEEE, 2000

[9] 冒伟，刘景宁等，基于相变存储器的存储技术研究综述[J]，计算机学报， 2015

[10] 中国信息通信技术研究院，中国算力发展指数白皮书[R]，2021