芯跑资本
iRUN VENTURE CAPITAL
立德家国使命感 点燃产业满天星

芯跑资本丨月报资讯

第36期：存算一体，后摩尔时代的新构架

2022.08

存算一体，后摩尔时代的新构架

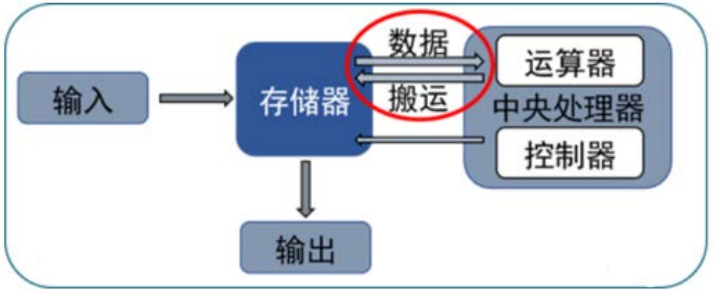
芯跑资本研究部 2022.8

1、“后摩尔时代”，摩尔定律已在物理、功耗、成本三个方面趋近极限

半个多世纪以来，集成电路一直遵循摩尔定律的技术轨迹发展，定律中指出集成电路芯片上所容纳的晶体管数目每隔18~24个月将增加一倍，同时处理器功能和处理速度会翻一番。但在2010年后，晶体管密度增速放缓，逐渐偏离摩尔定律预测的周期。**现阶段，摩尔定律已在物理、功耗、成本三个方面趋近极限，并被业界称为“后摩尔时代”**。

当前最先进的计算机采用基本都是冯·诺依曼架构，这种架构中，数据的处理和存储却是分离的。摩尔定律和冯·诺依曼架构的现状会引发 **存储墙与 功耗墙 两大问题**。

存储墙：冯·诺依曼架构的**存算分离会导致外部存储器运行速度远远小于处理器的运算速度，系统整体会受到传输带宽瓶颈的限制，导致算力会远低于处理器标定的理论算力。**目前，处理器的算力以每两年3.1倍的速度增长，而内存的性能每两年只有1.4倍的提升。后者的性能极大地影响了数据传输的速度，这也被认为是传统计算机的阿克琉斯之踵。



功耗墙：冯·诺依曼架构中，**数据在处理器和外部存储器中频繁高速传递，会导致系统功耗很高。**与此同时，摩尔定律接近瓶颈，芯片特征尺寸已进入量子效应显著的范围，引起一系列次级物理效应，包括栅隧穿泄漏、载流子界面散射、强场速度饱和、源漏寄生电阻占比增大等，导致功耗密度快速上升。在传统架构下，数据从内存单元传输到计算单元需要的功耗是计算本身的约200倍，因此真正用于计算的能耗和时间占比很低。

2、业界在解决“两堵墙”的问题，目前还没有最优解

面对两堵墙的挑战，为了解决“存储墙”问题，当前业内主要有三种方案：**用GDDR 或HBM来解决存储墙问题的冯·诺依曼架构策略；算法和芯片高度绑定在一起的DSA方案；以及存算一体的方案。**

HBM

• HBM就是High Bandwidth Memory的缩写，也就是高带宽内存，这是一项在2013年10月被JEDEC采纳为业界标准的内存技术，JEDEC又分别在2016年和2018把HBM2和HBM2E纳为行业标准。按照AMD的介绍，这种新型的 CPU/GPU 内存芯片（即 “RAM”），就像摩天大厦中的楼层一样可以垂直堆叠。基于这种设计，信息交换的时间将会缩短。这些堆叠的芯片通过称为“中介层 (Interposer)”的超快速互联方式连接至 CPU 或 GPU。将HBM的堆栈插入到中介层中，放置于 CPU 或 GPU 旁边，然后将组装后的模块连接至电路板。通过3D堆叠内存,可以以极小的空间实现高带宽和高容量需求。进一步，通过保持相对较低的数据传输速率，并使内存靠近处理器，总体系统功率得以维持在较低水位。英伟达在Tesla A100和谷歌在二代TPU都选择HBM2E作为内存方案。**因为HBM独特的设计，其复杂性、成本都高于其他方案，且功耗也是问题。**

GDDR

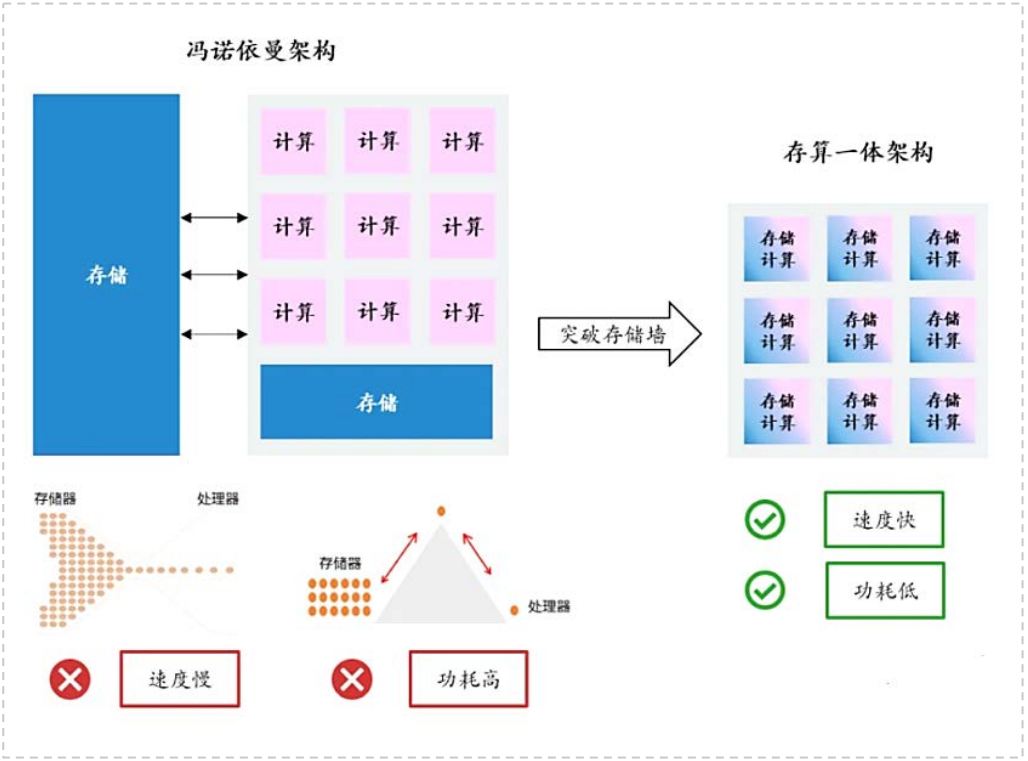
• 图形DDR SDRAM（GDDR SDRAM）最初是20多年前为游戏和显卡市场设计的。GDDR6 DRAM采用与生产标准DDR式DRAM的大批量制造和组装一样的技术。更具体地说，GDDR6采用传统的方法，通过标准PCB将封装和测试的DRAMs与SoC连接在一起。利用现有的基础架构和流程为系统设计者提供了熟悉度，从而降低了成本和实现的复杂性。GDDR6内存的优异性能特性建立久经考验的基础制造过程之上，是人工智能推理的理想内存解决方案。其出色的性价比使其适合在广泛的边缘网络和物联网终端设备上大量采用。采用GDDR6的主要设计挑战也来自于它最强大的特性之一：速度。在较低的电压条件，16 Gbps的信号速度下，**保持信号完整性需要大量的专业经验知识。设计人员面临更紧的时序和电压裕度量损失，这些损失来源与影响都在迅速增加。系统的接口行为、封装和电路板需要相互影响,需要采用协同设计方法来保证系统的信号完整性。**

DSA

• DSA（domain specific architecture，特定领域架构）是一种针对特定领域定制的可编程处理器，能够用于加速某些应用程序，实现更好的性能。 DSA与ASIC在同等晶体管资源下性能接近，两者最大的不同在于是否可软件编程。ASIC由于其功能确定，软件只能通过一些简单的配置控制硬件运行，其功能比较单一。而DSA则支持一些可编程能力，使得其功能覆盖的领域范围相比ASIC要大很多。如现在因为AI大行其道的GPU就是一个基于DSA思路设计的产品。包括谷歌、Tesla和Cerebras在内的厂商也针对其特定应用推行他们的DSA芯片。 **DSA是面向某个特定的领域定制优化的设计，和大规模落地相互是矛盾的，这就约束了DSA芯片的应用规模和商业价值。**

3、为了解决“两堵墙”的问题，存算一体是最具创新的技术方案

存算一体方案，这是一项诞生于实验室的新兴技术，其创新性在于打破了传统·冯诺伊曼架构局限性，实现了计算与存储模块一体化的整合创新，解决了传统芯片架构中计算与存储模块间巨大的数据传输延迟、能量损耗痛点，既增加了数据处理速度，又大大降低了数据传输的功耗，从而使芯片能效比（即每瓦能提供的算力）得到2-3个数量级（>100倍）的提升。



HBM、DSA、存算一体都属于芯片行业当前的技术创新路径，三者对比来看，存算一体可以算作是一条难度最大、颠覆性最强、风险最高，但差异化和创新性也最显著的路径。有观点指出，要想在纯市场化竞争中挑战英伟达等国际芯片巨头，必须另辟蹊径。于是差异化的技术创新成为芯片投资中的重要策略。

让芯片存算一体化拥有两种方案：**其一是将处理器和存储器放在同一芯片上，以减少数据交换、提升计算效率**，但处理器和存储器的制备工艺不兼容，且芯片中存储器密度受限，以目前及未来一段时间的技术水平来看，制造这种存算一体芯片的难度较大；**其二是基于新型存储材料和器件，是目前业界积极推进的一种方案。**

4、基于新型存储材料和器件，是目前存算一体最有前景的解决方案

存储器有许多种介质，不同介质实现存算一体的关键点也不同。从目前的存算一体发展技术路径来看，处于多种存储介质百花齐放的格局，包括各种易失性存储器件和非易失性存储器件（NVM），但**目前各个类型的存储器做存算一体都还存在自己的问题。**

类型	名称	全称	材料	片上单元大小/F^2	读取时间/ns	写入/擦除时间/ns	耐久性	写入功耗	其他能量消耗	高电压需求/V
易失性存储器 Volatile Memory	SRAM	静态随机存取存储器	硅晶体	50~120	1~100	1~100	1016	低	漏电流	无
	DRAM	动态随机存取存储器	硅晶体	6~10	30	15	1016	低	刷新电流	3
非易失性存储器 Non-volatile Memory	Flash（NOR）	非易失闪存（芯片内执行）	硅晶体	10	10	1μs/10ms	105	非常高	无	6~8
	Flash（NAND）	非易失闪存（数据）	硅晶体	5	50	1ms/0.1ms	105	非常高	无	16~20
	FRAM/FeRAM	铁电存储	铁电晶体材料	15~34	20~80	50	1012	低	无	2~3
	MRAM	非挥发性的磁性随机存储器	铁磁材料	16~40	3~20	3~20	>1015	高	无	3
	PCRAM/CRAM/PCM	相变随机存储器	相变材料（一种或多种硫系化合物薄膜）	6~12	20~50	60~120	108	高	无	1.5~3
	RRAM/ReRAM	阻变式存储器	非导性材料	6~10	10~50	10~50	108	低	无	1.5~3
	STT-RAM（第二代MRAM）	新型非易失性磁随机存储器	铁磁材料和非磁性隔离层	6~20	2~20	2~20	>1015	低	无	<1.5

参考资料丨沈志荣, 薛巍, 舒继武. 新型非易失存储研究[J]. 计算机研究与发展, 2014, 51(2): 445.

从产品成熟度来看，**目前相对比较成熟的存储产品选择是SRAM、DRAM和Flash**，三者的性能介绍如左图所示，其中SRAM、DRAM的主要优势体现在读写速度快，耐用性高，但缺点是**存储密度比较低**。为了实现更好的存算一体，目前产业界也在寻找基于新式存储打造解决方案，包括MRAM、PCM和RRAM，这三种类型的存储特性如左图所示。在参数方面，MRAM已经做到了很好的产品耐用性，RRAM在容量方面已经超越了DRAM，不过它们的**读写速度**还比不上传统存储器。

目前各个类型的存储器做存算一体都还存在自己的问题。传统存储产品SRAM和DRAM属于易失性存储器件，刷新的频率越高，**功耗的问题会越明显**，Flash 虽然是非易失性的，但其浮栅氧化层随着读写次数的增加有**失效的问题**，数据可靠性和寿命有待提高;新型存储产品方面，PCM、RRAM和MRAM都是非易失性存储，功耗不会再是挑战，不过PCM写入速度极慢，RRAM在容量方面最具优势，但写入速度同样是短板，MRAM虽然存储密度高，但容量提升还是个大问题，且写入速度也很慢。同时，几乎所有存储产品都要面临一个共同的问题——越是先进的工艺，存储产品良率提升越困难，但存内计算需要更出色的存储产品。

5、产业巨头在布局MRAM，但离落地量产还有距离

磁性随机存储器（MRAM）是一种基于自旋电子学的新型信息存储器件，其核心结构由一个磁性隧道结和一个访问晶体管构成。MTJ 呈现“三明治”结构，两层磁性固定层和自由层之间夹着一层隧穿层。这其中，铁磁层材料一般使用 CoFeB，隧穿层材料则为 MgO。固定层的磁化方向是不变的，而自由层的磁化方向可以被改变。当固定层和自由层磁化方向一致时，称为“平行状态”，MTJ 的隧道磁阻(Tunnel Magnetoresist-ance，TMR）为低；当磁化方向不一致时，称为“反平行状态”，TMR 为高。**它具有极快的开关速度、近乎为零的泄露功耗、极高的可靠性等显著优点，是实现存算一体化技术的理想器件之一。**

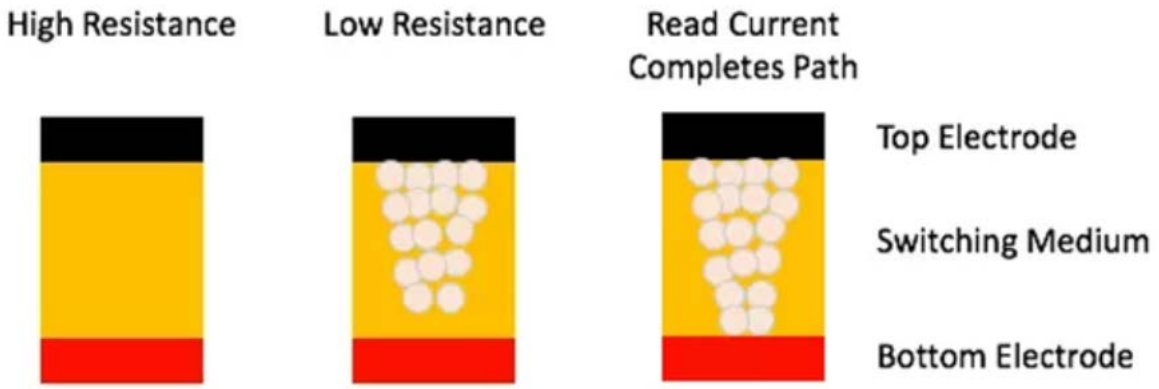
Year	Group	Capacity	CMOS (nm)	Cell (um2), Chip (mm2)	Speed (ns)	Power or Current
2017	SK Hynix & Toshiba [7]	4Gb	\	0.0081, 107.5	R:2.5	\
2018	U. Michigan & TSMC[8]	1Mb	28	0.0588, 0.214	R:2.8, W:20	W:3.6 mW, R:3.9 mW
2018	NTHU&TSMC[9]	32Kb	28	\, 0.18	R: 2	\
2019	Intel[10]	7Mb	22	0.0486, \	R:4 or 8	\
2019	Tohoku U.[11]	64Kb	40	\	\	47.14uW
2019	TSMC[12]	16Mb	40	0.066, 3.03	<17.5	
2020	NTHU[13]	1Mb	22	\, 3	2.75	R:0.23 pJ
2020	TSMC[14]	32Mb	22	0.0456, 3.6	R:10 or 6	R:0.8 uA, W:10 mA

2017-2020年的MRAM的公开芯片情况

MRAM虽然存储密度高，但容量提升还是个大问题，且写入速度也很慢。高通正在加速研究 STT-MRAM 技术，在过去几年里，包括**台积电、英特尔、三星、SK海力士**等晶圆代工厂和IDM，相继大力投入MRAM 研发。此外不少创新公司，如 Everspin，Avalanche, Crocus, Spin Transfer Technology 也已能够提供 MRAM 样品。整体上，STT-MRAM 已经在 2×nm 节点的嵌入式存储市场准备就绪。华为与 SNIA（全球网络存储工业协会）在2015 年共同举办的存储技术峰会上，Albert Fert（2007 年诺贝尔物理学奖获得者）表示：在存储技术底层是依靠核心技术和理论突破存储瓶颈。自旋转移矩技术取得了重大突破，基于 DMI 效应的方式使得 STT-MRAM 存储单元可以缩放到几纳米，显著的提高了磁存储器的集成度和性能。

6、新型存储器RRAM也被广泛看好，但在工艺成熟度和商业化上还需要一些耐心

阻变存储器（ReRAM或RRAM，Resistive RAM）全称是电阻式随机存取存储器，是以非导性材料的电阻在外加电场作用下，在高阻态和低阻态之间实现可逆转换为基础的非易失性存储器。ReRAM包括许多不同的技术类别，比如氧空缺存储器（OxRAM，Oxygen Vacancy Memories）、导电桥存储器 （CBRAM，Conductive Bridge Memories）、金属离子存储器（Metal Ion Memories）以及纳米碳管 （Carbon Nano-tubes）。以Crossbar和昕原半导体为例，其采用对CMOS友善的材料，能够使用标准的CMOS工艺与设备，对产线无污染，整体制造成本低，可以很容易地让半导体代工厂具备ReRAM的生产制造能力，这对于量产和商业化推动有很大优势。



在商业化上，Crossbar、昕原半导体、松下、Adesto、Elpida、东芝、索尼、美光、海力士、富士通等厂商都在开展ReRAM的研究和生产，其中专注IP授权的Crossbar对于ReRAM的基础技术研发走在了前列。在代工厂方面，中芯国际（SMIC）、台积电（TSMC）和联电（UMC）都已经将ReRAM纳入自己未来的发展版图中。

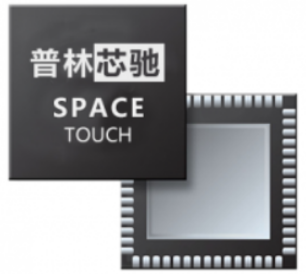
7、针对小算力、低功耗的AI应用场景，存算一体创业公司在快速发展

虽然基础的存储材料和器件、开发生态还不够成熟，但存算一体的创业公司在低功耗、低算力的AI应用场景中找到了一些市场，比如语音、AIoT、安防等领域。比如珠海普林芯驰在离线语音的产品线，通过超低功耗SRAM运算加速核的存算一体架构，能耗可降低85倍，芯片性能提升10倍以上。知存科技使用Flash存储器完成神经网络的储存和运算，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，适合可穿戴设备中的智能语音和智能健康服务，已完成批量生产。

智能离线语音交互MCU

SPV20系列芯片采用 CPU+DSP+NPU 三核架构，内置基于人工智能语音识别算法的NPU硬件加速核，通过神经网络对音频信号进行训练学习，提高语音信号的识别能力。CPU 与DSP 的代码存储于外部SPI NOR Flash，通过XIP方式执行，并通过4 Way-cache机制保证程序的高效执行。内置2路模拟麦克风codec。扩展I2S/DMIC接受最多支持4路音频信号输入，用于远场拾音的麦克风阵列方案，最多可支持100条左右的语音命令，且支持网站自定义词条生成，高效、快速；

SPV20系列芯片具有丰富的系统外设，包括 UART/I2C/SPI/PWM/RTC/Timer/ 断码管驱动/ADC/触控模块，是一款主控级别的离线语音识别MCU。广泛应用于家电的主控方案；



图：珠海普林芯驰官网

WTM8000

图：知存科技官网

8、布局新型存储材料和器件生产代工，持续迭代，谋求弯道超车

目前，国内也在布局**新型存储材料和器件的生产线，典型公司如**昕原半导体（杭州）有限公司、厦门半导体工业技术研发有限公司等。

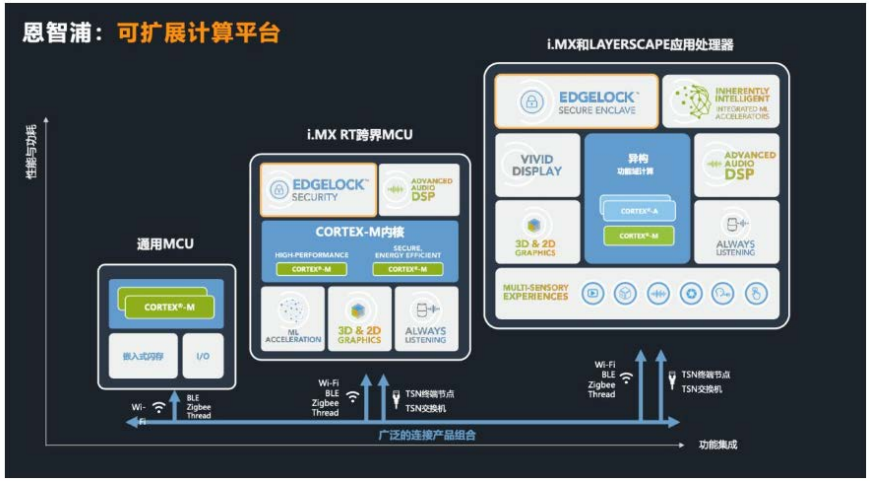
昕原半导体（杭州）有限公司，专注于ReRAM领域，致力于提供革新性的存储、存内计算等多种创新芯片产品和IPs，是一家集核心技术、工艺制程、芯片设计、IP授权和生产服务于一体的新型IDM公司。主导建设的**大陆首条28/22nm ReRAM（阻变存储器）12寸中试生产线**顺利完成了自主研发设备的装机验收工作，实现了中试线工艺流程的通线，并成功流片。新型存储器的核心，是在其开发中需要在传统CMOS工艺里增加一些特殊的材料或工艺，这些特殊材料或工艺的开发则需要经过大量实验及测试验证，昕原可加速国内各大科研院所在新型存储器的研发及量产能力。

昕原半导体技术人员对中试线进行设备调试

9、存算一体的发展路径及机会探讨

长期来看，存算一体方案的发展前景无疑非常有吸引力。但大的发展机会，必须依赖于基础存储材料和器件的技术发展。存算一体芯片按应用领域可分为**服务器、智能终端及AIOT领域**，在计算功能上可分**为主处理器芯片和协处理器芯片。**从发展阶段来看，存算一体芯片开始往往针对某种特定算法进行优化加速，存在比较高的客制化程度，存算一体芯片大概率是从协处理器起步。这个阶段，产品定义能力是创业团队最为关键的挑战，需要兼顾产品的应用场景、算力需求、性价比、方案的延展性等。

单纯作为一颗AI芯片，一定会面临通用集成芯片的竞争，方案的灵活性、针对场景的优化、性价比及市场规模就比较关键。可以看到，在AIOT领域，集成算力的MCU已经成为趋势。比如在通用 MCU 家族的大产品类别中，恩智浦推出了面向边缘计算的全新产品—— MCX 微控制器产品组合。MCX 的一大亮点就是拥有真正的硬件的神经处理单元，其在内部集成了 NPU，这也是恩智浦集成 NPU 的第一个产品家族。




THANK
YOU FOR
WATCHING
谢谢您的耐心观看