三星公司PIM进展：

MRAM路线

实现了基于MRAM(磁阻随机存取存储器)的存内计算

MRAM磁阻内存在运行速度、寿命、量产方面都有明显优势，功耗也远低于传统DRAM，还具有非易失的特点。但由于它在标准的存内计算架构中无法发挥低功耗优势，因此难以用于存内计算。

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存内计算架构的MRAM阵列芯片。

这一阵列成功解决了单个MRAM器件的小电阻问题，从而降低功耗，实现了基于MRAM的内存内计算。

按照三星的说法，在执行AI计算时，MRAM内存内计算可以做到98％的笔迹识别成功率、93％的人脸识别准确率。

A crossbar array of magnetoresistive memory devices for in-memory computing

可批量生产的四种NVMs：

1. 电阻性随机存取存储器 resistive random-access memory
2. 相变随机存取存储器 phase-change random-access memory
3. 闪存 flash memory
4. 自旋转移转矩磁阻随机存取存储器 spin-transfer-torque magnetoresistive random-access memory （STT-MRAM）

（磁性随机存储器（MRAM）的发展

磁性随机存储器（MRAM）作为新型NVM的一种，具有接近零的静态功耗，较高的读写速度，与互补金属氧化物半导体（CMOS）工艺相兼容等优点，在车用电子与穿戴设备等领域已实现商业化应用，被认为是最有希望的下一代存储器之一。目前MRAM技术的发展趋向于在先进技术节点提供高的写入速度和低的功耗。

第一代 MRAM 的写入方式是磁场写入式。它利用导体中流过的正反两个电流方向产生磁化感应，使得 STT 存储单元的 MTJ 自由层相对固定层磁化方向为同向和反向两个方向，从而表现出高低两种大小的阻抗状态用于存取数字逻辑“1”和“0”。由于这种写入方式需要足够强的磁场感应，因此需要足够大的写入电流，即需要利用高压来进行写，这导致了功耗的增大。同时随着工艺尺寸的缩减以及MTJ 结尺寸的降低，相邻 STT 存储单元的可靠性问题变得严重，通过增大间距减小相互干扰则会降低MRAM 的集成度。因此，可靠性和集成度之间是一对矛盾。

第二代 MRAM 的写入方式是自旋转移矩写入式。它是利用自旋转移矩(spin - transfer torque)效应诱导磁性材料发生磁化翻转，即利用流过隧道结中不同方向的自旋极化电流，驱动软磁体磁化方向的改变，实现 MTJ 结高低阻抗状态的写入。这种写入方式的存储器，称之为自旋转移矩磁随机存储器(STT-MRAM：Spin Transfer Torque MRAM)。与第一代 MRAM 依靠磁场写入的方式不同，这种方式是依靠自旋转移矩效应，它利用电流来完成磁性薄膜的磁化方向翻转，继而实现不同的阻抗状态。这种方式，最大的优势是降低了对写电压的要求，从而也降低了功耗，并且避免了磁场感应引起的串扰问题，有效提高了集成度。因此对磁存储器研究自然的过渡到了对自旋转移矩磁随机存储器的研究。

第三代MRAM的自旋道矩磁随机存储器（SOT-MRAM），电流通过底层重金属，产生自旋流并注入到自由层中，利用自旋轨道矩使自由层的磁化方向产生扰动，并结合多种方式让磁化方向产生确定性的翻转。相比于自旋转移矩的存储技术，基于自旋轨道矩的存储技术具有对称的读写能力、分离可优化的读写路径、亚纳秒的快速操作速度和低写入功耗等优点。

在商业化发展方面：

第一个商业化的磁随机存储器产品是飞思卡尔半导体公司（FreescaleSemiconductor）于2006 年生产 的4 Mb 容量Toggle-MRAM ，该部门是Everspin 的前身。之后Everspin 公司推出了具有SRAM 速度和 闪存结构的非易失的Toggle-MRAM ，其16 位32Mb并行MRAM 具有最高35 ns 的写入周期时间，工作温度范围为-40~125°C ，适用于工业和汽车应用。Everspin 在 2012 年就生产出了首个商业级的 64Mbit 的 STT-MRAM 存储器芯片并在2017年大批量生产了256Mb DDR3 STT-MRAM，并于2017年集成了40 nm CMOS，并在2019年批量生产了28 nm CMOS上的1Gb DDR4 STT-MRAM。

东芝公司在 2014 年的VLSI Symposia上提出了用 STT-MRAM 代替 SRAM，使得微处理器中的高速缓存功耗降低了近 60%。日本超低压元器件技术研究联盟（LEAP）在 VLSI Symposia 上同样实现了用 STT-MRAM 对传统 SRAM的替换。

高通公司的项目研究员 Seung H. Kang 2015 年发布了一款 8Mbit 混合 STT-MRAM 的SOC 芯片 。在没有软错误的情况下，工作频率可以达到 500MHZ。在 40nm 工艺下，适用于混合STT-MRAM 的 SOC，若工艺尺寸可以降到 28nm，那么其性能可以与SRAM 相媲美。

https://baijiahao.baidu.com/s?id=1731074817335671347&wfr=spider&for=pc

）

MRAM中，平行态在层状结构中表现出低电阻(RL)，而反平行态则表现出高电阻(RH)。

在标准的“电流总和”(current-sum)架构阵列中，由于平行导致低电阻，进而导致了功率消耗过大（U^2/R）

在实验时，三星团队采用了二进制神经网络(BNN)算法，该算法可以通过以网络大小为代价用多个二进制值表示每个实值权重或以计算速度为代价将每个实值输入数据表示为多个二进制数序列来提高精度。三星团队采用后者，将每个输入数据扩展为8-bit thermometer code。

潜在问题：采用新结构来进行计算可能存在模拟噪声，这对阵列规模有一定的限制。因此，如何构建一个集成了许多有限尺寸阵列的片上系统AI处理器是MRAM后续发展的一个重要挑战

DRAM路线

2021年开发出集成了人工智能 (AI) 处理能力的高带宽内存 (HBM)——HBM-PIM。新款内存内处理 (PIM) 架构将强大的 AI 计算能力引入高性能内存，加速数据中心、高性能计算 (HPC) 系统和支持 AI 的移动应用程序中的大规模处理。

HBM-PIM 通过将经过动态随机存取存储器 (DRAM) 优化的 AI 引擎置于每个内存库（存储子单元）内，将处理能力直接引入到数据存储位置，从而实现并行处理并尽可能减少数据移动。当应用于三星现有的 HBM2 Aquabolt 解决方案时，新架构能够提供超过两倍的系统性能，同时将能耗降低 70% 以上。HBM-PIM 也不需要任何硬件或软件变更，从而可以更快地集成到现有系统中。相关论文发表于2021年的ISSCC会议上

A 20nm 6GB Function-In-Memory DRAM, Based on HBM2 with a 1.2TFLOPS Programmable Computing Unit Using Bank-Level Parallelism, for Machine Learning Applications

每个存储库中一半的单元阵列被移除，并被可编程计算单元(PCU)取代。这些单元通过来自主机的常规存储命令进行控制，以启用DRAM中的计算功能，不同的是，它们可以执行FP16的计算。为了支持并发多bank操作，添加了FIMDRAM的特殊FIM模式。FIM模式下，所有pcu同时开机，多个bank从主机执行相同的命令，而普通模式下支持HBM2的基本操作。

三星开发了基于fpga的平台和基于应用模型的仿真环境，以验证FIMDRAM对系统性能的改善

据Business Korea报道，三星对AMD基于CDNA架构的Instinct MI100计算卡进行改造，加入了HBM-PIM芯片，然后使用96块经过改造的Instinct MI100计算卡构建了一个大型计算系统。相比于未改造前的系统，新系统在使用训练语言模型算法T5时性能提高了2.5倍，功耗降低至原来的2.67分之一，大大提高了GPU运行AI算法的效率。