

# 存算一体——重塑芯片架构新势力



## CIM/PIM/PNM: An Emerging Force Reshaping Chip Architecture

刘新阳/Liu Xinyang, 李婷宇/Li Tingyu, 巨鸣/Ju Ming

(深圳市中兴微电子技术有限公司, 中国 深圳 518055)  
(Sanechips Technology Co., Ltd, Shenzhen 518055, China)

DOI: 10.12142/ZTETJ.202604010

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260904.1059.002>

网络出版日期: 2026-08-24

收稿日期: 2026-06-02

**摘要:** 存算一体技术通过将计算单元与存储单元深度融合或物理级邻近部署, 从架构层面缓解了数据搬运瓶颈带来的压力。大语言模型 (LLMs) 及智能体 (Agentic AI) 的持续演进, 对计算架构的高带宽、低延迟与高效能提出了严苛要求。系统梳理了基于静态随机存取存储器 (SRAM)、动态随机存取存储器 (DRAM)、闪存 (Flash) 及新兴非易失性存储器的存算一体技术路径, 涵盖存内计算 (CIM)、存内处理 (PIM) 与近存处理 (PNM) 三大主流范式, 综述其研究进展、典型应用案例及面临的技术挑战。分析表明, 不同技术路径展现出差异化的性能与能效优势, 可通过场景驱动的技术融合、软硬件协同设计及产业链生态协同实现优势互补, 为下一代人工智能芯片的架构创新提供有益参考。

**关键词:** 大语言模型; 智能体; AI芯片架构; 存算一体

**Abstract:** Compute-in-Memory (CIM), Processing-in-Memory (PIM), and Processing-Near-Memory (PNM) hold the promise of fundamentally resolving the data movement bottleneck by integrating computing logic with memory units or placing them in close physical proximity. A systematic review is presented of CIM, PIM, and PNM technical paths implemented on static random-access memory (SRAM), dynamic random-access memory (DRAM), Flash, and emerging non-volatile memories, covering their research progress, representative use cases, and associated challenges. The analysis reveals that each technical path exhibits distinct performance and energy-efficiency advantages. It is argued that scenario-driven technology fusion, hardware-software co-design, and industry-wide ecosystem collaboration are essential for harnessing complementary strengths across these paths. This work is intended to serve as a valuable reference for architectural innovation in next-generation AI chips.

**Keywords:** large language models; Agentic AI; AI chip architecture; computing-in-memory

**引用格式:** 刘新阳, 李婷宇, 巨鸣. 存算一体——重塑芯片架构新势力 [J]. 中兴通讯技术, 2026, 32(4): 63–71. DOI: 10.12142/ZTETJ.202604010

**Citation:** Liu X Y, Li T Y, Ju M. CIM/PIM/PNM: an emerging force reshaping chip architecture [J]. ZTE technology journal, 2026, 32(4): 63–71. DOI: 10.12142/ZTETJ.202604010

## 1 内存墙成为 AI 系统核心瓶颈

### 1.1 从大语言模型到智能体

近年来, 随着生成式人工智能 (Generative AI) 的快速发展, 以 Transformer 架构为核心的大语言模型 (LLM) 凭借其强大的语言理解、内容生成与推理能力, 成为推动通用人工智能发展的核心引擎。进入 2026 年, AI 应用正从传统聊天机器人 (Chatbot) 的单次问答推理, 向多轮自主智能体 (Agent) 交互范式深度演进。然而, 算法与应用层面的繁荣, 正对底层计算硬件提出前所未有的挑战。在冯·诺依曼存算分离架构下, 经典的“内存墙 (Memory Wall)”问题<sup>[1]</sup>被 LLM 推理与 Agent 应用持续放大, 数据搬移而非计算本身成为制约系统性能与能效的核心瓶颈。

大语言模型推理面临以下几个方面的内存墙方面挑战:

1) 带宽瓶颈。在 Transformer 的自注意力计算中, 每个 Token 需与所有历史 Token 进行交互, 这意味着即使参数量相对适中, 其推理过程仍须以极高的带宽频繁访问大规模的权重数据和键值 (KV) 缓存<sup>[2]</sup>。

2) 容量瓶颈。大语言模型的全量参数往往远超单一芯片的存储容量, 必须依赖分布式存储与并行计算技术。这不仅增加了系统设计的复杂性, 也带来了严重的芯片间通信开销。

3) 能耗瓶颈。对于大语言模型这类访存密集型的负载, 数据搬运往往主导着系统总能耗<sup>[3]</sup>。从 Mark Horowitz 的经典数据<sup>[4]</sup>来看, 在 45 nm 工艺下, 一次动态随机存取存储器 (DRAM) 访问的能耗约为一次 64 位浮点运算能耗的 100~1 000 倍。由于工艺微缩对 DRAM 访存能耗的改善远滞后于对计算逻辑能耗的改善, 这一比例在更先进的工艺节点下可能进一步扩大。

如表1所示,智能体不仅是一个被动的对话模型,还具备环境感知、多轮迭代、长期记忆、工具调用及多智能体协作等能力<sup>[5]</sup>。这些特性进一步加剧了内存墙问题,主要体现在以下几个方面:

1) 多模态融合带来的数据规模膨胀。现代智能体系统通常需要感知并处理文本、图像、音频、视频等多种模态的信息。这种多模态融合的处理模式意味着系统所需存储与访问的数据规模相比纯文本模型将提升1~2个数量级。

2) 调用次数的倍增。智能体多轮迭代的执行机制使访存密集的解码阶段被反复触发。一个包含 $k$ 步推理的智能体任务,其总访存开销约为单次推理访存量的 $k$ 倍,呈线性甚至超线性增长。

3) 上下文长度的累积与生命周期的延长。智能体在执行过程中需不断将检索增强生成结果、历史观察、中间结果及工具返回信息追加至上下文,导致序列长度持续增长。与此同时,KV Cache管理需支持跨轮生命周期及长期记忆。这将带来KV Cache存储与访存量的同步膨胀,使长上下文下的存储容量和带宽压力进一步加剧。

4) 多模型或多智能体协作。智能体系统通常需并发处理多个子任务,多个智能体可能并行运行并通过消息传递进行协作。每个智能体实例均需独立的模型副本、上下文缓存及中间状态存储,因此当智能体数量增加时,系统对内存容量和带宽的需求呈线性增长。

综上,智能体应用将原本已属访存受限的LLM推理进一步推至“访存极度受限”的境地。可以预见,随着智能体自主性与复杂度的提升,内存墙将成为制约其规模化部署的核心障碍。

## 1.2 芯片架构的局限性和发展方向

过去20年,处理器算力的提升速度远高于内存DRAM带宽和互连带宽的增长速度。硬件峰值算力增长了约60 000倍,而DRAM带宽和互连带宽仅分别增长100倍和30倍<sup>[1]</sup>。这意味着,单纯增加计算单元无法带来等比例的性能提升,反而会导致硅片利用率下降和能耗剧增。

如图1所示,随着访问距离增大,传统存储层级延迟、带宽、功耗瓶颈加剧。当前,叠加模型与应用端范式的转变,AI芯片的竞争逻辑正在重构:竞争焦点不再局限于算力峰值,而拓展至内存容量、带宽、能耗及成本控制的综合较量。如何突破内存墙、优化数据移动路径、重构内存层级,已成为决定AI芯片成败的核心技术挑战。

为应对这一挑战,芯片架构设计正从“以计算为中心”转向“以存储为中心”,其核心原则是以存储容量最大化、

表1 从LLM Chatbot到Agent,内存墙瓶颈加剧

|               | LLM Chatbot | Agent                   |
|---------------|-------------|-------------------------|
| 模态            | 文本          | 文本、图像、音频、视频等            |
| 调用模式          | 单轮问答        | 多轮迭代                    |
| 上下文长度         | 数千Tokens    | 百万Tokens(累积工具调用、记忆、文档等) |
| KV Cache 生命周期 | 会话级,可丢弃     | 跨轮持久化,需长期驻留或反复加载        |
| 模型组合          | 单一大模型       | 多模型/多智能体协作等异构组合         |

KV: 键值 LLM: 大语言模型

数据访问带宽最大化、数据移动能耗最小化及系统成本最小化为目标,优化整个计算-存储数据流。主要实现路径包括:

1) 基于软硬协同的存储层级重构:根据LLM与Agent的网络结构与数据流向,重新设计由不同存储介质构成的存储层级,缩短或消除冗余访问路径。

2) 基于存算一体的数据就近计算:从器件、电路、芯片、系统等不同层面,拉近计算单元与数据存储的物理距离,从根本上消除数据在存储器与处理器之间的频繁搬运<sup>[6]</sup>。

## 2 存算一体技术路径分析

存算一体技术通过使计算单元与数据存储单元在物理上相互靠近乃至融合,使数据在存储位置即可完成计算,从而大幅减少数据搬运的距离与次数,提升访存带宽并降低访存功耗。依据以下3个维度,可形成多种存算一体技术路径。

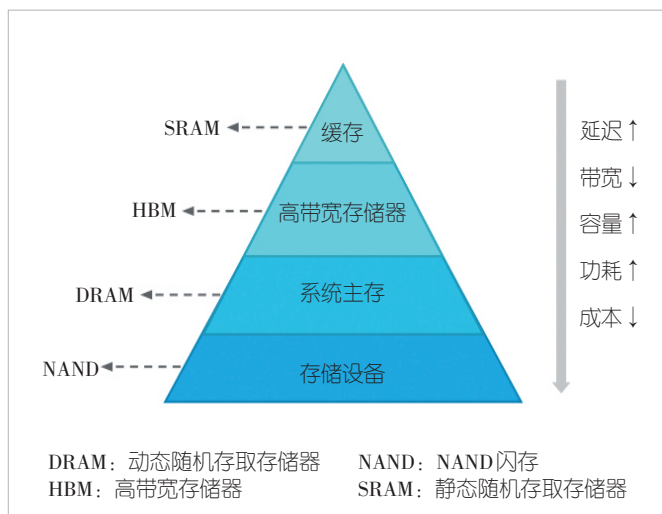


图1 典型内存层级架构

1) 按存储类型分类。依据所用存储介质的不同,可将存算一体技术路径分为两类:基于成熟存储介质的技术路径,包括静态随机存取存储器(SRAM)、DRAM、闪存(Flash)等;基于新型非易失性存储介质的技术路径,包括阻变随机存取存储器(ReRAM)、磁阻随机存取存储器(MRAM)、相变随机存取存储器(PCRAM)、铁电随机存取存储器(FeRAM)等。

2) 按存-算距离分类。依据计算单元与存储单元在物理上的距离与集成方式,可分为以下3种:

(1) 存内计算(CIM):存储单元与计算单元集成于同一芯片内,且两者紧密耦合,需对存储阵列进行改动,使存储单元可直接参与计算操作。

(2) 存内处理(PIM):存储单元与计算单元同样集成于同一芯片内,但两者相对独立,一般无须改动存储阵列,存储单元仅提供数据访问功能。

(3) 近存处理(PNM):存储单元与计算单元位于不同芯片,无须改动存储阵列,通过2.5D或3D先进封装技术缩短芯片间的物理距离。

3) 按封装方式分类。依据存储芯片与计算芯片的封装集成形式,可分为以下3种:

(1) 2D封装:该方式一方面代表传统存算分离的封装形式,如独立的低功耗双倍数据速率内存(LPDDR)、图形双倍数据速率内存(GDDR)、双倍数据速率内存(DDR)等;另一方面涵盖在同一芯片内实现的CIM。

(2) 2.5D封装:计算芯片与存储芯片通过中介层(Interposer)在水平方向进行互连,典型应用于高带宽存储器(HBM)或高带宽闪存(HBF)与计算芯片的封装集成。

(3) 3D封装:计算芯片与存储芯片通过混合键合和硅通孔(TSV)等技术,在垂直方向进行三维堆叠集成。

下文将依据上述分类维度,依次剖析各条技术路线的核心原理与特点。其中,基于SRAM、DRAM和Flash的存算一体方案,技术架构如图2<sup>[7-10]</sup>、图3<sup>[11-13]</sup>和图4<sup>[14]</sup>所示,后续分析将围绕这3类主流存储介质展开。

## 2.1 基于SRAM的存算一体技术

### 2.1.1 SRAM CIM

SRAM CIM架构在SRAM阵列内部将SRAM Bitcell与算术逻辑电路(如乘法器、加法树、移位累加单元等)紧密耦合,从而支持乘累加(MAC)计算。在相同先进工艺节点下,SRAM CIM Macro在实现相同功能时,相较于传统数字

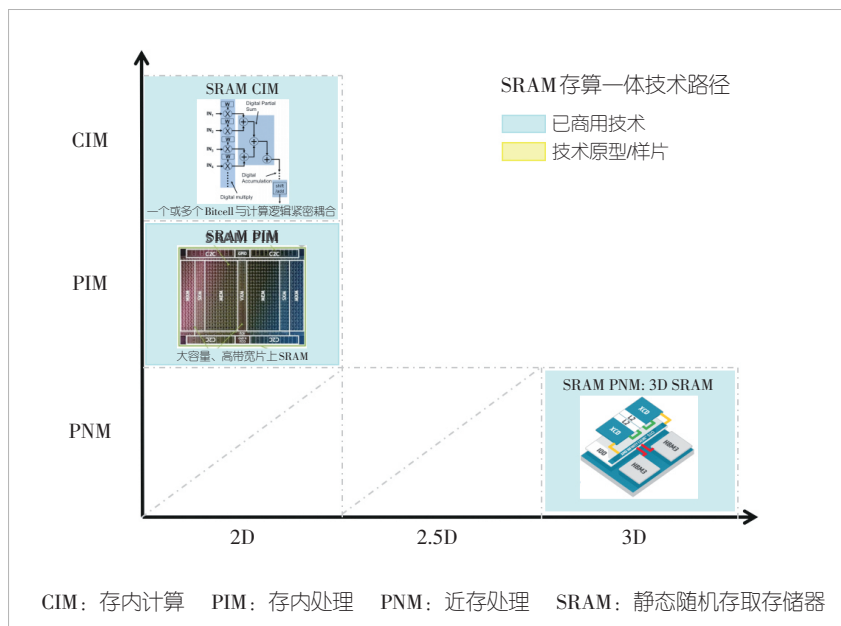


图2 SRAM存算一体技术路径<sup>[7-10]</sup>

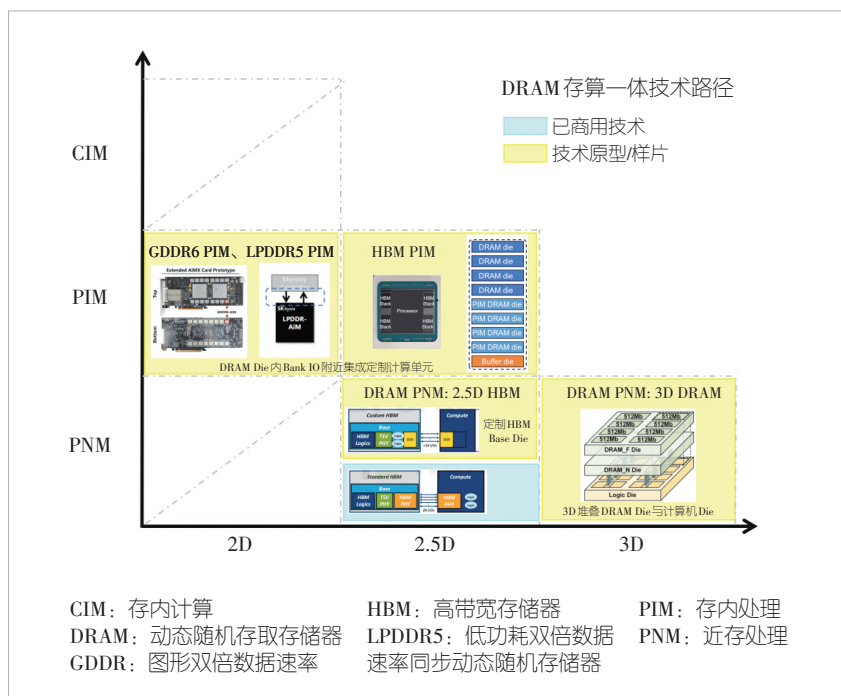
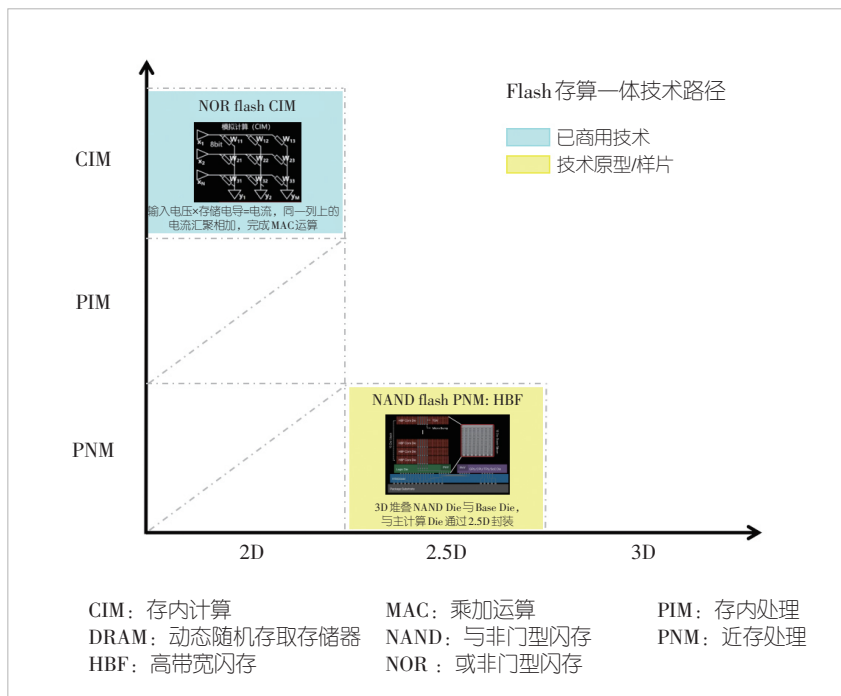


图3 DRAM存算一体技术路径<sup>[11-13]</sup>

图4 Flash存算一体技术路径<sup>[14]</sup>

计算单元可获得更高的能效与面积效率，如表2所示<sup>[15-16]</sup>。中兴微电子与北京大学合作了基于5 nm工艺的SRAM CIM Macro模型，在0.6 V供电电压、50%稀疏度条件下，能效达到111 8b-TOPS/W；同时，双方还开发了集成该模型的仿真套件，用于SRAM CIM架构的性能、功耗、面积（PPA）仿真预测。

SRAM CIM Macro的容量通常限于数十kb，其原因在于：当阵列规模增大时，输入并行度随之增加且走线距离变长，从而导致布线拥塞、延迟增加及可靠性下降。鉴于单个SRAM CIM Macro的容量受限，实际往往通过拼接多个Macro来扩展容量与算力，但拼接本身也引入了Macro间互连带宽、延迟及面积开销等问题。

d-Matrix基于6 nm工艺的Corsair计算卡是采用SRAM CIM架构的代表性产品。该产品结合Chiplet与精简指令集计算机体系结构（RISC-V）内核等技术，实现了8个Chiplet、总算力2.4千万亿次浮点运算每秒（PFLOPS）（MXINT8）、SRAM总带宽150 TB/s及SRAM总容量2 GB的规格。然而，

2 GB的片上容量对于大语言模型而言依然十分有限，仍需借助低功耗双倍数据速率同步动态随机存储器（LPDDR5）等片外存储进行扩展。受SRAM存储容量制约，该产品的应用场景主要面向10B以下的小语言模型（SLM）以及Whisper等语音模型，旨在使模型尽可能驻留于片上，减少数据搬移，从而实现极低时延的AI推理。在成本方面，采用SRAM加LPDDR5、不使用HBM的架构提供了一种低成本的推理方案。在能效方面，尽管SRAM CIM单元本身的能效可达约20~40 TOPS/W，但其功耗仅占计算卡整体功耗的20%左右，计算卡中的矢量计算单元、交叉开关、输入输出（IO）模块及片外DDR等部件会显著稀释SRAM CIM所带来的能效收益。

### 2.1.2 SRAM PIM

SRAM PIM架构通过减少控制与计算单元的面积开销，换取更大的片上SRAM存储。

业界典型方案包括Groq 3语言处理单元（LPU）、SambaNova SN50及Cerebras WSE-3等。以Groq 3 LPU为例，该芯片基于三星4 nm工艺，片上SRAM面积占比高达40%，容量可达500 MB，聚合带宽达150 TB/s，有效降低了片外访存的带宽压力与功耗。同时，其结合编译器针对确定性负载预先设定数据流动路径与指令排布，从而实现极低延迟推理。

然而，即使在此类芯片架构中配置如此高比例的SRAM，受限于SRAM的存储密度，单芯片仍难以满足大模型部署需求。针对这一问题，目前存在两种主要解决方案。一种是通过片间与机间互联实现多芯片、多机柜扩展，进而支持系统级并行切分与部署，该方案对芯片间互联带宽要求较高。Groq 3 LPU即采用此方案，而晶圆级芯片Cerebras WSE-3本质上是片间互联转换为Die间的互联。另一种方案是结合HBM、DDR等多级片外存储进行容量扩展，虽然在一定程度上牺牲了纯SRAM架构的高性能与确定性，但显著降低了最小部署规模所需的芯片数量与成本，SambaNova

表2 SRAM CIM Macro与传统数字计算单元收益对比<sup>[15-16]</sup>

|                     | 工艺   | 能效                                  | 面效                                 |
|---------------------|------|-------------------------------------|------------------------------------|
| SRAM CIM Macro(台积电) | 5 nm | ~150 4b-TOPS/W @0.5V (50% sparsity) | 221 4b-TOPS/mm <sup>2</sup> @0.9V  |
| 传统数字计算单元            |      | ~90 4b-TOPS/W @0.46V (50% sparsity) | ~100 4b-TOPS/mm <sup>2</sup> @1.1V |

CIM: 存内计算

是该方案的代表。

2.1.3 SRAM PNM

3D SRAM通过3D封装技术将计算Die与SRAM Die进行垂直堆叠。业界的代表产品包括AMD MI300/MI400 GPGPU系列，该系列通过优化末级缓存(LLC)的容量与带宽，在未牺牲片上算力资源的前提下将片上SRAM的容量扩展至256 MB。其3D SRAM带宽达17.2 TB/s。此外，计算Die与SRAM Die可采用不同工艺制程，有助于节约制造成本。然而，该技术对3D封装工艺与系统协同设计的要求较高，封装成本亦相对较高。

2.2 基于DRAM的存算一体技术

2.2.1 DRAM PIM

如表3所示，DRAM PIM架构在DRAM Die内的存储Bank附近集成计算单元，借助Bank级并行I/O大幅提升内部计算单元的访存带宽。相较于外部接口，内部带宽可提升约一个数量级，同时显著减少DRAM与AI芯片之间的数据搬运延迟与功耗。然而，DRAM Die基于DRAM工艺制造，其内部集成的计算单元在算力性能上面临限制。究其原因，DRAM工艺以存储密度最大化与漏电流最小化为核心目标，而计算单元的性能提升高度依赖于逻辑工艺所具备的高晶体管开关速度与多层互连能力。此外，计算单元的引入需占用DRAM存储面积，导致存储容量下降，散热与供电负担增加。因此，DRAM PIM中的计算单元通常采用较低的工作频率与数据传输速率，主要用于卸载算力需求较低但带宽需求较高的通用矩阵-向量乘法(GEMV)计算。三星与SK海力士等厂商已基于该技术路线研制出原型产品，但由于其计算能力有限且容量不足，难以满足日益增长的AI工作负载需求，商业应用推进缓慢。

2.2.2 DRAM PNM

标准HBM正逐渐逼近其性能瓶颈。一方面，为支持更高带宽，DRAM Die需容纳更多硅通孔输入/输出接口(TSV I/O)，导致其容量密度显著低于普通双倍数据速率同步动态随机存储器(DDR)；另一方面，为提升容量需增加堆叠层数，但封装与散热难度随之增大，良率下降，导致HBM成本过高，性价比比较低。因此，业界开始推动DRAM PNM架构研究。与PIM不同，DRAM PNM的计算与存储位于不同Die，工艺互不受限，可各自独立优化。按封装方式划分，DRAM PNM分为2.5D和3D两条技术路线。

2.5D路线在标准HBM基础上引入定制HBM Base Die。定制HBM Base Die的主要功能包括：1) 集成部分计算单元或控制逻辑，以节约AI芯片面积；2) 集成定制芯片对芯片物理层接口知识产权(D2D PHY IP)，优化处理器与HBM之间的接口IO密度与速率；3) 集成LPDDR控制器<sup>[7]</sup>，使AI芯片无须通过中央处理器(CPU)中转即可直接经AI芯片—HBM—LPDDR路径访问LPDDR数据，实现多层级存储融合。

以Marvell定制HBM Base Die方案为例<sup>[8]</sup>：1) 将原本位于主计算Die的HBM控制器集成至HBM Base Die，以节省主计算Die面积；2) 将标准HBM PHY转换为高密度、低功耗的定制D2D PHY，使线带宽密度从5 Tbit·s<sup>-1</sup>·mm<sup>-1</sup>提升至20~30 Tbit·s<sup>-1</sup>·mm<sup>-1</sup>，同时减少约75%的I/O功耗。2026年台积电技术研讨会展示了2023—2029年AI系统存储带宽的提升路径：通过采用基于先进工艺的定制HBM Base Die，结合更高代际的HBM颗粒与2.5D单颗晶圆上芯片封装(CoWoS)技术，预计CoWoS封装内总HBM带宽可提升34倍<sup>[9]</sup>。

3D路线通过3D封装垂直堆叠计算Die与DRAM Die。与2.5D封装相比，系统封装面积约减少一半。计算Die内通过定制的数据流计算架构、DRAM控制器及PHY等，实现计

表3 各DRAM-PIM内部带宽、算力对比

|                             | 内部带宽                           | 算力         | 容量或面积损失                                                | 收益                                 |
|-----------------------------|--------------------------------|------------|--------------------------------------------------------|------------------------------------|
| GDDR6-PIM <sup>[10]</sup>   | 16×外部带宽:512 GB/s               | 512 GFLOPS | 每个计算单元需0.19 mm <sup>2</sup> 换取32 GFLOPS算力,每Die共16个计算单元 | 执行MHA计算时,相比传统GPU系统,该方案计算延迟降至原来的2/5 |
| LPDDR5x-PIM <sup>[10]</sup> | 8×外部带宽:153.6 GB/s              | 307.2 GOPS | /                                                      | /                                  |
| HBM2-PIM <sup>[11]</sup>    | 4×外部带宽:<br>1.22 9 TB/s@300 MHz | 1.2 TFLOPS | 共8个Die,改造4个Die,每Die损失1/4容量                             | GEMV计算能效最高提升8.25倍                  |

GDDR: 图形双倍数据速率  
GEMV: 通用矩阵-向量乘法  
GFLOPS: 每秒十亿次浮点运算  
GOPS: 每秒十亿次整数运算

CPU: 图形处理器  
HBM: 高带宽存储器  
LPDDR: 低功耗双倍数据速率同步动态随机存储器  
MHA: 多头注意力

PIM: 存内处理  
TFLOPS: 每秒万亿次浮点运算

算单元对 DRAM Bank 的垂直并行访问，在单位面积内形成海量并行数据通道，大幅缩短 DRAM Die 的访问路径。与 2.5D 封装相比，在堆叠相同数量的 DRAM Die 条件下，带宽可提升约 10 倍，功耗降低至 60%<sup>[20]</sup>。该技术路径需解决堆叠架构带来的散热与供电压力，权衡封装工艺的良率与成本，并考虑 TSV 及堆叠层数引入的计算 Die 面积开销，在设计、制造、封装与测试全流程中均面临较高的系统复杂度。

由于上述系统工程挑战，3D DRAM PNM 预计将率先在边缘侧实现应用，并有望集成于中兴微电子边缘侧 AI 产品中。当前业界可堆叠 DRAM 层数为 1~4 层，单层面积约为 50~100 mm<sup>2</sup>，单层容量为 2~4 GB，堆叠后的总带宽可达 1~4 TB/s，动态功耗约为 0.6 pJ/bit（约为同代 LPDDR 功耗的五分之一）。由此可见，该技术路线的核心优势在于高带宽与低功耗，容量虽显著高于 SRAM，但仍受限于堆叠面积与层数，难以满足较大参数量模型的需求，因此需要结合 LPDDR 或 Flash 等存储介质进行系统级架构设计。

## 2.3 基于 Flash 的存算一体技术

### 2.3.1 Nor Flash CIM

基于 Flash 的模拟存内计算，存储密度高于 DRAM，器件成本低，但 Flash 读写速度差（几十到一百多 MB/s），耐久有限，只适合模型固定的小算力低功耗应用场景。例如知存科技 WTM2101 是基于 NOR Flash 的存算一体 AI 片上系统（SOC）芯片，算力仅 50 每秒十亿次整数运算（GOPS），面向智能语音和智能健康。

### 2.3.2 Nand Flash PNM

HBF 定位为介于 DRAM 与 Flash 之间的新型存储层级。根据闪存厂商（SanDisk）的产品规划，其第一代 HBF 产品

通过 3D 堆叠技术集成 16 个与非型闪存（NAND）Core Die，单颗容量可达 512 GB，两个 Stack 即可实现 TB 级以上容量，满足低成本大容量扩展需求。同时，通过底部 Base Die 对 NAND Core Die 的并行访问，HBF 可实现高达 1.6 TB/s 的带宽，性能与 HBM4 相当。然而，该技术仍面临写入寿命有限与访问延迟较高等短板，需在实际系统设计中予以权衡。

HBF 适合推理场景中只读或确定性访问模式的工作负载，但不适合训练等写密集型任务。HBF 的潜在应用方案如图 5 所示，主要包括以下 3 类：

1) 各种类型的处理器（xPU）直连 HBF：以 HBF 替代部分 HBM，在相同芯片边长下可实现 8~16 倍的读存储容量扩展。为避免因 HBF 延迟带来的性能损失，需在 HBF 内部对权重与 KV Cache 进行分区分配。

2) xPU - HBM - HBF 混合架构：该方案可减少 xPU 通过 CPU 访问内存或 SSD 所产生的延迟与功耗开销<sup>[21]</sup>。其中，HBM Base Die 负责处理 xPU 的访问请求路由与地址解码，并集成 SRAM Buffer 对 HBF 进行数据预取，以进一步隐藏访问延迟。仿真实验表明，在部署 LLaMa 3.1 405B 模型（上下文序列长度为 10M）的条件下，该方案相比纯 HBM 方案，吞吐量提升了 6.14 倍，能效提升了 2.69 倍。

3) 3D DRAM + HBF 架构：该方案无需 HBM，通过与 3D DRAM PNM 技术结合，利用 HBF 来突破 3D DRAM 容量不足的瓶颈。

## 2.4 基于新型存储的存算一体技术

新型存储主要包括 MRAM、ReRAM、PCRAM、FeRAM 等非易失性存储技术，如图 6（a）<sup>[22]</sup>所示，具备以下特点：

1) 容量密度高于 SRAM，但受限于良率与成本，难以扩展至 Gb 级以上规模；2) 写入延迟介于 SRAM 与 eFlash 之间，其中 MRAM 与 ReRAM 优于相变存储器（PCM）与铁电场效

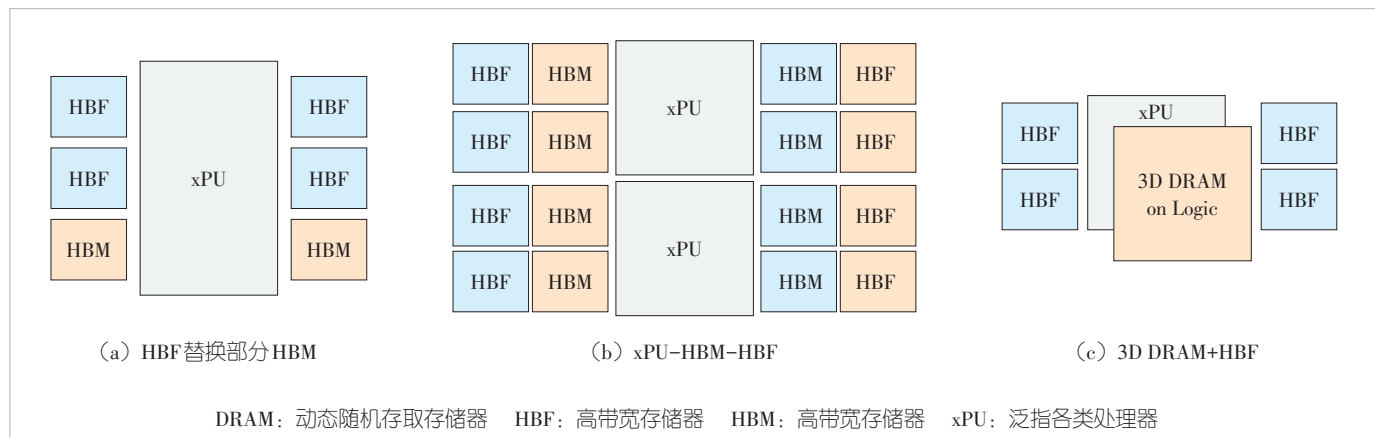
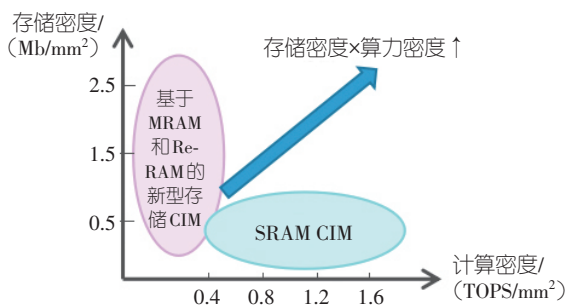


图5 HBF潜在应用方案

|      | SRAM               | MRAM               | ReRAM           | PCM             | FeFET            |
|------|--------------------|--------------------|-----------------|-----------------|------------------|
| 写延迟  | < 1 ns             | < 1 ns             | 100 ns          | 150 ns          | 175 ns           |
| 写功耗  | < 0.1 nJ           | < 1 nJ             | 2 nJ            | 6 nJ            | 0.1 nJ           |
| 静态功耗 | 高                  | 低                  | 低               | 低               | 低                |
| 容量密度 | 低                  | 中                  | 高               | 高               | 高                |
| 耐久   | > 10 <sup>16</sup> | > 10 <sup>15</sup> | 10 <sup>5</sup> | 10 <sup>7</sup> | 10 <sup>10</sup> |

(a) SRAM与新型存储介质的关键指标对比



(b) SRAM CIM与基于新型存储CIM的关键指标对比

CIM: 存内计算

MRAM: 磁阻随机存取存储器

ReRAM: 阻变随机存取存储器

TOPS: 每秒万亿次操作

FeFET: 铁电效应晶体管

PCM: 相变存储器

SRAM: 静态随机存取存储器

图6 SRAM与新型存储介质对比及其各自CIM技术<sup>[22]</sup>

应晶体管(FeFET); 3) 非易失性存储无须定期刷新, 静态功耗较低, 其中MRAM与ReRAM优于PCM; 4) MRAM与ReRAM相对PCM与FeFET, 与互补金属氧化物半导体(CMOS)工艺兼容性更佳。

目前, 业界将MRAM与ReRAM定位为替代eFlash(嵌入式闪存, 容量通常为数十Mb)的方案, 主要布局于车端微控制器(MCU)。台积电22 nm ReRAM节点已进入量产阶段, 12 nm ReRAM完成技术验证; 三星则计划于2026年实现8 nm MRAM的量产。

在CIM改造方面, MRAM和ReRAM相较于PCM和FeFET, 在写延迟与写功耗方面更具优势, 从而能够实现更高的计算能效<sup>[23]</sup>。其中: 1) MRAM与CMOS工艺兼容性更好, 适合用于数字CIM的改造; 2) ReRAM结构简单, 在外加电压下呈现连续的阻态变化, 能够支持多比特权重表示, 天然适用于模拟CIM。其工作原理基于欧姆定律与基尔霍夫定律: 输入电压与本地存储电导的乘积产生电流, 同一列上的电流汇聚相加, 完成乘加(MAC)运算<sup>[24]</sup>。

当前, 产业界与学术界正积极探索基于新型存储的存算一体技术。如图6(b)所示, 新型存储可与SRAM形成互补协同, 在容量与能效之间取得更好的平衡。例如, 台积电于2025年在Nature上发表了基于异构RRAM与SRAM的CIM AI边缘处理器设计, 在ResNet-20和MobileNet-V2上分别实现了40.91 TFLOPS/W与28.63 TFLOPS/W的能效。未来, 随着存储技术的持续演进, 有望出现同时具备非易失性、CMOS工艺兼容性、高容量密度的更理想存储介质。例如, 2026年IEEE国际固态电路会议(ISSCC)上中科院提出的基于电荷陷阱晶体管(CTT)的方案<sup>[22]</sup>, 即旨在兼顾高算力密度与高存储密度, 以适配更多应用场景需求。

### 3 存算一体技术趋势

如表4所示, 本文现对存算一体技术路径进行技术趋势总结<sup>[25]</sup>。其中, 带有颜色标注的为具备潜力的存算一体技术路径: 蓝色表示已商用技术, 黄色表示已有原型或样片的技术。

### 4 总结与展望

存算一体技术为突破云边端LLM与智能体(Agent)应用面临的“内存墙”瓶颈提供了多种技术路径。随着模型规模的持续增长与应用场景的日益复杂, 单点的硬件突破已难以满足产业需求, 亟需通过多层次深度协同, 实现系统级优化。

1) 存算一体技术协同。单一存储介质难以同时满足大容量、高速度、低功耗及低成本等多重极限需求, 因此, 基于多种存储介质协同组合的存算一体架构正成为重要发展趋势。例如, 在同一系统中可采用“SRAM CIM + 3D DRAM”或“SRAM CIM + 新型存储CIM”的异构存算架构, 利用SRAM处理高频访问的局部动态数据, 而将静态数据存储于大容量DRAM或高密度新型存储介质中, 实现优势互补, 从而突破单一介质的物理性能天花板。

2) 芯片与硬件基础设施系统协同。在芯片层面, 存算一体技术与先进封装、Chiplet等技术的结合是突破单芯片容量与带宽限制的关键路径; 在硬件基础设施层面, 存算一体芯片需与计算集群、算力网络、散热及供电等基础设施协同优化, 共同决定最终产品的能效表现。未来竞争将更多体现于系统级协同设计能力。

3) 软硬件协同。成熟的软件生态是创新架构应用推广的重要保障, 需构建涵盖算法层、系统层与芯片层的完整优

表4 存算一体技术趋势总结

| 存储介质              | 技术路径                           | 产品化进展                                    | 适用场景及技术趋势                                                                                                                      |
|-------------------|--------------------------------|------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| SRAM              | SRAM CIM                       | d-Matrix Corsair Chiplet、后摩智能 M50        | • 适合边侧和端侧的中小规模模型,尤其是对功耗敏感的场景<br>• 受限于 SRAM 容量,当前需要结合片外存储,导致 SRAM CIM 收益稀释;未来技术趋势为结合 DRAM PNM 技术,扩展容量的同时保持存算一体收益,已体现在此类产品下一代规划中 |
|                   | SRAM PIM                       | Groq 3 LPU、Sambanova SN50、Cerebras WSE-3 | • 适合云侧和边侧场景,尤其是对时延敏感的场景<br>• 受限于 SRAM 容量,往往多 Die、多芯片、多机柜规模部署,对互联带宽要求高;由于简化了硬件控制逻辑,软件编译和调度能力更加重要;技术趋势为整机软硬件系统深度协同优化             |
|                   | SRAM PNM                       | AMD MI300/MI400                          | • 适合云侧和边侧场景<br>• 承受了 3D 封装的代价和挑战,但相比 DRAM PNM,SRAM PNM 并不能从根本上解决 AI 存储容量限制                                                     |
| DRAM              | DRAM PIM                       | 三星 HBM PIM、SK 海力士 GDDR PIM 和 LPDDR PIM   | • 根据不同存储类型应用于云边缘场景<br>• 受限于算力主要用于 GEMV 计算;由存储厂商对存储颗粒产品内部进行改造                                                                   |
|                   | DRAM PNM: 2.5D HBM 定制 Base Die | 自 HBM4 起引入定制 Base Die                    | • 适合云侧和边侧场景,是实现大容量高带宽的主要路径<br>• 是 2026 年之后提升 HBM 带宽的重要驱动技术;需要设计厂商、晶圆代工厂商、存储厂商、封装厂商等协同设计                                        |
|                   | DRAM PNM: 3D DRAM              | d-Matrix 和后摩智能下一代产品、瑞芯微等国产 AI 协处理器       | • 适合云边缘各类场景,是实现大容量、高带宽、低功耗的主要路径<br>• 对于 3D 封装技术和复杂系统协同设计要求较高,预计边缘侧落地早于云侧落地                                                     |
| Flash             | Nor Flash CIM                  | 知存 WTM2101                               | • 只适合模型固定的端侧小算力、低功耗场景                                                                                                          |
|                   | Nand Flash PNM: HBF            | 闪迪和海力士战略合作等                              | • 适合云边缘各类场景的静态数据存储,构建介于 HBM 与传统 SSD 之间的新型存储层级<br>• 闪迪路线图显示,HBF 将沿着容量提升、带宽增长、功耗优化的方向持续迭代                                        |
| 新型存储              | 新型存储 CIM                       | 亿铸科技 ReRAM-CIM 原型,北京大学、香港科技大学等高校研究       | • 适合端侧小算力、功耗敏感场景<br>• 当前器件良率和成本面临挑战,容量扩展受限;随着材料和工艺的技术突破,有望发挥革命性作用                                                              |
| CIM: 存内计算         |                                | HBF: 高带宽闪存                               | PIM: 存内处理                                                                                                                      |
| DRAM: 动态随机存取存储器   |                                | HBM: 高带宽存储器                              | PNM: 近存处理                                                                                                                      |
| GDDR: 图形双倍数据速率存储器 |                                | LPDDR: 低功耗双倍数据速率同步动态随机存储器                | SRAM: 静态随机存取存储器                                                                                                                |
| GEMV: 通用矩阵-向量乘法   |                                | LPU: 语言处理单元                              | SSD: 固态硬盘                                                                                                                      |

化空间。具体包括：面向存算一体架构的模型结构设计、量化与剪枝等算法优化技术；面向软件栈的编译优化、运行时调度及内存管理机制；建立统一的编程模型与硬件抽象层，使存算一体硬件能够无缝接入主流 AI 框架，实现硬件透明、软件友好的开发体验。

4) 产业链生态协同。存算一体技术是一项涉及新材料与器件研发、芯片设计、晶圆制造、先进封装、系统集成及应用部署的系统工程，产业链各环节的紧密协作是推动该技术从实验室走向大规模落地的关键。

中兴通讯与中兴微电子以芯片与硬件基础设施层为核心发力点，以系统总拥有成本（TCO）最优的全栈算力解决方案为目标，通过产学研协同、创新架构研究及产业合作伙伴价值共创等方式，积极探索基于 SRAM、DRAM、Flash 及新型存储的多元化存算一体技术与产品应用。未来，通过底层异构存储介质的技术协同、系统级芯片与硬件基础设施的协同、全栈式软硬件协同以及全景式产业链生态协同，存算一

体有望成为重塑计算体系架构的关键驱动力，为后摩尔时代人工智能的广泛普及与应用提供坚实基础。

致谢

感谢深圳市中兴微电子有限公司陈桂显专家对本论文做出的贡献！

参考文献

[1] Gholami A, Yao Z, Kim S, et al. AI and memory wall [EB/OL]. (2024-03-25) [2026-06-10]. <https://ieeexplore.ieee.org/abstract/document/10477550>

[2] 韩炳涛, 刘涛. 大模型关键技术与应用 [J]. 中兴通讯技术, 2024, 30 (2): 76-88. DOI: 10.12142/ZTETJ.202402012

[3] Wolters C, Yang X X, Schlichtmann U, et al. Memory is all you need: an overview of compute-in-memory architectures for accelerating large language model inference [PP/OL]. arXiv(2024-06-12)[2026-06-13]. <https://arxiv.org/abs/2406.08413>

[4] Horowitz M. 1.1 computing's energy problem (and what we can do about it) [C]//Proceedings of IEEE International Solid-State

- Circuits Conference Digest of Technical Papers (ISSCC). IEEE, 2014: 10–14. DOI: 10.1109/isscc.2014.6757323
- [5] 包义明, 林阳, 屠要峰. 人工智能技术与应用前沿 [J]. 中兴通讯技术, 2025, 31(4): 64–74. DOI: 10.12142/ZTETJ.202504010
- [6] 何斯琪, 穆琛, 陈迟晓. 基于存算一体集成芯片的大语言模型专用硬件架构 [J]. 中兴通讯技术, 2024, 30(2): 37–42. DOI: 10.12142/ZTETJ.202402006
- [7] Chih Y D, Lee P H, Fujiwara H, et al. 16.4 an 89TOPS/W and 16.3TOPS/mm<sup>2</sup> all-digital SRAM-based full-precision compute-in memory macro in 22nm for machine-learning edge applications [C]//Proceedings of IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2021: 252–254. DOI: 10.1109/isscc42613.2021.9365766
- [8] Prabhakar R, Sivaramakrishnan R, Gandhi D, et al. SambaNova SN40L: scaling the AI memory wall with dataflow and composition of experts [PP/OL]. arXiv(2024–11–05) [2026–06–11]. <https://arxiv.org/abs/2405.07518>
- [9] NVIDIA Corporation. A decade of AI infrastructure innovation: from NVIDIA DGX-1 to NVIDIA vera rubin [EB/OL]. [2026–03–06] [2026–06–11]. <https://www.nvidia.com/en-us/on-demand/session/gtc26-d8006/>
- [10] Tom's Hardware. AMD shares new second-gen 3D v-cache chiplet details: up to 25 TB/s [EB/OL]. (2023–06–01)[2026–06–11]. <https://www.tomshardware.com/news/amd-shares-new-second-gen-3d-v-cache-chiplet-details-up-to-25-tbs>
- [11] Kim G, Kim J, Kim N, et al. SK hynix AI-specific computing memory solution: from AiM device to heterogeneous AiMX-xPU system for comprehensive LLM inference [C]//Proceedings of IEEE Hot Chips 36 Symposium (HCS). IEEE, 2024: 1–26. DOI: 10.1109/hcs61935.2024.10664793
- [12] Kim K. Advancements in co-packing and co-integration technologies for high-performance memory system [C]//2024 IEEE VLSI Symposium on Technology and Circuits. IEEE, 2024
- [13] Cao Y, Jiang J H, Jiang H J, et al. A 1.2GHz 12.77GB/s/mm<sup>2</sup> 3D two-DRAM-one-logic process-near-memory chip for edge LLM applications [C]//Proceedings of IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2026: 524–526. DOI: 10.1109/ISSCC49663.2026.11409184
- [14] SanDisk. High Bandwidth Flash (HBF) tutorial [R]. 2026 IEEE International Memory Workshop (IMS), 2026
- [15] Fujiwara H, Mori H, Zhao W C, et al. A 5-nm 254-TOPS/W 221-TOPS/mm<sup>2</sup> fully-digital computing-in-memory macro supporting wide-range dynamic-voltage-frequency scaling and simultaneous MAC and write operations [C]//Proceedings of IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2022: 1–3. DOI: 10.1109/ISSCC42614.2022.9731754
- [16] Keller B, Venkatesan R, Dai S, et al. A 95.6-TOPS/W deep learning inference accelerator with per-vector scaled 4-bit quantization in 5 nm [J]. IEEE journal of solid-state circuits, 2023, 58(4): 1129–1141. DOI: 10.1109/JSSC.2023.3234893
- [17] Kaist Teralab. HBM roadmap ver 1.7 workshop: overview of next generation HBM architectures [EB/OL]. [2026–06–13]. <https://tera.kaist.ac.kr/home>
- [18] Marvell Semiconductor, Inc. Marvell Custom AI Investor Event 2025 [EB/OL]. [2026–06–02]. <https://www.marvell.com/content/dam/marvell/en/company/assets/marvell-custom-ai-investor-event-2025.pdf>
- [19] TSMC. TSMC Technology symposium press briefing [EB/OL]. (2026–04–26) [2026–06–02]. <https://www.eet-china.com/mp/a490419.html>
- [20] Lim D. Packaging-aware design and optimization of HBM power delivery in AI platforms [R]. ISSCC 2026 Forum, 2026
- [21] Ha M, Kim E, Kim H. H3: hybrid architecture using high bandwidth memory and high bandwidth flash for cost-efficient LLM inference [J]. IEEE computer architecture letters, 2026, 25 (1): 49–52. DOI: 10.1109/LCA.2026.3660969
- [22] Shen J Z, Zhou Z D, Zha W F, et al. A 12nm 4Mb 104.56-to-137.75TFLOPS/W charge-trap transistor-based computing-in-memory macro using analog-predict-digital-compute for AI edge devices [C]//Proceedings of IEEE International Solid-State Circuits Conference (ISSCC). IEEE, 2026: 514–516. DOI: 10.1109/ISSCC49663.2026.11409272
- [23] Khwa W S, Wen T H, Hsu H H, et al. A mixed-precision memristor and SRAM compute-in-memory AI processor [J]. Nature, 2025, 639(8055): 617–623. DOI: 10.1038/s41586-025-08639-2
- [24] 朱炫鹏, 姚海东, 刘隽, 等. 大语言模型算法演进综述 [J]. 中兴通讯技术, 2024, 30(2): 9–20. DOI: 10.12142/ZTETJ.202402003
- [25] 廖俊淇, 魏昕, 周亮. 人工智能驱动的跨模态语义通信系统 [J]. 中兴通讯技术, 2024, 30(S1): 33–39. DOI: 10.12142/ZTETJ.2024S1005

## 作者简介



**刘新阳**, 深圳市中兴微电子技术有限公司副总经理、中国通信学会集成电路专业委员会副主任委员、中兴通讯股份有限公司多媒体与移动网络国家重点实验室IC研发中心主任; 负责中兴微电子芯片战略与技术规划、运营工作等, 拥有20余年通信集成电路规划、研发、经营管理经验。



**李婷宇**, 深圳市中兴微电子技术有限公司战略规划工程师; 主要从事人工智能芯片技术研究与规划工作, 包括人工智能芯片架构技术、存储技术、互联技术和先进封装技术等。



**巨鸣**, 深圳市中兴微电子技术有限公司战略规划工程师; 主要从事人工智能芯片技术研究与规划工作, 包括人工智能芯片存储技术、存算一体技术、新型存储技术等。