

新架构大模型推理的存算特征演变与近存计算适配技术研究



Evolution of Memory-Compute Characteristics in New-Architecture LLM Inference and Near-Memory Computing Adaptation Technologies

庄风云/Zhuang Fengyun, 孙菲/Sun Fei, 钱岭/Qian Ling

(中移(苏州)软件技术有限公司, 中国 苏州 215004)
(China Mobile (Suzhou) Software Technology Co., Ltd., Suzhou 215004, China)

DOI: 10.12142/ZTETJ.202604005

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260827.1355.002>

网络出版日期: 2026-08-27

收稿日期: 2026-05-18

摘要: 多头潜在注意力 (MLA) 与混合专家 (MoE) 正在重塑大模型推理的存算特征: MLA 通过低秩键值 (KV) 联合压缩, 使权重吸收模式下的注意力算术强度由传统多头注意力 (MHA) 解码的约 1 Op/B 提升至数百 Op/B; MoE 则因实际批量下的专家权重访问瓶颈, 成为近存计算的新核心场景。系统梳理解码阶段算术强度的解析模型以及 “MHA→分组查询注意力 (GQA)→MLA” 的存算特征演变规律, 对比分析 MLA 双执行模式的算术强度差异与现有存内处理 (PIM) 设备的算力密度约束, 按临界批量模型对 4 类主流 AI 加速器进行解析估算与横向研判, 并结合 FP8/INT4 等低精度计算特征, 提出面向 MLA+MoE 组合架构的分层近存计算适配方向并给出框架性建议。

关键词: 多头潜在注意力; 混合专家模型; 存算一体; 近存计算; 算术强度

Abstract: Multi-head latent attention (MLA) and mixture of experts (MoE) are reshaping the memory-compute characteristics of large language model (LLM) inference. Through low-rank joint compression of key-value (KV) representations, MLA raises the arithmetic intensity of attention under the weight-absorption mode from approximately 1 Op/B for conventional multi-head attention (MHA) decoding to several hundred Op/B, whereas MoE introduces an expert-weight memory-access bottleneck at practical batch sizes and becomes a new target for near-memory computing (NMC). This paper systematically reviews analytical models of arithmetic intensity in the decoding stage and investigates the evolution of memory-compute characteristics along the path of MHA → grouped-query attention (GQA) → MLA. It further compares the arithmetic-intensity characteristics of the two MLA execution modes and analyzes them in conjunction with the compute-density constraints of existing processing-in-memory (PIM) devices. Based on a critical-batch-size model, analytical estimations and cross-platform evaluations are conducted on four representative AI accelerators. Building upon these observations, the paper incorporates the characteristics of FP8/INT4 low-precision computation, proposes a hierarchical near-memory computing adaptation framework for combined MLA+MoE architectures, and provides corresponding architectural recommendations.

Keywords: multi-head latent attention; mixture of experts; processing-in-memory; near-memory computing; arithmetic intensity

引用格式: 庄风云, 孙菲, 钱岭. 新架构大模型推理的存算特征演变与近存计算适配技术研究 [J]. 中兴通讯技术, 2026, 32(4): 23-31. DOI: 10.12142/ZTETJ.202604005

Citation: Zhuang F Y, Sun F, Qian L. Evolution of memory-compute characteristics in new-architecture LLM inference and near-memory computing adaptation technologies [J]. ZTE technology journal, 2026, 32(4): 23-31. DOI: 10.12142/ZTETJ.202604005

大语言模型的大规模部署使推理效率问题日益凸显。冯·诺依曼架构中, 处理器与存储器间的数据搬运已成为性能瓶颈, 即“存储墙”问题。大模型自回归解码阶段的算术强度 (AI) 仅约 1 Op/B, 较主流图形处理器 (GPU) 数十至数百 Op/B 的脊点低 2 个数量级以上, 导致计算资源严重闲置^[1]。存内处理 (PIM) 技术通过将计算单元部署于存储器内部或近存储位置, 可将有效内存带宽提升数倍至数十倍, 被视为突破存储墙的关键路径; 三星高带宽存储器

(HBM) -PIM^[2]与 SK Hynix AiM 等均以注意力加速为核心目标场景。与此同时, 以 OpenAI o1/o3 和 DeepSeek-R1 为代表的推理时扩展 (Test-Time Scaling) 范式通过延长思维链显著拉长解码序列, 使本就访存受限的解码阶段在总推理开销中的占比进一步扩大, 存储墙矛盾愈发尖锐。

然而, 以 DeepSeek-V2 和 DeepSeek-V3 为代表的新一代大模型架构包含两项关键创新。其一, 多头潜在注意力 (MLA) 通过低秩 KV 联合压缩, 将键值缓存 (KV Cache)

规模压缩至同等参数规模多头注意力（MHA）的约 1/60，并在权重吸收执行模式下使注意力层 AI 由约 1 Op/B 提升至数百 Op/B（提升约 2 个数量级），达到甚至超越主流 GPU 脊点^[3-4]。其二，混合专家（MoE）通过稀疏条件激活实现参数规模扩展，但在实际推理批次下引入了新的专家权重访存瓶颈。现有以 MHA 为设计基准的 PIM 方案面临根本性的适配挑战。

现有工作对标准 Transformer 负载在 PIM 系统上的映射已进行较系统的研究^[2,5]，但针对 MLA 双执行模式、长上下文压缩稀疏注意力、低精度 MoE 推理以及多级异构存储组合代价的统一分析仍不充分。本文系统梳理解码阶段算术强度的解析模型及“MHA→分组查询注意力（GQA）→MLA”的存算特征演变规律；对比分析 MLA 双执行模式对 PIM 适配的差异化影响，并指出现有 PIM 设备的算力密度约束；归纳 MoE 专家层临界批量的估算方法，在 4 类主流 AI 加速器的公开规格参数下进行解析对比，分析其作为近存计算核心目标的可行性；最后，提出面向“MLA+MoE”组合架构的分层近存计算适配框架。

1 大模型推理存算特征基础

1.1 算术强度与 Roofline 模型

AI 定义为单位内存访问量对应的运算量。假设某个算子的总运算量为 W （单位：Op）、片外访存量为 Q （单位：B），则

$$AI = \frac{W}{Q} \quad (1)。$$

Roofline 模型^[1]是评估模型在特定硬件上潜在性能的有效理论框架：假设设备峰值算力为 π （Op/s）、峰值内存带宽为 β （B/s），则算子可达性能：

$$P_{\text{attain}} = \min(\pi, \beta \cdot AI) \quad (2)。$$

两段上界的交点即设备“脊点”：

$$RP = \frac{\pi}{\beta} \quad (3)。$$

算子在硬件设备上的执行过程如图 1 所示，数据从片外存储器（双倍数据速率存储器或 HBM）传输至片上缓冲区，再由片上处理单元完成计算，最终将结果写回内存。因此，性能评估需要同时考虑内存访问能力和处理单元的计算能力：当 $AI < RP$ （脊点）时，即算子的访存流量较大而计算操作数相对较少，则该算子为访存密集型，性能受带宽约束；当 $AI = RP$ 时处于性能拐点；当 $AI > RP$ 时则为计算密集

型，性能受算力约束。主流 AI 芯片的脊点参数如表 1 所示。

近几年，存储技术快速演进。HBM4 接口宽度由 1 024 位扩展至 2 048 位；2026 年公开商用产品已达 11.7 Gbit/s 稳定速率（最高 13 Gbit/s），单堆栈峰值带宽达 3.3 TB/s，12 层容量覆盖 24~36 GB，并通过低电压硅通孔（TSV）和供电网络优化提升能效^[6]。带宽代际跃升会降低同一精度下的 GPU 脊点（即进入计算密集区的门槛降低），并缓解访存压力，但 HBM 的容量、封装成本、逻辑底座面积及热设计约束仍然存在；同时，低精度 Tensor Core 算力的增长可重新抬高精度匹配的脊点。因此，“带宽增长即可消除存储墙”的论断并不成立，仍需结合算子精度和实际数据流分析。

1.2 低精度计算与精度感知 Roofline 修正

低精度量化会同时改变片外存储器访问字节总量与硬件峰值运算吞吐。设量化后权重、KV Cache、激活值的单元有效字节数分别为 p_w 、 p_{kv} 、 p_a ，对应元素数分别为 N_w 、 N_{kv} 和 N_a ，量化缩放因子、零点及对齐填充等引入的元数据访存流量为 Q_{meta} ；在精度 q 下，硬件峰值算力和脊点分别为 π_q 与 RP_q ，则总片外访存字节 Q_q 、算术强度 AI_q 、硬件脊点 RP_q 满足：

$$Q_q = p_w N_w + p_{kv} N_{kv} + p_a N_a + Q_{\text{meta}}, \quad AI_q = \frac{W}{Q_q}, \quad RP_q = \frac{\pi_q}{\beta} \quad (4)。$$

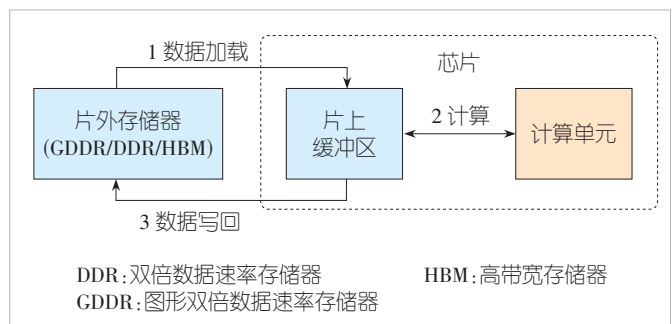


图1 算子执行过程

表1 主流AI加速器脊点

AI加速器	BF16/FP16理论算力/ TFLOPS	内存带宽/ (TB·s ⁻¹)	脊点/ (Op·B ⁻¹)
A100 SXM	312	2.00	156
H100 SXM	989	3.35	295
H20	148	4.00	37
Ascend 910B2	376	1.60	235

AI: 算术强度 FP: 浮点格式
BF: 脑浮点格式 TFLOPS: 每秒万亿次浮点运算

BF16/FP16、FP8/INT8 和 INT4/FP4 的理想单元数据载荷分别为 2 B、1 B 和 0.5 B。以 H100 SXM 为例，在不计结构化稀疏时，BF16 与 FP8 峰值算力分别约为 989 TFLOPS 和 1 979 TFLOPS；在 HBM 带宽 3.35 TB/s 条件下，硬件脊点由 BF16 下 295 Op/B 提升至 FP8 下 591 Op/B。当计算峰值与数据压缩比同步提高时，算术强度与硬件脊点近似同比右移，访存/计算属性未必改变。

Hopper 架构并不具备 INT4 原生矩阵计算能力，INT4 部署多采用仅权重量化方案，权重在计算单元内完成反量化后参与矩阵运算。TensorRT-LLM、NeMo 已提供 INT4-AWQ、FP8 等完整部署路径，但缩放因子读取、数据格式解包、离群值处理以及内核实现差异，都会削弱量化带来的理论性能收益。AMD MI300X 等主流 AI 加速器均原生支持 FP8 与 INT8 运算，低精度已经成为现代加速器的通用系统设计变量。

1.3 MHA 解码阶段的存算特征

标准 MHA 的自回归解码按照逐 Token 生成，假设缓存序列长度为 L 、注意力头数为 n_h 、头维度为 d_h 、模型层数为 l 、每元素字节数为 p ，则单请求每 Token 的 KV Cache 占用为：

$$m_{KV} = 2 \cdot l \cdot n_h \cdot d_h \cdot p \quad (5),$$

其中， n_{KV} 为 KV 头数，MHA 下 $n_{KV} = n_h$ ，FP16/BF16 精度下 $p=2$ ，FP8 下 $p=1$ ；总占用随批量 B 与序列长度线性增长。解码时每个请求的 KV Cache 为该请求私有，在不存在共享前缀缓存复用的情况下，批量维度难以摊薄 KV Cache 的读取开销；而前馈网络（FFN）权重为模型静态参数，可在批量样本间复用。每生成一个 Token，单层注意力核心运算（ QK^T 与 SV 操作）的运算量约为 $4n_h L d_h$ ，而须从显存读取的 KV 缓存约为 $2n_h L d_h \cdot p$ ，因此 MHA 的算术强度为：

$$AI_{MHA} \approx \frac{4n_h L d_h}{2n_h L d_h \cdot p} = \frac{2}{p} \quad (6).$$

也就是说，FP16/BF16 精度下 $AI_{MHA} \approx 1 \text{ Op/B}$ ，且与批量和上下文长度基本无关，与主流 AI 芯片脊点相差 2 个数量级以上，计算单元绝大部分时间在等待数据。这是三星 HBM-PIM^[5]、SK Hynix AiM 以及 LoL-PIM^[7] 等系统选择以注意力加速为核心设计目标的根本动机。

1.4 GQA 的改善与局限

分组查询注意力（GQA）^[8] 令每 G 个查询（Query）头共享同一组 KV 头，即 $n_{KV} = n_h / G$ ，KV Cache 规模降至 MHA 的 $1/G$ 。由于运算量不变而访存量按 G 倍缩减，解码的算术强

度相应提升为：

$$AI_{GQA} \approx \frac{2G}{p} \quad (7).$$

当 $G=8$ 且处于 FP16 精度时， $AI_{GQA} \approx 8 \text{ Op/B}$ 。以 LLaMA-3-70B（ $G=8$ ）为例，FP16 精度下每 Token 的 KV 存储量由 MHA 的约 2.6 MB 降至约 320 KB，但 AI 仍处于访存密集区，尚未突破存储墙约束的本质。

2 MLA 存算特征深度分析

2.1 MLA 架构原理与 KV Cache 压缩效果

MLA 由 DeepSeek-V2^[3] 首次提出，核心思想是通过下投影矩阵 W_{DKV} 将所有注意力头的 KV 联合压缩为低维潜变量 C_{KV} ，压缩维度 $d_c \ll n_h \cdot d_h$ ，推理时仅缓存 C_{KV} 及解耦的旋转位置编码分量（维度为 d_r ），按需通过上投影矩阵 W_{UK} 、 W_{UV} 还原完整 KV 矩阵。其每 Token 缓存占用为：

$$m_{MLA} = l \cdot (d_c + d_r) \cdot p \quad (8).$$

以 DeepSeek-V3 为例，其 $l=61$ ， $d_c=512$ ， $d_r=64$ ，在 BF16 精度下，每 Token 的 KV Cache 仅约 70 KB^[9]；按 DeepSeek-V3 等效配置估算（ $l=61$ ， $n_h=128$ ， $d_h=128$ ），MHA 每 Token 的 KV Cache 约为 4 MB，压缩比高达约 60 倍；采用 GQA（ $G=8$ ），每 Token 的 KV Cache 约为 500 KB，压缩比约 7 倍。表 2 从 KV Cache 规模与算术强度角度对各注意力机制进行横向比较。

2.2 MLA 双执行模式的算术强度量化

MLA 存在两种代数等价的执行路径，具有截然不同的存算特性，如图 2 所示。

Absorb（吸收）模式：利用矩阵结合律将上投影矩阵 W_{UK} 、 W_{UV} 分别与查询投影、输出投影预先融合（“矩阵吸收”），推理时直接在压缩潜变量空间内执行注意力运算，无须展开完整 KV。此时全部 n_h 个查询头共享读取同一份潜变量缓存 C_{KV} ，且该缓存同时承担 K 与 V 的双重角色，单层

表 2 不同注意力机制 KV Cache 规模与算术强度对比

注意力机制	单位 Token KV Cache	解码 AI/ (Op·B ⁻¹)	近存计算适用性
MHA	约 4 MB	约 1	高
GQA($G=8$)	约 500 KB	约 8	较高
MLA(Naive 模式)	约 70 KB	1 ~ 8	中
MLA(Absorb 模式)	约 70 KB	256 ~ 512	低

AI: 算术强度 MHA: 多头注意力
GQA: 分组查询注意力 MLA: 多头潜在注意力

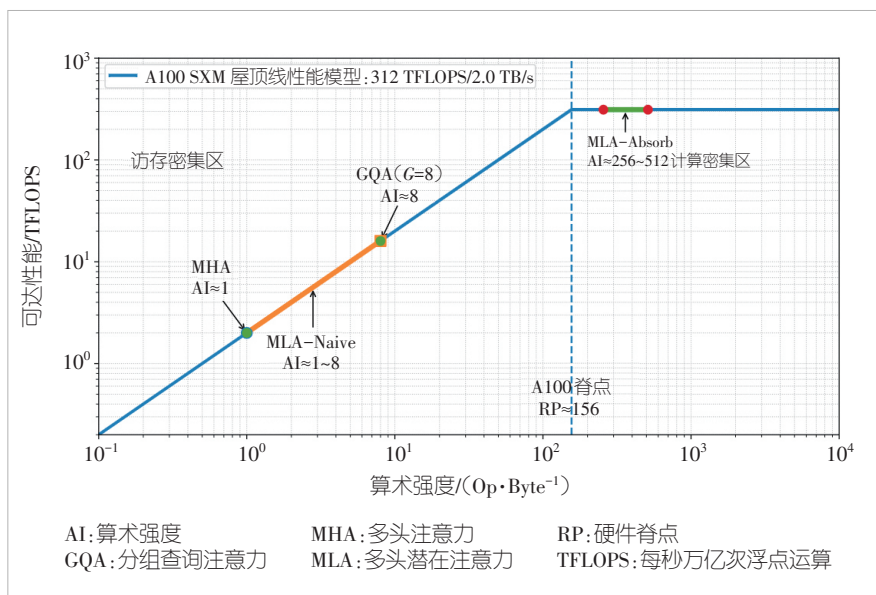


图2 不同注意力机制在Roofline模型中的算术强度分布 (A100 SXM)

核心运算量约为 $4n_h L_d c$ ，而潜变量缓存读取量仅为 $L_d p$ ，故：

$$AI_{\text{Absorb}} \approx \frac{4n_h L_d c}{L_d \cdot p} = \frac{4n_h}{p} \quad (9)$$

以 DeepSeek-V3 ($n_h=128$) 代入，在 BF16 精度下， $AI_{\text{Absorb}} \approx 256 \text{ Op/B}$ ，FP8 精度下约为 512 Op/B ，已达到或超越 A100 乃至 H100 的脊点。注意力层由此脱离深度访存密集区，进入计算密集区域或其边缘^[4]。

Naive (原始) 模式：运行时从 C_{KV} 在线展开完整 KV 后执行标准注意力运算。其访存特征近似于 GQA，小批量 ($B < 8$) 场景下 AI 处于约 $1 \sim 8 \text{ Op/B}$ 区间，呈访存密集特性，且随批量增大 AI 逐渐升高。

需要特别指出的是，Absorb 模式并非所有场景下的默认选项，在实际部署中两种模式常混合使用。TyphoonMLA^[10] 的工程实践表明，对于共享前缀 (Shared Prefix) 段，已有历史 Token 的 KV 可直接从 KV Cache 复用，此段采用 Naive 模式无须重新展开上投影矩阵，仅对新增 Token 段切换至 Absorb 模式以获得更高吞吐。此外，受出口管制影响，部分场景只能使用 H20 等低算力高带宽卡，其 MLA-Absorb 模式的绝对吞吐上限受算力瓶颈制约，而 Naive 模式配合高内存带宽的组合可能更具竞争力。因此，模式选择本质上是一个依赖硬件算力/带宽比的动态决策。

2.3 MLA 对近存计算设计的新挑战

MLA-Absorb 模式使注意力层脱离访存密集区，从根本

上动摇了 PIM 加速注意力的传统价值主张。与此同时，MLA 还带来 4 类新的硬件设计挑战：

1) PIM 算力密度制约 Absorb 模式执行：现有三星 HBM-PIM 和 SK Hynix AiM 等存内计算设备，其片内算力密度通常为 $1 \sim 10 \text{ TFLOPS}$ 量级，而 MLA-Absorb 模式要求执行规模较大的矩阵乘，例如，其在 A100 上饱和运行对应的算力需求约 312 TFLOPS ，与 PIM 片内算力相差约 2 个数量级。即使内存带宽充裕，PIM 设备也无法充分驱动 Absorb 模式所需的计算量。PIM 对 MLA-Absorb 注意力的加速价值从根本上受到算力约束，而非带宽约束。

2) 张量并行 (TP) 分片瓶颈： C_{KV} 为单个共享潜变量头，无法按头维度进行

TP 切分。多卡推理时每张卡必须冗余加载完整 KV Cache，且带宽开销随设备数量线性累积^[11]。

3) 混合执行的灵活性需求：在实际部署中，可能采用灵活的混合执行模式。例如，共享前缀段采用 Naive 模式复用已缓存 KV，新增词元段采用 Absorb 模式以获更高吞吐。混合执行模式下，近存硬件需支持两种模式的动态切换，对可编程性提出更高要求。

4) 新变种持续演进：分组查询潜在注意力 (GQLA)^[12] 将 MLA 与 GQA 融合，同一套训练权重可同时支持 MQA-Absorb 和 GQA 展开两类解码路径，分别适配 H100 等高计算/带宽比硬件和 H20 等低算力、高带宽硬件；DeepSeek-V3.2 则引入 DSA 稀疏注意力机制^[13]，通过设计闪电索引器模块和细粒度 Token 选择机制，将核心注意力复杂度由 $O(L^2)$ 降至 $O(Lk) (k \ll L)$ 。这些进展表明，注意力机制正沿“潜空间化压缩”和“细粒度稀疏化”方向演进，并逐步走向由硬件算力带宽比、上下文长度和推理批量共同决定执行路径的硬件自适应形态。

3 MoE 存算特征分析

3.1 MoE 推理机制与内存规模

MoE 通过稀疏条件激活实现参数规模扩展，每个输入 Token 经门控网络选择 Top-K 个专家子网络参与计算，其余专家权重在当前步骤不被访问。

以 DeepSeek-V3 为例，其总参数量为 6 710 亿，每 Token 激活参数约 370 亿；除前 3 个 dense FFN 层外，每个 MoE 层包

含 1 个共享专家与 256 个路由专家，每 Token 激活 8 个路由专家。虽然 MoE 降低了单 Token 计算量，但低时延推理通常仍要求专家权重常驻或快速可达，在 BF16 精度下约占 1.3 TB，远超当前 GPU 单机 HBM 容量上限（例如单机 8 卡 H100 80G 总容量 640 GB）^[9]，形成超出单机容量的结构性挑战。

3.2 MoE 专家访问的访存特征

1) 长尾分布规律：专家访问频率高度不均匀，遵循 Zipf 型长尾分布——约前 20% 的“热专家”承担约 80% 的激活请求，其余为低频“冷专家”^[14]。这一规律为热冷分层的近存计算策略提供了量化依据。

2) 批量大小与存算特性的动态关系：MoE 专家层本质上是 FFN，其专家权重可在多个 Token 间复用，因此算术强度随有效批量增大而提升。设总专家数为 n_e 、每 Token 激活专家数为 n_k ，批量 B 下平均每个专家处理 Bn_k/n_e 个 Token；专家权重读取量恒定而有效运算量随路由 Token 数线性增长，故专家层算术强度近似为：

$$AI_{MoE} \approx \frac{2Bn_k}{p \cdot n_e} \quad (10)。$$

令 $AI_{MoE} = RP_{acc}$ （AI 加速器的脊点），即可得使 MoE 专家层由访存密集转为计算密集的临界有效批量：

$$B_{th} = \frac{p}{2} \cdot RP_{acc} \cdot \frac{n_e}{n_k} \quad (11)。$$

在 BF16/FP16 精度下 $p=2$ ，公式（11）可简化为 $B_{th} = RP_{acc} \cdot \frac{n_e}{n_k}$ 。以 A100（ $RP \approx 156$ Op/B）和 DeepSeek-V3（ $n_e = 256$ ， $n_k = 8$ ， $n_e/n_k = 32$ ）代入，则 $B_{th} \approx 4\,896$ 。这意味着在绝大多数实际推理批量（ $B \leq 512$ ）下，MoE 专家层均持续处于访存密集状态，是近存计算的天然目标。

表 3 基于各 AI 加速器的公开规格，对 A100、H100、H20 及 Ascend 910B2 共 4 种主流 AI 加速器的 B_{th} 进行解析估算，其 B_{th} 均远高于典型推理批量上限。因此，MoE 专家层的访存密集特征并不会因加速器差异而消失，但其程度会随

硬件 Roofline 脊点、批量大小和路由分布而变化。

3.3 MLA+MoE 组合架构的存算解析

DeepSeek-V2/V3 将 MLA 与 MoE 结合，形成异构的存算特征分布。在 MLA-Absorb 执行模式下，注意力层的算术强度显著提升，由传统意义上的访存密集逐步转向脊点附近或计算密集区域，对 PIM 的带宽缓解需求降低；与此同时，当前 PIM 架构受限于算力密度、数据复用能力和矩阵计算效率，也难以高效承载高算术强度矩阵乘。

相比之下，MoE 专家层在典型批量下仍需频繁读取大规模专家权重，权重复用不足，持续呈现访存密集特征，因而成为新的近存计算核心目标。由此可见，大模型推理的存算瓶颈重心正在从传统注意力层向 MoE 专家层发生结构性迁移。

4 面向 MLA/MoE 的近存计算适配方向

4.1 现有 PIM 方案的局限性

当前主流 PIM 方案多以 MHA/GQA 解码访存瓶颈为设计基准（其核心负载特征为 KV Cache 密集访问），面临双重局限：一是 MLA-Absorb 等高算术强度执行模式的出现，使原有算术强度匹配关系不再成立；二是现有 PIM 计算单元的算力密度难以匹配 GPU Tensor Core/AI Core 上大规模通用矩阵乘法（GEMM）的计算需求。

例如，三星 HBM-PIM 和 SK Hynix AiM 分别在动态随机存取存储器（DRAM）存储体附近集成面向通用矩阵-向量乘法（GEMV）的乘加运算（MAC）单元或者 AI 引擎，其单器件或单芯片级算力通常处于约 1 TFLOPS 量级，而 A100 的 BF16/FP16 Tensor Core 稠密峰值为 312 TFLOPS，二者相差约 2.5 个数量级。因此，传统 PIM 更适合 MHA/GQA 解码中的低算术强度 KV 读取与 GEMV 类操作，而难以直接高效承载 MLA-Absorb 类高算术强度 GEMM 负载。

面向长上下文 MHA 解码，LoL-PIM^[15]与 L3^[16]分别探索了可扩展的 DRAM-PIM 系统与双列直插式内存模块

表 3 主流 AI 加速器 MoE 专家层临界批量（基于 DeepSeek-V3）

AI 加速器	理论算力/TFLOPS	内存带宽/(TB·s ⁻¹)	脊点/(Op·B ⁻¹)	B_{th}	典型批量	MoE 专家层状态
A100 SXM	312	2.0	156	4 896	≤512	访存密集
H100 SXM	989	3.35	295	9 440	≤512	访存密集
H20	148	4.0	37	1 184	≤512	访存密集
Ascend 910B2	376	1.6	235	7 520	≤512	访存密集

AI：算术强度 MoE：混合专家 TFLOPS：每秒万亿次浮点运算

(DIMM) PIM/GPU 协同架构; STARC^[9]进一步通过语义聚类将 KV 对映射到与 PIM Bank 结构对齐的连续存储区域, 在 KV Cache 规模为 1 024 词元时, 相比完整 KV 检索实现 54%~74% 的注意力层延迟降低和 45%~67% 的能耗降低。然而, 这些方案总体仍基于标准 K/V Cache 数据结构, 尚未针对 MLA 的低维潜变量缓存结构, 在布局、聚类与执行路径层面实现联合优化。

值得注意的是, SK Hynix 于 2025 年展示 AiMX 加速卡与 vLLM 框架协同的注意力卸载方案, 将访存受限的注意力计算卸载至 AiMX, 而将计算密集部分保留在 GPU, 印证了按算子存算特征进行异构分工的方向, 但其底层计算形态仍主要面向 GEMV/MAC, 对 MLA-Absorb 类 GEMM 负载的支撑能力仍受限于其 MAC/GEMV 导向的计算架构。综上, 现有方案的设计基准 (MHA/GQA) 与新架构 (MLA/MoE) 的负载特征之间仍存在不匹配, 构成共同挑战。

4.2 面向 MLA 的近存适配方向

面向 MLA 的近存适配思路需要从可重构计算、潜变量感知布局和多设备缓存协同 3 个层面进行重新设计。

1) 可重构计算单元支持双模式动态切换: MLA-Naive 模式在小批量解码下仍属访存密集型, 适合利用现有 HBM-PIM 的 GEMV/MAC 单元加速; 而 MLA-Absorb 模式算术强度显著提升, 更接近 GEMM 型计算, 对近存侧算力密度提出更高要求。因此, 可在 HBM 逻辑层或 3D 堆叠逻辑 Chiplet 中引入可重构矩阵计算单元, 支持 GEMV 与 GEMM 近存原语的动态切换。

2) 潜变量感知的 KV 存储布局优化: 传统 PIM 以完整注意力头为粒度分配 DRAM Bank, 而 MLA 的 C_{KV} 维度 d_c 远小于 $n_h \times d_h$, 建议以 d_c 为粒度重新设计 Bank 对齐策略, 使 C_{KV} 在物理上连续存储, 消除跨 Bank 冲突; 同时, 低维潜变量空间的语义聚类质量优于原始高维 KV 空间, STARC 类聚类方案在 MLA 布局下有望获得更高的预取命中率。

3) 多设备广播缓存缓解 TP 分片瓶颈: 针对 C_{KV} 单头结构的 TP 分片瓶颈, 可在近存层设置跨设备共享广播缓存, 令多张加速卡共享同一 C_{KV} 副本, 避免各设备独立从 HBM 重复加载。在广播缓存命中率较高、跨设备访问延迟可控的理想条件下, 多卡 KV 带宽开销可由 $O(N_{\text{device}})$ 降至 $O(1)$ ^[11]。

4.3 面向 MoE 的近存计算适配方向

在典型在线推理场景下, 单个专家接收的有效 Token 数有限, 专家权重读取难以被批量充分摊销, 使 MoE 专家层持续呈现访存密集特征。因此, 面向 MoE 的近存计算适配

应围绕专家热冷分层、近存并行调度和路由感知预取 3 个方向展开。

1) 热冷专家分层近存计算: MoNDE^[14]提出将少量高频热专家常驻 GPU HBM, 将大量低频冷专家保留在宿主 DRAM 或扩展内存侧 (如计算快速互连 (CXL) 内存池), 由近数据处理 (NDP) 单元在近数据侧就地执行专家计算。该方法以规模较小的激活数据传输替代大规模冷专家参数在高速串行计算机扩展总线 (PCIe) 等链路上的往返搬运, 从而降低专家卸载开销。在典型批量配置下, 实测相对于 GPU 卸载框架, 编码器推理提速最高 7.5 倍, 解码器提速最高 3.7 倍, NDP 单元面积开销仅 3.0 mm²。

2) 3D 近存处理混合并行调度: HD-MoE^[17]通过混合键合将逻辑计算层与 DRAM 堆叠紧密集成; 设计了离线混合并行映射与在线动态调度策略, 融合张量并行 (TP) 与专家并行 (EP), 以应对专家激活的实时不均匀性。在典型 MoE 模型配置下, 实测相比纯 TP 方案提速 1.1~1.8 倍, 相比纯 EP 提速 1.1~1.5 倍。

3) 路由感知预取策略: 利用 MoE 路由器输出的 Top-K 概率分布预判下一步热专家集合, 提前将冷专家数据从主存预取至近存缓冲区。可与投机解码结合, 以草稿模型预测词元序列, 从而将专家预取窗口从 1 步延伸至多步, 提升命中率。

4.4 MLA+MoE 组合架构的分层近存框架

综合以上分析, 图 3 给出了面向 MLA+MoE 组合架构的分层近存计算框架建议。框架自上而下分为 4 层: GPU 核心层承担 MLA-Absorb 路径下的高算术强度 GEMM 及大批量 MoE 专家计算, 发挥 Tensor Core/AI Core 的高算力密度优势; HBM 逻辑近存层承担 MLA-Naive 小批量注意力的 GEMV 加速、 C_{KV} 跨卡广播缓存, 以及热专家预取缓冲, 并支持 Naive/Absorb 双模式的动态切换; 3D 堆叠 DRAM 近存层执行冷专家 NDP 就地计算、KV Cache Token 淘汰评分, 以及路由感知预取; 主机 DRAM/CXL 扩展层负责冷专家权重的大规模存储及长上下文 KV Cache 溢出扩展, 并可结合 DIMM-PIM 等协同架构实现近存侧 Token 选择、缓存压缩与预取管理。

上述 4 层存储体系并非简单叠加即可获得性能收益, 其实际部署还伴随着相应的硬件代价与软件开销。

1) 硬件集成代价: HBM 逻辑层集成矩阵单元将占用逻辑底座面积与功耗预算, 并与物理层、可靠性/可用性/可维护性 (RAS) 机制及内存控制器形成资源竞争; DRAM 工艺对高频逻辑的支持有限, 使得“带宽充裕而算力密度不足”的矛盾依然存在。3D 混合键合虽能提供更高的垂直带宽,

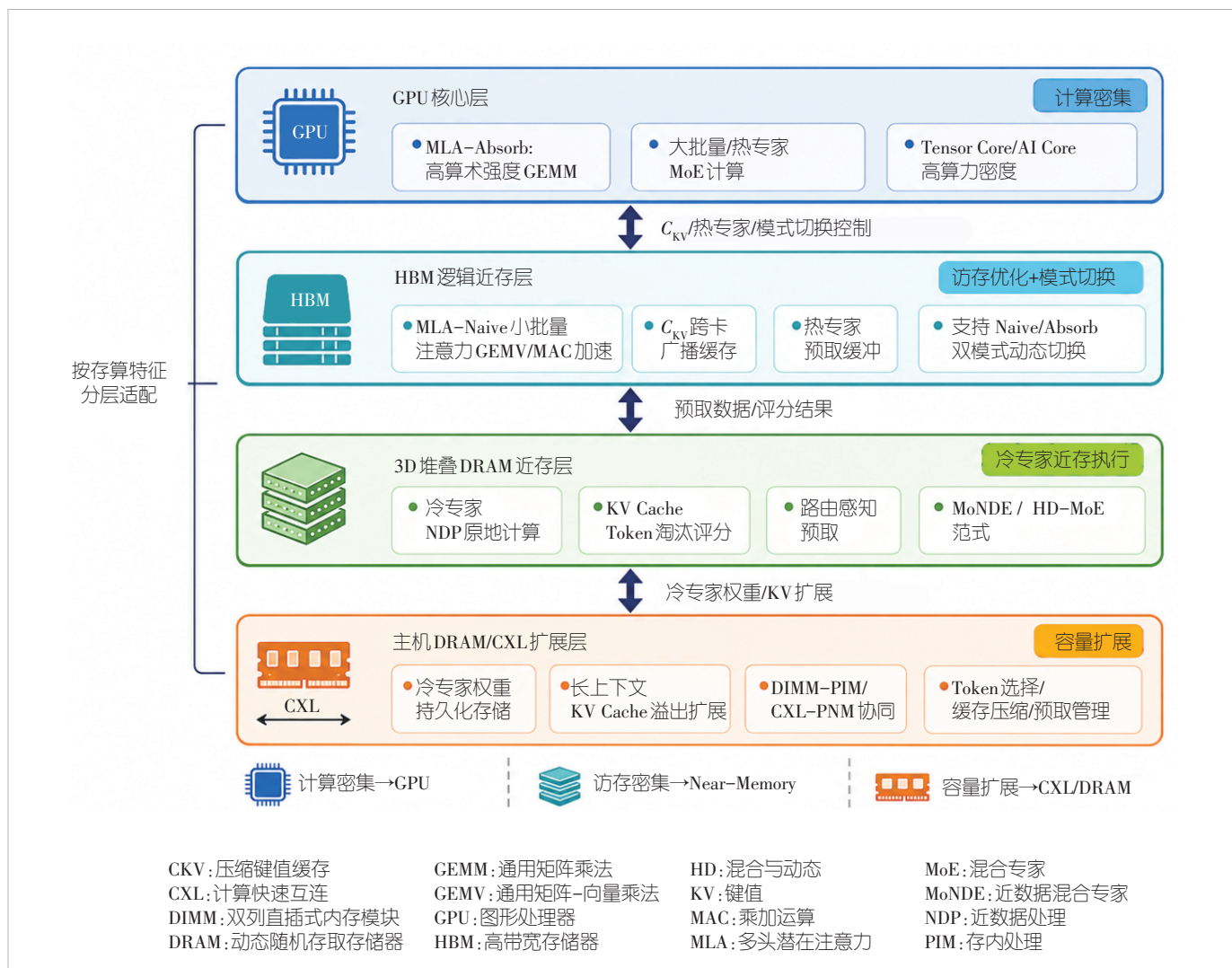


图3 面向MLA+MoE组合架构的分层近存计算框架

但也引入了堆叠良率、热密度、供电、测试及封装成本等多重工程挑战。

2) 跨介质带宽与延迟落差：HBM、3D DRAM、主机DDR/CXL在带宽、粒度和访问时延上相差显著，若按Token细粒度同步访问慢层，协议开销与往返延迟可能抵消近存计算带来的性能收益。CXL提供一致性与内存语义，可支持内存池化和跨节点共享^[18]，但其链路带宽远低于HBM。对于KV Cache等难以压缩的数据，跨CXL传输时的数据压缩还会受到布局、编解码效率及有效带宽的多重限制，因此需与预测搬移、异步流水及访问热度分层等策略进行联合设计。

3) 软件对象化与调度开销：运行时需同时维护专家块和共享前缀等不同对象；每类对象的生命周期、页粒度、量化格式及一致性域各不相同。MoE路由还存在动态不均衡

性，预取过早会占用近存容量，预取过晚则暴露CXL/PCIe延迟。HD-MoE和ELMoE-3D的结果表明，离线映射必须与在线动态调度、缓存及投机执行进行联合设计。

4) 低精度与编译开销：需处理异常值、不同层精度、累加精度和跨层格式转换。编译器必须联合选择数据布局、量化分组、GEMV/GEMM内核与放置层级，否则反量化、转置和Padding流量会抵消带宽节省。

5 总结与展望

本文围绕大模型推理阶段的存算特征演变，系统分析了MLA与MoE架构对近存计算适配的影响。通过对传统MHA、GQA、MLA-Naive、MLA-Absorb以及MoE专家层的算术强度建模可以看出，新一代大模型推理已不再呈现单一的访存瓶颈，而是形成了注意力层、专家层与硬件算力带宽比共同

塑造的异构存算特征分布。

MLA-Absorb 模式通过权重吸收显著提高注意力层算术强度,使其由传统 MHA 解码阶段约 1 Op/B 的强访存密集区域,提升至数百 Op/B 量级,并达到甚至跨越主流 AI 加速器的 Roofline 脊点。这意味着,传统面向 MHA KV Cache 读取瓶颈设计的 PIM 注意力加速器,在 MLA-Absorb 场景下面临算术强度匹配失效与算力密度不足的双重挑战。相比之下,MLA-Naive 模式在小批量、长上下文场景下仍具有较强访存密集特征,因而仍存在近存加速空间。这表明,未来面向 MLA 的近存计算架构需要从以 GEMV 为核心的注意力加速,转向同时支持 Naive/Absorb 双路径的可重构计算体系。

MoE 架构虽然通过稀疏激活降低了单 Token 计算量,但其大规模专家权重需要常驻或快速可达,且在线解码场景下有效批量通常难以充分摊薄专家权重读取开销。解析估算表明,在 A100、H100 等典型 AI 加速器上,MoE 专家层由访存密集转向计算密集所需的临界批量普遍高于常见在线推理批量。因此,MoE 专家层在实际部署中仍持续表现为访存主导,并逐步成为近存计算新的核心优化对象。热冷专家分层、路由感知预取、3D 近存处理以及 TP/EP 混合并行调度等方法,为缓解专家权重搬移瓶颈提供了重要方向。

从发展趋势看,近存计算架构需要从传统注意力加速器逐步演进为面向 MLA 与 MoE 的异构支撑体系。一方面,应面向 MLA 低维潜变量缓存特征,设计潜变量感知的 Bank 对齐、连续存储、聚类预取和跨设备广播缓存机制;另一方面,应面向 MoE 专家访问的长尾分布,构建热专家 HBM 常驻、冷专家近存执行、路由预测预取,以及与 CXL/DRAM 扩展层协同的分层存储与计算体系。随着 3D/3.5D Chiplet、HBM 逻辑层计算、CXL 内存扩展以及 RRAM-CIM^[19]等新型存储与封装技术的发展,近存计算有望在 MoE 冷专家存储、长上下文 KV 管理和低频专家原地计算等场景中发挥更大价值。

以 2026 年上半年发布的 DeepSeek-V4^[20]为代表,新一代 MoE 模型进一步融合压缩稀疏注意力(CSA)、高度压缩注意力(HCA)与多词元预测(MTP)机制,将百万级超长上下文变为推理标配。该类架构并非简单消除 KV 缓存压力,而是将瓶颈进一步分化:长序列下原始 KV 被压缩块替代,有效降低原生 KV 容量,但引入压缩索引、块元数据、稀疏检索计算等新开销;同时万亿级总参 MoE 模型的专家集合进一步扩张,FP4 混合量化成为权重部署主流,Group-wise 缩放因子及路由索引的元数据访存占比持续抬升。这一演变预示存算架构将面临三重新的适配需求:一是近存单元不仅要处理原始 KV,还需要支持压缩 KV 块的原位检索、块筛选

与索引遍历,硬件需兼容稠密潜变量与稀疏压缩块两套缓存访问范式;二是针对 FP4 等超低位宽权重量化,近存通路需要原生兼容 Scale 元数据并行加载、快速解包,缓解元数据流量对理论带宽收益的侵蚀;三是 MTP 多令牌预测会改变单步 Token 的算子调用模式,专家层与注意力层的计算访存交织关系进一步复杂化,固定流水线的近存硬件容易出现利用率下降。与此同时,细粒度专家划分、共享专家复用等模型层优化,会改变专家访问的热-冷分布,热专家集合的动态漂移现象,对近存侧的专家预取及缓存置换策略提出更高要求。

与此同时,该方向仍面临软件栈与编译器支持不足、低精度量化误差控制、MoE 路由预测不确定性,以及压缩注意力索引检索的硬件加速缺失等问题。总体而言,大模型推理的存算瓶颈正在从传统 MHA 注意力层,迁移至 MLA 潜变量缓存、压缩式长上下文 KV 块访问与 MoE 专家权重访问共同主导的新格局。未来近存计算不应继续沿用单一的 MHA-PIM 设计范式,而应围绕“按算子存算特征分层适配”的原则,形成可重构矩阵计算、潜变量与压缩块双模式感知的存储布局、热冷专家分层执行、路由感知预取以及多设备缓存协同的一体化架构,适配新一代融合压缩注意力、MTP 与超低位宽 MoE 模型的推理负载。

参考文献

- [1] Yuan Z H, Shang Y Z, Zhou Y, et al. LLM inference unveiled: survey and roofline model insights [PP/OL]. arXiv (2024-02-26) [2026-05-25]. <https://arxiv.org/abs/2402.16363>
- [2] Lee S, Kang S H, Lee J, et al. Hardware architecture and software stack for PIM based on commercial DRAM technology: industrial product [C]//Proceedings of ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). IEEE, 2021: 43-56. DOI: 10.1109/ISCA52012.2021.00013
- [3] DeepSeek-AI, Liu A X, Feng B, et al. DeepSeek-V2: a strong, economical, and efficient mixture-of-experts language model [PP/OL]. arXiv (2024-05-07) [2026-05-25]. <https://arxiv.org/abs/2405.04434>
- [4] Geens R, Verhelst M. Hardware-centric analysis of DeepSeek's multi-head latent attention [J]. Electronics letters, 2025, 61: e70504. DOI: 10.1049/ell2.70504
- [5] Fan Z H, Gagnon G, Liu Z Y, et al. Sparse attention remapping with clustering for efficient LLM decoding on PIM [PP/OL]. arXiv (2025-05-09) [2026-05-25]. <https://arxiv.org/abs/2505.05772>
- [6] Kim E A. Samsung ships industry-first commercial HBM4 with ultimate performance for AI computing: Samsung electronics [EB/OL]. (2026-02-10) [2026-05-25]. <https://news.samsung.com/global/samsung-ships-industry-first-commercial-hbm4-with-ultimate-performance-for-ai-computing>
- [7] Kwon H, Koo K, Kim J, et al. PIMphony: overcoming bandwidth and capacity inefficiency in PIM-based long-context LLM inference system [C]//Proceedings of IEEE International

- Symposium on High Performance Computer Architecture (HPCA). IEEE, 2026: 1–21. DOI: 10.1109/HPCA68181.2026.11408592
- [8] Ainslie J, Lee-Thorp J, de Jong M, et al. GQA: training generalized multi-query transformer models from multi-head checkpoints [C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2023: 4895–4901. DOI: 10.18653/v1/2023.emnlp-main.298
- [9] DeepSeek-AI, Liu A X, Feng B, et al. DeepSeek-V3 technical report [PP/OL]. arXiv (2024–12–27) [2026–05–25]. <https://arxiv.org/abs/2412.19437>
- [10] Yüzügüler A C, Çelik A, Zhuang J W, et al. TyphoonMLA: a mixed naïve-absorb MLA kernel for shared prefix [PP/OL]. arXiv (2025–09–25)[2026–05–25]. <https://arxiv.org/abs/2509.21081>
- [11] Zhao C G, Deng C Q, Ruan C, et al. Insights into DeepSeek-V3: scaling challenges and reflections on hardware for AI architectures [C]//Proceedings of the 52nd Annual International Symposium on Computer Architecture. ACM, 2025: 1731–1745. DOI: 10.1145/3695053.3731412
- [12] Meng F X. GQLA: group-query latent attention for hardware-adaptive large language model decoding [PP/OL]. arXiv (2026–05–13)[2026–05–25]. <https://arxiv.org/abs/2605.15250>
- [13] DeepSeek-AI, Liu A X, Mei A X, et al. DeepSeek-V3.2: pushing the frontier of open large language models [PP/OL]. arXiv (2025–12–02)[2026–05–25]. <https://arxiv.org/abs/2512.02556>
- [14] Kim T, Choi K, Cho Y, et al. MoNDE: mixture of near-data experts for large-scale sparse models [C]//Proceedings of the 61st ACM/IEEE Design Automation Conference. ACM, 2024: 1–6. DOI: 10.1145/3649329.3655951
- [15] Kwon H, Koo K, Kim J, et al. Lol-pim: long-context LLM decoding with scalable DRAM-PIM system [PP/OL]. arXiv (2024–12–28)[2026–05–25]. <https://arxiv.org/abs/2412.20166>
- [16] Liu Q, Chen L, Yang Y, et al. L3: DIMM-PIM integrated architecture and coordination for scalable long-context LLM inference [PP/OL]. arXiv (2025–04–24) [2026–05–25]. <https://arxiv.org/abs/2504.17584>
- [17] Huang H C, Zhong S Z, Zhang Z, et al. HD-MoE: hybrid and dynamic parallelism for mixture-of-expert LLMs with 3D near-memory processing [C]//Proceedings of IEEE/ACM International Conference on Computer Aided Design (ICCAD). IEEE, 2025: 1–9. DOI: 10.1109/ICCAD66269.2025.11240984
- [18] Natu M. Introducing the CXL 3. X specification [EB/OL]. [2026–05–25]. [https://computeexpresslink.org/wp-content/uploads/](https://computeexpresslink.org/wp-content/uploads/2025/02/CXL_Q1-2025-Webinar-Presentation_FINAL.pdf)

- 2025/02/CXL_Q1-2025-Webinar-Presentation_FINAL.pdf
- [19] 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
- [20] DeepSeek-AI, Xu A Y, Lin B C, et al. DeepSeek-V4: towards highly efficient million-token context intelligence [PP/OL]. arXiv (2026–04–26)[2026–05–25]. <https://arxiv.org/abs/2606.19348>

作者简介



庄凤云, 中移(苏州)软件技术有限公司工程师; 主要研究方向为大模型训推加速及智能体技术。



孙菲, 中移(苏州)软件技术有限公司高级工程师; 主要研究方向为AI推理加速。



钱岭, 中移(苏州)软件技术有限公司副总工程师, 教授级高工, 国务院政府特殊津贴获得者, 电信行业“大国工匠”; 主要研究方向为量子计算应用算法、量子计算云服务创新、量子计算与传统计算的异构算力融合、多样化量子算力并网等; 主持并负责国资委未来产业量子计算任务、国家自然科学基金、中国移动集团重点研发等多项课题; 曾获中国通信学会自然科学奖一等奖、北京市自然科学奖二等奖; 发表论文20余篇, 获授权发明专利10余项。