

面向智能体时代大语言模型推理的新型存算解聚合架构



A Novel Storage-Compute Disaggregated Architecture for Large Language Model Inference in the Agent Era

熊先奎/Xiong Xiankui, 王程晨/Wang Chengchen,
蔡文豪/Cai Wenhao

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202604004

网络出版地址: <https://link.cnki.net/urlid/34.1228.TN.20260901.0957.002>

网络出版日期: 2026-08-24

收稿日期: 2026-06-02

摘要: 针对传统大模型推理在长上下文场景及新兴智能体时代推理中所面临的键值 (KV) Cache 膨胀、本地显存容量受限及跨节点通信开销高等问题, 提出一种基于光输入/输出 (OIO) 拉远且以内存为中心的新型存算解聚合架构。该架构在物理层面实现计算与存储资源的解耦, 支持二者的独立伸缩、池化共享与按需协同; 在计算侧将 Prefill 与 Decode 阶段划分为异构计算簇, 在存储侧构建基于异质材料的全局内存池, 并借助近存计算单元对集合通信原语进行卸载。基于 8 个异构计算处理单元 (8XPU) 与 8 个内存处理单元 (8MPU) 全互联配置的系统级仿真分析表明, 该架构在长上下文及高并发推理场景下, 可有效降低首 Token 时延与逐 Token 时延, 减少数据搬运开销, 提升系统吞吐量, 并在满足同等服务水平协议 (SLA) 的条件下显著降低对计算资源的需求, 验证了其在智算基础设施中的潜在应用价值。

关键词: 智能体时代; 存算解聚合; 内存为中心; 光输入/输出; 近存计算

Abstract: To address the challenges of Key-Value (KV) cache explosion, limited local video memory capacity, and high inter-node communication overhead in long-context inference for traditional large language models as well as emerging agent-based reasoning scenarios, a novel compute-storage disaggregated architecture centered on optical I/O (OIO)-attached remote memory is proposed. At the physical level, this architecture decouples compute and storage resources, enabling independent scaling, pooled sharing, and on-demand collaborative orchestration. On the compute side, heterogeneous clusters are designated separately for the Prefill and Decode phases, while on the storage side, a global memory pool composed of heterogeneous materials is constructed, with near-memory processing units offloading collective communication primitives. System-level simulation results based on a fully interconnected configuration of 8 X processing units (8XPU) and 8 memory processing units (8MPU) demonstrate that the proposed architecture effectively reduces both time-to-first-token and per-token latency, minimizes data movement overhead, and improves system throughput in long-context and high-concurrency inference scenarios. Moreover, under equivalent service-level agreement (SLA) constraints, the required compute resources are significantly reduced, confirming its potential applicability in intelligent computing infrastructure.

Keywords: agent era; compute-storage disaggregation; memory-centric; optical I/O; near-memory computing

引用格式: 熊先奎, 王程晨, 蔡文豪. 面向智能体时代大语言模型推理的新型存算解聚合架构 [J]. 中兴通讯技术, 2026, 32(4): 17-22. DOI: 10.12142/ZTETJ.202604004

Citation: Xiong X K, Wang C C, Cai W H. A novel storage-compute disaggregated architecture for large language model inference in the agent era [J]. ZTE technology journal, 2026, 32(4): 17-22. DOI: 10.12142/ZTETJ.202604004

1 大语言模型推理现状与智能体时代的核心挑战

随着大语言模型 (LLM) 迈入规模化落地阶段, 智算基础设施的发展重心正从基础模型训练转向代理式人工智能 (Agentic AI) 驱动的推理服务^[1]。多模态应用的普及、长思维链能力的提升等, 使得 Agentic AI 应用中的核心瓶颈逐渐由算力限制转向内存约束。键值 (KV) Cache、记忆读写与状态管理所带来的访存开销, 已取代计算开销, 成为制约大规模智能体集群部署的关键因素。

1.1 大语言模型算法规模、上下文长度对内存容量的需求

随着 LLM 参数规模从 Llama-3 时期的 32B、DeepSeek V3 时期的 671B^[2], 快速扩展至 DeepSeek-V4 Pro^[3]、Kimi K2.5、Qwen3-Max 等万亿级规模, 基础设施面临的计算与存储压力日益严峻。长上下文指 LLM 在单次推理中处理超过 12.8 万超长文本的能力, 使模型能够一次性理解长篇书籍或完整代码库, 实现跨文档的精准推理与记忆。例如, 基于长上下文能力的推理阶段扩展范式由 OpenAI O1 模型首次提出, 并迅

速成为 LLM 的核心能力之一^[4]。进入智能体时代后，长程任务固有的复杂性、多智能体协作中的知识拓扑结构，以及多跳推理与证据链追溯等需求，对长上下文能力提出了更高要求^[5]。当前主流模型的上下文长度已达 25.6 万个（开源）至 200 万个（闭源）Token。然而，Transformer Decode-only 架构的计算与存储开销随上下文长度 N 呈 $O(N^2)$ 增长。尽管可通过量化、稀疏化、压缩等算法手段减轻负担，但长上下文带来的 KV Cache 容量需求依然可观。例如，具有混合注意力层的 Qwen3.5-235B 模型在 100 万 Token 长文本推理时，KV Cache 容量需求约为 90 GB^[6]；而 DeepSeek-V4-Pro 模型中，即便引入了压缩稀疏注意力（CSA）与重度压缩注意力（HCA）机制^[3]，其单个百万级 Token 上下文的 KV Cache 大小仍达 10.33 GB。KV Cache 正逐步取代模型权重，成为智算基础设施中主要的访存瓶颈。

未来智能体基础模型的重要能力“持续学习”和“可解释性”对存储系统提出了新要求。DeepSeek 提出的 Engram 架构将静态知识存储从高带宽内存（HBM）转移到主机内存（DRAM）中，在当前 Engram-27B 架构中，内存开销约为 5 GB^[7]，在 DeepSeek-V4 架构参数下，以 $\rho=0.7$ 估算，内存开销约为百 GB 级别；Google 提出的 Titan 和 Nest-Learning 架构^[8]，引入了推理过程中的可学习参数，考虑到参数读写和优化器存储开销，将使当前参数存储需求提升 3~5 倍。

1.2 智能体应用特征对内存带宽与访问模式的需求

Agentic AI 是一种能够自主运作、规划并执行复杂多步骤任务的智能系统。其核心特征在于从传统 LLM 的单一指令推理与被动响应模式，转向“自主决策、目标导向、持续学习”的主动式框架。Agentic AI 具备多步规划、工具调用及长期记忆等重要能力，单个任务可包含数百轮 LLM 请求^[5]，这使得推理负载从单次请求演变为连续的长链任务，模型存储开销与复用特征随之发生明显变化。

以 KV Cache 为例，在智能体典型工作流 ReAct 中，LLM 请求以“增量追加”的方式构建，每次执行动作或观察环境后，模型均在前文基础上继续推理。因此，多轮对话中的 KV Cache 命中率显著提升，典型复用概率约为 70%~84%^[9]。为支撑复杂记忆与长链推理能力，存储系统不仅需要大容量 HBM 以支持 Decode 计算，还需在前缀缓存命中时频繁从主机内存甚至固态硬盘（SSD）中读取 KV Cache，高频复用特征进一步强化了对智算基础设施存储容量与带宽的需求。

学术界和工业界已从推理框架、微架构及底层软件栈等多个层面对智算基础设施开展针对性优化。例如，vLLM 等主流推理框架围绕高效内存管理与低时延推理持续演进^[10]，

通过 PagedAttention、Speculative Decoding 等关键技术缓解显存压力与吞吐受限问题；PD 微架构设计将计算密集型的 Prefill 阶段与访存受限的 Decode 阶段解耦，从而提升系统推理效率；LMCache 等分层存储架构通过跨节点缓存迁移与持久化机制，有效降低 PD 场景下的通信开销与计算冗余。国产化异构算力平台亦从算子库、工具链等方面推进优化，着力解决框架兼容性、长尾应用场景及推理性能提升等问题。上述技术共同推动智算基础设施从传统一体化部署模式，向“阶段分离、存储分层、资源弹性调度”的新范式转变。

上述优化仅能解决特定场景下 LLM 推理的瓶颈问题。面向万亿级模型与 Agentic AI 场景，需应对百万级上下文处理、KV Cache 动态管理与高频复用等关键技术挑战，架构级创新将成为智算基础设施的核心能力。因此，未来 Agentic AI 计算集群的设计需从传统的以算力为中心，转向以内存与数据为中心的“Memory-centric”范式。

2 以内存为中心的存算解聚合架构发展现状

工业界围绕内存为中心的优化方向，已从共享内存编排、光互联扩展和内存池化等多个维度展开探索，以应对大模型推理中的内存墙问题。例如，微软 Feng Huang 通过多层共享内存、张量优化与近内存计算，缓解本地显存容量与跨图形处理器（GPU）通信压力；Celestial AI 借助光输入/输出（OIO）将计算芯片与更大规模的共享内存资源互联，试图在更大物理范围内提供接近本地 HBM 的访问性能；三星 CMM-D 则基于计算快速链接（CXL）总线提供面向数据中心的内存扩展与池化能力，以提高容量弹性与资源利用率。然而，上述方案仍主要聚焦于远端内存扩展、互联增强或通用池化能力，对大模型推理场景中计算负载特性、不同类型数据的访存需求，以及存储、通信与计算之间的协同优化，考虑仍不充分。

当前，业界对 OIO 技术的探索已从实验室验证迈入工程化落地阶段。作为一种先进的芯片级光互连技术，OIO 的核心在于利用光信号替代传统电信号进行片间数据传输，旨在将光互连功能直接集成至计算芯片内部，从而突破电互连在带宽密度、传输能耗和物理距离等方面的固有瓶颈。在产业界，主流芯片与半导体厂商正加速推进该技术的布局。2024 年，英特尔率先公开展示了其集成的光计算互连（OCI）芯片组原型；英伟达则明确将在下一代 Feynman 架构中大规模引入硅光子互连，以实现传统铜线互连的彻底替代。台积电研发的 COUPE（紧凑型通用光子引擎）技术采用 3D 异质光电集成方案，已于 2026 年实现规模化量产。在中国，以光联芯科为代表的本土企业亦积极推动 OIO 技术从研发验证

走向实用化部署，展现出良好的产业化前景。

为此，本文提出一种以内存为中心的新型存算解聚合架构。该架构借助 OIO 可将总线物理距离延展的特性，在物理层面将计算资源与存储资源彻底解耦，实现算力与存力的独立扩展、池化共享与按需协同。与现有方案不同，本架构不仅关注远端内存容量的扩展，更强调围绕推理流程构建分层访问语义与协同执行机制：在计算侧设置本地高速内存，并配备面向 Prefill 与 Decode 两类负载特征优化的异构计算簇；在存储侧构建由多个内存处理单元（MPU）节点组成的全局内存池，并结合异质存储材料形成兼顾性能与容量的层次化异构存储结构，从而实现对模型权重、KV Cache、元数据及过程数据的统一管理 with 高效支撑。

进一步地，本方案在 MPU 中引入近存计算能力，使部分数据处理与集合通信原语能够在靠近数据的位置完成，从而降低大模型推理过程中的数据搬移开销以及互联带宽压力，提升系统整体的数据访问效率与资源利用率。基于上述设计，所提出架构能够更好地适配智能体推理场景下的弹性部署需求，为该场景提供一种兼顾容量、带宽、时延与可伸缩性的新型架构支撑。

本方案面向智能体推理场景，主要特征如下：

- 1）提出一种以内存为中心的存算解聚合架构，面向智能体场景设计，突破传统计算与存储强绑定的部署模式，为大规模智能体系统提供高灵活性的基础设施支撑。
- 2）构建面向推理执行特征的异质融合内存与异构计算协同机制，提升系统对智能体推理不同阶段及不同负载类型的适配能力与执行效率。
- 3）结合动态随机存取存储器（DRAM）与 Flash 构建层

次化大容量异质融合内存体系，在保障热点访问性能的同时，突破本地高速内存的容量瓶颈，增强对超大模型参数及过程数据的承载能力。

- 4）在存储侧引入近存计算能力，使集合通信与部分数据处理能够在靠近数据的位置完成，从而提升系统级协同处理效率。

3 新型存算解聚合智算架构分析

如图 1 所示，该系统主要由本地内存与近端计算层、OIO 互联层以及全局内存与远端存储层 3 部分组成。其中，本地内存层采用字节级访问方式，容量为 GB 级别，主要用于存放指令、元表、元数据等高频访问的小规模数据，并为上层计算簇提供低时延的数据支撑。

近端计算资源根据大模型推理流程的不同阶段，划分为 Prefill 计算簇与 Decode 计算簇，分别适配高吞吐的上下文计算任务与低时延的自回归生成任务。该功能划分使系统能够针对不同负载特征进行差异化的资源配置，从而提升推理过程中的吞吐性能与响应效率。

中间的 OIO 互联层作为连接计算侧与存储侧的核心通道，承担大规模数据交换与远程访存请求的转发功能。相较于传统电互连方式，光互联具备更高带宽、更低时延及更强的扩展能力，能够有效支撑多计算节点与多存储节点之间的高速通信。同时，OIO 互联层为资源池化提供了基础，使不同计算簇可按需访问远端存储资源，从而实现跨节点的数据共享与协同处理。

在存储侧，全局内存层采用分块或 Tile 级访问方式，容量可扩展至百 TB 甚至 PB 级，由多个 MPU 存储节点构成，其

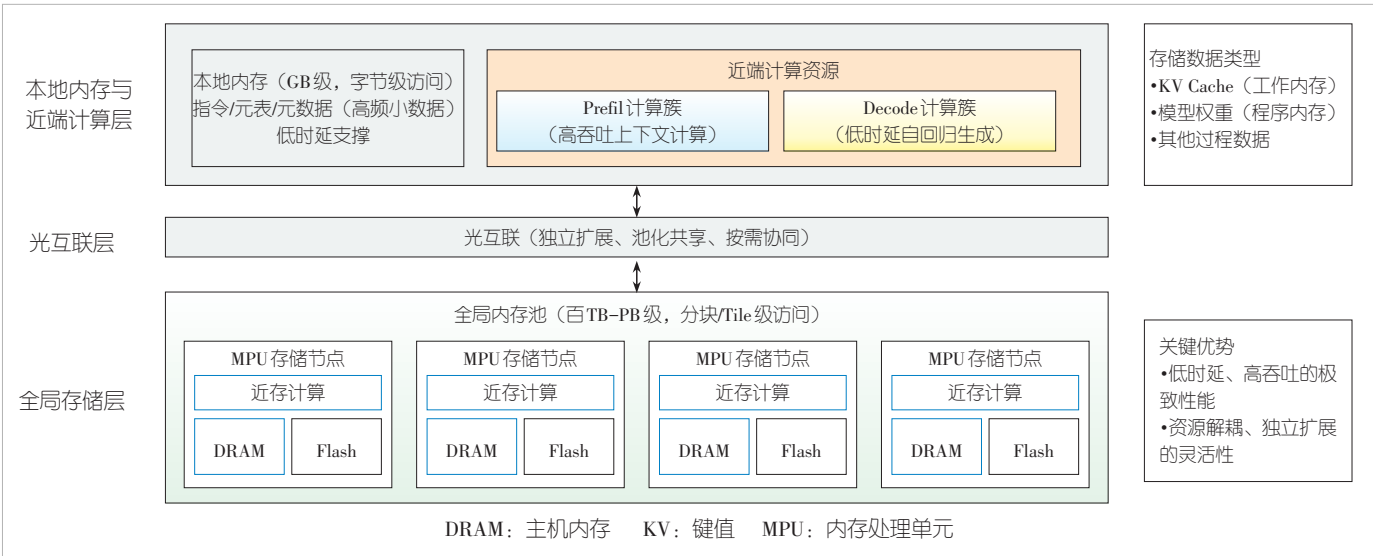


图1 新型存算解聚合架构

中 MPU 是一个定制化 ASIC。每个节点内部结合 DRAM 与 Flash 介质，形成兼顾访问性能与存储容量的层次化存储结构。MPU 还具备近存计算能力，通过自定义近存指令集，可在靠近数据的位置完成部分数据处理与集合通信原语操作，从而减少数据在存储侧与计算侧之间的往复搬移，降低互联带宽压力，并提升系统整体的数据访问与协同处理效率。按照功能划分，全局内存主要存放 3 类数据：一是作为工作内存的 KV Cache 数据，用于支撑推理过程中的上下文状态维护；二是作为程序内存的模型权重参数，用于支持模型计算执行；三是其他过程数据，包括中间激活、调度信息等。通过对不同类型数据的统一组织与分层管理，系统能高效地满足大模型推理过程中的数据访问需求。

该架构的关键优势在于，能够根据业务需求弹性扩展计算与存储资源，避免传统融合式部署中资源按固定比例绑定所带来的扩展受限与利用率不足问题。尤其在 Agent 场景下，模型参数、KV Cache 等大容量数据可集中部署于全局存储侧，而推理任务由不同计算簇动态承担，从而支撑多智能体任务下的弹性部署、高效共享与资源优化配置。

4 新型存算解聚合智算架构的性能分析

为评估本文提出的存算解聚合架构在智能体推理场景下的性能收益，我们通过整合专用加速器、高带宽异构池化内存等仿真库，构建了标准化交互模型，并实现了算网存联合仿真平台。本文以 8 张 H800 为基线配置（如表 1 所述），对比新型存算解聚合架构（如表 2 所述）与常规融合式 GPU 推理架构的性能差异。两种架构采用相同的计算侧资源配置，区别在于存算解聚合架构将模型权重、KV Cache 等大容量数据部署于远端全局内存，并借助 OIO 互联与 MPU 近存计算能力，降低计算侧的通信与显存压力。

新型存算解聚合架构采用 8 个异构计算处理单元（8XPU）+ 8 个内存处理单元（8MPU）的单节点配置，上层计算层由 8 个 XPU 组成，存储层由 8 个 MPU 组成，二者通过全互联的方式协同工作。单个 XPU-MPU 链路的基准带宽为 2 Tbit/s，系统聚合带宽为 2 TB/s。每个 MPU 配置 256 GB DRAM 与 1 TB Flash 存储器，形成总计 2 TB DRAM 与 8 TB Flash 的分层全局存储资源。基于 Qwen 与 Llama 3 模型，在两种配置下分别从首 Token 时延、逐 Token 时延、系统吞吐率、数据搬移量与资源需求等多个维度进行仿真分析。

图 2 给出了 Batch = 50 条件下，不同上下文长度下两种架构的首 Token 时延（TTFT）对比结果。可以看出，在短上下文场景下，两种架构的 TTFT 差异相对有限；而随着上下文长度持续增加，存算解聚合架构表现出更好的扩展性，二

者差距逐步拉大。以 3.2 万输入为例，Qwen235B 的 TTFT 相比常规融合架构下降 80.54%，Llama3 70B 下降 64.50%。这表明在长上下文推理场景下，存算解聚合架构通过将模型权重、KV Cache 及过程状态从本地设备侧解耦，并依托高带宽互联实现按需访问，能够有效缓解本地存储压力，从而显著降低首 Token 生成时延。

图 3 显示，在固定上下文长度为 4 000 的条件下，随着 batch 数目的增加，两种模型的 TPOT 均有所上升，但存算解聚合架构在所有测试点上均优于常规融合架构。其中，Qwen235B 在 batch=100 时 TPOT 由 0.041 3 s 降至 0.023 5 s，降幅达 43.1%；Llama3 70B 在 batch=10、50、100 时分别获得 62.1%、48.3% 和 54.0% 的时延优化。结果表明，存算解聚合架构能够有效降低解码阶段的单 Token 时延，并在高 batch 场景下展现出更优的性能优势与扩展能力。

图 4 显示，在固定 1 000 上下文和 batch=10 的条件下，Qwen 系列模型的吞吐率随参数量增大而下降，但存算解聚合架构在所有参数规模下均显著优于常规融合架构。具体而言，Qwen78B、Qwen130B 和 Qwen235B 的吞吐率分别提升约 138.6%、184.8% 和 138.5%。结果表明，存算解聚合设计能够有效提升大模型推理吞吐率，并在不同参数规模下保持稳定的性能收益。原因在于，传统融合式部署在面对更大规模模型时，往往需要依赖更细粒度的并行切分来适配本地高速内存容量，这不仅增加了跨设备通信与调度管理开销，也降低了系统的有效计算效率。本文架构则通过将模型参数、

表 1 仿真测试硬件基准

参数	规格
GPU 型号	NVIDIA H800
本地 HBM 容量	80 GB
本地 HBM 带宽	2 TB/s
单 GPU 聚合互联带宽	900 GB/s
节点间网络	400 Gbit/s InfiniBand

GPU：图形处理器 HBM：高带宽内存

表 2 新型存算解聚合架构配置参数

参数	规格
XPU 型号	NVIDIA H800
单 MPU 上行带宽(对应 8XPU)	2 Tbit/s(2 TB/s)可配
单内存节点 DRAM 容量	256 GB
单内存节点 FLASH 容量	1 TB
MPU 个数	8
近存计算支持功能	广播、聚合、累加等

DRAM：动态随机存取存储器 XPU：异构处理单元
MPU：内存处理单元

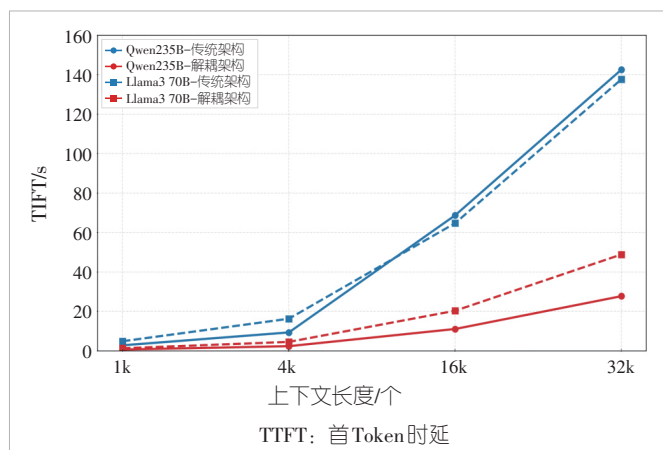


图2 不同上下文的TTFT

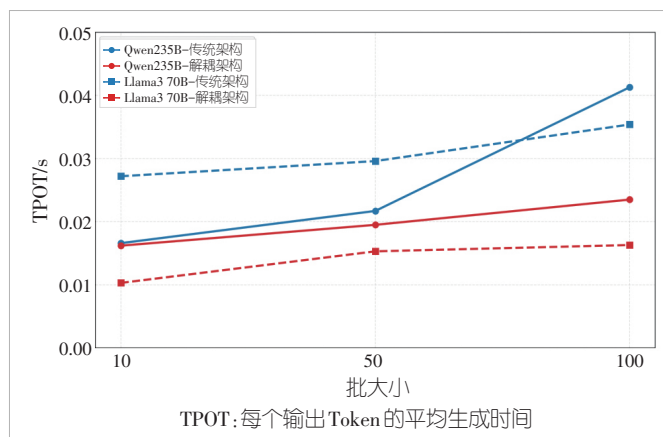


图3 不同batch的TPOT

KV Cache 及中间状态数据从计算侧剥离，统一存放于全局存储层，使计算节点的扩展不再主要受限于本地存储容量，而更多由实际推理算力需求决定。

图5比较了不同模型规模下单请求数据搬移量的差异。结果表明，本文架构在不同模型规模下的单请求数据搬移量均低于常规融合架构，且随着模型规模增大，这种优势愈加明显。本文架构的性能收益不仅源于远端存储的扩容能力，更来自对数据流动路径的系统性优化。通过将热点数据保留于DRAM层、冷数据驻留于Flash层，并在MPU侧完成部分数据处理与集合通信操作，本文架构将原本需频繁发生在计算层的跨设备搬移与同步，转化为更高效、更可控的远端访问与局部处理过程，从而降低了系统内部的数据搬移开销。

图6进一步给出了在满足同等服务质量约束（SLA）条件下，两种架构所需XPU数量的对比结果。可以看出，本文架构在不同模型规模下均需更少的XPU即可达到与常规融合架构相近的服务能力，且在405B等超大模型场景下这一优势尤为显著。该结果表明，在传统融合式部署中，计算资源的扩展不仅受算力需求驱动，也往往受限于本地存储容

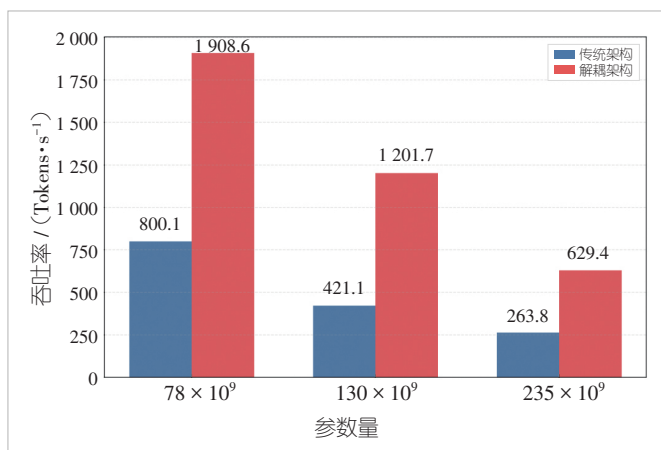


图4 不同参数数量下的吞吐量

量，因而需增加更多XPU以分摊模型参数与KV Cache的驻留压力。本文架构通过在存储层提供2 TB DRAM与8 TB Flash的分层全局内存，使计算资源扩展回归至“按算力需求扩展”的模式，从而在同等SLA约束下有效减少计算层资源需求，体现出更高的资源利用效率与部署弹性。

本文提出的存算解聚合架构在长上下文、大模型及高并发推理场景下，能够显著降低首Token与逐Token时延，减少系统内部数据搬移量，提升整体吞吐量，并在满足同等SLA条件下有效降低计算资源需求。上述结果表明，该架构能够有效突破传统融合式部署中由本地存储容量与跨设备通信共同造成的性能瓶颈，为面向大模型推理任务的弹性部署与高效共享提供了有效的体系结构支撑。

5 结束语

本文针对大模型长上下文及智能体推理场景下日益突出的内存瓶颈问题，提出了一种基于OIO互联拉远、以内存为中心的新型存算解聚合架构。该架构利用OIO互联低损耗、

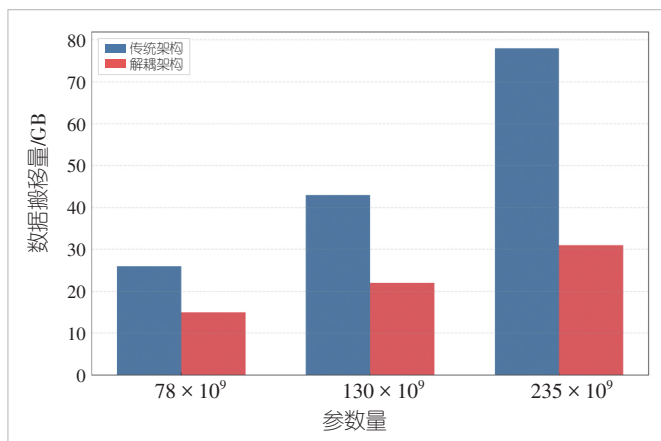


图5 不同模型规模下数据搬移量

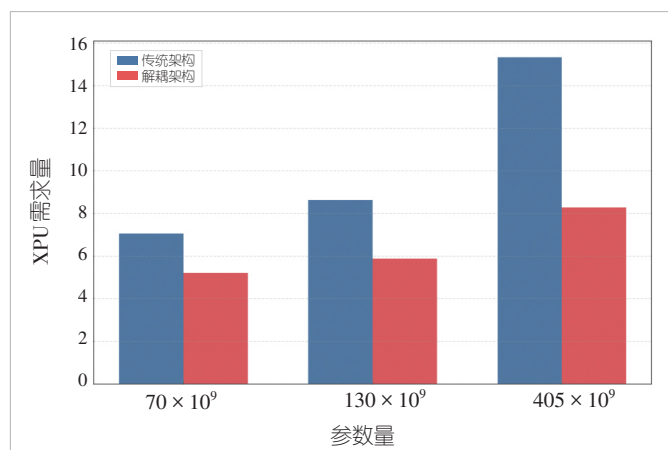


图6 服务质量约束条件下两种架构所需XPU数量

大带宽且与距离解耦的特性，将传统封装内受电互连严重制约的本地短距IO拉远，从根本上打破单芯片的IO密度墙，使计算资源与存储资源实现物理解耦、独立扩展与池化共享。在此基础上，构建了面向Prefill与Decode的异构计算组织，以及由MPU节点组成的DRAM-Flash分层全局异构内存，并引入近存计算能力，降低了推理过程中的数据搬移与互联压力。仿真表明，该架构通过IO拉远真正释放存算解耦的潜力，为大模型推理系统提供了一种高效能的新范式。

未来工作将围绕以下3个方面展开：一是结合原型系统实现，对远端访存、近存计算与光互联协同机制开展更贴近真实部署环境的验证；二是进一步研究面向模型权重、KV Cache及过程数据的自适应分层管理与动态迁移机制；三是面向国产异构算力平台，探索体系结构、运行时系统与推理框架之间的软硬件协同优化，以提升该架构在自主可控智算基础设施中的工程落地能力。

参考文献

- [1] Yao S Y. The second half [EB/OL]. (2025-04-10) [2025-06-05]. <https://ysmyth.github.io/The-Second-Half/>
- [2] DeepSeek-AI, Liu A X, Feng B, et al. DeepSeek-V3 technical report [PP/OL]. arXiv(2025-02-18) [2026-06-11]. <https://arxiv.org/abs/2412.19437>
- [3] DeepSeek-AI, Xu A X, Lin B C, et al. Deepseek-V4: towards highly efficient million-token context intelligence [PP/OL]. arXiv(2026-04-

- 26)[2026-06-14]. <https://arxiv.org/abs/2606.19348>
- [4] OpenAI. Learning to reason with LLMs [EB/OL]. (2024-09-12) [2026-06-14]. <https://openai.com/index/learning-to-reason-with-llms/>
- [5] Park G, Lee S, Park Y. Minimizing response latency in LLM-based agent systems: a comprehensive survey [J]. IEEE access, 2026, 14: 26140-26168. DOI: 10.1109/ACCESS.2026.3664226
- [6] Yang A, Li A F, Yang B S, et al. Qwen3 technical report [PP/OL]. arXiv(2025-05-14)[2026-06-18]. <https://arxiv.org/abs/2505.09388>
- [7] Cheng X, Tian R, Zeng W D, et al. Conditional memory via scalable lookup: a new axis of sparsity for large language models [PP/OL]. arXiv(2026-01-12)[2026-06-06]. <https://arxiv.org/abs/2601.07372>
- [8] Behrouz A, Razaviyayn M, Zhong P L, et al. Nested learning: the illusion of deep learning architectures [PP/OL]. arXiv(2025-12-31) [2026-06-13]. <https://arxiv.org/abs/2512.24695>
- [9] Ganjihal S R. Predictive multi-tier memory management for KV cache in large-scale GPU inference [PP/OL]. arXiv(2026-04-19) [2026-06-15]. <https://arxiv.org/abs/2604.26968>
- [10] Kwon W, Li Z H, Zhuang S Y, et al. Efficient memory management for large language model serving with PagedAttention [C]//Proceedings of the 29th Symposium on Operating Systems Principles. ACM, 2023: 611-626. DOI: 10.1145/3600006.3613165

作者简介



熊先奎，中兴通讯股份有限公司智算首席架构师及前瞻团队负责人；长期从事计算系统与体系结构、先进计算范式及异构计算加速器研究；主导了中兴通讯ATCA高级电信计算平台、服务器存储平台、智能网卡及AI加速器的系统设计；发表论文10余篇，拥有专利数十项。



王程晨，中兴通讯股份有限公司工程师；从事先进计算存储领域的相关工作；发表论文10余篇，申请专利9篇。



蔡文豪，中兴通讯股份有限公司技术预研工程师；主要研究方向包括深度学习算法、大语言模型、具身智能、模拟计算、无线通信系统等；发表论文10余篇，拥有专利6篇。