混合专家架构下大语言模型推理性能优化

李志勇[1]，田雯[2]
1. 中国移动通信集团有限公司，北京 100032
2. 中国移动通信集团设计院有限公司，北京 100032
DOI: 10.61369/TACS.2025100008

摘　　要： 针对大语言模型推理服务面临的高并发、低延迟与内存开销激增等挑战，本文探究了专家并行（Expert Parallelism, EP）策略在推理集群中的适配机制。剖析混合专家（Mixture of Experts, MoE）稀疏架构、KV Cache 缓存优化及预填充 – 解码分离（Prefill-Decode Separation, PD Separation）三大关键技术的性能增益原理，构建 " 架构 – 机制 – 部署 " 三位一体的优化体系。基于 DeepSeek R1模型的 W8A8量化版本，在昇腾910B2集群环境下设计多维度对比实验，验证不同节点规模、序列长度及并发场景下的性能表现。最终提出面向低时延与高吞吐两类核心场景的集群资源配置方案与动态调度策略，为大规模大模型推理服务的工程化落地提供理论支撑与实践指导。

关 键 词： 大语言模型；推理服务；专家并行；预填充 – 解码分离

Optimization of Inference Performance for Large Language Models under the Mixture of Experts Architecture

Li Zhiyong[1], Tian Wen[2]
1.China Mobile Communications Group Co., Ltd., Beijing 100032
2.China Mobile Group Design Institute Co., Ltd., Beijing 100032

Abstract: In response to the challenges faced by large language model inference services, such as high concurrency, low latency, and soaring memory overhead, this paper explores the adaptation mechanism of the Expert Parallelism (EP) strategy in inference clusters. It analyzes the performance enhancement principles of three key technologies: the Mixture of Experts (MoE) sparse architecture, KV Cache optimization, and Prefill-Decode Separation (PD Separation), and constructs a trinity optimization system encompassing "architecture-mechanism-deployment". Based on the W8A8 quantized version of the DeepSeek R1 model, multi-dimensional comparative experiments are designed in the Ascend 910B2 cluster environment to verify performance under different node scales, sequence lengths, and concurrent scenarios. Finally, cluster resource allocation plans and dynamic scheduling strategies tailored for two core scenarios—low latency and high throughput—are proposed, providing theoretical support and practical guidance for the engineering implementation of large-scale large language model inference services.

Keywords: large language model; inference service; expert parallelism; prefill-decode separation

引言

大语言模型（Large Language Model，LLM）技术的持续迭代推动人工智能进入规模化应用阶段，其参数规模从千亿级突破至万亿级，上下文窗口扩展至百万 token 级别，模型结构也从传统稠密架构演进为混合专家（MoE）等稀疏架构，在语义理解、长文档分析等场景中展现出卓越能力[1-3]。然而，模型能力的跃升也使推理服务面临严峻的工程化挑战，形成 " 高精度、低延迟、长序列、高并发、多模态 " 五大核心需求闭环：高精度要求驱动模型参数与结构复杂度提升，直接加剧集群资源调度压力；低延迟要求与传统推理架构存在本质矛盾，难以满足实时交互场景需求；长序列处理带来内存复杂度指数级增长，引发内存延迟与资源浪费问题；高并发场景对集群带宽与扩展性提出更高要求，传统并行策略难以适配 MoE 架构的动态特性；多模态融合则需要解决多源数据协同与异构算力调度的协同难题[4-5]。

现有研究中，MoE 架构通过条件计算实现参数规模与计算开销的解耦，专家并行技术解决大规模 MoE 模型的内存瓶颈，KV Cache 优化降低长序列计算复杂度，PD 分离策略缓解两阶段资源竞争[1,5]。但这些技术的协同优化机制、不同硬件环境下的适配规则及场景化部署策略仍缺乏系统验证。为此，本文聚焦 MoE 架构下推理服务的性能优化问题，深入分析三大关键技术的作用机理，通过昇腾910B2集群的实证研究，提出场景化的资源配置与调度方案，为突破大模型推理服务的性能瓶颈提供解决方案。

一、大模型推理服务关键技术

（一）MoE 稀疏模型架构

MoE 架构通过”路由器－专家网络”的核心结构实现计算稀疏化，是支撑万亿级参数模型部署的关键技术。如图1所示，MoE 层的核心组件包括路由器（Router）与多组专家网络（Experts）：路由器接收输入 token 后，通过权重计算与 Top-K 筛选机制，动态选择少量专家参与计算；专家网络通常采用前馈网络（FFN）结构，部分场景下可通过嵌套 MoE 形成层次化专家系统，实现更精细的任务适配 [2-3]。

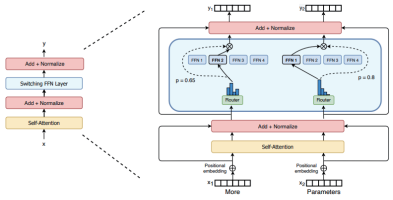

图1 MoE 模型架构核心

相较于传统 Transformer 稠密架构，MoE 模型具备三大核心优势：一是计算稀疏化，仅激活 Top-K 专家（通常 K=2）参与计算，在参数规模提升一个数量级的前提下，计算开销仅增加约 10%-20%；二是参数－计算解耦，通过增加专家数量扩展模型容量时无需同步提升计算量，使万亿级参数模型的推理部署成为可能；三是任务泛化性增强，不同专家可自适应学习不同数据分布特征，在多任务场景中表现出更优的适配能力 [3]。

（二）专家并行技术

MoE 架构的专家数量增长虽能提升模型容量，但会导致内存消耗线性增加。以 DeepSeek V3模型为例，其671B 参数规模对应的专家权重需 TB 级显存支撑，远超单卡 GPU 的显存上限 [3]。专家并行（EP）技术通过分布式存储与协同计算机制破解这一瓶颈，核心逻辑是将 MoE 模型的专家子网络按物理设备拆分，每张卡仅存储部分专家参数，通过动态路由与跨设备通信实现全局协同计算，如图2所示。

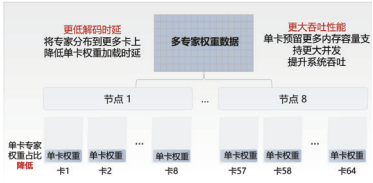

图2 专家并行示意图

专家并行架构下的推理流程可分为三个关键阶段：①路由分发阶段，基于路由器输出的专家权重与 Top-K 选择结果，将输入 token 的特征向量分发至对应专家所在的计算节点；②分布式计算阶段，各节点独立执行本地专家的前向计算，无需依赖外部节点数据，降低跨节点通信压力；③结果聚合阶段，通过 All-to-All 通信协议汇总各节点的计算结果，经融合后生成最终输出。该技术本质是通过＂空间换资源＂策略，将集中式显存压力分摊至多设备，同时保持计算稀疏化带来的效率优势 [5]。

（三）KV Cache 缓存机制

传统 Transformer 架构的自注意力机制存在计算冗余问题，每步解码需重复计算历史 token 的键值对（Key-Value），导致计算复杂度随序列长度呈 $O(n^2)$ 增长。KV Cache 作为针对性优化手段，通过缓存前序 token 的 Key 与 Value 向量，将自注意力计算复杂度降至 O(n)，大幅提升长序列解码效率 [4]。

KV Cache 的存储结构可定义为四维张量 [模型层数，注意力头数，序列长度，特征向量维度]，其中各维度分别对应自注意力层总数、单层级注意力头数量、输入 token 序列长度及单个注意力头的输出特征维度。以2层、2注意力头、序列长度2、特征维度4的简化模型为例，其 KV Cache 存储结构如下表所示：

层级	注意力头	序列位置	Key 向量	Value 向量
Layer 0	Head 0	Pos 0	[0.2，0.3，0.4，0.5]	[0.6，0.7，0.8，0.9]
Layer 0	Head 0	Pos 1	[0.3，0.4，0.5，0.6]	[0.7，0.8，0.9，1.0]
Layer 0	Head 1	Pos 0	[0.5，0.6，0.7，0.8]	[0.1，0.2，0.3，0.4]
…	…	…	…	…
Layer 1	Head 0	Pos 0	[1.1，1.2，1.3，1.4]	[1.5，1.6，1.7，1.8]
…	…	…	…	…

结合 MoE 架构特性，KV Cache 在推理过程中呈现两阶段差异化特征：预填充阶段（Prefill）：处理输入 prompt 序列时需激活多组专家并跨卡并行计算，KV Cache 需存储所有专家层的中间键值对，数据量庞大且对跨节点同步效率要求高；解码阶段（Decode）：逐 token 生成时仅依赖最新 token 的计算结果，专家激活范围收缩，KV Cache 的访问模式从＂批量写入＂转为＂高频只读＂，对显存带宽的局部性能提出极高要求 [5]。

（四）PD 分离策略

Prefill 与 Decode 阶段的计算特性差异导致资源竞争问题突出：Prefill 阶段为计算密集型任务，需大批次处理输入序列以完成 KV Cache 填充；Decode 阶段为内存密集型任务，需高频访问缓存并逐 token 生成输出。Yu 等 [5] 的实验证明，二者混部时，单个 Prefill 作业会导致 TTFT（首字时延）与 TPOT（字间时延）显著增加，且输入序列越长，干扰效应越明显（图3）。

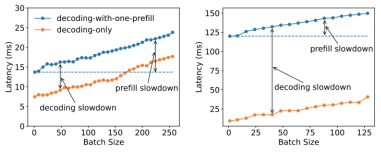

图3 Prefill 和 Decode 相互干扰，增大时延

PD 分离策略通过软硬件资源解耦解决这一问题：软件层面将模型实例拆分为 Prefill 实例与 Decode 实例，硬件层面划分专用 Prefill 节点与 Decode 节点，结合两阶段特性设计差异化资源配置方案——Prefill 节点配置高算力硬件以提升批量计算效率，Decode 节点配置高访存带宽硬件以优化缓存访问性能（图4）。这种架构设计使两阶段任务无需分时复用资源，Prefill 生成的 KV Cache 可通过高速互联直接传输至 Decode 节点，避免资源竞争导致的时延增加（图5）。

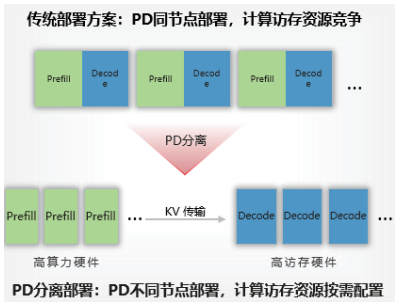

图4 PD 分离工作原理

图5 PD 分离与 PD 混部的计算状态对比

因此，PD 实例的分配以及占用资源的配比是保障系统性能的核心关键。实际部署中需结合模型结构、芯片性能、业务并发量及序列长度等参数动态调整配比。

二、实验与分析

（一）实验设置

实验采用 DeepSeek R1模型的 W8A8量化版本，硬件环境为华为昇腾910B2服务器（FP16算力376TFLOPS，显存64GB）集群，分别配置2机、8机、16机三种节点规模。数据集采用工具自动生成，覆盖1k1k（输入1024token/ 输出1024token）、2k2k、4k1k、6k1k、8k1k 五种序列长度组合，以模拟不同业务场景。实验对比 PD 混部与 PD 分离两种部署模式的性能差异，核心控制变量为节点规模、序列长度及并发数。

（二）性能指标与测试方法

实验采用 " 输入 – 输出 " 双维度指标体系：输入参数包括输入序列长度、输出序列长度、并发数及数据集规模；输出指标分为时延类（TTFT 首字时延、TPOT 字间时延）与吞吐类（TPS，token 生成速率），其中 TTFT 反映用户等待首字符的响应速度，TPOT 反映连续生成的流畅度，TPS 反映系统单位时间的处理能力。

测试方法采用 " 时延约束下的最大吞吐寻优 " 策略：设定时延阈值（TTFT ⩽1s，TPOT ⩽100ms），通过梯度增加并发数与序列长度，记录不同组合下的性能指标，确保实验结果的工程参考价值。

（三）实验结果

2机集群（PD 混部模式）性能结果

输入长度	输出长度	并发数	单卡输出吞吐 TPS
1024	1024	3072	55.9
2048	2048	2048	51.1
4096	1024	4096	35.5
6144	1024	2048	26.2
8192	1024	384	16.7

8机集群（2×2P+1×4D 模式）性能结果

输入长度	输出长度	并发数	TTFT 平均 (ms)	TPOT 平均 (ms)	单卡输出吞吐 TPS
1024	1024	3072	641.70	71.50	197.23
2048	2048	2048	998.04	105.50	146.63
4096	1024	4096	1319.02	46.21	48.50
6144	1024	2048	2637.21	50.39	55.55
8192	1024	384	4026.91	89.56	39.38

16机集群（4×2P+1×8D 模式）性能结果

输入长度	输出长度	并发数	TTFT 平均 (ms)	TPOT 平均 (ms)	单卡输出吞吐 TPS
1024	1024	3072	625.41	89.34	212.68
2048	2048	4096	561.37	92.20	196.00
4096	1024	4096	1823.10	55.73	92.70
6144	1024	2048	2122.98	46.71	61.09
8192	1024	2048	4803.48	47.37	45.50

（四）实验结论

1. 部署模式影响：PD 分离策略的性能优势显著，16机集群在1k1k场景下的单卡 TPS 达212.68，较2机混部模式提升2.8倍，且 TPOT 稳定在89.34ms 以内，验证了资源解耦对算力释放的关键作用。

2. 序列长度影响：输入序列长度与 TTFT 呈正相关，8k 输入场景的 TTFT 较1k 场景增加6.4倍（16机集群），但 TPOT 受输入长度影响较小，8k 场景的 TPOT 仅47.37ms，说明 KV Cache 对长序列解码的优化效果显著。

3. 并发与吞吐关系：固定序列长度下，吞吐随并发数增加呈先升后稳趋势，但并发过量会导致时延突破阈值（如8机集群8k输入场景，并发从384增至512时 TTFT 超1.2s）。

4. 场景化配置建议：①低时延场景（如实时对话）：推荐2+8机1P1D 方案，1k 输入场景 TTFT 可控制在500ms 以内；②高吞吐场景（如批量生成）：推荐8+8机4P1D 方案，1k1k 场景单卡 TPS 可达200以上，满足规模化服务需求。

三、未来研究方向

大模型推理性能优化正从单一技术突破迈向 " 场景 – 算法 – 硬件 – 通信 " 的全栈协同阶段，未来可聚焦三大方向：①动态适配调度：基于强化学习构建负载感知模型，实现专家并行策略、PD 配比的实时动态调整；②硬件协同优化：探索异构算力池化技术，结合 CXL 共享内存协议降低跨节点通信时延；③稀疏化深度优化：研究 MoE 专家的动态激活阈值调整机制，结合量化技术进一步降低内存开销。此外，多模态推理场景的资源调度、边缘设备的轻量化部署也将成为重要研究分支。

参考文献

[1] Lepikhin D, et al. GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding[J]. arXiv preprint arXiv:2006.16668, 2020.

[2] Fedus W, et al. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity[J]. arXiv preprint arXiv:2101.03961, 2022.

[3] DeepSeek AI. DeepSeekMoE: Towards Ultimate Expert Specialization in Mixture-of-Experts Language Models[J]. arXiv preprint arXiv:2401.06066, 2024.

[4] 王某某，等 . 大模型时代的混合专家系统优化综述 [J]. 计算机研究与发展 ,2024.

[5] Yu G, et al. DistServe: Disaggregating Prefill and Decoding for Goodput-optimized Large Language Model Serving[J]. arXiv preprint arXiv:2405.20070, 2024.

