PR #48788 完整报告

vllm-project/vllm

[ROCm][Perf][DSV4] Improve sparse decode reduction occupancy on gfx950

合并时间：2026-07-18 01:25

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/48788

执行摘要

- 一句话：优化 gfx950 稀疏解码 reduce 算子 occupancy
- 推荐动作：该 PR 是一个针对特定 GPU 架构（gfx950）的轻量级性能优化，改动量小（仅 4 行）、风险可控、收益明确。建议关注后续 PR #46275 合并时对该 reduce 几何形状在 gfx942 上的验证。值得学习的是如何通过调整 workgroup 粒度来提升 GPU occupancy 的优化思路。

功能与动机

现有的 split-K 稀疏解码 reduce 算子每个 workgroup 处理 16 个头，保持一个 [16, COMB_DIM] FP32 累加器，限制了 workgroup 级别并行度。通过将每个 workgroup 处理一个 head，可以降低每 workgroup 的累加器和寄存器压力，暴露最多 16 倍的独立 workgroup。PR body 明确指出该更改仅影响 gfx950 的 split-K 解码路径，gfx942 仍走 fallback 路径，需单独验证。

实现拆解

1. 修改 reduce kernel 启动网格维度：在 vllm/v1/attention/ops/rocm_aiter_mla_sparse.py 的 _rocm_sparse_attn_decode_ragged_triton 函数中，将 _sparse_attn_decode_reduce_kernel 的 grid 从 (num_queries, heads_blocks) 改为 (num_queries, num_heads)，使每个 workgroup 处理单个 head 而非 16 个 head 的块。
2. 调整 BLOCK_H 常量：将 BLOCK_H 从 block_h 改为 1，对应每个 workgroup 只处理一个 head 的新粒度。
3. 验证与测试：在 MI355X (gfx950) 和 MI300X (gfx942) 上进行了性能和准确率验证，GSM8K 准确率超过 94% 阈值。MI300X 的测试结果显示该改动对该架构也有性能收益，但作者明确指出该路径在 gfx942 上未开放，需后续 PR 合并时注意。

关键文件：

- vllm/v1/attention/ops/rocm_aiter_mla_sparse.py（模块 注意力；类别 source；类型 core-logic；符号 _rocm_sparse_attn_decode_ragged_triton）：包含所有变更：修改 reduce kernel 启动网格和 BLOCK_H 常量，是唯一的变更文件。

关键符号：_rocm_sparse_attn_decode_ragged_triton

评论区精华

Review 中的核心讨论聚焦于是否需要对 gfx950 添加额外防护门控。tjtanaa 建议用 _ON_GFX950 宏保护该变更，但作者 Fangzhou-Ai 回应称整个解码路径已经仅为 gfx950 门控，因此无需额外门控。双方未就此产生进一步分歧，PR 最终被 tjtanaa 批准。另外，shen-shanshan 在 MI300X 上验证了该改动，发现性能也有所提升，但作者指出该架构当前不走 split-K 路径，需要后续 PR #46275 合并时单独验证。

- 是否需要为 gfx950 添加额外门控宏 (design): 作者说明无需额外门控，审查者接受并批准。
- MI300X 上的性能验证 (testing): 数据有价值，但 PR 作者指出需在 #46275 合并时单独验证。

风险与影响

- 风险:
  1. 回归风险（低）：仅修改了 gfx950 专属路径中的 reduce kernel 网格配置，且通过了 GSM8K 准确率验证（95.679%）。
  2. 兼容性风险（中）：如果未来 PR #46275 将 split-K 路径扩展到 gfx942，当前改动的 reduce 几何形状可能在 gfx942 上未经验证。PR 作者特别指出需要在该 PR 合并前进行验证或添加门控。
  3. 数值精度风险（低）：改动仅改变 workgroup 划分方式，不改变 reduction 数学逻辑和顺序，因此数值结果一致。 - 影响：用户：DeepSeek V4 用户在 AMD MI355X (gfx950) GPU 上使用 vLLM 部署时，推理吞吐量提升约 4-5%，TTFT 延迟降低最多 16.88%。 系统：无外部依赖变更，无需配置调整。 团队：为未来 gfx942 启用 split-K 路径提供了铺垫和验证要求。
- 风险标记：架构特定优化，依赖后续兼容性验证

关联脉络

- PR #46275 Broaden split sparse decode to gfx942: 该 PR 计划将 split-K 路径扩展到 gfx942，与本 PR 的 reduce 几何形状紧密相关，需验证兼容性。
- PR #46172 Optimize the sparse indexer path: 同一作者在同一功能区域（DSV4 稀疏解码）的另一个优化，但针对不同子路径。