PR #21213 完整报告

sgl-project/sglang

[AMD]: Support MLA with nhead<16 and FP8 KV cache for TP=8 (Kimi K2.5…

合并时间：2026-04-05 13:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21213

执行摘要

本 PR 通过 head-repeat 策略扩展了 AITER MLA 注意力后端，支持头数小于 16（如 4 或 8）的配置，使 Kimi K2.5 等模型能在 TP=8 下运行，提升 AMD GPU 的兼容性和性能。变更涉及核心内核逻辑和测试更新，影响范围适中，已通过 review 并合并。

功能与动机

主要动机是解决 Kimi K2.5 模型在 TP=8 时每个 GPU 只有 8 个注意力头，而 AITER MLA 内核要求头数必须是 16 的倍数的问题。PR body 中描述，通过 head-repeat 策略将头数扩展到 16，可以重用现有优化 ASM 内核，避免开发新变体，从而支持更灵活的配置。这直接回应了 AMD 平台上模型部署的限制。

实现拆解

主要改动集中在 python/sglang/srt/layers/attention/aiter_backend.py：

1. 头数断言放宽：更新 __init__ 中的断言，接受头数为 4、8 或 16 的倍数（16 到 128）。
2. padding 逻辑：引入 num_head_padded 和 head_repeat_factor 变量，当头数小于 16 时，通过 repeat_interleave 扩展到头数 16。
3. wrapper 函数：新增 _mla_decode_fwd_with_head_pad 函数，处理头数 padding 和解码输出切片。
4. 集成到 MLA 路径：在 forward_extend 和 forward_decode 方法中集成新逻辑，确保不同模式下的兼容性。 测试文件更新 TP 值从 4 到 8，并移除过时注释以反映新支持。

评论区精华

review 讨论中，gemini-code-assist[bot] 指出 forward_extend 方法存在代码重复，建议重构为私有 helper 函数以提高维护性，作者未完全采纳但进行了其他优化。kkHuang-amd 建议优化 new_empty 使用以减少内存分配冗余，作者响应并更新代码。此外，过时测试注释被指出并移除。讨论焦点是代码清晰度和性能微调。

> gemini-code-assist[bot]: "This block of logic... is duplicated three times in forward_extend... reduces maintainability."
>
> kkHuang-amd: "Do we need to do new_empty twice... Maybe we can do the similar below changes..."

风险与影响

技术风险：头数 padding 可能引入轻微性能开销；断言放宽后，异常配置可能导致未定义行为；FP8 与非 FP8 路径的兼容性需仔细测试。例如，aiter_backend.py 中的逻辑变更若未全面覆盖，可能引发回归。

影响评估：对用户，允许 Kimi K2.5 在 TP=8 下运行，提升硬件利用率，基准显示 TPOT 改进显著；对系统，扩展了 MLA 后端支持范围；对团队，提供了 head-repeat 策略的设计案例。影响程度中等，主要限于 AMD 平台和特定模型。

关联脉络

从近期历史 PR 看，本 PR 与 PR 22372（FP8 FlashMLA KV padding）和 PR 21166（AMD GLM-5 优化）相关，它们都涉及注意力内核优化和 AMD 平台支持，显示出仓库在扩展内核兼容性和性能优化上的持续努力。本 PR 是这一趋势的一部分，聚焦于头数限制的突破。