

PR #29816 完整报告

sgl-project/sglang

[AMD] Update ROCm AITER pin to 9127c94

合并时间: 2026-07-01 16:09

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29816>

执行摘要

- 一句话: 更新 ROCm Docker AITER 版本并适配新 API
- 推荐动作: 此 PR 为常规基础设施更新, 建议阅读 docker/rocm.Dockerfile 中的版本变更, 了解 AMD GPU 构建的最新依赖。aiter_backend.py 的 1 行改动值得留意: 当第三方库 API 变更时, 如何以最小改动兼容新签名。

功能与动机

PR body 明确说明: 更新 ROCm Docker AITER pin 到 9127c94, 并且包含了来自 SGLang commit 4bb8f00bcb291bec8243f013dabb9482c1eedbc6 的补丁, 用于修复 AITER MLA reduce 签名变更。Issue 评论进一步解释: mla_reduce_v1 新增了 num_kv_splits 参数, 而 prefill PS 路径不需要额外的 split cap, 因此传入 0 以保持默认 CU 基数 reduce 尺寸。

实现拆解

1. Dockerfile 版本更新 - 修改 docker/rocm.Dockerfile, 将 AITER_COMMIT_DEFAULT 环境变量在四个构建阶段 (base image 942, 942+rocm720, 950, 950+rocm720) 中均从 7d604afe5fa7efba63c0dce323b95d9daf2db112 更新为 9127c94a18e4398e1eba91f6639e910f0994ad02。
2. 适配新 AITER API 签名 - 在 python/sglang/srt/layers/attention/aiter_backend.py 的 mla_fp8_prefill_attn 函数中, 调用 mla_reduce_v1 时新增一个参数: 0, 对应 num_kv_splits。该参数由上游 ROCm/AITER#3391 引入, 用于在 decode 阶段按批次限制 reduce 分片数量, 而 prefill 路径中传入 0 可沿用旧有的基于 multiprocessor count 的默认分片策略。

关键文件:

- docker/rocm.Dockerfile (模块 部署脚本; 类别 infra; 类型 infrastructure) : 更新了四个构建阶段的 AITER 依赖版本, 是本次变更的核心基础设施文件。
- python/sglang/srt/layers/attention/aiter_backend.py (模块 注意力层; 类别 source; 类型 core-logic; 符号 mla_fp8_prefill_attn) : 适配新 AITER API 签名的关键源码文件, 改动虽小但直接影响 FP8 MLA 预填充路径的正确性。

关键符号: mla_fp8_prefill_attn

关键源码片段

python/sglang/srt/layers/attention/aiter_backend.py

适配新 AITER API 签名的关键源码文件，改动虽小但直接影响 FP8 MLA 预填充路径的正确性。

```
# python/sglang/srt/layers/attention/aiter_backend.py
# 在 mla_fp8_prefill_attn 函数内，调用 mla_reduce_v1 时新增参数
# Prefill PS 路径不需要 split cap; 0 保持 AITER 默认 reduce 尺寸
mla_reduce_v1(
    logits,
    attn_lse,
    reduce_indptr,
    reduce_final_map,
    reduce_partial_map,
    tile_q,
    0, # num_kv_splits: 0 表示使用默认 CU 基数 reduce 分片
    output,
    final_lse,
)
```

评论区精华

讨论主要在 Issue 评论中进行，作者 [bingxche](#) 详细说明了额外参数 0 的来源：上游 AITER PR [ff61bb399](#) (Fix HK MLA decode fwd) 新增了 `num_kv_splits` 参数，prefill 路径无需 split cap，因此 0 保持默认行为。

- 暂无高价值评论线程

风险与影响

- 风险：

1. 回归风险：AITER_COMMIT_DEFAULT 的更新涉及多个构建阶段，若新 AITER 版本存在兼容性问题，可能影响所有 ROCm Docker 镜像的构建及运行时。
2. API 适配风险：mla_reduce_v1 的参数变化仅通过 0 简单适配，如果未来版本对 num_kv_splits 的处理逻辑发生改变，可能需再次调整。
3. 覆盖范围：变更仅涉及 AMD GPU 平台，对 NVIDIA 等后端无影响。

- 影响：

- 用户影响：使用 ROCm Docker 镜像的 AMD GPU 用户将自动获得新版本 AITER 的修复与特性，同时推理不受 API 签名变更影响。
- 系统影响：Docker 构建可能因上游 commit 不存在或新版本引入的问题而失败。但作者已确认 [9127c94](#) 存在于 ROCm/AITER origin/main。
- 团队影响：后续维护需跟随上游 AITER 版本更新，保持 API 一致性。
- 风险标记：外部依赖变更，缺少测试覆盖

关联脉络

- PR #29827 [Doc] Tiny update dsv4 doc: 同属 AMD 相关变更, DSv4 文档更新可能涉及 AITER 使用说明。