

PR #49361 完整报告

vllm-project/vllm

[ROCm]: bump AITER to 0.1.19

合并时间: 2026-07-31 08:15

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49361>

执行摘要

- 一句话: AITER 依赖升级至 v0.1.19
- 推荐动作: 值得快速浏览以了解 ROCm 依赖升级流程; 核心关注点是 AITER 0.1.19 的 gpt-oss 回归是否已在后续 PR 修复。若使用 ROCm 平台, 阅读时应结合 test_gpt_oss.py 与 AITER 内核变更记录。

功能与动机

PR body 仅为模板, 动机来自提交信息“Bump to aiter 0.1.19”与评审讨论: 跟随 AITER 上游发布节奏, 获取对 gfx950 等新平台的内核改进与修复。fxmarty-amd 的反馈进一步指出, AITER bump 到 0.1.19 后 gpt-oss 量化测试出现内存访问错误。

实现拆解

实现拆解

1. 修改版本引用: 在 docker/Dockerfile.rocm_base 中, 将 ARG AITER_BRANCH 从 v0.1.16.post5 改为 v0.1.19 (提交历史上先改到 v0.1.18, 随后改到 v0.1.19)。这是全仓库唯一的变更, AITER_REPO 固定指向 <https://github.com/ROCm/aiter.git>, 因此镜像构建时会拉取对应分支的最新代码。
2. 同步主分支: 作者通过 5 次 merge main (其中 AndreasKaratzas 参与同步) 保持分支与主干一致, 未引入其他代码冲突。
3. 依赖 CI 验证: 依靠 amd-ci 与标准 CI 覆盖验证, 无新增单元测试或文档; MI355 上的部分失败因 CI 从 DinD 迁移导致, 重跑后通过, 剩余失败被认为与 nightly 同源。
4. 回归暴露: 现有 ROCm 测试集覆盖不足, tests/models/quantization/test_gpt_oss.py 的量化用例在 MI355 上触发 memory access fault, 表明 AITER 0.1.19 可能引入 MXFP4_FP8 MoE 内核回归。

关键文件:

- docker/Dockerfile.rocm_base (模块 部署镜像; 类别 infra; 类型 infrastructure): 唯一变更文件, 定义 ROCm 基础镜像中 AITER 版本引用, 直接决定所有 ROCm 构建使用的内核库版本。

关键符号: 未识别

关键源码片段

docker/Dockerfile.rocm_base

唯一变更文件，定义 ROCm 基础镜像中 AITER 版本引用，直接决定所有 ROCm 构建使用的内核库版本。

```
# ROCm 基础镜像依赖版本声明 (docker/Dockerfile.rocm_base)
ARG PYTORCH_AUDIO_BRANCH="v2.9.0"
ARG PYTORCH_AUDIO_REPO="https://github.com/pytorch/audio.git"
ARG FA_BRANCH="0e60e394"
ARG FA_REPO="https://github.com/Dao-AILab/flash-attention.git"
# AITER 是 AMD 定制推理内核库，这里从 v0.1.16.post5 滚动升级到 v0.1.19,
# 为 gfx950 等新平台带来内核改进；固定分支而非哈希，升级不会自动漂移
ARG AITER_BRANCH="v0.1.19"
ARG AITER_REPO="https://github.com/ROCm/aiter.git"
ARG MORI_BRANCH="v1.1.0"
ARG MORI_REPO="https://github.com/ROCm/mori.git"
```

评论区精华

fxmarty-amd: AITER bump to 0.1.19 causes memory access fault in the test `test_gpt_oss_attention_quantization[amd/gpt-oss-20b-WFP8-AFP8-KVFP8-0.89-1]` on MI355. Likely MXFP4_FP8 MOE.

AndreasKaratzas: This looks good, only ones that I would double check are... But I think that they are also in our nightly. (确认除 DinD 迁移导致的失败外，其余失败与 nightly 一致)

dllehr-amd: Lets rock out like it's v0.1.19! (批准合并)

- AITER 0.1.19 导致 gpt-oss 量化测试内存访问故障 (correctness): 疑似 MXFP4_FP8 MoE 内核回归，需在后续 PR 中修复；本 PR 合并时该问题尚未解决。
- MI355 CI 从 DinD 迁移导致测试失败 (other): 失败归因于 CI 基础设施迁移，重跑后通过，不是 AITER 0.1.19 本身的问题。

风险与影响

- 风险：
 - gpt-oss 回归：fxmarty-amd 明确报告 AITER 0.1.19 在 MI355 上导致 `test_gpt_oss.py` 量化测试 memory access fault，疑似 MXFP4_FP8 MOE 内核问题，影响使用该量化配置的模型。
 - ROCm 全平台影响：AITER 版本跳跃改变了 ROCm 镜像中 MLA、MoE、sampler 等内核行为，可能引发精度或性能变化，也可能影响 gfx950 等新平台。
 - 测试覆盖不足：仅一行依赖升级，无配套测试与基准对比，回归范围无法自动隔离，只能依赖现有 CI 发现。

- CI 基础设施耦合：MI355 的 DinD 迁移导致部分失败，干扰了回归判断，存在环境不确定性。
- 影响：影响所有使用 ROCm 基础镜像的用户与 CI，范围取决于 AITER 0.1.19 内核改动；当前证据指向 gpt-oss (MXFP4_FP8 MOE) 存在内存访问故障，属于高影响但具体的回归点。合并后团队需跟踪后续修复 PR，避免量化配置用户踩坑。
- 风险标记：gpt-oss 测试内存访问故障，ROCM 平台回归风险，无配套测试

关联脉络

- PR #50328 [CI/Build][AMD] Install triton_kernels via CMake: 同属 ROCm 构建与 gpt-oss MoE 内核相关，gpt-oss 量化故障与其内核选择可能存在耦合。
- PR #50517 [ROCM][CI] Update Transformers AR+RMS fusion expectation: 同为 ROCm 内核依赖变化后的测试期望调整，AITER 版本升级常伴随此类后续 PR。