

PR #49737 完整报告

vllm-project/vllm

[ROCm][Docker] Drop MORI_GPU_ARCHS so MoRI autodetects the device arch

合并时间: 2026-07-25 05:14

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49737>

执行摘要

- 一句话: 移除硬编码的 GPU 架构环境变量以修复 MoRI 启动崩溃
- 推荐动作: 该 PR 改动简洁、目标明确, 建议合并。值得关注的是 MOI 自动检测架构的设计逻辑, 避免类似的环境变量覆盖陷阱。

功能与动机

PR body 明确指出: MoRI 在 gfx950 主机上因 MORI_GPU_ARCHS 被设为首个支持的架构 gfx942 而加载错误的代码对象, 导致启动崩溃。报错日志显示 device kernel image is invalid。

实现拆解

1. 编辑 docker/Dockerfile.rocm_base 第 33 行: 删除 ENV MORI_GPU_ARCHS=gfx942;gfx950, 改为注释 # Note: Do not set MORI_GPU_ARCHS here, it is automatically inferred at runtime。
2. 使得 MoRI 的 mori.jit.config.detect_gpu_arch() 不再被环境变量覆盖, 而是自动从当前 GPU 设备获取正确架构。
3. 该变更只影响 Docker 构建时的环境变量设置, 不涉及任何源码逻辑改动。

关键文件:

- docker/Dockerfile.rocm_base (模块 部署脚本; 类别 infra; 类型 infrastructure): 删除了 ENV MORI_GPU_ARCHS=gfx942;gfx950, 解决 MoRI 因 GPU 架构硬编码导致的启动崩溃。

关键符号: 未识别

关键源码片段

docker/Dockerfile.rocm_base

删除了 ENV MORI_GPU_ARCHS=gfx942;gfx950, 解决 MoRI 因 GPU 架构硬编码导致的启动崩溃。

```
# 删除 ENV MORI_GPU_ARCHS=gfx942;gfx950 以防止 MoRI 在运行时错误加载 gfx942 的代码对象
# MoRI 会自动检测 GPU 架构
ENV PYTORCH_ROCM_ARCH=${PYTORCH_ROCM_ARCH}
ENV AITER_ROCM_ARCH=gfx942;gfx950
```

```
# Note: Do not set MORI_GPU_ARCHS here, it is automatically inferred at runtime
# Required for RCCL in ROCm7.1
ENV HSA_NO_SCRATCH_RECLAIM=1
```

评论区精华

无 review 评论或额外讨论。PR 由 tjtanaa 审核通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅移除一行环境变量设置，使 MoRI 回退到自动检测行为，与官方推荐用法一致。如果后续 base image 中缺少 MoRI 自动检测所需的运行时库，可能会导致 MoRI 无法正确识别架构，但这种情况在标准 ROCm Docker 环境中不会出现。
- 影响：影响范围限定于使用 ROCm Docker 镜像的 CI 和用户部署。修复了 MoRI 在 gfx950 设备上的关键启动崩溃，使得 ROCm 上的 Mamba2/SSM 等依赖 MoRI 的推理功能恢复正常。无其他功能影响。
- 风险标记：低风险

关联脉络

- PR #44374 Fix MoRI for ROCm: PR body 提及此 PR 是对 #44374 的精简替代，由于基础镜像不兼容问题无法使用该 PR 中的 wheel 方案。