

PR #46386 完整报告

vllm-project/vllm

Run DeepSeek-V2-Lite prefetch-offload eval eager on ROCm

合并时间: 2026-06-25 01:25

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46386>

执行摘要

- 一句话: ROCm 上强制 eager 模式避免 HIP graph 解码异常
- 推荐动作: 该 PR 是典型的临时规避方案, 逻辑简单清晰。值得关注的是其根因分析过程展示了如何通过 bisect 定位 HIP graph 回归, 以及如何用环境检测实现平台差异化行为。维护者应注意在 ROCm 7.14 后回退。

功能与动机

MI300X 上 `deepseek-v2-lite-prefetch-offload-accuracy` 夜间测试约 1/5 服务器产生 ~45% 垃圾解码 token, 导致 GSM8K 准确率低于 0.25 阈值。作者 bisect 发现 ROCm CLR 7.2.0/7.2.2/7.2.3 存在 HIP graph 捕获 / 回放回归 (问题已在 ROCm/clr#279 跟踪), 而 eager 模式始终正确。

实现拆解

1. 检测 ROCm 平台: 在 `.buildkite/scripts/scheduled_integration_test/deepseek_v2_lite_prefetch_offload.sh` 中, 通过检查 `rocm-smi`、`amd-smi`、`/opt/rocm` 目录或 `ROCM_PATH` 环境变量判断是否为 ROCm 环境。
2. 添加 `--enforce-eager` 标志: 检测到 ROCm 时, 向 `EXTRA_ARGS` 数组追加 `--enforce-eager`, 使 `vllm serve` 命令在 ROCm 上强制使用 eager 模式, 绕过 HIP graph 的回放问题。
3. 添加注释与 TODO: 在代码中添加注释说明原因和跟踪链接, 并标记 TODO 要求在 ROCm 7.14 修复后恢复。
4. 无其他变更: 仅修改了一个测试脚本, 未改动 vLLM 核心代码或 CUDA 相关路径。

关键文件:

- `.buildkite/scripts/scheduled_integration_test/deepseek_v2_lite_prefetch_offload.sh` (模块 测试脚本; 类别 test; 类型 test-coverage): 唯一的变更文件, 在 ROCm 环境下添加 `--enforce-eager` 标志以绕过 HIP graph 解码异常。

关键符号: 未识别

关键源码片段

.buildkite/scripts/scheduled_integration_test/deepseek_v2_lite_prefetch_offload.sh

唯一的变更文件，在 ROCm 环境下添加 `--enforce-eager` 标志以绕过 HIP graph 解码异常。

ROCm 平台检测：优先使用常见命令或目录判断 # 详见 <https://github.com/ROCm/clr/issues/279> # TODO(aarushjain29): ROCm 7.14 修复后移除此逻辑

```
if command -v rocm-smi &> /dev/null || command -v amd-smi &> /dev/null || [[ -d /opt/rocm ]] || [[ -n "${ROCM_PATH:-}" ]]; then echo "ROCm platform detected: adding --enforce-eager to avoid HIP-graph decode corruption" EXTRA_ARGS+=(--enforce-eager) fi
```

该片段位于 `EXTRA_ARGS` 构建块内，随后与 `vllm serve` 命令合并，确保仅在 ROCm 上强制 `eager` 模式。

评论区精华

审核人 AndreasKaratzas 批准了该 PR，并提醒在 ROCm 7.14 发布后回退此更改。无其他 review 评论或争议。

- 回退时机 (other): PR 已合并，需要在 ROCm 7.14 发布后回退此变更。

风险与影响

- 风险：低风险。变更仅作用于 ROCm 平台的测试脚本，且只在 ROCm 环境下生效，CUDA 路径完全不变。强制 `eager` 模式会降低解码吞吐量，但测试超时窗口仍然满足。主要风险是后续忘记回退，但代码中有明确的 TODO 注释。
- 影响：影响范围极小：仅影响 DeepSeek-V2-Lite prefetch-offload 测试在 ROCm 平台的运行方式。`eager` 模式牺牲部分性能，但确保了准确率的确定性，修复了夜间测试的不稳定性。对其他模型、功能或平台无影响。
- 风险标记：临时性修复，低风险

关联脉络

- PR #46535 [Model Runner V2][MM] Support EVS: 同为 vLLM 主仓库的变更，但无直接功能关联。