

PR #46636 完整报告

vllm-project/vllm

[ROCm] Begin Deprecation Window for CUDA_VISIBLE_DEVICES on ROCm

合并时间: 2026-06-25 04:40

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46636>

执行摘要

- 一句话: 废弃 ROCm 上的 CUDA_VISIBLE_DEVICES 环境变量
- 推荐动作: 可快速合入, 属于低风险废弃流程。值得在内部沟通中提醒 ROCm 用户注意该变化。

功能与动机

CUDA_VISIBLE_DEVICES 在 ROCm 上支持不佳, 尤其在分布式环境 (如 Ray) 中反复引发 bug (参见 PR #45998、#35069、#33305、#33308)。弃用此变量可减少后续兼容性问题, 推动用户迁移到正确的 HIP_VISIBLE_DEVICES。

实现拆解

1. 在 vllm/platforms/rocm.py 的 _sync_hip_cuda_env_vars() 函数中, 在原有的环境变量同步逻辑之前, 增加对 CUDA_VISIBLE_DEVICES 的检查。
2. 如果 cuda_val 不为 None, 则调用 logger.warning_once 打印一次弃用警告, 警告内容明确提示该变量将在 vLLM v0.26.0 中被移除, 并建议使用 HIP_VISIBLE_DEVICES。
3. 警告使用 scope="process" 确保每个进程只输出一次, 避免日志洪泛。
4. 原有同步逻辑保持不变: 若两个变量同时存在且不一致则报错, 否则同步两者。

关键文件:

- vllm/platforms/rocm.py (模块 平台层; 类别 source; 类型 core-logic; 符号 _sync_hip_cuda_env_vars): 唯一变更文件, 增加弃用警告日志, 推动用户从 CUDA_VISIBLE_DEVICES 迁移到 HIP_VISIBLE_DEVICES。

关键符号: _sync_hip_cuda_env_vars

关键源码片段

vllm/platforms/rocm.py

唯一变更文件, 增加弃用警告日志, 推动用户从 CUDA_VISIBLE_DEVICES 迁移到 HIP_VISIBLE_DEVICES。

```
def _sync_hip_cuda_env_vars():  
    """Ensure HIP_VISIBLE_DEVICES and CUDA_VISIBLE_DEVICES are consistent.  
    Treats empty string as unset. Raises on genuine conflicts."""
```

```

hip_val = os.environ.get("HIP_VISIBLE_DEVICES") or None
cuda_val = os.environ.get("CUDA_VISIBLE_DEVICES") or None

# 新增：当检测到 CUDA_VISIBLE_DEVICES 时，打印弃用警告，
# 鼓励用户迁移到 HIP_VISIBLE_DEVICES（每个进程仅一次）。
if cuda_val is not None:
    logger.warning_once(
        "Using CUDA_VISIBLE_DEVICES on ROCm is deprecated and support "
        "will be removed in vLLM v0.26.0. Please use HIP_VISIBLE_DEVICES "
        "instead.",
        scope="process",
    )

# 以下为原有同步逻辑，保持不变
if hip_val is not None and cuda_val is not None:
    if hip_val != cuda_val:
        raise ValueError(
            f"Inconsistent GPU visibility env vars: "
            f"HIP_VISIBLE_DEVICES='{hip_val}' vs "
            f"CUDA_VISIBLE_DEVICES='{cuda_val}'. "
            f"Please set only one, or ensure they match."
        )
    elif hip_val is not None:
        os.environ["CUDA_VISIBLE_DEVICES"] = hip_val
    elif cuda_val is not None:
        os.environ["HIP_VISIBLE_DEVICES"] = cuda_val

```

评论区精华

该 PR 无 review 讨论，仅由 claude[bot] 自动评论和 tjtnaa 批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅在检测到已弃用的环境变量时增加一行警告日志，不影响任何逻辑路径。原有同步行为完全保留。未添加测试，但属于简单日志添加，可接受。
- 影响：对用户：当前使用 CUDA_VISIBLE_DEVICES 的 ROCm 用户会在进程启动时看到一次警告，需在 v0.26.0 前迁移到 HIP_VISIBLE_DEVICES。对系统：无性能或功能影响。对团队：属渐进式废弃，为后续彻底移除做准备。
- 风险标记：暂无

关联脉络

- PR #45998 示例：CUDA_VISIBLE_DEVICES 相关 bug：在 PR body 中被引用，是推动此废弃的直接原因之一。
- PR #35069 示例：CUDA_VISIBLE_DEVICES 相关 bug：在 PR body 中被引用，是推动此废弃的直接原因之一。

- PR #33305 示例: CUDA_VISIBLE_DEVICES 相关 bug: 在 PR body 中被引用, 是推动此废弃的直接原因之一。
- PR #33308 示例: CUDA_VISIBLE_DEVICES 相关 bug: 在 PR body 中被引用, 是推动此废弃的直接原因之一。
- PR #39448 示例: 后续跟进 PR: 在 PR body 中标记为后续跟进, 可能是废弃流程的下一步。