PR #23960 完整报告

sgl-project/sglang

ci: clean up stale-CUDA mooncake variant in install_extra_deps

合并时间：2026-04-29 10:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23960

执行摘要

- 一句话：清理 CI 中淘汰的 mooncake 变体残留
- 推荐动作：值得合并，解决了一个隐蔽的 CI 环境污染问题。注意后续若有新 mooncake 变体需要更新此逻辑。

功能与动机

PR #23119 为 CUDA 13 引入了 mooncake-transfer-engine-cuda13 包，但基于旧 base commit 的 PR 仍会安装旧包 mooncake-transfer-engine。两个包共享同一 python 包目录，导致旧变体残留在运行器上。全集群审计发现约 60 个容器受到污染，其中 17 个由 11 个不同旧 PR 引起。

实现拆解

1. 确定对立变体：在 scripts/ci/cuda/ci_install_dependency.sh 的 install_extra_deps 函数中，根据 CU_MAJOR 变量同时设置 MOONCAKE_PKG（正确变体）和 MOONCAKE_STALE_PKG（对立变体）。若 CU_MAJOR=13，则正确包为 mooncake-transfer-engine-cuda13，淘汰包为 mooncake-transfer-engine；否则相反。
2. 最佳卸载淘汰包：在安装正确包之前，使用 pip show 检查淘汰包是否存在，若存在则以 pip uninstall 卸载（错误忽略）。这删除所有共享文件。
3. 强制重装正确包：由于卸载也删除了正确包的文件，使用 --force-reinstall --no-deps 重新安装正确包，确保文件被恢复。然后用原命令安装正确包及依赖。
4. 对称处理：该方法对称处理 CUDA 13→12 和 12→13 回退场景。

关键文件：

- scripts/ci/cuda/ci_install_dependency.sh（模块 CI 脚本；类别 infra；类型 infrastructure）：PR 的唯一变更文件，在 install_extra_deps 中添加了预卸载对立 CUDA 变体的逻辑，并强制重装正确版本。

关键符号：install_extra_deps

关键源码片段

scripts/ci/cuda/ci_install_dependency.sh

PR 的唯一变更文件，在 install_extra_deps 中添加了预卸载对立 CUDA 变体的逻辑，并强制重装正确版本。

```
install_extra_deps() {
  if [ "$CU_MAJOR" = "13" ]; then
    MOONCAKE_PKG="mooncake-transfer-engine-cuda13==0.3.10.post2"
    MOONCAKE_STALE_PKG="mooncake-transfer-engine"
    EXTRA_NVIDIA_SPECS="nvidia-cuda-nvrtc"
  else
    MOONCAKE_PKG="mooncake-transfer-engine==0.3.10.post2"
    MOONCAKE_STALE_PKG="mooncake-transfer-engine-cuda13"
    EXTRA_NVIDIA_SPECS="nvidia-cuda-nvrtc-cu12"
  fi

  # 两个变体拥有相同的 mooncake/ 包文件和 bin/ 脚本
  # （如 mooncake_master 等）。卸载淘汰变体会删除共享文件，
  # 导致存活变体的 RECORD 引用缺失，因此使用 --force-reinstall
  # 恢复文件 —— 否则 pip 会因为 "already satisfied" 而跳过。
  if pip show ${MOONCAKE_STALE_PKG} >/dev/null 2>&1; then
    $PIP_UNINSTALL_CMD ${MOONCAKE_STALE_PKG} $PIP_UNINSTALL_SUFFIX || true
    $PIP_CMD install ${MOONCAKE_PKG} --force-reinstall --no-deps $PIP_INSTALL_SUFFIX
  fi

  $PIP_CMD install ${MOONCAKE_PKG} ${EXTRA_NVIDIA_SPECS} py-spy scipy huggingface_
  hub[hf_xet] pytest $PIP_INSTALL_SUFFIX
  ...
}
```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：无实质风险：该变更仅影响 CI 安装流程，且使用 || true 忽略卸载 / 重装错误，不会阻塞 CI。少量增加安装时间（额外 pip 操作），但影响几乎可忽略。
- 影响：直接修复 CI 运行器上 mooncake 变体残留问题，确保所有 PR 使用正确的 mooncake 版本运行测试。对用户无影响；对团队降低了偶发失败的排查成本。影响范围仅限于 CI 基础设施。
- 风险标记：少量安装时间增加

关联脉络

- PR #23119 Add CUDA 13 conditional for mooncake-transfer-engine: 该 PR 引入了 mooncake-transfer-engine-cuda13 包，是当前 PR 要修复的残留问题的根源。当前 PR 的卸载逻辑直接针对该 PR 带来的变体冲突。