PR #21711 完整报告

sgl-project/sglang

Remove flashinfer wheel cache cleanup that deletes other versions

合并时间：2026-03-31 07:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21711

执行摘要

本 PR 移除了 FlashInfer JIT 缓存脚本中的清理逻辑，以避免不同 CI 作业版本冲突导致的重复下载大文件问题。变更简单但显著提升 CI 效率，减少了带宽消耗和构建超时风险，但引入了磁盘使用增长和版本选择不确定性的潜在风险。建议关注未解决的review建议，以备未来优化。

功能与动机

此 PR 旨在解决 CI 构建中的缓存冲突问题。原脚本在下载 FlashInfer wheel 时，会自动删除缓存中不匹配当前所需 Python 版本的其他 wheel 文件。当多个 CI 作业或 PR 需要不同版本时，这种清理逻辑导致缓存频繁被清空，迫使每次重新下载 1.6-1.9GB 的大文件。例如，在 RTX 5090 runner 上因带宽限制（约 3MB/s），下载超时导致构建失败。PR body 引用了具体失败链接，强调优化构建流程以减少下载开销。

实现拆解

仅修改一个文件：scripts/ci/cuda/ci_download_flashinfer_jit_cache.sh。关键改动是删除以下行：

```
find "${FLASHINFER_CACHE_DIR}" -name "flashinfer_jit_cache-*.whl" ! -name "flashinfer_jit_cache-${FLASHINFER_PYTHON_REQUIRED}*" -type f -delete 2>/dev/null || true
```

移除后，脚本不再主动清理旧版本 wheel，允许缓存目录中保留多个版本共存。但 wheel 选择逻辑未变，仍使用模式 flashinfer_jit_cache-${FLASHINFER_PYTHON_REQUIRED}*.whl 和 head -n 1 选取第一个匹配文件。

评论区精华

gemini-code-assist[bot] 在 review 中提出两个核心讨论点：

1. wheel 选择逻辑的鲁棒性问题：

建议包含 CUDA 版本并使用 sort -V | tail -n 1 确保选择正确 wheel。当前模式宽泛且 head -n 1 是非确定性的，可能导致跨环境选取错误版本。

2. 磁盘清理策略的风险：

建议添加时间基清理（如 30 天）防止磁盘无限增长，每个 wheel 约 1.2GB，长期运行可能导致存储问题。

PR 作者未响应这些建议，Fridge003 直接批准合并，表明决策是优先解决缓存冲突，而将优化建议留待后续处理。

风险与影响

风险分析：

- 磁盘使用可能无限增长，因为旧版本 wheel 未被清理，对持久化 runner 构成存储压力。
- 当前 wheel 选择逻辑可能错误选取不匹配 CUDA 版本的 wheel，由于模式仅基于 Python 版本且使用非确定性 head -n 1，存在跨 runner 共享缓存时的正确性风险。

影响评估：

- 积极影响：大幅减少 CI 构建时间，避免大文件重复下载，提升开发效率和 CI 稳定性，尤其在带宽受限环境下。
- 负面影响：增加磁盘占用，需监控缓存大小；对最终用户无直接影响，但内部 CI 成本和维护复杂性可能增加。

关联脉络

与此 PR 相关的历史 PR 包括 #21682（放松 CI 测试阈值），两者均为 CI 流程优化，共同目标提升构建可靠性和效率。结合仓库近期历史，sglang 项目在多平台（如 NPU、AMD）支持中频繁涉及 CI 调整，本次变更反映了对基础设施稳定性的持续改进趋势，未来可能需跟进 review 建议以完善缓存管理。