PR #21441 完整报告

sgl-project/sglang

Upgrade CI default CUDA version from 12.9 to 13.0

合并时间：2026-04-13 12:48

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21441

执行摘要

- 一句话：将 CI 默认 CUDA 版本从 12.9 升级到 13.0，以匹配 PyTorch 2.11 默认。
- 推荐动作：对于技术管理者和工程师，建议快速浏览此 PR 以了解 CUDA 版本升级的 CI 配置变更，重点关注 ci_install_dependency.sh 中的 wheel 安装逻辑和测试计划执行情况，确保 CI 稳定后再推进 Docker 镜像更新。

功能与动机

根据 PR body，动机是 'Upgrade CI default CUDA from 12.9 to 13.0 to match Torch 2.11's default'。此外，所有 CI runner 已具备驱动 580+（支持 CUDA 13.0），B200 Novita 升级到驱动 590（支持 CUDA 13.1），为升级提供硬件支持。

实现拆解

实现涉及五个关键文件更新：1) .github/workflows/pr-test.yml：将 CI 构建矩阵中的 cuda-version 从 12.9 改为 13.0，并更新所有 wheel artifact 模式以匹配新版本。2) python/pyproject.toml：更新 cuda-python 依赖为 >=13.0，并将 torch 索引从 cu129 改为 cu130，确保包安装与 CUDA 13.0 兼容。3) scripts/ci/cuda/ci_download_flashinfer_jit_cache.sh：更新 CU_VERSION 引用从 cu129 到 cu130，保持脚本一致性。4) scripts/ci/cuda/ci_install_deepep.sh：移除针对 CUDA >12.8 时丢弃 sm_103 的 workaround，启用 Blackwell 架构支持，优化 DeepEP 构建。5) scripts/ci/cuda/ci_install_dependency.sh：更新 CU_VERSION 为 cu130，并修复 sgl-kernel wheel 安装逻辑以处理 +cu130 后缀文件名，避免安装失败。

关键文件：

- .github/workflows/pr-test.yml（模块 CI/Infrastructure）：核心 CI 工作流文件，定义构建矩阵和测试步骤，变更影响所有 CI 运行的 CUDA 版本和 artifact 模式。
- python/pyproject.toml（模块 Dependencies）：Python 项目依赖配置，更新 cuda-python 和 torch 索引，直接影响包安装和版本兼容性。
- scripts/ci/cuda/ci_install_dependency.sh（模块 CI/Infrastructure）：CI 依赖安装脚本，更新 CU_VERSION 和修复 sgl-kernel wheel 安装逻辑，关键于避免构建失败。
- scripts/ci/cuda/ci_install_deepep.sh（模块 CI/Infrastructure）：DeepEP 构建脚本，移除 workaround 并启用 sm_103 支持，优化 Blackwell 架构兼容性。

关键符号：未识别

评论区精华

Review 中没有具体评论，但提交历史显示基于 reviewer 请求移除了 Dockerfile 变更（提交消息：'Per reviewer request — test CI/script changes first before updating Docker images.'）。这表明 review 过程中采取了谨慎的测试策略，先验证 CI 脚本变更再更新 Docker 镜像，以减少风险。

- 测试策略优化与 Dockerfile 变更移除 (design): 移除了 Dockerfile 变更，专注于验证 CI 脚本升级，确保稳定性后再考虑镜像更新。

风险与影响

- 风险：风险包括：1) 兼容性问题：CUDA 13.0 可能不兼容某些现有 sgl-kernel 或依赖，如 torchaudio 版本冲突（在 ci_install_dependency.sh 中提及）。2) CI 构建失败：版本变更可能导致构建错误或测试失败，特别是 sgl-kernel wheel 安装逻辑修改（ci_install_dependency.sh）若文件名模式不匹配会失败。3) 依赖冲突：更新 torch 索引和 cuda-python 可能影响其他包安装，需确保所有环境一致。
- 影响：影响范围：1) CI 环境：所有使用默认 CUDA 版本的 CI runner 将切换到 13.0，影响构建和测试流程，可能提升性能或引入新特性支持。2) 开发者：需要确保本地开发环境与 CI 一致，可能需升级 CUDA 工具链以避免差异。3) 系统：sgl-kernel 构建将基于 CUDA 13.0，可能优化 Blackwell 架构兼容性，但需验证内核稳定性。影响程度中等，主要局限于基础设施和测试流程。
- 风险标记：依赖版本升级，CI 构建失败风险，兼容性风险

关联脉络

- PR #22727 Revert "Upgrade CI default CUDA version from 12.9 to 13.0": 此前尝试升级 CUDA 13.0 但遇到内核测试问题被回滚，本 PR 是重新尝试并修复了相关脚本问题。