PR #23733 完整报告

sgl-project/sglang

chore: bump sglang-kernel version to 0.4.1.post1

合并时间：2026-04-26 14:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23733

执行摘要

- 一句话：升级 sglang-kernel 至 0.4.1.post1 并恢复 hiCache 测试
- 推荐动作：建议快速浏览，重点关注测试恢复时的 CI 注册套件选择。版本升级模式可作为后续依赖同步的参考。

功能与动机

根据 PR body，对 sglang-kernel 的版本要求需与 sgl-kernel/pyproject.toml 中定义的一致。此外，在 PR #23119 中由于 CUDA 13 升级导致部分 hiCache 测试出现段错误（cudaMemcpyBatchAsync segfault），这些测试被临时移至 manual 目录。新版本 0.4.1.post1 已修复该问题，因此可重新激活这些测试。

实现拆解

1. 版本号更新：在 python/sglang/srt/entrypoints/engine.py 中的版本断言、python/pyproject.toml 的依赖声明、docker/Dockerfile 的构建参数中将 0.4.1 改为 0.4.1.post1。
2. 测试文件恢复：将 7 个测试文件从 test/manual/ 迁移到 test/registered/ 下的对应子目录，并删除文件头部的多行 TODO 注释（内容关于 cu13 升导致的段错误）。
3. CI 注册增强：对于 test/registered/hicache/test_hicache_storage_mooncake_backend.py，额外添加了 register_cuda_ci(est_time=236, suite="stage-b-test-2-gpu-large") 调用，将其注册到指定的 CI 套件；其余测试文件仅做路径移动和注释清理，保留了原有的 CI 注册语句。

关键文件：

- python/sglang/srt/entrypoints/engine.py（模块 引擎核心；类别 source；类型 core-logic）：核心版本断言所在，确保运行时 sglang-kernel 版本符合要求。
- python/pyproject.toml（模块 项目配置；类别 config；类型 configuration）：依赖声明文件，控制 Python 包安装时的版本约束。
- docker/Dockerfile（模块 容器构建；类别 infra；类型 infrastructure）：镜像构建时使用构建参数指定版本，保持一致性。
- test/registered/hicache/test_hicache_storage_mooncake_backend.py（模块 Mooncake 测试；类别 test；类型 rename-or-move）：测试文件从 manual 迁回 registered，并新增 CI 注册，是最重要的测试变更。

- test/registered/4-gpu-models/test_qwen35_hicache.py（模块 Qwen3.5 测试；类别 test；类型 rename-or-move）：Qwen3.5 hiCache 测试文件恢复，是重要的回归测试。
- test/registered/hicache/test_hicache_storage.py（模块 存储测试；类别 test；类型 rename-or-move）：通用存储测试恢复，覆盖基础 hiCache 功能。
- test/registered/hicache/test_hicache_storage_3fs_backend.py（模块 3FS 测试；类别 test；类型 rename-or-move）：3FS 后端测试恢复，确保第三方存储兼容。
- test/registered/hicache/test_hicache_storage_file_backend.py（模块 文件存储测试；类别 test；类型 rename-or-move）：文件后端测试恢复，覆盖基础文件存储。
- test/registered/hicache/test_hicache_storage_runtime_attach_detach.py（模块 运行时测试；类别 test；类型 rename-or-move）：运行时挂载 / 卸载测试恢复，覆盖动态存储操作。
- test/registered/hicache/test_hicache_variants.py（模块 变体测试；类别 test；类型 rename-or-move）：hiCache 变体测试恢复，覆盖不同配置场景。

关键符号：未识别

关键源码片段

python/sglang/srt/entrypoints/engine.py

核心版本断言所在，确保运行时 sglang-kernel 版本符合要求。

```
# 位于 _set_envs_and_config() 函数中
if _is_cuda:
    assert_pkg_version(
        "sglang-kernel",
        "0.4.1.post1", # 从 0.4.1 升至 0.4.1.post1，对齐 sgl-kernel 版本
        "Please reinstall the latest version with `pip install sglang-kernel --force-reinstall`",
    )
```

评论区精华

本 PR 仅有一条来自 gemini-code-assist[bot] 的自动评论，确认变更一致正确，无实质性讨论。Kangyan-Zhou 在合并前于 Issue 中评论 /tag-and-rerun-ci 以触发 CI 重跑。

- 版本号变更验证 (other): 无问题

风险与影响

- 风险：风险较低。版本号提升可能导致旧版本残留，但 CI 依赖安装会覆盖。测试恢复可能因 CI 环境差异偶发失败，但测试已在 manual 模式验证。主要风险是 CUDA 13 兼容性修复是否完整，但测试覆盖增加了信心。
- 影响：对用户无直接影响，但后续安装 sglang-kernel 时会拉取 0.4.1.post1 版本。对开发 / 测试团队：hiCache 测试重新加入 CI，增强了回归覆盖，简化了测试维护（不再需要手动管理禁用状态）。
- 风险标记：版本兼容性风险（低），测试覆盖恢复

关联脉络

- PR #23720 chore: bump sgl-kernel version to 0.4.1.post1: 该 PR 先更新了 sgl-kernel 自身版本，本 PR 同步更新 sglang 对 sglang-kernel 的依赖版本。
- PR #23119 （推测）CUDA 13 升级导致测试禁用：该 PR 因 CUDA 13 升级将 hiCache 测试移至 manual，本 PR 修复后重新启用。