

PR #30874 完整报告

sgl-project/sglang

chore: bump tilelang to 0.1.12

合并时间: 2026-08-20 08:01

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30874>

执行摘要

- 一句话: 升级 tilelang 至 0.1.12 修复 FlashInfer 融合初始化
- 推荐动作: 值得合并, 但建议关注升级后 FlashInfer 融合的实测性能。可精读 [tile-ai/tilelang#2541](#) 了解根因。

功能与动机

tilelang 0.1.11 打包的 lib cudart_stub.so 缺少 cudaDeviceReset 符号, 导致 FlashInfer 的 trtllm allreduce 融合 workspace 初始化失败, 错误信息为 'undefined symbol: cudaDeviceReset', 从而永久禁用 flashinfer allreduce fusion。该问题已在 [tile-ai/tilelang#2541](#) 修复并随 0.1.12 发布。

实现拆解

1. 修改 python/pyproject.toml 中 tilelang 版本号从 0.1.11 为 0.1.12。
2. 合并 main 分支并触发 CI, 确保新版本与当前代码兼容。
3. 无需其他代码改动, 纯依赖升级。

关键文件:

- python/pyproject.toml (模块 依赖配置; 类别 config; 类型 configuration) : 升级 tilelang 依赖版本, 修复 FlashInfer 初始化问题

关键符号: 未识别

评论区精华

维护者 hnyls2002 指出兼容性已由 SiriusNEO 在 [tile-ai/tilelang#2795](#) 修复, 建议升级并 @ 相关人员。未发现其他讨论。

- 暂无高价值评论线程

风险与影响

- 风险: 依赖升级可能引入 API 变更或行为差异, 但 tilelang 0.1.12 为补丁版本, 且 CI 已通过。需关注 FlashInfer 融合在不同后端的稳定性和性能是否受影响。

- 影响：影响使用 tilelang 和 FlashInfer 的用户，修复了 allreduce 融合无法初始化的问题，提升了多卡通信效率。对团队而言，消除了依赖中的潜在缺陷，降低了后续集成风险。
- 风险标记：依赖升级，FlashInfer 兼容性

关联脉络

- PR #2541 [CUDA][ROCm] Rename GPU stub library artifacts: tilelang 0.1.12 修复了 0.1.11 中 libcudart_stub.so 缺失符号的问题，该问题由本 PR 升级解决。