

PR #33678 完整报告

sgl-project/sglang

chore: bump sgl-kernel version to 0.4.6

合并时间: 2026-08-06 08:46

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33678>

执行摘要

- 一句话: sgl-kernel 版本 0.4.5 统一升级至 0.4.6
- 推荐动作: 不值得精读, 属于自动发版流程的一部分, 可快速审查版本号一致性后合入。值得关注的细节是: CUDA、CPU、MUSA、ROCm 四个平台的构建配置与运行时版本常量必须同步更新, 可在已发布的 0.4.6 包上核对 `sgl_kernel.__version__` 与 `importlib.metadata.version('sglang-kernel')` 是否一致。

功能与动机

PR 标题与 body 明确指出目标是 "bump the sgl-kernel version to 0.4.6 across all relevant files"。该 PR 由 GitHub Actions 自动生成 (🔗 Generated with GitHub Actions), 没有关联 Issue 或人工补充说明, 属于内核发版流水线的一部分: 在 sgl-kernel 0.4.6 发布时, 需要同步 4 个平台的构建配置和运行时版本号, 确保安装包元数据与 `sgl_kernel.__version__` 保持一致。

实现拆解

1. 更新 CUDA 主构建配置: 修改 `python/sglang/kernels/aot/pyproject.toml` 中 `[project].version` 为 0.4.6, 该值决定 sglang-kernel wheel 的包版本, 影响依赖解析与安装元数据。
2. 同步其他平台构建配置: `pyproject_cpu.toml` (包名 `sglang-kernel-cpu`)、`pyproject_musa.toml`、`pyproject_rocm.toml` 中的版本号一并从 0.4.5 改为 0.4.6, 保证不同硬件平台产出的包版本对齐。
3. 更新运行时版本常量: `python/sglang/kernels/aot/python/sgl_kernel/version.py` 中的 `__version__` 同步更新, 使运行时可查询的版本与包元数据一致。
4. 无功能与测试配套: 本次仅版本号改动, 不涉及编译逻辑、API 变更或测试; PR body 中 CI 状态记录为 `:no_entry_sign:` (PR Test 与 PR Test Extra), 但改动本身没有回归面。

关键文件:

- `python/sglang/kernels/aot/python/sgl_kernel/version.py` (模块 内核包; 类别 `source`; 类型 `core-logic`): 运行时版本常量, 是所有平台构建共享的版本来源; 若与打包版本不一致, 会出现元数据与运行时版本不匹配的问题。
- `python/sglang/kernels/aot/pyproject.toml` (模块 构建配置; 类别 `config`; 类型 `configuration`): CUDA 平台主构建配置, `[project] version` 决定 sglang-kernel wheel 的

包版本和依赖解析行为。

- python/sglang/kernels/aot/pyproject_cpu.toml (模块 构建配置; 类别 config; 类型 configuration) : CPU 平台构建配置 (包名 sglang-kernel-cpu) , 版本号需与主包保持一致, 否则 CPU wheel 安装版本会漂移。
- python/sglang/kernels/aot/pyproject_musa.toml (模块 构建配置; 类别 config; 类型 configuration) : 摩尔线程 MUSA 平台构建配置, 使用 setuptools backend, 版本号需与其它平台保持一致。
- python/sglang/kernels/aot/pyproject_rocm.toml (模块 构建配置; 类别 config; 类型 configuration) : AMD ROCm 平台构建配置, 使用 setuptools backend, 与 AMD CI 通道相关, 版本号需保持一致。

关键符号: 未识别

评论区精华

本 PR 没有任何 review 评论、讨论线程或 issue 评论。唯一人工痕迹是合并者 Fridge003 执行了两次 Merge branch 'main' 来同步基线, 说明这是一条由 bot 发起、人工合并的常规发版 PR。

- 暂无高价值评论线程

风险与影响

- 风险:
 - 版本号一致性 (低风险) : 5 处版本号必须同步, 若某平台遗漏会导致对应 wheel 的安装版本与 `__version__` 不匹配。当前所有文件已同步为 0.4.6, 风险集中在未来手工升级时。
 - CI 状态待确认 (低风险) : PR body 中 PR Test 与 PR Test Extra 状态快照均为: `no_entry_sign:`, 考虑到改动仅为常量升级且已合入, 推测后续有补测; 如果依赖版本稳定性, 可等待官方 0.4.6 发布后再使用。
 - 兼容性影响 (很低) : 版本号不改变任何内核行为; 若下游存在 `hard pin` 到 `sgl-kernel==0.4.5` 的依赖约束, 需要同步更新, 但仓库内未见此类约束证据。
 - 影响: 对用户和下游的影响是安装 `sgl-kernel 0.4.6` 时包元数据与运行时版本号保持一致; 对团队而言, 这是内核发版流程的版本载体, 后续发布与依赖解析以该版本号为基准。影响范围仅限 `sgl-kernel` 包元数据与版本常量, 不涉及运行时行为。
- 风险标记: 版本号一致性, 多平台同步, CI 状态待确认

关联脉络

- PR #33469 kernels: scalar scale A support for fp8_gemm: 同为 `sgl-kernel` 内核改动, 0.4.6 版本很可能汇聚了此类 fp8 GEMM 内核更新, 属于同一内核演进线。
- PR #33205 [Kernel] Unify BaseFusedOp and MultiPlatformOp dispatch: 重构了 `sgl-kernel` 算子分派体系, 与本次版本升级属于同一条内核基础设施演进脉络。
- PR #33618 Enable MoE deferred finalize by default and drop its expert_weights dtype workaround: 涉及 `kernels/jit/csrc/moe` 目录下的内核变更, 可能包含在 0.4.6 发布

范围内。