

PR #36433 完整报告

sgl-project/sglang

[NPU] Update sgl-kernel-npu version

合并时间: 2026-08-29 00:42

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36433>

执行摘要

- 一句话: 更新 sgl-kernel-npu 版本至 20260826
- 推荐动作: 该 PR 为常规维护性更新, 值得快速浏览以确认版本号一致, 无需深入精读。

功能与动机

PR 描述指出更新内核版本以与上游最新变更保持一致, 并提高与当前环境的兼容性。

实现拆解

1. 修改 .github/workflows/release-docker-npu-nightly.yml 和 .github/workflows/release-docker-npu.yml 中的 SGLANG_KERNEL_NPU_TAG 值, 从 2026.8.13 更新为 20260826, 以确保 Docker 构建使用新内核版本。
2. 修改 scripts/ci/npu/npu_ci_install_dependency.sh 中的 SGLANG_KERNEL_NPU_TAG 值, 保持 CI 环境与构建环境一致。
3. 触发 CI 重新运行 (通过 bot 命令), 确保更新后构建和测试通过。

关键文件:

- .github/workflows/release-docker-npu-nightly.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure): 更新夜间 NPU Docker 发布工作流中的内核版本, 确保构建产物与最新内核一致。
- .github/workflows/release-docker-npu.yml (模块 CI 工作流; 类别 infra; 类型 infrastructure): 更新正式 NPU Docker 发布工作流中的内核版本, 与 nightly 保持一致。
- scripts/ci/npu/npu_ci_install_dependency.sh (模块 CI 脚本; 类别 infra; 类型 infrastructure): 更新 NPU CI 依赖安装脚本中的内核版本, 确保 CI 测试使用与发布一致的新内核。

关键符号: 未识别

关键源码片段

scripts/ci/npu/npu_ci_install_dependency.sh

更新 NPU CI 依赖安装脚本中的内核版本, 确保 CI 测试使用与发布一致的新内核。

```
# scripts/ci/npu/npu_ci_install_dependency.sh
# 设置内核版本标签, 用于从 GitHub Releases 下载指定版本的 sgl-kernel-npu 包
```

SGLANG_KERNEL_NPU_TAG="20260826"

```
# 下载并解压 sgl-kernel-npu 安装包，使用该标签构建下载 URL
(cd sgl-kernel-npu && wget "${GITHUB_PROXY_URL:=}"https://github.com/sgl-project/sgl-kernel-npu/releases/download/${SGLANG_KERNEL_NPU_TAG}/sgl-kernel-npu-${SGLANG_KERNEL_NPU_TAG}-torch${PYTORCH_VERSION}-py311-cann9.0.0-${DEVICE_TYPE}-${arch}.zip" \
&& unzip ./sgl-kernel-npu-${SGLANG_KERNEL_NPU_TAG}-torch${PYTORCH_VERSION}-py311-cann9.0.0-${DEVICE_TYPE}-${arch}.zip)
```

评论区精华

无实质性代码审查讨论，审查由 bot 和合并者批准，评论仅包含 CI 重新运行的命令。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，仅涉及版本号更新，但需确保新版本内核与构建环境兼容，否则可能导致 NPU 推理失败。CI 中已确认构建和测试通过，但 PR Test (Extra) 曾失败后通过重跑通过。
- 影响：影响范围限于 NPU 的 Docker 镜像构建和 CI 依赖安装，对用户推理功能无直接影响，但新内核版本可能包含 bug 修复或特性，长期影响兼容性。
- 风险标记：依赖版本更新，CI 配置变更

关联脉络

- PR #36434 [AMD] Add ROCm 10 (gfx942 / gfx950) release images: 同为 CI/ 发布 workflow 变更，涉及依赖版本更新，可能关联基础设施演进。