PR #24176 完整报告

sgl-project/sglang

[CI] Publish nightly sglang wheel under both cu129 and cu130 indexes

合并时间：2026-05-01 07:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24176

执行摘要

- 一句话：nightly wheel 同时注册 cu129 和 cu130
- 推荐动作：建议合入。改动简洁、风险低，且提升了用户便利性。值得关注的是 update_nightly_whl_index.py 中错误处理的改进（从静默失败到显式抛出），这种模式值得在其他 CI 脚本中推广。

功能与动机

根据 PR body 和变更内容，目的是让用户可以通过 cu129 或 cu130 任一索引安装 sglang 的 nightly wheel，提高灵活性和兼容性。此前只发布到 cu129，cu130 用户需要额外操作。

实现拆解

1. 修改 CI 工作流配置文件：在 .github/workflows/release-pypi-nightly.yml 中的 release-nightly job 中添加 strategy.matrix，cuda_version 包含 ['129', '130']，并设置 max-parallel: 1 避免两个并行 runner 同时操作同一 gh-pages 分支导致冲突。同时移除已废弃的 cuda_version workflow_dispatch input。
2. 强化脚本错误处理：在 scripts/update_nightly_whl_index.py 中将原本的 print+return 的静默返回改为显式抛出异常：
   - 当 dist/ 目录不存在时抛出 FileNotFoundError
   - 当 dist/ 目录下没有 wheel 文件时抛出 RuntimeError
   - 移除原有的 try-except 包装，让哈希计算错误直接传播
3. 调整提交信息：在 release-nightly job 的 push 步骤中，将 git commit message 改为包含 cu${{ matrix.cuda_version }} 以区分两个索引的更新。

关键文件：

- .github/workflows/release-pypi-nightly.yml（模块 CI 工作流；类别 infra；类型 infrastructure）：定义 CI 工作流，添加 matrix 策略实现 dual-index 发布
- scripts/update_nightly_whl_index.py（模块 发布脚本；类别 source；类型 core-logic）：核心脚本增强错误处理，从静默失败改为显式异常

关键符号：update_wheel_index

关键源码片段

scripts/update_nightly_whl_index.py

核心脚本增强错误处理，从静默失败改为显式异常

```
# scripts/update_nightly_whl_index.py（关键变更片段）
# update_wheel_index 函数内部变更

def update_wheel_index(...):
    dist_dir = pathlib.Path("dist")
    whl_repo_dir = pathlib.Path("sgl-whl")

    # 原代码：log warning + return，CI 会显示绿色但索引未更新
    # 改为：直接抛出异常，让 CI 工作流明确失败
    if not dist_dir.exists():
        raise FileNotFoundError(
            f"{dist_dir} does not exist — the download-artifact step did not "
            f"populate it; refusing to silently no-op the index update"
        )

    # ... 其他逻辑 ...

    new_links = []
    for wheel_path in sorted(dist_dir.glob("*.whl")):
        # 原代码：try-except 包裹整个循环体，异常时 print + continue
        # 改为：直接执行，让异常自然传播（例如哈希计算失败）
        filename = wheel_path.name
        sha256 = compute_sha256(wheel_path)
        wheel_url = f"{base_url}/{release_tag}/{filename}#sha256={sha256}"
        link = f'<a href="{wheel_url}">{filename}</a><br>'
        new_links.append(link)
        print(f"  Added: {filename}")

    if not new_links:
        # 原代码：print + return，导致空提交
        # 改为：抛出异常
        raise RuntimeError(
            f"No wheels found in {dist_dir} — index update for {cuda_version} "
            f"would be a no-op; failing loudly instead of pushing an empty change"
        )
    # ... 后续合并去重并写入索引 ...
```

评论区精华

该 PR 无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。主要风险在于并行推送索引分支可能导致冲突，但通过 max-parallel: 1 已序列化处理。另一个风险是如果 wheel 实际上包含 CUDA 相关代码，同时注册两个索引可能导致安装不兼容的版本，但 PR 明确说明 wheel 是 CUDA agnostic 的，且历史经验表明 sglang wheel 对 CUDA 版本不敏感。
- 影响：对用户而言，安装 nightly wheel 时可以选择 cu129 或 cu130 的 extra-index-url，更灵活。对维护者而言，CI 构建时间略有增加（因为需要运行两次索引更新步骤），但构建本身只跑一次，影响很小。
- 风险标记：并行操作分支风险（已通过序列化缓解）

关联脉络

- PR #24183 [Misc] Redirect default sglang nightly wheel to cuda 130: 同为调整 nightly wheel 的 CUDA 版本策略，本 PR 实现了同时支持 cu129 和 cu130，而 PR#24183 仅将默认指向改为 cu130
- PR #24170 chore: bump sgl-kernel version to 0.4.2: 近期涉及 sgl-kernel 版本升级和 PyPI 默认切换的 CI 调整，与 nightly wheel 发布流程相关
- PR #24162 [sgl-kernel] Prep for torch 2.11 upgrade and switch PyPI default to cu130: 同样涉及 CUDA 版本切换和 wheel 发布流程，与本 PR 的 dual-index 策略形成演进链条