PR #24162 完整报告

sgl-project/sglang

[sgl-kernel] Prep for torch 2.11 upgrade and switch PyPI default to cu130

合并时间：2026-05-01 05:54

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24162

执行摘要

- 一句话：sgl-kernel 升级 torch 2.11，PyPI 默认发布切至 cu130
- 推荐动作：该 PR 以最小的改动完成了 torch 2.11 的适配和 PyPI 发布版本切换，适合作为基础设施变更的参考案例。reviewer 指出的两个问题（Dockerfile 冗余、pyproject.toml 版本未同步）虽已合并，但作者应尽快跟进修复，否则可能影响下游用户。建议团队建立构建镜像版本与 pyproject.toml 的自动校验机制。

功能与动机

PR body 明确指出本变更来自 #21247 的子集，目标是完成 sgl-kernel 侧的 torch 2.11 升级。torch 2.11 原生提供了 cu129 wheels（torch 2.9 没有），因此 12.9 CUDA 行可以正确解析为 cu129 而非回退到 cu128。同时，团队希望将 PyPI 默认发布版本从 cu129 切换至 cu130，因为 cu130 是更新的 CUDA 版本且能顺利通过 PyPI 上传检查。

实现拆解

1. 升级 torch 版本（sgl-kernel/Dockerfile、sgl-kernel/README.md）：将 sgl-kernel 构建镜像中所有 CUDA 行（12.8、12.9、13.0 及默认）的 TORCH_VER 从 2.9.1 改为 2.11.0，并同步更新 README 中的环境要求。12.9 行的 CU_TAG 从 cu128 更正为 cu129，以匹配 torch 2.11 的新标签。
2. 调整默认 CUDA 版本（scripts/update_kernel_whl_index.py）：将 DEFAULT_CUDA_VERSION 常量从 "129" 改为 "130"，使 cu130 成为无后缀 wheel 的默认版本。check_wheel_cuda_version 函数的逻辑据此判断 wheel 是否应被纳入索引。
3. 重定向 PyPI 发布流程（.github/workflows/release-whl-kernel.yml）：将原本在 build-cu129-matrix 中的“Strip +cu local version”和“Upload to PyPI”步骤整体移动至 build-cu130-matrix。cu129 构建矩阵不再执行 PyPI 上传，仅将 wheel 上传至 sgl-project/whl 索引（通过 upload-artifact 和 update whl index 步骤）。cu130 构建矩阵继承相同的 strip 脚本（正则 +cu[0-9]\+$ 通用），确保上传到 PyPI 的 wheel 版本号不带 +cu130 后缀。

关键文件：

- .github/workflows/release-whl-kernel.yml（模块 发布脚本；类别 infra；类型 infrastructure）：核心发布流程变更：将 Strip+Upload 步骤从 cu129 移至 cu130，使 cu130 成为 PyPI 默认发布变体。改动量最大（+34/-38），涉及构建矩阵控制流。

- scripts/update_kernel_whl_index.py（模块 脚本工具；类别 source；类型 core-logic；符号 check_wheel_cuda_version, update_wheel_index）：默认 CUDA 版本从 cu129 切换为 cu130，影响 wheel 索引生成时的文件筛选逻辑。虽仅改动一行，但直接影响索引生成行为。
- sgl-kernel/Dockerfile（模块 构建镜像；类别 config；类型 configuration）：构建镜像中 torch 版本从 2.9.1 全面升级至 2.11.0，并修正 cu129 的标签映射。直接影响所有 CUDA 后端的 sgl-kernel 编译环境。
- sgl-kernel/README.md（模块 文档；类别 docs；类型 documentation）：更新安装文档中的 torch 要求版本，提醒用户匹配新依赖。

关键符号：check_wheel_cuda_version, update_wheel_index

关键源码片段

scripts/update_kernel_whl_index.py

默认 CUDA 版本从 cu129 切换为 cu130，影响 wheel 索引生成时的文件筛选逻辑。虽仅改动一行，但直接影响索引生成行为。

```
# All the CUDA versions that the wheels will cover
SUPPORTED_CUDA_VERSIONS = ["129", "130"]
DEFAULT_CUDA_VERSION = "130" # 默认从 cu129 切换为 cu130


def check_wheel_cuda_version(path_name, target_cuda_version):
    # 跳过 ROCm wheel
    if re.search(r"rocm", path_name):
        return False
    # 对于非默认 CUDA 版本，wheel 路径名必须包含对应版本号
    if target_cuda_version != DEFAULT_CUDA_VERSION:
        return target_cuda_version in path_name
    # 对于默认版本，wheel 路径名不应包含其他 CUDA 版本后缀
    for cuda_version in SUPPORTED_CUDA_VERSIONS:
        if cuda_version != DEFAULT_CUDA_VERSION and cuda_version in path_name:
            return False
    return True
```

sgl-kernel/Dockerfile

构建镜像中 torch 版本从 2.9.1 全面升级至 2.11.0，并修正 cu129 的标签映射。直接影响所有 CUDA 后端的 sgl-kernel 编译环境。

```
# 安装 Python 依赖（torch + 构建工具）
RUN --mount=type=cache,id=sgl-kernel-pip,target=/root/.cache/pip \
    set -eux; \
    case "${CUDA_VERSION}" in \
      13.0) TORCH_VER=2.11.0; CU_TAG=cu130 ;; \
      12.9) TORCH_VER=2.11.0; CU_TAG=cu129 ;; \  # torch 2.11 首次提供 cu129 标签
      12.8) TORCH_VER=2.11.0; CU_TAG=cu128 ;; \
      *) TORCH_VER=2.11.0; CU_TAG=cu126 ;; \
```

```
esac; \
${PYTHON_ROOT_PATH}/bin/pip install torch==${TORCH_VER} --index-url https://${PYTORCH_MIRROR}/whl/${CU_TAG};
```

评论区精华

1. Dockerfile 中 TORCH_VER 重复与默认 CU_TAG 映射（gemini-code-assist[bot]）：reviewer 指出 TORCH_VER=2.11.0 在每个 case 分支重复，且默认分支的 CU_TAG=cu126 可能不适用于所有 CUDA 版本（如 12.1 或 12.4）。建议重构以减少冗余并确保标签映射鲁棒。未回复，但 PR 已合并。
2. pyproject.toml 的 torch 最低版本未同步更新（gemini-code-assist[bot]）：reviewer 提出 sgl-kernel/pyproject.toml 中 build-system.requires 仍指定 torch>=2.8.0，与新的 2.11.0 要求不一致，可能导致构建时使用不兼容 torch 版本。建议更新最低版本。未修改，但 PR 合并，可能留待单独处理。

- Dockerfile 中 TORCH_VER 重复及默认 CU_TAG 映射 (design): 未获得作者回复，PR 已合并，未修改。
- pyproject.toml 中 torch 最低版本需同步更新 (correctness): 未修改，PR 合并但风险未消除。

风险与影响

- 风险：
  1. 依赖版本不匹配：sgl-kernel/pyproject.toml 中的 torch>=2.8.0 尚未更新，用户安装 sgl-kernel 时若使用 torch 2.8~2.10 可能因 ABI 不兼容或 API 缺失导致构建失败。
  2. 默认 CUDA 版本切换：依赖 cu129 构建的用户在升级后若不明确指定索引，将默认安装 cu130 版本，可能因本地 CUDA 12.9 环境不兼容 cu130 的二进制而产生运行时错误。
  3. Dockerfile case 结构冗余：当前每个分支都重复 TORCH_VER=2.11.0，未来升级需修改多处，容易遗漏。默认 case 的 CU_TAG=cu126 覆盖了众多 CUDA 版本，可能并非预期。 - 影响：用户影响：安装或升级 sgl-kernel 的用户必须确保本地 torch 版本为 2.11.0（或指定兼容索引）。从 sgl-project/whl 索引安装 cu129 版本的用户需要添加 +cu129 后缀 pip 安装参数。系统影响：sgl-kernel 构建镜像不再支持 torch 2.9.x，CI 发布矩阵同步切换。团队影响：后续维护时需注意 pyproject.toml 的最低 torch 版本应与构建镜像保持一致。 - 风险标记：依赖版本不匹配，默认 CUDA 版本切换

关联脉络

- PR #24170 chore: bump sgl-kernel version to 0.4.2: 同为 sgl-kernel 的版本发布相关 PR，但本 PR 改变的是依赖版本和发布目标，而 24170 仅是版本号升级（0.4.1.post1→0.4.2）。二者共同影响 sgl-kernel 的发布流程。