

PR #38763 完整报告

sgl-project/sglang

[AMD][CI] Publish ROCm 10 release images and kernel wheel

合并时间: 2026-09-10 14:57

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/38763>

执行摘要

- 一句话: 为 AMD ROCm 10 发布 Docker 镜像和内核 wheel, 同步 CI 与发布产物。
- 推荐动作: 本 PR 是一次重要的基础设施补全, 虽然代码逻辑相对直接 (主要是 YAML 和 shell 脚本的条件扩展), 但其背景清晰、验证充分、兼容性考虑周全。对于关注 AMD ROCm 支持或 CI/CD 流水线的工程师, 值得精读, 尤其是 build_rocm.sh 中为 ROCm 10 pip 安装 SDK 构建环境的补丁部分, 它展示了解决跨版本环境差异的实用技巧。PR 的提交历史清晰地展示了从“发布 Docker 镜像”到“发布内核 wheel”再到“优化构建范围”的演进思路, 是一个良好的基础设施 PR 范例。

功能与动机

动机来自 PR body 的明确表述: #38659 将 ROCm 10 设为 AMD PR 和 nightly 测试的默认 flavor, 但发布侧的工作流仍然只构建 7.x 的产物。这导致 [lmsysorg/sglang](#) 没有 `v<version>-rocm10-mi30x/-mi35x` 标签, 且 AMD CI 所需的 `sglang-kernel` wheel 没有发布版本。本 PR 填补了这一缺口。

实现拆解

1. 更新 Docker 发布工作流 ([release-docker-amd.yml](#)): 在 workflow_dispatch 的 rocm_version 选项中增加 rocm10, 并将默认矩阵从 ['rocm700', 'rocm720', 'rocm724'] 扩展为包含 ['rocm1000']。在构建步骤的条件分支中增加 elif ["\${matrix.rocm_version}" = "rocm10"] 分支, 映射 gfx942 -> rocm10-mi30x 标签, gfx950 -> rocm10-mi35x 标签, 对应 Dockerfile 阶段 gfx942-rocm1000 / gfx950-rocm1000。这确保了版本标签发布时生成完整的 ROCm 10 镜像。
2. 更新内核 wheel 发布工作流 ([release-whl-kernel.yml](#)): 在 target 输入参数的 allowed_values 中增加 rocm1000。在 build-rocm-matrix 任务中, 将 rocm-version 矩阵的静态列表替换为条件表达式, 根据 inputs.target 动态选择构建哪些 flavor (700、720、1000 或全部)。为 release-rocm700 和 release-rocm720 任务添加 if 条件, 使其仅在对应 flavor 被构建时运行。新增 release-rocm1000 任务, 其流程与现有任务相同: 下载 artifact、创建 GitHub release、克隆 wheel index 仓库、运行 update_kernel_whl_index.py --rocm 1000 更新索引、推送。这确保了 ROCm 10 wheel 能被正确发布和索引。
3. 适配 ROCm 10 构建环境 ([3rdparty/amd/wheel/sgl-kernel/build_rocm.sh](#)): 在脚本开头, 为 ROCM_VERSION 为 1000 的情况新增镜像选择:

rocm/pytorch:rocm10.0_ubuntu24.04_py3.12_pytorch_release_2.11.0。由于 ROCm 10 SDK 通过 pip 安装到 site-packages，构建容器缺少 CMake 所需的 /opt/rocm 目录和编译工具链。因此，脚本在 docker run 命令前注入一段 ROCM_SETUP shell 脚本，用于在容器内安装 rocm-sdk-devel、设置 ROCM_HOME、创建 /opt/rocm 符号链接并修复 libc.so 路径问题。此修复与生产镜像构建（docker/rocm.Dockerfile）中的 rocm1000-base 阶段保持一致。

4. 适配 wheel 重命名脚本 (3rdparty/amd/wheel/sgl-kernel/rename_wheels_rocm.sh)：脚本接受一个新的可选参数 ROC_WHEEL_VERSION。对于 ROCm 10，由于没有 /opt/rocm-<version> 目录，该参数由 build_rocm.sh 传入（值为 1000）。脚本优先使用该参数，若未提供则回退到原有的 /opt/rocm-* 探测逻辑，确保了 ROCm 7.x flavor 的兼容性。
5. 变更的影响与配套：所有变更仅限于发布工作流和构建脚本，不影响运行时代码或模型输出。变更已通过手动触发的 CI 验证（PR body 中引用了成功的 workflow run），并确认了生成的 wheel 和镜像符合预期。兼容性方面，ROCm 7.x 的构建流程未受影响，仅新增了 ROCm 10 的支持。

关键文件：

- .github/workflows/release-whl-kernel.yml（模块 发布工作流；类别 infra；类型 infrastructure）：核心变更文件，将 ROCm 10 内核 wheel 的构建和发布逻辑集成到现有工作流中，并优化了按 flavor 构建的调度逻辑。
- 3rdparty/amd/wheel/sgl-kernel/build_rocm.sh（模块 内核构建脚本；类别 other；类型 core-logic）：实现 ROCm 10 构建环境适配的关键脚本，处理了 SDK 安装、环境设置和路径修复等核心逻辑。
- .github/workflows/release-docker-amd.yml（模块 发布工作流；类别 infra；类型 infrastructure）：将 ROCm 10 Docker 镜像的发布集成到版本标签发布流程中，补全了镜像发布链。
- 3rdparty/amd/wheel/sgl-kernel/rename_wheels_rocm.sh（模块 内核构建脚本；类别 other；类型 core-logic）：适配 ROCm 10 环境下 wheel 重命名逻辑，接受外部传入的版本参数以替代不可用的目录探测。

关键符号：未识别

评论区精华

Review 评论仅有一条，来自维护者 HaiShaw 的批准（LGTM），并附带一个未来建议：“Please purge rocm700 next.”。这表明对本 PR 的实现没有争议，并暗示了后续清理旧 ROCm 版本的维护计划。PR body 中包含了详尽的验证信息，展示了对构建结果的仔细检查，这本身也是一种有效的“讨论”，但在此 PR 中未引发公开的交锋。

- 暂无高价值评论线程

风险与影响

- 风险：

1. CI/CD 资源开销：每个 ROCm flavor 的构建都是在共享 AMD 节点上的一个独立镜像构建任务。新增 rocm1000 leg 意味着每次完整发布 (target: all) 会多一个构建和发布任务，增加了 CI 时间和资源消耗。但 PR 已通过逻辑优化 (if 条件) 确保了按需构建，仅在需要时运行对应 flavor，缓解了此问题。
 2. 构建环境一致性：ROCm 10 的构建环境依赖在容器内动态安装 rocm-sdk-devel 和设置环境变量。这与 Dockerfile 中的固定阶段不同，可能引入细微差异。PR 作者通过引用 docker/rocm.Dockerfile 的 rocm1000-base 阶段并复制其逻辑来最小化风险，且验证通过。
 3. Wheel 索引一致性：发布 job 会重复执行 update_kernel_whl_index.py --rocm 1000。如果多个 flavor 的发布 job 并发运行，理论上可能导致索引文件冲突。但根据 workflow 结构，release-rocm1000 是一个独立 job，且 git push 操作是顺序的，实际风险较低。
 4. 向后兼容性：变更保持了对现有 ROCm 7.x flavor 的完全兼容。build_rocm.sh 的 WHEEL_ROCM_VERSION 和 rename_wheels_rocm.sh 的参数处理都设计了回退逻辑。
- 影响：直接影响：
 - AMD 用户与部署者：现在可以从官方发布渠道获取到包含 ROCm 10 支持的 Docker 镜像和 sglang-kernel wheel，简化了在最新 AMD 硬件 (MI30x, MI35x) 上的部署。
 - AMD CI 与测试：确保了 CI 环境中使用的 ROCm 10 flavor 有匹配的官方发布产物，提高了测试结果的可信度。
 - 间接影响：
 - 项目维护：为移除过时的 ROCm 7.x 发布产物 (如 HaiShaw 建议的 [purge rocm700](#)) 铺平了道路，有助于简化维护负担。
 - 生态：sgl-project/whl 仓库的 rocm1000/sglang-kernel/index.html 索引被更新，使更多工具链可以自动发现和安装 wheel。
 - 风险标记：CI 资源开销增加，构建环境动态配置，跨版本兼容性维护

关联脉络

- PR #38659 [AMD] Make ROCm 10 the default flavor for PR and nightly tests: 本 PR 是 #38659 的直接后续，后者将 ROCm 10 设为默认测试 flavor，但未更新发布 workflow，导致发布产物不匹配。
- PR #38694 [AMD][CI] Drop the dead miles ROCm 7.0 nightly image build: 同属 AMD CI 维护线，清理过时的 ROCm 7.0 构建，与本 PR 补全 ROCm 10 发布形成演进关系。