

PR #37225 完整报告

sgl-project/sglang

[AMD] build gfx1250 release image from main

合并时间: 2026-08-31 17:04

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/37225>

执行摘要

- 一句话: AMD gfx1250 镜像改从 main 构建并转正式发布
- 推荐动作: 值得快速浏览而非精读。对维护 AMD 发布流水线的成员有参考价值, 重点关注两个设计决策: (1) 用独立 job 隔离 bring-up 构建失败, 避免影响正式发布; (2) tag 从 dev 到 release 的切换时机以构建 stage 是否合入 main 为准。若后续 gfx1250 镜像被 MI450 集群实际消费, 应重新评估是否仍需要独立 job, 并考虑为发布镜像增加冒烟验证步骤。

功能与动机

PR 正文明确指出: PR #36871 已把 gfx1250-rocm1000 stage 合入 main, 因此 release job 不再需要 checkout amd_helios 分支; 随之可以删除分支 pin 及其所需的 BRANCH_SHA 传递, 并把镜像提升为与 gfx942/gfx950 一致的正式 release tag, 镜像同步到 lmsysorg/sglang-rocm。此前 gfx1250 镜像只能从 bring-up 分支构建并带有 -dev 后缀, 用户无法从常规发布通道获取, 且维护分支 pin 需要额外的 BRANCH_SHA 传递逻辑, 增加流水线复杂度与 tag 语义歧义。

实现拆解

1. 解除分支锁定: publish-gfx1250 job 的 Checkout 步骤删除 ref: amd_helios (保留 fetch-depth: 0), 改为构建当前 workflow 触发的 ref (main)。这是整个 PR 的前提, 依赖 #36871 已合入 main。
2. 删除 BRANCH_SHA 传递链路: version 步骤不再计算并输出 branch_sha; docker build 的 --build-arg SGL_BRANCH 从 \${BRANCH_SHA} 改为 \${GITHUB_SHA}, 镜像内源码版本与触发发布时的 main commit 绑定。
3. Tag 语义升级: image tag 从 v\${VERSION}-rocm10-mi45x-dev-\${DATE} 改为 v\${VERSION}-rocm10-mi45x-\${DATE}, 去掉 -dev 后缀, 并把 IMAGE_TAG 写入 GITHUB_ENV 供后续同步步骤复用。
4. 镜像同步到公开仓库: 新增“Login to Docker Hub (lmsys)” (使用 DOCKERHUB_USERNAME / DOCKERHUB_TOKEN) 与“Push to lmsysorg/sglang-rocm”两步, 将 rocm/sgl-dev 镜像 tag 后推送到 lmsysorg/sglang-rocm, 与 gfx942/gfx950 路径一致。

5. 保留独立 job 结构: gfx1250 仍与 gfx942/gfx950 分开运行; workflows 注释说明原因——目前没有 MI450 runner 池消费该镜像, bring-up 构建失败不应 gate 已发布的正式镜像。

配套说明: 无测试改动、无配置 schema 变更; 属于 CI 发布流水线调整, 验证只能通过触发 release job 实际运行完成。

关键文件:

- .github/workflows/release-docker-amd-rocm10.yml (模块 发布流水线; 类别 infra; 类型 infrastructure; 符号 publish-gfx1250): 唯一的变更文件, 承载了 gfx1250 发布 job 的全部改动: 移除 amd_helios 分支 pin、删除 BRANCH_SHA 传递、改用 GITHUB_SHA 构建、去掉 -dev 标签后缀, 并新增向 lmsysorg/sglang-rocm 的镜像同步步骤, 使 gfx1250 的发布通道与 gfx942/gfx950 对齐。

关键符号: publish-gfx1250, Push to lmsysorg/sglang-rocm

评论区精华

该 PR 没有任何实质 review 讨论: 0 条 comment、0 条 review comment, 唯一审核为 bingxche 的 APPROVED (正文为空), 说明变更符合 reviewer 预期。值得提炼的决策全部写在代码注释里: gfx1250 保持独立 job, 避免 bring-up 构建失败阻断 gfx942/gfx950 发布; tag 从 -dev 升级为正式语义并写入 IMAGE_TAG 供同步步骤复用。PR 快照中 Extra CI 显示失败状态 (Run #33375062164), 但材料未提供失败详情, 无法归因。

- 暂无高价值评论线程

风险与影响

• 风险:

1. 镜像质量风险: 去掉 -dev 后缀后, gfx1250 镜像以正式 release tag 出现在 lmsysorg/sglang-rocm 公共仓库; 当前没有 MI450 runner 池消费该镜像, 质量验证可能不足, 用户可能拉取到未充分验证的镜像。缓解: tag 仍含 mi45x 与日期标识, 且该 job 失败不 gate 其他镜像。
2. 构建来源变更风险: SGL_BRANCH 从固定分支 head 改为 GITHUB_SHA 后与 main 实时状态绑定, 若 docker/rocm.Dockerfile 的 gfx1250 stage 有回归, 构建会失败, 但不会波及其他 job。
3. 配置 / 密钥风险: 新增 lmsys 登录步骤依赖 DOCKERHUB_USERNAME / DOCKERHUB_TOKEN secrets, 缺失或权限不足会导致发布失败; 此类配置错误无单元测试覆盖, 只能运行时暴露。
4. CI 状态: PR 快照中 Extra CI 运行显示失败 (Run #33375062164)、AMD ROCm 7.2 无对应运行; 材料中无失败详情, 无法判断是否与本 PR 相关。
5. 兼容性风险: -dev tag 不再生成, 若此前有消费者依赖该 tag 会受影响 (bring-up 阶段镜像, 影响面很小)。- 影响: 影响范围集中在发布流水线, 无运行时、性能或 API 影响。- 用户侧: AMD MI45x (gfx1250, ROCm 10) 用户可从 lmsysorg/sglang-rocm 获取与 gfx942/gfx950 同语义的正式 tag 镜像, 不再依赖 amd_helios 分支手工构建。- 流水线侧: 移除对 amd_helios 分支的运行时依赖, 减少分支管理与 BRANCH_SHA 传

递复杂度；后续 gfx1250 修复只需改 main 上的 docker/rocm.Dockerfile。 - 团队侧：AMD 多架构发布流程统一到 main；job 独立性使 gfx1250 的 bring-up 风险与正式镜像发布隔离。 - 影响程度：中等偏低，改动量小（+18/-19），但属于生产发布通道变更。 - 风险标记：发布流水线变更，镜像标签语义变化（-dev 移除），无自动化验证，依赖前置 PR #36871, Extra CI 失败待查

关联脉络

- PR #36871 （材料未提供标题，PR 正文引用）：本 PR 的直接前置依赖：PR 正文明确说明 #36871 已把 gfx1250-rocm1000 stage 合入 main，使 release job 无需再 checkout amd_helios，本 PR 据此删除 BRANCH_SHA 传递并将镜像升级为正式 tag。
- PR #37132 [AMD] Fix the QuickReduce bf16 cast failing to build for CDNA: 同属 AMD 平台构建 / 发布链路：近期 AMD bring-up 与 CDNA 构建修复持续合入，gfx1250 转为正式发布后这些修复会直接影响 release 镜像质量。