

PR #29219 完整报告

sgl-project/sglang

[AMD-miles] Add a ROCm 7.0 MI35x miles nightly image

合并时间: 2026-06-25 11:27

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29219>

执行摘要

为 AMD MI35x (gfx950) 在 ROCm 7.0 夜间 Docker 构建流水线中增加一个构建条目，以避免 ROCm 7.2 的 VMM 内存释放回归问题。变更仅涉及 CI 配置，无代码逻辑改动。

功能与动机

AMD MI35x 用户需要基于 ROCm 7.0 的 Docker 镜像：ROCm 7.2 存在已知的 VMM 物理内存释放回归问题，会导致 `torch_memory_saver` 无法释放 VRAM，在 `colocate sleep/wake` 场景下触发 OOM。通过提供 ROCm 7.0 镜像可避免此问题。

实现拆解

- 在 `.github/workflows/release-docker-amd-miles-rocm700-nightly.yml` 的 build job 矩阵中添加如下条目：
 - `gpu_arch: gfx950`
 - `rocm_tag: rocm700-mi35x`
 - `sglang_image_tag: v0.5.10-rocm700-mi35x`
- 在 `push_local_registry` job 的 `gpu_arch` 矩阵中增加 `gfx950` 值，从 `["gfx942"]` 变为 `["gfx950", "gfx942"]`。
- 变更后，ROCm 7.0 nightly 流水线会同时构建以下两个镜像：
 - `rocm/sgl-dev:miles-rocm700-mi35x-<DATE>` (`gfx950 / MI350X•MI355X`)
 - `rocm/sgl-dev:miles-rocm700-mi30x-<DATE>` (`gfx942 / MI300`)

无关键源码片段：变更仅为 YAML 配置条目添加。

评论区精华

无实质 review 讨论。该 PR 由 bingxche 批准，无 review 评论。

风险与影响

- 风险：低风险。依赖基础镜像 `v0.5.10-rocm700-mi35x` 必须存在且可构建，否则该构建条目会失败，但不影响其他条目。
- 影响：仅影响 AMD MI35x 用户，为其提供 ROCm 7.0 夜间镜像；对现有 MI30x 构建无影响。

关联脉络

该 PR 属于 AMD 平台持续支持的一部分，与近期其他 AMD 相关 PR（如 #28872 NPU 修复、#29042 DeepSeek 精度修复）无直接关联，更多是基础设施层面的补充。