

PR #2230 完整报告

radixark/miles

fix(ci): disable MI300X runner jobs

合并时间: 2026-08-06 16:51

原文链接: <http://prhub.com.cn/radixark/miles/pull/2230>

执行摘要

- 一句话: 禁用 MI300X CI 调度, 8 个 ROCm 用例标记 disabled
- 推荐动作: 该 PR 本身不值得精读 (单行重复变更), 但值得关注其背后的运维决策: 用一个注册层的 disabled 参数实现整条 CI 流水线的开关, 既保留配置又避免改动 workflow。建议团队为恢复动作建一个跟踪 issue, 并在 MI300X runner 稳定后逐例恢复并回填禁用原因。

功能与动机

PR body 指出症状是授权门允许每个 ROCm 触发下产生两个 stage-c-4-gpu-mi300x shard; 根因是 stage-c-4-gpu-mi300x.if 只检查 self-hosted 授权, 而 pr-test-rocm.yml 把 PR、nightly、manual 全部暴露给该 job。作者明确这是 operational containment and does not diagnose runner failures, 即不诊断 runner 故障, 先切断调度让 CI 恢复可用。

实现拆解

1. 变更入口: 8 个修改文件全部是 tests/ 下的 ROCm e2e 注册文件, 统一在 register_rocm_ci(...) 调用中追加 disabled="Disable due to failure"。
2. 作用机制: register_rocm_ci 是 tests/ci/ci_register.py 提供的注册函数 (该文件不在本次改动内), disabled 参数使该用例在 stage-c-4-gpu-mi300x 收集时被排除, 从调度源头上不生成 MI300X shard。
3. 覆盖范围: DeepSeek V4 Flash、GLM-5.2、Inkling、Qwen3-4B FSDP hybrid shard、GLM-4.7-Flash R3 MTP、Qwen3-30B-A3B DeepEP FP8 bridge、Qwen3-4B offload disk stream、Qwen3.5-35B-A3B MTP1 共 8 个用例。
4. 演进过程: 3 个 commit 显示最初还包含 workflow 文档与 run-suite 回归测试改动, 最终 commit 明确 "Remove the workflow documentation and run-suite regression-test changes", 收敛为纯注册层的最小变更。
5. 测试配套: PR body 声称新增 TestRocmWorkflowScopeSeam::test_stage_is_disabled_and_preserves_configuration 并跑过 71 个测试, 但最终 diff 未包含该测试文件 (可能已存在于代码库或未合入), 需要确认。

关键文件:

- tests/e2e/megatron/model_scripts/test_deepseek_v4_flash_4layer_ci.py (模块 模型脚本; 类别 test; 类型 test-coverage) : DeepSeek V4 Flash 4 layer CI 的 ROCm 注册入口

，追加 disabled 后 MI300X 不再调度，是本次变更的代表文件。

- tests/e2e/megatron/model_scripts/test_glm5_2_744b_a40b_5layer_ci.py (模块 模型脚本; 类别 test; 类型 test-coverage) : GLM-5.2 744B A40B 5 layer smoke 测试的 ROCm 注册被禁用，属于模型脚本 suite 内的高价值 AMD 用例。
- tests/e2e/megatron/model_scripts/test_inkling_small_4layer_ci.py (模块 模型脚本; 类别 test; 类型 test-coverage) : Inkling Small 4 layer 全参数训练 CI 的 ROCm 注册被禁用，影响多模态模型在 AMD 侧的覆盖。
- tests/e2e/fsdp/test_qwen3_4B_fsdp_hybrid_shard_r2s2.py (模块 FSDP; 类别 test; 类型 test-coverage) : FSDP hybrid shard 的 R2S2 e2e 用例 ROCm 注册被禁用，FSDP 后端在 MI300X 上的验证暂时缺席。
- tests/e2e/megatron/test_glm47_flash/test_amd_r3_mtp.py (模块 AMD; 类别 test; 类型 test-coverage) : GLM-4.7-Flash 的 AMD 专属 R3 + MTP 4 GPU 用例被禁用，该用例是 MI300X 双 4 GPU runner 拆分后的专项覆盖。
- tests/e2e/megatron/test_qwen3_30B_A3B/test_amd_deepep_fp8_bridge.py (模块 AMD ; 类别 test; 类型 test-coverage) : Qwen3-30B-A3B 的 AMD 专属 DeepEP + FP8 bridge 用例被禁用，直接影响 AMD 上 MoE 与 FP8 关键路径的 CI 验证。
- tests/e2e/megatron/test_qwen3_4B_offload_disk_stream.py (模块 卸载流; 类别 test; 类型 test-coverage) : Qwen3-4B offload disk stream 用例的 ROCm 注册被禁用，磁盘流式 offload 在 MI300X 上的覆盖暂时中断。
- tests/e2e/megatron/test_qwen3_5_35B_A3B_mtp/test_amd_mtp1_spec_v2_r3.py (模块 AMD; 类别 test; 类型 test-coverage) : Qwen3.5-35B-A3B 的 AMD 专属 MTP1 + Spec V2 + R3 用例被禁用，该用例验证 AMD 上的 spec decoding 行为。

关键符号：未识别

关键源码片段

tests/e2e/megatron/model_scripts/test_deepseek_v4_flash_4layer_ci.py

DeepSeek V4 Flash 4 layer CI 的 ROCm 注册入口，追加 disabled 后 MI300X 不再调度，是本次变更的代表文件。

```
from tests.ci.ci_register import register_cuda_ci, register_rocm_ci
```

```
# CUDA (H200) 用例保持启用，继续在 stage-c-4-gpu-h200 运行
register_cuda_ci(
    est_time=1900,
    suite="stage-c-4-gpu-h200",
    labels=["megatron", "model-scripts"],
)
```

```
# ROCm (MI300X) 用例: `disabled` 参数在注册层把该 stage 标记为禁用,
# PR / nightly / manual 三种触发下都不会再调度 stage-c-4-gpu-mi300x 分片,
# 但用例定义与 runner 配置原样保留，便于后续单独恢复。
register_rocm_ci(
    est_time=1900,
```

```
suite="stage-c-4-gpu-mi300x",
labels=["megatron", "model-scripts", "amd"],
disabled="Disable due to failure",
)

# 训练指标门保持不变，等恢复调度后继续兜底
register_ci_gate(metric_key="train/grad_norm")
```

评论区精华

本 PR 无 review 评论 (`comments_count=0`、`review_comments_count=0`)。PR body 中作者自定的审查焦点是：① 确认 `stage-c-4-gpu-mi300x.if` 在每种触发下都跳过 MI300X runner；② 确认 `docs/ci/00-stage.md` 的 dormant-stage 措辞与 false gate 一致。

- 暂无高价值评论线程

风险与影响

- 风险：
 1. 覆盖空洞：所有 MI300X 相关回归（含 AMD 专用 MTP、DeepEP FP8 bridge、Qwen3.5 35B A3B MTP 等）在 CI 中静默跳过，AMD 侧回归将无人值守，属于典型 CI 静默降级。
 2. 故障信息丢失：8 个用例统一使用 "Disable due to failure" 作为禁用理由，未区分各自的具体故障，后续逐例恢复时需要重新诊断。
 3. 恢复成本：8 个文件分散在 model-scripts、FSDP、AMD 等不同 suite，未来恢复需逐一移除 disabled 参数，若没有跟踪手段容易遗漏。
 4. 影响面：仅影响 CI 调度，对训练、rollout 等运行时无影响，风险等级较低。- 影响：影响范围集中在 CI 基础设施：MI300X 自托管 runner 将不再被任何 ROCm 触发（PR / nightly / manual）调度，相关 8 个 AMD e2e 用例进入休眠状态。对用户而言，AMD 硬件上的回归验证缺口变大；对系统而言，CI 队列压力下降但覆盖完整性受损；对团队而言，需要一个显式的恢复计划，否则 MI300X 支持会长期处于无人验证状态。- 风险标记：CI 静默降级，AMD 回归覆盖空洞，恢复需逐例处理

关联脉络

- PR #1606 ci(rocm): add ROCm CI workflow for MI300X self-hosted runners: 该 PR 引入 ROCm CI workflow 与 MI300X 自托管 runner，并包含本 PR 所禁用的多个 AMD e2e 用例；本 PR 是对其调度能力的运营停用，构成「引入→停用」闭环。
- PR #2222 [AMD] Remove PR#301 FileSystemWriterAsync swap — superseded by Megatron-LM#74: 同属 AMD/ROCm 维护线，移除 ROCm 检查点写入器替换，体现 ROCm 路径仍在持续稳定化清理。
- PR #2139 [AMD] retune AMD 4-node dsv4 config: AMD 4 节点 DSv4 配方 retune，依赖 MI300X CI 验证；本 PR 后该验证将暂时缺席，需要外部恢复计划。