

PR #3184 完整报告

radixark/miles

ci: declare hardware on the two CUDA tests that landed after #2701

合并时间: 2026-09-11 02:19

原文链接: <http://prhub.com.cn/radixark/miles/pull/3184>

执行摘要

- 一句话: 为两个遗漏的 CUDA 测试补 hardware 声明, 解除 PR CI 全局阻塞。
- 推荐动作: 变更微小 (2 个文件各 1 行), 不构成精读对象, 但值得作为 #2701 硬件感知调度演进链条中的一环来理解其上下文。值得关注的决策是: 大规模必填参数迁移类 PR (如 #2701) 存在 "同天合并窗口" 漏网问题, 且由于 main 上无 PR 测试执行而未及时暴露, 后续此类基础设施变更应考虑在合并后增加全量注册扫描的自动校验。

功能与动机

PR body 明确指出: #2701 使 hardware= 成为 register_cuda_ci 的必填参数, 而 #2876 与 #2825 两个测试于同一天合并、未带该参数。在 main 上 --list-only 命令会在列出任何测试之前抛错: `ValueError: ... hardware in register_cuda_ci() must list at least one arch from ['hopper', 'blackwell']`。由于 Resolve suite plan 步骤会运行该命令, 每个 PR 的 stage-a-cpu 作业都会停在这一步; 且自 f211937b5 之后没有 PR Test run 执行过 main, 因此问题一直没有在 CI 上显示为红色。

实现拆解

2701 让 register_cuda_ci 的 hardware 参数变为必填, 并在当时对 main 上所有已注册文件做了扫清; 但 #2876 与 #2825 两个 PR 在同一天合并, 成为漏网之鱼, 导致后续每次套件计划解析都在注册阶段失败。

在 main 上运行 `python tests/ci/run_suite.py --hw cpu --suite stage-a-cpu --match-all-labels --list-only` 会先在 `tests/e2e/agentice/test_harbor_e2b_rollout.py` 的注册处抛出 `ValueError`; 修复该文件后, 又在 `tests/fast-gpu/test_true_on_policy_logprob_chunking.py` 处抛出同样的错误。两个文件的注册解析依次失败, 阻塞了套件计划生成。

`tests/e2e/agentice/test_harbor_e2b_rollout.py` 的 `register_cuda_ci(est_time=1200, suite="stage-c-2-gpu-h200", labels=["agentice"], disabled=...)` 调用新增 `hardware=["hopper"]`。`tests/fast-gpu/test_true_on_policy_logprob_chunking.py` 的 `register_cuda_ci(est_time=180, suite="stage-b-2-gpu-h200", labels=["megatron"])` 调用同样新增 `hardware=["hopper"]`。

无新增测试，注册解析器本身即门禁。作者手动执行 `--list-only` 验证：stage-a-cpu 成功解析 436 个测试；stage-b-2-gpu-h200 解析 5 个测试（含 chunking 测试）；stage-c-2-gpu-h200 解析 15 个测试（harbor 测试被列为 disabled 但仍参与解析）。

两个测试均位于 H200 套件，故只声明 hopper；待有人在 Blackwell 硬件上跑通后再扩展为 blackwell。ROCm 的注册函数不接受 `hardware` 参数，因此 mi350 相关文件不受影响。

关键文件：

- tests/e2e/agent/c/test_harbor_e2b_rollout.py（模块 沙箱测试；类别 test；类型 configuration）：来自 #2876 的 Harbor E2B 端到端测试，是触发注册解析器首个 ValueError 的文件；补齐 hardware 声明后，stage-c-2-gpu-h200 套件可正常解析（该测试当前处于 disabled 状态但仍参与注册校验）。
- tests/fast-gpu/test_true_on_policy_logprob_chunking.py（模块 损失计算；类别 test；类型 configuration）：来自 #2825 的 true-on-policy logprob 分块计算测试，是第二个触发注册解析器错误的文件；补齐 hardware 声明后 stage-b-2-gpu-h200 套件可正常列出该测试。

关键符号：未识别

关键源码片段

tests/e2e/agent/c/test_harbor_e2b_rollout.py

来自 #2876 的 Harbor E2B 端到端测试，是触发注册解析器首个 ValueError 的文件；补齐 hardware 声明后，stage-c-2-gpu-h200 套件可正常解析（该测试当前处于 disabled 状态但仍参与注册校验）。

```
# 该文件来自 PR #2876 (Harbor E2B 沙箱端到端测试) 。
# register_cuda_ci 在 #2701 之后要求 hardware 必填：缺失时解析器会在
# --list-only 阶段直接抛 ValueError，导致所有 PR 的 stage-a-cpu 作业在
# Resolve suite plan 步骤失败；即使本测试当前被 disabled，也必须补齐
# 声明才能通过注册门禁（解析器对 disabled 用例同样校验）。
```

```
register_cuda_ci(
    est_time=1200,
    suite="stage-c-2-gpu-h200",
    hardware=["hopper"], # 套件是 H200 机群，仅声明 hopper；在 Blackwell 跑通后再扩展
    labels=["agent/c"],
    disabled="needs a network route to the sandbox service and its key on the runner; run
    manually on a GPU devbox that has both",
)
```

tests/fast-gpu/test_true_on_policy_logprob_chunking.py

来自 #2825 的 true-on-policy logprob 分块计算测试，是第二个触发注册解析器错误的文件；补齐 hardware 声明后 stage-b-2-gpu-h200 套件可正常列出该测试。

```
# 该文件来自 PR #2825 (true-on-policy logprob 分块计算修复) 。
# 与 Harbor 测试同因：#2701 之后 hardware 必填，缺省会触发注册解析器
# 报错并阻塞套件计划生成；补上声明后 stage-b-2-gpu-h200 可正常列出。
register_cuda_ci(
```

```
est_time=180,  
suite="stage-b-2-gpu-h200",  
labels=["megatron"],  
hardware=["hopper"], # H200 套件，故仅声明 hopper  
)
```

评论区精华

该 PR 没有实质性的技术讨论。claude[bot] 自动回复说明该 PR 来自 fork、自动审查被禁用，维护者可评论 [@claude review](#) 触发一次性审查；最终未触发 AI 审查，由维护者 nblintao 人工 APPROVED，并留言 "Thanks for the fix!"。

- fork PR 自动审查被禁用 (other): 未触发自动 AI 审查，由维护者 nblintao 人工 APPROVED 合并。
- 修复确认与合并 (other): 无未解决疑虑，PR 被合并。

风险与影响

- 风险：风险极低：改动仅涉及两个测试文件里的 CI 注册元数据，不触碰任何测试逻辑或源码路径，回归面为零。唯一值得注意的取舍是 hardware=["hopper"] 意味着这两个测试未来不会被 Blackwell 套件调度，作者已在 PR body 中明示这是有意决策 ("widen to blackwell once someone has run them there")，若 Blackwell 成为主要 runner 需回访补充声明。另外，disabled 的 Harbor 测试同样参与注册解析，说明该门禁对 disabled 用例同样生效，未来新增 disabled 测试时也需同步补齐必填参数。
- 影响：对 CI 系统的影响是直接的：解除所有 PR 的 stage-a-cpu 作业在 Resolve suite plan 阶段的全局阻塞，使套件计划恢复正常生成。对测试调度而言，两个 GPU 测试被正确纳入 #2701 引入的硬件感知调度体系，可在 H200 套件中被识别与分配。对团队而言，这是一个低代码量、高影响面的及时修复，避免了后续 PR 持续在入口处失败。
- 风险标记：CI 注册必填遗漏，仅声明 Hopper 架构

关联脉络

- PR #2701 feat(ci): declare supported GPU generations on every CUDA test: 引入 hardware 必填要求并扫清当时 main 文件的源头 PR，本 PR 是对其同天合并窗口遗漏的补充。
- PR #2876 ci(e2e): Harbor trials on real e2b sandboxes through the full rollout path (manual for now): 本 PR 修复的 Harbor E2B 测试文件即来自此 PR，其注册缺少 hardware 声明。
- PR #2825 fix: chunk true-on-policy logprob computation: 本 PR 修复的 chunking 测试文件即来自此 PR，其注册缺少 hardware 声明。