PR #28344 完整报告

sgl-project/sglang

[AMD] register 4 2-gpu tests to stage-b-test-2-gpu-large-amd

合并时间：2026-06-17 12:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28344

执行摘要

- 一句话：4 个 2-GPU 测试注册到 AMD CI
- 推荐动作：本 PR 是纯粹的 CI 注册变更，技术含量低，但作为 AMD CI 覆盖建设的一部分，值得团队关注其渐进式覆盖策略和从候选筛选中暴露的平台差异（如 kv-canary 内核仅 CUDA、模型权重复制不可用）。对于 CI 维护者和 AMD 平台开发者，建议阅读 PR body 中的决策记录，了解哪些测试被排除以及原因，以便后续在对应模块完成 ROCm 移植后重新注册。

功能与动机

作者在 PR body 中指出这是 NV→AMD CI 覆盖差距审计的第四批（延续 #25208、#25939、#27817），目的是将已有的 2-GPU CUDA 测试注册到 AMD 双 GPU CI 通道，以增加 AMD 硬件上的 CI 覆盖，确保这些后端无关或 ROCm 支持的路径在 AMD 上持续验证。

实现拆解

本 PR 的实现极为简洁，仅涉及对 4 个测试文件的细微修改，每个文件都遵循相同的两步模式：

1. 导入调整：在每个文件中，将原本只导入 register_cuda_ci 的语句改为同时导入 register_amd_ci 和 register_cuda_ci。
2. 注册调用：在已有的 register_cuda_ci(...) 行之后，新增一行 register_amd_ci(est_time=<估计用时>, suite="stage-b-test-2-gpu-large-amd")。其中 est_time 参数为预期运行时间（秒），suite 指定目标 CI 套件名，该名称在 AMD CI 的工作流配置中定义（mi3xx 和 rocm720 两条 AMD 通道均使用同一套件名）。

四个变更文件及其变更内容如下：

- test/registered/scheduler/test_load_snapshot_server.py：第 13 行导入增加 register_amd_ci，第 23 行新增 register_amd_ci(est_time=450, suite="stage-b-test-2-gpu-large-amd")。该测试进行服务器快照加载集成测试（无 DP/ 正常 DP × ZMQ/SHM 组合）。
- test/registered/attention/test_gemma4_swa_triton_oob_regression.py：第 18 行导入增加 register_amd_ci，第 27 行新增 register_amd_ci(est_time=160, suite="stage-b-test-2-gpu-large-amd")。该测试针对 Gemma4 模型在 Triton 注意力后端下启用确定性推理时的高并发越界回归。

- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_full.py：第 4 行导入增加 register_amd_ci，第 15 行新增 register_amd_ci(est_time=400, suite="stage-b-test-2-gpu-large-amd")。该测试使用 Qwen3-32B 模型验证统一基数树（Unified Radix Cache）在 full attention 模式下的 KL 散度阈值。
- test/registered/rl/test_patch_torch.py：第 10 行导入增加 register_amd_ci，第 13 行新增 register_amd_ci(est_time=30, suite="stage-b-test-2-gpu-large-amd")。该测试验证 Torch 分布式通信补丁（monkey-patch）的正确性。

此外，PR 的提交历史反映了从 10 个候选测试缩减到最终 4 个的过程：第一版注册了 10 个测试，第二版因 kv_canary 内核是 CUDA-only 而移除了 4 个依赖于它的测试，第三版因 MLA 测试模型在 AMD 运行器上不可用而再移除 2 个。

关键文件：

- test/registered/scheduler/test_load_snapshot_server.py（模块 调度器；类别 test；类型 test-coverage）：注册 AMD CI 的 4 个测试之一，测试负载快照服务器功能
- test/registered/attention/test_gemma4_swa_triton_oob_regression.py（模块 注意力；类别 test；类型 test-coverage）：注册 AMD CI 的 4 个测试之一，测试 Gemma4 SWA Triton OOB 回归
- test/registered/radix_cache/unified_radix_tree/test_unified_radix_cache_kl_full.py（模块 缓存层；类别 test；类型 test-coverage）：注册 AMD CI 的 4 个测试之一，测试统一基数树 full attention KL 散度
- test/registered/rl/test_patch_torch.py（模块 强化学习；类别 test；类型 test-coverage）：注册 AMD CI 的 4 个测试之一，测试 Torch 分布式补丁

关键符号：未识别

评论区精华

本 PR 的讨论主要来自 PR body 和机器人评论，无人工 review 评论或争议。

- 作者在 PR body 中详细说明了从 10 个候选测试中筛选到 4 个的决策过程：6 个测试因技术原因被排除，包括 kv_canary 内核仅 CUDA（导致 4 个测试无法通过）和 MLA 测试模型在 AMD 运行器上返回 404（导致 2 个测试失败）。
- gemini-code-assist 机器人提示达到每日配额限制，但未影响后续操作。
- amd-bot 自动生成了一个 Claude Code Review，确认了变更的正确性，指出 rocm720 通道使用相同的 stage-b-test-2-gpu-large-amd 套件名，因此一个注册调用即可覆盖两条 AMD CI 通道。
- 审核者 HaiShaw 直接批准，无其他评论。
- amd-bot 自动确认注册逻辑正确 (other): 变更正确，无需调整。

风险与影响

- 风险：风险极低。
- 本 PR 仅涉及 CI 注册代码的变更，不修改任何生产逻辑、测试内容或基础设施配置。

- 所有 4 个测试已在两条 AMD CI 通道（mi3xx 和 rocm720）上成功通过验证（参见 PR body 中引用的 Actions 运行链接）。
- 唯一潜在风险是注册时指定的 *est_time* 估计可能偏短，导致 CI 超时误报，但目前设置的估计值（450s、160s、400s、30s）均大于验证运行中的实际耗时（235s、273s、506s、未报告），因此该风险极小。
- 另外，导入语句的修改与已有代码风格一致，无语法风险。
- 影响：影响范围小，但意义重要。
- 对系统：无影响（仅 CI 配置变更）。
- 对用户：无影响（用户不会直接感知）。
- 对团队：增强了 AMD 硬件平台上的 CI 覆盖，有助于在 AMD GPU 上提前发现回归问题，降低 AMD 特定 bug 流入发布版本的风险。这是 NVIDIA→AMD CI 覆盖差距审计的第四批，展现了团队对多平台支持的持续投入。
- 影响的测试覆盖范围：负载快照服务器集成、Gemma4 SWA Triton OOB 回归、统一基数树 KL 散度 full attention、Torch 分布式补丁。
- 风险标记：测试配套变更，CI 配置变更

关联脉络

- PR #25208 [AMD] ci: register first batch of 2-gpu tests for AMD: 同系列 AMD CI 覆盖审计的第一批
- PR #25939 [AMD] ci: register second batch of 2-gpu tests for AMD: 同系列 AMD CI 覆盖审计的第二批
- PR #27817 [AMD] ci: register third batch of 2-gpu tests for AMD: 同系列 AMD CI 覆盖审计的第三批