PR #20294 完整报告

sgl-project/sglang

[AMD] Add 4-GPU test suite for MI325 runners

合并时间：2026-03-25 05:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20294

执行摘要

该 PR 为 AMD MI325 runner 新增 4-GPU CI 测试套件，包括 3 个 per-commit 测试和 1 个 nightly 测试，通过修改 CI 工作流和测试注册逻辑实现，严格隔离 AMD 更改以保持 NVIDIA 路径不变，提升 AMD 平台测试覆盖和可靠性。

功能与动机

PR body 中明确指出动机是 'Add 4-GPU AMD CI test coverage on MI325 runners'，解决 AMD 平台在 4-GPU 配置下的测试缺失问题。作者 michaelzhang-ai 在 Issue 评论中强调需 ' 严格隔离 AMD 更改 '，确保 NVIDIA 代码路径不受影响，这反映了跨平台 CI 测试的设计需求。

实现拆解

实现分为两个主要模块：

- CI 工作流模块：在 .github/workflows/ 下的四个 YAML 文件中新增 jobs，如 stage-c-test-4-gpu-amd 和 nightly-4-gpu，配置 AMD MI325 runner 和 ROCm 环境。关键代码片段： ```yaml nightly-4-gpu-rocm720: runs-on: linux-mi325-4gpu-sglang steps:
  - name: Nightly Test ROCm 7.2 (4-GPU) run: bash scripts/ci/amd/amd_ci_exec.sh python3 run_suite.py --hw amd --suite nightly-amd-4-gpu ```
- 测试逻辑模块：在多个测试文件中使用 register_amd_ci() 和 is_in_amd_ci() 实现 AMD-specific 调整： | 文件 | AMD 调整 | NVIDIA 路径 | |------|---------|------------| | test_pp_single_node.py | 阈值从 0.74 降至 0.70 | 保持 0.74 | | test_eagle_dp_attention.py | 后端切换为 triton | 保持 fa3 | | test_dp_attention_large.py | 跳过 MLA 类测试 | 正常运行 |

评论区精华

review 中无具体讨论，但 Issue 评论中作者 michaelzhang-ai 与 HaiShaw 交互，强调设计原则：

> "I've cleaned up the PR to strictly isolate AMD changes: No NVIDIA code path changes." 这揭示了在跨平台 CI 测试中，通过条件分支和注册系统隔离更改的重要性，避免引入回归风险。

风险与影响

- 技术风险：AMD-specific 逻辑可能引入假阳性，如 test_eagle_dp_attention.py 中 accuracy 阈值降低可能掩盖性能问题；新增 CI jobs 增加资源消耗，但使用专用 runner 控制影响。
- 影响范围：对用户，提升 AMD 4-GPU 环境下的软件可靠性；对团队，建立标准化 AMD 测试流程，便于后续扩展；对系统，CI 流水线增加 4 个 jobs，但未改变 NVIDIA 测试行为。

关联脉络

与历史 PR 关联显示 AMD 平台测试的持续演进：

- PR #21193 修复 AMD 夜间测试的不兼容性，与本 PR 共同完善 AMD CI 覆盖。
- PR #21239 重构 JIT 内核 CI 使用 run_suite.py 注册系统，与本 PR 的 register_amd_ci() 方法类似，体现仓库向中央化测试注册的演进趋势。 结合 commit 历史（44 次提交），本 PR 经过多次迭代调整阈值和配置，最终采用严格隔离设计，反映了在多平台环境中平衡测试覆盖与代码维护性的技术权衡。