PR #24572 完整报告

sgl-project/sglang

[AMD] Register 5 server-style 1-GPU tests for AMD PR CI

合并时间：2026-05-13 13:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24572

执行摘要

- 一句话：为 AMD CI 注册 5 个服务端 1-GPU 测试
- 推荐动作：值得合并。这是一次规范的 CI 测试注册实践：明确筛选条件、逐个验证、缩小范围。对于关注跨平台 CI 基础设施的读者，可以学习如何安全地将 NVIDIA 测试迁移到 AMD 平台。

功能与动机

Register 5 server-style 1-GPU PR-CI tests on AMD's stage-b-test-1-gpu-small-amd suite. All 5 tests are already running on NVIDIA per-commit CI; they use only Triton attention / CPU torch / mocks / default small models — none of FA3, FlashInfer, ModelOpt, Marlin, EAGLE, GDN/KDA kernels, or NSA-FP8 paths that already have AMD exclusions.

实现拆解

1. 测试筛选：从 12 个候选测试中筛选出 5 个可以在 AMD 上安全运行的测试。这些测试仅依赖 Triton attention、CPU torch、mock 对象或默认小模型，不涉及 FA3、FlashInfer、ModelOpt 等已知 AMD 不兼容的路径。
2. 添加 AMD CI 注册：在 5 个测试文件（test_parallel_state.py、test_input_embeds_chunked.py、test_http2_server.py、test_embed_overrides.py、test_radix_cache_hit.py）中，导入 register_amd_ci，并在已有 register_cuda_ci 调用之后添加对应的 register_amd_ci(est_time=..., suite="stage-b-test-1-gpu-small-amd")。具体改动模式：修改 import 行和新增一行注册调用。
3. 验证：在 AMD MI300 上运行 CI 并确认 5 个测试全部 PASS（CI run 25480011932）。同时确认未对其他 CI 产生负面影响。
4. 移除不可用的候选：在提交历史中，最初有 12 个测试被注册，但后续移除了 6 个因 ROCm 不兼容而失败的测试，最终只保留 5 个。

关键文件：

- test/registered/distributed/test_parallel_state.py（模块 并行状态；类别 test；类型 test-coverage；符号 register_amd_ci, register_cuda_ci）：注册分布式并行状态测试到 AMD CI，该测试使用纯 mock 验证组构建逻辑，无 GPU 需求。

- test/registered/input_embedding/test_input_embeds_chunked.py（模块 输入嵌入；类别 test；类型 test-coverage）：注册分块输入嵌入回归测试到 AMD CI，测试 chunked prefill 和 retraction 形状不匹配 bug。
- test/registered/openai_server/basic/test_http2_server.py（模块 HTTP2 服务器；类别 test；类型 test-coverage）：注册 HTTP/2 服务器测试到 AMD CI，验证 Granian 服务器启动和健康检查 / 模型信息 / 完成接口。
- test/registered/prefill_only/test_embed_overrides.py（模块 嵌入覆盖；类别 test；类型 test-coverage）：注册 token 嵌入覆盖单元测试到 AMD CI，测试 PositionalEmbeds、convert_embeds_to_tensors 等逻辑。
- test/registered/radix_cache/test_radix_cache_hit.py（模块 Radix 缓存；类别 test；类型 test-coverage）：注册 Radix 缓存命中测试到 AMD CI，验证多轮 cache hit 场景。

关键符号：未识别

关键源码片段

test/registered/distributed/test_parallel_state.py

注册分布式并行状态测试到 AMD CI，该测试使用纯 mock 验证组构建逻辑，无 GPU 需求。

```
"""
Tests for distributed parallel state initialization with mocked backend.
"""
from unittest.mock import Mock, patch
import pytest

# 同时导入 AMD 和 NVIDIA 的 CI 注册函数
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci

# 注册到 NVIDIA CI（已存在）
register_cuda_ci(est_time=8, suite="stage-b-test-1-gpu-small")
# 注册到 AMD CI（新增行）
register_amd_ci(est_time=8, suite="stage-b-test-1-gpu-small-amd")

parallel_state = pytest.importorskip("sglang.srt.distributed.parallel_state")
```

评论区精华

由于这是纯测试注册变更，review 评论极少，只有 amd-bot 自动回复 CI 状态和 gemini-code-assist 的配额提示。但提交历史揭示了重要设计决策：初始从 12 个测试开始，作者在 AMD CI 上验证后删除了 6 个有 ROCm 兼容性问题的测试，最终只保留 5 个确认为安全的测试。这一决策体现了对 AMD 平台兼容性的谨慎态度，确保不会引入假阳性失败。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改测试注册文件，不涉及任何生产代码。所有测试已在 NVIDIA CI 上稳定运行，且只使用 CPU/mock/ 默认小模型，不依赖 AMD 不兼容的 GPU kernel。主要风险是 AMD CI 环境本身的不稳定性（PR body 提及已有 pre-existing failures），但与本 PR 无关。
- 影响：对用户无直接影响。对系统的影响是扩大了 AMD PR CI 的测试覆盖，确保 AMD MI300 等平台的基本服务器功能（健康检查、模型信息、完成接口、Radix 缓存命中、分块输入嵌入等）在 CI 中得到验证。对团队来说，减少了 AMD 平台回归漏测的风险，并统一了 NVIDIA 和 AMD 的测试标准。
- 风险标记：仅测试文件变更，依赖 AMD CI 环境稳定，无生产代码改动

关联脉络

- PR #24569 [AMD] Register 8 unit tests for AMD PR CI: 首批 1-GPU 测试注册，本 PR 是第二批服务端测试注册。