

PR #29058 完整报告

sgl-project/sglang

[AMD] Register scripted-core chunked-prefill test for AMD extra-a CI

合并时间: 2026-06-25 16:50

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29058>

执行摘要

- 一句话: 为 AMD CI 注册 scripted-core 测试
- 推荐动作: 该 PR 简单明确, 变更安全, 可快速合并。值得关注的是 CI 注册模式和 `est_time` 设定方法, 对于后续 AMD 测试注册有参考价值。

功能与动机

该 PR 补充了 AMD 上 **extra-a** CI 套件的测试覆盖, 原本只有 CUDA 注册。PR body 指出 “This follows the extra-a AMD onboarding established in #28850 (mirror the CUDA stage; do not promote into per-commit stage-b-*-amd)”。目标是在不提升 CI 频率的前提下, 使 AMD 也能运行关键的调度器模型无关测试。

实现拆解

1. 修改导入语句: 在 `test/registered/chunked_prefill/test_scripted_core_1gpu.py` 中, 将 `from sglang.test.ci.ci_register import register_cuda_ci` 改为 `from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci`。
2. 添加 AMD 注册: 在现有的 `register_cuda_ci(est_time=300, stage="extra-a", runner_config="1-gpu-small")` 之后, 新增一行 `register_amd_ci(est_time=198, stage="extra-a", runner_config="1-gpu-small-amd")`。`est_time` 设置为两个 AMD 平台 (MI325 110s, ROCm 720 198s) 的最大实测耗时 198s, 以便 LPT 自动分区器根据真实 AMD 耗时进行平衡。
3. 验证: 分别在 AMD MI325 和 ROCm 7.2 上运行该测试, 确认通过且未跳过。NVIDIA CI 不受影响 (AMD marker 在 CUDA 上是运行时无操作)。

关键文件:

- `test/registered/chunked_prefill/test_scripted_core_1gpu.py` (模块 脚本调度器; 类别 test; 类型 test-coverage): 唯一的变更文件, 添加了 `register_amd_ci` 导入和调用, 将脚本式调度器测试注册到 AMD extra-a CI 套件。

关键符号: 未识别

关键源码片段

`test/registered/chunked_prefill/test_scripted_core_1gpu.py`

唯一的变更文件，添加了 `register_amd_ci` 导入和调用，将脚本式调度器测试注册到 AMD extra-a CI 套件。

```
import unittest

from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.scripted_runtime.context import ScriptedContext
from sglang.test.scripted_runtime.test_case import ScriptedTestCase
from sglang.test.scripted_runtime_chunked_helpers import (
    LIFECYCLE_STAGES,
    advance_to_lifecycle_stage,
    base_engine_kwargs,
    run_until_finished,
)

# CUDA 注册保持不变: 300s 预估时间, extra-a 阶段, 1-gpu-small 配置
register_cuda_ci(est_time=300, stage="extra-a", runner_config="1-gpu-small")
# AMD 注册: 198s 预估时间 (取 MI325 110s 和 ROCm 720 198s 的最大值),
# 同样放在 extra-a 阶段, 使用 1-gpu-small-amd 配置
register_amd_ci(est_time=198, stage="extra-a", runner_config="1-gpu-small-amd")

_CHUNK_SIZE = 64
_PROMPT_LEN = 4 * _CHUNK_SIZE - 3
_NUM_MIDDLE_CHUNKS = (_PROMPT_LEN - 1) // _CHUNK_SIZE
_LIFECYCLE_MAX_NEW_TOKENS = 4

def _advance_to_stage(r, stage: str):
    yield from advance_to_lifecycle_stage(
        r,
        stage,
        num_middle_chunks=_NUM_MIDDLE_CHUNKS,
        max_new_tokens=_LIFECYCLE_MAX_NEW_TOKENS,
    )

class TestScriptedCore(ScriptedTestCase):
    ENGINE_KWARGS = base_engine_kwargs(chunked_prefill_size=_CHUNK_SIZE)

    def test_chunked_prefill_smoke(self):
        self.server.execute_script(self._script_chunked_prefill_smoke)

    @staticmethod
    def _script_chunked_prefill_smoke(t: ScriptedContext):
        r = t.start_req(prompt_len=_PROMPT_LEN, max_new_tokens=3)
        yield from run_until_finished(r)
        assert r.finished, "req did not finish"
```

评论区精华

没有实质性 review 讨论，PR 已由 HaiShaw 批准。提交者 michaelzhang-ai 在 issue 评论中解释了 CI 红色标记是预存的 NVIDIA/infra 问题，而非本 PR 导致。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。改动仅添加一行注册调用和一个导入，不修改任何测试逻辑或生产代码。AMD 测试在真实硬件上验证通过，NVIDIA CI 不受影响。
- 影响：对用户无直接影响。对 CI 系统：AMD extra-a-test-1-gpu-small-amd 套件从 30 个测试增加到 31 个，AMD PR CI 可以捕获 scripted 调度器相关的回归。对团队：AMD CI 覆盖更完整，有助于提前发现调度器问题。
- 风险标记：暂无

关联脉络

- PR #28850 Previous extra-a AMD onboarding: PR 正文明确提到“follows the extra-a AMD onboarding established in #28850”，该 PR 建立了 AMD extra-a 注册模式。