PR #36307 完整报告

sgl-project/sglang

[AMD] [CI] Stabilize PyTorch sampling backend test on ROCm

合并时间：2026-08-27 09:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36307

执行摘要

- 一句话：ROCm 测试服务器限制并发请求至 64
- 推荐动作：建议查看该 PR 以了解如何在测试中针对特定平台调整资源配置，避免类似资源不足问题。值得关注的是使用 is_hip() 条件参数化测试启动，保持 CUDA 行为不变的设计。

功能与动机

在 ROCm 7.2.4 环境中，test_pytorch_sampling_backend 会因 HSA_STATUS_ERROR_OUT_OF_RESOURCES 而中止。服务器自动解析出 max_running_requests=4029，导致 AITER 从该请求池容量分配注意力工作区，即使测试仅驱动 32 个客户端线程，仍分配约 32 GiB 内存，触发资源不足。

实现拆解

1. 修改测试服务器启动参数：在 test/registered/sampling/test_pytorch_sampling_backend.py 的 setUpClass 中，将原本固定的 other_args 列表改为变量，并根据 is_hip() 条件追加 --max-running-requests 64。
2. 引入 is_hip 导入：从 sglang.srt.utils 导入 is_hip，用于判断当前平台是否为 HIP/ROCm。
3. 保持 CUDA 行为不变：非 HIP 平台仍使用自动推导的请求数限制，确保 CUDA/NVIDIA 测试参数与原先一致。
4. 测试覆盖：仅修改测试文件，无源码改动，不影响采样覆盖率（MMLU 使用 32 线程，greedy 批次为 10）。

关键文件：

- test/registered/sampling/test_pytorch_sampling_backend.py（模块 采样后端；类别 test；类型 test-coverage）：这是本 PR 唯一修改的文件，通过条件化启动参数限制 ROCm 下的并发请求数，避免 AITER 工作区过度分配导致资源不足。

关键符号：setUpClass, test_mmlu, test_greedy

关键源码片段

test/registered/sampling/test_pytorch_sampling_backend.py

这是本 PR 唯一修改的文件，通过条件化启动参数限制 ROCm 下的并发请求数，避免 AITER 工作区过度分配导致资源不足。

```
# test/registered/sampling/test_pytorch_sampling_backend.py
import unittest
from types import SimpleNamespace

import requests

from sglang.srt.utils import is_hip, kill_process_tree
from sglang.test.ci.ci_register import register_amd_ci, register_cuda_ci
from sglang.test.run_eval import run_eval
from sglang.test.test_utils import (
    DEFAULT_MODEL_NAME_FOR_TEST,
    DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
    DEFAULT_URL_FOR_TEST,
    CustomTestCase,
    is_in_amd_ci,
    popen_launch_server,
)

register_cuda_ci(est_time=80, stage="base-b", runner_config="1-gpu-small")
register_amd_ci(est_time=66, suite="stage-b-test-1-gpu-small-amd")


class TestPyTorchSamplingBackend(CustomTestCase):
    @classmethod
    def setUpClass(cls):
        cls.model = DEFAULT_MODEL_NAME_FOR_TEST
        cls.base_url = DEFAULT_URL_FOR_TEST
        other_args = ["--sampling-backend", "pytorch", "--disable-radix-cache"]
        # 仅在 HIP/ROCm 平台限制最大并发请求数，避免 AITER 工作区过度分配导致资源不足
        if is_hip():
            other_args.extend(["--max-running-requests", "64"])
        cls.process = popen_launch_server(
            cls.model,
            cls.base_url,
            timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
            other_args=other_args,
        )
```

评论区精华

Review 评论为空，审核人 HaiShaw 批准（APPROVED）。作者在 Issue 评论中说明了 ROCm 7.2.4 / MI300 验证已分发，并将目标测试分配至 shard 8。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。改动仅影响测试配置，且仅在 HIP/ROCm 平台生效，CUDA/NVIDIA 测试参数不变。限制并发请求至 64 可能影响测试的负载模拟，但根据 PR 描述，实际并发峰值

约 34，留有足够空间，不会限制测试覆盖。潜在风险是如果未来测试增加并发需求，可能需要调整此上限。

- 影响：影响范围限于 ROCm 平台的 CI 测试稳定性，使 test_pytorch_sampling_backend 在 ROCm 7.2.4 上不再因资源不足而中止。对用户无影响，对团队而言减少了 CI 失败。
- 风险标记：仅影响测试配置，ROCm 平台特定

关联脉络

- PR #36396 [AMD][CI] Add DeepSeek-V4-Flash FP8 accuracy coverage on MI30x: 同为 AMD CI 相关测试稳定性改进。
- PR #36602 [CI] Remove GLM-4.1V-9B-Thinking from nightly VLM MMMU eval: 同为 CI 测试稳定性调整，属于同类型维护。