PR #29345 完整报告

sgl-project/sglang

test(dp-attn): drop --enable-torch-compile from TestDPAttentionDP2TP2

合并时间：2026-06-26 07:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29345

执行摘要

- 一句话：DP Attention 测试移除 torch.compile 掩码
- 推荐动作：此 PR 改动虽小但意义重要：修复了一个测试用例因 torch.compile 掩盖真正竞争风险的问题。建议在双流 MoE 相关 PR 中参考此调整，确保测试标志不隐藏并发问题。

功能与动机

PR #27366 的根因分析 (issue #29142) 显示，*--enable-torch-compile* 会导致 Inductor 将双流 MoE 块编译为单个子图，扁平化 torch.cuda.stream(self.alt_stream) 上下文，使所有 kernel 在默认流上执行，从而暴露 inplace_fused_experts 的原地写后读竞争。移除该标志后，测试以 *eager* kernel 启动 + CUDA graph 保留逐 kernel stream 归属的方式运行，更接近用户实际配置。

实现拆解

1. 在 *test/registered/dp_attn/test_dp_attention.py* 的 TestDPAttentionDP2TP2.setUpClass 中，从 popen_launch_server 的 other_args 列表中删除 *--enable-torch-compile* 和 --torch-compile-max-bs 2 两个参数。
2. 保留 *cuda_graph_config.decode.backend='full'*（可通过其他方式设置），确保端到端 *decode* 覆盖不变。
3. 经本地验证，移除后 TestDPAttentionDP2TP2.test_ebnf_generate_complex_json 连续运行 10 次全部通过，平均约 89 秒。

关键文件：

- test/registered/dp_attn/test_dp_attention.py（模块 DP Attention；类别 test；类型 test-coverage；符号 TestDPAttentionDP2TP2.setUpClass）：移除了 TestDPAttentionDP2TP2 启动参数中的 torch.compile 选项，避免因 Inductor 扁平化 stream 上下文而掩盖双流 MoE 的原地写后读竞争风险。

关键符号：TestDPAttentionDP2TP2.setUpClass

关键源码片段

test/registered/dp_attn/test_dp_attention.py

移除了 TestDPAttentionDP2TP2 启动参数中的 torch.compile 选项，避免因 Inductor 扁平化 stream 上下文而掩盖双流 MoE 的原地写后读竞争风险。

```
# test/registered/dp_attn/test_dp_attention.py
# 变更前 (base): 带有 torch.compile 标志
@classmethod
def setUpClass(cls):
    cls.model = DEFAULT_MODEL_NAME_FOR_TEST_MLA
    cls.base_url = DEFAULT_URL_FOR_TEST
    cls._env_override = envs.SGLANG_DISABLE_CONSECUTIVE_PREFILL_OVERLAP.override(
        True
    )
    cls._env_override.__enter__()
    cls.process = popen_launch_server(
        cls.model,
        cls.base_url,
        timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
        other_args=[
            "--trust-remote-code",
            "--tp", "2",
            "--enable-dp-attention",
            "--dp", "2",
            "--enable-torch-compile", # <-- 移除：该标志会导致 Inductor 扁平化
            "--torch-compile-max-bs", "2", # CUDA stream 上下文，掩盖竞争条件
        ],
    )

# 变更后 (head): 不再传递 torch.compile 相关参数
@classmethod
def setUpClass(cls):
    cls.model = DEFAULT_MODEL_NAME_FOR_TEST_MLA
    cls.base_url = DEFAULT_URL_FOR_TEST
    cls._env_override = envs.SGLANG_DISABLE_CONSECUTIVE_PREFILL_OVERLAP.override(
        True
    )
    cls._env_override.__enter__()
    cls.process = popen_launch_server(
        cls.model,
        cls.base_url,
        timeout=DEFAULT_TIMEOUT_FOR_SERVER_LAUNCH,
        other_args=[
            "--trust-remote-code",
            "--tp", "2",
            "--enable-dp-attention",
            "--dp", "2",
            # 保留 full CUDA graph 捕获（通过其他配置），覆盖端到端 decode
        ],
    )
```

评论区精华

PR 作者与 Oasis-Git 和 ch-wan 线下讨论后直接合并。gemini-code-assist[bot] 自动生成无实质内容的评论。无其他审核争议。

- 移除 torch.compile 以暴露双流 MoE 竞争条件 (correctness): 线下讨论后同意移除该标志，使测试更贴近实际运行配置。

风险与影响

- 风险：风险极低。变更仅删除测试启动参数，不影响任何生产代码逻辑。保留的 full CUDA graph 捕获仍能覆盖 decode 路径，不会降低测试质量。
- 影响：影响范围仅限于 TestDPAttentionDP2TP2 测试类的启动配置。移除 torch.compile 后，该测试类将不再受 Inductor 编译行为影响，能够暴露双流 MoE 中潜在的 stream 竞争问题。对其他测试类和功能无影响。
- 风险标记：暂无

关联脉络

- PR #29142 Root cause issue for dual-stream MoE flush hazard: 此 PR 的根因分析来源于 issue #29142，该 issue 定位了 torch.compile 对双流 MoE 的竞争掩盖问题。