

PR #34779 完整报告

sgl-project/sglang

[CI] Disable the prefill CUDA graph on the P worker of test_kimi_linear_pd_dcp4

合并时间: 2026-08-14 06:26

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34779>

执行摘要

- 一句话: 禁用 Kimi PD 测试中 P worker 的 prefill CUDA graph
- 推荐动作: 该 PR 是纯测试一致性修复, 值得快速合入, 但无深度技术洞察, 可作为测试配置最佳实践的参考。

功能与动机

在 `test/registered/disaggregation/test_kimi_linear_pd_dcp4.py` 中, 单体参考服务器 (`_monolithic_reference_args`) 和解码 worker (`start_decode`) 均已使用 `--cuda-graph-backend-prefill disabled`, 唯独预填充 worker (`start_prefill`) 未设置, 导致它捕获了可破坏的预填充 CUDA graph (BCG), 而与之对比的其他进程都没有使用。为保持测试一致性, 需要为 P worker 添加相同的禁用参数。

实现拆解

1. 变更入口: 在 `test/registered/disaggregation/test_kimi_linear_pd_dcp4.py` 的 `start_prefill` 方法中, 向 `args` 列表添加 `--cuda-graph-backend-prefill disabled` 两个元素。
2. 位置: 放在 `--chunked-prefill-size` 参数之后、`--mem-fraction-static` 之前, 确保启动参数正确传递。
3. 配套测试: 无新增测试, 本 PR 是测试调整, 已通过 CI 重新运行该测试。

关键文件:

- `test/registered/disaggregation/test_kimi_linear_pd_dcp4.py` (模块 PD 测试; 类别 test; 类型 test-coverage; 符号 `start_prefill`): 添加 `--cuda-graph-backend-prefill disabled` 到 P worker 启动参数, 消除与参考服务器和 D worker 的配置不一致。

关键符号: `start_prefill`

关键源码片段

`test/registered/disaggregation/test_kimi_linear_pd_dcp4.py`

添加 `--cuda-graph-backend-prefill disabled` 到 P worker 启动参数, 消除与参考服务器和 D worker 的配置不一致。

```
# test/registered/disaggregation/test_kimi_linear_pd_dcp4.py
```

```
@classmethod
def start_prefill(cls):
    args = [
        "--trust-remote-code",
        "--disaggregation-mode", "prefill",
        # ... 其他参数 ...
        "--chunked-prefill-size", str(CHUNKED_PREFILL_SIZE),
        # 新增参数: 禁用 prefill CUDA graph, 与参考服务器和 D worker 保持一致
        "--cuda-graph-backend-prefill", "disabled",
        "--mem-fraction-static", "0.80",
    ]
```

评论区精华

无 review 评论或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低，仅修改测试启动参数，不影响运行时逻辑。但需注意，禁用 prefill CUDA graph 可能降低该测试场景下的性能模拟真实度，若测试意图是验证带 CUDA graph 的 prefill 性能，则需重新评估。
- 影响：影响范围很小，仅改动 test_kimi_linear_pd_dcp4 测试，使其 P worker 配置与参考服务器和 D worker 一致，消除测试环境不一致导致的潜在误判。
- 风险标记：测试配置调整

关联脉络

- PR #32039 [AMD][Fix] Route MoRI through the Qwen MoE all-to-all path: 涉及 PD 测试和模型路由，与本 PR 的 PD 测试配置调整有潜在关联。