PR #24259 完整报告

sgl-project/sglang

[AMD] enable sdma for moriep unittest

合并时间：2026-05-02 14:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24259

执行摘要

- 一句话：启用 SDMA 并调整 chunked prefill 大小以节省显存
- 推荐动作：该 PR 为常规测试维护，无需精读。但可作为 AMD MoE 测试配置的参考。

功能与动机

PR body 指出：

1) 在 moriep TBO（two-batch overlap）用例中启用 SDMA； 2) 降低 chunked prefill size 以节省 GPU 显存。目的是提高测试覆盖率和资源利用率。

实现拆解

该 PR 仅修改了一个测试文件，变更分为两部分：

1. 降低 chunked prefill size 并新增 max-total-tokens 参数：在 common_args 列表中，将 --chunked-prefill-size 从 32768 改为 16384，并新增 --max-total-tokens 参数设置为 131072。此举在保证上下文长度（12288）不变的前提下，通过减小 chunk size 来节省 GPU 显存，同时限制总 token 数以控制资源使用。
2. 启用 SDMA 环境变量：在 TestMTPwithTBONormal 和 TestMTPwithTBOLowLatency 两个测试类的 setUpClass 方法中，添加了 env["MORI_ENABLE_SDMA"] = "true"。这启用了 DeepEP MoE 通信中的 SDMA（可能是 Shared DMA 或某种 DMA 优化），用于验证 SDMA 在 speculative decoding（MTP）与 two-batch overlap 结合场景下的正确性。

关键文件：

- test/registered/amd/test_moriep_small.py（模块 测试；类别 test；类型 test-coverage）：唯一的变更文件，包含所有配置调整和 SDMA 启用。

关键符号：未识别

关键源码片段

test/registered/amd/test_moriep_small.py

唯一的变更文件，包含所有配置调整和 SDMA 启用。

```
# 在 common_args 中降低 chunk size 并新增 max-total-tokens
common_args = [
```

```
    # ... 其他参数 ...
    "--chunked-prefill-size",
    "16384", # 从 32768 降低为 16384，节省显存
    "--max-running-requests",
    "128",
    "--context-length",
    "12288",
    "--max-total-tokens",
    "131072", # 新增参数，限制最大 token 数
    # ... 其他参数 ...
]

# 在 TestMTPwithTBONormal.setUpClass 中启用 SDMA
env["SGLANG_ENABLE_SPEC_V2"] = "false"
env["MORI_ENABLE_SDMA"] = "true" # 启用 SDMA，用于 TBO 场景

# 在 TestMTPwithTBOLowLatency.setUpClass 中同样启用 SDMA
# env["MORI_DISABLE_P2P"] = "1" # 原注释保留
env["MORI_ENABLE_SDMA"] = "true" # 启用 SDMA
```

评论区精华

无 review 评论，仅有 HaiShaw 的批准。无争议或讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅涉及测试配置参数和环境变量，不修改核心逻辑。降低 chunked prefill size 可能影响 prefill 阶段的吞吐，但测试中上下文长度未变，且新增了 max-total-tokens 作为保护。SDMA 启用依赖底层环境，但环境变量为布尔值，未启用时会回退。
- 影响：影响范围：仅限 AMD CI 中 8-GPU 场景下的 MoE+EP 测试用例。用户无感知。团队收益：节省 GPU 显存（chunk size 减半），并验证 SDMA 在 TBO 下的兼容性。
- 风险标记：测试配置变更，仅 AMD 环境

关联脉络

- PR #24241 [bugfix] Support MIXED forward mode in TBO splitter for DP attention: 同为 TBO 相关修复，该 PR 中的 TBO 测试依赖于正确的 TBO splitter 实现。