PR #25301 完整报告

sgl-project/sglang

[AMD] fix moriep unittest oom on mi300x ci

合并时间：2026-05-18 15:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25301

执行摘要

- 一句话：修复 AMD CI 上 moriep 单测 OOM
- 推荐动作：该 PR 为纯粹的测试与 CI 配置调整，不涉及核心逻辑，普通读者无需精读。但 AMD 相关的开发者可关注参数调整逻辑，理解如何在有限显存下配置 MoE 测试。设计决策：通过降低 context length 和 dispatch tokens 有效减少显存占用。

功能与动机

PR body 指出："This patch is to reduce the memory use for moriep unittest to avoid potential oom CI failure found on AMD GPUs with smaller HBM volume"。

实现拆解

1. 调整测试参数(test/registered/amd/test_moriep_small.py)：
   - 引入 DEFAULT_DEEPEP_MODEL_NAME_FOR_TEST_NEXTN 导入，为 MTP 测试指定 draft 模型路径。
   - 降低 --mem-fraction-static 从 0.72 到 0.7，以减少预分配显存。
   - 大幅降低 --chunked-prefill-size 从 16384 到 1024、--context-length 从 12288 到 4096、--max-total-tokens 从 131072 到 32768，以降低峰值内存。
   - 将每个 rank 的最大 dispatch tokens 从 4096 降至 128（通过环境变量 SGLANG_MORI_NUM_MAX_DISPATCH_TOKENS_PER_RANK），减少通信缓冲。
   - 调整准确率阈值：TestPureDP.test_gsm8k 从 >=0.935 降至 >=0.90；TestMTP.test_gsm8k 从 >=0.92 降至 >=0.90。
   - 更新 CI 注册时间 est_time 从 1200 秒到 5400 秒，反映测试耗时增加。
2. 延长 CI job 超时(.github/workflows/pr-test-amd-rocm720.yml 和 pr-test-amd.yml)：
   - 将 job 超时从 60 分钟提升至 120 分钟。
   - 将 per-file 超时从 3600 秒提升至 5400 秒，防止测试因耗时过长被 kill。

关键文件：

- test/registered/amd/test_moriep_small.py（模块 MoE 测试；类别 test；类型 test-coverage）：核心测试文件，通过降低多个内存敏感参数修复 OOM，并调整准确率阈值。

- .github/workflows/pr-test-amd-rocm720.yml（模块 CI 配置；类别 infra；类型 infrastructure）：CI 工作流文件，延长了测试 job 超时和 per-file 超时，避免测试因耗时过长被终止。
- .github/workflows/pr-test-amd.yml（模块 CI 配置；类别 infra；类型 infrastructure）：相同超时调整的另一个 CI 工作流文件。

关键符号：未识别

关键源码片段

test/registered/amd/test_moriep_small.py

核心测试文件，通过降低多个内存敏感参数修复 OOM，并调整准确率阈值。

```
# 测试参数配置，通过降低各项内存占用避免 Mi300X 上 OOM
common_args = [
    "--tp-size", "8",
    "--ep-size", "8",
    "--dp-size", "8",
    "--enable-dp-attention",
    "--moe-a2a-backend", "mori",
    "--trust-remote-code",
    "--load-balance-method", "round_robin",
    "--moe-dense-tp-size", "1",
    "--enable-dp-lm-head",
    "--mem-fraction-static", "0.7", # 从 0.72 降至 0.7，减少预分配显存
    "--chunked-prefill-size", "1024", # 从 16384 降至 1024，降低 prefill 阶段内存
    "--max-running-requests", "128",
    "--context-length", "4096", # 从 12288 降至 4096，减少 KV cache
    "--max-total-tokens", "32768", # 从 131072 降至 32768，限制总 token 数
    "--attention-backend", "aiter",
    "--cuda-graph-max-bs", "32",
]

# 环境变量设定：限制每个 rank 最大 dispatch tokens 并启用隔离模式
env = dict(os.environ)
env["SGLANG_USE_AITER"] = "1"
env["SGLANG_MORI_DISPATCH_DTYPE"] = "bf16"
env["SGLANG_MORI_NUM_MAX_DISPATCH_TOKENS_PER_RANK"] = "128" # 从 4096 大幅降低
env["MORI_SHMEM_MODE"] = "ISOLATION" # 避免对称堆内存溢出
```

评论区精华

审核者 bingxche 和 HaiShaw 直接批准，无实质性讨论。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险在于降低了准确率阈值（0.935→0.90，0.92→0.90），可能掩盖精度回归。但考虑到测试仅为快速验证，且生产环境有独立精度测试，此风险可控。内存参数收紧可能导致测试在更大模型或更完整设置下仍 OOM，但当前参数已验证通过 CI。CI 超时延长不会影响其他功能。
- 影响：影响范围仅限 AMD CI 上的 stage-c-test-large-8-gpu-amd 测试套件中的 moriep 测试。不会影响其他硬件平台或生产推理。对系统无性能影响。团队需要关注测试通过后是否仍能有效捕获精度问题。
- 风险标记：测试阈值放宽，内存参数收紧，CI 超时延长

关联脉络

- PR #25285 Fix EPLB mapping for TopK paths: 同为 AMD MoE 相关的 bugfix PR，涉及 TopK 路径修正。
- PR #23760 [MoE] Unify DeepEPMoE+MoriEPMoE through AITER MoeRunner pre/post-permute: 涉及 moriep 的代码统一重构，本 PR 测试的便是 moriep 路径。