PR #27188 完整报告

sgl-project/sglang

[AMD] Fix TP2 DeepSeek-R1 nhead=64 MLA decode crash and add nightly coverage

合并时间：2026-06-04 07:56

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27188

执行摘要

- 一句话：修复 DeepSeek-R1 TP2 时 nhead=64 MLA decode 崩溃并添加夜间测试
- 推荐动作：值得精读。PR 展示了处理多 GPU 内核选择时的边界情况（head count 门控），并提供了完整的回归测试设计。建议关注 persistent 模式与非 persistent 模式的切换条件，以及如何通过 CI 配置覆盖不同 TP 场景。

功能与动机

来自 PR body：'Fixes a Memory access fault crash when running DeepSeek-R1-MXFP4 with TP2 and the AITER persistent MLA decode path.' 以及关联的 ROCm/aiter#3496 报告了在 TP2 下使用 native qh64 persistent kernel 时多 GPU 并发崩溃。

实现拆解

1. 在 python/sglang/srt/layers/attention/aiter_backend.py 中，将 __init__ 方法内条件从 if self.num_head == 32 or self.num_head == 128: 改为 if self.num_head in (32, 64, 128):，使 nhead=64 时启用 fast_mode=True 和 intra_batch_mode=False，从而使用正确的 persistent metadata 模式。
2. 新增两个 nightly 测试文件：test_deepseek_r1_mxfp4_tp2_mi35x.py 和 test_deepseek_r1_mxfp4_tp4_mi35x.py，分别覆盖 TP2 (nhead=64) 和 TP4 (nhead=32) 的 GSM8K 准确率测试，注册到 nightly 测试套件。
3. 在 test/run_suite.py 的 HWBackend.AMD 套件列表中添加两个新套件名称。
4. 在两个 AMD nightly 工作流文件（nightly-test-amd.yml 和 nightly-test-amd-rocm720.yml）中分别添加对应的 job 定义，触发条件、运行环境和步骤。
5. 本地验证：在 MI35X 上 TP2 配置之前 5/5 次崩溃，修复后 3/3 通过；TP4 控制组也通过；非 FP8 KV cache 和 MTP 场景也通过 smoke test。

关键文件：

- python/sglang/srt/layers/attention/aiter_backend.py（模块 注意力层；类别 source；类型 core-logic；符号 AiterMlaBackend.init）：核心修复，改动一行条件，使 nhead=64 启用 persistent 模式，解决多 GPU 崩溃。
- test/registered/amd/accuracy/mi35x/test_deepseek_r1_mxfp4_tp2_mi35x.py（模块 回归测试；类别 test；类型 test-coverage；符号 TestDeepSeekR1MXFP4TP2MI35x, get_model_path, run_gsm8k_benchmark）：新增 TP2 回归测试，直接覆盖崩溃路径，确

保 nhead=64 路径被 nightly CI 测试。

- test/registered/amd/accuracy/mi35x/test_deepseek_r1_mxfp4_tp4_mi35x.py（模块 回归测试；类别 test；类型 test-coverage；符号 TestDeepSeekR1MXFP4TP4MI35x, get_model_path, run_gsm8k_benchmark）：新增 TP4 回归测试，覆盖 nhead=32 的控制路径，与 TP2 测试互补。
- .github/workflows/nightly-test-amd.yml（模块 CI 配置；类别 infra；类型 infrastructure）：在 AMD nightly 工作流中添加两个新 job，定义触发条件、运行环境和步骤。
- .github/workflows/nightly-test-amd-rocm720.yml（模块 CI 配置；类别 infra；类型 infrastructure）：在 ROCm 7.20 nightly 工作流中同步添加两个新 job，确保不同 ROCm 版本都覆盖。
- test/run_suite.py（模块 测试配置；类别 test；类型 test-coverage）：在 NIGHTLY_SUITES 的 HWBackend.AMD 列表中注册两个新套件名称。

关键符号：AiterMlaBackend.init, TestDeepSeekR1MXFP4TP2MI35x.setUpClass, TestDeepSeekR1MXFP4TP4MI35x.setUpClass, run_gsm8k_benchmark

关键源码片段

python/sglang/srt/layers/attention/aiter_backend.py

核心修复，改动一行条件，使 nhead=64 启用 persistent 模式，解决多 GPU 崩溃。

```
# 设置 persistent MLA 解码元数据模式
# 当前 mla_decode_fwd 只支持 fake-nps 在 self.num_head == 16
# 因此所有 num_head 大小都不使用 qh16 内核来模拟
# 它不应该使用 fake-nps (fast_mode=False, intra_batch_mode=True)
# 否则会导致 GPU 故障或精度问题
if self.num_head in (32, 64, 128): # 修复前：只有 32 和 128；修复后：加入 64
    fast_mode = True
    intra_batch_mode = False

# 当前 persistent a16w16 mla_decode 内核不支持 head_num=128
# 需要回退到非 persistent 模式
# 仅当 fp8 kv_cache 时使用 mla_ps_kernel
if (
    self.num_head_padded == 16 or self.num_head_padded == 128
) and self.kv_cache_dtype is not fp8_dtype:
    _use_mla_ps_kernel = False
    fast_mode = False
    intra_batch_mode = False
```

test/registered/amd/accuracy/mi35x/test_deepseek_r1_mxfp4_tp2_mi35x.py

新增 TP2 回归测试，直接覆盖崩溃路径，确保 nhead=64 路径被 nightly CI 测试。

```
# 注册夜间测试套件
register_amd_ci(
    est_time=1800,
    suite="nightly-amd-2-gpu-mi35x-deepseek-r1-mxfp4-tp2",
```

```
    nightly=True,
)

# 常量定义
DEEPSEEK_R1_MXFP4_LOCAL_PATH = "/data2/models/amd-DeepSeek-R1-MXFP4-Preview"
DEEPSEEK_R1_MXFP4_HF_MODEL_ID = "amd/DeepSeek-R1-MXFP4-Preview"
SERVER_LAUNCH_TIMEOUT = 3600
GSM8K_ACCURACY_THRESHOLD = 0.93

# 测试类：TP=2 时每 rank 64 个 heads
class TestDeepSeekR1MXFP4TP2MI35x(unittest.TestCase):
    @classmethod
    def setUpClass(cls):
        cls.model = get_model_path()
        cls.base_url = DEFAULT_URL_FOR_TEST
        cls.num_questions = int(os.environ.get("GSM8K_NUM_QUESTIONS", "200"))

        # 强制使用 AITER 和 persistent MLA 模式
        env = os.environ.copy()
        env["SGLANG_USE_AITER"] = "1"
        env["SGLANG_AITER_MLA_PERSIST"] = "1"

        cls.process = popen_launch_server(
            model=cls.model,
            base_url=cls.base_url,
            timeout=SERVER_LAUNCH_TIMEOUT,
            other_args=[
                "--attention-backend", "aiter",
                "--tp", "2",
                "--chunked-prefill-size", "131072",
                "--disable-radix-cache",
                "--mem-fraction-static", "0.85",
                "--trust-remote-code",
                "--kv-cache-dtype", "fp8_e4m3",
                "--model-loader-extra-config", '{"enable_multithread_load": true}',
            ],
            env=env,
        )
```

评论区精华

Review 中 HaiShaw 直接批准，Lzy17 给出 LGTM。bingxche 在 CI 中观察到准确率问题（TP2 准确率 0.945、TP4 0.965）并提供了链接，但未进一步讨论。该问题未在此 PR 中解决，但准确率阈值（0.93）已达标。

- CI 准确率问题观察 (testing): 未在本次 PR 中解决；准确率阈值 (0.93) 已达标，但偏差需要后续调查。

风险与影响

- 风险：核心更改仅修改一行条件表达式，风险极低。但新增的 nightly 测试依赖于特定硬件（MI35X）和模型路径，若环境配置不正确可能失败。此外，准确率偏差提示 persistent kernel 在 nhead=64 时的浮点行为可能与之前非 persistent 路径略有差异，但 GSM8K 准确率仍在阈值内，且崩溃问题已修复。
- 影响：直接影响所有在 AMD GPU 上使用 AITER 后端、DeepSeek-R1 模型且 TP=2 的用户，解决了一直以来多 GPU 崩溃的问题。对其他模型和配置无影响。新增的 nightly 测试将确保该路径持续被 CI 覆盖。
- 风险标记：AMD 特定修复，测试环境依赖，潜在准确率漂移

关联脉络

- 暂无明显关联 PR