

PR #44946 完整报告

vllm-project/vllm

[Test] Fix one-sided MNNVL alltoall test workspace under-reservation

合并时间: 2026-06-11 00:43

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44946>

执行摘要

- 一句话: 修复 MNNVL one-sided 测试 workspace 不足
- 推荐动作: 此 PR 为一个测试 bugfix, 代码简单, 对理解 FlashInfer one-sided 机制有一定参考价值, 但无需深度精读。

功能与动机

`test_one_sided_dispatch_combine` 测试在初始化 FlashInfer one-sided `MoeAlltoAll` workspace 时未指定 fp8 block-scale payload (`a1q_scale`) 的字节数, 导致 dispatch 区域预留不足, 后续 combine 段溢出触发 FlashInfer 内部断言失败: "workspace insufficient for combine payload region"。

实现拆解

在 `tests/distributed/test_mnnvl_alltoall.py` 的 `_one_sided_data_worker` 函数中, 调用 `manager.initialize()` 时增加参数 `dispatch_scale_bytes_per_token=hidden_size // 16`, 使 workspace 预留大小与后续 dispatch 的实际 payload 匹配。

关键文件:

- `tests/distributed/test_mnnvl_alltoall.py` (模块 分布式通信; 类别 test; 类型 test-coverage): 修复测试用例 `_one_sided_data_worker` 中 workspace 初始化参数缺失问题, 是唯一变更文件。

关键符号: 未识别

关键源码片段

`tests/distributed/test_mnnvl_alltoall.py`

修复测试用例 `_one_sided_data_worker` 中 workspace 初始化参数缺失问题, 是唯一变更文件。

```
# File: tests/distributed/test_mnnvl_alltoall.py
```

```
def _one_sided_data_worker(rank, world_size):
    from vllm.distributed.device_communicators.all2all import (
        FlashInferNVLinkOneSidedManager,
    )
    # ... 省略上下文代码 ...
```

```
hidden_size = 256
tokens_per_rank = 32
experts_per_token = 2
num_experts = world_size * 8

# Initialize the one-sided manager
manager = FlashInferNVLinkOneSidedManager(cpu_group)
manager.initialize(
    max_num_tokens=tokens_per_rank,
    top_k=experts_per_token,
    num_experts=num_experts,
    hidden_size=hidden_size,
    # 关键修复：声明 fp8 block-scale payload 按每 token 的字节数，
    # 之前缺省为 0 导致 dispatch 区域预留不足，combine 段溢出。
    dispatch_scale_bytes_per_token=hidden_size // 16,
)
# ... 后续创建 a1q, a1q_scale, topk_ids 等数据进行 dispatch/combine ...
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险：仅改动测试文件逻辑，不涉及生产代码。但如果 `dispatch_scale_bytes_per_token` 的计算方式（`hidden_size // 16`）与未来实际 payload 不一致，可能再次出现类似问题。
- 影响：仅限于 `tests/distributed/test_mnnvl_alltoall.py::test_one_sided_dispatch_combine` 这一测试用例，修复后该用例可正常通过。不影响其他测试或功能。
- 风险标记：测试配套，低影响

关联脉络

- 暂无明显关联 PR