PR #22972 完整报告

sgl-project/sglang

[NPU] fix normal DeepEP mode num_tokens_per_rdma_rank error caused by none

合并时间：2026-06-01 15:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22972

执行摘要

- 一句话：修复 NPU 单机 DeepEP 模式下 RDMA 参数空指针异常
- 推荐动作：建议合并。变更简洁且目标明确，修复了特定配置下的崩溃。可考虑补充单元测试覆盖单机模式下的 on_deepep_dispatch_normal 调用。

功能与动机

当 NPU 单服务器部署且 expert_distribution_recorder_mode = "per_token" 时，DeepEP 分发的正常路径中 num_tokens_per_rdma_rank 为 None，导致 on_deepep_dispatch_normal 方法内调用 .cpu().tolist() 时触发 AttributeError。PR 描述及截图确认该问题。

实现拆解

1. 定位问题：在 python/sglang/srt/eplb/expert_distribution.py 的 DetailSinglePassGatherer.on_deepep_dispatch_normal 方法中，第 418 行直接对 num_tokens_per_rdma_rank 调用 .cpu().tolist()，未考虑单机场景下该参数为 None 的情况。
2. 修改逻辑：将赋值改为使用条件表达式——若 num_tokens_per_rdma_rank is not None，则调用 .cpu().tolist() 转换；否则直接记录 None。
3. 影响范围：该变更仅影响 misc_objects 字典中的键值对，下游 collect 方法返回的字典中包含该字段，但在 RDMA 未启用的环境下，None 值不会导致后续逻辑错误。无其他文件或配置变更。

关键文件：

- python/sglang/srt/eplb/expert_distribution.py（模块 专家分发；类别 source；类型 core-logic）：唯一变更文件，修复了 DetailSinglePassGatherer 中 on_deepep_dispatch_normal 方法对 num_tokens_per_rdma_rank 的空指针异常。

关键符号：on_deepep_dispatch_normal

关键源码片段

python/sglang/srt/eplb/expert_distribution.py

唯一变更文件，修复了 DetailSinglePassGatherer 中 on_deepep_dispatch_normal 方法对 num_tokens_per_rdma_rank 的空指针异常。

```
# python/sglang/srt/eplb/expert_distribution.py
class DetailSinglePassGatherer:
    # ... 其他方法

    def on_deepep_dispatch_normal(
        self,
        layer_idx: int,
        local_physical_count_of_layer: List[int],
        num_tokens_per_rank,
        num_tokens_per_rdma_rank, # 单机环境下可能为 None
        num_tokens_per_expert,
    ):
        self._misc_objects.append(
            dict(
                layer_id=layer_idx,
                num_tokens_per_rank=num_tokens_per_rank.cpu().tolist(),
                # 修复：当 num_tokens_per_rdma_rank 为 None 时直接记录 None
                num_tokens_per_rdma_rank=(
                    num_tokens_per_rdma_rank.cpu().tolist()
                    if num_tokens_per_rdma_rank is not None
                    else None
                ),
                num_tokens_per_expert=num_tokens_per_expert.cpu().tolist(),
            )
        )
```

评论区精华

无 review 讨论线程。gemini-code-assist[bot] 的评论仅确认变更内容，无反馈。最终由 sglang-npu-bot 批准合并。

- 暂无高价值评论线程

风险与影响

- 风险：变更范围极小（仅 6 行），逻辑简单明确。风险较低：若下游代码直接对 num_tokens_per_rdma_rank 进行列表操作而未处理 None，可能在后续环节引发问题。但当前代码中 misc_objects 仅用于收集和返回，未发现依赖此字段为列表的强假设。建议确认下游消费者（如测试或日志）能正确处理 None。
- 影响：仅影响 NPU 单机部署场景且使用 per_token 专家分发记录模式的用户。修复后该类用户不会因 AttributeError 而崩溃，恢复正常运行。对其他硬件（如 GPU）或配置无影响。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR