

PR #29571 完整报告

sgl-project/sglang

fix: include CP size in PP rank offset

合并时间: 2026-06-29 16:01

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29571>

执行摘要

- 一句话: 修复 PP 通信中缺失 CP 维度导致的 rank 偏移错误
- 推荐动作: 值得精读, 尤其是了解分布式 rank 布局和多维度并行交织的通信设计。对理解 sglang 的 (dp, cp, tp) 维度布局有启发。

功能与动机

来自 issue #29567: 当 $pp_size > 1$, $attn_cp_size > 1$, $attn_dp_rank > 0$ 时, PP point-to-point 通信中 dp_offset 缺失 $attn_cp_size$ 因子, 导致 rank 不匹配和潜在挂起。参考官方 rank 布局定义于 `dp_attention.py:250-252`。

实现拆解

1. 修改 `python/sglang/srt/managers/scheduler_pp_mixin.py` 中 `_pp_send_pyobj_to_next_stage` 和 `_pp_recv_pyobj_from_prev_stage` 的 dp_offset 计算, 从 $attn_dp_rank * attn_tp_size$ 改为 $attn_dp_rank * attn_cp_size * attn_tp_size$ 。
2. 修改 `python/sglang/srt/managers/scheduler_components/request_receiver.py` 中 `_pull_raw_reqs` 的 dp_offset 计算, 同样加入 $attn_cp_size$ 。
3. 新增 `test/registered/unit/managers/test_pp_cp_rank_offsets.py`, 通过两个测试用例分别验证 `request_receiver` 和 `scheduler_pp_mixin` 在模拟的 (dp1, cp2, tp2) 布局下产生的预期 rank 值。

关键文件:

- `test/registered/unit/managers/test_pp_cp_rank_offsets.py` (模块 秩偏移; 类别 test; 类型 test-coverage; 符号 `_make_ps`, `_fake_group`, `_make_receiver`, `TestPPCPRankOffsets`): 新增测试, 覆盖 `request_receiver` 和 `scheduler_pp_mixin` 中 rank 计算的修正逻辑。
- `python/sglang/srt/managers/scheduler_pp_mixin.py` (模块 PP 调度; 类别 source; 类型 core-logic; 符号 `_pp_send_pyobj_to_next_stage`, `_pp_recv_pyobj_from_prev_stage`): 核心修改, 修复 PP 对象发送和接收路径中的 rank 偏移计算。
- `python/sglang/srt/managers/scheduler_components/request_receiver.py` (模块 请求接收; 类别 source; 类型 core-logic; 符号 `_pull_raw_reqs`): 修改 PP 接收方在 `pull_raw_reqs` 中的 rank 偏移计算。

关键符号: `_pp_send_pyobj_to_next_stage`, `_pp_recv_pyobj_from_prev_stage`,
`_pull_raw_reqs`

评论区精华

ShangmingCai 在 review 中指出: "Logic looks good. We used to reuse the DP field for CP, since they are separated now, this fix adapts to the new impl." 这揭示此前 CP 和 DP 复用了—个字段, 现在分离后需要显式相乘。

- CP 和 DP 字段分离后的适配 (design): 设计合理, 批准合并。

风险与影响

- 风险: 涉及分布式 P2P 通信核心路径, 若计算错误会导致集群挂起或数据损坏。由于改动简单 (仅一行因子乘法) 且已有单元测试覆盖两个主要路径, 风险可控。但测试仅在 CPU 模拟环境运行, 生产环境多卡组合仍需验证。
- 影响: 仅影响同时启用 PP、CP 和 DP 的部署场景 (如 `pp_size>1`, `attn_cp_size>1`, `attn_dp_rank>0`), 对默认配置 (`pp=1` 或 `cp=1`) 无影响。修复后此类拓扑的 point-to-point 通信不再错位, 减少挂起几率。
- 风险标记: 核心路径变更, 分布式通信

关联脉络

- 暂无明显关联 PR