PR #32210 完整报告

sgl-project/sglang

[NPU] Fix MTP IndexShare warm-up for attention DP and prefill CP

合并时间：2026-07-27 19:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32210

执行摘要

- 一句话：修复 NPU PD 下 MTP IndexShare 的 warm-up 挂起与 CP top-k 形状错误
- 推荐动作：建议精读 can_run_graph() 的实现，尤其是通过 all-reduce 同步跨 rank 决策的模式，这种模式在异构硬件或分布式 warm-up 场景中具有通用性。代码注释充分，commit 历史清晰，适合作为 NPU 调试的参考。

功能与动机

NPU PD disaggregation 场景中，PR #30839 添加的 DSA top-k seed missing fallback 在 attention DP 下导致 decode warm-up hang，因为只有真实 rank 缺少 seed 而空闲 rank 无请求，它们会走不同代码路径，而 draft forward 含 TP/EP 集体通信，路径不一致造成死锁。另一问题是 PR #30992 引入的 prefill CP IndexShare 假设 topk_indices 是 tensor，但 NPU CP 路径返回 tuple，引发 AttributeError。

实现拆解

1. 同步解码图 / eager 决策：在 python/sglang/srt/hardware_backend/npu/graph_runner/eagle_draft_npu_graph_runner.py 中新增 can_run_graph() 覆盖，保留父类局部检查，添加 DSA seed 是否就绪的本地判断（空闲 rank 或 seed 存在则允许），然后通过 torch.distributed.all_reduce(decision, op=MIN) 在整个 TP group 上取最小值，确保所有 rank 统一决策。
2. 归一化 CP topk_indices 形状：在 python/sglang/srt/layers/attention/dsa/dsa_indexer.py 的 do_npu_cp_balance_indexer() 中，修改返回值为 torch.cat([topk_indices_prev[0], topk_indices_next[0]], dim=0).squeeze(1)，将 tuple 拼接为 tensor 并移除 singleton head 维度。
3. 调整 CP 平衡注意力接口：在 python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py 的 do_cp_balance_attn() 中，调用新工具函数 _expand_dsa_sparse_indices() 来标准化 topk_indices 后再进行 torch.split。

关键文件：

- python/sglang/srt/hardware_backend/npu/graph_runner/eagle_draft_npu_graph_runner.py（模块 NPU 图运行器；类别 source；类型 core-logic；符号 can_run_graph）：核心改动：新增 can_run_graph() 方法，通过 MIN all-reduce 同步跨 rank 的图 / eager 决策，避免 DP 下的 deadlock。

- python/sglang/srt/layers/attention/dsa/dsa_indexer.py（模块 DSA 索引器；类别 source；类型 core-logic）：修改 do_npu_cp_balance_indexer() 返回值为拼接后的 tensor，修复 CP 路径下 topk_indices 形状错误导致的 AttributeError。
- python/sglang/srt/hardware_backend/npu/attention/ascend_backend.py（模块 Ascend 注意力；类别 source；类型 core-logic）：调整 do_cp_balance_attn 使用 _expand_dsa_sparse_indices 归一化 topk_indices，配合索引器改动。

关键符号：can_run_graph, do_cp_balance_attn, do_npu_cp_balance_indexer

关键源码片段

python/sglang/srt/hardware_backend/npu/graph_runner/eagle_draft_npu_graph_runner.py

核心改动：新增 can_run_graph() 方法，通过 MIN all-reduce 同步跨 rank 的图 / eager 决策，避免 DP 下的 deadlock。

```
def can_run_graph(self, forward_batch: ForwardBatch) -> bool:
    # 先调用父类检查基础条件（如 batch size、模式等）
    can_run_graph = super().can_run_graph(forward_batch)
    # 如果 DSA seed 不从 draft extend 获取，或者 attention DP 大小为 1，
    # 则无需多 rank 同步，直接返回父类决策
    if (
        not self.eagle_worker.seed_dsa_topk_from_draft_extend
        or self.attn_dp_size <= 1
    ):
        return can_run_graph

    # PR #30839 在 seed 不可用时回退 eager。在 attention DP 下，
    # seed 可用性按 request 计算：有请求的 rank 可能缺少 seed，
    # 而空闲 rank 没有请求。但 draft forward 包含 TP/EP 集体通信，
    # 所有 rank 必须选择相同的执行路径。因此通过 TP group 的
    # MIN all-reduce 来同步最终决策：任何 rank 的 can_run_graph
    # 或 seed_ready 为 False，则所有 rank 都走 eager。
    spec_info = forward_batch.spec_info
    seed_ready = forward_batch.forward_mode.is_idle() or (
        spec_info is not None and spec_info.dsa_topk_indices is not None
    )
    decision = torch.tensor(
        int(can_run_graph and seed_ready),
        dtype=torch.int32,
        device=self.device,
    )
    torch.distributed.all_reduce(
        decision,
        op=torch.distributed.ReduceOp.MIN,
        group=self.model_runner.tp_group.device_group,
    )
    return bool(decision.item())
```

评论区精华

主要讨论集中在 CI 触发。维护者 iforgetmyname 评论“this pr only affects npu”后触发 /tag-run-ci-label 运行 CI。无设计争议。

- CI 触发 (other): CI 仅运行 NPU 相关测试，维护者确认影响面。

风险与影响

- 风险：风险低：改动仅限 NPU 后端三个文件，不涉及 GPU 或其他硬件；can_run_graph() 的重写使用明确守卫（仅在 seed_dsa_topk_from_draft_extend 启用且 attn_dp_size > 1 时走新路径），不影响默认路径。缺少单元测试覆盖，但 NPU 硬件的 CI 测试可以验证行为。
- 影响：直接影响：NPU 上使用 MTP + PD disaggregation + attention DP 的 DeepSeek 模型（如 deepseek-v3/v4）的 decode warm-up 不再 hang，prefill CP 的 IndexShare 不再崩溃。间接影响：引入新的工具函数 _expand_dsa_sparse_indices，可能被其他 CP 路径复用。
- 风险标记：缺少测试覆盖

关联脉络

- PR #30839 add a rank-local graph-to-eager fallback when DSA top-k seed is missing: 该 PR 的 fallback 逻辑在 attention DP 下引入 deadlock，本 PR 修复此问题。
- PR #30992 Add prefill CP IndexShare path for DSA: 该 PR 引入的 CP topk_gather 假设 topk_indices 为 tensor，但 NPU 返回 tuple，本 PR 修复此假设。