

PR #33523 完整报告

sgl-project/sglang

[npu] [bugfix] Fix PD-disaggregation error

合并时间: 2026-08-05 09:54

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33523>

执行摘要

- 一句话: 修复 NPU PD 分离 send_kvcache 缺参错误
- 推荐动作: 值得快速浏览: 这是一个典型的后端子类与基类 API 漂移的修复样例, 可作为多后端模块 (Mooncake、Ascend、HiSparse) 接口同步维护的参考。建议后续为该共享传输接口补充一个契约测试或基类抽象方法, 防止类似问题再次发生。

功能与动机

PR body 指出: Fix error: got an unexpected keyword argument 'dst_device_kv_indices'。该错误由 PR#31901 引入。AscendKVManager 继承自 MooncakeKVManager, 其 send_kvcache 方法缺少 dst_device_kv_indices 参数, 从而触发错误。本 PR 的目的就是让子类签名与基类保持一致, 恢复 NPU PD-disaggregation 的可用性。

实现拆解

1. 定位问题: 在 python/sglang/srt/disaggregation/ascend/conn.py 中, AscendKVManager.send_kvcache 方法签名比基类 MooncakeKVManager 缺少 dst_device_kv_indices 参数。
2. 补全签名: 新增 dst_device_kv_indices: Optional[npt.NDArray[np.int32]] = None 默认参数, 确保向后兼容, 已有调用不受影响。
3. 显式拒绝不支持场景: 当该参数非 None 时抛出 NotImplementedError, 提示 Ascend PD 传输不支持 HiSparse destination device KV indices, 防止数据被静默截断。
4. 验证方式: 未附带单元测试, 依赖 sglang-npu-bot 触发 run-ci 标签的 NPU 冒烟 CI; 合并者为 sglang-npu-bot。

关键文件:

- python/sglang/srt/disaggregation/ascend/conn.py (模块 传输层; 类别 source; 类型 core-logic; 符号 send_kvcache): 本 PR 的唯一修改文件, 修复 AscendKVManager.send_kvcache 与 MooncakeKVManager 基类签名不一致导致的运行时错误。

关键符号: send_kvcache

关键源码片段

python/sglang/srt/disaggregation/ascend/conn.py

本 PR 的唯一修改文件，修复 AscendKVManager.send_kvcache 与 MooncakeKVManager 基类签名不一致导致的运行时错误。

```
def send_kvcache(
    self,
    mooncake_session_id: str,
    prefill_kv_indices: npt.NDArray[np.int32],
    dst_kv_ptrs: list[int],
    dst_kv_indices: npt.NDArray[np.int32],
    executor: concurrent.futures.ThreadPoolExecutor,
    dst_layer_ids: Optional[List[int]] = None,
    dst_device_kv_indices: Optional[npt.NDArray[np.int32]] = None,
):
    # Ascend 侧暂不支持 HiSparse 目标设备 KV 索引，
    # 传入时显式报错，避免与 Mooncake 基类接口漂移
    if dst_device_kv_indices is not None:
        raise NotImplementedError(
            "Ascend PD transfer does not support HiSparse "
            "destination device KV indices"
        )

    # 将连续 KV 块分组，便于批量传输
    prefill_kv_blocks, dst_kv_blocks = group_concurrent_contiguous(
        prefill_kv_indices, dst_kv_indices
    )

    # 后续按 pp_size 是否大于 1 分支处理 MLA / MHA 的指针切片，
    # 这部分逻辑在本次变更中未改动，保持与之前一致。
    ...
```

评论区精华

没有人工 review 讨论。sglang-npu-bot 直接批准并触发 /tag-and-rerun-ci;

gemini-code-assist[bot] 提示其消费者版审查服务已停用。值得注意的设计点是：显式抛出 NotImplementedError 而不是静默接受参数，避免 Ascend 路径在未来误用 HiSparse 传输语义。

- NPU 改动自动合并 (other): 改动较小，由 bot 直接合并。

风险与影响

- 风险：主要风险是缺少测试配套：仅一个源码文件 +7 行，无单元测试，回归依赖 NPU CI 冒烟覆盖。若后续 Ascend 启用 HiSparse 传输，需要移除该 NotImplementedError 并实现真正的复数索引处理。另外，基类 MooncakeKVManager 若继续演化 send_kvcache 签名，可能再次出现同类漂移。
- 影响：用户影响：恢复 NPU 平台上 PD-disaggregation 的可用性，修复前 NPU 用户会在启用了该功能时直接崩溃。系统影响：改动仅影响 AscendKVManager.send_kvcache 的调

用路径，默认参数 `None` 时行为不变，对已有请求无性能或语义影响。团队影响：提示维护者多后端共享接口的抗漂移意识。

- 风险标记：缺少测试覆盖，基类接口漂移，平台专属修复

关联脉络

- PR #31901（上下文未提供标题）：PR body 明确指出本 bug 由该 PR 引入：它为 `send_kvcache` 增加了 `dst_device_kv_indices` 参数，但 `AscendKVManager` 未同步更新。