

PR #27534 完整报告

sgl-project/sglang

[PD] Downgrade propagated rank failure logs from error to debug

合并时间: 2026-06-08 16:56

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27534>

执行摘要

- 一句话: 降低传播 KV 传输失败日志等级至调试级别
- 推荐动作: PR 设计简洁有效, 降低了生产噪音, 同时保留了原始失败信息。建议关注异常类型变更对第三方扩展的影响, 并未来考虑添加单元测试覆盖传播检测逻辑。

功能与动机

在 PD 分离服务中, 当一个 rank 的 KV 传输失败时, 失败会传播到同一传输房间的所有其他 rank。当前每个 rank 都会记录 error (decode 侧) 或 warning (prefill 侧) 级别日志, 即使该 rank 并非失败源头。这产生了大量重复日志, 使生产环境中的根因定位困难。本 PR 旨在通过区分传播失败与本地发起失败, 将传播失败日志降级为 debug 级别, 从而减少日志噪声。

实现拆解

1. 提取公共异常类: 将原本定义在 mooncake/conn.py 中的 KVTransferError 类迁移到 common/conn.py, 并添加 is_from_another_rank 布尔属性 (默认 False), 用于标记该异常是否由另一个 rank 传播而来。
2. 后端的传播检测: 修改 mooncake/conn.py、mori/conn.py、nixl/conn.py 中 *Connector.failure_exception() 方法的实现。利用 failure_records.pop(room, None)——若弹出 None 则表示当前 room 没有本地记录, 判定为传播失败, 设置 is_from_another_rank=True。同时将 mori 和 nixl 后端的 RuntimeError 替换为 KVTransferError, 以携带传播标记。
3. 日志级别条件化: 在 decode.py 的 _update_handshake_waiters 和 pop_transferred 方法、prefill.py 的 process_disagg_prefill_inflight_queue 和 handle_bootstrap_failure 方法中, 捕获 KVTransferError 后通过 getattr(e, 'is_from_another_rank', False) 判断。若为传播失败, 使用 logger.debug 代替原来的 logger.error (decode 侧) 或 logger.warning (prefill 侧); 否则保持原日志级别。
4. 配套导入调整: 在各后端的 conn.py 中增加 KVTransferError 的导入 (来自 common/conn.py), 并更新原有异常抛出处。

关键文件:

- python/sglang/srt/disaggregation/common/conn.py (模块 分离服务; 类别 source; 类型 core-logic; 符号 KVTransferError, init, str): 新增 KVTransferError 异常类定义, 包含 is_from_another_rank 属性, 作为所有后端传播异常的基础。

- python/sglang/srt/disaggregation/mooncake/conn.py (模块 分离服务; 类别 source; 类型 core-logic; 符号 KVTransferError, init, str) : 移除 KVTransferError 类, 改为从 common 导入; 修改 failure_exception 方法以检测传播失败。
- python/sglang/srt/disaggregation/decode.py (模块 分离服务; 类别 source; 类型 core-logic) : 修改了两处 KV 传输失败处理: _update_handshake_waiters 和 pop_transferred, 根据异常中的 is_from_another_rank 属性决定错误日志级别。
- python/sglang/srt/disaggregation/prefill.py (模块 分离服务; 类别 source; 类型 core-logic) : 修改了两处 prefill 侧失败处理, 同样根据 is_from_another_rank 降级 warning 日志为 debug。
- python/sglang/srt/disaggregation/mori/conn.py (模块 分离服务; 类别 source; 类型 core-logic) : 导入 KVTransferError 并替换原有的 RuntimeError 抛出, 实现传播检测。
- python/sglang/srt/disaggregation/nixl/conn.py (模块 分离服务; 类别 source; 类型 core-logic) : 导入 KVTransferError 并替换原有的 RuntimeError 抛出, 同时 sender 端直接标记为传播。

关键符号: KVTransferError.init, KVTransferError.str,
MooncakeKVSender.failure_exception, MooncakeKVReceiver.failure_exception,
MoriKVSender.failure_exception, MoriKVReceiver.failure_exception,
NixlKVSender.failure_exception, NixlKVReceiver.failure_exception,
DecodeHandler._update_handshake_waiters, DecodeHandler.pop_transferred,
Scheduler.process_disagg_prefill_inflight_queue, Scheduler.handle_bootstrap_failure

关键源码片段

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

新增 KVTransferError 异常类定义, 包含 is_from_another_rank 属性, 作为所有后端传播异常的基础。

```
class KVTransferError(Exception):
    """KV 传输过程中发生的异常, 支持标记是否由另一个 rank 传播而来。"""
    def __init__(
        self,
        bootstrap_room: int,
        failure_reason: str,
        is_from_another_rank: bool = False,
    ):
        super().__init__(failure_reason)
        self.bootstrap_room = bootstrap_room # 传输房间号
        self.failure_reason = failure_reason # 失败原因描述
        self.is_from_another_rank = is_from_another_rank # 是否由其他 rank 传播

    def __str__(self):
        return f"KVTransferError(bootstrap_room={self.bootstrap_room}): {self.failure_reason}"
```

评论区精华

本 PR 的 review 过程中没有产生实质性讨论，仅有一条来自 [gemini-code-assist] 的每日配额警告。CI 重跑指令 `/rerun-group disaggregation` 后所有测试通过。

- 日志降级策略讨论 (design): 无需修改，按原始设计合并。

风险与影响

- 风险:

1. 异常类型变更: mori 和 nixl 后端原先抛出 `RuntimeError`，现在改为 `KVTransferError`。如果上游代码通过 `except RuntimeError` 捕获这些异常，会漏接；但经查看所有异常捕获点均已被更新。
2. 日志完整性: 传播失败日志降级为 `debug` 后，若用户依赖 `error` 级别日志进行告警，可能会漏掉传播类失败。但传播失败最终会由产生失败的原始 rank 记录 `error` 级别，整体告警覆盖仍完整。
3. 测试覆盖: 本次改动未新增测试用例，缺少对传播检测逻辑的自动化验证，后续修改可能引入回归。
 - 影响: 影响范围: 所有使用 PD 分离模式的后端 (mooncake、mori、nixl) 的 KV 传输失败处理路径。影响效果: 生产环境中传播类失败日志量显著降低，避免重复日志淹没真正的根因；但需要运维人员适应这一变化，调试传播类问题时需启用 `debug` 级别日志。对系统性能无影响，仅改动日志调用和异常构造路径。
 - 风险标记: 核心路径变更，缺少测试覆盖，异常类型变更

关联脉络

- PR #26922 [PD][MoRI] Drive KV transfers with a sharded synchronous worker pool: 同样涉及 PD 分离架构中 KV 传输失败处理，修改了 `mori/conn.py` 等文件。