PR #25287 完整报告

sgl-project/sglang

[PD] Un-blacklist mooncake sessions when probe succeeds

合并时间：2026-05-20 11:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25287

执行摘要

- 一句话：Mooncake session 故障黑名单自动恢复
- 推荐动作：值得精读：该 PR 实现简洁而稳健，展示了如何在分布式系统中处理临时故障自动恢复的典型模式：守护线程 + 轻量探测 + 配置化。特别值得学习的是 getattr 回退策略，确保了与旧版 API 的兼容性。建议 reviewer 关注 mooncake 新版本的发布节奏，以便启用该功能。

功能与动机

MooncakeKVManager 首次 KV 传输失败即加入黑名单且无重试机制，若 decode 端因临时故障（网络抖动、GC 暂停等）不再重新注册，后续 prefill→decode 请求会永久失败，直到重启任一方。PR body 明确描述了这一场景及修复目标。

实现拆解

1. 配置环境变量（python/sglang/srt/environ.py）：新增 SGLANG_ENABLE_FAILED_SESSION_PROBE（默认 False）和 SGLANG_FAILED_SESSION_PROBE_INTERVAL_S（默认 30.0），使功能可通过环境变量开关，避免不必要的后台线程开销。
2. 启动探测守护线程（python/sglang/srt/disaggregation/mooncake/conn.py 的 MooncakeKVManager.__init__）：在 prefill 模式下读取配置，若启用则创建一个 threading.Event 作为关闭信号，并启动名为 "MooncakeFailedSessionProbe" 的守护线程运行 _failed_session_probe_loop 方法。
3. 实现单轮探测逻辑（MooncakeKVManager._run_one_probe_pass）：加锁获取 failed_sessions 快照，对每个 session 使用 getattr(self.engine, "send_probe", None) 回退检查，若 send_probe 不存在则返回 -1，否则调用并捕获异常。若返回码为 0，则从 failed_sessions 和 session_failures 中移除该 session，记录恢复日志并递增 Prometheus Counter sglang:failed_session_recoveries_total。
4. 实现探测循环（MooncakeKVManager._failed_session_probe_loop）：以配置的间隔循环，每次调用 _run_one_probe_pass，通过 Event.wait() 实现可中断的睡眠，支持优雅关闭。
5. 传递探测接口（python/sglang/srt/distributed/device_communicators/mooncake_transfer_engine.py）：新增 send_probe(self, peer_session_id: str) -> int 方法，直接调用底层引擎的 send_probe，为探测提供调用入口。

关键文件：

- python/sglang/srt/disaggregation/mooncake/conn.py（模块 PD 分离；类别 source；类型 core-logic；符号 _run_one_probe_pass, _failed_session_probe_loop）：核心变更文件：添加了失败的 session 探测循环和单轮探测逻辑，以及 Prometheus 指标计数器和环境变量开关启用逻辑。
- python/sglang/srt/distributed/device_communicators/mooncake_transfer_engine.py（模块 分布式通信；类别 source；类型 core-logic；符号 send_probe）：新增 send_probe 方法，作为底层引擎探测接口的包装，供探测循环调用。
- python/sglang/srt/environ.py（模块 环境配置；类别 source；类型 configuration）：新增两个环境变量：功能开关和探测间隔，用于控制探测线程的启用和频率。

关键符号：_run_one_probe_pass, _failed_session_probe_loop, send_probe

关键源码片段

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

核心变更文件：添加了失败的 session 探测循环和单轮探测逻辑，以及 Prometheus 指标计数器和环境变量开关启用逻辑。

```
# 关键方法：单轮探测逻辑
# 遍历 failed_sessions 快照，对每个 session 发送 probe
# 使用 getattr 兼容旧版 mooncake（无 send_probe 时 rc=-1）
def _run_one_probe_pass(self) -> None:
    with self.session_lock:
        snapshot = list(self.failed_sessions)
    for session_id in snapshot:
        # getattr 回退：若引擎无 send_probe 属性，返回 -1 表示不可用
        send_probe = getattr(self.engine, "send_probe", None)
        if send_probe is None:
            rc = -1
        else:
            try:
                rc = send_probe(session_id)
            except Exception as e:
                logger.warning("send_probe(%s) raised: %s", session_id, e)
                continue
        if rc == 0:
            with self.session_lock:
                was_blacklisted = session_id in self.failed_sessions
                self.failed_sessions.discard(session_id)
                self.session_failures.pop(session_id, None)
            if was_blacklisted:
                logger.info(
                    "Session %s recovered via probe; un-blacklisted", session_id
                )
                FAILED_SESSION_RECOVERIES.inc()
        else:
```

```
                logger.debug("Probe still failing for %s (rc=%d)", session_id, rc)

    # 守护线程循环：以配置间隔定期执行探测
    # 使用 Event.wait() 实现可中断的睡眠，支持优雅关闭
    def _failed_session_probe_loop(self) -> None:
        logger.info(
            "Starting failed-session probe loop (interval=%.1fs)",
            self.failed_session_probe_interval,
        )
        while not self._failed_session_probe_shutdown.wait(
            self.failed_session_probe_interval
        ):
            self._run_one_probe_pass()
```

评论区精华

Reviewer ShangmingCai 提出了两点关键反馈：

1. 环境变量开关：建议增加 SGLANG_ENABLE_FAILED_SESSION_PROBE 开关，因为并非所有场景都需要此功能，后台线程和 CPU 开销不应由无关用户承担。作者在第二次提交中添加了该开关。
2. API 兼容性：由于 send_probe 是新 API，旧版 mooncake 引擎不存在该方法，建议使用 getattr 降级，让 rc==-1 表示不可用。作者在第三次提交中用 getattr 实现了回退逻辑。最终 Reviewer 批准了 PR，但指出需要等待新版本 mooncake 发布。

- 环境变量开关 (design): 作者在第二次提交中添加了 SGLANG_ENABLE_FAILED_SESSION_PROBE 开关，默认关闭。
- API 兼容性回退 (correctness): 作者在第三次提交中使用 getattr 实现了回退逻辑，当 send_probe 不存在时返回 -1。

风险与影响

- 风险：
  1. mooncake 版本依赖风险：send_probe 需要 mooncake-transfer-engine >= 对应版本（PR body 提及 Mooncake#2088），否则 AttributeError 被捕获后功能降级为无操作，但不会影响现有行为。
  2. 后台线程 CPU 开销：默认 30s 轮询一次，开销极低；但仍需环境变量开关避免非必要场景。
  3. 锁竞争风险：_run_one_probe_pass 中获取 session_lock 两次（快照时和移除时），可能与其他传输线程产生短暂竞争，但锁持有时间短、频率低，风险可控。
  4. 无测试覆盖：当前 PR 未包含单元测试或集成测试，对于新功能的可靠性验证不足。 - 影响：影响范围：仅影响 mooncake 后端的 PD 分离部署（prefill 节点），且默认不启用。用户需要通过设置 SGLANG_ENABLE_FAILED_SESSION_PROBE=1 显式开启。影响程度：中等。解决了生产环境中偶发性网络抖动导致的永久性 KV 传输黑屏问题，提高了系统韧性。新增的 Prometheus 指标 sglang:failed_session_recoveries_total 便于监控恢复情况。

- 风险标记：依赖外部版本，无测试覆盖，默认关闭

关联脉络

- PR #25677 [PD] Clean early abort logic in PD module: 同属 PD 分离模块的清理工作，涉及相同文件 conn.py 的修改，与本 PR 共同提升 PD 模块的健壮性。
- PR #25774 drop output ids: 涉及 schedule_batch 重构，虽然是不同模块，但同样影响 PD 分离路径的稳定性。