PR #25125 完整报告

sgl-project/sglang

[Disagg] Add retry with exponential backoff for prefill bootstrap register

合并时间：2026-05-15 16:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25125

执行摘要

- 一句话：为 prefill bootstrap 注册添加指数退避重试
- 推荐动作：该 PR 修复了实际竞态问题，重试实现稳健（指数退避 + 抖动，异常链遍历），测试用例设计完整。建议学习其测试 mock 策略和日志分级设计。对于最终失败是否崩溃的讨论，可后续考虑添加配置项或强制退出选项。

功能与动机

在 PD 分离模式下，prefill 调度器在初始化时调用 register_to_bootstrap() 发布自身信息，但 bootstrap 服务器在 HTTP 服务器启动时才创建，两者没有同步。对于小模型（如 Qwen3-0.6B），prefill 初始化在几秒内完成，赶在 bootstrap 服务器启动前发起注册，导致连接被拒绝且无重试，最终所有请求 500，唯一恢复方式是重启 prefill 实例。详见 PR 描述中的日志序列。

实现拆解

1. 修改 register_to_bootstrap 方法（conn.py），引入重试循环：设置 max_retries=5、initial_delay=1.0、max_delay=30.0。
2. 每次循环先尝试 HTTP PUT 请求，若状态码为 200 则直接返回；若失败或抛出异常，记录 warning 日志并进入退避逻辑。
3. 退避策略：delay = min(initial_delay * 2^attempt, max_delay) * (0.75 + 0.25 * (time.monotonic() % 1))，确保抖动因子在 [0.75, 1.0) 内，最大延迟不超过 30 秒。
4. 异常处理时遍历 __cause__ 链找到最底层异常，避免 urllib3 包装消息干扰日志可读性。
5. 最后尝试后不 sleep，直接跳出循环并记录 error 日志。
6. 修正 docstring 中 HTTP 方法描述（PUT 而非 POST）。
7. 新增测试文件 test_register_to_bootstrap.py，注册到 CI stage-a-test-cpu 套件（预估 5 秒）。
8. 通过 mock requests.put 和 time.monotonic 覆盖 7 个用例：首次成功、重试成功、全部失败、嵌套异常、无嵌套异常、指数延迟验证、抖动不越界。

关键文件：

- python/sglang/srt/disaggregation/common/conn.py（模块 连接层；类别 source；类型 core-logic；符号 register_to_bootstrap）：核心修改文件，实现重试逻辑

- test/registered/unit/disaggregation/test_register_to_bootstrap.py（模块 连接层；类别 test；类型 test-coverage；符号 TestRegisterToBootstrap, test_succeeds_on_first_attempt, test_succeeds_after_retries, test_all_retries_exhausted）：新增单元测试文件，覆盖所有重试和异常场景

关键符号：register_to_bootstrap

关键源码片段

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

核心修改文件，实现重试逻辑

```
def register_to_bootstrap(self):
    """Register prefill server info to bootstrap server via HTTP PUT."""
    if self.dist_init_addr:
        host = NetworkAddress.parse(self.dist_init_addr).resolved().host
    else:
        host = self.bootstrap_host

    bootstrap_na = NetworkAddress(host, self.bootstrap_port)
    url = f"{bootstrap_na.to_url()}/route"
    payload = {
        "attn_tp_size": self.attn_tp_size,
        # ... 其他 payload 字段省略 ...
        "load_balance_method": self.server_args.load_balance_method,
    }

    max_retries, initial_delay, max_delay = 5, 1.0, 30.0 # 最多重试 5 次，退避基准 1s，上限 30s
    for attempt in range(max_retries):
        try:
            response = requests.put(url, json=payload, timeout=5)
            if response.status_code == 200:
                logger.debug("Prefill successfully registered to bootstrap server.")
                return # 成功则立即返回
            logger.warning(
                f"Prefill register attempt {attempt + 1}/{max_retries} failed: status {response.status_
                code}"
            )
        except Exception as e:
            # 遍历 __cause__ 链以暴露根本原因（如 Connection refused），避免 urllib3 包装误导
            cause = e
            while cause.__cause__ is not None:
                cause = cause.__cause__
            logger.warning(
                f"Prefill register attempt {attempt + 1}/{max_retries} failed: {cause}"
            )
        if attempt == max_retries - 1:
            break # 最后一次尝试后不 sleep
        delay = min(initial_delay * (2**attempt), max_delay) * (
```

```
            0.75 + 0.25 * (time.monotonic() % 1) # 抖动系数 0.75~1.0，保证最大延迟不超过 max_
            delay
        )
        time.sleep(delay)
    logger.error(
        f"Prefill instance failed to register to bootstrap server after {max_retries} retries"
    )
```

评论区精华

审阅者 ShangmingCai 在 conn.py 第 393 行评论 "Should we crash the server here?"，质疑在重试全部耗尽后是否应该终止进程。该评论未收到回复，PR 随后获得批准合并，表明当前行为（仅记录 error 并继续运行）被接受，但设计选择未明确讨论。

- Should the server crash after all retries exhausted? (design): 未收到回复，PR 被批准合并，当前行为是记录 error 并继续运行。

风险与影响

- 风险：
  1. 启动延迟增加：最多可能增加约 31 秒（1+2+4+8+16 秒 + 抖动）的启动时间，但对于小模型可接受。
  2. 最终失败无防护：若 bootstrap 服务器永久不可达，prefill 会启动但未注册，后续请求必然失败，但日志已告知管理员。
  3. 无稳态影响：重试仅发生在启动阶段，不影响推理路径。
  4. 日志变更：失败日志从 error 降级为 warning，可能被监控忽略，但最终 error 仍保留。 - 影响：影响范围：仅 PD 分离模式下 prefill 实例的启动阶段；大模型（如 Qwen3-8B）因初始化时间长，通常不会触发重试，行为不变。影响程度：修复了小模型在 PD 分离下完全不可用的问题，提升了系统可靠性。对团队：测试覆盖完善，回归风险低。对用户：小模型 PD 分离部署不再 500。 - 风险标记：启动延迟增加，最终失败未终止进程

关联脉络

- 暂无明显关联 PR