

# PR #30053 完整报告

sgl-project/sglang

[BugFix] Release HiCache prefetch resources on disagg-prefill bootstrap-queue abort

合并时间: 2026-07-06 00:07

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30053>

## 执行摘要

- 一句话: 修复 HiCache prefetch 资源在 bootstrap 队列中止时泄漏
- 推荐动作: 值得合并: 这是一个典型的资源泄漏 bugfix, 逻辑清晰、风险低。推荐精读 `abort_request()` 方法以理解 PD 分离模式下不同队列的中止路径。

## 功能与动机

当 `abort_request()` 处理仍处于 `disagg_prefill_bootstrap_queue` 的请求时, 中止了 KV sender 但未调用 `release_aborted_request()`。由于 `_add_request_to_queue()` 在将请求加入 bootstrap 队列之前调用了 `_prefetch_kvcache()` (这会递增 `prefetch_tokens_occupied` 并创建 `ongoing_prefetch` 条目), 每个被中止的 bootstrap 请求都会永久泄漏 `len(prefetch_key)` 个 token。一旦计数器超过 `prefetch_capacity_limit`, `prefetch_rate_limited()` 会为后续所有请求返回 `True`, storage prefetch 将永久关闭。该问题由 v0.5.12 的 PR #23631 (commit 233048212a) 引入。

## 实现拆解

1. 定位问题: 在 `python/sglang/srt/managers/scheduler.py` 的 `abort_request()` 方法中, 处理 `disagg_prefill_bootstrap_queue` 的逻辑缺少 `release_aborted_request()` 调用。
2. 添加修复代码: 在 abort 循环内部, 中止 KV sender 之前, 先检查 `self.enable_hicache_storage` 是否为 `True`, 若是则调用 `self.tree_cache.release_aborted_request(req.rid)` 来释放 prefetch 占用的资源。
3. 保持风格一致: 该调用模式完全参考了 `waiting-queue abort` 路径 (line 3857 附近) 的现有写法, 确保修复与既有代码风格统一。
4. 无配置 / 测试改动: 本次是 3 行添加的纯源码修复, 未修改配置、测试或部署配套。

关键文件:

- `python/sglang/srt/managers/scheduler.py` (模块 调度器; 类别 source; 类型 core-logic)  
: 核心调度器文件, 实现 `abort_request()` 方法。本次在此文件中添加 3 行代码, 修复 bootstrap 队列中止时未释放 HiCache prefetch 资源的遗漏。

关键符号: `abort_request`

## 关键源码片段

## python/sglang/srt/managers/scheduler.py

核心调度器文件，实现 abort\_request() 方法。本次在此文件中添加 3 行代码，修复 bootstrap 队列中止时未释放 HiCache prefetch 资源的遗漏。

```
# python/sglang/srt/managers/scheduler.py
# 以下为 abort_request() 方法中处理 disagg_prefill_bootstrap_queue 的关键分支
# 新增代码在启用 HiCache 存储时释放 prefetch 资源，防止计数器永久泄漏

if self.disaggregation_mode == DisaggregationMode.PREFILL:
    # Abort requests that have not yet been bootstrapped
    for req in self.disagg_prefill_bootstrap_queue.queue:
        if recv_req.abort_all or req.rid.startswith(recv_req.rid):
            logger.debug(f"Abort bootstrap queue request. {req.rid}")
            # [ 新增 ] 释放 HiCache prefetch 占用的 token 计数器
            # 与 waiting-queue 分支 (line ~3857) 的 release_aborted_request() 调用保持一致
            if self.enable_hicache_storage:
                self.tree_cache.release_aborted_request(req.rid)

            # 原有的 KV sender 中止逻辑保持不变
            if hasattr(req.disagg_kv_sender, "abort"):
                req.disagg_kv_sender.abort()
```

## 评论区精华

无 review 讨论发生（0 条 review 评论）。仅有一条机器人评论提示 quota 已用完。审核者 ispobock 直接批准，未留下评论。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低：
  - 修复逻辑完全参照已有的 waiting-queue 路径，经过生产验证。
  - 新增代码在 self.enable\_hicache\_storage 开关保护下执行，不影响未启用 HiCache 的场景。
  - 仅 3 行添加，无删除或修改，回归可能性小。
  - 但缺少针对此场景的单元测试，未来重构可能被遗漏。
- 影响：影响范围窄：仅影响启用 PD 分离模式且使用 HiCache storage 的部署。修复前，一旦有 bootstrap 队列中的请求被中止（例如客户端取消或超时），storage prefetch 功能会永久失效，直到进程重启。修复后，资源正确释放，prefetch 功能持续可用。
- 风险标记：缺少测试覆盖

## 关联脉络

- PR #23631 引入了该 bug 的原始 PR: 本次修复的问题由 PR #23631 (commit 233048212a) 在 v0.5.12 中引入, 该 PR 添加了 bootstrap-queue abort 路径但遗漏了 `release_aborted_request()` 调用。
- PR #26886 与 HiCache prefetch 相关: PR body 中提到的关联 issue, 涉及 HiCache prefetch 功能。
- PR #27619 与 HiCache prefetch 相关: PR body 中提到的关联 issue, 涉及 HiCache prefetch 功能。