

# PR #38935 完整报告

sgl-project/sglang

[PD] Do not admit intake-rejected requests to a PD handoff

合并时间: 2026-09-16 12:44

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/38935>

## 执行摘要

- 一句话: PD 准入门口直接退休被拒请求, decode 超时降为即时 400
- 推荐动作: 值得精读。核心设计决策包括: a) 以「资源所有权」判别 intake 拒绝与请求重入——prefill 门口嗅探资源、decode 门口信任调用方标志的差异化策略; b) 退休收尾集中到 Scheduler.retire\_unadmitted\_request 单一入口, 与 #36988 的 post-forward 路径共享同一所有权测试从而保证互斥; c) 测试用 spy 包装真实退休方法, 让门口与退休两半代码相遇; d) 明确拒绝「统一 \_check\_if\_req\_exceed\_kv\_capacity 与退休逻辑」这类表面合理的合并。该 PR 也展示了高价值自审 review 的写法: 先验证两个 load-bearing 论断, 再列出「明确不是问题」的排除项, 避免后人重复争论。

## 功能与动机

PR body 明确指出这是「a waste-and-latency fix, not a correctness fix」:

`Req.set_finish_with_abort()` 记录拒绝在 `to_finish` 而非 `finished_reason`, `finished()` 读后者, 因此请求仍报告为 live, 所有准入门放行。关联 issue #36988 已添加 post-forward 退休路径, 客户端最终能拿到 400——但在此之前请求已完成 bootstrap 握手、预留 metadata buffer、初始化 RDMA sender、在 stub 上跑 forward pass, decode worker 还把它放入 prealloc 队列。目标是把退休提前到「请求尚未取得任何资源」的准入门口, 消除浪费与延迟。

## 实现拆解

1. 新增 `is_unadmitted_reject` 谓词 (python/sglang/srt/disaggregation/utils.py): 判定「intake 阶段被拒且未持有任何资源」。表达式为「`is_aborted(req)` 且不持有 {`req.is_retracted` 属性、KV row (`kv.holds_kv`)、mamba state (`kv.holds_mamba`)、metadata buffer (`metadata_buffer_index >= 0`)、host retraction backup (`kv.retraction_backup`) }」。资源标记区分「新被拒请求」与「抢占 / 恢复重入请求」; `seqlen <= 1` 抢占例外由 `req.is_retracted` 属性兜底——`release_req` 总是调用 `reset_for_retract` 设置它, 且其三个清除点 (`decode.py:862`、`decode_schedule_batch_mixin.py:73`、`schedule_batch.py:2746`) 都位于两个门口之后。
2. 两个 PD 准入门口加守卫: `PrefillBootstrapQueue.add` (python/sglang/srt/disaggregation/prefill.py) 在 `create_sender` 之前拦截, 跳过握手、metadata buffer 分配与 RDMA sender 初始化; `DecodePreallocQueue.add` (python/sglang/srt/disaggregation/decode.py) 不能嗅探资源——`retracted` 请求此时已释放全部标记——因此改用调用方声明的 `is_retracted` / `is_rebootstrap` 标志门控, 守卫条件

为「非 retracted 且非 rebootstrap 且 is\_unadmitted\_reject」，保持 retracted\_queue 的释放契约。

3. 新增 `Scheduler.retire_unadmitted_request` (python/sglang/srt/managers/scheduler.py)：一次完成全部收尾——streaming session 的 pre-abort detach (否则 session 永久 in-flight、后续请求全部失败)、beam\_coordinator.retire\_group(req) (validate\_and\_init 提前计数, 无 group 时 no-op)、\_release\_aborted\_request(req) (释放缓存侧 paced-retry 状态; review 发现首版缺失此行会造成 \_storage\_prefetch\_missed\_rids 进程级泄漏)、以真实原因关闭 trace span、update\_finish\_state() 与 stream\_output 让客户端拿到保留 BAD\_REQUEST 状态码的 400。
4. `_add_request_to_queue` 保持纯 dispatcher: 它是 retracted / preempted / resumed 请求的重入路径, 这些请求可能已持有 KV、metadata buffer 或 host retraction backup, 在 dispatcher 层退休会丢资源。
5. 测试配套: 新增 test/registered/unit/disaggregation/test\_admission\_abort\_not\_enqueue.py (326 行, 注册进 base-a-test-cpu CPU CI 套件), 覆盖两门口拒绝即退休、prefill 重入 (metadata buffer / retraction backup 两种标记)、decode retracted / rebootstrap 重入 (含 backup=None 短序列)、无标志抢占重入不被误退休、有效请求仍准入、NULL 模式刻意不碰、退休副作用全链路; \_make\_scheduler 用 spy 包装真实 retire\_unadmitted\_request, 保证门口测试同时执行退休体。
6. 演进中的适配: commit 2 落实 review 的缓存侧释放与 is\_retracted 兜底; commit 3 适配 #39318 把 \_release\_aborted\_request 从接收 rid 改为接收 Req 的签名变化——初版仍传 req.rid (str), 会在调度器事件循环抛 AttributeError。

关键文件:

- python/sglang/srt/managers/scheduler.py (模块 调度器; 类别 source; 类型 core-logic; 符号 retire\_unadmitted\_request) : 新增 Scheduler.retire\_unadmitted\_request——门口退休的全部收尾集中于此: streaming session 的 pre-abort detach、beam group 退休、缓存侧 \_storage\_prefetch\_missed\_rids 释放、以真实原因关闭 trace span 并输出 400。review 发现首版缺失 \_release\_aborted\_request 会造成进程级 rid 泄漏。
- python/sglang/srt/disaggregation/utils.py (模块 准入判断; 类别 source; 类型 core-logic; 符号 is\_unadmitted\_reject) : 新增 is\_unadmitted\_reject 谓词, 定义「intake 被拒且未持有任何资源」的资源所有权判定; 区分新被拒请求与抢占 / 恢复重入请求, 是 prefill 门口守卫的判定核心。
- python/sglang/srt/disaggregation/prefill.py (模块 预填队列; 类别 source; 类型 core-logic; 符号 add) : PrefillBootstrapQueue.add 增加门口守卫: intake 被拒请求直接退休, 跳过 create\_sender——bootstrap 握手、metadata buffer 预留与 RDMA sender 初始化全部不再发生。
- python/sglang/srt/disaggregation/decode.py (模块 解码队列; 类别 source; 类型 core-logic; 符号 add) : DecodePreallocQueue.add 以调用方声明的 is\_retracted / is\_rebootstrap 标志门控 (而非资源嗅探) : 重入的 retracted 请求此时已不持有任何 is\_unadmitted\_reject 读取的标记, 嗅探会误伤; 同时保持 retracted\_queue 的释放契约。

- test/registered/unit/disaggregation/test\_admission\_abort\_not\_enqueued.py (模块 准入测试; 类别 test; 类型 test-coverage; 符号 \_make\_req, \_make\_scheduler, \_prefill\_queue, \_decode\_queue) : 326 行回归测试覆盖两门口拒绝退休、三类重入不被误退休 (prefill 资源标记、decode 标志、无标志抢占)、NULL 模式不碰与退休副作用全链路; 用 spy 包装真实 retire\_unadmitted\_request, 确保门口测试同时执行退休体, 避免两半代码永远不相遇。

关键符号: retire\_unadmitted\_request, is\_unadmitted\_reject, PrefillBootstrapQueue.add, DecodePreallocQueue.add

## 关键源码片段

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

新增 `Scheduler.retire_unadmitted_request`——门口退休的全部收尾集中于此: streaming session 的 pre-abort detach、beam group 退休、缓存侧 `_storage_prefetch_missed_rids` 释放、以真实原因关闭 trace span 并输出 400。review 发现首版缺失 `_release_aborted_request` 会造成进程级 rid 泄漏。

```
def retire_unadmitted_request(self, req: Req) -> None:
    """Finish a request the disaggregation queues rejected at their door."""
    # `create_req` 会把 streaming session 标记为 in-flight, 而 pre-abort
    # detach 只在 `StreamingSession.find_active_slot` 中执行 (调度路径内);
    # 被退休的请求到不了那里, 不分离则 session 永久 in-flight,
    # 该 session 上的后续请求会全部报 "already has an active request".
    if req.session is not None and req.session.streaming:
        req.session.abort_req()
        req.session = None
    # `beam_coordinator.validate_and_init` 在拒绝检查之前就把 group 计入,
    # 这里对应退休; 请求没有 group 时是 no-op。
    self.beam_coordinator.retire_group(req)
    # PREFILL 门口之前已运行 `_prefetch_kvcache`, 单 token stub 也会被注册
    # 进缓存侧 paced-retry 集合 (`_storage_prefetch_missed_rids`), 只有
    # `finish`/ABORT 与 `waiting_queue` 扫描能清除; 缺失此行会进程级泄漏。
    self._release_aborted_request(req)
    # `update_finish_state` 在 `finished()` 为真时提前返回, 客户端实际收到
    # 的是已设置的 `finished_reason` (400); 这里以真实原因关闭 trace span,
    # 避免 span 在 GC 时以无状态方式被自动关闭。
    reason = req.finished_reason or req.to_finish
    req.time_stats.trace_ctx.abort(abort_info={"reason": reason.message})
    req.update_finish_state()
    self.output_streamer.stream_output([req], req.return_logprob)
```

### python/sclang/srt/disaggregation/prefill.py

`PrefillBootstrapQueue.add` 增加门口守卫: intake 被拒请求直接退休, 跳过 `create_sender`——bootstrap 握手、metadata buffer 预留与 RDMA sender 初始化全部不再发生。

```
def add(self, req: Req, num_kv_heads: int) -> None:
```

```

# intake 阶段被拒的请求: `set_finish_with_abort()` 把判决留在 `to_finish`
# (`finished()` 不读它) 并把 prompt 换成单 token stub。若放行, 一次
# bootstrap 会消耗握手、metadata buffer、RDMA sender 初始化与一次
# stub forward pass 之后才会被 #36988 的 post-forward 路径回卷。
if is_unadmitted_reject(req):
    self.scheduler.retire_unadmitted_request(req)
    return
if not self.create_sender(req, num_kv_heads):
    return
self.queue.append(req)

```

## 评论区精华

审核全部由作者本人以「Agent review」形式完成, 质量较高。阻塞项是 `retire_unadmitted_request` 缺失缓存侧释放: 「Every sibling retire path releases the request's cache-side state: `_retire_aborted_prefill_result` (prefill.py:1120), `handle_bootstrap_failure` (prefill.py:1152), abort method 1 (scheduler.py:5222). This one doesn't.」 PREFILL 下 `_prefetch_kvcache` 已在门口之前把 rid 放入 `_storage_prefetch_missed_rids` (单 token stub 必走 too-short 分支, 且该分支在 `prefetch_length > 0` 守卫之外), 该集合只有 `release_aborted_request` 与 `waiting_queue` 扫描能清, 门口退休的请求两者都到不了——commit 2 修复。另一个关键修正是 `req.is_retracted` 属性: 「The PR description rules `is_retracted` out, but that's the parameter passed to `add()`. The attribute on the request is set by `reset_for_retract()`, which `release_req` always calls」——priority preemption 与 pause/retract-all 走裸 `_add_request_to_queue` 不带任何标志, 只有该属性能兜住 `seqlen <= 1` 抢占的「资源标记全空」形状, 否则 `retire_group` 的「member rows were released by the caller」契约被破坏。作者还静态验证了两条承重论断: prefill 嗅探与 `_retire_aborted_prefill_result` (prefill.py:1118) 是同一所有权测试, 两退休路径天然互斥; `metadata_buffer_index` 只由 `ensure_metadata_buffer` 写入、`maybe_release_metadata_buffer` 清除, 重入论证成立。同时明确排除两个误判: `trace_ctx.abort()` 后置的 `set_prefill_bootstrap_queue_entry_time()` 无害 (trace.py:819 置 `thread_context = None` 后 span 短路); 「统一 `_check_if_req_exceed_kv_capacity` 与退休逻辑」是错误方向——该路径也会命中重入请求, session detach 与 `retire_group` 会造成本 PR 试图避免的伤害。

- `retire_unadmitted_request` 缺失缓存侧释放, rid 泄漏至进程结束 (correctness): 在 `retire_unadmitted_request` 中加入 `self._release_aborted_request(req)`; commit 2 「release cache state」落实。
- `req.is_retracted` 属性兜住无标志重入路径 (correctness): 在 `is_unadmitted_reject` 中加入 `req.is_retracted` 判断, 并新增 `test_preempted_reentry_is_not_retired` 用例。
- 门口测试与退休测试分离导致漏测, 用 spy 合体 (testing): `_make_scheduler` 采用 spy 包装 (`side_effect=lambda req: Scheduler.retire_unadmitted_request(sched, req)`), `SimpleNamespace` 补上 `_release_aborted_request = MagicMock()`。
- 两条 load-bearing 论断的静态验证与两个「明确不是问题」(design): 无代码变更; 排除两个误判, 划定「不合并 KV 容量检查与退休逻辑」的边界。

## 风险与影响

- 风险：依赖请求状态属性契约：is\_unadmitted\_reject 的正确性依赖 req.is\_retracted 由 reset\_for\_retract 设置、且清除点全部位于两个门口之后；未来若在门之前新增清除点，会误退休重入请求，破坏 retire\_group 契约。新调用方暴露的签名敏感：\_release\_aborted\_request 已被 #39318 改过一次签名 (rid → Req)，初版传 req.rid 会在调度器事件循环抛 AttributeError；同类接口变更需同步检查 retire\_unadmitted\_request。不对称拒绝场景未覆盖：仅 prefill 或仅 decode 单侧拒绝时，对端持有效请求仍等满传输超时 (500 @ 300 s)，PR 自述不在范围内，需配合 #38704 的 abort\_before\_send 才能快速失败。验证范围：无更广 GPU/ 传输套件验证 (checklist 明示未完成)，仅 1P1D 实测与 CPU 单元测试 (1011 passed / 4 skipped)。
- 影响：用户在 PD 模式下提交超长 prompt (如超过 max\_req\_input\_len) 的请求：decode 侧从「300 s 传输超时」变为「即时 400 (0.0 s)」，prefill 侧省去 bootstrap 握手、metadata buffer、RDMA sender 初始化与一次 stub forward pass。NULL 模式刻意不受影响：stub 单次 forward 成本低、batch 边界即返回 400，且在该门口加守卫会跳过 find\_active\_slot 的 pre-abort detach，反而引入回归。与 #38704 都编辑 DecodePreallocQueue.add，后落地者需 rebase 而非 take-theirs。
- 风险标记：PD 准入核心路径变更，依赖 is\_retracted 状态属性契约，不对称拒绝场景仍等满超时，与 #38704 同函数冲突，新增路径缺 GPU 级回归验证

## 关联脉络

- PR #36988 [VLM] Retire aborted disaggregated prefill results: 本 PR 的语义前身：为被 abort 的分离式 prefill 结果增加 post-forward 退休路径，让客户端最终能拿到 400。本 PR 把同一退休提前到准入门口，并复用其 \_retire\_aborted\_prefill\_result 的资源所有权测试，保证两条退休路径互斥、不会重复触发。
- PR #38704 Catch is\_aborted in pop\_bootstrapped with transport abort\_before\_send (题名取自 PR #38935 body 描述)：PR #38935 body 明确列为 related：攻击同一类浪费的下一层——在 pop\_bootstrapped 捕获 is\_aborted 并加 transport 级 abort\_before\_send，覆盖本 PR 明确排除的不对称拒绝场景。两者意图互补但都编辑 DecodePreallocQueue.add，后落地者需 rebase。
- PR #39318 Rescope cache prefetch state to the request attempt (题名取自 commit 8f3d69 描述)：把 \_release\_aborted\_request 从接收 rid 改为接收 Req、prefetch 状态按请求 attempt 作用域化。commit 3 明确说明 #39318 改签名后，retire\_unadmitted\_request 初版仍传 req.rid (str) 会在调度器事件循环抛 AttributeError，本 PR 为此适配为传 Req。