

PR #36150 完整报告

sgl-project/sglang

fix(mini-lb): forward the flush_cache timeout param to workers

合并时间: 2026-08-24 15:32

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36150>

执行摘要

- 一句话: 修复 mini-lb 转发 flush_cache 的 timeout 参数, 恢复延迟排空路径
- 推荐动作: PR body 值得精读: 对 " 参数丢弃 → scheduler 立即 flush → KV 释放 → Mooncake 单次失败永久拉黑 " 的跨层根因分析是高质量排障范例, 量化数据 (12 次真实失败放大为约 1465 个请求失败, 吞吐 2,039→9,994 tok/s) 很有说服力。作者主动更正 PR 归因的严谨做法也值得参考。代码本身仅 8 行改动、阅读成本极低; 管理上应跟进三个遗留项: #36152 的真实修复、Rust router 同缺陷、Mooncake 黑名单与恢复探针策略。

功能与动机

PR body 指出 mini_lb 的 /flush_cache handler 原本 "took no parameters and posted to a bare {server}/flush_cache, dropping any ?timeout= the caller sent"。而调度器侧 scheduler_components/flush_wrapper.py 将缺失或非正的 timeout 视为 "flush now, skip the idle check", 直接执行 _flush_cache(), 导致 SchedulerFlushWrapper 基于 is_fully_idle() 的延迟路径 (已纳入 disagg_prefill_inflight_queue、disagg_decode_transfer_queue 及 prealloc/retracted 队列) 完全不可达。实际后果是破坏性 KV-cache flush 落在在途 PD 传输上: MooncakeKVManager 读到刚释放的 buffer 而失败, 再经单次失败永久拉黑 peer session 的机制放大为 86.9% 空响应、吞吐从 9,994 tok/s 跌到 2,039 tok/s。

实现拆解

1. 变更入口: 仅 1 个文件, sgl-model-gateway/bindings/python/src/sglang_router/mini_lb.py, 改动 +8/-2。
2. 签名改造: /flush_cache handler 从无参的 flush_cache() 改为 flush_cache(timeout: Optional[float] = None), 使 FastAPI 能把 ?timeout=60 绑定为 60.0, 缺省时为 None。
3. 转发逻辑: 构造 params = None if timeout is None else {"timeout": timeout}; 遍历 chain(lb.prefill_urls, lb.decode_urls) 时把 params 传给 aiohttp ClientSession.post。关键设计是仅在调用方提供 timeout 时拼接 query string, 省略时 params 为 None、aiohttp 不产生查询串, 旧调用方行为完全不变。
4. 测试配套: 没有新增测试文件。mini_lb 无 in-tree 单测框架, 作者用行为验证 (FastAPI 参数绑定、aiohttp 查询串拼接、ruff/black/py_compile 检查) 取代, 并多次 /rerun-test 运行 test/registered/disaggregation/test_disaggregation_dp_attention.py, 最终在 8-gpu-h20 上通过。

关键文件：

- `sgl-model-gateway/bindings/python/src/sglang_router/mini_lb.py`（模块 负载均衡；类别 source；类型 data-contract；符号 `flush_cache`）：本 PR 唯一改动文件：`/flush_cache handler` 从无参改为接受可选 `timeout` 参数，并在调用方提供时通过 `aiohttp params` 转发给全部 `prefill/decode worker`，是修复的核心。

关键符号：`flush_cache`

关键源码片段

`sgl-model-gateway/bindings/python/src/sglang_router/mini_lb.py`

本 PR 唯一改动文件：`/flush_cache handler` 从无参改为接受可选 `timeout` 参数，并在调用方提供时通过 `aiohttp params` 转发给全部 `prefill/decode worker`，是修复的核心。

```
# sgl-model-gateway/bindings/python/src/sglang_router/mini_lb.py
# /flush_cache 在 FastAPI 应用中注册为 POST 端点。
```

```
@app.post("/flush_cache")
```

```
async def flush_cache(timeout: Optional[float] = None):
```

```
    # 关键语义：调度器的 SchedulerFlushWrapper 会把 timeout_s <= 0 解释为
    # "立即 flush、跳过 idle 检查"，因此这里绝不能丢弃调用方传入的 timeout，
    # 否则上游只会在 PD KV 传输进行中释放缓存，导致真实的传输失败并被
    # MooncakeKVManager 拉黑 peer session。
    # timeout 未提供时保持旧行为：params 为 None，aiohttp 不拼接 query string。
    params = None if timeout is None else {"timeout": timeout}
```

```
    async with aiohttp.ClientSession() as session:
```

```
        # 对 prefill 与 decode 全量 worker 做扇出，逐台转发 flush_cache 请求。
```

```
        tasks = []
```

```
        for server in chain(lb.prefill_urls, lb.decode_urls):
```

```
            tasks.append(session.post(f"{server}/flush_cache", params=params))
```

```
        for i, response in enumerate(asyncio.as_completed(tasks)):
```

```
            await response
```

```
    return Response(status_code=200)
```

评论区精华

没有 review 评论，讨论集中在 issue 评论区，核心是作者自己的归因更正：PR body 声称的 CI 失败其实不是本 PR 修复的——PD disaggregation 测试由 `PDDisaggregationServerBase.launch_lb` 传 `--mini-lb` 前置的是另一处 `mini_lb`，`experimental/sgl-router` 不在该路径上。作者明确说 "The defect here is real and independent"，真实修复在 #36152。此外作者多次发起 `/rerun-test`，github-actions 回报 8-gpu-h20 上 `dp_attention` 测试通过。

- PR 归因更正：本修复并非 `dp_attention` CI 失败的真因 (question)：本 PR 修复独立缺陷，不改变 `dp_attention` 失败归因；读者应把 #36152 视为该 CI 失败的实际修复。

- 测试覆盖与验证方式：mini_lb 无 in-tree 测试框架 (testing)：本次以行为验证代替单测，未新增测试文件；后续可考虑为 mini_lb 建立测试框架。

风险与影响

- 风险：
 1. 兼容性风险低：timeout 缺省时 params 为 None，行为与旧版一致。
 2. 测试缺失：mini_lb 无 in-tree 测试框架，修复依赖手工行为验证，同类缺陷后续可能再次放行。
 3. Rust router (experimental/sgl-router/src/server/routes/cache.rs) 存在同一个缺陷——无 query extractor、裸 worker URL，PR 明确留给后续单独修复，因此被该 router 前置的部署仍会命中同样的 race。
 4. 调度器 timeout_s <= 0 的立即 flush fast path 仍是 footgun：任何其他省略 timeout 的调用方依然会触发立即 flush，语义设计未改变。
 5. MooncakeKVManager 单次 ret != 0 永久拉黑 peer session、恢复探针默认关闭（30s 间隔）的放大机制未在本 PR 处理，偶发失败仍可能被放大为大规模请求失败。- 影响：用户侧：经该 router 的 Python binding (sglang_router/mini_lb.py) 前置的 PD 部署，flush_cache 的 timeout 语义得到修复，SchedulerFlushWrapper 的延迟排空路径 (is_fully_idle 门控，含 disagg_prefill_inflight_queue、disagg_decode_transfer_queue、prealloc/retracted 队列) 变为可达，避免 KV flush 打断在途传输。系统侧：影响面限于 admin 端点的一次查询参数透传，不涉及模型输出、kernel 或 forward 路径，属于低风险小改动。团队侧：作者在 PR body 和评论区厘清了从参数丢弃到 Mooncake 黑名单放大的完整缺陷链，为后续修复 Rust router 同缺陷和黑名单恢复策略提供了分析基础。- 风险标记：无自动化测试覆盖，Rust router 同缺陷未修，调度器立即 flush fast path 仍是陷阱，Mooncake 黑名单放大未处理

关联脉络

- PR #36152 评论中指出的真正修复（标题未提供）：作者在评论中明确指出的实际 CI 失败修复，针对 PD 测试路径中的另一处 mini_lb；本 PR 与其处于同一缺陷面，但目标是 model-gateway 侧独立副本，二者互补而不是替代关系。