PR #25641 完整报告

sgl-project/sglang

Fix flush_cache AttributeError on is_stats_logging_rank

合并时间：2026-05-18 19:23

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25641

执行摘要

- 一句话：修复 flush_cache 因 is_stats_logging_rank 报错
- 推荐动作：快速审核并合并。此 PR 修复了因之前重构引起的关键路径回归，对保证 CI 正常和服务器稳定性至关重要。

功能与动机

PR #25630 重构了 metrics 相关的 Mixin，将 is_stats_logging_rank 属性移动到 SchedulerMetricsReporter 中，但 Scheduler.flush_cache 方法中的对应引用未被更新，导致调用 /flush_cache 时抛出 AttributeError，进而引发 SIGQUIT 终止服务器进程。多个 CI 作业（如 base-b-test-1-gpu-large/small、extra-a-test-1-gpu-large 等）均因此失败。PR body 中提供了复现 traceback 和 CI 失败链接。

实现拆解

在 python/sglang/srt/managers/scheduler.py 文件中的 flush_cache 方法内，将第 3323 行的 self.is_stats_logging_rank 改为 self.metrics_reporter.is_stats_logging_rank，使属性访问路由到新的 SchedulerMetricsReporter 实例。改动仅一行，与同一方法中已有的 self.metrics_reporter.reset_metrics() 访问模式一致。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 flush_cache）：核心调度器文件；修复了 flush_cache 方法中对 is_stats_logging_rank 属性的错误引用

关键符号：flush_cache

关键源码片段

python/sglang/srt/managers/scheduler.py

核心调度器文件；修复了 flush_cache 方法中对 is_stats_logging_rank 属性的错误引用

```
def flush_cache(self, empty_cache: bool = True):
    """Flush memory pools (e.g., KV cache, Mamba cache) and optionally empty device allocator
    cache."""
    if self.is_fully_idle():
        self.cur_batch = None
```

```
            self.last_batch = None
            self.tree_cache.reset()
            self.req_to_token_pool.clear()
            self.token_to_kv_pool_allocator.clear()
            self.grammar_manager.clear()
            self.metrics_reporter.reset_metrics()

            if self.draft_worker:
                self.draft_worker.clear_cache_pool()

            if empty_cache:
                current_platform.empty_cache()
            # 修复：is_stats_logging_rank 已在 PR #25630 中从 Scheduler 属性迁移到了
            # SchedulerMetricsReporter，因此需要通过 self.metrics_reporter 访问。
            if self.metrics_reporter.is_stats_logging_rank:
                logger.info("Cache flushed successfully!")
            success = True
        else:
            logging.warning(
                f"Cache not flushed because there are pending requests. "
                f"#queue-req: {len(self.waiting_queue)}, "
                f"#running-req: {len(self.running_batch.reqs)}"
            )
            success = False
        return success
```

评论区精华

该 PR 无 review 评论，仅有一条作者自己的 /tag-and-rerun-ci 命令用于触发 CI。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅一行变更，且与同一方法中已存在的 self.metrics_reporter.reset_metrics() 访问模式完全一致。不会引入回归。
- 影响：影响范围限于修复 flush_cache 的 AttributeError，对使用 /flush_cache 接口的客户端（如 bench_serving 脚本）有直接影响。修复后 CI 恢复正常，服务器不再因该错误崩溃。
- 风险标记：暂无

关联脉络

- PR #25630 Move metrics reporting to SchedulerMetricsReporter and retire metrics mixin: 本 PR 修复了该重构引入的遗漏回归：is_stats_logging_rank 属性被迁移到 SchedulerMetricsReporter，但 flush_cache 中的引用未更新。