PR #22559 完整报告

sgl-project/sglang

[metrics] Add `PoolStats.update_scheduler_stats` to deduplicate metrics assignment

合并时间：2026-04-11 12:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22559

执行摘要

- 一句话：重构调度器指标统计逻辑，统一池相关字段赋值并修复四舍五入一致性。
- 推荐动作：该 PR 值得精读，特别是对于关注代码质量和可维护性的工程师。关注点包括：1. update_scheduler_stats 方法如何统一处理不同池统计字段。2. round(..., 2) 的引入如何修复四舍五入不一致问题。3. 如何通过单一方法调用替换多个重复块，这是典型的 DRY 原则应用。

功能与动机

PR body 中明确指出，本次变更旨在消除三个重复的赋值块（prefill stats、decode stats 和 idle check_memory），通过统一的 pool_stats.update_scheduler_stats(self.stats) 调用来集中池相关统计字段的赋值。同时，修复了 token_usage 字段的四舍五入不一致问题，此前仅 idle 路径进行了四舍五入，现在所有三个路径都保持一致。这是对 #22554 的后续改进。

实现拆解

实现分为两个关键文件：1. 在 python/sglang/srt/managers/scheduler_runtime_checker_mixin.py 中，为 PoolStats 类新增 update_scheduler_stats 方法，该方法接收 SchedulerStats 对象，统一设置 num_used_tokens、token_usage、full_token_usage、swa_token_usage 和 mamba_usage 字段，其中 token_usage 使用 round(..., 2) 进行四舍五入。2. 在 python/sglang/srt/observability/scheduler_metrics_mixin.py 中，移除 report_prefill_stats、report_decode_stats 和 check_memory 方法中重复的字段赋值代码，替换为对 pool_stats.update_scheduler_stats(self.stats) 的调用。

关键文件：

- python/sglang/srt/managers/scheduler_runtime_checker_mixin.py（模块 scheduler）：新增 update_scheduler_stats 方法，统一池统计字段赋值逻辑，是本次重构的核心。
- python/sglang/srt/observability/scheduler_metrics_mixin.py（模块 observability）：移除三个重复的赋值块，替换为对 update_scheduler_stats 的调用，实现了代码去重。

关键符号：update_scheduler_stats, report_prefill_stats, report_decode_stats, check_memory

评论区精华

由于 review_comments_count 为 0，没有 review 讨论记录。从 commit 历史看，有三个提交：第一个添加了 update_scheduler_stats 方法并去重；第二个添加了类型注解；第三个修复了 token_usage 的四舍五入一致性。这表明作者在实现过程中自行完善了代码，没有外部 review 反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低：1. 核心变更集中在指标统计逻辑，不涉及核心推理路径，因此回归风险有限。2. 新增的 round(..., 2) 可能引入微小的精度差异，但 PR body 指出这是为了修复不一致性，因此是预期行为。3. 由于去除了重复代码，如果 update_scheduler_stats 方法存在逻辑错误，将同时影响 prefill、decode 和 idle 三个路径，但测试计划覆盖了这些路径。4. 文件 python/sglang/srt/managers/scheduler_runtime_checker_mixin.py 和 python/sglang/srt/observability/scheduler_metrics_mixin.py 都属于调度器模块，变更范围可控。
- 影响：影响范围：1. 对用户：无直接影响，因为这是内部指标统计重构。2. 对系统：提升了代码可维护性，减少了重复逻辑，可能降低未来 bug 引入风险。3. 对团队：统一了指标赋值逻辑，便于后续开发和调试。影响程度：中等，因为涉及调度器核心统计逻辑，但属于重构而非功能变更。
- 风险标记：核心统计逻辑变更，四舍五入一致性修复

关联脉络

- PR #22554 [mem] Fix idle token_usage missing mamba_usage; add FIXME for naming: PR body 提到本次变更是对 #22554 的后续改进，两者都涉及调度器指标统计逻辑的修复和优化。
- PR #22555 [mem] Fix idle token_usage missing mamba_usage; add FIXME for naming: 同 #22554，但根据提供的近期历史 PR 分析，这是同一个 PR，可能为重复引用。