PR #25617 完整报告

sgl-project/sglang

Move on_idle from runtime_checker mixin into Scheduler

合并时间：2026-05-18 18:34

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25617

执行摘要

- 一句话：将 on_idle 方法从 Mixin 移至 Scheduler 主类
- 推荐动作：值得精读，作为理解调度器组件化重构路线的关键步骤。

功能与动机

PR body 指出，迁移依据是 components/scheduler/index.md 中对该方法归属的显式人类决策，旨在将空闲管家编排逻辑归入 Scheduler 主类，保持职责清晰。

实现拆解

1. 从 scheduler_runtime_checker_mixin.py 中删除 on_idle 方法定义（-31 行）。
2. 在 scheduler.py 中、is_fully_idle 方法之前插入 on_idle 方法（+31 行），方法体与之前字节一致，仅去掉了 self: Scheduler 参数注解（因 now redundant）。
3. 指标刷新辅助 _maybe_log_idle_metrics 未迁移，仍通过 Mixin 继承解析，后续将由下游提交直接移至 SchedulerMetricsReporter。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 on_idle）：接收 on_idle 方法，将其置于 is_fully_idle 之前，保持空闲相关代码连续。
- python/sglang/srt/managers/scheduler_runtime_checker_mixin.py（模块 调度器；类别 source；类型 core-logic；符号 on_idle）：移除 on_idle 方法定义（31 行），去除了冗余代码。

关键符号：on_idle

关键源码片段

python/sglang/srt/managers/scheduler.py

接收 on_idle 方法，将其置于 is_fully_idle 之前，保持空闲相关代码连续。

```
def on_idle(self):
    """Idle housekeeping orchestrator:  guard, check, metrics, reset, sleep."""
    if not self.is_fully_idle():
        return

    # 内存泄漏检查（hisparse 启用时跳过——host-backup 期间 pool 计数器
```

```
    # 会有意偏差，详见 _get_swa_token_info clamp）。
    if not self.enable_hisparse:
        has_leak, messages = self._check_all_pools(self.get_pool_stats())
        if has_leak:
            self._report_leak("pool", "\n".join(messages))
        self._check_req_pool()

    # tree cache 健全性检查
    self._check_tree_cache()

    # 每 30 秒记录一次指标
    self._maybe_log_idle_metrics()

    # 发布 KV 事件
    self._publish_kv_events()

    # 重置 token 比率
    self.new_token_ratio = self.init_new_token_ratio

    # 重置设备计时窗口，避免空闲时间被计入
    self.reset_device_timer_window()

    # 若无新事件则休眠
    self.maybe_sleep_on_idle()
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：由于是纯方法迁移且逻辑字节一致，回归风险极低。但需注意 on_idle 内部调用了 _maybe_log_idle_metrics 等仍在 Mixin 中的方法，在 Mixin 完全退役前需保持继承链。
- 影响：影响范围仅限于调度器内部重构，外部用户无感知。后续组件化重构（pool-stats-observer、invariant-checker 等）的基础步骤。
- 风险标记：核心路径变更，无测试覆盖

关联脉络

- PR #25621 Move pool-stats sampling to SchedulerPoolStatsObserver: 同一重构链中的下游提交，负责将 get_pool_stats 调用迁移至新组件。
- PR #25624 Move invariant checks to SchedulerInvariantChecker and retire runtime_checker mixin: 同一重构链中的下游提交，负责将 _check_all_pools 等检查逻辑迁移至新组件。

- PR #25631 Move idle-metrics logging to SchedulerMetricsReporter: 同一重构链中的下游提交，负责将 _maybe_log_idle_metrics 迁移至新组件。