PR #25718 完整报告

sgl-project/sglang

Stop returning the unused prefix-computed flag from priority calc

合并时间：2026-05-19 09:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25718

执行摘要

- 一句话：移除 calc_priority 无用的 bool 返回值
- 推荐动作：建议合入。该 PR 是简单的死代码清理，没有风险。对于关注代码整洁度和模块精简的工程师，可以借此回顾调度优先级计算模块的整体设计。

功能与动机

所有生产调用方（scheduler、dllm mixin、disagg decode）和唯一测试调用方都丢弃了 calc_priority 的布尔返回值，因此 prefix_computed 局部变量及相关 bookkeeping 是死代码。删除它们以减少维护负担和代码复杂度。

实现拆解

1. 修改方法签名：将 calc_priority 的返回类型注解从 -> bool 改为 -> None。
2. 移除局部变量：删除 prefix_computed = False 声明及其赋值（在 CacheAwarePolicy 分支中设为 True）。
3. 移除 return 语句：删除最后的 return prefix_computed，并在 FCFS 分支中将 return False 改为不带值的 return。
4. 调整分支结构：由于不再需要维护 prefix_computed，isinstance 分支中的 prefix_computed = True 被删除，分支结构保持原样。

关键文件：

- python/sglang/srt/managers/schedule_policy.py（模块 调度器；类别 source；类型 core-logic；符号 calc_priority）：唯一变更文件，核心调度策略类 SchedulePolicy 的方法 calc_priority 移除了无用返回值。

关键符号：calc_priority

关键源码片段

python/sglang/srt/managers/schedule_policy.py

唯一变更文件，核心调度策略类 SchedulePolicy 的方法 calc_priority 移除了无用返回值。

```
def calc_priority(
    self, waiting_queue: List[Req], running_batch: Optional[ScheduleBatch] = None
) -> None:
```

```
    # 原先返回 bool 但所有调用方均未使用，故改为 None
    if self.policy == CacheAgnosticPolicy.FCFS:
        if self.enable_priority_scheduling:
            SchedulePolicy._sort_by_priority_and_fcfs(
                waiting_queue, self.priority_sign
            )
        return # 原为 return False，现改为无值 return

    policy = self._determine_active_policy(waiting_queue)

    # 移除了 prefix_computed 局部变量及其赋值
    if isinstance(policy, CacheAwarePolicy):
        # prefix_computed = True 已删除
        temporary_deprioritized = self._compute_prefix_matches(
            waiting_queue, policy
        )
        if policy == CacheAwarePolicy.LPM:
            SchedulePolicy._sort_by_longest_prefix(
                waiting_queue, temporary_deprioritized
            )
        elif policy == CacheAwarePolicy.DFS_WEIGHT:
            SchedulePolicy._sort_by_dfs_weight(waiting_queue, self.tree_cache)
        else:
            raise ValueError(f"Unknown CacheAware Policy: {policy=}")
    else:
        # ... CacheAgnosticPolicy 各分支省略，无变动
        pass
    # 原 return prefix_computed 已删除
```

评论区精华

无 review 讨论。该 PR 只包含一个 commit，无 review 评论，变更简单直接，属于机械性死代码删除。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。该变更仅删除未使用的返回值和相关局部变量，不改变任何逻辑行为。所有调用方（scheduler、dllm mixin、disagg decode 及测试）均忽略原返回值，因此去除后不影响运行时行为。无回归或兼容性风险。
- 影响：影响范围极小，仅涉及 python/sglang/srt/managers/schedule_policy.py 这一文件，变更量为 +2/-5。对用户无感知，对系统功能无影响，对团队维护有利（减少死代码）。
- 风险标记：低风险变更

关联脉络

- PR #25719 Confine max-prefix-len to where it is used and drop the leftover variable: 同样是针对调度优先级计算模块的死代码清理，移除无用的 max_prefix_len 变量赋值，与本次 PR 类似地减少无效状态。
- PR #25728 Pull the max-prefix-len computation into its own helper and rename the matched-token argument: 对同一模块（调度优先级计算）进行重构，提取辅助方法，属于同一系列清理演进。