PR #22555 完整报告

sgl-project/sglang

[mem] Fix idle token_usage missing mamba_usage; add FIXME for naming

合并时间：2026-04-11 07:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22555

执行摘要

该 PR 修复了调度器空闲路径在混合 SSM（Mamba）场景下 token_usage 统计缺失 Mamba 使用量的 bug，通过统一使用 max(full_token_usage, mamba_usage) 计算确保与预填充 / 解码路径一致。同时添加 FIXME 注释澄清字段命名问题，为未来 API 重构铺垫。变更影响限于内部统计准确性，风险较低。

功能与动机

根据 PR body，修复动机源于历史不一致性：空闲路径在 #12014 引入，而预填充 / 解码路径在 #17862 重构后已正确包含 mamba_usage，但空闲路径未同步更新。这导致在 is_hybrid_ssm=True（混合 SSM 场景）时，self.stats.token_usage 仅使用 full_token_usage，忽略了 mamba_usage，使得内存使用统计不准确。修复目标是确保所有路径（空闲、预填充、解码）的统计逻辑一致。

实现拆解

实现分为两个关键部分：

1. 核心逻辑修复（scheduler_runtime_checker_mixin.py）：
   - 在 check_memory 方法中，当 is_hybrid_ssm=True 时，将 token_usage 计算从直接使用 full_token_usage 改为 max(full_token_usage, mamba_usage)。
   - 代码片段：python full_token_usage, mamba_usage, _, _, _, _, _ = self._get_mamba_token_info() token_usage = max(full_token_usage, mamba_usage)
2. 文档注释添加（io_struct.py 和 metrics_collector.py）：
   - 在 token_usage 字段定义处添加 FIXME 注释，说明该字段实际表示所有内存池（KV、SWA、mamba）的最大使用率，而非仅 KV token 使用率。
   - 注释示例：# FIXME: token_usage is actually max usage across all pools (KV, SWA, mamba), not just KV token usage. Rename requires API deprecation.

评论区精华

由于该 PR 没有公开的 review 评论（review_comments_count: 0），讨论有限。从提交历史看，作者 hnyls2002 分两次提交：先修复逻辑（提交消息："fix idle token_usage missing mamba_usage"），后添加注释（提交消息："add FIXME for token_usage naming"），表明是自审自合并的 PR。关联 Issue 评论中，作者通过 /rerun-test 和 /tag-and-rerun-ci 命令触

发 CI 测试，确保修复通过现有 Mamba 测试套件。

风险与影响

- 技术风险：核心变更仅影响统计计算，不涉及内存分配或调度算法，回归风险低。但 token_usage 字段的命名误导性可能影响监控数据解读，不过注释已明确说明。
- 影响范围：
  - 对用户：无直接影响，纯内部修复。
  - 对系统：提升混合 SSM 场景下内存使用统计的准确性，有助于运维监控。
  - 对团队：澄清字段含义为未来可能的 API 重命名（需弃用流程）提供上下文，但当前保持向后兼容。

关联脉络

- 历史 PR 关联：
  - 12014：引入了空闲路径，是本修复的起源点。
  - 17862：重构了预填充 / 解码路径以包含 mamba_usage，但未更新空闲路径，导致不一致性，是本修复的直接原因。
- 演进趋势：该 PR 是 SGLang 内存管理演进中的小步修复，反映了项目对混合 SSM（如 Mamba）支持逐步完善。近期历史 PR 中，如 #22340（修复多层 EAGLE 草案扩展）、#22380（优化 Mamba 跟踪索引性能），也涉及推测解码和 Mamba 优化，显示团队在该领域的持续投入。