PR #24396 完整报告

sgl-project/sglang

[sgl] expose swa and mamba cache metrics

合并时间：2026-05-05 20:19

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24396

执行摘要

- 一句话：暴露 SWA 和 Mamba 混合缓存容量指标
- 推荐动作：该 PR 属于小范围增强，逻辑清晰，变更量小，适合快速阅读以了解 SGLang 缓存指标扩展模式。

功能与动机

PR body 明确指出需要暴露 SWA 和 Mamba 缓存容量指标，以便在负载下更轻松地监控混合缓存。现有指标仅覆盖 full-attention KV 缓存池，缺少对 hybrid-SWA 和 hybrid-SSM 缓存池的细粒度暴露。

实现拆解

1. 数据结构扩展：在 python/sglang/srt/observability/metrics_collector.py 的 SchedulerStats 数据类中新增 6 个字段：swa_available_tokens、swa_evictable_tokens、swa_used_tokens、mamba_available_tokens、mamba_evictable_tokens、mamba_used_tokens，用于记录各自池的绝对容量。
2. Prometheus Gauge 注册与日志：在 MetricsCollector.__init__ 中添加对应的 6 个 Gauge 指标（如 sglang:swa_available_tokens），并在 log_stats 方法中调用 _log_gauge 进行数值上报。
3. 调度器填充逻辑：在 python/sglang/srt/managers/scheduler_runtime_checker_mixin.py 的 update_scheduler_stats 方法中，当 is_hybrid_swa 或 is_hybrid_ssm 为真时，将调度器内部属性（如 self.swa_available_size）赋值给 stats 的对应字段。

关键文件：

- python/sglang/srt/observability/metrics_collector.py（模块 可观测性；类别 source；类型 core-logic；符号 SchedulerStats, MetricsCollector）：定义了 SchedulerStats 的字段和 Prometheus Gauge 注册，是核心变更文件。
- python/sglang/srt/managers/scheduler_runtime_checker_mixin.py（模块 调度器；类别 source；类型 core-logic；符号 update_scheduler_stats）：负责将调度器内部缓存容量数据填充到 SchedulerStats 结构，是数据来源。

关键符号：update_scheduler_stats, log_stats

关键源码片段

python/sglang/srt/observability/metrics_collector.py

定义了 SchedulerStats 的字段和 Prometheus Gauge 注册，是核心变更文件。

```
# SchedulerStats 数据类新增 SWA 和 Mamba 的绝对容量字段
# 位于 python/sglang/srt/observability/metrics_collector.py

class SchedulerStats:
    # ... existing fields ...
    # 新增的 SWA 池指标
    swa_available_tokens: int = 0
    swa_evictable_tokens: int = 0
    swa_used_tokens: int = 0
    # 新增的 Mamba 池指标
    mamba_available_tokens: int = 0
    mamba_evictable_tokens: int = 0
    mamba_used_tokens: int = 0

# MetricsCollector.__init__ 中注册 Prometheus Gauge
# 位于同一文件，紧跟在 kv_used_tokens Gauge 之后
self.swa_available_tokens = Gauge(
    name="sglang:swa_available_tokens",
    documentation="Number of free token slots in the SWA pool (hybrid-SWA only).",
    labelnames=labels.keys(),
    multiprocess_mode="mostrecent",
)
self.swa_evictable_tokens = Gauge(
    name="sglang:swa_evictable_tokens",
    documentation="Number of evictable (radix-cached) token slots in the SWA pool.",
    labelnames=labels.keys(),
    multiprocess_mode="mostrecent",
)
self.swa_used_tokens = Gauge(
    name="sglang:swa_used_tokens",
    documentation="Number of actively used token slots in the SWA pool.",
    labelnames=labels.keys(),
    multiprocess_mode="mostrecent",
)
# 类似地定义 mamba_available_tokens、mamba_evictable_tokens、mamba_used_tokens
```

python/sglang/srt/managers/scheduler_runtime_checker_mixin.py

负责将调度器内部缓存容量数据填充到 SchedulerStats 结构，是数据来源。

```
# 在 update_scheduler_stats 方法中，新增的赋值逻辑（位于同一文件）
def update_scheduler_stats(self, stats: SchedulerStats) -> None:
    """Update pool-related fields on SchedulerStats."""
    # ... existing assignments ...
    if self.is_hybrid_swa:
        stats.swa_token_usage = self.swa_token_usage
        stats.swa_available_tokens = self.swa_available_size # 新增
```

```
        stats.swa_evictable_tokens = self.swa_evictable_size # 新增
        stats.swa_used_tokens = self.swa_num_used # 新增
    if self.is_hybrid_ssm:
        stats.mamba_usage = self.mamba_usage
        stats.mamba_available_tokens = self.mamba_available_size # 新增
        stats.mamba_evictable_tokens = self.mamba_evictable_size # 新增
        stats.mamba_used_tokens = self.mamba_num_used # 新增
    # ... rest unchanged ...
```

评论区精华

代码审查由 ispobock 通过，无额外讨论。但 base 版本的注释（FIXME: misleadingly named "token_usage"）提示了命名历史遗留问题，本次变更未涉及。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅新增字段和 Gauge 注册，不修改现有逻辑；所有新增字段在写之前都受 is_hybrid_swa/is_hybrid_ssm 守卫，不会影响非混合模型。未引入测试，但变更模式简单且只读，回归风险小。
- 影响：对用户：所有混合模型（如 Gemma2、Jamba）的 Prometheus 监控中新增 6 个指标，便于容量规划。对系统：零性能影响，因为只是统计值传递。对团队：统一了 full/SWA/Mamba 的指标暴露模式，为后续指标扩展提供了样板。
- 风险标记：暂无

关联脉络

- PR #24389 consolidate NSA pool construction: 同一周内的 PR，涉及 pool 构建的简化，与缓存池管理相关。
- PR #24411 [diffusion] Fuse LTX2 split rotary embedding: 虽属 diffusion 模块，但同为性能 / 缓存优化相关 PR。