PR #25637 完整报告

sgl-project/sglang

Move batch-result processing to SchedulerBatchResultProcessor and retire output_processor mixin

合并时间：2026-05-18 18:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25637

执行摘要

- 一句话：将 batch 结果处理逻辑从 mixin 迁移至独立组件
- 推荐动作：本 PR 是调度器重构链中的一环，建议相关开发者阅读以了解解耦模式。由于变更机械，无需深入审查逻辑，但可关注如何通过逐步提取实现大型 mixin 的拆解。

功能与动机

PR body 明确指出这是 'Mechanical cut + paste for the introduce-batch-result-processor mech move (final extract from SchedulerOutputProcessorMixin)'，目的是将调度器的输出处理逻辑从混入类彻底迁移到独立的 batch 结果处理器组件中，进一步解耦 Scheduler 类。

实现拆解

1. 在 batch_result_processor.py 中新增方法：从 scheduler_output_processor_mixin.py 复制所有 process_batch_result_* 及辅助方法（如 _maybe_collect_routed_experts），去掉 @staticmethod 装饰器，并将 self: "SchedulerBatchResultProcessor" 简化为普通 self。
2. 补充导入：在 batch_result_processor.py 中添加 torch、Req、ScheduleBatch、LogitsProcessorOutput 等依赖项。
3. 删除 scheduler_output_processor_mixin.py 文件（722 行）。
4. 更新 scheduler.py：移除对 SchedulerOutputProcessorMixin 的导入和多重继承，将所有调用方式从 self.<method>(self.batch_result_processor, ...) 改为 self.batch_result_processor.<method>(...)。
5. 更新 decode.py：同样将 process_batch_result_prebuilt 的调用方式改为 self.batch_result_processor.process_batch_result_prebuilt(...)。 注：未涉及测试、配置或部署配套变更。

关键文件：

- python/sglang/srt/managers/scheduler_output_processor_mixin.py（模块 调度器；类别 source；类型 deletion；符号 SchedulerOutputProcessorMixin, process_batch_result_prebuilt, _maybe_collect_routed_experts, _maybe_collect_indexer_topk）：该文件是整个 batch 结果处理逻辑的原始容器，被完全删除，是本次重构的源点。

- python/sglang/srt/managers/scheduler_components/batch_result_processor.py（模块 结果处理器；类别 source；类型 dependency-wiring；符号 process_batch_result_prebuilt, _maybe_collect_routed_experts, _maybe_collect_indexer_topk, _maybe_collect_customized_info）：该文件是 batch 结果处理的新家，新增了全部处理方法，是重构后的核心。
- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 dependency-wiring）：调度器主文件，移除了对 SchedulerOutputProcessorMixin 的依赖和调用方式调整。
- python/sglang/srt/disaggregation/decode.py（模块 分离解码；类别 source；类型 core-logic）：分离解码模块中有对 process_batch_result_prebuilt 的调用，需同步更新。

关键符号：process_batch_result_prebuilt, _maybe_collect_routed_experts, _maybe_collect_indexer_topk, _maybe_collect_customized_info, process_batch_result_prefill, _resolve_spec_overlap_tokens, process_batch_result_idle, process_batch_result_decode

关键源码片段

python/sglang/srt/managers/scheduler_components/batch_result_processor.py

该文件是 batch 结果处理的新家，新增了全部处理方法，是重构后的核心。

```
# 以下展示 process_batch_result_prebuilt 方法，它从 SchedulerOutputProcessorMixin
# 迁移过来，去掉了 @staticmethod 并将 self 类型注解简化为普通实例方法。
def process_batch_result_prebuilt(self, batch: ScheduleBatch):
    # 断言当前为分离 DECODE 模式
    assert self.disaggregation_mode == DisaggregationMode.DECODE
    use_free_group = self.server_args.disaggregation_decode_enable_radix_cache
    if use_free_group:
        self.token_to_kv_pool_allocator.free_group_begin()
    for req in batch.reqs:
        req.time_stats.set_decode_prebuilt_finish_time()
        req.check_finished()
        if req.finished():
            req.time_stats.set_quick_finish_time()
            if self.server_args.enable_hisparse:
                self.hisparse_coordinator.request_finished(req)
            release_kv_cache(req, self.tree_cache)
    # Logprobs 由 prefill 引擎处理，此处只做流式输出
    self.output_streamer.stream_output(batch.reqs, batch.return_logprob)
    if use_free_group:
        self.token_to_kv_pool_allocator.free_group_end()
```

评论区精华

该 PR 无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：该 PR 为纯机械移动，方法体无任何逻辑变化，回归风险极低。但需要注意以下两点：
  - 若存在外部脚本或未包含在本次变更中的文件也调用了 SchedulerOutputProcessorMixin 的方法，可能导致运行时错误（经查，仅 scheduler.py 和 decode.py 引用了相关方法，已全部更新）。
  - 缺少针对 SchedulerBatchResultProcessor 的独立单元测试，后续重构中可考虑补充。
  - 影响：对用户透明，功能完全一致。对开发团队而言，Scheduler 的多重继承减少，职责边界更清晰；结果处理逻辑集中到 SchedulerBatchResultProcessor 后更易于单元测试和后续优化。影响范围仅限调度器核心模块，其他模块无感知。
  - 风险标记：核心路径变更，无变更测试

关联脉络

- PR #25638 Move module-level helpers out of scheduler.py: 同一重构链的前一步，将 scheduler.py 中的模块级辅助函数移到独立文件，为本次移动 mixin 做准备。
- PR #25639 Delete the now-unused is_work_request from scheduler.py: 清理调度器中已无用的函数，与本次删除 mixin 同为清理遗留逻辑。