PR #20845 完整报告

sgl-project/sglang

Fix prefill batch iter logging under overlap

合并时间：2026-05-07 17:10

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20845

执行摘要

- 一句话：修复 overlap 调度下 prefill batch 日志迭代器号错误
- 推荐动作：值得精读，特别是其设计抉择：在 run_batch 冻结快照而非 process_batch_result。理解此区别有助于掌握 overlap 调度下全局状态与局部状态的交互。

功能与动机

Under overlap scheduling, the scheduler can launch the current batch before processing the previous batch result and increase forward_ct. The old prefill logging used the current global forward_ct when emitting Prefill batch lines, so completed batches could be mislabeled as a later iter. 当下一批次无任务且不增加 forward_ct 时，前后连续完成的 batches 甚至会打印相同的 iter id。

实现拆解

1. 在 ScheduleBatch 类中新增 forward_iter: Optional[int] 属性，并在 copy() 方法中同步复制该属性，确保快照沿 batch 生命周期传递。
2. 在 scheduler.py 的 run_batch() 中，递增 self.forward_ct 后立即将新值赋给 batch.forward_iter，冻结该批次的迭代身份标识。
3. 在 scheduler_metrics_mixin.py 的 report_prefill_stats() 和 report_decode_stats() 中，优先读取 batch.forward_iter（仅当其为非 None 时使用），否则回退到 self.forward_ct。
4. 在三个调用点（dllm mixin、scheduler_output_processor_mixin、disaggregation prefill）向 report_prefill_stats 新增 batch 参数，使可观测性函数能够访问到 batch 对象。

关键文件：

- python/sglang/srt/observability/scheduler_metrics_mixin.py（模块 可观测性；类别 source；类型 core-logic；符号 report_prefill_stats, report_decode_stats）：修改了日志迭代器生成逻辑，是修复的核心文件。
- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 run_batch）：在 run_batch 方法中赋值 forward_iter，决定了快照时机。
- python/sglang/srt/managers/schedule_batch.py（模块 调度批次；类别 source；类型 data-contract；符号 ScheduleBatch, ScheduleBatch.copy）：定义 forward_iter 属性并在 copy 方法中复制它，确保快照随 batch 传递。

- python/sglang/srt/managers/scheduler_output_processor_mixin.py（模块 调度器输出；类别 source；类型 core-logic；符号 process_batch_result_prefill）：在 process_batch_result_prefill 中向 report_prefill_stats 传入 batch 参数。
- python/sglang/srt/dllm/mixin/scheduler.py（模块 DLLM 调度器；类别 source；类型 core-logic；符号 process_batch_result_dllm）：在 DLLM 混合调度器中向 report_prefill_stats 传入 batch 参数。
- python/sglang/srt/disaggregation/prefill.py（模块 分离式 prefill；类别 source；类型 core-logic；符号 process_batch_result_disagg_prefill）：在分离式 prefill 处理中向 report_prefill_stats 传入 batch 参数。

关键符号：report_prefill_stats, report_decode_stats, run_batch, ScheduleBatch.copy

关键源码片段

python/sglang/srt/observability/scheduler_metrics_mixin.py

修改了日志迭代器生成逻辑，是修复的核心文件。

```
# scheduler_metrics_mixin.py — 使用 batch 级 forward_iter 替代全局 forward_ct

def report_prefill_stats(
    self: Scheduler,
    batch: Optional[ScheduleBatch], # 新增 batch 参数
    prefill_stats: PrefillStats,
    can_run_cuda_graph: bool,
    dp_cooperation_info: Optional[DPCooperationInfo] = None,
):
    ...
    # 优先使用 batch 自带的 forward_iter，否则回退到全局 forward_ct
    batch_iter = (
        batch.forward_iter
        if batch is not None and batch.forward_iter is not None
        else self.forward_ct
    )
    iter_msg = f" [{batch_iter}]" if LOG_FORWARD_ITERS else ""
    ...
```

python/sglang/srt/managers/scheduler.py

在 run_batch 方法中赋值 forward_iter，决定了快照时机。

```
# scheduler.py•run_batch — 冻结 batch 的迭代号

def run_batch(
    self,
    batch: ScheduleBatch,
    pp_proxy_tensors: Optional[PPProxyTensors] = None,
) -> Union[GenerationBatchResult, EmbeddingBatchResult]:
    self.forward_ct += 1
    batch.forward_iter = self.forward_ct # 快照当前 forward_ct
```

```
    ...
```

python/sglang/srt/managers/schedule_batch.py

定义 forward_iter 属性并在 copy 方法中复制它，确保快照随 batch 传递。

```
# schedule_batch.py — ScheduleBatch 数据属性扩展

class ScheduleBatch(ScheduleBatchDisaggregationDecodeMixin):
    ...
    forward_iter: Optional[int] = None # 新增，用于稳定日志迭代号
    ...
    def copy(self) -> "ScheduleBatch":
        return ScheduleBatch(
            ...
            forward_iter=self.forward_iter, # 复制快照
            ...
        )
```

评论区精华

- gemini-code-assist[bot] 指出 report_decode_stats 同样应使用 batch-specific forward_iter，以避免 overlap 下 decode 日志错标；该建议被采纳，最终实现覆盖了 decode 分支。
- ShangmingCai 询问 prefill 日志中应使用 self.forward_ct 还是 self.forward_ct+1；作者未直接回应，但最终代码采用 batch.forward_iter（值等于 run_batch 时的 forward_ct），即修正了原来 prefill 日志 +1 的 off-by-one 问题。
- sufeng-buaa 质疑为什么不在 process_batch_result 中赋值 forward_iter；作者解释：在 process_batch_result 赋值相当于读取处理时的全局计数器，不能排除后续 run_batch 的影响；在 run_batch 赋值确保快照是启动时的值。sufeng-buaa 回复 'ok' 表示理解。
- decode 日志也应使用 batch.forward_iter (correctness): 被采纳，最终实现中 report_decode_stats 同样使用 batch_iter 逻辑。
- prefill 日志应使用 forward_ct 还是 forward_ct+1 (question): 最终代码采用 batch.forward_iter（等于 run_batch 时的 forward_ct），修正了原 prefill 日志 +1 的 off-by-one 问题。
- forward_iter 赋值时机：run_batch vs process_batch_result (design): 作者解释后，sufeng-buaa 回复 'ok' 确认理解。

风险与影响

- 风险：
  1. 向后兼容：当 batch 为 None 或 forward_iter 为 None 时，日志回退到 self.forward_ct，对于 prefill 日志的迭代号会比旧版少 1（原使用 forward_ct+1），但所有现有调用路径均传入非空 batch，且旧 off-by-one 应视为 bug 而非特性，风险可控。
  2. 测试覆盖：本次改动未包含对应测试，若未来重构 batch 生命周期可能遗漏 forward_iter 复制，建议增加单元测试覆盖。

3. 性能：仅新增一个整型属性赋值与读取，无性能影响。 - 影响：影响范围：仅当开启 overlap 调度（enable_overlap=True）且 LOG_FORWARD_ITERS 为 True 时日志行为改变。非 overlap 模式下，若 batch 传入非 None，prefill 日志迭代号从原 forward_ct+1 变为 forward_ct（修正 off-by-one）；decode 日志迭代号不变。整体改善了调度可观测性数据的准确性，便于调试和性能分析。 - 风险标记：缺少测试覆盖，快照时机依赖性

关联脉络

- 暂无明显关联 PR