PR #24858 完整报告

sgl-project/sglang

multi_layer_eagle: add tracing hooks

合并时间：2026-05-14 06:29

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24858

执行摘要

- 一句话：multi-layer eagle 添加追踪钩子
- 推荐动作：值得合并，改动明确、风险低，提升了 multi-layer eagle 的可观测性。

功能与动机

为 multi_layer_eagle_worker.py 添加与 V1 EAGLE 相同的可观测性钩子，使 multi-layer eagle 的追踪能力与 V1 对齐，便于调试和性能分析。

实现拆解

1. 在 multi_layer_eagle_worker.py 中导入 set_time_batch 和 get_global_tracing_enabled。
2. 在 forward_batch_generation 方法的 decode 分支中，在 draft/verify 前后插入 set_time_batch 调用以记录各阶段时间。
3. 在 verify 完成后，若追踪启用，遍历 batch.reqs 并调用 req.time_stats.set_spec_verify_end_time 记录每个请求的正确 draft 数量。
4. 在返回的 GenerationBatchResult 中新增 num_correct_drafts_per_req_cpu 字段，透传 verify_output 中的对应数组。

关键文件：

- python/sglang/srt/speculative/multi_layer_eagle_worker.py（模块 投机解码；类别 source；类型 dependency-wiring；符号 forward_batch_generation）：唯一变更文件，添加了追踪钩子和透传字段。

关键符号：forward_batch_generation

关键源码片段

python/sglang/srt/speculative/multi_layer_eagle_worker.py

唯一变更文件，添加了追踪钩子和透传字段。

```
# 新增导入
from sglang.srt.observability.req_time_stats import set_time_batch
from sglang.srt.observability.trace import get_global_tracing_enabled

# 在 forward_batch_generation 的 decode 分支中
```

```
# 设置 draft 开始时间（只用于 trace）
set_time_batch(batch.reqs, "set_spec_draft_start_time", trace_only=True)

with (
    self.draft_tp_context(self.mtp_model_runner(0).tp_group),
    speculative_moe_backend_context(),
):
    verify_input = self.draft(batch)

# 记录 draft 结束和 verify 开始时间
set_time_batch(batch.reqs, "set_spec_draft_end_time", trace_only=True)
set_time_batch(batch.reqs, "set_spec_verify_start_time", trace_only=True)

batch.spec_info = verify_input
logits_output, verify_output, can_run_cuda_graph = self.verify(batch)

# 若全局追踪启用，逐个请求记录 verify 结束时间和正确 draft 数
if get_global_tracing_enabled():
    for idx, req in enumerate(batch.reqs):
        num_correct_drafts = verify_output.num_correct_drafts_per_req_cpu[idx]
        req.time_stats.set_spec_verify_end_time(
            num_correct_drafts=num_correct_drafts
        )

# 记录 draft extend 开始时间
set_time_batch(
    batch.reqs, "set_spec_draft_extend_start_time", trace_only=True
)
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：较低的回归风险：仅添加可观测性钩子和透传字段，不修改原有逻辑。但需确保 set_time_batch 和 get_global_tracing_enabled 在调用上下文中正确可用。
- 影响：对用户透明，无接口变化。对开发者有益：multi-layer eagle 的时间追踪和 draft 统计现在与 V1 EAGLE 一致，便于调试和性能分析。
- 风险标记：暂无

关联脉络

- 暂无明显关联 PR