

PR #40698 完整报告

vllm-project/vllm

[BugFix] Correct OTEL span start time for Dynamo compilation

合并时间: 2026-07-14 07:25

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/40698>

执行摘要

- 一句话: 修复 Dynamo OTEL span 起始时间错误
- 推荐动作: PR 修复逻辑正确, 改动简洁, 已获 reviewer 确认和独立复现。建议合并。

功能与动机

Dynamo 编译的 OpenTelemetry span 起始时间使用了 `time.perf_counter()`, 该值非 Unix 时间戳, 导致 span 显示为 1970 年, 持续时间约 56 年。需要修正以正确记录 span 起始时间。

实现拆解

1. 计算当前单调时间: 在 `vllm/compilation/backends.py` 的 `__call__` 方法中, 新增 `current_perf = time.perf_counter()` 记录当前单调时间。
2. 获取当前 Unix 时间戳: 新增 `current_epoch = time.time()` 获取当前准确 Unix 时间戳。
3. 计算编译耗时并反推起始时间: `dynamo_time = current_perf - torch_compile_start_time` 得到编译耗时 (基于单调时钟), 然后 `real_start_time = current_epoch - dynamo_time` 得到准确的 Unix 起始时间。
4. 更新 OTEL span 调用: 将 `instrument_manual` 的起始时间参数从 `int(torch_compile_start_time * 1e9)` 改为 `int(real_start_time * 1e9)`, 确保 OTEL 正确解释时间戳。

关键文件:

- `vllm/compilation/backends.py` (模块 编译后端; 类别 source; 类型 core-logic): 核心修复文件, 修改了 Dynamo bytecode 转换 span 的起始时间计算方式。

关键符号: 未识别

关键源码片段

`vllm/compilation/backends.py`

核心修复文件, 修改了 Dynamo bytecode 转换 span 的起始时间计算方式。

```
# vllm/compilation/backends.py ( 部分 __call__ 方法 )
```

```
# ... 前面代码不变 ...
```

```
compilation_counter.num_graphs_seen += 1
```

```

from .monitor import torch_compile_start_time

# 修复前：直接使用 torch_compile_start_time（来自 time.perf_counter()）
# 作为 OTEL 起始时间，导致时间戳错误（1970 年附近）

current_perf = time.perf_counter() # 当前单调时间
current_epoch = time.time() # 当前 Unix 时间戳
dynamo_time = current_perf - torch_compile_start_time # 编译耗时（秒）

logger.info_once(
    "Dynamo bytecode transform time: %.2f s",
    dynamo_time,
)

# 反推出真实的起始 Unix 时间戳
real_start_time = current_epoch - dynamo_time
start_time_ns = int(real_start_time * 1e9) # 转换为纳秒

attributes = {"dynamo.time_seconds": dynamo_time}
# 使用修正后的起始时间进行 OTEL 打点
instrument_manual("Dynamo bytecode transform", start_time_ns, None, attributes)
# ... 后续代码不变 ...

```

评论区精华

reviewer 确认了该 bug 的存在（在 H100 上复现了 55 年 span），并认可修复方案。无其他争议或未解决问题。

- 独立确认 bug 并一致认可修复方案 (correctness): 确认该修复正确，关闭了重复 PR #47994。

风险与影响

- 风险：风险极低。仅修改了一处计算逻辑，且新逻辑使用标准库函数 `time.time()` 和 `time.perf_counter()`，行为明确。修复后 span 时间正确，不会影响编译功能或性能。
- 影响：影响范围仅限于 Dynamo 编译的 OpenTelemetry tracing 数据。修复后 span 起始时间恢复正常，对用户无功能影响，对运维可观测性有正面改进。
- 风险标记：暂无

关联脉络

- PR #47994 [BugFix] Fix OTEL span start time for Dynamo compilation: reviewer 独立提交的相同修复 PR，因本 PR 更早而被关闭。