

PR #45905 完整报告

vllm-project/vllm

[KV Offloading] Remove dummy worker-side stats from OffloadingConnector

合并时间: 2026-06-18 13:13

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45905>

执行摘要

- 一句话: 移除 KV Offload 中 worker 侧虚拟统计
- 推荐动作: 简单、安全的清理 PR, 值得快速合并, 可作为 kv-connector 模块代码整洁的参考案例。

功能与动机

该临时代码是 RFC #44008 中约定的清理项。PR #43877 在调度器侧独立实现了统计收集, worker 侧返回空统计对象不再有意义。删除可避免误导性代码路径, 并让测试更加直接。

实现拆解

1. 删除 worker 分支的空统计返回(vllm/distributed/kv_transfer/kv_connector/v1/offloading_connector.py): 移除 get_kv_connector_stats() 方法中 if self.connector_worker is not None: return OffloadingConnectorStats() 及注释。方法简化为: 优先返回调度器统计, 否则返回 None。
2. 移除测试中的轮询变通(tests/v1/kv_connector/unit/test_offloading_connector.py): 删除 test_cpu_offloading_metrics 中用于等待 worker 统计刷新的 dummy generate 循环及对应的 deadline 变量。由于 worker 不再产生空统计, 测试可直接检查调度器统计。
3. 验证测试通过: 运行三个相关测试套件全部通过, 无需新增测试(移除的路径不携带数据, 已由 #43877 的测试覆盖)。

关键文件:

- vllm/distributed/kv_transfer/kv_connector/v1/offloading_connector.py (模块 KV 连接器; 类别 source; 类型 core-logic; 符号 get_kv_connector_stats): 核心变更文件, 删除了 get_kv_connector_stats 中 worker 侧的空统计返回及 TODO 注释。
- tests/v1/kv_connector/unit/test_offloading_connector.py (模块 测试; 类别 test; 类型 test-coverage; 符号 test_cpu_offloading_metrics): 移除了因 worker 统计而引入的 dummy generate 轮询循环, 使测试更简洁。

关键符号: get_kv_connector_stats, test_cpu_offloading_metrics

关键源码片段

[vllm/distributed/kv_transfer/kv_connector/v1/offloading_connector.py](#)

核心变更文件，删除了 `get_kv_connector_stats` 中 worker 侧的空统计返回及 TODO 注释。

```
# vllm/distributed/kv_transfer/kv_connector/v1/offloading_connector.py
```

```
def get_kv_connector_stats(self) -> KVConnectorStats | None:
    # 优先使用 scheduler 侧统计；worker 侧不再返回空统计
    # (PR #43877 已解耦调度器统计收集)
    if self.connector_scheduler is not None:
        return self.connector_scheduler.get_stats()
    return None
```

tests/v1/kv_connector/unit/test_offloading_connector.py

移除了因 worker 统计而引入的 dummy generate 轮询循环，使测试更简洁。

```
# tests/v1/kv_connector/unit/test_offloading_connector.py
# 此前此处有一段 retry 循环（已删除），其作用是通过 dummy generate
# 驱动异步统计刷新。删除后测试直接检查调度器侧统计值。
store_bytes = _get_counter_value("vllm:kv_offload_store_bytes")
assert store_bytes > 0, f"Expected store_bytes > 0, got {store_bytes}"
# ... 后续断言保持不变
```

评论区精华

没有 review 评论，只有 reviewer orozery 的审批通过。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：删除的代码是此前临时标记的 TODO，且实际不产生有效数据。若将来重新需要 worker 侧统计，可基于 #43877 的调度器路径扩展。无性能或兼容性影响。
- 影响：影响范围小，仅涉及 `OffloadingConnector` 的 `get_kv_connector_stats` 方法及其一个测试函数。不影响任何公共指标或用户可见行为。
- 风险标记：暂无

关联脉络

- PR #43877 [] Decouple scheduler-side stats collection from worker-side stats existence: 本 PR 的依赖和前置条件；解耦后使得删除 worker 侧空统计成为可能。