

PR #43877 完整报告

vllm-project/vllm

[Core][KV Connector] fix scheduler KV connector stats aggregation

合并时间: 2026-06-12 22:51

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43877>

执行摘要

- 一句话: 修复 scheduler 端 KV connector 统计聚合顺序
- 推荐动作: 建议审核并合并, 这是一个低风险、高价值的 bugfix, 且已有充分测试覆盖。值得关注的决策是将统计收集延迟到 connector 状态更新之后, 这是编写类似统计逻辑时的一个通用设计模式。

功能与动机

从 PR #35669 中提取的独立修复。原代码在 `update_from_output` 开始时提取 worker 端统计并与 scheduler 端聚合, 但此时 `update_connector_output()` 尚未被调用, 导致 scheduler 端统计不准确或丢失。需要保证 scheduler 端在其状态更新后再收集统计, 并正确处理仅 scheduler 端存在的情况。

实现拆解

步骤 1: 在 `scheduler.py` 的 `update_from_output` 方法中, 将原本位于函数顶部的 `kv_connector_stats` 提取与聚合代码移至 `_update_from_kv_xfer_finished` 调用之后。步骤 2: 调整后的逻辑: 只在 `self.connector` 存在时才获取 scheduler 端统计, 并通过 `is_empty()` 检查避免聚合空对象; 当 worker 端统计为 `None` 时, 直接使用 scheduler 端统计。步骤 3: 新增 `tests/v1/core/test_scheduler.py` 中的 `test_scheduler_kv_connector_stats` 测试, 使用 `GenericKVConnectorStats` 模拟类验证三种组合情况 (仅 worker、仅 scheduler、两者)。步骤 4: 修改 `tests/v1/kv_connector/unit/test_multi_connector.py`, 为 `MockConnector` 和 `MockHMAConnector` 的 mock 添加 `get_kv_connector_stats.return_value = None` 默认值, 并在事件过滤列表中增加 `get_kv_connector_stats`, 避免该轮查询调用干扰断言。步骤 5: 根据 review 反馈回退对 `tests/v1/e2e/general/test_mamba_prefix_cache.py` 的无关缩进改动。

关键文件:

- `vllm/v1/core/sched/scheduler.py` (模块 调度器; 类别 source; 类型 core-logic; 符号 `update_from_output`) : 核心修复文件: 调整 `kv_connector_stats` 聚合逻辑到 connector 更新之后。
- `tests/v1/core/test_scheduler.py` (模块 测试; 类别 test; 类型 test-coverage; 符号 `test_scheduler_kv_connector_stats`, `GenericKVConnectorStats`) : 新增 `test_scheduler_kv_connector_stats` 测试, 覆盖三种场景。

- tests/v1/kv_connector/unit/test_multi_connector.py (模块 连接器测试; 类别 test; 类型 test-coverage; 符号 MockConnector, MockHMAConnector, _ignore_event_collection)
: 修复 mock 默认返回值和事件过滤, 适配新的聚合逻辑。

关键符号: Scheduler.update_from_output, GenericKVConnectorStats.reset, GenericKVConnectorStats.aggregate, GenericKVConnectorStats.reduce, GenericKVConnectorStats.is_empty

关键源码片段

vllm/v1/core/sched/scheduler.py

核心修复文件: 调整 kv_connector_stats 聚合逻辑到 connector 更新之后。

```
# vllm/v1/core/sched/scheduler.py - update_from_output 方法片段

# ... 之前代码已处理 invalid_blocks、routed_experts 等 ...

# KV Connector: update state for finished KV Transfers.
if kv_connector_output:
    self._update_from_kv_xfer_finished(kv_connector_output)

# Worker-side KV connector stats from the model runner output.
kv_connector_stats: KVConnectorStats | None = (
    kv_connector_output.kv_connector_stats if kv_connector_output else None
)
if self.connector:
    # Scheduler-side KV connector stats collected after connector update.
    scheduler_kv_connector_stats = self.connector.get_kv_connector_stats()
    if (
        scheduler_kv_connector_stats is not None
        and not scheduler_kv_connector_stats.is_empty()
    ):
        # 若存在 worker 端统计 (来自 model_runner), 则与调度器端聚合;
        # 否则直接使用调度器端统计
        kv_connector_stats = (
            kv_connector_stats.aggregate(scheduler_kv_connector_stats)
            if kv_connector_stats is not None
            else scheduler_kv_connector_stats
        )
```

tests/v1/core/test_scheduler.py

新增 test_scheduler_kv_connector_stats 测试, 覆盖三种场景。

```
# tests/v1/core/test_scheduler.py - 新增测试函数

def test_scheduler_kv_connector_stats():
    """Test worker-side, scheduler-side, and combined KV connector stats."""

    # 内联模拟类, 继承 KVConnectorStats 并实现必要方法
```

```

class GenericKVConnectorStats(KVConnectorStats):
    def reset(self):
        self.data = {}

    def aggregate(self, other):
        self.data.update(other.data)
        return self

    def reduce(self):
        return {}

    def is_empty(self):
        return not self.data

# 测试场景三元组 : (worker_data, scheduler_data, expected_aggregated_data)
test_cases = (
    ({"worker": 1}, None, {"worker": 1}), # 仅 worker
    (None, {"scheduler": 2}, {"scheduler": 2}), # 仅 scheduler
    ({"worker": 1}, {"scheduler": 2}, {"worker": 1, "scheduler": 2}), # 两者
)

for worker_data, scheduler_data, expected_data in test_cases:
    scheduler = create_scheduler()
    worker_stats = (
        GenericKVConnectorStats(data=worker_data) if worker_data else None
    )
    scheduler_stats = (
        GenericKVConnectorStats(data=scheduler_data) if scheduler_data else None
    )
    scheduler.connector = Mock()
    # 初始返回 None (模拟 connector 尚未更新状态)
    scheduler.connector.get_kv_connector_stats.return_value = None
    scheduler.connector.take_events.return_value = []

    # 通过 side_effect 模拟 update_connector_output 调用后 connector 更新
    def update_connector_output(kv_connector_output, scheduler=scheduler,
                               scheduler_stats=scheduler_stats):
        scheduler.connector.get_kv_connector_stats.return_value = scheduler_stats

    scheduler.connector.update_connector_output.side_effect = update_connector_output

# 构造 ModelRunnerOutput, 携带 worker 端统计
model_output = ModelRunnerOutput(
    req_ids=["req_0"],
    req_id_to_index={"req_0": 0},
    sampled_token_ids=[[123]],
    logprobs=None,
    prompt_logprobs_dict={},
    pooler_output=[None],

```

```

        kv_connector_output=KVConnectorOutput(kv_connector_stats=worker_stats)
        if worker_stats else None,
    )
    scheduler_output = SchedulerOutput(
        scheduled_new_reqs=[],
        scheduled_cached_reqs=None,
        num_scheduled_tokens={"req_0": 1},
        total_num_scheduled_tokens=1,
        scheduled_spec_decode_tokens={},
        scheduled_encoder_inputs={},
        num_common_prefix_blocks=[0],
        finished_req_ids=set(),
        free_encoder_mm_hashes=[],
    )

    engine_core_outputs = scheduler.update_from_output(
        scheduler_output, model_output
    )

    final_stats = next(
        iter(engine_core_outputs.values())
    ).scheduler_stats.kv_connector_stats
    assert final_stats == expected_data

```

评论区精华

- orozery建议在代码中添加注释，明确区分 worker-side 和 scheduler-side 的统计变量，并建议将变量重命名以提高可读性（已采纳）。
- orozery提议将多个测试用例合并为一个，并嵌套 GenericKVConnectorStats 类（已采纳）。
- markmc指出对 tests/v1/e2e/general/test_mamba_prefix_cache.py 的缩进修正是无关更改，应还原（已修复），并提到 #45345 可能是该问题的真正修复。
- orozery和 markmc均审批通过（APPROVED）。
- 添加注释区分 worker 和 scheduler 统计 (style): 已采纳: head 版本中已添加注释 '# Worker-side KV connector stats ...' 和 '# Scheduler-side KV connector stats ...'。
- 测试用例合并与类内联 (testing): 已采纳: 最终版本为单一 test_scheduler_kv_connector_stats 函数，内部使用嵌套类并循环测试三个 case。
- 无关文件更改应回退 (design): 已回退: head 版本中不包含该文件更改。

风险与影响

- 风险：核心路径变更：update_from_output 是 scheduler 主循环的关键方法，修改统计聚合顺序可能影响依赖 kv_connector_stats 的上层指标收集和监控。但改动范围小（24 行），且新增单元测试覆盖了主要分支，风险可控。多连接器测试影响：修改 MockConnector 的默认返回值和事件过滤器，可能屏蔽对其他 get_kv_connector_stats 调用的验证，但事件过滤已明确为不关心的轮询调用，不会遗漏必要断言。

- 影响：影响范围：所有使用 KV Connector 的 v1 推理场景（分布式 KV 传输）。修复后 `kv_connector_stats` 将正确反映 scheduler 端和 worker 端的聚合数据，提升监控和调试准确性。对系统中无 KV Connector 的配置无影响。团队影响：减少统计相关 bug 排查成本。
- 风险标记：核心路径变更（`update_from_output`），多 mock 默认值调整可能掩藏问题

关联脉络

- PR #35669 Feature/offloading manager stats: 本 PR 从中提取 scheduler KV connector stats 聚合修复，是该特性的直接子 PR。
- PR #45345 Unknown (mentioned in review): review 中 markmc 提到该 PR 可能是 `test_mamba_prefix_cache.py` 真正问题的修复，关联此 PR 中尝试的无关修改。