

PR #49582 完整报告

vllm-project/vllm

[EC Connector] Add has_pending_push_work

合并时间: 2026-07-29 17:04

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49582>

执行摘要

- 一句话: EC Connector 新增 pending push 检测
- 推荐动作: 此 PR 为简单接口扩展, 设计清晰, 实现安全。值得关注的点是如何在子类中实际利用该钩子。

功能与动机

让 EC Connector 能够像 KV Connector 那样, 通过 `has_pending_push_work()` 通知引擎主循环仍有推送传输工作正在进行, 防止引擎在推送传输完成前退出。

实现拆解

1. 在 `vllm/distributed/ec_transfer/ec_connector/base.py` 的 `ECConnectorBase` 类中添加 `has_pending_push_work()` 方法, 返回 `False`, 供子类覆盖。
2. 在 `vllm/v1/core/sched/scheduler.py` 的 `has_requests()` 方法中增加对 `ec_connector.has_pending_push_work()` 的检查, 与已有 `connector.has_pending_push_work()` 并列。

关键文件:

- `vllm/distributed/ec_transfer/ec_connector/base.py` (模块 分布式连接器; 类别 `source`; 类型 `core-logic`; 符号 `has_pending_push_work`): 定义 `has_pending_push_work` 接口, 默认返回 `False`
- `vllm/v1/core/sched/scheduler.py` (模块 调度器; 类别 `source`; 类型 `core-logic`): 在 `has_requests` 中增加对 `ec_connector` 的 `pending push` 检查, 保持引擎活跃

关键符号: `has_pending_push_work`, `has_requests`

关键源码片段

`vllm/v1/core/sched/scheduler.py`

在 `has_requests` 中增加对 `ec_connector` 的 `pending push` 检查, 保持引擎活跃

```
def has_requests(self) -> bool:
    # Override the interface default to also keep the engine alive while a
    # connector still has pending push work (e.g. push-mode WRITE transfers
    # in flight after all "live" requests have finished). Without this hook
```

```
# the engine would quiesce before the connector can drain completions.
# TODO: replace with a more general mechanism for connectors to keep
# the scheduler alive.
return (
    self.has_unfinished_requests()
    or self.has_finished_requests()
    or (self.connector is not None and self.connector.has_pending_push_work())
    or (
        self.ec_connector is not None
        and self.ec_connector.has_pending_push_work()
    )
)
```

评论区精华

无 Review 讨论。NickLucche 直接批准，无评论。

- 暂无高价值评论线程

风险与影响

- 风险：改动极小，新增方法默认返回 False，不影响现有 EC Connector 行为。Scheduler 侧增加的条件是安全短路逻辑，不会有性能影响。风险极低。
- 影响：对使用 EC Connector 并实现推送传输的模型（如 EPD）有益，能正确保持引擎活跃直到传输完成。其他模型无影响。
- 风险标记：暂无

关联脉络

- PR #49114 Add CachePolicyFactory for pluggable/external eviction policies: 同为 connector/policy 相关的接口扩展，本 PR 与 KV Connector 的 has_pending_push_work 对齐。