

PR #45294 完整报告

vllm-project/vllm

Update hidden states extraction integration test triggers

合并时间: 2026-06-12 16:05

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45294>

执行摘要

- 一句话: 扩展现有 hidden states 集成测试触发条件
- 推荐动作: 建议精读。此 PR 展示了维护大型项目中 CI 测试触发策略的典型实践——当有新功能测试文件或关键依赖文件引入时, 需要同步更新 CI 触发规则。值得关注的决策是将 `example_hidden_states_connector.py` 显式加入依赖 (而非笼统依赖目录), 以及为多 GPU 测试单独设立 job, 避免与单 GPU 测试混淆。

功能与动机

根据 PR body, `example_hidden_states_connector.py` 被 hidden states 提取系统大量使用, 但修改该文件不会触发集成测试; 同时 #43805 添加了需要 TP=2 的测试, 但未添加对应的 CI job。此次变更旨在填补这些 CI 触发缺失。

实现拆解

1. 修改 `extract-hidden-states-integration` job 的 `source_file_dependencies`: 在 `.buildkite/test_areas/misc.yaml` 中, 向现有 hidden states 集成测试 job 的依赖列表添加 `vllm/distributed/kv_transfer/kv_connector/v1/example_hidden_states_connector.py`, 使得该文件变更时自动触发该 job。
2. 新增 `Extract Hidden States Integration (2 GPUs)` job: 在同一个 YAML 文件中, 添加一个新 job, 其 key 为 `extract-hidden-states-integration-2-gpus`, 使用 2 个 GPU (`num_devices: 2`), 依赖与第一个 job 相同, 但运行命令增加了 `-m 'distributed'` 标记, 仅执行分布式相关的测试用例。新 job 在第一个 job 之后定义 (位于 Regression job 之前)。
3. 确保两个 job 共享相同的 timeout (20 分钟) 和设备 (`h200_18gb`), 使用 spawn 模式启动 worker, 保持一致性。

关键文件:

- `.buildkite/test_areas/misc.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 唯一变更的文件; 修改了 hidden states 集成测试的依赖列表, 新增了多 GPU 测试 job。

关键符号: 未识别

评论区精华

无 review 评论; PR 由 mgoin 直接批准, 无争议。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅涉及 CI 配置文件，不会影响任何业务代码。新增 job 在原有硬件资源上运行，需注意 2 GPU job 可能增加 CI 资源消耗，但 timeout 和 device 配置与现有 job 类似，不会额外抢占总线。实际运行结果（PR 作者确认）显示新 job 成功通过。
- 影响：
 - CI 系统：增加了一个需要 2 GPU 的并行 job，但整体 CI pipeline 时间可能略微增加（约 20 分钟），因为新 job 与现有 job 并行运行。资源消耗增加，但对于 h200_18gb 队列影响有限。
 - 开发团队：今后修改 `example_hidden_states_connector.py` 或 `tests/v1/kv_connector/extract_hidden_states_integration` 下的测试会自动触发集成测试，提高回归覆盖。分布式测试不再被遗漏。
 - 下游用户：无直接影响。
 - 风险标记：CI 资源增加，无业务风险

关联脉络

- PR #43805 [Add TP=2 hidden states extraction test]: 此 PR 引入了需要 2 GPU 的测试，但未添加相应的 CI job 触发；当前 PR 正是为了补齐该遗漏。