

PR #46456 完整报告

vllm-project/vllm

[CI] intel CI: add quantization and awq case for xpu

合并时间: 2026-07-03 09:51

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/46456>

执行摘要

- 一句话: Intel CI 新增量化和 AWQ 测试
- 推荐动作: 该 PR 是典型的 CI 增强变更, 逻辑简单。如果读者关注 Intel GPU 上的量化支持状态, 可以了解新增的测试覆盖。对于大多数开发者, 无需深入阅读。

功能与动机

根据 PR 标题和提交消息, 目的是为 Intel GPU (XPU) CI 添加量化和 AWQ 测试覆盖, 以验证 Intel 平台上的量化推理功能是否正常工作。PR body 仅包含模板, 未提供详细动机, 但从变更内容可判断是基建层面的测试增强。

实现拆解

1. 新建独立的量化 CI Job: 新增 `.buildkite/intel_jobs/quantization.yaml`, 定义了一个名为 "Quantization" 的 Buildkite pipeline step。该 job 在 `intel_gpu` 设备上运行, 依赖 `image-build-xpu`, 通过 `run-intel-test.sh` 执行 `tests/quantization/test_per_token_kv_cache.py` 中的测试, 同时排除了 `test_triton_unified_attention_per_token_head_scale[int4-16-128-num_heads0-seq_lens1]` (该用例已知错误)。
2. 在现有 XPU Example Test 中增加 AWQ 模型推理: 在 `.buildkite/intel_jobs/test-intel.yaml` 的 "XPU example Test" 步骤中, 新增一行命令, 使用 `TheBloke/TinyLlama-1.1B-Chat-v0.3-AWQ` 模型运行基础的 `examples/basic/offline_inference/generate.py`, `block-size=64`, 模式为 `enforce-eager`。
3. 同步更新 CI 脚本: 在 `.buildkite/scripts/hardware_ci/run-intel-ci-test.sh` 的 `example test` 分支中, 同样新增上述 AWQ 模型推理命令, 保持与 YAML 配置一致。

关键文件:

- `.buildkite/intel_jobs/quantization.yaml` (模块 CI 配置; 类别 config; 类型 configuration) : 新增的量化 CI job 配置文件, 定义了 Quantization 测试步骤, 包括硬件要求、源文件依赖和具体的测试命令。
- `.buildkite/intel_jobs/test-intel.yaml` (模块 CI 配置; 类别 config; 类型 configuration) : 在 XPU Example Test 步骤中新增了一行 AWQ 模型推理命令, 是测试覆盖的核心补充。
- `.buildkite/scripts/hardware_ci/run-intel-ci-test.sh` (模块 CI 脚本; 类别 infra; 类型 infrastructure) : 同步更新脚本, 与 YAML 配置保持一致。

关键符号：未识别

评论区精华

该 PR 没有收到任何审查评论，仅有两次 approved，说明变更简单直接，未引起争议。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及 CI 配置和脚本，不修改任何业务逻辑代码。新增的 AWQ 用例执行了 `examples/basic/offline_inference/generate.py` 脚本，若 Intel 硬件上 AWQ 量化模型推理存在兼容性问题，可能导致该 CI job 失败，但不会影响主流程。量化 job 排除了一个已知失败的 triton 测试，避免了误报。
- 影响：影响范围限定在 Intel GPU 的 CI 管道。增量覆盖了 Intel XPU 上的量化和 AWQ 测试，有助于提前发现 Intel 平台上的量化推理回归问题。对其他 GPU 平台（NVIDIA/AMD）无影响。
- 风险标记：仅有配置变更，风险极低

关联脉络

- 暂无明显关联 PR