PR #43663 完整报告

vllm-project/vllm

[XPU][CI] Add more test cases in Intel GPU CI

合并时间：2026-06-08 14:21

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43663

执行摘要

- 一句话：扩展 Intel GPU CI 测试覆盖
- 推荐动作：该 PR 为纯 CI 配置变更，无核心逻辑改动。值得关注的是作者坚持 --deselect 前缀的做法以及 worker 测试的缺失；建议后续迭代中验证 deselect 有效性并补齐 worker 覆盖。

功能与动机

PR 目的为在 Intel GPU CI 中添加更多测试用例，以提升硬件平台测试覆盖率，确保功能正确性。参见 PR body：'Add more test cases in Intel GPU CI'。

实现拆解

实现分四步：

1. 新增多模态模型测试配置(models_multimodal_intel.yaml)：定义 4 个标准多模态 step（qwen2、qwen3+gemma、llava+qwen2_vl、其他 +whisper）和一个 processor step，后者设置 parallelism: 4 以支持 sharding。
2. 扩展现有 misc 配置(misc_intel.yaml)：在原 V1 Executor、V1 Sample+Logits、XPU CPU Offload 基础上，新增 Regression、Metrics & Tracing（2 GPU）、Async Engine/Inputs/Utils/Worker 步骤；同时将 V1 Sample + Logits 的 source_file_dependencies 从宽泛的 vllm/ 精确至具体子目录以精准触发。
3. 新增专家并行测试配置(expert_parallelism_intel.yaml)：加入 EPLB Algorithm 测试步骤，验证专家并行负载均衡逻辑。
4. 更新 runner 脚本(run-intel-test.sh)：添加 BUILDKITE_PARALLEL_JOB 和 BUILDKITE_PARALLEL_JOB_COUNT 环境变量传递，支持 test sharding。

关键文件：

- .buildkite/intel_jobs/models_multimodal_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：新增多模态模型 CI 测试配置，定义 5 个 step 覆盖主要多模态模型，是本次扩展的核心文件。
- .buildkite/intel_jobs/misc_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：扩展现有 CI 配置，新增 Regression、Metrics/Tracing、Async Engine 等 step，并细化依赖范围，是覆盖范围扩大的关键。

- .buildkite/intel_jobs/expert_parallelism_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：新增专家并行 EPLB 算法测试，确保 XPU 上 MLP 负载均衡逻辑正确。
- .buildkite/scripts/hardware_ci/run-intel-test.sh（模块 CI 脚本；类别 infra；类型 infrastructure）：修改测试 runner 脚本以传递 Buildkite 并行环境变量，支持 test sharding，是 processor step 正常工作的前提。

关键符号：未识别

评论区精华

- --deselect 路径前缀争议：gemini-code-assist[bot] 指出 --deselect 路径包含 tests/ 前缀，但 pytest 在 tests 目录内执行会导致 deselect 失效。作者回复“per CI results, it needs tests/”坚持保留，最终未修改。
- Worker 测试缺失：Async Engine, Inputs, Utils, Worker 步骤中缺少 worker 相关依赖和测试，作者以“aligned with misc.yaml”回应，表示与主仓库保持一致。
- 依赖范围收窄：将 vllm/ 拆分为具体子目录后，vllm/worker/、vllm/attention/ 等不再触发 V1 Sample/Logits 等步骤，作者同样以对齐主仓库为由。
- Basic Models 测试移除：jikunshang 认为 basic 模型测试可移除，最终该文件从 PR 中删除。
- EPLB 支持确认：jikunshang 询问 EPLB 是否已在 XPU 支持，未获公开回复但 PR 仍合并，可能已线下确认。
  - --deselect 路径前缀争议 (correctness): 作者未修改，最终合并时保留 tests/ 前缀。
  - Worker 测试缺失 (testing): 未添加 worker 测试，保持与主仓库 misc.yaml 一致。
  - 依赖范围收窄导致触发遗漏 (design): 接受当前依赖列表，与主仓库对齐。
  - Basic Models 测试配置移除 (other): 该文件（models_basic_intel.yaml）最终从 PR 中删除。
  - EPLB 在 XPU 上的支持确认 (question): 未收到公开回复，但 PR 已合并，推测已线下确认或暂时跳过。

风险与影响

- 风险：
  - 测试覆盖不完整：Async Engine, Inputs, Utils, Worker 步骤未包含 worker 测试，可能遗漏 Worker 回归。
  - 依赖遗漏风险：依赖从 vllm/ 改为精确子目录后，若核心代码变动（如 vllm/core/、vllm/attention/）但未列入依赖，相应测试不会被触发。
  - Deselect 可能失效：尽管作者依据 CI 结果保留 tests/ 前缀，但与常规 pytest 行为冲突，未来版本或配置变化可能导致大模型被错误运行。
  - 整体风险低：变更仅影响 CI 配置，不影响运行时逻辑。
- 影响：
  - CI 系统：Intel GPU CI 管道增加约 6 个新测试组，总耗时可能上升，但通过并行和 sharding 控制。

- 开发者：提交涉及 XPU 的变更时更多测试自动运行，及早发现回归。
- 用户：无直接影响，但间接提升 Intel GPU 平台发布的稳定性。
- 团队：CI 配置与主仓库（.buildkite/test_areas/）对齐，便于后续维护。
- 风险标记：Worker 测试覆盖缺失，依赖精确化可能遗漏触发，deselect 路径不一致，EPLB 支持未确认

关联脉络

- PR #36423 [XPU] Support cpu kv offloading and tiering offloading on XPU platform: 同样修改了 .buildkite/intel_jobs/misc_intel.yaml，属于同一 CI 配置文件的演进。
- PR #37149 [XPU][Feature] transparent sleep mode support for XPU platform: 之前添加了 XPU CI 配置（basic_correctness.yaml），本 PR 进一步扩展 CI 覆盖。
- PR #43087 Modify torch dependency in xpu.txt: 修改 XPU 依赖配置，与当前 PR 同属 Intel GPU CI 改进系列。