PR #49340 完整报告

vllm-project/vllm

[CI] Wire untethered test files into CI jobs

合并时间：2026-07-28 22:02

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49340

执行摘要

- 一句话：将 30+ 个未运行的测试文件接入 CI 任务
- 推荐动作：值得精读。该 PR 展现了系统性审计 CI 测试覆盖的方法，可作为 CI 维护的参考。尤其关注 kernels.yaml 中 catch-all job 的设计模式（使用 --ignore 排除已有 job 的文件），以及如何安全地将长期未运行的测试接入 CI。

功能与动机

PR body 指出：An audit of .buildkite pytest targets against tests/**/test_*.py found dozens of test files that no CI job ever runs. Wire in the ones that pass on current main (every set validated on B200-class hardware, or CPU-only for CPU tests, before wiring).

实现拆解

1. Kernels 根目录 catch-all job：在 .buildkite/test_areas/kernels.yaml 中新增 Kernels Root Misc Test (B200) job，收集 tests/kernels/ 根目录下所有测试文件，通过 --ignore 排除已有专属 job 或当前 broken 的文件。同时将 tests/models/test_deepseek_v4_mega_moe.py 接入 Deepseek V4 Kernel Test (B200)。
2. NIXL PD edge case 测试：在 .buildkite/test_areas/disaggregated.yaml 中新增 NixlConnector PD edge case test (2 GPUs) job，并修改 tests/v1/kv_connector/nixl_integration/run_edge_case_test.sh 将硬编码的 GPU ID（4/5）改为通过环境变量覆盖（默认 4/5，CI 中对 job 设置 PREFILL_GPU_ID=0、DECODE_GPU_ID=1），同时添加 --max-model-len 8192 参数以适应 CI 显存限制。修改 python 为 python3。
3. Batch invariance 测试：在 .buildkite/test_areas/misc.yaml 的 Batch Invariance (B200) job 中添加 test_matmul_batch_invariant.py、test_cutlass_batch_invariance.py、test_online_batch_invariance.py，并将超时从 35 分钟增加到 45 分钟。
4. 模型测试：在 .buildkite/test_areas/models_basic.yaml 的 Basic Models Test (Other CPU) job 中添加 tests/models/test_adapters.py（回归测试 #39650 fixes）。
5. 其他单测 job：在 .buildkite/test_areas/cuda.yaml 添加 cuda/test_cuda_compatibility_path.py，.buildkite/test_areas/spec_decode.yaml 添加 spec_decode/test_custom_proposer.py，.buildkite/test_areas/engine.yaml 添加

v1/test_tensor_ipc_queue.py，.buildkite/test_areas/misc.yaml 的 CPU job 添加 v1/test_kv_cache_spec_registry.py 和 tracing/test_loading_tracing.py。

6. 测试代码修复：在 tests/kernels/test_fused_minimax_m3_qknorm_rope_kv_insert.py 中，将 assert_close 的 tolerance 从 1e-2 放宽到 2e-2，并添加注释解释数值差异的原因为 fused kernel 在 norm->rope 之间使用 fp32 中间值，而 reference 在 norm 后截断为 bf16。

关键文件：

- .buildkite/test_areas/kernels.yaml（模块 CI 配置；类别 config；类型 configuration）：核心变更：新增 catch-all job 覆盖 kernels 根目录测试，并接入 deepseek_v4_mega_moe 模型测试。
- tests/kernels/test_fused_minimax_m3_qknorm_rope_kv_insert.py（模块 内核测试；类别 test；类型 test-coverage）：测试代码修改：放宽 tolerance 并添加注释说明数值差异原因，是 PR 中唯一的测试源代码变更。
- tests/v1/kv_connector/nixl_integration/run_edge_case_test.sh（模块 KV 连接器；类别 test；类型 test-coverage）：Shell 脚本修改：允许通过环境变量覆盖 GPU ID，并添加 --max-model-len 参数，使脚本可在 CI 环境中灵活运行。
- .buildkite/test_areas/disaggregated.yaml（模块 CI 配置；类别 config；类型 configuration）：新增 NixlConnector PD edge case test job，正式将之前未运行的 edge case 测试脚本纳入 CI。
- .buildkite/test_areas/misc.yaml（模块 CI 配置；类别 config；类型 configuration）：Batch Invariance 测试增加三个新的测试文件，并调整超时；CPU job 和 tracing job 也新增测试文件。
- .buildkite/test_areas/models_basic.yaml（模块 CI 配置；类别 config；类型 configuration）：在 CPU 基础模型测试中添加 test_adapters.py 回归测试。
- .buildkite/test_areas/engine.yaml（模块 CI 配置；类别 config；类型 configuration）：添加 v1/test_tensor_ipc_queue.py 到 Engine (1 GPU) job。
- .buildkite/test_areas/spec_decode.yaml（模块 CI 配置；类别 config；类型 configuration）：添加 spec_decode/test_custom_proposer.py 到 Spec Decode Ngram + Suffix job。
- .buildkite/test_areas/cuda.yaml（模块 CI 配置；类别 config；类型 configuration）：添加 cuda/test_cuda_compatibility_path.py 到 Platform Tests (CUDA) job。

关键符号：未识别

关键源码片段

tests/kernels/test_fused_minimax_m3_qknorm_rope_kv_insert.py

测试代码修改：放宽 tolerance 并添加注释说明数值差异原因，是 PR 中唯一的测试源代码变更。

```
# The fused kernel keeps an fp32 intermediate across norm->rope, while the
# reference materializes bf16 after the norm (the unfused boundary), so
# rounding-boundary elements can differ by ~1 bf16 ulp.
```

```
torch.testing.assert_close(q_out, q_ref, rtol=2e-2, atol=2e-2)
torch.testing.assert_close(k_out, k_ref, rtol=2e-2, atol=2e-2)
# V is untouched (no norm/rope applied).
torch.testing.assert_close(v_out, v_in, rtol=0, atol=0)
```

tests/v1/kv_connector/nixl_integration/run_edge_case_test.sh

Shell 脚本修改：允许通过环境变量覆盖 GPU ID，并添加 --max-model-len 参数，使脚本可在 CI 环境中灵活运行。

```
#!/bin/bash
set -xe

KV_BUFFER_DEVICE="cuda" # Default to cuda
# GPU IDs 现在可通过环境变量覆盖，用于 CI 环境（默认值保留 4/5 兼容原 8-GPU 假设）
PREFILL_GPU_ID="${PREFILL_GPU_ID:-4}"
DECODE_GPU_ID="${DECODE_GPU_ID:-5}"
# 解析命令行参数（略）
...
# vllm serve 命令中新增 --max-model-len 8192 以限制模型长度
BASE_CMD="CUDA_VISIBLE_DEVICES=$PREFILL_GPU_ID ... --max-model-len 8192 --kv-transfer-config '$KV_CONFIG'"
```

评论区精华

Review 中主要讨论：

- AndreasKaratzas建议同时在 test-amd.yaml 中添加这些新测试文件，但为了加速合并，当前变更已经 LGTM。作者 njhill 承诺将另开 PR 处理 AMD 镜像问题。
- njhill在 kernels.yaml 评论中说明将单独开 PR 修复 / 删除当前标记为 broke 的测试排除项。
- AMD 测试镜像适配 (testing)：作者回复将另开 PR 处理 AMD 镜像问题，当前变更先合并以加速覆盖。
- Kernels 排除项清理计划 (other)：同意当前 PR 不处理，后续跟进。

风险与影响

- 风险：风险较低，但需关注：
  1. CI 时间增加：新增多个 job 和已有 job 中增加测试用例，可能延长 CI 总时长。作者已验证所有测试在当前 main 上通过，并适当调整了超时时间（Batch Invariance 从 35 分钟增加到 45 分钟）。
  2. Tolerance 放宽：test_fused_minimax_m3_qknorm_rope_kv_insert.py 中将 rtol/atol 从 1e-2 放宽到 2e-2，虽然给出了合理理由，但可能掩盖未来精度回归。但该测试仅用于核函数验证，风险可控。
  3. NIXL edge case 测试依赖 GPU：使用 2 个 GPU 的设备，需确保 CI 有可用资源。
- 影响：影响范围：
  - 用户：无直接影响。

- CI 系统：新增约 2 个 CI job（Kernels Root Misc 和 NixlConnector PD edge case），并增强了多个已有 job 的测试覆盖。总 CI 时间估计增加 10-15%。
- 团队：提升测试覆盖率，降低未运行测试文件的回归风险。为未来在 tests/kernels/ 根目录添加新测试文件提供了默认 CI 覆盖。
- 风险标记：测试覆盖补全，CI 时间增加，tolerance 放宽

关联脉络

- PR #49423 [Bugfix] Fix kernel test failures: 提交日志提到 #49423 修复了某些内核测试，使得 kernels-root 中的部分测试可以重新启用。
- PR #49427 [Bugfix] Fix kernel test failures: 同上，提交日志提到 #49427 修复了内核测试。
- PR #16799 [Kernel] Categorize kernels tests into subdirectories: PR body 提到 kernels/ 子目录分类后，根目录不再被收集，因此需要新增 catch-all job。
- PR #39650 [Bugfix] Fix silent weight corruption in adapters: test_adapters.py 是针对 #39650 修复的回归测试，本 PR 将其接入 CI。