PR #37657 完整报告

vllm-project/vllm

[CI][PD] Add Hybrid SSM integration tests to CI

合并时间：2026-03-23 23:58

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/37657

执行摘要

本 PR 在 vLLM 仓库的 CI 中添加了 Hybrid SSM NixlConnector 的集成测试，通过更新 Buildkite 配置和测试脚本，扩展了 PD integration 的测试覆盖。变更简单直接，风险低，主要影响内部 CI 流程，无用户端影响。

功能与动机

动机是扩展 PD integration coverage，通过在 CI 中运行 Hybrid SSM 测试来验证 kv_connector 在混合 SSM 场景下的正确性。PR body 中明确表示："expand PD integration coverage by running these tests on CI." 这解决了测试覆盖不足的问题，确保新模型架构的稳定性。

实现拆解

关键改动按模块拆解如下：

- CI pipeline 模块：在 .buildkite/test_areas/distributed.yaml 中添加新步骤，配置为运行 Hybrid SSM NixlConnector PD accuracy tests，使用 4 GPUs。
- 测试配置模块：修改 tests/v1/kv_connector/nixl_integration/config_sweep_accuracy_test.sh 中的 hybrid_ssm_configs，将模型从 NVIDIA-Nemotron-3-Nano-30B-A3B-FP8 改为 ibm-granite/granite-4.0-h-tiny，以避免模型过大问题。
- 测试验证模块：在 tests/v1/kv_connector/nixl_integration/test_accuracy.py 中添加新模型的精度阈值 "ibm-granite/granite-4.0-h-tiny": 0.80，确保测试准确性。

评论区精华

review 中仅有 gemini-code-assist[bot] 的正面评论，指出变更直当且配置正确。例如，bot 提到："The configuration is consistent with existing test jobs in the file." DarkLight1337 批准，无争议讨论，表明变更被团队认可。

风险与影响

风险具体分析：

- 技术风险：低。CI 配置变更可能导致测试失败，但已更新阈值适配新模型；无核心代码变更，回归风险有限。

- 影响范围：对用户无直接影响；对系统增加 CI 运行时间，但资源消耗轻微；对团队提升测试覆盖，有助于早期发现 kv_connector 问题。

关联脉络

与近期历史 PR 的关联揭示测试演进方向：

- PR #37816（更新 LoRA 测试）和 PR #37834（测试目录重构）都是测试基础设施改进，与本 PR 共同反映仓库对测试覆盖和组织的持续重视。这些 PR 协同增强了整体测试能力，特别是在模型和连接器场景下。