PR #44174 完整报告

vllm-project/vllm

[CI] Align PD tests to HMA on by default

合并时间：2026-06-04 00:04

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44174

执行摘要

- 一句话：CI 测试移除显式 HMA 标记，对齐默认启用
- 推荐动作：可安全合并。建议后续关注 HMA 功能演进，确保 CI 持续对齐。

功能与动机

HMA 已通过 PR#41847 完成实验阶段并转为默认启用，CI 测试需要跟进，测试自动启用路径而非显式传参，避免遗漏回归。

实现拆解

1. 移除环境变量检查: 在 config_sweep_accuracy_test.sh, run_accuracy_test.sh, spec_decode_acceptance_test.sh 中删除 ENABLE_HMA_FLAG 检查及 ENABLE_HMA_VAR 构造。
2. 从测试配置中删除显式标记: 所有 hybrid_ssm_configs, sw_attn_configs, mtp_config 中的 ENABLE_HMA_FLAG=1 字段被移除。
3. 删除启动命令中的 HMA 参数: run_accuracy_test.sh 和 spec_decode_acceptance_test.sh 中不再拼接 --no-disable-hybrid-kv-cache-manager。
4. 简化测试开关逻辑: config_sweep_accuracy_test.sh 中移除了根据 ENABLE_HMA_FLAG 追加标志的循环。
5. 清理单元测试: test_nixl_connector_hma.py 中移除了强制关闭 HMA 的 disable_hybrid_kv_cache_manager=False 参数，依赖默认行为。

关键文件：

- tests/v1/kv_connector/nixl_integration/config_sweep_accuracy_test.sh（模块 CI 扫配置；类别 test；类型 test-coverage）：核心 sweep 脚本，移除了所有 ENABLE_HMA_FLAG 相关配置和条件逻辑，对齐默认行为。
- tests/v1/kv_connector/nixl_integration/run_accuracy_test.sh（模块 精度测试；类别 test；类型 test-coverage）：删除了 ENABLE_HMA_VAR 变量及其在启动命令中的拼接，简化流程。
- tests/v1/kv_connector/nixl_integration/spec_decode_acceptance_test.sh（模块 推测解码测试；类别 test；类型 test-coverage）：删除文档注释中关于 ENABLE_HMA_FLAG 的描述，移除 HMA 变量和启动参数。

- tests/v1/kv_connector/nixl_integration/config_sweep_spec_decode_test.sh（模块 推测解码扫配置；类别 test；类型 test-coverage）：从 MTP 测试配置中移除 ENABLE_HMA_FLAG=1。
- tests/v1/kv_connector/unit/test_nixl_connector_hma.py（模块 HMA 单元测试；类别 test；类型 test-coverage）：移除单元测试中显式关闭 HMA 的参数，依赖默认值。

关键符号：未识别

关键源码片段

tests/v1/kv_connector/nixl_integration/config_sweep_accuracy_test.sh

核心 sweep 脚本，移除了所有 ENABLE_HMA_FLAG 相关配置和条件逻辑，对齐默认行为。

```
# 原配置包含 ENABLE_HMA_FLAG=1，现移除，因为 HMA 默认启用
hybrid_ssm_configs=(
  "VLLM_SSM_CONV_STATE_LAYOUT=DS GPU_MEMORY_UTILIZATION=0.8 MODEL_NAMES=
  ibm-granite/granite-4.0-h-tiny VLLM_SERVE_EXTRA_ARGS=--max-model-len,8192,--trust-
  remote-code"
  # ...
)
sw_attn_configs=(
  "GPU_MEMORY_UTILIZATION=0.8 MODEL_NAMES=google/gemma-3-4b-it VLLM_SERVE_
  EXTRA_ARGS=--max-model-len,8192"
  # ...
)
# ENABLE_HMA_FLAG 条件分支被整体移除
```

tests/v1/kv_connector/nixl_integration/run_accuracy_test.sh

删除了 ENABLE_HMA_VAR 变量及其在启动命令中的拼接，简化流程。

```
# 删除以下代码块（约 16 行）
# ENABLE_HMA_VAR=""
# if [[ -n "${ENABLE_HMA_FLAG:-}" ]]; then
# ENABLE_HMA_VAR="--no-disable-hybrid-kv-cache-manager"
# fi
# ... 以及在 BASE_CMD 后添加 $ENABLE_HMA_VAR 的部分
# 现在不再需要显式传递 HMA 参数
```

评论区精华

无实质性讨论，仅 Mergify 机器人自动提示 pre-commit 失败。

- 暂无高价值评论线程

风险与影响

- 风险：极低。变更为纯测试清理，未涉及产品代码。若未来 HMA 默认值再次变更，需同步更新这些测试。当前对齐默认行为，更贴近实际部署。

- 影响：仅影响 CI 中 Nixl 连接器的 PD 测试路径。测试不再覆盖显式开启 / 关闭 HMA 的场景，但默认开启已是主流。无用户可见影响。
- 风险标记：仅测试变更，低风险

关联脉络

- PR #41847 [KVCache] Make HMA default-on (opt-out): 此 PR 是 #41847 的后续，对齐 CI 测试到新的默认行为。