PR #40623 完整报告

vllm-project/vllm

[CI] Split disaggregated tests into own test-area

合并时间：2026-04-23 23:20

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40623

执行摘要

- 一句话：拆分 disaggregated CI 测试到独立测试区域
- 推荐动作：建议在合并后立即跟进一个修复 PR：添加设置 SW_ATTN=1 的 CI 步骤，或将 gemma-3 配置移回 tp_configs 并增加 FlashInfer 兼容性检查。同时考虑采纳 gemini-code-assist 的依赖文件建议。

功能与动机

PR 发起人指出 PD disaggregation 代码和相关性在过去一年中已增长到需要独立测试框架的程度。参考了此前 PR #36945 的做法，使分布式任务更小、更独立运行，同时为 PD 设置分配更多预算（因为这些测试本来就成本高昂）。另外补充了之前缺失的 FlashInfer 评估运行，这对 Blackwell+ 开发至关重要。

实现拆解

1. 创建新测试区域文件：在 .buildkite/test_areas/ 下新建 disaggregated.yaml，包含所有原 distributed.yaml 中与 PD disaggregation 相关的测试步骤（共 8 个步骤），包括 NixlConnector 精度、DP/EP、CrossLayer、Hybrid SSM、MultiConnector、Spec Decode 等配置，并新增一个 Distributed FlashInfer NixlConnector PD accuracy 步骤。
2. 从 distributed.yaml 移除旧步骤：删除 distributed.yaml 中对应的 7 个测试步骤（约 85 行），替换为一个新的 Pipeline + Context Parallelism 步骤和一个 RayExecutorV2 步骤，保持文件结构。
3. 调整 gemma-3 测试配置：在 tests/v1/kv_connector/nixl_integration/config_sweep_accuracy_test.sh 中，将 gemma-3 滑动窗口测试从 tp_configs 数组移到新创建的 sw_attn_configs 数组，并添加注释说明 gemma3 不兼容 FlashInfer。这样确保了 gemma-3 只在 SW_ATTN 环境变量设置时运行（如 CI 中），而不是在默认路径下。

关键文件：

- .buildkite/test_areas/disaggregated.yaml（模块 测试配置；类别 config；类型 configuration）：新创建的 disaggregated 测试区域文件，包含了所有 PD disaggregation 相关的 CI 步骤，是本次 PR 的核心变更。
- .buildkite/test_areas/distributed.yaml（模块 测试配置；类别 config；类型 configuration）：原始 distributed.yaml 中移除了 PD disaggregation 相关的 7 个测试步骤（约 85 行），并新增了 Pipeline+Context Parallelism 和 RayExecutorV2 步骤。

- tests/v1/kv_connector/nixl_integration/config_sweep_accuracy_test.sh（模块 测试脚本；类别 test；类型 test-coverage）：调整了 gemma-3 滑动窗口测试的配置归属，从 tp_configs 移到 sw_attn_configs，并添加了兼容性注释。

关键符号：未识别

关键源码片段

tests/v1/kv_connector/nixl_integration/config_sweep_accuracy_test.sh

调整了 gemma-3 滑动窗口测试的配置归属，从 tp_configs 移到 sw_attn_configs，并添加了兼容性注释。

```
# config_sweep_accuracy_test.sh 部分片段
# 原 tp_configs 中删除了 gemma-3 项（第 15 行被移除）
tp_configs=(
  "GPU_MEMORY_UTILIZATION=0.6 PREFILLER_TP_SIZE=2 DECODER_TP_SIZE=2"
  ...
  # 移除了 "GPU_MEMORY_UTILIZATION=0.8 MODEL_NAMES=google/gemma-3-4b-it VLLM_
  SERVE_EXTRA_ARGS=--max-model-len,8192" # SW model
)
# 新增 sw_attn_configs 数组，包含 gemma-3 测试及 HMA 变体
sw_attn_configs=(
  # NOTE: gemma3 does not work with FlashInfer
  "GPU_MEMORY_UTILIZATION=0.8 MODEL_NAMES=google/gemma-3-4b-it VLLM_SERVE_
  EXTRA_ARGS=--max-model-len,8192" # SW model
  "ENABLE_HMA_FLAG=1 GPU_MEMORY_UTILIZATION=0.8 MODEL_NAMES=google/gemma-3-
  4b-it PREFILLER_TP_SIZE=1 DECODER_TP_SIZE=2 VLLM_SERVE_EXTRA_ARGS=--max-model-
  len,8192"
  "ENABLE_HMA_FLAG=1 GPU_MEMORY_UTILIZATION=0.8 MODEL_NAMES=google/gemma-3-
  4b-it PREFILLER_TP_SIZE=2 DECODER_TP_SIZE=1 VLLM_SERVE_EXTRA_ARGS=--max-model-
  len,8192"
)
```

评论区精华

1. gemini-code-assist 机器人提出的改进建议（未解决）：
   - 建议为所有测试步骤添加 vllm/v1/worker/kv_connector_model_runner_mixin.py 到 source_file_dependencies，以确保模型运行器变更时能触发这些测试。
   - 建议为 FlashInfer 测试步骤指定 device: h100 以运行在适当硬件上。
   - 指出 CROSS_LAYERS_BLOCKS=True 与其他步骤使用 1 不一致，建议统一改为 CROSS_LAYERS_BLOCKS=1。
   - 这些建议未被采纳或回复，但 PR 仍被批准合并。
2. orozery 的 Issue 评论（关键问题）：指出 gemma-3 配置被从 tp_configs 移到了 sw_attn_configs，但 CI 的 disaggregated.yaml 中没有设置 SW_ATTN=1 的步骤，导致 gemma-3 测试实际上不再在 CI 中运行。这是一个回归。建议要么添加 SW_ATTN=1 的步骤，要么保持 gemma-3 在 tp_configs 中并只在不使用 FlashInfer 时排除。

- gemma-3 测试配置移动导致 CI 覆盖丢失 (correctness): 此问题未在 PR 中得到解决，但 PR 仍被批准合并。这是明显的回归风险。
- source_file_dependencies 不完整 (other): 建议未被采纳，但已被记录。
- FlashInfer 测试未指定目标设备 (other): 未采纳，风险较低但可能影响测试覆盖率。
- 环境变量值不一致 (style): 未采纳，但建议改为 1。

风险与影响

- 风险:
  1. gemma-3 测试可能完全丢失：如 orozery 所指出，将 gemma-3 移到 sw_attn_configs 后，由于 CI 中没有设置 SW_ATTN 环境变量的步骤，gemma-3 测试将不再自动运行。这可能导致滑动窗口注意力相关的回归无法被捕获。
  2. 依赖文件列表可能不完整：gemini-code-assist 指出 source_file_dependencies 遗漏了关键的 kv_connector_model_runner_mixin.py，可能导致部分变更无法触发相关测试。
  3. 环境变量值不统一：CROSS_LAYERS_BLOCKS=True 与其他步骤的 =1 格式不一致，可能对某些解析逻辑造成问题。
- 影响:
  1. 对用户：无直接影响，纯 CI 重构。
  2. 对系统：CI 配置更清晰，disaggregated 测试独立运行，缩减了 distributed 测试的总执行时间，同时新增了 FlashInfer 专用测试步骤，提升了对 Blackwell+ 硬件的覆盖。
  3. 对团队：需要维护两个测试区域文件，但提高了 CI 的可扩展性和可维护性。gemma-3 测试的遗漏需尽快修复。 - 风险标记：gemma-3 测试完全丢失（CI 覆盖回归），source_file_dependencies 不完整，环境变量值不统一

关联脉络

- PR #36945 类似拆分的前例：PR body 中引用该 PR 作为拆分测试区域的先例。