PR #40168 完整报告

vllm-project/vllm

[CI][EPLB] Add Async EPLB end-to-end integration test to CI

合并时间：2026-04-20 22:22

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/40168

执行摘要

- 一句话：添加异步 EPLB 端到端集成测试到 CI，验证 Qwen3-30B 模型的准确性。
- 推荐动作：该 PR 主要面向测试工程师和 EPLB 开发者，展示了如何配置和运行端到端集成测试。建议关注测试脚本中的 EPLB 参数设置（如 window_size 和 step_interval），以理解异步行为验证方式。

功能与动机

根据 PR body 描述，目的是运行 Qwen3-30B-A3B-FP8 模型通过 lm_eval，同时启用异步 EPLB 并定期执行，以验证功能准确性，测试耗时约 4 分钟。

实现拆解

1. 新增测试脚本：创建文件 .buildkite/scripts/scheduled_integration_test/qwen30b_a3b_fp8_dp4_async_eplb.sh，编写 Shell 脚本启动 vLLM 服务器，配置数据并行大小 4、启用专家并行和异步 EPLB（参数为 --eplb-config '{"window_size":20, "step_interval":100, "use_async":true}'），并运行 GSM8K 准确性评估。
2. 修改 CI 配置：更新文件 .buildkite/test_areas/e2e_integration.yaml，在 CI 管道中添加新测试步骤，指定标签为“Qwen3-30B-A3B-FP8 DP4 Async EPLB Accuracy”，使用 H100 设备、4 个 GPU，并调用上述脚本。
3. 测试配套：脚本包含服务器启动、健康检查、清理逻辑和准确性断言，确保测试可重复且失败时能正确终止。

关键文件：

- .buildkite/scripts/scheduled_integration_test/qwen30b_a3b_fp8_dp4_async_eplb.sh（模块 集成测试；类别 test；类型 test-coverage）：新增的核心测试脚本，包含服务器启动、EPLB 配置和准确性评估逻辑，是实现测试功能的关键文件。
- .buildkite/test_areas/e2e_integration.yaml（模块 CI 配置；类别 config；类型 configuration）：修改的 CI 配置文件，添加了新测试任务到端到端集成测试区域，是测试在 CI 中执行的入口。

关键符号：未识别

关键源码片段

.buildkite/scripts/scheduled_integration_test/qwen30b_a3b_fp8_dp4_async_eplb.sh

新增的核心测试脚本，包含服务器启动、EPLB 配置和准确性评估逻辑，是实现测试功能的关键文件。

```
#!/usr/bin/env bash
set -euxo pipefail

# 参数：[THRESHOLD] [NUM_QUESTIONS] [START_PORT]
THRESHOLD=${1:-0.8} # 准确性阈值，默认 0.8
NUM_Q=${2:-1319} # 问题数量，默认 1319
PORT=${3:-8050} # 服务器端口，默认 8050
OUT_DIR=${OUT_DIR:-/tmp/vllm-scheduled}
mkdir -p "${OUT_DIR}" # 创建输出目录

# 等待服务器启动函数，超时 600 秒
wait_for_server() {
  local port=$1
  timeout 600 bash -c '
    until curl -sf "http://127.0.0.1:'"$port"'/health" > /dev/null; do
      sleep 1
    done'
}

MODEL="Qwen/Qwen3-30B-A3B-FP8" # 测试使用的模型
BACK="allgather_reducescatter" # 全到全后端类型

# 清理函数，确保测试后停止服务器进程
cleanup() {
  if [[ -n "${SERVER_PID:-}" ]] && kill -0 "${SERVER_PID}" 2>/dev/null; then
    kill "${SERVER_PID}" 2>/dev/null || true
    for _ in {1..20}; do
      kill -0 "${SERVER_PID}" 2>/dev/null || break
      sleep 0.5
    done
    kill -9 "${SERVER_PID}" 2>/dev/null || true
  fi
}
trap cleanup EXIT # 注册清理函数，在脚本退出时执行

# 启动 vLLM 服务器，启用异步 EPLB 配置
VLLM_DEEP_GEMM_WARMUP=skip \
vllm serve "$MODEL" \
--enforce-eager \ # 强制 eager 模式，避免图优化干扰
--data-parallel-size 4 \ # 数据并行大小为 4
--enable-expert-parallel \ # 启用专家并行
--enable-eplb \ # 启用 EPLB 功能
--all2all-backend "$BACK" \ # 设置全到全后端（review 中确认参数正确）
```

```
--eplb-config '{"window_size":20, "step_interval":100, "use_async":true}' \  # EPLB
配置，窗口大小 20，步间隔 100，启用异步
--trust-remote-code \  # 信任远程代码，用于加载模型
--max-model-len 2048 \  # 最大模型长度 2048
--port "$PORT" & # 指定端口并后台运行
SERVER_PID=$! # 记录服务器进程 ID
wait_for_server "$PORT" # 等待服务器健康检查通过

# 运行准确性评估，使用 GSM8K 数据集
TAG=$(echo "$MODEL" | tr '/: \\n' '_____') # 生成模型名称标签，用于文件名
OUT="${OUT_DIR}/${TAG}_${BACK}.json" # 输出结果文件路径
python3 tests/evals/gsm8k/gsm8k_eval.py --host http://127.0.0.1 --port "$PORT" --num-
questions "${NUM_Q}" --save-results "${OUT}" # 调用评估脚本（host 参数格式正确）

# 检查准确性是否达到阈值，失败则抛出断言错误
python3 - <<PY
import json; acc=json.load(open('${OUT}'))['accuracy']
print(f"${MODEL} ${BACK}: accuracy {acc:.3f}")
assert acc >= ${THRESHOLD}, f"${MODEL} ${BACK} accuracy {acc}"
PY
```

评论区精华

review 中主要讨论了测试脚本的参数正确性：

- gemini-code-assist[bot] 错误地建议将 --all2all-backend 改为 --moe-all2all-backend，并将 --host 参数从 http://127.0.0.1 改为 127.0.0.1，声称这会避免错误。
- 作者 SageMoore 引用源码（vllm/engine/arg_utils.py）和实际测试结果反驳，指出原参数正确，修改后反而会导致失败。
- tlrmchlsmth 支持作者，并询问是否降低 step_interval 以增加异步 EPLB 运行次数，作者解释手动设置可提高测试覆盖，避免准确性漏检。
- 结论：原实现无误，参数配置符合 vLLM 接口，无需修改。
- 测试脚本参数正确性验证 (correctness): 原实现正确，参数配置符合 vLLM 实际接口，无需修改；step_interval 保持 100 以增加异步 EPLB 运行次数，提高测试可靠性。

风险与影响

- 风险：
  - 回归风险：测试脚本参数依赖 vLLM 命令行接口，若未来接口变更（如 --all2all-backend 重命名），可能导致 CI 失败。但 review 中已验证当前正确性。
  - 性能风险：测试使用 4 个 H100 GPU，可能增加 CI 资源消耗和执行时间（约 4 分钟），但 PR body 已说明并标记为 optional，影响可控。
  - 兼容性风险：脚本硬编码模型名称和参数，若模型或 EPLB 配置更新，需同步调整测试阈值或设置。
- 影响：

- 对系统影响：增强 EPLB 特性的集成测试覆盖，有助于早期发现功能或准确性回归，提升系统稳定性。
- 对用户影响：无直接影响，但间接确保 EPLB 在生产环境中的可靠性。
- 对团队影响：为测试和开发团队提供标准化的 EPLB 验证流程，简化后续测试维护。
- 风险标记：测试配置依赖正确性，CI 执行时间增加

关联脉络

- PR #36276 [EPLB] Add nixl-based eplb communicator: 同为 EPLB 特性开发，该 PR 添加了 EPLB 通信器基础功能，本 PR 的集成测试用于验证其端到端正确性。
- PR #39616 [ROCm][Feature] Enable AITER MLA attention backend to work with Eagle3 speculative decoding on ROCm: 涉及 EPLB 与推测解码的集成，本 PR 的测试可间接验证类似复杂场景的稳定性。