

# PR #47480 完整报告

vllm-project/vllm

[ROCm][CI] Adding qwen3 dp4 eplb

合并时间: 2026-07-03 08:58

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47480>

## 执行摘要

- 一句话: ROCm CI 新增 Qwen3 DP4 异步 EPLB 测试
- 推荐动作: 该 PR 为常规 CI 覆盖补充, 无复杂技术决策, 无需精读。

## 功能与动机

PR body 说明: Adds the Qwen3-30B-A3B FP8 DP4 async EPLB accuracy job to the ROCm AMD CI pipeline on the 4-GPU MI300 pool. The script now disables ROCm MoE padding when ROCm is detected, avoiding the EPLB failure mode seen on AMD.

## 实现拆解

1. 配置 CI job: 在 `.buildkite/test-amd.yaml` 中新增一个 CI job, 标签为 "Qwen3-30B-A3B-FP8 DP4 Async EPLB Accuracy (4xH100-4xMI300)", 使用 `mi300_4 agent pool`, 4 GPU, 超时 180 分钟, 设置为可选 (`optional: true`)。命令执行新建的测试脚本。
2. 新增测试脚本: 创建 `.buildkite/scripts/scheduled_integration_test/qwen30b_a3b_fp8_dp4_async_eplb.sh`, 基于已有的 `block sync EPLB` 测试脚本改编, 核心变化:
  - 添加 ROCm 检测逻辑 (通过 `rocm-smi` 命令、`/opt/rocm` 目录或 `ROCM_PATH` 环境变量判断), 若检测到 ROCm 则设置 `VLLM_ROCM_MOE_PADDING=0` 环境变量, 禁用 MoE padding 以避免 EPLB 失败。
  - 使用 DP4 (`data-parallel-size 4`) 和异步 EPLB 配置 (`--eplb-config '{"window_size":20, "step_interval":100, "use_async":true}'`)。
3. 测试内容: 启动 `vllm server` 后, 运行 `GSM8K` 评估脚本 `tests/evals/gsm8k/gsm8k_eval.py`, 默认 1319 道题, 阈值 0.8。

关键文件:

- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 `config`; 类型 `configuration`): 注册新的 CI job, 定义标签、硬件要求、依赖文件和执行命令。
- `.buildkite/scripts/scheduled_integration_test/qwen30b_a3b_fp8_dp4_async_eplb.sh` (模块 测试脚本; 类别 `test`; 类型 `test-coverage`): 新增的测试脚本, 包含 ROCm 特定补丁 (禁用 MoE padding) 和 DP4 异步 EPLB 启动逻辑。

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](#)

新增的测试脚本，包含 ROCm 特定补丁（禁用 MoE padding）和 DP4 异步 EPLB 启动逻辑。

```
# Script: qwen30b_a3b_fp8_dp4_async_eplb.sh
set -euxo pipefail

# 检查是否运行在 ROCm 环境（通过多种方式检测）
if command -v rocm-smi &> /dev/null || [[ -d /opt/rocm ]] || [[ -n "${ROCM_PATH:-}" ]]; then
    # ROCm 上 MoE padding 会导致 EPLB 失败，故禁用它
    export VLLM_ROCM_MOE_PADDING=0
fi

# 启动 vllm server，使用 DP4 和异步 EPLB
VLLM_DEEP_GEMM_WARMUP=skip \
vllm serve "Qwen/Qwen3-30B-A3B-FP8" \
    --enforce-eager \
    --data-parallel-size 4 \
    --enable-expert-parallel \
    --enable-eplb \
    --all2all-backend "allgather_reducescatter" \
    --eplb-config '{"window_size":20, "step_interval":100, "use_async":true}' \
    --trust-remote-code \
    --max-model-len 2048 \
    --port "$PORT" &
```

## 评论区精华

无 review 讨论，仅 khluu 批准且无评论。

- 暂无高价值评论线程

## 风险与影响

- 风险：风险极低。本次变更仅为 CI 配置和测试脚本的新增，不涉及生产代码。新增的测试为 optional，不会阻塞流水线。唯一风险是测试脚本中的 ROCm 检测逻辑可能误判，但检测条件较为保守，且禁用 MoE padding 仅影响 ROCm 环境，不干扰其他平台。
- 影响：仅影响 ROCm CI 流水线，在 4-GPU MI300 pool 上新增一个可选的精度测试任务。对用户无影响，对系统无副作用。
- 风险标记：暂无

## 关联脉络

- 暂无明显关联 PR