

PR #50110 完整报告

vllm-project/vllm

[Test] Make EPD correctness tests configurable for XPU

合并时间: 2026-07-29 07:47

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/50110>

执行摘要

- 一句话: EPD 正确性测试 XPU 可配置化
- 推荐动作: 该 PR 是典型的测试基础设施改进, 适合快速合并。建议在合并后, XPU 团队实际运行一次该测试以验证兼容性。对于关注测试可移植性和跨平台 CI 的工程师, 值得精读以了解如何通过环境变量和参数化适配不同硬件。

功能与动机

PR body 明确指出目的为: “Add XPU device affinity support to the EPD correctness test. Make `gpu-memory-utilization max-num-seqs max-model-len` configurable.” 该 PR 旨在让 EPD 正确性测试能够在 Intel XPU 上运行, 通过引入设备平台变量和可配置参数, 避免硬编码 CUDA 环境变量和 GPU 参数。

实现拆解

1. 在 shell 测试脚本中引入设备平台和亲和性环境变量 (`run_epd_correctness_test.sh`): 添加 `DEVICE_PLATFORM` 默认为 `cuda`, 并根据其值自动设置 `DEVICE_AFFINITY_ENV`: `xpu` 时使用 `ZE_AFFINITY_MASK`, 否则使用 `CUDA_VISIBLE_DEVICES`。所有 vLLM 实例启动命令从 `CUDA_VISIBLE_DEVICES=...` 改为 `env "$DEVICE_AFFINITY_ENV=..."`, 从而兼容 XPU。
2. 在 shell 测试脚本中添加可配置参数 (`run_epd_correctness_test.sh`): 新增 `MAX_MODEL_LEN` (默认 10240)、`GPU_MEMORY_UTILIZATION` (默认 0.7)、`MAX_NUM_SEQS` (默认 128) 环境变量, 并应用到 baseline 和 EPD 的每个 vLLM serve 启动参数中。同时将默认的 `GPU_MEMORY_UTILIZATION` 从 0.7 改为 0.9 以提高 baseline 的显存利用率。
3. 在 shell 测试脚本中优化多模态提示处理 (`run_epd_correctness_test.sh`): 引入 `USE_TWO_IMAGE_PROMPT` 变量 (默认 1), 允许跳过需要两张图像的提示, 以适应 XPU 显存限制。将原有的简单 `MM_FLAG` 变量改为 `TEST_FLAGS` 数组, 动态添加 `--use_mm_prompts` 和 `--skip_two_image_prompt` 参数。
4. 在 Python 测试程序中添加 `--skip_two_image_prompt` 参数 (`test_epd_correctness.py`): 当指定该参数时, 在 `use_mm_prompts` 模式下仅使用 `SAMPLE_PROMPTS_MM` 的第一个提示 (单张图像), 从而降低 XPU 上的显存需求。

关键文件:

- tests/v1/ec_connector/integration/run_epd_correctness_test.sh (模块 测试脚本; 类别 test; 类型 test-coverage) : 核心变更文件, 引入了设备平台亲和性选择、可配置参数和双图像提示跳过逻辑, 是 EPD 测试脚本的主体改造
- tests/v1/ec_connector/integration/test_epd_correctness.py (模块 测试程序; 类别 test; 类型 test-coverage) : 新增 --skip_two_image_prompt 参数, 用于在 XPU 显存受限时跳过双图像 prompt

关键符号: 未识别

关键源码片段

tests/v1/ec_connector/integration/run_epd_correctness_test.sh

核心变更文件, 引入了设备平台亲和性选择、可配置参数和双图像提示跳过逻辑, 是 EPD 测试脚本的主体改造

```
# 新增可配置参数 (默认值兼容原 CUDA 行为)
MAX_MODEL_LEN="${MAX_MODEL_LEN:-10240}"
GPU_MEMORY_UTILIZATION="${GPU_MEMORY_UTILIZATION:-0.7}"
MAX_NUM_SEQS="${MAX_NUM_SEQS:-128}"

# 多模态提示控制: 允许跳过双图像 prompt
USE_TWO_IMAGE_PROMPT="${USE_TWO_IMAGE_PROMPT:-1}"
TEST_FLAGS=()
if [[ "$USE_MM_PROMPTS" == "1" ]]; then
    TEST_FLAGS+=(--use_mm_prompts)
fi
if [[ "$USE_TWO_IMAGE_PROMPT" != "1" ]]; then
    TEST_FLAGS+=(--skip_two_image_prompt) # 仅用于 XPU 显存受限场景
fi

# 设备平台与亲和性环境变量 (支持 CUDA 和 XPU)
DEVICE_PLATFORM="${DEVICE_PLATFORM:-cuda}"
if [[ -z "${DEVICE_AFFINITY_ENV:-}" ]]; then
    if [[ "${DEVICE_PLATFORM,,}" == "xpu" ]]; then
        DEVICE_AFFINITY_ENV="ZE_AFFINITY_MASK" # Intel XPU
    else
        DEVICE_AFFINITY_ENV="CUDA_VISIBLE_DEVICES" # NVIDIA GPU
    fi
fi

# 启动 baseline 实例, 使用动态设备和参数
run_baseline() {
    # ...
    env "$DEVICE_AFFINITY_ENV=$GPU_SINGLE" vllm serve "$MODEL" \
        --port "$PORT" \
        --max-model-len "$MAX_MODEL_LEN" \
        --enforce-eager \
        --gpu-memory-utilization "$GPU_MEMORY_UTILIZATION" \
```

```

--max-num-seqs "$MAX_NUM_SEQS" \
--allowed-local-media-path "${GIT_ROOT}/tests/v1/ec_connector/integration \
> "$LOG_PATH"/baseline.log 2>&1 &
# ...
# 传递动态 TEST_FLAGS 数组
python "${GIT_ROOT}/tests/v1/ec_connector/integration/test_epd_correctness.py" \
--service_url "http://localhost:$PORT" \
--model_name "$MODEL" \
--mode baseline \
--baseline_file "$BASELINE_FILE" \
"${TEST_FLAGS[@]}"
}

```

tests/v1/ec_connector/integration/test_epd_correctness.py

新增 `--skip_two_image_prompt` 参数，用于在 XPU 显存受限时跳过双图像 prompt

```

# 在参数解析部分新增
parser.add_argument(
    "--skip_two_image_prompt",
    action="store_true",
    help="Skip the two-image multimodal prompt",
)

# 在 prompt 选择逻辑中判断
if args.use_mm_prompts:
    # 如果指定了 --skip_two_image_prompt，则只取第一个 prompt（单张图像）
    test_prompts = (
        SAMPLE_PROMPTS_MM[:1] if args.skip_two_image_prompt else SAMPLE_PROMPTS_MM
    )
    print("Using multimodal prompts")
else:
    test_prompts = SAMPLE_PROMPTS_TEXT
    print("Using text-only prompts for quick testing")

```

评论区精华

本 PR 的 review 讨论较少，主要由作者 @zhenwei-intel 提交后请求审核，审核者 @jikunshang 直接批准。Claude bot 的评论仅说明此 PR 来自 fork 因此自动 review 关闭，需要维护者手动触发。没有公开的实质性讨论或争议。

- 暂无高价值评论线程

风险与影响

- 风险：本 PR 仅修改测试脚本和测试程序，不涉及生产代码，因此回归风险极低。主要风险是：如果 `DEVICE_PLATFORM` 环境变量设置错误（例如拼写错误 `xpu` 的大小写），可能导致设备亲和性变量选择错误，进而使用错误的 GPU 掩码；另外 `GPU_MEMORY_UTILIZATION` 默认值从 0.7 改为 0.9 可能影响 CUDA 环境下的 OOM 测试结果，但测试脚本允许通过环境变量覆盖。

- 影响：影响范围仅限于 EPD 正确性测试的执行方式。对用户（vLLM 库使用者）无影响；对 Intel XPU 团队，测试脚本现在可无缝在 XPU 上运行；对现有 CUDA 测试流程，由于所有新增变量都有与原来一致的默认值（除 GPU_MEMORY_UTILIZATION 改为 0.9），兼容性较好。整体影响程度低。
- 风险标记：测试配置默认值变更（GPU_MEMORY_UTILIZATION 0.7->0.9），依赖环境变量正确设置，XPU 分支未在 CI 中验证

关联脉络

- PR #49903 [Core] Warm up runner-owned Triton kernels before the first request: 同属测试基础设施改进，都涉及测试脚本的配置化调整