# PR #45977 完整报告

vllm-project/vllm

[XPU][CI] Enable shared loader test

合并时间：2026-06-30 19:20

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45977

## 执行摘要

- 一句话：启用 XPU 上 sharded state loader 测试
- 推荐动作：建议合入。该 PR 改动简洁、目的明确，通过极小的测试参数调整和 CI 配置新增，填补了 XPU 平台在分布式模型加载测试上的空白。值得关注的是其设计模式——复用已有的 max_num_seqs=1 策略来应对显存受限平台，而非引入全新逻辑，体现了可维护性。

## 功能与动机

PR body 明确说明："This PR improves stability of test_sharded_state_loader on XPU by forcing max_num_seqs=1 in the XPU test path." 此前测试在 XPU 上可能因显存不足失败，通过限制并发序列数量增强稳定性，并填补 XPU CI 中的测试覆盖空白。

## 实现拆解

1. 修改测试参数适配：在 tests/model_executor/model_loader/test_sharded_state_loader.py 中，将 if current_platform.is_rocm(): 改为 if current_platform.is_rocm() or current_platform.is_xpu():，使 XPU 平台也设置 max_num_seqs=1。这一配置通过 platform_args 传递给 _run_writer 和 _run_generate 进程，限制同一时间处理的序列数量，从而降低显存压力。
2. 新增 Intel CI 配置：新增 .buildkite/intel_jobs/models_distributed_intel.yaml 文件，定义名为 "Distributed Model Tests (2 GPUs)" 的 CI 步骤。步骤配置包括：依赖 image-build-xpu 镜像；使用 Intel GPU 设备，要求 2 张 GPU；设置 VLLM_TEST_DEVICE: "xpu"；通过 source_file_dependencies 关联 sharded_state_loader 源码、模型目录和测试文件；执行命令通过 run-intel-test.sh 运行 pytest -v -s model_executor/model_loader/test_sharded_state_loader.py -m "not slow_test"，过滤掉慢测试，确保 CI 效率。

关键文件：

- tests/model_executor/model_loader/test_sharded_state_loader.py（模块 模型加载测试；类别 test；类型 test-coverage；符号 test_sharded_state_loader）：测试文件，修改条件判断以在 XPU 平台设置 max_num_seqs=1，解决显存不足问题。
- .buildkite/intel_jobs/models_distributed_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：新增 Intel CI 配置，定义分布式模型测试步骤，将测试集成到 XPU CI 流水线。

关键符号：test_sharded_state_loader

## 关键源码片段

### tests/model_executor/model_loader/test_sharded_state_loader.py

测试文件，修改条件判断以在 XPU 平台设置 max_num_seqs=1，解决显存不足问题。

```
# tests/model_executor/model_loader/test_sharded_state_loader.py
@pytest.mark.parametrize("enable_lora", [False, True])
@pytest.mark.parametrize("tp_size", [1, 2])
def test_sharded_state_loader(
    enable_lora, tp_size, num_gpus_available, llama_3p2_1b_files
):
    # ... 前面的跳过逻辑和参数准备 ...
    platform_args = {}
    # 在 ROCm 和 XPU 平台上限制 max_num_seqs=1 以节省显存，避免 OOM
    if current_platform.is_rocm() or current_platform.is_xpu():
        platform_args["max_num_seqs"] = 1

    # 将 platform_args 传递给 writer 和 generate 进程
    with TemporaryDirectory() as output_dir:
        p = ctx.Process(
            target=_run_writer,
            args=(input_dir, output_dir, weights_patterns),
            kwargs=dict(
                tensor_parallel_size=tp_size,
                gpu_memory_utilization=gpu_memory_utilization,
                enforce_eager=True,
                **platform_args, # 关键：platform_args 解包传递
            ),
        )
        p.start()
        p.join()
        # ... generate 进程类似
```

## 评论区精华

Review 中主要围绕 CI 配置细节展开：

- yma11 询问 ZE_AFFINITY_MASK=0,1 是否适用于 model_loader 下的其他测试，暗示环境变量可能过于宽泛。
- jikunshang 明确反对将 ZE_AFFINITY_MASK 写入配置，认为不应在此处硬编码设备亲和性。
- zxd1997066 建议将依赖镜像从 image-build 改为 image-build-xpu（已采纳），确保 CI 使用合适的 Intel XPU 基础镜像。 讨论均集中在基础设施细节，核心测试逻辑无争议，最终获得 approved。
- ZE_AFFINITY_MASK 环境变量的适用性 (design)：最终 CI 配置中移除了 ZE_AFFINITY_MASK，避免硬编码设备亲和性。

- 依赖镜像名称修正 (other): 已采纳建议，最终配置使用 image-build-xpu。

## 风险与影响

- 风险：低风险。变更仅涉及测试行为和 CI 配置：
  - 测试条件中增加 is_xpu() 分支，已有 is_rocm() 先例，语义清晰，不会影响其他平台。
  - CI 配置独立于现有测试流水线，新增步骤不会干扰已有任务。
  - 潜在风险：若 Intel CI 环境（如资源限制）导致测试超时，可能需要调整超时时间或资源分配；但已通过 timeout_in_minutes: 50 和 -m "not slow_test" 做了缓解。
  - 影响：
  - 对用户：无直接用户影响。
  - 对系统：XPU 平台测试覆盖增强，sharded_state_loader 的 CI 验证从无到有，有助于提前发现回归。
  - 对团队：Intel CI 流水线新增一个分布式测试步骤，维护成本低。
  - 风险标记：低影响 - 测试 /CI 变更

## 关联脉络

- PR #46433 [XPU] Optimize XPU worker shutdown logic to prevent resource leak: 同为 Intel XPU 平台的 CI 和稳定性改进，共享相同的测试环境和基础设施关注点。