PR #45140 完整报告

vllm-project/vllm

[Kernel][XPU] Adjust kernel unit tests for XPU

合并时间：2026-06-30 17:57

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/45140

执行摘要

- 一句话：为 XPU 放宽 Mamba SSM 测试精度阈值
- 推荐动作：本 PR 是常规的测试适配变更，改动小且经过 review 确认，建议快速合并。值得关注的是使用 current_platform 统一平台检查的实践，可在其他测试中推广。

功能与动机

XPU 上的 Triton selective_state_update 核函数内部以 fp32 累积，相比纯 PyTorch bf16 参考实现（bf16 内累积）在某些元素上更精确，导致默认 bf16 容差下部分元素不满足 allclose。PR body 指出 ROCm 已因相同原因加倍 atol，本次扩展至 XPU，使测试通过。

实现拆解

1. 修改测试容差逻辑：在 tests/kernels/mamba/test_mamba_ssm.py 中，将 5 处 if torch.version.hip: 条件扩展为 if current_platform.is_rocm() or current_platform.is_xpu():，使 ROCm 和 XPU 均享受 bf16 atol 翻倍的宽松阈值。
2. 平台检查方式优化：应 review 建议，将 torch.version.hip 替换为更规范的 current_platform.is_rocm()，与新增的 current_platform.is_xpu() 保持风格一致。
3. 无其他文件变更：本 PR 仅修改测试文件，不涉及源码、配置或部署。

关键文件：

- tests/kernels/mamba/test_mamba_ssm.py（模块 测试；类别 test；类型 test-coverage；符号 test_selective_state_update, test_selective_state_update_varlen, test_selective_state_update_with_batch_indices, test_selective_state_update_with_num_accepted_tokens）：唯一变更文件，修改 5 处 bf16 atol 条件判断，新增 XPU 支持并统一平台检查方式。

关键符号：test_selective_state_update, test_selective_state_update_varlen, test_selective_state_update_with_batch_indices, test_selective_state_update_with_num_accepted_tokens, test_selective_state_update_varlen_with_num_accepted

关键源码片段

tests/kernels/mamba/test_mamba_ssm.py

唯一变更文件，修改 5 处 bf16 atol 条件判断，新增 XPU 支持并统一平台检查方式。

```
# 在 tests/kernels/mamba/test_mamba_ssm.py 中，
# 多个测试函数都包含类似的 bf16 容差调整逻辑。
# 以下是 test_selective_state_update 中的示例：

if itype == torch.bfloat16:
    rtol, atol = 1e-2, 5e-2
    # 对于 ROCm 和 XPU 平台，放大 atol 以避免因 fp32 累积精度
    # 高于 bf16 参考实现而导致的误报失败。
    if current_platform.is_rocm() or current_platform.is_xpu():
        atol *= 2

# 其余 4 个测试函数（test_selective_state_update_varlen、
# test_selective_state_update_with_batch_indices、
# test_selective_state_update_with_num_accepted_tokens、
# test_selective_state_update_varlen_with_num_accepted）
# 均使用完全相同的条件结构。
```

评论区精华

Review 中 jikunshang 建议将 torch.version.hip or current_platform.is_xpu() 改为 current_platform.is_rocm() or current_platform.is_xpu()，以使用更统一的平台检查 API。adobrzyn 接受并完成修改（commit 778a7474d）。

- 平台检查方式改进 (design): adobrzyn 接受建议并在后续 commit 中修改了所有 5 处条件。

风险与影响

- 风险：风险极低。仅修改测试容差逻辑，且仅针对 bf16 类型放大 atol，不影响 CUDA 或其他平台行为。放大 atol 可能掩盖真实的精度问题，但幅度（2 倍）较小，且 ROCm 已长期使用相同倍数，风险可控。
- 影响：直接影响：XPU 上 Mamba SSM 的 5 个 bf16 测试用例从失败变为通过，消除 CI 误报。间接影响：增强 vLLM 对 Intel XPU 平台的支持完整性，降低开发者适配成本。影响范围局限于测试环境。
- 风险标记：测试容差放宽，仅影响 XPU/ROCm

关联脉络

- PR #46381 [Bugfix][ROCm] Preserve MoE weight padding for unquantized Triton path: 同一作者 adobrzyn 参与的 ROCm 相关 PR，体现对 Intel XPU 和 AMD ROCm 平台测试一贯的适配策略。
- PR #46433 [XPU] Optimize XPU worker shutdown logic to prevent resource leak: 同为 XPU 平台优化 PR，体现当前 vLLM 对 Intel GPU 的持续支持。