PR #43646 完整报告

vllm-project/vllm

[XPU] Fix fused MoE LoRA kernel crash on XPU by using platform-agnos num_compute_units

合并时间：2026-05-26 18:40

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43646

执行摘要

- 一句话：修复 XPU 上 fused MoE LoRA kernel 崩溃
- 推荐动作：值得合并，修复清晰且无副作用。建议精读以确保后续平台无关化改造遵循相同模式。

功能与动机

XPU 设备上运行 fused MoE LoRA 内核时，由于直接调用 torch.cuda.get_device_properties() 而崩溃，报错 AssertionError: Torch not compiled with CUDA enabled。需要改用平台无关的 current_platform.num_compute_units() 来获取计算单元数量，该函数已为各平台提供统一接口。

实现拆解

1. 添加导入：在 vllm/lora/ops/triton_ops/fused_moe_lora_op.py 中增加 from vllm.platforms import current_platform。
2. 替换第一处调用：_run_fused_moe_lora_one_shot 函数中，将 torch.cuda.get_device_properties(device).multi_processor_count 替换为 current_platform.num_compute_units(device.index)。
3. 替换第二处调用：_run_fused_moe_lora_small_batch 函数中，进行完全相同替换。注意传递的是 device.index（整数设备索引）而非 device 对象，以符合 num_compute_units 的整数参数类型要求。

关键文件：

- vllm/lora/ops/triton_ops/fused_moe_lora_op.py（模块 LoRA；类别 infra；类型 infrastructure；符号 _run_fused_moe_lora_one_shot, _run_fused_moe_lora_small_batch）：核心变更文件，修复了 XPU 上 fused MoE LoRA 内核崩溃问题，替换了两处 CUDA 直接调用为平台无关接口。

关键符号：_run_fused_moe_lora_one_shot, _run_fused_moe_lora_small_batch

关键源码片段

vllm/lora/ops/triton_ops/fused_moe_lora_op.py

核心变更文件，修复了 XPU 上 fused MoE LoRA 内核崩溃问题，替换了两处 CUDA 直接调用为平台无关接口。

```
# vllm/lora/ops/triton_ops/fused_moe_lora_op.py

# 新增导入平台无关接口
from vllm.platforms import current_platform

def _run_fused_moe_lora_one_shot(...):
    # ...
    # NPID_FACTOR heuristic: scale N-axis parallelism when base CTA count is
    # short of saturating the SM array. Cap by the cost of redundant shrink.
    # 原代码使用 torch.cuda.get_device_properties(device).multi_processor_count
    # 在 XPU 上因 CUDA 不可用而崩溃，替换为平台无关调用
    sm_count = current_platform.num_compute_units(device.index)
    base_programs = max(M_blocks * num_slices * grid_lora_dim, 1)
    # ...

def _run_fused_moe_lora_small_batch(...):
    # ...
    N_tiles = triton.cdiv(N_per_slice, BLOCK_N)
    pair_slices = M_grid * num_slices
    # 同样替换第二处调用，传递 device.index 以符合 num_compute_units 的整数参数要求
    sm_count = current_platform.num_compute_units(device.index)
    n_tiles_per_program = _pick_small_batch_chunk(pair_slices, N_tiles, sm_count)
    # ...
```

评论区精华

Review 中 gemini-code-assist[bot] 指出 num_compute_units 期望整数设备 ID，若直接传递 torch.device 对象可能在部分平台上引发运行时错误。实际改动中已使用 device.index 传递整数索引，避免了该问题。审核者 jikunshang 和 jeejeelee 均 approves PR。

- num_compute_units 参数类型兼容性 (correctness): 实际代码已使用 device.index 传递整数索引，正确修复。

风险与影响

- 风险：低风险。仅替换两处平台相关的 API 调用为已封装好的平台无关接口，且已在 XPU 平台验证通过。但需确保 current_platform.num_compute_units() 在非 CUDA、非 XPU 平台上行为正确（降级或报错应有明确提示）。
- 影响：影响范围仅限于 fused MoE LoRA kernel 在 XPU 及未来可能的新硬件平台上的运行；对 CUDA 平台无行为变化（num_compute_units 内部会正确返回 CUDA SM 数量）。修复后 XPU 用户可正常使用 fused MoE LoRA 功能。
- 风险标记：平台兼容性依赖

关联脉络

- 暂无明显关联 PR