

PR #44468 完整报告

vllm-project/vllm

[XPU] Fix test_spec_decode_logprobs: use FLASH_ATTN for XPU in GPU_DETERMINISM_KWARGS

合并时间: 2026-06-17 11:07

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44468>

执行摘要

- 一句话: 修复 XPU 上 spec-decode logprobs 测试因注意力后端不匹配导致的数值差异
- 推荐动作: 值得快速审查和合并。变更小且聚焦, 解决了 Intel XPU 上的 CI 测试失败问题。对于维护多后端的跨平台测试, 这个模式 (提取平台特定配置) 是一个很好的实践。

功能与动机

在 Intel XPU 平台上, `test_spec_decode_logprobs` 测试因 logprobs 数值超出 `abs_tol=0.1` 而失败。根本原因与 ROCm 类似: 非确定性浮点归约导致参考模型和 speculative decoding 模型之间的 logprobs 不一致。此 PR 将原有的 ROCm 专用解决方法泛化, 并添加了 XPU 所需的确定性注意力后端配置。

实现拆解

1. 重构确定性配置变量: 在 `tests/v1/sample/test_logprobs.py` 中, 将硬编码的 `ROCM_DETERMINISM_KWARGS` 替换为平台感知的 `GPU_DETERMINISM_KWARGS` 字典, 使用 `if-elif-else` 结构为不同后端分配不同的注意力后端。
2. 添加 XPU 条目: 为 Intel XPU 平台设置 `max_num_seqs=1` 和 `attention_backend="FLASH_ATTN"`, 强制使用 Flash Attention 后端以确保数值确定性。
3. 更新引用点: 将测试函数 `test_spec_decode_logprobs` 中所有使用 `**ROCM_DETERMINISM_KWARGS` 的地方改为 `**GPU_DETERMINISM_KWARGS`。
4. 启用 CI 测试: 在 `.buildkite/intel_jobs/misc_intel.yaml` 的 XPU 测试命令列表中新增 `pytest -v -s v1/sample/test_logprobs.py`, 确保该测试在 Intel CI 中运行。

关键文件:

- `tests/v1/sample/test_logprobs.py` (模块 测试; 类别 `test`; 类型 `test-coverage`): 核心变更文件: 将 ROCm 专用的确定性 kwargs 重构为跨平台变量, 并添加 XPU 的 Flash Attention 后端配置。
- `.buildkite/intel_jobs/misc_intel.yaml` (模块 CI 配置; 类别 `config`; 类型 `configuration`): 在 XPU CI 测试套件中启用 `test_logprobs.py`, 确保该测试在 Intel CI 中运行。

关键符号: 未识别

关键源码片段

tests/v1/sample/test_logprobs.py

核心变更文件：将 ROCm 专用的确定性 kwargs 重构为跨平台变量，并添加 XPU 的 Flash Attention 后端配置。

```
# 之前：仅 ROCm 平台使用 TRITON_ATTN 来确保确定性执行
# 现在：通过 if-elif-else 为不同平台选择确定性注意力后端
if current_platform.is_rocm():
    GPU_DETERMINISM_KWARGS: dict = dict(
        max_num_seqs=1,
        attention_backend="TRITON_ATTN" # ROCm 上使用 TRITON_ATTN 以保证数值确定性
    )
elif current_platform.is_xpu():
    GPU_DETERMINISM_KWARGS = dict(
        max_num_seqs=1,
        attention_backend="FLASH_ATTN" # XPU 上使用 FLASH_ATTN 以保证数值确定性
    )
else:
    GPU_DETERMINISM_KWARGS = {} # 其他平台（如 NVIDIA）不需要特殊配置
```

评论区精华

该 PR 没有 review 评论，仅有一次来自维护者的批准 ("LGTM. thanks for fixing.")，表明变更直接且被接受。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更是局部的，仅修改测试文件和 CI 配置。测试使用 max_num_seqs=1 会降低并行度，但仅用于确定性测试，不影响生产性能。其他平台（如 NVIDIA）保持不变，因为它们使用空字典，不受影响。
- 影响：影响范围有限，仅影响 Intel XPU 平台上的 spec-decode logprobs 测试。变更确保了该测试在 XPU 上通过，提高了 CI 的可靠性和覆盖范围。对生产用户无影响。
- 风险标记：暂无

关联脉络

- PR #45722 [ROCm][CI] Patch conftest to resolve occasional OOMs: 类似的 ROCm 平台 CI 修复，显示了跨平台 CI 稳定性工作的持续模式。
- PR #45720 [Bugfix][ROCm] Fix MiniMax-M3 FP8 KV cache dtype: 同样解决了 ROCm 上的数值一致性问题，与本 PR 的跨平台确定性考虑相关。