PR #49380 完整报告

vllm-project/vllm

[CI][Bugfix] Fix ROCm FP8 KV cache dtype in attention backend test

合并时间：2026-07-22 09:34

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49380

执行摘要

- 一句话：修复 ROCm FP8 KV cache dtype 导致测试 NaN 失败
- 推荐动作：值得快速合并，修复了跨平台测试一致性问题，无潜在风险。

功能与动机

在 ROCm gfx94x / MI300 上，test_causal_backend_correctness[fp8*] 测试失败，报错 '[AttentionBackendEnum.TRITON_ATTN] produced non-finite values'。根本原因是测试硬编码了 torch.float8_e4m3fn，但 backends 运行时通过 current_platform.fp8_dtype() 重新解释字节，该函数在 ROCm 上返回 e4m3fnuz，两种格式的指数偏置和 NaN 编码不同（如 0x80 在 e4m3fn 中是 -0.0，在 e4m3fnuz 中是 NaN）。修复使测试存储 dtype 与运行时一致。

实现拆解

在 tests/v1/attention/test_attention_backends.py 中，将 FP8_KV_CACHE_DTYPES 字典的两个值从硬编码的 torch.float8_e4m3fn 替换为 current_platform.fp8_dtype()，并添加注释说明原因。该函数在 CUDA 上返回 e4m3fn，在 ROCm gfx94x 上返回 e4m3fnuz，从而匹配 backends 运行时的 reinterpret 逻辑。

关键文件：

- tests/v1/attention/test_attention_backends.py（模块 注意力；类别 test；类型 test-coverage）：修复了 FP8 KV cache dtype 硬编码问题，使测试在 ROCm 上正确运行。

关键符号：未识别

关键源码片段

tests/v1/attention/test_attention_backends.py

修复了 FP8 KV cache dtype 硬编码问题，使测试在 ROCm 上正确运行。

```
# Use the platform's preferred FP8 type so the stored cache matches what the
# backends reinterpret at runtime. On ROCm gfx94x this is e4m3fnuz, not e4m3fn;
# storing e4m3fn bytes there would be re-read as fnuz and produce NaNs.
FP8_KV_CACHE_DTYPES = {
    "fp8": current_platform.fp8_dtype(),
    "fp8_e4m3": current_platform.fp8_dtype(),
```

}

评论区精华

无审核评论讨论，仅 AndreasKaratzas 批准。

- 暂无高价值评论线程

风险与影响

- 风险：变更极小（5 行），仅在测试中修改 dtype 值。CUDA 上 function 返回相同 dtype，因此无回归风险。ROCm 上修复了 NaN 问题。风险极低。
- 影响：影响限于 ROCm 平台下 FP8 attention 测试的正确性。CUDA 用户无感知。
- 风险标记：暂无

关联脉络

- PR #49329 [ROCm][CI] Fix order-dependent failure in test_flash_attn_accepts_handled_fp8_variants (MI355)：同为 ROCm attention 测试修复，但问题不同（顺序依赖 vs dtype），本 PR 特意声明非重复。