PR #50190 完整报告

vllm-project/vllm

[ROCm][CI] Stabilize ngram and suffix correctness test

合并时间：2026-07-29 10:29

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/50190

执行摘要

- 一句话：修复 ROCm 上 spec decode 测试的间歇性故障
- 推荐动作：该 PR 是典型的 CI 稳定性修复，技术价值有限但对 ROCm 持续集成很重要。建议合并。

功能与动机

ROCm CI 在 Buildkite build 11367 中，test_ngram_and_suffix_correctness 测试（ngram 案例完成、引擎关机超时后，suffix 案例出现异步 GPU 故障）间歇性失败，需要提升测试稳定性。

实现拆解

1. 新增 fixture disable_vllm_compile_cache_on_rocm：在 tests/v1/e2e/spec_decode/test_spec_decode.py 中定义一个自动使用的 fixture，仅在 ROCm 平台下通过 request.getfixturevalue("disable_vllm_compile_cache") 调用已有的 disable_vllm_compile_cache fixture，避免 vLLM 编译缓存影响。
2. 替换手动 LLM 生命周期为 VllmRunner 上下文管理器：将 test_ngram_and_suffix_correctness 中的手动 LLM 创建、删除、缓存清除和分布式清理替换为 vllm_runner 上下文管理器，利用其内置的 60 秒关机超时、分布式清理和 ROCm 内存等待机制。
3. 调整 import 和装饰器：新增 from vllm.config import CompilationConfig 导入，并给测试函数添加 @pytest.mark.usefixtures("disable_vllm_compile_cache_on_rocm") 装饰器。通过 compilation_config=CompilationConfig() 保持默认编译配置，避免 VllmRunner 注入缩小的测试编译大小。

关键文件：

- tests/v1/e2e/spec_decode/test_spec_decode.py（模块 推测解码；类别 test；类型 test-coverage；符号 disable_vllm_compile_cache_on_rocm）：唯一变更文件，通过新增 ROCm 平台编译缓存隔离和替换 VllmRunner 上下文管理器来稳定测试。

关键符号：disable_vllm_compile_cache_on_rocm, test_ngram_and_suffix_correctness

关键源码片段

tests/v1/e2e/spec_decode/test_spec_decode.py

唯一变更文件，通过新增 ROCm 平台编译缓存隔离和替换 VllmRunner 上下文管理器来稳定测试。

```
# tests/v1/e2e/spec_decode/test_spec_decode.py
# ( 省略 import 部分 )

@pytest.fixture
def disable_vllm_compile_cache_on_rocm(request: pytest.FixtureRequest) -> None:
    """仅在 ROCm 平台下禁用 vLLM 编译缓存，避免缓存冲突引起故障"""
    if current_platform.is_rocm():
        request.getfixturevalue("disable_vllm_compile_cache")

@pytest.mark.parametrize(
    "speculative_config",
    [
        {
            "method": "ngram",
            "prompt_lookup_max": 5,
            "prompt_lookup_min": 3,
            "num_speculative_tokens": 3,
        },
        {
            "method": "suffix",
            "suffix_decoding_max_spec_factor": 2.0,
        },
    ],
)
@pytest.mark.usefixtures("disable_vllm_compile_cache_on_rocm")
@single_gpu_only
@large_gpu_mark(min_gb=20)
def test_ngram_and_suffix_correctness(
    speculative_config: dict,
    model_name: str,
    vllm_runner,
):
    """使用 VllmRunner 上下文管理器保证引擎优雅关闭和内存清理"""
    with vllm_runner(
        model_name,
        trust_remote_code=False,
        enable_chunked_prefill=None,
        speculative_config=speculative_config,
        max_model_len=4096,
        # 保持 LLM 默认编译 / cudagraph 配置，防止 VllmRunner 注入测试专用缩小尺寸
        compilation_config=CompilationConfig(),
    ) as runner:
        evaluate_llm_for_gsm8k(runner.llm)
```

评论区精华

无 review 评论。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅限于测试文件，未改动生产代码。新增 fixture 仅在 ROCm 平台生效，不影响其他平台。VllmRunner 上下文管理器的使用在 vLLM 测试中已有广泛先例，可靠性可接受。
- 影响：直接影响 ROCm CI 中 ngram/suffix spec decode 正确性测试的稳定性，降低间歇性故障率。对用户和生产路径无影响。
- 风险标记：仅 CI 影响

关联脉络

- 暂无明显关联 PR