

PR #56555 完整报告

vllm-project/vllm

[XPU][CI] Skip ROCm test on non-ROCm platforms

合并时间: 2026-09-12 21:26

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/56555>

执行摘要

- 一句话: 非 ROCm 平台跳过 ROCm 专属测试, 修复 CI 失败
- 推荐动作: 可作为「平台专属测试门控」的最小范例快速浏览, 不值得深度精读。值得关注的设计决策是: 用统一的 `current_platform.is_rocm()` 抽象做平台判定, 而非散落的平台名判断; 新增平台专属测试时应沿用 `skipif` 门控模式, 避免跨平台 CI 被无关用例污染。

功能与动机

PR body 基本空白, 动机主要来自标题与提交信息「Skip ROCm tests on non-ROCm platforms」。被门控的测试通过 monkeypatch 模拟 `rocm.on_gfx950` 与 `rocm_aiter_ops` 开关, 断言 DeepSeek-V4 在 ROCm gfx950 上的 CUDA Graph 默认配置, 这些符号在非 ROCm 平台不存在或行为不一致, 用例在 XPU / CUDA 平台必然失败。作者来自 Intel 并带有 `intel-gpu` 标签, 说明 XPU 平台的 CI 流程因此类用例产生失败告警, 需要按平台收敛测试范围。

实现拆解

1. 变更入口: `tests/test_config.py`, 在 `test_rocm_gfx950_deepseek_v4_cudagraph_default` 的参数化装饰器正上方插入一行 `@pytest.mark.skipif(not current_platform.is_rocm(), reason="ROCm-specific test")`, 位置紧随前一个测试 `test_rocm_keeps_compiled_deepseek_defaults` 的结束处。
2. 生效机制: `pytest` 在收集阶段求值 `skipif` 条件, 条件为真时整组 (5 个参数化组合) 用例被标记跳过, `reason` 会在跳过报告中明确展示, 便于 CI 日志排查。
3. 平台判定: `current_platform.is_rocm()` 是 vLLM 统一的平台抽象接口, 与仓库其他平台分支代码保持同一套判定逻辑, 避免硬编码 `sys.platform` 判断。
4. 配套验证: 无源码、配置或文档改动; CI 由维护者 jikunshang 触发 Buildkite 构建 #88506 并通过后合并。ROCm 平台上该用例运行行为完全不变, 覆盖不丢失。

关键文件:

- `tests/test_config.py` (模块 配置测试; 类别 `test`; 类型 `test-gating`; 符号 `test_rocm_gfx950_deepseek_v4_cudagraph_default`): 唯一变更文件: 为 ROCm 专属的 gfx950 CUDA Graph 默认配置测试追加 `skipif` 平台门控, 是本次 CI 修复的载体。

关键符号: `test_rocm_gfx950_deepseek_v4_cudagraph_default`

关键源码片段

tests/test_config.py

唯一变更文件：为 ROCm 专属的 gfx950 CUDA Graph 默认配置测试追加 skipif 平台门控，是本次 CI 修复的载体。

```
# tests/test_config.py
# 该用例只针对 ROCm gfx950 上的 DeepSeek-V4 CUDA Graph 默认配置，
# 内部 monkeypatch 依赖 ROCm 专属符号 (rocm.on_gfx950、rocm_aiter_ops) ，
# 在 XPU 等非 ROCm 平台执行必然失败，因此在参数化之前整组跳过。
@pytest.mark.skipif(not current_platform.is_rocm(), reason="ROCM-specific test")
@pytest.mark.parametrize(
    ("architecture", "use_v2", "mode", "expected"),
    [
        ("DeepseekV4ForCausalLM", True, None, CUDAGraphMode.NONE),
        ("DeepseekV4ForConditionalGeneration", True, None, CUDAGraphMode.NONE),
        ("DeepseekV4ForCausalLM", False, None, None),
        ("LlamaForCausalLM", True, None, None),
        (
            "DeepseekV4ForCausalLM",
            True,
            CUDAGraphMode.FULL_DECODE_ONLY,
            CUDAGraphMode.FULL_DECODE_ONLY,
        ),
    ],
)
def test_rocm_gfx950_deepseek_v4_cudagraph_default(
    monkeypatch, architecture, use_v2, mode, expected
):
    # 引用 rocm 平台模块与 AITER 算子开关，全部是 ROCm 专属能力；
    # 注意 skipif 条件在 pytest 收集阶段求值，此处的局部导入不影响跳过判断。
    from vllm._aiter_ops import rocm_aiter_ops
    from vllm.platforms import rocm

    monkeypatch.setattr(rocm, "on_gfx950", lambda: True)
    monkeypatch.setattr(rocm_aiter_ops, "is_fused_moe_enabled", lambda: False)
    monkeypatch.setattr(rocm_aiter_ops, "is_linear_fp8_enabled", lambda: False)
    monkeypatch.setattr(
        rocm_aiter_ops, "is_fusion_moe_shared_experts_enabled", lambda: False
    )
    config = SimpleNamespace(
        compilation_config=CompilationConfig(cudagraph_mode=mode),
        model_config=SimpleNamespace(architecture=architecture),
        use_v2_model_runner=use_v2,
    )
    # 后续按 mode/expected 校验 CUDA Graph 模式的默认取值，此处省略。
```

评论区精华

该 PR 无内联 review 评论，主要交流集中在 issue 评论区：

- 作者 zhenwei-intel 请求 review：「@AndreasKaratzas, @jikulshang, PTAL」。
- 作者尝试自行触发 CI 时被 github-actions bot 拦截：「A reviewer with write access must run /ci run, approve the PR, or add the ready label first」——体现 fork PR 的 CI 权限约束。
- 维护者 jikulshang 代为触发 Buildkite CI #88506 并直接 approve。
- claude[bot] 提示该 PR 来自 fork，自动 review 关闭，需维护者手动触发。
- Fork PR 的 CI 触发与审批流程 (other): 维护者触发 Buildkite CI #88506 并通过，PR 合入。
- fork PR 自动 review 被禁用 (other): 未触发额外深度 review，由维护者 jikulshang 直接审批通过。

风险与影响

- 风险：回归风险几乎为零：无运行时源码改动，影响面仅限 pytest 收集阶段；skipif 条件在收集期求值，要求 current_platform 在模块级可解析（本次 CI 已验证通过）。覆盖收敛是唯一值得注意的点：该用例在非 ROCm 平台被有意跳过，若未来 CUDA Graph 默认配置出现跨平台回归，不会由本用例暴露，需依赖平台无关用例。无安全性、性能或兼容性影响，pytest.mark.skipif 为通用成熟机制。
- 影响：影响范围限于测试与 CI：XPU、CUDA 等非 ROCm 平台的 CI 不再被必然失败的 ROCm 专属用例干扰，减少噪音性失败排查成本；ROCM 平台测试覆盖完全不变。对用户与线上服务无任何运行时影响。对团队而言是 1 行改动、1 commit、无返工的低成本维护，合入风险极低，影响程度小。
- 风险标记：测试收集期条件求值，非 ROCm 侧覆盖收敛

关联脉络

- PR #57044 [CI] Keep Qwen3 Omni DSpark config fixture complete: 同为 CI 测试夹具 / 平台适配的稳定性维护，目标都是消除与平台无关的 CI 失败。
- PR #56743 [ROCM][Perf] Optimize DSV4.1 K=512 decode top-k on gfx950: 同属 ROCm gfx950 平台优化与测试链路，被门控用例正是在 gfx950 上验证 DeepSeek-V4 的 CUDA Graph 默认配置。