PR #44369 完整报告

vllm-project/vllm

[ROCm][CI] Skip fp8 reload tests on gfx90a (MI250)

合并时间：2026-06-03 11:27

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44369

执行摘要

- 一句话：对 gfx90a 跳过 FP8 reload 测试
- 推荐动作：建议精读 _fp8_reload_unsupported() 的实现，作为处理平台特定测试跳过的良好范例——它展示了如何在不修改全局平台 API（如 supports_fp8()）的前提下，通过本地化函数解决特定硬件的测试问题。

功能与动机

在 ROCm（MI250 / gfx90a）平台上，supports_fp8() 对于通用（upcast）量化路径返回 True，但 gfx90a 没有原生 FP8，无法运行这些 reload 模型。因此 FP8 reload / online-quantize / kv-scale 测试未被跳过，导致 CI 失败。需要添加 gfx90a 的特定守卫，在不影响 supports_fp8() 的前提下跳过这些测试。

实现拆解

1. 新增辅助函数 _fp8_reload_unsupported()：定义在 tests/model_executor/model_loader/test_reload.py 中，位于导入之后、第一个测试类之前。函数逻辑：如果 current_platform.supports_fp8() 返回 False，则直接返回 True（FP8 不支持）；否则，如果是 ROCm 平台，则通过 vllm.platforms.rocm.on_gfx90a() 检查是否为 gfx90a 架构，若是则返回 True；其他情况返回 False。该函数通过 docstring 解释了为何需要额外检查 gfx90a。
2. 修改三处测试的 skip 条件：
   - test_reload_weights：将 "FP8" in base_model and not current_platform.supports_fp8() 改为 "FP8" in base_model and _fp8_reload_unsupported()。
   - test_kv_scale_reload：将 not current_platform.supports_fp8() 改为 _fp8_reload_unsupported()。
   - test_online_quantize_reload：将 quantization == "fp8" and not current_platform.supports_fp8() 改为 quantization == "fp8" and _fp8_reload_unsupported()。 所有修改均在同一文件内，只涉及测试条件的调整，不改变任何生产代码或测试逻辑的其他部分。

关键文件：

- tests/model_executor/model_loader/test_reload.py（模块 重载；类别 test；类型 test-coverage；符号 _fp8_reload_unsupported）：唯一变更文件，新增 _fp8_reload_unsupported() 函数并修改了三处测试的 skip 条件，是整个 PR 的核心所在。

关键符号：_fp8_reload_unsupported

关键源码片段

tests/model_executor/model_loader/test_reload.py

唯一变更文件，新增 _fp8_reload_unsupported() 函数并修改了三处测试的 skip 条件，是整个 PR 的核心所在。

```
def _fp8_reload_unsupported() -> bool:
    """Whether the FP8 reload / online-quantize tests should be skipped.

    ``supports_fp8()`` returns True on MI250 (gfx90a) because the general
    quantization paths upcast FP8 weights, but gfx90a has no native FP8 and
    cannot run these reload models, so treat it as unsupported here.
    """
    # 如果平台本身不支持 FP8，直接返回 True
    if not current_platform.supports_fp8():
        return True
    # 如果是 ROCm 平台，进一步检查是否为 gfx90a
    if current_platform.is_rocm():
        from vllm.platforms.rocm import on_gfx90a
        return on_gfx90a()
    # 其他情况（如 NVIDIA 平台）返回 False
    return False
```

评论区精华

无 review 评论，仅有一人（AndreasKaratzas）批准，评论为“LGTM”。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：变更仅涉及测试文件中的 skip 条件，且函数逻辑清晰：在 supports_fp8() 为 False 时直接返回 True，简化了原条件；对于 ROCm 平台仅增加 on_gfx90a() 的本地导入和调用，不影响其他平台。不涉及任何生产代码、性能路径或配置变更。
- 影响：仅影响 ROCm CI 中的 FP8 reload/online-quantize/kv-scale 测试：在 gfx90a（MI250）上这些测试将被跳过，不再因假阳性而失败。对其他平台（如 NVIDIA）无影响，因为这些平台 on_gfx90a() 返回 False 或根本不会进入 ROCm 分支。对用户无影响。
- 风险标记：仅测试变更，低风险

关联脉络

- PR #44042 [CI] Reject out-of-vocabulary before they reach the GPU logprob path: 同为 ROCm CI 稳定性修复，处理平台特定问题；修改了不同测试文件但目标一致。