PR #38566 完整报告

vllm-project/vllm

[Bugfix][CI] Skip flaky `test_eagle` test

合并时间：2026-03-31 21:42

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38566

执行摘要

此 PR 通过将测试 test_eagle 标记为预期失败来临时修复 CI 中的 flaky 问题，针对 EAGLE + DP > 1 与异步 spec decode 组合的错误输出问题，但未解决根本原因，旨在稳定 CI 流程。

功能与动机

动机源于 issue #31913 报告的 flaky 测试，作者尝试避免 CI 失败干扰开发。PR body 中提到这是试探性修复，并讨论禁用 async with DP>1 可能太严厉，但希望获取更多上下文确认是否需要此限制。

实现拆解

实现集中在单个文件 tests/v1/distributed/test_eagle_dp.py 中：

- 添加新的 @pytest.mark.xfail 装饰器，条件为 not current_platform.is_rocm()，原因字符串说明 EAGLE + DP > 1 产生错误输出且 root cause 在调查中。
- 更新现有 ROCm 相关 xfail 装饰器的原因字符串。 代码变更示例：

评论区精华

review 讨论中突出两点：

1. 设计权衡：MatthewBonanni 在 issue 评论中表示禁用 async spec decode for DP > 1 可能不正确，因为测试 flaky 可能源于 batch invariance。> MatthewBonanni: "Like you mention, I don't think disabling async spec decode for DP > 1 is the right move..."
2. 文档建议：gemini-code-assist[bot] 建议在 gpu_model_runner.py 中添加注释和 TODO 以记录限制并跟踪未来工作，但此建议未在本 PR 中实施。

风险与影响

- 风险：跳过测试可能掩盖底层逻辑问题（如 EAGLE 与异步 spec decode 的组合 bug），导致未发现的回归；未解决 root cause 可能影响系统正确性。
- 影响：短期提升 CI 稳定性，减少 false 失败信号；长期需关注根本修复，否则可能延迟相关功能开发和测试。

关联脉络

与此 PR 紧密相关的是 PR #38584，它同样处理 Eagle DP 测试的 flaky 行为，通过禁用 AOT 调度来修复，表明该测试区域存在持续问题。结合近期历史 PR，vLLM 项目在 v1 模块和 speculative decoding 功能上有较多测试和修复工作，此 PR 是临时措施，可能为后续根本性修复铺路。