

PR #49916 完整报告

vllm-project/vllm

[CI][ROCm] Reduce V1 attention test runtime

合并时间: 2026-07-27 11:12

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49916>

执行摘要

- 一句话: 缩减 ROCm V1 attention 测试耗时
- 推荐动作: 建议阅读, 特别是 collection-time 参数验证的设计, 可推广到其他需要根据运行时能力跳过测试的场景。

功能与动机

PR body 明确表示要'Reduce V1 attention test runtime', 通过从 MI250 移除测试、对 MI300/MI355 分片、以及优化参数收集时跳过不可运行的组合来减少 CI 耗时。

实现拆解

1. 移除 MI250 测试 job (.buildkite/test-amd.yaml) : 删除 V1 attention (H100-MI250) 标签块。
2. 分片 MI300 和 MI355 (.buildkite/test-amd.yaml) : 添加 parallelism: 2 和分片命令。
3. 添加 ROCm AITER prefill backend (tests/v1/attention/test_mla_backends.py) : 在 PREFILL_BACKENDS_TO_TEST 中加入 MLAPrefillBackendEnum.ROCM_AITER_FA。
4. 实现 collection-time 参数验证 (tests/v1/attention/test_mla_backends.py) : 新增 _prefill_backend_dimension_params 函数, 通过 validate_configuration 提前跳过不支持的组合。
5. 重构参数化 (tests/v1/attention/test_mla_backends.py) : test_backend_correctness 改为调用 _prefill_backend_dimension_params。

关键文件:

- tests/v1/attention/test_mla_backends.py (模块 MLA 测试; 类别 test; 类型 test-coverage; 符号 _prefill_backend_dimension_params, test_backend_correctness) : 核心测试文件, 实现了 collection-time 参数验证、新增 ROCm AITER backend 支持、重构参数化逻辑。
- .buildkite/test-amd.yaml (模块 CI 配置; 类别 config; 类型 configuration) : CI 配置文件, 移除了 MI250 的 V1 attention job, 为 MI300 和 MI355 增加了分片配置。

关键符号: _prefill_backend_dimension_params, test_backend_correctness

评论区精华

PR 没有较多讨论。审核者 tjtanana 评论 'LGTM. Let's check the CUDA CI.' 并批准，无未解决疑虑。

- 暂无高价值评论线程

风险与影响

- 风险：主要风险包括： 1) 移除 MI250 测试后可能遗漏该硬件上的回归； 2) 分片依赖 BUILDKITE_PARALLEL_JOB 环境变量，配置错误可能导致 job 失败； 3) collection-time 跳过基于 validate_configuration，该方法如果有 bug 可能跳过本该运行的测试。
- 影响：直接减少 AMD CI 中 V1 attention 测试的总 wall time，提升 CI 资源利用率。对 MI300 和 MI355 用户正面，MI250 用户部分测试覆盖降低，但未移除源码支持。
- 风险标记：移除 MI250 测试覆盖，分片配置正确性

关联脉络

- PR #49516 [ROCm][CI] Keep native datasets cache off shared NFS: 同为 ROCm CI 优化，关注 CI 运行稳定性。
- PR #49911 [CI][ROCm] Keep global GPU memory cleanup opt-in: 同为 ROCm CI 配置调整，改进测试环境。