

PR #53602 完整报告

vllm-project/vllm

[ROCm][CI] Split MI300 Distributed Compile by graph partition mode

合并时间: 2026-09-10 06:07

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/53602>

执行摘要

- 一句话: 拆分 MI300 分布式编译 CI 作业以缩短关键路径。
- 推荐动作: 此 PR 是一个高质量的 CI 优化。它针对明确的性能瓶颈 (关键路径 CI 耗时过长) 提出了一个经过实测数据支撑的、具体且低风险解决方案。PR body 提供了充分的决策依据 (测量数据、超时计算逻辑、权衡分析)。建议维护者合并, 以加速 AMD ROCm 平台的开发迭代。

功能与动机

该 CI 作业因测试 `test_tp2_ar_rms_fusions` 展开为 64 个参数组合 (每个组合需启动新的 2-GPU 引擎) 而成为 AMD CI 的耗时瓶颈 (77.7 分钟), 严重影响 CI 反馈速度。目标是将其拆分以并行运行, 缩短总耗时。

实现拆解

1. 变更入口: 修改 `.buildkite/test-amd.yaml` 配置文件中的 (MI300) Distributed Compile 作业定义。
2. 核心逻辑拆解: a. 将单个作业步骤拆分为两个步骤: (MI300) Distributed Compile (Inductor Partition) 和 (MI300) Distributed Compile (Dynamo Partition)。b. 通过为 `pytest` 命令添加 `-k "inductor_partition"` 和 `-k "dynamo_partition"` 参数过滤器, 将 64 个参数化的测试用例平均分配到两个作业中运行。c. 两个新作业共享相同的测试矩阵、依赖路径和 agent pool (`mi300_2`), 但运行不同的测试子集。d. 原作业中包含的 `test_async_tp.py` 测试在两个新作业中均会运行, 以保证每个作业的自包含性。
3. 配置配套调整: 将两个新作业的 `timeout_in_minutes` 从原有的 95 分钟调整为 65 分钟, 该值是基于实测数据 (约 44.5 分钟) 加上 20 分钟余量并向上取整到 5 的倍数计算得出。
4. 无测试、Schema 或部署配套改动: 本次变更仅涉及 CI 配置, 不涉及应用代码或测试逻辑的变更。

关键文件:

- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 config; 类型 configuration): 本次 PR 的唯一变更文件, 是 CI 作业定义的核心配置。所有变更 (拆分作业、添加过滤器、调整超时) 均在此文件中完成。

关键符号: 未识别

评论区精华

1. 测试覆盖度与正确性：PR body 中详细论证了 -k 过滤器的正确性，指出过滤器匹配的字面量 inductor_partition 和 dynamo_partition 在测试矩阵的参数 ID 中是唯一且互斥的，确保了两个作业的测试用例集合不相交且完备（ $32 + 32 = 64$ ）。
 2. 资源与时间权衡：PR 明确指出了拆分的权衡：需要同时空闲两个 mi300_2 代理来并行运行，总 GPU 时间会增加（从 2 GPU 增至 4 GPU，且 test_async_tp.py 会重复运行）。这是为了换取关键路径（wall-clock time）的缩短。
 3. 超时设置决策：讨论了超时值的选择逻辑，基于实际观测时间和 PR #54695 建立的规则（平均值 + 20 分钟余量），并解释了不再沿用旧值（95 分钟）的原因。
- 测试过滤器的完备性与正确性 (testing): 基于对测试参数化 ID 的分析，认为过滤器是正确且完备的，两个作业的测试集不相交且联合等于原测试集。
 - 作业拆分带来的资源与时间权衡 (performance): 决策是接受更高的 GPU 资源消耗，以换取 CI 反馈速度的提升（从 77.7 分钟降至约 44.5 分钟）。
 - 超时值的设置依据 (correctness): 超时设置基于实测和既定规则，旨在提供足够缓冲但又不过度宽松。

风险与影响

- 风险：
 1. CI 资源竞争风险：新方案要求同时占用两个 mi300_2 代理，如果 AMD CI 资源池不足，可能导致作业排队时间增加，抵消并行化带来的时间收益。
 2. 测试覆盖遗漏风险：虽然 source_file_dependencies 配置一致，但若未来测试矩阵的参数 ID 发生变化（例如 inductor_partition 子串出现在其他组件中），可能导致过滤器误匹配或遗漏。不过，当前分析认为此风险较低。
 3. 无应用层面风险：由于变更仅限于 CI 配置，不影响 vLLM 的核心功能、性能或安全性。
- 影响：
 1. 对团队的影响：显著改善了 AMD ROCm 开发者的工作流。将 CI 反馈时间从近 80 分钟缩短至约 45 分钟，加快了代码合并和问题发现的速度，提升了团队效率。
 2. 对系统的影响：仅影响 CI 系统，无功能影响。可能增加 CI 基础设施的 GPU 资源消耗（峰值 GPU 数量翻倍），但总 GPU 时间增幅可控。
 3. 对用户的影响：间接影响。更快的 CI 反馈意味着 vLLM 的 ROCm 版本能够更稳定、更频繁地发布。 - 风险标记：CI 资源依赖增加

关联脉络

- PR #54695 [ROCm][CI] Right-size AMD MI300 CI timeouts: 本 PR 的超时设置逻辑（平均值 + 20 分钟余量）明确引用了 PR #54695 建立的规则，属于同一优化框架下的延续工作。