

PR #52322 完整报告

vllm-project/vllm

[CI] Shard extended pooling model tests

合并时间: 2026-08-15 06:11

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52322>

执行摘要

- 一句话: 扩展池化测试拆 4 分片, CI 提速约 3.7 倍
- 推荐动作: 值得快速浏览, 作为 CI 分片模板的参考。关注点: Buildkite parallelism + %N label + 环境变量传参的组合, 以及通过 ci-infra 提供 mirror.amd.parallelism 覆盖的向后兼容机制。合并前务必确认 ci-infra #473 已合入。

功能与动机

该 job 是 CI 关键路径的主要瓶颈。PR body 明确说明 'Buildkite build #83851 took 72 minutes for this job. Its recent passed-run distribution is 70 minutes median, 82 minutes p90, and 112 minutes max.', 并指出 'Four shards leave substantially more headroom than three shards for the 30-minute target while preserving the same test selection', 目标是将该 job 压到 30 分钟以内。

实现拆解

1. NVIDIA 主 job 分片: 在 `.buildkite/test_areas/models_language.yaml` 的 `Language Models Test (Extended Pooling)` 步骤上增加 `parallelism: 4`, label 追加 `%N` 占位符, 并在 `pytest` 命令末尾追加 `--num-shards=$$BUILDKITE_PARALLEL_JOB_COUNT --shard-id=$$BUILDKITE_PARALLEL_JOB`, 利用 `pytest` 的 nodeID 确定性分片实现并行。
2. AMD mirror 分片: 根据 review 中 AndreasKaratzas 的建议, 作者在第二笔提交 `f391ffb` 中为 AMD mirror 设置 `parallelism: 2`, 并显式定义 mirror 下的 `commands`, 同样传入 `pytest` 分片参数, 避免 AMD 侧过度占用 MI300 资源。
3. 依赖与验证: 该改动依赖 ci-infra #473 提供的向后兼容 `mirror.amd.parallelism` 覆盖能力, 因此合并顺序有约束。作者在 PR body 中列出了完整验证: 本地渲染 pipeline 确认 NVIDIA 4 片、AMD 2 片; targeted Buildkite run #83898 通过, 4 个分片墙钟时间分别为 14.128 / 15.324 / 19.486 / 18.860 分钟, 最大 19.486 分钟, 比 #83851 快 3.697 倍; 同时用脚本核对分片清单两两不相交且并集为 120/120 个测试节点。

关键文件:

- `.buildkite/test_areas/models_language.yaml` (模块 测试配置; 类别 config; 类型 configuration): 唯一变更文件, 集中定义 extended pooling 测试的 NVIDIA 主 job 与 AMD mirror 的分片配置, 是本次 CI 提速的全部改动。

关键符号: 未识别

评论区精华

核心讨论点在于 AMD mirror 是否分片以及并行度选择。AndreasKaratzas 在 diff 上留言 'we can shard that here too but probably with parallelism 2 (i think 4 would be too much for the amd mirror)', 作者 khluu 回复 'Done in f391ffb38b. The AMD mirror now uses parallelism 2 and passes the Buildkite shard count and shard ID to pytest.'。结论是 AMD mirror 采用 2 分片，兼顾提速与资源占用，讨论已解决。

- AMD mirror 是否分片以及并行度选择 (design): AMD mirror 采用 parallelism 2，兼顾收益与 MI300 资源占用，讨论已解决。

风险与影响

- 风险：
 - 合并顺序风险：本 PR 依赖 ci-infra #473，若其未合并，AMD mirror 的 parallelism 覆盖不被支持，可能导致 pipeline 行为不符合预期。PR body 已声明不可先合并。
 - 资源占用：NVIDIA 侧 4 个并行 job 同时占用 4 个 h200_35gb 实例，AMD 侧 2 个 mi300_1，瞬时 GPU 配额需求上升，需确认 CI 容量足够。
 - 分片确定性：pytest shard 基于测试收集顺序划分，若未来测试文件组织变化，各分片负载可能不均，但测试覆盖范围不变。
 - 超时配置：timeout 仍为 120 分钟，分片后最长 19.5 分钟，余量充足，即使分片失效回退单 job 也不会超时。
- 影响：
 - 对用户：无直接影响，纯 CI 内部改进。
 - 对开发团队：该 job 从 CI 关键路径缩短约 50 分钟，显著提升 Merge 队列和 PR 验证效率，且该分片模式可推广到其他长耗时 job。
 - 对 CI 资源：并行度上升，需要关注 GPU 配额；AMD mirror 分 2 片是资源与收益的折中。
 - 风险标记：依赖外部 ci-infra 合并顺序，并行 job 增加 GPU 资源消耗，pytest 分片确定性依赖，AMD mirror 配置向后兼容性

关联脉络

- PR #52326 [CI] Shard Humming H100 eval: 同属 CI 测试分片提速系列，改动 buildkite/test_areas/lm_eval.yaml，与本 PR 目标一致。
- PR #52327 [CI] Shard MoE refactor B200 eval: 同属 CI 测试分片提速系列，拆分 MoE eval 为 4 分片，与本 PR 采用相同分片模式。
- PR #52323 [CI] Shard multimodal extended generation 2: 类似地将多模态扩展生成测试拆为 4 并行分片，同属 CI 关键路径提速工作。
- PR #52252 [CI] Increase extended generation test timeout: 修改同一文件 buildkite/test_areas/models_language.yaml，调整同区域测试 job 的超时配置。