

PR #52326 完整报告

vllm-project/vllm

[CI] Shard Humming H100 eval

合并时间: 2026-08-15 06:12

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52326>

执行摘要

- 一句话: Humming H100 eval 拆 3 个并行 Job, 关键路径提速约 3 倍
- 推荐动作: 建议 CI 维护者阅读: 这是 Buildkite parallelism + BUILDKITE_PARALLEL_JOB 环境变量 + 配置列表分片的简洁示例, 对长耗时、易并行化的测试步骤很有参考价值。但注意三份 shard 文件是手工维护的, 后续新增或删除配置时需同步重平衡, 可以考虑后续脚本化自动生成。

功能与动机

PR body 明确说明: lm-eval-humming-f16-h100 在完整 CI build 83851 中耗时 62.6 分钟, 14 个模型配置虽然相互独立却串行执行。为了把该步骤压到 30 分钟以内, 改为 3 个 Job 并行分片, 并基于 build 83851 的实测单配置耗时做分区。

实现拆解

实现分 3 步:

1. 步骤级并行化: 修改 `.buildkite/test_areas/lm_eval.yaml` 中 `lm-eval-humming-f16-h100` 步骤, 新增 `parallelism: 3`, label 追加 `%N` 以区分并行 Job, 并把 `pytest` 的 `--config-list-file` 从固定的 `config.txt` 改为按 `BUILDKITE_PARALLEL_JOB` 环境变量动态选择 `config-h100-shard-$$BUILDKITE_PARALLEL_JOB.txt`, 这是整个改动的入口。
2. 配置分片: 新增 3 个 shard 配置文件 (`tests/evals/gsm8k/configs/humming/config-h100-shard-0.txt`、`shard-1.txt`、`shard-2.txt`), 按 build 83851 中 14 个配置各自的实测耗时手工分区, 使每 shard 的负载约为 20.5、20.4、20.7 分钟, 且在 30 分钟目标下留有余量, 14 个配置无重无漏。
3. 验证与配套: 通过 targeted build 83895 验证三个 shard 全部通过、覆盖 14/14 配置且无重复无遗漏; 关键路径从 62.63 分钟降到 20.79 分钟 (约 3.01 倍), 总 GPU 时间基本不变 (60.45 vs 62.63 GPU 分钟), 未涉及任何产品代码或测试逻辑改动。

关键文件:

- `.buildkite/test_areas/lm_eval.yaml` (模块 CI 流水线; 类别 config; 类型 configuration): 核心变更: 给 `lm-eval-humming-f16-h100` 步骤加 `parallelism: 3` 和 `%N`, 并把命令改为按 `BUILDKITE_PARALLEL_JOB` 加载 shard 配置, 是并行化的入口。
- `tests/evals/gsm8k/configs/humming/config-h100-shard-0.txt` (模块 评测配置; 类别 docs; 类型 documentation): 新增 shard 0 配置列表, 包含 5 个 Humming 模型配置,

是按耗时划分的第 1 片。

- tests/evals/gsm8k/configs/humming/config-h100-shard-1.txt (模块 评测配置; 类别 docs; 类型 documentation) : 新增 shard 1 配置列表, 包含 5 个 Humming 模型配置, 是按耗时划分的第 2 片。
- tests/evals/gsm8k/configs/humming/config-h100-shard-2.txt (模块 评测配置; 类别 docs; 类型 documentation) : 新增 shard 2 配置列表, 包含 4 个 Humming 模型配置, 是按耗时划分的第 3 片。

关键符号: 未识别

评论区精华

本 PR 没有人工 review 评论。claude[bot] 自动评论指出这是 fork PR, 自动审核被禁用, 维护者可以评论 @claude review 触发一次性审核。没有关于分片策略或时长的公开讨论。

- Fork PR 自动 review 已禁用 (other): 无实质技术讨论, 合入前由维护者自行审核。

风险与影响

- 风险: 风险集中在维护与资源层面:
- 手工分片易失衡: 三个 shard 文件是静态手工划分的, 后续新增或删除 Humming 评测配置、或某个模型耗时明显变化时, 需要重新平衡, 否则可能出现某个 Job 逼近 70 分钟超时或覆盖率缺口。
- 并行资源占用上升: 从 1 个 Job 独占 H100 变为 3 个 Job 同时占 3 块 H100, 总 GPU 时间基本不变 (60.45 vs 62.63 GPU 分钟), 但瞬时并发压力变大, optional: true 仍保留, 失败不会阻塞主流水线。
- 环境变量依赖: 命令中 \$\$BUILDKITE_PARALLEL_JOB 依赖 Buildkite 并行 Job 注入的序号, 若在其他 CI 系统或本地重放该步骤, 该变量不存在时会有问题。
 - 未改动测试逻辑本身, 评估正确性由原 pytest 用例保证。
- 影响: 对开发者: Humming/H100 eval 的关键路径从约 62.6 分钟降至约 20.8 分钟, PR 合并等待时间明显缩短, 但之后维护 shard 列表需要额外注意。对系统: H100 由 1 个 Job 独占变为 3 个 Job 并行, 瞬时资源占用变大。对团队: 提供了一种 Buildkite parallelism + 环境变量分片的标准做法, 可复用到其他长耗时的 eval 步骤。
- 风险标记: 并行 GPU 资源占用上升, 手工分片易失衡, 配置维护成本

关联脉络

- PR #50589 [CI][Test] Seed the DeepEP v2 MoE workers, not just the parent: 同为 CI 测试稳定性与并行效率调整, 涉及 MoE 评测类测试的随机性与可靠性。
- PR #52252 [CI] Increase extended generation test timeout: 同为 CI 步骤超时与流水线时间预算管理, 与本 PR 的 timeout_in_minutes 调整同属 CI 治理。
- PR #52237 [UT] fix device of test_outputs.py: 同为测试与 CI 平台适配类小改动, 与本 PR 同属测试基础设施维护。