

PR #52327 完整报告

vllm-project/vllm

[CI] Shard MoE refactor B200 eval

合并时间: 2026-08-15 06:11

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52327>

执行摘要

- 一句话: MoE B200 eval 拆 4 分片, 关键路径提速 3.65x
- 推荐动作: 若关注 CI 可靠性和 Buildkite pipeline 写法, 本 PR 值得快速精读, 重点学习 %N + parallelism + BUILDKITE_PARALLEL_JOB 的组合用法。由于分区是静态的, 后续新增模型或评估口径变化时需要主动更新 4 个分片文件; 建议把静态分片当作技术债跟进, 后续可考虑按配置时长自动均衡分片。

功能与动机

PR 正文说明 moe-refactor-integration-test-b200-temporary 在完整 CI build 83851 中耗时 81.3 分钟, 19 个模型配置虽然相互独立却串行执行; 作者提前搜索过开放 PR, 确认没有重复分片实现。本次变更旨在把该步骤拆成多个并行 Job, 缩短验证 MoE refactor 回归的反馈时间。

实现拆解

1. 在 .buildkite/test_areas/lm_eval.yaml 中修改 moe-refactor-integration-test-b200-temporary 步骤: label 追加 %N, 新增 parallelism: 4, 使 Buildkite 生成 4 个并行实例。
2. 将命令改为 --config-list-file=.../config-b200-shard-\$\$BUILDKITE_PARALLEL_JOB.txt, 运行时按注入的分片序号加载对应配置。
3. 新增 config-b200-shard-0~3.txt 四个配置列表, 共同覆盖 19 个模型配置; 分区依据 build 83851 各配置的实测耗时静态分配, 避免最慢配置 (单跑 13.31 分钟) 拖垮某个分片。
4. 验证: 基线 83851 中 19/19 配置通过; 分区检查确认无遗漏 / 重复; 目标构建 83896 四分片全部通过, 各分片耗时 19.47/21.54/20.23/22.26 分钟。除 CI 配置和 txt 列表外无代码改动。

关键文件:

- .buildkite/test_areas/lm_eval.yaml (模块 CI 配置; 类别 config; 类型 configuration): 唯一的核心配置变更, 通过 parallelism 与 BUILDKITE_PARALLEL_JOB 将原本串行的 B200 eval 步骤拆成 4 个并行分片 Job, 是缩短关键路径的直接原因。
- tests/evals/gsm8k/configs/moe-refactor/config-b200-shard-0.txt (模块 测试配置; 类别 test; 类型 test-config): 分片 0 包含最耗时的 DeepSeek-V4-Flash-deepgemm-mega-moe 等 3 个配置, 是静态分片均衡策略的关键样例。

- tests/evals/gsm8k/configs/moe-refactor/config-b200-shard-1.txt (模块 测试配置; 类别 test; 类型 test-config) : 分片 1 包含 5 个配置, 与其他分片一起构成 19 个配置的无重复覆盖。
- tests/evals/gsm8k/configs/moe-refactor/config-b200-shard-2.txt (模块 测试配置; 类别 test; 类型 test-config) : 分片 2 包含 5 个配置, 用于平衡 4 个并行 Job 的总耗时。
- tests/evals/gsm8k/configs/moe-refactor/config-b200-shard-3.txt (模块 测试配置; 类别 test; 类型 test-config) : 分片 3 包含 6 个配置, 用于吸收剩余负载并维持分片数量最少的原则。

关键符号: 未识别

评论区精华

PR 没有产生人工 review 评论 (comments_count=0)。claude[bot] 提示该 PR 来自 fork, 自动 review 被禁用; pavanimajety 直接批准。PR 正文中作者点明了关键设计取舍: 不用按数量自动分片而采用静态列表, 因为最慢配置单独就要 13.31 分钟, 静态列表才能精确控制每个分片的整体耗时, 避免一个重模型拖垮某个并行 Job。

- 暂无高价值评论线程

风险与影响

- 风险: 静态分片依赖 build 83851 的测量结果, 模型加载或编译时间变化后不会自动重新平衡, 分片时长可能逐渐漂移。并行 4 Job x 2 卡使单次运行 GPU 总耗时增加约 2.7% (167.01 vs 162.63 GPU-minutes), 重跑次数多时会放大资源占用。b200-k8s 资源池若配额不足, 并行 Job 可能因排队或抢占而延长等待。另外命令依赖 Buildkite 注入的 BUILDKITE_PARALLEL_JOB, 若脱离 CI 手动执行会因变量缺失而失败。
- 影响: 对 vLLM 运行时零影响, 纯 CI 基础设施改动。对 MoE refactor 验证链路: 关键路径从 81.32 分钟降到 22.26 分钟, 开发者迭代验证的反馈速度提升约 3.65 倍。对 CI 资源池: 单次运行的 GPU 占用量小幅上升, 调度压力略有增加。对同仓库其他 eval 步骤形成可复用的分片模板, 是一系列 CI 提速工作的组成部分。
- 风险标记: 静态分片易失衡, GPU 占用小幅上升, 并行变量依赖平台

关联脉络

- PR #52326 [CI] Shard Humming H100 eval: 同属 .buildkite/test_areas/lm_eval.yaml 下的 eval 步骤分片改造, 与本 PR 共用 parallelism + %N + 分片配置的实现模式, 是同一 CI 提速系列的姊妹 PR。
- PR #52252 [CI] Increase extended generation test timeout: 同属围绕 eval/ 生成测试 CI 时长与稳定性的调整, 反映该阶段 CI 基础设施的持续优化方向。