

PR #51068 完整报告

vllm-project/vllm

Prune redundant tests points in `correctness\_e2e/[test\_sequence\_parallel,test\_async\_tp]`

合并时间: 2026-08-05 10:18

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51068>

执行摘要

- 一句话: 精简 SP/AsyncTP 端到端测试点并迁至 nightly
- 推荐动作: 值得阅读, 尤其是负责 CI/ 测试矩阵的工程师。优点: 用 `_build_sp_args` + `_compare_sp_settings` 把环境过滤与参数构建抽离, 配合 `compare_all_settings` 做批量对比, 是很实用的测试重构范式; 同时把耗时功能从 PR 门禁移到 nightly, 体现测试分级思想。但要注意覆盖面取舍, 后续若有 SP + eager/chunked prefill 相关改动应单独补测。

功能与动机

PR body 明确说明动机: 「Sequence Parallel Correctness Tests takes over an hour and isn't a default feature users run into. Let's greatly simplify the duplicated test points and run it on the nightly. Similar to AsyncTP」。即测试成本与收益不匹配: SP 不是默认路径却占用超过 1 小时 CI, 且大量参数组合重复, 需要大幅精简并移到夜间任务。

实现拆解

1. `tests/compile/correctness_e2e/test_sequence_parallel.py`: 移除 `ParallelSetup` 中的 `chunked_prefill` 字段, 删除 `SPTTestSettings.detailed` 工厂方法; fast 组合从 `eager × pp × chunked` 降为仅 `pp` 两种配置, 固定 `eager_mode=False`; `fp8_quant` 只遍历融合开关。模型矩阵改为 `hmlor/tiny-random-LlamaForCausalLM` 与 `RedHatAI/Llama-3.2-1B-Instruct-FP8`。
2. 把 `_compare_sp` 重构为 `_build_sp_args` (返回 `(tp_args, tp_sp_args)` 或 `None`) 和 `_compare_sp_settings` (按相同 baseline 分组后调用 `compare_all_settings`); `test_tp_sp_generation` 根据 PyTorch 版本动态决定 `use_inductor_graph_partition` 候选, 并按设备能力统一过滤跳过条件。
3. `tests/compile/correctness_e2e/test_async_tp.py`: 移除 `eager` 模式参数化, FP8 模型改为 1B 的 `RedHatAI/Llama-3.2-1B-Instruct-FP8`。
4. `tests/compile/passes/distributed/test_async_tp.py`: 测试模型默认 dtype 改为 `bfloat16`; dtype 参数化只保留 `bf16`, 删除 FP16 分支及对应的 skip 逻辑 (FP8 `_scaled_mm` 行缩放只支持 `bf16` 精度)。
5. `.buildkite/test_areas/compile.yaml` 与 `tests/models/registry.py`: 删除固定执行的 2 GPU SP job 和 H100 AsyncTP job, 合并为 B200 的可选步骤 (optional: true), SP job 超时从 80 分钟降到 45 分钟; registry 为 `LlamaForCausalLM` 增加 `fp8_1b` 额外项指

向新的 1B FP8 模型。整套改动属于测试资源与 CI 门禁策略调整。

关键文件：

- tests/compile/correctness_e2e/test_sequence_parallel.py (模块 序列并行; 类别 test; 类型 test-coverage; 符号 SPTestSettings.fast, SPTestSettings.fp8_quant, _build_sp_args, _compare_sp_settings) : SP 正确性测试矩阵的核心改动: 删除 detailed、缩减 fast 组合、用 compare_all_settings 聚合对比, 是本 PR 的变更主体。
- tests/compile/passes/distributed/test_async_tp.py (模块 异步 TP; 类别 test; 类型 test-coverage; 符号 TestMMRSMModel.init, TestAGMMModel.init, _BaseScaledMMModel.init, test_async_tp_pass_replace) : FP8 _scaled_mm 只支持 bf16 行缩放, 因此把模型默认 dtype 和参数化都收敛到 bf16, 删除 FP16 分支。
- .buildkite/test_areas/compile.yaml (模块 CI 配置; 类别 config; 类型 configuration) : 把 SP/AsyncTP 正确性测试的 CI 步骤合并为 2xB200 可选任务, 删除固定执行的 H100/2GPU 步骤。
- tests/compile/correctness_e2e/test_async_tp.py (模块 异步 TP; 类别 test; 类型 test-coverage; 符号 test_async_tp_pass_correctness) : 移除 eager 模式参数化, FP8 模型换成 1B 版本, 减少重复测试点。
- tests/models/registry.py (模块 模型注册; 类别 test; 类型 test-coverage) : 为新的 1B FP8 模型添加 fp8_1b 额外项, 供 async TP 测试使用。

关键符号: SPTestSettings.fast, SPTestSettings.fp8_quant, _build_sp_args, _compare_sp_settings, test_tp_sp_generation, test_tp_sp_generation_prompt_embeds, test_async_tp_pass_correctness, test_async_tp_pass_replace

评论区精华

PR 没有任何人工 review 评论或内联讨论, 唯一的审核记录是 claude[bot] 的自动提示: 由于 PR 来自 fork, 自动化 review 被关闭, 维护者可手动触发 @claude review。虽然零讨论, 但变更清楚呈现了测试矩阵与 CI 门禁的取舍: 用更小的模型、更少的组合换取合并门禁速度, 把完整覆盖交给 nightly。

- fork 的自动化 review 被关闭 (other): 没有人工 reviewer 参与讨论; 由于改动只涉及测试与 CI 配置, 风险可控。

风险与影响

- 风险:
 - 可选化导致门禁覆盖缺口: SP 与 AsyncTP 全量正确性从 merge 前必跑改为 optional: true, 默认不再阻塞 PR, 若 nightly 故障未被及时关注, 回归可能在合并后才暴露。
 - 组合删减可能漏掉 eager/chunked-prefill 与 SP 的交互: chunked_prefill 字段被整体删除, fast 只保留 eager_mode=False, 未来若影响 eager + chunked + SP 的交互路径, 现有测试无法捕获。
 - FP8 模型更换降低负载代表性: Llama-3.2-1B-Instruct-FP8 比原先 8B 模型更小, 权重加载和 FP8 量化路径的集成覆盖弱于 8B 模型; 不过 FP8 算子本身仍有

passes/distributed 单元测试补位。

- compare_all_settings 行为改变：从两两对比变为分组多对一，同一 baseline 下多个变体失败时定位不如两两对比直观。
- .buildkite 步骤移除 source_file_dependencies：按源码变更自动调度该 job 的行为被移除，改为固定 B200 可选任务，触发时机更不透明。
- 影响：覆盖范围：5 个文件、-199/+113 行，全部为测试与 CI 配置，无运行时源码改动。影响主体是编译模块的测试矩阵与 buildkite compile group：SP e2e 测试从固定 2 GPU job + H100 job 变为单个 2xB200 可选 job，AsyncTP e2e 同步缩减。对用户无功能影响；对团队影响集中在 CI 时间成本和回归发现周期——合并门禁更快，但夜间回归发现会延后。
- 风险标记：CI 门禁覆盖收窄，测试矩阵缩减，可选 nightly 任务，无人工审查

关联脉络

- PR #51087 [CI] Add run-all comment commands: 同一 CI 基础设施演进线：把 CI 触发方式做成可配置命令，与本 PR 把重测试移出门禁的目的之一致。
- PR #51079 [ci] Update CI notify workflow with PR write permissions: CI 通知工作流权限调整，同属 CI 基础设施稳定化。
- PR #51015 [CI] Stabilize GLM-5.2 PCP evaluation: CI/ 测试稳定化（GLM eval 配置），说明团队在持续调整测试运行策略。