PR #7242 完整报告

PaddlePaddle/FastDeploy

[CI] Reduce execution time for ngram kernel tests

合并时间：2026-04-08 16:54

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7242

执行摘要

- 一句话：大幅缩减 ngram kernel 测试执行时间，从 6 分钟降至 20 秒，提升 CI 效率。
- 推荐动作：该 PR 展示了 CI 优化中测试时间与覆盖率的典型权衡。建议精读 fastdeploy-bot 的评论，思考如何平衡 CI 效率与测试有效性。对于性能基准测试，可考虑在 CI 配置中排除或仅在特定触发条件下运行，而非修改参数使其失效。

功能与动机

PR body 明确指出：PR #7136 添加的两个测试文件执行时间过长（test_ngram_gpu_kernel.py 约 2 分钟，test_benchmark_ngram_kernel.py 约 4 分钟），影响 CI 效率。其中 test_benchmark_ngram_kernel.py 是性能基准测试，用于验证 GPU kernel 加速比，不适合在 CI 中作为功能验证运行。

实现拆解

修改两个测试文件：

1. test_benchmark_ngram_kernel.py：将 NUM_ITERS 从 1000 降至 1，WARMUP 从 5 降至 1。
2. test_ngram_gpu_kernel.py：
   - 将随机种子测试从 [0,7,123,999] 四个种子缩减为仅 [42]。
   - 将大序列测试的 seq_len 从 128k 降至 16k。
   - 将 latency 测试的迭代次数从 100 降至 1，warmup 从 5 降至 1。
   - 将 latency_scaling 测试的 n_runs 从 50 降至 1。
   - 将 latency_extreme 测试的 seq_len 从 128k 降至 16k。

关键文件：

- tests/spec_decode/test_benchmark_ngram_kernel.py（模块 spec_decode）：性能基准测试文件，NUM_ITERS 从 1000 降至 1，使基准测试失去统计意义，是 review 讨论焦点。
- tests/spec_decode/test_ngram_gpu_kernel.py（模块 spec_decode）：GPU kernel 功能测试文件，缩减迭代次数、测试数据规模和随机种子覆盖，是 CI 时间优化的主要改动点。

关键符号：test_correctness_varied_seeds, test_large_batch_long_seq, test_latency, test_latency_scaling

评论区精华

fastdeploy-bot 提出两点建议：

1. 针对 test_benchmark_ngram_kernel.py：NUM_ITERS=1 后基准测试失去性能测量统计意义，建议从 CI 中移除或跳过，而非修改参数使其失去作用。
2. 针对 test_ngram_gpu_kernel.py：随机种子从 4 个减少到 1 个，降低了对随机数据情况的测试覆盖率，建议至少保留 2-3 个不同种子。 CSWYF3634076 批准了 PR，未回应这些建议。

- 基准测试参数调整导致性能测量失效 (testing): 未采纳建议，PR 保持 NUM_ITERS=1 的修改。
- 随机种子覆盖减少可能降低测试覆盖率 (testing): 未采纳建议，PR 保持仅使用种子 42。

风险与影响

- 风险：1. 测试覆盖风险：随机种子减少可能遗漏某些边界条件，降低 GPU kernel 在不同随机输入下的正确性验证强度。 2. 性能验证失效：test_benchmark_ngram_kernel.py 的 NUM_ITERS=1 使性能基准测试失去统计意义，无法可靠测量 GPU kernel 加速比。 3. 大序列测试缩减：seq_len 从 128k 降至 16k，可能无法充分测试极端规模下的 kernel 稳定性。
- 影响：1. CI 效率提升：测试执行时间从 6 分钟降至 20 秒，显著加快 CI 流水线。 2. 功能验证保留：核心正确性测试仍运行，确保 GPU kernel 基本功能正常。 3. 性能监控弱化：基准测试失去原有性能监控价值，需依赖其他机制（如手动触发）进行性能回归检测。
- 风险标记：测试覆盖降低，性能验证失效，极端场景测试缩减

关联脉络

- PR #7136 [Optimization]【Hackathon 10th Spring No.49】GPU ngram_match: BlockScan Phase 2 -optimized: 本 PR 优化的测试文件正是由 PR #7136 引入，该 PR 添加了 ngram GPU kernel 优化及相应测试。
- PR #7231 [Speculative Decoding] Remove arctic_inference deps: 同属 Speculative Decoding 模块的优化，关注依赖管理和部署简化。
- PR #7201 [OP][Optimization] Remove ENABLE_PREFILL template parameter in multi_query_append_attention_warp1_4_kernel: 同属 GPU kernel 优化相关 PR，涉及模板参数简化和内存布局统一。