

PR #34882 完整报告

sgl-project/sglang

[CI] Trim Qwen3.5 FP8 GB300 performance batches

合并时间: 2026-08-15 07:45

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34882>

执行摘要

- 一句话: 裁剪 Qwen3.5 FP8 GB300 性能批次, 缩短夜间 CI 耗时
- 推荐动作: 值得快速浏览, 不需要精读。如果团队在维护 Blackwell/GB300 性能回归, 可以借鉴按变体裁剪 batch size、抽取公共 accuracy_params、循环内捕获 AssertionError 并聚合并最终统一抛出的模式; 但本 PR 本身只是测试编排小改动, 且真实环境验证未通过, 合入前应确认重跑失败的具体原因。

功能与动机

PR body 明确说明动机是 Reduce redundant performance coverage and runtime in the Qwen3.5 FP8 GB300 nightly test。原先两个变体共用整套 batch size, 会重复覆盖低区分度的性能场景, GB300 4 卡夜间环境资源昂贵, 测试耗时影响整个夜间回归队列, 因此按变体语义裁剪 batch size 集。

实现拆解

1. 新增 PERFORMANCE_BATCH_SIZES 常量字典, 指明 TP4+MTP 只跑 [1, 4], TP4+DP4+DPA+MTP 只跑 [16]。
2. 将 AccuracyTestParams 提取到循环外复用, 精度参数 (mmmu_pro_vision、baseline_accuracy=0.76、max_tokens=32768、temperature=0.7 等) 原样保留, 并在注释中强调这些参数是 ns eval --benchmarks=mmmu-pro:1 隐式基线, 不要随意“简化”, 防止回归标准漂移。
3. 将原先一次 run_combined_tests 同时跑两个变体改为 for 循环逐个变体执行, test_name 追加变体名以区分结果, 并为每个变体传入各自的 batch_sizes。
4. 循环内捕获 AssertionError 并把失败信息追加到 failures 列表, 循环结束后统一抛出聚合错误, 保证第一个变体失败不会中断后续变体的执行和数据收集。
5. 未新增其他配置或部署改动; 作者只做了 pre-commit 和隔离编排检查, 端到端测试依赖 4 GPU GB300 夜间环境, 后续通过 /rerun-test 重跑仍失败 (Run #31847173570), 真实环境验证尚未闭环。

关键文件:

- test/registered/gb300/test_qwen35_fp8.py (模块 夜间测试; 类别 test; 类型 test-coverage; 符号 PERFORMANCE_BATCH_SIZES, test_qwen35_fp8): 唯一的变更文件, 是 Qwen3.5-397B-A17B-FP8 在 4x GB300 NVL4 上的夜间性能与精度回归测试。

本次改动决定哪些性能批次被裁剪、两个变体如何拆分执行、失败如何聚合，直接影响夜间 CI 耗时与性能覆盖。

关键符号: test_qwen35_fp8

评论区精华

该 PR 没有任何 Review 评论 (review_comments_count=0)。唯一的外部反馈来自 Issue 评论区: 作者发起 /rerun-test test/registered/gb300/test_qwen35_fp8.py, github-actions[bot] 回复在 4-gpu-gb300 运行器上 1 个测试仍 ❸ 失败 (Run #31847173570)。作者也在 PR body 中承认 Full end-to-end test not run locally because it requires the 4-GPU GB300 nightly environment。也就是说裁剪逻辑本身没有引发技术争论, 但真实环境验证未通过是主要缺口。

- 裁剪后 GB300 重跑仍失败 (testing): 未解决: 真实 GB300 环境下测试仍失败, PR 内未定位失败原因; 合入前应确认是否为环境或基线问题。

风险与影响

- 风险:
 1. 性能覆盖缩减: TP4+DP4+DPA+MTP 变体只剩 batch 16 一个性能数据点, 若该配置在低 batch 下存在性能退化将无法被发现; 不过这正是裁剪目的, 属于可接受的权衡。
 2. 失败聚合逻辑只捕获 AssertionError, 若变体执行中发生其他异常 (如启动超时、OOM、进程崩溃) 会直接中断循环, 后续变体不会继续执行。
 3. test/registered/gb300/test_qwen35_fp8.py 仍注册为 est_time=7200 的 nightly 任务, 裁剪后实际耗时可能远小于估算, register_cuda_ci 的调度 slot 未随之调整。
 4. 合入前该测试在真实 GB300 环境重跑仍失败, 说明端到端有效性未被证明, 存在合入后夜间回归继续红的风险。
 5. 无产品代码改动, 因此对运行时没有风险。 - 影响: 影响仅限 GB300 nightly 测试队列: 裁剪后 Qwen3.5 FP8 的性能数据点减少 (batch 2/8/32 等不再产出), 若下游基准表依赖这些数据会受影响; 夜间队列整体耗时下降, 释放 4 卡 GB300 资源; 对最终用户无 Runtime 影响; 对团队而言, 该 PR 展示了如何在测试编排层按变体差异化配置性能批次、拆分执行与聚合失败的写法。 - 风险标记: 真实环境验证缺失, 性能覆盖缩减, 失败聚合仅捕获 AssertionError, 夜间队列调度时长未校准

关联脉络

- PR #34863 [Docs] Add GB300 cells and benchmarks for Qwen3.8-27B: 同样围绕 GB300 性能基准维护, 说明团队在持续补充和调整 Blackwell/GB300 平台的性能数据与测试覆盖。
- PR #34019 [SM12x] Default the fused MHC post+pre path on: 在 SM12x/Blackwell 平台默认开启性能优化路径, 会改变 GB300 上模型性能基准的实际行为, 可能影响此类夜间性能测试基线的可比性。