PR #2578 完整报告

radixark/miles

fix: Improve Qwen3 FSDP long CI perf

合并时间：2026-08-17 02:35

原文链接：http://prhub.com.cn/radixark/miles/pull/2578

执行摘要

- 一句话：Qwen3 FSDP CI 改用动态 token 批处理，超时测试降至 2295 秒
- 推荐动作：建议快速浏览该 PR，重点学习其验证方法：先定位 micro-batch 爆炸根因，再通过实测 token 上限边界（32768 通过、65536/100000 OOM）确定安全值，最后用 unpinned main CI 全量验证。这种『根因 - 边界 - 全链路验证』的排查思路值得借鉴；但变更本身仅为测试参数，无需精读。

功能与动机

PR body 明确指出：*release* CI run 31862426780 的 *stage-c-2-gpu-h200* shard 1 超时，test_qwen3_0.6B_fsdp_colocated_2xGPU.py 在 3750 秒时仅完成部分 rollout。根因拆解为三点：*execute* 配置 global batch 256 但未启用动态 token batching；--micro-batch-size 默认为 1，导致每 rank 128 个 micro-batch；actor_train 在 60-rollout 工作负载完成前耗尽预算。

实现拆解

本次变更为单文件测试参数调整，按以下步骤完成：

1. 诊断与定位：在 tests/e2e/short/test_qwen3_0.6B_fsdp_colocated_2xGPU.py 的 *execute*() 中发现训练参数组合导致 micro-batch 数量爆炸，结合 *release* CI 超时日志确认瓶颈。
2. 引入动态 token 批处理：新增 perf_args = "--use-dynamic-batch-size --max-tokens-per-gpu 32768 "，并在 train_args 中 fsdp_args 之后插入 f"{perf_args} "。该参数让训练器按 token 数打包 micro-batch，而不是固定 micro-batch size，从而将每 rank micro-batch 数从 128 降至 3-4。
3. 验证 token 上限边界：在 2xH200 上实测，32768 时仍有至少 4744 MiB 空闲内存；65536 和 100000 均在首次 log-prob 前向中 OOM，因此确定 32768 为安全上限。
4. 验证 CI 全链路：未 pin 的 main 栈 CI run 31932513930 中，专属 job 36m45s 通过，整体 workflow 1h05m12s 完成，28 pass / 7 skipped / 0 fail；CI registry 与 policy 测试 87 项通过，*pre-commit* 通过。
5. 演进过程：提交历史显示曾尝试为 *colocated* FSDP 添加专属 CI selector（commit c6123d2），但最终 revert（commit 48bd44c），保持本 PR 只改动态 token batching，避免影响 nightly 和 *image* 的测试注册范围。

无产品源码、配置或部署配套改动。

关键文件：

- tests/e2e/short/test_qwen3_0.6B_fsdp_colocated_2xGPU.py（模块 E2E 测试；类别 test；类型 test-coverage）：唯一变更文件。在 execute() 中新增 perf_args 并插入 train_args，启用动态 token 批处理，是本次性能修复的全部内容。

关键符号：execute

关键源码片段

tests/e2e/short/test_qwen3_0.6B_fsdp_colocated_2xGPU.py

唯一变更文件。在 execute() 中新增 perf_args 并插入 train_args，启用动态 token 批处理，是本次性能修复的全部内容。

```
# tests/e2e/short/test_qwen3_0.6B_fsdp_colocated_2xGPU.py 中 execute() 的核心配置段

def execute(): # 2xGPU 共置 FSDP 训练的入口，组装并启动训练
    # ...（ckpt_args、rollout_args、optimizer_args、grpo_args 等在上方定义）

    fsdp_args = (
        # SHARDED_STATE_DICT 模式（默认），更新权重 buffer 512 MB
        "--update-weight-buffer-size 536870912 "
    )

    # 本次修复核心：启用动态 token 批处理，将每 rank 的 micro-batch 数
    # 从 128（global batch 256 / micro-batch 1）降到 3-4 个，
    # 避免 actor 前向在 60 个 rollout 完成前耗尽 CI 时间预算。
    # 32768 是 2xH200 上经实测的 token 上限：更高（65536/100000）首轮 log-prob 即 OOM。
    perf_args = "--use-dynamic-batch-size --max-tokens-per-gpu 32768 "

    ci_args = (
        "--ci-test "
        "--ci-disable-kl-checker "
        "--ci-metric-checker-key eval/gsm8k "
        "--ci-metric-checker-threshold 0.71 " # 60 步下的宽松阈值
    )

    misc_args = "--actor-num-nodes 1 " "--actor-num-gpus-per-node 2 " "--colocate " "--train-
    backend fsdp "

    train_args = (
        f"{ckpt_args} "
        f"{rollout_args} "
        f"{optimizer_args} "
        f"{grpo_args} "
        f"{sglang_args} "
        f"{U.get_default_wandb_args(__file__)} "
        f"{eval_args} "
```

```
        f"{fsdp_args} "
        f"{perf_args} " # 插入 train_args，替代固定 micro-batch 的默认调度
        f"{ci_args} "
        f"{misc_args} "
    )

    U.execute_train(
        train_args=train_args,
        num_gpus_per_node=2,
        megatron_model_type=None,
    )
```

评论区精华

该 PR 没有实质性的 reviewer 讨论：唯一的 review 来自 claude[bot]，仅提示此仓库配置为手动 review，并告知可评论 @claude review 触发。PR body 的 Review Focus 指出 perf_args 是唯一需要关注的点：它改变了调度方式但不减少 workload 或验收标准。作者在 Issue 评论中补充了 unpinned main CI 的全量验证结果（28 pass / 7 skipped / 0 fail），作为主要说服依据。

- 暂无高价值评论线程

风险与影响

- 风险：风险集中在以下几点：
  - OOM 边界依赖硬件：32768 的 token 上限是在 2xH200 上实测的，若 CI 环境更换 GPU 型号或显存缩水，可能复现 OOM，需要重新校准该值。
  - 测试参数与生产配置漂移：该参数对齐了生产 Qwen3 FSDP launcher 的 9216 token 配置（后提升至 32768），但测试脚本与生产启动器若后续不同步演进，可能掩盖真实性能回归。
  - 无源码逻辑覆盖：本次仅调整测试参数，不涉及产品代码，因此不影响训练框架本身，但也不增加对动态 batching 功能的单元测试覆盖。
  - 回归风险：改动极小且经验证，风险低；但若 CI 时间预算后续收紧，仍需重新评估。
  - 影响：对用户无直接影响；对团队而言，Qwen3 0.6B FSDP colocated 2xGPU 的 E2E 测试从频繁超时变为稳定通过（2295 秒 vs 3750 秒上限），显著降低 release CI 的等待和重试成本。同时，该 PR 明确了 2xH200 在 FSDP 共置模式下的动态 token 内存边界，为后续类似 CI 测试的参数选择提供了可复用的经验数据。
  - 风险标记：OOM 边界依赖硬件，仅测试参数变更，无源码逻辑覆盖

关联脉络

- PR #2538 release: miles version release workflow: PR body 中引用的超时发生在该 release 工作流的 CI run 中，本次修复即为该工作流中的 Qwen3 FSDP colocated 测试恢复时间余量。

- PR #2219 [feat] Add training log-prob reuse to skip the redundant forward-only pass: 同为 FSDP/PPO 训练性能优化方向，但 2219 在产品逻辑层减少前向计算，2578 在 CI 测试层调整调度参数，属于同一性能谱系的不同层面。