PR #7172 完整报告

PaddlePaddle/FastDeploy

fix MTP bugs in TP and overlap

合并时间：2026-04-03 14:19

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7172

执行摘要

- 一句话：修复 MTP 在 TP 并行和重叠调度场景下的两个关键 bug
- 推荐动作：该 PR 值得精读，特别是关注推测解码在 TP 并行下的输出同步机制设计，以及重叠调度中 token 预测算法的优化思路。建议重点查看：1）rank 检查如何避免冗余通信；2）token 预测公式从复杂计算简化的设计权衡。

功能与动机

根据 PR body 描述，需要修复 MTP 在 TP>1 和 overlap scheduling 场景下的两个问题：1）在 TP 场景下，非 rank 0 的进程重复发送 sampling 输出到消息队列，造成冗余通信；2）overlap scheduling 中预测下一批次 token 数量的计算逻辑错误，影响调度准确性。这些 bug 会影响 MTP 在分布式环境下的性能和正确性。

实现拆解

实现分为两个关键修改点：

1. 在推测解码的输出保存函数中（speculate_save_output.cc 和 speculate_save_output_with_topk.cc），添加 rank_id>0 的检查，跳过非 rank 0 的进程，避免重复发送采样输出到消息队列。
2. 在 GPU 模型运行器的预测函数（_predict_next_launch_token_num）中，修正 token 数量预测公式：从基于序列长度和步数的复杂计算改为简单的 next_real_bsz * token_num_one_step，即批次大小乘以每步 token 数。

关键文件：

- custom_ops/gpu_ops/speculate_decoding/speculate_save_output.cc（模块 Speculative Decoding）：修改了推测解码的输出保存逻辑，添加 rank 检查避免 TP 并行下的冗余通信
- custom_ops/gpu_ops/speculate_decoding/speculate_save_output_with_topk.cc（模块 Speculative Decoding）：同样修改了带 topk 的输出保存逻辑，确保 TP 并行下的一致性
- fastdeploy/worker/gpu_model_runner.py（模块 Scheduler）：修正了重叠调度中预测下一批次 token 数量的计算逻辑，影响调度性能

关键符号：SpeculateSaveWithOutputMsg, SpeculateSaveOutMmsgTopK, _predict_next_launch_token_num

评论区精华

review 讨论主要集中在 fastdeploy/model_executor/pre_and_post_process.py 文件的修改上：

- Sunny-bot1 指出“不能提前 return，不然这一行执行不到，重调度可能会有问题”，担心提前 return 会跳过清理操作。
- fastdeploy-bot 进一步分析指出提前 return 会跳过第 592 行的 share_inputs["last_preempted_idx"][:] = 0 清理操作，可能导致后续批次使用脏数据，建议将清理操作移到函数开头或在 return 前执行清理。
- 从最终提交的文件列表看，这个讨论可能导致了实现方案的调整，最终修改的是 custom_ops/gpu_ops/speculate_decoding/ 下的 C++ 文件而非 Python 文件，说明设计决策发生了变化。
- 提前 return 可能跳过清理操作 (correctness): 从最终提交看，可能调整了实现方案，未在 pre_and_post_process.py 中修改，而是集中在 C++ 文件中

风险与影响

- 风险：风险点包括：
  1. 核心路径变更：修改了推测解码的核心输出保存逻辑（speculate_save_output.cc 等），如果 rank 检查逻辑错误，可能导致 rank 0 未正确发送输出或非 rank 0 错误发送输出。
  2. 调度逻辑变更：修改了 _overlap scheduling 的 token 预测逻辑（_predict_next_launch_token_num），如果新公式计算错误，可能影响调度性能和资源利用率。
  3. 清理操作缺失：根据 review 讨论，在 pre_and_post_process.py 中的提前 return 可能跳过清理操作，但最终提交未包含该文件修改，说明此风险可能已通过其他方式解决或仍存在。
- 影响：影响范围：
  1. 对用户：修复后 MTP 在 TP 并行和重叠调度场景下能正确工作，提升分布式推理的稳定性和性能。
  2. 对系统：避免非 rank 0 进程的冗余通信，减少消息队列压力；修正调度预测逻辑，提高资源利用率。
  3. 对团队：涉及推测解码、TP 并行和调度器多个模块，需要相关开发者关注变更。影响程度中等，主要影响使用 MTP+TP+overlap scheduling 的特定场景。
- 风险标记：核心路径变更，调度逻辑变更，清理操作缺失风险

关联脉络

- PR #7133 Revert "[BugFix][Speculative Decoding] Correct index calculation in speculate decoding operators": 都涉及推测解码算子的 bug 修复，属于同一功能模块
- PR #7121 [BugFix][Speculative Decoding] Correct index calculation in speculate decoding operators: 都涉及推测解码的 bug 修复，可能共享相似的技术背景

- PR #6993 [XPU] Refactor pre process: 都涉及推测解码的前处理 / 后处理逻辑优化