PR #21519 完整报告

sgl-project/sglang

[Bugfix] Fix incorrect dp-attention parallel info in bench_one_batch

合并时间：2026-04-03 11:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21519

执行摘要

本 PR 修复了 bench_one_batch 性能分析工具中 DP-Attention 并行元数据传递错误，通过使用 get_attention_tp_size 函数并添加 attn_cp_size 参数，确保在启用 DP-Attention 时性能分析路径与正常调度器路径一致，从而提升分析准确性。

功能与动机

动机源于 bench_one_batch 在启用 DP-Attention 时未正确传递注意力并行元数据，导致性能分析不一致。PR body 明确指出："attn_tp_size is hardcoded to 1 and attn_cp_size is not forwarded, which makes the DP-Attention preparation path inconsistent with the normal scheduler path"。这在使用 GLM-4.7-Flash 等模型进行性能分析时会造成问题。

实现拆解

仅修改了 python/sglang/bench_one_batch.py 文件，具体改动在 _maybe_prepare_mlp_sync_batch 函数中：

- 将 attn_tp_size 从硬编码 1 改为调用 get_attention_tp_size() 动态获取。
- 添加 attn_cp_size=model_runner.attn_cp_size 参数传递。 关键代码变更如下：

评论区精华

review 讨论中，alexnails 建议使用现有函数简化代码：

> "can't this just be an import and use: get_attention_tp_size" 作者 lviy 回应并测试后更新，采纳了这一建议，体现了对代码设计和一致性的重视。

风险与影响

风险较低，修复了已知错误，但需确保 get_attention_tp_size 函数在 bench_one_batch 上下文中正常工作。影响限于性能分析工具用户，提升分析准确性，对系统其他部分无影响。

关联脉络

与 PR #21840（调度器批次修复）可能共享并行计算逻辑，反映了项目对调度和性能分析一致性的持续优化。近期历史 PR 中多见 run-ci 标签，表明 CI 测试是重要环节，本 PR 也通过 run-ci 确保改动稳定性。