

PR #27721 完整报告

sgl-project/sglang

Add TP server GPU process regression test

合并时间: 2026-06-10 07:25

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/27721>

执行摘要

- 一句话: 新增 TP 服务进程 GPU 上下文回归测试
- 推荐动作: 建议精读此测试的实现思路, 特别是 `_wait_for_server_gpu_processes` 中的稳定等待策略以及如何利用进程树过滤 GPU 进程; 该模式可用于其他与 GPU 资源分配相关的回归测试。

功能与动机

确保 TP 模式下服务器父进程不会在 fork worker 之前意外初始化 CUDA 上下文, 从而浪费 GPU 显存或导致后续 fork 异常。PR body 提到通过人为注入 `torch.ones(10).cuda()` 到 `ServerArgs.__post_init__` 可验证测试能有效捕获此类问题。

实现拆解

1. CI 注册与依赖: 使用 `register_cuda_ci` 注册测试, 指定 `stage="base-b"`、`runner_config="2-gpu-large"` 和 `est_time=120`, 确保该测试在 2 GPU 大节点上运行。
2. 测试夹具: `setUpClass` 启动一个 TP=2 的 SGLang 服务器, 禁用 CUDA 图和分段 CUDA 图以减少干扰; `tearDownClass` 使用 `kill_process_tree` 清理进程树。
3. 核心测试方法: `test_tp_server_has_only_worker_gpu_processes` 首先检查 `nvidia-smi` 可用性, 然后调用 `_wait_for_server_gpu_processes` 等待直至出现至少 `tp_size` 个属于服务器进程树的 GPU 进程并稳定 3 秒, 最后断言父进程不在 GPU 进程列表中且 GPU 进程数量恰好等于 `tp_size`。
4. 辅助方法: `_server_process_tree_pids` 通过 `psutil` 递归获取进程树 PID 集合; `_query_gpu_processes` 执行 `nvidia-smi --query-compute-apps` 解析 CSV 输出; `_format_rows` 用于格式化错误消息。
5. 测试配置: 服务器使用 `DEFAULT_SMALL_MODEL_NAME_FOR_TEST`、`mem-fraction-static=0.70` 及其他禁用加速的选项, 以保持轻量。

关键文件:

- `test/registered/core/test_no_extra_forked_cuda_context.py` (模块 回归测试; 类别 `test`; 类型 `test-coverage`; 符号 `TestTPServerGPUProcesses`, `setUpClass`, `tearDownClass`, `test_tp_server_has_only_worker_gpu_processes`): 该文件是本次 PR 唯一变更, 新增了完整的回归测试类, 覆盖 TP 服务进程 GPU 上下文的正确性断言。

关键符号: setUpClass, tearDownClass, test_tp_server_has_only_worker_gpu_processes, _wait_for_server_gpu_processes, _server_process_tree_pids, _query_gpu_processes, _format_rows

评论区精华

该 PR 无 review 讨论, 仅包含自动 CI 评论和作者触发的 /tag-and-rerun-ci 命令。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低。该 PR 仅新增测试文件, 不修改任何生产代码。测试需要 2 块 GPU 和 nvidia-smi 工具, 所注册的 CI 阶段也明确要求 2 卡大型节点。若 CI 环境缺失 nvidia-smi 或 GPU, 测试会正确 skip。潜在风险是测试运行时长时间占用两卡资源, 但 est_time=120 已在 CI 调度中考虑。
- 影响: 对用户无直接影响。对 CI 系统, 将新增一个 base-b 阶段的回归测试, 为 TP 服务进程的 GPU 上下文管理提供基线保障。测试命名和注册方式可被后续类似回归测试复用。
- 风险标记: 新增回归测试, 无生产代码变更, 依赖 nvidia-smi, CI 消耗 2 卡资源

关联脉络

- 暂无明显关联 PR