PR #22083 完整报告

sgl-project/sglang

dp: add profile req hook

合并时间：2026-04-04 11:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22083

执行摘要

- 一句话：修复 DP 注意力模式下 ProfileReq 分发缺失导致的死锁问题。
- 推荐动作：该 PR 值得快速浏览，以了解 DP 注意力模式下控制消息分发的设计模式。关注点：1) 分发器如何映射消息类型到处理方法。2) send_to_all_workers 与 send_control_message 的差异。3) 通信器扇出预期与分发策略的匹配。

功能与动机

PR body 明确指出：当启用 enable_dp_attention 时，ProfileReq 没有明确的分发器条目，会回退到 send_control_message() 方法。在 tp_size=8、dp_size=2 的配置下，该方法仅将请求发送给 workers[::8]（即仅 worker[0]），而 tokenizer 的 _Communicator 期望 dp_size=2 个响应。由于只有一个调度器响应，通信器会永远阻塞等待第二个响应，导致 /start_profile HTTP 处理程序死锁。

实现拆解

在 data_parallel_controller.py 的 init_dispatcher 方法中，向 dispatcher 列表添加 (ProfileReq, self.send_to_all_workers) 条目。这确保 ProfileReq 消息通过 send_to_all_workers 方法发送给所有工作进程，而不是仅发送给部分进程。

关键文件：

- python/sglang/srt/managers/data_parallel_controller.py（模块 srt/managers）：这是唯一修改的文件，包含 DP 控制器的核心分发逻辑。添加 ProfileReq 到分发器解决了死锁问题。

关键符号：init_dispatcher, send_to_all_workers

评论区精华

由于 review 评论为空，没有公开的技术讨论。但从关联 Issue 的评论中可见，维护者 hnyls2002 通过 /rerun-test 命令验证了修复效果：首先运行 registered/profiling/test_start_profile.py 测试，然后运行 registered/distributed/test_dp_attention.py 测试，两者均通过，表明修复解决了死锁问题。

- ProfileReq 分发缺失导致死锁 (correctness): 通过添加分发器条目，确保 ProfileReq 发送给所有工作进程。

风险与影响

- 风险：风险较低：1) 变更仅涉及消息分发逻辑，不修改核心算法或数据结构。2) 单文件改动，仅添加两行代码，影响范围有限。3) 潜在风险是 send_to_all_workers 可能增加网络开销，但 ProfileReq 本身是低频控制消息，影响可忽略。4) 缺少单元测试直接验证此分发逻辑，但已有集成测试覆盖。
- 影响：影响范围：1) 用户：修复了启用 DP 注意力时的死锁问题，确保 /profile 端点正常工作，提升系统可靠性。2) 系统：解决了分布式配置下的调度阻塞，避免资源泄漏。3) 团队：代码变更极小，易于理解和维护，但揭示了 DP 注意力模式下的消息分发机制需保持一致性。
- 风险标记：缺少测试覆盖

关联脉络

- PR #21917 Fix DP attention worker port binding for IPv6 support: 同样修改 data_parallel_controller.py，涉及 DP 注意力模式下的网络通信问题。
- PR #20273 fix: pause_generation should not populate running_batch on prefill nodes: 同为调度相关的 bugfix，涉及请求处理和死锁避免。