PR #20908 完整报告

sgl-project/sglang

fix(PD): respect pause_generation in disagg event loops

合并时间：2026-04-13 09:07

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20908

执行摘要

此 PR 修复了 disaggregation 事件循环中 pause_generation 功能失效的 bug，通过在解码和预填充循环中添加暂停检查，确保调度器在暂停时停止生成，并添加了回归测试验证。改动小但关键，恢复了系统的可控制性。

功能与动机

动机源于 issue #20906，报告在 disaggregation 的 *decode.py* 和 prefill.py 中缺少 self._engine_paused 检查，导致调用 /pause_generation 时生成不暂停。PR 旨在恢复这一关键控制功能，引用 issue 描述：'Generation does not pause when /pause_generation is called.'

实现拆解

- 解码模块（*decode.py*）：在 *event_loop_normal_disagg_decode* 和 *event_loop_overlap_disagg_decode* 函数中添加 if self._engine_paused: continue，跳过批次运行当引擎暂停时。
- 预填充模块（prefill.py）：在 *event_loop_normal_disagg_prefill* 和 *event_loop_overlap_disagg_prefill* 函数中添加类似检查，并调用 self.process_disagg_prefill_inflight_queue() 以确保 bootstrap 和 transfer 队列的处理在暂停时仍能完成。
- 测试模块：在 test_disaggregation_basic.py 中添加 test_pause_resume_in_place 测试，模拟暂停场景，验证请求在暂停期间无进展，恢复后正常完成。

评论区精华

主要讨论集中在测试文件的放置。reviewer hnyls2002 指出：

> 'Do not add new E2E test files. Just put the pause generation test inside test_disaggregation_basic.'

作者据此调整，将测试集成到现有文件中，体现了团队对测试维护性和一致性的重视。没有其他技术争议。

风险与影响

风险：改动简单，风险较低。但需注意：1) 在 prefill.py 中添加的队列处理调用可能引入轻微性能开销；2) 修改调度循环需确保不影响其他功能，回归测试覆盖了基本场景。 影响：对用户，pause_generation 在 disaggregation 中现在正常工作，提升调试能力；对系统，修复了调度逻辑，避免资源浪费；对团队，加强了测试覆盖。

关联脉络

与 PR #22647（提取 pause_resume_in_place 测试工具包）相关，表明团队在系统调度和控制功能上持续演进。近期历史 PR 如 #22597（修复 SWA 输入长度限制）也涉及调度改进，显示该模块的重要性。