PR #20273 完整报告

sgl-project/sglang

fix: pause_generation should not populate running_batch on prefill nodes

合并时间：2026-04-04 07:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/20273

执行摘要

- 一句话：修复 pause_generation 在 prefill 节点泄漏请求导致调度停顿的 bug。
- 推荐动作：建议工程师精读此 PR，特别是 scheduler.py 中 pause_generation 方法的变更，以理解调度器中 prefill 节点的特殊处理逻辑和避免泄漏的设计决策。关注条件检查的添加位置和原因，以及测试如何模拟泄漏场景验证修复。对于学习调度机制和边界条件处理有参考价值。

功能与动机

根据关联 Issue #20272，用户报告在 disaggregated 模式下使用 pause_generation 时，当 running_batch 为空时会引发 AttributeError 崩溃。PR body 进一步指出，现有逻辑会将 prefill 请求泄漏到 running_batch，这些请求永远不会被清理，最终导致调度器在达到最大运行请求数时停顿。修复目标是确保调度器的稳定性和正确性，防止资源浪费和系统停滞。

实现拆解

实现分为两部分：首先，在 python/sglang/srt/managers/scheduler.py 的 pause_generation 方法中，添加条件检查 self.disaggregation_mode != DisaggregationMode.PREFILL 来跳过 prefill 节点的合并，防止泄漏；同时，在 retract 模式路径添加 not self.running_batch.is_empty() 检查，避免空 batch 时的崩溃。其次，在 test/registered/disaggregation/test_disaggregation_basic.py 中添加新的测试类 TestDisaggregationPauseResumePrefillLeak，覆盖 retract 模式下的 pause/resume 场景，通过异步客户端模拟请求验证修复效果。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 srt/managers）：修改核心调度方法 pause_generation，添加 disaggregation_mode 检查和 running_batch 空检查，修复泄漏和崩溃问题
- test/registered/disaggregation/test_disaggregation_basic.py（模块 test/disaggregation）：添加回归测试类 TestDisaggregationPauseResumePrefillLeak，覆盖 retract 模式下的 pause/resume 场景，验证修复效果

关键符号：pause_generation

评论区精华

review 中仅有一条来自 gemini-code-assist[bot] 的评论，指出修复不完整，建议在合并前检查 self.last_batch 是否为空，以防止不一致状态。评论强调：'If self.last_batch is empty, its internal state can be inconsistent... which can cause merge_batch to fail.' 但提交历史显示，后续提交通过迭代完善修复（如添加测试和额外检查），未直接回应此评论。讨论焦点在于正确性，结论是通过多次提交解决了核心问题，但潜在边缘条件可能仍需关注。

- 检查 last_batch 是否为空以防止不一致状态 (correctness): 提交历史通过迭代完善修复，但未明确采纳建议；PR 最终合并，表明问题已基本解决

风险与影响

- 风险：风险主要在于边界条件处理：变更涉及调度核心逻辑，如果条件检查不充分，可能引入新的崩溃或泄漏，尤其是在 prefill 节点和空 batch 场景下。例如，review 评论指出的 last_batch 为空问题未在代码中显式处理，可能存在隐患。文件 scheduler.py 的修改较集中，但影响调度路径，需确保所有模式（如 in_place、retract）和 disaggregation 状态都得到正确覆盖。测试覆盖增加降低了回归风险，但新测试可能未涵盖所有并发或边缘情况。
- 影响：对用户：修复了使用 pause_generation 时的崩溃问题，提升了系统在 disaggregated 模式下的稳定性和可用性。对系统：防止请求泄漏，确保调度器正常运作，避免因泄漏导致的资源浪费和性能下降，特别是当 --max-running-requests 限制较小时。对团队：增加了回归测试覆盖，有助于未来维护和类似 bug 的预防，但需注意测试可能增加 CI 运行时间。
- 风险标记：核心路径变更，边界条件处理，测试覆盖增加

关联脉络

- PR #20272 [Bug] pause_generation with mode in_place or retract crashes when running_batch is empty: 直接关联的 Issue，描述了本 PR 修复的 bug，包括崩溃和泄漏问题，为修复提供背景和验证依据