PR #32379 完整报告

sgl-project/sglang

Fix SWA admission livelock on cached-prefix resumes

合并时间：2026-07-25 22:37

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32379

执行摘要

- 一句话：修复 SWA 缓存前缀恢复时的 admission livelock
- 推荐动作：该 PR 值得精读，因为揭示了一个精细的预算 double-count 问题，以及修复中如何平衡缓存前缀与 decode headroom。设计决策（capped at window）清晰且有测试佐证。

功能与动机

在 hybrid-SWA 模型且 SWA KV pool 大小约为两个 sliding window 且使用 LPM 调度时，缓存前缀 >= 一个窗口的请求恢复时，调度器会 100% CPU 循环拒绝同一请求，导致 GPU 空闲。根因是 PrefillAdder._swa_budget_for_req 对每个候选都预留整个 sliding window 的 decode headroom，但 cached-prefix resume 的窗口已被锁定的前缀填充，二次预留导致 double-count。

实现拆解

1. 修改 _swa_reserved_tokens 方法：将 decode headroom 从常值 sliding_window_size 改为 min(extend_input_len + max_new_tokens, window)，避免对缓存前缀的窗口 double-count。
2. 新增 _swa_new_tokens 方法：以正确的顺序（先减后 clip）计算剩余 decode token，防止 clip 后减为零导致欠预留。
3. 更新 _swa_budget_for_req 和 _swa_chunk_cap 调用：传递 max_new_tokens 参数，使预留计算基于实际新增 token。
4. 扩展单元测试：在 test_prefill_adder.py 中增加了 _swa_budget_for_req 的新用例（short resume, multichunk over-window）、admission 边界测试（验证 cached-prefix resume 在约 2-window pool 下可被接受）以及 clip 顺序测试。

关键文件：

- python/sglang/srt/managers/schedule_policy.py（模块 调度器；类别 source；类型 core-logic；符号 _swa_reserved_tokens, _swa_budget_for_req, _swa_new_tokens, _swa_chunk_cap）：核心调度策略实现，修改了 SWA 预算计算方法，修复 livelock
- test/registered/unit/managers/test_prefill_adder.py（模块 单元测试；类别 test；类型 test-coverage；符号 test_swa_admission_admits_short_cached_resume_at_two_window_pool, test_swa_new_tokens_clamps_remaining_not_total, test_swa_budget_for_req）：新增用例验证 livelock 修复和新预算公式

关键符号：_swa_reserved_tokens, _swa_budget_for_req, _swa_new_tokens, _swa_chunk_cap

关键源码片段

test/registered/unit/managers/test_prefill_adder.py

新增用例验证 livelock 修复和新预算公式

```
def test_swa_admission_admits_short_cached_resume_at_two_window_pool(self):
    # Livelock regression: at SWA pool ~= 2 sliding windows, a cached-prefix
    # resume matches >= 1 window and has only a short uncached tail + decode.
    # Pre-fix constant-window reservation charged a second full window, causing
    # rejection every scheduler iteration -> 100% CPU, idle GPU.
    # After fix: min(extend+decode, window) admits it.
    WINDOW, PAGE, REM_SWA = 128, 8, 100
    PREFIX, EXTEND = 200, 16 # cached prefix > window, short extend
    adder, _ = self._build_hybrid_swa_chunked_req(
        page_size=PAGE, rem_swa=REM_SWA
    )
    # mock add_one_req with proper prefix hit length
    with patch.object(adder, 'tree_cache') as mock_tree:
        mock_tree.sliding_window_size = WINDOW
        mock_tree.match_prefix.return_value = (PREFIX, [])
        req = self._create_req(extend_input_len=EXTEND, max_new_tokens=10)
        result = adder.add_one_req(req, has_chunked_req=False, truncation_align_size=None)
        self.assertEqual(result, AddReqResult.ADDED)
```

评论区精华

PR 提交后，作者发现 H100 上的端到端 livelock 测试 test_swa_admission_livelock.py 不稳定，多次 rerun 后仍失败。最终决定移除该 e2e 测试，由新增的单元测试 test_swa_admission_admits_short_cached_resume_at_two_window_pool 和 test_swa_new_tokens_clamps_remaining_not_total 覆盖核心逻辑，CI 全部通过。

- H100 端到端 livelock 测试失败 (testing): 移除 e2e 测试，保留单元测试 test_swa_admission_admits_short_cached_resume_at_two_window_pool 和 test_swa_new_tokens_clamps_remaining_not_total 确保正确性。
- 单元测试覆盖充分性 (testing): 单元测试覆盖充分，无需保留不稳定的 e2e 测试。

风险与影响

- 风险：
  1. 回归风险：修改了 SWA 预算计算，可能影响非 cached-prefix 请求的 admission 行为。但 long decode 请求仍预留 full window（min 上限），且测试覆盖了多种边界情况。
  2. clip 顺序变化：_swa_new_tokens 的计算顺序与 add_one_req 一致，不会导致新的欠预留或 OOM。

3. 测试覆盖：新增的单元测试验证了关键场景，*e2e* 测试被移除，但单元测试足以保障正确性。

- 影响：
  - 用户：修复了使用 hybrid-SWA 并且缓存命中率高的场景下的调度器死锁，提升系统可用性和 GPU 利用率。
  - 系统：修改仅影响 SWA 预算计算路径，其他 KV 管理逻辑不变。
  - 团队：需要关注新计算的正确性，尤其是与现有 _swa_chunk_cap 的联动。
  - 风险标记：核心调度逻辑变更，double-count 修复，边界 *case* 敏感

关联脉络

- 暂无明显关联 PR