

PR #22495 完整报告

sgl-project/sglang

Add page_size to admission token budget check

合并时间: 2026-04-10 16:16

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/22495>

执行摘要

- 一句话: 修复 SWA 准入检查中页面大小预算不一致导致的过度准入问题。
- 推荐动作: 该 PR 值得精读, 尤其关注 SWA 内存预算管理中的页面对齐开销处理。设计决策简单直接, 但揭示了预算检查与实际扣除一致性的重要性。建议工程师理解 `_update_prefill_budget` 与 `add_one_req` 的协作逻辑, 以及页面大小在内存分配中的作用。

功能与动机

根据 PR body 描述, 当前准入检查 (`PrefillAdder.add_one_req`) 未考虑页面大小对齐开销, 而预算扣除 (`_update_prefill_budget`) 已计入此开销。这导致准入检查比实际预算扣除更乐观, 当内存池接近满载时, 请求可能通过准入但分配器失败 (返回 `None`), 触发不必要的重试。PR body 明确指出: “Without this, the admission check is more optimistic than the actual budget deduction. When the pool is nearly full, a request can pass admission but cause the allocator to fail (returning `None`), triggering an unnecessary retraction.”

实现拆解

仅修改一个文件: `python/sglang/srt/managers/schedule_policy.py` 中的 `add_one_req` 方法。关键改动是在计算 `total_tokens` 时添加 `self.page_size`, 以匹配 `_update_prefill_budget` 中已考虑的页面对齐开销。具体变更: 在原有 `total_tokens` 计算 (基于 `req.extend_input_len` 和 `max_new_tokens`) 基础上, 增加 `+ self.page_size`, 确保准入检查与预算扣除逻辑一致。

关键文件:

- `python/sglang/srt/managers/schedule_policy.py` (模块 `scheduling`): 唯一修改的文件, 包含 SWA 准入检查的核心逻辑 `add_one_req` 方法, 修复预算不一致问题。

关键符号: `add_one_req`

评论区精华

无 review 评论, PR 由作者直接合并。从 PR body 和 issue 评论看, 作者通过 CI 测试验证了修复: 运行了 `test_prefill_adder.py` 等测试, 确保变更不影响现有功能。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低，但需注意：
 1. 核心路径变更：修改了调度策略中的准入检查逻辑，可能影响所有 SWA 相关请求的准入决策。
 2. 兼容性：变更使准入检查更严格，可能拒绝之前可通过的请求，需确保不影响正常负载下的吞吐量。
 3. 测试覆盖：尽管运行了相关测试，但未明确是否有针对边缘情况（如 `page_size` 变化、内存池满载）的专项测试。
- 影响：影响范围：
 1. 用户：修复后，SWA 内存管理更准确，减少因分配失败导致的重试，提升系统稳定性。
 2. 系统：防止内存池过度准入，避免分配器失败和潜在的性能抖动。
 3. 团队：变更较小，但涉及核心调度模块，需在部署后监控准入率和内存使用情况。 - 风险标记：核心路径变更，预算逻辑调整

关联脉络

- PR #22470 Fix SWA eviction boundary and page-align chunked prefill: 同样涉及 SWA 和页面对齐问题，修改了 `schedule_batch.py` 和 `swa_radix_cache.py`，可能共享类似的内存管理上下文。
- PR #22239 [sgl] Fix mamba tracking calculation in spec dec: 同属调度和推测解码相关 bugfix，涉及核心路径的边界错误修复。