PR #25126 完整报告

sgl-project/sglang

Fix scheduler admission for near-full KV requests

合并时间：2026-05-14 06:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25126

执行摘要

- 一句话：修复接近满 KV 请求无法被调度的问题
- 推荐动作：值得精读，尤其是涉及调度准入边界条件的逻辑。建议在后续开发中考虑将准入预算计算提取为公共函数，避免重复。

功能与动机

当 KV 预算接近上限时，请求可能被接收进入等待队列，但随后永远无法通过预填充准入（PrefillAdder），从而阻塞队列，最终导致健康检查失败。具体复现场景：GLM-5.1-FP8，TP8，max_total_num_tokens=64384，随机负载约 61200/3000 和 61200/6800。

实现拆解

1. 定位问题：在 scheduler.py 的 init_req_max_new_tokens 方法中，原有的 max_new_tokens 仅受 max_req_len - input_len - 1 限制。
2. 分析预填充准入逻辑：PrefillAdder 的准入预算为 ceil_page(input_len) + max_new_tokens + page_size < max_total_num_tokens。
3. 修改约束：在 init_req_max_new_tokens 中增加 self.max_total_num_tokens - paged_input_len - self.page_size - 1 这一上限，其中 paged_input_len 为页面对齐后的输入长度。同时使用 max(0, min(...)) 确保 max_new_tokens 不会变为负数。
4. 保持向后兼容：原有的 max_req_len 约束仍然保留，以确保模型支持的最大生成长度不被突破。

关键文件：

- python/sglang/srt/managers/scheduler.py（模块 调度器；类别 source；类型 core-logic；符号 init_req_max_new_tokens）：核心修复文件。在 init_req_max_new_tokens 方法中增加了与 PrefillAdder 一致的 KV 预算约束，确保请求在进入等待队列前就能通过预填充准入检查。

关键符号：init_req_max_new_tokens

关键源码片段

python/sglang/srt/managers/scheduler.py

核心修复文件。在 init_req_max_new_tokens 方法中增加了与 PrefillAdder 一致的 KV 预算约束，确保请求在进入等待队列前就能通过预填充准入检查。

```
def init_req_max_new_tokens(self, req):
    input_len = len(req.origin_input_ids)
    # Keep this bound consistent with PrefillAdder's admission budget:
    # ceil_page(input_len) + max_new_tokens + page_size must be strictly
    # smaller than max_total_num_tokens. Otherwise a request can be accepted
    # into the waiting queue but can never be scheduled, blocking the queue
    # and eventually making health checks fail.
    paged_input_len = -(-input_len // self.page_size) * self.page_size # 向上取整到页大小
    req.sampling_params.max_new_tokens = max(
        0,
        min(
            (
                req.sampling_params.max_new_tokens
                if req.sampling_params.max_new_tokens is not None
                else 1 << 30
            ),
            self.max_req_len - input_len - 1, # 模型支持的最大生成长度
            self.max_total_num_tokens - paged_input_len - self.page_size - 1, # KV 预算约束
        ),
    )
```

评论区精华

无实质性讨论。评论仅包含机器人提示和请求审核。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅修改了初始化请求最大生成 token 数的逻辑，不涉及模型前向计算、采样或输出。但需要确保 self.page_size 和 self.max_total_num_tokens 在调用该函数时已正确初始化。新增的 paged_input_len 计算与 PrefillAdder 中的逻辑保持一致，若未来改动 PrefillAdder 的准入预算计算，需同步更新此函数。
- 影响：用户影响：修复了特定负载下服务不可用的 bug，提升服务稳定性。系统影响：无性能回归，仅增加一次整数计算。团队影响：维护者需保持此函数与 PrefillAdder 预算逻辑的一致性。
- 风险标记：核心路径变更

关联脉络

- PR #25062 [PD Disaggregation] Fix priority scheduling in PD disaggregation mode: 同为调度器相关的 bugfix，修改 scheduler.py，涉及准入逻辑。