PR #25741 完整报告

sgl-project/sglang

[Scheduler] fix chunked prefill not always being full

合并时间：2026-05-21 06:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25741

执行摘要

- 一句话：修复分段预填充在批次非空时错误拒绝新请求
- 推荐动作：该 PR 改动小（+16/-2），但解决了关键性能问题，值得阅读源码以理解调度预算交互。特别关注 add_one_req 中的条件演变和 rem_chunk_tokens 的作用。

功能与动机

在分段预填充模式下，如果当前预填充批次中已有请求，调度器会因剩余输入 token 不足而拒绝新请求，即使这些请求可以被分块适配剩余预算。这导致预填充批次无法充分利用，留下空闲 token 槽位，降低了整体吞吐量。PR body 中的日志显示修改前『full token usage』仅为 0.08-0.11，队列积压高达 24-26 个请求；修改后『full token usage』提升至 0.16-0.18，队列积压降至 3-4 个请求。

实现拆解

1. 在 add_one_req 方法中定位两处基于 rem_input_tokens 和 can_run_list 的判断分支（锁前和锁内）。
2. 将原条件 if real_input_tokens >= self.rem_input_tokens and len(self.can_run_list) != 0 修改为 if self.rem_chunk_tokens is None and len(self.can_run_list) != 0 and real_input_tokens >= self.rem_input_tokens，即仅在非分段预填充模式下才执行此拒绝。
3. 锁内对应的 input_tokens 检查做同样修改，并调整条件顺序（先检查 rem_chunk_tokens is None）。
4. 添加注释说明两种模式下的意图：非分段预填充时若批次非空则满足 max_prefill_tokens 约束后拒绝；若批次为空则始终接受首个请求。分段预填充时此限制被取消，由 rem_chunk_tokens 做分块控制。
5. 测试验证：依赖现有 CI 测试，未新增单独测试用例。

关键文件：

- python/sglang/srt/managers/schedule_policy.py（模块 调度策略；类别 source；类型 core-logic；符号 add_one_req）：核心调度策略文件，修改了预填充批次接纳逻辑，对吞吐量有直接影响。

关键符号：add_one_req

关键源码片段

python/sglang/srt/managers/schedule_policy.py

核心调度策略文件，修改了预填充批次接纳逻辑，对吞吐量有直接影响。

```
if (
    self.rem_chunk_tokens is None # 仅在非分段预填充模式下才进行此限制
    and len(self.can_run_list) != 0 # 如果批次已有请求
    and real_input_tokens >= self.rem_input_tokens # 且剩余输入预算不足
):
    # 非分段预填充：批次非空时满足 max_prefill_tokens 约束后拒绝
    # 若批次为空，则始终接受首个请求以避免死锁
    return AddReqResult.OTHER

with self._lock_node(req.last_node):
    # 锁内对应检查
    if (
        self.rem_chunk_tokens is None
        and len(self.can_run_list) != 0
        and input_tokens >= self.rem_input_tokens
    ):
        return AddReqResult.OTHER
```

评论区精华

merrymercy 在 review 中建议将条件重组，并强调“仅当未启用分段预填充时才应用 rem_input_tokens 约束”。核心设计决策：当启用分段预填充时，无需通过 rem_input_tokens 拒绝请求，因为分块会自然适配预算；当未启用分段预填充且批次非空时，仍应用 max_prefill_tokens 限制。最终提交合并了这些建议。

- 区分分段预填充和非分段预填充的拒绝策略 (design): 采纳建议，调整了条件顺序和注释。

风险与影响

- 风险：主要风险：该修改移除了分段预填充模式下基于 rem_input_tokens 的早期拒绝，但分段预填充有 rem_chunk_tokens 限制每次分块大小，且锁内仍有 rem_total_tokens 和 rem_input_tokens 检查作为安全网。对于非分段预填充模式，行为与原版一致（首个请求始终允许）。潜在问题是当 rem_chunk_tokens 配置很大时，早期接受可能导致大量 token 涌入，但预算控制仍会逐步生效。缺少测试覆盖（无新增单元测试），回归风险由 CI 覆盖。
- 影响：用户：主要影响使用 chunked prefill 的用户（如长场景或 PD 分离场景），预填充吞吐量提升，排队延迟降低。系统：更高效的批次填充可能增加单次预填充的 token 总数，但 budget 机制会限制总体。团队：此修改澄清了调度逻辑的意图，但需要与相关工作（如 #23048）协调，避免重复或冲突的修复。
- 风险标记：核心路径变更（调度策略核心方法），缺少测试覆盖

关联脉络

- PR #23048 Fix chunked prefill scheduling issue: Baichuan7 在 issue 评论中指出本 PR 与其之前的 PR #23048 重叠，可能涉及相同的调度问题。
- PR #22984 Chunked prefill scheduling bug: Baichuan7 在 issue 评论中引用了该 issue 作为相关工作。