PR #27399 完整报告

sgl-project/sglang

Respect explicit --max-running-requests instead of clamping to heuristic

合并时间：2026-06-13 03:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27399

执行摘要

- 一句话：修复显式 --max-running-requests 被裁剪的问题
- 推荐动作：值得快速合并。该 PR 修复了一个影响 beam search 等场景的静默限制问题，变更极小且逻辑正确，已获两位 reviewer 批准。建议合并后通知使用 --max-running-requests 的团队，特别是 beam search 用户。

功能与动机

PR body 指出：max_running_requests 显式设置时被 estimated 上限（min(..., 4096)）无声裁剪，如传入 12800 实际得到 4096 且无警告。beam search 场景下单个请求会分裂为 num_beams 个并发序列，需要远超 4096 的请求槽位，因此原有的裁剪行为使此类工作负载不可调度。

实现拆解

1. 修改上限绑定逻辑：在 _resolve_max_num_reqs 方法中，当用户显式设置了 max_running_requests 时，将原用的 estimated（启发式估计值）替换为 token_capacity // 2（真实的 KV 缓存容量一半），从而允许用户请求更大的并发数。自动路径（未显式设置时）保持不变，仍使用 estimated。
2. 添加告警日志：当用户的请求值因 KV 容量限制而被降低时，记录一条 logger.warning，明确告知用户实际生效的值和原因，避免之前的静默行为。
3. 变量引入：引入 requested_per_worker 局部变量存储除以 dp_size 后的请求值，使代码逻辑更清晰，并用于后续的告警判断。
4. 仅修改一个文件：所有变更集中在 python/sglang/srt/model_executor/model_runner_kv_cache_mixin.py 的 _resolve_max_num_reqs 方法中。

关键文件：

- python/sglang/srt/model_executor/model_runner_kv_cache_mixin.py（模块 调度器；类别 source；类型 data-contract）：唯一修改的文件，包含核心方法 _resolve_max_num_reqs 的边界条件变更和告警日志添加。

关键符号：_resolve_max_num_reqs

关键源码片段

python/sglang/srt/model_executor/model_runner_kv_cache_mixin.py

唯一修改的文件，包含核心方法 _resolve_max_num_reqs 的边界条件变更和告警日志添加。

```
# python/sglang/srt/model_executor/model_runner_kv_cache_mixin.py
# 方法 _resolve_max_num_reqs 中关键变更部分

def _resolve_max_num_reqs(self: ModelRunner, token_capacity: int) -> int:
    # ... 省略估计值计算 ...
    max_num_reqs = self.server_args.max_running_requests
    if max_num_reqs is not None:
        # 用户显式设置了 max_running_requests
        requested_per_worker = max_num_reqs // self.dp_size
        # 改为使用 token_capacity // 2 作为上限，尊重用户设置
        max_num_reqs = min(requested_per_worker, token_capacity // 2)
    else:
        # 未显式设置，继续使用启发式值 estimated
        requested_per_worker = None
        max_num_reqs = min(estimated, token_capacity // 2)

    # ... 省略 mamba 分支处理 ...

    # 当实际值低于用户请求时，发出警告（不再静默截断）
    if requested_per_worker is not None and max_num_reqs < requested_per_worker:
        logger.warning(
            "max_running_requests was reduced from the requested %d to %d "
            "(per dp worker) due to the available KV cache capacity.",
            requested_per_worker,
            max_num_reqs,
        )
    # ... 后续逻辑 ...
```

评论区精华

PR 未触发大量的 review 讨论，评论数为 0，两位 reviewer 均直接批准。这表明变更逻辑清晰、争议小。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。变更仅限于 _resolve_max_num_reqs 方法的一行边界条件，且显式设置路径和自动路径区分明确，不会影响未设置 --max-running-requests 的默认行为。唯一的风险点是：如果用户设置的 max_running_requests 超过 KV 容量实际可承受的范围，可能因过度调度导致 OOM，但新的 KV 容量上限 token_capacity // 2 已提供了合理的安全边界，且告警日志可以帮助用户感知。
- 影响：用户影响：显式设置 --max-running-requests 的用户（尤其是 beam search 等场景）现在可以看到自己的请求值被尊重，不会再被静默减小到 4096。系统影响：变更后，显式设置时上限更高，可能增加请求并发度，从而增加 KV 缓存压力，但上限为

token_capacity // 2 已基于内存 profiling 结果，理论上是安全的。团队影响：无。

- 风险标记：缺少测试覆盖

关联脉络

- PR #23862 Fix --mem-fraction-static not accounting for EAGLE draft model KV cache: 同样涉及 _resolve_max_num_reqs 方法和请求数限制逻辑，关联到调度与内存配置。