

PR #33133 完整报告

sgl-project/sglang

[PD] Add a queues.prealloc_ready counter to the load snapshot

合并时间: 2026-08-04 00:44

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33133>

执行摘要

- 一句话: PD 负载快照新增 prealloc_ready 计数指标
- 推荐动作: 值得快速浏览, 重点学习其“把两层等待语义拆开暴露为独立指标”的可观测性设计思路。这是低侵入、纯增量扩展指标的良好范例, 适合作为后续在 /v1/loads 上增加精细指标的模板。如果团队依赖 PD 扩缩容决策, 建议跟进补充计数逻辑的单元测试, 并确认 waiting_for_input 语义在队列状态机中的准确性。

功能与动机

PR body 明确指出: decode-side load snapshot reports the prealloc queue as a single depth (`disaggregation.decode_prealloc_queue_reqs`), 它混合了两类状态——一类是仍在等待 prefill 侧完成 KV 交接的请求, 另一类是交接已完成、仅被 decode 侧准入阻塞的请求。因此 consumers of /v1/loads (autoscalers, routers, dashboards) cannot tell a decode pool that is genuinely out of capacity from one that is simply waiting on prefill。新增 `prealloc_ready` 计数器即是为了把“仅等 decode 准入”的子集独立暴露, 避免 prefill 积压被误读为 decode 侧压力。

实现拆解

1. 数据契约扩展: 在 `python/sglang/srt/managers/load_snapshot.py` 的 `QueueMetrics` (`msgspec.Struct, array_like=True`) 末尾新增 `prealloc_ready: int` 字段, 保持既有字段顺序不变, 确保 SHM/zmq 序列化兼容; 该字段无默认值, 与 `waiting`、`grammar` 等字段风格一致。
2. 指标计算: 在 `python/sglang/srt/managers/scheduler_components/load_inquirer.py` 的 `SchedulerLoadInquirer.get_loads()` 中新增 `decode_prealloc_ready = 0` 初始化; 仅在 `DisaggregationMode.DECODE` 分支中通过 `sum(1 for decode_req in get_disagg_decode_prealloc_queue().queue if decode_req.waiting_for_input)` 统计“仅等准入”的请求数, 并在构造 `QueueMetrics` 时传入。prefill 与非 PD 引擎自然保持 0。
3. 测试配套: 更新 `test/registered/unit/entrypoints/test_v1_loads_aggregate.py` 的 `test_reads_snapshot_and_filters_sections`, 构造 `QueueMetrics` 时补充 `prealloc_ready=1`, 覆盖新增字段的 SHM 读写路径; 但未添加针对计数逻辑本身的断言。
4. 配套变更: PR 未更新文档, 也明确声明不涉及模型或内核代码, 无 accuracy/benchmark 测试。

关键文件：

- python/sglang/srt/managers/scheduler_components/load_inquirer.py（模块 负载采集；类别 source；类型 core-logic；符号 SchedulerLoadInquirer.get_loads）：核心改动文件：在 get_loads() 中新增 decode_prealloc_ready 统计逻辑，是 /v1/loads 新指标的唯一天然数据来源。
- python/sglang/srt/managers/load_snapshot.py（模块 负载快照；类别 source；类型 data-contract；符号 QueueMetrics）：定义 /v1/loads 的 wire 契约：QueueMetrics 新增 prealloc_ready 字段，直接影响 SHM/zmq 序列化格式。
- test/registered/unit/entrypoints/test_v1_loads_aggregate.py（模块 负载聚合；类别 test；类型 test-coverage；符号 test_reads_snapshot_and_filters_sections）：回归测试：验证 QueueMetrics 新增字段后 SHM 读写与 /v1/loads section 过滤逻辑仍正常。

关键符号：SchedulerLoadInquirer.get_loads, QueueMetrics

关键源码片段

python/sglang/srt/managers/scheduler_components/load_inquirer.py

核心改动文件：在 `get_loads()` 中新增 `decode_prealloc_ready` 统计逻辑，是 /v1/loads 新指标的唯一天然数据来源。

```
# 仅 decode 模式需要 prealloc_ready; prefill / 非 PD 引擎恒为 0
mode_str = "null"
prefill_bootstrap = prefill_inflight = 0
decode_prealloc = decode_transfer = decode_retracted = 0
decode_prealloc_ready = 0

if self.disaggregation_mode == DisaggregationMode.PREFILL:
    mode_str = "prefill"
    prefill_bootstrap = len(self.get_disagg_prefill_bootstrap_queue().queue)
    prefill_inflight = len(self.get_disagg_prefill_inflight_queue())
elif self.disaggregation_mode == DisaggregationMode.DECODE:
    mode_str = "decode"
    decode_prealloc = len(self.get_disagg_decode_prealloc_queue().queue)
    decode_transfer = len(self.get_disagg_decode_transfer_queue().queue)
    decode_retracted = len(self.get_disagg_decode_prealloc_queue().retracted_queue)
    # 单次遍历 prealloc 队列，统计“等准入”的请求数，复杂度 O(n)
    decode_prealloc_ready = sum(
        1
        for decode_req in self.get_disagg_decode_prealloc_queue().queue
        if decode_req.waiting_for_input
    )

# queues 汇总：prealloc_ready 只反映 decode 侧压力
queues = QueueMetrics(
    waiting=len(self.get_waiting_queue()),
    grammar=stats.num_grammar_queue_reqs,
    paused=stats.num_paused_reqs,
```

```
retracted=stats.num_retracted_reqs,  
prealloc_ready=decode_prealloc_ready,  
)
```

python/sclang/srt/managers/load_snapshot.py

定义 `/v1/loads` 的 wire 契约：QueueMetrics 新增 `prealloc_ready` 字段，直接影响 SHM/zmq 序列化格式。

```
class QueueMetrics(msgspec.Struct, array_like=True):  
    """Detailed queue info breakdown."""  
  
    waiting: int  
    grammar: int  
    paused: int  
    retracted: int  
    # 新增：仅 decode 模式有意义，表示已完成 KV 握手、  
    # 只等 decode 侧准入的 prealloc 队列请求数  
    prealloc_ready: int
```

评论区精华

该 PR 没有 code review 评论。issue 中仅有两条操作记录：Gemini Code Assist 的停止服务提示，以及作者触发 `/tag-and-rerun-ci` 重跑 CI。主 CI 通过，extra CI 有一次失败记录，但可见材料中未说明失败原因，最终 PR 由作者合并。没有出现针对实现方案的实质技术讨论。

- CI 状态与 rerun 触发 (other): 未产生设计争议，PR 已由作者合并。

风险与影响

- 风险：数据契约风险：QueueMetrics 新增无默认值字段，所有构造点都必须同步更新，否则 msgspec.Struct 构造会失败；当前生产构造点仅有 `load_inquirer.py` 一处，风险较低，但后续新增同类指标时需注意。字段语义风险：指标准确性依赖 `waiting_for_input` 的语义，PR 描述与 commit message 均将其理解为“KV 握手完成、仅等准入”，但字段名本身有歧义，建议后续补充针对队列状态转移的单元测试。兼容性风险：`/v1/loads` 响应新增字段对普通消费者是向后兼容的增量变化，但严格 schema 校验的 autoscaler 需同步更新。性能风险：decode 模式每次快照多一次 $O(n)$ 遍历， n 为 prealloc 队列长度，开销可忽略。
- 影响：影响范围集中在 PD 部署的可观测性链路：`/v1/loads` 接口的 `queues` 部分新增 `prealloc_ready` 字段。对使用该端点的 autoscaler、路由器和监控看板，这是一个更有区分度的 decode 侧容量信号；对系统运行时行为无任何影响。团队运维侧受益，尤其是 PD 场景的容量规划与扩缩容策略制定。修改面小，仅 3 个文件、11 行新增，无模型或内核代码变更。
- 风险标记：数据契约扩展，新增字段无默认值，`waiting_for_input` 语义需核实，未更新文档

关联脉络

- PR #33118 [PD] Fix false health-503 during decode retraction re-admission: 同为 PD 解码侧队列 / 准入逻辑的修复，与本 PR 一起完善 decode 侧的健康判断与负载观测。

- PR #33337 observability: publish the generated forward-pass-metrics endpoint to the bags: 同为负载 / 观测指标体系的演进方向，揭示 SGLang 持续把可观测性指标做细做准的脉络。