PR #22331 完整报告

sgl-project/sglang

[HiSparse] Clarify decode token usage logs

合并时间：2026-04-14 09:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22331

执行摘要

本 PR 澄清了 HiSparse 稀疏注意力场景下的解码令牌使用日志，将原有单一的令牌统计细分为 GPU 令牌和 CPU 令牌的详细使用情况，增强了资源监控的清晰度；变更涉及 hisparse_coordinator 和 scheduler_runtime_checker_mixin 两个模块，通过新增统计数据和日志逻辑实现，风险较低，适合关注 HiSparse 性能监控的开发者参考。

功能与动机

PR 的动机源于 HiSparse 环境中日志信息不够清晰的问题。在 before 日志中，仅显示 #token: 20800, token usage: 0.29，无法区分 GPU 和 CPU 的令牌使用；after 日志新增了 #gpu-token: 20800, gpu token usage: 0.29, #cpu-token: 41422, cpu token usage: 0.29，从而提供更精确的资源利用率洞察，帮助用户更好地调试和优化 HiSparse 配置。

实现拆解

实现主要包括两个文件的关键改动：

- python/sglang/srt/managers/hisparse_coordinator.py：
  - 新增 HiSparseTokenStats NamedTuple，包含 device_tokens、device_token_usage、host_tokens、host_token_usage 字段。
  - 新增 get_token_stats 方法，通过分配器计算设备（GPU）和主机（CPU）的令牌使用量和利用率。python def get_token_stats(self) -> HiSparseTokenStats: device_allocator = self.token_to_kv_pool_allocator.hisparse_attn_allocator device_capacity = device_allocator.size device_tokens = device_capacity - device_allocator.available_size() host_capacity = self.mem_pool_host.size host_tokens = host_capacity - self.mem_pool_host.available_size() return HiSparseTokenStats(device_tokens=device_tokens, device_token_usage=(device_tokens / device_capacity if device_capacity > 0 else 0.0), host_tokens=host_tokens, host_token_usage=(host_tokens / host_capacity if host_capacity > 0 else 0.0),)
- python/sglang/srt/managers/scheduler_runtime_checker_mixin.py：
  - 扩展 PoolStats 类，添加 is_hisparse、hisparse_device_tokens、hisparse_device_token_usage、hisparse_host_tokens、hisparse_host_token_usage 字段。

- 新增 _get_hisparse_token_info 方法，调用 hisparse_coordinator.get_token_stats() 并更新 PoolStats。
- 修改 get_decode_usage_msg_parts 方法，在 HiSparse 启用时输出细分统计到日志。

评论区精华

review 讨论较少，核心点是字段命名建议：

- hzh0425评论："can we rename to host_tokens_usage"，针对 cpu_tokens 字段。
- 最终代码已使用 host_tokens 和 host_token_usage，表明建议被采纳，命名更统一，无其他争议。

风险与影响

- 风险：日志格式变更可能破坏依赖旧格式的监控工具，需协调更新；新增统计计算引入轻微性能开销（除法运算），但仅在日志路径，影响可忽略；代码改动集中，回归风险低，但建议补充测试确保正确性。
- 影响：用户获得更详细的 HiSparse 资源使用日志，便于性能分析；系统可观测性提升，无功能副作用；团队受益于增强的监控能力。

关联脉络

从同仓库历史 PR 看，暂无直接相关的 HiSparse PR；但 PR 18016（新增 HiCache 存储后端）涉及类似缓存和监控主题，可间接参考架构模式。本 PR 独立于其他功能演进，专注于 HiSparse 模块的日志改进。