PR #21884 完整报告

sgl-project/sglang

revert: remove TTL-based hard pin from HiRadixCache

合并时间：2026-04-02 07:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21884

执行摘要

- 一句话：移除 HiRadixCache 中基于 TTL 的硬钉功能及相关代码。
- 推荐动作：建议关注此 revert 的原因，理解原有 TTL-based 硬钉设计的不足，并跟踪后续 PR 以学习新的实现方案。对于涉及缓存管理或 admin 接口的开发者，需注意 API 变更和配置调整。

功能与动机

PR body 中提到"Will approach this a different way in follow up PRs"，表明原有基于 TTL 的硬钉实现可能存在问题或设计不佳，因此先移除以准备新的方案。

实现拆解

删除所有与 pin_prefix 功能相关的代码：在 http_server.py 中移除 API 端点；在 environ.py 中删除环境变量 SGLANG_HICACHE_MAX_PINNED_RATIO；在 io_struct.py 中删除 PinPrefixReqInput 和 PinPrefixReqOutput 数据结构；在 scheduler.py 中移除 pin_prefix_wrapped 方法和相关导入；在 tokenizer_communicator_mixin.py 中删除 pin_prefix_communicator 和相关方法；在 hiradix_cache.py 中删除 pin_prefix 方法、pin 状态管理、预算初始化等；在 radix_cache.py 中移除 TreeNode 的 pin_expiry 和 pin_ttl 字段。

关键文件：

- python/sglang/srt/mem_cache/hiradix_cache.py（模块 mem_cache/hicache）：移除了 pin_prefix 方法、pin 状态管理（如 _is_pinned、_clear_pin）和预算初始化，是核心缓存逻辑的关键变更。
- python/sglang/srt/entrypoints/http_server.py（模块 entrypoints/http_server）：删除了 pin_prefix API 端点（/hicache/pin_prefix），影响 admin 接口的可访问性。
- python/sglang/srt/managers/scheduler.py（模块 managers/scheduler）：移除了 pin_prefix_wrapped 方法和相关调度器处理，影响请求分发和缓存管理集成。
- python/sglang/srt/mem_cache/radix_cache.py（模块 mem_cache/hicache）：移除了 TreeNode 的 pin_expiry 和 pin_ttl 字段，影响缓存节点数据结构和状态跟踪。

关键符号：pin_prefix, pin_prefix_wrapped, pin_prefix_communicator, PinPrefixReqInput, PinPrefixReqOutput

评论区精华

无 review 讨论，PR 由作者直接合并，未经过代码评审。

- 暂无高价值评论线程

风险与影响

- 风险：风险包括：1) 暂时移除了前缀钉住功能，依赖此功能的用户或客户端（如通过 admin API）可能无法使用；2) 移除环境变量 SGLANG_HICACHE_MAX_PINNED_RATIO 可能影响系统配置，需更新相关文档；3) 代码删除可能引入回归，如果其他模块错误地引用了被移除的代码（如通过导入或反射），但变更集中，风险可控。
- 影响：影响范围：admin API 失去 pin_prefix 端点，高级用户无法手动钉住前缀以抵抗缓存淘汰；缓存管理逻辑不再支持硬钉，可能影响特定优化场景。影响程度：中等，主要影响管理员和高级缓存配置，不影响正常推理流程和核心性能。
- 风险标记：功能移除，临时功能缺失，配置变更

关联脉络

- PR #18941 推测为引入 TTL-based 硬钉的 PR: 此 PR reverts 了 PR 18941（基于 head_ref 'ishan/revert-pr-18941'），移除了其中引入的 TTL-based 硬钉功能。