

PR #25122 完整报告

sgl-project/sglang

Restructure mem_cache auto-labels by layer

合并时间: 2026-08-21 23:30

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/25122>

执行摘要

- 一句话: 重构 mem_cache 自动标签, 按池 / radix / 主机三层拆分
- 推荐动作: 值得快速浏览。如果要扩展 labeler 规则, 本 PR 提供了两个可借鉴的设计决策: 用目录 glob 而非显式文件列表以抵御目录迁移; 按子系统所有权而非文件名子串组织匹配。对于关注 HiCache / 缓存层演进的读者, 重点是 ispobock 的合并说明与 RFC #25371 的关系, 以及 memory-pool 与 unified-radix-cache 两个新标签组如何为未来的 mem_cache 拆分铺路。

功能与动机

PR body 明确指出原规则的核心缺陷: The hicache auto-label rule in .github/labeler.yml only matched files whose path contains the substring hicache. This misses the files that define and own HiCache, 并列出 cache_controller.py (class HiCacheController)、hiradix_cache.py (class HiRadixCache)、memory_pool_host.py、storage/**、hybrid_cache/**、test/registered/hicache/** 等所有权文件, 估算约 50 个核心文件漏标。合并者 ispobock 在评论中进一步说明标签结构须对齐 RFC #25371 的 pool / radix / hicache 三层划分。

实现拆解

变更入口是 .github/labeler.yml, 这是 GitHub Actions labeler 的规则文件, 所有 PR 打开或更新时都会按这里的 glob 自动打标签。实现分四步:

1. 扩展并按层重组 hicache 规则。原规则只有 '**/*hicache*' 一条 glob。新规则保留该 glob, 并新增 '**/*hiradix*'、'**/HiCache/**/*' 两个命名变体, 再显式加入所有权路径: managers/cache_controller.py (class HiCacheController 的定义者)、mem_cache/pool_host/**、mem_cache/hybrid_cache/**、mem_cache/storage/**、mem_cache/buffer_mode/**、mem_cache/**/*host*、mem_cache/**/*storage*、mem_cache/l2_transfer.py、benchmark/hicache/** 与 test/registered/hicache/**。
2. 新增 memory-pool 标签组。覆盖 mem_cache/pool/**、mem_cache/allocator/**、mem_cache/layout/**、mem_cache/**/*memory_pool* 与 test/registered/mem_cache/**。该标签在改动前不存在, 是本次为设备侧 KV / 状态池与槽位分配新建的。

3. 新增 unified-radix-cache 标签组。覆盖 `mem_cache/unified_cache/**`、`mem_cache/cpp_radix_tree/**`、`mem_cache/**/*radix*` 以及三个具体文件 `chunk_cache.py`、`evict_policy.py`、`base_prefix_cache.py`，并包含 `test/registered/radix_cache/**`，对应 radix 树与匹配层。
4. 刻意排除共享调用点。`scheduler.py`、`server_args.py`、`model_runner.py`、`http_server.py`、可观测性 mixin、disagg 模块都不加入任何一组，避免任何一个触碰这些热点文件的 PR 继承 hicache 标签。

关于测试配套：PR body 的测试计划是三项手工验证（用 draft PR 分别触碰 `cache_controller.py`、`storage/hf3fs/storage_hf3fs.py`、`scheduler.py`，检查标签命中或未命中），且三项均未勾选完成；由于这是 CI 配置而非代码，没有单元测试配套。合入后实际打标行为只能通过后续 PR 的标签结果来观察。

关键文件：

- `.github/labeled.yml`（模块 标签配置；类别 infra；类型 infrastructure）：本 PR 唯一变更文件，是 GitHub Actions labeler 的规则配置。原 hicache 规则只匹配文件名含 hicache 子串的路径，重构后拆为 `memory-pool` / `unified-radix-cache` / `hicache` 三组，覆盖控制器、存储后端、混合缓存等所有权路径，并刻意排除共享调用点，修复约 50 个核心文件漏标问题。

关键符号：未识别

评论区精华

本 PR 没有正式的 review 线程（`review_comments_count = 0`），最有价值的讨论来自合并者 ispobock 在关闭时留下的说明，相当于设计决策的最终记录：

Restructured the `mem_cache` labels into three, aligned to the `pool / radix / hicache` layering in RFC #25371 ... Kept it to keyword and directory globs (no explicit per-file lists) so it survives the RFC #25371 moves that rename these files. Dropped the `mem-cache` rule since that label does not exist in the repo; `memory-pool` was created for this.

其中最关键的一条设计权衡是：主机池与存储后端归入 `hicache` 而不是 `memory-pool`，理由是 RFC 的划分把 `pool/`（设备侧）与 `pool_host/`（主机镜像）分开。另有一条 `gemini-code-assist` 的配额警告评论，与变更内容无关。

- `mem_cache` 标签分层与所有权路径划分 (design): 采用三层标签结构；主机侧（`pool_host` / `hybrid_cache` / `storage` / `buffer_mode` / `cache_controller`）全部归入 `hicache`；设备侧归入 `memory-pool`；radix 树相关归入 `unified-radix-cache`。

风险与影响

- 风险：
 - glob 过宽：`mem_cache/**/*host*` 与 `mem_cache/**/*storage*` 会命中未来新加入 `mem_cache` 下的一切主机 / 存储相关文件，若后续出现不属于 HiCache 的主机或存储模块，会被误标；这是用通配换取抗目录迁移能力的代价。

- 依赖未来目录结构：pool/、unified_cache/、cpp_radix_tree/、pool_host/ 这些路径是 RFC #25371 规划中的目标目录，如果实际重命名与 glob 假设不一致，规则会再次失效，需要跟随 RFC 落地情况做二次校对。
- 缺少自动化验证：打标行为只能靠手工 draft PR 验证，且合入时三项测试计划均未勾选，存在规则写错而未被发现的窗口期。
- 新标签可行性：memory-pool 是本次新建的标签，若仓库标签权限或命名规范不允许 labeler 自动创建标签对象，该组规则会静默失效。
- 影响：
 - 对维护者：HiCache / 缓存层相关 PR 的筛选与追溯更准确，约 50 个此前漏标的文件将正确获得 hicache 标签。
 - 对 CI 流程：影响所有后续 PR 的自动打标行为，属于全局性的工程流程变更，但无运行时影响、无用户可见影响。
 - 对团队协作：标签结构成为 RFC #25371 缓存分层改革在 CI 侧的先行落地，为后续 mem_cache 目录大迁移做了铺垫。
 - 风险标记：CI 配置变更，glob 匹配范围过宽，依赖 RFC 目录迁移，缺少自动化验证

关联脉络

- PR #35906 config: project the config bags from the resolution result: 该 PR 改动了 python/sglang/srt/mem_cache/hybrid_cache/hybrid_pool_assembler.py，正位于本 PR 新增 hicache 规则覆盖的 hybrid_cache 目录；其标签中也包含 hicache，可作为新规则命中样例的佐证。