PR #7125 完整报告

PaddlePaddle/FastDeploy

[Feature] Config eviction_duration

合并时间：2026-04-01 16:46

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7125

执行摘要

本 PR 新增了 radix tree 缓存驱逐时间的可配置性，将默认值从 5 分钟调整为 30 分钟以提升缓存命中率，并扩展 token counter 至 mixed worker 类型。变更涉及配置管理、缓存核心实现和测试覆盖，对调度性能有积极影响，需关注默认值变更的兼容性风险。

功能与动机

主要动机是解决默认驱逐时间过短导致的缓存命中率低问题。PR body 指出“更新 radix tree 默认驱逐时间，新增用户可配置驱逐时间”，fastdeploy-bot 在 review 中补充“原默认驱逐时间(5分钟)过短，导致缓存命中率较低”。通过支持用户自定义，增强系统灵活性和性能优化能力。

实现拆解

实现分为以下关键模块：

1. 配置模块(config.go): 新增 EvictionDurationMins 字段，默认值设为 30 分钟，在 Load 函数中初始化。
2. 缓存模块(prefill_cache_aware.go): 修改 newRadixPrefixCache 函数，传入 evictionDuration 参数，移除硬编码的 5 分钟默认值。
3. 调度模块(handler.go): 传递配置并扩展 SelectWorker 函数，使 token counter 逻辑支持 mixed worker 类型。
4. 文档与测试: 更新 router.md 文档添加配置项说明；在 completions_test.go 中添加大量单元测试，模拟超时和挂起场景，确保功能稳健性。

评论区精华

review 讨论由 fastdeploy-bot 主导，重点指出配置不一致问题：

> “文档注释标注 default: 5，但代码中 config.go 的默认值已改为 30 分钟。”

bot 建议统一默认值为 30 分钟或修正文档，但未明确结论；PR 最终被合并，暗示作者已解决此问题，强调了配置正确性的重要性。

风险与影响

- 风险: 默认值变更可能影响现有系统缓存行为，需评估向后兼容性；延长驱逐时间可能增加内存使用，需监控性能；新增配置项若传递错误可能导致功能失效。

- 影响: 用户可自定义驱逐时间以优化缓存策略，系统缓存命中率提升可能改善性能，团队需维护新增配置和测试覆盖。

关联脉络

本 PR 与近期历史 PR 共同构成调度和缓存优化脉络:

- PR 7001（调度优化）和 PR 7107（KVCache 管理）均涉及类似模块，显示团队持续改进调度策略和缓存效率。
- PR 6680（调度性能优化）与本 PR 的 token counter 扩展相辅相成，共同提升系统响应能力。 这些关联表明 FastDeploy 在在线服务路由和调度方面进行系统性增强。