# PR #1735 完整报告

THUDM/slime

[slime-router] support pd disaggregation and remove radix tree middleware

合并时间：2026-03-18 11:48

原文链接：http://prhub.com.cn/THUDM/slime/pull/1735

## PR 1735 分析报告

### 1. 执行摘要

本次 PR 在 slime 路由器中支持 Prefill-Decode（PD）分离以提升性能，并移除 radix tree 中间件以简化架构。变更影响路由器核心逻辑和用户配置，属于重大架构调整，旨在优化 rollout 流程并减少维护负担。

### 2. 功能与动机

动机源于优化 rollout 性能的需求：PD 分离通过将 prefill 和 decode 任务分发到不同类型 worker（如 prefill、decode、regular），提升高并发下的吞吐量。同时，移除了过时的 radix tree 中间件，该中间件原先用于文本到令牌缓存，但可能已不再必要或被替代，以简化系统设计和用户接口。从文档更新（如 docs/en/advanced/slime-router.md）可见，新功能强调流式代理和 PD 双发路由，而移除中间件减少了配置复杂性。

### 3. 实现拆解

按模块拆解关键改动：

- 路由器核心（slime/router/router.py）：添加 WorkerType 枚举和 WorkerInfo 类以支持 PD 分离，重构路由逻辑，并移除中间件加载代码。关键代码如下：python class WorkerType(str, Enum): REGULAR = "regular" PREFILL = "prefill" DECODE = "decode"
- 中间件移除：删除 slime/router/middleware_hub/ 目录下的 radix_tree.py 和 radix_tree_middleware.py 文件，彻底废弃 radix tree 功能。
  - rollout 流程（slime/rollout/sglang_rollout.py）：移除对 radix tree 中间件的调用，简化 token 更新逻辑。
  - 参数配置（slime/utils/arguments.py）：删除 --slime-router-middleware-paths 参数，影响用户命令行选项。
- 文档同步：更新中英文文档以反映 PD 分离和移除中间件的变化。

### 4. 评论区精华

PR 中没有 review 评论，表明变更可能经过内部讨论后直接合并，没有公开的技术争议或设计权衡。这提示变更可能被视为低风险或已充分验证。

### 5. 风险与影响

风险：

- 移除 radix tree 中间件可能导致依赖此功能的用户出现回归，影响 token 对齐和训练精度。
- PD 分离引入新的路由逻辑，增加系统复杂性，需充分测试负载均衡和容错机制。
- 配置变更可能引起用户错误，例如旧脚本中使用 --slime-router-middleware-paths 参数会失效。

影响：

- 用户需更新配置，不再支持自定义中间件，但简化了使用门槛。
- 系统吞吐量可能提升，但性能增益依赖于 worker 类型注册和路由策略的正确实现。
- 代码库精简，减少了长期维护成本，但需确保无遗留依赖。

## 6. 关联脉络

从同仓库历史 PR 看，本 PR 与以下变更相关：

- PR 1770（切换 sglang-router 依赖）：同样涉及路由器外部依赖更新，共同推动路由器架构演进。
- PR 1746（placeholder worker 类型和指标路由）：在 worker 类型管理方面有重叠，表明仓库正持续优化路由器功能以支持多样化 rollout 场景。 这些关联显示 slime 路由器处于活跃开发阶段，重点在性能优化和架构简化。