PR #43097 完整报告

vllm-project/vllm

[Docs][PD][NIXL] Bidirectional kv-cache transfer

合并时间：2026-05-20 15:02

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/43097

执行摘要

本 PR 为 NIXL KV Connector 新增双向 KV 传输的完整文档，解释如何通过 RDMA 实现预填充（P）从解码（D）拉取已有 KV 块，以降低多轮对话的首 token 延迟。文档包含序列图、配置参数和代理设置示例。review 中引发了关于代理协议标准化的讨论。

功能与动机

在标准单向 KV 传输中，每轮多轮对话预填充端必须重新计算整个 KV 缓存，导致高延迟。双向传输允许预填充端从解码端拉取已有块，仅计算新增 token，显著降低 TTFT。该文档帮助用户理解并部署此功能。

实现拆解

本文档无代码变更，仅对 docs/features/nixl_connector_usage.md 新增章节：

1. 特性概述：对比单向与双向传输的差异。
2. 序列图：展示两轮交互中代理、预填充、解码的角色，以及 RDMA 拉取操作。
3. 配置参数：列出 kv_connector_extra_config 中的 decoder_kv_blocks_ttl 等。
4. 代理设置：启动状态代理和运行示例脚本的步骤。
5. 注意事项：需要 InfiniBand 等 RDMA 硬件，且与移除思考痕迹的推理模型可能不兼容。

本文档关键内容为序列图和配置示例。以下是配置示例（json 格式）：

// 启用双向传输的 kv_transfer_config 示例 { "kv_connector_extra_config":{ "decoder_kv_blocks_ttl":600// 解码端缓存 TTL（秒） } } 注意：JSON 内注释仅用于说明，实际配置应移除注释。

评论区精华

- markmc：对代理自定义 kv_transfer_params 表示忧虑，认为缺乏标准，希望有可依赖的协议。
- NickLucche：同意该问题，并指向 #43094 继续讨论路由器责任方案。该设计讨论暂未解决。

风险与影响

- 风险：文档可能随功能迭代过时；代理协议未标准化可能导致兼容性问题。
- 影响：用户获得清晰指导，降低部署门槛；对系统无运行时影响。

关联脉络

本 PR 与 issue #43094（路由器责任方案）直接相关，是双向 KV 传输的文档配套。未来标准化讨论可能推动代理协议的统一。