PR #42994 完整报告

vllm-project/vllm

[Docs] Fix MooncakeStoreConnector role in disaggregated example

合并时间：2026-05-20 02:14

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42994

执行摘要

本次 PR 修复了 MooncakeStoreConnector 分散式前缀缓存共享文档中的配置示例错误，将预填充实例的角色从 kv_producer 改为 kv_both，使其实际行为与文档描述一致。同时将 PYTHONHASHSEED 注意事项从仅限 DP 扩展到所有跨进程共享场景。

功能与动机

文档 docs/features/mooncake_store_connector_usage.md 中的分散式 Prefill-Decode 示例宣传“通过分布式 store 实现跨实例前缀缓存共享”，但将 inner MooncakeStoreConnector 设为 kv_producer，这意味着预填充器只保存 KV 缓存而不从 store 中加载，导致跨实例前缀缓存命中无法生效。PR 作者对照了 connector 源码（scheduler.py 和 worker.py），确认 kv_both 才能同时启用双向操作。

实现拆解

1. 修复角色配置：在分散式示例中，将 inner MooncakeStoreConnector 的 kv_role 从 "kv_producer" 改为 "kv_both"，同时保留 outer MultiConnector 的 kv_producer 角色以维护点对点连接的行为。
2. 更新角色选项说明：在 KV Role Options 表格中，将 kv_both 的描述从“用于单节点 CPU 卸载”更新为“用于单节点 CPU 卸载或预填充实例”。
3. 泛化 PYTHONHASHSEED 说明：将原本仅针对数据并行的 PYTHONHASHSEED 注意事项扩展为适用于所有共享同一个分布式 store 的进程（包括 DP 秩、分离的预填充 / 解码节点等），并更新小节标题为“跨进程的可重现块哈希”。

无需关键源码片段，本次变更为纯文档修改。

评论区精华

无实质性讨论；机器人评论和 approvals 均无实质内容。

风险与影响

- 风险：无，纯文档修改，不涉及代码。
- 影响：提升文档准确性，帮助用户正确配置 MooncakeStoreConnector 以实现跨实例前缀缓存共享。

关联脉络

- 关联 PR #42828 引入了 MooncakeStoreConnector 的 HMA 支持，本 PR 修正了该功能线的文档示例。
- 文档改动与源码逻辑 (scheduler.py, worker.py 中的 kv_role 检查) 一致，确保配置示例的有效性。