

PR #44055 完整报告

vllm-project/vllm

[Docs][KV Connector][NIXL] document KV Transfer stat logging and Prometheus metrics

合并时间: 2026-06-12 23:39

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/44055>

执行摘要

该 PR 为 NIXL KV Connector 添加了 Metrics Reference 文档, 详细解释了定期日志行 **KV Transfer metrics** 中的各字段以及所有 7 个 Prometheus 指标。内容直接来自源码中的统计实现, 帮助用户监控 KV 传输性能。

功能与动机

部分解决 Issue #41230。用户在使用 NIXL KV Connector 时需要理解传输延迟、吞吐量等指标的含义, 以便有效监控和排障。该文档填补了这一空白。

实现拆解

在 `docs/features/nixl_connector_usage.md` 文件中新增 **## Metrics Reference** 小节:

1. 日志示例: 展示一行 KV Transfer metrics 输出, 包含 8 个字段的数值。
2. 指标表: 逐一解释每个 CLI 指标, 包括含义、单位和典型用途 (如 P90 用于识别尾部延迟)。
3. Prometheus 指标表: 列出所有 7 个 Prometheus 指标, 标注类型和描述。
4. 使用提示: 说明 `vllm:nixl_num_kv_expired_reqs` 峰值与 `kv_lease_duration` 参数的关系。

无 (纯文档变更)。

评论区精华

markmc: "Merging this because it's a good start, I think. Doesn't fully address @NickLucche points in #41230 about cross-TP aggregation though."

审阅者认为该 PR 是好的起点, 但跨 Tensor Parallelism 聚合的指标文档尚未覆盖, 留待后续完善。

风险与影响

- 风险: 无。纯文档变更, 无代码改动。
- 影响: 对使用 NIXL KV Connector 的用户有帮助, 能更准确地解读指标数据; 对系统无影响。

关联脉络

- 该文档与 PR #35264 (Nixl KV Connector 实现) 紧密相关, 为该功能提供使用说明。

- 部分响应 Issue #41230, 但跨 TP 聚合指标文档尚未完成。