

PR #43603 完整报告

vllm-project/vllm

[Docs][ROCm] MoRI-IO Connector Usage Guide

合并时间: 2026-05-26 21:52

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/43603>

执行摘要

- 一句话: 添加 MoRI-IO 连接器使用指南
- 推荐动作: 建议 ROCm 平台上进行 PD 分解的用户阅读此文档; 对于开发者, 文档中的配置模式和调试技巧 (如环境变量) 值得参考。

功能与动机

PR body 指出目的是为 MoRI-IO KV Connector 添加使用指南, 类似已有的 P2pNccl 文档, 帮助用户在 ROCm 平台上进行 PD 分解部署时的 KV 缓存传输配置。同时为后续更详尽的说明 (如工作原理) 做准备。

实现拆解

1. 在 docs/features/ 下创建 moriio_connector_usage.md, 包含前提条件 (Docker 安装、NIC 库)、基本使用 (单主机生产者 / 消费者 / 代理)、多节点部署、性能调优 (QP 数量、重试)、附录 (排错、NIC 库安装、配置参考)。
2. 更新 disagg_prefill.md 中的支持连接器列表从 6 增加到 9, 新增 MoRIIOConnector 条目并补充其他遗漏的连接器 (OffloadingConnector、FlexKVConnectorV1、MultiConnector)。
3. 根据 review 修正 CLI 示例中的语法错误 (gpu-memory-utilization 连字符) 和配置方式说明, 简化 Docker 示例, 调整连接器计数为 9。

关键文件:

- docs/features/moriio_connector_usage.md (模块文档; 类别 docs; 类型 documentation): 核心新增, 完整介绍 MoRIIOConnector 的使用, 包含 CLI 示例、性能调优和排错指南。
- docs/features/disagg_prefill.md (模块文档; 类别 docs; 类型 documentation): 更新受支持连接器列表, 添加 MoRIIOConnector 条目并修正计数。

关键符号: 未识别

关键源码片段

[docs/features/moriio_connector_usage.md](#)

核心新增, 完整介绍 MoRIIOConnector 的使用, 包含 CLI 示例、性能调优和排错指南。

```
# 示例：启动 MoRI-IO 生产者 (prefiller)
# 必须设置 VLLM_ROCM_USE_AITER=1 以启用优化引擎
# kv-transfer-config 中指定连接器类型、角色和额外配置（代理 IP、端口等）
export VLLM_ROCM_USE_AITER=1
export CUDA_VISIBLE_DEVICES=0,1,2,3
vllm serve Qwen/Qwen3-235B-A22B-FP8 \
  -tp 4 \
  --port 20005 \
  --gpu-memory-utilization 0.9 \
  --kv-transfer-config '{
    "kv_connector": "MoRIIOConnector",
    "kv_role": "kv_producer",
    "kv_connector_extra_config": {
      "proxy_ip": "127.0.0.1",
      "proxy_ping_port": "36367",
      "http_port": "20005",
      "handshake_port": "6301",
      "notify_port": "6105"
    }
  }'
```

评论区精华

- gemini-code-assist[bot] 指出连接器数量应更新为 9（原始改为 7），以及 `--gpu_memory_utilization` 下划线错误应为连字符，嵌套配置不能使用点表示法。
- simondanielsson 反驳说 FlexibleArgumentParser 支持点表示法，因此部分点表示法建议未被采纳。
- AndreasKaratzas 和 tjtananaa 建议简化 Docker 示例中的卷挂载，simondanielsson 移除了多余的 `-v /sys:/sys`。
- tjtananaa 指出文档渲染存在语法问题，simondanielsson 修正了 `!note` 语法。
- 连接器数量计数 (correctness): simondanielsson 采纳建议改为 9。
- CLI 参数语法纠正 (correctness): simondanielsson 修正了该错误。
- 嵌套配置点表示法支持 (question): 部分点表示法建议未被采纳，说明现行解析器确实支持。
- Docker 卷挂载简化 (design): simondanielsson 简化了示例，只保留必要卷（`/dev`、`/run`、网络配置等）。

风险与影响

- 风险：主要风险是文档中的配置示例可能过时或不准确，导致用户部署失败。但经过多次 review 和测试验证，风险较低。软依赖 PR #40344 尚未合并，该文档在 #40344 合并前可能不完整（例如多节点脚本），但文档已声明未来会补充。
- 影响：对使用 ROCm MoRI-IO 连接器的用户有直接帮助，降低学习成本。更新 `disagg_prefill.md` 使文档清单保持一致，避免了用户困惑。影响范围限定在 ROCm 平台和 KV 连接器文档。
- 风险标记：文档依赖未合并 PR，用户配置错误风险

关联脉络

- PR #43303 [Misc][Refactor][ROCm] Convert MoRI-related envvars to extra config args: 将 MoRI 环境变量迁移至 kv_connector_extra_config, 与本 PR 使用的配置方式一致。