PR #39400 完整报告

vllm-project/vllm

[Doc] Switch K8S examples to default MP mode

合并时间：2026-06-11 02:17

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/39400

执行摘要

- 一句话：K8s 示例切换为默认 MP 分布式后端
- 推荐动作：建议阅读此 PR 以了解 vLLM 从 Ray 到 MP 的迁移决策及其文档化方式。对于 K8s 部署用户，可直接参考更新后的示例。

功能与动机

Fix https://github.com/vllm-project/vllm/issues/38113 per @ed-pai, previous k8s yaml won't work since default docker images has removed ray. so switch the k8s yaml example to default mp(multiprocessing) as distributed backend.

实现拆解

1. 更新 LWS 部署文档(docs/deployment/frameworks/lws.md)：将 lws.yaml 示例中的 leader/worker 命令从调用 multi-node-serving.sh 改为直接运行 vllm serve，使用 --nnodes、--node-rank、--master-addr 等参数。同时使用 pymdown tabbed 语法将旧 Ray 配置作为可选选项卡保留。
2. 更新 Kthena 集成文档(docs/deployment/integrations/kthena.md)：类似地，将示例命令和完整 YAML 切换到 MP 模式，并用 tabbed 语法展示 MP 和 Ray 两个版本，统一标签为 Multiprocessing (default) 和 Ray。
3. 调整示例脚本注释(examples/ray_serving/multi-node-serving.sh)：在注释行中添加 --distributed-executor-backend ray 标志，提醒使用 Ray 后端的用户显式指定。

关键文件：

- docs/deployment/integrations/kthena.md（模块 部署文档；类别 docs；类型 documentation）：更新了 Kthena 集成部署文档，核心变更是将分布式后端示例从 Ray 切换为 MP，并统一术语。
- docs/deployment/frameworks/lws.md（模块 部署文档；类别 docs；类型 documentation）：更新了 LWS 部署文档，将示例命令从 Ray 脚本切换为 MP 模式，并补充 Ray 选项卡。
- examples/ray_serving/multi-node-serving.sh（模块 示例脚本；类别 other；类型 core-logic）：在注释中添加了 --distributed-executor-backend ray 标志，提醒使用 Ray 后端的用户显式指定。

关键符号：未识别

评论区精华

主要讨论来自 reviewer @hmellor:

- 使用 tabbed 语法：建议在代码块中使用 === "Multiprocessing" 和 === "Ray" 选项卡同时展示两种配置，作者已采纳。
- 标签命名统一：@hmellor 要求将选项卡名称统一为 Multiprocessing (default) 和 Ray，作者已修改。
- 格式一致性：@gemini-code-assist 建议参数使用空格而非等号（如 --nnodes 2），与文档其他部分保持一致。 最终 @hmellor 给予 approval。
- 使用 tabbed 语法同时展示 MP 和 Ray 示例 (design): 作者采纳建议，修改了 LWS 和 Kthena 文档。
- 选项卡标签命名一致性 (style): 作者修改为统一命名。
- 参数格式（等号 vs 空格） (style): 该建议未被采纳，最终版本仍使用等号格式。

风险与影响

- 风险：低风险。主要风险在于用户可能未注意到默认后端变更导致原有 Ray 依赖的部署失败。但文档同时保留了 Ray 选项卡，且示例中明确标记了默认值。另外，MP 模式在某些环境下可能需要额外的端口配置，文档中已包含 --master-addr 和 --headless 等参数。整体风险可控。
- 影响：用户：K8s 用户可直接使用官方 Docker 镜像进行多节点推理，无需手动安装 Ray。团队：明确了 vLLM 分布式后端的默认选择为 MP，减少维护两套脚本的成本。系统：无代码变更，仅文档更新。
- 风险标记：用户需注意默认后端变更，文档与用户既有配置可能冲突

关联脉络

- 暂无明显关联 PR