PR #49651 完整报告

vllm-project/vllm

[XPU][CI] add heterogeneous TP UT

合并时间：2026-07-26 10:01

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49651

执行摘要

- 一句话：为 Intel GPU 添加异构 TP 测试
- 推荐动作：该 PR 可快速合并，变更简单透明。建议确认 4 卡 Intel GPU 资源在 CI 中充足，并关注后续 run_xpu_disagg_accuracy_test.sh 测试脚本是否兼容新环境变量。

功能与动机

功能上电后，需要确保异构 TP 场景（即 Prefiller 和 Decoder 使用不同 TP 大小）的正确性。PR 标题和 body 明确说明目的是“Add heterogeneous TP UT”，这是对现有 PD accuracy 测试的补充，防止回归。

实现拆解

1. 资源提升：在 .buildkite/intel_jobs/misc_intel.yaml 中，将 NixlConnector PD accuracy 测试的 num_devices 从 2 改为 4，gpu 代理标签从 2+ 改为 4+，以支持 4 卡异构 TP 测试。
2. 命令扩展：在原有单命令基础上，追加 3 个异构 TP 组合的测试命令，通过环境变量 PREFILLER_TP_SIZE 和 DECODER_TP_SIZE 控制 Prefiller 和 Decoder 的 TP 大小，覆盖 (2,1)、(1,2)、(2,2) 三种组合。
3. 标签更新：将测试名称从 NixlConnector PD accuracy (2 GPUs) 改为 NixlConnector PD accuracy (4 GPUs)，反映资源需求变化。

关键文件：

- .buildkite/intel_jobs/misc_intel.yaml（模块 CI 配置；类别 config；类型 configuration）：唯一变更文件，修改了 Intel GPU CI 的测试配置，增加异构 TP 测试场景。

关键符号：未识别

评论区精华

无 review 评论。仅 claude[bot] 自动评论表示 fork PR 跳过审核，jikunshang 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低。变更仅涉及 CI 配置文件和测试命令，不触及任何源码逻辑。潜在风险包括：4 卡资源可能不可用导致 CI 排队或超时；测试命令 run_xpu_disagg_accuracy_test.sh 如果未正确处理 PREFILLER_TP_SIZE 和 DECODER_TP_SIZE 环境变量，可能导致测试失败，但这属于测试脚本内部的正确性问题，PR 本身未修改该脚本。
- 影响：影响范围小，仅限 Intel GPU CI 管线中的 NixlConnector PD accuracy 测试。该测试将覆盖更多异构 TP 组合，提升回归测试的充分性，但不会影响用户功能或系统性能。
- 风险标记：CI 资源消耗增加，依赖测试脚本兼容性

关联脉络

- PR #49226 [Bugfix][KVConnector] Disable cross-layer KV blocks for per-token-head quant: 同一功能域（kv-connector）和 CI 配置相关，修改了相同的 KV 传输管线。
- PR #49671 [Bugfix][KV Offloading] Defer request finalization until final store: 也涉及 kv-connector 的 CI 测试配置（但主要修改 offloading scheduler），同属 Intel GPU CI 测试扩展趋势。