

PR #47758 完整报告

vllm-project/vllm

[XPU][CI]Adjust memory request for tests in Intel GPU CI

合并时间: 2026-07-07 13:56

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47758>

执行摘要

- 一句话: 调整 Intel GPU CI 测试的内存请求配置
- 推荐动作: 该 PR 属于 CI 配置微调, 技术复杂度低, 无需精读。可以关注其后续 CI 运行稳定性, 若仍有失败可进一步调整。对非 Intel GPU 团队参考价值有限。

功能与动机

根据 PR 描述, 目的是调整 Intel GPU CI 中测试任务的内存请求, 以优化资源配置。没有关联具体 Issue, 但可以推测是 CI 运行中观察到内存不足或资源浪费的情况, 需要微调各任务的 agent_tags.mem 值。

实现拆解

变更涉及 5 个 Buildkite YAML 配置文件, 每个文件只修改了 agent_tags 下的 mem 键值。具体调整如下: 1..buildkite/intel_jobs/models_multimodal_intel.yaml: 三个步骤 (qwen2、llava+qwen2_vl、other+whisper) 及 Multi-Modal Processor 步骤的 mem 从 16+ 改为 24+。 2..buildkite/hardware_tests/intel_xpu_ci/test-intel.yaml: XPU V1 test 步骤的 mem 从 16+ 改为 24+。 3..buildkite/intel_jobs/misc_intel.yaml: V1 Sample + Logits 步骤 mem 从 16+ 改为 24+; Basic Models Tests (Initialization) 从 24+ 降为 16+; NixlConnector PD accuracy 从 24+ 降为 16+。 4..buildkite/intel_jobs/models_distributed_intel.yaml: Distributed Model Tests (2 GPUs) 步骤 mem 从 24+ 降为 16+。 5..buildkite/intel_jobs/test-intel.yaml: 一个步骤的 mem 从 16+ 改为 24+。 没有测试、文档或部署相关的配套改动。

关键文件:

- .buildkite/intel_jobs/models_multimodal_intel.yaml (模块 CI 配置; 类别 config; 类型 configuration) : 修改了三个步骤 (qwen2、llava+qwen2_vl、other+whisper) 及 Multi-Modal Processor 步骤的 mem 从 16+ 改为 24+, 共 4 处变更。
- .buildkite/intel_jobs/misc_intel.yaml (模块 CI 配置; 类别 config; 类型 configuration) : 包含 3 处调整: V1 Sample + Logits 从 16+ 升到 24+, Basic Models 从 24+ 降到 16+, NixlConnector PD accuracy 从 24+ 降到 16+。
- .buildkite/hardware_tests/intel_xpu_ci/test-intel.yaml (模块 CI 配置; 类别 test; 类型 test-coverage) : XPU V1 test 步骤的 mem 从 16+ 改为 24+

- .buildkite/intel_jobs/models_distributed_intel.yaml (模块 CI 配置; 类别 config; 类型 configuration) : Distributed Model Tests (2 GPUs) 步骤的 mem 从 24+ 降为 16+
- .buildkite/intel_jobs/test-intel.yaml (模块 CI 配置; 类别 config; 类型 configuration) : 一个步骤的 mem 从 16+ 改为 24+

关键符号: 未识别

评论区精华

无 review 讨论。Claude 机器人自动评论表示由于 PR 来自 fork 而跳过了代码审查, 但仓库维护者 jikunshang 批准了该 PR。

- 暂无高价值评论线程

风险与影响

- 风险: 风险极低, 仅涉及 CI 配置中的内存请求标签。如果某个任务实际内存需求超过新设定的值, 可能导致测试因资源不足而失败; 反之, 如果降低太多, 可能影响并行执行。但由于这些调整是基于实际运行经验 (推测), 风险可控。没有代码逻辑变更, 不会有回归或安全问题。
- 影响: 影响范围仅限于 Intel GPU CI 任务调度: 提升部分任务的内存请求 (16+ → 24+) 可减少因内存不足导致的失败; 降低部分任务的内存请求 (24+ → 16+) 可释放节点资源, 提高 CI 并行度。对用户、系统其他部分无影响。影响程度低。
- 风险标记: 暂无

关联脉络

- 暂无明显关联 PR