PR #42356 完整报告

vllm-project/vllm

[CI] Migrate more B200 jobs to b200-k8s queue

合并时间：2026-05-12 15:38

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42356

执行摘要

- 一句话：迁移 4 个 B200 CI 任务到新 k8s 队列
- 推荐动作：该 PR 属常规基础设施迁移，技术复杂度低。建议关注后续 PR #42387 中剩余 3 个任务的迁移和测试修复。

功能与动机

将 B200 CI 任务迁移到 k8s 队列，以统一基础设施、提升调度效率和运维一致性。PR body 指出这 4 个任务已在 build #65711 中通过 b200-k8s 队列验证，剩余 3 个任务因预置失败留待 #42387 处理。

实现拆解

1. 修改 CI 配置文件中的设备标签：在三个 YAML 文件中将 *device*: b200 改为 *device*: b200-k8s。 - .buildkite/test_areas/lm_eval.yaml：迁移 MoE Refactor Integration Test (B200 DP - TEMPORARY) 和 GPQA Eval (GPT-OSS) (B200)。 - .buildkite/test_areas/kernels.yaml：迁移 Kernels FusedMoE Layer Test (2 B200s)。 - .buildkite/test_areas/spec_decode.yaml：迁移 Spec Decode MTP hybrid (B200)。
2. 缩窄变更范围：第一个 commit 原本迁移了 7 个任务，第二个 commit 回退了 3 个因预置失败的任务，仅保留 4 个已验证的任务。
3. 验证：在 build #65711 中通过 NOAUTO=1 触发验证，所有 4 个任务在 b200-k8s 队列上通过。

关键文件：

- .buildkite/test_areas/lm_eval.yaml（模块 CI 配置；类别 config；类型 configuration）：迁移 2 个 B200 任务到 k8s 队列：MoE Refactor Integration Test (B200 DP) 和 GPQA Eval (B200)。
- .buildkite/test_areas/kernels.yaml（模块 CI 配置；类别 config；类型 configuration）：迁移 Kernels FusedMoE Layer Test (2 B200s) 到 k8s 队列。
- .buildkite/test_areas/spec_decode.yaml（模块 CI 配置；类别 config；类型 configuration）：迁移 Spec Decode MTP hybrid (B200) 到 k8s 队列。

关键符号：未识别

评论区精华

该 PR 没有人工 review 评论，仅有 bot 自动评论。gemini-code-assist[bot] 指出无反馈。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：
  - 仅修改 CI 配置文件中的 device 标签，不影响任何源码逻辑。
  - 4 个任务均已在新队列上通过验证（build #65711）。
  - 变更回退了 3 个有预置失败的任务，避免影响 CI 稳定性。
  - 影响：用户 / 系统：无直接影响；开发者 CI 体验不变。CI 团队：B200 任务逐步迁移到 k8s 队列，有助于统一基础设施。范围：仅 4 个 B200 测试任务，影响面小。
- 风险标记：仅配置变更，已验证通过

关联脉络

- PR #42387 [CI] Migrate remaining B200 jobs to b200-k8s with test fixes: 继续迁移剩余 3 个 B200 任务，并包含测试修复，是本 PR 的后续工作。