PR #42446 完整报告

vllm-project/vllm

[CI] Migrate 6 verified jobs from gpu_1_queue to h200_18gb MIG

合并时间：2026-05-13 02:52

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/42446

执行摘要

该 PR 将 6 个已在 H200 18GB MIG 分区上验证通过的 CI 测试任务从 L4（gpu_1_queue）迁移到 h200_18gb 设备队列，以缓解 L4 资源瓶颈。同时修复了 Acceptance Length Test 中非标准的设备字段用法。由于 PyTorch 的 CUDACachingAllocator NVML 断言问题，其余约 47 个任务暂无法迁移。

功能与动机

PR body 明确指出，目的是利用 H200 MIG 分区的闲置算力，将已验证通过的 6 个任务迁移到该分区。迁移前已在 Build #65789 上完成验证。核心限制是 PyTorch 的 CUDA 内存分配器在 MIG 分区上会触发 NVML 断言失败，影响大部分需要初始化引擎的任务。

实现拆解

1. 纯内核测试（Kernels KDA Test）：在 kernels.yaml 中添加 device: h200_18gb，该测试不涉及引擎初始化，最适合 MIG 环境。
2. 预量化模型评测（LM Eval TurboQuant KV Cache）：在 lm_eval.yaml 中添加 device: h200_18gb，使用预量化模型路径，避免重新初始化引擎。
3. 非标准字段修复（Acceptance Length Test）：在 misc.yaml 中将原来的 gpu: h100 和 num_gpus: 1 替换为 device: h200_18gb，统一设备指定方式。
4. 无需 GPU 的安装检查（Python-only Installation）：在 misc.yaml 中添加 device: h200_18gb，但该任务无需 GPU，引发资源浪费的讨论。
5. 轻量级初始化测试（Basic Models Tests）：在 models_basic.yaml 中添加 device: h200_18gb，仅运行模型初始化测试的小子集。
6. 核心语言模型测试（Language Models Tests）：在 models_language.yaml 中添加 device: h200_18gb，运行标准语言模型的核心测试。

评论区精华

> gemini-code-assist[bot]：Python-only Installation 任务无需 GPU，分配到 h200_18gb 浪费资源，建议迁移到 CPU-only 设备（如 cpu-small）。

该评论未在 PR 中得到作者或维护者的回复，最终 PR 保持原配置合并。

风险与影响

- 资源浪费：Python-only Installation 任务占用 H200 MIG 分区，属于昂贵的 GPU 资源浪费。
- MIG 兼容性风险：PyTorch 的 NVML 断言问题意味着未来扩展需慎重选择任务类型，仅纯内核测试或预量化模型评测适合 MIG 环境。
- 影响范围：仅 CI 配置文件变更，对用户无直接影响。可缓解 L4 队列压力，但长期可靠性和效率需持续观察。

关联脉络

该 PR 与 #42401（MIG 分区兼容性分析）密切相关，前者提供了 NVML 断言失败的完整分析，决定了本次迁移的范围。后续可能继续探索如何解决MIG上的兼容性问题，以扩大迁移范围。