

PR #54037 完整报告

vllm-project/vllm

[CI][ROCm] Expand weight loading test coverage on AMD and cap its KV cache

合并时间: 2026-09-01 15:57

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/54037>

执行摘要

- 一句话: 扩展 AMD 权重加载测试模型列表并为共享测试限制 KV 缓存大小以优化 CI 运行时。
- 推荐动作: 该PR主要属于CI基础设施优化和测试覆盖扩展, 不涉及核心运行时逻辑变更。
对于维护 ROCm 支持或 CI 系统的工程师, 值得 精读以理解模型列表筛选逻辑和 KV 缓存优化策略。对于关注核心功能的工程师, 可以 快速浏览其动机和结果。关键的设计决策是:
1) 通过分析 models.txt 筛选出 ROCm 支持的子集, 而非新建独立列表; 2) 在共享测试中引入一个保守的 KV 缓存上限以优化运行时, 同时避免影响测试有效性。

功能与动机

根据 PR body, 动机主要有两点: 1) AMD 的权重加载作业仅测试 3 个模型, 而 NVIDIA 测试 18 个, 导致许多 AMD 支持的量化格式 (如 TP=2 时的压缩张量和 AWQ) 未被覆盖。2) 测试本身仅生成 20 个 token, 但 vLLM 会根据 `gpu_memory_utilization` 为每个模型分配完整的 KV 缓存, 这在大内存设备 (如 MI300/MI355) 上主导了运行时间, 需要进行优化。

实现拆解

1. 扩展测试模型列表: 在 tests/weight_loading/models-amd.txt 文件中新增 14 个模型, 涵盖 GPTQ (无激活重排)、compressed-tensors (w8a8/w8a16/w4a16 和 FP8) 以及 AWQ 格式。此变更的目的是确保 ROCm 平台在 TP=2 时能测试到其支持的多种量化格式。
2. 优化测试运行时: 修改共享的测试脚本 tests/weight_loading/test_weight_loading.py, 导入 GiB_bytes 常量, 并在 vllm_runner 调用中新增 kv_cache_memory_bytes=2 * GiB_bytes 参数, 将 KV 缓存限制为 2 GiB。此改动适用于所有使用此测试的 CI 作业, 原因在于测试仅需生成少量 token, 而分配过大缓存是性能瓶颈。
3. 更新 CI 配置: 修改 .buildkite/test_areas/weight_loading.yaml, 将 AMD 镜像作业的设备从 mi300_2 切换到更新的 mi355_2, 标签相应更新为 (MI355), 并将超时时间从 35 分钟提高到 40 分钟以适应新的模型列表。这确保了测试在最新的 AMD 硬件上运行。
4. 配套调整: 在提交历史中, 有一个提交 (f6c9441e) 移除了 models-amd.txt 中的 awq_marlin 条目, 因为它在 ROCm 上解析为 auto_awq, 与 awq 条目重复, 不提供额外覆盖。

关键文件:

- tests/weight_loading/test_weight_loading.py (模块 权重加载测试; 类别 test; 类型 test-coverage): 共享测试脚本的核心修改, 引入了 KV 缓存大小限制以优化所有相关 CI

作业的运行时。

- `.buildkite/test_areas/weight_loading.yaml` (模块 CI 配置; 类别 config; 类型 configuration) : CI 配置文件, 定义了权重加载测试的设备、超时和执行命令, 本次变更新增了 AMD 作业的硬件和超时设置。
- `tests/weight_loading/models-amd.txt` (模块 模型列表; 类别 docs; 类型 documentation) : AMD 测试的模型列表, 新增了 14 个模型以大幅扩展测试覆盖范围, 是本次变更的主要内容之一。

关键符号: `test_weight_loading`

关键源码片段

`tests/weight_loading/test_weight_loading.py`

共享测试脚本的核心修改, 引入了 KV 缓存大小限制以优化所有相关 CI 作业的运行时。

```
# 导入 GiB 字节数常量
from vllm.utils.mem_constants import GiB_bytes

MAX_MODEL_LEN = 1024 # 保留原有模型长度限制
# ... 其他环境变量获取 ...

def test_weight_loading(vllm_runner):
    """
    测试 TP>1 时的参数权重加载。
    """
    # MoE 模型需要 fp16
    NEEDS_FP16 = (
        QUANTIZATION == "gptq"
        or MODEL_NAME == "nm-testing/test-w4a16-mixtral-actorder-group"
    )
    with vllm_runner(
        model_name=MODEL_NAME,
        revision=REVISION,
        dtype=torch.half if NEEDS_FP16 else "auto",
        quantization=None if QUANTIZATION == "None" else QUANTIZATION,
        max_model_len=MAX_MODEL_LEN,
        tensor_parallel_size=2,
        # 关键变更: 限制 KV 缓存为 2 GiB。此测试仅生成 20 个 token,
        # 而在大型设备上分配完整缓存会成为运行时瓶颈。
        kv_cache_memory_bytes=2 * GiB_bytes,
    ) as model:
        output = model.generate_greedy("Hello world!", max_tokens=20)
        print(output)
        assert output
```

评论区精华

核心讨论

tjtanaa: @AndreasKaratzas can you check if we want to create this as another mi355 tests and keep the mi300 test in parallel? stefankoncarevic: I discussed this with @AndreasKaratzas the plan is to move this group to MI355 upstream, while our internal CI keeps running it on both MI355 and MI300.

分析：讨论聚焦于 CI 设备的升级决策。质疑者（tjtanaa）考虑是否应保留 MI300 测试以确保兼容性。作者（stefankoncarevic）澄清，与维护者（AndreasKaratzas）的决策是将上游 CI 统一迁移到更新的 MI355 设备，而 AMD 的内部 CI 将继续在两种设备上运行以确保双重覆盖。这表明了升级是经过协调的，旨在使用更新硬件并优化 CI 资源，同时通过内部流程保持对旧设备的验证。

未解决的疑虑

评论中未明确讨论 KV 缓存限制（2 GiB）对 NVIDIA 测试的影响。虽然 PR 说明提到 `max_model_len=1024` 在 TP=2 时所需空间远小于 2 GiB，但并未展示在 NVIDIA 硬件上的具体验证结果。这是一个微小的潜在风险点。

- CI 设备升级策略（MI300 vs MI355）（design）：决策已定：上游 CI 升级到 MI355 以使用更新硬件并可能优化资源，兼容性通过 AMD 内部 CI 在两种设备上的并行测试来保证。

风险与影响

- 风险：### 技术风险 1. 测试覆盖风险（低）：扩大模型列表主要是增加测试用例，旨在提升覆盖率，而非改变核心逻辑。风险在于新添加的模型（特别是 GPTQ 模型）是否真的在所有目标 CI 环境中稳定通过。PR 作者通过实际测试（18 个模型在 MI355 上全部通过）进行了验证。 2. CI 配置风险（低）：从 MI300 切换到 MI355 设备并调整超时时间属于 CI 基础设施变更。风险是 MI355 设备的可用性或稳定性问题可能导致 CI 失败，但这属于运维范畴，且超时时间已相应增加。 3. 共享测试影响风险（低）：kv_cache_memory_bytes 的修改应用于共享的 test_weight_loading.py 脚本，影响 NVIDIA 和 AMD 的所有 CI 作业。由于限制值（2 GiB）远大于测试所需（max_model_len=1024），且修改前测试已能通过，因此引入正确性回归的风险极低。主要影响是积极的——加速了所有相关 CI 作业。

影响范围

- 直接影响：CI 流程。AMD 权重加载 CI 作业将测试更多模型并在 MI355 上运行；所有权重加载 CI 作业（包括 NVIDIA）的运行时间将缩短。
- 间接影响：项目对 ROCm 平台量化格式支持的信心。更广泛的测试覆盖有助于提前发现兼容性问题。
- 影响：### 对用户的影响 无直接影响。这是 CI 层面的变更，不会改变 vLLM 的功能或 API。

对系统的影响

1. CI 稳定性与速度：AMD CI 作业的测试覆盖范围显著扩大，有助于更早地捕获 ROCm 平台相关的回归。KV 缓存限制优化减少了所有权重加载测试的运行时间（据称 MI355 上节省约 16%），提升了 CI 效率。
2. 硬件支持：CI 从 MI300 迁移到 MI355，反映了项目向更新 AMD 硬件过渡的趋势，可能影响未来的硬件兼容性优先级。

对团队的影响

- CI/Infra 团队：需要维护更新的 MI355 设备配置，并可能观察测试运行时间的改善。
- AMD 支持团队：测试覆盖的扩展为验证 ROCm 上更多量化格式提供了保障，减轻了手动测试负担。
- 所有贡献者：共享测试的运行时间缩短，可能间接加快了 CI 反馈循环。
- 风险标记：CI 配置变更，共享测试修改

关联脉络

- PR #54403 [ROCm][CI] Stabilize the sqrt-softplus top-k tie oracle: 同为 ROCm 平台的 CI 测试稳定性改进，关注点从模型权重加载测试扩展到了内核测试。
- PR #54408 [CI][ROCm] Avoid redundant image pulls during smoke validation: 同为优化 ROCm CI 流程效率的 PR，通过调整构建步骤减少重复拉取镜像来节省时间。