

PR #45955 完整报告

vllm-project/vllm

[ROCm][CI] Enable kv_connector unit tests on ROCm

合并时间: 2026-06-22 10:08

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45955>

执行摘要

- 一句话: 在 ROCm CI 上启用 KV Connector 单元测试
- 推荐动作: 值得关注的设计决策: 如何在对平台敏感的测试中优雅跳过不适用部分 (将平台判断内聚在测试辅助函数中)。CI 配置的 mirror 用法可作为 AMD CI 作业的参考模式。

功能与动机

PR body 指出: MRV2 默认启用后 `test_offloading_connector` 在 ROCm 上出现断言失败 (`assert 0 >= 0.8 * 10`) , 原因是 V2 实现缺少跨层 KV 缓存支持, 导致 ROCm 上微小副本过多影响性能。`test_multi_example_connector_consistency` 在异步调度下本质不稳定, 需要显式禁用异步调度。

实现拆解

1. 修改 `tests/v1/kv_connector/unit/test_offloading_connector.py`: 在 `_latency_test` 函数开头添加 `if current_platform.is_rocm(): return`, 跳过整个延迟测试, 并添加 TODO 引用 #45947 跟踪后续恢复。
2. 修改 `tests/v1/kv_connector/unit/test_multi_connector.py`: 在 `test_multi_example_connector_consistency` 的 LLM 初始化中添加 `async_scheduling=False`, 避免异步调度导致日志事件顺序混乱。
3. 修改 `.buildkite/scripts/install-kv-connectors.sh`: 在安装依赖前检测是否 ROCm (通过检查 `torch` 的 `hip`) , 若是则安装 `kv_connectors_rocm.txt` 中的依赖。
4. 修改 `.buildkite/test_areas/misc.yaml`: 为 V1 Others (GPU) 步骤添加 `mirror.amd` 配置, 指定设备 `mi325_1`、超时和基础镜像依赖, 使这些测试在 AMD CI 上作为 gating 运行。

关键文件:

- `tests/v1/kv_connector/unit/test_offloading_connector.py` (模块 卸载测试; 类别 test; 类型 test-coverage) : 核心测试文件: 在 `_latency_test` 中添加 ROCm 跳过逻辑, 并清理冗余的 `monkeypatch` 参数, 使延迟测试在 ROCm 上安全跳过。
- `.buildkite/scripts/install-kv-connectors.sh` (模块 安装脚本; 类别 other; 类型 core-logic) : 安装脚本: 添加 ROCm 检测分支, 确保在 AMD CI 上安装 `kv_connectors_rocm.txt` 依赖, 而非通用依赖。
- `tests/v1/kv_connector/unit/test_multi_connector.py` (模块 多连接器测试; 类别 test; 类型 test-coverage) : 测试文件: 为 `test_multi_example_connector_consistency` 添加

`async_scheduling=False`, 避免异步调度导致日志事件顺序混乱, 使测试稳定。

- `.buildkite/test_areas/misc.yaml` (模块 CI 配置; 类别 `config`; 类型 `configuration`) : CI 配置: 为 V1 Others (GPU) 步骤添加 AMD 镜像配置 (`mi325_1`、超时等), 使 KV Connector 测试在 AMD CI 上作为 `gating` 运行。

关键符号: `_latency_test`

关键源码片段

`tests/v1/kv_connector/unit/test_offloading_connector.py`

核心测试文件: 在 `_latency_test` 中添加 ROCm 跳过逻辑, 并清理冗余的 `monkeypatch` 参数, 使延迟测试在 ROCm 上安全跳过。

```
def _latency_test(llm: LLM, subscriber: MockSubscriber | None):
    # 在 ROCm 上跳过延迟测试, 因为 MRV2 尚不支持跨层 KV 缓存
    # 这会导致大量小维度 <-> H 的副本, 在 AMD MI300 上性能不佳
    # TODO: 在 MRV2 实现跨层 KV 缓存后恢复 (#45947)
    if current_platform.is_rocm():
        return
    sampling_params = SamplingParams(max_tokens=1)
    # ... 后续冷启动、GPU 命中、CPU 命中测量逻辑保持不变 ...
```

`.buildkite/scripts/install-kv-connectors.sh`

安装脚本: 添加 ROCm 检测分支, 确保在 AMD CI 上安装 `kv_connectors_rocm.txt` 依赖, 而非通用依赖。

```
# 如果是 ROCm 环境 (通过检查 torch 的 hip 版本), 则使用特定的 requirements 文件
if python3 -c "import torch; raise SystemExit(0 if torch.version.hip is not None else 1)"; then
    uv pip install --system -r /vllm-workspace/requirements/kv_connectors_rocm.txt
    exit 0
fi
# 非 ROCm 平台继续使用通用 requirements
REQUIREMENTS_FILE="${KV_CONNECTORS_REQUIREMENTS:-/vllm-workspace/requirements/kv_connectors.txt}"
uv pip install --system -r "${REQUIREMENTS_FILE}"
```

评论区精华

Reviewer orozery 建议将 ROCm 跳过检查从三个调用点移到 `_latency_test` 内部以消除重复; 作者采纳。orozery 还指出 `monkeypatch` 参数未被使用, 建议移除; 作者移除。orozery 询问 `test_multi_example_connector_consistency` 的 `async_scheduling=False` 及其他 CI 改动是否意图明确; 作者确认是有意为之, 并调整 PR 标题反映更广泛目的。最终两个 reviewer 都批准。

- 将 ROCm 跳过检查移到 `_latency_test` 内部 (design): 作者采纳, 改为在 `_latency_test` 函数开头检查并 `return`。
- 移除未使用的 `monkeypatch` 参数 (style): 作者确认并移除这些参数。

- test_multi_example_connector_consistency 添加 async_scheduling=False 的意图 (question): 作者确认是有意改动, 并解释禁用异步调度是为了避免测试日志顺序混乱, CI 改动是为了在 AMD 上运行这些测试。

风险与影响

- 风险: 延迟测试 (_latency_test) 在 ROCm 上被完全跳过, 这意味着 ROCm 将不再检测 KV offloading 的延迟性能回归, 但核心功能测试 (_accuracy_test) 仍然保留。禁用异步调度使 test_multi_example_connector_consistency 不再测试并发调度行为, 可能遗漏调度器相关 bug。CI 配置改动仅影响 AMD 构建, 无引入风险。
- 影响: 仅影响 ROCm CI 测试流程: 确保 KV Connector 单元测试在 AMD GPU 上通过, 停止因 MRV2 兼容问题导致的持续失败。对用户无影响。对维护者: 需要在后续添加跨层 KV 缓存支持后恢复延迟测试 (跟踪 #45947)。
- 风险标记: 跳过 ROCm 延迟测试, 禁用异步调度

关联脉络

- PR #43458 [Core] Enable MRV2 by default: 导致 offloading 测试在 ROCm 上失败, 是本 PR 要解决的主要回归来源。
- PR #45947 [Core] Add cross-layer KV cache support for MRV2: 未来的恢复依赖, 本 PR 中跳过延迟测试的 TODO 关联到此 PR。