

PR #49909 完整报告

vllm-project/vllm

[ROCm] Use backend-default dot precision for ReplaySSM

合并时间: 2026-07-27 18:11

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/49909>

执行摘要

- 一句话: 修复 ROCm Triton dot 精度参数不兼容
- 推荐动作: 值得集成的低风险修复。PR 展示了一种通过平台 API 处理 CUDA/ROCm 差异的良好实践, 可推广到其他 Triton 内核。

功能与动机

ROCm Triton 不支持 CUDA-only 的 tf32x3 精度参数。PR body 中说明“Let ROCm Triton select its native dot precision instead of passing the unsupported CUDA-only tf32x3 value”。需修复以在 ROCm 上正常运行 Mamba 模型。

实现拆解

1. 引入平台抽象: 在 `selective_state_update_replayssm_output_only.py` 中导入 `vllm.platforms.current_platform`, 通过平台 API 判断当前是否为 ROCm, 动态设置 `dot_input_precision` 为 `None` (ROCm 默认) 或 `"tf32x3"` (CUDA)。
2. 修改内核调用: 将硬编码的 `input_precision="tf32x3"` 替换为内核参数 `DOT_INPUT_PRECISION`, 并在 `selective_state_update_replayssm_output_only` 函数中根据平台传入该参数。
3. 调整 CI 配置: 在 `.buildkite/test-amd.yaml` 中将 Mamba 测试从 MI250 (`amdgfx90anightly, amdmi250`) 迁移到 MI300 (`amdgfx942nightly, amdmi300`), 确保 ROCm 默认路径被覆盖; 同时更新注释明确解释精度选择逻辑。
4. 格式化改进: 对 `HAS_STATE_BATCH_INDICES` 启发式函数进行了格式化调整 (括号换行), 无逻辑变更。

关键文件:

- `vllm/model_executor/layers/mamba/ops/selective_state_update_replayssm_output_only.py` (模块 Mamba 内核; 类别 `infra`; 类型 `infrastructure`; 符号 `selective_state_update_replayssm_output_only, _replayssm_output_only_kernel`): 核心变更文件, 引入平台 API 并动态设置 dot 精度参数, 修复 ROCm Triton 不兼容问题。
- `.buildkite/test-amd.yaml` (模块 CI 配置; 类别 `config`; 类型 `configuration`): 将 Mamba 测试从 MI250 迁移至 MI300, 确保 ROCm 默认路径被覆盖。

关键符号: selective_state_update_replayssm_output_only,
_replayssm_output_only_kernel

评论区精华

无 review 评论。仅 [claude\[bot\]](#) 自动注释说明 fork PR 跳过自动审查, [tjtanaa](#) 直接批准。

- 暂无高价值评论线程

风险与影响

- 风险: 低风险。变更仅影响 ROCm 平台的 Triton dot 精度, CUDA 侧无功能变化。CI 配置将 Mamba 测试从 MI250 迁移至 MI300, 可能覆盖现有测试依赖; 但测试仍保持可选 (optional: true), 不影响主干 CI。
- 影响: 直接修复 ROCm Triton dot 精度不兼容问题, 使 ReplaySSM 内核在 ROCm 上可正常运行。对 CUDA 用户无感。CI 配置调整后, Mamba 测试在 MI300 上运行, MI250 不再执行此测试, 可能减少测试覆盖范围但符合硬件分组优化。
- 风险标记: CI 测试迁移可能覆盖丢失

关联脉络

- 暂无明显关联 PR