PR #38504 完整报告

vllm-project/vllm

[Kernels][MoE] Fix legacy_routing to use bitmatrix-based routing path

合并时间：2026-04-07 10:57

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/38504

执行摘要

- 一句话：修复 MoE 路由中 HIP 平台整数除法导致的 bitmatrix 错误，避免 GPU 内存访问故障。
- 推荐动作：建议技术管理者和工程师精读此 PR，关注 pack_bitmatrix 的 valid guard 设计和平台条件化对齐策略，这些是处理跨平台差异的典型模式。同时，可参考相关 MoE refactor PR 以了解路由演进脉络。

功能与动机

PR body 指出：'Three issues prevented test_gpt_oss_triton_kernels and GPQA serving from working on gfx950 (CDNA4) after the legacy routing deprecation'，详细描述了 pack_bitmatrix 中的未定义行为导致专家 31 污染 bitmatrix，引发 GPU 内存访问故障。根因是 HIP 使用 C 风格整数除法，-1 // 32 = 0 且 1 << -1 设置比特 31，破坏路由元数据。

实现拆解

实现分为四部分：1) 在 vllm/model_executor/layers/fused_moe/gpt_oss_triton_kernels_moe.py 的 pack_bitmatrix 函数中添加 valid = indices >= 0 guard，防止负索引产生伪比特；2) 在 tests/kernels/moe/test_gpt_oss_triton_kernels.py 中条件化 padding 对齐，ROCm 使用 256/512 匹配 CDNA4 硬件要求，CUDA 保持 64/128；3) 更新四个 GPQA eval 配置文件，添加 --tokenizer openai/gpt-oss-20b 和 --tensor-parallel-size 2 参数；4) 移除 legacy_routing_from_sparsematrix 等未使用代码，统一使用 bitmatrix-based 路由路径。

关键文件：

- vllm/model_executor/layers/fused_moe/gpt_oss_triton_kernels_moe.py（模块 MoE）：核心路由逻辑修改，添加 valid guard 防止 HIP 平台伪比特，并移除 legacy 函数，统一路由路径。
- tests/kernels/moe/test_gpt_oss_triton_kernels.py（模块 测试）：测试调整，平台条件化 padding 对齐，确保 CDNA4 和 Hopper 硬件兼容性。
- tests/evals/gpt_oss/configs/gpt-oss-20b-rocm-quark-mxfp4-bf16-aiter.yaml（模块 配置）：serving 配置文件更新，添加 tokenizer 和 TP 参数，修复 GPU 内存故障。

关键符号：pack_bitmatrix, legacy_routing

评论区精华

review 中核心讨论：1) tjtanaa 询问更改是否来自另一个 PR #38503，AndreasKaratzas 确认并恢复相关 scheduler 代码，聚焦于 MoE 修复；2) tjtanaa 建议移除未使用的 legacy_routing_from_sparsematrix 函数并要求披露模型准确性，AndreasKaratzas 回应已移除并启用 GPQA eval 测试，提供了通过结果。讨论强调了代码清理和测试验证。

- legacy 路由代码移除 (design): 作者已移除该函数，并启用 GPQA eval 测试验证准确性。
- 模型准确性验证 (testing): 测试通过，准确性在允许范围内，确保修复不引入回归。

风险与影响

- 风险：风险包括：1) 核心路径变更：pack_bitmatrix 修改可能影响所有 MoE 路由逻辑，需确保 CUDA 平台行为不变（作者说明为 no-op）；2) 跨平台不一致：条件化对齐引入维护负担，未来硬件变更需调整；3) legacy 代码移除可能破坏潜在依赖，但鉴于 deprecation，风险较低；4) 配置文件修改可能导致其他环境 serving 错误，需验证参数适用性。
- 影响：影响范围：1) 用户：在 AMD GPU（如 gfx950）上运行 GPT-OSS 模型的用户将避免 GPU 崩溃，提升 serving 稳定性和准确性；2) 系统：MoE 路由更健壮，测试套件通过率从 1/5 提升至 5/5，GPQA eval 全部通过；3) 团队：代码更简洁，legacy 代码减少，跨平台支持增强。影响程度中等，主要针对特定硬件和模型。
- 风险标记：核心路径变更，跨平台不一致，legacy 代码移除

关联脉络

- PR #38503 未知：review 中提及，可能涉及相关 scheduler 更改，但本 PR 已恢复那些修改。
- PR #38251 [Quantization] Add FlashInfer CuteDSL batched experts backend for NVFP4 MoE: 同为 MoE 相关 PR，涉及量化后端，可能有代码结构交叉。
- PR #35326 [MoE Refactor] Split of DefaultMoERunner class: 同为 MoE 重构 PR，体现了路由和 runner 模块的演进趋势。