PR #36799 完整报告

vllm-project/vllm

[Sparse24] [Deprecation] Remove Sparse24 CT integration and kernels

合并时间：2026-03-24 04:03

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/36799

执行摘要

本次 PR 移除了 vLLM 中对 Sparse24 稀疏量化模型的支持，包括压缩张量集成和 CUDA 内核实现，以降低维护负担并减小二进制体积。变更通过直接删除相关代码和在关键类抛出 NotImplementedError 实现，review 讨论确认了无需弃用期的决策。

功能与动机

Sparse24 模型因使用不广泛，成为维护负担。PR body 明确指出：" 减少 vLLM 和 Compressed Tensors 的维护负担，移除对不广泛使用的 Sparse24 模型的支持 " 和 " 移除内核以减少二进制大小 "。这体现了团队聚焦核心功能、优化代码健康的意图。

实现拆解

变更涵盖多个层次：

- 量化方案层：compressed_tensors_24.py 中 CompressedTensors24 类的 __init__ 和 create_weights 方法改为抛出 NotImplementedError。
- 构建系统：CMakeLists.txt 移除稀疏内核的编译配置，避免生成冗余代码。
- 内核实现：删除 csrc/sparse/cutlass/ 目录下的所有 CUDA 文件，如 sparse_scaled_mm_c3x.cu。
- 测试与基准：移除 Python 测试文件 test_cutlass_2of4_sparse.py 和基准脚本 sparse_benchmarks.py。
- 接口绑定：更新 csrc/torch_bindings.cpp 和 _custom_ops.py 移除 C++/Python 函数暴露。

评论区精华

Review 中仅有一次简短讨论：

> yewentao256: "Should we have a period of deprecation, or just deleting this? @mgoin CC" mgoin: "No need, there are no downloads on the models"

这明确了直接删除的依据是模型无实际使用数据，简化了弃用流程。

风险与影响

风险：主要影响尝试加载 Sparse24 模型的极少数用户，会立即遇到 NotImplementedError；但因模型无下载，回归风险低。影响：用户侧需迁移或放弃使用；系统侧二进制减小、编译加速；团队侧维护负担降低。

关联脉络

从近期 PR 看，无直接相关的变更；但量化领域（如 PR 32929 的 FP8 抽象）和性能优化（如 PR 36725 的 MoE 修复）展示了 vLLM 持续演进中对效率的追求，本次清理与之协同，保持代码库精简。