PR #32694 完整报告

vllm-project/vllm

[Quantization][Deprecation] Remove Petit NVFP4

合并时间：2026-04-05 08:07

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/32694

执行摘要

本 PR 移除 vLLM 中已被废弃的 Petit NVFP4 量化支持，全面删除相关代码和配置，以减少维护负担并遵循版本废弃策略。变更影响仅限于使用该量化方法的 AMD GPU 用户，需迁移到其他方案。

功能与动机

Petit NVFP4 是 vLLM 中针对 AMD GPU 的量化方法，已在 0.14 版本标记为废弃。PR body 明确指出：'now that 0.14 is released with deprecation notice, remove Petit NVFP4'，动机是清理不再维护的功能，降低代码库复杂性。

实现拆解

实现方案按模块拆解如下：

- 依赖管理：从 setup.py 中移除 petit-kernel 可选依赖。
- 配置验证：从 vllm/config/model.py 的量化验证列表移除 petit_nvfp4。
- 线性方法注册：从 vllm/model_executor/layers/linear.py 的线性方法列表移除 PetitNvFp4LinearMethod。
- 量化模块核心：删除 vllm/model_executor/layers/quantization/petit.py 和 vllm/model_executor/layers/quantization/utils/petit_utils.py，移除 PetitNvFp4Config 类和相关工具函数。
- 平台支持：从 vllm/platforms/rocm.py 的量化方法列表移除 petit_nvfp4。 所有变更均为删除操作，无新增代码逻辑。

评论区精华

review 讨论极为简短，仅 reviewer yewentao256 给出批准评论：

> LGTM, thanks for the work! 无争议点或设计权衡，表明变更被团队迅速接纳。

风险与影响

风险：

- 回归风险：若遗留代码依赖 Petit NVFP4，可能导致 ImportError，但 PR 已全面移除引用，需通过测试覆盖验证。

- 兼容性风险：使用该量化的模型将无法加载，用户需迁移到其他量化方法（如 FP8），但鉴于已废弃，影响可控。

影响：

- 用户：AMD GPU 用户需切换量化方案，可能带来短期不便。
- 系统：简化代码库，减少维护开销。
- 团队：开发者可聚焦于支持的量化功能。

关联脉络

从近期历史 PR 看，量化模块频繁维护（如 PR 38870 修复 FP8 量化 bug），本 PR 是量化功能演进的清理步骤。同时，ROCM 平台相关 PR（如 PR 38959）显示基础设施持续优化，本 PR 与之协同确保平台配置一致性。整体脉络指向量化支持的简化和标准化。