PR #44626 完整报告

vllm-project/vllm

[ROCm][AITER][Quark] Tag per-channel FP8 weights as PER_CHANNEL so AITER pre-shuffled GEMM is selected

合并时间：2026-06-17 22:05

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/44626

执行摘要

- 一句话：修复 Quark FP8 权重 scale 标签错误以启用 AITER 预洗牌 GEMM
- 推荐动作：建议精读此 PR，尤其是讨论中关于 vLLM 内部权重 scale 标签不一致的深入分析。该修复展示了正确的方案：当框架内部表示不统一时，优先在源头改正（采用压缩张量规范），而不是在消费端做兼容 hack。

功能与动机

在 ROCm 上启用 AITER 时，Quark 量化的 FP8 attention projection 权重 scale 被错误标记为 GroupShape.PER_TOKEN，导致 AITER 门控函数无法识别为 per-channel，从而回退到通用 hipBLASLt 路径。这是一个 vLLM 侧的标记 bug（per-channel 权重应使用 GroupShape.PER_CHANNEL），需在源头修复。

实现拆解

1. 修改 import 语句：在 quark_w8a8_fp8.py 中将 kFp8StaticTokenSym 替换为 kFp8StaticChannelSym，并删除不再使用的旧符号。
2. 调整变量命名与标签逻辑：将 per_token_weight 重命名为 per_channel_weight，并将 self.weight_quant_key 从 kFp8StaticTokenSym 改为 kFp8StaticChannelSym，以正确反映 per-channel 语义。
3. 保持其他逻辑不变：process_weights_after_loading 等方法无需修改，因为物理布局和 scale 数值完全不变，仅是标签对齐。
4. 回退 AITER kernel 门控的临时放宽：由于源头修复后门控自动生效，之前为绕过 bug 而尝试放宽 AITER 门控的变更被放弃。
5. 测试与验证：在 Kimi-K2.5-MXFP4-AttnFP8 模型上进行了性能基准测试（TP=4、不同并发度）和精度验证（GSM8K 准确率 ~94%），确认加速效果且无精度下降。

关键文件：

- vllm/model_executor/layers/quantization/quark/schemes/quark_w8a8_fp8.py（模块 量化；类别 source；类型 data-contract）：修改了 FP8 per-channel 权重量化标签，从 kFp8StaticTokenSym 改为 kFp8StaticChannelSym，是 PR 核心变更文件。
- vllm/model_executor/kernels/linear/scaled_mm/aiter.py（模块 kernel；类别 source；类型 reverted）：该文件原本计划放宽门控条件以绕过 Per_TOKEN 限制，但 review 后决定

回退所有修改，改为在源头修复。最终版本未修改此文件。

关键符号：QuarkW8A8Fp8.init

关键源码片段

vllm/model_executor/layers/quantization/quark/schemes/quark_w8a8_fp8.py

修改了 FP8 per-channel 权重量化标签，从 kFp8StaticTokenSym 改为 kFp8StaticChannelSym，是 PR 核心变更文件。

```
# vllm/model_executor/layers/quantization/quark/schemes/quark_w8a8_fp8.py (partial)

# ... (imports)
from vllm.model_executor.layers.quantization.utils.quant_utils import (
    GroupShape,
    kFp8DynamicTokenSym,
    kFp8StaticChannelSym, # 新增：per-channel 权重 scale 的标识符
    kFp8StaticTensorSym,
    # kFp8StaticTokenSym removed — 此标识符已被 kFp8StaticChannelSym 取代
)

class QuarkW8A8Fp8(QuarkScheme):
    def __init__(self, weight_config, input_config=None):
        self.weight_qscheme = cast(str, weight_config.get("qscheme"))
        # ... (activation 侧处理不变 )

        # 将变量名从 per_token_weight 改为 per_channel_weight，更准确地表达语义
        per_channel_weight = self.weight_qscheme == "per_channel"

        # 关键修复：per-channel 权重的 quant_key 从 kFp8StaticTokenSym 改为
        # kFp8StaticChannelSym，匹配 compressed-tensors 的 CHANNEL strategy。
        # 这样 AITER 门控函数 is_per_channel() 就能正确识别，自动选择
        # 预洗牌 GEMM（gemm_a8w8_bpreshuffle）路径，从而获得 1.13-1.15x 加速。
        self.weight_quant_key = (
            kFp8StaticChannelSym if per_channel_weight else kFp8StaticTensorSym
        )
        self.out_dtype = torch.get_default_dtype()
```

评论区精华

核心讨论发生在 review 中，tjtanaa 指出这是一个 vLLM 的 labelling bug——per-channel 权重本应使用 GroupShape.PER_CHANNEL，并指出应遵循 compressed-tensors 的约定，而不是在 kernel 门控中做 workaround。作者 xaguilar-amd 同意并重构了 PR，仅修改 Quark 源头标签。此外，tjtanaa 和 dllehr-amd 要求提供更广泛的 benchmark 覆盖，作者补充了多并发度下吞吐和 TPOT 数据，并展示了 profiler 截图确认 AITER 预洗牌路径被正确选中。BowenBao 确认本地已有类似修复，LGTM 并批准。

- Per-channel 权重 scale 标记错误应在源头修复而非 kernel 门控中做 workaround (design): 作者同意并重构 PR，仅修改 Quark 源头标签，回退 AITER 门控修改。
- 是否需要更全面的性能 benchmark 验证 (testing): 补充了足够的数据，确认加速效果且无精度下降。

风险与影响

- 风险：该 PR 仅更改了权重 quant_key 的标签常量，不修改任何数值、scale 或 kernel 代码，回归风险极低。主要风险在于：Quark 之外的 ModelOpt 和 fbgemm 存在相同的 mislabeling，但已在 scope 外声明，可能被误认为一并修复。
- 影响：影响范围：仅影响使用 Quark 量化的 FP8 per-channel 权重场景（主要在 ROCm + AITER 环境下）。对用户端可见的性能提升为 decode/prefill FP8 attention GEMM 约 13% 加速，端到端吞吐提升约 2-7%。未修改配置接口或 API，对非 Quark 用户无影响。
- 风险标记：缺少测试覆盖，仅覆盖 Quark（ModelOpt/fbgemm 待修复）

关联脉络

- PR #45863 [DSv4 Perf] DSv4 flashinfer sparse index cache for metadata, 2%~4% TTFT improvement: 同属 ROCm/AMD 生态的性能优化 PR，涉及 kernel 路径选择。
- PR #45782 [ROCm][Bugfix]: Fallback GFX942 sparse MLA ops to Triton: 同为 ROCm 上的 bugfix，涉及 kernel 回退路径。