PR #52987 完整报告

vllm-project/vllm

Revert "[Kernel] Gemma-4 FA4 FP8 Kernel"

合并时间：2026-08-20 01:48

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/52987

执行摘要

- 一句话：回退 Gemma-4 FA4 FP8 内核支持，恢复 FA3 默认路径
- 推荐动作：值得精读，尤其是 Gemma-4 模型支持、flash-attention 集成和推测解码方向的开发者。可以关注两点：一是回退选择了删除配置感知接口、改用 set_current_vllm_config 上下文对象，这种避免 API 扩散的做法可复用；二是 _copy_target_kv_scales 连同 FP8 KV scale 共享修复一起被移除，说明该问题在 MRV2 阶段被有意搁置。若团队计划重新引入 FA4，建议先建立覆盖 SM90 FP8 KV 与 MTP 的回归测试基线。

功能与动机

PR body 只有一句 “Reverts vllm-project/vllm#48666”，没有展开说明回退原因。被回退的 #48666 原本为 Gemma-4 在 SM90 上接入 FA4 FP8-KV-dequant 内核、支持 FP8 KV cache 并修复 MTP 场景下 draft 层 KV scale 默认 1.0 的问题。本次回退将这些改动整体撤销，公开信息中无法确认具体触发点，但从 MRV2 标签与直接合并的行为看，大概率是发布前稳定性收敛，或 FA4 FP8 路径在验证中暴露了回归。

实现拆解

1. 回退 FlashAttentionBackend 的 FA4 特判：在 vllm/v1/attention/backends/flash_attn.py 中删除 _get_sm90_fa4_fp8_kv_block_size、get_supported_kernel_block_sizes_for_config、get_preferred_block_size_for_config，get_supported_kernel_block_sizes() 恢复为固定返回 [MultipleOf(16)]；同时删除 __init__里对 supports_quant_query_input 的 FA4 限制，恢复 flash_attn_supports_quant_query_input() 的原始结果。
2. 删除 AttentionBackend 上的配置感知接口：vllm/v1/attention/backend.py 移除 get_supported_kernel_block_sizes_for_config 与 get_preferred_block_size_for_config 两个方法，vllm/model_executor/layers/attention/attention.py 的 _largest_kernel_block_within 相应去掉 vllm_config 参数，回退为调用无参的 get_supported_kernel_block_sizes()。
3. 平台层改用上下文方式取配置：vllm/platforms/interface.py 的 update_block_size_for_backend 不再传 vllm_config 给后端，而是用 set_current_vllm_config(vllm_config) 上下文包装后调用 get_preferred_block_size，在保持原接口不变的前提下仍能拿到当前配置，这是本次回退中唯一值得保留的适配方式。

4. 撤销 Gemma-4 MTP 的 KV scale 共享修复：vllm/v1/worker/gpu/spec_decode/gemma4/speculator.py 删除 _copy_target_kv_scales，_setup_gemma4_kv_sharing 从基于 target_names_by_index 的映射恢复为基于 layer index 构造 target 层名，不再拷贝 _k_scale/_v_scale 等缓冲。
5. 回退 flash-attn 接口与构建依赖：vllm/vllm_flash_attn/flash_attn_interface.py 删除 fa4_fp8_kv_dequant 分支，flash_attn_varlen_func 直接透传 q/k/v descale；cmake/external_projects/vllm_flash_attn.cmake 中的版本指针也一并回退。本次没有配套测试文件变更。

关键文件：

- vllm/v1/attention/backends/flash_attn.py（模块 注意力后端；类别 source；类型 core-logic；符号 FlashAttentionBackend.get_supported_kernel_block_sizes, FlashAttentionBackend.get_preferred_block_size, FlashAttentionBackend._get_sm90_fa4_fp8_kv_block_size）：核心回退点：删除 SM90 FA4 FP8-KV 的 block size 协商、量化 query 特判，恢复 FA3 默认行为，直接影响 Gemma-4 在 SM90 上的注意力路径。
- vllm/v1/worker/gpu/spec_decode/gemma4/speculator.py（模块 推测解码；类别 source；类型 core-logic；符号 Gemma4Speculator._setup_gemma4_kv_sharing, Gemma4Speculator._copy_target_kv_scales）：回退 Gemma-4 MTP 的 KV scale 共享修复，删除 _copy_target_kv_scales，影响 FP8 target 下 draft 模型读取共享 KV cache 的数值正确性。
- vllm/v1/attention/backend.py（模块 注意力后端；类别 source；类型 core-logic；符号 AttentionBackend.get_supported_kernel_block_sizes_for_config, AttentionBackend.get_preferred_block_size_for_config）：删除 FA4 引入的两个配置感知抽象方法，恢复无参接口，是理解本次回退 API 收敛方式的关键文件。
- vllm/model_executor/layers/attention/attention.py（模块 注意力层；类别 source；类型 data-contract；符号 _largest_kernel_block_within）：_largest_kernel_block_within 去掉 vllm_config 参数并回退到无参接口，是配置感知 API 回退的调用侧配套。
- vllm/vllm_flash_attn/flash_attn_interface.py（模块 FA 接口；类别 source；类型 core-logic；符号 flash_attn_varlen_func）：移除 FA4 fp8-kv-dequant 分支，flash_attn_varlen_func 直接透传 q/k/v descale，回退 flash-attn 集成层。
- vllm/platforms/interface.py（模块 平台层；类别 source；类型 dependency-wiring；符号 Platform.update_block_size_for_backend）：用 set_current_vllm_config 上下文替代显式传参，是回退后保持 block size 选择正确性的关键适配。

关键符号：FlashAttentionBackend.get_supported_kernel_block_sizes, FlashAttentionBackend.get_preferred_block_size, Gemma4Speculator._setup_gemma4_kv_sharing, _largest_kernel_block_within, Platform.update_block_size_for_backend

关键源码片段

vllm/v1/attention/backends/flash_attn.py

核心回退点：删除 SM90 FA4 FP8-KV 的 block size 协商、量化 query 特判，恢复 FA3 默认行为，直接影响 Gemma-4 在 SM90 上的注意力路径。

```
class FlashAttentionBackend(AttentionBackend):
    supported_dtypes: ClassVar[list[torch.dtype]] = [torch.float16, torch.bfloat16]
    supported_kv_cache_dtypes: ClassVar[list[CacheDType]] = [
        "auto",
        "float16",
        "bfloat16",
        "fp8",
        "fp8_e4m3",
    ]

    # 回退后不再按 SM90 FA4 FP8-KV 的特殊 page 契约上报 64 token 块，
    # 统一回到 FA3 的 multiple-of-16 能力声明，避免 sliding-window
    # cache spec 匹配到 64 token 的固定块大小。
    @staticmethod
    def get_supported_kernel_block_sizes() -> list[int | MultipleOf]:
        return [MultipleOf(16)]

    forward_includes_kv_cache_update: bool = False

    @classmethod
    def get_preferred_block_size(cls, default_block_size: int) -> int:
        # XPU 依然保持 64 的下界；SM90 FA4 的 64 token 特判被移除。
        if current_platform.is_xpu():
            return max(default_block_size, 64)
        return super().get_preferred_block_size(default_block_size)

    @staticmethod
    def get_name() -> str:
        return "FLASH_ATTN"
```

vllm/v1/attention/backend.py

删除 FA4 引入的两个配置感知抽象方法，恢复无参接口，是理解本次回退 API 收敛方式的关键文件。

```
class AttentionBackend(ABC):
    # 回退后不再提供按 VllmConfig 精确查询的
    # get_supported_kernel_block_sizes_for_config 接口；
    # 平台层改在 set_current_vllm_config 上下文中调本方法。
    @staticmethod
    def get_supported_kernel_block_sizes() -> list[int | MultipleOf]:
        return [MultipleOf(1)]

    @classmethod
    def get_preferred_block_size(cls, default_block_size: int) -> int:
        supported_sizes = cls.get_supported_kernel_block_sizes()
        if not supported_sizes:
```

```
            return default_block_size
        if cls.supports_block_size(default_block_size):
            return default_block_size
        # 取最小支持的块大小作为 fallback。
        return min(s.base if isinstance(s, MultipleOf) else s for s in supported_sizes)
```

评论区精华

PR 没有人工 review 评论，只有 claude[bot] 的自动提示，说明本仓库配置为人工 code review 模式。回退决定由作者 ywang96 直接作出并合并，公开讨论中没有出现针对 FA4 FP8 路径的争议、性能数据或替代方案。回退动机和是否出现具体回归均未在 PR 内说明。

- 回退原因未说明且无人工 review (question): 公开信息不足以确认回退动机，推测与 MRV2 发布期稳定性收敛有关；合并由作者直接执行。

风险与影响

- 风险：
  1. 性能回退：回退后 SM90 上启用 fp8/fp8_e4m3 KV cache 的 Gemma-4 full attention 层不再走 FA4 FP8-KV-dequant 内核，可能明显损失 KV cache 压缩带来的显存 / 带宽收益（vllm/v1/attention/backends/flash_attn.py）。
  2. MTP 数值风险回归：删除 _copy_target_kv_scales 后，Gemma-4 MTP 在 FP8 target 模型下 draft 层 K/V scale 恢复默认 1.0，_setup_gemma4_kv_sharing 也不再显式复制 scale，未来若重新启用 FA4 需要补回该修复（vllm/v1/worker/gpu/spec_decode/gemma4/speculator.py）。
  3. 接口兼容风险：AttentionBackend 上删除的两个 config-aware 方法可能仍被仓库其它调用点引用，需全局搜索确认没有遗漏；platforms/interface.py 已改用 set_current_vllm_config 适配，但其它 import 未排查。
  4. 构建依赖回退：cmake/external_projects/vllm_flash_attn.cmake 的版本指针回退后，已基于 FA4 构建的本地环境可能出现二进制与源码不一致。
  5. 缺少测试配套：本次回退没有新增或调整测试，相关回归风险无法通过 CI 增量验证。 - 影响：影响范围集中在 Nvidia SM90 平台、启用 FP8 KV cache 的 Gemma-4 推理，以及 Gemma-4 MTP 推测解码场景；默认配置（BF16 KV、非 Gemma-4 模型）用户基本无感。对团队而言，这意味着 MRV2 周期内 FA4 FP8 能力被暂时搁置，后续需评估是否以更成熟的方式重新引入，并在重新提交前补齐测试与回退原因记录。 - 风险标记：核心注意力路径变更，无测试配套，回退原因未公开说明，MTP KV scale 修复被移除，潜在性能回退

关联脉络

- PR #48666 [Kernel] Gemma-4 FA4 FP8 Kernel: 本 PR 是 #48666 的直接回退，改动内容全部来自该 PR 引入的 SM90 FA4 FP8 路径、MTP KV scale 共享与配置感知接口。