

PR #53161 完整报告

vllm-project/vllm

[ROCm][Perf][DeepSeek V4] Fuse native FP8 shared expert with MXFP4 routed experts

合并时间: 2026-09-07 10:42

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/53161>

执行摘要

- 一句话:
- 推荐动作:

功能与动机

实现拆解

1. 核心源码路径 - `vllm/models/deepseek_v4/amd/model.py` (data-contract) : 源码主路径 ; 涉及符号 `_use_heterogeneous_fmoe`, `_validate_heterogeneous_routes`, `_heterogeneous_shared_expert_enabled`, `_prepare_native_fp8_shared_expert`; 包含 导入关系调整、控制流调整、配置键调整; +377/-5; 关键符号: `_use_heterogeneous_fmoe`, `_validate_heterogeneous_routes`, `_heterogeneous_shared_expert_enabled`, `_prepare_native_fp8_shared_expert` - `vllm/_aiter_ops.py` (dependency-wiring) : 源码主路径; 涉及符号 `_validate_rocm_aiter_fused_moe_shared_expert_args`, `_probe_dsv4_i384_fmoe_capability`, `fused_moe_supports_heterogeneous_shared_expert`, `shuffle_scale`; 包含 导入关系调整、控制流调整、配置键调整; +107/-0; 关键符号: `_validate_rocm_aiter_fused_moe_shared_expert_args`, `_probe_dsv4_i384_fmoe_capability`, `fused_moe_supports_heterogeneous_shared_expert`, `shuffle_scale` - `vllm/model_executor/layers/fused_moe/experts/rocm_aiter_moe.py` (data-contract) : 源码主路径; 包含 控制流调整、配置键调整; +15/-0
2. 测试与验证 - `tests/model_executor/layers/test_fused_shared_expert.py` (test-coverage) : 测试配套; 涉及符号 `test_deepseek_v4_heterogeneous_fmoe_keeps_native_intermediate_width`, `test_deepseek_v4_heterogeneous_fmoe_token_policy`, `supports`, `_supported_fmoe_signature`; 包含 测试覆盖调整、导入关系调整、控制流调整; +340/-1; 关键符号: `test_deepseek_v4_heterogeneous_fmoe_keeps_native_intermediate_width`, `test_deepseek_v4_heterogeneous_fmoe_token_policy`, `supports`, `_supported_fmoe_signature`

关键文件:

- `vllm/models/deepseek_v4/amd/model.py` (模块 `vllm/models`; 类别 `source`; 类型 `data-contract`; 符号 `_use_heterogeneous_fmoe`, `_validate_heterogeneous_routes`, `_heterogeneous_shared_expert_enabled`, `_prepare_native_fp8_shared_expert`) : 源码

主路径；涉及符号 `_use_heterogeneous_fmoe`, `_validate_heterogeneous_routes`, `_heterogeneous_shared_expert_enabled`, `_prepare_native_fp8_shared_expert`；包含导入关系调整、控制流调整、配置键调整；+377/-5

- `vllm/_aiter_ops.py`（模块 `vllm`；类别 `source`；类型 `dependency-wiring`；符号 `_validate_rocm_aiter_fused_moe_shared_expert_args`, `_probe_dsv4_i384_fmoe_capability`, `fused_moe_supports_heterogeneous_shared_expert`, `shuffle_scale`）：源码主路径；涉及符号 `_validate_rocm_aiter_fused_moe_shared_expert_args`, `_probe_dsv4_i384_fmoe_capability`, `fused_moe_supports_heterogeneous_shared_expert`, `shuffle_scale`；包含 导入关系调整、控制流调整、配置键调整；+107/-0
- `tests/model_executor/layers/test_fused_shared_expert.py`（模块 `fused/shared/expert`；类别 `test`；类型 `test-coverage`；符号 `test_deepseek_v4_heterogeneous_fmoe_keeps_native_intermediate_width`, `test_deepseek_v4_heterogeneous_fmoe_token_policy`, `supports`, `_supported_fmoe_signature`）：测试配套；涉及符号 `test_deepseek_v4_heterogeneous_fmoe_keeps_native_intermediate_width`, `test_deepseek_v4_heterogeneous_fmoe_token_policy`, `supports`, `_supported_fmoe_signature`；包含 测试覆盖调整、导入关系调整、控制流调整；+340/-1
- `vllm/model_executor/layers/fused_moe/experts/rocm_aiter_moe.py`（模块 `vllm/model_executor`；类别 `source`；类型 `data-contract`）：源码主路径；包含 控制流调整、配置键调整；+15/-0

关键符号：`_use_heterogeneous_fmoe`, `_validate_heterogeneous_routes`, `_heterogeneous_shared_expert_enabled`, `_prepare_native_fp8_shared_expert`, `scale_bytes`, `DeepseekV4HeterogeneousMx4MoEMethod`, `process_weights_after_loading`, `DeepseekV4HeterogeneousSharedRoutedExperts`, `_validate_rocm_aiter_fused_moe_shared_expert_args`, `_probe_dsv4_i384_fmoe_capability`

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：
- 影响：
 - 风险标记：暂无

关联脉络

- 暂无明显关联 PR