

PR #45999 完整报告

vllm-project/vllm

Revert "[Kernel] Add PDL support for DeepGEMM kernel"

合并时间: 2026-06-18 13:59

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45999>

执行摘要

- 一句话: 回滚 DeepGEMM PDL 支持, 修复 AMD 构建
- 推荐动作: 该 PR 是纯回滚操作, 逻辑简单, 无需精读。但可作为跨平台兼容性问题的案例参考: 新特性引入时需在 AMD/Intel GPU 等非 NVIDIA 平台上验证。

功能与动机

PR body 明确指出 'This broke AMD build', 即 PR #42996 引入的改动导致 AMD 平台构建失败, 需要紧急回滚以恢复 AMD 平台的稳定性。

实现拆解

1. 移除 `_apply_pdl` 函数及相关调用: 在 `vllm/utils/deep_gemm.py` 中, 删除了整个 `_apply_pdl` 函数 (约 13 行), 该函数用于设置 DeepGEMM 模块的 PDL 开关。同时删除了 `_lazy_init` 中调用 `_apply_pdl` 的代码块 (`if current_platform.is_arch_support_pdl(): _apply_pdl(dg, True)`), 恢复为不主动启用 PDL 的行为。
2. 调整 Triton kernel 的 PDL 相关逻辑: 在 `vllm/models/deepseek_v4/common/ops/fused_inv_rope_fp8_quant.py` 中, 移除了 Triton kernel 中的 `USE_GDC` 编译时常量和 `launch_pdl` 参数, 以及 kernel 内的 `gdc_launch_dependents/gdc_wait` 调用。将 `pdl_kwargs` 的生成逻辑从基于架构判断改为默认空字典 (AMD/XPU 平台) 或传入 `launch_pdl=False` 的字典 (其他平台), 避免在 AMD 上触发不支持的功能。
3. 还原 CUDA kernel 的启动方式: 在 `csrc/libtorch_stable/quantization/w8a8/fp8/per_token_group_quant.cu` 中, 移除了 CUDA kernel 内部使用 `griddepcontrol.wait/launch_dependents` 的 asm 指令, 并将原本使用 `cudaLaunchKernelEx` 和 `cudaLaunchAttributeProgrammaticStreamSerialization` 的复杂启动方式, 回退为标准的 `<<<grid, block, 0, stream>>>` 语法, 消除对 CUDA 9.0+ PDL 特性的依赖。

关键文件:

- `vllm/utils/deep_gemm.py` (模块 工具层; 类别 source; 类型 core-logic; 符号 `_apply_pdl`): 核心变更文件, 删除了 `_apply_pdl` 函数及其在 `_lazy_init` 中的调用, 这是 PDL 支持的主要入口。
- `vllm/models/deepseek_v4/common/ops/fused_inv_rope_fp8_quant.py` (模块 模型层; 类别 source; 类型 core-logic): Triton kernel 中移除了 PDL/GDC 相关参数和调用, 涉及

对 DeepSeek V4 模型推理路径的修改。

- `csrc/libtorch_stable/quantization/w8a8/fp8/per_token_group_quant.cu` (模块 CUDA 内核; 类别 source; 类型 core-logic) : CUDA kernel 中移除了 `griddepcontrol` 指令, 并将启动方式从 `cudaLaunchKernelEx` 回退为标准 `<<<...>>>`, 消除对 CUDA 9.0+ 特性的依赖。

关键符号: `_apply_pdl`, `_lazy_init`, `_fused_inv_rope_fp8_quant_kernel_impl`,
`per_token_group_quant_8bit_packed`

关键源码片段

`vllm/utils/deep_gemm.py`

核心变更文件, 删除了 `_apply_pdl` 函数及其在 `_lazy_init` 中的调用, 这是 PDL 支持的主要入口。

```
# 以下代码已被删除 (PR #45999 回滚后不再存在)
# def _apply_pdl(mod, enable: bool = True) -> None:
#     """ 启用 / 禁用 DeepGEMM PDL (Programmatic Dependent Launch)。
#     此函数在 AMD 平台上调用会引发构建错误, 因此被完全移除。"""
#     mod_name = getattr(mod, "__name__", str(mod))
#     try:
#         set_pdl_fn = getattr(mod, "set_pdl", None)
#         if set_pdl_fn is None:
#             return
#         set_pdl_fn(enable)
#         logger.info_once("DeepGEMM PDL %s on %s.", "enabled" if enable else "disabled", mod_name)
#     except Exception as e:
#         logger.warning_once("Failed to set DeepGEMM PDL on %s: %s", mod_name, e)
```

`vllm/models/deepseek_v4/common/ops/fused_inv_rope_fp8_quant.py`

Triton kernel 中移除了 PDL/GDC 相关参数和调用, 涉及对 DeepSeek V4 模型推理路径的修改。

```
# 回滚后, pdl_kwargs 不再根据架构启用 PDL, 而是默认禁用
pdl_kwargs = (
    {}
    if current_platform.is_rocm() or current_platform.is_xpu()
    else {"launch_pdl": False}
)
# 注意: 移除了 USE_GDC 和 launch_pdl 参数, kernel 签名中不再包含它们
_fused_inv_rope_fp8_quant_per_head[grid](
    ...,
    ROPE_START=rope_start,
    HALF_ROPE=half_rope,
    TMA_ALIGNED_SCALES=tma_aligned_scales,
    # USE_GDC=use_gdc, # 已删除
    num_stages=1,
    **pdl_kwargs, # 现在为空字典或 {"launch_pdl": False}
```

```
    num_warps=1,  
)
```

csrc/libtorch_stable/quantization/w8a8/fp8/per_token_group_quant.cu

CUDA kernel 中移除了 `griddepcontrol` 指令，并将启动方式从 `cudaLaunchKernelEx` 回退为标准 `<<<...>>>`，消除对 CUDA 9.0+ 特性的依赖。

```
// 回滚后，不再使用 cudaLaunchKernelEx 和 programmatic stream serialization  
// 而是使用标准的 CUDA kernel launch 语法，避免了 PDL 特性导致的 AMD 构建失败  
#define LAUNCH_REG_KERNEL_INST(T, DST_DTYPE, KX, RY) \  
do { \  
    dim3 grid(static_cast<unsigned int>(blocks_x), static_cast<unsigned int>(blocks_y)); \  
    dim3 block(num_threads); \  
    per_token_group_quant_8bit_packed_register_kernel<T, DST_DTYPE, 128, KX, RY> \  
    <<<grid, block, 0, stream>>>( \  
        static_cast<const T*>(input.data_ptr()), output_q.data_ptr(), \  
        reinterpret_cast<unsigned int*>(output_s_packed.data_ptr()), \  
        static_cast<int>(padded_groups_per_row), \  
        static_cast<int>(groups_per_row), static_cast<int>(mn), \  
        static_cast<int>(output_q_mn_extent), \  
        static_cast<int>(tma_aligned_mn), num_scale_elems, \  
        static_cast<float>(eps), static_cast<float>(min_8bit), \  
        static_cast<float>(max_8bit)); \  
} while (0)
```

评论区精华

PR 无 review 评论，仅由 [DarkLight1337](#) 快速批准。决策依据直接来自 PR body 中的 'This broke AMD build'，属于紧急回滚，无技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：回滚后，DeepGEMM 在支持 PDL 的 NVIDIA SM90+ 架构上将失去 PDL 优化，可能轻微影响该平台上的性能。但原 PR #42996 的 PDL 支持本身是可选优化，回滚后恢复到未启用 PDL 的状态，不会引入正确性问题。CUDA kernel 的启动方式回退为标准语法，功能等价，无回归风险。AMD 构建恢复，是本次回滚的主要收益。
- 影响：用户影响：AMD 平台用户恢复可用性；NVIDIA SM90+ 用户将失去 DeepGEMM PDL 加速，但功能不受影响。系统影响：回滚是干净的反向操作，不影响其他模块。团队影响：需在后续重新评估 PDL 支持方案，确保兼容 AMD 后再合并。
- 风险标记：AMD 构建修复，NVIDIA SM90+ 性能退化

关联脉络

- PR #42996 [Kernel] Add PDL support for DeepGEMM kernel: 被回滚的原 PR，引入 DeepGEMM PDL 支持，但破坏了 AMD 构建。

- PR #45823 [Fix][KV offload] Defer on_request_finished until in-flight transfers drain:
同属近期对稳定性的修复，且与 DeepSeek V4 及 KV Connector 相关。