

PR #45781 完整报告

vllm-project/vllm

[Misc] Rename VLLM_TRITON_ATTN_USE_TD to VLLM_TRITON_USE_TD

合并时间: 2026-07-14 18:32

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/45781>

执行摘要

- 一句话: 重命名环境变量 VLLM_TRITON_ATTN_USE_TD 为 VLLM_TRITON_USE_TD
- 推荐动作: 建议尽快合并, 因为多个下游 PR 依赖此变量注册。值得关注的是弃用策略的设计, 参考了此前 HOST_IP 的处理方式。

功能与动机

RFC #42545 提议将 Tensor Descriptor 功能从注意力模块扩展至整个 vLLM 的 Triton 内核, 需要一个统一的开关变量。此 PR 是后续 PR (如 #42436、#47152 等) 的基础, 确保新变量在 envs.py 中注册, 后续 PR 无需再为每个内核引入独立变量。

实现拆解

1. 更新环境变量定义: 在 vllm/envs.py 中, 将 VLLM_TRITON_ATTN_USE_TD 替换为 VLLM_TRITON_USE_TD, 并保留旧变量作为弃用别名。
2. 添加弃用警告函数: 新增 _deprecated_triton_attn_use_td() 函数, 在旧变量被使用时打印警告并返回 None。
3. 更新引用处: 在 triton_attn.py 和 triton_unified_attention.py 中, 将 VLLM_TRITON_ATTN_USE_TD 引用替换为新变量名。
4. 注册旧变量: 在环境变量解析字典中将旧变量映射到弃用函数, 确保不会触发未知变量检查。

关键文件:

- vllm/envs.py (模块 环境配置; 类别 source; 类型 core-logic; 符号 _deprecated_triton_attn_use_td): 核心变更文件, 完成环境变量重命名和弃用机制。
- vllm/v1/attention/backends/triton_attn.py (模块 注意力后端; 类别 source; 类型 core-logic): 更新了注意力后端中对环境变量的引用。
- vllm/v1/attention/ops/triton_unified_attention.py (模块 注意力算子; 类别 infra; 类型 infrastructure): 更新了错误信息中的变量名引用。

关键符号: _deprecated_triton_attn_use_td

关键源码片段

vllm/envs.py

核心变更文件，完成环境变量重命名和弃用机制。

```
# vllm/envs.py

# 新增通用变量 VLLM_TRITON_USE_TD
VLLM_TRITON_USE_TD: bool | None = None
# 保留旧变量作为弃用别名 (v0.25 移除)
VLLM_TRITON_ATTN_USE_TD: bool | None = None

# 弃用警告函数
def _deprecated_triton_attn_use_td() -> None:
    """
    当用户仍使用 VLLM_TRITON_ATTN_USE_TD 时发出警告。
    该变量将被忽略，新变量 VLLM_TRITON_USE_TD 生效。
    """
    if "VLLM_TRITON_ATTN_USE_TD" in os.environ:
        logger.warning(
            "VLLM_TRITON_ATTN_USE_TD is deprecated and will be removed in "
            "v0.25. Use VLLM_TRITON_USE_TD instead."
        )
    return None

# 在环境变量解析字典中注册新变量和旧变量的弃用处理
os.environ["VLLM_TRITON_USE_TD"] = lambda: {"1": True, "0": False}.get(
    os.getenv("VLLM_TRITON_USE_TD", "").strip()
),
# 旧变量注册为弃用函数，避免触发未知变量检查
"VLLM_TRITON_ATTN_USE_TD": lambda: _deprecated_triton_attn_use_td(),
```

vllm/v1/attention/backends/triton_attn.py

更新了注意力后端中对环境变量的引用。

```
# vllm/v1/attention/backends/triton_attn.py

# 在 __init__ 方法中，使用新变量名获取 Tensor Descriptor 开关
# VLLM_TRITON_USE_TD 是三态变量：
# - unset (None): auto-select (TD on for XPU, off elsewhere)
# - ``1``: force TD on
# - ``0``: force TD off
td_override = envs.VLLM_TRITON_USE_TD
if td_override is None:
    self.use_td = current_platform.is_xpu()
else:
    self.use_td = td_override
```

vllm/v1/attention/ops/triton_unified_attention.py

更新了错误信息中的变量名引用。

```
# vllm/v1/attention/ops/triton_unified_attention.py
```

```
# 错误信息中提示用户设置新变量名
raise AssertionError(
    f"USE_TD_QO requires contiguous query heads "
    f"(q.stride(1) = {q.stride(1)} != head_size = {head_size}); "
    f"set VLLM_TRITON_USE_TD=0 or pad the query layout."
)
```

评论区精华

无显式 review 评论，但审阅者 jikunshang 已批准。Mergify 提示过合并冲突，后已解决。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更仅为重命名，旧变量仍受支持（带警告）。未影响任何运行时逻辑，三态语义完全保留。但用户若依赖旧变量且未切换，将收到警告，需在 v0.25 前更新。
- 影响：对用户：使用者若设置了 VLLM_TRITON_ATTN_USE_TD，将看到弃用警告，功能不受影响。对系统：无运行时性能或行为变化。对团队：为后续统一 TD 开关铺平道路。
- 风险标记：弃用兼容性

关联脉络

- PR #40327 Tensor Descriptor pilot in attention: 此 PR 是 Tensor Descriptor 试点，引入了最初的 VLLM_TRITON_ATTN_USE_TD。
- PR #42545 TD-adoption RFC: RFC 提议将 TD 扩展到整个 vLLM，触发此重命名。
- PR #42436 Use VLLM_TRITON_USE_TD in fused MoE kernel: 依赖此 PR 的变量注册，避免自注册新变量。
- PR #47152 Use VLLM_TRITON_USE_TD in W8A8 block-INT8 matmul: 依赖此 PR 的变量注册。
- PR #47084 Use VLLM_TRITON_USE_TD in Mamba2 bmm_chunk_fwd: 依赖此 PR 的变量注册。
- PR #45816 Use VLLM_TRITON_USE_TD in Mamba chunk-cumsum store: 依赖此 PR 的变量注册。