PR #49964 完整报告

vllm-project/vllm

[Bugfix][KV Offload] Keep Mamba block span unscaled under DCP

合并时间：2026-07-28 22:41

原文链接：http://prhub.com.cn/vllm-project/vllm/pull/49964

执行摘要

- 一句话：修复 Mamba 块 span 在 DCP 下被错误缩放
- 推荐动作：建议精读。此 PR 虽小，但修复了一个关键的缓存一致性问题，设计决策（仅在 AttentionSpec 时缩放）值得借鉴。测试覆盖完整，包括 alignment 和 store reachable 验证。

功能与动机

Native KV offloading currently multiplies every cache group's token span by the DCP world size, which is correct for attention KV (sharded across DCP ranks) but incorrect for Mamba state (replicated). This causes offload keys to associate with wrong Mamba state blocks, leading to restoration of recurrent state from an earlier boundary upon cache hit. The PR author states: 'With 16 token attention and Mamba blocks under DCP=2, the current configuration derives 32 tokens per block for both groups. The correct spans are 32 tokens for attention and 16 tokens for Mamba.'

实现拆解

1. 修改核心缩放逻辑(vllm/distributed/kv_transfer/kv_connector/v1/offloading/config.py): 在 build_offloading_config 函数中，将 tokens_per_block 的缩放条件从无条件乘以 decode_context_parallel_size 改为仅当 group.kv_cache_spec 是 AttentionSpec 实例时才缩放，否则缩放系数为 1。
2. 新增 Mamba 混合缓存配置辅助函数(tests/v1/kv_connector/unit/offloading_connector/test_config.py): 添加 _make_mamba_hybrid_kv_cache_config 函数，创建包含 FullAttentionSpec 和 MambaSpec 的 KVCacheConfig，用于测试混合场景。
3. 新增测试用例(tests/v1/kv_connector/unit/offloading_connector/test_config.py): 添加 test_dcp_scales_attention_but_not_mamba_group_blocks 测试，验证在 DCP=2 时 attention 组 token_per_block 为 32，Mamba 组为 16，并进一步测试 SchedulerOffloadConfig 的 alignment 和 store reachable 逻辑。
4. 导入调整: 在测试文件中新增对 MockOffloadingSpec、SchedulerOffloadConfig、is_store_reachable_swa_chunk 的导入；在源码中新增对 AttentionSpec 的导入。

关键文件：

- vllm/distributed/kv_transfer/kv_connector/v1/offloading/config.py（模块 卸载配置；类别 source；类型 core-logic；符号 build_offloading_config）：核心源码文件，修改了 build_offloading_config 函数中的 tokens_per_block 缩放逻辑，是本次 bug fix 的核心变更。
- tests/v1/kv_connector/unit/offloading_connector/test_config.py（模块 测试；类别 test；类型 test-coverage；符号 _make_mamba_hybrid_kv_cache_config, test_dcp_scales_attention_but_not_mamba_group_blocks）：测试文件，新增了混合缓存配置辅助函数和核心测试用例，确保修复正确且不会回归。

关键符号：build_offloading_config, _make_mamba_hybrid_kv_cache_config, test_dcp_scales_attention_but_not_mamba_group_blocks

关键源码片段

vllm/distributed/kv_transfer/kv_connector/v1/offloading/config.py

核心源码文件，修改了 build_offloading_config 函数中的 tokens_per_block 缩放逻辑，是本次 bug fix 的核心变更。

```
# vllm/distributed/kv_transfer/kv_connector/v1/offloading/config.py

from vllm.v1.kv_cache_interface import (
    AttentionSpec,
    FullAttentionSpec,
    MLAAttentionSpec,
)

def build_offloading_config(
    vllm_config: "VllmConfig",
    kv_cache_config: "KVCacheConfig",
) -> OffloadingConfig:
    """Translate vLLM configuration into the native offloading boundary."""
    # ... ( 省略前置代码 ) ...
    parallel_config = vllm_config.parallel_config

    # 关键修复：仅对 AttentionSpec 子类应用 DCP 缩放，
    # Mamba 等非注意力组保持原始 block_size
    groups = tuple(
        OffloadingGroupConfig(
            tokens_per_block=(
                group.kv_cache_spec.block_size
                * (
                    parallel_config.decode_context_parallel_size
                    if isinstance(group.kv_cache_spec, AttentionSpec)
                    else 1
                )
            ),
            layer_names=tuple(group.layer_names),
        )
```

```
        for group in kv_cache_config.kv_cache_groups
    )
    # ... ( 后续代码 ) ...
```

tests/v1/kv_connector/unit/offloading_connector/test_config.py

测试文件，新增了混合缓存配置辅助函数和核心测试用例，确保修复正确且不会回归。

```
# tests/v1/kv_connector/unit/offloading_connector/test_config.py

# 辅助函数：创建一个包含 FullAttention 和 Mamba 的混合 KVCacheConfig
def _make_mamba_hybrid_kv_cache_config() -> KVCacheConfig:
    return KVCacheConfig(
        num_blocks=4,
        kv_cache_tensors=[],
        kv_cache_groups=[
            # Attention 组：使用默认 block_size (16)
            KVCacheGroupSpec(["full_layer"], _full_attention_spec()),
            # Mamba 组：指定 block_size=16, mamba_cache_mode="align"
            KVCacheGroupSpec(
                ["mamba_layer"],
                MambaSpec(
                    block_size=16,
                    shapes=((1, 1),),
                    dtypes=(torch.float32,),
                    mamba_cache_mode="align",
                ),
            ),
        ],
    )

# 核心测试：验证 DCP 只缩放 Attention 组
def test_dcp_scales_attention_but_not_mamba_group_blocks():
    # DCP=2, TP=2
    config = _make_vllm_config(
        tensor_parallel_size=2, decode_context_parallel_size=2
    )
    config.speculative_config = None

    offloading_config = build_offloading_config(
        config, _make_mamba_hybrid_kv_cache_config()
    )

    # Attention 组：block_size * DCP = 16 * 2 = 32
    # Mamba 组：block_size * 1 = 16
    assert tuple(
        group.tokens_per_block for group in offloading_config.groups
    ) == (32, 16)

    # 进一步验证 scheduler 配置中的 alignment 和 store reachable 逻辑
```

```
    scheduler_config = SchedulerOffloadConfig.from_spec(
        MockOffloadingSpec(offloading_config),
        config,
        _make_mamba_hybrid_kv_cache_config(),
    )
    mamba_group = scheduler_config.kv_group_configs[1]
    assert mamba_group.alignment_chunk_count == 2
    assert [
        chunk_idx
        for chunk_idx in range(4)
        if is_store_reachable_swa_chunk(
            chunk_idx,
            4,
            mamba_group.alignment_chunk_count,
            mamba_group.sliding_window_size_in_chunks,
            mamba_group.is_eagle_group,
        )
    ] == [1, 3]
```

评论区精华

无 review 评论。只有 claude[bot] 自动评论（来自 fork 无法自动审查）和 maintainer orozery 的批准。

- 暂无高价值评论线程

风险与影响

- 风险：低风险。变更范围小，仅修改了 config.py 中的 12 行，并添加了 56 行测试。核心逻辑改动是增加了一个 isinstance 判断，条件明确（仅对 AttentionSpec 缩放），不会影响非混合模型。但需注意：如果未来新增其他需要 DCP 缩放的 Spec 类型，需要同步更新此条件。
- 影响：影响范围局限于使用原生 KV 卸载（OffloadingConnector）的混合模型（如 Mamba+Attention）在 DCP 场景下的缓存正确性。修复后，Mamba 状态的卸载和命中恢复行为将正确，避免缓存污染和错误的状态恢复。
- 风险标记：核心路径变更，条件分支新增

关联脉络

- PR #50297 [BugFix] Fix P/D preemption race condition: 同一仓库近期修改了 KV 连接器相关代码，涉及调度器和 offloading 连接器，共享相关上下文。
- PR #50326 [PD][Bugfix] Rebase KV lease deadlines onto worker clock: 同样是 KV 连接器 bugfix，涉及跨节点 KV 租约和时钟同步，与本 PR 同属 KV offload 领域。