

PR #48361 完整报告

vllm-project/vllm

[Test] Regression test for hybrid-Mamba eagle cache-peek in Mooncake connector (#43559)

合并时间: 2026-07-28 03:03

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/48361>

执行摘要

- 一句话: 为 hybrid-Mamba EAGLE 缓存查找添加回归测试
- 推荐动作: 建议所有涉及 KV 连接器和缓存查找逻辑的开发者仔细阅读此测试用例, 理解其针对的 bug 场景和测试设计思路。后续修改 `find_longest_cache_hit`、EAGLE 弹出逻辑或 Mamba 相关行为时, 确保此测试通过。此 PR 体现了测试驱动修复的典型实践, 值得作为回归测试的范例。

功能与动机

原始问题 #43559 中, `MambaManager` 未执行 `drop_eagle_block` (Mamba 块在位置 `p` 本身就是 $(p + 1) * \text{block_size}$ 个 token 后的循环状态, 无需重算), 导致 EAGLE 缓存查找逻辑为 Mamba 组授予了与 FullAttention 相同的 peek margin。这一 margin 使 Mamba 组匹配到比 attention-verified 结果更长的前缀, 从而在恢复循环状态时跳过已验证的 token 前缀, 引发静默输出损坏。生产修复 #49502 已通过 `not isinstance(spec, MambaSpec)` 条件关闭了 Mamba 组的额外 margin, 本 PR 添加回归测试以锁定该行为, 防止未来重构或疏忽导致回归。

实现拆解

1. 添加辅助函数 `_mamba` (`test_mooncake_store_coordinator.py`): 封装 `MambaSpec` 的创建, 便于测试复用。
2. 新增测试用例 `test_lookup_with_eagle_hybrid_full_plus_mamba_no_ouerrun`: 配置一组混合 KV 缓存组 (FullAttention + Mamba), 在 EAGLE 模式下调用 `find_longest_cache_hit`, 并断言返回的 hit 长度为 48 (FullAttention 的 4 块被 eagle 弹出 1 块后得到 48 个 token)。由于 Mamba 组没有 peek margin, 它无法匹配到第 5 块, 从而验证了门控正确性。
3. 在源文件中添加注释 (`coordinator.py`): 在 `if drop_eagle_block and not isinstance(spec, MambaSpec)` 条件上方插入解释性注释, 说明循环状态组为什么不能获得额外 margin, 为后续维护者提供设计原由。
4. 测试负向控制: 作者手动移除了门控条件后该测试失败 (`assert 64 == 48`), 证明测试能够有效检测回归。

关键文件:

- `tests/v1/kv_connector/unit/test_mooncake_store_coordinator.py` (模块 KV 连接器; 类别 test; 类型 test-coverage; 符号 `_mamba`, `test_lookup_with_eagle_hybrid_full_plus_`

mamba_no_ouerrun)：核心回归测试文件，新增了测试函数

test_lookup_with_eagle_hybrid_full_plus_mamba_no_ouerrun 和辅助函数 _mamba，用于验证 Mamba 组在 EAGLE 模式下不会因 peek margin 而越界。这是本 PR 的主要产出。

- vllm/distributed/kv_transfer/kv_connector/v1/mooncake/store/coordinator.py（模块 KV 连接器；类别 source；类型 documentation）：在生产代码中添加关键注释，解释 Mamba 组被排除在 eagle peek margin 之外的原因，为后续维护者提供设计上下文。该注释直接对应 #49502 的门控逻辑。

关键符号：_mamba, test_lookup_with_eagle_hybrid_full_plus_mamba_no_ouerrun

关键源码片段

tests/v1/kv_connector/unit/test_mooncake_store_coordinator.py

核心回归测试文件，新增了测试函数 test_lookup_with_eagle_hybrid_full_plus_mamba_no_ouerrun 和辅助函数 _mamba，用于验证 Mamba 组在 EAGLE 模式下不会因 peek margin 而越界。这是本 PR 的主要产出。

辅助函数：创建 MambaSpec，方便测试用例复用

```
def _mamba(block_size=16):
    return MambaSpec(
        block_size=block_size,
        shapes=((1, 1),),
        dtypes=(torch.float32,),
        mamba_cache_mode="align",
    )
```

```
def test_lookup_with_eagle_hybrid_full_plus_mamba_no_ouerrun():
```

```
    """Full+Mamba with eagle must not overrun the attention-verified hit.
```

```
    ``MambaManager`` ignores ``drop_eagle_block`` (a Mamba block at position
    p IS the recurrent state after (p + 1) * block_size tokens; there is
    nothing to recompute), so granting the Mamba group the one-block eagle
    peek margin lets it match one block PAST the eagle-pruned full-attention
    hit, and adopting that length resumes the recurrent state ahead of the
    verified token prefix (#43559). Gating the margin on ``not
    isinstance(spec, MambaSpec)`` pins the hit to the attention-verified 48.
    """
```

```
    groups = [
        KVCacheGroupSpec(["L0"], _full(16)),
        KVCacheGroupSpec(["L1"], _mamba(16)),
    ]
```

```
    coord = _make_coord(groups, hash_block_size=16, use_eagle=True)
```

```
    hs = _hashes(4)
```

```
    exists = {(g, bytes(h)) for g in (0, 1) for h in hs}
```

```
    cmap = ExternalCachedBlockPool(16, exists)
```

```
    _masks, hit = coord.find_longest_cache_hit(
```

```
        hs, max_length=64, cached_block_pool=cmap
```

```
)  
# FullAttn 匹配 4 个块, eagle 弹出 1 个 -> 48 个已校验 token;  
# Mamba 组必须提供 state@48 的快照, 而不是偷看到 state@64  
assert hit == 48
```

评论区精华

核心讨论: PR 范围简化与生产修复关系

- njhill指出 #46384 和 #47782 已合并, 部分修复逻辑已被覆盖, 建议简化 PR。作者随后 rebase 并移除了重复的修复代码, 仅保留测试和注释。
- njhill最终批准, 并要求获得 ivanium的确认。ivanium随后也批准, 认为 LGTM。
- 讨论中确认了测试的正确性和必要性, 并解决了合并冲突。
- PR 范围简化与生产修复合并 (design): PR 最终仅包含测试和注释, 生产修复部分已由 #49502 独立合并。ivanium 和 njhill 均批准了最终版本。

风险与影响

- 风险: 低风险。
- 变更仅包含新增测试代码和一条注释, 不涉及现有逻辑修改。
- 测试依赖内部辅助函数 (_make_coord、_hashes、ExternalCachedBlockPool), 若这些接口在未来重构中改变, 测试需要同步更新。但这是测试本身应有的维护成本。
- 负向控制已验证测试能捕获门控缺失, 因此回归检测能力可靠。
- 影响: 影响范围: 低。
- 用户: 无直接影响, 不改变运行时行为。
- 测试套件: 新增一个回归测试用例, 扩展了 Mooncake 连接器的测试覆盖, 增强了对 hybrid-Mamba 场景的保障。
- 开发者: 为后续修改 find_longest_cache_hit 或 EAGLE 缓存查找逻辑提供安全网, 降低引入相同 bug 的风险。
- 风险标记: 仅测试变更, 低回归风险, 依赖内部测试工具函数

关联脉络

- PR #43559 Bug: hybrid-Mamba eagle cache-peek overrun: 原始问题报告, 本测试旨在防止该问题回归
- PR #49502 [Bugfix] Gate EAGLE cache-peek for Mamba in Mooncake connector lookup: 已合并的生产修复, 本测试为其提供回归覆盖