

PR #47754 完整报告

vllm-project/vllm

[Test] Enable KV cache events for HMA models in CPU offloading test

合并时间: 2026-07-14 17:22

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/47754>

执行摘要

- 一句话: 移除 CPU 卸载测试中 HMA 模型的 KV 事件限制
- 推荐动作: 此 PR 值得一看, 因为它演示了如何随着底层依赖的演进 (如 #37688) 清理过时的测试防护。同时, 其变更模式对理解测试条件的演进有参考价值。

功能与动机

KV 缓存事件不再强制关闭混合 KV 缓存管理器 (该限制已在 #37688 中移除, 连接器侧组支持在 #43468 中添加), 因此删除 `test_cpu_offloading` 中过时的 `uses_hma` 防护, 并同时为 HMA 模型启用 KV 事件断言。

实现拆解

1. 移除条件分支: 在 `tests/v1/kv_connector/unit/test_offloading_connector.py` 中, 删除围绕 `KVEventsConfig` 构建的 `if not uses_hma:` 条件, 使 KV 事件配置不再依赖于 `uses_hma` 参数。
2. 简化变量定义: 移除了 `kv_events_config` 的 `None` 初始化和条件赋值, 将其提升为无条件创建, 并移除 `subscriber` 的 `None` 初始化 (但保留定义)。
3. 调整断言逻辑: 将 `if kv_events_config is not None:` 防护改为直接执行 `events_endpoint.replace` 和 `MockSubscriber` 构造, 因为 `kv_events_config` 现在始终非空。
4. 保留其他逻辑: 对 `force_prefix_caching` 等 HMA 特定逻辑做了保留, 以确保测试对 HMA 的后续行为仍然正确。

关键文件:

- `tests/v1/kv_connector/unit/test_offloading_connector.py` (模块 KV 连接器; 类别 test; 类型 test-coverage): 唯一变更文件, 移除 HMA 模型的 KV 事件条件限制, 简化了测试逻辑并扩大了测试覆盖。

关键符号: `test_cpu_offloading`

关键源码片段

`tests/v1/kv_connector/unit/test_offloading_connector.py`

唯一变更文件, 移除 HMA 模型的 KV 事件条件限制, 简化了测试逻辑并扩大了测试覆盖。

```
# tests/v1/kv_connector/unit/test_offloading_connector.py
```

```
@pytest.mark.parametrize("model, attn_backend, cpu_block_size, uses_hma", MODEL_PARAMS)
```

```
def test_cpu_offloading(
```

```
    model: str,
```

```
    attn_backend: str | None,
```

```
    cpu_block_size: int | None,
```

```
    uses_hma: bool,
```

```
) -> None:
```

```
    """Tests OffloadingConnector with CPUOffloadingSpec."""
```

```
    # 配置 OffloadingConnector (默认使用 CPUOffloadingSpec)
```

```
    extra_config: dict = {"cpu_bytes_to_use": 500 << 20}
```

```
    if cpu_block_size is not None:
```

```
        extra_config["block_size"] = cpu_block_size
```

```
    kv_transfer_config = KVTransferConfig(
```

```
        kv_connector="OffloadingConnector",
```

```
        kv_role="kv_both",
```

```
        kv_connector_extra_config=extra_config,
```

```
)
```

```
    # 移除 old guard: KV 事件现在兼容 HMA 模型, 因此始终创建 KVEventsConfig
```

```
    subscriber: MockSubscriber | None = None
```

```
    port: int
```

```
    with socket.socket(socket.AF_INET, socket.SOCK_STREAM) as s:
```

```
        s.bind(("0.0.0.0", 0))
```

```
        port = s.getsockname()[1]
```

```
    events_endpoint = f"tcp://*: {port}"
```

```
    kv_events_config = KVEventsConfig(
```

```
        enable_kv_cache_events=True,
```

```
        publisher="zmq",
```

```
        endpoint=events_endpoint,
```

```
        topic="test",
```

```
)
```

```
    # HMA 模型默认禁用 prefix caching; 强制启用以确保 offloading 正常工作
```

```
    force_prefix_caching = uses_hma
```

```
    llm = LLM(
```

```
        model=model,
```

```
        max_model_len=4096,
```

```
        gpu_memory_utilization=0.5,
```

```
        kv_events_config=kv_events_config,
```

```
        kv_transfer_config=kv_transfer_config,
```

```
        **({"attention_config": {"backend": attn_backend}} if attn_backend else {}),
```

```
        **({"disable_hybrid_kv_cache_manager": False} if uses_hma else {}),
```

```
        **({"enable_prefix_caching": True} if force_prefix_caching else {}),
```

```
        **({"max_num_seqs": 1} if current_platform.is_rocm() else {}),
```

```
)
```

```
# 始终创建 subscriber (因为 kv_events_config 现在一定非空)
events_endpoint = events_endpoint.replace("*", "127.0.0.1")
subscriber = MockSubscriber(events_endpoint, topic=kv_events_config.topic)

try:
    _latency_test(llm, subscriber)
    _accuracy_test(llm, subscriber)
finally:
    if subscriber is not None:
        subscriber.close()
    del llm
```

评论区精华

该 PR 没有引发实质性讨论。CI 预提交检查失败（由 mergify bot 提醒），但已通过合并分支解决。审阅者 orozery 直接批准了该 PR。

- 暂无高价值评论线程

风险与影响

- 风险：该 PR 仅修改测试文件，无生产代码变更，风险很低。主要风险是 HMA 模型可能隐含对 kv_events_config 的不兼容预期，导致测试失败，但该关联已在 #37688 和 #43468 中确认解决。
- 影响：直接影响仅限于 test_cpu_offloading 测试，现在 HMA 模型（如 Mamba）也将被测试 KV 缓存事件的正确发布。这提高了测试覆盖率，但不影响生产运行时的行为。
- 风险标记：仅测试变更，低风险

关联脉络

- PR #37688 Remove KV events forcing HMA off: 此 PR 所依赖的上游更改，解除了 KV 事件对 HMA 的限制。
- PR #43468 Add connector-side group support for KV events: 提供了连接器侧对 KV 事件组的支持，使 HMA 模型可以使用 KV 事件。