PR #7046 完整报告

PaddlePaddle/FastDeploy

[BugFix] Add lock to avoid generating nan when using storage cache

合并时间：2026-03-30 14:50

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7046

执行摘要

- 一句话：为 KVCache storage cache 读写任务加锁，防止并发导致 NaN 生成。
- 推荐动作：该 PR 值得精读，特别是锁机制的设计和异常处理。建议关注锁的粒度选择、assert 的替代方案（如显式异常），以及同步模式限制对现有代码的影响。

功能与动机

根据 PR body 中的表述，动机是“Add lock to avoid generating nan when using storage cache”，即在特定并发场景下防止 NaN 生成，确保缓存操作的可靠性。

实现拆解

实现包括三个主要部分：1) 在 prefix_cache_manager.py 的 issue_write_back_storage_task 方法中，添加 _acquire_kvcache_lock() 和 _release_kvcache_lock() 调用，并用 try/finally 包裹核心逻辑，同时加入 assert is_sync, "Only support is_sync=True for now."；2) 在 issue_prefetch_storage_task 方法中采用相同的锁机制和 assert；3) 在 mooncake_store.py 的 warmup 方法中，将 warmup 值从 1 MB 改为 4 KB，优化存储初始化。

关键文件：

- fastdeploy/cache_manager/prefix_cache_manager.py（模块 cache_manager）：核心修改文件，添加锁机制以避免并发导致的 NaN 生成，直接影响缓存管理的正确性。
- fastdeploy/cache_manager/transfer_factory/mooncake_store/mooncake_store.py（模块 storage）：次要修改文件，调整 warmup 值优化存储初始化，但对核心功能影响较小。

关键符号：issue_write_back_storage_task, issue_prefetch_storage_task, warmup

评论区精华

Review 中 Copilot 指出了两个关键问题：一是锁可能因异常未释放，导致 worker 永久阻塞，建议使用 try/finally 确保释放；二是使用 assert 限制 is_sync=True 可能在 Python -O 下失效，并影响现有测试。最终代码采纳了 try/finally 建议，但未修改 assert 部分，这可能导致约束不稳健。

- 锁释放保证 (correctness): 最终代码使用 try/finally 包裹锁操作，确保 _release_kvcache_lock() 一定执行，解决了释放问题。

- assert 使用问题 (correctness): 未修改 assert 部分，可能仍需处理以提升稳健性，建议改为显式异常或更新测试。

风险与影响

- 风险：技术风险包括：1) 锁未正确释放的风险，尽管已添加 try/finally，但在复杂异常路径下仍需验证；2) assert 使用可能导致在优化模式下约束失效，影响系统稳定性和测试用例；3) 修改 warmup 值可能对存储性能有轻微影响，但变化较小。此外，Codecov 报告显示有 7 行代码缺失覆盖，可能增加回归风险。
- 影响：影响范围限于使用 storage cache 的并发读写操作，可能提高数据一致性，但引入锁可能略微增加延迟。对用户透明，解决潜在的 NaN 问题，提升系统可靠性。团队需关注 assert 的潜在失效风险和测试更新需求。
- 风险标记：锁未释放风险，assert 失效风险，测试覆盖不足

关联脉络

- PR #6929 [BugFix][KVCache] Fix mm hash boundary comparison in get_block_hash_extra_keys: 修改了同一个文件 prefix_cache_manager.py，同为 KVCache 相关的 bugfix，揭示该模块在持续优化中。