PR #7241 完整报告

PaddlePaddle/FastDeploy

[Optimization] 移除 num_blocks 上限限制

合并时间：2026-04-13 22:07

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7241

执行摘要

- 一句话：移除 KV Cache 块数上限限制，提升高显存设备的显存利用率。
- 推荐动作：建议精读以理解 KV Cache 分配机制和移除限制的权衡。关注 iluvatar_worker.py 的未同步修改，以及测试 baseline 的普适性问题，可作为学习风险管理的案例。

功能与动机

PR body 中说明：" 之前为规避 ' 块数过多导致非法内存访问 ' 问题，在 worker_process.py 中对 num_blocks_local 硬编码了 40000 的上限。随着问题根因得到修复，该限制已无必要，且会错误地压低实际可用 KV Cache 块数，影响显存利用率。"

实现拆解

在 fastdeploy/worker/worker_process.py 和 fastdeploy/worker/iluvatar_worker.py 中注释掉 num_blocks_local > 40000 的检查和截断逻辑，移除上限限制。同时，更新两个 E2E 测试文件的 baseline 值：tests/e2e/4cards_cases/test_Qwen3_30b_tp4.py 从 40000 改为 74000，tests/e2e/test_EB_VL_Lite_serving.py 从 40000 改为 65400，以匹配移除上限后的实际块数。

关键文件：

- fastdeploy/worker/worker_process.py（模块 Worker）：包含核心 KV Cache 块数计算逻辑，移除上限直接影响显存分配
- fastdeploy/worker/iluvatar_worker.py（模块 Worker）：平台特定 worker 文件，未同步移除限制导致不一致
- tests/e2e/4cards_cases/test_Qwen3_30b_tp4.py（模块 Testing）：测试文件，更新 baseline 以匹配新块数
- tests/e2e/test_EB_VL_Lite_serving.py（模块 Testing）：测试文件，更新 baseline 以匹配新块数

关键符号：PaddleDisWorkerProc.initialize_kv_cache

评论区精华

Review 中，Copilot 和 fastdeploy-bot 指出移除限制可能导致非法内存访问风险，原注释明确警告 'Too many block will lead to illegal memory access'。建议删除注释代码而非保留，并同步修改 iluvatar_worker.py 以保持平台一致性。测试 baseline 值的确定缺乏说明。最终 PR 被 gongshaotian 批准，但风险未完全解决，平台不一致问题被标记为未解决。

- 非法内存访问风险 (correctness): 风险被提及但 PR 仍被批准，未添加替代保护
- 平台一致性 (design): 未在 PR 中解决，建议未来处理
- 测试 baseline 说明 (testing): baseline 已更新但未添加说明

风险与影响

- 风险：技术风险包括：1) 移除上限可能触发 GPU 非法内存访问错误（如 CUDA error 700），尤其是在极端显存配置下；2) iluvatar_worker.py 未同步修改，导致天数智芯平台仍受限，行为不一致；3) 测试 baseline 值为硬编码魔数，在不同 CI 环境中可能导致测试不稳定。
- 影响：对用户：显存利用率提升，高显存设备可分配更多 KV Cache 块，支持更大模型或更长上下文。对系统：可能增加 GPU 错误风险，需监控非法内存访问。对团队：需注意平台一致性，未来应考虑动态限制机制替代硬编码。
- 风险标记：核心路径变更，平台不一致，缺少动态保护机制

关联脉络

- PR #7299 [Optim] Remove IPCLock between CacheManager and WorkerProcess: 同样涉及 WorkerProcess 和 Cache 优化，共享 KV Cache 相关模块
- PR #7313 [Optimization] [OP] [Models] dsk del prefill mask: 涉及 KV Cache 优化和 GPU 算子，技术领域相关