PR #7049 完整报告

PaddlePaddle/FastDeploy

[XPU] Fix speculate schedule

合并时间：2026-03-27 18:28

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7049

执行摘要

本 PR 修复了 FastDeploy 中 XPU 设备上推测调度缓存内核的一个 bug，通过优化线程分配和内存处理，提升调度正确性和防止内存泄漏。变更涉及两个文件，已被快速批准合并，影响范围限于 XPU 特定功能，建议关注内核逻辑调整。

功能与动机

基于提交消息 '[BugFix] xpu fix speculate schedule cache kernel'，本 PR 旨在修复 XPU 推测调度缓存中的问题，以提高在 XPU 设备上执行推测解码时的稳定性和正确性。PR body 未提供具体细节，但推测是解决内存管理或调度逻辑错误。

实现拆解

实现分为两个关键文件：

- custom_ops/xpu_ops/src/ops/mtp/speculate_schedule_cache.cc：添加条件语句，在 stop_flags.is_cpu() 时删除上下文指针，防止内存泄漏。cpp if (stop_flags.is_cpu()) { delete ctx; }
- custom_ops/xpu_ops/src/plugin/src/kernel/kunlun3cpp/mtp_kernel/speculate_schedule_cache.xpu：重构内核代码，包括：
  - 线程索引从 tid 改为 cid，并调整 nthreads 计算。
  - 固定数组大小，如 max_draft_tokens = 6，减少动态内存分配。
  - 修改 block_tables 访问逻辑，从数组索引改为单个变量。
  - 优化内存复制，如调整 GM2LM_ASYNC 调用顺序。

评论区精华

Review 过程中仅有一名审核者 zhupengyang 批准，评论为：

> LGTM

这表明变更被快速接受，但缺乏深度技术讨论，可能问题较明确或影响有限。

风险与影响

- 风险：内核中数组大小固定为 6，若输入超限可能导致溢出；线程逻辑变更可能影响并发性能；CPU 上下文删除条件可能不完整。建议添加测试验证边界情况。
- 影响：仅影响 XPU 设备的推测调度缓存功能，对用户透明，但能提高内部系统稳定性和效率，属于中等影响改进。

关联脉络

从近期历史 PR 分析中，未发现直接相关 PR；本 PR 属于 bugfix 类别，与仓库中其他 XPU 或调度相关 PR（如 PR 6680 涉及调度优化）无直接关联，但反映了对特定设备功能稳定性的持续维护。