PR #7029 完整报告

PaddlePaddle/FastDeploy

[XPU] Refactor get_padding_offset to single kernel.

合并时间：2026-04-13 11:04

原文链接：http://prhub.com.cn/PaddlePaddle/FastDeploy/pull/7029

执行摘要 本 PR 将 XPU 的 get_padding_offset 内核重构为单内核实现，合并了原有的两个内核以对齐 GPU 接口，优化 XPU 推理性能，但引入了 batch size 限制等潜在风险。

功能与动机 动机是重写 XPU 的 get_padding_offset 内核以对齐 GPU 实现，提升跨平台一致性和性能。PR body 中明确表示 "Rewrite get_padding_offset kernel to align with GPU implementation"。准确性测试结果显示基准和解码测试无明显异常，支持变更的可靠性。

实现拆解

- Python 接口层：更新 fastdeploy/model_executor/xpu_pre_and_post_process.py，移除 cum_offsets 和 token_num 输入，改为传入 cpu_token_num 属性，简化调用逻辑。
- 内核层：在 custom_ops/xpu_ops/src/plugin/src/kernel/kunlun3cpp/get_padding_offset.xpu 中，将 get_padding_offset 和 remove_padding 合并为单个内核，使用 shared memory 存储序列长度，并通过 cluster 并行计算累积偏移。
- 插件层：修改 wrapper 和插件代码（如 custom_ops/xpu_ops/src/plugin/src/wrapper/get_padding_offset.cpp），移除旧参数，适配新接口。
- 测试层：更新单元测试 custom_ops/xpu_ops/test/test_get_padding_offset.py，验证输出正确性。

评论区精华 Review 中重点讨论了以下问题：

- "MAX_BATCH_SIZE 硬编码为 1024 无边界检查 " – Copilot 和 fastdeploy-bot 指出这可能引发 shared memory 越界，建议添加断言。
- "token_num_cpu 类型问题 " – Copilot 提到从 Tensor 提取标量的类型错误风险，需显式转换。
- " 算法复杂度较高 " – Copilot 评论新实现可能导致 O(bs^2) 复杂度，性能可能退化。 讨论结论是部分问题被认可，但提交中未完全解决，凸显了设计权衡。

风险与影响

- 风险：batch size 超过 1024 时可能崩溃；token_num 截断可能出错；同步调用可能死锁；接口变更可能影响现有代码。
- 影响：提升 XPU 性能，减少内核调用开销；但需用户注意 batch size 限制；团队需维护新接口并加强测试覆盖。

关联脉络 从历史 PR 看，近期有多个 XPU 相关优化（如 PR #7320 修复 CI），但本 PR 是首次对 get_padding_offset 进行重大重构，标志着 XPU 算子对齐 GPU 的演进趋势，与仓库中其他 OP 优化 PR（如 #7313）共同体现性能优化主线。