PR #25077 完整报告

sgl-project/sglang

Fix(spec): Fix the crash issue in the FA3 backend when running with top-k > 1 and page_size > 1

合并时间：2026-06-09 08:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25077

执行摘要

- 一句话：修复 FA3 EAGLE draft decode 的 page_table scatter OOB
- 推荐动作：本 PR 已合并，无需额外操作。建议开发者在涉及 FA3 后端和 speculative decoding 时，关注 #27360 的修复逻辑，并确保类似场景在其他后端也得到测试。

功能与动机

根据 PR body 中的描述，在 EAGLE replay 步骤中，只有当前步骤所需的 decode span 加一个额外缓存槽是有效的，后续 speculative 步骤不应参与本轮 metadata 生成，否则 scatter_ 操作会导致 out-of-bounds 写入。作者通过运行命令重现了 CUDA_LAUNCH_BLOCKING=1 下的崩溃，并提供了 backtrace 截图。

实现拆解

1. 确认修复已被覆盖：在 PR 讨论中，维护者 hnyls2002 指出原 fix 已在 #27360 中合入，因此移除了原补丁中的 expand-slice 修复。
2. 新增回归测试：在 test/manual/attention/test_flashattn_backend.py 中添加了 test_draft_decode_set_expand_metadata_page_crossing 方法，模拟 topk=2、page_size=4、decode_length=2 的场景，cache_loc 中的两个 draft token 跨越不同页面。
3. 测试验证：测试构造 (bs=1, topk=2) 的输入，调用 draft_decode_set_expand_metadata，验证 page_table 的形状为 (2, decode_length+1) 且最后一列保持为 0，同时 cache_seqlens_int32 正确。

关键文件：

- test/manual/attention/test_flashattn_backend.py（模块 测试；类别 test；类型 test-coverage；符号 test_draft_decode_set_expand_metadata_page_crossing）：新增了回归测试 test_draft_decode_set_expand_metadata_page_crossing，验证 page_size > 1 且 topk > 1 时 draft_decode_set_expand_metadata 不越界写入。

关键符号：test_draft_decode_set_expand_metadata_page_crossing

关键源码片段

test/manual/attention/test_flashattn_backend.py

新增了回归测试 test_draft_decode_set_expand_metadata_page_crossing，验证 page_size > 1 且 topk > 1 时 draft_decode_set_expand_metadata 不越界写入。

```
    def test_draft_decode_set_expand_metadata_page_crossing(self):
        """
        Regression for fa3 EAGLE draft decode with topk > 1 and page_size > 1.
        cache_loc arrives num_steps-wide; callers pre-slice it to `decode_length`
        (the live draft tokens) before this helper runs, so the dedup'd scatter
        never writes past the (decode_length + 1)-wide expand page_table row even
        when consecutive draft tokens land on distinct pages.
        """
        bs, topk, page_size = 1, 2, 4
        decode_length = 2
        last_page_lens = torch.tensor([3], dtype=torch.int32)
        # 2 live draft tokens per (batch, topk) crossing into distinct pages.
        cache_loc = torch.tensor([[23, 28], [31, 36]], dtype=torch.int32)
        cache_seqlens_int32 = torch.zeros(bs * topk, dtype=torch.int32)
        # page_table is (decode_length + 1) wide (extra slot for the last partial
        # page); the trailing column must stay zero.
        page_table = torch.zeros(bs * topk, decode_length + 1, dtype=torch.int32)
        draft_decode_set_expand_metadata(
            cache_seqlens_int32=cache_seqlens_int32,
            page_table=page_table,
            last_page_lens=last_page_lens,
            decode_length=decode_length,
            cache_loc=cache_loc,
            topk=topk,
            page_size=page_size,
        )
        expected_page_table = torch.tensor([[5, 7, 0], [7, 9, 0]], dtype=torch.int32)
        expected_cache_seqlens = torch.tensor([5, 5], dtype=torch.int32)
        self.assertTrue(torch.equal(page_table, expected_page_table))
        self.assertTrue(torch.equal(cache_seqlens_int32, expected_cache_seqlens))
```

评论区精华

PR 中仅有一条来自维护者 hnyls2002 的评论，指出该崩溃已在 #27360 中修复，因此将 PR 转为添加回归测试以保留作者的贡献。无其他争议或讨论。

- 修复已被 #27360 覆盖，PR 转为回归测试 (other): PR 变更为仅添加回归测试。

风险与影响

- 风险：该 PR 本身仅新增测试，不修改生产代码，风险极低。但需注意测试覆盖的是 FA3 后端，如果其他 attention 后端（如 FA2、TRTLLM）存在类似问题，该测试不会覆盖。
- 影响：对用户无直接功能影响。对开发团队：新增的回归测试能防止类似回归再次出现，特别是在 EAGLE 和 FA3 组合使用时。影响范围限于测试文件，不影响生产路径。
- 风险标记：仅测试变更，依赖外部修复 #27360

关联脉络

- PR #27360 Fix FA3 draft decode scatter OOB with topk>1 and page_size>1: 该 PR 包含了实际的修复代码，本 PR 的测试用于验证该修复的正确性。
- PR #27235 refactor: replace oversized 1.3MB tiny_tokenizer.json fixture with a genuinely tiny byte-level BPE fixture: 同仓库近期历史 PR，虽无直接关联，但体现了测试基础设施的优化趋势。