PR #26628 完整报告

sgl-project/sglang

Revert "Fix FA DRAFT_EXTEND_V2 cache extent"

合并时间：2026-05-29 09:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26628

执行摘要

- 一句话：回滚 DRAFT_EXTEND_V2 cache extent 修复并推迟 CUDA graph 测试
- 推荐动作：建议读者重点关注此回滚的背景：当前 DRAFT_EXTEND_V2 的 cache extent 问题尚未解决，团队选择了暂时回退。如果使用相关特性，应切换到其他 attention 后端（如 Triton）或等待后续修复。同时值得阅读 KNOWN_FAILURES.md 中的详细记录，了解根因和所需变更。

功能与动机

PR body 仅表明回滚 PR#26512，未给出具体原因。推测原修复可能引入了其他问题或与现有流式处理冲突，团队决定回退并推迟测试，等待更完善的方案。

实现拆解

1. 回滚核心逻辑：在 python/sglang/srt/layers/attention/flashattention_backend.py 中恢复 init_forward_metadata 和 init_forward_metadata_replay_cuda_graph 函数中关于 DRAFT_EXTEND_V2 的分支，去除 effective_cache_seqlens = seqlens_in_batch + extend_seq_lens 的计算，直接使用 seqlens_in_batch（prefix 长度）。
2. 推迟 CUDA graph 测试：在 test/registered/attention/unittest/dense/test_fa3.py 和 test/registered/attention/unittest/dense/test_fa4.py 中，对 test_runner_mode_eagle_draft_extend_v2_cuda_graph_cases 方法添加 self.skipTest(...)，并附上原因说明：init_forward_metadata_replay_cuda_graph 缺少 cache_seqlens = prefix + extend 的支持，导致 kernel 读取错误。
3. 更新文档：修改 test/registered/attention/unittest/dense/README.md，将 FA3/FA4 的 DRAFT_EXTEND_V2 CUDA graph 状态从 deferred 更新为更明确的描述；在 test/registered/attention/unittest/KNOWN_FAILURES.md 中新增条目，记录 FA3/FA4 的 DRAFT_EXTEND_V2 CUDA-graph replay 问题及根因。

关键文件：

- python/sglang/srt/layers/attention/flashattention_backend.py（模块 注意力层；类别 source；类型 core-logic；符号 init_forward_metadata, init_forward_metadata_replay_cuda_graph）：核心变更文件，回滚了 DRAFT_EXTEND_V2 的 cache extent 修复，恢复了错误的 metadata 计算逻辑。

- test/registered/attention/unittest/dense/test_fa3.py（模块 FA3 测试；类别 test；类型 test-coverage；符号 test_runner_mode_eagle_draft_extend_v2_cuda_graph_cases）：为 FA3 的 DRAFT_EXTEND_V2 CUDA graph 测试添加 skipTest，避免 CI 失败。
- test/registered/attention/unittest/dense/test_fa4.py（模块 FA4 测试；类别 test；类型 test-coverage；符号 test_runner_mode_eagle_draft_extend_v2_cuda_graph_cases）：为 FA4 的 DRAFT_EXTEND_V2 CUDA graph 测试添加 skipTest，同步 FA3 操作。
- test/registered/attention/unittest/dense/README.md（模块 文档；类别 docs；类型 documentation）：更新测试矩阵表，反映 FA3/FA4 DRAFT_EXTEND_V2 状态由 deferred 改为明确的错误说明。
- test/registered/attention/unittest/KNOWN_FAILURES.md（模块 文档；类别 docs；类型 documentation）：新增 FA3/FA4 DRAFT_EXTEND_V2 CUDA-graph replay 的 known failure 条目。

关键符号：init_forward_metadata, init_forward_metadata_replay_cuda_graph, test_runner_mode_eagle_draft_extend_v2_cuda_graph_cases

关键源码片段

python/sglang/srt/layers/attention/flashattention_backend.py

核心变更文件，回滚了 DRAFT_EXTEND_V2 的 cache extent 修复，恢复了错误的 metadata 计算逻辑。

```
# init_forward_metadata 中 DRAFT_EXTEND_V2 分支（回滚后） # 注意：对于 DRAFT_EXTEND_V2，seq_lens 仅为 prefix 长度， # 而 extend 部分已通过 set_kv_buffer 写入，因此有效 cache 长度应为 prefix + extend。 # 此处只使用 prefix 长度，导致 kernel 无法访问 extend 的 KV。 metadata.cache_seqlens_int32 = seqlens_in_batch.to(torch.int32) metadata.max_seq_len_k = forward_batch.seq_lens_cpu.max().item() metadata.cu_seqlens_k = torch.nn.functional.pad(     torch.cumsum(seqlens_in_batch, dim=0, dtype=torch.int32), (1, 0) ) # init_forward_metadata_replay_cuda_graph 中 DRAFT_EXTEND_V2 分支（回滚后） # 同样只使用 seq_lens（prefix），未加上 extend_seq_lens， # 导致 CUDA graph replay 时 page_table 和 cache_seqlens 都缺少 extend 部分。 metadata.cache_seqlens_int32.copy_(seq_lens) metadata.max_seq_len_k = seq_lens_cpu.max().item() metadata.cu_seqlens_k[1:].copy_( torch.cumsum(metadata.cache_seqlens_int32, dim=0, dtype=torch.int32) )
```

评论区精华

Codex 审核机器人（chatgpt-codex-connector[bot]）在两条评论中提出了 P1 级别的正确性问题：

- 在 init_forward_metadata 中，对于 DRAFT_EXTEND_V2，仅使用 seqlens_in_batch（prefix 长度）会导致 cache_seqlens_int32、cu_seqlens_k 和 max_seq_len_k 不包含 extend 部分，从而 FA kernel 无法访问刚写入的 KV。

- 同样在 init_forward_metadata_replay_cuda_graph 中，使用 seq_lens 而不加 extend_seq_lens 会导致 CUDA graph replay 时 kernel 只读取 prefix KV，产生错误 attention。 两条评论均未被采纳，作者直接合并了 PR。
- DRAFT_EXTEND_V2 *cache extent* 正确性 (correctness): 未采纳，作者直接合并 PR。

风险与影响

- 风险：
  1. 正确性回归：回滚后，FlashAttention 在 DRAFT_EXTEND_V2 模式下（如 EAGLE v2 多草稿模型）的 *eager* 路径会出现 ~0.55 的绝对差异（vs HF 参考），CUDA graph 路径错误率达 ~82%。生产环境中使用该特性将输出错误结果。
  2. 测试覆盖缺失：相关测试被显式跳过，导致该问题在 CI 中不再被检测，增加了回归风险。
  3. 文档记录不充分：虽然 KNOWN_FAILURES.md 记录了问题，但用户可能不会主动查看，仍会误以为功能正常。 - 影响：影响范围：使用 FlashAttention 后端并启用 DRAFT_EXTEND_V2 的推理场景（如 EAGLE v2、Frozen-KV-MTP 等投机解码方法）。影响程度：严重正确性问题，会导致输出质量显著下降，甚至完全不可用。需要紧急修复。用户感知：在 CUDA graph 路径下，模型输出将明显错误；*eager* 路径误差较小但仍有影响。
- 风险标记：CUDA graph 路径测试被跳过，DRAFT_EXTEND_V2 正确性回归，生产环境输出错误

关联脉络

- PR #26512 Fix FA DRAFT_EXTEND_V2 *cache extent*: 本 PR 回滚了该修复，两者直接对立。