

PR #30439 完整报告

sgl-project/sglang

Fix FA3 prefill CP NaNs

合并时间: 2026-07-08 11:16

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30439>

执行摘要

- 一句话: 修复 FA3 prefill CP NaN 崩溃
- 推荐动作:
 - 该 PR 值得精读, 特别是 `_should_disable_scheduler_metadata_precompute` 的设计思路——通过单一函数封装分布式注意力下的调度器元数据安全条件, 避免类似问题扩散。
 - eager runner 中的元数据清理模式也值得借鉴, 确保回退路径不会残留过时状态。

功能与动机

修复 FA3 在 split-KV combine 阶段因过时的预计算调度器元数据导致的越界访问和 NaN 崩溃。当 CP 活跃时, `cache_seq_lens/num_splits` 可能跨 rank 变化, 预计算元数据变得不一致, 导致 `flash_fwd_combine_launch_template.h:52` 处 OOB 读取。

实现拆解

1. 提取公共条件函数 (`flashattention_backend.py`): 新增 `_should_disable_scheduler_metadata_precompute(server_args)`, 返回 `server_args.enable_prefill_cp` or `server_args.enable_dp_attention`。将原先仅针对 DP attention 的硬编码条件扩展为包含 prefill CP。
2. 修改初始化逻辑 (`flashattention_backend.py`): 在 `FlashAttentionBackend.__init__` 中, 将 `_disable_scheduler_metadata_precompute` 的赋值改为调用新函数, 并更新注释说明意图。
3. 清理过时 CP 元数据 (`eager_runner.py`): 在 `_execute_extend` 中, 当 `cp_v2_active` 为 `False` 时, 显式将 `forward_batch.attn_cp_metadata` 置为 `None`。防止混合模式下残留的 zigzag CP-v2 元数据干扰后续计算。
4. 验证: 本地单元测试和 4xH200 开发机上的 8 序列 burst 测试 (96 轮) 确认无 NaN/assert/ 非法访问。

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 `_should_disable_scheduler_metadata_precompute`): 核心修改: 新增 `_should_disable_scheduler_metadata_precompute` 函数, 并将预计算禁用条件从仅 DP attention 扩展到包含 prefill CP, 修复了 CP + FA3 的 NaN 问题。

- `python/sglang/srt/model_executor/runner/eager_runner.py` (模块 执行器; 类别 `source`; 类型 `data-contract`): 修复了当回退到非 CP-v2 路径时, 清除过时的 `attn_cp_metadata`, 防止后续使用错误的元数据导致 NaN。

关键符号: `_should_disable_scheduler_metadata_precompute`

关键源码片段

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

核心修改: 新增 `_should_disable_scheduler_metadata_precompute` 函数, 并将预计算禁用条件从仅 DP attention 扩展到包含 prefill CP, 修复了 CP + FA3 的 NaN 问题。

```
# 新增函数: 当启用 prefill CP 或 DP attention 时, 禁用调度器元数据预计算
# 以避免跨 rank 的 cache_seqlens/num_splits 不一致导致 OOB 访问
def _should_disable_scheduler_metadata_precompute(server_args) -> bool:
    """Return True if scheduler metadata precompute should be skipped."""
    # 使用 getattr 安全获取属性, 默认为 False
    enable_prefill_cp = getattr(server_args, "enable_prefill_cp", False)
    enable_dp_attention = getattr(server_args, "enable_dp_attention", False)
    return bool(enable_prefill_cp or enable_dp_attention)

# 在 __init__ 中替换原来的硬编码条件
# ... 其他初始化代码 ...
# 原来: self._disable_scheduler_metadata_precompute = bool(
#     getattr(server_args, "enable_dp_attention", False)
# )
# 现在: 通过函数统一判断
self._disable_scheduler_metadata_precompute = (
    _should_disable_scheduler_metadata_precompute(server_args)
)
```

`python/sglang/srt/model_executor/runner/eager_runner.py`

修复了当回退到非 CP-v2 路径时, 清除过时的 `attn_cp_metadata`, 防止后续使用错误的元数据导致 NaN。

```
cp_v2_active = is_cp_v2_active(forward_batch)
forward_positions = forward_batch.positions
if cp_v2_active:
    # 执行 CP-v2 相关准备
    prepare_cp_forward(forward_batch)
    # ... 省略分片逻辑 ...
    kwargs["input_embeds"] = sharded_hidden_states
    forward_positions = sharded_positions
else:
    # 关键修复: 当不处于 CP-v2 路径时, 清除可能残留的 attn_cp_metadata
    # 防止后续 FA3 等模块使用过时的 CP 元数据
    forward_batch.attn_cp_metadata = None
```

评论区精华

- 关于 assert 的取舍: Fridge003 最初建议在函数中添加 assert 确保 enable_prefill_cp 等属性存在, 但 ch-wan 回复“not needed”, 最终移除了 assert, 简化实现。
- 与现有代码的关系: kpham-sgl 询问此修复与 flashattention_backend.py L847-L863 部分是否冲突或冗余。结论是那段代码在 CP-v2 路径下已被跳过, 因此无冲突。
- 测试文件变更: Fridge003 在 test_cp_strategy_unit.py 上评论“No need to change this test”, 测试文件最终未修改。
- 是否保留 assert (style): 最终移除了 assert。
- 修复与现有代码的冲突 (design): 确认那部分代码被 CP-v2 路径跳过, 无冲突。
- 测试文件是否需要修改 (testing): 测试文件未修改。

风险与影响

- 风险:
 - 性能风险: 禁用调度器元数据预计算可能对 decode 的调度效率有微小影响, 但仅在启用 prefill CP 或 DP attention 时生效, 且这些场景下通信开销占主导, 影响可忽略。
 - 回归风险: 改动仅限于禁用条件扩展和元数据清理, 非 CP/DP 路径完全不受影响。eager runner 的清理仅作用于回退路径, 安全。
 - 兼容性: 通过 getattr 安全访问属性, 确保新旧配置兼容。
- 影响:
 - 用户影响: 修复了 Qwen3 MoE 等模型在使用 FA3 和 prefill CP 时的 NaN 崩溃, 提升可靠性。
 - 系统影响: 增强了对 GQA 混合前缀命中 / 长预填充批次的支持, 使 zigzag CP 能更稳定工作。
 - 团队影响: 提供了清晰的分布式注意力元数据管理策略, 为后续扩展 (如多节点 CP) 奠定基础。
 - 风险标记: 分布式一致性, 性能退化风险, 核心路径变更

关联脉络

- 暂无明显关联 PR