PR #30512 完整报告

sgl-project/sglang

[DSA] Fix IMA in fused top-k v2: write all output slots on tie overflow

合并时间：2026-07-08 19:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/30512

执行摘要

本 PR 修复了 fused DSA top-k v2 内核在 tie 溢出时未写入所有输出槽位导致的 CUDA 非法内存访问 (IMA) 问题，通过简单的 9 行 padding 逻辑确保每个槽位都被初始化，彻底消除了 GLM-5.2 with MTP 场景下的崩溃。

功能与动机

PR 的动机是 GLM-5.2 with MTP (EAGLE) 在使用 fused DSA top-k v2 路径（默认开启 SGLANG_OPT_USE_TOPK_V2=1）时，在 CUDA graphs 下约 1 分钟内即因 CUDA 非法内存访问崩溃。作者通过 GPU coredump 将故障定位到 sparse-MLA decode FMHA 的 UTMALDG.2D.GATHER4 指令，该指令使用了由 top-k v2 产生的垃圾 paged-KV 索引。

根因分析：collect 阶段最多保留 kMaxNumTie = 1024 个阈值 bin 的并列结果。当 above_count < topk - 1024 时，handle_tie 实际写入的输出槽位数少于 topk，但下游的 transform pass 会对所有 topk 个槽位进行页表转换，未写入的槽位包含未初始化的 staging memory，导致垃圾 KV 索引，最终在内核中触发 IMA。

实现拆解

1. 定位修复点：在 python/sglang/jit_kernel/include/sgl_kernel/deepseek_v4/topk_impl.cuh 中的 handle_tie 函数，该函数被所有子内核共享。
2. 添加 padding 逻辑：在 if (num_ties <= topk) 分支内，已有的部分写入（只写了前 num_ties 个槽位）之后，增加一个循环将 [num_ties, topk) 范围内的所有剩余槽位填充为 -1u（已有的“空 token”哨兵值）。每个线程至多执行 2 次循环迭代。
3. 避免性能损失：填充操作仅在 tie 溢出且 num_ties <= topk 时激活，不影响正常路径，因此吞吐量和接受长度保持不变。

python/sglang/jit_kernel/include/sgl_kernel/deepseek_v4/topk_impl.cuh

核心修改文件：在 handle_tie 函数中增加 padding 循环，确保所有 topk 输出槽位都被初始化。

```
// 文件：topk_impl.cuh
// handle_tie 函数的 num_ties <= topk 分支中
if (num_ties <= topk) {
    if (tx < num_ties)
        problem.emit(base + tx, tie_buffer[tx].idx);

    // 当 ties 数量少于 topk 时（因为 collect 阶段最多保留 kMaxNumTie 个），
```

```
    // 剩余未写入的槽位必须被填充为 -1 ("no token") 哨兵。
    // 否则下游 transform pass 会读到未初始化的 staging memory，
    // 转换出垃圾页表索引，导致 sparse attention kernel 中发生非法内存访问 (IMA)。
    for (uint32_t t = num_ties + tx; t < topk; t += kBlockSize) {
        problem.emit(base + t, -1u);
    }
} else if (num_ties <= kWarpSize) {
    // 其他分支保持不变 ...
}
```

评论区精华

- DarkSharpness：建议将 padding 初始化代码从函数起始位置移入 if (num_ties <= topk) 分支内部，以提升代码可读性和逻辑清晰度。
- Jiminator：接受建议并执行移动，同时解释“效果等价，但放在分支内可读性更好”。
- DarkSharpness（最终）："This fix should be 100% safe." —— 表明该修复经过了深入分析和确认。

风险与影响

风险：极低。改动仅限于 tie overflow 且 num_ties <= topk 时的 padding 操作，使用已有的 -1u 哨兵值，不影响其他代码路径。逻辑简单，但未附带单元测试（作者提供了手动 kernel 测试和真实 workload 验证）。

影响：

- 正向影响：彻底修复了 GLM-5.2 with MTP 在 fused top-k v2 路径下的稳定性问题，使全量 agentic sweep 达到 0 IMA（之前 concurrency-16 必崩）。
- 性能影响：无。填充操作几乎零开销，基准测试显示吞吐量和接受长度不变。
- 准确性：无退化。GSM8K 从 0.9500 微升至 0.9545，AIME25 表现正常。

关联脉络

本 PR 与历史 PR #30310（Increase KV cache pool when using indexShare by 15%）类似，都属于 DeepSeek/GLM 系列模型在 KV cache 管理或 indexer 路径上的稳定性修复。此外，本 PR 作者通过 GPU coredump 和 device-side printf 进行了深入的诊断，这种问题定位方法值得团队推广，对后续类似 debug 有参考价值。