PR #26313 完整报告

sgl-project/sglang

Fix stale forward_metadata leak in DP attn unpadded idle batch

合并时间：2026-05-26 07:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26313

执行摘要

- 一句话：修复 DP Attention 空闲批次 stale metadata 引发的 UAF
- 推荐动作：这是一个高价值、低风险的关键 bugfix，建议快速合入。根因分析详实，可作为调试复杂并发 bug 的范例。

功能与动机

PR #26157 添加的调试 canary 在 DSv4 分解 + DP Attention 场景下触发了 mapping_stack 的哨兵值，表明存在 use-after-free。进一步分析发现，当 batch_size == 0 时 forward_idle 未初始化 forward_metadata，导致 stale metadata 被后续 forward 使用，引发 gsm8k 精度从 0.975 骤降至 0.010。该问题仅 DP Attention 可复现，EAGLE、PD 分解等均非必要条件。

实现拆解

在 python/sglang/srt/model_executor/model_runner.py 的 forward_idle 方法中，增加 else 分支：当 forward_batch.batch_size == 0 时，将 self.attn_backend.forward_metadata 显式设为 None。

- 变更前：只有 batch_size > 0 时调用 init_forward_metadata，batch_size == 0 时什么也不做，残留上次 forward 的 metadata。
- 变更后：batch_size == 0 时手动清除，避免下游 _maybe_upgrade_forward_metadata 误用。
  - 该修改是安全的，因为 AttentionBackend.forward 在 idle 模式下会提前返回，不会解引用 None 的 metadata；_maybe_upgrade_forward_metadata 对 None 也会直接返回。

关键文件：

- python/sglang/srt/model_executor/model_runner.py（模块 执行引擎；类别 source；类型 data-contract；符号 forward_idle）：唯一的变更文件，修改了 forward_idle 方法，在 unpadded idle 分支清除 stale forward_metadata。

关键符号：forward_idle

关键源码片段

python/sglang/srt/model_executor/model_runner.py

唯一的变更文件，修改了 forward_idle 方法，在 unpadded idle 分支清除 stale forward_metadata。

```
# python/sglang/srt/model_executor/model_runner.py
# 修复 DP Attention 中 unpadded idle batch 的 stale metadata 问题

def forward_idle(
    self, forward_batch: ForwardBatch, pp_proxy_tensors=None
) -> Union[LogitsProcessorOutput, PPProxyTensors]:
    # 在 DP Attention 中，idle batch 可能被填充（batch_size > 0）用于 MLP 同步。
    # 填充时需要重新初始化 metadata，否则 attention kernel 看到的 batch_size 不对。
    # 对于未填充的情况（batch_size == 0），显式丢弃前一次 forward 留下的 stale metadata。
    # 否则 DeepseekV4Model.forward 中的 _maybe_upgrade_forward_metadata 会误用
    # 旧 batch 的 req_pool_indices，导致 SWA mapping 的 use-after-free。
    #
    # 安全说明：AttentionBackend.forward 在 idle 模式会提前返回，不会解引用 None；
    # _maybe_upgrade_forward_metadata 对 None 也直接返回。
    if forward_batch.batch_size > 0:
        self.attn_backend.init_forward_metadata(forward_batch)
    else:
        # 将 forward_metadata 设为 None，防止下游误用
        self.attn_backend.forward_metadata = None

    kwargs = {}
    if self.support_pp:
        kwargs["pp_proxy_tensors"] = pp_proxy_tensors
    ctx = (
        self.device_timer.wrap(metadata={"category": "idle"})
        if self.device_timer
        else contextlib.nullcontext()
    )
    with ctx:
        return self.model.forward(
            forward_batch.input_ids,
            forward_batch.positions,
            forward_batch,
            **kwargs,
        )
```

评论区精华

无 review 评论，仅通过 CI 验证。作者在原 PR body 中详细分析了根因链条，并提供了精度恢复数据（gsm8k: 0.010→0.975）。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。修改仅 2 行，且通过条件分支保护，不会影响正常 forward 路径。非 DP Attention 后端已有 idle 早期返回保护，None 值不会被解引用。对于 DSv4 后端，_maybe_upgrade_forward_metadata 对 None 类型检查会失败，直接返回。但需注意：如果未来新增后端在 idle 分支中假设 forward_metadata 不为 None，则可能引入新的问题。
- 影响：直接影响 DSv4 模型在 DP Attention 分解推理场景的正确性和稳定性（gsm8k 精度从 0.010 恢复到 0.975）；间接消除了一个潜在的跨 rank use-after-free 漏洞。对其他模型和后端无影响。
- 风险标记：核心路径变更，并发安全风险

关联脉络

- PR #26157 Add debug canary on SWA mapping_stack: 该 PR 暴露了本 bug（canary 触发），是根因分析的前提。
- PR #26292 Zero req_pool_indices padding in cuda-graph populate: 同属 kv-cache 管理层的内存 /metadata 清理修复，体现近期对类似问题的重视。