PR #34213 完整报告

sgl-project/sglang

[DCP] Reuse partial output in natural-log LSE merge

合并时间：2026-08-11 01:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34213

执行摘要

- 一句话：DCP LSE 合并改为原地复用 partial 输出，消除 OOM
- 推荐动作：改动很小但触及 DCP 核心合并路径，值得快速阅读 cp_lse_ag_out_rs_mha 的原地复用写法；关注重点是调用方张量生命周期约定和 LSE 基数遗留问题。后续若有统一 MHA/MLA LSE 合并测试或 base 处理，应跟踪。

功能与动机

PR body 明确指出原先 cp_lse_ag_out_rs_mha 在合并后 partial 输出不再使用的情况下仍分配第二个完整 FP32 [tokens, heads, output_dim] 张量；该合并并非 MHA 专属，Kimi-K3 走 absorbed-MLA 时 output_dim = kv_lora_rank，partial 形状为 [16384, 96, 512]，额外分配 3.000 GiB 导致 MI350X OOM。PR 关联 DCP 路线图 #29736。作者在评论区澄清 Triton DCP 的 cached-extend 路径仍调用该 helper 做 cross-rank 合并，因此 MLA 也会命中。目标是峰值内存修复而非吞吐提升。

实现拆解

1. 定位问题来源：在 python/sglang/srt/layers/dcp/comm.py 的 cp_lse_ag_out_rs_mha 中，torch.nan_to_num(cp_attn_out, ...) * scale 会先为清理结果分配同形状 FP32 张量，再与 scale 相乘产生第二个临时张量，在 [tokens, heads, kv_lora_rank] 大形状下成为峰值内存来源。
2. 改为原地运算：将表达式拆为 out = cp_attn_out.nan_to_num_(...) 与 out.mul_(scale)，直接复用调用方传入的 cp_attn_out 存储；同一个 out 随后进入 cp_group.all_reduce(out)，不再持有无用的 partial 副本。
3. 保持语义不变：单 rank 分支、_ag_lse + logsumexp 的 LSE 计算、scale 的 nan_to_num 防护、按 rank 的 head 切片与 return_lse 分支全部未改。
4. 配套与验证：最终合并仅包含 comm.py 一个文件（+2/-2）；作者曾新增单测但被 reviewer 要求移除，后续计划统一覆盖 MHA/MLA 合并路径。验证包括真实权重 Kimi-K3 DCP8 cap8/cap16 精度 A/B、16K cached extend 从 OOM 变为 HTTP 200，以及 reviewer 触发的 DCP 套件全部通过。

关键文件：

- python/sglang/srt/layers/dcp/comm.py（模块 DCP 通信；类别 source；类型 core-logic；符号 cp_lse_ag_out_rs_mha）：DCP 核心通信辅助函数所在文件，

cp_lse_ag_out_rs_mha 的原地复用改造直接决定峰值内存与调用方张量生命周期语义，是本次变更的唯一文件。

关键符号：cp_lse_ag_out_rs_mha

关键源码片段

python/sglang/srt/layers/dcp/comm.py

DCP 核心通信辅助函数所在文件，cp_lse_ag_out_rs_mha 的原地复用改造直接决定峰值内存与调用方张量生命周期语义，是本次变更的唯一文件。

```
def cp_lse_ag_out_rs_mha(
    cp_attn_out: torch.Tensor,
    cp_attn_lse: torch.Tensor,
    cp_group: GroupCoordinator,
    return_lse: bool = False,
):
    # 单 rank 场景直接返回，不进入合并逻辑，行为与之前完全一致。
    if cp_group.world_size == 1:
        return (cp_attn_out, cp_attn_lse) if return_lse else cp_attn_out

    cp_attn_lse = cp_attn_lse.contiguous()
    # all-gather 得到 [world_size, tokens, heads] 的 LSE 栈
    lses = _ag_lse(cp_attn_lse, cp_group)
    global_lse = torch.logsumexp(lses, dim=0) # 跨 rank 的自然对数域合并
    scale = torch.exp(cp_attn_lse - global_lse).unsqueeze(-1)
    scale = torch.nan_to_num(scale, nan=0.0, posinf=0.0, neginf=0.0)

    # 关键改动：不再新分配 [tokens, heads, output_dim] 的 FP32 临时张量，
    # 而是原地清理并缩放 cp_attn_out，随后直接作为 all-reduce 输入。
    # 调用方必须知道：函数返回后 rank-local partial 已被覆盖。
    out = cp_attn_out.nan_to_num_(nan=0.0, posinf=0.0, neginf=0.0)
    out.mul_(scale)
    out = cp_group.all_reduce(out)

    # 按 DCP rank 切回本 rank 负责的 head 片段。
    cp_num_heads = global_lse.shape[1] // cp_group.world_size
    cp_rank = cp_group.rank_in_group
    head_start = cp_num_heads * cp_rank
    head_end = cp_num_heads * (cp_rank + 1)
    out = out[:, head_start:head_end, :].contiguous()
    if return_lse:
        return out, global_lse[:, head_start:head_end].contiguous()
    return out
```

评论区精华

kpham-sgl 对函数名中的 MHA 却影响 Kimi-K3 表示困惑，tanth47 解释该 helper 按 query head 合并，不区分 MHA 的 v_head_dim 与 MLA 的 kv_lora_rank，Kimi-K3 的 absorbed

MLA 经 Triton DCP 也走此路径。kpham-sgl 还指出部分 attention backend 的 LSE 是 base-2、部分是 base-e，询问 MHA 侧是否需要类似 is_mla_dcp_lse_base_on_e 的开关，本 PR 明确不处理，作遗留观察。代码风格上，reviewer 要求移除 AI 生成的注释与冗余 docstring，并删除新增的单测，等待其提供覆盖 MHA/MLA 的统一合并测试。reviewer 触发的 test/registered/dcp/* 全套在 Ubuntu/H200/B200 全部通过。

- 为何 MHA 合并函数影响 Kimi-K3（MLA） (question): 已澄清：函数并非 MHA 专属，MLA 的 kv_lora_rank 同样作为 output_dim 参与合并。
- LSE 基数（base-2 vs base-e）在 MHA 侧是否需开关 (design): 本 PR 明确不处理，作为遗留观察项；若未来 MHA backend 引入 base-2 LSE 需重新评估。
- 移除新增单测，等待统一合并测试 (testing): 测试已从 PR 中移除，统一测试由 reviewer 后续补充。
- 清理 docstring 与 AI 生成注释 (style): 按 reviewer 要求完成清理，最终 diff 更简洁。

风险与影响

- 风险：
  - 原地覆写语义：cp_lse_ag_out_rs_mha 现在会覆写调用方传入的 cp_attn_out，任何在函数返回后继续复用该张量的调用方都会读到被缩放 / 合并后的数据。当前调用点未见问题，但形成了隐式契约，未来新调用者容易踩坑。
  - 测试缺口：最终合并没有新增持久化单测（作者添加的单测被 reviewer 移除），回归防护依赖现有 test/registered/dcp/* 集成测试，对峰值内存下降这类资源属性缺少直接断言。
  - LSE 基数遗留：MHA 路径若出现 base-2 LSE 的 backend，现有自然对数公式会算错跨 rank softmax 权重；本 PR 不改变现状。
  - 性能声明局限：单个 32K prefill 仍可能受 DCP Q all-gather 限制，本 PR 只移除了 3 GiB 临时分配，不能保证更大长度不 OOM。
  - 影响：用户 / 系统层面：修复了 Kimi-K3 等 MLA 模型在 TP8/DCP8 cached-extend 场景的 OOM，16384 tokens 时峰值显存下降约 3 GiB；对 MHA/GQA 的 DCP 路径同样生效，属于通用峰值内存优化。团队层面：为后续 DCP 长上下文支持扫清一个显存瓶颈，并为同类“合并后 partial 即死亡”的 kernel 提供了原地复用范式；测试移除意味着后续需要补齐统一的 LSE 合并测试。
  - 风险标记：原地覆写调用方张量，缺少新增单测覆盖，LSE 基数问题遗留，长 prefill 仍受其他限制

关联脉络

- PR #25090 DCP base for MHA/GQA on Triton / AMD-HIP: 首次引入 cp_lse_ag_out_rs_mha 合并逻辑，本 PR 是对该 helper 的原地复用优化。
- PR #32541 Kimi-K3 day-0, landing the kimi-k3 branch DCP work on main: Kimi-K3 DCP 落地 PR，本 PR 的复现场景基于 Kimi-K3 TP8/DCP8 cached-extend。