PR #33871 完整报告

sgl-project/sglang

[Performance] Reduce idle DP work in breakable prefill CUDA graphs

合并时间：2026-08-27 10:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33871

执行摘要

- 一句话：空闲 DP rank 跳伪 token 计算，MegaMoE 支持 per-rank prefill 图桶
- 推荐动作：值得精读。这是 SGLang DP attention + breakable prefill CUDA graph 路径上典型的“性能优化牵出正确性修复”案例：伪 token 掩码语义跨越 MoE top-k、attention 入口、TBO 子 batch 与 CUDA graph capture 四层，任何一层丢失都会静默引入未初始化值或让优化失效。ch-wan 的三轮 review 与 PR 的迭代修复（LSE 双模式契约、extra_kwargs 清零、TBO 掩码传播、capture fallback）是很好的分布式执行引擎调试范例。关注重点：prefill_graph_tolerates_sum_len 的豁免条件需随新 A2A backend / CP 模式演进持续维护。

功能与动机

PR body 指出：DP attention + breakable prefill CUDA graph 下，空闲 DP rank 会被改写成伪造 EXTEND batch 以保持多 rank 执行序列一致，但此前伪造行被当作真实 token 处理，导致 MoE top-k/dispatch 处理 dummy 行、忙碌 rank 为空闲 token 做多余 expert 计算、无真实 token 时仍执行 attention、且 prefill BCG 强制所有 rank 用 MAX_LEN 造成显著 padding。padding 开销在多轮对话与 agentic 负载中尤为明显（cache 亲和性、工具调用、到达时间不均常让部分 DP rank 空闲）。优化目标是消除这些无效工作并允许 MegaMoE 用 SUM_LEN 桶；review 中 BBuf 要求补充真实数据集精度结果，PR 随后给出 MMLU/GSM8K/AIME25 数据。

实现拆解

变更入口：核心改动集中在 ForwardBatch.prepare_mlp_sync_batch（python/sglang/srt/model_executor/forward_batch_info.py），该方法是空闲 rank 伪造 EXTEND batch 的地方。

1. 对伪造 batch 施加 dummy-token 掩码（forward_batch_info.py） - 旧逻辑在伪造 batch 时把伪 token 计为真实 token（num_token_non_padded.fill_(num_tokens)），目的是避免 MoE topk/A2A 把该 rank 当空而饿死后续层。 - 新逻辑：对非 hybrid SSM 且原始 forward mode 为 IDLE 的情况，将 num_token_non_padded 与 num_token_non_padded_cpu 置 0，使 MoE top-k 跳过 dummy 行、attention 可提前返回；hybrid SSM 因状态更新需要伪造行而保留旧行为。
2. attention 0-token 跳过与输出清零（layers/radix_attention.py、models/deepseek_v4.py） - 新增 _zero_skipped_attn_outputs 帮助函数，在 0 真实 token 时把预分配输出清零，避

免未初始化值（NaN/Inf）流入残差与 MoE routing。 - 在 _unified_attention_with_output_impl、unified_sparse_attention_with_output、attention_with_output_extra_kwargs 及 DeepSeek-V4 专用入口加 real_num_tokens == 0 早退；LSE 路径返回 zeroed padded LSE 以满足上层断言，output-only 路径返回 None。

3. MegaMoE per-rank SUM_LEN 桶（forward_batch_info.py、model_executor/runner/prefill_cuda_graph_runner.py） - 新增 prefill_graph_tolerates_sum_len()：仅 MegaMoE 且未启用 DSA/MLA prefill CP 时返回 True。原因是 MegaMoE 对称 dispatch buffer 按 num_max_tokens_per_rank 分配、NVLink barrier 依赖 rank 数而非本地 token 数，各 rank 可选不同 capture bucket。 - prepare_mlp_sync_batch 的 MAX_LEN 强制条件加上 and not prefill_graph_tolerates_sum_len()；_has_inactive_dp_rank 对 MegaMoE 豁免 sparse-DP eager fallback（idle rank 以 0 token 执行仍 collective-safe），并同步更新注释。
4. TBO 子 batch 掩码传播（batch_overlap/two_batch_overlap.py） - TboForwardBatchPreparer.prepare 改用 _get_num_token_non_padded_cpu(batch) 读取父级掩码（CPU mirror 未设置时回退 len(input_ids) 物理行数），用 _split_num_token_non_padded 拆成 CPU pair 传入 prepare_raw。 - filter_batch 新增 out_num_token_non_padded_cpu 参数，子 batch 不再硬编码 None，确保 attention 的 real_num_tokens == 0 跳过在 TBO 子 batch 上仍触发。
5. 测试与精度验证 - 新增 test/registered/unit/batch_overlap/test_tbo_children_dummy_token_mask.py（注册 CPU CI）：覆盖 idle 父级拆分为 (0,0)、padding 不计真实 token、CPU/device pair 一致、filter_batch 传播 CPU count。 - test/registered/unit/layers/test_radix_attention.py 新增 3 个用例：0 真实 token 时 LSE 路径返回 zeroed LSE、output-only 路径返回 None、extra_kwargs 清零输出。 - 精度验证：GSM8K 96.97% vs 96.59%、MMLU 90.00%/89.87% vs 90.13%、AIME25 97.08%±0.60% vs 95.83%±0.78%；DP attention BCG KL 测试 48 样本 prefill/decode cache-hit KL 均为 0.0。

关键文件：

- python/sglang/srt/model_executor/forward_batch_info.py（模块 批次信息；类别 source；类型 data-contract；符号 prefill_graph_tolerates_sum_len, prepare_mlp_sync_batch）：核心数据契约变更：空闲非 hybrid rank 的伪造 EXTEND batch 被掩码为 0 真实 token；新增 prefill_graph_tolerates_sum_len 判定使 MegaMoE 保留 SUM_LEN 而非强制 MAX_LEN。
- python/sglang/srt/batch_overlap/two_batch_overlap.py（模块 双批重叠；类别 source；类型 core-logic；符号 _get_num_token_non_padded_cpu, _split_num_token_non_padded, prepare, prepare_raw）：TBO 子 batch 掩码传播：修复 prepare_mlp_sync_batch 掩码后被 padded input_ids 重算撤销的问题，将父级已掩码的 CPU count 拆分传给子 batch。
- python/sglang/srt/layers/radix_attention.py（模块 注意力层；类别 source；类型 core-logic；符号 _zero_skipped_attn_outputs, _unified_attention_with_output_impl, unified_sparse_attention_with_output, attention_with_output_extra_kwargs）：

0-token 跳过的核心实现：新增 _zero_skipped_attn_outputs 清零预分配输出，并在多个注意力入口加 real_num_tokens == 0 早退，LSE 路径返回 zeroed padded LSE。

- test/registered/unit/batch_overlap/test_tbo_children_dummy_token_mask.py（模块 双批重叠；类别 test；类型 test-coverage；符号 TestTboChildrenDummyTokenMask, _make_extend_batch, _make_decode_capture_batch, test_masked_idle_parent_yields_zero_token_children）：新增 CPU 回归测试，锁定 TBO 分割不得复活被掩码的空闲 rank 伪 token，是 review 中发现的交互 bug 的直接守护。
- test/registered/unit/layers/test_radix_attention.py（模块 注意力层；类别 test；类型 test-coverage；符号 test_impl_zero_real_tokens_returns_zeroed_lse, test_impl_zero_real_tokens_output_only_returns_none, test_extra_kwargs_zero_real_tokens_zeroes_output）：新增 0 真实 token 场景回归测试，覆盖 LSE 返回契约、output-only None 契约与 extra_kwargs 输出清零。
- python/sglang/srt/model_executor/runner/prefill_cuda_graph_runner.py（模块 图执行器；类别 source；类型 data-contract；符号 _has_inactive_dp_rank）：_has_inactive_dp_rank 对 MegaMoE 豁免 sparse-DP eager fallback，确保 idle rank 在 SUM_LEN 桶下可安全进入 MegaMoE 0-token 执行，并更新注释。
- python/sglang/srt/models/deepseek_v4.py（模块 模型实现；类别 source；类型 data-contract）：DeepSeek-V4 专用注意力入口补充 0-token 跳过与输出清零，与该 PR 的统一注意力入口行为保持一致。

关键符号：prepare_mlp_sync_batch, prefill_graph_tolerates_sum_len, _unified_attention_with_output_impl, _zero_skipped_attn_outputs, _get_num_token_non_padded_cpu, _split_num_token_non_padded, _has_inactive_dp_rank

关键源码片段

python/sglang/srt/batch_overlap/two_batch_overlap.py

TBO 子 batch 掩码传播：修复 prepare_mlp_sync_batch 掩码后被 padded input_ids 重算撤销的问题，将父级已掩码的 CPU count 拆分传给子 batch。

```
@classmethod
def prepare(cls, batch: ForwardBatch, is_draft_worker: bool = False):
    if batch.tbo_split_seq_index is None or is_draft_worker:
        return

    tbo_children_num_token_non_padded = (
        cls.compute_tbo_children_num_token_non_padded(batch)
    )
    cls.prepare_raw(
        batch,
        tbo_children_num_token_non_padded=tbo_children_num_token_non_padded,
        # eager 分割时子 batch 继承父级已掩码的 CPU 计数，否则 attention
        # 的 0-token 跳过（读 num_token_non_padded_cpu）因 None == 0 不成立
        # 而失效，MAX_LEN 下的 dummy 行会被“复活”。CUDA graph plugin 路径
        # 传 None，其设备 buffer 每次 replay 自行刷新。
```

```
            tbo_children_num_token_non_padded_cpu=cls._split_num_token_non_padded(
                tbo_split_token_index=cls._compute_split_token_index(batch),
                num_token_non_padded=cls._get_num_token_non_padded_cpu(batch),
            ),
        )
```

python/sglang/srt/layers/radix_attention.py

0-token 跳过的核心实现：新增 _zero_skipped_attn_outputs 清零预分配输出，并在多个注意力入口加 real_num_tokens == 0 早退，LSE 路径返回 zeroed padded LSE。

```
def _zero_skipped_attn_outputs(*bufs: Optional[torch.Tensor]) -> None:
    """空闲 DP rank 跳过 attention 时清零输出，防止未初始化值传播。"""
    for buf in bufs:
        if buf is not None:
            buf.zero_()


# _unified_attention_with_output_impl 内、query narrowing 之前的 0-token 早退
if real_query_num_tokens == 0:
    _zero_skipped_attn_outputs(output)
    if return_lse:
        # 上层 unified_attention_with_output_and_lse 断言 lse 非 None，
        # 必须返回与 fake impl 声明一致的 zeroed padded LSE：
        # 形状取 query 前两维（padded 行数），dtype 为 float32。
        return query.new_zeros(
            (query.shape[0], query.shape[1]), dtype=torch.float32
        )
    # output-only 路径注册为 inplace schema，必须返回 None。
    return None

query = query[:real_query_num_tokens]
```

评论区精华

ch-wan 进行了三轮 review，核心交锋集中在 0-token 跳过的完整性与 TBO 掩码传播：

1. _unified_attention_with_output_impl 的 0-token 早退只处理了 output-only 返回 None，未处理 return_lse=True 时 unified_attention_with_output_and_lse 的 assert lse is not None（chunked-prefix MHA merge 会 AssertionError）。
2. attention_with_output_extra_kwargs 仍只 copy output[:0]，预分配输出含未初始化垃圾值，会流入残差与 MoE routing；仅 ROCm 清零 padded tail，其他平台泄漏。
3. TBO 子 batch 用 padded len(input_ids) 重算计数并硬编码 cpu=None，导致 --enable-two-batch-overlap 下 MAX_LEN 路径静默丢失优化。
4. 改用父级 CPU count 后，decode CUDA-graph capture 路径 num_token_non_padded_cpu=None 触发 min(split, None) TypeError；修复为缺失时回退物理行数。

5. 遗留 gap：cpu=None fallback 回归测试在 test/manual/ 未注册 CPU CI，capture 路径仍无 CI 守护。BBuf 在 issue 评论中要求真实数据集精度结果，PR 补充了 MMLU/GSM8K/AIME25 三组数据。

- 0-token 跳过必须尊重 LSE 返回契约 (correctness): 修复为 return_lse=True 时返回 zeroed padded LSE（query.new_zeros((query.shape[0], query.shape[1]), float32)），output-only 路径保持返回 None。
- attention_with_output_extra_kwargs 输出含未初始化垃圾值 (correctness): 补上 real_num_tokens == 0 早退 + _zero_skipped_attn_outputs(output)，测试 test_extra_kwargs_zero_real_tokens_zeroes_output 锁定。
- TBO 子 batch 用 padded input_ids 重算撤销 dummy-token 掩码 (correctness): 新增 _get_num_token_non_padded_cpu / _split_num_token_non_padded，把已掩码的父级 CPU count 传播给子 batch；新测试 test_tbo_children_dummy_token_mask.py 锁定。
- decode CUDA graph capture 的 None CPU count 触发 TypeError (correctness): _get_num_token_non_padded_cpu 优先父级 CPU count，未设置时回退 len(input_ids) 物理行数；round-3 确认修复。
- capture fallback 回归测试未注册 CI (testing): PR 合入时该 gap 未完全解决：作者未将 capture fallback 用例迁入 registered 测试文件。
- 真实数据集精度验证 (testing): PR 补充 MMLU / GSM8K / AIME25 三组结果，MegaMoE 与 DeepEP 配置均在噪声范围内（AIME25 略升）。
- _has_inactive_dp_rank 注释与 MegaMoE 豁免不符 (documentation): PR 更新注释，说明 MegaMoE 经 prefill_graph_tolerates_sum_len 豁免，idle rank 仍以 0 token 执行 MegaMoE。

风险与影响

- 风险：
  1. 正确性：0-token 跳过横跨多个注意力入口（_unified_attention_with_output_impl、unified_sparse_attention_with_output、attention_with_output_extra_kwargs、deepseek_v4 专用入口），若新增入口漏接，预分配输出将携带未初始化值（NaN/Inf）流入残差与 MoE routing；LSE 路径若返回 None 会直接 AssertionError。
  2. 兼容性：num_token_non_padded_cpu 语义从“无真实 token”扩展为“被掩码的空闲”，依赖该字段做非零判断的代码（如 TBO 子 batch 分割、capture batch 构造）存在 None 处理风险；_get_num_token_non_padded_cpu 的回退依赖“CPU mirror 未设置 = capture 路径”这一隐式约定。
  3. 回归面：MegaMoE SUM_LEN 豁免由 prefill_graph_tolerates_sum_len() 集中判定，新增 A2A backend 或 prefill CP 模式时若漏更新，会导致 DP rank 图桶不一致引发集体通信形状不匹配或挂起。
  4. 测试盲区：cpu=None fallback 回归测试位于 test/manual/ 而非注册测试，CI 无法拦截同类 TypeError。
  5. 性能波动：单轮满负载场景收益趋近于 0（所有 DP rank 均忙），仅多轮 /agentic 负载显著，需按负载特征评估部署价值。 - 影响：影响范围：DP attention（

--enable-dp-attention）+ breakable prefill CUDA graph + DeepSeek-V4 部署场景，DeepEP 与 MegaMoE A2A 后端均涉及。对多轮对话、agentic 工具调用这类 per-ranktoken 分布 skew 的负载收益明显（多轮 MegaMoE QPS +12.9%，单轮 TTFT 中位数-18.9%），单轮高吞吐场景几乎无变化。系统层面修改了 ForwardBatch 的公共数据契约（num_token_non_padded_cpu 现在可以表示“被掩码的空闲”），TBO 子 batch 与capture batch 构造逻辑同步调整，任何依赖该字段的注意力后端或 MoE dispatcher 均需感知这一语义。团队层面确立了“idle rank 伪 token 掩码 + 0-token 跳过”的优化模式，后续 hybrid SSM / 新 MoE backend 接入需遵循相同门控。 - 风险标记：核心推理路径变更，DP 同步语义敏感，共享注意力入口契约变更，MegaMoE 豁免依赖后端判定 ,capture 回退测试未入 CI

关联脉络

- 暂无明显关联 PR