PR #29413 完整报告

sgl-project/sglang

[DSA] Enable draft-extend CUDA graph for DeepSeek Sparse Attention

合并时间：2026-06-27 14:53

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29413

执行摘要

- 一句话：DSA draft-extend 支持 CUDA Graph，消除 host sync
- 推荐动作：推荐特别关注 _apply_cuda_graph_metadata 中 draft_extend_v2 分支的重写：如何用静态 page_table_1 宽度替代动态 seq_lens_cpu 计算，以及如何用设备上常量张量消除 .tolist() 同步。此模式可推广到其他后端的图捕获优化。同时讨论中 bot 的缓存建议可反向考虑，但当前实现已满足性能目标。

功能与动机

Spec-v2 decode 阶段之前从 host 读取 seq_lens_cpu 导致 CUDA Graph 回放非法（.tolist() 和数据依赖的 repeat_interleave 形状），且每次都要 host sync，限制性能。PR body: 'Makes the DSA draft-extend path CUDA-graphable, so spec-v2 decode stops reading seq_lens_cpu on the host.'

实现拆解

1. 标记后端不再需要 CPU seq_lens：在 dsa_backend.py 中添加类属性 needs_cpu_seq_lens = False，表明图捕获路径可以省略 host 镜像。
2. 兼容 seq_lens_cpu 为 None 的 eager 路径：在 init_forward_metadata 中，当 seq_lens_cpu 为 None 时，从 GPU 上的 seq_lens 推导 max_seqlen_k，确保 eager fallback 正确。
3. 固定解码与验证的 page table 宽度：在 _apply_cuda_graph_metadata 中，将 decode 和 target_verify 分支的 max_len / max_seqlen_k 改为 metadata.page_table_1.shape[1]（静态预分配宽度），彻底去除对 seq_lens_cpu 的引用。
4. 重写 draft_extend_v2 分支：使用固定的 self.speculative_num_draft_tokens 构造 extend_seq_lens 张量，并直接 expand 而非 repeat_interleave，消除动态形状和 host 同步。
5. 注册 DSA 后端到 draft-extend 图白名单：在 eagle_worker_v2.py 中条件导入并添加 DeepseekSparseAttnBackend 到 graph_supported_backend_types，使该后端能被 draft-extend CUDA Graph 使用。
6. MTP 预计算适配：在 dsa_backend_mtp_precompute.py 中，使 _precompute_replay_metadata 接受 seq_lens_cpu 为 None，并在 _precompute_decode_mode 中使用预分配的 page_table_1 宽度。

关键文件：

- python/sglang/srt/layers/attention/dsa_backend.py（模块 注意力层；类别 source；类型 core-logic；符号 DeepseekSparseAttnBackend.needs_cpu_seq_lens, DeepseekSparseAttnBackend.init_forward_metadata, DeepseekSparseAttnBackend._apply_cuda_graph_metadata, DeepseekSparseAttnBackend.init_forward_metadata_out_graph）：核心变更文件：添加 needs_cpu_seq_lens 标记，修改 init_forward_metadata 和 _apply_cuda_graph_metadata 以消除 CPU seq_lens 依赖，重写 draft_extend_v2 分支。
- python/sglang/srt/speculative/eagle_worker_v2.py（模块 推测解码；类别 source；类型 dependency-wiring；符号 EAGLEWorkerV2._capture_cuda_graphs）：将 DeepseekSparseAttnBackend 添加到 draft-extend CUDA Graph 白名单，使 EAGLE 推测解码使用该后端时能捕获图。
- python/sglang/srt/layers/attention/dsa/dsa_backend_mtp_precompute.py（模块 MTP 预计算；类别 source；类型 core-logic；符号 DeepseekSparseAttnBackendMTPPrecomputeMixin._precompute_replay_metadata, DeepseekSparseAttnBackendMTPPrecomputeMixin._precompute_decode_mode）：适配 MTP 预计算：使 seq_lens_cpu 可选，并在 _precompute_decode_mode 中使用静态 page table 宽度。

关键符号：DeepseekSparseAttnBackend._apply_cuda_graph_metadata, DeepseekSparseAttnBackend.init_forward_metadata, DeepseekSparseAttnBackend.init_forward_metadata_out_graph, EAGLEWorkerV2._capture_cuda_graphs, DeepseekSparseAttnBackendMTPPrecomputeMixin._precompute_replay_metadata, DeepseekSparseAttnBackendMTPPrecomputeMixin._precompute_decode_mode

关键源码片段

python/sglang/srt/layers/attention/dsa_backend.py

核心变更文件：添加 needs_cpu_seq_lens 标记，修改 init_forward_metadata 和 _apply_cuda_graph_metadata 以消除 CPU seq_lens 依赖，重写 draft_extend_v2 分支。

```
# python/sglang/srt/layers/attention/dsa_backend.py
class DeepseekSparseAttnBackend(
    DeepseekSparseAttnBackendMTPPrecomputeMixin, AttentionBackend
):
    # Decode/verify/draft 图回放从静态缓冲区重建 metadata，不再读取
    # seq_lens_cpu / seq_lens_sum；该标记用于跳过 D2H 同步。
    # eager fallback 从 GPU seq_lens 推导长度。
    needs_cpu_seq_lens: bool = False

    # 在 _apply_cuda_graph_metadata 中 draft_extend_v2 分支重写
    elif forward_mode.is_draft_extend_v2():
        # V2 draft-extend 使用固定的 speculative_num_draft_tokens 宽度
        # （类似 target_verify），保证图回放无 host sync。
        max_seqlen_k = metadata.page_table_1.shape[1] # 静态预分配宽度
        # seq_lens 已经包含由 prefill 写入的 draft KV
```

```
        cache_seqlens = (seq_lens + self.speculative_num_draft_tokens).to(torch.int32)
        metadata.cache_seqlens_int32.copy_(cache_seqlens)
        dsa_cache_seqlens = compute_dsa_seqlens(
            cache_seqlens, dsa_index_topk=self.dsa_index_topk
        )
        metadata.dsa_cache_seqlens_int32.copy_(dsa_cache_seqlens)
        # 将 page_indices 按 draft token 数重复（expand 不拷贝数据）
        page_indices = self.req_to_token[req_pool_indices, :max_seqlen_k]
        page_indices = page_indices.unsqueeze(1).expand(
            -1, self.speculative_num_draft_tokens, -1
        ).contiguous()
        metadata.page_table_1[:, :max_seqlen_k].copy_(page_indices)
        # 使用设备上的常量 tensor 替代之前的 .tolist() host 同步
        extend_seq_lens = torch.tensor(
            [self.speculative_num_draft_tokens] * bs,
            device=self.device,
        )
        metadata.extend_seq_lens.copy_(extend_seq_lens)
```

python/sglang/srt/speculative/eagle_worker_v2.py

将 DeepseekSparseAttnBackend 添加到 draft-extend CUDA Graph 白名单，使 EAGLE 推测解码使用该后端时能捕获图。

```
# python/sglang/srt/speculative/eagle_worker_v2.py
graph_supported_backend_types = [
    TritonAttnBackend,
    TRTLLMMLABackend,
    TRTLLMHAAttnBackend,
    TokenspeedMLABackend,
    FlashInferAttnBackend,
]
if _is_cuda or _is_musa:
    # DSA 仅支持 CUDA; 延迟导入避免非 CUDA 环境引入 deep_gemm
    from sglang.srt.layers.attention.dsa_backend import (
        DeepseekSparseAttnBackend,
    )
    graph_supported_backend_types.append(DeepseekSparseAttnBackend)

graph_supported_backend = isinstance(
    self.draft_extend_attn_backend,
    tuple(graph_supported_backend_types),
)
```

评论区精华

gemini-code-assist[bot]指出，每次 graph 回放时创建 extend_seq_lens 张量会引入多余的 H2D 复制，建议在 metadata 上缓存该张量。此建议未被作者采纳，但 PR 仍合并。zhendonghua询问 max_seqlen_k = int(seq_lens_cpu.max().item()) 是否是他 profiling 中看到的巨大 CUDA sync 原因，Fridge003确认 'seems so'。

- 创建 extend_seq_lens 张量的 H2D 开销 (performance): 作者未采纳该建议，PR 仍合并。该开销相比消除的 host sync 影响较小。
- seq_lens_cpu 导致的 CUDA sync 确认 (performance): 确认该行是 host sync 的来源，本 PR 通过移除该行解决。

风险与影响

- 风险：性能回归风险：使用静态 page table 宽度可能增加显存带宽消耗（多余行拷贝），但此模式已在 decode/verify 分支中使用，风险可控。 正确性风险：完全移除 seq_lens_cpu 依赖后，eager 模式下若未提供该参数，需确保后备逻辑（从 GPU seq_lens 计算）同样正确。代码已处理 None 情况，但需覆盖 overflow 边界。 兼容性风险：非 CUDA 环境通过 lazy import 避免引入 deep_gemm，但需确保条件判断 (_is_cuda or _is_musa) 准确。 未解决优化：bot 提出的 extend_seq_lens 缓存未实现，少量 H2D 复制仍然存在，但相比消除的 host sync 影响极小。
- 影响：用户影响：使用 DeepSeek DSA + spec-v2 推测解码的用户将获得显著性能提升（profile 显示 sync 减少）。 系统影响：CUDA Graph 回放路径更稳定，不再因 host sync 导致图重录。 团队影响：DSA 成为官方支持 draft-extend 的后端之一，后续需维护相关图捕获代码。
- 风险标记：静态宽度增加带宽消耗，eager fallback 正确性依赖，未缓存 extend_seq_lens 张量

关联脉络

- PR #29379 [Fix] DSA: size cudagraph page_table to req_to_token width: 此前修复了 DSA CUDA Graph 中 page_table 宽度计算错误导致越界，本 PR 进一步利用其固定宽度的设计来消除 seq_lens_cpu 依赖。
- PR #29414 [DSA] Make seq_lens_cpu optional for DeepSeek Sparse Attention: 同一功能栈的 follow-up，在本 PR stack 中后续合并，进一步使 seq_lens_cpu 可选。
- PR #29415 [DSA] Remove host H2D sync in _apply_cuda_graph_metadata: 同一功能栈的 follow-up，进一步消除 _apply_cuda_graph_metadata 中的 H2D 同步。