PR #29105 完整报告

sgl-project/sglang

[NPU] perf: precompute mamba conv-state track indices once per batch

合并时间：2026-06-25 14:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29105

执行摘要

- 一句话：NPU GDN 预填充中预计算 mamba 跟踪索引
- 推荐动作：值得合并的优化，改动量小、收益明确、验证充分。可作为 NPU 后端预填充性能优化的参考模式。

功能与动机

在 Ascend GDN 后端中，prefill/extend 路径在每一层线性注意力层都重新计算 mamba conv-state 跟踪索引，而该索引在整个批处理中保持不变。nonzero() 在 NPU 上产生数据依赖的形状并强制设备 - 主机同步，造成不必要的开销。PR 旨在将索引计算提前到批处理开始时一次完成，减少首 token 延迟。

实现拆解

1. 新增 _prepare_mamba_track_metadata 方法：在 AscendGDNAttnBackend 类中添加该方法，当 forward_metadata.has_mamba_track_mask 为真时，预计算 mamba_track_mask_indices = forward_batch.mamba_track_mask.nonzero(as_tuple=True)[0] 和 conv_states_mask_indices = forward_batch.mamba_track_indices[mamba_track_mask_indices]，并存入 forward_metadata.conv_states_mask_indices。
2. 在初始化入口调用预计算方法：在 init_forward_metadata 和 init_forward_metadata_out_graph 中，原 prepare_gdn_inputs 调用后追加 _prepare_mamba_track_metadata(forward_batch)，确保每个批处理仅执行一次。
3. 简化 forward_extend 中的逻辑：将原每层执行的 forward_batch.mamba_track_mask is not None and forward_batch.mamba_track_mask.any() 条件和 nonzero() 调用替换为直接读取 forward_metadata.has_mamba_track_mask 和预计算的 forward_metadata.conv_states_mask_indices，减少重复计算。
4. 行为保持验证：has_mamba_track_mask 定义为 bool(mamba_track_mask is not None and mamba_track_mask.any())，与原守卫条件相同；conv_states_mask_indices 计算结果与之前相同。通过 GPQA-Diamond 测试确认精度无回归。

关键文件：

- python/sglang/srt/hardware_backend/npu/attention/ascend_gdn_backend.py（模块 NPU 后端；类别 source；类型 core-logic；符号 _prepare_mamba_track_metadata）：新增 _prepare_mamba_track_metadata 方法预计算 mamba conv-state 索引，并在初始

化入口调用；简化 forward_extend 中逻辑。

关键符号：_prepare_mamba_track_metadata

关键源码片段

python/sglang/srt/hardware_backend/npu/attention/ascend_gdn_backend.py

新增 _prepare_mamba_track_metadata 方法预计算 mamba conv-state 索引，并在初始化入口调用；简化 forward_extend 中逻辑。

```
# python/sglang/srt/hardware_backend/npu/attention/ascend_gdn_backend.py

class AscendGDNAttnBackend(AscendMambaAttnBackendBase):
    # ... __init__ unchanged ...

    def _prepare_mamba_track_metadata(self, forward_batch: ForwardBatch):
        # 预先计算 mamba conv-state 跟踪索引，避免每层重复 nonzero()
        if self.forward_metadata.has_mamba_track_mask:
            mamba_track_mask_indices = forward_batch.mamba_track_mask.nonzero(
                as_tuple=True
            )[0]
            # 根据 mask 索引从 mamba_track_indices 中选出实际需要更新的 conv state 索引
            self.forward_metadata.conv_states_mask_indices = (
                forward_batch.mamba_track_indices[mamba_track_mask_indices]
            )

    def init_forward_metadata(self, forward_batch: ForwardBatch):
        # ... 原有逻辑 ...
        self.prepare_gdn_inputs(
            forward_batch.batch_size,
            forward_batch.forward_mode,
            forward_batch.spec_info,
        )
        self._prepare_mamba_track_metadata(forward_batch) # 新增：批处理开始时预计算一次
        self.graph_mode = False

    def init_forward_metadata_out_graph(self, forward_batch, in_capture=False):
        # ... 原有逻辑 ...
        self.prepare_gdn_inputs(...)
        self._prepare_mamba_track_metadata(forward_batch) # 新增：CUDA graph
        捕获时也预计算
        self.graph_mode = True

    # forward_extend 中原先每层执行的重复代码简化为：
    # def forward_extend(self, ...):
    # ...
    # if forward_metadata.has_mamba_track_mask:
    # mixed_qkv_to_track = ...
```

```
# conv_states.transpose(1, 2)[
# forward_metadata.conv_states_mask_indices
# ] = mixed_qkv_to_track
# ...
```

评论区精华

Reviewer iridiumine 提出两点疑问：

1. 删除原 forward_batch.mamba_track_mask is not None 条件是否会影响未来 CUDA graph replay 场景。作者回复 forward_extend 目前仅 eager 执行（仅 decode/target-verify 走 CUDA graph），因此安全。
2. 建议使用局部变量而非存储在 forward_metadata 上。作者回应已修改（标记为"done"）。未发现其他争议，pr 最终被 sglang-npu-bot 批准。

- 删除 forward_batch.mamba_track_mask is not None 条件是否影响 CUDA graph replay (correctness): 作者澄清当前 forward_extend 仅 eager 执行，decode/target-verify 才走 CUDA graph，因此无需担心。改动被接受。
- 建议使用局部变量而非存储在 forward_metadata 上 (design): 作者回应"done"，即改为使用局部变量（但最终提交显示仍存储在 forward_metadata 上，可能后续另有调整）。

风险与影响

- 风险：回归风险（低）：改动仅提取重复计算，逻辑等价。精度测试已覆盖。 兼容性风险（低）：仅修改 AscendGDNAttnBackend 类，未影响其他后端或 decode 路径。 CUDA graph 风险（讨论中提及）：当前 forward_extend 不走 graph，但未来若启用，需确保 ForwardMetadata 在 replay 时仍包含 has_mamba_track_mask。作者已确认当前设计安全。
- 影响：用户影响：NPU 上使用 GDN 后端的混合模型（如 Qwen3.6-27B）TTFT 降低 5-14%，吞吐提升约 4.8%。 系统影响：减少每层 nonzero() 调用，降低预填充阶段的设备 - 主机同步开销。 团队影响：约 22 行改动，单文件修改，评审迅速。
- 风险标记：缺少测试覆盖

关联脉络

- PR #28757 [AMD] [GLM5] skip redundant -inf pre-fill of HIP indexer MQA-logits: 同为后端性能优化，跳过冗余预填充操作。
- PR #29117 [CPU] optimize GDN prefill performance: 同一功能模块（GDN prefill）的 CPU 性能优化，思路类似。