PR #23112 完整报告

sgl-project/sglang

Add fmha_v2 attention backend for SM90/120

合并时间：2026-08-19 09:41

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23112

执行摘要

- 一句话：trtllm_mha 新增 fmha_v2 预填充后端，支持 SM90/120
- 推荐动作：值得精读，尤其是后端分发与依赖版本校验结合的设计模式。建议结合 PR body 中引用的 #32272（KV 交织优化）和 #32268/#32269（MTP 精度上报）一起阅读，才能看到 fmha_v2 的完整性能收益；同时关注后续是否为该分支补充单元测试。

功能与动机

PR body 明确说明：Enables TRT-LLM's fmha_v2 prefill attention backend for SM90/120. This backend is more performant than the current default (FA3), and also enables skip-softmax feature。作者还补充了 MTP 精度对比与 Kernel 级测量（小问题规模下比 FA3 快约 15%，可达 20%），以及 skip-softmax 长上下文收益（128k 时 bf16 +12.9%，fp8 +3.5%）。

实现拆解

1. 新增后端开关：在 TRTLLMHAAttnBackend.__init__ 中根据 is_sm90_supported()/is_sm120_supported() 设置 self.use_fmha_v2，并更新模块 docstring 说明各 GPU 平台的分发策略（SM90/SM120 走 fmha_v2 prefill，SM100 走 batch_context）。
2. 调整 forward_extend 的布局处理：fmha_v2 需要 Q 连续视图 (num_tokens, tp_q_head_num, head_dim)；K/V 保持池内 NHD 布局（[num_pages, page_size, num_kv_heads, head_dim]），decode 与 SM100 batch_context 仍走 _reshape_paged_kv_cache 换成 HND 布局。
3. 新增 prefill 分发分支：在非 decode 且 use_fmha_v2 且未启用 CP-v2 时调用 flashinfer.prefill.trtllm_fmha_v2_prefill，传入 Q_PAGED_KV_NHD layout、causal mask、window_left、sinks 以及从环境变量 SGLANG_SKIP_SOFTMAX_PREFILL_THRESHOLD_SCALE_FACTOR 读取的 skip-softmax 阈值；CP-v2 分支被显式排除以避免 per-shard mask 错误。
4. 参数校验配套：在 ServerArgs._handle_attention_backend_compatibility 中把 trtllm_mha prefill 支持范围从 SM100 扩展到 SM90/SM120；新增 SM120 与 fp8_e4m3 KV cache 或正 skip-softmax 阈值的互斥校验，以及 SM90/SM120 上 prefill context parallel 的禁止。未新增单元测试，依赖现有 trtllm_mha 相关覆盖。

关键文件：

- python/sglang/srt/layers/attention/trtllm_mha_backend.py（模块 注意力后端；类别 source；类型 core-logic；符号 TRTLLMHAAttnBackend, forward_extend, use_fmha_v2）：核心后端分发逻辑：新增 use_fmha_v2 开关、Q/K/V 布局切换与 fmha_v2 prefill 调用分支，改动量最大（+61/-19）。
- python/sglang/srt/server_args.py（模块 参数校验；类别 source；类型 configuration；符号 _handle_attention_backend_compatibility）：SM120 兼容性校验与 CP 限制，防止 FlashInfer 0.6.17 内核不支持导致运行时崩溃。

关键符号：TRTLLMHAAttnBackend.forward_extend, TRTLLMHAAttnBackend.use_fmha_v2, ServerArgs._handle_attention_backend_compatibility

关键源码片段

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

核心后端分发逻辑：新增 use_fmha_v2 开关、Q/K/V 布局切换与 fmha_v2 prefill 调用分支，改动量最大（+61/-19）。

```
# trtllm_mha_backend.py（head 版本核心片段）

class TRTLLMHAAttnBackend(...):
    def __init__(self, ...):
        ...
        # fmha_v2 prefill 内核仅支持 SM90 / SM120
        self.use_fmha_v2 = is_sm90_supported() or is_sm120_supported()
        ...

    def forward_extend(self, ...):
        ...
        q_scale = 1.0
        if (
            self.data_type == torch.float8_e4m3fn
            and (not self.is_xqa_impl or not forward_batch.forward_mode.is_target_verify())
            and not use_fused_qkv
        ):
            q = q.to(torch.float8_e4m3fn)

        # fmha_v2 要求 Q 连续且按 (token, head, dim) 视图传入
        if self.use_fmha_v2:
            q = q.contiguous().view(-1, layer.tp_q_head_num, layer.head_dim)
        else:
            q = q.reshape(-1, layer.tp_q_head_num, layer.head_dim)

        # NHD layout（池内原生格式）：[num_pages, page_size, num_kv_heads, head_dim]
        k_cache_raw, v_cache_raw = self.token_to_kv_pool.get_kv_buffer(layer.layer_id)
        is_decode_mode = (
```

```
        forward_batch.forward_mode.is_target_verify()
        or forward_batch.forward_mode.is_draft_extend_v2()
    )

    if not self.use_fmha_v2 or is_decode_mode:
        # decode 与 SM100 batch_context 需要 HND 布局
        k_cache, v_cache = self._reshape_paged_kv_cache(
            k_cache_raw, v_cache_raw, layer, layer.head_dim
        )
    else:
        # fmha_v2 预填充直接使用 NHD 布局（无需交换 head 与 page 维）
        k_cache = k_cache_raw.view(-1, self.page_size, layer.tp_k_head_num, layer.head_dim)
        v_cache = v_cache_raw.view(-1, self.page_size, layer.tp_v_head_num, layer.head_dim)

    kv_cache = (k_cache, v_cache)
    ...
    if is_decode_mode:
        # 原有 decode 分支：XQA（SM90/SM120）或 TRTLLM-GEN（SM100）
        pass
    elif self.use_fmha_v2 and not cp_v2_active:
        # CP-v2 必须走 cp_strategy.run_attention（按分片 mask），
        # 直接调用纯 causal 的 fmha_v2 会算错，故此分支显式排除
        paged_kv = torch.stack([k_cache, v_cache], dim=1)
        o = flashinfer.prefill.trtllm_fmha_v2_prefill(
            (q, paged_kv),
            input_layout="Q_PAGED_KV_NHD",
            workspace_buffer=self.workspace_buffer,
            seq_lens=self.forward_metadata.cache_seqlens_int32,
            max_q_len=self.forward_metadata.max_seq_len_q,
            max_kv_len=self.max_context_len,
            bmm1_scale=bmm1_scale,
            bmm2_scale=bmm2_scale,
            batch_size=forward_batch.batch_size,
            cum_seq_lens_q=self.forward_metadata.cu_seqlens_q,
            cum_seq_lens_kv=self.forward_metadata.cu_seqlens_k,
            block_tables=page_table,
            out_dtype=self.q_data_type,
            mask_mode="causal",
            window_left=layer.sliding_window_size,
            sinks=attention_sink,
            # skip-softmax 阈值环境变量，仅在支持的内核上生效
            skip_softmax_threshold_scale_factor=envs.SGLANG_SKIP_SOFTMAX_PREFILL_
            THRESHOLD_SCALE_FACTOR.get() or 0.0,
        )
    else:
        # SM100 的 trtllm-gen batch_context 回退路径
        ...
```

python/sglang/srt/server_args.py

SM120 兼容性校验与 CP 限制，防止 FlashInfer 0.6.17 内核不支持导致运行时崩溃。

```
# server_args.py：SM120 与 CP 兼容性校验（head 版本核心片段）
...
prefill_backend, decode_backend = self._resolved_attention_backends()
if "trtllm_mha" in (prefill_backend, decode_backend):
    if prefill_backend == "trtllm_mha" and not (
        is_sm90_supported() or is_sm100_supported() or is_sm120_supported()
    ):
        raise ValueError(
            "TRTLLM MHA backend for prefill requires Hopper (SM90), Blackwell (SM100), or
            SM120 GPUs. "
            "Please use a different prefill backend."
        )
    # FlashInfer 0.6.17 的 FMHAv2 在 SM120 上既不支持 E4M3 KV cache,
    # 也没有生成 skip-softmax 内核，提前拒绝避免首次 prefill 时报错
    if (
        prefill_backend == "trtllm_mha"
        and is_sm120_supported()
        and (
            self.kv_cache_dtype == "fp8_e4m3"
            or (envs.SGLANG_SKIP_SOFTMAX_PREFILL_THRESHOLD_SCALE_FACTOR.get() or 0.0)
            > 0
        )
    ):
        raise ValueError(
            "TRTLLM FMHAv2 prefill on SM120 does not support "
            "fp8_e4m3 KV cache or skip-softmax."
        )
    if (
        prefill_backend == "trtllm_mha"
        and not is_sm100_supported()
        and (self.enable_prefill_context_parallel or self.attn_cp_size > 1)
    ):
        raise ValueError(
            "Prefill context parallelism with the TRTLLM MHA prefill backend "
            "requires SM100 (trtllm-gen context kernel): the SM90/SM120 "
            "fmha_v2 prefill path does not implement CP shard masking."
        )
...
```

评论区精华

YAMY1234 在 review 中提出两条 SM120 兼容性意见，并被第二个 commit 落实：

- fp8_e4m3 KV cache：pinned FlashInfer 0.6.17 的 FMHAv2 在 SM120 上明确拒绝 E4M3 query，否则首次 prefill 即失败。
- skip-softmax：pinned 版本只为 SM90 生成 enable_skip_softmax=True 内核，SM120 上正阈值会命中 unsupported-config 断言。 另外，b8zhong 在 issue 评论中同意 Po-Han

的建议——相关改动不复杂时尽量放同一个 PR；akhilg-nv 解释本 PR 保持独立是因为 #32272 依赖尚未合入的 flashinfer API 变更。

- SM120 FMHAv2 不支持 E4M3 KV cache (correctness): 第二个 commit 在 server_args 中增加 SM120 + fp8_e4m3 的启动期拒绝，避免运行时崩溃。
- SM120 不支持 skip-softmax 内核 (correctness): server_args 在 SM120 上拒绝 SGLANG_SKIP_SOFTMAX_PREFILL_THRESHOLD_SCALE_FACTOR > 0 的配置。
- PR 拆分策略：是否并入 #32272 的 KV 交织优化 (question): 保持两个 PR 独立，本 PR 可随时合入，性能提升留待后续 PR。

风险与影响

- 风险：
  1. 依赖 FlashInfer 版本行为：新增分支直接调用 trtllm_fmha_v2_prefill，一旦升级 FlashInfer 导致 Q_PAGED_KV_NHD 布局或 kernel 支持范围变化，server_args 中的硬编码校验可能失效或误拦截。
  2. SM120 配置被硬性拒绝：fp8_e4m3 KV cache、skip-softmax、prefill context parallel 在 SM120 上直接 raise ValueError，属于启动期快速失败；若用户期望回退到 FA3，会看到明确的报错而非自动降级。
  3. 缺少直接单元测试：本次变更未新增测试文件，新分支（尤其 NHD 布局与 CP-v2 排除逻辑）主要依赖现有 trtllm_mha 集成测试覆盖。
  4. q.contiguous().view 可能引入额外拷贝：PR body 已说明单独合入本 PR 时 e2e 不提升，正是由于 kernel 前 KV 交织拷贝开销。 - 影响：影响使用 --attention-backend trtllm_mha 且运行在 SM90/SM120 上的用户：prefill 阶段可以启用 fmha_v2，并通过环境变量 SGLANG_SKIP_SOFTMAX_PREFILL_THRESHOLD_SCALE_FACTOR 开启 skip-softmax；SM120 用户在 fp8 与 skip-softmax 场景会被阻止使用该后端。对团队而言，后续 FlashInfer 升级需要同步复核这里的兼容性断言；对 MTP 场景，body 中报告了与 FA3 相当的精度（GPQA-20 0.700-0.750 区间）与可接受的 accept rate。 - 风险标记：核心注意力后端路径变更，依赖 FlashInfer 版本特定行为，缺少直接单元测试，SM120 配置被启动期硬性拒绝

关联脉络

- PR #32272 Improve fmha_v2 performance (KV interleaving): PR body 指明该 PR 用于消除 fmha_v2 prefill 前的 KV 交织大拷贝，是完整性能提升的后续依赖。
- PR #32268 MTP accuracy reporting support: PR body 中用于 MTP 精度上报，与本 PR 的 fmha_v2 + MTP 测试相关。
- PR #32269 MTP accuracy reporting support (follow-up): PR body 中用于 MTP 精度上报，与本 PR 的 fmha_v2 + MTP 测试相关。