

PR #52148 完整报告

vllm-project/vllm

[Attention] Fix FlashInfer SM12x prefill with sinks

合并时间: 2026-08-14 05:25

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/52148>

执行摘要

- 一句话: 修复 FlashInfer SM12x 带 sinks 的 prefill
- 推荐动作: 建议结合 #49718 与 #51987 一起精读: 本 PR 展示了如何用“平台能力族 + 特性开关 + wrapper 类型分派”在保留性能优化 (XQA) 的同时修复功能错误, 值得作为 FlashInfer 后端针对特定架构做分叉处理的参考案例。关注点在于 wrapper 的 run 签名差异与 scale 合并公式, 后续扩展 sink 支持时可复用该模式。

功能与动机

PR body 说明: 通用 FA2 prefill 路径接受 `sinks` 参数但不应用它, 所以 #49718 可以让 SM12x 用 XQA 做 decode, 却产生错误的 prefill 输出。该 PR 的目标是在保留 XQA 的前提下修复 sink-attention 模型的 prefill 路径, 而不像自动回退 #51987 那样直接移除 SM12x XQA 支持。stecasta 的实机验证也显示 stock 版本在 DGX Spark 上 GPT-OSS 的 GSM8K 得分为 0, 进一步佐证了问题。

实现拆解

1. 引入 sink 感知 wrapper: 在 `vllm/v1/attention/backends/flashinfer.py` 的导入中新增 `BatchAttentionWithAttentionSinkWrapper`, 为后续分支判断准备类型。
2. 修改 wrapper 工厂 `_get_prefill_wrapper`: 在 `self.has_sinks` 且 `current_platform.is_device_capability_family(120)` 时, 非因果与因果两个分支都改为构造 `BatchAttentionWithAttentionSinkWrapper`, 并显式传入 `q_data_type_prefill`、`kv_cache_dtype`、`head_dim` 与 `window_left`; 其中因果分支额外增加 `assert not self.is_kvcache_nvfp4`, 明确该 wrapper 不支持 NVFP4 KV cache。其余场景仍走 `BatchPrefillWithPagedKVCacheWrapper` (NVFP4 时后端为 `trtllm-gen`)。
3. 调整 forward 的调用分派: 根据 `prefill_wrapper` 的实际类型选择 `run` 的签名。sink wrapper 以位置参数传入 `self.sinks`, 并把 `q_scale * k_scale` 合并进总 `scale` (即 `self.scale * layer._q_scale_float * layer._k_scale_float`), `v_scale` 仍单独传; 通用 wrapper 保持原有 `q_scale/k_scale/kv_cache_sf` 参数, 不再传 `sinks`。
4. 测试配套: `tests/v1/attention/test_attention_backends.py` 中 `test_flashinfer_native_prefill_with_sinks` 增加 `default_vllm_config` fixture, 使测试在默认引擎配置下运行, 覆盖 SM120 + sink 的 prefill 路径。

关键文件:

- vllm/v1/attention/backends/flashinfer.py (模块 注意力后端; 类别 source; 类型 core-logic) : 核心修复文件: 在 SM12x + sinks 场景切换 prefill wrapper 并调整调用分派。
- tests/v1/attention/test_attention_backends.py (模块 注意力测试; 类别 test; 类型 test-coverage; 符号 test_flashinfer_native_prefill_with_sinks) : 为 test_flashinfer_native_prefill_with_sinks 增加 default_vllm_config, 确保默认配置下覆盖该场景。

关键符号: FlashInferImpl._get_prefill_wrapper, FlashInferImpl.forward

关键源码片段

vllm/v1/attention/backends/flashinfer.py

核心修复文件: 在 SM12x + sinks 场景切换 prefill wrapper 并调整调用分派。

```
def _get_prefill_wrapper(
    self,
    causal: bool = True,
) -> BatchPrefillWithPagedKVCacheWrapper | BatchDCPPrefillWrapper:
    # 返回 prefill 用的 wrapper。
    # SM12x 且启用 attention sink 时, 使用 FlashInfer 的
    # BatchAttentionWithAttentionSinkWrapper。通用 FA2 路径虽然接受
    # sinks 参数, 却不会真正应用它们, 会导致 prefill 输出错误。
    if not causal:
        # 非因果 prefill 不支持 DCP 与 NVFP4 KV cache
        if self.use_dcp:
            raise NotImplementedError(
                "FlashInfer non-causal prefill is not supported with DCP yet."
            )
        if self.is_kvcache_nvfp4:
            raise NotImplementedError(
                "FlashInfer non-causal attention is not supported with "
                "NVFP4 KV cache."
            )
        if self._noncausal_prefill_wrapper is None:
            if self.has_sinks and current_platform.is_device_capability_family(120):
                # sink wrapper 需要显式指定 dtype 与 head dim, 并按
                # window_left 处理滑动窗口
                self._noncausal_prefill_wrapper = (
                    BatchAttentionWithAttentionSinkWrapper(
                        self._get_workspace_buffer(),
                        get_kv_cache_layout(),
                        backend="auto",
                        q_data_type=self.q_data_type_prefill,
                        kv_data_type=self.kv_cache_dtype,
                        head_dim_qk=self.head_dim,
                        head_dim_vo=self.head_dim,
                        window_left=self.window_left,
```

```

        )
    )
else:
    self._noncausal_prefill_wrapper = (
        BatchPrefillWithPagedKVCacheWrapper(
            self._get_workspace_buffer(),
            get_kv_cache_layout(),
            backend="auto",
        )
    )
return self._noncausal_prefill_wrapper

if self._prefill_wrapper is None:
    if self.use_dcp:
        self._prefill_wrapper = BatchDCPPrefillWrapper(
            workspace_buffer=self._get_workspace_buffer(),
            dcp_a2a=self.dcp_a2a,
        )
    else:
        if self.has_sinks and current_platform.is_device_capability_family(120):
            # sink wrapper 目前不支持 NVFP4 KV cache, 因此显式断言
            assert not self.is_kvcache_nvfp4
            self._prefill_wrapper = BatchAttentionWithAttentionSinkWrapper(
                self._get_workspace_buffer(),
                get_kv_cache_layout(),
                backend="auto",
                q_data_type=self.q_data_type_prefill,
                kv_data_type=self.kv_cache_dtype,
                head_dim_qk=self.head_dim,
                head_dim_vo=self.head_dim,
                window_left=self.window_left,
            )
        else:
            # NVFP4 KV cache 需要 trtllm-gen backend, fa2/fa3 不支持
            backend = "trtllm-gen" if self.is_kvcache_nvfp4 else "auto"
            self._prefill_wrapper = BatchPrefillWithPagedKVCacheWrapper(
                self._get_workspace_buffer(),
                get_kv_cache_layout(),
                backend=backend,
            )
    assert self._prefill_wrapper is not None
    return self._prefill_wrapper

```

评论区精华

review 中没有针对实现的争议评论, 核心证据来自 stecasta 的实机验证: 在 DGX Spark (GB10, sm121) 上, GPT-OSS-20b GSM8K 5-shot (limit 200) stock 全部失败 (flexible-extract 0.000, Harmony 非终止错误 196), 应用本 PR 后升至 0.875 且错误为 0

, 接近 **TRITON_ATTN** 对照的 0.880/0; Qwen3-8B 与 Nemotron 3.5 Lightning NVFP4 (带 DSpark-3 drafter) 对比均无回归。mgoin 确认 DGX Spark 测试通过后合并, LucasWilkinson 与 mgoin 均批准。

- DGX Spark 实机验证与无回归确认 (testing): 修复有效且无回归, mgoin 确认 DGX Spark 测试通过后合并。

风险与影响

- 风险: 风险主要集中在新增的 wrapper 分支:
 - 影响范围被 self.has_sinks 与 current_platform.is_device_capability_family(120) 双重限定, SM90/SM100、无 sink 与 DCP/NVFP4 路径不变, 回归面较小。
 - forward 中新增的 isinstance 分派依赖两个 wrapper 的参数语义; 若 FlashInfer 版本间 BatchAttentionWithAttentionSinkWrapper 的 run 签名或 scale 合并公式变化, 可能导致数值差异或运行时错误。
 - NVFP4 场景通过 assert not self.is_kvcache_nvfp4 显式排除, 未来若要在 SM12x 上同时使用 NVFP4 与 sink 需要重新设计。
 - 测试仅覆盖默认配置与 Llama-3-8B 小批量场景, 量化、自定义 head_dim、多模态等组合未直接覆盖, 实机验证覆盖了 GPT-OSS、Qwen3 与 Nemotron。
 - 影响: 对用户: SM12x (如 DGX Spark、GB10) 上启用 attention sink 的模型 (GPT-OSS、Nemotron 等) prefill 输出从错误恢复正确, 同时保留 XQA decode 的性能收益。对系统: 仅 FlashInfer 后端在 SM120 平台上的 prefill wrapper 选择发生变化, 其他后端 (如 Triton) 与平台无影响。对团队: 避免了 #51987 回退 XQA 的性能损失, 提供了一条保留 XQA 的修复路线; 新增的类型分派模式也为后续在 SM12x 扩展 sink 支持提供了示例。
- 风险标记: SM12x 专属路径, wrapper 类型分派, NVFP4 限制断言, 依赖 FlashInfer 新 wrapper

关联脉络

- PR #49718 标题不在上下文中 (PR body 引用): PR body 指出该 PR 让 SM12x 可用 XQA decode, 但 prefill 在 sink 场景输出错误; 本 PR 补全 sink 感知的 prefill 路径。
- PR #51987 标题不在上下文中 (PR body 引用): 该 PR 是自动回退, 移除 SM12x XQA 支持; 本 PR 保留 XQA 并修复 prefill, 是另一条修复路线。