PR #22918 完整报告

sgl-project/sglang

[FlashInfer v0.6.11] [RL] Support FlashInfer per-token NVFP4 MoE

合并时间：2026-05-19 16:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22918

执行摘要

- 一句话：支持 FlashInfer per-token NVFP4 MoE 激活量化
- 推荐动作：建议合入。实现简洁，上游 FlashInfer 已合并相应 kernel。建议在正式版本中开启默认测试覆盖，并关注 CI 时间。

功能与动机

在 RL 场景中需要一种无需 QAT 的 NVFP4 部署方式。FlashInfer 提供了 per-token NVFP4 MoE kernel，该 PR 将其集成，实现 "activation is always online per-token quantized"。

实现拆解

1. 环境变量定义：在 sglang/srt/environ.py 中新增 SGLANG_FLASHINFER_NVFP4_PER_TOKEN_ACTIVATION（默认关闭）。
2. 量化参数调整：在 sglang/srt/layers/quantization/modelopt_quant.py 中，当开启该变量且使用 flashinfer_trtllm MoE 后端时，将 w13_input_scale 和 w2_input_scale 设置为全 1，忽略 checkpoint 中的激活 scale。
3. 核心量化路径修改：在 sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py 的 fused_experts_none_to_flashinfer_trtllm_fp4 函数中添加分支：启用时调用 flashinfer.nvfp4_quantize 进行 per-token 量化并传递 per_token_scale；否则走原有 quantize_hidden_states_fp4 路径。
4. 测试更新：在 test/registered/backends/test_flashinfer_trtllm_gen_moe_backend.py 中，为 NVFP4 基类添加 extra_env 支持，新增 TestFlashinferTrtllmGenMoeBackendPerTokenNVFP4Routed 测试，并移除部分不必要的 fused 测试以缩减 CI 时间。
5. 文档同步：更新 docs_new/docs/references/environment_variables.mdx 和 docs/references/environment_variables.md，记录新环境变量。

关键文件：

- python/sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py（模块 MoE 内核；类别 source；类型 core-logic）：核心变更：在 MoE forward 中添加 per-token NVFP4 量化分支，调用 FlashInfer 新 API 并传递 per_token_scale。
- python/sglang/srt/layers/quantization/modelopt_quant.py（模块 量化层；类别 source；类型 data-contract）：调整权重加载逻辑：per-token 激活启用时忽略 checkpoint 中的激活 scale。

- python/sglang/srt/environ.py（模块 环境变量；类别 source；类型 core-logic）：新增环境变量定义，控制 per-token NVFP4 特性的开关。
- test/registered/backends/test_flashinfer_trtllm_gen_moe_backend.py（模块 MoE 后端测试；类别 test；类型 test-coverage；符号 TestFlashinferTrtllmGenMoeBackendMXFP8, TestFlashinferTrtllmGenMoeBackendBF16, TestFlashinferTrtllmGenMoeBackendFP8Routed, TestFlashinferTrtllmGenMoeBackendNVFP4Routed）：测试更新：添加 per-token NVFP4 路由测试，通过 extra_env 支持注入环境变量。同时减少 CI 负担。
- docs_new/docs/references/environment_variables.mdx（模块 文档；类别 other；类型 documentation）：新环境变量文档更新。
- docs/references/environment_variables.md（模块 文档；类别 docs；类型 documentation）：旧文档同步更新。

关键符号：fused_experts_none_to_flashinfer_trtllm_fp4, quantize_hidden_states_fp4

关键源码片段

python/sglang/srt/layers/moe/moe_runner/flashinfer_trtllm.py

核心变更：在 MoE forward 中添加 per-token NVFP4 量化分支，调用 FlashInfer 新 API 并传递 per_token_scale。

```
    # 如果环境变量启用 per-token activation，则使用 flashinfer 的 nvfp4_quantize
    if envs.SGLANG_FLASHINFER_NVFP4_PER_TOKEN_ACTIVATION.get():
        from flashinfer import SfLayout, nvfp4_quantize

        hs_fp4_bytes, hs_sf_bytes, per_token_scale = nvfp4_quantize(
            hidden_states,
            1.0 / (448.0 * 6.0),
            sfLayout=SfLayout.layout_linear,
            per_token_activation=True, # 启用 per-token 量化模式
        )

        seq_len, hidden_size = hidden_states.shape
        hs_fp4 = hs_fp4_bytes.reshape(seq_len, hidden_size // 2)
        hs_scale_linear = hs_sf_bytes.view(torch.float8_e4m3fn).reshape(
            seq_len, hidden_size // 16
        )
    else:
        per_token_scale = None
        hs_fp4, hs_scale_linear = quantize_hidden_states_fp4(
            hidden_states, quant_info.w13_input_scale_quant
        )
    hs_scale = hs_scale_linear.view(torch.float8_e4m3fn).reshape(
        *hs_scale_linear.shape[:-1], -1
    )
```

评论区精华

- 安全 guard 确认：b8zhong 质疑 modelopt_quant.py 中 scale 覆盖是否会被其他后端误用，zianglih 确认有 enable_flashinfer_trtllm_moe guard。
- 测试删减决策：Fridge003 询问为何移除 MXFP8/BF16/FP8Routed 测试类，zianglih 解释 fused 测试非必需且耗费 CI，最终保留 routed 测试。
- 量化函数设计：Fridge003 建议将 per-token 量化提取为独立函数，zianglih 展开为内联 if-else，认为单调用者场景下内联更清晰。
- 文档迁移：zijiexia 要求将文档变更迁移到 docs_new，zianglih 已在后续 commit 中补充。
  - modelopt_quant.py 中 scale 覆盖的安全性 (correctness): 已确认路径安全，只对 flashinfer_trtllm 后端生效。
  - 测试删减决策 (testing): 决定移除这些测试，只保留 routed 测试。
  - per-token 量化函数设计：内联 vs 独立函数 (design): 采用内联分支，保留原有函数不变。
  - 文档迁移到 docs_new (documentation): 已完成迁移。

风险与影响

- 风险：主要风险包括： 1) modelopt_quant.py 中的 scale 覆盖逻辑虽设有 guard，但未来其他后端若共用此代码可能引发意外覆盖； 2) per-token 量化可能略微改变模型精度（实际测试显示精度持平或提升）； 3) 测试删减可能遗漏回归，但 NVFP4 routed 测试已保留。整体风险低，因环境变量默认关闭且代码路径隔离。
  - 影响：影响范围：仅限于使用 NVFP4 量化模型且采用 flashinfer_trtllm 或 flashinfer_trtllm_routed MoE 后端的用户。默认不开启，升级无感知。为 RL 场景提供更简洁的部署选项。团队需维护新环境变量及对应代码路径。
  - 风险标记：核心路径变更，环境变量开关，测试覆盖调整

关联脉络

- PR #24452 未知（依赖 PR）：该 PR 等待 #24452 合并后再合入。