PR #28347 完整报告

sgl-project/sglang

Revert Gemma4 modelopt fp4 MoE backend change

合并时间：2026-06-25 10:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28347

执行摘要

- 一句话：回退 Gemma4 modelopt fp4 MoE 注意力后端默认值
- 推荐动作：该 PR 是配合上游修复的配置回归，变更简单直接，值得了解但对工程师学习价值有限。用于确认依赖修复完成后恢复默认配置的流程。

功能与动机

PR body 说明 "Restore Gemma4 MoE NVFP4 SM10X default trtllm_mha"，依赖 #28144 修复了底层 trtllm_mha 在 SM10X 上的准确性问题，因此原临时兜底的 TODO 和特殊逻辑不再需要，应恢复默认行为。

实现拆解

1. 在 server_args.py 的 _handle_model_specific_adjustments 方法中，删除了 is_gemma4_modelopt_fp4、is_gemma4_moe、is_gemma4_modelopt_fp4_moe 三个局部变量的计算。
2. 将 default_attention_backend 的条件从 is_sm100_supported() and not is_gemma4_modelopt_fp4_moe 简化为仅判断 is_sm100_supported()，恢复所有 SM10X Gemma4 默认使用 trtllm_mha。
3. 在 MoE runner 后端选择逻辑中，将 if is_gemma4_modelopt_fp4 改为直接检查 self.get_model_config().quantization == "modelopt_fp4"，去除了对局部变量的依赖。
4. 删除了关于 TODO 的注释，因为底层的准确性问题已在 #28144 中修复。
5. 顺便清理了 model_config = self.get_model_config() 临时变量，改为直接内联调用，属于小型重构。

关键文件：

- python/sglang/srt/server_args.py（模块 配置；类别 source；类型 core-logic）：唯一变更文件，包含 Gemma4 注意力后端默认值的回退逻辑，以及几处内联优化。

关键符号：_handle_model_specific_adjustments

关键源码片段

python/sglang/srt/server_args.py

唯一变更文件，包含 Gemma4 注意力后端默认值的回退逻辑，以及几处内联优化。

```
# python/sglang/srt/server_args.py (Gemma4 特异性默认值调整部分 )
        elif model_arch in (
            "Gemma4ForConditionalGeneration",
            "Gemma4ForCausalLM",
            "Gemma4UnifiedForConditionalGeneration",
        ):
            # 之前为了绕过 SM10X trtllm_mha 对 modelopt_fp4 MoE 的准确性问题
            # 临时回退为 triton，现在问题已在 PR#28144 中修复，恢复统一使用 trtllm_mha
            default_attention_backend = (
                "trtllm_mha" if is_sm100_supported() else "triton"
            )
            if self.is_attention_backend_not_set():
                self.attention_backend = default_attention_backend
                logger.info(
                    f"Use {self.attention_backend} as default attention backend for Gemma4"
                )
            else:
                if self.attention_backend is None:
                    self.attention_backend = default_attention_backend

            prefill_backend, decode_backend = self.get_attention_backends()
            accepted_backends = ("trtllm_mha", "triton", "ascend", "intel_xpu")
            ...
            if is_sm100_supported() and self.moe_runner_backend == "auto":
                # 内联检查量化类型，不再依赖外部变量
                if self.get_model_config().quantization == "modelopt_fp4":
                    self.quantization = "modelopt_fp4"
                    self.moe_runner_backend = "flashinfer_trtllm"
                    ...
```

评论区精华

无 review 评论，只有一条 gemini-code-assist 的配额警告，没有涉及技术讨论。

- 暂无高价值评论线程

风险与影响

- 风险：如果 PR#28144 的修复不完整或存在未覆盖的情况，Gemma4 modelopt_fp4 MoE 在 SM10X 上使用 trtllm_mha 可能再次遇到准确性问题。由于该 PR 仅在默认配置上做切换，风险较低，但若用户未升级到包含 #28144 的版本，则可能回退前行为不一致。没有增加测试覆盖，回归风险由上游修复承担。
- 影响：仅影响使用 Gemma4 模型、modelopt_fp4 量化、且启用了 MoE 块的 SM10X 用户：他们在此 PR 后默认将获得 trtllm_mha 注意力后端，而非 triton。性能可能有改善，但准确性问题由 #28144 保证。不影响其他模型或硬件后端。
- 风险标记：上游修复完整性依赖

关联脉络

- PR #28144 Fix SM10X trtllm_mha accuracy issue: 本 PR 依赖此 PR 修复了底层准确性问题，才得以恢复默认后端。
- PR #29252 Tune Gemma4 26B-A4B B200 memory recipe: 同为 Gemma4 相关调整，文档侧内存配比调优，但无直接逻辑依赖。