PR #25006 完整报告

sgl-project/sglang

Enable trtllm_mha as gemma4 default attn backend.

合并时间：2026-05-18 05:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25006

执行摘要

- 一句话：Gemma4 默认注意力后端切换到 trtllm_mha
- 推荐动作：该 PR 变更简洁高效，性能收益显著且经过充分讨论和验证。值得关注其性能基准测试方法和 trtllm_mha 后端在 Gemma4 上的兼容性处理。对于使用 Blackwell GPU 的 Gemma4 用户，建议及时合并此变更。

功能与动机

为了在 SM100 平台（Blackwell GPU）上利用 TensorRT-LLM 生成的高效 MHA 内核来优化 Gemma4 模型的注意力计算性能，通过默认启用 trtllm_mha 获得显著的端到端加速，而无需用户手动指定后端。

实现拆解

1. 在 python/sglang/srt/server_args.py 的 _handle_model_specific_adjustments 方法中，修改 Gemma4ForConditionalGeneration 模型的分支：将原先统一的 self.attention_backend = 'triton' 替换为条件选择：若 is_sm100_supported() 返回 True，则设为 'trtllm_mha'，否则仍为 'triton'。
2. 日志信息更新为动态显示所选择的后端。
3. 该 PR 仅涉及一个文件的 8 行改动，无其他配置或测试配套。

关键文件：

- python/sglang/srt/server_args.py（模块 服务配置；类别 source；类型 core-logic；符号 _handle_model_specific_adjustments）：核心控制文件，修改了 Gemma4 的默认注意力后端选择逻辑，是此 PR 的唯一变更文件。

关键符号：_handle_model_specific_adjustments

关键源码片段

python/sglang/srt/server_args.py

核心控制文件，修改了 Gemma4 的默认注意力后端选择逻辑，是此 PR 的唯一变更文件。

```
# python/sglang/srt/server_args.py 关键片段
elif model_arch == "Gemma4ForConditionalGeneration":
    # SM100 (Blackwell) 平台优先使用 TensorRT-LLM MHA 后端，
```

```
    # 否则回退到 Triton 后端保持兼容。
    if is_sm100_supported():
        self.attention_backend = "trtllm_mha"
    else:
        self.attention_backend = "triton"
    logger.info(
        f"Use {self.attention_backend} as default attention backend for Gemma4"
    )
```

评论区精华

核心讨论围绕 trtllm_mha 对 Gemma4 子型号（E2B、E4B）的兼容性。Reviewer kpham-sgl 最初担心这些变体存在 KV cache reuse 的特殊路径，可能存在兼容问题，建议仅对大模型（31B 和 26B-A4B）启用。作者 wenscarl 回复确认 trtllm_mha 已在这些子模型上通过验证，因为 trtllm-gen 内核直接通过 page_table 访问分页 KV cache，不需要额外的窗口检索路径，KV 共享重定向已在模型侧处理。此外还讨论了 flashinfer 版本支持 headdim=512 的版本号（v0.6.10.post1）。

- E2B/E4B 变体 KV cache reuse 兼容性 (design): wenscarl 验证了 trtllm_mha 在这些变体上正常工作，因为 trtllm-gen 内核直接读取分页 KV cache，无需额外窗口路径，且 KV 共享重定向已在模型层处理。
- flashinfer 版本依赖 (question): wenscarl 回复需要 v0.6.10.post1。编辑者计划在合并前升级版本。

风险与影响

- 风险：
  1. 回归风险：低，因为仅在用户未显式指定后端时改变默认值，且已在 SM100 上验证所有 Gemma4 变体（31B、26B、E2B、E4B）。
  2. 性能风险：低，基准测试显示显著提升。但吞吐量测试中 trtllm_mha 下有一请求被静默丢弃（999/1000），原因不明，需关注。
  3. 兼容性风险：trtllm_mha 依赖 TensorRT-LLM 和特定 flashinfer 版本（≥v0.6.10.post1），但 SM100 平台已预期满足。非 SM100 平台行为不变。 - 影响：对用户：Gemma4 用户在 SM100 平台无需任何配置即可自动获得 13-22% 的性能提升。对系统：仅变更默认值，无新增依赖或配置项。对团队：需要确保 flashinfer 版本升级到 v0.6.10.post1，否则 trtllm_mha 可能不可用（但 PR 已说明在合并前会升级）。 - 风险标记：吞吐量测试偶发请求丢弃，依赖 flashinfer v0.6.10.post1

关联脉络

- PR #25488 Revert "[attn backend] avoid initing parent class's workspace buffer": 涉及 Blackwell 平台 trtllm_mla_backend 的修复，与当前 PR 共享 Blackwell 上下文和类似的 TensorRT-LLM 后端工作。
- PR #25457 [diffusion] add memory-aware component load order: 同样涉及 SM100/Blackwell 相关的性能优化和配置默认值变更，展示了团队近期对 Blackwell 平台的

持续投入。