PR #22098 完整报告

sgl-project/sglang

Revert "[Bugfix] Temporarily skip TRTLLM attention on (G)B300 (SM103) to avoid high-concurrency hang"

合并时间：2026-04-04 17:17

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22098

执行摘要

本 PR 撤销了针对 SM103 GPU 高并发挂起的临时规避代码，恢复 TRTLLM attention 后端在 Blackwell 系列 GPU 上的默认使用，以提升性能和代码简洁性。变更基于 flashinfer v0.6.7.post2 的外部修复，影响注意力模块和服务器配置逻辑，建议关注核心路径变更和依赖协调。

功能与动机

原 PR #21906 为临时 workaround，因 flashinfer 在 SM103 GPU 上存在 TRTLLM attention 高并发挂起问题（issue #21904）。现 flashinfer v0.6.7.post2 已修复此问题（PR #22097），故撤销临时措施以回归正常逻辑。PR body 明确说明："This temporary fix can be reverted due to release of flashinfer v0.6.7.post2"。

实现拆解

- nsa_backend.py：移除条件分支，将 SM103 特殊处理（回退到 FA4）改为统一使用 flashinfer TRTLLM attention，代码从 50 行缩减至 30 行左右。
- server_args.py：删除 is_sm103_supported 相关检查，简化 _set_default_nsa_backends、_handle_model_specific_adjustments 等函数中的后端选择逻辑，影响多模型默认配置。
- common.py：删除 _check_cuda_device_exact 函数和 is_sm103_supported 常量，减少代码冗余。

评论区精华

Review 中仅 gemini-code-assist[bot] 提出文档改进建议：

> "For clarity, it would be better to be more precise in this error message. The is_sm100_supported() check covers the entire SM10x family of Blackwell GPUs, not just SM100."

建议使用"SM10x"术语提高错误消息准确性，但此建议未被采纳，PR 直接合并。

风险与影响

风险：

1. 依赖 flashinfer 外部修复，若新版本仍有缺陷，可能导致 SM103 GPU 回归挂起。
2. 错误消息未更新，用户可能误解支持架构范围。
3. 变更涉及核心注意力路径，需通过 CI 测试确保 SM103 稳定性。

影响：

- 正面：SM103 GPU 用户恢复高性能 TRTLLM attention 后端，减少工作负载切换开销。
- 负面：若外部修复不彻底，可能引入新问题；影响范围限于 Blackwell 系列 GPU 配置。

关联脉络

本 PR 是技术债务清理的一部分，与近期 PR 形成关联链：

- PR #21906 引入临时规避，标签为 bugfix、run-ci、performance。
- PR #22097 升级 flashinfer 依赖，标签为 dependencies、run-ci，为本 PR 提供前提。

这反映团队通过外部依赖协调快速响应硬件问题，并在修复后及时回退临时代码，保持代码库健康。