PR #21595 完整报告

sgl-project/sglang

Change default mm-attention backend from triton_attn to fa4

合并时间：2026-04-01 14:29

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21595

执行摘要

本次 PR 将 Blackwell datacenter GPUs（SM100）上的默认多模态注意力后端从 triton_attn 切换为 fa4（Flash Attention 4），基于基准测试显示 TTFT 和延迟显著提升（如 P99 TTFT 降低 57%）。同时规避了 RTX 5090（SM120）上的编译问题，确保变更仅影响稳定平台，是针对性性能优化。

功能与动机

变更动机源于性能基准数据：在模型 Qwen/Qwen3.5-35B-A3B-FP8 上，fa4 后端在 Blackwell GPU（SM100）相比 triton_attn，在中等并发下 TTFT 降低 18.7%、P99 TTFT 降低 57.0%。目标是通过后端优化提升多模态服务响应速度，同时由于 FA4 CUTE 内核在 SM120 上的编译问题，限制了变更范围以保持稳定性。

实现拆解

实现集中于文件 python/sglang/srt/layers/attention/vision.py 的 _determine_attention_backend 函数。核心改动如下：

- 更新注释，说明平台默认逻辑：CUDA Hopper（SM90）为 "fa3"、Blackwell（SM100）为 "fa4"、其他架构为 "triton_attn"。
- 修改代码逻辑，通过 get_device_capability() 检测设备能力，动态选择后端：
- 变更仅影响默认后端选择，服务器参数和构造函数参数优先级不变。

评论区精华

review 讨论中，b8zhong提出兼容性疑虑：

> "If I remember does FA4 work in CU12.x Docker? Since we switch to upstream FA4 and it has some compatibility issue with CU12 or something." 作者 JustinTong0323回复澄清： "The vision fa4 backend uses flash-attn-4>=4.0.0b4 ... it's already a required dependency ... so if there were a CU12.x compatibility issue, it would already be failing today for fa3 too." 讨论聚焦于依赖稳定性，结论是变更安全，无新风险。

风险与影响

风险：

- 编译风险：FA4 CUTE 内核在 SM120 上有已知编译问题，PR 通过限制 SM100 默认来规避，但未来扩展需测试。
- 依赖风险：fa4 基于上游 flash-attn-4，版本更新可能引入行为变化。
- 回归风险：变更仅影响默认选择，用户显式指定后端时无影响，但需确保 fa4 在 SM100 上稳定。

影响：

- 用户：Blackwell GPU 用户自动获得性能提升，改善多模态服务体验。
- 系统：优化后端选择逻辑，增强平台适配性。
- 团队：变更微小易维护，但需监控依赖和平台兼容性。

关联脉络

从历史 PR 看，PR #17122修复 GLM-4V 模型的多模态注意力问题，标签含 "multimodal"，与本 PR 共享功能模块；PR #21314优化 SM120 性能，涉及 GPU 架构特定优化，与本 PR 规避 SM120 问题的策略相关。整体上，这反映了仓库在多模态和 GPU 性能优化上的持续演进，强调平台适配和基准驱动决策。