PR #22079 完整报告

sgl-project/sglang

[nvidia] Gemma4 nvfp4 fix

合并时间：2026-04-10 08:44

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22079

执行摘要

- 一句话：修复 Gemma 4 NVFP4 模型在 GB200 上 Triton attention kernel 因 PTX 寄存器耗尽导致的崩溃问题。
- 推荐动作：建议工程师精读此 PR 以了解 Triton kernel 硬件适配模式，关注块大小调优对寄存器压力的影响。设计决策中值得注意：为不同 CUDA 能力添加专用分支以避免寄存器耗尽，但可考虑扩展更细粒度优化以适应不同场景。

功能与动机

根据 PR body，Gemma 4 NVFP4 checkpoints 在 GB200 上使用 Triton attention 后端时崩溃，错误为 'PTXAS error: Register allocation failed with register count of 255'。根本原因是 _get_block_sizes_for_extend_attention 函数缺少针对 CUDA_CAPABILITY[0] == 10（Blackwell 架构）的分支，导致使用 Hopper 的块大小配置时寄存器压力过大，特别是在 KV 缓存为 fp8 时加剧。

实现拆解

在 python/sglang/srt/layers/attention/triton_ops/extend_attention.py 的 _get_block_sizes_for_extend_attention 函数中，添加一个新的条件分支：当 CUDA_CAPABILITY[0] == 10（Blackwell 架构）时，根据查询长度 Lq 设置块大小：若 Lq <= 256，BLOCK_M, BLOCK_N = (64, 64)；否则为 (16, 64)。这替代了原先的 Hopper 分支，以适配 sm_100a 的寄存器约束，避免 PTX 寄存器耗尽。

关键文件：

- python/sglang/srt/layers/attention/triton_ops/extend_attention.py（模块 attention/triton_ops）：核心修复文件，修改了 Triton attention kernel 的块大小选择逻辑以适配 Blackwell 架构，避免 PTX 寄存器耗尽，直接影响 Gemma 4 NVFP4 模型在 GB200 上的运行。

关键符号：_get_block_sizes_for_extend_attention

评论区精华

reviewer alexnails 在代码第 77 行评论：'can you also include Lq <= 128 case? (e.g I believe 128x64, but it could be 128x128 if I am missing something from Blackwell tuning guide)'。此建议旨在进一步优化块大小选择以提升性能，但最终代码未采纳该修改，可

能因修复优先级或测试覆盖不足，无其他深入讨论。

- 为 Blackwell 架构添加更优块大小分支 (design): 建议未采纳，代码保持原修改，可能因时间紧迫或已有配置足够。

风险与影响

- 风险：技术风险包括：1) 回归风险：新分支可能影响其他 Blackwell 模型或配置的性能，需测试覆盖；2) 兼容性风险：仅针对 sm_100a，可能未覆盖其他 Blackwell 变体如 sm_120a（如评论提及）；3) 性能风险：缺少更细粒度分支可能导致某些场景性能次优。
- 影响：影响范围：直接受益于 Gemma 4 NVFP4 模型在 GB200 上的用户，确保模型可运行，提升硬件兼容性。系统层面，修复了 Triton attention 后端在 Blackwell 架构上的一个崩溃 bug，增强对新兴硬件的适配能力。团队需注意此硬件特定调优可能需后续迭代优化。
- 风险标记：硬件特定调优，缺少细粒度优化，潜在回归风险

关联脉络

- PR #21952 未知：PR body 中提及基于此 PR，但历史分析中未提供详细信息，可能为相关前置修复或依赖。
- PR #22323 [Lora] Lora quat info re-factor and support deepseekv3 mla lora: 共享 quant 标签，涉及量化相关优化，反映团队对量化模型的持续关注。