PR #19945 完整报告

sgl-project/sglang

[AMD] Tilelang sparse fwd for dsv32 mi355/mi300

合并时间：2026-03-24 17:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/19945

执行摘要

- 一句话：为 AMD MI300/MI355 GPU 优化稀疏注意力内核，提升性能。
- 推荐动作：建议精读此 PR 以学习 AMD GPU 内核优化策略，特别是条件内存分配和网格划分设计。关注安全漏洞修复的实现细节，以及性能基准测试方法。

功能与动机

根据 PR body，动机是“Enable the faster/new tilelang kernel on MI300”和“Improve longer-context kernel performance on MI355”，旨在通过优化内核提升 AMD GPU 上的推理速度，特别是针对 DeepSeek-V3.2 等模型的长上下文场景。

实现拆解

实现主要集中在修改 sparse_mla_fwd_decode_partial 函数：1）添加 inner_iter 参数，将网格计算改为 N_GROUPS = topk // (block_I * inner_iter)，以处理多个 KV tile；2）根据 inner_iter 值条件分配 Q 缓冲区（inner_iter == 1 时在共享内存，inner_iter > 1 时在 fragment），以重用 Q 提升占用率；3）添加安全索引逻辑处理负索引，避免越界读取；4）保留 v1 内核用于文档目的，作为设计权衡。

关键文件：

- python/sglang/srt/layers/attention/nsa/tilelang_kernel.py（模块 attention/nsa）：这是唯一修改的文件，包含稀疏注意力内核的核心优化，直接影响 AMD GPU 性能。

关键符号：sparse_mla_fwd_decode_partial

评论区精华

Review 讨论核心包括：1）gemini-code-assist[bot] 指出安全漏洞，索引未验证可能导致 GPU 越界读取，作者回应添加安全索引逻辑；2）gemini-code-assist[bot] 识别冗余计算循环，作者移除以优化性能；3）HaiShaw 建议回滚 v1 内核删除以保留文档用途，作者执行。决策包括安全修复和性能优化，所有疑虑已解决。

- 安全漏洞：索引未验证导致越界读取 (security): 作者添加 T.if_then_else 逻辑处理负索引，漏洞已修复。
- 冗余计算循环优化 (performance): 作者移除冗余循环，代码简化并优化。
- 保留 v1 内核用于文档目的 (design): v1 内核被保留，设计决策基于维护和文档考虑。

风险与影响

- 风险：技术风险：1）安全漏洞修复可能引入错误，如果安全索引逻辑不正确；2）内核参数变更（如 inner_iter）可能影响现有配置的兼容性，但添加了断言验证；3）性能调优针对特定硬件（MI300/MI355），在其他环境可能未充分测试。风险局部于单个文件，回归测试通过基准和准确性验证。
- 影响：影响范围：用户端，AMD GPU 用户将体验到推理速度提升（MI300 高达 2 倍加速）；系统端，优化资源利用率，尤其长上下文场景；团队端，展示硬件特定内核优化模式，促进类似工作。影响程度中等，针对性强但限于稀疏注意力模块。
- 风险标记：核心路径变更，安全漏洞修复，性能调优

关联脉络

- PR #21188 [AMD] Add fused GemmaRMSNorm forward_hip to use aiter/vllm kernels for qwen3.5: 同属 AMD GPU 性能优化，涉及内核代码和 JIT-kernel 标签，展示硬件特定调优趋势。
- PR #20438 [Perf] Overlap NSA-CP key all-gather with query computation for DeepSeek-V3.2: 涉及 NSA（稀疏注意力）性能优化，关联相同模块，揭示注意力内核演进方向。