PR #22791 完整报告

sgl-project/sglang

[MoE] Add LFM2 MoE tuning support + tuned configs for H100/B200/MI325X

合并时间：2026-04-22 09:32

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22791

执行摘要

- 一句话：为 LFM2 MoE 模型添加 Triton 内核调优支持及针对 H100/B200/MI325X 的优化配置。
- 推荐动作：建议技术管理者和工程师精读此 PR，重点关注：
  - common_utils.py 中的模型配置提取逻辑，了解如何适配新 MoE 架构。
  - JSON 配置文件的结构和参数选择，学习 Triton 内核调优的最佳实践。
  - 讨论中提到的代码重复问题，可作为重构机会以提升代码库可维护性。

功能与动机

LFM2 MoE 模型使用 num_experts / moe_intermediate_size 配置键，而默认的 Mixtral 回退逻辑期望 num_local_experts / intermediate_size，导致调优脚本崩溃或生成错误内核形状。没有调优配置时，融合 MoE Triton 内核会回退到通用默认值，远非 LFM2 专家形状的最优解，因此需要添加支持并生成优化配置以提升性能。

实现拆解

1. 扩展模型支持入口：在 benchmark/kernels/fused_moe_triton/common_utils.py 的 get_model_config 函数中添加针对 Lfm2MoeForCausalLM 架构的分支，正确读取 num_experts 和 moe_intermediate_size 键，确保调优脚本能处理 LFM2 模型配置。
2. 生成硬件特定配置：新增 24 个 JSON 配置文件，分别存放在 python/sglang/srt/layers/moe/fused_moe_triton/configs/triton_3_5_1/（针对 NVIDIA H100 和 B200）和 triton_3_6_0/（针对 AMD MI325X）目录下。这些文件覆盖了 LFM2-8B-A1B（E=32）和 LFM2-24B-A2B（E=64）模型在 TP=1,2,4,8 时的不同批次大小（1 到 8192），包含优化后的 Triton 内核参数如 BLOCK_SIZE_M、num_warps 等。
3. 版本管理：配置文件按 Triton 版本拆分，triton_3_5_1 对应 NVIDIA 环境（默认 PyPI 安装），triton_3_6_0 对应 AMD 容器环境，以应对不同版本下的内核性能差异。
4. 配套调整：提交中包括修复 JSON 文件结尾换行符的提交，确保文件格式一致性；没有直接添加测试文件，但 PR body 提及了端到端验证和数值正确性检查。

关键文件：

- benchmark/kernels/fused_moe_triton/common_utils.py（模块 基准测试；类别 source；类型 core-logic；符号 get_model_config）：核心入口文件，扩展了对 LFM2 MoE 模型的

支持，确保调优脚本能正确读取模型配置键。

- python/sglang/srt/layers/moe/fused_moe_triton/configs/triton_3_5_1/E=32,N=1792,device_name=NVIDIA_H100_80GB_HBM3.json（模块 MoE 层；类别 config；类型 configuration）：关键配置文件之一，定义了 LFM2-8B-A1B 模型在 TP=1 时针对 NVIDIA H100 的优化 Triton 内核参数，直接影响性能。
- python/sglang/srt/layers/moe/fused_moe_triton/configs/triton_3_6_0/E=32,N=1792,device_name=AMD_Instinct_MI325X.json（模块 MoE 层；类别 config；类型 configuration）：针对 AMD MI325X 的关键配置文件，确保在 Triton 3.6.0 环境下 LFM2 模型能获得优化性能，特别是支持 LoRA 服务。

关键符号：get_model_config

关键源码片段

benchmark/kernels/fused_moe_triton/common_utils.py

核心入口文件，扩展了对 LFM2 MoE 模型的支持，确保调优脚本能正确读取模型配置键。

```
def get_model_config(model_name, ep_size=1):
    # 从模型配置中提取 MoE 相关参数，支持多种架构
    config = AutoConfig.from_pretrained(model_name)
    architecture = config.architectures[0]

    if architecture == "MixtralForCausalLM":
        E = config.num_local_experts // ep_size
        topk = config.num_experts_per_tok
        intermediate_size = config.intermediate_size
    elif architecture == "Qwen2MoeForCausalLM":
        E = config.num_experts // ep_size
        topk = config.num_experts_per_tok
        intermediate_size = config.moe_intermediate_size
    # 添加 LFM2 支持分支
    elif architecture == "Lfm2MoeForCausalLM":
        E = config.num_experts // ep_size # 使用 num_experts 键
        topk = config.num_experts_per_tok # 专家数量 per token
        intermediate_size = config.moe_intermediate_size # 使用 moe_intermediate_size 键
    else:
        raise ValueError(f"Unsupported architecture: {architecture}")

    return E, topk, intermediate_size
```

评论区精华

review 中主要关注两个问题：

- 代码重复：gemini-code-assist[bot] 指出 common_utils.py 中添加的 elif 块与其他架构（如 BailingMoEForCausalLM）逻辑相同，建议重构以减少重复，提升可维护性。讨论未明确是否采纳，但 PR 已合并，可能被接受或暂缓处理。

- 文件格式：同一审核者发现新增的 AMD JSON 配置文件缺少结尾换行符，可能引发工具兼容性问题；后续提交 5d49f5479c1f8576cba68ab1206fac94223c2cdc 修复了此问题，体现了对代码风格的重视。
- 代码重复问题 (design): 讨论未明确是否采纳重构建议，但 PR 已合并，可能暂时接受重复以快速交付功能。
- 文件格式一致性 (style): 后续提交修复了此问题，添加了结尾换行符，确保文件格式符合规范。

风险与影响

- 风险：
  1. 配置版本依赖风险：配置文件针对特定 Triton 版本（3.5.1 和 3.6.0）调优，未来 Triton 升级可能导致性能下降或需要重新调优，增加维护负担。
  2. 兼容性风险：新增模型支持可能引入未覆盖的边缘情况，例如其他 LFM2 变体模型配置差异，但 PR body 中已验证了常见模型。
  3. 测试覆盖不足：本次变更未添加单元测试，依赖已有的端到端验证；若内核参数配置错误，可能影响推理正确性或性能。
  4. 硬件特定优化风险：配置文件针对特定 GPU（如 H100、B200、MI325X），在其他硬件上可能无法加载或性能不佳，但系统应回退到默认配置。
- 影响：
  1. 用户影响：LFM2 模型用户将获得显著性能提升，在 NVIDIA 硬件上吞吐量最高增加 47%，改善高并发场景体验；AMD 用户在使用 --moe-runner-backend triton 运行 LoRA 服务时，性能接近 aiter CK-MoE 后端，填补了关键功能缺口。
  2. 系统影响：新增配置文件和代码扩展了 MoE 模块的功能范围，但未改动核心调度或缓存逻辑，对系统其他部分影响有限；配置文件按版本管理有助于未来升级。
  3. 团队影响：提供了硬件特定调优的范例，可作为未来模型支持的参考；但代码重复问题提示需要关注代码结构优化。 - 风险标记：配置版本依赖，缺少测试覆盖

关联脉络

- PR #22911 [perf] support return_routed_experts with overlap scheduling: 同样涉及 MoE 性能优化和调度改进，可对比学习 MoE 模块的演进。
- PR #22933 [CPU] expand the interface of shared_expert without scaling factor: 涉及 MoE 内核扩展和重构，与本 PR 的模型支持增强相关。