PR #24004 完整报告

sgl-project/sglang

fix(moe): relocate orphan tuned configs after #23019

合并时间：2026-04-29 17:00

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24004

执行摘要

PR #24004 修复了 MoE 调优配置未被运行时加载的 bug：由于 #23019 移动了配置加载器目录，后续 PR 将 33 个调优 JSON 文件误放到旧路径，导致它们从未被读取（回退到默认值）。本次仅通过 git-mv 将配置移动到正确路径，并更新了 tuning README 中的路径指向，无代码或内容变更。修复低风险，恢复预期性能。

功能与动机

PR #23019 将 MoE 配置加载器和 configs/ 目录从 fused_moe_triton/ 移至 moe_runner/triton_utils/。之后，#22791（LFM2，24 个配置）和 #23533（Hy3 preview，9 个配置）在提交时未注意路径变更，将调优 JSON 添加到了旧路径。运行时加载器通过 os.path.dirname(os.path.realpath(__file__)) 确定自身位置，然后搜索 configs/ 子目录，因此旧路径下的配置从未被加载，实际回退到 get_default_config()。这些配置本身经过正确调优，只是文件位置不对。

实现拆解

1. 移动 33 个 JSON 配置文件：使用 git mv 将 fused_moe_triton/configs/triton_3_5_1/ 和 triton_3_6_0/ 下的全部 33 个文件原样移动到 moe_runner/triton_utils/configs/ 对应子目录。覆盖 E=32,64,192，设备包括 H100、B200、MI325X、H20、H20-3e 等。
2. 更新文档路径：在 benchmark/kernels/fused_moe_triton/README.md 中将配置目录引用更新为新的 moe_runner/triton_utils/configs/ 路径。
3. 无代码变更：不涉及 Python/CUDA/Triton 代码，仅文件位置修复。

由于变更仅涉及 JSON 文件移动，内容完全一致，此处展示其中一个配置的完整内容，并注明路径变化：

```
// 路径由 fused_moe_triton/configs/triton_3_5_1/ 移至
// moe_runner/triton_utils/configs/triton_3_5_1/E=192,N=192,device_name=NVIDIA_B200,dtype=fp8_w8a8.json
{
    "1": {
        "BLOCK_SIZE_M": 16, // 每块行数
        "BLOCK_SIZE_N": 32,
        "BLOCK_SIZE_K": 128, // 每块 K 维度
        "GROUP_SIZE_M": 64, // 行组大小
        "num_warps": 4, // 线程束数
```

```
        "num_stages": 4 // 流水线级数
    },
    "2": {
        "BLOCK_SIZE_M": 16,
        "BLOCK_SIZE_N": 64,
        "BLOCK_SIZE_K": 128,
        "GROUP_SIZE_M": 1,
        "num_warps": 4,
        "num_stages": 4
    },
    ... // 更多专家数（4,8,16,24,32,48,64）配置
}
```

评论区精华

唯一审查者 Qiaolin-Yu 评论 nice catch，认可此修复。无其他讨论。

风险与影响

- 风险：极低。文件内容未变动，仅改变位置。运行时加载器基于文件自身所在目录搜索 configs，移动后即可正常加载。若存在符号链接或自定义路径依赖，需额外确认。
- 影响：对于使用 LFM2 或 Hy3 preview 模型的用户，Triton 内核将重新加载调优后的配置，提升性能；其他用户无影响。开发者日后添加新配置需注意放置到 moe_runner/triton_utils/configs/ 下。

关联脉络

- 23019 引发路径变更，是本次修复的根因。
- 22791 和 #23533 错误地将配置放入旧路径，是本次修复的直接目标。
- 此 PR 清理了遗留问题，并使后续 CI 中的 MoE benchmark 能正确反映调优效果。