# PR #31793 完整报告

sgl-project/sglang

[Fix][AMD] Qwen3.5 MoE: disable global-slot shared-expert fusion under per-rank EP backends (MoRI + dp-attention init crash)

合并时间：2026-07-26 14:59

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/31793

## 执行摘要

- 一句话：禁用 Qwen3.5 在 MoRI+dp-attention 下的融合专家
- 推荐动作：值得精读。该 PR 清晰展示了模型架构（Qwen 全局槽位设计）与通信后端（MoRI per-rank 槽位设计）之间的不兼容问题，以及如何在模型代码层优雅地检测和回退。设计决策（保留独立 MLP 路径而非强制适配 per-rank 槽位）是务实的权衡，适合作为跨模块冲突处理的参考案例。

## 功能与动机

修复 Issue #31336：在 AMD MI355X 上使用 Qwen3.5-397B-A17B-FP8 模型时，同时开启 --moe-a2a-backend mori 和 --enable-dp-attention 导致模型初始化崩溃，断言 (num_experts - num_shared_slots) % moe_ep_size == 0 失败。根本原因是 Qwen 的共享专家融合采用全局槽位设计（在 num_experts 位置附加一个共享专家，并配有 per-token gate），而 MoRI 等 per-rank EP 后端要求每个 EP rank 有一个独立的共享槽位，两者布局不兼容。

## 实现拆解

1. 检测不兼容组合：在 Qwen2MoeSparseMoeBlock.__init__ 中，启用共享专家融合（self.enable_shared_expert_fusion = True）之后，增加条件判断：如果 uses_per_rank_fused_shared_slots() 返回 True（即后端为 DeepEP/MoRI 等 per-rank 共享槽位后端）且 get_parallel().moe_ep_size > 1，则自动禁用融合。
2. 日志警告：通过 logger.warning_once 打印清晰说明，指出全局槽位设计不兼容，并回退到独立 shared-expert MLP 路径。
3. 导入新工具函数：在文件头部导入 uses_per_rank_fused_shared_slots 函数，该函数定义在 sglang.srt.layers.moe.utils 中，用于检查当前使用的 EP 后端是否需要 per-rank 共享槽位。
4. 无其他文件修改：仅修改 python/sglang/srt/models/qwen2_moe.py 一个文件，增加 15 行代码，无删除。该改动对非 per-rank 后端（如单 EP rank 或非 deepep 类后端）无影响，保持原有融合行为。

关键文件：

- python/sglang/srt/models/qwen2_moe.py（模块 模型层；类别 source；类型 core-logic；符号 Qwen2MoeSparseMoeBlock.init）：唯一修改的文件，在

Qwen2MoeSparseMoeBlock 初始化中增加了对 per-rank EP 后端的检测，当检测到不兼容组合时禁用共享专家融合并回退到独立 MLP 路径。

关键符号：get_num_shared_experts, can_fuse_shared_expert, uses_per_rank_fused_shared_slots, Qwen2MoeSparseMoeBlock.init

## 关键源码片段

### python/sglang/srt/models/qwen2_moe.py

唯一修改的文件，在 Qwen2MoeSparseMoeBlock 初始化中增加了对 per-rank EP 后端的检测，当检测到不兼容组合时禁用共享专家融合并回退到独立 MLP 路径。

```
# 文件：python/sglang/srt/models/qwen2_moe.py
# 在 Qwen2MoeSparseMoeBlock.__init__ 中，启用融合后增加守卫逻辑

if support_shared_expert_fusion and (
    _use_aiter or (_is_cuda and enable_cuda_shared_expert_fusion)
):
    self.enable_shared_expert_fusion = (
        self.num_shared_experts > 0
        and can_fuse_shared_expert(config, quant_config)
    )

# 新增守卫：检测 per-rank 共享槽位后端 (DeepEP/MoRI) 且 EP size > 1
# Qwen 的共享专家融合使用全局槽位（单个 global id) + per-token gate,
# 这与 per-rank EP 后端的连续块掩码布局不兼容，会导致专家路由错误或初始化崩溃。
if (
    self.enable_shared_expert_fusion
    and uses_per_rank_fused_shared_slots() # 检查当前 EP 后端是否需要 per-rank 槽位
    and get_parallel().moe_ep_size > 1
):
    logger.warning_once(
        "Disabling Qwen shared-expert fusion: it uses a single global "
        "shared slot with a per-token gate, which is incompatible with "
        "per-rank EP shared-slot backends (e.g. DeepEP/MoRI) at "
        "moe_ep_size=%d. Using the separate shared-expert MLP instead.",
        get_parallel().moe_ep_size,
    )
    self.enable_shared_expert_fusion = False

if self.enable_shared_expert_fusion:
    self.num_fused_shared_experts = self.num_shared_experts
# 后续逻辑不变：当融合禁用时，num_fused_shared_experts 保持为 0,
# 因此 TopK 配置使用原始 num_experts，共享专家由独立 MLP 计算。
```

## 评论区精华

- gemini-code-assist[bot] 提议：建议将 get_parallel().moe_ep_size 存入局部变量以避免重复调用，提升可读性。作者 yichiche 回应表示保持当前形式，因为该代码在初始化时只执

行一次，重复调用开销可忽略，且内联写法使条件更加直观。

- HaiShaw 提问：要求作者说明问题是 AMD 特有还是也影响 NVIDIA，并建议在代码中增加注释。yichiche 回应确认问题是 Qwen 模型本身的设计缺陷，不限于 AMD，在 NVIDIA 上使用 per-rank EP 后端时同样会触发。
- HaiShaw 进一步要求：提供在 NVIDIA 上复现的配置，以帮助澄清 MoRI 描述和通用代码变更的混合。但看 diff 中未看到后续更新，可能已离线确认。
  - 是否应将 get_parallel().moe_ep_size 存入局部变量 (style): 作者 yichiche 表示保持当前形式，因为初始化只执行一次，重复调用开销可忽略，且内联写法更直观。
  - 问题是否仅限 AMD，代码变更是否覆盖 NVIDIA (question): yichiche 确认问题是 Qwen 模型设计导致，不限于 AMD，在 NVIDIA 上使用同类后端时同样会触发。
  - 提供 NVIDIA 上的复现配置 (question): 未看到后续回复，可能已离线确认。

## 风险与影响

- 风险：低风险。改动仅为在初始化时增加一个条件判断和日志，不影响正常路径。当检测到不兼容组合时，禁用融合并回退到已有且经过验证的独立 shared-expert MLP 路径，该路径在非融合场景下已稳定运行。对于 moe_ep_size == 1 或非 per-rank 后端，行为完全不变。潜在风险是依赖 uses_per_rank_fused_shared_slots() 函数的正确性，该函数定义在 moe/utils.py 中，若后续修改其逻辑可能影响此守卫。
- 影响：
  - 用户影响：修复了 AMD MI355X 用户使用 Qwen3.5-397B 模型时启用 MoRI + dp-attention 的初始化崩溃，使得该配置可以正常运行（尽管 MoRI 的 all-to-all 正确性问题仍需单独跟踪 #21886）。对于不使用 dp-attention 或使用其他 EP 后端的用户无影响。
  - 系统影响：仅影响模型初始化阶段的决策逻辑，不改变运行时路径。
  - 团队影响：为后续 MoRI 对 Qwen 模型的完全支持扫清了初始化障碍，提供了清晰的错误信息和回退机制。
  - 风险标记：依赖外部函数 uses_per_rank_fused_shared_slots 的正确性，未添加直接测试

## 关联脉络

- PR #31336 [Bug] [AMD] Qwen3.5-397B-A17B-FP8 + MoRI + dp-attention crashes at init: 该 PR 直接修复的 issue，详细描述了崩溃的根因和复现步骤。
- PR #21886 [bug] ROCm DI MoRI for non-standard attention doesn't work: 关联的更大范围 MoRI 兼容性问题跟踪 issue，本 PR 是解决 Qwen 模型初始化崩溃的必要前提，但 MoRI 的 all-to-all 正确性问题仍在跟踪中。