PR #28378 完整报告

sgl-project/sglang

[AMD] Fix Always mask padded topk_ids on HIP to prevent garbage MoE routing (DeepSeek-R1-MXFP4 accuracy regression)

合并时间：2026-06-18 14:39

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28378

执行摘要

- 一句话：修复 AMD HIP 上 MoE topk 填充区域掩码被错误跳过导致精度崩溃
- 推荐动作：此 PR 值得重点关注，尤其是：
  1. 条件逻辑正确放置的重要性：一个看似无害的移动（将 mask 放入 EPLB 条件）足以摧毁模型的端到端精度，凸显了在复杂代码中对条件守卫进行严格审查的必要性。
  2. 回归修复的范式：PR 以最小 diff（+9/-8）解决了重大回归，并附有清晰的动机、根因分析和交叉验证，是高质量 bugfix 的范例。
  3. 跨 PR 依赖管理：作者在 PR body 中明确标识了 MTP 失败为独立问题（#28410），避免了归因混淆。

功能与动机

修复 AMD MI35X 上 DeepSeek-R1-MXFP4 的精度回归：CI 测试 test_deepseek_r1_mxfp4_8gpu.py 报告 AssertionError: 0.1379833206974981 not greater than 0.94，精度从 >0.94 降至 ~0.09–0.14。根因是 #28188 将 padded region 掩码错误地置于 EPLB 条件守卫内，导致非 EPLB 模型无法触发掩码，padded token 的 topk_ids 保留越界值。

实现拆解

修改仅涉及 python/sglang/srt/layers/moe/topk.py 中的 _post_process_topk_ids 函数，具体步骤如下：

1. 将 padded region 掩码无条件移到 HIP 分支顶部：_mask_topk_ids_padded_region(topk_ids, num_token_non_padded, fill_value=0) 从 if _eplb_remap_enabled(): 块内移出，放置在该块之前，确保无论 EPLB 是否启用，所有 padded token 的 expert id 都被填充为 0（合法值）。
2. 保留 EPLB 逻辑：topk_ids_logical_to_physical 仍包裹在 if _eplb_remap_enabled(): 内，因为该 remap 仅在有真实 expert 映射时才有意义。
3. 更新注释：在代码中添加详细注释说明掩码的必要性（aiter MoE kernel 无法处理 -1，0 是安全回退）以及历史回归原因。
4. 测试验证：通过 CI（Run #27598431399）确认精度恢复（0.950）、速度正常（152.01 token/s），且 MTP 测试的精度子项通过（0.950/0.960）。单独的 MTP 接受长

度退化被确定为 #28410 导致的独立问题，不影响本 PR 的正确性。

关键文件：

- python/sglang/srt/layers/moe/topk.py（模块 MoE 层；类别 source；类型 core-logic；符号 _post_process_topk_ids, _mask_topk_ids_padded_region）：唯一修改的文件，包含 MoE routing 的后处理逻辑 _post_process_topk_ids，负责将 padded token 的 topk_ids 和 topk_weights 清零。本 PR 调整了 _mask_topk_ids_padded_region 的调用位置，修复了 #28188 引入的回归。

关键符号：_post_process_topk_ids, _mask_topk_ids_padded_region

关键源码片段

python/sglang/srt/layers/moe/topk.py

唯一修改的文件，包含 MoE routing 的后处理逻辑 _post_process_topk_ids，负责将 padded token 的 topk_ids 和 topk_weights 清零。本 PR 调整了 _mask_topk_ids_padded_region 的调用位置，修复了 #28188 引入的回归。

```
# python/sglang/srt/layers/moe/topk.py (head branch)
# 位于 _post_process_topk_ids 函数的 HIP 分支中
elif _is_hip:
    # On AMD HIP the aiter MoE kernels do not handle topk_ids=-1 safely
    # (negative indices cause illegal memory access). Always fill the padded
    # region with 0 so every kernel sees a valid in-range expert id.
    # Routing weights for padded tokens are zeroed below so their
    # contribution to the hidden state is still zero regardless of the id.
    # Regression: skipping this mask when EPLB is disabled caused garbage
    # MoE routing for models like DeepSeek-R1-MXFP4 (accuracy ~0.09 vs 0.94+).
    _mask_topk_ids_padded_region(topk_ids, num_token_non_padded, fill_value=0)
    # The logical->physical remap is only meaningful when a real
    # expert-location mapping exists. With a trivial placement and EPLB off
    # the map is identity so the remap can be skipped safely.
    if _eplb_remap_enabled():
        topk_ids = topk_ids_logical_to_physical(
            topk_ids, expert_location_dispatch_info
        )
    # On AMD HIP the aiter MoE kernels do not handle topk_ids=-1 safely, so
    # padded tokens are neutralized by zeroing their routing weights.
    _zero_topk_weights_padded_region(topk_weights, num_token_non_padded)
```

评论区精华

PR 的 review 评论数为 0，但有两条 issue 评论值得关注：

- 关于 MTP 测试失败：CI 报告者 @bingxche 指出 TestDeepseekR1MXFP4MTP 中 avg_spec_accept_length 过低（1.45 < 2.04）。作者 @kangwangamd 回应称，该问题是 #28043 引入的预存跨流竞争，与本 PR 无关，且已在 #28410 中修复。经在空闲机器上联合验证 #28378 + #28410 后确认 MTP 精度正常。

- 结论：团队接受说明，未引发其他讨论。
  - MTP avg_spec_accept_length 过低是否为本 PR 引入 (correctness): 确认是独立问题，本 PR 不负责修复，将在 #28410 中处理。

风险与影响

- 风险：此 PR 风险极低：
  - 仅移动一行掩码调用，调整条件逻辑，无新增代码路径。
  - 已在真实 AMD MI355X（gfx950）上通过端到端测试，精度和速度均回归正常。
  - 更改不涉及其他平台（如 CUDA）或非 AMD 模型。
  - 潜在风险：若某些 HIP MoE kernel 依赖未掩码的 padded token 行为（理论上不应有），但现有测试已覆盖常规配置。
- 影响：正面影响：
  - 恢复 AMD MI35X 上 DeepSeek-R1-MXFP4 的 GSM8K 精度至 0.950（原 >0.94），直接解除模型可用性阻塞。
  - 影响范围仅限于使用 AMD HIP MoE + MXFP4 且不启用 EPLB 的部署场景。
  - 无性能退化（batched speed 维持 152.01 token/s）。
  - 对非 AMD 平台或启用 EPLB 的模型无影响（逻辑保持不变）。
  - 文档无变更要求。
  - 风险标记：仅 AMD 平台影响，无新增单元测试

关联脉络

- PR #28188 [AMD] Skip eplb bookkeeping and topk remap when EPLB is not in use on mori-ep / HIP (#22985): 引入回归的 PR：将 _mask_topk_ids_padded_region 错误地放入 _eplb_remap_enabled() 条件内，导致本 PR 修复的问题。
- PR #28410 [Bugfix] Fix MTP acceptance regression on plan stream by moving int64 cast before plan stream context: 与 MTP 测试失败相关：本 PR body 指出 MTP 接受长度退化由 #28043 引入，#28410 正是修复该问题的 PR，且作者验证了两者联合后的正确性。
- PR #28043 [Bugfix][DeepSeek-V4] Fix Spec V2 Draft Input ID Dtype for DP Collectives: 引入了重叠计划流中的数据类型竞争，导致 MTP 接受长度退化，本 PR 中作者明确标记该问题为独立。