PR #27528 完整报告

sgl-project/sglang

Fix GPT-OSS MXFP4 hidden size reshape on SM10X

合并时间：2026-06-09 04:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27528

执行摘要

- 一句话：修复 GPT-OSS MXFP4 在 SM10X 上的 hidden size 对齐问题
- 推荐动作：该 PR 值得合并，是一个定位清晰、修复精准的 bugfix。虽然没有新增测试，但手动验证已覆盖关键场景。建议合并后关注是否有其他使用 self.experts.hidden_size 的模型也存在类似问题。

功能与动机

在 SM10X GPU 上使用 MXFP4 量化运行 GPT-OSS 20B 模型时，服务器 warmup 因 RuntimeError 'shape [4096, 3072] is invalid for input of size 11796480' 而崩溃。该 regression 由 PR #27063 引入，该 PR 为 AMD 优化了 GPT-OSS 性能，但导致 MoE 块的 hidden size 使用了后端内部填充后的值而非原始配置值。

实现拆解

1. 在 __init__ 中保存原始 hidden size：在 GptOssSparseMoeBlock.__init__ 方法中，添加 self.hidden_size = config.hidden_size 一行，将模型的原始 hidden size（2880）保存在实例属性中。
2. 修改 forward_normal 中的 reference：将 forward_normal 方法中原本使用 self.experts.hidden_size 获取 unpadded hidden size 的代码，改为使用新保存的 self.hidden_size。这使得路由器和后续 reshape 始终基于模型原始的 hidden size，而不是被后端填充后的值。
3. 影响范围：仅修改了 python/sglang/srt/models/gpt_oss.py 文件中的 3 行代码（+2/-1），变更极小。修复后，服务器 warmup 成功，且 GPT-OSS evals（如 GPQA）可正常完成。

关键文件：

- python/sglang/srt/models/gpt_oss.py（模块 模型定义；类别 source；类型 data-contract；符号 GptOssSparseMoeBlock.init, GptOssSparseMoeBlock.forward_normal）：修复的核心文件：新增 self.hidden_size 属性保存原始 hidden size，并修改 forward_normal 中使用该值替代 self.experts.hidden_size，避免 MXFP4 后端的 padding 导致形状不匹配。

关键符号：GptOssSparseMoeBlock.init, GptOssSparseMoeBlock.forward_normal

关键源码片段

python/sglang/srt/models/gpt_oss.py

修复的核心文件：新增 self.hidden_size 属性保存原始 hidden size，并修改 forward_normal 中使用该值替代 self.experts.hidden_size，避免 MXFP4 后端的 padding 导致形状不匹配。

```
# 在 __init__ 中新增一行，保存原始 hidden size（例如 2880）
class GptOssSparseMoeBlock(nn.Module):
    def __init__(self, layer_id, config, quant_config=None, prefix=""):
        super().__init__()
        self.tp_size = get_tensor_model_parallel_world_size()
        self.layer_id = layer_id
        self.hidden_size = config.hidden_size # 新增：保存模型原始 hidden size，不受后端 padding 影响
        self.activation = config.hidden_act
        # ...

    def forward_normal(self, hidden_states, should_allreduce_fusion=False):
        # ...
        # 修复前：hidden_dim_unpadded = self.experts.hidden_size # 可能被后端 padding 为 3072
        # 修复后：self.hidden_size = config.hidden_size = 2880
        hidden_dim_unpadded = self.hidden_size
        is_prepadded = hidden_states.shape[-1] != hidden_dim_unpadded
        if is_prepadded:
            router_input = hidden_states[..., :hidden_dim_unpadded]
        else:
            router_input = hidden_states
        # ... 后续逻辑不变
```

评论区精华

PR 没有 review 评论，但 PR 作者在 body 中清晰描述了问题根源和验证过程。关联的 Issue #27063（AMD 优化 PR）引入了该 regression，但作者已确认修复。

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险低：仅修改了 hidden size 的 reference，且新保存的 self.hidden_size 直接来自 config.hidden_size，功能上等同于修复前（SM10X 外）模型一直使用的值。
  2. 其他硬件无影响：作者已验证 H200 上使用 triton_kernel 后端时没有 padding，因此不存在此问题。修复不会破坏该路径。
  3. 测试覆盖缺失：当前没有针对 SM10X+MXFP4 的自动化测试，此修复仅在手动验证下确认有效。
- 影响：

- 用户：在 SM10X（Blackwell）上使用 MXFP4 量化和 FlashInfer 后端运行 GPT-OSS 模型的用户可以正常启动服务器。
- 系统：修复了一个导致 warmup 失败的 regression，确保模型在 SM10X 上可用。
- 团队：无负面影响。
- 风险标记：缺少测试覆盖，核心路径变更

关联脉络

- PR #27063 [AMD] Optimize gpt-oss-120B performance: 该 PR 引入了 regression：在 MXFP4 后端中 pad hidden size，导致 self.experts.hidden_size 不再是原始值。#27528 修复此问题。
- PR #27289 [ROCm] dsv4: remove the redundant fp8 scale transpose-copy on decode: 另一个与 DeepSeek 模型和量化相关的性能优化 PR，但未直接关联此修复。