PR #27150 完整报告

sgl-project/sglang

Support Waterfill with dynamic EPLB

合并时间：2026-06-06 07:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27150

执行摘要

- 一句话：修复 Waterfill 与动态 EPLB 在 fused shared experts 下的兼容性
- 推荐动作：此 PR 是 DeepEP+EPLB 兼容路径的关键修复，维护者应快速合入。代码设计清晰（通过分离 recorder ID 避免统计污染），可作为处理类似混合专家 ID 空间的参考实现。

功能与动机

PR body 明确指出需要修复 fused shared expert MoE weight views 被在线 EPLB 更新错误修改，以及确保 EPLB 统计仅跟踪逻辑路由专家 ID。MMLU 测试显示不加修复时可能存在 NCCL 超时。

实现拆解

1. 修正权重视图 (python/sglang/srt/models/deepseek_v2.py)：在 get_moe_weights 中对返回的权重张量进行切片，只取前 num_local_experts - num_fused_shared_experts 行，避免 EPLB 在线更新影响到 fused shared expert 的权重。
2. 分离 recorder topk_ids (python/sglang/srt/layers/moe/topk.py)：修改 _post_process_topk_ids 返回值，新增 recorder_topk_ids。当使用 DeepEP 后端且 fused shared experts > 0 时，recorder_topk_ids 只包含经过 EPLB 重映射的 routed 列，排除后续 DeepEP 插入的 shared slot。其他情况下与 topk_ids 保持一致。
3. 变更调用点 (python/sglang/srt/layers/moe/topk.py:select_experts)：将 on_select_experts 的输入从原始的 topk_ids 改为新的 recorder_topk_ids，确保 EPLB 统计基于正确的逻辑路由 ID。
4. 单元测试 (test/registered/unit/eplb/test_deepep_waterfill_eplb.py)：新增 4 个测试方法，覆盖 fused shared expert 排除、保持完整形状、DeepEP 与非 DeepEP 后端下的 recorder ID 正确性。

关键文件：

- python/sglang/srt/layers/moe/topk.py（模块 路由逻辑；类别 source；类型 core-logic；符号 _post_process_topk_ids, select_experts）：核心逻辑变更：修改 _post_process_topk_ids 返回类型并分离 recorder_topk_ids，变更调用 select_experts 以使用正确的统计 ID。
- python/sglang/srt/models/deepseek_v2.py（模块 模型定义；类别 source；类型 data-contract；符号 DeepseekV2MoE.get_moe_weights）：权重视图修复：

get_moe_weights 切片排除 fused shared expert，避免 EPLB 在线更新破坏共享专家权重。

- test/registered/unit/eplb/test_deepep_waterfill_eplb.py（模块 测试；类别 test；类型 test-coverage；符号 _FakeExpertParam, TestDeepEPWaterfillEPLB, test_deepseek_moe_get_moe_weights_excludes_fused_shared_slot, test_deepseek_moe_get_moe_weights_keeps_full_shape_without_fusion）：新增完整单元测试套件，验证 get_moe_weights 切片正确性以及 recorder_topk_ids 在不同后端下的行为。

关键符号：DeepseekV2MoE.get_moe_weights, _post_process_topk_ids, select_experts

关键源码片段

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

核心逻辑变更：修改 _post_process_topk_ids 返回类型并分离 recorder_topk_ids，变更调用 select_experts 以使用正确的统计 ID。

```
# 关键变更：_post_process_topk_ids 新增 recorder_topk_ids 返回值
# 用于 EPLB 统计时排除 fused shared expert 的 slot

def _post_process_topk_ids(
    topk_ids: torch.Tensor,
    topk_weights: torch.Tensor,
    topk_config: TopKConfig,
    router_logits: torch.Tensor,
    layer_id: int,
    num_token_non_padded: Optional[torch.Tensor] = None,
    expert_location_dispatch_info: Optional[ExpertLocationDispatchInfo] = None,
) -> tuple[torch.Tensor, torch.Tensor, torch.Tensor]:
    # ... 省略无关部分
    # 声明 recorder 变量
    recorder_topk_ids = None
    if _is_cuda:
        if num_fused_shared_experts > 0 and is_deepep_class_backend():
            shared_cols = topk_ids[:, -num_fused_shared_experts:]
            routed_cols = topk_ids[:, :-num_fused_shared_experts]
            # EPLB 仅对 routed 列进行重映射
            routed_cols = _biased_grouped_topk_postprocess(
                routed_cols, expert_location_dispatch_info, num_token_non_padded
            )
            topk_ids = torch.cat([routed_cols, shared_cols], dim=-1)
            # recorder 只保留重映射后的 routed 列，排除 shared slot
            recorder_topk_ids = routed_cols
        else:
            topk_ids = _biased_grouped_topk_postprocess(
                topk_ids, expert_location_dispatch_info, num_token_non_padded
            )
    if recorder_topk_ids is None:
        recorder_topk_ids = topk_ids
```

```
    # ... 后续 DeepEP remap 仅修改 topk_ids，不影响 recorder_topk_ids
    return topk_ids, topk_weights, recorder_topk_ids

# 调用点调整：
def select_experts(...) -> StandardTopKOutput:
    # ...
    topk_ids, topk_weights, recorder_topk_ids = _post_process_topk_ids(...)
    # 使用 recorder_topk_ids 传入统计，避免 fused shared slot 污染负载均衡数据
    get_global_expert_distribution_recorder().on_select_experts(
        topk_ids=recorder_topk_ids
    )
```

python/sglang/srt/models/deepseek_v2.py

权重视图修复：get_moe_weights 切片排除 fused shared expert，避免 EPLB 在线更新破坏共享专家权重。

```
# DeepseekV2MoE.get_moe_weights 修改：只返回 routed expert 部分的权重
def get_moe_weights(self):
    # EPLB 仅重平衡物理路由专家，fused shared expert 必须保持不变
    num_local_experts_for_eplb = (
        self.experts.num_local_experts - self.num_fused_shared_experts
    )
    return [
        x.data[:num_local_experts_for_eplb] # 切片丢弃 fused shared 参数
        for name, x in self.experts.named_parameters()
        if name not in ["correction_bias"]
        and filter_moe_weight_param_global_expert(name, x, self.experts.num_local_experts)
    ]
```

test/registered/unit/eplb/test_deepep_waterfill_eplb.py

新增完整单元测试套件，验证 get_moe_weights 切片正确性以及 recorder_topk_ids 在不同后端下的行为。

```
# 测试 fused shared expert 情况下 get_moe_weights 返回切片后的权重
class TestDeepEPWaterfillEPLB(CustomTestCase):
    def test_deepseek_moe_get_moe_weights_excludes_fused_shared_slot(self):
        experts = _FakeExpertParam() # 5 个本地 expert，权重 shape (5,2)
        moe = SimpleNamespace(num_fused_shared_experts=1, experts=experts)
        shared_before = experts.weight.data[-1].clone() # 保存 fused shared 权重

        weights = DeepseekV2MoE.get_moe_weights(moe)
        # 应只返回 routed 部分：5-1=4 个 expert
        self.assertEqual(len(weights), 1)
        self.assertEqual(weights[0].shape, (4, 2))

        # 修改返回权重的最后一行为零，验证不影响原始 fused shared 权重
        weights[0][-1].zero_()
        self.assertTrue(torch.equal(experts.weight.data[-2], torch.zeros(2)))
        self.assertTrue(torch.equal(experts.weight.data[-1], shared_before))
```

```
    # ... 其他测试方法类似，验证 recorder_topk_ids 的正确性
```

评论区精华

仅有 gemini-code-assist[bot] 自动代码审查，无人工讨论。审阅者 ch-wan 直接批准，未提出异议。

- 代码审查自动评论 (other): 无需处理，bot 评论仅为信息性。

风险与影响

- 风险：
  1. 回归风险低：变更仅影响 DeepEP 后端 + fused shared experts + EPLB 的组合路径，非该路径的行为不受影响。
  2. 测试覆盖风险：测试全部在 CPU 上通过 mock 和 patching 执行，未在真实 GPU 上验证 DeepEP 调度流程，可能遗漏运行时内存 / 布局问题。
  3. 兼容性：_post_process_topk_ids 返回类型从 2 元组改为 3 元组，所有调用点均已更新（仅 select_experts 一处），无兼容断裂。
  4. 性能影响：增加了一次切片和广播操作，开销可忽略。 - 影响：直接使 DeepEP Waterfill 用户能够安全启用动态 EPLB，修复了在线权重更新可能导致 fused shared expert 偏移或统计错误的缺陷。影响范围限于使用 --enable-ep-moe 且开启 num_fused_shared_experts > 0 的 DeepSeek-V2 类模型。间接为未来 EPLB 统一统计逻辑奠定基础。 - 风险标记：Fused shared experts 路径，测试无 GPU 端集成验证，返回类型变更需确保调用点更新

关联脉络

- PR #27329 [LoRA] Experimental fast LoRA path with experimental_sgl_trtllm MoE backend for FP8 and NVFP4 models: 同为 MoE 性能相关 PR，涉及 fast LoRA 和 trtllm 后端，但与本 PR 无直接依赖关系。
- PR #27166 Reland "Support NextN = 2/4 in DSV32": 同为 DeepEP 后端改进，涉及 DSV32 和 DG 原生路径，与本 PR 在 DeepEP 动态 EPLB 上有潜在关联。