PR #25948 完整报告

sgl-project/sglang

[dsv4] support eplb

合并时间：2026-05-25 01:09

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25948

执行摘要

- 一句话：支持 DeepSeek-V4 EPLB 专家负载均衡
- 推荐动作：值得 merge。修复了里程碑模型 DSV4 的 EPLB 功能，改动小而精准。建议补充 EPLB 相关的集成测试以覆盖未来回归。

功能与动机

DeepSeek-V4 在启用 EPLB 时，由于缺少 eplb 上下文，layer_idx 为 None，导致 prefill 节点崩溃和 decode 节点统计不准确。PR body 明确描述："Without the context, the layer_idx used in eplb distribution gatherer is None which causes crash in prefill node and stats inaccuracy issue in decode node。"

实现拆解

1. 修改 deepseek_v4.py：在 DeepseekV4Model.forward 中，为每层 MoE 调用包裹 eplb 上下文管理器。新增导入 nullcontext 和 get_global_expert_distribution_recorder。使用条件表达式：如果 get_global_server_args().disable_piecewise_cuda_graph 为真（即启用 eplb），则使用 get_global_expert_distribution_recorder().with_current_layer(i) 作为上下文，否则使用 nullcontext()。这样每层 MoE 前都会调用 with_current_layer(i) 记录当前层索引。
2. 修改 hash_topk.py：在 HashTopK.forward 中，于 topk_ids_logical_to_physical 和 _mask_topk_ids_padded_region 之后，调用 get_global_expert_distribution_recorder().on_select_experts(topk_ids=topk_ids)，记录当前 token 选中的 expert IDs。新增导入 get_global_expert_distribution_recorder。
3. 无测试 / 配置 / 部署配套改动。变更集中于两处源码文件。

关键文件：

- python/sglang/srt/models/deepseek_v4.py（模块 模型模块；类别 source；类型 data-contract）：核心改动：在 MoE 层 forward 前注入 eplb 上下文，修复了 layer_idx 缺失导致的崩溃。
- python/sglang/srt/layers/moe/hash_topk.py（模块 MoE 模块；类别 source；类型 dependency-wiring）：在 hash topk 路由后记录选中的 expert IDs，支持 eplb 统计。

关键符号：DeepseekV4Model.forward, HashTopK.forward

关键源码片段

python/sglang/srt/models/deepseek_v4.py

核心改动：在 MoE 层 forward 前注入 eplb 上下文，修复了 layer_idx 缺失导致的崩溃。

```
# File: python/sglang/srt/models/deepseek_v4.py
# 在 DeepseekV4Model.forward 的循环中为每层 MoE 调用包裹 eplb 上下文

# 前置导入（新增部分）：
from contextlib import nullcontext
from sglang.srt.eplb.expert_distribution import (
    get_global_expert_distribution_recorder,
)

# 在 forward 方法内部的循环（原始代码第 1137 行附近）：
for i in range(self.start_layer, self.end_layer):
    layer = self.layers[i]
    # 根据 disable_piecewise_cuda_graph 决定上下文：
    # - 如果禁用 piecewise cuda graph（即启用 eplb），则使用 with_current_layer(i)
    # 来设置当前层索引，使得 expert distribution gatherer 能正确获取 layer_idx
    # - 否则，使用 nullcontext() 保持原行为
    ctx = (
        nullcontext()
        if not get_global_server_args().disable_piecewise_cuda_graph
        else get_global_expert_distribution_recorder().with_current_layer(i)
    )
    with ctx:
        hidden_states = layer(
            positions=positions,
            hidden_states=hidden_states,
            forward_batch=forward_batch,
            input_ids=input_ids,
            input_ids_global=input_ids_global,
        )
```

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

在 hash topk 路由后记录选中的 expert IDs，支持 eplb 统计。

```
# File: python/sglang/srt/layers/moe/hash_topk.py
# 在 HashTopK.forward 方法的末尾，逻辑 - 物理映射和掩码之后

# 前置导入（新增部分）：
from sglang.srt.eplb.expert_distribution import (
    get_global_expert_distribution_recorder,
)

# 在 forward 方法中，topk_ids 是当前 token 选中的逻辑 expert IDs（形状 [num_tokens, topk]）
topk_ids = topk_ids_logical_to_physical(topk_ids, expert_location_dispatch_info)
_mask_topk_ids_padded_region(topk_ids, num_token_non_padded)
```

```
# 记录当前 batch 选中的 expert IDs，供 eplb 统计和 rebalancing 使用
get_global_expert_distribution_recorder().on_select_experts(topk_ids=topk_ids)
topk_output = StandardTopKOutput(
    topk_weights=topk_weights, topk_ids=topk_ids, router_logits=router_logits
)
return topk_output
```

评论区精华

- 暂无高价值评论线程

风险与影响

- 风险：
  1. 回归风险：当 disable_piecewise_cuda_graph 为 False（默认）时，使用 nullcontext()，行为与原先一致，无回归风险。当为 True 时，引入 with_current_layer 和 on_select_experts 调用，若这些函数有 bug 可能影响 MoE 前向，但属于增量添加。
  2. 性能风险：with_current_layer(i) 和 on_select_experts 调用在热路径上，但调用开销极低（通常只是记录层索引或专家 ID），对整体吞吐影响可忽略。
  3. 缺少测试覆盖：PR 未添加单元测试或集成测试，回归风险依赖 review 信任。 - 影响：影响范围：仅影响 DeepSeek-V4 模型且启用 EPLB 的场景（--enable-eplb）。影响程度：修复了 prefill 崩溃和 decode 统计不准的关键问题，使 EPLB 在 DSV4 上正常工作。未启用 EPLB 的用户无影响。 - 风险标记：缺少测试覆盖

关联脉络

- 暂无明显关联 PR