PR #25571 完整报告

sgl-project/sglang

[Benchmark] Add SGLANG_SIMULATE_UNIFORM_EXPERTS for balanced expert routing with dummy weights

合并时间：2026-05-19 00:16

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25571

执行摘要

- 一句话：新增均匀专家路由环境变量用于基准测试
- 推荐动作：该 PR 值得快速合并。实现简洁、文档清晰，且对基准测试工作流程有明显提升。建议后续添加单元测试验证均匀分配的正确性。

功能与动机

使用 --load-format dummy 基准测试 MoE 模型时，随机门控权重会导致严重的专家不均衡，部分专家接收所有 token 而其他专家为空，进而触发 DeepEP 调度缓冲区溢出和 OOM。该标志强制均匀分配，代表最优的 token 路由情况，用于设定服务性能的上界。

实现拆解

1. 注册环境变量：在 python/sglang/srt/environ.py 的 Envs 类中新增 SGLANG_SIMULATE_UNIFORM_EXPERTS = EnvBool(False)，默认关闭，用户可通过设置该环境变量为 1 来启用。
2. 核心覆盖逻辑：在 python/sglang/srt/layers/moe/topk.py 的 *select_experts* 函数中，在原始门控计算完成后、调用 _post_process_topk_ids 之前，检查环境变量是否启用。若启用，则用确定性轮询分配覆盖 topk_ids 和 topk_weights：每个 token 选择 k 个专家，专家间隔为 num_experts // k，并引入随机 per-token 偏移以保证跨 token 负载均衡；所有选中专家赋予均匀权重 1/k。
3. 兼容性考虑：覆盖逻辑放在 _post_process_topk_ids 之前，确保后续的 EP 重映射、fused shared expert 处理和逻辑到物理 ID 转换仍能正常工作，且不改变 dispatch/combine tensor 形状，与 overlap schedule 兼容。
4. 无测试配套变更：本次改动未添加单元测试或集成测试。

关键文件：

- python/sglang/srt/layers/moe/topk.py（模块 MoE 路由；类别 source；类型 core-logic；符号 *select_experts*）：核心逻辑改动：在 *select_experts* 中插入均匀专家分配覆盖，是 PR 的主体。
- python/sglang/srt/environ.py（模块 环境配置；类别 source；类型 configuration）：注册新环境变量 SGLANG_SIMULATE_UNIFORM_EXPERTS，作为功能开关。

关键符号：*select_experts*

关键源码片段

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

核心逻辑改动：在 select_experts 中插入均匀专家分配覆盖，是 PR 的主体。

```
# python/sglang/srt/layers/moe/topk.py 中 select_experts 函数的关键片段
# 原始门控计算后，在 _post_process_topk_ids 之前插入覆盖逻辑
if envs.SGLANG_SIMULATE_UNIFORM_EXPERTS.get():
    # 基准测试专用：用均匀轮询专家分配覆盖门控输出
    # 避免 dummy/random 权重导致的专家不均衡。禁止在生产中使用。
    num_tokens, k = topk_ids.shape # (batch * seq_len, topk)
    num_experts = router_logits.shape[1] # 专家总数
    # 每个 token 生成随机偏移，保证跨 token 负载均衡
    offsets = torch.randint(0, num_experts, (num_tokens, 1), device=topk_ids.device)
    # 步长 = 专家总数 // k，确保专家均匀分布
    steps = torch.arange(k, device=topk_ids.device).unsqueeze(0)
    # 计算均匀分布的 expert_ids
    topk_ids = ((offsets + steps * (num_experts // k)) % num_experts).to(
        topk_ids.dtype
    )
    # 所有选中专家赋予均匀权重
    topk_weights = torch.ones_like(topk_weights) / k
```

评论区精华

此 PR 的 review 中无讨论线程。机器人评论因达到每日配额限制。维护者 ishandhanani 评论“This is very helpful”表示认可。

- 暂无高价值评论线程

风险与影响

- 风险：仅影响 MoE 模型的 select_experts 路径，且仅在设置环境变量时生效（默认关闭）。风险极低：
  - 无回归风险，因为覆盖逻辑是新增的且完全隔离，不影响默认路径。
  - 无性能风险，因为覆盖操作为 CPU 上的少量张量操作，开销可忽略。
  - 无安全风险，该特性仅用于基准测试。
  - 主要风险是用户可能在生产环境中意外启用该标志，但 PR 文档和代码注释已明确警告。
  - 影响：
  - 用户影响：为基准测试用户提供简单方法获得 MoE 模型的上界性能，无需手动修改代码。
  - 系统影响：无运行时影响，除非用户主动设置环境变量。
  - 团队影响：有助于性能分析和容量规划，但增加一个需要维护的测试用环境变量。
  - 影响程度：低。仅 2 文件 13 行新增，完全隔离的调试特性。
  - 风险标记：仅基准测试使用 ，默认关闭无回归风险

关联脉络

- 暂无明显关联 PR