

PR #29356 完整报告

sgl-project/sglang

fix(bench): pass DCP_RANK/DCP_WORLD_SIZE to set_mla_kv_buffer_kernel

合并时间: 2026-06-26 09:13

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29356>

执行摘要

- 一句话: 修复 MLA KV 缓存基准测试缺少 DCP 参数
- 推荐动作: 值得快速合并以恢复 CI 健康。此 PR 也提醒了未来在修改内核接口时需同步更新所有直接调用点。

功能与动机

修复 jit-kernel-benchmark CI 因 `set_mla_kv_buffer_kernel` 在 PR #14194 中新增 `DCP_RANK` 和 `DCP_WORLD_SIZE` 参数后, 基准测试未同步更新导致的 `TypeError` 失败。

实现拆解

在 `test/registered/jit/benchmark/bench_set_mla_kv_buffer.py` 的 `_triton_baseline` 函数中, 于调用 `sglang_triton_kernel` 时将 `DCP_RANK=0`, `DCP_WORLD_SIZE=1` 作为常量参数传入。这模拟了单进程基准测试下 DCP 禁用时的默认值, 与生产中 `get_attention_dcp_rank()/get_attention_dcp_world_size()` 的返回值一致。变更仅增加两行。

关键文件:

- `test/registered/jit/benchmark/bench_set_mla_kv_buffer.py` (模块 MLA KV 缓存; 类别 test; 类型 test-coverage; 符号 `_triton_baseline`): 修复基准测试中直接调用内核时缺少两个必需参数的 bug, 恢复 CI 健康。

关键符号: `_triton_baseline`

关键源码片段

`test/registered/jit/benchmark/bench_set_mla_kv_buffer.py`

修复基准测试中直接调用内核时缺少两个必需参数的 bug, 恢复 CI 健康。

```
def _triton_baseline(kv_buffer, loc, cache_k_nope, cache_k_rope):
    nope_dim = cache_k_nope.shape[-1]
    rope_dim = cache_k_rope.shape[-1]
    total_dim = nope_dim + rope_dim
    BLOCK = 128
    n_loc = loc.numel()
    grid = (n_loc, triton.cdiv(total_dim, BLOCK))
    pdl_kwargs = {"USE_GDC": True, "launch_pdl": True} if is_arch_support_pdl() else {}
```

```
# 调用 set_mla_kv_buffer_kernel, 传入 DCP 参数
# DCP_RANK=0, DCP_WORLD_SIZE=1 模拟单进程无 DCP 场景
sglang_triton_kernel[grid](
    kv_buffer,
    cache_k_nope,
    cache_k_rope,
    loc,
    kv_buffer.stride(0),
    cache_k_nope.stride(0),
    cache_k_rope.stride(0),
    nope_dim,
    rope_dim,
    BLOCK=BLOCK,
    DCP_RANK=0,
    DCP_WORLD_SIZE=1,
    **pdl_kwargs,
)
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险极低：仅修改基准测试的常量参数传递，不影响生产逻辑。单进程场景下 DCP_RANK=0, DCP_WORLD_SIZE=1 是正确且无害的。
- 影响：仅影响 jit-kernel-benchmark CI 的运行流程，修复后 CI 可正常执行该基准测试。不对用户或生产系统产生影响。
- 风险标记：暂无

关联脉络

- PR #14194 [feature] implement dcp for deepseek_v2: 本 PR 修复了 #14194 引入的接口变更导致的基准测试兼容性问题。