

PR #23528 完整报告

sgl-project/sglang

[CPU] remove RECORD_FUNCTION

合并时间: 2026-04-24 09:18

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23528>

执行摘要

本 PR 从 21 个 CPU C++ 源文件中移除了 `RECORD_FUNCTION` 宏调用，消除了由于 `torch.ops.sgl_kernel.xxx` 调度与宏内部记录导致的 Profiling 重复条目。变更纯属机械删除，不影响任何内核逻辑、数值输出或系统行为。

功能与动机

目前，CPU 内核通过 `torch.ops.sgl_kernel.xxx(...)` 注册，PyTorch 的调度层已经记录了内核调用。但每个内核实现内部还包含了 `RECORD_FUNCTION` 宏，导致在 Profiling 结果中同一个内核名称出现两次（如 PR body 中的示例）。这不仅使 Profiling 输出混乱，还轻微增加了记录开销。删除这些宏可以简化性能分析，并减少不必要的开销。

实现拆解

1. 定位文件：在 `sgl-kernel/csrc/cpu/` 目录下识别出所有包含 `RECORD_FUNCTION` 的 21 个源文件，包括 `extend.cpp`、`norm.cpp`、`mamba/fla.cpp`、`qkv_proj.cpp`、`decode.cpp`、`moe.cpp`、`topk.cpp`、`flash_attn.cpp`、`gemm_int8.cpp`、`gemm.cpp`、`gemm_fp8.cpp`、`interface.cpp` 等。
2. 删除宏：在每个内核函数体开头删除 `RECORD_FUNCTION(...)` 语句。例如，`extend_attention_cpu` 中原有：`RECORD_FUNCTION("sgl-kernel::extend_attention_cpu", std::vector<c10::IValue>({...}));` 删除后函数直接从 `CHECK_LAST_DIM_CONTIGUOUS_INPUT` 开始。所有其他代码保持不变。
3. 验证无副作用：仅涉及删除操作，不修改任何业务逻辑、校验或内存访问。

sgl-kernel/csrc/cpu/extend.cpp

核心 Attention 扩展内核，删除了 `extend_attention_cpu` 中的 `RECORD_FUNCTION`，是改动量最大的文件之一。

sgl-kernel/csrc/cpu/norm.cpp

包含多种归一化内核（`l2norm`，`rmsnorm`，`layernorm` 等），共删除了 7 个 `RECORD_FUNCTION`，数量最多。

sgl-kernel/csrc/cpu/mamba/fla.cpp

包含 Flash Linear Attention 相关内核（`chunk_gated_delta_rule` 等），删除了 3 个 `RECORD_FUNCTION`。

关键源码片段

sgl-kernel/csrc/cpu/extend.cpp

核心 Attention 扩展内核，删除了 extend_attention_cpu 中的 RECORD_FUNCTION，是改动量最大的文件之一。

```
// 删除前:
// RECORD_FUNCTION("sgl-kernel::extend_attention_cpu", ...);
// 删除后: 函数体直接开始校验输入
void extend_attention_cpu(
    at::Tensor& q_extend, at::Tensor& k_extend, at::Tensor& v_extend, at::Tensor& o_extend,
    at::Tensor& k_buffer, at::Tensor& v_buffer, at::Tensor& req_to_token,
    at::Tensor& req_pool_indices, at::Tensor& seq_lens, at::Tensor& extend_seq_lens,
    at::Tensor& extend_start_loc, int64_t max_len_extend, double sm_scale, double logit_cap) {
// RECORD_FUNCTION 已被移除
CHECK_LAST_DIM_CONTIGUOUS_INPUT(q_extend);
CHECK_INPUT(o_extend);
CHECK_LAST_DIM_CONTIGUOUS_INPUT(k_extend);
// ... 后续代码完全不变
}
```

sgl-kernel/csrc/cpu/norm.cpp

包含多种归一化内核 (l2norm, rmsnorm, layernorm 等)，共删除了 7 个 RECORD_FUNCTION，数量最多。

```
// 示例: rmsnorm_cpu
at::Tensor rmsnorm_cpu(at::Tensor& input, at::Tensor& weight, double eps) {
    // RECORD_FUNCTION("sgl-kernel::rmsnorm_cpu", {input, weight}); // 已删除
    CHECK_LAST_DIM_CONTIGUOUS_INPUT(input);
    CHECK_INPUT(weight);
    // ... 完全不变
}
```

sgl-kernel/csrc/cpu/mamba/fla.cpp

包含 Flash Linear Attention 相关内核 (chunk_gated_delta_rule 等)，删除了 3 个 RECORD_FUNCTION。

```
// chunk_gated_delta_rule_cpu
std::tuple<at::Tensor, at::Tensor> chunk_gated_delta_rule_cpu(
    const at::Tensor& query, const at::Tensor& key, const at::Tensor& value,
    const at::Tensor& g, const at::Tensor& beta, const at::Tensor& initial_state,
    bool output_final_state, const at::Tensor& cu_seqlens, bool head_first,
    bool use_qk_l2norm_in_kernel, double eps) {
// RECORD_FUNCTION(...) 已移除
TORCH_CHECK(head_first == false, "...");
// ...
}
```

评论区精华

无 Review 评论。PR body 中通过 Profiling 示例清晰说明了重复记录的问题，无需额外讨论。

风险与影响

- 风险：极低。宏仅用于 Profiling，不影响计算正确性。后续若需要调试，仍可通过外部工具捕获。
- 影响：受益于 Profiling 的用户将获得更清晰的输出。对普通推理无影响。团队应禁止在内核实现中再次添加 RECORD_FUNCTION。

关联脉络

该 PR 与近期其他性能优化（如 #23470 缓存派生张量）同属性能调优方向，但本 PR 更侧重于 Profiling 工具链的清理，是 SGLang 内核注册方式演进后的自然清理动作。