PR #22670 完整报告

sgl-project/sglang

Migrate Intel CPU cases to the test/registered.

合并时间：2026-05-12 13:27

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22670

执行摘要

- 一句话：迁移 Intel CPU 测试至统一注册目录
- 推荐动作：由于该 PR 已被 revert（#25044），建议在重新合并前必须修复所有已指出的问题。具体包括：修复 test_causal_conv1d 中的布尔张量生成错误；清理 test_binding 的类名；删除 test_bmm 中未使用的类；将 test_decode 的默认设备改为 "cpu"；修正 test_extend 中的拼写错误；更新 test_mamba 的类型提示；移除 test_rope 的冗余断言。同时建议强化测试代码审查及自动化 lint 检查。

功能与动机

原有的 CPU 测试分散在 test/srt/cpu/ 下，缺乏统一的注册和调度机制，导致 CI 覆盖不完整且难以扩展。该 PR 旨在将 CPU 测试迁移至 test/registered/cpu/，利用统一的测试注册框架提升 CI 覆盖的可维护性和触发的灵活性（PR body: 'improve CPU CI coverage and refine Xeon CI triggering'）。

实现拆解

1. 创建共享测试工具模块 test/registered/cpu/utils.py，封装精度阈值、参数化装饰器、量化辅助函数等，统一断言标准并减少代码重复。
2. 为每个 CPU 算子 / 内核路径创建独立的测试文件（共 22 个），置于 test/registered/cpu/ 下，涵盖 activation、bmm、causal_conv1d、decode/extend attention、flash_attn、gemm、mamba、mla、moe、norm、qkv_proj_with_rope、qwen3、rope、shared_expert、topk 等。
3. 在 CI 注册框架中通过 register_cpu_ci(est_time=10, suite="stage-b-test-cpu") 注册每个测试类，指定预估运行时间和所属套件。
4. 更新 CI 工作流文件（如 .github/workflows/slash-command-handler.yml），添加 /tag-run-cpu-ci-label 和 /tag-cpu-and-rerun-ci 两条新的斜杠命令，允许按需触发 CPU 测试套件。
5. 更新 CPU per-commit 流水线配置，将 stage-b-test-cpu 纳入 per-commit 自动运行套件。

关键文件：

- test/registered/cpu/utils.py（模块 共享工具；类别 test；类型 test-coverage；符号 parametrize, decorator, wrapper, SiluAndMul）：共享测试工具模块，提供参数化装饰器、精度阈值、量化辅助函数等，被所有新增测试文件依赖，是本次迁移的基础设施。

- test/registered/cpu/test_causal_conv1d.py（模块 因果卷积；类别 test；类型 test-coverage；符号 causal_conv1d_ref, causal_conv1d_update_ref, TestCausalConv1d, test_causal_conv1d）：包含严重的布尔张量生成错误，是代码审查中重点讨论的缺陷文件，反映了整体测试质量问题。
- test/registered/cpu/test_qkv_proj_with_rope.py（模块 QKV 投影；类别 test；类型 test-coverage；符号 layernorm, rotary_emb, native_torch, native_torch_int8）：覆盖了包含 RoPE 的 QKV 投影复杂算子，涉及多种量化模式（bf16、int8、fp8），是本次新增测试中复杂度最高的文件之一。
- .github/workflows/slash-command-handler.yml（模块 部署脚本；类别 infra；类型 configuration）：CI 工作流配置变更，新增 CPU 专用斜杠命令标签，是 CI 集成部分的核心改动。

关键符号：parametrize, per_token_quant_int8, native_w8a8_per_token_matmul, causal_conv1d_ref, causal_conv1d_update_ref, fused_moe, _bf16_gemm, test_norm, test_causal_conv1d, test_chunk_gated_delta_rule, test_bf16_moe, test_bf16_qkv_proj_with_rope

关键源码片段

test/registered/cpu/test_causal_conv1d.py

包含严重的布尔张量生成错误，是代码审查中重点讨论的缺陷文件，反映了整体测试质量问题。

```
class TestCausalConv1d(CustomTestCase):
    activation = "silu"

    @parametrize(
        batch=[1, 1024],
        dim=[96, 512],
        seqlen=[2, 36],
        width=[4],
        has_bias=[True, False],
        has_initial_state=[True, False],
    )
    def test_causal_conv1d(
        self,
        batch,
        dim,
        seqlen,
        width,
        has_bias,
        has_initial_state,
        dtype=torch.bfloat16,
        prepack=True,
    ):
        # ... 初始化 x, weight, bias ...

        # BUG: torch.randint 不支持 dtype=torch.bool，会抛出 RuntimeError；
```

```
    # 且随后的 .fill_(False) 使随机初始化失效。
    # 应改为：
    # has_initial_state_tensor = torch.randint(0, 2, (batch,)).to(torch.bool)
    has_initial_state_tensor = torch.randint(
        0, 2, (batch,), dtype=torch.bool # 此处会报错
    ).fill_(False) # 此覆盖无意义

    # ... 后续调用 causal_conv1d_fwd ...
```

评论区精华

- mingfeima询问 test/srt/cpu/ 下已有测试文件的处理方式，但未得到直接回应，PR 仍被合并。
- gemini-code-assist指出多个代码问题：test_causal_conv1d.py 中布尔张量生成错误（torch.randint 不支持 dtype=bool，且随后的 fill_(False) 使随机初始化失效）；test_activation.py 中在循环内冗余调用 set_global_server_args_for_scheduler；test_binding.py 中类名 TestGemm 应为 TestBinding（复制粘贴错误）；test_bmm.py 中未使用的 Mod 类；test_decode.py 中默认设备参数为 "cuda" 在 cpu 目录下不合理；test_extend.py 中变量名 redudant 拼写错误；test_mamba.py 中过时类型提示 torch.FloatTensor；test_rope.py 中冗余断言。
- mingfeima在 test_extend.py 中针对具体行给出了拼写修正建议。
- 这些讨论中除 test_extend.py 的拼写得到 maintainer 的直接建议外，其余大部分自动评审发现的问题未在合并前修复。
- 旧测试文件处理 (question): 部分旧测试保留，未完全清理，但 PR 仍然合并。
- 布尔张量生成错误 (correctness): 建议使用 .to(torch.bool) 并删除 fill_。
- 类名复制粘贴错误 (correctness): 建议重命名。
- 默认设备参数误导 (design): 建议改为 "cpu"。
- 拼写错误 (style): mingfeima 直接给出了 unsqueeze 修正建议。
- 冗余断言 (correctness): 建议删除冗余断言。

风险与影响

- 风险：
  - 测试代码质量缺陷：部分新增测试存在逻辑错误（如 test_causal_conv1d 中布尔张量生成异常）可能导致运行时 CI 失败或测试结果无效。
  - 类名复制粘贴错误：test_binding.py 中的 TestGemm 类名与 test_gemm.py 冲突，可能影响测试发现和结果归因。
  - 配置变更风险：CI 工作流新增斜杠命令标签，可能与其他已有标签冲突或触发意外的流水线执行。
  - 维护负担：22 个测试文件大量重复样板代码（尽管由 utils.py 缓解），但若后续内核接口变更需要同步更新所有测试文件。
  - 后续 revert：这些质量问题可能是后续 revert PR #25044 的直接原因。

- 影响：
  - 用户影响：无，仅影响测试和 CI 流程。
  - 系统影响：增强了 CPU 后端的测试覆盖，但引入了不稳定的测试用例，可能降低 CI 可靠性。
  - 团队影响：增加了 22 个测试文件和共享工具模块，提升了测试代码复用性和可维护性；但也带来了需修复的代码债务。
  - 影响程度：中等。
  - 风险标记：测试代码质量缺陷，CI 配置变更，潜在运行时错误，已回滚风险

关联脉络

- PR #25044 Revert "Migrate Intel CPU cases to the test/registered.": 该 PR 直接 revert 了 #22670，推测是由于 #22670 中测试质量问题或 CI 失败导致需要回退。