

PR #25007 完整报告

sgl-project/sglang

Add Arm64 INT8 MoE test coverage

合并时间: 2026-06-10 10:36

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/25007>

执行摘要

- 一句话: 为 Arm64 CI 添加 INT8 MoE 测试并迁移至注册系统
- 推荐动作: 值得精读其中的 CI 迁移模式: 通过 `register_cpu_ci` 声明式注册测试, 配合集中式 `PER_COMMIT_SUITES` 管理套件, 可显著简化多平台 CI 配置。后续清理 `test/srt/cpu/` 目录时可参考此 PR 的副本策略。

功能与动机

现有 `test_moe.py` 包含 BF16/FP8/INT4 等 Arm64 尚未支持的量化路径, 直接复用会导致 CI 假失败, 因此需要一份专属的 INT8 MoE 测试文件。同时, Arm64 测试需要跟随 Xeon 的迁移路径, 从 `test/srt/run_suite.py` 迁移至 `test/run_suite.py` 的注册系统。

实现拆解

1. 新增 INT8 MoE 测试文件: 在 `test/registered/cpu/arm64/test_moe.py` 创建 `TestFusedExpertsInt8` 类, 通过 `_int8_moe` 方法构造随机 W8A8 输入, 调用 `kernel.fused_experts_cpu` 与 PyTorch 参考实现对比, 使用宽松容差 (Arm64 上 `atol=0.03`)。测试维度覆盖 `M={1,6,32,64}`、`N/K={256,512}`、`E=8`、`topk=4`。
2. 在旧目录保留副本: 为兼容 `test/srt/run_suite.py` 的固定 glob 检查, 同时保留 `test/srt/cpu/arm64/test_moe.py` (仅 suite 名称不同)。后续旧目录完全移除后该副本可删除。
3. 注册现有 CPU 测试到 Arm64 套件: 在 `test/registered/cpu/` 下的 `activation`、`decode`、`norm`、`qwen3`、`rope`、`server_args_backend`、`topk` 等测试文件中各添加一行 `register_cpu_ci(est_time=10, suite="base-b-test-cpu-arm64")`, 使它们也在 Arm64 CI 中运行。
4. 更新 CI 配置与套件定义: 在 `test/run_suite.py` 的 `PER_COMMIT_SUITES` 中添加 "`base-b-test-cpu-arm64`"; 在 `.github/workflows/pr-test-arm64.yml` 中将运行命令从 `test/srt/run_suite.py` 切换到 `test/run_suite.py --hw cpu --suite base-b-test-cpu-arm64`。
5. 清理旧套件: 从 `test/srt/run_suite.py` 中移除旧的 `suite_arm64` 定义及其 `suites.update(suite_arm64)` 调用。

关键文件:

- `test/registered/cpu/arm64/test_moe.py` (模块 MoE 测试; 类别 test; 类型 test-coverage; 符号 `TestFusedExpertsInt8`, `_int8_moe`, `test_int8_moe`): 新增的

Arm64 INT8 MoE 核心测试文件，覆盖 W8A8 量化路径，是本次 PR 的主要目标。

- `test/srt/cpu/arm64/test_moe.py` (模块 MoE 测试; 类别 test; 类型 test-coverage; 符号 `TestFusedExpertsInt8, _int8_moe, test_int8_moe`) : 为兼容旧 `test/srt/run_suite.py` 的 glob 检查而保留的副本, 仅在 suite 名称上不同 (`per-commit-cpu-arm64`), 后续可删除。
- `test/run_suite.py` (模块 CI 调度; 类别 test; 类型 test-coverage) : 在 `PER_COMMIT_SUITES` 中添加 `base-b-test-cpu-arm64` 套件, 使注册的 Arm64 测试能被 CI 发现和执行。
- `test/srt/run_suite.py` (模块 CI 调度; 类别 test; 类型 test-coverage) : 移除旧的 `suite_arm64` 定义, 清理已迁移的测试套件, 避免与新注册系统冲突。
- `.github/workflows/pr-test-arm64.yml` (模块 CI 配置; 类别 infra; 类型 infrastructure) : CI 工作流切换到 registry-based 运行器, 使用 `test/run_suite.py` 和新的套件名称。
- `test/registered/cpu/test_activation.py` (模块 激活测试; 类别 test; 类型 test-coverage) : 作为代表, 说明现有 CPU 测试通过添加一行 `register_cpu_ci` 被纳入 Arm64 套件。

关键符号: `TestFusedExpertsInt8._int8_moe`, `TestFusedExpertsInt8.test_int8_moe`

关键源码片段

`test/run_suite.py`

在 `PER_COMMIT_SUITES` 中添加 `base-b-test-cpu-arm64` 套件, 使注册的 Arm64 测试能被 CI 发现和执行。

修改前:

```
PER_COMMIT_SUITES = {
    HWBackend.CPU: ["base-a-test-cpu", "base-b-test-cpu"],
    # ...
}
```

修改后:

```
PER_COMMIT_SUITES = {
    HWBackend.CPU: ["base-a-test-cpu", "base-b-test-cpu", "base-b-test-cpu-arm64"],
    # ...
}
```

评论区精华

- `cyb70289` 建议目录结构: “Since this file is arm64 only, it's better to create a subdir 'arm64', and move it there, with the name `test_moe.py`”。`ranimandepudi` 同意并执行。
- `mingfeima` 推动注册系统迁移: “could you please move test cases to `sglang/test/registered/cpu`? we have moved all the xeon ci test cases to `registered`”。随后在 review 中要求更新 CI 工作流以使用 `test/run_suite.py` 和新的套件命名。
- CI 基础设施适配: 多次 CI 失败后, 作者发现 `test/run_suite.py` 的 `validate_all_suites()` 要求套件名称需在 `PER_COMMIT_SUITES` 中, 遂添加 `"base-b-test-cpu-arm64"`; 同时旧 `test/srt/run_suite.py` 的 glob 检查需要文件仍在 `test/srt/` 下, 故保留副本。
- 测试文件目录位置 (design): 作者同意并实施, 创建了 `cpu/arm64` 子目录。

- 迁移到 test/registered/cpu (design): 作者后续提交了迁移，将测试置于 test/registered/cpu/arm64/test_moe.py，并在旧目录保留副本以兼容 CI 检查。
- CI 工作流与套件命名 (infra): 作者最终提交完善了 CI 配置，arm64 套件改用新命名并成功运行。

风险与影响

- 风险：变更全部为测试和 CI 配置，无产品代码改动。主要风险是新注册的测试在 Arm64 CI 中可能因环境差异失败，但已在 AWS Graviton3 验证通过。副本文件的存在可能导致后续目录清理时遗漏，需在 test/srt/cpu/ 正式废弃后手动删除。
- 影响：对 Arm64 CI：测试覆盖率显著提升，增加了 INT8 MoE 及更多 core kernel 测试（activation、decode 等）；CI 运行命令从旧 run_suite.py 切换到新注册系统，与 Xeon/AMD 等平台对齐。对其他平台无影响。团队层面，这展示了从旧测试框架向注册系统迁移的完整模式，可供 NPU、XPU 等参考。
- 风险标记：测试副本可能遗留，CI 配置变更可能影响其他平台

关联脉络

- PR #16045 W8A8 INT8 Arm CPU PR: 实现了 aarch64/moe.cpp 内核，是此测试覆盖的底层实现基础。
- PR #22123 Phase 1A bootstrap PR: 建立了 Arm64 CI 基础设施，是此 PR 的前序工作。
- PR #25139 Move cpu tests under new dir: 将 Xeon CPU 测试迁移到 test/registered/cpu/，此 PR 跟随其模式。
- PR #22670 Migrate CPU tests to test/registered/cpu/: 计划迁移 CPU 测试到新目录，此 PR 是具体执行的一部分。