PR #24649 完整报告

sgl-project/sglang

[Xeon] CPU CI enhancement for Intel Xeon platforms

合并时间：2026-05-28 10:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24649

执行摘要

- 一句话：增强 Intel Xeon CPU CI 稳定性和测试覆盖
- 推荐动作：该 PR 显著提升了 Xeon CI 的可靠性和覆盖范围，设计合理，讨论中的问题均已澄清或解决。建议合并，并考虑后续将 HF_TOKEN 迁移至 GitHub Secrets 以增强安全性。

功能与动机

之前的 Xeon 工作流存在几个实际问题：固定 HF 缓存路径不匹配当前环境、缺少 HF_TOKEN 注入导致门控模型测试失败、测试覆盖不足、单分片执行易超时。本 PR 旨在解决这些问题，使 Xeon CI 能运行更广泛且有代表性的 CPU 测试集。

实现拆解

1. 更新 CI 工作流(.github/workflows/pr-test-xeon.yml)：加入 workspace 清理步骤；将固定 HF 缓存路径改为使用 $HOME/.cache/huggingface；注入 HF_TOKEN 以访问门控模型；引入矩阵构建策略，将 base-b-test-cpu 套件分为 2 个自动分区并行执行。
2. 更新 Docker 镜像(docker/xeon.Dockerfile)：安装 pytest，确保所有扩展测试的运行时依赖满足。
3. 扩展测试注册：在多个测试文件中增加 register_cpu_ci(est_time=..., suite="base-b-test-cpu") 调用，涉及 LoRA、模型加载、路由调度、性能分析、调试工具、可观测性等模块，仅添加注册，不修改测试逻辑。
4. 调整测试参数(test/srt/cpu/test_cpu_graph.py)：将 --mem-fraction-static 从 0.05 提高到 0.5，以支持 torch-compile CPU graph 场景。

关键文件：

- .github/workflows/pr-test-xeon.yml（模块 CI 流程；类别 infra；类型 infrastructure）：核心 CI 工作流文件，实现了缓存路径修正、令牌注入、矩阵分区等关键改动，是本次 PR 的主干。
- docker/xeon.Dockerfile（模块 Docker；类别 infra；类型 infrastructure）：Xeon Docker 镜像文件，安装 pytest 以支持新注册的测试集。
- test/registered/lora/test_lora_openai_api.py（模块 测试注册；类别 test；类型 test-coverage）：代表性地展示了测试注册扩展模式：新增 register_cpu_ci 调用，将测试纳入 CPU CI 套件。

- test/registered/model_loading/test_external_models.py（模块 测试注册；类别 test；类型 test-coverage）：模型加载测试，加入后显著增加 CPU 测试覆盖率，且估计时间较长（203s），是覆盖范围扩大的典型。
- test/srt/cpu/test_cpu_graph.py（模块 CPU 测试；类别 test；类型 test-coverage）：调整内存分片参数以适配 Xeon 平台，直接涉及 CPU 推理测试的稳定性。

关键符号：未识别

关键源码片段

test/registered/lora/test_lora_openai_api.py

代表性地展示了测试注册扩展模式：新增 register_cpu_ci 调用，将测试纳入 CPU CI 套件。

```
"""
Unit tests for OpenAI-compatible LoRA API support.
"""

import unittest
from unittest.mock import MagicMock

from sglang.srt.entrypoints.openai.serving_base import OpenAIServingBase
from sglang.srt.server_args import ServerArgs
# 新增 register_cpu_ci 导入
from sglang.test.ci.ci_register import (
    register_amd_ci,
    register_cpu_ci,
    register_cuda_ci,
)

register_cuda_ci(est_time=30, suite="nightly-1-gpu", nightly=True)
register_amd_ci(est_time=30, suite="nightly-amd-1-gpu", nightly=True)
# 将本测试注册到 CPU CI 的 base-b-test-cpu 套件
register_cpu_ci(est_time=8, suite="base-b-test-cpu")

class MockTokenizerManager:
    """Mock TokenizerManager for testing."""
    def __init__(self, enable_lora=False):
        self.server_args = MagicMock(spec=ServerArgs)
        self.server_args.enable_lora = enable_lora
        self.server_args.tokenizer_metrics_allowed_custom_labels = None

class ConcreteServingBase(OpenAIServingBase):
    """Concrete implementation for testing abstract base class."""
    def _request_id_prefix(self) -> str:
        return "test-"
    def _convert_to_internal_request(self, request, raw_request=None):
        pass
    def _validate_request(self, request):
        pass
```

```
class TestParseModelParameter(unittest.TestCase):
    """Test _parse_model_parameter method."""
    def setUp(self):
        self.tokenizer_manager = MockTokenizerManager(enable_lora=True)
        self.serving = ConcreteServingBase(self.tokenizer_manager)
```

评论区精华

- 分区资源分配：mingfeima 询问两个分区是否运行在不同的机器上，1pikachu 确认会分布在两个 runner 上，以平衡负载。
- 测试套件重复注册：mingfeima 指出部分测试已注册 base-a-test-cpu 是否还需 base-b-test-cpu，1pikachu 说明 base-a-test-cpu 专用于 CUDA，因此需要分别注册。
- HF_TOKEN 安全性：mingfeima 质疑通过文件读取传递令牌的做法，1pikachu 建议使用 GitHub Actions Secrets，但当前方案受限于 runner 配置权限。
- test_cpu_graph.py 变更影响：mingfeima 担忧可能影响 ARM CI，1pikachu 解释该变更是专为 EMR/SPR 平台调整，不影响其他后端。
  - 分区是否在不同机器上运行 (question): 1pikachu 确认是的，会分布到两台 runner 上以分散负载。
  - 已有 base-a-test-cpu 注册是否还需 base-b-test-cpu (question): 1pikachu 解释 base-a-test-cpu 套件专用于 CUDA，因此 CPU 测试需要单独注册 base-b-test-cpu。
  - HF_TOKEN 传递方式的安全性 (security): 1pikachu 推荐使用 GitHub Actions Secrets，但当前 runner 权限限制了该方案，暂时接受现有方式。
  - test_cpu_graph.py 变更是否影响 ARM CI (question): 1pikachu 澄清该变更是专为 EMR/SPR 平台，不影响 ARM。

风险与影响

- 风险：
  - CI 稳定性风险：自托管 runner 的 workspace 清理可能与其他 job 冲突，虽然加了 || true 但仍存在偶发失败可能。
  - 令牌安全风险：通过文件读取 HF_TOKEN 并注入环境变量，若文件权限控制不当可能导致泄露，推荐迁移到 GitHub Secrets。
  - ARM CI 回归：test_cpu_graph.py 的 mem-fraction-static 调整仅针对 Xeon，但 ARM 测试套件若共享同一文件可能受影响，但评论确认不触发。
  - 测试覆盖膨胀：大量测试注册到 base-b-test-cpu 可能导致总执行时间增加，虽已分区，但仍需监控超时情况。
- 影响：
  - 用户 / 系统：仅影响 Xeon CPU CI 运行，对用户无直接感知。
  - 团队：开发者将获得更可靠的 Xeon CI 反馈，减少因环境配置导致的失败；测试覆盖提升有助于提前发现 CPU 相关回归。
  - 影响程度：中等，因为仅限于 CI 基础设施层，不涉及核心逻辑变更。
  - 风险标记：CI 分区可能资源竞争，HF_TOKEN 明文传递风险，ARM CI 可能受影响

关联脉络

- PR #25061 Fix MiniMax-M2.7 on CPU: 同属 CPU 平台测试与修复，本 PR 扩展了 CPU CI 覆盖，与之形成互补。