PR #21456 完整报告

sgl-project/sglang

[CPU] upgrade dependent torch ver to PT2.12

合并时间：2026-06-04 11:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21456

执行摘要

- 一句话：升级 CPU 端 PyTorch 系列依赖至 2.12
- 推荐动作：建议 CPU 平台开发者和依赖管理者阅读该 PR，了解版本升级细节和适配方式。对于仅关注 GPU 的读者，此 PR 无直接参考价值。设计决策方面，迁移 AMX 查询到公开 API 是良好的版本兼容实践。

功能与动机

根据 PR 作者说明，PyTorch 2.11 未包含所需的 FP8 优化，因此关闭原 targeting PT2.11 的 PR，转而升级到 PT2.12 以获取最新的 FP8 加速能力。

实现拆解

1. 依赖版本升级：修改 python/pyproject_cpu.toml 和 sgl-kernel/pyproject_cpu.toml，将 torch 从 2.9.0 升至 2.12.0，torchao 从 0.14.1 升至 0.17.0，torchaudio 从 2.9.0 升至 2.11.0，torchvision 从 0.24.0 升至 0.27.0，triton 从 3.5.0 升至 3.7.0。
2. AMX 查询 API 迁移：在 python/sglang/srt/utils/common.py 和 python/sglang/multimodal_gen/runtime/utils/common.py 中，将 torch._C._cpu._is_amx_tile_supported() 替换为 torch.cpu._is_amx_tile_supported()，以匹配 PyTorch 2.12 的公开 API。
3. 测试种子调整：因 torch 新版本的随机数生成行为变化，调整了 test/registered/cpu/test_topk.py、test/srt/cpu/test_topk.py、test/registered/cpu/test_moe.py、test/srt/cpu/test_moe.py、test/registered/cpu/test_rope.py、test/srt/cpu/test_rope.py 中的随机种子，消除随机性导致的断言差异。
4. CI 配置微调：更新 .github/workflows/pr-test-xeon.yml，适配新的依赖版本。

关键文件：

- python/pyproject_cpu.toml（模块 CPU 配置；类别 config；类型 configuration）：CPU 平台依赖版本声明，核心变更文件，升级了 torch/torchvision/torchaudio/torchao/triton 等关键包。
- python/sglang/srt/utils/common.py（模块 SRT 工具；类别 source；类型 core-logic；符号 is_amx_tile_supported）：SRT 公共模块，更新了 AMX 支持查询接口，确保在 PyTorch 2.12 下正确检测 AMX 能力。

- python/sglang/multimodal_gen/runtime/utils/common.py（模块 多模态工具；类别 source；类型 core-logic；符号 is_amx_tile_supported）：多模态生成模块公共工具，与 SRT 同步 AMX 检测接口变更，保持一致。
- test/registered/cpu/test_topk.py（模块 测试套件；类别 test；类型 test-coverage）：CPU topk 测试，调整随机种子以适应新 torch 版本行为，消除 CI 失败。
- test/srt/cpu/test_topk.py（模块 测试套件；类别 test；类型 test-coverage）：与 test/registered/cpu/test_topk.py 相同的种子调整，保持一致性。
- test/registered/cpu/test_moe.py（模块 测试套件；类别 test；类型 test-coverage）：CPU MoE 测试种子调整，从 1234 改为 1183。
- test/srt/cpu/test_moe.py（模块 测试套件；类别 test；类型 test-coverage）：与 test/registered/cpu/test_moe.py 相同的种子调整。
- test/registered/cpu/test_rope.py（模块 测试套件；类别 test；类型 test-coverage）：CPU RoPE 测试种子调整。
- test/srt/cpu/test_rope.py（模块 测试套件；类别 test；类型 test-coverage）：与 test/registered/cpu/test_rope.py 同步种子调整。
- sgl-kernel/pyproject_cpu.toml（模块 内核配置；类别 config；类型 configuration）：sgl-kernel 的 CPU 依赖配置，与主 CPU 配置同步升级 torch 等版本。
- .github/workflows/pr-test-xeon.yml（模块 CI 配置；类别 infra；类型 infrastructure）：CPU CI 工作流配置，可能更新了容器镜像标签或缓存键以匹配新依赖。

关键符号：is_amx_tile_supported

关键源码片段

python/pyproject_cpu.toml

CPU 平台依赖版本声明，核心变更文件，升级了 torch/torchvision/torchaudio/torchao/triton 等关键包。

```
# pyproject_cpu.toml 依赖版本更新片段
[project]
dependencies = [
  "torch==2.12.0", # 之前 2.9.0，跳过大版本以获取 FP8 优化
  "torchao==0.17.0", # 之前 0.14.1
  "torchaudio==2.11.0", # 之前 2.9.0
  "torchvision==0.27.0",# 之前 0.24.0
  "triton==3.7.0", # 之前 3.5.0
  # ... 其余依赖不变
]
```

python/sglang/multimodal_gen/runtime/utils/common.py

多模态生成模块公共工具，与 SRT 同步 AMX 检测接口变更，保持一致。

```
try:
    # move torch.cpu._is_amx_tile_supported() from cpu_has_amx_support
    # to support torch compile
```

```
    is_amx_tile_supported = torch.cpu._is_amx_tile_supported()
except:
    is_amx_tile_supported = False
```

评论区精华

1. 作者 ZailiWang 说明关闭旧 PR 原因："Closing the PR since some FP8 optimizations in torch is not included in PT2.11 Targeting PT2.12"。
2. 合并者 mingfeima 要求先修复 Xeon CI："@1pikachu @MingxuZh first fix xeon ci issue. @ZailiWang then rebase this one."。
3. 修复后作者确认 Xeon CI 通过："Xeon CI passed with the tuned torch seeds in some CI cases: ... It can be merged if the model level E2E acc tests pass."。

- 关闭旧 PR 原因 (other): 决定迁移到 PT2.12 以获取 FP8 优化。
- Xeon CI 修复和 rebase (testing): 通过调整测试种子修复 CI，然后成功 rebase 并合并。
- 测试种子调整 (testing): 最终通过为每个测试方法单独设置种子（如 topk 改为 12，moe 改为 1183）解决 CI 失败。

风险与影响

- 风险：
  1. 依赖兼容性风险：PyTorch 2.12 可能引入与旧版本不完全兼容的 API 变化，但由于仅涉及 CPU 平台且已通过 CI，风险可控。
  2. 测试覆盖不足：测试种子调整后可能掩盖部分数值边界差异，尤其是 topk 和 moe 测试，需关注后续随机性相关 failure。
  3. AMX 查询接口风险：torch.cpu._is_amx_tile_supported() 是推荐 API，但若未来 PyTorch 修改该接口，CPU 平台 AMX 检测可能失效。
  4. 回归范围：变更集中在 CPU 配置和测试，对 GPU 路径无影响。 - 影响：用户影响：CPU 平台用户需安装 PyTorch 2.12 及配套包（torchao、triton 等），升级后可利用新的 FP8 优化提升性能。系统影响：CPU 相关的推理与训练路径将依赖新版本，务必保持依赖一致性。团队影响：后续 CPU 开发需基于此版本依赖，维护时需注意与主分支 torch 版本的同步。影响程度中等，主要限于 CPU 子项目。
- 风险标记：依赖升级可能引入回归，测试种子调整后覆盖可能不足，AMX API 变更仅影响 CPU

关联脉络

- PR #21455 [CPU] upgrade dependent torch ver to PT2.11: 同一作者之前 targeting PT2.11 的 PR，因缺少 FP8 优化被关闭，本 PR 是替代方案。