PR #25971 完整报告

sgl-project/sglang

[CPU Doc]Add Xeon CPU info in Qwen3 Cookbook

合并时间：2026-05-27 03:14

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25971

执行摘要

本 PR 在 Qwen3 Cookbook 中添加了 Intel Xeon CPU 的部署支持信息，包括交互式命令生成器的硬件配置和文档中的安装指导。变更主要涉及两个文件，但 review 中出现了关于 TP 数值正确性和 FP8 支持的争议，作者部分采纳了建议，但在关键参数上坚持己见。合并时这些争议未完全解决，可能对用户造成误导。

功能与动机

PR body 明确说明动机是 'Adding Xeon support information into SGLang Cookbook. Starting with Qwen3 page.' 旨在帮助用户在 Intel Xeon CPU 上部署 Qwen3 模型，填补文档中 CPU 部分的空白。

实现拆解

1. 部署配置组件 (JSX) 扩展: 在 qwen3-deployment.jsx 的 modelConfigs 对象中，为每个模型大小（235B 到 0.6B）新增 xeon 键，设置 TP 值（大模型 6，中小模型 3）和 FP8 true。同时删除了一段硬编码 hardware default 的冗余逻辑，该逻辑被 review 指出是无效的。
2. 文档 Markdown 更新: 在 Qwen3.mdx 中添加指向 CPU 安装指南的相对链接，更新硬件描述以包含 Intel Xeon CPU，并在部署参数章节新增 CPU 配置提示，建议用户参考 CPU 服务器文档了解 TP 和 NUMA 绑定。
3. 配套修复: 将 review 中提到的两个绝对链接改为相对链接，确保在预览和生产环境的一致行为。

docs_new/src/snippets/autoregressive/qwen3-deployment.jsx

核心变更文件，在硬件配置字典中为每个模型添加了 xeon 键和 TP/FP8 参数，同时清理了冗余 UI 逻辑。

```
// 文件：docs_new/src/snippets/autoregressive/qwen3-deployment.jsx
// 在 modelConfigs 的每个模型项中添加 xeon 配置
const modelConfigs = {
  '235b': {
    baseName: '235B-A22B',
    hasThinkingVariants: true,
    h100: { tp: 8, ep: 0, bf16: true, fp8: true },
    // ... 其他 GPU 配置
    mi355x: { tp: 4, ep: 0, bf16: true, fp8: true },
```

```
      xeon: { tp: 6, ep: 0, bf16: true, fp8: true } // 新增，TP=6 针对 6 代 Xeon
    },
    '30b': {
      // ...
      xeon: { tp: 3, ep: 0, bf16: true, fp8: true } // 新增，TP=3
    },
    // 其他模型类似（32b TP=6，14b/8b/4b/1.7b/0.6b TP=3）
  };

  // 被移除的冗余逻辑（已删除）：
  // if (values.hardware === 'xeon') {
  // options.hardware.items.map(...)
  // }
```

评论区精华

- TP 值正确性：gemini-code-assist[bot] 提出 'The tp size for Xeon should ideally be a power of 2' 且 '128 is not divisible by 6'，建议改为 tp:4。作者回应 'TP 3/6 is required for 6th Gen Xeon Processors'，未接受建议，亦未提供整除证据。
- FP8 支持：review 指出 'SGLang's CPU backend currently does not support FP8 quantization'，建议设为 false，但作者未回应，最终 merged 版本仍保留 fp8: true。
- 冗余逻辑和 链接相对化：作者接受了这两项建议并修改。

风险与影响

- TP 配置风险：若 Qwen3 模型的 attention head 数不能被 6 或 3 整除，用户按文档配置将直接运行失败。虽然作者声称 TP 3/6 是 6 代 Xeon 必须，但未提供验证数据，风险较高。
- FP8 误导风险：CPU 后端不支持 FP8，但配置显示 fp8: true，可能使用户尝试无效选项或产生困惑。
- 影响范围：仅影响阅读文档的 Qwen3 CPU 部署用户，无代码逻辑变更。

关联脉络

该 PR 与 #12662（CPU Qwen3-VL/Omni 支持）属于同一功能线，前者实现 CPU 支持功能，后者补充部署文档，逐步完善 CPU 平台的用户指引。后续可能需要对 Qwen3 各模型 attention head 数进行验证并修正 TP 值。