

PR #23547 完整报告

sgl-project/sglang

[Intel CPU/XPU] SGL doc updates

合并时间: 2026-04-24 09:23

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/23547>

执行摘要

本 PR 更新了 Intel CPU 和 XPU 平台的部署文档，将旧启动命令统一替换为新的 `sglang serve` 入口，并为 XPU 添加了完整的 Docker 环境搭建指南。但 `xpu.mdx` 中未包含 `--host 0.0.0.0` 参数，可能导致服务无法外部访问；此外 CPU 文档中部分示例的 `--tp` 值被修改，可能与模型不匹配。建议在合并后补充修复。

功能与动机

根据 PR body，更新 CPU/XPU 文档的动机包括：添加 XPU Docker 环境搭建示例命令、将旧命令（`python -m sglang.launch_server`）替换为新的 `sglang serve` 脚本、移除 `--torch-compile-max-bs` 的约束描述（该约束已在 #13881 合并后解除）。

实现拆解

仅涉及文档文件，无代码变更：

cpu_server.mdx (+44/-44)

- 替换所有 `python -m sglang.launch_server` 为 `sglang serve`，`--model` 改为 `--model-path`，调整参数顺序（如 `--host` 移到最后）。
- 删除关于 `--torch-compile-max-bs` 的额外说明（“can be up to 16”），简化注释。
- 修改部分示例的 `--tp` 值：Llama-3.2-3B 从 `--tp 2` 改为 `--tp 3`。

xpu.mdx (+28/-3)

- 新增 Docker 安装章节，包含完整的构建和运行命令（挂载 `/dev/dri`、设置 `HF_TOKEN` 等）。
- 启动命令同样替换为新格式，但未包含 `--host 0.0.0.0` 参数。

评论区精华

gemini-code-assist[bot]: “The `--host 0.0.0.0` argument is missing in the new `sglang serve` command, which may affect external accessibility...”

该评论未被作者回复或修复，PR 已合并，问题遗留。

风险与影响

- `xpu.mdx` 缺少 `--host` 参数：用户按照文档启动服务将只监听 `localhost`，无法从外部访问，这是中优先级风险。

- cpu.mdx 中 `--tp` 值变更：例如 Llama-3.2-3B 从 `tp=2` 改为 `tp=3`，若模型实际不支持 3 路张量并行，将导致启动失败。
- 影响范围：仅影响阅读文档的 Intel CPU/XPU 平台部署用户，但错误命令可能导致部署故障。

关联脉络

本 PR 是文档更新，关联的 #13881 移除了 `--torch-compile-max-bs` 的约束，本 PR 据此删除了相关注释。与近期 CPU/XPU 后端的持续支持（如 #23528、#23470）一致，体现了对 Intel 平台的完善。