PR #22446 完整报告

sgl-project/sglang

[NPU] add qwen3-30b-a3b low latency example

合并时间：2026-04-11 15:52

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22446

执行摘要

本 PR 为 sglang 仓库的 Ascend NPU 平台文档添加了 Qwen3-30B-A3B 模型的低延迟部署示例，通过更新最佳实践文档提供详细配置指南，旨在帮助用户优化模型性能，变更仅涉及文档，风险较低。

功能与动机

PR 的动机直接来自作者描述：'add qwen3-30b-a3b low latency example'。目的是扩展 Ascend NPU 最佳实践文档，覆盖新模型 Qwen3-30B-A3B 的低延迟配置案例，为用户提供现成的部署参考，减少调优时间。

实现拆解

PR 仅修改一个文件 docs/platforms/ascend/ascend_npu_best_practice.md，具体变更包括：

- 表格更新：在性能汇总表中新增两行： | 模型 | 硬件 | 卡数 | 部署模式 | 输入输出长度 | TPOT | 量化 | 链接 | |------|------|------|----------|--------------|------|------|------| | Qwen3-30B-A3B | Atlas 800I A3 | 1 | PD Mixed | 6K+1.5K | 10ms | W8A8 INT8 | 链接 | | Qwen3-30B-A3B | Atlas 800I A3 | 1 | PD Mixed | 1K+0.3K | 8ms | W8A8 INT8 | 链接 |
- 新章节添加：添加 'Qwen3-30B-A3B 6K-1_5K 10ms on A3 1 Cards Mixed Mode' 章节，包含环境变量设置、启动命令和基准测试脚本，例如：shell export SGLANG_SET_CPU_AFFINITY=1 python -m sglang.launch_server --model-path $MODEL_PATH ...

评论区精华

Review 讨论由 iforgetmyname 主导：

- 冗余内容移除：多个 'remove' 评论指示初始提交中存在重复文本，作者在后续提交中可能已清理。
- 配置逻辑质疑：iforgetmyname 提问 'when not enabling any tp here, why do we still need 4 cards?'，这引发对部署配置一致性的关注，提示需要确保文档准确性。

风险与影响

- 风险：文档准确性是主要风险，如环境变量错误或 TP 设置不匹配可能误导用户；由于无代码变更，回归风险低。
- 影响：用户可直接参考新示例加速部署，但对系统性能无直接影响；团队需维护文档更新以保持时效性。

关联脉络

从近期历史 PR 看，本 PR 专注于文档更新，与同仓库中其他 NPU 相关 PR（如 PR 17920 关于 Intel XPU 支持）无直接关联，但反映了仓库持续扩展平台兼容性和优化文档的趋势。