PR #23459 完整报告

sgl-project/sglang

[NPU] [DOC] Update Ascend NPU best practice

合并时间：2026-04-22 17:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23459

PR 分析报告 #23459: [NPU] [DOC] Update Ascend NPU best practice

执行摘要

该 PR 更新了 Ascend NPU 最佳实践文档，将 Qwen3-Next-A3B-Instruct 模型的推荐配置从双卡改为单卡，并同步调整了基准测试命令参数。变更仅涉及一个文档文件，无代码逻辑修改，属于轻量级文档维护。

功能与动机

根据 PR 描述，本次更新旨在反映 Ascend NPU 平台的最新最佳实践。具体而言，将 Qwen3-Next 模型在 Atlas 800I A3 硬件上的推荐卡数从 2 张调整为 1 张，并更新了相应的基准测试配置和部署命令。

实现拆解

变更集中在 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_best_practice.mdx 文件中，主要包含以下修改：

1. 硬件配置表更新：将 Qwen3-Next-A3B-Instruct 行的卡数列从 "2" 改为 "1"，同时将对应的锚点链接从 #qwen3-next-3_5k-1_5k-20ms-on-a3-2-cards-mixed-mode 更新为 #qwen3-next-3_5k-1_5k-20ms-on-a3-1-cards-mixed-mode。
2. 基准测试命令更新：调整了 bench_serving 命令的参数，以匹配单卡配置：--max-concurrency 从 768 改为 1024，--random-input-len 从 3500 改为 3584，--random-output-len 从 1500 改为 1536，--num-prompts 从 3072 改为 7168，--request-rate 从 16 改为 40。
3. 部署配置章节更新：将章节标题从 "Qwen3-Next 3_5K-1_5K 20ms on A3 2 Cards Mixed Mode" 改为 "1 Cards Mixed Mode"，并将硬件描述从 "2Card" 改为 "1Card"。同时更新了启动命令中的 --tp-size 从 4 改为 2，并移除了 --dp-size 2 和相关标志。

评论区精华

• gemini-code-assist[bot]提出了两个语法一致性建议：

" 锚点链接使用复数 'cards' 但对应单卡配置，建议改为 '1-card'" " 章节标题中 '1 Cards' 应改为 '1 Card'" 这两个建议均未被采纳，但属于合理改进。

风险与影响

- 风险：极低。仅文档变更，无代码逻辑修改。主要风险是文档中的配置信息可能不准确，但鉴于这是官方最佳实践，风险可控。
- 影响：影响范围限定于阅读该文档的用户，特别是部署 Qwen3-Next 模型的 Ascend NPU 用户。影响程度低，不会对系统功能产生任何影响。

关联脉络

- 与 #23378 同为 NPU 相关文档更新，反映了 NPU 文档维护的持续性。
- 当前 PR 修复了之前文档中双卡配置的推荐，使其更符合实际测试结果。

- 报告生成时间：2025-04-11*