PR #25830 完整报告

sgl-project/sglang

[NPU] Docs op performance optimize

合并时间：2026-05-22 09:20

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25830

执行摘要

- 一句话：新增 NPU 算子性能优化文档
- 推荐动作：值得 NPU 开发者和性能调试人员精读，特别是 msProf 输出指标解读部分，可快速定位算子瓶颈。

功能与动机

开发者需要一份系统性的算子性能优化指导，以便在 Ascend NPU 上高效地分析和提升 Triton 算子的性能。PR 正文说明这是 'a guidance about how to obtain and analyze profiling data, and thus enhance performance on the operator level'。

实现拆解

1. 创建文档页面：在 docs_new/docs/hardware-platforms/ascend-npus/ 下新建 ascend_npu_operator_performance_optimizing.mdx，内容涵盖性能数据获取（msProf 使用）、输出指标解读（Duration、Block Dim、aicore_time 等）、瓶颈定位方法以及优化策略（如充分利用 UB、使用双缓冲等）。
2. 更新导航配置：修改 docs_new/docs.json，将新页面加入 Ascend NPU 子导航列表中，同时将 FAQ 页面调整到 profiling 之前，使文档顺序更合理。
3. 质量修复：根据 review 反馈（gemini-code-assist[bot]），修复了代码块语法（单引号改为反引号）、标题格式（去掉下划线改用空格）、拼写错误（I0 → L0）、标点（中文句号改英文）及语法问题，并删除了内部开发备注。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_operator_performance_optimizing.mdx（模块 NPU 文档；类别 docs；类型 documentation）：核心新增文档，提供 Ascend NPU 算子性能优化的完整指导
- docs_new/docs.json（模块 导航配置；类别 config；类型 configuration）：导航配置文件，将新文档页面加入 Ascend NPU 分组导航

关键符号：未识别

关键源码片段

docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_operator_performance_optimizing.mdx

核心新增文档，提供 Ascend NPU 算子性能优化的完整指导

```
# 使用 msProf 采集指定算子的性能数据（替换 kernel-name 和脚本）
msprof op --kernel-name=DequantSwigluQuant python3 test_dequant_swiglu.py

# 输出包含以下关键指标：
# - Duration：算子执行耗时（us）
# - Block Dim：任务分块数，对应使用的核数
# - Input/Output Shapes：输入输出张量形状
# - aicore_time / aiv_time：AI Core 和 Vector Core 的执行时间
# - aic_mac_ratio / aiv_vec_ratio：MAC/ 向量单元利用率
```

评论区精华

gemini-code-assist[bot]指出了多处格式与语法问题，包括代码块使用三单引号而非三反引号、标题包含下划线、硬件层级 'I0' 应大写为 'L0'、以及存在不应出现的内部开发者备注。作者逐一确认并应用了所有建议，最终文档质量得到提升。

- 文档格式与语法问题修复 (style)：作者逐一回应并应用了所有建议，最终文档符合规范。

风险与影响

- 风险：纯文档变更，无任何代码或配置风险。但需确保导航链接正确，避免 404 错误。经 review 确认无此问题。
- 影响：对 Ascend NPU 用户提供清晰的操作指南，降低性能优化入门门槛；对团队文档体系进行了补充，扩展了 NPU 专区内容。
- 风险标记：暂无

关联脉络

- PR #25995 docs: delete deprecated args from npu supported features：同为 NPU 文档改进，共同丰富 NPU 专区内容