# PR #25114 完整报告

sgl-project/sglang

[NPU] [DOC] add performance testing and optimization docs for npu

合并时间：2026-05-14 09:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25114

## 执行摘要

- 一句话：新增 NPU 性能测试与优化文档
- 推荐动作：建议所有 Ascend NPU 用户阅读这两份文档，尤其是 ascend_npu_optimization.mdx 中的参数解析表。设计决策方面，选择将参数分为“必选”和“优化建议”两类值得借鉴，便于用户优先关注关键配置。

## 功能与动机

根据 PR 描述：'add performance testing and optimization docs for npu'。目的是为 Ascend NPU 用户提供标准化的性能测试流程和优化指导，填补该硬件平台的文档空白，使用户能够更高效地进行部署和调优。

## 实现拆解

1. 创建性能测试文档：在 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_performance_testing.mdx 中，覆盖文本生成、多模态和 Embedding 三种模型类型，分别给出在线（bench_serving）和离线（bench_offline_throughput）基准测试命令，并介绍 Evalscope、AISBench 等工具的使用。
2. 创建优化指南：在 docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_optimization.mdx 中，以 DeepSeek-V3.2 最佳实践为示例，详细解析系统级优化（CPU 调频、NUMA 等）、内存与设备配置（expandable_segments、NPU_DEVICE_ID 等）以及多节点通信等参数，并区分必选与可调参数。
3. 更新导航配置：在 docs_new/docs.json 的 Ascend NPU 子页面列表中添加对上述两个新文档的引用，使其出现在文档导航中。
4. 修复与完善：通过后续 commits 修复了死链接、将文档标题从“stress test”调整为“performance test”，并添加示例性免责声明。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_performance_testing.mdx（模块 NPU 文档；类别 other；类型 documentation-add）：核心新增：提供三类模型的性能测试方法，包括在线 / 离线基准命令，是 NPU 性能测试的入口文档。
- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_optimization.mdx（模块 NPU 文档；类别 other；类型 documentation-add）：核心新增：系统阐述 NPU 部署优化参数，包含系统级、内存设备级配置表格，是调优的主要参考文档。

- docs_new/docs.json（模块 导航配置；类别 config；类型 configuration-add）：导航配置文件更新，将两个新文档加入文档侧边栏层次结构中，是文档可发现性的必要变更。

关键符号：未识别

### 关键源码片段

#### docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_performance_testing.mdx

核心新增：提供三类模型的性能测试方法，包括在线 / 离线基准命令，是 NPU 性能测试的入口文档。

```
# 启动文本生成模型服务（Qwen2.5-7B-Instruct）
sglang serve --model-path Qwen/Qwen2.5-7B-Instruct

# 启动多模态模型服务（Qwen2.5-VL-7B-Instruct），需要指定 ascend_attn 后端
sglang serve --model-path Qwen/Qwen2.5-VL-7B-Instruct --mm-attention-backend ascend_attn

# 启动 Embedding 模型服务（Qwen3-Embedding-8B），使用 --is-embedding 标志
sglang serve --model-path Qwen/Qwen3-Embedding-8B --is-embedding
```

## 评论区精华

PR 无实质性 review 讨论，由 sglang-npu-bot 快速批准合并。仅有 gemini-code-assist[bot] 触发了每日配额警告，不影响内容。

- 暂无高价值评论线程

## 风险与影响

- 风险：
  - 文档内容准确性：优化指南中建议的参数值（如 PYTORCH_NPU_ALLOC_CONF=expandable_segments:True）可能依赖特定硬件版本（如 Atlas 800I A2 vs A3），用户需根据自身环境验证。风险可控。
  - 链接有效性：文档内引用其他页面，若目标路径更改可能导致死链接。但当前引用均指向同一仓库内的相对路径，风险低。
  - 无代码变更：仅涉及文档与配置，不存在回归或性能风险。
- 影响：
  - 用户影响：显著降低 Ascend NPU 用户的性能测试与调优门槛，提供可直接复用的命令与参数示例。
  - 系统影响：无。
  - 团队影响：增加了文档维护工作量，但提升了 NPU 平台的文档成熟度，有助于减少重复问题。
  - 风险标记：无代码变更，依赖硬件版本，链接有效性

## 关联脉络

- 暂无明显关联 PR