PR #26069 完整报告

sgl-project/sglang

[NPU]Ascend NPU Performance Profiling Guide and Ascend NPU Operator Development Guide

合并时间：2026-05-23 10:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26069

执行摘要

本 PR 为 SGLang 的 Ascend NPU 平台新增性能分析指南和算子开发指南，同时补充多节点部署示例和 Conda 镜像配置说明。纯文档变更，无代码修改，已于 2025-05-26 合并。

功能与动机

为 Ascend NPU 开发者提供系统化的性能分析和算子开发文档，帮助用户在 NPU 上实现 SGLang 的部署与性能优化。具体包括：

- 如何使用 CANN Profiling 工具进行基准测试和性能分析。
- 如何编写并注册自定义算子。
- 多节点部署 Qwen3.5 模型的操作指南。
- 解决 Anaconda 仓库限制的镜像配置方案。

实现拆解

1. 新增性能分析与算子开发文档

在 ascend_npu.mdx 中新增两个标题节，内容涵盖：

- Performance Profiling Guide：介绍 CANN Profiling 的配置、数据采集、结果解读。
- Operator Development Guide：说明自定义算子的开发流程、环境搭建与注册方法。

2. 完善快速开始文档

强调 Python 3.11 的唯一支持性（加粗），并补充 Anaconda 镜像配置示例，包括添加清华源和清除默认源的操作。

3. 多节点部署示例

在 ascend_npu_qwen3_5_examples.mdx 中新增 Multi-node Deployment 章节，提供 Qwen3.5-35B-A3B 模型的双节点启动脚本，包含环境变量设置（SGLANG_ENABLE_SPEC_V2, SGLANG_NPU_USE_MULTI_STREAM, HCCL_BUFFSIZE 等）和 launch_server 命令。

4. 格式修复

根据 review 意见，将 Anaconda 限制说明的标题改为 markdown 四级标题格式。

无涉及代码变更。

评论区精华

> gemini-code-assist[bot]:(medium) 将 HCCL_SOCKET_IFNAME 和 GLOO_SOCKET_IFNAME 设为 lo 对于多节点部署是错误的，应设为实际网络接口（如 eth0）。
>
> gemini-code-assist[bot]:(medium) your port 占位符建议替换为默认端口 20000，使示例即开即用。
>
> gemini-code-assist[bot]:(medium) IP 检测逻辑仅检查前两个 IP，若目标 IP 不在前两个则脚本失败，建议使用 hostname -I 完整列表匹配。

所有 review 评论均未得到回复或修改，但 sglang-npu-bot 给出了审批。

风险与影响

- 技术风险：无。仅文档变更，不影响运行时代码。
- 用户影响：提升 NPU 开发者在 SGLang 上的上手体验和调优能力。
- 系统影响：无。

关联脉络

无直接关联的 PR。属于单次文档补充，无功能演进依赖。