PR #22429 完整报告

sgl-project/sglang

[NPU]add Qwen3-32b and Qwen3-8b low latency md

合并时间：2026-04-09 16:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22429

PR 分析报告：为 NPU 平台添加 Qwen3 低延迟配置文档

执行摘要

本 PR 为 sglang 仓库的 Ascend NPU 最佳实践文档新增了 Qwen3-32B 和 Qwen3-8B 模型的低延迟部署配置，包括在 Atlas 800I A3 硬件上的 2 卡和 1 卡混合模式设置，提供详细命令和基准测试指南，是一个低风险的文档更新，旨在帮助用户优化 NPU 平台性能。

功能与动机

PR 动机基于标题和 body 中的表述："[NPU]add Qwen3-32b and Qwen3-8b low latency md"，即扩展 NPU 平台文档，添加这两个模型的低延迟配置部分。这解决了用户部署 Qwen3 模型时缺乏优化指南的问题，支持平台生态发展。

实现拆解

仅修改了一个文件：docs/platforms/ascend/ascend_npu_best_practice.md，添加了以下四个配置章节：

- Qwen3-32B 1K-0.3K 12ms on A3 2 Cards Mixed Mode
- Qwen3-32B 6K-1.5K 17ms on A3 2 Cards Mixed Mode
- Qwen3-8B 1K-0.3K 7ms on A3 1 Cards Mixed Mode
- Qwen3-8B 6K-1.5K 9ms on A3 1 Cards Mixed Mode

每个章节结构一致，包含：

- 模型与硬件：指定模型版本、硬件型号和卡数。
- 部署参数：如部署模式（PD Mixed）、数据集（random）、输入输出长度、TPOT（Time Per Output Token）。
- 部署命令：详尽的 shell 脚本，包括环境变量设置（如 SGLANG_SET_CPU_AFFINITY、HCCL_BUFFSIZE）和启动命令（使用 sglang.launch_server 带推测解码参数）。
- 基准测试命令：使用 sglang.bench_serving 进行性能测试。

关键代码片段示例（从 patch 摘录）：

```
export SGLANG_SET_CPU_AFFINITY=1
unset https_proxy
source /usr/local/Ascend/ascend-toolkit/set_env.sh
python -m sglang.launch_server --model-path $MODEL_PATH --host 127.0.0.1 --port 7339 --
```

```
attention-backend ascend --device npu --quantization modelslim --max-running-requests 16 --
speculative-algorithm EAGLE3 --speculative-draft-model-path xxx
```

评论区精华

review 讨论由 gemini-code-assist[bot] 主导，聚焦于文档风格改进：

- 硬件描述格式：建议添加空格，如将 "2Card" 改为 "2 Cards"，提升可读性。
- 占位符替换：建议将 MODEL_PATH=xxx 和 --speculative-draft-model-path xxx 中的 "xxx" 替换为更描述性的路径（如 /path/to/your/model），避免用户混淆。
- 未使用变量：指出 LOCAL_HOST1 和 LOCAL_HOST2 变量被定义但未使用，建议移除以减少噪音。

> gemini-code-assist[bot] 评论："For better readability, please add a space between the number and 'Card'."

讨论无技术争议，结论是建议被考虑，PR 最终由 sglang-npu-bot 批准合并。

风险与影响

风险分析：

- 文档误导风险：如果占位符未正确替换或命令有误，可能导致用户部署失败。例如，MODEL_PATH=xxx 需要用户自行填充实际路径。
- 环境依赖性：部署命令依赖于特定 NPU 驱动和环境变量，若系统配置不同可能需要调整。
- 低技术风险：无代码变更，因此无回归、性能或安全风险。

影响分析：

- 用户影响：直接受益，提供了现成的优化配置，降低部署门槛，尤其针对低延迟场景。
- 系统影响：无，纯文档更新不影响运行时行为。
- 团队影响：文档维护增强，支持 NPU 平台持续改进，可能减少后续支持成本。

关联脉络

从近期历史 PR 看，本 PR 是 NPU 平台文档演进的一部分：

- PR 22029（[NPU][CI] Use UV to improve pip install speed）：同属 NPU 改进，优化 CI 效率，而本 PR 补充文档，共同提升 NPU 生态。
- PR 22353（[SKILL] add torch profiler analysis workflow）：同为文档添加，引入新工作流程，反映仓库对文档维护的重视趋势。

整体上，sglang 仓库近期在 NPU 和文档领域有持续投入，本 PR 是这一脉络的自然延伸，旨在通过标准化文档支持更广泛的硬件和模型优化。