# PR #27353 完整报告

sgl-project/sglang

Update best practice for qwen3-next-80b-a3b-instruct

合并时间：2026-06-05 17:02

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/27353

## 执行摘要

- 一句话：更新 Qwen3-Next-80B-A3B 的 NPU 最佳实践文档
- 推荐动作：该 PR 内容清晰，文档质量较好，推荐用户参考其中的配置部署 Qwen3-Next-80B-A3B 模型。后续可关注 SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK 拼写是否修正，避免用户误用。

## 功能与动机

为最新的 Qwen3-Next-80B-A3B-Instruct 模型提供经过验证的 NPU 部署配置，帮助用户在高并发场景下获得最佳性能。PR body 中包含了详细的 Serving Benchmark 结果，表明该配置在 300 并发下达到了 2.69 req/s 的请求吞吐和 4037 tok/s 的输出吞吐。

## 实现拆解

1. 在 ascend_npu_best_practice.mdx 文件中，更新了服务器启动命令部分，针对 Qwen3-Next-80B-A3B-Instruct-W8A8 模型调整了环境变量和参数。
2. 新增了多项环境变量，如 DEEP_NORMAL_MODE_USE_INT8_QUANT=1、ASCEND_USE_FIA=1、SGLANG_NPU_USE_MULTI_STREAM=0、SGLANG_ENABLE_SPEC_V2=1、SGLANG_ENABLE_OVERLAP_PLAN_STREAM=1、FORCE_DRAFT_MODEL_NON_QUANT=1，以及多个 Zero Bubble 相关变量。
3. 删除了不再使用的变量，如 SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=20 调整为了 330，HCCL_BUFFSIZE 从 2000 降为 64。
4. 启动命令增加了 --speculative-algorithm NEXTN、--dp-size 2、--enable-dp-attention、--enable-dp-lm-head 等新参数，并调整了 --max-running-requests 为 300、--mem-fraction-static 为 0.75、--page-size 为 128。
5. 更新了 Benchmark 测试命令和结果数据，展示了该配置下的性能表现。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_best_practice.mdx（模块 文档；类别 other；类型 core-logic）：单一变更文件，更新了 Qwen3-Next-80B-A3B-Instruct 模型在 Ascend NPU 上的最佳实践环境变量和启动命令。

关键符号：未识别

## 关键源码片段

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

单一变更文件，更新了 Qwen3-Next-80B-A3B-Instruct 模型在 Ascend NPU 上的最佳实践环境变量和启动命令。

```
# 针对 Qwen3-Next-80B-A3B-Instruct 模型的 Ascend NPU 最佳实践配置
# 推荐在 `modelslim` 量化 + 推测解码 + 数据并行下使用

# 环境变量设置
export DEEP_NORMAL_MODE_USE_INT8_QUANT=1
export SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=330
export ASCEND_USE_FIA=1
export SGLANG_NPU_USE_MULTI_STREAM=0
export SGLANG_WARMUP_TIMEOUT=3600
export SGLANG_ENABLE_SPEC_V2=1
export SGLANG_ENABLE_OVERLAP_PLAN_STREAM=1
export FORCE_DRAFT_MODEL_NON_QUANT=1

# Zero Bubble 通信优化（减少 TP 通信等待）
export HCCL_BUFFSIZE=64
export SGLANG_ZBAL_LOCAL_MEM_SIZE=59648
export SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK=0 # 注意：原变量名为
INBALANCE，标准拼写应为 IMBALANCE
export SGLANG_ZBAL_BOOTSTRAP_URL="tcp://127.0.0.1:24669"

# 显存分配
export PYTORCH_NPU_ALLOC_CONF=expandable_segments:True
export ZBAL_NPU_ALLOC_CONF=use_vmm_for_static_memory:True
export ZBAL_ENABLE_GRAPH=1

MODEL_PATH=/home/weights/Qwen3-Next-80B-A3B-Instruct-W8A8

python3 -m sglang.launch_server --model-path ${MODEL_PATH} \
  --page-size 128 \
  --tp-size 4 \
  --trust-remote-code \
  --attention-backend ascend \
  --device npu \
  --watchdog-timeout 9000 \
  --host 127.0.0.1 --port 6699 \
  --mem-fraction-static 0.75 \
  --disable-radix-cache --max-prefill-tokens 14080 --context-length 26384 \
  --chunked-prefill-size -1 --max-running-requests 300 \
  --mamba-ssm-dtype bfloat16 \
  --quantization modelslim \
  --speculative-algorithm NEXTN --speculative-num-steps 3 \
  --speculative-eagle-topk 1 --speculative-num-draft-tokens 4 \
  --speculative-draft-model-quantization unquant \
```

```
--speculative-draft-model-path /home/weights/Qwen3-Next-80B-A3B-Instruct \
--dp-size 2 --enable-dp-attention --enable-dp-lm-head \
--moe-a2a-backend deepep --deepep-mode auto \
--cuda-graph-bs 1 2 3 4 5 6 7 8 10 12 14 16 18 20 22 24 26 28 30 32 40 44 48 52 56 60 64
72 80 88 96 104 112 120 128 136 144 150
```

## 评论区精华

Review 过程中，gemini-code-assist[bot] 指出环境变量 SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK 存在拼写错误，应为 SGLANG_ENABLE_TP_MEMORY_IMBALANCE_CHECK（缺少字母 M）。该评论被正确标注为 medium 优先级，但最终合并的代码中似乎并未修正该拼写（从 patch 看仍为 INBALANCE），需确认是否后续修复或该变量名本身允许此写法。

- 环境变量拼写错误 (other): 该评论未被采纳，合并的代码仍保留了 INBALANCE 的拼写。

## 风险与影响

- 风险：低风险。本次变更是纯文档更新，不涉及任何源码或运行时逻辑。唯一潜在风险是用户直接复制文档中的拼写错误的变量名可能导致环境变量不生效，但不会引起崩溃或安全问题。
- 影响：影响范围：仅影响 Ascend NPU 平台用户中部署 Qwen3-Next-80B-A3B-Instruct 模型的场景。 影响程度：中等，为用户提供了明确的性能最佳配置，有助于提升该模型在 NPU 上的推理效率。 用户影响：用户可直接参考文档中的命令启动服务，减少调优成本。
- 风险标记：拼写错误未修复

## 关联脉络

- PR #27352 [AMD] fix(ci): run partition 3 of stage-c-test-large-8-gpu-amd: 同为基础设施 / 文档类 PR，但具体内容无关。
- PR #27032 [NPU] add GLM model best practice docs: 同一位作者 McZyWu 之前提交的 NPU 最佳实践文档 PR，属于同一文档系列。