# PR #29492 完整报告

sgl-project/sglang

[NPU] update best practicce docs from testcase

合并时间：2026-06-29 11:30

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29492

## 执行摘要

- 一句话：更新 Ascend NPU 最佳实践文档，与实际测试配置保持一致
- 推荐动作：推荐 NPU 平台用户和运维团队精读此 PR 的文档变更，尤其是新增的 MiMo-V2-Flash 指南。对于其他读者，可快速了解团队如何将实测配置反哺文档，体现文档与代码对齐的工程文化。

## 功能与动机

PR 标题和 body 指出更新最佳实践文档，以反映实际测试案例中的经验和验证结果，确保用户获得准确、可用的部署配置。

## 实现拆解

1. 新增 MiMo-V2-Flash 文档：创建 mimo_v2_flash.mdx，提供低延迟和高吞吐两种场景下的配置表、环境变量设置、启动命令及 benchmark 示例。
2. 更新已有模型文档的启动参数：在 Qwen3.6-27B、Kimi-K2.6、GLM-5.1、Qwen3.5-397B 等文档中统一追加 --reasoning-parser qwen3 和 --tool-call-parser qwen3_coder 参数；调整 --max-prefill-tokens、--mem-fraction-static、--cuda-graph-bs 等配置以匹配测试结果。
3. 调整 benchmark 参数：更新 --max-concurrency、--num-prompts 等 benchmark 命令行参数，使压力测试更贴近实际负载。
4. 清理过时环境变量：移除多文档中已不再需要的 SGLANG_NPU_PROFILING、SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK、ZBAL_* 等环境变量，减少用户困惑。
5. 修复格式与命名规范：根据 review 建议增加换行、标准化模型名称写法（如 MiMo-V2-Flash → mimo_v2_flash 统一格式），修正量化类型标识。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/best_practice/mimo_v2_flash.mdx（模块 NPU 文档；类别 docs；类型 documentation）：新增文件，包含 MiMo-V2-Flash 模型在 Ascend NPU 上的完整最佳实践指南，包括多种部署模式的配置表和命令示例。
- docs_new/docs/hardware-platforms/ascend-npus/best_practice/qwen3_6_27b.mdx（模块 NPU 文档；类别 docs；类型 documentation）：修改量最大（+66/-62），涉及启动参数和环境变量的大幅更新，包括追加 reasoning-parser、更新量化配置和 benchmark 参数。

- docs_new/docs/hardware-platforms/ascend-npus/best_practice/kimi_k2_6.mdx（模块 NPU 文档；类别 docs；类型 documentation）：修改量大（+68/-51），表格式结构调整（新增 TTFT 列），更新多个部署模式的 benchmark 配置。
- docs_new/docs/hardware-platforms/ascend-npus/best_practice/glm5_1.mdx（模块 NPU 文档；类别 docs；类型 documentation）：修改量较大（+58/-21），新增 TTFT 列并更新具体数值（如 TPOT 从 20ms 改为 55.2ms），增加 benchmark 的 max-concurrency 调整。
- docs_new/docs/hardware-platforms/ascend-npus/best_practice/qwen3_5_397b.mdx（模块 NPU 文档；类别 docs；类型 documentation）：修改量中等（+37/-30），追加推理和工具调用解析器参数，调整 cuda-graph-bs 和 mem-fraction-static 等配置。
- docs_new/docs/hardware-platforms/ascend-npus/best_practice/qwen3_6_35b_a3b.mdx（模块 NPU 文档；类别 docs；类型 documentation）：修改量较小（+34/-13），但新增 --max-total-tokens 和推理参数，更新 mem-fraction-static 和 cuda-graph-bs。
- docs_new/docs/hardware-platforms/ascend-npus/best_practice/deepseek_r1.mdx（模块 NPU 文档；类别 docs；类型 documentation）：修改量较大（+36/-10），更新环境变量和启动参数，追加推理解析器。

关键符号：未识别

## 关键源码片段

### docs_new/docs/hardware-platforms/ascend-npus/best_practice/mimo_v2_flash.mdx

新增文件，包含 MiMo-V2-Flash 模型在 Ascend NPU 上的完整最佳实践指南，包括多种部署模式的配置表和命令示例。

```
# ============================================================
# Before running, update the following variables:
# P_IP: prefill node IP address
# D_IP: decode node IP address
# ASCEND_MF_STORE_URL: prefill node IP with port
# MODEL_PATH: path to the model weights directory
# HCCL_SOCKET_IFNAME: network interface name for HCCL
# GLOO_SOCKET_IFNAME: network interface name for Gloo
# ============================================================

echo performance | tee /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor
sysctl -w vm.swappiness=0
sysctl -w kernel.numa_balancing=0
sysctl -w kernel.sched_migration_cost_ns=50000

unset https_proxy
unset http_proxy
unset HTTPS_PROXY
unset HTTP_PROXY
unset ASCEND_LAUNCH_BLOCKING
```

```
source /usr/local/Ascend/ascend-toolkit/set_env.sh
source /usr/local/Ascend/nnal/atb/set_env.sh

export ASCEND_USE_FIA=1
export DEEPEP_NORMAL_LONG_SEQ_PER_ROUND_TOKENS=3584
export DEEPEP_NORMAL_LONG_SEQ_ROUND=32
export DEEP_NORMAL_MODE_USE_INT8_QUANT=1
export HCCL_CONNECT_TIMEOUT=1800
export HCCL_OP_EXPANSION_MODE=AIV
export PYTORCH_NPU_ALLOC_CONF=expandable_segments:True
export SGLANG_DEEPEP_BF16_DISPATCH=0
export SGLANG_DISAGGREGATION_BOOTSTRAP_TIMEOUT=3600
export SGLANG_DISAGGREGATION_WAITING_TIMEOUT=3600
export SGLANG_ENABLE_OVERLAP_PLAN_STREAM=0
export SGLANG_ENABLE_SPEC_V2=1
export SGLANG_SET_CPU_AFFINITY=1
```

## 评论区精华

Review 中 amote-i 提出了若干质量检查问题：(1) 确认新增的某个环境变量是否必要；(2) 检查量化类型是 BF16 还是 W8A8；(3) 模型名称格式应标准化；(4) 询问移除 --cuda-graph-bs 参数是否必要。作者 Hide-on-bushsh 逐一回应：环境变量经测试确认需要；量化类型已修正；名称格式已统一；移除的 cuda-graph-bs 在测试中未使用。此外 amote-i 还要求添加换行分隔，作者已修复。整体上讨论集中在文档严谨性和配置准确性，未产生重大分歧。

- 新增环境变量是否必需 (question): 作者确认该环境变量在测试中需要，保留。
- 模型名称格式标准化 (style): 作者将模型名称统一为规范格式。
- 量化类型标识正确性 (correctness): 作者修正了量化类型标识。
- 移除 --cuda-graph-bs 参数的必要性 (question): 作者答复测试中未使用该参数，因此移除。
- 某些配置删除是否影响文档准确性 (question): 作者确认测试后无问题。
- 添加换行分隔提升可读性 (style): 作者已添加换行。

## 风险与影响

- 风险：文档变更不直接影响代码运行，但存在以下风险：(1) 新增的 --reasoning-parser qwen3 参数若用户在其他模型上误用可能导致解析异常；(2) HCCL_BUFFSIZE 等环境变量调整在不同集群规模下可能影响通信性能；(3) 文档中部分 benchmark 配置（如 --max-concurrency 132）可能不适用于所有硬件规格。建议用户在非生产环境先验证配置。
- 影响：用户影响：为 Ascend NPU 用户提供了更准确、可复现的部署指导，降低试错成本。团队影响：减少了针对 NPU 配置的重复咨询，文档维护成本略有增加（需随测试更新）。影响范围：限 NPU 平台用户，影响力中等。
- 风险标记：配置参数可能因硬件变体失效，解析器参数仅适用于特定模型，环境变量调整可能影响通信性能

## 关联脉络

- PR #29029 [NPU][Bugfix] Fix a ModelSlim loading failure: 同是 NPU 平台的修复，与当前文档更新共同服务于 NPU 用户稳定性。
- PR #29378 [CPU] enable fused_sigmoid_mul on CPU device: 类似的硬件平台文档 / 代码更新，反映跨平台优化方向。