PR #36476 完整报告

sgl-project/sglang

[NPU] [DOC] update npu best practice

合并时间：2026-08-29 09:38

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36476

执行摘要

- 一句话：更新 NPU 最佳实践文档，新增 V4-Flash 教程与 GLM-5.2 指南
- 推荐动作：值得精读其中两个新页面（DeepSeek-V4-Flash best-practices 与 tutorial），尤其适合负责昇腾部署的工程师与文档维护者。值得学习的设计决策：以实测数据为准修正文档、删除无法支撑的性能声称；把环境变量收敛为顶层导出并去掉调试开关；在教程里提供 8-die 最小验证配置降低入门门槛；以及用 bot 自动化合并的文档流水线。

功能与动机

PR body 只有一句 'update npu best practice'，没有附带 issue。从提交历史看，主要动机有两层：一是为新支持的模型补齐官方验证配置（DeepSeek-V4-Flash 的 8/16 卡 W8A8、GLM-5.2 的 W4A8 都标注了 SGLang v0.5.16 验证）；二是修正既有最佳实践与实测结果的偏差，例如 GLM-5.1 的 PREFIX100 条目声称 100% 前缀缓存命中率但实测用 RANDOM 数据集，基准命令中 --seed 在 GSP 分支并不会被转发，以及 minimax_m2_5 里 --max-prefill-token 拼写错误，这些在提交信息中都有明确说明。

实现拆解

1. 新增 DeepSeek-V4-Flash 文档：best-practices/deepseek_v4_flash.mdx（+486 行）给出 16 卡 PD 分离（8k+1k、50ms TPOT）与 8 卡 PD Mixed 三组 W8A8 配置及完整启动脚本，脚本引入 DEEP_NORMAL_MODE_USE_INT8_QUANT、SGLANG_ENABLE_SPEC_V2、SGLANG_DSV4_FP4_EXPERTS 等模型专用开关；tutorials/deepseek_v4_flash.mdx（+285 行）提供功能表、特性兼容链接与多节点 PD 分离脚本，后续提交（f0ccae3）把最小示例从 16-die 降为 8-die，并对齐到 OpenAI 兼容后端，还补充了厂商 OPP/custom_transformer 环境的 set_env sourcing（24ec2c2）。
2. 新增 GLM-5.2 并整修 GLM-5.1：glm_5_2.mdx（+226 行）提供 3P1D 32 卡 W4A8 的 PD 分离脚本，首次使用多 prefill 节点 IP 数组（P_IP/D_IP）；glm_5_1.mdx（+78/-272 行）按评审意见删除与 RANDOM 数据集实测矛盾、无法支撑 100% 前缀缓存命中率声称的 48 卡 PREFIX100 章节，将 HCCL_BUFFSIZE 改成 deepep 的 DEEPEP_HCCL_BUFFSIZE，去掉 ENABLE_PROFILING 调试变量，加入 SGLANG_ZBAL_LOCAL_MEM_SIZE 等 ZBAL 内存均衡变量。
3. 修订 Qwen3.6-35B-A3B、Qwen3.6-27B、Qwen3.5-397B（改名 -A17B）等既有页面：调整 SGLANG_PREFILL_DELAYER_MAX_DELAY_PASSES（30→删除、150→300）、max-prefill-tokens、cuda-graph-bs 序列与 mem-fraction-static 取值；bench_serving 命

令统一补充实测使用的 --temperature/--top-p 与 --seed 1，并提示离线环境需要 --dataset-path。

4. 全库约定统一：ASCEND_MF_STORE_URL 收敛到单一顶层导出；去掉重复的 --disaggregation-bootstrap-port、硬编码数据集路径与调试环境变量；eab03eb 提交把 best-practices 导航顺序与 model tutorials 对齐；c37a7bb 修复 minimax_m2_5 的 --max-prefill-token 拼写。
5. 测试与 CI：纯文档变更，无源码或单元测试配套；通过 /tag-and-rerun-ci 重跑 CI 后由 sglang-npu-bot 批准合入。

关键文件：

- docs/docs/hardware-platforms/ascend-npus/model-deployment/best-practices/deepseek_v4_flash.mdx（模块 NPU 文档；类别 docs；类型 documentation）：本 PR 的核心新增页：给出 V4-Flash 在 Atlas 800I A3 上 16 卡 PD 分离、8 卡 PD Mixed 的实测 W8A8 配置与完整启动脚本，承载了大部分新环境变量开关。
- docs/docs/hardware-platforms/ascend-npus/model-deployment/tutorials/deepseek_v4_flash.mdx（模块 NPU 文档；类别 docs；类型 documentation）：新增 V4-Flash 模型部署教程：给出模型特性、功能支持表（dsv4 attention backend、EAGLE 投机解码、PD 分离、modelslim 量化）、环境准备与 8-die 最小验证配置，是用户落地的主要参考入口。
- docs/docs/hardware-platforms/ascend-npus/model-deployment/best-practices/glm_5_2.mdx（模块 NPU 文档；类别 docs；类型 documentation）：本次新增的另一个新模型配置页：提供 GLM-5.2 在 32 卡 W4A8 下的 3P1D PD 分离部署脚本，首次使用多 prefill 节点 IP 数组的写法。
- docs/docs/hardware-platforms/ascend-npus/model-deployment/best-practices/glm_5_1.mdx（模块 NPU 文档；类别 docs；类型 documentation）：评审意见落地最集中的文件：删除与实测矛盾的 PREFIX100 章节，围绕 deepep/ZBAL 更新环境变量，并统一基准命令 flag。

关键符号：未识别

关键源码片段

docs/docs/hardware-platforms/ascend-npus/model-deployment/best-practices/deepseek_v4_flash.mdx

本 PR 的核心新增页：给出 V4-Flash 在 Atlas 800I A3 上 16 卡 PD 分离、8 卡 PD Mixed 的实测 W8A8 配置与完整启动脚本，承载了大部分新环境变量开关。

```
# best-practices/deepseek_v4_flash.mdx 中 PD 分离部署脚本的前置环境准备
echo performance | tee /sys/devices/system/cpu/cpu*/cpufreq/scaling_governor
sysctl -w vm.swappiness=0
sysctl -w kernel.numa_balancing=0

# 清理代理与调试变量，避免影响 HCCL 集合通信
unset https_proxy http_proxy HTTPS_PROXY HTTP_PROXY ASCEND_LAUNCH_BLOCKING

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

```
source /usr/local/Ascend/nnal/atb/set_env.sh
# 厂商定制算子环境（OPP / custom_transformer），后续提交补充
source /usr/local/Ascend/ascend-toolkit/latest/opp/vendors/customize/bin/set_env.bash
source /usr/local/Ascend/ascend-toolkit/latest/opp/vendors/custom_transformer/bin/set_env.
bash

# 模型专用开关：INT8 激活量化 + 草稿模型非量化 + 关闭 INF NAN 检查
export DEEP_NORMAL_MODE_USE_INT8_QUANT=1
export FORCE_DRAFT_MODEL_NON_QUANT=1
export INF_NAN_MODE_FORCE_DISABLE=1
# 开启 spec v2 与 plan-stream 重叠；关闭无收益的 FP4 专家路径
export SGLANG_ENABLE_SPEC_V2=1
export SGLANG_ENABLE_OVERLAP_PLAN_STREAM=1
export SGLANG_DSV4_FP4_EXPERTS=False
```

docs/docs/hardware-platforms/ascend-npus/model-deployment/best-practices/glm_5_1.mdx

评审意见落地最集中的文件：删除与实测矛盾的 PREFIX100 章节，围绕 deepep/ZBAL 更新环境变量，并统一基准命令 flag。

```
# best-practices/glm_5_1.mdx 中按评审意见修正后的 prefill 节点环境变量
export DEEPEP_HCCL_BUFFSIZE=1200 # deepep 通信缓冲，替换旧的 HCCL_BUFFSIZE
export GLOO_SOCKET_IFNAME=<网络接口>
export HCCL_SOCKET_IFNAME=<网络接口>
# 关闭 TP 显存不均衡检查，并显式配置 ZBAL 本机内存
export SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK=0
export SGLANG_ZBAL_BOOTSTRAP_URL=tcp://127.0.0.1:24699
export SGLANG_ZBAL_LOCAL_MEM_SIZE=61184
export ZBAL_ENABLE_GRAPH=1
```

评论区精华

公开 review 区无实质讨论线程，唯一交互是 sglang-npu-bot 的 /tag-and-rerun-ci 命令。实质评审发生在提交迭代中：b3e7a82、de9601d 两条提交明确回应了内部评审意见，例如删除 GLM-5.1 PREFIX100 矛盾条目、去掉调试环境变量 ENABLE_PROFILING、清理重复的 bootstrap 端口参数和未使用的 HCCL/GLOO_SOCKET_IFNAME 行。最终 sglang-npu-bot APPROVED 后合入。

- CI rerun 命令 /tag-and-rerun-ci (other): CI 通过后 sglang-npu-bot APPROVED 并合入 PR；无代码相关讨论。

风险与影响

- 风险：纯文档变更，无代码回归风险。主要风险在内容层面： 1) 大量环境变量（DEEP_NORMAL_MODE_USE_INT8_QUANT、DEEPEP_HCCL_BUFFSIZE、SGLANG_ZBAL_*）与特定 NPU 固件、SGLang 版本绑定，后续版本升级或功能改名会让文档失效，页面虽然标注了 v0.5.16，但缺少过期检查机制； 2) 脚本中的 IP、端口、SGLANG_ZBAL_LOCAL_MEM_SIZE 等数值为具体环境调优值，用户复制到其他集群可能

会得到偏差结果； 3) 删除 PREFIX100 章节虽解决了矛盾，但同时也丢掉了 48 卡高吞吐配置的参考。建议后续维护时引入 markdownlint/link-check 对锚点与命令一致性做定期校验。

- 影响：影响范围限定在 Ascend NPU 用户与文档维护者：为使用 DeepSeek-V4-Flash、GLM-5.2 的用户提供了开箱即用的验证配置，降低了多节点 PD 分离与 EAGLE 投机解码场景的试错成本；对文档体系确立了统一约定，后续新增模型 best-practice 可复用同一脚本骨架。对系统的运行时性能、训练 / 推理代码无任何影响。
- 风险标记：纯文档变更，参数易随版本失效，配置与实测环境绑定，缺少自动化校验

关联脉络

- PR #36828 [AMD] Update v4 amd cookbook 0828: 与本次 NPU 侧更新 DeepSeek-V4-Flash 部署配置相呼应，形成 AMD/NPU 双平台同模型文档对齐；两 PR 均属模型 cookbook/best-practice 文档线。