PR #21468 完整报告

sgl-project/sglang

[NPU] Update DeepSeek-V3.2 model deployment instructions in documentation

合并时间：2026-03-30 15:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/21468

执行摘要

此 PR 更新了 DeepSeek-V3.2 模型在 Ascend NPU 上的部署最佳实践文档，基于新性能测试调整了配置参数和脚本，影响仅限于文档用户，无代码变更风险。

功能与动机

更新动机源于主分支启用了 atten-cp-size 函数并重新调优了 DeepSeek-V3.2 模型在长序列场景下的性能，如 PR body 所述：'Updates the documentation for deploying the DeepSeek-V3.2 model on Ascend NPU'，旨在提供最新的部署指南。

实现拆解

修改文件为 docs/platforms/ascend/ascend_npu_best_practice.md，具体改动包括：

- 配置表更新：将 DeepSeek-V3.2-Exp 改为 DeepSeek-V3.2，输入输出长度从 64K+3K 调整为 128K+1K，TPOT 时间从 30ms 缩短到 20ms。
- 部署脚本优化：
  - 分离预填充和解码实例的 IP 数组（P_IP 和 D_IP）。
  - 移除旧环境变量，如 ASCEND_HOME_PATH 和 HCCL_BUFFSIZE。
  - 更新 sglang.launch_server 参数以匹配新配置。

评论区精华

review 讨论主要关注脚本正确性和设计意图：

- gemini-code-assist[bot]指出：'HCCL_BUFFSIZE 变量被覆盖，导致冗余。' MichelleWu351回复：'already reduce redundant code. done'，解决了此问题。
- gemini-code-assist[bot]建议：' 占位符 P_IP1 和 D_IP1 未定义，应使用数组元素。' MichelleWu351解释：'P_IP1 and D_IP1 should be filled in by the user manually. And this part of code is a router script, not the rest of prefill or decode script.'，澄清了设计选择。

风险与影响

风险：

- 文档脚本中的占位符未明确定义，可能导致用户部署错误。
- 环境变量调整可能与其他 NPU 配置冲突，需用户确认兼容性。

影响：

- 直接影响 Ascend NPU 用户部署 DeepSeek-V3.2 模型的效率和正确性。
- 无系统级影响，仅为文档更新。

关联脉络

与历史 PR 的关联：

- PR #18461：'[Intel GPU] Enable DeepSeek R1 inference on XPU'，同样是 DeepSeek 模型的部署优化，但平台不同（XPU vs NPU），体现了跨平台模型支持的一致性更新趋势。
- 近期文档 PR 如 #21659 也更新了贡献指南，显示团队对文档维护的重视。