PR #22975 完整报告

sgl-project/sglang

[NPU] [DOC] Update npu best practice docs to match latest code

合并时间：2026-04-16 20:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22975

执行摘要

- 一句话：更新 Ascend NPU 最佳实践文档，同步最新配置和性能数据。
- 推荐动作：建议 NPU 平台用户和开发者关注此文档更新，特别是配置参数和性能指标的变化。对于工程团队，可注意 review 中提到的标准化问题，考虑在未来统一环境变量命名和移除已弃用标志，以提升文档一致性。

功能与动机

根据 PR body 描述，动机是“Update npu best practice docs to match latest code”，即更新 NPU 最佳实践文档以匹配最新代码，确保文档准确性。

实现拆解

1. 更新现有配置：修改 docs/platforms/ascend/ascend_npu_best_practice.md 文件中多个模型的性能指标和配置参数，例如将 DeepSeek-V3.2 的延迟从 20ms 更新为 26ms，调整 Qwen3-32B 的延迟从 12ms 改为 6ms 等。
2. 添加新模型配置：为 Qwen3-14B 等新模型添加完整的配置章节，包括环境变量和启动命令。
3. 修正 review 反馈的问题：根据 review 评论，调整了环境变量名（如 DP_ROUND_ROBIN 未改为 SGLANG_DP_ROUND_ROBIN，以保持与实际使用一致）和移除冗余参数（如 --speculative-draft-model-quantization unquant），但未采纳关于已弃用标志 --prefill-round-robin-balance 的移除建议。
4. 无测试或代码配套改动：本次变更仅涉及文档，没有源码、测试或配置文件的联动修改。

关键文件：

- docs/platforms/ascend/ascend_npu_best_practice.md（模块 平台文档；类别 docs；类型 documentation）：唯一变更文件，包含 NPU 最佳实践的全部配置和性能数据更新，是用户部署的关键参考文档。

关键符号：未识别

关键源码片段

docs/platforms/ascend/ascend_npu_best_practice.md

唯一变更文件，包含 NPU 最佳实践的全部配置和性能数据更新，是用户部署的关键参考文档。

以下片段展示了文档中一个典型的环境变量配置块，反映了本次更新中的参数调整：

```
# 环境变量设置示例（用于Qwen3-32B模型）
export SGLANG_ENABLE_OVERLAP_PLAN_STREAM=1 # 启用重叠计划和流式处理
export SGLANG_ENABLE_SPEC_V2=1 # 启用推测解码v2
export HCCL_BUFFSIZE=650 # 调整HCCL缓冲区大小以优化通信
export SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=16 #
增加每个rank的最大分发token数，提升吞吐
# 注意：DP_ROUND_ROBIN 环境变量在此处使用，但review建议改为 SGLANG_DP_ROUND_ROBIN
以保持项目一致性
export DP_ROUND_ROBIN=1 # 启用DP轮询负载均衡，根据实际代码使用保留此命名
```

评论区精华

review 中主要讨论点包括：

- 已弃用标志：gemini-code-assist[bot] 指出 --prefill-round-robin-balance 标志已弃用，建议从文档中移除，但作者未回应此建议，标志仍保留。
- 环境变量名拼写：gemini-code-assist[bot] 建议将 DP_ROUND_ROBIN 改为 SGLANG_DP_ROUND_ROBIN 以保持一致性，作者回复“Consistent with the actual case”，未作修改。
- 冗余配置参数：gemini-code-assist[bot] 建议移除冗余的 --speculative-draft-model-quantization unquant 参数，作者同样回复“Consistent with the actual case”，未作修改。 结论：作者坚持文档与实际代码使用情况一致，未完全采纳 review 中的标准化建议，可能存在文档与代码约定不一致的风险。
- 已弃用标志 --prefill-round-robin-balance 的处理 (design): 作者未回应或采纳此建议，标志仍保留在文档中。
- 环境变量名不一致和冗余参数 (correctness): 作者回复“Consistent with the actual case”，未作修改，坚持文档与实际使用情况一致。

风险与影响

- 风险：风险较低，但存在以下潜在问题：
- 配置过时风险：文档中仍包含已弃用标志（如 --prefill-round-robin-balance），可能导致用户混淆或错误配置。
- 不一致风险：环境变量命名（如 DP_ROUND_ROBIN）与项目其他部分不统一，可能影响用户理解和使用。
- 准确性风险：性能数据和配置参数基于最新代码更新，但若未来代码变更未同步文档，可能导致文档落后。
- 影响：影响范围主要针对使用 Ascend NPU 平台的用户和开发者：
- 用户影响：正面影响，确保用户能基于最新配置获得最佳性能，减少部署错误。
- 系统影响：无直接影响，不涉及代码逻辑变更。
- 团队影响：需确保文档持续与代码同步，避免类似不一致问题积累。
- 风险标记：配置过时，文档不一致

关联脉络

- PR #22923 docs: fix incorrect default max-payload-size in gateway config reference: 同为文档更新 PR，涉及配置修正，可参考文档维护模式。
- PR #20989 [Fix] eagle/eagle3 speculative decoding conflicts with xgrammar in NPU: 涉及 NPU 平台修复，与本 PR 的 NPU 文档更新相关，反映 NPU 功能的持续演进。