PR #28277 完整报告

sgl-project/sglang

[NPU] Docs op performance optimize

合并时间：2026-06-15 20:28

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28277

执行摘要

- 一句话：更新 NPU 文档：新增算子性能优化指南并调整 PD 分离示例
- 推荐动作：建议 PR 阅读者关注 ascend_npu_operator_performance_optimizing.mdx 中 msprof 输出字段的解释，这对性能调优有直接帮助；同时注意 send_request.mdx 中的 import 问题尚未彻底修复，若直接复制代码需自行添加 print_highlight 的 import。PD 分离文档的 MIMO 环境配置可以作为 NPU 部署的参考。

功能与动机

为 SGLang 在 Ascend NPU 上的用户提供更系统化的性能优化指导，包括使用 msprof 工具收集和解读算子性能数据，以及提供最佳实践的环境配置。此外，更新 PD 分离示例以反映当前推荐配置。

实现拆解

1. 新增算子性能优化文档(ascend_npu_operator_performance_optimizing.mdx)：详细解释 msprof 命令的使用方法，并新增一个完整的字段解释表格，涵盖基本标识、时序调度、核心配置等类别，帮助用户理解 DequantSwigluQuant 等融合算子的性能指标。
2. 重构 PD 分离文档(pd_disaggregation.mdx)：
   - 将原来的 Llama Single Node 示例替换为 MIMO Single Node 示例，并拆分为 prefill 和 decode 两个子章节。
   - 补充了完整的环境配置脚本（关闭代理、设置 HCCL 环境变量、CPU 调优等），并更新了启动命令以使用 --enable-mixed-input 等新参数。
   - 删除了 DeepSeek Multi-Node 示例（仅保留空标题，需后续清理）。
3. 修复 send_request 示例(send_request.mdx)：将代码示例中的 print_highlight 替换为内置的 print，以移除对 sglang.utils 中辅助函数的依赖，但未同步更新 import 语句，存在 NameError 风险。
4. 补充调试指南(msprobe_debugging_guide.mdx)：在 TensorBoard 启动说明前增加一句话，指出转储数据可使用 Matplotlib/Excel 进行可视化分析。

关键文件：

- docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_operator_performance_optimizing.mdx（模块 NPU 算子性能；类别 docs；类型 documentation）：新增文件，提

供了详细的 msprof 输出字段说明，是 PR 的核心新增内容，对 NPU 性能优化有重要指导意义。

- docs_new/docs/advanced_features/pd_disaggregation.mdx（模块 PD 分离；类别 docs；类型 documentation）：重构了 PD 分离示例，用 MIMO 单节点配置替代原来的 Llama 单节点，并补充大量环境变量设置，对 NPU 用户部署有指导意义。
- docs_new/docs/basic_usage/send_request.mdx（模块 发送请求；类别 docs；类型 documentation）：修改了 import 语句和函数调用，但存在未完成的替换，是 review 中重点指出的问题。
- docs_new/docs/developer_guide/msprobe_debugging_guide.mdx（模块 调试指南；类别 docs；类型 documentation）：补充了一句话关于数据可视化，增强了调试指南的实用性。

关键符号：未识别

关键源码片段

docs_new/docs/advanced_features/pd_disaggregation.mdx

重构了 PD 分离示例，用 MIMO 单节点配置替代原来的 Llama 单节点，并补充大量环境变量设置，对 NPU 用户部署有指导意义。

```
# 高性能 CPU 调优
# 设置 CPU 调频策略为 performance，降低系统延迟和抖动
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
# 绑定 CPU 核心，减少调度开销
export SGLANG_SET_CPU_AFFINITY=1

# CANN 环境变量来源
source /usr/local/Ascend/ascend-toolkit/set_env.sh

# 关键性能环境变量
# STREAMS_PER_DEVICE 控制设备流数量，默认值 32
export STREAMS_PER_DEVICE=32
# HCCL 缓冲区大小，影响通信性能
export HCCL_BUFFSIZE=1600
# 启用 Speculative Decoding V2
export SGLANG_ENABLE_SPEC_V2=1

# 启动 prefill 节点
python3 -m sglang.launch_server \
  --model-path /path/to/model \
  --disaggregation-mode prefill \
  --port 30000 \
  --enable-mixed-input
```

评论区精华

Review 作者 gemini-code-assist[bot] 指出了三个问题：

1) send_request.mdx 中删除了 print_highlight 的 import 但代码仍调用了该函数，会导致 NameError； 2) pd_disaggregation.mdx 中 Mimo 应改为全大写 MIMO 以保持术语一致性； 3) 删除 DeepSeek Multi-Node 后留下了空代码块，建议移除对应标题。测试人员 amote-i 还指出应避免硬编码 IP 地址。作者均回复接受建议并进行了相应修改。

- send_request.mdx 中 print_highlight 函数导入与调用不匹配 (correctness): 作者回复 'thanks'，表示接受，后续可能需要补充修复。
- Mimo 应改为 MIMO 以保持术语一致 (style): 作者回复 'THANKS'，已修改。
- 删除 DeepSeek Multi-Node 后遗留空标题和代码块 (other): 作者回复 'OK'，已移除。

风险与影响

- 风险：主要风险在于 send_request.mdx 中的 import 修改不完整：from sglang.utils import wait_for_server, terminate_process 移除了 print_highlight，但第 49 和 136 行仍调用了 print_highlight(response)，任何直接运行该代码的用户都会遭遇 NameError。该问题在 review 中被指出，但最终 commit 是否修复需确认（从 commit 历史看后续有 lint checks，但未明确提及修复）。其他变更风险较低，主要是文档准确性和一致性。
- 影响：影响范围限定于阅读 Ascend NPU 文档的用户：性能优化指南提供了更细致的 msprof 分析指导，PD 分离示例更新使配置更贴近实际部署。send_request 示例的 import 错误会影响按文档操作的用户。整体影响程度中等偏低，不涉及任何代码逻辑变更。
- 风险标记：导入不完整导致 NameError，文档一致性风险

关联脉络

- PR #28283 Fix inaccuracies and add NPU constraints in ascend_npu_profiling.mdx.: 同为 NPU profiling 文档更新，与本 PR 新增的算子性能优化指南共同构成性能分析文档体系。
- PR #28284 Update documentation for Ascend NPU Guide: 涉及 NPU 性能测试、快速开始等多个文档，与本 PR 的文档更新范围有重叠，体现 NPU 文档持续维护。