

PR #28283 完整报告

sgl-project/sglang

Fix inaccuracies and add NPU constraints in ascend_npu_profiling.mdx.

合并时间: 2026-06-15 21:12

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/28283>

PR #28283 分析报告: 修正 ascend_npu_profiling.mdx 中的不准确之处并补充 NPU 约束

执行摘要

该 PR 仅修改了一个文档文件 `ascend_npu_profiling.mdx`, 修正了其中关于 Ascend NPU 性能分析的多处表述不准确, 并补充了 NPU 平台特有的约束条件。变更包括移动 Note 卡片、澄清参数优先级、说明活动类型支持范围、纠正目录路径描述、修复 bash 代码块中未注释文本以及分布式合并限制的说明。这是一次典型的文档改进 PR, 无代码变更, 对 NPU 用户有实际帮助。

功能与动机

PR body 未详细说明具体动机, 但从变更内容看, 旨在修复 Ascend NPU 性能分析文档中的不准确之处, 并添加 NPU 平台特有的约束条件, 以确保用户获得正确的操作指导。

实现拆解

该 PR 仅修改了一个文档文件 `docs_new/docs/hardware-platforms/ascend-npus/ascend_npu_profiling.mdx`, 涉及 23 行新增和 15 行删除。具体变更步骤如下:

1. 章节结构调整: 将原位于“参数配置”部分的 Note 卡片 (关于 NPU 自动重定向 CUDA 活动) 移动到“收集方法”部分, 使其更贴合上下文。
2. 澄清参数优先级: 在参数配置表格后新增 Note, 明确“API 参数 > 环境变量 > 默认值”的优先级顺序。
3. 修正活动类型支持说明: 明确 NPU 上仅支持 CPU 和 GPU 活动, MEM 被静默忽略, RPD 会导致错误。
4. 修复bash代码块内未注释文本: 将终端2的说明行前加#, 避免用户直接复制运行时出错。
5. 修正目录路径描述: 修复之前错误的 `<output_dir>/<timestamp>/<output_dir>/...` 路径, 使其正确包含 `<hostname>_<pid>_<timestamp>_ascend_pt/`。
6. 更新分布式合并限制: 修改多节点部署中合并 trace 文件的说明, 指出 NPU 上 `*_ascend_pt` 格式支持有限, 应检查各节点独立 `trace_view.json`。

该 PR 仅涉及文档文件, 以下为部分变更前后的关键片段:

```
<!-- 在参数配置表格后新增优先级说明 -->
<Note>
```

Priority order: ****API parameter > Environment variable > Default value****.

</Note>

<!-- 修正后的活动类型说明 -->

<td><code>activities</code></td>

<td>

Activity types: CPU, GPU, MEM, RPD. On Ascend NPU, only CPU and GPU are supported.

MEM depends on CUDA memory APIs and will be silently ignored. RPD requires ROCm and will cause an error.

</td>

<td><code>["CPU", "GPU"]</code></td>

<!-- 修复 bash 代码块内未注释文本 --> ```bash Command # Terminal 2: Wait for "Waiting for N steps" output from Terminal 1, then send requests. # The profiler starts recording once /start_profile is received by the server. # Requests sent before the server receives /start_profile will not be captured. curl http://127.0.0.1:30000/generate \ -H "Content-Type: application/json" \ -d '{"text": "Hello", "sampling_params": {"max_new_tokens": 32}}' ```mdx <!-- 修正目录路径，移除重复的 <output_dir> --> <output_dir>/<timestamp>/<hostname>_<pid>_<timestamp>_ascend_pt/. Always check the server log for the exact path: `Profiling done. Traces are saved to: <path>`.

评论区精华

Review 中主要由 gemini-code-assist bot 和 amote-i 提出了 3 个问题：

1. bot: bash 代码块内未注释文本

"The explanation text inside the **bash** code block is not commented out. This will cause syntax errors if a user copies and runs the block." - 作者已为这些行添加了 # 前缀，问题已解决。

2. bot: 目录路径中 <output_dir> 重复

"There is a typo in the directory path where <output_dir> is duplicated: <output_dir>/<timestamp>/<output_dir>/..." - 作者已修正为 <output_dir>/<timestamp>/<hostname>_<pid>_<timestamp>_ascend_pt/, 问题已解决。

3. amote-i: Note 卡片应在代码块外

"Move the Note card out of code block" - 作者回复“Done.”，问题已解决。

所有 review 建议均在合并前得到处理，最终版本已修正所有问题。

风险与影响

- 风险：无技术风险，因为仅涉及文档变更。但如果文档内容仍有错误（例如路径格式或活动类型说明），可能误导用户。不过经过 review 修正后，可信度较高。
- 影响：影响范围限于使用 Ascend NPU 进行性能分析的用户。变更澄清了活动类型限制、参数优先级、目录结构和分布式合并限制，有助于用户避免常见错误。不影响其他模块或平台。

关联脉络

该 PR 与近期其它 NPU 文档 PR（如 #28295、#28284）属于同一轮文档改进工作，共同提升 Ascend NPU 平台的官方文档质量。这些 PR 均专注于修复错误、补充约束，反映了项目对 NPU 硬件支持的持续完善。