PR #22795 完整报告

sgl-project/sglang

[NPU] Offloading docs update

合并时间：2026-04-14 20:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/22795

执行摘要

本 PR 更新了 Ascend NPU 支持功能文档中的卸载参数表格，将多个参数从“计划支持”状态改为实际支持于 A2/A3 平台，并补充了关键使用约束。这是纯文档更新，风险极低，主要影响 NPU 用户的配置准确性，反映了 NPU 卸载功能从规划到实现的演进。

功能与动机

本次更新的动机是修正 Ascend NPU 卸载功能文档的过时信息。原文档中多个卸载参数（如 --offload-group-size）标记为“Planned”（计划支持），现已更新为“A2, A3”平台支持，表明这些功能已实际可用。同时，添加了具体约束：

- --cpu-offload-gb 必须与 --disable-cuda-graph 同时使用。
- --offload-mode 的 sharded_gpu 选项仅支持 DeepSeek 模型，且需搭配 --disable-cuda-graph。 这些修改旨在确保用户文档与最新实现保持一致，避免错误配置。

实现拆解

变更仅涉及一个文件 docs/platforms/ascend/ascend_npu_support_features.md，具体改动如下表所示：

参数	原支持状态	新支持状态	新增约束
--cpu-offload-gb	A2, A3	A2, A3	必须与 --disable-cuda-graph 同时使用
--offload-group-size	Planned	A2, A3	无
--offload-num-in-group	Planned	A2, A3	无
--offload-prefetch-step	Planned	A2, A3	无
--offload-mode	Planned	A2, A3	选项细化：cpu、meta、sharded_gpu；sharded_gpu 需搭配 --disable-cuda-graph 且仅支持 DeepSeek

评论区精华

review 中仅有一次讨论，由 gemini-code-assist[bot] 发起，聚焦文档语法和格式：

> “The phrase '(need used with --disable-cuda-graph)' is grammatically incorrect. It should be '(must be used with --disable-cuda-graph)'.” “The parenthetical note contains grammatical errors and missing spacing. 'need used' should be 'must be used', and 'only support for deepseek' should be 'only supported for DeepSeek'.”

讨论已通过提交采纳建议解决，修正了语法错误和格式问题，无争议点。

风险与影响

- 风险：极低。纯文档更新，无代码变更风险；主要风险是文档准确性，若更新内容与实际实现不符，可能导致用户配置错误，但基于近期 NPU 文档频繁维护（如 #22793、#22799），此风险可控。
- 影响：直接影响使用 Ascend NPU 卸载功能的用户，帮助他们准确了解参数支持状态和约束；对系统和团队无实质性影响，但提升了文档与实现的同步度。

关联脉络

从近期历史 PR 看，本 PR 是 NPU 文档维护系列的一部分：

- 22793 和 #22799 同样修复 Ascend NPU 文档的格式或内容错误。
- 22707 更新过时的 NPU 文档描述，本 PR 可视为其延续，将参数从“Planned”改为实际支持。

这些 PR 共同反映了团队对 NPU 平台文档准确性的持续投入，以及 NPU 功能从规划到落地的演进趋势。