

PR #24977 完整报告

sgl-project/sclang

fix gb envs in deployment guide

合并时间: 2026-05-11 21:06

原文链接: <http://prhub.com.cn/sgl-project/sclang/pull/24977>

执行摘要

将 GB200/GB300 部署文档中的 NCCL 环境变量从 `NCCL_IB_DISABLE=1` 替换为 `NCCL_MNNVL_ENABLE=1` 和 `NCCL_CUMEM_ENABLE=1`, 但 review 中建议保留原变量, 未采纳。

功能与动机

根据 PR body, 目的是为 GB200/GB300 部署设置 `NCCL_MNNVL_ENABLE=1` 和 `NCCL_CUMEM_ENABLE=1` 环境变量。

实现拆解

1. 修改文件 `docs_new/src/snippets/autoregressive/ling-25-1t-deployment.jsx` 第 84 行, 将 `envPrefix` 从 `'NCCL_IB_DISABLE=1 '` 改为 `'NCCL_MNNVL_ENABLE=1 NCCL_CUMEM_ENABLE=1 '`。

关键源码片段

`docs_new/src/snippets/autoregressive/ling-25-1t-deployment.jsx`

唯一变更文件, 修改了 GB200/GB300 部署的环境变量配置。

```
// 文件 : docs_new/src/snippets/autoregressive/ling-25-1t-deployment.jsx
// 生成部署命令时根据硬件类型决定环境变量前缀
const generateCommand = () => {
  const { hardware, parallelism, toolcall } = values;

  const isGB = hardware === 'gb200' || hardware === 'gb300';
  // 原值 : 'NCCL_IB_DISABLE=1 '
  // 变更为 : 'NCCL_MNNVL_ENABLE=1 NCCL_CUMEM_ENABLE=1 '
  // note: 移除了 NCCL_IB_DISABLE=1, 可能影响 NVLink 优先使用
  const envPrefix = isGB ? 'NCCL_MNNVL_ENABLE=1 NCCL_CUMEM_ENABLE=1 ' : '';
  // ... 后续生成命令
```

评论区精华

gemini-code-assist[bot]: 移除 `NCCL_IB_DISABLE=1` 可能导致 NCCL 默认使用 InfiniBand 而非 NVLink, 影响通信性能, 建议保留。

作者未采纳该建议，直接合并。

风险与影响

- 风险: 移除了 `NCCL_IB_DISABLE=1`，在 NVLink 与 InfiniBand 共存的环境中 NCCL 可能降级到 InfiniBand，导致通信性能下降。
- 影响: 仅影响一份文档，但可能对 GB200/GB300 用户的首次部署产生负面性能影响。

关联脉络

无关联的其他 PR。