

# PR #30614 完整报告

sgl-project/sglang

[Diffusion][Docs] Ascend A2, A3 add basic usage and benchmark results in diffusion cookbook

合并时间: 2026-07-29 16:50

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/30614>

## 执行摘要

本 PR 为 SGLang 的 Diffusion 模型文档添加了 Ascend A2 和 A3 NPU 硬件支持，修改了交互部署组件和 5 篇 Cookbook 文档。主要变更包括新增硬件选项、生成对应命令、自动切换 Benchmark Tab，以及插入 Ascend 的 Benchmark 结果。PR 经过 Review 修复了硬件 ID 不一致、本地路径硬编码等问题。

## 功能与动机

根据 PR Body，目的是“introduces documentation and interactive UI support for Ascend A2 and Ascend A3 hardware platforms”。让使用 Ascend 的用户能够方便地部署 FLUX、Wan 等 Diffusion 模型，并获取官方性能数据。

## 实现拆解

- 交互组件硬件选项扩展：在 flux-deployment.jsx、wan21-deployment.jsx、wan22-deployment.jsx、qwen-image-deployment.jsx、zimage-turbo-deployment.jsx 的 hardware 选项列表中追加 a2、a3 条目。
- 命令生成逻辑：每个组件的 generateCommand 添加判断，当 hardware === 'a2' 或 'a3' 时，根据模型规格输出 --tp-size、--sp-degree、--num-gpus 等参数。A3 命令前添加注释说明单卡包含 2 个 NPU chip。
- Tab 自动切换：使用 useEffect 监听 values.hardware，通过 DOM 选择器点击对应标题的 Tab，使 Benchmark 区域自动切换到 Ascend 结果页。
- 文档内容更新：5 篇 .mdx 文件在“基本配置”段落中补充 Ascend A2/A3 支持声明，Benchmark 区域由原始纯文本改为 <Tabs> 分区，分别展示不同硬件平台的性能结果。
- Review 后修复：将硬件 ID 由 ascend2/ascend3 统一为 a2/a3；移除命令中的 --port；替换本地路径为 Hugging Face 仓库 ID。

### docs\_new/src/snippets/diffusion/flux-deployment.jsx

核心交互组件，新增 A2/A3 硬件选项及对应的命令生成逻辑，体现了不同硬件规格的部署配置

```
export const FluxDeployment = () => {
  const config = {
    options: {
      hardware: {
        name: 'hardware',
        title: 'Hardware Platform',
```

```

    items: [
      { id: 'b200', label: 'B200', default: true },
      { id: 'h200', label: 'H200', default: false },
      // ... 其他选项 ...
      { id: 'a2', label: 'A2', default: false }, // 新增 Ascend A2
      { id: 'a3', label: 'A3', default: false } // 新增 Ascend A3
    ]
  },
  // ...
},
};

// 监听硬件变化，自动切换文档中对应的 Benchmark Tab
useEffect(() => {
  const isAscend = values.hardware === 'a2' || values.hardware === 'a3';
  const targetTabName = isAscend ? 'Ascend A3' : 'NVIDIA B200';
  const allTabs = document.querySelectorAll('button, [role="tab"]');
  allTabs.forEach((tab) => {
    const text = tab.textContent.trim();
    if (text === targetTabName && tab.getAttribute('aria-selected') !== 'true') {
      tab.click();
    }
  });
}, [values.hardware]);
};

```

## 评论区精华

Review 中主要讨论了以下问题：

- 硬件 ID 命名不统一：> 应该用 a2 而不是 ascend2 —— 最终全部统一为 a2/a3。
- Tab 切换匹配不充分：> useEffect 只匹配 'Ascend A2 / A3'，但部分文档 Tab 标题是 'Ascend A3' —— 各组件最终单独调整目标标题，但未完美泛化。
- 本地路径硬编码：> remove extra /FLUX.2-dev use huggingface path instead of local path in docs —— 已修正。
- 移除 --port 参数：> remove --port option in docs —— 已执行。
- A2 单卡能力确认：> FLUX.1-dev can be runned on 1 A2 device. should we show it here? —— 是的，已体现。

## 风险与影响

- 文档命令准确性：启动命令中的 --tp-size、--sp-degree 等参数基于开发环境验证，用户环境不同可能导致失败。建议用户在首次使用时参考官方指南调整。
- DOM 操作依赖：Tab 切换使用 useEffect + querySelectorAll，属于 hack。后续文档站点样式或 DOM 结构变化可能导致自动切换失效。
- 多 Tab 标题兼容：不同模型文档 Ascend Tab 标题不统一，部分用例下自动切换可能无法触发。

## 关联脉络

本 PR 与近期 Diffusion 模块的改进 PR 相关：

- #31538 增加了 Diffusion 模型的层常驻功能，提升性能；本 PR 为其补充了在 Ascend 上的文档。
- #31361 修复了 Diffusion Worker 在容器环境中的误杀问题，增加了模块的稳定性；本 PR 延续了对 Diffusion 全模块的文档覆盖。此外，本 PR 为后续支持更多硬件平台提供了交互式文档的参考模式。