

PR #39389 完整报告

sgl-project/sglang

[NPU] [DOC] Rename NPU hardware to Ascend A2/A3 Series product

合并时间: 2026-09-15 10:33

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39389>

PR 分析报告: #39389 [NPU] [DOC] Rename NPU hardware to Ascend A2/A3 Series product

执行摘要

该 PR 是一次覆盖 66 个文件的纯文档重命名: 把 NPU 文档与文档 Playground 中已退役的 Atlas 800I A2/A3、Ascend 910B/910C 等硬件旧名, 统一替换为华为官方的 Ascend A2/A3/950PR/DT Series 产品命名 (+759/-759)。改动不触及任何运行时逻辑, 但 `_deployment.jsx` 的硬件下拉 label 与各 Deployment 组件里自动切 Tab 的 `targetTabName` 属于跨文件命名契约, 需与 MDX 页面 Tab 文案逐字一致。该 PR 由 `sglang-npu-bot` 直接合并, 无 review 讨论。

功能与动机

PR body 明确说明动机: "Rename retired Atlas 800I A2/A3 and Ascend 910B/910C hardware names to the Ascend A2/A3/950PR/DT Series product naming across the docs"。即 Atlas 800I A2/A3、Ascend 910B/910C 属于已退役的旧产品名, 文档需要与华为官方产品线命名保持一致, 避免用户在 Ascend NPU 部署时因旧名造成混淆。

实现拆解

1. 全局替换硬件旧名 (docs/docs): 以 `support_features.mdx` (+310/-310) 为代表, 将 `reference`、`best-practices`、`tutorials`、`optimization` 等目录下 NPU 文档中的 "Atlas 800I A2/A3"、"Ascend 910B/910C" 等旧名改为 "Ascend A2/A3/950PR/DT Series" 官方产品名; `qwen3_5_397b.mdx` 等 `best-practices` 页面的 Hardware 列统一使用带 "Series Products" 后缀的完整产品名。
2. 同步部署组件的展示 label (docs/src/snippets/_deployment.jsx): 共享部署面板的 npu 硬件下拉选项 label 从 "Atlas 800I A3" 改为 "Ascend A3 Series", 并同步更新注释 (910C 已退役); 该 label 是全局命名契约, 所有配方页面复用它。
3. 同步模型配方 JSX (diffusion 部署组件): `qwen-image`、`flux`、`wan21`、`wan22`、`zimage-turbo` 五个 Deployment 组件的硬件 items label 从 "A2"/"A3" 改为 "A2 Series"/"A3 Series", 命令模板注释 "#One A3 card has 2 npu chips" 改为 "#One A3 Series card has 2 npu chips"。
4. 同步 Tab 点击逻辑的匹配字符串: 各 JSX 组件 `useEffect` 中 `targetTabName` 由 'Ascend A3'/'Ascend A2 / A3' 改为 'Ascend A3 Series'/'Ascend A2/A3 Series'。这是

与 MDX Tab 文案的跨文件契约，需 trim 后全等匹配，任何一边漏改都会导致自动切 Tab 静默失败。

5. 第二笔提交收敛命名形式 (b970d97)：对成对出现的 A2/A3 引用，采用紧凑写法 "A2/A3 Series" 或 "Ascend A2/A3 Series Products"; kimi-k3.jsx 等高复杂度配方配置的注释也统一为 "A3 Series" 表述。
6. 测试与 CI 配套：无测试文件变更，属于纯文档 / 展示层改动；PR Test 通过，PR Test (Extra) 显示 x，无 AMD 运行记录。

docs/src/snippets/_deployment.jsx

文档 Playground 的共享硬件下拉数据源，npu 硬件选项 label 从 "Atlas 800I A3" 改为 "Ascend A3 Series"。该 label 是全局命名契约，所有配方页面复用，是本次重命名在展示层的关键落点。

```
// 部署 Playground 的共享硬件选项表 (Deployment 组件内部) 。
// label 是用户可见的产品名，同时也是各配方 JSX 自动点击 Tab 时的匹配键，
// 命名改动必须与 MDX 页面上的 Tab 文案保持一致，否则切 Tab 会静默失败。
const hardwareTable = {
  amd: [
    { id: "mi350x", label: "MI350X", vram: "288GB" },
    { id: "mi355x", label: "MI355X", vram: "288GB" },
  ],
  // 旧产品名 Atlas 800I A3 (910C) 已退役，整体改用华为官方
  // Ascend A3 Series 产品命名；1 张卡 = 2 个 die，因此 --tp-size
  // 是卡数的 2 倍 (32 卡 -> --tp-size 64) 。
  npu: [
    { id: "a3", label: "Ascend A3 Series", vram: "64GB/die" },
  ],
};
```

docs/src/snippets/diffusion/qwen-image-deployment.jsx

代表一组 diffusion 部署组件 (flux/wan21/wan22/zimage-turbo 同构改动)：同时修改硬件选项 label 与自动切 Tab 的 targetTabName 字符串，完整展示了 JSX 与 MDX 之间的双端命名契约。

```
export const QwenImageDeployment = () => {
  const config = {
    modelFamily: 'Qwen-Image',
    options: {
      hardware: {
        name: 'hardware',
        title: 'Hardware Platform',
        items: [
          { id: 'b200', label: 'B200', default: true },
          { id: 'b300', label: 'B300', default: false },
          { id: 'h200', label: 'H200', default: false },
          { id: 'h100', label: 'H100', default: false },
          { id: 'mi300x', label: 'MI300X', default: false },
        ]
      }
    }
  }
}
```

```

    { id: 'mi325x', label: 'MI325X', default: false },
    { id: 'mi355x', label: 'MI355X', default: false },
    // 退役名 A2/A3 -> 官方产品名 A2 Series/A3 Series
    { id: 'a2', label: 'A2 Series', default: false },
    { id: 'a3', label: 'A3 Series', default: false },
  ],
},
},
};

// 硬件切换后自动点击与硬件匹配的 Tab。targetTabName 与 MDX 页面的
// Tab 文案必须 trim 后全等，任一端漏改都会导致自动切 Tab 静默失败。
useEffect(() => {
  let targetTabName = 'AMD MI300X';
  if (values.hardware === 'a2') targetTabName = 'Ascend A2 Series';
  if (values.hardware === 'a3') targetTabName = 'Ascend A3 Series';

  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

该 PR 没有任何 review 评论与讨论线程，由 sglang-npu-bot 直接合入。唯一可观察到的演进信号来自作者的第二笔提交（b970d97）：将第一笔提交中成对的硬件引用进一步收敛为 "A2/A3 Series" 与 "Ascend A2/A3 Series Products" 两种形式，说明命名规范在提交过程中被细化。最终同时存在三种写法：组件下拉 / 注释用 "A3 Series"，Tab 匹配用 "Ascend A3 Series"，文档表格列用 "Ascend A3 Series Products"，后续维护需按场景选用。

风险与影响

- 跨文件 Tab 匹配契约：qwen-image/flux/wan21/wan22/zimage-turbo 的 targetTabName 与对应 MDX 页面 Tab 文案必须逐字一致（trim 后全等比较），66 个文件跨文档改名一旦有遗漏，自动切 Tab 会静默失败，且该行为没有自动化测试覆盖。
- 三种命名形式并存："A3 Series"、"Ascend A3 Series"、"Ascend A3 Series Products" 同时存在，后续新增 NPU 文档容易用错，需要维护者记住各自适用场景（组件 / 注释 vs Tab 匹配 vs 表格列）。
- 旧名不可检索：910B/910C、Atlas 800I 等旧名被整体替换后，按旧名搜索的存量用户将无法命中文档。
- 零运行时影响：全部改动限于 docs 与展示层 JSX 的 label/ 注释字符串，不涉及任何 sglang serve 参数、内核或推理逻辑；对用户的主要影响是 Ascend 部署文档的硬件命名与

官方产品线一致，减少旧称带来的误解。

关联脉络

本 PR 与近期的文档类 PR 共同构成文档 Playground 基础设施的持续演进：PR#39396（DeepSeek-V4 MI355X PD 配方）同样改动了 `_playground.jsx` 与 `_deployment.jsx`，PR#39373（RTX 5090 H3 配方）与 PR#39406（GLM-5.2 MI355X 配方）则共享同一套 `configs JSX` 与 `cookbook` 路径。本 PR 为这条演进线建立了一条 NPU 硬件命名基线，后续新增 Ascend 配方文档时应沿用 "A2/A3 Series"（组件 / 注释）与 "Ascend A2/A3 Series Products"（表格 / 正文）的分层规范，并与 NPU CI 测试线（PR#38833）保持硬件命名口径一致。