PR #23936 完整报告

sgl-project/sglang

mimo v2.5 pro sglang-jax cookbook

合并时间：2026-04-29 16:58

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23936

PR 分析报告：mimo v2.5 pro sglang-jax cookbook

执行摘要

本 PR 为小米 MiMo-V2.5-Pro 模型新增 TPU 部署指南，支持通过 sgl-jax 在 TPU v7x 和 v6e 上运行。同时扩展了已有的交互式部署面板，用户可直接选择 TPU 硬件并生成启动命令。变更集中于文档和前端配置，不涉及核心推理代码。

功能与动机

MiMo-V2.5-Pro 是一个大规模 MoE 模型，此前仅有 CUDA（H200/H100/B200/GB300）部署方案。社区用户希望在 TPU 上部署该模型。PR 利用独立的 sgl-jax 提供一个已验证的 TPU 部署方案，并集成到现有 cookbook 中，使用户无需跳转到外部文档即可获得完整命令。

实现拆解

1. 扩展硬件选项：在 mimo-v25-deployment.jsx 的 hardware 选项中新增 tpu-v7x 和 tpu-v6e，并标注 sgl-jax, Pro only。
2. 补充拓扑配置：在 HW_VARIANT_SPEC 中添加 proltpu-v7x 和 proltpu-v6e，设置 jax: true 标志；tp 值分别为 32 和 64（对应 JAX 设备总数）。
3. 命令生成逻辑：在 generateCommand 函数中检测 jax 属性，切换为 python -m sgl_jax.launch_server，强制启用 EP，禁用 EAGLE MTP 和 DeepEP，并设置 TPU 特有参数（mem-fraction, swa 等）。公共 flag 如 --chunked-prefill-size 和 --max-running-requests 被提取到条件外部，减少重复。
4. 新增文档章节：在 MiMo-V2.5.mdx 中添加 3.3 TPU Deployment，包含拓扑表、JAX 容器镜像、sgl-jax 安装指南，并澄清端口使用。

docs_new/src/snippets/autoregressive/mimo-v25-deployment.jsx

核心逻辑文件：添加 TPU 硬件选项、拓扑配置及命令生成路径。

```
// 硬件选项新增 TPU 条目
items: [
  { id: "h200", label: "H200", default: true },
  { id: "h100", label: "H100", default: false },
  { id: "b200", label: "B200", default: false },
  { id: "gb300", label: "GB300", default: false },
  { id: "tpu-v7x", label: "TPU v7x", default: false, subtitle: "sgl-jax, Pro only" },
```

```
  { id: "tpu-v6e", label: "TPU v6e", default: false, subtitle: "sgl-jax, Pro only" },
],

// 拓扑配置扩展，jax: true 表示使用 sgl-jax 运行时
const HW_VARIANT_SPEC = {
  "proIh200": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 16, multinode: true, nnodes: 2, blackwell:
  false, jax: false },
  "proIh100": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 16, multinode: true, nnodes: 2, blackwell:
  false, jax: false },
  "proIb200": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 8, multinode: false, blackwell: true, jax:
  false },
  "proIgb300": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 8, multinode: true, nnodes: 2, blackwell:
  true, jax: false },
  "proItpu-v7x": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 32, multinode: true, nnodes: 4,
  blackwell: false, jax: true },
  "proItpu-v6e": { slug: "XiaomiMiMo/MiMo-V2.5-Pro", tp: 64, multinode: true, nnodes: 16,
  blackwell: false, jax: true },
  // ... base variants 保持不变
};
```

评论区精华

- TPU v7x tp 值疑问：gemini-code-assist[bot] 指出 tp=32 与 16 chips 矛盾。作者回应并更新了文档，加入 JAX Devices/Chip 列。
- 端口混淆：文档提及内部 JAX 端口 8471 但命令使用 30271。作者将启动端口统一为 30000，与 CUDA 路径一致。
- 重复标志：--chunked-prefill-size 和 --max-running-requests 在两个 TPU 分支重复。作者将它们提取到条件外部。

风险与影响

风险：TPU 拓扑或 tp 值错误可能导致部署失败；sgl-jax 版本与文档可能不同步；用户可能误解 JAX 设备与物理芯片的关系。 影响：TPU 用户可以直接使用 cookbook 部署 MiMo-V2.5-Pro，降低使用门槛。对项目而言，扩大了支持的硬件范围。

关联脉络

此 PR 是 cookbook 系列的一部分，类似于近期为 DeepSeek-V4 添加 H200 FP4 部署选项的 PR (#23980)。两者都展示了如何通过文档和交互面板支持新的部署环境。