PR #28437 完整报告

sgl-project/sglang

docs(cookbook): add GLM-5.2 deployment cookbook

合并时间：2026-06-16 21:49

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28437

执行摘要

本 PR 为 GLM-5.2 模型新增部署文档，涵盖 H200/B200/GB300/B300 硬件、三种服务策略以及 FP8/BF16 精度，并提供实测性能数据。文档基于 Mintlify 框架，包含交互式命令生成面板和 Playground 实验区域。合并后用户可一键生成启动命令并参考性能选型。

功能与动机

GLM-5.2 是 Z.ai 基于 DeepSeek Sparse Attention 和 256-expert MoE 的旗舰模型，需要一份清晰的部署指南帮助用户快速上手。PR 动机是"Add a deployment cookbook for GLM-5.2"（引自 PR 标题），为社区提供标准的部署方案和基准参考。

实现拆解

1. 配置面板(glm-5.2.jsx)：定义模型属性、支持的硬件、策略变体、命令模板、Docker 镜像和 Playground 可调参数（TP、CP、DP-Attention、EP、解析器等），是部署命令生成的配置中心。
2. 基准数据(glm-5.2-benchmarks.jsx)：存储每个硬件×策略的实测速度（TTFT、TPOT、吞吐）和准确率，用于文档中的性能表格渲染。B300 数据为推断（verified: false）。
3. 文档页面(GLM-5.2.mdx)：利用 <Deployment> 和 <Playground> 组件加载上述配置，并附加模型结构说明、推理推理配置（reasoning/tool-call parsers）、DSA Context Parallelism 说明和 HiCache 建议。
4. 导航注册(docs.json)：在 GLM 分组中插入新页面，并调整排序使其成为默认入口。
5. 入口更新(intro.mdx、GLM-5.1.mdx)：将 GLM 卡片链接指向新文档，移除旧版 NEW 标记。

docs_new/src/snippets/configs/zai-org/glm-5.2.jsx

核心配置文件，定义 GLM-5.2 的部署命令生成逻辑、支持硬件、策略、Docker 镜像、Playground 功能等所有交互元素。

```
// 单节点 8 GPU 下，--enable-dp-attention 使得 --tp 8 --dp 8 通过
// attention 层数据并行而其他层保持张量并行，实际无需 64 GPU。
export const config = {
  modelName: "GLM-5.2",
  supportedHardware: ["h200", "b200", "gb300", "b300"],
  quantizations: [
    { id: "fp8", label: "FP8" },
```

```
    { id: "bf16", label: "BF16" },
  ],
  strategies: [
    { id: "low-latency", label: "Low-Latency" },
    { id: "balanced", label: "Balanced" },
    { id: "high-throughput", label: "High-Throughput" },
  ],
  nodeOptions: [
    { id: "single", label: "Single Node" },
    { id: "multi-2", label: "Multi-Nodes" },
  ],
  // b300 已加入镜像映射，避免回退到 dev 镜像
  dockerImages: {
    h200: "lmsysorg/sglang:latest",
    b200: "lmsysorg/sglang:latest",
    gb300: "lmsysorg/sglang:latest",
    b300: "lmsysorg/sglang:latest",
  },
  // Playground 中禁用 b300 的 Context Parallelism，因 rope 内核未适配
  playgroundFeatures: {
    attention: {
      knobs: [{
        id: "cp",
        label: "CP (DSA prefill)",
        values: [null, 1, 2, 4, 8],
        disable: { hw: ["b200", "gb300", "b300"] },
        disableReason: "DSA prefill Context Parallel is verified on Hopper (H200); the Blackwell
        sm100 DSA-CP FP8 rope kernel is not yet adapted.",
      }],
    },
  },
};
```

docs_new/src/snippets/configs/zai-org/glm-5.2-benchmarks.jsx

包含每个硬件和策略组合的实测基准数据，是文档性能和准确度数据的来源。

```
// 每项包含 match 条件和对应版本、速度、准确度数据。
// 平衡策略使用了调优的 chunked-prefill（32768）和 max-running-requests 80。
export const benchmarks = [
  {
    match: { hw: "h200", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" },
    sglang_version: "0.5.13.post1",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 740, tpot_ms: 4.06, tokens_per_sec_per_gpu: 26 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 5980, tpot_ms: 13.97, tokens_per_sec_per_gpu: 98 },
    ],
  },
```

```
  // ... 其他组合省略，B300 数据标注为推断
];
```

docs_new/docs.json

文档导航配置文件，需要注册新页面并调整 GLM 分组顺序。

```
{
  "group": "GLM",
  "pages": [
    "cookbook/autoregressive/GLM/GLM-5.2", // 新增，置于首位
    "cookbook/autoregressive/GLM/GLM-5.1",
    // ... 其他 GLM 页面
  ]
}
```

评论区精华

- TP8+DP8 可行性：机器人认为单节点无法分配 64 GPU，作者澄清 --enable-dp-attention 下 attention 层 DP 而其余层 TP，实际单节点运行通过验证。这一讨论有助于用户理解 SGLang 中 DP-Attention 的独特语义。
- B300 Docker 镜像：机器人指出缺失，作者已补充。
- B300 CP 禁用：机器人建议与 B200/GB300 一起禁用，作者已采纳。

风险与影响

- 数据可靠性：B300 和 BF16 的基准数据为推断，用户可能未注意 verified: false 标记，建议添加醒目警告。
- 命令正确性风险：--enable-dp-attention 的行为依赖版本，若运行时改动可能导致命令失效，需与发布版本保持一致。
- 文档维护：后续更新基准数据需要同步修改 JSX 文件，可能遗漏。
- 无代码变更：不影响系统运行，风险可控。

关联脉络

本 PR 与近期 GLM-5.1 文档（PR#28330）一脉相承，是 GLM 系列 cookbook 的更新。同时，GLM-5.2 采用的 DeepSeek Sparse Attention 架构与近期多次出现的 DeepSeek-V4 优化（PR#27928、PR#28392）技术同源，文档中引用的 DSA CP 限制也反映了实际硬件适配状态。