PR #33820 完整报告

sgl-project/sglang

Add Intern-S2-Mobius cookbook

合并时间：2026-08-11 04:13

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33820

执行摘要

本 PR 为 SGLang 文档站新增 Intern-S2-Mobius 的配置驱动 cookbook 页面：以一个部署矩阵配置（intern-s2-mobius.jsx）+ 一个基准数据配置（intern-s2-mobius-benchmarks.jsx）驱动 Deployment / Playground 组件渲染启动命令、基准卡片与精度数据。全部 H200 数据来自 2×H200 实测，B200 配方为推断并以 Unverified 徽章标注；无运行时代码改动，影响范围限于 docs 站点与导航接线。

功能与动机

PR body 说明这是 "a config-driven cookbook page for internlm/Intern-S2-Mobius"（Apache-2.0、Mobius-v0、35B 混合 GDN + full-attention 多模态模型，256K 上下文，内置 MTP/NEXTN 投机解码头）。模型支持已在 PR #33691 合入 main（2026-08-08），因此需要一个可复现的部署指南，覆盖两种运行策略：

- Low-Latency：开启 MTP NEXTN 3-1-4 投机解码，追求单用户最低延迟；
- High-Throughput：关闭投机，追求多用户并发下的总吞吐峰值。

文档同时承载实测基准（8K-in / 1K-out p50、GSM8K 1319 全量、GPQA Diamond 198×8 重复），让用户能依据数据选择配方，而不是盲目开关投机解码。

实现拆解

1. 新增文档页面（docs/cookbook/autoregressive/InternLM/Intern-S2-Mobius.mdx）：页面按 tag: NEW 发布，包含 §1 模型介绍（Mobius 架构、GDN ×30 + full-attention ×10、MoE-2560 专家库、MTP 层）、§2 配置提示（--trust-remote-code 必需、Mamba 状态池划分、NEXTN 参数）、§3 示例输出（Reasoning / Tool-Calling / Vision 三个逐字真实捕获）。
2. 新增部署矩阵配置（docs/src/snippets/configs/internlm/intern-s2-mobius.jsx）：export const config 以反规范化 cells 描述 5 维组合（硬件 H200/B200 × 变体 × 量化 × 策略 × 节点），每个 cell 自带完整启动 flag；verified 字段区分实测与推断。Playground 部分明确禁用 EP/PP 旋钮（该架构 40 层共享 4 个全局专家库，--ep-size != 1 会被 server_args 拒绝），并给出 MTP 3-1-4 与 2-1-3 两档投机预设。
3. 新增基准数据配置（docs/src/snippets/configs/internlm/intern-s2-mobius-benchmarks.jsx）：export const benchmarks 与部署 cells 共用同一 match 元组，页面自动配对；含 H200 两策略的 p50 TTFT / TPOT / 吞吐与精度，B200 暂无数据。

4. 导航接线与去重：docs/docs.json 在 InternLM 分组顶部插入新页面；docs/cookbook/autoregressive/intro.mdx 的 vendor 卡片 href 从 Intern-S2-Preview 重指向新页；Intern-S2-Preview.mdx 移除 tag: NEW（同厂商只保留一个 NEW 标记）。
5. 验证配套：无单元测试，但 PR body 附完整复现命令与数据来源说明；mint validate 通过，CI 曾因基础镜像问题 /rerun-failed-ci、/tag-and-rerun-ci 重跑后通过。

docs/src/snippets/configs/internlm/intern-s2-mobius.jsx

部署矩阵核心配置：5 维组合（硬件 / 变体 / 量化 / 策略 / 节点）以反规范化 cells 定义启动 flag，verified 字段区分实测与推断，并包含 EP 禁用、MTP 预设等模型特定约束。

```
// 反规范化 cells：每个 cell 自带完整启动 flag，渲染引擎只注入占位符；
// 避免在渲染期做组合推导，Mintlify hydration 时直接取值。
// verified: true 表示已在 2×H200 TP=2 实测；false 表示从 H200 配方
// 推断的 Blackwell 对照，页面会渲染黄色 Unverified 徽章。
export const config = {
  modelName: "Intern-S2-Mobius",
  supportedHardware: ["h200", "b200"],
  strategies: [
    { id: "low-latency", label: "Low-Latency" },
    { id: "high-throughput", label: "High-Throughput" },
  ],

  // 默认精度数据：GPQA 只在 low-latency 点实测，通过 defaultAccuracy
  // 广播到所有 cell，个别 cell 可用 accuracy 字段覆盖。
  defaultAccuracy: {
    default: { gsm8k_pct: 96.66, gpqa_pct: 79.23 },
  },

  playgroundFeatures: {
    // 该架构没有 MoE 并行卡片：40 层共享 4 个全局专家库（meta_mlp,
    // num_blocks: 4），EP 无分片对象；server_args 对 --ep-size != 1
    // 会直接抛错，故配置里不出现 EP 旋钮。
    speculative: {
      options: [
        { id: "current", label: "Inherited from base" },
        { id: "off", label: "Off (greedy)" },
        { id: "mtp-314", label: "MTP / NEXTN 3-1-4 (recommended)",
          flags: ["--speculative-algorithm NEXTN", "--speculative-num-steps 3",
                  "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 4"] },
        { id: "mtp-213", label: "MTP / NEXTN 2-1-3 (lighter draft)",
          flags: ["--speculative-algorithm NEXTN", "--speculative-num-steps 2",
                  "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 3"] },
      ],
    },
  },

  cells: [
    // H200 / 2 GPU / BF16 / low-latency —— 已实测，开启 MTP NEXTN 3-1-4
```

```
    {
      match: { hw: "h200", variant: "default", quant: "bf16", strategy: "low-latency", nodes:
      "single" },
      verified: true,
      env: [],
      flags: [
        "--trust-remote-code",
        "--model-path {{MODEL_NAME}}",
        "--tp 2",
        "--mem-fraction-static 0.8",
        "--context-length 262144",
        "--reasoning-parser qwen3",
        "--speculative-algorithm NEXTN",
        "--speculative-num-steps 3",
        "--speculative-eagle-topk 1",
        "--speculative-num-draft-tokens 4",
        "--host {{HOST_IP}}",
        "--port {{PORT}}",
      ],
    },
    // B200 / 1 GPU / BF16 / high-throughput —— 推断配方：单卡 192 GB HBM
    // 可容纳 73 GB 权重 + KV，故 TP=1；未经真机验证
    {
      match: { hw: "b200", variant: "default", quant: "bf16", strategy: "high-throughput", nodes:
      "single" },
      verified: false,
      env: [],
      flags: [
        "--trust-remote-code",
        "--model-path {{MODEL_NAME}}",
        "--tp 1",
        "--mem-fraction-static 0.8",
        "--context-length 262144",
        "--reasoning-parser qwen3",
        "--host {{HOST_IP}}",
        "--port {{PORT}}",
      ],
    },
  ],
};
```

docs/src/snippets/configs/internlm/intern-s2-mobius-benchmarks.jsx

基准数据配置：与部署 cells 共用 match 元组，HTTP 页面据此自动配对，提供两策略的 p50 性能与精度数据。

```
// 基准数组与 config.cells 共用同一组 match 元组，页面据 key 自动配对；
// 没有 speed 数据的 cell（B200）表示该组合尚未在真机测量。
export const benchmarks = [
  // H200 / low-latency（MTP NEXTN 3-1-4）：conc=1 时 P50 TPOT 仅 3.13 ms，
```

```
  // 投机解码对单流延迟收益明显；conc=64 达该配方峰值 26033 tok/s
  {
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "low-latency", nodes: "single" }
    ,
    sglang_version: "main @ e0828ee3",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 177.91, tpot_ms: 3.13, tokens_per_sec_per_gpu: 1283.2 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 64 },
        ttft_ms: 4440.38, tpot_ms: 12.17, tokens_per_sec_per_gpu: 13016.5 },
    ],
    accuracy: { gsm8k_pct: 96.66, gpqa_pct: 79.23 },
  },
  // H200 / high-throughput（spec off）：conc=256 达饱和峰值 34786 tok/s,
  // 超过 spec 配方峰值 1.34 倍；conc=1024 被服务端 max_running 钳制到
  // ~730，TTFT 3 分钟实为排队等待而非计算耗时
  {
    match: { hw: "h200", variant: "default", quant: "bf16", strategy: "high-throughput", nodes:
    "single" },
    sglang_version: "main @ e0828ee3",
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 256 },
        ttft_ms: 16290.63, tpot_ms: 50.31, tokens_per_sec_per_gpu: 17393.4 },
    ],
    accuracy: { gsm8k_pct: 96.82 },
  },
];
```

评论区精华

官方 review 仅一条 APPROVED（zijiexia），无文字讨论；真正的质量把关发生在提交迭代里：

> 两次提交专门替换 §3 示例输出：先手写填充，后改为 verbatim 捕获真实服务器响应（含字段顺序、ID、模型自带 `cw` 尾标），并补充说明采样随机、重跑文本不同。
>
> PR body 的 Notes for reviewers 主动披露 B200 配方为推断（`verified: false`，渲染黄色 Unverified 徽章）、GPQA 仅在 low-latency 点测量并通过 `defaultAccuracy` 广播；两者均作为已知限制合入。

风险与影响

- B200 配方未实测：TP=2 / TP=1 推断配置可能与真机行为有偏差（尤其 KV 池与内存估算），页面用 Unverified 徽章缓解。
- GPQA 广播数据：spec-off 与 spec-on 推理行为不同，默认精度可能轻微漂移；配置已预留 per-cell override。
- verbatim 示例输出：随 SGLang 升级或 parser 行为变化，示例字段可能过期。
- AMD 轴留空：GDN 的 Triton 后端未在 ROCm 上验证，用户可能在 AMD 环境误用配置。

- 依赖 nightly 镜像：文档指向 lmsysorg/sglang:dev，发布节奏变化时表述需跟进。

总体影响仅限 docs 站点，运行时零风险；对部署用户的收益明确（可直接复制的启动命令与配方选择依据）。

关联脉络

- PR #33691：Intern-S2-Mobius 模型支持，本 PR 的验证基线（实测基于其 head，合并后文档改用 nightly 镜像）。
- PR #34283 / #34271：Ling-3.0-tiny、Muse Glimmer cookbook，同属"配置驱动 cookbook"系列，共同确立了 cells 反规范化、verified 标记、defaultAccuracy 广播等文档数据建模约定，后续新模型文档大概率沿用此模板。