PR #25370 完整报告

sgl-project/sglang

[NEW MODEL] Add H200 validation for Ring-2.6-1T cookbook

合并时间：2026-05-16 02:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25370

执行摘要

此 PR 为 Ring-2.6-1T 模型新增了 NVIDIA H200 x8 硬件的部署配置和已验证性能基准。变更仅涉及文档和 UI 组件，无后端逻辑更改，可安全合并。

功能与动机

Ring-2.6-1T 是新发布的巨型 MoE 模型，社区需要在不同硬件上部署的参考指南。此 PR 提供了 H200 x8 上经过烟雾测试、GSM8K 精度验证的部署参数和 latency/throughput 基准数据，方便用户直接参考使用。

实现拆解

1. 部署命令生成器(docs_new/src/snippets/autoregressive/ring-26-1t-deployment.jsx)：在硬件选择下拉框中新增 H200 x8 选项，并在配置字典中添加对应的 TP=8 和 mem_fraction=0.95 参数。
2. cookbook 文档(docs_new/cookbook/autoregressive/InclusionAI/Ring-2.6-1T.mdx)：
   - 更新硬件列表文字，加入 H200 x8。
   - 添加 --tp-size 8 和 --mem-fraction-static 0.95 的部署参数建议。
   - 在“Speed Benchmark”和“Throughput-Sensitive Benchmark”小节中分别插入 H200 的基准测试结果，包括请求吞吐量、E2E 延迟、TTFT、TPOT、ITL 等指标。

docs_new/src/snippets/autoregressive/ring-26-1t-deployment.jsx

Ring-2.6-1T 交互式命令生成器的 React 组件源码；变更增加了 H200 x8 硬件选项及其对应的 TP 与 mem_fraction 配置。

```
/* 在硬件选项数组中新增 `h200` 项，并添加其模型配置 */
export const Ring261TDeployment = () => {
  const options = {
    hardware: {
      name: 'hardware',
      title: 'Hardware Platform',
      items: [
        { id: 'gb300', label: 'GB300 x4', default: true },
        { id: 'b200', label: 'B200 x8', default: false },
        { id: 'h200', label: 'H200 x8', default: false }, // 新增 H200 x8 选项
      ],
    },
```

```
    // ... 其余选项保持不变
  };
  // H200 使用 TP=8、mem_fraction=0.95（与 B200 相同 TP，但 mem_fraction 更高）
  const modelConfigs = {
    gb300: { tp: 4, memFraction: '0.95' },
    b200: { tp: 8, memFraction: '0.8' },
    h200: { tp: 8, memFraction: '0.95' },
  };
  // ... 其余 UI state/effect 逻辑
};
```

评论区精华

gemini-code-assist[bot] 指出硬件列表已添加 H200，但 benchmark 结果部分还未包含对应数据，建议补充或添加 "TBD" 占位符。最终提交已包含完整 benchmark 数据，该 issue 已解决。

风险与影响

- 风险：极低。变更仅涉及文档和 UI 配置项，不触及推理核心或后端逻辑。文档中的参数值已通过实际验证。
- 影响：为用户提供 H200 上的权威部署参考，降低使用门槛。团队维护成本很低。

关联脉络

此 PR 是 Ring-2.6-1T 模型 cookbook 的硬件平台扩展，无直接关联的 issue 或 PR。