PR #25243 完整报告

sgl-project/sglang

[Docs] update dsv4 cookbook with H100 deployment commands

合并时间：2026-05-14 14:45

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25243

执行摘要

本 PR 为 DeepSeek-V4 部署文档新增 H100 (FP4) 硬件支持，包括部署命令生成器中的 H100 硬件选项、相应的部署参数配置，以及 Docker 镜像表的更新。同时将 Marlin FP4 路径的共享逻辑重构，便于未来扩展。

功能与动机

根据 PR body 的描述，本次变更是为了在 DeepSeek-V4 部署文档的交互式部署命令生成器中增加 NVIDIA H100 (FP4)硬件选项。H100 使用与 H200 (FP4) 相同的 Marlin FP4 runner，但具有更高的 TP 要求：Flash 模型需要 TP=8 单节点，Pro 模型需要 TP=16 双节点。此外，还需要在 Docker 镜像表中添加对应的镜像行。

实现拆解

1. 新增 H100 硬件选项：在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 的 hardware items 中添加 { id: "h100", label: "H100 (FP4)", default: false }，并在文件注释中增加 H100 说明。
2. 重构 Marlin 共享常量：将原先仅作用于 H200 (FP4) 的硬编码集合 H200_FP4_UNSUPPORTED_RECIPES 重构为三个共享常量：MARLIN_UNSUPPORTED_RECIPES（禁用 recipe 集合）、MARLIN_HARDWARE（Marlin 硬件 ID 集合）和 MARLIN_LABEL（ID 到名称的映射），使 H100 能自动继承相同的 UI 禁用和自动回退逻辑。
3. 更新条件判断：将 resolveItems 和 handleRadioChange 中的硬编码 vals.hardware === "h200-fp4" 改为 MARLIN_HARDWARE.has(vals.hardware)，禁用提示消息也改为动态引用 MARLIN_LABEL。
4. 新增 H100 部署配置：在 deploymentConfig 中添加：
   - h100lsmall：Flash 模型，TP=8，单节点
   - h100lbig：Pro 模型，TP=16，双节点，并设置环境变量 SGLANG_SHARED_EXPERT_TP1=1、--mem-fraction-static 0.9，且在 low-latency 和 balanced recipe 下额外添加 --cuda-graph-max-bs 8 --max-running-requests 32
5. 更新 Docker 镜像表：在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 的 Docker 镜像表格中新增一行 "NVIDIA H100 → lmsysorg/sglang:dev"。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

核心变更文件，新增 H100 硬件选项并重构 Marlin 路径共享常量

```
// 以下是重构后的核心逻辑片段：

// 定义 Marlin（FP4）硬件路径上不支持的 recipe
// 现为 H200 (FP4) 和 H100 (FP4) 共享
const MARLIN_UNSUPPORTED_RECIPES = new Set(["cp", "pd-disagg"]);

// 所有使用 Marlin runner 的硬件 ID 集合
const MARLIN_HARDWARE = new Set(["h200-fp4", "h100"]);

// 硬件 ID → 显示名称的映射，用于动态禁用提示
const MARLIN_LABEL = { "h200-fp4": "H200 (FP4)", h100: "H100 (FP4)" };

// 根据当前选中的硬件，返回可用 recipe 列表
// 若硬件属于 Marlin 路径，则禁用 cp 和 pd-disagg
const resolveItems = (option, vals) => {
  if (option.name === "recipe" && vals && MARLIN_HARDWARE.has(vals.hardware)) {
    return option.items.map((it) =>
      MARLIN_UNSUPPORTED_RECIPES.has(it.id)
        ? {
            ...it,
            disabled: true,
            disabledReason: `Not supported on ${MARLIN_LABEL[vals.hardware]}`
          }
        : it
    );
  }
  return option.items;
};

// 当用户切换到 Marlin 硬件时，若当前 recipe 不被支持，自动回退到 low-latency
const handleRadioChange = (optionName, value) => {
  setValues((prev) => {
    const next = { ...prev, [optionName]: value };
    if (
      optionName === "hardware" &&
      MARLIN_HARDWARE.has(value) &&
      MARLIN_UNSUPPORTED_RECIPES.has(next.recipe)
    ) {
      next.recipe = "low-latency";
    }
    // ... 其余处理
  });
};
```

评论区精华

本 PR 仅有一条机器人评论，提示每日配额限制，未产生人工 review 讨论或争议。最终由 wisclmy0611 审核通过。

风险与影响

风险：风险极低。变更仅涉及文档和前端 UI 组件，不涉及任何运行时代码或核心推理逻辑。但 H100 的部署参数（如 TP、mem_fraction_static 等）是基于经验值，可能需要根据真实环境后续微调。

影响：影响范围小。DeepSeek-V4 文档用户将能在交互式部署命令生成器中看到 H100 选项，并获得正确的命令行参数。Docker 镜像表也同步更新，方便用户选择镜像。对系统其他功能无影响。

关联脉络

本 PR 属于 DeepSeek-V4 文档的持续完善工作，与此前已有的 H200 (FP4) 支持形成互补，并在代码层面通过共享常量实现了统一的维护逻辑。没有观察到与其他近期 PR 的直接关联。