PR #25923 完整报告

sgl-project/sglang

[Docs] DeepSeek-V4: switch H200 FP4 Pro to flashinfer_mxfp4, Flash Balanced too

合并时间：2026-05-22 04:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25923

执行摘要

此 PR 修正 DeepSeek-V4 部署文档中 H200 FP4 模式下 MoE 后端的硬编码选择，将 Pro 和 Flash Balanced 从 Marlin 切换为 flashinfer_mxfp4，基于实际基准测试验证，解决了 Pro 变体无法启动和 Flash Balanced 吞吐性能不足的问题。

功能与动机

原先文档中的命令生成器在 H200 FP4 分支无条件使用 --moe-runner-backend marlin，导致两个问题：

1. Pro 变体在 TP=8 时服务器无法启动；
2. Flash Balanced 在吞吐基准测试中性能不佳。 需要根据实际验证结果为不同 recipe 选择最优后端。

实现拆解

1. 命令生成逻辑调整（docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx）：
   - 引入 useFlashinferMxfp4 变量，当模型为 Pro（isBig）或 recipe 为 balanced 时为真。
   - 据此选择 --moe-runner-backend flashinfer_mxfp4 或 --moe-runner-backend marlin。
   - 添加注释说明基准验证结果。
2. 文档同步（docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx）：
   - 在 Hopper 使用说明中明确提及 Pro 模型使用 Flashinfer。

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

核心变更文件：修改命令生成逻辑，根据 recipe 动态选择 MoE 后端。

```
// 从 generateCommand 函数中提取的 H200 FP4 分支逻辑
if (hardware === "h200-fp4") {
  const verifyKey = `${hardware}|${modelSize}|${recipe}`;
  if (TBD_RECIPES.has(verifyKey)) return TBD_PLACEHOLDER;

  // 核心变更：动态决定 MoE 后端
  // Pro（isBig）或 Balanced recipe 使用 flashinfer_mxfp4，
  // 其他 Flash recipe 仍使用 marlin。
  const useFlashinferMxfp4 = isBig || recipe === "balanced";
  const fp4Flags = [
```

```
    "  --trust-remote-code",
    `  --model-path ${slug}`,
    `  --tp ${tp}`,
    useFlashinferMxfp4
      ? "  --moe-runner-backend flashinfer_mxfp4"
      : "  --moe-runner-backend marlin",
  ];
  // 其余 flags 不变 ...
}
```

评论区精华

无 review 讨论。

风险与影响

- 风险：极低。仅影响文档生成，不涉及运行时代码。但需确保命令可执行性。
- 影响：正面影响——Pro 变体现在可正常启动，Flash Balanced 吞吐提升约 1.25 倍。负面无。

关联脉络

该 PR 独立，无关联 Issue 或相关 PR。