PR #23817 完整报告

sgl-project/sglang

docs: verify GB300 Pro DeepSeek V4 recipes

合并时间：2026-04-27 15:21

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23817

执行摘要

PR #23817 在 DeepSeek V4 部署配置文件中，将 GB300 Pro 硬件的 big|balanced 和 big|max-throughput 两种部署配方标记为已验证，并设置其 --mem-fraction-static 参数为 0.9。变更仅涉及一个 JSX 文件，无代码逻辑风险。

功能与动机

根据 PR 描述，目标是对 GB300 Pro 的 balanced 和 max-throughput 配方进行端到端验证，并设置合适的内存比例。用户可直接使用这些已验证的部署命令。

实现拆解

1. 验证标记：在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 中的 VERIFIED_RECIPES 集合里加入 "gb300|big|balanced" 和 "gb300|big|max-throughput" 两项。
2. 参数调整：在 balanced 和 max-throughput 配方分支中，新增 else if (isBig && hardware === "gb300") 条件，设置 --mem-fraction-static 0.9。此值高于 gb200 的 0.78 和 b200 的 0.82，推测与 GB300 Pro 内存架构相关。
3. 验证机制：该文件通过一个 JSX 组件渲染部署命令，未在 VERIFIED_RECIPES 中的配方会被默认注释掉，避免用户执行未经验证的命令。

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

唯一变更文件，既是文档也是部署配置的 JSX 代码，验证状态和参数调整均在此处。

```
// 已验证配方集合：添加了 GB300 Pro big balanced 和 max-throughput
const VERIFIED_RECIPES = new Set([
    // ... 已有条目 ...
    "gb300|big|balanced",
    "gb300|big|max-throughput",
    // ... 其他条目 ...
]);

// balanced 配方中设置 mem-fraction-static
if (hardware === "h200" && isBig) {
    flags.push("  --mem-fraction-static 0.88");
} else if (isBig && hardware === "gb300") {
    flags.push("  --mem-fraction-static 0.9"); // GB300 Pro 使用 0.9
```

```
} else if (isBig && hardware === "gb200") {
    flags.push("  --mem-fraction-static 0.78");
}

// max-throughput 配方中同理
if (hardware === "h200" && isBig) {
    flags.push("  --mem-fraction-static 0.88");
} else if (isBig && hardware === "gb300") {
    flags.push("  --mem-fraction-static 0.9"); // GB300 Pro 使用 0.9
}
```

评论区精华

无 reviewer 实质性评论，仅有一个 AI 自动评论和直接审批。

风险与影响

- 风险：极低。仅修改配置文件，不影响任何运行时逻辑。但若 mem-fraction-static 0.9 未经充分测试，在极端负载下可能引发 OOM，但鉴于已验证声明，风险可控。
- 影响：面向用户的部署文档更完善，GB300 Pro 用户可直接使用已验证命令。

关联脉络

- 关联 PR #23737：类似操作，标记 GB200 big 低延迟配方已验证。
- 属于 DeepSeek V4 部署文档的持续完善，为不同硬件提供已验证的推荐配置。