PR #29252 完整报告

sgl-project/sglang

Tune Gemma4 26B-A4B B200 memory recipe

合并时间：2026-06-25 11:31

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29252

执行摘要

此 PR 将 Gemma4 26B-A4B 模型在 B200 硬件上的默认静态内存占比从 0.9 调整为 0.75，避免因 Triton 融合 MoE 内核的工作空间不足导致服务不稳定。变更涉及部署配置文件和文档，风险极低。

功能与动机

Gemma4 命令生成器对 B200 上的 google/gemma-4-26B-A4B-it 模型生成了 --mem-fraction-static 0.9，但该数值对于 B200 MoE 方案过于激进，因为模型使用 Triton fused-MoE 路径，在服务时需要额外的工作空间。B200 cookbook 验证成功使用了 0.75，因此选择器应将该值调整为更可靠的工作值。

实现拆解

1. 修改默认内存占比：在 docs_new/src/snippets/autoregressive/gemma4-deployment.jsx 中，将 b200 型号下 '26b-a4b' 的 mem 值从 0.9 改为 0.75。该文件是自动命令生成的入口。
2. 添加配置提示：在 docs_new/cookbook/autoregressive/Google/Gemma4.mdx 中新增一条说明，提示用户对 Gemma 4 26B-A4B 使用 --mem-fraction-static 0.75 为 Triton MoE 路径预留工作空间。

docs_new/src/snippets/autoregressive/gemma4-deployment.jsx

修改了 B200 上 26B-A4B 模型的默认内存占比，这是自动命令生成的入口。

```
// 模型配置表，定义了不同硬件 / 模型组合的 TP 和内存占比
const modelConfigs = {
    h200: {
      e2b: { tp: 1, mem: 0.85 },
      e4b: { tp: 1, mem: 0.85 },
      '12b': { tp: 1, mem: 0.85 },
      '31b': { tp: 2, mem: 0.85 },
      '26b-a4b': { tp: 1, mem: 0.85 },
    },
    b200: {
      e2b: { tp: 1, mem: 0.9 },
      e4b: { tp: 1, mem: 0.9 },
      '12b': { tp: 1, mem: 0.9 },
      '31b': { tp: 1, mem: 0.9 },
```

```
      // 26B-A4B 使用 0.75 而不是 0.9，为 Triton 融合 MoE 内核预留工作空间
      '26b-a4b': { tp: 1, mem: 0.75 },
    },
    b300: {
      e2b: { tp: 1, mem: 0.9 },
      e4b: { tp: 1, mem: 0.9 },
      '12b': { tp: 1, mem: 0.9 },
      '31b': { tp: 1, mem: 0.9 },
      '26b-a4b': { tp: 1, mem: 0.9 },
    },
    mi300x: {
      '31b': { tp: 1, mem: 0.80 },
      '26b-a4b': { tp: 1, mem: 0.80 },
    },
  };
```

评论区精华

无 review 讨论。

风险与影响

- 风险：极低。仅修改了部署配置和文档，不影响核心代码。若后续模型需要更高内存，0.75 可能不足，但已通过验证确认其可行性。
- 影响：仅影响 Gemma4 26B-A4B 在 B200 上的自动命令生成用户，手动部署用户不受影响。

关联脉络

此 PR 是一个独立的配置调整，未关联其他 Issue 或 PR。