PR #26451 完整报告

sgl-project/sglang

[docs] Fix V4 Pro balanced recipe

合并时间：2026-05-27 14:50

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26451

执行摘要

- 一句话：修复 V4 Pro Balanced 配方的 MegaMoE 和 DeepEP 后端
- 推荐动作：可直接合并，逻辑清晰且无争议。

功能与动机

在 DeepSeek-V4 部署生成器中，Balanced 配方下 Flash 和 Pro 变体使用了不同的 MoE 后端：Flash 使用 --moe-a2a-backend deepep，而 Pro 使用 --moe-runner-backend flashinfer_mxfp4。此 PR 对齐两者，使 Pro 也使用 DeepEP，并在启用 MegaMoE 时通过现有逻辑自动替换为 MegaMoE。

实现拆解

1. 修改 MEGAMOE_UNSUPPORTED_RECIPES 集合：在 docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx 中将 "balanced" 加入该集合，使得 MegaMoE 在 Balanced 配方下自动禁用（灰显），只在 max-throughput 配方中可用。
2. 删除 Balanced 配方的 MegaMoE 环境变量：移除之前为 Balanced 配方特殊设置的 SGLANG_OPT_DEEPGEMM_MEGA_MOE_NUM_MAX_TOKENS_PER_RANK=4096 环境变量，因为 MegaMoE 已不在该配方下运行。
3. 更新文档：在 docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx 中同步更新 MegaMoE 支持说明，明确其仅支持 max-throughput 配方。

关键文件：

- docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx（模块 部署生成器；类别 source；类型 core-logic）：核心变更文件，修改了 MegaMoE 支持的配方集合和 Balanced 配方的环境变量。
- docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx（模块 文档；类别 other；类型 core-logic）：文档同步更新，说明 MegaMoE 仅支持 max-throughput 配方。

关键符号：未识别

关键源码片段

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

核心变更文件，修改了 MegaMoE 支持的配方集合和 Balanced 配方的环境变量。

```
// 修改后的 MegaMoE 不支持配方集合：MegaMoE 仅支持 max-throughput
const MEGAMOE_UNSUPPORTED_RECIPES = new Set([
  "low-latency",
  "balanced", // 新增：Balanced 配方也不支持 MegaMoE
  "cp",
  "pd-disagg"
]);

// 移除 Balanced 配方的 MegaMoE 环境变量（原代码已删除）
```

评论区精华

无 review 讨论。

- 暂无高价值评论线程

风险与影响

- 风险：风险较低。主要影响部署生成器的输出和文档描述，不涉及运行时逻辑。需确保用户依赖 Balanced 配方的 Pro 变体时，生成命令正确使用 DeepEP 后端并禁用 MegaMoE。
- 影响：影响范围小，仅影响 DeepSeek-V4 部署交互式生成器的 Balanced 配方和对应文档。用户此前若通过 Pro Balanced 配方使用 flashinfer_mxfp4 后端，更新后将改为 deepep 后端。
- 风险标记：暂无

关联脉络

- PR #26413 [docs] DeepSeek-V4 cookbook: note cu129 image for GB200 Pro DeepEP backend: 同为 DeepSeek-V4 部署文档改进，涉及 Pro 变体的 DeepEP 后端配置。