PR #23622 完整报告

sgl-project/sglang

Again update DeepSeek V4 cookbook

合并时间：2026-04-24 15:12

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/23622

执行摘要

本 PR 是 DeepSeek V4 部署文档的增量更新，扩展了已验证的部署命令组合（新增 B200 和 H200 的多种配方），简化了配方选择器 UI，修复了 cp 配方的参数生成逻辑，并补充了 Docker 运行示例。整体为低风险文档维护变更，对用户部署有实际指导价值。

功能与动机

继 #23605 和 #23617 初步建立 DeepSeek V4 部署指南后，团队通过端到端测试验证了更多硬件与配方组合（B200 Flash/Pro 的 balanced、max-throughput、cp；H200 Flash 的 low-latency、balanced、max-throughput），并将其标记为已验证，使生成的命令可直接复制运行。同时根据人类反馈，纠正了 cp 配方中 --max-running-requests 和 --mem-fraction-static 的参数处理，确保每个参数只出现一次，避免歧义。此外，为简化用户上手流程，在 Docker 镜像表格下方添加了最小 docker run 示例及安装文档链接。

实现拆解

步骤 1：简化配方选择器 UI

- 文件：docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx
- 变更：移除了 recipe 选项集合中每个条目（low-latency, balanced, max-throughput, cp, pd-disagg）的 subtitle 字段
- 原因：使界面更简洁，只显示配方名称，避免信息过载

步骤 2：扩展已验证配方集合

- 文件：同上
- 变更：在 VERIFIED_RECIPES Set 中添加了以下键：
  - b200|small|balanced、b200|small|max-throughput、b200|small|cp
  - b200|big|balanced、b200|big|max-throughput、b200|big|cp
  - h200|small|low-latency、h200|small|balanced、h200|small|max-throughput
- 原因：这些组合经过实际部署验证，确保命令参数正确可用

步骤 3：修复 cp 配方标志生成逻辑

- 文件：同上
- 变更：调整 buildCommand 函数中 recipe 值为 cp 的分支：
  - 移除 --mem-fraction-static 0.70（不再覆盖前面的 0.78）

  - 将 --max-running-requests 改为条件设置：Blackwell big（即 isBig && hardware 不是 h200 时）设为 256，其他情况（包括 H200）设为 1024
- 原因：根据人类最新指示，避免参数重复设置，使命令清晰一致

下面是该逻辑的核心片段（已包含详尽的注释）：

步骤 4：补充 Docker 使用文档

- 文件：docs_new/cookbook/autoregressive/DeepSeek/DeepSeek-V4.mdx
- 变更：在 Docker 镜像表格下方新增一段文字，包含指向安装文档的链接和最小 docker run 示例
- 原因：帮助用户快速启动容器部署，降低入门门槛

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

核心变更文件：更新了已验证配方集合、简化了 UI、调整了 cp 配方的参数生成逻辑。直接影响用户复制的命令。

关键源码片段

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

核心变更文件：更新了已验证配方集合、简化了 UI、调整了 cp 配方的参数生成逻辑。直接影响用户复制的命令。

```
// cp 配方标志生成部分（简化抽取）
if (recipe === 'cp') {
  // ... 之前已加入 --tp、--moe-a2a-backend deepep 等通用标志
  flags.push('--mem-fraction-static 0.78');
  // 人类指示（2026-04-24）：不再设置 --mem-fraction-static 0.70 覆盖
  // 只设置 --max-running-requests 一次：Blackwell big 用 256，其余用 1024
  if (isBig && hardware !== 'h200') {
    flags.push('--cuda-graph-max-bs 256');
    flags.push('--max-running-requests 256');
  } else {
    flags.push('--max-running-requests 1024');
  }
  // H200 cp 在非多节点时加上 DeepEP 大 SMS 标志
  if (!multinode) flags.push(DEEPEP_LARGE_SMS_FLAG);
}
```

评论区精华

本 PR 没有来自人类审核者的评论。仅有的两个评论来自 gemini-code-assist[bot]，内容为每日配额已达上限的通知，不涉及技术讨论。

风险与影响

- 风险：低风险。主要是文档和命令模板的更新，不影响运行时逻辑。cp 配方参数调整需与实际 sglang 版本保持同步，已人工核对。
- 影响：

- 用户：获得更多已验证的一键部署命令，减少调试时间；Docker 示例降低新手门槛。
- 系统：无运行时影响。
- 团队：文档维护更规范，已验证集合的管理更加清晰。

关联脉络

本 PR 是 DeepSeek V4 文档系列的一部分，前序 PR 包括：

PR#	标题	关联原因
#23605	Add DeepSeek V4 cookbook	首次建立 DeepSeek V4 交互部署文档
#23617	Further update Deepseek V4 docs	更新模型仓库地址等
#23634	Update pro fp8 checkpoint in DeepSeek V4 cookbook	同期文档更新，修改 H200 Pro 的模型地址

这些 PR 共同构建了 DeepSeek V4 在 sglang 上的部署指南体系，本 PR 在此基础上扩展了已验证配方的覆盖范围。