PR #26164 完整报告

sgl-project/sglang

[docs] DeepSeek-V4 cookbook: balanced MegaMoE cap, H200 Pro FP4 mem-frac, nsa-* compat, PD-disagg fixes

合并时间：2026-05-23 17:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/26164

PR 分析报告：[docs] DeepSeek-V4 cookbook 五项修复

执行摘要

本 PR 对 DeepSeek-V4 部署命令生成器（deepseek-v4-deployment.jsx）进行了五项兼容性和正确性修复，涵盖 Blackwell Balanced + MegaMoE 显存限制、H200 Pro FP4 低延迟显存比例调整、上下文并行标志名回退兼容、PD-Disagg 时禁用未实现的选项以及 parser 标志放置到正确 role。仅影响生成器输出的命令文本，无运行时代码变更。

功能与动机

DeepSeek-V4 部署命令生成器在多种硬件 / 配方组合下存在不准确或缺失的命令参数，可能导致部署失败或性能不佳。具体问题包括：Balanced 配方启用 MegaMoE 后 draft pass 显存不足；H200 Pro FP4 低延迟配方 --mem-fraction-static 过高导致 OOM；nsa-* 标志名在 :latest 镜像上不可用（PR #25821 重命名尚未发布）；PD-Disagg 配方下 HiCache 和 MegaMoE 选项被静默忽略；parser 标志未正确添加到 decode role。

实现拆解

1. Balanced + MegaMoE 显存上限：在 Balanced 配方中启用 MegaMoE 时，为命令添加环境变量 SGLANG_OPT_DEEPGEMM_MEGA_MOE_NUM_MAX_TOKENS_PER_RANK=4096，限制每 rank 的 dispatch 缓冲区大小，避免 draft pass 耗尽显存。
2. H200 Pro FP4 低延迟配方显存比例调整：将 --mem-fraction-static 从默认的 0.88 调低至 0.83，为 MTP 多 token 并行需求预留更多显存余量。其他 H200 FP4 Pro 配方保持不变。
3. 上下文并行标志名兼容性：由于 :latest 发布镜像在 PR #25821 之前构建，该 PR 将 --enable-nsa-prefill-context-parallel 重命名为 --enable-dsa-prefill-context-parallel。生成器现在输出旧名，并在命令前插入 5 行注释提醒使用 main 分支的用户手动替换 nsa- 为 dsa-。
4. PD-Disagg 时禁用 HiCache 和 MegaMoE：
   - 新增 HICACHE_UNSUPPORTED_RECIPES 常量（包含 "pd-disagg"）和 isHicacheUnsupported 函数。
   - 在 resolveItems 中为 hicache 选项添加禁用逻辑，显示工具提示说明原因。

- 将 "pd-disagg" 加入 MEGAMOE_UNSUPPORTED_RECIPES，并在禁用理由中区分说明“PD-Disagg 生成器尚未支持”。
- 在配方切换回调 handleRadioChange 中，当切换到 pd-disagg 时自动将 hicache 和 megamoe 重置为 disabled。

5. PD-Disagg 下 parser 标志放置到 decode role：在 buildRole("decode", …) 中添加 --reasoning-parser 和 --tool-call-parser 参数（当用户在 UI 中启用时），prefill role 不包含这些参数。这是因为 PD HTTP 路由器只返回 decode server 的响应，格式解析只在 decode 端执行。

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

唯一修改文件，包含所有五项修复的核心逻辑变更。

```
// 新增：HiCache 在 PD-Disagg 配方下不支持（生成器尚未发射相关标志）
const HICACHE_UNSUPPORTED_RECIPES = new Set(["pd-disagg"]);
const isHicacheUnsupported = (vals) =>
  HICACHE_UNSUPPORTED_RECIPES.has(vals.recipe);

// 在 resolveItems 中处理 hicache 选项禁用
if (option.name === "hicache" && vals && isHicacheUnsupported(vals)) {
  return option.items.map((it) =>
    it.id === "disabled"
      ? it
      : { ...it, disabled: true, disabledReason: "HiCache is not yet wired into the PD-Disagg
      cookbook command" }
  );
}

// 配方切换时自动将 hicache 重置为 disabled
if (
  optionName === "recipe" &&
  next.hicache !== "disabled" &&
  isHicacheUnsupported(next)
) {
  next.hicache = "disabled";
}
```

评论区精华

无 review 评论。本 PR 由一人提交、另一人直接批准，无讨论。

风险与影响

- 风险：仅修改文档工具生成器，无运行时代码变更，风险极低。唯一可能的风险是 nsa-* 回退导致 main 分支用户使用旧标志，但注释已明确提示替换方式。PD-Disagg 下禁用 HiCache 和 MegaMoE 是临时措施，未来需实现完整支持。
- 影响：直接改善使用 cookbook 的部署体验，确保生成命令在指定组合下正确无错。对现有部署无影响。

关联脉络

- 关联 PR #25821（nsa-→ dsa- 重命名），本 PR 的 nsa-* 兼容性修复直接依赖该 PR 的改动。
- 属于 DeepSeek-V4 部署工具链的持续完善，后续可能进一步实现 PD-Disagg 下的 HiCache/MegaMoE 完整支持。