PR #33157 完整报告

sgl-project/sglang

[Docs] Add RTX 5090 DeepSeek-V4 recipe

合并时间：2026-08-01 10:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/33157

执行摘要

本 PR 为 DeepSeek-V4 cookbook 增加 RTX 5090（32GB）硬件登记与一组 Flash Official / FP4 / low-latency / 单节点的部署 cell，使 8 卡 RTX 5090 用户可以直接生成 DeepSeek-V4-Flash-0731 的部署命令。变更仅触及 docs_new/src/snippets/configs/deepseek-ai/deepseek-v4.jsx 这一配置数据文件，不涉及运行时逻辑，benchmark 卡片保持原样，新 cell 明确标注 Not Verified。

功能与动机

PR body 的目标是 "Expose a working DeepSeek-V4-Flash-0731 recipe for users serving the FP4 checkpoint on a single node with eight RTX 5090 GPUs"。RTX 5090 是 32GB 显存的 Blackwell 桌面消费卡，与数据中心卡差异较大，原有 cookbook 缺乏针对它的单节点 TP8 配置，用户只能自行拼参数。该 PR 补齐这条部署入口，并且通过 verified: false 保持 " 命令可用但未跑基准 " 的透明状态。

实现拆解

1. 硬件登记：在 supportedHardware 数组中追加 rtx5090 标识，并在模型专属 hardware 列表中登记 RTX 5090（label RTX 5090、vram 32GB、vendor blackwell）。cookbook 引擎会把该列表合并进共享的 HARDWARE_CATALOG，因此只需配置数据，不改引擎。
2. 新增部署 cell：在 RTX 6000 cell 之后插入匹配 hw: rtx5090、variant: flash-official、quant: fp4、strategy: low-latency、nodes: single 的部署单元。命令包含 --tp 8、--moe-runner-backend flashinfer_mxfp4、--mem-fraction-static 0.90、--cuda-graph-max-bs-decode 32，verified 置为 false。
3. 配套校验：使用 Node.js 断言检查硬件元数据、部署元组唯一性、Not Verified 状态和具体 flags，并执行 mint validate 与 mint broken-links；由于是 cookbook 配置变更，没有新增运行时测试文件。
4. 提交演进：从设计 recipe、正式添加，到 "chore: keep PR scoped to cookbook config" 主动收窄改动范围，最终合并 main，体现作者控制 PR 边界的意图。

关键源码片段

docs_new/src/snippets/configs/deepseek-ai/deepseek-v4.jsx

唯一变更文件，承担 DeepSeek-V4 cookbook 全部配置数据：新增 RTX 5090 硬件登记和对应的 Flash Official / FP4 / low-latency 单节点部署 cell，直接决定文档页面生成的部署命令。

```
// DeepSeek-V4 cookbook 配置：新增 RTX 5090 硬件元数据
export const config = {
  modelName: "DeepSeek-V4",

  // 共享 HARDWARE_CATALOG 之外的模型专属 GPU 登记，
  // RTX 5090 是 SM120 / Blackwell Desktop 消费卡，不是数据中心卡
  supportedHardware: [
    "h100", "h200", "b200", "b300", "gb200", "gb300",
    "rtx6000", "rtx5090",
    // AMD ROCm — MI300X (Flash FP8) + MI355X (Flash/Pro, FP4/FP8)
    "mi300x", "mi355x",
  ],

  // 模型专属 GPU 定义：cookbook 引擎会合并进 HARDWARE_CATALOG
  hardware: [
    { id: "rtx6000", label: "RTX PRO 6000", vram: "96GB", vendor: "blackwell" },
    // 32GB 显存消费卡，8 卡合计 256GB，适合 FP4 284B 模型
    { id: "rtx5090", label: "RTX 5090", vram: "32GB", vendor: "blackwell" },
  ],
};

// 新增部署 cell：Flash Official / FP4 / low-latency / 单节点 8 卡
// verified 为 false —— 命令在 8 卡 RTX 5090 devbox 手动验证过
//  （BS=1、ISL=100000、OSL=1000），但未跑 benchmark，保持 Not Verified
{
  match: { hw: "rtx5090", variant: "flash-official", quant: "fp4", strategy: "low-latency", nodes:
  "single" },
  verified: false,
  env: [],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 8", // 8 卡张量并行
    "--moe-runner-backend flashinfer_mxfp4", // FP4 走 FlashInfer MXFP4 MoE 后端
    "--mem-fraction-static 0.90", // 32GB 卡上预留 KV 缓存上限
    "--cuda-graph-max-bs-decode 32",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
}
```

评论区精华

mintlify[bot]：Preview deployment … ❔ Failed（未给出失败原因，提示可启用 Workflows）

这是唯一有信息量的评论；zijiexia 直接 APPROVED，无人工 review 讨论。gemini-code-assist[bot] 提示其审查服务已停用，因此没有产生技术交锋。

风险与影响

- 命令未过 benchmark：cell 标记 verified: false，用户应自行验证负载表现；--mem-fraction-static 0.90 在 32GB 卡上偏激进，极端负载可能 OOM。
- 配置耦合：新 cell 依赖 modelNames 中 flash-officiallfp4 → deepseek-ai/DeepSeek-V4-Flash-0731 的既有映射，未来重构配置结构时容易产生隐藏依赖。
- 文档预览不可用：Mintlify preview 在合入前处于 Failed 状态，渲染效果未在预览站点核对，但未阻塞合并。
- 影响范围：仅文档站点与 cookbook 命令生成，无运行时影响；为 RTX 5090 用户提供了一条可参考的部署路径。

关联脉络

本 PR 是 #33083（DeepSeek-V4 Flash Official 0731 recipe）在消费级硬件上的扩展，两个 PR 共享同一个 deepseek-v4.jsx 配置源文件；同时它与 #33131（Inkling-Small DGX Spark cookbook）同属 cookbook 配置体系，展示了 " 硬件元数据 + 部署 cell" 的扩展模式。该模式未来可以继续覆盖其他 Blackwell 桌面 / 消费级 GPU。