PR #32834 完整报告

sgl-project/sglang

[docs] Kimi-K3: widen the H200 High-Throughput recipe to 4x8 TP32/EP32

合并时间：2026-07-30 08:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/32834

执行摘要

该 PR 修复了 Kimi-K3 文档中 H200 Unified High-Throughput 节点的配置错误：原配方是 Balanced 配置的复制（2×8 TP16/EP16），与实际运行的 TP32/EP32 跨 4 节点方案不符。更新了 JSX 配置文件中的节点数、TP/EP 大小、跨节点网络环境变量等，并同步修改了 MDX 文档中的相关描述。纯文档变更，无运行时代码修改，风险极低。

功能与动机

PR body 明确指出：Kimi-K3 cookbook 中 H200 Unified High-Throughput 单元格是 Balanced（2×8 TP16/EP16）的复制品，仅修改了 `--mem-fraction-static 0.90` 和 `extra_buffer_lazy`，而实际运行的操作点是跨 4 节点的 TP32/EP32，因此页面提供了无人使用的配方。本次修改将配方修正为实际运行的配置。

实现拆解

1. 更新 JSX 配置文件（docs_new/src/snippets/configs/moonshotai/kimi-k3.jsx）：
   - 将 H200 High-Throughput 单元格的 nnodes 从 2 改为 4，--tp-size 和 --ep-size 从 16 改为 32。
   - 在 env 数组中添加跨节点网络环境变量：SGLANG_HOST_IP={{LOCAL_IP}}、NCCL_SOCKET_IFNAME={{NETWORK_IFACE}}、GLOO_SOCKET_IFNAME={{NETWORK_IFACE}}，这些与 H100 4×8 单元格已有的变量一致。
   - 添加性能调优环境变量：PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True、SGLANG_K3_ATTN_RES_MODE=jit、SGLANG_MOE_FUSED_GATE_RADIX=1。
   - 在 flags 中添加 --dist-timeout 3600（分布式超时 1 小时），适应跨节点启动。
   - 更新单元格注释，记录实际运行结果：DSPARK 和 HiCache L1+L2 分层，ratio 0.058，每 GPU 剩余 12.54 GB 空闲，1940352 KV tokens。
   - 同时更新文件顶部的注释，H200 行增加 "or 4×8 TP32/EP32 for High-Throughput" 说明。
2. 更新 MDX 文档正文（docs_new/cookbook/autoregressive/Moonshotai/Kimi-K3.mdx）：
   - 硬件食谱行：在 H200 描述中增加 "4×8 on Unified High-Throughput"。
   - Strategy 列表：High-Throughput 条目中说明 "on H200 the cell itself widens to 4×8 TP32/EP32 at --mem-fraction-static 0.90"。

- 平台表格中的 H200 行：更新为 "2×8 (4×8 on Unified High-Throughput)"，并补充 High-Throughput 的具体配置。

3. 验证：PR 提交者确认 mint validate 通过，并在本地 mint dev 渲染验证生成的命令正确。

docs_new/src/snippets/configs/moonshotai/kimi-k3.jsx

核心变更文件：修复 H200 High-Throughput 单元格的节点数、TP/EP 大小、添加跨节点网络环境变量和性能调优参数。

```
{
  // The one H200 cell that widens past a single pair of nodes. As run — with
  // DSPARK and HiCache L1+L2 layered on, at ratio 0.058 — the static
  // allocation leaves 12.54 GB free per GPU and 1940352 KV tokens.
  match: { hw: "h200", pdMode: "unified", strategy: "high-throughput" },
  nnodes: 4, // 从 2 改为 4，对应跨 4 节点部署
  verified: false,
  verificationStatus: "in-progress",
  env: [
    "NCCL_MNNVL_ENABLE=1",
    "NCCL_CUMEM_ENABLE=1",
    "PYTORCH_CUDA_ALLOC_CONF=expandable_segments:True", // 新增：启用 expandable
    segments 避免内存碎片
    "SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK=0",
    "SGLANG_K3_ATTN_RES_MODE=jit", // 新增：K3 attention residual 模式设为 JIT
    "SGLANG_MOE_FUSED_GATE_RADIX=1", // 新增：MoE fused gate radix 设为 1
    "SGLANG_HOST_IP={{LOCAL_IP}}", // 新增：跨节点通信必需的本机 IP
    "NCCL_SOCKET_IFNAME={{NETWORK_IFACE}}", // 新增：指定 NCCL 使用的网络接口
    "GLOO_SOCKET_IFNAME={{NETWORK_IFACE}}", // 新增：指定 Gloo 使用的网络接口
  ],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp-size 32", // 从 16 改为 32
    "--ep-size 32", // 从 16 改为 32
    "--moe-runner-backend marlin",
    "--decode-attention-backend flashmla",
    "--enable-symm-mem",
    "--mem-fraction-static 0.90",
    "--mamba-radix-cache-strategy extra_buffer_lazy",
    "--dist-timeout 3600", // 新增：分布式超时设为 1 小时，适应跨节点启动
    "--reasoning-parser kimi_k3",
    "--tool-call-parser kimi_k3",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
```

评论区精华

该 PR 没有引发讨论。审核者 wisclmy0611 直接批准。Mintlify 预览部署机器人自动发布了预览链接。

风险与影响

- 风险：纯文档变更，无运行时代码修改，风险极低。唯一潜在风险是配置文本错误导致用户误用，但 PR 基于实际运行验证，且已验证 mint validate 通过和本地渲染正确。
- 影响：影响范围限定在 Kimi-K3 cookbook 文档的 H200 部分。用户将看到正确的 High-Throughput 配方（4×8 TP32/EP32），避免部署 2×8 错误配置。其他硬件（B200、H100 等）和策略（Low-Latency、Balanced）不受影响。

关联脉络

该 PR 是独立的文档修正，与近期其他 PR 无直接关联。但反映了 Kimi-K3 文档持续根据实际运行经验进行校正的趋势（类似 PR#32672 对 KV cache 配置的修复）。