

PR #32592 完整报告

sgl-project/sglang

docs(cookbook): update Kimi-K3 GB200 recipes from measured 4x4 runs

合并时间: 2026-07-29 07:26

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/32592>

执行摘要

本 PR 基于 4x4 GB200 实测数据更新了 Kimi-K3 的 PD 部署配置，新增 prefill cell，调整 decode 并行策略，并添加 multiNodeHints。所有 GB200 cell 仍标记为未验证，但提供了详尽的性能对比和配置选择说明。

功能与动机

Kimi-K3 的 GB200 配置之前是从 GB300 配方转录的，未经实测。本 PR 用实际测量结果替换了之前的估算，填补了缺失的 prefill cell 和 multiNodeHints，确保用户能够基于可靠数据部署。

实现拆解

1. 新增 prefill cell: 在 kimi-k3.jsx 中添加了两个 GB200 prefill 配置 (default 和 long-context)。Default cell 采用 TP1 x PP16 拓扑，根据实测在并发 32 时 prefill tok/s/GPU 比 PP8 x TP2 高 26%。Long-context cell 将 mem-fraction-static 提升至 0.90，保持 TP=1 以最大化 KV 容量。
2. 更新 decode cell: 基于四种拓扑的对比数据调整了配置。Low-latency 从 tp16 改为 tp8 pp2，以满足 decode pp_size 与 prefill pp_size 一致或为 1 的约束。Balanced 添加了 ep16，实测显示 EP16 在 DCP16 基础上提升 throughput 6-10%。High-throughput 将 DCP16 改为 dcp8 dp2 ep16，以利用 MegaMoE 在 dp>1 时的优势。
3. 添加 multiNodeHints: 新增 multiNodeHints.gb200 条目，包含 MNNVL 环境变量、NVL72 域和 JIT 缓存路径等设置。
4. 补充性能注释: 在每个 cell 中添加了详细的性能说明和权衡解释，例如 prefill 在低并发时推荐 TEP16，decode 在不同并发区的领先策略等。
5. 验证状态保持: 所有 GB200 cell 仍标记为 verified: false 和 verificationStatus: "in-progress"，因为 GSM8K 精度验证尚未完成。

关键源码片段

[docs_new/src/snippets/configs/moonshotai/kimi-k3.jsx](#)

唯一变更文件，包含了所有 GB200 配置的更新和新增，是 PR 的核心。

```
// GB200 prefill default cell: PP16 x TP1 spans all 16 ranks.  
// Prefill tok/s/GPU at ISL 8192, concurrency 32: 4550 here vs 3596 (PP8 x TP2),  
// 2407 (TEP16), 1652 (TP16); flat past 32. Below concurrency ~8 the pipeline
```

```
// cannot fill and TEP16 leads instead (1947 vs 1227).
match: { hw: "gb200", pdMode: "prefill", strategy: "default" },
nnodes: 4,
verified: false,
verificationStatus: "in-progress",
env: [
  "SGLANG_ENABLE_TP_MEMORY_INBALANCE_CHECK=0",
],
flags: [
  "--trust-remote-code",
  "--model-path {{MODEL_NAME}}",
  "--tp-size 1",
  "--pp-size 16",
  "--mem-fraction-static 0.85",
  "--chunked-prefill-size 16384",
  "--max-prefill-tokens 16384",
  "--disable-flashinfer-autotune",
  "--weight-loader-prefetch-checkpoints",
  "--reasoning-parser kimi_k3",
  "--tool-call-parser kimi_k3",
  "--disaggregation-mode prefill",
  "--disaggregation-transfer-backend nixl",
  "--disaggregation-bootstrap-port 8998",
  "--host {{HOST_IP}}",
  "--port {{PORT}}",
],
```

评论区精华

- kpham-sgl 询问是否包含 GSM8K 或 AIME26 验证结果，否则应保持 false。YAMY1234 回应已标记为 false 并解决了冲突。讨论确认了这些配置尚未经过精度验证。

风险与影响

- 风险：配置未经验证（verified: false），用户若直接用于生产可能遇到精度问题。但性能数据基于实测，且 GSM8K 在三种配置下准确率约为 0.975 附近，初步表明 accuracy-neutral。multiNodeHints 中的路径可能需要根据实际环境调整。
- 影响：仅影响 GB200 上的 Kimi-K3 部署配置，提供优化基准，减少手动调优成本。

关联脉络

本 PR 是 #32542 的后续，后者可能引入了最初的 Kimi-K3 recipe。此外，与近期 #32834（文档轮播模型）和 #32835（文档首页改进）无直接关系，但都属于 Kimi-K3 文档的持续完善。