

PR #33612 完整报告

sgl-project/sglang

Remove custom all-reduce disable from Kimi-K3 B300 recipe

合并时间: 2026-08-05 07:07

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/33612>

执行摘要

该 PR 从 Kimi-K3 的 B300 Unified + Balanced 部署配方 ([docs/src/snippets/configs/moonshotai/kimi-k3.jsx](#)) 中删除了一行 `--disable-custom-all-reduce`, 目的是让该配方生成的命令使用默认的自定义 all-reduce 路径。变更仅涉及一个文件的单行删除, 不影响其他 B300 策略或 PD 配方。

功能与动机

PR body 指出: The B300 Balanced cell still carried an explicit custom all-reduce opt-out, so the generated deployment command disabled that path for this recipe. 也就是说, B300 Balanced 配方仍然显式携带了 custom all-reduce 的退出开关, 导致基于该配方生成的部署命令禁用了这条路径。本 PR 的目标是让选中的命令——包括 HiCache 关闭时的 DSPARK 组合——使用默认自定义 all-reduce 路径, 修正配方与实际期望行为之间的不一致。

实现拆解

1. 定位变更入口: docs/src/snippets/configs/moonshotai/kimi-k3.jsx 中的 B300 Unified + Balanced 配置单元 (match: { hw: "b300", pdMode: "unified", strategy: "balanced" })。
2. 删除 flags 数组中的 --disable-custom-all-reduce 一项 (+0/-1)。这是本次唯一改动, 没有新增或修改其他逻辑。
3. 保持其他单元不变: B300 Unified 的 low-latency 和 high-throughput 单元, 以及其余 PD 配方均未改动; high-throughput 单元仍保留 --disable-custom-all-reduce。
4. 验证配套: PR body 提到对该 hw=b300, pdMode=unified, strategy=balanced, spec=dspark, hicache=off 命令组合做了 Node 断言, 并运行了 mint validate、mint broken-links 以及仓库 pre-commit hooks。不过本次提交没有附带测试文件变更, 相关断言可能位于文档工具链内部。

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

这是本次唯一变更文件, 删除了 B300 Unified + Balanced 配方中的 `--disable-custom-all-reduce`, 直接影响该配方生成部署命令的内容。

```
// docs/src/snippets/configs/moonshotai/kimi-k3.jsx
// B300 Unified + Balanced 配方: 删除 `--disable-custom-all-reduce`,
// 让 DSPARK + HiCache off 组合走默认 custom all-reduce 路径。
{
  match: { hw: "b300", pdMode: "unified", strategy: "balanced" },
```

```
nnodes: 1,
verified: false,
verificationStatus: "in-progress",
env: [],
flags: [
  "--trust-remote-code",
  "--model-path {{MODEL_NAME}}",
  "--tp-size 8",
  "--dcp-size 8",
  // 注意: high-throughput 单元仍保留 `--disable-custom-all-reduce`
  "--mem-fraction-static 0.85",
  "--reasoning-parser kimi_k3",
  "--tool-call-parser kimi_k3",
  "--host {{HOST_IP}}",
  "--port {{PORT}}",
],
}
```

评论区精华

本 PR 没有 review 评论或讨论线程，决策与验证均由作者 Fridge003 独立完成。从 body 看，作者明确声明影响范围仅限 B300 Unified + Balanced 单元，并指明其他 B300 策略与 PD 配方保持不变。

风险与影响

- 这是文档 / 配置示例变更，运行时代码未改动，整体风险低。
- 影响用户生成的部署命令：使用该配方时，命令将不再禁用 custom all-reduce，DSPARK + HiCache off 场景将走默认 custom all-reduce 路径；若该路径在 B300、DSPARK 组合下存在未覆盖的稳定性问题，可能需要回退。
- 需要注意 high-throughput 单元仍保留 --disable-custom-all-reduce，说明在大规模场景该开关仍被视为必要；移除 balanced 的开关后，两条策略的行为将不一致，建议后续观察线上表现。
- 无直接测试文件变更，验证依赖文档工具链与 Node 断言，覆盖面有限。

关联脉络

该 PR 是对 Kimi-K3 支持主线的后续收尾：PR #32541 落地了 Kimi K3 Day-0 支持（模型 / VLM / KDA 内核等），本 PR 修正其 B300 部署配方。另可关注 PR #33098（修复 DSpark 与 DP/EP），因为本 PR 涉及的正是 DSPARK + HiCache off 组合下的 all-reduce 配置，两者在组合场景上相关。