PR #35168 完整报告

sgl-project/sglang

docs: add NVFP4 quantization option to Kimi-K3 deploy panel

合并时间：2026-08-18 02:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35168

执行摘要

本 PR 为 Kimi-K3 cookbook 部署面板新增一个与网格正交的 Quantization维度：默认 MXFP4（Moonshot AI checkpoint），可选 NVFP4（NVIDIA ModelOpt checkpoint）。选中 NVFP4 后，面板自动切换模型 slug 为 nvidia/Kimi-K3-NVFP4、固定 --moe-runner-backend flashinfer_trtllm，并将 Docker 镜像切换为基于 #35077 构建的 lmsysorg/sglang:dev-dev-kimi-k3-nvfp4。该功能仅对 Blackwell（B200/GB200/B300/GB300）开放，Hopper/AMD 在面板层直接禁用。这是一次纯文档 / 面板配置变更，未触碰任何运行时代码，但对 GB300/B200 用户提供了可直接执行的 NVFP4 部署路径（AIME26 pass@1 = 0.900 vs bf16 0.908）。

功能与动机

Kimi-K3 的 NVFP4 checkpoint 由 NVIDIA ModelOpt 产出，其 routed experts 依赖 FlashInfer TRT-LLM 的 NVFP4 SiTU 内核，部署约束非常苛刻：auto runner 解析永远不会启用 TRT-LLM deferred finalize，flashinfer_cutlass 没有 SiTU 内核，NVFP4 MoE 在 CUDA graph 捕获时还会抛 NotImplementedError。PR body 的目标很直接：把这一选项和约束直接呈现到部署面板，让用户无需手工拼接命令即可生成正确的 NVFP4 部署配方，并附上 GB300 2×4（TP8/DCP8 + DSPARK）的精度验证数据。

实现拆解

1. 新增 quant overlay 维度：在 docs/src/snippets/configs/moonshotai/kimi-k3.jsx 的 overlayDims 顶部插入 id: "quant" 维度，与 matchDims 网格正交，避免选项增多导致单元格数量膨胀。mxfp4 为默认，nvfp4 通过 disabled 回调限定 b200/gb200/b300/gb300，并附 disableReason 说明 Hopper/AMD 为何不可用。
2. 命令改写逻辑：NVFP4 选项携带 stripPrefixes: ["--moe-runner-backend"] 与 flags: ["--moe-runner-backend flashinfer_trtllm"]。必须先剥离单元格固定的 runner（如 B200 Balanced/High-Throughput 单元的 flashinfer_mxfp4、Hopper 单元的 marlin），再注入唯一可用的 TRT-LLM runner，否则会在 CUDA graph 捕获阶段崩溃。
3. 模型名与镜像联动：modelNames 增加 nvfp4: "nvidia/Kimi-K3-NVFP4"；dockerImages 增加 b300|nvfp4、gb300|nvfp4、b200|nvfp4、gb200|nvfp4 四个组合键，统一指向从 #35077 head 构建的 CUDA 13 dev 镜像。面板根据 modelNames[quant] 与 dockerImages[`${hw}|${quant}`] 自动完成同步切换。
4. 文档正文提示：docs/cookbook/autoregressive/Moonshotai/Kimi-K3.mdx 的 intro 段落追加一行，提醒 NVFP4 checkpoint 用户改用专用镜像。

5. 验证数据：PR body 给出 GB300 2x4（TP8/DCP8 + DSPARK）实测 AIME26 pass@1 = 0.900，相比 bf16 0.908 退化约 0.9 个百分点；DSPARK overlay 无需额外改动，同一套 draft checkpoint 可直接叠在 NVFP4 base 之上。

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

核心变更文件：新增 Quantization overlay 维度，定义 NVFP4 的硬件禁用、runner 覆盖、模型 slug 与 Docker 镜像联动，并给出全部技术论证注释。

```
// ---- 1. Quantization 作为正交 overlay 维度 ----
// overlayDims 与部署网格（matchDims）正交：选择量化方式不会膨胀单元格数量，
// 而是把选中项以 flags 形式叠加到当前单元格的命令上。
{
  id: "quant",
  title: "Quantization",
  default: "mxfp4", // MXFP4 是 Moonshot AI 官方发布的默认 checkpoint
  options: [
    { id: "mxfp4", label: "MXFP4", subtitle: "Moonshot AI checkpoint" },
    {
      id: "nvfp4",
      label: "NVFP4",
      subtitle: "NVIDIA checkpoint",
      // NVFP4 是 NVIDIA ModelOpt 混合精度 checkpoint：routed experts 走
      // NVFP4 SiTU 内核，attention 走 FP8_PB_WO 128x128 block-FP8。
      // 这两个内核族只存在于 Blackwell，所以直接按硬件禁用 Hopper/AMD。
      disabled: (s) => !["b200", "gb200", "b300", "gb300"].includes(s.hw),
      disableReason:
        "The nvidia/Kimi-K3-NVFP4 checkpoint needs Blackwell: its routed experts run on
        FlashInfer TRT-LLM NVFP4 kernels (SiTU), which do not exist for Hopper or AMD.",
      // 为什么必须固定 flashinfer_trtllm：
      // - auto 解析永远不会启用 TRT-LLM deferred finalize；
      // - flashinfer_cutlass 没有 SiTU 内核；
      // - NVFP4 MoE 在 CUDA graph 捕获时会抛 NotImplementedError。
      // stripPrefixes 负责移除单元格原本固定的 runner（如 B200 的
      // flashinfer_mxfp4），避免与这里的 flags 冲突。
      stripPrefixes: ["--moe-runner-backend"],
      flags: ["--moe-runner-backend flashinfer_trtllm"],
      hints: [
        "Use docker image lmsysorg/sglang:dev-dev-kimi-k3-nvfp4 (CUDA 13).",
      ],
    },
  ],
},

// ---- 2. NVFP4 选择联动 checkpoint 与镜像 ----
// 面板根据 quant 值读取 modelNames[quant] 与 dockerImages[`${hw}|${quant}`]，
// 在同一套命令模板里无缝切换模型 slug 和容器镜像。
modelNames: {
  default: "moonshotai/Kimi-K3",
  nvfp4: "nvidia/Kimi-K3-NVFP4",
```

```
  },
  dockerImages: {
    // ... 其余硬件沿用 kimi-k3 镜像 ...
    // NVFP4 需要包含 #35077 变更的构建；dev 镜像即从该 PR 的 head 产出。
    "b300lnvfp4": "lmsysorg/sglang:dev-dev-kimi-k3-nvfp4",
    "gb300lnvfp4": "lmsysorg/sglang:dev-dev-kimi-k3-nvfp4",
    "b200lnvfp4": "lmsysorg/sglang:dev-dev-kimi-k3-nvfp4",
    "gb200lnvfp4": "lmsysorg/sglang:dev-dev-kimi-k3-nvfp4",
  },
```

评论区精华

本 PR 没有内联 review 评论，审批人 zijiexia 直接 APPROVED。技术论证全部沉淀在 PR body 与上述源码注释中，最有价值的三个结论：

风险与影响

- 面板渲染逻辑：modelNames.nvfp4 与 dockerImages["hwlnvfp4"] 组合键依赖 cookbook 渲染器对 overlay dim 的既有支持。若组合键不被识别，选中 NVFP4 后命令可能仍携带默认 checkpoint/ 镜像，需要人工点击面板验证。
- 镜像外部依赖：lmsysorg/sglang:dev-dev-kimi-k3-nvfp4 是从 #35077 的 head 构建的 dev 镜像，若该 PR 未合入、镜像被清理或未发布，文档中的镜像地址将不可用。
- 禁用范围：目前仅覆盖 b200/gb200/b300/gb300，未来其他支持 NVFP4 SiTU 内核的 Blackwell 衍生硬件会被面板错误禁用，需后续扩展。
- CI 状态：PR Test (Extra) 失败（Run #32039654868），虽主测试通过，仍建议确认失败项与本改动无关后再合并。

对运行时引擎、调度器、Kernel 等系统模块零影响；这是把从内核兼容矩阵与实测中沉淀的部署约束固化到文档工具链的典型实践。

关联脉络

该 PR 是 Kimi-K3 NVFP4 部署链路的最后一环：#35077 负责构建包含 NVFP4 支持的镜像，本 PR 负责把该镜像与正确的 runner/checkpoint 组合暴露给用户。它与 #35044（GB300 上 Kimi-K2.5 NVFP4 nightly 测试）同属 Kimi 系模型在 Blackwell 平台启用 NVFP4 的验证与部署大方向，也与仓库中持续的 FP8/NVFP4 量化内核演进（如 #30519 的 MLA absorbed bmm、#34277 的 TMA-aligned scale）一脉相承。从更广的视角看，这套 "overlay dim + stripPrefixes + flags + disableReason" 的模式，为其他具备多量化选项的模型 cookbook 提供了一个可复用的面板设计模板。