PR #37380 完整报告

sgl-project/sglang

Revert "[AMD] Add GLM-5.3-Flash recipes for MI300X, MI325X, and MI355X (#36608)"

合并时间：2026-09-01 16:25

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/37380

执行摘要

PR#37380 是一个纯文档回滚操作，移除了 GLM-5.3-Flash cookbook 文档中为 AMD MI300X、MI325X 和 MI355X 硬件添加的所有部署配置、基准测试和说明文本。变更不涉及任何运行时代码，旨在将文档状态恢复到仅支持 NVIDIA GPU 的原始配置，同时保留了所有后续添加的 NVIDIA 侧增强功能。

功能与动机

本 PR 的动机是完全回滚 PR#36608 ([AMD] Add GLM-5.3-Flash recipes for MI300X, MI325X, and MI355X) 引入的内容。如 PR body 所述：“This is a docs-only revert. It does not touch the ROCm engine support from #36607, which stays on main.” 目的是从文档层面撤销 AMD ROCm 的支持路径，但保留相关的运行时引擎代码。由于 #36608 合并后，多个后续 PR（#36544, #36660, #36719, #36740, #37109）已修改相同文件，作者进行了手动冲突解决，确保只移除 AMD 内容，保留所有 NVIDIA 的演进。

实现拆解

1. 核心配置清理：在 docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx 中，从 supportedHardware 数组中删除了三个 AMD 硬件 ID。移除了所有针对这些 ID 的 disabled 条件函数（例如在 low-latency 策略、fp8-trtllm 后端、deep_gemm MoE 后端、以及 eagle/dflash 推理选项上）。简化了 isRecommendedSelection 和 disabled 判断，使其仅适用于 NVIDIA 硬件。删除了整个 AMD ROCm 的部署配方块。清理了 dockerImages 映射和相关注释。
2. 基准测试数据移除：在 docs/src/snippets/configs/zai-org/glm-5.3-flash-benchmarks.jsx 中，删除了与 hw: "mi300x"、hw: "mi325x" 和 hw: "mi355x" 匹配的三行基准测试记录，包括附带的 GSM8K 准确率数据和详细说明。
3. 文档正文修订：在 docs/cookbook/autoregressive/GLM/GLM-5.3-Flash.mdx 中，更新了 front-matter 描述、安装说明、策略可用性说明、精度描述等段落，移除了所有提及 AMD ROCm、ROCm 引擎 PR#36607、以及 AMD 特定部署注意事项（如 BF16 KV、TileLang DSA、Triton MoE 后端、关闭 CUDA graphs）的句子和段落。

docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx

核心配置文件，定义了 GLM-5.3-Flash cookbook 支持的硬件、策略、选项和部署配置。本次回滚移除了所有 AMD 硬件相关的配置项和条件判断。

以下是回滚后配置的关键部分，展示了硬件支持范围和策略选项如何被简化以仅覆盖 NVIDIA GPU： // 回滚后的硬件支持与策略选项（关键片段） exportconstconfig={ modelName:"GLM-5.3-Flash", // 仅支持 NVIDIA GPU，移除了 mi300x, mi325x, mi355x supportedHardware:["gb300","h100","h200","b200","b300","gb200"], matchDims:[{ id:"strategy", title:"Strategy", options:[// 移除了 low-latency 选项上针对 AMD 硬件的 disabled 函数 {id:"low-latency",label:"Low Latency",subtitle:"Adaptive MTP 5/1/6"}, {id:"high-throughput",label:"High Throughput",subtitle:"Spec decode off"},], }, // ... 其他维度], // 推荐选择逻辑仅考虑 NVIDIA Hopper 硬件 isRecommendedSelection(s){ constpairing=["h100","h200"].includes(s.hw)?"bf16-tilelang":"fp8-trtllm"; // ... 返回值逻辑不变 }, // KV/DSA 后端选项：移除了针对 AMD 硬件的禁用条件 overlayDims:[{ id:"kvDsaPair", options:[{ id:"fp8-trtllm", label:"FP8 + TRT-LLM", // disabled 条件现在只针对 Hopper，移除了 AMD 硬件 disabled:(s)=>["h100","h200"].includes(s.hw), disableReason:"FP8 KV cache with TRT-LLM DSA is not supported on Hopper GPUs.", // ... flags }, // ...], }, // ...], // 推理选项：移除了针对 AMD 硬件的 disable 分支 playgroundFeatures:{ speculative:{ options:[// ... { id:"eagle", label:"EAGLE / Adaptive MTP 5-1-6", disable:[{ when:{dpAttnOn:[true]}, reason:"Adaptive MTP does not support DP-Attention...", }, // 移除了针对 AMD 硬件的禁用分支], }, // ...], }, }, // 部署配方：移除了整个 AMD ROCm 配方块 // ... };

评论区精华

本次变更无实质性的技术讨论。唯一的评审来自 wisclmy0611 的 APPROVE，表明这是一个直接、清晰的文档维护操作，符合预期。

风险与影响

1. 风险：极低。变更严格限制在文档文件内，不涉及任何运行时逻辑、核心数据结构或配置 schema。主要风险是手动解决冲突时可能误删非 AMD 内容，但 PR 声称已解决并通过了配置检查工具验证。
2. 影响范围：
   - 用户：仅影响那些参考 GLM-5.3-Flash cookbook 文档在 AMD MI300X/MI325X/MI355X 上部署模型的用户，他们将失去官方的配置指南。对 NVIDIA 用户无影响。
   - 系统：无运行时或部署影响。
   - 团队：明确了文档当前支持的硬件范围，简化了文档维护。相关的 AMD 运行时支持（PR#36607）仍在 main 分支的代码中，但未在用户面向的文档中体现。

关联脉络

- 直接关联 PR：本 PR 是 PR#36608的直接回滚。
- 衍生关联 PR：由于 PR#36608 为文档添加了 AMD 硬件 ID，后续的 PR#37109（添加 NVFP4）和 PR#36740（添加推理卡片）都在这些 ID 上添加了条件逻辑。本 PR#37380 在回滚时，将这些后来添加的、现在已变得不可达（dead code）的条件一并清理，确保了文档配置的干净状态。

- 持续演进：GLM-5.3-Flash 的文档持续在演进（如 PR#37109 的 NVFP4），但其硬件支持策略经历了这次调整。同时，AMD 的运行时支持（PR#36607）独立存在于代码库中，表明对 AMD 平台的支持可能处于一种“代码可用，文档未推广”的中间状态。