

PR #29194 完整报告

sgl-project/sglang

[AMD] [GLM5] GLM-5.1 MXFP4 (MI355X) + enable EAGLE for gfx950 in cookbook

合并时间: 2026-06-25 18:54

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29194>

PR 分析报告: GLM-5.1 MXFP4 + EAGLE 文档更新

执行摘要

本次 PR 为 GLM-5.1 部署文档补充了 MXFP4 量化选项 (MI355X/gfx950) 并更正了 EAGLE 投机解码在 AMD 上的支持状态。变更范围仅限两个文档配置文件, 无运行时核心代码改动, 风险极低。

功能与动机

PR body 指出: 现有的 GLM-5.1 Cookbook 页面缺少对 `amd/GLM-5.1-MXFP4` checkpoint 在 MI355X 上的部署说明, 且原来“EAGLE 在 AMD 上不支持”的表述对于 gfx950 已不准确 (已在硬件上经过 GSM8K 无损验证, 延迟和吞吐量均有显著提升)。本次变更旨在同步文档与实测能力, 消除用户歧义。

实现拆解

1. 交互式部署生成器 (`glm-51-deployment.jsx`)

- 新增 MXFP4 量化选项: 在 `getDynamicItems` 中添加 `mxfp4`, 仅当硬件选择 `mi355x` 时启用且默认选中; BF16 的默认逻辑从“所有 AMD”精确为“AMD 非 gfx950”。
- 删除 speculative UI 切换: 不再通过 on/off 开关控制 EAGLE, 改为在命令生成阶段直接根据硬件决定是否添加 EAGLE 参数 (gfx950 添加, gfx942 不添加)。
- 硬件配置扩展: `modelConfigs.mi355x` 增加 `mxfp4`: { `tp: 4`, `mem: 0.85` }, 命令生成时自动填入 `--model-path amd/GLM-5.1-MXFP4 --tp 4 --kv-cache-dtype fp8_e4m3` 及 AMD 特定的 `tilelang/chunked-prefill/watchdog` 标志。

2. Cookbook 页面 (`GLM-5.1.mdx`)

- 硬件表格新增 MXFP4 列: 为每个硬件平台新增对应列 (MI355X 显示 `tp=4`, 其余为“—”)。
- 添加 MXFP4 命令块: 在 4.2 节新增独立的命令行示例, 包含可选的 `SGLANG_DSA_TRITON_PREFILL=1` 环境变量和 EAGLE 参数。
- 更新 AMD 说明: 从“EAGLE 目前不受支持”改为“EAGLE 在 MI355X (gfx950) 上受支持, 在 MI300X/MI325X (gfx942) 上未验证”。

3. 无配套测试

本次变更仅涉及文档和 UI 配置, 未引入可执行的测试。

docs_new/src/snippets/autoregressive/glm-51-deployment.jsx

交互式部署生成器，核心变更文件：新增 MXFP4 量化选项、调整 EAGLE 生成逻辑、删除 speculative UI 切换。

// 该段选自 getDynamicItems 函数，展示量化选项的生成逻辑

```
getDynamicItems: (values) => {
  const hw = values.hardware;
  const isAMD = ['mi300x', 'mi325x', 'mi355x'].includes(hw);
  const isGfx950 = hw === 'mi355x'; // 仅 MI355X 支持 MXFP4
  const supportsNVFP4 = ['b300', 'gb300'].includes(hw);
  const isB300 = hw === 'b300';
  const isGB300 = hw === 'gb300';
  return [
    {
      id: 'mxfp4',
      label: 'MXFP4',
      subtitle: 'gfx950',
      default: isGfx950, // MI355X 默认选中
      disabled: !isGfx950,
      disabledReason: !isGfx950 ? 'MXFP4 verified on MI355X (gfx950)' : ''
    },
    {
      id: 'bf16',
      label: 'BF16',
      subtitle: 'Full Weights',
      default: isAMD && !isGfx950, // 非 gfx950 的 AMD 硬件默认 BF16
      disabled: !isAMD,
      disabledReason: supportsNVFP4 ? 'NVFP4 is recommended for this hardware' : 'FP8 is recommended for this hardware'
    },
    // fp8 和 nvfp4 选项保持不变（略）……
  ];
}
```

// modelConfigs 中为 MI355X 新增 mxfp4 配置

```
const modelConfigs = {
  // ...其他硬件...
  mi355x: {
    bf16: { tp: 8, mem: 0.80 },
    mxfp4: { tp: 4, mem: 0.85 } // 新增: MXFP4 使用 tp=4, 内存比例 0.85
  }
};
```

// 命令生成逻辑中，当选择 mxfp4 时输出对应参数

// （片段仅展示关键条件分支）

```
const generateCommand = () => {
  const { hardware, quantization } = values;
  const isGfx950 = hardware === 'mi355x';
  const cfg = modelConfigs[hardware][quantization];
```

```
let cmd = `python3 -m sglang.launch_server --model-path ${modelPath}`;  
cmd += `--tp ${cfg.tp}`;  
if (quantization === 'mxfp4') {  
  cmd += `--kv-cache-dtype fp8_e4m3`; // MXFP4 需要指定 KV 缓存 dtype  
}  
if (isGfx950) {  
  // EAGLE 参数: 仅对 gfx950 自动添加  
  cmd += `--enable-eagle --eagle-num-predict 3`;  
} else if (['mi300x', 'mi325x'].includes(hardware)) {  
  // gfx942 上不添加 EAGLE 参数 (未验证)  
}  
// .....  
};
```

评论区精华

无 reviewer 评论; PR 直接由 HaiShaw 批准通过。

风险与影响

- 风险极低: 仅修改文档和 UI 配置 JSX, 不涉及任何 Python 运行时或构建流程。
- 潜在误导: EAGLE 在 MI300X/MI325X 上的支持状态为“未验证”, 若后续验证通过但未及时更新文档, 可能导致用户低估可用性。
- 影响: 部署 GLM-5.1 到 MI355X 的用户可直接通过交互式生成器获得 MXFP4 配置并自动启用 EAGLE, 减少手动调参成本。

关联脉络

本 PR 是 PR #28975 (opt-in Triton fp8 sparse-MLA prefill kernel for gfx950) 的后续文档更新, 在 cookbook 中为 gfx950 启用相关标志。后续可能跟进 gfx942 EAGLE 验证文档。