

PR #29313 完整报告

sgl-project/sglang

[AMD] [GLM5] Mark EAGLE verified on MI300X/MI325X (gfx942) in GLM-5.1 cookbook

合并时间: 2026-06-26 10:49

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/29313>

执行摘要

此 PR 将 GLM-5.1 的 EAGLE 投机解码验证范围从 gfx950 (MI355X) 扩展到 gfx942 (MI300X/MI325X)，更新了 cookbook 文档和交互式部署生成器。同时记录了 AMD GPU 上必须使用 `--disable-custom-all-reduce` 以避免 aiter 自定义 all-reduce 内核死锁的要点。

功能与动机

作为 #29194 的跟进，该 PR 此前在 GLM-5.1 cookbook 中为 gfx950 启用了 EAGLE，但 MI300X/MI325X (gfx942) 仍被标记为“未验证”。EAGLE 投机解码已在 gfx942 上验证通过，因此需要更新文档和部署生成器，让用户能够在这类 GPU 上使用 EAGLE。

实现拆解

1. 更新文档 (GLM-5.1.mdx):

- 将 AMD GPU 的 EAGLE 支持描述从“仅 MI355X 支持，MI300X/MI325X 未验证”改为“所有 AMD GPU 均支持 (gfx942 和 gfx950)”。
- 添加说明：EAGLE 需要 `--disable-custom-all-reduce` 标志以避免 aiter 自定义 all-reduce 内核在并发高时死锁。
- 为 FP8 和 BF16 的静态服务器命令块添加了 EAGLE 标志 (`--speculative-algorithm` 等) 以及 `--disable-custom-all-reduce`。

2. 更新部署生成器 (glm-51-deployment.jsx):

- 移除之前防止 gfx942 硬件选择时生成 EAGLE 标志的条件判断，现在所有硬件默认生成 EAGLE 标志。
- 当选择 AMD 硬件时，额外添加 `--disable-custom-all-reduce` 标志。

[docs_new/src/snippets/autoregressive/glm-51-deployment.jsx](#)

交互式部署生成器的核心逻辑变更：移除 gfx942 的 EAGLE 排除条件，为 AMD 添加 `--disable-custom-all-reduce` 标志。

关键源码片段

[docs_new/src/snippets/autoregressive/glm-51-deployment.jsx](#)

交互式部署生成器的核心逻辑变更：移除 gfx942 的 EAGLE 排除条件，为 AMD 添加 `--disable-custom-all-reduce` 标志。

```
// docs_new/src/snippets/autoregressive/glm-51-deployment.jsx (关键片段)
```

```
// ... 之前代码省略 ...

// EAGLE MTP speculative decoding: emitted by default (recommended) on all
// hardware. 现已验证可通过 NVIDIA 和 AMD MI300X/MI325X (gfx942) 和 MI355X (gfx950)。
cmd += ' \\\
--speculative-algorithm EAGLE';
cmd += ' \\\
--speculative-num-steps 3';
cmd += ' \\\
--speculative-eagle-topk 1';
cmd += ' \\\
--speculative-num-draft-tokens 4';

// On AMD GPUs 的 aiter custom all-reduce 内核在 EAGLE 验证时高并发下会死锁，
// 所以需要禁用它来避免服务器挂起。
if (isAMD) cmd += ' \\\
--disable-custom-all-reduce';

cmd += ` \\\
--mem-fraction-static ${memFraction}`;
return cmd;
// ... 后续代码省略 ...
```

评论区精华

- HaiShaw审核时建议将文档中的 all AMD GPUs: 改为 AMD GPUs:，作者 Raiden-Makoto 已采纳并修改。

风险与影响

风险：风险较低，因为变更主要涉及文档和部署配置的生成逻辑，没有对运行时代码进行修改。但需要注意的是，如果用户忽略了 `--disable-custom-all-reduce` 标志，在 AMD GPU 上运行 EAGLE 时可能会导致服务器死锁。文档已明确警示此问题。

影响：对 AMD GPU 上使用 GLM-5.1 的用户有正面影响，特别是 MI300X/MI325X 用户，现在可以获得 EAGLE 投机解码的加速。负面影响是用户必须正确设置 `--disable-custom-all-reduce` 标志，否则可能遇到死锁。

关联脉络

此 PR 是 #29194 的后续跟进，将 EAGLE 支持从 gfx950 扩展到 gfx942，完成了 AMD GPU 上 GLM-5.1 EAGLE 的完整支持。