

PR #39106 完整报告

sgl-project/sglang

[AMD] Use the triton DSA backend for GLM-5.2 MXFP4 on MI355X

合并时间: 2026-09-12 02:48

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39106>

[AMD] GLM-5.2 MXFP4 on MI355X 切换 triton DSA 后端 (PR#39106) 分析

执行摘要

PR#39106 是一次纯配方与文档层的后端选择修正: 把 AMD MI355X 上 GLM-5.2 MXFP4 (`amd/GLM-5.2-MXFP4`, TP4) 四个 Deploy 配方单元的 `--dsa-prefill-backend / --dsa-decode-backend` 由 `tilelang` 切换为 `triton`, 并在 Cookbook 同步更新说明。依据是 TP4 + `--kv-cache-dtype fp8_e4m3` 时每 rank 16 个 query head, 正好落在 gfx950 Triton 稀疏 MLA 内核的调优形状内, decode 内核实测 1.53x–2.72x 加速, 输出与 `tilelang` 一致; TP8 的 FP8/BF16 单元与 gfx942 单元因形状不匹配、且已发布数字基于 `tilelang` 而保持不变。改动共 2 个文件、+12/-9, 无源码与测试变更, reviewer HaiShaw 直接 approve 后合入。

功能与动机

GLM-5.2 的 MXFP4 配方此前直接沿用 `amd/GLM-5.1-MXFP4` MI355X 的 `tilelang` DSA 后端。PR body 指出 TP4 场景的形状其实更匹配 Triton 内核:

At `tp=4` with `--kv-cache-dtype fp8_e4m3` the shape is 16 query heads per rank on an FP8 KV cache, which is what the gfx950 Triton sparse-MLA kernels target.

Decode 内核在 GLM-5.2 形状 (`H=16`、`D_V=512`、`KV_DIM=576`、`topk=2048`、FP8 KV) 、CUDA-graph 捕获、7 次交错运行取中位数下的对比:

batch	tilelang (us)	triton (us)	speedup
4	24.02	14.33	1.68x
64	41.87	27.40	1.53x
128	64.54	33.48	1.93x
256	109.72	51.06	2.15x
512	199.96	73.60	2.72x

输出一致性为 SNR 37–82 dB、bf16 输出最大绝对差 5e-3, 端到端 AgentX (TP4/EP4) 与 1,319 题 GSM8K 证据在 #39059。PR body 同时声明 `These cells are verified: false and carry no benchmark numbers, so no published result changes.`, 因此该切换不会改动任何

对外发布的性能数字，这也是它可以只改配方、不带 E2E 数据的直接原因。

实现拆解

1) 配置入口：[docs/src/snippets/configs/zai-org/glm-5.2.jsx](#)

改动严格限定在 `mi355x + quant: "mxfp4"` 的四个单元，每个单元只改两行 flag，其余参数原样保留，保证“单变量”切换、便于回退验证：

配方单元（ <code>match</code> 关键维度）	本次改动	未动的关键参数
<code>mi355x / mxfp4 / low-latency</code>	两行 DSA 后端 <code>tilelang</code> → <code>triton</code>	<code>--chunked-prefill-size 131072</code> 、 <code>--mem-fraction-static 0.80</code>
<code>mi355x / mxfp4 / balanced</code>	同上	<code>--chunked-prefill-size 32768</code> 、 <code>--cuda-graph-max-bs-decode 128</code>
<code>mi355x / mxfp4 / high-throughput</code>	同上	<code>--cuda-graph-max-bs-decode 256</code> 、 <code>--max-running-requests 256</code>
<code>mi355x / mxfp4 / mtp-314</code>	同上	<code>EAGLE steps=3</code> 、 <code>--watchdog-timeout 1800</code>

2) 把选择依据内联成注释

MXFP4 段落的注释块新增三行，说明 `triton` 是 SGLang 在 ROCm 上的默认后端、TP4 形状命中 `gfx950 FP8 sparse-MLA` 调优、TP8 FP8/BF16 单元继续保留 `tilelang`。这一动作的意义在于把“为什么只改这四个单元”写在配方旁边，防止后续维护者为了统一风格把后端改回去。

3) 刻意不动的范围（回归边界）

`mi355x` 的 `bf16 / fp8` (`--tp 8`) 单元与 `mi325x` 的 `fp8` (`gfx942`) 单元全部保持 `tilelang`：TP8 时每 rank 只有 8 个 query head，不满足 `gfx950` 稀疏 MLA 的调优形状；更关键的是这些单元的已发布数字以 `tilelang` 测得，改动会让文档数字与配方不一致。

4) 文档同步：[docs/cookbook/autoregressive/GLM/GLM-5.2.mdx](#)

MI355X MXFP4 段落由“the same DSA `tilelang` backends”改为 DSA `triton` 后端，并补上 16 query heads per rank 的调优依据与“`tp=8` 配方继续使用 `tilelang`”的说明；文件尾部原有的 `gfx950 block-FP8` 精度修复说明（issue #28685、ROCm/rocm-libraries#8639、约 0.96 的 GSM8K）本次未改动，属于同一文档里相邻的 AMD 背景信息。

5) 测试 / 部署配套

本 PR 没有新增或修改测试、CI workflow 与镜像 tag；四个单元维持 `verified: false`，Deploy 面板仍不展示基准数字，改动只体现在面板生成出来的启动命令上。

[docs/src/snippets/configs/zai-org/glm-5.2.jsx](#)

本次变更的唯一功能承载点：`mi355x + quant: "mxfp4"` 的四个配方单元把 DSA 前后端由 `tilelang` 改为 `triton`，并新增三行注释固化选择依据，直接决定 Cookbook / Deploy 面板生成

的启动命令。

```
// =====
// AMD MI355X + MXFP4 — amd/GLM-5.2-MXFP4 (Quark) 。TP4: 4-bit MoE 权重
// 正好放进 4 卡切片，沿用已验证的 amd/GLM-5.1-MXFP4 MI355X 配方（同一
// DSA 架构族）—— --trust-remote-code (Quark 自定义量化配置) 与
// --kv-cache-dtype fp8_e4m3 都来自该先例。
// DSA 后端: triton, 这也是 SGLang 在 ROCm 上的默认后端。TP4 下该 shape
// 正好命中 gfx950 FP8 sparse-MLA 的调优区间; 上面的 TP8 FP8/BF16 单元
// 继续保留 tilelang, 因为它们已发布的数字是用 tilelang 测出来的。
// =====
{
  match: { hw: "mi355x", variant: "default", quant: "mxfp4", strategy: "low-latency", nodes:
    "single" },
  verified: false, // 尚未在 GLM-5.2 上做基准验证, Deploy 面板不展示数字
  env: [],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 4", // 每 rank 16 个 query head, 正好命中 gfx950 稀疏 MLA 的调优形状
    "--kv-cache-dtype fp8_e4m3", // FP8 KV cache 是走 gfx950 稀疏 MLA 快路径的前提
    "--dsa-prefill-backend triton", // 本次变更: tilelang → triton
    "--dsa-decode-backend triton", // decode 内核在该 shape 下实测 1.53x–2.72x 加速
    "--chunked-prefill-size 131072",
    "--mem-fraction-static 0.80",
    "--watchdog-timeout 1200",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
{
  // MTP (mtp-314) : 已验证的 AgentX 配方; steps=3 仍在 gfx950 投机解码的
  // 构建范围内 (≤3), --mem-fraction-static 0.80 为 MTP draft buffer 留余量
  match: { hw: "mi355x", variant: "default", quant: "mxfp4", strategy: "mtp-314", nodes: "single" },
  verified: false,
  env: [],
  flags: [
    "--trust-remote-code",
    "--model-path {{MODEL_NAME}}",
    "--tp 4",
    "--kv-cache-dtype fp8_e4m3",
    "--dsa-prefill-backend triton", // 四个 MXFP4 单元统一到 triton
    "--dsa-decode-backend triton",
    "--speculative-algorithm EAGLE",
    "--speculative-num-steps 3",
    "--speculative-eagle-topk 1",
    "--speculative-num-draft-tokens 4",
    "--chunked-prefill-size 131072",
    "--mem-fraction-static 0.80",
```

```
    "--cuda-graph-max-bs-decode 160",
    "--max-running-requests 160",
    "--watchdog-timeout 1800",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
],
},
```

评论区精华

本次 PR 的 review 记录极简：HaiShaw 给出 APPROVED，review body 为空，没有任何行内评论，也没有 issue 评论。换言之，技术论证全部由 PR 描述自身承担：

At `tp=4` with `--kv-cache-dtype fp8_e4m3` the shape is 16 query heads per rank on an FP8 KV cache, which is what the gfx950 Triton sparse-MLA kernels target.

These cells are **verified: false** and carry no benchmark numbers, so no published result changes.

未在讨论中展开但客观存在的疑虑是覆盖度：PR body 的 CI 状态里“AMD ROCm 10” slot 显示 **No AMD PR run found for this commit**，Extra 测试为失败（Run #34629833525），也就是这次后端切换没有获得仓库内的 AMD CI 覆盖。

风险与影响

- 缺少自动化性能 / 精度回归：一致性证据（SNR 37–82 dB、bf16 最大绝对差 5e-3）只在 PR 描述的离线性测试中，仓库内没有对应测试；四个单元仍是 `verified: false`，若上游 gfx950 Triton 稀疏 MLA 内核回归，Cookbook 配方会静默劣化。
- MTP 单元的外推风险：mtp-314 走 EAGLE 投机解码路径，而基准数据来自纯 decode 场景，且该单元的 `--cuda-graph-max-bs-decode 160` / `--max-running-requests 160` 未随后端调整。
- 文档口径不一致：GLM-5.2.mdx 的 AMD 总述段落仍写“Use the DSA tilelang backend”，只有 MI355X MXFP4 那个 bullet 改为 triton；读者可能误把 triton 推到 gfx942 或 TP8 场景。
- 影响面：仅影响复制这四段配方的 MI355X MXFP4 用户，收益是在给定 shape 下 decode 内核 1.5x–2.7x；TP8 FP8/BF16 与 MI300X/MI325X 用户、引擎源码行为、已发布 benchmark 数字均不受影响；回退成本仅为改回 4 处 flag。

关联脉络

- 与 PR#38611（Qwen3.8-27B cookbook 增加 NVFP4 导出）属于同一条“docs/src/snippets/configs/<org>/*.jsx 配方片段与 cookbook 成对更新”的维护脉络：同样按量化格式与硬件维度组织 recipe 单元，并用 `verified: false` 隔离未验证结论。
- 与 PR#34432（AMD aiter 后端接入 DCP）、PR#38993（抽取注意力图变体契约，触及 dsa/dsa_indexer.py）共同说明 ROCm 上 DSA/ 稀疏 MLA 的后端矩阵仍在演进：aiter、triton、tilelang 各自覆盖不同的形状与硬件代际，本 PR 只是把“哪种形状该用哪个后端”的判断固化进配方。

- 本 PR 引用的 #39059 提供了端到端 AgentX (TP4/EP4) 与完整 GSM8K 证据，是这次纯文档切换的信任基础；GLM-5.2.mdx 中相邻的 gfx950 block-FP8 精度事故背景 (issue #28685、ROCm/rocm-libraries#8639) 也提醒：AMD FP8/MXFP4 路径上的内核问题往往是“逐层小误差、整体输出崩坏”，因此本 PR 这类后端切换值得保留可复测的精度门限。