

PR #39572 完整报告

sgl-project/sglang

[AMD] Align Qwen3.5 MI355X HiCache cookbook with kernel / page_first

合并时间: 2026-09-16 11:22

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/39572>

执行摘要

该 PR 将 Qwen3.5 部署 cookbook 中 MI355X + MXFP4 + HiCache 路径的文档化参数从 `direct / page_first_direct` 回退为 `kernel / page_first`，与 InferenceX 实际 serving 命令对齐，避免 recipe-link 的 MAJOR-arg 检查失败。改动仅涉及命令生成器 JSX 与 cookbook 正文两个文件，共 6 行，不影响任何运行时逻辑，NVIDIA 示例保持不变。

功能与动机

PR body 明确说明：InferenceX MI355X Qwen3.5 MXFP4 AgentX 已把 HiCache 切回 `--hicache-io-backend kernel` 与 `--hicache-mem-layout page_first`（镜像 lmsysorg/sglang-rocm:v0.5.19-rocm720-mi35x-20260915），而 cookbook 仍输出 #39358 引入的 `direct / page_first_direct`，会触发 InferenceX recipe-link 的 MAJOR-arg 检查失败。因此需要让文档生成器与正文跟随真实部署命令。

实现拆解

- 命令生成器（`docs/src/snippets/autoregressive/qwen35-deployment.jsx`）：在 Qwen35Deployment 组件的 `quantization === 'fp4'` 且 `hardware === 'mi355x'` 且 `kvOffload === 'hicache'` 分支中，把 `--hicache-io-backend` 从 `direct` 改为 `kernel`，`--hicache-mem-layout` 从 `page_first_direct` 改为 `page_first`。该文件是 cookbook 页面“复制命令”的生成源头，修改后生成的命令即与 InferenceX 镜像实际使用的参数一致。
- cookbook 正文（`docs/cookbook/autoregressive/Qwen/Qwen3.5.mdx`）：同步更新 MI355X KV-offload 说明中列出的 flag 序列，保证文字与生成器输出一致。
- 范围控制：`--hicache-ratio 1.5`、`--hicache-write-policy write_through`、`AttnFP8-V2 checkpoint ID` 均保持不变；NVIDIA HiCache 示例不受影响。无测试配套变更（PR body 声明不影响模型输出，Accuracy Tests 为 N/A）。

`docs/src/snippets/autoregressive/qwen35-deployment.jsx`

cookbook 命令生成器源头，修改后所有从页面复制的 MI355X MXFP4 HiCache 命令都会输出 `kernel / page_first`，是本次对齐的核心文件。

关键源码片段

`docs/src/snippets/autoregressive/qwen35-deployment.jsx`

cookbook 命令生成器源头，修改后所有从页面复制的 MI355X MXFP4 HiCache 命令都会输出 `kernel / page_first`，是本次对齐的核心文件。

```

// docs/src/snippets/autoregressive/qwen35-deployment.jsx
// Qwen35Deployment 命令生成器：FP4 量化分支按硬件区分 AMD / NVIDIA 后端
if (quantization === 'fp4') {
  if (hardware === 'mi355x') {
    // AMD MXFP4 on MI355X: backend / --page-size 16 与 INT4 量化 env 由上方 AMD 块输出
    if (kvOffload === 'hcache') {
      // HiCache 在设备 KV 之下加一层 host-DRAM, radix cache 必须保持开启,
      // 因为 --enable-hierarchical-cache 与 --disable-radix-cache 启动时互斥
      cmd += ' \
--enable-hierarchical-cache';
      cmd += ' \
--hcache-ratio 1.5';
      cmd += ' \
--hcache-write-policy write_through';
      // 与 InferenceX 实际 serving 命令对齐: kernel 后端 + page_first 布局
      cmd += ' \
--hcache-io-backend kernel';
      cmd += ' \
--hcache-mem-layout page_first';
    } else {
      // 无 host 层时直接禁用 radix cache
      cmd += ' \
--disable-radix-cache';
    }
    // MXFP4 场景统一使用 FP8 E4M3 的 KV cache dtype
    cmd += ' \
--kv-cache-dtype fp8_e4m3';
  } else {
    // NVIDIA NVFP4 on Blackwell (B200 / B300), 走 flashinfer 后端参数
    cmd += ' \
--moe-runner-backend flashinfer_trtllm';
  }
}

```

评论区精华

该 PR 无任何 inline review 评论，两位审阅人 1am9trash 与 sogalin 均直接批准 (LGTM)，说明变更争议小、方向明确。值得注意的是，该 PR 是对 #39358 刚引入的 `direct / page_first_direct` 的短期回退，两次一前一后的修正都没有留下设计讨论痕迹，反映出 AMD 侧 HiCache 后端选型仍处于镜像演进驱动的快速调整期。

风险与影响

- 配置方向反复：`direct / page_first_direct` 在 #39358 引入，本 PR 即回退，说明 AMD HiCache 后端 (kernel vs direct) 仍在波动；若 InferenceX 镜像后续再切回 `direct`，文档需再次同步，否则会与真实命令脱节。
- 无自动化校验：本次没有新增测试或 CI 检查来锁定 cookbook 生成命令与 InferenceX recipe 的一致性，仍依赖外部 MAJOR-arg 检查兜底。

- 影响面受限：改动被 fp4、mi355x、hicache 三个条件限定，NVIDIA 分支与其余 AMD 分支不受影响；kernel / page_first 为既有选项，无兼容性风险。

关联脉络

- 39358: PR body 明确指出当前 cookbook 中的 `direct` / `page_first_direct` 来自该 PR，本 PR 是它的回头修正，两者共同构成 AMD MI355X HiCache 文档的“改入一改出”闭环。
- 39702 ([AMD] Update deepseek-v4 PDI and cache policy setting for agentic workload) : 同属 AMD 部署文档跟随真实 serving 配置的维护线，说明 AMD 文档与镜像 /recipe 的绑定关系已成为常态。
- 与 #39631 等 AMD 配置默认值调整一脉相承：AMD 侧的文档与配置变更常由镜像版本与内部部署实践驱动。