

PR #36245 完整报告

sgl-project/sglang

[AMD] cookbook: add HiCache host-DRAM KV tier for Qwen3.5 MXFP4 on MI355X

合并时间: 2026-08-25 12:55

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/36245>

执行摘要

本 PR 为 Qwen3.5 部署配置生成器 (docs/src/snippets/autoregressive/qwen35-deployment.jsx) 新增了“KV Cache Offloading”选项, 使得在 MI355X + MXFP4 组合下可以选择启用 HiCache (host-DRAM KV 缓存), 并自动生成相应的启动参数, 同时解决了与 `--disable-radix-cache` 的冲突。文档 cookbook 也相应更新, 新增 HiCache 使用说明并修正了 FP8 KV Cache 的警告信息。改动仅涉及文档和配置脚本, 无运行时代码影响。

功能与动机

InferenceX #2693 为 Qwen3.5 MXFP4 MI355X 的 AgentX 扫描添加了 HiCache host-DRAM KV 层, 但 cookbook 的部署生成器无法生成该配置: MXFP4-on-MI355X 配方总是输出 `--disable-radix-cache`, 而 HiCache 需要 `--enable-hierarchical-cache`, 两者同时出现会导致 SGLang 启动时抛出 `ValueError`。因此需要将 HiCache 作为生成器的一等选项, 方便读者直接通过界面生成正确的命令。

实现拆解

1. 新增选项组: 在 `qwen35-deployment.jsx` 中, 在 `speculative` 和 `mambaCache` 之间插入 `kvOffload` 选项组。该选项组通过 `condition` 控制仅当 `hardware` 为 `mi355x` 且 `quantization` 为 `fp4` 时显示, 包含 `Disabled` (默认) 和 `Host DRAM (HiCache)` 两个选项。
2. 修改命令生成逻辑: 在 `generateCommand` 函数中解构 `values` 时加入 `kvOffload`, 并在 FP4 专属参数部分根据 `kvOffload` 的值决定追加 HiCache 参数还是保留 `--disable-radix-cache`。当选择 HiCache 时, 会追加 `--enable-hierarchical-cache` 以及一系列 `--hicache-*` 参数, 并省略 `--disable-radix-cache`。
3. 文档更新: 在 `Qwen3.5.mdx` 的 `Configuration Tips` 部分新增 HiCache 条目, 说明其作用、参数含义以及需要的主机内存预算; 同时修正 FP8 KV Cache 警告, 明确参数已包含在 MXFP4-on-MI355X 配方中。
4. 验证: 通过手动执行生成器确认了多个组合下的输出命令均有效, 并附上了冲突情况表格。

<docs/src/snippets/autoregressive/qwen35-deployment.jsx>

该文件是配置生成器的核心实现, 新增了 `kvOffload` 选项组并修改了 `generateCommand` 函数, 以支持生成 HiCache 参数并解决与 `--disable-radix-cache` 的冲突。

关键源码片段

<docs/src/snippets/autoregressive/qwen35-deployment.jsx>

该文件是配置生成器的核心实现，新增了 kvOffload 选项组并修改了 generateCommand 函数，以支持生成 HiCache 参数并解决与 --disable-radix-cache 的冲突。

```
// 新增 kvOffload 选项组，仅对 MI355X + FP4 配方显示
kvOffload: {
  name: 'kvOffload',
  title: 'KV Cache Offloading',
  // HiCache 在设备 KV cache 之下增加 host-DRAM 层，仅在 MI355X MXFP4 配方上启用
  condition: (values) => values.hardware === 'mi355x' && values.quantization === 'fp4',
  items: [
    { id: 'disabled', label: 'Disabled', default: true },
    { id: 'hichache', label: 'Host DRAM (HiCache)', default: false }
  ]
}

// generateCommand 中处理 HiCache 参数生成
if (kvOffload === 'hichache') {
  // HiCache 需要 radix cache 保持开启，因此追加 --enable-hierarchical-cache 并省略 --disable-
  radix-cache
  cmd += ' \\
--enable-hierarchical-cache';
  cmd += ' \\
--hichache-ratio 1.5';
  cmd += ' \\
--hichache-write-policy write_through';
  cmd += ' \\
--hichache-io-backend direct';
  cmd += ' \\
--hichache-mem-layout page_first_direct';
} else {
  // 默认行为：保留 --disable-radix-cache
  cmd += ' \\
--disable-radix-cache';
}
```

评论区精华

本 PR 无实质性讨论。两位 reviewer (1am9trash、sogalin) 直接批准 (LGTM)，未提出具体意见或疑虑。

风险与影响

- 风险极低：改动仅涉及文档和配置生成器 JSX，不涉及运行时代码，无回归风险。
- 潜在风险：若 HiCache 参数值与 SGLang 实际支持的参数不匹配，生成器可能产生无效命令；但已通过手动测试验证，且参数值来自 InferenceX 配方，风险可控。
- 影响范围：仅影响使用 Qwen3.5 部署生成器的用户，特别是 MI355X MXFP4 用户，帮助其正确配置 HiCache，避免启动失败；对其他用户无影响。

关联脉络

- 与 InferenceX #2693 关联：该 PR 增加了 HiCache 配置，本 PR 使其在 cookbook 中成为一等选项。
- 与 PR#35445 关联：该 PR 为 MXFP4-on-MI355X 配方添加了 `--kv-cache-dtype fp8_e4m3`，但未更新文档，本 PR 修正了该遗留问题。
- 整体方向：增强 cookbook 对 AMD MI355X 平台的部署支持，特别是面向长上下文的 Agent 场景。