PR #25369 完整报告

sgl-project/sglang

Add hicache feature in dsv4 cookbook

合并时间：2026-05-15 15:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/25369

执行摘要

本 PR 为 DeepSeek-V4 的部署文档和交互式命令生成器新增了 HiCache（层次化 KV 缓存）功能选项。用户通过 UI 切换即可启用 L2 模式，生成的命令自动包含所有必要参数和环境变量。cookbook 中同步增加了 HiCache 说明小节。

功能与动机

HiCache 是 SGLang 提供的 multi-tier KV cache offload 功能，可将冷 KV 页面从 GPU 卸载到 CPU（甚至存储），显著扩大有效上下文长度。此前用户需要手动添加 --enable-hierarchical-cache 等六个启动参数和三个环境变量，门槛较高。本 PR 将这些参数集成到部署命令生成器中，用户只需在 UI 中选择“HiCache → L2”即可生成正确命令。

实现拆解

1. 在部署命令生成器 UI 中添加选项（deepseek-v4-deployment.jsx）：在 options 对象末尾新增 hicache 字段，提供“Disabled”（默认）和“L2”两个选项。
2. 在命令生成逻辑中注入参数：在 generateCommand 函数中解构出 hicache 值，然后在 H200 FP4 和 Blackwell 两条命令路径中，当 hicache === 'l2' 时向 flags 数组 push 六个 --hicache-* 参数，并设置环境变量 SGLANG_ENABLE_UNIFIED_RADIX_TREE=1。
3. 更新 cookbook 文档（DeepSeek-V4.mdx）：新增 4.2.3 小节，简要介绍 HiCache 的分层架构（L2/L3）和与 UnifiedRadixTree 的协作，并引导用户使用上方 UI 切换。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

在交互式部署命令生成器 UI 中新增 HiCache 选项（disabled/l2），并在命令生成逻辑中为 H200 FP4 和 Blackwell 路径添加对应的启动参数和环境变量。

```
// 在 options 对象中新增 HiCache 选项（位于 toolcall 之后）
hicache: {
  name: "hicache",
  title: "HiCache",
  items: [
    { id: "disabled", label: "Disabled", default: true },
    { id: "l2", label: "L2", default: false, subtitle: "GPU+CPU" },
  ],
},

// 在 generateCommand 函数中解构出 hicache
```

```
const { hardware: rawHardware, modelSize, recipe, reasoningParser, toolcall, hicache } =
values;

// H200 FP4 路径中根据 hicache 选项添加参数
if (hicache === "l2") {
  fp4Flags.push(" --enable-hierarchical-cache");
  fp4Flags.push(" --hicache-ratio 2");
  fp4Flags.push(" --hicache-size 0");
  fp4Flags.push(" --hicache-write-policy write_through");
  fp4Flags.push(" --hicache-io-backend direct");
  fp4Flags.push(" --hicache-mem-layout page_first_direct");
}

// 同时设置环境变量（在拼接命令前）
const fp4Env = [];
if (hicache === "l2") fp4Env.push("SGLANG_ENABLE_UNIFIED_RADIX_TREE=1");
const fp4EnvBlock = fp4Env.length ? fp4Env.join(" \\\n") + " \\\n" : "";
const fp4Cmd = `${fp4EnvBlock}sglang serve \\\n${fp4Flags.join(" \\\n")}`;
```

评论区精华

无 review 讨论或评论。

风险与影响

- 风险：仅涉及文档和 UI 配置，不修改任何运行时代码，风险极低。
- 影响：用户可通过 UI 一键启用 HiCache，降低使用复杂度。功能默认关闭，不影响现有用户。

关联脉络

- 本 PR 为 PR#24691（DeepSeek V4 HiCache 支持）的配套文档，将新功能暴露给用户。
- 部署生成器文件 deepseek-v4-deployment.jsx 在近期被 PR#25317 修改过（后回退），需注意保持同步。