PR #36519 完整报告

sgl-project/sglang

GLM-5.3-Flash cookbook: default Blackwell recipes to FP8 KV + TRT-LLM DSA

合并时间：2026-08-27 00:51

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36519

执行摘要

GLM-5.3-Flash cookbook 的这次更新将 Blackwell 默认部署配方从 BF16 + TileLang DSA 切换为 FP8 KV + TRT-LLM DSA（实测吞吐提升 2.3-5.5%、KV 容量约 1.8 倍、GSM8K 精度在噪声范围内），同时修复了此前所有单元格缺失 --disable-shared-experts-fusion 的问题（该缺失会导致模型在最终权重上退化），并根据 HiCache 与 MTP 的启动崩溃问题在 Low Latency 策略下禁用了 HiCache 选项。变更还同步刷新了基准数据、披露了测量服务器使用的额外标志，并记录了 HiCache 的开销。作为文档配置类 PR，它不影响运行时代码，但对用户部署行为和模型可用性有直接影响。

功能与动机

PR body 说明这是 #36513 的后续 corrective batch：默认切换提交在合并时被遗漏，且之后发布的 cookbook 又暴露三个问题：

- 共享专家融合导致模型退化：所有单元格缺失 --disable-shared-experts-fusion，开启融合时最终权重上模型生成空内容、重复循环、永不停止；作者在 4x GB300 上用隔离实验确认仅此标志不同即可复现。
- HiCache + MTP 启动崩溃：AttributeError: DSATokenToKVPool has no attribute full_kv_pool，因此在 low-latency 策略下禁用 HiCache 选项，直到代码修复。
- HiCache L1+L2 数据缺失：补上了 High Throughput 下的开销测量。

实现拆解

1. 配置默认值切换（docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx）：kvDsaPair 默认值改为 fp8-trtllm，并将 isRecommendedSelection 改为硬件条件化：Hopper（H100/H200）上由于 FP8 KV + TRT-LLM 不受支持，推荐仍为 bf16-tilelang；其余硬件（GB300 等）推荐为新版默认。verificationStatus 也放宽为两种配对在有效组合下均显示 verified。
2. HiCache 禁用逻辑（同一文件）：为 hicache 的 l2 和 l3 选项引入 disabled 条件（s.strategy === "low-latency"）和 disableReason，说明 MTP 与 HiCache 的兼容问题，仅保留 High Throughput 下可用。
3. 恢复 fusion 禁用标志（同一文件 flags）：在 deploy 单元格命令中恢复 --disable-shared-experts-fusion，并同步更新基准 notes 说明该标志是测量时的必要条件。
4. 基准数据与文档同步（glm-5.3-flash-benchmarks.jsx、GLM-5.3-Flash.mdx、_deployment.jsx）：重新测量 LL 两行（数值微降，但一致性更好），HT 行补充额外启动

标志与 HiCache 开销；mdx 更新 Precision 描述、KV/DSA 配对段和徽章说明；Reproduce modal 摘要加入 kvDsaPair 显示。

5. 验证配套：作者完成多项验证（组合命令断言、GSM8K 对比、HiCache 启动测试、cookbook 配置检查），但未新增独立测试文件，主要依赖回归检查脚本。

docs/src/snippets/configs/zai-org/glm-5.3-flash.jsx

核心配置对象：切换默认 KV/DSA 配对、硬件条件化推荐、禁用 HiCache 于 Low Latency、恢复 fusion 标志，影响部署面板生成的命令。

```
export const config = {
  modelName: "GLM-5.3-Flash",

  supportedHardware: ["gb300", "h100", "h200", "b200", "b300", "gb200"],

  // 推荐选择现在按硬件区分：Hopper 上 FP8 KV + TRT-LLM 不受支持，
  // 因此回退到 BF16 + TileLang；其余 Blackwell 硬件默认推荐 FP8 + TRT-LLM。
  isRecommendedSelection(s) {
    const pairing = ["h100", "h200"].includes(s.hw) ? "bf16-tilelang" : "fp8-trtllm";
    return (
      s.kvDsaPair === pairing &&
      s.mmTransport === "auto" &&
      s.hicache === "off"
    );
  },

  overlayDims: [
    {
      id: "kvDsaPair",
      title: "KV Cache + DSA Backend",
      default: "fp8-trtllm",
      options: [
        {
          id: "fp8-trtllm",
          label: "FP8 + TRT-LLM",
          // Hopper 上禁用该组合，并给出明确的不可用提示。
          disabled: (s) => ["h100", "h200"].includes(s.hw),
          disableReason: "FP8 KV cache with TRT-LLM DSA is not supported on Hopper GPUs.",
          stripPrefixes: ["--kv-cache-dtype", "--dsa-prefill-backend", "--dsa-decode-backend"],
          flags: [
            "--kv-cache-dtype fp8_e4m3",
            "--dsa-prefill-backend trtllm",
            "--dsa-decode-backend trtllm",
          ],
          hints: ["Measured on GB300: faster than BF16 + TileLang with about 1.8x the KV token
          capacity."],
        },
        {
          id: "bf16-tilelang",
```

```
        label: "BF16 + TileLang",
        stripPrefixes: ["--kv-cache-dtype", "--dsa-prefill-backend", "--dsa-decode-backend"],
        flags: [
          "--kv-cache-dtype bfloat16",
          "--dsa-prefill-backend tilelang",
          "--dsa-decode-backend tilelang",
        ],
      },
    ],
  },
  {
    id: "hicache",
    title: "HiCache",
    default: "off",
    options: [
      { id: "off", label: "Off" },
      {
        id: "l2",
        label: "L1 + L2",
        subtitle: "Host memory",
        flags: ["--enable-hierarchical-cache", "--hicache-size 32"],
        // 当前构建中 HiCache 与 MTP 推测解码会在启动时崩溃，
        // 因此 Low Latency 策略下禁用，避免用户遇到 AttributeError。
        disabled: (s) => s.strategy === "low-latency",
        disableReason: "HiCache with MTP speculative decoding crashes at startup in the
        current build (DSA draft pool lacks full_kv_pool); use it with High Throughput only.",
        hints: ["32 GB host tier; the default ratio can demand more host RAM than the node has
        free."],
      },
      {
        id: "l3",
        label: "+ L3",
        subtitle: "Mooncake",
        flags: ["--enable-hierarchical-cache", "--hicache-size 32", "--hicache-storage-backend
        mooncake"],
        env: ["SGLANG_HICACHE_MOONCAKE_CONFIG_PATH={{MOONCAKE_CONFIG}}"],
        // 与 L1 + L2 相同的原因，Low Latency 下禁用。
        disabled: (s) => s.strategy === "low-latency",
        disableReason: "HiCache with MTP speculative decoding crashes at startup in the
        current build (DSA draft pool lacks full_kv_pool); use it with High Throughput only.",
        hints: ["Start Mooncake and place the configuration file on every serving node."],
      },
    ],
  },
],
// … 其余配置（modelNames、placeholders、curl 等）不变
};
```

评论区精华

本 PR 没有公开的 review 评论线程，两位维护者（zijiexia、ShangmingCai）直接通过。从提交历史看，作者在评审流程中自行完成了多项修正（补充 HT FP8 精度、披露额外标志、恢复 fusion 标志、重测 LL 行），体现了基于实测数据驱动文档收尾的严谨态度。

风险与影响

- 默认配置变更：Blackwell 用户部署时默认会使用 FP8 KV + TRT-LLM，虽然实测精度在噪声范围内，但不同负载下仍需验证，文档已用 hints 提示。
- fusion flag 修复：如果 flags 列表再次被精简或遗漏，模型会退化，需在后续维护中防止该标志被移除。
- HiCache 临时禁用：依赖 runtime 修复（DSA draft pool 增加 full_kv_pool），修复后需及时更新 cookbook，否则用户会错过可用的 HiCache 能力。
- 基准数据可复现性：数字均来自特定 rc2 commit 和附加启动标志，用户复现时可能因版本 / 环境不同而有偏差，notes 中已做披露。
    - 影响范围限于文档与部署面板配置，不涉及运行时代码，总体风险可控。

关联脉络

本 PR 是 GLM-5.3-Flash cookbook 系列（#36440、#36513）的收敛补丁，与 HiCache 修复线（#36317 等）相关——其禁用的根因是 DSA draft pool 缺少 full_kv_pool 属性，与 HiCache 的 pending ownership 修复同属统一缓存层演进方向。该系列也体现了 sglang 团队以 cookbook 为入口管理模型部署默认配置的模式，后续类似模型的 cookbook 可能沿用“硬件条件化推荐 + 实测数据背书”的写法。