PR #36544 完整报告

sgl-project/sglang

GLM-5.3-Flash cookbook: HiCache for LL, fusion-flag drop, EAGLE, default-cell numbers, DCP4 overlay

合并时间：2026-08-28 03:18

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/36544

GLM-5.3-Flash cookbook：基于实测证据的配置与基准全面修订

执行摘要

本 PR 是一次面向用户的 cookbook 修订：在 #36538 等运行时修复进入 release image 后，重新启用 Low Latency 策略下的 HiCache 选项，移除 --disable-shared-experts-fusion，统一 EAGLE 拼写，并新增 DCP4 上下文并行 overlay。基准表整体锚定到 release-image 树 d6ab04bdf1 的实测值，新增两条 DCP4 性能行，文档与配置面板的数值口径完全对齐。变更集中在 3 个文件，CI 与 cookbook 配置校验均通过，属于低风险高价值文档更新。

功能与动机

PR body 明确这是对 #36519 的 follow-up，核心动机是“用最终权重上的实测证据校正 cookbook”：

1. HiCache for Low Latency：裸 DSA draft pool 的启动崩溃已在 release image 与 #36538 中修复，重新启用 L1+L2 选项（L2 host pool 打包 target 11 + draft 1，GSM8K-20 95% 且 100% stop rate）。
2. --disable-shared-experts-fusion 不再需要：fusion-gate fix 使得无 flag 命令在 EP>1 时构建 unfused 路径，实测 3/3 探针 + GSM8K-20 20/20 通过。
3. EAGLE 拼写：上游把 NEXTN 折叠进 EAGLE，生成命令改用规范拼写（topk=1 行为不变）。
4. HT 数字改为默认 cell 实测：原先“不可复现”的注记被删除，因为自动 sizing 在 GB300 上可容纳 434 个 running requests，concurrency 256 行可直接按默认参数复现。

实现拆解

1. 重新启用 HiCache 低延迟选项：glm-5.3-flash.jsx 中删除 l2/l3 的 disabled 与 disableReason；依赖 #36538 的 DSATokenToKVPool unwrap 修复。
2. 移除 fusion flag 并统一 EAGLE：gsm8k 与 sgl-eval 命令模板删除 --disable-shared-experts-fusion，--speculative-algorithm 改为 EAGLE。
3. 新增 DCP4 overlay：overlayDims 增加 dcp 维度（off / 4），仅 gb300 可选；isRecommendedSelection 与 verificationStatus 同步纳入 dcp === "off" 或 ["off", "4"] 判断。

4. 重测并重写基准表：glm-5.3-flash-benchmarks.jsx 版本锚点更新为 d6ab04bdf1，新增 FP8 + TRT-LLM DCP4（1,680.61 out tok/s）与 BF16 + TileLang DCP4（1,565.8 out tok/s）两行；HT 三档数据改为发布命令默认 cell 测量（BF16 1,161 / 2,660 / 4,828，FP8 1,227 / 2,739 / 4,977 output tok/s）。配套校验：check_cookbook_configs.mjs 通过，PR CI 与 extra CI 通过。
5. 同步 MDX 叙述：新增 Decode context parallelism 小节，FP8 vs BF16 吞吐差统一为 2.9–5.7%，补充 EAGLE 拼写说明；无新增测试文件，依赖脚本与 CI 门禁。

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

配置面板核心：移除 HiCache 对 low-latency 的禁用、删除 fusion flag、新增 dcp 维度并联动推荐 / 验证状态，直接决定用户生成命令。

```
// isRecommendedSelection 与 verificationStatus 的联动（head 版）
// 新增 dcp 维度后，推荐 / 已验证状态都要求 DCP 为 off，
// 避免把只在 4x GB300 上验证的 DCP4 组合标记为 recommended / verified。
isRecommendedSelection(s) {
  const pairing = ["h100", "h200", "mi300x", "mi325x", "mi355x"].includes(s.hw)
    ? "bf16-tilelang"
    : "fp8-trtllm";
  return (
    s.kvDsaPair === pairing &&
    s.mmTransport === "auto" &&
    s.hicache === "off" &&
    s.dcp === "off"
  );
}

// overlayDims 中的 HiCache 与 DCP 维度（head 版）
// 1) HiCache 的 L1+L2 / L3 选项删除了对 low-latency 的 disabled 与 disableReason，
// 前提是 release image 已内置 #36538 的 DSA draft pool 修复
//  （L2 host pool 按 target 11 + draft 1 打包，GSM8K-20 95%、100% stop rate）。
// 2) 新增 dcp 维度：DCP4 仅 gb300 可选，flags 一次性注入
// --dcp-size / --dcp-comm-backend / --dcp-replicate-q-proj。
overlayDims: [
  // ...mmTransport 维度省略 ...
  {
    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"],
        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}}"],
        hints: ["Start Mooncake and place the configuration file on every serving node."],
      },
    ],
  },
  {
    id: "dcp",
    title: "Context Parallelism",
    default: "off",
    options: [
      { id: "off", label: "Off" },
      {
        id: "4",
        label: "DCP 4",
        // 目前只在 4x GB300 TP4/EP4 上完成验证，其他硬件直接置灰
        disabled: (s) => s.hw !== "gb300",
        disableReason: "DCP is validated only on 4x GB300 TP4/EP4 for now.",
        flags: ["--dcp-size 4", "--dcp-comm-backend a2a", "--dcp-replicate-q-proj"],
        hints: ["Measured on 4x GB300 with both KV/DSA pairings, adaptive MTP 5/1/6, full
        decode graph."],
      },
    ],
  },
]
```

评论区精华

该 PR 没有 inline review 评论或 issue 评论，仅有一条 zijiexia 的 APPROVED（空 body）审核。不过 commit 历史显式记录了 review follow-ups 的落地：

> Review follow-ups：fp8-vs-bf16 delta 统一为 2.9–5.7%（MDX 与 notes 一致），KV 池绝对值采用默认 cell 值，fp8 stop rate 采用 d6ab 门禁数据。

这说明作者的“实测数据必须与发布命令、release image 严格绑定”的修订原则已经过审核闭环。

风险与影响

- 运行时修复依赖：HiCache（LL）与 TileLang DSA DCP4 分别依赖 #36538 与 release image 的 LSE fix；旧 build 上生成的命令可能复现启动崩溃或精度偏差。
- 基准数据版本绑定：所有数字绑定 d6ab04bdf1 树与最终权重 c5b82b63e37b，其他版本不可直接引用。

- 配置联动面：isRecommendedSelection / verificationStatus 与 dcp 维度联动，未来扩展 DCP 值（如 2/8）时需同步更新判断。
- 测试覆盖：无直接测试文件，cookbook 依赖 check_cookbook_configs.mjs 与 CI 门禁，运行时 flag 组合未被单元化验证。

影响范围集中在 zai-org/GLM-5.3-Flash cookbook：用户获得更准确的可选项与基准参考，团队建立了“发布 image + 原样命令 + 默认 cell 测量”的文档数字基线。

关联脉络

- #36538（关联 Issue）：HiCache 启动崩溃修复，是本 PR 重新开放 Low Latency + HiCache 的前提；PR body 明确引用该修复。
- #36660：同一 cookbook 功能线的后续修正，改动同一组文件，继续打磨 flag 与基准数据。
- #36519：本 PR 的前置 PR（body 明示 follow-ups），提供初版配置面板与基准行。
  - 与该系列并行，仓库近期有大量 diffusion cookbook 与 DSA/HiCache 运行时修复，说明 cookbook 正逐步变成“运行时能力 + 验证数据”双绑定的选型工具。