PR #29486 完整报告

sgl-project/sglang

[Cookbook] GLM-5.2: tune GB300 NVFP4 recipes + fill benchmarks

合并时间：2026-06-27 16:01

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29486

PR 分析报告：GLM-5.2 NVFP4 食谱与基准数据更新

执行摘要

本 PR 为 GLM-5.2 模型的 NVFP4 量化变体更新了 GB300 平台上的部署食谱和基准数据。核心工作包括：优化 low-latency 策略的显存占比参数、为 balanced 和 high-throughput 策略引入 DP-Attention 及调整近似解码参数、移除冗余的 --trust-remote-code 标志，并填补了之前留空的性能基准数据（TTFT、TPOT、吞吐量）和 AIME25 准确率。变更仅涉及 Cookbook 文档的配置代码片段，不修改任何运行时代码。

功能与动机

PR Body 明确指出此 PR "Builds on the GB300 NVFP4 scaffolding already in main (#29380 + #29466) with the newly-measured single-node 4×GB300 (TP4) recipes and benchmark numbers for nvidia/GLM-5.2-NVFP4"。目的是为 GB300 NVFP4 提供经过实际测量验证的优化部署配置和性能参考数据，将原先的占位符替换为实测值，并统一移除不必要的 --trust-remote-code 标志以与其他量化方式保持一致。

实现拆解

1. 更新食谱文件（glm-5.2.jsx）：
   - low-latency 策略：将 mem-fraction-static 从 0.8 提升到 0.85，以更充分利用 GB300 的显存资源，同时保持 MTP 5-1-6 不变。
   - balanced 策略：从裸 TP4 升级为启用 DP-Attention（dp4），并调整 MTP 参数为 2-1-3，附注释说明“在此并发度下，长 MTP 的验证开销超过接受长度收益”。同时将 mem-fraction-static 提升至 0.92，设置 max-running-requests 为 256。
   - high-throughput 策略：新增一个完整的策略配置，同样使用 DP-Attention（dp4） 和 mem-fraction-static 0.92，并将 max-running-requests 提升到 512，不启用近似解码以最大化吞吐。
   - 从所有 B300 和 GB300 的 NVFP4 单元格中移除 --trust-remote-code 标志，使其与其他量化路径一致。
2. 填充基准数据文件（glm-5.2-benchmarks.jsx）：
   - 为 GB300 NVFP4 的三种策略（low-latency、balanced、high-throughput）填补了之前仅作为占位符存在的基准数据条目。

- 每个条目包含了 sglang_version（dev-glm52-nvfp4）、aime25_pct 准确率（89.58%）以及在 ISL/OSL 8192/1024 工作负载下不同并发度（1、16、64、256、1024）的 ttft_ms、tpot_ms 和 tokens_per_sec_per_gpu 实测值。
- 移除了"benchmarks pending"的注释标记。

3. AIME25 准确性校正：最后一个提交将 GB300 NVFP4 的 AIME25 得分从 91.04% 调整为 89.58%，与 PR Body 中描述的实测数据保持一致，并注明仍高于 FP8/BF16 变体默认值 87.7%，且处于运行间正常方差范围内。

docs_new/src/snippets/configs/zai-org/glm-5.2.jsx

核心食谱文件，包含 GB300 NVFP4 三种部署策略的完整配置参数变更，是部署者直接引用的关键配置源。

```
// ================================================================
// NVFP4 (Blackwell Ultra) — nvidia/GLM-5.2-NVFP4 (Model Optimizer). TP4.
// B300: low-latency + balanced (the 4-GPU GB300 node fits the ~381 GB build).
// GB300: low-latency / balanced / high-throughput measured on a single 4xGB300
// node — balanced & high-throughput add DP-Attention (dp4); low-latency uses MTP 5-1-6.
// ================================================================
{
  match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "low-latency", nodes: "single"
  },
  verified: true,
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp 4",
    "--quantization modelopt_fp4",
    "--speculative-algorithm EAGLE",
    "--speculative-num-steps 5",
    "--speculative-eagle-topk 1",
    "--speculative-num-draft-tokens 6",
    "--chunked-prefill-size 8192",
    "--mem-fraction-static 0.85", // 从 0.8 提升至 0.85，充分利用 GB300 显存
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
{
  match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },
  verified: true,
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp 4",
    "--quantization modelopt_fp4",
    "--dp 4", // 新增 DP-Attention 以提升并发吞吐
    "--enable-dp-attention",
    // 较短的 MTP 2-1-3：在此并发度下，长 MTP 的验证开销超过接受长度收益
```

```
      "--speculative-algorithm EAGLE",
      "--speculative-num-steps 2",
      "--speculative-eagle-topk 1",
      "--speculative-num-draft-tokens 3",
      "--chunked-prefill-size 8192",
      "--mem-fraction-static 0.92", // 更高显存利用率
      "--max-running-requests 256",
      "--host {{HOST_IP}}",
      "--port {{PORT}}",
    ],
  },
  {
    match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "high-throughput", nodes:
    "single" },
    verified: true,
    env: [],
    flags: [
      "--model-path {{MODEL_NAME}}",
      "--tp 4",
      "--quantization modelopt_fp4",
      "--dp 4", // DP-Attention 用于高吞吐场景
      "--enable-dp-attention",
      "--chunked-prefill-size 8192",
      "--mem-fraction-static 0.92",
      "--max-running-requests 512", // 最大并发请求数 512
      "--host {{HOST_IP}}",
      "--port {{PORT}}",
    ],
  },
```

docs_new/src/snippets/configs/zai-org/glm-5.2-benchmarks.jsx

基准数据文件，提供了 GB300 NVFP4 三种策略在不同并发度下的实测性能数据（TTFT、TPOT、吞吐量），是用户评估部署性能的关键参考。

```
  // ---- GB300 + NVFP4 ---- (4-GPU single node, TP4; nvidia/GLM-5.2-NVFP4 via --quantization
  modelopt_fp4,
  // measured on the lmsysorg/sglang:dev-glm52-nvfp4 preview image, flush-cache every run.
  // tokens_per_sec_per_gpu = total server output tok/s / 4 GPUs (337→84, 1248→312,
  1162→291, 1695→424, 1730→433).
  // aime25 overrides the variant default (87.7 → 89.58, measured on this NVFP4 build); gsm8k
  inherits the default.)
  {
    match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "low-latency", nodes: "single"
    },
    sglang_version: "dev-glm52-nvfp4",
    accuracy: { aime25_pct: 89.58 },
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 238, tpot_ms: 2.23, tokens_per_sec_per_gpu: 84 },
```

```
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 315, tpot_ms: 11.9, tokens_per_sec_per_gpu: 312 },
    ],
  },
  {
    match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },
    sglang_version: "dev-glm52-nvfp4",
    accuracy: { aime25_pct: 89.58 },
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 64 },
        ttft_ms: 1169, tpot_ms: 58, tokens_per_sec_per_gpu: 291 },
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 256 },
        ttft_ms: 6389, tpot_ms: 167, tokens_per_sec_per_gpu: 424 },
    ],
  },
  {
    match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "high-throughput", nodes:
    "single" },
    sglang_version: "dev-glm52-nvfp4",
    accuracy: { aime25_pct: 89.58 },
    speed: [
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1024 },
        ttft_ms: 156000, tpot_ms: 321, tokens_per_sec_per_gpu: 433 },
    ],
  },
```

评论区精华

本 PR 没有产生实质性 review 讨论，仅有一个批准和两个 bot 自动回复（Mintlify 预览部署和 Gemini Code Assist 配额警告），无人工技术交锋。

风险与影响

- 风险：变更局限于两个 .jsx 配置片段文件，不涉及任何 Python 或 CUDA 运行时代码，因此回归风险极低。移除 --trust-remote-code 若模型加载依赖此选项可能导致失败，但 PR 说明模型已原生集成，且其他量化路径未使用，风险可控。所有配置均标记为 verified: true。
- 影响：直接影响 GLM-5.2 模型在 GB300 NVFP4 场景下的文档化部署指引，用户可直接参考这些配置和基准进行部署。对系统运行时无影响。

关联脉络

本 PR 直接关联 #29466（更新 GLM-5.2 B300 和 GB300 NVFP4 的 Cookbook 设置）和 #29380（为 Cookbook 添加 NVFP4 框架和配方结构）。#29466 建立了基础的配置框架和 B300/GB300 的浅层配置，本 PR 在此基础上进一步用实测数据优化 GB300 的参数并填补了基准数据空白。这展示了 SGLang Cookbook 项目中对新硬件 / 量化组合进行“框架搭建 → 参数调优 → 基准填充”的典型演进模式。

与同仓库近期 PR 中的其他 Cookbook 相关变更（如 #29466）共同表明团队正系统性地为 GLM-5.2 NVFP4 模型在 Blackwell 架构（B300/GB300）上完善部署文档和性能数据。