PR #29674 完整报告

sgl-project/sglang

docs: add B200 NVFP4 recipes + benchmarks to GLM-5.2 cookbook

合并时间：2026-06-30 05:06

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/29674

执行摘要

本 PR 为 GLM-5.2 cookbook 新增了 B200 + NVFP4（TP8）的三种部署配方（低延迟、均衡、高吞吐）及其 P50 基准性能数据。纯文档变更，无代码逻辑改动，合入后用户可直接在 cookbook 页面查阅 B200 上的 NVFP4 量化部署方案。

功能与动机

原 cookbook 已支持 B300/GB300 上的 NVFP4 部署（#29557），但缺少 B200 硬件上的指导。本 PR 补全这一空白，为使用 nvidia/GLM-5.2-NVFP4 模型的用户提供经过验证的启动参数和实测性能数据，降低部署调优成本。

实现拆解

1. Docker 镜像映射：在 glm-5.2.jsx 的 dockerImages 对象中增加 "b200nvfp4" 键，指向专用预览镜像 lmsysorg/sglang:dev-glm52-nvfp4（该镜像包含 modelopt_fp4 支持）。
2. 部署配置单元：在 glm-5.2.jsx 的配置数组末尾插入三个对象，分别对应三种策略。关键区别：
   - 低延迟：TP8 + EAGLE MTP 5-1-6，chunked-prefill-size 8192，mem-fraction-static 0.85，无 DP-Attention。
   - 均衡：TP8 + --dp 8 --enable-dp-attention，MTP 2-1-3（较短草稿以降低验证开销），chunked-prefill-size 32768，mem-fraction-static 0.92，max-running-requests 256。
   - 高吞吐：TP8 + DP-Attention，无推测解码，chunked-prefill-size 32768，mem-fraction-static 0.92，max-running-requests 512。
3. 基准测试数据：在 glm-5.2-benchmarks.jsx 中追加三个 benchmark 条目，记录 P50 的 TTFT、TPOT 和每 GPU 吞吐量。所有数据在指定预览镜像上测得，数据集为随机文本（ISL=8192, OSL=1024）。

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

核心配置文件：添加了 B200 NVFP4 的 Docker 镜像映射和三个部署策略的启动参数。

```
// Docker 镜像映射：B200 NVFP4 需要 dev-glm52-nvfp4 预览镜像
dockerImages: {
  h200: "lmsysorg/sglang:latest",
  b200: "lmsysorg/sglang:latest",
  gb300: "lmsysorg/sglang:latest",
  b300: "lmsysorg/sglang:latest",
```

```
  // NVFP4 需要 modelopt_fp4 支持，使用 dev 镜像
  "b200lnvfp4": "lmsysorg/sglang:dev-glm52-nvfp4",
  "b300lnvfp4": "lmsysorg/sglang:dev-glm52-nvfp4",
  "gb300lnvfp4": "lmsysorg/sglang:dev-glm52-nvfp4",
},

// B200 NVFP4 部署策略单元（NVFP4 量化，TP8）
{
  match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "low-latency", nodes: "single" }
  ,
  verified: true,
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp 8",
    "--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",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
{
  match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },
  verified: true,
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp 8",
    "--quantization modelopt_fp4",
    "--dp 8",
    "--enable-dp-attention",
    // 均衡策略使用较短的 MTP 2-1-3 以避免长草稿的验证开销
    "--speculative-algorithm EAGLE",
    "--speculative-num-steps 2",
    "--speculative-eagle-topk 1",
    "--speculative-num-draft-tokens 3",
    // 增大 chunked-prefill 以平衡批处理
    "--chunked-prefill-size 32768",
    "--mem-fraction-static 0.92",
    "--max-running-requests 256",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
```

```
{
  match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "high-throughput", nodes:
  "single" },
  verified: true,
  env: [],
  flags: [
    "--model-path {{MODEL_NAME}}",
    "--tp 8",
    "--quantization modelopt_fp4",
    "--dp 8",
    "--enable-dp-attention",
    "--chunked-prefill-size 32768",
    "--mem-fraction-static 0.92",
    "--max-running-requests 512",
    "--host {{HOST_IP}}",
    "--port {{PORT}}",
  ],
},
```

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

基准数据文件：新增 B200 NVFP4 的三个 benchmark 条目，包含低延迟、均衡、高吞吐的 P50 性能数据。

```
// ---- B200 + NVFP4 ----  (8-GPU 单节点，TP8；nvidia/GLM-5.2-NVFP4 通过 --quantization
modelopt_fp4,
// 使用 dev-glm52-nvfp4 预览镜像，每次运行前刷新缓存。
// ttft_ms/tpot_ms 为 P50；tokens_per_sec_per_gpu = 输出 tok/s/GPU。
// 均衡和高吞吐额外启用 DP-Attention（dp8）；低延迟使用 MTP 5-1-6，均衡使用 MTP 2-1-3）
{
  match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "low-latency", nodes: "single" }
  ,
  sglang_version: "dev-glm52-nvfp4",
  speed: [
    { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
      ttft_ms: 295, tpot_ms: 1.85, tokens_per_sec_per_gpu: 58.6 },
    { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
      ttft_ms: 2491, tpot_ms: 5.43, tokens_per_sec_per_gpu: 254.3 },
  ],
},
{
  match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },
  sglang_version: "dev-glm52-nvfp4",
  speed: [
    { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 64 },
      ttft_ms: 5837, tpot_ms: 12.70, tokens_per_sec_per_gpu: 418.9 },
    { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 256 },
      ttft_ms: 16736, tpot_ms: 30.00, tokens_per_sec_per_gpu: 593.7 },
  ],
},
```

```
{
  match: { hw: "b200", variant: "default", quant: "nvfp4", strategy: "high-throughput", nodes:
  "single" },
  sglang_version: "dev-glm52-nvfp4",
  speed: [
    { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1024 },
      ttft_ms: 130174, tpot_ms: 67.12, tokens_per_sec_per_gpu: 589.4 },
  ],
},
```

评论区精华

无实质性讨论。PR 由 Fridge003 直接批准，review 评论数为 0。

风险与影响

- 风险：Docker 镜像 lmsysorg/sglang:dev-glm52-nvfp4 为预览版，后续可能变更或下架导致配方失效；基准数据在特定环境测得，其他用户可能无法复现。
- 影响：仅影响 cookbook 文档页面，用户可获取新的部署方案，无需适配现有逻辑。

关联脉络

本 PR 是 #29557（B300 NVFP4 配方）的自然扩展，将相同的三种策略部署模型适配到 B200 硬件（TP8 而非 TP4，对应 8-GPU 单节点）。与 #29470（GLM-5 router GEMM 阈值调优）无直接关联。整体属于 cookbook 持续完善 Blackwell 系列覆盖的一部分。