PR #28460 完整报告

sgl-project/sglang

docs(cookbook): verify GLM-5.2 single-node B300 (FP8 + BF16)

合并时间：2026-06-17 11:47

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/28460

执行摘要

本 PR 在 8×B300 节点上实测了 GLM-5.2 全部 6 种单节点部署配置，并更新了 cookbook 中的性能数据和验证状态。B300 当前因内核未针对 sm103 优化，单 GPU 性能比 B200 低 15-40%，但配置可直接使用。

功能与动机

继 PR #28448 后，用户需要确认 B300 上的实际部署性能。PR #28460 通过实测将之前推断的 pending 配置标记为 verified，并提供了准确的 TTFT、TPOT、吞吐量数据，同时向社区公开了 B300 与 B200 的性能差异原因（deep_gemm/DSA 内核调度降级）。

实现拆解

1. 填充 benchmark 数据(glm-5.2-benchmarks.jsx)：为 B300 的 FP8/BF16 各 3 种策略（low-latency、balanced、high-throughput）添加 6 个带有实测性能的对象，每个对象包含 2 个并发级别的 workload。注释解释了 B300 与 B200 差异的根因。
2. 更新配置验证状态(glm-5.2.jsx)：将 B300 的 6 个单节点配置的 verified 从 false 改为 true，更新注释描述实测结果及限制（BF16 未使用 DP-Attention/DeepEP 时高并发吞吐较低）。
3. 完善文档描述(GLM-5.2.mdx)：在简介中增加 B300，并在 CP 段落说明 B300 无需额外配置。
4. 保留多节点 pending：9 个 BF16 多节点配置因缺乏 InfiniBand 环境，继续保持 verified: false，避免误导用户。

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

核心变更：为 B300 的 FP8/BF16 各 3 种策略填充实测 TTFT/TPOT 吞吐量数据，替换原先的 pending 占位对象。

```
// 文件：glm-5.2-benchmarks.jsx（片段）
// 注意：B300 (sm103) 因 deep_gemm/DSA 内核未优化，单 GPU 性能低于 B200 (sm100)
…
  // ---- B300 + FP8 ---- (8-GPU single node, TP8; measured on v0.5.13.post1)
  {
    match: { hw: "b300", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" },
    sglang_version: "0.5.13.post1",
    speed: [
      // 低并发：TTFT 503 ms, TPOT 3.24 ms, 34 tokens/s/GPU
```

```
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 1 },
        ttft_ms: 503, tpot_ms: 3.24, tokens_per_sec_per_gpu: 34 },
      // 16 并发：TTFT 4731 ms, TPOT 9.56 ms, 140 tokens/s/GPU
      { workload: { dataset: "random", isl: 8192, osl: 1024, max_concurrency: 16 },
        ttft_ms: 4731, tpot_ms: 9.56, tokens_per_sec_per_gpu: 140 },
    ],
  },
...
```

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

关键变更：将 B300 的 FP8/BF16 各 3 种策略的 verified 字段从 false 改为 true，并更新注释说明实测状态和已知限制。

```
// 文件：glm-5.2.jsx（片段）
// B300 + FP8 已验证，注释说明内核优化状态
  {
    match: { hw: "b300", variant: "default", quant: "fp8", strategy: "low-latency", nodes: "single" },
    verified: true, // 之前是 false
    env: {},
    flags: [
      "--model-path {{MODEL_NAME}}",
      "--tp 8",
      "--speculative-algorithm EAGLE",
      "--speculative-num-steps 5",
      "--speculative-eagle-topk 1",
      "--speculative-num-draft-tokens 6",
      "--mem-fraction-static 0.8",
      "--cuda-graph-max-bs 32",
      "--host {{HOST_IP}}",
      "--port {{PORT}}",
    ],
  },
...
```

评论区精华

无实质 review 讨论。审核者 zijiexia 直接批准。

风险与影响

- 风险：benchmark 数据基于单一节点和特定版本（v0.5.13.post1），其他环境可能表现不同。多节点配置未验证，需注意文档提示。
- 影响：为社区用户提供 B300 部署 GLM-5.2 的直接参考，减少试错成本。无代码变更，不影响服务运行时。

关联脉络

- 前置 PR #28448 创建了 GLM-5.2 cookbook 的初始版本（H200、B200、GB300），本 PR 在其基础上补充了 B300 实测数据。

- 相关硬件支持 PR #28433 (Ascend GLM-5.2 部署文档) 补充了 NPU 平台的支持。
- 整体往 GLM-5.2 多平台覆盖方向演进，从 B200/H200 扩展到 GB300、B300 和 Ascend。