PR #24367 完整报告

sgl-project/sglang

[docs] Update B300 Pro cookbook with accuracy-verified serving configs

合并时间：2026-05-05 14:26

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/24367

PR 分析报告 (24367)

执行摘要

本 PR 为 DeepSeek-V4 deployment cookbook 中的 B300 Pro（B200 big）硬件更新了经过准确性验证的推荐配置，涵盖三种使用场景。变更主要发生在 JSX 命令生成器文件，新增了大量环境变量和参数调整。但 reviewer 指出条件范围过宽的风险，该问题未在合并前解决。

功能与动机

根据 PR body，更新 B300 Pro 的部署配置使其通过 SimpleQA-Verified（1000 个事实 QA 样本）验证，确保 Pass@1 准确率达到官方 57.9% 基准（±2% 内）。三个配方（Low Latency、Balanced、Max Throughput）都针对 B300 Pro 做了个性化调整，同时保持其他硬件路径不变。

实现拆解

1. 环境变量（recipeEnv）扩展：在 JSX 的 recipeEnv 构建逻辑中，为 B200 | big（B300 Pro）的每个配方增加了不同的环境变量集合。Low Latency 添加了 6 个与 JIT 和 SWA 相关的变量；Balanced 添加了 13 个变量，包括禁用 DeepEP、启用 fast mask EP 等；Max Throughput 添加了 13 个变量，使用 deepgemm mega moe 并设置高 token 上限。这些变量来自内部的准确性测试，用于微调运行时行为。
2. 命令行标志（flags）调整：三个配方的关键参数在 B300 Pro 场景下被修改。例如，Low Latency 的 chunked-prefill-size 从 4096 提升到 8192，mem-fraction-static 从 0.88 改为 0.90，并新增 swa-full-tokens-ratio 0.1。Balanced 配方将 MoE runner backend 从 deepep 切换到 flashinfer_mxfp4 以减少延迟。Max Throughput 配方增加了更大的 batch size 和 prefill-delayer 等优化。
3. 文档辅助更新：在生成的命令中添加 # flags will be simplified 注释，并在页面新增 MegaMoE 使用说明段落。

docs_new/src/snippets/autoregressive/deepseek-v4-deployment.jsx

核心变更文件，为 B300 Pro 添加准确性验证的配置环境变量和标志，并调整多个配方分支。

```
// Balanced 配方环境变量设置（B200/B300 Pro 专用）
if (recipe === "balanced") {
  if (hardware === "h200") {
    // H200：根据大小设置 dispatch-token 上限
```

```
    recipeEnv.push(isBig
      ? "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=128"
      : "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=256");
  } else if (isBig && hardware === "b200") {
    // B200/B300 Pro 经过 SimpleQA 验证的环境变量（提升准确度并优化 MegaMoE）
    recipeEnv.push(
      "SGLANG_JIT_DEEPGEMM_PRECOMPILE=0", // 避免预编译
      "SGLANG_OPT_SWA_SPLIT_LEAF_ON_INSERT=1", // SWA 叶节点分离
      "SGLANG_OPT_USE_JIT_NORM=1", // JIT 归一化
      "SGLANG_OPT_USE_JIT_INDEXER_METADATA=1", // JIT 索引元数据
      "SGLANG_OPT_USE_TOPK_V2=1", // TopK v2 实现
      "SGLANG_OPT_USE_CUSTOM_ALL_REDUCE_V2=1", // 自定义 AllReduce v2
      "SGLANG_OPT_SWA_EVICT_DROP_PAGE_MARGIN=1", // SWA 逐出优化
      "SGLANG_OPT_USE_DEEPGEMM_MEGA_MOE=0", // 禁用 deepgemm mega moe
      "SGLANG_OPT_FIX_HASH_MEGA_MOE=0", // 修复 hash mega moe
      "SGLANG_OPT_USE_FAST_MASK_EP=1", // 快速 mask EP
      "SGLANG_OPT_FIX_MEGA_MOE_MEMORY=1", // 修复 mega moe 内存
      "SGLANG_OPT_DEEPGEMM_MEGA_MOE_NUM_MAX_TOKENS_PER_RANK=4096", // token
      上限
      "SGLANG_OPT_FIX_NEXTN_MEGA_MOE=1", // 修复 NextN mega moe
      "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=0", // 禁用 DeepEP 调度
    );
  } else {
    // 其他 Blackwell 硬件（小 B200、GB200、GB300）使用统一设置
    recipeEnv.push(isBig
      ? "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=256"
      : "SGLANG_DEEPEP_NUM_MAX_DISPATCH_TOKENS_PER_RANK=1024");
  }
}

// Low Latency 配方标志调整（针对 B300 Pro 优化）
if (recipe === "low-latency") {
  // ... 其他标志 ...
  flags.push("  --speculative-eagle-topk 1");
  flags.push("  --speculative-num-draft-tokens 4");

  // 注意：hardware !== "h200" 条件可能过宽，影响其他 Blackwell 硬件
  if (hardware !== "h200") {
    // B200/B300 Pro 准确性验证：big 使用 8192，small 保持 4096
    flags.push(isBig
      ? " --chunked-prefill-size 8192"
      : " --chunked-prefill-size 4096");
    flags.push("  --disable-flashinfer-autotune");
    // 对 big 场景设置 SWA 全 tokens 比例 0.1
    flags.push(isBig ? " --swa-full-tokens-ratio 0.1" : "");
  }

  if (isBig) {
    // mem-fraction-static 从 0.88 提升到 0.90（B300 Pro 已验证）
```

```
      flags.push(" --mem-fraction-static 0.90");
    }
  }
```

评论区精华

- gemini-code-assist[bot]指出条件 hardware !== "h200" 过于宽泛：「使用 hardware !== "h200" 会导致 gb200 和 gb300 也收到增加的 chunked-prefill-size 和新的 swa-full-tokens-ratio」。建议改为 isBig && hardware === "b200"。
- gemini-code-assist[bot]同样质疑 mem-fraction-static 的条件，强调只有 B300 Pro 应该设置为 0.90，否则影响其他平台。

这两个评论均未在合并前解决，建议后续关注并修复。

风险与影响

- 风险：条件 hardware !== "h200" 导致配置意外泄漏到 GB200、GB300 等硬件，可能引发性能回退或准确性下降。环境变量改动依赖特定运行时优化，如果内核实现变化可能导致配置失效。缺乏自动化测试验证。
- 影响：B300 Pro 用户获得官方验证的配置，提升部署可靠性。非目标硬件可能受配置泄漏影响，但人为检查可规避。整体影响中等，可接受。

关联脉络

本 PR 是 DeepSeek-V4 cookbook 持续优化的一部分，与之前关于 B200/GB200 等硬件的配置更新一脉相承。历史 PR（如 #23146 AMD EAGLE、#23321 specdec 优化）也涉及 DeepSeek 模型部署优化，但本 PR 更聚焦于 B300 Pro 的准确性验证配置。建议后续统一条件检查逻辑，避免类似范围问题。