PR #34863 完整报告

sgl-project/sglang

[Docs] Add GB300 cells and benchmarks for Qwen3.8-27B

合并时间：2026-08-14 23:42

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/34863

Qwen3.8-27B 文档补齐 GB300 基准与 Spark 修复

执行摘要

本 PR 是 #34860 的后续补丁，为 Qwen3.8-27B cookbook 页面补齐两个在源分支上错过 #34860 合并窗口的变更：新增 GB300（Blackwell Ultra SM103）单卡部署 cells 与配套的 106 行实测基准数据，并为 DGX Spark recipes 追加 --disable-prefill-cuda-graph。变更全部位于文档站点的 jsx 配置与 mdx 文本，无运行时代码影响。作者自述两个遗留问题——benchmarks 数据未接线渲染、high-throughput 策略在 4/5 平台上惰性——均未在本 PR 内修复，需要后续文档维护跟进。

功能与动机

34860 打开之后，源分支上又合入了两个变更，导致该 cookbook 页面当时不完整。作者在 PR body 中明确说明：

> Two changes landed on the source branch after that PR was opened, so they missed the cut。

具体是：GB300 单卡 cells 及其支撑的 benchmarks 文件、DGX Spark recipes 的 prefill CUDA graph 修复。本 PR 的目的就是补上这两个遗漏，使 Qwen3.8-27B 文档覆盖 GB300 这一关键 Blackwell Ultra 平台，并修正 DGX Spark 上 prefill CUDA graph 的可用性问题。

实现拆解

1. 扩展部署配置主文件docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx（+118/-3）：
   - supportedHardware 增加 gb300；strategies 策略轴新增 high-throughput，页面从单一 balanced 策略扩展为双策略。
   - dockerImages 增加 gb300: lmsysorg/sglang:dev——GB300 没有平台专用镜像 tag，直接指向滚动镜像，这带来数据可复现性风险（见风险章节）。
   - 新增六个 GB300 单卡 cell（BF16 / FP8 / NVFP4 × Balanced / High-Throughput），统一使用 --mem-fraction-static 0.85 与 --chunked-prefill-size 2048；high-throughput 三个 cell 额外携带 EAGLE 投机参数（--speculative-algorithm EAGLE、--speculative-num-steps 3、--speculative-eagle-topk 1、--speculative-num-draft-tokens 4），启用 checkpoint 内建 MTP 头。

- DGX Spark 三个 cell 均追加 --disable-prefill-cuda-graph，这是源分支上的 prefill CUDA graph 修复。

2. 新增基准数据文件docs/src/snippets/configs/Qwen/qwen3.8-27b-benchmarks.jsx（+106）：
   - 导出 benchmarks 数组，六个数据行以与 cell 相同的 match 元组为键（hw / variant / quant / strategy / nodes），这是 _deployment.jsx 中 speed / accuracy schema 的数据契约。
   - 每行包含三档并发（max_concurrency 1/16/64）的 ttft_ms、tpot_ms、tokens_per_sec_per_gpu 实测值，并锚定二进制版本 lmsysorg/sglang:dev @ c4271c3fe；NVFP4 行附带 GSM8K 精度（fp8-KV 96.44% / bf16-KV 96.82%），FP8 / BF16 精度未测保持 null。
3. 同步 cookbook 正文docs/cookbook/autoregressive/Qwen/Qwen3.8-27B.mdx（+3/-3）：在 DGX Spark 段落补充 --disable-prefill-cuda-graph 说明，与三个 cell 的新 flag 保持一致。
4. 验证配套：作者声明 node docs/scripts/check_cookbook_configs.mjs、mint validate、mint broken-links 均通过；无单元测试变更（纯文档站点改动）。注意 PR Test (Extra) 在描述中标记为失败，合并前未确认原因。

docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx

页面配置主文件：新增 gb300 硬件支持、high-throughput 策略轴、gb300 镜像映射和六个单卡 cell，并为 DGX Spark 三个 cell 追加 --disable-prefill-cuda-graph；high-throughput cell 通过 EAGLE 参数启用 MTP 投机解码。

```
// 策略轴新增 high-throughput：仅在 GB300 上有对应 cell，
// 其余四个平台（H200 / RTX PRO 6000 / RTX 5090 / DGX Spark）只有 balanced，
// 选择 high-throughput 时引擎会通过 dimension-fallback 弹回 balanced（见 _deployment.jsx）。
strategies: [
  { id: "balanced", label: "Balanced" },
  { id: "high-throughput", label: "High-Throughput" },
],

// GB300 没有专用镜像 tag，直接指向滚动镜像 lmsysorg/sglang:dev，
// 因此基准数据与部署配置存在版本漂移风险，数据锚定 commit c4271c3fe。
dockerImages: {
  "dgx-spark": "lmsysorg/sglang:qwen38-27b",
  gb300: "lmsysorg/sglang:dev",
},

// 以 NVFP4 两个 cell 为例：high-throughput 通过 EAGLE + MTP 头
// （--speculative-num-steps 3 / --speculative-eagle-topk 1 /
// --speculative-num-draft-tokens 4）换取吞吐，balanced 则不启用投机。
{
  match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" },
  verified: true,
  flags: [
    "--trust-remote-code", "--model-path {{MODEL_NAME}}",
```

```
      "--mem-fraction-static 0.85", "--chunked-prefill-size 2048",
      "--reasoning-parser qwen3", "--tool-call-parser qwen3_coder",
      "--host {{HOST_IP}}", "--port {{PORT}}",
    ],
  },
  {
    match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "high-throughput", nodes:
    "single" },
    verified: true,
    flags: [
      "--trust-remote-code", "--model-path {{MODEL_NAME}}",
      "--mem-fraction-static 0.85", "--chunked-prefill-size 2048",
      "--speculative-algorithm EAGLE", "--speculative-num-steps 3",
      "--speculative-eagle-topk 1", "--speculative-num-draft-tokens 4",
      "--reasoning-parser qwen3", "--tool-call-parser qwen3_coder",
      "--host {{HOST_IP}}", "--port {{PORT}}",
    ],
  },
```

docs/src/snippets/configs/Qwen/qwen3.8-27b-benchmarks.jsx

新增的基准数据文件，以与 cell 相同的 match 元组为键记录 GB300 实测性能与精度，是六个 GB300 cell 的数据支撑；但当前未在 mdx 中 import 接线，数据不渲染。

```
// Qwen3.8-27B 在 GB300 上的按 cell 基准数据，keyed by 与 qwen3.8-27b.jsx
// 相同的 match 元组（hw / variant / quant / strategy / nodes）。
// 六个数据行均为单 batch 测量：sglang.bench_serving --flush-cache、random 数据集、
// ISL=1024 / OSL=1024、--random-range-ratio 1、request-rate inf、
// max_concurrency 1/16/64、n=64/64/256，全部跑在同一二进制
// lmsysorg/sglang:dev @ c4271c3fe 上，保证横向可比。
// 注意：该 commit 在 GB300 上 attention 解析为 triton，
// 而更新的 c7c03ec+ 解析为 trtllm_mha，数据会随版本漂移。
export const benchmarks = [
  {
    // NVFP4 balanced：KV 走 fp8_e4m3（checkpoint 声明的 kv_cache_quant_algo）
    match: { hw: "gb300", variant: "default", quant: "nvfp4", strategy: "balanced", nodes: "single" }
    ,
    sglang_version: "lmsysorg/sglang:dev @ c4271c3fe",
    speed: [
      { workload: { dataset: "random", isl: 1024, osl: 1024, max_concurrency: 1, num_prompts:
      64 },
        ttft_ms: 79, tpot_ms: 6.4, tokens_per_sec_per_gpu: 155 },
      { workload: { dataset: "random", isl: 1024, osl: 1024, max_concurrency: 16, num_prompts:
      64 },
        ttft_ms: 715, tpot_ms: 8.5, tokens_per_sec_per_gpu: 1742 },
      { workload: { dataset: "random", isl: 1024, osl: 1024, max_concurrency: 64, num_prompts:
      256 },
        ttft_ms: 1216, tpot_ms: 13.6, tokens_per_sec_per_gpu: 4316 },
    ],
    // GSM8K 只在 NVFP4 checkpoint 上测过：fp8-KV 96.44%、bf16-KV 96.82%；
```

```
    // FP8 / BF16 的 GB300 精度尚未测量，对应行 accuracy 为 null。
    accuracy: { gsm8k_pct: 96.44 },
  },
];
```

评论区精华

该 PR 没有任何 reviewer 评论，仅有一条空 body 的 APPROVED（JustinTong0323）。最有价值的讨论来自作者在 Notes for reviewers 中的自述：

> The new benchmarks export is not wired into the page. Kimi-K3 and Inkling both do import { benchmarks } from ".../<model>-benchmarks.jsx" and pass <Deployment config={config} benchmarks={benchmarks} />; Qwen3.8-27B.mdx has no such import, so these 106 lines of measured data currently render nowhere.
>
> high-throughput only has cells on GB300. The other four hardware families are balanced-only, so selecting High-Throughput there snaps back to Balanced via the engine's dimension-fallback. No broken panel, but the toggle is inert on 4 of 5 platforms.

作者刻意保留贡献者的原始写法（'rather than edit someone else's work in a port'），把两个 one-liner 修复决定权留给维护者；两者均未在本 PR 内处理。

风险与影响

- 基准数据版本漂移：全部数字锚定 lmsysorg/sglang:dev @ c4271c3fe，且不 pin --attention-backend；该 commit 在 GB300 上 attention 解析为 triton，更新版本（c7c03ec+）解析为 trtllm_mha，用户用新版复现会得到不同数值。
- benchmarks 未接线：qwen3.8-27b-benchmarks.jsx 的 106 行数据当前不会在页面渲染，属于功能完整性缺口；后续补 import 时还需确认 _deployment.jsx 的 speed / accuracy schema 严格匹配。
- 策略回退易误导：5 个平台中 4 个平台的 High-Throughput 切换静默回落到 Balanced，用户可能误以为已启用高吞吐配置。
- 私有 checkpoint 不可复现：benchmarks 依赖私有 revision（RadixArk W4A4-0811 及官方私有 checkpoint），社区用户无法完整复现精度与性能。
- CI Extra 未通过：PR Test (Extra) 在描述中标记为失败，合并前未确认失败原因是否与 docs 变更相关。

影响范围限于文档站点：为 GB300 / Blackwell 用户提供可直接复制的单卡部署配置与实测性能参考（尤其 NVFP4 + MTP 组合），对 sglang 运行时零影响。

关联脉络

- #34860（[Docs] Add Qwen3.8-27B cookbook page）：本 PR 的直接前身。cookbook 页面首次落地时，GB300 cells、benchmarks 与 DGX Spark CUDA graph 修复已在源分支上合入但错过合并窗口，本 PR 补齐。

- #34809（[Cookbook] Add DeepSeek-V4-Pro-0813 serving recipes）：同样采用 benchmarks.jsx 数据文件模式的 cookbook 页面，且正确执行了 import { benchmarks } + <Deployment benchmarks={...} /> 接线，可作为 Qwen3.8-27B 接线方式的直接对照参考。

这两条线索共同揭示了 cookbook 站点的演进方向：以 -benchmarks.jsx 数据文件 + _deployment.jsx 通用渲染引擎为模式，为每个主流模型沉淀可复现的部署配置与实测数据。Qwen3.8-27B 页面的接线缺口是这一模式下需要补上的最后一块拼图。