PR #35121 完整报告

sgl-project/sglang

docs(cookbook): add Qwen3.8-27B DGX Spark configs

合并时间：2026-08-18 06:04

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35121

执行摘要

本 PR 将 Qwen3.8-27B cookbook 中三个 DGX Spark 部署 cell 从定制 SM121 运行点改为逐字复用 RTX PRO 6000 配方，并在 GB10 上实测 36/36 配置 boot-and-serve 全部通过后标记 verified。变更同时修正了文档 Note 与源码注释中自相矛盾的验证声明，确立了 cookbook 的 " 弱验证标准清晰声明 " 先例，对实际部署命令有直接影响。

功能与动机

PR body 明确指出，三个 DGX Spark cell 承载了定制 SM121 运行点（`--mem-fraction-static 0.95`、`--chunked-prefill-size 8192`、`--disable-prefill-cuda-graph`），但同一注释却声称 "Unvalidated on SM121 / aarch64"，且定制设置没有测量依据；`--disable-prefill-cuda-graph` 在 #34863 引入时，注释同样声明配方未验证。

核心论据是两条：一是两个平台同属 SM12x Blackwell 架构，且 cell 早已共享 `--attention-backend flashinfer`；二是 GB10 的 128GB 统一内存大于 RTX PRO 6000 的 96GB，"a recipe that fits the smaller card has headroom on the larger one"。PR body 还修正了早期描述错误：移除 `--disable-prefill-cuda-graph` 并非 no-op，实测会让启动多花 54.09s 的 CUDA graph capture 时间，但 24 个已测量 cell 均正常服务，且 RTX PRO 6000 配方本就不带该 flag。

实现拆解

1. 替换部署配方（docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx）：将 DGX Spark 三个 cell 的 flags 从 `--mem-fraction-static 0.95`、`--chunked-prefill-size 8192`、`--disable-prefill-cuda-graph` 改为 `0.85`、`2048`、移除该 flag，使两个 SM12x 平台共享同一条命令；每个 cell 补充实测说明并设置 `verified: true`。
2. 更新验证元信息注释：把 `DGX Spark stays unverified` 改为 "measured across its whole overlay envelope … to a weaker standard (boot-and-serve only)"；后续 re-audit 又修正了 cells[] 头部注释残留的 unverified 声明，消除与徽章的对立。
3. 同步文档正文（docs/cookbook/autoregressive/Qwen/Qwen3.8-27B.mdx）：Deploy 面板 Note 为 DGX Spark 增加独立条款，声明 boot-and-serve 弱验证标准；§2 平台说明删除 8192 chunk 特例描述，补充 GB10 复现注意事项（Docker GPU 访问为 CDI-only、`nvidia-smi` 不支持统一内存查询、改用 `/proc/meminfo` 的 `MemAvailable` 判断）。
4. 校验与验证：本地执行 `node docs/scripts/check_cookbook_configs.mjs` 通过；全部 36 个配置在真实硬件上 boot-and-serve 成功后才翻转 `verified` 徽章。无专门自动化测试变更，验证依赖实测。

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

核心变更文件：DGX Spark 三个 cell 的部署命令从定制 SM121 运行点改为逐字复用 RTX PRO 6000 配方，并基于 36 项实测标记 verified。

```
// DGX Spark (GB10, SM121)：单节点，128GB 统一内存与 CPU 共享。
// 这些 cell 复用 RTX PRO 6000 配方，而不使用独立的 SM121 运行点：
// 两块卡同属 SM12x Blackwell，且 GB10 的 128GB 大于 6000 的 96GB，
// 小卡能跑通的配方在大卡上有余量。
//
// 已在 GB10 (SM121 / aarch64) 上验证：全部 36 个配置均完成 boot-and-serve
// （ISL 8192 / OSL 1024，并发 1）。这比 SM120 那对的验证标准弱——没有
// 吞吐量或接受长度数字——Deploy 面板的 Note 也如此声明。
{
  match: { hw: "dgx-spark", variant: "default", quant: "nvfp4", nodes: "single" },
  // 12 个 overlay 组合全部在 GB10 上服务成功；DSPARK 还覆盖了该
  // checkpoint 量化的 4-bit lm_head，未出现形状错误。
  verified: true,
  env: [],
  flags: [
    '--trust-remote-code',
    '--model-path {{MODEL_NAME}}',
    '--kv-cache-dtype fp8_e4m3',
    // 0.85 是统一内存池比例，与主机 OS / 页缓存共享，权衡不同于专用显存
    '--mem-fraction-static 0.85',
    '--attention-backend flashinfer',
    // 2048 与 RTX PRO 6000 一致；不再使用 8192 与 --disable-prefill-cuda-graph
    '--chunked-prefill-size 2048',
    '--reasoning-parser qwen3',
    '--tool-call-parser qwen3_coder',
    '--host {{HOST_IP}}',
    '--port {{PORT}}',
  ],
},
```

评论区精华

zijiexia：“This contradicts the <Note> under the Deploy panel (lines 59-65), which still tells readers that on platforms other than the 5090 / 6000 the non-default Speculative Decoding / Serving Strategy / SSM dtype picks are 'valid but unmeasured' — right next to the badges this PR flips to Verified. Please update it, and keep DGX Spark in its own clause.”

Jiminator：“Fixed in 91dcdd7. The Note now gives DGX Spark its own clause and states the weaker standard explicitly — boot-and-serve at ISL 8192 / OSL 1024, concurrency 1, with no throughput or acceptance-length numbers taken.”

Jiminator（re-audit）：“One more in 0141367 — a re-audit for the same class of problem found the cells[] header comment still saying 'DGX Spark stays unverified'.

> .. It now points at the boot-and-serve standard documented on the cell block."

风险与影响

- 验证口径偏弱：36/36 仅为并发 1 的 boot-and-serve，无吞吐量与接受长度数据，文档却展示 Verified 徽章；读者可能高估 DGX Spark 性能，需要依赖 Note 的弱标准声明来纠偏。
- 统一内存与专用显存差异：0.85 比例基于与主机 OS / 页缓存共享的 128GB 统一池，PR body 明确 "it is not a strictly-safer setting in every respect"，极端内存负载下可能不足。
- 启动延迟上升：移除 --disable-prefill-cuda-graph 后，prefill graph 后端从 disabled 变为 breakable，GB10 上启动多付 54.09s capture 时间，对频繁重启的部署有影响。
- EAGLE + bfloat16 状态精度风险：引擎会告警 closed-loop fold 重量化可能导致长序列漂移，该 caveat 未写入 §2，等待后续跟进。
- 文档配置直接影响用户部署命令：错误配方会被直接复制使用，属于低代码风险、高用户影响。

关联脉络

本 PR 是 Qwen3.8-27B 部署网格重构（#35065）的后续补充，与其配套的 mamba 缓存比例计算器修复（#35064）共同完善了该页面的多平台部署能力。它还与 Kimi-K3 部署面板（#35168）共用 verificationStatus 机制，体现了 cookbook 文档配置在验证状态管理上逐渐形成统一模式。未来值得关注的是 EAGLE + bfloat16 状态精度 caveat 是否会进入文档，以及该 " 配方复用 + 分层验证 " 范式是否会推广到其他多平台部署页面。