PR #35825 完整报告

sgl-project/sglang

[docs] Re-measure the Qwen3.8-27B RTX 5090, RTX PRO 6000 and DGX Spark grids on 1cf2b8c

合并时间：2026-08-22 13:03

原文链接：http://prhub.com.cn/sgl-project/sglang/pull/35825

Qwen3.8-27B 部署配方统一重测报告

执行摘要

本 PR 把 Qwen3.8-27B 文档页从“DFLASH2 需源码构建、其余走旧镜像、RTX 5090 pin 过时”的三套割裂状态，收敛为单一 commit 1cf2b8c（#35496）+ 单一多架构镜像 lmsysorg/sglang:dev-qwen38-27b-dflash2，并在 RTX 5090、RTX PRO 6000、DGX Spark 三张卡上重测全部发布 cell，让每个 pin 都来自实测而非继承。核心 pin 变化集中在 RTX 5090（EAGLE bf16 0.92→0.93、DSpark bf16 0.90→0.88、DFlash2 bf16 0.88→0.91 并追加 --chunked-prefill-size 1024、DSpark/DFlash2 的 fp32 组合置灰）与 DGX Spark（基础 mem-fraction 0.85→0.80，规避 earlyoom SIGTERM）。纯文档与配置变更，不涉及运行时代码。

功能与动机

PR body 说明动机：“The Qwen3.8-27B page was split across builds: DFLASH2 needed a source build, everything else used lmsysorg/sglang:qwen38-27b, and the RTX 5090 pins predated both. This unifies the page on one commit — 1cf2b8c (#35496) — and one image, and re-runs every published cell on the three cards that can be measured here, so each pin is a measurement rather than an inheritance.”

第二个动机与 checkpoint 换版强相关：NVFP4 cell 用 BF16-lm_head 导出（RadixArk/Qwen3.8-27B-NVFP4-BF16-LMHead）验证，其 head 增加约 3.2 GB 权重，正好解释了 32 GB 卡上 pin 的整体移动；密集头也原生解锁了 DFlash2/DSpark 的 selector 投影路径，是后续性能与精度收益的前提。

实现拆解

实现按 5 步推进（对应 7 个 commit 的演进顺序）：

1. 统一安装路径（docs/cookbook/autoregressive/Qwen/Qwen3.8-27B.mdx）：删除“DFLASH2 needs a build from 1cf2b8c”与“Picking DFLASH2 requires SGLang built from main”两个 Warning 块；Docker 路径改为直接 docker pull lmsysorg/sglang:dev-qwen38-27b-dflash2（多架构、基于 1cf2b8c 构建）；源码路径统一 git checkout 1cf2b8c54d81802abc15dcf23a29b9cc687bc01e 后 uv pip install。原因为新镜像已包含 DFlash2 selector 支持，DFLASH2 不再需要特殊构建。

2. 重测 RTX 5090 全部 cell 并重写 pin（docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx 的 spec 面板）：spec.eagle.flags 中 bf16 pin 从 0.92 升到 0.93——密集 lm_head 使权重更重，状态池需要更大静态预算；spec.dspark.flags 从按 dtype 分发的 0.90/0.92 收敛为固定 bf16 单 pin 0.88——草稿模型叠加自动 prefill CUDA-graph capture 后 0.90 已放不下；spec.dflash.flags 改为 0.91 + --chunked-prefill-size 1024——0.91 时各 pool 恰好放下，但 2048 token 的 prefill 激活值放不下，chunk 降到 1024 后成为全卡最快配方（TPOT 4.92 ms、accept 4.29），同时删除 fp32 专属的 --mamba-full-memory-ratio 10 覆盖路径。
3. 扩大 fp32 置灰范围（ssmDtype.float32.disabled）：从原 dflash && low-latency 扩大为 dflash || dspark。作者在 0.86-0.96 全区间、两种 prefill chunk 尺寸、balanced-ratio 覆盖到 20 的扫描中确认：低于约 0.92 状态池达不到目标 slot 数，高于则 graph capture 或首个请求 OOM，fp32 与草稿模型在 32 GB 卡上已无共存点；EAGLE（replayssm 使池极小）与 no-speculation（不加载草稿权重）不受影响。
4. 重测 DGX Spark 网格并降低基础 mem-fraction：48 个 cell（3 种量化 × none/EAGLE/DSPARK/DFLASH2 × 双 tier × 双 SSM dtype）在 1cf2b8c 上端到端 boot-and-serve 通过，DFLASH2 的“Final Verification In Progress”徽标移除；基础 mem-fraction 0.85→0.80。根因是统一内存同时计价宿主内存：0.85 的 128 GB 只给 OS 留约 8 GB，恰好撞上 DGX OS earlyoom 的 SIGTERM 阈值（exit -15、无 traceback），0.85 下 48 个 cell 有 15 个被杀，0.80 下全部通过；文档同步给出 journalctl -u earlyoom 排查指引。
5. 统一镜像映射并清理陈旧说明：dockerImages 中 rtx6000 / rtx5090 / dgx-spark 全部指向 dev-qwen38-27b-dflash2，h200 / gb300 保持原镜像（未重测）；最后提交删除过时的 H200/GB300 验证说明。配置与测试层面无配套改动（docs only），CI 只跑 mintlify 文档预览。

测量标准本身也有说明价值：所有 cell 在 ISL 8192 / OSL 1024、concurrency 1、4 个 prompt、--random-range-ratio 1.0 下，以“exit 0 且恰好 4 个成功请求、32768 输入 token、4096 生成 token”作为通过判据。

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

部署命令生成器的核心配置，承载全部 mem-fraction pin、fp32 置灰逻辑、dockerImages 镜像映射以及逐条实测依据注释，是本 PR 数据重测结果的主要落点。

```
// docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx（head 版本整理）
// 该文件是部署面板的命令生成配置：flags 函数负责把用户选择渲染成最终
// sglang serve 命令，stripPrefixes 负责在切换选项时剥离旧 pin 再重发。

// DFLASH2 草稿模型选项：RTX 5090 上重测后的唯一特殊 cell。
{
  id: "dflash", label: "DFLASH2",
  // 32GB 的 RTX 5090 上 DFlash2 只能叠加在 NVFP4 权重之上，其余卡全部放开。
  disabled: (sel) => sel.hw === "rtx5090" && sel.quant !== "nvfp4",
  disableReason:
    "On the 32GB RTX 5090 the DFlash2 draft model only fits on top of the NVFP4 weights",
  // 切换时统一剥离 mem-fraction，按本卡实测值重新下发，避免旧 pin 残留。
```

```
  stripPrefixes: (sel) =>
    sel.hw === "rtx5090" ? ["--mem-fraction-static"] : [],
  flags: (sel) => [
    "--speculative-algorithm DFLASH",
    "--speculative-draft-model-path incoai/Qwen3.8-27B-DFlash2",
    "--speculative-num-draft-tokens 8",
    // 1cf2b8c 实测：0.91 时各 pool 恰好放下，但 2048 token 的 prefill 激活值
    // 放不下，故这是全页面唯一需要把 prefill chunk 降到 1024 的 cell。
    // 该组合是全卡最快配方：TPOT 4.92 ms，accept length 4.29。
    ...(sel.hw === "rtx5090"
      ? ["--mem-fraction-static 0.91", "--chunked-prefill-size 1024"]
      : []),
  ],
}

// Mamba SSM Dtype 面板的 float32 选项：RTX 5090 上置灰范围覆盖两种草稿模型。
{
  id: "float32", label: "float32",
  // 密集 lm_head 让权重增加约 3.2 GB，把状态池顶进 prefill CUDA graph
  // capture 放不下的区间。作者扫过 0.86-0.96、两种 chunk 尺寸并叠加
  // balanced-ratio 覆盖到 20：低于约 0.92 时状态池达不到目标 slot 数，
  // 高于时 graph capture 或首个请求即 OOM。EAGLE 与 no-speculation 不受
  // 影响：replayssm 让 EAGLE 状态池很小，no-speculation 不加载草稿权重。
  disabled: (sel) =>
    sel.hw === "rtx5090" && (sel.spec === "dflash" || sel.spec === "dspark"),
  disableReason:
    "On the 32GB RTX 5090 an fp32 GDN state pool and a speculative draft model " +
    "do not fit together — use bfloat16",
  flags: ["--mamba-ssm-dtype float32"],
}

// Docker 镜像映射：三张 SM12x 卡统一到同一镜像，其构建 commit（1cf2b8c）
// 与本页所有 pin 的测量基准一致，DFLASH2 不再需要特殊构建。
dockerImages: {
  h200: "lmsysorg/sglang:qwen38-27b", // 未重测，保留原镜像
  rtx6000: "lmsysorg/sglang:dev-qwen38-27b-dflash2",
  rtx5090: "lmsysorg/sglang:dev-qwen38-27b-dflash2",
  // 多架构（linux/amd64 + linux/arm64），GB10 可直接拉取。
  "dgx-spark": "lmsysorg/sglang:dev-qwen38-27b-dflash2",
  gb300: "lmsysorg/sglang:dev", // 未重测，保留原镜像
}
```

评论区精华

审核人 zijiexia 直接 APPROVED 且无正文，页面也没有 review 内联评论；真正的技术讨论发生在 Issue 评论区（合入者 JustinTong0323 补充）：

> checkpoint 更新说明：RadixArk/Qwen3.8-27B-NVFP4 已发布 BF16 lm_head 的新 revision 496d00f2（旧 revision 554ebba 保留在仓库历史），动机是“a dense head

> unblocks the DFlash2/DSpark selector path natively and improves draft acceptance”。体积 +1.7 GB，这也是 RTX 5090 各 cell 配 --mamba-ssm-dtype bfloat16 的原因。A/B 测试（DFlash2 drafter、4x GB300）：新 ckpt + bf16 SSM 的 GSM8K 96.36%、Terminal-Bench 69.66%；旧 ckpt + fp32 SSM 为 95.98% / 67.42%。
>
> 数字口径更正：Terminal-Bench 应使用 harbor 报告的均值，官方口径为 69.4%（247/356, pass@4 87.6%） vs 66.7%（234/351，截掉 5 个 heavy long-tail trials），结论不变：bf16 lm_head + bf16 SSM 组合不差，约 +2.7 pts。

风险与影响

- pin 与 commit/checkpoint 强绑定：全部新 pin 基于 1cf2b8c 与 BF16-lm_head 的 NVFP4 checkpoint（revision 496d00f2）。用户若解析到旧 revision 554ebba（packed-head）或使用更新构建，mem-fraction 数字不再保证成立。PR body 自述 NVFP4 cells “a step ahead of the checkpoint”，需要与 HF 仓库换版同步合并。
- dev 前缀镜像 tag 漂移：统一指向的 lmsysorg/sglang:dev-qwen38-27b-dflash2 是浮动 dev tag，文档未钉死 digest；镜像一旦重建覆盖，用户会拿到与 pin 测量基准不一致的构建。
- DGX Spark 结论依赖宿主环境：0.80 的结论建立在 DGX OS earlyoom 阈值与 128 GB 统一内存假设上；换发行版、改内核参数或增大非模型内存占用后可能失效（文档已给出 journalctl -u earlyoom 排查手段，属缓解措施）。
- 复制粘贴命令的隐性约束：RTX 5090 全部 cell 带 --max-running-requests 1 / --cuda-graph-max-bs 1 单请求约束，--chunked-prefill-size 1024 偏离引擎默认；cell 内已有 warn 提示，但并发用户直接复用仍有 OOM 风险。
  - 无运行时代码变更，引擎回归风险为零；风险集中在文档准确性这一层。

关联脉络

- #35371：DFLASH2 首次落地，被本 PR 删除的 Warning 引用（“DFlash2 landed in #35371”）；本 PR 删除其特殊构建要求，正是因为该功能已进入统一镜像。
- #35496：commit 1cf2b8c 所来自的 PR，引入 DFlash2 的 dense-lm_head 支持；本 PR 全部新 pin 均在该 commit 上重测，是整页的测量基准（标题据上下文推断）。
- checkpoint 换版：RadixArk/Qwen3.8-27B-NVFP4 的 BF16-lm_head revision 是 RTX 5090 新 pin 的前提，两个仓库的换版必须与本文档合并同步完成；这是本 PR 比“普通文档更新”更值得注意的编排点。
  - 与近期 diffusion 系列 PR（如 #36067、#36036）无直接关联，但共享同一“测量驱动发布”的团队流程：新功能或新 checkpoint 进入文档前，先在目标硬件上按固定协议重测。