

PR #35065 完整报告

sgl-project/sglang

docs(cookbook): Qwen3.8-27B deployment grid rework

合并时间: 2026-08-17 15:56

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/35065>

执行摘要

本 PR 重做 Qwen3.8-27B 的部署配置矩阵，用 RTX 5090 和 RTX PRO 6000 的实测数据替换此前基于假设的配方，并引入 `overlayDims` 机制将推测解码、服务策略和 SSM dtype 从匹配维度中拆出，避免组合爆炸。所有 cell 固定 fp8 KV 以换容量，并记录了精度 sign-off；benchmark 文件被显式标记为结构不匹配，仅作参考来源。整体是纯文档变更，不涉及运行时代码。

功能与动机

Qwen3.8-27B 部署矩阵原来只有单一定点配方，且部分配置（如 32GB 卡上的 EAGLE）实际无法启动。PR 以实测数据（8192 ISL / 1024 OSL，并发 1）为准重构矩阵，并解决两个问题：

- 让未测量平台（h200、gb300、dgx-spark）的命令在默认选择下与原始配方语义一致（`overlay` 默认等于引擎默认）。
- 记录 12 个 cell 统一固定 `--kv-cache-dtype fp8_e4m3` 的精度取舍（BF16/FP8 检查点无 fp8 KV 校准），依据 PR body 中的 sign-off 和 GB300 GSM8K A/B 数据点（96.82% vs 96.44%）。

同时，矩阵依赖 #35064 修复的 ratio calculator，否则会为推测配置输出错误的 `--mamba-full-memory-ratio`。

实现拆解

1. 配置结构重构（qwen3.8-27b.jsx）：用 `matchDims`（variant/quant/nodes）统一匹配维度，新增 `overlayDims` 承载 Speculative Decoding、Serving Strategy、Mamba SSM Dtype 三个正交可选项，避免 3 x 2 将 12 cell 膨胀为 72。`overlay` 默认值均对齐引擎默认（tier=low-latency、ssmDtype=float32），保证未改动选择不改变任何平台的命令。
2. 固定 fp8 KV：所有 cell 显式写入 `--kv-cache-dtype fp8_e4m3`。NVFP4 为透明 no-op；BF16/FP8 为容量 / 质量权衡，PR body 记录 sign-off。
3. 按小显存卡校准：RTX 5090 cell 增加 `--max-running-requests 1` 与 `--cuda-graph-max-bs 1`，并针对 EAGLE/DSPARK 与 dtype 分别设置 `--mem-fraction-static`（EAGLE fp32 0.94 / bf16 0.92，DSPARK fp32 0.92 / bf16 0.90）；删除 8192 prefill chunk 与 `--mamba-backend triton` `--linear-attn-backend triton`（后者在 ServerArgs 中已是默认，为 no-op）。

4. 文档配套 (Qwen3.8-27B.mdx) : 新增部署面板 <Note> 限定验证范围; 新增 --mamba-ssm-dtype 配置提示, 说明状态槽大小、速度非单向、ReplaySSM fp32 auto-select 与 drift 警告; 澄清 DSpark 的 D 推导 ($\gamma+1$) 与 ReplaySSM 的 $D=0$ 。
5. benchmark 文件保护 (qwen3.8-27b-benchmarks.jsx) : 因数据行 key 在已删除的 strategy 维度上, 文件头新增 STRUCTURALLY UNMATCHED 注释, 警告不可直接复用, 仅作测量来源。

关键源码片段

[docs/src/snippets/configs/Qwen/qwen3.8-27b.jsx](#)

核心配置文件, 重做部署矩阵, 引入matchDims/overlayDims, 承载所有测量数据与平台约束。

```
// 推测解码 overlay 行: EAGLE / MTP 选项。
// 该行不参与 cell 匹配, 而是叠加在已匹配的 cell 上, 避免组合爆炸。
{
  id: "eagle",
  label: "EAGLE",
  // 32GB 的 RTX 5090 上, MTP 头只有搭配 NVFP4 权重才放得下,
  // 因此其它量化组合下直接禁用该选项。
  disabled: (sel) => sel.hw === "rtx5090" && sel.quant !== "nvfp4",
  disableReason:
    "On the 32GB RTX 5090 the MTP head only fits on top of the NVFP4 weights",
  // EAGLE 与 DSPARK 在 5090 上需要的 mem-fraction 修正方向相反:
  // EAGLE 启动时饿死 state pool, 需要调高; DSPARK 饿死运行时激活, 需要调低。
  // 因此每个选项先移除 cell 里的固定值, 再重新写入自己的实测值。
  stripPrefixes: (sel) =>
    sel.hw === "rtx5090" ? ["--mem-fraction-static"] : [],
  flags: (sel) => [
    "--speculative-algorithm EAGLE",
    "--speculative-num-steps 3",
    "--speculative-eagle-topk 1",
    "--speculative-num-draft-tokens 4",
    // ReplaySSM 仅在有实测的 SM120/SM121 (rtx5090/rtx6000/dgx-spark) 上启用;
    // 它把 D 的中间 SSM 状态放到固定环上, 是 MTP 能塞进 32GB 的关键。
    // h200 (SM90) 和 gb300 (SM103) 保持普通 MTP 配方。
    ...(["rtx5090", "rtx6000", "dgx-spark"].includes(sel.hw)
      ? ["--enable-linear-replayssm-spec"]
      : []),
    // 5090 上的 mem-fraction 实测: bf16 状态 0.92 即可, fp32 需要 0.94
    // (fp32 槽位 146.81 MiB, bf16 槽位 74.81 MiB) 。
    ...(sel.hw === "rtx5090"
      ? [sel.ssmDtype === "float32"
        ? "--mem-fraction-static 0.94"
        : "--mem-fraction-static 0.92"]
      : []),
  ],
}
```

评论区精华

“The `_deployment.jsx` engine change is risky and could impact other cookbook... If the engine change is mandatory, could you raise it in a separate PR?” —— zijiexia 要求隔离引擎改动，Jiminator 最终撤销 selection-aware badge (#35112)，回归静态 verified。

“With the overlay defaults equal to the engine defaults, an untouched selection is the validated recipe on every platform” —— 讨论中确认默认值原则。

“`--mamba-backend triton --linear-attn-backend triton ...` Both fields already default to triton in `ServerArgs` ... Suggest dropping both” —— zijiexia 指出 no-op flag，Jiminator 确认并从所有 cell 删除。

“A published recipe capped at one concurrent request” —— 对 RTX 5090 bs=1 配方的质疑，Jiminator 保留并更新注释说明设计意图。

风险与影响

- 精度：BF16/FP8 检查点固定 fp8 KV 无校准，唯一准确率数据点（GB300 GSM8K）显示 96.82%→96.44% 的下降，文档已记录但容量受限用户可能忽略。
- 可部署性：RTX 5090 单并发配方如果用户提高并发，内存可能不足，虽然文档提示需同时调整 pins。
- 数据复用：benchmarks 文件结构不匹配，若后续 re-key 未重新测量会绑定到不同命令上。
- 未测量平台：h200/gb300 等平台的非默认 overlay 组合未验证，文档 Note 已限定范围。

影响范围限于文档 cookbook，无运行时代码变更；团队文档维护流程因 #35112 撤销而简化。

关联脉络

本 PR 与 #35064（ratio calculator 修复）直接相关，依赖后者保证

`--mamba-full-memory-ratio` 正确。讨论中提到的 #35112（selection-aware badge 原型）被撤销，使得本 PR 不再依赖 `_deployment.jsx` 引擎改动。这一系列 PR 显示 cookbook 正在从“单一配方”转向“实测 + overlay 维度”的部署指南，同时严格控制未验证组合的表达。