

PR #34797 完整报告

sgl-project/sglang

docs: link dots3.note checkpoints, add H100 cells

合并时间: 2026-08-14 10:42

原文链接: <http://prhub.com.cn/sgl-project/sglang/pull/34797>

执行摘要

这是一个纯文档 PR: 为 Dots3-Note cookbook 补上已公开发布的 Hugging Face checkpoint ID ([dots-studio/dots3-note-prev / -fp8](#))，删除所有“尚未发布”的占位提示，并把 H200 两个单元标记为已验证，同时新增 H100 硬件单元。为支持无 variant 维度的配置，[_deployment.jsx](#) 与 [_playground.jsx](#) 中的 `resolveModelName` 解析链同步扩展了 `hwlquant` 与裸 `quant` 回退。不涉及运行时代码，风险低，可由文档维护者直接合并。

功能与动机

PR body 明确写道: "Fills in the dots3.note checkpoint IDs on the cookbook page, now that the weights are public" (权重公开后，在 cookbook 页面填上 checkpoint ID)。原先页面因 checkpoint 未发布，使用占位符 "`<dots-note-checkpoint>`"，并带有两处 **Note** 说明“模型未发布、配方只在 PR #33829 验证过”，用户无法直接使用。此外作者希望为 H100 用户提供部署指引，并主动说明 BF16 权重 (537 GiB) 在 H100 8-GPU (640 GiB 总量、`--mem-fraction-static 0.87` 约 557 GiB 静态池) 下无法支撑 `--context-length 524288` 的 KV pool，引导用户使用 FP8 单元。

实现拆解

1. 配置数据更新 ([docs/src/snippets/configs/rednote/dots3-note.jsx](#)) - `modelNames` 从占位 `default` 改为按精度分键: `bf16` / `fp8` 分别指向真实 HF 仓库。 - `supportedHardware` 扩展为 `["h200", "h100"]`，`dockerImages` 增加 `h100`。 - H200 的 BF16/FP8 单元 `verified` 从 `false` 改为 `true`。 - 新增两个 H100 cell，flag 复刻 H200 对应精度单元，但 `verified: false`；FP8 单元维持 `--moe-runner-backend auto` / `--deepdp-dispatcher-output-dtype auto`，BF16 单元维持 `deep_gemm` / `bf16`。
2. 渲染逻辑配套 ([docs/src/snippets/_deployment.jsx](#) + [docs/src/snippets/_playground.jsx](#)) - 两处 `resolveModelName` 原键为 `hwlvariantlquant` → `variantlquant` → `hw` → `default` (Deployment) 或直接空串 (Playground)。因 Dots3-Note 的 `matchDims` 只有 `quant` 维，`sel.variant` 为 `undefined`，前两级永远命中不了，所以新增 `hwlquant` 与裸 `quant` 键，插在旧键之后以保证兼容；Deployment 侧保留 `hw` 与 `default` 兜底。 - [_deployment.jsx](#) 头部大注释同步更新 `modelNames` 键优先级说明。
3. cookbook 正文清理 ([docs/cookbook/autoregressive/RedNote/Dots3-Note.mdx](#)) - 删除两处“checkpoint 未发布”的 `<Note>` 与 TODO 注释。 - Resources 增加 Hugging Face 链接，并保留 PR #33829 链接。 - 将“single 8-GPU H200 node”改为“single 8-GPU

Hopper node”，并新增 H100 BF16 显存不足的说明段落。

4. 测试与部署配套：无测试文件变更。文档站由 `_deployment.jsx` / `_playground.jsx` 渲染，验证依赖文档站构建；PR CI Extra 失败与本次改动无直接关联，reviewer 已放行。

docs/src/snippets/_deployment.jsx

部署面板核心渲染逻辑所在文件，`resolveModelName` 用于将选择映射为 HF 模型名。本次新增 `hwlquant` 与裸 `quant` 两级回退，是无 `variant` 维度配置（如 Dots3-Note）能够正确渲染 `{{MODEL_NAME}}` 的关键，同时保持旧键顺序以兼容既有配置。

```
// 模型名解析：按最具体到最不具体的顺序查找 HF slug。
// 支持 matchDims 中没有 variant 维度的配置（此时 sel.variant 为 undefined，
// 前两级 key 永远无法命中），所以新增 hwlquant 与裸 quant 两级回退。
// 新键插在原有键之后，保证既有配置的解析结果不变。
const resolveModelName = (sel) => {
  const keys = [
    `${sel.hwl}${sel.variant}${sel.quant}`, // 完整三元组：hwlvariantlquant
    `${sel.variant}${sel.quant}`, // 无 hw：variantlquant
    `${sel.hwl}${sel.quant}`, // 新增：无 variant 时按硬件 + 精度
    sel.quant, // 新增：仅按精度（如 bf16 / fp8）
    sel.hwl, // 仅按硬件（兼容旧行为的回退）
    "default", // 最终兜底
  ];
  for (const k of keys) {
    const hit = config.modelNames[k];
    if (hit) return hit; // 命中即返回，避免 undefined 被当成有效值
  }
  return "";
};
```

docs/src/snippets/_playground.jsx

Playground 组件中维护了与 `_deployment.jsx` 独立的一份 `resolveModelName`，本次必须同步修改以保持一致，否则页面底部 Playground 的模型名替换会与部署面板产生行为分叉。

```
// hwlvariantlquant → variantlquant → hwlquant → quant → ""。
// 与 _deployment.jsx 保持同步：matchDims 无 variant 维度时旧键无法命中。
const resolveModelName = (sel) => {
  const keys = [
    `${sel.hwl}${sel.variant}${sel.quant}`,
    `${sel.variant}${sel.quant}`,
    `${sel.hwl}${sel.quant}`, // 新增：hw + quant
    sel.quant, // 新增：裸 quant
  ];
  for (const k of keys) {
    const hit = config.modelNames[k];
    if (hit) return hit; // 找到真实值才返回，避免空值覆盖
  }
  return "";
};
```

评论区精华

本 PR 没有任何 review 评论或讨论线程（`comments_count` 与 `review_comments_count` 均为 0），两位 reviewer [zijiexia](#) 与 [wisclmy0611](#) 直接 APPROVED，未留下文字意见。因此没有可提炼的交锋。唯一可讨论的隐含决策来自 PR body：H100 BF16 单元虽然知道显存不足，但只放在正文提示，没有使用单元格级 `warn` 横幅。

风险与影响

- H100 BF16 单元仍渲染完整启动命令：`_deployment.jsx` 头注释显示单元格支持 `warn` 横幅与 `disableReason`，但该 PR 未使用，用户可能直接复制跑挂。建议后续在 H100 BF16 单元加 `warn`（mdx 正文已有说明，但单元格内更醒目）。
- 双份 `resolveModelName` 易漂移：`_deployment.jsx` 与 `_playground.jsx` 各维护一份，本次已同步修改；将来若再扩展键序，两处必须同时改，否则 Deployment 与 Playground 行为分叉。
- `verified: true` 缺少验证数据支撑：PR 未附带吞吐 / 精度数据，标记“已验证”的客观依据不可追溯；若组织要求可复现证据，这点需要补充。
- H100 参数未实测：`--mem-fraction-static 0.87` 等 flag 直接复刻 H200，H100 上是否最优未经验证（作者已声明 `unverified`）。
- 外部依赖：`dots-studio/dots3-note-prev` 的 `-prev` 命名暗示可能后续替换为正式版本号，改名会导致文档失效，且 `modelName`s 删除了 `default` 兜底，解析会直接回退空串。

影响面：仅文档站。用户获得可用的部署命令与 H100 指引；文档系统的 `resolveModelName` 回退链向后兼容扩展，其他无 `variant` 维度的新配置页可受益。

关联脉络

本 PR 是 PR #33829（新增 Dots3-Note cookbook，标签注明 `do not merge`）的后续维护：原配方基于未发布 checkpoint 编写，本 PR 在权重公开后补全 ID 并扩展 H100。这与近期历史 PR（如 #34658 将 #33829 配方文档化）共同构成“Dots3-Note cookbook 从搭建到公开可用的逐步完善”这一演进线。